รายงานการวิจัยฉบับสมบูรณ์
เรื่อง
วิธีการแบบผสมสำหรับการสกัดลักษณะของชุดข้อมูลบนเครือข่าย เพื่อระบุผู้บุกรุกแบบเวลาจริง

(A Hybrid Method for Feature Extraction in Real-time Intrusion Detection)

โครงการวิจัยนี้ได้รับการสนับสนุนทุนวิจัย
จาก
สำนักงานคณะกรรมการวิจัยแห่งชาติ
ปีงบประมาณ พ.ศ. ๒๕๕๕

คณะผู้วิจัย

นายกฤษณะ ชินสาร	หัวหน้าโครงงานวิจัย
นางสาวสุวรรณา รัศมีขวัญ	ผู้ร่วมวิจัย
นางสาวสุนิสา ริมเจริญ	ผู้ร่วมวิจัย
นายภูสิต กุลเกษม	ผู้ร่วมวิจัย
นางสาวเบญจภรณ์ จันทรกองกุล	ผู้ร่วมวิจัย
นางสาวจรรยา อ้นปันส์	ผู้ช่วยนักวิจัย

ศูนย์วิจัย Knowledge and Smart Technology
คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา

บทคัดย่อ

วิธีการของการตรวจจับการบุกรุกสามารถแบ่งออกได้เป็น 2 ชนิด คือ วิธีการตรวจจับการบุกรุกแบบอโนมาลี (Anomaly Intrusion Detection Method) และวิธีการตรวจจับการบุกรุกแบบมิสยูส (Misuse Intrusion Detection Method) โดยที่วิธีการตรวจจับแบบอโนมาลีเป็นวิธีการหาผู้บุกรุกโดยการวิเคราะห์การใช้งานของผู้ใช้งาน หรือตัวระบบเองที่เบี่ยงเบนไปจากระดับการใช้งานโดยปกติส่วนการตรวจจับการบุกรุกแบบมิสยูสนั้น เป็นวิธีการหาผู้บุกรุกโดยการเปรียบเทียบข้อมูลที่เข้ามากับรูปแบบของผู้บุกรุกที่มีอยู่เดิม ซึ่งทั้งสองวิธีนี้มีจุดแข็งและจุดอ่อนที่แตกต่างกัน ปัญหาที่เด่นชัดที่สุดของการตรวจจับการบุกรุกแบบมิสยูส คือ ไม่สามารถตรวจจับการบุกรุกแบบใหม่ หรือการบุกรุกที่ไม่มีในชุดรูปแบบของผู้บุกรุกที่มีได้ ส่วนการตรวจจับการบุกรุกแบบอโนมาลีนั้น จะสามารถตรวจจับการบุกรุกจากผู้บุกรุกที่ไม่มีในฐานข้อมูลการบุกรุกได้ แต่ปัญหาที่สำคัญในการตรวจจับการบุกรุกแบบอโนมาลี คือ ทำอย่างไรถึงจะสร้างเค้าโครงของการใช้งานปกติที่ดีได้

ในงานวิจัยนี้ คณะผู้วิจัยได้แสดงให้เห็นแล้วว่าการหาตัวแทนที่เหมาะสม ได้แก่ การสกัดลักษณะและการเลือกลักษณะของชุดข้อมูลบนเครือข่าย มีความสำคัญต่อการพัฒนาการระบุผู้บุกรุกเป็นอย่างมาก ในการได้มาซึ่งตัวแทนชุดลักษณะของชุดข้อมูลที่เหมาะสม เพื่อใช้ในการระบุผู้บุกรุกโดยอาศัยวิธีการแบบผสมในการสกัดลักษณะและเลือกลักษณะของชุดข้อมูลเครือข่าย ซึ่งจะเพิ่มความสามารถในการระบุผู้บุกรุกได้ การพัฒนาการหาตัวแทนที่เหมาะสมบนชุดข้อมูลเครือข่ายประกอบด้วย 2 ขั้นตอน คือ 1.การหาลักษณะของชุดข้อมูลที่สามารถแทนข้อมูลได้และมีจำนวนลักษณะที่เหมาะสม และขั้นตอนที่ 2.การรู้จำแบบการบุกรุกเพื่อระบุผู้บุกรุกจากชุดข้อมูลบนเครือข่ายจากลักษณะที่ได้จากการสกัดลักษณะเปรียบเทียบกับการเลือกลักษณะของชุดข้อมูล โดยวัดประสิทธิภาพจากค่าร้อยละของความถูกต้อง (Accuracy) อัตราการตรวจจับ (Detection Rate) อัตราความผิดพลาดเชิงบวก (False Alarm Rate) ค่าเฉลี่ยเรขาคณิต (Geometric Means) อัตราความเร็วในการตรวจจับการบุกรุก และจำนวนข้อมูลที่แบ่งประเภทได้ถูกต้องของคลาสคำตอบจากผลการทดลองข้อมูล KDDcup99 จำนวน 13,499 จุดข้อมูล (Patterns) 34 ลักษณะ พบว่า วิธีการวิเคราะห์องค์ประกอบหลักสามารถสกัดลักษณะเด่นออกมาได้จำนวน 19 ลักษณะ และ วิธีฮิวริสติกกรีดดี ได้ผลการเลือกลักษณะข้อมูลจำนวน 13 ลักษณะ ผลการแบ่งกลุ่มข้อมูลด้วยวิธีการที่เลือกใช้ พบว่าการเลือกลักษณะด้วยวิธีฮิวริสติกกรีดดีให้ค่าความถูกต้องสูงกว่าการสกัดลักษณะเด่นด้วยวิธีการวิเคราะห์องค์ประกอบหลักด้วยวิธีการรู้จำแบบโครงข่ายประสาทเทียมแบบรัศมีฐานร้อยละ 5.06

Abstract

Detection of Network Intrusion can be categorized into two groups. The First one is "Anomaly Intrusion Detection Method". The second one is "Misuse Intrusion Detection Method". For the first method, it will be used to inspect the irregular behavior on the usage of the network or on the computer systems. For the second method is to inspect the mismatching with those patterns store in the database. According to the characteristic of comparing with the existing database, "Misuse Intrusion Detection Method", led to a discussion of improper way to detect the intrusion. This is because of the fact that intruders keep on changing their way to intrude the networks or computer systems. So the research question is how to find out proper features that represent well the normal behavior of traffic data.

In this research report, we have demonstrated how the use of feature selection on those traffic data will help in improving the detection of anomaly intrusion more efficient. There are two steps in detecting anomaly intrusion on traffic data. The first step is to extract features and select features. Then the use of pattern recognition algorithm to validate whether there is any anomaly behavior for those traffic data. The performance has been evaluate by comparing the percentage of accuracy, detection rate, false alarm rate, geometric means, speed in detecting the intrusion and the number of classes that correctly classified. It can be seen that from the KDDcup99 (with 13,499 sampling patterns) with 34 data dimensions based on HGIS and PCA algorithms, there are 19 and 13 features that have been extracted respectively. In addition, the classification accuracies confirm that HGIS algorithm produces better performance than PCA by 5.06% based on RBF recognition algorithm.

สารบัญ

เดียวกับปัญหาการแบ่งกลุ่ม (Classification Problem) โดยที่ปัญหาการแบ่งกลุ่มนั้นวิธีการที่ให้ผลลัพธ์ที่ดี คือ วิธีการเครือข่ายประสาทเทียม แต่อย่างไรก็ตามการตรวจจับการบุกรุกนั้นควรจะประมวลผลข้อมูลที่ต้องการตรวจสอบทั้งที่เป็นกรณีที่เป็นการบุกรุก และกรณีที่ไม่ใช่การบุกรุก ซึ่งการตรวจสอบดังกล่าว ทำให้ข้อมูลที่ตรวจสอบมีปริมาณมากทั้งจำนวนข้อมูล และจำนวนลักษณะของข้อมูลเป็นผลทำให้เกิดความล่าช้าในการระบุผู้บุกรุก และอาจเป็นสาเหตุให้การบุกรุกบางชนิดสามารถบุกรุกเข้าสู่ระบบเครือข่ายได้

จากที่ได้กล่าวมาทั้งหมดนั้น ผู้วิจัยได้แสดงให้เห็นแล้วว่าการสกัดลักษณะหรือการเลือกลักษณะของชุดข้อมูลบนเครือข่ายเพื่อให้ได้ตัวแทนชุดลักษณะของข้อมูลที่เหมาะสม มีความสำคัญต่อการพัฒนาการระบุผู้บุกรุกเป็นอย่างมาก จึงจำเป็นที่จะต้องหาวิธีการที่ดีในการหาตัวตัวแทนของชุดข้อมูลบนเครือข่าย หลังจากที่ผู้วิจัยได้ศึกษาการสกัดลักษณะข้อมูลบนเครือข่ายด้วยการวิเคราะห์องค์ประกอบหลักจึงพบว่าข้อมูลที่ได้จากการวิเคราะห์องค์ประกอบหลักนั้นไม่เหมาะสมเนื่องจากข้อมูลบนเครือข่ายมีการกระจายตัวมาก อาจทำให้บางลักษณะที่มีความสำคัญนั้นเปลี่ยนไปเป็นผลทำให้ภาพในกระบวนการการรู้จำมีประสิทธิภาพที่น้อยลงได้ และผู้วิจัยได้ศึกษางานวิจัยในด้านการเลือกลักษณะด้วยขั้นตอนวิธีฮิวริสติกกรีดดี (Heuristic Greedy Algorithm) ซึ่งเป็นการแก้ปัญหาในลักษณะที่ไม่มีรูปแบบวิธีการขั้นตอนโดยตรง โดยจะพิจารณาว่าข้อมูลที่มีอยู่ในขณะนั้นมีทางเลือกใดที่ให้คำตอบที่ดีที่สุดของปัญหาโดยจะใช้หลักการ Apriori เพื่อเลือกลักษณะที่สำคัญที่เป็นตัวแทนของลักษณะทั้งหมดและนำโครงข่ายประสาทเทียมใช้ในการรู้จำเพื่อเพิ่มความสามารถในการระบุผู้บุกรุกได้เหมาะสมมากกว่าการพัฒนาการตรวจจับผู้บุกรุกบนชุดข้อมูลเครือข่าย ประกอบไปด้วย 2 ขั้นตอน คือ 1.ขั้นตอนการหาลักษณะของชุดข้อมูลที่สามารถแทนข้อมูลได้และมีจำนวนลักษณะที่เหมาะสม และขั้นตอนที่ 2. การรู้จำรูปแบบการบุกรุกเพื่อระบุผู้บุกรุกจากชุดข้อมูลบนเครือข่าย จากลักษณะที่ได้จากการเลือกลักษณะด้วยขั้นตอนวิธีฮิวริสติกกรีดดีและเปรียบเทียบกับการสกัดลักษณะของชุดข้อมูลด้วยการวิเคราะห์องค์ประกอบหลัก โดยวัดประสิทธิภาพจากค่าร้อยละของความถูกต้อง อัตราการตรวจจับ อัตราความผิดพลาดเชิงบวก ค่าเฉลี่ยเรขาคณิต อัตราความเร็วในการตรวจจับการบุกรุก และจำนวนข้อมูลที่แบ่งประเภทได้ถูกต้องของคลาสคำตอบ

1.2 วัตถุประสงค์ของโครงการวิจัย

1. เพื่อศึกษาเทคนิคการสกัดลักษณะเด่นของชุดข้อมูล
2. เพื่อศึกษาการเลือกลักษณะของข้อมูลที่เหมาะสม
3. เพื่อศึกษาการรู้จำลักษณะของการระบุผู้บุกรุกเครือข่าย
4. เพื่อเพิ่มประสิทธิภาพการการระบุผู้บุกรุกเครือข่าย
5. เพื่อให้ผู้ที่สนใจสามารถนำแนวความคิดที่นำเสนอ ไปศึกษาเพื่อทำการพัฒนาหรือประยุกต์ใช้ในงานวิจัยของตนเองต่อไป

1.3 ขอบเขตของโครงการวิจัย

การวิจัยครั้งนี้มุ่งที่จะศึกษาและพัฒนาการหาตัวแทนของข้อมูล โดยใช้วิธีการสกัดลักษณะและการเลือกลักษณะที่เหมาะสมกับชุดข้อมูลเครือข่ายและการรู้จำเพื่อระบุผู้บุกรุกเครือข่ายโดยการรู้จำผู้บุกรุกที่เป็นลักษณะของการบุกรุกเครือข่ายมีขอบเขตดังต่อไปนี้

ชุดข้อมูลเครือข่าย เป็นข้อมูลมาตรฐานที่ใช้ในการทำแบบทดลองการตรวจจับการบุกรุก สามารถดาวน์โหลดได้เว็บไซต์ http://kdd.ics.uci.edu/databases/KDDcup99/KDDcup99.html

ลักษณะของข้อมูลที่นำมาทำแบบทดสอบ จะประกอบด้วยข้อมูล 5 กลุ่มหลัก ซึ่งได้แก่ Normal, DoS, Probe, R2L และ U2R และมีรูปแบบของฟีเจอร์ (Feature) ดังแสดงตามรูปที่ 1-1

0,tcp,pop_3,RSTO,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,211,6,0.00,0.00,1.00,1.00,0.03,0.07,0.00,255,6,0.02,0.07,0.00,0.00,0.00,0.00,1.00,1.00,neptune.
0,tcp,pop_3,RSTO,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,231,16,0.00,0.00,1.00,1.00,0.07,0.06,0.00,255,16,0.06,0.07,0.00,0.00,0.00,0.00,1.00,1.00,neptune.
0,tcp,pop_3,RSTO,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,232,5,0.00,0.00,1.00,1.00,0.02,0.06,0.00,255,5,0.02,0.07,0.00,0.00,0.00,0.00,1.00,1.00,neptune.
0,tcp,pop_3,RSTO,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,254,15,0.00,0.00,1.00,1.00,0.06,0.07,0.00,255,15,0.06,0.07,0.00,0.00,0.00,0.00,1.00,1.00,neptune.
0,tcp,pop_3,RSTO,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,252,6,0.00,0.00,1.00,1.00,0.02,0.07,0.00,255,6,0.02,0.08,0.00,0.00,0.00,0.00,1.00,1.00,neptune.
0,tcp,pop_3,RSTO,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,272,16,0.00,0.00,1.00,1.00,0.06,0.06,0.00,255,16,0.06,0.07,0.00,0.00,0.00,0.00,1.00,1.00,neptune.
0,tcp,pop_3,SH,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1.00,1.00,0.00,0.00,1.00,0.00,0.00,255,1,0.00,1.00,1.00,0.00,1.00,1.00,0.00,0.00,nmap.
0,tcp,pop_3,SH,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1.00,1.00,0.00,0.00,1.00,0.00,0.00,255,1,0.00,1.00,1.00,0.00,1.00,1.00,0.00,0.00,nmap.
5,tcp,pop_3,SF,6,151,0,0,0,0,0,0,0,0,0,0,0,0,0,0,511,1,0.07,0.00,0.91,0.00,0.00,1.00,0.00,255,1,0.00,1.00,0.00,0.00,0.07,0.00,0.90,0.00,satan.
40339,tcp,pop_3,RSTR,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,1.00,1.00,1.00,0.00,0.00,255,2,0.01,0.44,0.86,0.00,0.00,0.00,0.86,1.00,portsweep.

รูปที่ 1-1 ตัวอย่างของข้อมูลที่ใช้ในงานวิจัย

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. ได้ขั้นตอนวิธีการสกัดลักษณะของชุดข้อมูลเครือข่าย โดยอาศัยวิธีการแบบผสมสำหรับการสกัดลักษณะและเลือกลักษณะชุดข้อมูลเครือข่าย และการรู้จำการระบุผู้บุกรุก ซึ่งเป็นการสกัดลักษณะและเลือกลักษณะในรูปแบบใหม่ที่แตกต่างจากแบบเดิม โดยผสมกับตัวแบบโครงข่ายประสาทเทียม
2. สามารถนำไปพัฒนาระบบตรวจจับการบุกรุกบนระบบเครือข่ายได้ ซึ่งจะทำให้ช่วยรักษาความปลอดภัยเครือข่ายและให้การใช้งานเครือข่ายมีประสิทธิภาพมากขึ้น
3. ขั้นตอนวิธีที่นำเสนอจะสามารถใช้เป็นต้นแบบในการศึกษาขั้นสูงต่อไป

1.5 ระยะเวลาทำการวิจัยและแผนการดำเนินงานตลอดโครงการวิจัย

ดำเนินการวิจัยแบบ 2 ปีต่อเนื่อง (ปีงบประมาณ 2554-2555)

แผนการดำเนินงานปีที่ 1 (ปีงบประมาณ 2554-2555)

แผนการดำเนินงานวิจัย	เดือนที่											
	1	2	3	4	5	6	7	8	9	10	11	12
การจัดเตรียมเครื่องมือ อุปกรณ์และวัสดุวิจัย	→											
การศึกษาการสกัดลักษณะชุดข้อมูล	→	→	→	→	→	→						
การจัดทำรายงานความก้าวหน้าครั้งที่ 1							→					
การศึกษาวิธีการแบบผสมสำหรับการสกัดลักษณะชุดข้อมูล							→	→	→	→	→	→
การจัดทำรายงานความก้าวหน้าครั้งที่ 2												→

แผนการดำเนินงานปีที่ 2 (ปีงบประมาณ 2555-2556)

แผนการดำเนินงานวิจัย	เดือนที่											
	1	2	3	4	5	6	7	8	9	10	11	12
การศึกษาระบบโครงข่ายประสาทเทียมเพื่อการรู้จำและระบุผู้บุกรุก	→	→	→	→	→	→						
การศึกษาการเลือกลักษณะชุดข้อมูล							→	→	→	→	→	
การจัดทำรายงานความก้าวหน้าครั้งที่ 3											→	
จัดส่งรายงาน												→

บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 การตรวจจับการบุกรุก

เนื่องจากการพัฒนาอย่างรวดเร็วของโครงข่ายอินเทอร์เน็ตนั้น การรักษาความปลอดภัยจึงเป็นสิ่งสำคัญ โดยปกติแล้วการบุกรุกจะเกิดขึ้นใน 3 ลักษณะคือ การละเมิดความเป็นส่วนตัวหรือความลับ การแก้ไขความถูกต้องของข้อมูล และการทำให้ไม่สามารถใช้งานระบบคอมพิวเตอร์ได้ วิธีการหนึ่งที่นิยมนำมาใช้ในการสร้างความปลอดภัยให้กับระบบเครือข่ายคอมพิวเตอร์ คือ การตรวจจับการบุกรุก (Intrusion Detection) ซึ่งการตรวจจับการบุกรุก สามารถแบ่งออกได้เป็น 2 ชนิดคือ ระบบการตรวจจับการบุกรุกแบบโฮสเบส (Host-based Intrusion Detection Systems) และระบบตรวจจับการบุกรุกแบบเน็ตเวิร์คเบส (Network-based Intrusion Detection Systems) โดยที่ระบบตรวจจับการบุกรุกแบบเน็ตเวิร์คเบสนั้นจะติดตั้งที่ระบบเครือข่ายเพื่อทำการตรวจสอบและวิเคราะห์ชุดข้อมูลที่ใช้งานบนเครือข่าย ซึ่งมีความแตกต่างจากระบบการตรวจจับการบุกรุกแบบโฮสเบส ที่จะทำงานอยู่บนระบบเพื่อตรวจสอบและวิเคราะห์ชุดคำสั่งเพื่อระบุการทำงานที่น่าสงสัย วิธีการของการตรวจจับการบุกรุกสามารถแบ่งออกได้เป็น 2 ชนิด คือ วิธีการตรวจจับการบุกรุกแบบอโนมาลี (Anomaly Intrusion Detection Method) และวิธีการตรวจจับการบุกรุกแบบมิสยูส (Misuse Intrusion Detection Method) โดยที่วิธีการตรวจจับการบุกรุกแบบอโนมาลี นั้นเป็นวิธีการหาผู้บุกรุกโดยการวิเคราะห์การใช้งานของผู้ใช้งาน หรือตัวระบบเองที่เบี่ยงเบนไปจากระดับการใช้งานโดยปกติ ส่วนการตรวจจับการบุกรุกแบบมิสยูสนั้น เป็นวิธีการหาผู้บุกรุกโดยการเปรียบเทียบข้อมูลที่เข้ามากับรูปแบบของผู้บุกรุกที่มีอยู่เดิม ซึ่งทั้งสองวิธีนี้มีจุดแข็งและจุดอ่อนที่แตกต่างกัน ปัญหาที่เด่นชัดที่สุดของการตรวจจับการบุกรุกแบบมิสยูส คือ ไม่สามารถตรวจจับการบุกรุกแบบใหม่ หรือการบุกรุกที่ไม่มีในชุดรูปแบบของผู้บุกรุกที่มีได้ ส่วนการตรวจจับการบุกรุกแบบอโนมาลีนั้น จะระบุว่าการใช้งานที่ตรวจสอบนั้นเป็นผู้บุกรุกหรือไม่นั้นจะตรวจสอบจากการใช้งานนั้นว่ามีการเบี่ยงเบนจากกิจกรรมปกติมากหรือไม่ ดังนั้นการตรวจจับการบุกรุกแบบอโนมาลีจะสามารถตรวจจับการบุกรุกจากผู้บุกรุกที่ไม่มีในฐานข้อมูลการบุกรุกได้

2.2 ลักษณะของข้อมูลที่ใช้ในการทำแบบทดลอง

ข้อมูลที่นำมาใช้ในการทำแบบทดลอง เป็นข้อมูลที่ได้จากฐานข้อมูลความรู้ (Knowledge Discovery in Database (KDD) Cup data) ซึ่งเป็นชุดข้อมูลในปี 1999 โดยชุดข้อมูลนี้ได้มาจากความร่วมมือของโครงการวิจัยและพัฒนาเพื่อการทหารของประเทศสหรัฐอเมริกา (Defense Advanced Research Projects Agency: DARPA) ซึ่งร่วมมือกับทางมหาวิทยาลัยเมตซาชูเซตส์ สหรัฐอเมริกา ชุดข้อมูลนี้ถูกสร้างตามการจำลองการโจมตีของผู้บุกรุกจาก U.S. Air Force local area network ตั้งขึ้นที่ Lincoln Labs โดยข้อมูลนั้นมีระยะเวลาในการจัดทำนานถึง 9 สัปดาห์ จากการเก็บข้อมูลจากแพ็กเก็ต TCP ผ่านโปรแกรม TCP Dump ประกอบด้วยข้อมูลขนาดใหญ่มาก ซึ่งมี 41 ลักษณะ (Feature) ที่ได้จากแพ็กเก็ต TCP (Raw TCP Packet) รวมถึงชนิดของโปรโตคอลซึ่งมีค่า “TCP”, “ICMP”, “UDP” ซึ่งเป็นแอทริบิวต์ที่มีความต่อเนื่องเป็นในลักษณะข้อความ (Nominal) ซึ่งจะมีสถานะ (Label)

กำกับไว้เสมอในแต่ละบรรทัด (Record) ว่าข้อมูลชุดนี้เป็นสถานะปกติ (Normal) หรือว่าเป็นชุดข้อมูลที่ถูกโจมตี (Malicious) ทางผู้จัดทำได้คัดเลือก ชุดข้อมูลที่มีมากถึง 5 ล้านบรรทัด ออกมาประมาณ 10% เท่านั้นเพื่อสะดวกในการจัดทำชุดข้อมูลเรียนรู้ (Training Set) และชุดข้อมูลทดสอบ (Test Set)

ชุดข้อมูลที่ได้นำมาทำแบบทดลองนี้ อยู่ในรูปแบบของเครื่องกลเรียนรู้ (Machine Learning Pattern) โดยสามารถแบ่งออกเป็นกลุ่มหลักๆ ได้ 5 กลุ่มคือ Normal, DoS, Probe, R2L และ U2R โดยแต่ละกลุ่มยังมีชนิดของข้อมูลย่อย ๆ อีก โดยมีรายละเอียดดังนี้

Normal คือ ข้อมูลมีลักษณะปกติหรือไม่มีการบุกรุก

Dos คือ ผู้บุกรุกพยายามโจมตีเพื่อทำให้เครื่องคอมพิวเตอร์ปลายทางหยุดทำงาน หรือสูญเสียเสถียรภาพ ซึ่งแบ่งออกเป็นประเภทย่อยๆอีก ได้แก่ back, land, neptune, pod, smurf และ teardrop

Probe คือ ผู้บุกรุกพยายามตรวจสอบหาจุดอ่อนของระบบ แบ่งเป็นประเภทย่อย ได้แก่ ipsweep, nmap, portsweep และ satan

R2L คือ ผู้บุกรุกไม่มี user ในระบบแต่พยายามเจาะเข้าไปในระบบ แบ่งเป็นประเภทย่อย ได้แก่ ftp_write, guess_passwd, imap multihop, phf, spy, warezclient และ warezmaster

U2R คือ ผู้บุกรุกพยายามเข้าสู่ระบบโดยการใช้สิทธิ์ของ super user แบ่งเป็นประเภทย่อย ได้แก่ buffer_overflow, loadmodule, perl และ rootkit

แต่ละแอทริบิวต์มีรายละเอียดดังตารางที่ 2-1 โดยแอทริบิวต์สุดท้ายคือ คลาสแอทริบิวต์ (Class Attribute) ที่เป็นตัวบอกสถานะว่าถูกโจมตีหรือไม่ ซึ่งหากไม่ถูกโจมตีจะมีค่าเป็นปกติ

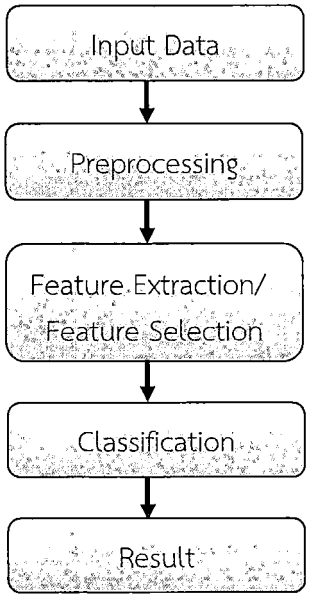
ตารางที่ 2-1 รายละเอียดลักษณะของข้อมูล KDDcup99

Feature Number	Feature Name
1	duration
2	protocol type
3	service
4	Flag
5	src_bytes
6	dst_bytes
7	land
8	wrong _ fragment
9	urgent
10	Hot
11	num_field_logins
12	logged_in
13	num_compromised
14	root_shell
15	su_attempted
16	num_root
17	num_file_creation
18	num_shells
19	num_access_files
20	num_outbounds_cmds
21	is_hist_login
22	is_guest_login
23	count
24	srv_count
25	serror_rate
26	srv_serror_rate
27	rerror_rate
28	srv_rerror_rate
29	same_srv_rate
30	diff_srv_rate
31	srv_diff_host_rate
32	dst_host_count
33	dst_host_srv_count
34	dst_hosdst_same_srv_rate
35	dst_host_diff_srv_rate
36	dst_host_same_src_port_rate
37	dst _host_ srv _diff_host_rate
38	dst_host_serror_rate
39	dst_host_srv_serror_rate
40	dst_host_rerror_rate
41	dst_host_srv_rerror_rate
42	class

2.3 กระบวนการรู้จำทางคอมพิวเตอร์

โดยทั่วไปกระบวนการรู้จำทางคอมพิวเตอร์นั้น จะมีขั้นตอนการทำงานหลักๆ คือ การประมวลผลเบื้องต้น (Preprocessing) การสกัดลักษณะหรือการเลือกลักษณะของข้อมูล (Feature Extraction/Feature Selection) และการจำแนกข้อมูล (Classification) แสดงดังรูปที่ 2-1 ซึ่งในแต่ละส่วนมีรายละเอียดเบื้องต้นดังนี้

Input Data เป็นข้อมูลที่มีลักษณะหลายรูปแบบตามความต้องการของระบบ เช่น ในการรู้จำตัวอักขระ อาจจะใช้ภาพที่มีลักษณะเป็นข้อความบรรทัดเดียว ข้อความหลายบรรทัด หรือ ภาพตัวอักขระจำนวน 1 อักขระ เป็นต้น ซึ่งข้อมูลอาจได้จากการเก็บข้อมูล หรือนำเอาเอกสารที่เป็นกระดาษไปสแกนเพื่อเปลี่ยนเป็นข้อมูลทางคอมพิวเตอร์ เช่น แฟ้มข้อมูลชนิด txt, xls, bitmap หรือ ได้จากการป้อนข้อมูลผ่านอุปกรณ์อินพุต เช่น เมาส์หรือปากกาอิเล็กทรอนิกส์



รูปที่ 2-1 กระบวนการรู้จำสำหรับปัญหา Intrusion Detection

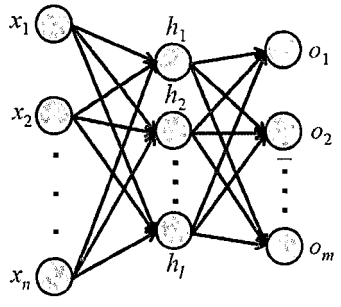
Preprocessing เป็นการประมวลผลเบื้องต้นเพื่อปรับเปลี่ยนลักษณะรูปแบบบางอย่างของข้อมูลอินพุต ทั้งนี้เพื่อปรับอินพุตให้มีความเหมาะสมและตรงตามที่ระบบต้องการ เช่นปรับขนาด (Resize) ปรับลดจำนวนลักษณะ (Reduce Dimension) หรือการกำจัดสัญญาณรบกวน (Noise Remove)

Feature Extraction/Feature Selection เป็นขั้นตอนของการสกัดเอาลักษณะเฉพาะหรือการเลือกลักษณะที่สำคัญของแต่ละอินพุตออกมาเป็นเวกเตอร์เพื่อนำไปใช้เป็นอินพุตในการเรียนรู้ระบบและทดสอบระบบ

Classification เป็นขั้นตอนในการจำแนกและตัดสินใจว่าอินพุตที่เข้ามานั้นเป็นการบุกรุกแบบใด โดยในขั้นตอนนี้มีหลายวิธีด้วยกัน เช่นการเปรียบเทียบอินพุตกับโครงสร้างของตัวต้นแบบการบุกรุกในฐานข้อมูลการเปรียบเทียบอินพุตกับกฎเพื่อการตัดสินใจการใช้โครงข่ายประสาทเทียม หรือการใช้ตัวแบบฮิดเดนมาร์คอฟ เป็นต้น

2.3.1 ระบบโครงข่ายประสาทเทียมแบบวิธีการแพร่กระจายย้อนกลับ *(Back propagation Algorithm)*

ขั้นตอนวิธีการแพร่กระจายย้อนกลับ เป็นขั้นตอนวิธีที่ใช้ในการเรียนรู้ของเครือข่ายประสาทเทียมวิธีหนึ่งที่นิยมใช้ในโครงข่ายประสาทเทียมหลายชั้น (Multilayer neural network) เพื่อใช้ในการปรับค่าน้ำหนักในเส้นเชื่อมต่อระหว่างโหนดให้เหมาะสม (Robert Hecht Nielsen, 1989) โดยการปรับค่านี้จะขึ้นกับความแตกต่างของค่าเอาต์พุตที่คำนวณได้กับค่าเอาต์พุตที่ต้องการ พิจารณารูปต่อไปนี้ประกอบ



รูปที่ 2-2 ตัวอย่างข่ายงานประสาทเทียมแบบหลายชั้น

ตัวอย่างในรูปด้านบนแสดงข่ายงานป้อนไปหน้าแบบหลายชั้นซึ่งประกอบไปด้วยชั้นอินพุต ชั้นฮิดเดนหรือชั้นซ่อน และชั้นเอาต์พุต ในรูปแสดงชั้นฮิดเดนเพียงชั้นเดียวแต่อาจมีมากกว่าหนึ่งชั้นก็ได้ เส้นเชื่อมจะเชื่อมต่อเป็นชั้น ๆ ถ้ามีชั้นฮิดเดนมากกว่าหนึ่งชั้นก็เชื่อมต่อกันไป และสุดท้ายจากชั้นฮิดเดนไปชั้นเอาต์พุต

ในการปรับค่าน้ำหนักโดยขั้นตอนวิธีการแพร่กระจายย้อนกลับนั้น เราต้องนิยามค่าผิดพลาดสำหรับการเรียนรู้ของข่ายงาน $MSE(\vec{w})$ จากนั้นจะหาค่าน้ำหนักที่ให้ค่าผิดพลาดต่ำสุด นิยามค่าผิดพลาดดังนี้

$$MSE(\vec{w}) = \frac{1}{2}\sum_{p\in P}\sum_{k\in outputs}(d_{p,k} - o_{p,k})^2 \qquad ...(1)$$

โดยที่ *Outputs* คือ เซตของเอาต์พุตโหนดในข่ายงานประสาทเทียม $d_{p,k}$ และ $o_{p,k}$ เป็นค่าเอาต์พุตเป้าหมายและเอาต์พุตที่ได้จากข่ายงานประสาทเทียมตามลำดับของเอาต์พุตโหนดที่ k ของตัวอย่างที่ p ขั้นตอนการแพร่กระจายย้อนกลับจะค้นหาค่าน้ำหนักที่ให้ค่าผิดพลาดกำลังสองเฉลี่ยต่ำสุด

ขั้นตอนของ Back-propagation Algorithmมีดังนี้

Algorithm Backpropagation;

Start with randomly chosen weights;

while MSE is unsatisfactory

and computational bounds are not exceeded, do

for each input pattern $x_p, 1 \leq p \leq P$,

Compute hidden node inputs $(net_{p,j}^{(1)})$;

Compute hidden node outputs $(x_{p,j}^{(1)})$;

Compute inputs to the output nodes $(net_{p,j}^{(2)})$;

Compute the network outputs $(o_{p,k})$;

Modify outer layer weights:

$$\Delta w_{k,j}^{(2,1)} = \eta(d_{p,k} - o_{p,k})S'(net_{p,j}^{(2)})x_{p,i}$$

end-for

end-while.

