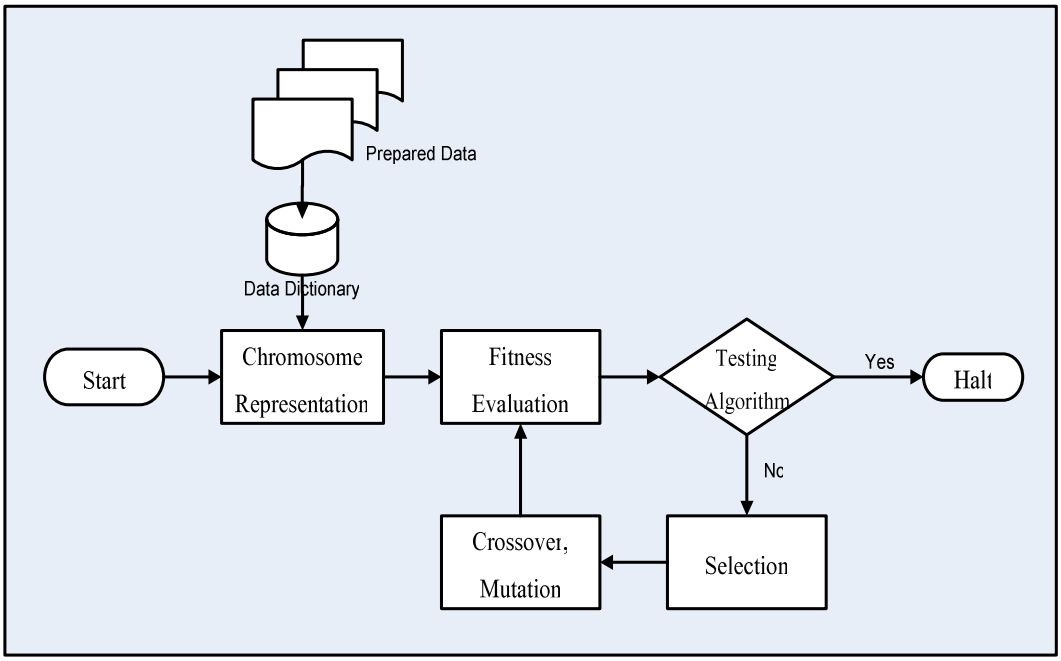
บทที่ 4
ตัวกรองอีเมล์ขยะด้วยทฤษฎีเจเนติกอัลกอริทึม

ตัวกรองอีเมล์ขยะโดยใช้เจเนติกอัลกอริทึม มีวัตถุประสงค์เพื่อพัฒนาระบบกำจัดอีเมล์ขยะให้มีประสิทธิภาพมากยิ่งขึ้น โดยใช้กระบวนการเจเนติกอัลกอริทึมซึ่งเป็นกลไกที่เลียนแบบวิวัฒนาการของสิ่งมีชีวิตในธรรมชาติ ตัวกรองอีเมล์ขยะที่สร้างขึ้นนี้จะใช้ตัวดำเนินการทางเจเนติก เช่น การคัดเลือก การครอสโอเวอร์ และการมิวเตชัน เพื่อสร้างรูปแบบของอีเมล์ที่หลากหลายขึ้นจากอีเมล์ที่มีอยู่เดิม ซึ่งจะช่วยให้เจเนติกอัลกอริทึมสามารถเรียนรู้ได้ว่ารูปแบบของอีเมล์แบบใดเป็นอีเมล์ขยะหรืออีเมล์ดี ในวิทยานิพนธ์นี้ เริ่มต้นจากการออกแบบระบบตัวกรองอีเมล์ขยะ การเตรียมอีเมล์ดีและอีเมล์ขยะเพื่อให้อัลกอริทึมเรียนรู้ ซึ่งประกอบไปด้วยขั้นตอนย่อยคือการตัดคำ การจัดกลุ่มคำและการเตรียมฐานข้อมูลของคำ เมื่อเตรียมข้อมูลเรียบร้อยแล้วจะเข้าสู่กระบวนการเจเนติกอัลกอริทึมเพื่อที่จะเรียนรู้รูปแบบของอีเมล์และดำเนินการจำแนกอีเมล์ต่อไป

