สื่อการสอนบนเครือข่ายอินเทอร์เน็ต
เรื่องการทำเหมืองข้อมูลขั้นพื้นฐาน

นายนิวัฒน์ ไทเศรษฐวัฒน์กุล
นายรติ เพิ่มพูล
นายสรรเพชญ ภุมรินทร์

ปัญหาพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต
ภาควิชาสถิติประยุกต์
คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ปีการศึกษา 2548

Web-based Media for e-Learning on the Introduction to Data Mining

Mr.NIWAT THAISETTAWATKUL

Mr.RATI POEMPOOL

Mr.SANPETCH POOMMARIN

A Special Project Submitted in Partial Fulfillment of the

Requirement for the Degree of Bachelor of Science

Department of Applied Statistics

Faculty of Science

King Mongkut's Institute of Technology Ladkrabang

Academic Year 2005

ปัญหาพิเศษเรื่อง	สื่อการสอนบนเครือข่ายอินเทอร์เน็ตเรื่องการทำเหมืองข้อมูลขั้นพื้นฐาน
นักศึกษา	นายนิวัฒน์ ไทเศรษฐวัฒน์กุล
	นายรติ เพิ่มพูล
	นายสรรเพชญ ภุมรินทร์
ภาควิชา	สถิติประยุกต์
สาขาวิชา	สถิติประยุกต์
อาจารย์ที่ปรึกษา	ดร.รุจิเรข บุศราวงศ์

ภาควิชาสถิติประยุกต์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง อนุมัติให้ปัญหาพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต

คณะกรรมการตรวจสอบ		ลายมือชื่อ
ประธานกรรมการ	ดร.รุจิเรข บุศราวงศ์	
กรรมการ	ผศ.ดร.วลัยลักษณ์ อัตธีรวงศ์	
กรรมการ	ดร.สมศรี บัณฑิตวิไล	

.....

ผศ.ดร.มนัส ไพฑูรย์เจริญลาภ

หัวหน้าภาควิชา

ปัญหาพิเศษเรื่อง สื่อการสอนบนเครือข่ายอินเทอร์เน็ตเรื่องการทำเหมืองข้อมูลขั้นพื้นฐาน

นักศึกษา นายนิวัฒน์ ไทเศรษฐวัฒน์กุล
นายรติ เพิ่มพูล
นายสรรเพชญ ภุมรินทร์

ภาควิชา สถิติประยุกต์ คณะวิทยาศาสตร์

สาขาวิชา สถิติประยุกต์

ปีการศึกษา 2548

อาจารย์ที่ปรึกษา ดร.รุจิเรข บุศราวงศ์

บทคัดย่อ

การศึกษาเรื่อง "สื่อการสอนบนเครือข่ายอินเทอร์เน็ตเรื่อง การทำเหมืองข้อมูลขั้นพื้นฐาน" เกิดจากการที่ผู้ศึกษาเห็นว่า Data Mining เป็นระบบฐานข้อมูลที่มีประโยชน์ในการสนับสนุนการตัดสินใจ และแก้ไขปัญหาต่างๆ ขององค์กร อย่างไรก็ตาม ยังไม่มีการแปลและเรียบเรียงเกี่ยวกับ Data Mining เป็นภาษาไทย

การศึกษานี้มีวัตถุประสงค์เพื่อรวบรวมข้อมูลที่เกี่ยวข้องกับ Data Mining ถอดความอย่างละเอียดเป็นภาษาไทย และบรรจุลงในเว็บไซต์ที่จัดทำขึ้น รวมทั้งเพื่อให้เป็นแหล่งข้อมูลทางวิชาการสำหรับผู้ที่สนใจได้ทำความเข้าใจเกี่ยวกับ Data Mining ขั้นพื้นฐาน และสามารถนำความรู้บนเครือข่ายอินเทอร์เน็ตไปใช้ในการปฏิบัติงานที่เกี่ยวข้องได้ด้วยตนเอง

สื่อการสอนบนเครือข่ายอินเทอร์เน็ตเรื่อง การทำเหมืองข้อมูลขั้นพื้นฐาน จัดทำขึ้นเป็นภาษาไทย โดยพัฒนาขึ้นจากโปรแกรม Macromedia Dreamweaver MX 2004, Photoshop CS และสามารถเรียกใช้งานได้บน Microsoft Internet Explorer Version 5.0 ขึ้นไป

เนื้อหาของสื่อการสอนแบ่งออกได้เป็น 2 ส่วน คือ ความรู้ทั่วไปเกี่ยวกับ Data Mining และประเภทของวิธีการเรียนรู้และเทคนิคต่างๆ ใน Data Mining พร้อมทั้งยกตัวอย่างของการทำ Data Mining ประกอบไว้ด้วย

Special Project Title	Web-based Media for e-Learning on the Introduction to Data Mining
Name	Mr. Niwat Thaisettawatkul
	Mr. Rati Poempool
	Mr. Sanpetch Poommarin
Department	Applied Statistics Faculty of Science
Program	Applied Statistics
Academic Year	2005
Special Project Advisor	Dr. Rujirek Boosarawongse

ABSTRACT

The study on the "Web-based Media for e-Learning on the Introduction to Data Mining" emerges from the idea that Data Mining is the database system that is very useful for an organization to make a decision related to data. However, there is no information about Data Miming in Thai.

This study therefore aims to collect data about Data Mining in English and translate it into Thai in order to put into the website to be the academic source of information for interested persons. The on-line information is hoped to be useful for users to apply Data Mining to works by him/herself.

The web-based media for e-Learning on the Introduction to Data Mining is created by using Macromedia Dreamweaver MX 2004, Photoshop CS and supported by Microsoft Internet Explorer Version 5.0 or above.

The outline of this study can be divided into 2 parts: the introduction to Data Mining and the type of learning method and technique of Data Mining. In order to make them understandable, examples of how to use Data Mining are added.

กิตติกรรมประกาศ

ขอขอบพระคุณ ดร.รุจิเรข บุศราวงศ์ อาจารย์ที่ปรึกษาในการทำปัญหาพิเศษฉบับนี้ ที่ได้กรุณาให้ความรู้ ให้คำปรึกษาและให้คำแนะนำต่างๆ ตลอดจนตรวจแก้ไขข้อบกพร่อง ด้วยความเอาใจใส่อย่างดียิ่งตลอดมา

ผู้จัดทำขอขอบพระคุณ ผศ.ดร.วลัยลักษณ์ อัตธีรวงศ์ และ ดร.สมศรี บัณฑิตวิไล กรรมการในการทำปัญหาพิเศษฉบับนี้ที่ได้ให้ความกรุณาตรวจ ปรับปรุง ให้คำแนะนำและแก้ไข จนสามารถนำมาเป็นแนวทางในการปรับปรุงให้สมบูรณ์

ขอขอบพระคุณอาจารย์ทุกท่านที่ได้ประสิทธิประสาทวิชาความรู้ตั้งแต่อดีตจนถึงปัจจุบัน โดยเฉพาะคณาจารย์ภาควิชาสถิติประยุกต์ทุกท่านที่ได้มอบความรู้ในศาสตร์ที่สำคัญ

ขอขอบคุณเจ้าหน้าที่ภาควิชาสถิติประยุกต์ทุกท่านที่ให้ความอนุเคราะห์ในการช่วยเหลือในด้านเอกสารและอุปกรณ์ในการทำปัญหาพิเศษฉบับนี้

ท้ายที่สุดนี้ขอกราบขอบพระคุณ บิดา – มารดา และผู้มีอุปการะของผู้จัดทำที่ให้ความอนุเคราะห์คอยให้ความช่วยเหลือและสนับสนุนงบประมาณ และคอยสนับสนุนในทุกด้าน ทำให้เป็นแรงใจในการฟันฝ่าอุปสรรคต่างๆ ผ่านพ้นไปด้วยดี

