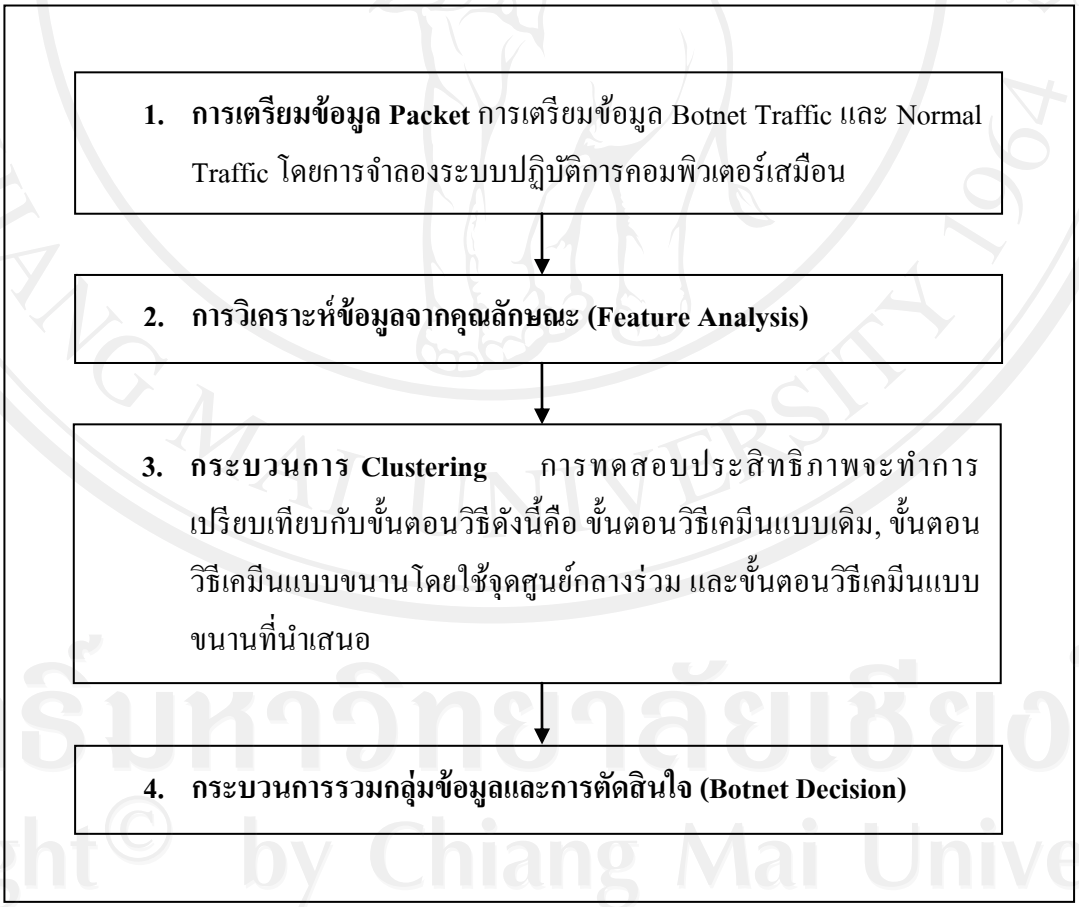
บทที่ 3
แนวคิดในการแก้ไขปัญหาและขั้นตอนการพัฒนา

ในบทนี้จะกล่าวถึงแนวคิดที่ใช้ในการพัฒนาขั้นตอนวิธี และการพัฒนาขั้นตอนวิธีการจำแนกบอตเน็ตโดยใช้ขั้นตอนวิธีเคมีนแบบขนานที่ปรับปรุงใหม่ สำหรับประมวลผลในตัวประมวลผลหลายตัว โดยมีรูปแบบการทำงานเป็นการแบ่งงานประมวลผลไปพร้อมๆกัน รวมถึงการนำขั้นตอนการเลือกจุดศูนย์กลางเริ่มต้นแบบใหม่แบบใหม่ มาประยุกต์ใช้ร่วมกันกับขั้นตอนวิธีเคมีนแบบขนาน เพื่อเพิ่มประสิทธิภาพทางด้านเวลาในการประมวลผลและการจัดกลุ่มที่ได้ผลดี โดยได้นำเอาทฤษฎีต่าง ๆ ในบทที่ 2 มาประยุกต์ใช้ ซึ่งลำดับขั้นตอนการพัฒนาแสดงดังรูปที่ 3.1



