บทที่ 1

บทนำ

ความรู้เป็นทรัพยากรที่มีความสำคัญสำหรับบุคลากรและองค์กรเป็นอย่างมาก ดังนั้นได้มีการนำความรู้มาใช้เพื่อให้เกิดประโยชน์สำหรับบุคคลและองค์กร เช่น การนำความรู้นั้น ๆ มาใช้ในเรื่องของการสร้างความได้เปรียบทางด้านการแข่งขันเมื่อมองระดับขององค์กร หรือนำความรู้มาใช้ในการตัดสินใจแก้ปัญหาต่าง ๆ หรือเพื่อเพิ่มพูนความรู้สำหรับแต่ละบุคคล [1, 2] เนื่องจากในปัจจุบันข้อมูลที่เก็บอยู่ในฐานข้อมูลนั้นมีเพิ่มขึ้นมาก ทำให้การนำข้อมูลในฐานข้อมูลมาใช้ประโยชน์นั้นเป็นไปได้ยากมากยิ่งขึ้น ดังนั้นจึงต้องมีการนำข้อมูลเหล่านั้นมาทำการสกัดหาความรู้ ซึ่งความรู้นั้นอยู่ในรูปของกฎ เพื่อทำให้เข้าใจข้อมูลเหล่านั้นได้ง่ายขึ้นและสามารถนำไปใช้ประโยชน์ได้ต่อไป

การทำเหมืองข้อมูล (Data Mining) เป็นแนวทางหนึ่งที่นำมาช่วยในการค้นหาความรู้จากฐานข้อมูลที่มีจำนวนข้อมูลมาก เพื่อให้ได้มาซึ่งความรู้ที่เป็นประโยชน์นั้น จึงได้มีการนำตัวอย่างเทคนิคการทำเหมืองข้อมูลมาใช้ ไม่ว่าจะเป็น เทคนิคการจัดกลุ่มข้อมูล (Clustering Technique) ที่มีการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) และเทคนิคการสกัดความรู้ เป็นต้น การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Clustering) เป็นการเรียนรู้เพื่อแบ่งกลุ่มข้อมูลเอง เป็นการจัดกลุ่มข้อมูลที่มีลักษณะเหมือนกันให้อยู่ในกลุ่มเดียวกัน โดยไม่จำเป็นต้องมีตัวอย่างข้อมูลในการสอน ตัวอย่างเทคนิคการเรียนรู้แบบไม่มีผู้สอน ได้แก่ แผนที่การจัดกลุ่มเอง (Self-Organizing Map : SOM) และการแบ่งกลุ่ม K กลุ่มด้วยค่าเฉลี่ย (K-Means) [3] สำหรับแผนที่การจัดกลุ่มเองนั้นได้มีการนำมาประยุกต์ใช้ในงานวิจัยทางด้านการทำเหมืองข้อมูลอย่างแพร่หลายในด้านต่าง ๆ เช่น นำวิธีการแผนที่การจัดกลุ่มเองมาใช้ในการแก้ปัญหาเกี่ยวกับการจัดกลุ่มข้อมูล นำมาประยุกต์ใช้ร่วมกับ Case-Based Reasoning เพื่อทำนายยอดขายหนังสือใหม่และทำนายการจัดการอัตราของพันธบัตร [4, 5] และนำวิธีการแผนที่การจัดกลุ่มเองมาวิเคราะห์ข้อมูลทางด้านการตลาด [6] เป็นต้น

การสกัดความรู้เป็นเทคนิคหนึ่งของการทำเหมืองข้อมูล ซึ่งเป็นแนวทางที่นำมาช่วยในการค้นหาความรู้จากฐานข้อมูลที่มีข้อมูลจำนวนมาก เพื่อให้ได้มาซึ่งความรู้ที่เป็นประโยชน์และสามารถเข้าใจลักษณะของข้อมูลได้ง่าย โดยเฉพาะอย่างยิ่งข้อมูลที่มีลักษณะเป็นตัวเลข ความรู้ที่สกัดได้นั้นอยู่ในรูปของกฎ “ถ้า-แล้ว” (“If-Then” Rules) ซึ่งมี 2 แบบคือ กฎทั่วไป (Crisp Rules) ซึ่งแทนด้วยค่าจำนวนจริง [7] และกฎภาษาธรรมชาติ (Linguistic Rules) ซึ่งแทนด้วยภาษาธรรมชาติ เช่น เล็ก กลาง และ ใหญ่ เป็นต้น [8] ลักษณะกฎทั้งสองแบบนั้นมีองค์ประกอบ 2 ส่วน โดยที่ส่วนแรกที่อยู่หลัง “ถ้า” หรือ “If” จะหมายถึงเหตุ และส่วนที่สองที่อยู่หลัง “แล้ว” หรือ “Then” หมายถึงผลลัพธ์ที่ตามมา รายละเอียดของแต่ละแบบมีดังนี้

