การเลือกลักษณะของข้อมูลผู้บุกรุกด้วยขั้นตอนวิธีฮิวริสติกกรีดดี

จรรยา อ้นปันส์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิขาวิทยาการคอมพิวเตอร์
คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา
พ.ศ. 2556

INTRUSION FEATURE SELECTION USING HEURISTIC GREEDY ALGORITHM OF ITEMSET

JANYA ONPANS

A THESIS SUBMITTED IN PARTAIL FULFILLMENT OF THE REQUIREMENT
FOR THE MASTER DEGREE OF SCIENCE IN COMPUTER SCIENCE
FACULTY OF INFORMATICS BURAPHA UNIVERSITY
2013.

คณะกรรมการควบคุมวิทยานิพนธ์และคณะกรรมการสอบวิทยานิพนธ์ได้พิจารณาวิทยานิพนธ์ของ จรรยา อ้นปันส์ ฉบับนี้แล้ว เห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ ของมหาวิทยาลัยบูรพาได้

คณะกรรมการควบคุมวิทยานิพนธ์

ผู้ช่วยศาสตราจารย์ ดร.กฤษณะ ชินสาร อาจารย์ที่ปรึกษา

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ศาสตราจารย์ ดร.ชิดชนก เหลือสินทรัพย์)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.กฤษณะ ชินสาร)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.สุวรรณา รัศมีขวัญ)

คณะวิทยาการสารสนเทศ อนุมัติให้รับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ ของมหาวิทยาลัยบูรพา

..... คณบดีคณะวิทยาการสารสนเทศ
(ผู้ช่วยศาสตราจารย์ ดร.สุวรรณา รัศมีขวัญ)
วันที่ 22 เดือน มิถุนายน พ.ศ. 2556

ประกาศคุณูปการ

วิทยานิพนธ์นี้สำเร็จลุล่วงไปได้ด้วยดี ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.กฤษณะ ชินสาร อาจารย์ผู้ควบคุมวิทยานิพนธ์ และ ผู้ช่วยศาสตราจารย์ ดร.สุวรรณา รัศมีขวัญ ที่ให้ความกรุณา ให้คำปรึกษา ความรู้ และความช่วยเหลือในทุก ๆ ด้าน ทำให้วิทยานิพนธ์นี้มีความคืบหน้าและสำเร็จภายในระยะเวลาที่กำหนด

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.อัณณ์นุพันธ์ รอดทุกข์ และคุณปิยตระกูล บุญทอง ที่คอยให้คำปรึกษา คำแนะนำ แนวทางในการแก้ไขปัญหาต่าง ๆ ที่ทำให้วิทยานิพนธ์นี้สำเร็จได้ด้วยดี

ขอขอบคุณห้องปฏิบัติการวิจัย Knowledge and Smart Technologies คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา ที่ให้การสนับสนุนในการค้นคว้าศึกษางานวิจัยต่าง ๆ และขอขอบคุณพี่ ๆ นักวิจัยที่ให้คำแนะนำ และคำปรึกษาที่ดีที่เป็นประโยชน์ในการจัดทำวิทยานิพนธ์นี้อย่างมาก

สุดท้ายนี้ ขอขอบพระคุณ คุณพ่อคุณแม่ และคุณวันชัย ตระกูลเขียว ที่คอยให้กำลังใจตลอดการทำวิทยานิพนธ์ในครั้งนี้ และเป็นแรงบัลดาลใจให้การทำงานสำเร็จลุล่วงไปด้วยดี จึงกราบขอบพระคุณเป็นอย่างสูง

จรรยา อ้นปันส์

55910163: สาขาวิชา: วิทยาการคอมพิวเตอร์; วท.ม. (วิทยาการคอมพิวเตอร์)
คำสำคัญ: การสกัดลักษณะเด่น/การเลือกลักษณะ/การรู้จำรูปแบบ/การตรวจจับการบุกรุกเครือข่าย
จรรยา อ้นปันส์: การเลือกลักษณะของข้อมูลผู้บุกรุกด้วยขั้นตอนวิธีฮิวริสติกกรีดดี
(INTRUSION FEATURE SELECTION USING HEURISTIC GREEDY ALGORITHM OF ITEMSET)
คณะกรรมการควบคุมวิทยานิพนธ์: กฤษณะ ขินสาร, Ph.D., 146 หน้า. ปี พ.ศ. 2556.

วิทยานิพนธ์นี้นำเสนอวิธีการเลือกลักษณะของข้อมูลผู้บุกรุกในเครือข่ายคอมพิวเตอร์ซึ่งมีจำนวนข้อมูลมาก โดยวิธีการที่นำเสนอ คือ การเลือกลักษณะด้วยขั้นตอนวิธีฮิวริสติกกรีดดี ซึ่งจะทดลองหาฟังก์ชันผิดพลาดที่เหมาะสมในการเลือกลักษณะด้วยวิธีนี้ จากนั้นจะเปรียบเทียบกับวิธีการเลือกลักษณะและสกัดลักษณะอื่น ๆ ได้แก่ การเลือกลักษณะด้วยค่าสถิติไคสแควร์ และการสกัดลักษณะเด่นการวิเคราะห์องค์ประกอบหลัก เมื่อได้ลักษณะตามที่ต้องการแล้วผู้วิจัยได้ทำการทดสอบผลการแบ่งกลุ่มข้อมูลด้วยวิธีการเรียนรู้แบบมีผู้สอนด้วยวิธีโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน โดยจะทดลองกับชุดข้อมูลการตรวจจับการบุกรุก KDDCup99 มีจำนวน 13,499 จุดข้อมูล (Patterns) และ 34 ลักษณะ เพื่อให้ได้ผลการทดลองที่มีประสิทธิภาพทั้งในด้านอัตราค่าความถูกต้อง ค่าความครบถ้วน ค่าความแม่นยำ ค่าเอฟเมเชอร์ อัตราความผิดพลาดเชิงบวก และเวลาที่ใช้ในการทดสอบ นอกจากนี้ยังทดสอบกับชุดข้อมูลอื่น ๆ ด้วย เพื่อให้มีความแน่ใจว่าวิธีการเลือกลักษณะด้วยขั้นตอนวิธีฮิวริสติกกรีดดีที่ได้นำเสนอนั้นมีประสิทธิภาพมากน้อยเพียงใด

จากการทดลองวิธีการเลือกลักษณะที่นำเสนอหรือการสกัดลักษณะด้วยวิธีการต่าง ๆ และทดสอบการแบ่งกลุ่มข้อมูลด้วยวิธีการเรียนรู้แบบมีผู้สอนด้วยวิธีโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐานกับชุดข้อมูลการตรวจจับการบุกรุก KDDCup99 ผลปรากฏว่าการเลือกลักษณะด้วยขั้นตอนวิธีฮิวริสติกกรีดดีได้ผลดีที่สุด และเมื่อทดลองกับข้อมูลชุดอื่น ๆ ก็มีผลที่ดีเช่นเดียวกัน แต่จะดีกว่าการสกัดลักษณะเด่นด้วยวิธีวิเคราะห์องค์ประกอบหลักเพียงเล็กน้อย

55910163: MAJOR: COMPUTER SCIENCE; M.Sc. (COMPUTER SCIENCE)
KEYWORD: FEATURE EXTRACTION/FEATURE SELECTION/PATTERN RECOGNITION/NETWORK INTRUSION DETECTION
JANYA ONPANS: INTRUSION FEATURE SELECTION USING HEURISTIC GREEDY ALGORITHM OF ITEMSET
THESIS ADVISOR: KRISANA CHINNASARN, Ph.D., 146 P. 2013.

This thesis proposes a feature selection method of network intrusion data which are the heuristic greedy algorithm (HGAIS) of item set. And I find appropriate function error for heuristic greedy algorithm. In our work, I compare with chi-square feature selection and popular method of feature extraction is the principal component analysis (PCA). After proposed feature selection and extraction steps, we use standard supervised learning algorithm that is radial basis function (RBF) for evaluating the significance of the selecting features. Evaluation of the propose method is performed by KDDCup99 intrusion detection dataset; 13,499 randomly sampling patterns with 34 data dimensions. Our experimental results have more performance in accuracy rate, recall, precision, F-measure, false alarm rate and processing times. Moreover, experiments with other datasets to make sure that feature selection with propose method has more or less effective.

The experiment results show that, the classification accuracies measure with feature selection by the heuristic greedy algorithm produces better accuracies as compare to the PCA feature extraction and chi-square feature selection.

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญตาราง (ต่อ)

สารบัญภาพ

สารบัญภาพ (ต่อ)

สารบัญภาพ (ต่อ)

สารบัญภาพ (ต่อ)

สารบัญภาพ (ต่อ)

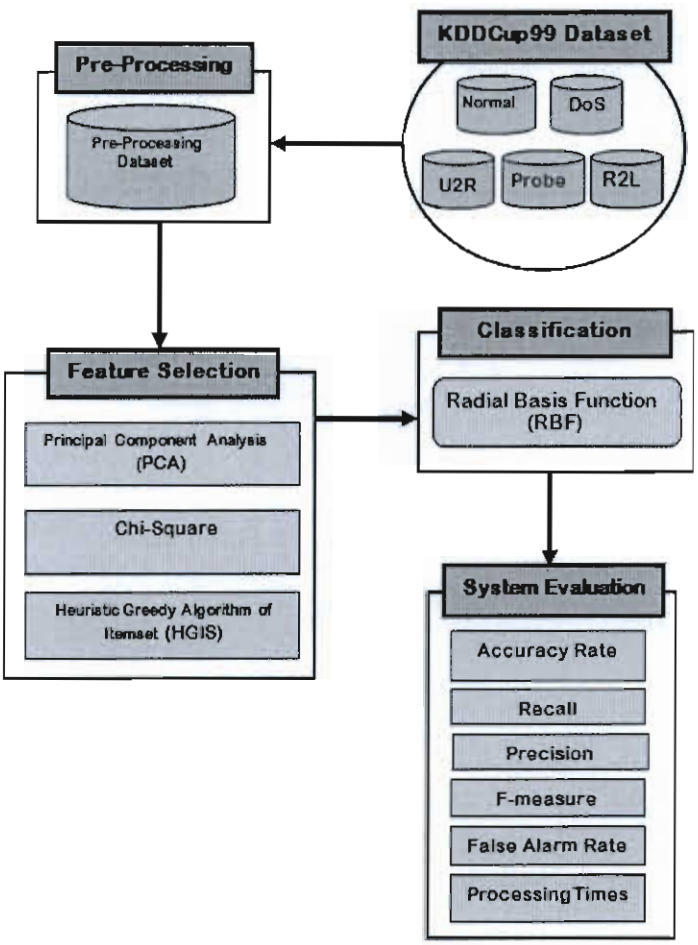
บทที่ 1
บทนำ

ที่มาและความสำคัญ

จากการพัฒนาอย่างรวดเร็วของเครือข่ายอินเทอร์เน็ตนั้น ทำให้คนส่วนใหญ่หันมาตระหนักถึงการรักษาความปลอดภัยกันมากขึ้น วิธีการหนึ่งที่นิยมนำมาใช้ในการสร้างความปลอดภัยให้กับระบบเครือข่ายคอมพิวเตอร์ คือ การตรวจจับการบุกรุก (Intrusion Detection) วิธีการของการดรวจจับการบุกรุกสามารถแบ่งออกได้เป็น 2 ชนิด คือ วิธีการตรวจจับการบุกรุกแบบมิสยูส (Misuse Intrusion Detection Method) และ วิธีการตรวจจับการบุกรุกแบบอโนมาลี (Anomaly Intrusion Detection Method) โดยที่การตรวจจับการบุกรุกแบบมิสยูสเป็นวิธีการหาผู้บุกรุกโดยการเปรียบเทียบข้อมูลที่เข้ามากับรูปแบบของผู้บุกรุกที่มีอยู่เดิมแต่ไม่สามารถดรวจจับการบุกรุกแบบใหม่ หรือการบุกรุกที่ไม่มีในขุดรูปแบบของผู้บุกรุกที่มีได้ ส่วนวิธีการตรวจจับการบุกรุกแบบอโนมาลีนั้นเป็นวิธีการหาผู้บุกรุกโดยการวิเคราะห์การใช้งานที่เบี่ยงเบนไปจากระดับการใช้งานโดยปกติโดยทั่ว ๆ ไปมีหลายวิธีถูกนำมาสร้างเป็นต้นแบบเพื่อระบุผู้บุกรุก และปัญหาการตรวจจับการบุกรุกสามารถพิจารณาได้ในลักษณะเดียวกับปัญหาการจำแนกกลุ่ม (Classification Problem) โดยจะประมวลผลข้อมูลที่ต้องการตรวจสอบเพื่อแบ่งกรณีที่เป็นการบุกรุก และที่ไม่ใช่การบุกรุกและเนื่องจากข้อมูลที่ส่งผ่านทางเครือข่ายอินเตอร์เน็ตหรือข้อมูลที่ตรวจสอบนั้นมีปริมาณมากทั้งจำนวนข้อมูลและจำนวนคุณลักษณะของข้อมูล ซึ่งเป็นผลทำให้เกิดความล่าข้าในการระบุผู้บุกรุก และอาจเป็นสาเหตุให้การบุกรุกบางชนิดสามารถบุกรุกเข้าสู่ระบบเครือข่ายได้

เนื่องจากข้อมูลผู้บุกรุกเครือข่ายมีขนาดของข้อมูลขนาดใหญ่ บางลักษณะของข้อมูลผู้บุกรุกอาจจะไม่มีความสำคัญ และอาจจะส่งผลกระทบให้การตรวจจับการบุกรุกผิดพลาดได้ หากมีวิธีในหาตัวแทนของข้อมูลที่ดีจะสามารถเพิ่มประสิทธิภาพในการตรวจจับการบุกรุก ลดความซ้ำซ้อนของข้อมูลและการลดจำนวนลักษณะข้อมูลทำให้ใช้เวลาในการตรวจจับน้อยลงด้วย สามารถแก้ปัญหานี้โดยการเลือกหรือสกัดลักษณะที่สำคัญเพื่อให้ได้ประสิทธิภาพการตรวจจับผู้บุกทีดีขึ้น หลังจากที่ผู้วิจัยได้ศึกษาการสกัดลักษณะข้อมูลด้วยการวิเคราะห์องค์ประกอบหลักจึงพบว่าข้อมูลที่ได้จากการวิเคราะห์องค์ประกอบหลักนั้นไม่เหมาะสมเนื่องจากข้อมูลผู้บุกรุกมีการกระจายตัวมาก อาจทำให้บางลักษณะที่มีความสำคัญนั้นเปลี่ยนไปเป็นผลทำให้ภาพในกระบวนการการรู้จำมีประสิทธิภาพที่น้อยลงได้ และผู้วิจัยได้ศึกษางานวิจัยในด้านการเลือกลักษณะด้วยขั้นตอนวิธีฮิวริสติกกรีดดี (Heuristic Greedy Algorithm) ซึ่งเป็นการแก้ปัญหาในลักษณะที่ไม่มีรูปแบบวิธีการขั้นตอนโดยตรง โดยจะพิจารณาว่าข้อมูลที่มีอยู่ในขณะนั้นมีทางเลือกใดที่ให้คำตอบที่ดีที่สุดของปัญหา และการตรวจจับการบุกรุกที่มีประสิทธิภาพควรได้ค่าความถูกต้องที่สูง

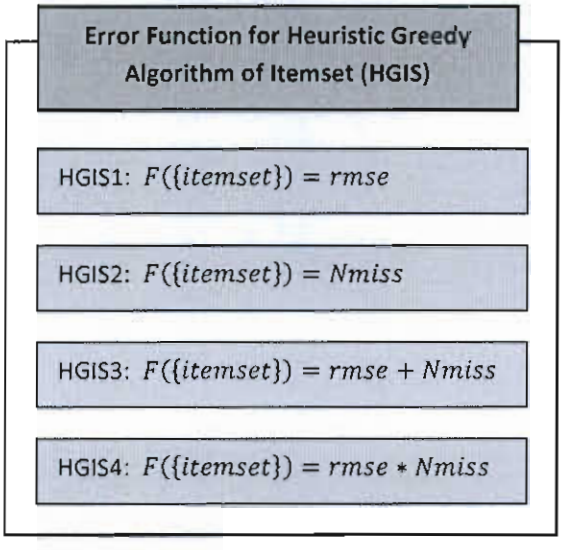
แนวทางในการแก้ปัญหา



ภาพที่ 1-1 ขั้นตอนการแก้ปัญหา

จากที่ได้กล่าวมาทั้งหมดนั้น ผู้วิจัยได้แสดงให้เห็นแล้วว่าการเลือกลักษณะที่สำคัญของชุดข้อมูลบนเครือข่ายมีความสำคัญต่อการพัฒนาการระบุผู้บุกรุกเป็นอย่างมาก จึงจำเป็นที่จะต้องหาวิธีการที่ดีในการเลือกลักษณะที่สำคัญของชุดข้อมูลบนเครือข่าย เพื่อให้ได้ตัวแทนชุดลักษณะของชุดข้อมูลที่เหมาะสมและเป็นการลดจำนวนลักษณะ เพื่อใช้ในการระบุผู้บุกรุกด้วยขั้นตอนวิธีฮิวริสติกกรีดดี ซึ่งเป็นขั้นตอนนี้ไม่มีรูปแบบวิธีการขั้นตอนโดยตรง แต่จะพิจารณาว่าข้อมูลที่มีอยู่ในขณะนั้นมีทางเลือกใดที่ให้คำตอบที่ดีที่สุดของปัญหา โดยการหาชุดลักษณะ (Itemset) จะใช้หลักการ Apriori และเลือกลักษณะที่ดีที่สุด ลักษณะที่ถูกเลือกมานั้นจะให้ค่าความผิดพลาดโดยรวมที่มีประสิทธิภาพ นอกจากนี้ควรจะได้ค่าความถูกต้องที่สูงกับคลาสที่มีจำนวนน้อยด้วย โดยจะเปรียบเทียบกับการสกัดลักษณะด้วยวิธีการวิเคราะห์

องค์ประกอบหลักและการหาค่าสถิติไคสแควร์ จากนั้นนำข้อมูลมาทำการรู้จำรูปแบบการบุกรุกเพื่อระบุผู้บุกรุกจากชุดข้อมูลบนโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน โดยวัดประสิทธิภาพจากอัตราค่าความถูกต้อง (Accuracy Rate) ค่าความครบถ้วน (Recall) ค่าความแม่นยำ (Precision) ค่าเอฟเมเชอร์ (F-measure) อัตราความผิดพลาดเชิงบวก (False Alarm Rate: FAR) และเวลาที่ใช้ในการทดสอบ จากที่กล่าวมาข้างต้นนี้แสดงขั้นตอนการแก้ปัญหาได้ดังภาพที่ 1-1



ภาพที่ 1-2 ฟังก์ชันผิดพลาดที่ใช้ทดสอบ HGIS

ในส่วนของการเลือกลักษณะด้วยวิธีฮิวริสติกกรีด จะทำการทดสอบหาฟังชันก์ที่เหมาะสมสำหรับใช้เป็นเกณฑ์ในการคัดเลือกลักษณะก่อน โดยฟังก์ชันผิดพลาดที่ใช้ทดสอบมี 4 ฟังก์ชันดังภาพที่ 1-2 และเลือกฟังก์ชันที่เหมาะสมที่สุดสำหรับดำเนินการตามขั้นตอนการแก้ปัญหา

วัตถุประสงค์ของโครงงาน

1. เพื่อศึกษาการตรวจจับการบุกรุกกับข้อมูลการบุกรุกที่มีจำนวนมากในเครือข่าย
2. เพื่อเพิ่มประสิทธิภาพการตรวจจับการบุกรุกเครือข่าย และการเลือกลักษณะของข้อมูลที่เหมาะสม
3. เพื่อพัฒนาขั้นตอนวิธีสำหรับเลือกลักษณะข้อมูลการบุกรุกเครือข่าย และการตรวจจับการบุกรุกที่มีประสิทธิภาพ

ขอบเขตของโครงงาน

ชุดข้อมูลที่นำมาทำการทดลองกับขั้นตอนวิธีข้างต้นคือชุดข้อมูล KDDCup99 ประกอบด้วย 13,499 จุดข้อมูล จำนวนคุณลักษณะมี 34 คุณลักษณะ แบ่งออกเป็น 5 กลุ่ม เป็นชุดข้อมูลที่สมบูรณ์ และการกระจายข้อมูลในแต่ละกลุ่มไม่เท่ากันโดยแต่ละกลุ่มมีจำนวนข้อมูลดังนี้

Normal	4,107	จุดข้อมูล
Dos	4,107	จุดข้อมูล
Probe	4,107	จุดข้อมูล
R2L	1,126	จุดข้อมูล
U2R	52	จุดข้อมูล

แผนการดำเนินโครงงาน

ตารางที่ 1-1 แผนการดำเนินโครงงาน

ระยะเวลา / ขั้นตอน	พฤศจิกายน				ธันวาคม				มกราคม				กุมภาพันธ์				มีนาคม			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1. รวบรวมข้อมูล		←	—	—	—	→														
2. ศึกษาขั้นตอนและวิเคราะห์ข้อมูล				←	—	—	—	—	—	→										
3. เขียนโปรแกรมพัฒนาโครงงาน										←	—	—	—	—	—	→				
4. ทดสอบและแก้ไข										←	—	—	—	—	—	—	—	→		
5. จัดทำเอกสารประกอบ				←	—	—	—	—	—	—	—	—	—	—	—	—	—	—	→	

ประโยชน์ที่คาดว่าจะได้รับ

1. ได้ศึกษาการตรวจจับการบุกรุกกับข้อมูลการบุกรุกที่มีจำนวนมากในเครือข่าย
2. เพิ่มประสิทธิภาพการตรวจจับการบุกรุกเครือข่าย และการเลือกลักษณะของข้อมูลที่เหมาะสม
3. ได้ขั้นตอนวิธีสำหรับเลือกลักษณะข้อมูลการบุกรุกเครือข่าย และการตรวจจับการบุกรุกที่มีประสิทธิภาพ

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีที่นำมาใช้ในการพัฒนาวิทยานิพนธ์ เพื่อให้เข้าใจถึงการบุกรุกเครือข่าย การเลือกลักษณะด้วยขั้นตอน**วิธีฮิวริสติกกรีดดี**โดยใช้หลักการ Apriori การสกัดคุณลักษณะด้วยการวิเคราะห์องค์ประกอบหลัก การเลือกลักษณะด้วยค่าสถิติไคสแควร์ ตรวจจับการบุกรุกด้วยการจำแนกกลุ่มวิธีการเรียนรู้แบบมีผู้สอนคือโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน และงานวิจัยที่ศึกษาซึ่งนำมาพัฒนาโครงงานนี้ด้วย ประกอบไปด้วยหัวข้อดังนี้

1. การตรวจจับการบุกรุก
2. ลักษณะของข้อมูลที่ใช้ในการทำแบบทดลอง
3. ขั้นตอนวิธีฮิวริสติกกรีดดี
4. การสร้าง Itemsets โดยใช้หลักการ Apriori
5. การวิเคราะห์องค์ประกอบหลัก
6. การเลือกลักษณะด้วยค่าสถิติไคสแควร์
7. โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน
8. งานวิจัยที่เกี่ยวข้อง

การตรวจจับการบุกรุก

เนื่องจากการพัฒนาอย่างรวดเร็วของโครงข่ายอินเทอร์เน็ตนั้น การรักษาความปลอดภัยจึงเป็นสิ่งสำคัญ โดยปกติแล้วการบุกรุกจะเน้น 3 ด้านคือ การละเมิดความเป็นส่วนตัวหรือความลับ การแก้ไขความถูกต้องของข้อมูล และการทำให้ไม่สามารถใช้งานระบบคอมพิวเตอร์ได้ วิธีการหนึ่งที่นิยมนำมาใช้ในการสร้างความปลอดภัยให้กับระบบเครือข่ายคอมพิวเตอร์ คือ การตรวจจับการบุกรุก (Intrusion Detection) ซึ่งการตรวจจับการบุกรุก สามารถแบ่งออกได้เป็น 2 ชนิดคือ ระบบการตรวจจับการบุกรุกแบบโฮสเบส (Host-based Intrusion Detection Systems) และระบบตรวจจับการบุกรุกแบบเน็ตเวิร์คเบส (Network-based Intrusion Detection Systems) (Scarfone Karen, 2010) โดยที่ระบบตรวจจับการบุกรุกแบบเน็ตเวิร์คเบสนั้นจะติดตั้งที่ระบบเครือข่ายเพื่อทำการตรวจสอบและวิเคราะห์ชุดข้อมูลที่ใช้งานบนเครือข่าย ซึ่งมีความแตกต่างจากระบบการตรวจจับการบุกรุกแบบโฮสเบส ที่จะทำงานอยู่บนระบบเพื่อตรวจสอบและวิเคราะห์ชุดคำสั่งในการระบุการทำงานที่น่าสงสัย วิธีการของการตรวจจับการบุกรุกสามารถแบ่งออกได้เป็น 2 ชนิด คือ วิธีการตรวจจับการบุกรุกแบบอโนมาลี (Anomaly Intrusion Detection Method) และวิธีการตรวจจับการบุกรุกแบบมิสยูส (Misuse Intrusion Detection Method) โดยที่วิธีการตรวจจับการบุกรุกแบบอโนมาลีนั้นเป็นวิธีการหาผู้บุกรุกโดยการวิเคราะห์การใช้งานของผู้ใช้งาน หรือตัวระบบเองที่เบี่ยงเบนไปจากระดับการใช้งานโดยปกติ ส่วนการตรวจจับการบุกรุกแบบมิสยูสนั้นเป็นวิธีการหาผู้บุกรุกโดยการเปรียบเทียบข้อมูลที่เข้ามากับรูปแบบของผู้

บุกรุกที่มีอยู่เดิม ซึ่งทั้งสองวิธีนี้มีจุดแข็งและจุดอ่อนที่แตกต่างกัน ปัญหาที่เด่นชัดที่สุดของการตรวจจับการบุกรุกแบบมิสยูส คือ ไม่สามารถตรวจจับการบุกรุกแบบใหม่ หรือการบุกรุกที่ไม่มีในขุดรูปแบบของผู้บุกรุกที่มีได้ ส่วนการตรวจจับการบุกรุกแบบอโนมาลีนั้น จะระบุว่าการใช้งานที่ตรวจสอบนั้นเป็นผู้บุกรุกหรือไม่นั้นจะตรวจสอบจากการใช้งานนั้นว่ามีการเบี่ยงเบนจากกิจกรรมปกติมากหรือไม่ ดังนั้นการตรวจจับการบุกรุกแบบอโนมาลีจะสามารถตรวจจับการบุกรุกจากผู้บุกรุกที่ไม่มีในฐานข้อมูลการบุกรุกได้

ลักษณะของข้อมูลที่ใช้ในการทำแบบทดลอง

ข้อมูลที่นำมาใช้ในการทำแบบทดลอง เป็นข้อมูลที่ได้จากฐานข้อมูลความรู้ (Knowledge Discovery in Database (KDD) Cup Data) (Irine, 1999) ซึ่งเป็นชุดข้อมูลในปี 1999 ชุดข้อมูลนี้ถูกสร้างจากการจำลองการโจมตีของผู้บุกรุกจาก U.S. Air Force Local Area Network มีจำนวนประมาณ 4,900,000 จุดข้อมูล มี 41 คุณลักษณะ ภาพที่ 2-1 แสดงตัวอย่างข้อมูล KDDCup99 ซึ่งข้อมูลอยู่ในรูปแบบของสัญลักษณ์ และจำนวนจริง โดยคุณลักษณะสุดท้ายคือคลาสที่บ่งบอกว่าข้อมูลชุดใดเป็นลักษณะปกติหรือบุกรุก ซึ่งแบ่งออกเป็น 5 ประเภทใหญ่ ดังนี้

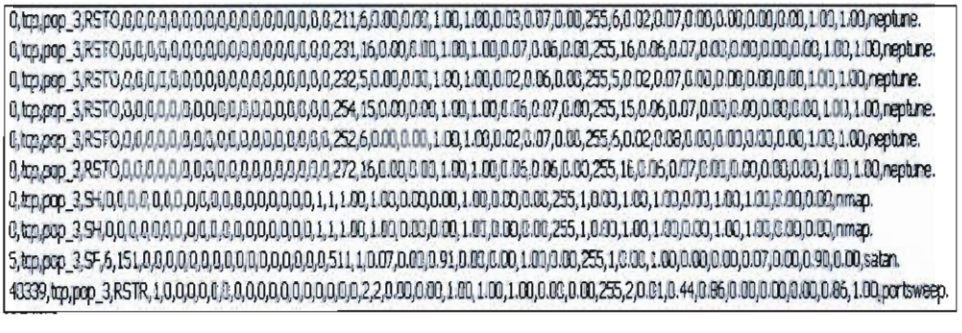
Normal คือ ข้อมูลมีลักษณะปกติหรือไม่มีการบุกรุก

Dos คือ ผู้บุกรุกพยายามโจมตีเพื่อทำให้เครื่องคอมพิวเตอร์ปลายทางหยุดทำงาน หรือสูญเสียเสถียรภาพ ซึ่งแบ่งออกเป็นประเภทย่อย ๆ อีก 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



ภาพที่ 2-1 ตัวอย่างข้อมูล KDDCup99

จากตารางที่ 2-1 แสดงชื่อของลักษณะของข้อมูล KDDCup99 จำนวน 41 ลักษณะ และในแต่ละชุดข้อมูลของ KDDCup99 นี้ จะแบ่งคุณลักษณะออกเป็น 3 กลุ่มคือ

Basic Features เป็นคุณลักษณะพื้นฐานที่ได้จากแพคเกจข้อมูลที่สื่อสารในเครือข่าย เช่น ชนิดของโปรโตคอล

Traffic Features เป็นคุณลักษณะที่แสดงถึงลักษณะของการสื่อสาร เช่น เวลาหรือจำนวนครั้งในการติดต่อ

Content Features เป็นคุณลักษณะที่บอกถึงลักษณะการบุกรุกหรือพฤติกรรมที่น่าสงสัย เช่น ความผิดพลาดในการล็อคอิน

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

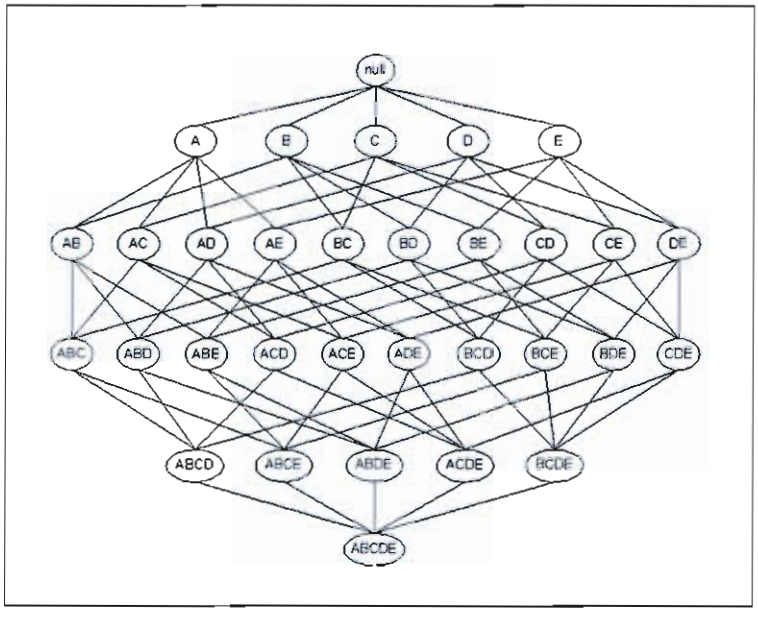
เนื่องจากข้อมูล KDDCup99 มีจำนวนมาก ดังนั้น ในงานวิจัยส่วนใหญ่จึงแนะนำให้เลือกข้อมูลเพียงร้อยละ 10 และเพื่อสะดวกในการสอนและทดสอบประสิทธิภาพของระบบการรู้จำจึงทำการสุ่มข้อมูลมาประมาณ 13,499 จุดข้อมูล (Patterns) โดยแบ่งเป็นประเภท Normal จำนวน 4,107 จุดข้อมูล Dos จำนวน 4,107 จุดข้อมูล Probe จำนวน 4,107 จุดข้อมูล R2L จำนวน 1,126 จุดข้อมูล และ U2R จำนวน 52 จุดข้อมูล และตัดบางคุณลักษณะที่ไม่มีผลต่อการรู้จำออกไป เช่น Basic Features และคุณลักษณะที่มีค่าเป็นศูนย์ทั้งหมด จึงเหลือจำนวนคุณลักษณะ 34 คุณลักษณะ

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 ซึ่งต่างจากการค้าหาข้อมูลแบบธรรมดาที่ต้องพิจารณาตรวจสอบข้อมูลทุกตัวจนครบ ทำให้ไม่เหมาะกับข้อมูลที่มีขนาดใหญ่ทำให้เสียเวลาได้ แต่ข้อเสียของการค้นหาคำตอบอาศัยวิธีการทางฮิวริสติกคือคำตอบที่ได้เป็นเพียงคำตอบที่ดี แต่ไม่รับรองว่าเป็นคำตอบที่ดีที่สุด สิ่งที่สำคัญในการแก้ปัญหาวิธีการทางฮิวริสติกว่าจะสามารถแก้ปัญหาได้ตามที่ต้องการหรือไม่ คือ ฮิวริสติกฟังก์ชัน ทำหน้าในการวัดความเป็นไปได้ของคำตอบ ซึ่งเป็นการกำกับทิศทางของกระบวนการค้นหา เพื่อให้อยู่ในทิศทางที่ได้ประโยชน์สูงสุด โดยพิจารณาจากน้ำหนักที่ให้กับการแก้ปัญหาของแต่ละวิธี น้ำหนักเหล่านี้จะถูกแสดงด้วยตัวเลขที่กำกับไว้กับโหนดต่าง ๆ และค่าเหล่านี้จะเป็นตัวที่ใช้ในการประมาณความเป็นไปได้ว่าเส้นทางที่ผ่านโหนดนั้นจะมีความเป็นไปได้ในการเข้าใกล้เป้าหมายมากน้อยเพียงใด ตัวอย่างของการค้นหาคำตอบอาศัยวิธีการทางฮิวริสติก เช่น การค้นหาแบบกรีดดี เป็นการค้นหาแบบดีที่สุดก่อน (Best First Search) ที่ง่ายที่สุด เป็นการนำข้อดีของการค้นหาตามแนวกว้าง และการค้นหาตามแนวลึกมารวมกัน โดยการค้นหาแบบดีที่สุดก่อน จะเลือกโหนดที่มีค่าดีสุดซึ่งอาศัยฮิวริสติกฟังก์ชัน ในการหาและหลักการของขั้นตอนวิธีกรีดดีเพื่อหาคำตอบที่เหมาะสมที่สุดในแต่ละสถานการณ์

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} แสดงได้ดังภาพที่ 2-2 ซึ่งมีความยาว Itemset จากระดับชั้น (Level) ที่ 1 (1-itemset) ถึงระดับชั้นที่ 5 (5-itemset)



ภาพที่ 2-2 Itemset lattice

ถ้าในชุดข้อมูลมีจำนวน Items เท่ากับ k items ดังนั้นจำนวน Itemsets ที่มีโอกาสเป็น Frequent Itemsets ทั้งหมดคำนวณได้จาก $2^k - 1$ (ไม่รวมเซตว่าง) ซึ่งจะเห็นได้ว่าเมื่อมีการใช้งานจริงในภาคธุรกิจ หรือลักษณะงานอื่น ๆ ค่าของ k จะสูงมาก ผลที่ตามมาคือการค้นหาและเปรียบเทียบรวมถึงการคำนวณจะมีจำนวนมากขึ้นเป็นทวีคูณ ในการหา Frequent Itemsets ขั้นตอนวิธีจะต้องแจง Itemsetes ทั้งหมดที่เป็นไปได้ตามโครงสร้างแลตทิซ เราเรียก Itemsets ชุดนี้ว่า Candidate Itemsets (K.P. Soman) เพื่อช่วยกำจัดหรือลดจำนวน Candidate Itemsets แทนที่จะแจกแจงทั้งหมดเหมือนที่แสดงด้วยโครงสร้างแลตทิซ เราจะใช้หลักการที่เรียกว่า Apriori ซึ่งมีหลักการดังนี้

ถ้า Itemset หนึ่ง ๆ เป็น Frequent แล้ว ทุก ๆ สับเซตของ Itemset นั้นจะต้องเป็น Frequent ด้วย

หลักการ Apriori เป็นหลักการที่นิยมใช้ในการหากฎความสัมพันธ์ (Association Rule) ซึ่งเป็นวิธีการที่ง่ายแต่มีประสิทธิภาพที่จะนำไปสู่การสร้าง Candidate itemset ที่น้อยลง โดยการใช้เซตที่มี

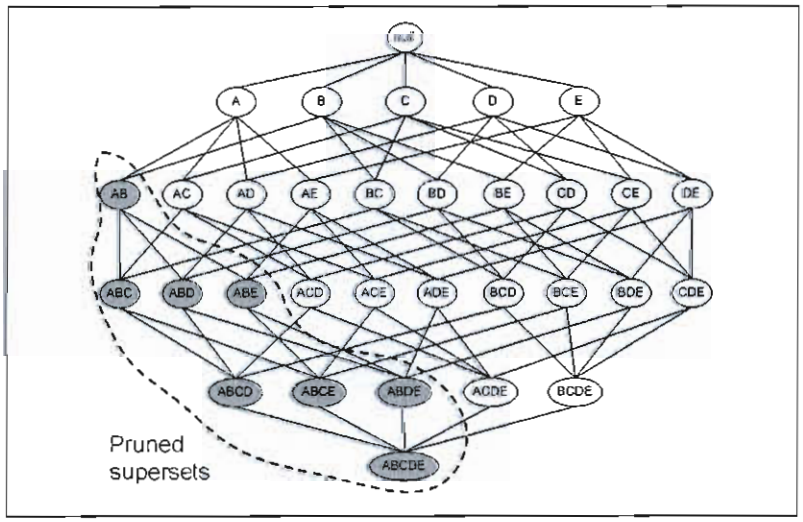
ขนาดใหญ่ที่หาได้ในขั้นตอนก่อนหน้า ขั้นตอน Apriori เป็นการทำงานที่ซ้ำ ๆ ในขั้นตอนแรกจะหาเซตที่มีขนาด 1-itemsets จากนั้นจะหาเซตที่มีขนาด 2-itemsets, 3-itemsets และต่อ ๆ ไป

ตัวอย่างจากข้อมูลรูปโครงสร้างแลตทิช ในภาพที่ 2-2 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 ดังนั้นทุก ๆ Superset ของ Itemset ดังกล่าวก็จะเป็น Infrequent itemset ด้วย ดังแสดงได้ด้วยโครงสร้างแลตทิชในภาพที่ 2-3 ซึ่งหลักการในการกำจัด Infrequent itemset นี้เรียกว่า support-based pruning กล่าวคือในการตรวจนับความถี่ของ Itemset ใด ๆ อยู่ แล้วพบว่าค่าสนับสนุนหรือค่าความถี่น้อยกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำ เราสามารถกำจัด Itemset นั้น ๆ ออกไป รวมถึงไม่จำเป็นต้องพิจารณาตรวจนับทุก ๆ Superset ของ Itemset นั้น ๆ ด้วยเช่นกัน



ภาพที่ 2-3 Itemset lattice กรณีกำจัด Infrequent Itemsets

ด้วยวิธีการเช่นนี้จะทำให้จำนวน Candidate Itemsets ลดขนาดลง คุณสมบัตินี้มีชื่อเรียกว่า Anti-Monotone ซึ่งมีนิยามดังนี้

$$\forall X, Y \in J : (X \subseteq Y) \rightarrow f(Y) \leq f(X) \qquad (2.1)$$

จาก (2.1) ซึ่งหมายความว่า ถ้า X เป็นสับเซตของ Y แล้ว $f(Y)$ จะต้องไม่มากกว่า $f(X)$

จากคุณสมบัติ Anti-Monotone ใน (2.1) เราสามารถนำมาประยุกต์ให้ f เป็นค่าสนับสนุน ก็จะสอดคล้องกับการกำจัด Infrequent Itemset ที่เรียกว่า Support-based Pruning ที่กล่าวแล้วข้างต้นนั่นเอง

การวิเคราะห์องค์ประกอบหลัก

วิธีการวิเคราะห์องค์ประกอบหลัก เป็นวิธีการทางสถิติ เพื่อใช้ในการสกัดปัจจัยที่อาศัยหลักความสัมพันธ์เชิงเส้นตรงระหว่างตัวแปรที่ใช้เป็นข้อมูล องค์ประกอบหลักตัวแปร (Jackson, 1991) คือ การการผสมเชิงเส้นตรง (Linear Combination) ของตัวแปรที่อธิบายการผันแปรของข้อมูลได้มากที่สุด จากนั้นหาการผสมเชิงเส้นครั้งที่สองที่สามารถอธิบายการผันแปรได้มากที่สุดเป็นอันดับที่สอง โดยที่ไม่สัมพันธ์กับการผสมครั้งแรก การวิเคราะห์องค์ประกอบหลักถูกนำไปประยุกต์ใช้งานต่าง ๆ เช่น การบีบอัดข้อมูล, การสร้างภาพใบหน้าไอเกนเพื่อใช้ในระบบจดจำ และการลบออกของพื้นหลังโดยใช้ไอเกน เป็นต้นวิธีการวิเคราะห์องค์ประกอบหลักสามารถนำมาใช้ในการลดมิติของข้อมูลโดย การวิเคราะห์ข้อมูลและเลือกเฉพาะข้อมูลที่มีความสำคัญเท่านั้น ส่วนข้อมูลที่ไม่สำคัญจะถูกตัดทิ้งไป ดังนั้นเมื่อข้อมูลผ่านกระบวนการ วิเคราะห์องค์ประกอบหลักแล้ว จะได้ผลลัพธ์เป็นไอเกนเวกเตอร์และค่าไอเกน ซึ่งไอเกนเวกเตอร์ที่มีค่าสมนัยกับค่าไอเกนที่มีค่าสูง ๆ จะเป็นการดึงข้อมูลที่มีความถี่ต่ำ ส่วนไอเกนเวกเตอร์ที่สมนัยกับค่าไอเกนที่ ต่ำ ๆ จะเป็นการดึงข้อมูลที่มีความถี่สูง

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 เป็นค่าคงที่ใด ๆ โดยที่

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

ที่ทำให้ $$Av = \lambda v \quad (2.2)$$

หรือ $$(A - \lambda I)v = 0 \quad (2.3)$$

เมื่อ A คือ ค่าเมทริกซ์

λ คือ เป็นค่าคงที่ใด ๆ เป็นสเกลาร์

v คือ ไอเกนเวกเตอร์

จากสมการจะเห็นว่า $v = 0$ ที่ทำให้สมการ เป็นจริงทุก ๆ ค่าของ λ

ตัวอย่างการคำนวณหาค่าไอเกนและไอเกนเวกเตอร์

กำหนดให้ A เป็นเมตริกซ์ขนาด 2x2

$$A = \begin{bmatrix} 2 & -4 \\ -1 & -1 \end{bmatrix}$$

จากสมการที่ (2.3) จะได้ว่า

$$\lambda I = \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

$$A - \lambda I = \begin{bmatrix} 2 & -4 \\ -1 & -1 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = \begin{bmatrix} 2-\lambda & -4 \\ -1 & -1-\lambda \end{bmatrix}$$

$$\det(A - \lambda I) = 0$$

$$\det(A - \lambda I) = (2-\lambda)(-1-\lambda) - (-1)(-4)$$

$$= (\lambda - 3)(\lambda + 2)$$

ซึ่งจะได้ $\lambda = 3, -2$

จาก $\begin{bmatrix} 2-\lambda & -4 \\ -1 & -1-\lambda \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

แทนค่า $\lambda = 3$

$$\begin{bmatrix} 2-3 & -4 \\ -1 & -1-3 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} -1 & -4 \\ -1 & -4 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} -4 \\ 1 \end{bmatrix}$$

แทนค่า $\lambda = -2$

$$\begin{bmatrix} 2-(-2) & -4 \\ -1 & -1-(-2) \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 4 & 4 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

ดังนั้นไอเกนเวกเตอร์ คือ $\begin{bmatrix} -4 & 1 \\ 1 & 1 \end{bmatrix}$ และเมื่อนำไอเกนเวกเตอร์ที่ได้นี้ไปคูณกับข้อมูลเดิม ซึ่งจะเป็นการหมุนแกนจึงทำให้ได้ข้อมูลใหม่

การเลือกลักษณะด้วยค่าสถิติไคสแควร์

การเลือกลักษณะแบบการคัดกรอง (Filter-Based Feature Selection) ส่วนใหญ่จะใช้เงื่อนไขทางสถิติหรือเงื่อนไขทางทฤษฎีสารสนเทศ (Information Theory) ในการคัดกรองลักษณะ วิธีการเลือกลักษณะแบบการคัดกรองที่ส่วนใหญ่นิยมใช้กัน เช่น Chi-Square Information Gain และ Relief Algorithm เป็นต้น ในงานวิทยานิพนธ์นี้ได้นำ Chi-Square Algorithm เป็นวิธีที่ใช้ในการเลือกลักษณะ โดยที่ค่าสถิติไคสแควร์ (Chi-Square) นี้จะวัดค่าความสัมพันธ์ระหว่างลักษณะกับคลาสคำตอบ เพื่อจัดลำดับลักษณะตามค่านัยสำคัญทางสถิติ ซึ่งสามารถคำนวณได้ตามสมการที่ (2.4)

$$\chi^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \tag{2.4}$$

เมื่อ O_{ij} คือ ความถี่ที่สังเกตได้

E_{ij} คือ ความถี่ที่คาดหวัง

โดยที่ E_{ij} คำนวณได้จากสมาการที่ (2.5)

$$E_{ij} = \frac{(R_{T_i})(C_{T_j})}{N} \tag{2.5}$$

เมื่อ R_{T_i} คือ ผลรวมของสมาชิกในแถว

C_{T_j} คือ ผลรวมของสมาชิกในคอลัมน์

N คือ จำนวนสมาชิกทั้งหมด

จากตารางที่ 2-2 เป็นตัวอย่างชุดข้อมูลฝึกที่ใช้ประกอบการตัดสินใจในการออกไปเล่นกอล์ฟ โดยพิจารณาจากสภาพอากาศต่าง ๆ ซึ่งมีลักษณะ 4 ลักษณะในประกอบการตัดสินใจ ได้แก่ Outlook, Temperature, Humidity และ Windy และคลาสที่เป็นคำตอบในการตัดสินใจมีค่าที่เป็นไปได้คือ Yes หรือ No

ตารางที่ 2-2 ตัวอย่างชุดข้อมูลฝึกที่ใช้ประกอบการตัดสินใจการออกไปเล่นกอล์ฟ

No.	Attributes				Class
	Outlook	Temperature	Humidity	Windy	
1	sunny	hot	high	false	No
2	sunny	hot	high	true	No
3	overcast	hot	high	false	Yes
4	rain	mild	high	false	Yes
5	rain	cool	normal	false	Yes
6	rain	cool	normal	true	No
7	overcast	cool	normal	true	Yes
8	sunny	mild	high	false	No
9	sunny	cool	normal	false	Yes
10	rain	mild	normal	false	Yes
11	sunny	mild	normal	true	Yes
12	overcast	mild	high	true	Yes
13	overcast	hot	normal	false	Yes
14	rain	mild	high	true	No

ตัวอย่างการคำนวณหาค่าสถิติไคสแควร์โดยใช้ข้อมูลจากตารางที่ 2-2

เมื่อพิจารณาลักษณะ Outlook

หาความถี่ที่สังเกตได้ (O)

จากลักษณะ Outlook มี 3 ค่า คือ sunny overcast และ rainy โดยที่แต่ละค่ามีความถี่ที่สังเกตได้ คือมีคำตอบ Yes จำนวน 2, 4 และ 3 ตามลำดับ และมีคำตอบ No จำนวน 3, 0 และ 2 ตามลำดับ สรุปได้ดังตารางที่ 2-3

ตารางที่ 2-3 ความถี่ที่สังเกตได้ของลักษณะ Outlook

	Yes	No
sunny	2	3
overcast	4	0
rainy	3	2

หาความถี่ที่คาดหวัง (E)

ตัวอย่างการหาความถี่ที่คาดหวังของลักษณะ Outlook เช่น หาความถี่คาดหวังของ sunny ที่มีคำตอบ Yes โดยผลรวมของจำนวนสมาชิกในแถวของ sunny เท่ากับ 5 ผลรวมของจำนวนสมาชิกในคอลัมน์ของคำตอบ Yes เท่ากับ 9 และผลรวมของจำนวนสมาชิกทั้งหมดเท่า 14 นำค่าต่าง ๆ เหล่านี้ไปแทนในสมการที่ (2.5) ได้ดังนี้

$$E(sunny, yes) = \frac{5 \times 9}{14} = 3.21$$

ดังนั้นความถี่ที่คาดหวังของลักษณะ Outlook สรุปดังตารางที่ 2-4

ตารางที่ 2-4 ความถี่ที่คาดหวังของลักษณะ Outlook

	Yes	No	Total
sunny	3.21	1.79	5
overcast	2.57	1.43	4
rainy	3.21	1.79	5
Total	9	5	14