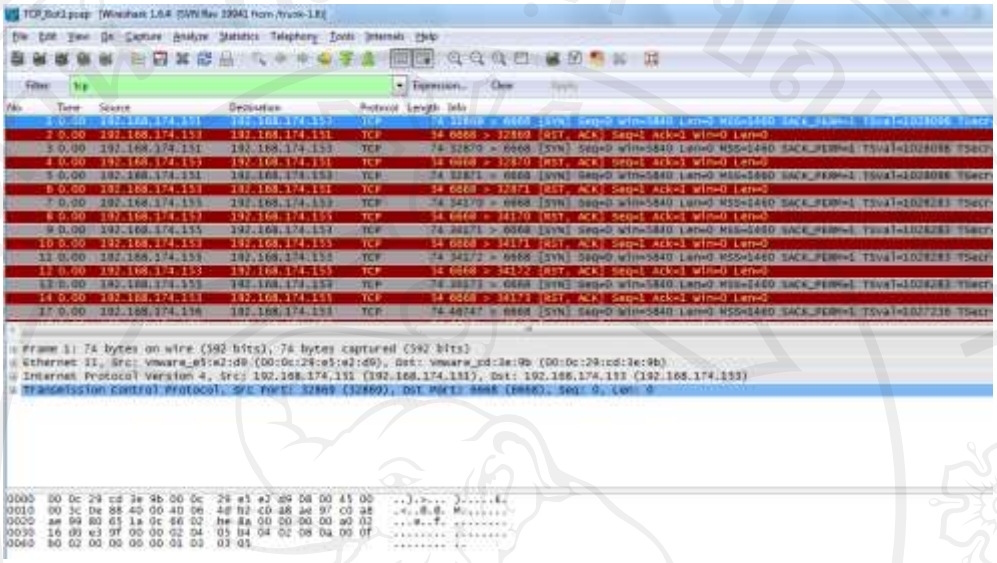
รูปที่ 3.1 กระบวนการจำแนกบอตเน็ตโดยใช้ขั้นตอนวิธีเคมีนแบบขนาน

3.1 การเตรียมข้อมูลแพ็กเก็ต (Packet)

ในการเตรียมข้อมูลได้นำ Source Code ของบอตเน็ตชื่อ q8bot [23] ซึ่งเป็นบอตเน็ตประเภท IRC bot ที่ทำงานบนระบบปฏิบัติการลีนุกซ์ (Linux) มีรูปแบบการโจมตีจะเป็นแบบ DDos (Distributed Denial of Service) สามารถ Download ได้จากเว็บไซต์ [23] http://www.security dot.net/exploits/bots/ จากไฟล์ที่ชื่อว่า q8bot.gz มาใช้เป็นข้อมูลที่เป็น Botnet Traffic ในการเก็บข้อมูลในงานวิจัย [20] ทำการ Botnet Traffic Generation โดยเริ่มต้นจากการจำลองระบบปฏิบัติการคอมพิวเตอร์เสมือน หรือการสร้าง Virtual Machine (VM) โดยใช้ VMWare [25] ซึ่งสามารถสร้างเครื่องคอมพิวเตอร์ได้หลายๆเครื่องภายในเครื่องคอมพิวเตอร์ที่ทำการทดลอง และทำการติดตั้งระบบปฏิบัติการ Ubuntu 10.04 ทั้งหมด 4 เวอร์ชวลแมชชีน โดยให้เวอร์ชวลแมชชีน 1 เครื่องทำหน้าที่เป็นตัวควบคุมหลักหรือที่เรียกว่า Bot Master เป็นตัวสั่งการไปยังอีก 3 เวอร์ชวลแมชชีน ซึ่งจะทำหน้าเป็นสมาชิกในเครือข่ายของบอตเน็ต เมื่อบอตเน็ตทำงานจะมีการเชื่อมต่อกันระหว่างตัวควบคุมหลักและสมาชิกในเครือข่ายของบอตเน็ต โดยตัวควบคุมหลักจะสั่งการให้ส่งข้อความ NICK IRC เพื่อทำการรับค่า IP เมื่อสมาชิกในเครือข่ายบอตเน็ตตอบรับแล้วจะรอคำสั่งจากตัวควบคุมหลัก เมื่อคำสั่งมาถึงสมาชิกในเครือข่ายจะทำหน้าที่ตามคำสั่งที่ได้รับมอบหมาย ถ้าไม่มีการได้รับคำสั่งเป็นเวลา 20 นาที สมาชิกในเครือข่ายจะทำการเชื่อมต่อไปยังตัวควบคุมหลัก เพื่อเริ่มปฏิบัติการจากที่กล่าวข้างต้นอีกครั้ง สำหรับข้อมูลที่เป็น Normal Traffic นั้นงานวิจัยนี้เก็บข้อมูลภายในห้องปฏิบัติการ ภาควิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเชียงใหม่

