บทที่ 4 ผลการทดลอง

ในบทนี้ จะกล่าวถึงผลการทดลองเบื้องต้นที่ได้ศึกษามา โดยผลการทดลองที่นำเสนอจะประกอบด้วย ผลการทดลองจากการสกัดคุณลักษณะข้อมูล และ ขั้นตอนการรู้จำประเภทของผู้บุกรุกโดยใช้โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ

4.1 การสกัดคุณลักษณะข้อมูล

จากตัวอย่างข้อมูล KDDcup99 ซึ่งเป็นฐานข้อมูลมาตรฐานและนำมาใช้เป็นตัวอย่างในการทดลองในงานวิจัยนี้ โดยข้อมูลชุดนี้มีจำนวนประมาณ 4,900,000 ระเบียน เป็นข้อมูล 41 มิติ โดยในแต่ละชุดข้อมูล (pattern) ของ KDDcup99 นี้ จะแบ่งออกเป็นสามกลุ่มใหญ่ คือ

1. Basic Features คุณลักษณะ (attributes/features) ของข้อมูลในกลุ่มนี้จะบอกได้ว่าเป็นข้อมูลที่ได้มาจาก TCP/IP
2. Traffic Features คุณลักษณะของข้อมูลในกลุ่มนี้จะเป็นค่าของเวลาที่ได้จากการคำนวณ โดยแบ่งออกเป็นสองประเภทย่อย คือ Same host features และ same service features
3. Content Features คุณลักษณะของข้อมูลในกลุ่มนี้จะไว้บอกลักษณะของการบุกรุก เนื่องจาก DoS และ Probing จะมีพฤติกรรมการบุกรุกที่แตกต่างจาก R2L และ U2L เช่น DoS และ Probing จะมีพฤติกรรมในการโจมตีเครื่องแม่ข่ายบางเครื่อง (some hosts) ในช่วงเวลาสั้น ๆ แต่ในขณะที่การโจมตีของ R2L และ U2L จะถูกรวมมากับส่วนของข้อมูล ซึ่งโดยปกติแล้วจะโจมตีมาเพียงครั้งเดียว

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

รูปที่ 4-1 ตัวอย่างข้อมูล Intrusion สำหรับการทดสอบการรู้จำ

เนื่องจากข้อมูล KDDcup99 มีจำนวนมาก ดังนั้น ในงานวิจัยส่วนใหญ่จึงแนะนำให้เลือกข้อมูลเพียงร้อยละ 10 สำหรับเป็นข้อมูลในการสอนและทดสอบประสิทธิภาพของระบบการรู้จำ และจาก

ข้อมูลที่เลือกมาร้อยละ 10 นั้น จะทำการสุ่มมาทำข้อมูลในการสอนประมาณ 13,499 ชุด (Patterns) ทำการสกัดคุณลักษณะโดยใช้ขั้นตอนวิธีการวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis) ซึ่งได้คุณลักษณะเด่นจากข้อมูลทั้งหมด 37 มิติ สำหรับการรู้จำด้วยโครงข่ายประสาทเทียม จำนวน 19 มิติ โดยเราจะพิจารณาจากค่าไอเกนของข้อมูล (จากที่กล่าวไว้ก่อนหน้านี้ว่าข้อมูล KDDcup99 เป็นข้อมูลทั้งหมด 41 มิติ แต่มิติที่เป็น Basic Features และมิติที่เป็นคำตอบจะไม่นำมาพิจารณา ดังนั้น จึงเหลือเพียง 37 มิติ)

4.2 การรู้จำประเภทของผู้บุกรุกเบื้องต้น

ในการทดลองเบื้องต้นสำหรับการรู้จำประเภทของผู้บุกรุกในงานวิจัยนี้ ผู้วิจัยเลือกวิธีการรู้จำแบบมีผู้สอน (Supervised learning) ที่ได้รับความนิยมในการใช้ทดสอบการรู้จำ คือ โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ และ Support Vector Machine โดยมีข้อมูลนำเข้าสำหรับการรู้จำ 2 ประเภท คือ ข้อมูลทั้งหมด 37 มิติ และ ข้อมูลที่ผ่านขั้นตอนการลดมิติข้อมูล (PCA) 19 มิติ ทำให้เราสามารถแบ่งการทดลองออกเป็น 4 การทดลอง ดังนี้

1. **All+BPNN** (ข้อมูลทั้งหมด 37 มิติ และรู้จำด้วยโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ)
   Number of hiddenLayers = (attribs + classes) / 2
   LearningRate=0.3
   Momentum=0.2
   TrainingTime=500
   Training 100%
2. **All+SVM** (ข้อมูลทั้งหมด 37 มิติ และรู้จำด้วย Support Vector Machine)
   The polynomial kernel
3. **PCA+BPNN** (ข้อมูลผ่าน PCA 19 มิติ และรู้จำด้วยโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ)
   Number of hiddenLayers = (attribs + classes) / 2
   LearningRate=0.3
   Momentum=0.2
   TrainingTime=500
   Training 100%
4. **PCA+SVM** (ข้อมูลผ่าน PCA 19 มิติ และรู้จำด้วย Support Vector Machine)
   The polynomial kernel

