บทที่ 2
ทฤษฎีที่ใช้ในการแก้ปัญหา

ในบทนี้จะกล่าวถึงรายละเอียดของทฤษฎีที่ใช้แก้ปัญหาในงานวิจัย ซึ่งแบ่งออกเป็นสามส่วนหลัก ได้แก่ งานวิจัยที่เกี่ยวกับการวิเคราะห์ข้อมูลจากคุณลักษณะ (Feature Analysis) [17-19] งานวิจัยที่เกี่ยวข้องกับขั้นตอนวิธีเคมีนและการเลือกจุดศูนย์กลางเริ่มต้น [22] และงานวิจัยที่ใช้ในการตัดสินใจบอตเน็ต (Botnet Decision) [8]

2.1 การวิเคราะห์ข้อมูลจากคุณลักษณะ (Feature Analysis) [17-19]

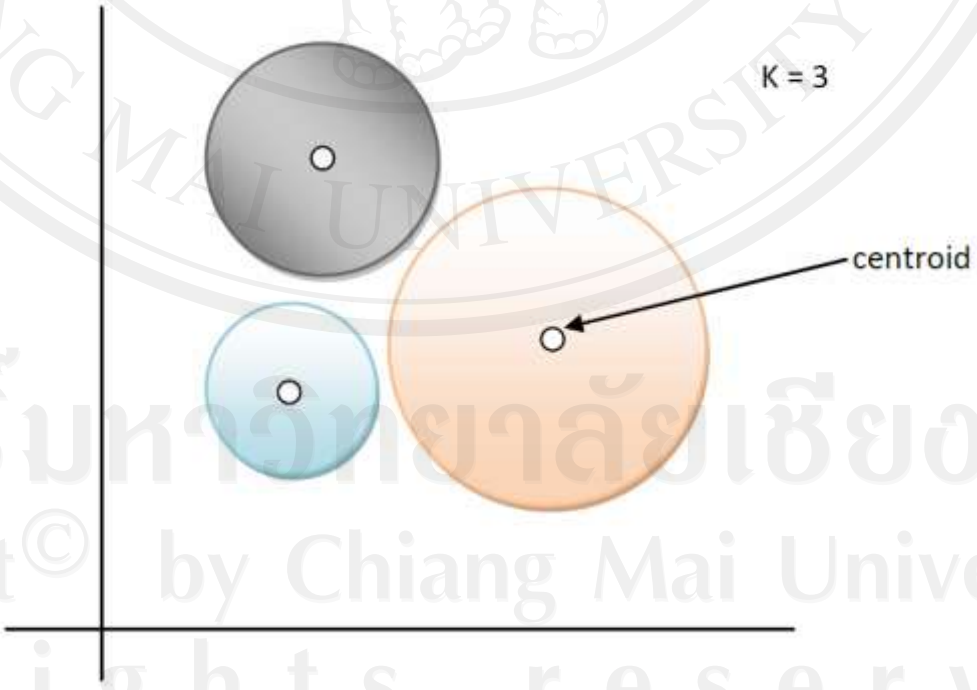
งานวิจัยที่เกี่ยวกับการวิเคราะห์ข้อมูลจากคุณลักษณะ (Feature Analysis) เพื่อใช้ศึกษาพฤติกรรมของทราฟฟิกที่เป็นบอตเน็ตและทราฟฟิกที่ปกติในระบบเครือข่าย ในงานวิจัย [17-19] ใช้การพิจารณาจาก 256 ASCII Characters ใน Packet Payload ซึ่งเป็นวิธีที่นิยมใช้ในปัจจุบัน หรือเรียกวิธีการนี้ว่า การหา Temporal-frequent Metric [8] กำหนดให้เมตริกซ์มีขนาด $n \times 256$ และกำหนดให้มีเวกเตอร์ $< f_1^{t_i}, f_2^{t_i}, \dots, f_{256}^{t_i} >$ เมื่อ $f_j^{t_i}$ คือ ความถี่ของจำนวนอักขระ ASCII ใน Packet Payload และในช่วงเวลา t_i เมื่อ $(j = 1,2,\dots,256 \ ; \ i = 1,2,\dots)$ ในงานวิจัยจะทำการศึกษาพฤติกรรมของบอตเน็ตและพร็อตกราฟ เพื่อดูความถี่เฉลี่ยของทราฟฟิกที่ปกติและทราฟฟิกที่เป็นบอตเน็ต มาใช้ในการพิจารณาแยกทราฟฟิกทั้งสองออกจากกัน ในผลการทดลองพบว่า ค่าความถี่เฉลี่ยในแต่ละไบต์ (Byte) ของตัวอักษรแอสกี (ASCII) ในทราฟฟิกที่เป็นบอตเน็ตและทราฟฟิกที่ปกติมีความแตกต่างกันอย่างชัดเจน ดังนั้นค่าความถี่เฉลี่ยในแต่ละไบต์ (Byte) ของตัวอักษรแอสกี (ASCII) จึงสามารถนำมาวิเคราะห์แยกทราฟฟิกที่เป็นบอตเน็ตจากทราฟฟิกที่ปกติได้ การหา Temporal-frequent Metric [8] หาได้จากสมการที่ (2.1)

