การตรวจจับการบุกรุกโดยการรวมกันของเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและเทคนิคการตรวจจับความผิดปกติ

INTRUSION DETECTION VIA COMBINING MISUSE AND ANOMALY DETECTION TECHNIQUES

โดย

สุระสิทธิ์ ทรงม้า

ดุษฎีนิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตาม

หลักสูตรปรัชญาดุษฎีบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

บัณฑิตวิทยาลัย มหาวิทยาลัยรังสิต

ปีการศึกษา 2557

INTRUSION DETECTION VIA COMBINING MISUSE AND ANOMALY DETECTION TECHNIQUES

BY

SURASIT SONGMA

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR

THE DEGREE OF DOCTOR OF PHILOSOPHY IN INFORMATION TECHNOLOGY

FACULTY OF INFORMATION TECHNOLOGY

GRADUATE SCHOOL, RANGSIT UNIVERSITY

ACADEMIC YEAR 2014

ดุษฎีนิพนธ์เรื่อง

การตรวจจับการบุกรุกโดยการรวมกันของเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและเทคนิคการตรวจจับความผิดปกติ

โดย

สุระสิทธิ์ ทรงม้า

ได้รับการพิจารณาให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

มหาวิทยาลัยรังสิต

ปีการศึกษา 2557

.....	.....
รศ.ดร.บุญวัฒน์ อัตชู	ดร.ศิริพร ฉิมพลี
ประธานกรรมการสอบ	กรรมการ
.....	.....
รศ.ดร.ชม กิ้มปาน	ผศ.ดร.สิริพร ศุภราทิตย์
กรรมการ	กรรมการ
.....	.....
ดร.เกียรติศักดิ์ ใหม่เจริญนุกูล	ดร.วิชชา ฉิมพลี
กรรมการและอาจารย์ที่ปรึกษา	กรรมการและอาจารย์ที่ปรึกษาร่วม

.....

ผศ.ดร.ปริญญา สงวนสัตย์

กรรมการและอาจารย์ที่ปรึกษาร่วม

บัณฑิตวิทยาลัยรับรองแล้ว

.....

(ผศ.ร.ต.หญิง ดร.วรรณี ศุขสาตร)

คณบดีบัณฑิตวิทยาลัย

15 พฤษภาคม 2558

Dissertation entitled

INTRUSION DETECTION VIA COMBINING MISUSE AND ANOMALY DETECTION TECHNIQUES

by

SURASIT SONGMA

was submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy in Information Technology

Rangsit University

Academic Year 2014

.....	.....
Assoc.Prof.Boonwat Attachoo, D.Eng.	Siriporn Chimphlee, Ph.D.
Examination Committee Chairperson	Member
.....	.....
Assoc.Prof.Chom Kimpan, D.Eng.	Asst.Prof.Siriporn Supratid, D.Tech.Sci.
Member	Member
.....	.....
Kiattisak Maichalernnukul, D.Eng.	Witcha Chimphlee, Ph.D.
Member and Advisor	Member and Co-Advisor
	.....
	Asst.Prof.Parinya Sanguansat, D.Eng.
	Member and Co-Advisor

Approved by Graduate School

.....

(Asst.Prof.Plt.Off. Vannee Sooksatra, D.Eng.)

Deam of Graduate School

May 15, 2015

ดุษฎีนิพนธ์เรื่อง

การตรวจจับการบุกรุกโดยการรวมกันของเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและเทคนิคการตรวจจับความผิดปกติ

โดย

สุระสิทธิ์ ทรงม้า

ได้รับการพิจารณาให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

มหาวิทยาลัยรังสิต

ปีการศึกษา 2557

.....

รศ.ดร.บุญวัฒน์ อัตชู

ประธานกรรมการสอบ

.....

ดร.ศิริพร ฉิมพลี

กรรมการ

.....

รศ.ดร.ชม กิ้มปาน

กรรมการ

.....

ผศ.ดร.สิริพร ศุภราทิตย์

กรรมการ

.....

ดร.เกียรติศักดิ์ ใหม่เจริญนุกูล

กรรมการและอาจารย์ที่ปรึกษา

.....

ดร.วิชชา ฉิมพลี

กรรมการและอาจารย์ที่ปรึกษาร่วม

.....

ผศ.ดร.ปริญญา สงวนสัตย์

กรรมการและอาจารย์ที่ปรึกษาร่วม

บัณฑิตวิทยาลัยรับรองแล้ว

.....

(ผศ.ร.ต.หญิง ดร.วรรณี ศุขสาตร)

คณบดีบัณฑิตวิทยาลัย

15 พฤษภาคม 2558

Dissertation entitled

INTRUSION DETECTION VIA COMBINING MISUSE AND ANOMALY DETECTION TECHNIQUES

by

SURASIT SONGMA

was submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy in Information Technology

Rangsit University

Academic Year 2014

.....	.....
Assoc.Prof.Boonwat Attachoo, D.Eng.	Siriporn Chimphlee, Ph.D.
Examination Committee Chairperson	Member
.....	.....
Assoc.Prof.Chom Kimpan, D.Eng.	Asst.Prof.Siriporn Supratid, D.Tech.Sci.
Member	Member
.....	.....
Kiattisak Maichalernnukul, D.Eng.	Witcha Chimphlee, Ph.D.
Member and Advisor	Member and Co-Advisor

.....

Asst.Prof.Parinya Sanguansat, D.Eng.

Member and Co-Advisor

Approved by Graduate School

.....

(Asst.Prof.Plt.Off. Vannee Sooksatra, D.Eng.)

Deam of Graduate School

15 May 2015

กิจกรรมประกาศ

ดุษฎีนิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดีจากความช่วยเหลือและกำลังใจของบุคคลหลายฝ่าย ผู้วิจัยขอขอบพระคุณไว้ ณ ที่นี้

ขอขอบพระคุณ ดร.เกียรติศักดิ์ ใหม่เจริญนุกูล ดร.วิชชา ฉิมพลี และผศ.ดร.ปริญญา สงวนสัตย์ อาจารย์ที่ปรึกษาดุษฎีนิพนธ์ที่กรุณาให้คำปรึกษาและตรวจสอบต้นฉบับดุษฎีนิพนธ์ อันเป็นผลให้การวิจัยนี้มีความสมบูรณ์ ยิ่งไปกว่านั้น ความเอาใจใส่ดูแลตลอดเวลาในการดำเนินงานเป็นสิ่งที่ทำให้ผู้ทำดุษฎีนิพนธ์มีความซาบซึ้งใจเป็นอย่างยิ่ง ขอขอบพระคุณ รศ.ดร.บุญวัฒน์ อัตชู ประธานกรรมการสอบดุษฎีนิพนธ์ที่ให้แนวคิดตลอดจนชี้แนะแนวทางที่เป็นประโยชน์ต่อการวิจัย ขอขอบพระคุณ รศ.ดร.ชม กิ้มปาน ผศ.ดร.สิริพร ศุภราทิตย์ และดร.ศิริพร ฉิมพลี คณะกรรมการสอบดุษฎีนิพนธ์ที่กรุณาให้ข้อแนะนำที่มีคุณค่า และขอขอบคุณเจ้าหน้าที่ คณะเทคโนโลยีสารสนเทศและบัณฑิตวิทยาลัย มหาวิทยาลัยรังสิต ที่ให้ความช่วยเหลือและอำนวยความสะดวกตลอดมา

ขอกราบขอบพระคุณบิดามารดา ครอบครัว และมหาวิทยาลัยราชภัฏสวนดุสิตที่สนับสนุนให้กำลังใจในการศึกษาเสมอมา คุณประโยชน์ใดอันพึงเกิดจากดุษฎีนิพนธ์ฉบับนี้ขอมอบแด่ผู้มีพระคุณทุกท่าน

ท้ายที่สุดนี้ ผู้วิจัยสำนึกในคุณค่าของตำรา เอกสารอ้างอิง และความรู้จากคณาจารย์ทุกท่านที่ได้นำมากล่าวอ้างไว้ด้วยความเคารพอย่างสูง ความอนุเคราะห์และเกื้อกูลจากคณาจารย์ทุกท่านที่ได้กล่าวไว้และไม่ได้กล่าวไว้ในที่นี้ จะถูกระลึกในจิตใจของผู้ทำวิจัยตลอดไป

สุระสิทธิ์ ทรงม้า

ผู้วิจัย

5407016 : **สาขาวิชาเอก : เทคโนโลยีสารสนเทศ; ปร.ด. (เทคโนโลยีสารสนเทศ)**

คำสำคัญ : **ระบบตรวจจับการบุกรุก, การตรวจจับการใช้สิทธิ์โดยมิชอบ, การตรวจจับความผิดปกติ และข้อมูล KDD Cup 1999**

สุระสิทธิ์ ทรงม้า: การตรวจจับการบุกรุกโดยการรวมกันของเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและเทคนิคการตรวจจับความผิดปกติ (INTRUSION DETECTION VIA COMBINING MISUSE AND ANOMALY DETECTION TECHNIQUES) อาจารย์ที่ปรึกษา: ดร.เกียรติศักดิ์ ใหม่เจริญนุกูล, อาจารย์ที่ปรึกษาร่วม: ดร.วิชชา ฉิมพลี และผศ.ดร.ปริญญา สงวนสัตย์, 127 หน้า.

ดุษฎีนิพนธ์นี้มีจุดประสงค์เพื่อศึกษาและเปรียบเทียบประสิทธิภาพของระบบตรวจจับการบุกรุก และปรับปรุงประสิทธิภาพของระบบดังกล่าวโดยการทำงานร่วมกันระหว่างเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและเทคนิคการตรวจจับความผิดปกติ ชุดข้อมูลที่ใช้ในการทดลอง คือ ชุดข้อมูล KDD Cup 1999 ผลลัพธ์ของระบบตรวจจับการบุกรุกสามารถจำแนกกลุ่มได้เป็น 6 กลุ่ม ได้แก่ กลุ่มพฤติกรรมปกติ กลุ่มการโจมแบบปฏิเสธการให้บริการ กลุ่มการโจมตีแบบเข้าถึงระบบโดยไม่ได้รับอนุญาต กลุ่มการโจมตีแบบพยายามทำตัวเป็นผู้มีสิทธิ์ กลุ่มการโจมตีแบบสแกนเป้าหมาย และกลุ่มไม่รู้จัก ประสิทธิภาพของระบบตรวจจับการบุกรุกประเมินจากอัตราการตรวจจับการบุกรุก อัตราการแจ้งเตือนผิดพลาด และอัตราความแม่นยำโดยรวม

จากการเฉลี่ยผลตรวจจับการบุกรุกจำนวน 10 รอบ พบว่าการตรวจจับการบุกรุกโดยการรวมกันของเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและเทคนิคการตรวจจับความผิดปกติมีประสิทธิภาพสูงกว่าเทคนิคอื่น โดยมีอัตราการตรวจจับการบุกรุกเท่ากับ 87.32% มีอัตราการแจ้งเตือนผิดพลาดเท่ากับ 0.67% และมีอัตราความแม่นยำโดยรวมเท่ากับ 94.80% และเมื่อเทียบกับผลการตรวจจับการบุกรุกด้วยเทคนิคการใช้สิทธิ์โดยมิชอบเพียงอย่างเดียว พบว่าการตรวจจับที่นำเสนอสามารถตรวจจับได้ถูกต้องเพิ่มจำนวน 610 ระเบียน คิดเป็น 0.79%

ลายมือชื่อนักศึกษา .....

ลายมือชื่ออาจารย์ที่ปรึกษา .....

ลายมือชื่ออาจารย์ที่ปรึกษาร่วม.....

ลายมือชื่ออาจารย์ที่ปรึกษาร่วม.....

5407016 : **MAJOR : INFORMATION TECHNOLOGY; Ph.D. (INFORMATION TECHNOLOGY)**

KEYWORD : **INTRUSION DECTION, MISUSE DETECTION, ANOMALY DETECTION AND KDD CUP 1999 DATA SET**

SURASIT SONGMA: INTRUSION DETECTION VIA COMBINING MISUSE AND ANOMALY DETECTION TECHNIQUES: ADVISOR: KIATTISAK MAICHALERNNUKUL, D.ENG., DISSERTATION CO-ADVISOR: WITCHA CHIMPHLEE, Ph.D. AND ASST. PROF. PARINYA SANGUANSAT,D.ENG., 127 p.

This dissertation aims at studying and comparing the performance of intrusion detection systems, as well as improving such performance via combining misuse and anomaly detection techniques. We use the so-called KDD Cup 1999 data set in our experiment, and the results can be classified into six groups: normal, denial of service, remote to local, user to root, probing, and unknown. We evaluate the performance of intrusion detection systems in terms of the detection rate, false positive rate, and overall accuracy rate.

By averaging 10 results of intrusion detection, we find that the intrusion detection via combining misuse and anomaly detection techniques outperforms that of other relevant techniques. Specifically, its detection rate, false positive rate, and overall accuracy rate are 87.32%, 0.67%, and 94.80%, respectively. Also, such intrusion detection can correctly detect data more 610 records (which is 0.79%) when compared to the misuse detection.

Student's Signature .....

Dissertation Advisor's Signature .....

Dissertation Co-Advisor's Signature .....

Dissertation Co-Advisor's Signature .....

สารบัญ

สารบัญ (ต่อ)

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญตาราง (ต่อ)

สารบัญรูป

สัญลักษณ์และคำย่อ

สัญลักษณ์	ความหมาย
DoS	Denial of Service
DTR	Detection Rate
FCM	Fuzzy *c*-Means
FPR	False Positive Rate
GA	Genetic Algorithm
HIDS	Host-Based Intrusion Detection Systems
ICMP	Internet Control Message Protocol
IDS	Intrusion Detection Systems
KDD	Knowledge Discovery in Database
MDL	Minimum Description Length
NIDS	Network-Based Intrusion Detection Systems
OA	Overall Accuracy Rate
PCA	Principal Component Analysis
R2L	Remote to Local
RBF	Radial Basis Function
ROC	Receiver Operating Characteristics
SVM	Support Vector Machine
TCP	Transmission Control Protocol
U2R	User to Root
UDP	User Datagram Protocol

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ระบบคอมพิวเตอร์และอินเทอร์เน็ตในปัจจุบันเจริญเติบโตรวดเร็วมากและถือได้ว่าเป็นส่วนสำคัญของการศึกษา การประกอบกิจการและถูกนำมาใช้ในชีวิตประจำวันจนกลายเป็นสิ่งจำเป็นมากขึ้น เช่น การเก็บข้อมูล การติดต่อสื่อสาร การค้นคว้าข้อมูล การซื้อขายสินค้า รวมถึงความบันเทิง เป็นต้น ซึ่งสามารถกระทำได้สะดวกและรวดเร็ว แต่ปัญหาสำคัญที่ตามมาคือเรื่องความปลอดภัยของข้อมูลต่างๆ ของผู้ใช้งานจากผู้บุกรุก เช่น การเข้าถึง แก้ไข ทำลายข้อมูลโดยบุคคลอื่นในระบบคอมพิวเตอร์ โดยการกระทำดังกล่าวมีปริมาณเพิ่มจำนวนมากขึ้นและใกล้ตัวมากขึ้น ทั้งนี้มิใช่เป็นเพียงด้านธุรกิจเท่านั้นแต่ยังเป็นการกระทำผิดที่มีผลกระทบในด้านสังคม และความมั่นคงของรัฐ รวมทั้งความสงบสุขและศีลธรรมอันดีของประชาชนอีกด้วย

การใช้คอมพิวเตอร์และอินเทอร์เน็ตในชีวิตประจำวันของบุคคลทั่วไป หรือภายในองค์กรส่วนใหญ่ต้องมีการเก็บข้อมูลที่จำเป็นและมีความสำคัญแทบทั้งสิ้น ดังนั้นการป้องกันและรักษาความปลอดภัยข้อมูลจึงเป็นสิ่งที่ต้องให้ความสำคัญอย่างมาก ซึ่งในปัจจุบันมีเครื่องมือหลากหลายชนิดที่ช่วยในการป้องกันและรักษาความปลอดภัย เช่น โปรแกรมไฟล์วอลล์ เทคโนโลยีการเข้ารหัสข้อมูล ระบบตรวจจับการบุกรุก (Intrusion Detection System) เป็นต้น อนึ่งเครื่องมือที่นิยมใช้งานกันในปัจจุบันคือระบบตรวจจับการบุกรุก ดังจะเห็นได้จากองค์กรต่างๆทั้งภาครัฐและเอกชน ได้ติดตั้งระบบตรวจจับการบุกรุก เพื่อเพิ่มเติมความแข็งแกร่งให้กับระบบคอมพิวเตอร์ รวมถึงการรักษาความปลอดภัยของข้อมูลต่างๆในองค์กร เนื่องจากการติดตั้งไฟล์วอล์อย่างเพียงอย่างเดียวนั้นไม่เพียงพอที่จะรักษาความปลอดภัยได้ทั้งหมด และยังไม่สามารถเข้าใจลักษณะการ บุกรุกในแบบต่างๆ ที่ผู้บุกรุกโจมตีได้ จึงได้มีการติดตั้งระบบตรวจจับการบุกรุกเข้ามาช่วยสำหรับการวิเคราะห์และตรวจสอบข้อมูลทราฟฟิกที่ผิดปกติ หรือต้องสงสัยว่าเป็นผู้บุกรุก โดยระบบตรวจจับการบุกรุกก็จะทำหน้าที่แจ้งเตือนให้ผู้ดูแลระบบทราบ และเข้ามาดำเนินการแก้ไขได้ทันที เปรียบเสมือนบ้านพักอาศัยที่มีการติดตั้งสัญญาณกันขโมยไว้ หากมีขโมยหรือบุคคลต้องสงสัยผ่านเข้าไปในส่วน

ที่มีสัญญาณติดตั้งไว้ ก็จะส่งเสียงดังเตือนภัยให้เจ้าของบ้านรับทราบ โดยในที่นี้ตัวสัญญาณกันขโมยก็เปรียบเสมือนระบบตรวจจับการการบุกรุกในระบบคอมพิวเตอร์นั่นเอง

เทคนิคที่ใช้ในการตรวจจับการบุกรุกในระบบคอมพิวเตอร์โดยทั่วไปมี 2 เทคนิค เทคนิคแรกคือเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบ (Misuse Detection) ซึ่งใช้หลักการวิเคราะห์กิจกรรมของระบบ โดยการพิจารณาเหตุการณ์หรือชุดของเหตุการณ์ที่ตรงกับรูปแบบเหตุการณ์ที่ได้กำหนดไว้แล้วว่าเป็นการบุกรุก รูปแบบของเหตุการณ์การบุกรุกที่รู้จักเหล่านี้จะเรียกว่า ร่องรอยการบุกรุก (Signature) เทคนิคที่สองคือเทคนิคการตรวจจับความผิดปกติ (Anomaly Detection) ซึ่งใช้หลักการตรวจจับจากเหตุการณ์ต่างๆ ที่แตกต่างไปจากเหตุการณ์ปกติโดยทั่วไป (Known-Good) เทคนิคนี้จำเป็นต้องมีการเรียนรู้จากชุดข้อมูลฝึกอบรม (Train Data) เหตุการณ์ปกติว่ามีลักษณะอย่างไร และเมื่อตรวจพบว่ามีเหตุการณ์ที่มีความแตกต่างออกไป ระบบตรวจจับการบุกรุกจะแจ้งว่าเป็นเหตุการณ์ที่ผิดปกติและมีความเสี่ยงในการบุกรุก

ระบบตรวจจับการบุกรุกที่ดีและมีประสิทธิภาพนั้นต้องมีอัตราการตรวจจับการบุกรุก (Detection Rate: DTR) สูง มีอัตราการแจ้งเตือนผิดพลาด(False Positive Rate: FPR) ต่ำ และมีอัตราความแม่นยำโดยรวม (Overall Accuracy Rate: OA) สูง จากงานวิจัยที่ผ่านมาพบว่านักวิจัยมุ่งศึกษาและพัฒนาวิธีการตรวจจับการบุกรุกที่เจาะจงเฉพาะแต่ละเทคนิคเท่านั้น นั่นคือไม่ได้นำเอาเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและเทคนิคการตรวจจับความผิดปกติมาใช้งานร่วมกันเพื่อปรับปรุงประสิทธิภาพของระบบตรวจจับการบุกรุก ซึ่งจากการศึกษาของผู้วิจัยพบว่าเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบ มีจุดเด่นในด้านการมีอัตราการตรวจจับการบุกรุกที่สูง และมีอัตราการแจ้งเตือนผิดพลาดที่ต่ำ แต่มีข้อจำกัดว่าสามารถตรวจจับการบุกรุกได้ดีเฉพาะชนิดการบุกรุกที่รู้จักเท่านั้น จึงจำเป็นต้องมีการเก็บระเบียนรูปแบบการบุกรุกไว้ในฐานข้อมูลเป็นจำนวนมาก ซึ่งเป็นเรื่องที่ทำได้ยากและส่งผลให้วิธีการนี้ไม่สามารถตรวจจับการบุกรุกชนิดใหม่ได้ ส่วนเทคนิคการตรวจจับความผิดปกติ มีจุดเด่นที่สามารถตรวจจับการบุกรุกชนิดใหม่ได้โดยไม่จำเป็นต้องมีความรู้เกี่ยวกับชนิดการบุกรุกนั้น แต่มีข้อจำกัดเรื่องอัตราการแจ้งเตือนผิดพลาดที่สูง เนื่องจากพฤติกรรมของผู้ใช้ระบบไม่สามารถคาดการณ์ได้ ซึ่งบางครั้งพบว่ามีความเป็นไปได้ที่จะเป็นการกระทำที่ผิดปกติแต่ไม่ใช่การบุกรุก

จากที่กล่าวมาจะเห็นว่าเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและเทคนิคการตรวจจับความผิดปกติ มีจุดเด่นที่แตกต่างกัน ผู้วิจัยจึงมีแนวคิดที่จะปรับปรุงประสิทธิภาพของระบบตรวจจับ

การบุกรุกให้สูงขึ้นโดยการนำเอาทั้งสองเทคนิคมาทำงานร่วมกัน เพื่อใช้ประโยชน์จากทั้งจุดเด่นด้านการมีอัตราการตรวจจับการบุกรุกที่สูงและการมีอัตราการแจ้งเตือนผิดพลาดที่ต่ำของเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและจุดเด่นของเทคนิคการตรวจจับความผิดปกติที่สามารถตรวจจับการบุกรุกชนิดใหม่ได้

1.2 วัตถุประสงค์ของการวิจัย

1.2.1 เพื่อศึกษาและเปรียบเทียบประสิทธิภาพของระบบตรวจจับการบุกรุก

1.2.2 เพื่อปรับปรุงประสิทธิภาพของระบบตรวจจับการบุกรุกโดยการทำงานร่วมกันระหว่างเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและเทคนิคการตรวจจับความผิดปกติ

1.3 ขอบเขตของการวิจัย

1.3.1 ใช้ชุดข้อมูล KDD Cup 1999 ซึ่งเป็นชุดข้อมูลมาตรฐานสำหรับการทดสอบประสิทธิภาพของระบบตรวจจับการบุกรุก

1.3.2 ผลลัพธ์ของระบบตรวจจับการบุกรุกที่นำเสนอจัดกลุ่มได้เป็น 3 กลุ่มใหญ่ คือ

1.3.2.1 กลุ่มปกติ (Normal)

1.3.2.2 กลุ่มผู้บุกรุก (Attacks) โดยสามารถจำแนกเป็นกลุ่มย่อยๆ ดังนี้

1) กลุ่มการโจมตีแบบปฏิเสธการให้บริการ (Denial of Service: DoS)

2) กลุ่มการโจมตีแบบเข้าถึงระบบโดยไม่ได้รับอนุญาต (Remote to Local: R2L)

3) กลุ่มการโจมตีแบบพยายามทำตัวเป็นผู้มีสิทธิ์ (User to Root: U2R)

4) กลุ่มการโจมตีแบบสแกนเป้าหมาย (Probing)

1.3.2.1 กลุ่มไม่รู้จัก (Unknown)

1.3.3 วัดประสิทธิภาพของระบบตรวจจับการบุกรุกจากความสามารถในการตรวจจับการบุกรุก (DTR) อัตราการแจ้งเตือนผิด (FPR) และอัตราความแม่นยำโดยรวม (OA)

1.4 นิยามศัพท์

การบุกรุก (Intrusion) หมายถึง ความพยายามหรือการกระทำต่างๆที่ส่งผลเสียต่อความลับ (Confidentiality) สภาพพร้อมใช้งาน (Availability) และบูรณภาพ (Integrity) ของทรัพยากรระบบคอมพิวเตอร์ หรือวิธีการกระทำเพื่อข้ามผ่านมาตรการในการควบคุมความปลอดภัยของระบบคอมพิวเตอร์ ซึ่งอาจเกิดจากธรรมชาติหรือตัวบุคคลผู้เกี่ยวข้องไม่ว่าจะเกิดด้วยความตั้งใจหรือไม่ก็ตาม

ชุดข้อมูล KDD Cup 1999 หมายถึง ชุดข้อมูลที่ใช้ในการวิจัย โดยข้อมูลสร้างขึ้นจากห้องทดลองของสถาบันเมสซาชูเซส และเก็บรวบรวมโดยใช้โปรแกรม TCP Dump เป็นเวลา 9 สัปดาห์ การติดต่อสื่อสารข้อมูลดังกล่าวจะอยู่ในลักษณะที่ซีพีแพ็คเก็ตจากต้นทางไปยังปลายทาง ซึ่งมีการกำหนดโปรโตคอลที่ใช้ในการสื่อสารหมายเลขไอพีแอดเดรสต้นทาง และหมายเลขไอพีแอดเดรสปลายทางไว้อย่างชัดเจน แต่ละระเบียนข้อมูลจะมีป้ายกำกับไว้ว่าเป็นระเบียนที่ปกติหรือระเบียนที่เป็นการบุกรุก

กลุ่มไม่รู้จัก (Unknown) หมายถึง กลุ่มพฤติกรรมที่ไม่รู้จักหรือรูปแบบเหตุการณ์ใหม่ที่ไม่แน่ชัดว่าว่าเป็นพฤติกรรมปกติหรือบุกรุก

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1.5.1 ได้เรียนรู้เกี่ยวกับประสิทธิภาพของการนำเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและเทคนิคการตรวจจับความผิดปกติที่นำมาทำงานร่วมกัน เพื่อใช้ในระบบตรวจจับการบุกรุก

1.5.2 ได้ระบบตรวจจับการบุกรุกแบบรวมระหว่างเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและเทคนิคการตรวจจับความผิดปกติ ซึ่งจะเป็นประโยชน์ในการพัฒนาระบบตรวจจับการบุกรุก

1.5.3 ได้กรอบงานของการทำงานร่วมกันระหว่างเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและเทคนิคการตรวจจับความผิดปกติใหม่

1.5.4 เป็นแนวทางสำหรับนักวิจัยที่สนใจในการพัฒนาระบบตรวจจับการบุกรุก

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ความมั่นคงของระบบคอมพิวเตอร์

ปัจจุบันมีการเก็บบันทึกข้อมูลต่างๆ ไว้ภายในระบบคอมพิวเตอร์เป็นจำนวนมาก เนื่องจากมีความสะดวกในการจัดเก็บและสามารถเรียกใช้ได้ง่าย ทำให้ระบบคอมพิวเตอร์จำเป็นต้องมีความมั่นคง เพื่อให้ข้อมูลมีความปลอดภัยนั่นเอง แต่จะทราบได้อย่างไรว่าข้อมูลที่เก็บอยู่ในระบบคอมพิวเตอร์ของเรานั้นมีความถูกต้องและปลอดภัยอยู่เสมอ สิ่งที่ช่วยรับประกันได้ว่าระบบคอมพิวเตอร์มีความมั่นคง จะพิจารณาจากองค์ประกอบที่สำคัญอยู่ 3 ส่วนซึ่งสอดคล้องกับแนวคิดของ จตุชัย แพงจันทร์ (2550) และ Bishop (2002) ที่ได้สรุปไว้ดังนี้

2.1.1 ความลับ

ในการทำให้เกิดความมั่นคงจะต้องมั่นใจว่าข้อมูลที่อยู่ในระบบคอมพิวเตอร์นั้นควรจะไม่มีบุคคลที่ไม่ได้รับอนุญาต (Unauthorized User) เข้าถึงได้และหากเข้าถึงได้ก็ไม่สามารถนำไปใช้งานได้ เนื่องจากมีวิธีการปกป้องความปลอดภัยของข้อมูลอย่างเพียงพอ ทำให้ไม่สามารถเข้าใจเนื้อหาหรือไม่สามารถเปิดดูเนื้อหาของข้อมูลนั้นได้ กลไกที่นิยมใช้งานกันในปัจจุบันเรียกว่าการเข้ารหัสลับ (Cryptography หรือ Encryption) เพื่อทำให้ข้อมูลไม่สามารถอ่านหรือเข้าใจความหมายได้ เปรียบเทียบได้กับการส่งจดหมายที่มีการลงทะเบียนและการปิดผนึกซองจดหมาย การใช้ซองจดหมายที่ทึบแสง การเขียนด้วยหมึกที่มองไม่เห็น เป็นต้น

2.1.2 บูรณภาพ

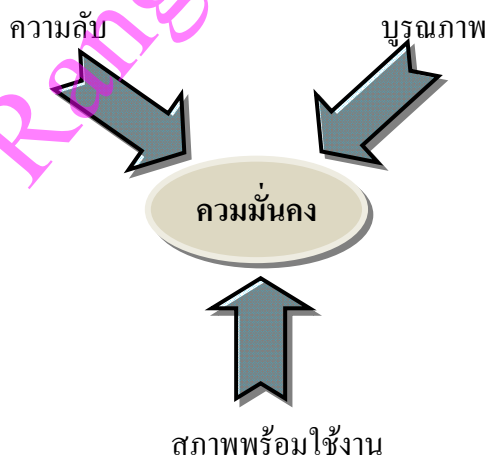
ต้องมั่นใจว่าข้อมูลที่อยู่ในระบบคอมพิวเตอร์ไม่ได้ถูกแก้ไข เปลี่ยนแปลง ด้วยวิธีการใดๆ โดยไม่ได้รับอนุญาต เรียกอีกอย่างว่าการทำให้ข้อมูลมีความน่าเชื่อถือ (ทั้งตัวข้อมูลและแหล่งที่มา) กลไกที่นิยมใช้เพื่อความมั่นคง คือ การป้องกัน (Prevention) และการตรวจจับ (Detection) โดยกลไกการป้องกันมีอยู่ 2 ลักษณะ คือ การป้องกันไม่ให้ผู้ที่ไม่ได้รับอนุญาตเข้าไป

เปลี่ยนแปลงแก้ไขข้อมูล (ป้องกันบุคคลภายนอกที่ไม่มีสิทธิ์) และการป้องกันไม่ให้ผู้ที่ได้รับอนุญาตเข้าไปเปลี่ยนแปลงแก้ไขข้อมูลในรูปแบบที่ไม่ถูกต้อง (ป้องกันบุคคลภายในที่มีสิทธ์) ส่วนกลไกการตรวจจับนั้น เป็นการวิเคราะห์ว่าข้อมูลยังคงสภาพเดิมอยู่หรือไม่ เปรียบเทียบได้กับการเขียนด้วยหมึกซึ่งถ้าถูกลบแล้วจะก่อให้เกิดรอยลบขึ้น การใช้โฮโลแกรม (Hologram) กำกับบนบัตรเครดิต เป็นต้น

2.1.3 สภาพพร้อมใช้งาน

ต้องมั่นใจว่าข้อมูลสามารถใช้งานได้ตามปกติและเต็มประสิทธิภาพ เมื่อผู้ที่ได้รับอนุญาตต้องการใช้งาน ซึ่งกลไกการป้องกันที่นิยมใช้กัน คือ การออกแบบระบบให้มีความเชื่อถือได้ (Reliability) เปรียบได้กับร้านสะดวกซื้อ เซเว่นอีเลฟเว่น ซึ่งเปิดให้บริการตลอด 24 ชั่วโมง คำว่า "สภาพพร้อมใช้งาน" ตรงกันข้ามกับคำว่า "การปฏิเสธการให้บริการ" ซึ่งหมายถึง ภาวะที่ผู้ที่ได้รับอนุญาตให้เข้าใช้งานไม่สามารถเข้าใช้งานได้

จะเห็นว่าองค์ประกอบความมั่นคงจะขาดส่วนใดส่วนหนึ่งไม่ได้ จำเป็นต้องมีครบทั้ง 3 องค์ประกอบ จึงจะถือได้ว่าระบบคอมพิวเตอร์นั้นมีความมั่นคง ดังรูปที่ 2.1 อย่างไรก็ตามไม่มีวิธีการรักษาความปลอดภัยใดที่สามารถป้องกันได้อย่างสมบูรณ์ ทั้งนี้เนื่องจากยังมีปัจจัยอื่นที่เกี่ยวข้องอีกมาก



รูปที่ 2.1 องค์ประกอบของความมั่นคงของระบบคอมพิวเตอร์

2.2 ภัยคุกคามความมั่นคง

ภัยคุกคามที่เกิดขึ้นในปัจจุบันนั้นมีรูปแบบที่หลากหลาย และยากต่อการป้องกัน เนื่องด้วยประสิทธิภาพของระบบคอมพิวเตอร์เอง ฮาร์ดแวร์และซอฟต์แวร์ที่มีประสิทธิภาพที่สูงขึ้น บุคคลที่เกี่ยวข้องกับคอมพิวเตอร์และสามารถใช้คอมพิวเตอร์ได้มีมากขึ้น

ภัยคุกคามความมั่นคงของระบบคอมพิวเตอร์ หมายถึง การโจมตี การบุกรุก หรือวิธีการที่ทำแล้วส่งผลทำให้องค์ประกอบความมั่นคงของระบบคอมพิวเตอร์ส่วนใดส่วนหนึ่งหรือมากกว่าหนึ่งส่วนขาดหายไป อาจเกิดจากธรรมชาติหรือตัวบุคคลที่เกี่ยวข้อง จะด้วยความตั้งใจหรือไม่ก็ตาม

พวงทิพย์ แท่นแสง (2550) สมเกียรติ รุ่งเรืองลดา (2551) และ Kendall (1998) ได้แบ่งกลุ่มของภัยคุกคามที่มีต่อระบบคอมพิวเตอร์ออกเป็นกลุ่มใหญ่ๆดังนี้

2.2.1 การโจมตีแบบปฏิเสธการให้บริการ (DoS)

เป็นวิธีการโจมตีเป้าหมายที่เน้นไปที่การทำให้คอมพิวเตอร์ไม่สามารถติดต่อสื่อสารหรือใช้งานได้ตามปกติ หรือทำให้ระบบคอมพิวเตอร์ใช้งานไม่ได้ โดยการเข้าถึงระบบและใช้ทรัพยากรที่มีอยู่อย่างจำกัดทำให้ไม่สามารถให้บริการได้ตามปกติ เช่น การโจมตีแบบ Neptune, Mailbomb หรือ Smurf Attack จะใช้วิธีการส่งแพ็คเก็ตจำนวนมาก (Flooding) เข้าไปยังระบบเป้าหมายทำให้ทราฟฟิก (Traffic) เพิ่มสูงขึ้นในเวลาอันรวดเร็วและการสื่อสารในระบบตามปกติช้าลงจนใช้การไม่ได้ หรือทำให้คอมพิวเตอร์เป้าหมายไม่สามารถให้บริการได้ การโจมตีแบบ Teardrop หรือ Ping of Death จะใช้วิธีการสร้างแพ็คเก็ตขนาดใหญ่แล้วทำการแตกกระจาย (Fragment) เป็นแพ็คเก็ตย่อยๆ โดยมีการปรับแต่งออฟเซต (Offset) ให้ผิดปกติไป ทำให้เมื่อส่งแพ็คเก็ตถึงเป้าหมายก็จะเกิดปัญหาในการรวมกันของแพ็คเก็ต

2.2.2 การโจมตีแบบเข้าถึงระบบโดยไม่ได้รับอนุญาต (R2L)

เป็นลักษณะการพยายามเข้าถึงระบบคอมพิวเตอร์เป้าหมายที่ไม่ได้รับอนุญาตให้เข้าถึง โดยใช้วิธีการส่งแพ็คเก็ตไปยังเครื่องเป้าหมายด้วยโปรแกรมเครือข่าย เช่น Imap, Named และ Sendmail หรือการพยายามเข้าใช้เครื่องเป้าหมายที่มีการตั้งค่าการรักษาความปลอดภัยไว้ไม่ดี

เช่น การโจมตีแบบ Guess_passwod, Ftp_write และ Xsnoop หรือการใช้ม้าโทรจัน (Trojan Horse) เพื่อดักรับรหัสผ่านของผู้อื่น

2.2.3 การโจมตีแบบพยายามทำตัวเป็นผู้มีสิทธิ์ (U2R)

เป็นลักษณะที่ผู้ใช้พยายามใช้งานระบบคอมพิวเตอร์ในสิ่งที่ไม่ได้รับอนุญาตให้เข้าถึงสิทธิ์นั้น กล่าวคือพยายามยกระดับสิทธิ์การเข้าถึงระบบให้เป็นรูท (Root) หรือซุปเปอร์ยูสเซอร์ (Super-User) เพื่อที่จะสามารถทำอะไรก็ได้ในระบบดังกล่าว วิธีการเปลี่ยนสิทธิ์ของผู้บุกรุกเพื่อเป็นรูทมีหลายวิธี เช่น การอาศัยข้อผิดพลาดของระบบปฏิบัติการ แต่วิธีที่นิยมมากที่สุดคือการทำให้โปรแกรมมีการคัดลอกข้อมูลเกินกว่าที่บัฟเฟอร์ (Buffer) จะสามารถรับได้จึงทำให้ข้อมูลล้นออกมา (Overflow) ซึ่งข้อมูลส่วนที่ล้นจะถูกบันทึกทับด้วยคำสั่งถัดไป ผู้บุกรุกจะใช้จุดบกพร่องนี้ในการเขียนคำสั่งต่างๆ ที่ตนเองต้องการใส่ลงไปแทนที่ หรือวิธีการโจมตีแบบ Loadmodule ซึ่งเป็นการคาดเดาสภาพแวดล้อมของระบบเป้าหมายเพื่อหาช่องโหว่หรือความผิดพลาดของโปรแกรมในระบบและใช้ประโยชน์จากสิ่งนั้น

2.2.4 การโจมตีแบบสแกนเป้าหมาย (Probing)

เป็นวิธีแรกที่นิยมใช้กันสำหรับการจะเข้าโจมตีเป้าหมาย เนื่องจากผู้บุกรุกจำเป็นต้องทราบเสียก่อนว่าระบบคอมพิวเตอร์เป้าหมายมีช่องโหว่บริเวณใดบ้าง โดยใช้เครื่องมือ (โปรแกรมคอมพิวเตอร์) ช่วยตรวจหาและเก็บข้อมูล เช่น Satan, Saint และ Mscan เมื่อผู้บุกรุกทราบถึงช่องโหว่แล้วจึงจะโจมตี วิธีการนี้ทำได้ง่าย รวดเร็ว และสามารถตรวจหาเครื่องคอมพิวเตอร์เป้าหมายได้ครั้งละเป็นร้อย ๆ เครื่อง

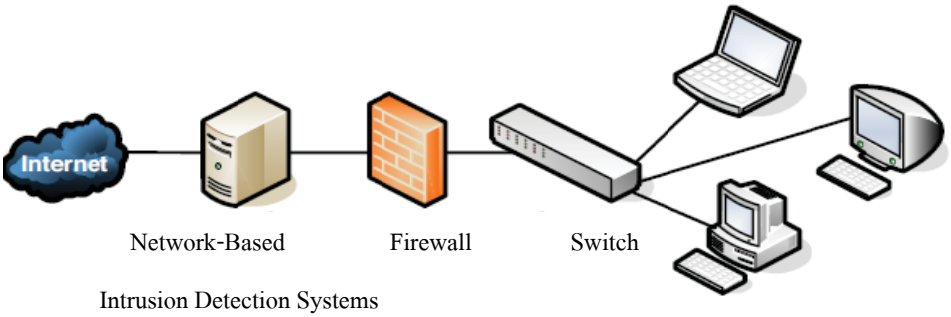
2.3 ระบบตรวจจับการบุกรุก

ระบบตรวจจับการบุกรุกเป็นเครื่องมือตรวจจับความพยายามที่จะบุกรุกระบบคอมพิวเตอร์โดยระบบแจ้งเตือนไปยังผู้ดูแลระบบเมื่อมีการบุกรุกหรือพยายามที่จะบุกรุก ระบบตรวจจับการบุกรุกนั้นไม่ใช่ระบบที่ใช้ป้องกันการบุกรุกระบบคอมพิวเตอร์ แต่เป็นระบบที่คอยแจ้งเตือนภัยคุกคามที่จะเกิดขึ้นกับระบบคอมพิวเตอร์เท่านั้น

Bace and Mell (2011) ได้ให้คำนิยามของระบบตรวจจับการบุกรุกว่าเป็นกระบวนการในการตรวจตราหรือตรวจสอบเหตุการณ์ต่างๆ ที่เกิดขึ้นภายในระบบคอมพิวเตอร์และวิเคราะห์ถึงเหตุการณ์นั้นว่าเป็นภัยคุกคามต่อความมั่นคงของระบบคอมพิวเตอร์หรือไม่ เช่น ทำให้ความลับบูรณภาพ หรือสภาพพร้อมใช้เสียหายไปไม่ว่าจะเป็นองค์ประกอบเดียวหรือหลายองค์ประกอบ

ระบบตรวจจับการบุกรุกถือได้ว่าเป็นเครื่องมือพื้นฐานที่สำคัญสำหรับการรักษาความปลอดภัยให้กับระบบคอมพิวเตอร์ โดยมีเป้าหมายว่าระบบตรวจจับการบุกรุกจะต้องสามารถแบ่งเหตุการณ์ที่เกิดขึ้นในระบบนั้นเป็นพฤติกรรมปกติ (Normal) หรือเป็นพฤติกรรมที่เป็นภัยคุกคาม (Attacks) ซึ่งสอดคล้องกับแนวคิดของ Denning (1987) โดยระบบตรวจจับการบุกรุกเป็นการรวมกันของซอฟต์แวร์และฮาร์ดแวร์ที่จะดำเนินการตรวจจับการบุกรุก และส่งสัญญาณเตือนทันทีเมื่อตรวจพบว่ามีการบุกรุก ระบบตรวจจับการบุกรุกยังช่วยเพิ่มระดับกลไกการป้องกันและรักษาความปลอดภัยขึ้นอีกระดับ เช่น ในการตรวจสอบไฟร์วอลล์หากเกิดข้อผิดพลาดขึ้นจะไม่สามารถตรวจสอบและแจ้งเตือนได้ว่าเป็นการบุกรุก ระบบตรวจจับการบุกรุกก็สามารถเป็นเครื่องมืออีกตัวที่จะช่วยตรวจสอบและแจ้งเตือนผู้ดูแลระบบได้ ดังรูปที่ 2.2

องค์ประกอบของระบบตรวจจับการบุกรุกแบ่งออกเป็น 2 ส่วน คือ ตัวแทน (Agent) ซึ่งคอยตรวจสอบเหตุการณ์หรือข้อมูลที่ส่งมายังเซิร์ฟเวอร์ที่ติดตั้งอยู่ และการวิเคราะห์ (Analysis) เหตุการณ์หรือข้อมูลเหล่านั้นว่าเป็นเหตุการณ์หรือข้อมูลที่ปกติ หรือเป็นการบุกรุก ซึ่งระบบตรวจจับการบุกรุกจะรายงานเฉพาะสิ่งที่กำหนดให้รายงานเท่านั้น สิ่งสำคัญที่ผู้ดูแลระบบจะต้องปรับแต่ง (Configuration) ได้แก่ รูปแบบ (Signature) ของการบุกรุก และเหตุการณ์ที่ผู้ดูแลระบบให้ความสำคัญหรือเหตุการณ์ที่คาดว่าจะนำไปสู่การบุกรุกในภายหน้า เหตุการณ์เหล่านี้อาจเป็นข้อมูลทราฟฟิกที่ไม่ปกติหรืออาจเป็นบางข้อความในล็อกไฟล์ (Log File) การปรับแต่งค่าให้กับระบบตรวจจับการบุกรุกของแต่ละองค์กรอาจจะไม่เหมือนกัน ขึ้นอยู่กับองค์กรนั้นให้ความสนใจกับการบุกรุกประเภทใด

