

บทที่ 5

ผลการทดลองและการวิเคราะห์

5.1 ขั้นตอนการทดลองตัวกรองอีเมลขยะโดยใช้เจเนติกอัลกอริทึม

ขั้นตอนในการกรองอีเมลขยะโดยใช้เจเนติกอัลกอริทึมมีส่วนหลักอยู่ 3 ส่วนคือ การเตรียมข้อมูล (Data Preparation) การสร้างฐานข้อมูลคำ (Creating Data Dictionary) และ การประยุกต์ใช้เจเนติกอัลกอริทึม (Adopting Genetic Algorithm)

1. การเตรียมอีเมลสำหรับงานวิทยานิพนธ์นี้ ใช้อีเมลขยะจำนวน 1,000 ฉบับ และใช้อีเมลดีจำนวน 1,000 ฉบับ และรวมทั้งสิ้น 2,000 ฉบับ โดยทำการทดลองแบบสลับชุดข้อมูล 5 ส่วน (5-folds Cross Validation) โดยแบ่งชุดข้อมูลออกเป็น 5 ส่วนเท่าๆกัน แล้วทำการทดลองทั้งสิ้น 5 ครั้ง แต่ทุกครั้งจะใช้ชุดข้อมูล 4 ส่วนเป็นชุดข้อมูลฝึกหัดและใช้ชุดข้อมูล 1 ส่วนเป็นชุดข้อมูลทดสอบ จากนั้นอีเมลจะถูกนำไปผ่านกระบวนการตัดคำ (ดูรายละเอียดของการตัดคำได้ในบทที่ 4) ก่อนเข้าสู่กระบวนการเรียนรู้ของตัวกรองต่อไป
2. กระบวนการเรียนรู้ของตัวกรอง ประกอบไปด้วย การแปลงอีเมลให้เป็นโครโมโซมการหาค่าความน่าจะเป็นให้ยีนในโครโมโซม การดำเนินการทางเจเนติกอัลกอริทึม ซึ่งเป็นการทำให้ได้รูปแบบของอีเมลที่หลากหลายขึ้นมา แล้วคัดเลือกโครโมโซมที่เหมาะสมมาสร้างเป็นโครโมโซมแม่แบบเพื่อใช้จำแนกอีเมลต่อไป (สามารถดูรายละเอียดของแต่ละส่วนได้ในบทที่ 4)
3. กระบวนการจำแนกอีเมล ประกอบไปด้วย การแปลงอีเมลทดสอบให้เป็นโครโมโซมแล้วหาค่าความน่าจะเป็นให้ยีนในโครโมโซม เพื่อที่จะนำค่าความน่าจะเป็นของยีนในโครโมโซมที่แปลงมาจากอีเมลทดสอบไปเปรียบเทียบกับค่าความน่าจะเป็นของยีนในโครโมโซมซึ่งเป็นแม่แบบซึ่งได้จากกระบวนการเรียนรู้ของตัวกรอง (สามารถดูรายละเอียดของการจำแนกได้ในบทที่ 4)

5.2 การปรับพารามิเตอร์ต่างๆ ในงานวิทยานิพนธ์

เนื่องจากตัวดำเนินการในเจเนติกอัลกอริทึมประกอบไปด้วย การคัดเลือก การครอสโอเวอร์ และการมิวเตชัน ซึ่งแต่ละตัวดำเนินการก็สามารถทำได้หลายวิธี เช่น การคัดเลือก มีแบบการคัดเลือกแบบจัดลำดับ (Ranking) การคัดเลือกแบบวงล้อถ่วงน้ำหนัก (Roulette Wheel) หรือในครอสโอเวอร์ มีเปอร์เซ็นต์ของการครอสเป็นต้น สิ่งเหล่านี้เป็นพารามิเตอร์ที่เราจะต้องหาค่าที่เหมาะสมเพื่อให้ได้ผลลัพธ์ที่ดีที่สุด โดยผลลัพธ์ที่ดีที่สุดสำหรับวิทยานิพนธ์นี้จะเปรียบเทียบจากค่า Accuracy ซึ่งเป็นค่าที่บอกความถูกต้องของการจำแนกอีเมลขยะและอีเมลดี

Recall คือค่าที่บอกความถูกต้องในการจำแนกอีเมลขยะ และ Precision คือค่าที่บอกความถูกต้องในการจำแนกอีเมลดี ซึ่งทั้งสามค่าสามารถคิดได้จากสมการ (5.1), (5.2) และ (5.3) [22]

$$\text{Accuracy} = \frac{N_{S \rightarrow S} + N_{H \rightarrow H}}{N_{S \rightarrow S} + N_{H \rightarrow H} + N_{S \rightarrow H} + N_{H \rightarrow S}} \quad (5.1)$$

$$\text{Precision} = \frac{N_{S \rightarrow S}}{N_{S \rightarrow S} + N_{H \rightarrow S}} \quad (5.2)$$

$$\text{Recall} = \frac{N_{S \rightarrow S}}{N_{S \rightarrow S} + N_{S \rightarrow H}} \quad (5.3)$$

เมื่อ $N_{S \rightarrow S}$ คือ จำนวนของอีเมลขยะที่ถูกทายว่าเป็นอีเมลขยะ
 $N_{H \rightarrow H}$ คือ จำนวนของอีเมลดีที่ถูกทายว่าเป็นอีเมลดี
 $N_{S \rightarrow H}$ คือ จำนวนของอีเมลขยะที่ถูกทายเป็นเมลล์ดี
 $N_{H \rightarrow S}$ คือ จำนวนของอีเมลดีที่ถูกทายเป็นเมลล์ขยะ

ในงานวิทยานิพนธ์ได้ทำการทดลองเพื่อหาค่าพารามิเตอร์ที่เหมาะสมของแต่ละตัวดำเนินการ โดยมีพารามิเตอร์กลางที่ได้จากการทดลองอย่างคร่าวๆ ดังนี้

- จำนวนคู่โครโมโซมพ่อแม่ 10%
- วิธีการคัดเลือกแบบวงล้อถ่วงน้ำหนัก
- การครอสโอเวอร์แบบสุ่ม
- ไม่ใช้มิวเตชันพารามิเตอร์
- ยีนเรโซว์ 5 ยีน

จากนั้นนำพารามิเตอร์กลางเหล่านี้ไปใช้ในการหาพารามิเตอร์ที่เหมาะสมทีละตัว โดย Fix พารามิเตอร์ตัวอื่นๆ เอาไว้ โดยมีรายละเอียดของการหาพารามิเตอร์ที่เหมาะสมแต่ละตัวดังต่อไปนี้

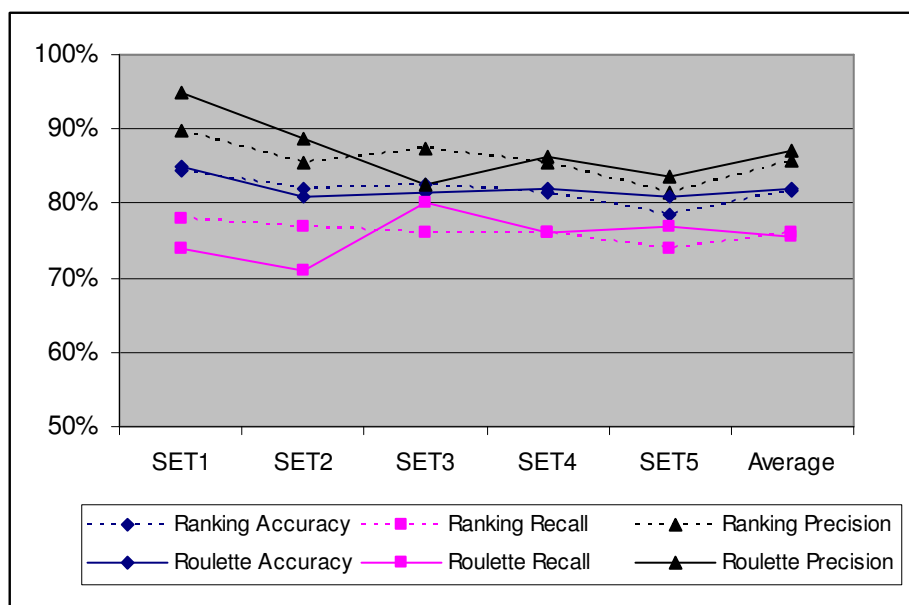
5.2.1 การเปรียบเทียบวิธีการคัดเลือกโครโมโซมระหว่างการคัดเลือกแบบวงล้อถ่วงน้ำหนัก (Roulette Wheel) และการคัดเลือกแบบจัดลำดับ (Ranking)