ขั้นตอนการสร้างตัวกรองอีเมล์ขยะมี 3 ขั้นตอนหลัก คือ การเตรียมข้อมูล (Data Preparation) การสร้างฐานข้อมูลคำ (Creating Data Dictionary) และ การประยุกต์ใช้เจเนติกอัลกอริทึม (Adopting Genetic Algorithm) ซึ่งแสดงดังรูป 4.1



รูปที่ 4.1 แสดงบล็อกไดอะแกรมของระบบ

4.1 การเตรียมข้อมูล (Data Preparation)

1. อีเมล์ที่นำมาใช้ประกอบไปด้วยอีเมล์ขยะจำนวน 900 ฉบับจาก [3], [4] และอีก 100 ฉบับผู้เขียนได้รวบรวมเองจาก Hotmail และ Yahoo Mail และอีเมล์ดีจำนวน 1,000 ฉบับจาก [4]
2. แบ่งอีเมล์ออกเป็นชุดข้อมูลสำหรับเรียนรู้ (Training Set) จำนวน 1,600 ฉบับ และชุดข้อมูลสำหรับทดสอบ (Testing Set) จำนวน 400 ฉบับ
3. ตัดคำและนับความถี่ของคำ อาศัยหลักการใช้ช่องว่างระหว่างคำเป็นตัวแบ่งแยกคำออกจากกัน ในวิทยานิพนธ์นี้ได้ใช้เครื่องมือช่วยในการตัดคำและนับความถี่ของคำจาก [6] เมื่อตัดคำแล้ว จะคัดคำที่เป็น URL, คำที่อยู่ในส่วนหัวเรื่อง (Header) ของอีเมล์, คำที่ไม่มีความหมาย และตัวเลข ทิ้งไป โดยคำอื่นที่เหลือทั้งหมดต่อไปนี้จะถูกเรียกว่าเป็นคำสำคัญ (Keywords)

4.2 การสร้างฐานข้อมูลของคำ (Creating Data dictionary)

ฐานข้อมูลคำประกอบไปด้วย

1. คำสำคัญ (Keywords)และกลุ่มของคำสำคัญ
   เนื่องจากเราต้องการสร้างโครโมโซมโดยกำหนดให้แต่ละยีนของโครโมโซมเก็บคำสำคัญในอีเมล์ซึ่งแยกเป็นกลุ่มๆ เอาไว้ ดังนั้นจึงต้องมีการจัดกลุ่มให้แก่คำสำคัญทั้งหมด โดยการแบ่งกลุ่มมีหลักการคือ จะรวบรวมคำสำคัญที่มีความหมายใกล้เคียงกัน หรือมาจากรากศัพท์เดียวกัน ให้อยู่ตระกูลเดียวกัน[7] โดยเกณฑ์ในการแบ่งกลุ่มนั้น แบ่งตามการสำรวจกลุ่มของอีเมล์โดยไมโครซอฟต์ [2] และจากการระบุกลุ่มของอีเมล์จากคลังอีเมล์ [4] รวมกับการพิจารณาเนื้อหาในอีเมล์เพิ่มเติม ทำให้สามารถแบ่งคำสำคัญในอีเมล์ออกเป็น 8 กลุ่มหลัก และตัวอย่างของคำสำคัญในกลุ่ม แสดงดังตารางที่ 4.1
2. ความน่าจะเป็นของคำสำคัญ
   การหาความน่าจะเป็นของคำสำคัญทำเพื่อเตรียมความน่าจะเป็นเพื่อนำไปใช้ในการสร้างรูปแบบของโครโมโซม โดยจะอธิบายรายละเอียดของการคำนวณในหัวข้อถัดไป