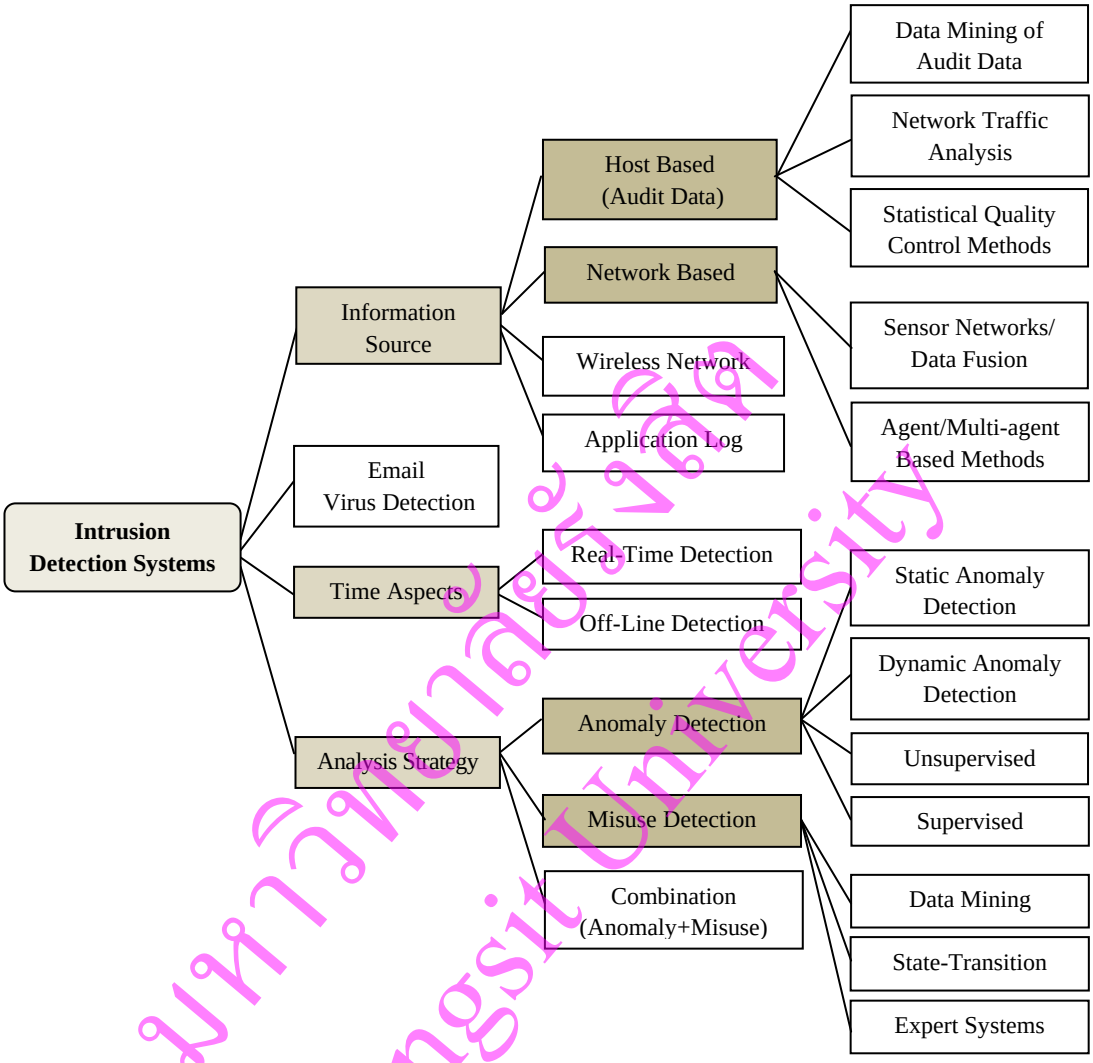


รูปที่ 2.2 การติดตั้งระบบตรวจจับการบุกรุก

ที่มา : (Soldier, http://thuansoldier.net, 3 มีนาคม 2555)

2.3.1 การจำแนกระบบตรวจจับการบุกรุก

Axelsson (2000, อ้างถึงใน Witcha Chimphlee, 2008) ได้อธิบายถึงภาพรวมของการจำแนกระบบตรวจจับการบุกรุกออกเป็น 4 ประเภท คือ แหล่งสารสนเทศ (Information Source) กลยุทธ์วิเคราะห์ (Analysis Strategy) ลักษณะของเวลา (Time Aspects) และการตรวจจับไวรัสทางอีเมล์ (Email Virus) รายละเอียดดังรูปที่ 2.3



รูปที่ 2.3 การจัดหมวดหมู่ของระบบตรวจจับการบุกรุก

ที่มา : Witcha Chimphlee, 2008

จากรูปนี้จะเห็นว่าหากแบ่งระบบตรวจจับการบุกรุกตามกลยุทธ์วิเคราะห์ จะแบ่งออกได้เป็นวิธีการตรวจจับความผิดปกติ วิธีการใช้สิทธิ์โดยมิชอบ และวิธีการตรวจจับแบบรวมกันระหว่างวิธีการตรวจจับความผิดปกติและวิธีการใช้สิทธิ์โดยมิชอบ หากแบ่งตามแหล่งสารสนเทศ จะแบ่งออกได้เป็นเอชไอดีเอส (Host-Based Intrusion Detection System) และเอ็นไอดีเอส (Network-Based Intrusion Detection Systems) ระบบตรวจจับการบุกรุกสามารถแบ่งตามลักษณะของเวลาได้เช่นกัน ได้แก่ ระบบตรวจจับการบุกแบบเรียลไทม์และระบบตรวจจับการบุกรุกแบบออฟไลน์หรือ

อาจแบ่งตามการตรวจจับไวรัสทางอีเมล์ ซึ่งเหมาะกับสำนักงานขนาดเล็ก ถ้านำไปใช้กับสำนักงานขนาดใหญ่จะทำให้ประสิทธิภาพในการตรวจจับการบุกรุกลดลง

2.4 เทคนิคการตรวจจับการบุกรุก

ในการตรวจจับการบุกรุกโดยทั่วไปมี 2 เทคนิค คือ การตรวจจับการใช้สิทธิ์โดยมิชอบ และการตรวจจับความผิดปกติ เทคนิคแรกเป็นการตรวจจับการบุกรุกโดยอาศัยคุณลักษณะหรือรูปแบบที่ได้กำหนดไว้ก่อนว่าเป็นการบุกรุก แล้วนำมาเปรียบเทียบกับเหตุการณ์ที่เกิดขึ้นในระบบคอมพิวเตอร์เพื่อหาการบุกรุกโดยตรง ดังนั้นแนวทางนี้จึงต้องอาศัยความรู้เกี่ยวกับการบุกรุกและพฤติกรรมที่ไม่เหมาะสม เพื่อที่จะสามารถกำหนดรูปแบบที่ใช้ในการค้นหาการบุกรุกได้อย่างถูกต้องและคลอบคลุมการบุกรุกทั้งหมดที่รู้จัก ส่วนเทคนิคที่สองเป็นการตรวจจับที่ตั้งอยู่บนพื้นฐานการตรวจหาเหตุการณ์หรือพฤติกรรมต่างๆ ที่เกิดขึ้นแตกต่างไปจากสภาวะการใช้งานปกติ เช่น ตามปกติผู้ใช้ ก. ใช้คอมพิวเตอร์ในที่ทำงานในช่วงเวลา 9.00 น. ถึง 17.00 น. เท่านั้น แต่ถ้ามีการใช้งานในช่วงกลางคืนก็จะถือว่าผิดปกติและอาจจะเป็นการบุกรุก แนวทางในการตรวจจับ คือ พยายามแยกพฤติกรรมปกติหรือยอมรับได้ออกมาและให้พฤติกรรมที่เหลือเป็นการบุกรุก ซึ่งสอดคล้องกับแนวคิดของ Carter (2001), Teng, et al. (2010) และ Wang and Zhou (2011)

2.4.1 เทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบ

เป็นเทคนิคตรวจจับการบุกรุกซึ่งจะวิเคราะห์กิจกรรมของระบบ โดยการพิจารณาเหตุการณ์ที่ตรงกับรูปแบบเหตุการณ์ที่ได้กำหนดไว้แล้วว่าเป็นการบุกรุก ดังนั้นบางครั้งจึงเรียกเทคนิคการตรวจจับนี้ว่า การตรวจจับโดยอาศัยรูปแบบ

ประเด็นสำคัญของเทคนิคการตรวจจับนี้ คือ จะเขียนรูปแบบการบุกรุกอย่างไรให้ครอบคลุมถึงความหลากหลายทั้งหมดที่เป็นไปได้ของการบุกรุก เทคนิคนี้ไม่สามารถรับประกันได้ว่าจะตรวจจับการบุกรุกได้อย่างสมบูรณ์ เนื่องจากจำเป็นจะต้องมีการปรับปรุงฐานข้อมูลหรือฐานความรู้ให้ทันสมัยอยู่เสมอ เพื่อตรวจจับการบุกรุกได้อย่างสมบูรณ์ รูปที่ 2.4 แสดงให้เห็นว่ารู้ "Bad" แนวทางที่ใช้ในการตรวจจับการใช้สิทธิ์โดยมิชอบคือ ระบบผู้เชี่ยวชาญ (Expert Systems) และการทวนสอบรูปแบบ (Signature Verification) ระบบตรวจจับการบุกรุกแบบนี้ เช่น โปรแกรมป้องกันไวรัสและไฟล์วอลล์

รูปที่ 2.4 การตรวจจับการใช้สิทธิ์โดยมิชอบ

2.4.2 เทคนิคการตรวจจับความผิดปกติ

เป็นเทคนิคตรวจจับการบุกรุกซึ่งตรวจจับเหตุการณ์หรือพฤติกรรมต่างๆที่แตกต่างไปจากเหตุการณ์ปกติโดยทั่วไป "Good" ดังรูปที่ 2.5 เทคนิคนี้มีความผันแปรไปตามลักษณะของเหตุการณ์ หรือพฤติกรรมปกติ จึงจำเป็นต้องมีการเรียนรู้จากชุดข้อมูลฝึกอบรม เหตุการณ์หรือพฤติกรรมปกติว่ามีลักษณะอย่างไร เมื่อตรวจพบว่ามีเหตุการณ์หรือพฤติกรรมที่เกิดขึ้นซึ่งมีความแตกต่างไปจากเหตุการณ์หรือพฤติกรรมปกตินั้น ระบบตรวจจับการบุกรุกจะแจ้งว่าเป็นพฤติกรรมที่ผิดปกติและมีความเสี่ยงในการบุกรุก

อนึ่งหากพิจารณาถึงการกระทำที่เป็นการบุกรุกและการกระทำที่ผิดปกติจะพบว่ามีความเป็นไปได้ที่จะเป็นการกระทำที่ผิดปกติแต่ไม่ใช่การบุกรุก ซึ่งระบบจะตรวจจับและแจ้งเตือนว่าเป็นการบุกรุก ทำให้เกิดปัญหาฟอลต์โพสซิทีฟ (False Positive) ส่วนการกระทำที่เป็นการบุกรุกจริงแต่ไม่ถูกตรวจจับว่าเป็นการกระทำที่ผิดปกติ ทำให้เกิดปัญหาฟอลต์เนกาทีฟ (False Negative) ซึ่งเป็นปัญหาที่สำคัญมาก ประเด็นที่สำคัญอีกประการหนึ่ง คือ การเลือกระดับหรือเกณฑ์ของการกระทำที่จะถือว่าเป็นการบุกรุกและการเลือกคุณลักษณะที่จะตรวจจับ นอกจากนี้ยังมีค่าใช้จ่ายสูงเนื่องจากต้องทำการเก็บข้อมูลสำหรับการตรวจสอบและอาจต้องคอยปรับปรุงเกณฑ์ในการชี้วัดพฤติกรรมปกติของระบบอยู่เสมอ เพื่อให้การตรวจจับมีความถูกต้องมากขึ้น แนวทางที่ใช้ในการตรวจจับความผิดปกติ คือ การตรวจจับด้วยขีดเริ่มเปลี่ยน (Threshold) วิธีการทางสถิติ (Statistical Measures) และวิธีการอาศัยกฎ (Rule-Based Measures)

รูปที่ 2.5 การตรวจจับความผิดปกติ

การตรวจจับการบุกรุกทั้ง 2 เทคนิคมีทั้งข้อดีและข้อด้อย จากการศึกษาพบว่าเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบ มีจุดเด่นในด้านการมีอัตราการตรวจจับการบุกรุกที่สูง และอัตราการแจ้งเตือนผิดพลาดที่ต่ำ แต่มีข้อจำกัดว่าสามารถตรวจจับการบุกรุกได้ดีเฉพาะชนิดการบุกรุกที่รู้จักเท่านั้น จึงทำให้จำเป็นต้องมีการเก็บข้อมูลรูปแบบการบุกรุกไว้ในฐานข้อมูลเป็นจำนวนมาก ซึ่งเป็นเรื่องที่ทำได้ยากจึงส่งผลให้วิธีการนี้ไม่สามารถตรวจจับการบุกรุกชนิดใหม่ได้ ส่วนเทคนิคการตรวจจับความผิดปกติ มีจุดเด่นที่สามารถตรวจจับการบุกรุกชนิดหรือรูปแบบใหม่ได้โดยไม่จำเป็นต้องมีความรู้เกี่ยวกับชนิดการบุกรุกนั้น แต่มีข้อจำกัดเรื่องอัตราการแจ้งเตือนผิดพลาดที่สูงเนื่องจากพฤติกรรมของผู้ใช้ระบบไม่สามารถคาดการณ์ได้ ซึ่งบางครั้งพบว่ามีความเป็นไปได้ที่จะเป็นการกระทำที่ผิดปกติแต่ไม่ใช่การบุกรุก รายละเอียดแสดงได้ดังตารางที่ 2.1

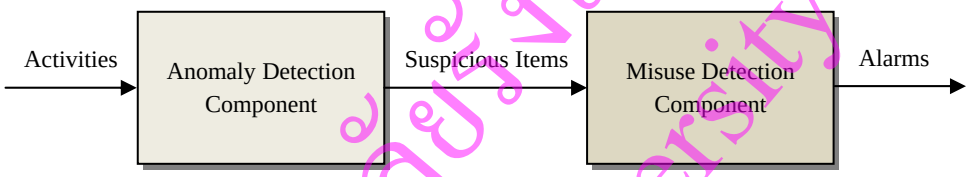
ตารางที่ 2.1 เปรียบเทียบข้อดีและข้อเสียของเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบ และเทคนิคการตรวจจับความผิดปกติ

	เทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบ	เทคนิคการตรวจจับความผิดปกติ
ข้อดี	1. มีประสิทธิภาพสูงในการตรวจจับการบุกรุกที่เป็นรู้จักดี และการแจ้งเตือนผิดพลาดต่ำ 2. เป็นวิธีการที่ทราบรายละเอียดของการบุกรุกแต่ละประเภท ทำให้ง่ายต่อการรักษาความปลอดภัยและการป้องกัน 3. เป็นวิธีการที่ช่วยในการบริหารจัดการระบบ ไม่ต้องคำนึงถึงผู้ใช้ว่าต้องเป็นผู้เชี่ยวชาญ	1. ไม่จำเป็นต้องมีความรู้หรือรายละเอียดเกี่ยวกับการบุกรุกดังกล่าวมากนัก และยังสามารถตรวจจับความพยายามที่จะใช้ประโยชน์จากช่องโหว่ใหม่และที่คาดไม่ถึง ซึ่งนำไปสู่การค้นพบโดยอัตโนมัติจากการโจมตีใหม่ๆ 2. ตรวจจับการบุกรุกหรือลักษณะผิดปกติที่ไม่มีอยู่ข้อมูลหรือฐานความรู้ได้
ข้อเสีย	1. มีความสามารถตรวจจับการบุกรุกที่มีอยู่ในฐานข้อมูลหรือฐานความรู้เท่านั้น 2. การปรับปรุงฐานข้อมูลหรือฐานความรู้ต้องระมัดระวังในการวิเคราะห์และจำเป็นต้องใช้ผู้ที่มีความรู้ความสามารถจึงใช้เวลานาน 3. ผลลัพธ์การตรวจจับการบุกรุกจะดีหรือไม่ขึ้นอยู่กับสภาพแวดล้อมของระบบ เนื่องจากฐานข้อมูลหรือฐานความรู้มีอยู่อย่างจำกัด หากมีการบุกรุกรูปแบบใหม่เกิดขึ้นในสภาพแวดล้อมดังกล่าว จะทำให้ไม่สามารถตรวจจับได้ 4. หากเป็นการบุกรุกที่เกิดขึ้นภายในจะทำให้ยากต่อการตรวจสอบ เนื่องจากไม่มีข้อมูลดังกล่าวอยู่ในฐานข้อมูลหรือฐานความรู้	1. มักจะมีการแจ้งเตือนผิดพลาด เนื่องจากพฤติกรรมที่ไม่แน่นอนของผู้ใช้ ทำให้ระบบตรวจจับการบุกรุกมองพฤติกรรมปกติว่าเป็นพฤติกรรมบุกรุกได้ ทั้งนี้ขึ้นอยู่กับการเรียนรู้พฤติกรรมปกติว่ามีมากเพียงพอหรือไม่และควรมีการเรียนรู้พฤติกรรมปกติใหม่เพิ่มเติมเพื่อลดการแจ้งเตือนที่ผิดพลาด 2. จำเป็นต้องมีการเรียนรู้พฤติกรรมปกติหรือเหตุการณ์ปกติของระบบให้มากขึ้นเพื่อตรวจจับการบุกรุกได้อย่างถูกต้อง 3. ไม่สามารถตรวจจับการบุกรุกได้หากระบบตรวจจับการบุกรุกดังกล่าวอยู่ในขั้นตอนการเรียนรู้

Zhang and Zulkernine (2006) ได้มีการนำเอาข้อดีของทั้ง 2 เทคนิคมารวมกัน เพื่อให้ได้ประโยชน์สูงสุด เรียกว่า การรวมกันของเทคนิคการตรวจจับแบบผิดปกติและเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบ ซึ่งจะได้นำเสนอรายละเอียดในลำดับถัดไป

2.4.3 การรวมกันของเทคนิคการตรวจจับความผิดปกติและเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบ

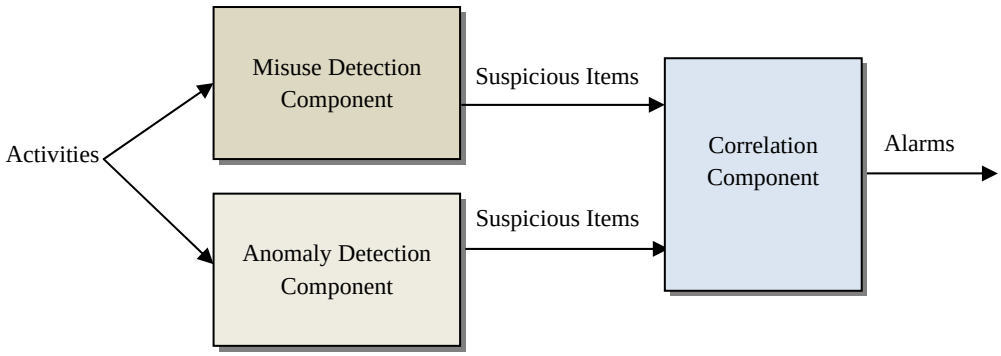
การรวมกันแบบแรกเป็นการตรวจจับการบุกรุกแบบเรียงต่อกันระหว่างเทคนิคการตรวจจับความผิดปกติและเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบ ดังรูปที่ 2.6



รูปที่ 2.6 การตรวจจับความผิดปกติตามด้วยการตรวจจับการใช้สิทธิ์โดยมิชอบ
ที่มา : Zhang and Zulkernine, 2006

จากรูปจะเห็นว่าเมื่อเหตุการณ์หรือพฤติกรรมเข้าสู่เทคนิคการตรวจจับความผิดปกติจะได้ผลออกเป็นพฤติกรรมที่ต้องสงสัยหรือไม่แน่ชัดว่าเป็นการบุกรุก (Suspicious) ซึ่งจะถูกส่งต่อเข้าสู่เทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบ เมื่อผ่านการตรวจแล้วหากตรงกับรูปแบบการบุกรุกที่มีอยู่ก็จะแจ้งเป็นผู้บุกรุกจริง แต่หากไม่ตรงก็จะแจ้งเป็นพฤติกรรมปกติหรือเหตุการณ์หรือพฤติกรรมที่ไม่รู้จัก

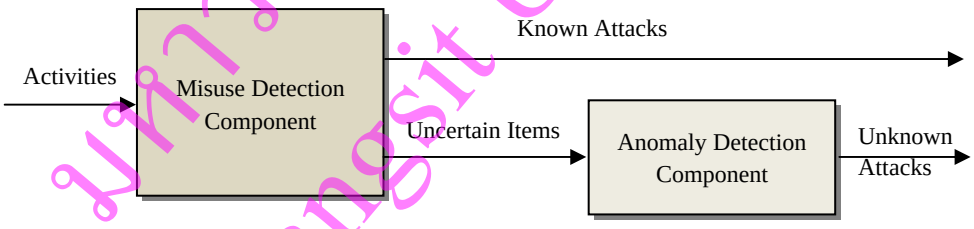
การรวมกันแบบที่สองเป็นการตรวจจับการบุกรุกแบบคู่ขนานกันระหว่างเทคนิคการตรวจจับความผิดปกติและเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบ เมื่อเหตุการณ์หรือพฤติกรรมเข้าสู่เทคนิคการตรวจจับความผิดปกติและเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบ จะได้ผลออกมาเป็นพฤติกรรมที่ต้องสงสัยหรือไม่แน่ชัดว่าเป็นการบุกรุก จากนั้นนำเอาผลลัพธ์ที่ได้จากทั้งสองเทคนิคมารวมกันเพื่อหาความสัมพันธ์ ระบุการบุกรุกและแจ้งเตือนผู้ดูแลระบบ ดังรูปที่ 2.7



รูปที่ 2.7 การตรวจจับความผิดปกติขนานกับการตรวจจับการใช้สิทธิ์โดยมิชอบ

ที่มา : Zhang and Zulkernine, 2006

ในการรวมกันแบบสุดท้าย เริ่มเหตุการณ์หรือพฤติกรรมส่งเข้าสู่เทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบก่อน หากพบว่าเป็นการบุกรุกก็จะระบุเป็นการบุกรุกทันที ส่วนที่ไม่แน่ชัดให้ส่งต่อเข้าสู่เทคนิคการตรวจจับความผิดปกติต่อไป ซึ่งจะได้ผลออกมาเป็นพฤติกรรมปกติและพฤติกรรมที่ไม่แน่ชัดว่าเป็นการบุกรุกหรือพฤติกรรมที่ไม่รู้จัก ดังรูปที่ 2.8



รูปที่ 2.8 การตรวจจับการใช้สิทธิ์โดยมิชอบตามด้วยการตรวจจับความผิดปกติ

ที่มา : Zhang and Zulkernine, 2006

2.5 เทคนิคการจัดกลุ่ม (Clustering Technique)

(Gomez, http://www.dis.unal.edu.co/~jgomezpe, 10 ตุลาคม 2557) และสิรภัทร เชี่ยวชาญวัฒนา (2552) ได้อธิบายถึงเทคนิคการจัดกลุ่มข้อมูล (Data Clustering) ว่าเป็นเทคนิคที่จัดอยู่ในกลุ่มการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) หรือระเบียนที่นำมาใช้ไม่มีเฉลย โดยพยายามจัดกลุ่มของระเบียนที่มีความคล้ายคลึงกันหรืออยู่ใกล้กันมากที่สุด (Intra-Cluster Distance) ไว้กลุ่ม

หนึ่ง และกลุ่มของระเบียนที่มีความแตกต่างกันหรืออยู่ไกลกันมากที่สุด (Inter-Cluster Distance) ไว้อีกกลุ่มหนึ่ง การจัดข้อมูลเข้ากลุ่มพิจารณาจากการนำข้อมูลที่มีคุณสมบัติหรือลักษณะเหมือนกันหรือมีความคล้ายกันมาจัดไว้ในกลุ่มเดียวกัน ซึ่งมีการคำนวณความเหมือนของกลุ่มข้อมูลหรือความใกล้กันของข้อมูล โดยใช้มาตรวัดความคล้าย (Similarity Measure) หรือมาตรวัดระยะห่าง (Distance Measure) มาตรวัดที่ใช้ในการจัดกลุ่มนั้นมีหลายมาตรวัด เช่น มาตรวัดค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน (Pearson Correlation Coefficient) มาตรวัดยูคลิด (Euclidean) และมาตรวัดแมนฮัตตัน (Manhattan) โดย Chimphlee (2008) กล่าวว่ามาตรวัดที่นิยมใช้กันมากที่สุด คือ มาตรวัดยูคลิดซึ่งนำมาใช้วัดระยะห่าง (Distance) เพื่อจัดกลุ่มข้อมูล

Zhong, Khoshgoftaar and Seliya (2004) ได้กล่าวว่าโดยทั่วไปเทคนิคการจัดกลุ่มข้อมูลสามารถแบ่งออกเป็นการจัดกลุ่มทุกคู่ (Pairwise) ซึ่งใช้กันในช่วงแรก โดยพิจารณาจากการวัดระยะห่างตามมาตรวัดที่กำหนด การจัดกลุ่มจากจุดเซนทรอยด์ (Centroid) ซึ่งใช้จุดดังกล่าวเป็นตัวแทนของกลุ่มข้อมูล การจัดกลุ่มแบบนี้มักจะมีประสิทธิภาพดีกว่าการจัดกลุ่มจากจำนวนกลุ่มที่ต้องการ เช่น การจัดกลุ่มแบบเคมีนส์ (*k*-Means) นอกจากนี้ยังมีการจัดกลุ่มแบบคลุมเครือ (Fuzzy) ซึ่งนำเอาความน่าจะเป็น (Probability) หรือค่าความเป็นสมาชิก (Membership) ของแต่ละกลุ่มมาใช้ในการคิดด้วย เช่น การจัดกลุ่มแบบฟัซซี่ซีมีนส์ (Fuzzy *c*-Means)

2.5.1 การจัดกลุ่มแบบเคมีนส์

เทคนิคการจัดกลุ่มแบบเคมีนส์ได้รับความนิยมตั้งแต่อดีตจนถึงปัจจุบันและมีการใช้งานอย่างแพร่หลาย (กาญจนา จรัญศิริโพศาล, สิรภัทร์ เชี่ยวชาญวัฒนา และคำรณ สุนัติ, 2554) ซึ่งพัฒนาและนำเสนอโดย MacQueen ในปี ค.ศ. 1967 การจัดกลุ่มแบบนี้เป็นวิธีการแบ่งกลุ่มข้อมูลทั้งหมดออกตามจำนวนกลุ่มที่ต้องการ ค่าเฉลี่ยของกลุ่มจะเรียกว่า เซนทรอยด์ ซึ่งใช้เป็นจุดศูนย์กลางในการวัดระยะห่างของข้อมูลในกลุ่มเดียวกัน โดยวัดจากระยะห่างที่น้อยที่สุดระหว่างข้อมูลกับจุดดังกล่าว เพื่อจัดข้อมูลเข้าสู่กลุ่มต่างๆ ตามจำนวนกลุ่มที่ต้องการ ระยะห่างที่นิยมใช้คือระยะห่างแบบยูคลิด

Wilkin and Huang (2007) ได้กล่าวว่าการจัดกลุ่มแบบเคมีนส์เป็นที่นิยมใช้กันในหลายงานวิจัย และ Watthananon (2012) ได้สนับสนุนคำกล่าวข้างต้นว่าเนื่องจากการจัดกลุ่มแบบเคมีนส์มีขั้นตอนวิธีในการปฏิบัติที่ง่ายและได้ผลดี โดยใช้ระยะห่างของข้อมูลเป็นตัวกำหนด แต่

เทคนิคนี้มีข้อจำกัดในเรื่องของเวลาซึ่งสัมพันธ์กับมิติข้อมูลและจำนวนกลุ่มต่อการวนซ้ำ ซึ่งทำให้ต้องใช้ทรัพยากรสิ้นเปลืองมากเมื่อต้องทำงานกับชุดข้อมูลขนาดใหญ่ ศุภาวีร์ มากดี (2554) ได้สรุปขั้นตอนการจัดกลุ่มแบบเคมีนส์ไว้ดังนี้

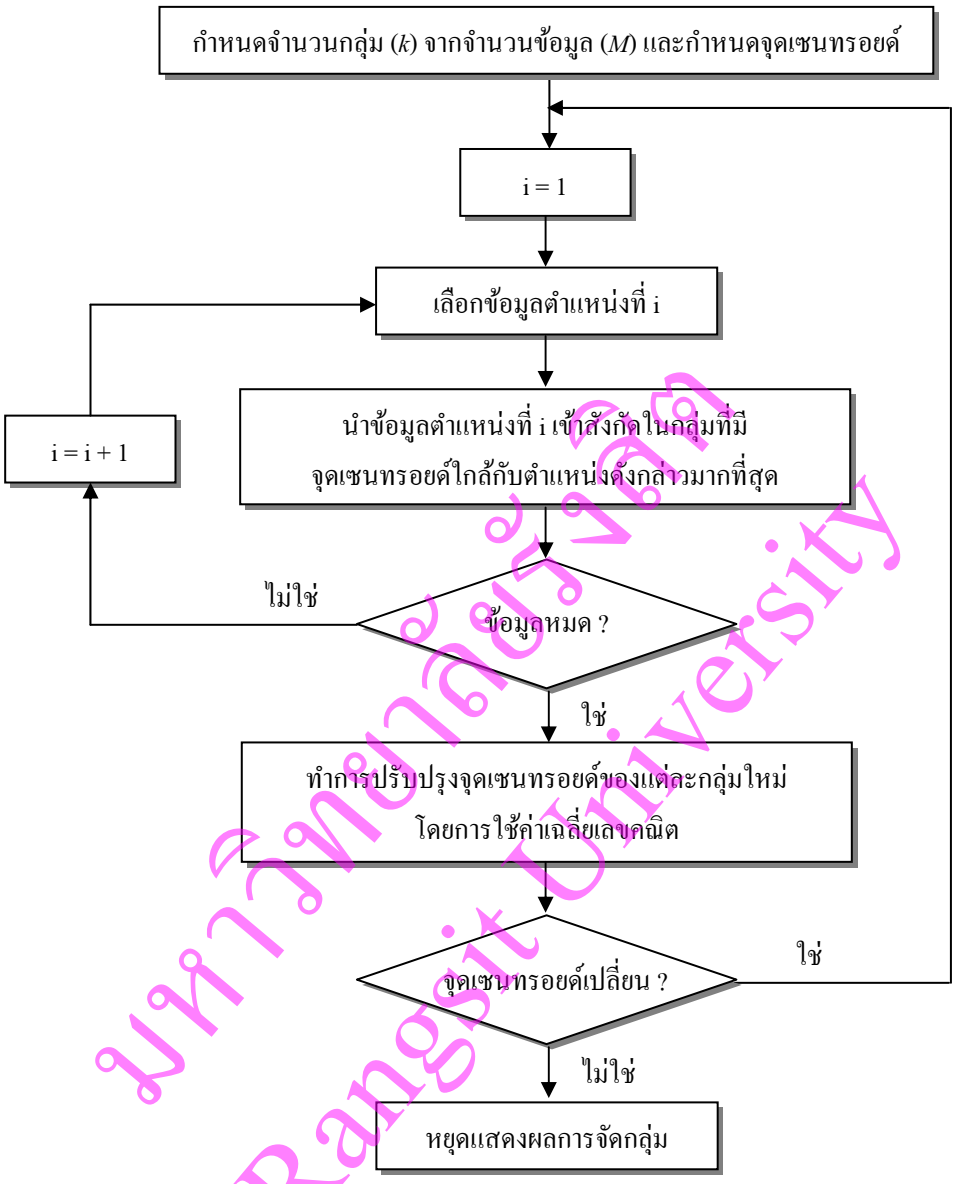
1) สุ่มข้อมูลออกเป็นกลุ่มย่อย k กลุ่ม โดยที่ค่า k ต้องเป็นเลขจำนวนเต็มและมีค่าน้อยกว่าจำนวนข้อมูลทั้งหมด

2) เมื่อได้กลุ่มข้อมูลแต่ละกลุ่มแล้ว ทำการคำนวณหาจุดเซนทรอยด์ โดยใช้ค่าเฉลี่ยเลขคณิต (Arithmetic Mean)

3) ตรวจดูระยะห่างระหว่างข้อมูลกับจุดเซนทรอยด์ โดยข้อมูลใดที่มีระยะห่างสั้นที่สุดก็นำข้อมูลไปสังกัดกลุ่มข้อมูลดังกล่าว (ทำให้เกิดการย้ายกลุ่มข้อมูล)

4) ทำการคำนวณจุดเซนทรอยด์และกำหนดข้อมูลให้กับกลุ่มข้อมูลใหม่ไปเรื่อยๆ จนกว่าข้อมูลในแต่ละกลุ่มไม่สามารถเปลี่ยนกลุ่มได้ดีกว่าเดิม

Niknam et al. (2011) ได้สรุปขั้นตอนการจัดกลุ่มแบบเคมีนส์ออกมาในรูปของผังงาน ดังรูปที่ 2.9



รูปที่ 2.9 ผังงานการจัดกลุ่มแบบเคมีนส์

ที่มา : Niknam et al., 2011

2.5.2 การจัดกลุ่มแบบฟัซซี่ซีมีนส์

Brezdek (1981) ได้อธิบายว่าการจัดกลุ่มแบบฟัซซี่ซีมีนส์เป็นวิธีการที่พัฒนาเพิ่มเติมจากการจัดกลุ่มแบบดั้งเดิม โดยคำนวณค่าความเป็นสมาชิก (Membership Value) ของข้อมูลในแต่

ละกลุ่ม ทำให้ในการจัดกลุ่มแบบนี้ไม่ได้แบ่งแยกกลุ่มข้อมูลอย่างชัดเจน ข้อมูลแต่ละตัวจะมีค่าความเป็นสมาชิกอยู่ในทุกกลุ่ม ซึ่งอาจจะมากน้อยแตกต่างกัน

การจัดกลุ่มแบบฟัซซี่ซีมีนส์เป็นเทคนิคที่ยอมให้ข้อมูลในแต่ละกลุ่มมีการซ้อนทับกันได้ โดยอาศัยการให้ค่าความเป็นสมาชิกของข้อมูลต่อกลุ่มข้อมูลต่างๆ ในการหาค่าความเป็นสมาชิกจะต้องทราบระยะห่างระหว่างข้อมูลและจุดเซนทรอยด์ของกลุ่มข้อมูลเหล่านั้น การวัดระยะทางจึงมีความสำคัญต่อการจัดกลุ่มแบบนี้ ซึ่งอาจเป็นการวัดระยะทางแบบยูคลีด หรือการวัดระยะทางแบบมาฮาลาโนบิส (Mahalanobis Distance) การวัดระยะทางแบบยูคลิดจะไม่เหมาะกับข้อมูลที่มีความเกี่ยวเนื่องกัน ส่วนการวัดระยะทางแบบมาฮาลาโนบิสจะเหมาะสำหรับกลุ่มข้อมูลที่มีค่านอกกลุ่ม (Outlier) และแต่ละกลุ่มข้อมูลที่มีความหนาแน่นแตกต่างกัน การจัดกลุ่มแบบฟัซซี่ซีมีนส์นั้นเป็นเทคนิคในการจัดกลุ่มที่แก้ไขข้อด้อยของการจัดกลุ่มแบบเคมีนส์ เนื่องจากการจัดกลุ่มแบบเคมีนส์ไม่เหมาะสมกับข้อมูลที่มีความสัมพันธ์กัน (Correlation) ขั้นตอนการจัดกลุ่มแบบฟัซซี่ซีมีนส์สามารถสรุปได้ดังนี้

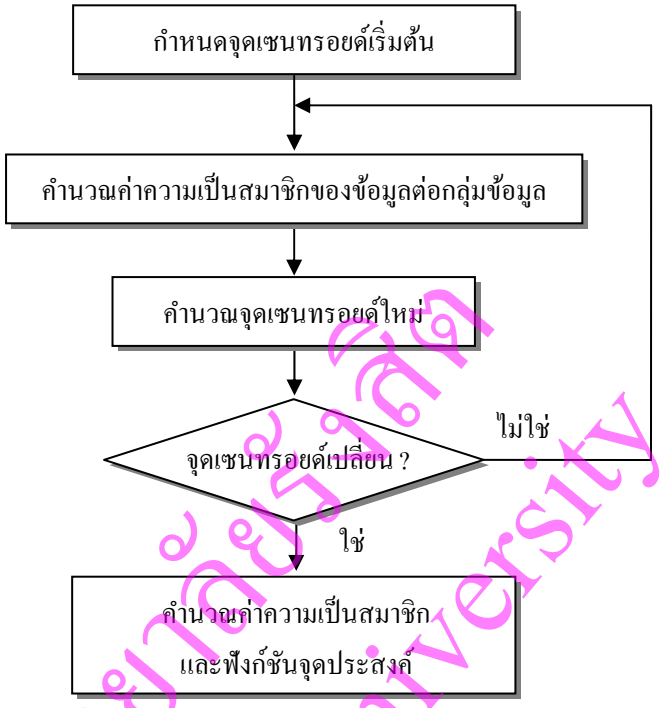
1) กำหนดกลุ่มข้อมูลที่ต้องการแบ่งกลุ่ม เงื่อนไขในการให้ข้อมูลหยุดการจัดกลุ่ม ค่าฟัซซี่พารามิเตอร์ และจุดเซนทรอยด์เริ่มต้น

2) คำนวณค่าความเป็นสมาชิกของข้อมูลต่อกลุ่มข้อมูลต่างๆ

3) คำนวณจุดเซนทรอยด์ของกลุ่มข้อมูลใหม่และตรวจสอบเงื่อนไขจากการลบค่าความเป็นสมาชิกใหม่ด้วยค่าความเป็นสมาชิกก่อนหน้า

4) ถ้าเงื่อนไขเป็นจริงให้คำนวณค่าความเป็นสมาชิกและค่าฟังก์ชันจุดประสงค์ (Objective Function) ถ้าเงื่อนไขเป็นเท็จให้คำนวณค่าความเป็นสมาชิกจากจุดเซนทรอยด์ล่าสุดและคำนวณรอบต่อไป

ศุภาวีร์ มากดี (2554) ได้แสดงผังงานของการจัดกลุ่มแบบฟัซซี่ซีมีนส์ ดังรูปที่ 2.10



รูปที่ 2.10 ผังงานการจัดกลุ่มแบบฟัซซี่ซีมีนส์

ที่มา : ศุภาวีร์ มากดี, 2554

2.6 เทคนิคการตรวจหาค่านอกกลุ่มจากการวัดระยะทาง (Distance Based Outlier Detection)

Freedman, Pisani, and Purves (1997), Hawkins (1980) และ Kantardzic (2003) ได้ให้คำจำกัดความว่า เป็นวิธีการหาข้อมูลที่มีความแตกต่างไปจากข้อมูลส่วนใหญ่ เพื่อช่วยให้การทำงานเร็วขึ้นหรือช่วยแยกข้อมูลออกเป็นกลุ่มได้โดยใช้สถิติ (Statistics)

เทคนิคการตรวจหาค่านอกกลุ่มนั้นแตกต่างกันไปตามลักษณะของข้อมูล เช่น ถ้ามีข้อมูลหนึ่งมิติ เป็น 1, 4, 5, 40, 50, 55, 67, 88, 94, 96, 176 และ 200 ซึ่งจะเห็นได้ว่าเลข 176 และ 200 มีค่ามากกว่าข้อมูลส่วนใหญ่ ดังนั้น 176 และ 200 ถือได้ว่าเป็นค่านอกกลุ่ม หรือถ้านำเอาเทคนิคการตรวจหาค่านอกกลุ่มมาใช้ในการจัดกลุ่มข้อมูลก็สามารถทำได้เช่นกัน กล่าวคือ เราแบ่งข้อมูล

ออกเป็น 2 กลุ่มดังนี้ กลุ่มที่หนึ่งมีสมาชิกเป็น 1, 4, 5, 40, 50, 55, 67, 88, 94 และ 96 กลุ่มที่สองมีสมาชิกเป็น 176 และ 200 เป็นต้น

Han และ Kamber (2000) ได้กล่าวว่าในกรณีที่ข้อมูลมีขนาดใหญ่และมีมากกว่าหนึ่งมิติ เทคนิคที่นิยมใช้สำหรับการหาค่านอกกลุ่ม คือ การหาค่านอกกลุ่มจากการวัดระยะห่างระหว่างข้อมูล และ Kantardzic (2003) ได้อธิบายว่ามีพารามิเตอร์ที่สำคัญอยู่ 2 พารามิเตอร์ ได้แก่ D และ P โดย D คือ ระยะห่างของข้อมูลที่กำหนด หากระยะห่างของข้อมูลมีค่ามากกว่า D จะถือว่าเป็นส่วนหนึ่งในการพิจารณาค่านอกกลุ่ม และ P คือ จำนวนของข้อมูล ซึ่งต้องนำมาพิจารณาด้วย เนื่องจากข้อมูลมีมากกว่าหนึ่งมิติ ฉะนั้นถ้าข้อมูลมีระยะห่างมากกว่า D และมีจำนวนมากกว่า P แล้วจะถือว่าเป็นค่านอกกลุ่ม พารามิเตอร์ทั้งสองต้องกำหนดให้เหมาะสม เพื่อให้ได้ผลลัพธ์ที่ดี

2.7 เทคนิคต้นไม้ตัดสินใจ (Decision Tree)

Negnevitsky (2011) ได้กล่าวว่าความรู้ หมายถึง ทฤษฎีหรือการประยุกต์ความเข้าใจในเรื่องใดเรื่องหนึ่งอย่างชัดเจน ยิ่งมีความรู้มากกว่าผู้อื่นมากเท่าไรยิ่งจะได้เปรียบมากขึ้นเท่านั้น เราเรียกผู้รู้เหล่านั้นว่า ผู้เชี่ยวชาญ (Experts) ซึ่งถือว่ามีความสำคัญอย่างมาก ดังจะเห็นได้ว่าองค์กรที่จะประสบความสำเร็จได้อย่างน้อยต้องมีผู้เชี่ยวชาญอยู่ด้วยเสมอ ความรู้ต่างๆของผู้เชี่ยวชาญสามารถถ่ายทอดออกมาในรูปแบบของกฎ (Rule) ได้ เช่น ชาวต่างด้าวเข้ามาในประเทศและต้องการข้ามถนนให้ปลอดภัย เราก็จะอธิบายว่าหากต้องการข้ามถนนให้ปลอดภัยจะต้องดูสัญญาณไฟจราจรก่อนข้ามถนนทุกครั้ง นั่นคือถ้าสัญญาณไฟเป็นสีเขียวให้ข้ามถนนได้ แต่ถ้าสัญญาณไฟเป็นสีแดงให้หยุดห้ามข้ามถนน ความรู้ดังกล่าวสามารถถ่ายทอดออกมาในรูปแบบ IF-THEN ซึ่งจะเรียกว่า การสร้างกฎ (Rule Production) ได้ดังนี้

IF "สัญญาณไฟจราจรสีเขียว"
THEN ข้ามถนนได้

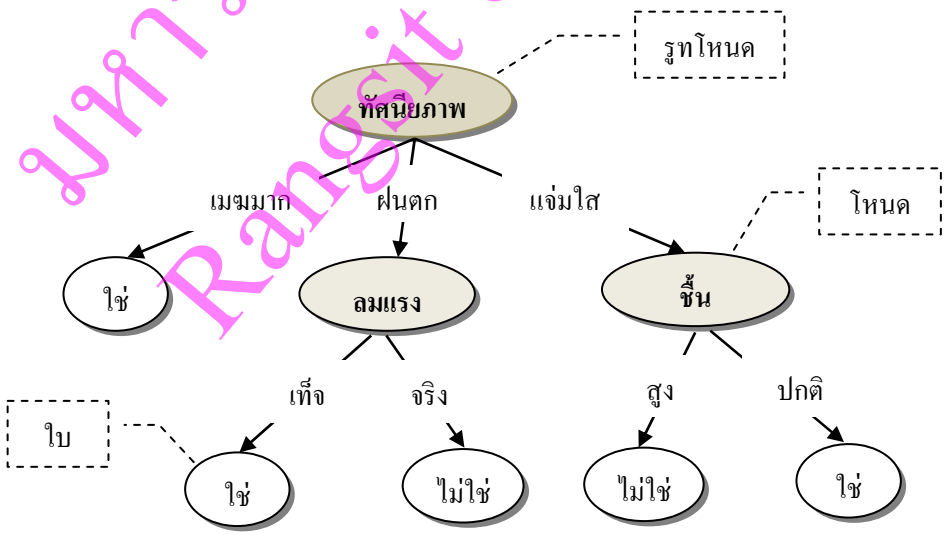
IF "สัญญาณไฟจราจรสีแดง"
THEN หยุดห้ามข้ามถนน

