

การรู้จำอักขรตัวพิมพ์ภาษาไทยโดยใช้พีชคณิตและราฟเซต

PRINTED THAI CHARACTERS RECOGNITION  
USING FUZZY AND ROUGH SET THEORY

ธีรราช สืบสนิท  
THIRAT SUEBSANIT

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2546

ISBN 974-324-652-5

สำนักหอสมุดกลาง พระจอมเกล้าลาดกระบัง

การรู้จำอักษรตัวพิมพ์ภาษาไทยโดยใช้ฟัซซี่เซตและราฟเซต

PRINTED THAI CHARACTERS RECOGNITION  
USING FUZZY AND ROUGH SET THEORY



ธีราช สืบสนิท

THIRAT SUEBSANIT

เลขหม.....  
เลขทะเบียน 47715  
วัน, เดือน, ปี 2.2.๕๕ 2546

.b.....  
.i.....

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ.2546

ISBN 974-324-652-5

**COPYRIGHT 2003**

**SCHOOL OF GRADUATE STUDIES**

**KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG**

บัณฑิตวิทยาลัย  
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง  
ใบรับรองวิทยานิพนธ์

---

หัวข้อวิทยานิพนธ์      การรู้จำอักขรตัวพิมพ์ภาษาไทยโดยใช้ฟัซซี่เซตและราฟเซต  
PRINTED THAI CHARACTERS RECOGNITION USING FUZZY  
AND ROUGH SET THEORY

ชื่อนักศึกษา              นายธีราช              สืบสนิท


รหัสประจำตัว              43061626

ปริญญา                      วิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชา                    วิศวกรรมคอมพิวเตอร์

อาจารย์ผู้ควบคุมวิทยานิพนธ์      รศ.ดร.เอื้อน              ปิ่นเงิน

อาจารย์ผู้ควบคุมวิทยานิพนธ์ร่วม      รศ.ดร.ชม                  กิมปาน

คณะกรรมการสอบวิทยานิพนธ์		ลายมือชื่อ
รศ.ดร.บุญวัฒน์	อัครชู	
รศ.ดร.บุญธีร์	เกรือตราชู	
ดร.วัชระ	ฉัตรวิริยะ	
รศ.ดร.ชม	กิมปาน	
รศ.ดร.เอื้อน	ปิ่นเงิน	

วัน/เดือน/ปี ที่สอบ 26 พฤษภาคม 2546 เวลา 14.00-16.00 น.

สถานที่สอบ ณ อาคาร 12 ชั้น ชั้น 4 (ห้อง E12-403)

  
บัณฑิตวิทยาลัยรับรองแล้ว  
(รศ.ดร.บุญวัฒน์ อัครชู)  
คณบดีบัณฑิตวิทยาลัย

วันที่...30...เดือน...พฤษภาคม...พ.ศ...2546...

หัวข้อวิทยานิพนธ์	การรู้จำอักษรตัวพิมพ์ภาษาไทยโดยใช้ฟัซซี่เซตและกราฟเซต
นักศึกษา	นายธีราช สืบสนิท
รหัสประจำตัว	43061626
ปริญญา	วิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
พ.ศ.	2546
อาจารย์ผู้ควบคุมวิทยานิพนธ์	รศ.ดร.เอื้อน ปิ่นเงิน
อาจารย์ผู้ควบคุมวิทยานิพนธ์ร่วม	รศ.ดร.ชม กิมปาน

### บทคัดย่อ

งานวิจัยนี้เสนอระบบการรู้จำอักษรตัวพิมพ์ภาษาไทยโดยใช้ฟัซซี่เซตและกราฟเซต ซึ่งฟัซซี่เซตและกราฟเซตจะช่วยจัดการกับความคลุมเครือและความไม่แน่นอนของข้อมูลในระบบปัญญาประดิษฐ์ ระบบการรู้จำตัวอักษรนี้ได้ประยุกต์ใช้ฟัซซี่เซตและกราฟเซตในการสร้างกฎการตัดสินใจสำหรับการจำแนกตัวอักษร ในส่วนของกระบวนการการรู้จำแบ่งเป็นสี่ขั้นตอนคือ ขั้นตอนแรกหาลักษณะเด่นของตัวอักษร เพื่อนำไปสร้างข้อมูลสำหรับขั้นตอนถัดไป ขั้นตอนที่สองแบ่งช่วงลักษณะเด่นที่มีความต่อเนื่องของข้อมูลโดยใช้ฟัซซี่เซต ขั้นตอนที่สามนำเอาทริบิวของข้อมูลที่ได้มาลดข้อมูลที่ไม่จำเป็นสำหรับการอธิบายลักษณะของตัวอักษร แต่จะยังคงความสามารถในการรู้จำและขั้นตอนสุดท้ายสร้างกฎสำหรับการจำแนกตัวอักษร จากการทดสอบความถูกต้องของการรู้จำ ระบบการรู้จำนี้สามารถจำแนกตัวอักษรได้ถูกต้อง 96.34%

<b>Thesis Title</b>	Printed Thai Characters Recognition using Fuzzy and Rough Set Theory
<b>Student</b>	Mr.Thirat Suebsanit
<b>Student ID.</b>	43061626
<b>Degree</b>	Master of Engineering
<b>Programme</b>	Computer Engineering
<b>Year</b>	2003
<b>Thesis Advisor</b>	Assoc. Prof. Dr. Ouen Pinngern
<b>Thesis Co-Advisor</b>	Assoc. Prof. Dr. Chom Kimpan

## ABSTRACT

The printed Thai characters recognition system that implemented using Fuzzy and Rough sets is proposed. Both Fuzzy and Rough sets have been introduced as tools to deal with vagueness and uncertain data in artificial intelligence applications. For the recognition system, we use Fuzzy and Rough to create a rule-based for classify the characters. The recognition process consists of four stages. First, we find the features of printed Thai characters to create the data for the next stage. Second, discrete continuous data to subintervals. Third, reduce the redundant input attributes of data and all the while minimizing information loss for the recognition system. Finally, recognize the printed Thai characters. From the testing results, the average of the recognition rate is 96.34% accuracy.

## กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จลุล่วงได้อย่างดีด้วยคำแนะนำและคำปรึกษาเกี่ยวกับการรู้จำอักษรตัวพิมพ์ภาษาไทยโดยใช้พีชชีเซตและราฟเซต จาก รศ.ดร.เอื้อน ปิ่นเงิน และ รศ.ดร. ชม กิมปาน ซึ่งเป็นอาจารย์ผู้ควบคุมวิทยานิพนธ์ ผู้วิจัยรู้สึกเคารพ และซาบซึ้งในความกรุณาเป็นอย่างยิ่ง และขอกราบขอบพระคุณเป็นอย่างสูง

ขอขอบพระคุณ ผศ.ดร.พิศิษฐ์ โภคารัตน์กุล ภาควิชาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ มหาวิทยาลัยมหิดล ที่คอยช่วยเหลือ และให้คำปรึกษาเกี่ยวกับงานวิจัย

ขอขอบพระคุณ ผศ.ไพโรบลย์ พันธรัญพงษ์ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ ที่คอยให้ความช่วยเหลือทางด้านงานวิจัย อุปกรณ์และคำปรึกษาในเรื่องการทำงาน และเรื่องต่างๆ ไป

ขอขอบพระคุณ อาจารย์วิสันต์ ตั้งวงษ์เจริญ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ ที่คอยให้คำปรึกษาในเรื่องการทำงาน และเรื่องต่างๆ ไป

ขอขอบพระคุณ พี่อุกฤษณ์ มารังค์ สำหรับข้อมูลต่างๆ ที่เกี่ยวกับงานวิจัย รวมถึงขอขอบพระคุณพี่ๆ ห้องวิจัย Information Science ทุกคน ที่ได้ให้ความช่วยเหลือ ถ่ายทอดประสบการณ์ต่างๆ และเป็นกำลังใจให้ผู้วิจัยตลอดมา

ขอขอบพระคุณ คุณพ่อคุณแม่ และน้องสาว รวมถึง เธอ ที่เป็นกำลังใจให้ผู้วิจัยในทุกๆ เรื่องอย่างใกล้ชิดตลอดมา

คุณค่าและประโยชน์อันพึงมีจากวิทยานิพนธ์ฉบับนี้ ผู้วิจัยขอมอบแด่ผู้มีพระคุณทุกท่าน

ธีราช สืบสนิท

# สารบัญ

	หน้า
บทคัดย่อภาษาไทย .....	I
บทคัดย่อภาษาอังกฤษ .....	II
กิตติกรรมประกาศ .....	III
สารบัญ .....	IV
สารบัญตาราง .....	VII
สารบัญรูป .....	IX
บทที่ 1 บทนำ .....	1
1.1 ความเป็นมาและความสำคัญของปัญหา .....	1
1.2 ความมุ่งหมายและวัตถุประสงค์ของการวิจัย .....	1
1.3 งานวิจัยที่เกี่ยวข้อง .....	2
1.3 ขอบเขตของการวิจัย .....	4
บทที่ 2 ฟิชชีเซต .....	5
2.1 ทฤษฎีฟิชชีเซต .....	5
2.2 การแทนข้อมูลในฟิชชีเซต .....	5
2.3 ฟังก์ชันความเป็นสมาชิก .....	7
2.4 ระบบการจัดกลุ่มแบบฟิชชี .....	13
2.4.1 การแบ่งกลุ่มข้อมูลที่ไม่เป็นฟิชชี .....	14
2.4.2 การแบ่งกลุ่มข้อมูล โดยวิธีฟิชชี .....	16
2.5 สรุป .....	19
บทที่ 3 ราฟเซต .....	20
3.1 ตัวอย่าง .....	21
3.2 คำจำกัดความของ Approximations .....	22
3.3 ค่าฟังก์ชันความเป็นสมาชิกของราฟเซต .....	24
3.4 การขึ้นต่อกันของแอดทริบิว .....	24

## สารบัญ (ต่อ)

	หน้า
3.5 การลดแอดทริบิวที่ไม่จำเป็น .....	26
3.5.1 ค่าความสำคัญของแอดทริบิว .....	26
3.5.2 การหา reduct ที่ดีที่สุด .....	26
3.5.3 Discernibility Matrix .....	27
3.5.4 Core และ Discernibility Matrix .....	27
3.5.5 อัลกอริทึมสำหรับคำนวณ reduct ที่ดีที่สุด หรือเซตของแอดทริบิวที่น้อยที่สุด .....	28
3.6 กฎการตัดสินใจ .....	29
3.7 สรุป .....	30
บทที่ 4 การประยุกต์พีชชีเซต และกราฟเซตร่วมกัน .....	31
4.1 การจัดกลุ่มข้อมูลโดยพีชชีเซต .....	31
4.2 การจำแนกโดยกราฟเซต .....	34
4.2.1 การหา core จาก Discernibility Matrix .....	35
4.2.2 reduct ที่ดีที่สุด หรือเซตของแอดทริบิวที่น้อยที่สุด .....	35
4.2.3 กฎการตัดสินใจ .....	37
4.3 สรุป .....	37
บทที่ 5 การเตรียมข้อมูล และกระบวนการรู้จำ .....	38
5.1 การหาลักษณะเด่นของตัวอักษร .....	38
5.1.1 การหาจุดศูนย์กลางของภาพ .....	38
5.1.2 การคำนวณค่าลักษณะขอบภาพรอบนอก .....	39
5.1.2.1 การแบ่งพื้นที่โดยอาศัยเส้นแบ่งในแนวนอน .....	39
5.1.2.2 การแบ่งพื้นที่โดยอาศัยเส้นแบ่งในแนวตั้ง .....	41
5.1.3 การคำนวณค่าความหนาแน่นของจุดตัด .....	43
5.1.3.1 การแบ่งพื้นที่เพื่อคำนวณค่าความหนาแน่นของจุดตัด โดยอาศัยเส้นแบ่งในแนวนอน .....	43

## สารบัญ (ต่อ)

	หน้า
5.1.3.2 การแบ่งพื้นที่เพื่อคำนวณค่าความหนาแน่นของจุดตัด โดยอาศัยเส้นแบ่งในแนวตั้งฉาก .....	44
5.1.4 การคำนวณสัดส่วนความกว้างต่อความสูง .....	45
5.1.5 การนับจำนวนของการเปลี่ยนจากพื้นหลังภาพไปเป็นเนื้อภาพ .....	46
5.1.6 การหาลักษณะเด่นด้วยการเข้ารหัสลูกโซ่ (Chain code) .....	47
5.1.6.1 การคำนวณค่าตัวเลขต่อเนื่อง (Connectivity number) .....	47
5.1.6.2 การหาขอบภาพ (Edge detection) .....	50
5.1.6.3 การหาลักษณะเด่นของตัวอักษรด้วยรหัสลูกโซ่ (Chain Code) .....	51
5.2 การประยุกต์ใช้พีชคณิตและกราฟเซตร่วมกัน .....	53
5.3 กฎการตัดสินใจ .....	62
บทที่ 6 การทดลองและผลการทดลอง .....	64
6.1 การเตรียมข้อมูลสำหรับการทดลอง .....	64
6.2 ขั้นตอนการทดลอง .....	64
6.2 ผลการทดลอง .....	65
บทที่ 7 สรุปและข้อเสนอแนะ .....	84
7.1 สรุป .....	84
7.2 ข้อเสนอแนะ .....	85
เอกสารอ้างอิง .....	86
ภาคผนวก ก. ผลงานวิจัยที่ได้รับการตีพิมพ์ .....	88
ประวัติผู้เขียน .....	89

# สารบัญตาราง

ตารางที่	หน้า
3.1 ตัวอย่างของตารางข้อมูล .....	21
3.2 แสดง Discernibility Matrix .....	27
4.1 ผลการประเมินจำนวนกลุ่ม .....	34
5.1 ตารางแสดงความสัมพันธ์ระหว่างจุดภาพ .....	49
5.2 แสดงตัวอย่างข้อมูลแอตทริบิวต์ .....	53
5.3 แสดงการแบ่งกลุ่มสำหรับแต่ละแอตทริบิวต์ .....	54
5.4 แสดงตัวอย่างผลที่ได้จากการแบ่งกลุ่มโดยพีชชีเซต .....	56
5.3 แสดงแอตทริบิวต์ core .....	57
5.4 แสดงแอตทริบิวต์ reduct .....	57
6.1 แสดงผลการทดสอบความถูกต้องของฟอนต์ AngsanaUPC .....	65
6.2 ตัวอย่างแอตทริบิวต์ลักษณะเด่นของตัวอักษร ช ฟอนต์ AngsanaUPC ที่จำแนกผิดพลาด .....	67
6.3 ตัวอย่างแอตทริบิวต์ลักษณะเด่นของตัวอักษร ฎ ฟอนต์ AngsanaUPC ที่จำแนกผิดพลาด .....	68
6.4 แสดงผลการทดสอบความถูกต้องของฟอนต์ BrowalliaUPC .....	68
6.5 ตัวอย่างแอตทริบิวต์ลักษณะเด่นของตัวอักษร ต ฟอนต์ BrowalliaUPC ที่จำแนกผิดพลาด .....	70
6.6 ตัวอย่างแอตทริบิวต์ลักษณะเด่นของตัวอักษร ร ฟอนต์ BrowalliaUPC ที่จำแนกผิดพลาด .....	71
6.7 ตัวอย่างแอตทริบิวต์ลักษณะเด่นของตัวอักษร ส ฟอนต์ BrowalliaUPC ที่จำแนกผิดพลาด .....	71
6.8 แสดงผลการทดสอบความถูกต้องของฟอนต์ CordiaUPC .....	71
6.9 ตัวอย่างแอตทริบิวต์ลักษณะเด่นของตัวอักษร ช ฟอนต์ CordiaUPC ที่จำแนกผิดพลาด .....	74
6.10 ตัวอย่างแอตทริบิวต์ลักษณะเด่นของตัวอักษร ต ฟอนต์ CordiaUPC ที่จำแนกผิดพลาด .....	74
6.11 ตัวอย่างแอตทริบิวต์ลักษณะเด่นของตัวอักษร ฎ ฟอนต์ CordiaUPC ที่จำแนกผิดพลาด .....	74
6.12 ตัวอย่างแอตทริบิวต์ลักษณะเด่นของตัวอักษร ส ฟอนต์ CordiaUPC ที่จำแนกผิดพลาด .....	75
6.13 แสดงผลการทดสอบความถูกต้องของฟอนต์ FreesiaUPC .....	75
6.14 ตัวอย่างแอตทริบิวต์ลักษณะเด่นของตัวอักษร ช ฟอนต์ FreesiaUPC ที่จำแนกผิดพลาด .....	77
6.15 ตัวอย่างแอตทริบิวต์ลักษณะเด่นของตัวอักษร ค ฟอนต์ FreesiaUPC ที่จำแนกผิดพลาด .....	78
6.16 แสดงผลการทดสอบความถูกต้องของฟอนต์ DilleniaUPC .....	78
6.17 ตัวอย่างแอตทริบิวต์ลักษณะเด่นของตัวอักษร ข ฟอนต์ DilleniaUPC ที่จำแนกผิดพลาด .....	81
6.18 ตัวอย่างแอตทริบิวต์ลักษณะเด่นของตัวอักษร ศ ฟอนต์ DilleniaUPC ที่จำแนกผิดพลาด .....	81
6.19 ตัวอย่างแอตทริบิวต์ลักษณะเด่นของตัวอักษร ฎ ฟอนต์ DilleniaUPC ที่จำแนกผิดพลาด .....	82

## สารบัญตาราง (ต่อ)

ตารางที่	หน้า
6.20 ตัวอย่างแอตทริบิวต์ลักษณะเด่นของตัวอักษร ฎ ฟอนต์ DilleniaUPC ที่จำแนกผิดพลาด.....	82
6.21 ตัวอย่างแอตทริบิวต์ลักษณะเด่นของตัวอักษร ภ ฟอนต์ DilleniaUPC ที่จำแนกผิดพลาด.....	82
6.22 ผลสรุปการทดสอบความถูกต้อง .....	83
6.23 ผลสรุปการทดสอบความถูกต้องของฟอนต์ DilleniaUPC .....	83

# สารบัญรูป

รูปที่	หน้า
2.1 ค่าความเป็นสมาชิกของเซต A สำหรับกรณีที่เป็นเซตธรรมดา	6
2.2 ค่าความเป็นสมาชิกของเซต A สำหรับกรณีที่เป็นฟuzzyเซต	7
2.3 กราฟฟังก์ชันความเป็นสมาชิกของตัวแปร "ความสูง" ที่ประกอบไปด้วย 3 เทอมเซต	8
2.4 ภาพฟuzzyรูปสามเหลี่ยม	9
2.5 ภาพฟuzzyรูปสี่เหลี่ยมคางหมู	10
2.6 เส้นโค้งรูปตัว S	11
2.7 ค่าฟuzzyเซตรูปประฆังคว่ำ	12
2.8 ค่าฟังก์ชัน Gaussian	13
2.9 กลุ่มข้อมูล [X] มีสมาชิก $\{x_1, x_2, x_3, \dots, x_n\}$	14
2.10 การแบ่งข้อมูล [X] เป็นกลุ่ม $A_i; i = 1, 2, \dots, c$	15
2.11 ข้อมูล [X] จำนวน 15 จุด	18
2.12 ผลของการจำแนกด้วยอัลกอริทึม FCM	18
4.1 ตัวอย่างข้อมูลของแอดทรีที่มีความต่อเนื่อง	32
4.2 กราฟค่าความเป็นสมาชิกของ U	33
5.1 จุดศูนย์กลางของภาพสองระดับ	39
5.2 การแบ่งพื้นที่โดยอาศัยเส้นแบ่งในแนวนอน	40
5.3 ระยะทางในแนวนอนจากกรอบภาพทางด้านซ้ายถึงขอบภาพด้านนอกของบริเวณ $S_h^{-1}$	40
5.4 การแบ่งพื้นที่โดยอาศัยเส้นแบ่งในแนวตั้งฉาก	41
5.5 ระยะทางในแนวตั้งฉากจากกรอบภาพทางด้านบนถึงขอบภาพด้านนอกของบริเวณ $S_v^{-1}$	42
5.6 ความหนาแน่นของการเปลี่ยนจากพื้นหลังภาพไปเป็นเนื้อภาพ โดยการแบ่งในแนวนอน	43
5.7 ความหนาแน่นของการเปลี่ยนจากพื้นหลังภาพไปเป็นเนื้อภาพ โดยการแบ่งในแนวตั้งฉาก	44
5.8 ความกว้างและความสูงของตัวอักษร	45
5.9 เส้นที่ลากผ่านจุดศูนย์กลางของภาพ	46
5.10 การแบ่งพื้นที่ของภาพเป็น 9 บริเวณเท่ากัน	47
5.11 ลักษณะของจุดภาพข้างเคียง	48
5.12 ค่า NC4 และ NC8 ของจุดภาพ	49
5.13 ตัวอย่างการคำนวณค่า NC4 และ NC8	50

## สารบัญรูป (ต่อ)

รูปที่	หน้า
5.14 ค่าตัวเลขต่อเนื่องที่เป็นขอบภาพตัวอักษร “ง” .....	51
5.15 ทิศทางของรหัสลูกโซ่ .....	51
5.16 ทิศรหัสลูกโซ่และการเข้ารหัสลูกโซ่ (Chain code) ตัวอักษร “ง” .....	52
5.17 แสดงการแบ่งพื้นที่รหัสลูกโซ่ของตัวอักษร “ง” .....	52
5.18 การจำลองกรอบภาพตัวอักษร และการแบ่งพื้นที่สำหรับหาลักษณะเด่น .....	59
5.19 ตัวอย่างอักษรต่างฟอนต์ ที่มีความแตกต่างกันของรูปร่าง .....	61

# บทที่ 1

## บทนำ

### 1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันนี้ ได้มีงานวิจัยที่พัฒนาระบบการรู้จำอักษรภาษาไทยโดยใช้วิธีต่างๆ เพื่อให้คอมพิวเตอร์สามารถวิเคราะห์ข้อมูลภาพตัวอักษร และแสดงผลจากการวิเคราะห์ให้อยู่ในรูปของ text file ที่สามารถนำไปใช้กับ word processor หรือประยุกต์ใช้กับงานอื่นๆ ได้

ทฤษฎีกราฟเซต สามารถนำมาวิเคราะห์ความคลุมเครือของข้อมูล การหาขอบเขตบน และขอบเขตล่างของกลุ่มข้อมูล เพื่อนำมาคำนวณค่าการขึ้นต่อกันของข้อมูล เราสามารถนำค่าการขึ้นต่อกันนี้มาวิเคราะห์เพื่อเลือกส่วนของข้อมูลที่มีความจำเป็นต่อการจำแนกกลุ่ม งานวิจัยที่เกี่ยวกับกราฟเซตได้มีการพัฒนาในหลายสาขา เช่น ทางการแพทย์และเภสัชศาสตร์ได้นำกราฟเซตไปใช้ในการวินิจฉัยโรค ทางการเงินและการธนาคารได้นำกราฟเซตไปวิเคราะห์ความเสี่ยงทางการเงินหรือการลงทุน หรือแม้แต่ทางด้านวิศวกรรม ก็ได้มีการนำกราฟเซตไปใช้ในระบบควบคุมต่างๆ ซึ่งงานวิจัยนี้จะได้นำกราฟเซตมาใช้ในการเรียนรู้เพื่อจำแนกตัวอักษร และได้ใช้ฟuzzyเซต ที่สามารถระบุถึงการเป็นสมาชิกโดยอ้างจากค่าความเป็นสมาชิกซึ่งอยู่ระหว่าง 0 ถึง 1 ทำให้เกิดความยืดหยุ่นในการนำไปใช้งานต่างๆ มากมาย เช่น การนำไปใช้ในระบบควบคุม การประมวลผลภาพ โดยงานวิจัยนี้ได้นำฟuzzyเซตมาใช้ในการแบ่งกลุ่มของข้อมูลที่มีความต่อเนื่อง เพื่อแสดงลักษณะของข้อมูลได้มากขึ้น

### 1.2 ความมุ่งหมายและวัตถุประสงค์ของการวิจัย

งานวิจัยนี้ได้เสนอการรู้จำตัวอักษรตัวพิมพ์ภาษาไทยโดยใช้ฟuzzyเซตและกราฟเซต เนื่องจากการหาลักษณะเด่นของข้อมูล เพื่อนำไปสู่กระบวนการในการรู้จำ มีวิธีในการหาหลายวิธี แต่ละวิธีก็มีจุดมุ่งหมายในการวิเคราะห์โครงสร้างของข้อมูลภาพตัวอักษรที่แตกต่างกัน ข้อมูลที่ได้มาจากลักษณะเด่นอาจจะมีส่วนที่ไม่จำเป็นต่อการรู้จำ ดังนั้นในงานวิจัยนี้จึงศึกษาวิธีการเลือกส่วนของข้อมูลลักษณะเด่นที่จำเป็นต่อการระบุตัวอักษรแต่ละตัว โดยใช้ทฤษฎีกราฟเซตในการลดแอดทริบิวหรือข้อมูลลักษณะเด่นที่ไม่จำเป็นต่อการรู้จำ โดยยังคงแอดทริบิวที่มีความสามารถในการจำแนกและมีความสามารถในการรู้จำ และแต่ละแอดทริบิวที่จะนำไปวิเคราะห์ด้วยกราฟเซต จะเป็นข้อมูลที่ไม่ต่อเนื่อง แต่ข้อมูลลักษณะเด่นบางอย่างที่นำมาวิเคราะห์อาจจะมีความต่อเนื่องของข้อมูล จึงได้นำการจัดกลุ่มข้อมูลโดยฟuzzyเซตมาแบ่งช่วงของข้อมูลที่มีความต่อเนื่อง เพื่อให้เกิดการแบ่งช่วงของข้อมูลที่เหมาะสมที่สุด การทำงานจะแบ่งออกเป็นสี่ส่วนคือ การหาลักษณะเด่นของตัวอักษร

โดยวิธีต่างๆ ส่วนที่สองคือการจัดกลุ่มข้อมูลที่มีความต่อเนื่อง และส่วนที่สามคือการวิเคราะห์ลักษณะเด่นของข้อมูลตัวอักษรเพื่อหาลักษณะเด่นที่มีความสามารถในการจำแนก และสุดท้ายสร้างกฎสำหรับการจำแนกตัวอักษร

### 1.3 งานวิจัยที่เกี่ยวข้อง

การวิจัยที่เกี่ยวข้องกับการรู้จำอักษรตัวพิมพ์ภาษาไทย ได้มีการวิจัยอย่างต่อเนื่องโดยใช้วิธีต่างๆ เพื่อที่จะทำให้คอมพิวเตอร์สามารถวิเคราะห์ข้อมูลภาพตัวอักษรได้ และจากการที่มีวิธีในการหาลักษณะเด่นหลายวิธี จึงได้ศึกษางานวิจัยเกี่ยวกับพีชชีเซตและกราฟเซต เพื่อนำมาประยุกต์มาใช้ในระบบการรู้จำตัวอักษร เพื่อเลือกลักษณะเด่นที่มีความจำเป็นต่อการจำแนก โดยมีบทความวิจัยที่เกี่ยวข้อง ดังนี้

- Recognition of Printed Thai Character Recognition using a matching method [1] งานวิจัยนี้ได้้นำวิธีการจดจำแบบการซ้อนทับ ซึ่งจะแบ่งกลุ่มตัวอักษรโดยวัดค่าสัมประสิทธิ์ของความเหมือนแต่ละรูปแบบ แล้วใช้การกระจายแบบคาร์ยูเนนโลบหารูปแบบมาตรฐานสำหรับเป็นตัวแทนแต่ละกลุ่ม และขั้นตอนการแยกตัวอักษรออกจากกลุ่ม จะใช้วิธีการซ้อนทับเป็นส่วนๆ เพื่อแยกอักษรออกจากกัน
- Printed Thai Character Recognition using topological properties method [2] งานวิจัยนี้ได้นำตัวอักษรมาทำให้บาง และหลังจากนั้นนำมาแบ่งกลุ่มตัวอักษร เทคนิคการทำตัวอักษรให้บางจะถูกนำมาใช้สำหรับกำจัดสัญญาณรบกวน และส่วนที่ปลีกย่อยที่ไม่จำเป็น และได้แบ่งกลุ่มตัวอักษรโดยใช้ฟังก์ชันการตัดสินใจแบบเชิงเส้น และในการแยกตัวอักษรออกจากกลุ่มจะใช้ค่าความสัมพันธ์ระหว่างจุดของแต่ละจุดในรูปร่างในการวิเคราะห์
- Printed Thai Character Recognition using Multifeature and Multilevel Classification คุณภาพผนวก ก. เพิ่มเติม งานวิจัยนี้ได้้นำการหาลักษณะเด่นได้แก่ การหาค่าลักษณะขอบนอกของตัวอักษร และความหนาแน่นของจุดตัดภายในตัวอักษร มาจัดกลุ่มหยาบโดยใช้ K-means algorithm ข้อมูลตัวอักษรที่มีความใกล้เคียงกันจะอยู่ในกลุ่มเดียวกัน และเมื่อพิจารณาในแต่ละกลุ่มจะประกอบด้วยข้อมูลตัวอักษรหลายชนิด หลังจากนั้นนำข้อมูลตัวอักษรที่อยู่ในกลุ่มไปเรียนรู้โดยนิวรอลเน็ตเวิร์ค เพื่อจำแนกชนิดของตัวอักษรที่อยู่ในกลุ่มหยาบ
- Why Rough Sets? [4] งานวิจัยนี้เสนองราฟเซต และอธิบายขอบเขตบน และขอบเขตล่างของข้อมูล ซึ่งนำไปใช้ในการวิเคราะห์ค่าการขึ้นต่อกันของแอตทริบิว สำหรับวิเคราะห์แอตทริบิวที่มีความจำเป็นต่อการจำแนก
- Knowledge-base Reduction based on Rough Set Techniques [6] งานวิจัยนี้เสนองการนำราฟเซตไปประยุกต์ใช้สำหรับเลือกแอตทริบิวที่จำเป็นต่อการวิเคราะห์ระบบไฟฟ้า

- Knowledge-base Reduction based on Rough Set Techniques [8] งานวิจัยนี้เสนออัลกอริทึมสำหรับการหา reduct ซึ่งเป็นเซตของแอตทริบิวต์ที่ถูกลดโดยกราฟเซต โดยที่ยังคงค่าการขึ้นต่อกันเหมือนข้อมูลที่ยังไม่ได้ถูกลด ซึ่งงานวิจัยนี้ได้นำมาประยุกต์ใช้ใน Knowledge Discovery
- Printed Thai Character Recognition using Rough Classification คุณภาพผนวก ก. เพิ่มเติม การหาลักษณะเด่นที่จำเป็นต่อการรู้จำมีความสำคัญ งานวิจัยนี้จึงเสนอทฤษฎีกราฟเซตเพื่อหาลักษณะเด่นที่จำเป็นต่อการจำแนกกลุ่มตัวอักษร ลักษณะเด่นที่ใช้ในงานวิจัยนี้คือ วิธีการโปรเจกชันแบบวงแหวน (Ring projection) ลักษณะเด่นที่น่าสนใจนี้มีความต่อเนื่องของข้อมูล จึงมีการแบ่งช่วงโดยกำหนดให้ทุกช่วงมีขนาดที่เท่ากัน และจัดกลุ่มหยาบจากตัวอย่างที่นำมาเรียนรู้ ตัวอย่างที่มีลักษณะเด่นใกล้เคียงกันจะถูกจัดให้อยู่กลุ่มเดียวกัน หลังจากนั้นนำไปเรียนรู้โดยกราฟเซต กำหนดแอตทริบิวต์เงื่อนไขคือลักษณะเด่นที่แบ่งช่วง และแอตทริบิวต์ของการตัดสินใจคือกลุ่มแต่ละกลุ่ม ซึ่งภายในกลุ่มประกอบด้วยตัวอักษรหลายชนิด หลังจากเรียนรู้โดยกราฟเซตแล้ว ได้นำแอตทริบิวต์ที่จำเป็นมาสร้างกฎสำหรับการจำแนกกลุ่มหยาบ จากนั้นจะจำแนกชนิดของตัวอักษร โดยใช้วิธี Min-Max Boundaries
- Recognition of Printed Thai Characters using Fuzzy and Rough Sets Theory คุณภาพผนวก ก. เพิ่มเติม การแบ่งช่วงของข้อมูล ทำให้สูญเสียค่าของข้อมูลไป แต่จะทำให้เห็นลักษณะของข้อมูลได้มากขึ้น ดังนั้นงานวิจัยนี้จึงเสนอการจัดกลุ่มข้อมูลโดยใช้วิธีฟuzzy เพื่อให้มีการแบ่งช่วงที่เหมาะสม ก่อนที่จะนำไปเรียนรู้โดยกราฟเซต โดยได้จัดกลุ่มหยาบกับข้อมูลลักษณะเด่นตัวอย่างที่ใช้ในการเรียนรู้ หลังจากนั้นนำไปเรียนรู้โดยกราฟเซตและสร้างกฎสำหรับการจำแนกกลุ่มหยาบ จากนั้นจะจำแนกชนิดของตัวอักษร โดยใช้วิธี Min-Max Boundaries

จากงานวิจัยดังกล่าวข้างต้น ได้มีการประยุกต์ฟuzzy เซตและกราฟเซตมาใช้ในการรู้จำตัวอักษร เพื่อให้เกิดความสามารถในการรู้จำ และลดข้อมูลบางอย่างที่ไม่จำเป็นต่อการรู้จำ แต่อย่างไรก็ตาม โอกาสที่จะวิเคราะห์การรู้จำผิดพลาดสามารถเป็นไปได้ตั้งแต่การจำแนกกลุ่มหยาบ ดังนั้นงานวิจัยนี้จึงได้กำหนดแอตทริบิวต์ของการตัดสินใจให้ตรงกับชนิดของตัวอักษร และเสนอวิธีการหาลักษณะเด่นของข้อมูลที่มีความหลากหลายมากขึ้น ซึ่งประกอบด้วยลักษณะเด่นที่ไม่มีความต่อเนื่อง และลักษณะเด่นที่มีความต่อเนื่อง ในส่วนของลักษณะเด่นที่มีความต่อเนื่องได้นำการจัดกลุ่มของข้อมูลโดยใช้วิธีฟuzzy มาแบ่งข้อมูลนี้ ซึ่งการหาลักษณะเด่นของข้อมูลที่มีความหลากหลายมากขึ้น ทำให้กราฟเซตสามารถหาลักษณะเด่นที่จำเป็นและสอดคล้องกับชนิดของตัวอักษรได้มากขึ้น

## 1.4 ขอบเขตของการวิจัย

ในวิทยานิพนธ์ฉบับนี้จะเป็นการนำเสนอการรู้จำตัวอักษรตัวพิมพ์ภาษาไทยโดยใช้ฟuzzyเซตและกราฟเซต เพื่อแบ่งกลุ่มของข้อมูลที่มีความต่อเนื่อง และเลือกส่วนของข้อมูลลักษณะเด่นที่จำเป็นต่อการระบุตัวอักษรแต่ละตัว โดยรายละเอียดต่างๆ ภายในวิทยานิพนธ์ฉบับนี้ได้จัดแบ่งในส่วนของเนื้อหาทั้งหมดออกเป็น 7 บท ซึ่งแต่ละบทจะมีหัวข้อและเนื้อหาดังต่อไปนี้