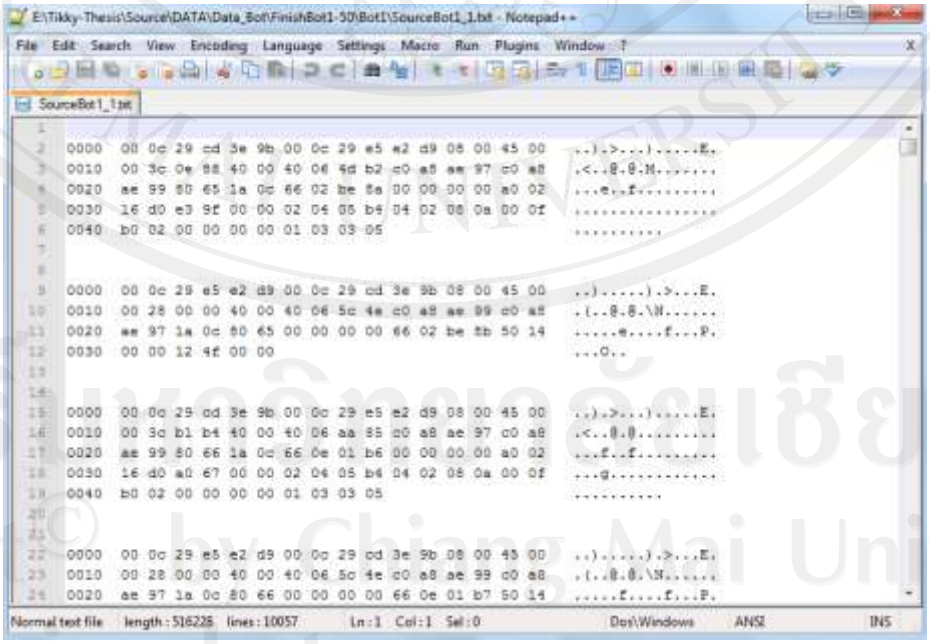
การดักจับข้อมูลแพ็กเก็ตของทั้ง Botnet Traffic และ Normal Traffic ใช้โปรแกรม Wireshark [24] หรือชื่อเดิมคือ Ethereal เป็นโปรแกรมประเภท Packet Sniffer ชนิดหนึ่ง ซึ่งในการอ่าน Packet นั้นประกอบด้วยส่วนของ Packet Capture และ Packet Analyzer โดยทำหน้าที่วิเคราะห์รูปแบบพฤติกรรมใน Packet ของระบบเครือข่าย (Network) โดยโปรแกรม Wireshark นั้นสามารถทำงานได้ทั้งบนระบบปฏิบัติการลีนุกซ์ (Linux), วินโดวส์ (Windows) และโอเอสเท็น (OSX) สามารถทำการวิเคราะห์ข้อมูลบนเครือข่ายได้หลากหลายรูปแบบ และโปรแกรม Wireshark ยังเป็นซอฟแวร์แบบ Open Source หรือ Freeware ซึ่งให้ใช้งานโดยไม่ต้องเสียค่าใช้จ่ายใดๆ

ขั้นตอนการดักจับข้อมูลแพ็กเก็ตของ Botnet Traffic และ Normal Traffic โดยใช้โปรแกรม Wireshark ในการอ่านข้อมูล Packet แสดงดังรูปที่ 3.2



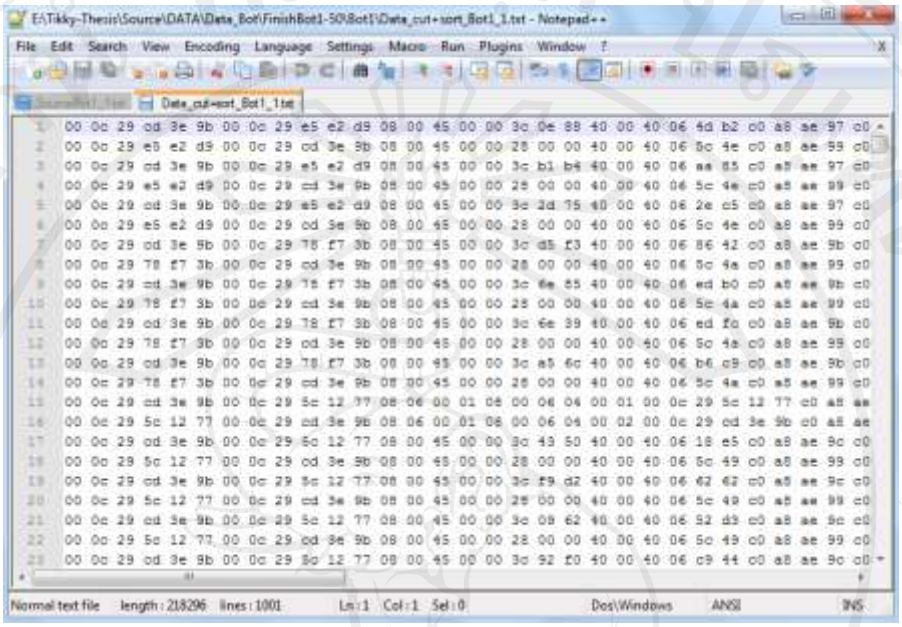
รูปที่ 3.2 การดักจับข้อมูลแพ็กเก็ตของ Botnet Traffic และ Normal Traffic ของโปรแกรม Wireshark