อนึ่งจะเห็นว่าเมื่อนำกฎนี้ไปใช้ ชาวต่างด้าวก็จะรู้และเข้าใจได้ว่าจะข้ามถนนอย่างไรให้ปลอดภัย แสดงว่าความรู้นั้นถ่ายทอดออกมาได้ด้วยกฎ IF-THEN ถ้าเปรียบเทียบการทำงานของ

คอมพิวเตอร์ ซึ่งความรู้เก็บอยู่ในฐานข้อมูล เมื่อพบ IF แล้วดูเงื่อนไข (Condition) ที่ตรงกับกฎที่ตั้งไว้ในฐานข้อมูลจริง ให้กระทำคำสั่งที่อยู่ภายหลัง THEN

Chan et al. (2004) ได้อธิบายถึงการนำเทคนิคระบบผู้เชี่ยวชาญมาใช้กับระบบตรวจจับการบุกรุกว่า เป็นกระบวนการในการวิเคราะห์ทราฟฟิกข้อมูล ซึ่งพิจารณาว่าเป็นพฤติกรรมปกติหรือบุกรุก โดยระบบตรวจจับการบุกรุกดังกล่าวจะมีความรู้เกี่ยวกับทราฟฟิกข้อมูลที่ผิดปกติ (Attack Behavior) และการกระทำที่เกิดขึ้นในภายหลังหรือกิจกรรมที่ไม่ได้รับอนุญาตให้ใช้งาน (Unauthorized Activites) ซึ่งจะแสดงออกมาในรูปของกฎ IF-THEN การสร้างกฎด้วยระบบผู้เชี่ยวชาญมีข้อจำกัดด้านบุคลากรที่ต้องมีจำนวนมากและใช้เวลานานในการวิเคราะห์และสร้างกฎ ซึ่งปัจจุบันนิยมใช้เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เข้ามาช่วยสร้างกฎแทนระบบผู้เชี่ยวชาญ

เทคนิคต้นไม้ตัดสินใจได้รับการพัฒนาขึ้นโดย J. R. Quinland ในปี ค.ศ. 1986 จัดอยู่ในกลุ่มการเรียนรู้แบบมีผู้สอน มีลักษณะโครงสร้างข้อมูลเป็นลำดับชั้น (Hierarchy) ซึ่งเป็นที่นิยมใช้กันอย่างแพร่หลาย และอาศัยหลักการสร้างกฎในรูปแบบ IF-THEN ต้นไม้ตัดสินใจประกอบด้วยโหนด (Node) กิ่ง (Branch) และใบ (Leaf) แสดงได้ดังรูปที่ 2.11



รูปที่ 2.11 ส่วนประกอบของต้นไม้ตัดสินใจ

หลักการพื้นฐานของการสร้างต้นไม้ตัดสินใจ คือ การสร้างรูทโหนด (Root Node) ลงมาก่อน แล้วตามด้วยโหนดจนถึงกิ่ง มีลักษณะการสร้างจากบนลงล่าง Han และ Kamber (2006) ได้อธิบายขั้นตอนการสร้างต้นไม้ตัดสินใจดังนี้

1) เริ่มต้นสร้างโหนดขึ้นมาหนึ่งโหนดจากชุดข้อมูล

2) ถ้าข้อมูลทั้งหมดอยู่ในกลุ่มเดียวกันแล้ว ให้โหนดที่สร้างขึ้นนั้นเป็นโหนดใบและตั้งชื่อแยกตามกลุ่มของข้อมูลนั้น

3) ถ้าข้อมูลไม่มีคุณลักษณะใดที่เหมาะสมในการแบ่งกลุ่ม ให้โหนดที่สร้างขึ้นนั้นเป็นโหนดใบและตั้งชื่อตามกลุ่มที่มีข้อมูลสนับสนุนมากที่สุด

4) ถ้าข้อมูลภายในโหนดมีหลากหลายกลุ่มปะปนกันให้ทำการเลือกคุณลักษณะที่มีความเหมาะสมที่สุด โดยพิจารณาจากอินฟอร์เมชันเกน (Information Gain: IG) ซึ่งเป็นตัวชี้วัดความสามารถในการจำแนกกลุ่ม

5) เมื่อได้ตัวทดสอบการตัดสินใจ ให้สร้างกิ่งของต้นไม้ด้วยค่าต่างๆที่เป็นไปได้ของตัวทดสอบและแบ่งข้อมูลออกตามกิ่งต่างๆที่สร้างขึ้น

6) พิจารณาข้อมูลในแต่ละกิ่ง ถ้าพบว่าข้อมูลทั้งหมดอยู่ในกลุ่มเดียวกัน ให้ต่อกิ่งด้วยโหนดใบและกำหนดค่าด้วยกลุ่มของข้อมูลนั้น แต่ถ้าพบว่าข้อมูลมีหลากหลายกลุ่มปะปนกัน ให้ทำการวนซ้ำการหาตัวทดสอบการตัดสินใจที่เหมาะสมต่อไป

7) ทำการวนซ้ำเพื่อแบ่งข้อมูลและแตกกิ่งของต้นไม้ไปเรื่อยๆ โดยการวนซ้ำจะสิ้นสุดก็ต่อเมื่อเงื่อนไขข้อใดข้อหนึ่งต่อไปนี้เป็นจริง

7.1) ข้อมูลทั้งหมดในโหนดอยู่ในกลุ่มเดียวกัน ให้โหนดที่สร้างขึ้นนั้นเป็นโหนดใบและตั้งชื่อกลุ่มของข้อมูลนั้น

7.2) ไม่มีคุณลักษณะใดที่เหมาะสมในการแบ่งกลุ่ม ให้โหนดที่สร้างขึ้นนั้นเป็นโหนดใบและกำหนดค่าด้วยกลุ่มที่มีข้อมูลสนับสนุนมากที่สุด

เอกสิทธิ์ พัชรวงศ์ศักดา (2557) อธิบายถึงการสร้างกฎด้วยเทคนิคต้นไม้ตัดสินใจว่า ทำการคัดเลือกคุณลักษณะที่มีความสัมพันธ์กับคลาสมากที่สุดเป็นรูทโหนด การหาความสัมพันธ์ของคุณลักษณะนี้จะใช้ตัววัดที่เรียกว่าอินฟอร์เมชันเกน ซึ่งถ้าคุณลักษณะใดมีค่าอินฟอร์เมชันเกนสูงแสดงว่าคุณลักษณะนั้นสามารถจำแนกกลุ่มได้ดี ช่วยลดจำนวนครั้งของการทดสอบในการแยกแยะข้อมูลและรับประกันว่าต้นไม้ตัดสินใจที่ได้จะไม่มีความซับซ้อนมากจนเกินไป โดยค่าอินฟอร์เมชันเกนคำนวณได้จากสมการ

$$IG(\text{parent},\ \text{child}) = Entropy\ (\text{parent}) - \sum_i [p(c_i) \times Entropy\ (c_i)] \qquad (2.1)$$

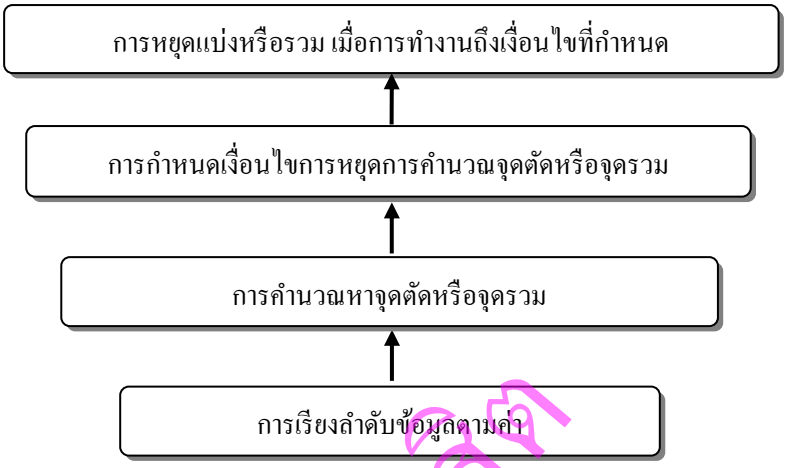
โดยที่ $Entropy(c_i) := -p(c_i)\log_2 p(c_i)$ คือ เอนโทรปีของ c_i

$p(c_i)$ คือ ค่าความน่าจะเป็นของ c_i

เอนโทรปีนี้จะใช้ในการวัดความแตกต่างของข้อมูล นั่นคือ ถ้าข้อมูลมีความแตกต่างกันน้อย เอนโทรปีจะมีค่าต่ำ แต่ถ้าข้อมูลมีความแตกต่างกันมาก เอนโทรปีจะมีค่าสูง ดังนั้นถ้าโหนดลูก (Child) สามารถแบ่งแยกข้อมูลได้ดีจะมีค่าเอนโทรปีต่ำ และจะทำให้ค่าอินฟอร์เมชันเกนสูงเมื่อเทียบกับโหนดแม่ (Parent) ในขั้นตอนการสร้างแบบจำลองของต้นไม้ตัดสินใจ จะคำนวณอินฟอร์เมชันเกนของแต่ละคุณลักษณะเทียบกับคลาสเพื่อหาคุณลักษณะที่มีค่าอินฟอร์เมชันเกนมากที่สุดและให้เป็นรูทโหนดของต้นไม้ตัดสินใจ

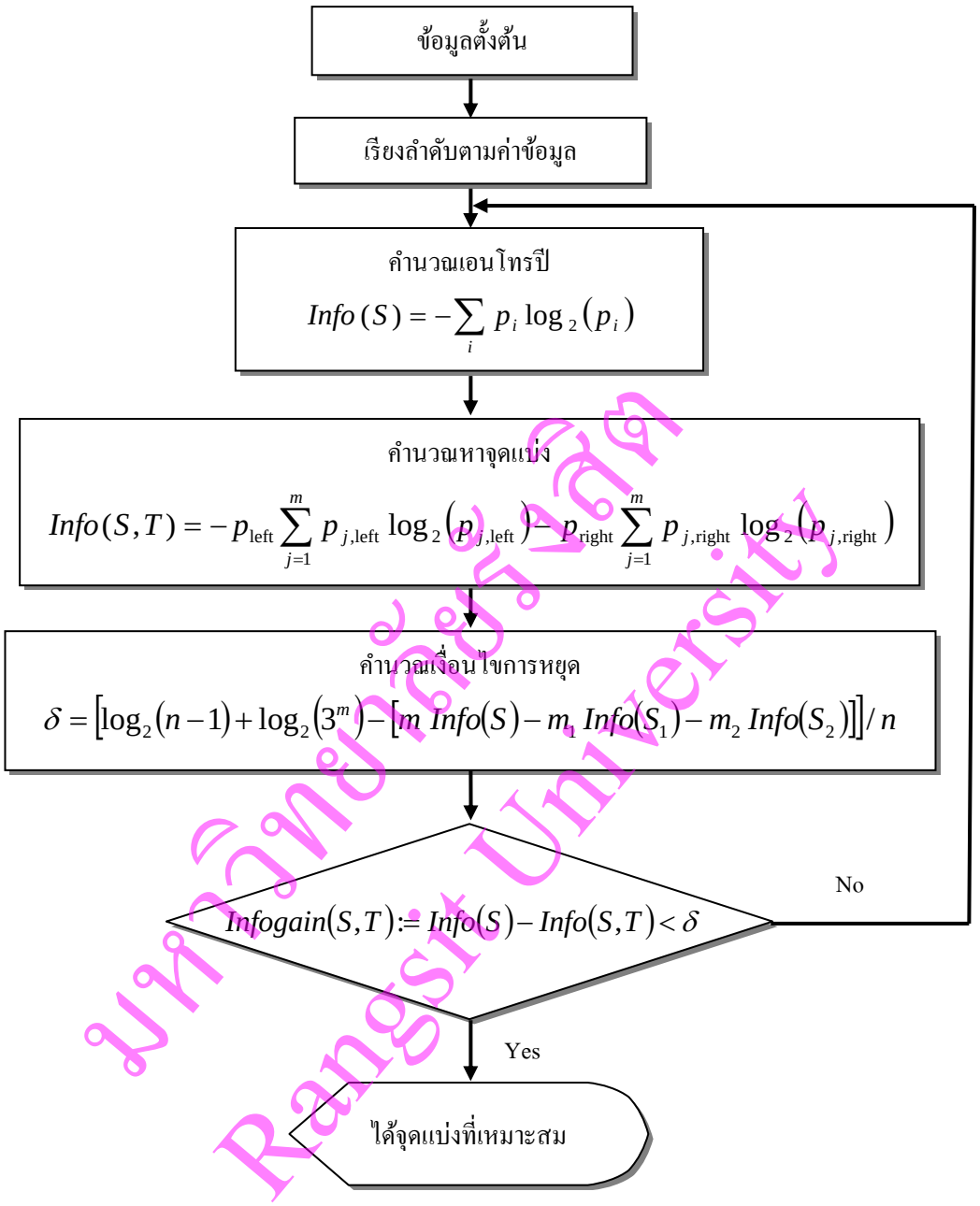
2.8 เทคนิคการแบ่งข้อมูลแบบเอ็มดีแอล (Minimum Description Length : MDL Discretization)

Dash, Paramguru, Dash (2011) และ Hemada, Lakshmi (2013) ได้อธิบายถึงเทคนิคการแบ่งข้อมูล (Discretization) ว่าเป็นกระบวนที่ใช้ในการแปลงข้อมูลของคุณลักษณะที่เป็นค่าต่อเนื่อง (Continuous) ให้เป็นค่าไม่ต่อเนื่อง (Discrete) โดยมีจุดมุ่งหมายเพื่อให้เข้าใจข้อมูลได้ง่ายขึ้น ลดเวลาและรายจ่ายอื่นๆในการนำไปใช้งาน อีกทั้งยังช่วยเพิ่มความแม่นยำและประสิทธิภาพในการจำแนกประเภท กระบวนการทำงานหลักของการแบ่งข้อมูลจะมี 4 ส่วน คือ การเรียงลำดับข้อมูลตามค่า การคำนวณหาจุดตัด (Cut Point) หรือจุดรวม (Merging Point) การกำหนดเงื่อนไขการหยุดการคำนวณจุดตัดหรือจุดรวม และการหยุดแบ่งหรือรวม เมื่อการทำงานถึงเงื่อนไขที่กำหนด ดังรูปที่ 2.12



รูปที่ 2.12 ขั้นตอนการทำงานของเทคนิคการแบ่งข้อมูล

เทคนิคการแบ่งข้อมูลแบบเอ็มดีแอล (Minimum Description Length : MDL) เป็นเทคนิคที่จัดอยู่ในกลุ่มการเรียนรู้แบบมีผู้สอนและใช้หลักการคำนวณเอนโทรปี เพื่อหาจุดตัดและกำหนดเงื่อนไขการหยุดจากค่าขีดเริ่มเปลี่ยน ขั้นตอนการทำงานของเทคนิคนี้สามารถแสดงในรูปแบบผังงานได้ดังรูปที่ 2.13



รูปที่ 2.13 ผังงานของเทคนิคการแบ่งข้อมูลแบบเอ็มดีแอล

จากผังงานของเทคนิคการแบ่งข้อมูลแบบเอ็มดีแอล สามารถอธิบายขั้นตอนต่างๆได้ดังนี้

1) นำเข้าข้อมูลชนิดที่เป็นค่าต่อเนื่อง

2) เรียงลำดับข้อมูลที่นำเข้ามาตามค่าข้อมูล

3) คำนวณหาเอนโทรปีของข้อมูลจากสมการ

$$Info(S) = -\sum_{i} p_i \log_2(p_i) \quad (2.2)$$

โดยที่ S คือ เซตของข้อมูลที่ประกอบด้วยชุดของตัวแปรต้นและตัวแปรตามหลายๆกรณี

p_i คือ ความน่าจะเป็นของคลาสแต่ละคลาส

เมื่อได้เอนโทรปีของข้อมูลจะนำไปใช้หาจุดแบ่งที่เหมาะสมต่อไป

4) หาจุดแบ่งของข้อมูลที่เป็นค่าต่อเนื่อง โดยทำการคำนวณหาค่าเอนโทรปีของทุกจุดแบ่งที่เป็นไปได้และเลือกจุดแบ่งที่มีค่าเอนโทรปีที่ต่ำที่สุด จากสมการ

$$Info(S,T) = -p_{\text{left}} \sum_{j=1}^{m} p_{j,\text{left}} \log_2\left(p_{j,\text{left}}\right) - p_{\text{right}} \sum_{j=1}^{m} p_{j,\text{right}} \log_2\left(p_{j,\text{right}}\right) \quad (2.3)$$

โดยที่ $Info(S,T)$ คือ เอนโทรปีของเซตข้อมูล S ที่จุดแบ่ง T

p_{left} คือ ค่าความน่าจะเป็นของช่วงข้อมูลด้านซ้าย

p_{right} คือ ค่าความน่าจะเป็นของช่วงข้อมูลด้านขวา

j คือ ลำดับในเซตข้อมูล S

m คือ จำนวนคลาสในเซตข้อมูล S

5) คำนวณหาค่าเดลต้า (δ) ที่ใช้สำหรับพิจารณาการหยุดแบ่งข้อมูล จากสมการ

$$\delta = \left[\log_2(n-1) + \log_2(3^m) - \left[m\ Info(S) - m_1\ Info(S_1) - m_2\ Info(S_2)\right]\right]/n \quad (2.4)$$

โดยที่ n คือ จำนวนข้อมูลทั้งหมดในเซต S

S_1 คือ เซตย่อยที่ได้จากการแบ่งเซต S ที่จุดแบ่ง T (ด้านซ้าย)

S_2 คือ เซตย่อยที่ได้จากการแบ่งเซต S ที่จุดแบ่ง T (ด้านขวา)

m_1 คือ จำนวนคลาสในเซตย่อย S_1

m_2 คือ จำนวนคลาสในเซตย่อย S_2

6) เมื่อคำนวณค่าต่างๆครบแล้ว ก็พิจารณาว่าการแบ่งข้อมูลเหมาะสมหรือไม่ โดยพิจารณาจากเงื่อนไข

$$Infogain(S,T) := Info(S) - Info(S,T) < \delta \quad (2.5)$$

โดยถ้า $Infogain(S,T)$ มีค่าน้อยกว่าค่าของเดลต้า แสดงว่า T เป็นจุดแบ่งที่เหมาะสมแล้ว แต่ถ้าไม่จริงก็ให้กลับไปทำงานซ้ำอีกเพื่อคำนวณหาจุดแบ่งต่อไป

2.9 งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวกับระบบตรวจจับการบุกรุก และมีการพัฒนาอย่างต่อเนื่อง ตัวอย่างมีดังนี้

Narayan, Singh and Tak (2012) ได้ศึกษาการจัดกลุ่มแบบฟัซซี่ซีมีนส์สำหรับระบบตรวจจับการบุกรุก และกล่าวอ้างถึงข้อจำกัดของการจัดกลุ่มแบบเคมีนส์ รวมทั้งนำเสนอการจัดกลุ่มแบบฟัซซี่ซีมีนส์ที่มีค่าคาดหมายมากที่สุด (Expectation-Maximization) และใช้ชุดข้อมูล KDD Cup 1999 โดยสุ่มมาเพียง 2,000 ระเบียน ผลการวิจัยพบว่ามีอัตราการแจ้งผิดพลาดที่ต่ำลง

Hoque, Mukit and Bikas (2012) ได้ประยุกต์ใช้ขั้นตอนวิธีทางพันธุกรรม (Genetic Algorithm) กับระบบตรวจจับการบุกรุก โดยใช้ชุดข้อมูล KDD Cup 1999 ผลการวิจัยพบว่าเทคนิคที่นำเสนอมีอัตราความแม่นยำโดยภาพรวม เท่ากับ 90.04%

Li and Zhao (2011) ได้เสนอเทคนิคการตรวจจับการบุกรุกที่ใช้เซตอย่างหยาบ (Rough Set) ทำการลดคุณลักษณะของชุดข้อมูล KDD Cup 1999 และใช้ซัพพอร์ตเวกเตอร์แมทชีน (Support Vector Machine) เพื่อตรวจจับการบุกรุก ผลการวิจัยพบว่าเทคนิคที่นำเสนอมีอัตราความแม่นยำโดยรวม (90%) สูงกว่าเทคนิคซัพพอร์ตเวกเตอร์แมทชีน (86.1%)

Kandeeban (2011) ได้นำเอาขั้นตอนวิธีทางพันธุกรรมและโครงข่ายประสาทเทียมไปใช้ในการตรวจจับการบุกรุก เพื่อลดคุณลักษณะของข้อมูลและสร้างกฎที่ใช้ตรวจจับการบุกรุก ตามลำดับ

อนึ่งจากงานวิจัยข้างต้น สามารถสรุปสาระสำคัญได้ดังตารางที่ 2.2

ตารางที่ 2.2 สาระสำคัญของงานวิจัยที่เกี่ยวกับระบบตรวจจับการบุกรุก

ผู้วิจัย	จุดเด่น	จุดด้อย	เทคนิค/เป้าหมาย
Narayan, Singh and Tak (2012)	เป็นการทดสอบประสิทธิภาพของเทคนิคการจัดกลุ่ม	ข้อมูลที่ใช้น้อย (2,000 จาก 311,029 ระเบียน)	ฟัซซี่ซีมีนส์ที่มีค่าคาดหมายมากที่สุด มีอัตราแจ้งเตือนความผิดพลาดน้อยลง
Hoque, Mukit and Bikas (2012)	แบ่งชนิดการบุกรุกเป็น 5 กลุ่ม	ใช้เทคนิคเดียวและไม่มีการเปรียบเทียบกับงานวิจัยอื่น	เทคนิคพันธุกรรม มีอัตราความแม่นยำโดยรวม 90.04%
Li and Zhao (2011)	ใช้ 2 เทคนิครวมกันเพื่อปรับปรุงประสิทธิภาพการตรวจจับการบุกรุก	มีการลดคุณลักษณะของข้อมูล	เทคนิคเซตอย่างหยาบและซัพพอร์ตเวกเตอร์แมทชีน มีอัตราความแม่นยำโดยรวม (90%) สูงกว่าเทคนิคซัพพอร์ตเวกเตอร์แมทชีน (86.1%)
Kandeeban (2011)	ใช้ 2 เทคนิครวมกันเพื่อปรับปรุงประสิทธิภาพการตรวจจับการบุกรุก	มีการลดคุณลักษณะของข้อมูล และใช้ข้อมูลน้อย (50,000 จาก 311,021 ระเบียน)	เทคนิคทางพันธุกรรมและโครงข่ายประสาทเทียม

บทที่ 3

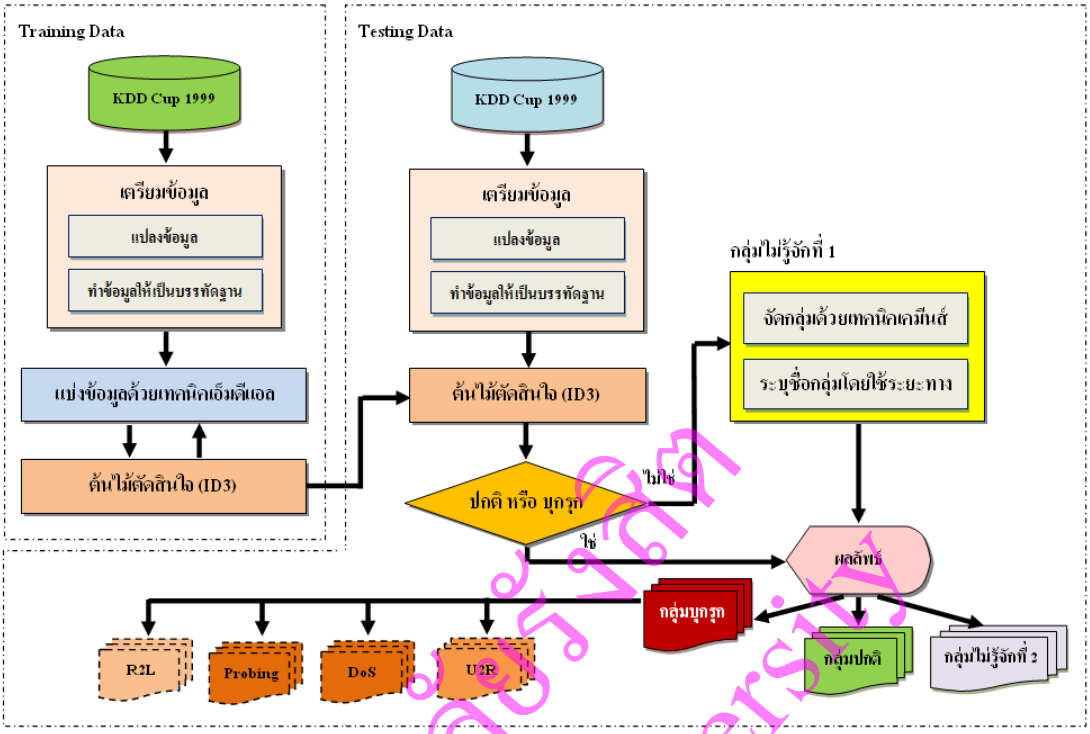
วิธีดำเนินงานวิจัย

ในบทนี้จะนำเสนอวิธีการดำเนินงานวิจัยของการตรวจจับการบุกรุกโดยการรวมกันของเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและเทคนิคการตรวจจับความผิดปกติ

3.1 กรอบการวิจัย (Research Framework)

จากการศึกษางานวิจัยด้านระบบตรวจจับการบุกรุกและของ Surasit Songma et al. (2012a, 2012b, 2013) ที่ใช้เทคนิคการตรวจจับความผิดปกติและใช้เทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบ ทำให้ทราบถึงจุดเด่นและจุดด้อยของแต่ละเทคนิค โดยพบว่าเทคนิคการตรวจจับความผิดปกติมีจุดเด่นที่สามารถตรวจจับการบุกรุกชนิดใหม่ได้โดยไม่จำเป็นต้องมีความรู้เกี่ยวกับชนิดการบุกรุกนั้น แต่มีข้อจำกัดเรื่องมีอัตราการแจ้งเตือนผิดพลาดที่สูง เนื่องจากพฤติกรรมของผู้ใช้ระบบไม่สามารถคาดการณ์ได้ ซึ่งบางครั้งพบว่ามีความเป็นไปได้ที่จะเป็นการกระทำที่ผิดปกติแต่ไม่ใช่การบุกรุก ส่วนเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบ มีจุดเด่นในด้านการมีอัตราการตรวจจับการบุกรุกที่สูง และมีอัตราการแจ้งเตือนผิดพลาดที่ต่ำ แต่มีข้อจำกัดว่าสามารถตรวจจับการบุกรุกได้ดีเฉพาะชนิดการบุกรุกที่รู้จักเท่านั้น จึงจำเป็นต้องมีการเก็บรูปแบบการบุกรุกไว้ในฐานข้อมูลเป็นจำนวนมาก ซึ่งเป็นเรื่องที่ทำได้ยากและส่งผลให้วิธีการนี้ไม่สามารถตรวจจับการบุกรุกชนิดใหม่ได้

ผู้วิจัยจึงได้มีแนวคิดในการนำเอาทั้งสองเทคนิคมาทำงานร่วมกันเพื่อใช้ประโยชน์จากทั้งจุดเด่นด้านการมีอัตราการตรวจจับการบุกรุกที่สูงและการมีอัตราการการแจ้งเตือนผิดพลาดที่ต่ำของเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและจุดเด่นของเทคนิคการตรวจจับความผิดปกติที่สามารถตรวจจับการบุกรุกชนิดใหม่ได้ ซึ่งจะทำให้ระบบตรวจจับการบุกรุกทำงานได้อย่างมีประสิทธิภาพมากขึ้น โดยนำเสนอเป็นกรอบงานวิจัยได้ดังรูปที่ 3.1



รูปที่ 3.1 กรอบงานวิจัยการตรวจจับการบุกรุกโดยการรวมกันของเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและเทคนิคการตรวจจับความผิดปกติ

3.2 การเตรียมข้อมูล

จากการศึกษาค้นคว้างานวิจัยที่เกี่ยวข้องทำให้ผู้วิจัยเลือกใช้ชุดข้อมูล KDD Cup 1999 ซึ่งเป็นชุดข้อมูลมาตรฐานที่นิยมใช้ในงานวิจัยด้านระบบตรวจจับการบุกรุกและสามารถแบ่งข้อมูลออกเป็น 2 ส่วน คือ ชุดข้อมูลฝึกสอน ซึ่งใช้เวลาในการเก็บข้อมูล 7 สัปดาห์ และประกอบด้วยข้อมูลประมาณ 5 ล้านระเบียน ดังตารางที่ 3.1 และชุดข้อมูลทดสอบ ซึ่งใช้เวลาในการเก็บข้อมูล 2 สัปดาห์ และประกอบด้วยข้อมูลประมาณ 2 ล้านระเบียน การติดต่อสื่อสารข้อมูลจะอยู่ในลักษณะทีซีพีแพ็คเก็ตจากต้นทางไปยังปลายทาง ซึ่งมีการกำหนดโปรโตคอลที่ใช้ในการสื่อสาร หมายเลขไอพีแอดเดรสต้นทาง และหมายเลขแอดเดรสปลายทางไว้อย่างชัดเจน แต่ละระเบียนจะมีขนาดประมาณ 100 ไบต์และมีป้ายกำกับไว้ด้วยว่าเป็นระเบียนที่ปกติหรือระเบียนที่เป็นการบุกรุก อนึ่งระเบียนของการบุกรุกสามารถแบ่งออกเป็น 4 กลุ่มดังนี้

1) การโจมตีแบบปฏิเสธการให้บริการ (DoS) เป็นการโจมตีที่เน้นไปที่การทำให้คอมพิวเตอร์ไม่สามารถติดต่อสื่อสารหรือใช้งานได้ตามปกติ หรือทำให้ระบบคอมพิวเตอร์ใช้งานไม่ได้ โดยการเข้าถึงระบบคอมพิวเตอร์และใช้ทรัพยากรที่มีอยู่อย่างจำกัดทำให้ไม่สามารถให้บริการได้ตามปกติ เช่น การโจมตีแบบ Apache2, Back, Land, Neptune, Pod, Smurf และ Teardrop

2) การโจมตีแบบเข้าถึงระบบโดยไม่ได้รับอนุญาต (R2L) เป็นการพยายามเข้าถึงระบบคอมพิวเตอร์เป้าหมายที่ไม่ได้รับอนุญาตในการเข้าถึง โดยวิธีการส่งแพ็คเก็ตไปยังเครื่องเป้าหมายด้วยโปรแกรมเครือข่าย เช่น Imap, Named และ Sendmail หรือการพยายามเข้าใช้เครื่องเป้าหมายที่มีการตั้งค่าการรักษาความปลอดภัยไว้ไม่ดี เช่น การโจมตีแบบ Guess_passwod, Ftp_write และ Xsnoop หรือการใช้ม้าโทรจันเพื่อดักรับรหัสผ่านของผู้อื่น

3) การโจมตีแบบพยายามทำตัวเป็นผู้มีสิทธิ์ (U2R) เป็นการพยายามใช้งานระบบคอมพิวเตอร์ในสิ่งที่ไม่ได้รับอนุญาตให้เข้าถึงสิทธิ์นั้น กล่าวคือพยายามยกระดับสิทธิ์การเข้าถึงระบบให้เป็นรูทหรือซุปเปอร์ยุสเซอร์ เพื่อที่จะสามารถทำอะไรก็ได้ในระบบดังกล่าว เช่น Buffer_Overflow, Loadmodule, Perl และ Rootkit

4) การโจมตีแบบแสกนเป้าหมาย (Probing) เป็นการโจมตีที่ใช้วิธีการตรวจหาและเก็บข้อมูลช่องโหว่ของระบบคอมพิวเตอร์เป้าหมาย เมื่อผู้บุกรุกทราบก็จะอาศัยช่องโหว่ดังกล่าวเข้าบุกรุกหรือโจมตีในครั้งต่อไป เช่น Ipsweep, Nmap, Portsweep และ Satan

ตารางที่ 3.1 รายละเอียดของชุดข้อมูล KDD Cup 1999

ประเภทข้อมูล	จำนวน	Normal	DoS	R2L	U2R	Probing
10% KDD	494,021	97,278	391,458	1,126	52	4,107
	(100 %)	(19.69 %)	(79.24 %)	(0.23 %)	(0.01 %)	(0.83 %)
Corrected	311,029	60,593	229,853	16,189	228	4,166
	(100 %)	(19.48 %)	(73.90 %)	(5.20 %)	(0.07 %)	(1.34 %)
Whole KDD	4,898,430	972,780	3,883,370	1,126	52	41,102
	(100 %)	(19.8 %)	(79.3 %)	(0.02 %)	(0.01 %)	(0.84 %)

ผู้วิจัยได้ทำการศึกษาชุดข้อมูล KDD Cup 1999 และพบว่าในแต่ละระเบียนประกอบด้วยคุณลักษณะที่เป็นเฉลยว่าระเบียนดังกล่าวเป็นปกติหรือการบุกรุก จำนวน 1 ตัว และคุณลักษณะอื่นๆ อีกจำนวน 41 ตัว ซึ่งแต่ละตัวจะมีความหมายแตกต่างกันออกไป เช่น ระยะเวลาในการเชื่อมต่อ (Duration) ชนิดของโปรโตคอล (Protocol Type) รวมไปถึงชนิดของการให้บริการ (Service) คุณลักษณะเหล่านี้มี 2 รูปแบบ คือ แบบต่อเนื่อง (Continuous) และแบบไม่ต่อเนื่อง (Discrete) และมีรายละเอียดดังตารางที่ 3.2 – 3.5

ตารางที่ 3.2 คุณลักษณะพื้นฐานของการเชื่อมต่อทีซีพี

ที่	ชื่อคุณลักษณะ	รายละเอียด	รูปแบบ
1	duration	Length (number of seconds) of the connection	ต่อเนื่อง
2	protocol type	Type of the protocol, e.g., tcp, udp, etc.	ไม่ต่อเนื่อง
3	service	Network service on the destination, e.g., http, telnet, etc.	ไม่ต่อเนื่อง
4	flag	Normal or error status of the connection	ไม่ต่อเนื่อง
5	src_bytes	Number of data bytes from source to destination	ต่อเนื่อง
6	dst_bytes	Number of data bytes from destination to source	ต่อเนื่อง
7	land	1 if connection is from/to the same host/port; 0 otherwise	ไม่ต่อเนื่อง
8	wrong_fragment	Number of "wrong" fragments	ต่อเนื่อง
9	urgent	Number of urgent packets	ต่อเนื่อง
10	hot	Number of "hot" indicators	ต่อเนื่อง
11	num_failed_logins	Number of failed login attempts	ต่อเนื่อง

ตารางที่ 3.3 คุณลักษณะเนื้อหาภายในการเชื่อมต่อที่แนะนำโดยความรู้โดเมน

ที่	ชื่อคุณลักษณะ	รายละเอียด	รูปแบบ
12	logged_in	1 if successfully logged in; 0 otherwise	ไม่ต่อเนื่อง
13	num_compromised	Number of "compromised" conditions	ต่อเนื่อง
14	root_shell	1 if root shell is obtained; 0 otherwise	ต่อเนื่อง
15	su_attempted	1 if "su root" command attempted; 0 otherwise	ต่อเนื่อง
16	num_root	Number of "root" accesses	ต่อเนื่อง
17	num_file_creations	Number of file creation operations	ต่อเนื่อง
18	num_shells	Number of shell prompts	ต่อเนื่อง
19	num_access_files	Number of operations on access control files	ต่อเนื่อง

ตารางที่ 3.3 คุณลักษณะเนื้อหาภายในการเชื่อมต่อที่แนะนำโดยความรู้โดเมน (ต่อ)

ที่	ชื่อคุณลักษณะ	รายละเอียด	รูปแบบ
20	num_otbound_cmds	Number of outbound commands in an ftp session	ต่อเนื่อง
21	is_host_login	1 if the login belongs to the "hot" list; 0 otherwise	ไม่ต่อเนื่อง
22	is_guest_login	1 if the login is a "guest" login; 0 otherwise	ไม่ต่อเนื่อง

ตารางที่ 3.4 คุณลักษณะทราฟฟิกที่คำนวณโดยใช้หน้าต่างเวลาสองวินาที

ที่	ชื่อคุณลักษณะ	รายละเอียด	รูปแบบ
23	count	Number of connections to the same host as the current connection in the past two seconds	ต่อเนื่อง
24	srv_count	Number of connections to the same service as the current connection in the past two seconds	ต่อเนื่อง
25	serror_rate	% of connections that have "SYN" errors	ต่อเนื่อง
26	srv_serror_rate	% of connections that have "SYN" errors	ต่อเนื่อง
27	rerror_rate	% of connections that have "REJ" errors	ต่อเนื่อง
28	srv_rerror_rate	% of connections that have "REJ" errors	ต่อเนื่อง
29	same_srv_rate	% of connections to the same service	ต่อเนื่อง
30	diff_srv_rate	% of connections to different services	ต่อเนื่อง
31	srv_diff_host_rate	% of connections to different hosts	ต่อเนื่อง

ตารางที่ 3.5 คุณลักษณะทราฟฟิกที่คำนวณโดยใช้หน้าต่างเวลาสองวินาทีจากปลายทางสู่ต้นทาง

ที่	ชื่อคุณลักษณะ	รายละเอียด	รูปแบบ
32	dst_host_count	Count for destination host	ต่อเนื่อง
33	dst_host_srv_count	srv_count for destination host	ต่อเนื่อง
34	dst_host_same_srv_rate	same_srv_rate for destination host	ต่อเนื่อง
35	dst_host_diff_srv_rate	dif_srv_rate for destination host	ต่อเนื่อง
36	dst_host_same_srv_port_rate	same_src_port_rate for destination host	ต่อเนื่อง
37	dst_host_srv_diff_host_rate	diff_host_rate for destination host	ต่อเนื่อง
38	dst_host_serror_rate	serror_rate for destination host	ต่อเนื่อง
39	dst_host_srv_serror_rate	srv_serror_rate for destination host	ต่อเนื่อง
40	dst_host_rerror_rate	rerror_rate for destination host	ต่อเนื่อง
41	dst_host_srv_rerror_rate	srv_serror_rate for destination host	ต่อเนื่อง

ชุดข้อมูล KDD Cup 1999 ที่นำมาใช้ในงานวิจัยนี้สามารถดาวน์โหลดได้จากเว็บไซต์ http://kdd.ics.uci.edu/databases/kddcup99/ ข้อมูลภายในระเบียนมีลักษณะเป็นตัวเลขผสมกับตัวอักษรและขั้นกลางด้วยเครื่องหมายมหัพภาค ซึ่งยกตัวอย่างได้ดังนี้

- ชุดข้อมูลที่ไม่มีการโจมตี (Normal) ประกอบด้วย 0, udp, private, SF, 105, 146, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 255, 254, 1, 0.01, 0, 0, 0, 0, 0, 0, normal

- การโจมตีแบบ DoS ประกอบด้วย 0, tcp, http, RSTR, 54540, 8314, 0, 0, 0, 2, 0, 1, 1, 0, 0 , 0, 0, 0, 0, 0, 0, 0, 0, 1, 8, 0, 0, 1, 0.12, 1, 0, 0.25, 255, 229, 0.9, 0.01, 0, 0, 0, 0, 0.01, 0, back

- การโจมตีแบบ R2L ประกอบด้วย 0, udp, private, SF, 105, 146, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 0, 0, 0, 0, 1, 0, 0, 255, 254, 1, 0.01, 0, 0, 0, 0, 0, 0, snmpgetattack

- การโจมตีแบบ Probing ประกอบด้วย 0,tcp,private,SH,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0, 1,1,1,1,0,0,1,0,0,4,1,0.25,1,1,0,1,1,0,0,nmap

- การโจมตีแบบ U2R ประกอบด้วย 154, tcp, telnet, SF, 1510, 3349, 0, 0, 0, 3, 0, 1, 2, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 255, 35, 0.14, 0.03, 0, 0, 0, 0, 0, 0, buffer_overflow

จากการวิเคราะห์ข้อมูลดังกล่าวโดยละเอียด ผู้วิจัยพบว่ามีระเบียนที่มีความซ้ำซ้อนกันอยู่จำนวนมาก จึงได้ทำการลดความซ้ำซ้อนของระเบียนออกก่อน เพื่อให้สะดวกต่อการดำเนินงานในขั้นตอนต่อไป ผู้วิจัยใช้หลักความน่าจะเป็นในการสุ่ม โดยสุ่มตัวอย่างแบบแบ่งชั้นภูมิ (Stratified Sampling) ผลการลดความซ้ำซ้อนของระเบียนในชุดข้อมูลโดยแยกตามชนิดการบุกรุกแสดงได้ดังตารางที่ 3.6 และ 3.7

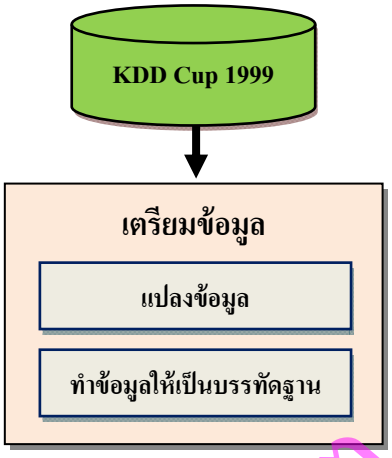
ตารางที่ 3.6 อัตราส่วนของระเบียนข้อมูล 10% KDD ที่จำแนกตามชนิดการบุกรุก

ที่	ชนิดการบุกรุก	ระเบียนเดิม		ระเบียนที่ลดความซ้ำซ้อนแล้ว	
		จำนวน	อัตราส่วน (%)	จำนวน	อัตราส่วน (%)
1	Normal	97,278	19.69	87,832	60.33
2	Denial of Service (DoS)	391,458	79.24	54,572	37.48
3	Remote to Local (R2L)	1,126	0.23	999	0.69
4	Probing	4,107	0.83	2,131	1.46
5	User to Root (U2R)	52	0.01	52	0.04
	รวมทั้งสิ้น	494,021	100	145,586	100

ตารางที่ 3.7 อัตราส่วนของระเบียนข้อมูล Corrected ที่จำแนกตามชนิดการบุกรุก

ที่	ชนิดการบุกรุก	ระเบียนเดิม		ระเบียนที่ลดความซ้ำซ้อนแล้ว	
		จำนวน	อัตราส่วน (%)	จำนวน	อัตราส่วน (%)
1	Normal	60,593	19.48	47,913	61.99
2	Denial of Service (DoS)	229,853	73.90	23,568	30.49
3	Remote to Local (R2L)	16,189	5.20	2,913	3.77
4	Probing	4,166	1.34	2,682	3.47
5	User to Root (U2R)	228	0.07	215	0.28
	รวมทั้งสิ้น	311,029	100	77,291	100

ผู้วิจัยแบ่งการเตรียมข้อมูลออกเป็น 2 ขั้นตอน คือ การแปลงข้อมูลชุดข้อมูล KDD Cup 1999 ที่เป็นตัวอักษรให้อยู่ในรูปของตัวเลข และการทำข้อมูลให้เป็นบรรทัดฐานโดยใช้ค่าต่ำสุดและค่าสูงสุด ดังรูปที่ 3.2 ซึ่งจะอธิบายรายละเอียดของแต่ละขั้นตอนในลำดับถัดไป



รูปที่ 3.2 กระบวนการเตรียมข้อมูล

3.3 การแปลงข้อมูล

ขั้นตอนนี้เป็นการปรับค่าคุณลักษณะของชุดข้อมูล KDD Cup 1999 ที่เป็นตัวอักษรให้อยู่ในรูปของตัวเลข จากการวิเคราะห์พบว่ามี 3 คุณลักษณะที่เป็นตัวอักษร ได้แก่

1) คุณลักษณะ Protocol Type จำนวน 3 ตัว คือ icmp, tcp และ udp

2) คุณลักษณะ Service จำนวน 65 ตัว เช่น http, telnet และ pop_2

3) คุณลักษณะ Flag จำนวน 11 ตัว เช่น REJ, OTH และ RSTO

คุณลักษณะเหล่านี้จะถูกแทนที่เป็นตัวเลขตั้งแต่ 1 ถึง N เมื่อ N คือจำนวนของตัวอักษรทั้งหมดที่มีในคุณลักษณะนั้น (Kandeeban and Rajesh, 2009) เช่น การแทนค่าคุณลักษณะของ Protocol Type สามารถกระทำได้ดังนี้ คือ ตัวเลข 1 แทนโปรโตคอล icmp ตัวเลข 2 แทนโปรโตคอล tcp และตัวเลข 3 แทนโปรโตคอล udp โดยผลการปรับคุณลักษณะแสดงไว้ในบทที่ 4