คำนวณหาค่าสถิติไคสแควร์ของลักษณะ Outlook

โดยนำความถี่ที่สังเกตและความถี่ที่คาดหวังที่คำนวณได้ในก่อนหน้า แทนในสมการที่ (2.4) ดังนี้

$$\chi^2(outlook) = \frac{(2-3.21)^2}{3.21} + \frac{(4-2.57)^2}{2.57} + \frac{(3-3.21)^2}{3.21} + \frac{(3-1.79)^2}{1.79} + \frac{(0-1.43)^2}{1.43} + \frac{(2-1.79)^2}{1.79}$$

$$\chi^2(outlook) = 3.55$$

ดังนั้นค่าสถิติไคสแควร์ของลักษณะ Outlook เท่ากับ 3.55

โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน

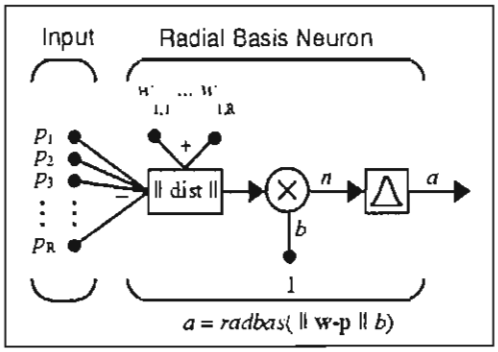
โดยแบบที่นิยมใช้โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน เป็นโครงข่ายประสาทเทียมป้อนไปข้างหน้าแบบหลายขั้น จะประกอบไปด้วย 3 ชั้น ได้แก่ ชั้นรับข้อมูลเข้า ขั้นซ่อน และชั้นข้อมูลออก (S.Chen, 1991) ดังภาพที่ 2-5 โดยเป็นฟังก์ชันการส่งระหว่างชั้นรับข้อมูลเข้า $p \in \mathbb{R}^{Nx1}$ ไปยังชั้นข้อมูลออก $y \in \mathbb{R}^{Mx1}$ จะได้ข้อมูลออกของเครือข่ายดังสมการที่ (2.6)

$$y_i = \sum_{k=i}^{s} w_{ik} \emptyset_k(\|p - c\|) \tag{2.6}$$

โดยที่ w_{ik} คือ ค่าน้ำหนักนิวรอนในชั้นซ่อน

S คือ จำนวนนิวรอนในชั้นซ่อน

C คือ เวกเตอร์จุดศูนย์กลาง



ภาพที่ 2-4 โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน
(ที่มา http://matlab.izmiran.ru/help/toolbox/nnet/radial74.html)

งานวิจัยที่เกี่ยวข้อง

Murat Karabatak และคณะ (2009) นำเสนองานวิจัยเรื่อง A New Feature Selection Method Based on Association Rules for Diagnosis of Erythemato-squamous Diseases ได้นำเสนองานวิธีการเลือกลักษณะบนพื้นฐานของกฎความสัมพันธ์ (Association Rules) และ โครงข่ายประสาทเทียม ถูกนำเสนอสำหรับการวินิจฉัยโรค Erythemato-Squamous ซึ่งคือโรคผิวหนังชนิดหนึ่ง โดยกฎความสัมพันธ์ใช้เพื่อลดจำนวนลักษณะของข้อมูล และโครงข่ายประสาทเทียมใช้สำหรับกระบวนการการจำแนกกลุ่ม และเปรียบประสิทธิภาพกับวิธีการเลือกลักษณะวิธีอื่น หลังจากใช้กฎความสัมพันธ์เลือกลักษณะสามารถลดจำนวนจาก 34 ลักษณะ เหลือ 24 ลักษณะ มีอัตราการจำแนกกลุ่มถูกต้อง 98.61% ซึ่งให้ค่าความถูกต้องมากกว่ากับข้อมูลที่ไม่ได้ผ่านการเลือกลักษณะและการเลือก

ลักษณะวิธีอื่น ๆ ผลการทดลองแสดงให้เห็นว่าการเลือกลักษณะมีความสำคัญ และทำให้การจำแนกกลุ่มข้อมูลเพื่อวินิจฉัยโรค Erythemato-Squamous ได้อย่างมีประสิทธิภาพ

Jianwen Xie และคณะ (2009) ได้นำเสนองานวิจัยเรื่อง Feature Selection Algorithm Based on Association Rules Mining Method โดยนำเสนอวิธีการเลือกลักษณะบนเทคนิคของกฎความสัมพันธ์ ซึ่งแนวคิดหลักของขั้นตอนวิธีที่นำเสนอคือการหาลักษณะที่มีความสัมพันธ์กันของลักษณะคลาสโดยใช้วิธีการหากฎความสัมพันธ์ ผลการทดลองกับหลาย ๆ ชุดข้อมูล และเปรียบเทียบกับวิธีการเลือกลักษณะด้วยวิธีอื่น ๆ แสดงให้เห็นว่าวิธีที่นำเสนอได้จำนวนลักษณะที่น้อยกว่าและผลลัพธ์เป็นที่น่าพอใจ

Mansour Sheikhan และคณะ (2009) นำเสนองานวิจัยเรื่อง Misuse Detection Using Hybrid of Association Rule Mining and Connectionist Modeling โดยวิธีการที่เสนอเป็นการรวมการจำแนกกลุ่มเพอร์เซ็ปตรอนแบบหลายชั้นกับกฎความสัมพันธ์ เพื่อจำแนกกลุ่มข้อมูลตรวจจับการบุกรุก KDDCup99 จำนวน 5 คลาส โดยเปรียบเทียบกับการจำแนกกลุ่มเพอร์เซ็ปตรอนแบบหลายชั้น ซึ่งผลการทดลอง การจำแนกกลุ่มด้วยเพอร์เซ็ปตรอนแบบหลายชั้นสามารถจำแนกกลุ่มได้ดีกับคลาส DoS และ Probe แต่ให้ผลไม่ดีกับคลาส R2L และ U2R แต่วิธีการที่นำเสนอสามารถจำแนกกลุ่มได้ดีกับทุกคลาส และได้ผลอัตราการตรวจจับที่ดีกว่า แม้ว่าค่าความผิดพลาดเชิงบวกจากวิธีการที่นำเสนอจะมากกว่าการจำแนกกลุ่มด้วยเพอร์เซ็ปตรอนแบบหลายชั้น แต่ยังได้ผลที่ดีกว่าการจัดกลุ่มโดยการหาค่าเฉลี่ยแบบเคและโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน

M. Anbarasi และคณะ (2010) นำเสนองานวิจัยเรื่อง Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm โดยที่จุดประสงค์ของงานวิจัยนี้คือการเพิ่มประสิทธิภาพการทำนายการวินิจฉัยโรคหัวใจด้วยการลดจำนวนจำนวนลักษณะ ซึ่งใช้ขั้นตอนวิธีเชิงพันธุกรรมในการลดจำนวนลักษณะของข้อมูลโรคหัวใจ จากจำนวนลักษณะเดิม 13 ลักษณะ เหลือ 6 ลักษณะ และเปรียบเทียบการทำนายด้วยสามวิธีคือ เนอีฟเบย์ ต้นไม้ตัดสินใจ และการจัดกลุ่ม ผลลัพธ์ที่ได้คือ ต้นไม้ตัดสินใจสามารถทำนายได้ถูกต้อง และค่าความผิดพลาดสัมบูรณ์เฉลี่ยน้อยกว่าทั้งสองวิธี แต่ใช้เวลามากกว่า

Onur Inan และคณะ (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 และเมื่อ

เปรียบกับวิธีการอื่น ๆ ปรากฏว่าวิธีการที่นำเสนอส่วนใหญ่ให้ค่าความถูกต้องในการวินิจฉัยโรคมะเร็งหน้าอกได้มากกว่าวิธีการอื่น ๆ

Jing Zhang และ คณะ (2003) ได้นำเสนองานวิจัยเรื่อง A New Heuristic Reduct Algorithm Base on Rough Sets Theory เนื่องจากการนำทฤษฎีเซตอย่างหยาบ เป็นวิธีทางคณิตศาสตร์ในการวิเคราะห์ขอมูลที่มีความคลุมเครือ และความไมแน่นอนจากทุกวัตถุใน ลดคุณสมบัติที่ไมจำเป็นหาค่าความสำคัญของคุณสมบัติ เพื่อหาซับเซตของลักษณะที่เหมาะสมที่สุดจากการเลือกลักษณะด้วยวิธีนี้ใช้เวลานาน จึงนำเสนอขั้นตอนวิธีฮิวริสติกบนพื้นฐานของทฤษฎีเซตอย่างหยาบเพื่อหาซับเซตของลักษณะที่เหมาะสมและใช้เวลาน้อย ผลการทดลองกับชุดข้อมูลหลาย ๆ ชุดแสดงให้เห็นว่าส่วนใหญ่วิธีการที่นำเสนอสามารถหาซับเซตของลักษณะได้เหมาะสมที่สุดได้อย่างรวดเร็วและมีประสิทธิภาพ

Khazaee Saeed และคณะ (2011) นำเสนองานวิจัยเรื่อง A Hybrid Model Based on Feature Extraction for Network Intrusion Detection งานวิจัยนี้มุ่งหมายเพื่อศึกษาความเป็นไปได้ของการประยุกต์วิธีการสกัดลักษณะเพื่องานการตรวจจับการบุกรุกแบบมิสยูส โดยขั้นตอนของการสกัดลักษณะจะใช้การจำแนกกลุ่มหลาย ๆ ตัว เพื่อเพิ่มลักษณะใหม่ที่ทำให้มีผลกระทบในเชิงบวก และเป็นอิสระจากชนิดของการจำแนกกลุ่มด้วย ทดสอบกับชุดข้อมูลบุกรุกเครือข่าย KDDCup99 จำนวน 11 คลาส เพื่อเพิ่มความแม่นยำในการรู้จำ ผลการทดลองแสดงให้เห็นว่าการตรวจจับการบุรุกด้วยวิธีการสกัดคุณลักษณะ มีประสิทธิภาพดีกว่าไม่มีขั้นตอนการสกัดคุณลักษณะ ทั้งในด้านของ อัตราการจำแนกกลุ่ม อัตราการตรวจจับผู้บุกรุก และค่าความผิดพลาดเชิงบวกเมื่อเปรียบเทียบกับวิธีการอื่น ๆ

Adel Jahanbani และคณะ (2012) นำเสนองานวิจัยเรื่อง A New Approach for Detecting Intrusions Based on the PCA Neural Networks โดยวิธีการที่นำเสนอคือระบบการจำแนกข้อมูลระบบใหม่ โดยใช้โครงข่ายประสาทเทียมการวิเคราะห์องค์ประกอบหลักสำหรับการตรวจจับการบุกรุกเพื่อตรวจจับการบุกรุกให้มีอัตราการตรวจจับและความผิดพลาดเชิงบวกเป็นที่น่าพอใจ จากการทดลองกับขุดข้อมูลการตรวจจับการบุกรุก KDDCup99 เปรียบเทียบกับวิธีการอื่น ๆ ได้ว่าวิธีการที่นำเสนอมีประสิทธิภาพที่ดีกว่าทั้งในด้านอัตราการตรวจจับ และความผิดพลาดเชิงบวก

Shailendra Singh และคณะ (2011) นำเสนองานวิจัยเรื่อง An Efficient Feature Reduction Technique for Intrusion Detection System ได้นำเสนอวิธีใหม่สำหรับการลดลักษณะซึ่งใช้ในด้านของการตรวจจับการบุกรุกเครือข่าย ซึ่งข้อมูลมีขนาดใหญ่และมีลักษณะจำนวนมาก และจากที่การวิเคราะห์องค์ประกอบหลักมีข้อจำกัดไม่เหมาะสมกับชุดข้อมูลที่มีลักษณะที่เส้นตรงไม่สามารถแบ่งกลุ่มได้ จึงได้นำเสนอวิธีการใหม่ที่มีประสิทธิภาพอยู่บนพื้นฐานของ Generalized Discriminant Analysis (GDA) โดยการเลือกลักษณะที่มี Discriminant มากที่สุด เพื่อเปรียบเทียบประสิทธิภาพจึงเปรียบกับการสกัดลักษณะด้วยการวิเคราะห์องค์ประกอบหลัก และจำแนกกลุ่มด้วยวิธีแผนผังการจัดระเบียบตัวเอง และต้นไม้ตัดสินใจ C4.5 ผลการทดลองแสดงให้เห็นว่าวิธีการที่นำเสนอให้ผลได้ดีว่าการ

วิเคราะห์องค์ประกอบหลัก ทั้งในด้านของอัตราการตรวจจับ ความผิดพลาดเชิงบวก และเวลาที่ใช้ในการเรียนรู้และทดสอบ

Iftikhar Ahmad และคณะ (2011) นำเสนองานวิจัยเรื่อง Optimized Intrusion Detection Mechanism using Soft Computing Techniques นำเสนอวิธีการการตรวจจับการบุกรุก กับชุดข้อมูล KDDCup99 โดยใช้การวิเคราะห์องค์ประกอบหลักเพื่อให้ได้ข้อมูลชุดใหม่และใช้ขั้นตอนวิธีเชิงพันธุกรรมในการเลือกลักษณะที่เหมาะสม จำแนกกลุ่มด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ประสิทธิภาพของวิธีการที่นำเสนอนี้จะนำไปเปรียบเทียบกับอื่น ๆ จากการทดลองสรุปได้ว่าวิธีการที่นำเสนอนี้เป็นวิธีที่ดีที่สุดในการตรวจจับการบุกรุกเมื่อเปรียบเทียบกับวิธีการอื่น ๆ ซึ่งสามารถลดจำนวนของลักษณะ เพิ่มอัตราการตรวจจับ และลดความผิดพลาดเชิงบวกในด้านของการตรวจจับการบุกรุกด้วย

บทที่ 3
วิธีการดำเนินงาน

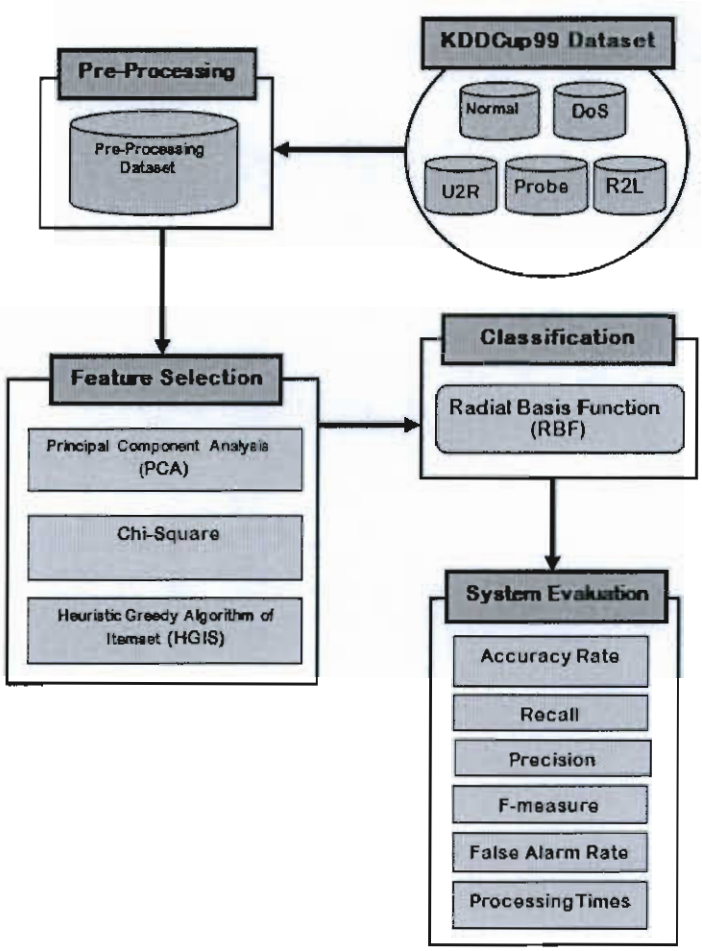
ในบทนี้จะกล่าวถึง ขั้นตอนวิธีการดำเนินงานการตรวจจับการบุกรุก ซึ่งมีหัวข้อต่อไปนี้

1. ศึกษาความเป็นไปได้
2. ขั้นตอนการดำเนินงาน
3. ขั้นตอนเตรียมข้อมูลนำเข้า
4. ขั้นตอนการเลือกลักษณะด้วยขั้นตอนฮิวริสติกกรีดดีโดยใช้หลักการ Apriori
   4.1 ขั้นตอนการออกแบบฟังก์ชันผิดพลาด
5. ขั้นตอนการสกัดลักษณะด้วยวิธีวิเคราะห์องค์ประกอบหลัก
6. ขั้นตอนการเลือกลักษณะด้วยค่าสถิติไคสแควร์
7. ขั้นตอนการจำแนกกลุ่มด้วยโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน
8. ขั้นตอนการวัดประสิทธิภาพ

ศึกษาความเป็นไปได้

วิทยานิพนธ์นี้มีวัตถุประสงค์เพื่อศึกษาการลดจำนวนของข้อมูลการบุกรุก เพื่อสามารถตรวจจับการบุกรุกได้อย่างรวดเร็วและมีประสิทธิภาพ มีความถูกต้องที่สูง ซึ่งจากการศึกษาทฤษฎี และงานวิจัย สรุปได้ว่าการเลือกลักษณะด้วยขั้นตอนวิธีฮิวริสติกกรีดดี หรือวิธีการเลือกลักษณะด้วยวิธีอื่น ๆ และการสกัดลักษณะเด่น มีความสำคัญเพื่อกำจัดความซับซ้อนของข้อมูล สามารถลดจำนวนลักษณะให้ได้ลักษณะที่มีความสำคัญต่อการตรวจจับการบุกรุก โดยจะทดสอบการตรวจจับการบุกด้วยวิธีการจำแนกกลุ่มโดยโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐานกับชุดข้อมูล KDDCup99 (Irvine, 1999) ซึ่งเป็นชุดข้อมูลการบุกรุกเครือข่ายที่มาตรฐาน เพื่อวิเคราะห์เปรียบเทียบผลลัพธ์ค่าความถูกต้อง จึงเห็นว่าวิทยานิพนธ์นี้มีความเป็นไปได้ที่จะบรรลุจุดประสงค์

ขั้นตอนการดำเนินงาน



ภาพที่ 3-1 ขั้นตอนการดำเนินงาน

นำข้อมูล KDDCup99 ผ่านกระการเตรียมข้อมูลโดยตัดคุณลักษณะบางคุณลักษณะที่ไม่จำเป็นออกไป จากนั้นนำข้อมูลที่ได้ผ่านกระบวนการเลือกลักษณะด้วยขั้นตอนวิธีฮิวริสติกกรีดดี เปรียบเทียบกับการสกัดลักษณะด้วยการวิเคราะห์องค์ประกอบหลัก และการเลือกลักษณะด้วยค่าสถิติไคสแควร์ และทดสอบประสิทธิภาพการตรวจจับการบุกรุกด้วยการจำแนกกลุ่มวิธีโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมี โดยวัดประสิทธิภาพด้วยวิธีการต่าง ๆ แสดงขั้นตอนการดำเนินงานได้ดังภาพที่ 3-1

ขั้นตอนเตรียมข้อมูลนำเข้า

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 41 คุณลักษณะ ซึ่งข้อมูลอยู่ในรูปแบบของสัญลักษณ์ และจำนวนจริง โดยคุณลักษณะสุดท้าย คือ Class ที่บ่งบอกว่าข้อมูลชุดใดเป็นลักษณะปกติหรือบุกรุก มีลักษณะของข้อมูลที่สมบูรณ์ ซึ่งแบ่งออกเป็น 5 ประเภทใหญ่ คือ Normal Dos Probe R2L และ U2R โดยที่แต่ละประเภทมีจำนวนไม่เท่ากัน

เนื่องจากข้อมูล KDDCup99 มีจำนวนมาก ดังนั้น ในงานวิจัยส่วนใหญ่จึงแนะนำให้เลือกข้อมูลเพียงร้อยละ 10 และเพื่อสะดวกในการสอนและทดลอบประสิทธิภาพของระบบการรู้จำจึงทำการสุ่มข้อมูลมาประมาณ 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 และตัดบางคุณลักษณะที่ไม่มีผลต่อการรู้จำออกไป เช่น Basic Features และคุณลักษณะที่มีค่าเป็นศูนย์ทั้งหมด จึงเหลือจำนวนคุณลักษณะ 34 คุณลักษณะ สำหรับใช้ในการทดลองขั้นต่อไป

ขั้นตอนการเลือกลักษณะด้วยวิธีการทางฮิวริสติกกรีดดี

จากข้อมูล KDDCup99 ที่ใช้ในการทดลองหลังจากผ่านกระบวนการเตรียมข้อมูลแล้วจะเหลือลักษณะทั้งหมด 34 ลักษณะ ดังนั้น Items จึงมีทั้งหมด 34 items คือ
{1,2,3,4,5,6,7,8,9,10,11,12,13,14,1,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34}

การเลือกชุดลักษณะ (Itemset) โดยวิธีฮิวริสติกกรีดดีของไอเท็มเซตโดยอไพรออริอัลกอริทึมจะแบ่งออกเป็น 2 ขั้นตอน คือ การหาชุดลักษณะที่เป็นฐาน และการเพิ่มลักษณะที่หลุดออกเข้าไปเพื่อเพิ่มประสิทธิภาพ ซึ่งมีวิธีการดังนี้

1. ขั้นตอนการหาชุดลักษณะฐาน

ขั้นตอนที่ 1.1: สร้าง 1-itemset โดยนำแต่ละลักษณะหาค่าผิดพลาดจากฟังก์ชันผิดพลาดโดยใช้โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน

ขั้นตอนที่ 1.2: สร้าง 2-candidate itemset โดยการนำแต่ละลักษณะมาจับคู่กันทุกๆ ลักษณะที่เป็นไปได้ และค่าผิดพลาดจากฟังก์ชันผิดพลาด โดยใช้โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน

ขั้นตอนที่ 1.3: สร้าง 2-itemset โดยนำ 2-candidate itemset ที่มีค่าผิดพลาดน้อยกว่าหรือเท่ากับซับเซตขนาด 1-itemset ของ 2-candidate itemset ดังสมการที่ (3.1) เมื่อ {a,b} คือชุดลักษณะ ค่าผิดพลาดของลักษณะ {a,b} ต้องน้อยกว่าหรือเท่ากับค่าผิดพลาดของลักษณะ {a} และ {b}

$$f(\{a,b\}) \leq f(\{a\}) \text{ และ } f(\{a,b\}) \leq f(\{b\}) \quad (3.1)$$

ขั้นตอนที่ 1.4: ทำซ้ำในขั้นตอนที่ 1.2 โดยเพิ่มขนาดของ itemset ไปเรื่อย ๆ จนกระทั่งไม่สามารถสร้าง itemset ได้อีก

ขั้นตอนที่ 1.5: เลือก itemset จาก itemsets ทั้งหมดที่มีค่าผิดพลาดต่ำที่สุดเป็นลักษณะฐานเพื่อใช้เป็นฐานในการเพิ่มลักษณะในขั้นต่อไป

2. ขั้นตอนการเพิ่มลักษณะ

ขั้นตอนที่ 2.1: นำแต่ละลักษณะที่เหลือมาเพิ่มให้กับลักษณะฐาน และหาค่าผิดพลาดจากฟังก์ชันผิดพลาด โดยใช้โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน

ขั้นตอนที่ 2.2: เลือกชุดลักษณะที่มีค่าความผิดพลาดต่ำที่สุดมาเป็นฐานต่อไป และนำแต่ละลักษณะที่เพิ่มเข้าไปในขั้นตอนที่ 2.1 ที่มีค่าความผิดพลาดต่ำกว่าลักษณะฐานของขั้นตอนที่ 2.1 มาเพิ่มให้กับลักษณะฐานใหม่ที่ได้ และหาค่าผิดพลาดจากฟังก์ชันผิดพลาด โดยใช้โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน

ขั้นตอนที่ 2.3: ทำขั้นตอนที่ 2.1 และ 2.2 จนกระทั่งไม่สามารถเพิ่มลักษณะได้อีก และเลือกชุดลักษณะที่มีค่าผิดพลาดต่ำที่สุด

โดยที่ฟังก์ชันผิดพลาดที่ใช้เป็นเกณฑ์การเลือกลักษณะนั้นจะทำการทดลองเพื่อหาฟังก์ชันที่เหมาะสมสำหรับการเลือกลักษณะโดยวิธีฮิวริสติกกรีดดี

ขั้นตอนการออกแบบฟังก์ชันผิดพลาดสำหรับการเลือกลักษณะโดยวิธีฮิวริสติกกรีดดีอัลกอริทึม

ฟังก์ชันที่ใช้คำนวณหาค่าผิดพลาดของลักษณะหรือชุดลักษณะในขั้นตอนของการเลือกลักษณะโดยใช้วิธีฮิวริสติกกรีดดีอัลกอริทึม ซึ่งในวิทยานิพนธ์นี้จะทำการทดลองกับ 4 ฟังก์ชัน ได้แก่ HGIS1 HGIS2 HGIS3 และ HGIS4 โดยมีรายละเอียดดังต่อไปนี้

เมื่อกำหนดให้ $rmse$ คือ ค่าความผิดพลาดเฉลี่ยกำลังสอง (Root Mean Square Error) ตามสมการที่ (3.2) และ $Nmiss$ คือจำนวนที่ตอบผิด ที่ได้จากการนำชุดลักษณะเข้ากระบวนการรู้จำด้วยโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน

$$rmse = \sqrt{\frac{\sum_{i=1}^{n}(d_i - o_i)^2}{n}} \tag{3.2}$$

เมื่อ d_i คือ เป็นค่าเอาต์พุตเป้าหมายตัวอย่างที่ i

o_i คือ เอาต์พุตที่ได้ตัวอย่างที่ i

n คือ จำนวนตัวอย่างทั้งหมด

1. HGIS1

ฟังก์ชันผิดพลาด HGIS1 คือค่าความผิดพลาดเฉลี่ยกำลังสอง ซึ่งเป็นการวัดค่าความผิดพลาดที่วัดว่าคำตอบที่ได้จากการนำชุดลักษณะผ่านกระบวนการรู้จำแล้ว ได้คำตอบที่ผิดหรือมีความใกล้เคียงกับคำตอบที่ถูกต้องมากน้อยเพียงใด ดังนี้

$$F(\{itemset\}) = rmse \tag{3.3}$$

2. HGIS2

ฟังก์ชันผิดพลาด HGIS2 คือจำนวนที่ตอบผิดเมื่อนำชุดลักษณะผ่านกระบวนการรู้จำ ดังนี้

$$F(\{itemset\}) = Nmiss \tag{3.4}$$

3. HGIS3

ฟังก์ชันผิดพลาด HGIS3 คือผลรวมของค่าความผิดพลาดเฉลี่ยกำลังสองและจำนวนที่ตอบผิด ซึ่งเป็นการนำ HGIS1 และ HGIS2 มารวมกัน เพื่อเพิ่มความแม่นยำของค่าความผิดพลาด และสำหรับกรณีที่ชุดลักษณะมากกว่าสองชุดขึ้นไปมี $rmse$ หรือ $Nmiss$ เท่ากัน ซึ่งค่าความผิดพาดที่ได้จากฟังก์ชัน HGIS3 นี้จะมีความแตกต่างกันเพียงเล็กน้อย ดังนี้

$$F(\{itemset\}) = rmse + Nmiss \tag{3.5}$$

4. HGIS4

ฟังก์ชันผิดพลาด HGIS3 คือผลคูณของค่าความผิดพลาดเฉลี่ยกำลังสองและจำนวนที่ตอบผิด ซึ่งเป็นการนำ HGIS1 และ HGIS2 มาคูณกัน เพื่อเพิ่มความแม่นยำของค่าความผิดพลาด ซึ่งค่าความผิดพาดของแต่ละชุดลักษณะจะมีความแตกต่างกันมาก ๆ ดังนี้

$$F(\{itemset\}) = rmse * Nmiss \tag{3.5}$$

โดยจะนำฟังก์ชันผิดพลาดทั้ง 4 สมการนี้ทดสอบหาฟังก์ชันที่เหมาะสมสำหรับการเลือกลักษณะด้วยวิธีฮิวริสติกกรีดดี

ขั้นตอนการวิเคราะห์องค์ประกอบหลัก

ให้ $x_1, x_2, \ldots x_M$ 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 $A = [\varphi_1, \varphi_2, \ldots \varphi_M]$ มีขนาด $N \times M$ แล้วคำนวณหาความแปรปรวนร่วม

$$C = \frac{1}{M}\sum_{N=1}^{M} \varphi_N \varphi_N$$

4. คำนวณค่าไอเกน $C: \lambda_1 > \lambda_2 > .. > \lambda_N$

5. คำนวณไอเกนเวกเตอร์ $C: u_1, u_2, \ldots u_N$

6. นำไอเกนเวกเตอร์คูณกับข้อมูลเดิม ซึ่งจะได้ข้อมูลใหม่ $xPca = u \times x'$

7. เลือกองค์ประกอบหลัก K

$$\frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{N} \lambda_i} > threshold$$

λ_i คือ ค่าไอเกนลำดับที่ i

N คือ จำนวนลักษณะทั้งหมด

K คือ จำนวนลักษณะที่ถูกเลือก

$threshold$ คือ ค่าเกณฑ์ที่บ่งบอกว่าต้องการให้องค์ประกอบหลักที่ได้มีค่าไอเกนสะสมใกล้เคียงกับค่าไอเกนสะสมทั้งหมดมากน้อยเพียงใด ในที่นี้กำหนดให้ $threshold$ เท่ากับ 0.95

ขั้นตอนการเลือกลักษณะด้วยค่าสถิติไคสแควร์

ค่าสถิติไคสแควร์ใช้สำหรับประเมินค่าของลักษณะ ซึ่งวัดค่าความสัมพันธ์ระหว่างลักษณะกับคลาสเพื่อจัดลำดับลักษณะตามค่านัยสำคัญทางสถิติ โดยจะเรียงลำดับค่าที่ได้จากมากไปน้อย ขั้นตอนการเลือกลักษณะโดยใช้ค่าสถิติไคสแควร์มีดังนี้

ขั้นตอนที่ 1: คำนวณหาค่าไคสแควร์ของแต่ละลักษณะกับคลาสคำตอบตามสมการที่ (2.4)

ขั้นตอนที่ 2: เรียงลำดับค่าไคสแควร์ที่คำนวณได้จากขั้นตอนที่ 1 จากมากไปน้อย

ขั้นตอนที่ 3: ตัดลักษณะที่มีไคสแควร์เท่ากับศูนย์ออกไป เนื่องจากไม่มีความสัมพันธ์ใด ๆ กับคลาสคำตอบ

ขั้นตอนที่ 4: ตัดลักษณะที่มีความสัมพันธ์กับคลาสคำตอบที่ต่ำออกไป ตามเงื่อนไขต่อไปนี้

$$\frac{\chi_i^2 \times \log(N^2)}{\sum \chi_i^2 \times N} \times 100 \ < \delta \qquad (3.6)$$

โดยกำหนดให้ δ = 0.1 คือค่าเกณฑ์ที่ใช้ตัดลักษณะที่มีความสัมพันธ์ต่ำ (Ranjit Abraham, 1993)

χ_i^2 คือ ค่าสถิติไคสแควร์ของลักษณะที่พิจารณา
N คือ จำนวนลักษณะทั้งหมด

ขั้นตอนการจำแนกกลุ่ม

ทดสอบประสิทธิภาพการตรวจจับการบุกรุกด้วยการจำแนกกลุ่มแบบโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน โดยแบ่งข้อมูลสำหรับเรียนรู้ร้อยละ 60 ข้อมูลสำหรับทดสอบร้อยละ 40 และกำหนดจำนวนรอบในกานเรียนรู้ 500 รอบ

ขั้นตอนการวัดประสิทธิภาพ

ในการวัดประสิทธิภาพของการตรวจจับการบุกรุกโดยการจำแนกกลุ่ม วิธีที่นิยมมากวิธีหนึ่งคือการวัดโดยใช้อัตราความถูกต้อง (Accuracy) ซึ่งบ่งบอกถึงความถูกต้องในการจำแนกกลุ่มโดยรวมทั้งหมด แต่สำหรับข้อมูล KDD Cup 1999 เป็นข้อมูลแบบไม่สมดุล หากวัดค่าความถูกต้องอาจจะเอนเอียงไปยังคลาสที่มีสมาชิกมาก จึงใช้วิธีการวัดประสิทธิภาพอื่นที่เหมาะสมกับข้อมูลแบบไม่สมดุลร่วมด้วย ซึ่งได้แก่ อัตราค่าความถูกต้อง (Accuracy Rate) ค่าความครบถ้วน (Recall) ค่าความแม่นยำ (Precision) ค่าเอฟเมเซอร์ (F-measure) และ อัตราความผิดพลาดเชิงบวก (False Alarm Rate: FAR) ตามสมการที่ (3.7) (3.8) (3.9) (3.10) และ (3.11) ตามลำดับ และนอกจากนี้ยังวัดประสิทธิภาพเวลาที่ใช้ในการทดสอบด้วย

การวัดประสิทธิภาพจากการจำแนกกลุ่มที่ถูกต้องโดยรวมได้จากสมการ (3.7)

$$\text{Accuracy } Rate = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \times 100\% \qquad (3.7)$$

การวัดค่าความครบถ้วนซึ่งเป็นค่าที่บ่งบอกถึงอัตราผลลัพธ์ที่ถูกต้องของคลาสนั้น ดังสมการที่ (3.8)

$$Recall = \frac{TP}{TP+FN} \times 100\% \qquad (3.8)$$

การวัดความแม่นยำในการจำแนกกลุ่มวัดได้จากสมการที่ (3.9)

$$Precision = \frac{TP}{TP+FP} \times 100\% \qquad (3.9)$$

การวัดค่าเอฟเมเซอร์คือการวัดประสิทธิภาพพื้นฐานในการจำแนกกลุ่ม ซึ่งเกิดจากการรวมค่าความครบถ้วนและค่าความแม่นยำมาคำนวณ ดังสมการที่ (3.10)

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \times 100\% \qquad (3.10)$$

อัตราความผิดพลาดเชิงบวก ซึ่งไม่ควรเกิดขึ้นหรือควรพยายามให้เกิดขึ้นน้อย วัดได้จากสมการที่ (3.11)

$$FAR = \frac{FP}{FP+TN} \times 100\% \tag{3.11}$$

โดยที่ ตัวแปรต่าง ๆ สรุปตามตารางที่ 3-1

ตารางที่ 3-1 Confusion matrix

		Predicted	
		True	False
Actual	True	True negative (TN)	False positive (FP)
	False	False negative (FN)	True positive (TP)

บทที่ 4
ผลการทดลอง

ในบทนี้จะกล่าวถึงผลการทดลอง โดยผลการทดลองที่นำเสนอจะประกอบด้วย ผลการทดสอบฟังก์ชันผิดพลาดสำหรับการเลือกลักษณะด้วยฮิวริสติกกรีด การเลือกลักษณะด้วยวิธีฮิวริสติกกรีดดี การเลือกลักษณะด้วยค่าสถิติไคสแควร์ และการสกัดลักษณะด้วยวิธีวิเคราะห์องค์ประกอบหลัก เมื่อทดสอบการจำแนกกลุ่มด้วยโครงข่ายประสาทเทียมฟังก์ชันรัศมีกับชุดข้อมูล KDDCup99 และนอกจากนี้เพื่อให้ผลการทดลองที่ได้มีความน่าเชื่อถือมากยิ่งขึ้นจึงนำวิธีที่ได้กล่าวมาทดสอบกับข้อมูลชุดอื่น ๆ ด้วย

ลักษณะข้อมูลของ KDDcup99 จำนวน 34 ลักษณะ

ชุดข้อมูล KDDCup99 เป็นฐานข้อมูลมาตรฐานที่นำมาใช้เป็นตัวอย่างในการทดลองในงานวิจัยนี้ โดยข้อมูลชุดนี้มีจำนวนประมาณ 4,900,000 จุดข้อมูล เป็นข้อมูล 41 ลักษณะ มีจำนวน 5 คลาส ได้แก่ Dos Probe U2R R2L และ Normal เนื่องจากข้อมูล KDDcup99 มีจำนวนมาก ดังนั้นในงานวิจัยส่วนใหญ่จึงแนะนำให้เลือกข้อมูลเพียงร้อยละ 10 สำหรับเป็นข้อมูลในการสอนและทดสอบประสิทธิภาพของระบบการรู้จำ และจากข้อมูลที่เลือกมาร้อยละ 10 นั้น จะทำการสุ่มมาทำข้อมูลในการสอนประมาณ 13,499 ชุด (Patterns) โดยแบ่งออกเป็น 5 คลาส ดังตารางที่ 4-1 ซึ่งมีลักษณะจำนวน 34 ลักษณะสำหรับการรู้จำด้วยโครงข่ายประสาทเทียม (จากที่กล่าวไว้ก่อนหน้านี้ว่าข้อมูล KDDcup99 เป็นข้อมูลทั้งหมด 41 ลักษณะ แต่ลักษณะที่เป็น Basic Features ลักษณะที่มีค่าเป็นศูนย์ทั้งหมด และลักษณะที่เป็นคำตอบจะไม่นำมาพิจารณา ดังนั้น จึงเหลือเพียง 34 ลักษณะ) ซึ่งในแต่ละลักษณะมีค่าสูงสุด ค่าต่ำสุด ค่าเฉลี่ย และค่าส่วนเบี่ยงเบนมาตรฐานตามตารางที่ 4-2 ในตารางที่ 4-3 จะแสดงค่าสหสัมพันธ์ระหว่างแต่ละลักษณะทั้ง 34 ลักษณะ และภาพที่ 4-1 ถึง 4-34 แสดงให้เห็นถึงการกระจายตัวของข้อมูลในแต่ละลักษณะ ในลักษณะที่ 1 ถึง ลักษณะที่ 34

ตารางที่ 4-1 จำนวนข้อมูล KDDCup99 ในแต่ละคลาส

ชื่อคลาส	จำนวนจุดข้อมูล
Normal	4,107
DoS	4,107
U2R	4,107
R2L	1,126
Probe	52
รวม	13,499

ตารางที่ 4-2 ค่าทางสถิติของข้อมูล KDDcup99 จำนวน 34 ลักษณะ

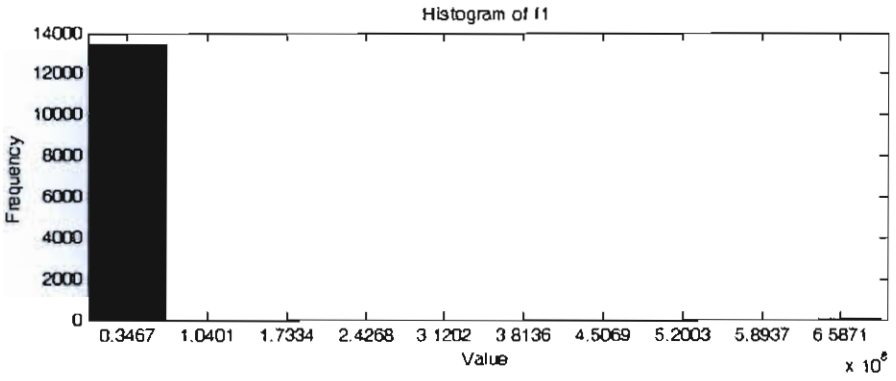
Features	Maximum	Minimum	Mean	Standard Deviation
f1	6.93E+08	0	74547.73	5977289
f2	5155468	0	7012.817	172422.5
f3	3	0	0.001852	0.073517
f4	2	0	0.000296	0.021081
f5	30	0	0.640121	4.039457
f6	5	0	0.004223	0.07541
f7	1	0	0.301207	0.4588
f8	38	0	0.012742	0.426195
f9	1	0	0.002371	0.048632
f10	1	0	7.41E-05	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 4-3 ค่าสหสัมพันธ์ระหว่างแต่ละลักษณะของข้อมูล KDDcup99 จำนวน 34 ลักษณะ

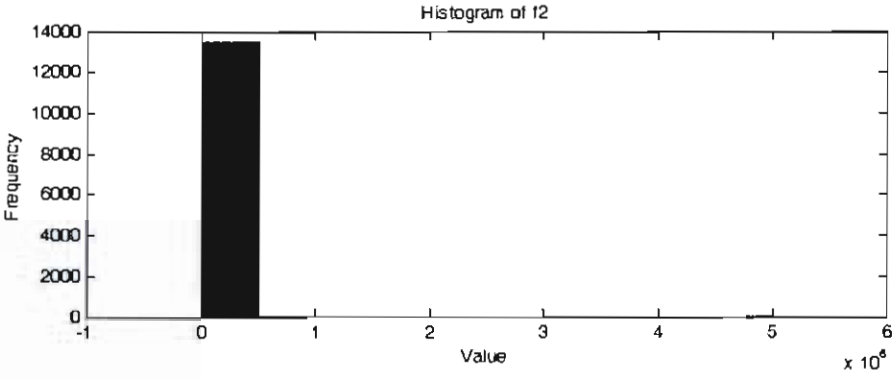
	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14	f15	f16	f17
f1	1																
f2	-0.00051	1															
f3	-0.00031	-0.00102	1														
f4	-0.00017	0.001262	-0.00035	1													
f5	0.002711	-0.00415	-0.00399	-0.00136	1												
f6	-0.0007	-0.00222	-0.00141	-0.00079	0.004503	1											
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 4-3 (ต่อ)

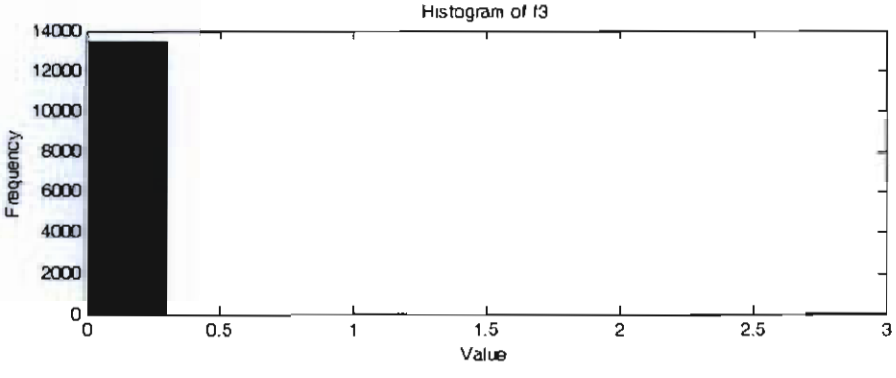
	f18	f19	f20	f21	f22	f23	f24	f25	f26	f27	f28	f29	f30	f31	f32	f33	f34
f18	1																
f19	0.927679	1															
f20	-0.07669	-0.06654	1														
f21	-0.07291	-0.15973	0.962791	1													
f22	-0.55315	-0.51906	-0.4886	-0.4698	1												
f23	0.060497	0.061472	0.625324	0.601707	-0.75729	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



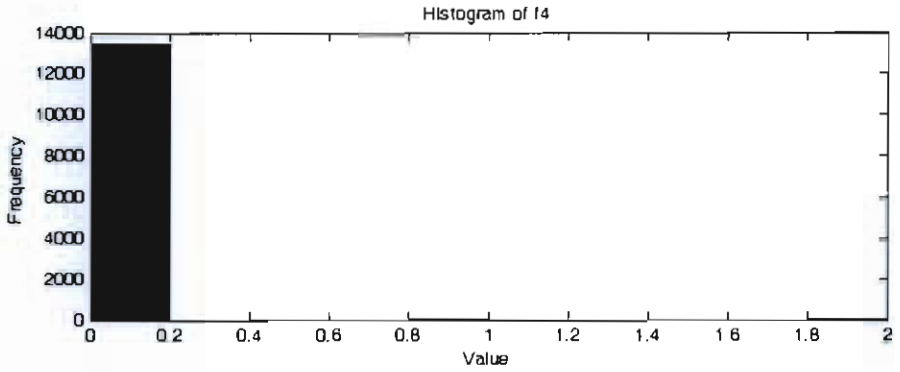
ภาพที่ 4-1 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 1 (f1)



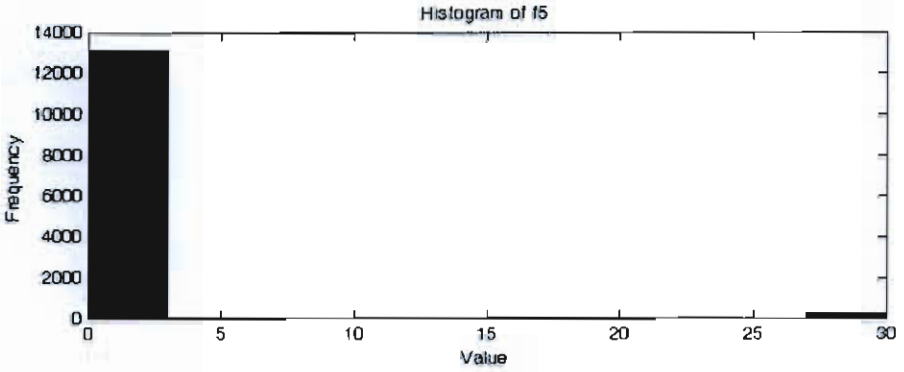
ภาพที่ 4-2 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 2 (f2)



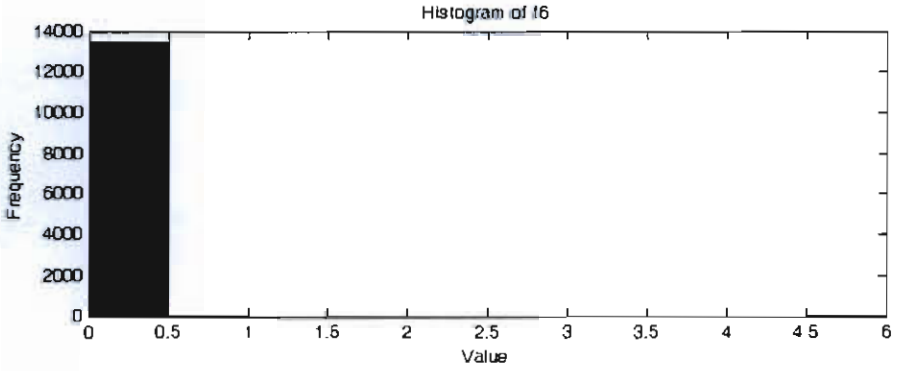
ภาพที่ 4-3 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 3 (f3)



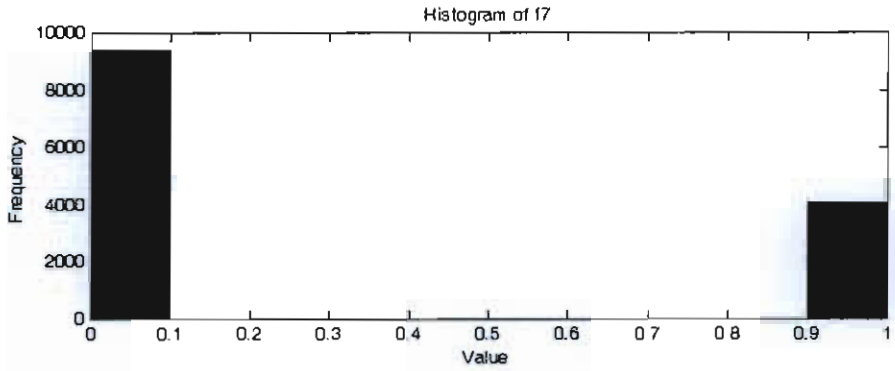
ภาพที่ 4-4 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 4 (f4)



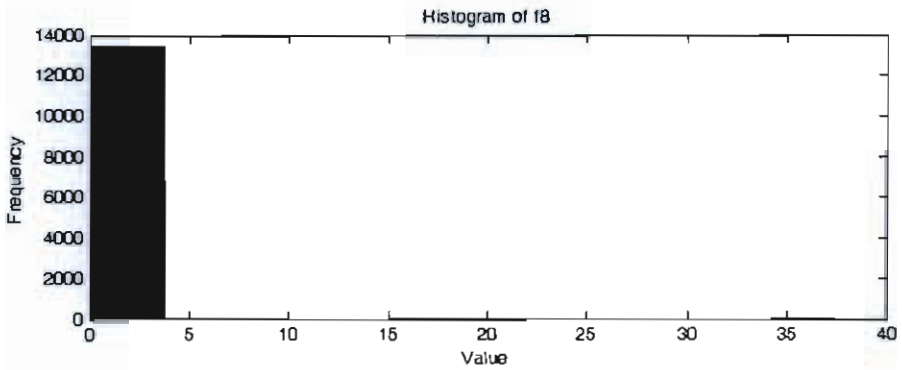
ภาพที่ 4-5 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 5 (f5)