งานวิทยานิพนธ์นี้ ได้ทำการเปรียบเทียบการคัดเลือกแบบวงล้อถ่วงน้ำหนักและการจัดลำดับกับชุดข้อมูล 5 ชุด ซึ่งได้ผลการทดลองดังแสดงในตารางที่ 5.1

ตารางที่ 5.1 แสดงการเปรียบเทียบค่า Accuracy, Recall และ Precision ระหว่างการคัดเลือกแบบวงล้อถ่วงน้ำหนัก (Roulette Wheel) และการคัดเลือกแบบจัดลำดับ (Ranking)

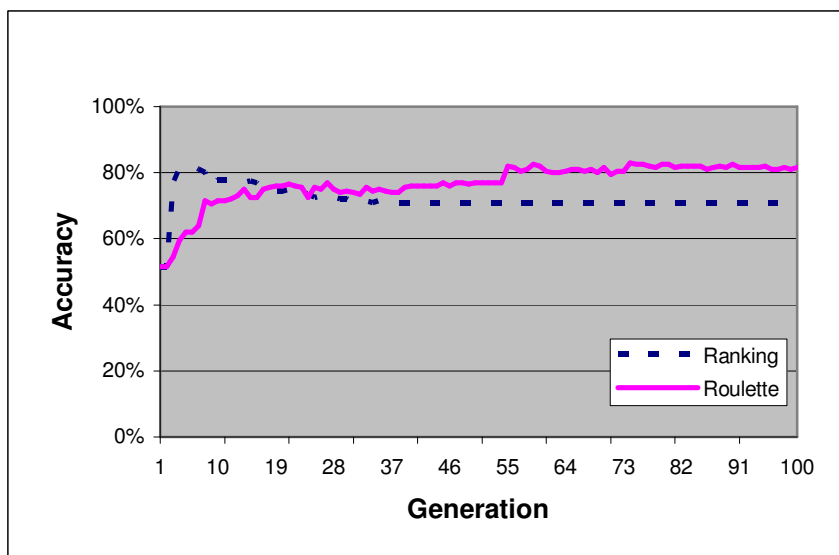
	Ranking			Roulette Wheel		
	Accuracy	Recall	Precision	Accuracy	Recall	Precision
SET1	84.50%	78.00%	89.66%	85.00%	74.00%	94.87%
SET2	82.00%	77.00%	85.56%	81.00%	71.00%	88.75%
SET3	82.50%	76.00%	87.36%	81.50%	80.00%	82.47%
SET4	81.50%	76.00%	85.39%	82.00%	76.00%	86.36%
SET5	78.50%	74.00%	81.32%	81.00%	77.00%	83.70%
Average	81.80%	76.20%	85.86%	82.10%	75.60%	87.23%

จากตารางที่ 5.1 พบว่าค่าเฉลี่ยของ Accuracy, Recall และ Precision ของวิธีการคัดเลือกแบบวงล้อถ่วงน้ำหนักและแบบจัดลำดับมีค่าใกล้เคียงกันและเมื่อนำค่า Accuracy, Recall และ Precision ของวิธีการคัดเลือกแบบวงล้อถ่วงน้ำหนักและแบบจัดลำดับมาสร้างเป็นกราฟ พบว่าทั้งสองวิธีมีค่าใกล้เคียงกันโดยค่าเฉลี่ยของวิธีการคัดเลือกแบบวงล้อถ่วงน้ำหนักมีค่าดีกว่าเล็กน้อยดังรูป 5.1

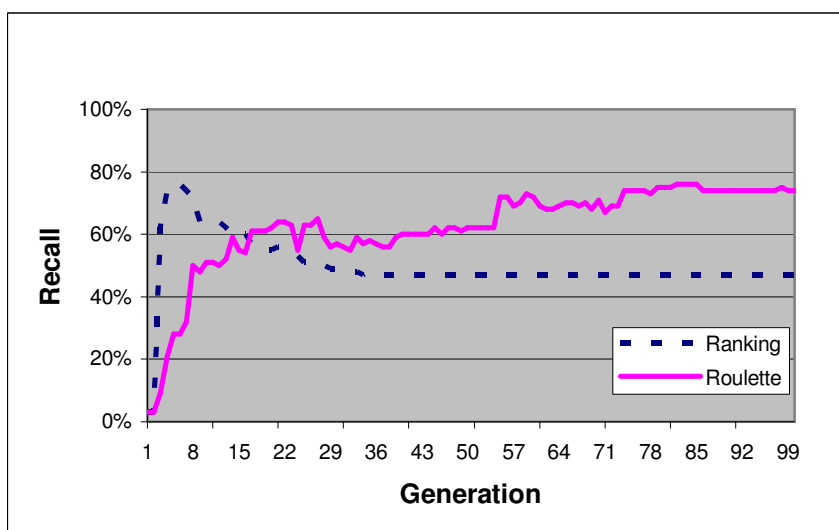


รูปที่ 5.1 แสดงการเปรียบเทียบค่า Accuracy, Recall และ Precision ระหว่างการคัดเลือกแบบวงล้อถ่วงน้ำหนัก (Roulette Wheel) และการคัดเลือกแบบจัดลำดับ (Ranking)

รูปที่ 5.2 ถึง 5.4 เป็นการแสดงค่า Accuracy, Recall และ Precision ของการคัดเลือกแบบวงล้อถ่วงน้ำหนัก โดยใช้เลือกชุดข้อมูลที่ใกล้เคียงกับค่าเฉลี่ย มาสร้างเป็นกราฟเพื่อให้เห็นภาพรวมของผลของวิธีการคัดเลือกทั้งสองแบบ

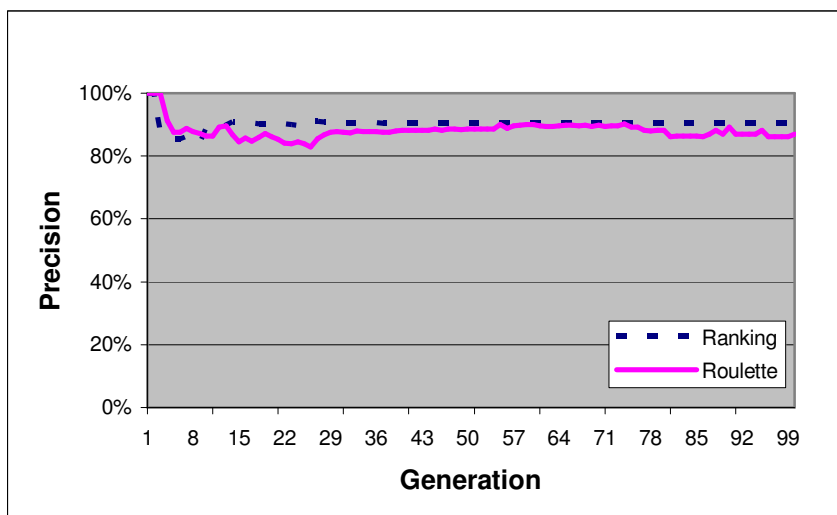


รูปที่ 5.2 แสดงค่า Accuracy ของการการคัดเลือกแบบวงล้อถ่วงน้ำหนัก และการคัดเลือกแบบจัดลำดับ



รูปที่ 5.3 แสดงค่า Recall ของการการคัดเลือกแบบวงล้อถ่วงน้ำหนัก และการคัดเลือกแบบจัดลำดับ

จากรูปที่ 5.2 และจากรูปที่ 5.3 จะเห็นว่าการคัดเลือกแบบจัดลำดับนั้นจะมีค่า Accuracy และ Recall ดีกว่าการคัดเลือกแบบวงล้อถ่วงน้ำหนักในรุ่นแรกๆ และจะลดต่ำลงเรื่อยๆ และคงที่ในที่สุด ส่วนค่า Precision ของการคัดเลือกทั้งสองแบบจะมีค่าใกล้เคียงกันดังรูปที่ 5.4