$$p_{n\times 256}^{app} = \begin{bmatrix} f_1^{t_1} & f_2^{t_1} & \cdots & f_{256}^{t_1} \\ f_1^{t_2} & f_2^{t_2} & \vdots & f_{256}^{t_2} \\ \vdots & \vdots & \vdots & \vdots \\ f_1^{t_n} & f_2^{t_n} & \cdots & f_{256}^{t_n} \end{bmatrix} \qquad (2.1)$$

2.2 ขั้นตอนวิธีเคมีนและการเลือกจุดศูนย์กลางเริ่มต้น [16]

การจัดกลุ่มข้อมูล (Clustering) คือการหารูปแบบที่สัมพันธ์กันของข้อมูลที่มีการรวมกลุ่มกัน มีจุดประสงค์และเป้าหมายสำคัญคือการค้นหาและจำแนกลักษณะเฉพาะของข้อมูลออกเป็นกลุ่มๆหรือที่เรียกว่าคลัสเตอร์ (Cluster) โดยการแบ่งข้อมูลออกเป็นคลัสเตอร์ (Cluster) นั้น มุ่งเน้นให้ข้อมูลที่อยู่ในคลัสเตอร์เดียวกันมีความคล้ายคลึงกันมากที่สุด และในขณะเดียวกันข้อมูลที่อยู่ต่างคลัสเตอร์กันจะต้องมีความคล้ายคลึงกันน้อยที่สุดหรือแตกต่างกันมากที่สุด ซึ่งความคล้ายคลึงกันหรือต่างกันสามารถเปรียบเทียบได้จากความใกล้ชิดกันของข้อมูลนั้นๆ โดยใช้ระยะทางเป็นตัวชี้วัด

การจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีเคมีน (K-means) เป็นวิธีการจัดกลุ่มข้อมูลให้อยู่ในจำนวนกลุ่มที่ต้องการ k กลุ่ม โดยจำนวนกลุ่มต้องมีค่าน้อยกว่าจำนวนของข้อมูลทั้งหมด ขั้นตอนวิธีเคมีนเป็นที่รู้จักและนิยมนำมาพัฒนาในโปรแกรมประยุกต์ส่วนใหญ่หลากหลายสาขา เพราะเป็นขั้นตอนวิธีที่ง่ายในทางปฏิบัติ อีกทั้งยังได้ผลดีในการจัดกลุ่ม เหมาะสมกับการรวมกลุ่มที่มีข้อมูลจำนวนมาก ใช้เวลาในการคำนวณน้อย ในการจัดกลุ่มข้อมูลจะพิจารณาจากความเหมือนของวัตถุ ซึ่งใช้วิธีการหาค่าระยะห่างที่น้อยที่สุดระหว่างข้อมูลกับจุดศูนย์กลางของข้อมูลในแต่ละกลุ่ม เพื่อจัดข้อมูลให้อยู่ในกลุ่มที่กำหนด โดยใช้เงื่อนไขของค่าที่ใกล้ที่สุด (Minimum Distance) หรือมีระยะห่างของจุดกึ่งกลางกลุ่มน้อยที่สุด ตัวอย่างการแบ่งกลุ่มของขั้นตอนวิธีเคมีนแสดงดังรูปที่ 2.1