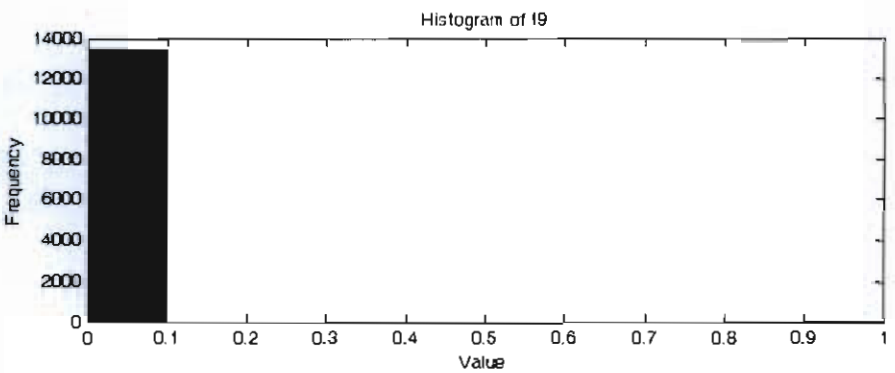
ภาพที่ 4-6 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 6 (f6)



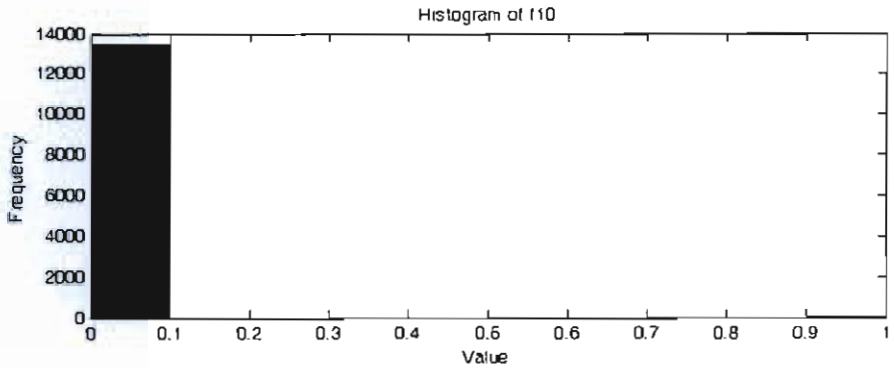
ภาพที่ 4-7 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 7 (f7)



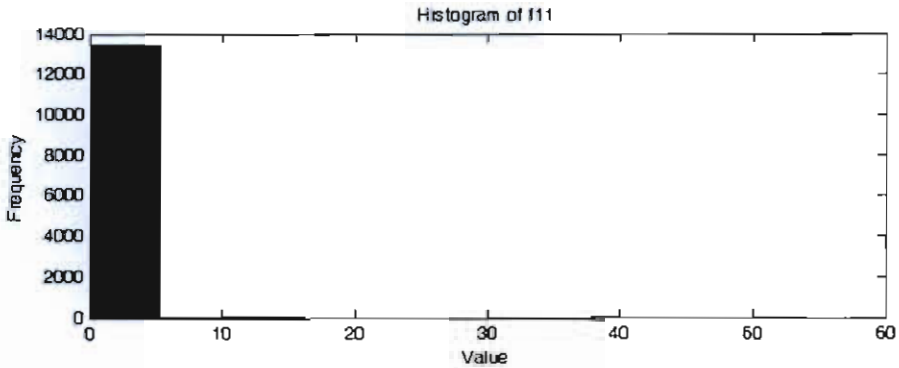
ภาพที่ 4-8 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 8 (f8)



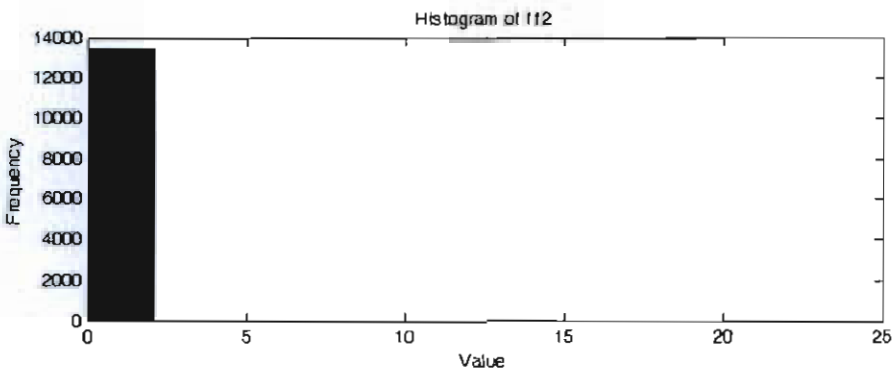
ภาพที่ 4-9 Histogram ของลักษณะข้อมูล KDDcup99ลักษณะที่ 9 (f9)



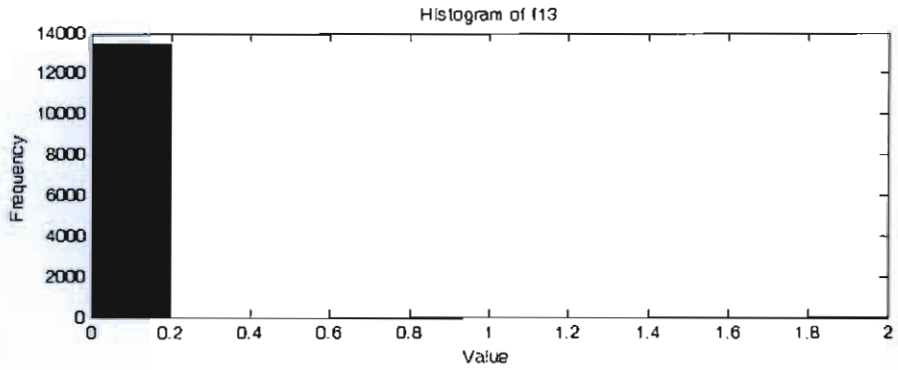
ภาพที่ 4-10 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 10 (f10)



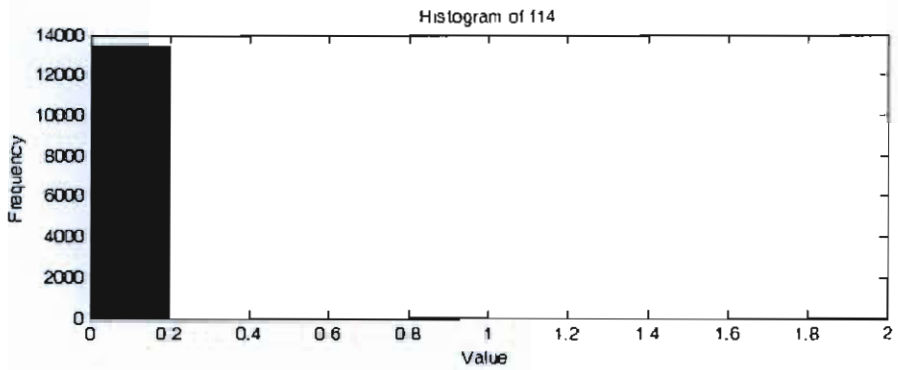
ภาพที่ 4-11 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 11 (f11)



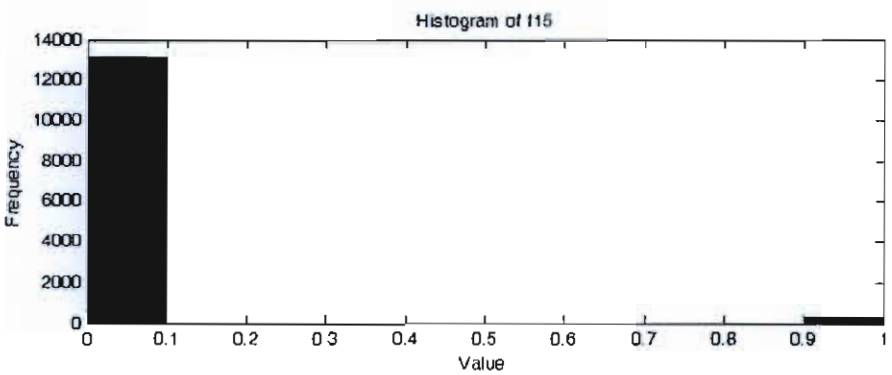
ภาพที่ 4-12 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 12 (f12)



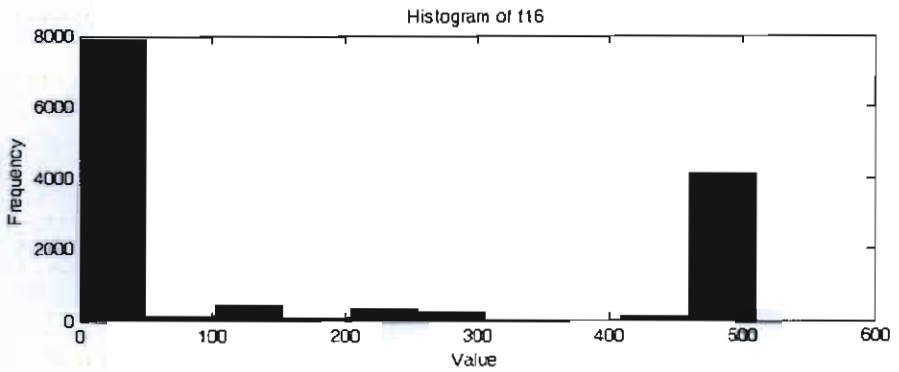
ภาพที่ 4-13 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 13 (f13)



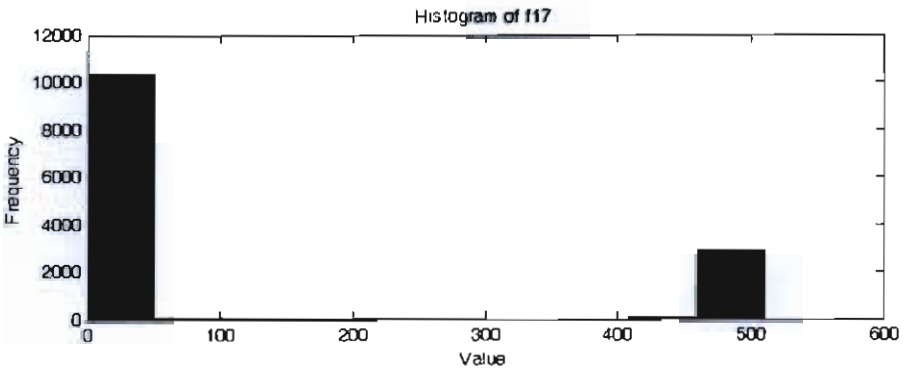
ภาพที่ 4-14 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 14 (f14)



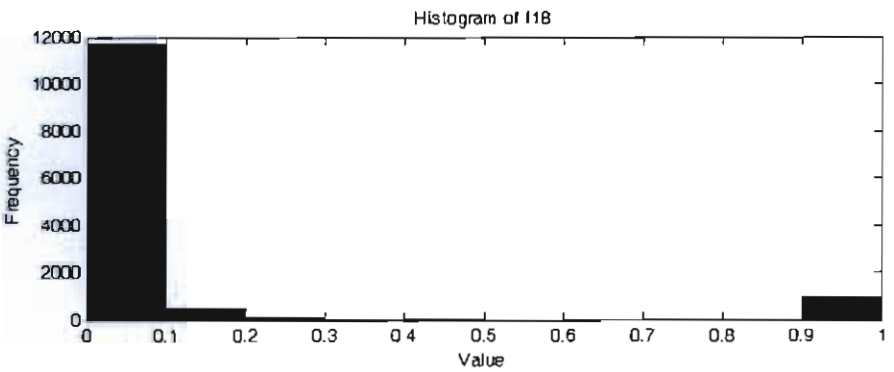
ภาพที่ 4-15 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 15 (f15)



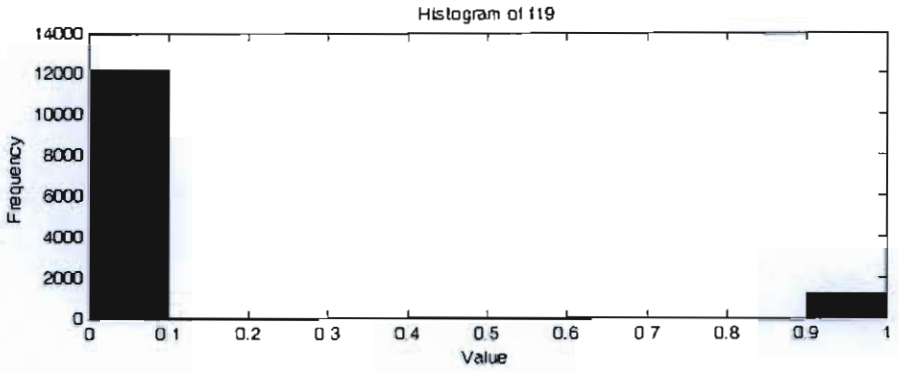
ภาพที่ 4-16 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 16 (f16)



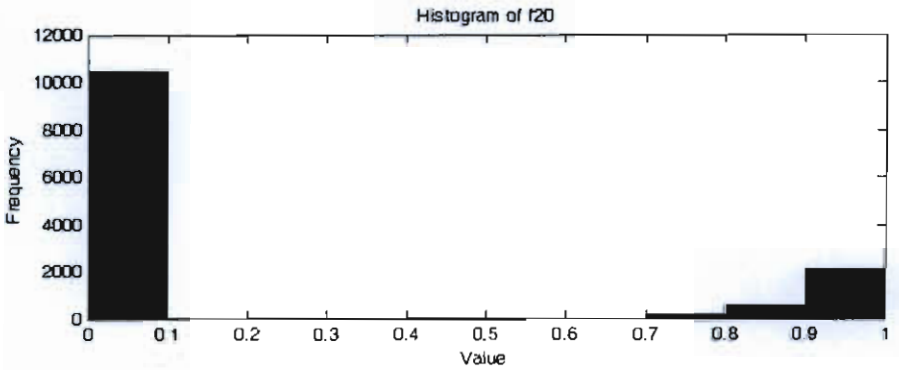
ภาพที่ 4-17 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 17 (f17)



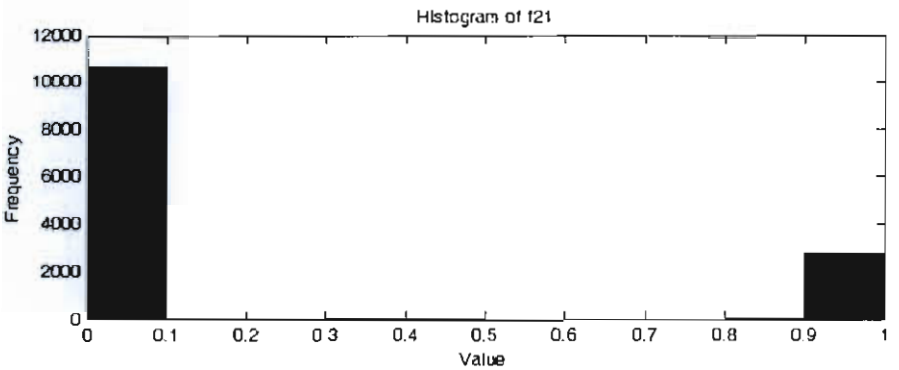
ภาพที่ 4-18 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 18 (f18)



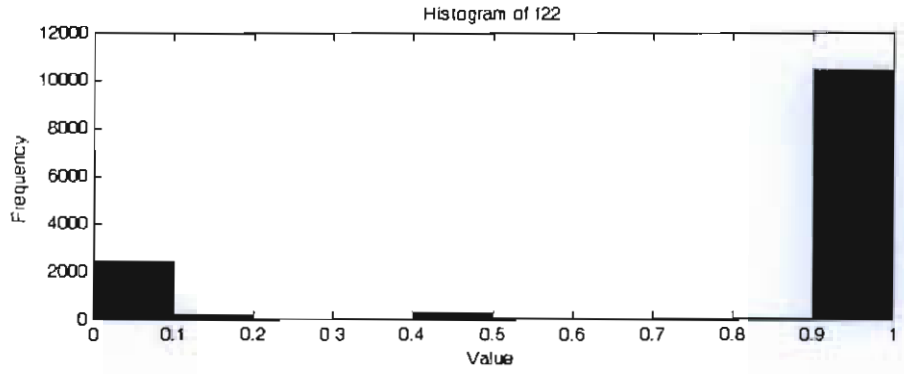
ภาพที่ 4-19 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 19 (f19)



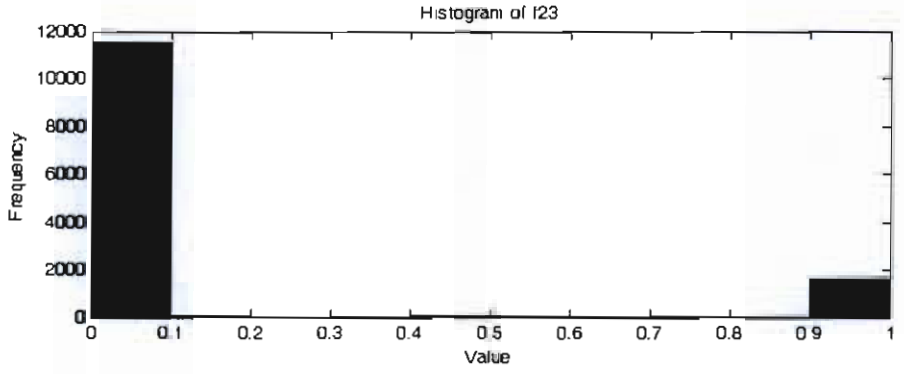
ภาพที่ 4-20 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 20 (f20)



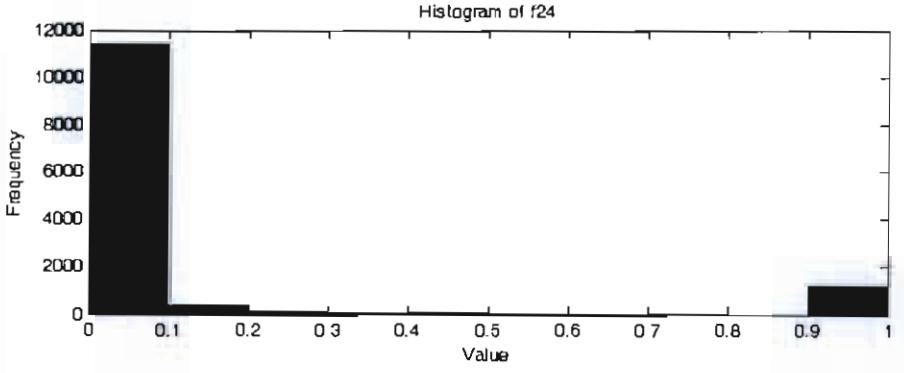
ภาพที่ 4-21 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 21 (f21)



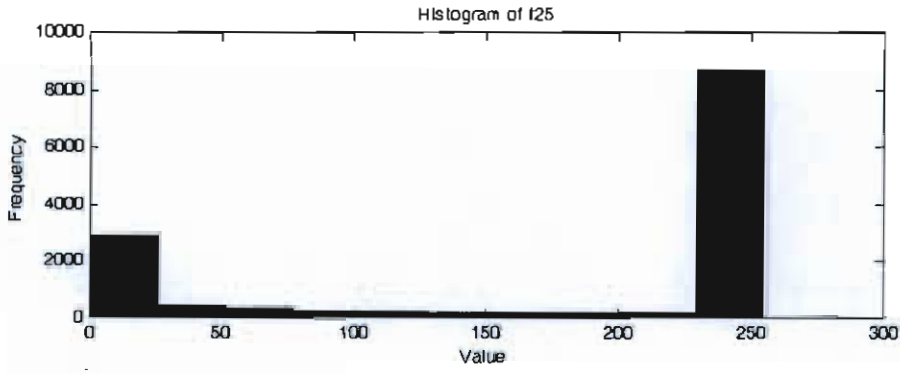
ภาพที่ 4-22 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 22 (f22)



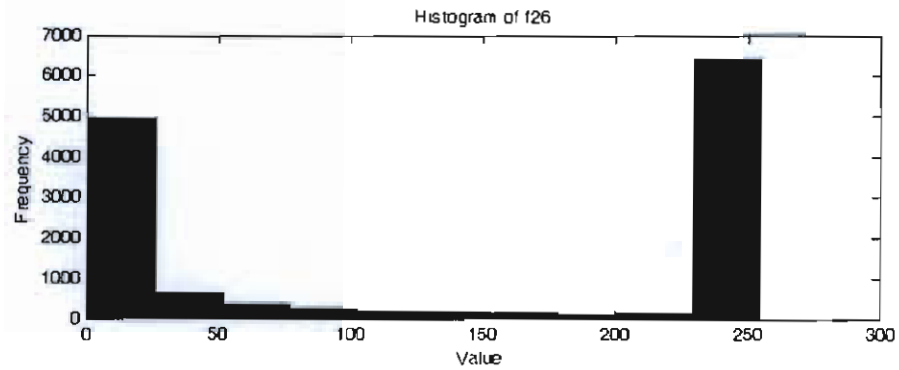
ภาพที่ 4-23 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 23 (f23)



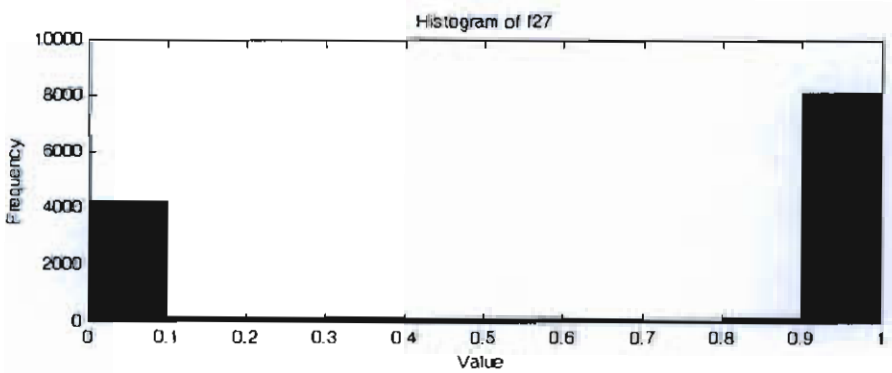
ภาพที่ 4-24 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 24 (f24)



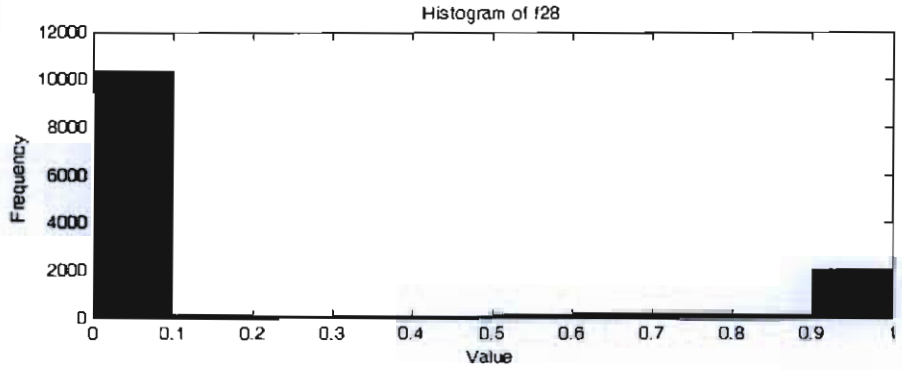
ภาพที่ 4-25 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 25 (f25)



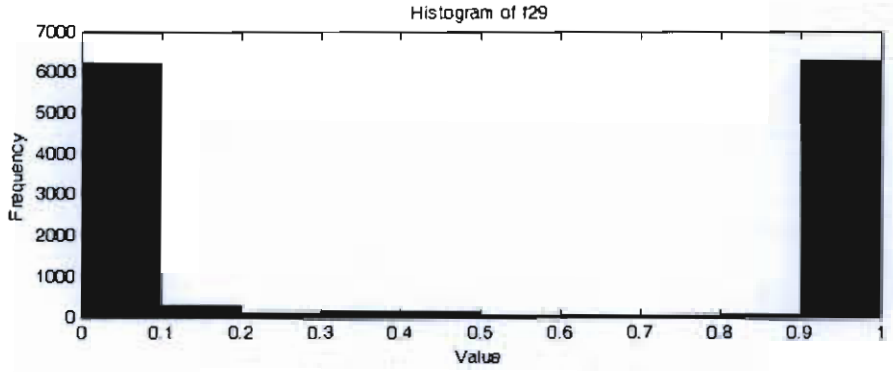
ภาพที่ 4-26 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 26 (f26)



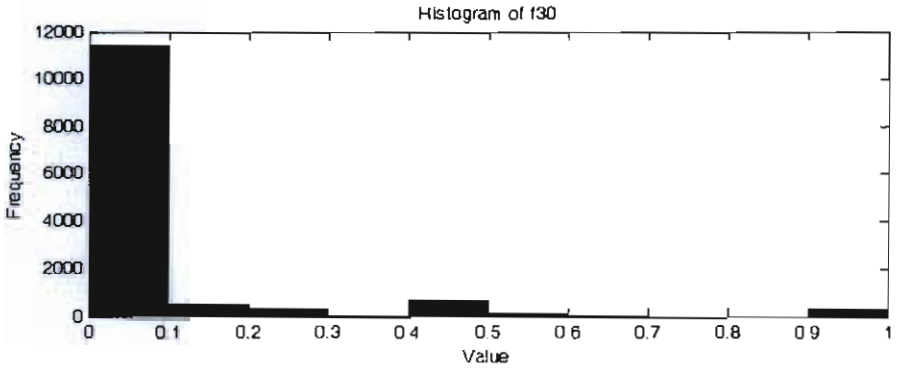
ภาพที่ 4-27 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 27 (f27)



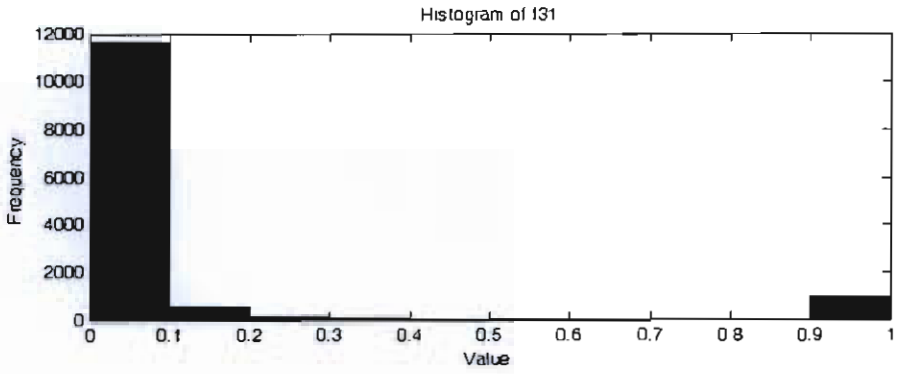
ภาพที่ 4-28 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 28 (f28)



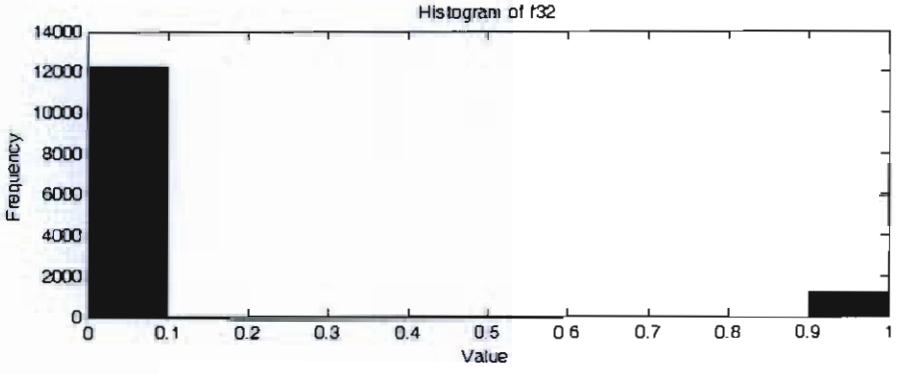
ภาพที่ 4-29 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 29 (f29)



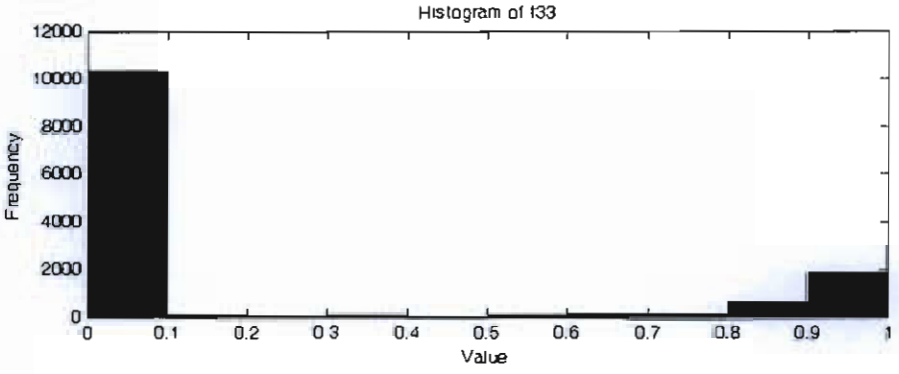
ภาพที่ 4-30 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 30 (f30)



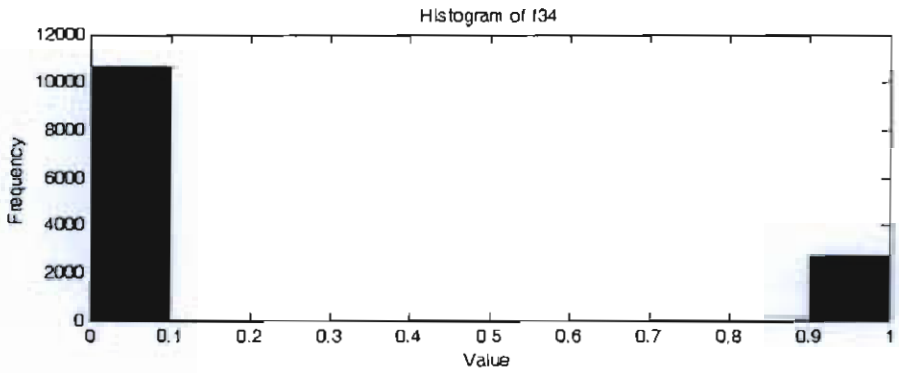
ภาพที่ 4-31 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 31 (f31)



ภาพที่ 4-32 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 32 (f32)



ภาพที่ 4-33 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 33 (f33)



ภาพที่ 4-34 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 34 (f34)

ผลการทดสอบฟังก์ชันผิดพลาดสำหรับการเลือกลักษณะด้วยวิธีฮิวริสติกกรีดี

จากการทดสอบฟังก์ชันโดยนำชุดข้อมูล KDDCup99 ที่ผ่านขั้นตอนกระบวนการเตรียมข้อมูลแล้ว ทดสอบการเลือกลักษณะโดยใช้ฮิวริสติกกรีดีอัลกอริทึมทั้ง 4 วิธี คือ HGIS1 HGIS2 HGIS3 และ HGIS4 โดยมีฟังก์ชันผิดพลาดตามตารางที่ 4-4 ซึ่งการเลือกลักษณะด้วยวิธีฮิวริสติกกรีดีอัลกอริทึมนี้แบ่งออกเป็น 2 ขั้นตอน คือ เลือกลักษณะฐาน และการเติมลักษณะที่หลุดออกเข้าไป โดยผลการทดลองในแต่ละฟังก์ชันเป็นดังนี้

ตารางที่ 4-4 ฟังก์ชันผิดพลาด

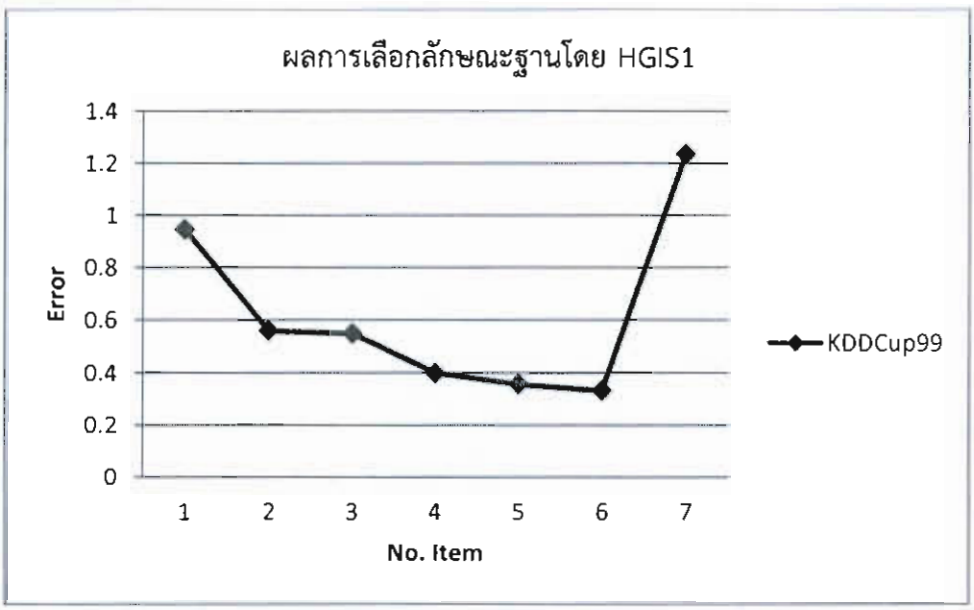
	Error Function
HGIS1	$F(\{itemset\}) = rmse$
HGIS2	$F(\{itemset\}) = Nmiss$
HGIS3	$F(\{itemset\}) = rmse + Nmiss$
HGIS4	$F(\{itemset\}) = rmse * Nmiss$

1. ผลการเลือกลักษณะของชุดข้อมูล KDDCup99 โดยใช้ HGIS1

จากการเลือกลักษณะโดยใช้ HGIS1 ผลการเลือกลักษณะฐานในแต่ละรอบแสดงดังตารางที่ 4-5 และกราฟแสดงค่าผิดพลาดของแต่ละชุดลักษณะที่ถูกเลือกมาในแต่ละรอบแสดงดังภาพที่ 4-35 จะเห็นว่า ชุดลักษณะที่ 6 ได้แก่ลักษณะ 5, 7, 16, 23, 29 แล 30 มีค่าความผิดพลาดต่ำสุด จึงเลือกชุดลักษณะนี้เป็นฐาน

ตารางที่ 4-5 ผลการเลือกลักษณะฐานชุดข้อมูล KDDCup99 ในแต่ละรอบโดย HGIS1

No. Itemset	Attribute Item	error
1	1	0.9456
2	16,29	0.5581
3	7,27,30	0.5481
4	5,7,16,29	0.3981
5	5,7,16,23,29	0.3556
6	5,7,16,23,29,30	0.3513
7	3,4,5,6,7,10,15	1.2346

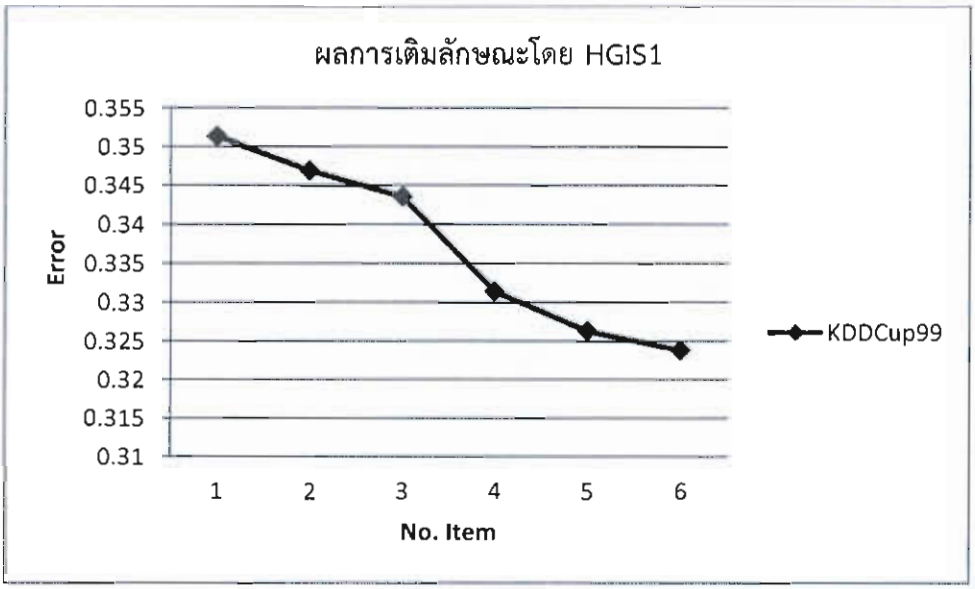


ภาพที่ 4-35 กราฟแสดงความสัมพันธ์ระหว่างค่าความผิดพลาดของแต่ละชุดลักษณะของชุดข้อมูล KDDCup99 ในแต่ละรอบโดย HGIS1 ในขั้นตอนการหาลักษณะฐาน

จากลักษณะฐานที่ได้นำมาเติมลักษณะที่หลุดออกไปเพื่อเพิ่มประสิทธิภาพที่ดีขึ้น จึงได้ผลการเติมลักษณะในแต่ละรอบดังตารางที่ 4-6 และแสดงค่าความผิดพลาดที่ลดลงเมื่อเติมลักษณะเข้าไปดังภาพที่ 4-36 จนกระทั่งได้ชุดของลักษณะที่มีค่าความผิดพลาดต่ำสุด

ตารางที่ 4-6 ผลการเติมลักษณะชุดข้อมูล KDDCup99 ในแต่ละรอบโดย HGIS1

No. Itemset	Attribute Item	error
1	5,7,16,23,29,30	0.3513
2	5,7,16,23,26,29,30	0.3469
3	5,7,16,22,23,26,29,30	0.3435
4	5,6,7,16,22,23,26,29,30	0.3314
5	3,5,6,7,16,22,23,26,29,30	0.3263
6	1,3,5,6,7,16,22,23,26,29,30	0.3238



ภาพที่ 4-36 กราฟแสดงความสัมพันธ์ระหว่างค่าความผิดพลาดของแต่ละชุดลักษณะของชุดข้อมูล KDDCup99 ในแต่ละรอบโดย HGIS2 ในขั้นตอนการเติมลักษณะ

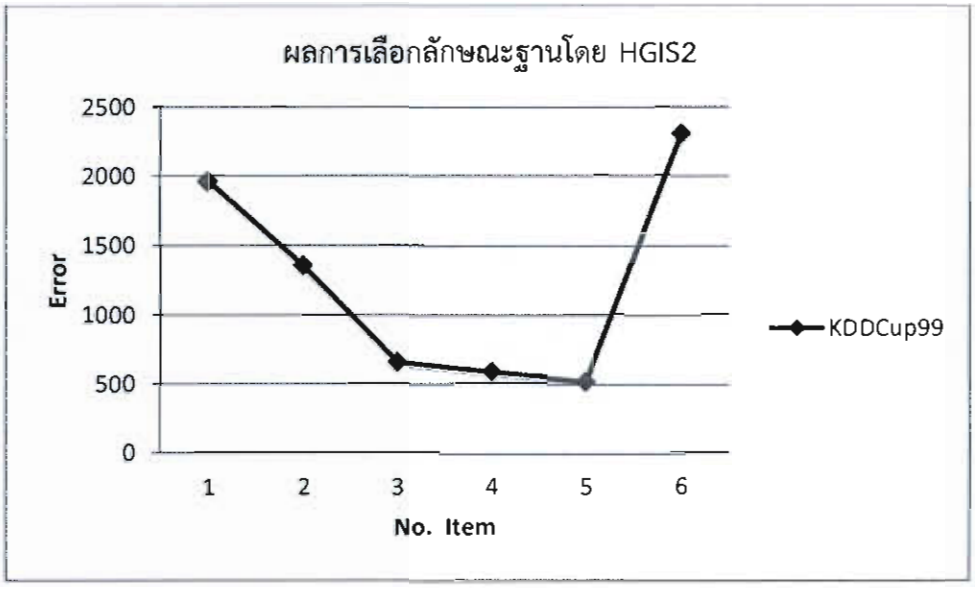
ดังนั้นการเลือกลักษณะของข้อมูล KDDCup99 โดยใช้ HGIS1 ได้จำนวนลักษณะ 11 ลักษณะ ได้แก่ 1, 3, 5, 6, 7, 16, 22, 23, 26, 29 และ 30

2. ผลการเลือกลักษณะของชุดข้อมูล KDDCup99 โดยใช้ HGIS2

จากการเลือกลักษณะโดยใช้ HGIS2 ผลการเลือกลักษณะฐานในแต่ละรอบแสดงดังตารางที่ 4-7 และแสดงดังภาพที่ 4-37 กราฟแสดงค่าผิดพลาดของแต่ละชุดลักษณะที่ถูกเลือกมาในแต่ละรอบ จะเห็นว่า ชุดลักษณะที่ 5 ได้แก่ลักษณะ 9, 15, 16, 28 และ 29 ให้ค่าความผิดพลาดต่ำสุด จึงเลือกชุดลักษณะนี้เป็นฐาน

ตารางที่ 4-7 ผลการเลือกลักษณะฐานชุดข้อมูล KDDCup99 ในแต่ละรอบโดย HGIS2

No. Itemset	Attribute Itemset	error
1	17	1958
2	17,29	1352
3	5,16,29	657
4	5,16,26,29	584
5	9,15,16,28,29	523
6	3,6,8,13,15,28	2302

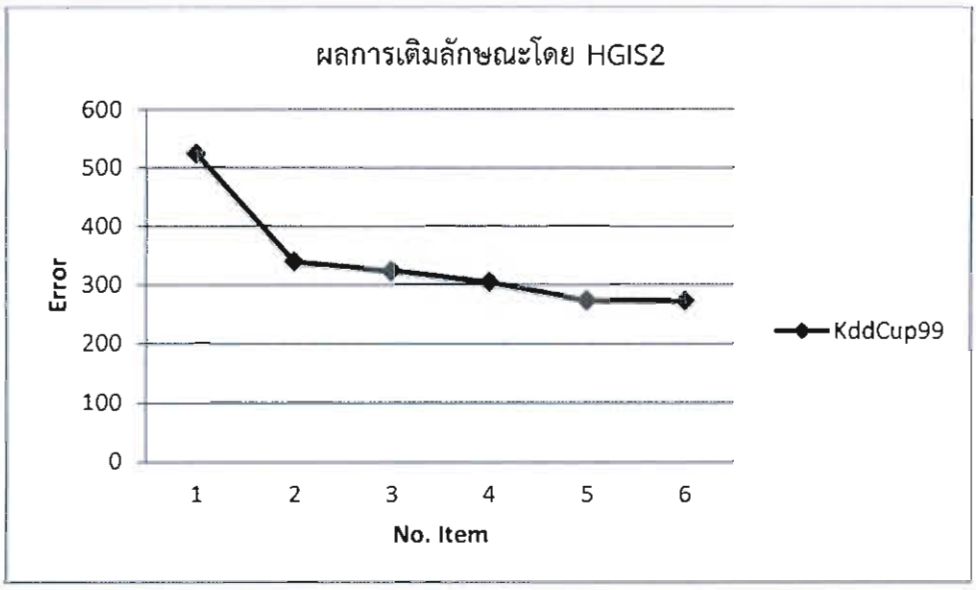


ภาพที่ 4-37 กราฟแสดงความสัมพันธ์ระหว่างค่าความผิดพลาดของแต่ละชุดลักษณะของชุดข้อมูล KDDCup99 ในแต่ละรอบโดย HGIS2 ในขั้นตอนการหาลักษณะฐาน

จากลักษณะฐานที่ได้นำมาเติมลักษณะที่หลุดออกไปเพื่อเพิ่มประสิทธิภาพที่ดีขึ้น จึงได้ผลการเติมลักษณะในแต่ละรอบดังตารางที่ 4-8 และแสดงค่าความผิดพลาดที่ลดลงเมื่อเติมลักษณะเข้าไปดังภาพที่ 4-38 จนกระทั่งได้ชุดของลักษณะที่มีค่าความผิดพลาดต่ำสุด

ตารางที่ 4-8 ผลการเติมลักษณะชุดข้อมูล KDDCup99 ในแต่ละรอบโดย HGIS2

No. Itemset	Attribute Itemset	error
1	9,15,16,28,29	523
2	7,9,15,16,28,29	338
3	6,7,9,15,16,28,29	323
4	3,6,7,9,15,16,28,29	303
5	3,6,7,9,15,16,26,28,29	273
6	1,3,6,7,9,15,16,26,28,29	271



ภาพที่ 4-38 กราฟแสดงความสัมพันธ์ระหว่างค่าความผิดพลาดของแต่ละชุดลักษณะของชุดข้อมูล KDDCup99 ในแต่ละรอบโดย HGIS2 ในขั้นตอนการเติมลักษณะ

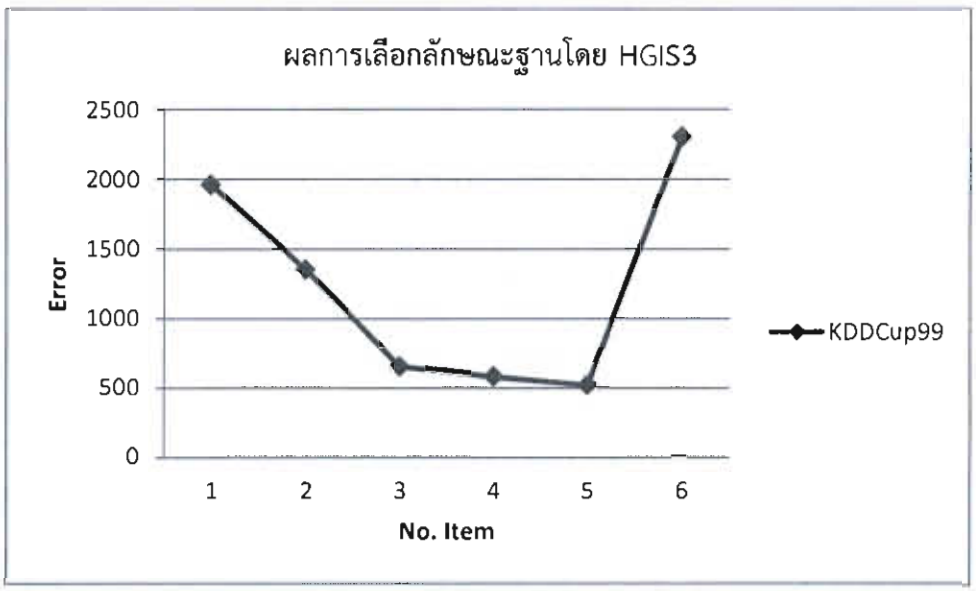
ดังนั้นการเลือกลักษณะของข้อมูล KDDCup99 โดยใช้ HGIS2 ได้จำนวนลักษณะ 10 ลักษณะ ได้แก่ 1, 3, 6, 7, 9, 15, 16, 26, 28 และ 29

3. ผลการเลือกลักษณะของชุดข้อมูล KDDCup99 โดยใช้ HGIS3

จากการเลือกลักษณะโดยใช้ HGIS3 ผลการเลือกลักษณะฐานในแต่ละรอบแสดงดังตารางที่ 4-9 และแสดงดังภาพที่ 4-39 กราฟแสดงค่าผิดพลาดของแต่ละชุดลักษณะที่ถูกเลือกมาในแต่ละรอบ จะเห็นว่า ชุดลักษณะที่ 5 ได้แก่ลักษณะ 9, 15, 16, 28 และ 29 ให้ค่าความผิดพลาดต่ำสุด จึงเลือกชุดลักษณะนี้เป็นฐาน

ตารางที่ 4-9 ผลการเลือกลักษณะฐานชุดข้อมูล KDDCup99 ในแต่ละรอบโดย HGIS3

No. Itemset	Attribute Itemset	error
1	17	1958.64
2	17,29	1352.55
3	5,16,29	657.43
4	5,16,26,29	584.42
5	9,15,16,28,29	523.40
6	3,6,8,13,15,28	2302.80

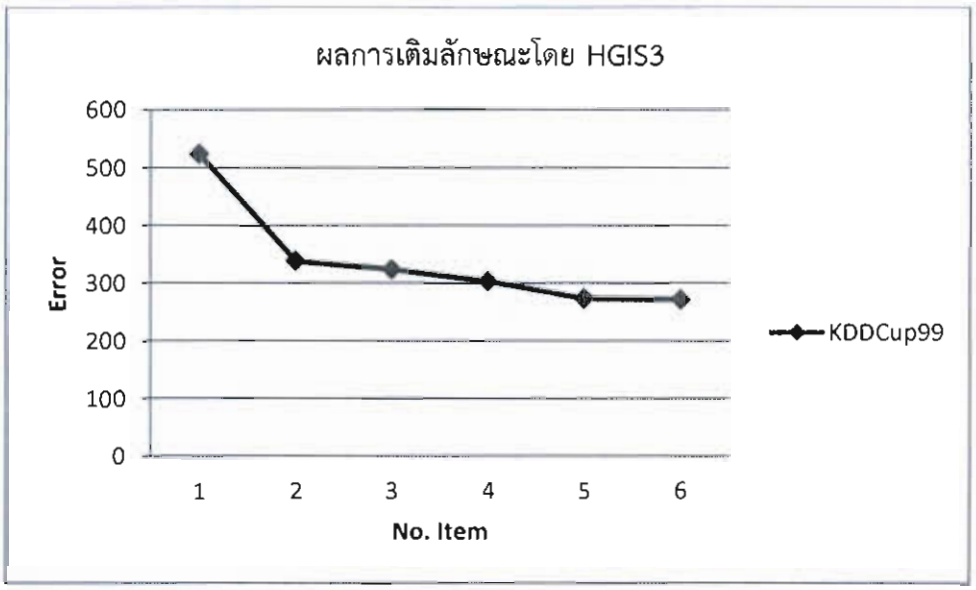


ภาพที่ 4-39 กราฟแสดงความสัมพันธ์ระหว่างค่าความผิดพลาดของแต่ละชุดลักษณะของชุดข้อมูล KDDCup99 ในแต่ละรอบโดย HGIS3 ในขั้นตอนการหาลักษณะฐาน

