ระบบตรวจจับการบุกรุกแบบปรับตัวได้โดยใช้ราฟเซตย่านจุดใกล้เคียงร่วมกับตัวเรียนรู้จำแนกประเภท

Adaptive Intrusion Detection Systems using Neighborhood Rough Set and eXtensive Learning Classifier System (XCS)

ไพฑูรย์ ศรีนิล[1] สมบัติ ฝอยทอง[2] และธารารัตน์ พวงสุวรรณ[1]

บทคัดย่อ

ปัจจุบันเป็นยุคของข้อมูลข่าวสาร การเติบโตของโลกอินเทอร์เน็ตสร้างผลกระทบทำให้การดำเนินการทางธุรกิจและชีวิตประจำวันของผู้คนเปลี่ยนไป ข้อมูลต่าง ๆ รวมถึงข้อมูลที่เป็นความลับได้ถูกสื่อสารแลกเปลี่ยนกันอย่างง่ายดายและรวดเร็ว ทำให้ประเด็นเรื่องความปลอดภัยของข้อมูลจึงได้รับความสนใจเป็นอย่างมาก ระบบตรวจจับการบุกรุกเป็นเครื่องมือหนึ่งที่ช่วยป้องกันข้อมูลจากผู้ประสงค์ร้าย โดยปกติระบบตรวจจับการบุกรุกจะอยู่บนหลักการพื้นฐานของการจดจำรูปแบบพฤติกรรมของการใช้งานปกติ หรือจดจำรูปแบบพฤติกรรมของการบุกรุก แต่อย่างไรก็ตามเนื่องจากผู้บุกรุกได้มีการคิดค้นรูปแบบของการบุกรุกใหม่ ๆ ตลอดเวลา จึงเป็นการยากในการสร้างระบบตรวจจับการบุกรุกให้ทันสมัยเท่าเทียมกันตลอดเวลา ดังนั้นแนวคิดในการนำเทคนิคทางการเรียนรู้เครื่องจักรและปัญญาประดิษฐ์มาสร้างระบบตรวจจับการบุกรุกจึงเป็นปัญหาที่น่าสนใจเป็นอย่างยิ่ง ในงานวิจัยนี้ได้นำเสนอการประยุกต์ให้ตัวเรียนรู้จำแนกประเภทมาทำงานร่วมกับราฟเซตย่านจุดใกล้เคียงเพื่อสร้างระบบการตรวจจับการบุกรุกที่สามารถปรับตัวเองได้ โดยได้นำราฟเซตย่านจุดใกล้เคียงมากำหนดระดับความผิดปกติของชุดข้อมูลก่อนที่จะใช้เป็นชุดข้อมูลสอนให้กับตัวเรียนรู้จำแนกประเภท กระบวนการนี้เป็นการทำให้ข้อมูลมีความเด่นชัดขึ้น ง่ายต่อการจำแนก ฟังก์ชันการกำหนดระดับความผิดปกติของข้อมูลที่ได้นำเสนอเป็นการสกัดจุดเด่นของข้อมูลโดยใช้ข้อมูลในพื้นที่เขตประมาณล่างและพื้นที่เขตประมาณบน ผลลัพธ์ของการทดลองแสดงให้เห็นว่าระบบที่นำเสนอมีประสิทธิภาพในการจำแนกพฤติกรรมการบุกรุกสูงเมื่อเปรียบเทียบกับตัวเรียนรู้จำแนกประเภทอื่น ๆ

คำสำคัญ : การแบ่งนับข้อมูล ระบบตรวจจับการบุกรุก ราฟเซตย่านจุดใกล้เคียง

[1] อาจารย์ คณะวิทยาศาสตร์และศิลปศาสตร์ มหาวิทยาลัยบูรพา วิทยาเขตจันทบุรี

[2] อาจารย์ ดร. คณะวิทยาศาสตร์และศิลปศาสตร์ มหาวิทยาลัยบูรพา วิทยาเขตจันทบุรี

ABSTRACT

The world has now entered into an era of digital revolution. The growth of the internet has been creating an impact on business operations and people daily life. Information is exchanged easily and quickly over the Internet, including confidential information. Therefore, the issue of data security has more interested. Intrusion detection systems (IDS) is one of the tools that protects the data from intruder. Usually, IDS is based on the basic principles of pattern recognition of normal use or intrusion behavior. However, the intrusion patterns are emerging all the time. It is therefore difficult to create an IDS that as up-to-date as intrusion patterns. So, the idea of applying machine learning and artificial intelligent techniques to create IDS is a very challenging problems. In this paper, we proposed a combined technique between neighborhood rough set (NRS) and learning classifier system (XCS) to create an IDS that can be adapted to the patterns of intrusion. NRS is used to determine the level of abnormality of the data set, a process called Abnormal Quantization (AQ), before feeding as a training set to XCS. AQ treats the data set to more distinct and easier to distinguish, it extracts the dominant feature using the concept of lower and upper approximation regions in NRS. Experimental results are illustrated that proposed system offers high performance than other classifiers used in the experiment, comparing the ability to classifiy intruder behavior from normal user behavior.

Keywords : Data Quantization, Intrusion Detection System, Neighborhood Rough Set

บทนำ

ปัจจุบันข้อมูลสารสนเทศมีความสำคัญอย่างมากในทุกองค์กร โดยเฉพาะข้อมูลสารเทศที่เป็นความลับภายในองค์กร อีกทั้งการเติบโตอย่างรวดเร็วของระบบเครือข่ายอินเทอร์เน็ตทำให้ข้อมูลต่าง ๆ ถูกแลกเปลี่ยนกันได้อย่างรวดเร็วและง่ายดาย Richardson (2009) ได้รายงานผลสำรวจการรักษาความมั่นคงปลอดภัยทางคอมพิวเตอร์ในปลายปี พ.ศ. 2552 แจ้งว่าภัยคุกคามจากโปรแกรมประสงค์ร้าย (Malware) มีแพร่หลาย ซึ่งเป็นตัวก่อให้เกิดการขโมยข้อมูลจากเครื่องคอมพิวเตอร์ส่วนบุคคลและโทรศัพท์มือถือ ดังนั้นระบบรักษาความปลอดภัยที่ปกป้องข้อมูลสารสนเทศจึงกลายมาเป็นส่วนสำคัญในทุกองค์กร ระบบตรวจจับการบุกรุก (Intrusion Detection System: IDS) เป็นระบบรักษาความปลอดภัยที่ทำหน้าที่จำแนกพฤติกรรมของผู้ใช้งานเครือข่ายคอมพิวเตอร์ที่เป็นผู้บุกรุกออกจากพฤติกรรมของผู้ใช้งานปกติ และดำเนินการกับผู้บุกรุกด้วยวิธีการที่เหมาะสม (Power, 2002) โดยทั่วไประบบตรวจจับการบุกรุกมี 2 ระบบหลัก คือ Misuse based IDS (MIDS) และ Anomaly based IDS (AIDS) (Roesch et al., 1999; Hossein, 2009; Richardson, 2009; Red et al., 2013) MIDS คือ ระบบตรวจจับการบุกรุกที่ใช้ข้อมูลพฤติกรรมการใช้งานระบบคอมพิวเตอร์ที่ผิดปกติต่าง ๆ ที่เคยเกิดขึ้นมาแล้วนำมาสร้างเป็นฐานข้อมูลของรูปแบบการทำงานที่ผิดปกติ (Habra et al., 1992; Sandeep et al., 1994; Ilgun et al., 1995) การจำแนกผู้บุกรุก MIDS จะนำข้อมูลพฤติกรรมของผู้ใช้งานระบบคอมพิวเตอร์มาเทียบในฐานข้อมูลว่าเทียบเคียงกันได้หรือไม่ ถ้าปรากฏว่ามีข้อมูลของการบุกรุกในรูปแบบนั้น ๆ อยู่ก็แสดงว่าเกิดความผิดปกติขึ้น ข้อเสียของ MIDS คือจะไม่สามารถตรวจจับการบุกรุกชนิดใหม่ ๆ ได้ เนื่องจากจำเป็นต้องมีข้อมูลรูปแบบของการบุกรุกอยู่ก่อนจึงจะตรวจจับความผิดปกติได้ ในขณะที่ AIDS จะมีแนวคิดแบบตรงกันข้ามกันกับ MIDS นั่นคือ การหาเซตของการทำงานที่เป็นปกติย่อย ๆ ขึ้นมาแล้วนำมารวมกันเพื่อให้ระบบตรวจจับการบุกรุกทราบข้อมูลของเซตการทำงานที่เป็น

ปกติทั้งหมดในระบบคอมพิวเตอร์ ถ้าเกิดกรณีที่ระบบตรวจจับการบุกรุกตรวจพบการทำงานที่ไม่ได้อยู่ในเซตของการทำงานที่เป็นปกติ ระบบตรวจจับการบุกรุกจะแจ้งเตือนต่อผู้ดูแลระบบทันที สำหรับการสร้างขอบเขตของระบบนั้นอาจสร้างได้โดยการหาข้อมูลการทำงานที่เป็นปกติในระบบโดยใช้ข้อมูล การทำงานของผู้ใช้งาน เวลาที่มีการใช้งาน ทรัพยากรที่ผู้ใช้งานมักจะใช้บ่อย ๆ เป็นต้น (Dowell et al., 1990; Vaccaro and Liepins 1990; Teresa et al., 1998)