ตารางที่ 4.1 แสดงตัวอย่างของคำสำคัญในแต่ละกลุ่ม

กลุ่ม	ชื่อกลุ่ม	ตัวอย่างของคำสำคัญในกลุ่ม
G1	**Adult** - กลุ่มคำสำคัญที่เกี่ยวกับสื่อลามก อนาจาร	sex, adult, viagra, hardcore, webcam, teen, girl, nude, gay, xxx, erection, etc.
G2	**Business and Financial** - กลุ่มคำสำคัญที่เกี่ยวกับธุรกิจ และการเงิน	enterprise, share, holder, investor, strategy, mortgage, obligate, fund, refund, loan, etc.
G3	**Commercial** - กลุ่มของคำสำคัญที่เกี่ยวกับการค้าขาย และข้อเสนอชวนเชื่อ	free, special, retail, resell, inexpensive, cartier, louis, buy, etc.
G4	**Medicine, Diet and Beauty** - กลุ่มของคำสำคัญที่เกี่ยวกับการขายยา และการบริการเพื่อรูปร่างและความสวยงาม	diet, fat, herb, weight, lose, age, medicine, nature, health, prescription, etc.
G5	**Traveling and Gambling** - กลุ่มของคำสำคัญที่เกี่ยวกับการท่องเที่ยว การพักผ่อนหย่อนใจ การพนัน และการเสี่ยงดวง	hotel, reserve, travel, trip, holiday, win, bonus, casino, extra, gambling, etc.
G6	**Internet and Home-Based Business** - กลุ่มของคำสำคัญที่เกี่ยวกับธุรกิจทางอินเตอร์เน็ต และการทำธุรกิจผ่านอินเตอร์เน็ตที่บ้าน	internet, home, based, email, earn, subscriber, home based etc.
G7	**Political, Social and Religion** - กลุ่มของคำสำคัญที่เกี่ยวกับการเมือง สังคม ศาสนา และความเชื่อ	policy, political, challenge, campaign, public, church, etc.
G8	**Common** - กลุ่มของคำสำคัญที่มีการใช้ร่วมกันอยู่ในหลายกลุ่ม	now, tell, today, year, monday, texas etc.

4.3 การประยุกต์ใช้เจเนติกอัลกอริทึม (Adopting Genetic Algorithm)

4.3.1 การกำหนดรูปแบบของโครโมโซม (Representation of a Chromosome)

การสร้างโครโมโซมแม่แบบคือการกำหนดรูปแบบของโครโมโซมจากการเลียนแบบรูปแบบของทั้งอีเมล์ขยะ และอีเมล์ดี โดยตัวดำเนินการทางเจเนติกจะทำการสร้างรูปแบบใหม่ๆ ขึ้นเพื่อให้อัลกอริทึมได้เรียนรู้ว่ารูปแบบใด (Template) มีความคล้ายคลึงกับอีเมล์แบบใดรูปแบบของโครโมโซมหรือโครโมโซมแม่แบบเกิดจากการประกอบกันของค่าความน่าจะเป็นเฉลี่ยในแต่ละกลุ่มซึ่งอยู่ในรูปของเลขฐานสองดังรูป 4.2

Gene1	Gene2	Gene3	Gene4	Gene5	Gene6	Gene7	Gene8
$P_{avg}(G1)$ = 0001100110	$P_{avg}(G2)$ = 0000000000	$P_{avg}(G3)$ = 0000110101	$P_{avg}(G4)$ = 0000000000	$P_{avg}(G5)$ = 0000000000	$P_{avg}(G6)$ = 0000000000	$P_{avg}(G7)$ = 0000000000	$P_{avg}(G8)$ = 0000000000

รูปที่ 4.2 แสดงรูปแบบของโครโมโซม

ค่าความน่าจะเป็นเฉลี่ยคำนวณได้จากความน่าจะเป็นของคำสำคัญใดๆในกลุ่มหารด้วยจำนวนคำในกลุ่มดังสมการ (4.1)

$$P_{avg}(G) = \sum_{i=1}^{\text{total words}} \frac{\text{จำนวนครั้งที่พบคำสำคัญ i x ความน่าจะเป็นของคำสำคัญ i}}{\text{จำนวนคำสำคัญทั้งหมดในกลุ่ม (total words)}} \quad (4.1)$$

โดยที่ $P(word_i)$ คำนวณได้จากสมการ (4.2)

$$P(word_i) = \frac{\text{จำนวนครั้งที่พบคำสำคัญ}}{\text{จำนวนคำสำคัญทั้งหมด}} \quad (4.2)$$