1. กฎทั่วไป (Crisp Rules/Coventional Rules) มีรูปแบบดังสมการ (1.1)

$$\text{If } x_1 \text{ op } t_1 \text{ and/or ... and/or } x_i \text{ op } t_j \text{ then } c_k \qquad (1.1)$$

โดยที่ x_i คือ ตัวแปรข้อมูลเข้าตัวที่ i
t_j คือ ค่าของตัวแปรข้อมูลเข้าที่เป็นจำนวนจริงที่ j
op คือ ตัวดำเนินการ =, <, >, <=, >= หรือ ≠
c_k คือ กลุ่มข้อมูล (Class) ที่ k

ตัวอย่างกฎทั่วไปจากฐานข้อมูลโรคมะเร็งเต้านม (Wisconsin Breast Cancer Database) เช่น If x_2 = 1 or 2 then benign (ไม่เป็นโรคมะเร็งเต้านม) [7] เป็นต้น

2. กฎภาษาธรรมชาติ (Linguistic Rules) มีรูปแบบดังสมการ (1.2)

$$\text{If } x_1 \text{ is } l_1 \text{ and/or ... and/or } x_i \text{ is } l_j \text{ then } c_k \qquad (1.2)$$

โดยที่ x_i คือ ตัวแปรข้อมูลเข้าตัวที่ i
l_j คือ ค่าของตัวแปรข้อมูลเข้าที่ j มีลักษณะเป็นรูปแบบภาษาธรรมชาติ (Linguistic Terms) เช่น เล็ก กลาง และใหญ่ เป็นต้น
c_k คือ กลุ่มข้อมูล (Class) ที่ k

ตัวอย่างกฎภาษาธรรมชาติจากฐานข้อมูลดอกไม้ไอริช (Iris Flower Database) เช่น If x_3 is small and x_4 is small then Setosa [8] เป็นต้น

วิทยานิพนธ์นี้ได้นำเสนอแบบจำลองการสกัดความรู้โดยใช้แผนที่การจัดกลุ่มเอง ซึ่งแผนที่การจัดกลุ่มเอง ใช้จัดกลุ่มข้อมูลที่มีลักษณะเหมือนกันให้อยู่ในกลุ่มเดียวกันโดยไม่มีตัวอย่างในการสอน ทำการลดมิติของข้อมูลเข้าให้มีจำนวนมิติเป็น 1 มิติ หรือ 2 มิติ [9, 10] มีประโยชน์ คือทำให้สามารถเข้าใจลักษณะภาพรวมของข้อมูลได้ง่าย ซึ่งความรู้ที่สกัดได้จากแผนที่การจัดกลุ่มเอง อยู่ในรูปของกฎ “ถ้า-แล้ว” ที่ผู้ใช้สามารถเข้าใจได้ง่าย และใช้ฐานข้อมูลที่เป็นมาตรฐานจาก University of California at Irvine มาเป็นข้อมูลทดสอบ

1.1 การตรวจเอกสาร

เทคนิคที่ใช้ในการสกัดกฎ คือ โครงข่ายประสาทเทียม (Neural Networks) ต้นไม้การตัดสินใจ (Decision Trees) ฟัซซี่เซต (Fuzzy Set) และทฤษฎีราฟเซต (Rough Set Theory) ดังรายละเอียดต่อไปนี้

1.1.1 โครงข่ายประสาทเทียม (Neural Networks)

โครงข่ายประสาทเทียมถูกสร้างขึ้นเพื่อเลียนแบบการทำงานของสมองมนุษย์ หลักการทำงานประกอบด้วยหน่วยประมวลผลย่อยหลาย ๆ หน่วยทำงานเชื่อมต่อกัน แต่ละหน่วยสามารถปรับค่าพารามิเตอร์ประจำหน่วยได้จากกระบวนการเรียนรู้ โครงข่ายประสาทเทียมที่ผ่านการฝึกฝนจะสามารถนำไปใช้แก้ปัญหาจริงได้ [11] ข้อดีของโครงข่ายประสาทเทียม คือ สามารถทำนายได้ค่าความถูกต้องสูง เหมาะกับข้อมูลที่มีลักษณะเป็นตัวเลข