ขั้นตอนการนำข้อมูล Packet Analyzer ที่ต้องการนำมาวิเคราะห์ของข้อมูล Botnet Traffic และ Normal Traffic จากโปรแกรม Wireshark แสดงดังรูปที่ 3.3



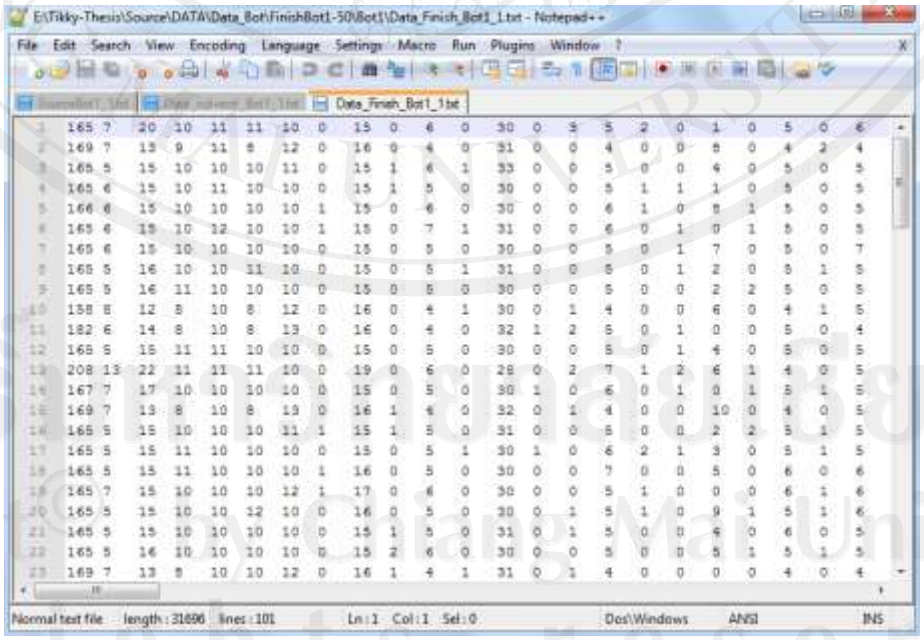
รูปที่ 3.3 ข้อมูล Packet Analyzer ของ Botnet Traffic และ Normal Traffic

ขั้นตอนการจัดเรียงข้อมูล Packet Analyzer ของ Botnet Traffic และ Normal Traffic โดยจัดเรียงตามลำดับของอักขระแอสกี (ASCII) ทั้งหมด 256 ตัว (0-255) แสดงดังรูปที่ 3.4



รูปที่ 3.4 การจัดเรียงข้อมูล Packet Analyzer ของ Botnet Traffic และ Normal Traffic

ขั้นตอนการนับข้อมูล Packet Analyzer ของ Botnet Traffic และ Normal Traffic โดยนับการเกิดซ้ำกันของอักขระแอสกี (ASCII) ทั้งหมด 256 ตัว (0-255) ใน Packet แสดงดังรูปที่ 3.5