Note: if S is a logistic function, then $S' = S(x)(1 - S(x))$

2.3.2 วิธีการรู้จำแบบซัพพอร์ตเวกเตอร์แมชชีน *(Support Vector Machine: SVM)*

Support Vector Machine หรือ SVM จุดมุ่งหมายที่สำคัญของแนวคิด SVM คือการหาเส้นแบ่ง Hyperplane (M. Hearst, 1998) ซึ่งใช้แบ่งข้อมูลออกเป็นคลาส เพื่อให้ได้ผลลัพธ์ที่ดี โดยพิจารณาจากสมการเส้นตรง Hyperplanes และ SVM จะทำการค้นหาจุดของข้อมูลที่อยู่ใกล้เส้นแบ่ง Hyper planes ซึ่งจุดนี้เรียกว่า "Support Vector" มีหลักการดังนี้

นำข้อมูลมาคำนวณหาค่า y ซึ่งค่า $y \in \{-1,1\}$ จากสมการ

$$y = w^T x + b \quad \text{...(2)}$$

คำนวณหาเส้นแบ่ง ซึ่งเรียกว่าเส้น Optimal Hyperplane จากสมการ

$$w^T x + b = 0 \quad \text{...(3)}$$

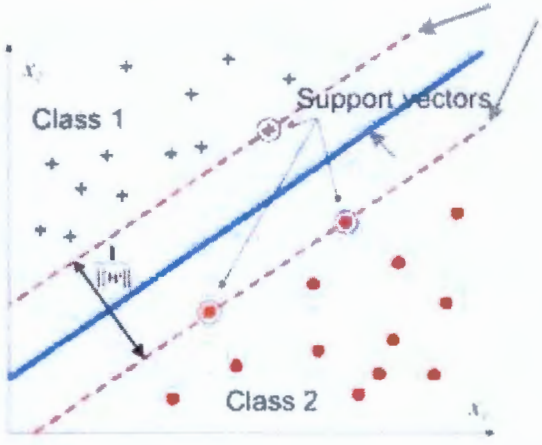
ระยะทาง (d) หรือ maximum margin จากเส้นขอบ ณ จุด x_i ไปยัง Hyperplane แสดงดังสมการ

$$d = \frac{|w^T x_i + b|}{\|w\|} \quad \text{...(4)}$$

w คือ เวกเตอร์น้ำหนัก (Weight Vector)

x_i คือ Input

b คือ ค่าคงที่ที่กำหนดขึ้นเพื่อให้เหมาะสมกับการจัดกลุ่ม



รูปที่ 2-3 การแบ่งกลุ่มข้อมูลโดย Support Vector Machine

เลือกจุดที่อยู่ใกล้เส้นตรง Optimal Hyperplane ทั้งเหนือเส้นซึ่งเรียกว่า "ขอบล่าง" ซึ่งเป็นขอบล่างสุดของคลาสเอกสารที่อยู่เหนือเส้นตรง Optimal Hyperplane และใต้เส้นเรียกว่า "ขอบบน" ซึ่งเป็นขอบบนสุดของคลาสเอกสารที่อยู่ใต้เส้นตรง Optimal Hyperplane เพื่อที่จะหาระยะทางระหว่างเส้นขอบทั้งสองโดยจะเลือกเอาค่าระยะทางที่ห่างจากเส้นตรง Optimal Hyperplane ที่น้อยที่สุดเป็นตัวเลือกในการจัดกลุ่มเอกสาร

2.3.3 โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน

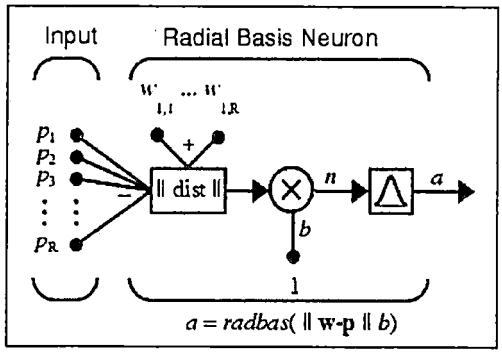
โดยแบบที่นิยมใช้โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน เป็นโครงข่ายประสาทเทียมป้อนไปข้างหน้าแบบหลายชั้นจะประกอบไปด้วย 3 ชั้น ได้แก่ ชั้นรับข้อมูลเข้า ชั้นซ่อน และชั้นข้อมูลออก (S.Chen, 1991) ดังรูปที่ 2-4 โดยเป็นฟังก์ชันการส่งระหว่างชั้นรับข้อมูลเข้า $p \in \mathbb{R}^{Nx1}$ ไปยังชั้นข้อมูลออก $y \in \mathbb{R}^{Mx1}$ จะได้ข้อมูลออกของเครือข่ายดังนี้

$$y_i = \sum_{k=i}^{s} w_{ik} \emptyset_k (\|p - c\|) \quad \ldots(5)$$

โดยที่ w_{ik} คือ ค่าน้ำหนักนิวรอนในชั้นซ่อน

s คือ จำนวนนิวรอนในชั้นซ่อน

c คือ เวกเตอร์จุดศูนย์กลาง



รูปที่ 2-4 โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน

(ที่มา http://matlab.izmiran.ru/help/toolbox/nnet/radial74.html)

2.4 การสกัดลักษณะและการเลือกลักษณะ (Feature Extraction and Feature Selection)

การหาตัวแทนข้อมูลเป็นอีกกระบวนหนึ่งที่มีความสำคัญเพื่อลดขนาดของข้อมูลโดยที่สูญเสียลักษณะสำคัญของข้อมูลน้อยที่สุดและสูญเสียความถูกต้องของผลลัพธ์น้อยที่สุด และเพื่อให้ตัวจำแนกประเภทสามารถทำงานได้ถูกต้องมากขึ้น ซึ่งวิธีการหาตัวแทนข้อมูลได้แก่ การสกัดลักษณะ และการเลือกลักษณะ ตำราส่วนใหญ่จะแยกส่วนนี้ออกจากการประมวลผลเบื้องต้นคือจะอยู่ระหว่างขั้นตอนการประมวลผลเบื้องต้นกับขั้นตอนการรู้จำในส่วนของการสกัดลักษณะเป็นการดึงเอาโครงสร้างพื้นฐานที่สำคัญของข้อมูลนั้นออกมาโดยโครงสร้างพื้นฐานที่ว่าจะต้องมีการกำหนดไว้ก่อนว่าจะมีอะไรบ้างมีการนิยามอย่างไร ตัวอย่างเช่นสำหรับภาษาไทยเราอาจกำหนดว่าตัวอักษรภาษาไทยทั้งหมดประกอบด้วยโครงสร้างพื้นฐานคือเส้นตรง (แนวตั้ง/นอน) เส้นเอียง หัว (วงกลม) ส่วนโค้ง ส่วนเว้า จุดแตกกิ่ง จุดตัด เป็นต้น เมื่อเราสามารถแยกเอาองค์ประกอบของตัวอักษรแต่ละตัวออกมาได้แล้วจากนั้นเราก็นำเสนอรูปภาพของตัวอักษรนั้นในรูปแบบของรายการขององค์ประกอบพื้นฐานต่างๆแทน ซึ่งจะถูกส่งต่อเป็นอินพุตสำหรับขั้นตอนการรู้จำต่อไป ในส่วนของการเลือกลักษณะเป็นขั้นตอนหนึ่งของกระบวนการทำเหมืองข้อมูลเพื่อค้นหาเซตย่อยของลักษณะที่เหมาะสมที่สุด ซึ่งการทำเหมืองข้อมูลเกี่ยวข้องกับการวิเคราะห์ข้อมูลจำนวนมหาศาลและมีความซับซ้อนมาก ซึ่งโดยส่วนใหญ่กระบวนการทำเหมืองข้อมูลนั้นประกอบไปด้วยคุณลักษณะที่ไม่สำคัญ (Irrelevant Feature) และคุณลักษณะที่ซ้ำซ้อน (Redundant Feature) ดังนั้นการคัดเลือกคุณลักษณะจึงเป็นงานที่สำคัญ ซึ่งจะช่วยปรับปรุงศักยภาพของการทำเหมืองข้อมูลโดยลดทั้งขนาดข้อมูล และเวลาที่ใช้ในการวิเคราะห์ข้อมูล ซึ่งหากเลือกคุณลักษณะได้อย่างมีประสิทธิภาพจะทำให้การวิเคราะห์ข้อมูลมีความแม่นยำสูง และทำให้ได้ตัวแทนที่มีศักยภาพ

2.5 วิธีการวิเคราะห์องค์ประกอบหลัก

วิธีการวิเคราะห์องค์ประกอบหลัก Principal Component Analysis (PCA) เป็นวิธีการทางสถิติ เพื่อใช้ในการสกัดปัจจัยที่อาศัยหลักความสัมพันธ์เชิงเส้นตรงระหว่างตัวแปรที่ใช้เป็นข้อมูล องค์ประกอบหลักตัวแปร คือ การผสมเชิงเส้นตรง (Linear Combination) ของตัวแปรที่อธิบายการผันแปรของข้อมูลได้มากที่สุด (Jackson, 1991) จากนั้นหาการผสมเชิงเส้นครั้งที่สองที่สามารถอธิบายการผันแปรได้มากที่สุดเป็นอันดับที่สอง โดยที่ไม่สัมพันธ์กับการผสมครั้งแรก การวิเคราะห์องค์ประกอบหลักถูกนำไปประยุกต์ใช้งานต่างๆ เช่น การบีบอัดข้อมูล การสร้างภาพใบหน้าไอเกนเพื่อใช้ในระบบจดจำ และ การลบออกของพื้นหลังโดยใช้ไอเกน เป็นต้นวิธีการวิเคราะห์องค์ประกอบหลักสามารถนำมาใช้ในการลดลักษณะของข้อมูลโดย การวิเคราะห์ข้อมูลและเลือกเฉพาะข้อมูลที่มีความสำคัญเท่านั้น ส่วนข้อมูลที่ไม่สำคัญจะถูกตัดทิ้งไป ดังนั้นเมื่อข้อมูลผ่านกระบวนการ PCA แล้ว จะได้ผลลัพธ์เป็นไอเกนเวกเตอร์และค่าไอเกน ซึ่งไอเกนเวกเตอร์ที่มีค่าสมนัยกับค่าไอเกนที่มีค่าสูงๆ จะเป็นการดึงข้อมูลที่มีความถี่ต่ำ ส่วนไอเกนเวกเตอร์ที่สมนัยกับค่าไอเกนที่ต่ำๆ จะเป็นการดึงข้อมูลที่มีความถี่สูง

2.5.1 การหาค่าไอเกน และไอเกนเวกเตอร์ *(Eigen Value and Eigen Vector)*

ความหมายของค่าไอเกน และไอเกนเวกเตอร์ กำหนดให้ A เป็นค่าเมทริกซ์จัตุรัส

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ a_{31} & a_{32} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

และ v เป็นเวกเตอร์หลัก (Column Vector) และ λ เป็นค่าคงที่ใดๆ โดยที่

$$v = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

ที่ทำให้

$$Av = \lambda v \qquad \text{...(6)}$$

เมื่อ A แทน ค่าเมตทริกซ์

λ แทน เป็นค่าคงที่ใดๆ เป็นสเกลาร์

v แทน ค่าไอเกนเวกเตอร์

จากสมการจะเห็นว่า $v = 0$ ที่ทำให้สมการ เป็นจริงทุกๆ ค่าของ สมการที่ (6) อาจเขียนให้อยู่ในอีกรูปหนึ่งคือ

$$(\lambda I - A)v = 0 \quad \text{...(7)}$$

เมื่อ A แทน ค่าเมทริกซ์
I แทน เมทริกซ์เอกลักษณ์
λ แทน เป็นค่าคงที่ใดๆ เป็นสเกลาร์
v แทน ค่าไอเกนเวกเตอร์

เราจะคำนวณค่าไอเกน และเวคเตอร์ไอเกน ของสมการ (8) โดย

$$\det(\lambda I - A) = 0 \quad \text{...(8)}$$

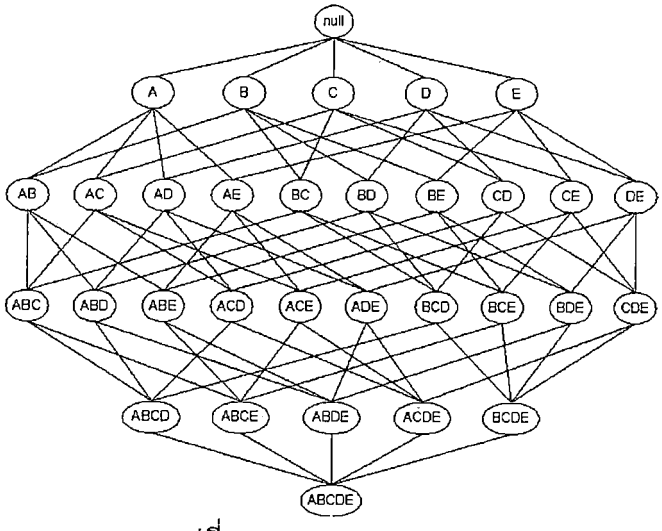
จากนั้นก็ใช้วิธีแก้สมการแบบปกติ

2.6 ขั้นตอนวิธีฮิวริสติกกรีดดี

ขั้นตอนวิธีฮิวริสติกกรีดดี (Heuristic Greedy Algorithm) เป็นขั้นตอนวิธีการแก้ปัญหาที่คิดแบบง่าย ๆ และตรงไปตรงมา (T.H. Cormen, 2001) ซึ่งเป็นการแก้ปัญหาในลักษณะที่ไม่มีรูปแบบวิธีการขั้นตอนโดยตรง โดยจะพิจารณาว่าข้อมูลที่มีอยู่ในขณะนั้นมีทางเลือกใดที่ให้คำตอบที่ดีที่สุดของปัญหาโดยการเลือกคำตอบที่ดีที่สุดในขณะนั้น ซึ่งถ้าข้อมูลนั้นเพียงพอที่จะทำให้สรุปคำตอบที่ดีที่สุด เราจะได้ขั้นตอนวิธีที่มีประสิทธิภาพ การค้นหาคำตอบอาศัยวิธีการทางฮิวริสติก สามารถทำการค้นหาคำตอบจากข้อมูลที่มีขนาดใหญ่มาก ๆ ได้ เพราะเป็นการค้นหาคำตอบที่ไม่ต้องดูข้อมูลทุกตัว เนื่องจากใช้ฮิวริสติกฟังก์ชัน (Heuristic Function) ซึ่งเป็นฟังก์ชันในการวัดความเป็นไปได้ในการแก้ปัญหาซึ่งจะแสดงด้วยตัวเลข ซึ่งต่างจากการค้นหาข้อมูลแบบธรรมดาที่ต้องพิจารณาตรวจสอบข้อมูลทุกตัวจนครบทำให้ไม่เหมาะกับข้อมูลที่มีขนาดใหญ่ทำให้เสียเวลาได้ แต่ข้อเสียของการค้นหาคำตอบอาศัยวิธีการทางฮิวริสติกคือคำตอบที่ได้เป็นเพียงคำตอบที่ดี แต่ไม่รับรองว่าเป็นคำตอบที่ดีที่สุด สิ่งที่สำคัญในการแก้ปัญหาวิธีการทางฮิวริสติกว่าจะสามารถแก้ปัญหาได้ตามที่ต้องการหรือไม่ คือ ฮิวริสติกฟังก์ชัน ทำหน้าในการวัดความเป็นไปได้ของคำตอบ ซึ่งเป็นการกำกับทิศทางของกระบวนการค้นหา เพื่อให้อยู่ในทิศทางที่ได้ประโยชน์สูงสุด โดยพิจารณาจากน้ำหนักที่ให้กับการแก้ปัญหาของแต่ละวิธี น้ำหนักเหล่านี้จะถูกแสดงด้วยตัวเลขที่กำกับไว้กับโหนดต่าง ๆ และค่าเหล่านี้จะเป็นตัวที่ใช้ในการประมาณความเป็นไปได้ว่าเส้นทางที่ผ่านโหนดนั้นจะมีความเป็นไปได้ในการเข้าใกล้เป้าหมายมากน้อยเพียงใด ตัวอย่างของการค้นหาคำตอบอาศัยวิธีการทางฮิวริสติก เช่น การค้นหาแบบกรีดดี เป็นการค้นหาแบบดีที่สุดก่อน (Best First Search) ที่ง่ายที่สุด เป็นการนำข้อดีของการค้นหาตามแนวกว้าง และการค้นหาตามแนวลึกมารวมกัน โดยการค้นหาแบบดีที่สุดก่อน จะเลือกโหนดที่มีค่าดีสุดซึ่งอาศัยฮิวริสติกฟังก์ชันในการหาและหลักการของขั้นตอนวิธีกรีดดีเพื่อหาคำตอบที่เหมาะสมที่สุดในแต่ละสถานการณ์

2.7 การสร้าง Itemsets โดยใช้หลักการ Apriori

การสร้าง Itemsets สามารถใช้โครงสร้างแลตทิซ (Lattice Structure) ในการแจกแจง Itemsets ทั้งหมดที่เป็นไปได้ จากจำนวน Items ที่มีอยู่ เช่นตัวอย่างโครงสร้างแลตทิซของ 5 Items คือ I = {A, B, C, D, E} แสดงได้ดังรูปที่ 2-5 ซึ่งมีความยาว Itemsets จากระดับชั้น (Level) ที่ 1 (1-itemset) ถึงระดับชั้นที่ 5 (5-itemset)



รูปที่ 2-5 Itemsets Lattice

ถ้าในชุดข้อมูลมีจำนวน Items เท่ากับ k items ดังนั้นจำนวน Itemsets ที่มีโอกาสเป็น Frequent Itemsets ทั้งหมดคำนวณได้จาก (ไม่รวมเซตว่าง) ซึ่งจะเห็นได้ว่าเมื่อมีการใช้งานจริงในภาคธุรกิจหรือลักษณะงานอื่นๆ ค่าของ k จะสูงมาก ผลที่ตามมาคือการค้นหาและเปรียบเทียบ รวมถึงการคำนวณจะมีจำนวนมากขึ้นเป็นทวีคูณ ในการหา Frequent Itemsets ขั้นตอนวิธีจะต้องแจง Itemsets ทั้งหมดที่เป็นไปได้ตามโครงสร้างแลตทิซ เราเรียก Itemsets ชุดนี้ว่า Candidate Itemsets (P. N. Tan, 2006) เพื่อช่วยกำจัดหรือลดจำนวน Candidate Itemsets แทนที่จะแจกแจงทั้งหมดเหมือนที่แสดงด้วยโครงสร้างแลตทิซ เราจะใช้หลักการที่เรียกว่า Apriori ซึ่งมีหลักการดังนี้

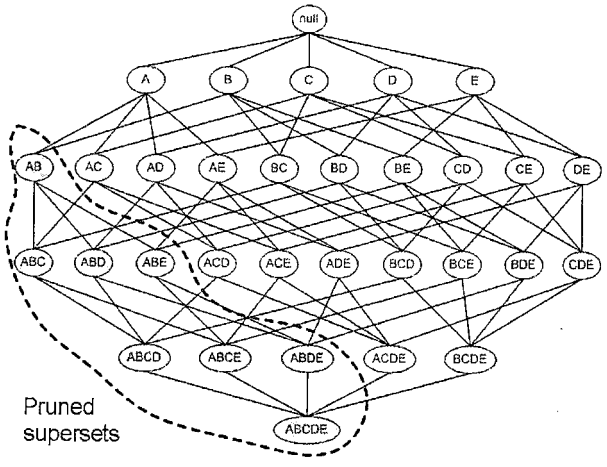
ถ้า Itemset หนึ่งๆ เป็น Frequent แล้ว ทุกๆ สับเซตของ Itemset นั้นจะต้องเป็น Frequent ด้วย

หลักการ Apriori เป็นหลักการที่นิยมใช้ในการหากฎความสัมพันธ์ (Association Rules) ซึ่งเป็นวิธีการที่ง่ายแต่มีประสิทธิภาพที่จะนำไปสู่การสร้าง Candidate Set ที่น้อยลง โดยการใช้เซตที่มีขนาดใหญ่ที่หาได้ในขั้นตอนก่อนหน้า ขั้นตอน Apriori เป็นการทำงานที่ซ้ำ ๆ ในขั้นตอนแรกจะหาเซตที่มีขนาด 1-itemsets จากนั้นจะหาเซตที่มีขนาด 2-itemsets, 3-itemsets และต่อ ๆ ไป

ตัวอย่างจากข้อมูลรูปโครงสร้างแลตทิซ ในรูปที่ 2-6 สมมติให้ทรานแซคชันหนึ่ง ๆ ประกอบด้วย Items สามตัวคือ {C, D, E} ดังนั้นทรานแซคชันดังกล่าวจะประกอบด้วยสับเซตดังนี้ {C, D}, {C, E}, {D, E}, {C}, {D}, {E} ซึ่งหาก 3-itemset = {CDE} เป็น Frequent แล้ว ดังนั้นสับเซตขนาด 2 และขนาด 1 ของ Itemset ดังกล่าวต้องเป็น Frequent ด้วย ดังนี้

เซตของ Frequent 2-itemsets = {CD, CE, DE}
เซตของ Frequent 1-itemsets = {C, D, E}

ในทางตรงกันข้าม ถ้า {AB} ไม่เป็น Frequent Itemset หรือเรียกว่าเป็น Infrequent Itemset ดังนั้นทุกๆ Superset ของ Itemset ดังกล่าวก็จะเป็น Infrequent Itemset ด้วย ดังแสดงได้ด้วยโครงสร้างแลตทิซในรูปที่ 2-6 ซึ่งหลักการในการกำจัด Infrequent Itemset นี้เรียกว่า Support-based Pruning กล่าวคือในการตรวจนับความถี่ของ Itemset ใดๆ อยู่ แล้วพบว่าค่าสนับสนุนหรือค่าความถี่น้อยกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำ เราสามารถกำจัด Itemset นั้นๆ ออกไป รวมถึงไม่จำเป็นต้องพิจารณาตรวจนับทุกๆ Superset ของ Itemset นั้นๆ ด้วยเช่นกัน



รูปที่ 2-6 Itemset Lattice กรณีกำจัด Infrequent Itemsets

ด้วยวิธีการเช่นนี้จะทำให้จำนวน Candidate Itemset ลดขนาดลง คุณสมบัตินี้มีชื่อเรียกว่า anti-monotone ซึ่งมีนิยามดังนี้

$$\forall X,Y \in J : (X \subseteq Y) \rightarrow f(Y) \leq f(X) \qquad \text{...(9)}$$

จาก (9) ซึ่งหมายความว่า ถ้า X เป็นสับเซตของ Y แล้ว f(Y) จะต้องไม่มากกว่า f(X) จากคุณสมบัติ Anti-monotone ใน (9) เราสามารถนำมาประยุกต์ให้ f เป็นค่าสนับสนุน ก็จะสอดคล้องกับการกำจัด Infrequent Itemset ที่เรียกว่า Support-based Pruning ที่กล่าวแล้วข้างต้นนั่นเอง

2.8 การทบทวนวรรณกรรม/สารสนเทศ (*Information*) ที่เกี่ยวข้อง

Murat Karabatak และ M. Cevdet Ince (2009) นำเสนองานวิจัยเรื่อง A New Feature Selection Method Based on Association Rules for Diagnosis of Erythemato-quamous Diseases ได้นำเสนองานวิธีการเลือกลักษณะบนพื้นฐานของกฎความสัมพันธ์และโครงข่ายประสาทเทียม ถูกนำเสนอสำหรับการวินิจฉัยโรค Erythemato-squamous ซึ่งคือโรคผิวหนังชนิดหนึ่ง โดยกฎความสัมพันธ์ใช้เพื่อลดจำนวนลักษณะของข้อมูล และโครงข่ายประสาทเทียมใช้สำหรับกระบวนการการจำแนกกลุ่ม และเปรียบประสิทธิภาพกับวิธีการเลือกลักษณะวิธีอื่น หลังจากใช้กฎความสัมพันธ์เลือกลักษณะสามารถลดจำนวนจาก 34 ลักษณะ เหลือ 24 ลักษณะ มีอัตราการจำแนกกลุ่มถูกต้อง 98.61% ซึ่งให้ค่าร้อยละของความถูกต้องมากกว่ากับข้อมูลที่ไม่ได้ผ่านการเลือกลักษณะและการเลือกลักษณะวิธีอื่นๆ ผลการทดลองแสดงให้เห็นว่าการเลือกลักษณะมีความสำคัญ และทำให้การจำแนกกลุ่มข้อมูลเพื่อวินิจฉัยโรค Erythemato-squamous ได้อย่างมีประสิทธิภาพ

Mansour Sheikhan และ Zahra Jadidi (2009) นำเสนองานวิจัยเรื่อง Misuse Detection Using Hybrid of Association Rule Mining and Connectionist Modeling โดยวิธีการที่เสนอเป็นการรวมการจำแนกกลุ่มเพอร์เซ็ปตรอนแบบหลายชั้นกับกฎความสัมพันธ์ เพื่อจำแนกกลุ่มข้อมูลตรวจจับการบุกรุก KDDcup99 จำนวน 5 คลาส โดยเปรียบเทียบกับการจำแนกกลุ่มเพอร์-เซ็ปตรอนแบบหลายชั้น ซึ่งผลการทดลอง การจำแนกกลุ่มด้วยเพอร์เซ็ปตรอนแบบหลายชั้นสามารถจำแนกกลุ่มได้ดีกับคลาส DoS และ Probe แต่ให้ผลไม่ดีกับคลาส R2L และ U2R แต่วิธีการที่นำเสนอสามารถจำแนกกลุ่มได้ดีกับทุกคลาส และได้ผลอัตราการตรวจจับที่ดีกว่า แม้ว่าค่าความผิดพลาดเชิงบวกจากวิธีการที่นำเสนอจะมากกว่าการจำแนกกลุ่มด้วยเพอร์เซ็ปตรอนแบบหลายชั้น แต่ยังได้ผลที่ดีกว่าการจัดกลุ่มโดยการหาค่าเฉลี่ยแบบเคและโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน

Onur Inan Mustafa Serter Uzer, และ Nihat Y.lmaz (2013) นำเสนองานวิจัยเรื่อง A New Hybrid Feature Selection Method Based on Association Rules and PCA for Detection of Breast Cancer ซึ่งนำเสนอวิธีการผสมในการเลือกลักษณะสำหรับการตรวจจับโรคมะเร็งหน้าอก โดยวิธีการที่นำเสนอเป็นการรวมกันระหว่างกฎความสัมพันธ์และการวิเคราะห์องค์ประกอบหลัก กฎความสัมพันธ์หรือขั้นตอนวิธี Apriori เป็นเทคนิคในการเลือกลักษณะที่เหมาะสม จากนั้นนำลักษณะที่ได้ผ่านการวิเคราะห์องค์ประกอบหลักและจำแนกกลุ่มด้วยโครงข่ายประสาทเทียม ซึ่งได้ทดสอบกับชุดข้อมูล Wisconsin Breast Cancer จำนวน 9 ลักษณะ เมื่อเลือกลักษณะด้วยขั้นตอนวิธีที่นำเสนอจะเหลือลักษณะจำนวน 6 ลักษณะ จากการทดสอบวิธีการที่นำเสนอสามารถลดจำนวนลักษณะ ใช้เวลาการเรียนรู้ในการจำแนกกลุ่มได้รวดเร็ว และเมื่อเปรียบกับวิธีการอื่น ๆ ปรากฏว่าวิธีการที่นำเสนอส่วนใหญ่ให้ค่าความถูกต้องในการวินิจฉัยโรคมะเร็งหน้าอกได้มากกว่าวิธีการอื่น ๆ

Dong Seong Kim, Ha-nam Nguyen, Thanda Thein และ Jong Sou Park (2005) ได้นำเสนองานวิจัยเรื่อง An Optimized Intrusion Detection System Using PCA and BPNN โดยได้นำเสนอการหาค่าที่เหมาะสมสำหรับการตรวจจับการบุกรุกโดยอาศัยการวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis: PCA) และโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ

(Backpropagation Neural Network: BPNN) โดยมุ่งเน้นในการแก้ปัญหา 2 ปัญหาด้วยกันคือ การกำหนดจำนวนของ Hidden Layer และการจัดการค่าของน้ำหนักเพื่อใช้ในการกำหนดรูปแบบของโครงข่ายประสาทเทียม และการประมวลผลข้อมูลที่ตรวจสอบที่มีปริมาณมาก โดยพิจารณาถึงการเพิ่มอัตราการตรวจจับและลดเวลาการประมวลผล โดยนำข้อดีของ Genetic Algorithm (GA) มาใช้ โดยการทำงานของ GA จะทำงานบนการทำงานที่รวมกันระหว่าง PCA และ BPNN แต่ผลการทดลองยังออกมาไม่เป็นที่น่าพอใจตามที่คาดหวังไว้ ในส่วนงานในอนาคตได้มีการชี้ถึงประเด็นว่า ถ้ามีการปรับเปลี่ยนตัว PCA และ BPNN น่าจะทำให้ได้ผลการทดลองที่ดีขึ้น

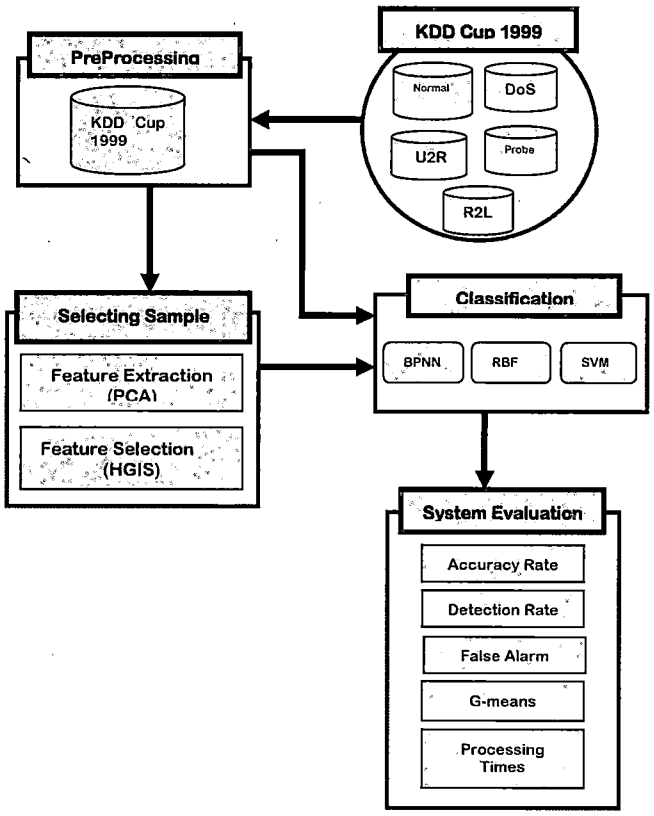
Hai-Hua Gao, Hui-Hua Yang และ Xing-Yu Wang (2005) ได้นำเสนองานวิจัยเรื่อง Kernel PCA Based Network Intrusion Feature Extraction and Detection Using SVM โดยได้นำเสนอวิธีการใหม่ในการตรวจจับการบุกรุกด้วยการประยุกต์ Kernel Principal Component Analysis: KPCA สำหรับการสกัดลักษณะและใช้ Support Vector Machine: SVM ในการแบ่งประเภท โดยทำการเปรียบเทียบผลกับข้อมูลที่ไม่ได้ผ่านการสกัดลักษณะ และการสกัดลักษณะด้วยวิธีการ PCA โดยผลการทดลองชี้ให้เห็นว่าการสกัดลักษณะของข้อมูลสามารถลดขนาดของข้อมูลนำเข้าโดยไม่ทำให้ประสิทธิภาพในการแบ่งกลุ่มลดลง ซึ่งการทดลองด้วย SVM ใช้ข้อมูลเพียง 4 ลักษณะหลักที่สกัดได้จาก KPCA ก็ทำให้ได้ผลลัพธ์ที่ดีกว่าชุดข้อมูลที่ไม่ผ่านการสกัด และชุดข้อมูลที่ผ่านการสกัดด้วย PCA

Hai-Hua Gao, Hui-Hua Yang และ Xing-Yu Wang (2005) ได้นำเสนองานวิจัยเรื่อง Principal Component Neural Networks Based Intrusion Feature Extraction and Detection Using SVM โดยได้นำเสนอวิธีการใหม่ในการสกัดลักษณะชุดข้อมูลการบุกรุก โดยการประยุกต์ใช้ Principal Component Neural Network: PCNN และนำผลลัพธ์ที่ได้จากการสกัดลักษณะมาทำการแบ่งกลุ่มด้วย SVM โดยที่ใช้อัลกอริทึม Adaptive Principal Component Extraction: APEX มาดัดแปลงให้เหมาะสมในการทำงานของ PCNN โดยผลที่ได้จากการทดลองนำมาเปรียบเทียบกับ SVM ที่ไม่ได้ทำการสกัดลักษณะชุดข้อมูล ผลการทดลองแสดงให้เห็นชัดว่า การสกัดลักษณะด้วย PCNN สามารถลดจำนวนลักษณะของข้อมูลนำเข้า และไม่ทำให้ประสิทธิภาพในการตรวจจับการบุกรุกลดลง

Zhu Xiaorong, Wang Dianchun และ Ye Changguo (2009) ได้นำเสนองานวิจัยเรื่อง A New Feature Extraction Method of Intrusion Detection โดยได้นำเสนอวิธีการนำเอา Kernel Principal Component Analysis: KPCA มาทำการสกัดลักษณะจากการตัวอย่างของข้อมูลการบุกรุกที่จะใช้ฝึกฝน โดยที่วิธีการนี้สกัดลักษณะและลดจำนวนลักษณะของข้อมูลได้อย่างมีประสิทธิผล โดยได้นำเสนอวิธีการนำ Reduce SVM: RSVM ร่วมกับวิธีการ Nonlinear Proximal SVM ซึ่งวิธีการที่นำเสนอนี้สามารถลดความซับซ้อนในการคำนวณของ Kernel Matrix ได้ และยังส่งผลให้ความเร็วในการฝึกฝนและผลลัพธ์ของการแบ่งกลุ่มดีขึ้น

บทที่ 3 วิธีดำเนินการวิจัย

การดำเนินการวิจัยมีขอบเขตการวิจัย ดังแสดงในรูปที่ 3-1 เพื่อให้เกิดความเข้าใจในวิธีการดำเนินการวิจัย ผู้วิจัยจะอธิบายการดำเนินการวิจัยเป็นส่วนๆ ดังต่อไปนี้



รูปที่ 3-1 Intrusion Detection Framework ซึ่งใช้ในงานวิจัยนี้

3.1 การจัดการชุดข้อมูล

ข้อมูลที่นำมาใช้ในการทำแบบทดลอง เป็นข้อมูลที่ได้จากฐานข้อมูลความรู้ (Knowledge Discovery in Database (KDD) Cup data) ซึ่งเป็นชุดข้อมูลในปี 1999 ชุดข้อมูลนี้ถูกสร้างจากการจำลองการโจมตีของผู้บุกรุกจาก U.S. Air Force Local Area Network มีจำนวนประมาณ 4,900,000 จุดข้อมูล มี 42 ลักษณะ ซึ่งข้อมูลอยู่ในรูปแบบของสัญลักษณ์ และจำนวนจริง โดยลักษณะสุดท้ายคือคลาสที่บ่งบอกว่าข้อมูลชุดใดเป็นลักษณะปกติหรือบุกรุก มีลักษณะของข้อมูลที่สมบูรณ์ ซึ่งแบ่งออกเป็น 5 ประเภทใหญ่ คือ 1) Normal 2) Dos 3) Probe 4) R2L และ 5) U2R โดยที่แต่ละประเภทมีจำนวนไม่เท่ากัน

เนื่องจากข้อมูล KDDcup99 มีจำนวนมาก ดังนั้น ในงานวิจัยส่วนใหญ่จึงแนะนำให้เลือกข้อมูลเพียงร้อยละ 10 และเพื่อสะดวกในการสอนและทดสอบประสิทธิภาพของระบบการรู้จำจึงทำการสุ่มข้อมูลมาประมาณ 13,499 จุดข้อมูล (Patterns) โดยแบ่งเป็นประเภท Normal จำนวน 4,107 จุดข้อมูล Dos จำนวน

4,107 จุดข้อมูล Probe จำนวน 4,107 จุดข้อมูล R2L จำนวน 1,126 จุดข้อมูล และ U2R จำนวน 52 จุดข้อมูล และตัดบางลักษณะที่ไม่มีผลต่อการรู้จำออกไป เช่น Basic Features และลักษณะที่มีค่าเป็นศูนย์ทั้งหมด จึงเหลือจำนวนลักษณะ 34 ลักษณะเมื่อได้ข้อมูลมาแล้วทำการแบ่งข้อมูลออกเป็นสองกลุ่มเพื่อทำแบบทดลอง โดยกลุ่มที่ 1 ใช้ในการฝึกฝน และกลุ่มที่ 2 ใช้ในการทดสอบ ซึ่งวิธีการแบ่งจะใช้วิธีการสุ่มข้อมูลจากข้อมูลทั้งหมด ออกเป็นข้อมูลเพื่อทำแบบทดลอง

3.2 การเลือกลักษณะชุดข้อมูล

การเลือกลักษณะชุดข้อมูล จะใช้วิธีการที่ได้ศึกษาจากวิธีการสกัดลักษณะชุดข้อมูลด้วยวิธีวิเคราะห์องค์ประกอบหลักเปรียบเทียบกับการเลือกลักษณะชุดข้อมูลด้วยวิธีฮิวริสติกกรีดของไอเท็มเซตโดยใช้หลักการ Apriori มาทำการสกัดลักษณะและเลือกชุดข้อมูล เพื่อนำไปทดสอบในขั้นตอนการฝึกฝนต่อไป

3.2.1 การสกัดลักษณะด้วยวิธีวิเคราะห์องค์ประกอบหลัก *(PCA)*

ขั้นตอนการสกัดลักษณะด้วยวิธีวิเคราะห์องค์ประกอบมีดังนี้

ให้ $x_1, x_2, \dots x_M$ คือ เวกเตอร์ $N \times 1$

1. $\bar{x} = \frac{1}{M}\sum_{i=1}^{M} x_i$
2. $\varphi_i = x_i - \bar{x}$
3. เมทริกซ์ $A = [\varphi_1, \varphi_2, \dots \varphi_M]$ มีขนาด $N \times M$ แล้วคำนวณหาความแปรปรวนร่วม

$$C = \frac{1}{M}\sum_{N=1}^{M} \varphi_N \varphi_N \qquad \dots(10)$$

4. คำนวณค่าไอเกน $C: \lambda_1 > \lambda_2 > .. > \lambda_N$
5. คำนวณไอเกนเวกเตอร์$C: u_1, u_2, \dots u_N$
6. นำไอเกนเวกเตอร์คูณกับข้อมูลเดิม ซึ่งจะได้ข้อมูลใหม่$xPca = u \times x'$
7. เลือกองค์ประกอบหลัก K

$$\frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{N} \lambda_i} > threshold \qquad \dots(11)$$

λ_iคือ ค่าไอเกนลำดับที่ i

N คือ จำนวนลักษณะทั้งหมด

Kคือ จำนวนลักษณะที่ถูกเลือก

$threshold$ คือ ค่าเกณฑ์ที่บ่งบอกว่าต้องการให้องค์ประกอบหลักที่ได้มีค่าไอเกนสะสมใกล้เคียงกับค่าไอเกนสะสมทั้งหมดมากน้อยเพียงใด ในที่นี้กำหนดให้ $threshold$ เท่ากับ 0.95

3.2.2 การเลือกลักษณะด้วยวิธีฮิวริสติกกรีดดีของไอเท็มเซต (Heuristic Greedy Item Sets: HGIS) โดยใช้หลักการ Apriori

การเลือกลักษณะชุดข้อมูล ใช้วิธีการเลือกวิธีฮิวริสติกกรีดดีของไอเท็มเซตโดยใช้หลักการ Apriori ซึ่งมีวิธีการดังนี้

ขั้นตอนที่ 1: สร้าง 1-itemset โดยนำแต่ละลักษณะหาค่า RMSE (Root Mean Square Error) โดยใช้โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน

ขั้นตอนที่ 2: สร้าง 2-candidate itemset โดยการนำแต่ละลักษณะมาจับคู่กันทุกๆลักษณะที่เป็นไปได้ และหาค่า RMSE โดยใช้โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน

ขั้นตอนที่ 3: สร้าง 2-itemset โดยนำ 2-candidate itemset ที่มีค่า RMSE น้อยกว่า 1-itemset ของตัวมันเอง

ขั้นตอนที่ 4: สร้าง 3-candidate itemset โดยนำ 2-itemset จำนวน 3 เซตมายูเนียนกันทุกๆ 3 เซต ที่เป็นไปได้และหาค่า RMSE โดยใช้โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน

ขั้นตอนที่ 5: นำ 3-candidate itemset หาค่า RMSE โดยใช้โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน

ขั้นตอนที่ 6: เลือกเซตลักษณะโดยการสุ่มเลือกจาก 2-itemset และ 3-candidate itemset หาก itemset ใด มีค่า RMSE ต่ำ จะมีโอกาสสุ่มเลือกมากกว่า

3.3 การรู้จำด้วยโครงข่ายประสาทเทียม

ในขั้นตอนนี้จะทำการเรียนรู้ด้วยโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ ซัพพอร์ตเวกเตอร์แมชชีน และ โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐานที่มีการปรับปรุงการทำงานให้สามารถทำงานได้ดีในการฝึกฝนและการทดสอบ

3.4 การประเมินระบบ

นำผลจากการเรียนรู้ที่ได้มาทำการประเมินระบบโดยนำข้อมูลที่ผ่านการเรียนรู้และไม่เคยผ่านการเรียนรู้มาทดสอบระบบ จากนั้นหาค่าร้อยละของความถูกต้อง (Accuracy) อัตราการตรวจจับ (Detection Rate) อัตราความผิดพลาดเชิงบวก (False Alarm Rate) ค่าเฉลี่ยเรขาคณิต (Geometric Means) ตามสมการที่ (12) (13) (14) และ (15) ตามลำดับ เวลาที่ใช้ในการทดสอบ และจำนวนข้อมูลที่แบ่งประเภทได้ถูกต้องของคลาสคำตอบ

การวัดประสิทธิภาพจากการจำแนกกลุ่มที่ถูกต้องโดยรวมได้จากสมการ (12)

$$\text{Accuracy } Rate = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad \text{...(12)}$$

การวัดการตรวจจับผู้บุกรุกที่ถูกจับได้ถูกต้องทั้งหมดวัดได้จากสมการที่ (13)

$$Detection\ Rate = \frac{TP}{TP+FN} \quad ...(13)$$

การวัดการตรวจจับที่ถูกตรวจจับว่าเป็นผู้บุกรุกซึ่งที่จริงคือปกติ เป็นการตรวจจับที่ไม่ควรเกิดขึ้นหรือควรพยายามให้เกิดขึ้นน้อย วัดได้จากสมการที่ (14)

$$False\ Alarm\ Rate = \frac{FP}{FP+TN} \quad ...(14)$$

เนื่องจากในแต่ละคลาสมีจำนวนที่แตกต่างกัน ดังนั้นเพื่อวัดประสิทธิของการจำแนกกลุ่มที่ถูกต้องในแต่ละคลาส จึงใช้การหาค่าเฉลี่ยเรขาคณิตของอัตราความถูกต้องของการจำแนกกลุ่มในแต่ละคลาสดังสมการที่ (15)

$$GM = \sqrt[N]{\Pi TPR_i} \quad ...(15)$$

โดยที่ True Positive (TP) คือ จำนวนที่ถูกตรวจจับว่าเป็นผู้บุกรุก ซึ่งที่จริงคือผู้บุกรุก
True Negative (TN) คือ จำนวนที่ถูกตรวจจับว่าปกติ ซึ่งที่จริงคือปกติ
False Positive (FP) หรือ false alarm คือ จำนวนที่ถูกตรวจจับว่าเป็นผู้บุกรุก ซึ่งที่จริงคือ ปกติ
False Negative (FN) คือ จำนวนที่ถูกตรวจจับว่าปกติ ซึ่งที่จริงคือผู้บุกรุก
N คือ จำนวนของคลาสทั้งหมด
TPR_i คือ อัตรา True Positive Rate ของคลาสที่ i

เพื่อไม่ให้เกิดความสับสนจึงสรุปตามตารางที่ 3-1

ตารางที่ 3-1 Confusion Matrix

Actual \ Predicted	Normal	Attack
Normal	True Negative (TN)	False Positive (FP)
Attack	False Negative (FN)	True Positive (TP)

บทที่ 4 ผลการทดลอง

ในบทนี้จะกล่าวถึงผลการทดลอง โดยผลการทดลองที่นำเสนอจะประกอบด้วย ผลการทดลองจากการสกัดลักษณะข้อมูล การเลือกลักษณะข้อมูล และ ขั้นตอนการรู้จำประเภทของผู้บุกรุกโดยใช้โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ ซัพพอร์ตเวกเตอร์แมชชีน และโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน

4.1 การสกัดลักษณะและการเลือกลักษณะข้อมูล

จากตัวอย่างข้อมูล KDDcup99 ซึ่งเป็นฐานข้อมูลมาตรฐานและนำมาใช้เป็นตัวอย่างในการทดลองในงานวิจัยนี้โดยข้อมูลชุดนี้มีจำนวนประมาณ 4,900,000 จุดข้อมูล เป็นข้อมูล 41 ลักษณะ โดยในแต่ละชุดข้อมูล (Pattern) ของ KDDcup99 นี้ จะแบ่งออกเป็นสามกลุ่มใหญ่คือ

1. Basic Features ลักษณะ (Attributes/Features) ของข้อมูลในกลุ่มนี้จะบอกได้ว่าเป็นข้อมูลที่ได้มาจาก TCP/IP
2. Traffic Features ลักษณะของข้อมูลในกลุ่มนี้จะเป็นค่าของเวลาที่ได้จากการคำนวณ โดยแบ่งออกเป็นสองประเภทย่อย คือ Same Host Features และ Same Service Features
3. Content Features ลักษณะของข้อมูลในกลุ่มนี้จะไว้บอกลักษณะของการบุกรุก เนื่องจาก DoS และ Probing จะมีพฤติกรรมการบุกรุกที่แตกต่างจาก R2L และ U2L เช่น DoS และ Probing จะมีพฤติกรรมในการโจมตีเครื่องแม่ข่ายบางเครื่อง (Some Hosts) ในช่วงเวลาสั้น ๆ แต่ในขณะที่การโจมตีของ R2L และ U2L จะถูกรวมมากับส่วนของข้อมูล ซึ่งโดยปกติแล้วจะโจมตีมาเพียงครั้งเดียว

```
0,tcp,pop_3,RSTO,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,211,6,0.00,0.00,1.00,1.00,0.03,0.07,0.00,255,6,0.02,0.07,0.00,0.00,0.00,0.00,1.00,1.00,neptune.
0,tcp,pop_3,RSTO,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,231,16,0.00,0.00,1.00,1.00,0.07,0.06,0.00,255,16,0.06,0.07,0.00,0.00,0.00,0.00,1.00,1.00,neptune.
0,tcp,pop_3,RSTO,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,232,5,0.00,0.00,1.00,1.00,0.02,0.06,0.00,255,5,0.02,0.07,0.00,0.00,0.00,0.00,1.00,1.00,neptune.
0,tcp,pop_3,RSTO,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,254,15,0.00,0.00,1.00,1.00,0.06,0.07,0.00,255,15,0.06,0.07,0.00,0.00,0.00,0.00,1.00,1.00,neptune.
0,tcp,pop_3,RSTO,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,252,6,0.00,0.00,1.00,1.00,0.02,0.07,0.00,255,6,0.02,0.08,0.00,0.00,0.00,0.00,1.00,1.00,neptune.
0,tcp,pop_3,RSTO,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,272,16,0.00,0.00,1.00,1.00,0.06,0.06,0.00,255,16,0.06,0.07,0.00,0.00,0.00,0.00,1.00,1.00,neptune.
0,tcp,pop_3,SH,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1.00,1.00,0.00,0.00,1.00,0.00,0.00,255,1,0.00,1.00,1.00,0.00,1.00,1.00,0.00,0.00,nmap.
0,tcp,pop_3,SH,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1.00,1.00,0.00,0.00,1.00,0.00,0.00,255,1,0.00,1.00,1.00,0.00,1.00,1.00,0.00,0.00,nmap.
5,tcp,pop_3,SF,6,151,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,511,1,0.07,0.00,0.91,0.00,0.00,1.00,0.00,255,1,0.00,1.00,0.00,0.00,0.07,0.00,0.90,0.00,satan.
40339,tcp,pop_3,RSTR,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,1.00,1.00,1.00,0.00,0.00,255,2,0.01,0.44,0.86,0.00,0.00,0.00,0.86,1.00,portsweep.
```

รูปที่ 4-1 ตัวอย่างข้อมูล Intrusion สำหรับการทดสอบการรู้จำ

4.1.1 ลักษณะข้อมูลของ KDDcup99 จำนวน 34 ลักษณะ

เนื่องจากข้อมูล KDDcup99 มีจำนวนมาก ดังนั้น ในงานวิจัยส่วนใหญ่จึงแนะนำให้เลือกข้อมูลเพียงร้อยละ 10 สำหรับเป็นข้อมูลในการสอนและทดสอบประสิทธิภาพของระบบการรู้จำ และจากข้อมูลที่เลือกมาร้อยละ 10 นั้น จะทำการสุ่มมาทำข้อมูลในการสอนประมาณ 13,499 ชุด (Patterns) ทำการสกัดลักษณะโดยใช้ขั้นตอนวิธีการวิเคราะห์องค์ประกอบหลักจำนวน 19 ลักษณะโดยเราจะพิจารณาจากค่าไอเกนของข้อมูลและเปรียบเทียบจากการเลือกลักษณะด้วยวิธีฮิวริสติกกรีดดีของไอเท็มเซตโดยใช้หลักการ Apriori จำนวน 13 ลักษณะซึ่งทั้งสองได้ลักษณะเด่นจากข้อมูลทั้งหมด 34 ลักษณะสำหรับการรู้จำด้วยโครงข่ายประสาทเทียม

(จากที่กล่าวไว้ก่อนหน้านี้ว่าข้อมูล KDDcup99 เป็นข้อมูลทั้งหมด 41 ลักษณะ แต่ลักษณะที่เป็น Basic Featuresลักษณะที่มีค่าเป็นศูนย์ทั้งหมด และลักษณะที่เป็นคำตอบจะไม่นำมาพิจารณา ดังนั้น จึงเหลือเพียง 34 ลักษณะ) ซึ่งในแต่ละลักษณะมีค่าสูงสุด ค่าต่ำสุด ค่าเฉลี่ย และค่าส่วนเบี่ยงเบนมาตรฐานตามตารางที่ 4-1 ในตารางที่ 4-2 (ก) และ 4-2 (ข) แสดงค่าสหสัมพันธ์ระหว่างแต่ละลักษณะทั้ง 34 ลักษณะ และรูปที่ 4-2 ถึง 4-35 แสดงให้เห็นถึงการกระจายตัวของข้อมูลในแต่ละลักษณะ ในลักษณะที่ 1 ถึง ลักษณะที่ 34

ตารางที่ 4-1 ค่าทางสถิติของข้อมูล KDDcup99 จำนวน 34 ลักษณะ

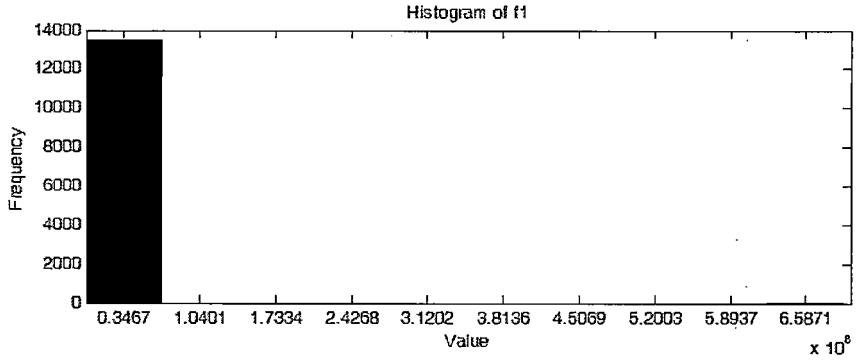
Features	Maximum	Minimum	Mean	Standard Deviation
f1	6.93E+08	0	74547.73	5977289
f2	5155468	0	7012.817	172422.5
f3	3	0	0.001852	0.073517
f4	2	0	0.000296	0.021081
f5	30	0	0.640121	4.039457
f6	5	0	0.004223	0.07541
f7	1	0	0.301207	0.4588
f8	38	0	0.012742	0.426195
f9	1	0	0.002371	0.048632
f10	1	0	7.41E-05	0.008607
f11	54	0	0.018816	0.650588
f12	21	0	0.007704	0.243328
f13	2	0	0.000889	0.036507
f14	2	0	0.002593	0.056382
f15	1	0	0.024446	0.154436
f16	511	0	182.1567	229.4395
f17	511	0	118.6734	207.4745
f18	1	0	0.08962	0.265435
f19	1	0	0.090148	0.284962
f20	1	0	0.207335	0.391144
f21	1	0	0.20698	0.404462
f22	1	0	0.794391	0.388862
f23	1	0	0.135309	0.327396
f24	1	0	0.109987	0.29354
f25	255	1	180.9078	106.9052
f26	255	1	138.053	117.0673
f27	1	0	0.64521	0.455144
f28	1	0	0.203488	0.371824
f29	1	0	0.497511	0.481672
f30	1	0	0.073159	0.194441
f31	1	0	0.090221	0.263441
f32	1	0	0.090133	0.284373
f33	1	0	0.201948	0.378601
f34	1	0	0.205992	0.401915

ตารางที่ 4-2 (ก) ค่าสหสัมพันธ์ระหว่างแต่ละลักษณะของข้อมูล KDDcup99 จำนวน 34 ลักษณะ

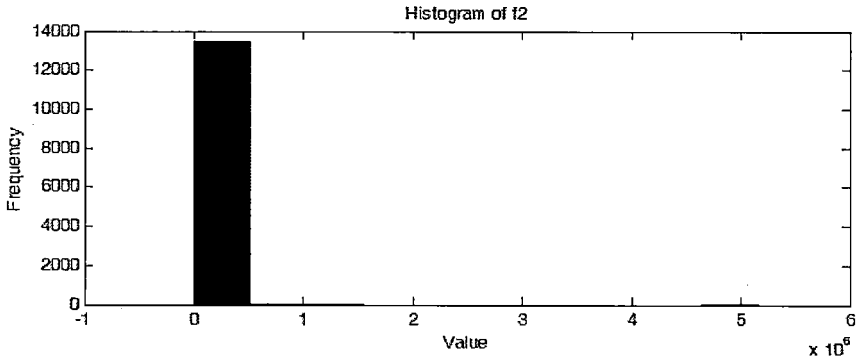
	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14	f15	f16	f17
f1	1																
f2	-0.00051	1															
f3	-0.00031	-0.00102	1														
f4	-0.00017	0.001262	-0.00035	1													
f5	0.002711	-0.00415	-0.00399	-0.00136	1												
f6	-0.0007	-0.00222	-0.00141	-0.00079	0.004503	1											
f7	0.000182	-0.01356	-0.01654	0.02141	0.238899	-0.03676	1										
f8	-0.00033	0.001165	-0.00075	0.057299	0.031237	-0.00167	0.045538	1									
f9	-0.0006	-0.00046	-0.00123	0.071576	0.02433	-0.00273	0.074247	0.384572	1								
f10	-0.00011	-0.00027	-0.00022	-0.00012	-0.00136	-0.00048	0.01311	-0.00026	-0.00042	1							
f11	-0.00035	0.000546	-0.00073	0.08602	0.022479	-0.00162	0.044054	0.886198	0.279574	-0.00025	1						
f12	-0.00039	0.000419	-0.0008	0.013997	0.035834	-0.00177	0.048228	0.249086	0.192534	0.035102	0.213889	1					
f13	-0.0003	-0.00013	-0.00061	-0.00034	0.01071	-0.00136	0.037091	0.437335	0.416098	0.235571	0.44847	0.291128	1				
f14	-0.00057	-0.00039	-0.00116	0.124012	0.007025	-0.00258	0.070046	0.248353	0.186889	0.15227	0.224875	0.079545	0.178845	1			
f15	-0.00194	-0.0043	-0.00399	-0.00223	0.956204	-0.00886	0.241114	-0.00473	-0.00772	-0.00136	-0.00458	0.046246	-0.00385	0.009737	1		
f16	-0.00767	-0.03196	-0.0155	-0.0111	-0.12504	-0.04409	-0.50251	-0.02357	-0.03844	-0.0068	-0.02281	-0.02469	-0.01919	-0.03594	-0.12499	1	
f17	-0.0069	-0.02283	-0.00943	-0.00797	-0.08981	-0.03163	-0.34901	-0.01691	-0.0276	-0.00488	-0.01637	-0.01795	-0.01377	-0.02528	-0.08978	0.748042	1
f18	0.021607	-0.01265	-0.00851	-0.00475	-0.05265	-0.0115	-0.2149	-0.00996	-0.01646	-0.00291	-0.00977	-0.00842	-0.00822	-0.01553	-0.05345	0.043823	-0.17812
f19	0.016516	-0.01186	-0.00797	-0.00445	-0.04925	-0.01082	-0.20137	-0.00934	-0.01542	-0.00272	-0.00915	-0.01002	-0.0077	-0.01455	-0.05008	0.041531	-0.16715
f20	-0.00192	-0.02154	-0.01335	-0.00745	-0.0804	0.093389	-0.34329	-0.01563	-0.02389	-0.00456	-0.01533	-0.01678	-0.01291	-0.02438	-0.08207	0.112904	-0.29873
f21	0.000707	-0.02078	-0.01289	-0.00719	-0.07831	0.090363	-0.3316	-0.01509	-0.02306	-0.0044	-0.0148	-0.01545	-0.01246	-0.02126	-0.07983	0.10902	-0.28834
f22	-0.01448	0.021448	0.013321	0.007433	0.082067	0.029609	0.337956	0.015585	0.022837	0.004551	0.01423	0.014018	0.012876	0.019789	0.08247	-0.32361	0.290055
f23	0.005279	-0.01669	-0.01041	-0.00581	-0.06178	-0.02314	-0.25457	-0.01183	-0.01317	-0.00356	-0.00939	-0.01017	-0.01006	-0.01363	-0.0625	0.363572	-0.23339
f24	-0.00392	-0.01212	-0.00714	-0.00527	-0.05824	-0.02098	-0.00216	-0.01076	-0.01308	-0.00323	-0.00578	-0.00772	-0.00912	-8.7E-05	-0.05768	-0.29435	-0.17942
f25	0.000666	-0.05889	0.011531	-0.01521	0.059787	-0.08019	-0.36791	-0.02896	-0.05143	0.005966	-0.02266	-0.02114	-0.02215	-0.01241	0.066741	0.540264	0.361069
f26	-0.01321	-0.03416	-0.01186	-0.01637	-0.09659	-0.05482	0.189343	-0.02709	-0.04225	-0.00669	-0.02843	-0.03185	-0.02627	0.006063	-0.10096	0.158605	0.54897
f27	-0.0099	0.03004	-0.01537	-0.00819	-0.11794	0.041493	0.315514	0.010634	0.021164	-0.0088	-0.00828	-0.01225	-0.00068	0.022829	-0.12926	-0.02116	0.431606
f28	-0.0045	-0.02209	0.002555	0.02085	-0.0756	-0.03059	-0.33084	-0.01557	-0.02602	-0.00425	-0.01292	-0.01236	-0.01262	-0.01877	-0.07428	0.173425	-0.309
f29	-0.00152	0.030534	-0.00679	-0.00358	-0.15413	-0.03834	-0.3764	0.003334	-0.00114	-0.00889	0.000599	-0.01161	0.00594	-0.03059	-0.16019	0.210185	0.554466
f30	-0.00266	-0.01464	-0.00922	-0.00529	-0.05722	0.004193	-0.10666	-0.00692	0.0043	-0.00324	-0.00927	-0.00812	-0.00259	-0.01663	-0.05956	-0.29641	-0.19166
f31	0.001758	-0.01321	-0.0053	-0.00481	-0.0452	0.000736	-0.21317	-0.00908	-0.01669	0.00424	-0.00985	-0.0091	-0.00665	-0.01445	-0.04098	0.043235	-0.1809
f32	0.016615	-0.01223	-0.00799	-0.00446	-0.0492	0.000699	-0.20343	-0.00902	-0.01411	0.006958	-0.00875	-0.00855	-0.00543	-0.0131	-0.05007	0.041601	-0.16748
f33	-0.00546	-0.0216	-0.00106	-0.0075	-0.07768	0.088613	-0.33863	-0.01468	-0.0184	-0.00459	-0.01386	-0.01367	-0.00929	-0.02453	-0.07883	0.12433	-0.30017
f34	0.000671	-0.02074	-0.01291	-0.0072	-0.07872	0.082909	-0.32686	-0.01419	-0.02188	-0.00441	-0.01454	-0.01552	-0.01147	-0.02357	-0.07998	0.111673	-0.28867

