การพัฒนาระบบการแบ่งกลุ่มข้อมูลด้วย ทูสเต็ปคลัสเตอร์ริ่ง

DEVELOPMENT OF CLUSTERING SYSTEM USING TWO STEP CLUSTERING

โดย

ทวีชัย ปิยะตานนท์

THAWEECHAI PIYATANON

อาจารย์ที่ปรึกษา

รศ.ดร.วรพจน์ กรีสุระเดช

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ภาคฤดูร้อน ปีการศึกษา 2553

PROPOSAL OF DEVELOPMENT OF CLUSTERING SYSTEM USING TWO STEP CLUSTERING

THAWEECHAI PIYATANON

A REPORT SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS OF THE COURSE

SYSTEM DEVELOPMENT PROJECT

MASTER OF SCIENCE PROGRAM IN INFORMATION TECHNOLOGY

FACULTY OF INFORMATION TECHNOLOGY

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

SUMMER / 2010

ใบรับรองโครงการพัฒนาระบบงาน (System Development Project)

เรื่อง

การพัฒนาระบบการแบ่งกลุ่มข้อมูลด้วย ทูสเต็ปคลัสเตอร์ริ่ง

DEVELOPMENT OF CLUSTERING SYSTEM USING TWO STEP CLUSTERING

นายทวีชัย ปิยะตานนท์

รหัสประจำตัว 49066428

ขอรับรองว่ารายงานฉบับนี้ ข้าพเจ้าไม่ได้คัดลอกมาจากที่ใด

รายงานฉบับนี้ได้รับการตรวจสอบและอนุมัติให้เป็นส่วนหนึ่งของ

การศึกษาวิชาโครงการพัฒนาระบบงาน หลักสูตรวิทยาศาสตรมหาบัณฑิต (เทคโนโลยีสารสนเทศ)

ภาคฤดูร้อน ปีการศึกษา 2553

.....อาจารย์ที่ปรึกษา

(รศ.ดร.วรพจน์ กรีสุระเดช)

.....กรรมการสอบ

(ผศ.ดร.พรฤดี เนติโสภากุล)

.....กรรมการสอบ

(รศ.ดร.อาริต ธรรมโน)

หัวข้อ	การพัฒนาระบบการแบ่งกลุ่มข้อมูลด้วย ทูสเต็ปคลัสเตอร์ริ่ง
นักศึกษา	นายทวีชัย ปิยะตานนท์
รหัสนักศึกษา	49066428
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2553
อาจารย์ที่ปรึกษา	รศ.ดร.วรพจน์ กรีสุระเดช

บทคัดย่อ

ในโลกของการดำเนินธุรกิจในปัจจุบันยากที่จะหลีกเลี่ยงการแข่งขัน ไม่ว่าจะเนินธุรกิจใดๆ ก็จำเป็นเข้าสู่เวทีการแข่งขัน การเก็บรวมข้อมูลข่าวสารรวมไปถึงการนำข้อมูลที่เก็บไว้อยู่มาใช้ให้เกิดประโยชน์จึงเป็นส่วนสำคัญเพื่อนำไปใช้ในการวางแผนกลยุทธ์และสร้างความได้เปรียบเหนือคู่แข่ง ดังนั้นการนำเครื่องมือดาต้าไมนิ่งมาใช้จึงเป็นอีกทางเลือกที่สามารถช่วยปรับปรุงการทำงานให้เกิดประสิทธิภาพมากยิ่งขึ้น

โครงงานนี้จึงได้เสนอกระบวนการทำเหมืองข้อมูล โดยอัลกอริทึมที่นำมาใช้ในการทำเหมืองข้อมูลคือ Two Step Clustering เพื่อแบ่งกลุ่มข้อมูลโดยใช้หลักงานทำงาน 2 ขั้นตอน ขั้นตอนแรกจะทำการสร้าง CF-Tree ขึ้นมาเพื่อเป็นการกำหนดกลุ่มย่อยแล้วนำตัวแทนของกลุ่มย่อยที่ได้ไปเป็นค่าเริ่มต้นในขั้นตอนที่ 2 โดยขั้นตอนที่ 2 จะทำงานโดยการแบ่งกลุ่มตัวแทนของกลุ่มย่อยในขั้นตอนที่ 1 โดยใช้อัลกอริทึม K-Mean หลังจากนั้นจะทำการคำนวณหาจำนวนกลุ่มที่เหมาะสมเพื่อกำหนดกลุ่มที่เหมาะสมให้กับข้อมูลนั้น

Title	Development of Data Clustering System Using Two Step Clustering
Student	Mr. Thaweechai Piyatanon
Student ID.	49066428
Degree	Master of Science
Program	Information Technology
Major	Information Science
Academic Year	2010
Advisor	Assoc. Prof. Dr. Worapoj Kreesuradej

ABSTRACT

In the modern world of the business operation, it is difficult to avoid competitions. Any activity needs to enter into a competition. Collecting information and using information are important parts which are used for the best benefit to plan a strategy and to gain an advantage with a competitor. Data mining is one of the choice that can improve the effectiveness of systematic operation.

This project offers data-mining process. The algorithm used in data mining is a Two Step Clustering for grouping data using primary job function two steps The first step is to create a CF-Tree up to a specified sub-groups and the representatives of Small groups that have gone by default in the Step 2, Step 2 is done by segment represents a small group in step 1 using the algorithm K-Mean and then will calculate the amount The right to define the appropriate data.

กิตติกรรมประกาศ

โครงการศึกษาพัฒนาระบบการแบ่งกลุ่มข้อมูลโดยใช้ อัลกอริทึม ทูสเต็ปคลัสเตอร์ริ่ง จะสำเร็จลุล่วงไปไม่ได้เลย ถ้าไม่ได้รับการช่วยเหลือและแรงสนับสนุนจากบุคคลสำคัญหลายท่าน ดังต่อไปนี้

ข้าพเจ้าขอกราบขอบพระคุณ รศ.ดร. วรพจน์ กรีสุระเดช ซึ่งเป็นอาจารย์ที่ปรึกษาโครงการนี้ ที่ให้ความกรุณาให้คำแนะนำ และปรึกษา ข้าพเจ้ารู้สึกซาบซึ้งและขอขอบพระคุณเป็นอย่างสูง

ขอกราบพระคุณคณาจารย์ของสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ทุก ๆ ท่านที่ได้ประสิทธิ์ประสาทวิชาให้กับข้าพเจ้า

ขอขอบคุณ นายณัฐพงษ์ ฉัตรแก้วโพธิ์ทองที่เป็นที่ปรึกษาที่ดีในการพัฒนาระบบ

ขอขอบคุณเพื่อน ๆ พี่ ๆ น้อง ๆ ทุกคนที่ให้คำแนะนำ

สุดท้ายนี้ข้าพเจ้าขอกราบขอบพระคุณ บิดา มารดา และครอบครัวของข้าพเจ้าที่เป็นกำลังใจ และให้การสนับสนุนในทุกเรื่องๆ ทำให้ข้าพเจ้าสามารถทำโครงการพัฒนาระบบงานฉบับนี้สำเร็จลุล่วงด้วยดี

คุณค่าและประโยชน์จากโครงการพัฒนาระบบงานฉบับนี้ ข้าพเจ้าขอมอบแด่ผู้มีพระคุณทุกท่าน

ทวีชัย ปิยะตานนท์

สารบัญ

สารบัญ (ต่อ)

หน้า

สารบัญรูป

หน้า

สารบัญรูป(ต่อ)

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันเทคโนโลยีสารสนเทศเข้ามามีบทบาทอย่างกว้างขวางในทุกวงการ และส่งผลให้กลายมาเป็นเครื่องมือสำคัญในการทำงานทุกด้าน นับตั้งแต่ด้านเศรษฐกิจ พานิชยกรรม อุตสาหกรรม การวิจัยและการพัฒนาตลอดจนด้านการเมืองและราชการ ทั้งนี้ได้มีการนำเทคโนโลยีสารสนเทศมาประยุกต์ใช้กับข้อมูลที่มีอยู่ภายในองค์กรนั้นๆ เพื่อพัฒนาเป็นระบบการตัดสินใจที่ช่วยวิเคราะห์ข้อมูลต่างๆ เพื่อนำสารสนเทศที่ได้มาใช้ประโยชน์ในการตัดสินใจเพื่อให้บรรลุวัตถุประสงค์ขององค์กร เนื่องจากหากผู้บริหารมีข้อมูลที่มีประสิทธิเพื่อใช้ในการประกอบการตัดสินใจของผู้บริการ ผลลัพธ์ของการตัดสินใจย่อมมีประสิทธิภาพต่อหน่วยงานด้วยเช่นกัน

การศึกษาและพัฒนาโครงงานนี้จึงได้นำเทคโนโลยีการทำเหมืองข้อมูลมาใช้ในการวิเคราะห์ข้อมูล เพื่อใช้วิเคราะห์ข้อมูลและมุ่งเน้นอธิบายความสัมพันธ์ของข้อมูลที่ต้องการ เพื่อนำความสัมพันธ์ของข้อมูลมาใช้ให้เกิดประโยชน์มากยิ่งขึ้น ทั้งนี้ได้มีการนำข้อมูลเหล่านั้นมาใช้ประโยชน์อย่างเต็มที่ โดยกระบวนการในการวิเคราะห์ข้อมูลในการทำการแบ่งกลุ่มข้อมูล

โดยในการศึกษาโครงการพัฒนาระบบนี้ได้มุ่งเน้นไปที่การจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม Twostep Clustering

1.2 ความมุ่งหมายและวัตถุประสงค์

โครงการพัฒนาระบบงานเรื่องการศึกษาอัลกอริทึมในการแบ่งกลุ่มข้อมูลมีวัตถุประสงค์ดังนี้

1. ศึกษาเทคโนโลยีการทำเหมืองข้อมูล เพื่อนำมาใช้ออกแบบและพัฒนาระบบ
2. ศึกษาวิธีการและขั้นตอนในการทำเหมืองข้อมูลโดยที่เน้นไปที่Twostep Clustering
3. สร้างระบบแบ่งกลุ่มด้วยTwostep Clustering

1.3 ขอบเขตการวิจัย

การทำงานของระบบการแบ่งกลุ่มด้วย Twostep Clustering มีขอบเขตการศึกษาและพัฒนาดังนี้

1. เตรียมข้อมูลเพื่อเข้าสู่กระบวนการประมวลผล
2. สร้างอัลกอริทึมเพื่อนำกลุ่มของข้อมูลมาประมวลผล

3. นำผลลัพธ์ของการประมวลผลกลุ่มข้อมูลมาแสดงผล

1.4 ขั้นตอนของการศึกษา

การพัฒนาโครงงานนี้มีการนำข้อมูลจากแหล่งข้อมูลที่น่าเชื่อถือ และต้องมีการจัดเตรียมการศึกษา ออกแบบ และพัฒนา โดยมีขั้นตอนดังนี้

1. ศึกษาทฤษฎีที่เกี่ยวข้องเกี่ยวกับการทำเหมืองข้อมูลเพื่อนำมาประยุกต์ใช้ในการพัฒนาระบบ
2. การจัดการให้ข้อมูลอยู่รูปแบบที่เหมาะสมและมีการจัดกลุ่มเปลี่ยนแปลงค่าเพื่อให้สามารถนำไปวิเคราะห์หาผลลัพธ์ได้ตรงตามวัตถุประสงค์
3. สร้างรูปแบบจำลองเพื่อหาความสัมพันธ์ของข้อมูลที่ผ่านการจัดรูปแบบให้ถูกต้องแล้ว โดยใช้Twostep Clustering ในการวิเคราะห์เพื่อหาผลลัพธ์
4. จัดเตรียมรายงานผลความสัมพันธ์ที่ผ่านการประมวลผลจากอัลกอริทึมในรูปแบบรายงานที่ผู้ใช้งานสามารถเข้าใจง่าย
5. ทดสอบการใช้งานโปรแกรม
6. สรุปผลการศึกษาและข้อเสนอแนะ

1.5 ประโยชน์ที่คาดว่าจะได้รับ

ทำให้ทราบวิธีการพัฒนาระบบที่มีการนำกระบวนการการแบ่งกลุ่มข้อมูล(Clustering) โดยการจัดกลุ่มข้อมูลแบบ Twostep Clustering มาประยุกต์ใช้กับการจัดกลุ่มข้อมูลในระบบได้มีประสิทธิภาพในการจำแนกกลุ่มที่แม่นยำและมีความเหมาะสมที่จะนำรูปแบบใดรูปแบบหนึ่งไปใช้เป็นเครื่องมือในการจำแนกกลุ่มข้อมูล

บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

ในหัวข้อนี้จะกล่าวถึงทฤษฎีพื้นฐานต่างๆ ที่เกี่ยวข้องในการพัฒนาระบบ เนื้อหาในบทนี้จะกล่าวถึง ทฤษฎีดาต้าไมนิ่ง อัลกอริทึมทูสเต็ปคลัสเตอร์ริ่งเป็นอัลกอริทึมที่ใช้แบ่งกลุ่มข้อมูล โดยมีรายละเอียดที่เกี่ยวข้องกับการพัฒนาระบบงาน ดังนี้

2.1 แผนภูมิต้นไม้แบบCF

Cf tree หรือ Clustering Feature เป็นแกนหลักของอัลกอริทึมของ Birch โดยเป็นการแบ่งกลุ่มแบบเพิ่ม โดยในแต่ละข้อมูลสามารถเก็บค่าได้หลายมิติ โดยในแต่ละ CF มีการกำหนดค่าตัวแปรหลัก 3 ค่า : CF = (N , LS , SS)

โดยที่ N คือจำนวนข้อมูลในกลุ่มย่อย

LS คือ ค่าผลรวมเชิงเส้น

SS คือ ค่าผลรวมกำลังสอง

CF มีคุณสมบัติการบวก

สมมติให้ CF1 = (N1 , LS1 , SS1) และ CF2 = (N2 , LS2 , SS2) เมื่อ CF ทั้งสองมารวมกันจะทำให้ได้

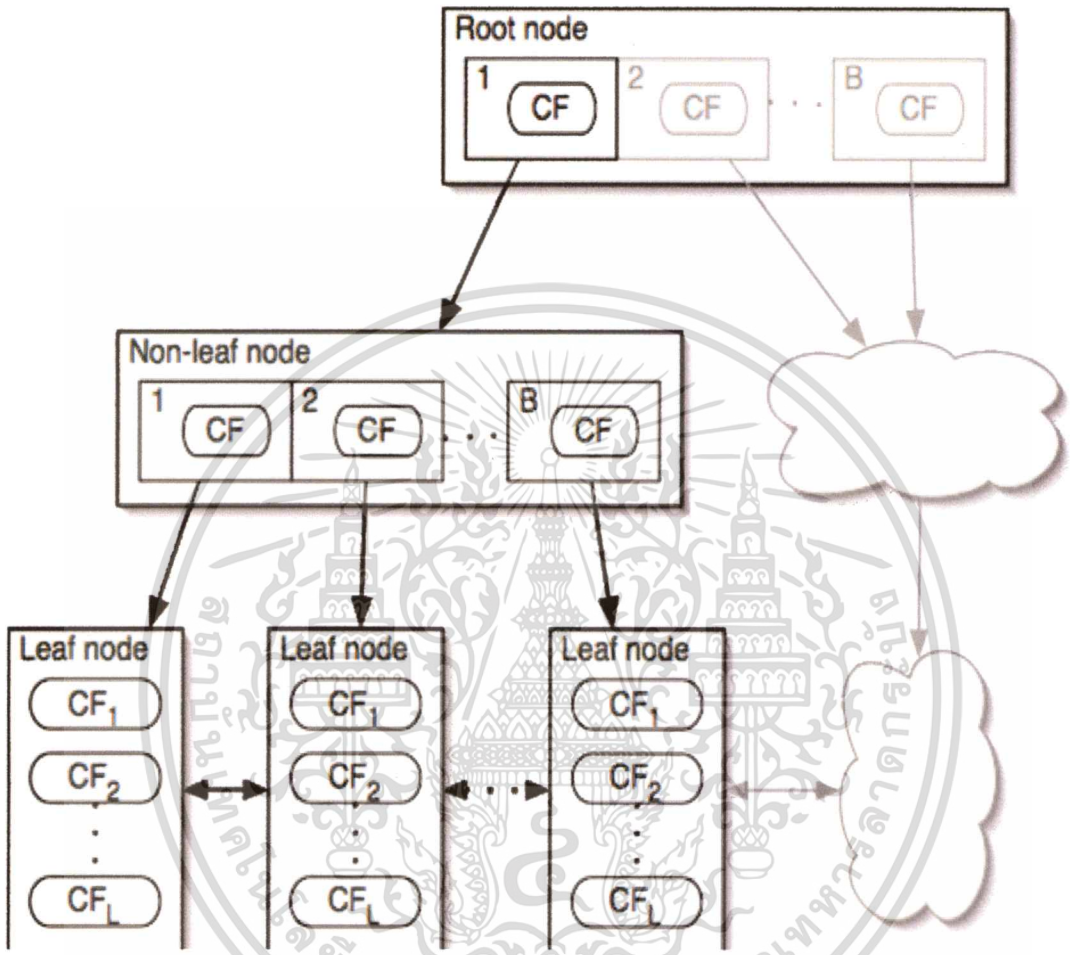
$$CF1+CF2 = (N1+N2 , LS1+LS2 , SS1+SS2)$$

ส่วนประกอบของ CF tree

ตารางที่ 2.1 แสดงความหมายของส่วนประกอบของ CF-Tree

รายการ	ความหมาย
Root node	โหนดบนสุด เป็นจุดเริ่มต้นในการเข้าสู่ Tree ของข้อมูล
Non leaf node	โหนดที่ไม่ใช่ Leaf node ใช้นำทางและเป็นตัวแทนของ Leaf node
Leaf node	โหนดที่ไม่มีโหนดลูก ทำหน้าที่เก็บข้อมูลสมาชิก
Level of Tree	ความสูงของแผนภูมิต้นไม้