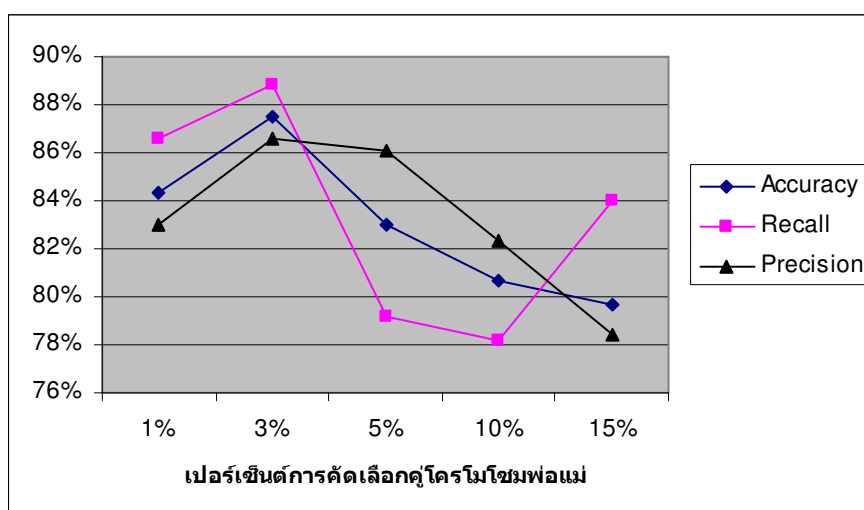


รูปที่ 5.4 แสดงค่า Precision ของการคัดเลือกแบบวงล้อถ่วงน้ำหนัก และการคัดเลือกแบบจัดลำดับ

การคัดเลือกแบบจัดลำดับนั้นอาจทำให้เกิดโครโมโซมที่มีลักษณะคล้ายคลึงกันมากเกินไปได้ (Over Crowding) เนื่องจากการพยายามเลือกแต่โครโมโซมที่มีค่าความเหมาะสมสูงๆ เอาไว้ ส่งผลให้การไม่เกิดโครโมโซมที่มีความหลากหลาย ดังนั้น สำหรับการคัดเลือกในงานในวิทยานิพนธ์นี้ จะใช้การคัดเลือกแบบวงล้อถ่วงน้ำหนักซึ่งสามารถแก้ปัญหาข้างต้นได้

5.2.2 การปรับเปอร์เซ็นต์การเลือกคู่โครโมโซมพ่อแม่ (Parent Pair)

การปรับเปอร์เซ็นต์การเลือกคู่โครโมโซมพ่อแม่ ได้ทำการทดลองกับข้อมูล 5 ชุด โดยในแต่ละชุดมีจำนวนโครโมโซมเป็น 1%, 3%, 5%, 10%, 15% ได้ผลการทดลองดังรูปที่ 5.5 และตารางที่ 5.2



รูปที่ 5.5 แสดงค่า Accuracy, Recall และ Precision ที่การคัดเลือกคู่โครโมโซมพ่อแม่ที่เปอร์เซ็นต์ต่างๆ

ตารางที่ 5.2 แสดงค่า Accuracy, Recall และ Precision ที่การคัดเลือกคู่โครโมโซมพ่อแม่ที่เปอร์เซ็นต์ต่างๆ

Parent Pair	Percent	Accuracy	Recall	Precision
SET 1	1%	85.00%	89.00%	82.41%
	3%	87.50%	87.00%	87.88%
	5%	86.50%	87.00%	86.14%
	10%	85.00%	87.00%	83.65%
	15%	84.00%	84.00%	84.00%
SET 2	1%	81.00%	76.00%	84.44%
	3%	88.50%	90.00%	87.38%
	5%	78.00%	64.00%	88.89%
	10%	79.00%	71.00%	84.52%
	15%	79.50%	81.00%	78.64%
SET 3	1%	83.50%	93.00%	78.15%
	3%	87.50%	91.00%	85.05%
	5%	82.00%	75.00%	87.21%
	10%	82.50%	84.00%	81.55%
	15%	79.50%	98.00%	71.53%
SET 4	1%	85.00%	89.00%	82.41%
	3%	86.50%	90.00%	84.11%
	5%	84.00%	80.00%	86.96%
	10%	78.00%	76.00%	79.17%
	15%	82.50%	75.00%	88.24%
SET 5	1%	87.00%	86.00%	87.76%
	3%	87.50%	86.00%	88.66%
	5%	84.50%	90.00%	81.08%
	10%	79.00%	73.00%	82.95%
	15%	73.00%	82.00%	69.49%
Average	1%	84.30%	86.60%	83.03%
	3%	87.50%	88.80%	86.62%
	5%	83.00%	79.20%	86.06%
	10%	80.70%	78.20%	82.37%
	15%	79.70%	84.00%	78.38%

จากการทดลองพบว่า ค่าเฉลี่ย Accuracy, Recall และ Precision ของเปอร์เซ็นต์การคัดเลือกคู่โครโมโซมพ่อแม่ที่ดีที่สุดคือ ที่การเลือกโครโมโซมพ่อแม่ที่ 3% ดังตารางที่ 5.2 ทั้งนี้เนื่องจากการเลือกคู่โครโมโซมพ่อแม่จำนวนน้อยๆ เพื่อมาทำการตามกระบวนการเจเนติกนั้นจะทำให้เกิดการเปลี่ยนแปลงอย่างค่อยเป็นค่อยไป จึงมีโอกาสพบคำตอบของปัญหาได้หลากหลายกว่าการเลือกคู่โครโมโซมพ่อแม่ที่เปอร์เซ็นต์สูงๆ และเมื่อนำมาสร้างเป็นกราฟจะเห็นได้อย่างชัดเจนว่า การเลือกคู่โครโมโซมพ่อแม่ที่ 3% นั้นให้ผลลัพธ์ที่ดีมากกว่าที่เปอร์เซ็นต์อื่นๆ ดังรูป 5.5

5.2.3 การปรับค่ามิวเตชันพารามิเตอร์

การปรับค่ามิวเตชันพารามิเตอร์ในงานวิทยานิพนธ์นี้มี 2 พารามิเตอร์ด้วยกันคือ จำนวน Generation ที่ผ่านไวก่อนทำการมิวเตชัน และ เปอร์เซนต์ของโครโมโซมที่จะถูกมิวเตชัน (Chromosome) การปรับค่าพารามิเตอร์ทั้งสอง เพื่อหาค่าที่เหมาะสมสำหรับการมิวเตชันนั้นแสดงดังตารางที่ 5.3 และ 5.4

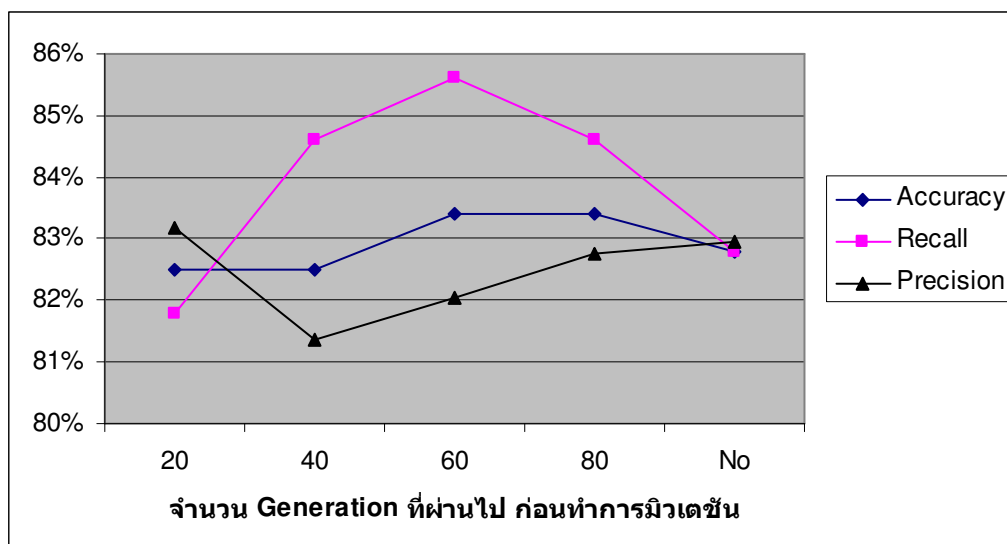
ตารางที่ 5.3 แสดงค่า Accuracy, Recall และ Precision ของการมิวเตชันเมื่อผ่านไวก 20, 40, 60, 80 Generation และเมื่อไม่ทำการมิวเตชันเลย (No)