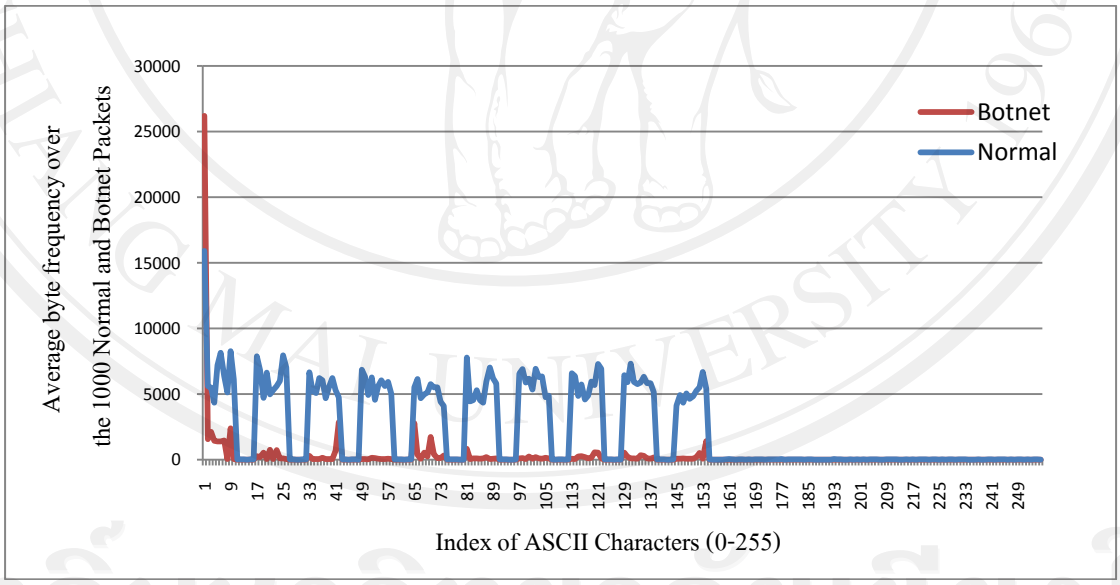


รูปที่ 3.5 การนับข้อมูล Packet Analyzer ของ Botnet Traffic และ Normal Traffic

3.2 การวิเคราะห์ข้อมูลจากคุณลักษณะ (Feature Analysis)

เมื่อเก็บข้อมูลของ Botnet Traffic และNormal Traffic มาในรูปแบบของแพ็กเก็ตได้แล้ว การวิเคราะห์ข้อมูลทั้งหมดนั้นจะพิจารณาจากคุณลักษณะ(Feature Analysis) โดย พิจารณาจาก 256 ASCII Characters ใน Packet Payload เพื่อดูค่าความถี่เฉลี่ยในแต่ละไบต์ (Byte) ของตัวอักษรแอสกี (ASCII) ของทราฟฟิกทั้งสองแบบมาใช้เป็นข้อมูลสำหรับขั้นตอนของการจำแนกบอตเน็ตต่อไป โดยใช้การหา Temporal-frequent Metric เหตุผลที่ใช้ Temporal-frequent Metric นั้น ได้มีงานวิจัย [17-19] ซึ่งทำการศึกษาพฤติกรรมของบอตเน็ตทราฟฟิก และทำการพร็อตกราฟเพื่อดูความถี่ของทราฟฟิกที่เป็นบอตเน็ตและทราฟฟิกที่ปกติ ในผลการทดลองพบว่า ค่าความถี่เฉลี่ยในแต่ละไบต์ (Byte) ของตัวอักษรแอสกี (ASCII) ในทราฟฟิกที่เป็นบอตเน็ตและทราฟฟิกที่ปกติมีความแตกต่างกันอย่างเห็นได้ชัดเจน ทำให้สามารถสรุปได้ว่า ค่าความถี่เฉลี่ยในแต่ละไบต์ (Byte) ของตัวอักษรแอสกี (ASCII) สามารถนำมาวิเคราะห์แยกทราฟฟิกที่เป็นบอตเน็ตจากทราฟฟิกที่ปกติได้

ในงานวิจัยได้ทำการพร็อตกราฟความถี่เฉลี่ยในแต่ละไบต์ (Byte) ของตัวอักษรแอสกี (ASCII) ข้อมูล 1,000 แพ็กเก็ตของ Botnet Traffic และ Normal Traffic แสดงดังรูปที่ 3.6



รูปที่ 3.6 กราฟแสดงค่าความถี่เฉลี่ยในแต่ละไบต์ (Byte) ของตัวอักษรแอสกี (ASCII) ข้อมูล 1,000 แพ็กเก็ตของ Botnet Traffic และ Normal Traffic

3.3 กระบวนการแบ่งกลุ่ม (Clustering)

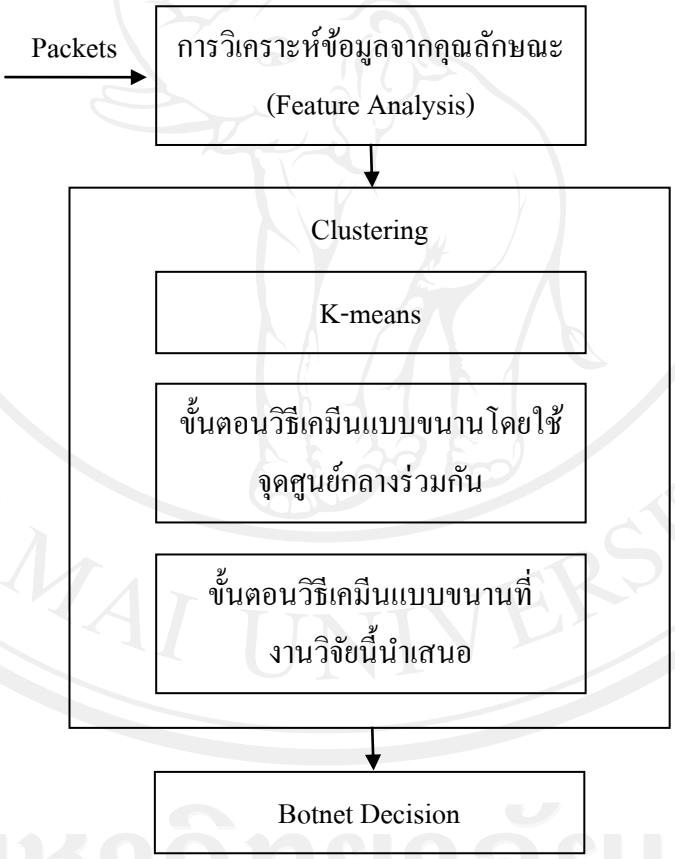
ในงานวิจัยนี้ได้เสนอวิธีการจำแนกบอตเน็ตโดยใช้ขั้นตอนวิธีเคมีนแบบขนานที่ปรับปรุงใหม่ สำหรับประมวลผลในตัวประมวลผลหลายตัว และรูปแบบการทำงานจะเป็นการแบ่งงานกันประมวลผลไปพร้อมๆกัน การทดสอบประสิทธิภาพจะทำการเปรียบเทียบกับขั้นตอนวิธีดังนี้