### บทที่ 1 บทนำ

อธิบายถึงความเป็นมา และความสำคัญของปัญหา งานวิจัยที่เกี่ยวข้อง ตลอดจนขอบเขตของการวิจัย

### บทที่ 2 ทฤษฎีฟuzzyเซต

อธิบายถึงทฤษฎีของฟuzzyเซต การแทนข้อมูลในฟuzzyเซต ฟังก์ชันความเป็นสมาชิกและระบบการจัดกลุ่มแบบฟuzzy

### บทที่ 3 ทฤษฎีของกราฟเซต

อธิบายถึงทฤษฎีของกราฟเซต คำจำกัดความของ approximations ค่าฟังก์ชันความเป็นสมาชิกของกราฟเซต การขึ้นต่อกันของแอดทริบิว การลดแอดทริบิวที่ไม่จำเป็นและกฎการตัดสินใจ

### บทที่ 4 การประยุกต์ฟuzzyเซต และกราฟเซตร่วมกัน

อธิบายถึงการจัดกลุ่มข้อมูล โดยใช้ฟuzzyเซต และการเรียนรู้โดยกราฟเซต

### บทที่ 5 การเตรียมข้อมูลและกระบวนการรู้จำ

อธิบายถึงการหาลักษณะเด่นของตัวอักษร และขั้นตอนการทำงานของระบบการรู้จำที่ใช้ในวิทยานิพนธ์นี้

### บทที่ 6 การทดลองและผลการทดลอง

การเตรียมข้อมูลสำหรับการทดลอง ขั้นตอนการทดลอง และผลการทดสอบความถูกต้องของการรู้จำอักษรตัวพิมพ์ภาษาไทย ฟอนต์และขนาดตามแบบและขนาดที่ได้กำหนด

### บทที่ 7 สรุปและข้อเสนอแนะ

สรุปและวิจารณ์ผลที่ได้จากการทดลอง พร้อมทั้งปัญหาที่เกิดขึ้นตลอดจนข้อเสนอแนะต่างๆ ในการทำวิจัยต่อไป

## บทที่ 2

# ฟัซซีเซต

ทฤษฎีฟัซซีเซตได้ถูกเสนอโดย Lofti Zadeh ในปี ค.ศ. 1965 และถูกนำมาใช้ในการอธิบายระบบที่มีความคลุมเครือ ฟัซซีเซตมีพื้นฐานมาจากทฤษฎีเซต โดยค่าความเป็นสมาชิกของข้อมูลจะอยู่ในช่วง 0 ถึง 1 ซึ่งมีความแตกต่างจากค่าความเป็นสมาชิกของเซตธรรมดาที่มีค่า 0 และ 1 เท่านั้น ซึ่งทฤษฎีของฟัซซีเซตได้ถูกพัฒนาไปใช้ในการหลายๆ ด้าน เช่น งานทางด้านระบบควบคุมในทางวิศวกรรมศาสตร์ ทางด้านประมวลผลภาพ หรือการจัดกลุ่มของข้อมูล

### 2.1 ทฤษฎีฟัซซีเซต

ในระบบที่เป็นเซตธรรมดา การระบุถึงการเป็นสมาชิกของเซต จะมีเพียง “เป็นสมาชิก” และ “ไม่เป็นสมาชิก” โดยแทนด้วยค่าความเป็นสมาชิก 0 หรือ 1 แต่ในฟัซซีเซต การระบุถึงการเป็นสมาชิกจะอ้างจากค่าความเป็นสมาชิกที่มีค่าอยู่ในช่วง 0 ถึง 1 ถ้ากำหนดให้  $U$  เป็นเซตเอกภพสัมพัทธ์ และฟัซซีเซต  $A$  มีสมาชิกของเซตเป็น  $X_1, X_2, \dots, X_n$  โดย  $X_i (i = 1, 2, 3, \dots, n)$  เป็นค่าฟัซซีเซต ดังนั้นฟัซซีเซต  $A$  สามารถเขียนให้อยู่ในรูปความสัมพันธ์กับฟัซซีเซตได้เป็น  $\mu_A(X_i)$  โดยจะเรียกว่า ฟังก์ชันความเป็นสมาชิก (Membership function) สามารถเขียนได้เป็น

$$\mu_A(X_i); U \rightarrow [0,1] \quad (2.1)$$

โดยที่  $\mu_A(X_i)$  คือฟังก์ชันความเป็นสมาชิกของเซต  $A$

ตัวแปรฟัซซี (Fuzzy variable) คือฟัซซีเซตของระบบที่เราสนใจ ตัวอย่างเช่น ถ้าเรากำหนดให้อุณหภูมิเป็นตัวแปรฟัซซี และกำหนดเซตค่าของตัวแปรเป็น {ต่ำ, ปานกลาง, สูง} ซึ่งเราอาจจะเรียกเซตของค่าตัวแปรเหล่านี้ว่าค่าตัวแปรฟัซซี หรือเทอมเซต (Term set) นอกจากนี้ตัวแปรฟัซซีแต่ละตัวอาจมีส่วนขยาย (Hedges) เพื่อปรับค่าตัวแปรฟัซซีให้มีความยืดหยุ่นมากยิ่งขึ้น เช่น จากค่าของตัวแปรอุณหภูมิ “ต่ำ - ต่ำมาก” หรือ “สูง - ก่อนข้างสูง” เป็นต้น คำว่า “มาก” หรือ “ค่อนข้าง” ในที่นี้เป็นส่วนขยายของตัวแปรฟัซซี

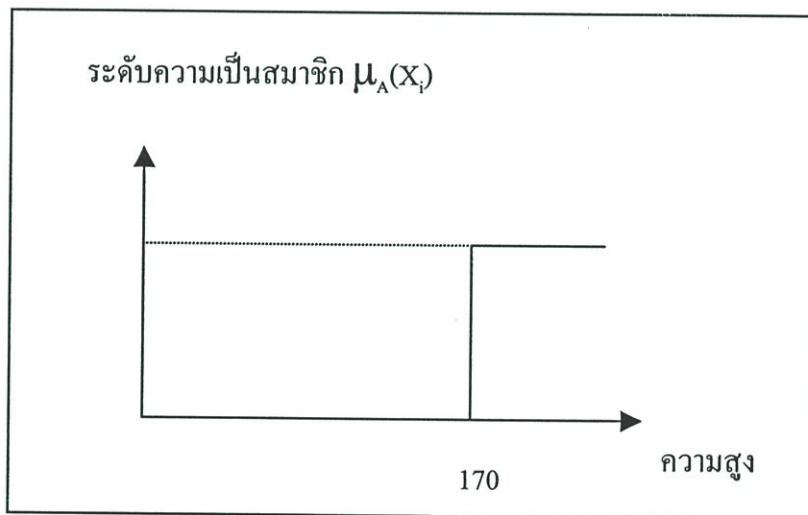
### 2.2 การแทนข้อมูลในฟัซซีเซต

การแทนข้อมูลในฟัซซีเซต เพื่อให้เข้าใจมากยิ่งขึ้นจะเปรียบเทียบกับระบบเซตธรรมดา ตัวอย่างเช่น “ความสูง” (ในระบบฟัซซีเซต ก็คือตัวแปรฟัซซีเซต) ที่เราสนใจอยู่ในช่วง 140-180 cm

(เป็นโดเมนของเอกภพสัมพัทธ์) แต่ถ้าเราสนใจเฉพาะคนที่สูง และกำหนดให้เซต A เป็นเซตของคนที่สูง (“สูง” ในพีชเซตจะเป็นทอมเซต หรือค่าตัวแปรพีชเซต) ดังนั้นเซต A ในกรณีของระบบเซตธรรมดาจะต้องมีการกำหนดค่าเริ่มเปลี่ยน หรือค่าเทรชโฮลด์ (Threshold) เพื่อเป็นตัวชี้ว่าข้อมูลจะเป็นสมาชิกของเซตใด ถ้าลองสมมุติกำหนดค่าเทรชโฮลด์มีค่าเท่ากับ 170 cm นั่นคือ ถ้าคนใดมีความสูงมากกว่า หรือเท่ากับ 170 cm จะถือว่าคนนั้นสูง (เป็นสมาชิกของเซต A) ซึ่งสามารถเขียนเป็นฟังก์ชันความเป็นสมาชิกได้ดังสมการที่ 2.2

$$\mu_A(x_i) = \begin{cases} 1; & x \geq 170 \\ 0; & \text{otther} \end{cases} \quad (2.2)$$

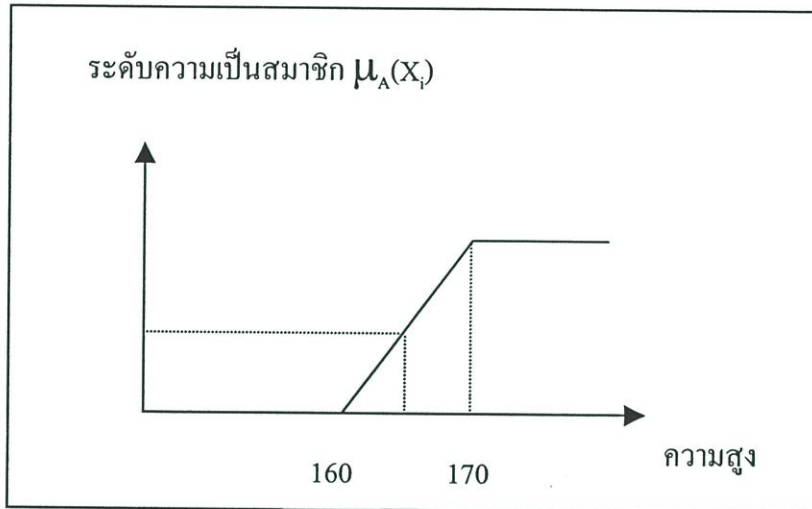
โดยที่  $x_i$  คือความสูงของคนที  $i$  ดังนั้นจากสมการที่ 2.2 เราสามารถนำมาเขียนเป็นกราฟแสดงความ เป็นสมาชิกของเซต ได้ดังรูปที่ 2.1



รูปที่ 2.1 ค่าความเป็นสมาชิกของเซต A สำหรับกรณีที่เป็นเซตธรรมดา

จากรูปที่ 2.1 จะเห็นได้ว่าคนที่มีความสูงตั้งแต่ 170 cm ขึ้นไปเท่านั้นจึงถือว่าเป็นคนสูง (มีค่าความเป็นสมาชิกเท่ากับ 1) นอกนั้นก็ถือว่าเป็นคนเตี้ยทั้งหมด แต่ในความเป็นจริงจะไม่ถูกต้องมากนัก เพราะคนที่มีความสูง 169.9 cm ก็มีความสูงใกล้เคียงกับคนที่สูง 170 cm มาก แต่ถือว่าเป็นคนเตี้ย จากปัญหานี้ สามารถนำพีชเซตที่มีฟังก์ชันความเป็นสมาชิกที่ยืดหยุ่นกว่า และสอดคล้องกับความเป็นจริงมากกว่าเซตธรรมดา มาอธิบายปัญหานี้ ดังในสมการที่ 2.3

$$\begin{aligned} & \frac{1}{(170 - 160)}(X_i - 160); & \text{ถ้า } 160 \leq 170 \\ \mu_A(X_i) &= 1; & \text{ถ้า } X_i > 170 \\ \mu_A(X_i) &= 0; & \text{other} \end{aligned} \quad (2.3)$$

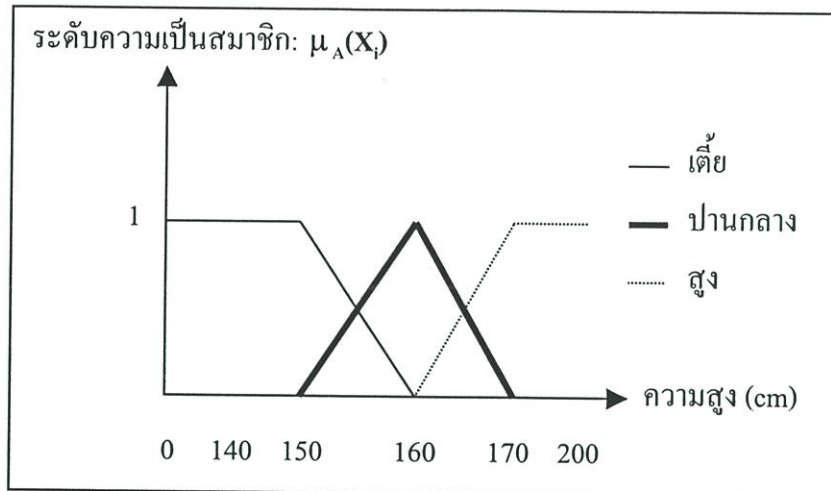


รูปที่ 2.2 ค่าความเป็นสมาชิกของเซต A สำหรับกรณีที่เป็นฟัซซี่เซต

จากรูปที่ 2.1 และรูปที่ 2.2 จะเห็นได้ว่ากราฟแสดงความเป็นสมาชิกของฟัซซี่เซตจะมีความต่อเนื่องในช่วง  $[0,1]$  เช่น คนที่มีความสูง 165 cm ก็จะถือว่าเป็นคนสูงเช่นกัน แต่เขาจะมีระดับความเป็นสมาชิกของเซตคนสูงเท่ากับ 0.5 ถ้าเปรียบเทียบกับเซตธรรมดาจะพบว่าคนนี้ไม่มีความเป็นสมาชิกของเซตคนสูง (เซต A) ทั้งๆ ที่ความสูงแตกต่างจากความสูงที่เป็นค่าเทรชโฮลด์เพียงไม่กี่เซนติเมตร

### 2.3 ฟังก์ชันความเป็นสมาชิก

เรามาศึกษา คุณสมบัติบางประการของฟัซซี่เซต เมื่อพิจารณาถึงกราฟรูปที่ 2.2 แกนนอนของกราฟจะแทน โดเมนของฟัซซี่เซต ส่วนแกนตั้งจะแทน ค่าระดับความเป็นสมาชิก แต่ในกราฟรูปที่ 2.2 มีเทอมเซตเพียงเทอมเดียว ซึ่งเป็นเทอมของคนสูงเท่านั้นเพื่อให้ครอบคลุมกลุ่มประชากรที่เราสนใจจึงเพิ่มเทอมเซตอีกสองเทอมคือ เทอมเซตของคน “เตี้ย” และ เทอมเซตของคนที่มีความสูง “ปานกลาง” ดังรูปที่ 2.3 ซึ่งในกรณีนี้โดเมนของตัวแปรฟัซซี่เซตที่ครอบคลุมประชากรทั้งหมดที่ทำการศึกษา เราจะเรียกโดเมนของตัวแปรนี้ว่า เซตเอกภพสัมพัทธ์ และในแต่ละเทอม เซตก็จะมีโดเมนของตัวเองอย่างเช่น เทอมเซตของคนเตี้ยจะมีโดเมนอยู่ในช่วง 140-150cm เทอมเซตของคนที่มีความสูงปานกลางจะมีโดเมนอยู่ในช่วง 150 – 170cm และเทอมเซตของคนที่สูงจะ โดเมนอยู่ในช่วง 170 – 200 cm



รูปที่ 2.3 กราฟฟังก์ชันความเป็นสมาชิกของตัวแปร “ความสูง” ที่ประกอบไปด้วย 3 เทอมเซต

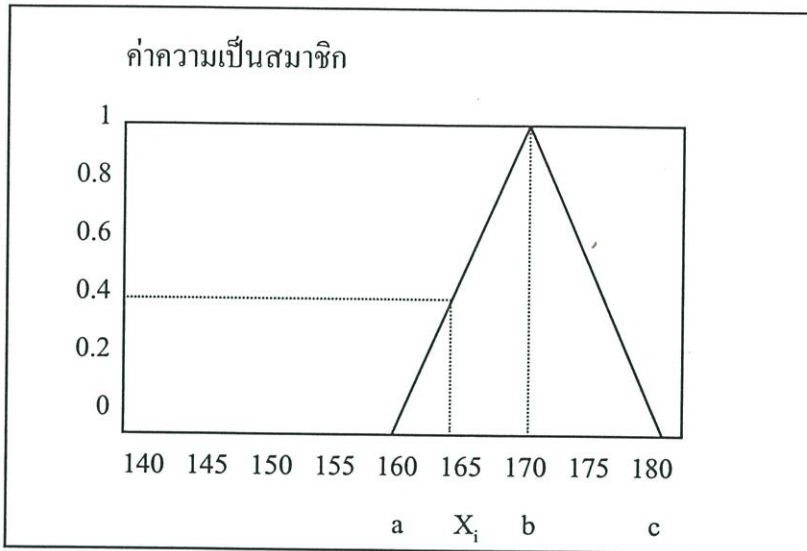
กราฟฟังก์ชันความเป็นสมาชิกมีหลายแบบที่จะใช้ในการประมาณค่าระดับความเป็นสมาชิกของแต่ละเทอมเซต โดยมีทั้งแบบที่ประมาณค่าเป็นเชิงเส้น เช่น รูปสามเหลี่ยม, รูปสี่เหลี่ยมคางหมู และแบบที่ประมาณค่าโดยฟังก์ชันต่อเนื่อง เช่น  $S, S^{-1}$  ระวังอย่าเป็นต้น ซึ่งมีรายละเอียดดังนี้

- สมการเชิงเส้น

ฟังก์ชันความเป็นสมาชิกแบบสมการเชิงเส้นที่ใช้ในแปลงค่าความสัมพันธ์ของโดเมนไปยังเรนจ์ สำหรับข้อมูลที่มีความสัมพันธ์เป็นเชิงเส้น ดังสมการที่ 2.3

- รูปสามเหลี่ยม

การกำหนดฟังก์ชันความเป็นสมาชิกโดยใช้รูปสามเหลี่ยม นิยมใช้กันมากในระบบควบคุมที่ใช้ฟuzzyเซต เทอมเซตที่กำหนดโดยฟังก์ชันนี้จะต้องมีค่าที่เหมาะสมที่สุดอยู่เพียงค่าเดียวที่ทำให้ค่าความเป็นสมาชิกเท่ากับหนึ่ง ส่วนค่าอื่น ๆ จะมีความเป็นสมาชิกลดลงเรื่อย ๆ เมื่อยิ่งห่างจากค่านี้มากขึ้น ในตัวอย่างของตัวแปรความสูง เทอมเซต “ปานกลาง” มีฟังก์ชันความเป็นสมาชิกเป็นรูปสามเหลี่ยม ถ้ากำหนดให้  $a \leq b \leq c$  เมื่อ  $a, b$  และ  $c$  เป็นเลขจำนวนจริงใด ๆ ดังนั้นสมการของฟังก์ชันความเป็นสมาชิกของรูปสามเหลี่ยมสามารถกำหนดได้ดังนี้

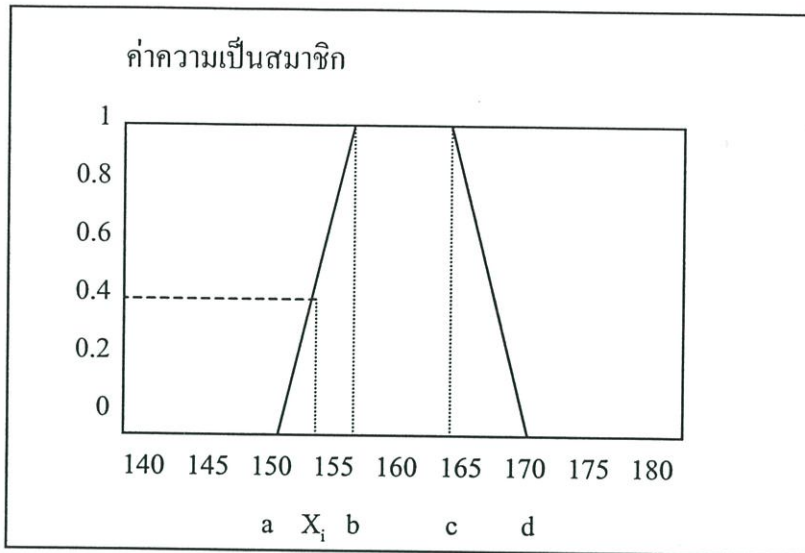


รูปที่ 2.4 ภาพฟังก์ชันรูปสามเหลี่ยม

$$\mu_A(X_i) = \begin{cases} \frac{1}{b-a}(x_i - a); & \text{ถ้า } a \leq x_i \leq b \text{ และ } a \neq b \\ \frac{1}{c-b}(c - x_i); & \text{ถ้า } b \leq x_i \leq c \text{ และ } a \neq b \\ 0; & \text{ในกรณีอื่นๆ} \end{cases} \quad (2.4)$$

- รูปสี่เหลี่ยมคางหมู

สำหรับฟังก์ชันความเป็นสมาชิกที่กำหนดด้วยสมการรูปสี่เหลี่ยมคางหมู นิยมใช้ใน ระบบควบคุมที่ใช้ฟังก์ชันเซตเช่นกัน แต่จะพบน้อยกว่าฟังก์ชันความเป็นสมาชิกที่กำหนดด้วยสมการ รูปสามเหลี่ยม ฟังก์ชันความเป็นสมาชิกที่จะกำหนดด้วยสมการรูปสี่เหลี่ยมคางหมูจะต้องมีช่วงของ ค่าที่เหมาะสมมากที่สุดอยู่ในกลุ่มหนึ่งที่ทำให้ค่าความเป็นสมาชิกของเทอมเซตนั้น ๆ นอกจากนั้นจะมีค่าความเป็นสมาน้อยลงเรื่อย ๆ เมื่อยิ่งห่างจากข้อมูลกลุ่มนี้ ถ้ากำหนดให้  $a \leq b \leq c$  เมื่อ  $a, b, c$  และ  $d$  เป็นตัวเลขจำนวนจริงใด ๆ ดังนั้นสมการของฟังก์ชันความเป็นสมาชิกของรูปสี่เหลี่ยมคาง หมูสามารถกำหนดได้ดังนี้

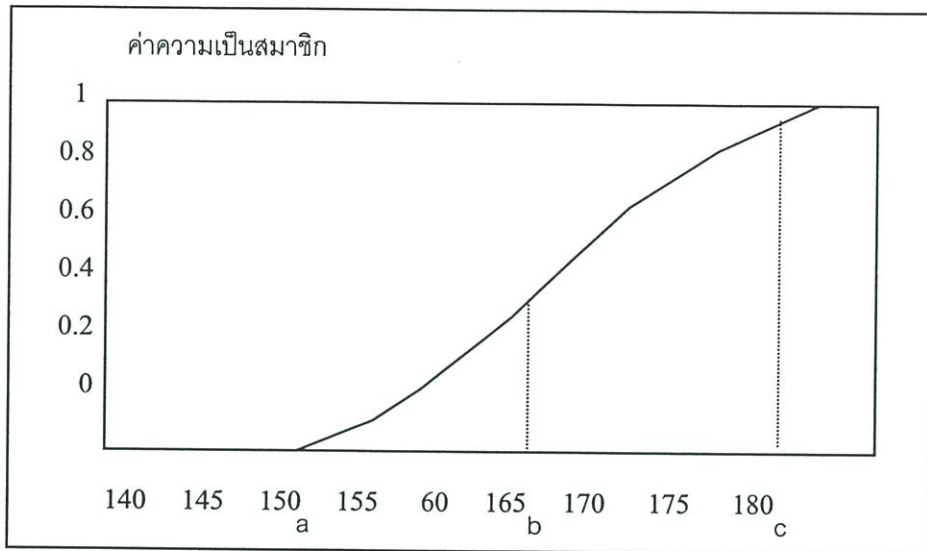


รูปที่ 2.5 ภาพฟัซซี่รูปสี่เหลี่ยมคางหมู

$$\mu_A(X_i) = \begin{cases} \frac{1}{b-a}(x_i - a); & \text{ถ้า } a \leq x_i \leq b \text{ และ } a \neq b \\ 1; & \text{ถ้า } b \leq x_i \leq c \\ \frac{1}{c-b}(c - x_i); & \text{ถ้า } c \leq x_i \leq d \text{ และ } c \neq b \\ 1; & \text{ในกรณีอื่นๆ} \end{cases} \quad (2.5)$$

- เส้นโค้งรูปตัว S

ฟังก์ชันความเป็นสมาชิกที่กำหนดด้วยเส้นโค้งรูปตัว S (S-Curve, Sigmoid หรือ Logistic) เป็นการกำหนดความสัมพันธ์ระหว่างโดเมนไปยังเรนจ์แบบไม่เป็นเชิงเส้น ตัวแปรที่เหมาะสมที่จะใช้ฟังก์ชันนี้คือพวกอายุการใช้งานของอุปกรณ์ต่าง ๆ ตัวแปรส่วนใหญ่ที่ใช้ในทางการประมวลผลทางภาพเป็นต้น ซึ่งตัวแปรพวกนี้มักจะมีความสัมพันธ์แบบไม่เป็นเชิงเส้นที่เป็นรูปตัว S หรือ  $S^{-1}$  ถ้ากำหนดให้  $a \leq b \leq c$  เมื่อ  $a$ ,  $b$  และ  $c$  เป็นเลขจำนวนจริงใด ๆ ดังนั้นสมการของฟังก์ชันเป็นสมาชิกของเส้นโค้งรูปตัว S สามารถกำหนดให้ดังสมการที่ 2.6 และ  $S^{-1} = 1 - \mu_A(X_i)$  โดยที่  $\mu_A(X_i)$  ได้จากสมการที่ 2.6

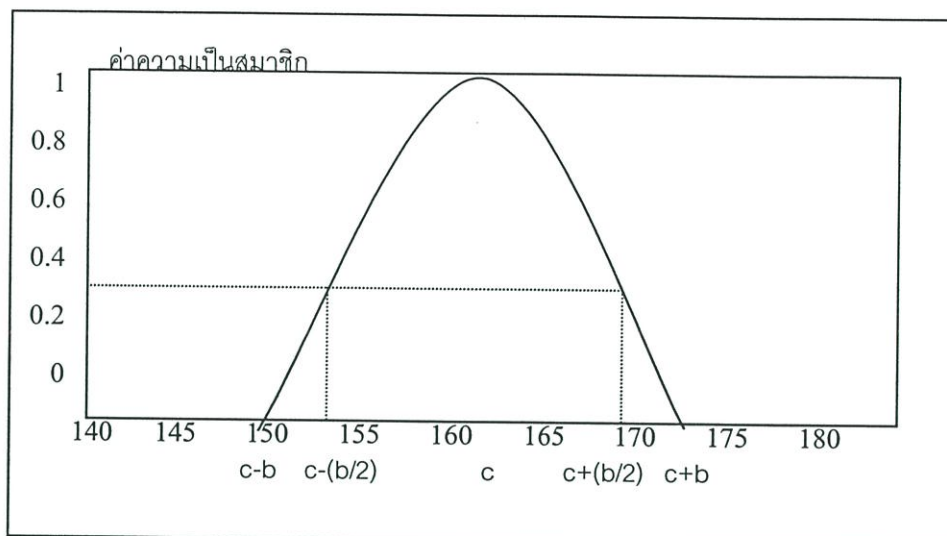


รูปที่ 2.6 เส้นโค้งรูปตัว S

$$\mu_A(X_i) = \begin{cases} 0; & \text{ถ้า } x_i \leq a \\ 2((x_i - a)/(c - a))^2; & \text{ถ้า } a \leq x_i \leq b \\ 1 - 2((x_i - a)/(c - a))^2; & \text{ถ้า } b \leq x_i \leq c \\ 1; & \text{ถ้า } x_i \geq c \end{cases} \quad (2.6)$$

• รูปประฆังคว่ำ

เทอมเซตที่ใช้ฟังก์ชันความเป็นสมาชิกแบบรูปประฆังคว่ำ จะมีคุณสมบัติของข้อมูล คล้ายกับสมการรูปร่างสามเหลี่ยม แต่ข้อมูลของโดเมนที่แปลงไปยังเรนจ์จะมีความสัมพันธ์ในลักษณะ ไม่เป็นเชิงเส้น สมการของรูปประฆังคว่ำมีด้วยกันหลายสมการแต่ที่ใช้กันมากคือ ฟังก์ชัน  $\pi[1,30]$  ฟังก์ชัน Gaussian และฟังก์ชัน Beta ซึ่งสามารถกำหนดได้ดังนี้



รูปที่ 2.7 ค่าพีชชีเซตรูประฆังคว่ำ

ฟังก์ชัน จำนวน ได้จากสมการที่ 2.7

$$S(X_i; a, b, c) = \begin{cases} 0; & \text{ถ้า } x_i \leq a \\ 2((x_i - a)/(c - a))^2; & \text{ถ้า } a < x_i \leq b \\ 1 - 2((x_i - a)/(c - a))^2; & \text{ถ้า } b < x_i \leq c \\ 1; & \text{ถ้า } x_i > c \end{cases} \quad (2.7)$$

$$\text{โดยที่ } b = \frac{c - a}{2}$$

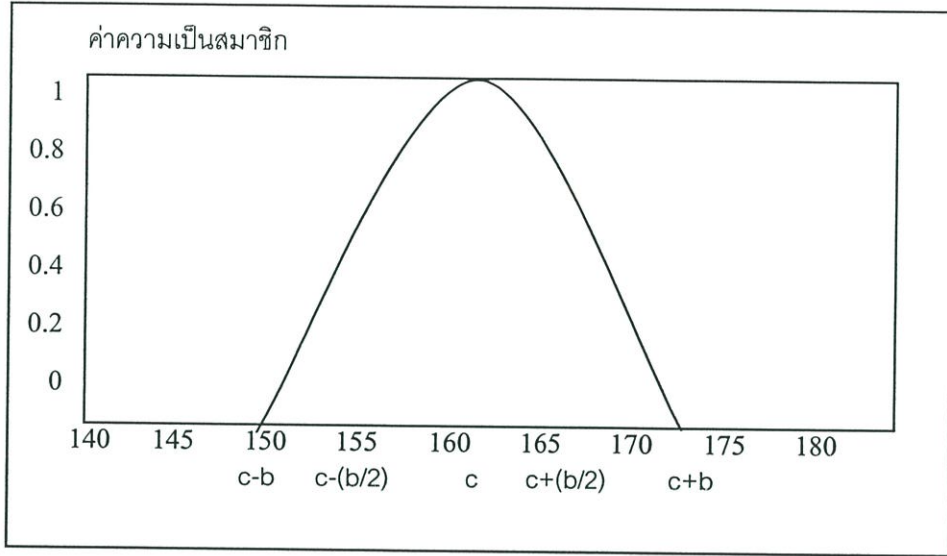
ดังนั้นฟังก์ชัน  $\pi$  สามารถกำหนดได้โดย

$$\mu_A(x_i) = \begin{cases} S(X_i; c - b, c - b/2, c); & \text{ในกรณี } X_i \leq c \\ 1 - S(X_i; c - b, c - b/2, c + b); & \text{ในกรณี } X_i > c \end{cases} \quad (2.8)$$

ฟังก์ชัน Gaussian

$$\mu_A(X_i) = e^{K(x - x_i)^2} \quad (2.9)$$

โดยที่  $K = 1/(2\sigma)^2$  ซึ่งเป็นโคเมนของพีชชีเซต A ที่มีค่าเฉลี่ยเป็น  $\bar{X}$  และส่วนเบี่ยงเบนมาตรฐาน  $\sigma$



รูปที่ 2.8 ค่าฟังก์ชัน Gaussian

#### ฟังก์ชัน Beta

ฟังก์ชันนี้จะมีคุณสมบัติคล้ายกับสมการรูปสี่เหลี่ยมคางหมู คือจะมีช่วงของข้อมูลที่เหมาะสมกับเทอมเซตนั้น ๆ อยู่กลุ่มหนึ่งที่ทำให้ค่าความเป็นสมาชิกของเทอมเซตมีค่าเข้าใกล้ 1 ซึ่งข้อมูลกลุ่มนี้จะมีความกว้างของช่วงเท่ากับค่า  $\beta$  ซึ่งถ้า  $\sigma \rightarrow \infty$  จะทำให้ข้อมูลในช่วง  $\beta$  มีค่าเข้าใกล้ 1 มากยิ่งขึ้น โดยมีจุดศูนย์กลางอยู่ที่  $\bar{X}$  ดังแสดงในสมการที่ 2.10

$$\mu_A(X_i) = 1 / (1 + (X_i - \bar{x}) / \beta)^p \quad (2.10)$$

## 2.4 การจัดกลุ่มแบบพีชชี

การจัดกลุ่มของข้อมูล (Clustering) เป็นส่วนที่สำคัญสำหรับการแบ่งเซตของข้อมูลให้อยู่ในกลุ่มที่เหมือนกัน โดยอาศัยการวัดความคล้ายคลึงกัน นั่นคือข้อมูลที่อยู่ในกลุ่มเดียวกันจะมีความคล้ายคลึงกัน และข้อมูลที่แตกต่างกันจะจัดอยู่ในคนละกลุ่ม วิธีการวัดความคล้ายคลึงกันมีความสำคัญต่อการจำแนกกลุ่ม เพราะแต่ละวิธีการจะแสดงคุณสมบัติแตกต่างกัน ตัวอย่างเช่น การวัดระยะทาง, การวัดการเชื่อมต่อ หรือความเข้มของข้อมูล โดยการเลือกวิธีการจะขึ้นอยู่กัลักษณะของการจัดกลุ่ม ในระบบการจัดกลุ่มที่ไม่เป็นพีชชี ขอบเขตของการแบ่งกลุ่มจะเป็นลักษณะของคริสปี (0 หรือ 1) นั่นคือระบุว่าข้อมูลนั้นจะอยู่ในกลุ่มใดกลุ่มหนึ่งเท่านั้น ในทางตรงกันข้าม, ระบบ

การจัดกลุ่มที่เป็นพีชซึ่งจะมีความสามารถในการจัดการกับข้อมูลที่มีความคลุมเครือของการแบ่งกลุ่มได้มากกว่าระบบการจัดกลุ่มที่ไม่เป็นพีชซึ่ง จะมีการแสดงค่าที่เป็นตัวบอกว่าข้อมูลนั้นมีโอกาสที่จะอยู่ในกลุ่มใดได้บ้าง ซึ่งเป็นค่าความเป็นสมาชิกของข้อมูลแต่ละกลุ่ม