ตารางที่ 4-2 (ข) ค่าสหสัมพันธ์ระหว่างแต่ละลักษณะของข้อมูล KDDcup99 จำนวน 34 ลักษณะ

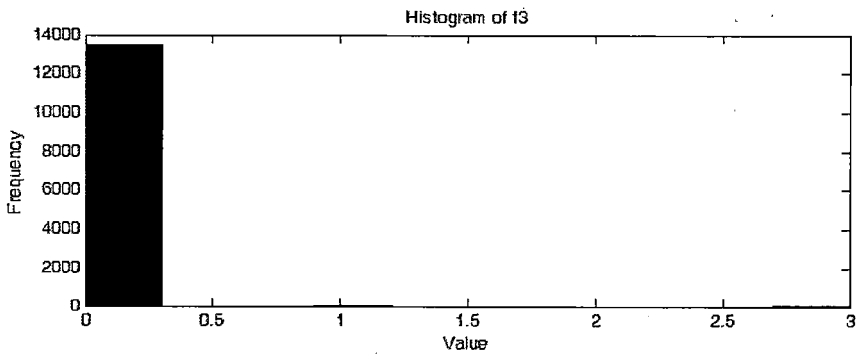
	f18	f19	f20	f21	f22	f23	f24	f25	f26	f27	f28	f29	f30	f31	f32	f33	f34
f18	1																
f19	0.927679	1															
f20	-0.07669	-0.06654	1														
f21	-0.07291	-0.15973	0.962791	1													
f22	-0.55315	-0.51906	-0.4886	-0.4698	1												
f23	0.060497	0.061472	0.625324	0.601707	-0.75729	1											
f24	-0.12361	-0.11574	-0.16304	-0.15638	0.194309	-0.14779	1										
f25	0.212846	0.201607	0.211832	0.204599	-0.35714	0.273199	-0.47505	1									
f26	-0.369	-0.34517	-0.53852	-0.51905	0.595353	-0.47284	0.14542	-0.05286	1								
f27	-0.44102	-0.41433	-0.62938	-0.60671	0.721102	-0.57138	0.242756	-0.41479	0.824219	1							
f28	0.064028	0.062622	0.770066	0.743525	-0.53585	0.739235	-0.19628	0.310536	-0.627	-0.7527	1						
f29	-0.27852	-0.26051	-0.16194	-0.15664	0.450316	-0.31187	0.128222	-0.12002	0.163082	0.325947	-0.11822	1					
f30	-0.12647	-0.11849	-0.10207	-0.09852	0.197615	-0.15341	0.444948	-0.607	-0.00364	0.227313	-0.13214	0.275446	1				
f31	0.987209	0.918195	-0.07349	-0.07058	-0.55285	0.061445	-0.12521	0.216912	-0.37489	-0.44757	0.065054	-0.28513	-0.12804	1			
f32	0.925022	0.996408	-0.06481	-0.15772	-0.51985	0.061593	-0.11664	0.20146	-0.34884	-0.4166	0.062772	-0.26062	-0.11814	0.919704	1		
f33	-0.07548	-0.06716	0.966496	0.93238	-0.49571	0.616781	-0.165	0.216166	-0.53974	-0.63102	0.795291	-0.15713	-0.10221	-0.07551	-0.06513	1	
f34	-0.0732	-0.15995	0.955675	0.991991	-0.47419	0.606586	-0.16055	0.20836	-0.52247	-0.6116	0.749228	-0.15765	-0.09852	-0.07054	-0.15824	0.932962	1



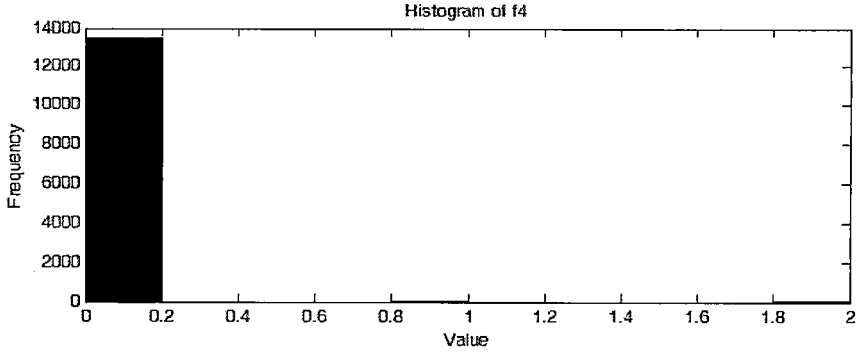
รูปที่ 4-2 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 1 (f1)



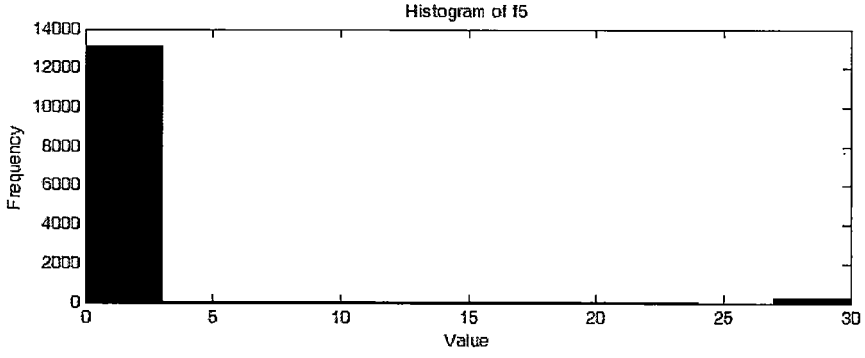
รูปที่ 4-3 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 2 (f2)



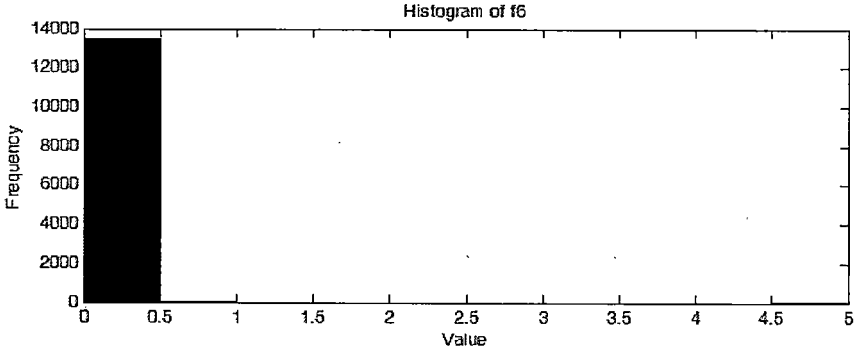
รูปที่ 4-4 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 3 (f3)



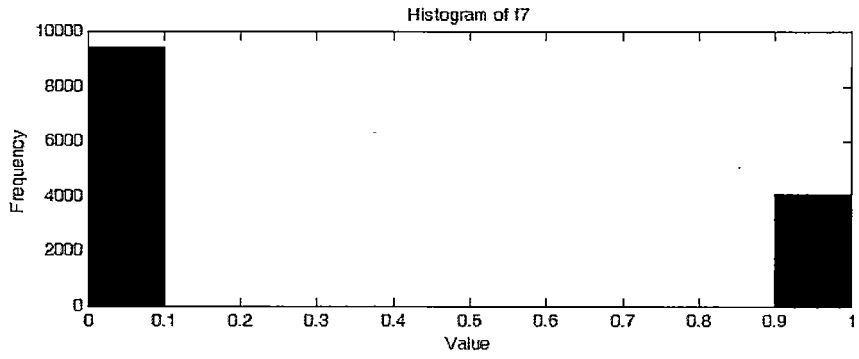
รูปที่ 4-5 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 4 (f4)



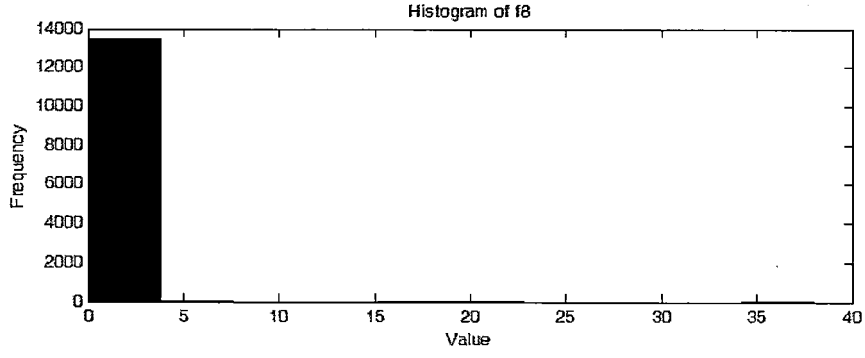
รูปที่ 4-6 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 5 (f5)



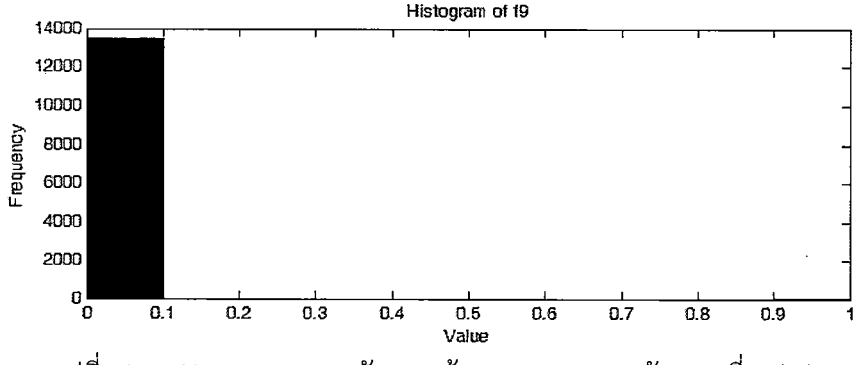
รูปที่ 4-7 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 6 (f6)



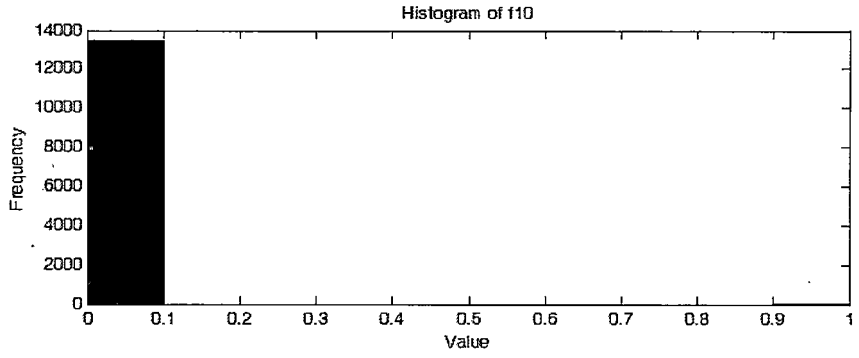
รูปที่ 4-8 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 7 (f7)



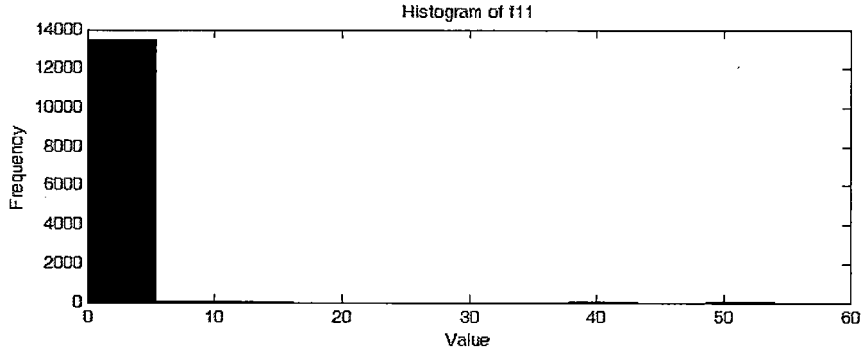
รูปที่ 4-9 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 8 (f8)



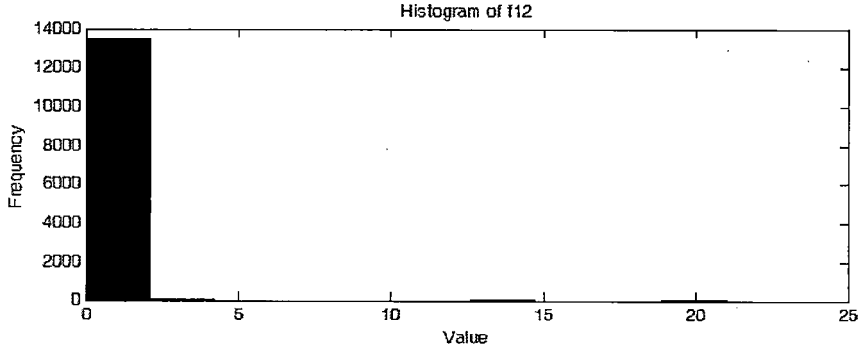
รูปที่ 4-10 Histogram ของลักษณะข้อมูล KDDcup99ลักษณะที่ 9 (f9)



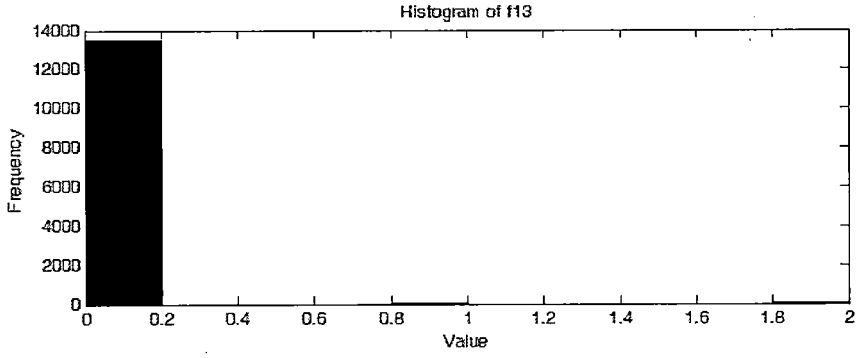
รูปที่ 4-11 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 10 (f10)



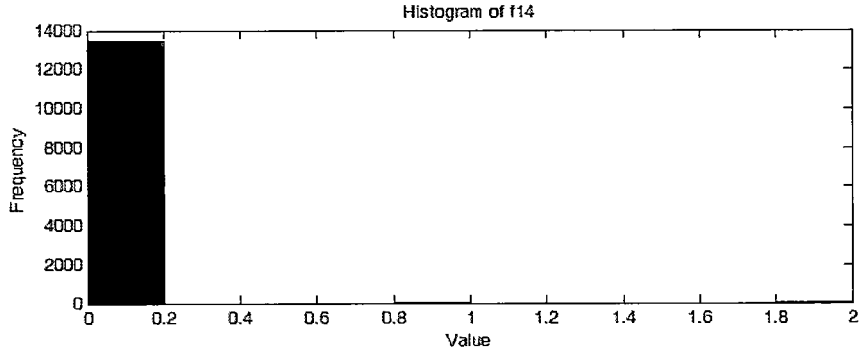
รูปที่ 4-12 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 11 (f11)



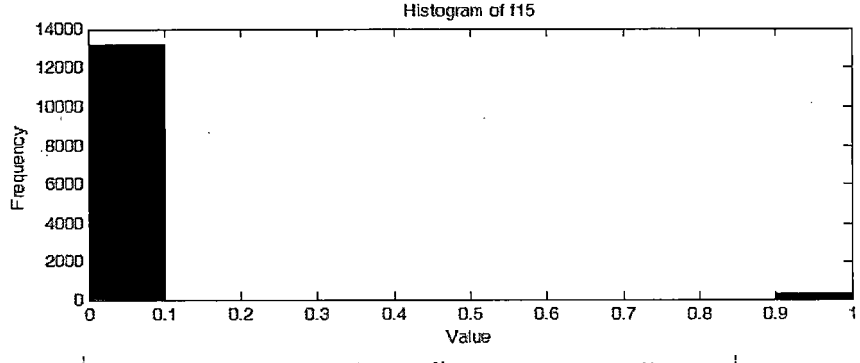
รูปที่ 4-13 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 12 (f12)



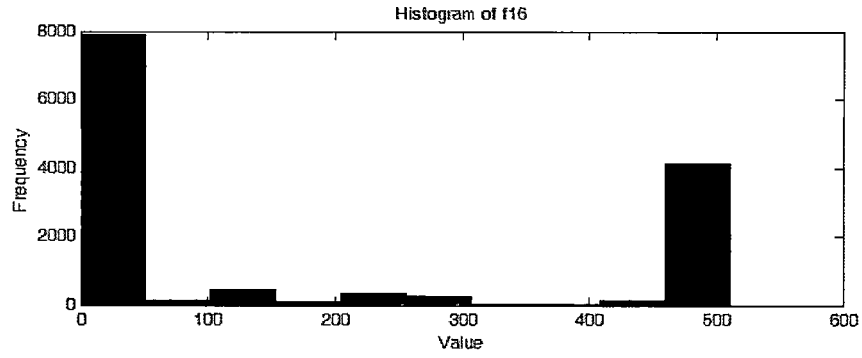
รูปที่ 4-14 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 13 (f13)



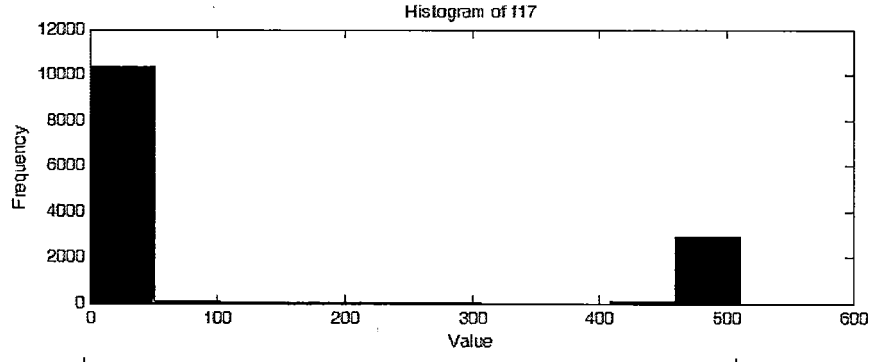
รูปที่ 4-15 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 14 (f14)



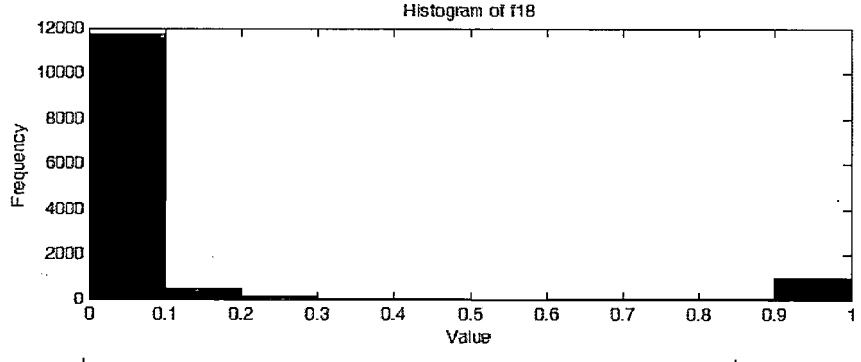
รูปที่ 4-16 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 15 (f15)



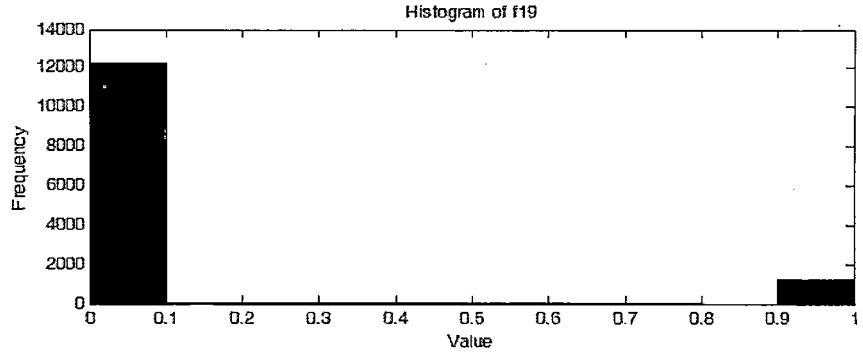
รูปที่ 4-17 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 16 (f16)



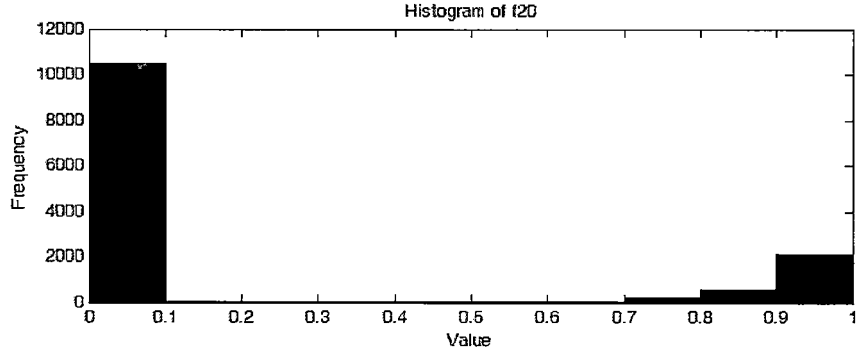
รูปที่ 4-18 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 17 (f17)



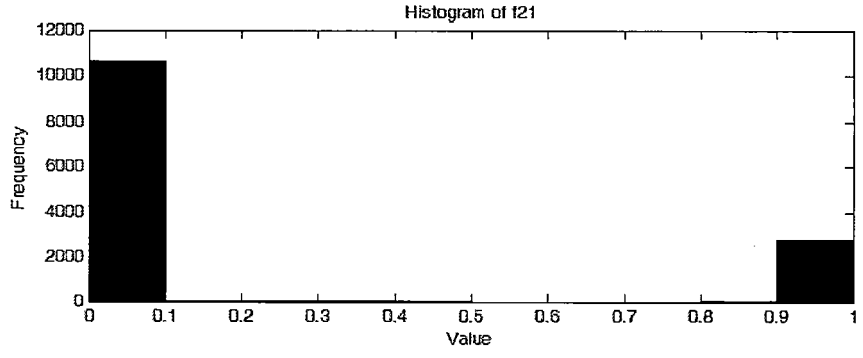
รูปที่ 4-19 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 18 (f18)



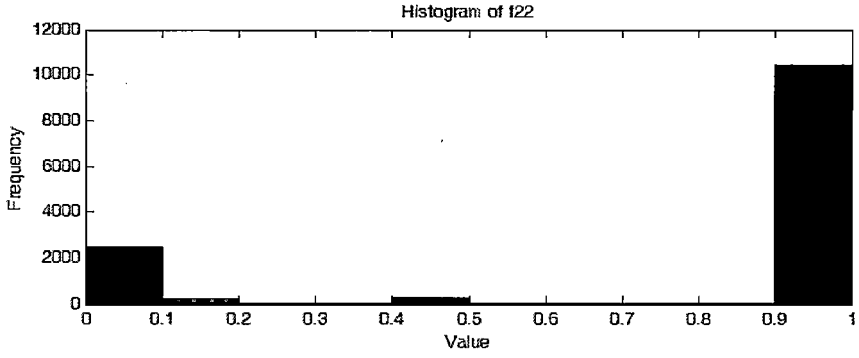
รูปที่ 4-20 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 19 (f19)



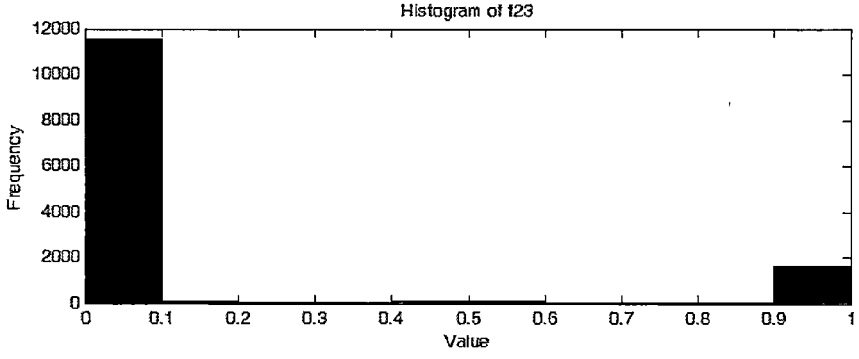
รูปที่ 4-21 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 20 (f20)



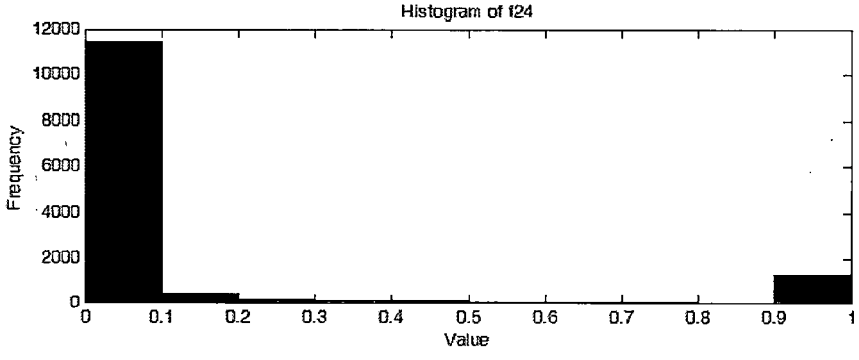
รูปที่ 4-22 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 21 (f21)



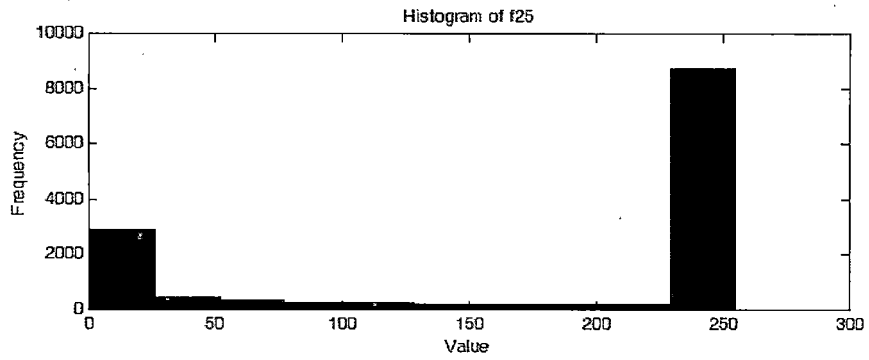
รูปที่ 4-23 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 22 (f22)



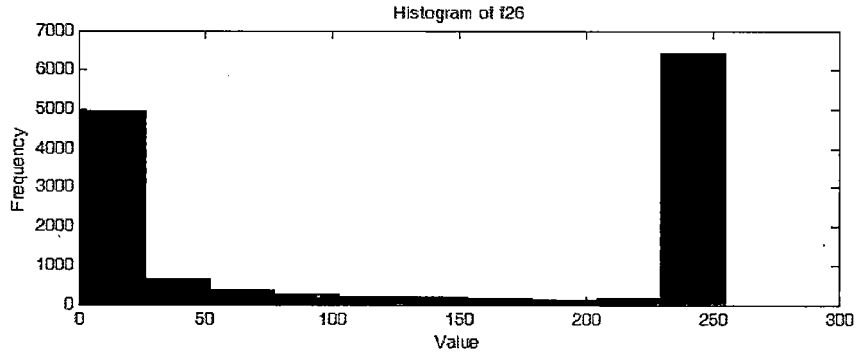
รูปที่ 4-24 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 23 (f23)



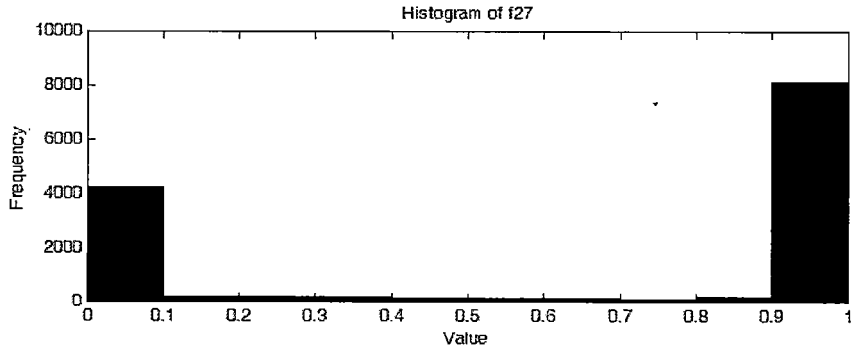
รูปที่ 4-25 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 24 (f24)