ตัวอย่างเช่น คำสำคัญว่า “sex” มีจำนวนครั้งที่พบ 457 ครั้ง จากความถี่ที่ปรากฏของคำทุกคำรวมกัน 1,273 ครั้ง จะได้ว่า P(sex) = 457/1,273 = 0.359 เป็นต้น

และถ้ากำหนดให้อีเมล์ 1 ฉบับ เมื่อผ่านขั้นตอนการตัดคำแล้ว พบว่ามีคำสำคัญว่า “sex”, “adult” ซึ่งเป็นคำสำคัญในกลุ่ม G1 และคำว่า “free”, “special” ซึ่งคำสำคัญในกลุ่ม G3 โดยคำสำคัญ “sex”, “adult”, “free” และ “special” มีค่าความน่าจะเป็น 0.395, 0.005, 0.230 และ 0.090 ตามลำดับ การสร้างโครโมโซมสำหรับอีเมล์ฉบับนี้ ทำได้โดยการหาความน่าจะเป็นของคำสำคัญแต่ละคำแล้วคำนวณความน่าจะเป็นเฉลี่ยของกลุ่ม สำหรับอีเมล์ฉบับนี้มีผลการคำนวณเป็นไปดังตารางที่ 4.2

ตารางที่ 4.2 แสดงการคิดค่าเฉลี่ยความน่าจะเป็นของคำสำคัญในอีเมล์ตัวอย่าง

Group	Word	$P(word_i)$	Probability of Gene
G1	sex	0.395	((1x0.395)+(1x0.005)/2) = 0.200
G1	adult	0.005	
G3	free	0.230	((1x0.230)+(1x0.090)/2) = 0.160
G3	special	0.090	

เนื่องจากค่าความน่าจะเป็นของคำสำคัญที่ต่ำสุดในฐานข้อมูลคือ 0.002 และสูงสุดคือ 1.000 ซึ่งในวิทยานิพนธ์นี้ เราพิจารณาใช้ค่าความแตกต่างด้วยค่าจุดทศนิยม 3 ตำแหน่ง ค่าในช่วง 0.004 – 1.000 จะพบว่ามีค่าที่เป็นจำนวนเต็มทั้งหมด 998 ค่า ซึ่งจะต้องนำมาแปลงเป็นเลขฐานสองเพื่อจะนำไปใช้ในกระบวนการเจเนติกอัลกอริทึม ทำให้เราจึงต้องใช้เลขฐานสองจำนวน 10 บิท ซึ่งมี 1,024 ค่าในการแสดงค่า แต่ก่อนอื่นเราต้องทำการปรับช่วงของค่า 998 ค่า เพื่อแสดงด้วยเลขฐานสอง 10 บิทเสียก่อน ซึ่งสามารถทำได้ดังสมการที่ (4.3)

$$\text{ความน่าจะเป็นที่ปรับแล้ว} = \left[\left(\frac{\text{ความน่าจะเป็นที่นำมาปรับ} - \text{ความน่าจะเป็นต่ำสุด}}{\text{ความน่าจะเป็นสูงสุด} - \text{ความน่าจะเป็นต่ำสุด}}\right) \times 1{,}024\right] + 1 \qquad (4.3)$$

ตัวอย่างเช่น ค่าความน่าจะเป็นที่นำมาปรับคือ 0.200 เมื่อคำนวณตามสมการ (4.3) จะได้ว่า

$$\text{ความน่าจะเป็นที่ปรับแล้ว} = [(0.200-0.004/1.000-0.004)*1024]+1$$
$$= 202_{10}$$
$$= (0001100110)_2$$

เมื่อนำทุกยีนมาคำนวณตามสมการที่ 4.2 จะได้ค่าของยีนที่แปลงจากเลขฐานสิบเป็นเลขฐานสองดังแสดงในตารางที่ 4.3

ตารางที่ 4.3 แสดงการแปลงค่าความน่าจะเป็นก่อนปรับและหลังปรับและค่าความน่าจะเป็นหลังปรับเมื่อแปลงเป็นเลขฐานสอง