รูปที่ 2.1 ตัวอย่างการแบ่งกลุ่มของขั้นตอนวิธีเคมีน

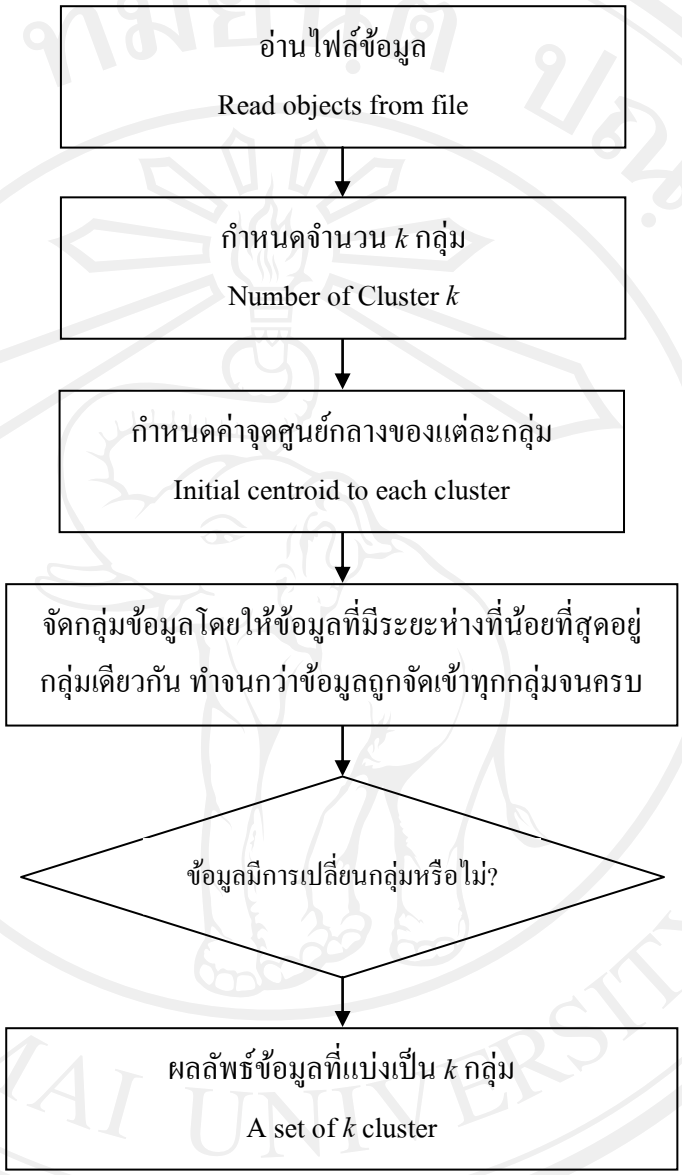
จากตัวอย่างในรูปที่ 2.1 ประกอบด้วยคลัสเตอร์จำนวน 3 กลุ่ม (k=3) โดยแต่ละกลุ่มจะมีจุดศูนย์กลาง (Centroid) 3 จุด จุดศูนย์กลางทั้งสามจุดนี้จะใช้สำหรับเปรียบเทียบกับข้อมูลที่ไม่รู้ว่าอยู่กลุ่มไหน จะใช้วิธีการหาค่าระยะห่างที่น้อยที่สุดระหว่างข้อมูลกับจุดศูนย์กลางของข้อมูลในแต่ละกลุ่มในการจัดข้อมูลให้อยู่ในกลุ่มใดๆ การหาค่าระยะห่างที่น้อยที่สุดนี้ใช้การวัดระยะแบบยูคลิด (Euclidean Distance) ดังแสดงในสมการที่ (2.2)