นายนิวัฒน์ ไทเศรษฐวัฒน์กุล

นายรติ เพิ่มพูล

นายสรรเพชญ ภุมรินทร์

สารบัญ

สารบัญ (ต่อ)

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญรูป

สารบัญรูป (ต่อ)

หน้า

สารบัญรูปภาคผนวก

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มา

ในปัจจุบัน Data Mining มีการใช้กันอย่างแพร่หลายในวงการอุตสาหกรรมสาขาต่างๆ ซึ่งการแพร่หลายอย่างรวดเร็วของ Data Mining เป็นสิ่งสำคัญที่สะท้อนให้เห็นถึงขอบเขตของการแพร่หลาย และเทคนิคต่างๆที่สามารถนำไปประยุกต์สู่ปัญหา โดยเราสามารถจะนำผลลัพธ์ที่ได้มาประยุกต์ใช้ในการแก้ไขปัญหาต่างๆของธุรกิจและองค์กร อีกทั้งยังเป็นสิ่งที่ช่วยสนับสนุนในการตัดสินใจขององค์กร ด้วยการสกัดความรู้ที่มีประโยชน์ซึ่งซ่อนอยู่ในฐานข้อมูลขนาดใหญ่ โดยใช้เทคนิคต่างๆของ Data Mining ดังนั้น Data Mining จึงเป็นสิ่งสำคัญที่ควรค่าแก่การศึกษาเรียนรู้

เนื่องจากหนังสือเกี่ยวกับ Data Mining ในปัจจุบันยังไม่มีการแปลและเรียบเรียงเป็นภาษาไทย จึงทำให้เป็นการจำกัดขอบเขตของความรู้ดังกล่าว คณะผู้จัดทำได้เล็งเห็นถึงปัญหานี้จึงได้จัดทำสื่อการสอนเรื่อง Data Mining ขั้นพื้นฐาน ผู้ที่สนใจจะศึกษาสามารถทำความเข้าใจได้ง่ายและรวดเร็วขึ้น โดยคณะผู้จัดทำจะทำการผลิตสื่อการสอนดังกล่าวบนเครือข่ายอินเทอร์เน็ต เนื่องจากอินเทอร์เน็ตได้ถูกนำมาใช้งานเป็นสื่อกลางในการสื่อสารอย่างแพร่หลาย จึงเห็นสมควรที่จะนำมาใช้เป็นสื่อกลางในการศึกษา เพื่อเป็นการเพิ่มขีดความสามารถของการศึกษาในเรื่องดังกล่าว

1.2 วัตถุประสงค์ที่ศึกษา

1. สร้างเว็บไซต์เพื่อใช้เป็นสื่อการสอนเรื่อง การทำเหมืองข้อมูลขั้นพื้นฐาน
2. เพื่อให้เป็นแหล่งข้อมูลทางวิชาการสำหรับผู้ที่สนใจได้ทำความเข้าใจเกี่ยวกับเนื้อหาเรื่องการทำเหมืองข้อมูลขั้นพื้นฐาน และใช้ความรู้ดังกล่าวมาช่วยในการวิเคราะห์ข้อมูลด้วยตนเอง

1.3 ประโยชน์ที่คาดว่าจะได้รับ

เพื่อเป็นแหล่งค้นคว้าสำหรับเรื่อง การทำเหมืองข้อมูลและสามารถนำความรู้ดังกล่าวมาใช้ในการวิเคราะห์ข้อมูลด้วยตนเองได้

1.4 ขอบเขตของการศึกษา

สื่อการสอนจะรวบรวมเนื้อหาเกี่ยวกับการทำเหมืองข้อมูลขั้นพื้นฐาน โดยจะจัดแบ่งเนื้อหาที่แสดงดังต่อไปนี้

1. ความรู้เบื้องต้นเกี่ยวกับการทำเหมืองข้อมูล
2. ประเภทของวิธีการเรียนรู้และเทคนิคต่างๆในการทำเหมืองข้อมูล
   - 2.1 Supervised Learning
     - 2.1.1 การแบ่งหมวดหมู่เป็น 2 class (Classification into 2 classes)
       - 2.1.1.1 Division Regression
       - 2.1.1.2 Logistic Regression
     - 2.1.2 การแบ่งหมวดหมู่มากกว่า 2 class (Classification into more than 2 classes)
       - 2.1.2.1 Predictive Regression
       - 2.1.2.2 Bayesian Inference
       - 2.1.2.3 ต้นไม้ตัดสินใจ (Decision Tree) เช่น ID3, C 5.0, CART
   - 2.2 Unsupervised Learning
     - 2.2.1 การจัด Cluster (Clustering)
       - 2.2.1.1 การจัด Cluster แบบลำดับชั้น (Hierarchical Algorithms) เช่น การจัด Cluster แบบลำดับชั้นจากล่างขึ้นบน (Agglomerative Algorithms) และ การจัด Cluster แบบลำดับชั้นจากบนลงล่าง (Divisive Clustering)
       - 2.2.1.2 การจัด Cluster แบบแบ่งเป็นส่วน (Partitional Algorithms) เช่น Minimum Spanning Tree, K - Means Clustering
     - 2.2.2 Summarization เช่น Mean, Median, Mode, Variance, Standard Deviation
     - 2.2.3 กฎของความสัมพันธ์ (Association Rules) เช่น Market – Basket Analysis, Apriori Algorithms

1.5 ขั้นตอนการดำเนินงาน

1. ศึกษาความเป็นไปได้ของโครงร่างปัญหาพิเศษ
   1.1 กำหนดหัวข้อเรื่องที่จะศึกษา
   1.2 ศึกษางานวิจัยและทฤษฎีที่เกี่ยวข้อง
2. นำเสนอโครงร่างปัญหาพิเศษและ ปรับปรุงแก้ไข
3. ศึกษาและฝึกวิธีการใช้เครื่องมือในการพัฒนาโปรแกรมและขอบเขตความสามารถของโปรแกรม
4. จัดทำและส่งโครงร่างปัญหาพิเศษฉบับสมบูรณ์
5. รวบรวมเนื้อหาและบทความที่เกี่ยวข้องสำหรับการจัดทำสื่อการสอน
6. ออกแบบ ลงมือสร้างและพัฒนาสื่อการสอน
7. ตรวจสอบความถูกต้องและแก้ไขสื่อการสอน
8. นำโปรแกรมไปติดตั้งบนเว็บไซต์
9. จัดทำรายงานและรูปเล่มปัญหาพิเศษ

1.6 อุปกรณ์ที่ใช้ในการทำปัญหาพิเศษ

1. โปรแกรม Microsoft Office 2003
2. โปรแกรม Microsoft Visio 2000
3. โปรแกรม Adobe Photoshop CS
4. โปรแกรม Adobe Illustrator CS
5. โปรแกรม Macromedia Dreamweaver MX 2004
6. เครื่องคอมพิวเตอร์

บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

2.1 e – Learning

e - Learning ย่อมาจากคำว่า Electronic Learning หมายถึง การใช้ทรัพยากรต่างๆ ในระบบอินเทอร์เน็ตมาออกแบบและจัดระบบเพื่อสร้างระบบการเรียนการสอน โดยการสนับสนุนและส่งเสริมให้เกิดการเรียนรู้อย่างมีความหมายตรงกับความต้องการของผู้สอนและผู้เรียน เชื่อมโยงระบบเป็นเครือข่ายที่สามารถเรียนรู้ได้ทุกคน ทุกที่ และทุกเวลา (ที่มา http:/www.nectec.or.th /courseware/cai/0018.html)โดยสามารถพิจารณาได้จากคุณลักษณะของ e – Learning ซึ่งมี ดังนี้