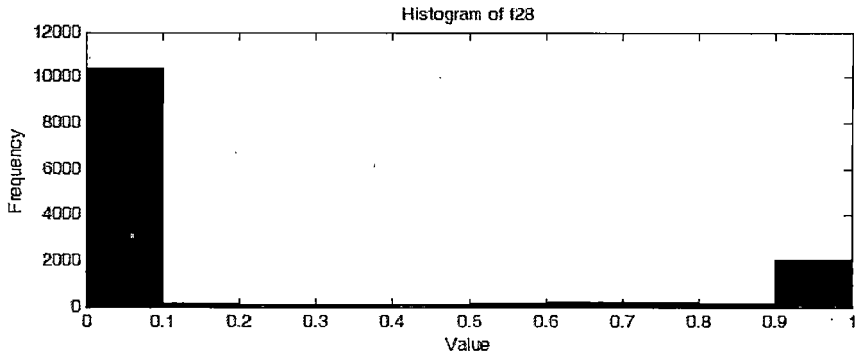
รูปที่ 4-26 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 25 (f25)



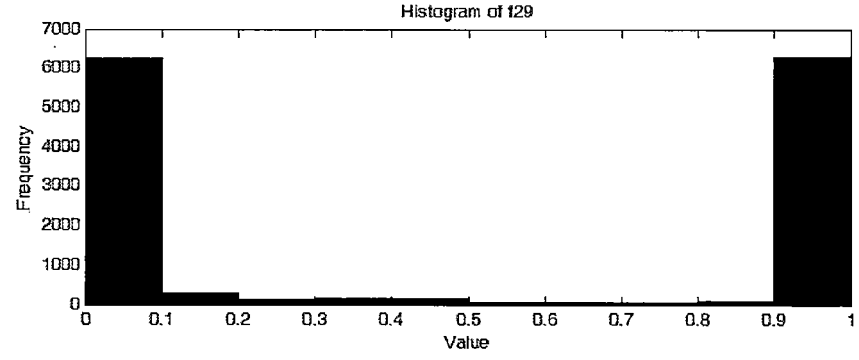
รูปที่ 4-27 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 26 (f26)



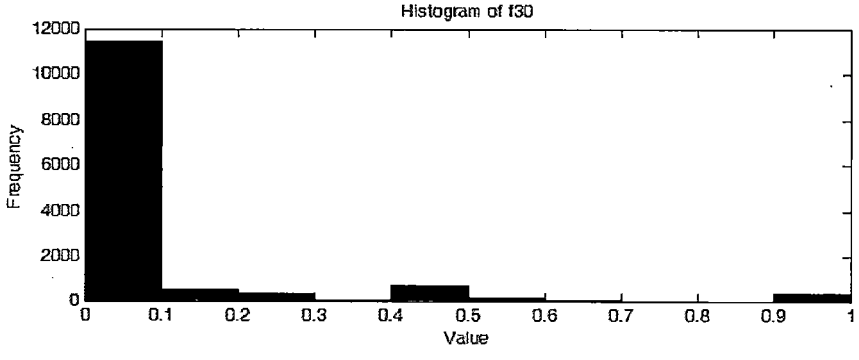
รูปที่ 4-28 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 27 (f27)



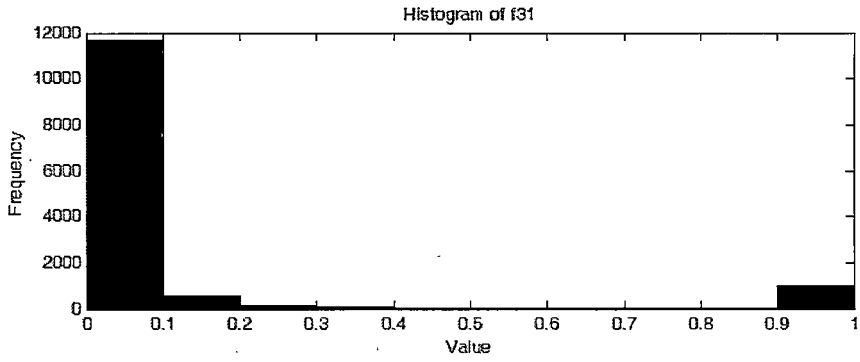
รูปที่ 4-29 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 28 (f28)



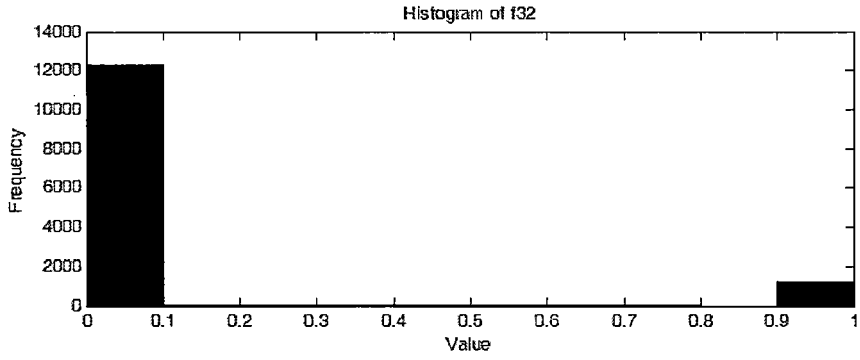
รูปที่ 4-30 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 29 (f29)



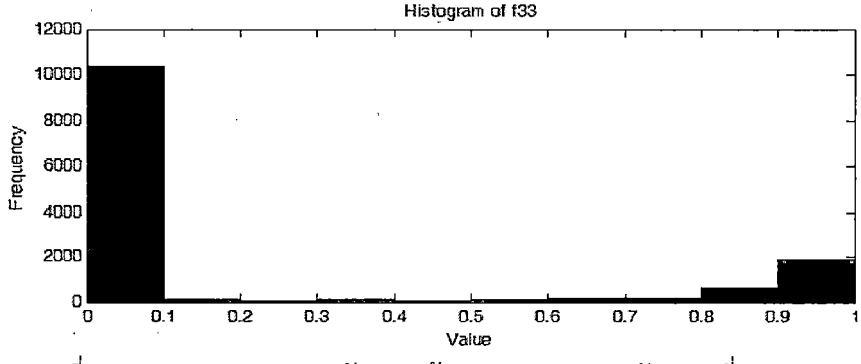
รูปที่ 4-31 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 30 (f30)



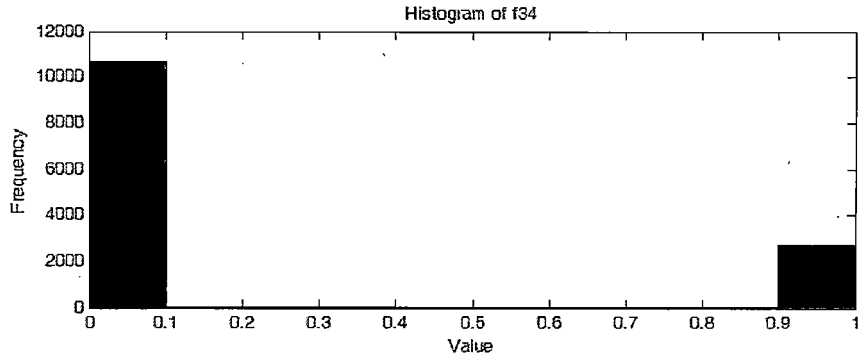
รูปที่ 4-32 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 31 (f31)



รูปที่ 4-33 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 32 (f32)



รูปที่ 4-34 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 33 (f33)



รูปที่ 4-35 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 34 (f34)

4.1.2 ลักษณะข้อมูล KDDcup99 เมื่อสกัดลักษณะด้วย PCA จำนวน 19 ลักษณะ

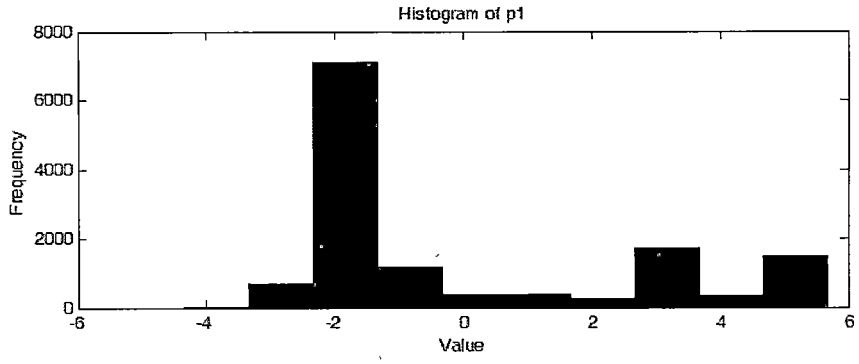
หลังจากที่ได้นำข้อมูลที่ผ่านขั้นตอนการเตรียมข้อมูล 34 ลักษณะ มาสกัดลักษณะด้วยวิธีวิเคราะห์องค์ประกอบหลัก ซึ่งสกัดออกมาได้จำนวน 19 ลักษณะ โดยแต่ละลักษณะมีค่าสูงสุด ค่าต่ำสุด ค่าเฉลี่ย และค่าส่วนเบี่ยงเบนมาตรฐานตามตารางที่ 4-3 ส่วนในตารางที่ 4-4 จะแสดงถึงความสัมพันธ์ระหว่างแต่ละลักษณะทั้ง 19 ลักษณะ และ รูปที่ 4-36 ถึง 4-54 แสดงให้เห็นถึงการกระจายตัวของข้อมูลในแต่ละลักษณะในลักษณะที่ 1 ถึง ลักษณะที่ 19

ตารางที่ 4-3 ค่าทางสถิติของข้อมูล KDDcup99 เมื่อสกัดลักษณะด้วย PCA จำนวน 19 ลักษณะ

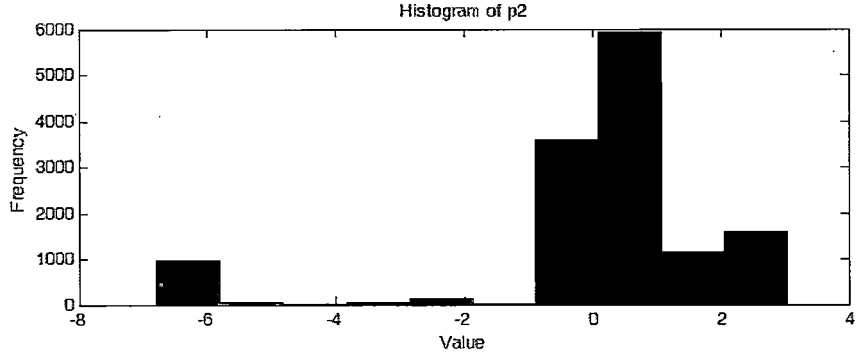
Features	Maximum	Minimum	Mean	Standard Deviation
p1	5.670465	-4.34379	-6.7E-08	2.704464
p2	3.033078	-6.78915	3.57E-08	2.108844
p3	2.677554	-48.5059	-4E-08	1.748789
p4	2.634652	-110.555	-5.1E-08	1.639518
p5	8.209626	-12.9017	-3.9E-08	1.505454
p6	6.763351	-6.99998	-4.6E-08	1.203837
p7	25.28668	-108.391	-5.3E-08	1.064877
p8	48.07232	-46.7354	-7.2E-08	1.033677
p9	69.50105	-21.0231	-8.4E-08	1.029671
p10	17.08883	-25.2092	4.82E-08	1.008519
p11	36.47585	-14.2204	5.3E-08	1.001889
p12	114.8038	-3.39208	6.7E-08	1.000048
p13	37.50314	-54.6732	-3.6E-09	0.955812
p14	6.847732	-22.5184	2.18E-08	0.935022
p15	33.42737	-19.7254	-8.8E-08	0.897046
p16	23.70268	-39.4791	6.08E-08	0.880227
p17	5.327129	-10.4938	5.47E-08	0.791106
p18	12.57152	-3.63403	-7.6E-09	0.721796
p19	31.14863	-20.4477	6.67E-09	0.687534

ตารางที่ 4-4 ค่าสหสัมพันธ์ระหว่างแต่ละลักษณะของข้อมูล KDDcup99 เมื่อสกัดลักษณะด้วย PCA จำนวน 19 ลักษณะ

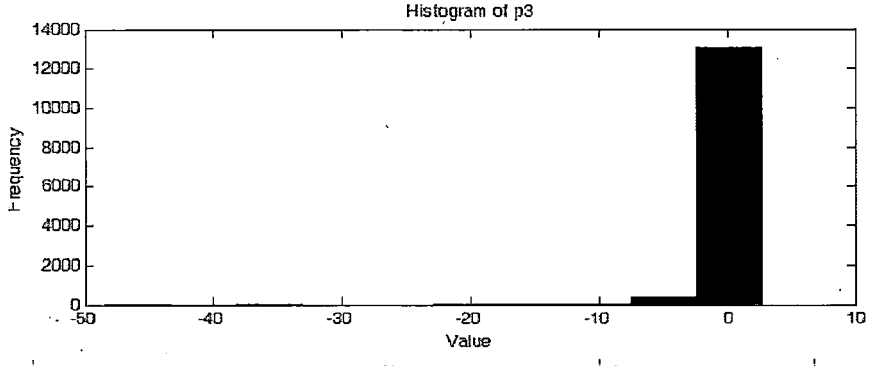
	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12	p13	p14	p15	p16	p17	p18	p19
p1	1																		
p2	-2.4E-08	1																	
p3	-3.5E-08	2.13E-08	1																
p4	5.11E-08	-4.4E-09	-4.1E-08	1															
p5	1.33E-08	1.61E-09	-4.3E-08	1.3E-08	1														
p6	6.65E-08	5.03E-09	-7.7E-08	4.91E-08	-9.7E-09	1													
p7	5.43E-08	-3.2E-09	-1E-07	4.65E-08	-1.5E-08	7.21E-09	1												
p8	4.21E-08	-9.5E-09	-1.1E-07	4.37E-08	-1.9E-08	2.05E-08	4.05E-09	1											
p9	4.38E-08	-5.7E-09	-1.1E-07	4.54E-08	-6.8E-09	3.12E-08	7.15E-09	-8.5E-09	1										
p10	-5.8E-09	1.64E-08	4.97E-08	-2.7E-08	4.12E-09	-2.8E-08	-1.4E-08	-1.5E-08	5.41E-09	1									
p11	-3.1E-08	1.12E-08	7.01E-08	-2.8E-08	1.87E-08	-1.1E-08	-1.9E-09	-3.2E-09	7.75E-09	6.22E-09	1								
p12	-2.9E-08	1.33E-08	8E-08	-3.3E-08	8.64E-09	-3E-08	4.12E-09	-8.3E-09	6.17E-09	-1.9E-09	1.66E-09	1							
p13	-1.7E-10	3.92E-09	-1.3E-08	4.86E-09	-7.1E-09	-9.5E-09	-6.4E-09	-2.1E-09	1.59E-09	-7.7E-09	5.99E-10	-2.6E-09	1						
p14	7.69E-09	8.62E-10	1.61E-08	-7E-10	7.37E-09	1.75E-08	2.91E-08	-1.3E-08	7.98E-09	2.34E-08	2.5E-09	1.48E-08	-2E-09	1					
p15	6.37E-08	-1.5E-08	-1.5E-07	6.91E-08	-2.2E-08	3.75E-08	2.85E-09	1.27E-08	-7.4E-09	-2.6E-09	-4.5E-09	-1E-09	1.14E-09	8.56E-09	1				
p16	-7.6E-08	3.54E-09	1.35E-07	-6.3E-08	2.57E-08	-2.1E-08	-5.3E-09	2.59E-09	2.27E-09	8.91E-09	-3.8E-09	-1.5E-09	3.52E-09	-4.3E-08	-5.9E-09	1			
p17	-3.3E-08	1.61E-08	8.13E-08	-2.4E-08	3.82E-09	-2.4E-08	8.43E-11	-1.7E-08	1.73E-08	8.12E-09	9.04E-09	2.29E-09	-1.3E-08	3.08E-08	7.75E-09	-5.2E-09	1		
p18	-3.1E-08	-2.2E-08	2.02E-08	-4.9E-09	1.56E-08	3.27E-08	2.12E-08	1.6E-08	2.33E-09	2.08E-08	-1.6E-08	4.72E-09	9.48E-09	-4.6E-08	1.9E-08	-5.2E-09	-1.1E-08	1	
p19	-4.6E-08	-1.9E-08	3.91E-08	-1.7E-08	4.08E-09	-6.9E-09	2.01E-09	1.06E-08	-5.5E-09	4.53E-09	-3.8E-10	2.89E-09	1.14E-08	-2.9E-08	7.2E-11	4.5E-09	5.01E-09	-1E-08	1



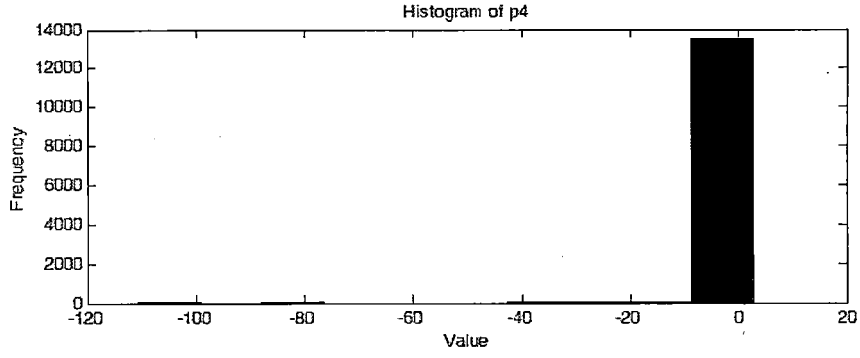
รูปที่ 4-36 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 1 (p1)



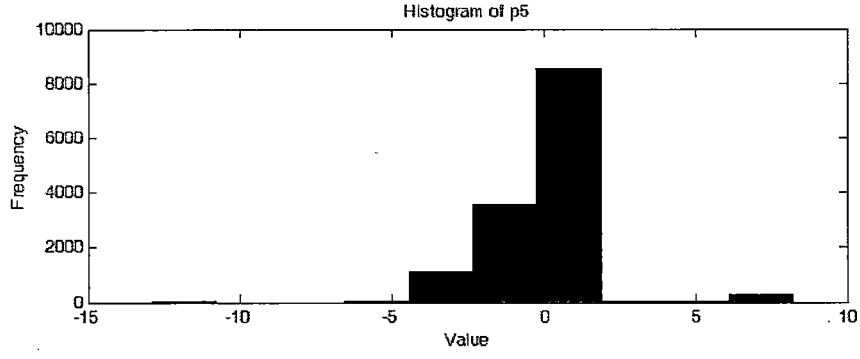
รูปที่ 4-37 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 2 (p2)



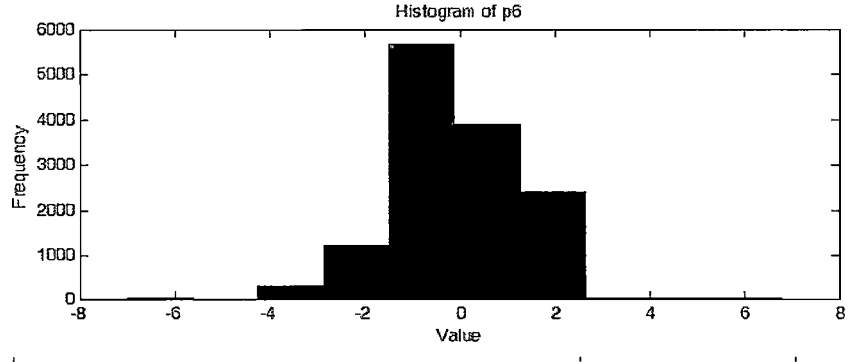
รูปที่ 4-38 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 3 (p3)



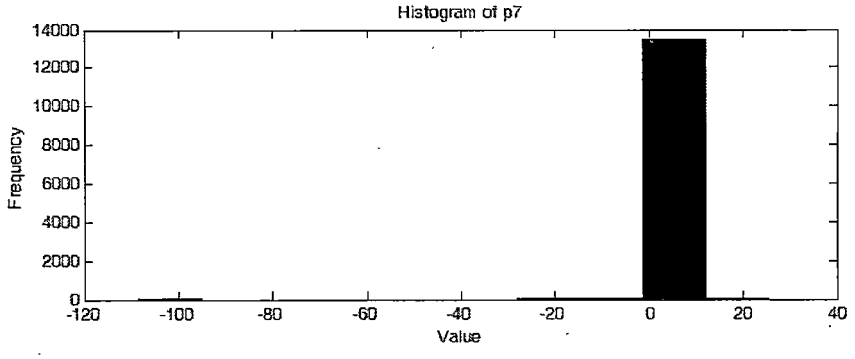
รูปที่ 4-39 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 4 (p4)



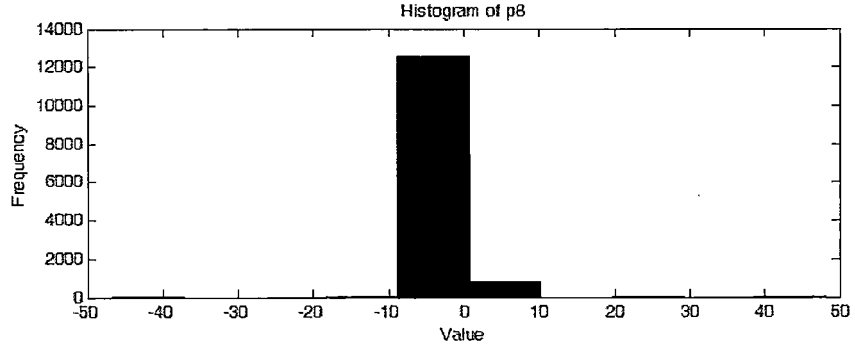
รูปที่ 4-40 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 5 (p5)



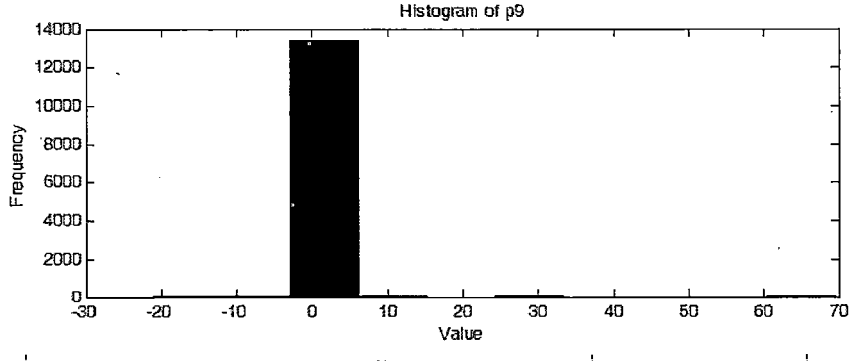
รูปที่ 4-41 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 6 (p6)



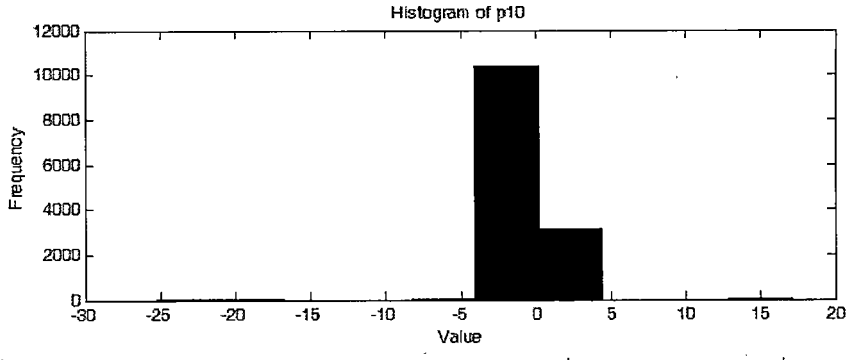
รูปที่ 4-42 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 7 (p7)



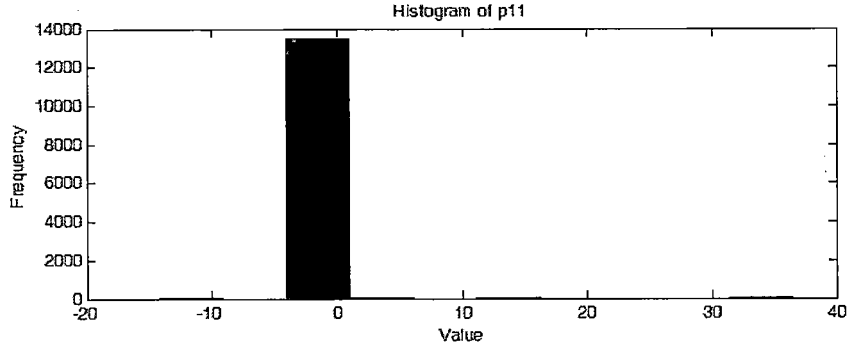
รูปที่ 4-43 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 8 (p8)



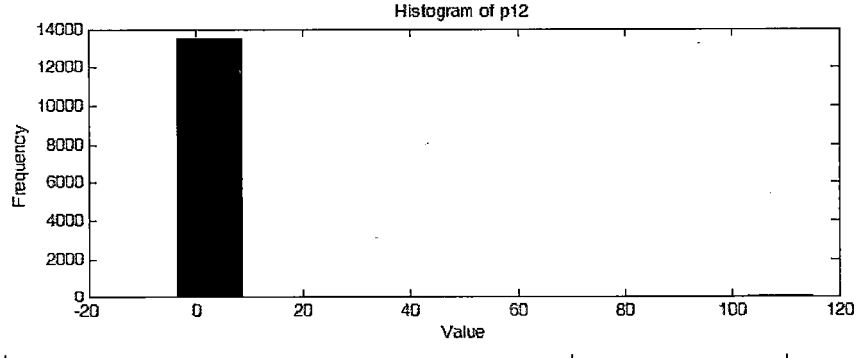
รูปที่ 4-44 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 9 (p9)



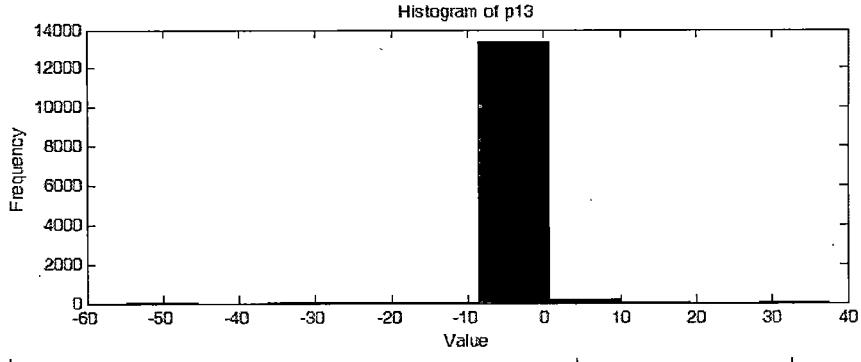
รูปที่ 4-45 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 10 (p10)



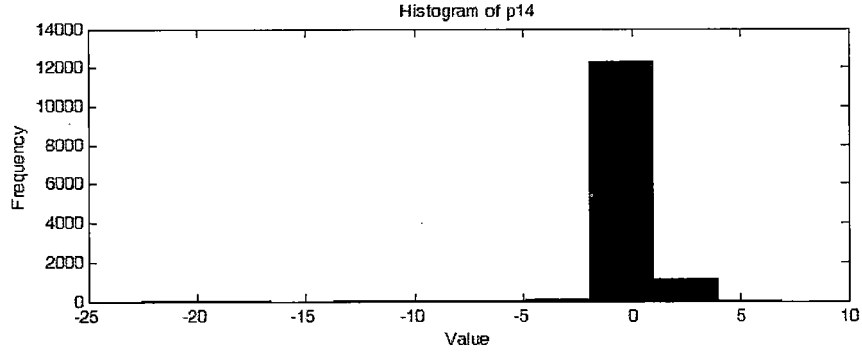
รูปที่ 4-46 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 11 (p11)



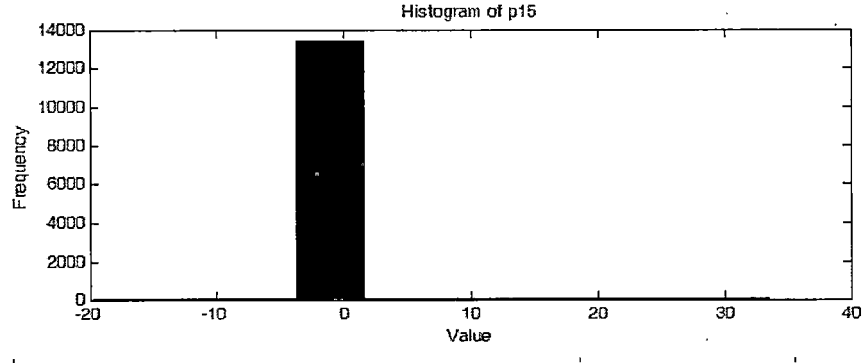
รูปที่ 4-47 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 12 (p12)



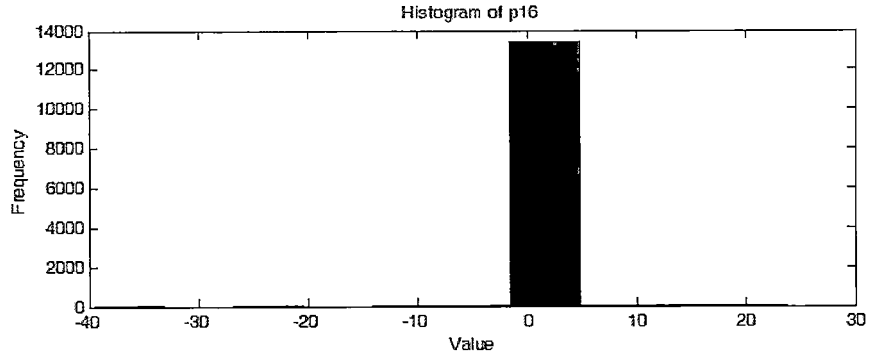
รูปที่ 4-48 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 13 (p13)



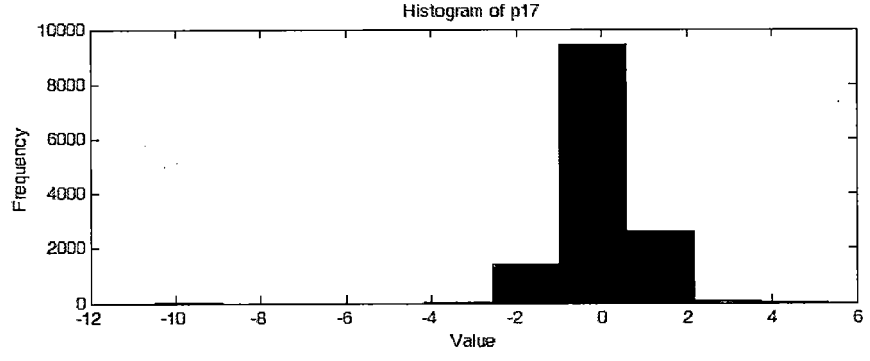
รูปที่ 4-49 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 14 (p14)