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \qquad (2.2)$$

จากสมการที่ (2.2) คือระยะห่างแบบยูคลิดระหว่างจุด p และ จุด q ขั้นตอนวิธีแบบเคมีนเป็นขั้นตอนที่มีการทำซ้ำไปเรื่อยๆในการเลือกจุดศูนย์กลางและการจัดเข้ากลุ่มให้กับข้อมูล จะจบการทำงานเมื่อค่าจุดศูนย์กลางของแต่ละกลุ่ม ไม่มีการเปลี่ยนแปลงและข้อมูลในแต่ละกลุ่มไม่มีการเปลี่ยนแปลง โดยมีกระบวนการทำงานดังนี้

1. เริ่มต้นอ่านไฟล์ข้อมูล
2. กำหนดจำนวนกลุ่มออกเป็น k กลุ่ม
3. กำหนดจุดศูนย์กลาง (Centroid) ของแต่ละกลุ่ม
4. นำข้อมูลทั้งหมดจัดเข้ากลุ่มที่มีจุดศูนย์กลาง (Centroid) อยู่ใกล้ข้อมูลนั้นมากที่สุด
5. คำนวณจุดศูนย์กลาง (Centroid) ใหม่สำหรับกลุ่มที่มีการเพิ่มหรือเสียข้อมูลไป
6. ทำซ้ำตั้งแต่ขั้นตอนที่ 4 จนกระทั่งข้อมูลในกลุ่มไม่มีการเปลี่ยนแปลง

รูปที่ 2.2 แสดงการทำงานในรูปแบบของ Flow Chart ดังต่อไปนี้



รูปที่ 2.2 ขั้นตอนวิธีเคมีน

2.2.1 ขั้นตอนการเลือกจุดศูนย์กลางเริ่มต้น [16]

D. Arthur และ S. Vassilvitskii [16] ได้เสนอวิธีการเลือกจุดศูนย์กลางเริ่มต้น โดยเลือกจุดศูนย์กลางจุดแรกจากการสุ่มข้อมูลทั้งหมด และเลือกจุดศูนย์กลางถัดไปที่ความน่าจะเป็นแปรผันตามระยะห่างของวัตถุกับจุดศูนย์กลางยกกำลังสอง $D(x)^2$ เมื่อ $x \epsilon X$ และ $D(x)$ คือ ระยะห่างระหว่างวัตถุกับจุดศูนย์กลางของวัตถุนั้น ซึ่งถูกเรียกว่า "D^2 Weighting" ดังแสดงในสมการที่ (2.3)

$$\frac{D(x)^2}{\sum_{x \epsilon X} D(x)^2} \qquad (2.3)$$

ในการเลือกจุดศูนย์กลางมีความสำคัญมากต่อการจัดกลุ่มข้อมูล เพราะการเลือกชุดของจุดศูนย์กลางเริ่มต้นที่ดี ทำให้เวลาในการประมวลผลของอัลกอริทึมลดลง เนื่องจากจำนวนรอบของการทำงานลดลง

2.2.2 ตัวอย่างการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีเคมีน

กำหนดให้มีตัวแปร 2 ตัวแปรคือ x และ y มีข้อมูลอยู่ 6 ชุดดังแสดงในตารางที่ 2.1 และต้องการจัดกลุ่มข้อมูลออกเป็น 2 กลุ่ม

ข้อมูล	x	y
1	1.00	1.50
2	1.00	4.50
3	2.00	1.50
4	2.00	3.50
5	3.00	2.50
6	5.00	3.00

ตารางที่ 2.1 ตัวอย่างข้อมูลในการจัดกลุ่มข้อมูลของขั้นตอนวิธีเคมีน

กำหนดให้ข้อมูลที่ 1 เป็นจุดศูนย์กลาง (Centroid) ของกลุ่มที่ 1 และให้ข้อมูลที่ 3 เป็นจุดศูนย์กลาง (Centroid) ของกลุ่มที่ 2 โดยใช้สัญลักษณ์ C_1 และ C_2 แทนกลุ่มที่ 1 และกลุ่มที่ 2 ดังนี้