Mutation	Generation	Accuracy	Recall	Precision
SET 1	20	82.50%	84.00%	81.55%
	40	84.00%	87.00%	82.08%
	60	84.00%	90.00%	80.36%
	80	83.50%	85.00%	82.52%
	No	84.50%	90.00%	81.08%
SET 2	20	81.00%	74.00%	86.05%
	40	80.00%	74.00%	84.09%
	60	84.00%	84.00%	84.00%
	80	82.00%	77.00%	85.56%
	No	80.50%	75.00%	84.27%
SET 3	20	83.50%	89.00%	80.18%
	40	83.00%	88.00%	80.00%
	60	83.00%	85.00%	81.73%
	80	84.00%	86.00%	82.69%
	No	80.50%	75.00%	84.27%
SET 4	20	84.00%	86.00%	82.69%
	40	83.00%	84.00%	82.35%
	60	83.00%	85.00%	81.73%
	80	84.50%	86.00%	83.50%
	No	84.50%	85.00%	84.16%
SET 5	20	81.50%	76.00%	85.39%
	40	82.50%	90.00%	78.26%
	60	83.00%	84.00%	82.35%
	80	83.00%	89.00%	79.46%
	No	84.00%	89.00%	80.91%
Average	20	82.50%	81.80%	83.17%
	40	82.50%	84.60%	81.36%
	60	83.40%	85.60%	82.03%
	80	83.40%	84.60%	82.75%
	No	82.80%	82.80%	82.94%

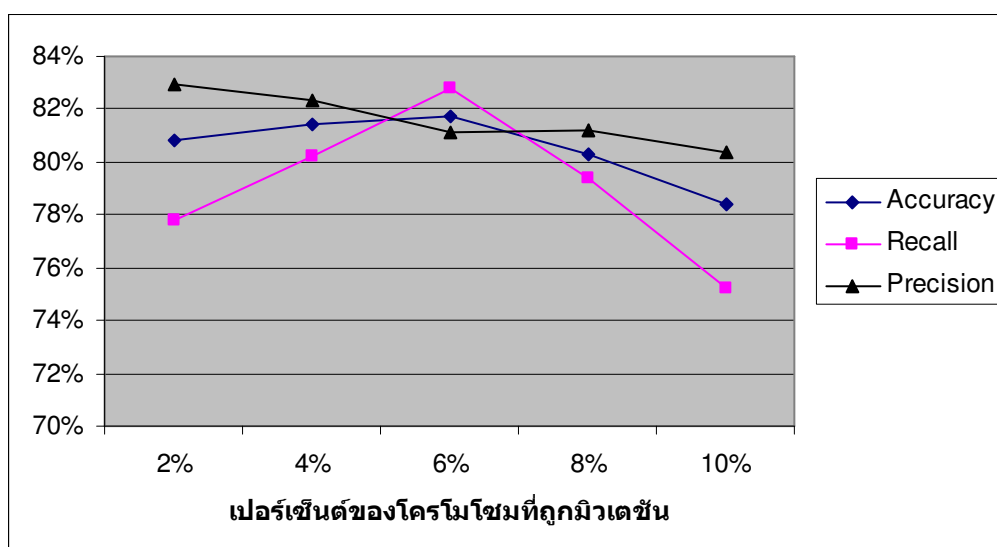
ตารางที่ 5.4 แสดงค่า Accuracy, Recall และ Precision ของการมิวเตชันเมื่อปรับจำนวน โครโมโซมที่เปอร์เซ็นต์ต่างๆ

Mutation	Chromosome	Accuracy	Recall	Precision
SET 1	2%	82.00%	85.00%	80.19%
	4%	83.50%	84.00%	83.17%
	6%	80.00%	74.00%	84.09%
	8%	79.00%	68.00%	87.18%
	10%	78.00%	74.00%	80.43%
SET 2	2%	80.00%	73.00%	84.88%
	4%	82.00%	85.00%	80.19%
	6%	81.50%	86.00%	78.90%
	8%	79.50%	79.00%	79.80%
	10%	76.50%	70.00%	80.46%
SET 3	2%	80.00%	74.00%	84.09%
	4%	79.50%	72.00%	84.71%
	6%	83.00%	86.00%	81.13%
	8%	80.00%	81.00%	79.41%
	10%	78.50%	77.00%	79.38%
SET 4	2%	79.50%	73.00%	83.91%
	4%	79.50%	74.00%	83.15%
	6%	82.50%	84.00%	81.55%
	8%	82.00%	86.00%	79.63%
	10%	77.00%	70.00%	81.40%
SET 5	2%	82.50%	84.00%	81.55%
	4%	82.50%	86.00%	80.37%
	6%	81.50%	84.00%	80.00%
	8%	81.00%	83.00%	79.81%
	10%	82.00%	85.00%	80.19%
Average	2%	80.80%	77.80%	82.92%
	4%	81.40%	80.20%	82.32%
	6%	81.70%	82.80%	81.14%
	8%	80.30%	79.40%	81.17%
	10%	78.40%	75.20%	80.37%

จากตารางที่ 5.3 ถึงตารางที่ 5.4 และรูป 5.6 ถึงรูปที่ 5.7 แสดงให้เห็นว่า ค่าเฉลี่ยค่า Accuracy, Recall และ Precision ที่ดีคือ ทำการมิวเตชันเมื่อผ่านการดำเนินการทางเจเนติกไป 60 Generation และเปอร์เซ็นต์ของโครโมโซมที่ถูกมิวเตชันซึ่งให้ค่าเฉลี่ยของค่า Accuracy, Recall และ Precision ที่ดีคือ 6 % แต่อย่างไรก็ตาม ค่าเฉลี่ยของ ค่า Accuracy, Recall และ Precision เมื่อทำการมิวเตชัน ไม่สามารถกำหนดแนวโน้มที่แน่นอนได้และไม่ได้ทำให้ผลการทดลองดีขึ้นเสมอไป ในบางกรณี อาจทำให้ผลการทดลองแย่ลง เนื่องจากการมิวเตชันนั่นเอง ดังนั้นในงานวิทยานิพนธ์นี้จึงเลือกที่จะ ไม่ใช้มิวเตชันพารามิเตอร์



รูปที่ 5.6 แสดงค่า Accuracy, Recall และ Precision ของการมิวเตชันเมื่อผ่านไป 20, 40, 60, 80 Generation และเมื่อไม่ทำการมิวเตชันเลย (No)



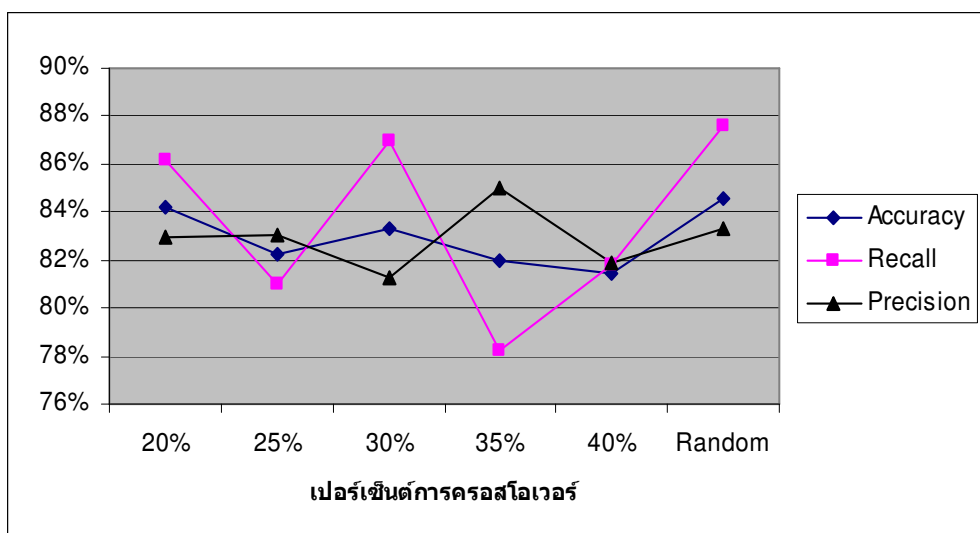
รูปที่ 5.7 แสดงค่า Accuracy, Recall และ Precision ของการมิวเตชันเมื่อปรับจำนวนโครโมโซมที่เปอร์เซ็นต์ต่างๆ