โครงข่ายประสาทเทียมที่นำมาใช้ในการสกัดกฎ แบ่งตามสถาปัตยกรรม มี 3 แบบคือ แผนที่การจัดกลุ่มเอง (Self-Organizing Map: SOM) โครงข่ายประสาทเทียมแบบย้อนกลับ (Recurrent Neural Networks) และโครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น (Multilayer Perceptron Neural Networks: MLP) มีรายละเอียดดังต่อไปนี้

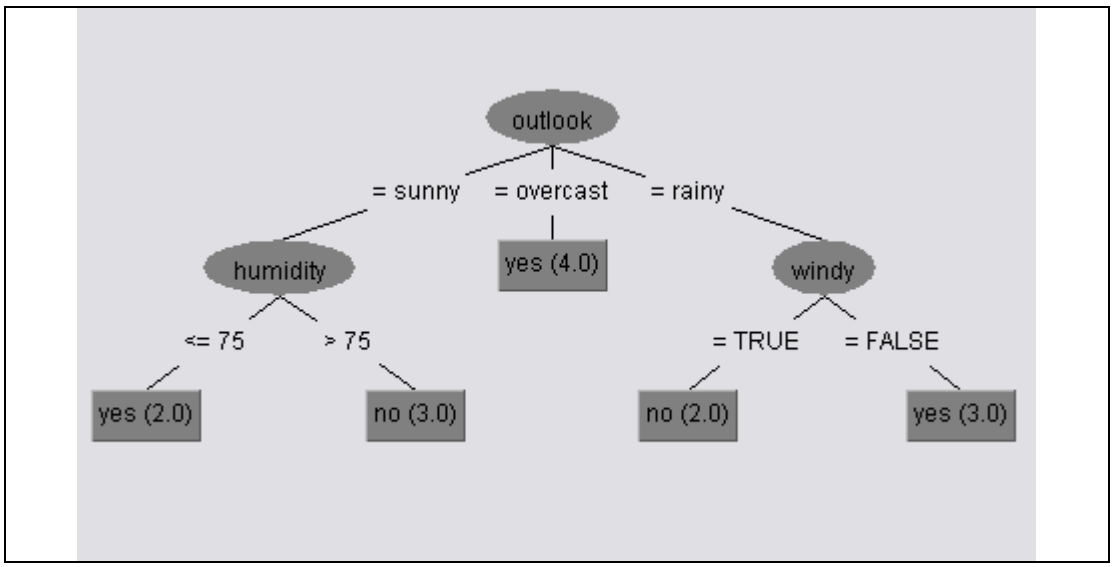
1. แผนที่การจัดกลุ่มเอง (Self-Organizing Map: SOM) เป็นวิธีการหนึ่งของโครงข่ายประสาทเทียมที่มีการเรียนรู้แบบไม่มีผู้สอนซึ่งใช้กันอย่างแพร่หลาย นำมาใช้ในเรื่องของการจัดกลุ่มข้อมูลที่มีลักษณะเหมือนกันให้อยู่ในกลุ่มเดียวกัน ซึ่งประโยชน์ของแผนที่การจัดกลุ่มเองนั้น จะทำการลดมิติของข้อมูลเข้าให้มีจำนวนมิติเป็น 1 มิติ หรือ 2 มิติ และสามารถเข้าใจลักษณะข้อมูลได้ในภาพรวม [10, 12] สำหรับการสกัดความรู้โดยใช้แผนที่การจัดกลุ่มเองเพียงอย่างเดียวนั้น ได้มีการนำ U-matrix และ Component plane ที่เป็นเครื่องมือของแผนที่การจัดกลุ่มเอง และขั้นตอนวิธี sig* มาทำการวิเคราะห์หาขอบเขตของประเภทข้อมูล รวมทั้งใช้เพื่ออธิบายความสัมพันธ์ของข้อมูลให้อยู่ในรูปของกฎ “ถ้า-แล้ว” ของแต่ละข้อมูลเข้า [12, 13] สำหรับกฎที่สกัดได้จากแผนที่การจัดกลุ่มเองนั้นอยู่ในรูปของกฎทั่วไป (Crisp Rules) นอกจากนี้ได้นำแผนที่การจัดกลุ่มเองมาประยุกต์ใช้ร่วมกับวิธีการอื่น ๆ เพื่อเพิ่มประสิทธิภาพและความถูกต้องในการสกัดความรู้ เช่น การสกัดความรู้โดยใช้แผนที่การจัดกลุ่มเองร่วมกับกฎฟัซซี่ [14] ซึ่งกฎที่ได้นั้นอยู่ในรูปของกฎภาษาธรรมชาติ (Linguistic Rules) การหากฎความสัมพันธ์โดยใช้แผนที่การจัดกลุ่มเองร่วมกับทฤษฎีราฟเซต [15] การนำแผนที่การจัดกลุ่มเองมาทำการจัดกลุ่มข้อมูล และใช้ทฤษฎีราฟเซตเพื่อทำการสกัดความรู้ให้อยู่ในรูปของกฎ “ถ้า-แล้ว” เพื่อสกัดความรู้ที่เป็นประโยชน์จากฐานข้อมูล [16]