	ค่าความน่าจะเป็นก่อนปรับ	ค่าความน่าจะเป็นหลังปรับ	ค่าความน่าจะเป็นหลังปรับ(ฐานสอง)
Gene1	0.200	202	0001100110
Gene2	0	0	0000000000
Gene3	0.106	105	0000110101
Gene4	0	0	0000000000
Gene5	0	0	0000000000
Gene6	0	0	0000000000
Gene7	0	0	0000000000
Gene8	0	0	0000000000

4.3.2 การประเมินค่าความเหมาะสม (Fitness Evaluation)

คือการหาชุดของโครโมโซมที่เหมาะสมเพื่อนำไปใช้ในกระบวนการจำแนกอีเมล์ โดยเริ่มจากการกำหนดฟังก์ชันความเหมาะสม (Defining Fitness Function) โดยมีแนวคิดที่ว่าฟังก์ชันที่กำหนดขึ้นมานั้นต้องมีความสามารถในการคัดเลือกโครโมโซมที่ดี และก่อให้เกิดประสิทธิภาพแก่ระบบให้มากที่สุด การกำหนดฟังก์ชันความเหมาะสม (Fitness Function) สำหรับงานวิทยานิพนธ์นี้ ได้กำหนดให้ค่าความเหมาะสมของแต่ละโครโมโซมแม่แบบคือค่าที่บ่งบอกความแม่นยำในการจำแนกอีเมล์ทดสอบ โดยนำชุดข้อมูลสำหรับเรียนรู้มาเป็นตัวทดสอบกับโครโมโซมแม่แบบ (Template) ว่าโครโมโซมแม่แบบแต่ละตัวมีความแม่นยำมากน้อยเท่าไร ซึ่งมีขั้นตอนการทำงานตามชุดคำสั่งเทียม (Pseudo Code) ที่แสดงได้ดังรูปที่ 4.3

```
แต่ละ Template_i  (i = 1 ~ จำนวนของ Template)
  แต่ละ Training_j  (j= 1~ จำนวนของ Training set)
        ถ้า Class ของ Template_i = Spam แล้ว /* Template_i ทำนายว่า Training_j เป็น Spam*/
          ถ้า Training_j = Spam แล้ว
             สถิติการ ClassifySpamAsSpam ของ Template_i ++
          ถ้า Training_j = Ham แล้ว
             สถิติการ ClassifyHamAsSpam ของ Template_i ++
        ถ้า Class ของ Template_i = Ham แล้ว /* Template_i ทำนายว่า Training_j เป็น Ham*/
          ถ้า Training_j = Spam แล้ว
             สถิติการ ClassifySpamAsHam ของ Template_i ++
          ถ้า Training_j = Ham แล้ว
             สถิติการ ClassifyHamAsHam ของ Template_i ++
```

รูปที่ 4.3 แสดงการเก็บผลการจำแนกของแต่ละแม่แบบ (Template) สำหรับใช้คำนวณหาค่าความเหมาะสม

ค่าที่ได้จากการนับข้างต้นจะนำมาใช้ในการคำนวณค่าความเหมาะสม (Fitness Value) ซึ่งเป็นไปดังสมการที่ (4.4)

$$\text{ค่าความเหมาะสม} = \left(N_{S\to S} + N_{H\to H}\right) - \left(N_{S\to H} + N_{H\to S}\right) \qquad (4.4)$$