1. เว็บไซต์ที่เกี่ยวข้องกับการศึกษา
2. เว็บไซต์ที่เกี่ยวข้องการเนื้อหารายวิชาใด วิชาหนึ่งเป็นอย่างน้อย
3. ผู้เรียนสามารถเรียนรู้ได้ตนเองโดยอิสระ ทุกที่ทุกเวลา
4. ผู้เรียนมีอิสระในการเรียน การบรรลุจุดประสงค์การเรียนรู้แต่ละเนื้อหาไม่จำเป็นต้องเรียนพร้อมกับผู้เรียนรายอื่น
5. มีเครื่องมือที่วัดผลการเรียนได้
6. มีการออกแบบการเรียนการสอนอย่างมีระบบ

ดังนั้นจะเห็นได้ว่า e - Learning เป็นระบบการเรียนการสอนที่เกี่ยวข้องกับเทคโนโลยีเว็บและเครือข่ายอินเทอร์เน็ต มีสภาวะแวดล้อมที่สนับสนุนการเรียนรู้อย่างมีชีวิตชีวา (Active Learning) และการเรียนที่เน้นผู้เรียนเป็นศูนย์กลาง (Child Center Learning) ผู้เรียนเป็นผู้คิดตัดสินใจเรียน โดยการสร้างความรู้และความเข้าใจใหม่ๆด้วยตนเอง สามารถเชื่อมโยงกระบวนการเรียนรู้ให้เข้ากับชีวิตจริง ครอบคลุมการเรียนทุกรูปแบบทั้งการเรียนทางไกลและการเรียนผ่านเครือข่ายระบบต่างๆ

ประโยชน์ของ e – Learning

1. สามารถเรียนผ่านเว็บไซต์ได้ไม่จำกัดอายุ สถานที่ และเวลา เป็นการเรียนที่ยึดถือความสะดวกของผู้เรียนเป็นหลัก ช่วยประหยัดเวลาในการเดินทาง และส่งเสริมการเรียนรู้ตลอดชีวิต ไม่ว่าวัยไหนก็สามารถเรียนได้ ดังคำกล่าวที่ว่า ไม่มีใครแก่เกินเรียน

2. e - Learning เป็นการเรียนโดยเน้นผู้เรียนเป็นศูนย์กลาง ดังนั้น ผู้เรียนสามารถเลือกเรียนรู้ได้ตามที่ตัวเองต้องการ ตัดสินใจเลือกสิ่งที่เหมาะสมกับตัวเองได้ ทั้งเนื้อหาที่จะเรียน สถานที่ในการเรียน และเวลาที่ตัวเองสะดวก ซึ่งมีผลให้ผู้เรียนมีความกระตือรือร้นที่จะเรียนรู้มากขึ้น

3. ในปัจจุบัน มีบทเรียนออนไลน์มากมายจากฐานข้อมูลทั่วโลก และในประเทศไทย ก็มีการเรียนการสอนแบบ e - Learning มากขึ้นเรื่อยๆ รวมทั้งการนำบทเรียนมาเผยแพร่ทางอินเทอร์เน็ตมากขึ้น ทำให้มีการแลกเปลี่ยนความรู้ ความคิดเห็นระหว่างผู้คนมากมาย

4. การเรียนออนไลน์ ช่วยประหยัดค่าใช้จ่ายได้มาก ไม่ว่าจะเป็นค่าเอกสาร ค่าบทเรียน ค่าเดินทาง ฯลฯ และยังมีเว็บไซต์หลายๆ แห่งที่เปิดบริการ e - Learning โดยไม่ต้องเสียค่าใช้จ่ายแต่อย่างใด นับว่าเป็นการเรียนรู้ที่ทุกคนเข้าถึงได้มากขึ้น

2.2 อินเทอร์เน็ต

อินเทอร์เน็ต (Internet) หมายถึง ระบบของการเชื่อมโยงข่ายงานคอมพิวเตอร์ขนาดใหญ่ โดยอาศัยการนำสัญญาณภายใต้กฎเกณฑ์และมาตรฐานเดียวกัน และยังสามารถที่จะทำให้คนจำนวนมากสื่อสารข้อมูลทั้งในรูปแบบของตัวอักษร ข้อความ ภาพ และเสียง ได้อย่างสะดวกรวดเร็วด้วยคอมพิวเตอร์ที่ต่างระบบและต่างชนิดกันได้

อินเทอร์เน็ตในวงการศึกษา ในปัจจุบันนั้นนับว่ามีประโยชน์อย่างมากเนื่องจากมีบนข้อมูลข่าวสารอยู่มากมาย ทำให้สามารถศึกษาค้นคว้าได้มากขึ้นได้ความรู้กว้างยิ่งขึ้น อย่างไรก็ตามผู้ใช้อินเทอร์เน็ตควรที่จะมีการเรียนรู้เกี่ยวกับการใช้บริการอินเทอร์เน็ตและเลือกใช้ให้เหมาะสมเพื่อที่จะใช้ค้นหาความรู้ในการเรียนรู้ด้วยตนเองอย่างมีประสิทธิภาพและสามารถใช้บริการบนอินเทอร์เน็ตในการสืบค้นข้อมูล

2.3 Macromedia Dreamweaver MX 2004

โปรแกรม Macromedia Dreamweaver MX 2004 เป็นโปรแกรมขั้นพื้นฐานที่พัฒนาขึ้นโดยบริษัท Macromedia ซึ่งในปัจจุบันนี้ได้พัฒนาเรื่อยมาจนถึงเวอร์ชัน Macromedia Dreamweaver MX 2004 โดยมีวัตถุประสงค์เพื่อให้การสร้างเว็บไซต์เป็นเรื่องง่าย เนื่องจากโปรแกรม Macromedia Dreamweaver MX 2004 สามารถสร้างInterfaceและยังสามารถแทรกโค้ดเพื่อควบคุมการทำงานหรือใส่ลูกเล่นอื่นๆที่น่าสนใจให้กับเว็บเพจได้ โดยแยกคุณสมบัติที่เพิ่มขึ้นของ Macromedia Dreamweaver MX 2004 ได้ดังนี้

1. สนับสนุนความปลอดภัยในการส่งข้อมูลผ่าน FTP
2. มีการตรวจสอบผ่านคำสั่งแท็ก และขอการใช้คำสั่งผ่านบราวเซอร์ได้
3. เขียนโค้ดได้รวดเร็วขึ้นเพราะมีเครื่องช่วยเมื่อคลิกเมาส์ขวา
4. สามารถเพิ่มการปฏิสัมพันธ์ระหว่างหน้าเว็บเพจด้วย Macromedia Flash ภายใน Macromedia Dreamweaver ได้

2.4 ทฤษฎีเกี่ยวกับ Data Mining

ในปัจจุบันองค์การส่วนใหญ่จะประสบกับปัญหาของการที่มีข้อมูลดิบเป็นจำนวนมากแต่สามารถนำสารสนเทศมาใช้ประโยชน์ได้น้อย เนื่องจากยังขาดความรู้ ความเข้าใจในการนำข้อมูลดังกล่าวมาวิเคราะห์ ทั้งนี้วิธีการวิเคราะห์ข้อมูลทางสถิติเป็นวิธีการที่ยุ่งยาก ซับซ้อนผู้วิเคราะห์ต้องมีความรู้ ความเข้าใจเป็นอย่างดี อีกทั้งวิธีการวิเคราะห์ทางสถิติยังมีข้อจำกัดบางประการ เช่น ข้อสมมติเกี่ยวกับข้อมูล ทำให้เกิดวิธีการข้อมูลแนวใหม่ เรียกว่า การทำเหมืองข้อมูล (Data Mining) ซึ่งในปัจจุบันมีการนำแนวคิดนี้มาประยุกต์ใช้ในทางธุรกิจและอื่นๆอีกมากมาย เช่น การจัดการความสัมพันธ์กับลูกค้า (CRM), การวิเคราะห์ข้อมูลในชีวสารสนเทศ (Bioinformatics) เป็นต้น