ปัจจุบันกลวิธีทางเหมืองข้อมูลและปัญญาประดิษฐ์ได้ถูกนำมาใช้ในระบบตรวจจับการบุกรุกอย่างกว้างขวาง เนื่องจากรูปแบบการบุกรุกมีมากและเพิ่มขึ้นอย่างรวดเร็ว ในปี ค.ศ. 2014 Memon and Chandel (2014) ได้นำเสนองานวิจัยเรื่อง Hybrid IDS ซึ่งเป็นการนำกลวิธีทางด้านการทำเหมืองข้อมูล 3 กลวิธีมาทำงานร่วมกัน ได้แก่ k-Means, k-Nearest Neighbor (k-NN), และ DTM (Decision Tree Majority) เพื่อสร้างระบบตรวจจับการบุกรุกแบบ Anomaly โดยทดสอบประสิทธิภาพของการตรวจจับโดยใช้ชุดข้อมูล KDD-99 (Hettich, 1999) นำกลวิธี k-means มาจัดกลุ่มข้อมูลใน KDD-99 จากนั้นนำผลที่ได้ไปประมวลผลต่อด้วย k-NN และ DTM ตามลำดับ ผลการทดสอบพบว่าระบบที่นำเสนอสามารถจำแนกประเภทของการบุกรุกได้ดี โดยจำแนกประเภทของการบุกรุกออกเป็น 4 ประเภท คือ R2L, DoS, U2R, และ Probe ในปี ค.ศ. 2013 Dhakar and Tiwari (2013) ได้นำเสนอผลงานเรื่อง แบบจำลองการตรวจจับการบุกรุกด้วยกลวิธี REP (Reduced Error Pruning) มีการนำอัลกอริทึมในการจำแนกประเภท 2 ตัวมาทำงานร่วมกัน คือ K2 (BayesNet) และ DT (Decision Tree) เพื่อสร้างต้นไม้ตัดสินใจแบบรวดเร็วและมีประสิทธิภาพในการจำแนกสูง ในปี ค.ศ. 2012 Om and Kundu (2012) นำเสนอการประยุกต์ใช้หลักการของการคัดเลือกปัจจัยเด่นด้วยค่า Entropy มาสร้างระบบตรวจจับการบุกรุก โดยใช้ k-Means จัดกลุ่มข้อมูลในชุดข้อมูล KDD-99 ออกเป็น 5 กลุ่ม จากนั้นใช้ Naïve Bays และ k-NN สำหรับจำแนกประเภทของการบุกรุก การทดสอบระบบพบว่า ระบบที่นำเสนอมีประสิทธิภาพในการจำแนกสูงกว่าการใช้ k-Means ทำงานแบบเดี่ยว และสูงกว่าการใช้ k-means ทำงานร่วมกับ k-NN ในปี ค.ศ. 2011 Muda et al. (2011) นำเสนอระบบตรวจจับการบุกรุกที่นำ k-means มาทำงานร่วมกับ Naïve Bays โดยนำ k-means มาจัดกลุ่มข้อมูลใน KDD-99 ตามรูปแบบของการบุกรุก จากนั้นใช้ Naïve Bays มาจำแนกกลุ่มของผลลัพธ์ที่ได้จาก k-means อีกครั้ง เนื่องจากข้อมูลที่จัดกลุ่มผิดในขั้นตอนของ k-means อาจจะถูกจำแนกกลุ่มได้ถูกต้องในขั้นตอนของ Naïve Bays ในปี ค.ศ. 2010 Farid et al. (2010) ได้นำเสนออัลกอริทึมสำหรับระบบตรวจจับการบุกรุกแบบปรับตัวเองได้โดยใช้ Naïve Bays ทำงานร่วมกับ ID3 ที่มีประสิทธิภาพในการตรวจจับการบุกรุกสูง และมีค่า False Positive ต่ำ อีกทั้งยังสามารถลดความซ้ำซ้อนของคุณลักษณะ (attribute) สามารถดำเนินการกับคุณลักษณะที่มีข้อมูลเป็นค่าแบบต่อเนื่อง คุณลักษณะที่มีค่าขาดหาย (missing value) และคุณลักษณะที่มีข้อมูลรบกวน (noisy data) ได้

ระบบเรียนรู้จำแนกประเภท (eXtensive Learning Classifier System: XCS) (Wilson, 1995) เป็นระบบเรียนรู้แบบกฎเป็นพื้นฐาน (Rule-Based System) โดยอาศัยเทคนิคการเรียนรู้รีอินฟอซเมนต์ (Reinforcement Learning) (Barto et al., 1983) และจีนีติกอัลกอริทึม (Genetic Algorithms) (Holland, 1975) การเรียนรู้รีอินฟอซเมนต์ทำหน้าที่ปรับค่าพารามิเตอร์ของกฎโดยใช้ข้อมูลค่าตอบกลับจากสภาพแวดล้อม (reward) จากการทดลองทำแบบลองผิดลองถูก (trial-and-error) และจีนีติกอัลกอริทึมทำหน้าที่ในการสร้างและค้นหากฎใหม่ ๆ ที่ดีกว่าที่มีอยู่ ระบบเรียนรู้จำแนกประเภท XCS มีจุดเด่นเรื่องการสร้างกฎที่มีลักษณะความเป็นทั่วไปสูง (Generalization) ปัจจุบันจึงได้ถูกนำมาประยุกต์ใช้อย่างแพร่หลายทั้งในงานระบบควบคุม (Bull, 2014) และงานด้านระบบตรวจจับการบุกรุก (Shafi et al., 2006; Shafi et al., 2006; Shafi et al., 2007;Shafi et al., 2007; Shafi et al., 2009) อย่างไรก็ตามในการนำ XCS มาประยุกต์ใช้ในงานตรวจจับผู้บุกรุกยังประสบปัญหาในเรื่องประสิทธิภาพการทำงานทั้งขั้นตอนการสอนระบบ

และขั้นตอนการจำแนกพฤติกรรมที่ผิดปกติ เพราะพฤติกรรมของการบุกรุกมีหลากหลายและในแต่ละรูปแบบมีการเหลื่อมกันหรือคาบเกี่ยวกัน (Overlap)

ทฤษฎีราฟเซต (Rough Set: RS) ถูกนำเสนอโดย Palawk (1982, 1991) โดยถูกออกแบบมาให้ทำงานกับชุดข้อมูลที่อยู่ในเซตจำกัด (finite set) และมีค่าแบบไม่ต่อเนื่อง (discrete value) ปัจจุบันราฟเซตได้ถูกนำไปประยุกต์ใช้งานอย่างแพร่หลายในหลากหลายสาขา เช่น การคัดเลือกปัจจัยเด่น (Hassanien, 2004; Jensen, 2004; Parthalain et al., 2010) การเรียนรู้ของเครื่องจักร (Blum et al., 1997; Kohavi and John 1997) การทำเหมืองข้อมูล (Dash and Liu 1997; Kim et al., 2000) การจำแนกประเภทเอกสาร (Shang et al., 2007; Aghdam et al., 2009) หรือ การตรวจจับผู้บุกรุก (Lee et al., 2000; Cui-Juan Liu, 2007; Mun et al., 2009) เป็นต้น ในเวลาต่อมา Hu et al.(2008) ได้นำเสนออัลกอริทึมที่มีชื่อว่า Neighborhood Rough Set (NRS) ซึ่งเป็นอัลกอริทึมที่ทำให้ราฟเซตสามารถใช้งานกับชุดข้อมูลที่มีคุณลักษณะผสมกันระหว่างค่าต่อเนื่องและค่าไม่ต่อเนื่องได้

งานวิจัยนี้ได้นำเสนอระบบตรวจจับการบุกรุกโดย NRS ทำงานร่วมกับ XCS โดยนำ NRS มาใช้สำหรับกำหนดระดับความผิดปกติให้กับข้อมูลตัวอย่างในชุดข้อมูล KDD-99 (Hettich et al., 1999) ซึ่งเป็นชุดข้อมูลมาตรฐานที่ใช้สำหรับทดสอบระบบตรวจจับการบุกรุก โดยแบ่งกลุ่มตัวอย่างในชุดข้อมูลออกเป็น 3 กลุ่มแยกตามการเป็นสมาชิกในพื้นที่ คือ 1) พื้นที่เขตปลอดภัย (Safety Region) 2) พื้นที่เขตเฝ้าระวัง (Monitoring Region) และ 3) พื้นที่เขตผิดปกติ (Abnormal Region) จากนั้นทำการเข้ารหัสค่าคุณลักษณะของกลุ่มตัวอย่างพฤติกรรมการบุกรุกตามพื้นที่ที่แบ่งได้โดยใช้ฟังก์ชันที่นำเสนอที่เรียกว่า Abnormal Quantization function (AQ) จากนั้นนำชุดข้อมูลที่ผ่านการเข้ารหัสแล้วใช้เป็นชุดข้อมูลสอนระบบเรียนรู้จำแนกประเภท XCS ทำการทดสอบประสิทธิภาพของระบบที่นำเสนอเปรียบเทียบกับตัวเรียนรู้จำแนกประเภทอื่น ๆ ได้แก่ RF (Random Forest) C4.5 MLP (Multi-Layer Perceptron) RBF (Radial Basis Function Network) และ Naïve bays โดยใช้โปรแกรม Weka (Lan and Eibe 2005) เวอร์ชัน 3.4 ด้วยวิธี 10-fold ผลการทดสอบพบว่าระบบตรวจจับการบุกรุกที่นำเสนอมีประสิทธิภาพดีที่สุดเมื่อเทียบกับตัวเรียนรู้จำแนกประเภทตัวอื่นที่ใช้ในการทดลอง

องค์ประกอบที่เหลือของบทความในงานวิจัยนี้ประกอบด้วยโครงสร้างดังต่อไปนี้ ส่วนที่ 2 อธิบายเกี่ยวกับชุดข้อมูล KDD-99 ทฤษฎีพื้นฐานของ NRS และการทำงานของ XCS ส่วนที่ 3 วิธีดำเนินการวิจัย กรรมวิธีที่เป็นแนวคิดใหม่สำหรับการตรวจจับการบุกรุกบนหลักการของ NRS และ XCS ส่วนที่ 4 ผลการวิจัย อธิบายถึงผลลัพธ์ของการทดสอบประสิทธิภาพของการจำแนกพฤติกรรมการบุกรุก และส่วนที่ 5 บทสรุปและวิจารณ์ของงานวิจัยที่นำเสนอ

ทฤษฎีพื้นฐาน

ในส่วนนี้จะอธิบายเกี่ยวกับชุดข้อมูลผู้บุกรุก KDD-99 ทฤษฎีพื้นฐานของ NRS และการทำงานของ XCS ดังนี้

KDD-99 Data set

ชุดข้อมูล KDD-99 (Hettich and Bay 1999) เป็นชุดข้อมูลที่นิยมใช้ในงานวิจัยที่เกี่ยวกับการตรวจจับการบุกรุก ข้อมูลดังกล่าวได้จากการวิเคราะห์ระบบเครือข่ายของ US Air Force Research Lab ในช่วงเวลา 1998 ซึ่งแต่ละข้อมูลประกอบไปด้วยคุณลักษณะที่ได้จากการเชื่อมต่อระบบเครือข่าย จำนวน 41 คุณลักษณะดังแสดงในตารางที่ 1-3