เมื่อ $N_{S\to S}$ คือ จำนวนการทายอีเมล์ขยะเป็นอีเมล์ขยะ

$N_{H\to H}$ คือ จำนวนการทายอีเมล์ดีเป็นอีเมล์ดี

$N_{S\to H}$ คือ จำนวนการทายอีเมล์ขยะเป็นอีเมล์ดี

$N_{H\to S}$ คือ จำนวนการทายอีเมล์ดีเป็นเมล์ขยะ

ค่าความเหมาะสมของโครโมโซมแม่แบบทั้งหมดหลังจากคำนวณเสร็จแล้วแสดงดังตารางที่4.4

ตารางที่ 4.4 แสดงตัวอย่างโครโมโซมแม่แบบ และค่าความเหมาะสมของโครโมโซม

ลำดับที่	โครโมโซม	ค่าความเหมาะสม
1	[0100101100,1100101000,0000000000,0000000000, 1110010000, 0000000000, 1010001111,0000000000]	15
2	[0001011010, 0011001100, 0000000000, 000000000, 1010111001, 0000000000, 1010100000, 0000000000]	70
3	[0000000000, 000000010, 1001001011, 1000000000, 0000000000,0100110011, 0001000100, 1000100100]	2
4	[0000000000,1010001111, 1100001100,0000000000, 0000111100, 0000000000, 0011000100, 0100011000]	40
5	[0010000000, 1000000000, 0000000000,1111000000, 0011001110, 0101000000,1100110011, 0000000000]	83
…	…	…
1,600	[0000000000, 0011000011,0000000000,0101001111, 0000000000,0000000000,0000000000,0000000000]	25

4.3.3 การคัดเลือกประชากร (Selection of Population)

คือการคัดเลือกโครโมโซมที่เหมาะสมไว้สำหรับใช้เป็น โครโมโซมแม่แบบ(Template) ในรุ่นถัดไป โดยในรอบแรกจะต้องกำหนดประชากรตั้งต้นขึ้นมาจำนวนหนึ่ง ซึ่งได้จากการนำชุดข้อมูลสำหรับเรียนรู้ มาสร้างเป็นโครโมโซมแม่แบบ(Template) โดยในวิทยานิพนธ์นี้ได้กำหนดโครโมโซมสำหรับเป็นประชากรตั้งต้น 1,600 ตัว จากนั้นกำหนดจำนวน และเลือกโครโมโซมพ่อแม่ที่จะใช้ในการครอสโอเวอร์ขึ้นมา โดยสามารถกำหนดจำนวนโครโมโซมพ่อแม่เป็นเปอร์เซ็นต์เมื่อคิดจากประชากรตั้งต้นได้ เช่น กำหนดโครโมโซมพ่อแม่ 10% หมายความว่า จากโครโมโซมตั้งต้น 1,600 ตัว จะใช้โครโมโซมพ่อแม่ 160 คู่ ในการครอสโอเวอร์ในรอบถัดไป ภายหลังจากการครอสโอเวอร์ จะได้โครโมโซมใหม่ขึ้นมาจำนวน 160 คู่ หรือ 320 ตัว เมื่อรวมกับโครโมโซมตั้งต้นจะได้ 1,920 ตัว จากนั้นจะต้องทำการคัดเลือก

โครโมโซมที่เหมาะสมกว่าเอาไว้เพียง 1,600 ตัว เท่ากับจำนวนโครโมโซมตั้งต้นโดยการคัดเลือกนั้น จะใช้วงล้อถ่วงน้ำหนักเป็นตัวคัดเลือกว่าโครโมโซมไหนจะผ่านเข้าสู่กระบวนการเจเนติกในรุ่นถัดไป (อ่านรายละเอียดของการคัดเลือกแบบวงล้อถ่วงน้ำหนักได้ในบทที่ 3)

4.3.4 การครอสโอเวอร์และการมิวเตชัน (Crossover and Mutation)

การครอสโอเวอร์ทำขึ้นเพื่อสร้างประชากรรุ่นใหม่ให้มีความหลากหลายขึ้นในระบบ ในวิทยานิพนธ์นี้ ใช้การครอสโอเวอร์แบบหลายจุด (Multiple-Point Crossover) ซึ่งเราสามารถกำหนดจำนวนบิทที่ต้องการให้เกิดการครอสโอเวอร์เป็นเปอร์เซ็นต์ หรือจะทำการสุ่มเลือกจำนวนบิทที่จะนำมาครอสโอเวอร์กันก็ได้ โดยการครอสจะเกิดขึ้นเฉพาะยีนในกลุ่มเดียวกันเท่านั้น ไม่สามารถกระทำข้ามกลุ่มได้ ส่วนการมิวเตชันทำเพื่อประกันการสูญหาย และเพื่อความหลากหลายของข้อมูล สำหรับไบนารีมิวเตชัน เป็นการปรับเปลี่ยนข้อมูล ณ ตำแหน่งที่กำหนดนั้น โดยเปลี่ยนข้อมูลจาก 0 เป็น 1 สำหรับวิทยานิพนธ์นี้ เราสามารถกำหนดได้ว่าจะให้มีการมิวเตชันเกิดขึ้นหลังจากผ่านกระบวนการทางเจเนติกไปกี่รุ่น และจะมิวเตชันกี่เปอร์เซ็นต์ของบิททั้งหมดในโครโมโซม โดยจะสุ่มเลือกโครโมโซมที่จะมิวเตชัน จากนั้นจึงสุ่มเลือกบิตที่จะทำมิวเตชัน โดยจะทำการกลับบิตเพียงบิตเดียวเพื่อให้ไม่เกิดการกลายพันธุ์ของโครโมโซมมากเกินไป