Data Mining หรือ การทำเหมืองข้อมูลเป็นกระบวนการค้นหาความรู้ ความสัมพันธ์ และรูปแบบของข้อมูลที่มีอยู่ในฐานข้อมูลขนาดใหญ่โดยอาศัยการผสมผสานแนวคิดทางสารสนเทศและวิธีทางสถิติ ซึ่งประเภทของการเรียนรู้ (Type of Learning Method) สามารถแบ่งได้เป็น 2 ประเภท คือ Supervised Learning และ Unsupervised Learning ดังนี้

1. Supervised Learning

Supervised Learning เป็นกระบวนการเรียนรู้เพื่อหาความสัมพันธ์ของข้อมูลจากข้อมูล input – output ที่ทราบค่าซึ่งเรียกว่า ชุดข้อมูลฝึกฝน (Training Data) ทำหน้าที่เสมือนเป็นผู้สอนและสร้างเป็นระบบการเรียนรู้ (Learning System) เพื่อใช้ในการประมาณค่า output จากข้อมูล input ชุดใหม่ โดยงานที่เกี่ยวข้องกับการเรียนรู้ประเภทนี้คือ

1.1 การจัดแบ่งหมวดหมู่ (Classification)

การจัดแบ่งหมวดหมู่หรือการแยกประเภท ซึ่งใช้ในการจำแนกประเภทของข้อมูล Output ที่เราต้องการโดยอาศัยเทคนิคต่างๆใน Data Mining ดังต่อไปนี้

1.1.1 การแบ่งหมวดหมู่เป็น 2 class (Classification into 2 classes) มี 2 วิธีดังนี้

1.1.1.1 Division Regression

วิธีการนี้ได้นำแนวคิดของ การถดถอย ซึ่งเป็นเทคนิคในทางสถิติในการหาจุดแบ่ง หรือเกณฑ์เพื่อแบ่งข้อมูลออกเป็น 2 class

1.1.1.2 Logistic Regression

วิธีการนี้ได้นำแนวคิดของ Logistic Regression ซึ่งเป็นเทคนิคในทางสถิติมาใช้ในการประมาณความน่าจะเป็นที่จะเกิดความสำเร็จของตัวแปร output จากตัวแปร input เพื่อนำความน่าจะเป็นดังกล่าวมาใช้ในการจัดแบ่งข้อมูลออกเป็น 2 class

1.1.2 การแบ่งหมวดหมู่มากกว่า 2 class (Classification into more than 2 classes) มี 3 วิธีดังนี้

1.1.2.1 Predictive Regression

วิธีการนี้ได้นำแนวคิดของ การถดถอย ซึ่งเป็นเทคนิคในทางสถิติในการหาตัวแบบที่ดีที่สุดเพื่อนำมาใช้ในการพยากรณ์ class จากข้อมูล input ที่กำหนดให้

1.1.2.2 Bayesian Inference

วิธีการนี้ได้นำแนวคิดของ เบย์ ซึ่งเป็นเทคนิคทางสถิติมาใช้ในการหาความน่าจะเป็น เพื่อนำความน่าจะเป็นดังกล่าวไปจัดข้อมูลว่าควรจะอยู่ใน class ใด

1.1.2.3 ต้นไม้การตัดสินใจ (Decision Tree)

วิธีการนี้ได้นำแนวคิดของ ต้นไม้การตัดสินใจ ซึ่งเป็นเทคนิคทางสถิติมาใช้ในการสร้างกฎเกณฑ์เพื่อใช้ในการแบ่ง class ของข้อมูล

2. Unsupervised Learning

Unsupervised Learning เป็นกระบวนการเรียนรู้เพื่อหารูปแบบที่เหมาะสมว่า output ควรเป็นอย่างไร จากลักษณะของข้อมูลในชุดข้อมูลฝึกฝนซึ่งมีเพียงข้อมูล input เท่านั้น โดยงานที่เกี่ยวกับการเรียนรู้ประเภทนี้ คือ

2.1 การจัด Cluster (Clustering)

การจัด Cluster หรือ การแบ่ง Cluster ซึ่งใช้ในการจัดกลุ่มของข้อมูล ซึ่งศึกษาจากข้อมูล inputโดยภายในกลุ่ม (cluster) เดียวกันจะมีความคล้ายคลึงกันมากที่สุดและระหว่าง cluster ต่างกันจะมีความแตกต่างกันมากที่สุด ซึ่งการจัดกลุ่มจะแตกต่างจากการจัดแบ่งหมวดหมู่ (Classification) คือ การจัดกลุ่มพยายามหาความคล้ายคลึงของข้อมูล input เพื่อสร้างกลุ่ม โดยไม่มีการกำหนด class ของข้อมูลเอาไว้ก่อน ซึ่ง Algorithms ในการจัด Cluster ของข้อมูลโดยทั่วไปแบ่งเป็น 2 ประเภท คือ

2.1.1 การจัด Cluster แบบลำดับชั้น (Hierarchical Algorithms)

การจัด Cluster แบบลำดับชั้นนี้จะกำหนดให้ข้อมูลแต่ละค่าเป็น 1 cluster แล้วจัด cluster ที่อยู่ใกล้กันเข้าไว้ด้วยกันเป็น cluster ใหม่ 1 cluster แล้วทำเช่นนี้ต่อไป ซึ่งจะกล่าวถึง 2 วิธีดังนี้

2.1.1.1 การจัด Cluster แบบลำดับชั้นจากล่างขึ้นบน (Agglomerative Algorithms)

2.1.1.2 การจัด Cluster แบบลำดับชั้นจากบนลงล่าง (Divisive Clustering)

2.1.2 การจัด Cluster แบบแบ่งเป็นส่วน (Partitional Algorithms)

การจัด Cluster แบบแบ่งเป็นส่วนเริ่มจากให้ข้อมูลทั้งหมดเป็น cluster เดียว แล้วจึงแบ่งออกเป็น cluster ตามที่ต้องการด้วยเทคนิคต่างๆ

2.2 Summarization

Summarization เป็นกระบวนการที่ใช้ในการวิเคราะห์หาลักษณะเบื้องต้นของข้อมูลในฐานข้อมูลเช่น ค่าเฉลี่ย ค่ามัธยฐาน ค่าความแปรปรวน และส่วนเบี่ยงเบนมาตรฐาน

2.3 กฎของความสัมพันธ์ (Association Rules)

การวิเคราะห์หาความสัมพันธ์ของข้อมูลในฐานข้อมูลเพื่อหากฎความสัมพันธ์หรือความน่าจะเป็นของข้อมูลที่เราสนใจจากฐานข้อมูลที่จะเกิดขึ้นพร้อมกัน

จากที่กล่าวมานี้จะเห็นได้ว่า เทคนิคการวิเคราะห์ข้อมูลของ Data Mining เป็นอีกทางเลือกที่ผู้ใช้สามารถนำมาวิเคราะห์ข้อมูล โดยผู้ใช้จะต้องเลือกให้เหมาะสมกับชนิดของข้อมูล วัตถุประสงค์ในการวิเคราะห์ และลักษณะของข้อมูลที่มีอยู่ อย่างไรก็ตาม Data Mining ยังมี

เทคนิคอื่นๆในการวิเคราะห์ข้อมูลอื่นๆอีก เช่น Neural Network ซึ่งจะเป็นประโยชน์ในการศึกษาต่อไป

บทที่ 3
วิธีการดำเนินงาน

ในการดำเนินงานสร้างสื่อการสอนบนเครือข่ายอินเทอร์เน็ตเรื่อง การทำเหมืองข้อมูลขั้นพื้นฐาน นี้แบ่งการดำเนินการออกเป็นขั้นตอนต่างๆ ดังนี้