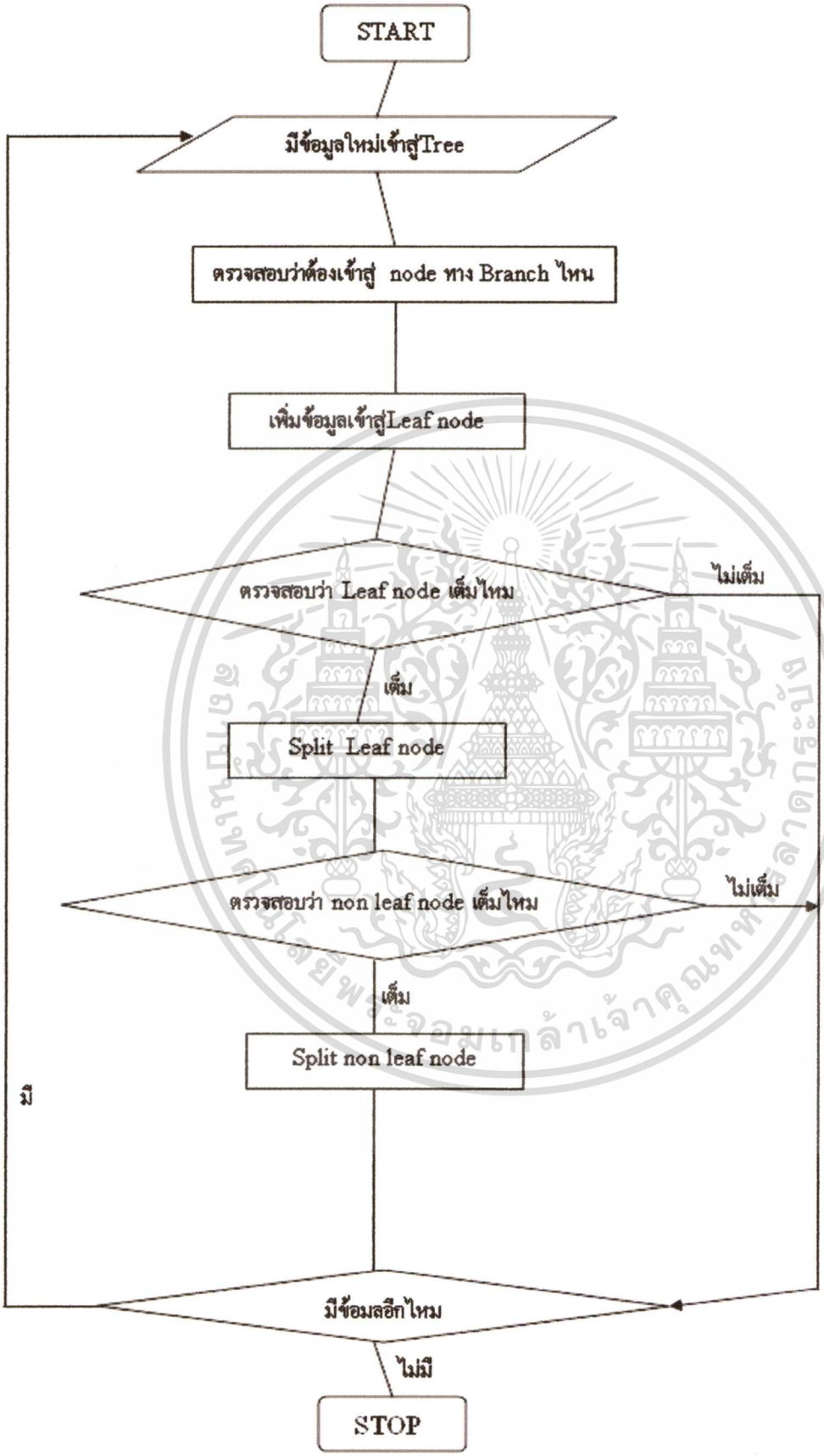
โดยจำนวน non-leaf node มีจำนวน B- Branch ทำให้ ภายใน non leaf node มีจำนวน B รายการย่อยในแต่ละCF Tree และจำนวน leaf node เป็น จำนวน L ทำให้ภายในแต่ละรายการจะมีจำนวนข้อมูลได้ L รายการ



รูปที่ 2.1 แสดงลักษณะทั่วไปของ CF-Tree

การเพิ่มข้อมูลลงไปใน CF Tree

การเพิ่มข้อมูลลงไปใน CF-Tree ทำได้โดยการเพิ่มข้อมูลลงไปในCF tree ทีละตัวเพิ่มข้อมูลเข้าสู่ Tree ผ่านทาง Root node หลังจากนั้นจะทำการวัดระยะหาเส้นทางไปตาม Branch เพื่อไปยังข้อมูลที่มีค่าใกล้เคียงโดยในการวัดระยะห่างเพื่อกำหนดเส้นทางการเข้าสู่ Tree ในที่นี้จะใช้การวัดระยะห่าง Euclidean Distance ทำการวัด โดยข้อมูลจะถูกเก็บเข้าสู่ leaf node โดยมีลำดับขั้นตอนการทำงานดังรูป 2.2



รูปที่ 2.2 แสดงการทำงานการเพิ่มข้อมูลเข้าสู่ CF-Tree

มีการนำข้อมูลเข้าสู่ TREE ทาง Root node หรือ non leaf node ที่อยู่บนสุด ทำการตรวจสอบว่า เข้าสู่โหนดถัดไปผ่าน branch ใดใน non leaf node จนเข้าสู่ leaf node ที่มีความใกล้เคียงที่ โดยสมการที่ใช้ในการวัดระยะห่างเพื่อหาความใกล้เคียงใช้สมการ ระยะห่าง Euclidean Distance

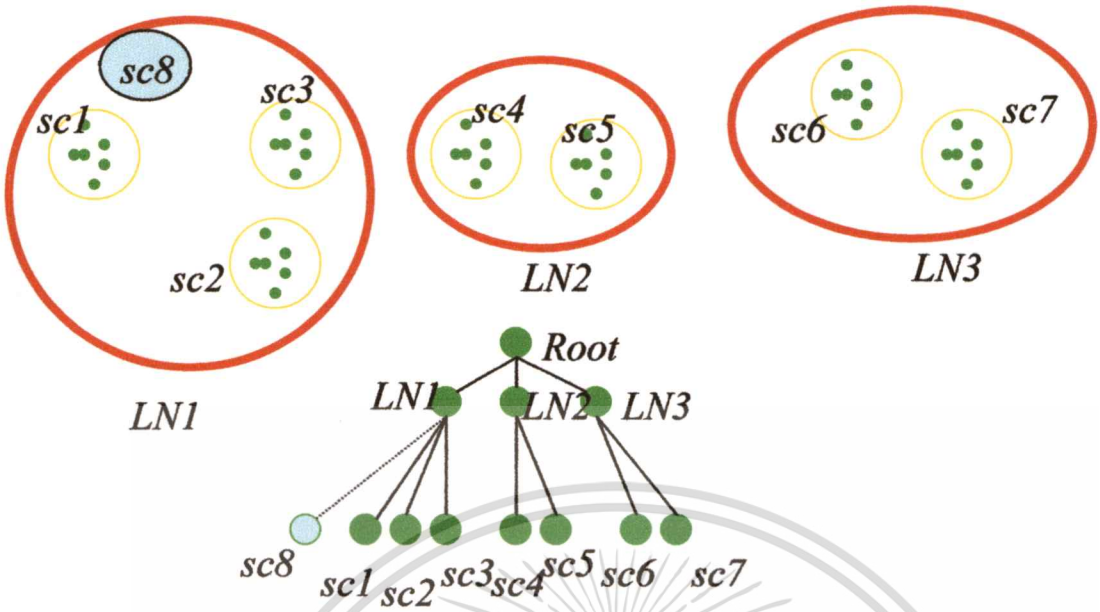
$$\text{distance}(\vec{X}_{01}, \vec{X}_{02}) = \sqrt{(\vec{X}_{01} - \vec{X}_{02})^2} \qquad (2.1)$$

ใช้สมการ

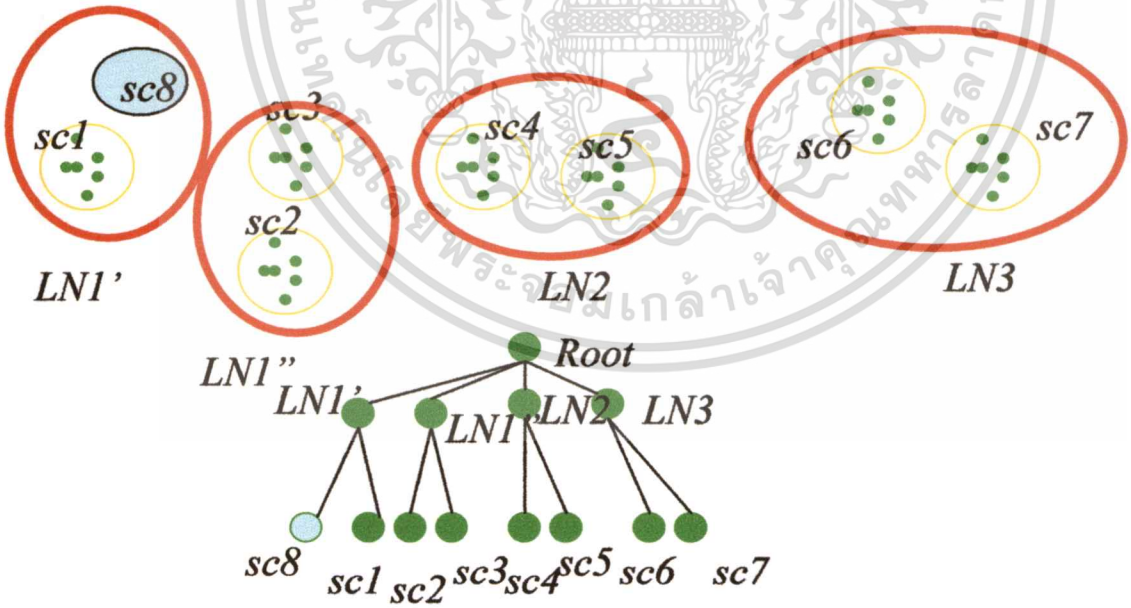
$$X_0 = \frac{\sum_{i=1}^{N} X_i}{N} \qquad (2.2)$$

ในหารหาจุดศูนย์กลางของกลุ่ม

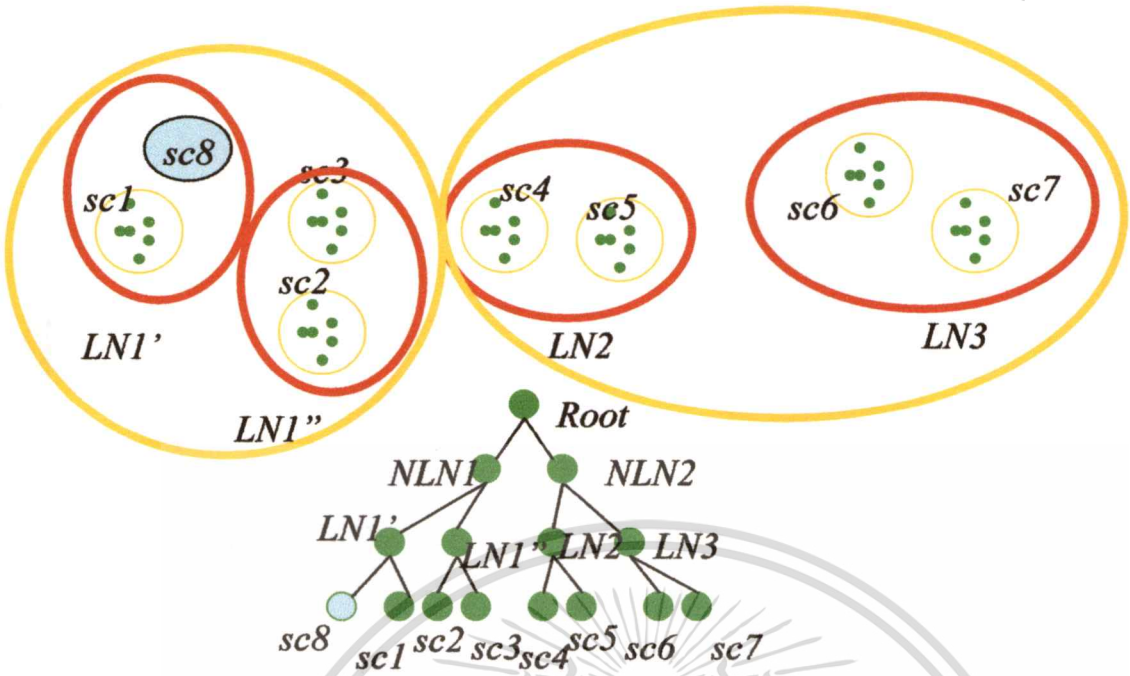
เมื่อเพิ่มข้อมูลเข้าสู่ leaf node หลังจากนั้นจะทำการตรวจสอบว่าข้อมูลใน leaf node เกินค่า L ที่กำหนดไว้หรือไม่ ถ้าไม่เกินก็จะทำการเพิ่มข้อมูลตัวถัดไปจนข้อมูลหมด แต่ถ้าเกินก็ทำการ Split โดยวิธีการ Split จะทำการ Split โดยการวัดระยะห่างของข้อมูลภายในกลุ่มที่จะ split เพื่อหาข้อมูล 2 ตัวที่มีระยะห่างมากที่สุดมาเป็นจุดเริ่มต้นของ leaf node ใหม่ในการ split จากนั้นนำข้อมูลแต่ตัวในกลุ่มเดิมแบ่งเข้าสู่กลุ่มใหม่โดยการหาระยะห่างจากจุดเริ่มต้นของ2 กลุ่มเพื่อระบุว่าสมาชิกแต่ละตัวควรอยู่กับสมาชิกกลุ่มใหม่กลุ่มใด จากการ split จะทำให้ จำนวน Branch ใน non leaf node มีขนาดเพิ่มขึ้นจึงทำให้ต้องมีการตรวจสอบว่า สมาชิกใน non leaf node เกินขนาดของ Branch ที่กำหนดไว้ไหม ถ้าไม่เกินก็จะทำการเพิ่มข้อมูลตัวถัดไปจนข้อมูลหมด แต่ถ้าเกินก็ทำการ Split non leaf node โดยการ split ของ non leaf node ทำคล้ายกับ การsplit ของ leaf node คือวัดระยะห่างของจุดศูนย์กลางของข้อมูลภายในกลุ่มที่จะ split เพื่อหาข้อมูล 2 ตัวที่มีระยะห่างมากที่สุดมาเป็นจุดเริ่มต้นของ leaf node ใหม่ในการ split จากนั้นนำจุดศูนย์กลางของข้อมูลแต่ตัวในกลุ่มเดิมแบ่งเข้าสู่กลุ่มใหม่โดยการหาระยะห่างจากจุดเริ่มต้นของ2 กลุ่มเพื่อระบุว่าสมาชิกแต่ละตัวควรอยู่กับสมาชิกกลุ่มใหม่กลุ่มใด โดยจะทำการตรวจสอบขึ้นไปเรื่อยตามลำดับขั้นถ้ามีการเพิ่มขึ้นของ non leaf node เมื่อมีการ split ในส่วนของ non leaf node แล้วจะมีการทำ Balance Tree โดยการ Balance Tree นั้นจะพิจารณาข้อมูลศูนย์กลางของ non leaf node ที่อยู่ใกล้กันว่าสามารถรวมกันโดยที่ ไม่ทำให้เกินค่า B หรือค่า L ที่กำหนดไว้



รูปที่ 2.3 แสดงการเพิ่มข้อมูล(sc8)แล้วทำให้ขนาดของ non leaf node เกิน (กำหนด B=3)



รูปที่ 2.4 แสดงการsplit เมื่อ non leaf node เกินในกรณีนี้ต้องทำการsplit non leaf node ขั้นต่อไป



รูปที่ 2.5 แสดงการ split ในขั้นถัดมาจะทำให้ ลำดับชั้นของ Tree สูงขึ้น

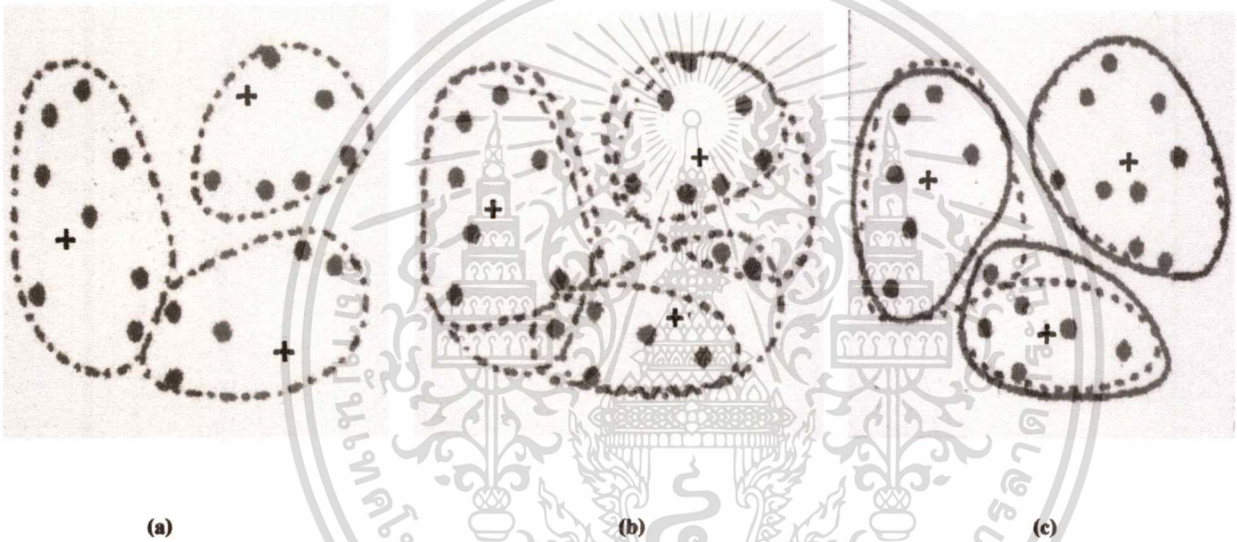
รูปที่ 2.3-2.5 แสดงการเพิ่มข้อมูลเข้าสู่ Tree โดยข้อมูลที่เข้ามาใหม่ (SC8) ในรูป 2.3 ทำให้ LN1 มีขนาดเกินกว่า 3 ทำให้ต้อง Split หลังจาก Split แล้วทำให้เกิด LN1' และ LN1" รูป 2.4 ทำให้จำนวนเกิน ต้องทำการ Split หลังจาก Split ในขั้นตอนนี้ จะทำให้ ระดับชั้นของTree (Level) เพิ่มสูงขึ้นดังรูป 2.5

2.2 K- means

K-means อัลกอริทึมเป็นอัลกอริทึมหนึ่งใน Partitioning Methods เป็นการจัดกลุ่มข้อมูลที่มีความคล้ายคลึงกันอยู่กลุ่มเดียวกันโดยการใช้ค่าเฉลี่ยของข้อมูลที่ถูกจัดให้อยู่กลุ่มเดียวกันเป็น โดยตัวแทนของกลุ่มนั้น การจัดกลุ่มข้อมูลจะทำการแบ่งข้อมูลออกเป็น k กลุ่ม โดยค่า k ในที่นี้จะต้องกำหนดขึ้นมาตามขนาดของกลุ่มที่เราสนใจ และใช้ค่าเฉลี่ยของข้อมูลสร้างจุดศูนย์กลางใหม่ โดยจะพยายามปรับปรุงการแบ่งกลุ่มข้อมูล จากการเคลื่อนย้ายข้อมูลจากกลุ่มจนกระทั่งได้ข้อมูลที่มีความคล้ายคลึงกันอยู่ในกลุ่มเดียวกันและข้อมูลที่มีความแตกต่างกันอยู่คนละกลุ่มกันจนกระทั่งได้ข้อมูลที่มีความคล้ายคลึงกันอยู่ในกลุ่มเดียวกันและข้อมูลที่มีความแตกต่างกันอยู่คนละกลุ่มกัน

การทำงานของ K- means

1. กำหนดจำนวน k กลุ่มเพื่อเริ่มจุดศูนย์กลางของข้อมูล
2. สุ่มเลือกข้อมูลมา k ข้อมูลตามจำนวนกลุ่มเพื่อเป็นจุดศูนย์กลางของข้อมูล
3. กำหนดสมาชิกของข้อมูลให้อยู่ในกลุ่มข้อมูล โดยการวัดระยะห่างจากสูตร (Euclidean distance) ของสมาชิกกับศูนย์กลางของกลุ่ม หากสมาชิกอยู่ใกล้จุดศูนย์กลางของกลุ่มใดมากสุดก็นำไปรวมกับกลุ่มนั้น
4. ในแต่ละกลุ่มคำนวณค่าเฉลี่ยใหม่เพื่อสร้างจุดศูนย์กลางของกลุ่มใหม่
5. ทำขั้นตอน 3-4 ซ้ำจนกระทั่งศูนย์กลางของกลุ่มไม่เปลี่ยนแปลง



รูปที่ 2.6 ขั้นตอนการทำงานของ K- means

จากรูป อธิบายการทำงานของ K- means ได้ดังนี้ ภาพ (a) เราเลือกข้อมูลมา 3 ตัวเพื่อเป็นจุดศูนย์กลางของกลุ่มข้อมูลที่เราเลือกเป็นจุดศูนย์กลางของกลุ่มข้อมูลถูกแทนที่ด้วยเครื่องหมาย "+" หลังจากนั้นวัดระยะห่างจากแต่ละข้อมูลกับจุดศูนย์กลางของกลุ่มนั้น ข้อมูลใดอยู่ใกล้ศูนย์กลางไหนมากที่สุดก็จะไปอยู่ในกลุ่มนั้น การกระจายสมาชิกข้อมูลแทนที่ด้วย จุดสีดำ ภาพ (b) เมื่อทำการรวมกลุ่มเรียบร้อยแล้ว จะคำนวณค่าเฉลี่ยของข้อมูลแต่ละกลุ่มใหม่เพื่อที่จะปรับปรุงศูนย์กลางของแต่ละกลุ่ม ถ้าข้อมูลใกล้เคียงกับกลุ่มอื่นก็จะทำการย้ายกลุ่ม ภาพ(c) กระบวนการนี้จะทำการวนซ้ำจนกระทั่งศูนย์กลางกลุ่มไม่เปลี่ยนแปลง