3.3.1 ขั้นตอนวิธีเคมีนแบบเดิม

3.3.2 ขั้นตอนวิธีเคมีนแบบขนาน โดยใช้จุดศูนย์กลางร่วมกัน [16]

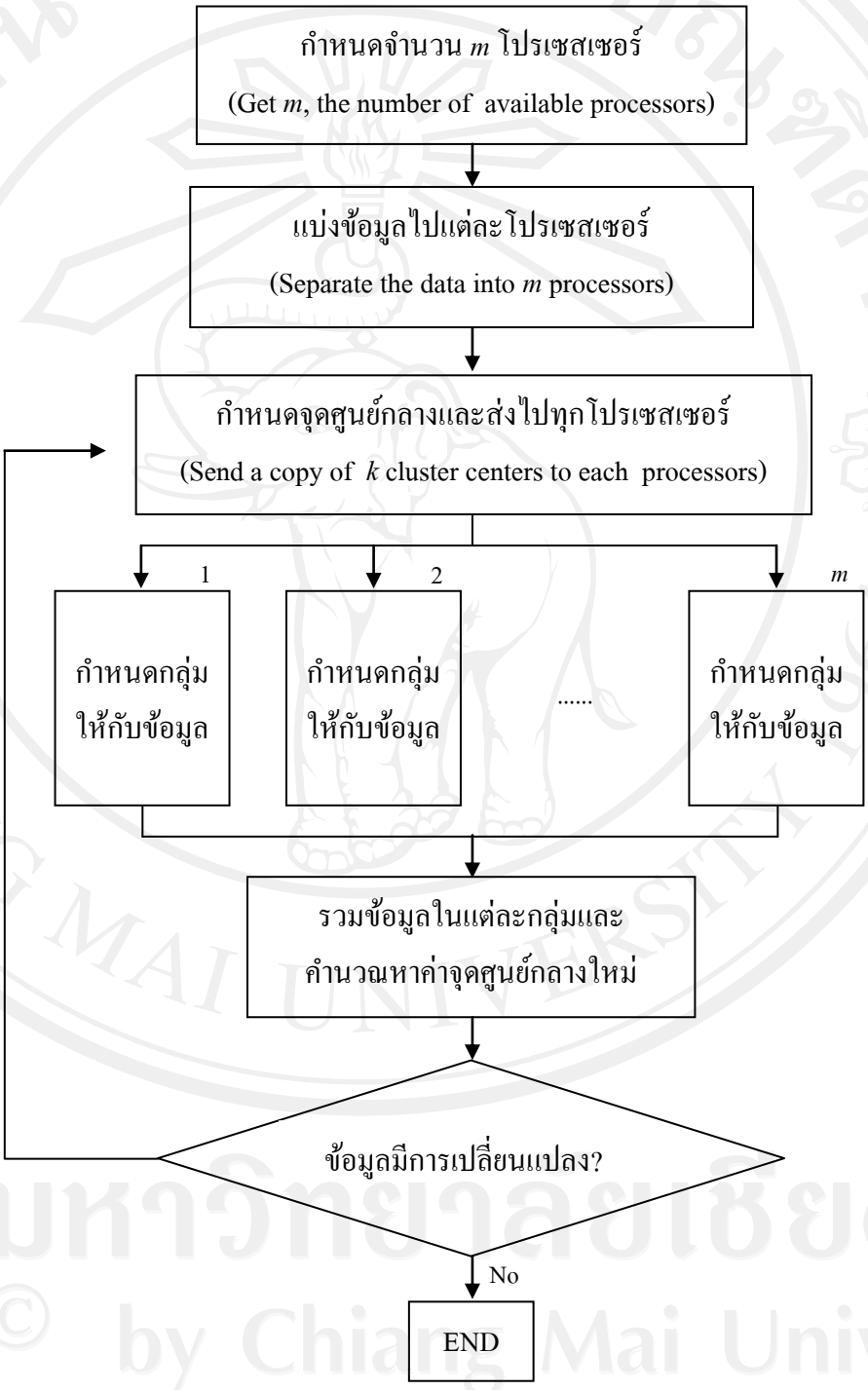
3.3.3 ขั้นตอนวิธีเคมีนแบบขนานที่งานวิจัยนี้นำเสนอ