3.4 การทำข้อมูลให้เป็นบรรทัดฐาน

เนื่องจากช่วงของข้อมูลที่จะนำเข้ามีความแตกต่างกันมาก จำเป็นต้องปรับให้อยู่ในช่วงเดียวกัน โดย Yang and Karahoca (2006) ได้กล่าวถึงวิธีการทำข้อมูลให้เป็นบรรทัดฐานว่าเป็น

วิธีการแปลงข้อมูลเชิงเส้น จากช่วงที่เป็นไปได้เดิมของค่าข้อมูลนำเข้าให้เป็นช่วงข้อมูลระหว่าง 0 ถึง 1 ซึ่งนิยมเรียกว่า Min-Max Normalization ดังสมการ

$$V' = (V - Min_{\text{old}})/(Max_{\text{old}} - Min_{\text{old}}) \quad (3.1)$$

โดยที่ V' คือ ข้อมูลใหม่

V คือ ข้อมูลเก่า

Min_{old} คือ ค่าต่ำสุดของข้อมูลเก่า

Max_{old} คือ ค่าสูงสุดของข้อมูลเก่า

3.5 ขั้นตอนการทำงานของเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบ

ผู้วิจัยเลือกเทคนิคต้นไม้ตัดสินใจมาทำการสร้างกฎ โดยเริ่มสร้างกฎขึ้นมาจากชุดข้อมูลฝึกอบรมแล้วนำกฎที่ได้ไปใช้กับชุดข้อมูลทดสอบ เพื่อคัดกรองข้อมูลที่ไม่มีอยู่ในกฎที่สร้างขึ้นหรือที่เรียกว่ากลุ่มไม่รู้จัก ซึ่งจะถูกส่งไปยังขั้นตอนของเทคนิคการตรวจจับการบุกรุกแบบผิดปกติต่อไป

ชุดข้อมูล KDD Cup 1999 ที่ใช้ในการทดลองเป็นข้อมูลที่มีลักษณะผสมกันระหว่างข้อมูลแบบไม่ต่อเนื่องและข้อมูลแบบต่อเนื่อง เทคนิคต้นไม้ตัดสินใจที่เลือกใช้ (Iterative Dichotomiser 3: ID3) นั้นมีข้อจำกัดในด้านข้อมูลที่จะต้องอยู่ในรูปแบบไม่ต่อเนื่องเท่านั้น ฉะนั้นจำเป็นต้องแปลงข้อมูลให้อยู่ในรูปแบบไม่ต่อเนื่องก่อนจึงจะสามารถนำไปใช้กับเทคนิคต้นไม้ตัดสินใจได้ ทั้งนี้ผู้วิจัยเลือกทำการแบ่งข้อมูลให้อยู่ในรูปแบบไม่ต่อเนื่องด้วยเทคนิคเอ็มดีแอล ซึ่งเป็นเทคนิคการเรียนรู้แบบมีผู้สอนที่ให้ผลการแบ่งข้อมูลที่ดีและไม่จำเป็นต้องมีการทดลองผลหลายๆ ครั้งเพื่อให้ได้ผลที่ดีที่สุด ดังรูปที่ 2.13 เมื่อได้จุดแบ่งที่เหมาะสมแล้วจะนำข้อมูลฝึกอบรมไปแบ่งและแปลงให้อยู่ในรูปแบบไม่ต่อเนื่อง เพื่อเข้าสู่การสร้างกฎด้วยเทคนิคต้นไม้ตัดสินใจ โดยผลลัพธ์ที่ได้แบ่งเป็น 6 กลุ่ม ดังนี้

1) กลุ่มพฤติกรรมปกติ
2) กลุ่มการโจมตีแบบปฏิเสธการให้บริการ
3) กลุ่มการโจมตีแบบเข้าถึงระบบโดยไม่ได้รับอนุญาต
4) กลุ่มการโจมตีแบบพยายามทำตัวเป็นผู้มีสิทธิ์

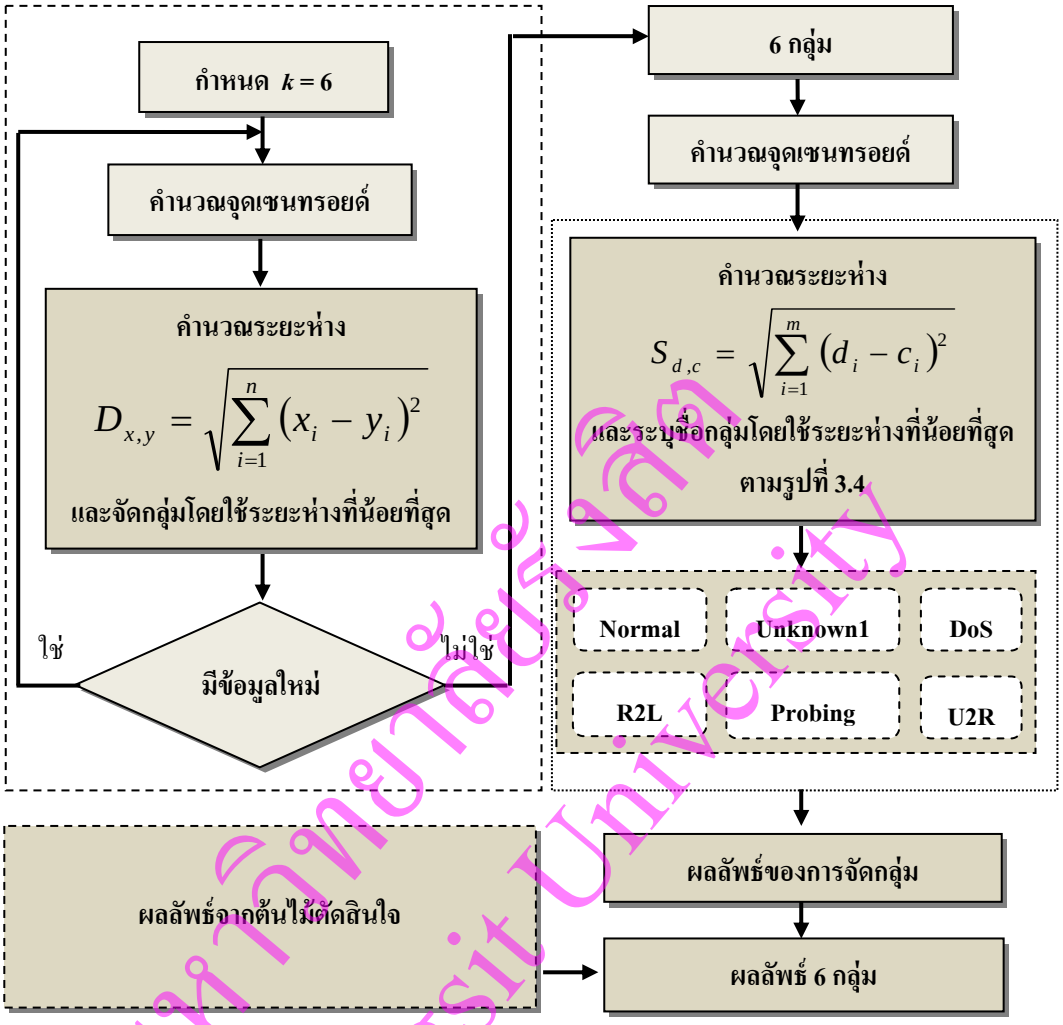
5) กลุ่มการโจมตีแบบสแกนเป้าหมาย

6) กลุ่มไม่รู้จัก

3.6 ขั้นตอนการทำงานของเทคนิคการตรวจจับความผิดปกติ

เป็นขั้นตอนที่นำผลลัพธ์ที่ได้จากการตรวจจับการใช้สิทธิ์โดยมิชอบ เฉพาะส่วนที่เป็นกลุ่มไม่รู้จักมาทำการตรวจจับด้วยเทคนิคการตรวจจับความผิดปกติ ซึ่งผู้วิจัยเลือกใช้เทคนิคการจัดกลุ่มแบบเคมีนส์ เนื่องจากเป็นเทคนิคที่ใช้ได้ผลดี และได้กำหนดผลลัพธ์ออกเป็น 6 กลุ่ม ดังรูปที่ 3.3

การจัดกลุ่มแบบเคมีนส์นั้นต้องระบุให้ได้ว่าแต่ละกลุ่มผลลัพธ์ที่ได้ คือ การโจมตีประเภทใดใน 6 ประเภทข้างต้น ผู้วิจัยจึงเลือกใช้การวัดระยะทางแบบยูคลิดระหว่างจุดเซนทรอยด์ของกลุ่มที่จัดใหม่กับจุดเซนทรอยด์ของกลุ่มที่จำแนกแล้ว (ผลจากการตรวจจับการใช้สิทธิ์โดยมิชอบ) ซึ่งรายละเอียดแสดงดังรูปที่ 3.4 ประสิทธิภาพในการตรวจจับการบุกรุกของระบบตรวจจับการบุกรุกโดยการรวมกันของเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและเทคนิคการตรวจจับความผิดปกติจะคิดจากการนำผลการจัดกลุ่มที่ถูกต้องด้วยเทคนิคการตรวจจับความผิดปกติไปรวมกับผลที่ทายถูกของเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบซึ่งรายละเอียดจะแสดงในบทที่ 4



รูปที่ 3.3 ขั้นตอนการทำงานของการตรวจจับความผิดปกติ

โดยที่

x คือ จุดเซนทรอยด์เดิม

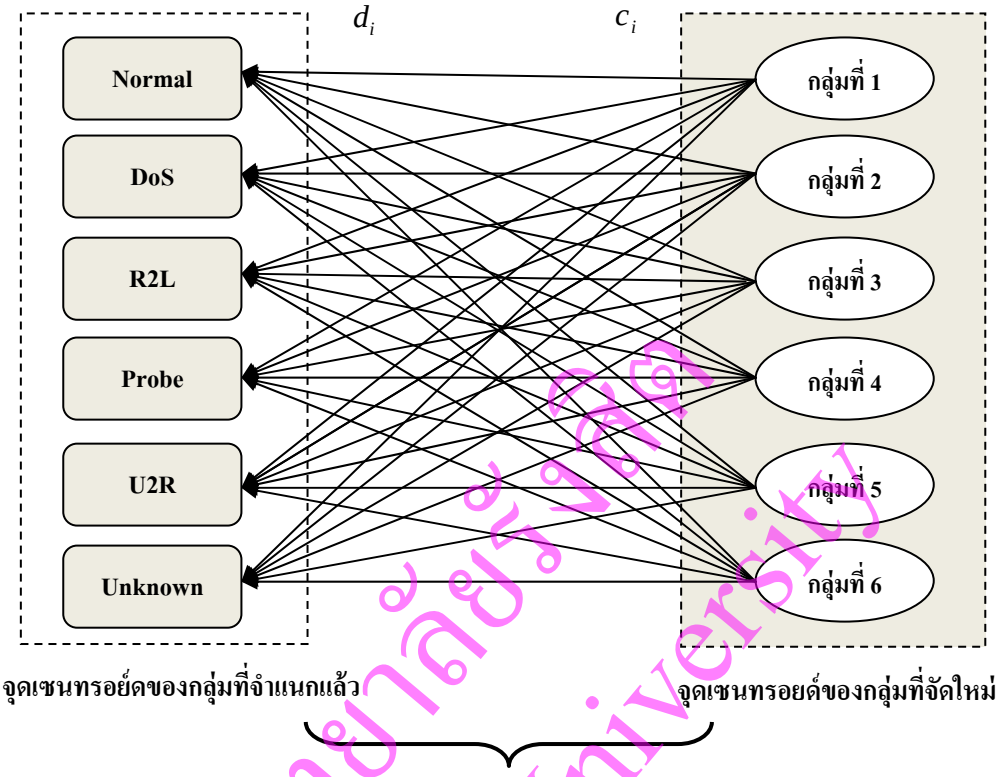
y คือ ข้อมูลที่ต้องการเปรียบเทียบ

n คือ จำนวนข้อมูลทั้งหมด

d คือ จุดเซนทรอยด์ของกลุ่มที่จำแนกแล้ว

c คือ จุดเซนทรอยด์ของกลุ่มที่จัดใหม่

m คือ จำนวนคุณลักษณะทั้งหมด



รูปที่ 3.4 ขั้นตอนการระบุชื่อกลุ่ม

3.7 การวัดประสิทธิภาพของระบบตรวจจับการบุกรุก

จากการศึกษาค้นคว้างานวิจัยที่เกี่ยวข้องกับระบบตรวจจับการบุกรุก พบว่านักวิจัยโดยส่วนใหญ่จะมีหลักการวัดประสิทธิภาพ ซึ่งสอดคล้องกับ Chimphlee (2008) ที่ได้กล่าวว่าประสิทธิภาพของระบบตรวจจับการบุกรุกนั้น พิจารณาองค์ประกอบ 3 ประการ ได้แก่ อัตราการตรวจจับการบุกรุก อัตราการแจ้งเตือนผิดพลาด และอัตราความแม่นยำโดยรวม โดยระบบตรวจจับการบุกรุกที่ดีและมีประสิทธิภาพจะต้องมีอัตราการตรวจจับการบุกรุกที่สูง มีการอัตราการแจ้งเตือนผิดพลาดที่ต่ำ และมีอัตราความแม่นยำโดยรวมที่สูง ซึ่งมีสมการสำหรับการคำนวณดังนี้

$$DTR = [TP/(TP + FN)] \times 100\% \quad (3.2)$$

$$FPR = [FP/(TN + FP)] \times 100\% \quad (3.3)$$

$$OA = [(TP + TN)/(TP + TN + FP + FN)] \times 100\% \quad (3.4)$$

โดยที่

TP คือ จำนวนของการบุกรุกที่ระบบตรวจจับได้ถูกต้องว่าเป็นการบุกกรุก

TN คือ จำนวนของผู้ใช้ปกติที่ระบบตรวจจับได้ถูกต้องว่าเป็นผู้ใช้ปกติ

FP คือ จำนวนของผู้ใช้ปกติที่ระบบตรวจจับได้ผิดพลาดว่าเป็นการบุกรุก

FN คือ จำนวนของการบุกรุกที่ระบบตรวจจับได้ผิดพลาดว่าเป็นผู้ใช้ปกติ

อนึ่งนอกจากการวัดประสิทธิภาพของระบบตรวจจับการบุกรุกตามสมการข้างต้นแล้ว เราสามารถวิเคราะห์ระบบตรวจจับการบุกรุกให้ละเอียดมากขึ้นโดยการสร้างคอนฟิวชันเมตริกซ์ (Confusion Matrix)

คอนฟิวชันเมตริกซ์ คือ ตารางที่แสดงผลการแบ่งแยกข้อมูลจริงและข้อมูลที่เกิดจากการทำนาย เช่น การตรวจสอบข้อมูลของระบบตรวจจับการบุกรุกว่าเป็นปกติหรือการบุกรุก ดังตารางที่ 3.8

ตารางที่ 3.8 ผลการตรวจจับการบุกรุกในรูปคอนฟิวชันเมตริกซ์

ค่าที่ทำนายได้ / ค่าจริง	การบุกรุก	ปกติ
การบุกรุก	*TP*	*FN*
ปกติ	*FP*	*TN*

บทที่ 4

ผลการวิจัย

การวิจัยนี้มีวัตถุประสงค์เพื่อปรับปรุงประสิทธิภาพของระบบตรวจจับการบุกรุก โดยการนำเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและเทคนิคการตรวจจับความผิดปกติมาทำงานร่วมกัน ผู้วิจัยขอนำเสนอผลการวิจัยตามลำดับ ดังนี้

4.1 ผลการเตรียมข้อมูล

ผลการเตรียมข้อมูลแบ่งออกเป็นผลการแปลงข้อมูล และผลการทำข้อมูลให้เป็นบรรทัดฐาน

4.1.1 ผลการแปลงข้อมูล

จากการศึกษาชุดข้อมูล KDD Cup 1999 พบว่าคุณลักษณะ Protocol Type จำนวน 3 ค่า คุณลักษณะ Service จำนวน 67 ค่า และคุณลักษณะ Flag จำนวน 11 ค่า จำเป็นต้องมีการปรับค่าจากเดิมที่เป็นตัวอักษรให้อยู่ในรูปของตัวเลข ผลการปรับค่าคุณลักษณะแสดงได้ดังตารางที่ 4.1 – 4.3

ตารางที่ 4.1 ผลการปรับค่าคุณลักษณะ Protocol Type

ค่าที่ใช้	รายละเอียด
1	icmp
2	tcp
3	udp

ตารางที่ 4.2 ผลการปรับค่าคุณลักษณะ Service

ค่าที่ใช้	รายละเอียด
1	IRC
2	Z39_50
3	auth
4	bgp
5	courier
6	csnet_ns
7	ctf
8	daytime
9	discard
10	domain
11	domain_u
12	echo
13	eco_i
14	ecr_i
15	efs
16	exec
17	finger
18	ftp
19	ftp_data
20	gopher
21	hostnames
22	http
23	http_443
24	imap4
25	iso_tsap
26	klogin
27	kshell
28	ldap

ตารางที่ 4.2 ผลการปรับค่าคุณลักษณะ Service (ต่อ)

ค่าที่ใช้	รายละเอียด
29	link
30	login
31	mtp
32	name
33	netbios_dgm
34	netbios_ns
35	netbios_ssn
36	netstat
37	nnsp
38	nntp
39	ntp_u
40	Other
41	pop_2
42	pop_3
43	printer
44	Private
45	remote_job
46	rje
47	shell
48	smtp
49	sql_net
50	ssh
51	sunrpc
52	supdup
53	systat
54	telnet
55	time
56	urp_i

ตารางที่ 4.2 ผลการปรับค่าคุณลักษณะ Service (ต่อ)

ค่าที่ใช้	รายละเอียด
57	uucp
58	uucp_path
59	vmnet
60	whois
61	X11
62	icmp
63	pm_dump
64	tftp_u
65	tim_i
66	red_i
67	urh_i

ตารางที่ 4.3 ผลการปรับค่าคุณลักษณะ Flag

ค่าที่ใช้	รายละเอียด
1	REJ
2	RSTO
3	RSTOS0
4	RSTR
5	S0
6	S1
7	S2
8	S3
9	SF
10	SH
11	OTH

4.1.2 ผลการทำข้อมูลให้เป็นบรรทัดฐาน

จากผลการวิเคราะห์ค่าทางสถิติของชุดข้อมูลฝึกอบรมและชุดข้อมูลทดสอบ ดังตารางที่ 4.4 และ 4.5 พบว่าชุดข้อมูลทั้งสองมีช่วงข้อมูลที่มีความแตกต่างกันมาก จึงจำเป็นต้องทำข้อมูลให้เป็นบรรทัดฐาน โดยปรับให้อยู่ในช่วงค่าระหว่าง 0 ถึง1

ตารางที่ 4.4 ผลการวิเคราะห์ค่าทางสถิติของชุดข้อมูล 10% KDD ที่ใช้ในการวิจัย

ชื่อคุณลักษณะ	ค่าต่ำสุด	ค่าสูงสุด	ค่าเฉลี่ย	ส่วนเบี่ยงเบนมาตรฐาน
1. duration (ต่อเนื่อง)	0	58,329	132.025	1,224.157
2. protocol type (ไม่ต่อเนื่อง)	1	3	2.068	0.31
3. service (ไม่ต่อเนื่อง)	1	67	31.632	12.422
4. flag (ไม่ต่อเนื่อง)	1	11	6.986	2.708
5. Src_bytes (ต่อเนื่อง)	0	693,375,640	7,995.700	1,820,383.255
6. dst_bytes (ต่อเนื่อง)	0	5,155,468	2,859.780	60,809.791
7. land (ไม่ต่อเนื่อง)	0	1	0.000	0.012
8. wrong_fragment (ต่อเนื่อง)	0	3	0.020	0.239
9. urgent (ต่อเนื่อง)	0	3	0.000	0.010
10. hot (ต่อเนื่อง)	0	30	0.100	1.427
11. num_failed_logins (ต่อเนื่อง)	0	5	0.001	0.029
12. logged_in (ไม่ต่อเนื่อง)	0	1	0.491	0.500
13. lum_compromised (ต่อเนื่อง)	0	884	0.026	3.311
14. root_shell (ต่อเนื่อง)	0	1	0.000	0.019
15. su_attempted (ต่อเนื่อง)	0	2	0.000	0.014
16. num_root (ต่อเนื่อง)	0	993	0.039	3.707
17. num_file_creations (ต่อเนื่อง)	0	28	0.004	0.178
18. num_shells (ต่อเนื่อง)	0	2	0.000	0.020
19. num_access_files (ต่อเนื่อง)	0	8	0.003	0.067
20. num_otbound_cmds (ต่อเนื่อง)	0	0	0	0
21. is_host_login (ไม่ต่อเนื่อง)	0	0	0	0
22. is_guest_login (ไม่ต่อเนื่อง)	0	1	0.005	0.068
23. count (ต่อเนื่อง)	0	511	74.386	100.336
24. srv_count (ต่อเนื่อง)	0	511	13.011	30.736

ตารางที่ 4.4 ผลการวิเคราะห์ค่าทางสถิติของชุดข้อมูล 10% KDD ที่ใช้ในการวิจัย (ต่อ)

ชื่อคุณลักษณะ	ค่าต่ำสุด	ค่าสูงสุด	ค่าเฉลี่ย	ส่วนเบี่ยงเบนมาตรฐาน
25. serror_rate (ต่อเนื่อง)	0	1	0.291	0.453
26. srv_serror_rate (ต่อเนื่อง)	0	1	0.292	0.454
27. rerror_rate (ต่อเนื่อง)	0	1	0.108	0.309
28. srv_rerror_rate (ต่อเนื่อง)	0	1	0.108	0.309
29. same_srv_rate (ต่อเนื่อง)	0	1	0.655	0.447
30. diff_srv_rate (ต่อเนื่อง)	0	1	0.041	0.120
31. srv_diff_host_rate (ต่อเนื่อง)	0	1	0.093	0.242
32. dst_host_count (ต่อเนื่อง)	0	255	181.468	99.098
33. dst_host_srv_count (ต่อเนื่อง)	0	255	129.935	114.715
34. dst_host_same_srv_rate (ต่อเนื่อง)	0	1	0.553	0.456
35. dst_host_diff_srv_rate (ต่อเนื่อง)	0	1	0.061	0.147
36. dst_host_same_srv_port_rate (ต่อเนื่อง)	0	1	0.093	0.241
37. dst_host_srv_diff_host_rate (ต่อเนื่อง)	0	1	0.019	0.060
38. dst_host_serror_rate (ต่อเนื่อง)	0	1	0.292	0.452
39. dst_host_srv_serror_rate (ต่อเนื่อง)	0	1	0.291	0.453
40. dst_host_rerror_rate (ต่อเนื่อง)	0	1	0.110	0.306
41. dst_host_srv_rerror_rate (ต่อเนื่อง)	0	1	0.108	0.304

ตารางที่ 4.5 ผลการวิเคราะห์ค่าทางสถิติของชุดข้อมูล Corrected ที่ใช้ในการวิจัย

ชื่อคุณลักษณะ	ค่าต่ำสุด	ค่าสูงสุด	ค่าเฉลี่ย	ส่วนเบี่ยงเบนมาตรฐาน
1. duration (ต่อเนื่อง)	0	57,715	70.762	815.287
2. protocol type (ไม่ต่อเนื่อง)	1	3	2.042	0.279
3. dervice (ไม่ต่อเนื่อง)	1	65	29.714	12.218
4. flag (ไม่ต่อเนื่อง)	1	11	7.005	3.195
5. src_bytes (ต่อเนื่อง)	0	62,825,648	3,427.815	255,452.832
6. dst_bytes (ต่อเนื่อง)	0	5,203,179	2,870.84	32,229.826
7. land (ไม่ต่อเนื่อง)	0	1	0.0001	0.011
8. wrong_fragment (ต่อเนื่อง)	0	3	0.003	0.078
9. urgent (ต่อเนื่อง)	0	3	0.0002	0.020

ตารางที่ 4.5 ผลการวิเคราะห์ค่าทางสถิติของชุดข้อมูล Corrected ที่ใช้ในการวิจัย (ต่อ)

ชื่อคุณลักษณะ	ค่าต่ำสุด	ค่าสูงสุด	ค่าเฉลี่ย	ส่วนเบี่ยงเบนมาตรฐาน
10. hot (ต่อเนื่อง)	0	101	0.040	0.594
11. num_failed_logins (ต่อเนื่อง)	0	4	0.007	0.084
12. logged_in (ไม่ต่อเนื่อง)	0	1	0.585	0.493
13. num_compromised (ต่อเนื่อง)	0	796	0.036	3.927
14. root_shell (ต่อเนื่อง)	0	1	0.001	0.028
15. su_attempted (ต่อเนื่อง)	0	2	0.00009	0.012
16. num_root (ต่อเนื่อง)	0	878	0.034	4.343
17. num_file_creations (ต่อเนื่อง)	0	100	0.004	0.387
18. num_shells (ต่อเนื่อง)	0	5	0.0003	0.026
19. num_access_files (ต่อเนื่อง)	0	4	0.003	0.059
20. num_otbound_cmds (ต่อเนื่อง)	0	0	0.000	0.000
21. is_host_login (ไม่ต่อเนื่อง)	0	1	0.0002	0.012
22. is_guest_login (ไม่ต่อเนื่อง)	0	1	0.010	0.097
23. count (ต่อเนื่อง)	0	511	68.089	105.097
24. srv_count (ต่อเนื่อง)	0	511	18.336	53.464
25. serror_rate (ต่อเนื่อง)	0	1	0.097	0.290
26. srv_serror_rate (ต่อเนื่อง)	0	1	0.098	0.293
27. rerror_rate (ต่อเนื่อง)	0	1	0.209	0.402
28. srv_rerror_rate (ต่อเนื่อง)	0	1	0.209	0.403
29. same_srv_rate (ต่อเนื่อง)	0	1	0.730	0.423
30. diff_srv_rate (ต่อเนื่อง)	0	1	0.047	0.161
31. srv_diff_host_rate (ต่อเนื่อง)	0	1	0.097	0.230
32. dst_host_count (ต่อเนื่อง)	0	255	178.380	99.731
33. dst_host_srv_count (ต่อเนื่อง)	0	255	158.263	112.549
34. dst_host_same_srv_rate (ต่อเนื่อง)	0	1	0.661	0.438
35. dst_host_diff_srv_rate (ต่อเนื่อง)	0	1	0.046	0.134
36. dst_host_same_srv_port_rate (ต่อเนื่อง)	0	1	0.077	0.221
37. dst_host_srv_diff_host_rate (ต่อเนื่อง)	0	1	0.016	0.053
38. dst_host_serror_rate (ต่อเนื่อง)	0	1	0.095	0.282

ตารางที่ 4.5 ผลการวิเคราะห์ค่าทางสถิติของชุดข้อมูล Corrected ที่ใช้ในการวิจัย (ต่อ)

ชื่อคุณลักษณะ	ค่าต่ำสุด	ค่าสูงสุด	ค่าเฉลี่ย	ส่วนเบี่ยงเบนมาตรฐาน
39. dst_host_srv_serror_rate (ต่อเนื่อง)	0	1	0.096	0.288
40. dst_host_rerror_rate (ต่อเนื่อง)	0	1	0.208	0.392
41. dst_host_srv_rerror_rate (ต่อเนื่อง)	0	1	0.206	0.397

4.2 ผลการแบ่งข้อมูลให้อยู่ในรูปแบบไม่ต่อเนื่องด้วยเทคนิคเอ็มดีแอล

เนื่องจากชุดข้อมูลฝึกอบรมมีข้อมูลเป็นจำนวนมาก ผู้วิจัยจึงขอนำเสนอผลการแบ่งข้อมูลทั้งหมดด้วยเทคนิคเอ็มดีแอลไว้ในภาคผนวก ในที่นี้จะขอยกตัวอย่างผลการแบ่งข้อมูลเฉพาะคุณลักษณะที่ 1 ซึ่งมีค่าต่ำสุดคือ 0 และค่าสูงสุดคือ 58,329 ดังตารางที่ 4.6 โดยสามารถอธิบายประกอบได้ว่า ถ้าข้อมูลมีค่าอยู่ระหว่าง 0 - 0.99 จะถูกปรับเปลี่ยนให้เป็นข้อมูลไม่ต่อเนื่องกลุ่มที่ I1 หรือถ้าข้อมูลมีค่าอยู่ระหว่าง 14 - 30.99 จะถูกปรับเปลี่ยนให้เป็นข้อมูลไม่ต่อเนื่องกลุ่มที่ I7 เป็นต้น

ตารางที่ 4.6 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 1 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.99	I1
2	1 - 1.99	I2
3	2 - 3.99	I3
4	4 - 4.99	I4
5	5 - 11.99	I5
6	12 - 13.99	I6
7	14 - 30.99	I7
8	31 - 197.99	I8
9	198 - 1,030.99	I9
10	1,031 - 4,989.99	I10
11	4,990 - 5,024.99	I11
12	5,025 - 5,063.99	I12
13	5,064 - 5,084.99	I13

ตารางที่ 4.6 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 1 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
14	5,085 - 10,038.99	I14
15	10,039 - 10,133.99	I15
16	10,134 - 12,704.99	I16
17	12,705 - 14,681.99	I17
18	14,682 - 15,126.99	I18
19	15,127 - 15,167.99	I19
20	15,168 - 20,939.99	I20
21	20,940 - 30,189.99	I21
22	30,190 - 58,328.99	I22
23	58,329.00	I23

4.3 ผลการสร้างกฎโดยการรวมกันของเทคนิคเอ็มดีแอลและเทคนิคต้นไม้ตัดสินใจ

นำผลการแบ่งข้อมูลที่อยู่ในรูปแบบไม่ต่อเนื่องจากเทคนิคเอ็มดีแอลมาใช้ในการสร้างกฎด้วยเทคนิคต้นไม้ตัดสินใจ ซึ่งสามารถสร้างกฎในลักษณะ IF-THEN โดยใช้ตัวเชื่อม "และ" (AND) ผู้วิจัยได้แทนชื่อคุณลักษณะที่ 1 ถึงคุณลักษณะที่ 41 ด้วย *F1 - F41* ตามลำดับ เพื่อให้สะดวกต่อการนำเสนอ

อนึ่งผู้วิจัยขอยกตัวอย่างกฎที่สร้างขึ้นเพื่อใช้ในการตรวจจับการใช้สิทธิ์โดยมิชอบ ดังตารางที่ 4.7

ตารางที่ 4.7 ตัวอย่างการสร้างกฎโดยการรวมกันของเทคนิคเอ็มดีแอลและเทคนิคต้นไม้ตัดสินใจจากชุดข้อมูลฝึกอบรม

ผลลัพธ์		กฎที่ได้
Normal	IF	*F30* = " I1" AND *F5* = "I7" AND *F3* = "I8" AND *F32* = "I11"
	THEN	*Class_type* = *"Normal"*
DoS	IF	*F30* = "I1" AND *F5* = "I7" AND *F3* = "I8"
	THEN	*Class_type* = *"DoS"*
R2L	IF	*F30* = "I1" AND *F5* = "I7" AND *F3* = "I16"
	THEN	*Class_type* = *"R2L"*
Probing	IF	*F30* = "I1" AND *F5* = "I7" AND *F3* = "I8" AND *F23* = "I3"
	THEN	*Class_type* = *"Probe"*
U2R	IF	*F30* = "I1" AND *F5* = "I7" AND *F3* = "I8" AND *F32* = "I6"
	THEN	*Class_type* = *"U2R"*
Unknown	IF	*F30* = "I1" AND *F5* = "I7" AND *F3* = "I10"
	THEN	*Class_type* = *"Unknown"*

4.4 ผลการตรวจจับการบุกรุกโดยใช้เทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบ

ผู้วิจัยได้ใช้กฎที่สร้างขึ้นในหัวข้อ 4.3 กับชุดข้อมูลทดสอบเพื่อจำแนกข้อมูลออกเป็น 6 กลุ่ม คือ กลุ่มพฤติกรรมปกติ กลุ่มการบุกรุกแบบปฏิเสธการให้บริการ กลุ่มการบุกรุกแบบการเข้าถึงระบบโดยไม่ได้รับอนุญาต กลุ่มการบุกรุกแบบแสกนเป้าหมาย กลุ่มการบุกรุกแบบพยายามทำตัวเป็นผู้มีสิทธิ์ และกลุ่มไม่รู้จัก ผลการทดลองพบว่ามีอัตราการตรวจจับการบุกรุกเท่ากับ 87.37% มี

อัตราการแจ้งเตือนผิดพลาดเท่ากับ 0.56% และมีอัตราความแม่นยำโดยรวมเท่ากับ 94.96% ทั้งนี้ได้นำผลการทดลองมาแสดงในรูปคอนฟิวชันเมตริกซ์ ยกเว้นกลุ่มข้อมูลที่ไม่รู้จักไม่นำมาแสดง ดังตารางที่ 4.8

ตารางที่ 4.8 ผลการตรวจจับการบุกรุกโดยใช้เทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบ จากการสร้างกฎโดยการรวมกันของเทคนิคเอ็มดีแอลและเทคนิคต้นไม้ตัดสินใจ ในรูปคอนฟิวชันเมตริกซ์

ค่าที่ทำนายได้ / ค่าจริง	Normal	DoS	R2L	Probing	U2R	% Correct
Normal	47,454	74	11	184	1	99.43
Dos	454	22,207	255	47	5	96.69
R2L	2,585	2	19	128	3	0.69
Probing	368	72	1	1,745	21	79.07
U2R	142	0	3	16	32	16.58
% Correct	93.04	99.34	6.57	82.31	51.61	

จากการวิเคราะห์ผลการทดลองโดยละเอียด พบว่าการตรวจจับการใช้สิทธิ์โดยมิชอบ มีความสามารถในการตรวจจับได้ถูกต้องจำนวน 71,457 ระเบียน คิดเป็น 92.45% กลุ่มไม่รู้จัก จำนวน 1,462 ระเบียน คิดเป็น 1.89% และกลุ่มที่ทายผิดจำนวน 4,372 ระเบียน คิดเป็น 5.66%

4.5 ผลการตรวจจับการบุกรุกโดยใช้เทคนิคการตรวจจับความผิดปกติ

กระบวนการตรวจจับความผิดปกติในขั้นตอนนี้จะใช้ข้อมูลในส่วนของกลุ่มที่ไม่รู้จัก จำนวน 1,462 ระเบียน ที่ได้จากเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบมาทำการจัดกลุ่มด้วยเทคนิคเคมีนส์และเทคนิคฟัซซี่ซีมีนส์ เพื่อเปรียบเทียบประสิทธิภาพการตรวจจับการบุกรุกของทั้งสองเทคนิค โดยทำการทดสอบจำนวน 10 รอบ ซึ่งแต่ละรอบกำหนดจุดเซนทรอยด์เริ่มต้นใหม่ให้เหมือนกันและแบ่งออกเป็น 6 กลุ่ม คือ กลุ่มพฤติกรรมปกติ กลุ่มการบุกรุกแบบปฏิเสธการให้บริการ กลุ่มการบุกรุกแบบการเข้าถึงระบบโดยไม่ได้รับอนุญาต กลุ่มการบุกรุกแบบแสกน

เป้าหมาย กลุ่มการบุกรุกแบบพยายามทำตัวเป็นผู้มีสิทธิ์ และกลุ่มไม่รู้จัก ผลการตรวจจับการบุกรุกแสดงได้ดังตารางที่ 4.9

ตารางที่ 4.9 ผลการตรวจจับการบุกรุกโดยใช้เทคนิคการตรวจจับความผิดปกติในรูปคอนฟิวชันเมตริกซ์ รอบที่ 1-10

ค่าที่ทำนายได้ / ค่าจริง	เคมีนส์ (รอบที่ 1)						ฟัซซี่ซีมีนส์ (รอบที่ 1)					
	Normal	DoS	R2L	Probing	U2R	% Correct	Normal	DoS	R2L	Probing	U2R	% Correct
Normal	10	10	20	4	3	21.28	2	3	70	18	95	1.06
Dos	0	5	279	0	187	1.06	296	187	44	3	66	31.38
R2L	123	4	20	0	2	13.42	22	2	122	3	26	69.71
Probe	0	17	1	259	124	64.59	1	121	0	276	63	59.87
U2R	0	1	3	0	1	20.00	1	0	5	2	14	63.64
% Correct	7.52	13.51	6.19	98.48	0.32		0.62	59.74	50.62	91.39	5.3	
ค่าที่ทำนายได้ / ค่าจริง	เคมีนส์ (รอบที่ 2)						ฟัซซี่ซีมีนส์ (รอบที่ 2)					
	Normal	DoS	R2L	Probing	U2R	% Correct	Normal	DoS	R2L	Probing	U2R	% Correct
Normal	3	17	25	2	1	6.25	3	17	53	87	3	1.84
Dos	309	5	65	0	92	1.06	299	7	49	26	187	1.23
R2L	1	4	142	0	2	95.3	23	4	95	19	2	66.43
Probe	1	157	0	121	124	30.02	1	162	0	181	124	38.68
U2R	1	2	2	0	0	0.00	1	2	2	9	0	0.00
% Correct	0.95	2.7	60.68	98.37	0.00		0.92	3.65	47.74	56.21	0.00	
ค่าที่ทำนายได้ / ค่าจริง	เคมีนส์ (รอบที่ 3)						ฟัซซี่ซีมีนส์ (รอบที่ 3)					
	Normal	DoS	R2L	Probing	U2R	% Correct	Normal	DoS	R2L	Probing	U2R	% Correct
Normal	5	0	0	14	1	25.00	5	1	7	19	2	14.71
Dos	101	80	0	5	6	41.67	89	98	308	7	0	19.52
R2L	0	2	139	4	0	95.86	0	2	23	4	1	76.67
Probe	0	0	0	279	124	69.23	0	106	1	288	19	69.57
U2R	0	0	0	2	0	0.00	0	0	1	2	0	0.00
% Correct	4.72	97.56	100	91.78	0.00		5.32	47.34	6.76	90	0.00	
ค่าที่ทำนายได้ / ค่าจริง	เคมีนส์ (รอบที่ 4)						ฟัซซี่ซีมีนส์ (รอบที่ 4)					
	Normal	DoS	R2L	Probing	U2R	% Correct	Normal	DoS	R2L	Probing	U2R	% Correct
Normal	86	16	16	3	3	69.35	11	19	3	3	3	28.21
Dos	65	5	13	0	187	1.85	198	7	287	0	0	1.42
R2L	17	4	3	0	2	11.54	2	4	22	0	2	73.33
Probe	1	21	64	254	124	54.74	121	285	1	9	0	2.16
U2R	10	2	0	0	0	0.00	0	2	1	0	2	40.00
% Correct	48.04	10.42	3.13	98.83	0.00		3.31	2.21	7.01	75	28.57	

ตารางที่ 4.9 ผลการตรวจจับการบุกรุกโดยใช้เทคนิคการตรวจจับความผิดปกติในรูปคอนฟิวชันเมตริกซ์ รอบที่ 1-10 (ต่อ)

ค่าที่ทำนายได้ / ค่าจริง	เคมีนส์ (รอบที่ 5)						ฟัซซี่ซีมีนส์ (รอบที่ 5)					
	Normal	DoS	R2L	Probing	U2R	% Correct	Normal	DoS	R2L	Probing	U2R	% Correct
Normal	3	7	10	5	9	8.82	86	11	42	34	0	49.71
Dos	9	169	4	149	74	41.73	65	93	0	8	369	17.38
R2L	1	3	80	7	20	72.07	17	3	131	4	17	76.16
Probe	71	135	44	142	36	33.18	2	117	6	339	1	72.9
U2R	1	2	1	1	6	54.55	10	0	3	2	0	0.00
% Correct	3.53	53.48	57.55	46.71	4.14		47.78	41.52	71.98	87.6	0.00	
ค่าที่ทำนายได้ / ค่าจริง	เคมีนส์ (รอบที่ 6)						ฟัซซี่ซีมีนส์ (รอบที่ 6)					
	Normal	DoS	R2L	Probing	U2R	% Correct	Normal	DoS	R2L	Probing	U2R	% Correct
Normal	86	17	3	3	0	78.9	86	17	0	3	5	77.48
Dos	65	5	400	0	0	1.06	65	9	381	0	4	1.96
R2L	19	4	2	0	0	8.00	19	4	23	0	0	50.00
Probe	3	157	125	83	46	20.05	2	162	5	137	38	39.83
U2R	10	2	0	0	0	0.00	10	2	0	0	0	0.00
% Correct	46.99	2.7	0.38	96.51	0.00		47.25	4.64	5.62	97.86	0.00	
ค่าที่ทำนายได้ / ค่าจริง	เคมีนส์ (รอบที่ 7)						ฟัซซี่ซีมีนส์ (รอบที่ 7)					
	Normal	DoS	R2L	Probing	U2R	% Correct	Normal	DoS	R2L	Probing	U2R	% Correct
Normal	15	3	67	19	64	8.93	21	1	50	19	97	11.17
Dos	65	187	64	5	18	55.16	65	187	49	7	72	49.21
R2L	3	2	133	4	16	84.18	3	2	130	4	20	81.76
Probe	0	124	10	275	65	58.02	0	121	8	286	59	60.34
U2R	2	0	17	2	0	0.00	2	0	8	2	9	42.86
% Correct	17.65	59.18	45.7	90.16	0.00		23.08	60.13	53.06	89.94	3.5	
ค่าที่ทำนายได้ / ค่าจริง	เคมีนส์ (รอบที่ 8)						ฟัซซี่ซีมีนส์ (รอบที่ 8)					
	Normal	DoS	R2L	Probing	U2R	% Correct	Normal	DoS	R2L	Probing	U2R	% Correct
Normal	42	17	60	2	3	33.87	3	5	28	17	62	2.61
Dos	80	5	3	0	187	1.82	244	9	66	0	171	1.84
R2L	4	4	16	0	2	61.54	21	3	11	1	70	10.38
Probe	1	158	64	119	124	25.54	1	42	3	253	122	60.10
U2R	10	2	0	0	0	0.00	1	2	5	0	2	20.00
% Correct	30.66	2.69	11.19	98.35	0.00		1.11	14.75	9.73	93.36	0.47	

ตารางที่ 4.9 ผลการตรวจจับการบุกรุกโดยใช้เทคนิคการตรวจจับความผิดปกติในรูปคอนฟิวชันเมตริกซ์ รอบที่ 1-10 (ต่อ)

ค่าที่ทำนายได้ / ค่าจริง	เคมีนส์ (รอบที่ 9)						ฟัซซี่ซีมีนส์ (รอบที่ 9)					
	Normal	DoS	R2L	Probing	U2R	% Correct	Normal	DoS	R2L	Probing	U2R	% Correct
Normal	86	2	1	19	1	78.9	2	1	100	19	2	1.61
Dos	65	178	214	5	9	37.79	127	72	68	7	12	25.17
R2L	19	2	17	4	0	40.48	0	2	20	4	0	76.92
Probe	23	0	1	278	124	65.26	57	67	64	279	1	59.62
U2R	9	0	1	2	0	0.00	0	1	9	2	4	25.00
% Correct	42.57	97.8	7.26	90.26	0		1.08	50.35	7.66	89.71	21.05	
ค่าที่ทำนายได้ / ค่าจริง	เคมีนส์ (รอบที่ 10)						ฟัซซี่ซีมีนส์ (รอบที่ 10)					
	Normal	DoS	R2L	Probing	U2R	% Correct	Normal	DoS	R2L	Probing	U2R	% Correct
Normal	1	3	0	19	0	4.35	1	0	0	19	71	1.10
Dos	214	187	0	5	0	46.06	0	170	229	7	155	30.30
R2L	1	2	44	3	96	30.14	111	0	36	4	9	22.50
Probe	1	124	0	276	0	68.83	0	61	1	292	55	71.39
U2R	1	0	0	2	0	0.00	0	0	0	2	15	88.24
% Correct	0.46	59.18	100	90.49	0		0.89	73.59	13.53	90.12	4.92	

หมายเหตุ กำหนดให้เลขชี้กำลังถ่วงน้ำหนัก (Weighting Exponent) สำหรับเทคนิคฟัซซี่ซีมีนส์เท่ากับ 15 เนื่องจากถ้าใช้ค่าที่น้อยกว่านี้จะทำให้ไม่สามารถกำหนดสมาชิกในกลุ่มได้ครบทุกกลุ่ม ซึ่งสอดคล้องกับผลการวิจัยของ Madhukumara and Santhiyakumari (2015)