Algorithm: *k*-means. The *k*-means algorithm for partitioning based on the mean value of the object in the cluster.
Input: The number of cluster *k* and a database containing *n* objects.
Output: A set of *k* clusters that minimizes the squared-error criterion.
Method:
(1) arbitrarily choose *k* objects as the initial cluster center;
(2) repeat
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, i.e., calculate the mean value of the objects for each cluster;
(5) until no change;

รูปที่ 2.7 อัลกอริทึมการทำงานของ K- means

ตัวอย่างการจัดกลุ่มของ K- means

ชุดข้อมูลมีทั้งหมด 10 ชุด จำนวนกลุ่ม 2 กลุ่ม

จำนวนกลุ่มเท่ากับ k = 2

ชุดข้อมูล

X1 [2 6]

X2 [3 4]

X3 [3 8]

X4 [4 7]

X5 [6 2]

X6 [6 4]

X7 [7 3]

X8 [7 4]

X9 [8 5]

X10[7 6]

ขั้นตอนที่ 1 สุ่มชุดข้อมูลเท่ากับจำนวนกลุ่ม k=2

จุดศูนย์กลางที่สุ่มได้

C1[3 4]

C2[7 4]

ขั้นตอนที่2 หาระยะห่างระหว่างข้อมูลแต่ละชุดกับจุดศูนย์กลาง (โดยจะใช้สมการหาระยะห่าง Euclidean Distance) ได้ดังตาราง 2.2 โดยข้อมูลที่มีสีน้ำเงินคือข้อมูลที่มีระยะใกล้กว่า ทำให้พิจารณาเลือกใช้ชุดข้อมูลที่มีสีน้ำเงิน

ตารางที่ 2.2 แสดงค่าระยะห่าง Euclidean Distance ระหว่างคู่ของข้อมูลกับจุดศูนย์กลาง

ข้อมูล	จุดศูนย์กลาง	ระยะห่าง
X1	C1	2.23
X1	C2	5.38
X2	C1	0
X2	C2	4
X3	C1	4
X3	C2	5.65
X4	C1	3.16
X4	C2	4.24
X5	C1	3.6
X5	C2	2.23
X6	C1	3
X6	C2	1
X7	C1	4.12
X7	C2	1
X8	C1	4
X8	C2	0
X9	C1	5.09
X9	C2	1.41
X10	C1	4.47
X10	C2	2

ขั้นตอนที่ 3 คำนวณหาค่าเฉลี่ยใหม่เพื่อสร้างจุดศูนย์กลางของกลุ่มใหม่

กลุ่ม C1

X1 [2 6]

X2 [3 4]

X3 [3 8]

X4 [4 7]

C1=[(2+3+3+4)/4 ,(6+4+8+7)/4]

C2=[(6+6+7+7+8+7)/6 , (2+4+3+4+5+6)/6]

จุดศูนย์กลางใหม่ที่คำนวณได้

C1=[3 6.25]

C2=[6.83 4]

ขั้นตอนที่ 4 ทำขั้นตอนที่สองและสามซ้ำ จนกระทั่งกลุ่มไม่เปลี่ยนแปลง ระยะห่างระหว่างข้อมูลแต่ละชุดกับจุดศูนย์กลาง ดังแสดงในตาราง 2.2 โดยข้อมูลที่มีสีน้ำเงินคือข้อมูลที่มีระยะใกล้กว่า ทำให้พิจารณาเลือกใช้ชุดข้อมูลที่มีสีน้ำเงิน

ตารางที่ 2.2 แสดงค่าระยะห่าง Euclidean Distance ระหว่างคู่ของข้อมูลกับจุดศูนย์กลาง

ข้อมูล	จุดศูนย์กลาง	ระยะห่าง
X1	C1	1.03
X1	C2	5.22
X2	C1	2.25
X2	C2	3.83
X3	C1	1.75
X3	C2	5.53
X4	C1	1.25
X4	C2	4.12
X5	C1	5.2
X5	C2	2.16

ตารางที่ 2.2 (ต่อ)

ข้อมูล	จุดศูนย์กลาง	ระยะห่าง
X6	C1	3.75
X6	C2	0.68
X7	C1	5.15
X7	C2	1.01
X8	C1	4.58
X8	C2	0.17
X9	C1	5.15
X9	C2	1.53
X10	C1	4
X10	C2	2

โดยข้อมูลที่มีสีน้ำเงินคือข้อมูลที่มีระยะใกล้กว่า ทำให้พิจารณาเลือกใช้ชุดข้อมูลที่มีสีน้ำเงิน

หลังจากพิจารณากลุ่มไม่มีการเปลี่ยนแปลงกลุ่มจะได้การจัดกลุ่มได้ดังนี้

X1 [2 6] กลุ่มที่ 1

X2 [3 4] กลุ่มที่ 1

X3 [3 8] กลุ่มที่ 1

X4 [4 7] กลุ่มที่ 1

X5 [6 2] กลุ่มที่ 2

X6 [6 4] กลุ่มที่ 2

X7 [7 3] กลุ่มที่ 2

X8 [7 4] กลุ่มที่ 2

X9 [8 5] กลุ่มที่ 2

X10[7 6] กลุ่มที่ 2

2.3 Two Step Clustering Algorithm

วิธีการทำงานของอัลกอริทึม ทูสเต็ปครัสเตอร์ริ่ง ทูสเต็ปครัสเตอร์ริ่งเป็นการแบ่งกลุ่มข้อมูลโดยมีลักษณะการทำงาน 2 ขั้นตอน

ขั้นตอนที่ 1 Pre-Cluster เป็นกระบวนแบ่งกลุ่มแบบเรียงลำดับข้อมูลเข้าทีละตัว โดยเริ่มเข้าที่ละข้อมูลโดยตัดสินใจว่าแต่ข้อมูลควรอยู่รวมกับกลุ่มใดก่อนหน้านี้หรือจะต้องสร้างกลุ่มใหม่ โดยที่จะต้องอยู่ภายใต้พื้นฐานของระยะห่างกันของข้อมูล โดยจะทำการสร้าง CF tree โดยที่ Cf tree ที่ ได้จะนำเสนอ มาก่อนหน้านี้

ขั้นตอนที่ 2 Cluster จะทำการเลือกจำนวนกลุ่มของข้อมูลที่เหมาะสม โดยใช้ค่าจะแผนภูมิต้นไม้ในขั้นตอนที่ 1 มาเป็นค่าเริ่มต้น โดยทำการแบ่งกลุ่มข้อมูลโดยใช้อัลกอริทึม K-Mean ที่ค่า K ที่ 1,2,3..... นำค่าที่ได้จากการแบ่งกลุ่ม K-Mean ไปหาค่า BIC เพื่อใช้ในการคำนวณหาจำนวนกลุ่มที่เหมาะสมสำหรับ K-Mean

การคำนวณค่า BIC

จำนวนกลุ่ม : Auto-Cluster

การหาจำนวนกลุ่มที่เหมาะสม โดยในขั้นตอนนี้จะเป็นการหา กลุ่มที่เหมาะสมสำหรับข้อมูล โดยวิธีการในขั้นตอนนี้ทำโดยการแบ่งกลุ่มข้อมูลจากข้อมูลใน CF tree โดยใช้โหนดที่ เป็น non leaf node มาเป็นข้อมูลเริ่มต้นในการแบ่งกลุ่ม โดยการแบ่งกลุ่มในขั้นตอนแรกจะใช้อัลกอริทึม K-Mean โดยจะแบ่งที่จำนวนกลุ่มเท่ากับ 1,2,3,4,....โดยในขั้นตอนที่ 2 ของ Two Step Clustering ทำโดยการนำข้อมูลจากการเบ่งกลุ่มข้อมูลด้วย K-Mean มาทำการคำนวณด้วยสูตร BIC ดังนี้

$$BIC(J) = -2\sum_{j=1}^{J} \xi_j + m_j \log(N) \tag{2.3}$$

โดยที่

$$\xi_v = -N_v \left[\sum_{k=1}^{K^A} \frac{1}{2} \log(\hat{\sigma}_k^2 + \hat{\sigma}_{vk}^2) + \sum_{k=1}^{K^B} \hat{E}_{vk} \right] \tag{2.4}$$

$$\hat{E}_{vk} = -\sum_{l=1}^{L_k} \frac{N_{vkl}}{N_v} \log \frac{N_{vkl}}{N_v} \tag{2.5}$$

$$m_j = J\left[2K^A + \sum_{k=1}^{K^B}(L_k - 1)\right] \quad (2.6)$$

J คือ จำนวนกลุ่ม ของ K-Mean ที่เราสนใจ

N คือ จำนวน เรคคอร์ดทั้งหมด

N_v คือ จำนวน เรคคอร์ดในกลุ่มย่อย K-Mean ที่ V

K^A คือ จำนวนตัวแปรที่เป็น Continuous

K^B คือ จำนวนตัวแปรที่เป็น Categorical

$\hat{\sigma}_k^2$ คือ ค่าความแปรปรวนของตัวแปร Continuous ที่ K ของข้อมูลทั้งหมด

$\hat{\sigma}_{vk}^2$ คือ ค่าความแปรปรวนของตัวแปร Continuous ที่ K ของข้อมูลในกลุ่มที่ V

N_{vkl} คือ จำนวน เรคคอร์ดในกลุ่มย่อย K-Mean ที่ V ตัวแปร Categorical ที่ K ของ ชุด j categories ที่ L

L_k คือ จำนวน categories ในตัวแปร Categorical ที่ K

โดยในการพิจารณาเลือกจำนวนกลุ่มที่เหมาะสมนั้นจะนำค่า BIC Change มาใช้ในการพิจารณา โดยค่า BIC change สามารถคำนวณได้จาก

$$BIC(J)_{change} = BIC(J) - BIC(J-1) \quad (2.7)$$

ค่า BIC Change ที่ได้เราจะเลือกค่าเพิ่มขึ้นมากที่สุดของค่า BIC Change มาใช้ในการพิจารณาเลือกจำนวนกลุ่มที่เหมาะสม เพราะค่า BIC เกิดจากการใช้ค่าความแปรปรวนของข้อมูลมาใช้ในการคำนวณดังนั้นค่าทีเลือกนำมาพิจารณาจึงควรจะเป็นค่าที่ เกิดจากการคำนวณแล้วจึงใช้ค่าที่เพิ่มขึ้นมากที่สุดระหว่าง 2 กลุ่มใกล้เคียงกัน

บทที่ 3

วิธีการดำเนินการศึกษา

ในบทนี้จะกล่าวถึงวิธีการพัฒนาโครงการโดยที่จะแบ่งเป็นเครื่องมือที่ใช้ใน การพัฒนาโปรแกรม การทำงานของโปรแกรม ลักษณะอินพุทและเอาท์พุทของโปรแกรม ขั้นตอนการทำงานของโปรแกรมรูปแบบการทำงานและส่วนติดต่อผู้ใช้ของโปรแกรม

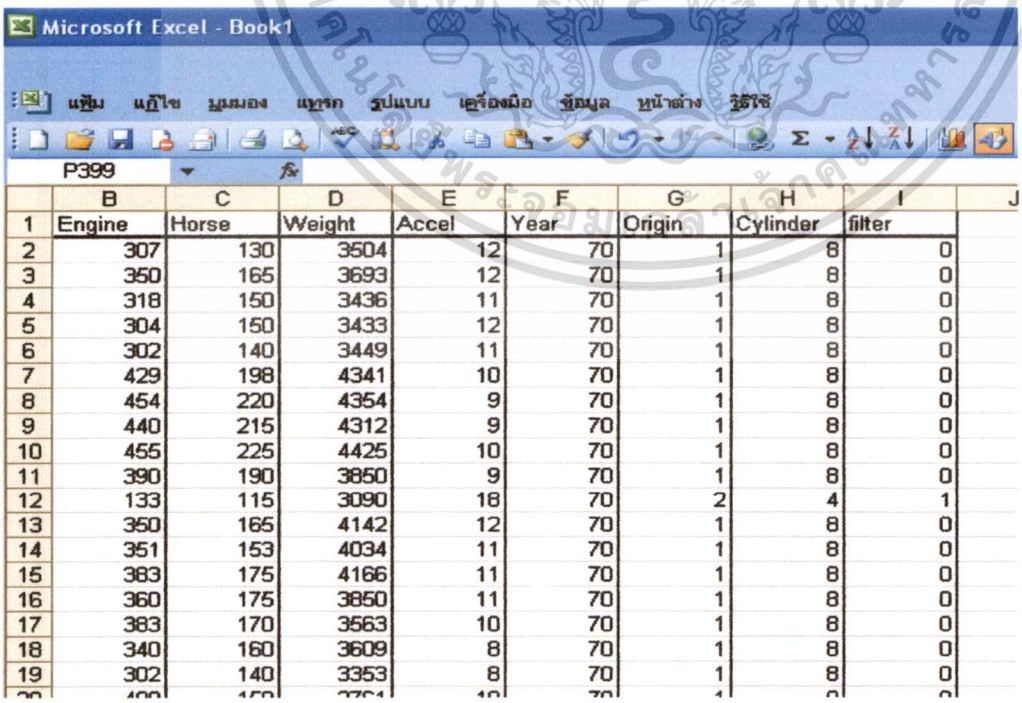
3.1 เครื่องมือที่ใช้ในการพัฒนาโปรแกรม

โครงการนี้เป็นการพัฒนาโปรแกรมประยุกต์ขึ้น เพื่อใช้ในการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม Twostep Clustering โดยใช้ BEA Weblogic 10.2 ในการพัฒนาระบบ

3.2 ลักษณะอินพุทและเอาท์พุทของโปรแกรม

3.2.1 ลักษณะอินพุทของโปรแกรม

อินพุทของโปรแกรมคือ ข้อมูลจำนวนหลายรูปแบบ โดยที่แต่ละรูปแบบต้องเป็นข้อมูลประเภทเดียวกันและมีลักษณะของข้อมูลเป็นตัวเลข โดยในโครงการนี้มุ่งเน้นไปที่การพัฒนาข้อมูลอินพุทในรูปแบบของไฟล์ Excel ตัวอย่างลักษณะอินพุทของโปรแกรมแสดงดังรูป



	B	C	D	E	F	G	H	I
1	Engine	Horse	Weight	Accel	Year	Origin	Cylinder	filter
2	307	130	3504	12	70	1	8	0
3	350	165	3693	12	70	1	8	0
4	318	150	3436	11	70	1	8	0
5	304	150	3433	12	70	1	8	0
6	302	140	3449	11	70	1	8	0
7	429	198	4341	10	70	1	8	0
8	454	220	4354	9	70	1	8	0
9	440	215	4312	9	70	1	8	0
10	455	225	4425	10	70	1	8	0
11	390	190	3850	9	70	1	8	0
12	133	115	3090	18	70	2	4	1
13	350	165	4142	12	70	1	8	0
14	351	153	4034	11	70	1	8	0
15	383	175	4166	11	70	1	8	0
16	360	175	3850	11	70	1	8	0
17	383	170	3563	10	70	1	8	0
18	340	160	3609	8	70	1	8	0
19	302	140	3353	8	70	1	8	0

รูปที่3.1 ลักษณะของอินพุทโปรแกรม

จากรูปสามารถอธิบายลักษณะของอินพุทได้คือ

- Row ที่ 1 จะเป็นตัวบอกถึงชื่อตัวแปรที่นำไปเป็นอินพุท
- ในแต่ละบรรทัดคือ อินพุท 1 รูปแบบ
- ข้อมูลในแต่ละคุณสมบัติจะมีค่าเป็นตัวเลขเท่านั้น

3.2.2 ลักษณะเอาท์พุทของโปรแกรม

เอาท์พุทของโปรแกรมคือ แสดงกลุ่มของข้อมูลอินพุทข้อมูลแต่ละตัว จำนวนกลุ่มที่ได้ สมาชิกในข้อมูลแต่ละกลุ่ม

ine	accel	mpg	Cluster
.0	15.0	27.0	1
.0	18.0	25.0	1
.0	19.0	31.0	1
.0	17.0	28.0	1
.0	20.0	29.0	1
.0	19.0	31.0	1
.0	17.0	29.0	1
.0	19.0	32.0	1
.0	18.0	33.0	1
.0	18.0	26.0	1
.0	14.0	29.0	1
.0	16.0	28.0	1
.0	15.0	30.0	1
.0	15.0	29.0	1
.0	17.0	31.0	1
.0	14.0	36.0	1
.0	15.0	31.0	1
.0	14.0	32.0	1
.0	13.0	34.0	1
.0	15.0	42.0	1
.0	16.0	34.0	1
.0	15.0	32.0	1
.0	22.0	44.0	1

Cluster	BIC	BIC Change
1	11030.126	
2	9830.333	-1199.793
3	9462.419	-367.914
4	9384.144	-78.275
5	9513.965	129.821
6	9671.102	157.136
7	9771.038	99.937
8	10038.989	267.951

Cluster	N	%of Total
1	131	32.27%
2	73	17.98%
3	100	24.63%
4	102	25.12%
Total	406	100%

Cluster	horse		engine		accel	
	STD	MEAN	STD	MEAN	STD	MEAN
1	94.06	187.02	71.80	155.93	16.95	6.22
2	233.26	392.22	99.48	169.54	16.88	3.23
3	137.62	299.72	94.48	188.21	15.98	5.92
4	349.70	1844.90	160.54	677.03	12.72	3.54

รูปที่ 3.2 ลักษณะของเอาท์พุทโปรแกรม

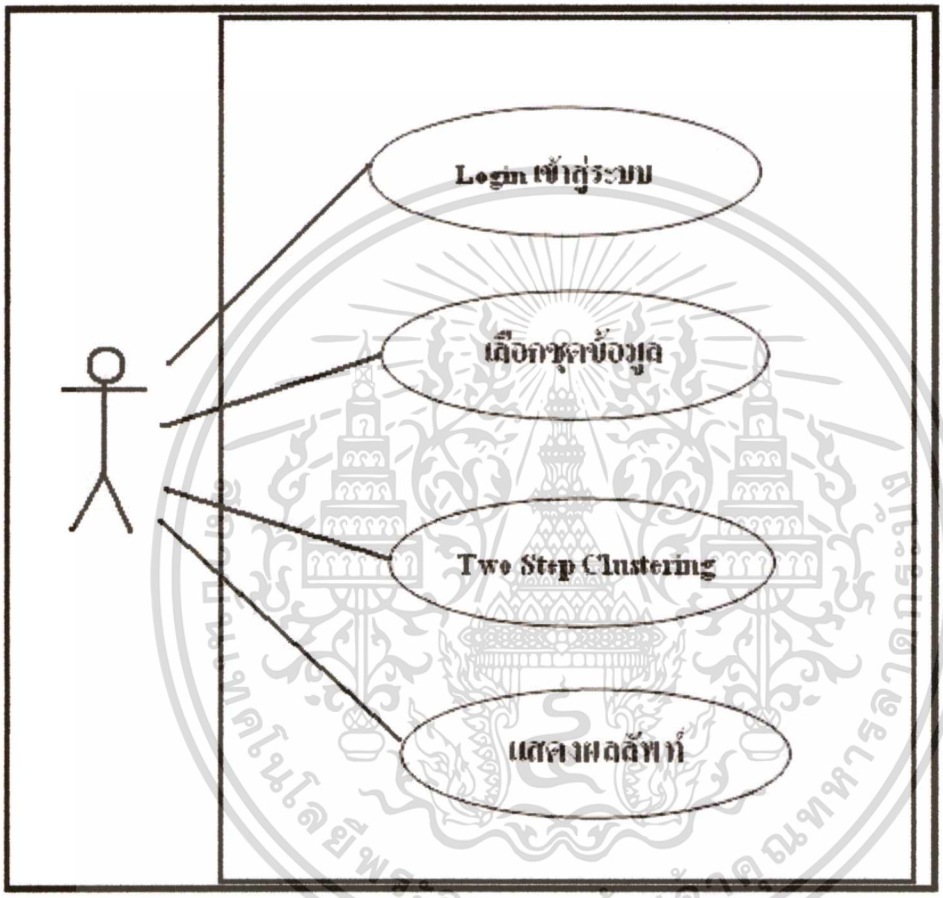
จากรูปสามารถอธิบายลักษณะของเอาท์พุทได้คือ