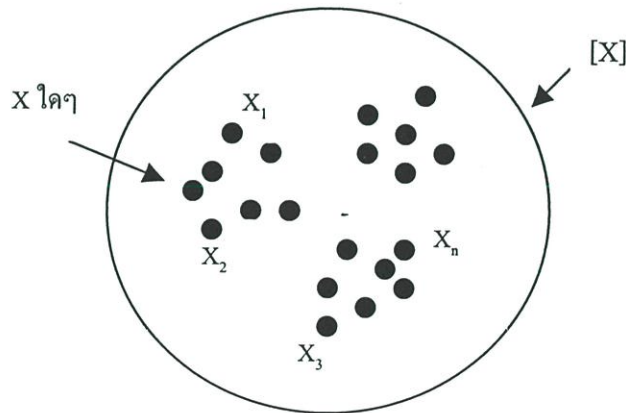
ในการจัดกลุ่มข้อมูลของทั้งระบบที่ไม่เป็นพีชซึ่ง และเป็นพีชซึ่งนั้น สิ่งที่จะเป็นตัวกำหนดลักษณะการจัดกลุ่มคือ วิธีการวัดความคล้ายคลึงกัน ซึ่งจะใช้วิธีการวัดระยะทางระหว่างข้อมูลกับจุดศูนย์กลางของกลุ่มข้อมูลเป็นตัวกำหนดลักษณะในการจัดกลุ่ม โดยที่ระยะทางดังกล่าวเรียกว่า “ระยะทางยูคลีเดียน” (Euclidean distance) โดยสามารถระยะทางยูคลีเดียนได้ดังสมการที่ 2.11

$$d_{ij}^2 = \|x_j - v_i\|^2 \quad (2.11)$$

โดย  $d_{ij}$  เป็นระยะทางยูคลีเดียนระหว่างข้อมูล  $x_j$  และ  $v_i$  ซึ่งเป็นจุดศูนย์กลางของกลุ่มข้อมูลที่  $i$  โดยที่ค่า  $j$  มีค่าอยู่ระหว่าง  $1 \leq j \leq n$  เมื่อ  $n$  คือจำนวนของข้อมูลทั้งหมด และค่า  $i$  มีค่าอยู่ระหว่าง  $1 \leq i \leq c$  เมื่อ  $c$  คือจำนวนกลุ่มที่ต้องการแบ่ง

#### 2.4.1 การแบ่งกลุ่มข้อมูลที่ไม่เป็นพีชซึ่ง

พิจารณา เซตข้อมูล  $X = \{x_1, x_2, \dots, x_n\}$  ซึ่งเป็นเซตจำกัด ซึ่ง  $n$  เป็นจำนวนข้อมูลทั้งหมดที่จะนำมาแบ่งกลุ่ม (Cluster) ดังแสดงในรูปที่ 2.9



รูปที่ 2.9 กลุ่มข้อมูล  $[X]$  มีสมาชิก  $\{x_1, x_2, x_3, \dots, x_n\}$

จากรูปที่ 2.9 กำหนด  $x_j$  เป็นข้อมูลใดๆ ที่อยู่ใน  $X$  โดยกำหนดให้  $P(x)$  คือ เพาเวอร์เซตของ  $X$  หรือเซตของทุกสับเซตใน  $X$

ถ้ากำหนดให้  $\{A_i \in P(x) ; i = 1, 2, \dots, c\}$  โดย  $A_i$  เป็นเซตของข้อมูลที่อยู่ในกลุ่มเดียวกัน ซึ่งจะมีคุณสมบัติดังนี้

1. การยูเนียนของ  $A_i$

$$\bigcup_{i=1}^c A_i = X \quad (2.12)$$

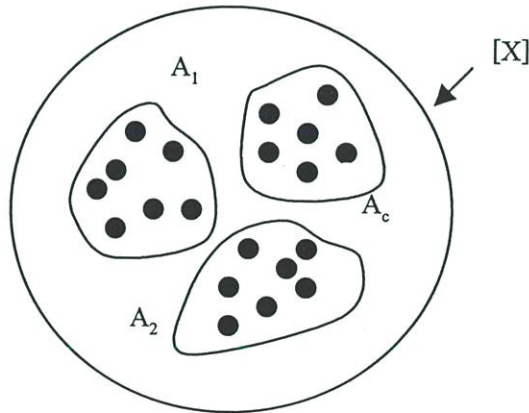
เมื่อนำข้อมูลที่อยู่ในแต่ละ  $A_i$  ทั้งหมดมายูเนียนกันจะเท่ากับข้อมูลทั้งหมด ดังสมการที่ 2.12

2. การอินเตอร์เซกของ  $A_i$

$$A_i \cap A_j = 0; \quad 1 \leq i \neq j \leq c \quad (2.13)$$

จากคุณสมบัติของการอินเตอร์เซกกัน ดังสมการที่ 2.13 จะเห็นได้ว่าสมาชิกของ  $A_i$  จะไม่เป็นสมาชิกของ  $A_j$

ถ้ากำหนดให้  $c$  คือจำนวนกลุ่มที่ต้องการจะแบ่ง ดังนั้นการแบ่งข้อมูลทั้งหมดออกเป็นกลุ่ม ซึ่งเท่ากับจำนวน  $c$  กลุ่ม หรือเขียนได้เป็น  $\{A_1, A_2, \dots, A_c\}$  ดังแสดงในรูป 2.10



รูปที่ 2.10 การแบ่งข้อมูล  $[X]$  เป็นกลุ่ม  $A_i; i = 1, 2, \dots, c$

การแบ่งกลุ่มข้อมูลที่ไม่เป็นฟิชซึ่งสามารถเขียนฟังก์ชันความเป็นสมาชิกของ  $x_j$  ในกลุ่ม  $A_i$  ได้ดังสมการที่ 2.14 โดย  $u_{ij}$  คือค่าความเป็นสมาชิก ซึ่งเป็นค่าความเป็นสมาชิกของ  $x_j$  ในเซต  $A_i$

$$u_{ij} = \begin{cases} 1, & x_j \in A_i \\ 0, & x_j \notin A_i \end{cases} \quad (2.14)$$

จากสมการที่ 2.14 ถ้า  $u_{ij}$  เป็น 1 แสดงว่า  $x_j$  เป็นสมาชิกของกลุ่ม  $A_i$  แต่ถ้า  $u_{ij}$  เป็น 0 แสดงว่า  $x_j$  ไม่ได้เป็นสมาชิกของกลุ่ม  $A_i$  และค่า  $u_{ij}$  มีคุณสมบัติคือ

1. ผลรวมของค่าความเป็นสมาชิกทั้งหมดของข้อมูล  $x_j \in X$  ในทุกกลุ่ม จะมีค่าเท่ากับ 1.0 ดังสมการที่ 2.15

$$\sum_{i=1}^c u_{ij} = 1, \text{ สำหรับทุก } j = 1, 2, \dots, n \quad (2.15)$$

2. ในแต่ละกลุ่มจะมีค่าความเป็นสมาชิก และแตกต่างกันในแต่ละเซต ดังสมการ 2.16

$$0 < \sum_{j=1}^n u_{ij} < n, \text{ สำหรับทุก } i = 1, 2, \dots, c \quad (2.16)$$

และจะเห็นได้ว่า  $x_j$  จะต้องเป็นสมาชิกของกลุ่มใดกลุ่มหนึ่งเท่านั้น ดังนั้นผลบวกในแต่ละคอลัมน์หรือภายในกลุ่มจะต้องเท่ากับ 1 เสมอ และผลบวกของ  $u_{ij}$  ในแต่ละแถว หรือในแต่ละกลุ่มจะมีค่ามากกว่า 0 แต่ไม่เกิน  $n$  ซึ่งหมายความว่า ทุกกลุ่มจะต้องมีสมาชิกอย่างน้อย 1 ตัว แต่ต้องน้อยกว่าจำนวนข้อมูลทั้งหมด สมการที่ 2.17 เป็นการจัดกลุ่มข้อมูลที่ไม่เป็นแบบฟัซซี่

$$U = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad (2.17)$$

#### 2.4.2 การแบ่งกลุ่มข้อมูลโดยวิธีฟัซซี่

ผลลัพธ์ของการแบ่งกลุ่มโดยวิธีฟัซซี่ สามารถแสดงอยู่ในรูปของ partition matrix  $U$  ดังสมการ 2.18

$$U = [u_{ij}]_{i=1 \dots c, j=1 \dots n} \quad (2.18)$$

โดยค่า  $u_{ij}$  คือค่าความเป็นสมาชิกอยู่ระหว่าง 0 ถึง 1 และแสดงค่าที่ข้อมูล  $x_j$  จะอยู่ในกลุ่ม  $i$  และค่า  $u_{ij}$  มีคุณสมบัติคือ

1. ผลรวมของค่าความเป็นสมาชิกทั้งหมดของข้อมูล  $x_j \in X$  ในทุกกลุ่ม จะมีค่าเท่ากับ 1.0 ดังสมการที่ 2.19

$$\sum_{i=1}^c u_{ij} = 1, \text{ สำหรับทุก } j = 1, 2, \dots, n \quad (2.19)$$

2. ในแต่ละกลุ่มจะมีค่าความเป็นสมาชิก และแตกต่างกันในแต่ละเซต ดังสมการ 2.20

$$0 < \sum_{j=1}^n u_{ij} < n, \text{ สำหรับทุก } i = 1, 2, \dots, c \quad (2.20)$$

สำหรับการแบ่งกลุ่มโดยวิธีฟัซซี่ เราสามารถเขียนค่า  $u$  ได้ใหม่ จากสมการที่ 2.21 จะเห็นได้ว่าค่าของ  $u$  สามารถจะมีค่าอยู่ในทั้งสองกลุ่มได้ ซึ่งแตกต่างกับค่าของ  $u$  ในสมการที่ 2.17 ซึ่งจะมีค่าอยู่ในกลุ่มใดกลุ่มหนึ่งเท่านั้น แสดงให้เห็นว่าระบบที่เป็นฟัซซี่จะมีความยืดหยุ่นมากกว่าระบบที่ไม่เป็นฟัซซี่

$$u = \begin{bmatrix} 0.9 & 0.2 & 0.9 \\ 0.1 & 0.8 & 0.1 \end{bmatrix} \quad (2.21)$$

**อัลกอริทึม FCM : Fuzzy c-Means Algorithm**

ขั้นที่ 1: กำหนดจำนวนกลุ่ม  $c$  ( $2 \leq c \leq n$ ) และค่าพารามิเตอร์  $m$  ซึ่งเป็นค่า exponential weight ( $1 < m < \infty$ ) กำหนดค่าเมตริกซ์เริ่มต้น  $U^{(0)}$  และค่า  $\epsilon$  ซึ่งเป็นค่าผิดพลาดที่ยอมรับได้ โดยเริ่มต้นการวนรอบแรกเป็น 0

ขั้นที่ 2: คำนวณจุดศูนย์กลางของแต่ละกลุ่ม  $\{v_i^1 \mid i = 1, 2, \dots, c\}$  โดยอาศัยค่า  $U^{(0)}$  และสมการที่ 2.22 ซึ่งทุกกลุ่มเป็นการจำแนกแบบฟัซซี่

$$v_i = \frac{1}{\sum_{j=1}^n (u_{ij})^m} \sum_{j=1}^n (u_{ij})^m x_j, \quad i = 1, 2, \dots, c \quad (2.22)$$

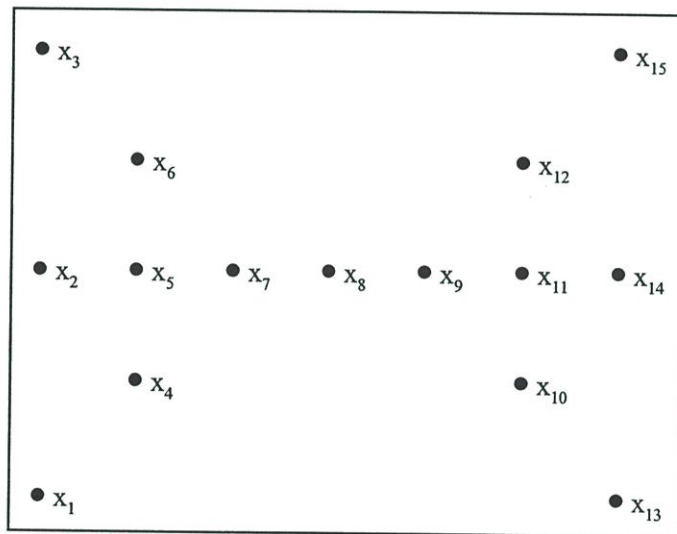
ขั้นที่ 3: คำนวณหาค่าความเป็นสมาชิกของค่า  $U^{(t+1)}$  ซึ่งเป็นค่า  $U$  ในการวนรอบถัดไป โดยใช้จุดศูนย์กลาง  $\{v_i^1 \mid i = 1, 2, \dots, c\}$  และสมการที่ 2.23

$$u_{ij} = \frac{(1/\|x_j - v_i\|^2)^{1/m-1}}{\sum_{k=1}^c (1/\|x_j - v_k\|^2)^{1/m-1}}, \quad i = 1, 2, \dots, c; j = 1, 2, \dots, n \quad (2.23)$$

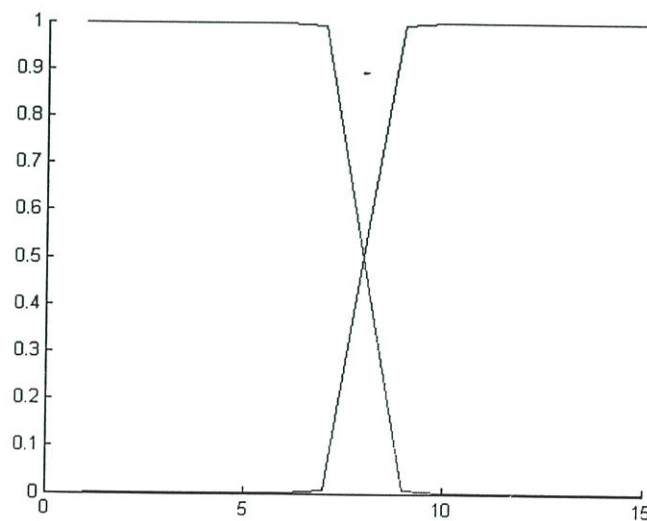
ขั้นที่ 4: คำนวณค่า  $\Delta = \|U^{(l+1)} - U^{(l)}\| = \max_{i,j} |u_{ij}^{(l+1)} - u_{ij}^{(l)}|$  ถ้า  $\Delta > \varepsilon$  ให้  $l = l+1$  และกลับไปทำขั้นที่ 2 แต่ถ้า  $\Delta \leq \varepsilon$  ให้สิ้นสุดการวนรอบการทำงาน

จากสมการ 2.22 และ 2.23 จะมีค่า  $m$  (exponential weight) ซึ่งเป็นค่าที่มีผลต่อค่าความเป็นสมาชิกใน partition matrix  $U$  ทำให้ค่าความเป็นสมาชิกของ  $u_{ij}$  มีค่าอยู่ในช่วง 0 ถึง 1 นั่นคือ  $m$  จะเป็นตัวควบคุมความคลุมเครือ (Fuzziness) ของการประมวลผลทั้งหมด กระบวนการวนรอบ

ตัวอย่าง สมมติให้  $X$  ประกอบด้วยข้อมูลจำนวน 15 จุดซึ่งมีลักษณะคล้ายผีเสื้อ โดยข้อมูล  $x_1$  ถึง  $x_7$  อยู่ทางด้านซ้าย ส่วน  $x_8$  ถึง  $x_{15}$  อยู่ทางด้านขวา โดยมีข้อมูล  $x_8$  อยู่ตรงกลาง ดังแสดงในรูปที่ 2.11 เมื่อนำมาแบ่งเป็น 2 กลุ่ม ( $c = 2$ ) และค่า  $m = 1.40$  สามารถแสดงได้ดังรูป 2.12



รูปที่ 2.11 ข้อมูล  $[X]$  จำนวน 15 จุด



รูปที่ 2.12 ผลของการจำแนกด้วยอัลกอริทึม FCM

ผลของการจำแนกด้วยอัลกอริทึม Fuzzy c-means แสดงการแบ่งกลุ่ม และค่าความเป็นสมาชิกตามจำนวนกลุ่มที่กำหนด ยังไม่มีวิธีการที่มาวิเคราะห์จำนวนกลุ่มที่ถูกต้อง แต่สามารถวัดค่าความเป็นสมาชิกของ  $U$  ในการแบ่งแต่ละกลุ่มเพื่อประเมินการแบ่งข้อมูลตามจำนวนกลุ่ม ด้วยคุณสมบัติตามสมการที่ 2.24 และ 2.25

Partition Entropy

$$H(U, c) = \frac{1}{n} \sum_{j=1}^c \sum_{i=1}^n |u_{ij} \ln u_{ij}| \quad (2.24)$$

Partition coefficient

$$F(U, c) = \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n u_{ij}^2 \quad (2.25)$$

จากคุณสมบัติในสมการที่ 2.24 และ 2.25 จะเห็นได้ว่า ค่า  $H = 0$  และ  $F = 1$  ถ้า  $u_{ij} \in \{0, 1\}$  (การแบ่งกลุ่มข้อมูลโดยวิธี Hard) ค่า  $H = \ln c$  และ  $F = 1/c$  ถ้า  $u_{ij} = 1/c$  สำหรับ  $i$  และ  $j$  ทั้งหมด นั่นคือ  $0 \leq H \leq \ln c$  และ  $1/c \leq F \leq 1$  จะเห็นได้ว่า  $H$  และ  $F$  ขึ้นอยู่กับใน  $cn$  ทั้งหมด การประเมินคุณสมบัติของ  $H$  และ  $F$  ในการแบ่งกลุ่มจำนวน  $c$  กลุ่มสามารถตรวจสอบจำนวนกลุ่มที่เหมาะสมได้ดังนี้

$$\min_{c=2}^{n-1} \left\{ \min_{U \in \Omega_c} [H(U, c)] \right\}$$

$$\max_{c=2}^{n-1} \left\{ \max_{U \in \Omega_c} [F(U, c)] \right\},$$

โดย  $\Omega_c$  คือเซตของคำตอบที่เหมาะสมที่สุดในการแบ่งจำนวน  $c$  กลุ่ม

## 2.5 สรุป

ทฤษฎีฟัซซี่เซต จะระบุการเป็นสมาชิกโดยอาศัยค่าความเป็นสมาชิกซึ่งอยู่ในช่วง 0 ถึง 1 แตกต่างจากค่าความเป็นสมาชิกของเซตธรรมดาที่เป็นได้เพียง 0 หรือ 1 ทำให้เกิดความยืดหยุ่นของข้อมูล โดยได้มีการพัฒนาฟัซซี่เซตไปใช้ในงานต่างๆ มากมาย เช่น ทางด้านระบบควบคุม ทางด้านประมวลผลภาพ โดยในงานวิจัยนี้ได้นำฟัซซี่เซตมาใช้ในการแบ่งกลุ่มข้อมูลที่เหมาะสม

### บทที่ 3

## กราฟเซต

ทฤษฎีกราฟเซตถูกนำเสนอเป็นครั้งแรกโดย Zdzislaw Pawlak ในปี ค.ศ. 1980 เป็นวิธีการคณิตศาสตร์ในการวิเคราะห์กับข้อมูลที่มีความคลุมเครือ และความไม่แน่นอน กราฟเซตอยู่บนสมมุติฐานที่ว่า เราสามารถหาฐานความรู้ (ข้อมูล, ความรู้) จากทุก objects ในเอกภพสัมพัทธ์ได้ ตัวอย่างเช่น ถ้า objects คือกลุ่มคนไข้ที่ป่วยด้วยโรคที่มีความชัดเจน อาการต่างๆ ของโรคนี้จะเป็นฐานความรู้ที่เกี่ยวข้องกับคนไข้

คนไข้ทุกคนที่แสดงอาการเหมือนกันหรือเรียกว่า “Indiscernible (คล้ายคลึงกัน)” เป็นรูปของฐานความรู้ที่ตรวจสอบได้ และกลุ่มของฐานความรู้เกี่ยวกับคนไข้ (หรือประเภทของคนไข้) เรียกว่า “elementary sets” หรือ “concepts” โดย elementary นี้เป็นหน่วยความรู้ที่ได้จากกลุ่มของข้อมูลที่มี เราสามารถนำ Elementary concepts มารวมกันได้เรียกว่า “compound concepts” นั่นคือ concepts ที่เป็นเทอมของ elementary concepts เพียงอันใดอันหนึ่ง ความสัมพันธ์ที่เป็น indiscernible เป็นพื้นฐานของทฤษฎีกราฟเซต

เนื่องจากกราฟเซตจะมีขอบเขต หรือ boundary-line ซึ่ง Objects ที่อยู่ในขอบเขตจะไม่สามารถระบุได้แน่นอนว่าเป็นสมาชิกของเซตหรือไม่ เห็นได้ว่าในคริปเซตจะไม่มีส่วนที่เป็นขอบเขต นั่นก็หมายความว่าในบริเวณขอบเขตจะไม่มีความสามารถในการจำแนกได้ด้วยฐานความรู้ที่มี

ดังนั้นจากข้อมูลที่มี เราสามารถสร้างฐานความรู้ขึ้นมาได้ ฐานความรู้เกิดขึ้นมาจาก objects ที่มีความเหมือนกัน และผลของ concepts ที่มีความคลุมเครือ ทำให้ไม่สามารถจำแนก elements ภายในได้ด้วยฐานความรู้ที่มี เพราะฉะนั้นจึงได้มีการกำหนด lower และ upper approximation ขึ้นมา lower approximation ประกอบด้วย object ทั้งหมดที่อยู่ใน concepts อย่างแน่นอน ส่วน upper approximation ประกอบไปด้วย objects ทั้งหมดที่เป็นไปได้ว่าจะอยู่ใน concepts และผลต่างระหว่าง upper และ lower approximation จะเป็น boundary region ซึ่งเป็นบริเวณที่ concept มีความคลุมเครือ ประกอบด้วย objects ทั้งหมดที่ไม่สามารถจำแนกได้อย่างแน่นอน

### 3.1 ตัวอย่าง

ข้อมูลดิบจะแสดงอยู่ในรูปของตารางข้อมูล ในแต่ละคอลัมน์คือ แอตทริบิวต์, แต่ละแถวคือ objects และค่าที่อยู่ในตารางคือ ค่าแอตทริบิวต์ (attribute-values) ในตารางที่ 3.1 แสดงตัวอย่างของตารางข้อมูล

ตารางที่ 3.1 ตัวอย่างของตารางข้อมูล

คนไข้	ปวดหัว	ปวดกล้ามเนื้อ	อุณหภูมิ	ไข้หวัดใหญ่
p1	ไม่ใช่	ใช่	สูง	ใช่
p2	ใช่	ไม่ใช่	สูง	ใช่
p3	ใช่	ใช่	สูงมาก	ใช่
p4	ไม่ใช่	ใช่	ปกติ	ไม่ใช่
p5	ใช่	ไม่ใช่	สูง	ไม่ใช่
p6	ไม่ใช่	ใช่	สูงมาก	ใช่

คอลัมน์ของตารางคือ แอตทริบิวต์(อาการ) และแถวคือ objects (คนไข้) โดยค่าในตารางทั้งหมดคือ ค่าแอตทริบิวต์ ดังนั้นในแต่ละแถวของตาราง ทำให้เราทราบข้อมูลเกี่ยวกับคนไข้แต่ละคน ตัวอย่างเช่น คนไข้ p2 ในตารางมีค่าแอตทริบิวต์ดังนี้

(ปวดหัว, ใช่), (ปวดกล้ามเนื้อ, ไม่ใช่), (อุณหภูมิ, สูง), (ไข้หวัดใหญ่, ใช่)

ซึ่งเป็นฐานของความรู้เกี่ยวกับคนไข้

จากตารางที่ 3.1 เมื่อพิจารณาจากอาการปวดหัว จะเห็นได้ว่าคนไข้ p2, p3 และ p5 เป็น indiscernible กัน, เมื่อพิจารณาจากอาการปวดกล้ามเนื้อ และไข้หวัดใหญ่ จะเห็นได้ว่าคนไข้ p3 และ p6 เป็น indiscernible กัน และเมื่อพิจารณาจากอาการปวดหัว, ปวดกล้ามเนื้อ และอุณหภูมิ จะเห็นได้ว่าคนไข้ p2 และ p5 เป็น indiscernible กัน เมื่อพิจารณาจากอาการปวดหัวจะประกอบไปด้วย 2 elementary sets คือ {p2, p3, p5} และ {p1, p4, p6} แต่เมื่อพิจารณาอาการปวดหัว และปวดกล้ามเนื้อ จะประกอบไปด้วย elementary sets คือ {p1, p4, p6}, {p2, p5} และ {p3} ในทำนองเดียวกันสามารถสร้าง elementary sets จากสับเซตของอาการต่างๆ ได้

เมื่อพิจารณาอาการปวดหัว, ปวดกล้ามเนื้อ และ อุณหภูมิ จะเห็นได้ว่าคนไข้ p2 เป็นไข้หวัดใหญ่ ในขณะที่คนไข้ p5 ไม่ได้เป็นไข้หวัดใหญ่ และทั้งสอง indiscernible กัน ดังนั้นอาการไข้หวัดใหญ่ไม่

สามารถวิเคราะห์ได้โดยอาศัยทฤษฎีของอาการปวดหัว, ปวดกล้ามเนื้อ และ อุณหภูมิ เพราะฉะนั้น  $p_2$  และ  $p_5$  เป็นกรณี boundary-line ซึ่งไม่สามารถจำแนกได้ด้วยแอตทริบิวของข้อมูลที่มี ส่วนคนไข้  $p_1$ ,  $p_3$  และ  $p_6$  เมื่อสังเกตอาการเหล่านี้แล้วสามารถในการระบุได้แน่นอนว่าเป็นไข้หวัดใหญ่ และคนไข้  $p_2$  และ  $p_5$  ไม่สามารถระบุได้ว่าเป็นไข้หวัดใหญ่ และคนไข้  $p_4$  ไม่เป็นไข้หวัดใหญ่อย่างแน่นอน เมื่อสังเกตจากอาการที่แสดง ดังนั้น lower approximation ของเซตของคนไข้ที่เป็นไข้หวัดใหญ่คือ  $\{p_1, p_3, p_6\}$  และ upper approximation ของเซตนี้คือ  $\{p_1, p_2, p_3, p_5, p_6\}$  โดยกรณี boundary-line คือคนไข้  $p_2$  และ  $p_5$  ในลักษณะเดียวกันคนไข้  $p_4$  ไม่ได้เป็นไข้หวัดใหญ่ และคนไข้  $p_2, p_5$  ไม่สามารถบอกได้ว่าเป็นไข้หวัดใหญ่ ดังนั้น lower approximation ของ concept นี้คือเซต  $\{p_4\}$  ส่วน upper approximation คือเซต  $\{p_2, p_4, p_5\}$  และขอบเขตของ concept “ไม่เป็นไข้หวัดใหญ่” คือเซต  $\{p_2, p_5\}$  เหมือนกับกรณีเซตของคนไข้ที่เป็นไข้หวัดใหญ่

### 3.2 คำจำกัดความของ Approximations

กำหนดให้มี 2 เซตจำกัดและไม่เป็นเซตว่างคือ  $U$  และ  $A$ , โดย  $U$  คือเอกภพสัมพัทธ์ และ  $A$  คือเซตของแอตทริบิว แอตทริบิว  $a \in A$  เราสามารถหาค่าแอตทริบิว  $V_a$  เรียกว่า “โดเมนของ  $a$ ” สับเซต  $B$  ใดๆ ของ  $A$  เป็น binary relation  $I(B)$  บน  $U$  ซึ่งเรียกว่า “indiscernibility relation” จะต้องเป็นไปตามเงื่อนไขดังนี้

$xI(B)y$  ก็ต่อเมื่อ  $a(x) = a(y)$  สำหรับทุก  $a \in A$ , โดย  $a(x)$  คือค่าแอตทริบิว  $a$  ของ element  $x$

$I(B)$  คือ equivalence relation, กลุ่มของ equivalence classes ทั้งหมดของ  $I(B)$  ซึ่งเป็นการจัดกลุ่มโดยอาศัยสับเซต  $B$  สามารถเขียนได้เป็น  $U/I(B)$  หรือเขียนในรูปอย่างง่ายคือ  $U/B$  สำหรับ equivalence class ของ  $I(B)$  คือกลุ่มของการแบ่ง  $U/B$  จะมีสมาชิก  $x$  ซึ่งแสดงในรูปของ  $B(x)$

ถ้า  $(x, y)$  เป็นส่วนหนึ่งของ  $I(B)$  เรากล่าวได้ว่า  $x$  และ  $y$  เป็น  $B$ -indiscernible กัน เราเรียก equivalence classes ของความสัมพันธ์  $I(B)$  (หรือกลุ่มของการแบ่ง  $U/B$ ) ว่า “ $B$ -elementary sets” กำหนดตัวดำเนินการของเซตได้ดังนี้

$$B_*(X) = \{x \in U : B(x) \subseteq X\}$$

$$B^*(X) = \{x \in U : B(x) \cap X \neq \emptyset\}$$

กำหนดให้ทุกสับเซต  $X$  ในเอกภพสัมพัทธ์  $U$  ประกอบไปด้วยเซต  $B_*(X)$  และ  $B^*(X)$  เรียกว่า “B-upper” และ “B-lower approximation ของ  $X$ ” ตามลำดับ และเซต boundary region ของ  $X$  คือ

$$BN_B(X) = B^*(X) - B_*(X)$$

ถ้าบริเวณ boundary ของ  $X$  เป็นเซตว่าง นั่นคือ ถ้า  $BN_B(X) = \emptyset$  แล้วเซต  $X$  เป็น crisp โดยอาศัยสับเซต  $B$  ในทางกลับกัน ถ้าบริเวณ boundary ของ  $X$  ไม่เป็นเซตว่าง นั่นคือ ถ้า  $BN_B(X) \neq \emptyset$  แล้วเซต  $X$  เป็น rough โดยอาศัยสับเซต  $B$

เราสามารถคำนวณค่าสัมประสิทธิ์ของราฟเซต จากสมการที่ 3.1

$$\alpha_B(X) = \frac{|B_*(X)|}{|B^*(X)|} \quad (3.1)$$

เรียกว่า ค่าประมาณของความถูกต้อง(accuracy of approximation) โดย  $|X|$  คือจำนวนสมาชิกของ  $X$  ค่าประมาณของความถูกต้องมีค่าอยู่ระหว่าง 0 ถึง 1 ( $0 \leq \alpha_B(X) \leq 1$ ) ถ้า  $\alpha_B(X) = 1$  แล้ว  $X$  เป็น crisp เมื่ออาศัยสับเซต  $B$  ( $X$  มีความถูกต้องเมื่อพิจารณาจากสับเซต  $B$ ) และถ้า  $\alpha_B(X) < 1$  แล้ว  $X$  เป็น rough เมื่อพิจารณาจากสับเซต  $B$  ( $X$  มีความคลุมเครือเมื่อพิจารณาจากสับเซต  $B$ )

จากตัวอย่างตารางที่ 3.1 เมื่อพิจารณา concept “ไข้หวัดใหญ่” ประกอบด้วยสมาชิก  $\{p1, p2, p3, p6\}$  และเซตของแอตทริบิว  $B = \{\text{ปวดหัว, ปวดกล้ามเนื้อ, อุณหภูมิ}\}$  จะได้ว่า concept “ไข้หวัดใหญ่” เป็น roughly B-definable เพราะว่า  $B_*(X) = \{p1, p3, p6\} \neq \emptyset$  และ  $B^*(X) = \{p1, p2, p3, p5, p6\} \neq U$  สำหรับกรณีนี้ จะได้  $\alpha_B(\text{“ไข้หวัดใหญ่”}) = 3/5$  หมายความว่าสามารถแสดง concept “ไข้หวัดใหญ่” ได้เพียงบางส่วนเมื่อพิจารณาจากอาการปวดหัว, ปวดกล้ามเนื้อ และอุณหภูมิ เมื่อเราพิจารณาเพียง 1 อาการ คือ  $B = \{\text{ปวดหัว}\}$  จะได้  $B_*(X) = \emptyset$  และ  $B^*(X) = U$  ซึ่งหมายความว่า concept “ไข้หวัดใหญ่” เป็น totally indefinable ต่อเทอมของอาการปวดหัว แสดงว่าแอตทริบิวนี้ไม่มีคุณลักษณะในการจำแนกอาการไข้หวัดใหญ่ อย่างไรก็ตามสำหรับกรณีแอตทริบิว  $B = \{\text{อุณหภูมิ}\}$  เราจะได้  $B_*(X) = \{p3, p6\}$  และ  $B^*(X) = \{p1, p2, p3, p5, p6\}$  ดังนั้น concept “ไข้หวัดใหญ่” สำหรับกรณีนี้ จะเป็น roughly definable แต่ในกรณีนี้ เราได้ค่า  $\alpha_B(\text{“ไข้หวัดใหญ่”}) = 2/5$  หมายความว่า การวิเคราะห์อุณหภูมิเพียงอย่างเดียว จะสามารถวินิจฉัยอาการไข้หวัดใหญ่ได้น้อยกว่าการวิเคราะห์ด้วยเซตของอาการทั้งหมด และคนไข้  $p1$  ไม่สามารถจำแนกว่าเป็นไข้หวัดใหญ่ได้ในกรณีนี้

### 3.3 ค่าฟังก์ชันความเป็นสมาชิกของกราฟเซต

concepts ที่มีความคลุมเครือซึ่งเป็นกรณีของ boundary line นั่นคือ elements ในเอกภพสัมพัทธ์ที่ไม่สามารถจำแนกได้แน่นอนว่าเป็น elements ของ concept นั้นหรือไม่ เพราะฉะนั้น ความไม่แน่นอนนี้ จะมีความเกี่ยวข้องกับค่าฟังก์ชันความเป็นสมาชิกของ elements ในเซต ซึ่งเราสามารถหาค่าฟังก์ชันความเป็นสมาชิกของกราฟเซตได้จากสมการ 3.2

$$\mu_x^B(x) = \frac{|X \cap B(x)|}{|B(x)|} \quad (3.2)$$

และ  $\mu_x^B \in [0, 1]$  เราสามารถนำค่าฟังก์ชันความเป็นสมาชิกของกราฟเซตมาจำกัดความ approximations และขอบเขตของเซตได้

$$\begin{aligned} B_*(X) &= \{x \in U : \mu_x^B = 1\} \\ B^*(X) &= \{x \in U : \mu_x^B > 0\} \\ BN_B(X) &= \{x \in U : 0 < \mu_x^B < 1\} \end{aligned}$$

### 3.4 การขึ้นต่อกันของแอดทริบิว

สิ่งหนึ่งที่สำคัญในการวิเคราะห์ข้อมูลคือ การขึ้นต่อกันระหว่างแอดทริบิวต่างๆ จะเห็นได้ว่าเซตของแอดทริบิว D ขึ้นอยู่กับเซตของแอดทริบิว C ทั้งหมด สามารถเขียนได้เป็น  $C \Rightarrow D$  เมื่อแต่ละค่าของแอดทริบิว D สามารถวิเคราะห์ได้จากค่าของแอดทริบิวที่มีใน C เพียงค่าใดค่าหนึ่ง ตัวอย่างในตารางที่ 3.1 ไม่ได้เป็นการขึ้นต่อกันทั้งหมด ถ้าในตารางที่ 3.1, อุณหภูมิของคนไข้ p5 เป็น “ปกติ” แทนที่ “สูง” จะทำให้ {อุณหภูมิ}  $\Rightarrow$  {ไข้หวัดใหญ่} เพราะว่าคุณสมบัติแต่ละค่าสามารถนำไปวิเคราะห์ว่ามีอาการไข้หวัดใหญ่หรือไม่เพียงอย่างเดียว