จากลักษณะฐานที่ได้นำมาเติมลักษณะที่หลุดออกไปเพื่อเพิ่มประสิทธิภาพที่ดีขึ้น จึงได้ผลการเติมลักษณะในแต่ละรอบดังตารางที่ 4-10 และแสดงค่าความผิดพลาดที่ลดลงเมื่อเติมลักษณะเข้าไปดังภาพที่ 4-40 จนกระทั่งได้ขุดของลักษณะที่มีค่าความผิดพลาดต่ำสุด

ตารางที่ 4-10 ผลการเติมลักษณะชุดข้อมูล KDDCup99 ในแต่ละรอบโดย HGIS3

No. Itemset	Attribute Itemset	error
1	9,15,16,28,29	523.40
2	7,9,15,16,28,29	338.35
3	6,7,9,15,16,28,29	323.33
4	3,6,7,9,15,16,28,29	303.31
5	3,6,7,9,15,16,26,28,29	273.30
6	1,3,6,7,9,15,16,26,28,29	271.29



ภาพที่ 4-40 กราฟแสดงความสัมพันธ์ระหว่างค่าความผิดพลาดของแต่ละขุดลักษณะของชุดข้อมูล KDDCup99 ในแต่ละรอบโดย HGIS3 ในขั้นตอนการเติมลักษณะ

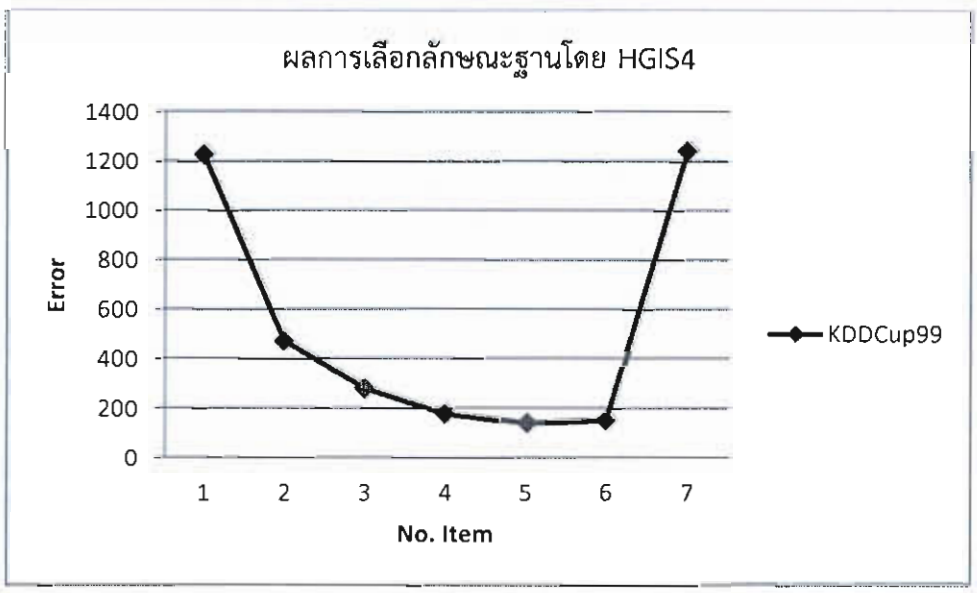
ดังนั้นการเลือกลักษณะของข้อมูล KDDCup99 โดยใช้ HGIS3 ได้จำนวนลักษณะ 10 ลักษณะ ได้แก่ 1, 3, 6, 7, 9, 15, 16, 26, 28 และ 29

4. ผลการเลือกลักษณะของชุดข้อมูล KDDCup99 โดยใช้ HGIS4

จากการเลือกลักษณะโดยใช้ HGIS4 ผลการเลือกลักษณะฐานในแต่ละรอบแสดงดังตารางที่ 4-11 และกราฟแสดงค่าผิดพลาดของแต่ละชุดลักษณะที่ถูกเลือกมาในแต่ละรอบแสดงดังภาพที่ 4-41 จะเห็นว่า ชุดลักษณะที่ 5 ได้แก่ลักษณะ 5, 7, 16, 23 และ 29 มีค่าความผิดพลาดต่ำสุด จึงเลือกชุดลักษณะนี้เป็นฐาน

ตารางที่ 4-11 ผลการเลือกลักษณะฐานชุดข้อมูล KDDCup99 ในแต่ละรอบโดย HGIS4

No. Item	Attribute Item	error
1	1	1228.276
2	16,29	471.5989
3	5,16,29	280.1717
4	5,7,16,29	177.5722
5	5,7,16,23,29	137.9892
6	5,7,15,16,23,29	152.0447
7	6,7,9,15,21,22,23	1242.516

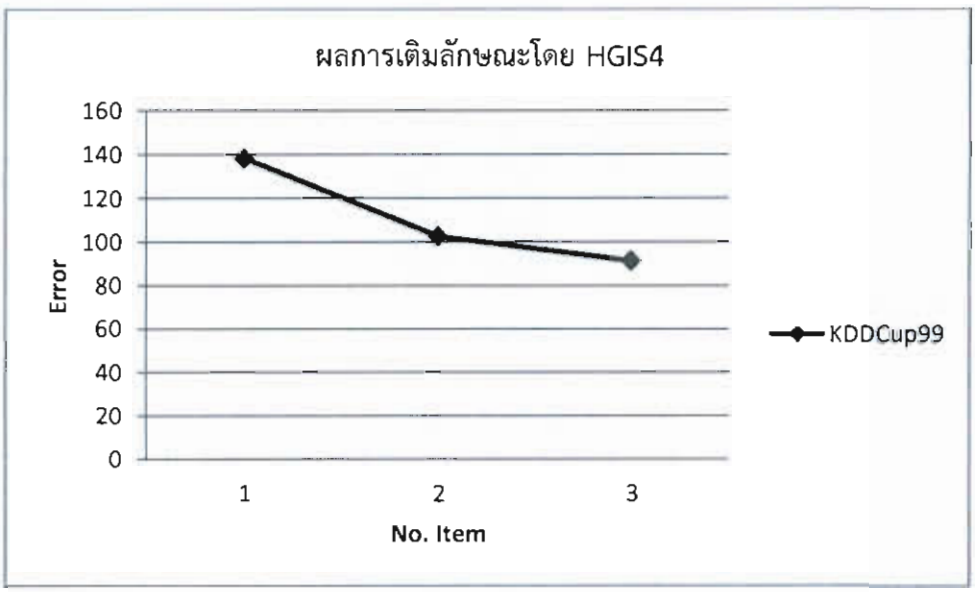


ภาพที่ 4-41 กราฟแสดงความสัมพันธ์ระหว่างค่าความผิดพลาดของแต่ละชุดลักษณะของชุดข้อมูล KDDCup99 ในแต่ละรอบโดย HGIS4 ในขั้นตอนการหาลักษณะฐาน

จากลักษณะฐานที่เลือกในขั้นตอนก่อนหน้า นำมาเติมลักษณะที่หลุดออกไปเพื่อเพิ่มประสิทธิภาพที่ดีขึ้น จึงได้ผลการเติมลักษณะในแต่ละรอบดังตารางที่ 4-12 และแสดงค่าความผิดพลาดที่ลดลงเมื่อเติมลักษณะเข้าไปดังภาพที่ 4-42 จนกระทั่งได้ชุดของลักษณะที่มีค่าความผิดพลาดต่ำสุด

ตารางที่ 4-12 ผลการเติมลักษณะชุดข้อมูล KDDCup99 ในแต่ละรอบโดย HGIS4

No. Item	Attribute Item	error
1	5,7,16,23,29	137.9892
2	5,7,16,23,26,29	102.4037
3	3,5,7,16,23,26,29	91.28709



ภาพที่ 4-42 กราฟแสดงความสัมพันธ์ระหว่างค่าความผิดพลาดของแต่ละชุดลักษณะของชุดข้อมูล KDDCup99 ในแต่ละรอบโดย HGIS4 ในขั้นตอนการเติมลักษณะ

ดังนั้นการเลือกลักษณะของข้อมูล KDDCup99 โดยใช้ HGIS4 ได้จำนวนลักษณะ 7 ลักษณะ ได้แก่ 3, 5, 7, 16, 23, 26 และ 29

5. สรุปผลการทดสอบฟังก์ชันผิดพลาดสำหรับการเลือกลักษณะด้วยวิธีฮิวริสติกกรีดี

สำหรับ HGIS1 ลักษณะที่เลือกได้มีดังนี้ 1, 3, 5, 6, 7, 16, 22, 23, 26, 29, 30 จำนวน 11 ลักษณะ ซึ่งมีค่าความถูกต้องโดยรวมร้อยละ 94.9074 มีค่าความครบถ้วน ค่าความแม่นยำ ค่าเอฟเมเซอร์ และอัตราความผิดพลาดเชิงบวก ดังตารางที่ 4-13

ตารางที่ 4-13 ประสิทธิภาพของชุดลักษณะที่เลือกด้วย HGIS1

Class	Recall	Precision	F-measure	FAR
DoS	98.6553	98.5950	98.6251	0.6111
Normal	90.9747	95.3943	93.1321	1.9529
Probe	97.4893	90.9714	94.1176	4.1943
R2L	90.4018	95.2941	92.7835	0.4039
U2R	9.5238	95.2941	17.3169	0.0186
Weighted Avg.	94.9074	95.0178	94.7706	2.0882

สำหรับ HGIS2 ลักษณะที่เลือกได้มีดังนี้ 1, 3, 6, 7, 9, 15, 16, 26, 28, 29 จำนวน 10 ลักษณะ ซึ่งมีค่าความถูกต้องร้อยละ 95.1482 มีค่าความครบถ้วน ค่าความแม่นยำ ค่าเอฟเมเซอร์ และอัตราความผิดพลาดเชิงบวก ดังตารางที่ 4-14

ตารางที่ 4-14 ประสิทธิภาพของชุดลักษณะที่เลือกด้วย HGIS2

Class	Recall	Precision	F-measure	FAR
DoS	98.7783	99.2025	98.9899	0.3455
Normal	90.1324	95.7801	92.8705	1.7657
Probe	98.1017	90.6621	94.2353	4.3801
R2L	91.0515	95.7647	93.3486	0.3634
U2R	66.6667	95.7647	78.6094	0
Weighted Avg.	95.1482	95.2685	95.1217	2.0028

สำหรับ HGIS3 ลักษณะที่เลือกได้มีดังนี้ 1, 3, 6, 7, 9, 15, 16, 26, 28, 29 จำนวน 10 ลักษณะ ซึ่งมีค่าความถูกต้องร้อยละ 95.1482 มีค่าความครบถ้วน ค่าความแม่นยำ ค่าเอฟเมเซอร์ และอัตราความผิดพลาดเชิงบวก ดังตารางที่ 4-15

ตารางที่ 4-15 ประสิทธิภาพของชุดลักษณะที่เลือกด้วย HGIS3

Class	Recall	Precision	F-measure	FAR
DoS	98.7783	99.2025	98.9899	0.3455
Normal	90.1324	95.7801	92.8705	1.7657
Probe	98.1017	90.6621	94.2353	4.3801
R2L	91.0515	95.7647	93.3486	0.3634
U2R	66.6667	95.7647	78.6094	0
Weighted Avg.	95.1482	95.2685	95.1217	2.0028

สำหรับ HGIS4 ลักษณะที่เลือกได้มีดังนี้ 3, 5, 7, 16, 23, 26, 29 จำนวน 7 ลักษณะ ซึ่งมีค่าความถูกต้องร้อยละ 94.44 ค่าความครบถ้วน ค่าความแม่นยำ ค่าเอฟเมเซอร์ และอัตราความผิดพลาดเชิงบวก ดังตารางที่ 4-16

ตารางที่ 4-16 ประสิทธิภาพของชุดลักษณะที่เลือกด้วย HGIS4

Class	Recall	Precision	F-measure	FAR
DoS	98.5941	99.3839	98.9874	0.2657
Normal	88.2671	95.3216	91.6589	1.9262
Probe	98.5915	89.4942	93.8228	5.0173
R2L	90.8482	93.9954	92.3950	0.5250
U2R	14.2857	93.9954	24.8020	0.0558
Weighted Avg.	94.4444	94.6749	94.3346	2.2344

จากผลการทดลองหาฟังก์ชันผิดพลาดที่ใช้สำหรับเป็นเกณฑ์ในการเลือกลักษณะด้วยวิธีฮิวริสติกกรีดดีอัลกอริทึมที่เหมาะสม จึงสรุปได้ว่าฟังก์ชัน HGIS2 และ HGIS3 ให้ผลลัพธ์ที่เหมือนกันและดีกว่า HGIS1 และ HGIS4 ดังนั้นจึงเลือกฟังก์ชัน HGIS2 นั่นคือ คือ $F(\{itemset\}) = Nmiss$ ใช้สำหรับเป็นเกณฑ์ในการเลือกลักษณะด้วยวิธีฮิวริสติกกรีดดีอัลกอริทึม

ลักษณะข้อมูลของ KDDcup99 เมื่อเลือกลักษณะด้วย HGIS2 จำนวน 10 ลักษณะ

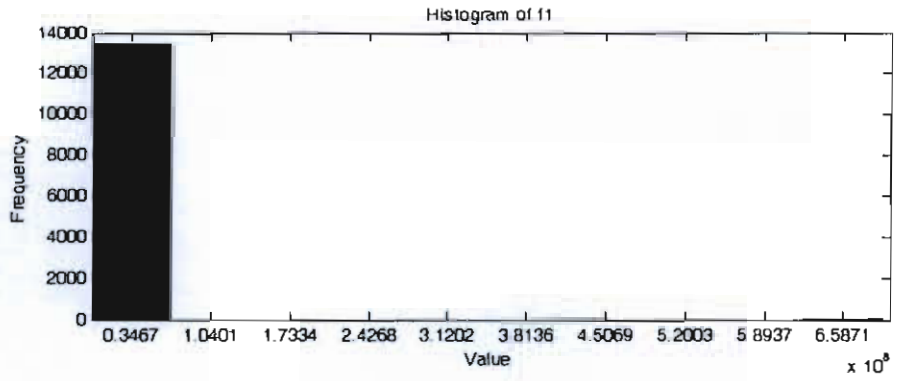
เมื่อนำข้อมูลที่ผ่านขั้นตอนการเตรียมข้อมูล 34 ลักษณะ มาเลือกลักษณะด้วยวิธีฮิวริสติกกรีดดี ซึ่งเลือกลักษณะออกมาได้จำนวน 10 ลักษณะได้แก่ 1, 3, 6, 7, 9, 15, 16, 26, 28 และ 29 โดยแต่ละลักษณะมีค่าสูงสุด ต่ำสุด ค่าเฉลี่ย และค่าส่วนเบี่ยงเบนมาตรฐานตามตารางที่ 4-17 ความสัมพันธ์ระหว่างแต่ละลักษณะทั้ง 10 ลักษณะแสดงในตารางที่ 4-18 โดยแต่ละลักษณะในลักษณะที่ 1 ถึง ลักษณะที่ 10 มีการกระจายตัวของข้อมูลดังภาพที่ 4-43 ถึง 4-52

ตารางที่ 4-17 ค่าทางสถิติของข้อมูล KDDcup99 จำนวน 10 ลักษณะ ที่เลือกลักษณะโดย HGIS2

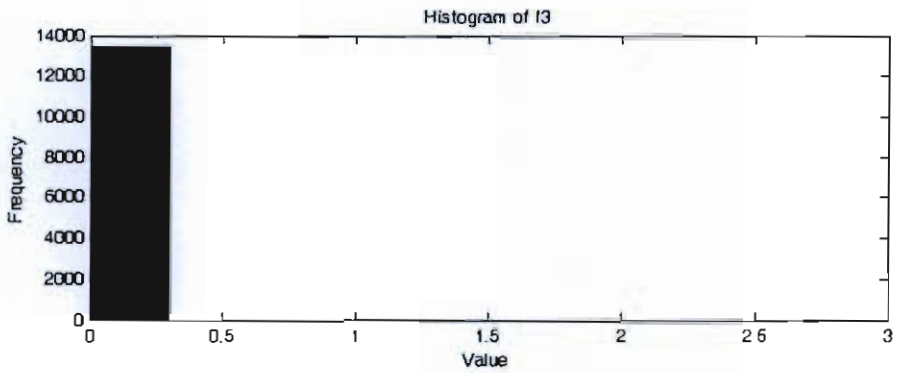
Features	Maximum	Minimum	Mean	Standard Deviation
1	6.93E+08	0	74547.73	5977289
3	3	0	0.001852	0.073517
6	5	0	0.004223	0.07541
7	1	0	0.301207	0.4588
9	1	0	0.002371	0.048632
15	1	0	0.024446	0.154436
16	511	0	182.1567	229.4395
26	255	1	138.053	117.0673
28	1	0	0.203488	0.371824
29	1	0	0.497511	0.481672

ตารางที่ 4-18 ค่าสหสัมพันธ์ระหว่างแต่ละลักษณะของข้อมูล KDDcup99 จำนวน 10 ลักษณะ ที่เลือกลักษณะโดย HGIS

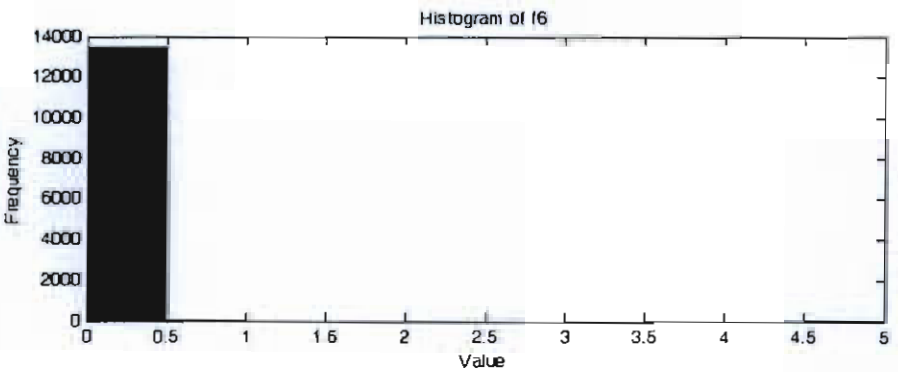
	1	3	6	7	9	15	16	26	28	29
1	1									
3	-0.00031	1								
6	-0.0007	-0.00141	1							
7	0.000182	-0.01654	-0.03676	1						
9	-0.0006	-0.00123	-0.00273	0.074247	1					
15	-0.00194	-0.00399	-0.00886	0.241114	-0.00772	1				
16	-0.00767	-0.0155	-0.04409	-0.50251	-0.03844	-0.12499	1			
26	-0.01321	-0.01186	-0.05482	0.189343	-0.04225	-0.10096	0.158605	1		
28	-0.0045	0.002555	-0.03059	-0.33084	-0.02602	-0.07428	0.173425	-0.627	1	
29	-0.00152	-0.00679	-0.03834	-0.3764	-0.00114	-0.16019	0.210185	0.163082	-0.11822	1



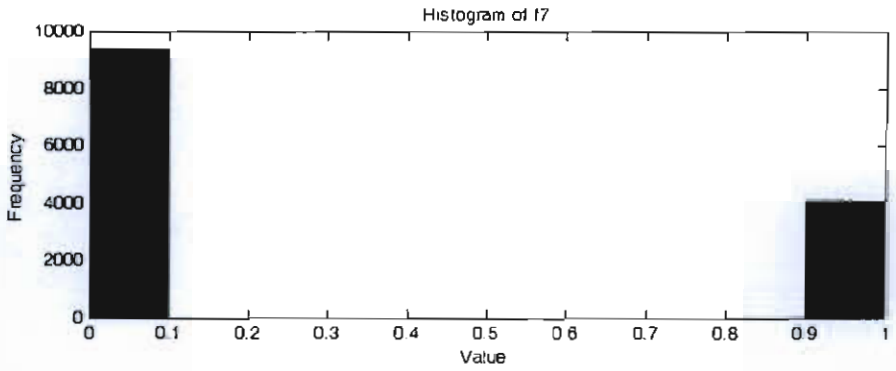
ภาพที่ 4-43 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 1 (f1)



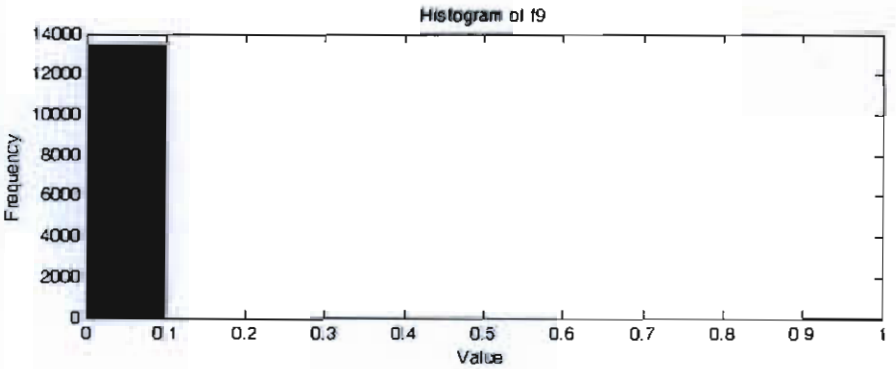
ภาพที่ 4-44 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 3 (f3)



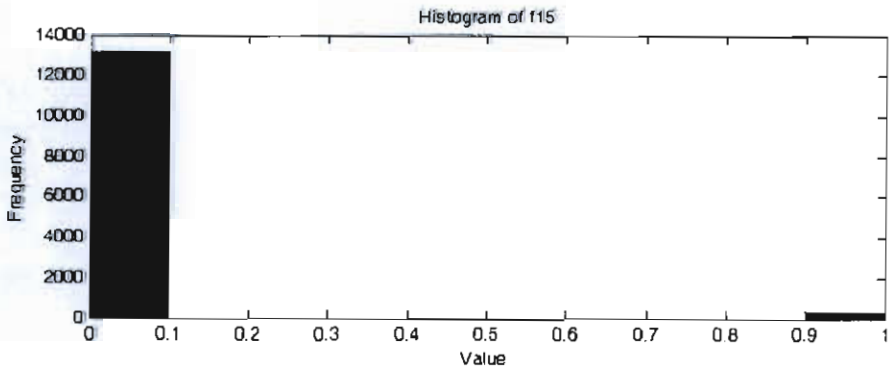
ภาพที่ 4-45 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 6 (f6)



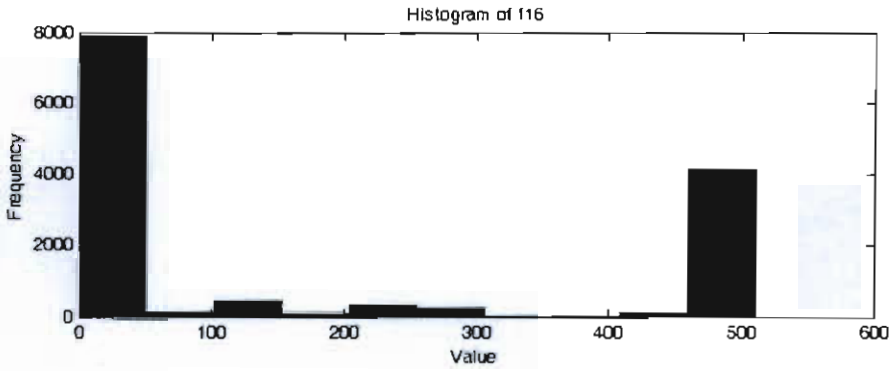
ภาพที่ 4-46 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 7 (f7)



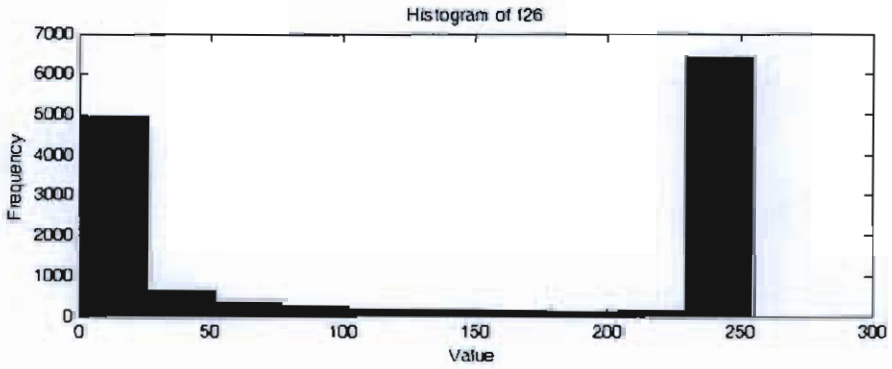
ภาพที่ 4-47 Histogram ของลักษณะข้อมูล KDDcup99ลักษณะที่ 9 (f9)



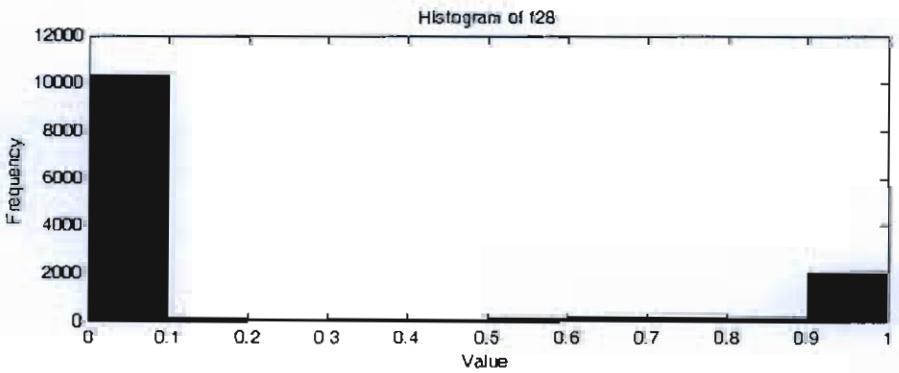
ภาพที่ 4-48 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 15 (f15)



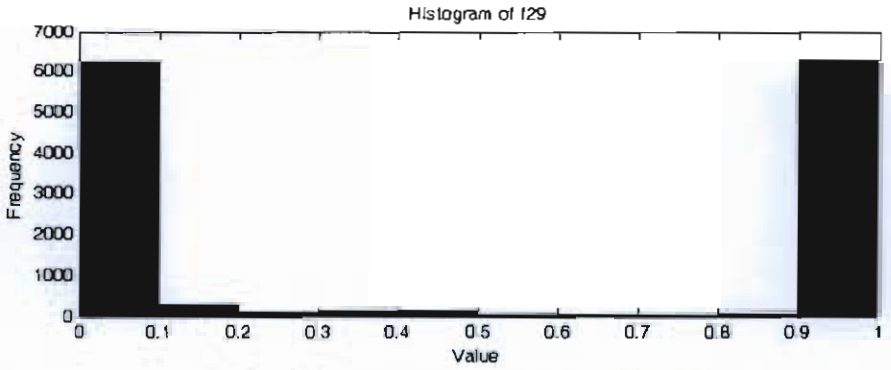
ภาพที่ 4-49 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 16 (f16)



ภาพที่ 4-50 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 26 (f26)



ภาพที่ 4-51 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 28 (f28)



ภาพที่ 4-52 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 29 (f29)

ลักษณะข้อมูล KDDCup99 เมื่อสกัดลักษณะด้วย PCA จำนวน 19 ลักษณะ

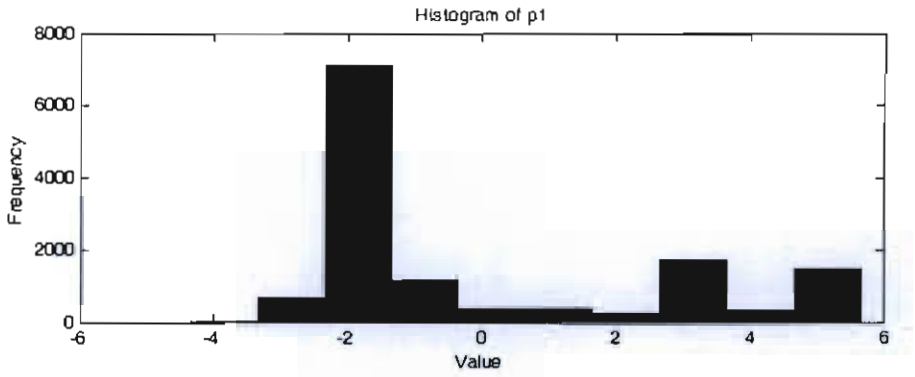
หลังจากที่ได้นำข้อมูลที่ผ่านขั้นตอนการเตรียมข้อมูล 34 ลักษณะ มาสกัดลักษณะด้วยวิธีวิเคราะห์องค์ประกอบหลัก ซึ่งสกัดออกมาได้จำนวน 19 ลักษณะ โดยแต่ละลักษณะมีค่าสูงสุด ค่าต่ำสุด ค่าเฉลี่ย และค่าส่วนเบี่ยงเบนมาตรฐานตามตารางที่ 4-19 ส่วนในตารางที่ 4-20 จะแสดงถึงความสัมพันธ์ระหว่างแต่ละลักษณะทั้ง 19 ลักษณะ และ ภาพที่ 4-53 ถึง 4-71 แสดงให้เห็นถึงการกระจายตัวของข้อมูลในแต่ละลักษณะ ในลักษณะที่ 1 ถึง ลักษณะที่ 19

ตารางที่ 4-19 ค่าทางสถิติของข้อมูล KDDcup99 เมื่อสกัดลักษณะด้วย PCA จำนวน 19 ลักษณะ

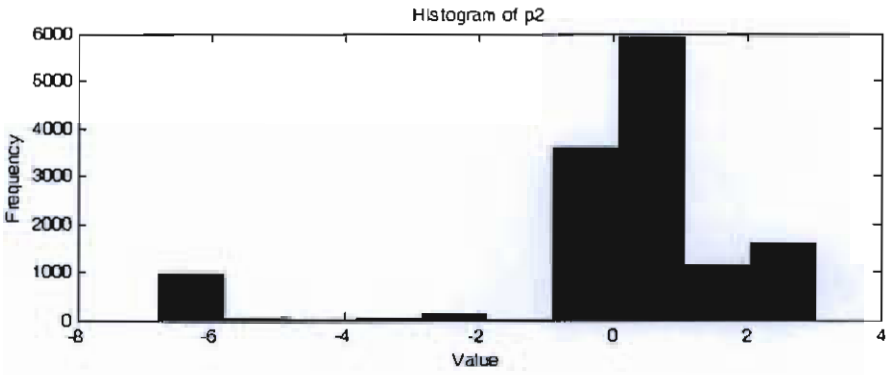
Features	Maximum	Minimum	Mean	Standard Deviation
p1	5.670465	-4.34379	-6.7E-08	2.704464
p2	3.033078	-6.78915	3.57E-08	2.108844
p3	2.677554	-48.5059	-4E-08	1.748789
p4	2.634652	-110.555	-5.1E-08	1.639518
p5	8.209626	-12.9017	-3.9E-08	1.505454
p6	6.763351	-6.99998	-4.6E-08	1.203837
p7	25.28668	-108.391	-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 4-20 ค่าสหสัมพันธ์ระหว่างแต่ละลักษณะของข้อมูล KDDcup99 เมื่อสกัดลักษณะด้วย PCA จำนวน 19 ลักษณะ

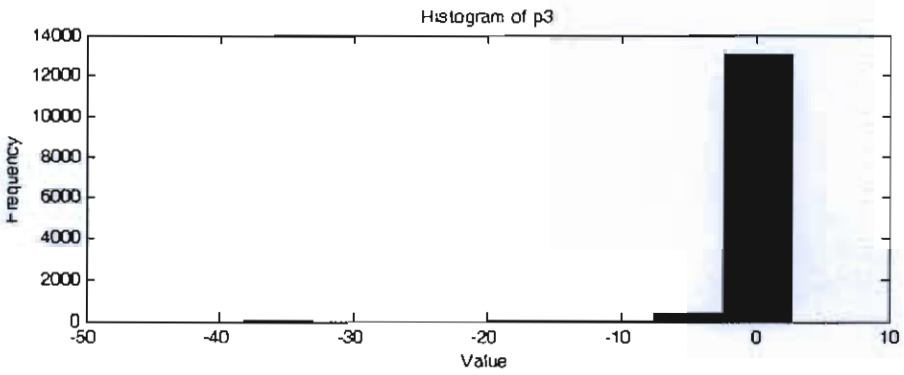
	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12	p13	p14	p15	p16	p17	p18	p19
p1	1																		
p2	-2.4E-08	1																	
p3	-3.5E-08	2.13E-08	1																
p4	5.11E-08	-4.4E-09	-4.1E-08	1															
p5	1.33E-08	1.61E-09	-4.3E-08	1.3E-08	1														
p6	6.65E-08	5.03E-09	-7.7E-08	4.91E-08	-9.7E-09	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	5.12E-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-06	-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



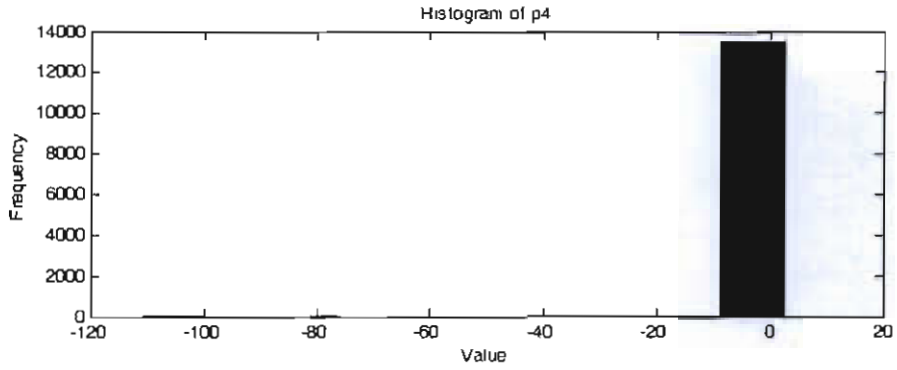
ภาพที่ 4-53 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 1 (p1)



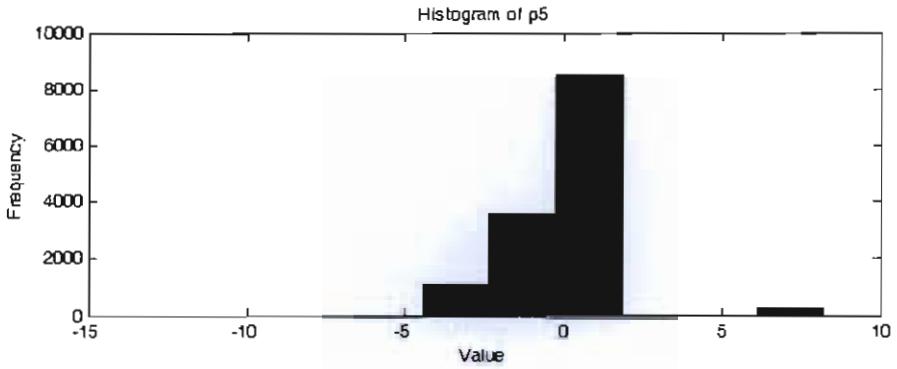
ภาพที่ 4-54 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 2 (p2)



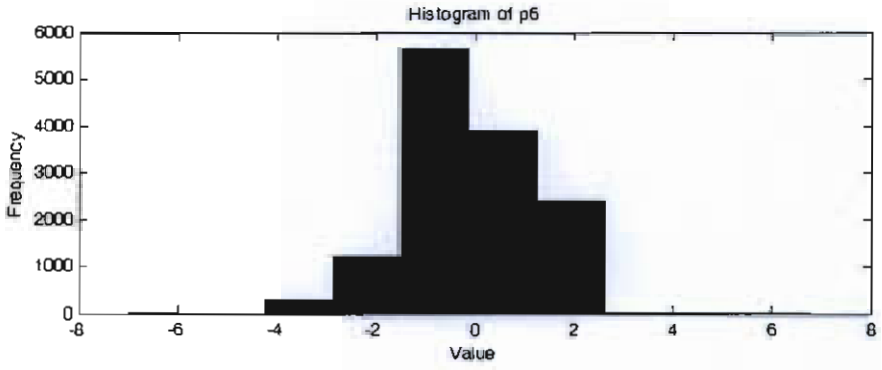
ภาพที่ 4-55 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 3 (p3)



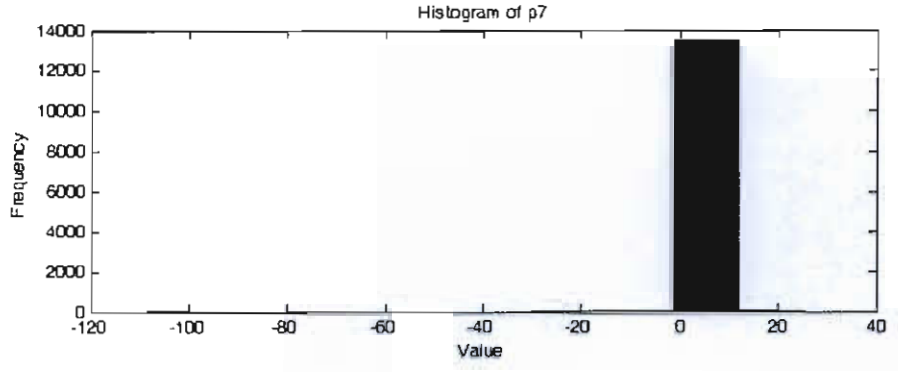
ภาพที่ 4-56 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 4 (p4)



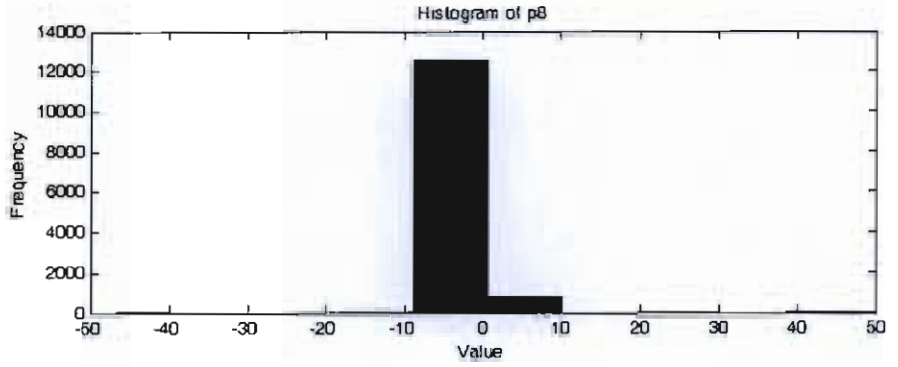
ภาพที่ 4-57 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 5 (p5)



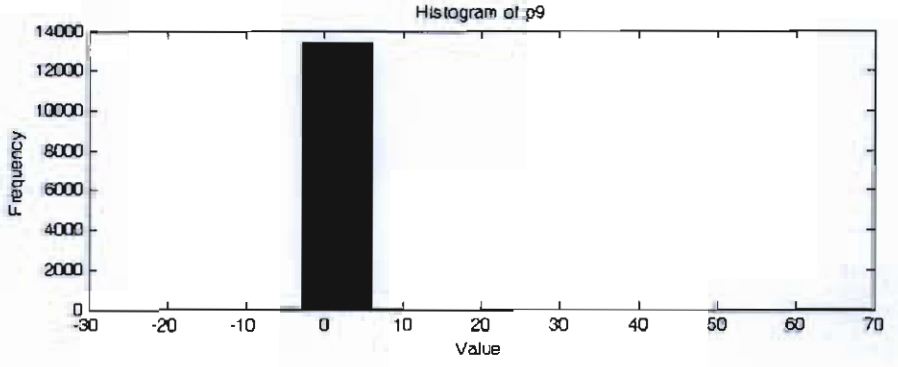
ภาพที่ 4-58 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 6 (p6)



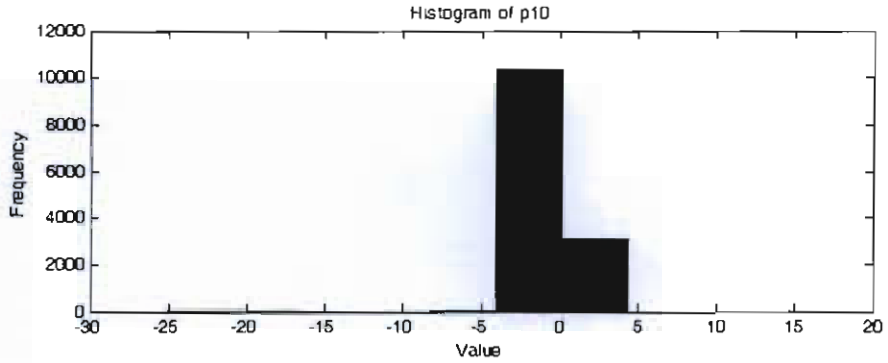
ภาพที่ 4-59 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 7 (p7)



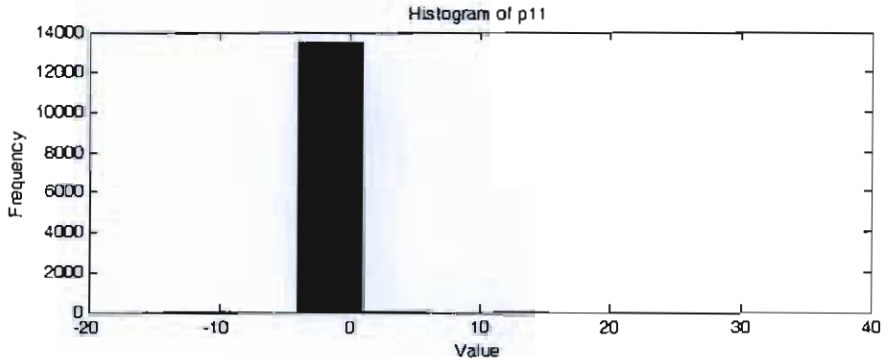
ภาพที่ 4-60 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 8 (p8)



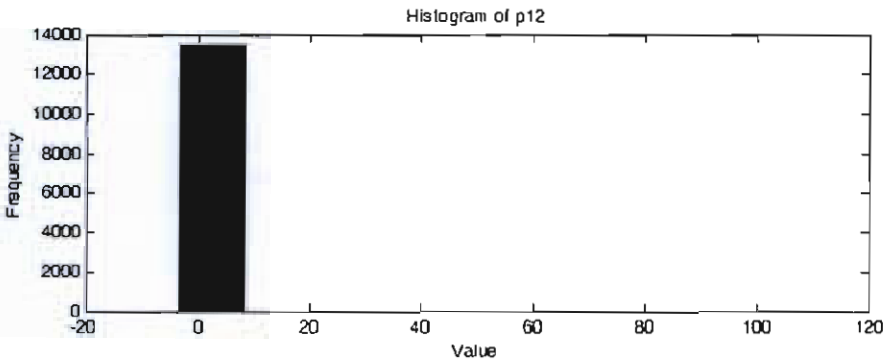
ภาพที่ 4-61 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 9 (p9)



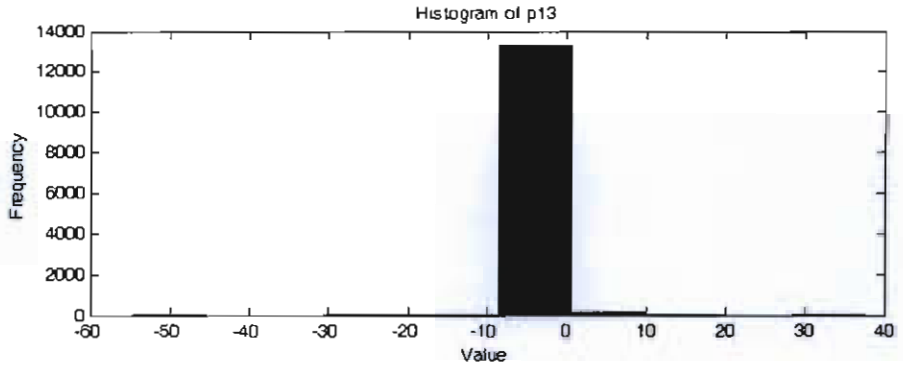
ภาพที่ 4-62 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 10 (p10)



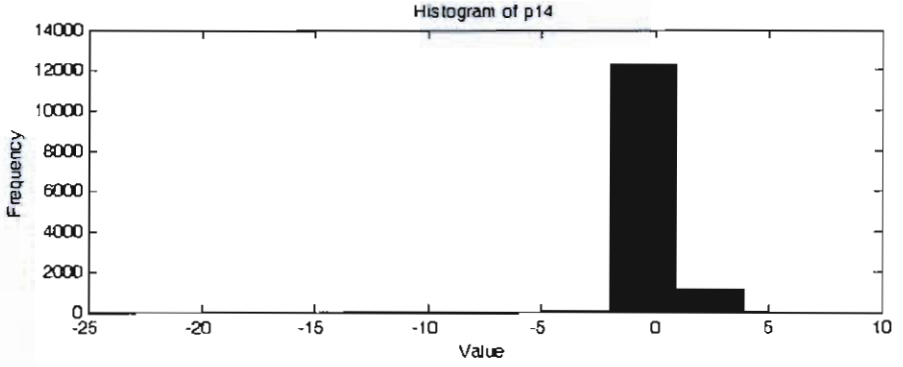
ภาพที่ 4-63 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 11 (p11)



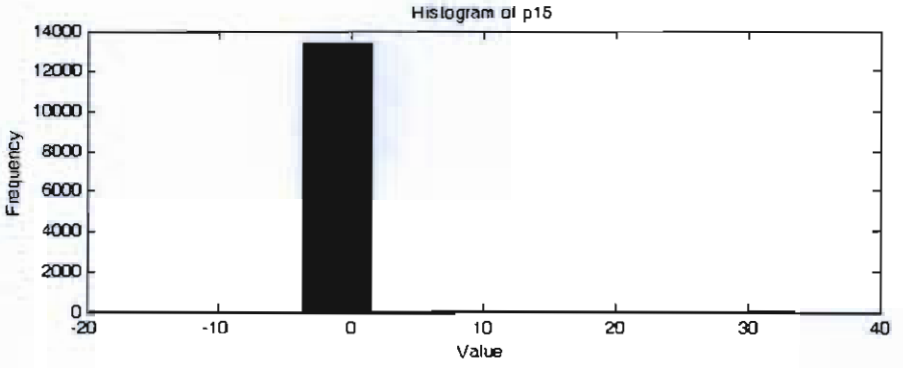
ภาพที่ 4-64 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 12 (p12)



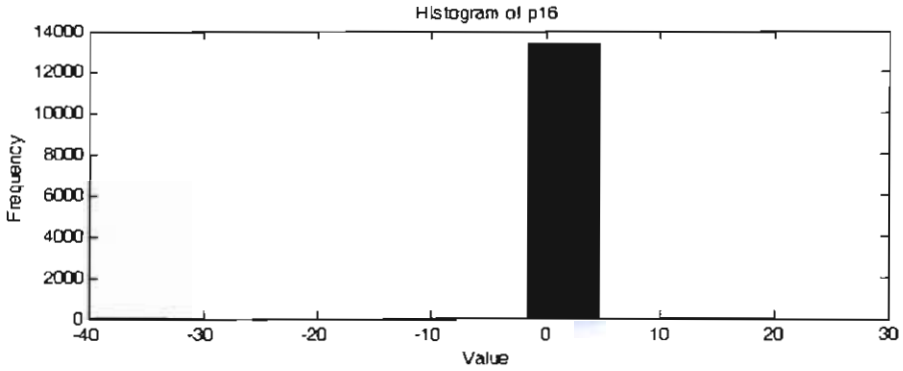
ภาพที่ 4-65 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 13 (p13)



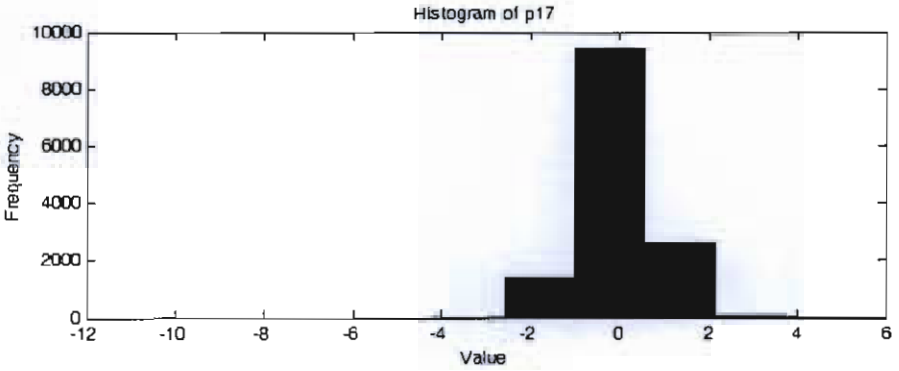
ภาพที่ 4-66 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 14 (p14)