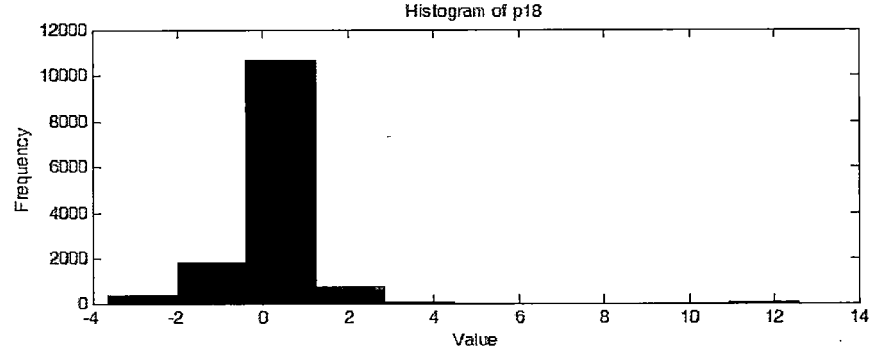
รูปที่ 4-50 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 15 (p15)



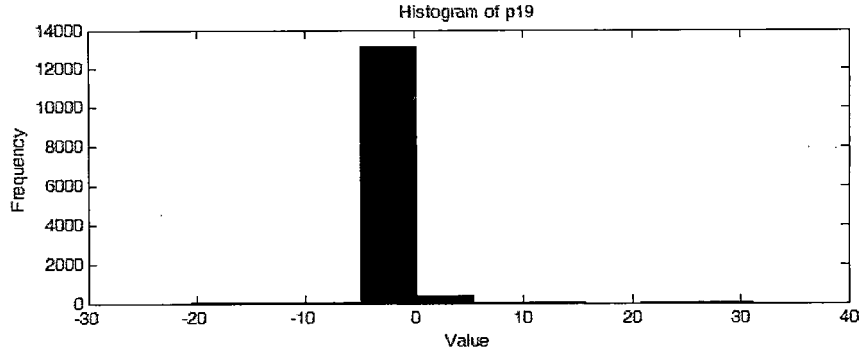
รูปที่ 4-51 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 16 (p16)



รูปที่ 4-52 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 17 (p17)



รูปที่ 4-53 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 18 (p18)



รูปที่ 4-54 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 19 (p19)

4.1.3 ลักษณะข้อมูล KDDcup99 เมื่อเลือกลักษณะด้วย HGIS จำนวน 13 ลักษณะ

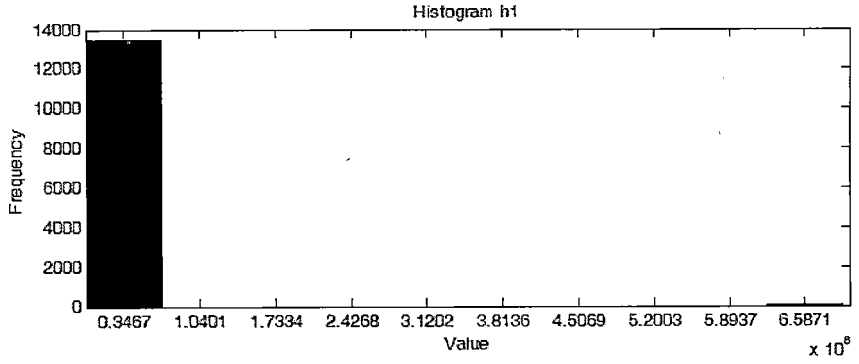
เมื่อนำข้อมูลที่ผ่านขั้นตอนการเตรียมข้อมูล 34 ลักษณะ มาเลือกลักษณะด้วยวิธีฮิวริสติกกรีดดี ซึ่งเลือกลักษณะออกมาได้จำนวน 13 ลักษณะได้แก่ f1, f2, f7, f9, f15, f16, f17, f19, f20, f27, f28, f29 และ f34 โดยแต่ละลักษณะมีค่าสูงสุด ค่าต่ำสุด ค่าเฉลี่ย และค่าส่วนเบี่ยงเบนมาตรฐานตามตารางที่ 4-5 ความสัมพันธ์ระหว่างแต่ละลักษณะทั้ง 13 ลักษณะแสดงในตารางที่ 4-6 โดยแต่ละลักษณะในลักษณะที่ 1 ถึงลักษณะที่ 13 มีการกระจายตัวของข้อมูลดังรูปที่ 4-55 ถึง 4-67

ตารางที่ 4-5 ค่าทางสถิติของข้อมูล KDDcup99 เมื่อเลือกลักษณะด้วย HGIS จำนวน 13 ลักษณะ

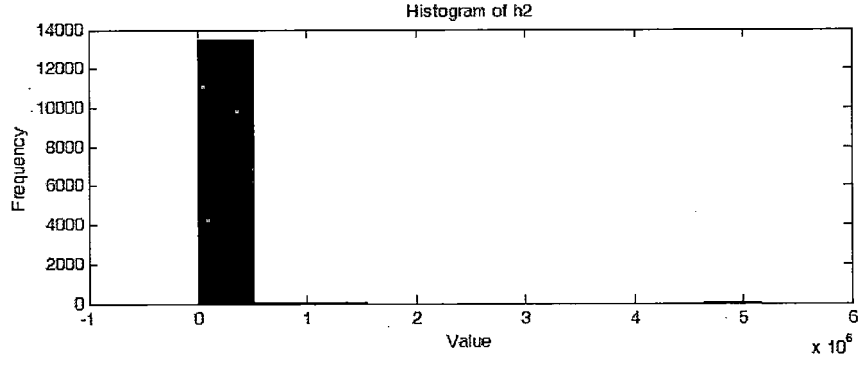
Feature	Maximum	Minimum	Mean	Standard Deviation
h1	6.93E+08	0	74547.73	5977289
h2	5155468	0	7012.817	172422.5
h3	1	0	0.301207	0.4588
h4	1	0	0.002371	0.048632
h5	1	0	0.024446	0.154436
h6	511	0	182.1567	229.4395
h7	511	0	118.6734	207.4745
h8	1	0	0.090148	0.284962
h9	1	0	0.207335	0.391144
h10	1	0	0.64521	0.455144
h11	1	0	0.203488	0.371824
h12	1	0	0.497511	0.481672
h13	1	0	0.205992	0.401915

ตารางที่ 4-6 ค่าสหสัมพันธ์ระหว่างแต่ละลักษณะของข้อมูล KDDcup99 เมื่อเลือกลักษณะด้วย HGIS จำนวน 13 ลักษณะ

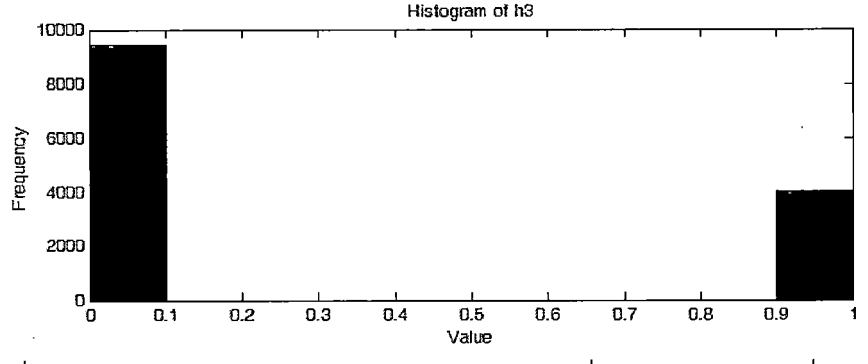
	h1	h2	h3	h4	h5	h6	h7	h8	h9	h10	h11	h12	h13
h1	1												
h2	-0.00051	1											
h3	0.000182	-0.01356	1										
h4	-0.0006	-0.00046	0.074247	1									
h5	-0.00194	-0.0043	0.241114	-0.00772	1								
h6	-0.00767	-0.03196	-0.50251	-0.03844	-0.12499	1							
h7	-0.0069	-0.02283	-0.34901	-0.0276	-0.08978	0.748042	1						
h8	0.016516	-0.01186	-0.20137	-0.01542	-0.05008	0.041531	-0.16715	1					
h9	-0.00192	-0.02154	-0.34329	-0.02389	-0.08207	0.112904	-0.29873	-0.06654	1				
h10	-0.0099	0.03004	0.315514	0.021164	-0.12926	-0.02116	0.431606	-0.41433	-0.62938	1			
h11	-0.0045	-0.02209	-0.33084	-0.02602	-0.07428	0.173425	-0.309	0.062622	0.770066	-0.7527	1		
h12	-0.00152	0.030534	-0.3764	-0.00114	-0.16019	0.210185	0.554466	-0.26051	-0.16194	0.325947	-0.11822	1	
h13	0.000671	-0.02074	-0.32686	-0.02188	-0.07998	0.111673	-0.28867	-0.15995	0.955675	-0.6116	0.749228	-0.15765	1



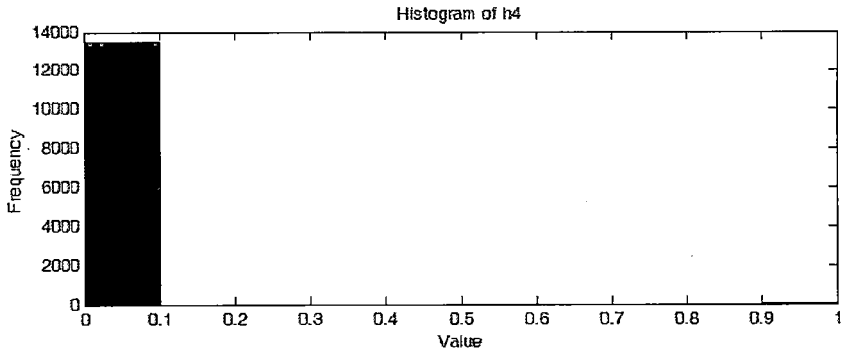
รูปที่ 4-55 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน HGIS ลักษณะที่ 1 (h1)



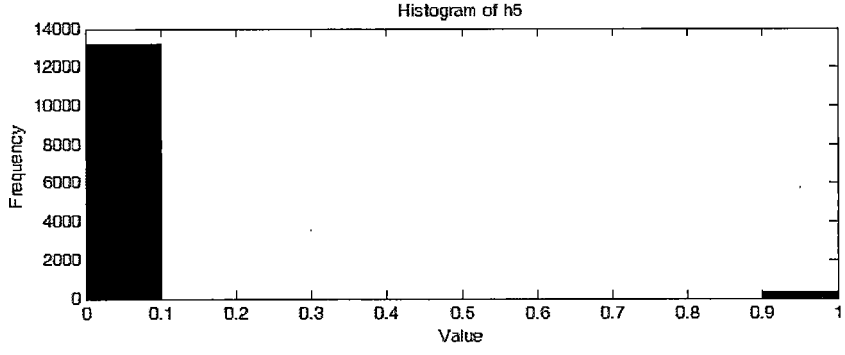
รูปที่ 4-56 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน HGIS ลักษณะที่ 2 (h2)



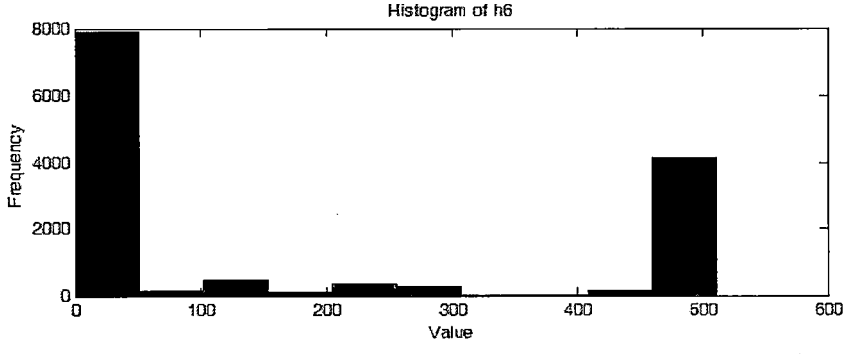
รูปที่ 4-57 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน HGIS ลักษณะที่ 3 (h3)



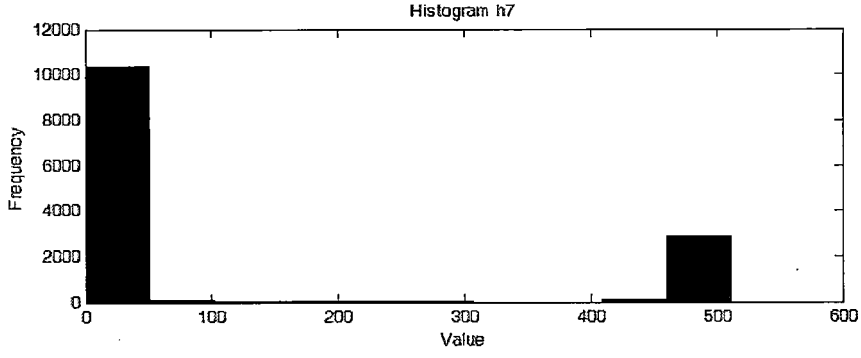
รูปที่ 4-58 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน HGIS ลักษณะที่ 4 (h4)



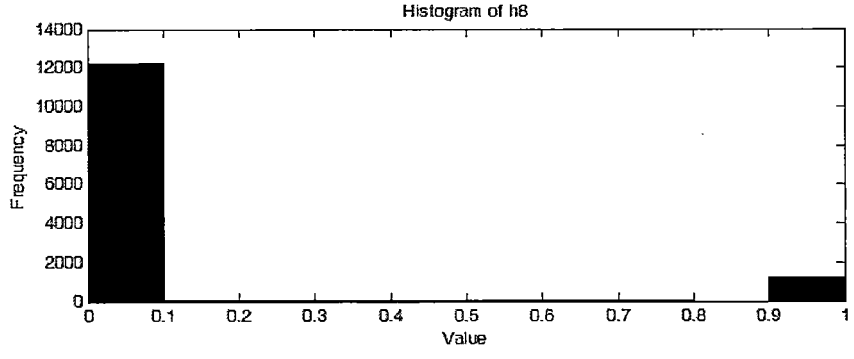
รูปที่ 4-59 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน HGIS ลักษณะที่ 5 (h5)



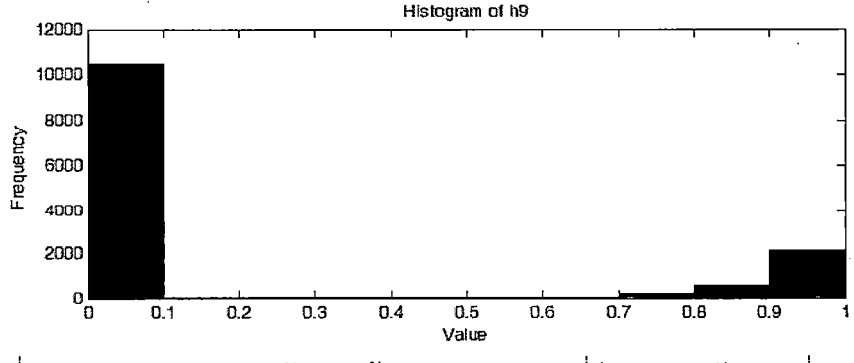
รูปที่ 4-60 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน HGIS ลักษณะที่ 6 (h6)



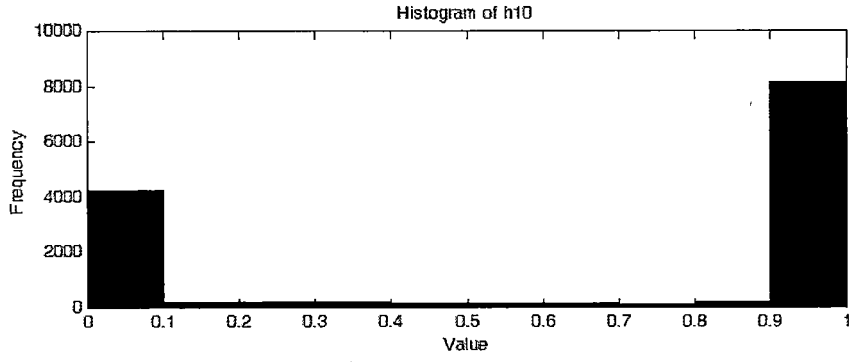
รูปที่ 4-61 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน HGIS ลักษณะที่ 7 (h7)



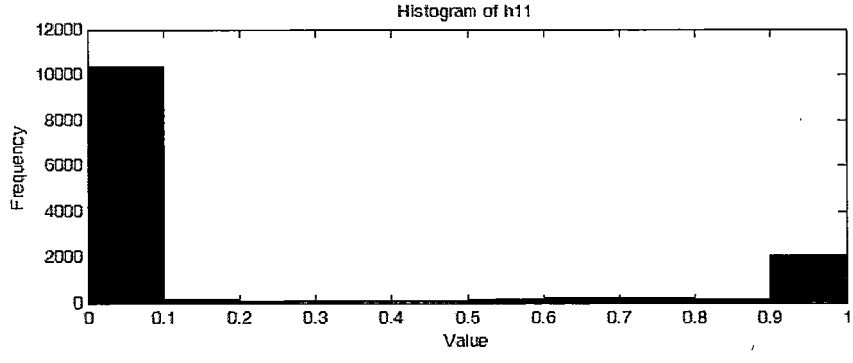
รูปที่ 4-62 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน HGIS ลักษณะที่ 8 (h8)



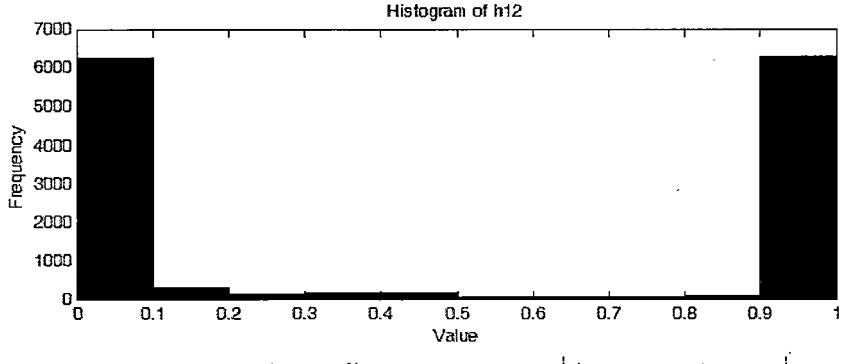
รูปที่ 4-63 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน HGIS ลักษณะที่ 9 (h9)



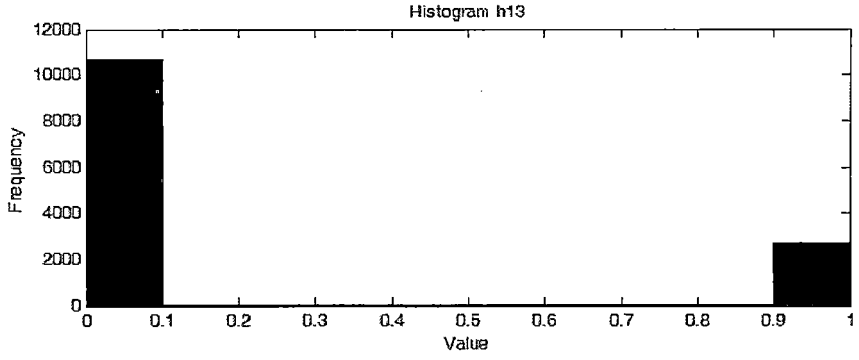
รูปที่ 4-64 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน HGIS ลักษณะที่ 10 (h10)



รูปที่ 4-65 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน HGIS ลักษณะที่ 11 (h11)



รูปที่ 4-66 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน HGIS ลักษณะที่ 12 (h12)



รูปที่ 4-67 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน HGIS ลักษณะที่ 13 (h13)

4.2 การรู้จำประเภทของผู้บุกรุก

ในการทดลองเบื้องต้นสำหรับการรู้จำประเภทของผู้บุกรุกในงานวิจัยนี้ ผู้วิจัยเลือกวิธีการรู้จำแบบมีผู้สอน (Supervised Learning) ที่ได้รับความนิยมในการใช้ทดสอบการรู้จำ คือ โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ ซัพพอร์ตเวกเตอร์แมชชีน และ โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน โดยมีข้อมูลนำเข้าสำหรับการรู้จำ 3 ประเภท คือ ข้อมูลทั้งหมด 34 ลักษณะข้อมูลที่ผ่านขั้นตอนการสกัดลักษณะข้อมูล (PCA) 19 ลักษณะและข้อมูลที่ผ่านขั้นตอนการเลือกลักษณะข้อมูล (HGIS) 13 ลักษณะ ทำให้เราสามารถแบ่งการทดลองออกเป็น 9 การทดลอง ดังนี้

1. **All+BPNN** (ข้อมูลทั้งหมด 34 ลักษณะ และรู้จำด้วยโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ)
   Number of hiddenLayers = (attribs + classes) / 2
   LearningRate=0.3
   Momentum=0.2
   TrainingTime=500
   Training 100%
2. **All+SVM** (ข้อมูลทั้งหมด 34 ลักษณะ และรู้จำด้วยซัพพอร์ตเวกเตอร์แมชชีน)
   The polynomial kernel
3. **All+RBF** (ข้อมูลทั้งหมด 34 ลักษณะ และรู้จำด้วย โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน)
   Gaussian function
4. **PCA+BPNN** (ข้อมูลผ่าน PCA 19 ลักษณะ และรู้จำด้วยโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ)
   Number of hiddenLayers = (attribs + classes) / 2
   LearningRate=0.3
   Momentum=0.2
   TrainingTime=500
   Training 100%
5. **PCA+SVM** (ข้อมูลผ่าน PCA 19 ลักษณะ และรู้จำด้วยซัพพอร์ตเวกเตอร์แมชชีน)
   The polynomial kernel

6. PCA+RBF (ข้อมูลผ่าน PCA 19 ลักษณะ และรู้จำด้วยโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน)
   Gaussian function
7. HGIS+BPNN (ข้อมูลผ่าน HGIS 13 ลักษณะ และรู้จำด้วยโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ)
   Number of hiddenLayers = (attribs + classes) / 2
   LearningRate=0.3
   Momentum=0.2
   TrainingTime=500
   Training 100%
8. HGIS+SVM (ข้อมูลผ่าน HGIS 13 ลักษณะ และรู้จำด้วยซัพพอร์ตเวกเตอร์แมชชีน)
   The polynomial kernel
9. HGIS+RBF (ข้อมูลทั้งหมด HGIS 13 และรู้จำด้วยโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน)
   Gaussian function

ตารางที่ 4-7 นำเสนอรายละเอียดของข้อมูลที่ใช้ในการทดลองนี้ โดยข้อมูลที่ใช้นี้ ชื่อ KDDcup99 ทำการสุ่มเลือกมาทั้งหมด 13,499 ชุดทดสอบ โดยข้อมูลมี 34 ลักษณะ โดยรายละเอียดของข้อมูลแต่ละคลาสเป็นดังนี้

คลาสที่ 1 (ประเภทข้อมูล Normal) จำนวนข้อมูลที่สุ่มมาได้ 4,107 ชุดข้อมูล
คลาสที่ 2 (ประเภทผู้บุกรุก DoS) จำนวนข้อมูลที่สุ่มมาได้ 4,107 ชุดข้อมูล
คลาสที่ 3 (ประเภทผู้บุกรุก Probe) จำนวนข้อมูลที่สุ่มมาได้ 4,107 ชุดข้อมูล
คลาสที่ 4 (ประเภทผู้บุกรุก R2L) จำนวนข้อมูลทั้งหมด 1,126 ชุดข้อมูล
คลาสที่ 5 (ประเภทผู้บุกรุก U2L) จำนวนข้อมูลทั้งหมด 52 ชุดข้อมูล

ตารางที่ 4-7 รายละเอียดข้อมูลที่ใช้ในการทดลอง

ประเภทข้อมูล	จำนวนข้อมูล/ลักษณะ	จำนวนข้อมูล (Patterns) ในแต่ละคลาส
KDDcup99	13499/34	4107/4107/4107/1126/52

ตารางที่ 4-8 นำเสนอค่าสถิติที่ได้จากการทำการทดลองประกอบด้วยค่าร้อยละของความถูกต้องอัตราการตรวจจับผู้บุกรุก และค่าความผิดพลาดเชิงบวก พบว่าในการทดลองวิธีการเลือกลักษณะด้วยขั้นตอนวิธีฮิวริสติกกรีดดีของไอเท็มเซตส่วนใหญ่ให้ค่าร้อยละของความถูกต้องดีกว่าการสกัดลักษณะด้วยวิธีวิเคราะห์องค์ประกอบหลัก และให้ผลที่ใกล้เคียงกับข้อมูลทั้งหมด

ตารางที่ 4-8 ค่า Accuracy จากการประมวลผล

Learning Method	ข้อมูลทั้งหมด (34)			ข้อมูลที่ผ่านขั้นตอน PCA (19)			ข้อมูลที่ผ่านขั้นตอน HGIS (13)		
	Accuracy	Detection Rate	False Alarm rate	Accuracy	Detection Rate	False Alarm rate	Accuracy	Detection Rate	False Alarm rate
BPNN	98.7406	0.9948	0.0119	97.4739	0.9901	0.0224	97.3405	0.9827	0.0397
SVM	96.9479	0.9873	0.0285	94.081	0.9561	0.1018	94.3181	0.9567	0.1008
RBF	91.01	0.9521	0.1093	90.47	0.9656	0.0738	95.53	0.9749	0.0572

เมื่อเปรียบเทียบค่าความถูกต้องของแต่ละคลาสโดยเฉลี่ยด้วยวิธีวัดค่าเฉลี่ยเรขาคณิตดังตารางที่ 4-9 แสดงให้เห็นว่าวิธีการเลือกลักษณะด้วยขั้นตอนวิธีฮิวริสติกกรีดดีของไอเท็มเซตมีค่าเฉลี่ยเรขาคณิตใกล้เคียงกับข้อมูลทั้งหมด และส่วนใหญ่ดีกว่าการสกัดลักษณะด้วยวิธีวิเคราะห์องค์ประกอบหลัก

ตารางที่ 4-9 ค่า G-means จากการประมวลผล

Learning Method	G-Means		
	ข้อมูลทั้งหมด (34)	PCA (19)	Heuristic Greedy Algorithm (13)
BPNN	0.8652	0.8215	0.8104
SVM	0.8458	0.8058	0.8234
RBF	0.7626	0.7015	0.7445

จากการทดลองสรุปเวลาที่ใช้ในการประมวลผลแสดงดังตารางที่ 4-10 ซึ่งเห็นได้ว่าการเลือกลักษณะด้วยขั้นตอนวิธีฮิวริสติกกรีดดีใช้เวลาในการประมวลผลน้อยที่สุด

ตารางที่ 4-10 เวลาที่ใช้ในการประมวลผล

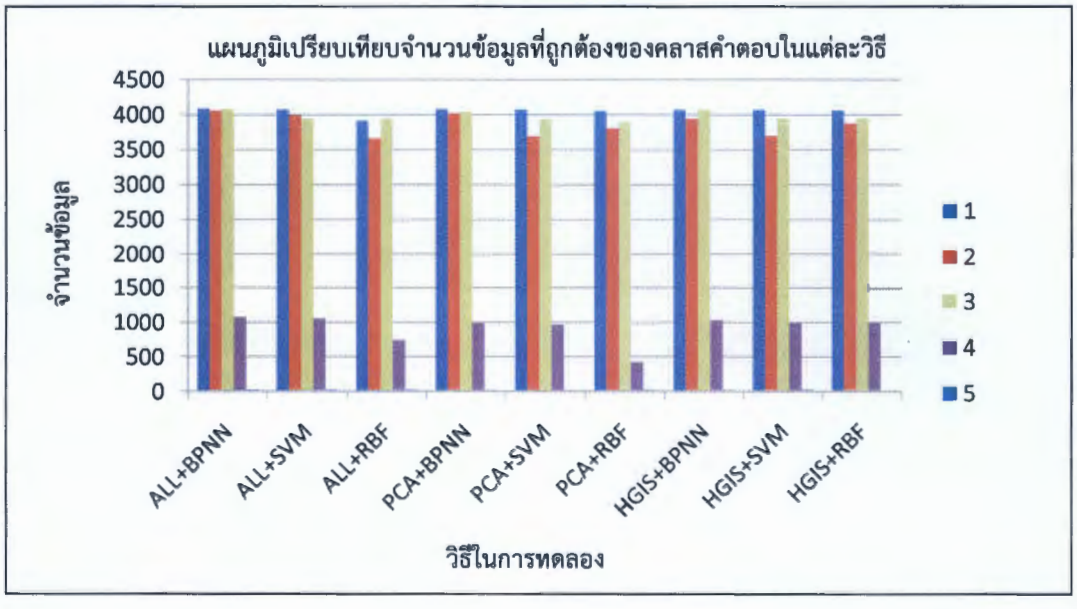
Learning Method	Processing Time(s)		
	ข้อมูลทั้งหมด(34)	PCA(19)	Heuristic Greedy Algorithm(13)
BPNN	168.2811	80.6536	52.0124
SVM	43.2272	25.4783	18.7926
RBF	49.1635	30.6124	22.1979

ตารางที่ 4-11 และรูปที่ 4-68 ได้แสดงจำนวนชุดข้อมูลที่วิธีการเรียนรู้แต่ละวิธีทำการแบ่งประเภทได้อย่างถูกต้อง ซึ่งพบว่าในการทดลองนี้ โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับให้ผลการทดลองกับข้อมูลทั้งหมดที่ดีกว่า และโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐานให้ผลการทดลองกับ

ข้อมูลที่ผ่านการเลือกลักษณะด้วยฮิวริสติกกรีดดีดีกว่ากับข้อมูลทั้งหมดและข้อที่มูลที่ผ่านวิธีวิเคราะห์องค์ประกอบหลัก

ตารางที่ 4-11 จำนวนข้อมูลที่แบ่งประเภทได้ถูกต้องของคลาสคำตอบในแต่ละวิธี

Learning Method	คลาส				
	1 (4107)	2 (4107)	3 (4107)	4 (1126)	5 (52)
ALL+BPNN	4088	4058	4078	1078	27
ALL+SVM	4075	3990	3943	1053	26
ALL+RBF	3916	3658	3947	740	25
PCA+BPNN	4079	4015	4045	996	23
PCA+SVM	4075	3689	3944	968	24
PCA+RBF	4058	3804	3899	425	27
HGIS+BPNN	4068	3944	4075	1032	21
HGIS+SVM	4070	3693	3949	994	26
HGIS+RBF	4059	3872	3953	996	15