ดังนั้นสามารถให้คำจำกัดความของการขึ้นต่อกันได้ว่า ให้ D และ C เป็นสับเซตของ A เรากล่าวได้ว่าสับเซต D ขึ้นต่อสับเซต C ทั้งหมด, ก็ต่อเมื่อ  $I(C) \subseteq I(D)$  หมายความว่า การแบ่งกลุ่มของสมาชิกโดยอาศัยสับเซต C มีความละเอียดกว่าการแบ่งกลุ่มของสมาชิกโดยอาศัยสับเซต D เราสามารถหาการขึ้นต่อกันเพียงส่วนย่อย จากตัวอย่างในตารางที่ 3.1 จะเห็นได้ว่ามีเพียงบางค่าของอุณหภูมิที่สามารถตรวจสอบอาการไข้หวัดใหญ่ได้ นั่นคือ (อุณหภูมิ, สูงมาก) แสดงว่า (ไข้หวัดใหญ่, ไข้) ในทำนองเดียวกัน (อุณหภูมิ, ปกติ) แสดงว่า (ไข้หวัดใหญ่, ไม่ไข้) แต่ (อุณหภูมิ, สูง) ไม่ได้หมาย

ความว่า (ไข้หวัดใหญ่, ไข้) เสมอไป ดังนั้นการขึ้นต่อกันเพียงส่วนย่อยแสดงว่าค่าของ  $D$  บางค่าสามารถหาได้จากค่าของ  $C$

ให้  $D$  และ  $C$  เป็นสับเซตของ  $A$ , สับเซต  $D$  ขึ้นอยู่กับสับเซต  $C$  ด้วยค่าคิกริ  $k$ ,  $0 \leq k \leq 1$  เขียนได้เป็น  $C \Rightarrow_k D$  ถ้า

$$k = \frac{|POS_C(D)|}{|U|} \quad (3.3)$$

โดย

$$POS_C(D) = \bigcup_{X \in U/I(D)} C_*(X)$$

เรียกว่า “positive region” ของการแบ่ง  $U/D$  ซึ่งสอดคล้องกับสับเซต  $C$  นั่นคือเซตของ elements ในเอกภพสัมพัทธ์ทั้งหมดจะจำแนกไปอยู่กลุ่มใดกลุ่มหนึ่งของการแบ่ง  $U/D$  โดยอาศัยสับเซต  $C$

ดังนั้นค่าสัมประสิทธิ์  $k$  แสดงอัตราส่วนของ elements ทั้งหมดในเอกภพสัมพัทธ์ที่สามารถแบ่งกลุ่มได้เป็น  $U/D$  โดยการใช้แอตทริบิว  $C$  สังเกตว่า สำหรับ  $k = 1$ , จะเป็นการขึ้นต่อกันทั้งหมด

สำหรับ {ปวดหัว, ปวดกล้ามเนื้อ, อุณหภูมิ}  $\Rightarrow$  {ไข้หวัดใหญ่} เราจะได้  $k = 4/6 = 2/3$  เพราะวคนไข้ 4 ใน 6 คนสามารถจำแนกได้ว่าเป็นไข้หวัดใหญ่หรือไม่ โดยใช้อาการปวดหัว, ปวดกล้ามเนื้อ และอุณหภูมิ

ถ้าเราต้องการวินิจฉัยคนไข้โดยอาศัยเพียงอุณหภูมิอย่างเดียว จะได้ {อุณหภูมิ}  $\Rightarrow$  {ไข้หวัดใหญ่} มีค่า  $k = 3/6 = 1/2$  เพราะวในกรณีนี้มีเพียงคนไข้  $p_3, p_4$  และ  $p_6$  ในจำนวนทั้งหมดที่สามารถจำแนกได้ว่าเป็นไข้หวัดใหญ่หรือไม่ เมื่อเปรียบเทียบกับกรณีที่ผ่านมา จะเห็นได้ว่าไม่สามารถวินิจฉัยได้ว่าคนไข้  $p_1$  เป็นไข้หวัดใหญ่หรือไม่ ดังนั้นกรณีที่วิเคราะห์จากอุณหภูมิเพียงอย่างเดียว จะมีความสามารถในการจำแนกต่ำกว่าการวิเคราะห์จากเซตของอาการปวดหัว, ปวดกล้ามเนื้อ และอุณหภูมิ กรณีที่สังเกตเฉพาะอาการปวดหัว หรือปวดกล้ามเนื้อเพียงอย่างเดียวเพื่อวินิจฉัยอาการไข้หวัดใหญ่ นั่นคือ {อาการปวดหัว}  $\Rightarrow$  {ไข้หวัดใหญ่} และ {อาการปวดกล้ามเนื้อ}  $\Rightarrow$  {ไข้หวัดใหญ่} จะได้ค่าสัมประสิทธิ์  $k = 0$  สรุปได้ว่าสับเซต  $D$  ขึ้นต่อสับเซต  $C$  ทั้งหมด (หรือส่วนย่อย) ถ้า elements ทั้งหมด (หรือบางส่วน) ของเอกภพสัมพัทธ์  $U$  สามารถจำแนกได้อยู่กลุ่มของ  $U/D$  เพียงกลุ่มใดกลุ่มหนึ่ง เมื่ออาศัยสับเซต  $C$  ในการวิเคราะห์

### 3.5 การลดแอดทริบิวที่ไม่จำเป็น

เรามักจะพบคำถามว่า เราสามารถลบข้อมูลบางอย่างจากตารางข้อมูล โดยยังคงคุณสมบัติพื้นฐานเหมือนเดิมได้หรือไม่ นั่นคือ ตารางข้อมูลนั้นมีข้อมูลบางตัวที่ไม่จำเป็น ตัวอย่างเช่น ในตารางที่ 3.1 ถ้าเราลบอาการปวดหัวหรือปวดกล้ามเนื้อแล้วเรายังได้เซตเหมือนเดิมซึ่งมีความถูกต้องของค่า approximation และค่าความสัมพันธ์เหมือนเดิม โดยมีจำนวนของแอดทริบิวลดลง

กำหนดให้สับเซต  $D$  ขึ้นอยู่กับสับเซต  $C$  ถ้าหาสับเซตที่เล็กที่สุดของ  $C$  คือ  $C'$  โดยที่สับเซต  $D$  ยังคงขึ้นอยู่กับ  $C'$  เราจะเรียก  $C'$  ว่า “reduct ของสับเซต  $C$ ”

ดังนั้น reduct คือเซตของแอดทริบิวที่ยังคงรักษาค่าความสัมพันธ์ หมายความว่า reduct เป็นสับเซตที่มีจำนวนแอดทริบิวน้อยที่สุดที่มีความสามารถในการจำแนก elements ในเอกภพสัมพัทธ์ได้เหมือนเซตของแอดทริบิวทั้งหมดที่ยังไม่ได้ถูกลด

reduct อาจจะมีมากกว่าหนึ่ง แอดทริบิวที่มีอยู่ร่วมกันของ reduct ทั้งหมดเรียกว่า “core” จากตารางที่ 3.1 มี core คือ อุณหภูมิ ดังนั้นจะเห็นได้ว่า core คือแอดทริบิวที่สำคัญ ซึ่งไม่สามารถลดออกไปได้จากตารางข้อมูล โดยที่ไม่เปลี่ยนการขึ้นต่อกันของข้อมูล

#### 3.5.1 ค่าความสำคัญของแอดทริบิว

แอดทริบิวที่ต่างกันจะมีการขึ้นต่อกันระหว่างแอดทริบิวเงื่อนไข และแอดทริบิวการตัดสินใจที่แตกต่างกัน

ค่าความสำคัญของแต่ละแอดทริบิว  $a$  ที่เพิ่มเข้าไปในสับเซต  $C$  สามารถหาได้จากสมการที่ 3.4

$$SGF(a,C,D) = k(C+\{a\},D) - k(C,D) \quad (3.4)$$

$SGF(a,C,D)$  คือค่าการขึ้นต่อกันที่เพิ่มขึ้น เมื่อเพิ่มแอดทริบิว  $a$  เข้าไปในแอดทริบิวเงื่อนไข

#### 3.5.2 การหา reduct ที่ดีที่สุด

reduct อาจจะมีมากกว่าหนึ่ง ซึ่งแต่ละ reduct สามารถนำมาใช้แทนกลุ่มของแอดทริบิวทั้งหมดที่ยังไม่ได้ถูกลด โดยที่ไม่เปลี่ยนแปลงค่าการขึ้นต่อกันของข้อมูล ดังนั้นจึงเกิดคำถามขึ้นว่า reduct ใดดีที่สุด การหา reduct ที่ดีที่สุดขึ้นอยู่กับเกณฑ์ในการเลือกแอดทริบิว โดยเกณฑ์ที่ใช้ในการ reduct คือจะต้องมีจำนวนของแอดทริบิวน้อยที่สุด และแต่ละแอดทริบิวที่ถูกเลือกจะมีจำนวนค่าแอดทริบิวน้อยที่สุด

### 3.5.3 Discernibility Matrix

Discernibility Matrix เป็นเมตริกซ์ที่แต่ละ element ของเมตริกซ์แสดงความแตกต่างกันของ 2 objects ในตารางข้อมูล เพราะฉะนั้นการแสดงความแตกต่างกันของ object ที่อยู่คนละกลุ่มในรูปของ Discernibility matrix ทำให้เราสามารถหา Core ได้สะดวกขึ้น ซึ่งนิยาม Discernibility matrix ของแอตทริบิวต์เงื่อนไข C ในตารางข้อมูล S,  $M(C) = \{m_{ij}\}_{n \times n}$  คือ

$$(m_{ij}) = \begin{cases} \Phi & \text{if } x_i, x_j \in \text{same concepts} \\ \{c \in C : f(c, x_i) \neq f(c, x_j)\} & \text{if } x_i, x_j \in \text{different concepts} \end{cases} \quad (3.5)$$

สมการที่ 3.5,  $f(c, x_i)$  คือค่าของแอตทริบิวต์ จะเห็นได้ว่า  $m_{ij}$  แสดงความแตกต่างกันของ  $x_i$  และ  $x_j$  โดยที่  $x_i$  และ  $x_j$  อยู่คนละกลุ่ม  $M(S) = (m_{ij})$  เป็นเมตริกซ์ที่ symmetric จากตารางที่ 3.1 เราสามารถคำนวณค่า  $m_{ij}$  ได้ดังแสดงในตารางที่ 3.2

### 3.5.4 Core และ Discernibility Matrix

Core เป็นส่วนที่สำคัญในกราฟเซตซึ่ง core เป็นแอตทริบิวต์ร่วมกันของทุก reduct ดังนั้นสามารถนำ core มาคำนวณหา reduct และเราสามารถหา core ได้จาก Discernibility Matrix โดย core เป็นค่าของ element ที่มีความแตกต่างกันเพียงหนึ่งค่า นั่นคือเป็นแอตทริบิวต์ที่ไม่สามารถลดได้ และ core สามารถที่จะเป็นเซตว่างได้ จากตารางที่ 3.1 เราสามารถหา core ได้ จากตารางที่ 3.2 จะเห็นได้ว่า  $m_{5,1} = \{t\}$  และ  $m_{5,4} = \{t\}$  ดังนั้น core ของแอตทริบิวต์ทั้งหมดคือ {อุณหภูมิจ}

ตารางที่ 3.2 แสดง Discernibility Matrix

Obj.	P1	p2	p3	p6	p4	p5
P1						
P2						
P3						
P6						
P4	t	hmt	ht	t		
P5	hm		mt	hmt		

ตัวย่อ: h: อาการปวดหัว, m: อาการปวดกล้ามเนื้อ, t: อุณหภูมิ

### 3.5.5 อัลกอริทึมสำหรับคำนวณ reduct ที่ดีที่สุด หรือเซตของแอตทริบิวทีวี่น้อยที่สุด

เราสามารถหา reduct ที่ดีที่สุด หรือเซตของแอตทริบิวทีวี่น้อยที่สุด (ซึ่งมีการกำหนดแอตทริบิวทีวี่ที่มีความสำคัญ) ได้จากค่าการขึ้นต่อกัน และค่าความสำคัญของแอตทริบิวทีวี่ ดังนั้น reduct ที่ดีที่สุดนี้ จะประกอบไปด้วยแอตทริบิวทีวี่จำนวนที่น้อยที่สุดที่ไม่เปลี่ยนแปลงค่าการขึ้นต่อกัน

อัลกอริทึมสำหรับคำนวณ reduct ที่ดีที่สุด จะเริ่มต้นจาก core ถ้าไม่กำหนดแอตทริบิวทีวี่ที่มีความสำคัญใดๆ แล้วอัลกอริทึมนี้จะคำนวณหา reduct ซึ่งประกอบไปด้วยแอตทริบิวทีวี่ที่มีค่าความสำคัญมากที่สุด แต่ถ้ากำหนดแอตทริบิวทีวี่ที่มีความสำคัญแล้วอัลกอริทึมนี้จะหาจำนวนแอตทริบิวทีวี่น้อยที่สุดที่รวมแอตทริบิวทีวี่ซึ่งกำหนดเพิ่มเข้าไป โดยผลที่ได้จะเป็นฐานความรู้ที่ไม่เปลี่ยนแปลงฐานความรู้เดิม

**Reduct Algorithm:** คำนวณหา reduct ที่ดีที่สุด หรือเซตของแอตทริบิวทีวี่น้อยที่สุด

Input: (i) เซตของแอตทริบิวทีวี่ AR ซึ่งประกอบไปด้วยแอตทริบิวทีวี่เงื่อนไข และการตัดสินใจ

(ii) CO ของ AR คือแอตทริบิวทีวี่ที่เป็น core คำนวณได้จาก Discernibility Matrix

(iii) แอตทริบิวทีวี่ UA คือแอตทริบิวทีวี่ที่กำหนดให้มีความสำคัญ (UA อาจจะเป็นเซตว่าง ถ้า UA เป็นเซตว่าง นั่นคือไม่มีการกำหนดแอตทริบิวทีวี่ใดๆ ในการหา Reduct)

Output: เซตของแอตทริบิวทีวี่ REDU

ขั้นตอน

ขั้นที่ 1:  $REDU = CO \cup UA;$

ขั้นที่ 2:  $AR' = AR - REDU$

ขั้นที่ 3: คำนวณค่าความสำคัญของแต่ละแอตทริบิวทีวี่  $a \in AR'$  เรียงลำดับเซตของแอตทริบิวทีวี่  $AR'$  ตามค่าความสำคัญ

ขั้นที่ 4: หาเซต REDU ของแอตทริบิวทีวี่ AR ด้วยการเพิ่มทีละแอตทริบิวทีวี่

ถ้า  $K(REDU, D) \neq K(AR, D)$  แล้ว เลือกแอตทริบิวทีวี่  $a$  ใน  $AR'$  ที่มีความสำคัญมากที่สุด (ถ้า  $a_i$  ( $i = 1, \dots, m$ ) มีค่า  $SGF(a, REDU, D)$  มากที่สุด แล้วเลือกแอตทริบิวทีวี่  $a_i$  ที่มีจำนวนค่าแอตทริบิวทีวี่น้อยที่สุด) คำนวณค่าการขึ้นต่อกัน  $K(REDU, D)$  แล้วกลับไปทำขั้นที่ 4

ขั้นที่ 5:  $|REDU| \rightarrow N$

ขั้นที่ 6: หาเซต reduct ที่ดีที่สุด หรือเซตของแอตทริบิวทีวี่น้อยที่สุด โดยทดสอบลดแอตทริบิวทีวี่

วนรอบ  $i = 0$  ถึง  $N-1$

ถ้า  $a_i$  ไม่ใช่แอตทริบิวทีวี่  $CO \cup UA$  แล้วลบ  $a_i$  ออกจาก REDU

หลังจากนั้น ถ้า  $K(REDU, D) \neq K(AR, D)$  แล้ว  $REDU \cup a_i \rightarrow REDU$

สิ้นสุดกระบวนการ

อัลกอริทึมนี้จะคำนวณหาค่าความสำคัญของแต่ละแอดทริบิวและเรียงลำดับแอดทริบิวตามค่าความสำคัญ โดยผลของอัลกอริทึมนี้จะได้เซตของแอดทริบิวที่น้อยที่สุด ซึ่งมีความสามารถในการจำแนกเหมือนเซตของแอดทริบิวทั้งหมด

### 3.6 กฎการตัดสินใจ

ในตารางข้อมูลจะสังเกตได้ว่ามีแอดทริบิวสองกลุ่ม คือ แอดทริบิวของเงื่อนไข และการตัดสินใจ เมื่อพิจารณาตัวอย่างตารางที่ 3.1 อาการปวดหัว, ปวดกล้ามเนื้อ และ อุณหภูมิ เป็นแอดทริบิวของเงื่อนไข โดยอาการไข้หวัดใหญ่ เป็นแอดทริบิวของการตัดสินใจ

แต่ละแถวของตารางนี้จะนำมาสร้างกฎการตัดสินใจ ซึ่งมีการตัดสินใจเมื่อเงื่อนไขที่ถูกต้อง ตัวอย่างในตารางที่ 3.1, จากเงื่อนไข (ปวดหัว, ไม่ใช่), (ปวดกล้ามเนื้อ, ใช่), (อุณหภูมิ, สูง) จะตัดสินใจว่า (ไข้หวัดใหญ่, ใช่) กฎการตัดสินใจที่ 2 และ 5 ในตารางที่ 3.1 มีเงื่อนไขเดียวกัน แต่ตัดสินใจต่างกัน กฎลักษณะเรียกว่า “กฎที่ไม่ถูกต้อง” (จำแนกไม่ได้, เกิดความขัดแย้ง) ส่วนกฎอื่นเรียกว่า “กฎที่ถูกต้อง” (จำแนกได้, ไม่เกิดความขัดแย้ง) บางครั้ง กฎการตัดสินใจที่ถูกต้องเรียกว่า “กฎที่แน่นอน” และกฎที่ไม่ถูกต้องเรียกว่า “กฎที่อาจจะเป็นไปได้” ตารางที่มีการตัดสินใจซึ่งมีกฎการตัดสินใจที่ไม่ถูกต้องเรียกว่า “ตารางที่มีความไม่ถูกต้อง” ส่วนกรณีอื่นจะเรียกว่า “ตารางที่มีความถูกต้อง”

จำนวนของกฎที่ถูกต้องในจำนวนกฎทั้งหมดในตารางที่มีการตัดสินใจสามารถนำมาวัดความถูกต้องของตาราง และใช้เครื่องหมาย  $\gamma(C, D)$  แทน โดย C และ D คือแอดทริบิวของเงื่อนไข และการตัดสินใจ ตามลำดับ ดังนั้นถ้า  $\gamma(C, D) = 1$  แล้วตารางการตัดสินใจมีความถูกต้อง และถ้า  $\gamma(C, D) \neq 1$  แล้วตารางการตัดสินใจมีความไม่ถูกต้อง จะเห็นได้ว่า

$$\gamma(C, D) = \frac{POS_C(D)}{|U|} \quad (3.6)$$

กฎการตัดสินใจจะเป็นลักษณะของ “ถ้า... แล้ว...” สำหรับตัวอย่างกฎของที่ 1 ของตารางที่ 1 สามารถแสดงได้เป็น

ถ้า (ปวดหัว, ไม่ใช่) และ (ปวดกล้ามเนื้อ, ใช่) และ (อุณหภูมิ, สูง) แล้ว (ไข้หวัดใหญ่, ใช่) เพราะฉะนั้น ตารางที่มีการตัดสินใจสามารถแสดงให้อยู่ในรูปเซตของกฎการตัดสินใจ สำหรับทุกการขึ้นต่อกันไม่ว่าจะเป็นทั้งหมด หรือส่วนย่อย เราสามารถสร้างเซตของกฎการตัดสินใจที่หาการขึ้นต่อกันได้ นอกจากการลดคุณสมบัติของเงื่อนไขในตารางที่มีการตัดสินใจ เราสามารถลดคุณสมบัติในแต่ละกฎการตัดสินใจเพื่อให้มีกฎการตัดสินใจที่เล็กที่สุด สำหรับตัวอย่างการขึ้นต่อกันในส่วนย่อย

{ปวดหัว, ปวดกล้ามเนื้อ, อุณหภูมิ}  $\Rightarrow$  {ไข้หวัดใหญ่}

เราสามารถหากฎการตัดสินใจที่มีขนาดเล็กที่สุด

ถ้า (อุณหภูมิ, ปกติ) แล้ว (ไข้หวัดใหญ่, ไข้)

ถ้า (ปวดหัว, ไม่ใช่) และ (ปวดกล้ามเนื้อ, ไข้) หรือ (ปวดกล้ามเนื้อ, ไข้) และ (อุณหภูมิ, สูง) หรือ (อุณหภูมิ, สูงมาก) แล้ว (ไข้หวัดใหญ่, ไข้)

### 3.7 สรุป

ทฤษฎีกราฟเซต เป็นวิธีทางคณิตศาสตร์ที่สามารถวิเคราะห์ความคลุมเครือของข้อมูล การหาขอบเขตบน และขอบเขตล่างของกลุ่มข้อมูล เพื่อนำมาคำนวณค่าการขึ้นต่อกันของข้อมูล และจากค่าของการขึ้นต่อกันนี้ เราสามารถนำไปวิเคราะห์หาแอดทริบิวต์ที่ไม่มีความจำเป็นต่อการจำแนกกลุ่มได้ ได้มีการพัฒนางานวิจัยทางด้านกราฟเซตไปใช้ในหลายสาขา เช่น ทางการแพทย์และเภสัชศาสตร์ได้นำกราฟเซตไปใช้ในการวินิจฉัยโรค ทางการเงินและการธนาคารได้นำกราฟเซตไปวิเคราะห์ความเสี่ยงทางการเงินหรือการลงทุน หรือแม้แต่ทางด้านวิศวกรรม ก็ได้มีการนำกราฟเซตไปใช้ในระบบควบคุมต่างๆ ซึ่งงานวิจัยนี้จะได้นำกราฟเซตมาใช้ในการเรียนรู้เพื่อจำแนกตัวอักษร

## บทที่ 4

# การประยุกต์ฟัซซีเซต และกราฟเซตร่วมกัน

งานวิจัยนี้เสนอการเลือกส่วนของข้อมูลลักษณะเด่นที่จำเป็นต่อการระบุตัวอักษรแต่ละตัว โดยใช้กราฟเซตในการกำจัดแอดทริบิวต์ที่ไม่จำเป็นต่อการรู้จำทิ้งไปและยังคงแอดทริบิวต์ที่มีความสามารถในการจำแนกไว้ สำหรับแอดทริบิวต์บางอย่างที่มีความต่อเนื่องของข้อมูลจะนำวิธีการจัดกลุ่มโดยฟัซซีเซตมาแบ่งช่วงของแอดทริบิวต์เพื่อให้เกิดการแบ่งช่วงของข้อมูลที่เหมาะสมที่สุด

### 4.1 การจัดกลุ่มข้อมูลโดยฟัซซีเซต

ข้อมูลที่มีความต่อเนื่อง ตัวอย่างเช่น อุณหภูมิ ซึ่งมีการวัดค่าเป็นเลขจำนวนจริง สามารถนำมาแบ่งระดับได้มากกว่าหนึ่งระดับ หรืออัตราการเต้นของหัวใจ ถึงแม้ว่าจะมีค่าเป็นเลขจำนวนเต็ม แต่แพทย์วินิจฉัยว่าการเต้นของหัวใจไม่ว่าจะเป็น 68 หรือ 72 ครั้งต่อนาที ก็ถือว่าเป็นอัตราการเต้นของหัวใจที่เป็นปกติ แต่ถ้าการเต้นของหัวใจเป็น 48 หรือ 56 ครั้งต่อนาที จะวินิจฉัยได้ว่าเป็นอัตราการเต้นของหัวใจที่ต่ำ ในขณะที่ 120 ถึง 140 ครั้งต่อนาทีจะเป็นอัตราการเต้นของหัวใจที่สูงมาก จะเห็นได้ว่าผลของการแบ่งช่วงจะทำให้สูญเสียค่าของข้อมูลไป แต่จะทำให้เห็นคุณลักษณะของข้อมูลได้มากขึ้น ดังนั้นการกำหนดช่วงของข้อมูลมีความสำคัญ การแบ่งกลุ่มข้อมูลโดยฟัซซีเซตจะใช้อัลกอริทึม Fuzzy c-Means ซึ่งได้อธิบายรายละเอียดไว้ในหัวข้อที่ 3.4

**อัลกอริทึม FCM : Fuzzy c-Means Algorithm**

ขั้นที่ 1: กำหนดจำนวนกลุ่ม  $c$  ( $2 \leq c \leq n$ ) และค่าพารามิเตอร์  $m$  ซึ่งเป็นค่า exponential weight ( $1 < m < \infty$ ) กำหนดค่าเมตริกซ์เริ่มต้น  $U^{(0)}$  และค่า  $\epsilon$  ซึ่งเป็นค่าผิดพลาดที่ยอมรับได้ โดยเริ่มต้นการวนรอบแรกเป็น 0

ขั้นที่ 2: คำนวณจุดศูนย์กลางของแต่ละกลุ่ม  $\{v_i^1 \mid i = 1, 2, \dots, c\}$  โดยอาศัยค่า  $U^{(0)}$  และสมการที่ 4.1 ซึ่งทุกกลุ่มเป็นการจำแนกแบบฟัซซี

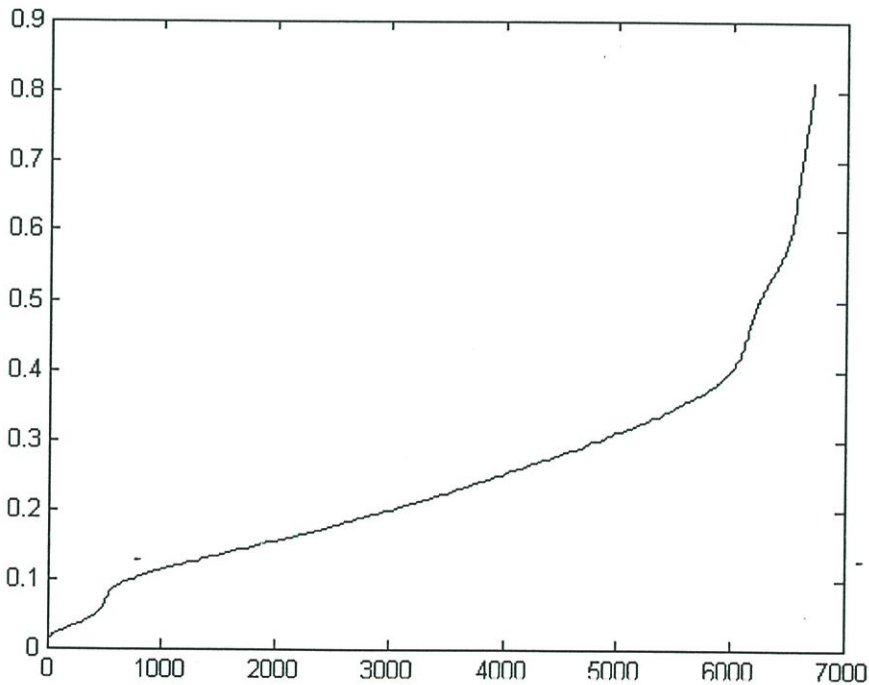
$$v_i = \frac{1}{\sum_{j=1}^n (u_{ij})^m} \sum_{j=1}^n (u_{ij})^m x_j, \quad i = 1, 2, \dots, c \quad (4.1)$$

ขั้นที่ 3: คำนวณหาค่าความเป็นสมาชิกของค่า  $U^{(l+1)}$  ซึ่งเป็นค่า  $U$  ในการวนรอบถัดไป โดยใช้จุดศูนย์กลาง  $\{v_i \mid i = 1, 2, \dots, c\}$  และสมการที่ 4.2

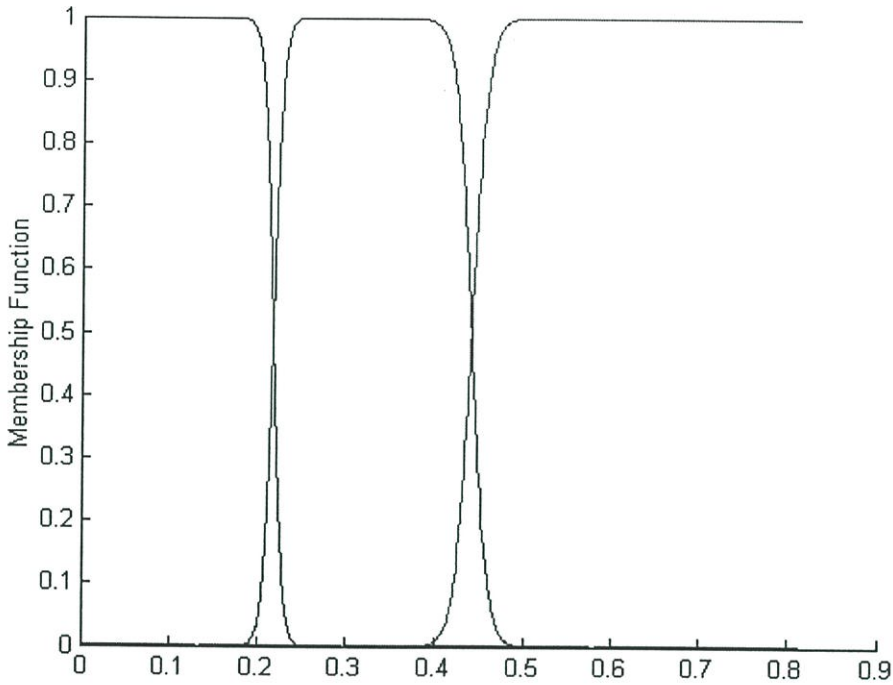
$$u_{ij} = \frac{(1/\|x_j - v_i\|^2)^{1/m-1}}{\sum_{k=1}^c (1/\|x_j - v_k\|^2)^{1/m-1}}, \quad i = 1, 2, \dots, c; j = 1, 2, \dots, n \quad (4.2)$$

ขั้นที่ 4: คำนวณค่า  $\Delta = \|U^{(l+1)} - U^{(l)}\| = \max_{i,j} |u_{ij}^{(l+1)} - u_{ij}^{(l)}|$  ถ้า  $\Delta > \epsilon$  ให้  $l = l+1$  และกลับไปทำขั้นที่ 2 แต่ถ้า  $\Delta \leq \epsilon$  ให้สิ้นสุดการวนรอบการทำงาน

รูปที่ 4.1 เป็นตัวอย่างข้อมูลของแอตทริบิวต์ที่ข้อมูลมีความต่อเนื่อง เมื่อนำมาจัดกลุ่มโดยใช้ฟังก์ชัน Fuzzy c-means สามารถแสดงกราฟค่าความเป็นสมาชิกของ  $U$  ดังรูปที่ 4.2



รูปที่ 4.1 ตัวอย่างข้อมูลของแอตทริบิวต์ที่มีความต่อเนื่อง



รูปที่ 4.2 กราฟค่าความเป็นสมาชิกของ U

ผลของการจำแนกด้วยอัลกอริทึม Fuzzy c-means แสดงแบ่งกลุ่มและแสดงค่าความเป็นสมาชิกตามจำนวนกลุ่มที่กำหนด แต่สามารถวัดค่าความเป็นสมาชิกของ U ในการแบ่งแต่ละกลุ่มเพื่อประเมินการแบ่งกลุ่มที่มีความเหมาะสม ด้วยคุณสมบัติตามสมการที่ 4.3 และ 4.4

Partition Entropy

$$H(U, c) = \frac{1}{n} \sum_{j=1}^n \sum_{i=1}^c |u_{ij} \ln u_{ij}| \quad (4.3)$$

Partition coefficient

$$F(U, c) = \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n u_{ij}^2 \quad (4.4)$$

จากคุณสมบัติในสมการที่ 4.3 และ 4.4 จะเห็นได้ว่า ค่า  $H = 0$  และ  $F = 1$  ถ้า  $u_{ij} \in \{0,1\}$  (การแบ่งกลุ่มข้อมูลโดยวิธี Hard) ค่า  $H = \ln c$  และ  $F = 1/c$  ถ้า  $u_{ij} = 1/c$  สำหรับ  $i$  และ  $j$  ทั้งหมด นั่นคือ  $0 \leq H \leq \ln c$  และ  $1/c \leq F \leq 1$  จะเห็นได้ว่า  $H$  และ  $F$  ขึ้นอยู่กับใน  $cn$  ทั้งหมด การประเมินคุณสมบัติของ  $H$  และ  $F$  ในการแบ่งกลุ่มจำนวน  $c$  กลุ่มสามารถตรวจสอบจำนวนกลุ่มที่เหมาะสมได้ดังนี้

$$\min_{c=2}^{n-1} \left\{ \min_{U \in \Omega_c} [H(U, c)] \right\}$$

$$\max_{c=2}^{n-1} \left\{ \max_{U \in \Omega_c} [F(U, c)] \right\},$$

โดย  $\Omega_c$  คือเซตของคำตอบที่เหมาะสมที่สุดในการแบ่งจำนวน  $c$  กลุ่ม

จากตัวอย่างข้อมูลของแอดทรีที่มีความต่อเนื่องในรูปที่ 4.1 สามารถประเมินจำนวนกลุ่มที่เหมาะสมได้โดยวัดค่า Partition Entropy และ Partition coefficient ดังสมการที่ 4.3 และ 4.4 ตามลำดับ ตารางที่ 4.1 แสดงผลการประเมินจำนวนกลุ่มแต่ละกลุ่ม

ตารางที่ 4.1 ผลการประเมินจำนวนกลุ่ม

จำนวนกลุ่ม $c$	Partition Entropy	Partition coefficient
2	0.0630	0.9621
3	0.0595 ←	0.9640 ←
4	0.0869	0.9466
5	0.0916	0.9446
6	0.0875	0.9467
7	0.0804	0.9516

## 4.2 การจำแนกโดยกราฟเซต

จากทฤษฎีกราฟเซตสามารถนำมาเป็นวิธีการเลือกข้อมูลลักษณะเด่นที่จำเป็นต่อการระบุตัวอักษรแต่ละตัว ในการลดแอดทรีวิว หรือข้อมูลลักษณะเด่นที่ไม่จำเป็นต่อการรู้จำ โดยยังคงแอดทรีวิวที่มีความสามารถในการจำแนก

กำหนดให้สับเซต  $D$  ขึ้นอยู่กับสับเซต  $C$  ถ้าหาสับเซตที่เล็กที่สุดของ  $C$  คือ  $C'$  โดยที่สับเซต  $D$  ยังคงขึ้นอยู่กับ  $C'$  เราจะเรียก  $C'$  ว่า “reduct ของสับเซต  $C$ ” จากคุณสมบัติดังกล่าว จะสามารถวิเคราะห์หา reduct ซึ่งเป็นเซตของแอดทรีวิวที่ยังคงรักษาค่าความขึ้นต่อกัน นั่นคือ reduct เป็นสับเซตที่มีจำนวนแอดทรีวิวน้อยที่สุดที่มีความสามารถในการจำแนก elements ในเอกภพสัมพัทธ์ได้เหมือนเซตของแอดทรีวิวทั้งหมดที่ยังไม่ได้ถูกลด

reduct อาจจะมีมากกว่าหนึ่ง แอตทริบิวที่มีอยู่ร่วมกันของ reduct ทั้งหมดเรียกว่า “core” โดย core คือแอตทริบิวที่สำคัญ ซึ่งไม่สามารถลดยกออกไปได้จากตารางข้อมูลโดยที่ไม่เปลี่ยนการขึ้นต่อกันของข้อมูล เราสามารถหา core ได้ด้วย Discernibility Matrix