ซึ่งลำดับกระบวนการจำแนกบอตเน็ตมีขั้นตอนการพัฒนาแสดงดังรูปที่ 3.7



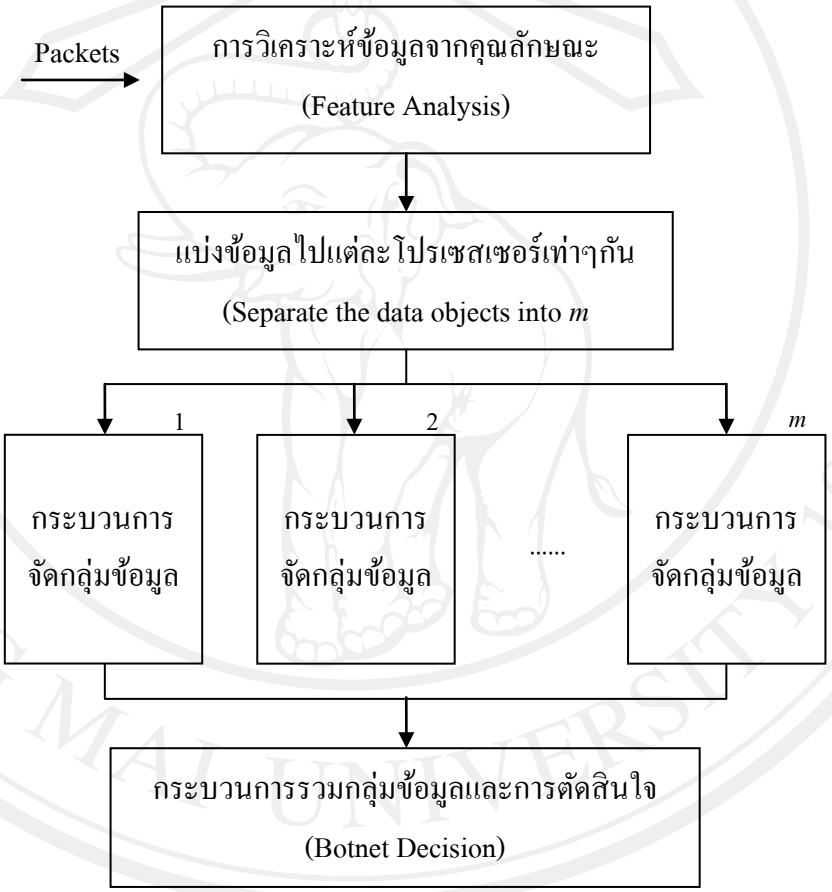
รูปที่ 3.7 กระบวนการจำแนกบอตเน็ต

วิธีการจำแนกบอตเน็ตโดยใช้ขั้นตอนวิธีเคมีนแบบขนาน โดยใช้จุดศูนย์กลางร่วมกัน [16] มีขั้นตอนการทำงานดังแสดงในรูปที่ 3.8



รูปที่ 3.8 ขั้นตอนวิธีเคมีนแบบขนานโดยใช้จุดศูนย์กลางร่วมกัน

วิธีการจำแนกบอตเน็ตโดยใช้ขั้นตอนวิธีเคมีนแบบขนานที่ปรับปรุงใหม่จะแตกต่างกับขั้นตอนวิธีเคมีนแบบขนานในรูปที่ 3.8 ตรงที่การคำนวณจุดศูนย์กลาง โดยขั้นตอนการทำงานทั้งหมดแบ่งออกเป็น 4 ขั้นตอนหลักๆได้แก่ การเตรียมและเก็บข้อมูลในรูปแพ็กเก็ต (Packets) การวิเคราะห์ข้อมูลจากคุณลักษณะ(Feature Analysis) การแบ่งกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีเคมีนแบบขนานปรับปรุงใหม่ (Parallel K-means Algorithm) และกระบวนการรวมกลุ่มข้อมูลและการตัดสินใจ (Botnet Decision) ซึ่งกระบวนการทำงานทั้งหมดแสดงในรูปที่ 3.9