ตารางที่ 1 คุณลักษณะแสดงเนื้อหาทางโปรโตคอล

ลำดับ	ชื่อคุณลักษณะ	รายละเอียด	ชนิดข้อมูล
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	src_bytes	number of data bytes from source to destination	continuous
5	dst_bytes	number of data bytes from destination to source	continuous
6	flag	normal or error status of the connection	discrete
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

ตารางที่ 2 คุณลักษณะแสดงเนื้อหาของการเชื่อมต่อ

ลำดับ	ชื่อคุณลักษณะ	รายละเอียด	ชนิดข้อมูล
1	hot	number of ``hot'' indicators	continuous
2	num_failed_logins	number of failed login attempts	continuous
3	logged_in	1 if successfully logged in; 0 otherwise	discrete
4	num_compromised	number of ``compromised'' conditions	continuous
5	root_shell	1 if root shell is obtained; 0 otherwise	discrete
6	su_attempted	1 if ``su root'' command attempted; 0 otherwise	discrete
7	num_root	number of ``root'' accesses	continuous
8	num_file_creations	number of file creation operations	continuous
9	num_shells	number of shell prompts	continuous
10	num_access_files	number of operations on access control files	continuous
11	num_outbound_cmds	number of outbound commands in an ftp session	continuous
12	is_hot_login	1 if the login belongs to the ``hot'' list; 0 otherwise	discrete
13	is_guest_login	1 if the login is a ``guest''login; 0 otherwise	discrete

ตารางที่ 3 ลักษณะแสดงปริมาณการจราจรของแพ็กเกตเมื่อคำนวณด้วยวิธี two-second time window

ลำดับ	ชื่อคุณลักษณะ	รายละเอียด	ชนิดข้อมูล
1	count	number of connections to the same host as the current connection in the past two seconds Note: The following features refer to these same-host connections.	continuous
2	serror_rate	% of connections that have ``SYN'' errors	continuous
3	rerror_rate	% of connections that have ``REJ'' errors	continuous
4	same_srv_rate	% of connections to the same service	continuous
5	diff_srv_rate	% of connections to different services	continuous
6	srv_count	number of connections to the same service as the current connection in the past two seconds Note: The following features refer to these same-service connections.	continuous
7	srv_serror_rate	% of connections that have ``SYN'' errors	continuous
8	srv_rerror_rate	% of connections that have ``REJ'' errors	continuous
9	srv_diff_host_rate	% of connections to different hosts	continuous

แบ่งตามรูปแบบการบุกรุกได้เป็น 5 รูปแบบ คือ 1) Normal 2) DoS (denial-of-service) 3) U2R (unauthorized access to local supervisor privileges) 4) R2L (unauthorized access from a remote machine) และ 5) Probe (surveillance and other probing) ดังแสดงในตารางที่ 4 และ 5

ตารางที่ 4 การกระจายตัวของชุดข้อมูลใน KDD-99 ตามชนิดการบุกรุก

Normal		Dos		R2L		Probe		U2R	
ชนิด	จำนวน	ชนิด	จำนวน	ชนิด	จำนวน	ชนิด	จำนวน	ชนิด	จำนวน
normal	60,593	apache2	794	guess_passwd	4,367	Ipsweep	306	buffer_overflow	22
		pod	87	multihop	18	portsweep	354	loadmodule	2
		smurt	164,091	named	17	saint	736	perl	2
		back	1,098	Phf	2	mscan	1,053	ps	16
		land	9	sendmail	17	nmap	84	rootkit	13
		mailbomb	5,000	snmpgetattack	7,741	satan	1,633	sqlattack	2
		neptune	58,001	xlock	9			xterm	13
		processta	759	xsnoop	4				
		teardrop	12	ftp_write	3				
		udpstorm	2	httptunnel	158				
				imap	1				
				snmpguess	2,406				
				Spv	-				
				warezclient	-				
				warezmaster	1,602				
				worm	2				

ตารางที่ 5 การกระจายตัวของชุดข้อมูลใน KDD-99 ตามประภทการบุกรุก

การกระจายตัวของข้อมูล	การกระจายตัว					รวม
	Normal	DoS	R2L	Probe	U2R	
จำนวนแถว	60,593	22,853	16,347	4,166	70	111,146
คิดเป็นร้อยละ	58.25	21.97	15.71	4.00	0.07	100

Neighborhood Rough Set

ราฟเซต (rough set: RS) นำเสนอโดย Palawk (1982) อัลกอริทึมราฟเซตจะแบ่งกลุ่มของตัวอย่าง (Samples) ในชุดข้อมูลออกเป็น 2 กลุ่ม คือ พื้นที่บวก (Positive Region) และพื้นที่ขอบเขต (Boundary Region) ราฟเซตถูกออกแบบมาให้ทำงานกับข้อมูลที่มีค่าไม่ต่อเนื่อง (Discrete value) ที่มีค่าอยู่ในเซตจำกัด (Finite set) เท่านั้น แต่เนื่องจากราฟเซตถูกนำมาประยุกต์ใช้อย่างแพร่หลาย ดังนั้นในเวลาต่อมาจึงมีการนำเสนออัลกอริทึมที่ทำให้ราฟเซตสามารถประยุกต์ใช้กับชุดข้อมูลที่มีค่าต่อเนื่องและค่าไม่ต่อเนื่องได้ เรียกว่า “ราฟเซตย่านจุดใกล้เคียง” หรือ Neighborhood Rough Set (NRS) (Hu et al., 2008) ซึ่งสามารถอธิบายการทำงานได้ดังนี้

นิยามที่ 1: กำหนดให้ข้อมูลอยู่รูปแบบตารางตัดสินใจ $IS =< U, A >$ โดย U คือยูนิเวิร์สที่ไม่เป็นเซตว่าง ซึ่งมีสมาชิกเป็น $\{x_1, x_2, \dots, x_N\}$ และ $A = \{C \cup D\}$ คือเซตของคุณลักษณะที่ไม่เป็นเซตว่าง ซึ่งมีสมาชิกเป็น $\{a_1, a_2, \dots, a_m\}$ ใช้สำหรับจำแนกสมาชิกในยูนิเวิร์ส เมื่อ C คุณลักษณะเงื่อนไข (conditional attributes) และ D คือคุณลักษณะตัดสินใจ (decision attributes)

นิยามที่ 2: กำหนดให้ $x \in U$ และ $B \subseteq C$ แล้ว ฟังก์ชันย่านจุดใกล้เคียง (Neighborhood) $\delta_B(x_i)$ ของตัวอย่าง x_i ในเซตย่อย B สามารถนิยามได้ดังนี้

$$\delta_{\mathrm{B}}(x_i) = \{x_j | x_j \in \mathrm{U}, \Delta_{\mathrm{B}}(x_i, x_j) \leq \delta\} \qquad (1)$$

เมื่อ δ คือค่าคงที่ และ Δ คือเมทริกซ์ โดยที่ทุก x_1, x_2, และ x_3 ในยูนิเวิร์ส U จะบรรลุ 3 เงื่อนไขต่อไปนี้ 1) $\Delta_{\mathrm{B}}(x_1, x_2) \geq 0$, และ $\Delta_{\mathrm{B}}(x_1, x_2) = 0$ ก็ต่อเมื่อ $x_1 = x_2$ 2) $\Delta_{\mathrm{B}}(x_1, x_2) = \Delta_{\mathrm{B}}(x_2, x_1)$ และ 3) $\Delta_{\mathrm{B}}(x_1, x_3) \leq \Delta_{\mathrm{B}}(x_1, x_2) + \Delta_{\mathrm{B}}(x_2, x_3)$

พิจารณา x_1 และ x_2 เป็นตัวอย่างที่มีคุณลักษณะขนาด m มิติ $A = \{a_1, a_2, \dots, a_m\}$ และกำหนดให้ $f(x_i, a_i)$ คือ ฟังก์ชันที่ส่งคืนค่าของตัวอย่าง x_i ในมิติที่ *i*-th ของคุณลักษณะ a_i สามารถอธิบายระยะทางระหว่าง x_1 และ x_2 ด้วยฟังก์ชัน Minkowski ดังนี้

$$\Delta_P(x_1, x_3) = (\sum_{i=1}^{m} |f(x_1, a_i) - f(x_2, a_i)|^P)^{1/P} \qquad (2)$$

ซึ่งฟังก์ชันระยะทางนี้ถ้ากำหนดให้ $P = 1$ (Δ_1) จะเทียบเท่ากับฟังก์ชันระยะทาง Manhattan ถ้ากำหนดให้ $P = 2$ (Δ_2) จะเทียบเท่ากับฟังก์ชันระยะทาง Euclidean และเทียบเท่าฟังก์ชันระยะทาง Tehebyshev ถ้ากำหนดให้ $P = \infty$ อย่างไรก็ตาม ข้อมูลใน KDD-99 เป็นข้อมูลที่มีชนิดคุณลักษณะผสมกันระหว่างประเภทตัวเลข (numerical data) กับประเภทกลุ่ม (categorical data) ดังนั้นฟังก์ชันระยะทางที่เหมาะสมกับข้อมูลในลักษณะผสมนี้จึงถูกนำเสนอ (Wang, 1999; Pan and Billings, 2008) เช่น ฟังก์ชัน HEOM (Heterogeneous Euclidean Overlap Metric), VDM (Value Difference Metric), HVDM (Heterogeneous VDM), และ IVDM (Interpolated VDM) เป็นต้น ฟังก์ชัน HEOM เป็นตัวอย่างของฟังก์ชันที่สามารถใช้วัดระยะทางระหว่าง 2 ตัวอย่างที่มีคุณลักษณะแบบผสมได้ มีนิยามดังนี้ (Wang, 1999)

$$HEOM = \sqrt{\sum_{i=1}^{m} w_{a_i} \times d_{a_i}^2(x_{a_i}, y_{a_i})} \qquad (3)$$

เมื่อ m คือ จำนวนของคุณลักษณะ w_{a_i} คือ ค่าน้ำหนักของคุณลักษณะ a_i และ $d_{a_i}(x,y)$ คือ ระยะทางระหว่างตัวอย่าง x และ y โดยที่

$$d_{a_i}(x,y) = \begin{cases} 1, & \text{หากไม่ทราบค่าของ } x \text{ หรือ } y \\ overlap_{a_i}(x,y), & \text{หาก } a_i \text{ คุณลักษณะประเภทกลุ่ม} \\ rn_{diff}{}_{a_i}(x,y), & \text{หาก } a_i \text{ คุณลักษณะประเภทตัวเลข} \end{cases} \quad (4)$$