- ข้อมูลในตาราง Results จะมีข้อมูลที่ใช้ในการจัดกลุ่ม และแสดงข้อมูลในกลุ่มแต่ละชุดอยู่ที่กลุ่มใดบ้างได้
- ข้อมูลในตาราง Cluster จะมีข้อมูลแยกตามจำนวนกลุ่มที่ได้ โดยบอกจำนวนสมาชิกในแต่ละกลุ่ม

- ตารางคำนวณค่า *BIC*
- ตารางแสดงค่าจุดศูนย์กลาง ค่าความแปรปรวนของแต่ละกลุ่มแยกตามตัวแปรแต่ละตัว

3.3 การวิเคราะห์ระบบโดยใช้ยูสเคสวิวและแอคทิวิตีไดอะแกรม

3.3.1 ยูสเคสไดอะแกรม (Use Case Diagram)



รูปที่ 3.3 USE CASE การทำงานของระบบ TwoStep Clustering

อธิบายแอคเตอร์และยูสเคส

จากรูปที่ 3.3 เป็นการใช้ยูสเคสไดอะแกรมแสดงให้เห็นถึงภาพรวมของการพัฒนาระบบ โดยมีแอคเตอร์และยูสเคส ดังนี้

แอคเตอร์และบทบาทหน้าที่ของแอคเตอร์ที่ติดต่อกับโปรแกรม มีดังนี้

- ผู้ใช้งาน

เป็นบุคคลที่ป้อนอินพุตเข้าสู่ระบบ กำหนดตัวแปรควบคุม และเป็นผู้รับผลลัพธ์การแบ่งกลุ่มที่ได้จากระบบ

กิจกรรมหลักและรายละเอียดของกิจกรรมที่ทำงานภายในโปรแกรม ดังนี้

- การเข้าสู่ระบบ

  เพื่อยืนยันสิทธิ์การเข้าใช้ระบบ โดยผู้ใช้งานต้องทำการระบุ Username และ Password ให้ถูกต้อง

- กำหนดชุดข้อมูลที่ต้องการประมวลผล

  การเลือกชุดข้อมูลที่นำมาใช้ในการประมวลผล โดยการกำหนด path ที่ไปยังที่เก็บไฟล์ชุดข้อมูลที่จะนำมาประมวลผล

- TwoStepClustering

  การกำหนดตัวแปรควบคุมการแบ่งกลุ่มของข้อมูล ได้แก่ Maximum Branch , Maximum Tree Depth , Maximum Leaf node , ระบุประเภทของตัวแปรในการประมวลผล และ จำนวนกลุ่มสูงสุด

- การแสดงผลลัพธ์

  นำข้อมูลมาแสดงผลลัพธ์ โดยแสดงรายละเอียดการแบ่งกลุ่มข้อมูลเพื่อบอกว่าข้อมูลแต่ละชุดอยู่กลุ่มใด มีการแสดงค่า BIC ที่ได้จากการคำนวณในขั้นตอนที่ 2 ตารางแสดงรายละเอียดจำนวนกลุ่มและสมาชิกในกลุ่ม แสดงค่าจุดศูนย์กลางและค่าความแปรปรวนของกลุ่ม

รายละเอียดยูสเคส

ตารางที่ 3.1 รายละเอียดยูสเคส การเลือกLogin เข้าสู่ระบบ

ยูสเคส การLoginเข้าสู่ระบบ
รายละเอียด เป็นยูสเคสที่อธิบายวิธีการเข้าใช้งานระบบต้องทำการLogin เข้าสู่ระบบ
แพกเกจ Username Password

ตารางที่ 3.1(ต่อ)

แอคเตอร์ 1.ผู้ใช้งาน
เงื่อนไขก่อนเข้ายูสเคส ไม่มี
ลำดับเหตุการณ์หลัก 1. ผู้ใช้งานระบุ user password เข้าสู่ระบบ 2. ระบบทำการตรวจสอบ user 3. ระบบแสดงหน้าจอถัดไป หรือ ถามหา username password ใหม่
ลำดับเหตุการณ์ย่อย ระบบทำการตรวจสอบusername, password กับระบบ
เงื่อนไขก่อนออกยูสเคส ระบุ username password ให้ถูกต้อง

ตารางที่ 3.2 รายละเอียดยูสเคส การเลือกชุดข้อมูล

ยูสเคส เลือกชุดข้อมูล
รายละเอียด เป็นยูสเคสที่อธิบายวิธีการเลือกชุดข้อมูลเข้าสู่ระบบ
แพกเกจ ชุดข้อมูล
แอคเตอร์ 1.ผู้ใช้งาน
เงื่อนไขก่อนเข้ายูสเคส ต้องทำ Login เข้าสู่ระบบก่อน

ตารางที่ 3.2(ต่อ)

ลำดับเหตุการณ์หลัก 1. ผู้ใช้งานระบุชุดข้อมูล เข้าสู่ระบบ 2. ระบบแสดงชุดข้อมูลที่ถูกเลือก 3. ระบบแสดงหน้าจอถัดไป
ลำดับเหตุการณ์ย่อย ไม่มี
เงื่อนไขก่อนออกยูสเคส ต้องเลือกชุดข้อมูล

ตารางที่ 3.3 รายละเอียดยูสเคส Two Step Clustering

ยูสเคส Two Step Clustering
รายละเอียด เป็นยูสเคสที่อธิบายขั้นตอนการทำงานของอัลกอริทึม Two Step Clustering
แพกเกจ Max Branch Maximum Tree Maximum Leaf node Max Cluster ระบุประเภทตัวแปร
แอคเตอร์ 1.ผู้ใช้งาน
เงื่อนไขก่อนเข้ายูสเคส ต้องทำ การเลือกชุดข้อมูลเข้าสู่ระบบก่อน

ตารางที่ 3.3(ต่อ)

ลำดับเหตุการณ์หลัก 1. ผู้ใช้งานระบุประเภทของตัวแปรที่ใช้ในการประมวลผลระบบ 2. ผู้ใช้งานระบุตัวแปร Max Cluster , Max Branch , Max level , Max leaf 3. ระบบทำการประมวลผล
ลำดับเหตุการณ์ย่อย ระบบทำการตรวจสอบความเป็นไปได้ของการสร้าง Tree
เงื่อนไขก่อนออกยูสเคส ต้องระบุตัวแปรทั้งหมดครบถ้วน

ตารางที่ 3.4 รายละเอียดยูสเคส แสดงผลลัพธ์

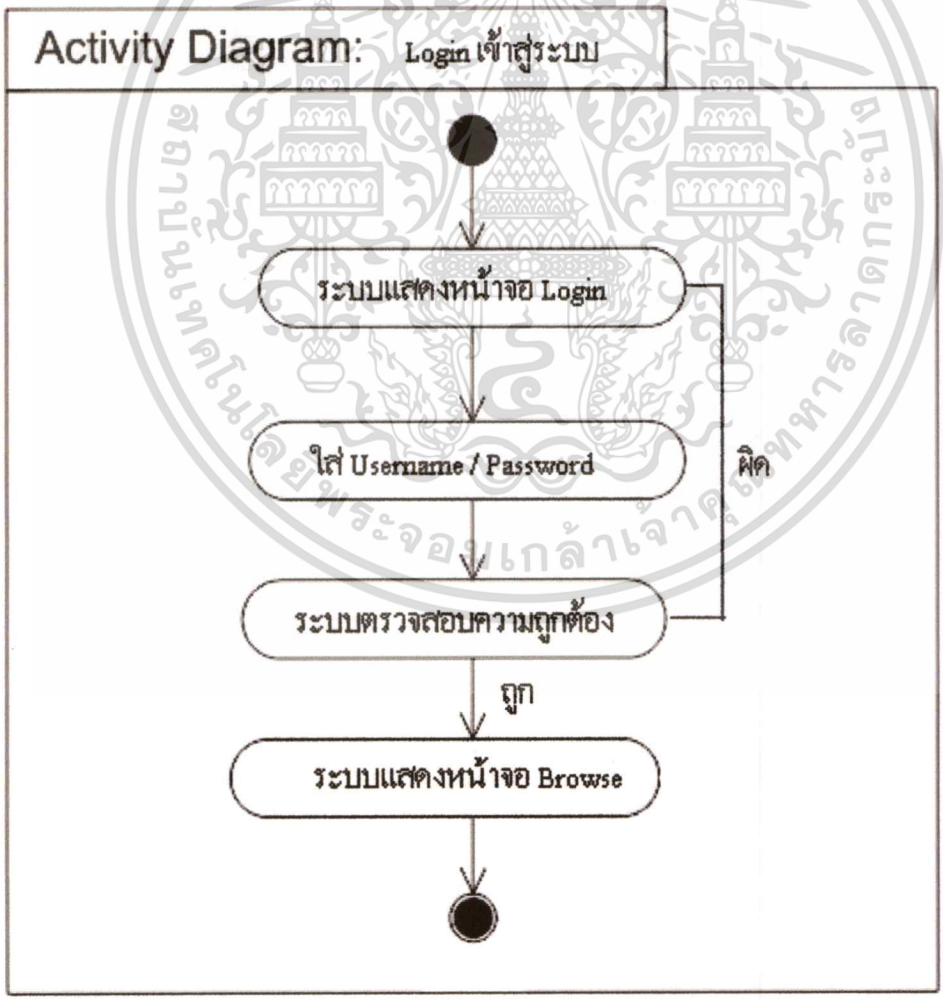
ยูสเคส แสดงผลลัพธ์
รายละเอียด เป็นยูสเคสที่อธิบายการแสดงผลลัพธ์
แพกเกจ ข้อมูลการประมวลผลจากระบบ
แอคเตอร์ 1.ผู้ใช้งาน
เงื่อนไขก่อนเข้ายูสเคส ต้องทำการประมวลผล Two Step Clustering ก่อน
ลำดับเหตุการณ์หลัก 1. ผู้ใช้งานเลือกประเภทของการแสดงผลลัพธ์ (open,save) 2. ระบบแสดงผลลัพธ์

ตารางที่ 3.4(ต่อ)

ลำดับเหตุการณ์ย่อย ไม่มี
เงื่อนไขก่อนออกยูสเคส ต้องเลือกประเภทของการจัดการผลลัพธ์แล้ว

3.3.2 Activity Diagram

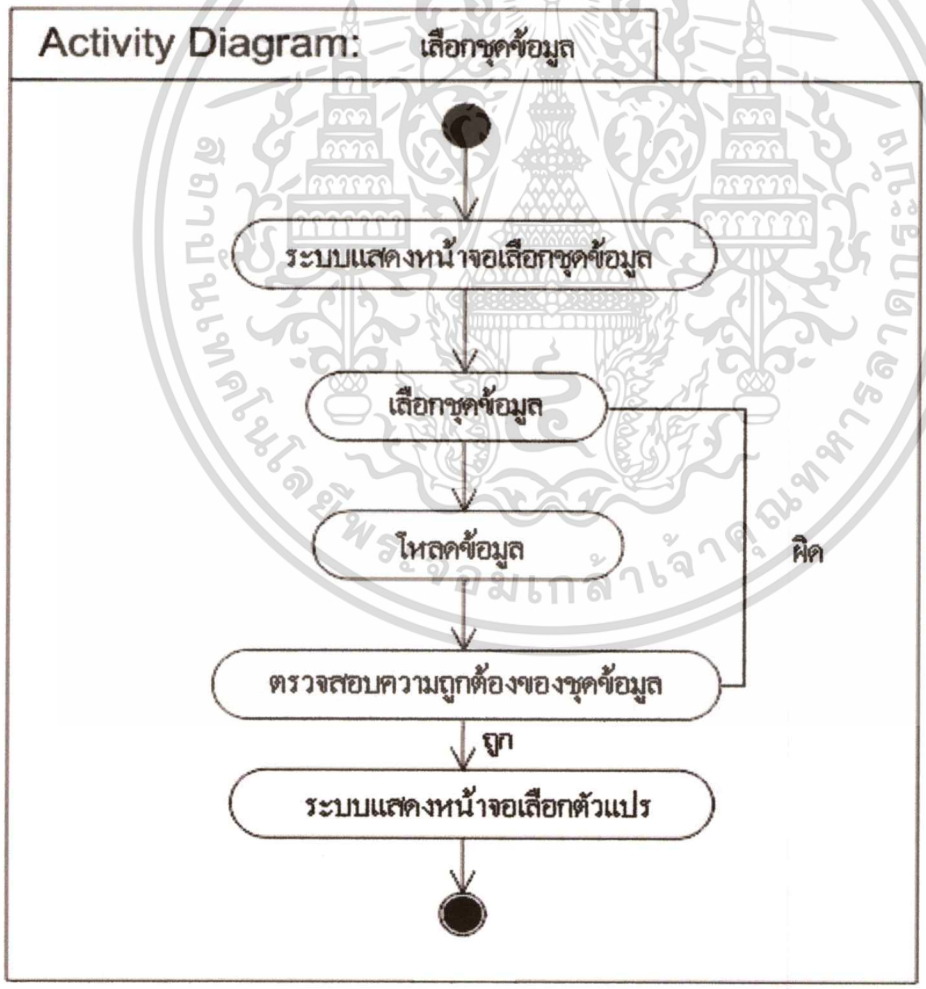
ขั้นตอนของกิจกรรมที่เกิดขึ้นในการพัฒนาระบบสามารถอธิบายได้ด้วย Activity Diagram โดย Activity Diagram แรก Activity Diagram การเข่าสู่ระบบ แสดงการเข้าสู่ระบบของผู้ใช้งาน โดยระบบจะมีการตรวจสอบความถูกต้องของ Username และ Password ดังแสดงในรูป 3.4



รูปที่ 3.4 Activity Diagram การเข้าสู่ระบบ

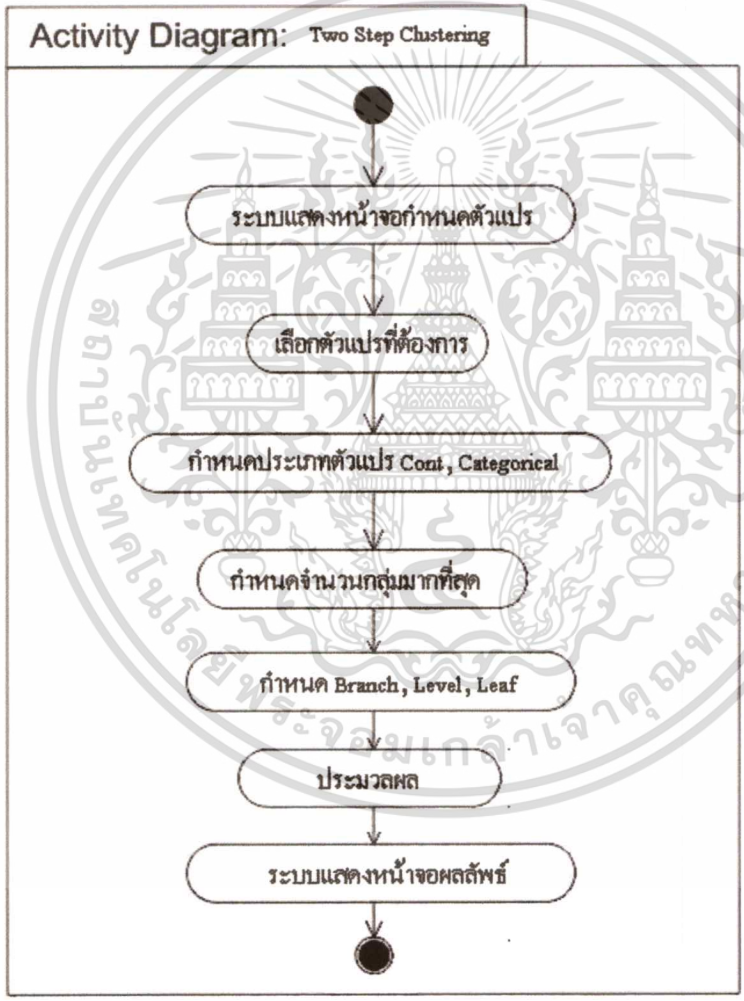
Activity Diagram แสดงขั้นตอนการทำงานของระบบย่อยการเลือกชุดข้อมูล โดยในการเลือกชุดข้อมูลนั้นทำได้โดยการระบุ เส้นทางไปยังที่เก็บชุดข้อมูลจากนั้นดึงค่ามาเพื่อแสดงผลหลังจากนั้นระบบจะทำการแสดงผลเพื่อให้ผู้ใช้งานได้ตรวจสอบความถูกต้องของข้อมูล ดังที่ได้แสดงในรูป 3.5

Activity Diagram แสดงขั้นตอนการทำงานของระบบย่อยTwo Step Clustering โดยในระบบย่อยTwo Step Clustering นี้จะทำการ ระบุประเภทของตัวแปรให้กับตัวแปรอันได้แก่ ตัวแปร Categorical และตัวแปร Continuous ระบุค่าต่างๆภายใน Tree อันได้แก่ Maximum Branch , Maximum Tree Depth , Maximum Leaf node และ ระบุค่าจำนวนกลุ่มสูงสุด หลังจากนั้นจะประมวลผลตามขั้นตอนการทำงานของ Two Step Clustering ดังแสดงในรูป 3.6

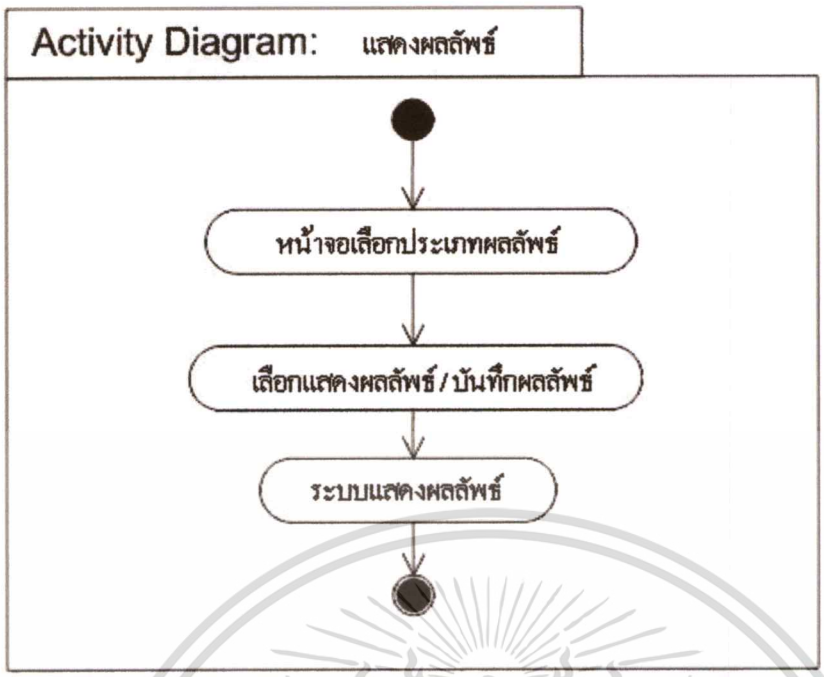


รูปที่ 3.5 Activity Diagram การเลือกชุดข้อมูล

Activity Diagram แสดงขั้นตอนการทำงานของระบบย่อยการแสดงผลลัพธ์ โดยจะนำผลลัพธ์จากการประมวลผลของระบบย่อย Two Step Clustering เข้ามาใช้งาน โดยมีการแสดงผลลัพธ์ในรูปแบบ Excel โดยจะต้องทำการระบุว่าจะทำการเก็บข้อมูลเป็นไฟล์ หรือจะทำการเปิดไฟล์ ผลลัพธ์ได้จะมีการแสดงการแบ่งกลุ่มของข้อมูลแยกตามกลุ่ม , แสดงค่า BIC , ค่าผลสรุปในการแบ่งกลุ่ม , และค่าสรุปข้อมูลตัวแปรแต่ ละตัวแยกตามประเภทของตัวแปรมีการคำนวณหาค่าเฉลี่ยของข้อมูลต่างๆที่น่าสนใจ โดยจะแสดงดังรูป 3.7