รูปที่ 3.9 กระบวนการจำแนกบอตเน็ตโดยใช้ขั้นตอนวิธีเคมีนแบบขนานในรูปแบบที่นำเสนอ

ในงานวิจัยได้มีการปรับปรุงการจำแนกข้อมูลโดยใช้ขั้นตอนวิธีเคมีนแบบขนานขึ้นมาใหม่สำหรับประมวลผลในตัวประมวลผลหลายตัว และรูปแบบการทำงานจะเป็นการแบ่งงานกันประมวลผลไปพร้อมๆกัน เพื่อเป็นการเพิ่มประสิทธิภาพทางด้านเวลาในการประมวลผลและความถูกต้องในการแบ่งกลุ่ม เนื่องจากปริมาณของข้อมูลที่มีขนาดใหญ่ ผู้วิจัยได้ทำการปรับปรุงขั้นตอนการเลือกจุดศูนย์กลางเริ่มต้นของขั้นตอนวิธีการเคมีนขึ้นใหม่ โดยในงานวิจัย [21] แสดงการเลือกจุดศูนย์กลางจุดแรกจากการสุ่มข้อมูลทั้งหมด และเลือกจุดศูนย์กลางถัดไปที่ความน่าจะเป็นแปรผันตามระยะห่างของวัตถุกับจุดศูนย์กลางยกกำลังสอง $D(x)^2$ เมื่อ $x \in X$ และ $D(x)$ คือ ระยะห่างระหว่างวัตถุกับจุดศูนย์กลางของวัตถุนั้น ในการเลือกจุดศูนย์กลางมีความสำคัญมากต่อการจัดกลุ่มข้อมูล เพราะการเลือกชุดของจุดศูนย์กลางเริ่มต้นที่ดี ทำให้เวลาในการประมวลของอัลกอริทึมลดลง เนื่องจากจำนวนรอบของการทำงานลดลง ขั้นตอนการเลือกจุดศูนย์กลางเริ่มของขั้นตอนวิธีเคมีนแบบขนานและกระบวนการทั้งหมดที่งานวิจัยนี้ใช้คือ

1. เริ่มต้นอ่านไฟล์ข้อมูล และทำการแบ่งข้อมูลไปแต่ละโปรเซสเซอร์เท่าๆกัน
2. แต่ละโปรเซสเซอร์ทำการหาจุดศูนย์กลางจุดแรก โดยเลือกจากการสุ่มข้อมูลทั้งหมด
3. เลือกจุดศูนย์กลางถัดไปที่ความน่าจะเป็นแปรผันตามระยะห่างของวัตถุกับจุดศูนย์กลางยกกำลังสอง
4. ทำงานซ้ำขั้นตอนที่ 3 จนกว่าจุดศูนย์กลางจะครบ k จุดศูนย์กลาง
5. คำนวณหาค่าระยะห่างของข้อมูลที่พิจารณากับจุดศูนย์กลางของแต่ละคลัสเตอร์
6. จัดกลุ่มให้กับข้อมูล
7. ทำการปรับปรุงจุดศูนย์กลาง
8. กลับไปขั้นตอนที่ 5 จนกว่าไม่มีการเปลี่ยนแปลงของกลุ่มข้อมูล

3.4 กระบวนการรวมกลุ่มข้อมูลและการตัดสินใจ (Botnet Decision)

ในส่วนสุดท้ายงานวิจัยที่ใช้ในการตัดสินใจบอตเน็ต (Botnet Decision) ใช้การพิจารณาจากการหาค่าส่วนเบี่ยงเบนมาตรฐาน (Standard deviation metric) ในการลาเบลข้อมูล ถ้าข้อมูลเป็นความปกติหรือความผิดปกติ ตัวข้อมูลควรจะได้รับลาเบลข้อมูลที่ถูกต้อง และเป็นตัวแทนประเภทพฤติกรรมทั้งหมด โดยส่วนเบี่ยงเบนมาตรฐาน [8] หาได้จากสมการที่ (3.1) ดังนี้

$$\sigma = \frac{\sum_{i=1}^{256} \sigma_i}{256} \qquad (3.1)$$

เมื่อ σ_i คือส่วนเบี่ยงเบนมาตรฐานของตัวอักษรแอสกีในชุดข้อมูลนั้นๆ

วิธีการคำนวณหาค่าร้อยละของข้อมูล Botnet (BPR) ในสมการที่ (3.2) และค่าร้อยละของข้อมูล Normal (NPR) ในสมการที่ (3.3)

$$\text{BPR} = \frac{\text{number of botnet instances in the cluster}}{\text{total number of botnet instances}} \quad (3.2)$$

$$\text{NPR} = \frac{\text{number of normal instances in the cluster}}{\text{total number of normal instances}} \quad (3.3)$$

กระบวนการรวมกลุ่มข้อมูลและการตัดสินใจ (Botnet Decision) ใช้การพิจารณาจากการหาค่าส่วนเบี่ยงเบนมาตรฐาน (Standard deviation metric) ของข้อมูลทุกกลุ่มแบ่งเรียบร้อยแล้ว จากนั้นจะทำการรวมกลุ่มของข้อมูลทุกโปรเซสเซอร์ที่ได้รับการลาเบลว่าเป็นกลุ่มเดียวกันไว้ด้วยกัน