เมื่อสิ้นสุดการดำเนินการทางเจเนติกแล้ว จะได้โครโมโซมที่ผ่านการคัดเลือกในรอบนี้มาจำนวน 1,000 โครโมโซม ซึ่งโครโมเหล่านี้จะถูกเรียกว่า โครโมโซมแม่แบบ (Template) ซึ่งจะมี 1,000 แม่แบบ แม่แบบอีเมล์ขยะเหล่านี้เปรียบเสมือนชุดของกฎ (Rule set) ตัวอย่างของโครโมโซมแม่แบบที่ได้แสดงดังรูปที่ 4.3 ซึ่งจะนำไปใช้ในขบวนการทดสอบเพื่อการจำแนกอีเมล์ต่อไป

	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6	Gene7	Gene8
Template 1	0100101100	1100101000	0000000000	0000000000	1110010000	0000000000	1010001111	0000000000

	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6	Gene7	Gene8
Template 2	0001011010	0011001100	0000000000	000000000	1010111001	0000000000	1010100000	0000000000

	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6	Gene7	Gene8
Template 3	0000000000	000000010	1001001011	1000000000	0000000000	0100110011	0001000100	1000100100

.

.

.

	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6	Gene7	Gene8
Template 1,600	0000000000	0011000011	0000000000	0101001111	0000000000	0000000000	0000000000	0000000000

รูปที่ 4.4 แสดงแม่แบบของอีเมล์ทั้งหมดที่ได้จากกระบวนการเจเนติกอัลกอริทึมในรุ่นนี้

จากรูปที่ 4.4 จะเห็นว่า แม่แบบจะมีค่าในแต่ละยีนแตกต่างกันเนื่องจากแม่แบบแต่ละตัวมาจากลักษณะของอีเมล์ที่ต่างกัน ซึ่งลักษณะเหล่านี้สามารถนำไปใช้ในการบ่งชี้รูปแบบ (Pattern) ของอีเมล์ที่มีลักษณะแตกต่างกันได้

4.3.5 การทดสอบประสิทธิภาพของระบบ (Testing Algorithm)