5.2.4 การปรับครอสโอเวอร์พารามิเตอร์

การปรับพารามิเตอร์ของการครอสโอเวอร์ทำได้ 2 แบบ คือการครอสโอเวอร์แบบสุ่ม (Random) และการครอสโอเวอร์แบบกำหนดเปอร์เซ็นต์คงที่ ซึ่งในวิทยานิพนธ์นี้ได้กำหนดเปอร์เซ็นต์ที่ 20%, 25%, 30%, 35%, 40% และได้ผลการทดลองดังแสดงในตารางที่ 5.5

ตารางที่ 5.5 แสดงค่า Accuracy, Recall และ Precision ของการครอสโอเวอร์ทั้งสองแบบ

Crossover	Percent	Accuracy	Recall	Precision
SET 1	20%	85.50%	91.00%	81.98%
	25%	87.50%	88.00%	87.13%
	30%	88.00%	88.00%	88.00%
	35%	84.00%	77.00%	90.00%
	40%	80.50%	96.00%	73.28%
	Random	80.00%	95.00%	73.08%
SET 2	20%	83.50%	84.00%	83.17%
	25%	81.50%	84.00%	80.00%
	30%	83.00%	85.00%	81.73%
	35%	80.50%	80.00%	80.81%
	40%	79.00%	78.00%	79.59%
	Random	84.00%	86.00%	82.69%
SET 3	20%	83.00%	88.00%	80.00%
	25%	79.00%	71.00%	84.52%
	30%	85.00%	96.00%	78.69%
	35%	83.00%	76.00%	88.37%
	40%	83.50%	81.00%	85.26%
	Random	87.50%	90.00%	85.71%
SET 4	20%	84.50%	82.00%	86.32%
	25%	84.00%	83.00%	84.69%
	30%	82.00%	82.00%	82.00%
	35%	81.50%	84.00%	80.00%
	40%	83.00%	81.00%	84.38%
	Random	85.50%	82.00%	88.17%
SET 5	20%	84.50%	86.00%	83.50%
	25%	79.00%	79.00%	79.00%
	30%	78.50%	84.00%	75.68%
	35%	81.00%	74.00%	86.05%
	40%	81.00%	73.00%	86.90%
	Random	86.00%	85.00%	86.73%
Average	20%	84.20%	86.20%	82.99%
	25%	82.20%	81.00%	83.07%
	30%	83.30%	87.00%	81.22%
	35%	82.00%	78.20%	85.05%
	40%	81.40%	81.80%	81.88%
	Random	84.60%	87.60%	83.28%



รูปที่ 5.8 แสดงค่า Accuracy, Recall และ Precision ของการครอสโอเวอร์ทั้งสองแบบ

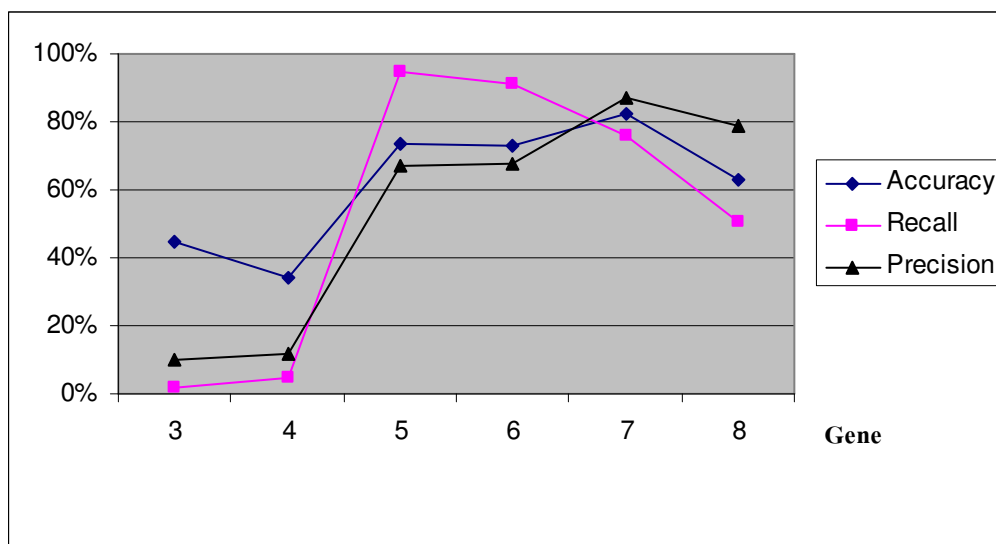
เมื่อทำการทดลองกับชุดข้อมูลทั้ง 5 เซตพบว่า ค่าเฉลี่ย Accuracy, Recall และ Precision ของการครอสโอเวอร์แบบกำหนดเปอร์เซ็นต์ครั้งที่เปอร์เซ็นต์ต่างๆ กัน ไม่สามารถระบุแนวโน้มที่แน่นอนได้และมีค่าเฉลี่ยดังกล่าวน้อยกว่าการครอสโอเวอร์แบบสุ่มเล็กน้อย ดังตารางที่ 5.5 และรูปที่ 5.8 สำหรับในงานวิทยานิพนธ์นี้จึงเลือกการครอสโอเวอร์แบบสุ่ม (Random)

5.2.5 การปรับค่าอินเทอร์โซว

การปรับค่าอินเทอร์โซวคือการปรับพารามิเตอร์จำนวนของอินที่ตรงตามเงื่อนไขระหว่างอินในโครโมโซมแม่แบบและอินในโครโมโซมของอีเมลทดสอบ ซึ่งสามารถปรับได้ตั้งแต่ 1 ถึง 8 (1 โครโมโซมมี 8 อิน) เมื่อทำการทดลองกับชุดข้อมูลทั้ง 5 เซตพบว่า ค่าเฉลี่ยของการปรับค่าอินเท่ากับ 7 ให้ผลลัพธ์ของการกรองที่ดี ดังนั้นจึงเลือกปรับค่าอินเทอร์โซวเท่ากับ 7 เนื่องจากให้ผลลัพธ์ที่ดีกว่าค่าอื่นๆ ดังแสดงในตารางที่ 5.6

ตารางที่ 5.6 แสดงค่า Accuracy, Recall และ Precision ของยีนเครือข่ายที่แตกต่างกัน

Gene Threshold	Gene	Accuracy	Recall	Precision
SET 1	3	45.00%	1.00%	8.33%
	4	36.00%	2.00%	6.25%
	5	66.50%	99.00%	60.00%
	6	79.00%	95.00%	71.97%
	7	85.00%	74.00%	94.87%
	8	38.50%	76.00%	43.43%
SET 2	3	43.50%	3.00%	15.79%
	4	34.50%	6.00%	13.95%
	5	78.50%	88.00%	73.95%
	6	67.50%	91.00%	61.90%
	7	81.00%	71.00%	88.75%
	8	66.00%	35.00%	92.11%
SET 3	3	49.00%	0.00%	0.00%
	4	18.00%	8.00%	10.00%
	5	78.00%	99.00%	69.72%
	6	85.00%	93.00%	80.17%
	7	81.50%	80.00%	82.47%
	8	69.50%	43.00%	91.49%
SET 4	3	46.00%	2.00%	16.67%
	4	37.00%	5.00%	13.89%
	5	73.50%	93.00%	66.91%
	6	70.50%	84.00%	66.14%
	7	82.00%	76.00%	86.36%
	8	74.50%	56.00%	88.89%
SET 5	3	39.50%	2.00%	8.00%
	4	45.00%	2.00%	14.29%
	5	72.50%	95.00%	65.52%
	6	63.00%	94.00%	58.02%
	7	81.00%	77.00%	83.70%
	8	66.00%	43.00%	79.63%
Average	3	44.60%	1.60%	9.76%
	4	34.10%	4.60%	11.68%
	5	73.80%	94.80%	67.22%
	6	73.00%	91.40%	67.64%
	7	82.10%	75.60%	87.23%
	8	62.90%	50.60%	79.11%