และ

$$overlap_{a_i}(x,y) = \begin{cases} 0, & \text{ถ้าหาก } x = y \\ 1, & otherwise \end{cases}$$

$$rn_{diff}{}_{a_i}(x,y) = |x-y| / \max_{a_i} - \min_{a_i}$$

นิยามที่ 3: กำหนดให้ $< U, R >$ เป็นปริภูมิของเซตประมาณย่านจุดใกล้เคียง (Neighborhood Approximation) เมื่อ U คือ ยูนิเวิร์สของกลุ่มตัวอย่าง R คือ ความสัมพันธ์ย่านจุดใกล้เคียงของตัวอย่างใน U โดยที่

$$\forall x, y \in U, R(x,y) = 1 \text{ ถ้าหาก } y \in \delta(x) \text{ ถ้าไม่ } R(x,y) = 0 \quad (5)$$

ดังนั้น $< U, N >$ สำหรับแต่ละ $x \subseteq U$ สามารถแบ่งตัวอย่างออกเป็น 2 เซตย่อย คือ เซตประมาณล่าง (Lower Approxi-mation: $\underline{N}\mathrm{X}$)) และเซตประมาณบน (Upper Approximation: $\overline{N}\mathrm{X}$) ของ X มีนิยามดังนี้

$$\underline{N}\mathrm{X} = \{x_i | \delta(x_i) \subseteq X, x_i \in U\} \quad (6)$$

$$\overline{N}\mathrm{X} = \{x_i | \delta(x_i) \cap X \neq \emptyset, x_i \in U\} \quad (7)$$

นิยามที่ 4: กำหนดให้ตารางตัดสินใจย่านจุดใกล้เคียง (Neighborhood Decision Table) $NDT =< U, C, D >$ และ $X_1, X_2, \ldots X_N$ คือ ปริภูมิย่อยของ U/D แล้วเขตประมาณล่างตัดสินใจ $\underline{N_B}D$ และเขตประมาณบนตัดสินใจ $\overline{N_B}D$ ของคุณลักษณะตัดสินใจ D เทียบเคียงกับคุณลักษณะ $B \subseteq C$ มีนิยามดังนี้

$$\underline{N_B}D = \bigcup_{i=1}^{N} \underline{N_B} X_i \quad (8)$$

$$\overline{N_B}D = \bigcup_{i=1}^{N} \overline{N_B} X_i \quad (9)$$

เมื่อ $\underline{N_B}X = \{x_j | \delta_B(x_j) \subseteq X, x_j \in U\}$ และ $\overline{N_B}X = \{x_j | \delta_B(x_j) \cap X \neq \emptyset, x_j \in U\}$

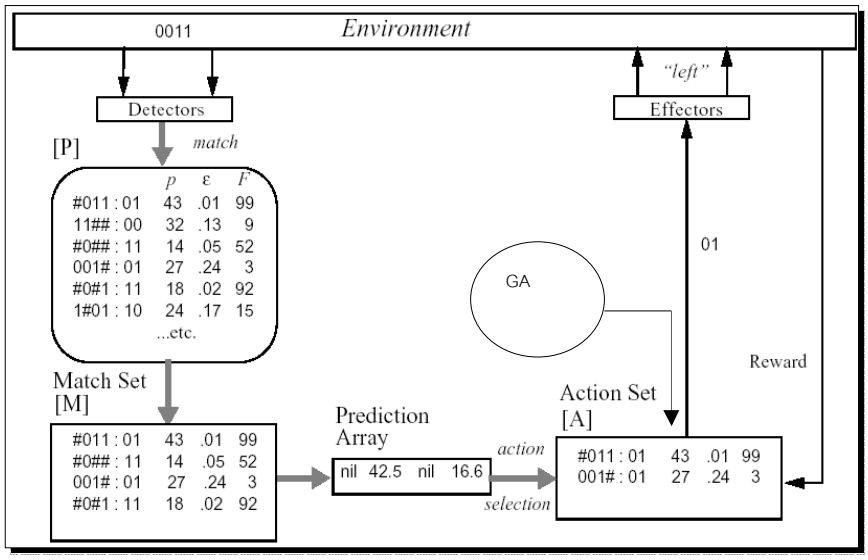
ดังนั้นพื้นที่บริเวณขอบเขตตัดสินใจ (Decision Boundary Region) ของคุณลักษณะตัดสินใจ D เทียบเคียงกับคุณลักษณะเงื่อนไข B มีนิยามดังนี้

$$BN(D) = \overline{N_B}D - \underline{N_B}D \quad (10)$$

อาจกล่าวได้ว่าบริเวณขอบเขตตัดสินใจ $BN(D)$ คือ ปริภูมิย่อยของกลุ่มตัวอย่างที่อยู่เขตบริเวณ $\delta_B(x)$ เดียวกันแต่มีค่าของคุณลักษณะตัดสินใจต่างกัน ในขณะที่เขตประมาณล่างตัดสินใจ $\underline{N_B}D$ คือ ปริภูมิย่อยของกลุ่มตัวอย่างที่อยู่เขตบริเวณ $\delta_B(x)$ เดียวกัน และมีค่าของคุณลักษณะตัดสินใจเหมือนกัน กลุ่มตัวอย่างที่ถูกจัดอยู่ในกลุ่มบริเวณขอบเขตตัดสินใจ $BN(D)$ มีโอกาสที่จะถูกจำแนกผิดได้ง่าย ในขณะที่ตัวอย่างที่ถูกจัดอยู่ในเขตประมาณล่างตัดสินใจ $\underline{N_B}D$ จะถูกจำแนกถูกต้องเสมอ

eXtensive Learning Classifier (XCS)

ระบบเรียนรู้จำแนกประเภท (Learning Classifier System : LCS) นำเสนอโดย Holland (1976) เป็นระบบที่ทำงานแบบกฎเป็นพื้นฐาน (Rule-Based Sstem) ในเวลาต่อมา Wilson (1995) ได้นำมาพัฒนาต่อยอด เรียกว่า (eXtensive Learning Classifier System : XCS) โดยกฎหรือตัวจำแนกประเภท (classifier) จะประกอบไปด้วย 2 ส่วนหลัก ส่วนแรก คือ ส่วน condition-action ของกฎ โดยที่ condition จะถูกแทนด้วยรหัสเทอนารี (ternary) ได้แก่ “0”, “1”, หรือ “#” และ action จะถูกแทนด้วยค่าที่สามารถแบ่งแยกได้ชัดเจนหรือสัญลักษณ์ระบุการกระทำ เช่น “เปิด”, “ปิด”, “เลี้ยวซ้าย”, หรือ “เลี้ยวขวา” เป็นต้น และส่วนหลัง คือ พารามิเตอร์เรียนรู้ของกฎ มีดังต่อไปนี้ 1) prediction payoff (p) เป็นค่าตัวเลขทำนาย



ภาพที่ 1 การทำงานของ XCS

ผลตอบแทนที่จะได้รับ (reward) จากการกระทำ action ของตัวจำแนกประเภทตัวนั้นออกไปยังสิ่งแวดล้อม 2) error (ε) เป็นค่าเฉลี่ยของความผิดพลาดในการทำนายผลตอบแทน 3) fitness (F) เป็นค่าระบุความน่าเชื่อถือของตัวจำแนกประเภท และ 4) niche size estimate (σ) เป็นพารามิเตอร์ที่บ่งบอกการมีส่วนร่วมของตัวจำแนกประเภท การเรียนรู้ของตัวจำแนกประเภท XCS ดังแสดงตามภาพที่ 1 มีขั้นตอนดังต่อไปนี้

1) เริ่มต้นที่ข้อมูลสถานะของสิ่งแวดล้อม $X = \{x_1, x_1, \ldots, x_n\}$ ถูกส่งผ่าน Detector เข้ามาเปรียบเทียบกับส่วน condition=$\{c_1, c_1, \ldots, c_n\}$ ของแต่ละตัวจำแนกประเภทที่มีอยู่ในฐานข้อมูลของกฎ [P] (หรือเรียกว่าประชากรของตัวจำแนกประเภท) ถ้ามีตัวจำแนกประเภทใด ๆ ที่มีความสอดคล้องกับอินพุต X ระบบจะนำตัวจำแนกประเภทตัวดังกล่าวไปใส่ไว้ใน Match Set [M] ถ้าผลการเปรียบเทียบไม่มีตัวจำแนกประเภทใด ๆ ใน [P] ที่มีความสอดคล้องกับอินพุต X ระบบก็จะสร้างตัวจำแนกประเภทที่มีความสอดคล้องกับอินพุต X ขึ้นมาใหม่แล้วนำไปใส่ใน [P] และ [M]

2) ระบบจะเลือก action ที่เหมาะสมจากกลุ่มของตัวจำแนกประเภทที่เป็นสมาชิกใน [M] โดยพิจารณาจากค่าพารามิเตอร์เรียนรู้ของตัวจำแนกประเภทนั้น ๆ เมื่อได้ action a(t) ที่เหมาะสม ระบบจะนำตัวจำแนกประเภททุกตัวใน [M] ที่มี action = a(t) ไปใส่ลงใน Action Set [A]

3) นำ action a(t) ส่งไปกระทำกับสิ่งแวดล้อม ระบบจะได้รับค่าตอบแทน r(t) กลับมา

4) นำค่าตอบแทน a(t) มาปรับปรุงพารามิเตอร์เรียนรู้ของแต่ละตัวจำแนกประเภทที่เป็นสมาชิกใน [A] ณ เวลาปัจจุบันโดยอัลกอริทึมของ Widrow-Hoff delta ดังแสดงตามสมการต่อไปนี้

$$p \leftarrow p + \beta(R - p) \tag{11}$$

จากนั้นจะทำการปรับค่า ε ดังนี้

$$\varepsilon \leftarrow \varepsilon + \beta(|R - p| - \varepsilon) \tag{12}$$

ปรับค่า σ

$$\sigma_j \leftarrow \sigma_j + \beta(|[A]| - \sigma_j) \tag{13}$$

ปรับปรุงค่าความเหมาะสม F เริ่มจากหาค่าความ accuracy ของแต่ละตัวจำแนกประเภทคือ κ หลังจากได้ค่า κ จะนำมาหาค่า relative accuracy κ' ซึ่งทั้งสองสามารถคำนวณได้จากสมการดังนี้

$$\kappa = \begin{cases} 1, & if(\varepsilon < \varepsilon_0) \\ \alpha\left(\frac{\varepsilon}{\varepsilon_0}\right)^{-\nu}, & otherwise \end{cases} \quad (14)$$

$$\kappa' = \frac{\kappa}{\sum_{x \in [A]} \kappa_x} \quad (15)$$

หลังจากได้ค่า κ' แล้วก็จะทำการปรับปรุงค่าความเหมาะสม *F* ดังนี้

$$F \leftarrow F + \beta(\kappa' - F) \quad (16)$$

5) XCS จะมีกลไกในการค้นหาความรู้ใหม่สองแบบ คือ discovery และ covering โดยแต่ละแบบจะมีเงื่อนไขในการทำงานในแต่ละรอบของการเรียนรู้ดังนี้