3.1 ศึกษาและรวบรวมข้อมูลเนื้อหาเรื่อง การทำเหมืองข้อมูลขั้นพื้นฐาน

สื่อการสอนจะรวบรวมเนื้อหาเกี่ยวกับเรื่อง การทำเหมืองข้อมูลขั้นพื้นฐาน โดยจะจัดแบ่งเนื้อหาที่แสดงดังต่อไปนี้

1 ความรู้เบื้องต้นเกี่ยวกับการทำเหมืองข้อมูล (Data Mining)

2 ประเภทของวิธีการเรียนรู้และเทคนิคต่างๆใน Data Mining

2.1 Supervised Learning / Predictive

สำหรับ Supervised Learning จะกล่าวถึงเทคนิคต่างๆ ของ การจัดแบ่งหมวดหมู่ (Classification) ซึ่งแบ่งเป็น

2.1.1 การแบ่งหมวดหมู่เป็น 2 class (Classification into 2 classes)

2.1.2 การแบ่งหมวดหมู่มากกว่า 2 class (Classification into more than 2 classes)

2.2 Unsupervised Learning / Descriptive

สำหรับ Unsupervised Learning จะกล่าวถึงเทคนิคต่างๆ ดังนี้

การจัด Cluster (Clustering)

2.2.1 Summarization

2.2.2 กฎของความสัมพันธ์ (Association Rules)

3.2 ศึกษาโปรแกรมคอมพิวเตอร์ที่ใช้ในงานวิจัย

ศึกษาโปรแกรมคอมพิวเตอร์ที่ใช้ในงานปัญหาพิเศษ โดยจะประกอบด้วยโปรแกรม Macromedia Dreamweaver MX 2004, Adobe Photoshop CS และAdobe Illustrator CS

3.3 การออกแบบและพัฒนาสื่อการสอน

ในการออกแบบและพัฒนาสื่อการสอนจะแบ่งเป็นขั้นตอนได้ ดังนี้

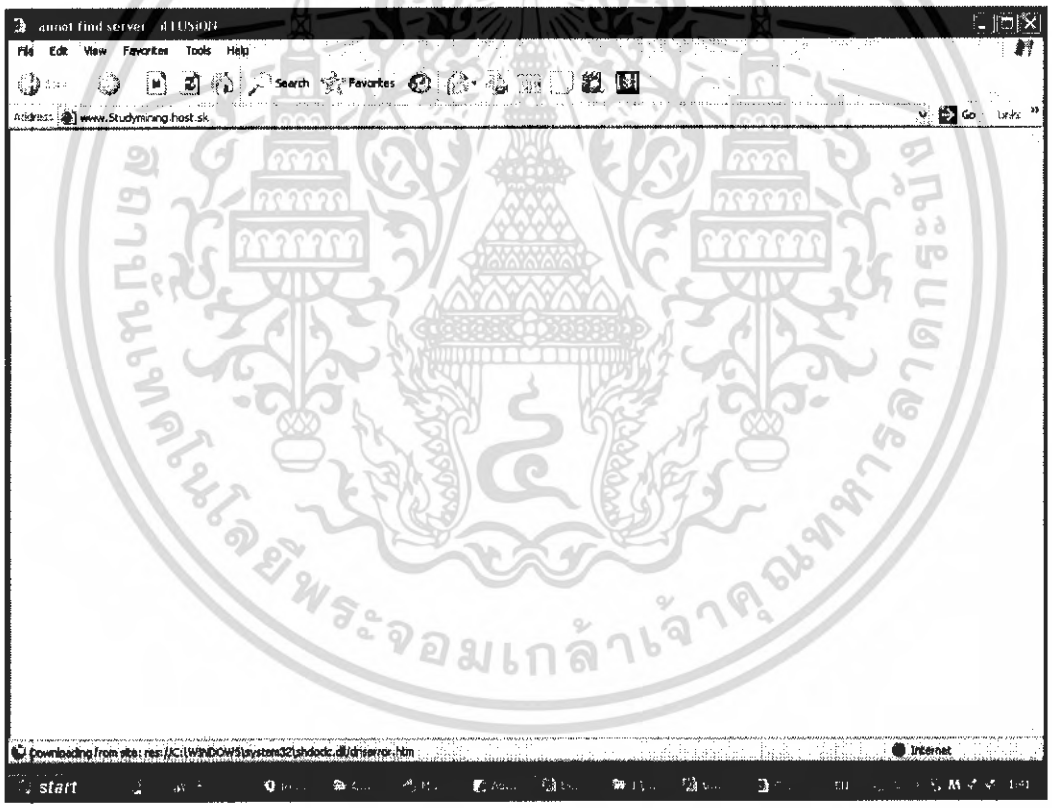
1. กำหนดจุดประสงค์และขอบเขตเพื่อให้สามารถสร้างสื่อการสอนได้อย่างมีประสิทธิภาพ
2. การออกแบบสื่อการสอน ให้มีความสวยงาม น่าสนใจและผู้ใช้สามารถเข้าศึกษาในหัวข้อต่างๆ ได้ง่าย

บทที่ 4
ผลการศึกษา

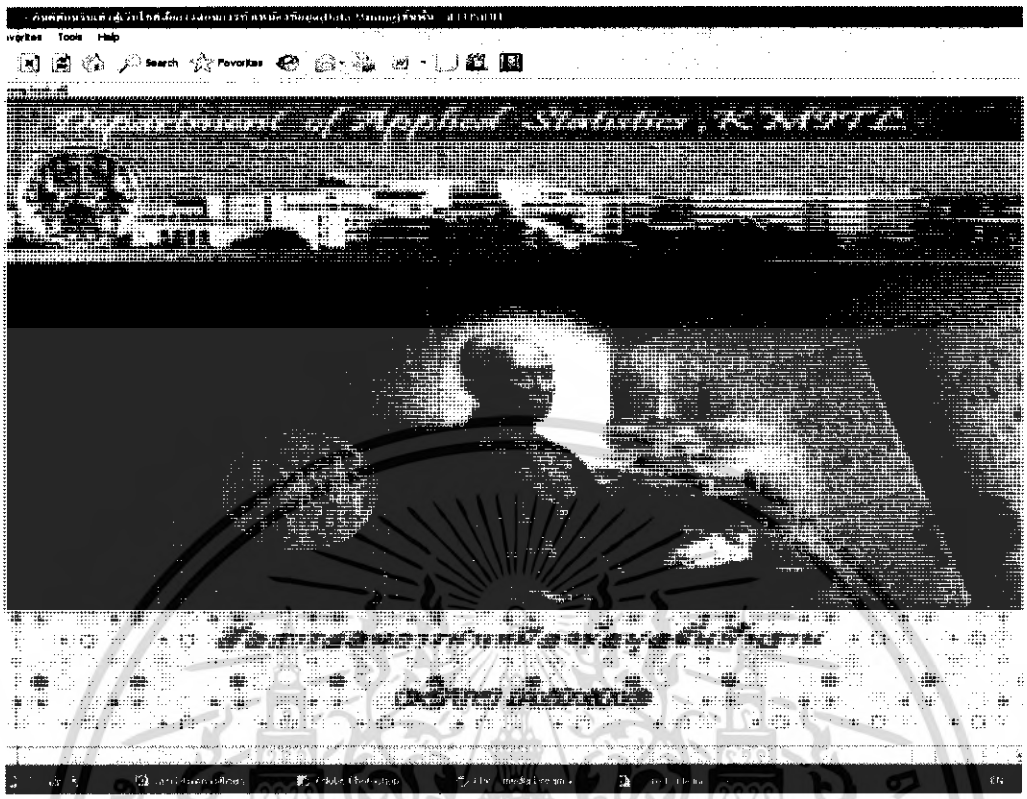
ส่วนประกอบต่างๆที่สร้างขึ้นในสื่อการสอนบนเครือข่ายอินเทอร์เน็ตเรื่อง การทำเหมืองข้อมูลขั้นพื้นฐาน มีดังนี้