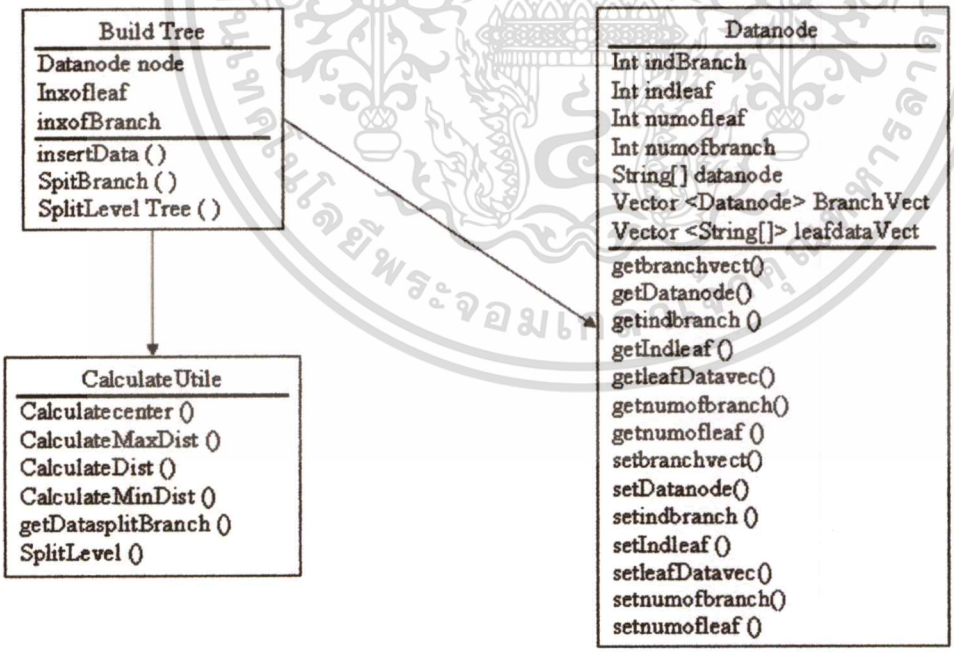
รูปที่ 3.6 Activity Diagram Two Step Clustering



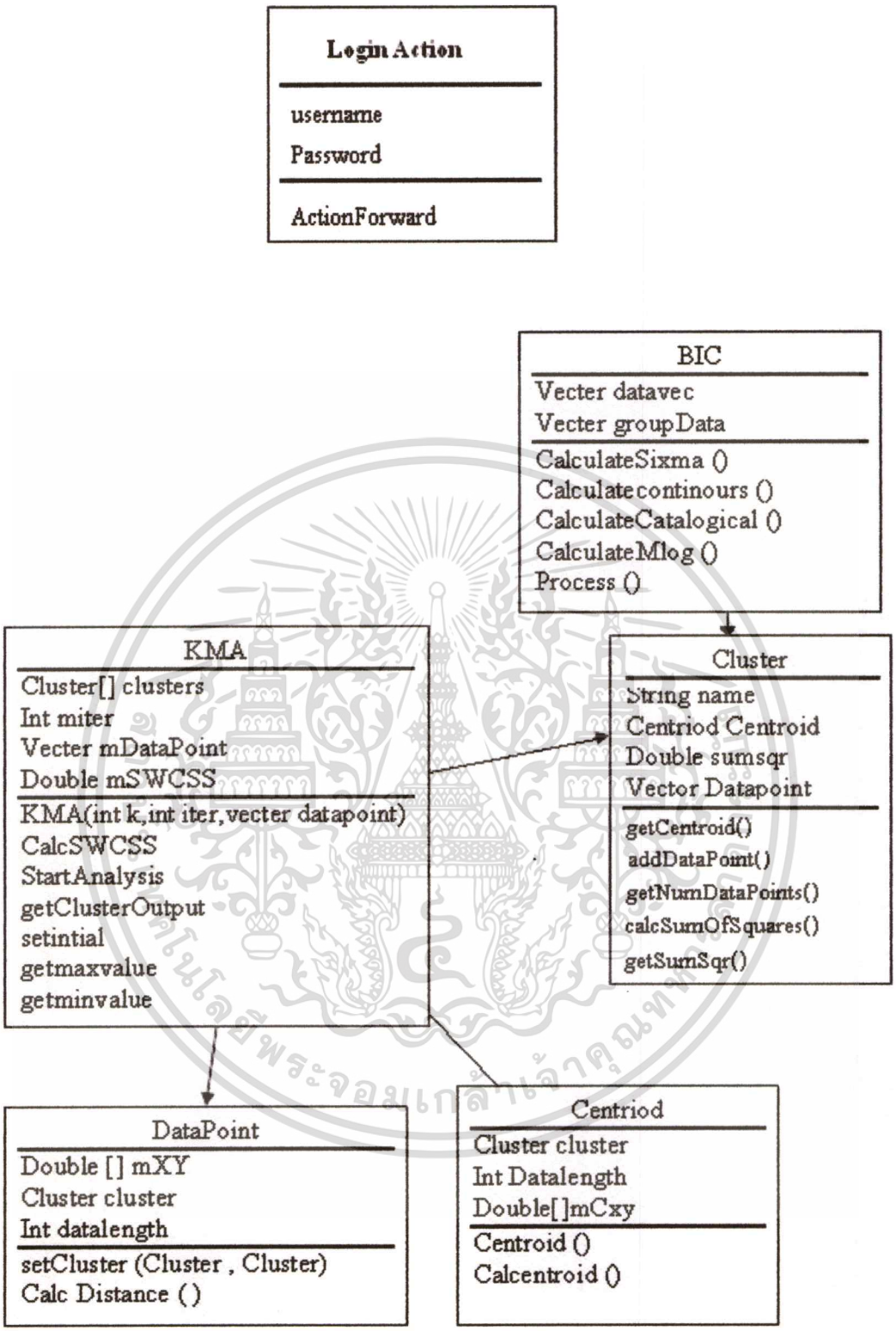
รูปที่ 3.7 Activity Diagram ประมวลผล

3.3.4 การออกแบบระบบโดยใช้คลาสไดอะแกรม

รูปแสดงการทำงานของคลาสต่างๆ ที่มีในระบบ โดยมีรายละเอียดดังรูปต่อไปนี้



รูปที่ 3.8 แสดงคราสไดอะแกรมของระบบการสร้าง Tree



รูปที่ 3.9 แสดงคราสไดอะแกรมระบบK-Mean และการ Login

คลาสไดอะแกรมของระบบประกอบด้วย

คลาส DataNode เป็น data object คลาสในการเก็บข้อมูลการสร้าง Tree

คลาส BuildTree เป็นคลาสประมวลผลและทำการเพิ่มข้อมูลลงใน tree โดยเก็บข้อมูลเป็น dataNode

คลาส CalculateUtil เป็นคลาสที่ถูกเรียกใช้จากคลาส buildTree เพื่อทำการคำนวณข้อมูลต่างๆ

คลาส KMA เป็น คลาสในการคำนวณ K-mean โดยจะแบ่งข้อมูลเป็น cluster และเก็บข้อมูลย่อยเป็น DataPoint

คลาส Cluster แทนการเก็บชุดข้อมูล

คลาส DataPoint แทนข้อมูลนำเข้า

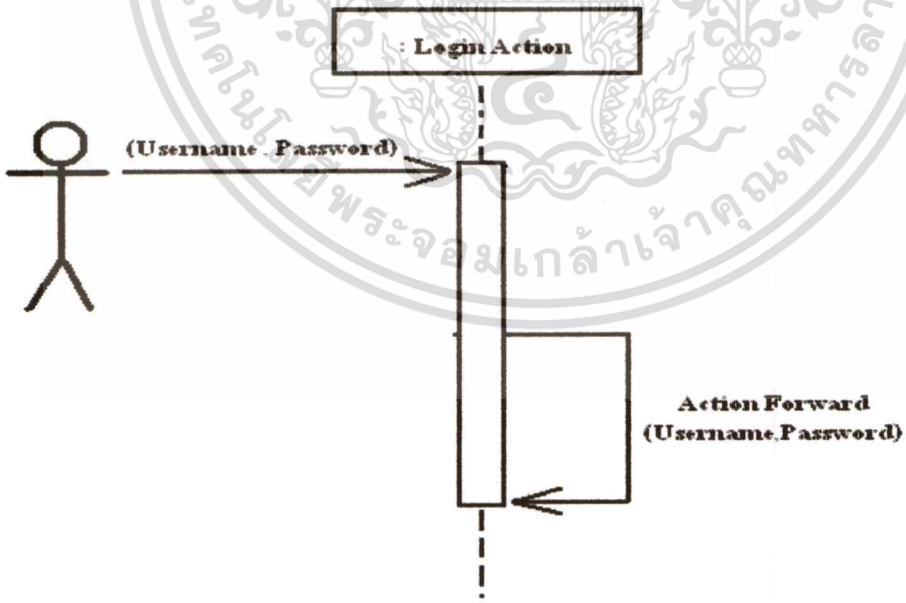
คลาส Centroid ใช้ในการคำนวณข้อมูลแล ะเปรียบเทียบข้อมูลชุดข้อมูลต่างๆ

คลาส Login Action ทำหน้าที่ระบุการเข้าถึงการใช้งาน

คลาส BIC เป็นคลาสที่ใช้ในการคำนวณสูตร BIC

3.3.5 Sequence Diagram

Sequence Diagram คือการจำลองกระบวนการที่ทำให้เกิดกิจกรรมของระบบ เกิดจากชุดของกิจกรรม ซึ่งกิจกรรมหนึ่งๆ ที่เกิดขึ้นนั้นจะถูกระบุเพิ่มไว้บนเส้นที่ใช้เพื่อแสดงลำดับเวลา และเส้นที่ใช้เพื่อแสดงกิจกรรมที่เกิดกิจกรรมต่างๆ

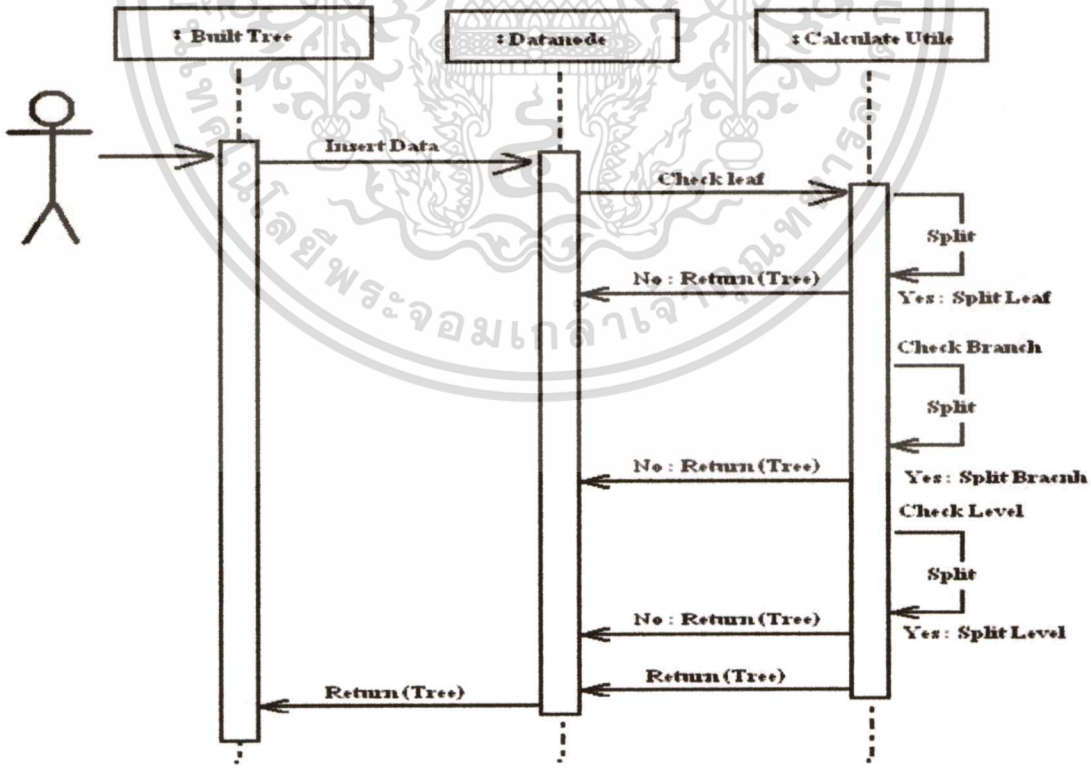


รูปที่ 3.10 Sequence Diagram การ Login เข้าสู่ระบบ

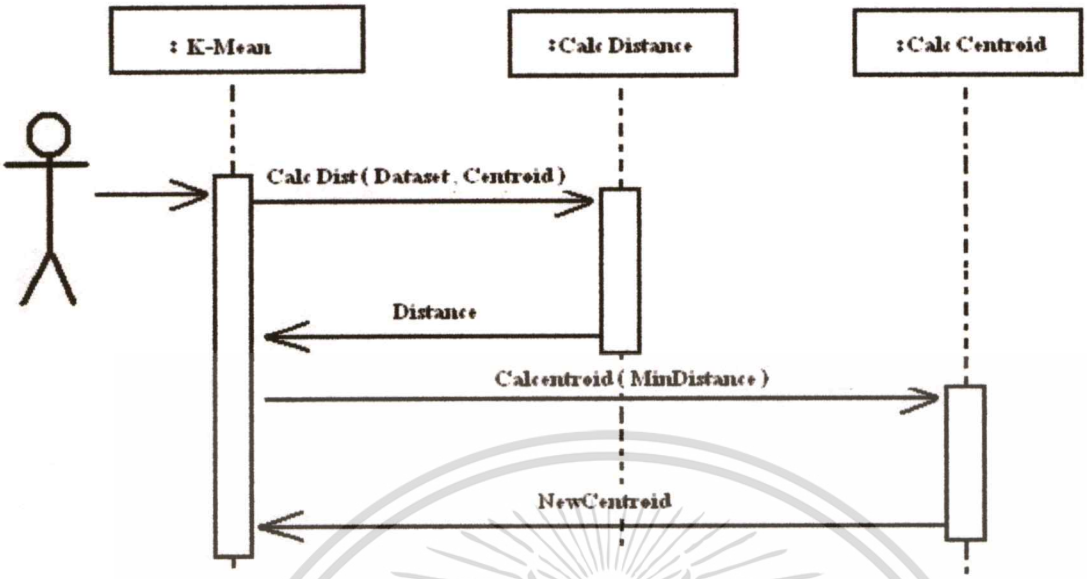
Sequence Diagram การ Login เข้าสู่ระบบ ผู้ใช้งานระบบจัดการกรอกข้อมูลเพื่อ Login เข้าสู่ระบบ ระบบทำการส่งข้อมูล Username Password ไปตรวจสอบก่อนส่งผลกลับมา ดังแสดงในรูป 3.10

Sequence Diagram Build CF-Tree ทำการสร้าง Tree โดยผู้ใช้ทำการระบุตัวแปรและกำหนดตัวแปรในการใช้ Tree หลังจากนั้นจะเป็นหน้าที่ของระบบ โดยระบบจะทำการตรวจสอบหาข้อมูลที่สนใจควรจะอยู่ leaf node ไหน หลังจากนั้นระบบจะทำการเพิ่มข้อมูลลงใน leaf node จะมีการตรวจสอบว่า leaf node เต็มไหม ถ้าเต็ม Split ถ้าไม่เต็มก็ทำต่อข้อมูลตัวใหม่เข้าสู่ระบบ ถ้ามีการ split จะมีการตรวจสอบ Branch ว่าเต็มไหม ถ้าเต็มก็ทำการ Split เหมือนกัน หลังจากนั้นจะมีการส่งค่า Tree ที่ได้กลับดังที่ได้แสดงในรูป 3.12

Sequence Diagram K-MEAN ทำการแบ่งกลุ่มข้อมูล non leaf node จาก ขั้นตอนสร้าง Tree มาใช้เป็นข้อมูลตั้งต้น หลังจากนั้นทำการคำนวณหา K-Mean โดยการหาระยะห่างจากจุด Centroid ที่สุ่มขึ้นมา เพื่อว่าข้อมูลอยู่จุดศูนย์กลางใดมากที่สุดหลังจากนั้นก็จัดกลุ่มใหม่เพื่อหาจุดศูนย์กลางใหม่นำมาทำซ้ำขั้นตอนก่อนหน้านี้ ดังที่ได้แสดงในรูป 3.13



รูปที่ 3.11 Sequence Diagram การสร้าง Tree



รูปที่ 3.11 Sequence Diagram K-Mean

บทที่ 4

การพัฒนาระบบ

4.1 เครื่องมือที่ใช้ในการพัฒนาระบบ

4.1.1 ฮาร์ดแวร์

ในการพัฒนาระบบงานใช้เครื่องคอมพิวเตอร์ที่มีคุณสมบัติดังนี้

- CPU : Intel Centrino Duo T2300 1.66GHz.
- Hard disk 80 GB.
- RAM 1.25 GB.