รูปที่ 5.9 แสดงค่า Accuracy, Recall และ Precision ของยีนเครือข่ายที่แตกต่างกัน

การปรับค่ายีนเครือข่ายน้อยเกินไป เช่น 3 ยีนหรือ 4 ยีน จะทำให้ค่าเฉลี่ยของ Accuracy, Recall และ Precision ต่ำมาก ๆ เนื่องจากมีความเป็นไปได้สูงที่ยีนของโครโมโซมที่เป็นอีเมลทดสอบจะตรงกับยีนในโครโมโซมแม่แบบเพียง 3 ถึง 4 โครโมโซม ซึ่งยังไม่มากเพียงพอที่จะระบุประเภทของอีเมลได้ ทำให้ผลลัพธ์ต่ำมาก การปรับค่ายีนเครือข่ายที่ 5 หรือ 6 นั้นพบว่าผลลัพธ์ที่ดีขึ้นมาก แต่จะดีที่สุดเมื่อปรับค่ายีนเครือข่ายที่ 7 ยีน และผลจะต่ำลงเมื่อปรับค่ายีนเครือข่ายที่ 8 ซึ่งมากเกินไป เพราะเป็นไปได้ยากที่โครโมโซมที่เป็นอีเมลทดสอบจะตรงกับยีนในโครโมโซมแม่แบบถึง 8 ยีน ดังตารางที่ 5.6 และรูปที่ 5.9

จากผลการทดลองการปรับพารามิเตอร์ข้างต้น จะเห็นว่า การปรับวิธีการคัดเลือก, มิวเตชัน พารามิเตอร์และครอสโอเวอร์พารามิเตอร์นั้น ไม่ทำให้เห็นความแตกต่างของผลลัพธ์มากนัก ส่วนพารามิเตอร์ที่มีผลต่อการกรองอย่างมากคือ จำนวนคู่โครโมโซมพ่อแม่และจำนวนยีนเครือข่าย ดังนั้น เราจะทำการปรับพารามิเตอร์ทั้งสองตัวอย่างละเอียดอีกครั้ง โดยจะทำการ Fix พารามิเตอร์ตัวอื่นๆ ที่เหลือ โดยจำนวนคู่โครโมโซมพ่อแม่ประกอบด้วย 3, 5, 10, 15 และ 30 เปอร์เซนต์ ส่วนค่ายีนเครือข่ายประกอบด้วย 5, 6 และ 7 ยีน ซึ่งเมื่อทำการทดลองกับข้อมูล 5 ชุดได้ค่า Accuracy, Recall และ Precision ดังตารางที่ 5.7

ตารางที่ 5.7 แสดงค่า Accuracy, Recall และ Precision ของการปรับจำนวนคู่โครโมโซมพ่อแม่
และยีนเรโซว์ที่ค่าต่างๆ

	Parent Pair	Gene Threshold	Accuracy	Recall	Precision
SET 1	3	5	85.00%	87.50%	83.33%
	5	5	85.25%	86.50%	84.39%
	10	5	85.50%	85.50%	85.50%
	15	5	82.00%	88.00%	78.57%
	30	5	84.00%	88.00%	81.48%
	3	6	87.75%	88.00%	87.56%
	5	6	87.00%	86.00%	87.76%
	10	6	81.75%	83.50%	80.68%
	15	6	83.50%	81.50%	84.90%
	30	6	78.50%	96.00%	71.11%
	3	7	78.50%	64.00%	90.14%
	5	7	85.25%	81.00%	88.52%
	10	7	85.75%	82.00%	88.65%
	15	7	86.25%	82.00%	89.62%
	30	7	86.75%	89.00%	85.17%
SET 2	3	5	88.00%	87.50%	88.38%
	5	5	87.00%	86.00%	87.76%
	10	5	87.00%	87.00%	87.00%
	15	5	78.00%	91.00%	72.22%
	30	5	65.50%	91.50%	60.20%
	3	6	90.50%	91.50%	89.71%
	5	6	83.75%	96.00%	77.11%
	10	6	87.75%	87.50%	87.94%
	15	6	82.75%	86.50%	80.47%
	30	6	71.75%	100.00%	63.90%
	3	7	84.75%	76.00%	92.12%
	5	7	87.25%	87.50%	87.06%
	10	7	84.50%	83.00%	85.57%
	15	7	87.00%	89.50%	85.24%
	30	7	83.75%	83.00%	84.26%
SET 3	3	5	87.00%	86.50%	87.37%
	5	5	86.25%	85.00%	87.18%
	10	5	86.50%	85.00%	87.63%
	15	5	84.75%	85.50%	84.24%
	30	5	64.50%	100.00%	58.48%
	3	6	87.50%	85.00%	89.47%
	5	6	83.25%	77.50%	87.57%
	10	6	85.75%	84.50%	86.67%
	15	6	87.00%	86.50%	87.37%
	30	6	84.00%	85.00%	83.33%
	3	7	85.50%	82.00%	88.17%
	5	7	82.75%	73.50%	90.18%
	10	7	85.25%	83.00%	86.91%
	15	7	85.25%	80.50%	88.95%
	30	7	83.75%	83.00%	84.26%

ตาราง 5.7 (ต่อ)

SET 4	3	5	86.00%	87.00%	85.29%
	5	5	84.75%	85.00%	84.58%
	10	5	86.50%	89.50%	84.43%
	15	5	84.75%	84.50%	84.92%
	30	5	86.00%	84.50%	87.11%
	3	6	86.50%	84.50%	88.02%
	5	6	88.00%	88.50%	87.62%
	10	6	74.50%	99.50%	66.33%
	15	6	78.00%	94.00%	71.21%
	30	6	73.00%	88.00%	67.69%
	3	7	82.75%	77.50%	86.59%
	5	7	83.50%	84.00%	83.17%
	10	7	84.25%	79.00%	88.27%
	15	7	83.25%	77.00%	88.00%
	30	7	70.75%	51.00%	84.30%
SET 5	3	5	88.00%	91.00%	85.85%
	5	5	86.75%	84.50%	88.48%
	10	5	87.00%	88.50%	85.92%
	15	5	86.50%	87.00%	86.14%
	30	5	85.25%	85.50%	85.07%
	3	6	87.25%	87.00%	87.44%
	5	6	88.25%	88.50%	88.06%
	10	6	86.25%	83.50%	88.36%
	15	6	87.75%	87.50%	87.94%
	30	6	86.25%	83.50%	88.36%
	3	7	80.25%	68.50%	89.54%
	5	7	86.25%	86.00%	86.43%
	10	7	84.75%	81.00%	87.57%
	15	7	72.75%	51.50%	89.57%
	30	7	82.75%	80.00%	84.66%
Average	3	5	86.80%	87.90%	86.04%
	5	5	86.00%	85.40%	86.48%
	10	5	86.50%	87.10%	86.10%
	15	5	83.20%	87.20%	81.22%
	30	5	77.05%	89.90%	74.47%
	3	6	87.90%	87.20%	88.44%
	5	6	86.05%	87.30%	85.62%
	10	6	83.20%	87.70%	82.00%
	15	6	83.80%	87.20%	82.38%
	30	6	78.70%	90.50%	74.88%
	3	7	82.35%	73.60%	89.31%
	5	7	85.00%	82.40%	87.07%
	10	7	84.90%	81.60%	87.39%
	15	7	82.90%	76.10%	88.28%
	30	7	81.55%	77.20%	84.53%

จากตารางที่ 5.7 จะเห็นว่าจากค่าเฉลี่ยของชุดข้อมูลทั้ง 5 ชุด พบว่าจะให้ผลลัพธ์ที่ดีที่สุดเมื่อปรับจำนวนคู่โครโมโซมพ่อแม่ 3% และเมื่อปรับค่าอินเทอร์โรว์ที่ 6 ขึ้น

5.3 ผลการทดลองการกรองอีเมลขยะโดยใช้เจเนติกอัลกอริทึม

จากการทดลองในหัวข้อ 5.2 จะได้ค่าพารามิเตอร์ที่เหมาะสมที่จะนำมาใช้ในการวิจัย ซึ่งได้แก่

- วิธีการคัดเลือกใช้แบบวงล้อถ่วงน้ำหนัก
- การครอสโอเวอร์ เลือกแบบสุ่ม
- การมิวเตชัน ไม่ใช้พารามิเตอร์นี้
- จำนวนคู่โครโมโซมพ่อแม่ 3%
- ยีนเรโซว์ 6 ยีน

พารามิเตอร์เหล่านี้จะถูกนำมาใช้ในการทดลองประสิทธิภาพของตัวกรอง ซึ่งผลที่ได้เป็นดัง