4.1 การเข้าสู่สื่อการสอน

ในการเข้าสู่สื่อการสอนเรื่อง การทำเหมืองข้อมูลขั้นพื้นฐานผ่านระบบอินเทอร์เน็ตนั้นผู้ศึกษาจะต้องทำการพิมพ์ URL ของเว็บไซต์ คือ http://www.studymining.host.sk ดังรูปที่ 4-1 โดยหน้าต่างที่จะปรากฏขึ้นเป็นหน้าแรกของสื่อการสอนมีหน้าต่างดังรูปที่ 4-2



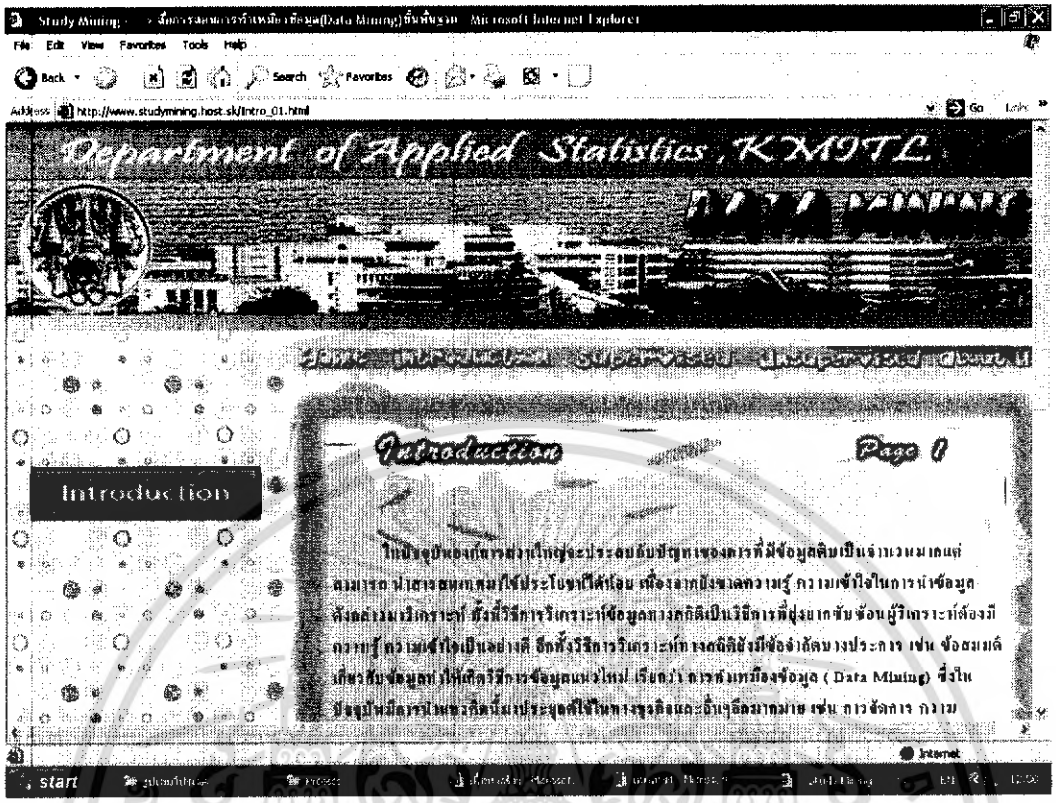
รูปที่ 4-1 หน้าจอ URL ของเว็บไซต์



รูปที่ 4-2 หน้าต่างโฮมเพจเมื่อระบบเข้าสู่สื่อการสอน

จากรูปที่ 4-2 เมื่อผู้ศึกษาสามารถเข้าสู่ส่วนต่างๆของเนื้อหาโดยคลิกที่คำว่า Data Mining ดังรูปที่ 4-3 จะปรากฏหน้าต่างในส่วนของเนื้อหาโดยหน้าต่างแรกที่ปรากฏขึ้นบนหน้าจอคือ หน้า Introduction ดังรูปที่ 4-4

รูปที่ 4-3 ปุ่มตัวอักษร Data Mining ในหน้าต่างแรกที่ปรากฏขึ้นเมื่อเข้าสู่โฮมเพจ



รูปที่ 4-4 หน้าต่างของหน้าเว็บเพจด้านบน

โดยหน้าต่างทุกหน้าต่างจะประกอบไปด้วยปุ่มหลักที่เชื่อมโยงไปยังเนื้อหาในหัวข้ออื่นที่เกี่ยวข้องซึ่งจะปรากฏเป็นแถบอยู่ด้านบนของเนื้อหาดังรูปที่ 4-5

รูปที่ 4-5 แถบเมนูหลักด้านบน

โดยรายละเอียดการเชื่อมโยงของปุ่มที่อยู่บนแถบเมนูหลักด้านบนมีดังนี้

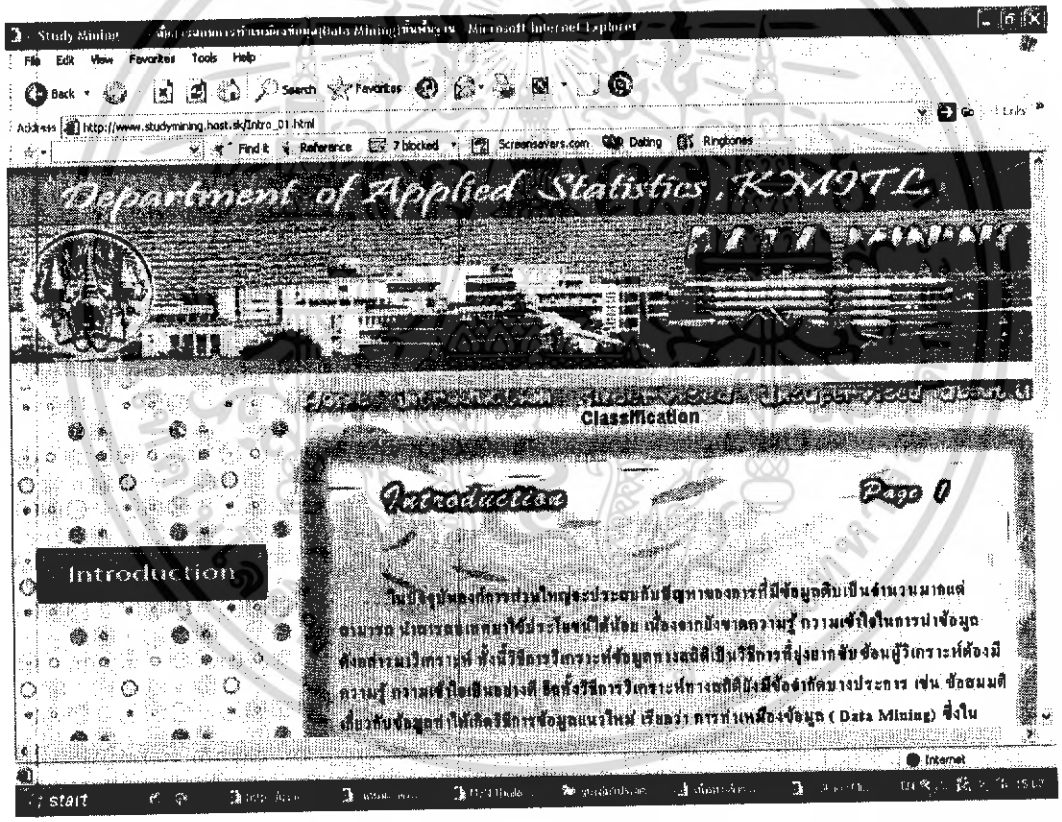
ปุ่ม Home เป็นปุ่มที่ทำหน้าที่เชื่อมหน้าต่างอื่นๆ ในส่วนเนื้อหา กับ หน้าแรกของเว็บไซต์