5.1) ถ้าค่าเฉลี่ย time-step (เวลาที่ตัวจำแนกประเภทอยู่ในระบบ) จากทุกๆ ตัวจำแนกประเภทที่อยู่ใน [A] ณ เวลาปัจจุบันมากกว่า θ_{ga} ที่กำหนดไว้ กลไกของ discovery จะทำงาน โดย Genetic Algorithm (GA) จะทำการเลือกตัวจำแนกประเภทต้นแบบ (parents) จำนวน 2 ตัวโดยพิจารณาจากค่าความเหมาะสม (*F*) เพื่อใช้สำหรับสร้างลูกสองตัว (offspring) กระบวนการ GA จะกระทำ mutation ด้วยการเปลี่ยนเป็นค่า allele เป็น wildcard (#) ด้วยค่าความน่าจะเป็น p_μ และกระทำ crossover แบบ single point ด้วยค่าความน่าจะเป็น p_c สำหรับค่า parameter ต่างๆ จะถ่ายทอดเหมือนกับ parent หรือแทนด้วยค่าเฉลี่ย จากนั้นนำ Offspring ที่ได้ไปเพิ่มลงใน [P] แต่ถ้าหากไม่มีที่ว่างใน [P] ก็ให้นำไปแทนที่ตัวจำแนกประเภทตัวเดิมที่มีอยู่ใน [P] โดยการเลือกตัวจำแนกประเภทตัวที่จะถูกแทนที่ด้วย Offspring นั้นจะพิจารณาจากค่า estimated niche size

5.2) กรณีที่ไม่มีตัวจำแนกประเภทตัวใดๆ เลยใน [P] ที่สอดคล้องกับอินพุตเวคเตอร์ กลไกของ covering จะทำงาน โดยจะสร้างตัวจำแนกประเภทขึ้นมาใหม่ โดยที่กำหนดให้ส่วน condition มีความสอดคล้องกับอินพุตเวคเตอร์ (กำหนดให้มี wildcards ตามอัตราส่วน) สุ่มค่า action ให้กับตัวจำแนกประเภทที่สร้างขึ้นมาใหม่ กำหนดค่าเริ่มต้นให้กับพารามิเตอร์ต่างๆ และนำตัวจำแนกประเภทที่สร้างขึ้นมาใหม่ไปแทนที่สมาชิกใน [P] โดยที่การเลือกตัวจำแนกประเภทที่จะถูกแทนที่จะใช้วิธีเดียวกันกับที่ใช้ในขั้นตอน 5.1

6) ทำซ้ำขั้นตอน (1)-(5) จนกว่าระบบจะมีค่าความผิดพลาดของการทำนายต่ำกว่าค่าที่กำหนด หรือครบจำนวน-รอบการทำซ้ำที่กำหนดไว้

วิธีดำเนินการวิจัย

ในหัวข้อนี้จะกล่าวถึงรายละเอียดเกี่ยวกับระบบการตรวจจับการบุกรุกที่คณะผู้วิจัยได้นำเสนอ ดังที่ได้กล่าวไปแล้วราฟเซตถูกได้ถูกนำมาประยุกต์ใช้ในการตรวจจับการบุกรุก (Lee et al., 2000; Cui-Juan Liu, 2007; Mun et al., 2009) อย่างไรก็ตามราฟเซตได้ถูกออกแบบมาสำหรับใช้กับข้อมูลที่ไม่ต่อเนื่อง ดังนั้นการนำราฟเซตมาใช้กับชุดข้อมูลการบุกรุก KDD-99 ซึ่งเป็นข้อมูลที่ประกอบด้วยคุณลักษณะ (attribute) ที่มีค่าต่อเนื่องและไม่ต่อเนื่อง จึงต้องมีกระบวนการแปลงค่าหรือการทำดีสครีตไตเซชันให้เป็นค่าไม่ต่อเนื่อง (discretization) ก่อนการนำไปใช้งาน ซึ่งกระบวนการดังกล่าวอาจจะเกิดการสูญเสียข่าวสารของข้อมูลได้ และอาจจะส่งผลกระทบต่อประสิทธิภาพของการตรวจจับผู้บุกรุกได้ นอกจากนั้นชุดข้อมูล KDD-99 เป็นข้อมูลที่มีการเหลื่อมทับกันของกลุ่มชนิดการบุกรุกค่อนข้างมาก ทำให้ตัวจำแนกประเภทต้องใช้เวลาในการเรียนรู้ค่อนข้างมากและมีผลกระทบต่อประสิทธิภาพของการจำแนกผู้บุกรุกออกจากผู้ใช้งานปกติได้เช่นกัน (Hu, 2010) ด้วยเหตุนี้คณะวิจัยจึงได้นำเสนอเทคนิคการทำดิสครีตไตเซชันโดยใช้ NRS เนื่องจาก NRS ถูกออกแบบมาให้ใช้กับข้อมูลที่มีคุณลักษณะที่มีค่าทั้งแบบต่อเนื่องและแบบไม่ต่อเนื่อง และยังมีกลไกในการใช้ข่าวสารที่มีประโยชน์ของบริเวณขอบเขตอีกด้วย ดังนั้นคณะวิจัยจึงมีสมมติฐานว่าการใช้ NRS มาช่วยในกระบวนทำดิสครีตไตเซชันจะทำให้ได้ชุดข้อมูลที่มีลักษณะที่ดีสำหรับป้อนให้ตัวเรียนรู้จำแนกประเภทได้

การกำหนดเขตบุกรุก

งานวิจัยนี้มีวัตถุประสงค์คือต้องการแยกพฤติกรรมของผู้บุกรุก (Abnormal behavior) ออกจากผู้ใช้งานปกติ (Normal behavior) โดยใช้ตัวเรียนรู้

จำแนกประเภท XCS จากที่ได้กล่าวไว้ในหัวข้อที่ผ่านมาเป็นที่เข้าใจตรงกันว่าความซับซ้อนของการจำแนกข้อมูลมีผลมาจากอาณาเขตของข้อมูลที่มีความขัดแย้งกัน (inconsistent regions) เช่น มีการเหลื่อมทับกัน (overlap regions) หรือ มีบริเวณขอบเขตตัดสินใจ (decision boundary regions) เป็นต้น (Hu et al., 2010) ดังนั้นในงานวิจัยนี้มุ่งเน้นการสร้างความเด่นชัดของข้อมูลด้วยการกำหนดระดับของบริเวณผิดปกติ (Abnormal Regions) ก่อนนำข้อมูลดังกล่าวเข้าสู่ขั้นตอนการเรียนรู้ของตัวเรียนรู้จำแนกประเภท XCS ต่อไป

นิยามที่ 5: กำหนดให้ $NDT =< U, C, D >$, $x_i \in U$, $\delta(x_i)$ คือ ย่านจุดใกล้เคียงของ x_i ตามนิยามในสมการ (1) และ $P\left(\omega_j \middle| \delta(x_i)\right), j = 1,2, \dots, c$ คือ ค่าความน่าจะเป็นของคลาส ω_j ดังนั้นคลาสตัดสินใจของย่านจุดใกล้เคียง $\delta(x_i)$ สามารถเขียนแทนด้วย $ND_B(x_i) = \omega_l$ ถ้าหาก $P(\omega_l|\delta(x_i)) = max_j P(\omega_j|\delta(x_i))$ เมื่อ $P(\omega_l|\delta(x_i)) = n_j/K$, K คือ จำนวนสมาชิกในย่านจุดใกล้เคียง $\delta(x_i)$ และ n_j คือ จำนวนสามาชิกในย่านจุดใกล้เคียง $\delta(x_i)$ ที่มีคลาสเป็น ω_j

จากนิยามที่ 5 สรุปได้ว่า $ND_B(x_i)$ คือ คลาสที่ถูกกำหนดให้กับ x_i ในกรณีที่ $x_i \notin \underline{N_B}D$ ไม่อยู่ในเขตประมาณล่าง โดยพิจารณาจากความน่าจะเป็นสูงสุดในย่านจุดใกล้เคียง $\delta(x_i)$ และ $ND_B(x_i) = \omega(x_i)$ เสมอ ถ้าหาก $x_i \in \underline{N_B}D$ อยู่ในเขตประมาณล่าง เมื่อ $\omega(x_i)$ คือ คลาสจริงของ x_i

นิยามที่ 6: กำหนดให้เซตของคลาสตัดสินใจ $D = \{Normal, DoS, U2R, R2L, Probe\}$ แล้ว x_i จะอยู่ในเขตปลอดภัย (Safety Regions) ถ้าหาก $x_i \in \underline{N_B}D$ และ $\omega(x_i) = Normal$, กำหนดให้ x_i อยู่ในเขตเฝ้าระวัง (Monitoring Regions) ถ้าหาก $x_i \in \overline{N_B}D$ และ $ND_B(x_i) = Normal$, และกำหนดให้ x_i อยู่ในเขตผิดปกติ (Abnormal Regions) ถ้า x_i ไม่เป็นสมาชิกทั้งในเขตปลอดภัยและเขตเฝ้าระวัง