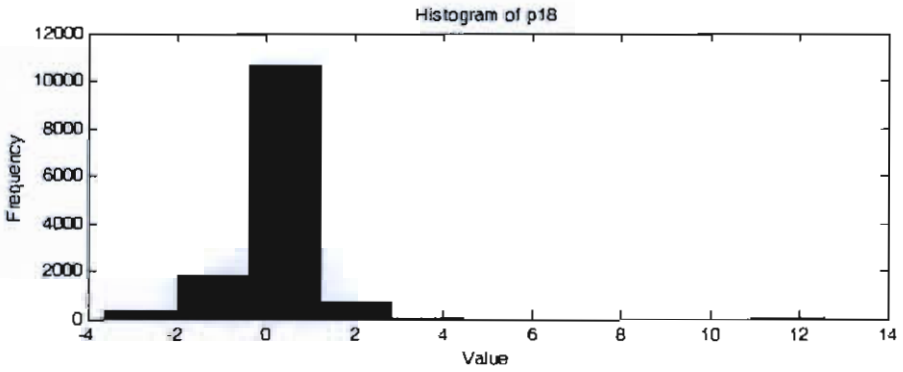
ภาพที่ 4-67 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 15 (p15)



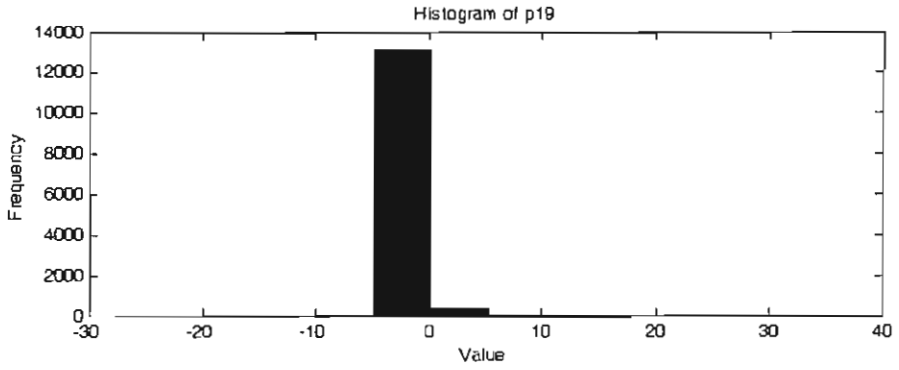
ภาพที่ 4-68 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 16 (p16)



ภาพที่ 4-69 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 17 (p17)



ภาพที่ 4-70 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 18 (p18)



ภาพที่ 4-71 Histogram ของลักษณะข้อมูล KDDcup99 ที่ผ่าน PCA ลักษณะที่ 19 (p19)

ลักษณะข้อมูล KDDCup99 เมื่อสกัดลักษณะด้วย Chi-Square จำนวน 26 ลักษณะ

เมื่อนำข้อมูลที่ผ่านขั้นตอนการเตรียมข้อมูล 34 ลักษณะ มาเลือกลักษณะด้วยวิธีฮิวริสติกกรีดดี ซึ่งเลือกลักษณะออกมาได้จำนวน 26 ลักษณะได้แก่ 1, 2, 5, 7, 8, 9, 12, 15, 16, 17, 18, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33 และ 34 โดยแต่ละลักษณะมีค่าสูงสุด ต่ำสุด ค่าเฉลี่ย และค่าส่วนเบี่ยงเบนมาตรฐานตามตารางที่ 4-21 ความสัมพันธ์ระหว่างแต่ละลักษณะทั้ง 10 ลักษณะแสดงในตารางที่ 4-22 โดยแต่ละลักษณะในลักษณะที่ 1 ถึง ลักษณะที่ 26 มีการกระจายตัวของข้อมูลดังภาพที่ 4-72 ถึง 4-97

ตารางที่ 4-21 ค่าทางสถิติของข้อมูล KDDcup99 จำนวน 26 ลักษณะ ที่เลือกลักษณะโดย Chi-Square

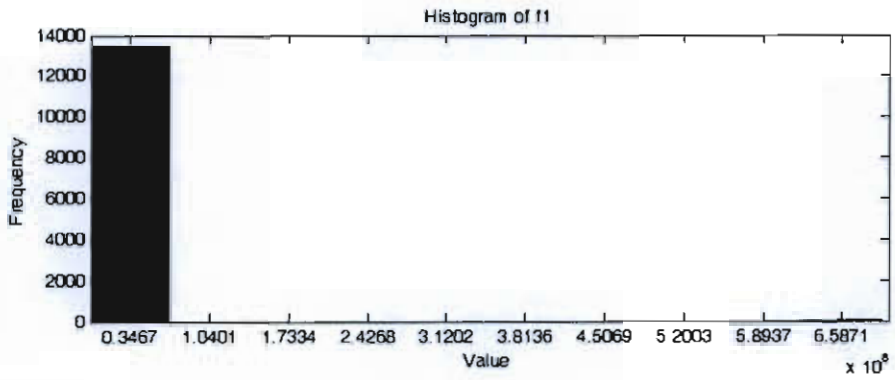
Features	Maximum	Minimum	Mean	Standard Deviation
f1	6.93E+08	0	74547.73	5977289
f2	5155468	0	7012.817	172422.5
f5	30	0	0.640121	4.039457
f7	1	0	0.301207	0.4588
f8	38	0	0.012742	0.426195
f9	1	0	0.002371	0.048632
f12	21	0	0.007704	0.243328
f15	1	0	0.024446	0.154436
f16	511	0	182.1567	229.4395
f17	511	0	118.6734	207.4745
f18	1	0	0.08962	0.265435
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 4-22 ค่าสหสัมพันธ์ระหว่างแต่ละลักษณะของข้อมูล KDDcup99 จำนวน 34 ลักษณะ

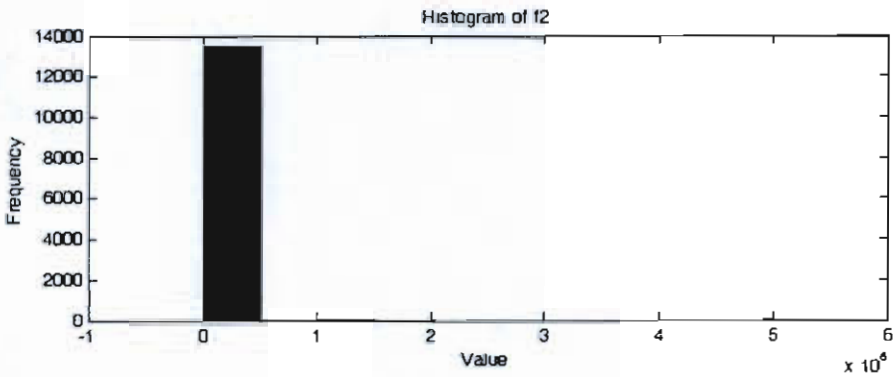
	f1	f2	f5	f7	f8	f9	f12	f15	f16	f17	f18	f20	f21
f1	1												
f2	-0.00051	1											
f5	0.002711	-0.00415	1										
f7	0.000182	-0.01356	0.238899	1									
f8	-0.00033	0.001165	0.031237	0.045538	1								
f9	-0.0006	-0.00046	0.02433	0.074247	0.384572	1							
f12	-0.00039	0.000419	0.035834	0.048228	0.249086	0.192534	1						
f15	-0.00194	-0.0043	0.956204	0.241114	-0.00473	-0.00772	0.046246	1					
f16	-0.00767	-0.03196	-0.12504	-0.50251	-0.02357	-0.03844	-0.02469	-0.12499	1				
f17	-0.0069	-0.02283	-0.08981	-0.34901	-0.01691	-0.0276	-0.01795	-0.08978	0.748042	1			
f18	0.021607	-0.01265	-0.05265	-0.2149	-0.00996	-0.01646	-0.00842	-0.05345	0.043823	-0.17812	1		
f20	-0.00192	-0.02154	-0.0804	-0.34329	-0.01563	-0.02389	-0.01678	-0.08207	0.112904	-0.29873	-0.07669	1	
f21	0.000707	-0.02078	-0.07831	-0.3316	-0.01509	-0.02306	-0.01545	-0.07983	0.10902	-0.28834	-0.07291	0.962791	1
f22	-0.01448	0.021448	0.082067	0.337956	0.015585	0.022837	0.014018	0.08247	-0.32361	0.290055	-0.55315	-0.4886	-0.4698
f23	0.005279	-0.01669	-0.06178	-0.25457	-0.01183	-0.01317	-0.01017	-0.0625	0.363572	-0.23339	0.060497	0.625324	0.601707
f24	-0.00392	-0.01212	-0.05824	-0.00216	-0.01076	-0.01308	-0.00772	-0.05768	-0.29435	-0.17942	-0.12361	-0.16304	-0.15638
f25	0.000666	-0.05889	0.059787	-0.36791	-0.02896	-0.05143	-0.02114	0.066741	0.540264	0.361069	0.212846	0.211832	0.204599
f26	-0.01321	-0.03416	-0.09659	0.189343	-0.02709	-0.04225	-0.03185	-0.10096	0.158605	0.54897	-0.369	-0.53852	-0.51905
f27	-0.0099	0.03004	-0.11794	0.315514	0.010634	0.021164	-0.01225	-0.12926	-0.02116	0.431606	-0.44102	-0.62938	-0.60671
f28	-0.0045	-0.02209	-0.0756	-0.33084	-0.01557	-0.02602	-0.01236	-0.07428	0.173425	-0.309	0.064028	0.770066	0.743525
f29	-0.00152	0.030534	-0.15413	-0.3764	0.003334	-0.00114	-0.01161	-0.16019	0.210185	0.554466	-0.27852	-0.16194	-0.15664
f30	-0.00266	-0.01464	-0.05722	-0.10666	-0.00692	0.0043	-0.00812	-0.05956	-0.29641	-0.19166	-0.12647	-0.10207	-0.09852
f31	0.001758	-0.01321	-0.0452	-0.21317	-0.00908	-0.01669	-0.0091	-0.04098	0.043235	-0.1809	0.987209	-0.07349	-0.07058
f32	0.016615	-0.01223	-0.0492	-0.20343	-0.00902	-0.01411	-0.00855	-0.05007	0.041601	-0.16748	0.925022	-0.06481	-0.15772
f33	-0.00546	-0.0216	-0.07768	-0.33863	-0.01468	-0.0184	-0.01367	-0.07883	0.12433	-0.30017	-0.07548	0.966496	0.93238
f34	0.000671	-0.02074	-0.07872	-0.32686	-0.01419	-0.02188	-0.01552	-0.07998	0.111673	-0.28867	-0.0732	0.955675	0.991991

ตารางที่ 4-22 (ต่อ)

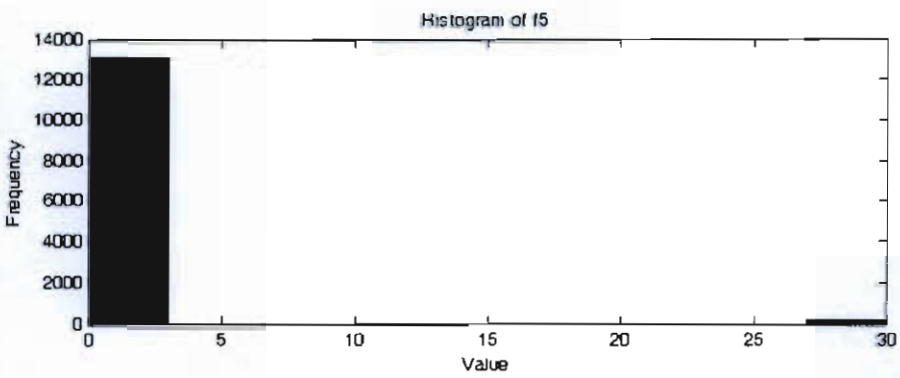
	f22	f23	f24	f25	f26	f27	f28	f29	f30	f31	f32	f33	f34
f22	1												
f23	-0.75729	1											
f24	0.194309	-0.14779	1										
f25	-0.35714	0.273199	-0.47505	1									
f26	0.595353	-0.47284	0.14542	-0.05286	1								
f27	0.721102	-0.57138	0.242756	-0.41479	0.824219	1							
f28	-0.53585	0.739235	-0.19628	0.310536	-0.627	-0.7527	1						
f29	0.450316	-0.31187	0.128222	-0.12002	0.163082	0.325947	-0.11822	1					
f30	0.197615	-0.15341	0.444948	-0.607	-0.00364	0.227313	-0.13214	0.275446	1				
f31	-0.55285	0.061445	-0.12521	0.216912	-0.37489	-0.44757	0.065054	-0.28513	-0.12804	1			
f32	-0.51985	0.061593	-0.11664	0.20146	-0.34884	-0.4166	0.062772	-0.26062	-0.11814	0.919704	1		
f33	-0.49571	0.616781	-0.165	0.216166	-0.53974	-0.63102	0.795291	-0.15713	-0.10221	-0.07551	-0.06513	1	
f34	-0.47419	0.606586	-0.16055	0.20836	-0.52247	-0.6116	0.749228	-0.15765	-0.09852	-0.07054	-0.15824	0.932962	1



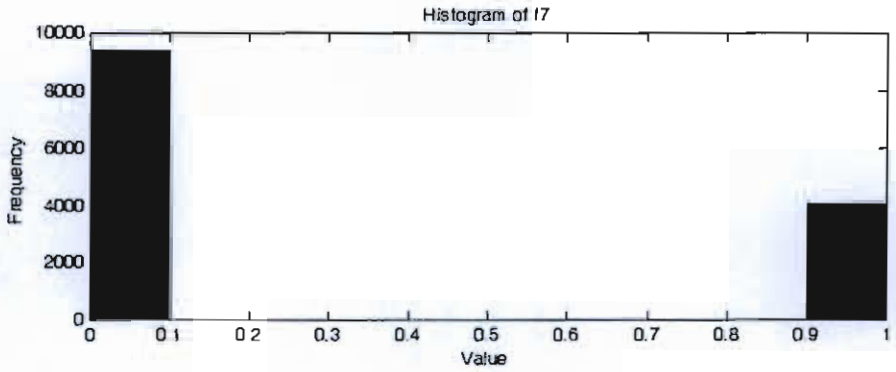
ภาพที่ 4-72 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 1 (f1)



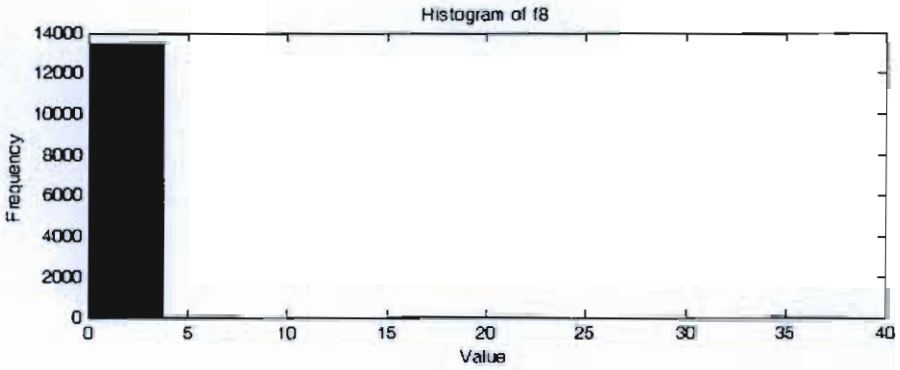
ภาพที่ 4-73 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 2 (f2)



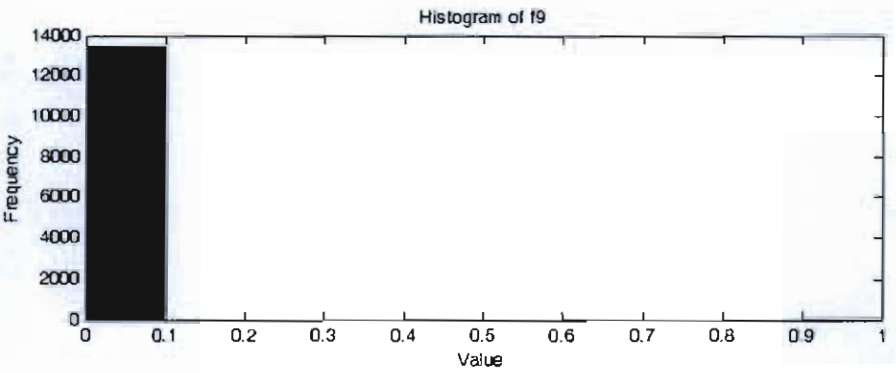
ภาพที่ 4-74 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 5 (f5)



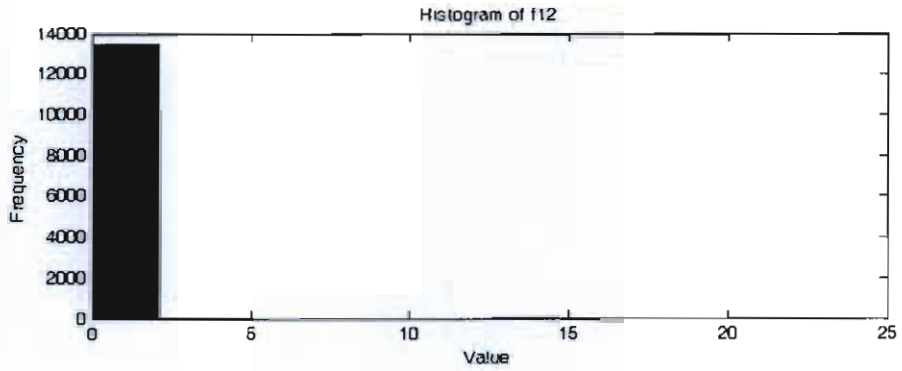
ภาพที่ 4-75 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 7 (f7)



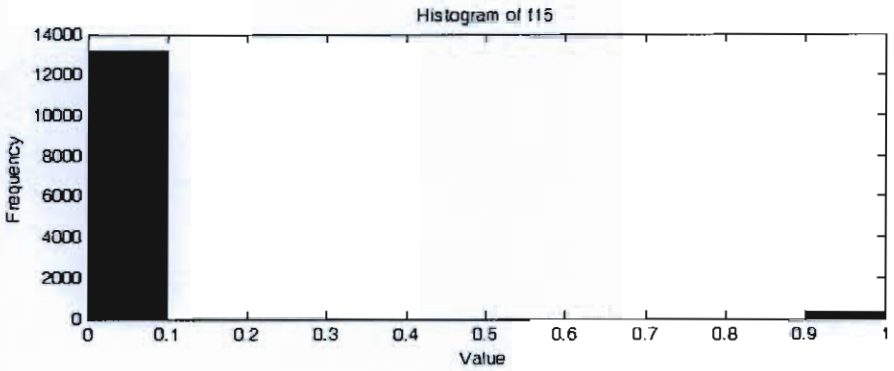
ภาพที่ 4-76 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 8 (f8)



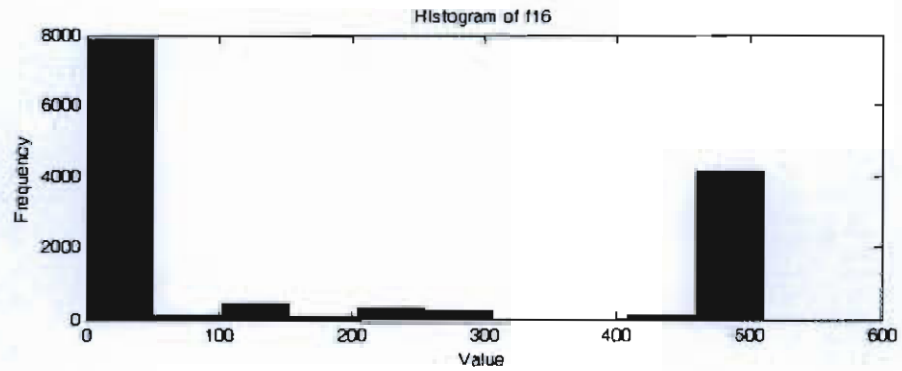
ภาพที่ 4-77 Histogram ของลักษณะข้อมูล KDDcup99ลักษณะที่ 9 (f9)



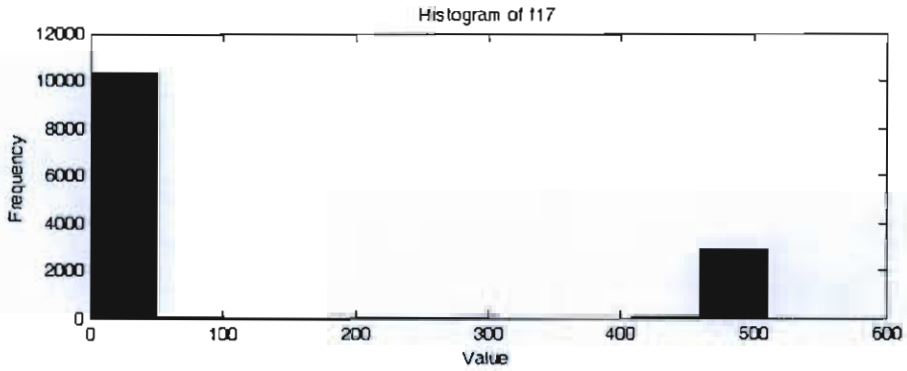
ภาพที่ 4-78 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 12 (f12)



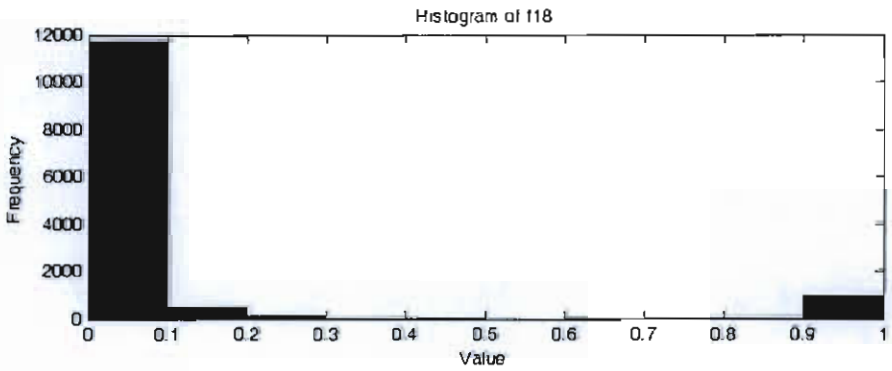
ภาพที่ 4-79 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 15 (f15)



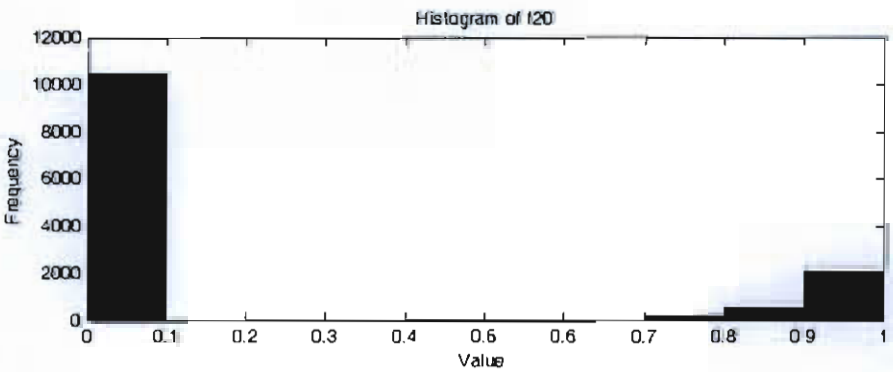
ภาพที่ 4-80 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 16 (f16)



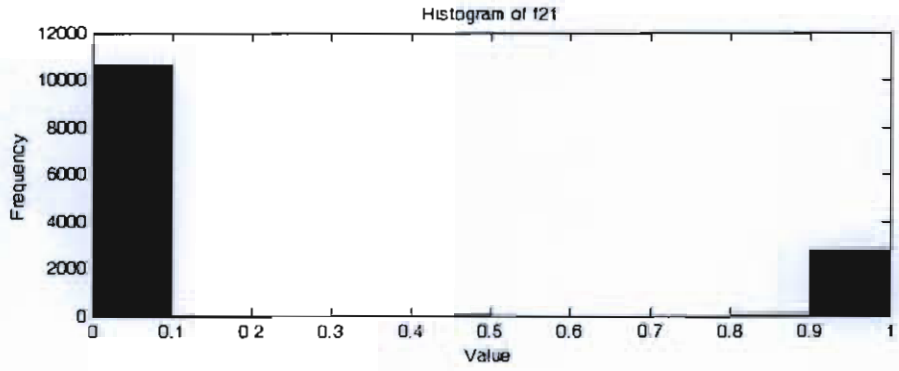
ภาพที่ 4-81 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 17 (f17)



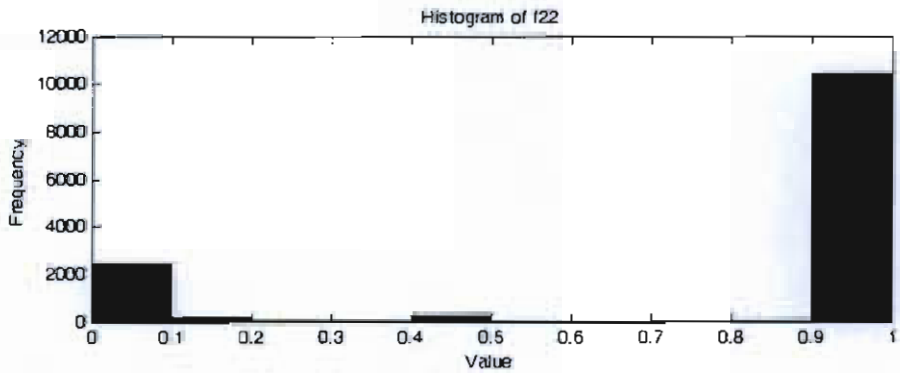
ภาพที่ 4-82 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 18 (f18)



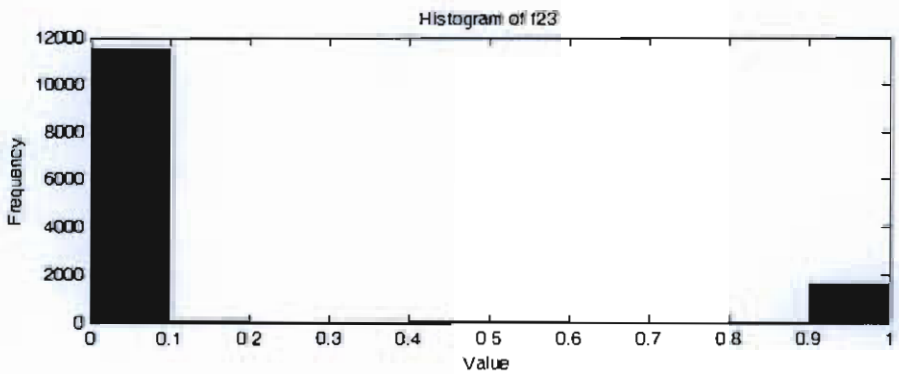
ภาพที่ 4-83 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 20 (f20)



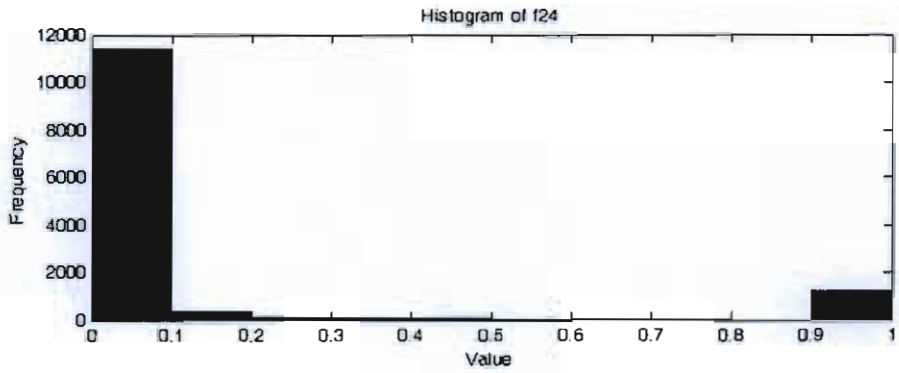
ภาพที่ 4-84 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 21 (f21)



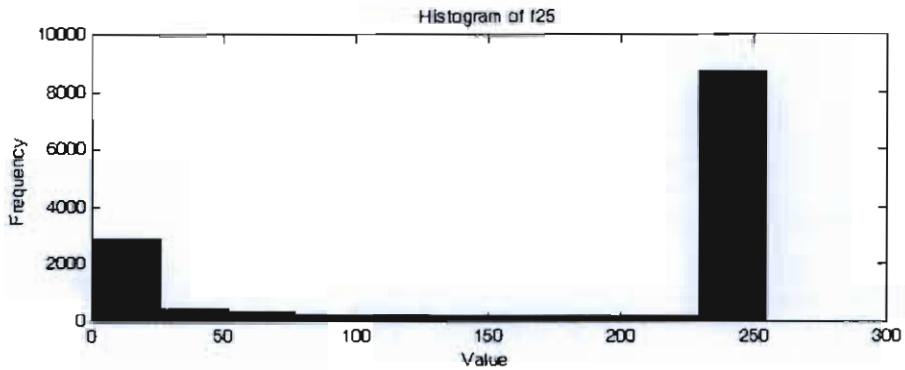
ภาพที่ 4-85 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 22 (f22)



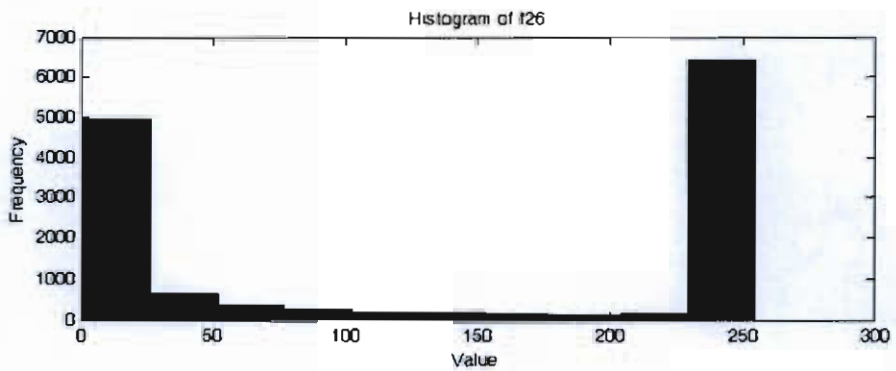
ภาพที่ 4-86 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 23 (f23)



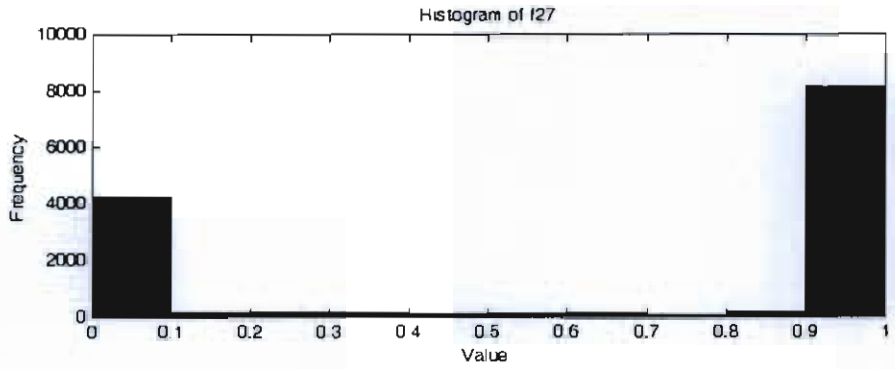
ภาพที่ 4-87 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 24 (f24)



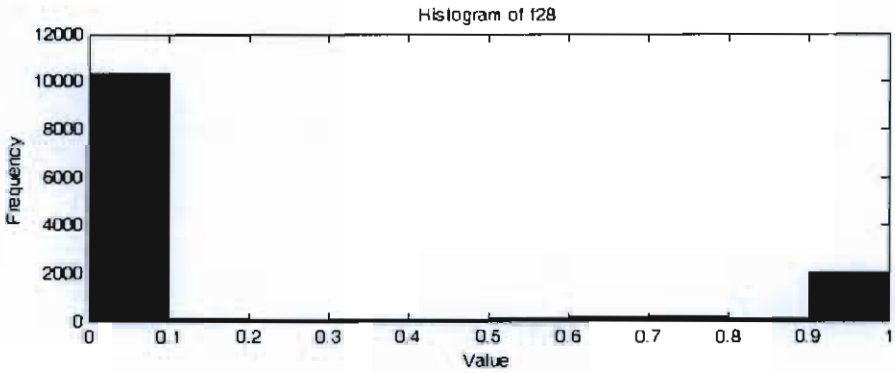
ภาพที่ 4-88 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 25 (f25)



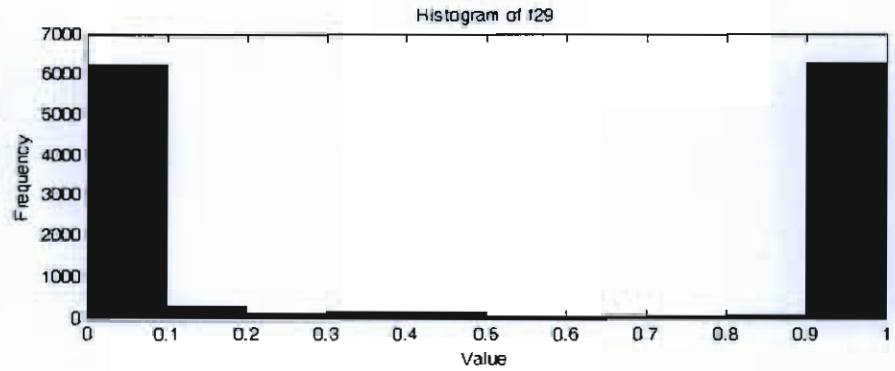
ภาพที่ 4-89 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 26 (f26)



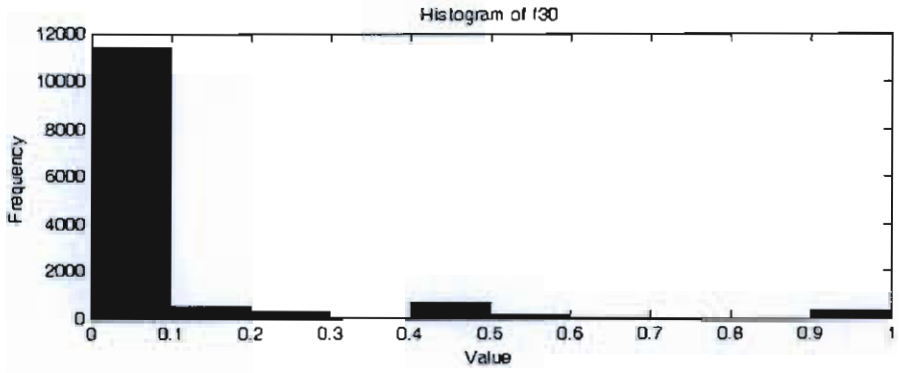
ภาพที่ 4-90 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 27 (f27)



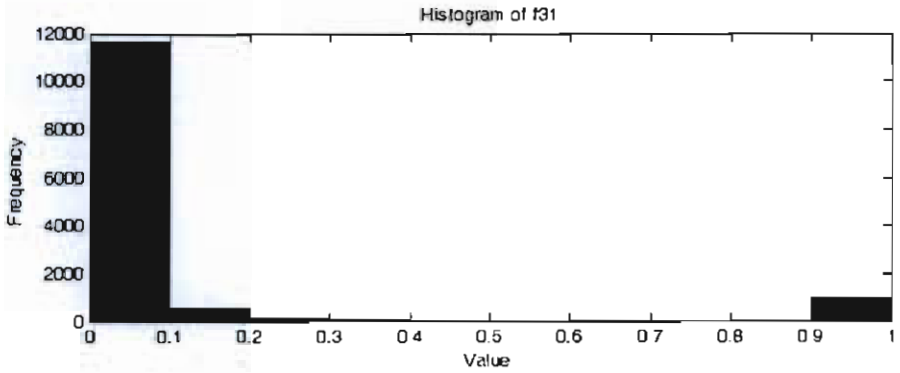
ภาพที่ 4-91 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 28 (f28)



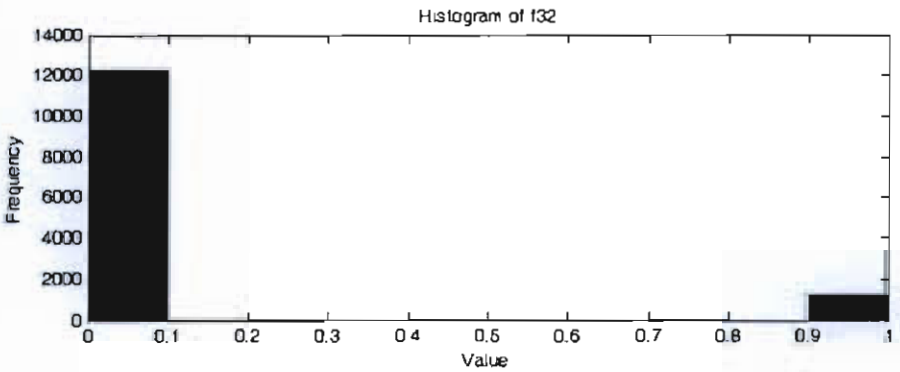
ภาพที่ 4-92 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 29 (f29)



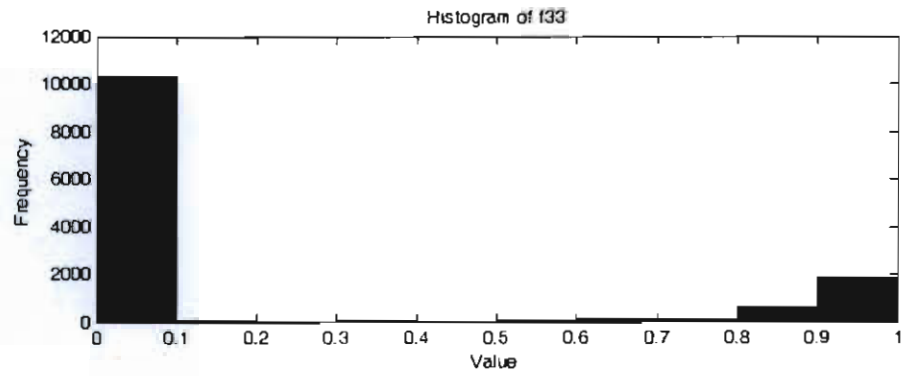
ภาพที่ 4-93 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 30 (f30)



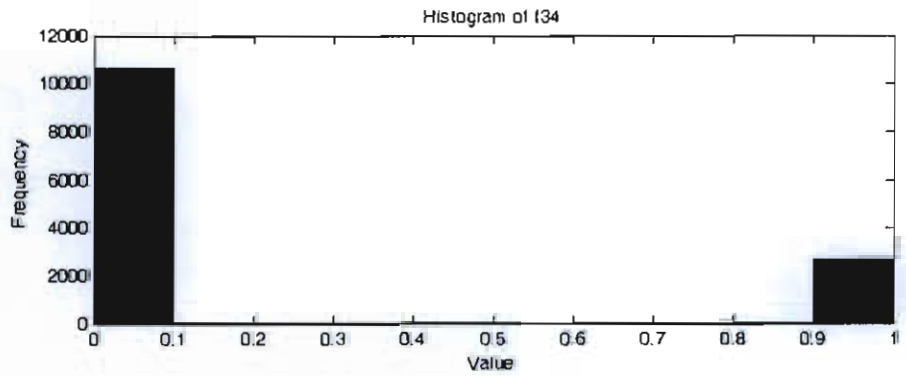
ภาพที่ 4-94 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 31 (f31)



ภาพที่ 4-95 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 32 (f32)



ภาพที่ 4-96 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 33 (f33)



ภาพที่ 4-97 Histogram ของลักษณะข้อมูล KDDcup99 ลักษณะที่ 34 (f34)

ผลการทดลอง

จากการทดลองการรู้จำด้วยโครงข่ายประสามเทียมแบบฟังก์ชันรัศมีฐานกับชุดข้อมูล KDDCup99 ด้วยการวัดประสิทธิภาพค่าความครบถ้วน ค่าความแม่นยำ ค่าเอฟเมเซอร์ และอัตราความผิดพลาดเชิงบวกของแต่ละคลาสคำตอบจำนวน 5 คลาส โดยแบ่งออกเป็น 5 การทดลอง ได้แก่ ชุดข้อมูล KDDCup99 ทั้งหมด เลือกลักษณะด้วย HGIS2 สกัดลักษณะด้วย PCA และ เลือกลักษณะด้วย Chi-Square มีผลการทดลองดังนี้

ตารางที่ 4-23 ผลการทดลองกับชุดข้อมูล KDDCup99

Class	Recall	Precision	F-measure	FAR
DoS	93.4597	99.0927	96.1938	0.3719
Normal	83.5138	95.3952	89.0600	1.7924
Probe	95.6522	88.8005	92.0991	5.2296
R2L	92.8571	67.3139	78.0488	4.0792
U2R	61.9048	67.3139	64.4961	0.2231
Weighted Avg.	90.8889	90.8889	91.1313	2.5851

ผลการทดลองกับชุดข้อมูล KDDCup99 ทั้งหมด จำนวน 34 ลักษณะ ของแต่ละคลาส 5 คลาส ดังตารางที่ 4-23 โดยผลเฉลี่ยค่าความครบถ้วน ค่าความแม่นยำ ค่าเอฟเมเซอร์ และอัตราความผิดพลาดเชิงบวก 90.8889% 90.8889% 91.1313% และ 2.5851% ตามลำดับ

ตารางที่ 4-24 ผลการทดลองเลือกลักษณะด้วย HGIS2 กับชุดข้อมูล KDDCup99

Class	Recall	Precision	F-measure	FAR
DoS	98.7783	99.2025	98.9899	0.3455
Normal	90.1324	95.7801	92.8705	1.7657
Probe	98.1017	90.6621	94.2353	4.3801
R2L	91.0515	95.7647	93.3486	0.3634
U2R	66.6667	95.7647	78.6094	0
Weighted Avg.	95.1482	95.2685	95.1217	2.0028

ผลการทดลองกับชุดข้อมูล KDDCup99 ที่ผ่านการเลือกลักษณะด้วย HGIS2 จำนวน 10 ลักษณะ ของแต่ละคลาส 5 คลาส ดังตารางที่ 4-24 โดยผลเฉลี่ยค่าความครบถ้วน ค่าความแม่นยำ ค่าเอฟเมเซอร์ และอัตราความผิดพลาดเชิงบวก 95.1482% 95.2685% 95.1217% และ 2.0028% ตามลำดับ

ตารางที่ 4-25 ผลการทดลองสกัดลักษณะด้วย PCA กับชุดข้อมูล KDDCup99

Class	Recall	Precision	F-measure	FAR
DoS	93.4597	0.9909	96.1938	0.3719
Normal	91.4561	0.9389	92.6547	2.6485
Probe	95.2235	0.8553	90.1188	6.9817
R2L	75.4464	0.8325	79.1569	1.3732
U2R	47.6191	0.8325	60.5843	0.0744
Weighted Avg.	91.7037	0.9201	91.7155	3.1534

ผลการทดลองกับชุดข้อมูล KDDCup99 ที่การผ่านการสกัดลักษณะด้วย PCA จำนวน 19 ลักษณะ ของแต่ละคลาส 5 คลาส ดังตารางที่ 4-25 โดยผลเฉลี่ยค่าความครบถ้วน ค่าความแม่นยำ ค่าเอฟเมเซอร์ และอัตราความผิดพลาดเชิงบวก 91.7037% 0.9201% 91.7155% และ 3.1534% ตามลำดับ

ตารางที่ 4-26 ผลการทดลองเลือกลักษณะด้วย Chi-Square กับชุดข้อมูล KDDCup99

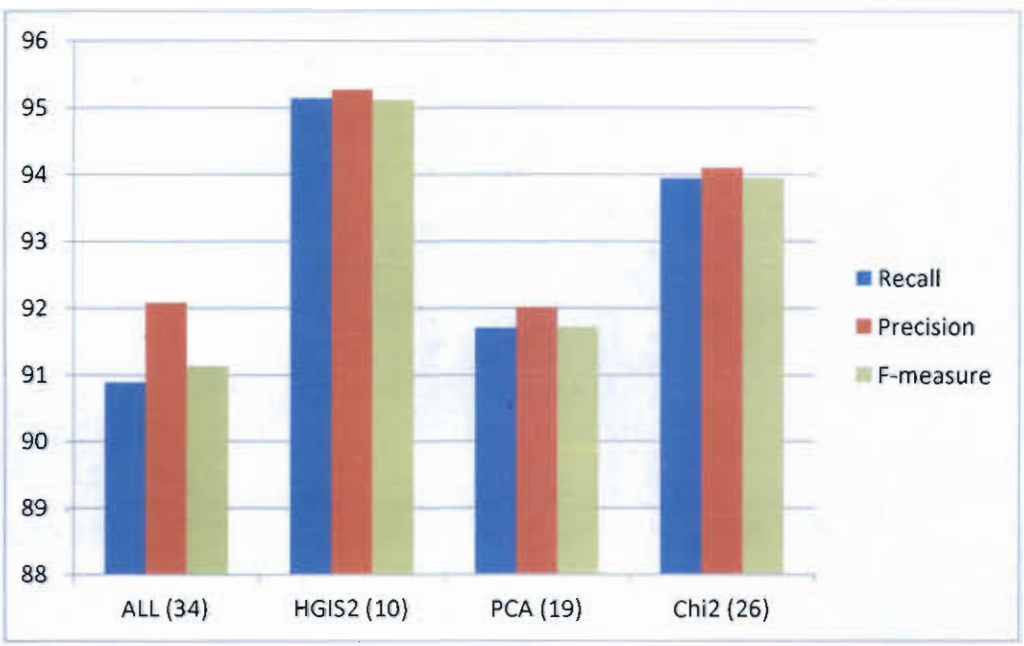
Class	Recall	Precision	F-measure	FAR
DoS	93.1540	98.8968	95.9396	0.4516
Normal	96.6907	92.3563	94.4738	3.5581
Probe	94.8561	92.2024	93.5104	3.4776
R2L	85.2679	90.0943	87.6147	0.8481
U2R	52.3810	90.0943	66.2463	0.0744
Weighted Avg.	93.9444	94.0948	93.9477	2.3542

ผลการทดลองกับชุดข้อมูล KDDCup99 ที่ผ่านการเลือกลักษณะด้วย Chi-Square จำนวน 26 ลักษณะ ของแต่ละคลาส 5 คลาส ดังตารางที่ 4-26 โดยผลเฉลี่ยค่าความครบถ้วน ค่าความแม่นยำ ค่าเอฟเมเซอร์ และอัตราความผิดพลาดเชิงบวก 93.9444% 94.0948% 93.9477% และ 2.3542% ตามลำดับ

ตารางที่ 4-27 ผลเฉลี่ยการทดลองกับชุดข้อมูล KDDCup99

	Recall	Precision	F-measure	FAR
ALL (34)	90.8889	92.0822	91.1313	2.5851
HGIS2 (10)	95.1482	95.2685	95.1217	2.0028
PCA (19)	91.7037	92.0137	91.7155	3.1534
Chi^2 (26)	93.9444	94.0948	93.9477	2.3542

ดังนั้นเมื่อเปรียบเทียบผลเฉลี่ยในการวัดประสิทธิภาพด้วยวิธีต่าง ๆ ของการเลือกลักษณะด้วย HGIS2 Chi-Square การสกัดลักษณะด้วย PCA และกับชุดข้อมูล KDDCup99 ทั้งหมด ผลดังตารางที่ 4-27 และภาพที่ 4-98 แสดงให้เห็นว่า การเลือกลักษณะด้วย HGIS2 ได้ผลดีที่สุด โดยได้ผลที่ดีขึ้นกว่าชุดข้อมูลทั้งหมด ทั้งค่าความครบถ้วน ค่าความแม่นยำ ค่าเอฟเมเซอร์ และอัตราความผิดพลาดเชิงบวก 4.2593% 3.1863% 3.9904% และ 0.5823% ตามลำดับ



ภาพที่ 4-98 กราฟผลเฉลี่ยการทดลองกับชุดข้อมูล KDDCup99

ตารางที่ 4-28 ผลค่าความถูกต้องและเวลาที่ใช้กับชุดข้อมูล KDDCup99

	Accuracy	Processing Times (s)
ALL (34)	90.8889	48.1209
HGIS2 (10)	95.1482	22.4036
PCA (19)	91.7037	29.8472
Chi^2 (26)	93.9444	39.2032