Introduction

ปุ่ม introduction เป็นปุ่มที่ทำหน้าที่เชื่อมหน้าต่างอื่นๆ ในส่วนเนื้อหากับหน้า Introduction ซึ่งเป็นหน้าที่แนะนำเกี่ยวกับความรู้เบื้องต้นของวิธีการทำเหมืองข้อมูลขั้นพื้นฐาน

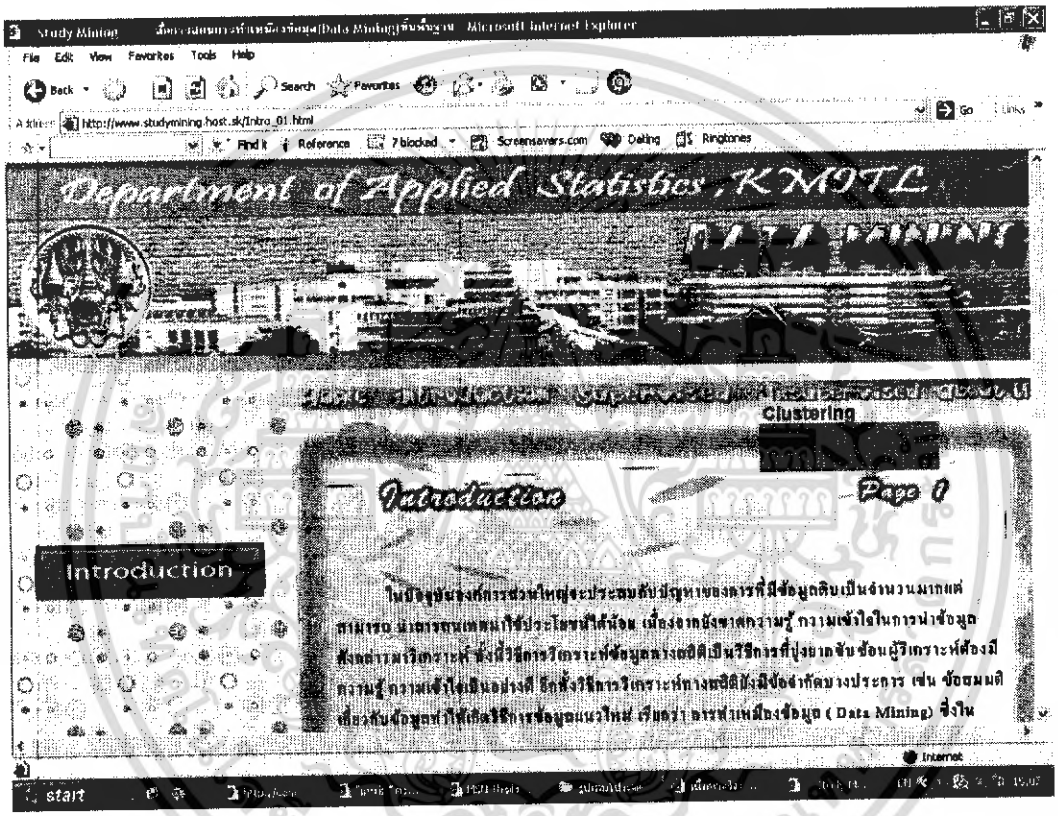
Supervised

ปุ่ม Supervised เป็นปุ่มที่ทำหน้าที่เชื่อมหน้าต่างอื่นๆ ในส่วนเนื้อหากับวิธีการต่างๆ ที่ใช้ในการทำเหมืองข้อมูลแบบ Supervised Learning ที่ถูกนำมาจัดอยู่ในรูปแบบ Popup Menu ซึ่งประกอบด้วยวิธีต่างๆ ให้ผู้ศึกษาเลือก คือ Classification ดังรูปที่ 4-6



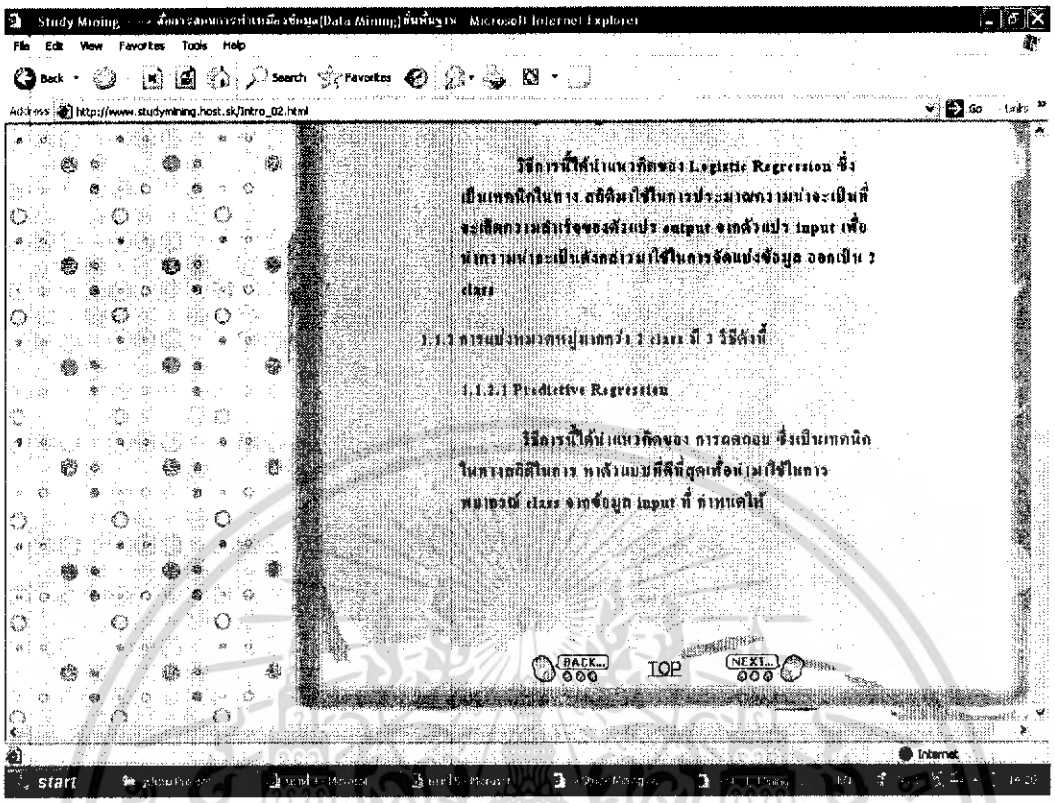
รูปที่ 4-6 Popup Menu ของ Supervised Learning

ปุ่ม Unsupervised เป็นปุ่มที่ทำหน้าที่เชื่อมหน้าต่างอื่นๆในส่วนเนื้อหากับวิธีการต่างๆ ที่ใช้ในการทำเหมืองข้อมูลแบบ Unsupervised Learning ที่ถูกนำมาจัดอยู่ในรูปแบบ Popup Menu ซึ่งประกอบด้วยวิธีต่างๆ ให้ผู้ศึกษาเลือก คือ Clustering , Summarization และ Association Rules ดังรูปที่ 4-7



รูปที่ 4-7 Popup Menu ของ Unsupervised Learning

ปุ่ม About Us เป็นปุ่มที่ทำหน้าที่เชื่อมหน้าต่างส่วนเนื้อหากับหน้าต่าง About Us ซึ่งแสดงข้อมูลของผู้จัดทำสื่อการสอน



รูปที่ 4-8 ปุ่มเชื่อมระหว่างหน้าเว็บเพจแต่ละหน้าในวิธีเดียวกัน

จากรูปที่ 4-8 เราสามารถอธิบายปุ่มที่ทำการเชื่อมหน้าเว็บเพจแต่ละหน้าในวิธีการหรือเทคนิคเดียวกันได้ดังนี้