จากตารางที่ 4.9 ทำให้ทราบถึงความสามารถในการตรวจจับการบุกรุกตามชนิดการบกรุก ทั้งนี้เพื่อเปรียบเทียบประสิทธิภาพการตรวจจับการบุกรุกจากการจัดกลุ่มด้วยเทคนิคเคมีนส์และเทคนิคฟัซซี่ซีมีนส์ ผู้วิจัยจึงแสดงอัตราการตรวจจับการบุกรุก อัตราการแจ้งเตือนที่ผิดพลาดและอัตราความแม่นยำโดยรวมที่ได้จากการใช้เทคนิคทั้งสอง จำนวน 10 รอบ โดยกำหนดจุดเซนทรอยด์เริ่มต้นให้เหมือนกันในแต่ละรอบ ดังตารางที่ 4.10

ตารางที่ 4.10 อัตราการตรวจจับการบุกรุก อัตราการแจ้งเตือนผิดพลาด และอัตราความแม่นยำโดยรวม ที่ได้จากการจัดกลุ่มด้วยเทคนิคเคมีนส์และเทคนิคฟัซซี่ซีมีนส์ จำนวน 10 รอบ

	เคมีนส์			ฟัซซี่ซีมีนส์		
	DTR (%)	FPR (%)	OA (%)	DTR (%)	FPR (%)	OA (%)
รอบที่ 1	88.01	78.72	85.09	74.48	98.94	64.91
รอบที่ 2	69.65	93.75	66.82	72.84	98.16	64.31
รอบที่ 3	86.39	75.00	84.78	90.62	85.29	88.00
รอบที่ 4	87.95	30.65	85.38	65.96	71.79	64.46
รอบที่ 5	91.41	91.18	88.57	92.08	50.29	86.69
รอบที่ 6	89.47	21.10	88.35	88.85	22.52	87.55
รอบที่ 7	92.94	91.07	80.78	93.23	88.83	80.61
รอบที่ 8	87.80	66.13	80.40	74.00	97.39	66.81
รอบที่ 9	87.80	21.10	86.89	76.88	98.39	66.74
รอบที่ 10	77.30	95.65	75.59	90.32	98.90	83.76

จากตารางที่ 4.10 จะเห็นได้ว่าการจัดกลุ่มด้วยเทคนิคเคมีนส์มีอัตราความแม่นยำโดยรวมสูงกว่าการจัดกลุ่มด้วยเทคนิคฟัซซี่ซีมีนส์ จำนวน 8 รอบใน 10 รอบ ซึ่งคิดเป็น 80% ดังนั้นเทคนิคเคมีนส์น่าจะมีประสิทธิภาพการตรวจจับการบุกรุกที่สูงกว่า ผู้วิจัยจึงนำเทคนิคดังกล่าวไปใช้งานร่วมกับเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบสำหรับระบบตรวจจับการบุกรุก และจะเรียกว่าการตรวจจับการบุกรุกโดยการรวมกันของเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและเทคนิคการตรวจจับความผิดปกติ โดยแสดงผลลัพธ์ในลำดับถัดไป

4.6 ผลการตรวจจับการบุกรุกโดยการรวมกันของเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและเทคนิคการตรวจจับความผิดปกติ

การตรวจจับการบุกรุกนี้เป็นการทำงานร่วมกันระหว่างเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบที่ใช้กฎที่สร้างขึ้นโดยการรวมกันของเทคนิคเอ็มดีแอลกับเทคนิคต้นไม้ตัดสินใจ และเทคนิคการตรวจจับความผิดปกติที่ใช้การจัดกลุ่มด้วยเทคนิคเคมีนส์ โดยแบ่งออกเป็น 6 กลุ่ม ได้แก่ กลุ่มพฤติกรรมปกติ กลุ่มการโจมตีแบบปฏิเสธการให้บริการ กลุ่มการโจมตีแบบการเข้าถึงระบบโดยไม่ได้รับอนุญาต กลุ่มการโจมตีแบบสแกนเป้าหมาย กลุ่มการโจมตีแบบพยายามทำตัวเป็นผู้มีสิทธิ์

และกลุ่มไม่รู้จัก และทำการทดสอบจำนวน 10 รอบ ผลการตรวจจับการบุกรุกในรูปคอนฟิวชันเมตริกซ์ ยกเว้นส่วนข้อมูลที่ไม่รู้จักแสดงได้ดังตารางที่ 4.11

ตารางที่ 4.11 ผลการตรวจจับการบุกรุกโดยการรวมกันของเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและเทคนิคการตรวจจับความผิดปกติ ในรูปคอนฟิวชันเมตริกซ์ รอบที่ 1-10

ค่าที่ทำนายได้ / ค่าจริง	(รอบที่ 1)					
	Normal	DoS	R2L	Probing	U2R	% Correct
Normal	47,454	89	36	188	4	99.34
Dos	454	22,212	534	47	192	94.77
R2L	2,708	6	39	128	5	1.35
Probing	368	89	2	2,004	145	76.84
U2R	142	2	6	16	32	16.16
% Correct	92.82	99.17	6.32	84.1	8.47	
ค่าที่ทำนายได้ / ค่าจริง	(รอบที่ 2)					
	Normal	DoS	R2L	Probing	U2R	% Correct
Normal	47,457	91	36	186	2	99.34
Dos	763	22,212	320	47	97	94.77
R2L	2,586	6	161	128	5	5.58
Probing	369	229	1	1,866	145	71.49
U2R	143	2	5	16	32	16.16
% Correct	92.48	98.54	30.78	83.19	11.39	
ค่าที่ทำนายได้ / ค่าจริง	(รอบที่ 3)					
	Normal	DoS	R2L	Probing	U2R	% Correct
Normal	47,454	74	11	203	2	99.39
Dos	555	22,287	255	52	11	96.23
R2L	2,585	4	158	132	3	5.48
Probing	368	72	1	2,024	145	77.55
U2R	142	0	3	18	32	16.41
% Correct	92.86	99.33	36.92	83.33	16.58	

ตารางที่ 4.11 ผลการตรวจจับการบุกรุกโดยการรวมกันของเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบ และเทคนิคการตรวจจับความผิดปกติ ในรูปคอนฟิวชันเมตริกซ์ รอบที่ 1-10 (ต่อ)

ค่าที่ทำนายได้ / ค่าจริง (รอบที่ 4)	Normal	DoS	R2L	Probing	U2R	% Correct
Normal	47,540	90	27	187	4	99.36
Dos	519	22,212	268	47	192	95.58
R2L	2,602	6	22	128	5	0.8
Probing	369	93	65	1,999	145	74.84
U2R	152	2	3	16	32	15.61
% Correct	92.88	99.15	5.71	84.1	8.47	

ค่าที่ทำนายได้ / ค่าจริง (รอบที่ 5)	Normal	DoS	R2L	Probing	U2R	% Correct
Normal	47,457	81	21	189	10	99.37
Dos	463	22,376	259	196	79	95.73
R2L	2,586	5	99	135	23	3.48
Probing	439	207	45	1,887	57	71.61
U2R	143	2	4	17	38	18.63
% Correct	92.89	98.7	23.13	77.85	18.36	

ค่าที่ทำนายได้ / ค่าจริง (รอบที่ 6)	Normal	DoS	R2L	Probing	U2R	% Correct
Normal	47,540	91	14	187	1	99.39
Dos	519	22,212	655	47	5	94.77
R2L	2,604	6	21	128	3	0.76
Probing	371	229	126	1,828	67	69.74
U2R	152	2	3	16	32	15.61
% Correct	92.88	98.54	2.56	82.86	29.63	

ตารางที่ 4.11 ผลการตรวจจับการบุกรุกโดยการรวมกันของเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบ และเทคนิคการตรวจจับความผิดปกติ ในรูปคอนฟิวชันเมตริกซ์ รอบที่ 1-10 (ต่อ)

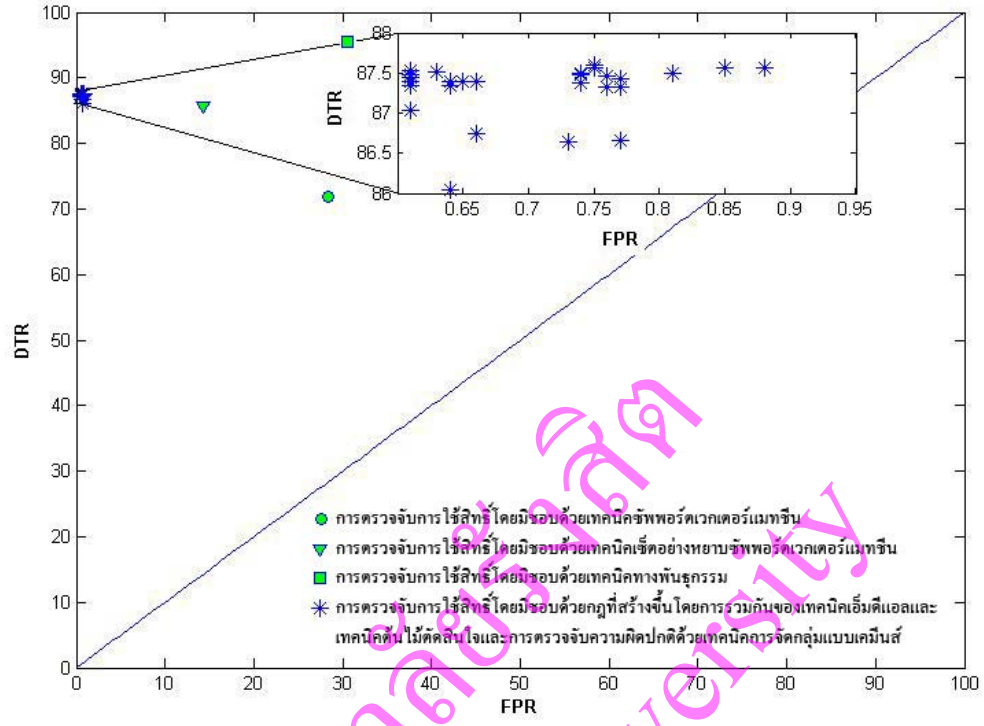
ค่าที่ทำนายได้ / ค่าจริง	(รอบที่ 7)					
	Normal	DoS	R2L	Probing	U2R	% Correct
Normal	47,469	77	78	203	65	99.12
Dos	519	22,394	319	52	23	96.08
R2L	2,588	4	152	132	19	5.25
Probing	368	196	11	2,020	86	75.35
U2R	144	0	20	18	32	14.95
% Correct	92.92	98.78	26.21	83.3	14.22	
ค่าที่ทำนายได้ / ค่าจริง	(รอบที่ 8)					
	Normal	DoS	R2L	Probing	U2R	% Correct
Normal	47,496	91	71	186	4	99.26
Dos	534	22,212	258	47	192	95.56
R2L	2,589	6	35	128	5	1.27
Probing	369	230	65	1,864	145	69.73
U2R	152	2	3	16	32	15.61
% Correct	92.87	98.54	8.1	83.18	8.47	
ค่าที่ทำนายได้ / ค่าจริง	(รอบที่ 9)					
	Normal	DoS	R2L	Probing	U2R	% Correct
Normal	47,540	76	12	203	2	99.39
Dos	519	22,385	469	52	14	95.5
R2L	2,604	4	36	132	3	1.3
Probing	391	72	2	2,023	145	76.83
U2R	151	0	4	18	32	15.61
% Correct	92.84	99.33	6.88	83.32	16.33	

ตารางที่ 4.11 ผลการตรวจจับการบุกรุกโดยการรวมกันของเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบ และเทคนิคการตรวจจับความผิดปกติ ในรูปคอนฟิวชันเมตริกซ์ รอบที่ 1-10 (ต่อ)

ค่าที่ทำนายได้ / ค่าจริง	(รอบที่ 10)					
	Normal	DoS	R2L	Probing	U2R	% Correct
Normal	47,455	77	11	203	1	99.39
Dos	668	22,394	255	52	5	95.81
R2L	2,586	4	63	131	99	2.19
Probing	369	196	1	2,021	21	77.49
U2R	143	0	3	18	32	16.33
%Correct	92.65	98.78	18.92	83.34	20.25	

จากการเฉลี่ยผลตรวจจับการบุกรุก จำนวน 10 รอบ พบว่าการตรวจจับการบุกรุกโดยการรวมกันของเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและเทคนิคการตรวจจับความผิดปกติมีอัตราการตรวจจับการบุกรุกเท่ากับ 87.32% มีอัตราการแจ้งเตือนผิดพลาดเท่ากับ 0.67% และมีอัตราความแม่นยำโดยรวมเท่ากับ 94.80% ทั้งนี้ผู้วิจัยได้นำผลการตรวจจับการบุกรุกรอบที่ 7 ซึ่งมิใช่รอบที่มีอัตราความแม่นยำโดยรวมที่สูงสุดมาวิเคราะห์โดยละเอียด พบว่าสามารถตรวจจับได้ถูกต้องจำนวน 72,067 ระเบียน คิดเป็น 93.24% ทายผิดจำนวน 4,922 ระเบียน คิดเป็น 6.37% และกลุ่มไม่รู้จักจำนวน 302 ระเบียน คิดเป็น 0.39% โดยเมื่อเทียบกับผลการตรวจจับการบุกรุกด้วยเทคนิคการใช้สิทธิ์โดยมิชอบเพียงอย่างเดียว พบว่าการตรวจจับที่นำเสนอสามารถตรวจจับได้ถูกต้องเพิ่มจำนวน 610 ระเบียน คิดเป็น 0.79%

อนึ่งผู้วิจัยได้แสดงผลการตรวจจับการบุกรุกโดยการรวมกันของเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและเทคนิคการตรวจจับความผิดปกติ จากการทดสอบจำนวน 30 รอบ ในรูปกราฟอาร์โอซี (Receiver Operating Characteristics: ROC) ซึ่งแสดงความสัมพันธ์ระหว่างอัตราการตรวจจับการบุกรุก (แกนตั้ง) และอัตราการแจ้งเตือนผิดพลาด (แกนนอน) จากรูปจะเห็นได้ว่าระบบที่นำเสนอเป็นระบบตรวจจับการบุกรุกที่ดี เนื่องจากข้อมูลในกราฟอยู่ใกล้กับมุมบนซ้าย และเมื่อเปรียบเทียบกับระบบตรวจจับการบุกรุกของนักวิจัยท่านอื่นที่ปรากฏในวรรณกรรม พบว่ามีอัตราการแจ้งเตือนผิดพลาดต่ำที่สุด ดังรูปที่ 4.1



รูปที่ 4.1 กราฟอาร์โอซีของระบบตรวจจับการบุกรุกที่ได้จากการทำงานร่วมกันของเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและเทคนิคการตรวจจับความผิดปกติ

4.7 ผลการเปรียบเทียบประสิทธิภาพของระบบตรวจจับการบุกรุก

ผู้วิจัยได้ทำการเปรียบเทียบประสิทธิภาพของระบบตรวจจับการบุกรุกโดยการรวมกันของเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบที่ใช้กฎที่สร้างขึ้นโดยการรวมกันของเทคนิคเอ็มดีแอลกับเทคนิคต้นไม้ตัดสินใจ และเทคนิคการตรวจจับความผิดปกติที่ใช้การจัดกลุ่มด้วยเทคนิคเคมีนส์กับระบบตรวจจับการบุกรุกของนักวิจัยท่านอื่นที่ปรากฏในวรรณกรรม พบว่าระบบตรวจจับการบุกรุกที่ผู้วิจัยนำเสนอ มีประสิทธิภาพสูงกว่าเทคนิคอื่น โดยมีอัตราความแม่นยำโดยรวมเท่ากับ 94.80% และมีอัตราการแจ้งเตือนผิดพลาดเท่ากับ 0.67% ดังตารางที่ 4.12

ตารางที่ 4.12 ผลการเปรียบเทียบประสิทธิภาพของระบบตรวจจับการบุกรุก

เทคนิค	DTR	FPR	OA
1. การตรวจจับการใช้สิทธิ์โดยมิชอบด้วยเทคนิคซัพพอร์ตเวกเตอร์แมทชีน (Lei Li and Ke-Nan Zhao, 2011)	71.60%	28.40%	86.10%
2. การตรวจจับการใช้สิทธิ์โดยมิชอบด้วยเทคนิคเซ็ตอย่างหยาบซัพพอร์ตเวกเตอร์แมทชีน (Lei Li and Ke-Nan Zhao, 2011)	85.76%	14.24%	90.00%
3. การตรวจจับการใช้สิทธิ์โดยมิชอบด้วยเทคนิคทางพันธุกรรม (Mohammad Sazzadul Hoque, Md. Abdul Mukit and Md. Abu Naser Bikas, 2012)	95.00%	30.46%	90.04%
5. การตรวจจับการใช้สิทธิ์โดยมิชอบด้วยกฎที่สร้างขึ้นโดยการรวมกันของเทคนิคเอ็มดีแอลและเทคนิคต้นไม้ตัดสินใจและการตรวจจับความผิดปกติด้วยเทคนิคการจัดกลุ่มแบบเคมีนส์	87.32%	0.67%	94.80%

บทที่ 5

บทสรุปและข้อเสนอแนะ

การวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบประสิทธิภาพของระบบตรวจจับการบุกรุกและปรับปรุงประสิทธิภาพของระบบดังกล่าวโดยการทำงานร่วมกันระหว่างเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและเทคนิคการตรวจจับความผิดปกติ ชุดข้อมูลที่ใช้ในการวิจัย คือ ชุดข้อมูล KDD Cup 1999 ผลลัพธ์ของระบบตรวจจับการบุกรุกที่นำเสนอสามารถจัดกลุ่มได้เป็น 3 กลุ่ม คือ กลุ่มปกติ กลุ่มบุกรุก และกลุ่มไม่รู้จัก โดยกลุ่มบุกรุกสามารถจำแนกเป็นกลุ่มย่อยๆ ดังนี้ 1) กลุ่มการโจมแบบปฏิเสธการให้บริการ 2) กลุ่มการโจมตีแบบเข้าถึงระบบโดยไม่ได้รับอนุญาต 3) กลุ่มการโจมตีแบบพยายามทำตัวเป็นผู้มีสิทธิ์ และ 4) กลุ่มการโจมตีแบบสแกนเป้าหมาย ประสิทธิภาพของระบบตรวจจับการบุกรุกประเมินจากอัตราการตรวจจับการบุกรุก อัตราการแจ้งเตือนผิดพลาด และอัตราความแม่นยำโดยรวม

5.1 สรุปผล

ผลการศึกษาและเปรียบเทียบประสิทธิภาพของระบบตรวจจับการบุกรุกที่ใช้เทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและระบบตรวจจับการบุกรุกที่ใช้เทคนิคการตรวจจับความผิดปกติ Surasit Songma et al. (2012a, 2012b, 2013) พบว่าเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบ มีจุดเด่นในด้านการมีอัตราการตรวจจับการบุกรุกที่สูง และมีอัตราการแจ้งเตือนผิดพลาดที่ต่ำ แต่มีข้อจำกัดว่าสามารถตรวจจับการบุกรุกได้ดีเฉพาะชนิดการบุกรุกที่รู้จักเท่านั้น จึงจำเป็นต้องมีการเก็บระเบียนรูปแบบการบุกรุกไว้ในฐานข้อมูลเป็นจำนวนมาก ซึ่งเป็นเรื่องที่ทำได้ยากและส่งผลให้วิธีการนี้ไม่สามารถตรวจจับการบุกรุกชนิดใหม่ได้ ส่วนเทคนิคการตรวจจับความผิดปกติ มีจุดเด่นที่สามารถตรวจจับการบุกรุกชนิดใหม่ได้โดยไม่จำเป็นต้องมีความรู้เกี่ยวกับชนิดการบุกรุกนั้น แต่มีข้อจำกัดเรื่องอัตราการแจ้งเตือนผิดพลาดที่สูง เนื่องจากพฤติกรรมของผู้ใช้ระบบไม่สามารถคาดการณ์ได้ ซึ่งบางครั้งพบว่ามีความเป็นไปได้ที่จะเป็นการกระทำที่ผิดปกติแต่ไม่ใช่การบุกรุก

จากการปรับปรุงประสิทธิภาพของระบบตรวจจับการบุกรุกโดยการทำงานร่วมกันระหว่างเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบซึ่งใช้กฎที่สร้างขึ้นจากการรวมกันของเทคนิคเอ็มดีแอลกับเทคนิคต้นไม้ตัดสินใจ และเทคนิคการตรวจจับความผิดปกติซึ่งใช้การจัดกลุ่มด้วยเทคนิคเคมีนส์ พบว่าระบบตรวจจับการบุกรุกที่ได้มีอัตราความแม่นยำโดยรวมสูงและอัตราการแจ้งเตือนผิดพลาดต่ำ และสามารถตรวจจับการบุกรุกชนิดใหม่ได้ จึงทำให้สรุปได้ว่าการทำงานร่วมกันระหว่างเทคนิคการตรวจจับการใช้สิทธิ์โดยมิชอบและเทคนิคการตรวจจับความผิดปกติสามารถมีประสิทธิภาพในการตรวจจับการบุกรุกที่ดีกว่าการใช้เทคนิคการตรวจจับความผิดปกติเพียงอย่างเดียว

5.2 ข้อเสนอแนะ

ระบบตรวจจับการบุกรุกที่นำเสนอจัดเป็นแบบออฟไลน์ และชุดข้อมูลใดๆที่จะนำมาใช้กับระบบดังกล่าวจะต้องมีการแปลงในลักษณะเดียวกันกับชุดข้อมูล KDD Cup 1999 เสียก่อนจึงจะสามารถนำมาใช้ได้ แนวทางการวิจัยในอนาคต ได้แก่

1) ทำให้ระบบตรวจจับการบุกรุกสามารถระบุกลุ่มไม่รู้จักได้มากขึ้น

2) ทำให้ระบบตรวจจับการบุกรุกสามารถตรวจจับการบุกรุกประเภทการโจมตีแบบเข้าถึงระบบโดยไม่ได้รับอนุญาตได้มากขึ้น

บรรณานุกรม

กาญจนา จรัญศิริโพศาล, สิรภัทร์ เชี่ยวชาญวัฒนา, และคำรณ สุนัติ. "การจัดกลุ่มข้อมูลด้วยขั้นตอนวิธีเคอินเวอร์สฮาร์โมนิกมีน." *Proceedings of the Graduate Research Conference.* (มกราคม 2554) : 228-238.

จตุชัย แพงจันทร์. *Master in Security*. นนทบุรี : อินโฟเพรส, 2550.

ปริญญา สงวนสัตย์. *คู่มือ MATLAB ฉบับสมบูรณ์*. นนทบุรี : ไอดีซี พรีเมียร์, 2553.

พลอยพรรณ สอนสุวิทย์. "การพัฒนาขั้นตอนวิธีสำหรับการตรวจจับสิ่งผิดปกติบนการจราจรระบบเครือข่าย." วิทยานิพนธ์ปริญญาวิศวกรรมศาสตรมหาบัณฑิต, สาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์, มหาวิทยาลัยเชียงใหม่, 2552.

พวงทิพย์ แท่นแสง. "การทดสอบประสิทธิภาพการทำงานของขั้นตอนวิธีการไม่นิ่งกฎสำหรับจำแนก." วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต, สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาวิทยาการคอมพิวเตอร์และสารสนเทศ, สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2550.

มนัส สังวรศิลป์ และวรรัตน์ ภัทรอมรกุล. *คู่มือการใช้ MATLAB ฉบับสมบูรณ์* , กรุงเทพ : อินโฟเพรส, 2543.

มหาวิทยาลัยรามคำแหง. "การแบ่งกลุ่มข้อมูลอัตโนมัติ" [ออนไลน์] เข้าถึงได้จาก : http://e-book.ram.edu/e-book/c/CT477/CT477-3.pdf, 23 มกราคม 2556.

ศุภาวีร์ มากดี. "การรู้จำภาพระยะใบหน้าหลากมุมมองจำกัดพิสัยด้วยฟัซซี่แอนท์อัลกอริทึมเซ็นเตอร์ ออฟการ์วิตี้เซิร์ชและแมมเบอร์ชิฟแมชชิงสกอร์." วิทยานิพนธ์ปริญญาวิทยาศาสตรดุษฎีบัณฑิต, สาขาวิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ, มหาวิทยาลัยรังสิต, 2554.

สมเกียรติ รุ่งเรืองลดา. *คู่มือดูแลระบบ Network ฉบับมืออาชีพ*. กรุงเทพ : โปรวิชั่น, 2551.

สิรภัทร เชี่ยวชาญวัฒนา. *เอกสารประกอบการสอนวิชาข่ายงานประสาทเทียม (Artificial Neural Networks) เรื่อง ระบบการเรียนรู้แบบไม่มีผู้สอน.*, ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น, 2552.

สุรพล ดีขำ. "ระบบตรวจสอบและแกะรอยผู้บุกรุกในระบบเครือข่าย." *National Conference on Computing and Information Technology.* (มิถุนายน 2553) : 264-269.

เอกสิทธิ์ พัชรวงศ์ศักดา. *การวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมน์นิง เบื้องต้น*. กรุงเทพ : เอเชียดิจิตอลการพิมพ์, 2557.

บรรณานุกรม (ต่อ)

Axelsson, Stefan. "Intrusion Detection Systems: A Taxonomy and Survey." *Technical Report No.* 99-15. Department of Computer Engineering, Chalmers University of Technology, 2000.

Bace, Rebecca., and Mell, Peter. Intrusion Detection Systems. *NIST Special Publications on Intrusion Detection Systems*. MD : National Institute of Standards and Technology, 2011.

Bezdek, C. James. *Partition Structures: A Tutorial.* Boca Raton, FL., 1987.

_____**.** *Pattern Recognition with Fuzzy Objective Function Algorithms.* NY : Plenum Press, 1981.

Bishop, Matt. *Computer Security : Art and Science*. 1st ed. MA : Pearson, 2002.

Carter, Earl. *Cisco Intrusion Detection System*. IN : Cisco Press , 2001.

Chan, P.F. Aki., Ng, W.Y. Wing., Yeung, S. Daniel., and Tsang, C.C. Eric. "Refinement of Rule-Based Intrusion Detection System for Denial of Service Attacks by Support Vector Machine." *Proceedings of the International Conference on Machine Learning and Computing.* 7 (August 2004) : 4252-4256.

Chimphlee, Witcha. "*Hybrid Fuzzy Techniques of Unsupervised Intrusion Detection System*." Dissertation in Computer Science, Department of Faculty of Computer Science and Information Systems, University of Technology, Malaysia, 2008.

Chimphlee, Witcha., Abdullah, H. Abdul., MdSap, N. Mohd., Chimphlee, Siriporn., and Srinoy. "To Misuse and Anomaly Attacks through Induction Analysis and Fuzzy Methods." *WSEAS Transactions on Computers*. 5 (January 2006) : 49-54.

_____. "Anomaly Detection of Intrusion Based on Integration of Rough Sets and Fuzzy *c*-Means." *Journal of Information Technology*. 17 (December 2005) : 1-14.

Dash, Rajashree., Paramguru, Lochan Rajib., and Dash, Rasmita. "Comparative Analysis of Supervised and Unsupervised Discretization Techniques." *International Journal of Advances in Science and Technology.* 2 (2011) : 29-37.

Denning, E. Dorothy. "An Intrusion-Detection Model." *IEEE Transactions on Software Engineering*. SE-13 (1987) : 222-232.

Elkan Charles. "Result of the KDD'99 Classifier Learning." *SIGKDD Explorations, ACM SIGKDD*. 1 (January 2000) : 63-64.

บรรณานุกรม (ต่อ)

Freedman, David., Pisani, Robert., and Purves, Roger. *Statistics*. NY: W. W. Norton and Company, 1997.

Gomez, Jonatan. "Soft Computing Techniques for Intrusion Detection." [Online] available : http://www.dis.unal.edu.co/~jgomezpe/docs/books/dissertation.pdf, 10 October 2014.

Hammouda, Khaled. "A Comparative Study of Data Clustering Techniques." [Online] available : http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.126.3224&rep=rep1&type=pdf, 5 October 2013.

Han, Jiawei., and Kamber, Micheline. *Data Mining: Concepts and Techniques*. MA : Morgan Kaufmann, 2000.

_____. *Data Mining: Concepts and Techniques*. MA : Morgan Kaufmann, 2006.

Hawkins, M. Douglas. *Identification of Outliers*. NY : Chapman and Hall, 1980.

Hemada, B. and Lakshmi, K.S. Vijaya. "A Study On Discretization Technique." *International Journal of Engineering Research & Technology*. 2 (2013): 1887-1892.

Hettich, Seth., and Bay, D. Stephen. "The UCI KDD Archive." [Online]. available at : http://kdd.ics.uci.edu., 1 January 2012.

Hoque, S. Mohammad., Mukit, A. Md., and NaserBikas, A. Md. "An Implementation of Intrusion Detection System Using Genetic Algorithm." *International Journal of Network Security and Its Applications*. 4 (March 2012) : 109-120.

Jahanbani, Adel. and Karimi, Hossein. "A New Approach for Detecting Intrusion Based on the PCA Neural Networks." *Journal of Basic and Applied Scientific Research*. 2 (2012) : 672-679.

Jones , K. Anita., and Sielken, S. Robert. "Computer System Intrusion Detection: A Survey." [Online] available at : http://www.princeton.edu/~rblee/ELE572Papers/Fall04Readings/IntrusionDetection_jones-sielken-survey-v11.pdf, 1 January 2014.

Kandeeban, S. Selvakani. "A Genetic Algorithm Based Elucidation for Improving Intrusion Detection through Condensed Feature Set by KDD 99 Data Set." *Journal of Information and Knowledge Management*. 1 (2011) : 1-10.

บรรณานุกรม (ต่อ)

Kandeeban, Selvakani., and Rajesh, R. S. "Integrated Intrusion Detection Using SCT." *International Journal of Computer and Network Security*. 1 (November 2009) : 128-133.

Kantardzic, Mehmed. *Data Mining: Concepts, Models, and Algorithms.* NJ : IEEE Press, 2003.

Kendall, Kristopher. "A Database of Computer Attacks for the Evaluation of Intrusion Detection Systems." Master's Thesis in Engineering, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Boston, 1998.

Larose, T. Daniel. *Data Mining Methods and Models.* Hoboken, NJ : Wiley, 2006.

Li, Lei., and Zhao, Ke-nan. "A New Intrusion Detection System Based on Rough Set Theory and Fuzzy Support Vector Machine." *Proceedings of the Third International Workshop on Intelligent Systems and Applications.* (May 2011) : 1-5.

Lu, Wei. "An Unsupervised Anomaly Detection Framework for Multiple-Connection Based Network Intrusions." Dissertation in Electrical and Computer Engineering, Department of Electrical and Computer Engineering, University of Victoria, Canada, 2005.

MacQueen, J. B. "Some Methods for Classification and Analysis of Multivariate Observations." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. 1 (1967) : 281-297.

Madhukumar, S., and Santhiyakumari, N. "Evaluation of k-Means and fuzzy C-means segmentation on MR images of brain." *The Egyptian Journal of Radiology and Nuclear Medicine*. 46 (March 2015) : 475-479.

Narayan, Esh., Singh, Pankaj., and Tak, Kumar. Gaurav. "Intrusion Detection System Using Fuzzy *c*- Means Clustering with Unsupervised Learning via EM Algorithms." *Journal of Computer Science and Information Technology*. 2 (2012) : 502-510.

Negnevitsky, Michael. *Artificial Intelligence: A Guide to Intelligent Systems.* 3rd ed., Essex, UK : Pearson, 2011.

Niknam, Taher., Fard, Taherian Elahe., Pourjafarian, Narges., and Rousta, Alireza. "An Efficient Hybrid Algorithm based on Modified Imperialist Competitive Algorithm and *k*-Means for Data Clustering." *Journal Engineering Applications of Artificial Interlligence*. 24 (2001) : 306-317.

บรรณานุกรม (ต่อ)

Olusola, A. Adetunmbi., Oladele, S. Adeola., and Abosede, O. Daramola. "Analysis of KDD'99 Intrusion Detection Dataset for Selection of Relevance Features." *Proceedings of the World Congress on Engineering and Computer Science.* 1 (October 2010) : 162-168.

Sani, Yusuf., Mohamedou, Ahmed., Ali, Khalid., Farjamfar, Anahita., Azman, Mohamed., and Shamsuddin, Solahuddin. "An Overview of Neural Use in Anomaly Intrusion Detection Systems." *Proceedings of 2009 IEEE Student Conference on Research and Development*. (November 2009) : 89-82.

Secure-Byte. "Network Designing Services." [Online]. available at : http://www.secure-bytes.com/designing+services.php., 2 February 2013.

Soldier, Thuan. "IDS-Diagram." [Online] available at : http://thuansoldier.net/?attachment_id=669., 3 March 2012.

Songma, Surasit., Chimphlee, Witcha., Maichalernnukul, Kiattisak., and Sanguansat, Parinya. "Implementation of Fuzzy *c*-Means and Outlier Detection for Intrusion with KDD Cup 1999 Data Set." *International Journal of Engineering Research and Development*. 2 (July 2012) : 44-48.

_____. "Classification via *k*-Means Clustering and Distance-Based Outlier Detection." *Proceedings of the Tenth International Conference on ICT and Knowledge Engineering.* (November) : 125-128.

_____. "Intrusion Detection through Rule Induction Analysis." *Advances in information Sciences and Service Sciences*. 5 (June 2013) : 187-194.

Teng, Shaohua., Du, Hongle., Wu, Naiqi., Zhang, Wei., and Su, Jiangyi. "A Cooperative Network Intrusion Detection Based on Fuzzy SVMs." *Journal of Networks*. 5 (April 2010) : 475-483.

Tsai, S. Flora. "Network Intrusion Detection Using Association Rules." *International Journal of Recent Trends in Engineering.* 2 (November 2009): 202-204.

Vanderbilt, R.E. "Anomaly Detection in Computer Networks Using Type-2 Fuzzy Logic." Dissertation in Computer Engineering, Department of Computer Engineering, Florida Institute of Technology, FL, USA, 2005.

บรรณานุกรม (ต่อ)

Wang, Deguang. and Zhou, Zhigang. "Application of a Collaborative Filtering Recommendation Algorithm Based on Cloud Model in Intrusion Detection." *Journal of Networks.* 6 (February 2011) : 214-221.

Watthananon, Julaluk., and Mingkhwan, Anirach. "Comparative Efficiency of Correlation Plot Data Classification." *Journal of KMUTNB*. 22 (2012) : 77-89.

Wilkin, A. Gregory., and Huang, Xiuzhen. "K-Means Clustering Algorithms: Implementation and Comparison." *Proceeding of the Second International Multi-Symposium of Computer and Computational Sciences*. (August 2007): 133-136.

Yang, Zhongxue., and Karahoca, Adem. "An Anomaly Intrusion Detection Approach Using Cellular Neural Networks." *Proceedings of the International Symposium on Computer and Information Sciences*. (November 2006) : 908-917.

Zhang, Jiong., and Zulkernine, Mohammad. "A Hybrid Network Intrusion Detection Technique Using Random Forests." *Proceedings of International Conference on Availability Reliability and Security*. (April 2006): 262-269.

Zhong, Shi., Khoshgoftaar, Taghi., and Seliya, Naeem. "Evaluating Clustering Techniques for Unsupervised Network Intrusion Detection." *Proceedings of the Tenth ISSAT International Conference on Reliability and Quality in Design*. (August 2004) : 178-182.

ภาคผนวก

ภาคผนวก ก

ผลงานตีพิมพ์เผยแพร่ในการประชุมวิชาการนานาชาติและวารสารวิชาการนานาชาติ

ผลงานตีพิมพ์เผยแพร่ในการประชุมวิชาการนานาชาติและวารสารวิชาการนานาชาติ

S. Songma, W. Chimphlee, K. Maichalernnukul and P. Sanguansat. "Intrusion Detection through Rule Induction Analysis." Advances in information Sciences and Service Sciences, vol.5 (2013), 187-194.

S. Songma, W. Chimphlee, K. Maichalernnukul and P. Sanguansat. "Implementation of Fuzzy *c*-Means and Outlier Detection for Intrusion Detection with KDD Cup 1999 Data Set." International Journal of Engineering Research and Development. Vol. 2 (2012), 44-48.

S. Songma, W. Chimphlee, K. Maichalernnukul and P. Sanguansat. "Classification via *k*-Means Clustering and Distance-Based Outlier Detection." In Proceeding of 2012 Tenth International Conference on ICT and Knowledge Engineering." (2012), 125-128.

Contents:

Intrusion Detection through Rule Induction Analysis

[1] Surasit Songma, [2] Witcha Chimphlee, [3] Kiattisak Maichalernnukul, [4] Parinya Sanguansat
[*1] *Faculty of Information Technology, Rangsit University, Phathumthani, Thailand, surasit.songma@gmail.com*
[2] *Faculty of Science and Technology, Suan Dusit Rajabhat University, Bangkok, Thailand, witcha.chi@gmail.com*
[3] *Faculty of Information Technology, Rangsit University, Phathumthani, Thailand, kiattisak.m@rsu.ac.th*
[4] *Faculty of Engineering and Technology, Panyapiwat Institute of Management, Nonthaburi, Thailand, sanguansat@yahoo.com*

Abstract

In this paper, we present intrusion detection by means of rule induction analysis. Specifically, the rule induction analysis is used to generate signatures from the Knowledge Discovery Databases (KDD) Cup 1999 testing data set and find a set of rules that satisfy some predefined criteria. The results show that this kind of intrusion detection is not only able to achieve extremely high detection rate (99.99%) but also to yield very low false positive rate (1.65%).

Keywords: *Intrusion Detection, KDD Cup 1999 Data Set, Rule Induction Analysis.*

1. Introduction

Intrusion detection is the process of monitoring the events occurring in a computer system or network and analysing them for signs of intrusions [1]. An Intrusion Detection System (IDS) is a combination of software and hardware that attempt to perform intrusion detection and raise alarm when possible intrusions are detected. The objective of the IDS is to act like a burglar alarm, which monitors the network and the connected systems to find evidence of intrusion.

Traditionally, intrusion detection approaches can be classified into two main categories: misuse detection and anomaly detection [2]. Misuse detection systems use *a prior* knowledge on attacks to look for attack traces. In other words, they detect intrusions by knowing that the misuse is the signature (rule) based system. Since signatures are associated with specific misuse behaviour, it is easy to determine the attack type. Anomaly detection systems adopt the opposite approach, which is, to know what is normal, and then find the deviations from the normal behaviour. These deviations are considered as anomalies or possible intrusions.

In this paper, we present misuse detection via rule induction analysis. First, a set of signatures, describing the characteristics of possible attacks, are generated from the Knowledge Discovery Databases (KDD) Cup 1999 data set [3], and the corresponding rules are stored in the IDS. Then, these rules are used to detect the incoming hostile traffic passing through the IDS.

The rest of the paper is organized as follows: Section 2 describes the rule induction analysis. Section 3 discusses the experimental results. Finally, Section 4 presents the conclusion.

2. Rule induction analysis

A rule-based IDS analyzes the traffic data passing through it and differentiates the traffic behaviors to be intrusive or normal. The rule-based IDS uses the rules stored in its knowledge base to detect and take actions when anomaly occurs in the traffic and/or undergoing some unauthorized activities [4]. The discovered rule are typically expressed in the form IF-THEN.

An example of an IF-THEN rule is shown in Fig. 1. Such a rule can have multiple conditions joined by conjunction (AND), disjunction (OR) or combination of both. In this figure, A, B, C, and D are conditions. If A, B, C, and D are true, then the action X is taken. Before proceeding, we give a brief overview of the traffic data used in this paper.

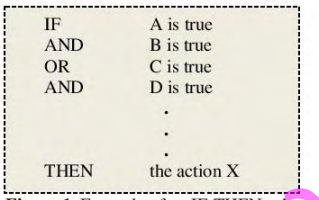


Figure 1. Example of an IF-THEN rule

The KDD Cup 1999 data set was originally prepared by MIT Lincoln labs for the 1998 Defense Advanced Research Projects Agency (DARPA) Intrusion Detection Evaluation Program, with the objective of evaluating research in intrusion detection. This data set consists of 494,021 training instances and 311,029 testing instances. Each instance represents a network connection, i.e., a sequence of network packets starting and ending at some well defined times, between which data flows to and from a source IP address to a target IP address under some well defined protocol. Such a connection instance is described by a 41-dimensional feature vector and labeled with respect to five classes:

- Normal;
- Denial of service (DoS), where an attacker makes some computing or memory resource too busy or too full to handle legitimate requests, thus denying legitimate users access to a machine, e.g., SYN flood;
- Remote to local (R2L), where an attacker sends packets to a machine over a network, then exploits machine's vulnerability to illegally gain local access as a user, e.g., guessing password;
- User to root (U2R), where an attacker starts out with access to a normal user account on the system and is able to exploit vulnerability to gain root access to the system, e.g., buffer overflows;
- Probing, where an attacker scans a network to gather information or find known vulnerabilities, e.g., port scanning.

For clarity, all features are listed in Table 1 on the last page, and their descriptive statistics are shown in Tables 2-5 where C and D denote Cont. and discrete, respectively. Examples of data contained in the features are as follows:

- *"icmp"*, *"tcp"*, and *"udp"* in protocol_type;
- *"IRC"*, *"ftp"*, *"pop_3"*, *"http"*, *"telnet"*, and *"login"* in service;
- *0* and *1* in logged_in;
- *0-255* in dst_host_count.

Table 2. Descriptive statistics of basic features of individual TCP connection

Feature name	Min	Max	Mean	Standard deviation
1. duration (C)	0	57,715	70.762	815.287
2. protocol type (D)	1	3	2.042	0.279
3. service (D)	1	65	29.714	12.218
4. flag (D)	1	11	7.005	3.195
5. src_bytes (C)	0	62,825,648	3,427.815	255,452.832
6. dst_bytes (C)	0	5,203,179	2,870.84	32,229.826
7. land (D)	0	1	0.0001	0.011
8. wrong_fragment (C)	0	3	0.003	0.078
9. urgent (C)	0	3	0.0002	0.020

Table 3. Descriptive statistics of content features within a connection suggested by domain knowledge

Feature name	Min	Max	Mean	Standard deviation
10. hot (C)	0	101	0.040	0.594
11. num_failed_logins (C)	0	4	0.007	0.084
12. logged_in (D)	0	1	0.585	0.493
13. num_compromised (C)	0	796	0.036	3.927
14. root_shell (C)	0	1	0.001	0.028
15. su_attempted (C)	0	2	0.00009	0.012
16. num_root (C)	0	878	0.034	4.343
17. num_file_creations (C)	0	100	0.004	0.387
18. num_shells (C)	0	5	0.0003	0.026
19. num_access_files (C)	0	4	0.003	0.059
20. num_otbound_cmds (C)	0	0	0.000	0.000
21. is_host_login (D)	0	1	0.0002	0.012
22. is_guest_login (D)	0	1	0.010	0.097

Table 4. Descriptive statistics of traffic features computed using a two-second time window

Feature name	Min	Max	Mean	Standard deviation
23. count (C)	0	511	68.089	105.097
24. srv_count (C)	0	511	18.336	53.464
25. serror_rate (C)	0	1	0.097	0.290
26. srv_serror_rate (C)	0	1	0.098	0.293
27. rerror_rate (C)	0	1	0.209	0.402
28. srv_rerror_rate (C)	0	1	0.209	0.403
29. same_srv_rate (C)	0	1	0.730	0.423
30. diff_srv_rate (C)	0	1	0.047	0.161
31. srv_diff_host_rate (C)	0	1	0.097	0.230

Table 5. Descriptive statistics of traffic features computed using a two-second time window from destination to host

Feature name	Min	Max	Mean	Standard deviation
32. dst_host_count (C)	0	255	178.380	99.731
33. dst_host_srv_count (C)	0	255	158.263	112.549
34. dst_host_same_srv_rate (C)	0	1	0.661	0.438
35. dst_host_diff_srv_rate (C)	0	1	0.046	0.134
36. dst_host_same_srv_port_rate (C)	0	1	0.077	0.221
37. dst_host_srv_diff_host_rate (C)	0	1	0.016	0.053
38. dst_host_serror_rate (C)	0	1	0.095	0.282
39. dst_host_srv_serror_rate (C)	0	1	0.096	0.288
40. dst_host_rerror_rate (C)	0	1	0.208	0.392
41. dst_host_srv_rerror_rate (C)	0	1	0.206	0.397

Table 6. Distribution and ratio of the KDD Cup 1999 testing data set

Class	Original		Subset	
	Amount of data	Ratio (%)	Amount of data	Ratio (%)
normal	60,593	19.48	47,913	61.99
DoS	229,853	73.90	23,568	30.49
R2L	16,189	5.20	2,913	3.77
U2R	228	0.07	215	0.27
probing	4,166	1.34	2,682	3.47
Total	311,029	100	77,291	100

In this paper, we consider the testing data set only. Its class distribution and ratio are shown in Table 6. Before we could do any experiment, data preprocessing has to be undertaken. We handle missing and incomplete data by involves mapping symbolic-valued attributes to numeric-valued attributes. Because the KDD Cup 1999 data set has a huge number of duplicated instances, we instead use its subset which comprises 77,291 distinct instances (see Table 6). Based on this subset, the corresponding class rules can be generated and exemplified as follows (where variable labels are shown in Table 1):

2.1 Normal Class Rule

IF *B = "icmp"* OR *B = "tcp"* OR *B = "udp"* AND *C = "IRCr"* OR *C = "auth"* OR *C = "domain_u"* OR *C = "eco_i"* OR *C = "ecr_i"* OR *C = "finger"* OR *C = "ftp"* OR *C = "ftp_data"* OR *C = "http"* OR *C = "link"* OR *C = "ntp_u"* OR *C = "other"* OR *C = "pop_3"* OR *C = "private"* OR *C = "remote_job"* OR *C = "smtp"* OR *C = "telnet"* OR *C = "time"* OR *C = "urp_i"* OR *C = "X11"* OR *C = "tcomp"* OR *C = "tftp_u"* OR *C = "tim_i"* AND *D ~= "S0"* AND *E >= 0* AND *G = 0* AND *H = 0* OR *H = 1* OR *H = 3* AND *I = 0* AND *K = 0* OR *K = 1* OR *K = 3* AND *L = 0* OR *L = 1* AND *M >= 0* AND *M <= 7* OR *M = 16* OR *M = 381* OR *M = 611* OR *M = 796* AND *N = 0* OR *N = 1* AND *O = 0* OR *O = 1* OR *O = 2* AND *P >= 0* AND *P <= 4* OR *P = 6* OR *P = 9* OR *P = 401* OR *P = 684* OR *P = 878* *Q >= 0* AND *Q <= 2* OR *Q = 12* OR *Q = 13* OR *Q = 30* OR *Q = 100* AND *R = 0* AND *S >= 0* AND *S <= 3* AND *T = 0* AND *U = 0* OR *U = 1* AND *V = 0* OR *V = 1* AND *Y = 0* OR *Y >= 0.02* AND *Y <= 0.17* OR *Y = 0.2* OR *Y = 0.25* OR *Y = 0.33* OR *Y = 0.5* OR *Y = 1* AND *AA = 0* OR *AA = 0.2* OR *AA = 0.33* OR *AA = 0.5* OR *AA = 0.98* OR *AA = 1* AND *AH >= 0* AND *AH <= 1* AND *AI ~= 0.34* AND *AJ >= 0* AND *AJ <= 1* THEN *Class_type* = ***"normal"***.