ชุดข้อมูล FTP-Only

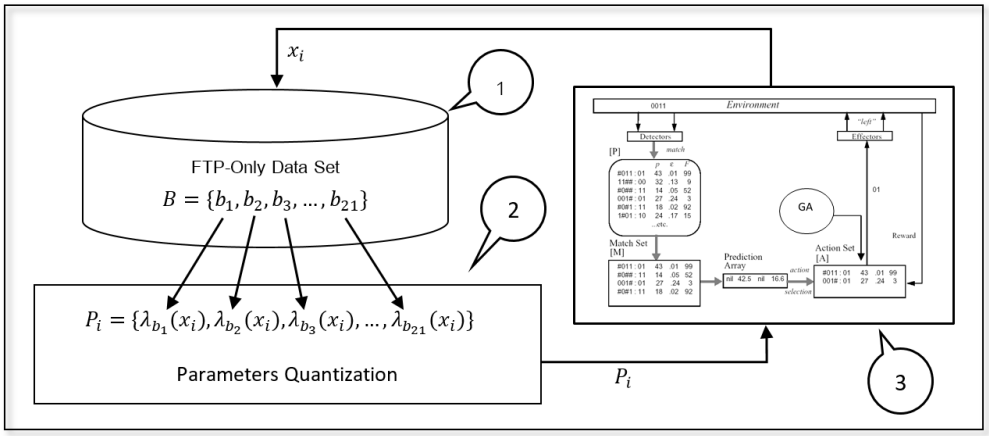
ในงานวิจัยนี้ใช้ชุดข้อมูล FTP-only (Shafi et al., 2006) ซึ่งเป็นชุดข้อมูลชุดข้อมูลที่ย่อยมาจากชุดข้อมูล KDD-99 โดยนำเฉพาะข้อมูลที่มีการเชื่อมต่อระบบเครือข่ายด้วยโปรโตคอล FTP (port 21) ในงานวิจัยของ Shafi และคณะ (2006) ได้ทำการลดจำนวนคุณลักษณะลดลงเหลือ 29 คุณลักษณะ $X = \{x_1, x_2, \dots, x_{21}\}$ (จากเดิม 41 คุณลักษณะ) โดยประกอบด้วยคุณลักษณะแบบสัญลักษณ์ (symbolic) แบบตัวเลขต่อเนื่อง (continuous) และแบบตัวเลข (discrete) คุณลักษณะที่เป็นแบบสัญลักษณ์ได้ทำการปรับให้เป็นตัวเลขก่อน หลังจากนั้นทำการปรับค่าในทุกคุณลักษณะให้มีค่าอยู่ในช่วงตั้งแต่ 0 ถึง 1 (normalization) จำนวนระเบียนของข้อมูลในชุดข้อมูล FTP-Only มีการกระจายตัวดังรายละเอียดในตารางที่ 6

ตารางที่ 6 การกระจายตัวของชุดข้อมูลใน FTP-Only

การกระจายตัวของข้อมูล	การกระจายตัว					รวม
	Normal	DoS	R2L	Probe	U2R	
จำนวนแถว	495	161	954	7	18	1,635
คิดเป็นร้อยละ	30.28	9.85	58.35	0.43	1.10	100

ระบบตรวจจับการบุกรุกที่นำเสนอ

โครงสร้างและการทำงานของระบบตรวจจับการบุกรุกที่ผู้วิจัยได้นำเสนอดังแสดงในภาพที่ 2



ภาพที่ 2 การทำงานของระบบการตรวจจับการบุกรุกที่นำเสนอ

การทำงานของระบบแบ่งออกเป็น 3 ส่วนดังแสดงต่อไปนี้

1) ชุดข้อมูล FTP-Only ซึ่งประกอบด้วย 21 คุณลักษณะ $B = \{b_1, b_2, \dots, b_{21}\}$ และมี 1,635 ตัวอย่าง

2) ขั้นตอนการกำหนดระดับความผิดปกติ (Abnormal Quantization: AQ) ให้กับทั้ง 21 คุณลักษณะ โดยนำแต่ละคุณลักษณะ (ข้อมูล 1 มิติ) ไปประมวลผลด้วย NRS เพื่อกำหนดขอบเขตพฤติกรรมการบุกรุก (เขตปลอดภัย เขตเฝ้าระวัง และเขตผิดปกติ) ตามนิยามที่ 6 จากนั้นใช้ฟังก์ชัน $\lambda_B(x_i)$ เป็นตัวดำเนินการเข้ารหัสดังนี้

$$\lambda_B(x_i) = \begin{cases} 000, & \text{หาก } x_i \text{ อยู่ในเขตปลอดภัย} \\ 001, & \text{หาก } x_i \text{ อยู่ในเขตเฝ้าระวัง และ } \omega(x_i) = Normal \\ 010, & \text{หาก } x_i \text{ อยู่ในเขตเฝ้าระวัง และ } \omega(x_i) \neq Normal \\ 011, & \text{หาก } x_i \text{ อยู่ในเขตผิดปกติ และ } ND_B(x_i) = DoS \\ 100, & \text{หาก } x_i \text{ อยู่ในเขตผิดปกติ และ } ND_B(x_i) = Probe \\ 101, & \text{หาก } x_i \text{ อยู่ในเขตผิดปกติ และ } ND_B(x_i) = U2R \\ 110, & \text{หาก } x_i \text{ อยู่ในเขตผิดปกติ และ } ND_B(x_i) = R2L \end{cases} \quad (17)$$

เมื่อ $\lambda_B(x_i)$ คือ ฟังก์ชันกำหนดระดับความผิดปกติ (Abnormal Quantization: AQ) ของตัวอย่าง x_i ในคุณลักษณะ B

3) ตัวเรียนรู้จำแนกประเภท ทำหน้าที่รับอินพุตเวคเตอร์ P เข้าไปประมวลผลตามขั้นตอนการเรียนรู้ของ XCS เมื่อ อินพุตเวคเตอร์ P สร้างตามอัลกอริทึมดังแสดงในภาพที่ 3

```
1  C, the conditional attributes
2  D, the decision attributes
3  x_i, the sample in universe
4
5  P ← { }
6  For each B ∈ C
7    compute N_B D
8    compute ¬N_B D
9    compute ND_B(x_i)
10   P ← P ∪ {λ_B(x_i)}
11 Return P
```

ภาพที่ 3 อัลกอริทึมของการสร้างชุดพารามิเตอร์

อัลกอริทึมการสร้างชุดพารามิเตอร์ *P* เริ่มต้นด้วยกำหนดให้ *P* เป็นเซตว่าง (บรรทัดที่ 5) รอบวนในแต่ละรอบการทำงานจะดำเนินการสำหรับแต่ละคุณลักษณะ (บรรทัดที่ 6) สำหรับแต่ละคุณลักษณะ *B* ดำเนินการคำนวณเขตประมาณล่าง (บรรทัดที่ 7) เขตประมาณบน (บรรทัดที่ 8) และคลาสตัดสินใจของย่านจุดใกล้เคียง $ND_B(x_i)$ เพื่อใช้ในการกำหนดเขตปลอดภัย เขตเฝ้าระวัง และเขตผิดปกติ (บรรทัดที่ 9) ค่าของคุณลักษณะผ่านการกำหนดรหัสแล้วถูกเพิ่มเข้าไป *P* (บรรทัด 10) ค่าสุดท้ายของ *P* จะถูกส่งคืนไปยังผู้เรียก (บรรทัด 11)

ผลการวิจัย

ในการวิจัยนี้ได้ทำการทดสอบประสิทธิภาพในการจำแนกพฤติกรรมของผู้ใช้งานปกติ (Normal User) และผู้บุกรุก (Abnormal User) ชุดข้อมูลที่นำมาใช้ในการทดสอบ คือ ชุดข้อมูล FTP-Only โดยแบ่งรูปแบบ

พฤติกรรมออกเป็น 5 รูปแบบ (Normal, DoS, R2L, Probe, และ U2R) ดังแสดงในตารางที่ 6 ทำการทดสอบบนตัวเรียนรู้จำแนกประเภทจำนวน 6 ตัว ได้แก่ RF (Random Forest) C4.5 MLP(Multi-Layer Perceptron) Naïve bays RBF(Radial Basis Function Network) และ XCS (eXtensive Learning Clasifier System) การทดลองแบ่งออกเป็น 2 กรรมวิธี คือ 1) นำชุดข้อมูล FTP-Only ต้นฉบับให้ตัวเรียนรู้จำแนกประเภททั้ง 6 ตัว ทำการเรียนรู้และจำแนก เรียกกรรมวิธีนี้ว่า Original Method (Org.) และ 2) นำชุดข้อมูลมาผ่านกระบวนการกำหนดระดับความผิดปกติที่นำเสนอ (ดูสมการที่ 16) โดยกำหนดค่า δ = 0.002 (ดูสมการที่ 1) ก่อนนำส่งให้ตัวเรียนรู้จำแนกประเภททั้ง 6 ตัวทำการเรียนรู้และจำแนก เรียกกรรมวิธีนี้ว่า Proposed Method (Prp.) การทดสอบระบบใช้โปรแกรม weka (Lan and Eibe 2005) เวอร์ชัน 3.4 ด้วยวิธี 10-fold cross validation

การวัดประสิทธิภาพของแต่ละตัวเรียนรู้จำแนกประเภทวัดจาก 7 ตัวชี้วัด ได้แก่ 1) TP Rate (True Positive Rate) หรืออาจจะเรียกว่าค่า Recall เป็นการวัดความสามารถในการค้นคืนข้อมูลที่อยู่ในคลาสจริง โดยหาได้จากอัตราส่วนของการทำนายข้อมูลที่อยู่ในคลาสจริงได้ถูกต้องเทียบกับจำนวนข้อมูลทั้งหมดของคลาสจริง สามารถคำนวณได้จากสูตร TP Rate = TP/(TP+FN) เมื่อ TP คือค่า True Positive และ FN คือค่า False Negative 2) FP Rate (False Positive Rate) เป็นการวัดอัตราส่วนความผิดพลาดในการทำนายข้อมูลที่ไม่ได้อยู่ในคลาสจริง โดยคำนวณได้จากสูตร FP Rate = FP/(FP+TN) เมื่อ FP คือค่า False Positive และ TN คือค่า True Negative 3) Precision เป็นการวัดความแม่นยำของการทำนายข้อมูลที่อยู่ในคลาสจริง โดยหาได้จากอัตราส่วนของการทำนายข้อมูลที่อยู่ในคลาสจริงได้ถูกต้องเทียบกับจำนวนข้อมูลที่ทำนายว่าเป็นคลาสจริงทั้งหมด สามารถคำนวณได้จากสูตร Precision = TP/(TP+FP) 4) F-measure เป็นการวัดค่าความแม่นยำโดยดูจากผลเฉลี่ยของ Precision เทียบกับ Recall สามารถคำนวณได้จากสูตร F-meature = (2 * Precision * Recall)/(Precision + Recall) 5) Corrected Classify (CC) คือจำนวนข้อมูลที่อยู่ในคลาสจริงที่ทำนายได้ถูกต้อง 6) Incorrected Classify (IC) คือ จำนวนข้อมูลที่อยู่ในคลาสจริงที่ทำนายไม่ถูกต้อง (ทำนายว่าอยู่คลาสอื่น) และ 7) Accuracy เป็นการประเมินประสิทธิภาพการจำแนกประเภทข้อมูลโดยรวมทุกคลาสของโมเดล สามารถคำนวณได้จากสูตร Accuracy = CC/IC