#### 4.2.1 การหา Core จาก Discernibility Matrix

การหา Discernibility matrix ของ  $C$  ใน  $S$ ,  $M(C) = \{m_{ij}\}_{n \times n}$  ดังสมการที่ 4.5

$$m_{ij} = \begin{cases} \Phi & \text{if } x_i, x_j \in \text{same concepts} \\ \{c \in C : f(c, x_i) \neq f(c, x_j)\} & \text{if } x_i, x_j \in \text{different concepts} \end{cases} \quad (4.5)$$

$m_{ij}$  แสดงความแตกต่างกันของ  $x_i$  และ  $x_j$  โดยที่  $x_i$  และ  $x_j$  อยู่คนละกลุ่ม  $M(S) = (m_{ij})$  เป็นเมทริกซ์ที่ symmetric

#### 4.2.2 reduct ที่ดีที่สุด หรือเซตของแอตทริบิวที่น้อยที่สุด

เราสามารถหา reduct ที่ดีที่สุด หรือเซตของแอตทริบิวที่น้อยที่สุด (ซึ่งมีการกำหนดแอตทริบิวที่มีความสำคัญ) ได้จากการวิเคราะห์ค่าการขึ้นต่อกัน และค่าความสำคัญของแอตทริบิว ดังนั้น reduct ที่ดีที่สุดนี้ จะประกอบไปด้วยแอตทริบิวจำนวนที่น้อยที่สุดที่ไม่เปลี่ยนแปลงค่าการขึ้นต่อกัน

เราสามารถคำนวณค่าการขึ้นต่อกันได้จากสมการที่ 4.6

$$k = \frac{|POS_C(D)|}{|U|} \quad (4.6)$$

โดยกำหนดให้  $D$  และ  $C$  เป็นสับเซตของแอตทริบิวทั้งหมดจำนวน  $A$  และ  $U$  เป็นเซตของเอกภพสัมพัทธ์ และ

$$POS_C(D) = \bigcup_{X \in U/II(D)} C_*(X)$$

เรียกว่า “positive region” ของการแบ่ง  $U/D$  ซึ่งสอดคล้องกับสับเซต  $C$  นั่นคือเซตของ elements ในเอกภพสัมพัทธ์ทั้งหมดจะจำแนกไปอยู่กลุ่มใดกลุ่มหนึ่งของการแบ่ง  $U/D$  โดยอาศัยสับเซต  $C$

แอตทริบิวต์ต่างกันจะมีค่าการขึ้นต่อกันระหว่างแอตทริบิวต์เงื่อนไข และแอตทริบิวต์การตัดสินใจที่แตกต่างกัน เราสามารถหาค่าความสำคัญของแต่ละแอตทริบิวต์  $a$  ที่เพิ่มเข้าไปในสับเซต  $C$  ได้จากสมการที่ 4.7

$$SGF(a,C,D) = k(C+\{a\},D) - k(C,D) \quad (4.7)$$

$SGF(a,C,D)$  คือค่าการขึ้นต่อกันที่เพิ่มขึ้น เมื่อเพิ่มแอตทริบิวต์  $a$  เข้าไปในแอตทริบิวต์เงื่อนไข

การคำนวณค่าการขึ้นต่อกัน และค่าความสำคัญของแอตทริบิวต์เป็นค่าที่สำคัญสำหรับคำนวณหา reduct ที่ดีที่สุด โดยอัลกอริทึมการหา reduct ที่ดีที่สุดจะเริ่มต้นจาก core ถ้าไม่กำหนดแอตทริบิวต์ที่มีความสำคัญใดๆ แล้วอัลกอริทึมนี้จะคำนวณหา reduct ซึ่งประกอบไปด้วยแอตทริบิวต์ที่มีค่าความสำคัญมากที่สุด แต่ถ้ากำหนดแอตทริบิวต์ที่มีความสำคัญแล้วอัลกอริทึมนี้จะหาจำนวนแอตทริบิวต์น้อยที่สุดที่รวมแอตทริบิวต์ซึ่งกำหนดเพิ่มเข้าไป โดยผลที่ได้จะเป็นฐานความรู้ที่ไม่เปลี่ยนแปลงฐานความรู้เดิม

**อัลกอริทึมการหา Reduct :** คำนวณหา reduct ที่ดีที่สุด หรือเซตของแอตทริบิวต์น้อยที่สุด

Input: (i) เซตของแอตทริบิวต์  $AR$  ซึ่งประกอบไปด้วยแอตทริบิวต์เงื่อนไข และการตัดสินใจ

(ii)  $CO$  ของ  $AR$  คำนวณได้จาก Discernibility Matrix

(iii) แอตทริบิวต์  $UA$  การกำหนดแอตทริบิวต์ที่มีความสำคัญ ( $UA$  อาจจะเป็นเซตว่าง ถ้า  $UA$  เป็นเซตว่าง นั่นคือไม่มีการกำหนดแอตทริบิวต์ใดๆ ในการหา Reduct)

Output: เซตของแอตทริบิวต์  $REDU$

ขั้นตอน

ขั้นที่ 1:  $REDU = CO \cup UA;$

ขั้นที่ 2:  $AR' = AR - REDU$

ขั้นที่ 3: คำนวณค่าความสำคัญของแต่ละแอตทริบิวต์  $a \in AR'$  เรียงลำดับเซตของแอตทริบิวต์  $AR'$  ตามค่าความสำคัญ

ขั้นที่ 4: หาเซต  $REDU$  ของแอตทริบิวต์  $AR$  ด้วยการเพิ่มทีละแอตทริบิวต์

ถ้า  $K(REDU,D) \neq K(AR,D)$  แล้ว เลือกแอตทริบิวต์  $a$  ใน  $AR'$  ที่มีค่าความสำคัญมากที่สุด (ถ้า  $a_i$  ( $i = 1, \dots, m$ ) มีค่า  $SGF(a_i, REDU, D)$  มากที่สุด แล้วเลือกแอตทริบิวต์  $a_j$  ที่มีจำนวนค่าแอตทริบิวต์น้อยที่สุด) คำนวณค่าการขึ้นต่อกัน  $K(REDU,D)$  แล้วกลับไปทำขั้นที่ 4

ขั้นที่ 5:  $|REDU| \rightarrow N$

ขั้นที่ 6: หาเซต reduct ที่ดีที่สุด หรือเซตของแอตทริบิวต์น้อยที่สุด โดยทดสอบลดแอตทริบิวต์

วนรอบ  $i = 0$  ถึง  $N-1$

ถ้า  $a_i$  ไม่ใช่แอตทริบิวต์  $CO \cup UA$  แล้วลบ  $a_i$  ออกจาก REDU

หลังจากนั้น ถ้า  $K(REDU, D) \neq K(AR, D)$  แล้ว  $REDU \cup a_i \rightarrow REDU$

สิ้นสุดกระบวนการ

อัลกอริทึมนี้จะคำนวณหาค่าความสำคัญของแต่ละแอตทริบิวต์และเรียงลำดับแอตทริบิวต์ตามค่าความสำคัญ โดยผลของอัลกอริทึมนี้จะได้เซตของแอตทริบิวต์ที่น้อยที่สุด ซึ่งมีความสามารถในการจำแนกเหมือนเซตของแอตทริบิวต์ทั้งหมด

#### 4.2.3 กฎการตัดสินใจ

ในตารางข้อมูลจะสังเกตได้ว่ามีแอตทริบิวต์สองกลุ่ม คือ แอตทริบิวต์ของเงื่อนไข และการตัดสินใจแต่ละแถวของตารางนี้จะนำมาสร้างกฎการตัดสินใจ ซึ่งมีการตัดสินใจเมื่อเงื่อนไขที่ถูกต้อง นั่นคือ ไม่มีความขัดแย้งเกิดขึ้น แต่กฎการตัดสินใจที่ต่างกัน นั่นคือ มีเงื่อนไขเดียวกันแต่มีการตัดสินใจที่ต่างกัน จะเป็นกฎที่เกิดความขัดแย้ง

จำนวนของกฎที่ถูกต้องในจำนวนกฎทั้งหมดในตารางที่มีการตัดสินใจสามารถนำมาวัดความถูกต้องของตาราง และใช้เครื่องหมาย  $\gamma(C, D)$  ในสมการที่ 4.8 แทน โดย  $C$  และ  $D$  คือแอตทริบิวต์ของเงื่อนไข และการตัดสินใจ ตามลำดับ ดังนั้นถ้า  $\gamma(C, D) = 1$  แล้วตารางการตัดสินใจมีความถูกต้อง และถ้า  $\gamma(C, D) \neq 1$  แล้วตารางการตัดสินใจมีความไม่ถูกต้อง จะเห็นได้ว่า

$$\gamma(C, D) = \frac{POS_C(D)}{|U|} \quad (4.8)$$

กฎการตัดสินใจจะเป็นลักษณะของ “ถ้า... แล้ว...” ตัวอย่างของกฎ สามารถแสดงได้เป็น

ถ้า (ปวดหัว, ไม่ใช่) และ (ปวดกล้ามเนื้อ, ใช่) และ (อุณหภูมิ, สูง) แล้ว (ใช้หัวดใหญ่, ใช่)

#### 4.3 สรุป

เนื่องจากข้อมูลลักษณะเด่นของตัวอักษรที่วัดมาได้ ประกอบด้วยลักษณะเด่นที่ไม่มีความต่อเนื่องของข้อมูลและลักษณะเด่นที่มีความต่อเนื่องของข้อมูลซึ่งการแบ่งกลุ่มของแอตทริบิวต์ที่มีความต่อเนื่องโดยฟuzzyเซต ทำให้ได้ช่วงการแบ่งข้อมูลที่เหมาะสม และลักษณะเด่นที่ไปได้ไปวิเคราะห์โดยกราฟเซต เพื่อหาแอตทริบิวต์ที่มีความจำเป็นต่อการจำแนกตัวอักษร

## บทที่ 5

# การเตรียมข้อมูล และกระบวนการรู้จำ

การเตรียมข้อมูลภาพของตัวอักษรตัวพิมพ์ภาษาไทย เป็นกระบวนการเบื้องต้นก่อนที่จะนำข้อมูลที่ได้สู่กระบวนการรู้จำตัวอักษร เมื่อป้อนข้อมูลตัวอักษรผ่านทางเครื่องสแกนภาพ ข้อมูลตัวอักษรที่ได้จะอยู่ในรูปของข้อมูลบิตแมป (Bitmap) ก่อนที่จะดำเนินกระบวนการต่อไป จะต้องทำการแปลงข้อมูลภาพตัวอักษรให้อยู่ในรูปข้อมูลแบบไบนารี (Binary) คือส่วนที่เป็นเนื้อภาพมีค่าเป็น 1 ส่วนที่เป็นพื้นหลังมีค่าเป็น 0 จากนั้นทำการวิเคราะห์หาลักษณะเด่นของตัวอักษรแบบต่างๆ ซึ่งลักษณะเด่นที่มีความต่อเนื่องของข้อมูลจะนำวิธีการจัดกลุ่มนำโดยพีชคณิตมาแบ่งช่วงของลักษณะเด่น เพื่อให้เกิดการแบ่งช่วงของข้อมูลที่เหมาะสมที่สุด และนำลักษณะเด่นทั้งหมดไปวิเคราะห์โดยใช้ราฟเซตเพื่อลดข้อมูลที่ไม่จำเป็นต่อการจำแนก เป็นขั้นตอนต่อไป

### 5.1 การหาลักษณะเด่นของตัวอักษร

การหาลักษณะเด่นของตัวอักษร จะวิเคราะห์ลักษณะเด่นหลายๆ แบบ สำหรับนำไปใช้ในกระบวนการรู้จำเพื่อเลือกลักษณะเด่นที่จำเป็นต่อการจำแนกตัวอักษร

#### 5.1.1 การหาจุดศูนย์กลางของภาพ

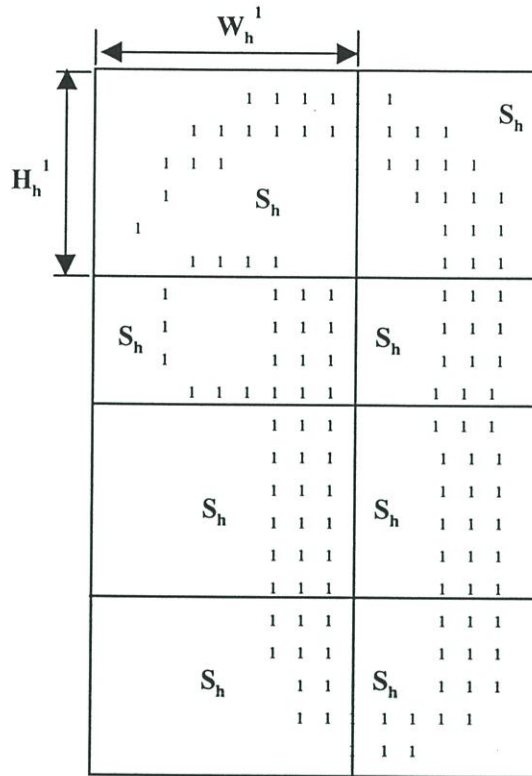
ภาพสองระดับที่มีจุดภาพมีค่าเป็น 1 เป็นเนื้อของภาพ และมีค่าเป็น 0 เป็นพื้นหลังของภาพ เราสามารถหาตำแหน่งที่เป็นจุดศูนย์กลางของภาพโดย คำนวณจากสมการที่ 5.1 และ 5.2

$$I_X = \frac{\sum_j \sum_i j f(i, j)}{\sum_i \sum_j f(i, j)} \quad (5.1)$$

$$I_Y = \frac{\sum_i \sum_j i f(i, j)}{\sum_i \sum_j f(i, j)} \quad (5.2)$$

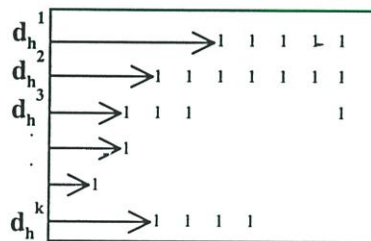
โดย  $I_X$  คือตำแหน่งของแถวที่เป็นตำแหน่งของจุดศูนย์กลาง และ  $I_Y$  คือตำแหน่งคอลัมน์ที่เป็นตำแหน่งของจุดศูนย์กลาง และ  $f(i, j)$  เป็นฟังก์ชันแสดงจุดภาพที่ตำแหน่ง  $(i, j)$  บนภาพ 2 ระดับ ดังสมการที่ 5.3





รูปที่ 5.2 การแบ่งพื้นที่โดยอาศัยเส้นแบ่งในแนวนอน

ในแต่ละพื้นที่จะคำนวณหาพื้นที่พื้นหลังของภาพ และผลต่างของระยะทางในแต่ละเส้นที่ลากในแนวนอนจากกรอบภาพถึงขอบภาพ ดังตัวอย่างในรูปที่ 5.3 สำหรับการคำนวณในบริเวณ  $S_h^1, S_h^2, S_h^3$  และ  $S_h^4$  ลักษณะเด่นทั้งสองนี้จะคำนวณได้จากระยะทางในแนวนอนจากกรอบภาพทางด้านซ้ายถึงขอบภาพด้านบนนอก และสำหรับการคำนวณในบริเวณ  $S_h^5, S_h^6, S_h^7$  และ  $S_h^8$  ลักษณะเด่นทั้งสองนี้จะคำนวณได้จากระยะทางในแนวนอนจากกรอบภาพทางด้านขวาถึงขอบภาพด้านบนนอก



รูปที่ 5.3 ระยะทางในแนวนอนจากกรอบภาพทางด้านซ้ายถึงขอบภาพด้านบนนอกของบริเวณ  $S_h^1$

เราสามารถคำนวณพื้นที่พื้นหลังของแต่ละบริเวณ ได้จากสมการที่ 5.4

$$PA_h^i = \left( \sum_{j=1}^k d_h^j \right) / (W_h^i \times H_h^i) \tag{5.4}$$

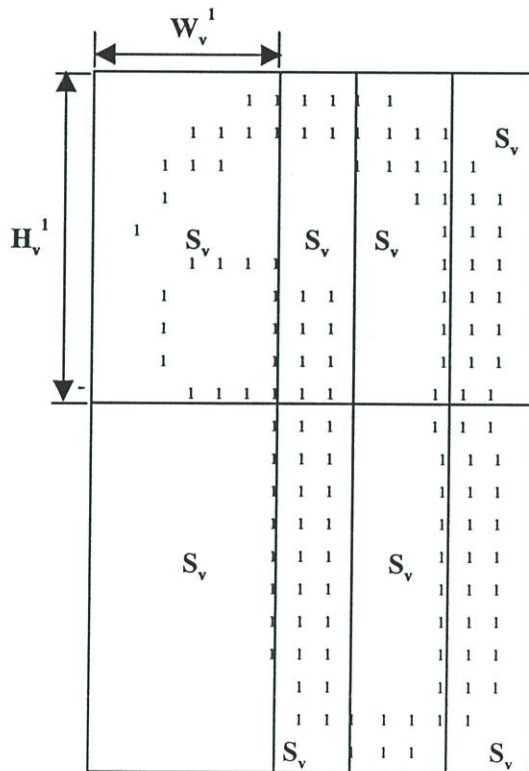
เมื่อ  $PA_h^i$  คือพื้นที่พื้นหลังภาพของบริเวณ  $S_h^i$  แต่ละบริเวณจะประกอบด้วยเป็นระยะทาง  $d_h^j$  เมื่อ  $j = 1, 2, \dots, k$  ซึ่งเป็นระยะทางในแนวนอนจากกรอบภาพจนถึงขอบภาพ โดย  $W_h^i$  เป็นความกว้างของบริเวณ  $S_h^i$  และ  $H_h^i$  เป็นความสูงของบริเวณ  $S_h^i$  และคำนวณผลต่างของแต่ละระยะทางจากกรอบภาพไปถึงขอบภาพของแต่ละบริเวณ สามารถคำนวณได้จากสมการที่ 5.5

$$PD_h^i = \left\{ \sum_{j=1}^k \left( |d_h^j| - |d_h^{j+1}| \right) \right\} / (W_h^i \times H_h^i) \tag{5.5}$$

เมื่อ  $PD_h^i$  คือผลต่างของแต่ละระยะทางที่อยู่ติดกัน โดยระยะทาง  $d_h^j$  กับระยะทาง  $d_h^{j+1}$  เป็นระยะทางในแนวนอนจากกรอบภาพไปถึงขอบภาพของบริเวณ  $S_h^i$

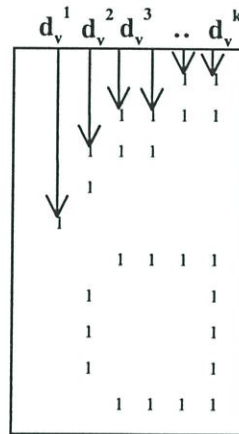
**5.1.2.2 การแบ่งพื้นที่โดยอาศัยเส้นแบ่งในแนวตั้ง**

การแบ่งพื้นที่โดยอาศัยเส้นแบ่งในแนวตั้งจากจะแบ่งบริเวณออกเป็น 8 พื้นที่ การแบ่งบริเวณจะแบ่งโดยใช้เส้นที่ลากผ่านจุดศูนย์กลางของภาพเป็นเส้นแบ่งแต่ละบริเวณ ซึ่งมีการแบ่งดังแสดงในรูปที่ 5.4 และกำหนดชื่อเรียกบริเวณต่างๆ เป็น  $S_v^1, S_v^2, S_v^3, S_v^4, S_v^5, S_v^6, S_v^7, S_v^8$



รูปที่ 5.4 การแบ่งพื้นที่โดยอาศัยเส้นแบ่งในแนวตั้งจาก

ในแต่ละพื้นที่จะคำนวณหาพื้นที่พื้นหลังของภาพ และผลต่างของระยะทางในแต่ละเส้นที่ลากในแนวตั้งฉากจากกรอบภาพถึงขอบภาพ ดังตัวอย่างในรูปที่ 5.5 สำหรับการคำนวณในบริเวณ  $S_v^1, S_v^2, S_v^3$  และ  $S_v^4$  ลักษณะเด่นทั้งสองนี้จะคำนวณได้จากระยะทางในแนวตั้งฉากจากกรอบภาพทางด้านบนถึงขอบภาพด้านบน และสำหรับการคำนวณในบริเวณ  $S_v^5, S_v^6, S_v^7$  และ  $S_v^8$  ลักษณะเด่นทั้งสองนี้จะคำนวณได้จากระยะทางในแนวตั้งฉากจากกรอบภาพทางด้านล่างถึงขอบภาพด้านล่าง



รูปที่ 5.5 ระยะทางในแนวตั้งฉากจากกรอบภาพทางด้านบนถึงขอบภาพด้านบนของบริเวณ  $S_v^1$

เราสามารถคำนวณพื้นที่พื้นหลังของแต่ละบริเวณ ได้จากสมการที่ 5.6

$$PA_v^i = \left( \sum_{j=1}^k d_v^j \right) / (W_v^i \times H_v^i) \quad (5.6)$$

เมื่อ  $PA_v^i$  คือพื้นที่พื้นหลังภาพของบริเวณ  $S_v^i$  แต่ละบริเวณจะประกอบด้วยเป็นระยะทาง  $d_v^j$  เมื่อ  $j = 1, 2, \dots, k$  ซึ่งเป็นระยะทางในแนวตั้งฉากจากกรอบภาพจนถึงขอบภาพ โดย  $W_v^i$  เป็นความกว้างของบริเวณ  $S_v^i$  และ  $H_v^i$  เป็นความสูงของบริเวณ  $S_v^i$  และคำนวณผลต่างของแต่ละระยะทางจากกรอบภาพไปถึงขอบภาพของแต่ละบริเวณ สามารถคำนวณได้จากสมการที่ 5.7

$$PD_v^i = \left\{ \sum_{j=1}^k \left( \|d_v^j\| - \|d_v^{j+1}\| \right) \right\} / (W_v^i \times H_v^i) \quad (5.7)$$

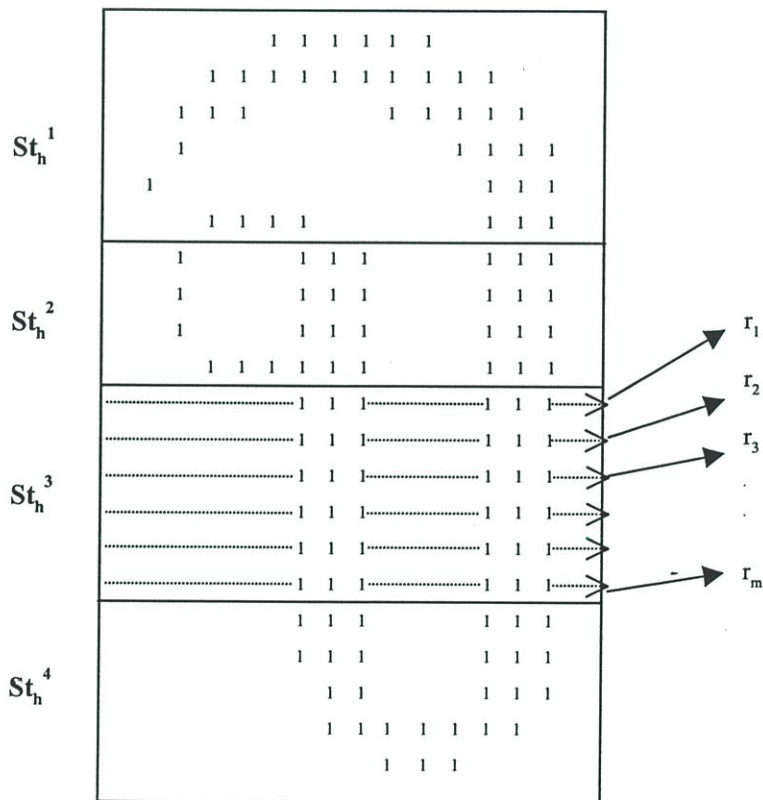
เมื่อ  $PD_v^i$  คือผลต่างของแต่ละระยะทางที่อยู่ติดกัน โดยระยะทาง  $d_v^j$  กับระยะทาง  $d_v^{j+1}$  เป็นระยะทางในแนวตั้งฉากจากกรอบภาพไปถึงขอบภาพของบริเวณ  $S_v^i$

5.1.3 การคำนวณค่าความหนาแน่นของจุดตัด

ค่าความหนาแน่นของการเปลี่ยนจากพื้นหลังภาพไปเป็นเนื้อภาพเมื่อลากเส้นผ่านแสดง ลักษณะโครงสร้างภายในของตัวอักษร เริ่มต้นหาได้จากการแบ่งพื้นที่เป็น 4 บริเวณ โดยอาศัยเส้นที่ ลากผ่านจุดศูนย์กลางของภาพ การแบ่งพื้นที่นี้จะแยกเป็นแบ่งพื้นที่ในแนวนอนและแนวตั้ง

5.1.3.1 การแบ่งพื้นที่เพื่อคำนวณค่าความหนาแน่นของจุดตัด โดยอาศัยเส้นแบ่งในแนวนอน

การแบ่งพื้นที่เพื่อคำนวณค่าความหนาแน่นของการเปลี่ยนจากพื้นหลังภาพไปเป็นเนื้อ ภาพโดยอาศัยเส้นแบ่งในแนวนอนโดยแบ่งบริเวณออกเป็น 4 พื้นที่ ซึ่งเส้นแบ่งแต่ละพื้นที่ เป็นเส้น ที่ลากผ่านจุดศูนย์กลางของภาพในแนวนอนเท่านั้น และกำหนดชื่อเรียกบริเวณต่างๆ เป็น  $St_h^1, St_h^2, St_h^3, St_h^4$  ดังแสดงในรูปที่ 5.6



รูปที่ 5.6 ความหนาแน่นของการเปลี่ยนจากพื้นหลังภาพไปเป็นเนื้อภาพ โดยการแบ่งในแนวนอน

ค่าความหนาแน่นของการเปลี่ยนจากพื้นหลังภาพไปเป็นเนื้อภาพ เมื่อลากเส้นผ่านในแนวนอน คำนวณได้จากการนับจำนวนของการเปลี่ยนจากพื้นหลังภาพไปเป็นเนื้อภาพของแต่ละ



ในการทำงานเกี่ยวกับการหาค่าความหนาแน่นของการเปลี่ยนจากพื้นหลังภาพไปเป็นเนื้อภาพโดยอาศัยเส้นแบ่งในแนวนอน เราสามารถคำนวณค่าความหนาแน่นของการเปลี่ยนจากพื้นหลังภาพไปเป็นเนื้อภาพ เมื่อลากเส้นผ่านภาพโดยอาศัยเส้นแบ่งในแนวตั้งฉาก ในแต่ละบริเวณได้จากสมการที่ 5.9

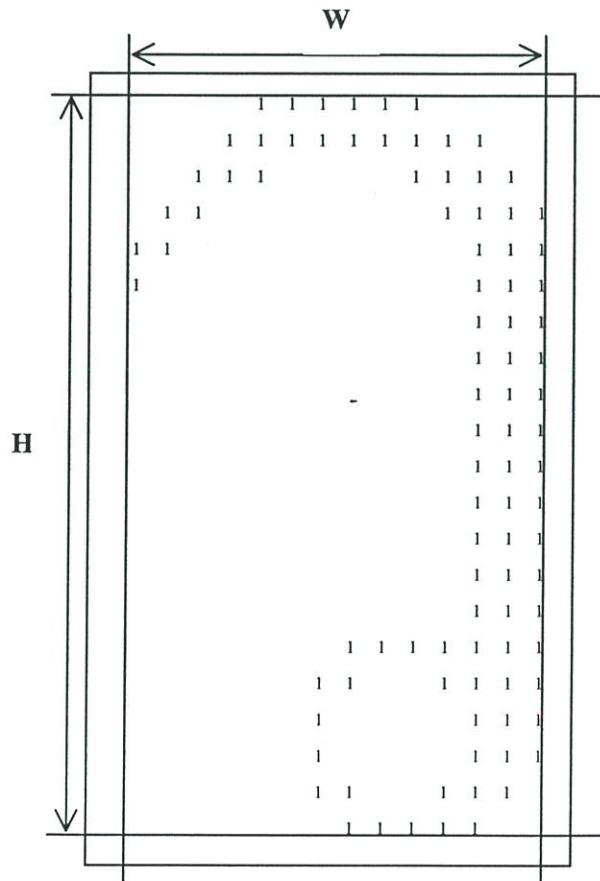
$$SD_v^j = \sum_{i=1}^m \sum_x \overline{s(x, y)} \cdot s(x+1, y) / m \quad (5.9)$$

โดย  $SD_v^j$  คือค่าความหนาแน่นของการเปลี่ยนจากพื้นหลังภาพไปเป็นเนื้อภาพของบริเวณ  $St_v^j$  ซึ่ง  $s(x, y)$  คือค่าของจุดภาพที่ตำแหน่ง  $(x, y)$  โดยที่  $m$  เท่ากับจำนวนเส้นที่ลากผ่านในบริเวณที่วัดค่า

#### 5.1.4 การคำนวณสัดส่วนความกว้างต่อความสูง

เราสามารถหาอัตราส่วนระหว่างความกว้างต่อความสูงของภาพได้จากสมการที่ 5.10 และแสดงได้ดังรูปที่ 5.8

$$\text{อัตราส่วนระหว่างความกว้างต่อความสูง} = W/H \quad (5.10)$$



รูปที่ 5.8 ความกว้างและความสูงของตัวอักษร



		1 1 1 1 1 1	1	
	1	1 1 1 1 1 1	1 1 1	
	1 1		1 1 1 1	
1 1				1 1 1
1 1 1 1	1 1 1			1 1 1
1 1 1 1	1 1 1			1 1 1
	1 1	1		1 1 1
	1 1	1		1 1 1
	1 1	1		1 1 1
	1 1	1		1 1 1
	1 1	1		1 1 1
	1 1	1		1 1 1
	1 1	1		1 1 1
	1 1 1 1	1		1 1 1
1 1 1 1 1 1	1			1 1 1
1 1	1 1 1	1		1 1 1
1		1 1		1 1 1
1 1	1 1	1		1 1
1 1 1 1				1 1

V2 = 2      V3 = 1

H2 = 2  
H3 = 2

รูปที่ 5.10 การแบ่งพื้นที่ของภาพเป็น 9 บริเวณเท่ากัน

### 5.1.6 การหาลักษณะเด่นด้วยการเข้ารหัสลูกโซ่ (Chain code)

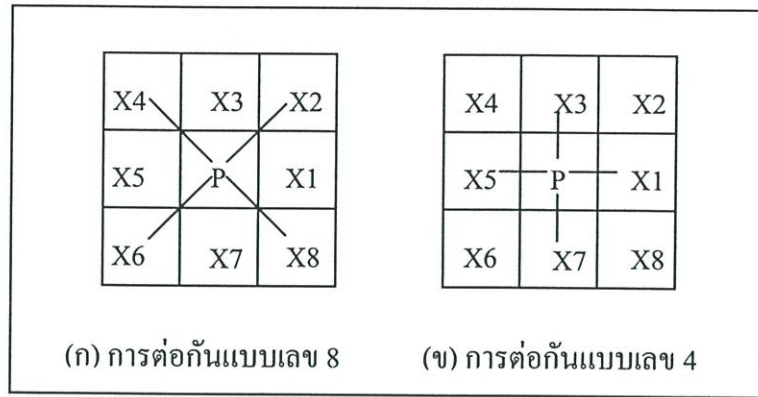
การหาลักษณะเด่นด้วยการเข้ารหัสลูกโซ่ สามารถหาได้โดยแบ่งเป็น 2 ขั้นตอน คือ ขั้นตอนของการหาขอบภาพโดยการคำนวณค่าตัวเลขต่อเนื่อง (Connectivity number) ส่วนขั้นตอนที่ 2 เป็นการเข้ารหัสตัวอักษรด้วยรหัสลูกโซ่

#### 5.1.6.1 การคำนวณค่าตัวเลขต่อเนื่อง (Connectivity number)

ค่าตัวเลขต่อเนื่องเป็นคุณสมบัติทาง โทโพโลยี (Topological property) ของจุดภาพที่อยู่ติดกันในรูปแบบสองระดับ ค่าตัวเลขต่อเนื่องนี้จะแสดงความสัมพันธ์ระหว่างจุดภาพข้างเคียง (Neighborhood pixel) จุดภาพข้างเคียงบนรูปแบบสองระดับใดๆ จะแบ่งได้เป็น 2 แบบคือ

1. จุดภาพข้างเคียงแบบเลข 4 จุดภาพ (4-Connectivity)
2. จุดภาพข้างเคียงแบบเลข 8 จุดภาพ (8-Connectivity)

กำหนดให้ P เป็นจุดภาพใด ๆ บนรูปแบบสองระดับ จุดภาพข้างเคียงของจุด P ทั้ง 2 แบบแสดงไว้ในรูปที่ 5.11(ก) และ (ข) ตามลำดับ



รูปที่ 5.11 ลักษณะของจุดภาพข้างเคียงแบบ (ก) 8-Connectivity (ข) 4-Connectivity

ค่าตัวเลขต่อเนื่องของจุดภาพข้างเคียงทั้ง 2 แบบสามารถหาได้จากสมการ 5.11 และสมการ 5.12

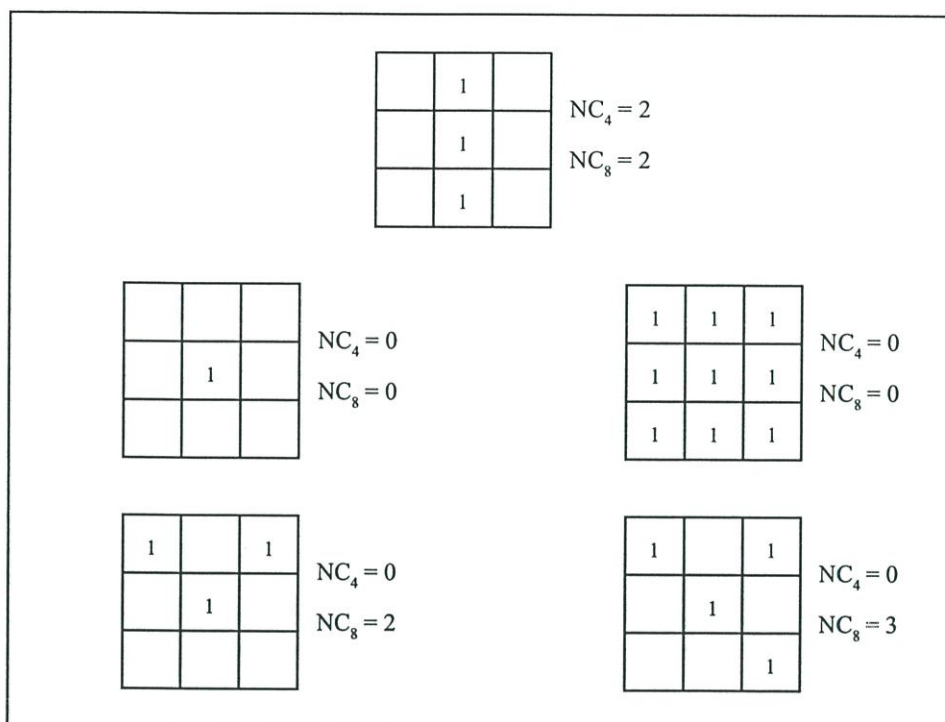
$$NC_4 = \sum (X_k - X_k X_{k+1} X_{k+2}) \quad (5.11)$$

$$NC_8 = \sum (\bar{X}_k - \bar{X}_k \bar{X}_{k+1} \bar{X}_{k+2}) \quad (5.12)$$

$$S_i = \{1, 3, 5, 7\}, k \geq 9 \rightarrow k = k - 8$$

$$\bar{X} = (1 - X)$$

ผลที่ได้จากการคำนวณ โดยสมการ 5.11 และ 5.12 จะได้เป็นเลขจำนวนเต็มบวกที่มีค่าอยู่ระหว่าง 0 ถึง 4 ดังแสดงในรูป 5.12