2.2 DoS Class Rule

IF *A = 0* AND *B = "tcp"* AND *C = "finger"* OR *C = "telnet"* AND *D = "S0"* AND *E = 0* AND *F = 0* AND *G = 1* AND *H = 0* AND *I = 0* AND *J = 0* AND *K = 0* AND *L = 0* AND *M = 0* AND *N = 0* AND *O = 0* AND *P = 0* AND *Q = 0* AND *R = 0* AND *S = 0* AND *T = 0* AND *U = 0* AND *V = 0* AND *W = 1* AND *X = 1* OR *X = 2* OR *X = 3* AND *Y = 1* AND *Z = 1* AND *AA = 0* AND *AB = 0* AND *AC = 1* AND *AD = 0* AND *AE = 0* OR *AE = 1* AND *AN = 0* OR *AN = 0.02* AND *AO = 0* THEN *Class_type* = ***"land"***.

2.3 R2L Class Rule

IF *A = 0* AND *B = "tcp"* AND *C = "http"* AND *D = "REJ"* OR *D = "SF"* AND *E = 0* OR *E = 51* AND *F = 0* OR *F = 8371* AND *G = 1* AND *H = 0* AND *I = 0* AND *J = 0* OR *J = 2* AND *K = 0* AND *L = 0* OR *L = 1* AND *M = 0* AND *N = 0* OR *N = 1* AND *O = 0* AND *P = 0* AND *Q = 0* AND *R = 0* AND *S = 0* OR *S = 1* AND *T = 0* AND *U = 0* AND *V = 0* AND *W = 1* AND *X = 1* AND *Y = 0* AND *Z = 0* AND *AA = 0* OR *AA = 1* AND *AB = 0* OR *AB = 1* AND *AC = 1* AND *AD = 0* AND *AE = 0* AND *AF = 17* OR *AF = 255* AND *AG = 17* OR *AG = 227* AND *AH = 0.89* OR *AH = 1* AND *AI = 0* OR *AI = 0.02* AND *AJ = 0* OR *AJ = 0.06* AND *AK = 0* AND *AL = 0* AND *AM = 0* AND *AN = 0* OR *AN = 0.06* AND *AO = 0* OR *AO = 0.06* THEN *Class_type* = ***"phf"***.

2.4 U2R Class Rule

IF *A = 75* OR *A = 84* AND *B = "tcp"* AND *C = "telnet"* AND *D = "SF"* AND *E = 227* AND *F = 1087* OR *F = 1089* AND *G = 0* AND *H = 0* AND *I = 0* AND *J = 2* AND *K = 0* AND *L = 1* AND *M = 1* AND *N = 1* AND *O = 0* AND *P = 0* AND *Q = 4* AND *R = 2* AND *S = 0* AND *T = 0* AND *U = 0* AND *V = 0* AND *W = 1* AND *X = 1* AND *Y = 0* AND *Z = 0* AND *AA = 0* AND *AB = 0* AND *AC = 1* AND *AD = 0* AND *AE = 0* AND *AF = 137* OR *AF = 255* AND *AG = 1* OR *AG = 101* AND *AH = 0* OR *AH = 0.74* AND *AI = 0.04* OR *AI = 0.07* AND *AJ = 0* OR *AJ = 0.01* AND *AK = 0* AND *AL = 0.14* OR *AL = 0.73* AND *AM = 0* OR *AM = 0.99* AND *AN = 0* OR *AN = 0.86* AND *AO = 0* THEN *Class_type* = ***"loadmodule"***.

2.5 Probing Class Rule

IF *A* = 0 AND *B = "tcp"* AND *C = "courier"* OR *C = "domain"* OR *C = "nnsp"* OR *C = "private"* OR *C = "uucp"* AND *D = "SH"* AND *E = 0* AND *F = 0* AND *G = 0* AND *H = 0* AND *I = 0* AND *J = 0* AND *K = 0* AND *L = 0* AND *M = 0* AND *N = 0* AND *O = 0* AND *P = 0* AND *Q = 0* AND *R = 0* AND *S = 0* AND *T = 0* AND *U = 0* AND *V = 0* AND *W = 1* OR *W = 2* AND *X = 1* AND *Y = 1* AND *Z = 1* AND *AA = 0* AND *AB = 0* AND *AC = 0.5* OR *AC = 1* AND *AD = 0* OR *AD = 1* AND *AE = 0* AND *AG= 1* OR *AG = 3* AND *AI = 0.99* OR *AI = 1* AND *AJ >= 0.98* AND *AK = 0* OR *AK = 0.67* AND *AL >=0.98* AND *AM = 1* AND *AN = 0* AND *AO = 0* THEN *Class_type* = ***"nmap"***.

Through the above rule induction analysis, the misuse detection for the data subset in Table 6 can be achieved. The overall approach is illustrated in Fig. 2

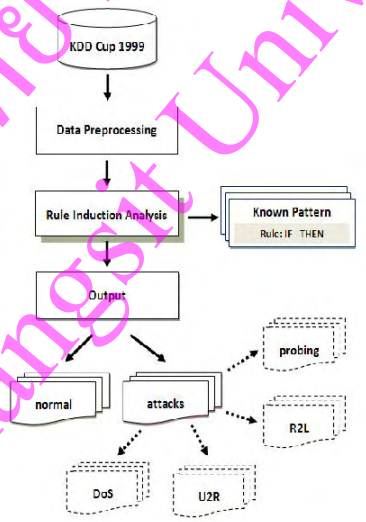


Figure 2. Proposed intrusion detection

3. Experimental results

The results of testing the proposed intrusion detection are summarized in Table 7. The top left entry in the confusion matrix shows that 47,122 of the actual "normal" instances were predicted to be normal by this entry. The last column indicates that in total 98.35% of the "normal" instances were recognized correctly. The bottom row shows that 99.99% of instances said to be normal were indeed "normal" in reality. The same analysis can be applied to the other four classes.

Table 7. Confusion matrix

	Predicted Class					
Actual Class	**normal**	**DoS**	**R2L**	**U2R**	**Probing**	**% Correct**
normal	47,122	456	300	0	35	98.35
DoS	0	23,316	1	0	251	98.93
R2L	1	0	2,912	0	0	99.97
U2R	0	0	0	215	0	100
probing	0	1	2	0	2,679	99.89
% Correct	99.99	98.08	90.58	100	90.35	

In Fig. 3, we compare the detection rate (DTR), false positive rate (FPR), and overall accuracy (OA) of our rule-based approach with those of other approaches in the literature, i.e., the support vector machine, rough set-fuzzy support vector machine [5], genetic algorithm [6], and two-phase classification [7]. The details of how DTR, FPR, and OA are calculated can be found in [8]. It is seen from this figure that the rule-based approach is the best as it achieves the highest DTR and OA, and the lowest FPR.

Figure 3. Comparison of DTR, FPR, and OA of the proposed rule-based approach with those of other approaches in the literature

4. Conclusion

In this paper, we have proposed a rule-based intrusion detection approach. The experimental results have shown that the proposed approach is very promising in terms of its DTR, FPR, and OA performances.

5. References

[1] Rebecca Bace and Peter Mell, "Intrusion Detection Systems", NIST Special Publications on Intrusion Detection Systems, SP 800-31, 2011.
[2] Chen Xiao-Ming, "Research on the Network Intrusion Detection Model with Neural Network Parameters Optimized by Ant-Colony Algorithm", Journal of Convergence Information Technology, vol. 8, no. 5, pp. 619-627, 2013.
[3] KDD Data Set. (1999) [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
[4] Aki P. F. Chan, Wing W. Y. Ng, Daniel S. Yeung, and Eric C. C. Tsang, "Refinement of Rule-Based Intrusion Detection System for Denial of Service Attacks by Support Vector Machine", In Proceedings of the International Conference on Machine Learning and Cybernetics, pp. 4252-4256, 2004.
[5] Lei Li and Ke-Nan Zhao, "A New Intrusion Detection System based on Rough Set Theory and Fuzzy Support Vector Machine", In Proceedings of the International Workshop on Intelligent Systems and Applications, pp. 1-5, 2011.
[6] Mohammea S. Hoque, Md. A. Mukit, and Md. A. N. Bikas, "An Implementation of Intrusion Detection System Using Genetic Algorithm", International Journal of Network Security and Its Applications, vol. 4, no. 2, pp. 109-120, 2012.
[7] Surasit Songma, Witcha Chimphlee, Kiattisak Maichalernnukul, and Parinya Sanguansat, "Classification via k-Means Clustering and Distance-Based Outlier Detection," In Proceedings of the International Conference on ICT and Knowledge Engineering, pp. 125-128, 2012.
[8] Witcha Chimphlee, Mohd N. Md Sap, Abdul H. Abdullah, Siriporn Chimphlee, and Surat Srinoy, "Anomaly Detection of Intrusion Based on Integration of Rough Sets and Fuzzy c-Means", Journal of Information Technology, vol. 17, no. 2, pp. 1-14, 2005.

Table 1. Feature description of the KDD Cup 1999 data set

Feature name	Description	Variable label
1. duration (C)	length (number of seconds) of the connection	A
2. protocol type (D)	type of the protocol, e.g., tcp, udp, etc.	B
3. service (D)	network service on the destination, e.g., http, telnet, etc.	C
4. flag (D)	normal or error status of the connection	D
5. src_bytes (C)	number of data bytes from source to destination	E
6. dst_bytes (C)	number of data bytes from destination to source	F
7. land (D)	1 if connection is from/to the same host/port; 0 otherwise	G
8. wrong_fragment (C)	number of "wrong" fragments	H
9. urgent (C)	number of urgent packets	I
10.hot (C)	number of "hot" indicators	J
11. num_failed_logins (C)	number of failed login attempts	K
12. logged_in (D)	1 if successfully logged in; 0 otherwise	L
13. num_compromised (C)	number of "compromised" conditions	M
14. root_shell (C)	1 if root shell is obtained; 0 otherwise	N
15. su_attempted (C)	1 if "su root" command attempted; 0 otherwise	O
16. num_root (C)	number of "root" accesses	P
17. num_file_creations (C)	number of file creation operations	Q
18. num_shells (C)	number of shell prompts	R
19. num_access_files (C)	number of operations on access control files	S
20. num_otbound_cmds (C)	number of outbound commands in an ftp session	T
21. is_host_login (D)	1 if the login belongs to the "hot" list; 0 otherwise	U
22. is_guest_login (D)	1 if the login is a "guest" login; 0 otherwise	V
23. count (C)	number of connections to the same host as the current connection in the past two seconds	W
24. srv_count (C)	number of connections to the same service as the current connection in the past two seconds	X
25. serror_rate (C)	% of connections that have "SYN" errors	Y
26. srv_serror_rate (C)	% of connections that have "SYN" errors	Z
27. rerror_rate (C)	% of connections that have "REJ" errors	AA
28. srv_rerror_rate (C)	% of connections that have "REJ" errors	AB
29. same_srv_rate (C)	% of connections to the same service	AC
30. diff_srv_rate (C)	% of connections to different services	AD
31. srv_diff_host_rate (C)	% of connections to different hosts	AE
32. dst_host_count (C)	count for destination host	AF
33. dst_host_srv_count (C)	srv_count for destination host	AG
34. dst_host_same_srv_rate (C)	same_srv_rate for destination host	AH
35. dst_host_diff_srv_rate (C)	dif_srv_rate for destination host	AI
36. dst_host_same_src_port_rate (C)	same_src_port_rate for destination host	AJ
37. dst_host_srv_diff_host_rate (C)	diff_host_rate for destination host	AK
38. dst_host_serror_rate (C)	serror_rate for destination host	AL
39. dst_host_srv_serror_rate (C)	srv_serror_rate for destination host	AM
40. dst_host_rerror_rate (C)	rerror_rate for destination host	AN
41. dst_host_srv_rerror_rate (C)	srv_serror_rate for destination host	AO

International Journal of Engineering Research and Development
eISSN : 2278-067X, pISSN : 2278-800X, www.ijerd.com
Volume 2, Issue 2 (July 2012), PP. 44-48

Implementation of Fuzzy c-Means and Outlier Detection for Intrusion Detection with KDD Cup 1999 Data Set

S. Songma[1], W. Chimphlee[2], K. Maichalernnukul[3], P. Sanguansat[4]
[1,3]Faculty of Information Technology, Rangsit University, Muang-Ake, Phathumthani 12000, Thailand
[2]Faculty of Science and Technology, Suan Dusit Rajabhat University, Bangkok 10300, Thailand
[4]Faculty of Engineering and Technology, Panyapiwat Institute of Management, Nonthaburi 10200, Thailand

Abstract— In this paper, a two-phase method for computer network intrusion detection is proposed. In the first phase, a set of patterns (data) are clustered by the fuzzy c-means algorithm. In the second phase, outliers are constructed by a distance-based technique and a class label is assigned to each pattern. The KDD Cup 1999 data set is used for the experiment. The results show that, for binary classification (i.e., normal or attack), the proposed method achieves a higher detection rate and a greater overall accuracy than the fuzzy c-means algorithm.

Keywords— Clustering, fuzzy c-means, intrusion detection, KDD Cup 1999 data set, outlier detection

I. INTRODUCTION

As defined in [1], intrusion detection is the process of monitoring the events occurring in a computer network and analyzing them for signs of intrusions. It is also defined as attempts to compromise the confidentiality, integrity, availability, or to bypass the security mechanisms of a computer network. Anomaly intrusion detection systems (IDSs) aim at distinguishing an abnormal activity from an ordinary one.

The current state of computer networks is vulnerable; they are prone to an increasing number of attacks. These attacks are seldom previously seen. It is very hard to detect them before subsequent damage is done. Therefore, securing such a network from unwanted malicious traffic is of prime concern.

In this paper, a two-phase method for intrusion detection, called 2PID, is proposed. The Knowledge Discovery in Databases (KDD) Cup 1999 data set [2], which has been utilized extensively for development of IDSs, is used as a representative sample of data.

The rest of the paper is organized as follows: Section II introduces the proposed method. Section III describes the experimental setup. Section IV provides the results, and Section V concludes the paper.

II. METHOD DESCRIPTION

In this section, we first review fuzzy c-means (FCM) clustering and distance-based outlier detection. Then, we present our proposed method.

A. FCM Clustering

Clustering is an unsupervised classification mechanism where a set of patterns (data), usually multidimensional, are classified into groups (clusters) such that members of one group are similar according to a predefined criterion [3].

FCM is an unsupervised fuzzy clustering algorithm that has been applied successfully to a number of problems involving feature analysis, clustering, and classifier design. It takes unlabeled intrusion data points and tries to group them according to their similarity; points assigned to the same cluster have high similarity, while the similarity between points assigned to different clusters is low [4].

The FCM algorithm partition a set of N patterns $\{X_k\}$ into c clusters by minimizing the objective function

$$J = \sum_{k=1}^{N}\sum_{i=1}^{c}(\mu_{ik})^{m'}\left\|X_k - m_i\right\|^2 \quad (1)$$

where $1 \le m' < \infty$ is the fuzzifier, m_i is the i^{th} cluster center, $\mu_{ik} \in [0,1]$ is the membership of the k^{th} pattern to it, and $\|\cdot\|$ is the distance norm. The parameters m_i and μ_{ik} are calculated as

$$m_i = \frac{\sum_{k=1}^{N}(\mu_{ik})^{m'} X_k}{\sum_{k=1}^{N}(\mu_{ik})^{m'}}, \quad (2)$$

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}} \right)^{\frac{2}{m'-1}}} \tag{3}$$

with $d_{ik} = \|X_k - m_i\|^2$, subject to $\sum_{i=1}^{c} \mu_{ik} = 1$ and $0 < \sum_{k=1}^{N} \mu_{ik} < N$. The algorithm proceeds as follows [5]:

(i) Pick the initial means m_i, $i = 1,2,...,c$. Choose the values for the fuzzifier m' and the threshold ε. Set the iteration counter $t = 1$;

(ii) Compute μ_{ik} for c clusters and N data points, by (3);

(iii) Update m_i by (2);

(iv) Repeat steps (ii) and (iii), by incrementing t, until $\|\mu_{ik}(t) - \mu_{ik}(t-1)\| > \varepsilon$.

B. Distance-Based Outlier Detection

Outlier is defined as an observation that appears to be inconsistent with other observations in a data set. Many data-mining algorithms try to minimize the influence of outliers on the final model, or to eliminate them in the preprocessing phases. Outlier detection and potential removal from the data set can be described as a process of the selection of L out of N samples that are considerably dissimilar, exceptional, or inconsistent with respect to the remaining data.

Distance-based technique is a class of outlier-detection method. The basic computational complexity of this technique is the evaluation of distance measures between all samples in a given data set. Then, a sample in a data set $\{X_k\}$ is an outlier if at least a fraction p of the samples in $\{X_k\}$ lies at a distance greater than r. Clearly, the criterion for outlier detection is based on p and r. These two parameters may be given beforehand using knowledge about the data. Further details are in [6].

C. Proposed Method

The 2PID consists of two phases. In the first phase, a set of patterns are classified by FCM clustering. In the second phase, outliers are constructed by a distance-based technique, and a class label is assigned to each pattern.

Binary classification is the task of classifying the members of a given data set into two groups on the basis of whether they have some property or not. The binary classification task in the context of intrusion detection is to differentiate between normal connections and attack situations. In this paper, we focus on such a task.

III. EXPERIMENTAL SETUP

A. KDD Cup 1999 Data Set

The data set provided for the 1999 KDD Cup was originally prepared by MIT Lincoln labs for the 1998 Defense Advanced Research Projects Agency (DARPA) Intrusion Detection Evaluation Program, with the objective of evaluating research in intrusion detection, and it has become a benchmark data set for the evaluation of IDSs. Attacks fall into four main categories:

- Denial of service (DoS), where an attacker makes some computing or memory resource too busy or too full to handle legitimate requests, thus denying legitimate users access to a machine, e.g., SYN flood;
- Remote to local (R2L), where an attacker sends packets to a machine over a network, then exploits machine's vulnerability to illegally gain local access as a user, e.g., guessing password;
- User to root (U2R), where an attacker starts out with access to a normal user account on the system and is able to exploit vulnerability to gain root access to the system, e.g., buffer overflows;
- Probing, where an attacker scans a network to gather information or find known vulnerabilities, e.g., port scanning.

The KDD Cup 1999 data set has a huge number of duplicated records as shown in Table I on the next page. This data set lies with the distribution of its five classes. The DoS attack comprises 79.24% in training and 73.90% in testing, respectively. Meanwhile, normal connection consists of 19.69% in training and 19.48% in testing, respectively. This imbalance makes it very difficult to train classifiers on the training set, and results in having extremely poor detection rates. In this paper, we use a subset of the original data set which consists of distinct records only.

Table I: Data Distribution and Ratio in the Original Data Set

Class	Training		Testing	
	Amount of Data	Ratio (%)	Amount of Data	Ratio (%)
Normal	97,278	19.69	60,593	19.48
DoS	391,458	79.24	229,853	73.90
R2L	1,126	0.23	16,189	5.20
U2R	52	0.01	228	0.07
Probing	4,107	0.83	4,166	1.34
Total	494,021	100	311,029	100

B. Data Preprocessing

Data preprocessing has to be undertaken before we could do any experiment. It is carried out in two steps. The first step involves mapping symbolic-valued attributes to numeric-valued attributes. The second step implements non-zero numerical features.

The redundancy in the KDD Cup 1999 data set is surprisingly high. By deleting the repeated data, the size of the data set is reduced from 311,029 to 77,291 as shown in Table II.

Table II: Data Distribution and Ratio in the Reduced Data Set

Class	Amount of Data	Ratio (%)
Normal	47,913	61.99
DoS	23,568	30.49
R2L	2,913	9.77
U2R	215	0.27
Probing	2,682	3.47
Total	77,291	100

IV. RESULTS

Standard measures which were developed for evaluating IDSs include detection rate (DTR), false positive rate (FPR), and overall accuracy (OA). These three performance metrics may be defined as follows [7]:

$$\mathrm{DTR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \times 100\%, \tag{4}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN+FP}} \times 100\%, \tag{5}$$

$$\mathrm{OA} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \times 100\%, \tag{6}$$

where TP, TN, FP, and FN are the numbers of malicious executables correctly classified as malicious, benign programs correctly classified as benign, benign programs falsely classified as malicious, and malicious executables falsely classified as benign, respectively. An IDS requires high DTR, low FPR, and high OA.

The block diagram of our experiment is shown in Fig.1 on the next page. We consider all attacks as a whole, and all 41 features are shown in Table III on the next page. We choose $c = 2$, $p = 230$, and $r = 1.5$. The DTRs, FPRs, and OAs for the FCM and the 2PID are shown in Table IV. Obviously, the 2PID yields a higher DTR and a greater OA than the FCM, while the FPRs for both methods are equal. This demonstrates the effectiveness of our proposed method.

Table IV: Result of the Experiment

Method	Detection Rate (DTR)	False Positive Rate (FPR)	Overall Accuracy (OA)
FCM	81.07%	2.50%	91.26%
2PID	90.35%	2.50%	94.78%

V. CONCLUSION

This paper has proposed a two-phase approach to intrusion detection, where the KDD Cup 1999 data set has been considered. The experimental results have shown that the proposed method is superior to the FCM. In future work, we plan to include a feature selection algorithm to help build efficient and practical intrusion detection.

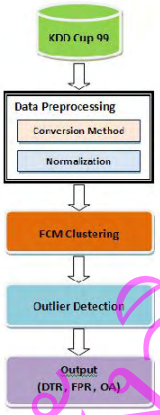


Fig. 1 Block diagram of the experiment

Table III: Feature Description of the KDD Cup 1999 Data Set

Feature Name	Description	Type
1. Duration	Length (number of seconds) of the connection	Continuous
2. Protocol type	Type of the protocol, e.g. tcp, udp, etc.	Discrete
3. Service	Network service on the destination, e.g., http, telnet, etc.	Discrete
4. Flag	Normal or error status of the connection	Discrete
5. Src_bytes	Number of data bytes from source to destination	Continuous
6. Dst_bytes	Number of data bytes from destination to source	Continuous
7. Land	1 if connection is from/to the same host/port; 0 otherwise	Discrete
8. Wrong_fragment	Number of "wrong" fragments	Continuous
9. Urgent	Number of urgent packets	Continuous
10. Hot	Number of "hot" indicators	Continuous
11. Num_failed_logins	Number of failed login attempts	Continuous
12. Logged_in	1 if successfully logged in; 0 otherwise	Discrete
13. Num_compromised	Number of "compromised" conditions	Continuous
14. Root_shell	1 if root shell is obtained; 0 otherwise	Continuous
15. Su_attempted	1 if "su_root" command attempted; 0 otherwise	Continuous
16. Num_root	Number of "root" accesses	Continuous
17. Num_file_creations	Number of file creation operations	Continuous
18. Num_shells	Number of shell prompts	Continuous
19. Num_access_files	Number of operations on access control files	Continuous
20. Num_otbound_cmds	Number of outbound commands in an ftp session	Continuous
21. Is_host_login	1 if the login belongs to the "hot" list; 0 otherwise	Discrete
22. Is_guest_login	1 if the login is a "guest" login; 0 otherwise	Discrete
23. Count	Number of connections to the same host as the current connection in the past two seconds	Continuous
24. Srv_count	Number of connections to the same service as the current connection in the past two seconds	Continuous
25. Serror_rate	% of connections that have "SYN" errors	Continuous
26. Srv_serror_rate	% of connections that have "SYN" errors	Continuous
27. Rerror_rate	% of connections that have "REJ" errors	Continuous

Table III (Continued): Feature Description of the KDD Cup 1999 Data Set

Feature Name	Description	Type
28. Srv_rerror_rate	% of connections that have "REJ" errors	Continuous
29. Same_srv_rate	% of connections to the same service	Continuous
30. Diff_srv_rate	% of connections to different services	Continuous
31. Srv_diff_host_rate	% of connections to different hosts	Continuous
32. Dst_host_count	Count for destination host	Continuous
33. Dst_host_srv_count	Srv_count for destination host	Continuous
34. Dst_host_same_srv_rate	Same_srv_rate for destination host	Continuous
35. Dst_host_diff_srv_rate	Dif_srv_rate for destination host	Continuous
36. Dst_host_same_srv_port_rate	Same_src_port_rate for destination host	Continuous
37. Dst_host_srv_diff_host_rate	Diff_host_rate for destination host	Continuous
38. Dst_host_serror_rate	Serror_rate for destination host	Continuous
39. Dst_host_srv_serror_rate	Srv_serror_rate for destination host	Continuous
40. Dst_host_rerror_rate	Rerror_rate for destination host	Continuous
41. Dst_host_srv_rerror_rate	Srv_serror_rate for destination host	Continuous

REFERENCES

[1] R. Bace and P. Mell, "Intrusion Detection Systems," NIST Special Publications on Intrusion Detection Systems. SP 800.31, Nov. 2001.

[2] KDD Data Set. (1999) [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[3] M. K. Pakhira, "A modified k-means algorithm to avoid empty clusters," International Journal of Recent Trends in Engineering, vol. 1, no. 1, pp. 220-226, May 2009.

[4] D. E. Denning, "An intrusion-detection model," IEEE Transactions on Software Engineering, vol. SE-13, no. 2, pp. 222-232, Feb. 1987.

[5] R. Jensen and Q. Shen, "Fuzzy-rough data reduction with ant colony optimization," Fuzzy Sets and Systems, vol. 149, pp. 5-20, 2005.

[6] M. Kantardzic, Data Mining: Concepts, Models, and Algorithms, New Jersey: IEEE Press, 2003.

[7] W. Chimphlee, M. N. M. Sap, A. H. Abdullah, S. Chimphlee, and S. Srinoy, "Anomaly detection of intrusion based on integration of rough sets and fuzzy c-means," Journal of Information Technology, vol. 17, no. 2, pp. 1-14, Dec. 2005.

Proceedings
2012 Tenth International Conference on ICT and Knowledge Engineering

November 21-23, 2012 Bangkok, Thailand

IEEE Catalog Number: **CFP1228H-PRT**
ISBN: **978-1-4673-2314-7**
ISSN: **2157-0981**
IEEE Conference: **#20935**

Jointly organized by

CONTENTS (Cont.)

Page

2012 Tenth International Conference on ICT and Knowledge Engineering

Classification via *k*-Means Clustering and Distance-Based Outlier Detection

Surasit Songma
Faculty of Information Technology
Rangsit University
Phathumthani, Thailand

Witcha Chimphlee
Faculty of Science and Technology
Suan Dusit Rajabhat University
Bangkok, Thailand

Kiattisak Maichalernnukul
Faculty of Information Technology
Rangsit University
Phathumthani, Thailand

Parinya Sanguansat
Faculty of Engineering and Technology
Panyapiwat Institute of Management
Nonthaburi, Thailand

***Abstract*—We propose a two-phase classification method. Specifically, in the first phase, a set of patterns (data) are clustered by the *k*-means algorithm. In the second phase, outliers are constructed by a distance-based technique and a class label is assigned to each pattern. The Knowledge Discovery Databases (KDD) Cup 1999 data set, which has been utilized extensively for development of intrusion detection systems, is used in our experiment. The results show that the proposed method is effective in intrusion detection.**

***Keywords*—Classification, *k*-means, intrusion detection, KDD Cup 1999 data set, outlier detection**

I. Introduction

As defined in [1], intrusion detection is the process of monitoring the events occurring in a computer network and analyzing them for signs of intrusions. Also, it is defined as attempts to compromise the confidentiality, integrity, availability, or to bypass the security mechanisms of a computer network. Anomaly intrusion detection systems (IDSs) aim at distinguishing an abnormal activity from an ordinary one.

The current state of computer networks is vulnerable; they are prone to an increasing number of attacks. These attacks are seldom previously seen. It is very hard to detect them before subsequent damage is done. Therefore, securing such a network from unwanted malicious traffic is of prime concern.

In this paper, a two-phase classification method is proposed and applied to intrusion detection where the Knowledge Discovery Databases (KDD) Cup 1999 data set [2] is considered.

The rest of the paper is organized as follows: Section II introduces the proposed method. Section III describes the experimental setup. Section IV provides the results, and Section V concludes the paper.

II. Method Description

In this section, we begin with a review of *k*-means clustering and distance-based outlier detection. Next, we present our proposed method.

A. *k-Means Clustering*

k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple way to classify a given data set through a certain number of clusters (*k* clusters assumed) fixed a priori [3]. This algorithm aims at minimizing an objective function, or, in our context, the following squared error function:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \quad (1)$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the corresponding cluster center c_j. This function is an indicator of the distance of the n data points from their respective cluster centers.

The *k*-means algorithm is composed of the following steps:

1) Initialize k center locations (i.e., $c_1,...,c_k$);
2) Assign each $x_i^{(j)}$ to its nearest cluster center c_j.
3) Update each cluster center c_j as the mean of all $x_i^{(j)}$ that have been assigned closest to it.
4) Recalculate the positions of the k centers.
5) Repeat steps 2, 3, and 4 until the centers no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

B. *Distance-Based Outlier Detection*

Outlier is defined as an observation that appears to be inconsistent with other observations in a data set. Many data-mining algorithms try to minimize the influence of outliers on the final model, or to eliminate them in the preprocessing phases. Outlier detection and potential removal from the data set can be described as a process of the selection of L out of N samples that are considerably dissimilar, exceptional, or inconsistent with respect to the remaining data.

Distance-based technique is a class of outlier-detection method. The basic computational complexity of this technique is the evaluation of distance measures between all samples in a given data set. Then, a sample in a data set $\{X_k\}$ is an outlier if

978-1-4673-2317-8/12/$31.00©2012 IEEE

at least a fraction p of the samples in $\{X_k\}$ lies at a distance greater than r. Clearly, the criterion for outlier detection is based on p and r. These two parameters may be given beforehand using knowledge about the data. Further details are in [3].

C. Proposed Method

Our classification method consists of two phases. In the first phase, a set of patterns are classified by k-means clustering. In the second phase, outliers are constructed by the distance-based technique, and a class label is assigned to each pattern.

Binary classification is the task of classifying the members of a given data set into two groups on the basis of whether they have some property or not. The binary classification task in the context of intrusion detection is to differentiate between normal connections and attack situations. In this paper, we focus on such a task.

III. Experimental Setup

A. KDD Cup 1999 Data Set

The data set provided for the 1999 KDD Cup was originally prepared by MIT Lincoln labs for the 1998 Defense Advanced Research Projects Agency (DARPA) Intrusion Detection Evaluation Program, with the objective of evaluating research in intrusion detection, and it has become a benchmark data set for the evaluation of IDSs. Attacks fall into four main categories:

- Denial of service (DoS), where an attacker makes some computing or memory resource too busy or too full to handle legitimate requests, thus denying legitimate users access to a machine, e.g., SYN flood;
- Remote to local (R2L), where an attacker sends packets to a machine over a network, then exploits machine's vulnerability to illegally gain local access as a user, e.g., guessing password;
- User to root (U2R), where an attacker starts out with access to a normal user account on the system and is able to exploit vulnerability to gain root access to the system, e.g., buffer overflows;
- Probing, where an attacker scans a network to gather information or find known vulnerabilities, e.g., port scanning.

The KDD Cup 1999 data set has a huge number of duplicated records as shown in Table I. This data set lies with the distribution of its five classes. The DoS attack comprises 79.24% in training and 73.90% in testing, respectively. Meanwhile, normal connection consists of 19.69% in training and 19.48% in testing, respectively. This imbalance makes it very difficult to train classifiers on the training set, and results in having extremely poor detection rates. In this paper, we use a subset of the original data set which consists of distinct records only.

TABLE I
Data Distribution and Ratio in The Original Data Set

Class	Training		Testing	
	Amount of Data	Ratio (%)	Amount of Data	Ratio (%)
Normal	97,278	19.69	60,593	19.48
DoS	391,458	79.24	229,853	73.90
R2L	1,126	0.23	16,189	5.20
U2R	52	0.01	228	0.07
Probing	4,107	0.83	4,166	1.34
Total	494,021	100	311,029	100

TABLE II
Data Distribution and Ratio in The Reduced Data Set

Class	Amount of Data	Ratio (%)
Normal	47,913	61.99
DoS	23,568	30.49
R2L	2,913	3.77
U2R	215	0.27
Probing	2,682	3.47
Total	77,291	100

B. Data Preprocessing

Data preprocessing has to be undertaken before we could do any experiment. It is carried out in two steps. The first step involves mapping symbolic-valued attributes to numeric-valued attributes. The second step implements non-zero numerical features.

The redundancy in the KDD Cup 1999 data set is surprisingly high. By deleting the repeated data, the size of data set is reduced from 311,029 to 77,291 as shown in Table II.

IV. Results

Standard measures which were developed for evaluating IDSs include detection rate (DTR), false positive rate (FPR), and overall accuracy (OA). These three performance metrics may be defined as follows [4]:

$$\mathrm{DTR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \quad 100\%, \tag{2}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN+FP}} \quad 100\%, \tag{3}$$

$$\mathrm{OA} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \times 100\%, \tag{4}$$

where TP, TN, FP, and FN are the numbers of malicious executables correctly classified as malicious, benign programs correctly classified as benign, benign programs falsely classified as malicious, and malicious executables falsely classified as benign, respectively. An IDS requires high DTR, low FPR, and high OA.

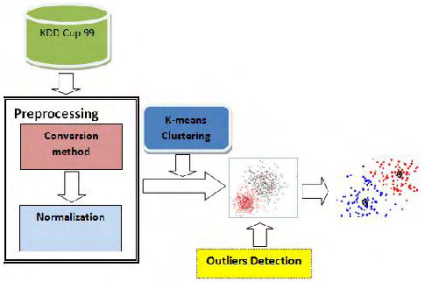


Fig. 1. Block diagram of the experiment

TABLE IV
RESULT OF EXPERIMENT

Method	DTR	FPR	OA
Support vector machine	71.6%	28.4%	86.1%
Rough set-fussy support vector machine	85.76%	14.24%	90%
Two-phase classification	91.35%	1.80%	95.59%

The block diagram of our experiment is shown in Fig. 1. We consider all attacks as a whole, and all 41 features are shown in Table III on the next page. In Table IV, we compare the DTR, FPR, and OA of our two-phase classification method with those of support vector machine and rough set-fussy support vector machine [5]. The two-phase classification method achieves the highest DTR and the lowest FPR.

V. CONCLUSION

This paper proposes a two-phase classification method. The experimental results show that the proposed method is not only able to achieve low FPR but also to yield high DTR for the IDS.

REFERENCES

[1] R. Bace and P. Mell, "Intrusion Detection Systems," NIST Special Publications on Intrusion Detection Systems. SP800-31, Nov. 2011.
[2] KDD data set. (online). http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
[3] M. Kantardzic, *Data Mining: Concepts, Models, and Algorithms*, New Jersey: IEEE Press, 2003.
[4] W. Chimphlee, M. N. M. Sap, A. H. Abdullah, S. Chimphlee, and S. Srinoy, "Anomaly detection of intrusion based on integration of rough sets and fuzzy *c*-means," *Journal of Information Technology*, vol. 17, no. 2, pp. 1-14, Dec. 2005.
[5] L. Li and K. Zhao, "A New Intrusion Detection System Based on Rough Set Theory and Fuzzy Support Vector Machine Fuzzy Support Vector Machine," in *Proceedings of International Workshop on Intelligent Systems and Applications*, pp. 1-5, 2011.