รูปที่ 4-68 แผนภูมิเปรียบเทียบจำนวนข้อมูลที่ถูกต้องของคลาสคำตอบในแต่ละวิธี

บทที่ 5 สรุปผลการทดลอง

5.1 สรุปผลการทดลอง

ในงานวิจัยนี้เสนอวิธีการแบบผสมสำหรับการหาตัวแทนจากชุดข้อมูลบนเครือข่ายที่เหมาะสม เพื่อระบุผู้บุกรุกแบบเวลาจริงกับชุดข้อมูล KDDcup99 จำนวนประมาณ 4,900,000 จุดข้อมูล 41 ลักษณะ ซึ่งได้นำข้อมูล 10% ของข้อมูลทั้งหมดสุ่มออกมาอีกจำนวน 13,499 จุดข้อมูล เพื่อสะดวกในการทดลอง จากนั้นตัดลักษณะที่ไม่มีผลต่อการทดลองออกไปจึงเหลือลักษณะจำนวน34ลักษณะ และหาตัวแทนของข้อมูลที่เหมาะสม เพื่อลดความซ้ำซ้อนของข้อมูลและเพิ่มประสิทธิภาพในการระบุผู้บุกรุก ในงานวิจัยนี้จะใช้วิธีการหาตัวแทนข้อมูลด้วยการสกัดลักษณะด้วยวิธีการวิเคราะห์องค์ประกอบหลัก โดยเปรียบเทียบกับการเลือกลักษณะด้วยวิธีฮิวริสติกกรีดดีของไอเท็มเซตโดยใช้หลักการ Apriori จากนั้นนำลักษณะที่ได้นั้นเข้าสู่กระบวนการรู้จำเพื่อระบุผู้บุกรุก โดยจะทดสอบ 3 วิธีคือโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ ซัพพอร์ตเวกเตอร์แมชชีน และ โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน

จากการทดลองด้วยการสกัดลักษณะโดยวิธีการวิเคราะห์องค์ประกอบหลัก ซึ่งจะใช้หลักความสัมพันธ์เชิงเส้นระหว่างตัวแปร โดยการผสมเชิงเส้นตรงได้องค์ประกอบหลักที่สามารถอธิบายความแปรปรวนของชุดข้อมูลได้มากที่สุดเป็นอันดับหนึ่ง และองค์ประกอบที่สามารถอธิบายความแปรปรวนของชุดข้อมูลได้มากที่สุดอันดับสอง โดยที่ทั้งสองนี้ไม่มีความสัมพันธ์กัน เมื่อนำชุดข้อมูล KDDcup99 จำนวน 34 ลักษณะหาองค์ประกอบหลักแล้วจึงเลือกองค์กระกอบหลักที่ผลรวมค่าไอเกนไม่น้อยกว่า 0.95 ซึ่งเป็นเกณฑ์ที่งานวิจัยส่วนใหญ่นิยมใช้กัน ซึ่งจะได้ 19 องค์ประกอบ และนำไปคูณกับชุดข้อมูลดั้งเดิมทำให้ได้ข้อมูลชุดใหม่ ดังนั้นการหาตัวแทนชุดข้อมูลด้วยวิธีการวิเคราะห์องค์ประกอบหลักสามารถสกัดลักษณะออกมาได้จำนวน 19 ลักษณะ และเมื่อนำข้อมูลชุดใหม่นี้ไปเข้ากระบวนการรู้จำประเภทของผู้บุกรุกทั้ง 3 วิธี แสดงให้เห็นว่าเมื่อลดลักษณะลงด้วยการสกัดลักษณะด้วยวิธีการวิเคราะห์องค์ประกอบหลัก ประสิทธิภาพของความถูกต้องของข้อมูลต่ำลงเล็กน้อยเมื่อเปรียบเทียบกับข้อมูลทั้งหมด เว้นแต่การรู้จำโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐานที่มีอัตราการตรวจจับผู้บุกรุก และค่าความผิดพลาดเชิงบวกเพิ่มขึ้นเล็กน้อยและเวลาที่ใช้ในการตรวจจับผู้รุกจะน้อยกว่าเนื่องจากจำนวนลักษณะข้อมูลมีจำนวนลดลงจากเดิมซึ่งสามารถสรุปได้ว่าการสกัดคุณลักษณะด้วยวิธีการวิเคราะห์องค์ประกอบหลักยังไม่เหมาะสมสำหรับข้อมูลผู้บุกรุกในระบบเครือข่ายเนื่องจากข้อมูลมีลักษณะที่กระจายตัวมาก ทั้งนี้อาจเนื่องจากการกำหนดค่าไอเกนสำหรับการเลือกลดลักษณะข้อมูลและวิธีที่ใช้ในการรู้จำผู้บุกรุกด้วย

การทดลองหาตัวแทนข้อมูลด้วยการเลือกลักษณะโดยวิธีฮิวริสติกกรีดดีของไอเท็มเซตโดยใช้หลักการ Apriori ซึ่งเป็นวิธีการหาลักษณะที่มีความสำคัญในการตรวจจับผู้บุกรุกโดยไม่มีการเปลี่ยนแปลงข้อมูลใดๆ แต่จะเลือกลักษณะบางลักษณะจากลักษณะทั้งหมด 34 ลักษณะ โดยวิธีการเลือกนั้นขั้นตอนวิธีการแก้ปัญหาที่คิดแบบง่าย ๆโดยจะพิจารณาว่าข้อมูลที่มีอยู่ในขณะนั้นมีทางเลือกใดที่ให้คำตอบที่ดีที่สุดสามารถทำการค้นหาคำตอบจากข้อมูลที่มีขนาดใหญ่มาก ๆ ได้ เพราะเป็นการค้นหาคำตอบที่ไม่ต้องดูข้อมูลทุกตัวซึ่งเรียกวิธีนี้ว่าฮิวริสติกโดยจะสร้างไอเท็มเซตหรือเซตของลักษณะที่

เป็นไปได้ และในการทดลองนั้นเราจะนำหลักการ Apriori มาช่วยในการลดจำนวนไอเท็มเซตลง เนื่องจากสร้างไอเท็มเซตจากการใช้เซตที่มีขนาดใหญ่ที่หาได้ในขั้นตอนก่อนหน้าซึ่งจะนำแต่ละเซตของลักษณะมาหาค่า RMSE โดยใช้โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐานในการคัดเลือกเพื่อที่จะสร้างไอเท็มเซตถัดไปการสร้างไอเท็มได้นั้นทุกๆเซตย่อยจะต้องมีค่า RMSE น้อยกว่าเซตที่กำลังสร้างในการทดลองนี้เราได้สร้างไอเท็มเซตทั้งหมดจำนวน 3 candidate-itemsets เป็นการหาเซตลักษณะที่มีค่า RMSE ต่ำๆ จากนั้นทำการสุ่มนำลักษณะอื่นมารวมด้วย โดยสุ่มจากลักษณะที่มีค่า RMSE ต่ำๆ ก็จะมีโอกาสถูกสุ่มมาก จนกระทั่งได้ผลเป็นที่น่าพอใจ ดังนั้นจึงได้ลักษณะที่ผ่านการเลือกลักษณะด้วยวิธีฮิวริสติกกรีดดีของไอเท็มเซตจำนวน 13 ลักษณะ เมื่อนำตัวแทนข้อมูลที่ผ่านการเลือกลักษณะผ่านกระบวนการตรวจจับผู้บุกรุกโดยการรู้จำเปรียบเทียบทั้ง 3 วิธี แสดงให้เห็นว่าเมื่อใช้กระบวนการรู้จำโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐานให้ประสิทธิภาพที่ดีขึ้นทั้งในด้านของค่าร้อยละของความถูกต้อง อัตราการตรวจจับผู้บุกรุก และค่าความผิดพลาดเชิงบวก เมื่อเปรียบเทียบกับข้อมูลดั้งเดิมทั้งหมด และข้อมูลที่ผ่านการสกัดลักษณะด้วยวิธีการวิเคราะห์องค์ประกอบหลักเนื่องจากเราใช้โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐานในการคัดเลือกเซตลักษณะที่มีค่า RMSE ต่ำๆและมีอัตราความเร็วในการตรวจจับการบุกรุกที่ดีกว่าเพราะมีจำนวนลักษณะที่น้อยกว่า

สรุปผลการทดลองการหาตัวแทนข้อมูลที่เหมาะสมสำหรับการตรวจจับผู้บุกรุกในเครือข่ายด้วยการสกัดลักษณะได้ 19 ลักษณะ เปรียบเทียบกับการเลือกลักษณะข้อมูลได้ 13 ลักษณะ และทำการรู้จำประเภทของผู้บุกรุกโดยใช้โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ ซัพพอร์ตเวกเตอร์แมชชีน และ โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐานผลที่ได้จากการทดลองแสดงให้เห็นว่าส่วนใหญ่เมื่อลดข้อมูลลงแล้วทำให้ผลการทดลองในส่วนการประมวลผลเพื่อการรู้จำในขั้นตอนที่สองมีประสิทธิภาพที่ต่ำลง ซึ่งในส่วนของการสกัดลักษณะด้วยวิธีวิเคราะห์องค์ประกอบหลักนั้นอาจเกิดจากการกำหนดค่าไอเกนสำหรับการเลือกลักษณะที่ยังไม่เหมาะสม แต่ในส่วนของการเลือกลักษณะด้วยวิธีฮิวริสติกกรีดดีของไอเท็มเซตนั้นให้ผลที่ดีกว่าทั้งหมดด้วยการรู้จำแบบโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน เนื่องจากขั้นตอนในการเลือกลักษณะได้ใช้การรู้จำแบบโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐานในการเลือกลักษณะที่เหมาะสมได้จำนวนลักษณะที่น้อยกว่าการสกัดลักษณะของข้อมูลด้วยวิธีการวิเคราะห์องค์ประกอบหลัก และให้ผลที่ใกล้เคียงกับข้อมูลทั้งหมดและข้อมูลที่ผ่านการสกัดลักษณะและมีอัตราความเร็วในการตรวจจับการบุกรุกน้อยที่สุดด้วย

5.2 ปัญหาและข้อเสนอแนะ

ในขั้นตอนของการเลือกลักษณะด้วยวิธีฮิวริสติกกรีดดีของไอเท็มเซตจะใช้เวลานานมาก เนื่องจากต้องนำเซตลักษณะที่สร้างได้ผ่านกระบวนการรู้จำหา RMSE เพื่อให้ได้เซตของลักษณะที่เหมาะสม และการสกัดลักษณะด้วยวิธีการวิเคราะห์องค์ประกอบหลักควรศึกษาในเรื่องของการหาค่าไอเกนที่เหมาะสมในการเลือกองค์ประกอบหลักสำหรับชุดข้อมูลผู้บุกรุก

5.3 งานที่จะทำต่อไปในอนาคต

1. ทำขั้นตอนวิธีการเลือกลักษณะด้วยวิธีฮิวริสติกกรีดดีของไอเท็มเซตจนครบกระบวนการ ซึ่งจะไม่สามารถสร้างไอเท็มเซตได้อีก

2. ศึกษาวิธีการหรือเกณฑ์ที่ใช้ในการคัดเลือกลักษณะที่เหมาะสมในส่วนของการเลือกลักษณะด้วยวิธีฮิวริสติกกรีดดีของไอเท็มเซต

3. ศึกษาค่าไอเกนที่เหมาะสมในการเลือกองค์ประกอบหลักสำหรับชุดข้อมูลผู้บุกรุกในการสกัดลักษณะข้อมูลด้วยวิธีการวิเคราะห์องค์ประกอบหลัก

4. ศึกษาการรู้จำประเภทผู้บุกรุกที่มีประสิทธิภาพและเหมาะสมกับชุดข้อมูลผู้บุกรุก

5. นำเสนอผลงานวิจัยในการประชุมวิชาการนานาชาติวารสารวิจัยระดับนานาชาติ

6. พัฒนาโปรแกรมประยุกต์เพื่อให้ทดลองนำไปทดสอบกับข้อมูลผู้บุกรุกชุดอื่น

บรรณานุกรม

KDD'99 datasets, The UCI KDD Archive, http://kdd.ics.uci.edu/databases/KDDcup99/, Irvine, CA, USA, 1999.

Hai-Hua Goa, Hui-Hua Yang, and Xing-Yu Wang (2005), "Kernel PCA Based Network Intrusion Feature Extraction and Detection Using SVM", *Proceedings of ICNC 2005, LNCS 3611*, pp. 89-94.

Hai-Hua Goa, Hui-Hua Yang, and Xing-Yu Wang (2005), "Principal Component Neural Networks Based Intrusion Feature Extraction and Detection Using SVM", *Proceedings of ICNC 2005, LNCS 3611*, pp. 21-27.

Dong Seong Kim, Ha-Nam Nguyen, T. Thein, and JongSouPark (2005), "An Optimized Intrusion Detection System Using PCA and BNN", *Proceedings of The 6th Asia-Pacific Sym. on Information and Telecommunication Technologies, IEICE Communications Society*, pp. 356-359.

Zhu Xiaorong, Wang Dianchun, Ye Changguo (2009), "A New Feature Extraction Method of Intrusion Detection," *2009 First International Workshop on Education Technology and Computer Science vol. 2*, pp.504-507.

Murat Karabatak, M. Cevdet Ince, "A New Feature Selection Method Based on Association Rules for Diagnosis of Erythemato-squamous Diseases", *Expert Systems with Applications*, Volume 36, pp. 12500–12505, 2009.

Mansour Sheikhan and Zahra Jadidi, "Misuse Detection Using Hybrid of Association Rule Mining and Connectionist Modeling", *World Applied Sciences Journal 7 (Special Issue of Computer & IT): 31-37*, 2009.

Onur Inan, Mustafa Serter Uzer, and Nihat Y.lmaz, "A New Hybrid Feature Selection Method Based on Association Rules and PCA for Detection of Breast Cancer", *International Journal of Innovative Computing, Information and Control*, Volume 9, Number 2, 2013.

Jackson, J. E., *A User's Guide to Principal Components*, John Wiley and Sons, p. 592, 1991.

Scarfone Karen, Mell Peter (February 2007). "Guide to Intrusion Detection and Prevention Systems (IDPS)". *Computer Security Resource Center (National Institute of Standards and Technology) (800–94)*. Retrieved 1 January 2010.

Robert Hecht Nielsen, *Theory of the back propagation neural network in Proceedings 1989 IEEE IJCNN*, pp. 1593–1605, IEEE Press, New York, 1989.

S.Chen, C. F. N. Cowan, P. M. Grant, "Orthogonal Least Squares Learning Algorithm for Radial Basis Function Networks", *IEEE transactions on neural networks*, vol. 2, no.2, 1991.

M. Hearst, "Support Vector Machines," *IEEE Intelligent Systems Magazine*, Trends and Controversies, Marti Hearst, ed., vol 13, no 4, 1998.

Jianwen Xie, Jianhua Wu, Qingquan Qian, "Feature Selection Algorithm Based on Association Rules Mining Method", *International Conference on Computer and Information Science*, 2009.

Hari Om , Aritra Kundu, "A Hybrid System for Reducing the False Alarm Rate of Anomaly Intrusion Detection System", *1st Int'l Conf. on Recent Advances in Information Technology*, 2012.

Iftikhar Ahmad, Azween Abdullah, Abdullah Alghamdi and Muhammad Hussain , "Optimized intrusion detection mechanism using soft computing techniques". *Telecommunication Systems*, 47 . pp. 1-9, 2011.

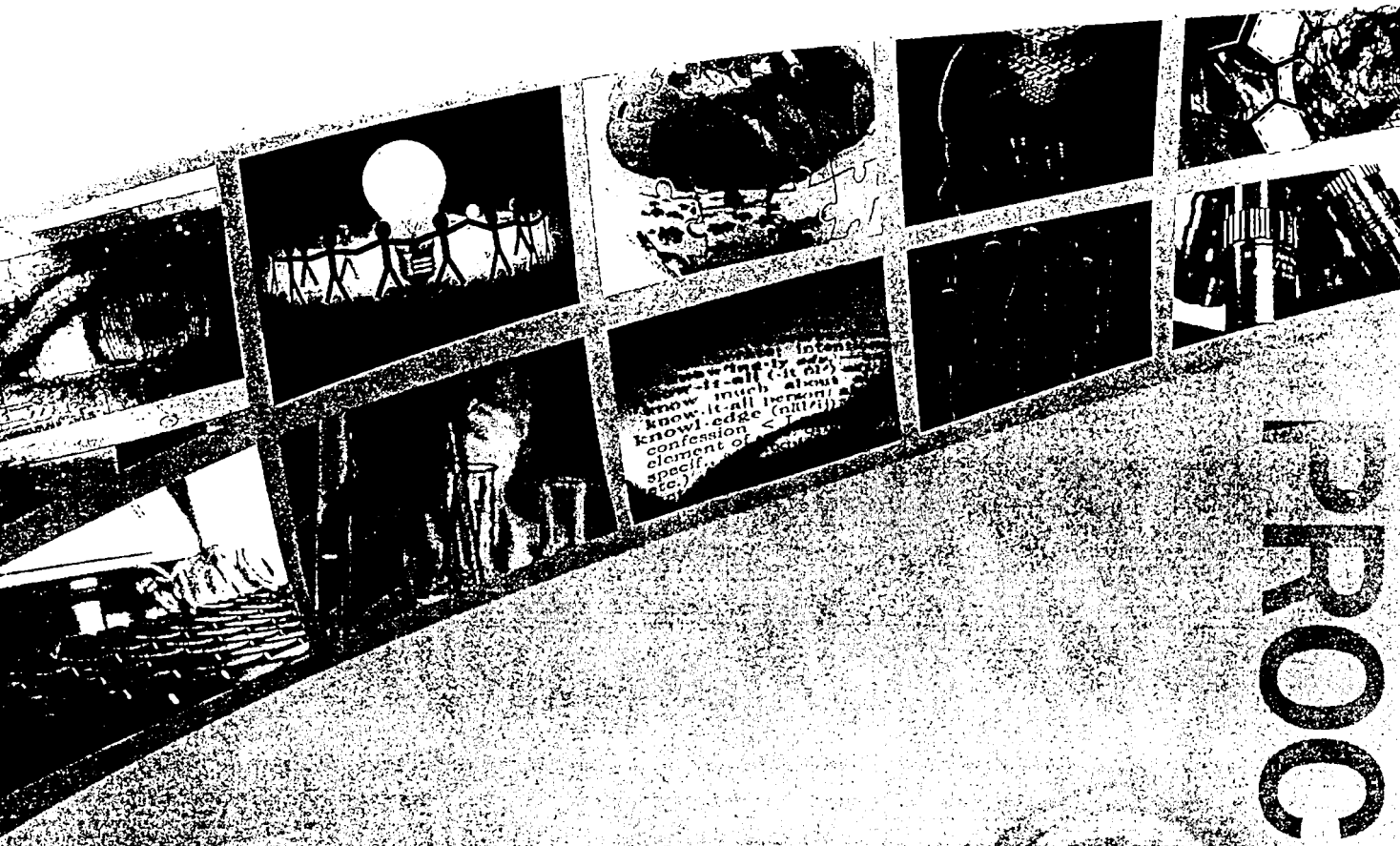
Iftikhar Ahmad, Azween Abdulah, Abdullah Alghamdi, Khaled Alnfajan and Muhammad Hussain , "Feature Subset Selection for Network Intrusion Detection Mechanism Using Genetic Eigen Vectors", *International Conference on Telecommunication Technology and Applications Proc .of CSIT vol.5*, 2011.

Shailendra Singh, Sanjay Silakari and Ravindra Patel, "An efficient feature reduction technique for intrusion detection system", *International Conference on Machine Learning and Computing IPCSIT vol.3*, 2011.

T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein, Introduction to Algorithms (3e), p.360, 2001.

P. N. Tan, M. Steinbach, and V. Kumar. Introduction to Data Mining, Pearson International Edition, ISBN: 0-321-42-52-7, 2006.

ภาคผนวก



เอกสารประกอบการประชุมวิชาการ

Knowledge and Smart Technology

ครั้งที่ ๕ (KST-2556)

๓๑ มกราคม - ๑ กุมภาพันธ์ ๒๕๕๖

สารบัญ

การเลือกลักษณะของข้อมูลผู้บุกรุกด้วย Heuristic Greedy Algorithm of Item Set

Intrusion Feature Selection using Heuristic Greedy Algorithm of Item Set

จรรยา อ้นปันส์[1] อัณณ์นุพันธ์ รอดทุกข์[2] สุวรรณา รัศมีขวัญ[1] เบญจภรณ์ จันทรกองกุล[1] และกฤษณะ ชินสาร[1]
[1]สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา อ.เมือง จ.ชลบุรี 20131
[2]ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยรามคำแหง แขวงหัวหมาก กรุงเทพมหานคร 10240
Email: mai.janya@gmail.com

บทคัดย่อ

บทความนี้นำเสนอวิธีการเลือกและวิธีการสกัดคุณลักษณะเด่นของข้อมูลผู้บุกรุกในเครือข่ายคอมพิวเตอร์ โดยวิธีการเลือกและวิธีการสกัดลักษณะเด่นที่เลือกใช้ในงานวิจัยนี้ คือ Heuristic Greedy Algorithm of Item Set และ การวิเคราะห์องค์ประกอบหลัก ตามลำดับ เมื่อได้ลักษณะตามที่ต้องการแล้วผู้วิจัยได้ทำการทดสอบผลการแบ่งกลุ่มข้อมูลด้วยวิธีการเรียนรู้แบบมีผู้สอน 3 วิธี คือ BPNN, RBF และ SVM จากผลการทดลองข้อมูล KDD99 จำนวน 13,499 จุดข้อมูล (patterns) 34 ลักษณะ พบว่าวิธีการวิเคราะห์องค์ประกอบหลักสามารถสกัดลักษณะเด่นออกมาได้จำนวน 19 ลักษณะ และ วิธี Heuristic Greedy Algorithm of Item Set ได้ผลการเลือกลักษณะข้อมูลจำนวน 13 ลักษณะ ผลการแบ่งกลุ่มข้อมูลด้วยวิธีการที่เลือกใช้ พบว่าการเลือกลักษณะด้วยวิธี Heuristic Greedy Algorithm of Item Set ให้ค่าความถูกต้องสูงกว่า การสกัดลักษณะเด่นด้วยวิธีการวิเคราะห์องค์ประกอบหลัก

คำสำคัญ: การสกัดลักษณะเด่น, การเลือกลักษณะ, การรู้จำรูปแบบ, การตรวจจับการบุกรุกเครือข่าย

Abstract

This paper proposes a feature selection and extraction methods of network intrusion data which are the heuristic greedy algorithm (HGAIS) of item set and principal component analysis (PCA), respectively. After proposed feature selection and extraction steps, we use three standard supervised learning algorithms which are BPNN, RBF and SVM for evaluating the significance of the selecting features. It can be seen that from the KDD99 (with 13,499 sampling patterns) with 34 data dimensions based on HGAIS and PCA algorithms, we obtain 19 and 13 features, respectively. In addition, the classification accuracies confirm that HGAIS algorithm produces better features than the PCA.

Key Words: *Feature Extraction, Feature Selection, Pattern Recognition, Network Intrusion Detection*

1. บทนำ

จากการพัฒนาอย่างรวดเร็วของเครือข่ายอินเทอร์เน็ตนั้น ทำให้คนส่วนใหญ่หันมาตระหนักถึงการรักษาความปลอดภัยกันมากขึ้น วิธีการหนึ่งที่นิยมนำมาใช้ในการสร้างความปลอดภัยให้กับระบบเครือข่ายคอมพิวเตอร์ คือ การตรวจจับการบุกรุก (Intrusion Detection) วิธีการของการตรวจจับการบุกรุกสามารถแบ่งออกได้เป็น 2 ชนิด คือ วิธีการตรวจจับการบุกรุกแบบมิสยูส (misuse intrusion detection method) และ วิธีการตรวจจับการบุกรุกแบบอโนมาลี (anomaly intrusion detection method) โดยที่การตรวจจับการบุกรุกแบบมิสยูสเป็นวิธีการหาผู้บุกรุกโดยการเปรียบเทียบข้อมูลที่เข้ามากับรูปแบบของผู้บุกรุกที่มีอยู่เดิมแต่ไม่สามารถตรวจจับการบุกรุกแบบใหม่ หรือการบุกรุกที่ไม่มีในชุดรูปแบบของผู้บุกรุกที่มีได้ ส่วนวิธีการตรวจจับการบุกรุกแบบอโนมาลีนั้นเป็นวิธีการหาผู้บุกรุกโดยการวิเคราะห์การใช้งานที่เบี่ยงเบนไปจากระดับการใช้งานโดย

ปกติโดยทั่วๆ ไปมีหลายวิธีถูกนำมาสร้างเป็นต้นแบบเพื่อระบุผู้บุกรุก และปัญหาการตรวจจับการบุกรุกสามารถพิจารณาได้ในลักษณะเดียวกับปัญหาการแบ่งกลุ่ม (Classification Problem) โดยจะประมวลผลข้อมูลที่ต้องการตรวจสอบเพื่อแบ่งกรณีที่เป็นการบุกรุก และที่ไม่ใช่การบุกรุก และเนื่องจากข้อมูลที่ส่งผ่านทางเครือข่ายอินเตอร์เน็ตหรือข้อมูลที่ตรวจสอบนั้นมีปริมาณมากทั้งจำนวนข้อมูล และจำนวนลักษณะของข้อมูลเป็นผลทำให้เกิดความล่าช้าในการระบุผู้บุกรุก และอาจเป็นสาเหตุให้การบุกรุกบางชนิดสามารถบุกรุกเข้าสู่ระบบเครือข่ายได้

จากปัญหาที่พบข้างต้น ได้มีความพยายามที่จะพัฒนาประสิทธิภาพของการตรวจจับการบุกรุกโดยนำวิธีการต่างๆ เพื่อมาช่วยในการลดลักษณะข้อมูลและเพิ่มประสิทธิภาพการรู้จำหรือระบุผู้บุกรุก การลดลักษณะข้อมูลที่ดีเมื่อลดจำนวนลักษณะลงแล้วควรจะให้ค่าความถูกต้องของการตรวจจับการบุกรุกได้ดี ซึ่งมี 2 วิธีการ คือ วิธีการเลือกลักษณะ และ วิธีการสกัดลักษณะเด่น โดยการเลือกลักษณะนั้นเป็นการเลือกลักษณะบางลักษณะจากข้อมูลเดิมที่มีความสำคัญ เช่น วิธีการเลือกลักษณะด้วยวิธีเชิงพันธุกรรม ซึ่งจะตัดลักษณะที่ไม่มีความสำคัญหรือมีความสำคัญน้อยออกไป ส่วนการสกัดลักษณะเด่น เช่น การวิเคราะห์องค์ประกอบหลัก จะช่วยลดความซ้ำซ้อนของข้อมูล ซึ่งจะได้ตัวแทนข้อมูลชุดใหม่ที่มีจำนวนลักษณะน้อยลง แต่เนื่องจากการสกัดลักษณะเด่นเป็นการหาตัวแทนข้อมูลชุดใหม่ซึ่งอาจจะทำให้ข้อมูลที่มีความสำคัญนั้นเปลี่ยนไปเป็นผลทำให้ภาพในกระบวนการการรู้จำมีประสิทธิภาพที่น้อยลงได้

จากที่ได้กล่าวมาทั้งหมดนั้น ผู้วิจัยได้แสดงให้เห็นแล้วว่าการเลือกลักษณะที่สำคัญของชุดข้อมูลบนเครือข่าย มีความสำคัญต่อการพัฒนาการระบุผู้บุกรุกเป็นอย่างมาก จึงจำเป็นที่จะต้องหาวิธีการที่ดีในการเลือกลักษณะที่สำคัญของชุดข้อมูลบนเครือข่าย เพื่อให้ได้ตัวแทนชุดลักษณะของชุดข้อมูลที่เหมาะสมและเป็นการลดจำนวนลักษณะเพื่อใช้ในการระบุผู้บุกรุกโดยอาศัยวิธีการ Heuristic Greedy algorithm ซึ่งขั้นตอนนี้ไม่มีรูปแบบวิธีการขั้นตอนโดยตรง แต่จะพิจารณาว่าข้อมูลที่มีอยู่ในขณะนั้นมีทางเลือกใดที่ให้คำตอบที่ดีที่สุดของปัญหา โดยการหา item set และเลือกลักษณะที่ดีที่สุด เมื่อนำมาทำการแบ่งกลุ่มข้อมูลในขณะนั้นการพัฒนาการเลือกลักษณะชุดข้อมูลเครือข่ายประกอบไปด้วย 2 ขั้นตอน คือ 1.หาลักษณะของชุดข้อมูลที่สามารถแทนข้อมูลได้และมีจำนวนลักษณะที่เหมาะสม และขั้นตอนที่ 2 การรู้จำรูปแบบการบุกรุกเพื่อระบุผู้บุกรุกจากชุดข้อมูลบนเครือข่าย จากลักษณะที่ได้จากการสกัดลักษณะของชุดข้อมูล โดยวัดประสิทธิภาพจากอัตราความเร็วในการตรวจจับผู้บุกรุก และเปอร์เซ็นต์ความผิดพลาดของการตรวจจับผู้บุกรุกบทความนี้นำเสนอการเลือกลักษณะของข้อมูลผู้บุกรุกด้วย Heuristic Greedy Algorithm of Item Set ซึ่งในส่วนที่ 2 กล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้อง ส่วนที่ 3 คือวิธีการที่นำเสนอ ส่วนที่ 4 วิธีในการวัดประสิทธิภาพ ส่วนที่ 5 การทดลองและผลการทดลอง และส่วนที่ 6 สรุปผลการทดลอง

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในงานวิจัยนี้ ผู้วิจัยได้นำเสนอวิธีการเลือกลักษณะด้วยวิธีการ Heuristic Greedy Algorithm of Item Set โดยใช้กฎความสัมพันธ์ (association rules) และใช้หลักการ apriori ในการสร้าง item set และอีกวิธีหนึ่ง คือ สกัดลักษณะเด่นด้วยการวิเคราะห์องค์ประกอบหลัก เพื่อวัดประสิทธิภาพจะทดสอบด้วยวิธีการเรียนรู้แบบมีผู้สอน 3 วิธี คือ BPNN, RBF และ SVM

Murat Karabatak และคณะ [1] ได้นำเสนองานวิธีการเลือกลักษณะบนพื้นฐานของกฎความสัมพันธ์ (association rules) และ โครงข่ายประสาทเทียม ถูกนำเสนอสำหรับการวินิจฉัยโรค erythemato-squamous กฎความสัมพันธ์ใช้เพื่อลดจำนวนลักษณะของข้อมูล และโครงข่ายประสาทเทียมใช้สำหรับกระบวนการการจำแนกกลุ่ม และเปรียบประสิทธิภาพกับวิธีการเลือกลักษณะวิธีอื่น หลังจากใช้กฎความสัมพันธ์เลือกลักษณะสามารถลดจำนวนจาก 34 ลักษณะ เหลือ 24 ลักษณะ มีอัตราการจำแนกกลุ่มถูกต้อง 98.61% ซึ่งให้ค่าความถูกต้องมากกว่ากับข้อมูลที่ไม่ได้ผ่านการเลือกลักษณะและการเลือกลักษณะวิธีอื่นๆ ผลการทดลองแสดงให้เห็นว่าการเลือกลักษณะมีความสำคัญและทำให้การจำแนกกลุ่มข้อมูลเพื่อวินิจฉัยโรค erythemato-squamous มีประสิทธิภาพ