ตารางที่ 4-1 นำเสนอรายละเอียดของข้อมูลที่ใช้ในการทดลองนี้ โดยข้อมูลที่ใช้นี้ ชื่อ KDDcup99 ทำการสุ่มเลือกมาทั้งหมด 13,499 ชุดทดสอบ โดยข้อมูลมี 37 มิติ โดยรายละเอียดของข้อมูลแต่ละคลาสเป็นดังนี้

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

ตารางที่ 4-1 รายละเอียดข้อมูลที่ใช้ในการทดลอง

ประเภทข้อมูล	จำนวนข้อมูล/มิติ (แอทริบิวต์)	จำนวนข้อมูล (Patterns) ในแต่ละคลาส
KDDcup99	13499/37	4107/4107/4107/1126/52

ตารางที่ 4-2 นำเสนอค่าสถิติที่ได้จากการทำการทดลองประกอบด้วย ค่าร้อยละของความถูกต้อง (Accuracy), ค่า F และ ค่า AUC พบว่า ในการทดลองเบื้องต้นในรายงานการวิจัยนี้ ชุดข้อมูลทั้งหมด 37 มิติ ให้ผลการทดลองที่สูงกว่าผลการทดลองที่ได้จากการลดมิติข้อมูลด้วยเทคนิค PCA

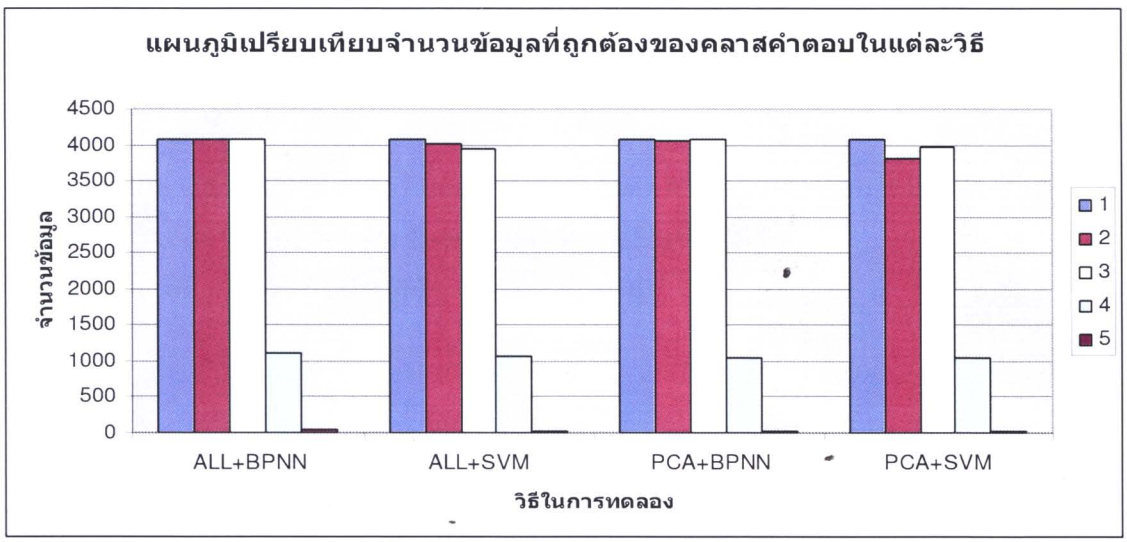
ตารางที่ 4-2 ค่า Accuracy จากการประมวลผล

Learning Method	ข้อมูลทั้งหมด			ข้อมูลที่ผ่านขั้นตอน PCA		
	Accuracy	F	AUC	Accuracy	F	AUC
BPNN	99.17	0.992	0.999	98.45	0.984	0.996
SVM	97.15	0.971	0.988	95.63	0.956	0.983

ตารางที่ 4-3 และ รูปที่ 4-2 ได้แสดงจำนวนชุดข้อมูลที่วิธีการเรียนรู้แต่ละวิธีทำการแบ่งประเภทได้อย่างถูกต้อง ซึ่งพบว่าในการทดลองเบื้องต้นนี้ โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับให้ผลการทดลองที่ดีกว่า Support Vector Machine

ตารางที่ 4-3 จำนวนข้อมูลที่แบ่งประเภทได้ถูกต้องของคลาสคำตอบในแต่ละวิธี

Learning Method	คลาส				
	1 (4107)	2 (4107)	3 (4107)	4 (1126)	5 (52)
ALL+BPNN	4087	4073	4082	1112	34
ALL+SVM	4075	4002	3953	1057	28
PCA+BPNN	4072	4060	4080	1050	28
PCA+SVM	4075	3820	3957	1032	26



รูปที่ 4-2 แผนภูมิเปรียบเทียบจำนวนข้อมูลที่ถูกต้องของคลาสคำตอบในแต่ละวิธี