TABLE III
FEATURE DESCRIPTION OF THE KDD CUP 1999 DATA SET

No.	Feature Name	Description	Type
1	Duration	Length (number of seconds) of the connection	Continuous
2	Protocol type	Type of the protocol, e.g., tcp, udp, etc.	Discrete
3	Service	Network service on the destination, e.g., http, telnet, etc.	Discrete
4	Flag	Normal or error status of the connection	Discrete
5	Src_bytes	Number of data bytes from source to destination	Continuous
6	Dst_bytes	Number of data bytes from destination to source	Continuous
7	Land	1 if connection is from/to the same host/port; 0 otherwise	Discrete
8	Wrong_fragment	Number of "wrong" fragments	Continuous
9	Urgent	Number of urgent packets	Continuous
10	Hot	Number of "hot" indicators	Continuous
11	Num_failed_logins	Number of failed login attempts	Continuous
12	Logged_in	1 if successfully logged in; 0 otherwise	Discrete
13	Num_compromised	Number of "compromised" conditions	Continuous
14	Root_shell	1 if root shell is obtained; 0 otherwise	Continuous
15	Su_attempted	1 if "su root" command attempted; 0 otherwise	Continuous
16	Num_root	Number of "root" accesses	Continuous
17	Num_file_creations	Number of file creation operations	Continuous
18	Num_shells	Number of shell prompts	Continuous
19	Num_access_files	Number of operations on access control files	Continuous
20	Num_otbound_cmds	Number of outbound commands in an ftp session	Continuous
21	Is_host_login	1 if the login belongs to the "hot" list; 0 otherwise	Discrete
22	Is_guest_login	1 if the login is a "guest" login; 0 otherwise	Discrete
23	Count	Number of connections to the same host as the current connection in the past two seconds	Continuous
24	Srv_count	Number of connections to the same service as the current connection in the past two seconds	Continuous
25	Serror_rate	% of connections that have "SYN" errors	Continuous
26	Srv_serror_rate	% of connections that have "SYN" errors	Continuous
27	Rerror_rate	% of connections that have "REJ" errors	Continuous
28	Srv_rerror_rate	% of connections that have "REJ" errors	Continuous
29	Same_srv_rate	% of connections to the same service	Continuous
30	Diff_srv_rate	% of connections to different services	Continuous
31	Srv_diff_host_rate	% of connections to different hosts	Continuous
32	Dst_host_count	Count for destination host	Continuous
33	Dst_host_srv_count	Srv_count for destination host	Continuous
34	Dst_host_same_srv_rate	Same_srv_rate for destination host	Continuous
35	Dst_host_diff_srv_rate	Dif_srv_rate for destination host	Continuous
36	Dst_host_same_src_port_rate	Same_src_port_rate for destination host	Continuous
37	Dst_host_srv_diff_host_rate	Diff_host_rate for destination host	Continuous
38	Dst_host_serror_rate	Serror_rate for destination host	Continuous
39	Dst_host_srv_serror_rate	Srv_serror_rate for destination host	Continuous
40	Dst_host_rerror_rate	Rerror_rate for destination host	Continuous
41	Dst_host_srv_rerror_rate	Srv_serror_rate for destination host	Continuous

ภาคผนวก ข

ผลการแบ่งข้อมูลฝึกอบรมในชุดข้อมูล KDD Cup 1999 โดยใช้เทคนิคเอ็มดีแอล

ตารางที่ 1 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 1 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.99	I1
2	1 - 1.99	I2
3	2 - 3.99	I3
4	4 - 4.99	I4
5	5 - 11.99	I5
6	12 - 13.99	I6
7	14 - 30.99	I7
8	31 - 197.99	I8
9	198 - 1,030.99	I9
10	1,031 - 4,989.99	I10
11	4,990 - 5,024.99	I11
12	5,025 - 5,063.99	I12
13	5,064 - 5,084.99	I13
14	5,085 - 10,038.99	I14
15	10,039 - 10,133.99	I15
16	10,134 - 12,704.99	I16
17	12,705 - 14,681.99	I17
18	14,682 - 15,126.99	I18
19	15,127 - 15,167.99	I19
20	15,168 - 20,939.99	I20
21	20,940 - 30,189.99	I21
22	30,190 - 58,328.99	I22
23	58,329.00	I23

ตารางที่ 2 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 2 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	1	I1
2	2	I2
3	3	I3

ตารางที่ 3 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 3 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	1 - 1.99	I1
2	2 - 2.99	I2
3	3 - 9.99	I3
4	10 - 10.99	I4
5	11 - 11.99	I5
6	12 - 12.99	I6
7	13 - 13.99	I7
8	14 - 15.99	I8
9	16 - 16.99	I9
10	17 - 17.99	I10
11	18 - 18.99	I11
12	19 - 19.99	I12
13	20 - 20.99	I13
14	21 - 21.99	I14
15	22 - 23.99	I15
16	24 - 37.99	I16
17	38 - 39.99	I17
18	40 - 40.99	I18

ตารางที่ 3 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 3 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
19	41 - 41.99	I19
20	42 - 42.99	I20
21	43 - 43.99	I21
22	44 - 46.99	I22
23	47 - 47.99	I23
24	48 - 52.99	I24
25	53 - 53.99	I25
26	54 - 54.99	I26
27	55 - 55.99	I27
28	56 - 59.99	I28
29	60 - 66.99	I29
30	67	I30

ตารางที่ 4 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 4 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 1.99	I1
2	2 - 3.99	I2
3	4 - 4.99	I3
4	5 - 7.99	I4
5	8 - 8.99	I5
6	9 - 10.99	I6
7	11	I7

ตารางที่ 5 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 5 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.99	I1
2	1 - 27.99	I2
3	28 - 114.99	I3
4	115 - 124.99	I4
5	125 - 125.99	I5
6	126 - 141.99	I6
7	142 - 206.99	I7
8	207 - 245.99	I8
9	246 - 333.99	I9
10	334 - 519.99	I10
11	520 - 831.99	I11
12	832 - 1,011.99	I12
13	1,012 - 1,031.99	I13
14	1,032 - 1,137.99	I14
15	1,138 - 1,189.99	I15
16	1,190 - 1,221.99	I16
17	1,222 - 1,243.99	I17
18	1,244 - 1,268.99	I18
19	1,269 - 1,478.99	I19
20	1,479 - 1,479.99	I20
21	1,480 - 7,021.99	I21
22	7,022 - 7,044.99	I22
23	7,045 - 16,828.99	I23
24	16,829 - 40,493.99	I24
25	40,494 - 52,292.99	I25
26	52,293 - 54,539.99	I26

ตารางที่ 5 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 5 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
27	54,540 - 2,194,618.99	I27
28	2,194,619 - 693,375,639.99	I28
29	693,375,640	I29

ตารางที่ 6 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 6 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 31.99	I1
2	32 - 177.99	I2
3	178 - 178.99	I3
4	179 - 196.99	I4
5	197 - 1183.99	I5
6	1184 - 1207.99	I6
7	1208 - 2186.99	I7
8	2187 - 2200.99	I8
9	2201 - 2442.99	I9
10	2443 - 2444.99	I10
11	2445 - 2446.99	I11
12	2447 - 2452.99	I12
13	2453 - 7297.99	I13
14	7298 - 7299.99	I14
15	7300 - 8312.99	I15
16	8313 - 8314.99	I16
17	8315 - 460349.99	I17
18	460350 - 5134217.99	I18
19	5134218 - 5155467.99	I19
20	5155468	I20

ตารางที่ 7 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 7 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0	I1
2	1	I2

ตารางที่ 8 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 8 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0-0.99	I1
2	1-1.99	I2
3	2.00-3.00	I3

ตารางที่ 9 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 9 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0-2.99	I1
2	3	I2

ตารางที่ 10 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 10 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.99	I1
2	1 - 1.99	I2
3	2 - 2.99	I3
4	3 - 23.99	I4
5	24 - 27.99	I5
6	28 - 29.99	I6
7	30	I7

ตารางที่ 11 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 11 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0-0.99	I1
2	1-3.99	I2
3	4-4.99	I3
4	5	I4

ตารางที่ 12 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 12 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0	I1
2	1	I2

ตารางที่ 13 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 13 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.99	I1
2	1 - 1.99	I2
3	2 - 5.99	I3
4	6 - 12.99	I4
5	13 - 883.99	I5
6	884	I6

ตารางที่ 14 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 14 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0	I1
2	1	I2

ตารางที่ 15 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 15 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 1.99	I1
2	2	I2

ตารางที่ 16 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 16 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0-992.99	I1
2	993	I2

ตารางที่ 17 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 17 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0-27.99	I1
2	28	I2

ตารางที่ 18 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 18 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0-1.99	I1
2	2	I2

ตารางที่ 19 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 19 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.99	I1
2	1 - 1.99	I2
3	2 - 7.99	I3
4	8	I4

ตารางที่ 20 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 20 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0	I1

ตารางที่ 21 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 21 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0	I1

ตารางที่ 22 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 22 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0	I1
2	1	I2

ตารางที่ 23 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 23 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.99	I1
2	1 - 1.99	I2
3	2 - 2.99	I3
4	3 - 3.99	I4
5	4 - 4.99	I5
6	5 - 5.99	I6
7	6 - 6.99	I7
8	7 - 7.99	I8
9	8 - 8.99	I9
10	9 - 9.99	I10
11	10 - 10.99	I11
12	11 - 11.99	I12
13	12 - 12.99	I13

ตารางที่ 23 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 23 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
14	13 - 13.99	I14
15	14 - 14.99	I15
16	15 - 15.99	I16
17	16 - 16.99	I17
18	17 - 17.99	I18
19	18 - 18.99	I19
20	19 - 19.99	I20
21	20 - 20.99	I21
22	21 - 21.99	I22
23	22 - 22.99	I23
24	23 - 23.99	I24
25	24 - 24.99	I25
26	25 - 25.99	I26
27	26 - 26.99	I27
28	27 - 27.99	I28
29	28 - 28.99	I29
30	29 - 29.99	I30
31	30 - 30.99	I31
32	31 - 31.99	I32
33	32 - 32.99	I33
34	33 - 33.99	I34
35	34 - 34.99	I35
36	35 - 35.99	I36
37	36 - 36.99	I37
38	37 - 37.99	I38
39	38 - 38.99	I39

ตารางที่ 23 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 23 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
40	39 - 40.99	I40
41	41 - 41.99	I41
42	42 - 42.99	I42
43	43 - 43.99	I43
44	44 - 44.99	I44
45	45 - 45.99	I45
46	46 - 46.99	I46
47	47 - 47.99	I47
48	48 - 48.99	I48
49	49 - 76.99	I49
50	77 - 85.99	I50
51	86 - 100.99	I51
52	101 - 148.99	I52
53	149 - 180.99	I53
54	181 - 200.99	I54
55	201 - 301.99	I55
56	302 - 325.99	I56
57	326 - 430.99	I57
58	431 - 507.99	I58
59	508 - 508.99	I59
60	509 - 509.99	I60
61	510 - 510.99	I61
62	511	I62

ตารางที่ 24 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 24 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.99	I1
2	1 - 1.99	I2
3	2 - 2.99	I3
4	3 - 3.99	I4
5	4 - 4.99	I5
6	5 - 5.99	I6
7	6 - 6.99	I7
8	7 - 7.99	I8
9	8 - 8.99	I9
10	9 - 9.99	I10
11	10 - 10.99	I11
12	11 - 11.99	I12
13	12 - 12.99	I13
14	13 - 13.99	I14
15	14 - 14.99	I15
16	15 - 15.99	I16
17	16 - 16.99	I17
18	17 - 17.99	I18
19	18 - 18.99	I19
20	19 - 19.99	I20
21	20 - 20.99	I21
22	21 - 21.99	I22
23	22 - 22.99	I23
24	23 - 23.99	I24
25	24 - 24.99	I25
26	25 - 39.99	I26

ตารางที่ 24 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 24 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
27	40 - 52.99	I27
28	53 - 63.99	I28
29	64 - 99.99	I29
30	100 - 162.99	I30
31	163 - 324.99	I31
32	325 - 510.99	I32
33	511	I33

ตารางที่ 25 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 25 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.009	I1
2	0.01 - 0.019	I2
3	0.02 - 0.019	I3
4	0.02 - 0.029	I4
5	0.03 - 0.039	I5
6	0.04 - 0.049	I6
7	0.05 - 0.059	I7
8	0.06 - 0.069	I8
9	0.07 - 0.079	I9
10	0.08 - 0.089	I10
11	0.09 - 0.099	I11
12	0.1 - 0.109	I12
13	0.11 - 0.119	I13
14	0.12 - 0.159	I14
15	0.16 - 0.169	I15
16	0.17 - 0.189	I16

ตารางที่ 25 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 25 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
17	0.19 - 0.199	I17
18	0.2 - 0.239	I18
19	0.24 - 0.249	I19
20	0.25 - 0.319	I20
21	0.32 - 0.329	I21
22	0.33 - 0.439	I22
23	0.44 - 0.499	I23
24	0.5 - 0.619	I24
25	0.62 - 0.809	I25
26	0.81 - 0.989	I26
27	0.99 - 0.999	I27
28	1	I28

ตารางที่ 26 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 26 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.139	I1
2	0.14 - 0.329	I2
3	0.33 - 0.499	I3
4	0.5 - 0.509	I4
5	0.51 - 0.579	I5
6	0.58 - 0.669	I6
7	0.67 - 0.949	I7
8	0.95 - 0.999	I8
9	1	I9

ตารางที่ 27 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 27 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.019	I1
2	0.02 - 0.029	I2
3	0.03 - 0.039	I3
4	0.04 - 0.139	I4
5	0.14 - 0.169	I5
6	0.17 - 0.189	I6
7	0.19 - 0.199	I7
8	0.2 - 0.239	I8
9	0.24 - 0.249	I9
10	0.25 - 0.429	I10
11	0.43 - 0.499	I11
12	0.5 - 0.749	I12
13	0.75 - 0.989	I13
14	0.99 - 0.999	I14
15	1	I15

ตารางที่ 28 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 28 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.019	I1
2	0.02 - 0.029	I2
3	0.03 - 0.039	I3
4	0.04 - 0.139	I4
5	0.14 - 0.169	I5
6	0.17 - 0.199	I6
7	0.2 - 0.219	I7
8	0.22 - 0.249	I8
9	0.25 - 0.499	I9

ตารางที่ 28 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 28 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
10	0.5 - 0.669	I10
11	0.67 - 0.999	I11
12	1	I12

ตารางที่ 29 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 29 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.009	I1
2	0.01 - 0.099	I2
3	0.1 - 0.109	I3
4	0.11 - 0.139	I4
5	0.14 - 0.199	I5
6	0.2 - 0.239	I6
7	0.24 - 0.249	I7
8	0.25 - 0.319	I8
9	0.32 - 0.329	I9
10	0.33 - 0.389	I10
11	0.39 - 0.399	I11
12	0.4 - 0.489	I12
13	0.49 - 0.499	I13
14	0.5 - 0.589	I14
15	0.59 - 0.599	I15
16	0.6 - 0.649	I16
17	0.65 - 0.669	I17
18	0.67 - 0.749	I18
19	0.75 - 0.939	I19
20	0.94 - 0.959	I20
21	0.96 - 0.969	I21
22	0.97 - 0.979	I22

ตารางที่ 29 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 29 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
23	0.98 - 0.989	I23
24	0.99 - 0.999	I24
25	1	I25

ตารางที่ 30 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 30 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.009	I1
2	0.01 - 0.019	I2
3	0.02 - 0.029	I3
4	0.03 - 0.039	I4
5	0.04 - 0.049	I5
6	0.05 - 0.089	I6
7	0.09 - 0.189	I7
8	0.19 - 0.199	I8
9	0.2 - 0.209	I9
10	0.21 - 0.249	I10
11	0.25 - 0.279	I11
12	0.28 - 0.499	I12
13	0.5 - 0.559	I13
14	0.56 - 0.599	I14
15	0.6 - 0.639	I15
16	0.64 - 0.669	I16
17	0.67 - 0.729	I17
18	0.73 - 0.749	I18
19	0.75 - 0.799	I19
20	0.8 - 0.889	I20

ตารางที่ 30 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 30 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
21	0.89 - 0.989	I21
22	0.99 - 0.999	I22
23	1	I23

ตารางที่ 31 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 31 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.029	I1
2	0.03 - 0.039	I2
3	0.04 - 0.049	I3
4	0.05 - 0.069	I4
5	0.07 - 0.079	I5
6	0.08 - 0.329	I6
7	0.33 - 0.399	I7
8	0.4 - 0.499	I8
9	0.5 - 0.669	I9
10	0.67 - 0.749	I10
11	0.75 - 0.999	I11
12	1	I12

ตารางที่ 32 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 32 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.99	I1
2	1 - 1.99	I2
3	2 - 2.99	I3
4	3 - 3.99	I4
5	4 - 4.99	I5
6	5 - 5.99	I6

ตารางที่ 32 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 32 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
7	6 - 6.99	I7
8	7 - 7.99	I8
9	8 - 10.99	I9
10	11 - 15.99	I10
11	16 - 51.99	I11
12	52 - 72.99	I12
13	73 - 103.99	I13
14	104 - 127.99	I14
15	128 - 172.99	I15
16	173 - 214.99	I16
17	215 - 253.99	I17
18	254 - 254.99	I18
19	255	I19

ตารางที่ 33 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 33 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.99	I1
2	1 - 1.99	I2
3	2 - 2.99	I3
4	3 - 3.99	I4
5	4 - 4.99	I5
6	5 - 5.99	I6
7	6 - 6.99	I7
8	7 - 7.99	I8
9	8 - 8.99	I9
10	9 - 9.99	I10
11	10 - 10.99	I11
12	11 - 11.99	I12

ตารางที่ 33 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 33 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
13	12 - 12.99	I13
14	13 - 13.99	I14
15	14 - 14.99	I15
16	15 - 15.99	I16
17	16 - 16.99	I17
18	17 - 17.99	I18
19	18 - 18.99	I19
20	19 - 19.99	I20
21	20 - 20.99	I21
22	21 - 21.99	I22
23	22 - 22.99	I23
24	23 - 23.99	I24
25	24 - 24.99	I25
26	25 - 48.99	I26
27	49 - 67.99	I27
28	68 - 99.99	I28
29	100 - 143.99	I29
30	144 - 150.99	I30
31	151 - 152.99	I31
32	153 - 222.99	I32
33	223 - 243.99	I33
34	244 - 253.99	I34
35	254 - 254.99	I35
36	255	I36

ตารางที่ 34 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 34 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.009	I1
2	0.01 - 0.019	I2
3	0.02 - 0.029	I3
4	0.03 - 0.039	I4
5	0.04 - 0.049	I5
6	0.05 - 0.059	I6
7	0.06 - 0.069	I7
8	0.07 - 0.079	I8
9	0.08 - 0.089	I9
10	0.09 - 0.099	I10
11	0.1 - 0.229	I11
12	0.23 - 0.239	I12
13	0.24 - 0.269	I13
14	0.27 - 0.399	I14
15	0.4 - 0.569	I15
16	0.57 - 0.819	I16
17	0.82 - 0.869	I17
18	0.87 - 0.989	I18
19	0.99 - 0.999	I19
20	1	I20

ตารางที่ 35 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 35 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.009	I1
2	0.01 - 0.019	I2
3	0.02 - 0.029	I3
4	0.03 - 0.039	I4

ตารางที่ 35 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 35 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
5	0.04 - 0.049	I5
6	0.05 - 0.059	I6
7	0.06 - 0.069	I7
8	0.07 - 0.079	I8
9	0.08 - 0.089	I9
10	0.09 - 0.099	I10
11	0.1 - 0.149	I11
12	0.15 - 0.509	I12
13	0.51 - 0.519	I13
14	0.52 - 0.639	I14
15	0.64 - 0.669	I15
16	0.67 - 0.679	I16
17	0.68 - 0.819	I17
18	0.82 - 0.899	I18
19	0.9 - 0.909	I19
20	0.91 - 0.989	I20
21	0.99 - 0.999	I21
22	1	I22

ตารางที่ 36 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 36 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.009	I1
2	0.01 - 0.019	I2
3	0.02 - 0.029	I3
4	0.03 - 0.139	I4
5	0.14 - 0.149	I5

ตารางที่ 36 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 36 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
6	0.15 - 0.159	I6
7	0.16 - 0.169	I7
8	0.17 - 0.179	I8
9	0.18 - 0.189	I9
10	0.19 - 0.199	I10
11	0.2 - 0.209	I11
12	0.21 - 0.239	I12
13	0.24 - 0.249	I13
14	0.25 - 0.259	I14
15	0.26 - 0.319	I15
16	0.32 - 0.329	I16
17	0.33 - 0.339	I17
18	0.34 - 0.389	I18
19	0.39 - 0.489	I19
20	0.49 - 0.499	I20
21	0.5 - 0.509	I21
22	0.51 - 0.699	I22
23	0.7 - 0.859	I23
24	0.86 - 0.979	I24
25	0.98 - 0.999	I25
26	1	I26

ตารางที่ 37 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 37 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.009	I1
2	0.01 - 0.039	I2

ตารางที่ 37 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 37 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
3	0.04 - 0.049	I3
4	0.05 - 0.069	I4
5	0.07 - 0.249	I5
6	0.25 - 0.259	I6
7	0.26 - 0.499	I7
8	0.5 - 0.519	I8
9	0.52 - 0.999	I9
10	1	I10

ตารางที่ 38 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 38 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.009	I1
2	0.01 - 0.019	I2
3	0.02 - 0.029	I3
4	0.03 - 0.039	I4
5	0.04 - 0.069	I5
6	0.07 - 0.089	I6
7	0.09 - 0.099	I7
8	0.1 - 0.319	I8
9	0.32 - 0.329	I9
10	0.33 - 0.499	I10
11	0.5 - 0.969	I11
12	0.97 - 0.989	I12
13	0.99 - 0.999	I13
14	1	I14

ตารางที่ 39 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 39 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.009	I1
2	0.01 - 0.019	I2
3	0.02 - 0.029	I3
4	0.03 - 0.039	I4
5	0.04 - 0.079	I5
6	0.08 - 0.479	I6
7	0.48 - 0.669	I7
8	0.67 - 0.999	I8
9	1	I9

ตารางที่ 40 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 40 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.009	I1
2	0.01 - 0.019	I2
3	0.02 - 0.029	I3
4	0.03 - 0.039	I4
5	0.04 - 0.049	I5
6	0.05 - 0.059	I6
7	0.06 - 0.079	I7
8	0.08 - 0.139	I8
9	0.14 - 0.509	I9
10	0.51 - 0.519	I10
11	0.52 - 0.529	I11
12	0.53 - 0.629	I12
13	0.63 - 0.839	I13
14	0.84 - 0.909	I14

ตารางที่ 40 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 40 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
15	0.91 - 0.919	I15
16	0.92 - 0.949	I16
17	0.95 - 0.959	I17
18	0.96 - 0.999	I18

ตารางที่ 41 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 41 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.009	I1
2	0.01 - 0.019	I2
3	0.02 - 0.029	I3
4	0.03 - 0.039	I4
5	0.04 - 0.049	I5
6	0.05 - 0.059	I6
7	0.06 - 0.079	I7
8	0.08 - 0.139	I8
9	0.14 - 0.289	I9
10	0.29 - 0.489	I10
11	0.49 - 0.499	I11
12	0.5 - 0.849	I12
13	0.85 - 0.909	I13
14	0.91 - 0.919	I14
15	0.92 - 0.939	I15
16	0.94 - 0.949	I16
17	0.95 - 0.959	I17
18	0.96 - 0.999	I18
19	1	I19

ภาคผนวก ก

ผลงานตีพิมพ์เผยแพร่ในการประชุมวิชาการนานาชาติและวารสารวิชาการนานาชาติ

ผลงานตีพิมพ์เผยแพร่ในการประชุมวิชาการนานาชาติและวารสารวิชาการนานาชาติ

S. Songma, W. Chimphlee, K. Maichalernnukul and P. Sanguansat. "Intrusion Detection through Rule Induction Analysis." Advances in information Sciences and Service Sciences, vol.5 (2013), 187-194.

S. Songma, W. Chimphlee, K. Maichalernnukul and P. Sanguansat. "Implementation of Fuzzy *c*-Means and Outlier Detection for Intrusion Detection with KDD Cup 1999 Data Set." International Journal of Engineering Research and Development. Vol. 2 (2012), 44-48.

S. Songma, W. Chimphlee, K. Maichalernnukul and P. Sanguansat. "Classification via *k*-Means Clustering and Distance-Based Outlier Detection." In Proceeding of 2012 Tenth International Conference on ICT and Knowledge Engineering." (2012), 125-128.

Contents:

Intrusion Detection through Rule Induction Analysis

[1] Surasit Songma, [2] Witcha Chimphlee, [3] Kiattisak Maichalernnukul, [4] Parinya Sanguansat
[*1] *Faculty of Information Technology, Rangsit University, Phathumthani, Thailand, surasit.songma@gmail.com*
[2] *Faculty of Science and Technology, Suan Dusit Rajabhat University, Bangkok, Thailand, witcha.chi@gmail.com*
[3] *Faculty of Information Technology, Rangsit University, Phathumthani, Thailand, kiattisak.m@rsu.ac.th*
[4] *Faculty of Engineering and Technology, Panyapiwat Institute of Management, Nonthaburi, Thailand, sanguansat@yahoo.com*

Abstract

In this paper, we present intrusion detection by means of rule induction analysis. Specifically, the rule induction analysis is used to generate signatures from the Knowledge Discovery Databases (KDD) Cup 1999 testing data set and find a set of rules that satisfy some predefined criteria. The results show that this kind of intrusion detection is not only able to achieve extremely high detection rate (99.99%) but also to yield very low false positive rate (1.65%).

Keywords: *Intrusion Detection, KDD Cup 1999 Data Set, Rule Induction Analysis.*

1. Introduction

Intrusion detection is the process of monitoring the events occurring in a computer system or network and analysing them for signs of intrusions [1]. An Intrusion Detection System (IDS) is a combination of software and hardware that attempt to perform intrusion detection and raise alarm when possible intrusions are detected. The objective of the IDS is to act like a burglar alarm, which monitors the network and the connected systems to find evidence of intrusion.

Traditionally, intrusion detection approaches can be classified into two main categories: misuse detection and anomaly detection [2]. Misuse detection systems use *a prior* knowledge on attacks to look for attack traces. In other words, they detect intrusions by knowing that the misuse is the signature (rule) based system. Since signatures are associated with specific misuse behaviour, it is easy to determine the attack type. Anomaly detection systems adopt the opposite approach, which is, to know what is normal, and then find the deviations from the normal behaviour. These deviations are considered as anomalies or possible intrusions.

In this paper, we present misuse detection via rule induction analysis. First, a set of signatures, describing the characteristics of possible attacks, are generated from the Knowledge Discovery Databases (KDD) Cup 1999 data set [3], and the corresponding rules are stored in the IDS. Then, these rules are used to detect the incoming hostile traffic passing through the IDS.

The rest of the paper is organized as follows: Section 2 describes the rule induction analysis. Section 3 discusses the experimental results. Finally, Section 4 presents the conclusion.

2. Rule induction analysis

A rule-based IDS analyzes the traffic data passing through it and differentiates the traffic behaviors to be intrusive or normal. The rule-based IDS uses the rules stored in its knowledge base to detect and take actions when anomaly occurs in the traffic and/or undergoing some unauthorized activities [4]. The discovered rule are typically expressed in the form IF-THEN.

An example of an IF-THEN rule is shown in Fig. 1. Such a rule can have multiple conditions joined by conjunction (AND), disjunction (OR) or combination of both. In this figure, A, B, C, and D are conditions. If A, B, C, and D are true, then the action X is taken. Before proceeding, we give a brief overview of the traffic data used in this paper.

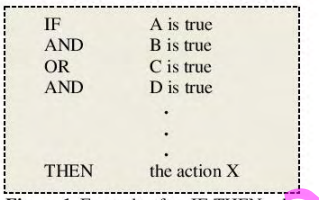


Figure 1. Example of an IF-THEN rule

The KDD Cup 1999 data set was originally prepared by MIT Lincoln labs for the 1998 Defense Advanced Research Projects Agency (DARPA) Intrusion Detection Evaluation Program, with the objective of evaluating research in intrusion detection. This data set consists of 494,021 training instances and 311,029 testing instances. Each instance represents a network connection, i.e., a sequence of network packets starting and ending at some well defined times, between which data flows to and from a source IP address to a target IP address under some well defined protocol. Such a connection instance is described by a 41-dimensional feature vector and labeled with respect to five classes:

- Normal;
- Denial of service (DoS), where an attacker makes some computing or memory resource too busy or too full to handle legitimate requests, thus denying legitimate users access to a machine, e.g., SYN flood;
- Remote to local (R2L), where an attacker sends packets to a machine over a network, then exploits machine's vulnerability to illegally gain local access as a user, e.g., guessing password;
- User to root (U2R), where an attacker starts out with access to a normal user account on the system and is able to exploit vulnerability to gain root access to the system, e.g., buffer overflows;
- Probing, where an attacker scans a network to gather information or find known vulnerabilities, e.g., port scanning.

For clarity, all features are listed in Table 1 on the last page, and their descriptive statistics are shown in Tables 2-5 where C and D denote Cont. and discrete, respectively. Examples of data contained in the features are as follows:

- *"icmp"*, *"tcp"*, and *"udp"* in protocol_type;
- *"IRC"*, *"ftp"*, *"pop_3"*, *"http"*, *"telnet"*, and *"login"* in service;
- *0* and *1* in logged_in;
- *0-255* in dst_host_count.

Table 2. Descriptive statistics of basic features of individual TCP connection

Feature name	Min	Max	Mean	Standard deviation
1. duration (C)	0	57,715	70.762	815.287
2. protocol type (D)	1	3	2.042	0.279
3. service (D)	1	65	29.714	12.218
4. flag (D)	1	11	7.005	3.195
5. src_bytes (C)	0	62,825,648	3,427.815	255,452.832
6. dst_bytes (C)	0	5,203,179	2,870.84	32,229.826
7. land (D)	0	1	0.0001	0.011
8. wrong_fragment (C)	0	3	0.003	0.078
9. urgent (C)	0	3	0.0002	0.020

Table 3. Descriptive statistics of content features within a connection suggested by domain knowledge

Feature name	Min	Max	Mean	Standard deviation
10. hot (C)	0	101	0.040	0.594
11. num_failed_logins (C)	0	4	0.007	0.084
12. logged_in (D)	0	1	0.585	0.493
13. num_compromised (C)	0	796	0.036	3.927
14. root_shell (C)	0	1	0.001	0.028
15. su_attempted (C)	0	2	0.00009	0.012
16. num_root (C)	0	878	0.034	4.343
17. num_file_creations (C)	0	100	0.004	0.387
18. num_shells (C)	0	5	0.0003	0.026
19. num_access_files (C)	0	4	0.003	0.059
20. num_otbound_cmds (C)	0	0	0.000	0.000
21. is_host_login (D)	0	1	0.0002	0.012
22. is_guest_login (D)	0	1	0.010	0.097

Table 4. Descriptive statistics of traffic features computed using a two-second time window

Feature name	Min	Max	Mean	Standard deviation
23. count (C)	0	511	68.089	105.097
24. srv_count (C)	0	511	18.336	53.464
25. serror_rate (C)	0	1	0.097	0.290
26. srv_serror_rate (C)	0	1	0.098	0.293
27. rerror_rate (C)	0	1	0.209	0.402
28. srv_rerror_rate (C)	0	1	0.209	0.403
29. same_srv_rate (C)	0	1	0.730	0.423
30. diff_srv_rate (C)	0	1	0.047	0.161
31. srv_diff_host_rate (C)	0	1	0.097	0.230

Table 5. Descriptive statistics of traffic features computed using a two-second time window from destination to host

Feature name	Min	Max	Mean	Standard deviation
32. dst_host_count (C)	0	255	178.380	99.731
33. dst_host_srv_count (C)	0	255	158.263	112.549
34. dst_host_same_srv_rate (C)	0	1	0.661	0.438
35. dst_host_diff_srv_rate (C)	0	1	0.046	0.134
36. dst_host_same_srv_port_rate (C)	0	1	0.077	0.221
37. dst_host_srv_diff_host_rate (C)	0	1	0.016	0.053
38. dst_host_serror_rate (C)	0	1	0.095	0.282
39. dst_host_srv_serror_rate (C)	0	1	0.096	0.288
40. dst_host_rerror_rate (C)	0	1	0.208	0.392
41. dst_host_srv_rerror_rate (C)	0	1	0.206	0.397

Table 6. Distribution and ratio of the KDD Cup 1999 testing data set

Class	Original		Subset	
	Amount of data	Ratio (%)	Amount of data	Ratio (%)
normal	60,593	19.48	47,913	61.99
DoS	229,853	73.90	23,568	30.49
R2L	16,189	5.20	2,913	3.77
U2R	228	0.07	215	0.27
probing	4,166	1.34	2,682	3.47
Total	311,029	100	77,291	100

In this paper, we consider the testing data set only. Its class distribution and ratio are shown in Table 6. Before we could do any experiment, data preprocessing has to be undertaken. We handle missing and incomplete data by involves mapping symbolic-valued attributes to numeric-valued attributes. Because the KDD Cup 1999 data set has a huge number of duplicated instances, we instead use its subset which comprises 77,291 distinct instances (see Table 6). Based on this subset, the corresponding class rules can be generated and exemplified as follows (where variable labels are shown in Table 1):

2.1 Normal Class Rule

IF *B = "icmp"* OR *B = "tcp"* OR *B = "udp"* AND *C = "IRCr"* OR *C = "auth"* OR *C = "domain_u"* OR *C = "eco_i"* OR *C = "ecr_i"* OR *C = "finger"* OR *C = "ftp"* OR *C = "ftp_data"* OR *C = "http"* OR *C = "link"* OR *C = "ntp_u"* OR *C = "other"* OR *C = "pop_3"* OR *C = "private"* OR *C = "remote_job"* OR *C = "smtp"* OR *C = "telnet"* OR *C = "time"* OR *C = "urp_i"* OR *C = "X11"* OR *C = "tcomp"* OR *C = "tftp_u"* OR *C = "tim_i"* AND *D ~= "S0"* AND *E >= 0* AND *G = 0* AND *H = 0* OR *H = 1* OR *H = 3* AND *I = 0* AND *K = 0* OR *K = 1* OR *K = 3* AND *L = 0* OR *L = 1* AND *M >= 0* AND *M <= 7* OR *M = 16* OR *M = 381* OR *M = 611* OR *M = 796* AND *N = 0* OR *N = 1* AND *O = 0* OR *O = 1* OR *O = 2* AND *P >= 0* AND *P <= 4* OR *P = 6* OR *P = 9* OR *P = 401* OR *P = 684* OR *P = 878* *Q >= 0* AND *Q <= 2* OR *Q = 12* OR *Q = 13* OR *Q = 30* OR *Q = 100* AND *R = 0* AND *S >= 0* AND *S <= 3* AND *T = 0* AND *U = 0* OR *U = 1* AND *V = 0* OR *V = 1* AND *Y = 0* OR *Y >= 0.02* AND *Y <= 0.17* OR *Y = 0.2* OR *Y = 0.25* OR *Y = 0.33* OR *Y = 0.5* OR *Y = 1* AND *AA = 0* OR *AA = 0.2* OR *AA = 0.33* OR *AA = 0.5* OR *AA = 0.98* OR *AA = 1* AND *AH >= 0* AND *AH <= 1* AND *AI ~= 0.34* AND *AJ >= 0* AND *AJ <= 1* THEN *Class_type* = ***"normal"***.

2.2 DoS Class Rule

IF *A = 0* AND *B = "tcp"* AND *C = "finger"* OR *C = "telnet"* AND *D = "S0"* AND *E = 0* AND *F = 0* AND *G = 1* AND *H = 0* AND *I = 0* AND *J = 0* AND *K = 0* AND *L = 0* AND *M = 0* AND *N = 0* AND *O = 0* AND *P = 0* AND *Q = 0* AND *R = 0* AND *S = 0* AND *T = 0* AND *U = 0* AND *V = 0* AND *W = 1* AND *X = 1* OR *X = 2* OR *X = 3* AND *Y = 1* AND *Z = 1* AND *AA = 0* AND *AB = 0* AND *AC = 1* AND *AD = 0* AND *AE = 0* OR *AE = 1* AND *AN = 0* OR *AN = 0.02* AND *AO = 0* THEN *Class_type* = ***"land"***.

2.3 R2L Class Rule

IF *A = 0* AND *B = "tcp"* AND *C = "http"* AND *D = "REJ"* OR *D = "SF"* AND *E = 0* OR *E = 51* AND *F = 0* OR *F = 8371* AND *G = 1* AND *H = 0* AND *I = 0* AND *J = 0* OR *J = 2* AND *K = 0* AND *L = 0* OR *L = 1* AND *M = 0* AND *N = 0* OR *N = 1* AND *O = 0* AND *P = 0* AND *Q = 0* AND *R = 0* AND *S = 0* OR *S = 1* AND *T = 0* AND *U = 0* AND *V = 0* AND *W = 1* AND *X = 1* AND *Y = 0* AND *Z = 0* AND *AA = 0* OR *AA = 1* AND *AB = 0* OR *AB = 1* AND *AC = 1* AND *AD = 0* AND *AE = 0* AND *AF = 17* OR *AF = 255* AND *AG = 17* OR *AG = 227* AND *AH = 0.89* OR *AH = 1* AND *AI = 0* OR *AI = 0.02* AND *AJ = 0* OR *AJ = 0.06* AND *AK = 0* AND *AL = 0* AND *AM = 0* AND *AN = 0* OR *AN = 0.06* AND *AO = 0* OR *AO = 0.06* THEN *Class_type* = ***"phf"***.

2.4 U2R Class Rule

IF *A = 75* OR *A = 84* AND *B = "tcp"* AND *C = "telnet"* AND *D = "SF"* AND *E = 227* AND *F = 1087* OR *F = 1089* AND *G = 0* AND *H = 0* AND *I = 0* AND *J = 2* AND *K = 0* AND *L = 1* AND *M = 1* AND *N = 1* AND *O = 0* AND *P = 0* AND *Q = 4* AND *R = 2* AND *S = 0* AND *T = 0* AND *U = 0* AND *V = 0* AND *W = 1* AND *X = 1* AND *Y = 0* AND *Z = 0* AND *AA = 0* AND *AB = 0* AND *AC = 1* AND *AD = 0* AND *AE = 0* AND *AF = 137* OR *AF = 255* AND *AG = 1* OR *AG = 101* AND *AH = 0* OR *AH = 0.74* AND *AI = 0.04* OR *AI = 0.07* AND *AJ = 0* OR *AJ = 0.01* AND *AK = 0* AND *AL = 0.14* OR *AL = 0.73* AND *AM = 0* OR *AM = 0.99* AND *AN = 0* OR *AN = 0.86* AND *AO = 0* THEN *Class_type* = ***"loadmodule"***.

2.5 Probing Class Rule

IF *A* = 0 AND *B = "tcp"* AND *C = "courier"* OR *C = "domain"* OR *C = "nnsp"* OR *C = "private"* OR *C = "uucp"* AND *D = "SH"* AND *E = 0* AND *F = 0* AND *G = 0* AND *H = 0* AND *I = 0* AND *J = 0* AND *K = 0* AND *L = 0* AND *M = 0* AND *N = 0* AND *O = 0* AND *P = 0* AND *Q = 0* AND *R = 0* AND *S = 0* AND *T = 0* AND *U = 0* AND *V = 0* AND *W = 1* OR *W = 2* AND *X = 1* AND *Y = 1* AND *Z = 1* AND *AA = 0* AND *AB = 0* AND *AC = 0.5* OR *AC = 1* AND *AD = 0* OR *AD = 1* AND *AE = 0* AND *AG= 1* OR *AG = 3* AND *AI = 0.99* OR *AI = 1* AND *AJ >= 0.98* AND *AK = 0* OR *AK = 0.67* AND *AL >=0.98* AND *AM = 1* AND *AN = 0* AND *AO = 0* THEN *Class_type* = ***"nmap"***.

Through the above rule induction analysis, the misuse detection for the data subset in Table 6 can be achieved. The overall approach is illustrated in Fig. 2

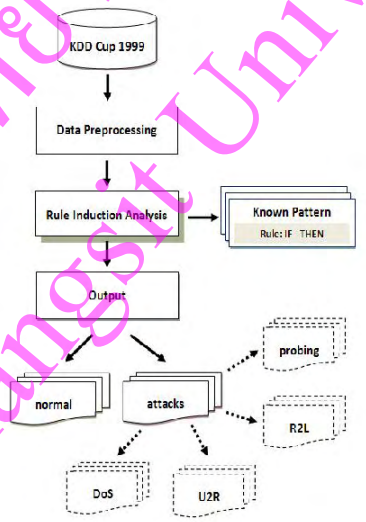


Figure 2. Proposed intrusion detection

3. Experimental results

The results of testing the proposed intrusion detection are summarized in Table 7. The top left entry in the confusion matrix shows that 47,122 of the actual "normal" instances were predicted to be normal by this entry. The last column indicates that in total 98.35% of the "normal" instances were recognized correctly. The bottom row shows that 99.99% of instances said to be normal were indeed "normal" in reality. The same analysis can be applied to the other four classes.

Table 7. Confusion matrix

	Predicted Class					
Actual Class	**normal**	**DoS**	**R2L**	**U2R**	**Probing**	**% Correct**
normal	47,122	456	300	0	35	98.35
DoS	0	23,316	1	0	251	98.93
R2L	1	0	2,912	0	0	99.97
U2R	0	0	0	215	0	100
probing	0	1	2	0	2,679	99.89
% Correct	99.99	98.08	90.58	100	90.35	

In Fig. 3, we compare the detection rate (DTR), false positive rate (FPR), and overall accuracy (OA) of our rule-based approach with those of other approaches in the literature, i.e., the support vector machine, rough set-fuzzy support vector machine [5], genetic algorithm [6], and two-phase classification [7]. The details of how DTR, FPR, and OA are calculated can be found in [8]. It is seen from this figure that the rule-based approach is the best as it achieves the highest DTR and OA, and the lowest FPR.

Figure 3. Comparison of DTR, FPR, and OA of the proposed rule-based approach with those of other approaches in the literature

4. Conclusion

In this paper, we have proposed a rule-based intrusion detection approach. The experimental results have shown that the proposed approach is very promising in terms of its DTR, FPR, and OA performances.

5. References

[1] Rebecca Bace and Peter Mell, "Intrusion Detection Systems", NIST Special Publications on Intrusion Detection Systems, SP 800-31, 2011.
[2] Chen Xiao-Ming, "Research on the Network Intrusion Detection Model with Neural Network Parameters Optimized by Ant-Colony Algorithm", Journal of Convergence Information Technology, vol. 8, no. 5, pp. 619-627, 2013.
[3] KDD Data Set. (1999) [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
[4] Aki P. F. Chan, Wing W. Y. Ng, Daniel S. Yeung, and Eric C. C. Tsang, "Refinement of Rule-Based Intrusion Detection System for Denial of Service Attacks by Support Vector Machine", In Proceedings of the International Conference on Machine Learning and Cybernetics, pp. 4252-4256, 2004.
[5] Lei Li and Ke-Nan Zhao, "A New Intrusion Detection System based on Rough Set Theory and Fuzzy Support Vector Machine", In Proceedings of the International Workshop on Intelligent Systems and Applications, pp. 1-5, 2011.
[6] Mohammea S. Hoque, Md. A. Mukit, and Md. A. N. Bikas, "An Implementation of Intrusion Detection System Using Genetic Algorithm", International Journal of Network Security and Its Applications, vol. 4, no. 2, pp. 109-120, 2012.
[7] Surasit Songma, Witcha Chimphlee, Kiattisak Maichalernnukul, and Parinya Sanguansat, "Classification via k-Means Clustering and Distance-Based Outlier Detection," In Proceedings of the International Conference on ICT and Knowledge Engineering, pp. 125-128, 2012.
[8] Witcha Chimphlee, Mohd N. Md Sap, Abdul H. Abdullah, Siriporn Chimphlee, and Surat Srinoy, "Anomaly Detection of Intrusion Based on Integration of Rough Sets and Fuzzy c-Means", Journal of Information Technology, vol. 17, no. 2, pp. 1-14, 2005.

Table 1. Feature description of the KDD Cup 1999 data set

Feature name	Description	Variable label
1. duration (C)	length (number of seconds) of the connection	A
2. protocol type (D)	type of the protocol, e.g., tcp, udp, etc.	B
3. service (D)	network service on the destination, e.g., http, telnet, etc.	C
4. flag (D)	normal or error status of the connection	D
5. src_bytes (C)	number of data bytes from source to destination	E
6. dst_bytes (C)	number of data bytes from destination to source	F
7. land (D)	1 if connection is from/to the same host/port; 0 otherwise	G
8. wrong_fragment (C)	number of "wrong" fragments	H
9. urgent (C)	number of urgent packets	I
10.hot (C)	number of "hot" indicators	J
11. num_failed_logins (C)	number of failed login attempts	K
12. logged_in (D)	1 if successfully logged in; 0 otherwise	L
13. num_compromised (C)	number of "compromised" conditions	M
14. root_shell (C)	1 if root shell is obtained; 0 otherwise	N
15. su_attempted (C)	1 if "su root" command attempted; 0 otherwise	O
16. num_root (C)	number of "root" accesses	P
17. num_file_creations (C)	number of file creation operations	Q
18. num_shells (C)	number of shell prompts	R
19. num_access_files (C)	number of operations on access control files	S
20. num_otbound_cmds (C)	number of outbound commands in an ftp session	T
21. is_host_login (D)	1 if the login belongs to the "hot" list; 0 otherwise	U
22. is_guest_login (D)	1 if the login is a "guest" login; 0 otherwise	V
23. count (C)	number of connections to the same host as the current connection in the past two seconds	W
24. srv_count (C)	number of connections to the same service as the current connection in the past two seconds	X
25. serror_rate (C)	% of connections that have "SYN" errors	Y
26. srv_serror_rate (C)	% of connections that have "SYN" errors	Z
27. rerror_rate (C)	% of connections that have "REJ" errors	AA
28. srv_rerror_rate (C)	% of connections that have "REJ" errors	AB
29. same_srv_rate (C)	% of connections to the same service	AC
30. diff_srv_rate (C)	% of connections to different services	AD
31. srv_diff_host_rate (C)	% of connections to different hosts	AE
32. dst_host_count (C)	count for destination host	AF
33. dst_host_srv_count (C)	srv_count for destination host	AG
34. dst_host_same_srv_rate (C)	same_srv_rate for destination host	AH
35. dst_host_diff_srv_rate (C)	dif_srv_rate for destination host	AI
36. dst_host_same_src_port_rate (C)	same_src_port_rate for destination host	AJ
37. dst_host_srv_diff_host_rate (C)	diff_host_rate for destination host	AK
38. dst_host_serror_rate (C)	serror_rate for destination host	AL
39. dst_host_srv_serror_rate (C)	srv_serror_rate for destination host	AM
40. dst_host_rerror_rate (C)	rerror_rate for destination host	AN
41. dst_host_srv_rerror_rate (C)	srv_serror_rate for destination host	AO

International Journal of Engineering Research and Development
eISSN : 2278-067X, pISSN : 2278-800X, www.ijerd.com
Volume 2, Issue 2 (July 2012), PP. 44-48

Implementation of Fuzzy c-Means and Outlier Detection for Intrusion Detection with KDD Cup 1999 Data Set

S. Songma[1], W. Chimphlee[2], K. Maichalernnukul[3], P. Sanguansat[4]
[1,3]Faculty of Information Technology, Rangsit University, Muang-Ake, Phathumthani 12000, Thailand
[2]Faculty of Science and Technology, Suan Dusit Rajabhat University, Bangkok 10300, Thailand
[4]Faculty of Engineering and Technology, Panyapiwat Institute of Management, Nonthaburi 10200, Thailand

Abstract— In this paper, a two-phase method for computer network intrusion detection is proposed. In the first phase, a set of patterns (data) are clustered by the fuzzy c-means algorithm. In the second phase, outliers are constructed by a distance-based technique and a class label is assigned to each pattern. The KDD Cup 1999 data set is used for the experiment. The results show that, for binary classification (i.e., normal or attack), the proposed method achieves a higher detection rate and a greater overall accuracy than the fuzzy c-means algorithm.

Keywords— Clustering, fuzzy c-means, intrusion detection, KDD Cup 1999 data set, outlier detection

I. INTRODUCTION

As defined in [1], intrusion detection is the process of monitoring the events occurring in a computer network and analyzing them for signs of intrusions. It is also defined as attempts to compromise the confidentiality, integrity, availability, or to bypass the security mechanisms of a computer network. Anomaly intrusion detection systems (IDSs) aim at distinguishing an abnormal activity from an ordinary one.

The current state of computer networks is vulnerable; they are prone to an increasing number of attacks. These attacks are seldom previously seen. It is very hard to detect them before subsequent damage is done. Therefore, securing such a network from unwanted malicious traffic is of prime concern.

In this paper, a two-phase method for intrusion detection, called 2PID, is proposed. The Knowledge Discovery in Databases (KDD) Cup 1999 data set [2], which has been utilized extensively for development of IDSs, is used as a representative sample of data.

The rest of the paper is organized as follows: Section II introduces the proposed method. Section III describes the experimental setup. Section IV provides the results, and Section V concludes the paper.

II. METHOD DESCRIPTION

In this section, we first review fuzzy c-means (FCM) clustering and distance-based outlier detection. Then, we present our proposed method.

A. FCM Clustering

Clustering is an unsupervised classification mechanism where a set of patterns (data), usually multidimensional, are classified into groups (clusters) such that members of one group are similar according to a predefined criterion [3].

FCM is an unsupervised fuzzy clustering algorithm that has been applied successfully to a number of problems involving feature analysis, clustering, and classifier design. It takes unlabeled intrusion data points and tries to group them according to their similarity; points assigned to the same cluster have high similarity, while the similarity between points assigned to different clusters is low [4].

The FCM algorithm partition a set of N patterns $\{X_k\}$ into c clusters by minimizing the objective function

$$J = \sum_{k=1}^{N}\sum_{i=1}^{c}(\mu_{ik})^{m'}\left\|X_k - m_i\right\|^2 \quad (1)$$

where $1 \le m' < \infty$ is the fuzzifier, m_i is the i^{th} cluster center, $\mu_{ik} \in [0,1]$ is the membership of the k^{th} pattern to it, and $\|\cdot\|$ is the distance norm. The parameters m_i and μ_{ik} are calculated as

$$m_i = \frac{\sum_{k=1}^{N}(\mu_{ik})^{m'} X_k}{\sum_{k=1}^{N}(\mu_{ik})^{m'}}, \quad (2)$$

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}} \right)^{\frac{2}{m'-1}}} \tag{3}$$

with $d_{ik} = \|X_k - m_i\|^2$, subject to $\sum_{i=1}^{c} \mu_{ik} = 1$ and $0 < \sum_{k=1}^{N} \mu_{ik} < N$. The algorithm proceeds as follows [5]:

(i) Pick the initial means m_i, $i = 1,2,...,c$. Choose the values for the fuzzifier m' and the threshold ε. Set the iteration counter $t = 1$;

(ii) Compute μ_{ik} for c clusters and N data points, by (3);

(iii) Update m_i by (2);

(iv) Repeat steps (ii) and (iii), by incrementing t, until $\|\mu_{ik}(t) - \mu_{ik}(t-1)\| > \varepsilon$.