รูปที่ 5.12 ค่า  $NC_4$  และ  $NC_8$  ของจุดภาพ

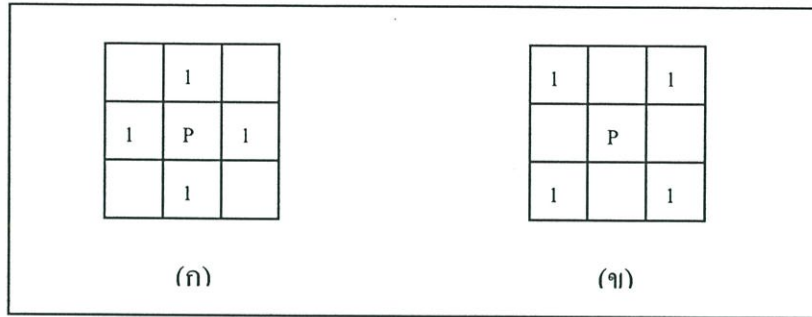
ค่าตัวเลขต่อเนื้องที่คำนวณได้ทั้ง 2 แบบ จะบ่งบอกถึงความสัมพันธ์ระหว่างจุดภาพ ออกมาดังที่แสดงในตารางที่ 5.1

ตารางที่ 5.1 ตารางแสดงความสัมพันธ์ระหว่างจุดภาพ

ค่าตัวเลขต่อเนื้อง	จุดอิสระหรือจุดภายใน	จุดปลาย		จุดต่อเนื้อง					จุดแยก			จุดตัด		
				1	2	2	2	0	3	0	1	2	4	0
$NC_4$	0	1	0	1	2	2	2	0	3	0	1	2	4	0
$NC_8$	0	1	1	2	2	2	1	2	1	3	3	2	0	4

ค่าตัวเลขต่อเนื้องของจุดอิสระหรือจุดภายในคือ 0, จุดปลายคือ 1, จุดต่อเนื้องคือ 2, จุดแยกคือ 3, จุดตัดคือ 4

ตัวอย่างการคำนวณ



รูปที่ 5.13 ตัวอย่างการคำนวณค่า  $NC_4$  และ  $NC_8$

จากรูปที่ 5.13 (ก) จะคำนวณค่า  $NC_4$  และ  $NC_8$  ได้ดังนี้

$$\begin{aligned}
 NC_4 &= X_1 + X_3 + X_5 + X_7 - (X_1 * X_2 * X_3) - (X_3 * X_4 * X_5) - (X_5 * X_6 * X_7) - (X_7 * X_8 * X_1) \\
 &= 1 + 1 + 1 + 1 - (1*0*1) - (1*0*1) - (1*0*1) - (1*0*1) \\
 &= 4
 \end{aligned}$$

$$\begin{aligned}
 NC_8 &= \overline{X_1} + \overline{X_3} + \overline{X_5} + \overline{X_7} - (\overline{X_1} * \overline{X_2} * \overline{X_3}) - (\overline{X_3} * \overline{X_4} * \overline{X_5}) - (\overline{X_5} * \overline{X_6} * \overline{X_7}) - (\overline{X_7} * \overline{X_8} * \overline{X_1}) \\
 &= 0 + 0 + 0 + 0 - (0*1*0) - (0*1*0) - (0*1*0) - (0*1*0) \\
 &= 0
 \end{aligned}$$

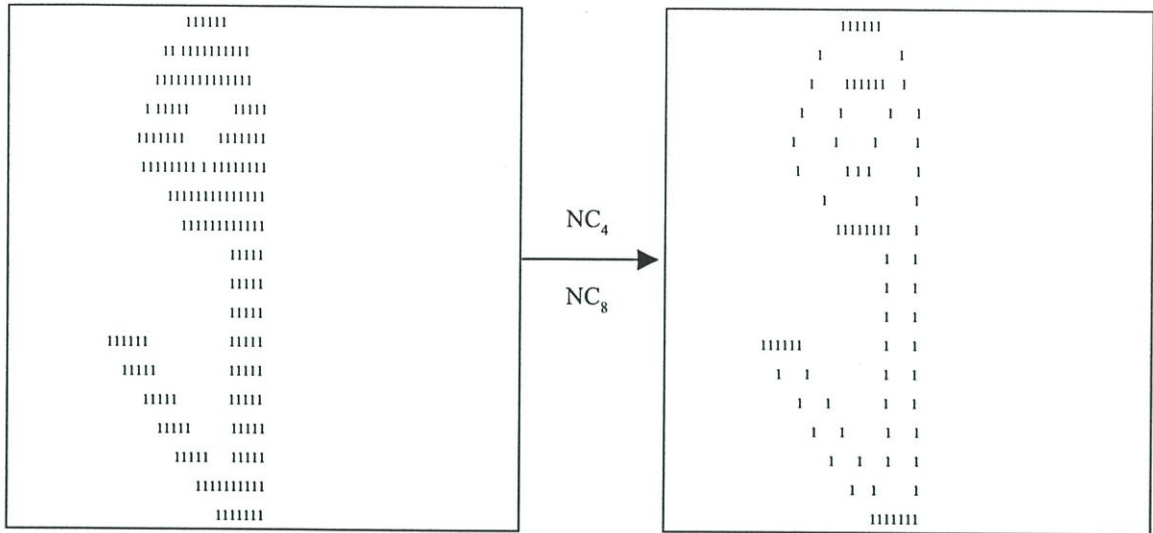
จากรูปที่ 5.13 (ข) จะคำนวณค่า  $NC_4$  และ  $NC_8$  ได้ดังนี้

$$\begin{aligned}
 NC_4 &= X_1 + X_3 + X_5 + X_7 - (X_1 * X_2 * X_3) - (X_3 * X_4 * X_5) - (X_5 * X_6 * X_7) - (X_7 * X_8 * X_1) \\
 &= 0 + 0 + 0 + 0 - (0*1*0) - (0*1*0) - (0*1*0) - (0*1*0) \\
 &= 0
 \end{aligned}$$

$$\begin{aligned}
 NC_8 &= \overline{X_1} + \overline{X_3} + \overline{X_5} + \overline{X_7} - (\overline{X_1} * \overline{X_2} * \overline{X_3}) - (\overline{X_3} * \overline{X_4} * \overline{X_5}) - (\overline{X_5} * \overline{X_6} * \overline{X_7}) - (\overline{X_7} * \overline{X_8} * \overline{X_1}) \\
 &= 1 + 1 + 1 + 1 - (1*0*1) - (1*0*1) - (1*0*1) - (1*0*1) \\
 &= 4
 \end{aligned}$$

#### 5.1.6.2 การหาขอบภาพ (Edge detection)

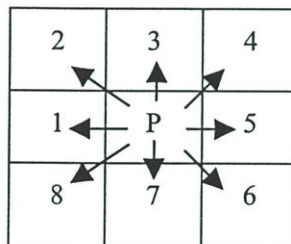
ขอบภาพของภาพในรูปแบบกับฉากจะแบ่งแยกจากกันโดยขอบภาพ (Edge) การหาขอบเขตของภาพในรูปแบบสองระดับใดๆ สามารถใช้ค่าตัวเลขต่อเนื่อง  $NC_4$  และ  $NC_8$  ที่คำนวณได้จากสมการที่ 5.11 และ 5.12 จุดภาพใดๆ ที่มีค่า  $NC_4$  และ  $NC_8$  เป็นไปตามลักษณะของจุดปลาย (End point) นั่นก็หมายความว่าจุดภาพ ณ ตำแหน่งนั้นเป็นขอบภาพ ดังแสดงในรูปที่ 5.14



รูปที่ 5.14 ค่าตัวเลขต่อเนื่องที่เป็นขอบภาพตัวอักษร “ง”

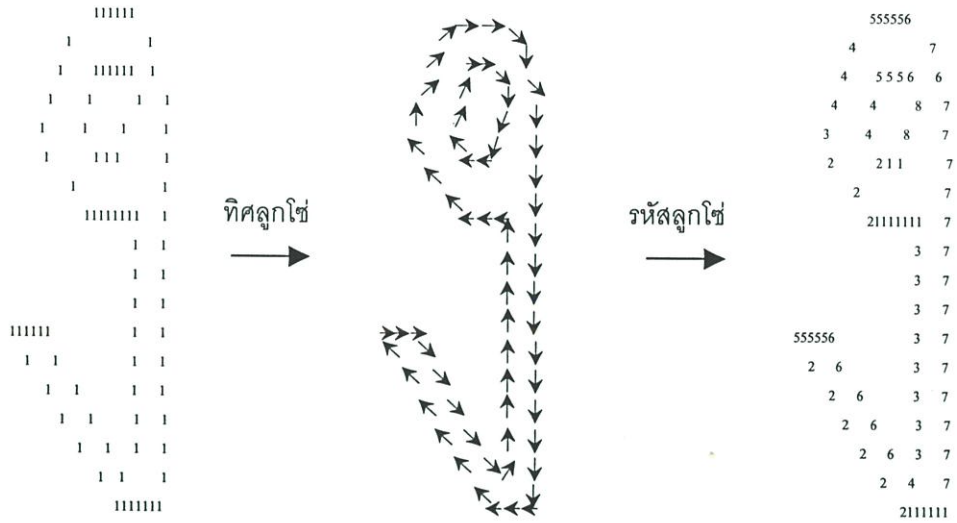
### 5.1.6.3 การหาลักษณะเด่นของตัวอักษรด้วยรหัสลูกโซ่ (Chain Code)

รหัสลูกโซ่ (Chain code) มีขนาด 1 พิกเซล และมีทิศทาง 8 ทิศทาง คือใช้รหัสลูกโซ่ ในการเคลื่อนที่ระหว่างพิกเซลซึ่งบอกทางดังรูปที่ 5.15



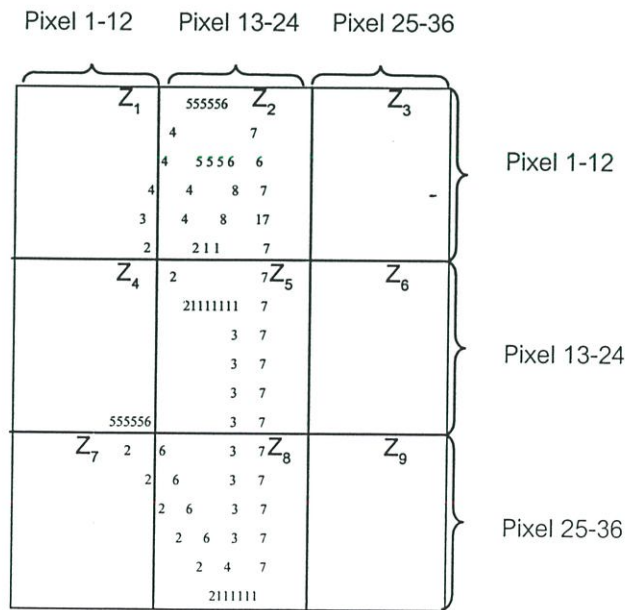
รูปที่ 5.15 ทิศทางของรหัสลูกโซ่

เพื่อเป็นการลดข้อมูลภาพตัวอักษรพิมพ์ภาษาไทยให้น้อยลง เรากำหนดลักษณะเด่นของตัวอักษรโดยเรารหัสลูกโซ่มาทำการประมวลผล ซึ่งสามารถแบ่งออกเป็นส่วน ๆ ดังรูปที่ 5.16



รูปที่ 5.16 แสดงทิศรหัสลูกโซ่และการเข้ารหัสลูกโซ่ (Chain code) ตัวอักษร “ง”

เมื่อได้รหัสลูกโซ่แล้วทำการนับรหัสลูกโซ่ในพื้นที่เฉพาะ กำหนดชื่อเรียกบริเวณต่างๆ เป็น  $Z_1, Z_2, Z_3, Z_4, Z_5, Z_6, Z_7, Z_8, Z_9$  การแบ่งพื้นที่เฉพาะเหล่านี้ลดจำนวนข้อมูลให้น้อยลงและยังสามารถระบุลักษณะของตัวอักษรได้ บริเวณที่กำหนดจะมีสัดส่วนดังแสดงในรูป 5.17 ในการกำหนดรหัสลูกโซ่เรานับรหัสลูกโซ่ในทิศตรงข้ามกันเป็นรหัสเดียวกัน ดังนั้นกำหนดให้รหัส 1 กับ 5 แทนด้วยรหัส 1, รหัส 2 กับ 6 แทนด้วยรหัส 2, รหัส 3 กับ 7 แทนด้วยรหัส 3, รหัส 4 กับ 8 แทนด้วยรหัส 4 เพื่อป้องกันรหัสผิดพลาดจากการเข้ารหัส



รูปที่ 5.17 แสดงการแบ่งพื้นที่รหัสลูกโซ่ของตัวอักษร “ง”

จากรหัสลูกโซ่ในแต่ละพื้นที่ที่ได้ มานับจำนวนแต่ละรหัส และแทนลักษณะเด่นของแต่ละพื้นที่ด้วยรหัสลูกโซ่ที่มีจำนวนมากที่สุด และพื้นที่ที่ไม่มีรหัสลูกโซ่ถูกแทนด้วยรหัส 0

## 5.2 การประยุกต์พีชชีเซต และราฟเซตร่วมกัน

ลักษณะเด่นที่คำนวณได้สำหรับงานวิจัยนี้ ประกอบด้วยลักษณะเด่นที่ไม่มีความต่อเนื่องของข้อมูล ได้แก่ จำนวนครั้งของการเปลี่ยนพื้นหลังไปเป็นเนื้อหาตามเส้นในแนวนอน และแนวตั้งฉาก, รหัสลูกโซ่ที่มีจำนวนมากที่สุดในแต่ละพื้นที่ ดังรูปที่ 5.17 ส่วนลักษณะเด่นที่มีความต่อเนื่องของข้อมูล ได้แก่ การหาพื้นหลังของภาพ, ค่าความหนาแน่นของการเปลี่ยนจากพื้นหลังภาพไปเป็นเนื้อหา โดยตัวอย่างข้อมูลที่ได้จากการหาลักษณะเด่นที่คำนวณได้ แสดงได้ดังตารางที่ 5.2 โดยประกอบไปด้วยแอตทริบิวต์ลักษณะเด่นที่มีความต่อเนื่อง และแอตทริบิวต์ลักษณะเด่นที่ไม่มีความต่อเนื่อง รวมทั้งหมด 56 แอตทริบิวต์

ตารางที่ 5.2 แสดงตัวอย่างข้อมูลแอตทริบิวต์แอตทริบิวต์

Att1	Att2	Att3	Att4	Att5	Att6	...	Att55	Att56	Char
0.3571	0.16	0.2143	0.2333	0.4821	0.125	...	0	3	ก
0.4286	0.2424	0.2727	0.2727	0.3571	0.125	...	0	0	ก
0.4	0.1667	0.2	0.2333	0.3214	0.125	...	0	0	ก
...	...	...	...	...	...	...	...	...	...
0.2273	0.1455	0.7403	0.7121	0.119	0.1429	...	0	0	ข
0.1688	0.1455	0.7273	0.697	0.1429	0.1429	...	0	0	ข
0.197	0.1091	0.6623	0.7143	0.1667	0.1429	...	0	3	ข
...	...	...	...	...	...	...	...	...	...
0.2821	0.3932	0.094	0.2615	0.2479	0.3162	...	0	0	ช
0.2735	0.359	0.0846	0.3077	0.2037	0.3333	...	1	3	ช
0.2821	0.3761	0.0846	0.3154	0.1944	0.3333	...	1	0	ช

ซึ่งจะนำการแบ่งกลุ่มโดยพีชชีเซตมาแบ่งกลุ่มลักษณะเด่นที่มีความต่อเนื่อง การแบ่งกลุ่มข้อมูลสำหรับแต่ละแอตทริบิวต์ แสดงได้ดังตารางที่ 5.3 และตัวอย่างผลที่ได้จากการแบ่งกลุ่มโดยพีชชีเซต แสดงได้ดังตารางที่ 5.4

ตารางที่ 5.3 แสดงการแบ่งกลุ่มสำหรับแต่ละแอตทริบิว

แอตทริบิว	ตัวแปร	คำอธิบาย
$PA_h^1$	{Low, Medium, High}	การหาพื้นหลังของภาพ โดยการแบ่งพื้นที่ของภาพ ดังรูปที่ 5.2
$PA_h^2$	{Low, High}	
$PA_h^3$	{Low, High }	
$PA_h^4$	{Low, High }	
$PA_h^5$	{Low, High }	
$PA_h^6$	{Low, High }	
$PA_h^7$	{Low, High }	
$PA_h^8$	{Low, High }	
$PD_h^1$	{Low, High }	ค่าลักษณะของภาพรอบนอก โดยคำนวณผลต่างของ แต่ละระยะทางจากกรอบภาพไปถึงขอบภาพ โดย การแบ่งพื้นที่ของภาพดังรูปที่ 5.2
$PD_h^2$	{Low, High }	
$PD_h^3$	{Low, High }	
$PD_h^4$	{Low, Medium, High}	
$PD_h^5$	{Low, High }	
$PD_h^6$	{Low, High }	
$PD_h^7$	{Low, Medium, High}	
$PD_h^8$	{Low, High }	
$PA_v^1$	{Low, High }	การหาพื้นหลังของภาพ โดยการแบ่งพื้นที่ของภาพ ดังรูปที่ 5.4
$PA_v^2$	{Low, High }	
$PA_v^3$	{Low, High }	
$PA_v^4$	{Low, Medium, High} -	
$PA_v^5$	{Low, High }	
$PA_v^6$	{Low, Medium, High}	
$PA_v^7$	{Low, High }	
$PA_v^8$	{Low, High }	

ตารางที่ 5.3 (ต่อ)

แอตทริบิว	ตัวแปร	คำอธิบาย
$PD_v^1$	{Low, High }	ค่าลักษณะของภาพรอบนอก โดยการผลต่างของแต่ละระยะทางจากกรอบภาพไปถึงขอบภาพ โดยการแบ่งพื้นที่ของภาพดังรูปที่ 5.4
$PD_v^2$	{Low, Medium, High}	
$PD_v^3$	{Low, High }	
$PD_v^4$	{Low, High }	
$PD_v^5$	{Low, High }	
$PD_v^6$	{Low, Medium, High}	
$PD_v^7$	{Low, High }	
$PD_v^8$	{Low, High }	
$SD_h^1$	{Low, High }	ค่าความหนาแน่นของการเปลี่ยนจากพื้นหลังภาพไปเป็นเนื้อหาภาพ โดยการแบ่งแถบพื้นที่ในแนวนอน ดังรูปที่ 5.6
$SD_h^2$	{Low, High }	
$SD_h^3$	{Low, High }	
$SD_h^4$	{Low1,Low2,High1,High2}	
$SD_v^1$	{Low, High }	ค่าความหนาแน่นของการเปลี่ยนจากพื้นหลังภาพไปเป็นเนื้อหาภาพ โดยการแบ่งแถบพื้นที่ในแนวตั้ง ดังรูปที่ 5.7
$SD_v^2$	{Low, High }	
$SD_v^3$	{Low, High }	
$SD_v^4$	{Low, High }	
W/H	{Low, Medium, High}	สัดส่วนความกว้างต่อความสูง ดังสมการที่ 5.9 และรูปที่ 5.8
H1	-	จำนวนของการเปลี่ยนจากพื้นหลังไปเป็นเนื้อหาภาพตามแนวนอน ดังรูปที่ 5.9
V1	-	จำนวนของการเปลี่ยนจากพื้นหลังไปเป็นเนื้อหาภาพตามแนวตั้ง ดังรูปที่ 5.9
H2	-	จำนวนของการเปลี่ยนจากพื้นหลังไปเป็นเนื้อหาภาพตามแนวนอนของเส้นด้านบน ดังรูปที่ 5.10
H3	-	จำนวนของการเปลี่ยนจากพื้นหลังไปเป็นเนื้อหาภาพตามแนวนอนของเส้นด้านล่าง ดังรูปที่ 5.10
V2	-	จำนวนของการเปลี่ยนจากพื้นหลังไปเป็นเนื้อหาภาพตามแนวตั้งฉากของเส้นด้านซ้าย ดังรูปที่ 5.10

ตารางที่ 5.3 (ต่อ)

แอตทริบิวต์	ตัวแปร	คำอธิบาย
V3	-	จำนวนของการเปลี่ยนจากพื้นหลังไปเป็นเนื้อภาพตามแนวตั้งฉากของเส้นด้านขวา ดังรูปที่ 5.10
Z <sub>1</sub>	{0, 1, 2, 3, 4}	จำนวนรหัสโซ่ที่มีมากที่สุดของแต่ละบริเวณ ดังรูปที่ 5.17
Z <sub>2</sub>	{0, 1, 2, 3, 4}	
Z <sub>3</sub>	{0, 1, 2, 3, 4}	
Z <sub>4</sub>	{0, 1, 2, 3, 4}	
Z <sub>5</sub>	{0, 1, 2, 3, 4}	
Z <sub>6</sub>	{0, 1, 2, 3, 4}	
Z <sub>7</sub>	{0, 1, 2, 3, 4}	
Z <sub>8</sub>	{0, 1, 2, 3, 4}	
Z <sub>9</sub>	{0, 1, 2, 3, 4}	

ตารางที่ 5.4 แสดงตัวอย่างผลที่ได้จากการแบ่งกลุ่มโดยฟัซซี่เซต

Att1	Att2	Att3	Att4	Att5	Att6	...	Att55	Att56	Char
Medium	Low	Low	Low	High	Low	...	0	3	ก
Medium	Low	Low	Low	High	Low	...	0	0	ก
Medium	Low	Low	Low	High	Low	...	0	0	ก
...	...	...	...	...	...	...	...	...	...
Medium	Low	High	High	Low	Low	...	0	0	ข
Low	Low	High	High	Low	Low	...	0	0	ข
Low	Low	High	High	Low	Low	...	0	3	ข
...	...	...	...	...	...	...	...	...	...
Medium	High	Low	Low	Low	High	...	0	0	ช
Medium	High	Low	Low	Low	High	...	1	3	ช
Medium	High	Low	Low	Low	High	...	1	0	ช

จากตารางที่ 5.3 ตัวอักษรหนึ่งตัวสามารถนำมาหาลักษณะเด่นได้ 56 อย่าง ลักษณะเด่นที่มีความต่อเนื่องของข้อมูล จะนำมาแบ่งกลุ่มด้วยอัลกอริทึม Fuzzy c-Means และประเมินจำนวนกลุ่มที่เหมาะสมที่สุด

จากนั้นคำนวณหา core และ reduct โดยอาศัยกราฟเซต เพื่อเลือกส่วนของข้อมูลลักษณะเด่นที่จำเป็นต่อการระบุตัวอักษรแต่ละตัว จากข้อมูลในตารางที่ 5.3 เราสามารถนำข้อมูลเหล่านี้ไปสร้างตา

วางข้อมูลสำหรับการเรียนรู้โดยกราฟเซตได้ ดังตัวอย่างในตารางที่ 5.4 ซึ่งแอตทริบิวเงื่อนไขคือ ลักษณะเด่นจำนวน 56 อย่าง โดยที่แบ่งกลุ่มลักษณะเด่นที่มีความต่อเนื่องแล้ว ส่วนแอตทริบิวสำหรับการตัดสินใจจะระบุตามชนิดของตัวอักษร ซึ่งสามารถแสดง core และ reduct ได้ดังตารางที่ 5.5 และ 5.6 ตามลำดับ

ตารางที่ 5.5 แสดงแอตทริบิว core

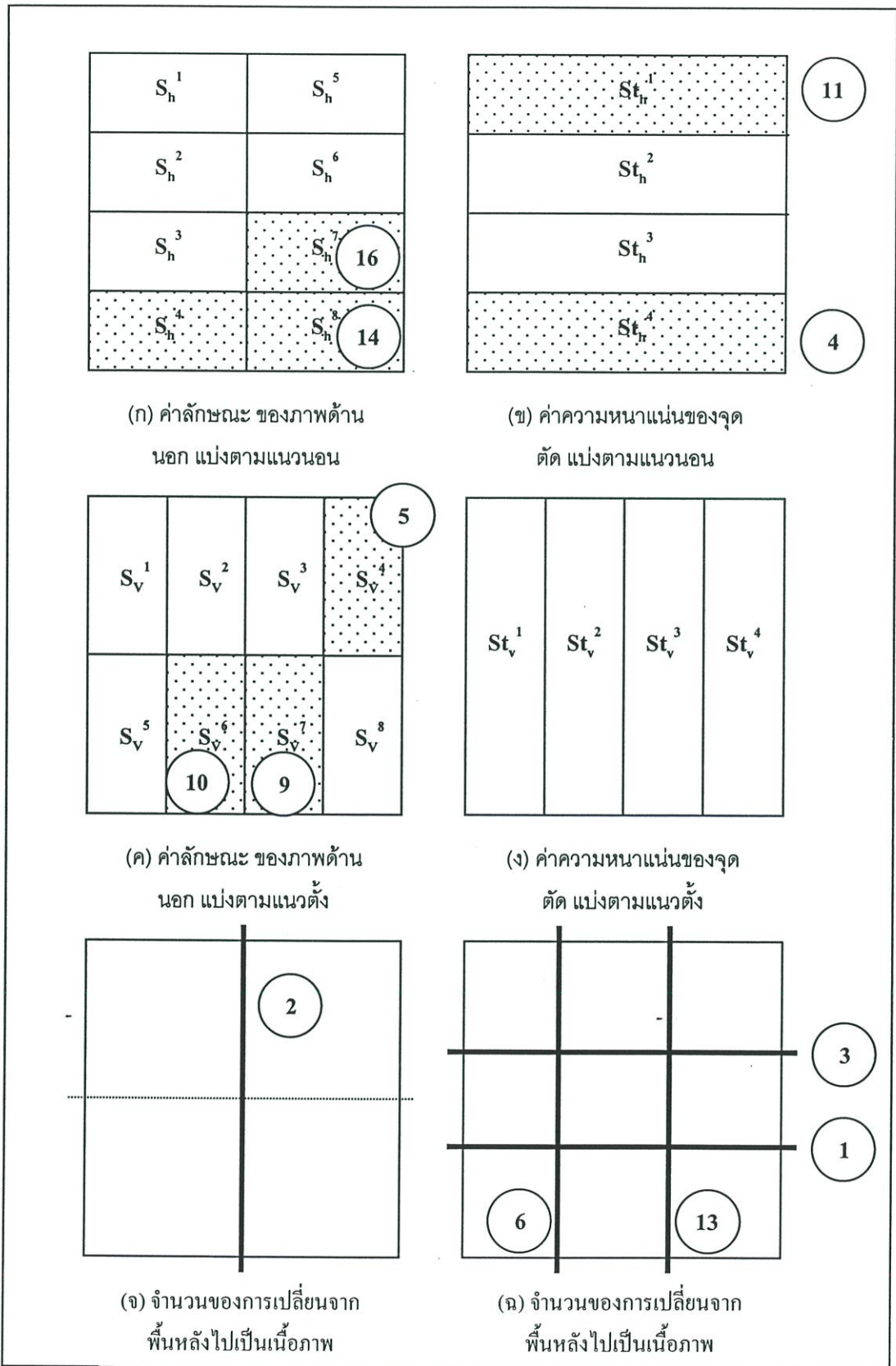
แอตทริบิว	ตัวแปร	คำอธิบาย
$PD_h^4$	{Low, Medium, High}	ค่าลักษณะของภาพรอบนอก โดยการผลต่างของแต่ละระยะทางจากกรอบภาพไปถึงขอบภาพของพื้นที่ $S_h^4$ ของภาพดังรูปที่ 5.2
$Z_4$	{0, 1, 2, 3, 4}	รหัสสีที่มีจำนวนมากที่สุดในบริเวณที่ 4 ดังรูปที่ 5.17

ตารางที่ 5.6 แสดงแอตทริบิว reduct

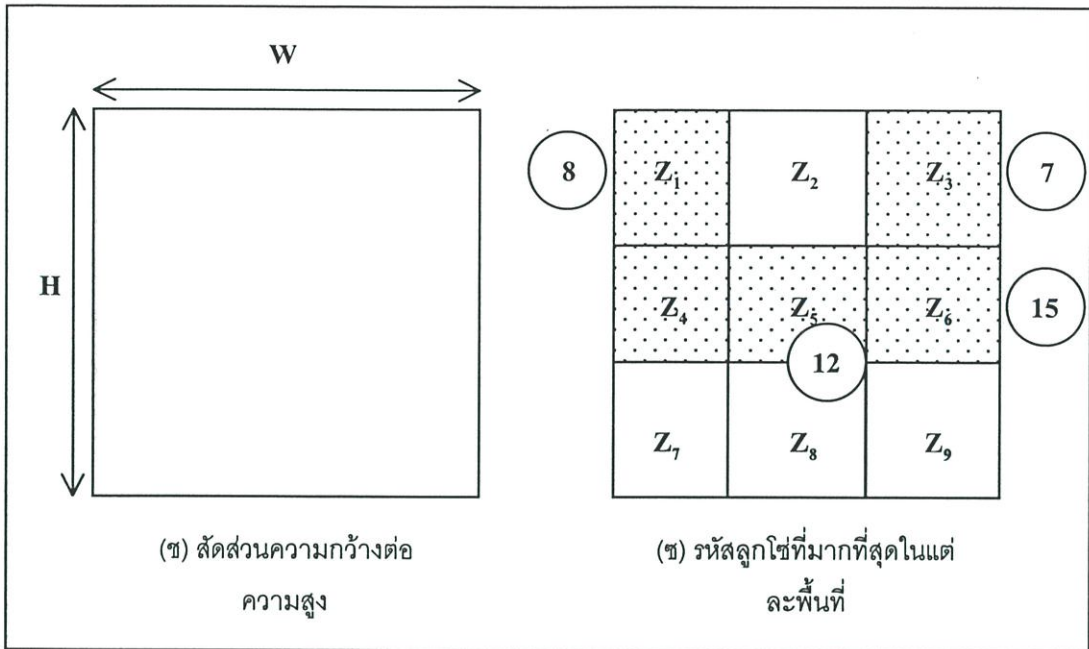
แอตทริบิว	ตัวแปร	คำอธิบาย
$PD_h^4$	{Low, Medium, High}	ค่าลักษณะของภาพรอบนอก โดยการผลต่างของแต่ละระยะทางจากกรอบภาพไปถึงขอบภาพของพื้นที่ $S_h^4$ ของภาพดังรูปที่ 5.2
$PD_h^7$	{Low, Medium, High}	ค่าลักษณะของภาพรอบนอก โดยการผลต่างของแต่ละระยะทางจากกรอบภาพไปถึงขอบภาพของพื้นที่ $S_h^7$ ของภาพดังรูปที่ 5.2
$PD_h^8$	{Low, High }	ค่าลักษณะของภาพรอบนอก โดยการผลต่างของแต่ละระยะทางจากกรอบภาพไปถึงขอบภาพของพื้นที่ $S_h^8$ ของภาพดังรูปที่ 5.2
$SD_h^1$	{Low, High }	ค่าความหนาแน่นของการเปลี่ยนจากพื้นหลังภาพไปเป็นเนื้อภาพ โดยการแบ่งแถบ $St_h^1$ ดังรูปที่ 5.6
$SD_h^4$	{Low1,Low2,High1,High2}	ค่าความหนาแน่นของการเปลี่ยนจากพื้นหลังภาพไปเป็นเนื้อภาพ โดยการแบ่งแถบ $St_h^4$ ดังรูปที่ 5.6
$PA_v^4$	{Low, Medium, High}	ค่าลักษณะของภาพรอบนอก โดยคำนวณค่าพื้นหลังพื้นภาพของพื้นที่ $S_v^4$ ดังรูปที่ 5.4

ตารางที่ 5.6 (ต่อ)

แอตทริบิว	ตัวแปร	คำอธิบาย
$PD_v^6$	{Low, Medium, High}	ค่าลักษณะของภาพรอบนอก โดยการผลต่างของแต่ละระยะทางจากกรอบภาพไปถึงขอบภาพของพื้นที่ $S_v^6$ ของภาพดังรูปที่ 5.4
$PD_v^7$	{Low, High }	ค่าลักษณะของภาพรอบนอก โดยการผลต่างของแต่ละระยะทางจากกรอบภาพไปถึงขอบภาพของพื้นที่ $S_v^7$ ของภาพดังรูปที่ 5.4
V1	-	จำนวนของการเปลี่ยนจากพื้นหลังไปเป็นเนื้อภาพตามแนวตั้ง ดังรูปที่ 5.9
H2	-	จำนวนของการเปลี่ยนจากพื้นหลังไปเป็นเนื้อภาพตามแนวนอนของเส้นด้านบน ดังรูปที่ 5.10
H3	-	จำนวนของการเปลี่ยนจากพื้นหลังไปเป็นเนื้อภาพตามแนวนอนของเส้นด้านล่าง ดังรูปที่ 5.10
V2	-	จำนวนของการเปลี่ยนจากพื้นหลังไปเป็นเนื้อภาพตามแนวตั้งฉากของเส้นด้านซ้าย ดังรูปที่ 5.10
V3	-	จำนวนของการเปลี่ยนจากพื้นหลังไปเป็นเนื้อภาพตามแนวตั้งฉากของเส้นด้านขวา ดังรูปที่ 5.10
$Z_1$	{0, 1, 2, 3, 4}	จำนวนรหัสโซ่ที่มีมากที่สุดในบริเวณที่ 1 ดังรูปที่ 5.17
$Z_3$	{0, 1, 2, 3, 4}	มีจำนวนรหัสโซ่ที่มีมากที่สุดในบริเวณที่ 3 ดังรูปที่ 5.17
$Z_4$	{0, 1, 2, 3, 4}	จำนวนรหัสโซ่ที่มีมากที่สุดในบริเวณที่ 4 ดังรูปที่ 5.17
$Z_5$	{0, 1, 2, 3, 4}	จำนวนรหัสโซ่ที่มีมากที่สุดในบริเวณที่ 5 ดังรูปที่ 5.17
$Z_6$	{0, 1, 2, 3, 4}	จำนวนรหัสโซ่ที่มีมากที่สุดในบริเวณที่ 6 ดังรูปที่ 5.17