$C_1 = (1.00, 1.50)$ $C_2 = (2.00, 1.50)$

คำนวณหาค่าระยะห่างที่น้อยที่สุดนี้ใช้การวัดระยะแบบยูคลิด (Euclidean distance) จากสมการ (2.4)

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} \qquad (2.4)$$

$$= \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

การคำนวณแสดงดังต่อไปนี้

$$C1,1\ \sqrt{(1.00-1.00)^2+(1.50-1.50)^2}=0.00$$

$$C2,1\ \sqrt{(2.00-1.00)^2+(1.50-1.50)^2}=1.00$$

$$C1,2\ \sqrt{(1.00-1.00)^2+(1.50-4.50)^2}=3.00$$

$$C2,2\ \sqrt{(2.00-1.00)^2+(1.50-4.50)^2}=3.16$$

$$C1,3\ \sqrt{(1.00-2.00)^2+(1.50-1.50)^2}=1.00$$

$$C2,3\ \sqrt{(2.00-2.00)^2+(1.50-1.50)^2}=0.00$$

$$C1,4\ \sqrt{(1.00-2.00)^2+(1.50-3.50)^2}=2.24$$

$$C2,4\ \sqrt{(2.00-2.00)^2+(1.50-3.50)^2}=2.00$$

$$C1,5\ \sqrt{(1.00-3.00)^2+(1.50-2.50)^2}=2.24$$

$$C2,5\ \sqrt{(2.00-3.00)^2+(1.50-2.50)^2}=1.41$$

$$C1,6\sqrt{(1.00-5.00)^2+(1.50-3.00)^2}=4.27$$

$$C2,6\ \sqrt{(2.00-5.00)^2+(1.50-3.00)^2}=3.35$$

จะได้ระยะแบบยูคลิด (Euclidean Distance) ดังนี้

$d(C_{1,1}) = 0.00$	$d(C_{1,2}) = 1.00$
$d(C_{1,2}) = 3.00$	$d(C_{2,2}) = 3.16$
$d(C_{1,3}) = 1.00$	$d(C_{2,3}) = 0.00$
$d(C_{1,4}) = 2.24$	$d(C_{2,4}) = 2.00$
$d(C_{1,5}) = 2.24$	$d(C_{2,5}) = 1.41$
$d(C_{1,6}) = 4.27$	$d(C_{2,6}) = 3.35$

และนำข้อมูลจัดเข้ากลุ่มที่มีจุดศูนย์กลาง (Centroid) อยู่ใกล้ข้อมูลนั้นมากที่สุด ดังนี้

กลุ่ม C_1 จะมีข้อมูลที่ 1 และ 2

กลุ่ม C_2 จะมีข้อมูลที่ 3, 4, 5 และ 6

ทำซ้ำรอบที่ 2 คำนวณ Centroid ของ C_1 และ C_2 ใหม่และทำการหาระยะแบบยูคลิด (Euclidean distance) เหมือนในตอนต้นและจัดเข้ากลุ่มใหม่จนกว่าข้อมูลในแต่ละกลุ่มจะไม่เปลี่ยนแปลง

	พิกัดของ Centroid	
กลุ่ม	ค่าเฉลี่ย x	ค่าเฉลี่ย y
C_1	1.0	3.0
C_2	3.0	2.63

ตารางที่ 2.2 การคำนวณจุดศูนย์กลาง (Centroid) รอบที่ 2

จะได้ระยะแบบยูคลิด (Euclidean Distance) ดังนี้

$d(C_{1,1}) = 1.50$ $d(C_{1,2}) = 2.29$

$d(C_{1,2}) = 1.50$ $d(C_{2,2}) = 2.74$

$d(C_{1,3}) = 1.80$ $d(C_{2,3}) = 1.51$

$d(C_{1,4}) = 1.19$ $d(C_{2,4}) = 1.33$

$d(C_{1,5}) = 2.06$ $d(C_{2,5}) = 0.13$

$d(C_{1,6}) = 4.00$ $d(C_{2,6}) = 2.03$

และนำข้อมูลจัดเข้ากลุ่มที่มีจุดศูนย์กลาง (Centroid) อยู่ใกล้ข้อมูลนั้นมากที่สุด ดังนี้