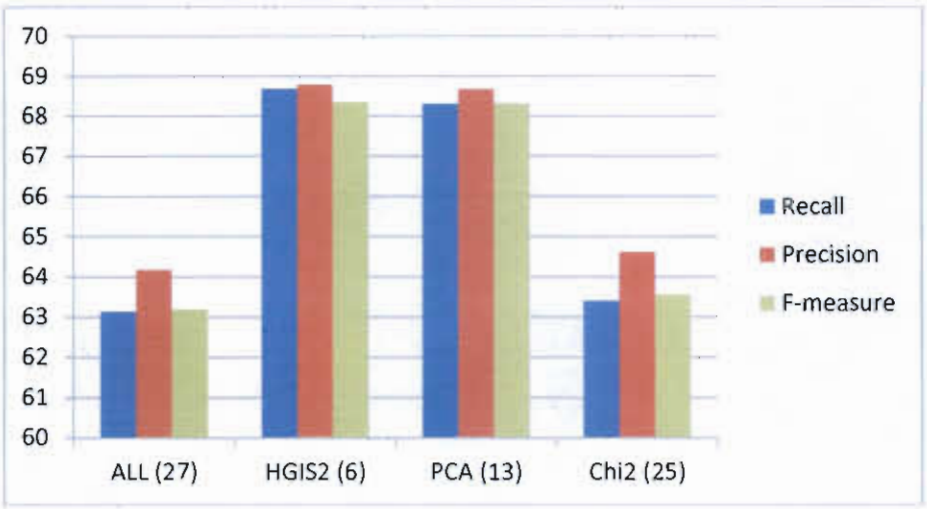
จากตารางที่ 4-28 แสดงผลค่าความถูกต้องและเวลาที่ใช้ประมวลผล กับชุดข้อมูล KDDCup99 กับวิธีการเลือกลักษณะวิธีต่าง ๆ โดยการเลือกลักษณะด้วย HGIS2 ได้ค่าความถูกต้อง 95.1482% ซึ่งดีกว่าวิธีการอื่น ๆ และใช้เวลาในการประมวลผลน้อยที่สุดคือ 22.4036 วินาที

นอกจากนี้ เพื่อให้มีความแน่ใจในการทดลองการเลือกลักษณะด้วยวิธี HGIS2 ว่ามีประสิทธิภาพกับลักษณะชุดข้อมูลอื่น ๆ อย่างไร จึงได้ทดลองกับชุดข้อมูลอื่น ๆ ด้วย ได้แก่ ชุดข้อมูล Statlog และ ชุดข้อมูล Faults โดยรายละเอียดของชุดข้อมูลเหล่านี้จะกล่าวใน ภาคผนวก ก

ตารางที่ 4-29 ผลเฉลี่ยการทดลองกับชุดข้อมูล Statlog

	Recall	Precision	F-measure	FAR
ALL (36)	83.4918	84.6382	85.8809	3.0161
HGIS2 (5)	86.2690	86.0484	86.1240	2.9685
PCA (6)	86.0885	85.7509	85.8809	3.0099
Chi^2 (36)	83.4918	84.6382	83.9270	3.0161

เมื่อเปรียบเทียบผลเฉลี่ยในการวัดประสิทธิภาพด้วยวิธีต่าง ๆ กับชุดข้อมูล Statlog ทั้งหมด จำนวน 36 ลักษณะ เลือกลักษณะด้วย HGIS2 ได้จำนวน 5 ลักษณะ สกัดลักษณะด้วย PCA ได้จำนวน 6 ลักษณะ และเลือกลักษณะด้วย Chi-Square ได้จำนวน 36 ลักษณะ ได้ผลการทดลองดังตารางที่ 4-29 และภาพที่ 4-99 แสดงให้เห็นว่า การเลือกลักษณะด้วย HGIS2 ได้ผลดีที่สุด โดยได้ผลที่ดีขึ้นกว่าชุดข้อมูลทั้งหมด ทั้งค่าความครบถ้วน ค่าความแม่นยำ ค่าเอฟเมเซอร์ และอัตราความผิดพลาดเชิงบวก 2.7772% 1.4102% 0.2431% และ 0.0476% ตามลำดับ แต่ดีกว่าการสกัดลักษณะด้วย PCA เพียงเล็กน้อย



ภาพที่ 4-99 กราฟผลเฉลี่ยการทดลองกับชุดข้อมูล Statlog

ตารางที่ 4-30 ผลค่าความถูกต้องและเวลาที่ใช้กับชุดข้อมูล Statlog

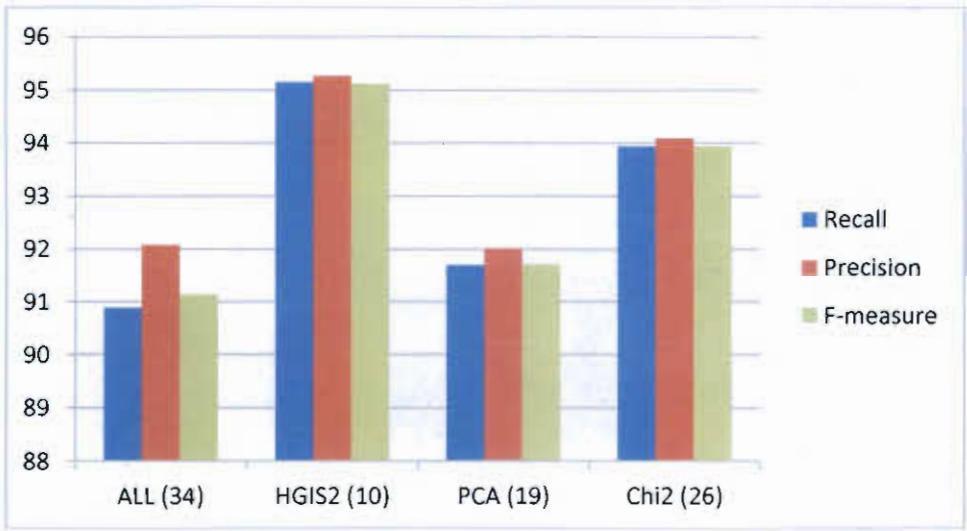
	Accuracy	Processing Times (s)
ALL (36)	83.6830	27.1133
HGIS2 (5)	86.4624	9.0417
PCA (6)	86.3248	10.1707
Chi^2 (36)	83.6830	27.1133

จากตารางที่ 4-30 แสดงผลค่าความถูกต้องและเวลาที่ใช้ประมวลผล กับชุดข้อมูล Statlog กับวิธีการเลือกลักษณะวิธีต่าง ๆ โดยการเลือกลักษณะด้วย HGIS2 ได้ค่าความถูกต้อง 86.4624% ซึ่งดีกว่าวิธีการอื่น ๆ และดีกว่ากับชุดข้อมูลทั้งหมด 2.7794% แต่ดีกว่าเลือกลักษณะด้วย PCA เพียง 0.1376% ส่วนใช้เวลาในการทดสอบการรู้จำน้อยที่สุดคือ 9.0417 วินาที ซึ่งได้จากการเลือกลักษณะวิธี HGSI2 จำนวน 5 ลักษณะ จากทั้งหมด 36 ทั้งลักษณะ

ตารางที่ 4-31 ผลเฉลี่ยการทดลองกับชุดข้อมูล Faults

	Recall	Precision	F-measure	FAR
ALL (27)	63.1443	64.1748	63.1950	14.5364
HGIS2 (6)	68.6856	68.7899	68.3508	12.6586
PCA (13)	68.2990	68.6743	68.3201	14.2684
Chi^2 (25)	63.4021	64.6073	63.5542	14.2997

เมื่อเปรียบเทียบผลเฉลี่ยในการวัดประสิทธิภาพด้วยวิธีต่าง ๆ กับชุดข้อมูล Faults ทั้งหมด จำนวน 27 ลักษณะ เลือกลักษณะด้วย HGIS2 ได้จำนวน 6 ลักษณะ สกัดลักษณะด้วย PCA ได้จำนวน 13 ลักษณะ และเลือกลักษณะด้วย Chi-Square ได้จำนวน 25 ลักษณะ ได้ผลการทดลองดังตารางที่ 4-31 และภาพที่ 4-100 แสดงให้เห็นว่า การเลือกลักษณะด้วย HGIS2 ได้ผลดีที่สุด โดยได้ผลที่ดีขึ้นกว่าชุดข้อมูลทั้งหมด ทั้งค่าความครบถ้วน ค่าความแม่นยำ ค่าเอฟเมเซอร์ และอัตราความผิดพลาดเชิงบวก 5.5413% 4.6151% 5.1558% และ 1.8778% ตามลำดับ แต่ดีกว่า PCA เพียงเล็กน้อย



ภาพที่ 4-100 กราฟผลเฉลี่ยการทดลองกับชุดข้อมูล Faults

ตารางที่ 4-32 ผลค่าความถูกต้องและเวลาที่ใช้กับชุดข้อมูล Faults

	Accuracy	Processing Times
ALL (27)	63.1443	8.7223
HGIS2 (6)	68.6856	5.5236
PCA (13)	68.2990	6.5685
Chi^2 (25)	63.4021	7.5577

จากตารางที่ 4-32 แสดงผลค่าความถูกต้องและเวลาที่ใช้ประมวลผล กับชุดข้อมูล Faults กับวิธีการเลือกลักษณะวิธีต่าง ๆ โดยการเลือกลักษณะด้วย HGIS2 ได้ค่าความถูกต้อง 68.6856% ซึ่งดีกว่าวิธีการอื่น ๆ และดีกว่ากับชุดข้อมูลทั้งหมด 5.5413% แต่ดีกว่าเลือกลักษณะด้วย PCA เพียง 0.3866% ส่วนเวลาที่ใช้ในการทดสอบการรู้จำน้อยที่สุดคือ 5.5236 วินาที ซึ่งได้จากการเลือกลักษณะวิธี HGSI2 จำนวน 6 ลักษณะ จากทั้งหมด 27 ทั้งลักษณะ

บทที่ 5
สรุปและอภิปรายผล

สรุปผลการทดลอง

ในงานวิจัยนี้เสนอวิธีการหาตัวแทนจากชุดข้อมูลบนเครือข่ายที่เหมาะสมเพื่อระบุผู้บุกรุกแบบเวลาจริงกับชุดข้อมูล KDDcup99 จำนวนประมาณ 4,900,000 จุดข้อมูล 41 ลักษณะ ซึ่งได้นำข้อมูล 10% ของข้อมูลทั้งหมดสุ่มออกมาอีกจำนวน 13,499 จุดข้อมูล เพื่อสะดวกในการทดลอง จากนั้นตัดลักษณะที่ไม่มีผลต่อการทดลองออกไปจึงเหลือลักษณะจำนวน 34 ลักษณะ และหาตัวแทนของข้อมูลที่เหมาะสม เพื่อลดความซ้ำซ้อนของข้อมูลและเพิ่มประสิทธิภาพในการระบุผู้บุกรุก ในงานวิจัยนี้นำเสนอวิธีการเลือกลักษณะด้วยฮิวริสติกรีดดีอัลกอริทึมโดยใช้หลักการ Apriori ซึ่งได้ทำการทดลองหาฟังก์ชันผิดพลาดที่เหมาะสมสำหรับวิธีการนี้ โดยแบ่งออกเป็น 4 ฟังก์ชัน เมื่อทดลองได้ฟังก์ชันที่เหมาะสมสำหรับวิธีการลักษณะด้วยฮิวริสติกรีดดีอัลกอริทึมแล้ว จึงเปรียบกับวิธีการสกัดลักษณะและวิธีการเลือกลักษณะวิธีอื่น ๆ ซึ่งในวิทยานิพนธ์นี้ได้เปรียบเทียบกับอีก 2 วิธี ได้แก่ การสกัดลักษณะด้วยวิธีการวิเคราะห์องค์ประกอบหลัก และการเลือกลักษณะด้วยค่าสถิติไคสแควร์ จากนั้นนำลักษณะที่ได้นั้นเข้าสู่กระบวนการรู้จำเพื่อระบุผู้บุกรุกด้วยโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐานเพื่อทดสอบประสิทธิภาพในด้านของอัตราค่าความถูกต้อง ค่าความครบถ้วน ค่าความแม่นยำ ค่าเอฟเมเซอร์ อัตราความผิดพลาดเชิงบวก และเวลาที่ใช้ในการรู้จำ

1. การหาฟังก์ชันผิดพลาดสำหรับการเลือกลักษณะด้วยฮิวริสติกกรีดดี

จากการทดลองหาฟังก์ชันผิดพลาดเพื่อเป็นเกณฑ์สำหรับการเลือกลักษณะด้วยวิธีการฮิวริสติกกรีดดีโดยใช้ชุดข้อมูล KDDCup99 ในการทดลอง และใช้การรู้จำแบบโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐานนั้น แสดงให้เห็นว่าการหาลักษณะที่คัดจากเกณฑ์ที่มีจำนวนที่ตอบผิดน้อยให้ผลลัพธ์ที่ดีกว่าฟังก์ชันอื่นๆที่ได้ทำการทดลองทั้ง 4 ฟังก์ชัน เนื่องจากหากใช้ rmse จะมีผลกับจำนวนของคลาส หากจำนวนคลาสมากจะทำให้การผลต่างของความผิดพลาดนั้นสูง ซึ่งไม่เหมาะกับการคัดเลือกลักษณะด้วยวิธีการฮิวริสติกกรีดดี หากใช้ผลรวมของ rmse และจำนวนที่ตอบผิดจะได้ผลเช่นเดียวกับจำนวนที่ตอบอย่างเดียว ไม่มีความแตกต่างกันเนื่องจากจำนวนที่ตอบผิดมีค่ามาก แต่ค่า rmse ที่ได้มีค่าน้อยมาก ๆ เมื่อมารวมกันแล้วจึงไม่มีผลใด ๆ ที่จะทำให้ค่าความผิดพลาดมีความแตกต่างกันอย่างชัดเจน ส่วน rmse คูณกับจำนวนที่จำแนกผิดพลาดทำให้เกิดค่าความผิดพลาดที่แปรปรวนได้

2. การเปรียบเทียบเลือกลักษณะด้วยฮิวริสติกกรีดดีกับวิธีการอื่น ๆ

การทดลองหาตัวแทนข้อมูลด้วยการเลือกลักษณะโดยวิธีฮิวริสติกกรีดดีโดยใช้หลักการ Apriori ซึ่งเป็นวิธีการหาลักษณะที่มีความสำคัญในการตรวจจับผู้บุกรุกโดยไม่มีการเปลี่ยนแปลงข้อมูลใด ๆ แต่จะเลือกลักษณะบางลักษณะจากลักษณะทั้งหมด 34 ลักษณะ จากชุดข้อมูล KDDCup99 ที่ผ่านขั้นตอนการเตรียมข้อมูลเรียบร้อยแล้ว โดยวิธีการเลือกนั้นเป็นขั้นตอนวิธีการแก้ปัญหาที่คิดแบบง่าย ๆ โดยจะพิจารณาว่าข้อมูลที่มีอยู่ในขณะนั้นมีทางเลือกใดที่ให้คำตอบที่ดีที่สุดสามารถทำการค้นหาคำตอบจาก

ข้อมูลที่มีขนาดใหญ่มาก ๆ ได้ เพราะเป็นการค้นหาคำตอบที่ไม่ต้องดูข้อมูลทุกตัวซึ่งเรียกวิธีนี้ว่าฮิวริสติก โดยจะแบ่งออกเป็น 2 ขั้นตอน ได้แก่ การหาลักษณะฐาน และการเติมลักษณะ ในขั้นตอนแรกจะสร้างไอเท็มเซตหรือชุดลักษณะที่เป็นไปได้ และในการทดลองนั้นเราจะนำหลักการ Apriori มาช่วยในการลดจำนวนการสร้างไอเท็มเซตลง เนื่องจากสร้างไอเท็มเซตจากการใช้เซตที่มีขนาดใหญ่ที่หาได้ในขั้นตอนก่อนหน้า ซึ่งจะนำแต่ละเซตของลักษณะมาหาค่า โดยใช้โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐานในการคัดเลือกเพื่อที่จะสร้างไอเท็มเซตถัดไป การสร้างไอเท็มเซตได้นั้นทุก ๆ เซตย่อยจะต้องมีจำนวนที่ตอบผิดน้อยกว่าเซตที่กำลังสร้าง จากนั้นสร้างไอเท็มเซตจนกว่าจะไม่มีสามารถสร้างได้อีก และเลือกไอเท็มเซตหรือชุดลักษณะที่มีจำนวนตอนผิดน้อยที่สุดมาใช้เป็นฐานสำหรับเติมลักษณะในขั้นต่อไป ขั้นตอนที่ 2 จะเป็นการเติมลักษณะที่หลุดออกไปเข้ามาใหม่ เนื่องจากการคัดเลือกลักษณะในขั้นตอนแรกอาจจะมีลักษณะที่สำคัญหลุดออกไปได้ ซึ่งจากการทดลองสามารถเลือกลักษณะออกมากได้จำนวน 10 ลักษณะ จากทั้งหมด 34 ลักษณะ และจะเห็นได้ว่าผลลัพธ์ที่ได้เป็นที่น่าพอใจ ได้ผลลัพธ์ที่ดีกว่าการสกัดลักษณะด้วยวิธีวิเคราะห์องค์ประกอบหลัก เลือกลักษณะด้วยค่าสถิติไคสแควร์ และกับข้อมูลทั้งหมด

จากการทดลองด้วยการสกัดลักษณะโดยวิธีการวิเคราะห์องค์ประกอบหลัก ซึ่งจะใช้หลักความสัมพันธ์เชิงเส้นระหว่างตัวแปร โดยการผสมเชิงเส้นตรงได้องค์ประกอบหลักที่สามารถอธิบายความแปรปรวนของชุดข้อมูลได้มากที่สุดเป็นอันดับหนึ่ง และองค์ประกอบที่สามารถอธิบายความแปรปรวนของชุดข้อมูลได้มากที่สุดอันดับสอง โดยที่ทั้งสองนี้ไม่มีความสัมพันธ์กัน เมื่อนำชุดข้อมูล KDDcup99 จำนวน 34 ลักษณะหาองค์ประกอบหลักแล้วจึงเลือกองค์กระกอบหลักที่ผลรวมค่าไอเกนไม่น้อยกว่า 0.95 ซึ่งเป็นเกณฑ์ที่งานวิจัยส่วนใหญ่นิยมใช้กัน ซึ่งจะได้ 19 องค์ประกอบ และนำไปคูณกับชุดข้อมูลดั้งเดิมทำให้ได้ข้อมูลชุดใหม่ ดังนั้นการหาตัวแทนชุดข้อมูลด้วยวิธีการวิเคราะห์องค์ประกอบหลักสามารถสกัดลักษณะออกมาได้จำนวน 19 ลักษณะ และเมื่อนำข้อมูลชุดใหม่นี้ไปเข้ากระบวนการรู้จำแบบโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน แสดงให้เห็นว่าเมื่อลดลักษณะลงด้วยการสกัดลักษณะด้วยวิธีการวิเคราะห์องค์ประกอบหลัก ประสิทธิภาพของความถูกต้องของข้อมูลดีขึ้นเล็กน้อยเมื่อเปรียบเทียบกับข้อมูลทั้งหมด แต่เวลาที่ใช้ในการตรวจจับผู้รุกจะน้อยกว่าเนื่องจากจำนวนลักษณะข้อมูลมีจำนวนลดลงจากเดิม ซึ่งสามารถสรุปได้ว่าการสกัดคุณลักษณะด้วยวิธีการวิเคราะห์องค์ประกอบหลักยังไม่เหมาะสมสำหรับข้อมูลผู้บุกรุกในระบบเครือข่ายเนื่องจากข้อมูลมีลักษณะที่กระจายตัวมาก ทั้งนี้อาจเนื่องจากการกำหนดค่าไอเกนสำหรับการเลือกลดลักษณะข้อมูล และวิธีที่ใช้ในการรู้จำผู้บุกรุกด้วย

สำหรับการเลือกลักษณะด้วยค่าสถิติไคสแควร์ เป็นการวัดค่าความสัมพันธ์ระหว่างลักษณะกับคลาสคำตอบ หากลักษณะใด ๆ มีนัยสำคัญกับคลาสคำตอบมาก ค่าสถิติไคสแควร์ก็จะมาก ในทางกลับกันหากหากลักษณะใด ๆ มีนัยสำคัญกับคลาสคำตอบน้อย ค่าสถิติไคสแควร์ที่วัดได้ก็จะน้อย และหากค่าสถิติไคสแควร์เป็นศูนย์ แสดงว่าลักษณะนั้นไม่มีนัยสำคัญกับคลาสคำตอบเลย สามารถตัดลักษณะออกไปได้ ในการเลือกลักษณะด้วยค่าสถิติไคสแควร์นี้ จะนำค่าสถิติไคสแควร์มาเรียงลำดับจากมากไปน้อย และตัดลักษณะที่มีค่าสถิติไคสแควร์น้อยกว่า 0.1 ออกไป จากการทดลองเลือกลักษณะกับชุดข้อมูล

KDDCup99 ได้จำนวนลักษณะ 26 ลักษณะ ซึ่งมีประสิทธิภาพดีกว่าการสกัดลักษณะด้วยวิธีวิเคราะห์องค์ประกอบหลักและข้อมูลทั้งหมด แต่น้อยกว่าการเลือกลักษณะด้วยฮิวริสติกกรีดดี

3. การทดสอบกับชุดข้อมูลอื่น ๆ เพิ่มเติม

จากการทดสอบการเลือกลักษณะด้วยวิธีฮวริสติกกรีดดี การสกัดลักษณะด้วยวิธีการวิเคราะห์องค์ประกอบหลัก และการเลือกลักษณะด้วยค่าสถิติไคสแควร์ กับชุดข้อมูล Statlog จำนวน 36 ลักษณะ และ ชุดข้อมูล Faults จำนวน 27 ลักษณะ ผลปรากฏว่ามีลักษณะไปในทิศทางเดียวกัน กล่าวคือการเลือกลักษณะด้วยวิธีฮวริสติกกรีดดีได้ผลดีที่สุด แต่มีประสิทธิภาพดีกว่าการสกัดลักษณะด้วยวิธีการวิเคราะห์องค์ประกอบหลักเพียงเล็กน้อย เนื่องลักษณะการกระจายตัวของชุดข้อมูล 2 ชุดนี้ มีลักษณะการกระจายตัวไม่มาก จึงเหมาะสำหรับการสกัดด้วยวิธีการวิเคราะห์องค์ประกอบหลัก แต่หากเลือกลักษณะด้วยวิธีฮวริสติกกรีดดีก็ให้ผลได้ใกล้เคียงแต่มีผลลัพธ์ที่ดีเช่นกัน

ปัญหาและข้อเสนอแนะ

ในขั้นตอนของการเลือกลักษณะด้วยวิธีฮิวริสติกกรีดดีจะใช้เวลานานมาก เนื่องจากต้องนำไอเท็มเซตหรือชุดลักษณะที่สร้างได้ผ่านกระบวนการรู้จำแบบโครงข่ายประสาทเทียมเพื่อหาจำนวนที่ตอบผิด เพื่อให้ได้ชุดลักษณะที่ดี ควรมีการออกแบบและทดสอบหาฟังก์ชันผิดพลาดที่หลากหลาย และควรมีการทดสอบกับชุดข้อมูลที่มีลักษณะหลายรูปแบบมากกว่านี้

งานที่จะทำต่อไปในอนาคต

1. แบ่งข้อมูลสำหรับการเรียนรู้และทดสอบเพิ่มเติม โดยแบ่งออกเป็นร้อยละ 70:30 และ 80:20

2. ปรับพารามิเตอร์ให้เหมาะสมในการเรียนรู้ด้วยวิธีโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐานสำหรับชุดข้อมูลนั้น ๆ เพื่อให้ผลการทดลองสูงสุด

บรรณานุกรม

Adel Jahanbani, Hossein Karimi. (2012). A New Approach for Detecting Intrusions Based on the PCA Neural Networks. *Journol of Basic and Applied Scientific Research*, 672-769.

Chih-Fong Tsaia, Yu-Feng Hsub, Chia-Ying Linc, Wei-Yang Lin. (2009). Intrusion detection by machine learning: A review", *Expert Systems with Applications: An International Journal, 36(10)*, 11994–12000.

Fangjun Kuang, Weihong Xu, Siyang Zhang, Yanhua Wang , and Ke Liu. (2012). A Novel Approach of KPCA and SVM for Intrusion Detection. *Journal of Computotional Informotion Systems*.

Guo-Liang Li, Tze-Yun Leong. (2005). Feature Selection for the Prediction of Translation Initiation Sites. *Genomics Proteomics & Bioinformotics, Volume 3, No.2*.

Hari Om , Aritra Kundu. (2012). *A Hybrid System for Reducing the Folse Alarm Rote of Anomaly Intrusion Detection System*. International Conference on Recent Advances in Information Technology.

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

Iftikhar Ahmad, Azween Abdullah, Abdullah Alghamdi and Muhammad Hussain. (2012). *Optimized Intrusion Detection Mechanism using Soft Computing Techniques.* Telecommunication Systems (pp. 1-9).

J. Ross Quinlan. (1993). *C4.5: Progroms for Mochine Learning*. San Francisco, CA, USA, Morgan Kaufmann Publishers Inc.

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

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

Jiawei Han, Micheline Kamber, and Jian Pei. (2011). *Dota Mining; Concepts and Techniques*. Morgan Kaufmann.

K. P. Soman, S. Diwakar, and V. Ajay. (2006). *Insight into Data Mining Theory and Proctice*. Prentice-Hall of India.

KDD Cup datasets. (1999). The UCI KDD Archive. From http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, Irvine, CA, USA.

Khazaee Saeed, Abadeh Mohammad Saniee. (2011). A Hybrid Model Based on Feature Extraction for Network Intrusion Detection. *Journal of Computing*, 3(9), 65.

Lei Xie, Jin Li. (2009). A Novel Feature Extraction Method Assembled with PCA and ICA for Network. *International Forum on Computer Science-Technology and Applications Intrusion Detection*.

M. Anbarasi, E. Anupriya , N.CH.S.N.Iyengar. (2010). Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm. *International Journal of Engineering Science and Technology*, 2(10).

Marti Hearst. (1998). Support Vector Machines. *IEEE Intelligent Systems Magazine*, Trends and Controversies, 13(4).

M. Revathi, T.Ramesh. (2011). Network Intrusion Detection System Using Reduced Dimensionality. *Indian Journal of Computer Science and Engineering (IJCSE)*. 2(1).

Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani. (2009). A Detailed Analysis of the KDD CUP 99 Data Set, *Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA)*.

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

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*, 36, pp. 12500–12505.

Noreen Kausar, Brahim Belhaouari Samir, Suziah Sulaiman, Iftikhar Ahmad, Muhammad Hussain. (2012). An Approach towards Intrusion Detection using PCA Feature Subsets and SVM. *International Conference on Computer & Information Science (ICCIS)*, pp. 569 – 574.

Onur Inan, Mustafa Serter Uzer, and Nihat Y.Imaz. (2013). 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*, 9(2), 727-739.

P. N. Tan, M. Steinbach, and V. Kumar. (2006). *Introduction to Data Mining*. Pearson International Edition.

Rakesh Agrawal, Tomasz Imielinski, Arun Swami. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Conference on Monagement of Data*, Washington.

Ranjit Abraham, Jay B.Simha and S.Sitharama Iyengar. (2009). Effective Discretization and Hybrid feature selection using Naïve Bayesian classifier for Medical Datamining. *International Journal of Computation Intelligence Research*, 5(2), 116-129.

Riti Lath, Manish Shrivastava. Analytical Study of Different Classification Technique for KDD Cup Data'99. *International Journal of Applied Information Systems, Foundation of Computer Science FCS*, 3(6).

Robert Hecht Nielsen. (1989). Theory of the back propagation neural network. Proceedings 1989 IEEE IJCNN, pp. 1593–1605.

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*, 2(2).

Scarfone Karen, Mell Peter. (2007). Guide to Intrusion Detection and Prevention Systems (IDPS). *Computer Security Resource Center (Notional Institute of Standards and Technology).*

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

Shilpa Lakhina, Sini Joseph, Bhupendra Verma. (2010). Feature Reduction using Principal Component Analysis for Effective Anomaly–Based Intrusion Detection on NSL-KDD. *International Journal of Engineering Science and Technology,* 2(6).

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

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

ภาคผนวก

ต้นฉบับไม่ปรากฏหน้า 95

ภาคผนวก ก

ลักษณะของชุดข้อมูลเพิ่มเติม

ลักษณะของชุดข้อมูล Statlog

ชุดข้อมูล Statlog เป็นข้อมูลภาพถ่ายดาวเทียมซึ่งเป็นข้อมูลที่ได้จากเหตุการณ์จริง จัดทำขึ้นโดย AshwinSrinivasan ภาควิชาแบบจำลองข้อมูลและสถิติ (Statistics and Data Modeling) มหาวิทยาลัย Strathclyde ประเทศอังกฤษโดยภาพถ่ายได้มาจากการสำรวจระยะไกลโดยองค์กรนาซ่าศูนย์ออสเตรเลียจากชุดข้อมูล Statlog ประกอบด้วย 6,435 จุดข้อมูล 36 แอทริบิวต์และ 6 คลาสคำตอบ ซึ่งแต่ละคลาสคำตอบมีรายละเอียดดังตารางที่โดยข้อมูลเป็นตัวเลขตั้งแต่ 0 ถึง 255

ตารางที่ ก-1 จำนวนข้อมูล Statlog ในแต่ละคลาส

Class	Class Name	Amount
1	red soil	1533
2	cotton crop	703
3	grey soil	1358
4	damp grey soil	626
5	soil with vegetation stubble	707
6	very damp grey soil	1508
รวม		6435

ตารางที่ ก-2 ค่าทางสถิติของข้อมูล Statlog

Features	Maximum	Minimum	Mean	Standard Deviation
1	104	39	69.4	13.60587
2	137	27	83.59487	22.88223
3	140	53	99.2906	16.64594
4	154	33	82.5927	18.89767
5	104	39	69.15027	13.5612
6	137	27	83.24351	22.88649
7	145	50	99.11064	16.66409
8	157	29	82.49713	18.94092
9	104	40	68.91235	13.4706
10	130	27	82.89308	22.86225
11	145	50	98.8533	16.63661

ตารางที่ ก-2 (ต่อ)

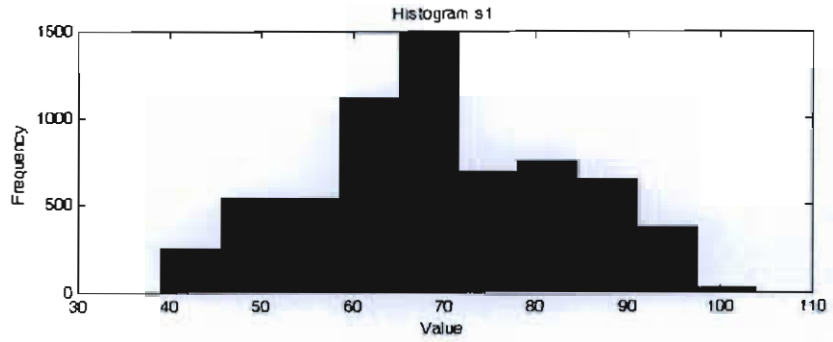
Features	Maximum	Minimum	Mean	Standard Deviation
12	157	29	82.38819	18.98111
13	104	39	69.28967	13.60269
14	137	27	83.47677	22.84989
15	145	50	99.31127	16.66787
16	154	29	82.64491	18.93199
17	104	40	69.04569	13.53762
18	130	27	83.1711	22.90506
19	145	50	99.14981	16.71767
20	157	29	82.60326	19.03554
21	104	39	68.83932	13.45923
22	130	27	82.86092	22.88438
23	145	50	98.94965	16.72962
24	157	29	82.46853	19.07075
25	104	39	69.16239	13.58052
26	131	27	83.37343	22.80274
27	140	50	99.21476	16.61251
28	154	29	82.66061	18.99128
29	104	39	68.94406	13.49268
30	130	27	83.14561	22.8472
31	145	50	99.11189	16.7043
32	157	29	82.61803	19.04366
33	104	39	68.72758	13.4016
34	130	27	82.8589	22.81696
35	145	50	98.92603	16.69549
36	157	29	82.50536	19.05427

ตารางที่ ก-3 ค่าสหสัมพันธ์ระหว่างแต่ละลักษณะของข้อมูล Statlog

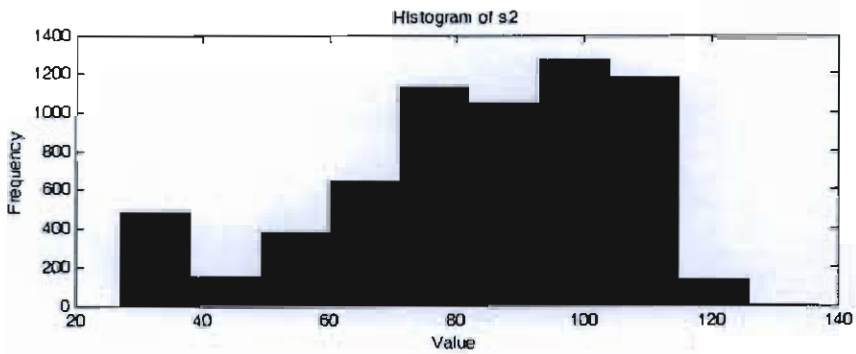
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1																	
2	0.810493	1																
3	0.212968	0.360041	1															
4	-0 16806	-0 10377	0.864817	1														
5	0.952848	0.788064	0.206825	-0.16654	1													
6	0.773463	0 959137	0.345801	-0.1029	0 810643	1												
7	0.188514	0.337845	0.934762	0.832632	0.2119	0.359	1											
8	-0.17554	-0.10866	0.818505	0.945849	-0.17053	-0.10545	0.864401	1										
9	0.882883	0.726034	0.177394	-0.16905	0.952409	0.787266	0.20242	-0.17167	1									
10	0.709774	0.887226	0.311209	-0.10518	0.7748	0.958873	0.342594	-0.10689	0.811493	1								
11	0.156923	0.299433	0.847483	0.761048	0.186794	0.332664	0.934742	0.835447	0.205905	0.350711	1							
12	-0.17608	-0.1112	0.739535	0.858751	-0.17687	-0.11258	0.817052	0.947009	-0.17568	-0.11223	0.86537	1						
13	0.935998	0.759909	0.191993	-0.1624	0.909904	0.728793	0.175463	-0.16543	0.860631	0.680754	0.150411	-0.1675	1					
14	0.774462	0.94226	0.337753	-0.09797	0.751747	0.910761	0.321804	-0.09902	0.705133	0.85704	0.289678	-0.10372	0.809606	1				
15	0 195549	0.338212	0.910392	0.804833	0.191505	0.329224	0.882411	0.775548	0.169936	0.304081	0.830216	0.726963	0.212607	0.357399	1			
16	-0.17039	-0.10417	0.807549	0.9214	-0.16468	-0.09851	0.786873	0.889546	-0.16582	-0.09921	0.744998	0.838605	-0 16784	-0.10469	0.865751	1		
17	0.917335	0.752691	0 186535	-0.16826	0.934789	0.759585	0.190401	-0.16549	0.90861	0.729573	0.17325	-0.16641	0.952629	0.787524	0.204809	-0 1681	1	
18	0.751151	0.927937	0.329067	-0.10267	0.771428	0.941398	0.337618	-0 09844	0.748022	0.909168	0 317652	-0.1005	0 771098	0.959591	0.343668	-0.10391	0.808398	1

ตารางที่ ก-3 (ต่อ)

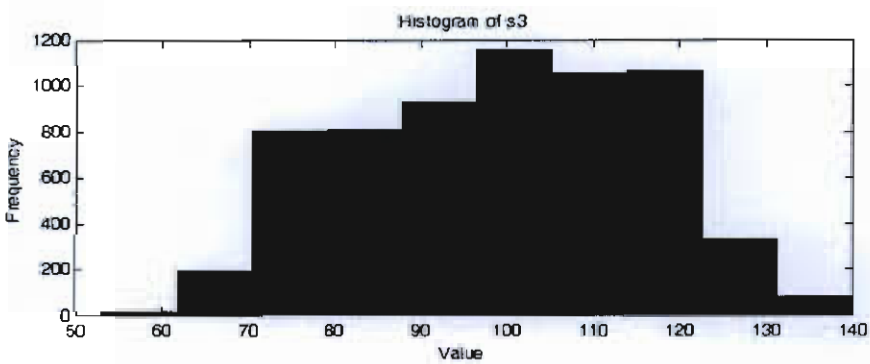
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
19	0.172774	0.320484	0.884432	0.789326	0.189364	0.335045	0.910283	0.805098	0.182897	0.323965	0.88173	0.77458	0.183298	0.332622	0.935746	0.835716	0.205374	0.354199
20	-0.1818	-0.11408	0.779649	0.900742	-0.17485	-0.10617	0.806942	0.921133	-0.17079	-0.10217	0.788089	0.888741	-0.17814	-0.11125	0.819968	0.947194	-0.1741	-0.10746
21	0.862778	0.708389	0.163854	-0.17323	0.915936	0.751239	0.183323	-0.17189	0.934208	0.760194	0.186518	-0.16802	0.883585	0.726918	0.175177	-0.17186	0.952953	0.785236
22	0.697657	0.874053	0.302241	-0.10664	0.750995	0.927836	0.327566	-0.1052	0.771566	0.941798	0.331292	-0.10314	0.708171	0.889726	0.309864	-0.10615	0.772316	0.959477
23	0.144779	0.28942	0.816299	0.730755	0.170549	0.316852	0.885764	0.791723	0.183914	0.329539	0.911801	0.806278	0.151114	0.294958	0.850164	0.764807	0.180032	0.328861
24	-0.18078	-0.11406	0.714556	0.829609	-0.18131	-0.11525	0.780209	0.901752	-0.17727	-0.10959	0.809997	0.921876	-0.17797	-0.11336	0.744295	0.862551	-0.17895	-0.11253
25	0.855716	0.677694	0.150966	-0.16223	0.834732	0.652065	0.142809	-0.16053	0.800096	0.618609	0.126897	-0.15908	0.93591	0.7583	0.187193	-0.16648	0.909525	0.725414
26	0.701209	0.853511	0.297027	-0.09306	0.680095	0.827725	0.289226	-0.08973	0.647849	0.789869	0.267187	-0.09106	0.774584	0.942776	0.333297	-0.10063	0.752218	0.911043
27	0.161407	0.3	0.798175	0.7038	0.159142	0.295427	0.786598	0.691534	0.148501	0.280445	0.759233	0.666941	0.190033	0.32994	0.910485	0.808887	0.185813	0.321851
28	-0.16779	-0.09816	0.705769	0.803828	-0.16045	-0.09067	0.697539	0.789089	-0.15831	-0.08856	0.680299	0.763085	-0.17341	-0.111	0.806736	0.923274	-0.16787	-0.10438
29	0.852749	0.68464	0.148515	-0.16989	0.854256	0.677944	0.1521	-0.16289	0.833233	0.652312	0.142837	-0.15852	0.918939	0.752329	0.183221	-0.17102	0.936047	0.757411
30	0.697374	0.859497	0.294101	-0.10169	0.69877	0.855075	0.298105	-0.09344	0.677514	0.827796	0.285631	-0.09118	0.752853	0.930531	0.32524	-0.10584	0.773728	0.943952
31	0.14749	0.293117	0.792605	0.702577	0.157147	0.300059	0.80257	0.706934	0.153777	0.293268	0.788504	0.692136	0.17059	0.316259	0.88702	0.794217	0.186411	0.331122
32	-0.17933	-0.10839	0.701348	0.806162	-0.17025	-0.09794	0.710546	0.808033	-0.16392	-0.092	0.702203	0.790815	-0.18175	-0.11752	0.783725	0.906191	-0.17556	-0.10951
33	0.822496	0.66599	0.134293	-0.1792	0.850864	0.683685	0.145695	-0.17369	0.852608	0.677721	0.148998	-0.16425	0.865008	0.709504	0.160034	-0.17763	0.91893	0.750552
34	0.666953	0.835833	0.279157	-0.1089	0.695877	0.86091	0.295016	-0.10197	0.696742	0.854961	0.29524	-0.09468	0.699928	0.879463	0.299766	-0.1098	0.752924	0.931597
35	0.126571	0.275642	0.759368	0.675805	0.14369	0.29225	0.79808	0.707288	0.151722	0.297166	0.806428	0.709645	0.141526	0.287577	0.823004	0.73886	0.166436	0.315065
36	-0.18389	-0.11264	0.668132	0.773961	-0.18126	-0.10867	0.705879	0.810112	-0.17294	-0.09942	0.71549	0.809841	-0.18399	-0.11761	0.722813	0.839058	-0.18455	-0.11783



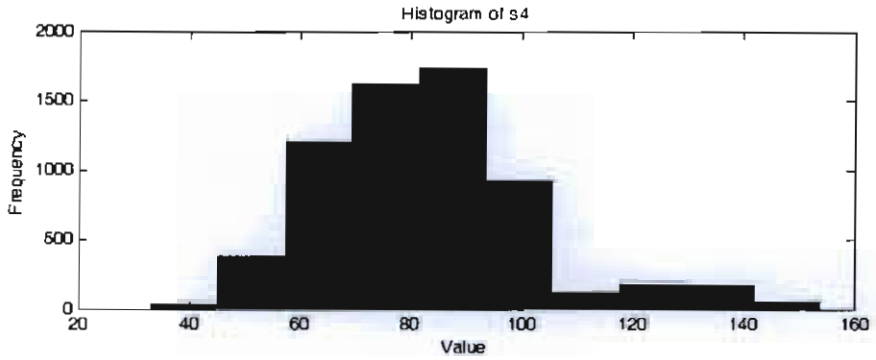
ภาพที่ ก-1 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 1 (s1)



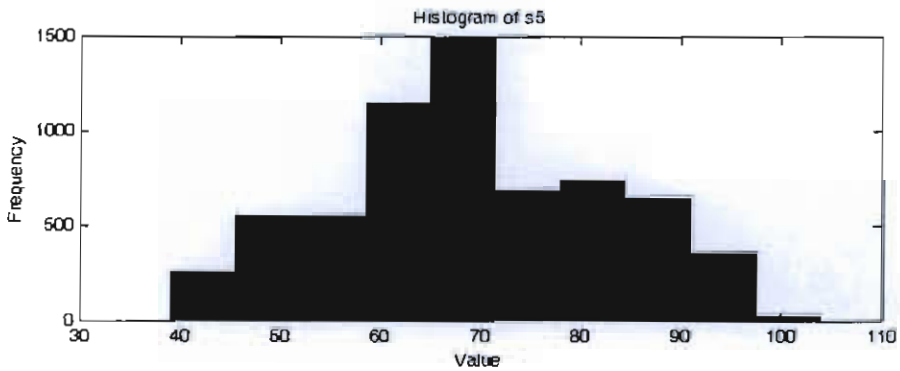
ภาพที่ ก-2 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 2 (s2)



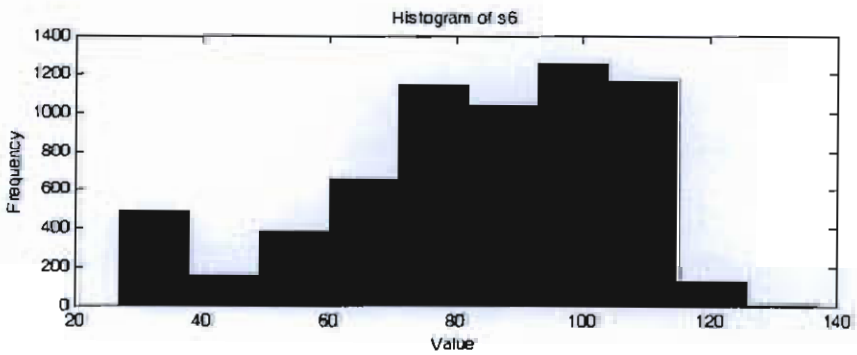
ภาพที่ ก-3 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 3 (s3)



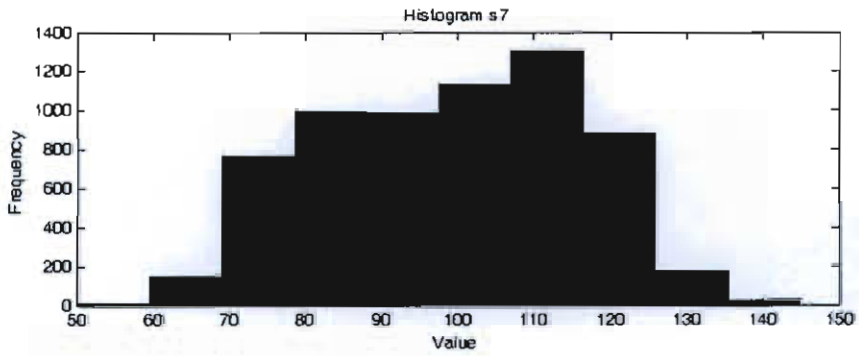
ภาพที่ ก-4 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 4 (s4)



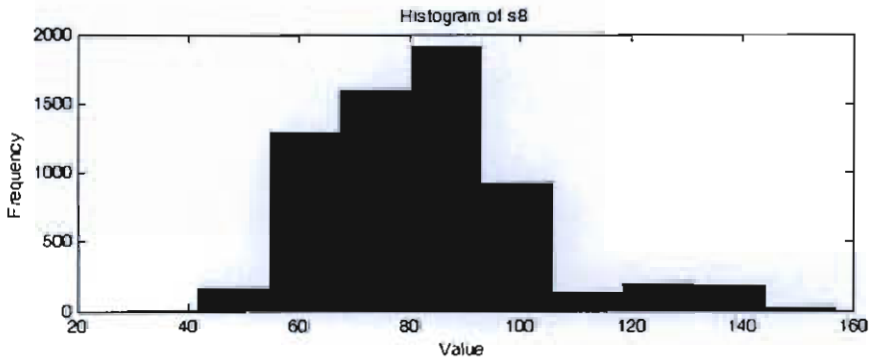
ภาพที่ ก-5 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 5 (s5)



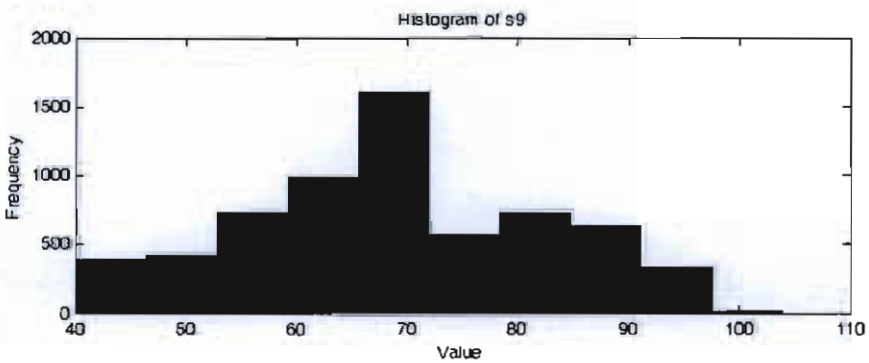
ภาพที่ ก-6 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 6 (s6)



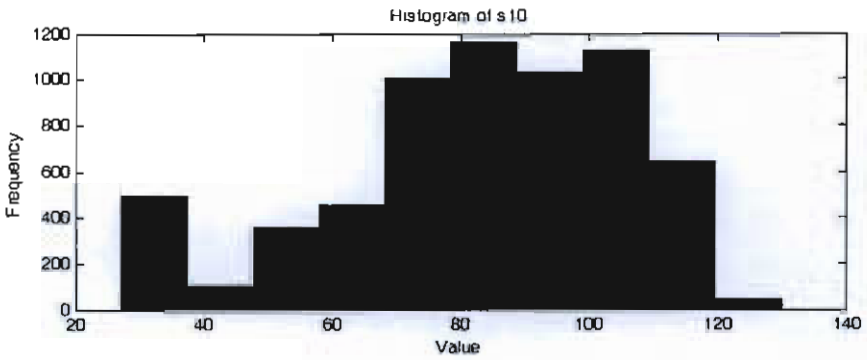
ภาพที่ ก-7 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 7 (s7)



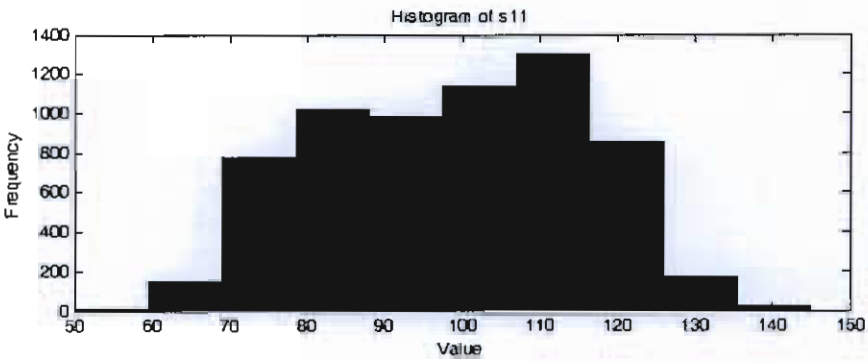
ภาพที่ ก-8 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 8 (s8)



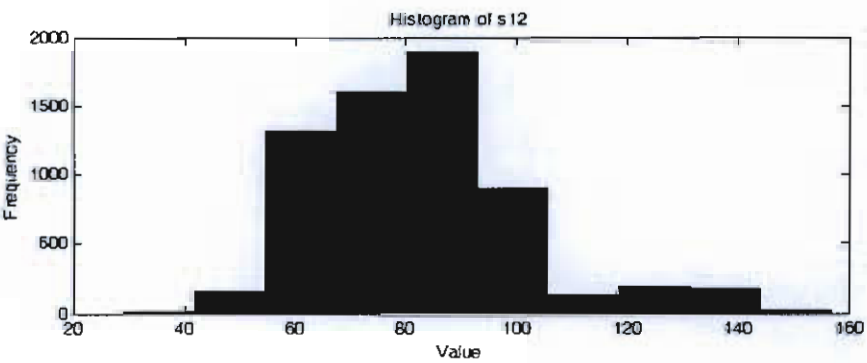
ภาพที่ ก-9 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 9 (s9)