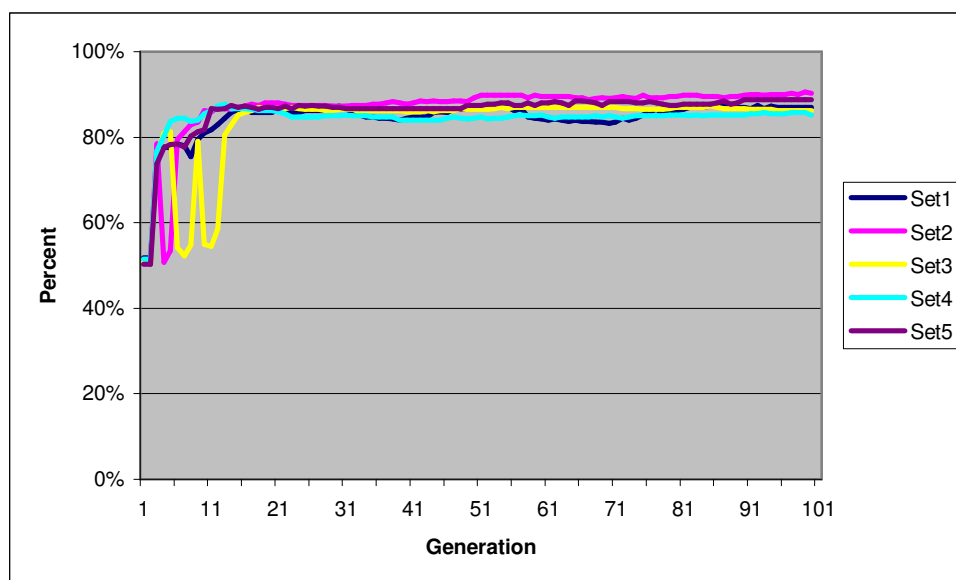
ตาราง 5.8

ตาราง 5.8 แสดงค่าเฉลี่ย Accuracy, Recall และ Precision ของการทดลองในแต่ละเซตและค่าเฉลี่ย

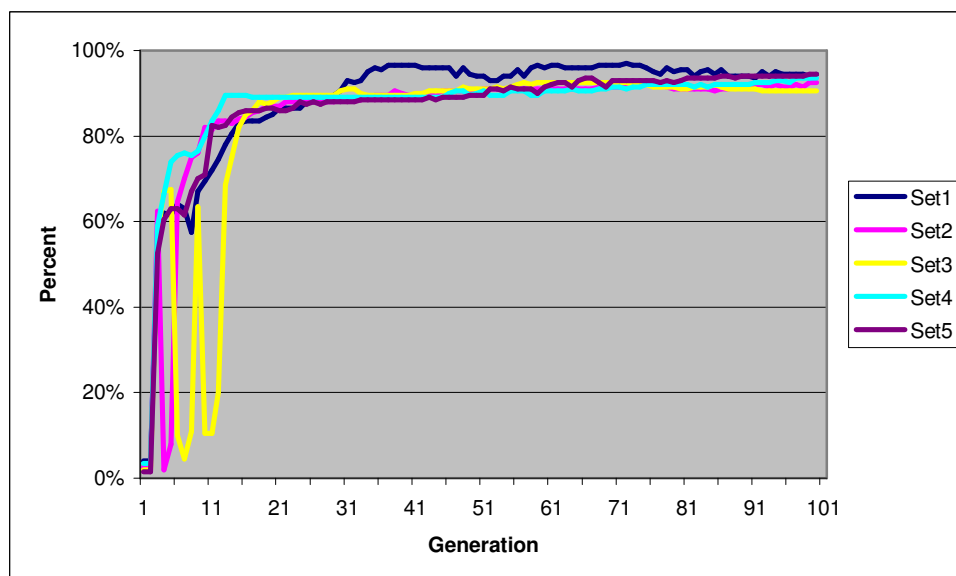
Accuracy, Recall และ Precision ของทุกๆ เซต

		Acc	Recall	Pre	Average of Each Set		
รอบที่ 1	SET1	86.50%	86.50%	86.50%	87.70%	88.20%	87.33%
	SET2	90.25%	92.50%	88.52%			
	SET3	87.00%	88.00%	86.27%			
	SET4	87.25%	86.00%	88.21%			
	SET5	87.50%	88.00%	87.13%			
รอบที่ 2	SET1	87.25%	87.50%	87.06%	87.15%	87.90%	86.60%
	SET2	88.25%	88.50%	88.06%			
	SET3	86.50%	87.50%	85.78%			
	SET4	85.50%	86.00%	85.15%			
	SET5	88.25%	90.00%	86.96%			
รอบที่ 3	SET1	88.50%	89.50%	87.75%	87.55%	87.50%	87.66%
	SET2	87.25%	84.50%	89.42%			
	SET3	86.75%	84.50%	88.48%			
	SET4	86.75%	86.00%	87.31%			
	SET5	88.50%	93.00%	85.32%			
รอบที่ 4	SET1	86.50%	82.00%	90.11%	84.75%	90.20%	82.46%
	SET2	78.75%	96.50%	71.22%			
	SET3	83.25%	96.50%	76.28%			
	SET4	87.75%	88.50%	87.19%			
	SET5	87.50%	87.50%	87.50%			
รอบที่ 5	SET1	88.25%	89.50%	87.32%	88.10%	88.70%	87.68%
	SET2	90.25%	94.50%	87.10%			
	SET3	87.00%	88.00%	86.27%			
	SET4	86.75%	84.00%	88.89%			
	SET5	88.25%	87.50%	88.83%			
Average of All Set					87.05%	88.50%	86.35%

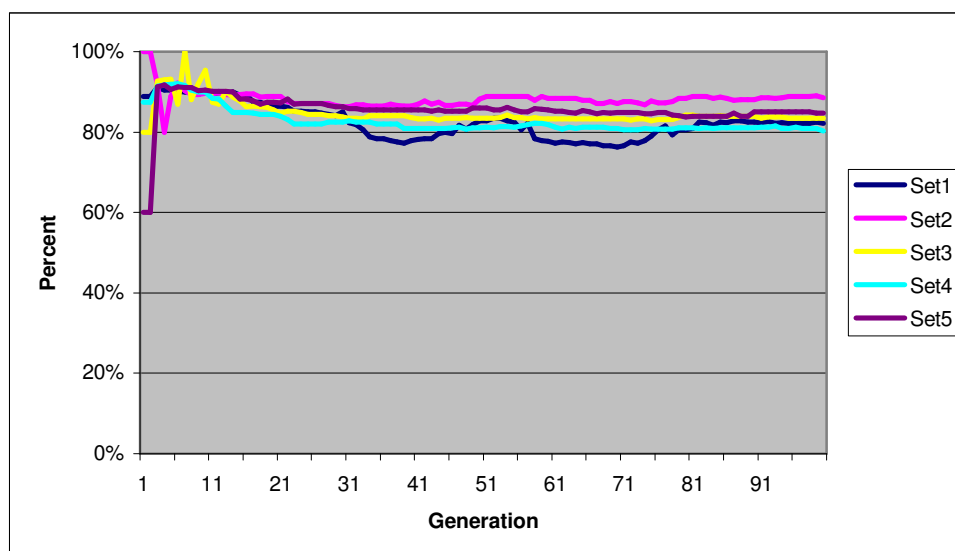
จากตารางที่ 5.8 จะเห็นว่าเมื่อทดลองด้วยพารามิเตอร์ที่เหมาะสมที่ได้จากการทดลองเมื่อปรับค่าอย่างละเอียดขึ้น พบว่าค่าเฉลี่ย Accuracy, Recall และ Precision ของทั้ง 5 เซตให้ผลลัพธ์ที่ดีขึ้น รูปที่ 5.10 ถึงรูปที่ 5.12 เป็นการแสดงค่า Accuracy, Recall และ Precision ของเซตข้อมูลทั้ง 5 เซตในรอบที่ 1 เมื่อใช้พารามิเตอร์ที่เหมาะสมที่ได้จากการทดลองในหัวข้อ 5.2 เพื่อให้เห็นภาพรวมของผลลัพธ์ที่ได้ในแต่ละเซตข้อมูล



รูปที่ 5.10 แสดงค่า Accuracy ของเซตข้อมูลทั้ง 5 เซตเมื่อใช้พารามิเตอร์ที่ได้จากการทดลอง



รูปที่ 5.11 แสดงค่า Recall ของเซตข้อมูลทั้ง 5 เซตเมื่อใช้พารามิเตอร์ที่ได้จากการทดลอง



รูปที่ 5.12 แสดงค่า Precision ของเซตข้อมูลทั้ง 5 เซตเมื่อใช้พารามิเตอร์ที่ได้จากการทดลอง