กลุ่ม C_1 จะมีข้อมูลที่ 1,2 และ 4

กลุ่ม C_2 จะมีข้อมูลที่ 3, 5 และ 6

ทำซ้ำรอบที่ 3

	พิกัดของ Centroid	
กลุ่ม	ค่าเฉลี่ย x	ค่าเฉลี่ย y
C_1	1.33	3.17
C_2	3.33	2.33

ตารางที่ 2.3 การคำนวณจุดศูนย์กลาง (Centroid) รอบที่ 3

จะได้ระยะแบบยูคลิด (Euclidean Distance) ดังนี้

$d(C_{1,1}) = 1.70$	$d(C_{1,2}) = 2.47$
$d(C_{1,2}) = 1.37$	$d(C_{2,2}) = 3.18$
$d(C_{1,3}) = 1.80$	$d(C_{2,3}) = 1.57$
$d(C_{1,4}) = 0.75$	$d(C_{2,4}) = 1.77$
$d(C_{1,5}) = 1.80$	$d(C_{2,5}) = 0.37$
$d(C_{1,6}) = 3.67$	$d(C_{2,6}) = 1.80$

และนำข้อมูลจัดเข้ากลุ่มที่มีจุดศูนย์กลาง (Centroid) อยู่ใกล้ข้อมูลนั้นมากที่สุด ดังนี้

กลุ่ม C_1 จะมีข้อมูลที่ 1,2 และ 4

กลุ่ม C_2 จะมีข้อมูลที่ 3, 5 และ 6

จะเห็นว่ารอบที่ 3 ไม่มีการเปลี่ยนแปลงของข้อมูลในกลุ่ม ดังนั้นจึงหยุดการทำงานของขั้นตอนวิธีเคมีนได้ว่า

กลุ่มที่ 1 (C_1) จะมีข้อมูลที่ 1,2 และ 4

กลุ่มที่ 2 (C_2) จะมีข้อมูลที่ 3, 5 และ 6

2.3 การตัดสินใจบอตเน็ต (Botnet Decision) [8]

งานวิจัยที่ใช้ในการตัดสินใจบอตเน็ต (Botnet Decision) ใช้การพิจารณาจากการหาค่าส่วนเบี่ยงเบนมาตรฐาน (Standard deviation metric) ในการลาเบลข้อมูล ถ้าข้อมูลเป็นความปกติหรือความผิดปกติ ตัวข้อมูลควรจะได้รับลาเบลข้อมูลที่ถูกต้อง และเป็นตัวแทนประเภทพฤติกรรมทั้งหมด โดยส่วนเบี่ยงเบนมาตรฐาน[8] หาได้จากสมการที่ (2.5) ดังนี้

$$\sigma = \frac{\sum_{i=1}^{256} \sigma_i}{256} \qquad (2.5)$$

เมื่อ σ_i คือส่วนเบี่ยงเบนมาตรฐานของตัวอักษรแอสกีในชุดข้อมูลนั้นๆ

วิธีการคำนวณหาค่าร้อยละของข้อมูล Botnet (BPR) ในสมการที่ (2.6) และค่าร้อยละของข้อมูล Normal (NPR) ในสมการที่ (2.7)

$$BPR = \frac{number\ of\ botnet\ instances\ in\ the\ cluster}{total\ number\ of\ botnet\ instances} \quad (2.6)$$

$$NPR = \frac{number\ of\ normal\ instances\ in\ the\ cluster}{total\ number\ of\ normal\ instances} \quad (2.7)$$