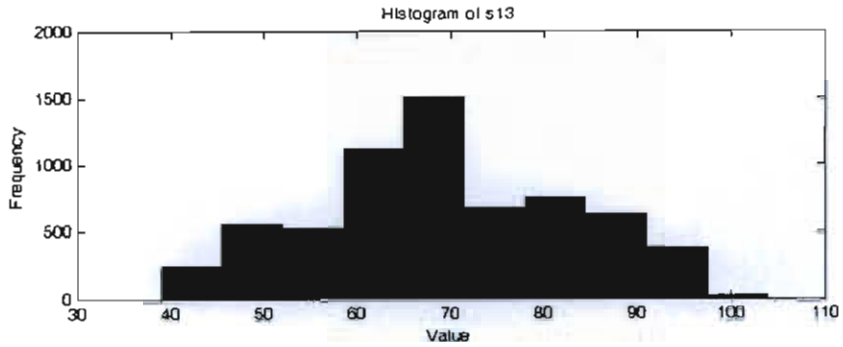
ภาพที่ ก-10 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 10 (s10)



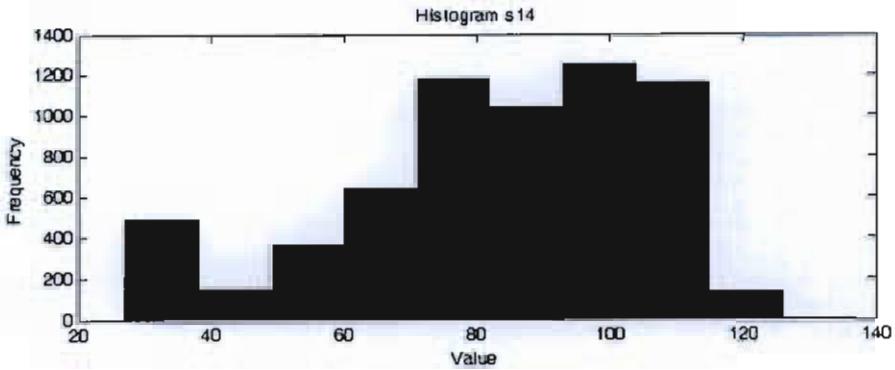
ภาพที่ ก-11 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 11 (s11)



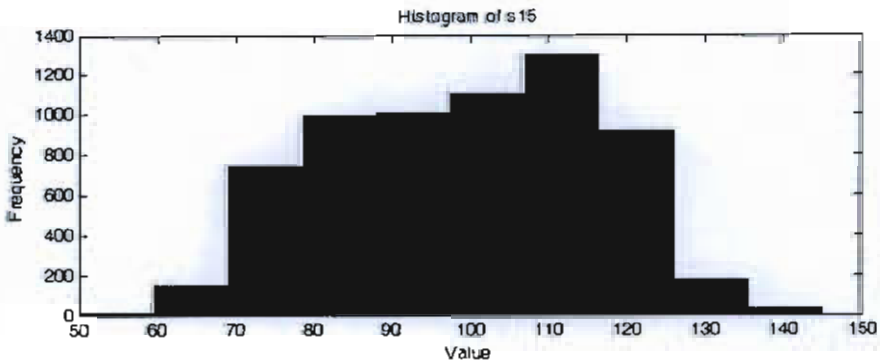
ภาพที่ ก-12 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 12 (s12)



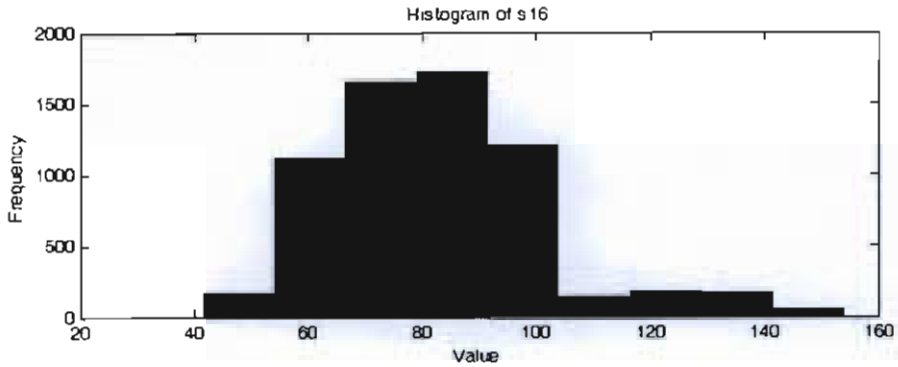
ภาพที่ ก-13 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 13 (s13)



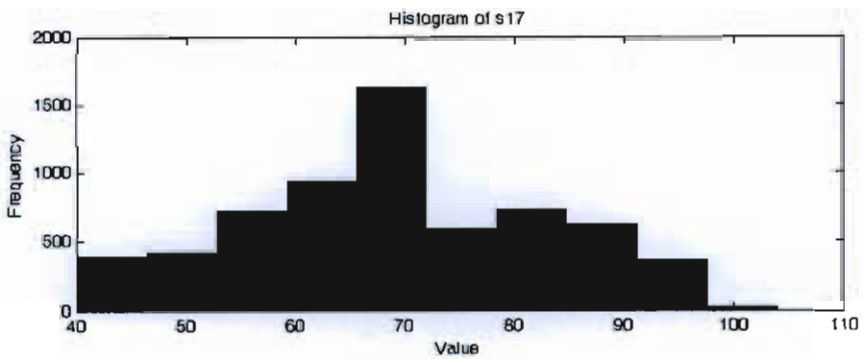
ภาพที่ ก-14 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 14 (s14)



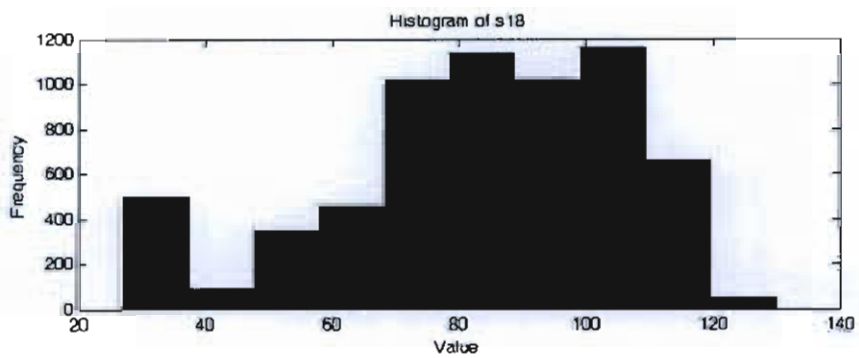
ภาพที่ ก-15 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 15 (s15)



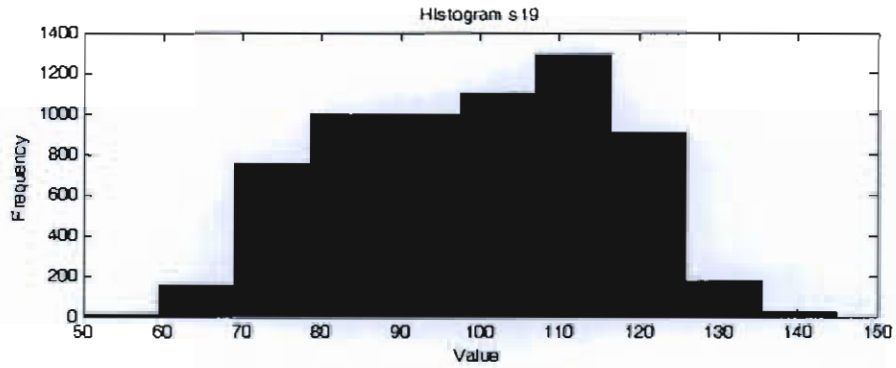
ภาพที่ ก-16 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 16 (s16)



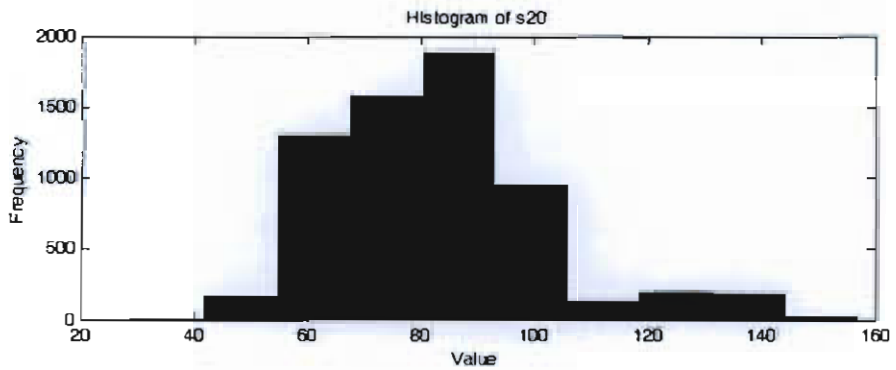
ภาพที่ ก-17 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 17 (s17)



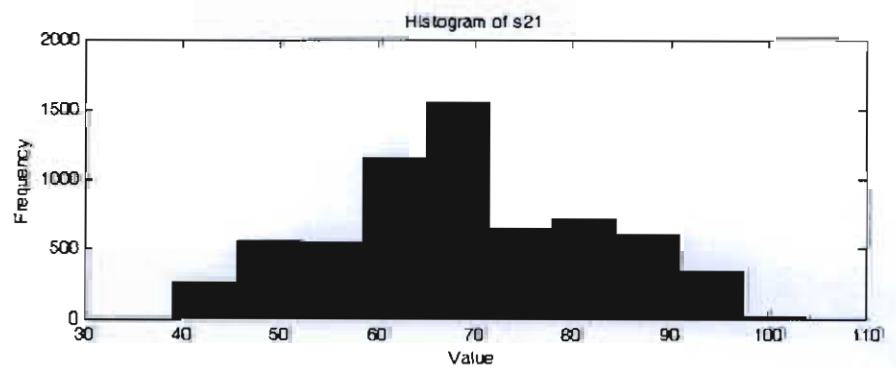
ภาพที่ ก-18 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 18 (s18)



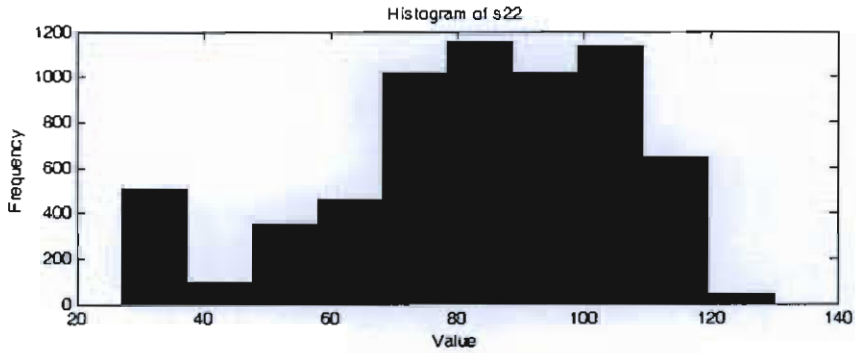
ภาพที่ ก-19 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 19 (s19)



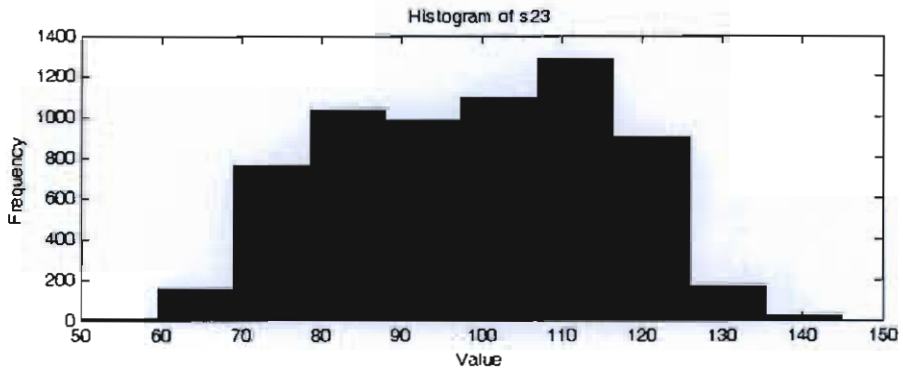
ภาพที่ ก-20 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 20 (s20)



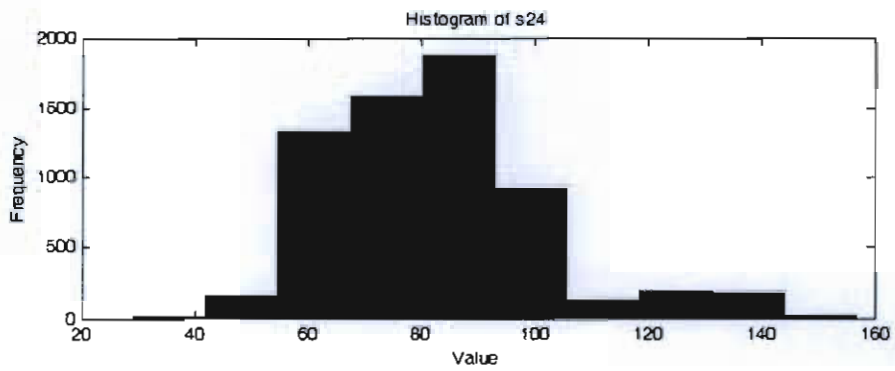
ภาพที่ ก-21 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 1 (s21)



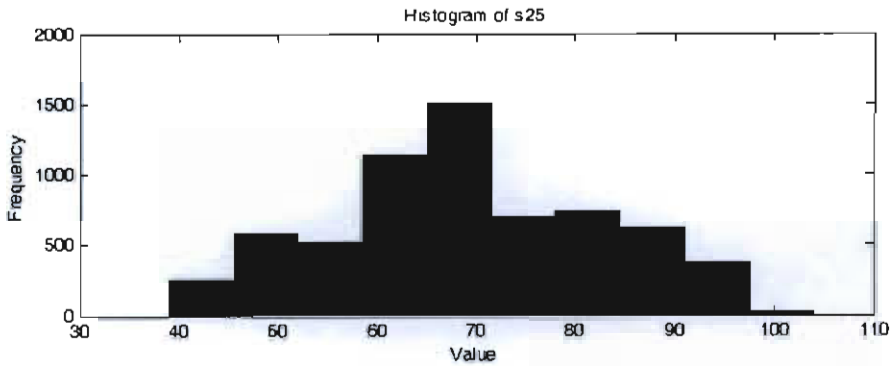
ภาพที่ ก-22 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 22 (s22)



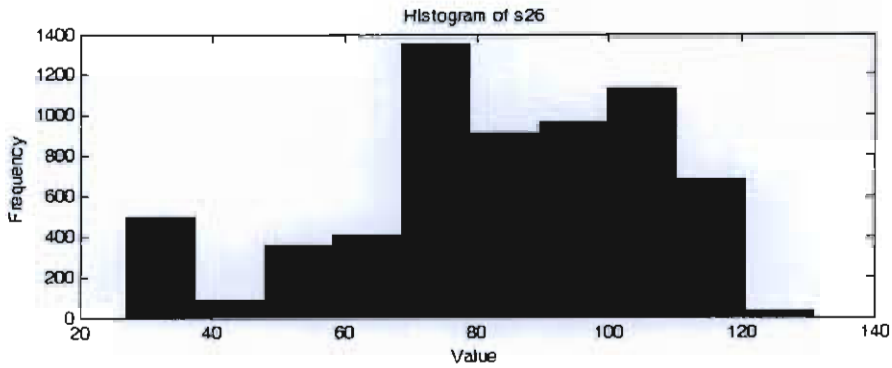
ภาพที่ ก-23 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 23 (s23)



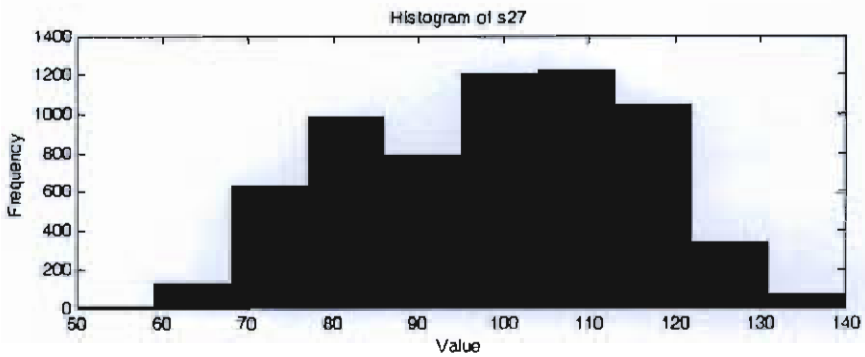
ภาพที่ ก-24 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 24 (s24)



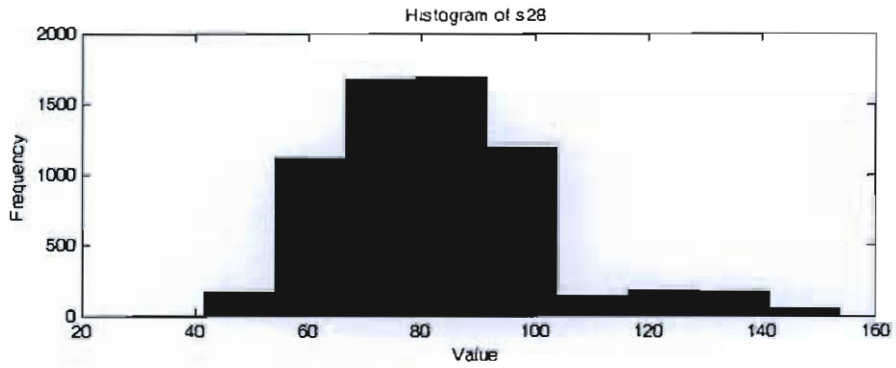
ภาพที่ ก-25 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 25 (s25)



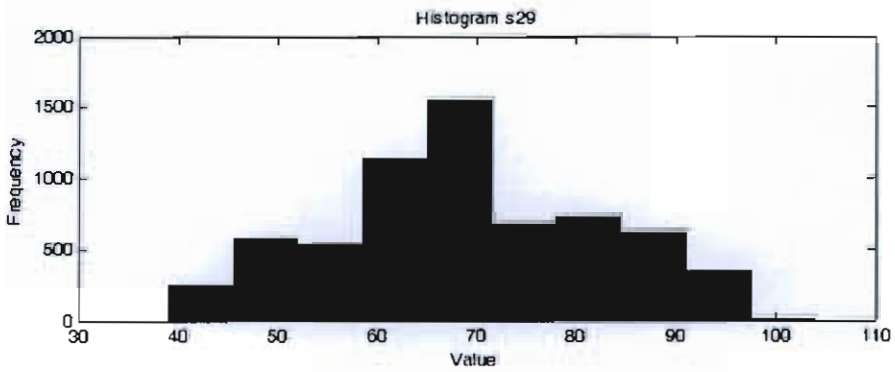
ภาพที่ ก-26 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 26 (s26)



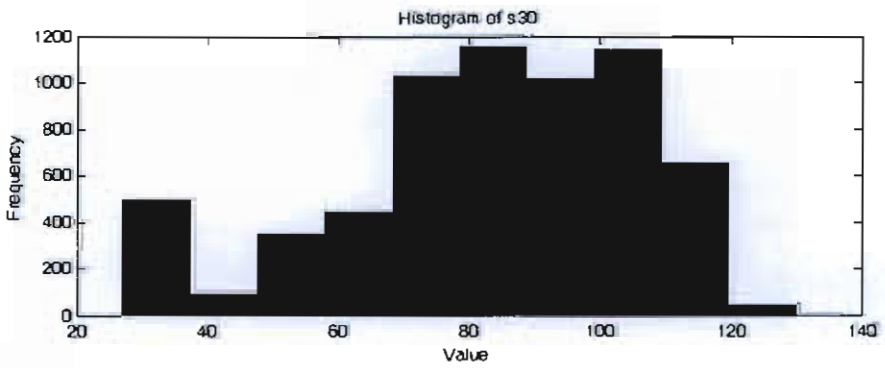
ภาพที่ ก-27 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 27 (s27)



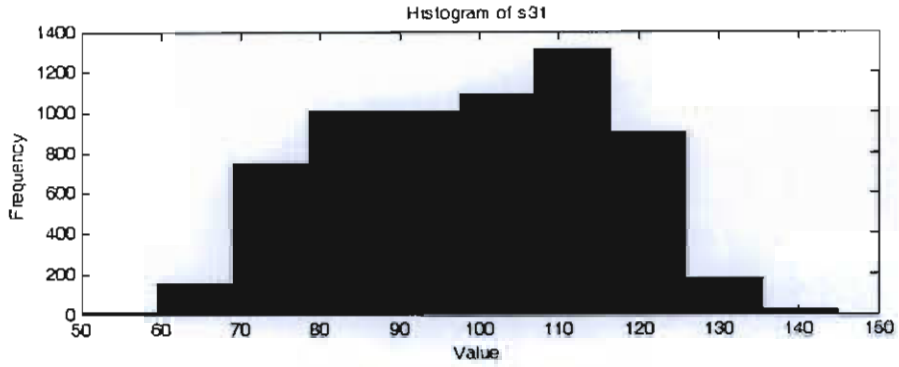
ภาพที่ ก-28 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 28 (s28)



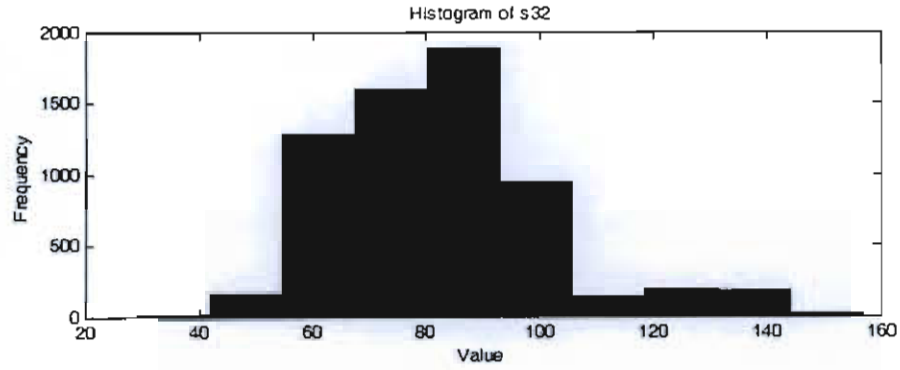
ภาพที่ ก-29 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 29 (s29)



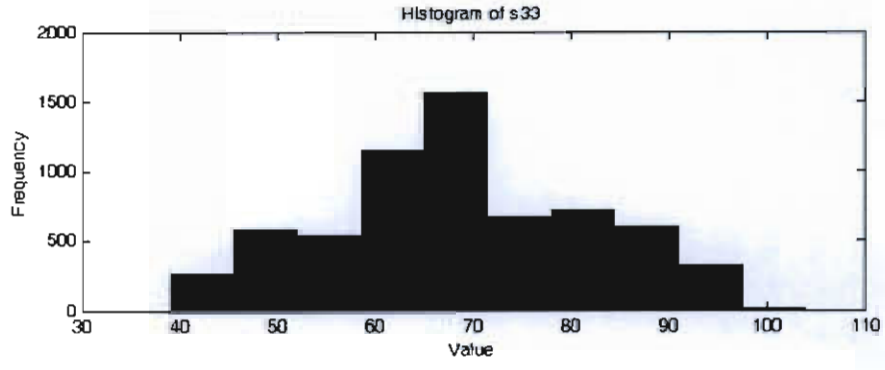
ภาพที่ ก-30 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 30 (s30)



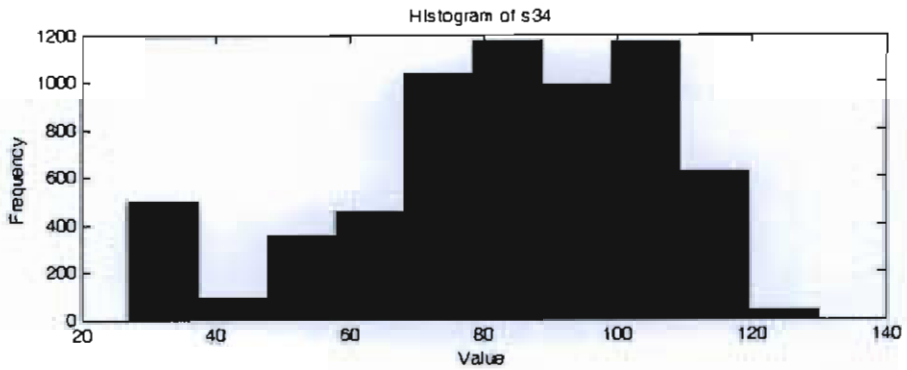
ภาพที่ ก-31 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 31 (s31)



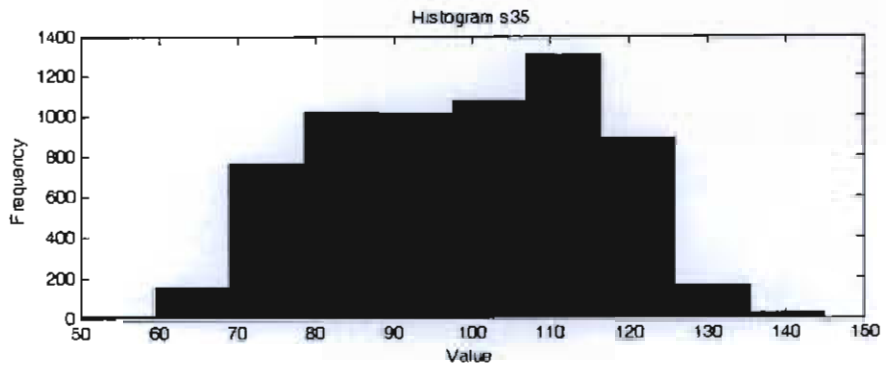
ภาพที่ ก-32 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 32 (s32)



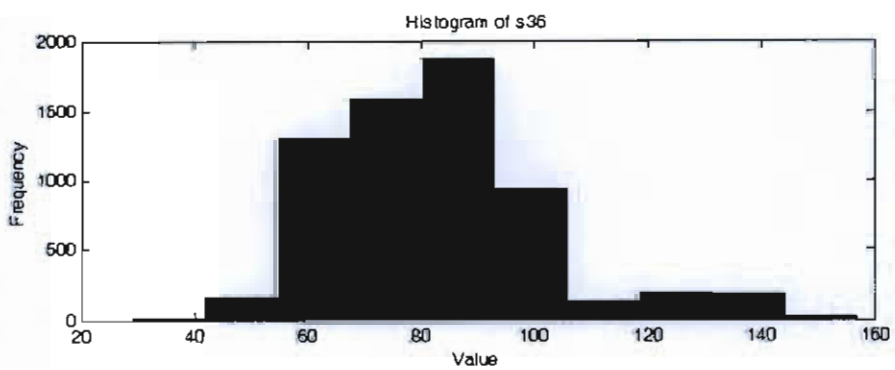
ภาพที่ ก-33 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 33 (s33)



ภาพที่ ก-34 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 34 (s34)



ภาพที่ ก-35 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 35 (s35)



ภาพที่ ก-36 Histogram ของลักษณะข้อมูล Statlog ลักษณะที่ 36 (s36)

ลักษณะของชุดข้อมูล Faults

ชุดข้อมูล Faults ประกอบด้วย 1,941 จุดข้อมูล 27 แอทริบิวต์และ 7 คลาสคำตอบ ซึ่งแต่ละคลาสคำตอบมีรายละเอียดดังตารางที่โดยข้อมูลเป็นเลขจำนวนเต็มและเลขจำนวนจริง นำข้อมูลนี้มาจาก

ตารางที่ ก-4 จำนวนข้อมูล Faults ในแต่ละคลาส

Class	Class Name	Amounts
1	Pastry	158
2	Z_Scratch	190
3	K_Scatch	391
4	Stains	72
5	Dirtiness	55
6	Bumps	402
7	Other_Faults	673
รวม		1,941

ตารางที่ ก-5 ค่าทางสถิติของข้อมูล Faults

Features	Maximum	Minimum	Mean	Standard Deviation
1	1705	0	571.136	520.6907
2	1713	4	617.9645	497.6274
3	12987661	6712	1650685	1774578
4	12987692	6724	1650739	1774590
5	152655	2	1893.878	5168.46
6	10449	2	111.8552	301.2092
7	18152	1	82.966	426.4829
8	11591414	250	206312.1	512293.6
9	203	0	84.54869	32.13428
10	253	37	130.1937	18.69099
11	1794	1227	1459.16	144.5778
12	1	0	0.400309	0.490087
13	1	0	0.599691	0.490087
14	300	40	78.73776	55.08603

ตารางที่ ก-5 (ต่อ)

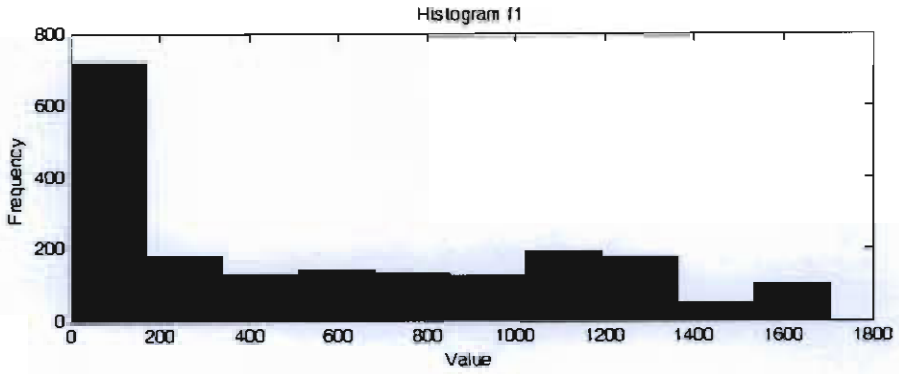
Features	Maximum	Minimum	Mean	Standard Deviation
15	0.9952	0	0.331715	0.299712
16	0.9439	0	0.414203	0.137261
17	1	0.0083	0.570767	0.271058
18	0.8759	0.0015	0.033361	0.058961
19	1	0.0144	0.610529	0.243277
20	1	0.0484	0.813472	0.234274
21	1	0	0.575734	0.482352
22	5.1837	0.301	2.492388	0.78893
23	3.0741	0.301	1.335686	0.481612
24	4.2587	0	1.403271	0.454345
25	0.9917	-0.991	0.083288	0.500868
26	0.6421	-0.9989	-0.13131	0.148767
27	1	0.119	0.58542	0.339452

ตารางที่ ก-6 ค่าสหสัมพันธ์ระหว่างแต่ละลักษณะของข้อมูล Faults

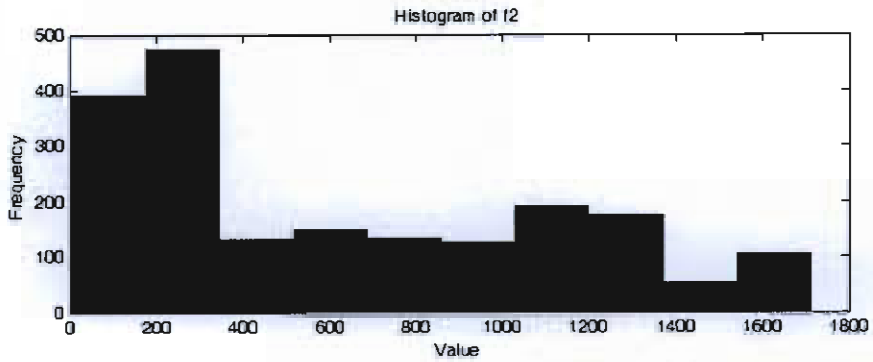
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1												
2	0.988314	1											
3	0.041821	0.052147	1										
4	0.041807	0.052135	1	1									
5	-0.30732	-0.2254	0.01767	0.01784	1								
6	-0.25894	-0.18633	0.023843	0.024038	0.966644	1							
7	-0.11876	-0.09014	0.02415	0.02438	0.827199	0.912436	1						
8	-0.33905	-0.24705	0.007362	0.007499	0.978952	0.912956	0.704876	1					
9	0.237637	0.168649	-0.0657	-0.06573	-0.4972	-0.40043	-0.21376	-0.54057	1				
10	-0.07555	-0.06239	-0.06779	-0.06778	0.110063	0.111363	0.061809	0.136515	0.429605	1			
11	0.316662	0.29939	-0.04921	-0.04922	-0.15585	-0.13424	-0.06382	-0.16933	-0.02358	-0.09801	1		
12	0.144319	0.112009	0.075164	0.075151	-0.23559	-0.18925	-0.09515	-0.26363	0.042048	-0.21634	0.378542	1	
13	-0.14432	-0.11201	-0.07516	-0.07515	0.235591	0.18925	0.095154	0.263632	-0.04205	0.216339	-0.37854	-1	1
14	0.136625	0.106119	-0.20764	-0.20764	-0.18374	-0.14771	-0.05889	-0.20481	0.103393	-0.1284	0.214769	0.125649	-0.12565
15	0.278075	0.242846	0.021314	0.0213	-0.27529	-0.22759	-0.11124	-0.30145	0.358915	0.149675	0.135152	0.11214	-0.11214
16	-0.19846	-0.15268	-0.04312	-0.04309	0.272808	0.306348	0.188825	0.293691	-0.04411	0.031425	-0.2306	-0.09195	0.091954
17	0.063658	0.048575	-0.00613	-0.00615	0.017865	0.004507	-0.04751	0.049607	0.066748	0.065517	0.073694	0.164156	-0.16416
18	-0.36116	-0.21493	0.054165	0.054185	0.588606	0.517098	0.20916	0.658339	-0.48757	0.0993	-0.21742	-0.24477	0.244765
19	0.154778	0.149259	0.066085	0.066051	-0.29467	-0.29304	-0.19516	-0.32773	0.252256	0.093522	0.123585	0.173836	-0.17384
20	0.367907	0.271915	-0.03654	-0.03655	-0.46357	-0.4121	-0.13672	-0.52975	0.31661	-0.16744	0.235732	0.240634	-0.24063
21	0.147282	0.099253	-0.06291	-0.0629	-0.10965	-0.07911	0.013438	-0.12109	0.035462	-0.12404	0.128663	0.022142	-0.02214
22	-0.42855	-0.33217	0.044952	0.044994	0.650234	0.563036	0.29404	0.712128	-0.67876	0.007672	-0.19325	-0.32961	0.329614
23	-0.43794	-0.32401	0.070406	0.070432	0.603072	0.524716	0.228485	0.667736	-0.56765	0.092823	-0.21997	-0.26695	0.266955
24	-0.32685	-0.26599	-0.00844	-0.00838	0.578342	0.523472	0.344378	0.618795	-0.58821	-0.06952	-0.15706	-0.3118	0.311796
25	0.178585	0.115019	-0.0865	-0.08648	-0.1376	-0.10173	0.031381	-0.15848	0.057123	-0.16975	0.120715	0.01063	-0.01063
26	-0.03158	-0.039	-0.09065	-0.09067	-0.04345	-0.03262	-0.04778	-0.01407	0.669534	0.87016	-0.14977	-0.25282	0.252818
27	-0.35525	-0.28674	0.025257	0.025284	0.422947	0.380605	0.191772	0.464248	-0.5148	-0.03965	-0.19754	-0.30891	0.30891

ตารางที่ ก-6 (ต่อ)

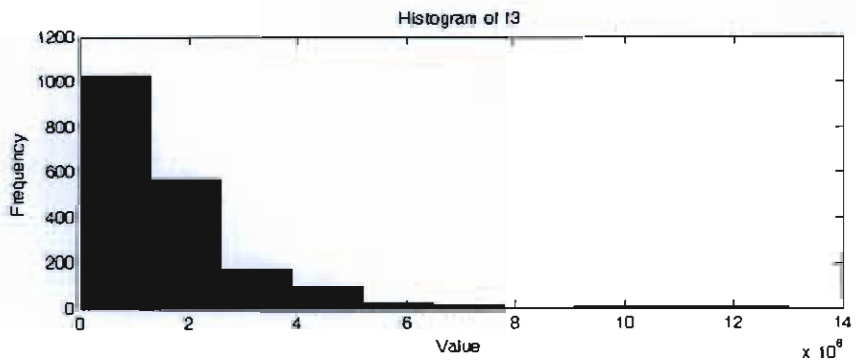
	14	15	16	17	18	19	20	21	22	23	24	25	26	27
14	1													
15	0.063449	1												
16	0.012526	-0.18074	1											
17	-0.12438	0.149498	-0.07644	1										
18	-0.22835	-0.29651	0.334996	-0.11363	1									
19	-0.07741	0.250178	-0.38934	0.242779	-0.07666	1								
20	0.251985	0.285302	-0.4598	0.081488	-0.68987	0.108144	1							
21	0.221244	0.008282	-0.16529	-0.06991	-0.33717	-0.41938	0.537565	1						
22	-0.17664	-0.40862	0.356685	-0.18934	0.710837	-0.49621	-0.64299	-0.09776	1					
23	-0.25282	-0.35585	0.448864	-0.08285	0.820223	-0.18926	-0.85541	-0.42806	0.888919	1				
24	-0.03729	-0.37199	0.397289	-0.25766	0.46486	-0.74889	-0.32189	0.241898	0.882974	0.598652	1			
25	0.274097	0.020548	-0.13942	-0.16203	-0.44036	-0.5503	0.658049	0.86267	-0.1239	-0.53663	0.316792	1		
26	-0.1165	0.207516	0.061608	0.111977	-0.03572	0.12646	-0.09437	-0.12232	-0.17588	-0.06492	-0.21911	-0.15346	1	
27	-0.08516	-0.33001	0.481738	-0.29225	0.51891	-0.55843	-0.54539	-0.05377	0.877768	0.757343	0.838188	-0.02398	-0.18484	1



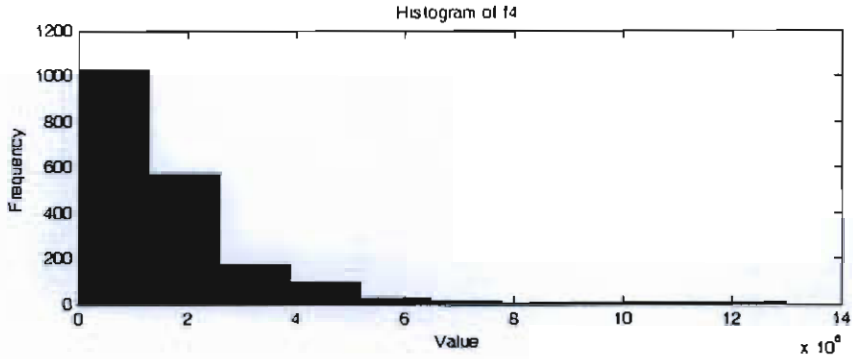
ภาพที่ ก-37 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 1 (f1)



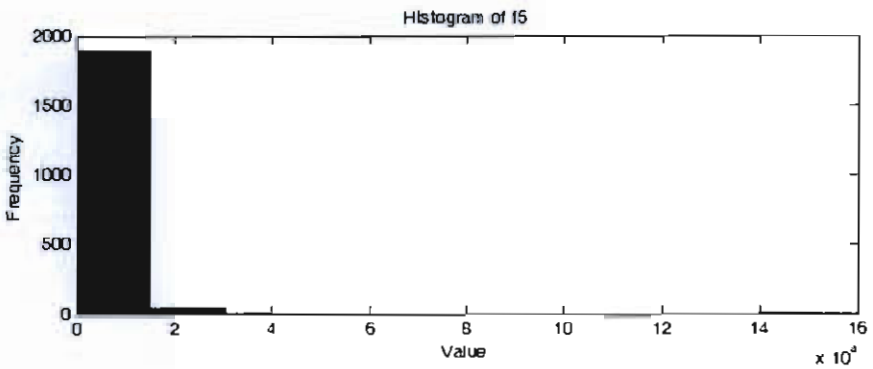
ภาพที่ ก-38 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 2 (f2)



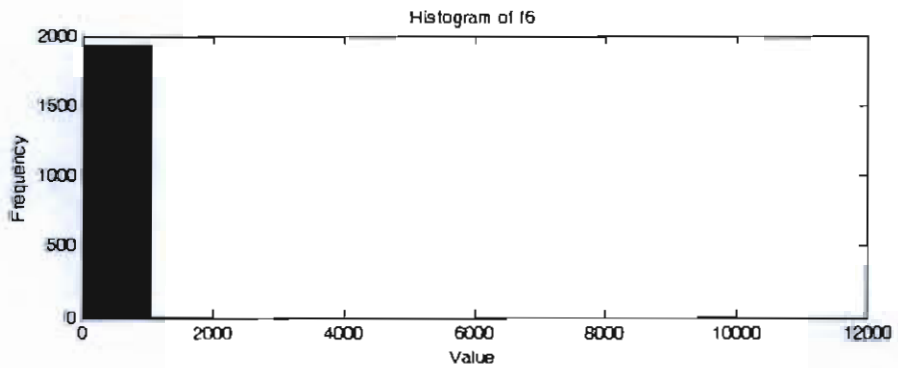
ภาพที่ ก-39 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 3 (f3)



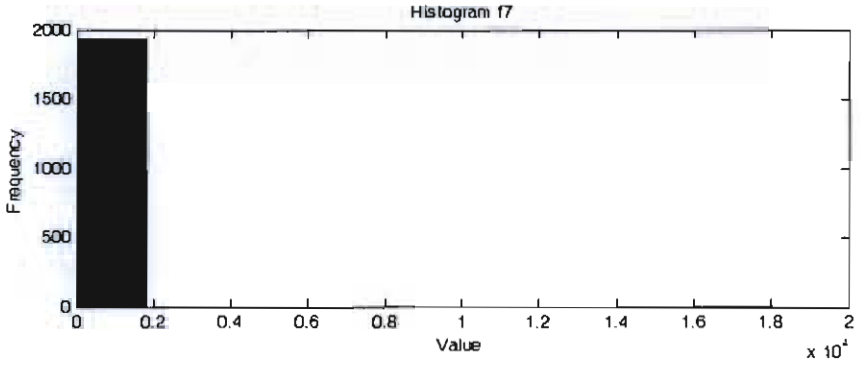
ภาพที่ ก-40 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 4 (f4)



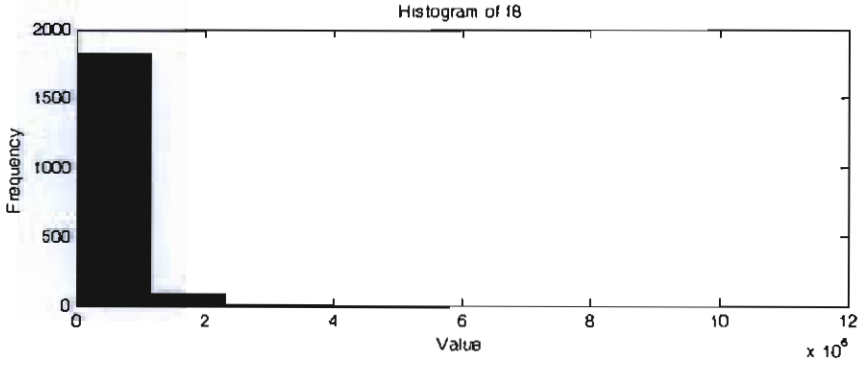
ภาพที่ ก-41 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 5 (f5)



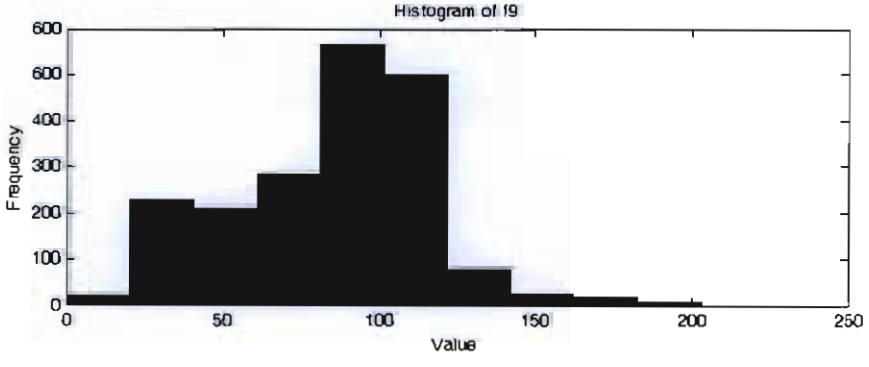
ภาพที่ ก-42 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 6 (f6)



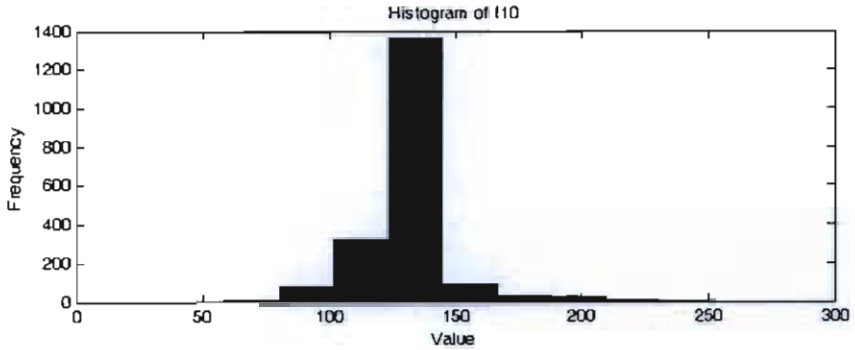
ภาพที่ ก-43 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 7 (f7)



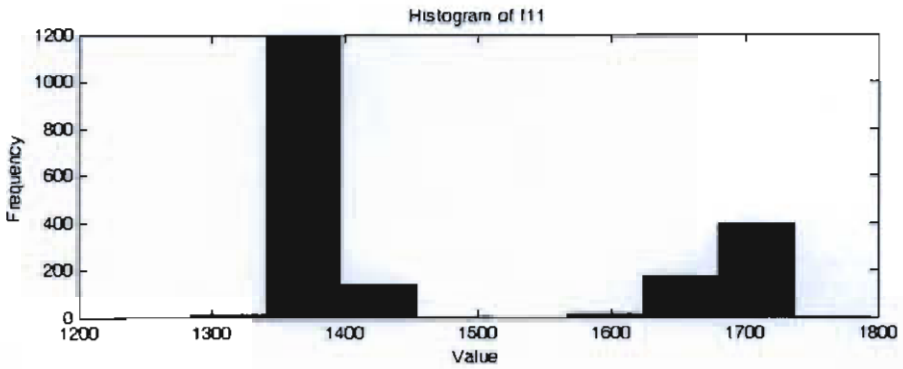
ภาพที่ ก-44 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 8 (f8)



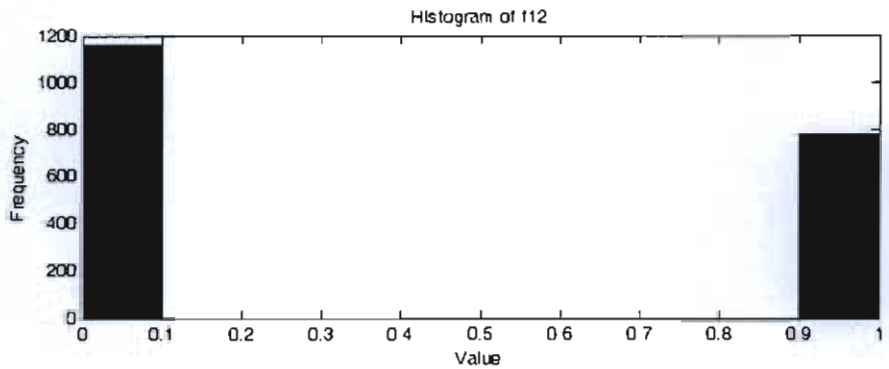
ภาพที่ ก-45 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 9 (f9)



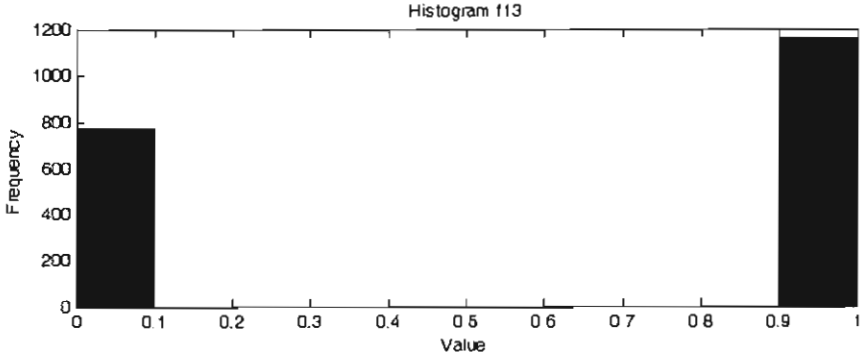
ภาพที่ ก-46 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 10 (f10)