รูปที่ 5.18 การจำลองกรอบภาพตัวอักษร และการแบ่งพื้นที่สำหรับหาลักษณะเด่น



### รูปที่ 5.18 (ต่อ)

รูปที่ 5.18 แสดงการจำลองกรอบภาพตัวอักษร และการแบ่งพื้นที่สำหรับหาลักษณะเด่นแต่ละแบบ และพื้นที่ที่เราแสดงส่วนที่เป็น reduct และลำดับแอดทริบิวต์ที่เพิ่มขึ้นซึ่งเป็นผลของการหา reduct โดยอาศัยอัลกอริทึมการหา reduct ดังได้อธิบายในหัวข้อ 3.5

ส่วนที่เรเงาและมีหมายเลขกำกับ ในรูปที่ 5.18 เป็นพื้นที่ของแอดทริบิวต์ที่เป็น reduct โดยหมายเลขที่กำกับ แสดงลำดับของการเพิ่มแอดทริบิวต์ซึ่งเป็นไปตามอัลกอริทึมการหาค่า reduct ส่วนที่เรเงา แต่ไม่มีหมายเลขกำกับ เป็นแอดทริบิวต์ที่เป็น core จากรูปที่ 5.18 แอดทริบิวต์ที่เป็น core คือ  $PD_h^4$  และ  $Z_4$  โดย  $PD_h^4$  คือ ค่าลักษณะของภาพรอบนอก จากการหาผลต่างของแต่ละระยะทางจากกรอบภาพด้านซ้ายไปถึงขอบภาพของพื้นที่  $S_h^4$  ดังแสดงในรูป 5.18(ก) และ  $Z_4$  จำนวนรหัสโซ่ที่มีมากที่สุดในบริเวณที่ 4 ดังแสดงรูปที่ 5.18(ข) ซึ่งทั้งสองแอดทริบิวต์เป็นบริเวณที่อยู่ทางด้านซ้ายตั้งแต่กลางภาพลงไปจนถึงด้านล่างสุดของภาพตัวอักษร จะเห็นได้ว่าบริเวณนี้จะมีความแตกต่างของกลุ่มตัวอักษรที่มีลักษณะเนื้อหาของบริเวณนี้ขนานกับกรอบภาพ เช่น ก, บ เป็นต้น และกลุ่มตัวอักษรที่มีลักษณะเนื้อหาของส่วนนี้มีความชัน เช่น ง, จ, ด เป็นต้น และความแตกต่างของตำแหน่งหัวของตัวอักษรที่อยู่ในส่วนนี้ ตัวอย่างเช่น ก, ถ, ฎ หรือ ม, ฃ, น เป็นต้น แต่ก็ยังไม่เพียงพอสำหรับการระบุตัวอักษรทั้งหมด และตัวอักษรชนิดเดียวกันแต่ต่างฟอนต์ ก็มีความแตกต่างกันของพื้นที่ในส่วนนี้ เช่น ค, ต เป็นต้น บางฟอนต์จะมีความชัน เช่น ฟอนต์ AngsanaUPC และ BrowalliaUPC แต่บางฟอนต์เนื้อหาของส่วนนี้จะขนานกับกรอบภาพ เช่น ฟอนต์ CordiaUPC และ FreesiaUPC ดังตัวอย่างในรูปที่ 5.19

# ค ค ค ค ต ต ต ต

รูปที่ 5.19 ตัวอย่างอักษรต่างฟอนต์ ที่มีความแตกต่างกันของรูปร่าง

ดังนั้นจึงมีการหา reduct โดยอัลกอริทึมการหา reduct จะเพิ่มแอดทริบิว และตรวจสอบค่าความสำคัญของการเพิ่มแอดทริบิวนั้นเข้าไป จะเห็นได้ว่า reduct อาจจะมีมากกว่าหนึ่ง ซึ่งแต่ละ reduct สามารถนำมาใช้แทนกลุ่มของแอดทริบิวทั้งหมดที่ยังไม่ได้ถูกลดโดยที่ไม่เปลี่ยนแปลงค่าการขึ้นต่อกันของข้อมูล การเลือก reduct จึงขึ้นอยู่กับความคิดเห็นว่าจะใช้ reduct ตัวใด กรณีที่แอดทริบิวซึ่งเพิ่มเข้าไปมีความสำคัญเท่ากันจะเลือกแอดทริบิวที่มีจำนวนค่าแอดทริบิวน้อยที่สุด ดังได้อธิบายในหัวข้อ 3.5 ในรูปที่ 5.18 จะมีหมายเลขกำกับเพื่อแสดงลำดับของการเพิ่มแอดทริบิว โดยการเพิ่มเข้าไปแต่ละแอดทริบิวและคำนวณค่าความสำคัญของการเพิ่มแต่ละแอดทริบิวเข้าไป แอดทริบิวของลักษณะเด่น H3, V1 และ H2 ตามลำดับ เป็นแอดทริบิวลักษณะเด่นที่เกิดจากการนับจำนวนของการเปลี่ยนจากพื้นหลังไปเป็นเนื้อหาตามแนวของเส้นที่ลากผ่าน ซึ่งเป็นแอดทริบิวที่ทำให้เห็นความแตกต่างของตัวอักษรที่มีหยักและไม่มีหยักบริเวณส่วนกลางเช่น บ กับ ผ, บ กับ พ หรือ ป กับ ฟ หรือ ป กับ ฝ เป็นต้น แอดทริบิวลักษณะเด่น  $SD_h^4$  ซึ่งเป็นค่าความหนาแน่นของการเปลี่ยนจากพื้นหลังภาพไปเป็นเนื้อหาในแถบ  $S_h^4$  ที่เป็นแถบล่างสุดของตัวอักษรดังรูปที่ 5.18(ข) โดยแอดทริบิวนี้จะตรวจสอบลักษณะโครงสร้างภายในของแถบล่างสุดของตัวอักษร และแอดทริบิว V2 และ V3 ซึ่งเป็นแอดทริบิวลักษณะเด่นที่เกิดจากการนับจำนวนของการเปลี่ยนจากพื้นหลังไปเป็นเนื้อหาตามแนวของเส้นที่ลากผ่าน ดังแสดงในรูปที่ 5.18(ค) เป็นการตรวจสอบลักษณะของตัวอักษรจากเส้นที่ลากตั้งฉาก-แอดทริบิวลักษณะเด่น  $PA_v^4$  และ  $Z_3$  แสดงลักษณะเด่นของตัวอักษรบริเวณแถบขวบน ซึ่ง  $PA_v^4$  คือค่าพื้นหลังพื้นภาพของพื้นที่  $S_v^4$  ของรูปที่ 5.18(ค) โดยคำนวณได้จากระยะทางในแนวตั้งฉากจากกรอบภาพทางด้านบนถึงขอบภาพด้านบน และ  $Z_3$  คือจำนวนรหัสโซ่ที่มากที่สุดที่สุดในบริเวณที่ 3 ดังแสดงในรูปที่ 5.18(ข) และแอดทริบิว  $Z_1$  คือจำนวนรหัสโซ่มากที่สุดที่สุดในบริเวณที่ 1 ดังแสดงในรูปที่ 5.18(ข) เป็นแอดทริบิวที่วิเคราะห์บริเวณด้านซ้ายบนของตัวอักษรได้มากที่สุดที่สุดในจำนวนแอดทริบิว reduct ซึ่งบริเวณนี้จะแสดงความเด่นชัดของส่วนหัวตัวอักษร และจะมีอีกหนึ่งแอดทริบิวที่มีผลต่อการวิเคราะห์ส่วนบนของตัวอักษรคือ  $SD_h^1$  ซึ่งเป็นค่าความหนาแน่นของการเปลี่ยนจากพื้นหลังภาพไปเป็นเนื้อหาในแถบ  $S_h^1$  ที่เป็นแถบบนสุดของตัวอักษรดังรูปที่ 5.18(ข) โดยจะตรวจสอบลักษณะโครงสร้างภายในของแถบบน แอดทริบิวลักษณะเด่น  $PD_v^6$  และ  $PD_v^7$  แสดงลักษณะเด่นของตัวอักษรในส่วนล่างของตัวอักษร ทั้งสองคำนวณผลต่าง

ของแต่ละระยะทางจากกรอบภาพไปถึงขอบภาพของพื้นที่  $S_v^6$  และ  $S_v^7$  ของรูปที่ 5.18(ค) ตามลำดับ ซึ่งมีพื้นที่ติดกัน และคำนวณค่าจากระยะทางในแนวตั้งฉากจากกรอบภาพทางด้านล่างถึงขอบภาพด้านนอก และแอดทริบิวต์ลักษณะเด่น  $PD_h^8$ ,  $PD_h^7$  และ  $Z_6$  จะแสดงลักษณะเด่นของตัวอักษรทางด้านแถบขวา โดย  $PD_h^7$ ,  $PD_h^8$  คือค่าลักษณะของภาพรอบนอก โดยการหาผลต่างของแต่ละระยะทางจากกรอบภาพด้านขวาไปถึงขอบภาพของพื้นที่  $S_h^7$  และ  $S_h^8$  ตามลำดับดังแสดงในรูปที่ 5.18(ก) และ  $Z_6$  คือจำนวนรหัสโซ่ที่มากที่สุดในบริเวณที่ 6 ดังแสดงในรูปที่ 5.18(ข) และในบริเวณส่วนกลางของตัวอักษรนอกจากจะมีแอดทริบิวต์ที่นับจำนวนของการเปลี่ยนจากพื้นหลังไปเป็นเนื้อภาพแล้ว  $Z_5$  ซึ่งจำนวนรหัสโซ่ที่มากที่สุดในบริเวณที่ 5 ดังแสดงในรูปที่ 5.18(ข) ก็เป็นแอดทริบิวต์หนึ่งของ reduct ที่ตรวจสอบส่วนกึ่งกลางของตัวอักษร จากแอดทริบิวต์ reduct ทั้งหมดจะเห็นได้ว่าครอบคลุมทุกส่วนของตัวอักษร โดยแอดทริบิวต์เพียงอย่างเดียวอาจไม่สามารถจำแนก หรือวิเคราะห์ส่วนที่แอดทริบิวต์นั้นวัดได้อย่างชัดเจน ดังนั้นจะเห็นได้ว่าบางส่วนของตัวอักษรจะมีการวัดแอดทริบิวต์มากกว่าหนึ่งอย่าง ตัวอย่างเช่น แอดทริบิวต์  $PA_v^4$ ,  $Z_3$  และ  $SD_h^1$  จะมีส่วนเกี่ยวข้องกับการหาลักษณะเด่นบริเวณด้านบนขวา นั่นก็หมายความว่า การเพิ่มแอดทริบิวต์ที่เกี่ยวข้องกับบริเวณเดียวกันเข้าไป ทำให้เกิดความเด่นชัดของบริเวณนั้นมากยิ่งขึ้นซึ่งจะส่งผลต่อการจำแนกของระบบได้มากยิ่งขึ้น อย่างไรก็ตาม การนำกราฟเชตมาใช้สำหรับเลือกแอดทริบิวต์ที่มีความจำเป็นจะขึ้นอยู่กับค่าความสำคัญของการเพิ่มแอดทริบิวต์แต่ละแอดทริบิวต์ ซึ่งเมื่อค่าความสำคัญของการเพิ่มแอดทริบิวต์มีค่าสูง นั่นก็หมายความว่า ระบบสามารถที่จะจำแนกได้ดียิ่งขึ้น และความเด่นชัดของแต่ละแอดทริบิวต์ลักษณะเด่นจะมีผลเมื่อมีการเพิ่มแอดทริบิวต์ และคำนวณค่าความสำคัญของการเพิ่มแอดทริบิวต์ แอดทริบิวต์ที่ถูกเลือกจะมีความเด่นชัดในการจำแนกมากที่สุด ดังนั้นการแบ่งกลุ่มของแอดทริบิวต์ที่มีความต่อเนื่องเพื่อให้ได้ช่วงที่มีความเหมาะสม จึงมีส่วนสำคัญต่อระบบการจำแนก

### 5.3 กฎการตัดสินใจ

เมื่อหาลักษณะเด่นแบบต่างๆ ของตัวอย่างตัวอักษรสำหรับการเรียนรู้ และนำมาผ่านขั้นตอนการเลือกลักษณะเด่นที่จำเป็นต่อการรู้จำโดยใช้กราฟเชต จะได้แอดทริบิวต์ reduct ซึ่งเป็นเซตของแอดทริบิวต์ลักษณะเด่นที่จำเป็นต่อจำแนกอักษรตัวพิมพ์ภาษาไทย แอดทริบิวต์ของ reduct ประกอบด้วยแอดทริบิวต์ทั้งหมด 18 แอดทริบิวต์ เราจะนำแอดทริบิวต์ของ reduct นี้มาสร้างเป็นกฎการตัดสินใจสำหรับจำแนกชนิดของตัวอักษร โดยจะอยู่ในรูปของ “if...then...rules” ในส่วนของเงื่อนไขประกอบด้วยค่าในแต่ละแอดทริบิวต์ของ reduct จำนวน 18 แอดทริบิวต์ และส่วนของการตัดสินใจจะเป็นผลการจำแนกชนิดตัวอักษร โดยเมื่อนำตัวอักษรมาทดสอบ ก็จะหาลักษณะเด่นแบบต่างๆ ซึ่งอยู่ในแอดทริบิวต์ reduct หลังจากนั้นก็นำค่าของแต่ละแอดทริบิวต์มาเปรียบเทียบกับแต่ละกฎการตัดสินใจ เพื่อหากฎที่มีความสัมพันธ์กับค่าที่วัดได้ของแต่ละแอดทริบิวต์ลักษณะเด่นของตัวอักษรนั้น

และจากกฎที่สัมพันธ์กับค่าที่วัดได้นี้ สามารถแสดงผลของการตัดสินใจในการระบุชนิดของตัวอักษร ตัวอย่างของกฎการตัดสินใจแสดงได้ดังนี้

If a1 = Medium and a2 = Low and a3 = High and .... and a18 = 3 Then char = 'ก'

If a1 = Low and a2 = Low and a3 = High and .... and a18 = 3 Then char = 'ก'

If a1 = Low and a2 = Medium and a3 = High and .... and a18 = 3 Then char = 'ก'

If a1 = Low and a2 = Medium and a3 = High and .... and a18 = 3 Then char = 'ก'

...

If a1 = High and a2 = Low and a3 = High and .... and a18 = 3 Then char = 'ข'

If a1 = Medium and a2 = Low and a3 = High and .... and a18 = 3 Then char = 'ข'

If a1 = High and a2 = Low and a3 = High and .... and a18 = 3 Then char = 'ข'

If a1 = Medium and a2 = Low and a3 = High and .... and a18 = 3 Then char = 'ข'

...

If a1 = Low and a2 = Low and a3 = Low and .... and a18 = 3 Then char = 'ค'

If a1 = Low and a2 = Medium and a3 = Low and .... and a18 = 3 Then char = 'ค'

If a1 = Low and a2 = Medium and a3 = Low and .... and a18 = 3 Then char = 'ค'

If a1 = Low and a2 = Low and a3 = Low and .... and a18 = 3 Then char = 'ค'

...

If a1 = Medium and a2 = Low and a3 = Low and .... and a18 = 3 Then char = 'ช'

If a1 = Medium and a2 = Low and a3 = Low and .... and a18 = 3 Then char = 'ช'

If a1 = High and a2 = Low and a3 = Low and .... and a18 = 3 Then char = 'ช'

If a1 = Medium and a2 = Medium and a3 = Low and .... and a18 = 3 Then char = 'ช'

## บทที่ 6

### การทดลองและผลการทดลอง

ข้อมูลตัวอักษรที่นำมาเรียนรู้อาจมีผลต่อการรู้จำ การนำแบบของตัวอักษรที่ต่างกัน ขนาดที่ต่างกัน มาเรียนรู้อาจหรือทดสอบ จึงส่งผลกระทบต่อประสิทธิภาพในการรู้จำ หรือถ้านำตัวอย่างที่ไม่ดี มีสัญญาณรบกวนมากมาเรียนรู้อาจส่งผลให้เกิดฐานความรู้ที่ไม่มีความสามารถในการจำแนกตัวอักษรอื่นๆ ได้ ดังนั้นจึงจำเป็นต้องเลือกข้อมูลตัวอักษรที่จะนำมาเรียนรู้อาจ และจัดสภาพแวดล้อมในการอ่านข้อมูลที่ดี เพื่อให้เป็นรูปแบบ หรือฐานความรู้สำหรับจำแนกตัวอักษรอื่นๆ ได้

#### 6.1 การเตรียมข้อมูลสำหรับการทดลอง

กลุ่มข้อมูลต้นแบบเป็นกลุ่มข้อมูลของตัวอักษรภาษาไทยตัวพิมพ์ ได้มาจากเครื่องพิมพ์เลเซอร์ HP LaserJet5L ความละเอียด 600 dpi การพิมพ์ตัวอักษรจะใช้โปรแกรม Microsoft word บนเครื่องคอมพิวเตอร์พีซี ฟอนต์ที่ใช้คือ AngsanaUPC, BrowalliaUPC, CordiaUPC, FreesiaUPC และ DilleniaUPC ขนาด 14, 16, 18, 22 พอยท์ โดยกระดาษแต่ละแผ่น เป็นกระดาษสำหรับเครื่องถ่ายเอกสารขนาด 80 แกรม เครื่องที่ประมวลผลเป็นเครื่องไมโครคอมพิวเตอร์ใช้ซีพียู PIII 933 เมกะเฮิร์ตซ์ หน่วยความจำ 128 MB เครื่องอ่านข้อมูลภาพเป็นเครื่องสแกนภาพรุ่น HP ScanJet 5470c ความละเอียดในการอ่าน 300 dpi ทั้งในแนวตั้งและแนวนอน โปรแกรมที่ใช้ในการอ่านภาพคือ Microsoft Paint เครื่องอ่านข้อมูลภาพสามารถกำหนดตัวแปรที่มีผลต่อการทดลองได้ดังนี้

- ก. ชนิดของการอ่านภาพ Type: จากการทดลองใช้ Black and White Drawing เนื่องจากเป็นแบบที่ให้ผลลัพธ์ที่ดีที่สุด
- ข. ความละเอียดในการอ่านข้อมูล ใช้ความละเอียดทางกายภาพสูงสุดคือ 300 จุดต่อนิ้วทั้งแนวนอน และแนวตั้ง

#### 6.2 ขั้นตอนการทดลอง

1. พิมพ์ตัวอักษรฟอนต์ AngsanaUPC, BrowalliaUPC, CordiaUPC, FreesiaUPC ขนาด 14, 16, 18, 22 พอยท์ และนำข้อมูลกระดาษแต่ละแผ่นมาอ่านด้วยเครื่องสแกนภาพ จำนวน 3 ครั้ง สำหรับข้อมูลที่จะนำไปเรียนรู้อาจ
2. พิมพ์ตัวอักษรฟอนต์ AngsanaUPC, BrowalliaUPC, CordiaUPC, FreesiaUPC, DilleniaUPC ขนาด 14, 16, 18, 22 พอยท์ และนำข้อมูลกระดาษแต่ละแผ่นมาอ่านด้วยเครื่องสแกนภาพ จำนวน 10 ครั้ง สำหรับข้อมูลที่จะนำไปทดสอบ
3. นำข้อมูลที่จะนำไปเรียนรู้อาจ ไปสู่ขั้นตอนของการเรียนรู้อาจ

## 4. นำข้อมูลทดสอบไปทดสอบ และวัดความถูกต้อง

## 6.3 ผลการทดลอง

ระบบการรู้จำอักษรตัวพิมพ์นี้ ในการเรียนรู้จะใช้ตัวอักษรจำนวน 4 ฟอนต์ คือ AngsanaUPC, BrowalliaUPC, CordiaUPC และ FreesiaUPC เมื่อผ่านกระบวนการเรียนรู้ และได้ถูกสำหรับ จำแนกตัวอักษรแล้ว นำมาทดสอบความถูกต้องกับตัวอักษรจำนวน 5 ฟอนต์ คือ AngsanaUPC, BrowalliaUPC, CordiaUPC, FreesiaUPC, DilleniaUPC แต่ละฟอนต์ที่นำมาทดสอบมี 4 ขนาดๆ ละ 10 ตัวอย่าง รวมทั้งหมด 8,460 ตัวอย่าง ผลการทดสอบความถูกต้อง และผลสรุปการทดสอบ ความถูกต้อง แสดงในตารางที่ 6.6

ตารางที่ 6.1 แสดงผลการทดสอบความถูกต้องของฟอนต์ AngsanaUPC

ตัวอักษร ทดสอบ	จำนวน		ผลการรู้จำผิดพลาด		
	ถูก	ผิด	ตัวอักษร	ขนาด	จำนวนที่ผิด
ก	40	-			
ข	40	-			
ค	40	-			
ฅ	39	1	ม	16	1
ง	39	1	ธ	14	1
จ	39	1	ฐ	14	1
ฉ	40	-			
ช	32	8	ข	14	2
				16	4
			ซ	22	1
			ย	16	1
ซ	40	-			
ฌ	40	-			
ญ	39	1	ณ	22	1
ฎ	37	3	ฎ	16	2
			ถ	14	1
ฏ	32	8	ฎ	14	5
				16	3

ตารางที่ 6.1 (ต่อ)

ตัวอักษร ทดสอบ	จำนวน		ผลการรู้จำผิดพลาด		
	ถูก	ผิด	ตัวอักษร	ขนาด	จำนวนที่ผิด
ฐ	35	5	จ	16	1
			ฐ	18	1
			ฮ	22	3
ฑ	40	-			
ฒ	40	-			
ณ	40	-			
ด	40	-			
ต	37	3	ค	14	3
ถ	38	2	ภ	14	2
ท	40	-			
ธ	39	1	ง	14	1
น	40	-			
บ	40	-			
ป	40	-			
ผ	40	-			
ฝ	40	-			
พ	40	-			
ฟ	40	-			
ภ	36	4	ก	14	1
			ถ	16	3
ม	36	4	น	22	2
			บ	18	2
ย	40	-			
ร	40	-			
ล	36	4	ฉ	16	2
			ถ	16	2
ว	40	-			
ศ	40	-			

ตารางที่ 6.1 (ต่อ)

ตัวอักษร ทดสอบ	จำนวน		ผลการรู้จำผิดพลาด		
	ถูก	ผิด	ตัวอักษร	ขนาด	จำนวนที่ผิด
ษ	40	-			
ส	38	2	ถ	22	1
			ล	18	1
ห	38	2	ท	22	2
พ	40	-			
อ	40	-			
ช	39	1	ฐ	14	1

จากตารางที่ 6.1 เมื่อวิเคราะห์ตัวอักษรที่เกิดความผิดพลาดในการรู้จำ ตัวอักษร ข และ ช มีความคลุมเครือในการรู้จำ ตัวอย่างของลักษณะเด่นของ ข ของฟอนต์ AngsanaUPC ซึ่งเมื่อนำไปเปรียบเทียบกับกฎแล้วได้ผลการจำแนกเป็นตัวอักษร ข แสดงได้ดังตารางที่ 6.2

ตารางที่ 6.2 ตัวอย่างแอดทริบิวต์ลักษณะเด่นของตัวอักษร ข ฟอนต์ AngsanaUPC ที่จำแนกผิดพลาด

$PD_h^4$	$PD_h^7$	$PD_h^8$	$SD_h^1$	$SD_h^4$	$PA_v^4$	$PD_v^6$	$PD_v^7$	V1
Low	Medium	Low	High	Low1	Medium	Low	Low	2
H2	H3	V2	V3	$Z_1$	$Z_3$	$Z_4$	$Z_5$	$Z_6$
2	3	2	3	2	0	0	1	0

แอดทริบิวต์ที่แสดงความแตกต่างระหว่าง ข และ ช คือ  $Z_3$  และ  $PA_v^4$  ซึ่ง  $PA_v^4$  เป็นข้อมูลที่มีความต่อเนื่อง การแบ่งกลุ่มของข้อมูลที่มีความต่อเนื่อง ถึงแม้จะทำให้แสดงลักษณะของข้อมูลได้มากขึ้น แต่ก็ต้องสูญเสียค่าของข้อมูลไป ทำให้ตัวอักษรบางตัวที่มาทดสอบแสดงผลของส่วนที่แตกต่างกันไม่เด่นชัด จะส่งผลให้การแบ่งกลุ่มของข้อมูลในส่วนที่แตกต่างกันไปอยู่ในกลุ่มเดียวกันได้ ในทำนองเดียวกับ ค และ ต ที่มีความแตกต่างกันที่แอดทริบิวต์  $SD_h^1$  แต่ถ้าข้อมูลบางตัวอย่างของ ค ที่มีส่วนหยักของหัวไม่เด่นชัดมาก ก็ทำให้มีผลต่อแบ่งกลุ่มของข้อมูลได้เช่นเดียวกัน ในกรณีของ ฎ และ ฏ จากลักษณะของฟอนต์ AngsanaUPC จะได้ชัดเจนว่ามีความใกล้เคียงกันมากในบริเวณหยักที่ส่วนหางของตัวอักษร ฎ มีลักษณะของหยักน้อยมาก ในส่วนของแอดทริบิวต์ที่เกี่ยวข้องกับส่วนหยักนี้คือ  $SD_h^4$ ,  $PD_v^6$ ,  $PD_v^7$  ส่วน  $PD_h^8$  นั้นจะตรวจสอบในส่วนของขอบทางด้านขวาของพื้นที่  $S_h^8$  ดังรูปที่ 5.18(ก) จึงไม่ผลต่อวิเคราะห์ส่วนหยักของหาง ฎ และ ฏ ตัวอย่างของลักษณะเด่นของ ฎ

ของฟอนต์ AngsanaUPC ซึ่งเมื่อนำไปเปรียบเทียบกับกฎแล้วได้ผลการจำแนกเป็นตัวอักษร ฎ แสดงได้ดังตารางที่ 6.3

ตารางที่ 6.3 ตัวอย่างแอตทริบิวของตัวอักษร ฎ ฟอนต์ AngsanaUPC ที่จำแนกผิดพลาด

PD <sub>h</sub> <sup>4</sup> Medium	PD <sub>h</sub> <sup>7</sup> Low	PD <sub>h</sub> <sup>8</sup> Low	SD <sub>h</sub> <sup>1</sup> Low	SD <sub>h</sub> <sup>4</sup> Low1	PA <sub>v</sub> <sup>4</sup> Low	PD <sub>v</sub> <sup>6</sup> Medium	PD <sub>v</sub> <sup>7</sup> High	V1 2
H2 2	H3 3	V2 2	V3 3	Z <sub>1</sub> 2	Z <sub>3</sub> 0	Z <sub>4</sub> 0	Z <sub>5</sub> 1	Z <sub>6</sub> 0

ตารางที่ 6.4 แสดงผลการทดสอบความถูกต้องของฟอนต์ BrowalliaUPC

ตัวอักษร ทดสอบ	จำนวน		ผลการรู้จำผิดพลาด		
	ถูก	ผิด	ตัวอักษร	ขนาด	จำนวนที่ผิด
ก	39	1	ภ	22	1
ข	36	4	ช	14	1
			ซ	14	3
ค	37	3	ศ	14	1
				16	2
ฅ	38	2	ญ	14	1
			ณ	14	1
ง	40	-			
จ	38	2	อ	16	2
ฉ	40	-			
ช	36	4	ซ	14	1
				16	2
			ฐ	14	1
ซ	38	2	ช	14	1
				ฐ	14
ฅ	40	-			
ญ	38	2	ฅ	16	1
				ณ	14
ฎ	39	1	ฎ	14	1

ตารางที่ 6.4 (ต่อ)

ตัวอักษร ทดสอบ	จำนวน		ผลการรู้จำผิดพลาด		
	ถูก	ผิด	ตัวอักษร	ขนาด	จำนวนที่ผิด
ฎ	36	4	ฎ	14	4
ฐ	37	3	ง	16	1
			ฮ	22	2
ฑ	40	-			
ฒ	40	-			
ณ	40	-			
ด	37	3	ค	14	2
			ต	14	1
ต	31	9	ก	22	1
			ค	14	4
				22	4
ถ	40	-			
ท	40	-			
ธ	40	-			
น	40	-			
บ	40	-			
ป	40	-			
ผ	40	-			
ฝ -	40	-		-	
พ	40	-			
ฟ	39	1	ฝ	18	1
ภ	37	3	ก	22	1
			ถ	16	2
ม	40	-			
ย	36	4	ข	22	1
			ญ	22	2
			บ	22	1
ร	37	3	ธ	22	3

ตารางที่ 6.4 (ต่อ)

ตัวอักษร ทดสอบ	จำนวน		ผลการรู้จำผิดพลาด		
	ถูก	ผิด	ตัวอักษร	ขนาด	จำนวนที่ผิด
ล	40	-			
ว	40	-			
ศ	40	-			
ษ	40	-			
ส	33	7	ฎ	14	1
				16	2
			ล	14	3
			ฮ	14	1
ห	40	-			
พ	37	3	พ	22	3
อ	40	-			
ฮ	37	3	ฐ	14	3

จากตารางที่ 6.4 เมื่อวิเคราะห์ตัวอักษรที่เกิดความผิดพลาดในการรู้จำ จะเห็นได้ว่าลักษณะตัวอักษร ค และ ต ของฟอนต์ BrowalliaUPC มีความใกล้เคียงกันมาก ซึ่งส่วนหยักบริเวณส่วนบนของตัวอักษร ต ไม่เด่นชัดมากโดยมีความเด่นชัดของส่วนนี้้น้อยกว่าฟอนต์ AngsanaUPC ตัวอย่างของลักษณะเด่นของ ต ของฟอนต์ AngsanaUPC ซึ่งเมื่อนำไปเปรียบเทียบกับกฎแล้วได้ผลการจำแนกเป็นตัวอักษร ค แสดงได้ดังตารางที่ 6.5

ตารางที่ 6.5 ตัวอย่างแอดทริบิวของตัวอักษร ต ฟอนต์ BrowalliaUPC ที่จำแนกผิดพลาด

$PD_h^4$	$PD_h^7$	$PD_h^8$	$SD_h^1$	$SD_h^4$	$PA_v^4$	$PD_v^6$	$PD_v^7$	V1
Low	Low	Low	Low	Low1	Low	Medium	High	2
H2	H3	V2	V3	Z <sub>1</sub>	Z <sub>3</sub>	Z <sub>4</sub>	Z <sub>5</sub>	Z <sub>6</sub>
4	3	3	3	2	1	0	0	3

และเมื่อพิจารณาจาก reduct จะมีความแตกต่างกันที่แอดทริบิว  $SD_h^1$  ทำให้เกิดความผิดพลาดในการจำแนกตัวอักษร ค ค และ ต ส่วน ฎ และ ฎ มีจำนวนตัวอักษรที่ผิดพลาดน้อยกว่าฟอนต์ AngsanaUPC และน้อยกว่าฟอนต์ที่มาทดสอบ เนื่องจากตัวอักษร ฎ ฟอนต์ BrowalliaUPC นี้ มี

ความเด่นชัดของส่วนหยักบริเวณหางมากที่สุด ในจำนวนฟอนต์ทั้งหมดที่นำมาทดสอบ และตัวอักษร ร บางตัวที่รู้จำผิดพลาดเป็น ฐ ตัวอย่างของลักษณะเด่นของ ร ของฟอนต์ BrowalliaUPC ซึ่งเมื่อนำไปเปรียบเทียบกับกฎแล้วได้ผลการจำแนกเป็นตัวอักษร ฐ แสดงได้ดังตารางที่ 6.6

ตารางที่ 6.6 ตัวอย่างแอดทริบิวของตัวอักษร ร ฟอนต์ BrowalliaUPC ที่จำแนกผิดพลาด

$PD_h^4$	$PD_h^7$	$PD_h^8$	$SD_h^1$	$SD_h^4$	$PA_v^4$	$PD_v^6$	$PD_v^7$	V1
Low	Medium	Low	Low	Low1	Low	Low	Low	2
H2	H3	V2	V3	$Z_1$	$Z_3$	$Z_4$	$Z_5$	$Z_6$
4	3	3	3	2	1	0	0	3

เมื่อตรวจสอบจากแอดทริบิว reduct ซึ่งมีผลต่อบริเวณที่ต่างกันของตัวอักษรทั้งสองนี้คือ บริเวณส่วนกลางลงมาจนถึงล่างสุด ได้แก่  $Z_5$  ซึ่งเป็นรหัสลูกโซ่ที่มากที่สุดในส่วนกลางของตัวอักษร และจำนวนของการเปลี่ยนจากพื้นหลังไปเป็นเนื้อภาพตามแนวเส้นที่ลากผ่านในแนวนอน ความผิดพลาดของการรู้จำตัว พ และ ฐ เป็นผลมาจากความผิดพลาดในส่วนการวิเคราะห์ส่วนหางของตัวอักษร ซึ่งเป็นบริเวณขาวบน โดยมีแอดทริบิวลักษณะเด่นที่ตรวจสอบในส่วนนี้ได้แก่  $Z_3$  และ  $PA_v^4$  ตัวอย่างของลักษณะเด่นของ ฐ ของฟอนต์ BrowalliaUPC ซึ่งเมื่อนำไปเปรียบเทียบกับกฎแล้วได้ผลการจำแนกเป็นตัวอักษร ฐ แสดงได้ดังตารางที่ 6.7

ตารางที่ 6.7 ตัวอย่างแอดทริบิวของตัวอักษร ฐ ฟอนต์ BrowalliaUPC ที่จำแนกผิดพลาด

$PD_h^4$	$PD_h^7$	$PD_h^8$	$SD_h^1$	$SD_h^4$	$PA_v^4$	$PD_v^6$	$PD_v^7$	V1
Low	Low	Low	Low	Low1	Low	Low	High	2
H2	H3	V2	V3	$Z_1$	$Z_3$	$Z_4$	$Z_5$	$Z_6$
3	2	2	4	2	1	3	3	1

ตารางที่ 6.8 แสดงผลการทดสอบความถูกต้องของฟอนต์ CordiaUPC

ตัวอักษร ทดสอบ	จำนวน		ผลการรู้จำผิดพลาด		
	ถูก	ผิด	ตัวอักษร	ขนาด	จำนวนที่ผิด
ก	40	-			
ข	31	9	ช	16	2
			ซ	18	3
				22	4

ตารางที่ 6.8 (ต่อ)

ตัวอักษร ทดสอบ	จำนวน		ผลการรู้จำผิดพลาด		
	ถูก	ผิด	ตัวอักษร	ขนาด	จำนวนที่ผิด
ก	36	4	ค	22	4
ฅ	39	1	ท	18	1
ง	40	-			
จ	40	-			
ฉ	40	-			
ช	40	-			
ซ	26	14	ข	14	2
			ซ	14	3
				16	9
ฌ	40	-			
ญ	40	-			
ฎ	26	14	ฎ	14	4
				18	7
				22	3
ฏ	31	9	จ	16	1
			ฎ	14	2
				22	6
ฐ	40	-			
ฑ	40	-			
ฒ	40	-			
ณ	40	-			
ด	39	1	ค	14	1
ต	33	7	ค	14	2
				16	5
ถ	39	1	ภ	14	1
ท	40	-			
ธ	40	-			
น	40	-			