2. โครงข่ายประสาทเทียมแบบย้อนกลับ (Recurrent Neural Networks) เป็นโครงประสาทเทียมที่มีการเรียนรู้แบบมีผู้สอน โดยสัญญาณจากข้อมูลออก (Output) สามารถวนกลับไปเป็นข้อมูลเข้า (Input) ได้ [17] ข้อเสียของโครงข่ายประสาทเทียมแบบย้อนกลับ คือใช้เวลาในการประมวลผลนาน เมื่อเทียบกับโครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น [7]

3. โครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น (Multilayer Perceptron Neural Networks: MLP) เป็นโครงประสาทเทียมที่มีการเรียนรู้แบบมีผู้สอน นิยมใช้ในการแบ่งกลุ่มข้อมูล ในการสกัดกฎจากโครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น มีทั้งได้กฎทั่วไป (Crisp Rules/Conventional Rules) และกฎภาษาธรรมชาติ (Linguistic Rules) [7]

1.1.2 ต้นไม้การตัดสินใจ (Decision Trees)

ต้นไม้การตัดสินใจเป็นการเรียนรู้เพื่อใช้แบ่งกลุ่มข้อมูลโดยใช้คุณสมบัติของข้อมูลในการจัดแบ่งกลุ่ม สำหรับการสร้างต้นไม้การตัดสินใจนั้นจะเป็นสร้างจากบนลงล่าง โดยเริ่มจากการเลือกตัวแปรข้อมูลเข้า (Input Attributes) ที่มีคุณสมบัติที่สามารถแบ่งกลุ่มได้ดีที่สุดมาสร้างเป็นโหนดราก หลังจากนั้น เมื่อข้อมูลผ่านการแบ่งแยกที่โหนดรากตามค่าตัวแปรข้อมูลเข้าของโหนดรากแล้ว หาตัวแปรข้อมูลเข้าที่ดีที่สุดของข้อมูลที่ผ่านการแบ่งแยกนั้นมาสร้างเป็นโหนดลูกของโหนดรากต่อไป และจะวนสร้างโหนดลูกและต้นไม้ย่อยของแต่ละกิ่งไปเรื่อย ๆ จนกว่าข้อมูลที่ผ่านการแบ่งแยกนั้นจะจัดอยู่ในกลุ่มเดียวกัน หรือจำนวนข้อมูลที่ผ่านการแบ่งแยกในกิ่งหนึ่ง ๆ มีค่าน้อยกว่าค่าที่กำหนดไว้ [18, 19] ตัวอย่างอัลกอริทึมต้นไม้การตัดสินใจ ได้แก่ อัลกอริทึม CART ใช้ค่าสัมประสิทธิ์จีนี (Gini) เพื่อกำหนดคุณสมบัติในการแบ่งข้อมูล อัลกอริทึม ID3 และ C4.5 จะใช้ค่าค่ามาตรฐานเกน (Gain Criteria) เพื่อกำหนดคุณสมบัติในการแบ่งข้อมูล ตัวอย่างต้นไม้การตัดสินใจของฐานข้อมูลอากาศจากวิธีการ J48 ซึ่งเป็นวิธีการที่พัฒนามาจาก C4.5 [20] แสดงดังภาพประกอบ 1.1



ภาพประกอบ 1.1 ตัวอย่างต้นไม้การตัดสินใจของฐานข้อมูลอากาศ

1.1.3 ฟัซซี่เซต (Fuzzy Set)

ฟัซซี่เซตเป็นแนวทางในการแสดงถึงลักษณะความคลุมเครือ หรือความไม่ชัดเจนของข้อมูล [21] ฟัซซี่เซตเป็นเซตที่แสดงถึงความสัมพันธ์ของสมาชิกภายในกลุ่ม โดยความสัมพันธ์นี้จะถูกแสดงในลักษณะของระดับความเป็นสมาชิกที่มีค่าอยู่ในช่วง [0, 1] เทคนิคฟัซซี่เซตนำมาใช้ในการแปลงข้อมูลเดิมที่มีลักษณะเป็นตัวเลขจำนวนจริงให้มีลักษณะเป็นรูปแบบภาษาธรรมชาติ (Linguistic Terms) หรือเรียกว่าเทอมเซต (Term Set) เช่น เล็ก กลาง และใหญ่ ซึ่งมีฟังก์ชันความเป็นสมาชิกที่ใช้ในการประมาณค่าระดับความเป็นสมาชิกของแต่ละเทอมเซต [22] สำหรับการสกัดความรู้โดยใช้ฟัซซี่เซตนั้นได้นำวิธีการอื่น ๆ มาประยุกต์ใช้ร่วมกันด้วย เช่น นำหลักการฟัซซี่เซตสร้างฟังก์ชันความเป็นสมาชิกเพื่อใช้ประมาณค่าระดับความเป็นสมาชิกของแต่ละเทอมเซต หลังจากนั้นใช้หลักการราฟเซตเพื่อทำการสกัดกฎ ซึ่งกฎที่ได้อยู่ในรูปแบบของกฎภาษาธรรมชาติ [23] นอกจากนี้ได้มีการนำฟัซซี่เซตมาประยุกต์ใช้ร่วมกับเทคนิคโครงข่ายประสาทเทียมสำหรับการสกัดกฎความรู้ทำให้ได้กฎที่มีค่าความถูกต้องสูง เช่น อัลกอริทึม Adaptive-Network-based Fuzzy Inference System (ANFIS) ซึ่งเป็นอัลกอริทึมที่ใช้โครงข่ายประสาทเทียม 5 ชั้น ใช้การเรียนรู้แบบแพร่ย้อนกลับ (Backpropagation) เพื่อเรียนรู้เหตุ (Antecedent) จากฟังก์ชันความเป็นสมาชิก (Membership Function) และใช้ค่าเฉลี่ยกำลังสองน้อยที่สุด (Least Mean Square) เพื่อกำหนดค่าสัมประสิทธิ์ในส่วนผลลัพธ์ของกฎที่ได้ [24]

1.1.4 ราฟเซต (Rough Set)

ราฟเซตเป็นทฤษฎีที่ใช้ในการจัดการเกี่ยวกับเรื่องความคลุมเครือและความไม่แน่นอนของข้อมูล หลักการของราฟเซตนั้นใช้วิธีการประมาณค่าจากข้อมูลที่มีอยู่ในสิ่งที่เราสนใจ โดยความสัมพันธ์ของข้อมูลกลุ่มแรกเป็นการประมาณค่าขอบเขตล่าง (Lower

Approximation) และความสัมพันธ์ของข้อมูลกลุ่มที่สองเป็นการประมาณค่าขอบเขตบน (Upper Aprroximation) [25] สำหรับการสกัดกฎความรู้ได้มีการประยุกต์ทฤษฎีราฟเซตร่วมกับวิธีการอื่น ๆ เช่น นำทฤษฎีราฟเซตมาประยุกต์ใช้ร่วมกับวิธีการคำนวณแบบเมทริกซ์เพื่อใช้ในการสกัดกฎความรู้ [26]

1.2 วัตถุประสงค์ของโครงการ

1.2.1 เพื่อสร้างแบบจำลองในการสกัดความรู้โดยใช้แผนที่การจัดกลุ่มเอง

1.2.2 พัฒนาโปรแกรมจากแบบจำลองสำหรับการสกัดความรู้โดยใช้แผนที่การจัดกลุ่มเอง

1.3 ขอบเขตของการดำเนินงาน

1.3.1 พัฒนาอัลกอริทึมและสร้างแบบจำลองในการสกัดความรู้โดยใช้แผนที่การจัดกลุ่มเอง

1.3.2 พัฒนาโปรแกรมเพื่อสกัดความรู้โดยใช้แผนที่การจัดกลุ่มเอง

1.3.3 ข้อมูลที่นำมาสกัดความรู้เป็นข้อมูลที่ได้มาจากข้อมูลทางด้านการแพทย์ คือ กลุ่มฐานข้อมูลที่เป็นมาตรฐานเป็นที่ยอมรับในระดับสากล ซึ่งทำการดาวน์โหลดมาจาก University of California at Irvine (UCI) [27]

1.4 ขั้นตอนและระยะเวลาการดำเนินงาน

1.4.1 ขั้นตอนการดำเนินงาน

1. ศึกษางานวิจัยและเอกสารที่เกี่ยวข้อง
2. ศึกษาเทคนิคการสกัดความรู้โดยใช้แผนที่การจัดกลุ่มเอง และเทคนิคอื่น ๆ ที่ใช้ในงานวิจัย
3. ศึกษาเครื่องมือและซอฟต์แวร์สำหรับทำงานวิจัย
4. วิเคราะห์และออกแบบโปรแกรมการสกัดความรู้โดยใช้แผนที่การจัดกลุ่มเอง
5. พัฒนาโปรแกรมการสกัดความรู้โดยใช้แผนที่การจัดกลุ่มเอง
6. ทดสอบและติดตั้งโปรแกรมการสกัดความรู้โดยใช้แผนที่การจัดกลุ่มเอง
7. จัดทำเอกสารประกอบโปรแกรมการสกัดความรู้โดยใช้แผนที่การจัดกลุ่มเอง
8. เขียนผลงานวิจัยและนำเสนองานวิจัย
9. จัดทำเอกสารวิทยานิพนธ์

1.4.2 ระยะเวลาดำเนินการ

ตุลาคม 2548 – มีนาคม 2550

1.4.3 แผนการดำเนินการวิจัย

ตารางที่ 1.1 แสดงระยะเวลาดำเนินการวิจัย

กิจกรรมขั้นตอนการดำเนินงาน	เดือน																	
	2548			2549												2550		
	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3
1. ศึกษางานวิจัยและเอกสารที่เกี่ยวข้อง	■	■																
2. ศึกษาเทคนิคการสกัดความรู้โดยใช้แผนที่การจัดกลุ่มเอง และเทคนิคอื่น ๆ ที่ใช้ในงานวิจัย		■	■															
3. ศึกษาเครื่องมือและซอฟต์แวร์สำหรับทำงานวิจัย				■														
4. วิเคราะห์และออกแบบโปรแกรมการสกัดความรู้โดยใช้แผนที่การจัดกลุ่มเอง					■	■	■	■										
5. พัฒนาโปรแกรมการสกัดความรู้โดยใช้แผนที่การจัดกลุ่มเอง							■	■										
6. ทดสอบและติดตั้งโปรแกรมการสกัดความรู้โดยใช้แผนที่การจัดกลุ่มเอง								■	■	■								
7. จัดทำเอกสารประกอบโปรแกรมการสกัดความรู้โดยใช้แผนที่การจัดกลุ่มเอง										■	■	■						
8. เขียนผลงานวิจัย											■	■	■	■				
9. จัดทำเอกสารวิทยานิพนธ์และสอบวิทยานิพนธ์														■	■	■	■	■

1.5 สถานที่และเครื่องมือที่ใช้

1.5.1 สถานที่

ห้องวิจัยปัญญาประดิษฐ์ CS207 ภาควิชาวิทยาการคอมพิวเตอร์ ตึกวิทยาการ-คอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่

1.5.2 เครื่องมือที่ใช้

ฮาร์ดแวร์

เครื่องไมโครคอมพิวเตอร์หน่วยความจำ 256 เมกะไบต์ ฮาร์ดดิสค์ 40 กิกกะไบต์หน่วยประมวลผลกลางรุ่นเพนเทียมอาร์โฟร์ 3.00 กิกกะเฮิร์ต จำนวน 1 เครื่อง

ซอฟต์แวร์

1. Microsoft Windows XP เป็นระบบปฏิบัติการ
2. MATLAB 7.0 สำหรับพัฒนาโปรแกรมจากแบบจำลอง และทดสอบแบบจำลอง
3. Java (TM) 2 SDK, Standard Edition สำหรับประมวลผลโปรแกรม Weka
4. Weka สำหรับเปรียบเทียบผลการทดลอง
5. Microsoft Excel 2000 สำหรับเตรียมข้อมูล

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1.6.1 ได้แบบจำลองในการสกัดความรู้โดยใช้แผนที่การจัดกลุ่มเอง

1.6.2 ได้โปรแกรมเพื่อสกัดความรู้โดยใช้แผนที่การจัดกลุ่มเอง