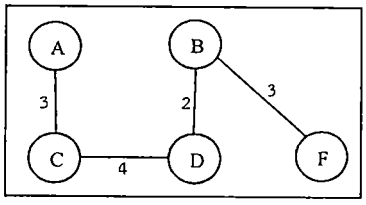
Jing Zhang, Jianmin Wang, Deyi Li, Huacan He, Jiaguang Sun (2003) [2] ได้นำเสนองานวิจัยเรื่อง A New Heuristic Reduct Algorithm Base on Rough Sets Theoryเนื่องจากการนำทฤษฎีเซตอย่างหยาบมาเพื่อหาซับเซตของลักษณะที่เหมาะสมที่สุดจากการเลือกลักษณะเป็น

วิธีที่ใช้เวลานาน จึงนำเสนอวิธีการ heuristic algorithm บนพื้นฐานของทฤษฎีเซตอย่างหยาบเพื่อหาซับเซตของลักษณะที่เหมาะสมและใช้เวลาน้อย ผลการทดลองกับชุดข้อมูลหลายๆชุดแสดงให้เห็นว่าส่วนใหญ่วิธีการที่นำเสนอสามารถหาซับเซตของลักษณะได้เหมาะสมที่สุดได้อย่างรวดเร็วและมีประสิทธิภาพ

Dong Seong Kim และคณะ [3] ได้นำเสนองานวิจัยเรื่อง An Optimized Intrusion Detection System Using PCA and BNN โดยได้นำเสนอการหาค่าที่เหมาะสมสำหรับการตรวจจับการบุกรุกโดยอาศัยการวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis: PCA) และโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ (Backpropagation Neural Network: BNN) โดยมุ่งเน้นในการแก้ปัญหา 2 ปัญหาด้วยกันคือ การกำหนดจำนวนของ Hidden Layer และการจัดการค่าของน้ำหนักเพื่อใช้ในการกำหนดรูปแบบของโครงข่ายประสาทเทียม และการประมวลผลข้อมูลที่ตรวจสอบที่มีปริมาณมาก โดยพิจารณาถึงการเพิ่มอัตราการตรวจจับและลดเวลาการประมวลผล โดยนำข้อดีของ Genetic Algorithm (GA) มาใช้ โดยการทำงานของ GA จะทำงานบนการทำงานที่รวมกันระหว่าง PCA และ BNN แต่ผลการทดลองยังออกมาไม่เป็นที่น่าพอใจตามที่คาดหวังไว้ ในส่วนงานในอนาคตได้มีการชี้ถึงประเด็นว่า ถ้ามีการปรับเปลี่ยนตัว PCA และ BPN น่าจะทำให้ได้ผลการทดลองที่ดีขึ้น

2.1 Heuristic Greedy Algorithm

Heuristic Greedy Algorithm เป็นขั้นตอนวิธีการแก้ปัญหาที่คิดแบบง่ายๆและตรงไปตรงมา [4] ซึ่งเป็นการแก้ปัญหาในลักษณะที่ไม่มีรูปแบบวิธีการขั้นตอนโดยตรง โดยจะพิจารณาว่าข้อมูลที่มีอยู่ในขณะนั้นมีทางเลือกใดที่ให้คำตอบที่ดีที่สุดของปัญหา โดยการเลือกคำตอบที่ดีที่สุดในขณะนั้น ซึ่งถ้าข้อมูลนั้นเพียงพอที่จะทำให้สรุปคำตอบที่ดีที่สุด เราจะได้ขั้นตอนวิธีที่มีประสิทธิภาพ เช่น การพิจารณาเลือกทางเลือกของกราฟต้นไม้ที่สามารถเชื่อมต่อกันได้ทุกโหนด แต่ไม่ก่อให้เกิดเป็นกราฟวงกลม และมีค่าน้ำหนักของเส้นเชื่อมรวมทุกโหนดน้อยที่สุดดังรูปที่ 1



รูปที่ 1 การทำงานของ Heuristic Greedy Algorithm

2.2 การวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis: PCA)

วิธีการวิเคราะห์องค์ประกอบหลัก เป็นวิธีการทางสถิติเพื่อใช้ในการสกัดปัจจัยที่อาศัยหลักความสัมพันธ์เชิงเส้นตรงระหว่างตัวแปรที่ใช้เป็นข้อมูล [5] องค์ประกอบหลักตัวแปร คือ การการผสมเชิงเส้นตรง (Linear Combination) ของตัวแปรที่อธิบายการผันแปรของข้อมูลได้มากที่สุด จากนั้นหาการผสมเชิงเส้นครั้งที่สองที่สามารถอธิบายการผันแปรได้มากที่สุดเป็นอันดับที่สอง โดยที่ไม่สัมพันธ์กับการผสมครั้งแรก การวิเคราะห์องค์ประกอบหลักถูกนำไปประยุกต์ใช้งานต่างๆ เช่น การบีบอัดข้อมูล, การสร้างภาพใบหน้าไอเกนเพื่อใช้ในระบบจดจำ และ การลบออกของพื้นหลังโดยใช้ไอเกน เป็นต้นวิธีการวิเคราะห์องค์ประกอบหลักสามารถนำมาใช้ในการลดมิติของข้อมูลโดย การวิเคราะห์ข้อมูลและเลือกเฉพาะข้อมูลที่มีความสำคัญเท่านั้น ส่วนข้อมูลที่ไม่สำคัญจะถูกตัดทิ้งไป ดังนั้นเมื่อข้อมูลผ่านกระบวนการ PCA แล้ว จะได้ผลลัพธ์เป็นไอเกนเวกเตอร์และค่าไอเกน ซึ่งไอเกนเวกเตอร์ที่มีค่าสมนัยกับค่าไอเกนที่มีค่าสูงๆ จะเป็นการดึงข้อมูลที่มีความถี่ต่ำ ส่วนไอเกนเวกเตอร์ที่สมนัยกับค่าไอเกนที่ต่ำๆ จะเป็นการดึงข้อมูลที่มีความถี่สูง

2.2.1 การหาค่าไอเกน และไอเกนเวกเตอร์ (Eigen Value and Eigen Vector)

ความหมายของค่าไอเกน และไอเกนเวกเตอร์ กำหนดให้ A เป็นค่าเมตริกซ์จัตุรัส

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ a_{31} & a_{32} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

และ v เป็นเวกเตอร์หลัก (Colum Vector) และ λ เป็นค่าคงที่ใดๆ โดยที่

$$v = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_3 \end{bmatrix}$$

ที่ทำให้ $Av = \lambda v$ (1)

หรือ $(\lambda I - A)v = 0$ (2)

เมื่อ A คือ ค่าเมทริกซ์

λ คือ เป็นค่าคงที่ใดๆ เป็นสเกลาร์

v คือ ค่าไอเกนเวกเตอร์

จากสมการจะเห็นว่า $v = 0$ ที่ทำให้สมการ เป็นจริงทุกๆ ค่าของ λ

2.3 โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ (Back Propagation Neural Network: BPNN)

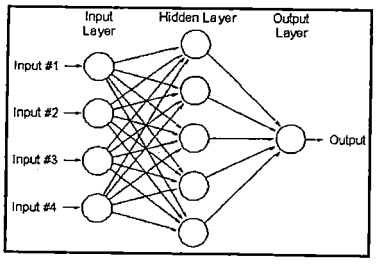
โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ เป็นขั้นตอนวิธีในการเรียนรู้ของเครือข่ายประสาทเทียมวิธีหนึ่งที่นิยมใช้ในโครงข่ายประสาทเทียมหลายชั้น [6] ประกอบไปด้วยชั้นข้อมูลเข้า ชั้นซ่อน และชั้นข้อมูลออก ดังรูปที่ 2 ซึ่งชั้นซ่อนอาจมีชั้นเดียวหรือมากกว่าก็ได้ เพื่อปรับค่าน้ำหนักในเส้นเชื่อมให้มีค่าผิดพลาดกำลังสองเฉลี่ยน้อยที่สุด ดังสมการ

$$MSE(\vec{w}) = \frac{1}{2}\sum_{p\in P}\sum_{k\in outputs}(d_{p,k} - o_{p,k})^2 \quad (3)$$

โดยที่ $outputs$ คือเซตโหนดข้อมูลออก

$d_{p,k}$ คือ ค่าข้อมูลออกเป้าหมาย โหนดที่ k ตัวอย่างที่ p

$o_{p,k}$ คือ ค่าข้อมูลออกที่ได้ โหนดที่ k ตัวอย่างที่ p



รูปที่ 2 โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ (ที่มา http://www.odec.ca/projects/2006/stag6m2/background.html)

2.4 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM)

ซัพพอร์ตเวกเตอร์แมชชีน มีจุดมุ่งหมายที่สำคัญคือการหาเส้นไฮเปอร์เพลน ซึ่งใช้แบ่งข้อมูลออกเป็นคลาส เพื่อให้ได้ผลลัพธ์ที่ดี [7] โดยพิจารณาจากสมการเส้นตรงไฮเปอร์เพลน เพื่อค้นหาจุดของข้อมูลที่อยู่ใกล้เส้นแบ่งไฮเปอร์เพลน เรียกจุดนี้ว่า ซัพพอร์ตเวกเตอร์ มีหลักการดังนี้

1.นำข้อมูลคำนวณหาค่า yซึ่งค่า $y \in \{-1,1\}$จากสมการ

$$y = w^T x + b \quad (4)$$

2.คำนวณหาเส้นแบ่ง Optimal Hyperplane จากสมการ

$$w^T x + b = 0 \quad (5)$$

3.ระยะทาง (d) หรือจากเส้นขอบ ณ จุด x_i ไปยังไฮเปอร์เพลนแสดงดังสมการ

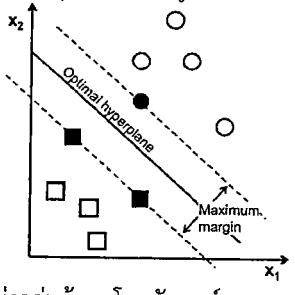
$$\mathrm{d} = \frac{|\mathrm{w^T x_i + b}|}{\|\mathrm{w}\|} \quad (6)$$

w คือ เวกเตอร์น้ำหนัก

x_i คือ ข้อมูลนำเข้า

b คือ ค่าคงที่ที่กำหนดขึ้นเพื่อให้เหมาะสมกับการจัดกลุ่ม

4. เลือกจุดที่อยู่ใกล้เส้นตร Optimal Hyperplane ทั้งเหนือเส้นซึ่งเรียกว่า ขอบล่าง ซึ่งเป็นขอบล่างสุดของคลาสเอกสารที่อยู่เหนือเส้นตรง Optimal Hyperplane และใต้เส้นเรียกว่า ขอบบน ซึ่งเป็นขอบบนสุดของคลาสเอกสารที่อยู่ใต้เส้นตรง Optimal Hyperplane เพื่อที่จะหาระยะทางระหว่างเส้นขอบทั้งสองโดยจะเลือกเอาค่าระยะทางที่ห่างจากเส้นตรง Optimal Hyperplane ที่น้อยที่สุดเป็นตัวเลือกในการจัดกลุ่มเอกสารดังรูปที่ 3



รูปที่ 3 การแบ่งกลุ่มข้อมูลโดยซัพพอร์ตเวกเตอร์แมชชีน (ที่มา *http://docs.opencv.org/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html)*

2.5 โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน (Radial Basis Function: RBF)

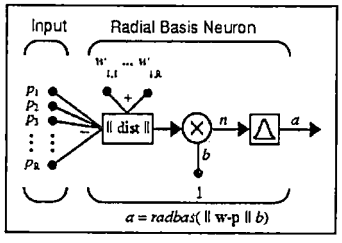
โดยแบบที่นิยมใช้โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน เป็นโครงข่ายประสาทเทียมป้อนไปข้างหน้าแบบหลายชั้น [8] จะประกอบไปด้วย 3 ชั้น ได้แก่ ชั้นรับข้อมูลเข้า ชั้นซ่อน และชั้นข้อมูลออก ดังรูปที่ 4 โดยเป็นฟังก์ชันการส่งระหว่างชั้นรับข้อมูลเข้า $p \in \mathbb{R}^{Nx1}$ ไปยังชั้นข้อมูลออก $y \in \mathbb{R}^{Mx1}$ จะได้ข้อมูลออกของเครือข่ายดังสมการที่ 7

$$y_i = \sum_{k=i}^{s} w_{ik} \emptyset_k(\|p - c\|) \qquad (7)$$

โดยที่ w_{ik} คือ ค่าน้ำหนักนิวรอนในชั้นซ่อน

s คือ จำนวนนิวรอนในชั้นซ่อน

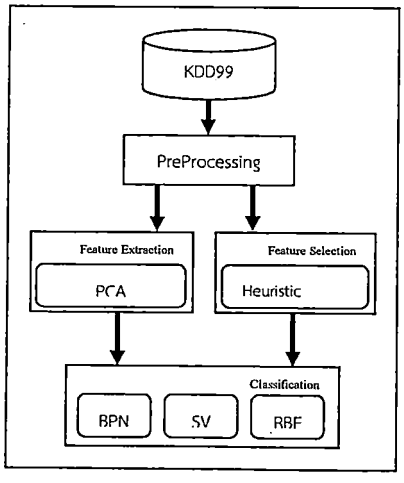
c คือ เวกเตอร์จุดศูนย์กลาง



รูปที่ 4 โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน
(ที่มา http://matlab.izmiran.ru/help/toolbox/nnet/radial74.html)

3. วิธีการที่นำเสนอ

งานวิจัยนี้ได้นำเสนอวิธีการในการดำเนินการวิจัย เป็นส่วนๆ ดังแสดงในรูปที่ 5



รูปที่ 5 ขั้นตอนการดำเนินงาน

การเตรียมชุดข้อมูล (PreProcessing)

ในงานวิจัยนี้จะใช้ข้อมูล 10% ของชุดข้อมูล KDD99 [10] ทั้งหมดมาทำการวิจัย โดยชุดข้อมูลนี้แบ่งออกได้เป็น 5 ชนิด คือ Normal, Dos, Probe, U2R และ R2L และสุ่มออกมาเพื่อให้ง่ายต่อการทดสอบจะได้จำนวนทั้งหมด 13,499 จุดข้อมูล จากนั้นตัดลักษณะข้อมูลในส่วนที่ลักษณะข้อมูลที่เป็นสัญลักษณ์ และมีค่าเป็นศูนย์เนื่องจากไม่มีผลในการทำวิจัย ซึ่งจะเหลือจำนวนลักษณะทั้งหมด 34 ลักษณะ

การสกัดลักษณะชุดข้อมูล(Feature extraction)

การสกัดลักษณะชุดข้อมูล จะใช้วิธีการวิเคราะห์องค์ประกอบหลักในการสกัด โดยเลือกลักษณะจากค่าไอเกนที่เรียงลำดับจากน้อยไปมากที่มีอัตราค่าไอเกนสะสมและค่าไอเกนสะสมของทั้งหมดมากกว่า 0.95 ดังสมการที่ 8

$$\frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{N} \lambda_i} > threshold \qquad (8)$$

λ_i คือ ค่าไอเกนลำดับที่ i

N คือ จำนวนลักษณะทั้งหมด

K คือ จำนวนลักษณะที่ถูกเลือก

$threshold$ คือ ค่าเกณฑ์ที่บ่งบอกว่าต้องการให้องค์ประกอบหลักที่ได้มีค่าไอเกนสะสมใกล้เคียงกับค่าไอเกนสะสมทั้งหมดมากน้อยเพียงใด ในที่นี้กำหนดให้ $threshold$ เท่ากับ 0.95

การเลือกลักษณะของข้อมูลขนาด 3-candidate item set (Feature Extraction)

การเลือกลักษณะชุดข้อมูล ใช้วิธีการเลือกโดย Heuristic Greedy Algorithm ของ item set ซึ่งมีวิธีการดังนี้

ขั้นตอนที่ 1: สร้าง 1-item set โดยนำแต่ละลักษณะหาค่า RMSE (root mean square error) โดยใช้โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน

ขั้นตอนที่ 2: สร้าง 2-candidate item set โดยการนำแต่ละลักษณะมาจับคู่กันทุกๆ ลักษณะที่เป็นไปได้ และหาค่า RMSE โดยใช้โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน

ขั้นตอนที่ 3: สร้าง 2-itemset โดยนำ 2-candidate item set ที่มีค่า RMSE น้อยกว่า 1-itemset ของตัวมันเอง

ขั้นตอนที่ 4: สร้าง 3-candidate item setโดยนำ 2-itemset จำนวน 3 เซตมายูเนียนกันทุกๆ 3 เซต ที่เป็นไปได้และหาค่า RMSEโดยใช้โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน

ขั้นตอนที่ 5: นำ 3-candidate item set หาค่า RMSEโดยใช้โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน

ขั้นตอนที่ 6: เลือกเซตลักษณะโดยการสุ่มเลือกจาก 2-itemset และ 3-candidate item set หาก item set ใด มีค่า RMSEต่ำ จะมีโอกาสสุ่มเลือกมากกว่า

การแบ่งกลุ่มด้วยวิธีการเรียนรู้แบบมีผู้สอน (Classification)

ในขั้นตอนนี้จะทำการทดสอบผลการแบ่งกลุ่มข้อมูลด้วยวิธีการเรียนรู้แบบมีผู้สอน 3 วิธี คือ โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ ซัพพอร์ตเวกเตอร์แมชชีน และโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน

การประเมินระบบ

ในการทดสอบแบ่งกลุ่มข้อมูล งานวิจัยนี้จะแบ่งข้อมูลออกเป็น 10 ชุด หรือ 10 fold cross validation เพื่อใช้ในการฝึกฝน และการทดสอบ จากนั้นวัดค่าความถูกต้อง และค่าเฉลี่ยเรขาคณิต และเวลาที่ใช้ในการประมวลผล เพื่อประเมินตัวระบบต้นแบบต่อไป เพื่อให้ได้ตัวต้นแบบที่เหมาะสมทั้งการเลือกลักษณะและการสกัดลักษณะชุดข้อมูลเครือข่าย และตัวแบบการรู้จำเพื่อระบุผู้บุกรุก

4. การวัดประสิทธิภาพ

วิธีการวัดประสิทธิภาพการจำแนกข้อมูลของชุดข้อมูลก่อนและหลังการเลือกลักษณะด้วยวิธีการที่นำเสนอ ใช้วิธีการวัดค่าความถูกต้อง (Accuracy) ค่าเฉลี่ยเรขาคณิต (Geometric Mean: G-Mean) และเวลาที่ใช้ในการจำแนกข้อมูล

การหาค่าความถูกต้องของการจำแนกกลุ่ม วัดได้จากอัตราส่วนระหว่างจำนวนข้อมูลที่แบ่งกลุ่มถูกต้องและจำนวนข้อมูลทั้งหมด ดังสมการที่ 9

$$AC = \frac{Cor}{Ins} \qquad (9)$$

โดยที่ Cor คือ จำนวนข้อมูลที่แบ่งกลุ่มถูกต้อง

Ins คือ จำนวนข้อมูลทั้งหมด

การหาค่าเฉลี่ยเรขาคณิต คือ การหาค่าเฉลี่ยเรขาคณิตของอัตราความถูกต้องของการจำแนกกลุ่มในแต่ละคลาสดังสมการที่ 10

$$GM = \sqrt[n]{\Pi TPR} \qquad (10)$$

โดยที่ TPR คือ อัตราการแบ่งกลุ่มที่ถูกต้องของคลาสแต่ละคลาส

n คือ จำนวนของคลาสทั้งหมด

5. การทดลองและผลการทดลอง

ข้อมูลที่นำมาใช้ในการทำแบบทดลอง เป็นข้อมูลที่ได้จากฐานข้อมูลความรู้ (Knowledge Discovery in Database (KDD) Cup data) ซึ่งเป็นชุดข้อมูลในปี 1999 ชุดข้อมูลนี้ถูกสร้างตาการจำลองการโจมตีของผู้บุกรุกจาก U.S. Air Force local area network มีจำนวนประมาณ 4,900,000 จุดข้อมูล มี 41 ลักษณะ ดังรูปที่ 6 ตัวอย่างข้อมูล KDD cup 1999 ซึ่งข้อมูลอยู่ในรูปแบบของสัญลักษณ์ และจำนวนจริง โดยลักษณะสุดท้ายคือ class ที่บ่งบอกว่าข้อมูลชุดใดเป็นลักษณะปกติหรือบุกรุก ซึ่งแบ่งออกเป็น 5 ประเภทใหญ่ ดังนี้

Normal คือ ข้อมูลมีลักษณะปกติหรือไม่มีการบุกรุก

Dos คือ ผู้บุกรุกพยายามโจมตีให้ระบบหยุดการทำงาน ซึ่งแบ่งออกเป็นประเภทย่อยๆอีก เช่น smurf

Probing คือ ผู้บุกรุกพยายามตรวจสอบหาจุดอ่อนของระบบ เช่น portsweep

R2L ผู้บุกรุกไม่มี user ในระบบแต่พยายามเจาะ เช่น guess password

U2R ผู้บุกรุกพยายามเข้าสู่ระบบโดยการใช้สิทธิ์ของ super user เช่น buffer overflow

ในแต่ละชุดข้อมูลของ KDD cup 1999 นี้ จะแบ่งลักษณะออกเป็น 3 กลุ่มคือ

Basic Features เป็นลักษณะพื้นฐานที่ได้จากแพคเกจข้อมูลที่สื่อสารในเครือข่าย เช่น ชนิดของโปรโตคอล

Traffic Features เป็นลักษณะที่แสดงถึงลักษณะของการสื่อสาร เช่น เวลาหรือจำนวนครั้งในการติดต่อ

Content Features เป็นลักษณะที่บอกถึงลักษณะการบุกรุกหรือพฤติกรรมที่น่าสงสัย เช่น ความผิดพลาดในการล็อคอิน

รูปที่ 6 ตัวอย่างข้อมูล KDD cup 1999

เนื่องจากข้อมูล KDD cup 1999 มีจำนวนมาก ดังนั้นในงานวิจัยส่วนใหญ่จึงแนะนำให้เลือกข้อมูลเพียงร้อยละ 10 และเพื่อสะดวกในการสอนและทดสอบประสิทธิภาพของระบบการรู้จำจึงทำการสุ่มข้อมูลมาประมาณ 13,499 จุดข้อมูล (Patterns) โดยแบ่งเป็นประเภท Normal จำนวน 4,107 จุดข้อมูล Dos จำนวน 4,107 จุดข้อมูล Prob จำนวน 4,107 จุดข้อมูล R2L จำนวน 1,126 จุดข้อมูล และ U2R จำนวน 52 จุดข้อมูล และตัดบางลักษณะที่ไม่มีผลต่อการรู้จำออกไป เช่น Basic Features และ ลักษณะที่มีค่าเป็นศูนย์ทั้งหมด จึงเหลือจำนวนลักษณะ 34 ลักษณะโดยทดสอบการแบ่งกลุ่มด้วยวิธีการเรียนรู้แบบมีผู้สอน 3 วิธี คือ โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ ซัพพอร์ตเวกเตอร์แมชชีน และโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน กับข้อมูลทั้งหมด ข้อมูลที่ผ่านการสกัดลักษณะด้วยวิธีการวิเคราะห์องค์ประกอบหลักได้ลักษณะจำนวน 19 ลักษณะ และข้อมูลที่ผ่านการเลือกลักษณะด้วยวิธี Heuristic Greedy Algorithm ได้ลักษณะจำนวน 13 ลักษณะ

Learning Method	Accuracy		
	ข้อมูลทั้งหมด(34)	PCA(19)	Heuristic Greedy Algorithm(13)
BPNN	98.7406	97.4739	97.3405
SVM	96.9479	94.081	94.3181
RBF	91.0141	90.4734	95.5256

ตารางที่ 1 ค่าความถูกต้องของการทดลอง

จากตารางที่ 1 ค่าความถูกต้องจากการทดลอง แสดงให้เห็นว่าวิธีการเลือกลักษณะด้วยวิธี Heuristic Greedy Algorithm ส่วนใหญ่ให้ค่าความถูกต้องดีกว่าการสกัดลักษณะด้วยวิธีวิเคราะห์องค์ประกอบหลัก และให้ผลที่ใกล้เคียงเมื่อใช้ข้อมูลทั้งหมด

Learning Method	G-Mean		
	ข้อมูลทั้งหมด(34)	PCA(19)	Heuristic Greedy Algorithm(13)
BPNN	0.8652	0.8215	0.8104
SVM	0.8458	0.8058	0.8234
RBF	0.7626	0.7015	0.7445

ตารางที่ 2 ค่าเฉลี่ยเรขาคณิตของการทดลอง

เมื่อเปรียบเทียบค่าความถูกต้องของแต่ละคลาสโดยเฉลี่ยด้วยวิธีวัดค่าเฉลี่ยเรขาคณิตดังตารางที่ 2 แสดงให้เห็นว่าวิธีการเลือกลักษณะด้วยวิธี Heuristic Greedy Algorithm ซึ่งได้จำนวนลักษณะ 13 ลักษณะ มีค่าเฉลี่ยเรขาคณิตใกล้เคียงกับข้อมูลทั้งหมดซึ่งมี 34 ลักษณะ และส่วนใหญ่ดีกว่าการสกัดลักษณะด้วยวิธีวิเคราะห์องค์ประกอบหลัก

Learning Method	Processing Time(s)		
	ข้อมูลทั้งหมด(34)	PCA(19)	Heuristic Greedy Algorithm(13)
BPNN	229.43	98.52	65.28
SVM	12.25	16.91	12.74
RBF	16.53	17.60	13.18

ตารางที่ 3 เวลาที่ใช้ในการประมวลผล

จากการทดลองสรุปเวลาที่ใช้ในการประมวลผลแสดงดังตารางที่ 3 ซึ่งเห็นได้ว่าส่วนใหญ่การเลือกลักษณะด้วยวิธี Heuristic Greedy Algorithm ใช้เวลาในการประมวลผลน้อยกว่า

6. สรุปผลการทดลองและข้อเสนอแนะ

งานวิจัยนี้นำเสนอวิธีการเลือกลักษณะด้วยวิธี Heuristic Greedy Algorithm of Item Set และเปรียบเทียบวิธีการสกัดลักษณะเด่นด้วยการวิเคราะห์องค์ประกอบหลักของข้อมูลผู้บุกรุกในเครือข่ายคอมพิวเตอร์ และได้ทำการทดสอบผลการแบ่งกลุ่มข้อมูลด้วยวิธีการเรียนรู้แบบมีผู้สอน 3 วิธี คือ BPNN, RBF และ SVM จากผลการทดลองข้อมูล KDD99 จำนวน 13,499 จุดข้อมูล (patterns) พบว่าวิธีการวิเคราะห์องค์ประกอบหลักสามารถสกัดลักษณะออกมาได้จำนวน 19 ลักษณะ และ วิธี Heuristic Greedy

Algorithm of Item Set ได้ผลการเลือกลักษณะข้อมูลจำนวน 13 ลักษณะ ผลการแบ่งกลุ่มข้อมูลด้วยวิธีการที่เลือกใช้ พบว่าการเลือกลักษณะด้วยวิธี Heuristic Greedy Algorithm of Item Set ส่วนให้ค่าความถูกต้องค่าเฉลี่ยเรขาคณิตที่มีประสิทธิภาพดีกว่า และใช้เวลาในการประมวลผลน้อยกว่าการสกัดลักษณะด้วยวิธีการวิเคราะห์องค์ประกอบหลักเพราะการสกัดลักษณะเด่นเป็นการหาตัวแทนข้อมูลชุดใหม่ซึ่งอาจจะทำให้ข้อมูลที่มีความสำคัญนั้นเปลี่ยนไปเป็นผลทำให้ภาพในกระบวนการการรู้จำมีประสิทธิภาพที่น้อยลงได้แต่วิธีที่นำเสนอจะใช้เวลาในเลือกลักษณะนาน และเนื่องจากข้อมูล KDD99 เป็นข้อมูลที่ไม่สมดุล มีบางคลาสที่มีจำนวนน้อยมากๆ เป็นผลทำให้การเลือกลักษณะด้วยวิธีที่นำเสนอสามารถจำแนกกลุ่มของคลาสน้อยมีค่าความถูกต้องน้อย

7. กิตติกรรมประกาศ

โครงการวิจัยนี้ได้รับการสนับสนุนทุนวิจัยจากสถาบันการวิจัยแห่งชาติ ปีงบประมาณ 2555

ขอขอบคุณ คุณปิยตระกูล บุญทอง ที่ช่วยแนะนำในการเลือกคุณลักษณะที่เหมาะสม

8. เอกสารอ้างอิง

[1] Murat Karabatak, M. Cevdet Ince (2009), "A new feature selection method based on association rules for diagnosis of erythemato-squamous diseases", Expert Systems with Applications, Volume 36, pp. 12500–12505, 2009

[2] Jing Zhang, Jianmin Wang, Deyi Li, Huacan He, Jiaguang Sun (2003), "A New Heuristic Reduct Algorithm Base on Rough Sets Theory", 2003 LNCS 2762, pp. 247–253, 2003.

[3] Dong Seong Kim, Ha-Nam Nguyen, T. Thein, and Jong Sou Park (2005), "An Optimized Intrusion Detection System Using PCA and BNN", Proceedings of The 6th Asia-Pacific Sym. on Information and Telecommunication Technologies, IEICE Communications Society, pp. 356-359.

[4] T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein, Introduction to Algorithms (3e), 2001, p.360.

[5] *Jackson, J. E.,* A User's Guide to Principal Components, *John Wiley and Sons, 1991, p. 592.*

[6] Robert Hecht Nielsen, Theory of the back propagation neural network in Proceedings 1989 IEEE IJCNN, pp. 1593–1605, IEEE Press, New York, 1989.

[7] M. Hearst, ed., "Support Vector Machines," IEEE Intelligent Systems Magazine, Trends and Controversies, Marti Hearst, ed., vol 13, no 4, 1998.

[8] S.Chen, C. F. N. Cowan, P. M. Grant (1991), "Orthogonal Least Squares Learning Algorithm for Radial Basis Function Networks" IEEE transactions on neural networks, vol. 2, no.2.

[9] KDD'99 datasets, The UCI KDD Archive, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, Irvine, CA, USA, 1999.

[10] จิราภรณ์ ถมแก้ว, "การจำแนกข้อมูลโดยคัดเลือกคุณลักษณะที่สำคัญ", (2554), การประชุมวิชาการเสนอผลงานวิจัยระดับบัณฑิตศึกษาแห่งชาติ ครั้งที่ 23.