จากรูปที่ 5.10 ถึงรูปที่ 5.12 จะเห็นว่าค่า Accuracy, Recall และ Precision ของเซตข้อมูลทั้ง 5 จะมีการเปลี่ยนแปลงอย่างมากในช่วง Generation แรกๆ และมีการเปลี่ยนแปลงน้อยมากในช่วง Generation ใหม่ๆ ซึ่งทำให้ค่า Accuracy, Recall และ Precision ของชุดข้อมูลแต่ละเซตจะเริ่มคงที่มากขึ้น ซึ่งเป็นรูปแบบทั่วไปของกระบวนการเจเนติกอัลกอริทึม และให้ค่า Accuracy, Recall และ Precision ที่ดี

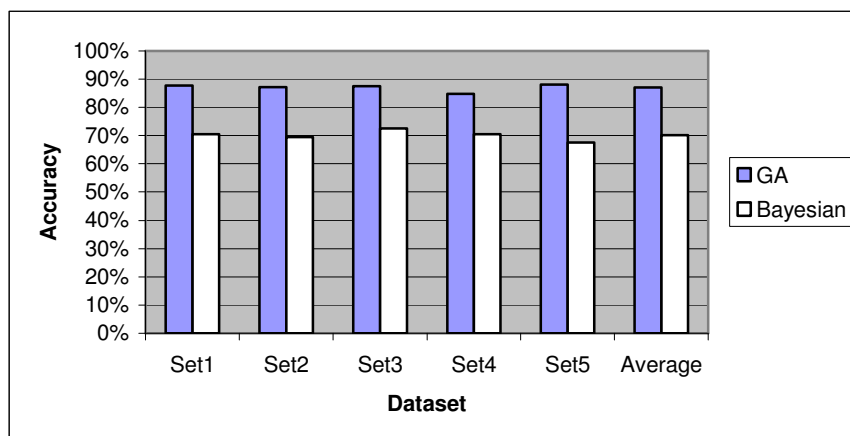
5.4 เปรียบเทียบตัวกรองอีเมลขยะที่นำเสนอด้วยตัวกรองอีเมลขยะซึ่งประยุกต์ใช้ทฤษฎีเบย์เซียน

การเปรียบเทียบค่า Accuracy, Recall และ Precision ที่ได้ระหว่างตัวกรองอีเมลขยะโดยใช้เจเนติกอัลกอริทึมและผลการทดลองที่ได้ระหว่างตัวกรองอีเมลขยะโดยใช้การประยุกต์ใช้ทฤษฎีเบย์เซียนมาเปรียบเทียบกันดังแสดงในตารางที่ 5.9

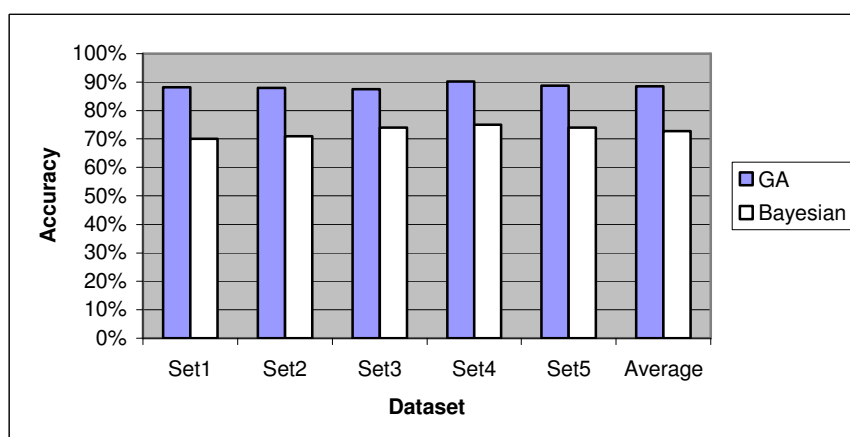
ตารางที่ 5.9 แสดงค่า Accuracy, Recall และ Precision ของเซตข้อมูลทั้ง 5 เซต

	Genetic Algorithm			Bayesian		
	Accuracy	Recall	Precision	Accuracy	Recall	Precision
Set1	87.70%	88.20%	87.33%	70.50%	70.00%	70.71%
Set2	87.15%	87.90%	86.60%	69.50%	71.00%	68.93%
Set3	87.55%	87.50%	87.66%	72.50%	74.00%	71.84%
Set4	84.75%	90.20%	82.46%	70.50%	75.00%	68.81%
Set5	88.10%	88.70%	87.68%	67.50%	74.00%	65.49%
Average	87.05%	88.50%	86.35%	70.10%	72.80%	69.16%

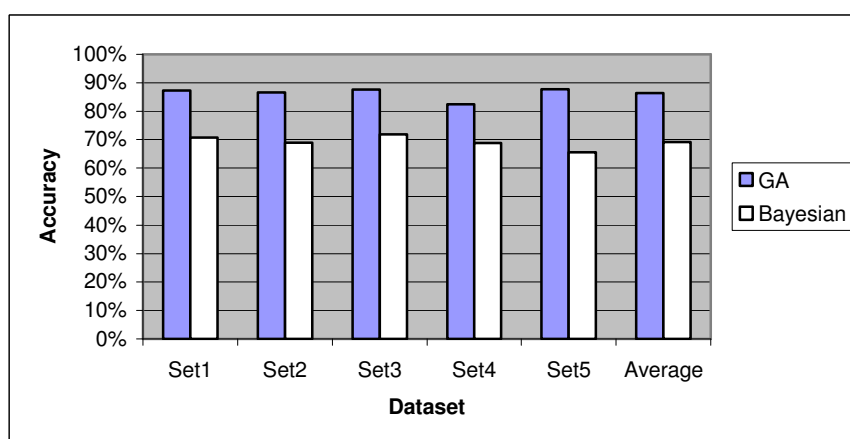
จากตารางที่ 5.9 พบว่าเมื่อนำค่า Accuracy, Recall และ Precision ของวิธีการที่นำเสนอ กับวิธีการกรองอีเมลขยะโดยใช้การประยุกต์ใช้ทฤษฎีเบย์เซียนมาเปรียบเทียบกันทีละค่าจะได้กราฟดังรูปที่ 5.13 ถึงรูปที่ 5.15



รูปที่ 5.13 แสดงการเปรียบเทียบค่า Accuracy ระหว่างเจเนติกอัลกอริทึมและเบย์เซียน



รูปที่ 5.14 แสดงการเปรียบเทียบค่า Recall ระหว่างเจเนติกอัลกอริทึมและเบย์เซียน



รูปที่ 5.15 แสดงการเปรียบเทียบค่า Precision ระหว่างเจเนติกอัลกอริทึมและเบย์เซียน

จากรูปที่ 5.13 ถึงรูปที่ 5.15 พบว่าค่า Accuracy, Recall และ Precision ของเจเนติกอัลกอริทึม ดีกว่าค่า Accuracy, Recall และ Precision ของเบย์เซียนในทุกชุดข้อมูลที่น่ามาทดลอง ทั้งนี้ เนื่องจากตัวกรองอีเมลขยะซึ่งประยุกต์ใช้ทฤษฎีเบย์เซียนนั้นจะเก็บสถิติของทั้งคำดีและคำขยะไว้ในฐานข้อมูลคำ เมื่อมีอีเมลทดสอบเข้ามา ตัวกรองจะนำคำในอีเมลมาแยกเป็นคำดีและคำขยะ จากนั้นเปรียบเทียบระหว่างความน่าจะเป็นของคำที่เป็นคำขยะและความน่าจะเป็นของคำที่เป็นคำดี ถ้าความน่าจะเป็นของคำประเภทใดมากกว่า ก็จะทำนายว่าอีเมลทดสอบเป็นอีเมลประเภทนั้น ซึ่งจะเห็นว่าเป็นการเปรียบเทียบเพียงแค่ครั้งเดียว ในขณะที่ตัวกรองอีเมลขยะที่สร้างจากเจเนติกอัลกอริทึมจะมีการดำเนินการทางเจเนติกไปหลายๆรุ่น เพื่อหารูปแบบที่เหมาะสมเพื่อจะนำรูปแบบเหล่านี้มาใช้ในการกรองอีเมลขยะ ทำให้ประสิทธิภาพของการกรองโดยใช้ตัวกรองอีเมลขยะโดยใช้เจเนติกอัลกอริทึมดีกว่าประสิทธิภาพของตัวกรองอีเมลขยะตามทฤษฎีเบย์เซียน