ตารางที่ 7 แสดงผลการทดสอบประสิทธิภาพของระบบทั้ง 7 ตัวชี้วัดของแต่ละกรรมวิธีการทดลอง (Org. และ Prp.) โดยระบุประสิทธิภาพด้านต่าง ๆ ของแต่ละรูปแบบการบุกรุก ได้แก่ Normal DoS R2L Probe และ U2R จากผลการทดลองพบว่า ความสามารถในการจำแนกพฤติกรรม Normal วิธีการที่นำเสนอ (Prp.) มีประสิทธิภาพเหนือกว่าวิธีการที่ใช้ชุดข้อมูล FTP-Only ต้นฉบับ (Org.) ถึงแม้ว่าในตัวเรียนรู้จำแนกประเภท C4.5 กรรมวิธี Org. จะมีค่า TP Rate (0.980) สูงกว่ากรรมวิธี Prp. (0.978) แต่เมื่อพิจารณาประสิทธิภาพการจำแนกข้อมูลโดยรวมทุกคลาสแล้ว กรรมวิธี Prp. จะมี Accurary (98.899) สูงกว่ากรรมวิธี Org. (98.532) ความสามารถในการจำแนกพฤติกรรมการบุกรุกแบบ DoS และ R2L กรรมวิธี Prp. มีประสิทธิภาพเหนือกว่ากรรมวิธี Org. ในทุกตัวเรียนรู้จำแนกประเภท ส่วนความสามารถในการจำแนกพฤติกรรมการบุกรุกแบบ Probe และ U2R ทั้งกรรมวิธี Org. และกรรมวิธี Prp. มีประสิทธิภาพไม่ต่างกันนัก สาเหตุเนื่องมาจากจำนวนข้อมูลตัวอย่างของรูปแบบพฤติกรรมการบุกรุกทั้ง 2 รูปแบบนี้มีจำนวนค่อนข้างน้อยเมื่อเทียบกับจำนวนข้อมูลทั้งหมดในชุดข้อมูล FTP-Only และเมื่อพิจารณาจากประสิทธิภาพในการจำแนกข้อมูลโดยรวมทุกคลาส (Accuracy) แล้ว กรรมวิธี Prp. ให้ค่าสูงกว่ากรรมวิธี Org. ในทุกตัวเรียนรู้จำแนกประเภท โดยตัวจำแนกประเภท XCS ให้ค่า Accuracy

ตารางที่ 7 การเปรียบเทียบประสิทธิภาพในการจำแนกพฤติกรรมการบุกรุก

Classifiers \ Attack type		Normal		DoS		R2L		Probe		U2R		Accuracy	
		Org	Prp.	Org	Prp.	Org	Prp	Org	Prp	Org	Prp	Org.	Prp.
RF	TP Rate	0.9	**0.99**	1.0	1.0	0.9	**0.9**	0.7	**0.8**	**0.7**	0.6	98.7	**99.2**
	FP Rate	0.0	**0.00**	0.0	0.00	0.0	**0.0**	0.0	0.0	0.0	**0.0**	77	**66**
	Precision	0.9	**0.99**	0.9	0.99	0.9	**0.9**	0.7	**0.8**	**0.6**	**0.7**		
	F-measure	0.9	0.99	0.9	0.99	0.9	**0.9**	0.7	**0.8**	**0.7**	0.6		
	Corrected	487	**494**	161	161	949	**951**	5	**6**	**13**	11		
	Incorrected	8	**1**	0	0	6	**4**	2	**1**	**4**	6		
C4.5	TP Rate	**0.9**	0.97	0.9	**1.0**	0.9	**0.9**	0.7	**1.0**	0.7	**1.0**	98.5	**98.8**
	FP Rate	0.0	**0.0**	0.0	**0.0**	0.0	**0.0**	0.0	0.0	**0.0**	0.0	32	**99**
	Precision	0.9	**1.0**	0.9	**1.0**	0.9	**0.9**	**0.8**	0.7	0.6	0.6		
	F-measure	0.9	**0.98**	0.9	**1.0**	0.9	**0.9**	0.7	**0.8**	0.7	0.7		
	Corrected	**485**	484	160	**161**	948	948	5	**7**	13	17		
	Incorrected	**10**	11	1	**0**	7	7	2	**0**	4	0		
Naïve bays	TP Rate	0.9	**0.98**	1.0	1.0	0.6	**0.9**	1.0	1.0	**0.8**	0.2	75.5	**98.7**
	FP Rate	0.1	**0.0**	0.0	0.0	0.0	**0.0**	**0.0**	0.0	0.1	**0.0**	35	**77**
	Precision	0.7	**1.0**	1.0	1.0	0.9	**0.9**	**1.0**	0.7	0.0	**0.5**		
	F-measure	0.8	**0.99**	1.0	1.0	0.7	**0.9**	**1.0**	0.8	0.1	**0.3**		
	Corrected	448	**489**	161	161	604	**954**	7	7	**15**	13		
	Incorrected	47	**6**	0	0	351	**1**	0	0	**2**	4		
MLP	TP Rate	0.9	**0.99**	1.0	**1.0**	0.9	**0.9**	0.7	**1.0**	0.8	**0.8**	98.2	**99.5**
	FP Rate	0.0	**0.0**	0.0	**0.0**	0.0	**0.0**	0.0	**0.0**	0.0	**0.0**	26	**11**
	Precision	0.9	**1.0**	0.9	**1.0**	0.9	**0.9**	0.8	**1.0**	0.6	**0.7**		
	F-measure	0.9	**0.99**	0.9	**1.0**	0.9	**0.9**	0.7	**1.0**	0.7	**0.8**		
	Corrected	482	**494**	161	161	944	**950**	5	**7**	14	**15**		
	Incorrected	13	**1**	0	0	11	**5**	2	**0**	3	**2**		
RBF	TP Rate	0.8	**0.99**	1.0	1.0	0.9	**0.9**	0.7	**0.8**	0.0	**0.4**	89.7	**98.6**
	FP Rate	0.0	**0.00**	0.0	0.0	0.1	**0.0**	**0.0**	0.0	**0.0**	0.0	25	**54**
	Precision	0.8	**0.99**	1.0	1.0	0.8	**0.9**	**1.0**	0.7	0.3	**0.4**		
	F-measure	0.8	**0.99**	1.0	1.0	0.9	**0.9**	**1.0**	0.8	0.1	**0.4**		
	Corrected	398	**490**	161	161	902	**949**	5	**6**	1	**7**		
	Incorrected	97	**5**	0	0	53	**6**	2	**1**	16	**10**		
XCS	TP Rate	0.9	**0.99**	1.0	1.0	0.9	**0.9**	0.8	**1.0**	0.7	**0.9**	98.3	**99.7**
	FP Rate	0.0	**0.0**	0.0	0.0	0.0	**0.0**	0.0	**0.0**	0.0	**0.0**	49	**55**
	Precision	0.9	**1.0**	1.0	1.0	0.9	**0.9**	0.8	**1.0**	0.6	**0.8**		
	F-measure	0.9	**0.99**	1.0	1.0	0.9	0.9	0.8	**1.0**	0.7	**0.9**		
	Corrected	488	**494**	161	161	940	**953**	6	**7**	13	**16**		
	Incorrected	7	**1**	0	0	15	**2**	1	**0**	4	**1**		

สูงสุด อีกทั้งเมื่อพิจารณาจากค่าเฉลี่ยของค่า F-measure โดยรวมทุกคลาส (Normal DoS R2L Probe และ U2R) แล้วตัว-เรียนรู้จำแนกประเภท XCS ยังให้ค่าที่สูงที่สุดอีกด้วย (อาจจะกล่าวได้ว่าว่า XCS มีความแม่นยำในการทำนายและมีความสามารถในการค้นคืนข้อมูลที่ครบถ้วนสูง) นั่นคือ XCS เป็นตัวเรียนรู้จำแนกประเภทที่เหมาะกับชุดข้อมูล FTP-Only มากที่สุดเมื่อเทียบกับตัวเรียนรู้จำแนกประเภท RF C4.5 Naïve bays MLP และ RBF

สรุปและวิจารณ์ผล

ระบบตรวจจับการบุกรุกนับว่ามีความสำคัญมากขึ้นตามลำดับ เนื่องจากข้อมูลข่าวสารได้ถูกสื่อสารกันผ่านเครือข่ายอินเทอร์เน็ต ซึ่งอาจจะมีผู้ประสงค์ร้ายปะปนอยู่ในกลุ่มผู้ใช้งานอินเทอร์เน็ต อีกทั้งรูปแบบของการบุกรุกนับวันก็ยิ่งมีรูปแบบที่หลากหลายมากขึ้นเป็นลำดับ ดังนั้นการประยุกต์ใช้เทคนิคทางการเรียนรู้เครื่องจักร (Machine Learning) มาสร้างระบบตรวจจับการบุกรุกที่สามารถปรับตัวเองไปตามรูปแบบการบุกรุกแบบใหม่ ๆ ได้จึงเป็นที่สนใจเป็นอย่างมาก จากผลการทดลองในหัวข้อหัวที่ผ่านมาพบว่าตัวเรียนรู้จำแนกประเภท XCS เป็นตัวเรียนรู้จำแนกประเภทที่มีประสิทธิภาพสูง ดังนั้นในงานวิจัยนี้จึงได้ประยุกต์ใช้ตัวเรียนรู้จำแนกประเภท XCS เพื่อสร้างระบบตรวจจับการบุกรุก อย่างไรก็ตามตัว-เรียนรู้จำแนกประเภท XCS ได้ถูกออกแบบมาให้ใช้กับชุดข้อมูลที่เป็นค่าไบนารี่ (Binary) หรือค่าที่ไม่ต่อเนื่องเท่านั้น ในขณะที่ชุดข้อมูลผู้บุกรุก KDD-99 เป็นชุดข้อมูลผสม (mixture feature) ระหว่างค่าต่อเนื่องและค่าไม่ต่อเนื่อง ดังนั้นการนำชุดข้อมูล KDD-99 มาใช้กับตัวเรียนรู้จำแนกประเภท XCS จึงต้องทำดิสครีตไตเซชัน (discretization) เสียก่อนเพื่อแปลงข้อมูลจากค่าต่อเนื่องไปเป็นข้อมูลไม่ต่อเนื่อง ซึ่งกระบวนการดังกล่าวอาจจะที่ให้เกิดการสูญเสียข่าวสารของชุดข้อมูลได้ ด้วยเหตุนี้ผู้วิจัยจึงได้นำเสนอการใช้ทฤษฎีราฟเซตย่านจุดใกล้เคียง (NRS) มาประยุกต์ใช้ในขั้นตอนของการทำดิสครีตไต-เซชันชุดข้อมูล โดยเรียกกระบวนการที่นำเสนอขึ้นมาใหม่นี้ว่า “การกำหนดระดับความผิดปกติ” (ดังแสดงในภาพที่ 2 และสมการที่ 16) ผลที่ได้จากวิธีการที่นำเสนอนี้นอกจากจะเป็นการแปลงข้อมูลต่อเนื่องไปเป็นข้อมูลไม่ต่อเนื่องแล้วมันยังเป็นการสร้างความเด่นชัดของข้อมูลในแต่ละคุณลักษณะอีกด้วย จากผลการทดลองแสดงให้เห็นว่าเมื่อนำข้อมูลที่ได้ผ่านกระบวนการกำหนดระดับความผิดปกติที่นำเสนอไปใช้ในตัวเรียนรู้จำแนกประเภทจำนวน 6 ตัว ได้แก่ RF C4.5 MLP RBF XCS และ Naïve bays ปรากฏว่าตัวเรียนรู้จำแนกประเภททั้ง 6 ตัวมีการเรียนรู้ที่ดีขึ้นและมีประสิทธิภาพในการจำแนกพฤติกรรมการบุกรุกเพิ่มมากขึ้น