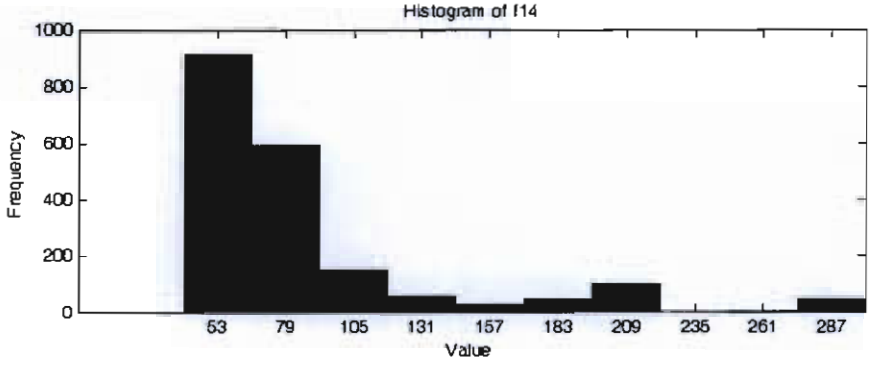
ภาพที่ ก-47 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 11 (f11)



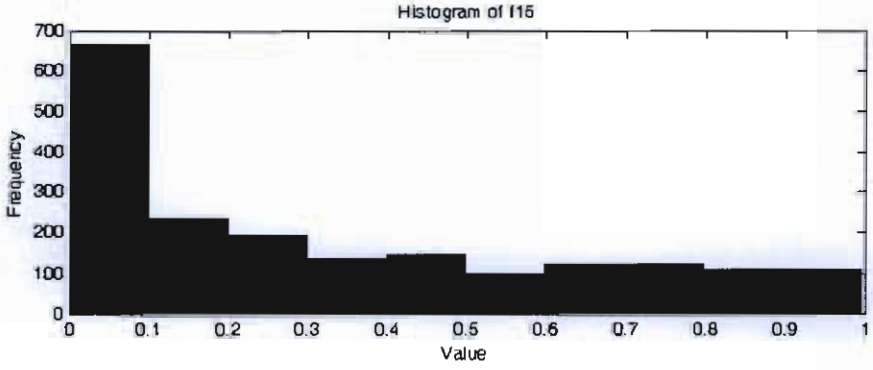
ภาพที่ ก-48 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 12 (f12)



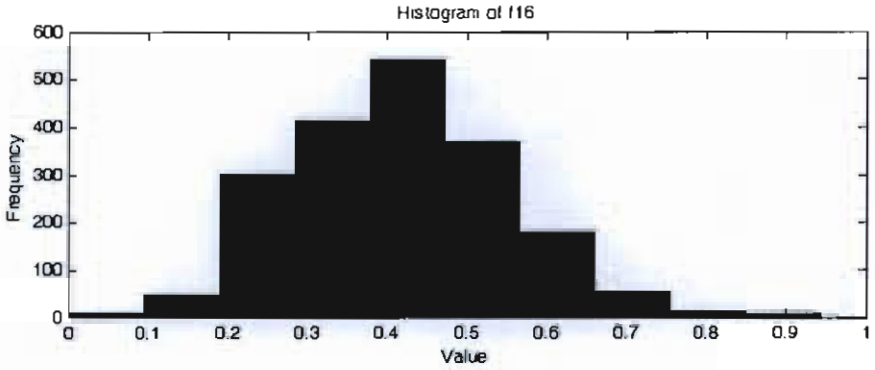
ภาพที่ ก-49 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 13 (f13)



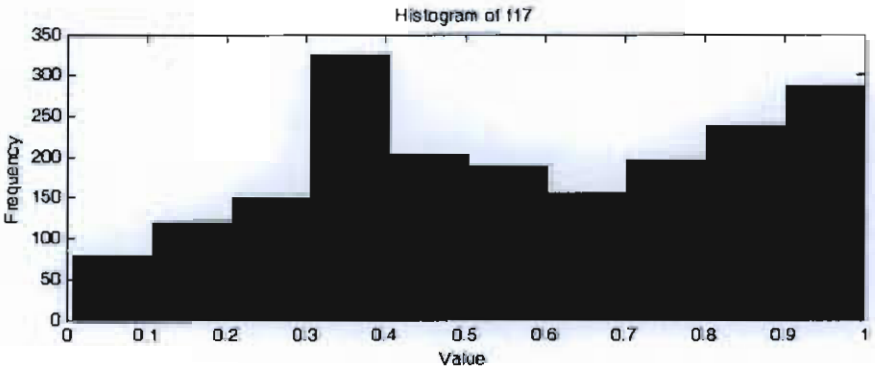
ภาพที่ ก-50 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 14 (f14)



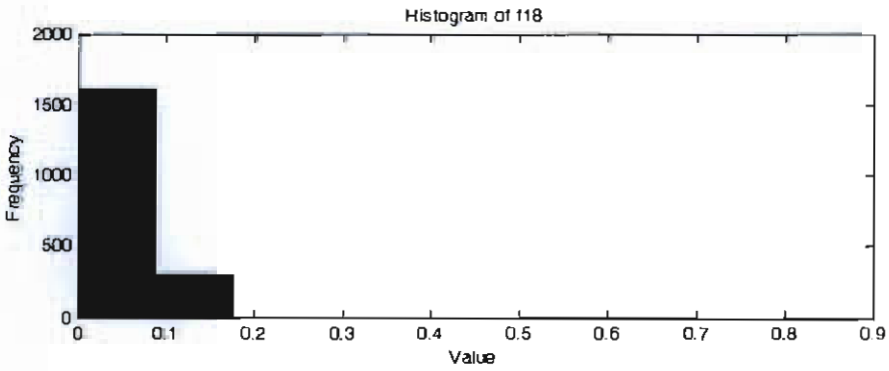
ภาพที่ ก-51 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 15 (f15)



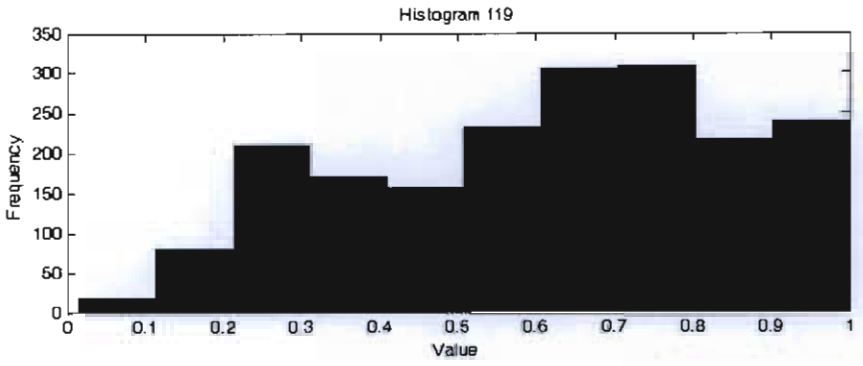
ภาพที่ ก-52 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 16 (f16)



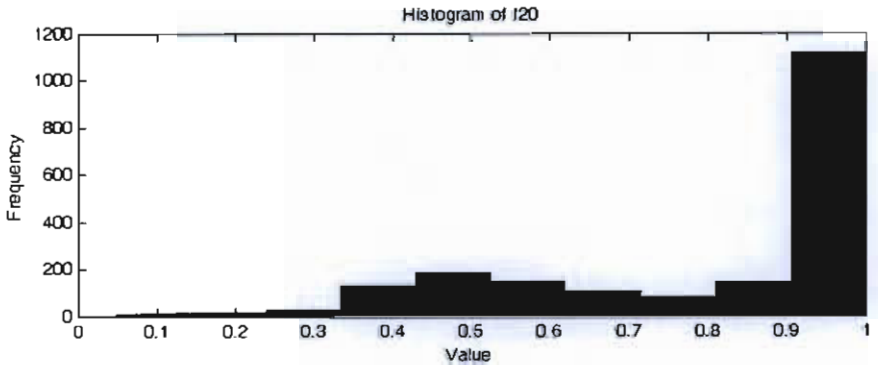
ภาพที่ ก-53 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 17 (f17)



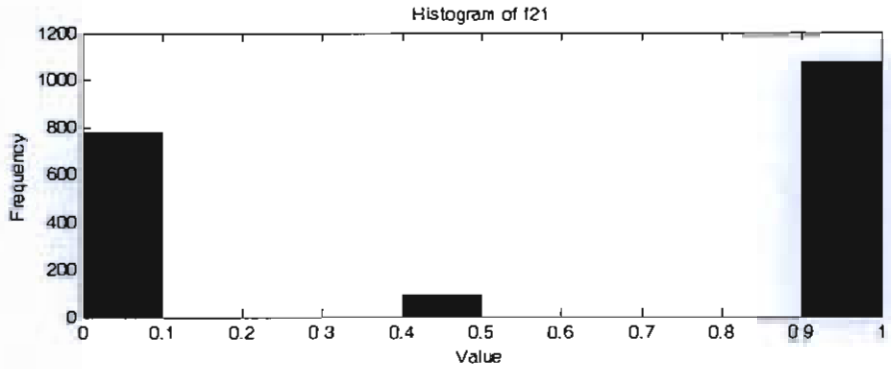
ภาพที่ ก-54 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 18 (f18)



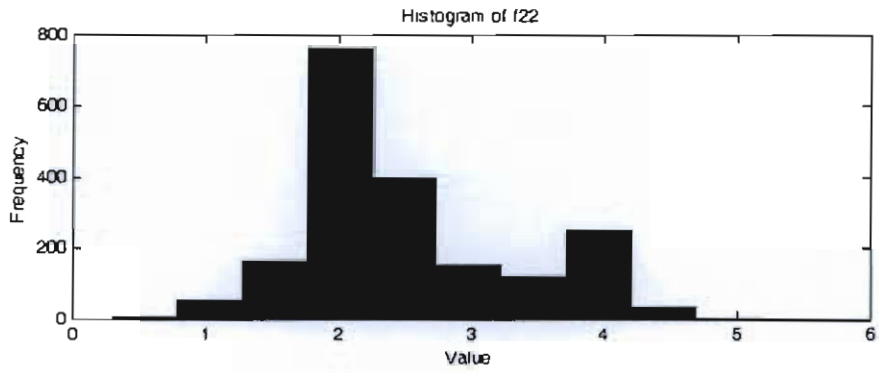
ภาพที่ ก-55 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 19 (f19)



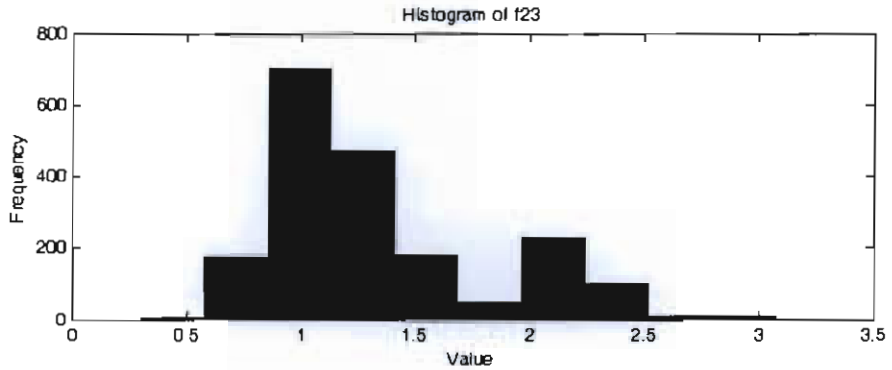
ภาพที่ ก-56 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 20 (f20)



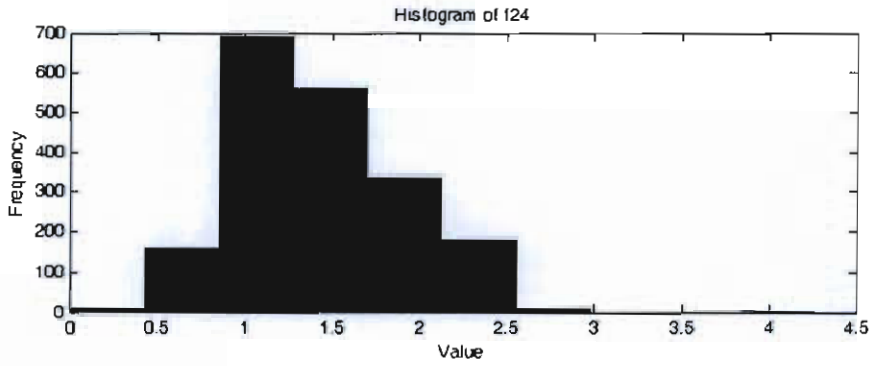
ภาพที่ ก-57 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 21 (f21)



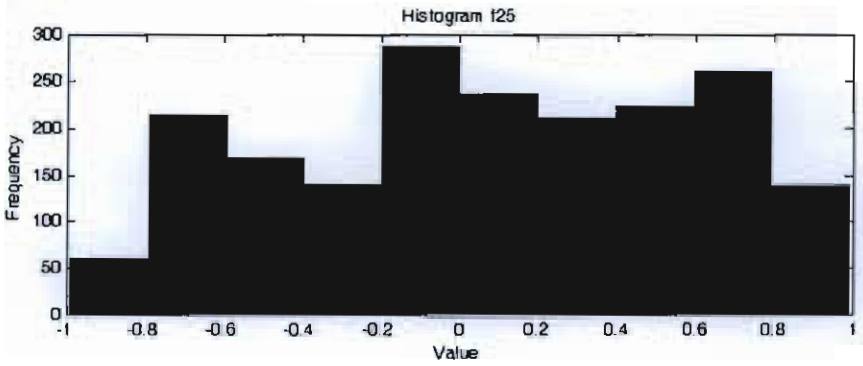
ภาพที่ ก-58 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 22 (f22)



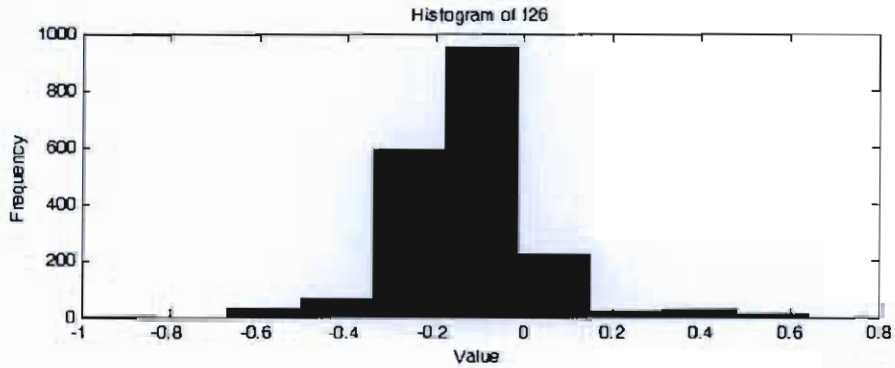
ภาพที่ ก-59 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 23 (f23)



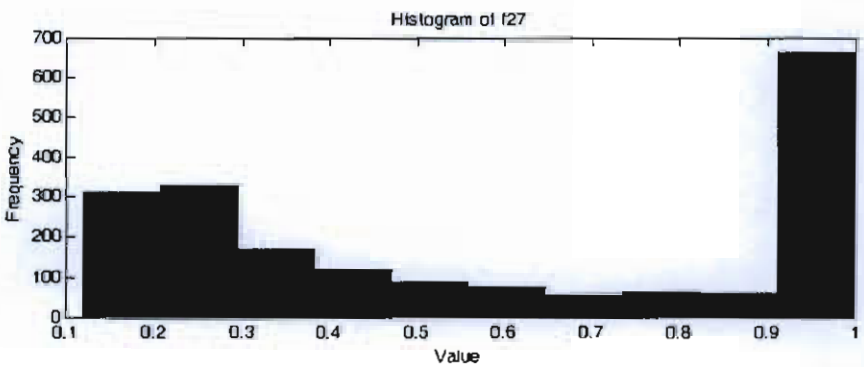
ภาพที่ ก-60 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 24 (f24)



ภาพที่ ก-61 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 25 (f25)



ภาพที่ ก-62 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 26 (f26)



ภาพที่ ก-63 Histogram ของลักษณะข้อมูล Faults ลักษณะที่ 27 (f27)

ภาคผนวก ข

การเผยแพร่ผลงานวิจัย

เอกสารประกอบการประชุมวิชาการ

Knowledge and Smart Technology

ครั้งที่ ๕ (KST-2556)

๓๑ มกราคม - ๑ กุมภาพันธ์ ๒๕๕๖

PROCEEDINGS

สารบัญ

การเลือกลักษณะของข้อมูลผู้บุกรุกด้วย Heuristic Greedy Algorithm of Item Set

Intrusion Feature Selection using Heuristic Greedy Algorithm of Item Set

จรรยา อันปันส์[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 o feature selection ond extraction methods of network intrusion doto which are the heuristic greedy algorithm (HGAIS) of item set ond principol component onalysis (PCA), respectively. After proposed feature selection and extroction steps, we use three standard supervised learning algorithms which ore BPNN, RBF ond 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 obtoin 19 ond 13 feotures, respectively. In addition, the clossificotion accuracies confirm thot HGAIS olgarithm produces better features than the PCA.

Key Words: Feoture Extroctian, Feature Selection, Pattern Recognition, Network Intrusian 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 และที่ไม่ใช่การบุกรุกและเนื่องจากข้อมูลที่ส่งผ่านทางเครือข่ายอินเตอร์เน็ตหรือข้อมูลที่ตรวจสอบนั้นมีปริมาณมากทั้งจำนวนข้อมูล 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 และเลือกลักษณะที่ดีที่สุดเมื่อนำมาทำการแบ่งกลุ่มข้อมูลในขณะนั้นการพัฒนาการเลือกลักษณะชุดข้อมูลเครือข่ายประกอบไปด้วย 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

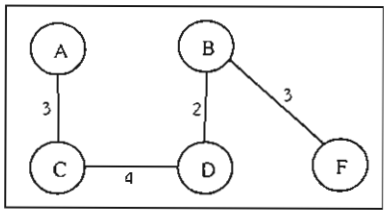
Jing Zhang, Jianmin Wang, Deyi Li, Huacan He, Jiaguang Sun (2003) [2] ได้นำเสนองานวิจัยเรื่อง A New Heuristic Reduct Algorithm Base on Rough Sets Theoryเนื่องจากการนำทฤษฎีเซตอย่างหยาบมาเพื่อหาซับเซตของลักษณะที่เหมาะสมที่สุดจากการเลือกลักษณะเป็น

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 โดยพิจารณาถึงการเพิ่มอัตราการตรวจจับและลดเวลาการประมวลผลโดยนำข้อดีของ 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

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

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)

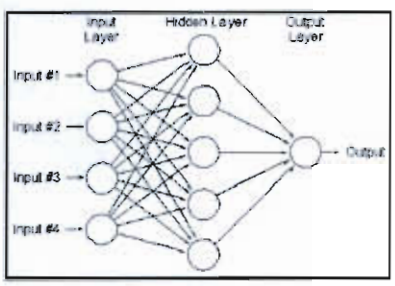
โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ เป็นขั้นตอนวิธีในการเรียนรู้ของเครือข่ายประสาทเทียมวิธีหนึ่งที่นิยมใช้ในโครงข่ายประสาทเทียมหลายชั้น [6] ประกอบไปด้วยชั้นข้อมูลเข้า ชั้นซ่อน และชั้นข้อมูลออก ดังรูปที่ 2 ซึ่งชั้นซ่อนอาจมีชั้นเดียวหรือมากกว่าก็ได้ เพื่อปรับค่าน้ำหนักในเส้นเชื่อมให้มีค่าผิดพลาดกำลังสองเฉลี่ยน้อยที่สุด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

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

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

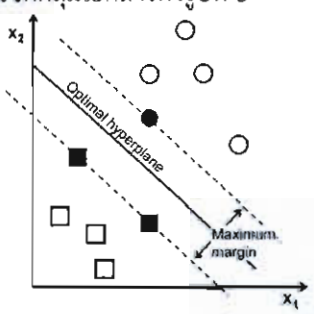
$$d = \frac{|w^T x_i + b|}{\|w\|} \quad (6)$$

w คือ เวกเตอร์น้ำหนัก

x_i คือ ข้อมูลนำเข้า

b คือ ค่าคงที่ที่กำหนดขึ้นเพื่อให้เหมาะสมกับการจัดกลุ่ม

4. เลือกจุดที่อยู่ใกล้เส้นตรง 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)

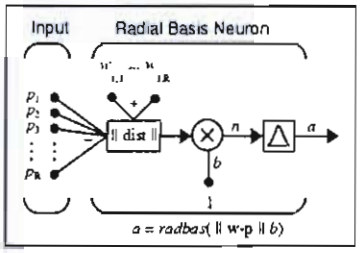
โดยแบบที่นิยมใช้โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน เป็นโครงข่ายประสาทเทียมป้อนไปข้างหน้าแบบหลายชั้น [8] จะประกอบไปด้วย 3 ชั้น ได้แก่ ชั้นรับข้อมูลเข้า ชั้นซ่อน และชั้นข้อมูลออก ดังรูปที่ 4 โดยเป็นฟังก์ชันการส่งระหว่างชั้นรับข้อมูลเข้า $p \in \mathbb{R}^{Nx1}$ ไปยังชั้นข้อมูลออก $y \in \mathbb{R}^{Mx1}$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

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

โดยที่ w_{ik} คือ ค่าน้ำหนักนิวรอนในชั้นซ่อน

s คือ จำนวนนิวรอนในชั้นซ่อน

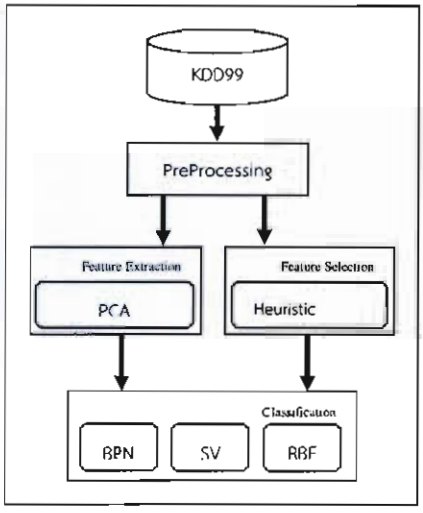
c คือ เวกเตอร์จุดศูนย์กลาง



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.

ใบตอบรับการส่งผลงานวิจัยในงานประชุมวิชาการ International Symposium on Communications and Information Technologies (ISCIT), 2013

Dear Ms.Janya Onpans

Affiliation: Burapha University

Paper ID: 1685

Title: Intrusion Feature Selection Using Modified Heuristic Greedy Algorithm of Itemset

The ISCIT 2013 Technical Program Committee has completed the review process, and we are pleased to inform you that the manuscript listed above has been ACCEPTED for presentation. Congratulations!

Total ISCIT 2013 submission is over 210 papers. Technical program committee carefully selected about 163 papers with 77.6% acceptance.

Information on many aspects of ISCIT 2013 are available on the conference web site http://www.iscit2013.org and more will come very soon. The conference information will be updated shortly to include the full technical program.

Included at the end of this e-mail message are the reviewers' comments on your paper. Please revise your paper according to reviewers' comments.

Also importance! The final manuscript MUST BE LIMITED to only 6 (six) pages. Violating this limit WILL cause your paper being unpublished***

The final manuscript, copyright form, and copy of e-mail response from IEEE PDF Express MUST be uploaded to the http://www.iscit2013.org web site ONLY by July 9, 2013. We do not accept any submission by e-mail or any other way.

CRITICAL INFORMATION:

1) Due to new regulation of IEEE, all papers to be included in IEEExplorer database must use the new IEEE conference template, located at:

http://www.ieee.org/conferences_events/conferences/publishing/templates.html

2) Registration Policy:

-At least ONE author of an accepted paper needs to register for the conference, and in case of using a Student Rate registration, this can cover ONE paper only when the student author is the first author of the paper.

-In the case of multiple accepted papers by one author, one Regular Rate registration will cover up to TWO papers.

3) Additionally, the paper MUST BE PRESENTED at the conference by one of the authors, otherwise it ("no show" paper) will NOT be included in IEEE Xplore.

4) There are two presentation types: Oral and Poster. You may check your presentation type, on June 22, 2013 onward at the following URL:

http://iscit2013.org/content/technicalprogram.php

In case that you want to change presentation type, please contact us before July 9, 2013.

Again, congratulations; we look forward to welcoming you to Samui, Thailand.

Sincerely,

Technical Program Chairs

---- Comments from the Reviewers: ----

-------- Review for Paper#1685--------

Self-Evaluation: Please assess your competence in the research problem addressed in the reviewed paper :

-Strong

Is this paper relevant to this conference? (If it is not, then please specify the reason and do not need to answer any following questions.)

-Yes

Evaluation: Evaluation of work and contribution :

-Good work of some importance

Originality: Novelty :

-New good idea combined with existing methods for better performance and/or cost

Readability: Readability and organization :

-Basically well written

Summary: Overall recommendation :

-Accept (good quality)

Award: Award Recommendation: Is this paper suitable for an award :

-No

Strengths of the paper:

Well written. Extensive simulation and result.

Weaknesses of the paper:

none

Comment for paper improvement:

--

Intrusion Feature Selection Using Modified Heuristic Greedy Algorithm of Itemset

Janya Onpans[1], Suwanna Rasmequan[2],
Benchaporn Jantarakongkul[3], Krisana Chinnasarn[4]
Faculty of Informatics, Burapha University
Chonburi, Thailand
[1]mai.janya@gmail.com, [2]rsuwanna@buu.ac.th,
[3]benchapornj@yahoo.com, [4]krisana@buu.ac.th

Annupan Rodtook
Department of Computer Science
Faculty of Science, Ramkhamhaeng University
Bangkok, Thailand
annupan@ru.ac.th

Abstract— **This paper proposes the Modified Heuristic Greedy Algorithm of Itemset (MHGIS) as a feature selection method for Network Intrusion Data. The proposed method can be use as an alternative method to gain the proper attributes for the proposed domain data: Network Intrusion Data. MHGIS is modified from original Heuristic Greedy Algorithm of Itemset (HGIS) to increase efficiency for finding proper feature. In our work, we compare our result with the common method of feature selection is which the Chi-Square (Chi^2) feature selection. There are 4 main steps in our experiment: Firstly, we start with data pre-processing to discard unnecessary attributes. Secondly, MHGIS feature selection and Chi^2 feature selection have been employed on the pre-processed data, to reduce the number of attributes. Thirdly, we measure the recognition performance by using supervised learning algorithms which are C4.5, BPNN, RBF and SVM. Lastly, we evaluate the results received from MHGIS and Chi^2. From the KDDCup99 dataset, we got 13,499 randomly sampling patterns with 34 data dimensions. With the use of MHGIS and Chi^2 algorithms, we obtain 14 and 26 features respectively. The result shows that, the classification accuracies measure by C4.5 over the MHGIS selection algorithm produces better accuracies as compare to the Chi^2 feature selection and HGIS feature selection over all types of classification methods.**

Keywords—Feature Selection; Pattern Recognition; Network Intrusion Detection; Heuristic Greedy

I. INTRODUCTION

With the growing demands of computer networks made many people realize more about the significant of the security in networks. Intrusion detection is a preferable choice for most people to use in computer network security. Intrusion detection system divides into two types: misuse detection and anomaly detection. Misuse detection system is a method which is use for detecting abnormal pattern by comparing them with well-known pattern. This made it unable to detect any unknown attack that has no matched pattern found in the system. In contrast, anomaly detection system is a method that detect deviated pattern from the normal behavior. So that it can detect any new intrusion behavior. However, this method will depend pretty much on the structure design. If the structure is not well designed, then some types of attack may not be detected.

Intrusion detection system can be seen as a classification problem in which it will distinguish the attracted activities from the normal activities. Data that transmit via network is very large; as a result, it will cause the delay the indentifying of intrusion and may allow any intruder to attack the network. This problem of enormous data that may slow down the detection process needs a method to eliminate useless features of the network data to be able to increase accuracy of classification while speeding up intrusion detection process. Furthermore, a good detection system should have higher detection rate and lower false alarm. The methods that can be use to solve these problems are either feature selection or feature extraction. Feature selection is a method that will produce a subset of features whereas feature extraction will create new features. Both of these methods will discard irrelevant or redundant features, so that only essential data will be left for further processes.

It has been known that finding proper representative of data attributes is very important to the intrusion detection system development. In other word, good data representations provide higher or better degree of recognition. In our previous work [3] we have proposed HGIS to compare with feature extraction method: PCA. The result shows an improvement of 5.05% with RBF. In this work we propose MHGIS to improve the result as compare to the feature selection: Chi^2. MHGIS is a method for select feature which considers error information of each itemsets in order to give the best answer in each iteration. Itemsets used in each iteration is constructed from apriori algorithm. MHGIS method divides into two steps: (1) finding based itemset and (2) adding discard item. With this method, the discard item will be included back to reconsider. This made an extensive cover of the itemsets. In classification step, selected features from MHGIS are compared with selected features from the Chi^2 algorithm. We verify the performance of our proposed feature selection, by calibrating with the other four classic classification methods: C4.5 decision tree, back-propagation neural network (BPNN), radial basis function (RBF) network and support vector machine (SVM). Performance metrics used in this paper are accuracy rate, detection rate, false-alarm rate and CPU processing time.

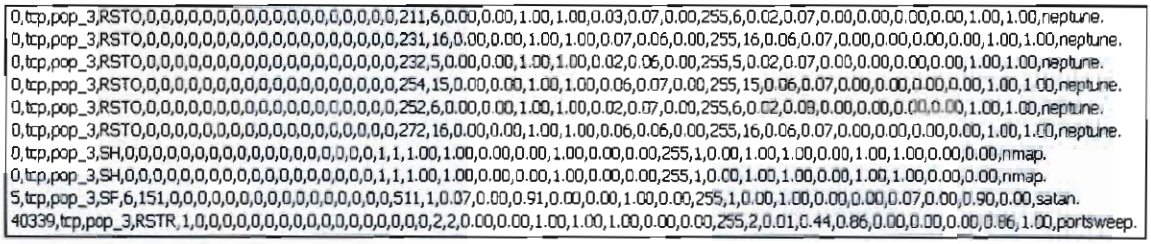


Fig. 1. Example of KDD Cup 99 Dataset

The paper is organized as follows: Related works are discussed in section II. Theoretical Background is discussed in section III. Proposed Method is presented in section IV. Experimental results are discussed in section V. Finally, the conclusion is summarized in the last section.

II. Related Works

In recent years, there is still a number of research works that try to improve the accuracy rate. The main stream of these works is focused on reducing the insignificant features as much as possible to be able to increase performance of rapid detection. So in this section, we will discuss related research works that are attempted to find proper features.

Murat Karabatake et al. [1] proposed a new feature selection base on association rules and neural network for the diagnosis of erythemato-squarmous diseases. The dimensions of input features space are reduced from 34 to 24 by using association rules. The correct classification rate of proposed system is 98.61%.

Hari Om et al. [2] proposed a hybrid system for reduction the false alarm rate of intrusion detection system. Feature selection use entropy and combines k-Means and two classifiers: K-nearest neighbor and Naïve Bayes for anomaly detection. Combination of tree classifiers, the detection rate reaches 98.18% and false alarm rate 0.83%.

Janya Onpans et al. [3] proposed the feature selection and extraction methods of network intrusion data which are the heuristic greedy algorithm of itemset (HGIS) and principal component analysis (PCA), respectively. After HGIS feature selection and PCA feature extraction steps that it obtain 19 and 13 features, respectively, and then use three standard supervised learning algorithms which are BPNN, RBF and SVM for evaluating with KDDCup99 dataset. Experimental results shown that HGIS algoritbm produces better features than the PCA.

III. Theoretical Background

In this section, we describe the essential background that use in the proposed method.

A. KDD Cup 99 Dataset

The Knowledge Discovery in Database (KDD) Cup data is a common benchmark for evaluation of intrusion detection system [4]. This dataset is prepared by MIT Lincoln Labs to simulate the attack to U.S. Air Force local area network. There are about 4,900,000 records. Each connection is labeled as either normal or attack that consists 41 features. Features are in form of continuous, discrete, and symbolic are shows in Fig. 1 and it divides into three categories:

- Basic Features: are fundamental features can be derived from packet data communication network such as protocol type.
- Traffic Features: are features that compute with respect to a window interval such as time in connect.
- Content Features: are features to be able to look for suspicious behavior in the data portion such as the number of failed login attempt.

The last feature is labeled as either normal or attack. Attack type that can be classified in four main categories [5]:

- DoS (Denial of Service): is class of attack that attacker makes memory resource too busy to accept legitimate request.
- U2R (User to Root): is class of attack that attacker starts with access to a normal user account and exploit some vulnerability to gain root access to the system.
- R2L (Remote to Local): is class of attack that attacker sends packets to a machine over a network, then exploit some vulnerability to gain local access to a machine.
- Probing: is class of attack that attacker attempts to gather information about a network and find the known vulnerabilities for the purpose of circumventing its security controls.

B. Heuristic Greedy Algorithm

Heuristic Greedy Algorithm is a simple and straightforward method for finding solutions such as finding the shortest path in graph, finding solution in dynamic programming problem, and so on [6]. It considers error information from current learning step in order to give the best answer at the time. It can be used for finding optimal solutions from very large database or big dataset because it does not consider all of data attributes. It is an iterative method which starts from finding the best solution in the first iteration. Then, the first solution is used for finding the best solution for the second iteration. It will be repeated until the optimal solutions are converged. Solutions from the first iteration, second iteration, third iteration, and so on are called 1-itemsets, 2-itemsets, 3-itemsets, and n-itemsets, respectively.

Heuristic Greedy Algorithm of itemset is feature selection technique by using apriori as following [3]:

Step1: generate 1-itemsets; find support value as in (1) of each item with RBF.

$$f(\{itemset\}) = rmse(\{itemset\}) \quad (1)$$

Where $rmse$ is root mean square error.

Step 2: generate 2-candidate itemsets; find support value as equation (1) with RBF of each pair 1-itemset.

Step 3: generate 2-itemset by union of 2-candidate itemsets that have support value less than or equal to 1-itemset subsets of the 2-candidate itemsets as in (2).

$$f(\{a,b\}) \leq f(\{a\}) \; and \; f(\{a,b\}) \leq f(\{b\}) \quad (2)$$

Step 4: repeat step 2 and increase size of itemsets until cannot generate itemsets.

Step 5: Select the best itemset that has lowest root mean square error ($rmse$).

C. Itemset Creation using Apriori Algorithm

Apriori Algorithm is the most commonly used in finding association rules between data attributes. It has been known that it works iteratively. First, it finds the set of attributes of size 1-itemsets. Then, set of attributes of size 1-itemsets are used as the base for finding set of attributes of size 2-itemsets. Next, set of attributes of size 3-itemsets will be caleulated based on set of attributes of size 2-itemsets. The algorithm will repeat until set of attributes of size n-itemsets could not be computed. As mention before, it can be said that Apriori algorithm is a simple technique but powerful method for creating smaller candidate subset from very large sets which were found from the previous iteration. In addition, it can be used for eliminating infrequent itemsets. Frequent itemsets are itemsets that their support values are less than or equal to support value of previous itemsets [7].

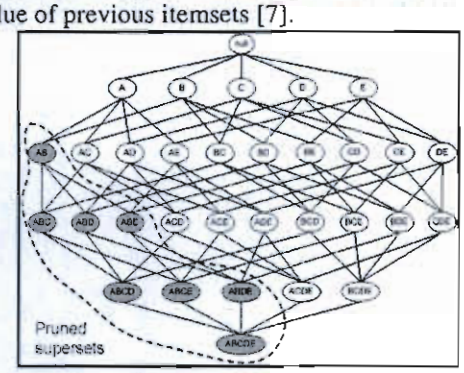


Fig. 2. Itemset lattice with eliminate Infrequent itemsets

On the other hand, support values for infrequent itemsets are higher than the previous support values. In [8] describes frequent itemsets as follow: if any itemsets are frequent itemsets then every subsets must be frequent itemsets. In other word, if a subset is infrequent itemsets then the following itemsets are infrequent itemset as well. For example in Fig.2, if {A,B} is infrequent itemset then {A,B,C}, {A,B,D}, {A,B,E}, {A,B,C,D}, {A,B,C,E}, {A,B,D,E} and {A,B,C,D,E} are infrequent itemset. Hence, we can eliminate or prune all subsets within that itemsets. This eliminating infrequent itemset step is sometime called support-based pruning algorithm. In Fig. 2 show itemset lattice with eliminate infrequent itemsets witch receive smaller candidate set. Therefore it can remove the itemsets and not consider all superset of itemsets.

D. Chi-square Feature Selection

The Chi-Square (χ^2) statistic is a common technique for find relationship between two variables. In this paper, we use for feature selection in data of high dimension. The Chi-Square feature selection algorithm evaluates the worth of a feature by computing the value of the Chi-Square statistic with respect to the class [9]. Then we remove all irrelevant and least relevant features from the dataset that considers from Chi-Square value. It is tested by Chi-Squared formula as is shown in equation (3):

$$\chi^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \quad (3)$$

Where: O_{ij} is the observed frequency,

E_{ij} is the expected frequeney.

$$E_{ij} = \frac{(R_{T_i})(C_{T_j})}{N} \quad (4)$$

Where: R_{T_i} is number patterns in the ith interval,

C_{T_i} is number patterns in the jth elass,

N is total number patterns.

IV. Proposed Method

In this section, we will elaborate on our proposed method for intrusion deteetion whieh consists of four phases as illustrated in Fig. 3. Four phases of our proposed method used in this paper are: data preprocessing, feature selection, classification, and system evaluation. Firstly, insignificant attributes will be removed and then applied with data sampling technique. Secondly, feature selection based on Chi^2, HGIS and MHGIS feature selection are to get the optimal attributes. Thirdly, data classification based on C4.5, BPNN, RBF and SVM are used to classify the network data. In the last phase, accuracy rate, detection rate, false alarm rate and CPU processing time are calculated to evaluate the performance of intrusion detection.

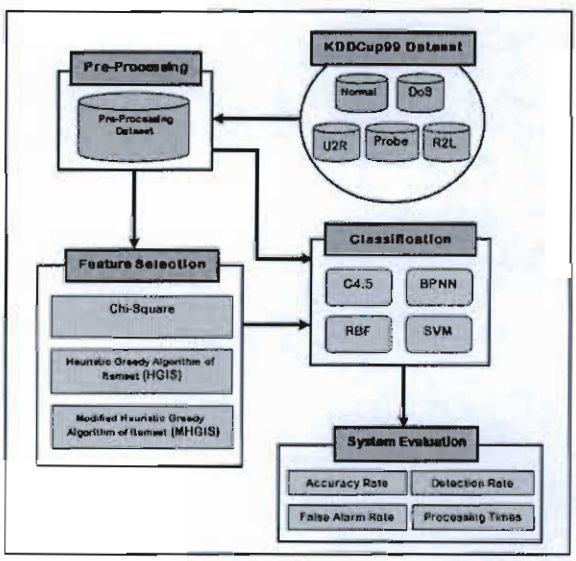


Fig. 3. Proposed model of intrution detection

A. *Data Preprocessing Phase*

The KDD Cup 99 dataset is standard dataset used for evaluating intrusion detection algorithms. It is about 5 million records. It can be said that it is a very large dataset. It will consume a lot of CPU time if we use all of them in recognition procedure. Then, many researchers recommend to choose only 10 percents and then sampling about 13,499 patterns for learning and testing the performance of the recognition system [11][12][13]. KDD Cup 99 consists of 5 groups. After the last sampling step, numbers of instants in each group are shown in TABLE II. In this paper, we discard some basic features and zero value features because these features have no significant effect to the learning performance. Hence, the remaining features for next step are only 34 features.

TABLE I. AMOUNT DATA OF EACH GROUP

Class Name	Amount
Normal	4,107
DoS	4,107
U2R	4,107
R2L	1,126
Probe	52

B. *Feature Selection Phase*

We propose two algorithms for extracting and selecting features, detailed as following:

- Feature Selection using Chi-square:

Step 1: Compute chi-square value every pair of attribute and class as in equation (3) then sort value in descending order.

Step 2: remove irrelevant features from dataset that there chi-square value equals zero.

Step 3: remove least relevant feature from the dataset that satisfy the condition

$$\frac{\chi_i^2 \times \log(N^2)}{\sum \chi_i^2 \times N} \times 100 < \delta \qquad (5)$$

Where we set $\delta = 0.1$ to satify our criterion, χ_i^2 is chi-square value for feature in consideration and N is the total number of attributes.

- Feature Selection using Modified Heuristic Greedy Algorithm of Itemset:

In feature selection step, we used MHGIS for select proper feature. Difference between HGIS and Modified HGIS criterion for itemsets generation and itemsets addition step that added to modified HGIS are shows in TABLE II. Improvement of criterion for itemset generation more elaborate in modified HGIS is added with delta. Final step of modified HGIS is item addition that dropped off by finding consequents item that it can reduce $rmse$.

TABLE II. DIFFERENCE BETWEEN HGIS AND MODIFIED HGIS

	HGIS	Modified HGIS
Criterion for itemset generation	$f(\{a,b\}) \leq f(\{a\})$ *and* $f(\{a,b\}) \leq f(\{b\})$	$f(\{a,b\}) \leq f(\{a\}) + \alpha$ *and* $f(\{a,b\}) \leq f(\{b\}) + \alpha$
Itemset dddition	-	Finding consequent item that it can reduce rmse

As described in previous section, 34 features have been selected. Then, feature selection using modified heuristic greedy algorithm by apriori algorithm consists two steps will be described as following:

1. Finding based itemset step

Step1: generate 1-itemsets; find support value as in (1) of each item with RBF.

Step 2: generate 2-candidate itemsets; find support value as equation (1) with RBF of each pair 1-itemset.

Step 3: generate 2-itemset by union of 2-candidate itemsets that have support value less than or equal to 1-itemset subsets of the 2-candidate itemsets as in (6).

$$f(\{a,b\}) \leq f(\{a\}) + \alpha \; and \; f(\{a,b\}) \leq f(\{b\}) + \alpha \qquad (6)$$

where we set $\alpha = 0.2$ to satify our criterion.

Step 4: repeat step 2 and increase size of itemsets until cannot generate itemsets.

2. Itemset addition step

The best itemset that derived from the finding based itemset step are used as the base for finding consequent set that it can reduce $rmse$.

Step 1: get 1-itemsets with $rmse$ is less than the based itemset add to based itemset and find support value as in (1).

Step 2: select lowest $rmse$ itemset is the base for finding consequent itemset.

Step 3: add item that its $rmse$ is less than the $rmse$ in previous step.

Step 4: repeat step 2 until get the best itemset that it is the lowest $rmse$.

C. *Classification Phase*

In classification phase, we compare the performance metrics with four classification methods named C4.5 [13], BPNN [14], RBF [15] and SVM [16]. Learning Parameters of each method are defined as follows:

- BPNN
  Number of hidden Layers = (attributes + classes) / 2
  Learning Rate = 0.3
  Momentum = 0.2
- SVM
  The polynomial kernel
- RBF
  Gaussian function
- C4.5
  Confidence threshold for pruning = 0.25

Minimum number of instances per leaf = 2
Number of folds for reduced error pruning = 3

In classification experiment, data is divide into 10 folds cross validation for training and testing.

D. Performance Evaluation

The standard metrics and most to use that have been developed for intrusion detection system evaluation are accuracy rate, detection rate and false alarm rate.

Accuracy rate is computed as the ratio between amounts of correctly classified and total number of all data can be found from:

$$\text{Accuracy } Rate = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

Detection rate is computed as the ratio between the number of correctly detected attacks and the total number of attacks can be represented by:

$$Detection\ Rate = \frac{TP}{TP+FN} \quad (8)$$

False alarm rate is computed as the ratio between the numbers of normal connections that is incorrectly misclassified as attack and total number normal connection by following equation:

$$False\ Alarm\ Rate = \frac{FP}{FP+TN} \quad (9)$$

Definitions of variables are show in TABLE III.

TABLE III. CONFUSION MATRIX

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

V. EXPERIMENTAL RESULTS

After proposed feature selection, we use four standard supervised learning algorithms which are C4.5, BPNN, RBF and SVM for evaluating the significance of the selecting features. From the KDDCup99 with 34 data dimensions based on Chi^2 and MHGIS algorithms, we obtain 26 and 14 features, respectively. Furthermore, we compared with original HGIS feature selection that it gets 13 features.

TABLE IV. ACCURACY RATE

Learning Method	Accuracy Rate (%)			
	All (34)	Chi^2(26)	HGIS (13) [4]	MHGIS (14)
C4.5	99.43	99.44	98.76	99.52
BPNN	98.69	96.54	97.34	97.43
RBF	93.69	94.52	95.53	95.59
SVM	97.02	95.04	94.32	95.24

In TABLE IV shows accuracy rate of experimental. It can be seen that most accuracy rate using MHGIS feature selection batter than other feature selection. And MHGIS with C4.5 decision trees classification is the best with accuracy rate 99.52%. MHGIS with RBF has accuracy rate more than original data 1.9%.

TABLE V. DETECTION RATE

Learning Method	Detection Rate (%)			
	All (34)	Chi^2(26)	HGIS (13) [4]	MHGIS (14)
C4.5	99.57	99.53	99.30	99.63
BPNN	99.33	98.17	98.41	99.05
RBF	93.05	95.50	96.40	95.37
SVM	98.75	97.85	96.87	97.72

Detection rate of experiment displays in TABLE V. that MHGIS is better than HGIS and resemble Chi^2. MHGIS feature selection and C4.5 classification has the best result detection rate 99.63%.

TABLE VI. FALSE ALARM RATE

Learning Method	False Alarm Rate (%)			
	All (34)	Chi^2(26)	HGIS (13) [4]	MHGIS (14)
C4.5	0.96	0.84	1.56	0.84
BPNN	1.50	2.87	3.55	2.11
RBF	16.49	10.52	11.49	10.29
SVM	2.77	4.91	7.47	5.31

In TABLE VI, Selected data with MHGIS 9 features and Chi^2 26 features have minimum false alarm rate 0.84%. The best performance of processing times in all pattern recognitions is HGIS with 13 features that show in TABLE VII.

TABLE VII. PROCESSING TIMES

Learning Method	Processing Times(s)			
	All (34)	Chi^2(26)	HGIS (13) [4]	MHGIS (14)
C4.5	38.63	31.39	16.61	18.43
BPNN	166.68	114.87	65.28	74.90
RBF	47.28	38.98	12.74	15.62
SVM	41.48	32.96	13.18	17.05

VI. CONCLUSION

In this paper, the proposed method of feature selection using modified Heuristic Greedy Algorithm of Itemset (MHGIS) implementing with apriori algorithm to improve the detection rate, accuracy rate and false alarm rate, as compare to feature selection using Chi^2 and HGIS. The experimental results indicate that the feature selection based on MHGIS yields a better all of assessment but processing times are a

little lower than HGIS. The most results have good result with decision tree C4.5 because KDDCup99 dataset has high spreading. But for the detection rate and the false alarm rate are improved only in some cases. So in the future works, we intend to improve an algorithm to have a better performance on both detection and false alarm rate. In our future work, we are interested in find out the algorithm that can improve the computing time during the selection process too.

ACKNOWLEDGMENT

This work is funded by the National Research Council of Thailand (NRCT), fiscal year 2012.

REFERENCES

[1] 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.

[2] 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.

[3] Janya Onpans, Annupan Rodtook , Suwanna Rasmequan, Benchaporn Jantarakongkul and Krisana Chinnasarn, "Intrusion Feature Selection using Heuristic Greedy Algorithm of Item Set", Knowledge and Smart Technology (KST) 5th, pp.22-29, 2013.

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

[5] Mahbod Tavllace, Ebrahim Bagheri, Wei Lu Ali A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data set", Proceeding of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA), 2009.

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

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

[8] K. P. Soman, S. Diwakar, and V. Ajay, "Insight into Data Mining Theory and Practice", Prentice-Hall of India, ISBN: 81-203-2897-3, 2006.

[9] H. Liu, R. Setiono, "Chi2: Feature selection and discretization of numeric attributes," IEEE 7th International Conference on Tools with Artificial Intelligence, pp. 338-391, 1995.

[10] Ranjit Abraham , Jay B. Simha, S. Sitharama Iyengar, "Effective Discretization and Hybrid feature selection using Naïve Bayesian classifier for Medical datamining", International Journal of Computational Intelligence Research, ISSN 0974-1259 Vol.5, No.2, pp. 116–129, 2009.

[11] Amir-Massoud Bidgoli, Mehdi Naseri Parsa, "A Hybrid Feature Selection by Resampling, Chi squared and Consistency Evaluation Techniques", World Academy of Science, Engineering and Technology 68, 2012.

[12] M.Revathi, T.Ramesh, "Network Intrusion Detecion System Using Reduced Dimensionality", Indian Journal of Computer Science and Engineering (IJCSE), ISSN: 0976-5166, vol. 2, no.1, 2011.

[13] J. Ross Quinlan, "C4.5: programs for machine learning", Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, ISBN: 1-55860-238-0, 1993.

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

[15] 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.

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

ผลงานทางวิชาการ

จรรยา อ้นปันส์ อัณณ์นุพันธ์ รอดทุกข์ สุวรรณา รัศมีขวัญ เบญจภรณ์ จันทรกองกุล และกฤษณะ ชินสาร. (2556). Intrusion Feature Selection using Heuristic Greedy Algorithm of Item Set. *Knowledge and Smart Technology (KST)* (หน้า 22-29).

Janya Onpans, Annupan Rodtook, Suwanna Rasmequan, Benchaporn Jantarakongkul, Krisana Chinnasarn. (2013). Intrusion Feature Selection Using Modified Heuristic Greedy Algorithm of Itemset. *International Symposium on Communications and Information Technologies (ISCIT).*