B. Distance-Based Outlier Detection

Outlier is defined as an observation that appears to be inconsistent with other observations in a data set. Many data-mining algorithms try to minimize the influence of outliers on the final model, or to eliminate them in the preprocessing phases. Outlier detection and potential removal from the data set can be described as a process of the selection of L out of N samples that are considerably dissimilar, exceptional, or inconsistent with respect to the remaining data.

Distance-based technique is a class of outlier-detection method. The basic computational complexity of this technique is the evaluation of distance measures between all samples in a given data set. Then, a sample in a data set $\{X_k\}$ is an outlier if at least a fraction p of the samples in $\{X_k\}$ lies at a distance greater than r. Clearly, the criterion for outlier detection is based on p and r. These two parameters may be given beforehand using knowledge about the data. Further details are in [6].

C. Proposed Method

The 2PID consists of two phases. In the first phase, a set of patterns are classified by FCM clustering. In the second phase, outliers are constructed by a distance-based technique, and a class label is assigned to each pattern.

Binary classification is the task of classifying the members of a given data set into two groups on the basis of whether they have some property or not. The binary classification task in the context of intrusion detection is to differentiate between normal connections and attack situations. In this paper, we focus on such a task.

III. EXPERIMENTAL SETUP

A. KDD Cup 1999 Data Set

The data set provided for the 1999 KDD Cup was originally prepared by MIT Lincoln labs for the 1998 Defense Advanced Research Projects Agency (DARPA) Intrusion Detection Evaluation Program, with the objective of evaluating research in intrusion detection, and it has become a benchmark data set for the evaluation of IDSs. Attacks fall into four main categories:

- Denial of service (DoS), where an attacker makes some computing or memory resource too busy or too full to handle legitimate requests, thus denying legitimate users access to a machine, e.g., SYN flood;
- Remote to local (R2L), where an attacker sends packets to a machine over a network, then exploits machine's vulnerability to illegally gain local access as a user, e.g., guessing password;
- User to root (U2R), where an attacker starts out with access to a normal user account on the system and is able to exploit vulnerability to gain root access to the system, e.g., buffer overflows;
- Probing, where an attacker scans a network to gather information or find known vulnerabilities, e.g., port scanning.

The KDD Cup 1999 data set has a huge number of duplicated records as shown in Table I on the next page. This data set lies with the distribution of its five classes. The DoS attack comprises 79.24% in training and 73.90% in testing, respectively. Meanwhile, normal connection consists of 19.69% in training and 19.48% in testing, respectively. This imbalance makes it very difficult to train classifiers on the training set, and results in having extremely poor detection rates. In this paper, we use a subset of the original data set which consists of distinct records only.

Table I: Data Distribution and Ratio in the Original Data Set

Class	Training		Testing	
	Amount of Data	Ratio (%)	Amount of Data	Ratio (%)
Normal	97,278	19.69	60,593	19.48
DoS	391,458	79.24	229,853	73.90
R2L	1,126	0.23	16,189	5.20
U2R	52	0.01	228	0.07
Probing	4,107	0.83	4,166	1.34
Total	494,021	100	311,029	100

B. Data Preprocessing

Data preprocessing has to be undertaken before we could do any experiment. It is carried out in two steps. The first step involves mapping symbolic-valued attributes to numeric-valued attributes. The second step implements non-zero numerical features.

The redundancy in the KDD Cup 1999 data set is surprisingly high. By deleting the repeated data, the size of the data set is reduced from 311,029 to 77,291 as shown in Table II.

Table II: Data Distribution and Ratio in the Reduced Data Set

Class	Amount of Data	Ratio (%)
Normal	47,913	61.99
DoS	23,568	30.49
R2L	2,913	9.77
U2R	215	0.27
Probing	2,682	3.47
Total	77,291	100

IV. RESULTS

Standard measures which were developed for evaluating IDSs include detection rate (DTR), false positive rate (FPR), and overall accuracy (OA). These three performance metrics may be defined as follows [7]:

$$\mathrm{DTR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \times 100\%, \tag{4}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN+FP}} \times 100\%, \tag{5}$$

$$\mathrm{OA} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \times 100\%, \tag{6}$$

where TP, TN, FP, and FN are the numbers of malicious executables correctly classified as malicious, benign programs correctly classified as benign, benign programs falsely classified as malicious, and malicious executables falsely classified as benign, respectively. An IDS requires high DTR, low FPR, and high OA.

The block diagram of our experiment is shown in Fig.1 on the next page. We consider all attacks as a whole, and all 41 features are shown in Table III on the next page. We choose $c = 2$, $p = 230$, and $r = 1.5$. The DTRs, FPRs, and OAs for the FCM and the 2PID are shown in Table IV. Obviously, the 2PID yields a higher DTR and a greater OA than the FCM, while the FPRs for both methods are equal. This demonstrates the effectiveness of our proposed method.

Table IV: Result of the Experiment

Method	Detection Rate (DTR)	False Positive Rate (FPR)	Overall Accuracy (OA)
FCM	81.07%	2.50%	91.26%
2PID	90.35%	2.50%	94.78%

V. CONCLUSION

This paper has proposed a two-phase approach to intrusion detection, where the KDD Cup 1999 data set has been considered. The experimental results have shown that the proposed method is superior to the FCM. In future work, we plan to include a feature selection algorithm to help build efficient and practical intrusion detection.

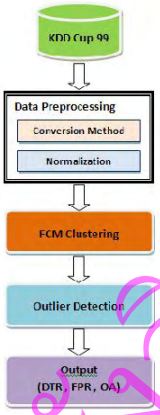


Fig. 1 Block diagram of the experiment

Table III: Feature Description of the KDD Cup 1999 Data Set

Feature Name	Description	Type
1. Duration	Length (number of seconds) of the connection	Continuous
2. Protocol type	Type of the protocol, e.g. tcp, udp, etc.	Discrete
3. Service	Network service on the destination, e.g., http, telnet, etc.	Discrete
4. Flag	Normal or error status of the connection	Discrete
5. Src_bytes	Number of data bytes from source to destination	Continuous
6. Dst_bytes	Number of data bytes from destination to source	Continuous
7. Land	1 if connection is from/to the same host/port; 0 otherwise	Discrete
8. Wrong_fragment	Number of "wrong" fragments	Continuous
9. Urgent	Number of urgent packets	Continuous
10. Hot	Number of "hot" indicators	Continuous
11. Num_failed_logins	Number of failed login attempts	Continuous
12. Logged_in	1 if successfully logged in; 0 otherwise	Discrete
13. Num_compromised	Number of "compromised" conditions	Continuous
14. Root_shell	1 if root shell is obtained; 0 otherwise	Continuous
15. Su_attempted	1 if "su_root" command attempted; 0 otherwise	Continuous
16. Num_root	Number of "root" accesses	Continuous
17. Num_file_creations	Number of file creation operations	Continuous
18. Num_shells	Number of shell prompts	Continuous
19. Num_access_files	Number of operations on access control files	Continuous
20. Num_otbound_cmds	Number of outbound commands in an ftp session	Continuous
21. Is_host_login	1 if the login belongs to the "hot" list; 0 otherwise	Discrete
22. Is_guest_login	1 if the login is a "guest" login; 0 otherwise	Discrete
23. Count	Number of connections to the same host as the current connection in the past two seconds	Continuous
24. Srv_count	Number of connections to the same service as the current connection in the past two seconds	Continuous
25. Serror_rate	% of connections that have "SYN" errors	Continuous
26. Srv_serror_rate	% of connections that have "SYN" errors	Continuous
27. Rerror_rate	% of connections that have "REJ" errors	Continuous

Table III (Continued): Feature Description of the KDD Cup 1999 Data Set

Feature Name	Description	Type
28. Srv_rerror_rate	% of connections that have "REJ" errors	Continuous
29. Same_srv_rate	% of connections to the same service	Continuous
30. Diff_srv_rate	% of connections to different services	Continuous
31. Srv_diff_host_rate	% of connections to different hosts	Continuous
32. Dst_host_count	Count for destination host	Continuous
33. Dst_host_srv_count	Srv_count for destination host	Continuous
34. Dst_host_same_srv_rate	Same_srv_rate for destination host	Continuous
35. Dst_host_diff_srv_rate	Dif_srv_rate for destination host	Continuous
36. Dst_host_same_srv_port_rate	Same_src_port_rate for destination host	Continuous
37. Dst_host_srv_diff_host_rate	Diff_host_rate for destination host	Continuous
38. Dst_host_serror_rate	Serror_rate for destination host	Continuous
39. Dst_host_srv_serror_rate	Srv_serror_rate for destination host	Continuous
40. Dst_host_rerror_rate	Rerror_rate for destination host	Continuous
41. Dst_host_srv_rerror_rate	Srv_serror_rate for destination host	Continuous

REFERENCES

[1] R. Bace and P. Mell, "Intrusion Detection Systems," NIST Special Publications on Intrusion Detection Systems. SP 800.31, Nov. 2001.

[2] KDD Data Set. (1999) [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[3] M. K. Pakhira, "A modified k-means algorithm to avoid empty clusters," International Journal of Recent Trends in Engineering, vol. 1, no. 1, pp. 220-226, May 2009.

[4] D. E. Denning, "An intrusion-detection model," IEEE Transactions on Software Engineering, vol. SE-13, no. 2, pp. 222-232, Feb. 1987.

[5] R. Jensen and Q. Shen, "Fuzzy-rough data reduction with ant colony optimization," Fuzzy Sets and Systems, vol. 149, pp. 5-20, 2005.

[6] M. Kantardzic, Data Mining: Concepts, Models, and Algorithms, New Jersey: IEEE Press, 2003.

[7] W. Chimphlee, M. N. M. Sap, A. H. Abdullah, S. Chimphlee, and S. Srinoy, "Anomaly detection of intrusion based on integration of rough sets and fuzzy c-means," Journal of Information Technology, vol. 17, no. 2, pp. 1-14, Dec. 2005.

Proceedings
2012 Tenth International Conference on ICT and Knowledge Engineering

November 21-23, 2012 Bangkok, Thailand

IEEE Catalog Number: **CFP1228H-PRT**
ISBN: **978-1-4673-2314-7**
ISSN: **2157-0981**
IEEE Conference: **#20935**

Jointly organized by

CONTENTS (Cont.)

Page

2012 Tenth International Conference on ICT and Knowledge Engineering

Classification via *k*-Means Clustering and Distance-Based Outlier Detection

Surasit Songma
Faculty of Information Technology
Rangsit University
Phathumthani, Thailand

Witcha Chimphlee
Faculty of Science and Technology
Suan Dusit Rajabhat University
Bangkok, Thailand

Kiattisak Maichalernnukul
Faculty of Information Technology
Rangsit University
Phathumthani, Thailand

Parinya Sanguansat
Faculty of Engineering and Technology
Panyapiwat Institute of Management
Nonthaburi, Thailand

***Abstract*—We propose a two-phase classification method. Specifically, in the first phase, a set of patterns (data) are clustered by the *k*-means algorithm. In the second phase, outliers are constructed by a distance-based technique and a class label is assigned to each pattern. The Knowledge Discovery Databases (KDD) Cup 1999 data set, which has been utilized extensively for development of intrusion detection systems, is used in our experiment. The results show that the proposed method is effective in intrusion detection.**

***Keywords*—Classification, *k*-means, intrusion detection, KDD Cup 1999 data set, outlier detection**

I. Introduction

As defined in [1], intrusion detection is the process of monitoring the events occurring in a computer network and analyzing them for signs of intrusions. Also, it is defined as attempts to compromise the confidentiality, integrity, availability, or to bypass the security mechanisms of a computer network. Anomaly intrusion detection systems (IDSs) aim at distinguishing an abnormal activity from an ordinary one.

The current state of computer networks is vulnerable; they are prone to an increasing number of attacks. These attacks are seldom previously seen. It is very hard to detect them before subsequent damage is done. Therefore, securing such a network from unwanted malicious traffic is of prime concern.

In this paper, a two-phase classification method is proposed and applied to intrusion detection where the Knowledge Discovery Databases (KDD) Cup 1999 data set [2] is considered.

The rest of the paper is organized as follows: Section II introduces the proposed method. Section III describes the experimental setup. Section IV provides the results, and Section V concludes the paper.

II. Method Description

In this section, we begin with a review of *k*-means clustering and distance-based outlier detection. Next, we present our proposed method.

A. *k-Means Clustering*

k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple way to classify a given data set through a certain number of clusters (*k* clusters assumed) fixed a priori [3]. This algorithm aims at minimizing an objective function, or, in our context, the following squared error function:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \quad (1)$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the corresponding cluster center c_j. This function is an indicator of the distance of the n data points from their respective cluster centers.

The *k*-means algorithm is composed of the following steps:

1) Initialize k center locations (i.e., $c_1,...,c_k$);
2) Assign each $x_i^{(j)}$ to its nearest cluster center c_j.
3) Update each cluster center c_j as the mean of all $x_i^{(j)}$ that have been assigned closest to it.
4) Recalculate the positions of the k centers.
5) Repeat steps 2, 3, and 4 until the centers no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

B. *Distance-Based Outlier Detection*

Outlier is defined as an observation that appears to be inconsistent with other observations in a data set. Many data-mining algorithms try to minimize the influence of outliers on the final model, or to eliminate them in the preprocessing phases. Outlier detection and potential removal from the data set can be described as a process of the selection of L out of N samples that are considerably dissimilar, exceptional, or inconsistent with respect to the remaining data.

Distance-based technique is a class of outlier-detection method. The basic computational complexity of this technique is the evaluation of distance measures between all samples in a given data set. Then, a sample in a data set $\{X_k\}$ is an outlier if

978-1-4673-2317-8/12/$31.00©2012 IEEE

at least a fraction p of the samples in $\{X_k\}$ lies at a distance greater than r. Clearly, the criterion for outlier detection is based on p and r. These two parameters may be given beforehand using knowledge about the data. Further details are in [3].

C. Proposed Method

Our classification method consists of two phases. In the first phase, a set of patterns are classified by k-means clustering. In the second phase, outliers are constructed by the distance-based technique, and a class label is assigned to each pattern.

Binary classification is the task of classifying the members of a given data set into two groups on the basis of whether they have some property or not. The binary classification task in the context of intrusion detection is to differentiate between normal connections and attack situations. In this paper, we focus on such a task.

III. Experimental Setup

A. KDD Cup 1999 Data Set

The data set provided for the 1999 KDD Cup was originally prepared by MIT Lincoln labs for the 1998 Defense Advanced Research Projects Agency (DARPA) Intrusion Detection Evaluation Program, with the objective of evaluating research in intrusion detection, and it has become a benchmark data set for the evaluation of IDSs. Attacks fall into four main categories:

- Denial of service (DoS), where an attacker makes some computing or memory resource too busy or too full to handle legitimate requests, thus denying legitimate users access to a machine, e.g., SYN flood;
- Remote to local (R2L), where an attacker sends packets to a machine over a network, then exploits machine's vulnerability to illegally gain local access as a user, e.g., guessing password;
- User to root (U2R), where an attacker starts out with access to a normal user account on the system and is able to exploit vulnerability to gain root access to the system, e.g., buffer overflows;
- Probing, where an attacker scans a network to gather information or find known vulnerabilities, e.g., port scanning.

The KDD Cup 1999 data set has a huge number of duplicated records as shown in Table I. This data set lies with the distribution of its five classes. The DoS attack comprises 79.24% in training and 73.90% in testing, respectively. Meanwhile, normal connection consists of 19.69% in training and 19.48% in testing, respectively. This imbalance makes it very difficult to train classifiers on the training set, and results in having extremely poor detection rates. In this paper, we use a subset of the original data set which consists of distinct records only.

TABLE I
Data Distribution and Ratio in The Original Data Set

Class	Training		Testing	
	Amount of Data	Ratio (%)	Amount of Data	Ratio (%)
Normal	97,278	19.69	60,593	19.48
DoS	391,458	79.24	229,853	73.90
R2L	1,126	0.23	16,189	5.20
U2R	52	0.01	228	0.07
Probing	4,107	0.83	4,166	1.34
Total	494,021	100	311,029	100

TABLE II
Data Distribution and Ratio in The Reduced Data Set

Class	Amount of Data	Ratio (%)
Normal	47,913	61.99
DoS	23,568	30.49
R2L	2,913	3.77
U2R	215	0.27
Probing	2,682	3.47
Total	77,291	100

B. Data Preprocessing

Data preprocessing has to be undertaken before we could do any experiment. It is carried out in two steps. The first step involves mapping symbolic-valued attributes to numeric-valued attributes. The second step implements non-zero numerical features.

The redundancy in the KDD Cup 1999 data set is surprisingly high. By deleting the repeated data, the size of data set is reduced from 311,029 to 77,291 as shown in Table II.

IV. Results

Standard measures which were developed for evaluating IDSs include detection rate (DTR), false positive rate (FPR), and overall accuracy (OA). These three performance metrics may be defined as follows [4]:

$$\mathrm{DTR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \quad 100\%, \tag{2}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN+FP}} \quad 100\%, \tag{3}$$

$$\mathrm{OA} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \times 100\%, \tag{4}$$

where TP, TN, FP, and FN are the numbers of malicious executables correctly classified as malicious, benign programs correctly classified as benign, benign programs falsely classified as malicious, and malicious executables falsely classified as benign, respectively. An IDS requires high DTR, low FPR, and high OA.

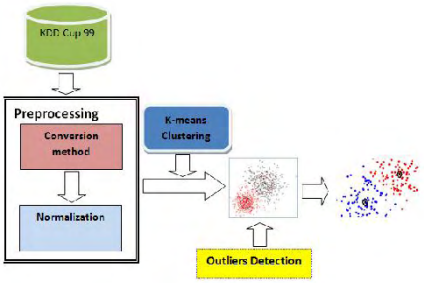


Fig. 1. Block diagram of the experiment

TABLE IV
RESULT OF EXPERIMENT

Method	DTR	FPR	OA
Support vector machine	71.6%	28.4%	86.1%
Rough set-fussy support vector machine	85.76%	14.24%	90%
Two-phase classification	91.35%	1.80%	95.59%

The block diagram of our experiment is shown in Fig. 1. We consider all attacks as a whole, and all 41 features are shown in Table III on the next page. In Table IV, we compare the DTR, FPR, and OA of our two-phase classification method with those of support vector machine and rough set-fussy support vector machine [5]. The two-phase classification method achieves the highest DTR and the lowest FPR.

V. CONCLUSION

This paper proposes a two-phase classification method. The experimental results show that the proposed method is not only able to achieve low FPR but also to yield high DTR for the IDS.

REFERENCES

[1] R. Bace and P. Mell, "Intrusion Detection Systems," NIST Special Publications on Intrusion Detection Systems. SP800-31, Nov. 2011.
[2] KDD data set. (online). http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
[3] M. Kantardzic, *Data Mining: Concepts, Models, and Algorithms*, New Jersey: IEEE Press, 2003.
[4] W. Chimphlee, M. N. M. Sap, A. H. Abdullah, S. Chimphlee, and S. Srinoy, "Anomaly detection of intrusion based on integration of rough sets and fuzzy *c*-means," *Journal of Information Technology*, vol. 17, no. 2, pp. 1-14, Dec. 2005.
[5] L. Li and K. Zhao, "A New Intrusion Detection System Based on Rough Set Theory and Fuzzy Support Vector Machine Fuzzy Support Vector Machine," in *Proceedings of International Workshop on Intelligent Systems and Applications*, pp. 1-5, 2011.

TABLE III
FEATURE DESCRIPTION OF THE KDD CUP 1999 DATA SET

No.	Feature Name	Description	Type
1	Duration	Length (number of seconds) of the connection	Continuous
2	Protocol type	Type of the protocol, e.g., tcp, udp, etc.	Discrete
3	Service	Network service on the destination, e.g., http, telnet, etc.	Discrete
4	Flag	Normal or error status of the connection	Discrete
5	Src_bytes	Number of data bytes from source to destination	Continuous
6	Dst_bytes	Number of data bytes from destination to source	Continuous
7	Land	1 if connection is from/to the same host/port; 0 otherwise	Discrete
8	Wrong_fragment	Number of "wrong" fragments	Continuous
9	Urgent	Number of urgent packets	Continuous
10	Hot	Number of "hot" indicators	Continuous
11	Num_failed_logins	Number of failed login attempts	Continuous
12	Logged_in	1 if successfully logged in; 0 otherwise	Discrete
13	Num_compromised	Number of "compromised" conditions	Continuous
14	Root_shell	1 if root shell is obtained; 0 otherwise	Continuous
15	Su_attempted	1 if "su root" command attempted; 0 otherwise	Continuous
16	Num_root	Number of "root" accesses	Continuous
17	Num_file_creations	Number of file creation operations	Continuous
18	Num_shells	Number of shell prompts	Continuous
19	Num_access_files	Number of operations on access control files	Continuous
20	Num_otbound_cmds	Number of outbound commands in an ftp session	Continuous
21	Is_host_login	1 if the login belongs to the "hot" list; 0 otherwise	Discrete
22	Is_guest_login	1 if the login is a "guest" login; 0 otherwise	Discrete
23	Count	Number of connections to the same host as the current connection in the past two seconds	Continuous
24	Srv_count	Number of connections to the same service as the current connection in the past two seconds	Continuous
25	Serror_rate	% of connections that have "SYN" errors	Continuous
26	Srv_serror_rate	% of connections that have "SYN" errors	Continuous
27	Rerror_rate	% of connections that have "REJ" errors	Continuous
28	Srv_rerror_rate	% of connections that have "REJ" errors	Continuous
29	Same_srv_rate	% of connections to the same service	Continuous
30	Diff_srv_rate	% of connections to different services	Continuous
31	Srv_diff_host_rate	% of connections to different hosts	Continuous
32	Dst_host_count	Count for destination host	Continuous
33	Dst_host_srv_count	Srv_count for destination host	Continuous
34	Dst_host_same_srv_rate	Same_srv_rate for destination host	Continuous
35	Dst_host_diff_srv_rate	Dif_srv_rate for destination host	Continuous
36	Dst_host_same_src_port_rate	Same_src_port_rate for destination host	Continuous
37	Dst_host_srv_diff_host_rate	Diff_host_rate for destination host	Continuous
38	Dst_host_serror_rate	Serror_rate for destination host	Continuous
39	Dst_host_srv_serror_rate	Srv_serror_rate for destination host	Continuous
40	Dst_host_rerror_rate	Rerror_rate for destination host	Continuous
41	Dst_host_srv_rerror_rate	Srv_serror_rate for destination host	Continuous

ภาคผนวก ข

ผลการแบ่งข้อมูลฝึกอบรมในชุดข้อมูล KDD Cup 1999 โดยใช้เทคนิคเอ็มดีแอล

ตารางที่ 1 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 1 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.99	I1
2	1 - 1.99	I2
3	2 - 3.99	I3
4	4 - 4.99	I4
5	5 - 11.99	I5
6	12 - 13.99	I6
7	14 - 30.99	I7
8	31 - 197.99	I8
9	198 - 1,030.99	I9
10	1,031 - 4,989.99	I10
11	4,990 - 5,024.99	I11
12	5,025 - 5,063.99	I12
13	5,064 - 5,084.99	I13
14	5,085 - 10,038.99	I14
15	10,039 - 10,133.99	I15
16	10,134 - 12,704.99	I16
17	12,705 - 14,681.99	I17
18	14,682 - 15,126.99	I18
19	15,127 - 15,167.99	I19
20	15,168 - 20,939.99	I20
21	20,940 - 30,189.99	I21
22	30,190 - 58,328.99	I22
23	58,329.00	I23

ตารางที่ 2 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 2 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	1	I1
2	2	I2
3	3	I3

ตารางที่ 3 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 3 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	1 - 1.99	I1
2	2 - 2.99	I2
3	3 - 9.99	I3
4	10 - 10.99	I4
5	11 - 11.99	I5
6	12 - 12.99	I6
7	13 - 13.99	I7
8	14 - 15.99	I8
9	16 - 16.99	I9
10	17 - 17.99	I10
11	18 - 18.99	I11
12	19 - 19.99	I12
13	20 - 20.99	I13
14	21 - 21.99	I14
15	22 - 23.99	I15
16	24 - 37.99	I16
17	38 - 39.99	I17
18	40 - 40.99	I18

ตารางที่ 3 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 3 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
19	41 - 41.99	I19
20	42 - 42.99	I20
21	43 - 43.99	I21
22	44 - 46.99	I22
23	47 - 47.99	I23
24	48 - 52.99	I24
25	53 - 53.99	I25
26	54 - 54.99	I26
27	55 - 55.99	I27
28	56 - 59.99	I28
29	60 - 66.99	I29
30	67	I30

ตารางที่ 4 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 4 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 1.99	I1
2	2 - 3.99	I2
3	4 - 4.99	I3
4	5 - 7.99	I4
5	8 - 8.99	I5
6	9 - 10.99	I6
7	11	I7

ตารางที่ 5 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 5 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.99	I1
2	1 - 27.99	I2
3	28 - 114.99	I3
4	115 - 124.99	I4
5	125 - 125.99	I5
6	126 - 141.99	I6
7	142 - 206.99	I7
8	207 - 245.99	I8
9	246 - 333.99	I9
10	334 - 519.99	I10
11	520 - 831.99	I11
12	832 - 1,011.99	I12
13	1,012 - 1,031.99	I13
14	1,032 - 1,137.99	I14
15	1,138 - 1,189.99	I15
16	1,190 - 1,221.99	I16
17	1,222 - 1,243.99	I17
18	1,244 - 1,268.99	I18
19	1,269 - 1,478.99	I19
20	1,479 - 1,479.99	I20
21	1,480 - 7,021.99	I21
22	7,022 - 7,044.99	I22
23	7,045 - 16,828.99	I23
24	16,829 - 40,493.99	I24
25	40,494 - 52,292.99	I25
26	52,293 - 54,539.99	I26

ตารางที่ 5 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 5 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
27	54,540 - 2,194,618.99	I27
28	2,194,619 - 693,375,639.99	I28
29	693,375,640	I29

ตารางที่ 6 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 6 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 31.99	I1
2	32 - 177.99	I2
3	178 - 178.99	I3
4	179 - 196.99	I4
5	197 - 1183.99	I5
6	1184 - 1207.99	I6
7	1208 - 2186.99	I7
8	2187 - 2200.99	I8
9	2201 - 2442.99	I9
10	2443 - 2444.99	I10
11	2445 - 2446.99	I11
12	2447 - 2452.99	I12
13	2453 - 7297.99	I13
14	7298 - 7299.99	I14
15	7300 - 8312.99	I15
16	8313 - 8314.99	I16
17	8315 - 460349.99	I17
18	460350 - 5134217.99	I18
19	5134218 - 5155467.99	I19
20	5155468	I20

ตารางที่ 7 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 7 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0	I1
2	1	I2

ตารางที่ 8 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 8 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0-0.99	I1
2	1-1.99	I2
3	2.00-3.00	I3

ตารางที่ 9 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 9 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0-2.99	I1
2	3	I2

ตารางที่ 10 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 10 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.99	I1
2	1 - 1.99	I2
3	2 - 2.99	I3
4	3 - 23.99	I4
5	24 - 27.99	I5
6	28 - 29.99	I6
7	30	I7

ตารางที่ 11 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 11 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0-0.99	I1
2	1-3.99	I2
3	4-4.99	I3
4	5	I4

ตารางที่ 12 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 12 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0	I1
2	1	I2

ตารางที่ 13 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 13 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.99	I1
2	1 - 1.99	I2
3	2 - 5.99	I3
4	6 - 12.99	I4
5	13 - 883.99	I5
6	884	I6

ตารางที่ 14 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 14 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0	I1
2	1	I2

ตารางที่ 15 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 15 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 1.99	I1
2	2	I2

ตารางที่ 16 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 16 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0-992.99	I1
2	993	I2

ตารางที่ 17 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 17 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0-27.99	I1
2	28	I2

ตารางที่ 18 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 18 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0-1.99	I1
2	2	I2

ตารางที่ 19 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 19 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.99	I1
2	1 - 1.99	I2
3	2 - 7.99	I3
4	8	I4

ตารางที่ 20 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 20 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0	I1

ตารางที่ 21 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 21 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0	I1

ตารางที่ 22 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 22 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0	I1
2	1	I2

ตารางที่ 23 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 23 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.99	I1
2	1 - 1.99	I2
3	2 - 2.99	I3
4	3 - 3.99	I4
5	4 - 4.99	I5
6	5 - 5.99	I6
7	6 - 6.99	I7
8	7 - 7.99	I8
9	8 - 8.99	I9
10	9 - 9.99	I10
11	10 - 10.99	I11
12	11 - 11.99	I12
13	12 - 12.99	I13

ตารางที่ 23 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 23 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
14	13 - 13.99	I14
15	14 - 14.99	I15
16	15 - 15.99	I16
17	16 - 16.99	I17
18	17 - 17.99	I18
19	18 - 18.99	I19
20	19 - 19.99	I20
21	20 - 20.99	I21
22	21 - 21.99	I22
23	22 - 22.99	I23
24	23 - 23.99	I24
25	24 - 24.99	I25
26	25 - 25.99	I26
27	26 - 26.99	I27
28	27 - 27.99	I28
29	28 - 28.99	I29
30	29 - 29.99	I30
31	30 - 30.99	I31
32	31 - 31.99	I32
33	32 - 32.99	I33
34	33 - 33.99	I34
35	34 - 34.99	I35
36	35 - 35.99	I36
37	36 - 36.99	I37
38	37 - 37.99	I38
39	38 - 38.99	I39

ตารางที่ 23 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 23 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
40	39 - 40.99	I40
41	41 - 41.99	I41
42	42 - 42.99	I42
43	43 - 43.99	I43
44	44 - 44.99	I44
45	45 - 45.99	I45
46	46 - 46.99	I46
47	47 - 47.99	I47
48	48 - 48.99	I48
49	49 - 76.99	I49
50	77 - 85.99	I50
51	86 - 100.99	I51
52	101 - 148.99	I52
53	149 - 180.99	I53
54	181 - 200.99	I54
55	201 - 301.99	I55
56	302 - 325.99	I56
57	326 - 430.99	I57
58	431 - 507.99	I58
59	508 - 508.99	I59
60	509 - 509.99	I60
61	510 - 510.99	I61
62	511	I62

ตารางที่ 24 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 24 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.99	I1
2	1 - 1.99	I2
3	2 - 2.99	I3
4	3 - 3.99	I4
5	4 - 4.99	I5
6	5 - 5.99	I6
7	6 - 6.99	I7
8	7 - 7.99	I8
9	8 - 8.99	I9
10	9 - 9.99	I10
11	10 - 10.99	I11
12	11 - 11.99	I12
13	12 - 12.99	I13
14	13 - 13.99	I14
15	14 - 14.99	I15
16	15 - 15.99	I16
17	16 - 16.99	I17
18	17 - 17.99	I18
19	18 - 18.99	I19
20	19 - 19.99	I20
21	20 - 20.99	I21
22	21 - 21.99	I22
23	22 - 22.99	I23
24	23 - 23.99	I24
25	24 - 24.99	I25
26	25 - 39.99	I26

ตารางที่ 24 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 24 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
27	40 - 52.99	I27
28	53 - 63.99	I28
29	64 - 99.99	I29
30	100 - 162.99	I30
31	163 - 324.99	I31
32	325 - 510.99	I32
33	511	I33

ตารางที่ 25 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 25 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.009	I1
2	0.01 - 0.019	I2
3	0.02 - 0.019	I3
4	0.02 - 0.029	I4
5	0.03 - 0.039	I5
6	0.04 - 0.049	I6
7	0.05 - 0.059	I7
8	0.06 - 0.069	I8
9	0.07 - 0.079	I9
10	0.08 - 0.089	I10
11	0.09 - 0.099	I11
12	0.1 - 0.109	I12
13	0.11 - 0.119	I13
14	0.12 - 0.159	I14
15	0.16 - 0.169	I15
16	0.17 - 0.189	I16

ตารางที่ 25 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 25 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
17	0.19 - 0.199	I17
18	0.2 - 0.239	I18
19	0.24 - 0.249	I19
20	0.25 - 0.319	I20
21	0.32 - 0.329	I21
22	0.33 - 0.439	I22
23	0.44 - 0.499	I23
24	0.5 - 0.619	I24
25	0.62 - 0.809	I25
26	0.81 - 0.989	I26
27	0.99 - 0.999	I27
28	1	I28

ตารางที่ 26 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 26 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.139	I1
2	0.14 - 0.329	I2
3	0.33 - 0.499	I3
4	0.5 - 0.509	I4
5	0.51 - 0.579	I5
6	0.58 - 0.669	I6
7	0.67 - 0.949	I7
8	0.95 - 0.999	I8
9	1	I9

ตารางที่ 27 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 27 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.019	I1
2	0.02 - 0.029	I2
3	0.03 - 0.039	I3
4	0.04 - 0.139	I4
5	0.14 - 0.169	I5
6	0.17 - 0.189	I6
7	0.19 - 0.199	I7
8	0.2 - 0.239	I8
9	0.24 - 0.249	I9
10	0.25 - 0.429	I10
11	0.43 - 0.499	I11
12	0.5 - 0.749	I12
13	0.75 - 0.989	I13
14	0.99 - 0.999	I14
15	1	I15

ตารางที่ 28 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 28 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.019	I1
2	0.02 - 0.029	I2
3	0.03 - 0.039	I3
4	0.04 - 0.139	I4
5	0.14 - 0.169	I5
6	0.17 - 0.199	I6
7	0.2 - 0.219	I7
8	0.22 - 0.249	I8
9	0.25 - 0.499	I9

ตารางที่ 28 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 28 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
10	0.5 - 0.669	I10
11	0.67 - 0.999	I11
12	1	I12

ตารางที่ 29 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 29 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.009	I1
2	0.01 - 0.099	I2
3	0.1 - 0.109	I3
4	0.11 - 0.139	I4
5	0.14 - 0.199	I5
6	0.2 - 0.239	I6
7	0.24 - 0.249	I7
8	0.25 - 0.319	I8
9	0.32 - 0.329	I9
10	0.33 - 0.389	I10
11	0.39 - 0.399	I11
12	0.4 - 0.489	I12
13	0.49 - 0.499	I13
14	0.5 - 0.589	I14
15	0.59 - 0.599	I15
16	0.6 - 0.649	I16
17	0.65 - 0.669	I17
18	0.67 - 0.749	I18
19	0.75 - 0.939	I19
20	0.94 - 0.959	I20
21	0.96 - 0.969	I21
22	0.97 - 0.979	I22

ตารางที่ 29 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 29 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
23	0.98 - 0.989	I23
24	0.99 - 0.999	I24
25	1	I25

ตารางที่ 30 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 30 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.009	I1
2	0.01 - 0.019	I2
3	0.02 - 0.029	I3
4	0.03 - 0.039	I4
5	0.04 - 0.049	I5
6	0.05 - 0.089	I6
7	0.09 - 0.189	I7
8	0.19 - 0.199	I8
9	0.2 - 0.209	I9
10	0.21 - 0.249	I10
11	0.25 - 0.279	I11
12	0.28 - 0.499	I12
13	0.5 - 0.559	I13
14	0.56 - 0.599	I14
15	0.6 - 0.639	I15
16	0.64 - 0.669	I16
17	0.67 - 0.729	I17
18	0.73 - 0.749	I18
19	0.75 - 0.799	I19
20	0.8 - 0.889	I20

ตารางที่ 30 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 30 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
21	0.89 - 0.989	I21
22	0.99 - 0.999	I22
23	1	I23

ตารางที่ 31 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 31 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.029	I1
2	0.03 - 0.039	I2
3	0.04 - 0.049	I3
4	0.05 - 0.069	I4
5	0.07 - 0.079	I5
6	0.08 - 0.329	I6
7	0.33 - 0.399	I7
8	0.4 - 0.499	I8
9	0.5 - 0.669	I9
10	0.67 - 0.749	I10
11	0.75 - 0.999	I11
12	1	I12

ตารางที่ 32 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 32 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.99	I1
2	1 - 1.99	I2
3	2 - 2.99	I3
4	3 - 3.99	I4
5	4 - 4.99	I5
6	5 - 5.99	I6

ตารางที่ 32 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 32 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
7	6 - 6.99	I7
8	7 - 7.99	I8
9	8 - 10.99	I9
10	11 - 15.99	I10
11	16 - 51.99	I11
12	52 - 72.99	I12
13	73 - 103.99	I13
14	104 - 127.99	I14
15	128 - 172.99	I15
16	173 - 214.99	I16
17	215 - 253.99	I17
18	254 - 254.99	I18
19	255	I19

ตารางที่ 33 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 33 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.99	I1
2	1 - 1.99	I2
3	2 - 2.99	I3
4	3 - 3.99	I4
5	4 - 4.99	I5
6	5 - 5.99	I6
7	6 - 6.99	I7
8	7 - 7.99	I8
9	8 - 8.99	I9
10	9 - 9.99	I10
11	10 - 10.99	I11
12	11 - 11.99	I12

ตารางที่ 33 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 33 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
13	12 - 12.99	I13
14	13 - 13.99	I14
15	14 - 14.99	I15
16	15 - 15.99	I16
17	16 - 16.99	I17
18	17 - 17.99	I18
19	18 - 18.99	I19
20	19 - 19.99	I20
21	20 - 20.99	I21
22	21 - 21.99	I22
23	22 - 22.99	I23
24	23 - 23.99	I24
25	24 - 24.99	I25
26	25 - 48.99	I26
27	49 - 67.99	I27
28	68 - 99.99	I28
29	100 - 143.99	I29
30	144 - 150.99	I30
31	151 - 152.99	I31
32	153 - 222.99	I32
33	223 - 243.99	I33
34	244 - 253.99	I34
35	254 - 254.99	I35
36	255	I36

ตารางที่ 34 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 34 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.009	I1
2	0.01 - 0.019	I2
3	0.02 - 0.029	I3
4	0.03 - 0.039	I4
5	0.04 - 0.049	I5
6	0.05 - 0.059	I6
7	0.06 - 0.069	I7
8	0.07 - 0.079	I8
9	0.08 - 0.089	I9
10	0.09 - 0.099	I10
11	0.1 - 0.229	I11
12	0.23 - 0.239	I12
13	0.24 - 0.269	I13
14	0.27 - 0.399	I14
15	0.4 - 0.569	I15
16	0.57 - 0.819	I16
17	0.82 - 0.869	I17
18	0.87 - 0.989	I18
19	0.99 - 0.999	I19
20	1	I20

ตารางที่ 35 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 35 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.009	I1
2	0.01 - 0.019	I2
3	0.02 - 0.029	I3
4	0.03 - 0.039	I4

ตารางที่ 35 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 35 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
5	0.04 - 0.049	I5
6	0.05 - 0.059	I6
7	0.06 - 0.069	I7
8	0.07 - 0.079	I8
9	0.08 - 0.089	I9
10	0.09 - 0.099	I10
11	0.1 - 0.149	I11
12	0.15 - 0.509	I12
13	0.51 - 0.519	I13
14	0.52 - 0.639	I14
15	0.64 - 0.669	I15
16	0.67 - 0.679	I16
17	0.68 - 0.819	I17
18	0.82 - 0.899	I18
19	0.9 - 0.909	I19
20	0.91 - 0.989	I20
21	0.99 - 0.999	I21
22	1	I22

ตารางที่ 36 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 36 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.009	I1
2	0.01 - 0.019	I2
3	0.02 - 0.029	I3
4	0.03 - 0.139	I4
5	0.14 - 0.149	I5

ตารางที่ 36 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 36 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
6	0.15 - 0.159	I6
7	0.16 - 0.169	I7
8	0.17 - 0.179	I8
9	0.18 - 0.189	I9
10	0.19 - 0.199	I10
11	0.2 - 0.209	I11
12	0.21 - 0.239	I12
13	0.24 - 0.249	I13
14	0.25 - 0.259	I14
15	0.26 - 0.319	I15
16	0.32 - 0.329	I16
17	0.33 - 0.339	I17
18	0.34 - 0.389	I18
19	0.39 - 0.489	I19
20	0.49 - 0.499	I20
21	0.5 - 0.509	I21
22	0.51 - 0.699	I22
23	0.7 - 0.859	I23
24	0.86 - 0.979	I24
25	0.98 - 0.999	I25
26	1	I26

ตารางที่ 37 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 37 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.009	I1
2	0.01 - 0.039	I2

ตารางที่ 37 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 37 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
3	0.04 - 0.049	I3
4	0.05 - 0.069	I4
5	0.07 - 0.249	I5
6	0.25 - 0.259	I6
7	0.26 - 0.499	I7
8	0.5 - 0.519	I8
9	0.52 - 0.999	I9
10	1	I10

ตารางที่ 38 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 38 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.009	I1
2	0.01 - 0.019	I2
3	0.02 - 0.029	I3
4	0.03 - 0.039	I4
5	0.04 - 0.069	I5
6	0.07 - 0.089	I6
7	0.09 - 0.099	I7
8	0.1 - 0.319	I8
9	0.32 - 0.329	I9
10	0.33 - 0.499	I10
11	0.5 - 0.969	I11
12	0.97 - 0.989	I12
13	0.99 - 0.999	I13
14	1	I14

ตารางที่ 39 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 39 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.009	I1
2	0.01 - 0.019	I2
3	0.02 - 0.029	I3
4	0.03 - 0.039	I4
5	0.04 - 0.079	I5
6	0.08 - 0.479	I6
7	0.48 - 0.669	I7
8	0.67 - 0.999	I8
9	1	I9

ตารางที่ 40 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 40 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.009	I1
2	0.01 - 0.019	I2
3	0.02 - 0.029	I3
4	0.03 - 0.039	I4
5	0.04 - 0.049	I5
6	0.05 - 0.059	I6
7	0.06 - 0.079	I7
8	0.08 - 0.139	I8
9	0.14 - 0.509	I9
10	0.51 - 0.519	I10
11	0.52 - 0.529	I11
12	0.53 - 0.629	I12
13	0.63 - 0.839	I13
14	0.84 - 0.909	I14

ตารางที่ 40 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 40 โดยใช้เทคนิคเอ็มดีแอล (ต่อ)

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
15	0.91 - 0.919	I15
16	0.92 - 0.949	I16
17	0.95 - 0.959	I17
18	0.96 - 0.999	I18

ตารางที่ 41 ผลการแบ่งข้อมูลฝึกอบรมเฉพาะคุณลักษณะที่ 41 โดยใช้เทคนิคเอ็มดีแอล

ที่	ข้อมูลต่อเนื่อง	ข้อมูลไม่ต่อเนื่อง
1	0 - 0.009	I1
2	0.01 - 0.019	I2
3	0.02 - 0.029	I3
4	0.03 - 0.039	I4
5	0.04 - 0.049	I5
6	0.05 - 0.059	I6
7	0.06 - 0.079	I7
8	0.08 - 0.139	I8
9	0.14 - 0.289	I9
10	0.29 - 0.489	I10
11	0.49 - 0.499	I11
12	0.5 - 0.849	I12
13	0.85 - 0.909	I13
14	0.91 - 0.919	I14
15	0.92 - 0.939	I15
16	0.94 - 0.949	I16
17	0.95 - 0.959	I17
18	0.96 - 0.999	I18
19	1	I19

ประวัติผู้วิจัย

ชื่อ – นามสกุล	สุระสิทธิ์ ทรงม้า
วัน เดือน ปีเกิด	10 ตุลาคม 2520
สถานที่เกิด	กรุงเทพมหานคร ประเทศไทย
ประวัติการศึกษา	สถาบันราชภัฏจะเชิงเทรา อนุปริญญาวิทยาศาสตร์ สาขาวิชาวิทยาการคอมพิวเตอร์, 2541 สถาบันราชภัฏสวนดุสิต ปริญญาวิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์, 2543 มหาวิทยาลัยรังสิต ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาโครงข่ายโทรคมนาคมและคอมพิวเตอร์, 2548 มหาวิทยาลัยรังสิต ปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ, 2558
ที่อยู่ปัจจุบัน	4/146 ซอยหมู่บ้านชัยพฤกษ์ ถนนมาเจริญ แขวงหนองแขม เขตหนองแขม จังหวัดกรุงเทพมหานคร รหัสไปรษณีย์ 10160
โทรศัพท์	081-372 3218
อีเมล	surasit.songma@gmail.com
สถานที่ทำงาน	มหาวิทยาลัยราชภัฏสวนดุสิต 295 ถนนราชสีมา แขวงดุสิต เขตดุสิต จังหวัดกรุงเทพมหานคร 10300
ตำแหน่งปัจจุบัน	อาจารย์ประจำหลักสูตรเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏสวนดุสิต