การปรับปรุงตัวแบบการให้คะแนนสินเชื่อโดยใช้การวิเคราะห์การเกาะกลุ่มของตัวทำนายหลากหลาย

นายวิวัช ชลไชยะ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาการคณนา ภาควิชาคณิตศาสตร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2550

IMPROVING CREDIT SCORING MODEL VIA CLUSTER ANALYSIS OF MULTI-PREDICTORS (CLAMP)

Mr. Vivach Chonchaiya

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computational Science

Department of Mathematics
Faculty of Science
Chulalongkorn University

Academic Year 2007

ข้อวิทยานิพนธ์	การปรับปรุงตัวแบบการให้คะแนนสินเชื่อโดยใช้การวิเคราะห์การเกาะกลุ่มของตัวทำนายหลากหลาย
ย	นายวิวัช ชลไชยะ
ขาวิชา	วิทยาการคณนา
จารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร. กรุง สินอภิรมย์สราญ

คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของรศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิทยาศาสตร์
(ศาสตราจารย์ ดร. สุพจน์ หารหนองบัว)

ะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(รองศาสตราจารย์ ดร. พีระพนธ์ โสพัศสถิตย์)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผู้ช่วยศาสตราจารย์ ดร. กรุง สินอภิรมย์สราญ)

..... กรรมการ
(อาจารย์ ดร. สิริพันธุ์ สงวนสินธุกุล)

วิวัช ชลไชยะ : การปรับปรุงตัวแบบการให้คะแนนสินเชื่อ โดยใช้การวิเคราะห์การเกาะกลุ่มของตัวทำนายหลากหลาย. (IMPROVING CREDIT SCORING MODEL VIA CLUSTER ANALYSIS OF MULTI-PREDICTORS (CLAMP)) อ. ที่ปรึกษา : ผศ. ดร. กรุง สินอภิรมย์ ลราญ, 109 หน้า.

ธนาคารใช้การให้คะแนนสินเชื่อ เพื่อจัดลำดับความเสี่ยงของลูกค้าที่ขอสินเชื่อตามศักยภาพของแต่ละคน คะแนนที่ได้ช่วยให้ธนาคารสามารถระบุลูกค้าที่มีความเสี่ยงสูงจากกลุ่มลูกค้าที่ขอสินเชื่อ ตัวแบบการให้คะแนนสินเชื่อได้จากข้อมูลที่สำคัญ ได้แก่ ข้อมูลส่วนตัวของลูกค้า ประวัติการชำระสินเชื่อ และพฤติกรรมของลูกค้า งานวิจัยนี้เสนอตัวแบบการให้คะแนนแบบผสม โดยมีพื้นฐานมาจาก 2 เทคนิคการทำเหมืองข้อมูล คือการวิเคราะห์การเกาะกลุ่มและการจำแนกประเภท ซึ่งจะเรียกวิธีการนี้ว่า การวิเคราะห์การเกาะกลุ่มของตัวทำนายหลากหลาย (CLAMP) วิธีการนี้พัฒนาตัวแบบ 2 ส่วน ส่วนแรกใช้กระบวนการวิเคราะห์การเกาะกลุ่ม ที่ใช้ขั้นตอนวิธีของการวิเคราะห์การเกาะกลุ่มแบบค่าเฉลี่ยเอ็กซ์ กระบวนการนี้แบ่งกั้นข้อมูลออกเป็น k กลุ่ม โดยค่า k ได้จากการพิจารณาค่าเกณฑ์การวัดข้อมูล ส่วนที่ 2 เลือกวิธีการจำแนกประเภทข้อมูลจาก J48 (ตัวแบบต้นไม้การตัดสินใจ) วิธีการจำแนกแบบเบย์อย่างง่าย (ตัวแบบเชิงความน่าจะเป็น) สมการถดถอยแบบโลจิสติก (ตัวแบบเชิงสถิติ) และ ข่ายงานประสาท (ตัวแบบปัญญาประดิษฐ์) เกณฑ์ที่ใช้ในการเลือกตัวแบบจำแนกประเภทแบ่ง 40% เป็นข้อมูลพัฒนาตัวแบบ สำหรับสร้างตัวแบบ 30% เป็นข้อมูลประเมิน สำหรับเลือกวิธีการจำแนกประเภทที่ดีที่สุดในกลุ่ม และ 30% เป็นข้อมูลทดสอบ เพื่อป้องกันปัญหาตัวแบบเหมาะสมเฉพาะข้อมูลพัฒนาตัวแบบ วิธีการจำแนกประเภทที่ใช้ CLAMP แสดงความถูกต้องที่ดีกว่าการใช้วิธีการจำแนกประเภทเพียงอย่างเดียว

ภาควิชาคณิตศาสตร์ ลายมือชื่อนิสิต วิวัช ชลไชยะ
สาขาวิชาวิทยาการคณนา ลายมือชื่ออาจารย์ที่ปรึกษา
ปีการศึกษา 2551

4772481823 : MAJOR COMPUTATIONAL SCIENCE
KEY WORD: CREDIT SCORING / DATA MINING / CLUSTER ANALYSIS / CLASSIFICATION METHOD

VIVACH CHONCHAIYA : IMPROVING CREDIT SCORING MODEL VIA CLUSTER ANALYSIS OF MULTI-PREDICTORS (CLAMP). THESIS ADVISOR : ASST. PROF. KRUNG SINAPIROMSARAN, Ph.D., 109 pp.

Banks use the credit score to rank potential individuals among loan customers. This score helps the banks to determine the high risk customers among loan customers. The scoring model incorporates essential data from personal customer data, credit histories and customer behavior. This research proposes a combined scoring model based on two data mining techniques, clustering analysis and classification, called "Cluster Analysis of multi-predictors (CLAMP)." This combined strategy constructs the model in two phases. The first phase is a clustering process which uses the X-mean clustering algorithm. This process will partition training data into *k* groups where *k* is determined by information measure criteria. The second phase is classifier selection from J48 (decision tree model), Naïve Bayes (probability model), logistic regression (statistical model) and multi-layer perceptron (artificial intelligent model). The criteria for selecting classifier are based on partitioning data into 40% training set for building a model, 30% validation set for selecting the best classifier within a group and 30% test set to reject overfitting model. The classifier using CLAMP shows a better accuracy than the classifier alone.

Department **Mathematics** Student's signature.....
Field of study **Computational Science** Advisor's signature.....
Academic year 2007

กิตติกรรมประกาศ

ผู้วิจัยขอกราบขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร. กรุง สินอภิรมย์สราญ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ท่านได้กรุณาให้ความรู้ คำแนะนำ และคำปรึกษาต่างๆที่ทำให้วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยดี

ขอกราบขอบพระคุณ รองศาสตราจารย์ ดร. พีระพนธ์ โสพัศสถิตย์ ประธานกรรมการ อาจารย์ ดร. สิริพันธุ์ สงวนสินธุกุล กรรมการ ที่ได้ให้คำปรึกษาและช่วยเหลือตลอดระยะเวลาในการทำงานวิจัยนี้ ซึ่งทำให้วิทยานิพนธ์ฉบับนี้มีความสมบูรณ์มากขึ้น

ขอกราบขอบพระคุณ บิดา มารดา ตลอดจนพี่น้องในครอบครัวที่คอยเป็นกำลังใจ และช่วยเหลือผู้วิจัยมาโดยตลอด และขอบคุณเพื่อนๆทุกคนที่เป็นกำลังใจให้

สารบัญ

หน้า

หน้า

สารบัญตาราง

หน้า

หน้า

หน้า

หน้า

หน้า

สารบัญภาพ

หน้า

หน้า

บทที่ 1
บทนำ

ในปัจจุบันธนาคาร และสถาบันทางการเงินที่บริการสินเชื่อ ใช้ข้อมูลที่ลูกค้าที่มายื่นขอรับสินเชื่อและข้อมูลที่ได้จากองค์กรกลางภายนอก มาตัดสินการให้สินเชื่อหรือการปฏิเสธการขอรับสินเชื่อ โดยทางธนาคารหรือสถาบันทางการเงินที่บริการสินเชื่อใช้เจ้าหน้าที่ในการพิจารณาสินเชื่อของลูกค้า แต่เนื่องจากจำนวนลูกค้าที่มาขอรับสินเชื่อมีจำนวนมาก จึงเกิดปัญหาความล่าช้าของกระบวนการพิจารณาสินเชื่อและการอนุมัติที่ไม่เหมาะสม ปัญหาที่เกิดขึ้นส่งผลให้ลูกค้าเกิดความไม่พอใจกับการบริการที่ล่าช้า และการพิจารณาที่ไม่สมเหตุสมผลเป็นเหตุให้ลูกค้าเปลี่ยนไปขอรับสินเชื่อจากธนาคารหรือสถาบันทางการเงินที่บริการสินเชื่ออื่น ดังนั้น ธนาคารและสถาบันทางการเงินที่บริการสินเชื่อจึงต้องการวิธีการพิจารณาสินเชื่อที่รวดเร็ว และมีประสิทธิภาพในการอนุมัติสินเชื่อ

ในปัจจุบันวิธีการหนึ่งที่ได้รับการยอมรับและใช้งานอย่างแพร่หลาย คือ การให้คะแนนสินเชื่อ (credit scoring) [1] ซึ่งเป็นการพัฒนาตัวแบบทางคณิตศาสตร์โดยใช้สถิติ ซึ่งได้จากการศึกษาพฤติกรรมการตัดสินการให้สินเชื่อของเจ้าหน้าที่พิจารณาสินเชื่อในอดีตร่วมกับประวัติการชำระเงินของลูกค้าที่ขอรับสินเชื่อ

ตัวแบบการพิจารณาการให้สินเชื่อที่ได้จากการตัดสินใจของเจ้าหน้าที่พิจารณาสินเชื่อมาจากข้อมูลที่ลูกค้ากรอกในแบบฟอร์มการขอรับสินเชื่อ ข้อมูลประวัติการขอสินเชื่อจากธนาคารแห่งประเทศไทย เป็นต้น จุดประสงค์ของการพัฒนาตัวแบบเพื่อจำลองพฤติกรรมการให้สินเชื่อของเจ้าหน้าที่ที่ประสบผลสำเร็จ กล่าวคือการให้สินเชื่อแก่ลูกค้าที่ส่งยอดชำระเงินทันเวลา กับการไม่ให้สินเชื่อแก่ลูกค้าที่มักเกิดการค้างชำระเป็นประจำ

ข้อมูลที่นำมาพัฒนาตัวแบบ คือ ข้อมูลเฉพาะของลูกค้า เช่น ชื่อ-นามสกุล เพศ อายุ ฯลฯ พฤติกรรมการชำระเงินของลูกค้า รวมถึงปัจจัยอื่นที่ส่งผลต่อความสามารถในการชำระเงินของลูกค้า เช่น สภาพเศรษฐกิจ ภูมิลำเนาของลูกค้า เป็นต้น จุดประสงค์ของการพัฒนาตัวแบบ เพื่อศึกษาพฤติกรรมของลูกค้าที่มีแนวโน้มจะค้างชำระที่ส่งเงินไม่ครบตามที่กำหนด

การพัฒนาตัวแบบการให้คะแนนสินเชื่อมีมาอย่างต่อเนื่องจากอดีตจนถึงปัจจุบัน โดยมีการเพิ่มกระบวนการที่เรียกว่า การทำเหมืองข้อมูล (data mining) [9,10] ร่วมตัดสินการให้สินเชื่อแก่ลูกค้า การทำเหมืองข้อมูล คือ กระบวนการค้นหาองค์ความรู้ที่แฝงอยู่ในข้อมูล โดยใช้วิธีการต่างๆ ซึ่งสามารถจำแนกออกเป็น 2 ลักษณะใหญ่ คือ

1. การเรียนรู้แบบมีเป้าหมาย (supervised learning) เป็นการพัฒนาตัวแบบโดยทราบผลลัพธ์ล่วงหน้าก่อนแล้ว การเรียนรู้แบบมีเป้าหมายใช้สำหรับการพัฒนาตัวแบบเพื่อจำลองพฤติกรรมของสิ่งที่สนใจหรือเพื่อใช้ในการทำนายอนาคต เช่น บริษัทที่ให้บริการโทรศัพท์มือถือต้องการทำนายลูกค้าที่จะเปลี่ยนไปใช้บริการกับบริษัทคู่แข่ง เป็นต้น หรือทำนายข้อมูลที่หายไปจากข้อมูลที่มีอยู่ เช่น การทำนายค่าน้ำฝนที่หายไป เนื่องจากอุปกรณ์ชำรุด เป็นต้น วิธีการที่ใช้ในการทำเหมืองข้อมูล ได้แก่ ข่ายงานประสาท (neural network) ต้นไม้การตัดสินใจ (decision tree) การจำแนกแบบเบย์อย่างง่าย (naive Bayes) สมการถดถอย (regression) เป็นต้น

2. การเรียนรู้แบบไม่มีเป้าหมาย (unsupervised learning) เป็นการพัฒนาตัวแบบโดยไม่ทราบผลลัพธ์ล่วงหน้า ผลลัพธ์ที่ได้เกิดจากการนำสมบัติของข้อมูลมาพิจารณาความสัมพันธ์ เพื่อจัดกลุ่มข้อมูลที่มีความสัมพันธ์ใกล้เคียงกันให้อยู่ร่วมกัน การเรียนรู้แบบไม่มีเป้าหมายนิยมใช้กับการแบ่งกั้นข้อมูล เพื่อให้เห็นลักษณะเฉพาะ โดยมีตัวแทนกลุ่มแสดงลักษณะที่มีความสำคัญออกมา เช่น บริษัทที่ให้บริการโทรศัพท์มือถือใช้การเรียนรู้แบบไม่มีเป้าหมายในการจัดกลุ่มลูกค้า โดยมีจุดประสงค์เพื่อให้บริการกับกลุ่มลูกค้าแต่ละกลุ่มที่แตกต่างกันตามความเหมาะสม โดยวิธีการที่ใช้ คือ การวิเคราะห์การเกาะกลุ่ม (cluster analysis) [15] เช่น การจัดกลุ่มแบบค่าเฉลี่ยเค (K-means clustering) [9,12] และการจัดกลุ่มแบบค่าเฉลี่ยเอ็ก (X-means clustering) [9,16,17] เป็นต้น

ตัวแบบทางคณิตศาสตร์ของการให้คะแนนสินเชื่อใช้วิธีการเรียนรู้แบบมีเป้าหมายในการพัฒนา โดยงานวิจัยนี้เสนอแนวคิดที่เรียกว่า การวิเคราะห์การเกาะกลุ่มของตัวทำนายหลากหลาย (cluster analysis of multi-predictors: CLAMP)

CLAMP เป็นการใช้วิธีการวิเคราะห์การเกาะกลุ่ม ร่วมกับวิธีการจำแนกข้อมูล กระบวนการทำงานคือนำข้อมูลไปจัดกลุ่มโดยวิธีการวิเคราะห์การเกาะกลุ่ม จะได้ข้อมูลที่ถูกแบ่งออกเป็นกลุ่มๆ นำแต่ละกลุ่มข้อมูลที่ได้ไปพัฒนาตัวแบบจำแนกประเภท โดยเลือกตัวแบบจำแนกประเภทที่ดีที่สุด

เหตุผลของการพัฒนาการให้คะแนนสินเชื่อตามแนวคิดของ CLAMP คือ ข้อมูลของลูกค้าเป็นข้อมูลที่มีความหลากหลาย โดยมีความคล้ายกันเป็นกลุ่มๆ ซึ่งสามารถแยกได้จากสมบัติบางประการที่ลูกค้าแสดงออกมา ส่งผลให้เมื่อพิจารณาตัวแบบที่ถูกพัฒนาจากข้อมูลทั้งหมดพร้อมกัน เกิดความผิดพลาดสูง ปัญหานี้สามารถแก้ไขจากการใช้วิธีการวิเคราะห์การเกาะกลุ่มของข้อมูลลูกค้าก่อนทำการจำแนกลูกค้า

อีกสาเหตุหนึ่ง คือ วิธีการจำแนกประเภทแต่ละวิธีมีข้อดีและข้อเสียที่แตกต่างกันไป เมื่อจัดกลุ่มข้อมูลเรียบร้อยแล้ว ลูกค้าแต่ละกลุ่มจะมีสมบัติที่เหมือนกันภายในกลุ่ม และเหมาะสมกับวิธีการพัฒนาตัวแบบที่ต่างกัน

จากเหตุผลทั้ง 2 ประการที่ได้กล่าวมาในข้างต้น ทำให้ผู้วิจัยสนใจศึกษาการวิเคราะห์การเกาะกลุ่มของตัวทำนายหลากหลาย โดยเลือกใช้วิธีการวิเคราะห์การเกาะกลุ่มแบบ X-means เนื่องจากการวิเคราะห์การเกาะกลุ่มแบบ X-means ไม่จำเป็นต้องกำหนดจำนวนกลุ่มที่ชัดเจน เพียงกำหนดจำนวนกลุ่มที่น้อยที่สุด (lower bound) และจำนวนกลุ่มที่มากที่สุด (upper bound) ของการแบ่งกลุ่ม

CLAMP อาจเกิดปัญหาตัวแบบเข้ากับข้อมูลพัฒนาตัวแบบเกินไป (overfitting) งานวิจัยนี้ใช้หลัก train-validate-test เพื่อพัฒนาตัวแบบ โดยแบ่งกลุ่มออกเป็น 3 กลุ่ม คือ ข้อมูลพัฒนาตัวแบบ ข้อมูลประเมิน และข้อมูลทดสอบ

วิธีการจำแนกประเภทลูกค้าที่เลือกใช้เพื่อพัฒนาตัวแบบสำหรับงานวิจัยนี้ คือ ต้นไม้การตัดสินใจ (decision tree) การจำแนกเบย์อย่างง่าย (naive bayes) ข่ายงานประสาท (neural network) สมการถดถอยแบบโลจิสติก (logistic regression) โดยวิธีการข่ายงานประสาทเลือกใช้วิธีการเพอร์เซ็ปตรอนหลายชั้น (multi-layer perceptron) เนื่องจากทั้ง 4 วิธีการเป็นวิธีการที่นิยมใช้ในการพัฒนาตัวแบบการให้คะแนนสินเชื่อ

จากวิธีการวิเคราะห์การเกาะกลุ่ม และวิธีการจำแนกข้อมูลทั้งหมดที่กล่าวมา ผู้วิจัยเลือกใช้คลาสของแต่ละวิธีที่มีอยู่ในโปรแกรมเวกา (WEKA) คือ

- วิธีการ J48 สำหรับต้นไม้การตัดสินใจ
- วิธีการ NaiveBayes สำหรับวิธีการจำแนกแบบเบย์อย่างง่าย
- วิธีการ MultilayerPerceptron สำหรับวิธีการข่ายงานประสาท
- วิธีการ Logistic สำหรับวิธีการสมการถดถอยแบบโลจิสติก

โดยตัววัดที่ใช้สำหรับเลือกวิธีการจำแนกที่ดีที่สุด คือ ความถูกต้องในการทำนาย (accuracy) ของกลุ่มข้อมูลทดสอบ

ในบทที่ 2 อธิบายถึงการให้คะแนนสินเชื่อ และการทำเหมืองข้อมูล โดยการให้คะแนนสินเชื่อจะกล่าวถึงความหมายการให้คะแนนสินเชื่อ และประโยชน์ที่ได้รับจากการใช้การให้คะแนนสินเชื่อ สำหรับการทำเหมืองข้อมูลจะกล่าวถึงความหมายของการทำเหมืองข้อมูล ประโยชน์ที่ได้รับจากการทำเหมืองข้อมูล และ CLAMP โดยจะกล่าวถึงวิธีการวิเคราะห์การเกาะกลุ่มจะพิจารณาวิธีการเกาะกลุ่มแบบค่าเฉลี่ยเค (K-means clustering) และวิธีการเกาะกลุ่มแบบค่าเฉลี่ยเอ็กซ์ (X-means clustering) สำหรับวิธีการจำแนกประเภทในบทนี้จะกล่าวถึงมี 4 วิธี คือ วิธีการต้นไม้การตัดสินใจ (decision tree) วิธีการจำแนกแบบเบย์อย่างง่าย (naive Bayes) วิธีสมการถดถอยแบบโลจิสติก (logistic regression) และวิธีข่ายงานประสาท (neural network)

บทที่ 3 อธิบายถึงผังการทำงานของ CLAMP และ การอธิบายโปรแกรมที่พัฒนา CLAMP

บทที่ 4 อธิบายถึงรายละเอียดของข้อมูลสินเชื่อ และผลการวิเคราะห์ที่ได้จากโปรแกรมที่พัฒนาขึ้น โดยในบทนี้มีการอธิบายผลลัพธ์ที่ได้จากการนำข้อมูล benchmark มาพิจารณา

บทที่ 5 อธิบายถึงผลสรุปจากงานวิจัย และแนวทางของงานวิจัยในอนาคต

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงพื้นฐานของงานวิจัยซึ่งประกอบด้วยความหมายของการให้คะแนนสินเชื่อ ประโยชน์ที่ได้รับจากการใช้การให้คะแนนสินเชื่อ ความหมายของการทำเหมืองข้อมูล ประโยชน์ที่ได้รับจากการทำเหมืองข้อมูล และ CLAMP

2.1. **การให้คะแนนสินเชื่อ** (credit scoring)

2.1.1. **ความหมายของการให้คะแนนสินเชื่อ**

ระบบการให้คะแนนสินเชื่อ [1] เป็นระบบงานที่มีการให้คะแนนกับลูกค้าแต่ละรายที่มาขอสินเชื่อจากสถาบันการเงิน และมีการประเมินผลเพื่อจัดอันดับหรือเกรดของลูกค้า จากลูกค้าคุณภาพดีหรือผู้ที่มีคะแนนสูงมาสู่ลูกค้าที่มีฐานะแย่หรือมีคะแนนต่ำ ทั้งนี้เพื่อช่วยให้การพิจารณาการให้สินเชื่อแก่ลูกค้ารายย่อยของสถาบันการเงินโดยเฉพาะธนาคารพาณิชย์เป็นไปด้วยความราบรื่น โดยระบบจะอาศัยฐานข้อมูลของลูกค้าบนบรรทัดฐานเดียวกัน ซึ่งจะทำให้การพิจารณาประเมินความเสี่ยงของลูกค้าและการปล่อยสินเชื่อให้กับลูกค้าของธนาคารเป็นไปในแนวทางเดียวกัน

2.1.2. **ประโยชน์ที่ได้รับจากใช้การให้คะแนนสินเชื่อ**

ประโยชน์ที่ได้รับจากการนำเทคนิคการให้คะแนนสินเชื่อมาใช้ [6] คือ

1. ช่วยลดเวลาที่ใช้พิจารณาอนุมัติสินเชื่อแก่ลูกค้าที่เข้าข่ายเป็นลูกค้าชั้นดี ซึ่งเป็นขั้นตอนต่อจากขั้นตอนการประเมินลูกค้าเบื้องต้นจากแบบฟอร์มใบคำขอสินเชื่อ เพราะแบบฟอร์มใบคำขอสินเชื่อของลูกค้าที่มีคะแนนรวมสูงสุด(ความเสี่ยงต่ำสุด) ถือได้ว่าเป็นลูกค้าคุณภาพดีเยี่ยมที่เข้าข่ายตามเงื่อนไขหรือหลักเกณฑ์ที่สถาบันการเงินนั้นได้กำหนดไว้ก่อนการประเมินผล ในกรณีนี้อาจจะทำการอนุมัติสินเชื่อโดยอัตโนมัติ ส่วนลูกค้าสินเชื่อที่มีคะแนนรวมรองลงมาจะต้องผ่านขั้นตอนการวิเคราะห์สินเชื่อและการอำนวยสินเชื่อก่อนที่จะนำเสนอสู่การตัดสินใจยอมรับหรือปฏิเสธลูกค้าสินเชื่อรายนั้น ต่อไป

2. ช่วยลดความสูญเสียจากภาระหนี้เสียที่อาจเกิดขึ้น ระบบการให้คะแนนสินเชื่อจะช่วยในการประเมินจัดอันดับคุณภาพของลูกค้าสินเชื่อ ทำให้

สถาบันการเงินซึ่งเป็นผู้ให้สินเชื่อทราบถึงความเสี่ยงของผู้ขอสินเชื่อ ส่งผลให้สถาบันการเงินประมาณการณ์รายได้หรือรายจ่ายที่ได้จากการกู้ยืมของผู้ขอสินเชื่อได้ทั้งหมด และ สามารถประเมินความเสี่ยงของผู้ขอสินเชื่อแต่ละรายได้ ขณะเดียวกันก็มีส่วนช่วยลดความลำเอียงในทางที่ให้ความช่วยเหลือเป็นพิเศษแก่ลูกค้าสินเชื่อหรือประเมินความเสี่ยงของลูกค้าที่ต่ำกว่าความเป็นจริงของเจ้าหน้าที่สินเชื่อด้วย ทำให้เจ้าหน้าที่สินเชื่อสามารถสร้างมูลค่าเพิ่ม (value-added) แก่องค์กรได้สูงสุด ระบบการให้คะแนนสินเชื่อจึงช่วยป้องกันความสูญเสียจากภาระหนี้เสียที่อาจจะเกิดขึ้นได้

3. เสริมการประมวลผลข้อมูลด้าน consumer credit แก่เครดิตบูโร (credit bureau) หรือบริษัทศูนย์ข้อมูลเครดิตแห่งประเทศไทย เพราะเครดิตบูโรไม่ได้มีหน้าที่อนุมัติ (approved) หรือปฏิเสธการให้เครดิตแก่ลูกค้า (rejected) แต่มีหน้าที่เพียงรายงานข้อมูลที่ได้จากสถาบันการเงินที่เป็นสมาชิกเท่านั้น การเสริมข้อมูลแก่เครดิตบูโรจะช่วยลดความเสี่ยงให้กับระบบสถาบันการเงินได้มากขึ้น และในระยะยาวจะเป็นการสร้างวัฒนธรรมให้ลูกค้าชำระหนี้ได้ตามกำหนดเวลา เพราะลูกค้าที่มีประวัติการเงินไม่ดีจะไม่สามารถขอกู้เงินจากสถาบันการเงินอื่นได้ ขณะที่ลูกหนี้ที่มีประวัติการชำระหนี้ที่ดีก็จะได้รับการตอบแทนทั้งในแง่ของการพิจารณาสินเชื่อที่สะดวกรวดเร็ว และอัตราดอกเบี้ยที่ต่ำ ทำให้สถาบันการเงินมีฐานข้อมูลของลูกค้าที่ดีและกว้างขึ้นด้วย

4. สามารถนำไปช่วยงานการวางแผนดำเนินการเพื่อสำรองหนี้สูญ และเป็นเครื่องมือที่ช่วยชี้คุณภาพของหนี้ที่มีในกลุ่มว่าน่าจะมีหนี้ และหนี้จัดชั้นซึ่งทำได้โดยการนำคะแนนรวมที่ได้จากการประเมินลูกค้าสินเชื่อทุกรายมาวิเคราะห์โดยการแยกกลุ่มลูกค้าหรือหาสัดส่วนของลูกค้าที่ธนาคารยอมให้สินเชื่อกับปฏิเสธการให้สินเชื่อ ระบบนี้จึงช่วยปรับปรุงระบบการควบคุมการบริหารงานตรวจสอบและติดตาม (monitoring) ให้กับผู้บริหารงานสินเชื่อ

2.2. **การทำเหมืองข้อมูล** (data mining)

2.2.1. **ความหมายของการทำเหมืองข้อมูล**

การทำเหมืองข้อมูล [10] หรือ การสืบค้นความรู้ในฐานข้อมูล เป็นกระบวนการ หรือ วิธีสกัดหรือสืบค้นความรู้จากฐานข้อมูลขนาดใหญ่อัตโนมัติ โดยความรู้ที่ได้อาจอยู่ในรูปแบบของความสัมพันธ์ ข้อสรุปแบบจำลอง หรือลักษณะเฉพาะของข้อมูลในฐานข้อมูล การทำเหมืองข้อมูลต้องมีกำหนดวัตถุประสงค์ หรือเป้าหมายในการทำเหมืองข้อมูล เพื่อให้สามารถกำหนดขั้นตอนและการวัดผลลัพธ์ที่ได้อย่างชัดเจน

2.2.2. **ประโยชน์ที่ได้รับจากการทำเหมืองข้อมูล**

การทำเหมืองข้อมูลเป็นศาสตร์ที่ได้รับความสนใจ และได้รับการยอมรับอย่างกว้างขวาง เช่น การทำเหมืองข้อมูลเพื่อค้นหาการซื้อสินค้าของลูกค้า การวางแผนการลงทุน การตรวจหาการลงทุนที่ผิดปรกติ เป็นต้น นอกจากนี้ยังสามารถนำไปประยุกต์กับการปรับปรุงแผนการโฆษณาสินค้า การวางแผนการตลาดเฉพาะกลุ่ม เพื่อเพิ่มความพึงพอใจให้แก่ลูกค้าในกลุ่ม

2.2.3. **วิธีการทำเหมืองข้อมูลสำหรับทดสอบการวิเคราะห์การเกาะกลุ่มของตัวทำนายหลากหลาย**

การวิเคราะห์การเกาะกลุ่มของตัวทำนายหลากหลาย (CLAMP) เป็นการใช้หลักการทำเหมืองข้อมูล 2 แบบ คือ การวิเคราะห์การเกาะกลุ่ม โดยจัดกลุ่มข้อมูลที่มีสมบัติใกล้เคียงกันไว้ในกลุ่มเดียวกัน และการพัฒนาตัวแบบสำหรับการทำนายผล

CLAMP แบ่งออกเป็น 2 ส่วนดังนี้

2.2.3.1. **การวิเคราะห์การเกาะกลุ่ม** (cluster analysis)

วิธีการวิเคราะห์การเกาะกลุ่ม [9,12,15] ที่สนใจใช้การวิเคราะห์การเกาะกลุ่มแบบค่าเฉลี่ยเอ็กซ์ (X-means clustering) [9,16,17] ซึ่งเป็นการขยายวิธีการวิเคราะห์การเกาะกลุ่มแบบค่าเฉลี่ยเค (K-means clustering) [9,12] ให้สามารถแบ่งกลุ่มได้อย่างอัตโนมัติ โดยไม่ต้องกำหนดจำนวนกลุ่มเหมือน วิธีการวิเคราะห์การเกาะกลุ่มแบบค่าเฉลี่ยเค เพียงแต่กำหนดช่วงของจำนวนกลุ่ม คือ จำนวนกลุ่มที่น้อยที่สุด มักกำหนดเป็น 2 เพราะว่า 2 เป็นจำนวนกลุ่มที่น้อยที่สุดที่สามารถแสดงให้เห็นการแบ่งกลุ่มได้ และจำนวนกลุ่มที่มากที่สุด เพื่อทำให้กลุ่มที่ได้จาก

การแบ่งกลุ่มไม่มีจำนวนมากเกินไป ในการทดลองจะมีการกำหนดช่วงของจำนวนกลุ่ม คือ จำนวนกลุ่มอย่างน้อย 2 กลุ่ม อย่างมากไม่เกิน 10 กลุ่ม ($2 \leq k \leq 10$)

- **วิธีการวิเคราะห์การเกาะกลุ่มแบบค่าเฉลี่ยเค** (K-means clustering) เป็นการกำหนดจุดสำหรับเป็นตัวแทนกลุ่มของข้อมูล (centroid) โดยตัวแทนกลุ่ม คือ ค่าเฉลี่ยของทุกจุดที่อยู่ในกลุ่มเดียวกัน หรือกล่าวได้ว่าตัวแทนกลุ่ม คือ จุดศูนย์กลางของกลุ่มข้อมูลนั่นเอง

  ข้อดีของการวิเคราะห์การเกาะกลุ่มแบบค่าเฉลี่ยเค คือ การคำนวณที่ง่ายทำให้ขั้นตอนวิธีมีความเร็วสูงถึงแม้จะใช้กับข้อมูลขนาดใหญ่ แต่การวิเคราะห์การเกาะกลุ่มยังมีข้อเสีย คือ วิธีการนี้ต้องมีการกำหนดจำนวนกลุ่มการแบ่งกั้น แล้วกำหนดจุดเริ่มต้น ถ้าจุดเริ่มต้นต่างกัน ผลลัพธ์ที่ได้อาจไม่เหมือนกัน วิธีการวิเคราะห์การเกาะกลุ่มแบบค่าเฉลี่ยเคเป็นการพิจารณาค่าความแปรปรวนที่ต่ำที่สุดภายในกลุ่ม แต่ไม่ได้พิจารณาค่าความแปรปรวนรวมของข้อมูล วิธีการวิเคราะห์การเกาะกลุ่มแบบค่าเฉลี่ยเคไม่เหมาะกับข้อมูลเกาะกลุ่มแบบมีส่วนเว้า ส่วนโค้ง และถ้าใช้กับตัวแปรไม่ต่อเนื่อง เช่น เพศ, ระดับการศึกษา ที่เปลี่ยนเป็นข้อมูลจำนวนแล้ว ค่ากลางที่ได้อาจจะไม่มีความหมายจริงตามค่าของตัวแปรไม่ต่อเนื่อง

- **วิธีการวิเคราะห์การเกาะกลุ่มแบบค่าเฉลี่ยเอ็กซ์** (X-means clustering) เป็นการเพิ่มความสามารถของ การวิเคราะห์การเกาะกลุ่มแบบค่าเฉลี่ยเค โดยเพิ่มการพิจารณาการแยกกลุ่ม โดยอาศัยหลักการของเกณฑ์การวัดข้อมูลแบบเบย์เซียน (Bayesian information criterion : BIC) [16] ซึ่งการวิเคราะห์การเกาะกลุ่มแบบค่าเฉลี่ยเค ต้องกำหนดจำนวนข้อมูลที่ชัดเจน แต่สำหรับการวิเคราะห์การเกาะกลุ่มแบบค่าเฉลี่ยเอ็กซ์ จำนวนกลุ่มของข้อมูลจะมีขอบเขต โดยขอบเขตล่างคือจำนวนกลุ่มน้อยสุดที่จะใช้แบ่งข้อมูล และขอบเขตบนคือจำนวนกลุ่มมากสุดที่จะแบ่งข้อมูล ขั้นตอนการ

แบ่งกลุ่มของวิธีการวิเคราะห์การเกาะกลุ่มแบบค่าเฉลี่ยเอ็กซ์เป็นดังนี้

- ขั้นตอนที่ 1 นำข้อมูลมาแบ่งกลุ่มตามค่าขอบเขตล่าง (จำนวนกลุ่มน้อยสุด) โดยให้ k เท่ากับขอบเขตล่าง
- ขั้นตอนที่ 2 พิจารณาแบ่งกลุ่มข้อมูลของแต่ละตัวแทนกลุ่มออกเป็น k ส่วน โดยใช้การวิเคราะห์การเกาะกลุ่มแบบค่าเฉลี่ยเค
- ขั้นตอนที่ 3 พิจารณา ค่าเกณฑ์การวัดข้อมูลแบบเบย์เซียนของแต่ละกลุ่มข้อมูลแล้วนำมาพิจารณาการแยกกลุ่มต่อ หากสามารถแยกกลุ่มได้ ให้ทำขั้นตอนที่ 2 และ ขั้นตอนที่ 3 อีกครั้งจนกระทั่งไม่สามารถแยกกลุ่มต่อได้ หรือแยกกลุ่มจนกระทั่งเท่ากับจำนวนกลุ่มที่มากที่สุดที่กำหนดไว้

- **เกณฑ์การวัดข้อมูลแบบเบย์เซียน** (Bayesian information criterion : BIC) เป็นเกณฑ์ที่ใช้ทางสถิติสำหรับเลือกตัวแบบ เกณฑ์การวัดข้อมูลแบบเบย์เซียนบางครั้งอาจเรียกว่า เกณฑ์การวัดแบบชวาร์ส (Schwarz criterion, หรือ Schwarz information criterion : SIC) เพราะว่า Gideon E. Schwarz (1978) เป็นผู้พัฒนาขึ้นมา

การคำนวณค่าเกณฑ์การวัดข้อมูลแบบเบย์เซียน

การคำนวณค่าเกณฑ์การวัดข้อมูลแบบเบย์เซียน เป็นการคำนวนการแจกแจงความน่าจะเป็นอย่างมีเงื่อนไข (Posterior probabilities : $Pr[M|D]$)เพื่อหาคะแนนของแต่ละกลุ่ม ค่าเกณฑ์การวัดข้อมูลแบบเบย์เซียน สามารถประมาณค่าได้จากสูตรของ Kass และ Wasserman ต่อไปนี้

$$BIC = \hat{l} - \frac{P}{2} \times \log(R)$$

โดย $\hat{l}$ คือ ความเป็นไปได้ที่จะอยู่ในกลุ่มเหมาะที่สุด โดย P คือ จำนวนตัวแปร และ R คือ จำนวนสมาชิกในกลุ่มของข้อมูล

2.2.3.2. **วิธีการการจำแนกประเภท** (classification method)

วิธีการจัดจำแนกประเภทที่การวิเคราะห์การเกาะกลุ่มของตัวทำนายหลากหลายรองรับมี 4 วิธีการดังนี้

- **การจำแนกแบบเบย์อย่างง่าย** (naive bayes) เป็นเทคนิคที่ถูกตั้งชื่อตาม Thomas Bayes (1702-1761) โดยใช้หลักความน่าจะเป็น ซึ่งอยู่บนพื้นฐานของ ทฤษฎีของเบย์ (Bayes' theorem) และสมมติฐานที่กำหนดการเกิดของเหตุการณ์ต่างๆ ที่ใช้ในการจัดกลุ่มนั้นเป็นอิสระต่อกัน (independence)

  การจำแนกแบบเบย์อย่างง่ายจะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอิสระแต่ละตัวกับตัวแปรตาม เพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็น ในทางทฤษฎีแล้วการทำนายผลของการจำแนกแบบเบย์อย่างง่ายจะถูกต้อง ถ้าตัวแปรอิสระทั้งหมดเป็นอิสระต่อกัน ไม่ขึ้นกับตัวแปรอิสระตัวใดตัวหนึ่ง การจำแนกแบบเบย์อย่างง่ายยังไม่รองรับข้อมูลที่เป็นข้อมูลต่อเนื่อง (continuous data) ดังนั้นตัวแปรอิสระ หรือตัวแปรตามที่เป็นค่าต่อเนื่องจะต้องถูกแบ่งเป็นช่วง การจำแนกแบบเบย์อย่างง่ายสามารถให้ผลลัพธ์ที่ดี รวดเร็ว และเป็นเครื่องมือที่ดีในการพัฒนาตัวแบบ

- **ต้นไม้การตัดสินใจ** (decision tree) เป็นเทคนิคที่ให้ผลลัพธ์ในลักษณะของโครงสร้างต้นไม้

  วิธีต้นไม้การตัดสินใจประกอบด้วยโหนด สำหรับเงื่อนไขในการตัดสินใจ กิ่งคือผลลัพธ์ที่ได้จากการพิจารณาเงื่อนไขที่โหนด โดยแต่ละกิ่งจะนำไปสู่ผลลัพธ์สุดท้ายคือใบ (leaf node หรือ decision node)

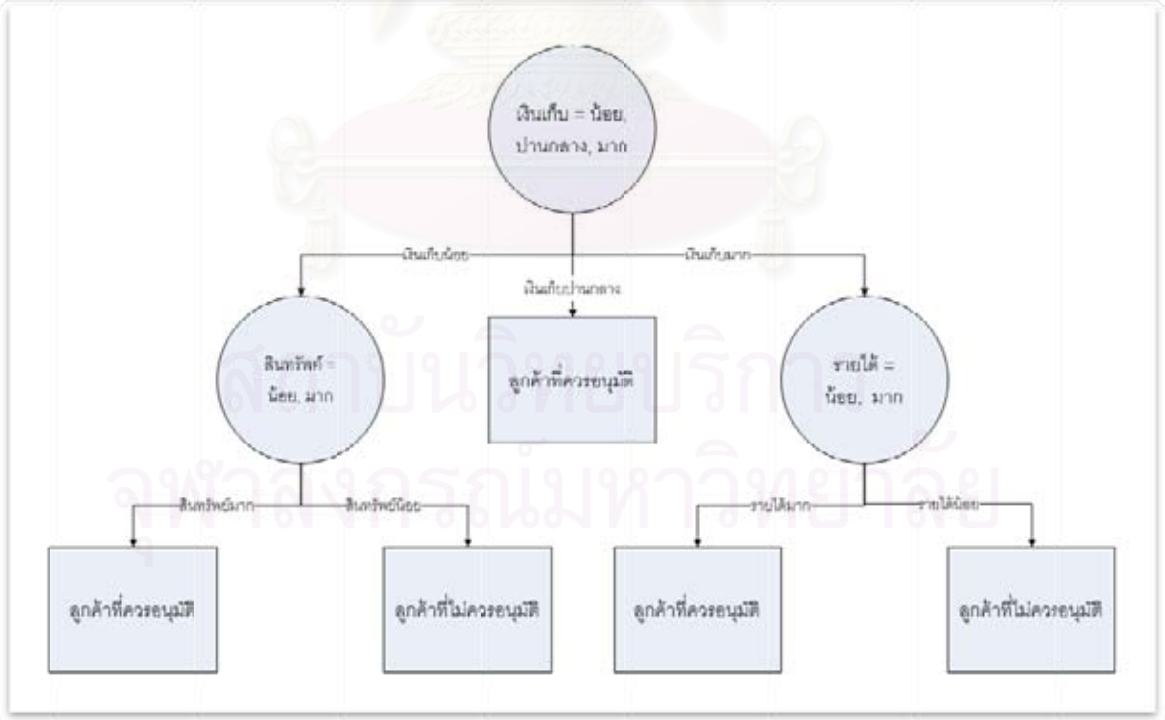
  วิธีการต้นไม้การตัดสินใจจะจำกัดข้อมูลที่เป็นตัวแปรตาม (dependent variable) หนึ่งตัวต่อหนึ่งต้น ถ้าต้องการทำนายตัวแปรตามหลายตัว ผู้วิเคราะห์ต้องสร้างตัวแบบสำหรับตัวแปรตามแต่ละตัว ขั้นตอนวิธีของวิธีการต้นไม้การตัดสินใจส่วนใหญ่ไม่รองรับข้อมูล

แบบต่อเนื่อง (continuous data) จะต้องมีการแบ่งให้เป็นข้อมูลแบบไม่ต่อเนื่อง (discretized data) เสียก่อน จึงใช้ขั้นตอนวิธีได้

วิธีการต้นไม้การตัดสินใจเป็นวิธีการจำแนกที่ค่อนข้างแพร่หลาย เนื่องจากความไม่ซับซ้อนของขั้นตอนการทำงาน และสามารถตีความและเข้าใจลักษณะของรูปแบบข้อมูลได้ง่าย เพราะมีการแยกออกเป็นกฎ หรือข้อกำหนดได้

จากรูปที่ 3.3 เป็นตัวอย่างของต้นไม้การตัดสินใจ ที่ได้จากการพิจารณาความเสี่ยงของสินเชื่อ (credit risk) จำแนกข้อมูลลูกค้าออกเป็น 2 กลุ่มคือ ลูกค้าที่ควรอนุมัติ และลูกค้าที่ไม่ควรอนุมัติ โดยพิจารณาตัวแปรในข้อมูลดังนี้

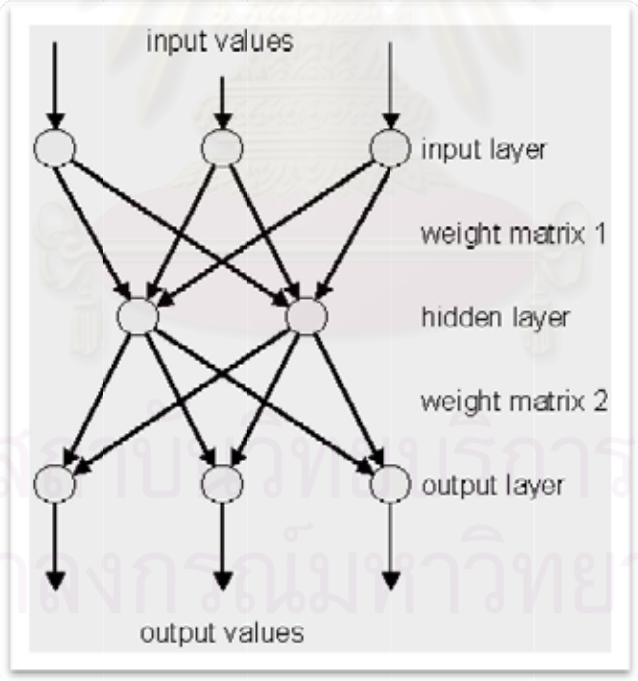
- ยอดในบัญชีเงินฝากของลูกค้า (savings)
- ปริมาณสินทรัพย์ของลูกค้า (assets)
- รายได้ของลูกค้า (income)



รูปที่ 2.1 : แสดงตัวอย่างของผลลัพธ์ที่ได้จากวิธีการจำแนกประเภทแบบต้นไม้การตัดสินใจ

- **ข่ายงานประสาท** (neural networks) เป็นแบบจำลองระบบการประมวลผลที่คล้ายกับข่ายงานเส้นใยประสาทของสมองมนุษย์ที่เชื่อมโยงกัน ที่เรียกว่า เส้นประสาท (neurous) ทำให้เกิดความสามารถ “เรียนรู้” จากข้อมูลที่ได้ผ่านการประมวลผล ในข่ายงานประสาทจะมีการเริ่มต้นใช้งานในส่วนของการปฏิบัติการเฉพาะ เช่น การประมวลผลรูปภาพ และการรับรู้ด้วยเสียง

ในโครงสร้างของข่ายงานประสาท จะประกอบด้วยโหนดสำหรับการรับข้อมูลเข้า (input) ส่งข้อมูลออก (output) และการประมวลผล กระจายอยู่ในโครงสร้างเป็นชั้น ๆ ได้แก่ ชั้นของการรับข้อมูลเข้า (input layer) ชั้นของการส่งข้อมูลออก (output layer) และ ชั้นซ่อน (hidden layers) การประมวลผลของข่ายงานประสาทจะอาศัยการส่งการทำงานผ่านโหนดต่าง ๆ ในชั้นเหล่านี้ แต่ละโหนดต่อกันโดยใช้การเชื่อมต่อ (connection link) ตามค่าถ่วงน้ำหนัก



รูปที่ 2.2 : แสดงระดับชั้นการทำงานของข่ายงานประสาทแบบเพอร์เซพตรอนหลายชั้น

วิธีการจำแนกแบบข่ายงานประสาทถูกใช้ในการแก้ปัญหาอย่างกว้างขวาง โดยการจำแนกแบบข่ายงานประสาทมีสมบัติที่ไวต่อ

รูปแบบของข้อมูลนำเข้า ถ้าเราแทนข้อมูลด้วยรูปแบบที่แตกต่างกัน ก็จะสามารถผลิตผลลัพธ์ที่แตกต่างกันออกมา

ถึงแม้ว่า วิธีการจำแนกแบบข่ายงานประสาทสามารถนำไปประยุกต์ใช้กับงานหลายๆ ชนิดได้อย่างมีประสิทธิภาพ แต่ผลลัพธ์ที่ได้ก็ยากต่อการทำความเข้าใจ เนื่องจากวิธีการจำแนกแบบข่ายงานประสาทเป็นวิธีการที่ค่อนข้างซับซ้อนกว่าวิธีการจำแนกแบบอื่น และการไม่สามารถอธิบายความสัมพันธ์ของค่าถ่วงน้ำหนัก กับผลลัพธ์ที่ได้ อย่างไรก็ตามวิธีนี้ก็มีข้อดีที่สำคัญ คือ วิธีการจำแนกนี้ไม่มีข้อจำกัดเกี่ยวกับชนิดของความสัมพันธ์ นอกจากนั้นวิธีการจำแนกแบบข่ายงานประสาทยังไม่มีปัญหา กับความสัมพันธ์ที่เป็นแบบตรีโกณมิติ (trigonometric) หรือ ลอการิทึม (logarithm) ด้วย

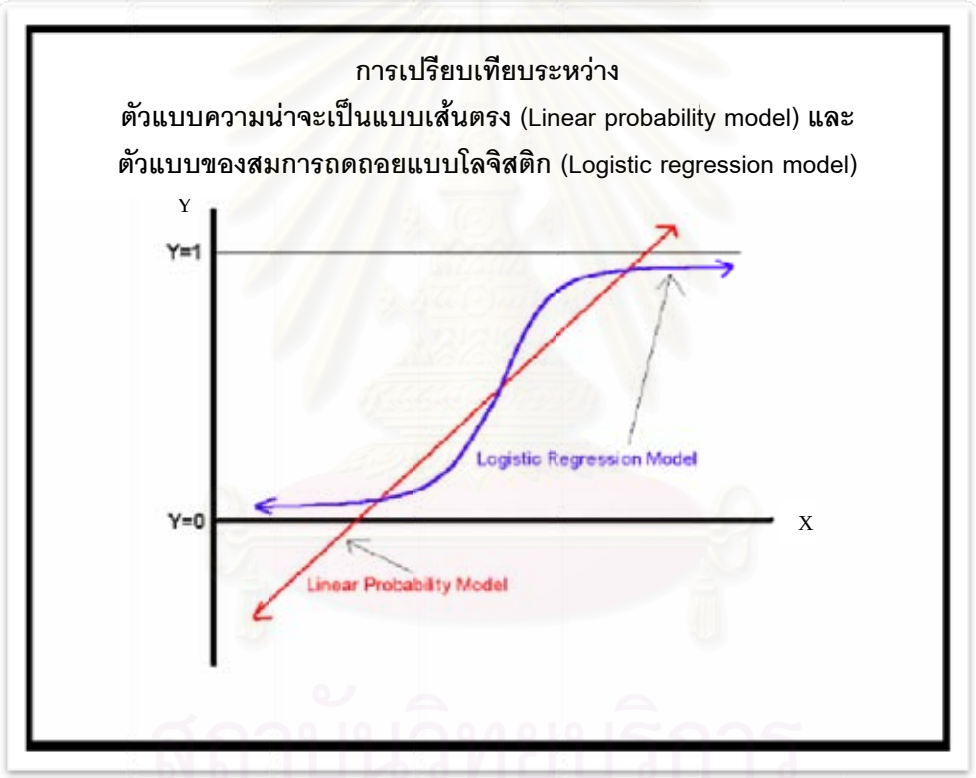
- **สมการถดถอยแบบโลจิสติก (logistic regression)** คือ สมการถดถอย เมื่อตัวแปรตามอยู่ในรูปของตัวแปรสองค่า (dichotomy) มีค่าของตัวแปรเป็น 0 หรือ 1 โดยผลลัพธ์ที่ได้จะอยู่ในรูปของความน่าจะเป็นจึงใช้วิเคราะห์ข้อมูลตอบคำถามเกี่ยวกับโอกาส หรือความน่าจะเป็นที่กลุ่มประชากรมีลักษณะอย่างใดอย่างหนึ่งตามเส้นโค้ง และเนื่องจากรูปแบบความสัมพันธ์แบบเส้นโค้งนี้เรียกว่า เส้นโค้งโลจิสติก วิธีการวิเคราะห์แบบนี้จึงมีชื่อเรียกว่า สมการถดถอยแบบโลจิสติก

สมการถดถอยแบบโลจิสติกมีวัตถุประสงค์ เพื่อศึกษาความสัมพันธ์ระหว่างตัวแปรตาม และตัวแปรอิสระ จากนั้นนำสมการถดถอยที่ได้ไปประมาณ หรือพยากรณ์ค่าตัวแปรตาม เมื่อกำหนดค่าตัวแปรอิสระ เราอาจเรียกเทคนิคนี้ว่า สมการถดถอยแบบโลจิสติกแบบทวินาม เนื่องจากมีความเกี่ยวข้องกับทฤษฎีความน่าจะเป็นแบบทวินาม

การพัฒนาตัวแบบสมการถดถอยแบบโลจิสติก โดยอาศัยขั้นตอนวิธีของความน่าจะเป็นของการเกิดเหตุการณ์ ซึ่งเรียกขั้นตอน

วิธีที่พัฒนาขึ้นมาว่าการพิจารณาลอการิทึมของโอกาสที่เป็นไปได้ (logit : log of the odds)

สมการถดถอยแบบโลจิสติก เป็นวิธีการที่มีประสิทธิภาพ แต่มีข้อจำกัดเกี่ยวกับความอิสระต่อกันของตัวแปรตาม (dependent variable) เนื่องจากตัวแปรตามกับตัวแปรอิสระ อาจไม่เป็นอิสระต่อกัน นอกจากนั้นผู้วิเคราะห์ตัวแบบต้องอาศัยประสบการณ์ของตนในการวิเคราะห์ และการเลือกตัวแปรที่จะนำมาวิเคราะห์อย่างเหมาะสมจึงจะทำให้ได้ผลลัพธ์ที่มีประสิทธิภาพ



รูปที่ 2.3 : แสดงการเปรียบเทียบระหว่าง
ตัวแบบความน่าจะเป็นแบบเส้นตรง (linear probability model) และ
ตัวแบบของสมการถดถอยแบบโลจิสติก (logistic regression model)

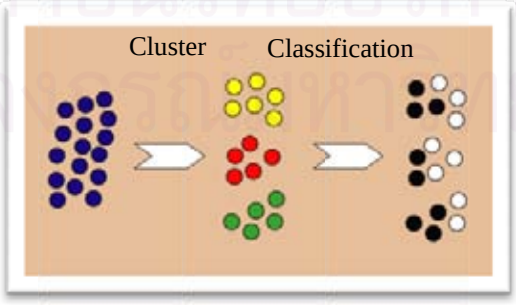
การทำงานของวิธีการทำเหมืองข้อมูลที่กล่าวมาทั้งหมดจะกล่าวในบทที่ 3

บทที่ 3
การวิเคราะห์การเกาะกลุ่มของตัวทำนายหลากหลาย
(Cluster analysis of multi-predictors : CLAMP)

การวิเคราะห์การเกาะกลุ่มของตัวทำนายหลากหลาย (cluster analysis of multi-predictors : CLAMP) เป็นกระบวนการที่ใช้ขั้นตอนวิธีจากการวิเคราะห์การเกาะกลุ่ม ซึ่งเป็นการเรียนรู้แบบไม่มีเป้าหมาย มาใช้ร่วมกับการจำแนกประเภท (classifications) ซึ่งเป็นการเรียนรู้แบบมีเป้าหมาย โดยวิธีการจำแนกกลุ่มมีหลายวิธี เช่น วิธีการจำแนกแบบเบย์อย่างง่าย (naive Bayes) ต้นไม้การตัดสินใจ (decision tree) ข่ายงานประสาท (neural network) สมการถดถอยแบบโลจิสติก (logistic regression)

เหตุผลของการสร้าง CLAMP คือ ข้อมูลที่นำมาใช้ในการพัฒนาตัวแบบมีความหลากหลาย ซึ่งภายในข้อมูลชุดเดียวอาจจะพบว่าสามารถจัดกลุ่มของข้อมูลที่มีค่าของลักษณะประจำใกล้เคียงกันได้หลายกลุ่ม ทำให้การใช้ตัวแบบจำแนกประเภทเพียงชนิดเดียวอาจไม่เหมาะสม เนื่องจากตัวแบบเกิดการเบี่ยงเบนของผลลัพธ์กับข้อมูลของกลุ่มลักษณะประจำกลุ่มอื่น ทำให้ความถูกต้องลดน้อยลง

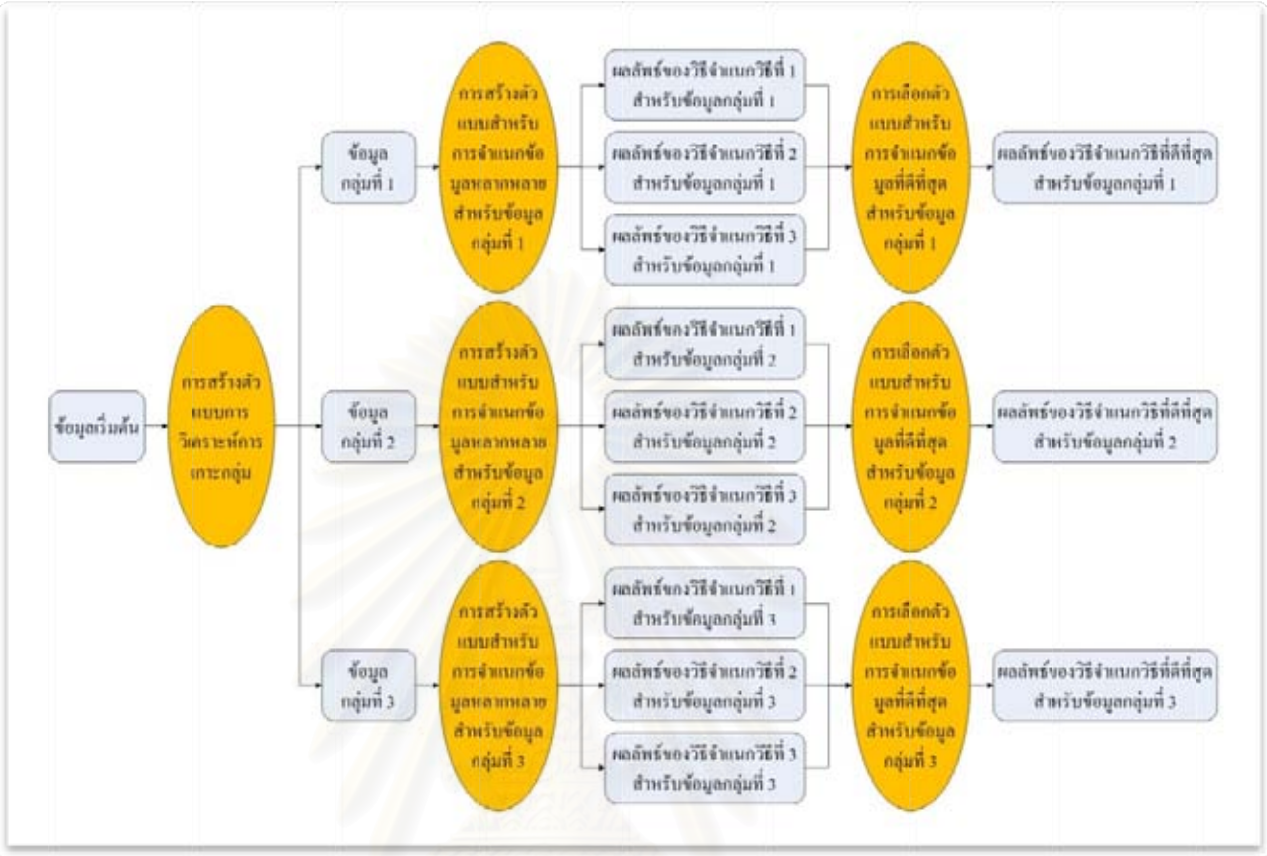
แนวคิดของ CLAMP คือ การนำข้อมูลมาแยกกลุ่ม ก่อนที่จะนำไปพิจารณาพัฒนาตัวแบบวิธีการจำแนกประเภทตามกลุ่มที่ได้จากการวิเคราะห์การเกาะกลุ่ม โดยวิธีการจำแนกของข้อมูลแต่ละกลุ่มนั้น ไม่จำเป็นต้องเป็นตัวแบบเดียวกันทั้งหมด เพราะข้อมูลแต่ละกลุ่มอาจจะเหมาะสมกับวิธีการจำแนกประเภทที่แตกต่างกันออกไป ดังนั้นหลังจากการวิเคราะห์การเกาะกลุ่มแล้ว จะมีการพัฒนาตัวแบบจำแนกประเภท โดยใช้ตัววัด สำหรับงานวิจัยนี้ใช้ความแม่นยำในการทำนายของตัวแบบ: accuracy of model เพื่อหาตัวแบบจำแนกประเภทที่เหมาะที่สุด



รูปที่ 3.1 : แสดงขั้นตอนการทำงานของ CLAMP

3.1. ขั้นตอนการสร้างตัวแบบโดยใช้การวิเคราะห์การเกาะกลุ่มของตัวทำนายหลากหลาย

การพัฒนาตัวแบบโดยใช้ CLAMP นั้นมีขั้นตอนการพิจารณาอยู่ 2 ขั้นตอนหลัก คือ



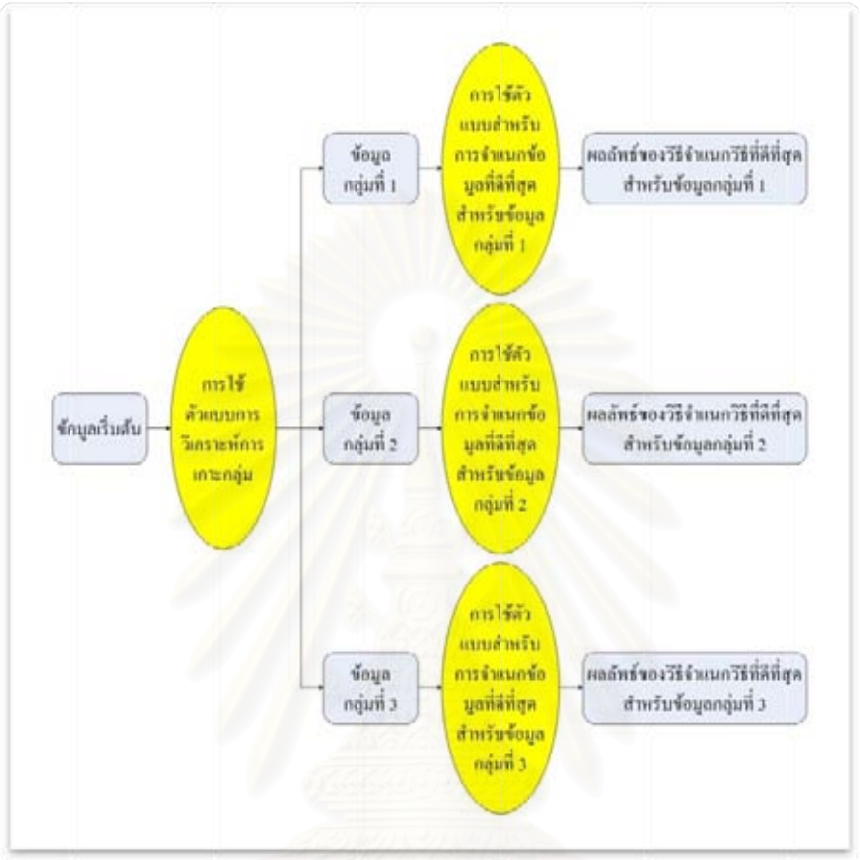
รูปที่ 3.2 : แสดงขั้นตอนการสร้างตัวแบบของ CLAMP

1. ขั้นตอนการวิเคราะห์การเกาะกลุ่มเป็นขั้นตอนการสร้างตัวแบบการวิเคราะห์การเกาะกลุ่มก่อนนำไปสร้างตัวแบบจำแนกประเภทต่อไป

2. ขั้นตอนการสร้างแบบจำแนกประเภท เป็นขั้นตอนการสร้างตัวแบบจำแนกประเภทตามวิธีที่เลือกไว้ โดยการพิจารณาตัวแบบจำแนกประเภทที่เหมาะสมกับข้อมูลแต่ละกลุ่ม คือ การพิจารณาผลลัพธ์จากตัววัดที่ได้จากการใช้ตัวแบบจำแนกประเภทแต่ละวิธีกับข้อมูลแต่ละกลุ่มที่แบ่งกั้น จากขั้นตอนการวิเคราะห์ข้อมูล โดยเลือกตัวแบบที่ให้ค่าของตัววัดดีที่สุด การพิจารณาตัวแบบจำแนกประเภทที่เหมาะสมวิธีนี้อาจเกิดปัญหาตัวแบบเข้ากับข้อมูลพัฒนาตัวแบบเกินไป (overfitting) จึงใช้หลัก train-validate-test โดยแบ่งข้อมูลออกเป็น 3 ส่วน คือ ข้อมูลพัฒนาตัวแบบ ข้อมูลประเมิน และข้อมูลทดสอบ ข้อมูลที่ใช้สำหรับการสร้างตัวแบบ คือ ข้อมูลพัฒนาตัวแบบ และข้อมูลประเมินใช้สำหรับการพิจารณาตัวแบบจำแนกประเภทที่เหมาะสม

3.2. ขั้นตอนการใช้ตัวแบบที่สร้างโดยใช้การวิเคราะห์การเกาะกลุ่มของตัวทำนายหลากหลาย

ขั้นตอนการใช้ CLAMP กับข้อมูลทดสอบ มีขั้นตอนการพิจารณาอยู่ 2 ขั้นตอนหลัก คือ



รูปที่ 3.3 : แสดงขั้นตอนการใช้ตัวแบบของ CLAMP

1. ขั้นตอนการแยกกลุ่ม ผลลัพธ์ที่ได้จากขั้นตอนนี้ คือ ข้อมูลทดสอบถูกจัดเป็นกลุ่มๆ ตามตัวแบบของการวิเคราะห์การเกาะกลุ่มที่ได้จาก CLAMP
2. ขั้นตอนการทำนาย เป็นการประมวลผลข้อมูลกับตัวแบบจำแนกประเภทของกลุ่ม ผลลัพธ์ของการจำแนกประเภทของข้อมูลทั้งหมดเรียกเป็นผลลัพธ์ของการจำแนกประเภทที่ได้จากการใช้ CLAMP

3.3. โปรแกรมสำหรับทดสอบการวิเคราะห์การเกาะกลุ่มของตัวทำนายหลากหลาย

โปรแกรมสำหรับทดสอบการวิเคราะห์การเกาะกลุ่มของตัวทำนายหลากหลายเป็นโปรแกรมที่พัฒนาตามหลักการของ CLAMP โปรแกรมนี้ถูกพัฒนาโดยใช้ภาษาจาวา ซึ่งเครื่องมือสำหรับการพัฒนา คือ โปรแกรมเนตบีน เวอร์ชัน 5.5 (Netbean5.5) [33] และใช้คลาสในโปรแกรมการทำเหมืองข้อมูลเวกา เวอร์ชัน 3.5.3 (Weka version 3.5.3) [32]

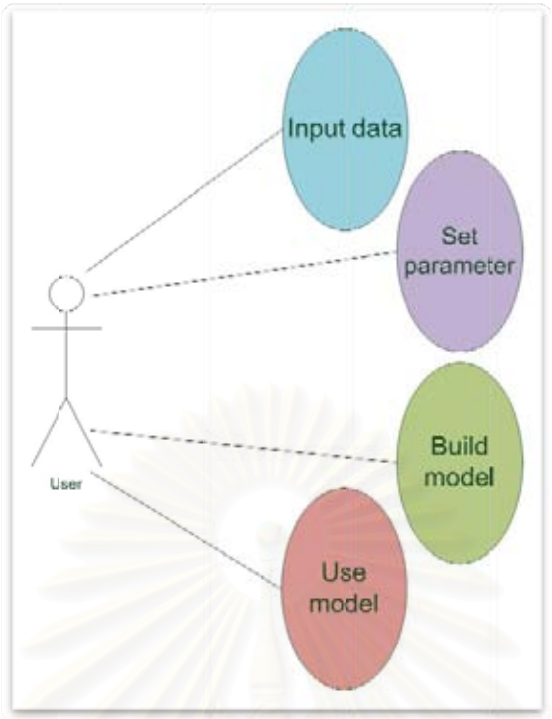
โปรแกรมการทำเหมืองข้อมูลเวกาเป็นโปรแกรมที่ประกอบด้วยคลาสของวิธีการวิเคราะห์ข้อมูลมากมายมีทั้งการวิเคราะห์การเกาะกลุ่ม และวิธีการจำแนกประเภท โดยวิธีการวิเคราะห์การเกาะกลุ่มของโปรแกรมเวกา มีวิธีการวิเคราะห์การเกาะกลุ่มแบบค่าเฉลี่ยเอ็กซ์ (X-means clustering) ซึ่งจะใช้ในงานวิจัยนี้ และสำหรับวิธีการจำแนกประเภทของโปรแกรมการทำเหมืองข้อมูลเวกามีหลายวิธี โดยวิธีที่สนใจในงานวิจัยนี้ คือ การจัดจำแนกแบบเบย์อย่างง่าย (naive Bayes) ต้นไม้การตัดสินใจ (decision tree : J48) เพอร์เซพตรอนหลายชั้น (neural network : multi-layer perceptron) และสมการถดถอยแบบโลจิสติก (regression : logistic regression)

การเลือกตัวทำนายจากผลลัพธ์ของวิธีการจำแนกประเภทแต่ละวิธีจะใช้ความแม่นยำในการทำนายของตัวแบบจำแนกประเภทแต่ละวิธีเปรียบเทียบกัน โดยใช้ข้อมูลประเมินเพื่อพิจารณาหาตัวแบบจำแนกประเภทที่มีความแม่นยำสูงที่สุดในแต่ละกลุ่ม แล้วกำหนดให้เป็นตัวแบบจำแนกประเภทที่เหมาะสมสำหรับกลุ่มนั้น

ลักษณะประจำของข้อมูลที่ใช้ในโปรแกรมสำหรับ CLAMP เป็นลักษณะประจำที่มีค่าต่อเนื่องทั้งหมด ยกเว้นลักษณะประจำเป้าหมายที่เป็นค่าไม่ต่อเนื่อง โปรแกรม CLAMP รองรับตัวกรองข้อมูล 4 วิธี คือ วิธีการแทนที่ข้อมูลที่ค่าขาดหายไป (replace missing value) ด้วยค่าเฉลี่ยหรือค่าฐานนิยม วิธีการเปลี่ยนช่วงของข้อมูลโดยใช้วิธีทำให้เป็นปกติ (normalize) [11] วิธีการเปลี่ยนช่วงของข้อมูลโดยใช้วิธีทำให้เป็นมาตรฐาน (standardize) [11] และ วิธีการเปลี่ยนค่าลักษณะประจำที่เป็นค่าไม่ต่อเนื่องไปเป็นรูปแบบของเลขฐาน 2 (nominal to binary)

โปรแกรม CLAMP กำหนดค่าของพารามิเตอร์ของวิธีการจัดจำแนกแต่ละวิธี 2 ลักษณะ คือ การกำหนดค่าเองจากผู้วิเคราะห์ และการกำหนดให้โปรแกรมหาค่าที่เหมาะสมอัตโนมัติ โดยระบุพารามิเตอร์ที่ต้องการให้โปรแกรมหาค่าที่เหมาะสม จำนวนครั้งที่ต้องการให้โปรแกรมคำนวณหรือช่วงที่ต้องการ

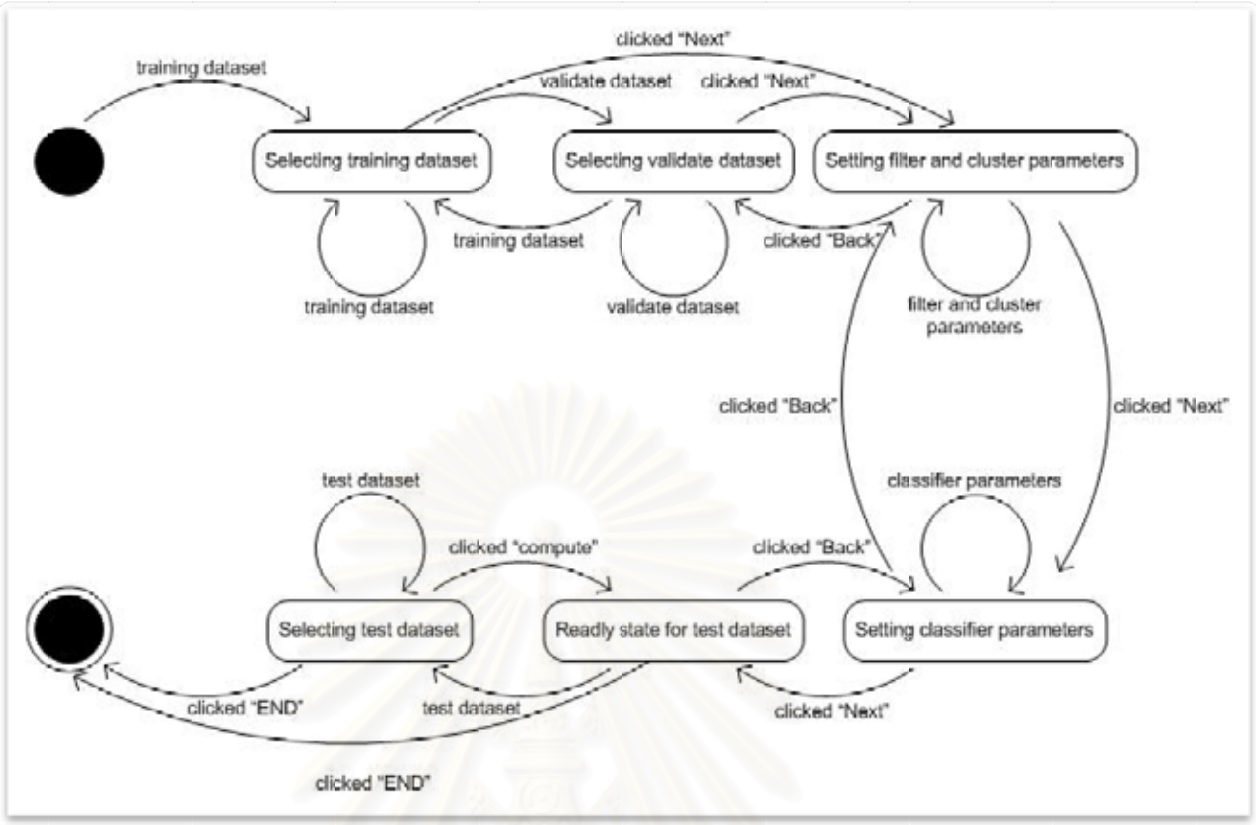
วิธีการเลือกค่าพารามิเตอร์ที่เหมาะสมทำได้โดยแบ่งค่าพารามิเตอร์ที่สนใจเป็นช่วงย่อยเท่าๆกัน แล้วประมวลผลทีละค่าจนกว่าจะผ่านเกณฑ์ที่ต้องการจึงหยุด



รูปที่ 3.4 : Use case diagram สำหรับโปรแกรม CLAMP

Use case diagram ของโปรแกรม CLAMP แสดงอยู่ในรูปที่ 3.4 มีฟังก์ชันการทำงานอยู่ 4 อย่าง คือ

- การนำข้อมูลเข้า (Input data) ข้อมูลที่นำเข้ามี 3 ชนิด คือ ข้อมูลพัฒนาตัวแบบ ข้อมูลประเมิน และข้อมูลทดสอบ
- การกำหนดค่าพารามิเตอร์ (Set parameter) การกำหนดพารามิเตอร์สามารถแบ่งออกเป็น 3 กลุ่ม คือ ตัวแบบสำหรับการกรองข้อมูล (Filter model) ตัวแบบการวิเคราะห์การเกาะกลุ่ม (Cluster model) และ ตัวแบบจำแนกประเภท (Classification model)
- การพัฒนาตัวแบบ (Build model) การพัฒนาตัวแบบสามารถแบ่งออกเป็น 3 กลุ่ม คือ ตัวแบบสำหรับการกรองข้อมูล ตัวแบบการวิเคราะห์การเกาะกลุ่ม และตัวแบบจำแนกประเภท
- การใช้ตัวแบบ (Use model) คือ ใช้ CLAMP กับข้อมูลทดสอบ



รูปที่ 3.5 : Statechart diagram สำหรับโปรแกรม CLAMP

Statechart diagram สำหรับโปรแกรม CLAMP แสดงในรูปที่ 3.5 มีขั้นตอนดังต่อไปนี้

ขั้นตอนการเลือกข้อมูลพัฒนาตัวแบบ (Selecting training data set) คือ ขั้นตอนการรับข้อมูลพัฒนาตัวแบบ ชุดข้อมูลที่เลือกเข้าโปรแกรมจะเก็บไว้ในแฟ้มข้อมูล (file) ที่มีนามสกุลเป็น ARFF

ขั้นตอนการเลือกข้อมูลประเมิน (Selecting validate data set) คือ ขั้นตอนการรับข้อมูลประเมินซึ่งอาจละได้ ถ้าไม่มีการเลือกข้อมูลประเมินโปรแกรมจะเลือกใช้ข้อมูลพัฒนาตัวแบบในการประเมินแทนข้อมูลประเมิน แต่ชุดข้อมูลที่เลือกเข้ามานั้นต้องมีโครงสร้างของชุดข้อมูลเหมือนกับชุดข้อมูลพัฒนาตัวแบบ

ขั้นตอนการเลือกข้อมูลพัฒนาตัวแบบ และขั้นตอนการเลือกข้อมูลประเมินเป็นหน้าต่างแรกสำหรับการทำงานของโปรแกรม โปรแกรมจะทำการเตรียมข้อมูลสำหรับใช้ในการวิเคราะห์การเกาะกลุ่ม และเปลี่ยนหน้าต่างไปเป็นหน้าต่างของขั้นตอนการกำหนดพารามิเตอร์สำหรับตัวแบบการกรองข้อมูล และตัวแบบการวิเคราะห์การเกาะกลุ่มต่อไป

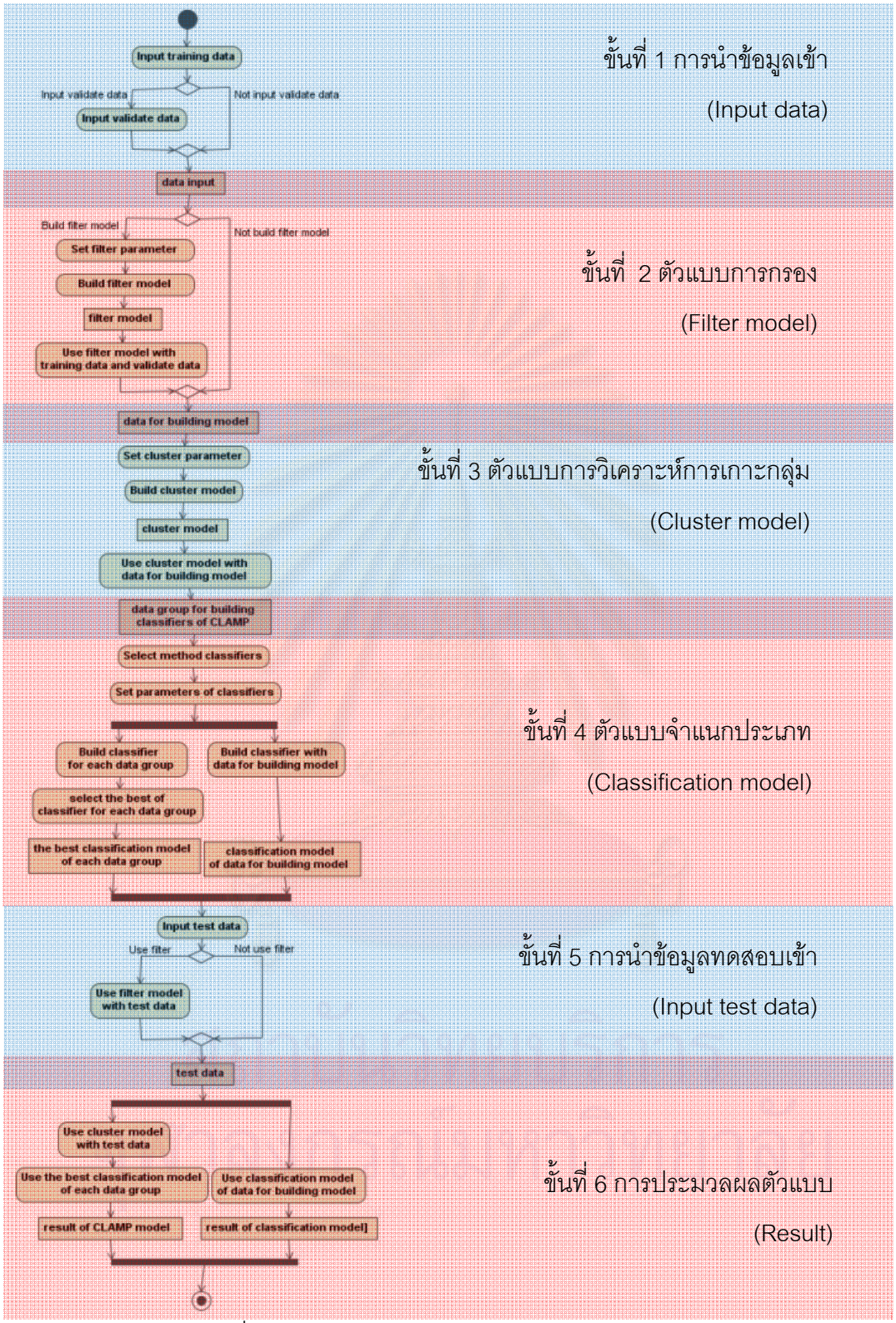
ขั้นตอนการกำหนดพารามิเตอร์สำหรับตัวแบบการกรองข้อมูล และตัวแบบการวิเคราะห์การเกาะกลุ่ม (Setting filter and cluster parameters) คือ ขั้นตอนที่ใช้ในการเลือกวิธีการกรองข้อมูลและวิธีการวิเคราะห์การเกาะกลุ่มพร้อมทั้งกำหนดพารามิเตอร์ที่จำเป็น โปรแกรมจะทำการพัฒนาตัวแบบการกรองข้อมูล และตัวแบบการวิเคราะห์การเกาะกลุ่มพร้อมทั้งแบ่งข้อมูลออกเป็นกลุ่มๆ สำหรับใช้ในการพัฒนาตัวแบบวิธีจำแนกประเภท แล้วจึงเปลี่ยนหน้าต่างไปเป็นหน้าต่างของขั้นตอนการกำหนดพารามิเตอร์สำหรับตัวแบบจำแนกประเภทต่อไป

ขั้นตอนการกำหนดพารามิเตอร์สำหรับตัวแบบจำแนกประเภท (Setting classifier parameters) คือ ขั้นตอนที่ใช้ในการเลือกวิธีการจำแนกประเภท พร้อมทั้งกำหนดพารามิเตอร์ที่จำเป็น แล้วโปรแกรมจะทำการพัฒนาตัวแบบจำแนกประเภทของ CLAMP เทียบกับวิธีการพัฒนาวิธีการจำแนกแบบปกติ แล้วจึงเปลี่ยนหน้าต่างไปเป็นหน้าต่างของขั้นตอนการรับข้อมูลทดสอบต่อไป

ขั้นตอนการรอรับข้อมูลทดสอบ (Ready state for test data set) คือ ขั้นตอนรอรับข้อมูลทดสอบ โดยขั้นตอนนี้จะรอรับชุดข้อมูลทดสอบจากผู้ใช้โปรแกรม และเข้าสู่ขั้นตอนการรับข้อมูลต่อไป

ขั้นตอนการรับข้อมูลทดสอบ (Selecting test data set) คือ ขั้นตอนการรับข้อมูลทดสอบโดยผู้ใช้โปรแกรมเลือกชุดข้อมูลทดสอบ และสั่งให้คำนวณตัวแบบจำแนกประเภทที่พัฒนาขึ้นพร้อมทั้งแสดงผลลัพธ์ แล้วเข้าสู่ขั้นตอนรอรับข้อมูล

Activity diagram สำหรับโปรแกรม CLAMP แสดงในรูปที่ 3.6



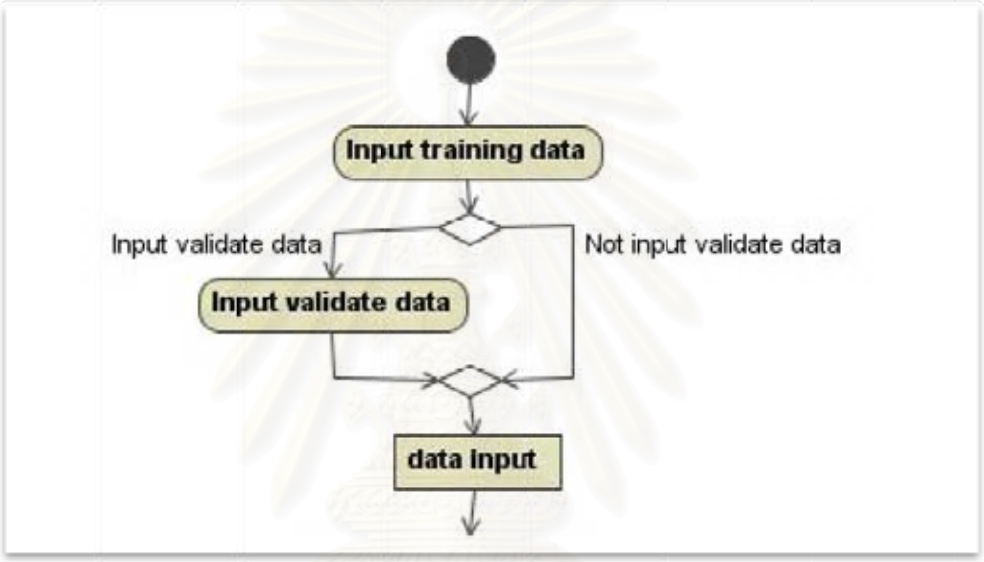
รูปที่ 3.6 : Activity diagram สำหรับโปรแกรม CLAMP

Activity diagram สำหรับโปรแกรม CLAMP แบ่งออกเป็น 6 ขั้นตอนหลัก คือ

ขั้นที่ 1 การนำข้อมูลเข้า (Input data)

การนำข้อมูลเข้าเป็นส่วนการรับข้อมูลพัฒนาตัวแบบ และประเมินตัวแบบ หากไม่มีข้อมูลประเมินตัวแบบจะใช้ข้อมูลพัฒนาตัวแบบแทน เมื่อนำข้อมูลเข้าโปรแกรมแล้ว โปรแกรมจะแสดงค่าทางสถิติเบื้องต้นของลักษณะประจำที่ถูกเลือก

ผลลัพธ์ที่ได้จากการนำข้อมูลเข้าคือข้อมูลพัฒนาตัวแบบ



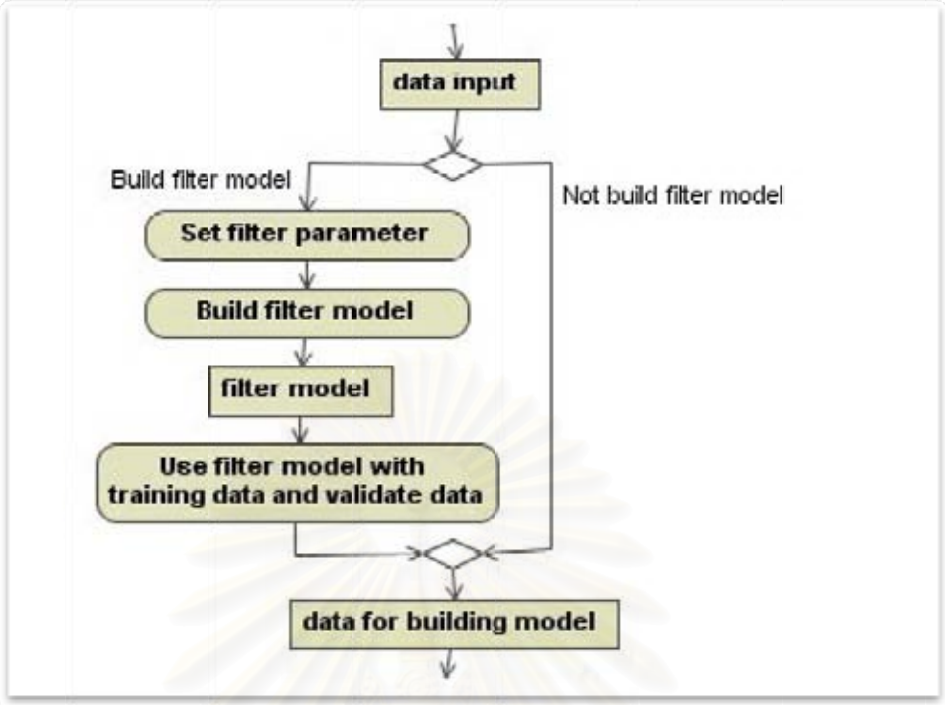
รูปที่ 3.7 : Activity diagram ขั้นที่ 1 การนำข้อมูลเข้า (Input data)

ขั้นที่ 2 ตัวแบบการกรอง (Filter model)

การพัฒนาตัวแบบการกรองข้อมูลเป็นการกำหนดพารามิเตอร์ เพื่อใช้กรองข้อมูลพัฒนาตัวแบบ และข้อมูลพัฒนาตัวแบบ วิธีการกรองที่สามารถเรียกใช้ได้กล่าวแล้วในบทที่ 3 ซึ่งเป็นการเรียกใช้คลาสของวิธีการกรองที่มีอยู่ในเวกา

ความแตกต่างของตัวแบบการกรองที่รองรับในโปรแกรม CLAMP กับวิธีการกรองข้อมูลในคลาสของโปรแกรมเวกา คือ วิธีการกรองข้อมูลในคลาสของโปรแกรมเวกาเป็นการกรองข้อมูลภายในชุดข้อมูลเท่านั้น แต่วิธีการกรองที่รองรับในโปรแกรม CLAMP เป็นการพัฒนาตัวแบบการกรองข้อมูลจากข้อมูลพัฒนาตัวแบบ และข้อมูลประเมิน

ผลลัพธ์ที่ได้จากการพัฒนาตัวแบบการกรองข้อมูล คือ ข้อมูลที่พร้อมสำหรับใช้พัฒนาตัวแบบจำแนกประเภทแต่ละวิธี

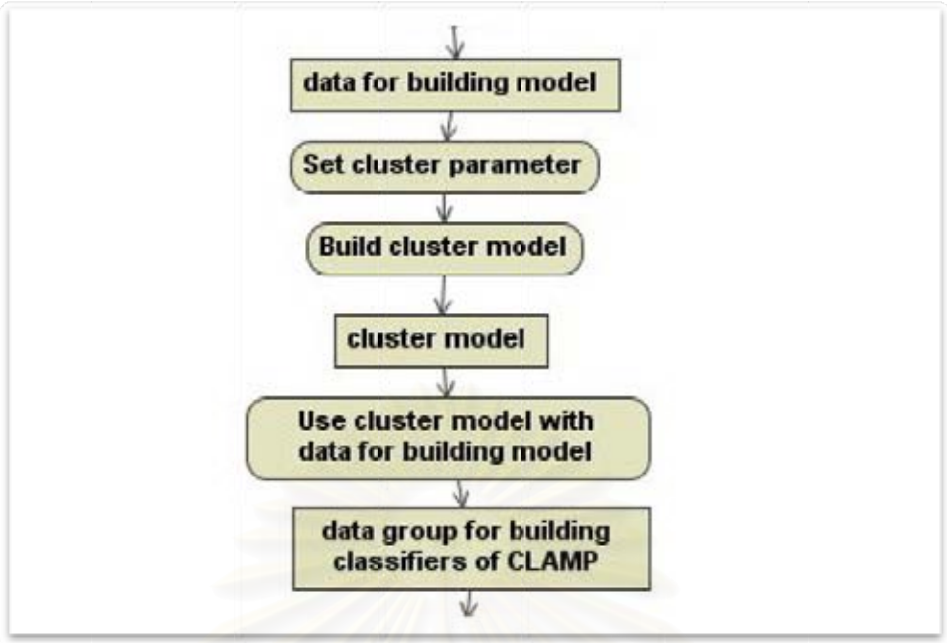


รูปที่ 3.8 : Activity diagram ขั้นที่ 2 ตัวแบบการกรอง (Filter model)

ขั้นที่ 3 ตัวแบบการวิเคราะห์การเกาะกลุ่ม (Cluster model)

การพัฒนาตัวแบบการวิเคราะห์การเกาะกลุ่มเป็นการเรียกใช้คลาสของขั้นตอนวิธีการวิเคราะห์การเกาะกลุ่มด้วยค่าเฉลี่ยเอ็กซ์ และการกำหนดค่าพารามิเตอร์ ข้อมูลที่ใช้สร้างตัวแบบการวิเคราะห์การเกาะกลุ่ม คือ ข้อมูลพัฒนาตัวแบบ ซึ่งตัดลักษณะประจำตัวสุดท้ายออก (เพราะเป็นลักษณะประจำเป้าหมาย)

ผลลัพธ์ที่ได้จากการพัฒนาตัวแบบการวิเคราะห์การเกาะกลุ่ม คือ กลุ่มของข้อมูลที่ผ่านการวิเคราะห์การเกาะกลุ่มที่พร้อมใช้ในการพัฒนาตัวแบบจำแนกประเภท



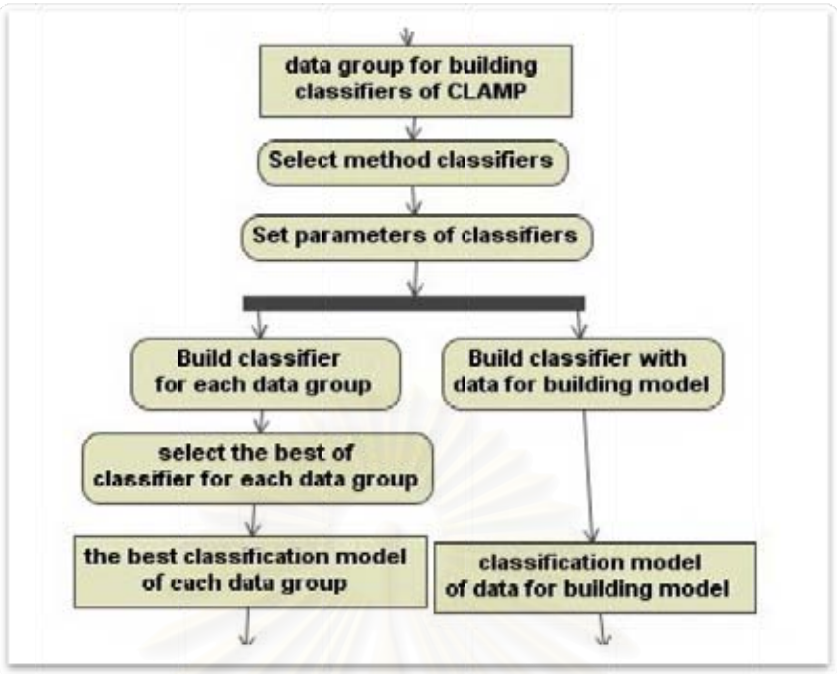
รูปที่ 3.9 : Activity diagram ขั้นที่ 3 ตัวแบบการวิเคราะห์การเกาะกลุ่ม (Cluster model)

ขั้นที่ 4 ตัวแบบจำแนกประเภท (Classification model)

ตัวแบบจำแนกประเภทเป็นการเลือกตัวแบบจำแนกประเภท และกำหนดค่าพารามิเตอร์ของตัวแบบจำแนกประเภทที่เหมาะสมของแต่ละกลุ่ม

CLAMP แบ่งกลุ่มข้อมูลออกเป็นส่วนๆ จากข้อมูลพัฒนาตัวแบบ แล้วใช้ข้อมูลประเมินเพื่อประเมินตัวแบบจำแนกประเภทที่เหมาะสมสำหรับแต่ละกลุ่มข้อมูล

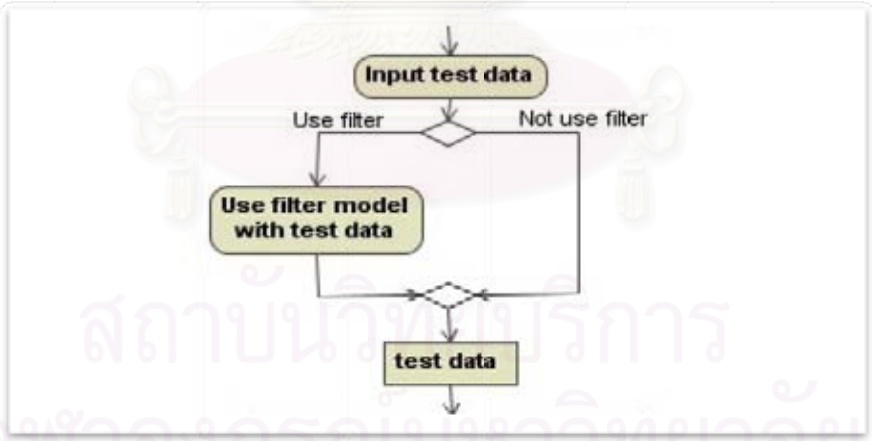
ผลลัพธ์ที่ได้ คือ ตัวแบบจำแนกประเภทสำหรับแต่ละกลุ่ม



รูปที่ 3.10 : Activity diagram ขั้นที่ 4 ตัวแบบจำแนกประเภท (Classification model)

ขั้นที่ 5 การนำข้อมูลทดสอบเข้า (Input test data)

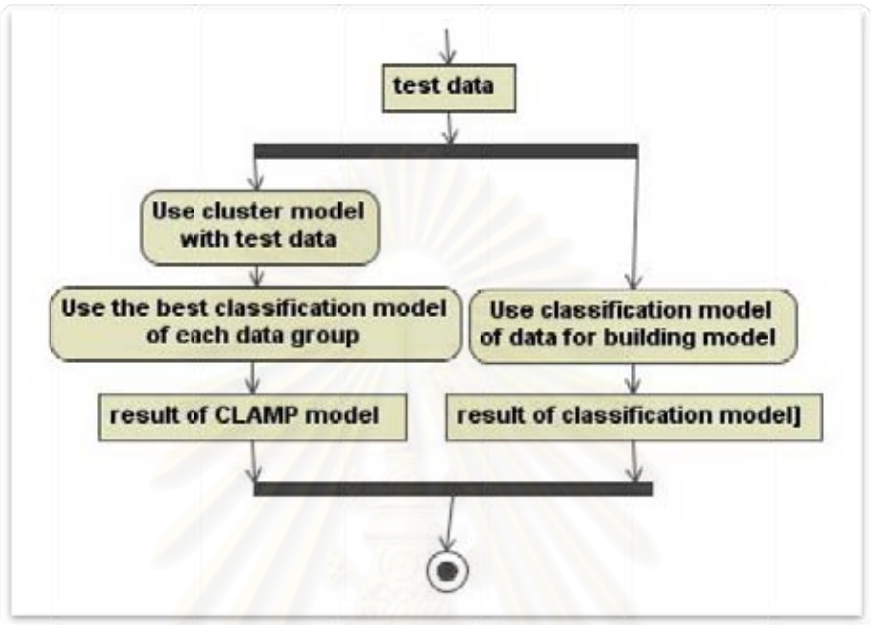
การนำข้อมูลทดสอบเข้าเป็นส่วนการรับข้อมูลทดสอบที่ประเมิน CLAMP โดยโปรแกรมจะแสดงค่าทางสถิติเบื้องต้นของลักษณะประจำตามที่ผู้ใช้โปรแกรมเลือก



รูปที่ 3.11 : Activity diagram ขั้นที่ 5 การนำข้อมูลทดสอบเข้า (Input test data)

ขั้นที่ 6 การประมวลผลตัวแบบ (Result)

การประมวลผลตัวแบบจำแนกประเภทกับข้อมูลทดสอบเป็นการประมวลผล CLAMP กับข้อมูลทดสอบ



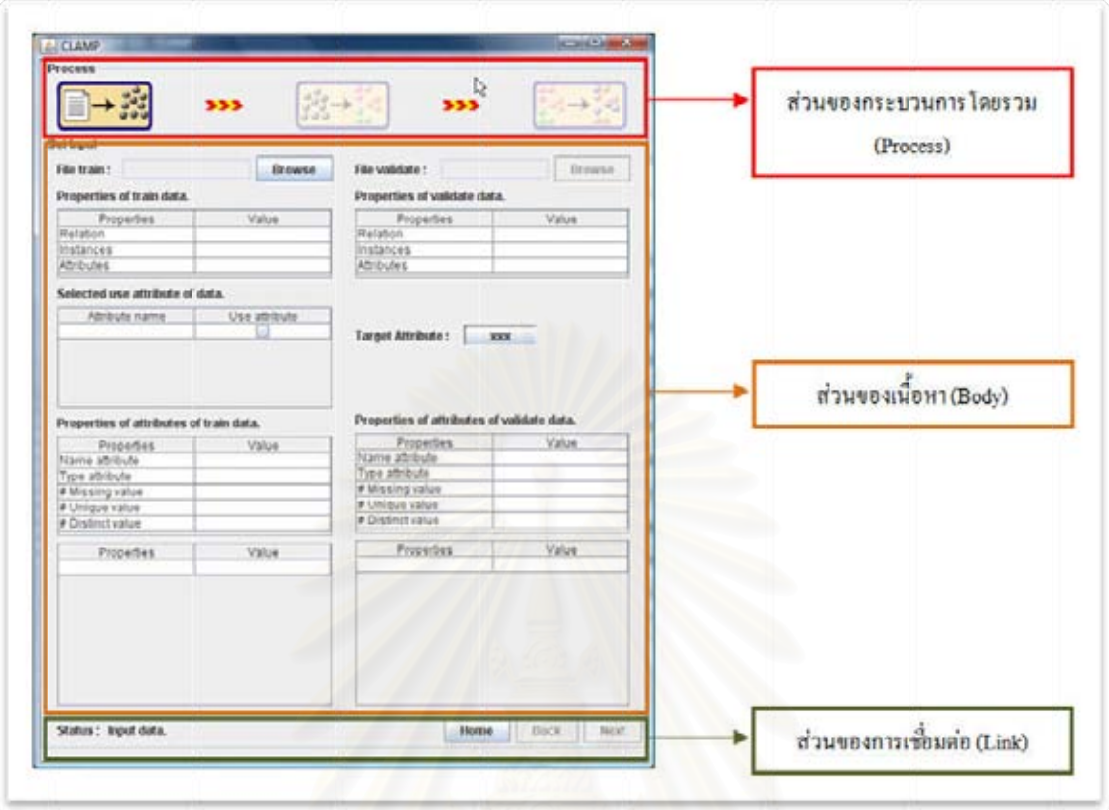
รูปที่ 3.12 : Activity diagram ขั้นที่ 6 การประมวลผลตัวแบบ (Result)

3.4. การทำงานของโปรแกรมสำหรับการวิเคราะห์การเกาะกลุ่มของตัวทำนายหลากหลาย

โมดูลของโปรแกรมสำหรับ CLAMP สามารถแบ่งออกเป็นส่วนๆได้ดังนี้

1. โมดูลของการรับข้อมูลพัฒนาตัวแบบ และข้อมูลประเมิน ขั้นตอนใน Activity diagram ที่เกี่ยวข้อง คือ ขั้นตอนที่ 1 การนำข้อมูลเข้า
2. โมดูลของการสร้างตัวแบบการวิเคราะห์การเกาะกลุ่ม ขั้นตอนใน Activity diagram ที่เกี่ยวข้อง คือ ขั้นที่ 2 ตัวแบบการกรอง และขั้นที่ 3 ตัวแบบการวิเคราะห์การเกาะกลุ่ม
3. โมดูลของการสร้างตัวแบบการจำแนกประเภท ขั้นตอนใน Activity diagram ที่เกี่ยวข้อง คือ ขั้นที่ 4 ตัวแบบจำแนกประเภท
4. โมดูลของการทดสอบตัวแบบกับข้อมูลทดสอบตัวแบบ ขั้นตอนใน Activity diagram ที่เกี่ยวข้อง คือ ขั้นที่ 5 การนำข้อมูลทดสอบเข้า และขั้นที่ 6 การประมวลผลตัวแบบ

แต่ละโมดูลจะมีหน้าต่างทำงานซึ่งจะถูกแบ่งออกเป็น 3 ส่วนใหญ่ๆ คือ



รูปที่ 3.13 : การแสดงส่วนประกอบในหนึ่งหน้าต่างทำงาน

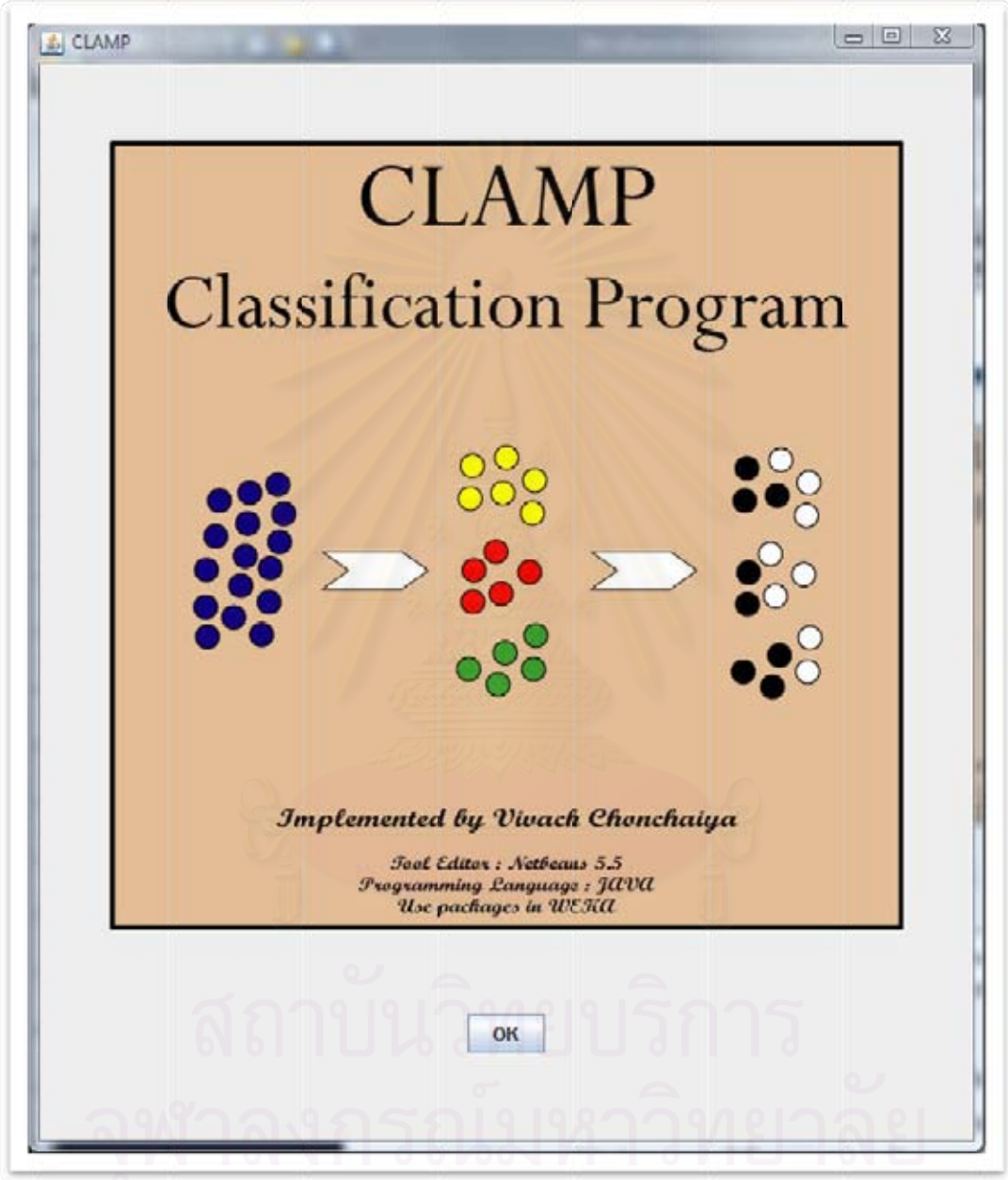
ส่วนแรก คือ ส่วนของการแสดงกระบวนการโดยรวม เป็นส่วนที่แสดงขั้นที่ประมวลผลด้วยรูปของกระบวนการ (process)

ส่วนที่สอง คือ ส่วนของการแสดงรายละเอียดต่างๆของโปรแกรม ในส่วนนี้ใช้ในการปรับเปลี่ยนวิธี และค่าพารามิเตอร์ และใช้ในการแสดงรายละเอียดของข้อมูลเบื้องต้นด้วย ในเอกสารนี้จะเรียกส่วนนี้ว่าส่วนของเนื้อหา (body)

ส่วนที่สามเป็นส่วนที่ใช้ในการเชื่อมต่อแต่ละหน้าต่าง โดยส่วนประกอบที่สำคัญจะมีเพียงปุ่ม และข้อความแสดงสถานะเท่านั้น ซึ่งสามารถคลิกได้เท่านั้น ในเอกสารนี้จะเรียกส่วนนี้ว่าส่วนของการเชื่อมต่อ (link)

โดยแต่ละหน้าต่างมีรายละเอียดดังนี้

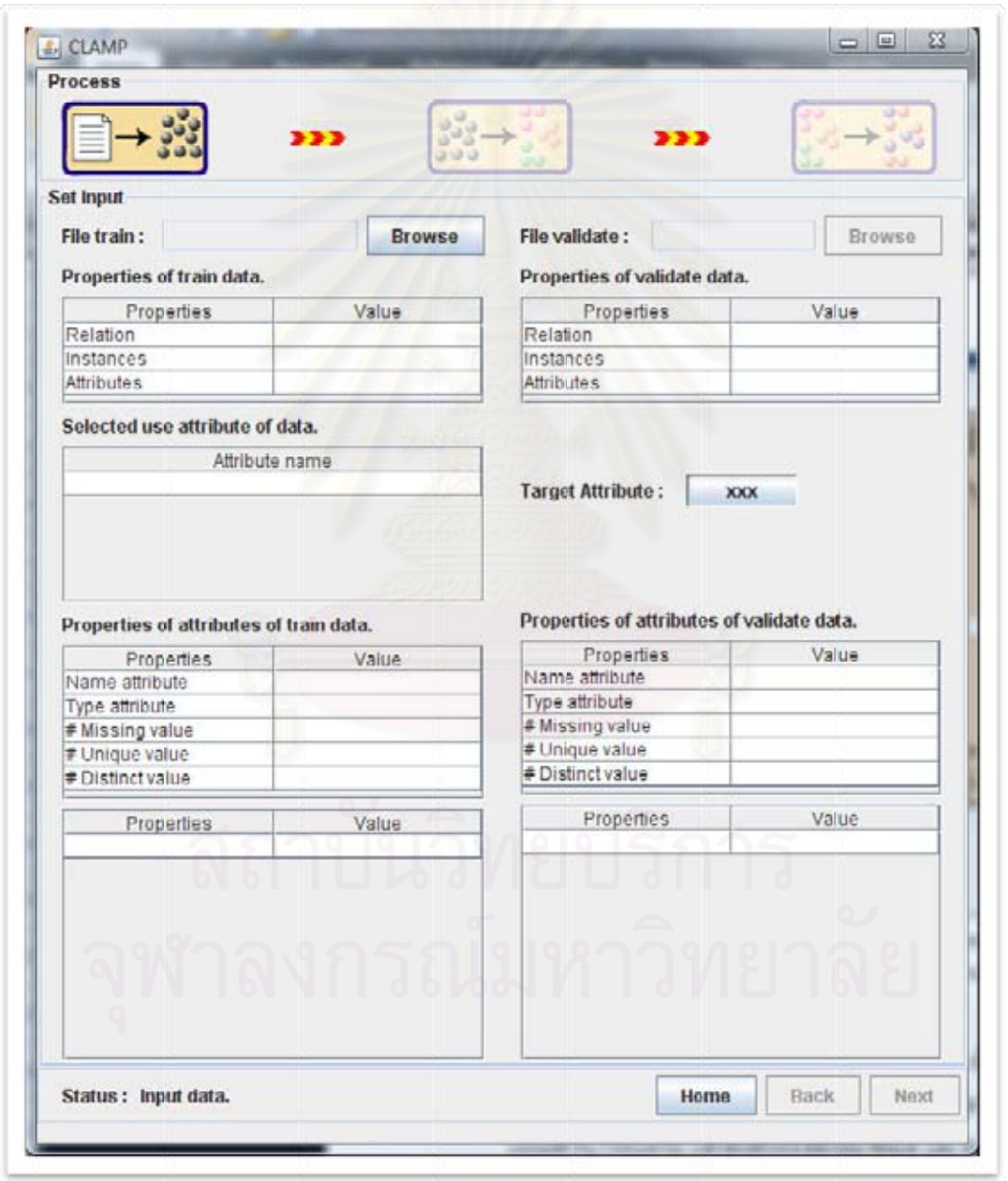
1. หน้าต่างแรกของการทำงานของโปรแกรมสำหรับ CLAMP ประกอบด้วย ชื่อของโปรแกรม รูปของโปรแกรม CLAMP ชื่อของผู้พัฒนาโปรแกรม โปรแกรมที่ใช้ในการพัฒนาโปรแกรม ภาษาที่ใช้พัฒนา คลาสที่ใช้ในการพัฒนาโปรแกรม



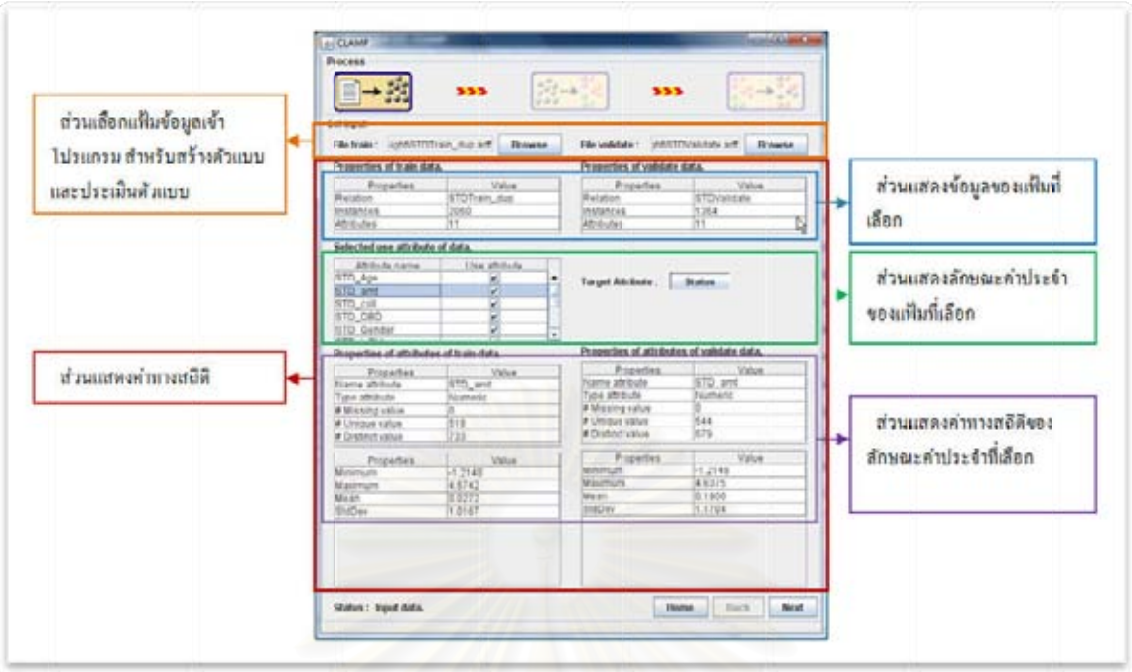
รูปที่ 3.14 : หน้าแรกของโปรแกรม CLAMP

โดยหน้าต่างเป็นหน้าแรกที่แสดงการใช้งานโปรแกรม CLAMP สำหรับพัฒนาตัวแบบ และข้อมูลสำหรับทดสอบตัวแบบเข้าในโปรแกรม ไปเป็นกระบวนการการวิเคราะห์การเกาะกลุ่ม ตามด้วยกระบวนการการจำแนกประเภทแล้วจบด้วยการนำไปทดสอบกับข้อมูลทดสอบที่ต้องการ

2. หน้าต่างของการรับข้อมูลพัฒนาตัวแบบ และข้อมูลประเมิน (โมดูลของการรับข้อมูลพัฒนาตัวแบบ และข้อมูลประเมิน)



รูปที่ 3.15 : หน้าต่างของการรับข้อมูลพัฒนาตัวแบบ และข้อมูลประเมิน



รูปที่ 3.16 : ส่วนประกอบในหน้าต่างของการรับข้อมูลพัฒนาตัวแบบ และข้อมูลประเมิน

ส่วนประกอบที่สำคัญ คือ

- ส่วนเลือกแฟ้มข้อมูลเข้าโปรแกรมแบ่งเป็น แฟ้มข้อมูลพัฒนาตัวแบบ และแฟ้มข้อมูลประเมิน โดยแฟ้มข้อมูลที่สามารถใช้ในโปรแกรมได้ต้องเป็นแฟ้มข้อมูลแบบ ARFF (แฟ้มข้อมูลที่มีนามสกุลเป็น .arff และโครงสร้างข้อมูลภายในตรงตามข้อกำหนดของแฟ้มข้อมูลแบบ ARFF) ในกรณีของข้อมูลประเมินและข้อมูลทดสอบต้องมีโครงสร้างการบันทึกข้อมูลเหมือนกับข้อมูลพัฒนาตัวแบบ

- ส่วนแสดงค่าสถิติเบื้องต้นของข้อมูลที่รับเข้าในโปรแกรม ประกอบไปด้วย 3 ส่วน คือ

  - ส่วนแสดงรายละเอียดของแฟ้มข้อมูลที่นำเข้าในโปรแกรม ประกอบไปด้วย ชื่อของข้อมูล (relation) จำนวนของข้อมูล(number of instance : Instance) จำนวนของลักษณะประจำ (number of attribute : Attribute)

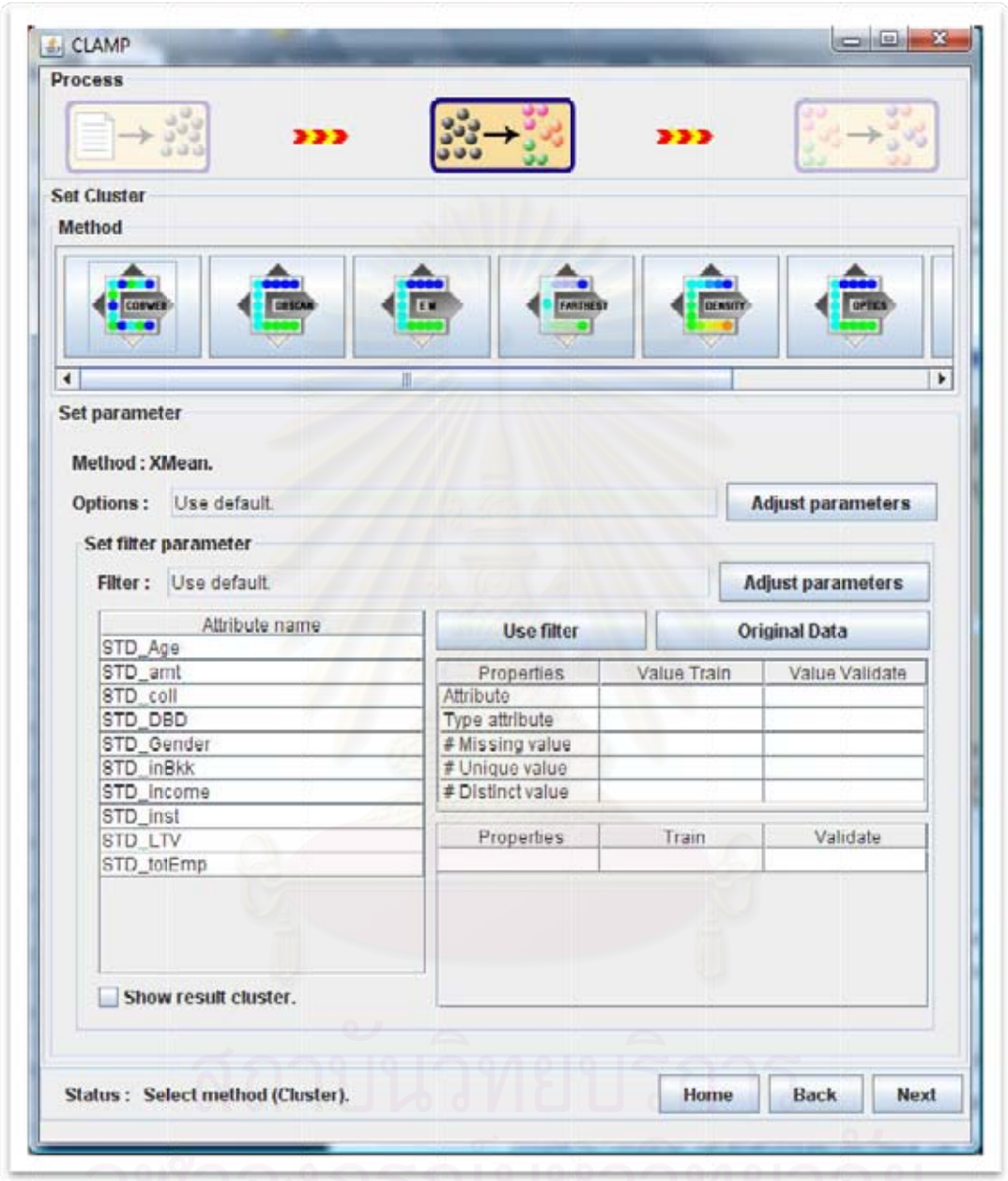
  - ส่วนแสดงลักษณะประจำทุกลักษณะประจำที่มีอยู่ในแฟ้มข้อมูล โดยลักษณะประจำตัวสุดท้ายจะถูกตัดออกจากตารางเพราะถือว่าว่าเป็น

ลักษณะประจำที่ถูกกำหนดให้เป็นลักษณะประจำเป้าหมาย (target) แต่จะแสดงชื่อลักษณะประจำไว้ที่ปุ่มข้างขวาของตาราง เพื่อให้สามารถเลือกดูค่าสถิติเบื้องต้นได้

- ส่วนแสดงค่าสถิติเบื้องต้นของลักษณะประจำที่ถูกเลือก ซึ่งค่าทางสถิตินี้แบ่งออกเป็น 2 ตาราง คือ ตารางแรกแสดงค่าสถิติที่ใช้ได้กับทุกลักษณะประจำ ทั้งลักษณะประจำที่ต่อเนื่อง และลักษณะประจำที่ไม่ต่อเนื่อง ค่าที่แสดงในตารางนี้ คือ ชื่อของลักษณะประจำที่เลือก (name attribute) ชนิดของค่าลักษณะประจำที่เลือก (type attribute) จำนวนข้อมูลสูญหาย (number of missing value : # Missing value) จำนวนของค่าของลักษณะประจำที่มีข้อมูลที่มีเพียงตัวเดียว (number of unique value : # Unique value) และจำนวนของข้อมูลที่แตกต่างกันทั้งหมด (number of distinct value : # Distinct value) ตารางที่ 2 แสดงค่าสถิติที่แตกต่างระหว่างลักษณะประจำแบบต่อเนื่อง และลักษณะประจำแบบไม่ต่อเนื่อง โดยลักษณะประจำแบบต่อเนื่องจะแสดงค่าสถิติ คือ ค่าน้อยที่สุด (minimum) ค่ามากที่สุด (maximum) ค่าเฉลี่ย (means) และค่าส่วนเบี่ยงเบนมาตราฐาน (standard deviation) ส่วนลักษณะประจำแบบไม่ต่อเนื่องจะแสดงค่าสถิติ คือ จำนวนระเบียนของแต่ละค่าของลักษณะประจำ

ในส่วนของการเชื่อมต่อจะแตกต่างจากหน้าต่างทำงานอื่น คือ หน้าต่างนี้จะต้องมีการนำข้อมูลพัฒนาตัวแบบเข้าในโปรแกรมก่อนจึงจะสามารถกดปุ่ม “Next” เพื่อเปิดหน้าต่างของการวิเคราะห์การเกาะกลุ่มได้ และไม่สามารถกดปุ่ม “Back” ได้ ส่วนปุ่ม “Home” เมื่อคลิกแล้วจะแสดงหน้าต่างแรกของโปรแกรม

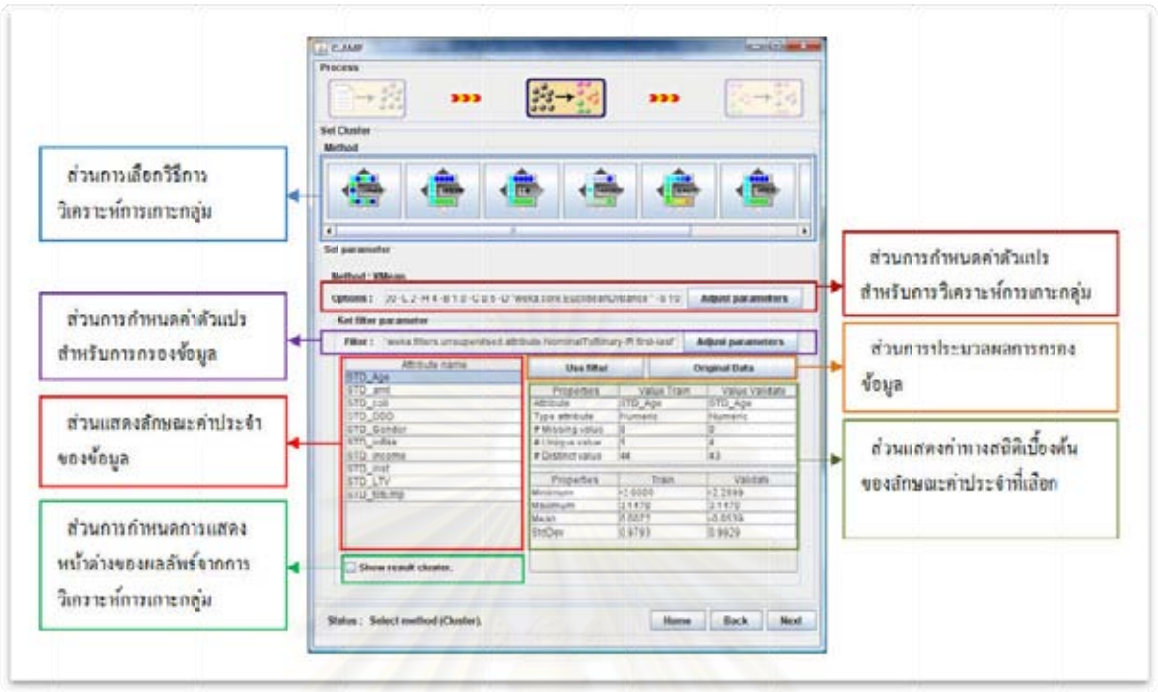
3. หน้าต่างของการเลือกวิธี และกำหนดค่าพารามิเตอร์ สำหรับการวิเคราะห์การเกาะกลุ่มของข้อมูล (โมดูลของการสร้างตัวแบบการวิเคราะห์การเกาะกลุ่ม)



รูปที่ 3.17 : หน้าต่างของการเลือกวิธี และกำหนดค่าพารามิเตอร์
สำหรับการวิเคราะห์การเกาะกลุ่มของข้อมูล

ในส่วนของกระบวนการรูปกลาง แสดงให้เห็นถึงการนำข้อมูลมาวิเคราะห์การเกาะกลุ่ม

ส่วนประกอบที่สำคัญ คือ



รูปที่ 3.18 : ส่วนประกอบของหน้าต่างของการเลือกวิธี
และกำหนดค่าพารามิเตอร์ สำหรับการวิเคราะห์การเกาะกลุ่ม

- ส่วนเลือกวิธีการวิเคราะห์การเกาะกลุ่ม ที่มีอยู่ทั้งหมด 8 ขั้นตอนวิธี ซึ่งแต่ละขั้นตอนวิธีเป็นวิธีการวิเคราะห์การเกาะกลุ่มที่มีอยู่ในเวกาเวอร์ชัน 3.5.3 โปรแกรมบันทึกค่าพารามิเตอร์สำหรับแต่ละวิธีไว้ ดังนั้นเมื่อกลับมาเลือกวิธีการวิเคราะห์การเกาะกลุ่มใหม่ โปรแกรมสามารถแสดงค่าพารามิเตอร์เดิมที่กำหนดไว้ได้

- ส่วนกำหนดค่าพารามิเตอร์สำหรับการวิเคราะห์การเกาะกลุ่มประกอบด้วย 2 ส่วนย่อย คือ ส่วนแสดงค่าพารามิเตอร์สำหรับการวิเคราะห์การเกาะกลุ่มกับส่วนที่ 2 คือปุ่ม “Adjust parameter” สำหรับเปลี่ยนแปลงค่าพารามิเตอร์ เมื่อคลิกที่ปุ่มนี้จะแสดงหน้าต่างที่ใช้กำหนดค่าพารามิเตอร์ ซึ่งได้มาจากคลาสของเวกา

- ส่วนกำหนดค่าพารามิเตอร์สำหรับการกรองข้อมูล ส่วนนี้ใช้หลักการเดียวกับการกำหนดค่าพารามิเตอร์สำหรับการวิเคราะห์การเกาะกลุ่มเพียงแต่การกรองข้อมูลจะใช้โมดูลการกรองข้อมูลแบบหลายวิธี (multi-filter) จาก

ซอฟต์แวร์เวกาเท่านั้น และผู้ใช้สามารถใช้วิธีการกรองข้อมูลกับข้อมูลประเมิน และข้อมูลทดสอบดังที่กล่าวไว้ข้างต้น

- ส่วนแสดงลักษณะประจำของข้อมูล ส่วนนี้มีไว้เพื่อแสดงค่าทางสถิติเบื้องต้นของลักษณะประจำที่เลือกหลังจากทำการประมวลผลการกรองข้อมูลแล้ว

- ส่วนประมวลผลการกรองข้อมูลประกอบด้วยปุ่ม 2 ปุ่ม ปุ่มแรกคือปุ่มที่ชื่อ “Use filter” ใช้สำหรับประมวลผลการกรองข้อมูล ปุ่มที่ 2 คือปุ่ม “Original Data” เป็นปุ่มที่ใช้สำหรับการเรียกข้อมูลต้นฉบับก่อนทำการประมวลผลการกรองข้อมูล

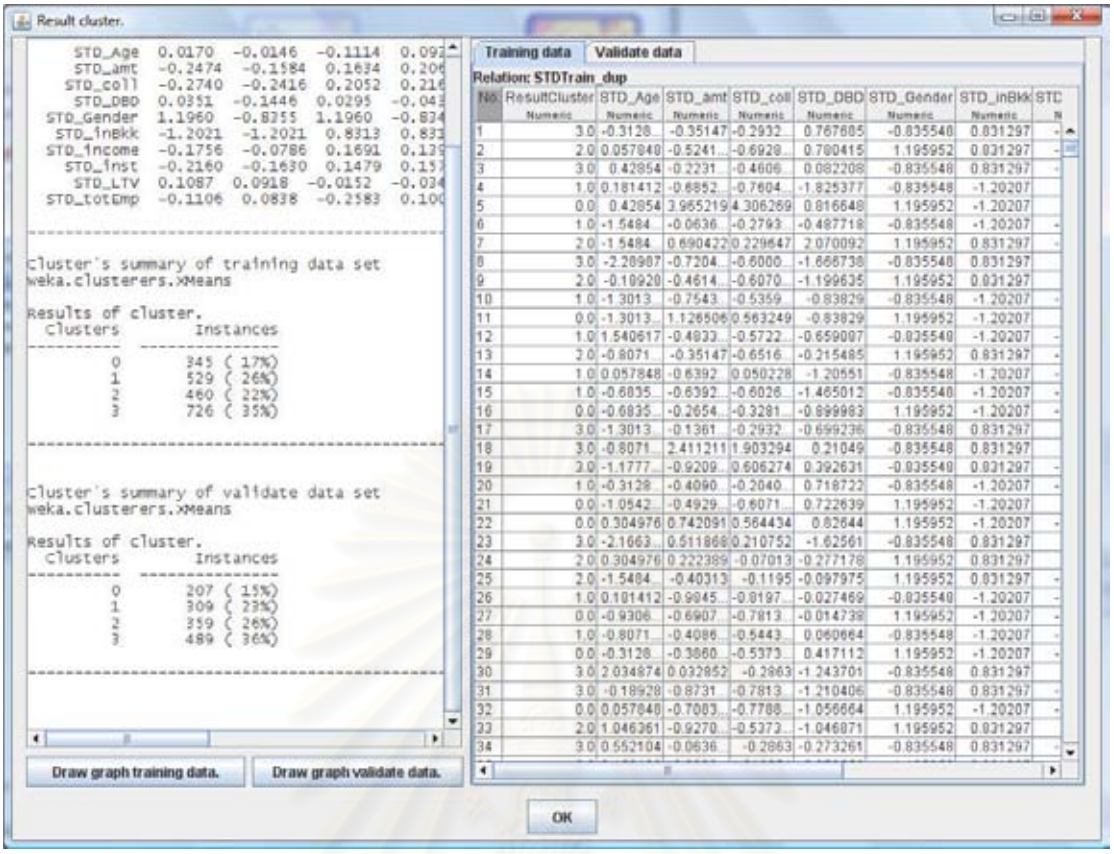
- ส่วนการแสดงค่าทางสถิติเบื้องต้นของลักษณะประจำที่เลือก เป็นส่วนที่ใช้สำหรับแสดงผลลัพธ์หลังจากผ่านการประมวลผลของการกรองข้อมูล

- ส่วนกำหนดการแสดงหน้าต่างผลลัพธ์จากการวิเคราะห์การเกาะกลุ่มใช้สำหรับกำหนดให้แสดงหน้าต่างของผลลัพธ์ที่ได้จากการวิเคราะห์การเกาะกลุ่ม โดยปกติจะไม่แสดงผลลัพธ์ทางจอภาพ

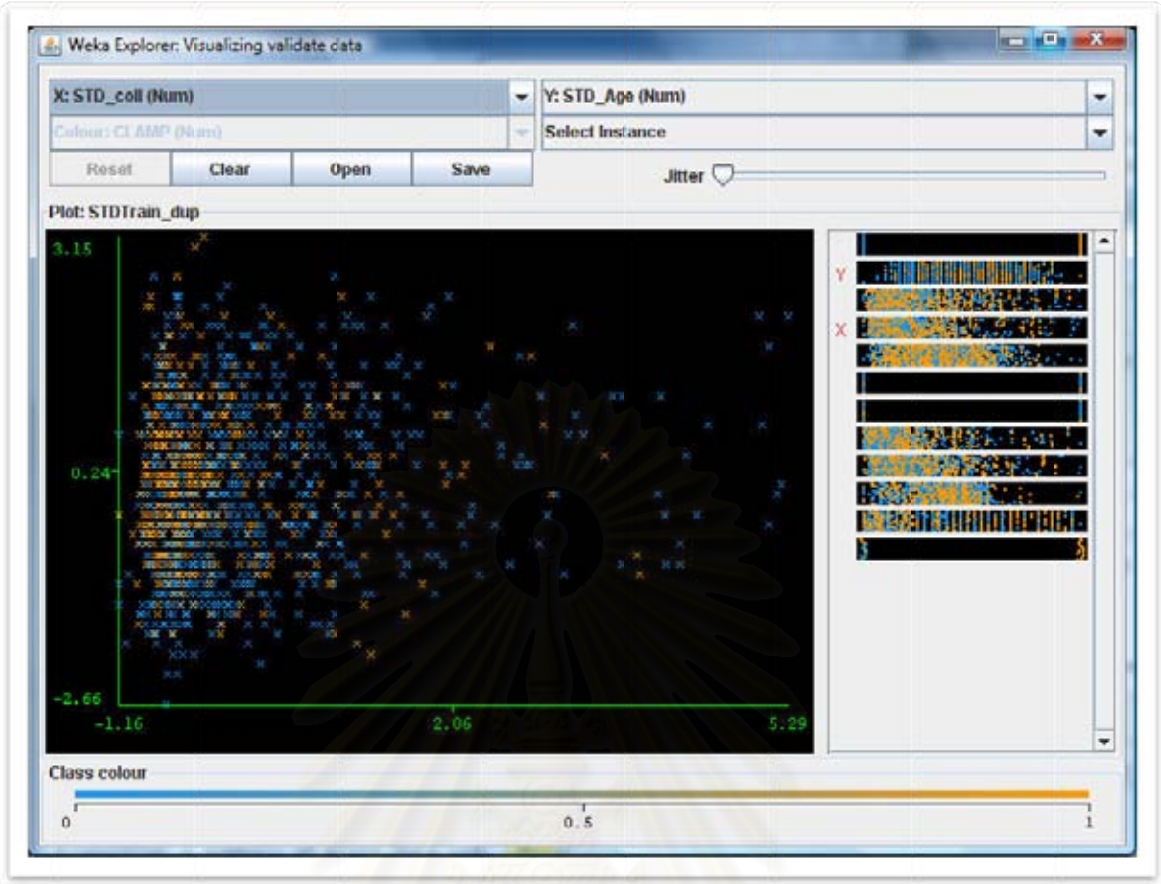
ในส่วนของการเชื่อมต่อมี 3 ปุ่ม คือ ปุ่มแรกปุ่ม “Home” ใช้สำหรับไปหน้าต่างแรก ปุ่มที่ 2 ปุ่ม “Back” ใช้สำหรับกลับไปหน้าต่างของการรับข้อมูล และปุ่มสุดท้ายปุ่ม “Next” เป็นปุ่มที่จะเชื่อมต่อไปยังหน้าต่างของการเลือกวิธี และกำหนดค่าพารามิเตอร์สำหรับวิธีการจำแนกประเภท ในกรณีที่ผู้วิเคราะห์เลือกการแสดงผลลัพธ์ โปรแกรมจะแสดงผลลัพธ์จากกระบวนการทำงานในการพัฒนาตัวแบบการวิเคราะห์การเกาะกลุ่มและทดสอบกับข้อมูลพัฒนาตัวแบบ และข้อมูลประเมิน หลังการกดปุ่ม “Next”

ผลลัพธ์ในหน้าต่างแสดงผลลัพธ์ที่ได้จากการประมวลผลของขั้นตอนการวิเคราะห์การเกาะกลุ่มประกอบด้วย 3 ส่วน คือ

- ส่วนแสดงค่าทางสถิติที่ได้จากการวิเคราะห์การเกาะกลุ่ม
- ส่วนแสดงตารางข้อมูลที่เพิ่มลักษณะประจำที่ระบุกลุ่มของข้อมูล
- ส่วนของปุ่มแสดงหน้าต่างกราฟของข้อมูลพัฒนาตัวแบบ และข้อมูลประเมินของคู่ของลักษณะประจำ

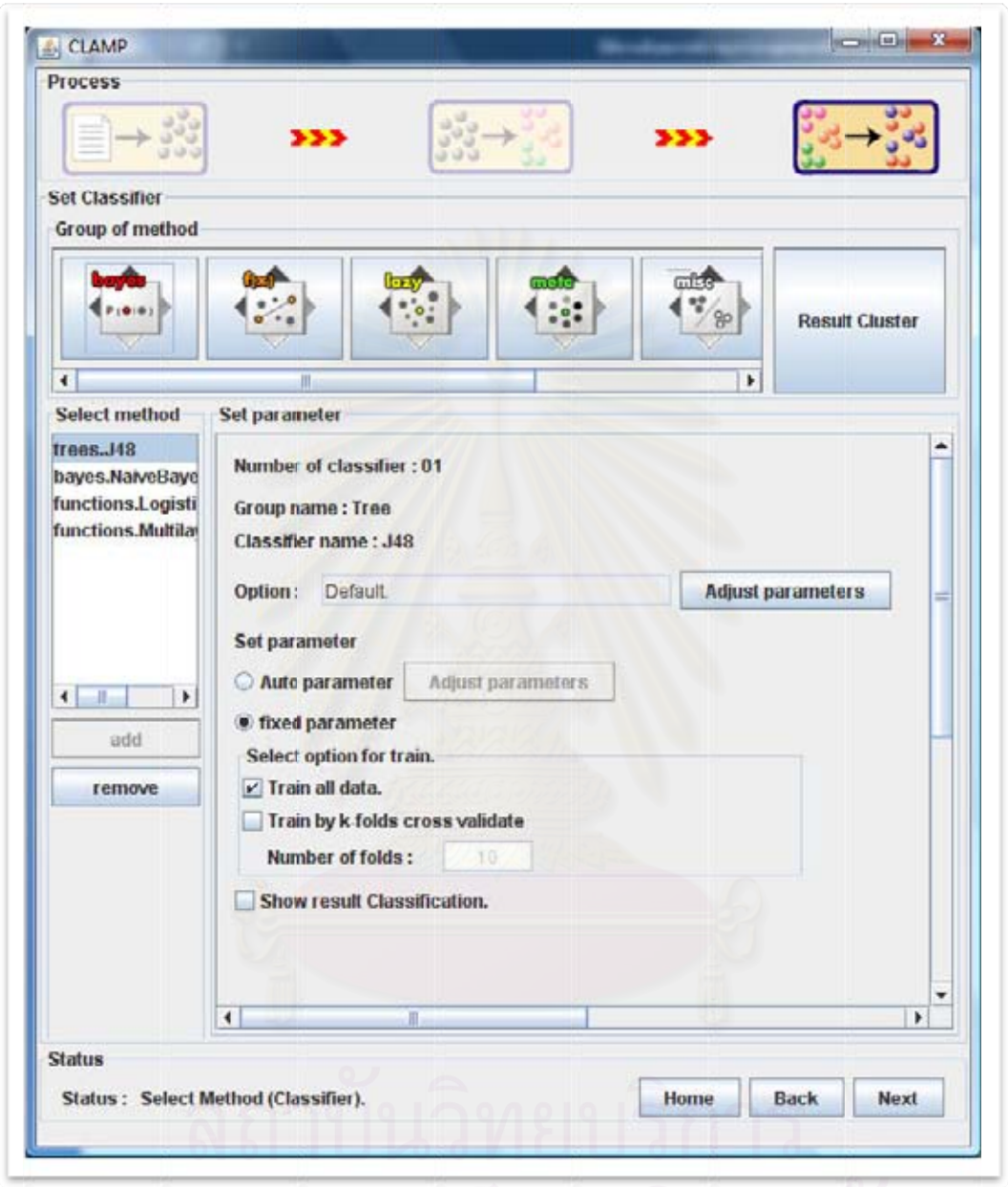


รูปที่ 3.19 : หน้าต่างแสดงผลลัพธ์ที่ได้จากการประมวลผลของ
ขั้นตอนการวิเคราะห์การเกาะกลุ่ม



รูปที่ 3.20 : หน้าต่างกราฟแสดงพิกัดระหว่างลักษณะประจำ 2 ค่า

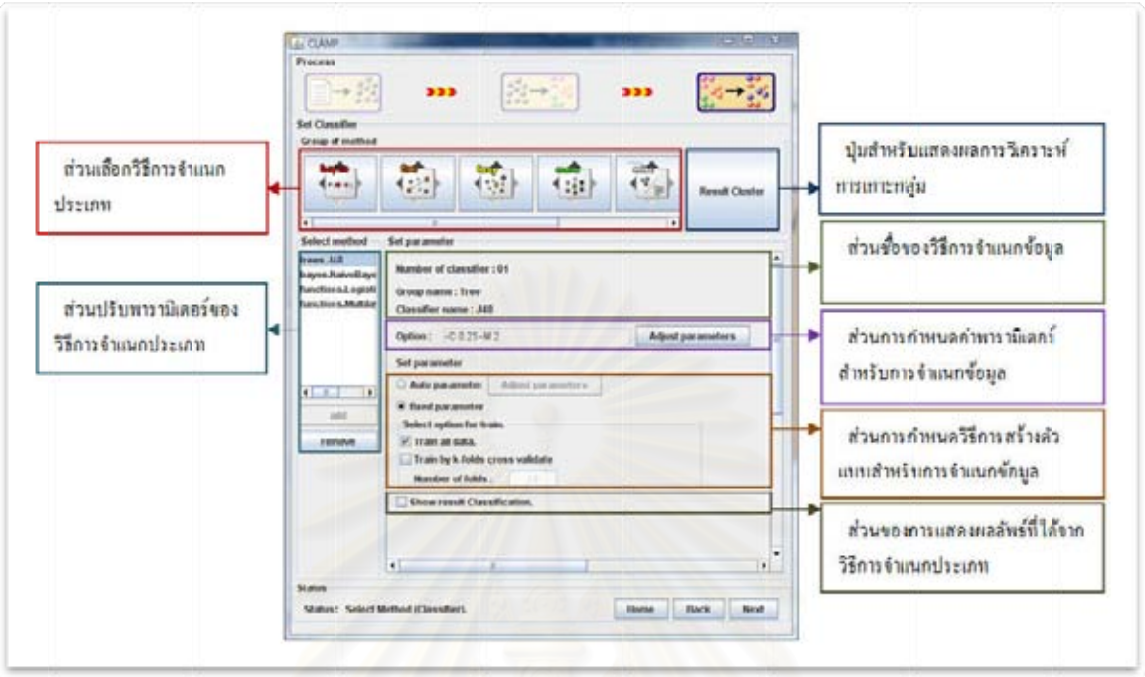
4. หน้าต่างของการเลือกวิธี และกำหนดค่าพารามิเตอร์ สำหรับวิธีการจำแนกประเภท (โมดูลของการสร้างตัวแบบการจำแนกประเภท)



รูปที่ 3.21 : หน้าต่างของการเลือกวิธี และกำหนดค่าพารามิเตอร์ สำหรับวิธีการจำแนกประเภท

ในส่วนของกระบวนการสุดท้ายคือ การพัฒนาตัวแบบจำแนกประเภทของข้อมูลที่ผ่านการวิเคราะห์การเกาะกลุ่ม

ส่วนประกอบของหน้าต่างประกอบไปด้วย



รูปที่ 3.22 : ส่วนประกอบในหน้าต่างของการเลือกวิธี และกำหนดค่าพารามิเตอร์สำหรับวิธีการจำแนกประเภท

- ส่วนการเลือกวิธีการจำแนกประเภท โดยโปรแกรมได้จัดวิธีการไว้เป็นกลุ่ม คือ กลุ่มของการจำแนกแบบเบย์ (Bayes) กลุ่มของการจำแนกแบบฟังก์ชัน (functions : f(x)) กลุ่มการจำแนกแบบขี้เกียจ (lazy) กลุ่มการจำแนกแบบหลายระเบียน (multiple-instance: MI) กลุ่มการจำแนกแบบหลักเกณฑ์เชื่อมโยง (rules) กลุ่มการจำแนกโดยใช้ต้นไม้การตัดสินใจ (trees) กลุ่มการจำแนกแบบอื่นๆ (misc) เมื่อคลิกปุ่มเพื่อเลือกกลุ่มของวิธีการจำแนกประเภทจะปรากฏหน้าต่างวิธีการจำแนกประเภทที่สามารถเรียกใช้ได้ในโปรแกรม
- ส่วนปรับพารามิเตอร์ของวิธีการจำแนกประเภทเป็นส่วนที่ใช้สำหรับแสดงปรับแต่งพารามิเตอร์สำหรับวิธีการจำแนกประเภทที่ถูกเลือก
- ส่วนแสดงผลลัพธ์ที่ได้จากขั้นตอนการวิเคราะห์การเกาะกลุ่มช่วยให้ผู้วิเคราะห์สามารถเรียกย้อนดูผลลัพธ์ของข้อมูลจากขั้นตอนการวิเคราะห์การเกาะกลุ่มได้

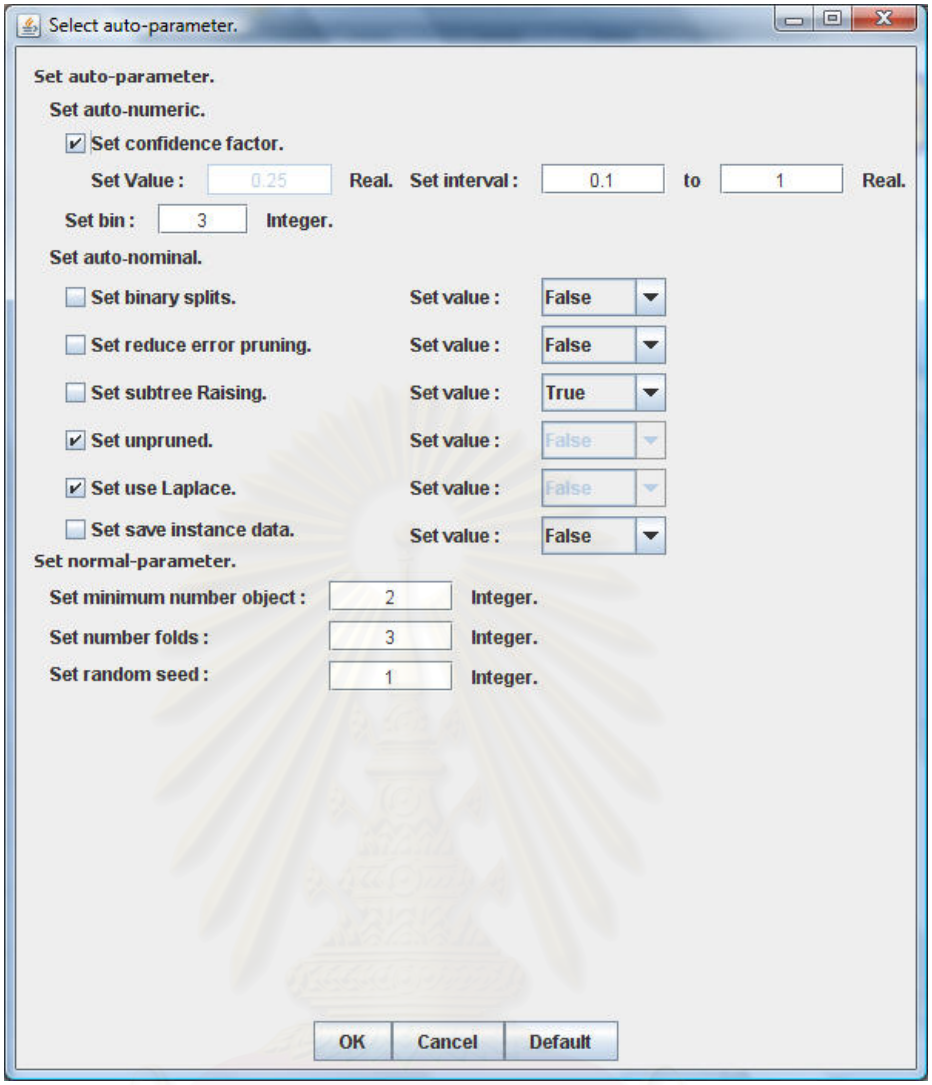
- ส่วนชื่อของวิธีการจำแนกประเภทเป็นส่วนที่แสดงลำดับของวิธีการจำแนกประเภท ชื่อกลุ่มของวิธีการจำแนกประเภท และชื่อวิธีการจำแนกประเภท

- ส่วนการกำหนดค่าพารามิเตอร์สำหรับการจำแนกประเภท ในส่วนนี้จะประกอบด้วย 2 ส่วนย่อย คือ ส่วนที่แสดงค่าพารามิเตอร์สำหรับวิธีการจำแนกประเภทที่ได้เลือกไว้ ส่วนที่ 2 คือปุ่ม “Adjust parameter” สำหรับเปลี่ยนแปลงค่าพารามิเตอร์ เมื่อคลิกที่ปุ่มนี้จะแสดงหน้าต่างที่ใช้กำหนดค่าพารามิเตอร์ ซึ่งได้มาจากคลาสของเวกา

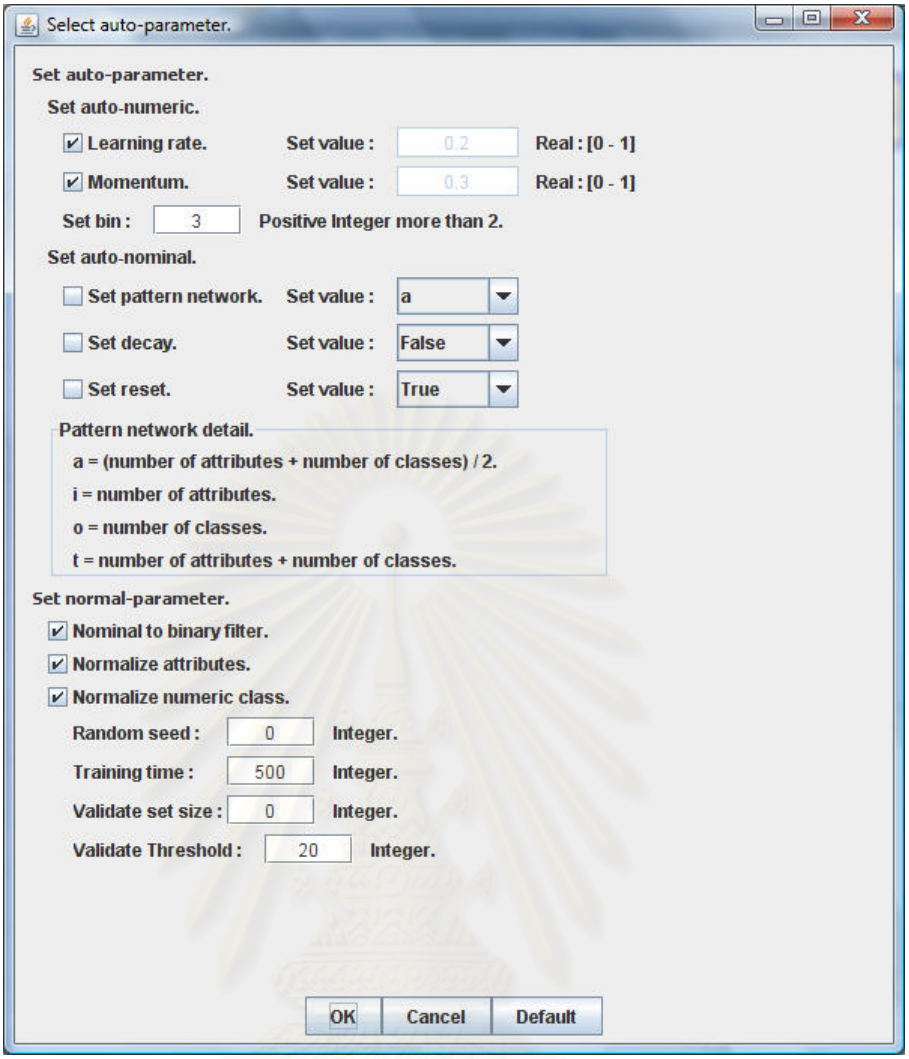
- ส่วนกำหนดวิธีการพัฒนาตัวแบบ สำหรับวิธีการจำแนกประเภทแยกออกเป็น 2 ส่วนคือ ส่วนแรกคือการกำหนดวิธีการพิจารณาพารามิเตอร์สามารถพิจารณาได้ 2 แบบ คือ การพิจารณาพารามิเตอร์แบบกำหนดเอง (fixed parameter) การกำหนดพารามิเตอร์จะใช้คลาสที่มีอยู่แล้วในเวกา กับการพิจารณาหาพารามิเตอร์แบบอัตโนมัติ (auto parameter) การกำหนดพารามิเตอร์จะใช้หน้าต่างที่พัฒนาขึ้นมา เพื่อกำหนดพารามิเตอร์อัตโนมัติ โดยหน้าต่างที่พัฒนาขึ้นจะอ้างอิงพารามิเตอร์จากคลาสในเวกา เพียงแต่เพิ่มช่วงกำหนดขอบเขต และพารามิเตอร์เสริมเพื่อการคำนวณพารามิเตอร์แบบอัตโนมัติ

- ส่วนแสดงผลลัพธ์ที่ได้จากวิธีการจำแนกประเภทใช้สำหรับกำหนดการแสดงหน้าต่างของผลลัพธ์ที่ได้จากวิธีการจำแนกประเภท โดยปกติแล้วจะไม่แสดงผลลัพธ์

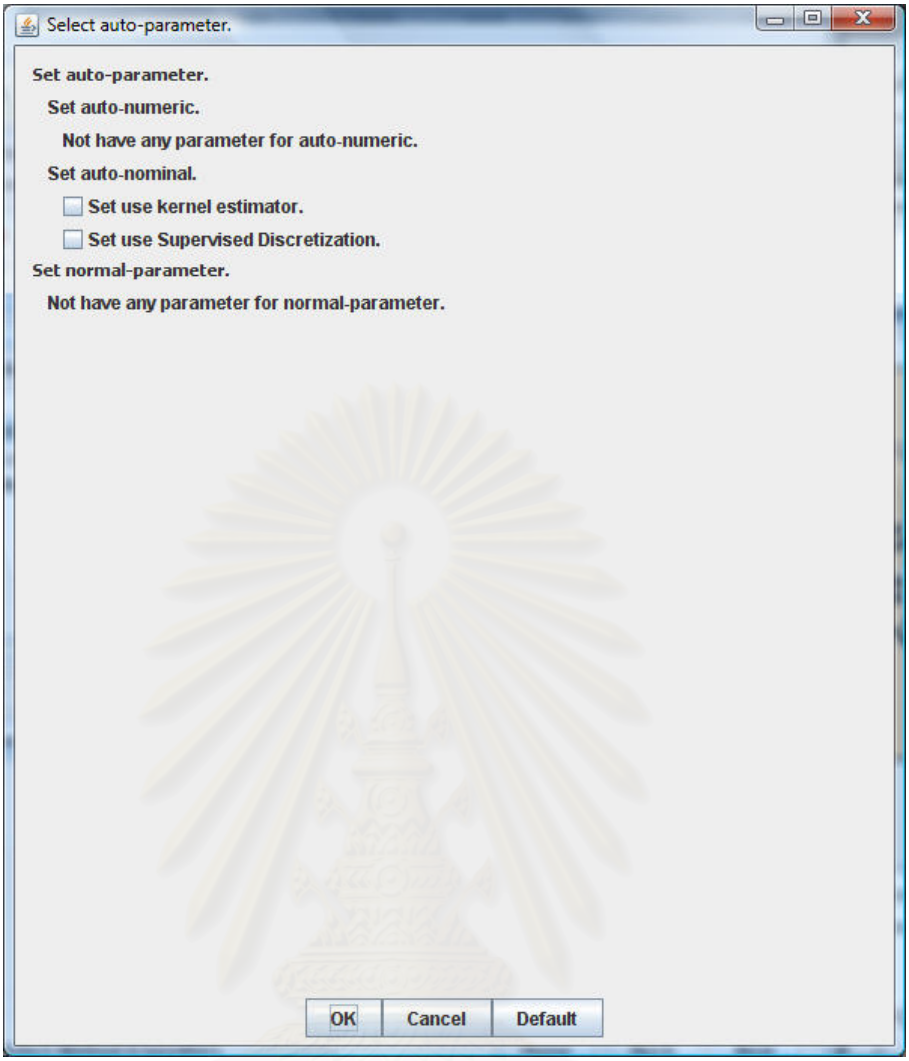
ในส่วนของการเชื่อมต่อมี 3 ปุ่ม คือ ปุ่มแรกปุ่ม “Home” ใช้สำหรับไปหน้าต่างแรก ปุ่มที่ 2 ปุ่ม “Back” ใช้สำหรับกลับไปหน้าต่างของการวิเคราะห์การเกาะกลุ่ม และปุ่มสุดท้ายปุ่ม “Next” เป็นปุ่มที่จะเชื่อมต่อไปยังหน้าต่างสำหรับรับข้อมูลทดสอบ และการแสดงการทดสอบตัวแบบกับข้อมูลทดสอบ ในกรณีที่ผู้วิเคราะห์เลือกการแสดงผลลัพธ์ โปรแกรมจะแสดงผลลัพธ์จากกระบวนการทำงานในการพัฒนาตัวแบบจำแนกประเภท และทดสอบตัวแบบจำแนกประเภทกับข้อมูลพัฒนาตัวแบบ และข้อมูลประเมิน หลังการกดปุ่ม “Next”



รูปที่ 3.23 : แสดงการกำหนดพารามิเตอร์อัตโนมัติ
ของวิธีต้นไม้การตัดสินใจ (J48)



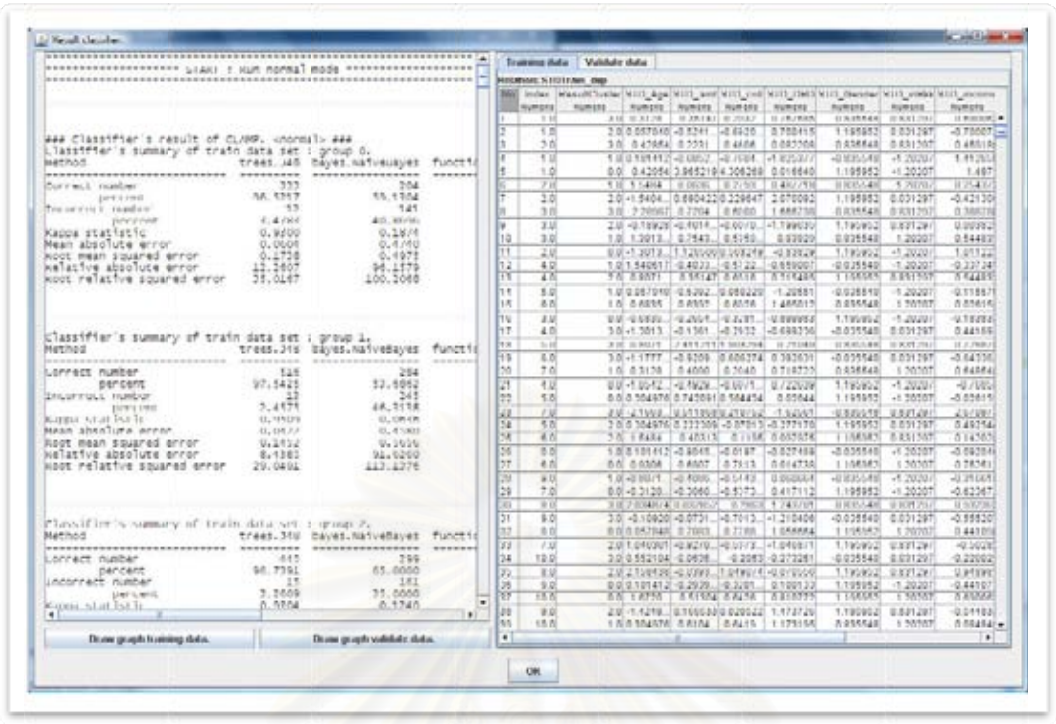
รูปที่ 3.24 : แสดงการกำหนดพารามิเตอร์อัตโนมัติ
ของวิธีเพอร์เซพตรอนหลายชั้น (multi-layer perceptron)



รูปที่ 3.25 : แสดงการกำหนดพารามิเตอร์อัตโนมัติ
ของวิธีการจำแนกแบบเบย์อย่างง่าย (naive Bayes)

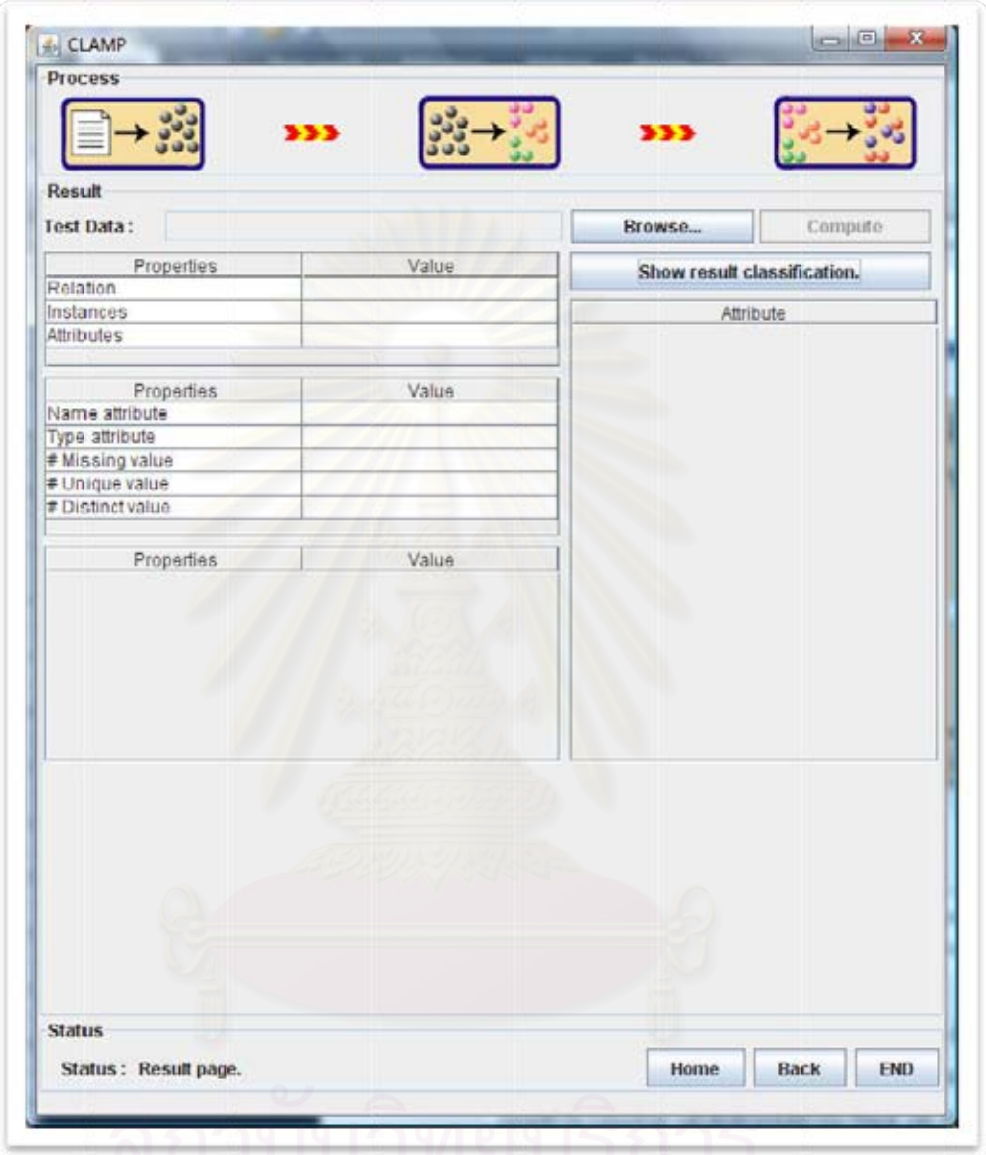
ผลลัพธ์ที่ได้จากการประมวลผลของตัวแบบจำแนกประเภทประกอบด้วย 3 ส่วน คือ

- ส่วนแสดงค่าสถิติที่ได้จากตัวแบบจำแนกประเภท
- ส่วนแสดงข้อมูลที่ได้จากขั้นตอนการวิเคราะห์การเกาะกลุ่ม แต่เพิ่มผลลัพธ์ตัวแบบจำแนกประเภทสำหรับข้อมูลแต่ละชุดข้อมูล
- ส่วนแสดงกราฟระหว่าง 2 ลักษณะประจำ



รูปที่ 3.26 : หน้าต่างผลลัพธ์ที่ได้จากการประมวลผลของตัวแบบจำแนกประเภท

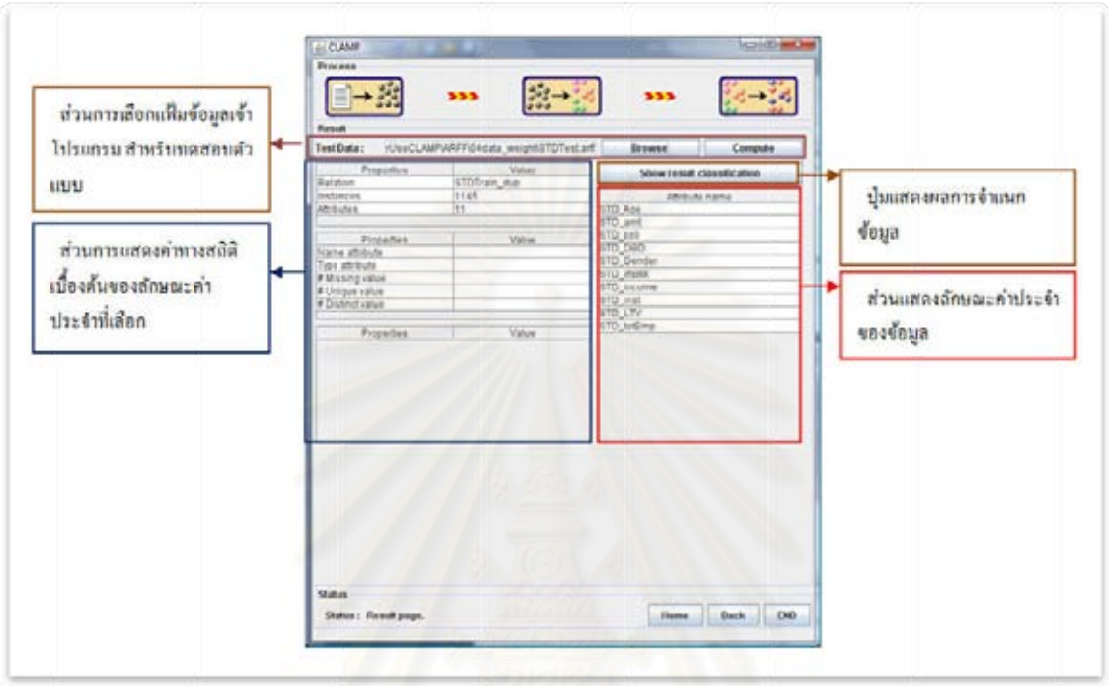
5. หน้าต่างสำหรับรับข้อมูลทดสอบ และประมวลผลตัวแบบกับข้อมูลทดสอบ (โมดูลของการทดสอบตัวแบบกับข้อมูลทดสอบตัวแบบ)



รูปที่ 3.27 : หน้าต่างสำหรับรับข้อมูลทดสอบ
และประมวลผลตัวแบบกับข้อมูลทดสอบ

ในส่วนของกระบวนการแสดงกระบวนการพัฒนาตัวแบบที่ผ่านครบทุกขั้นตอนของ CLAMP

ส่วนประกอบหน้าต่างสำหรับข้อมูลทดสอบและประมวลผลตัวแบบกับข้อมูลทดสอบ คือ

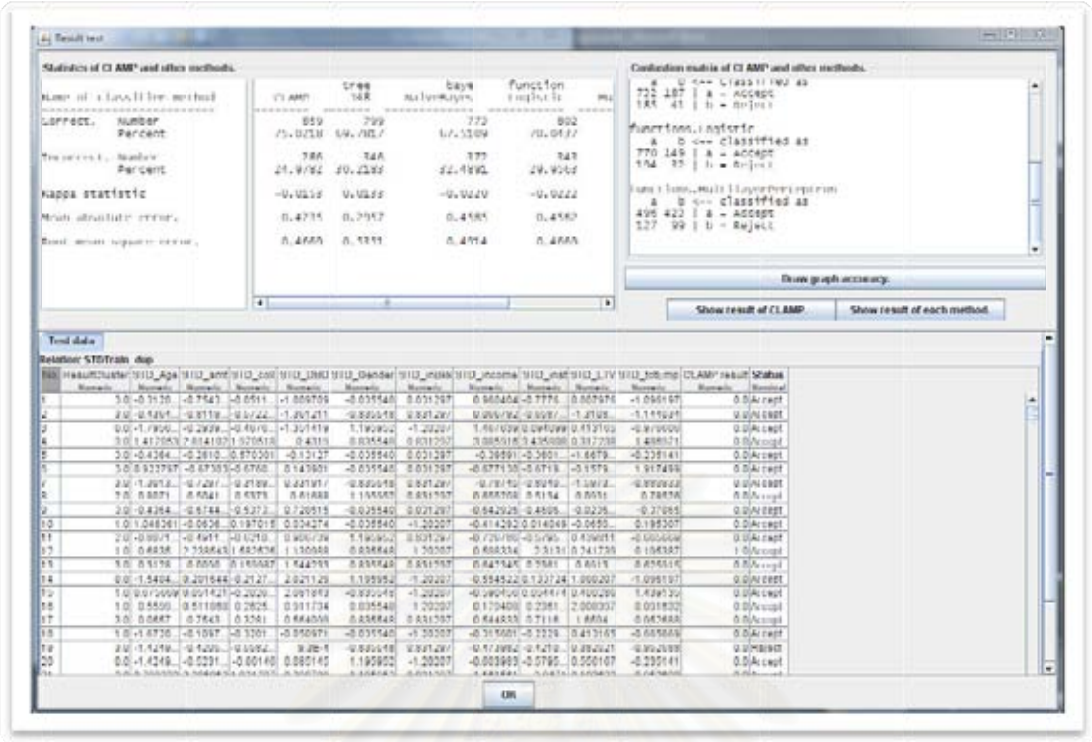


รูปที่ 3.28 : ส่วนประกอบในหน้าต่างสำหรับรับข้อมูลทดสอบ
และประมวลผลตัวแบบกับข้อมูลทดสอบ

- ส่วนการเลือกแฟ้มข้อมูลเข้าโปรแกรม สำหรับทดสอบตัวแบบ
- ส่วนการแสดงค่าสถิติเบื้องต้นของลักษณะประจำที่เลือกจากข้อมูลทดสอบ
- ปุ่มแสดงผลการจำแนกประเภท
- ส่วนแสดงลักษณะประจำของข้อมูล

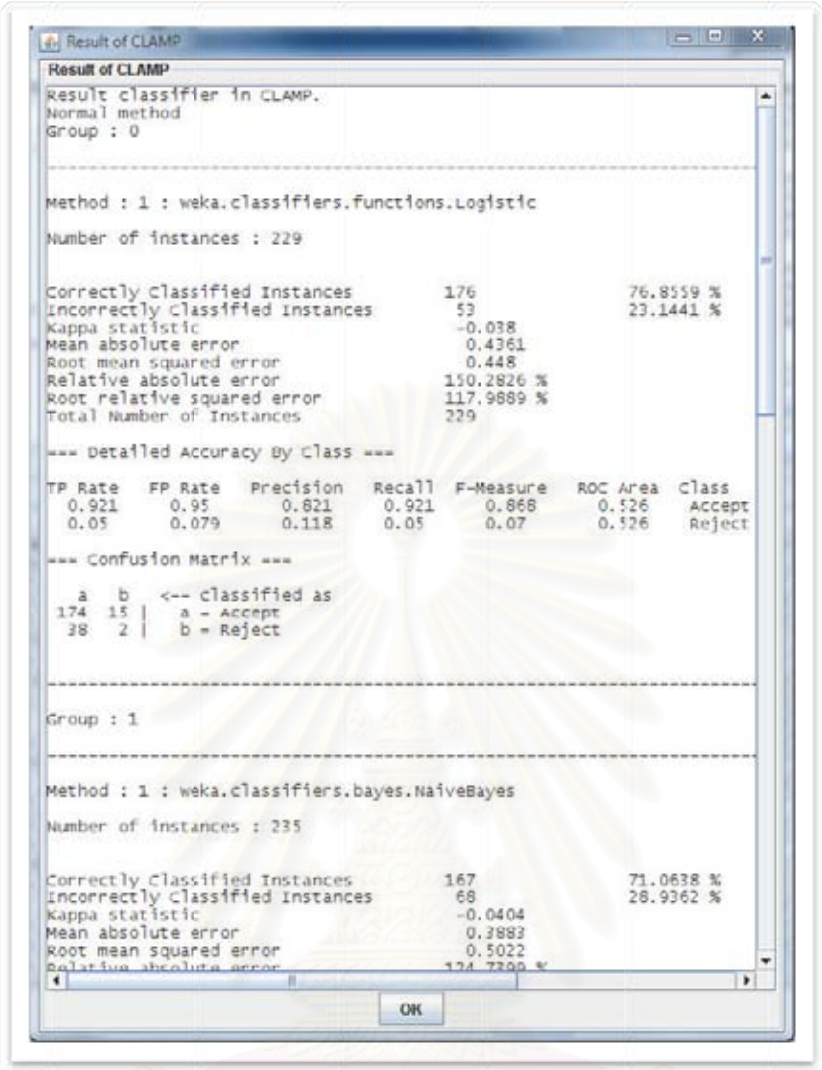
ในส่วนเชื่อมต่อมี 3 ปุ่ม คือ ปุ่มแรกปุ่ม “Home” ใช้สำหรับไปหน้าต่างแรก ปุ่มที่ 2 ปุ่ม “Back” ใช้สำหรับกลับไปหน้าต่างของการจำแนกประเภท และปุ่มสุดท้ายปุ่ม “END” เป็นปุ่มจบการทำงานของโปรแกรม

หน้าต่างแสดงผลลัพธ์ที่ได้จากการประมวลผลของขั้นตอนการทดสอบตัวแบบประกอบด้วย 4 ส่วน คือ

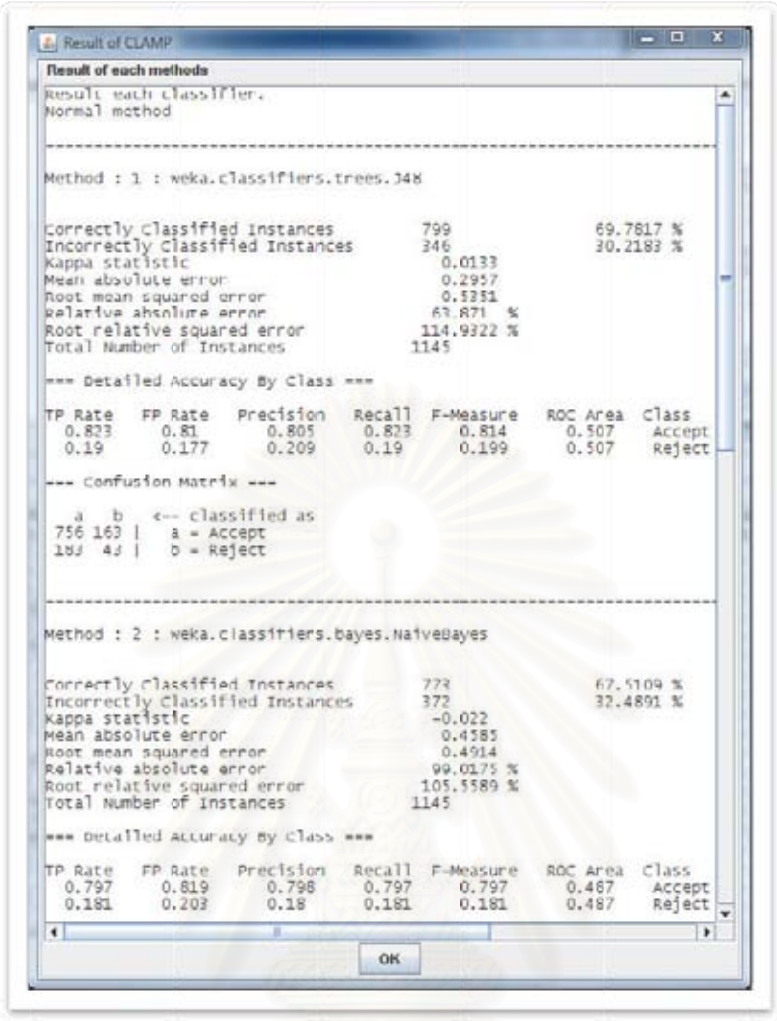


รูปที่ 3.29 : หน้าต่างแสดงผลลัพธ์ที่ได้จากการประมวลผลของขั้นตอนการทดสอบตัวแบบ

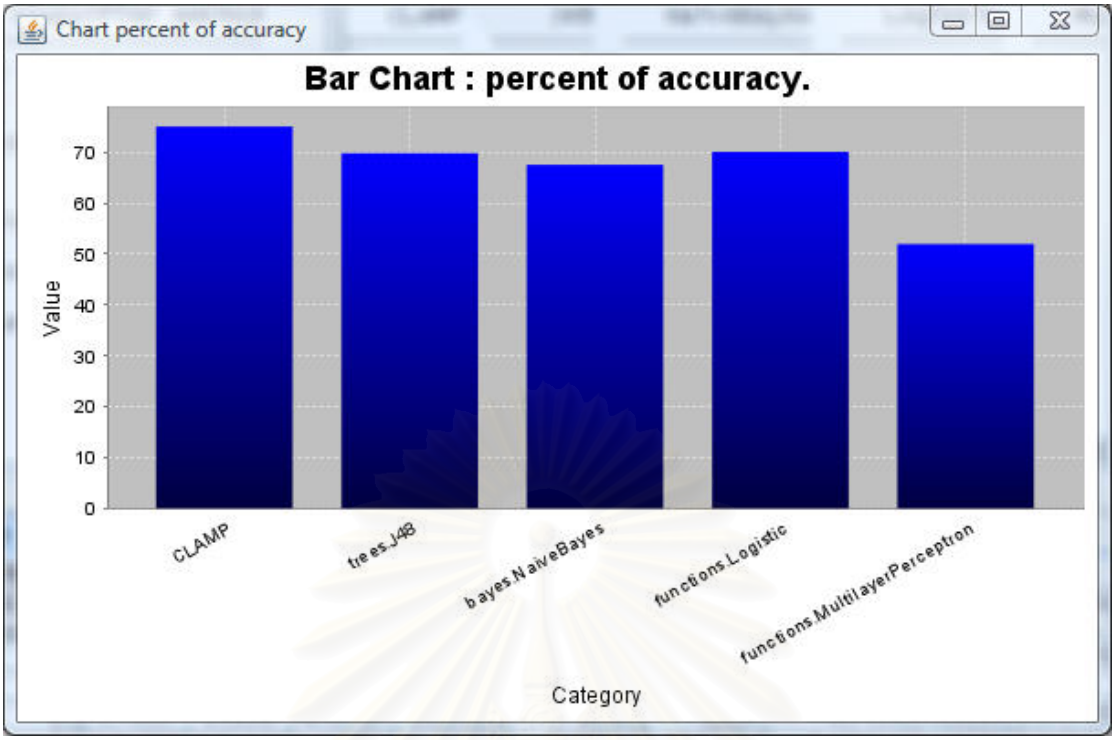
- ส่วนแสดงค่าสถิติที่ได้จากการทดสอบ ส่วนนี้เป็นการเปรียบเทียบระหว่างการใช้ CLAMP กับ การใช้วิธีการพัฒนาตัวแบบจากวิธีการจำแนกประเภทเพียงอย่างเดียว
- ส่วนแสดงตารางความสับสน (Confusion table) ของทุกตัวแบบ
- ส่วนกราฟของค่าความแม่นยำของตัวแบบจำแนกประเภทแต่ละวิธี
- ส่วนแสดงข้อมูลที่ได้จากขั้นตอนการวิเคราะห์การเกาะกลุ่ม แต่เพิ่มผลลัพธ์การจำแนกประเภทสำหรับข้อมูลแต่ละชุดข้อมูล ซึ่งจะมีปุ่มเชื่อมโยงไปยังผลลัพธ์ของวิธีการที่ถูกเลือกสำหรับข้อมูลแต่ละกลุ่ม และปุ่มเชื่อมโยงไปยังผลลัพธ์ที่ได้จากการใช้ตัวแบบจำแนกประเภทเพียงอย่างเดียว



รูปที่ 3.30 : หน้าต่างแสดงค่าสถิติของผลลัพธ์การวิเคราะห์การเกาะกลุ่มของการทำนายหลากหลายจากการประมวลผลของขั้นตอนการทดสอบตัวแบบ



รูปที่ 3.31 : หน้าต่างแสดงค่าสถิติของผลลัพธ์ที่ได้จากตัวแบบจำแนกประเภทแต่ละวิธีกับข้อมูลทดสอบ



รูปที่ 3.32 : หน้าต่างกราฟแสดงค่าความแม่นยำของตัวแบบจำแนกประเภทแต่ละวิธี

บทที่ 4
ผลลัพธ์จากการทดสอบการวิเคราะห์การเกาะกลุ่มของตัวทำนายหลากหลาย
(The experimental result of CLAMP)

ในบทนี้จะกล่าวถึงวิธีการทดลองการเปรียบเทียบผลการทดลองระหว่างวิธีการสร้างตัวแบบโดย CLAMP กับวิธีการสร้างตัวแบบโดยใช้วิธีการจำแนกประเภทแต่ละวิธี ซึ่งข้อมูลที่ใช้ในการทดลองเป็นข้อมูลการพิจารณาสินเชื่อของธนาคารแห่งหนึ่งในประเทศไทย และข้อมูลมาตราฐานกลางจากฐานข้อมูล UCI Machine Learning Repository [27] โดยแสดงผลลัพธ์ที่ได้ในรูปตารางเพื่อเปรียบเทียบประสิทธิภาพของแต่ละวิธี

4.1 วิธีการทดลอง

การทดลองเพื่อเปรียบเทียบประสิทธิภาพของโปรแกรม CLAMP กับตัวแบบจำแนกประเภทแต่ละวิธีจะใช้ข้อมูลการพิจารณาสินเชื่อของธนาคาร และข้อมูลที่นำมาทดสอบเพิ่มเติมซึ่งเป็นข้อมูลมาตราฐานกลางจาก UCI โดยแบ่งออกเป็นข้อมูลจากการพิจารณาสินเชื่อ และข้อมูลอื่นที่ไม่เกี่ยวกับสินเชื่อ ข้อมูลจากการพิจารณาสินเชื่อที่นำมาทดสอบ คือ ข้อมูลการพิจารณาสินเชื่อของธนาคาร ข้อมูลการพิจารณาสินเชื่อจากประเทศออสเตรเลีย (Australian credit approval) [28] และ ข้อมูลการพิจารณาสินเชื่อจากประเทศเยอรมัน (German credit data) [29] ดังตารางที่ 4.1 ข้อมูลอื่นที่ไม่เกี่ยวกับสินเชื่อที่นำมาทดสอบ คือ ข้อมูลการประเมินคุณภาพรถยนต์ (Car Evaluation Data Set) [30] และ ข้อมูล spambase [31] ดังตารางที่ 4.2 ซึ่งข้อมูลทั้งหมดจะต้องผ่านขั้นตอนการเตรียมข้อมูลเบื้องต้นก่อนนำมาทดลอง

ตารางที่ 4.1 : ข้อมูลสรุปบางส่วนของข้อมูลการพิจารณาสินเชื่อ

ชื่อของข้อมูล	จำนวนระเบียน	จำนวนลักษณะประจำต่อเนื่อง	จำนวนลักษณะประจำไม่ต่อเนื่อง	จำนวนลูกค้าดี	จำนวนลูกค้าไม่ดี
ข้อมูลการพิจารณาสินเชื่อของธนาคาร	3,891	8	2	3,152	739
ข้อมูลการพิจารณาสินเชื่อจากประเทศออสเตรเลีย	690	6	8	307	383
ข้อมูลการพิจารณาสินเชื่อจากประเทศเยอรมัน	1,000	7	13	700	300

ตารางที่ 4.2 : ข้อมูลสรุปบางส่วนของข้อมูลอื่น

ชื่อของข้อมูล	จำนวนระเบียน	จำนวนลักษณะประจำต่อเนื่อง	จำนวนลักษณะประจำไม่ต่อเนื่อง	จำนวนคลาส
ข้อมูลการประเมินคุณภาพรถยนต์	1,728	0	6	4
ข้อมูล spambase	4,601	57	0	2

1. **ข้อมูลที่เกี่ยวข้องกับการให้สินเชื่อ**

ข้อมูลชุดนี้เป็นข้อมูลของการจำแนกลูกค้าที่ขอรับสินเชื่อจากข้อมูลของลูกค้าเพื่อพิจารณาความเสี่ยงของลูกค้า คือลูกค้าดี และลูกค้าไม่ดี

1.1. **ข้อมูลการพิจารณาสินเชื่อจากธนาคาร**

ข้อมูลที่นำมาใช้ทดสอบกับ CLAMP เป็นข้อมูลการพิจารณาสินเชื่อที่เก็บอยู่ในช่วงเดือนธันวาคม 2548 ถึงเดือนมิถุนายน 2549

แหล่งข้อมูล

ธนาคารแห่งหนึ่งในประเทศไทย

จำนวนของลักษณะประจำ

สำหรับข้อมูลตั้งต้นมีลักษณะประจำทั้งหมดจำนวน 10 ลักษณะประจำ

- ลักษณะประจำแบบค่าต่อเนื่องจำนวน 8 ลักษณะประจำ
- ลักษณะประจำแบบค่าไม่ต่อเนื่องจำนวน 2 ลักษณะประจำ

คำอธิบายลักษณะประจำ

ลักษณะประจำที่ 1 เป็นลักษณะประจำอายุมีค่าต่อเนื่อง

ลักษณะประจำที่ 2 เป็นลักษณะประจำบริเวณพื้นที่ที่อาศัยอยู่ มีค่าไม่ต่อเนื่อง ค่าที่เป็นไปได้คือ

1 คืออาศัยในกรุงเทพฯและเขตปริมณฑล

0 คือไม่อาศัยในกรุงเทพฯและเขตปริมณฑล

ลักษณะประจำที่ 3 เป็นลักษณะประจำอายุการทำงาน เป็นลักษณะประจำแบบต่อเนื่อง

ลักษณะประจำที่ 4 เป็นลักษณะประจำเพศมีค่าไม่ต่อเนื่อง

ค่าที่เป็นไปได้คือ

1 คือ เพศชาย

0 คือ เพศหญิง

ลักษณะประจำที่ 5 เป็นลักษณะประจำจำนวนเงินที่ขอสินเชื่อมีค่าต่อเนื่อง

ลักษณะประจำที่ 6 เป็นลักษณะประจำค่างวดที่ต้องชำระต่อเดือนมีค่าต่อเนื่อง

ลักษณะประจำที่ 7 เป็นลักษณะประจำมูลค่าหลักทรัพย์ที่นำมาค้ำประกันมีค่าต่อเนื่อง

ลักษณะประจำที่ 8 เป็นลักษณะประจำรายได้สุทธิต่อเดือนมีค่าต่อเนื่อง

ลักษณะประจำที่ 9 เป็นลักษณะประจำอัตราส่วนของค่างวดที่ต้องชำระต่อเดือนต่อรายได้สุทธิต่อเดือน (DBD) มีค่าต่อเนื่อง

ลักษณะประจำที่ 10 เป็นลักษณะประจำอัตราส่วนของจำนวนเงินที่ขอสินเชื่อต่อมูลค่าหลักทรัพย์ที่นำมาค้ำประกัน (LTV) มีค่าต่อเนื่อง

จำนวนของคลาส

ลูกค้าดีมีทั้งหมด 3,152 ระเบียน

ลูกค้าไม่ดีมีทั้งหมด 739 ระเบียน

1.2. **ข้อมูลการพิจารณาสินเชื่อจากประเทศออสเตรเลีย** (Australian Credit Approval)

แหล่งข้อมูล

เป็นข้อมูลที่ได้มาจาก Quinlan

จำนวนข้อมูล 690 ระเบียน

จำนวนของลักษณะประจำ

สำหรับข้อมูลตั้งต้นมีลักษณะประจำทั้งหมดจำนวน 14 ลักษณะประจำ

- ลักษณะประจำแบบค่าต่อเนื่องจำนวน 6 ลักษณะประจำ
- ลักษณะประจำแบบค่าไม่ต่อเนื่องจำนวน 8 ลักษณะประจำ

คำอธิบายลักษณะประจำ

ลักษณะประจำที่ 1 (A1) เป็นลักษณะประจำที่ไม่ต่อเนื่อง
ค่าที่เป็นไปได้คือ a, b

ลักษณะประจำที่ 2 (A2) เป็นลักษณะประจำที่ต่อเนื่อง

ลักษณะประจำที่ 3 (A3) เป็นลักษณะประจำที่ต่อเนื่อง

ลักษณะประจำที่ 4 (A4) เป็นลักษณะประจำที่ไม่ต่อเนื่อง
ค่าที่เป็นไปได้คือ p, g, gg

ลักษณะประจำที่ 5 (A5) เป็นลักษณะประจำที่ไม่ต่อเนื่อง
ค่าที่เป็นไปได้คือ ff, d, i, k, j, aa, m, c, w, e, q, r, cc, x

ลักษณะประจำที่ 6 (A6) เป็นลักษณะประจำที่ไม่ต่อเนื่อง
ค่าที่เป็นไปได้คือ ff, dd, j, bb, v, n, o, h, z

ลักษณะประจำที่ 7 (A7) เป็นลักษณะประจำที่ต่อเนื่อง

ลักษณะประจำที่ 8 (A8) เป็นลักษณะประจำที่ไม่ต่อเนื่อง
ค่าที่เป็นไปได้คือ t, f

ลักษณะประจำที่ 9 (A9) เป็นลักษณะประจำที่ไม่ต่อเนื่อง
ค่าที่เป็นไปได้คือ t, f

ลักษณะประจำที่ 10 (A10) เป็นลักษณะประจำที่ต่อเนื่อง

ลักษณะประจำที่ 11 (A11) เป็นลักษณะประจำที่ไม่ต่อเนื่อง
ค่าที่เป็นไปได้คือ t, f

ลักษณะประจำที่ 12 (A12) เป็นลักษณะประจำที่ไม่ต่อเนื่อง
ค่าที่เป็นไปได้คือ s, g, p

ลักษณะประจำที่ 13 (A13) เป็นลักษณะประจำที่ต่อเนื่อง

ลักษณะประจำที่ 14 (A14) เป็นลักษณะประจำที่ต่อเนื่อง

ข้อมูลที่ขาดหาย 37 ระเบียน คิดเป็นประมาณ 5% ของข้อมูลทั้งหมด

ลักษณะประจำที่ 1 มีข้อมูลที่ขาดหายทั้งหมด 12 ระเบียน
ลักษณะประจำที่ 2 มีข้อมูลที่ขาดหายทั้งหมด 12 ระเบียน
ลักษณะประจำที่ 4 มีข้อมูลที่ขาดหายทั้งหมด 6 ระเบียน
ลักษณะประจำที่ 5 มีข้อมูลที่ขาดหายทั้งหมด 6 ระเบียน
ลักษณะประจำที่ 6 มีข้อมูลที่ขาดหายทั้งหมด 9 ระเบียน
ลักษณะประจำที่ 7 มีข้อมูลที่ขาดหายทั้งหมด 9 ระเบียน
ลักษณะประจำที่ 14 มีข้อมูลที่ขาดหายทั้งหมด 13 ระเบียน

โดยข้อมูลที่ขาดหายจะถูกแทนด้วยค่าฐานนิยมสำหรับลักษณะประจำที่ไม่ต่อเนื่อง และค่าเฉลี่ยสำหรับลักษณะประจำที่ต่อเนื่อง

จำนวนของคลาส

ลูกค้าดีมีทั้งหมด 307 ระเบียน
ลูกค้าไม่ดีมีทั้งหมด 383 ระเบียน

1.3. **ข้อมูลการพิจารณาสินเชื่อจากประเทศเยอรมัน** (German Credit Data)

แหล่งข้อมูล

Professor Dr. Hans Hofmann
Institute Statistic and Econometrics
University of Hamburg

จำนวนข้อมูล 1,000 ระเบียน

จำนวนของลักษณะประจำ

ข้อมูลมีลักษณะประจำทั้งหมดจำนวน 20 ลักษณะประจำ

- ลักษณะประจำแบบค่าต่อเนื่องจำนวน 7 ลักษณะประจำ
- ลักษณะประจำแบบค่าไม่ต่อเนื่องจำนวน 13 ลักษณะประจำ

คำอธิบายลักษณะประจำ

ลักษณะประจำที่ 1 (attribute 1) เป็นลักษณะประจำที่มีค่าไม่ต่อเนื่อง

ความหมายของลักษณะประจำ ข้อมูลสถานะการใช้งานของบัญชีลูกค้า

A11 x < 0 DM

A12 0 <= x < 200 DM

A13 x >= 200 DM

A14 ไม่มีการตรวจสอบบัญชี

เมื่อ x คือ เงินเดือนของลูกค้า

ลักษณะประจำที่ 2 (attribute 2) เป็นลักษณะประจำที่มีค่าต่อเนื่อง

ความหมายของลักษณะประจำ ระยะเวลาการใช้งาน

ลักษณะประจำที่ 3 (attribute 3) เป็นลักษณะประจำที่มีค่าไม่ต่อเนื่อง

ความหมายของลักษณะประจำ ประวัติการขอสินเชื่อ

A30 ไม่มีการขอสินเชื่อ หรือ ไม่มีการค้างค่างวด

A31 สินเชื่อที่ขอจากธนาคารได้ชำระครบ

A32 สินเชื่อที่ขอจากธนาคารยังชำระไม่หมด

A33 สินเชื่อที่ขอจากธนาคารมีการค้างชำระ

A34 ลูกค้ามีความเสี่ยงสูงหรือมีการใช้บริการสินเชื่อเงินนอกระบบที่ไม่ใช่ธนาคาร

ลักษณะประจำที่ 4 (attribute 4) เป็นลักษณะประจำที่มีค่าไม่ต่อเนื่อง

ความหมายของลักษณะประจำ จุดประสงค์ของการขอสินเชื่อ

A40 เพื่อซื้อรถใหม่

A41 เพื่อซื้อรถมือสอง

A42 เพื่อซื้อเฟอร์นิเจอร์ หรือ อุปกรณ์

A43 เพื่อซื้อวิทยุ หรือ โทรทัศน์

A44 เพื่อใช้สำหรับซื้อบ้าน

A45 เพื่อใช้สำหรับการซ่อมแซมบ้าน

A46 เพื่อใช้สำหรับการศึกษา

A47 เพื่อใช้สำหรับการท่องเที่ยว

A48 เพื่อใช้สำหรับการฝึกอบรม

A49 เพื่อใช้สำหรับธุรกิจ

A410 อื่นๆ

ลักษณะประจำที่ 5 (attribute 5) เป็นลักษณะประจำที่มีค่าต่อเนื่อง

ความหมายของลักษณะประจำ จำนวนสินเชื่อที่ขอ

ลักษณะประจำที่ 6 (attribute 6) เป็นลักษณะประจำที่มีค่าไม่ต่อเนื่อง

ความหมายของลักษณะประจำ จำนวนเงินเก็บ และมูลค่าของใบหุ้นกู้ที่ลูกค้ามีอยู่

A61 x < 100 DM

A62 100 <= x < 500 DM

A63 500 <= x < 1000 DM

A64 x >= 1000 DM

A65 ไม่ระบุจำนวน หรือ ไม่มีเงินเก็บ

เมื่อ x คือ จำนวนเงินเก็บ และมูลค่าของใบหุ้นกู้ที่ลูกค้ามีอยู่

ลักษณะประจำที่ 7 (attribute 7) เป็นลักษณะประจำที่มีค่าไม่ต่อเนื่อง

ความหมายของลักษณะประจำ อายุการทำงานของลูกค้า

A71 ไม่ได้ทำงาน

A72 x < 1

A73 1 <= x < 4

A74 4 <= x < 7

A75 x >= 7

เมื่อ x คือ อายุการทำงานมีหน่วย เป็น ปี (years)

ลักษณะประจำที่ 8 (attribute 8) เป็นลักษณะประจำที่มีค่าต่อเนื่อง

ความหมายของลักษณะประจำ อัตราส่วนของการผ่อนชำระคืน ในรูปของร้อยละของการจัดการรายได้

ลักษณะประจำที่ 9 (attribute 9) เป็นลักษณะประจำที่มีค่าไม่ต่อเนื่อง

ความหมายของลักษณะประจำ เพศ และสถานะของลูกค้า

A91 ชาย หย่า / แยกกันอยู่

A92 หญิง หย่า / แยกกันอยู่ / สมรส

A93 ชาย โสด

A94 ชาย สมรส / หม่าย

A95 หญิง โสด

ลักษณะประจำที่ 10 (attribute 10) เป็นลักษณะประจำที่มีค่าไม่ต่อเนื่อง

ความหมายของลักษณะประจำ ผู้กู้ร่วมหรือผู้ค้ำประกัน

A101 ไม่มี

A102 มีผู้กู้ร่วม

A103 มีผู้ค้ำประกัน

ลักษณะประจำที่ 11 (attribute 11) เป็นลักษณะประจำที่มีค่าต่อเนื่อง

ความหมายของลักษณะประจำ จำนวนที่อยู่อาศัยในปัจจุบัน

ลักษณะประจำที่ 12 (attribute 12) เป็นลักษณะประจำที่มีค่าไม่ต่อเนื่อง

ความหมายของลักษณะประจำ สินทรัพย์ของลูกค้า

A121 สินทรัพย์ที่ดิน

A122 ถ้าไม่มีสินทรัพย์ที่ดินพิจารณาสินทรัพย์ของสิ่งก่อสร้างที่มีค่า หรือ ประกันชีวิต

A123 ถ้าไม่มีสินทรัพย์ที่ดินหรือสินทรัพย์ของสิ่งก่อสร้างที่มีค่า หรือ ประกันชีวิตพิจารณาสินทรัพย์ของรถ หรือ อื่นๆ

A124 ไม่ระบุหรือไม่มีสินทรัพย์

ลักษณะประจำที่ 13 (attribute 13) เป็นลักษณะประจำที่มีค่าต่อเนื่อง

ความหมายของลักษณะประจำ อายุของลูกค้ามีหน่วยเป็นปี

ลักษณะประจำที่ 14 (attribute 14) เป็นลักษณะประจำที่มีค่าไม่ต่อเนื่อง

ความหมายของลักษณะประจำ แผนการชำระ

A141 ธนาคาร

A142 เงินเก็บ

A143 ไม่มี

ลักษณะประจำที่ 15 (attribute 15) เป็นลักษณะประจำที่มีค่าไม่ต่อเนื่อง

ความหมายของลักษณะประจำ ลักษณะที่อยู่อาศัยของลูกค้า

A151 เช่า

A152 เป็นเจ้าของ

A153 มรดก

ลักษณะประจำที่ 16 (attribute 16) เป็นลักษณะประจำที่มีค่าต่อเนื่อง

ความหมายของลักษณะประจำ จำนวนของการสินเชื่อที่ยังชำระธนาคารไม่ครบ

ลักษณะประจำที่ 17 (attribute 17) เป็นลักษณะประจำที่มีค่าไม่ต่อเนื่อง

ความหมายของลักษณะประจำ ลักษณะงานของลูกค้า

A171 ไม่ได้ทำงาน/ ไม่มีความเชี่ยวชาญ และไม่มีที่อยู่อาศัย

A172 ไม่มีความเชี่ยวชาญ และมีที่อยู่อาศัย

A173 มีความเชี่ยวชาญงานทั่วไป/เจ้าหน้าที่

A174 มีความเชี่ยวชาญการจัดการ/พนักงานคุณภาพสูง

ลักษณะประจำที่ 18 (attribute 18) เป็นลักษณะประจำที่มีค่าต่อเนื่อง

ความหมายของลักษณะประจำ จำนวนลูกค้าที่รับผิดชอบสินเชื่อ

ลักษณะประจำที่ 19 (attribute 19) เป็นลักษณะประจำที่มีค่าไม่ต่อเนื่อง

ความหมายของลักษณะประจำ การมีโทรศัพท์

A191 ไม่มี

A192 มี

ลักษณะประจำที่ 20 (attribute 20)

ความหมายของลักษณะประจำ การเป็นแรงงานต่างชาติ

A201 ลูกค้าเป็นแรงงานต่างชาติ

A202 ลูกค้าไม่เป็นแรงงานต่างชาติ

จำนวนของคลาส

ลูกค้าดีมีทั้งหมด 700 ระเบียน

ลูกค้าไม่ดีมีทั้งหมด 300 ระเบียน

2. ข้อมูลอื่น

2.1. ข้อมูลการประเมินคุณภาพรถยนต์

แหล่งข้อมูล

สร้างโดย Marko Bohanec

บริจาคโดย Marko Bohanec และ Blaz Zupan

เมื่อเดือนมิถุนายน ปี 1997

จำนวนข้อมูล 1,728 ระเบียน

จำนวนของลักษณะประจำ

ลักษณะประจำแบบค่าไม่ต่อเนื่องจำนวน 6 ลักษณะประจำ

คำอธิบายลักษณะประจำ

ลักษณะประจำที่ 1 ราคาที่ซื้อรถ (buying) เป็นลักษณะประจำแบบไม่ต่อเนื่อง
ค่าที่เป็นไปได้คือ v-high, high, med, low

ลักษณะประจำที่ 2 ราคาที่ซื้อรถ (buying) เป็นลักษณะประจำแบบไม่ต่อเนื่อง
ค่าที่เป็นไปได้คือ v-high, high, med, low

ลักษณะประจำที่ 3 จำนวนประตูของรถ (doors) เป็นลักษณะประจำแบบไม่ต่อเนื่อง
ค่าที่เป็นไปได้คือ two, three, four, more

ลักษณะประจำที่ 4 จำนวนความจุคนของรถ (persons) เป็นลักษณะประจำแบบไม่ต่อเนื่อง
ค่าที่เป็นไปได้คือ two, four, more

ลักษณะประจำที่ 5 เวลาที่ในการสตาร์ท (lug_boot) เป็นลักษณะประจำแบบไม่ต่อเนื่อง
ค่าที่เป็นไปได้คือ small, med, big

ลักษณะประจำที่ 6 ความปลอดภัย (safety) เป็นลักษณะประจำแบบไม่ต่อเนื่อง ค่าที่เป็นไปได้คือ low, med, high

จำนวนของคลาส

unacc มีทั้งหมด	1,210 ระเบียน
acc มีทั้งหมด	300 ระเบียน
good มีทั้งหมด	69 ระเบียน
v-good มีทั้งหมด	65 ระเบียน

2.2. **ข้อมูล** spambase

แหล่งข้อมูล

สร้างโดย Mark Hopkins, Erik Reeber, George Forman, Jaap Suermondt

บริจาคโดย George Forman

เมื่อเดือนมิถุนายน ปี 1999

จำนวนข้อมูล 4,601 ระเบียน

จำนวนของลักษณะประจำ

ลักษณะประจำแบบค่าต่อเนื่องจำนวน 57 ลักษณะประจำ

คำอธิบายลักษณะประจำ

ข้อมูลชุดนี้มีลักษณะประจำตามตาราง 4.3

ตารางที่ 4.3 : ตารางแสดงชื่อลักษณะประจำของข้อมูล spambase

word_freq_address	word_freq_font	word_freq_meeting
word_freq_all	word_freq_000	word_freq_original
word_freq_3d	word_freq_money	word_freq_project
word_freq_our	word_freq_hp	word_freq_re
word_freq_over	word_freq_hpl	word_freq_edu
word_freq_remove	word_freq_george	word_freq_table
word_freq_internet	word_freq_650	word_freq_conference
word_freq_order	word_freq_lab	char_freq_;
word_freq_mail	word_freq_labs	char_freq_(
word_freq_receive	word_freq_telnet	char_freq_[
word_freq_will	word_freq_857	char_freq_!
word_freq_people	word_freq_data	char_freq_$
word_freq_report	word_freq_415	char_freq_#
word_freq_addresses	word_freq_85	capital_run_length_average
word_freq_free	word_freq_technology	capital_run_length_longest
word_freq_business	word_freq_1999	capital_run_length_total
word_freq_email	word_freq_parts	
word_freq_you	word_freq_pm	
word_freq_credit	word_freq_direct	
word_freq_your	word_freq_cs	

ลักษณะประจำทั้ง 57 ลักษณะประจำมีค่าต่อเนื่อง

จำนวนของคลาส

ขยะไปรษณีย์อิเล็กทรอนิกส์มีทั้งหมด	1,813 ระเบียน
ไม่เป็นขยะไปรษณีย์อิเล็กทรอนิกส์มีทั้งหมด	2,788 ระเบียน

ตัววัดที่สนใจในการทดสอบวิธีการพัฒนาตัวแบบการให้คะแนนสินเชื่อโดยวิเคราะห์การเกาะกลุ่มของตัวทำนายหลากหลาย กับวิธีการพัฒนาตัวแบบการให้คะแนนสินเชื่อโดยวิธีการจำแนกแต่ละวิธี ในงานวิจัยนี้เลือกใช้ความแม่นยำในการทำนาย (accuracy)

การทดลองแบ่งข้อมูลออกเป็น 3 ส่วน ส่วนแรก คือ 40% ของข้อมูลเรียกว่าข้อมูลพัฒนาตัวแบบ (training data) ส่วนที่สอง คือ 30% ของข้อมูลจะเรียกว่าข้อมูลประเมิน (validate data) และส่วนที่สามคือ 30% ของข้อมูลจะเรียกว่า ข้อมูลทดสอบ (test data)

ขั้นตอนการทดลองของ CLAMP ใช้ข้อมูลพัฒนาตัวแบบมาพัฒนาตัวแบบกับการวิเคราะห์การเกาะกลุ่ม แล้วนำกลุ่มข้อมูลที่ได้แต่ละกลุ่มสร้างตัวแบบจำแนกประเภทที่เหมาะสมที่สุดจาก 4 รูปแบบ หลังจากนั้นทำการประเมินตัวแบบจำแนกประเภท โดยการเลือกตัวแบบจำแนกประเภทที่มีความถูกต้องในการทำนายดีที่สุดกับข้อมูลประเมินเรียกว่าตัวแบบตัวทำนายเหมาะสม ในขั้นสุดท้ายเมื่อได้ตัวแบบตัวทำนายเหมาะสมของแต่ละวิธีการแล้วนำ CLAMP ใช้กับข้อมูลทดสอบ เพื่อหาค่าความถูกต้องของ CLAMP

4.2 การเปรียบเทียบผลการทดลอง

การจำแนกประเภทแบบทวิภาค (binary classification) วัดประสิทธิภาพของตัวแบบการให้คะแนนสินเชื่อ

ตารางที่ 4.4 : การจำแนกประเภทแบบทวิภาค

		ข้อมูลจริง	
		ลูกค้าได้รับการอนุมัติ	ลูกค้าไม่ได้รับการอนุมัติ
ตัวแบบ	ลูกค้าได้รับการอนุมัติ	TP	FP
	ลูกค้าไม่ได้รับการอนุมัติ	FN	TN

สำหรับตัวแบบพิจารณาสินเชื่อสามารถอธิบายได้ดังนี้

- true positive คือ จำนวนของผลลัพธ์ที่ได้จากตัวแบบที่ทายว่าลูกค้าได้รับอนุมัติสินเชื่อ และผลลัพธ์จากเจ้าหน้าที่พิจารณาสินเชื่อคือลูกค้าได้รับการอนุมัติสินเชื่อ

- false positive คือ จำนวนของผลลัพธ์ที่ได้จากตัวแบบที่ทายว่าลูกค้าได้รับอนุมัติสินเชื่อ แต่ผลลัพธ์จากเจ้าหน้าที่พิจารณาสินเชื่อคือลูกค้าไม่ได้รับการอนุมัติสินเชื่อ
- false negative คือ จำนวนของผลลัพธ์ที่ได้จากตัวแบบที่ทายว่าลูกค้าไม่ได้รับอนุมัติสินเชื่อ แต่ผลลัพธ์จากเจ้าหน้าที่พิจารณาสินเชื่อคือลูกค้าได้รับการอนุมัติสินเชื่อ
- true negative คือ จำนวนของผลลัพธ์ที่ได้จากตัวแบบที่ทายว่าลูกค้าไม่ได้รับอนุมัติสินเชื่อ และผลลัพธ์จากเจ้าหน้าที่พิจารณาสินเชื่อคือลูกค้าไม่ได้รับการอนุมัติสินเชื่อ

ตัววัดที่ใช้ในการทดสอบ คือ ความถูกต้องในการทำนายซึ่งเป็นผลรวมของค่า true positive และ ค่า true negative

4.3 ผลการทดลอง

ในหัวข้อนี้แสดงผลการทดลองที่ได้จากการทดลองกับโปรแกรม CLAMP พร้อมประสิทธิภาพของตัวแบบทำนายของแต่ละวิธี

1. ข้อมูลการพิจารณาสินเชื่อจากธนาคารแห่งหนึ่งในประเทศไทย

ข้อมูลการพิจารณาสินเชื่อจากธนาคารแห่งหนึ่งในประเทศไทยมีจำนวนระเบียนทั้งหมด 3,891 ระเบียน แบ่งข้อมูลออกเป็น 3 ส่วน คือข้อมูลพัฒนาตัวแบบจำนวน 1,382 ระเบียน ข้อมูลประเมินจำนวน 1,364 ระเบียน และข้อมูลทดสอบจำนวน 1,145 ระเบียน โดยแสดงจำนวนข้อมูลที่อนุมัติ และไม่อนุมัติในตารางที่ 4.5

หลังจากใช้ CLAMP ได้ผลลัพธ์คือจำนวนกลุ่มทั้งหมด 4 กลุ่ม แต่ละกลุ่มมีจำนวนข้อมูลตามตารางที่ 4.6 ซึ่งตัวแบบจำแนกประเภทที่เหมาะสมแสดงในตารางที่ 4.12 โดยพิจารณาจากค่าความแม่นยำของตัวแบบจำแนกประเภทตามตารางที่ 4.7 ตารางที่ 4.8 ตารางที่ 4.9 และตารางที่ 4.10

ผลลัพธ์จากตัวแบบจำแนกประเภทเพียงอย่างเดียวที่ได้จากการทดสอบข้อมูลการพิจารณาสินเชื่อจากธนาคารแห่งหนึ่งในประเทศไทยแสดงรายละเอียดในตารางที่ 4.11

สรุปผลลัพธ์ที่ได้จากการทดสอบกับข้อมูลพิจารณาสินเชื่อจากธนาคารแห่งหนึ่งในประเทศไทย แสดงรายละเอียดในตารางที่ 4.13 พบว่าค่าความแม่นยำของ CLAMP ไม่แตกต่างจากตัวแบบจำแนกประเภทอื่น โดยวิธีการจำแนกประเภทที่ดีที่สุดคือการทำนายลูกค้าทั้งหมดเป็นลูกค้าดี โดยไม่สนใจลูกค้าไม่ดีที่ได้ผลดังนี้ก็เพราะข้อมูลการพิจารณาสินเชื่อจากธนาคารมีปัญหาขนาดของคลาสไม่สมดุล (class imbalance) หมายความว่าขนาดของคลาสที่ลูกค้าได้รับการอนุมัติมีมากกว่าขนาดของคลาสที่ลูกค้าไม่ได้รับการอนุมัติมาก

ตารางที่ 4.5 : ตารางแสดงจำนวนระเบียนหลังจากการแบ่งข้อมูลพิจารณาสินเชื่อจากธนาคารแห่งหนึ่งในประเทศไทย

	จำนวนระเบียน		
	Training data	Validate data	Test data
Accept	1,156	1,077	919
Reject	226	287	226
Total	1,382	1,364	1,145

ตารางที่ 4.6 : ตารางแสดงอัตราส่วนร้อยละของจำนวนระเบียนที่อยู่ในแต่ละกลุ่มที่ได้จากตัวแบบการวิเคราะห์การเกาะกลุ่มของข้อมูลการพิจารณาสินเชื่อจากธนาคารแห่งหนึ่งในประเทศไทย

	อัตราส่วนร้อยละ		
	Training data	Validate data	Test data
กลุ่มที่ 0	7	10	11
กลุ่มที่ 1	28	27	24
กลุ่มที่ 2	41	41	44
กลุ่มที่ 3	24	23	21

ตารางที่ 4.7 : ตารางแสดงผลลัพธ์จากตัวแบบจำแนกประเภทในข้อมูลกลุ่มที่ 0 กับข้อมูลการพิจารณาสินเชื่อจากธนาคารแห่งหนึ่งในประเทศไทย

Classifier's summary of train data set : group 0.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	79	81	79	85
percent	81.4433	83.5052	81.4433	87.6289
Incorrect number	18	16	18	12
percent	18.5567	16.4948	18.5567	12.3711
Kappa statistic	0.2977	0.4441	0.3746	0.6475
Mean absolute error	0.2889	0.2734	0.2916	0.2280
Root mean squared error	0.3800	0.3587	0.3794	0.3081
Relative absolute error	79.2326	74.9954	79.9794	62.5466
Root relative squared error	89.3486	84.3346	89.1956	72.4261

Classifier's summary of validate data set : group 0.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	96	90	92	81
percent	72.1805	67.6692	69.1729	60.9023
Incorrect number	37	43	41	52
percent	27.8195	32.3308	30.8271	39.0977
Kappa statistic	0.0372	-0.0067	-0.0214	-0.1166
Mean absolute error	0.3647	0.3936	0.3677	0.4323
Root mean squared error	0.4680	0.5146	0.4964	0.5381
Relative absolute error	93.6106	101.0291	94.3661	110.9531
Root relative squared error	103.1226	113.3970	109.3883	118.5750

ตารางที่ 4.8 : ตารางแสดงผลลัพธ์จากตัวแบบจำแนกประเภทในข้อมูลกลุ่มที่ 1 กับข้อมูลการพิจารณาสินเชื่อจากธนาคารแห่งหนึ่งในประเทศไทย

Classifier's summary of train data set : group 1.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	326	326	327	328
percent	84.2377	84.2377	84.4961	84.7545
Incorrect number	61	61	60	59
percent	15.7623	15.7623	15.5039	15.2455
Kappa statistic	0.0000	0.0000	0.0273	0.1125
Mean absolute error	0.2656	0.2782	0.2621	0.2155
Root mean squared error	0.3644	0.3666	0.3619	0.3406
Relative absolute error	99.5481	104.2763	98.2597	80.7801
Root relative squared error	99.9988	100.6037	99.3072	93.4738

Classifier's summary of validate data set : group 1.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	291	291	291	287
percent	80.3867	80.3867	80.3867	79.2818
Incorrect number	71	71	71	75
percent	19.6133	19.6133	19.6133	20.7182
Kappa statistic	0.0000	0.0000	0.0000	-0.0047
Mean absolute error	0.2919	0.3088	0.2947	0.2693
Root mean squared error	0.3989	0.4027	0.4026	0.4142
Relative absolute error	99.6349	105.4097	100.5968	91.9084
Root relative squared error	100.0416	100.9982	100.9640	103.8675

ตารางที่ 4.9 : ตารางแสดงผลลัพธ์จากตัวแบบจำแนกประเภทในข้อมูลกลุ่มที่ 2 กับข้อมูลการพิจารณาสินเชื่อจากธนาคารแห่งหนึ่งในประเทศไทย

Classifier's summary of train data set : group 2.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	488	479	488	491
percent	86.0670	84.4797	86.0670	86.5961
Incorrect number	79	88	79	76
percent	13.9330	15.5203	13.9330	13.4039
Kappa statistic	0.0000	0.0203	0.0000	0.1110
Mean absolute error	0.2398	0.2391	0.2345	0.1916
Root mean squared error	0.3463	0.3543	0.3426	0.3178
Relative absolute error	99.6202	99.3146	97.3845	79.5906
Root relative squared error	99.9993	102.3055	98.9258	91.7842

Classifier's summary of validate data set : group 2.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	428	430	428	424
percent	76.2923	76.6488	76.2923	75.5793
Incorrect number	133	131	133	137
percent	23.7077	23.3512	23.7077	24.4207
Kappa statistic	0.0000	0.0452	0.0000	0.0015
Mean absolute error	0.3103	0.3032	0.3046	0.2915
Root mean squared error	0.4364	0.4372	0.4349	0.4612
Relative absolute error	99.7857	97.4955	97.9406	93.7370
Root relative squared error	100.0647	100.2618	99.7258	105.7641

ตารางที่ 4.10 : ตารางแสดงผลลัพธ์จากตัวแบบจำแนกประเภทในข้อมูลกลุ่มที่ 3 กับข้อมูลการพิจารณาสินเชื่อจากธนาคารแห่งหนึ่งในประเทศไทย

Classifier's summary of train data set : group 3.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	268	255	268	273
percent	80.9668	77.0393	80.9668	82.4773
Incorrect number	63	76	63	58
percent	19.0332	22.9607	19.0332	17.5227
Kappa statistic	0.0000	0.0891	0.0000	0.1225
Mean absolute error	0.3082	0.2773	0.3021	0.3001
Root mean squared error	0.3926	0.4344	0.3886	0.3754
Relative absolute error	99.6277	89.6291	97.6447	97.0157
Root relative squared error	99.9989	110.6588	98.9806	95.6366

Classifier's summary of validate data set : group 3.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	263	237	263	255
percent	85.3896	76.9481	85.3896	82.7922
Incorrect number	45	71	45	53
percent	14.6104	23.0519	14.6104	17.2078
Kappa statistic	0.0000	-0.0280	0.0000	0.0092
Mean absolute error	0.2808	0.2823	0.2790	0.3075
Root mean squared error	0.3560	0.4470	0.3573	0.3874
Relative absolute error	99.5334	100.0477	98.8739	109.0021
Root relative squared error	99.9338	125.4895	100.3124	108.7588

ตารางที่ 4.11 : ตารางแสดงผลลัพธ์จากตัวแบบจำแนกประเภทสำหรับวิธีการจำแนกประเภทแต่ละวิธีที่ได้จากการทดสอบข้อมูลการพิจารณาสินเชื่อจากธนาคารแห่งหนึ่งในประเทศไทย

```
Classifier's summary of all train data set
weka.classifiers.trees.J48

Correctly Classified Instances        1156               83.6469 %
Incorrectly Classified Instances       226               16.3531 %
Kappa statistic                          0
Mean absolute error                      0.2736
Root mean squared error                  0.3698
Relative absolute error                 99.8805 %
Root relative squared error             99.9999 %

Classifier's summary of all train data set
weka.classifiers.bayes.NaiveBayes

Correctly Classified Instances        1120               81.042  %
Incorrectly Classified Instances       262               18.958  %
Kappa statistic                          0.0214
Mean absolute error                      0.2737
Root mean squared error                  0.3885
Relative absolute error                 99.9307 %
Root relative squared error            105.0504 %

Classifier's summary of all train data set
weka.classifiers.functions.Logistic

Correctly Classified Instances        1156               83.6469 %
Incorrectly Classified Instances       226               16.3531 %
Kappa statistic                          0
Mean absolute error                      0.2693
Root mean squared error                  0.367
Relative absolute error                 98.3299 %
Root relative squared error             99.2269 %

Classifier's summary of all train data set
weka.classifiers.functions.MultilayerPerceptron

Correctly Classified Instances        1166               84.3705 %
Incorrectly Classified Instances       216               15.6295 %
Kappa statistic                          0.098
Mean absolute error                      0.2395
Root mean squared error                  0.3561
Relative absolute error                 87.4314 %
Root relative squared error             96.2796 %
```

ตารางที่ 4.12 : ตัวทำนายเหมาะสมที่ได้จากการทดสอบข้อมูลการพิจารณาสินเชื่อจากธนาคารแห่งหนึ่งในประเทศไทย

กลุ่มของข้อมูล	วิธีการพัฒนาตัวแบบของตัวทำนายเหมาะสม
0	วิธีการต้นไม้การตัดสินใจ
1	วิธีการสมการถดถอยแบบโลจิสติก
2	วิธีการจำแนกแบบเบย์อย่างง่าย
3	วิธีการสมการถดถอยแบบโลจิสติก

ตารางที่ 4.13 : ผลลัพธ์ที่ได้จากการทดสอบข้อมูลการพิจารณาสินเชื่อจากธนาคารแห่งหนึ่งในประเทศไทย

Name of classifier method		n_CLAMP	tree n_J48	baye n_NaiveBayes	function n_Logistic	function n_MultilayerPerceptron
Correct.	Number	909	919	878	919	913
	Percent	79.3886	80.2620	76.6812	80.2620	79.7380
Incorrect.	Number	236	226	267	226	232
	Percent	20.6114	19.7380	23.3188	19.7380	20.2620
Kappa statistic		-0.0013	0.0000	-0.0017	0.0000	0.0105
means absolute error		0.2971	0.2964	0.3111	0.2948	0.2782
Root means square error		0.4110	0.3995	0.4308	0.4035	0.4145

ค่าสถิติที่แสดงในตารางที่ 4.13 มีดังต่อไปนี้

- Kappa statistic คือ ดัชนีของความสอดคล้องของข้อมูลจากการทำนายและข้อมูลจากการสังเกต โดยมีการปรับแก้ agreement มีสูตรที่ใช้ในการคำนวณดังนี้

$$Kappa = \frac{(Observed\ agreement - Chance\ agreement)}{(1 - Chance\ agreement)}$$

Observed agreement = *[*จำนวนของผลลัพธ์ที่ได้จากตัวแบบที่ทายว่าลูกค้าได้รับอนุมัติสินเชื่อ และ ผลลัพธ์จากเจ้าหน้าที่พิจารณาสินเชื่อคือลูกค้าได้รับการอนุมัติสินเชื่อ (true positive) + จำนวนของผลลัพธ์ที่ได้จากตัวแบบที่ทายว่าลูกค้าไม่ได้รับอนุมัติสินเชื่อ และผลลัพธ์จากเจ้าหน้าที่พิจารณาสินเชื่อคือลูกค้าไม่ได้รับการอนุมัติสินเชื่อ (true negative)*]* / จำนวนลูกค้าทั้งหมด

Change agreement = จำนวนข้อมูลที่ได้รับการอนุมัติ*จำนวนที่ทำนายว่าได้รับการอนุมัติ + จำนวนข้อมูลที่ไม่ได้รับการอนุมัติ*จำนวนที่ทำนายว่าไม่ได้รับการอนุมัติ

- Mean absolute error คือ ค่าสมบูรณ์ของความคลาดเคลื่อนเฉลี่ย

$$MAE = \frac{1}{n}\left(\sum|\hat{\theta} - \theta|\right)$$

- Root mean square error คือ ค่ารากที่สองของค่าความคลาดเคลื่อนยกกำลังสองเฉลี่ย

$$RMSD = \sqrt{\frac{\sum(\hat{\theta} - \theta)^2}{n}}$$

เมื่อ $\hat{\theta}$ คือค่าที่ได้จากการทำนายและ θ คือค่าที่ได้จากการสังเกต

วิธีการแก้ปัญหาขนาดของคลาสไม่สมดุลกันที่ใช้ในวิทยานิพนธ์นี้ คือการเพิ่มปริมาณข้อมูลที่น้อยกว่าให้มีขนาดเท่าๆ กับข้อมูลกลุ่มอื่น ผลลัพธ์คือจำนวนกลุ่มทั้งหมด 4 กลุ่ม แต่ละกลุ่มมีจำนวนข้อมูลตามตารางที่ 4.14 ซึ่งตัวแบบจำแนกประเภทที่เหมาะสมแสดงในตารางที่ 4.20 โดยพิจารณาจากค่าความแม่นยำของตัวแบบจำแนกประเภทตามตารางที่ 4.15 ตารางที่ 4.16 ตารางที่ 4.17 และตารางที่ 4.18 สำหรับตารางที่ 4.19 แสดงผลลัพธ์จากตัวแบบจำแนกประเภทเพียงอย่างเดียว ที่ได้จากการทดสอบข้อมูลการพิจารณาสินเชื่อจากธนาคารแห่งหนึ่งในประเทศไทย

สรุปผลลัพธ์ที่ได้จากการทดสอบข้อมูลการพิจารณาสินเชื่อ จากธนาคารแห่งหนึ่งในประเทศไทยที่แก้ปัญหาขนาดของคลาสไม่สมดุลที่แสดงในตารางที่ 4.21 ปรากฏว่าตัวแบบที่พัฒนาจาก CLAMP ให้ค่าความถูกต้องสูงที่สุด เมื่อเทียบกับตัวแบบที่พัฒนาจากวิธีการจำแนกของตัวแบบแต่ละวิธี

ตารางที่ 4.14 : ตารางแสดงอัตราส่วนร้อยละของจำนวนระเบียนที่อยู่ในแต่ละกลุ่มที่ได้จากตัวแบบการวิเคราะห์การเกาะกลุ่มของข้อมูลการพิจารณาสินเชื่อจากธนาคารแห่งหนึ่งในประเทศไทยที่แก้ปัญหาขนาดของคลาสไม่สมดุล

	อัตราส่วนร้อยละ		
	Training data	Validate data	Test data
กลุ่มที่ 0	17	15	20
กลุ่มที่ 1	26	23	21
กลุ่มที่ 2	22	26	25
กลุ่มที่ 3	35	36	35

ตารางที่ 4.15 : ตารางแสดงผลลัพธ์จากตัวแบบจำแนกประเภทในข้อมูลกลุ่มที่ 0 กับข้อมูลการพิจารณาสินเชื่อจากธนาคารแห่งหนึ่งในประเทศไทยที่แก้ปัญหาขนาดของคลาสไม่สมดุล

Classifier's summary of train data set : group 0.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	333	204	190	286
percent	96.5217	59.1304	55.0725	82.8986
Incorrect number	12	141	155	59
percent	3.4783	40.8696	44.9275	17.1014
Kappa statistic	0.9300	0.1874	0.0252	0.6650
Mean absolute error	0.0604	0.4740	0.4798	0.2382
Root mean squared error	0.1738	0.4975	0.4900	0.3560
Relative absolute error	12.2607	96.1579	97.3245	48.3169
Root relative squared error	35.0167	100.2068	98.7001	71.7016

Classifier's summary of validate data set : group 0.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	133	124	150	107
percent	64.2512	59.9034	72.4638	51.6908
Incorrect number	74	83	57	100
percent	35.7488	40.0966	27.5362	48.3092
Kappa statistic	-0.0034	-0.0130	0.1158	-0.1096
Mean absolute error	0.3521	0.4494	0.4488	0.4288
Root mean squared error	0.5863	0.4786	0.4623	0.5824
Relative absolute error	75.2654	96.0727	95.9527	91.6754
Root relative squared error	124.6590	101.7601	98.2772	123.8226

ตารางที่ 4.16 : ตารางแสดงผลลัพธ์จากตัวแบบจำแนกประเภทในข้อมูลกลุ่มที่ 1 กับข้อมูลการพิจารณาสินเชื่อจากธนาคารแห่งหนึ่งในประเทศไทยที่แก้ปัญหาขนาดของคลาสไม่สมดุล

Classifier's summary of train data set : group 1.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	516	284	313	335
percent	97.5425	53.6862	59.1682	63.3270
Incorrect number	13	245	216	194
percent	2.4575	46.3138	40.8318	36.6730
Kappa statistic	0.9509	0.0636	0.1810	0.2735
Mean absolute error	0.0422	0.4580	0.4809	0.4068
Root mean squared error	0.1452	0.5656	0.4904	0.4613
Relative absolute error	8.4385	91.6260	96.2124	81.3867
Root relative squared error	29.0491	113.1376	98.0930	92.2664

Classifier's summary of validate data set : group 1.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	211	225	191	129
percent	68.2848	72.8155	61.8123	41.7476
Incorrect number	98	84	118	180
percent	31.7152	27.1845	38.1877	58.2524
Kappa statistic	-0.0281	-0.0536	0.0348	0.0325
Mean absolute error	0.3120	0.3621	0.4829	0.5033
Root mean squared error	0.5528	0.4736	0.4916	0.5559
Relative absolute error	63.1681	73.3106	97.7492	101.8924
Root relative squared error	111.8898	95.8550	99.5090	112.5233

ตารางที่ 4.17 : ตารางแสดงผลลัพธ์จากตัวแบบจำแนกประเภทในข้อมูลกลุ่มที่ 2 กับข้อมูลการพิจารณาสินเชื่อจากธนาคารแห่งหนึ่งในประเทศไทย ที่แก้ปัญหาขนาดของคลาสไม่สมดุล

Classifier's summary of train data set : group 2.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	445	299	317	397
percent	96.7391	65.0000	68.9130	86.3043
Incorrect number	15	161	143	63
percent	3.2609	35.0000	31.0870	13.6957
Kappa statistic	0.9304	0.1240	0.2195	0.7011
Mean absolute error	0.0527	0.4134	0.4233	0.1831
Root mean squared error	0.1624	0.4742	0.4593	0.3350
Relative absolute error	11.4920	90.0694	92.2186	39.8935
Root relative squared error	33.9064	98.9971	95.8918	69.9363

Classifier's summary of validate data set : group 2.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	232	258	249	253
percent	64.6240	71.8663	69.3593	70.4735
Incorrect number	127	101	110	106
percent	35.3760	28.1337	30.6407	29.5265
Kappa statistic	-0.0061	0.0457	0.0394	0.0984
Mean absolute error	0.3423	0.3922	0.4098	0.3114
Root mean squared error	0.5726	0.4540	0.4474	0.5014
Relative absolute error	80.1389	91.8016	95.9377	72.8976
Root relative squared error	128.8114	102.1260	100.6438	112.8055

ตารางที่ 4.18 : ตารางแสดงผลลัพธ์จากตัวแบบจำแนกประเภทในข้อมูลกลุ่มที่ 3 กับข้อมูลการพิจารณาสินเชื่อจากธนาคารแห่งหนึ่งในประเทศไทยที่แก้ปัญหาขนาดของคลาสไม่สมดุล

Classifier's summary of train data set : group 3.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	702	421	412	506
percent	96.6942	57.9890	56.7493	69.6970
Incorrect number	24	305	314	220
percent	3.3058	42.0110	43.2507	30.3030
Kappa statistic	0.9337	0.1239	0.0631	0.4005
Mean absolute error	0.0549	0.4877	0.4870	0.3577
Root mean squared error	0.1657	0.4962	0.4935	0.4276
Relative absolute error	11.0816	98.4571	98.3107	72.2091
Root relative squared error	33.2894	99.7037	99.1564	85.9248

Classifier's summary of validate data set : group 3.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	322	328	366	277
percent	65.8487	67.0757	74.8466	56.6462
Incorrect number	167	161	123	212
percent	34.1513	32.9243	25.1534	43.3538
Kappa statistic	-0.0023	-0.0215	0.0334	-0.0435
Mean absolute error	0.3355	0.4700	0.4643	0.4195
Root mean squared error	0.5704	0.4809	0.4696	0.5183
Relative absolute error	70.9305	99.3741	98.1590	88.7063
Root relative squared error	120.1698	101.3187	98.9415	109.2016

ตารางที่ 4.19 : ตารางแสดงผลลัพธ์จากตัวแบบจำแนกประเภทสำหรับวิธีการจำแนกประเภทแต่ละวิธีที่ได้จากการทดสอบข้อมูลการพิจารณาสินเชื่อจากธนาคารแห่งหนึ่งในประเทศไทยที่แก้ปัญหาขนาดของคลาสไม่สมดุล

```
Classifier's summary of all train data set
weka.classifiers.trees.J48

Correctly Classified Instances        2007               97.4272 %
Incorrectly Classified Instances        53                2.5728 %
Kappa statistic                          0.9481
Mean absolute error                      0.0439
Root mean squared error                  0.1479
Relative absolute error                  8.9063 %
Root relative squared error             29.7983 %

Classifier's summary of all train data set
weka.classifiers.bayes.NaiveBayes

Correctly Classified Instances        1154               56.0194 %
Incorrectly Classified Instances       906               43.9806 %
Kappa statistic                          0.0416
Mean absolute error                      0.4734
Root mean squared error                  0.5029
Relative absolute error                 96.1124 %
Root relative squared error            101.3501 %

Classifier's summary of all train data set
weka.classifiers.functions.Logistic

Correctly Classified Instances        1196               58.0583 %
Incorrectly Classified Instances       864               41.9417 %
Kappa statistic                          0.098
Mean absolute error                      0.4795
Root mean squared error                  0.4896
Relative absolute error                 97.3503 %
Root relative squared error             98.6681 %

Classifier's summary of all train data set
weka.classifiers.functions.MultilayerPerceptron

Correctly Classified Instances        1195               58.0097 %
Incorrectly Classified Instances       865               41.9903 %
Kappa statistic                          0.1738
Mean absolute error                      0.4042
Root mean squared error                  0.4518
Relative absolute error                 82.0663 %
Root relative squared error             91.0538 %
```

ตารางที่ 4.20 : ตัวทำนายเหมาะสมที่ได้จากการทดสอบข้อมูลการพิจารณาสินเชื่อจากธนาคารแห่งหนึ่งในประเทศไทยที่แก้ปัญหาขนาดของคลาสไม่สมดุล

กลุ่มของข้อมูล	วิธีการพัฒนาตัวแบบของตัวทำนายเหมาะสม
0	วิธีการสมการถดถอยแบบโลจิสติก
1	วิธีการจำแนกแบบเบย์อย่างง่าย
2	วิธีการจำแนกแบบเบย์อย่างง่าย
3	วิธีการสมการถดถอยแบบโลจิสติก

ตารางที่ 4.21 : ผลลัพธ์ที่ได้จากการทดสอบข้อมูลการพิจารณาสินเชื่อจากธนาคารแห่งหนึ่งในประเทศไทยที่แก้ปัญหาขนาดของคลาสไม่สมดุล

Name of classifier method		n_CLAMP	tree n_J48	baye n_NaiveBayes	function n_Logistic	function n_MultilayerPerceptron
Correct.	Number	859	799	773	802	595
	Percent	75.0218	69.7817	67.5109	70.0437	51.9651
Incorrect.	Number	286	346	372	343	550
	Percent	24.9782	30.2183	32.4891	29.9563	48.0349
Kappa statistic		-0.0153	0.0133	-0.0220	-0.0222	-0.0149
means absolute error		0.4235	0.2957	0.4585	0.4562	0.4278
Root means square error		0.4669	0.5351	0.4914	0.4669	0.5052

2. ข้อมูลการพิจารณาสินเชื่อจากประเทศออสเตรเลีย

ข้อมูลการพิจารณาสินเชื่อจากประเทศออสเตรเลียมีจำนวนระเบียนทั้งหมด 690 ระเบียน แบ่งข้อมูลออกเป็น 3 ส่วน คือข้อมูลพัฒนาตัวแบบจำนวน 267 ระเบียน ข้อมูลประเมินจำนวน 199 ระเบียน และข้อมูลทดสอบจำนวน 224 ระเบียน โดยแสดงจำนวนข้อมูลที่อนุมัติ และไม่อนุมัติในตารางที่ 4.22

หลังจากใช้ CLAMP ได้ผลลัพธ์คือจำนวนกลุ่มทั้งหมด 2 กลุ่ม แต่ละกลุ่มมีจำนวนข้อมูลตามตารางที่ 4.23 ซึ่งตัวแบบจำแนกประเภทที่เหมาะสมแสดงในตารางที่ 4.27 โดยพิจารณาจากค่าความแม่นยำของตัวแบบจำแนกประเภทตามตารางที่ 4.24 และตารางที่ 4.25 สำหรับตารางที่ 4.26 แสดงผลลัพธ์จากตัวแบบจำแนกประเภทเพียงอย่างเดียวที่ได้จากการทดสอบข้อมูลการพิจารณาสินเชื่อจากประเทศออสเตรเลีย

สรุปผลลัพธ์ที่ได้จากการทดสอบกับข้อมูลพิจารณาสินเชื่อจากประเทศออสเตรเลีย แสดงรายละเอียดในตารางที่ 4.28 พบว่าตัวแบบที่มีค่าความแม่นยำสูงที่สุด คือ ตัวแบบต้นไม้การตัดสินใจ และตัวแบบเพอร์เซพตรอนหลายชั้น โดยตัวแบบ CLAMP มีความแม่นยำน้อยกว่าตัวแบบที่มีความแม่นยำสูงสุดเล็กน้อย

ตารางที่ 4.22 : ตารางแสดงจำนวนระเบียนหลังจากการแบ่งข้อมูลการพิจารณาสินเชื่อจากประเทศออสเตรเลีย

	จำนวนระเบียน		
	Training data	Validate data	Test data
Accept	122	100	138
Reject	145	99	86
Total	267	199	224

ตารางที่ 4.23 : ตารางแสดงอัตราส่วนร้อยละของจำนวนระเบียนที่อยู่ในแต่ละกลุ่มที่ได้จากตัวแบบการวิเคราะห์การเกาะกลุ่มของข้อมูลการพิจารณาสินเชื่อจากประเทศออสเตรเลีย

	อัตราส่วนร้อยละ		
	Training data	Validate data	Test data
กลุ่มที่ 0	73	79	79
กลุ่มที่ 1	27	21	21

ตารางที่ 4.24 : ตารางแสดงผลลัพธ์จากตัวแบบจำแนกประเภทในข้อมูลกลุ่มที่ 0 กับข้อมูลการพิจารณาสินเชื่อจากประเทศออสเตรเลีย

Classifier's summary of train data set : group 0.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	190	169	188	197
percent	94.5274	84.0796	93.5323	98.0100
Incorrect number	11	32	13	4
percent	5.4726	15.9204	6.4677	1.9900
Kappa statistic	0.8903	0.6854	0.8698	0.9599
Mean absolute error	0.0901	0.1763	0.1001	0.0277
Root mean squared error	0.2123	0.3750	0.2246	0.1267
Relative absolute error	18.1234	35.4510	20.1307	5.5665
Root relative squared error	42.5727	75.2046	45.0534	25.4068

Classifier's summary of validate data set : group 0.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	126	111	126	125
percent	84.5638	74.4966	84.5638	83.8926
Incorrect number	23	38	23	24
percent	15.4362	25.5034	15.4362	16.1074
Kappa statistic	0.6905	0.5011	0.6899	0.6786
Mean absolute error	0.1667	0.2534	0.1660	0.1576
Root mean squared error	0.3573	0.4701	0.3531	0.3737
Relative absolute error	33.5162	50.9551	33.3801	31.6861
Root relative squared error	71.6535	94.2828	70.8117	74.9437

ตารางที่ 4.25 : ตารางแสดงผลลัพธ์จากตัวแบบจำแนกประเภทในข้อมูลกลุ่มที่ 1 กับข้อมูลการพิจารณาสินเชื่อจากประเทศออสเตรเลีย

Classifier's summary of train data set : group 1.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	62	55	64	65
percent	93.9394	83.3333	96.9697	98.4848
Incorrect number	4	11	2	1
percent	6.0606	16.6667	3.0303	1.5152
Kappa statistic	0.8277	0.4590	0.9093	0.9535
Mean absolute error	0.0925	0.1472	0.0688	0.0205
Root mean squared error	0.2151	0.3521	0.1788	0.1236
Relative absolute error	27.2784	43.4080	20.2792	6.0310
Root relative squared error	52.5970	86.1198	43.7208	30.2206

Classifier's summary of validate data set : group 1.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	40	41	39	40
percent	80.0000	82.0000	78.0000	80.0000
Incorrect number	10	9	11	10
percent	20.0000	18.0000	22.0000	20.0000
Kappa statistic	0.5840	0.5975	0.4860	0.5575
Mean absolute error	0.2268	0.1796	0.2475	0.1990
Root mean squared error	0.4049	0.4200	0.4756	0.4282
Relative absolute error	52.3936	41.4815	57.1563	45.9563
Root relative squared error	79.2564	82.2132	93.0891	83.8135

ตารางที่ 4.26 : ตารางแสดงผลลัพธ์จากตัวแบบจำแนกประเภทสำหรับวิธีการจำแนกประเภทแต่ละวิธีที่ได้จากการทดสอบข้อมูลการพิจารณาสินเชื่อจากประเทศออสเตรเลีย

Classifier's summary of all train data set
weka.classifiers.trees.J48

Correctly Classified Instances	256	95.8801 %
Incorrectly Classified Instances	11	4.1199 %
Kappa statistic	0.9173	
Mean absolute error	0.0722	
Root mean squared error	0.19	
Relative absolute error	14.5415 %	
Root relative squared error	38.1344 %	

Classifier's summary of all train data set
weka.classifiers.bayes.NaiveBayes

Correctly Classified Instances	226	84.6442 %
Incorrectly Classified Instances	41	15.3558 %
Kappa statistic	0.6859	
Mean absolute error	0.1674	
Root mean squared error	0.3723	
Relative absolute error	33.7349 %	
Root relative squared error	74.7452 %	

Classifier's summary of all train data set
weka.classifiers.functions.Logistic

Correctly Classified Instances	250	93.633 %
Incorrectly Classified Instances	17	6.367 %
Kappa statistic	0.8721	
Mean absolute error	0.1099	
Root mean squared error	0.2283	
Relative absolute error	22.136 %	
Root relative squared error	45.8333 %	

Classifier's summary of all train data set
weka.classifiers.functions.MultilayerPerceptron

Correctly Classified Instances	261	97.7528 %
Incorrectly Classified Instances	6	2.2472 %
Kappa statistic	0.9547	
Mean absolute error	0.0292	
Root mean squared error	0.1394	
Relative absolute error	5.8755 %	
Root relative squared error	27.9891 %	

ตารางที่ 4.27 : ตัวทำนายเหมาะสมที่ได้จากการทดสอบข้อมูลการพิจารณาสินเชื่อจากประเทศออสเตรเลีย

กลุ่มของข้อมูล	วิธีการพัฒนาตัวแบบของตัวทำนายเหมาะสม
0	วิธีการสมการถดถอยแบบโลจิสติก
1	วิธีการจำแนกแบบเบย์อย่างง่าย

ตารางที่ 4.28 : ผลลัพธ์ที่ได้จากการทดสอบข้อมูลการพิจารณาสินเชื่อจากประเทศออสเตรเลีย

Name of classifier method		n_CLAMP	tree n_J48	baye n_NaiveBayes	function n_Logistic	function n_MultilayerPerceptron
Correct.	Number	181	185	176	183	185
	Percent	80.8036	82.5893	78.5714	81.6964	82.5893
Incorrect.	Number	43	39	48	41	39
	Percent	19.1964	17.4107	21.4286	18.3036	17.4107
Kappa statistic		0.5933	0.6359	0.5284	0.6189	0.6311
means absolute error		0.2083	0.1874	0.2110	0.2117	0.1700
Root means square error		0.4150	0.3899	0.4332	0.4042	0.3911

3. ข้อมูลการพิจารณาสินเชื่อจากประเทศเยอรมัน

ข้อมูลการพิจารณาสินเชื่อจากประเทศเยอรมันมีจำนวนระเบียนทั้งหมด 1,000 ระเบียน แบ่งข้อมูลออกเป็น 3 ส่วน คือข้อมูลพัฒนาตัวแบบจำนวน 416 ระเบียน ข้อมูลประเมินจำนวน 307 ระเบียน และข้อมูลทดสอบจำนวน 277 ระเบียน โดยแสดงจำนวนข้อมูลที่อนุมัติ และไม่อนุมัติในตารางที่ 4.29

หลังจากใช้ CLAMP ได้ผลลัพธ์คือจำนวนกลุ่มทั้งหมด 2 กลุ่ม แต่ละกลุ่มมีจำนวนข้อมูลตามตารางที่ 4.30 ซึ่งตัวแบบจำแนกประเภทที่เหมาะสมแสดงในตารางที่ 4.34 โดยพิจารณาจากค่าความแม่นยำของตัวแบบจำแนกประเภทตามตารางที่ 4.31 และตารางที่ 4.32 สำหรับตารางที่ 4.33 แสดงผลลัพธ์จากตัวแบบจำแนกประเภทเพียงอย่างเดียวที่ได้จากการทดสอบข้อมูลการพิจารณาสินเชื่อจากประเทศเยอรมัน

สรุปผลลัพธ์ที่ได้จากการทดสอบกับข้อมูลพิจารณาสินเชื่อจากประเทศเยอรมัน แสดงรายละเอียดในตารางที่ 4.35 พบว่าตัวแบบที่มีค่าความแม่นยำสูงที่สุดคือ ตัวแบบสมการถดถอยแบบโลจิสติก โดยตัวแบบ CLAMP มีความแม่นยำน้อยกว่าตัวแบบที่มีความแม่นยำสูงสุดเล็กน้อย

ตารางที่ 4.29 : ตารางแสดงจำนวนระเบียนหลังจากการแบ่งข้อมูลการพิจารณาสินเชื่อจากประเทศเยอรมัน

	จำนวนระเบียน		
	Training data	Validate data	Test data
Accept	290	97	200
Reject	126	210	77
Total	416	307	277

ตารางที่ 4.30 : ตารางแสดงอัตราส่วนร้อยละของจำนวนระเบียนที่อยู่ในแต่ละกลุ่มที่ได้จากตัวแบบการวิเคราะห์การเกาะกลุ่มของข้อมูลการพิจารณาสินเชื่อจากประเทศเยอรมัน

	อัตราส่วนร้อยละ		
	Training data	Validate data	Test data
กลุ่มที่ 0	49	49	50
กลุ่มที่ 1	51	51	50

ตารางที่ 4.31 : ตารางแสดงผลลัพธ์จากตัวแบบจำแนกประเภทในข้อมูลกลุ่มที่ 0 กับข้อมูลการพิจารณาสินเชื่อจากประเทศเยอรมัน

Classifier's summary of train data set : group 0.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	174	150	161	203
percent	85.7143	73.8916	79.3103	100.0000
Incorrect number	29	53	42	0
percent	14.2857	26.1084	20.6897	0.0000
Kappa statistic	0.6620	0.4126	0.4997	1.0000
Mean absolute error	0.2255	0.2944	0.2858	0.0071
Root mean squared error	0.3358	0.4431	0.3789	0.0091
Relative absolute error	52.1423	68.0862	66.0868	1.6308
Root relative squared error	72.2648	95.3623	81.5417	1.9628

Classifier's summary of validate data set : group 0.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	102	114	111	107
percent	68.4564	76.5101	74.4966	71.8121
Incorrect number	47	35	38	42
percent	31.5436	23.4899	25.5034	28.1879
Kappa statistic	0.2663	0.4706	0.3838	0.3479
Mean absolute error	0.3768	0.2876	0.3239	0.2912
Root mean squared error	0.5124	0.4254	0.4419	0.4911
Relative absolute error	86.1387	65.7438	74.0525	66.5753
Root relative squared error	109.0344	90.5255	94.0381	104.5082

ตารางที่ 4.32 : ตารางแสดงผลลัพธ์จากตัวแบบจำแนกประเภทในข้อมูลกลุ่มที่ 1 กับข้อมูลการพิจารณาสินเชื่อจากประเทศเยอรมัน

Classifier's summary of train data set : group 1.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	196	167	181	212
percent	92.0188	78.4038	84.9765	99.5305
Incorrect number	17	46	32	1
percent	7.9812	21.5962	15.0235	0.4695
Kappa statistic	0.8000	0.4912	0.6216	0.9886
Mean absolute error	0.1340	0.2217	0.1968	0.0096
Root mean squared error	0.2588	0.3948	0.3118	0.0689
Relative absolute error	32.3933	53.6017	47.5999	2.3186
Root relative squared error	56.9705	86.9070	68.6326	15.1693

Classifier's summary of validate data set : group 1.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	108	105	114	107
percent	68.3544	66.4557	72.1519	67.7215
Incorrect number	50	53	44	51
percent	31.6456	33.5443	27.8481	32.2785
Kappa statistic	0.2339	0.2207	0.2918	0.2324
Mean absolute error	0.3340	0.3314	0.3137	0.3120
Root mean squared error	0.5199	0.5040	0.4719	0.5148
Relative absolute error	79.7544	79.1283	74.8994	74.5056
Root relative squared error	113.0122	109.5654	102.5770	111.9180

ตารางที่ 4.33 : ตารางแสดงผลลัพธ์จากตัวแบบจำแนกประเภทสำหรับวิธีการจำแนกประเภทแต่ละวิธีที่ได้จากการทดสอบข้อมูลการพิจารณาสินเชื่อจากประเทศเยอรมัน

Classifier's summary of all train data set
weka.classifiers.trees.J48

Correctly Classified Instances	375	90.1442 %
Incorrectly Classified Instances	41	9.8558 %
Kappa statistic	0.7492	
Mean absolute error	0.168	
Root mean squared error	0.2899	
Relative absolute error	39.7553 %	
Root relative squared error	63.0795 %	

Classifier's summary of all train data set
weka.classifiers.bayes.NaiveBayes

Correctly Classified Instances	318	76.4423 %
Incorrectly Classified Instances	98	23.5577 %
Kappa statistic	0.452	
Mean absolute error	0.2691	
Root mean squared error	0.422	
Relative absolute error	63.6563 %	
Root relative squared error	91.8276 %	

Classifier's summary of all train data set
weka.classifiers.functions.Logistic

Correctly Classified Instances	328	78.8462 %
Incorrectly Classified Instances	88	21.1538 %
Kappa statistic	0.4706	
Mean absolute error	0.2892	
Root mean squared error	0.3802	
Relative absolute error	68.4218 %	
Root relative squared error	82.7337 %	

Classifier's summary of all train data set
weka.classifiers.functions.MultilayerPerceptron

Correctly Classified Instances	413	99.2788 %
Incorrectly Classified Instances	3	0.7212 %
Kappa statistic	0.9828	
Mean absolute error	0.0128	
Root mean squared error	0.0853	
Relative absolute error	3.0375 %	
Root relative squared error	18.5614 %	

ตารางที่ 4.34 : ตัวทำนายเหมาะสมที่ได้จากการทดสอบข้อมูลการพิจารณาสินเชื่อจากประเทศเยอรมัน

กลุ่มของข้อมูล	วิธีการพัฒนาตัวแบบของตัวทำนายเหมาะสม
0	วิธีการจำแนกแบบเบย์อย่างง่าย
1	วิธีการสมการถดถอยแบบโลจิสติก

ตารางที่ 4.35 : ผลลัพธ์ที่ได้จากการทดสอบข้อมูลการพิจารณาสินเชื่อจากประเทศเยอรมัน

Name of classifier method		n_CLAMP	tree n_J48	baye n_NaiveBayes	function n_Logistic	function n_MultilayerPerceptron
Correct.	Number	206	199	204	207	204
	Percent	74.3682	71.8412	73.6462	74.7292	73.6462
Incorrect.	Number	71	78	73	70	73
	Percent	25.6318	28.1588	26.3538	25.2708	26.3538
Kappa statistic		0.3537	0.2241	0.3133	0.3332	0.3512
means absolute error		0.2907	0.3205	0.2924	0.2957	0.2748
Root means square error		0.4576	0.4586	0.4208	0.3962	0.4764

4. ข้อมูลการประเมินคุณภาพรถยนต์

ข้อมูลการประเมินคุณภาพรถยนต์มีจำนวนระเบียนทั้งหมด 1,728 ระเบียน แบ่งข้อมูลออกเป็น 3 ส่วน คือข้อมูลพัฒนาตัวแบบจำนวน 648 ระเบียน ข้อมูลประเมินจำนวน 531 ระเบียน และข้อมูลทดสอบจำนวน 549 ระเบียน โดยแสดงจำนวนข้อมูลคุณภาพรถยนต์ในตารางที่ 4.36

หลังจากใช้ CLAMP ได้ผลลัพธ์คือจำนวนกลุ่มทั้งหมด 2 กลุ่ม แต่ละกลุ่มมีจำนวนข้อมูลตามตารางที่ 4.37 ซึ่งตัวแบบจำแนกประเภทที่เหมาะสมแสดงในตารางที่ 4.41 โดยพิจารณาจากค่าความแม่นยำของตัวแบบจำแนกประเภทตามตารางที่ 4.38 และตารางที่ 4.39 สำหรับตารางที่ 4.40 แสดงผลลัพธ์จากตัวแบบจำแนกประเภทเพียงอย่างเดียวที่ได้จากการทดสอบข้อมูลการประเมินคุณภาพรถยนต์

สรุปผลลัพธ์ที่ได้จากการทดสอบกับข้อมูลการประเมินคุณภาพรถยนต์แสดงรายละเอียดในตารางที่ 4.42 พบว่าตัวแบบที่มีค่าความแม่นยำสูงที่สุดคือ ตัวแบบ CLAMP

ตารางที่ 4.36 : ตารางแสดงจำนวนระเบียนหลังจากการแบ่งข้อมูลการประเมินคุณภาพรถยนต์

	จำนวนระเบียน		
	Training data	Validate data	Test data
Unacc	446	370	394
Acc	150	118	116
Vgood	31	20	14
Good	21	23	25
Total	648	531	549

ตารางที่ 4.37 : ตารางแสดงอัตราส่วนร้อยละของจำนวนระเบียนที่อยู่ในแต่ละกลุ่มที่ได้จากตัวแบบการวิเคราะห์การเกาะกลุ่มของข้อมูลการประเมินคุณภาพรถยนต์

	อัตราส่วนร้อยละ		
	Training data	Validate data	Test data
กลุ่มที่ 0	62	58	55
กลุ่มที่ 1	38	42	45

ตารางที่ 4.38 : ตารางแสดงผลลัพธ์จากตัวแบบจำแนกประเภทในข้อมูลกลุ่มที่ 0 กับข้อมูลการประเมินคุณภาพรถยนต์

Classifier's summary of train data set : group 0.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	383	334	383	400
percent	95.7500	83.5000	95.7500	100.0000
Incorrect number	17	66	17	0
percent	4.2500	16.5000	4.2500	0.0000
Kappa statistic	0.9257	0.7218	0.9257	1.0000
Mean absolute error	0.0340	0.1135	0.0289	0.0046
Root mean squared error	0.1304	0.2337	0.1214	0.0111
Relative absolute error	11.9070	39.7149	10.1255	1.6194
Root relative squared error	34.5599	61.9220	32.1764	2.9467

Classifier's summary of validate data set : group 0.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	268	252	260	286
percent	87.2964	82.0847	84.6906	93.1596
Incorrect number	39	55	47	21
percent	12.7036	17.9153	15.3094	6.8404
Kappa statistic	0.7824	0.7020	0.7413	0.8808
Mean absolute error	0.0700	0.1287	0.0774	0.0429
Root mean squared error	0.2253	0.2528	0.2578	0.1692
Relative absolute error	24.5261	45.0680	27.1248	15.0263
Root relative squared error	59.7465	67.0461	68.3773	44.8636

ตารางที่ 4.39 : ตารางแสดงผลลัพธ์จากตัวแบบจำแนกประเภทในข้อมูลกลุ่มที่ 1 กับข้อมูลการประเมินคุณภาพรถยนต์

Classifier's summary of train data set : group 1.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	244	228	248	248
percent	98.3871	91.9355	100.0000	100.0000
Incorrect number	4	20	0	0
percent	1.6129	8.0645	0.0000	0.0000
Kappa statistic	0.9278	0.7015	1.0000	1.0000
Mean absolute error	0.0137	0.0400	0.0000	0.0029
Root mean squared error	0.0828	0.1587	0.0000	0.0099
Relative absolute error	11.6300	33.9507	0.0001	2.4877
Root relative squared error	34.7085	66.5061	0.0004	4.1309

Classifier's summary of validate data set : group 1.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	209	195	210	212
percent	93.3036	87.0536	93.7500	94.6429
Incorrect number	15	29	14	12
percent	6.6964	12.9464	6.2500	5.3571
Kappa statistic	0.6981	0.4990	0.7539	0.7794
Mean absolute error	0.0331	0.0657	0.0305	0.0310
Root mean squared error	0.1699	0.2262	0.1712	0.1457
Relative absolute error	26.7988	53.1860	24.6525	25.0566
Root relative squared error	67.9113	90.4017	68.4415	58.2239

ตารางที่ 4.40 : ตารางแสดงผลลัพธ์จากตัวแบบจำแนกประเภทสำหรับวิธีการจำแนกประเภทแต่ละวิธีที่ได้จากการทดสอบข้อมูลการประเมินคุณภาพรถยนต์

Classifier's summary of all train data set
weka.classifiers.trees.J48

Correctly Classified Instances	633	97.6852 %
Incorrectly Classified Instances	15	2.3148 %
Kappa statistic	0.9511	
Mean absolute error	0.0186	
Root mean squared error	0.0964	
Relative absolute error	7.8969 %	
Root relative squared error	28.1526 %	

Classifier's summary of all train data set
weka.classifiers.bayes.NaiveBayes

Correctly Classified Instances	570	87.963 %
Incorrectly Classified Instances	78	12.037 %
Kappa statistic	0.7497	
Mean absolute error	0.0932	
Root mean squared error	0.2045	
Relative absolute error	39.589 %	
Root relative squared error	59.6863 %	

Classifier's summary of all train data set
weka.classifiers.functions.Logistic

Correctly Classified Instances	616	95.0617 %
Incorrectly Classified Instances	32	4.9383 %
Kappa statistic	0.8949	
Mean absolute error	0.0295	
Root mean squared error	0.1228	
Relative absolute error	12.534 %	
Root relative squared error	35.8438 %	

Classifier's summary of all train data set
weka.classifiers.functions.MultilayerPerceptron

Correctly Classified Instances	648	100 %
Incorrectly Classified Instances	0	0 %
Kappa statistic	1	
Mean absolute error	0.0037	
Root mean squared error	0.0097	
Relative absolute error	1.5814 %	
Root relative squared error	2.8446 %	

ตารางที่ 4.41 : ตัวทำนายเหมาะสมที่ได้จากการทดสอบข้อมูลการประเมินคุณภาพรถยนต์

กลุ่มของข้อมูล	วิธีการพัฒนาตัวแบบของตัวทำนายเหมาะสม
0	วิธีการเพอร์เซพตรอนหลายชั้น
1	วิธีการเพอร์เซพตรอนหลายชั้น

ตารางที่ 4.42 : ผลลัพธ์ที่ได้จากการทดสอบข้อมูลการประเมินคุณภาพรถยนต์

Name of classifier method		n_CLAMP	tree n_J48	baye n_NaiveBayes	function n_Logistic	function n_MultilayerPerceptron
Correct.	Number	526	525	480	505	524
	Percent	95.8106	95.6284	87.4317	91.9854	95.4463
Incorrect.	Number	23	24	69	44	25
	Percent	4.1894	4.3716	12.5683	8.0146	4.5537
Kappa statistic		0.9038	0.9019	0.7193	0.8151	0.8974
means absolute error		0.0277	0.0310	0.0885	0.0406	0.0278
Root means square error		0.1300	0.1477	0.2044	0.1708	0.1319

5. **ข้อมูล** spambase

ข้อมูล spambase มีจำนวนระเบียนทั้งหมด 4,601 ระเบียน แบ่งข้อมูลออกเป็น 3 ส่วน คือข้อมูลพัฒนาตัวแบบจำนวน 1,808 ระเบียน ข้อมูลประเมินจำนวน 1,390 ระเบียน และข้อมูลทดสอบจำนวน 1,403 ระเบียน โดยแสดงจำนวนข้อมูล spambase ในตารางที่ 4.43

หลังจากใช้ CLAMP ได้ผลลัพธ์คือจำนวนกลุ่มทั้งหมด 4 กลุ่ม แต่ละกลุ่มมีจำนวนข้อมูลตามตารางที่ 4.44 ซึ่งตัวแบบจำแนกประเภทที่เหมาะสมแสดงในตารางที่ 4.50 โดยพิจารณาจากค่าความแม่นยำของตัวแบบจำแนกประเภทตามตารางที่ 4.45 ตารางที่ 4.46 ตารางที่ 4.47 และตารางที่ 4.48 สำหรับตารางที่ 4.49 แสดงผลลัพธ์จากตัวแบบจำแนกประเภทเพียงอย่างเดียวที่ได้จากการทดสอบข้อมูล spambase

สรุปผลลัพธ์ที่ได้จากการทดสอบกับข้อมูล spambase แสดงรายละเอียดในตารางที่ 4.51 พบว่าตัวแบบที่มีค่าความแม่นยำสูงที่สุดคือ ตัวแบบ CLAMP

ตารางที่ 4.43 : ตารางแสดงจำนวนระเบียนหลังจากการแบ่งข้อมูล spambase

	จำนวนระเบียน		
	Training data	Validate data	Test data
Spam	704	565	544
non-spam	1,104	825	859
Total	1,808	1,390	1,403

ตารางที่ 4.44 : ตารางแสดงอัตราส่วนร้อยละของจำนวนระเบียนที่อยู่ในแต่ละกลุ่มที่ได้จากตัวแบบการวิเคราะห์การเกาะกลุ่มของข้อมูล spambase

	อัตราส่วนร้อยละ		
	Training data	Validate data	Test data
กลุ่มที่ 0	4	4	3
กลุ่มที่ 1	34	36	36
กลุ่มที่ 2	1	1	1
กลุ่มที่ 3	62	59	60

ตารางที่ 4.45 : ตารางแสดงผลลัพธ์จากตัวแบบจำแนกประเภทในข้อมูลกลุ่มที่ 0 กับข้อมูล spambase

Classifier's summary of train data set : group 0.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	67	67	67	67
percent	100.0000	100.0000	100.0000	100.0000
Incorrect number	0	0	0	0
percent	0.0000	0.0000	0.0000	0.0000
Kappa statistic	1.0000	1.0000	1.0000	1.0000
Mean absolute error	0.0000	0.0000	0.0000	0.0050
Root mean squared error	0.0000	0.0000	0.0000	0.0119
Relative absolute error	0.0000	0.0000	0.0006	7.1232
Root relative squared error	0.0000	0.0000	0.0008	6.9799

Classifier's summary of validate data set : group 0.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	56	52	53	53
percent	98.2456	91.2281	92.9825	92.9825
Incorrect number	1	5	4	4
percent	1.7544	8.7719	7.0175	7.0175
Kappa statistic	0.8995	0.2636	0.4722	0.4722
Mean absolute error	0.0175	0.0877	0.0713	0.0698
Root mean squared error	0.1325	0.2962	0.2650	0.2365
Relative absolute error	12.5683	62.8415	51.0651	50.0283
Root relative squared error	42.3106	94.6094	84.6462	75.5599

ตารางที่ 4.46 : ตารางแสดงผลลัพธ์จากตัวแบบจำแนกประเภทในข้อมูลกลุ่มที่ 1 กับข้อมูล spambase

Classifier's summary of train data set : group 1.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	590	559	581	554
percent	96.4052	91.3399	94.9346	90.5229
Incorrect number	22	53	31	58
percent	3.5948	8.6601	5.0654	9.4771
Kappa statistic	0.9050	0.7807	0.8658	0.7742
Mean absolute error	0.0657	0.0893	0.0721	0.1090
Root mean squared error	0.1812	0.2793	0.1864	0.2513
Relative absolute error	17.2031	23.3922	18.8759	28.5387
Root relative squared error	41.4976	63.9581	42.6769	57.5356

Classifier's summary of validate data set : group 1.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	452	453	453	450
percent	89.8608	90.0596	90.0596	89.4632
Incorrect number	51	50	50	53
percent	10.1392	9.9404	9.9404	10.5368
Kappa statistic	0.7413	0.7569	0.7434	0.7557
Mean absolute error	0.1325	0.1038	0.1162	0.1175
Root mean squared error	0.3135	0.3057	0.2872	0.2771
Relative absolute error	33.9329	26.5844	29.7465	30.0980
Root relative squared error	70.2094	68.4573	64.3278	62.0524

ตารางที่ 4.47 : ตารางแสดงผลลัพธ์จากตัวแบบจำแนกประเภทในข้อมูลกลุ่มที่ 2 กับข้อมูล spambase

Classifier's summary of train data set : group 2.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	14	14	14	14
percent	100.0000	100.0000	100.0000	100.0000
Incorrect number	0	0	0	0
percent	0.0000	0.0000	0.0000	0.0000
Kappa statistic	1.0000	1.0000	1.0000	1.0000
Mean absolute error	0.0000	0.0000	0.0000	0.0039
Root mean squared error	0.0000	0.0000	0.0000	0.0040
Relative absolute error	0.0000	0.0000	0.0000	6.2177
Root relative squared error	0.0000	0.0000	0.0000	6.3594

Classifier's summary of validate data set : group 2.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	6	6	6	6
percent	100.0000	100.0000	100.0000	100.0000
Incorrect number	0	0	0	0
percent	0.0000	0.0000	0.0000	0.0000
Kappa statistic	1.0000	1.0000	1.0000	1.0000
Mean absolute error	0.0000	0.0000	0.0000	0.0035
Root mean squared error	0.0000	0.0000	0.0000	0.0036
Relative absolute error	0.0000	0.0000	0.0000	5.6680
Root relative squared error	0.0000	0.0000	0.0000	5.6975

ตารางที่ 4.48 : ตารางแสดงผลลัพธ์จากตัวแบบจำแนกประเภทในข้อมูลกลุ่มที่ 3 กับข้อมูล spambase

Classifier's summary of train data set : group 3.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	1086	796	1058	1055
percent	97.3991	71.3901	94.8879	94.6188
Incorrect number	29	319	57	60
percent	2.6009	28.6099	5.1121	5.3812
Kappa statistic	0.9020	0.3682	0.8065	0.7830
Mean absolute error	0.0493	0.2848	0.0835	0.0582
Root mean squared error	0.1569	0.5301	0.2009	0.2148
Relative absolute error	17.8478	103.1818	30.2632	21.0714
Root relative squared error	42.2774	142.8201	54.1139	57.8722

Classifier's summary of validate data set : group 3.

Method	trees.J48	bayes.NaiveBayes	functions.Logistic	functions.MultilayerPerceptron
Correct number	750	596	755	747
percent	91.0194	72.3301	91.6262	90.6553
Incorrect number	74	228	69	77
percent	8.9806	27.6699	8.3738	9.3447
Kappa statistic	0.6711	0.3825	0.6924	0.6224
Mean absolute error	0.1104	0.2757	0.1058	0.0927
Root mean squared error	0.2941	0.5224	0.2540	0.2801
Relative absolute error	38.5446	96.2306	36.9394	32.3509
Root relative squared error	76.3613	135.6312	65.9458	72.7214

ตารางที่ 4.49 : ตารางแสดงผลลัพธ์จากตัวแบบจำแนกประเภทสำหรับวิธีการจำแนกประเภทแต่ละวิธีที่ได้จากการทดสอบข้อมูล spambase

Classifier's summary of all train data set
weka.classifiers.trees.J48

Correctly Classified Instances	1746	96.5708 %
Incorrectly Classified Instances	62	3.4292 %
Kappa statistic	0.9274	
Mean absolute error	0.0636	
Root mean squared error	0.1783	
Relative absolute error	13.3759 %	
Root relative squared error	36.5741 %	

Classifier's summary of all train data set
weka.classifiers.bayes.NaiveBayes

Correctly Classified Instances	1440	79.646 %
Incorrectly Classified Instances	368	20.354 %
Kappa statistic	0.6038	
Mean absolute error	0.2031	
Root mean squared error	0.4483	
Relative absolute error	42.7024 %	
Root relative squared error	91.9445 %	

Classifier's summary of all train data set
weka.classifiers.functions.Logistic

Correctly Classified Instances	1696	93.8053 %
Incorrectly Classified Instances	112	6.1947 %
Kappa statistic	0.8694	
Mean absolute error	0.0993	
Root mean squared error	0.2166	
Relative absolute error	20.8705 %	
Root relative squared error	44.4304 %	

Classifier's summary of all train data set
weka.classifiers.functions.MultilayerPerceptron

Correctly Classified Instances	1703	94.1925 %
Incorrectly Classified Instances	105	5.8075 %
Kappa statistic	0.8777	
Mean absolute error	0.0934	
Root mean squared error	0.2208	
Relative absolute error	19.6439 %	
Root relative squared error	45.2772 %	

ตารางที่ 4.50 : ตัวทำนายเหมาะสมที่ได้จากการทดสอบข้อมูล spambase

กลุ่มของข้อมูล	วิธีการพัฒนาตัวแบบของตัวทำนายเหมาะสม
0	วิธีการต้นไม้การตัดสินใจ
1	วิธีการจำแนกแบบเบย์อย่างง่าย
2	วิธีการต้นไม้การตัดสินใจ
3	วิธีการสมการถดถอยแบบโลจิสติก

ตารางที่ 4.51 : ผลลัพธ์ที่ได้จากการทดสอบข้อมูล spambase

Name of classifier method		n_CLAMP	tree n_J48	baye n_NaiveBayes	function n_Logistic	function n_MultilayerPerceptron
Correct.	Number	1291	1289	1121	1284	1270
	Percent	92.0171	91.8746	79.9002	91.5182	90.5203
Incorrect.	Number	112	114	282	119	133
	Percent	7.9829	8.1254	20.0998	8.4818	9.4797
Kappa statistic		0.8315	0.8286	0.6046	0.8213	0.8008
means absolute error		0.1007	0.1077	0.2007	0.1133	0.1217
Root means square error		0.2538	0.2748	0.4461	0.2483	0.2713

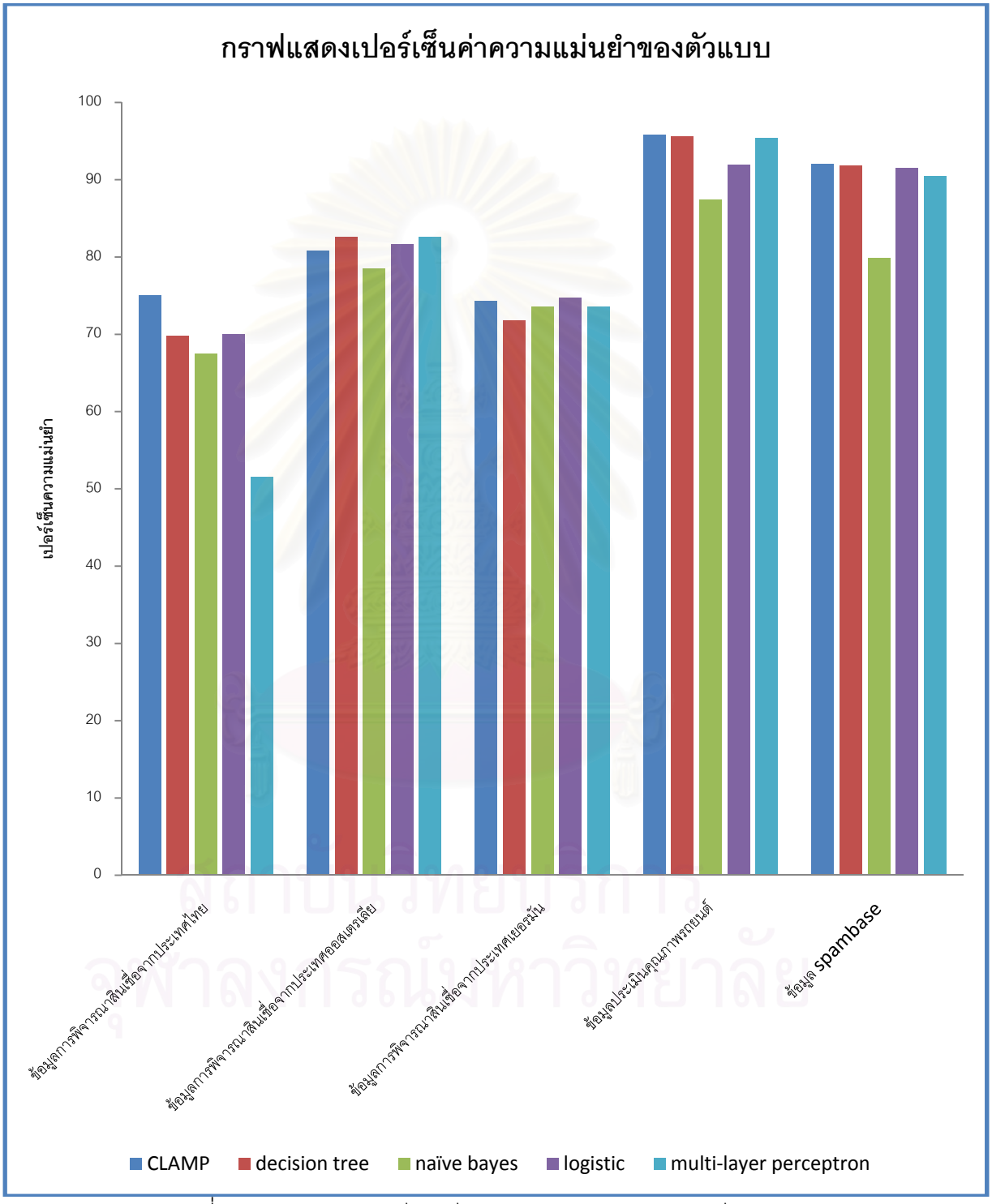
จากผลลัพธ์ที่ได้สามารถสรุปได้ดังนี้

1. ข้อมูลการพิจารณาสินเชื่อของธนาคารมีปัญหาความไม่สมดุลของคลาส เนื่องจาก ตัวแบบที่พัฒนาจากข้อมูลที่ยังไม่แก้ปัญหาความไม่สมดุลของคลาส ปรากฏว่าตัวแบบที่ดีที่สุดคือตัวแบบที่ทำนายว่าข้อมูลทุกระเบียนเป็นข้อมูลดีทั้งหมด ซึ่งไม่เหมาะกับการนำไปใช้ในการพิจารณาสินเชื่อได้

2. ข้อมูลที่ได้รับการแก้ปัญหาความไม่สมดุลของคลาส เมื่อทำการพิจารณาการพัฒนาตัวแบบปรากฏว่า ตัวแบบที่ได้จาก CLAMP มีความถูกต้องเท่ากับ 75.02 % ซึ่งมากกว่าการพัฒนาตัวแบบ โดยใช้วิธีการจำแนกประเภทต่างๆ ซึ่งได้ผลลัพธ์ไม่ถึง 70 %

3. จากข้อมูลการพิจารณาสินเชื่อจากประเทศออสเตรเลีย ข้อมูลการพิจารณาสินเชื่อจากประเทศเยอรมัน และข้อมูลการประเมินคุณภาพรถยนต์ เนื่องจากข้อมูลที่นำมาพิจารณามีจำนวนน้อยจึงทำให้มีจำนวนกลุ่มจากการวิเคราะห์การเกาะกลุ่มเท่ากับจำนวนกลุ่มที่กำหนดเป็นขั้นต่ำ คือ 2 กลุ่ม

4. เมื่อพิจารณาผลลัพธ์ที่ได้จากตัวอย่างข้อมูลจาก UCI ปรากฏว่า การพัฒนาตัวแบบจาก CLAMP ให้ผลลัพธ์ที่ดีกว่าหรือใกล้เคียงกับผลลัพธ์ที่ได้จากตัวแบบจำแนกประเภทเพียงหนึ่งตัว



รูปที่ 4.1 : แสดงกราฟเปรียบเทียบความแม่นยำของแต่ละวิธี

บทที่ 5
สรุปผลการทดลอง

งานวิจัยนี้นำเสนอการพัฒนาตัวแบบที่มีชื่อว่า การวิเคราะห์การเกาะกลุ่มของตัวทำนายหลากหลาย (CLAMP) โดยพัฒนาโปรแกรม CLAMP เพื่อทดสอบการทำงาน CLAMP ที่ใช้เป็นการนำข้อมูลพัฒนาตัวแบบมาวิเคราะห์การเกาะกลุ่มแบบค่าเฉลี่ยเอ็กซ์ แล้วจึงพัฒนาตัวแบบจำแนกประเภทสำหรับแต่ละกลุ่ม โดยวิธีการจำแนกประเภทที่พิจารณาคือ วิธีต้นไม้การตัดสินใจ สมการถดถอยแบบโลจิสติก วิธีการจำแนกแบบเบย์อย่างง่าย และวิธีการเพอร์เซพตรอนหลายชั้น เพื่อใช้กับข้อมูลประเมิน ให้ได้ตัวแบบตัวทำนายเหมาะสม แล้วจึงนำตัวแบบที่ได้มาทดสอบกับข้อมูลทดสอบ

ในการทดลองพบว่า จำนวนข้อมูลที่น้อยเกินไปทำให้การวิเคราะห์การเกาะกลุ่มแบบค่าเฉลี่ยเอ็กซ์ได้จำนวนกลุ่มต่ำสุดที่กำหนดให้ สำหรับงานวิจัยนี้กำหนดค่าจำนวนกลุ่มขั้นต่ำเป็น 2 กลุ่ม ทำให้ผลลัพธ์ที่ได้จาก CLAMP มีความถูกต้องมากกว่าหรือใกล้เคียงกับตัวแบบจำแนกประเภทเพียงตัวเดียว

ในกรณีที่ข้อมูลมีการแบ่งกลุ่มอย่างเหมาะสม ผลลัพธ์ที่ได้จาก CLAMP จะให้ค่าความแม่นยำสูงขึ้น ดังแสดงได้จากข้อมูลการพิจารณาสินเชื่อของธนาคารและข้อมูล spambase

ผลลัพธ์ของข้อมูลการพิจารณาสินเชื่อของธนาคาร ปรากฏว่าการใช้แผนการวิเคราะห์การเกาะกลุ่มของตัวทำนายหลากหลายให้ผลลัพธ์ที่ดีขึ้นอย่างชัดเจน คือ ผลลัพธ์ที่ได้จากตัวแบบจำแนกประเภทเพียงอย่างเดียว จะให้ค่าความถูกต้องประมาณ 70% แต่ CLAMP จะให้ค่าความถูกต้องประมาณ 75%

สำหรับผลลัพธ์ของข้อมูล spambase ถึงแม้ว่าการพัฒนาตัวแบบจำแนกประเภทเพียงอย่างเดียว จะให้ค่าความถูกต้องสูงถึง 91% แต่ CLAMP จะให้ค่าความถูกต้อง 92% ซึ่งเป็นผลลัพธ์ที่ดีขึ้นจากเดิม

กล่าวโดยสรุป การวิเคราะห์การเกาะกลุ่มของตัวทำนายหลากหลาย สามารถใช้ได้ดีกับข้อมูลที่มีลักษณะเฉพาะบางประการที่เหมือนกันเป็นกลุ่ม และมีจำนวนที่มากพอสำหรับการจัดกลุ่ม

วิทยานิพนธ์นี้ ผู้ทำวิทยานิพนธ์เลือกการวิเคราะห์การเกาะกลุ่มแบบค่าเฉลี่ยเอ็กซ์ ซึ่งผู้พัฒนาต่อสามารถเปลี่ยนวิธีการวิเคราะห์การเกาะกลุ่มเป็นวิธีการอื่นได้ นอกจากนี้ CLAMP ต้องการตัวแปรต่อเนื่องทั้งหมดก่อนการประมวลผล ยกเว้นตัวแปรเป้าหมายผู้พัฒนาต่อ สามารถขยาย CLAMP ให้ยอมรับตัวแปรต่อเนื่อง และไม่ต่อเนื่องต่อไปได้ด้วย

รายการอ้างอิง

1. Elizabeth, M. Handbook of credit scoring. United States of America: Glenlake publishing company, 2001.
2. Naeem, S. Credit risk scorecards: developing and implementing intelligent credit scoring. United States of America: John & Sons, Inc., 1969.
3. David, O. and Yong, S. Introduction to business data mining. Singapore: McGraw-Hill companies ,2007.
4. Naeem, S. Credit scorecard development and implementation course notes (Course notes). United States of America : SAS Software (Thailand), 2005.
5. Lyn, T., David, E. and Jonathan, C. Credit scoring and its applications. Philadelphia: Society for industrial and applied mathematics, 2002.
6. ธนาคารไทยพาณิชย์ (มหาชน) จำกัด, แบบแสดงรายการข้อมูลประจำปี 2547. Thailand, 2002.
7. ธนาคารกรุงเทพ (มหาชน) จำกัด, รายงานประจำปี 2549. Thailand, 2006
8. วิศิษฐ์ ลิ้มสมบุญชัย, การวิเคราะห์แบบจำลองคะแนนสินเชื่อสำหรับตลาดการเงินในชนบทไทย (An Analysis of Credit Scoring Model for Rural Financial Market in Thailand). ภาควิชาเศรษศาสตร์เกษตรและทรัพยากร คณะเศรษฐศาสตร์ มหาวิทยาลัยเกษตรศาสตร์, 2007.
9. Ian, W and Eibe, F. Data mining: Practical machine learning tool and techniques. Second edition. Morgan Kaufmann series in data management systems. United States of America : Morgan Kaufmann publishers, 2005.
10. แสงจันทร์ เรืองอ่อน, ประณต บุญไชยอภิสิทธิ์, ประสงค์ ปราณีตพลกรัง, ปิยวัฒน์ จิรพงษ์สุวรรณ. ดาต้าไมน์นิ่งเชิง XML (XML-based data mining). Thailand: Nectec, 2002.
11. Hassan, A. Data Mining (PowerPoint). Iran: Sharif University of Technology, 2007.
12. Kurt, T. An Introduction to Data Mining (PowerPoint). Virginia of America, 2007.
13. Herb, E. Building Profitable Customer Relationships with Data Mining. Maryland of America: Two Crows Consulting.
14. Adalbert, W. Data Mining Tasks [online]. (n.d.). Avalible from : http://www.quantlet.com/mdstat/scripts/csa/html/node205.html [cited 25 January 2008]

15. Wikipedia. Cluster analysis [online]. (n.d.). Avalible from : http://en.wikipedia.org/wiki/Data_clustering [cited 25 January 25 2008].

16. Dan, A. and Andrew M. X-means: extending K-means with efficient estimation of the number of clusters. Seventeenth International Conference on Machine Learning, Pittsburgh, 2000.

17. Yasir, J. and A, B. Emitter recognition based on modified X-means clustering. IEEE: International conference on emerging technology. Isalamabad, 2005.

18. George, J. and Pat, L. Estimating continuous distributions in bayesian classifiers. eleventh conference on uncertainty in artificial intelligence, San Mateo, 1995.

19. Daniel, L. and Pedro, D. Naive Bayes models for probability estimation. ACM International Conference Proceeding Series: Proceedings of the 22nd international conference on Machine learning. Germany: 2005

20. Wasan, P. and Kamthorn, P. Naive Bayes Model [online]. Thailand: Network technology lab (NTL). Avalible from : http://wiki.nectec.or.th/ntl/Project/NaiveBayesModel [cited 25 January 2008].

21. le, C., S., van H., J.C. . Ridge estimators in logistic regression. Applied Statistics, 1992.

22. Yong, W. and Ian, W. Modeling for optimal probability prediction. Proceedings of the Nineteenth International Conference on Machine Learning. San Francisco: 2002.

23. Ross, Q. C4.5: programs for machine learning. Morgan Kaufmann Publishers, San Mateo, 1994.

24. XLMiner. XLMiner help[online]. (n.d.). Avalible from : http://www.resample.com/xlminer/help/Index.htm [cited 2 February 2008].

25. Wikipedia. Bayesian information criterion [online]. (n.d.). Avalible from : http://en.wikipedia.org/wiki/Bayesian_information_criterion [cited 25 January 2008].

26. Richard, K. and Eibe F. WEKA explorer user guide for version 3-4-3. University of Waikato, 2004.

27. The UCI Machine Learning Repository. The UCI machine learning repository: data sets[online]. (n.d.). Avalible from : http://archive.ics.uci.edu/ml/datasets.html [cited 2 February 2008].

28. The UCI Machine Learning Repository. Statlog (Australian Credit Approval) Data Set. (n.d.): Ross Quinlan.

29. The UCI Machine Learning Repository. Statlog (German Credit Data) Data Set. Institute Statistic and Econometrics University of Hamburg: Professor Dr. Hans Hofmann, 1994.

30. The UCI Machine Learning Repository. Car Evaluation Data Set. (n.d.): Marko Bohanec and Blaz Zupan, 1997.

31. The UCI Machine Learning Repository. Spambase Data Set. Hewlett-Packard Labs: Mark Hopkins, Erik Reeber, George Forman and Jaap Suermondt, 1999.

32. WEKA. Weka 3 Data minig with open source machine learning software in Java[online]. (n.d.). Avalible from : http://www.cs.waikato.ac.nz/ml/weka/ [cited 2 February 2008].

33. Netbeans. Welcome to netbeans[online]. (n.d.). Avalible from : http://www.netbeans.org/ [cited 2 February 2008].

ประวัติผู้เขียนวิทยานิพนธ์

นายวิวัช ชลไชยะ เกิดเมื่อวันที่ 14 ธันวาคม พุทธศักราช 2524 สำเร็จการศึกษาระดับปริญญาวิทยาศาสตรบัณฑิต สาขาวิทยาการคอมพิวเตอร์ จากภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย เมื่อปีการศึกษา 2547 และเข้าศึกษาต่อในหลักสูตรปริญญามหาบัณฑิต สาขาวิทยากการคณนา ภาควิชาคณิตศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย เมื่อปีการศึกษา 2547