การทดสอบระบบว่าเป็นที่พอใจหรือไม่นั้นวัดได้จากความถูกต้องของการจำแนกอีเมล์ว่าเป็นอีเมล์ใดๆ เป็นอีเมล์ขยะหรืออีเมล์ดี โดยใช้ชุดทดสอบที่แยกไว้จำนวน 400 อีเมล์ โดยนำอีเมล์ดังกล่าวมาสร้างรูปแบบของโครโมโซมตามหัวข้อ 4.1.2 จากนั้นนำค่าความน่าจะเป็นเฉลี่ยของแต่ละยีนของโครโมโซมที่แปลงจากอีเมล์ดังกล่าวมาเปรียบเทียบกับค่าความน่าจะเป็นเฉลี่ยในยีนของโครโมโซมแม่แบบทีละโครโมโซม โดยแต่ละโครโมโซมแม่แบบจะทำการเปรียบเทียบไปทีละยีน ถ้าตรงตามเงื่อนไข ยีนนั้นจะได้ 1 คะแนน โดยจะทำการเปรียบเทียบเช่นนี้ไปจนครบ 8 ยีน เมื่อรวมคะแนนแล้วพบว่าจำนวนยีนที่มีเงื่อนไขตรงกัน มากกว่าหรือเท่ากับค่า Gene Threshold แสดงว่าอีเมล์ที่เข้ามาจำแนกนี้ มีรูปแบบที่คล้ายคลึงกันกับโครโมโซมแม่แบบนี้ เพราะฉะนั้น โครโมโซมแม่แบบนี้จะให้คะแนนอีเมล์ฉบับนี้ว่ามีชนิดเดียวกับโครโมโซมแม่แบบนี้ แต่ถ้าไม่ตรงตามเงื่อนไขก็จะให้คะแนนอีเมล์ว่ามีชนิดตรงกันข้ามกับโครโมโซมแม่แบบนี้ ซึ่งมีขั้นตอนการทำงานตามชุดคำสั่งเทียม (Pseudo Code) ที่แสดงได้ ดังรูปที่ 4.5

```
แต่ละ Template_i  (i = 1 ~ จำนวนของ Template)
   แต่ละ Testing_j  (j= 1~ จำนวนของ Testing set)
      แต่ละ Gene_k  (k = 1~ 8)
         ถ้า (ค่าของ Gene_k ของ Testing_j ) ≥ (ค่าของ Gene_k ของ Template_i ) แล้ว
            (GeneMatchCount ของ Testing_j ) ++
       ถ้า (GeneMatchCount ของ Testing_j ) ≥ Gene Threshold แล้ว
         ถ้า Class ของ Template_i = Spam แล้ว /* Template_i ทำนายว่า Testing_j เป็น Spam*/
            ถ้า Testing_j = Spam แล้ว
               สถิติการ ClassifySpamAsSpam ของ Template_i ++
            ถ้า Testing_j = Ham แล้ว
               สถิติการ ClassifyHamAsSpam ของ Template_i ++
         ถ้า Class ของ Template_i = Ham แล้ว /* Template_i ทำนายว่า Testing_j เป็น Ham*/
            ถ้า Testing_j = Spam แล้ว
               สถิติการ ClassifySpamAsHam ของ Template_i ++
            ถ้า Testing_j = Ham แล้ว
               สถิติการ ClassifyHamAsHam ของ Template_i ++
```

รูปที่ 4.5 แสดงการจำแนกอีเมล์โดยพิจารณาเปรียบเทียบจากค่าความน่าจะเป็นเฉลี่ยในยีน

เมื่อ $Template_i$ คือ Template ที่กำลังพิจารณา

$Testing_j$ คือ อีเมล์ที่กำลังพิจารณา

Gene Threshold คือจำนวนขั้นต่ำที่กำหนดไว้ให้ยีนของอีเมล์ที่นำมาจำแนกตรงกับยีนของโครโมโซมแม่แบบ เช่น ถ้า Gene Threshold =3 หมายความว่า ความน่าจะเป็นเฉลี่ยของยีนในอีเมล์ที่นำมาจำแนกต้องตรงตามเงื่อนไขกับค่าความน่าจะเป็นเฉลี่ยของยีนในโครโมโซมแม่แบบ 3 ยีนขึ้นไปโดยโครโมโซมแม่แบบนี้จะให้คะแนนอีเมล์ที่นำมาจำแนกนี้ว่ามีความคล้ายคลึงกับโครโมโซมแม่แบบ แต่ถ้าไม่ตรงตามเงื่อนไข ก็จะให้คะแนนที่ตรงกันข้ามแก่อีเมล์นี้แทน

จากนั้นนำอีเมล์ที่ต้องการจำแนกไปเปรียบเทียบกับโครโมโซมแม่แบบทั้งหมด 1,600 แม่แบบ แล้วรวมคะแนนการของการจำแนก เปรียบเทียบคะแนนระหว่างคะแนนอีเมล์ขยะและคะแนนอีเมล์ดี ก็จะสามารถทำนายว่าอีเมล์ที่นำมาจำแนกเป็นอีเมล์ชนิดใด

ถ้าผลลัพธ์ของการจำแนกอีเมล์ใน Generation นี้มีความเหมาะสมและเป็นที่พอใจ ซึ่งวัดจากค่า Accuracy, Recall และ Precision จึงทำการหยุดกระบวนการเจเนติกอัลกอริทึม ถ้ายังไม่

เป็นที่พอใจก็ดำเนินการทางเจเนติก เช่นการครอสโอเวอร์ การมิวเตชันใน Generation ถัดไป จนกระทั่งได้ผลลัพธ์ที่เหมาะสมและเป็นที่พอใจ