เอกสารอ้างอิง

Aghdam, M.H., Ghasem-Aghaee, N., and Basiri, M.E. 2009. Text feature selection using ant colony optimization. Expert Systems with Applications. 36: 6843-6853.

Barto, A. G., Sutton, R. S., Anderson, C. W. 1983. Neuron like adaptive elements that can solve difficult learning control problems. IEEE Transactions on Systems Man and Cybernetics. 13(5):834-846

Blum, A.L., and Langley, P. 1997. Selection of relevant features and examples in machine learning. Artificial Intelligence. 97: 245-271.

Bull, L. 2014. Applications of Learning Classifier Systems. Germany: Springer.

Cui-Juan Liu. 2007. The Application of Rough Sets on Network Intrusion Detection. Proceedings of the 6th International Conference on Machine Learning and Cybernetics. Hong Kong, pp.19-22

Dash, M., and Liu, H. 1997. Feature selection for classification. Intelligent Data Analysis: An Int’l J. 1(3): 131-156. Deogun, J.S., Raghavan, V.V. and Sever, H. 1995. Exploiting upper approximation in the rough set Methodology. Proc. First Int’l Conf. Knowledge Discovery and Data Mining.

Dhakar, M., Tiwari, A. 2013. A New Model for Intrusion Detection based on Reduced Error Pruning Technique. International Journal of Computer Network and Information Security: pp. 51-57.

Dowell, C., and Ramstedt, P. 1990. The Computer Watch Data Reduction Tool. Proceedings of the 13th National Computer Security Conference, Washington, D.C.

Farid, D. M., Harbi N. and Rahman M. Z. 2010. “Combining naive bayes and decision tree for adaptive intrusion detection”, International Journal of Network Security & Its Applications (IJNSA), 2(2): 12-25

Habra, J., Charlier, B., Mounji, A., Mathieu, I. 1992. ASAX: software architecture and rule based language for universal audit trail analysis. Computer security, Proc. Of ESORICS 92, Toulouse.

Hassanien, A. 2004. Rough set approach for attribute reduction and rule generation: A case of patients with suspected breast cancer. J. Am. Soc. Information Science and Technology. 55(11): 954-962.

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

Holland, J. H. 1975. Adaption in Natural and Artificial Systems. The University of Michigan Press. Ann Arbor.

Holland, J.H. (1976) Adaptation. In R. Rosen & F.M. Snell (eds) Progress in Theoretical Biology, 4. Plenum.

Hossein, S. 2009. Anomaly Intrusion Detection System Using Information Theory, K-NN and KMC Algorithms. Australian Journal of Basic and Applied Sciences: 2581-2597.

Hu, Q. H., Yu, D., Liu, J. F., Wu, C. 2008. Neighborhood-rough-setbased heterogeneous feature subset selection. Inf. Sci. 178(18): 3577-3594

Hu, Q., Pedrycz, W., Yu, D., Lang, J. 2010 SeSelecting Discrete and Continuous Features Based on Neighborhood Decision Error Minimization. IEEE Trans Syst Man Cybern B Cybern. 40(1): 137-150

Ilgun, K., Kemmerer, R. A., Porras, P. A. 1995. State Transition Analysis: A Rule-Based Intrusion Detection. IEEE transaction on software engineering 21 (3):181-199.

Jensen, R. and Shen, Q. 2004. Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based Approaches. IEEE Trans. Knowledge and Data Eng. 16(12): 1457-1471.

Kim, Y., Street, W. and Menczer, F. 2000. Feature selection for unsupervised learning via evolutionary Kira, K. and Rendell, L. 1992. A practical approach to feature selection. Proc. Ninth Int’l Conf. Machine Learning.

Kohavi, R. and John, G. H. 1997. Wrappers for feature subset selection. Artificial Intelligence. 97(1-2): 273-324.

Lan, H. W. and Eibe, F. 2005. Data mining: practical machine learning tools and techniques 2nd Edition. Morgan Kaufmann.

Lee, W., Stolfo, S. J. and Mok, K.W. 2000. Adaptive intrusion detection: a data mining approach. AI Rev. 14(6): 533-567.

Memon, V. I. and Chandel, G. S. 2014. “A Design and Implementation of New Hybrid System for Anomaly Intrusion Detection System to Improve Efficiency”, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, 4(5): 01-07.

Muda, Z., Yassin, W., Sulaiman, M. N., Udzir, N. I. 2011. Intrusion Detection based on K-Means Clustering and Naive Bayes Classification. 7th IEEE International Conference on IT in Asia (CITA)

Mun, G. J., Noh, B. N. and Kim, Y. M. 2009. Enhanced stochastic learning for feature selection in intrusion Classification. Int’l J. Innovative Computing, Information and Control. 5(11): 3625-3635.

Om, H. and Kundu A. 2012. “A hybrid system for reducing the false alarm rate of anomaly intrusion detection system”, Recent Advances in Information Technology (RAIT), IEEE International Conference on Print ISBN:978-1-4577-0694-3, March 15-17, pp. 131-136

Pan, Y., and Billings, S. A. 2008. Neighborhood detection for the identification of spatiotemporal systems. IEEE Trans. Syst., Man, Cybern. B, Cybern. 38(3): 846-854

Parthalain, N., Jensen, R. and Shen, Q. 2010. A distance measure approach to exploring the rough set boundary region for attribute reduction. IEEE Transactions on Knowledge and Data Engineering. 22(3): 305-317.

Pawlak, Z.1982. Rough sets. Int. J. Inf. Comput. Sci. 11: 314-356.

Pawlak, Z. 1991. Rough sets: Theoretical aspects of reasoning about data. Kluwer Academic Publishing.

Power, R. 2002. CSI/FBI computer crime & security survey. Computer Security Journal. 18(2): 7-30.

Red, S., Selvakumar, M., and Sureswaran, R. 2013. Intrusion Detection System in IPv6 Network Based on Data Mining Techniques-Survey. Proc. Of the Second Intl. Conf. On Advances in Computer and Information Technology (ACIT): 130-134.

Richardson. 2009. 14th Annual Computer Crime Security Survey Executive Summary, CSI Computer Crime and Security Survey: [Online] Available at: http://www.gdais.com/index.cfm?acronym =news, [Accessed: September 16, 2010].

Roesch, M., 1999. Snort - Lightweight Intrusion Detection for Networks. 13th USENIX Conference on System Administration, USENIX Association: 229-238.

Sandeep, K., Eugene H. S. 1994. A Pattern Matching Model for Misuse Intrusion Detection. Proceedings of the 17th National Computer Security Conference. pp. 11-21, Baltimore MD, USA.

Shafi, K. H., Abbass, H. A., Zhu, W. 2006. An adaptive rule-based intrusion detection architecture. The Security Technology Conference, the 5th Homeland Security Summit. Canberra, Australia, 19-21 September 2006: pp. 345-355

Shafi, K. H., Abbass, H. A., Zhu, W. 2006. The role of early stopping and population size in XCS for intrusion detection. T.-D. Wang, X. Li, S.-H. Chen, X. Wang, H. Abbass, H. Iba, G. Chen, X. Yao (Eds.), Simulated Evolution and Learning. Springer, Berlin/Heidelberg

Shafi, K. H., Abbass, H. A., Zhu, W. 2007. Intrusion detection with evolutionary learning classifier systems. Natural Computing, December 2007.

Shafi, K. H., Abbass, H. A., Zhu, W. 2007. Real time signature extraction during adaptive rule discovery using UCS. Proceedings of the IEEE Congress on Evolutionary Computation (CEC'07). Singapore, 25-28 September 2007. IEEE Press: pp. 2509-2516

Shafi, K. H., Abbass, H. A., Zhu, W. 2009. Intrusion detection with evolutionary learning classifier systems. Natural Computing. 8(1): 3-27

Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y. and Wang, Z. 2007. A novel feature selection algorithm for text categorization. Expert Systems with Applications. 33: 1-5.

Teresa, F. L., Jagannathan R., Rosanna Lee, Sherry Listgarten, David L. E., Peter G. N., Harold S. J., Al Valdes. 1998. IDES: The Enhanced Prototype - A RealTime Intrusion-Detection Expert System. Technical Report SRI Project 4 185-010, SRI-CSL-88.

Vaccaro H. S. and Liepins G. E. 1990. Detection of anomalous computer session activity. Proceedings IEEE Symposium on Security and Privacy: pp. 280-289

Wang, H. 1999. Nearest neighbors by neighborhood counting. IEEE Trans. Pattern Anal. Mach. Intell. 28(6): 942-953

Wilson, S. W. 1995. Classifier Fitness Based on Accuracy, Evolutionary Computation. 3(2): 149-175