ตารางที่ 6.8 (ต่อ)

ตัวอักษร ทดสอบ	จำนวน		ผลการรู้จำผิดพลาด		
	ถูก	ผิด	ตัวอักษร	ขนาด	จำนวนที่ผิด
บ	40	-			
ป	40	-			
ผ	40	-			
ฝ	40	-			
พ	39	1	ท	22	1
ฟ	40	-			
ภ	38	2	ก	16	1
			ถ	16	1
ม	38	2	ฅ	18	1
			บ	22	1
ย	40	-			
ร	39	1	ธ	16	1
ล	40	-			
ว	40	-			
ศ	39	1	ก	18	1
ษ	40	-			
ส	33	7	ล	14	3
				18	4
ห	40	-			
พ	40	-			
อ	38	2	ณ	14	1
			ถ	14	1
ช	40	-			

จากตารางที่ 6.8 เมื่อวิเคราะห์ตัวอักษรที่เกิดความผิดพลาดในการรู้จำ จะเห็นได้ว่า ตัวอักษร ช มีความคลุมเครือมาก ซึ่งแอดทริบิวต์ลักษณะเด่น  $Z_1$  จะแสดงความเด่นชัดของส่วนหัวของ ช ซึ่งมี

ส่วนหัยก็ได้ไม่ชัดเจน ตัวอย่างของลักษณะเด่นของ ช ของฟอนต์ CordiaUPC ซึ่งเมื่อนำไปเปรียบเทียบกับกฎแล้วได้ผลการจำแนกเป็นตัวอักษร ช แสดงได้ดังตารางที่ 6.9

ตารางที่ 6.9 ตัวอย่างแอดทริบิวของตัวอักษร ช ฟอนต์ CordiaUPC ที่จำแนกผิดพลาด

PD <sub>h</sub> <sup>4</sup>	PD <sub>h</sub> <sup>7</sup>	PD <sub>h</sub> <sup>8</sup>	SD <sub>h</sub> <sup>1</sup>	SD <sub>h</sub> <sup>4</sup>	PA <sub>v</sub> <sup>4</sup>	PD <sub>v</sub> <sup>6</sup>	PD <sub>v</sub> <sup>7</sup>	V1
Low	Low	Low	Low	Low1	Medium	Low	Low	1
H2	H3	V2	V3	Z <sub>1</sub>	Z <sub>3</sub>	Z <sub>4</sub>	Z <sub>5</sub>	Z <sub>6</sub>
3	4	2	2	2	1	1	1	0

ลักษณะตัวอักษร ค และ ต ของฟอนต์ CordiaUPC มีความใกล้เคียงกันมาก ซึ่งส่วนหัยบริเวณส่วนบนของตัวอักษร ค ไม่เด่นชัดมากโดยมีความเด่นชัดของส่วนนี้้น้อยกว่าฟอนต์ AngsanaUPC และเมื่อพิจารณาจาก reduct จะมีความแตกต่างกันที่แอดทริบิว SD<sub>h</sub><sup>1</sup> ทำให้เกิดความผิดพลาดในการจำแนกตัวอักษร ค ค และ ต ตัวอย่างของลักษณะเด่นของ ต ของฟอนต์ CordiaUPC ซึ่งเมื่อนำไปเปรียบเทียบกับกฎแล้วได้ผลการจำแนกเป็นตัวอักษร ค แสดงได้ดังตารางที่ 6.10

ตารางที่ 6.10 ตัวอย่างแอดทริบิวของตัวอักษร ค ฟอนต์ CordiaUPC ที่จำแนกผิดพลาด

PD <sub>h</sub> <sup>4</sup>	PD <sub>h</sub> <sup>7</sup>	PD <sub>h</sub> <sup>8</sup>	SD <sub>h</sub> <sup>1</sup>	SD <sub>h</sub> <sup>4</sup>	PA <sub>v</sub> <sup>4</sup>	PD <sub>v</sub> <sup>6</sup>	PD <sub>v</sub> <sup>7</sup>	V1
Low	Low	Low	Low	Low2	Low	Low	High	1
H2	H3	V2	V3	Z <sub>1</sub>	Z <sub>3</sub>	Z <sub>4</sub>	Z <sub>5</sub>	Z <sub>6</sub>
3	3	3	3	2	1	0	0	1

และมีความคลุมเครือของตัวอักษร ฎ และ ฏ ทำให้เกิดความผิดพลาดในการรู้จำมาก ตัวอย่างของลักษณะเด่นของ ฎ ของฟอนต์ CordiaUPC ซึ่งเมื่อนำไปเปรียบเทียบกับกฎแล้วได้ผลการจำแนกเป็นตัวอักษร ฎ แสดงได้ดังตารางที่ 6.11

ตารางที่ 6.11 ตัวอย่างแอดทริบิวของตัวอักษร ฎ ฟอนต์ CordiaUPC ที่จำแนกผิดพลาด

PD <sub>h</sub> <sup>4</sup>	PD <sub>h</sub> <sup>7</sup>	PD <sub>h</sub> <sup>8</sup>	SD <sub>h</sub> <sup>1</sup>	SD <sub>h</sub> <sup>4</sup>	PA <sub>v</sub> <sup>4</sup>	PD <sub>v</sub> <sup>6</sup>	PD <sub>v</sub> <sup>7</sup>	V1
Low	Low	Low	High	Low1	Low	Medium	High	1
H2	H3	V2	V3	Z <sub>1</sub>	Z <sub>3</sub>	Z <sub>4</sub>	Z <sub>5</sub>	Z <sub>6</sub>
2	2	1	2	2	1	0	1	0

และ ส ของฟอนต์ CordiaUPC มีส่วนทางที่ไม่เด่นชัด ซึ่งแอดทริบิว  $PA_v^4$  และ  $Z_3$  ซึ่งเป็นแอดทริบิว ลักษณะเด่นของบริเวณทางของ ส นี้ไม่สามารถแสดงความเด่นชัดได้ ทำให้เกิดความคล้ำยคลึงกับ ตัวอักษร ล ตัวอย่างของลักษณะเด่นของ ส ของฟอนต์ CordiaUPC ซึ่งเมื่อนำไปเปรียบเทียบกับกฎ แล้วได้ผลการจำแนกเป็นตัวอักษร ล แสดงได้ดังตารางที่ 6.12

ตารางที่ 6.12 ตัวอย่างแอดทริบิวของตัวอักษร ส ฟอนต์ CordiaUPC ที่จำแนกผิดพลาด

$PD_h^4$	$PD_h^7$	$PD_h^8$	$SD_h^1$	$SD_h^4$	$PA_v^4$	$PD_v^6$	$PD_v^7$	V1
Low	Low	Low	Low	Low1	Low	Low	Low	1
H2	H3	V2	V3	$Z_1$	$Z_3$	$Z_4$	$Z_5$	$Z_6$
4	3	2	4	2	1	3	3	1

ตารางที่ 6.13 แสดงผลการทดสอบความถูกต้องของฟอนต์ FreesiaUPC

ตัวอักษร ทดสอบ	จำนวน		ผลการรู้จำผิดพลาด		
	ถูก	ผิด	ตัวอักษร	ขนาด	จำนวนที่ผิด
ก	40	-			
ข	35	5	ช	14	1
			ซ	14	4
ค	39	1	ศ	14	1
ฅ	38	2	ฒ	22	1
ง	40	-			
จ	40	-			
- ฉ	40	-	-		
ช	31	9	ข	22	1
			ซ	14	3
				22	3
			ฮ	14	2
ซ	37	3	ข	22	2
			ช	14	1
ฅ	38	2	ฒ	14	1
			ญ	16	1
ญ	40	-			

ตารางที่ 6.13 (ต่อ)

ตัวอักษร ทดสอบ	จำนวน		ผลการรู้จำผิดพลาด		
	ถูก	ผิด	ตัวอักษร	ขนาด	จำนวนที่ผิด
ฎ	30	10	ฎ	14	4
				16	6
ฏ	33	7	ฎ	14	2
				22	5
ฐ	39	1	ช	22	1
ฑ	40	-			
ฒ	40	-			
ณ	40	-			
ด	35	5	ต	14	2
				16	2
			ก	18	1
ต	37	3	ด	14	3
ถ	40	-			
ท	40	-			
ธ	40	-			
น	40	-			
บ	39	1	ป	18	1
ป	40	-			
ผ	40	-	-		
ฝ	39	1	ก	16	1
พ	40	-			
ฟ	40	-			
ภ	40	-			
ม	40	-			
ย	39	1	อ	14	1
ร	40	-			
ล	40	-			
ว	40	-			

ตารางที่ 6.13 (ต่อ)

ตัวอักษร ทดสอบ	จำนวน		ผลการรู้จำผิดพลาด		
	ถูก	ผิด	ตัวอักษร	ขนาด	จำนวนที่ผิด
ศ	40	-			
ษ	40	-			
ส	39	1	ล	16	1
ห	40	-			
พ	40	-			
อ	40	-			
ช	36	4	ฐ	14	1
			ฑ	14	2
			ร	16	1

จากตารางที่ 6.13 เมื่อวิเคราะห์ตัวอักษรที่เกิดความผิดพลาดในการรู้จำ จะเห็นได้ว่า ตัวอักษร ช มีความคลุมเครือมาก ซึ่งแอดทริบิวต์ลักษณะเด่น  $Z_1$  จะแสดงความเด่นบริเวณนี้ ตัวอย่างของลักษณะเด่นของ ช ของฟอนต์ FreesiaUPC ซึ่งเมื่อนำไปเปรียบเทียบกับกฎแล้วได้ผลการจำแนกเป็นตัวอักษร ช แสดงได้ดังตารางที่ 6.14

ตารางที่ 6.14 ตัวอย่างแอดทริบิวต์ของตัวอักษร ช ฟอนต์ FreesiaUPC ที่จำแนกผิดพลาด

$PD_h^4$	$PD_h^7$	$PD_h^8$	$SD_h^1$	$SD_h^4$	$PA_v^4$	$PD_v^6$	$PD_v^7$	V1
Low	Low	Low	High	Low1	Medium	Low	High	1
H2	H3	V2	V3	$Z_1$	$Z_3$	- $Z_4$	$Z_5$	$Z_6$
2	4	2	2	2	1	3	1	0

ลักษณะตัวอักษร ค และ ต ของฟอนต์ FreesiaUPC มีความใกล้เคียงกันมาก ซึ่งส่วนหยักบริเวณส่วนบนของตัวอักษร ต ไม่เด่นชัดมาก และเมื่อพิจารณาจาก reduct จะมีความแตกต่างกันที่แอดทริบิวต์  $SD_h^1$  ทำให้เกิดความผิดพลาดในการจำแนกตัวอักษร ค และ ต ตัวอย่างของลักษณะเด่นของ ค ของฟอนต์ FreesiaUPC ซึ่งเมื่อนำไปเปรียบเทียบกับกฎแล้วได้ผลการจำแนกเป็นตัวอักษร ค แสดงได้ดังตารางที่ 6.15

ตารางที่ 6.15 ตัวอย่างแอตทริบิวของตัวอักษร ค ฟอนต์ FreesiaUPC ที่จำแนกผิดพลาด

PD <sub>h</sub> <sup>4</sup>	PD <sub>h</sub> <sup>7</sup>	PD <sub>h</sub> <sup>8</sup>	SD <sub>h</sub> <sup>1</sup>	SD <sub>h</sub> <sup>4</sup>	PA <sub>v</sub> <sup>4</sup>	PD <sub>v</sub> <sup>6</sup>	PD <sub>v</sub> <sup>7</sup>	V1
Low	Low	Low	Low	Low2	Low	Low	High	1
H2	H3	V2	V3	Z <sub>1</sub>	Z <sub>3</sub>	Z <sub>4</sub>	Z <sub>5</sub>	Z <sub>6</sub>
4	3	3	3	2	1	0	3	1

และมีความคลุมเครือของตัวอักษร ฎ และ ฏ ทำให้เกิดความผิดพลาดในการรู้จำ และตัวอักษร ช และ ซ

ตารางที่ 6.16 แสดงผลการทดสอบความถูกต้องของฟอนต์ DilleniaUPC

ตัวอักษร ทดสอบ	จำนวน		ผลการรู้จำผิดพลาด		
	ถูก	ผิด	ตัวอักษร	ขนาด	จำนวนที่ผิด
ก	40	-			
ข	29	11	ช	14	2
			ซ	18	1
				22	4
				18	2
				22	2
ค	40	-			
ฅ	37	3	ฅ	16	1
			ม	18	2
ง -	40	-		-	
จ	34	6	ก	18	2
			ค	22	4
ฉ	40	-			
ช	30	10	ข	16	2
			ซ	14	2
				16	3
				18	2
				22	1

ตารางที่ 6.16 (ต่อ)

ตัวอักษร ทดสอบ	จำนวน		ผลการรู้จำผิดพลาด		
	ถูก	ผิด	ตัวอักษร	ขนาด	จำนวนที่ผิด
ซ	27	13	ซ	14	5
				18	3
				22	3
			ฮ	16	2
ณ	39	1	ณ	18	1
ญ	39	1	ณ	14	1
ฉ	23	17	ฉ	14	6
				16	3
				18	3
				22	5
ฉ	21	19	ฉ	14	1
				16	6
				18	4
22	3				
ถ			ถ	18	2
อ			อ	22	3
ฐ	39	1	อ	14	1
ฑ	40	-			
ฒ	40	-			
ณ	40	-			
ด	32	8	ด	14	1
				16	1
				18	1
				22	5
ด	39	1	ด	14	1
ถ	40	-			
ท	40	-			
ธ	40	-			

ตารางที่ 6.16 (ต่อ)

ตัวอักษร ทดสอบ	จำนวน		ผลการรู้จำผิดพลาด		
	ถูก	ผิด	ตัวอักษร	ขนาด	จำนวนที่ผิด
น	40	-			
บ	40	-			
ป	40	-			
ผ	40	-			
ฝ	40	-			
พ	40	-			
ฟ	40	-			
ภ	28	12	ถ	14	1
				18	2
				22	9
ม	34	6	ฅ	18	2
				22	3
			ผ	18	1
ย	34	6	ข	16	1
				18	2
				22	3
ร	40	-			
ล	40	-			
ว	40	-			
ศ	19	21	ค	14	7
				16	6
				18	5
				22	3
ษ	40	-			
ส	39	1	ล	22	1
ห	40	-			
พ	36	4	ฟ	18	2
				22	2

ตารางที่ 6.16 (ต่อ)

ตัวอักษร ทดสอบ	จำนวน		ผลการรู้จำผิดพลาด		
	ถูก	ผิด	ตัวอักษร	ขนาด	จำนวนที่ผิด
อ	40	-			
ฮ	38	2	อ	14	1
				18	1

ตารางที่ 6.16 เป็นการทดสอบความถูกต้องของฟอนต์ DilleniaUPC ซึ่งไม่ได้เป็นฟอนต์ที่นำใช้ในการเรียนรู้ เมื่อวิเคราะห์ตัวอักษรที่เกิดความผิดพลาดในการรู้จำ จะเห็นได้ว่าตัวอักษร ข, ช และ ซ ของตัวอักษรฟอนต์นี้มีความคลุมเครือในการจำแนก ซึ่งแอตทริบิวต์ลักษณะเด่น  $Z_1$  จะแสดงความเด่นบริเวณนี้ ตัวอย่างของลักษณะเด่นของ ข ของฟอนต์ DilleniaUPC ซึ่งเมื่อนำไปเปรียบเทียบกับกฎแล้วได้ผลการจำแนกเป็นตัวอักษร ช แสดงได้ดังตารางที่ 6.17

ตารางที่ 6.17 ตัวอย่างแอตทริบิวต์ของตัวอักษร ข ฟอนต์ DilleniaUPC ที่จำแนกผิดพลาด

$PD_h^4$	$PD_h^7$	$PD_h^8$	$SD_h^1$	$SD_h^4$	$PA_v^4$	$PD_v^6$	$PD_v^7$	V1
Low	Low	Low	High	Low1	Low	Low	Low	1
H2	H3	V2	V3	$Z_1$	$Z_3$	$Z_4$	$Z_5$	$Z_6$
2	4	2	3	2	1	3	1	0

และตัวอักษร ศ ซึ่งมีความผิดพลาดในการรู้จำเป็น ค ก็มีสาเหตุมาจากความผิดพลาดเนื่องจากแอตทริบิวต์  $PA_v^4$  และ  $Z_3$  ซึ่งเป็นแอตทริบิวต์ลักษณะเด่นของบริเวณส่วนหาง ตัวอย่างของลักษณะเด่นของ ศ ของฟอนต์ DilleniaUPC ซึ่งเมื่อนำไปเปรียบเทียบกับกฎแล้วได้ผลการจำแนกเป็นตัวอักษร ค แสดงได้ดังตารางที่ 6.18

ตารางที่ 6.18 ตัวอย่างแอตทริบิวต์ของตัวอักษร ศ ฟอนต์ DilleniaUPC ที่จำแนกผิดพลาด

$PD_h^4$	$PD_h^7$	$PD_h^8$	$SD_h^1$	$SD_h^4$	$PA_v^4$	$PD_v^6$	$PD_v^7$	V1
Low	Low	Low	Low	Low1	Medium	Medium	High	2
H2	H3	V2	V3	$Z_1$	$Z_3$	$Z_4$	$Z_5$	$Z_6$
4	2	2	2	1	1	1	3	1

ตัวอักษร ฎ และ ฏ เกิดจากความผิดพลาดในการจำแนกส่วนหยักของตัวอักษรซึ่งแอดทริบิวที่เกี่ยวข้องกับบริเวณส่วนหยักนี้คือ  $SD_h^4$ ,  $PD_v^6$ ,  $PD_v^7$  ตัวอย่างของลักษณะเด่นของ ฎ ของฟอนต์ DilleniaUPC ซึ่งเมื่อนำไปเปรียบเทียบกับกฎแล้วได้ผลการจำแนกเป็นตัวอักษร ฎ แสดงได้ดังตารางที่ 6.19 และตัวอย่างของลักษณะเด่นของ ฏ ของฟอนต์ DilleniaUPC ซึ่งเมื่อนำไปเปรียบเทียบกับกฎแล้วได้ผลการจำแนกเป็นตัวอักษร ฎ แสดงได้ดังตารางที่ 6.20

ตารางที่ 6.19 ตัวอย่างแอดทริบิวของตัวอักษร ฎ ฟอนต์ DilleniaUPC ที่จำแนกผิดพลาด

$PD_h^4$	$PD_h^7$	$PD_h^8$	$SD_h^1$	$SD_h^4$	$PA_v^4$	$PD_v^6$	$PD_v^7$	V1
Low	Low	Low	Low	Low1	Low	Medium	Low	1
H2	H3	V2	V3	$Z_1$	$Z_3$	$Z_4$	$Z_5$	$Z_6$
2	2	2	2	2	1	0	1	0

ตารางที่ 6.20 ตัวอย่างแอดทริบิวของตัวอักษร ฎ ฟอนต์ DilleniaUPC ที่จำแนกผิดพลาด

$PD_h^4$	$PD_h^7$	$PD_h^8$	$SD_h^1$	$SD_h^4$	$PA_v^4$	$PD_v^6$	$PD_v^7$	V1
Medium	Low	Low	Low	Low2	Low	Medium	Low	1
H2	H3	V2	V3	$Z_1$	$Z_3$	$Z_4$	$Z_5$	$Z_6$
2	2	2	3	3	1	0	1	0

และตัวอักษร ภ ของฟอนต์ DilleniaUPC เกิดความผิดพลาดไปเป็น ฎ โดยลักษณะเด่นที่เกี่ยวข้องกับความผิดพลาดส่วนนี้คือ  $PD_h^4$  ซึ่งเป็นแสดงค่าลักษณะของภาพรอบนอกของบริเวณส่วนหัวของตัวอักษร ภ ตัวอย่างของลักษณะเด่นของ ภ ของฟอนต์ DilleniaUPC ซึ่งเมื่อนำไปเปรียบเทียบกับกฎแล้วได้ผลการจำแนกเป็นตัวอักษร ฎ แสดงได้ดังตารางที่ 6.21

ตารางที่ 6.21 ตัวอย่างแอดทริบิวของตัวอักษร ภ ฟอนต์ DilleniaUPC ที่จำแนกผิดพลาด

$PD_h^4$	$PD_h^7$	$PD_h^8$	$SD_h^1$	$SD_h^4$	$PA_v^4$	$PD_v^6$	$PD_v^7$	V1
Medium	Low	Low	Low	Low1	Low	Low	Low	2
H2	H3	V2	V3	$Z_1$	$Z_3$	$Z_4$	$Z_5$	$Z_6$
2	2	2	3	1	1	0	0	0

ตารางที่ 6.22 ผลสรุปการทดสอบความถูกต้อง

พอนต์ตัวอักษร	เปอร์เซ็นต์การรู้จำถูกต้อง
AngsanaUPC	97.02%
BrowalliaUPC	96.19%
CordiaUPC	95.48%
FreesiaUPC	96.67%
เฉลี่ยรวม	96.34%

ตารางที่ 6.23 ผลสรุปการทดสอบของพอนต์ DilleniaUPC

พอนต์ตัวอักษร	เปอร์เซ็นต์การรู้จำถูกต้อง
DilleniaUPC	91.45%

จากผลการทดสอบความถูกต้องของพอนต์ในตารางที่ 6.22 เมื่อวิเคราะห์ตัวอักษรที่มีความผิดพลาดมากจะเห็นได้ว่าเป็นความผิดพลาดอันเกิดจากความคลุมเครือของตัวอักษรที่ไม่สามารถแสดงความเด่นชัดของบริเวณที่แสดงให้เห็นถึงความแตกต่างกันของตัวอักษรทั้งสองชนิด นั่นคือ ตัวอักษร ช และ ซ มีความผิดพลาดของการรู้จำที่ชัดเจนเนื่องจากแอดทริบิวต์ลักษณะเด่น  $Z_1$  แสดงความเด่นชัดของส่วนหัวของ ช และ ซ ซึ่งมีส่วนหยักได้ไม่ชัดเจน ส่วนตัวอักษร ค และ ต มีความใกล้เคียงกันมาก โดยแตกต่างกันในส่วนหยักบริเวณส่วนบนของตัวอักษร ค ซึ่งไม่เด่นชัดมาก และเมื่อพิจารณาจาก reduct จะมีความแตกต่างกันที่แอดทริบิวต์  $SD_h^1$  ทำให้เกิดความผิดพลาดในการจำแนกตัวอักษร ค และ ต อีกกลุ่มตัวอักษรที่มีความคลุมเครือคือ ฎ และ ฏ จะมีความใกล้เคียงกันมากในบริเวณหยักที่ส่วนหางของตัวอักษร ในส่วนของแอดทริบิวต์ที่เกี่ยวข้องกับส่วนหยักนี้คือ  $SD_h^4$ ,  $PD_v^6$ ,  $PD_v^7$  ซึ่งถ้าส่วนนี้ไม่มีความเด่นชัด จะส่งผลต่อการจำแนกตัวอักษร ทำให้เกิดความผิดพลาดของการรู้จำ จากผลการทดสอบทั้งหมดจะเห็นได้ว่า ลักษณะเด่นมีผลต่อความสามารถในการจำแนกของระบบ ดังนั้นการหาลักษณะเด่นที่สามารถแสดงความแตกต่างของบริเวณที่มีความแตกต่างกันของตัวอักษรสองคนละชนิดที่มีความคล้ายคลึงกันมาก จะทำให้สามารถจำแนกตัวอักษรที่มีความคล้ายคลึงกันได้ สำหรับการทดสอบพอนต์ที่ไม่ได้นำมาใช้ในการเรียนรู้ดังแสดงผลการทดสอบความถูกต้องในตารางที่ 6.23 นอกจากความผิดพลาดของการจำแนกตัวอักษรที่พบบ่อยเหมือนกับพอนต์อื่นแล้ว ยังพบความผิดพลาดของการจำแนกตัวอักษรบางตัวได้แก่ตัวอักษร ศ ซึ่งความผิดพลาดเป็นผลมาจากแอดทริบิวต์  $PA_v^4$  และ  $Z_3$  ซึ่งเป็นแอดทริบิวต์ลักษณะเด่นของบริเวณส่วนหาง และตัวอักษร ภ โดยลักษณะเด่นที่เกี่ยวข้องกับความผิดพลาดส่วนนี้คือ  $PD_h^4$  ซึ่งเป็นแสดงค่าลักษณะของภาพรอบนอกของบริเวณส่วนหัวของตัวอักษร ภ

## บทที่ 7

# สรุปและข้อเสนอแนะ

### 7.1 สรุป

กราฟเซตเป็นวิธีในการวิเคราะห์ความคลุมเครือของข้อมูล โดยการหาขอบเขตบน และขอบเขตล่างของข้อมูล และวิเคราะห์การขึ้นต่อกันของข้อมูล เพื่อเลือกแอตทริบิวต์ที่มีความจำเป็นต่อการจำแนก โดยที่ไม่เปลี่ยนแปลงค่าความขึ้นต่อกันไปจากข้อมูลที่ยังไม่มีการลดแอตทริบิวต์ โดยในวิทยานิพนธ์ฉบับนี้ได้นำกราฟเซตมาใช้ในการเรียนรู้ระบบรู้จำอักษรตัวพิมพ์ภาษาไทย กราฟเซตมีคุณสมบัติเด่นในเรื่อง

1. การวิเคราะห์การขึ้นต่อกันของข้อมูล โดยกราฟเซตสามารถนำมาเลือกแอตทริบิวต์ที่มีความจำเป็นต่อการจำแนกมีความสามารถในการจำแนกเหมือนกับข้อมูลทั้งหมด
2. กราฟเซตวิเคราะห์คุณสมบัติจากตารางข้อมูล และไม่ต้องระบุค่าเริ่มต้นใดๆ
3. กฎที่สร้างจากกราฟเซตมีพื้นฐานมาจากข้อมูลจริง เพราะว่าแต่ละกฎสร้างได้จากตารางข้อมูลที่ผ่านขั้นตอนของการวิเคราะห์หาแอตทริบิวต์ที่มีความจำเป็นต่อการจำแนกโดยกราฟเซต

แต่ข้อมูลในแต่ละแอตทริบิวต์ที่จะนำมาเรียนรู้โดยกราฟเซต จะเป็นข้อมูลที่ไม่มีความต่อเนื่อง ซึ่งข้อมูลของบางแอตทริบิวต์ที่วัดได้จะมีความต่อเนื่องของข้อมูล ดังนั้นจึงจำเป็นต้องมีการแบ่งช่วงของข้อมูล โดยได้เสนอวิธีการแบ่งช่วงของข้อมูลโดยใช้ฟัซซี่เซต ซึ่งได้มีการนำอัลกอริทึม Fuzzy c-Means มาแบ่งช่วงของข้อมูลเพื่อให้เกิดการแบ่งช่วงที่มีความเหมาะสม ซึ่งมีผลอย่างมากต่อการวิเคราะห์ค่าการขึ้นต่อกันในกราฟเซต และความสามารถในการจำแนก

การรู้จำอักษรตัวพิมพ์ภาษาไทยโดยใช้ฟัซซี่เซตและกราฟเซต มีขั้นตอนคือ นำตัวอักษรมาหาลักษณะเด่นหลายอย่าง เพื่อใช้กราฟเซตในการเลือกแอตทริบิวต์ลักษณะเด่นที่มีความจำเป็นต่อการจำแนกตัวอักษร โดยลักษณะเด่นใดที่ข้อมูลมีความต่อเนื่อง จะนำมาแบ่งช่วงโดยใช้ฟัซซี่เซต และวิเคราะห์จำนวนกลุ่มที่มีความเหมาะสมที่สุด

การทดสอบการรู้จำ พบว่าตัวอักษรที่มีส่วนที่มีลักษณะโดดเด่นสามารถจำแนกโดยใช้ฟัซซี่เซตและกราฟเซตได้อย่างถูกต้อง แต่พบว่าอักษรของภาษาไทยบางตัวมีความคล้ายคลึงกันมาก และมีรายละเอียดที่แตกต่างกันเพียงเล็กน้อย ซึ่งอาจจะต้องหาลักษณะเด่นที่สามารถแสดงความแตกต่างของลักษณะรายละเอียดที่มีความแตกต่างกันนี้มาเรียนรู้ด้วย หรือนำวิธีพิเศษต่างๆ เพื่อนำมาวิเคราะห์ส่วนที่ต่างกันเพียงเล็กน้อยนั้น เพื่อให้มีความสามารถในการจำแนกตัวอักษรนั้นได้

## 7.2 ข้อเสนอแนะ

1. จากการทดลองเมื่อมีตัวอักษรที่ไม่ตรงกับเงื่อนไขใดในกฎ ในการตัดสินใจจะมีการเลือกกฎที่มีเงื่อนไขที่ตรงกับข้อมูลตัวอักษรมากที่สุด ซึ่งอาจจะมีได้มากกว่า 1 กฎ ทำให้มีความคลุมเครือในการจำแนกตัวอักษร ดังนั้นจึงควรหาวิธีสำหรับประเมินความสามารถของกฎในการจำแนก
2. จากการศึกษาในหลายๆ งานวิจัย พบว่ามีวิธีในการหา reduct หลายวิธี ดังนั้นน่าจะมีการทดลองหา reduct โดยวิธีอื่น เพื่อหา reduct ที่มีความสามารถในการจำแนกที่ดีที่สุด
3. เพิ่มลักษณะเด่นสำหรับการเรียนรู้โดยราฟเซตให้มากขึ้น เพื่อให้ราฟเซตสามารถเลือกลักษณะเด่นที่มีความเด่นชัดมากที่สุด ทำให้ระบบสามารถรู้จำตัวอักษรได้มากขึ้น
4. ศึกษาวิธีในการรวมกฎที่มีเงื่อนไขร่วมกัน กฎหลายกฎสามารถที่จะรวมเป็นกฎเดียวกันได้ โดยที่ยังคงความสามารถในการจำแนก

## เอกสารอ้างอิง

1. C. Kimpan, A.Itoh, and K. Kawanishi, "Recogniton of printed Thai character using a matching method", IEE Proc., Vol. 130, Pt.E, No. 6, P. 183-188, Nov. 1983.
2. C. Kimpan, "Printed Thai character recognition using topological properties method", INT.J. Electronics, Vol. 60, No. 3, P. 303-329, 1986.
3. ชม กิมปาน. "การจดจำรูปแบบตัวพิมพ์อักษรภาษาไทย", วิทยานิพนธ์วิศวกรรมศาสตรดุษฎีบัณฑิต สาขาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง. 2529.
4. Z. PAWLAK, "Why rough sets?", Proceedings of the Fifth IEEE International Conference on Fuzzy Systems, 1996., Volume: 2 , 8-11 Sep 1996, pp. 738 -743 vol.2.
5. Z. PAWLAK, "Rough Set Approach to Knowledge-Base Decision Support", 14<sup>th</sup> European Conference on Operational Research, pp. 1-12, 1995.
6. G. Lambert-Torres, A.P.A. da Silva, V.H. Quintana and L.E. Borges da Silva, "Knowledge-base reduction based on rough set techniques", Canadian Conference on Electrical and Computer Engineering, 1996., Volume: 1 , 26-29 May 1996, pp. 278 -281 vol.1.
7. G. Lambert-Torres, A.P.A. da Silva, V.H. Quintana and L.E. Borges da Silva, "Classification of power system operation point using rough set techniques", IEEE International Conference on Systems, Man, and Cybernetics, 1996., Volume: 3 , 14-17 Oct 1996, pp. 1898 -1903 vol.38.
8. X. Hu. "Knowledge Discovery in Database: An Attribute-Oriented Rough Set Approach" Ph.D. Thesis of University of Regina. 1995.
9. George J. Klir et. al., Fuzzy Set Theory: Foundations and Applications, Prentice Hall International, Inc.
10. CHIN-TENG LIN and C.S. GEORGE LEE, Neural Fuzzy Systems: A Neuro-Fuzzy Synergism to Intelligent Systems, Prentice Hall International, Inc.
11. ศิริชัย ปรีดโตทกพร. "การแยกกลุ่มข้อมูลภาพทางการแพทย์โดย Fuzzy-C-Means และการพิจารณา Histogram ของภาพ", วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง. 2542.
12. ศักดิ์พงษ์ จันทร์ไพบ, "ฟัซซี่ลอจิก." คอมพิวเตอร์รีวิว ฉบับที่ 111, พฤศจิกายน 2536. หน้า 236-282.

13. อุกฤษณ์ มารังค์. “การรู้จำอักษรตัวพิมพ์ภาษาไทยโดยใช้เจนนเน็ติก-นิวรอลเน็ตเวิร์ค”, วิทยานิพนธ์  
วิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย, สถาบันเทคโนโลยีพระ  
จอมเกล้าเจ้าคุณทหารลาดกระบัง. 2544.
14. Yuan Y. Tang, Lo-Ting Tu. “Off-line recognition of Chinese handwriting by multifeature and  
multilevel classification”, IEEE Trans. on Pattern analysis and Machine Intelligence, vol. 20.  
No. 5, May 1998, pp. 556-561.
15. E.M. de A. Neves, A. Gonzaga, A.F.F. Slaets, “A multi-font character recognition based on its  
fundamental features by artificial neural networks”, Cybernetic Vision, 1996. Proceedings.,  
Second Workshop on , 9-11 Dec 1996, pp. 196 -201.

## ภาคผนวก ก.

## ผลงานวิจัยที่ได้รับการตีพิมพ์

1. T. Suebsanit and C. Kimpan, "Printed Thai Character Recognition using Multifeature and Multilevel Classification", SCORED 2001, paper 167, Feb. 20-21, 2001.
2. P. Phokharatkul, T. Suebsanit, P. Pantaragphong, O. Pinngern and C. Kimpan, "Printed Thai Character Recognition using Rough Classification", Proceeding of The International Conference on Artificial Intelligence in Engineering and Technology 2002, Malaysia, Jun. 17-18, 2002, pp. 284-290.
3. T. Suebsanit, P. Phokharatkul, P. Pantaragphong, O. Pinngern and C. Kimpan, Recognition of Printed Thai Characters using Fuzzy and Rough Sets Theory", Proceedings of The International Conference on Robotics, Vision, Information and Signal Processing ROVISIP 2003, Malaysia, Jan. 22-24, 2003, pp.33-38.

## ประวัติผู้เขียน

นายธีราช สืบสนิท สำเร็จการศึกษาปริญญาตรีวิศวกรรมศาสตรบัณฑิตสาขาวิศวกรรมไฟฟ้า จากมหาวิทยาลัยมหิดล ปีการศึกษา 2541 ทำงานเป็นผู้ช่วยวิจัย ห้องวิจัย Information Science สำนักวิจัยสื่อสารและเทคโนโลยีสารสนเทศ (ReCCIT) 1 ตุลาคม 2543 ถึง 30 กันยายน 2544 และทำงานเป็นเจ้าหน้าที่ระบบคอมพิวเตอร์ คณะวิทยาศาสตร์ 17 มกราคม 2545 ถึงปัจจุบัน