4.1.2 ซอฟต์แวร์

ในการพัฒนาระบบงานใช้ซอฟต์แวร์ดังนี้

- Window XP
- BEA Weblogic 10.2

4.2 รายละเอียดของการทำงานของระบบ

โครงงานพัฒนาระบบ Two Step Clustering โดยมีระบบเว็บแอพพลิเคชั่นเป็นระบบที่ใช้ติดต่อกับผู้ใช้งาน รับชุดข้อมูลและการกำหนดตัวแปรควบคุมต่างๆ และนำมาประมวลผลตามรายละเอียดของโครงการ มีรายละเอียดหน้าจอการทำงาน ดังต่อไปนี้

หน้าจอหลักเข้าสู่ระบบ Two Step Cluster

เป็นหน้าจอหลักหน้าจอแรกในการเข้าสู่ระบบโดยในหน้าจอนี้จะมีการ Login เข้าสู่การใช้งานโดยจะมีการ ระบุ USERNAME และ PASSWORD เพื่อเข้าสู่การใช้งานของระบบ โดยในระบบที่พัฒนาขึ้นในเบื้องต้นนี้ได้มีการระบุ USERNAME และ PASSWORD ดังนี้

USERNAME : ADMIN

PASSWORD : ADMIN

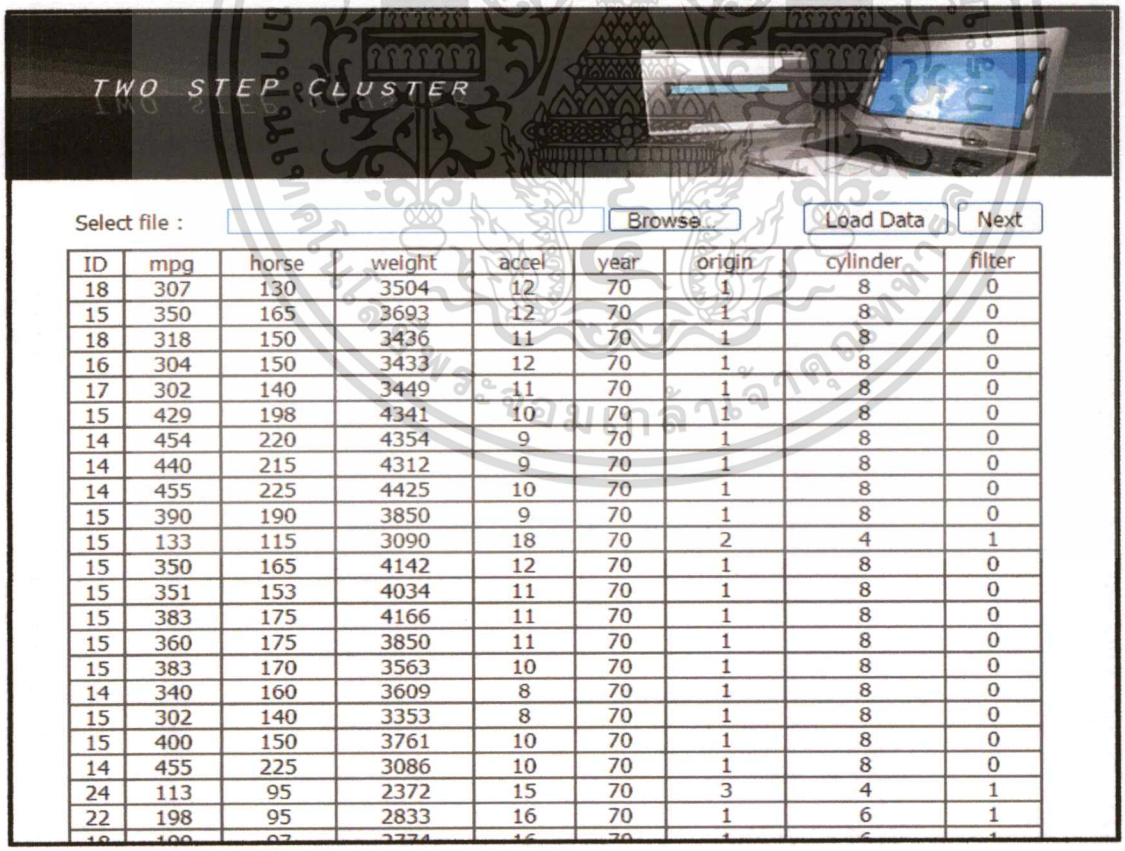
เพื่อให้ระบบมีความปลอดภัยมากขึ้นได้การใช้งานจึงได้มีการจัดเตรียมในส่วนของการ Login เข้าสู่ระบบ อีกทั้งยังเป็นการระบุผู้ใช้งานได้ โดยหน้าจอ Login นี้จะแสดงดังรูปที่ 4.1

รูปที่ 4.1 หน้าจอเข้าสู่ระบบ TwoStep Cluster

หน้าจอเลือกชุดข้อมูล

หน้าจอเลือกชุดข้อมูลหน้าจอนี้จะอยู่ถัดจากหน้าจอเข้าสู่ระบบ โดยหลังจาก Login ข้าสู่ระบบแล้ว ก็จะต้องทำการเลือกชุดข้อมูลเข้าสู่ระบบ โดยที่หน้าจอการเลือกชุดข้อมูลเข้าสู่ระบบเป็นหน้าจอที่เลือกชุดข้อมูลเป็นหน้าจอที่ผู้ใช้งานระบบ ใช้ในการระบุไฟล์ข้อมูลเส้นทางสำหรับชุดข้อมูลที่จะนำมาวิเคราะห์ โดยเลือกไฟล์ที่เลือกมาเป็นไฟล์ Microsoft Excel ที่เก็บชุดข้อมูลที่ต้องการนำมาใช้ในการประมวลผล หลังจากนั้นระบบจะแสดงชุดข้อมูลที่เลือกมา โดยหลังจากที่แสดงชุดข้อมูลแล้ว หากข้อมูลที่เลือกมาไม่ถูกต้องก็สามารถตรวจสอบความถูกต้องของชุดข้อมูลได้และยังสามารถเปลี่ยนชุดข้อมูลได้อีกด้วย ในการเลือกชุดข้อมูลมานั้น ระบบไม่สามารถที่จะทำการแก้ไขเพิ่ม หรือ ลบข้อมูลได้ หากผู้ใช้งานจำเป็นต้องมีการแก้ไขสามารถทำได้โดยใช้ Microsoft Excel หน้าจอเลือกชุดข้อมูลนี้ได้แสดงดังรูปที่ 4.2

รูปที่ 4.2 หน้าจอเลือกชุดข้อมูลเข้าสู่ระบบ

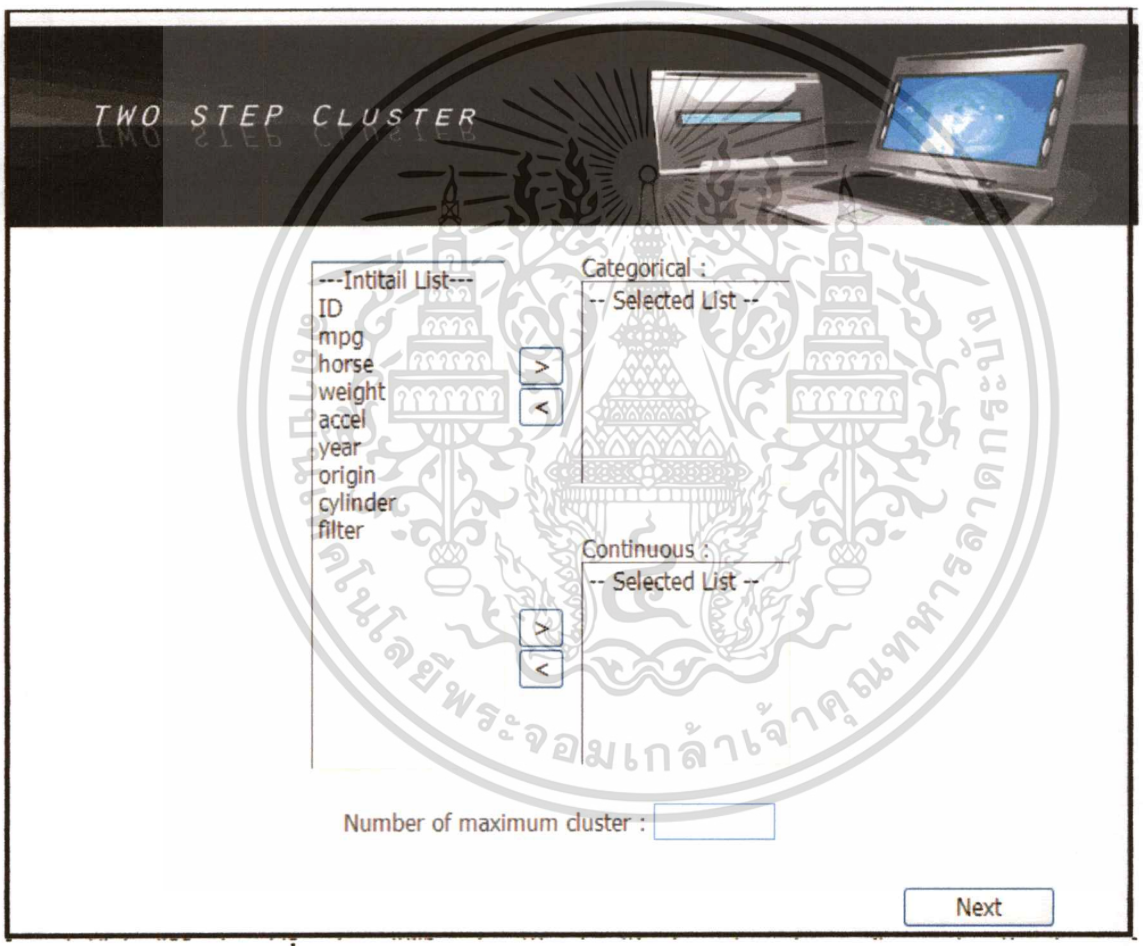


ID	mpg	horse	weight	accel	year	origin	cylinder	filter
18	307	130	3504	12	70	1	8	0
15	350	165	3693	12	70	1	8	0
18	318	150	3436	11	70	1	8	0
16	304	150	3433	12	70	1	8	0
17	302	140	3449	11	70	1	8	0
15	429	198	4341	10	70	1	8	0
14	454	220	4354	9	70	1	8	0
14	440	215	4312	9	70	1	8	0
14	455	225	4425	10	70	1	8	0
15	390	190	3850	9	70	1	8	0
15	133	115	3090	18	70	2	4	1
15	350	165	4142	12	70	1	8	0
15	351	153	4034	11	70	1	8	0
15	383	175	4166	11	70	1	8	0
15	360	175	3850	11	70	1	8	0
15	383	170	3563	10	70	1	8	0
14	340	160	3609	8	70	1	8	0
15	302	140	3353	8	70	1	8	0
15	400	150	3761	10	70	1	8	0
14	455	225	3086	10	70	1	8	0
24	113	95	2372	15	70	3	4	1
22	198	95	2833	16	70	1	6	1

รูปที่ 4.3 หน้าจอเลือกชุดข้อมูลและโหลดข้อมูลเข้าสู่ระบบ

หน้าจอกำหนดตัวแปรควบคุมการแบ่งกลุ่มระบบ

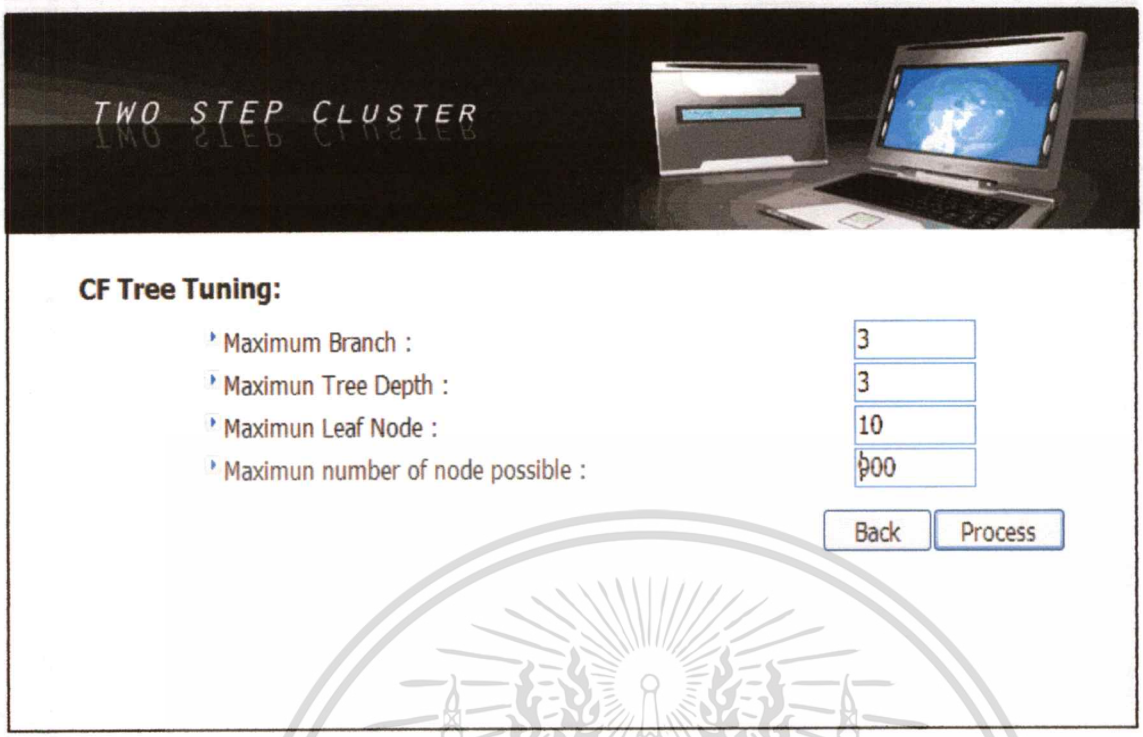
หน้าจอกำหนดตัวแปรควบคุมการแบ่งกลุ่มระบบเป็นหน้าจอที่ใช้ในการกำหนดตัวแปรต่างๆเพื่อใช้ในการคำนวณและจัดกลุ่มข้อมูล โดยในการเลือกตัวแปรมาใช้ในการคำนวณนี้จะต้องระบุประเภทของตัวแปรที่ใช้โดยในที่นี้ตัวแปรที่ใช้ในการคำนวณมีด้วยกัน 2 ประเภท คือตัวแปร Categorical และตัวแปร Continuous และในหน้าจอกำหนดตัวแปรควบคุมการแบ่งกลุ่มระบบนี้ยังต้องระบุค่า Number of maximum Cluster หรือ จำนวนกลุ่มสูงสุดที่สามารถมีได้ โดยหน้าจอดังกล่าวจะแสดงดังรูปที่ 4.4



รูปที่ 4.4 หน้าจอกำหนดตัวแปรควบคุมการแบ่งกลุ่มระบบ

หน้าจอกำหนดตัวแปรของ Tree

หน้าจอกำหนดตัวแปรของ Tree เป็นหน้าจอที่ระบุตัวแปรต่างๆที่ใช้ในการสร้าง Tree โดยจะมีการระบุถึงตัวแปรต่างๆได้แก่ ความสูงของTree , จำนวนสมาชิกมากที่สุดใน non leaf node , จำนวนสมาชิกมากที่สุดใน leaf node โดยจะนำค่า 3 ค่านี้มาคำนวณเพื่อระบุจำนวนของชุดข้อมูลสูงสุดที่เป็นไปได้ในการคำนวณโดยหน้าจอนี้จะแสดงในรูปที่ 4.5



รูปที่ 4.5 แสดงหน้าจอกำหนดตัวแปรของ Tree

หน้าจอผลลัพธ์ของระบบ

หน้าจอผลลัพธ์ของระบบเป็นหน้าจอเพื่อให้ที่แสดงชุดข้อมูลโดยมีการระบุหมายเลขกลุ่มของชุดข้อมูลนั้น แสดงดังรูปที่ 4.6

ine	accel	mpg	Cluster
.0	15.0	27.0	1
.0	18.0	25.0	1
.0	19.0	31.0	1
.0	17.0	28.0	1
.0	20.0	29.0	1
.0	19.0	31.0	1
.0	17.0	29.0	1
.0	19.0	32.0	1
.0	18.0	33.0	1
.0	18.0	26.0	1
.0	14.0	29.0	1
.0	16.0	28.0	1
.0	15.0	30.0	1
.0	15.0	29.0	1
.0	17.0	31.0	1
.0	14.0	36.0	1
.0	15.0	31.0	1
.0	14.0	32.0	1
.0	13.0	34.0	1
.0	15.0	42.0	1
.0	16.0	34.0	1
.0	15.0	32.0	1
.0	22.0	44.0	1

Cluster	BIC	BIC Change
1	11030.126	
2	9830.333	-1199.793
3	9462.419	-367.914
4	9384.144	-78.275
5	9513.965	129.821
6	9671.102	157.136
7	9771.038	99.937
8	10038.989	267.951

Cluster	N	%of Total
1	131	32.27%
2	73	17.98%
3	100	24.63%
4	102	25.12%
Total	406	100%

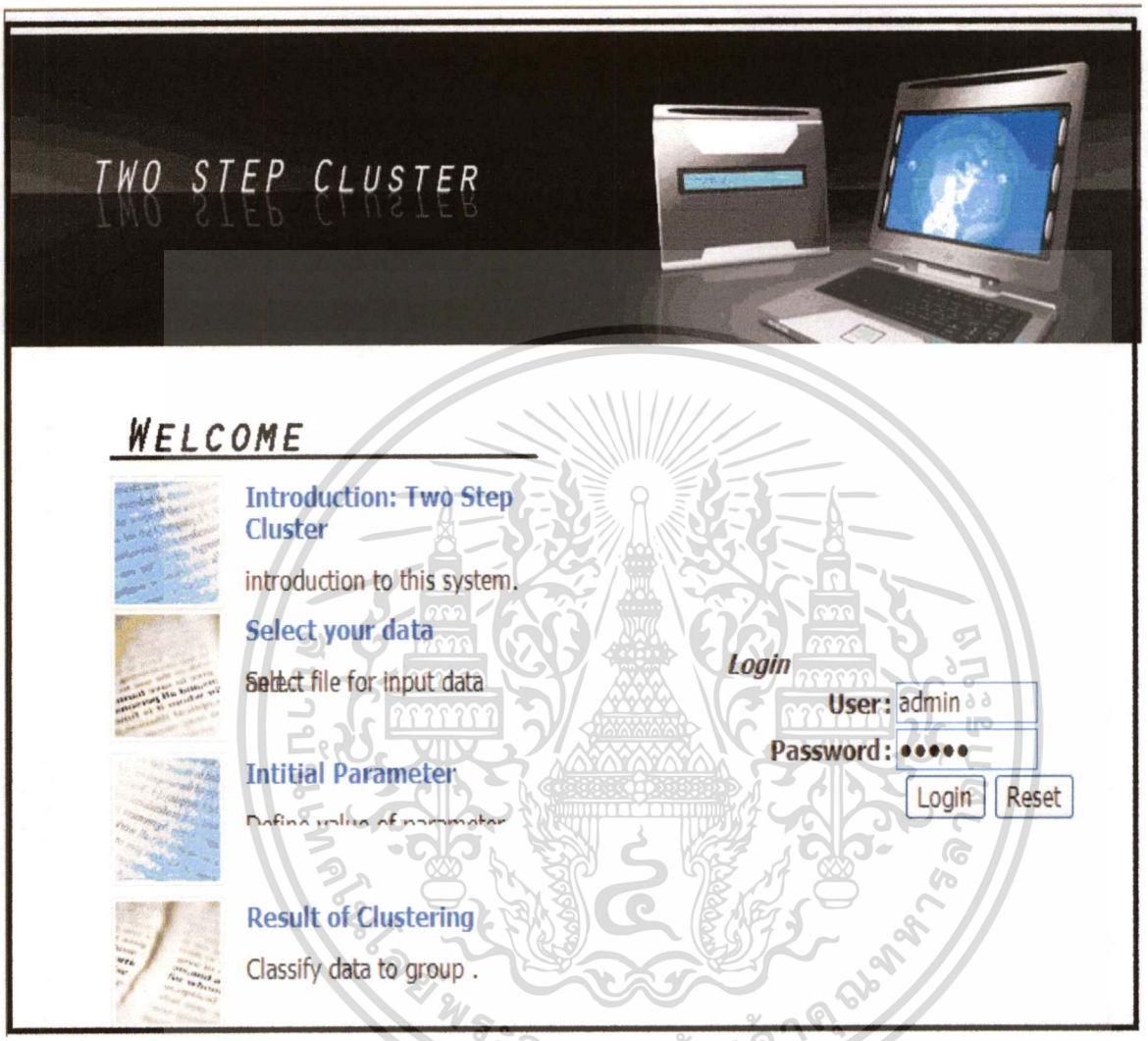
Cluster	horse		engine		accel	
	STD	MEAN	STD	MEAN	STD	MEAN
1	94.06	187.02	71.80	155.93	16.95	6.22
2	233.26	392.22	99.48	169.54	16.88	3.23
3	137.62	299.72	94.48	188.21	15.98	5.92
4	349.70	1844.90	160.54	677.03	12.72	3.54

รูปที่ 4.6แสดงหน้าจอผลลัพธ์ของระบบ

4.3 วิธีการใช้งานระบบ

ระบบจะมีหน้าจอหลักหน้าจอแรกในการเข้าสู่ระบบโดยในหน้าจอนี้จะมีการ Login เข้าสู่การใช้งานโดยจะมีการ ระบุ USERNAME และ PASSWORD เพื่อเข้าสู่การใช้งานของระบบ โดยใน

ระบบที่พัฒนาขึ้นในเบื้องต้นนี้ได้มีการระบุ USERNAME และ PASSWORD โดยหน้าจอนี้จะแสดงดังรูป 4.7



รูปที่ 4.7 แสดงหน้าจอเข้าสู่ระบบ จะมีให้ระบบ Login เข้าใช้งานโปรแกรม

ขั้นตอนการ Login เข้าสู่ระบบสามารถทำได้ดังนี้

ใส่ Username : Admin

Password : Admin

หลังจากนั้นกดปุ่ม Login เพื่อทำการ login เข้าสู่ระบบ หลังจากนั้นระบบจะแสดงหน้าจอการเลือกชุดข้อมูล ดังแสดงในรูปที่ 4.8

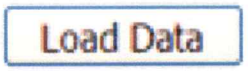
รูปที่ 4.8 หน้าจอ เลือกข้อมูลเข้าสู่โปรแกรมการแบ่งกลุ่มข้อมูล

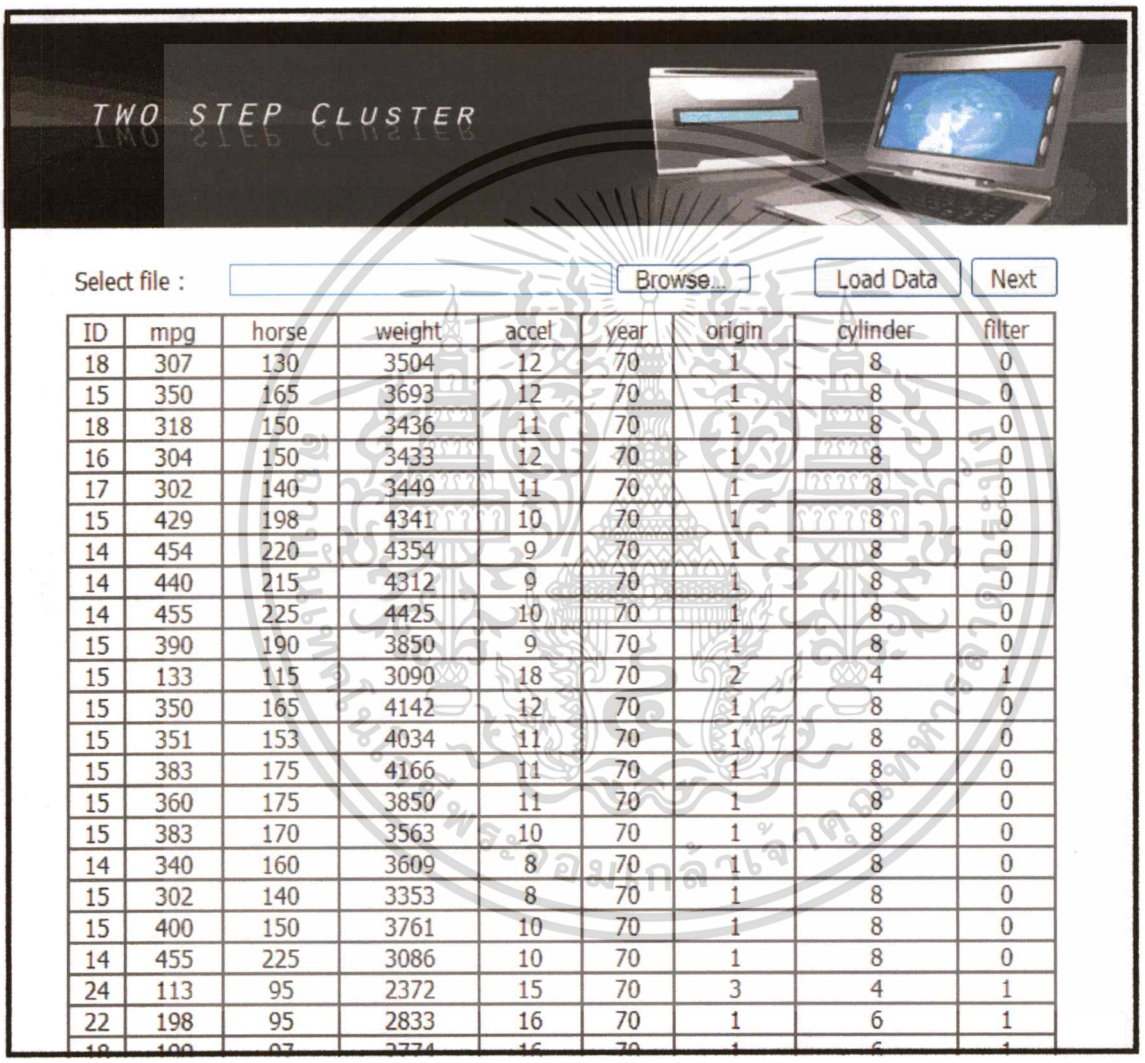
หลังจากนั้นผู้ใช้งานระบบจะต้องทำการกำหนดเส้นทางในการเข้าสู่ชุดข้อมูล ดังรูปที่ 4.9

รูปที่ 4.9 แสดงหน้าจอการเลือกชุดข้อมูลเข้าสู่ระบบ

กด browse เพื่อเลือกข้อมูลเข้าสู่ระบบ

กด load data เพื่อนำข้อมูลเข้าสู่ระบบ จะแสดงข้อมูลออกมาดังรูป 4.10

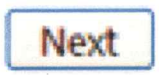




ID	mpg	horse	weight	accel	year	origin	cylinder	filter
18	307	130	3504	12	70	1	8	0
15	350	165	3693	12	70	1	8	0
18	318	150	3436	11	70	1	8	0
16	304	150	3433	12	70	1	8	0
17	302	140	3449	11	70	1	8	0
15	429	198	4341	10	70	1	8	0
14	454	220	4354	9	70	1	8	0
14	440	215	4312	9	70	1	8	0
14	455	225	4425	10	70	1	8	0
15	390	190	3850	9	70	1	8	0
15	133	115	3090	18	70	2	4	1
15	350	165	4142	12	70	1	8	0
15	351	153	4034	11	70	1	8	0
15	383	175	4166	11	70	1	8	0
15	360	175	3850	11	70	1	8	0
15	383	170	3563	10	70	1	8	0
14	340	160	3609	8	70	1	8	0
15	302	140	3353	8	70	1	8	0
15	400	150	3761	10	70	1	8	0
14	455	225	3086	10	70	1	8	0
24	113	95	2372	15	70	3	4	1
22	198	95	2833	16	70	1	6	1

รูปที่ 4.10 แสดงชุดข้อมูลที่เลือกเข้าสู่ระบบ

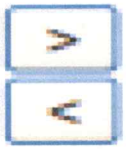
หลังจากโหลด Load Data เพื่อแสดงข้อมูลจากไฟล์ Excel ที่เลือก หลังจากนั้นกด Next เพื่อเข้าสู่ขั้นตอนถัดไปโดยจะแสดงหน้าจอถัดไปดังรูป 4.11

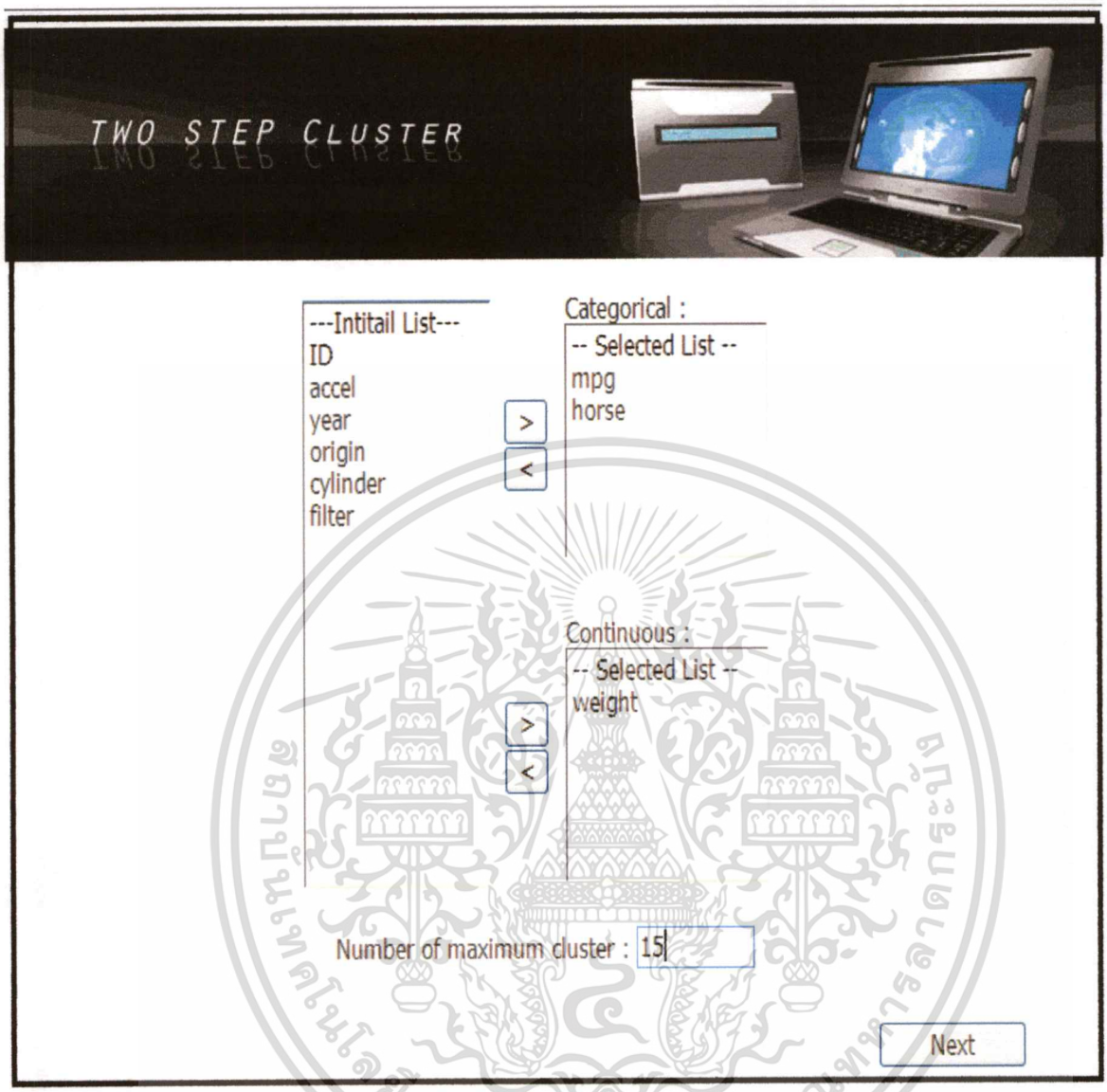


รูปที่ 4.11 แสดงหน้าจอการเลือกตัวแปรมาใช้ในการวิเคราะห์ข้อมูล

เลือกข้อมูล โดยข้อมูลจะถูกแบ่งออกเป็นข้อมูล 2 ประเภทคือข้อมูล ประเภท Categorical และ ประเภท Continuours

วิธีการเลือกสามารถทำได้โดยการ ไฮไลท์ ตัวแปรที่สนใจจากนั้นเลือกตัวแปรที่เราสนใจโดยระบุประเภทของตัวแปร โดยใช้ ปุ่ม > < ในการเลือกข้อมูลมาวิเคราะห์

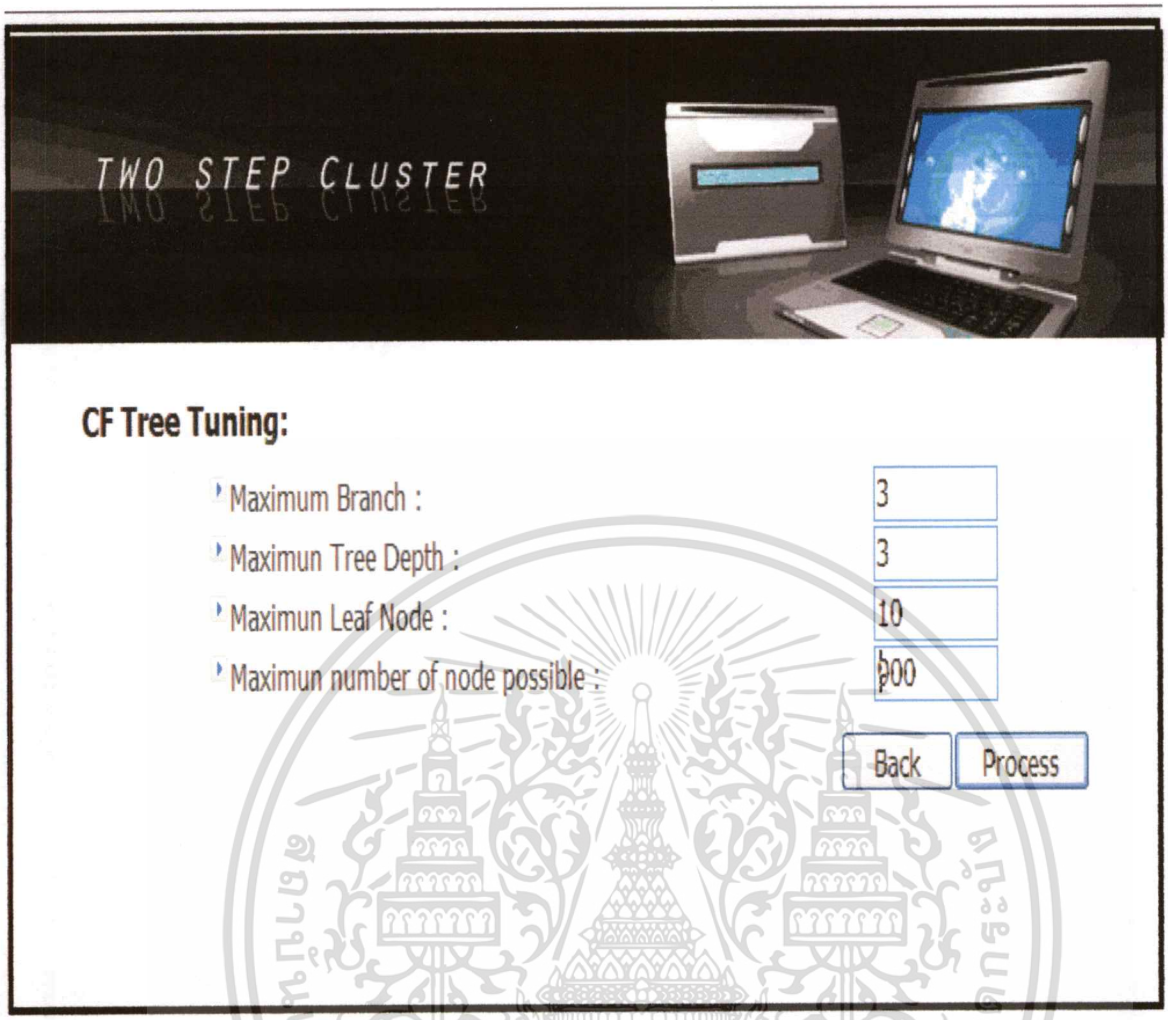




รูปที่ 4.12 แสดงหน้าจอส่วนของการเพิ่มข้อมูล จำนวนกลุ่มมากที่สุด

หลังจากที่กำหนดประเภทตัวแปรเรียบร้อยแล้ว จะต้องทำการระบุจำนวนกลุ่มที่ต้องการสูงที่สุด โดยจะต้องระบุไปในช่อง ดังแสดงรูปที่ 4.12

หลังจากนั้นกดปุ่ม NEXT เพื่อเข้าสู่หน้าจอถัดไป ดังรูปที่ 4.13



รูปที่ 4.13 แสดงหน้าจอการกำหนดข้อมูลตัวแปรต่างๆ

กำหนดตัวแปร

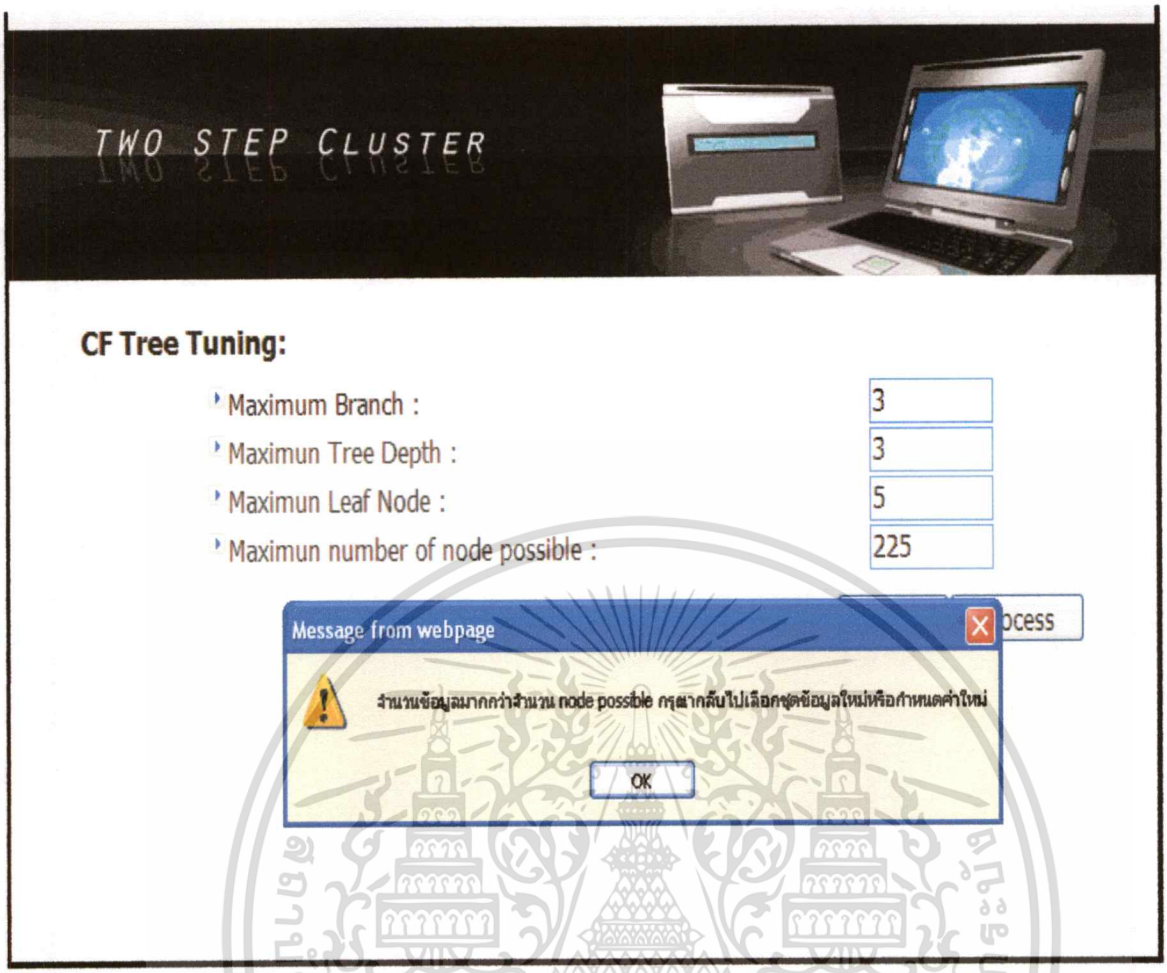
Maximum Branch คือ จำนวนสมาชิกสูงสุดที่เป็นไปได้ใน Non leaf node

Maximum Tree Depth คือ ความสูงของ Tree หรือ Tree Level

Maximum Leaf node คือ จำนวนสมาชิกสูงสุดที่เป็นไปได้ใน Non leaf node

โดนในส่วนของค่า Maximum number of node possible จะถูกคำนวณขึ้นมาเองโดยใช้สมการ

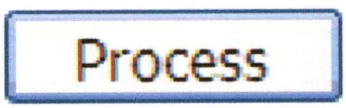
$$Possiblenode = (leaf - nonleaf)^{Level-1}$$



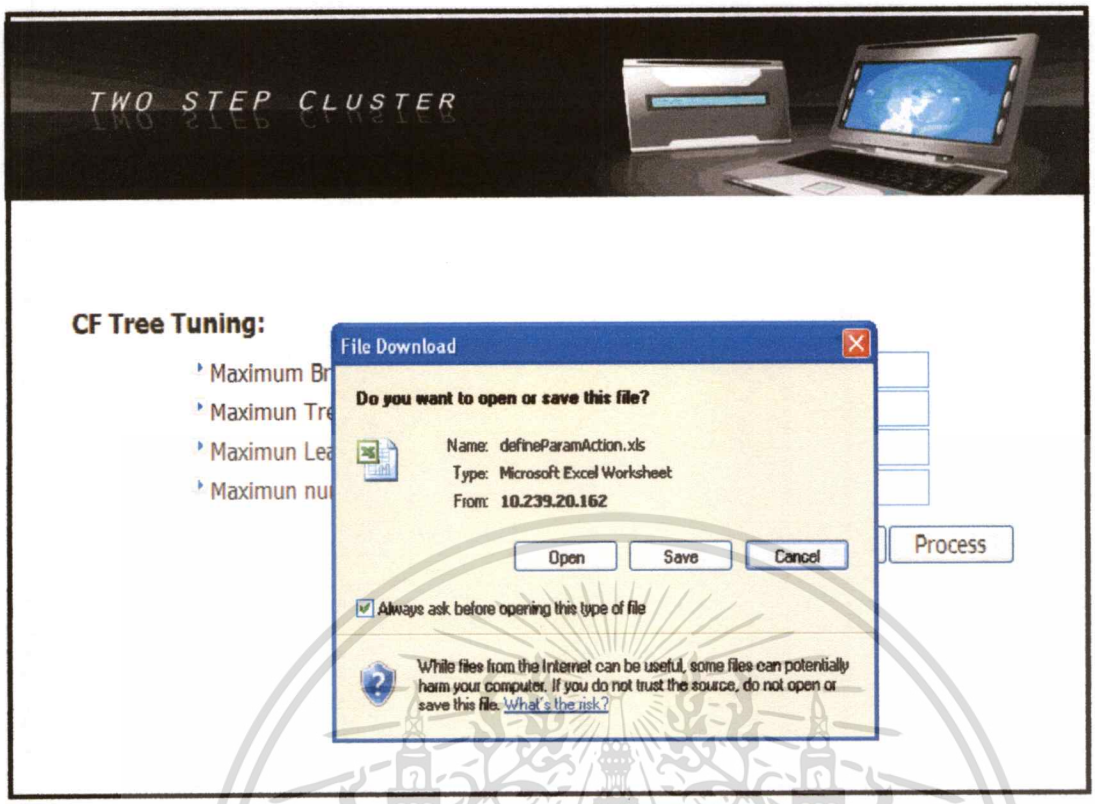
รูปที่ 4.14 แสดงหน้าจอการเตือนเมื่อตัวแปรที่กำหนดไม่สามารถนำไปคำนวณได้

โดยกรณีที่ ชุดข้อมูลมี 408 ตัว แต่กรอกข้อมูลแล้วคำนวน Possible node ได้ 225 จะทำให้ไม่สามารถคำนวณได้จะมีการเตือนดังแสดงในรูป 4.14

หลังจากการระบุค่าตัวแปรต่างเสร็จเรียบร้อยแล้วจะทำการประมวลผลโปรแกรมโดยการกดปุ่ม

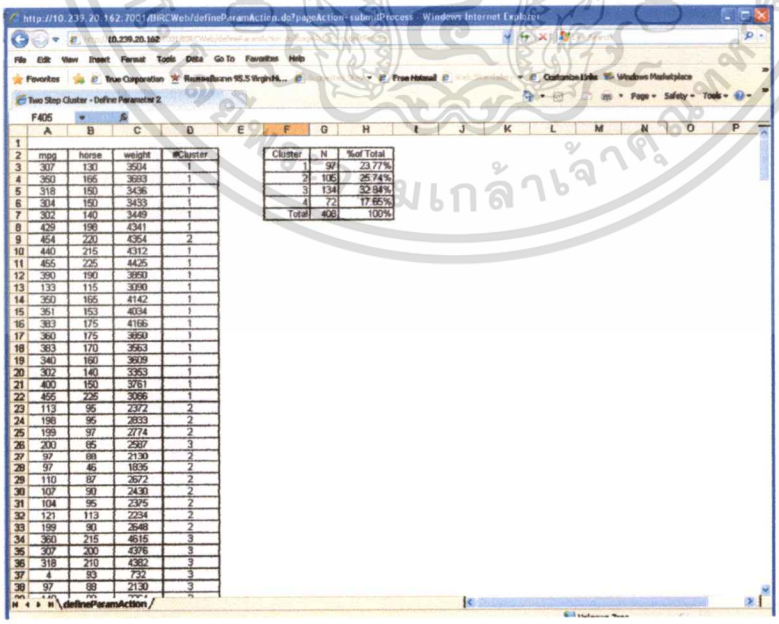


หลังจากนั้นระบบจะแสดงหน้าจอการเลือกจัดเก็บผลลัพธ์ หรือ จะแสดงผลลัพธ์ โดยจะแสดงดังรูปที่ 4.15



รูปที่ 4.15 รูปแสดงการแสดงค่าผลลัพท์ในรูปแบบ excel

แสดงหน้าจอผลลัพท์เป็น excel ใน internet explorer ให้เลือก open เพื่อเปิดไฟล์หรือ Save เพื่อบันทึกไฟล์



รูปที่ 4.16 แสดงผลลัทธ์ของระบบในรูปแบบ IE

ine	accel	mpg	Cluster
.0	15.0	27.0	1
.0	18.0	25.0	1
.0	19.0	31.0	1
.0	17.0	28.0	1
.0	20.0	29.0	1
.0	19.0	31.0	1
.0	17.0	29.0	1
.0	19.0	32.0	1
.0	18.0	33.0	1
.0	18.0	26.0	1
.0	14.0	29.0	1
.0	16.0	28.0	1
.0	15.0	30.0	1
.0	15.0	29.0	1
.0	17.0	31.0	1
.0	14.0	36.0	1
.0	15.0	31.0	1
.0	14.0	32.0	1
.0	13.0	34.0	1
.0	15.0	42.0	1
.0	16.0	34.0	1
.0	15.0	32.0	1
.0	22.0	44.0	1

Cluster	BIC	BIC Change
1	11030.126	
2	9830.333	-1199.793
3	9462.419	-367.914
4	9384.144	-78.275
5	9513.965	129.821
6	9671.102	157.136
7	9771.038	99.937
8	10038.989	267.951

Cluster	N	%of Total
1	131	32.27%
2	73	17.98%
3	100	24.63%
4	102	25.12%
Total	406	100%

Cluster	horse		engine		accel	
	STD	MEAN	STD	MEAN	STD	MEAN
1	94.06	187.02	71.80	155.93	16.95	6.22
2	233.26	392.22	99.48	169.54	16.88	3.23
3	137.62	299.72	94.48	188.21	15.98	5.92
4	349.70	1844.90	160.54	677.03	12.72	3.54

รูปที่ 4.17 แสดงผลลัพท์ของระบบ

ระบบจะแสดงผลลัทธ์ทางหน้าจอในรูปแบบ Excel ในรูปของ IE ดังรูปที่ 4.16-4.17

4.4 การทดสอบระบบ

การทดสอบประสิทธิภาพการใช้งานของระบบ ข้อมูลชุดที่ 1

การทดสอบประสิทธิภาพการใช้งานของระบบได้มีการตรวจสอบผลลัพธ์ของข้อมูลโดยใช้ข้อมูลของ UCI โดยชุดข้อมูลที่เลือกมาทดสอบโดยใช้ข้อมูลชุด Wine Quality ที่มี 4899 เร็ดคอร์ด โดยทำการทำสอบโดยการเปรียบเทียบข้อมูลระหว่าง ผลลัพธ์ที่ได้จากระบบที่พัฒนาขึ้นมาเปรียบเทียบกับผลลัพธ์ของข้อมูลซึ่งอยู่ในรูปของคะแนนคุณภาพของไวท์

ข้อมูลชุด Wine Quality เป็นชุดข้อมูล มี 12 Attribute ได้แก่

1 - fixed acidity ความเป็นกรดคงที่

2 - volatile acidity ค่าความเป็นกรดระเหย

3 - citric acid กรดซิตริก

4 - residual sugar น้ำตาลที่เหลือ

5 - chlorides คลอไรด์

6 - free sulfur dioxide ก๊าซซัลเฟอร์ไดออกไซด์อิสระ

7 - total sulfur dioxide ก๊าซซัลเฟอร์ไดออกไซด์ทั้งหมด

8 - density ความหนาแน่น

9 - pH ค่าความเป็นกรด-ด่าง

10 - sulphates ซัลเฟต

11 - alcohol แอลกอฮอล์

12 – quality คุณภาพ (มีค่าระหว่าง 0 ถึง10)

โดย Attribute ที่ 1-11 เป็น Attribute ที่บอกถึงปัจจัยในการผลิต ส่วน Attribute ที่ 12 quality คุณภาพ เป็น Attribute ที่บอกถึงคุณภาพของไวน์ที่ได้

วิธีการทดสอบจะทดสอบโปรแกรมโดยการกำหนดค่าตัวแปรต่างๆดังนี้

กำหนดให้

รายการ Attribute	ประเภทตัวแปร
fixed acidity ความเป็นกรดคงที่	Continuous
volatile acidity ค่าความเป็นกรดระเหย	Continuous
citric acid กรดซิตริก	Continuous
residual sugar น้ำตาลที่เหลือ	Continuous
chlorides คลอไรด์	Continuous
free sulfur dioxide ก๊าซซัลเฟอร์ไดออกไซด์อิสระ	Continuous
total sulfur dioxide ก๊าซซัลเฟอร์ไดออกไซด์ทั้งหมด	Continuous
density ความหนาแน่น	Continuous
pH ค่าความเป็นกรด-ด่าง	Continuous
Sulphate ซัลเฟต	Continuous
alcohol แอลกอฮอล์	Continuous

จำนวนกลุ่มสูงสุดเป็น 11 จำนวน (แทน 0-10)

Maximum Branch 5

Maximum Tree Depth 5

Maximum Leaf node 5

ได้ผลลัพธ์ดังรูปที่ 4.18

Cluster	Data	% of Total
1	912	18.62%
2	125	2.55%
3	633	12.92%
4	955	19.49%
5	2274	46.42%
Total	4899	100.00%

รูปที่ 4.18 แสดงผลลัพธ์ของระบบที่พัฒนาขึ้นจากข้อมูลชุด Wine Quality

Wine Quality	Cluster	Data	% of Data
9	1	5	0.10%
8	2	175	3.57%
7	3	880	17.96%
6	4	2198	44.87%
5	5	1457	29.74%
4	6	163	3.32%
3	7	20	0.40%
	Tatal	4899	100%

รูปที่ 4.19 แสดงผลรวมสรุปของกลุ่มข้อมูลไวท์ชุด Wine Quality

Result			Check Data
Cluster	Data	% of Data	Wine Quality
1	912	18.60%	6,7,8
2	125	2.60%	3,4,5,6,7
3	633	12.90%	6,7,8,9
4	955	19.50%	6,7,8
5	2274	46.40%	4,5,6,7,8
	4899	100%	

รูปที่ 4.20 แสดงตารางเปรียบเทียบค่ารวมสรุปของกลุ่มข้อมูลไวท์ชุด Wine Quality กับระบบที่พัฒนาขึ้น

อภิปรายผลการทดสอบโปรแกรมที่ได้พัฒนาขึ้นมาเปรียบเทียบกับข้อมูลชุด Wine Quality ที่ได้มีการแบ่งจัดกลุ่มเอาไว้แล้ว จากการทดสอบพบว่าโปรแกรมที่ได้ผลไม่ตรงกับข้อมูลตัวอย่างที่มีการแบ่งกลุ่มไว้แล้ว ความผิดพลาดในการประมวลผลเกิดจากจำนวนกลุ่มย่อยภายในกลุ่มที่ 1(Quality Wine = 9)และ 7(Quality Wine = 3) ที่มีปริมาณของสมาชิกในกลุ่มน้อยคือ 5,20 ตัวตามลำดับ ทำให้เกิดการผิดพลาดขึ้นในการประมวลผล และการกำหนดค่าตัวแปรในการจัดการแผนภูมิต้นไม้

การทดสอบประสิทธิภาพการใช้งานของระบบ ข้อมูลชุดที่ 2

การทดสอบประสิทธิภาพการใช้งานของระบบได้มีการตรวจสอบผลลัพธ์ของข้อมูลโดยใช้ข้อมูลของ UCI โดยชุดข้อมูลที่เลือกมาทดสอบโดยใช้ข้อมูลชุด Nursery ที่มี 12,960 เร็คคอร์ด โดยทำการทำสอบโดยการเปรียบเทียบข้อมูลระหว่าง ผลลัพธ์ที่ได้จากระบบที่พัฒนาขึ้นมาเปรียบเทียบกับผลลัพธ์ของข้อมูลซึ่งอยู่ในรูปของคะแนนคุณภาพของNursery

โดยข้อมูลชุด Nursery นี้เกิดจากแบบจำลองการตัดสินใจในการเลือกโรงเรียน Nursery ข้อมูลชุด Nursery เป็นชุดข้อมูล มี 8Attribute ได้แก่

Parents	อาชีพของผู้ปกครอง (ปกติ , ดี , ดีมาก)
Has_nurs	(เหมาะสม,ไม่ค่อยเหมาะสม,ไม่เหมาะสม,เสี่ยง,เสี่ยงอย่างมาก)
Form	ลักษณะของครอบครัว (สมบูรณ์, ไม่สมบูรณ์, อุปถัมภ์)
Children	จำนวนเด็ก (1,2,3,...)
Housing	สภาวะบ้านที่อยู่อาศัย (สะดวก , ไม่ค่อยสะดวก , ไม่สะดวก)
Finance	สถานะทางการเงิน (ดี , ไม่ดี)
Social	สภาพสังคมสังคม (มีปัญหา , ไม่ค่อยมีปัญหา , ไม่มีปัญหา)
Health	สภาวะสุขภาพอนามัย (ดี , ธรรมดา , ไม่ดี)

วิธีการทดสอบจะทดสอบโปรแกรมโดยการกำหนดค่าตัวแปรต่างๆดังนี้

กำหนดให้

ตัวแปร	ประเภท
Parents	Categorical
Has_nurs	Categorical
Form	Categorical
Children	Categorical

Housing	Categorical
Finance	Categorical
Social	Categorical
Health	Categorical

จำนวนกลุ่มสูงสุดเป็น15 จำนวน

Maximum Branch 5

Maximum Tree Depth 5

Maximum Leaf node 5

ได้ผลลัพธ์ดังรูปที่ 4.20

Cluster	Data	% of Data
1	400	2.53%
2	3891	31.20%
3	4517	33.33%
4	4152	32.92%
Total	12960	100%

รูปที่ 4.21 แสดงผลลัพธ์ของระบบที่พัฒนาขึ้นจากข้อมูลชุด Nursery

Cluster	Data	% of Data
very_recom	328	2.53
spec_prior	4044	31.2
not_recom	4320	33.33
priority	4266	32.92
recommend	2	.02
Total	12960	100

รูปที่ 4.22 แสดงผลรวมสรุปของกลุ่มข้อมูลไวท์ชุด Nursery

อภิปรายผลการทดสอบโปรแกรมที่ได้พัฒนาขึ้นมาเปรียบเทียบกับข้อมูลชุด Nursery ที่ได้มีการแบ่งจัดกลุ่มไว้ จากการทดสอบพบว่าโปรแกรมที่ได้ผลไม่ตรงกับข้อมูลตัวอย่างกล่าวคือที่มีการจัดกลุ่มไว้จริงดังผลลัพธ์ในรูป 4.20 – 4.21 โดยในกลุ่มที่ โดยใช้โปรแกรมที่พัฒนาขึ้นมาแบ่งกลุ่มได้ 4 กลุ่มซึ่งข้อมูลจริงแบ่งออกเป็น 5 กลุ่ม ความผิดพลาดในการประมวลผลเกิดจากจำนวนกลุ่มย่อยภายในการคำนวณ กลุ่มข้อมูล Recommend มีขนาดเล็กมากมีเพียง 2 เร็คคอร์ดหรือคิดเป็น 0.02 % ของข้อมูลทั้งหมด อาจจะทำให้เกิดความผิดพลาดอันเนื่องมาจากข้อจำกัดในการกำหนดตัวแปรต่างๆในการสร้างแผนภูมิต้นไม้

บทที่ 5
สรุป และข้อเสนอแนะ

ผลที่ได้จากการทำโครงงานพัฒนาระบบงานในหัวข้อเรื่อง "การพัฒนาระบบการแบ่งกลุ่มข้อมูลด้วย Two Step Clustering " สามารถสรุปผลได้ดังนี้

5.1 สรุปผลการดำเนินงาน

จากการศึกษาโครงการพัฒนาระบบงานเรื่องการศึกษาอัลกอริทึมในการแบ่งกลุ่มข้อมูลด้วย Two Step Clustering สามารถสรุปผลการดำเนินงานได้ดังนี้

1. ในการนำเทคโนโลยีการทำเหมืองข้อมูล เพื่อนำมาใช้ออกแบบและพัฒนาระบบ ทำให้เกิดความรู้ความเข้าใจในระบบการทำเหมืองข้อมูลที่ดีขึ้น
2. ได้มีการนำเอาความรู้จากการศึกษา เกี่ยวกับการแบ่งกลุ่มข้อมูล โดยใช้ Twostep Clustering ไปใช้ในการสามารถสร้างระบบแบ่งกลุ่มด้วย Twostep Clustering

5.2 ประโยชน์ที่ได้รับจากการศึกษาและพัฒนาระบบ

1. ทำให้เข้าใจการทำเหมืองข้อมูล เพื่อนำมาใช้ออกแบบและพัฒนาระบบ
2. การเรียนรู้วิธีการและขั้นตอนในการทำเหมืองข้อมูลโดยที่เน้นไปที่Two step Clustering ทำให้เกิดความรู้ความเข้าใจในระบบการทำงานของระบบ Two step Clustering ได้เป็นอย่างดี
3. สามารถสร้างระบบแบ่งกลุ่มด้วย Twostep Clustering

5.3 ปัญหาที่พบในการพัฒนา

เนื่องจากโปรแกรม BEA ที่นำมาใช้ในการพัฒนาระบบนี้ เป็นโปรแกรมเฉพาะทางที่ช่วยเพิ่มประสิทธิภาพในการพัฒนาให้ระบบให้มีประสิทธิภาพมากขึ้น ทำให้ต้องใช้เวลาในทำความเข้าใจกับโปรแกรมเป็นเวลามาก

5.4 ข้อเสนอแนะ

เพื่อปรับปรุงโปรแกรมประยุกต์นี้ในอนาคต ผู้ศึกษามีความคิดเห็นว่า ควรมีการตรวจสอบความถูกต้องของผลที่ได้ การแสดงผลในรูปแบบกราฟ แผนภูมิวงกลม ข้อมูลทางสถิติอื่นๆ ระยะเวลาในการประมวลผล ซึ่งยังใช้เวลาค่อนข้างนานซึ่งจำเป็นต้องใช้ผู้ที่มีความเชี่ยวชาญในการพัฒนาระบบ

บรรณานุกรม

Chiu, T., Fang, D., Chen, J., Wang, Y., and Jeris, C. 2001. **A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment.** Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining, 263.

Zhang, T., Ramakrishnon, R., and Livny M. 1996. **BIRCH: An Efficient Data Clustering Method for Very Large Datebases**. Proceedings of the ACM SIGMOD Conoference on Management of Data, p. 103-114, Montreal, Canada.

ประวัติผู้เขียน

ชื่อผู้เขียน	นายทวีชัย ปิยะตานนท์
วันที่เกิด	วันที่ 13 กรกฎาคม 2527
วุฒิการศึกษาระดับ	ปริญญาตรี วิทยาศาสตร์บัณฑิต
สถานที่สำเร็จการศึกษา	คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ปีที่สำเร็จการศึกษา	2548