ปุ่ม Back เป็นปุ่มที่เชื่อมหน้าปัจจุบันกับหน้าที่แล้วของเว็บเพจ

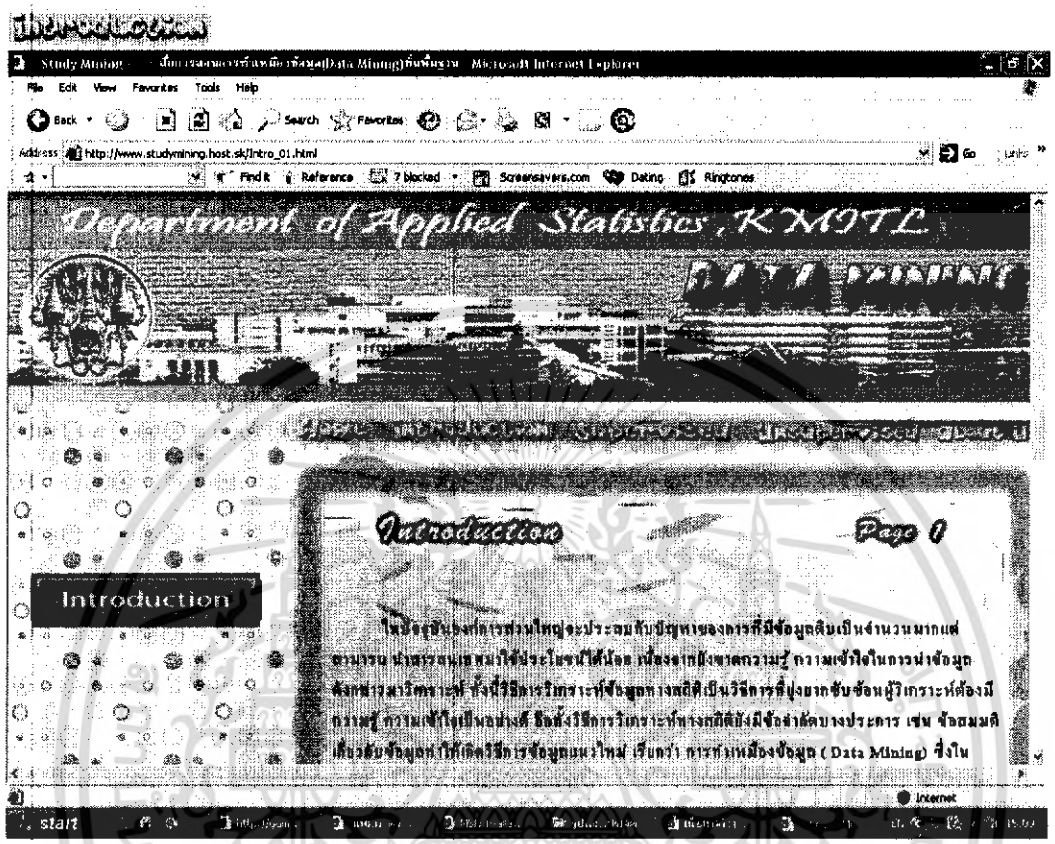
TOP ปุ่ม TOP เป็นปุ่มเชื่อมไปยังด้านบนสุดของเว็บเพจหน้าเดียวกัน

ปุ่ม Next เป็นปุ่มที่เชื่อมหน้าปัจจุบันกับหน้าถัดไปของเว็บเพจ

4.2 การเข้าสู่หน้าบทนำ (Introduction)

ผู้ศึกษาจะเข้าสู่หน้าบทนำได้เมื่อได้คลิกเลือกปุ่ม ที่แถบเมนูด้านบน และเมื่อได้เข้าสู่เว็บเพจแล้วภายในบทนำ (Introduction) จะกล่าวถึง ความหมายของการทำเหมืองข้อมูล ความหมายของประเภทของการเรียนรู้แบบ Supervised Learning และ แบบ Unsupervised Learning รวมทั้งรายละเอียดเบื้องต้นเกี่ยวกับเทคนิคต่างๆ

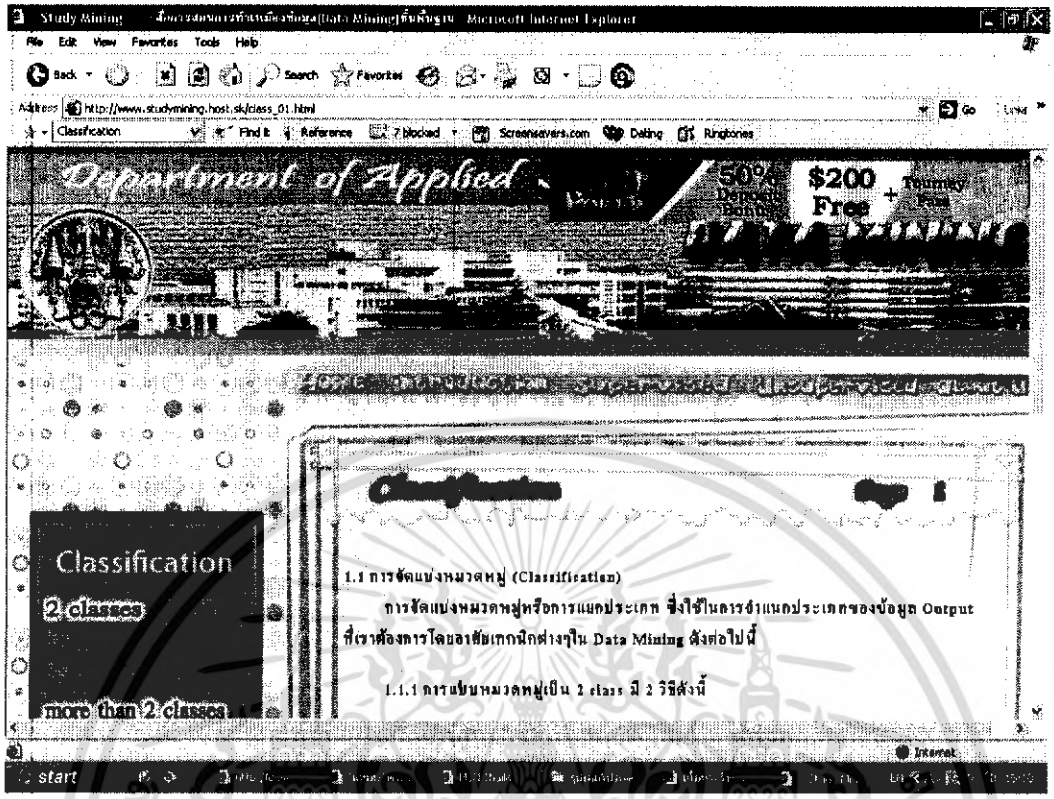
ดังที่กล่าวในหัวข้อ 4.1 ถ้าผู้ใช้คลิกที่ปุ่ม จะปรากฏหน้า Introduction อย่างไรก็ตามถ้าผู้ใช้ไปอยู่ที่หน้าต่างอื่นสามารถกลับมาที่หน้า Introduction โดยคลิกที่



รูปที่ 4-9 หน้าต่าง Introduction

4.3 การเข้าสู่หน้าการจัดแบ่งหมวดหมู่ (Classification)

ผู้ศึกษาจะเข้าสู่หน้าบทนำการจัดแบ่งหมวดหมู่ได้ก็ต่อเมื่อเลื่อนเมาส์ไปที่ปุ่ม แล้วทำการเลือกหัวข้อ Classification บน Popup Menu เมื่อได้เข้าสู่เว็บเพจแล้วภายในการจัดแบ่งหมวดหมู่จะกล่าวถึง วิธีการจัดแบ่งหมวดหมู่ (Classification) และเทคนิคการจัดแบ่งหมวดหมู่ ซึ่งมีหน้าต่างดังรูปที่ 4-10



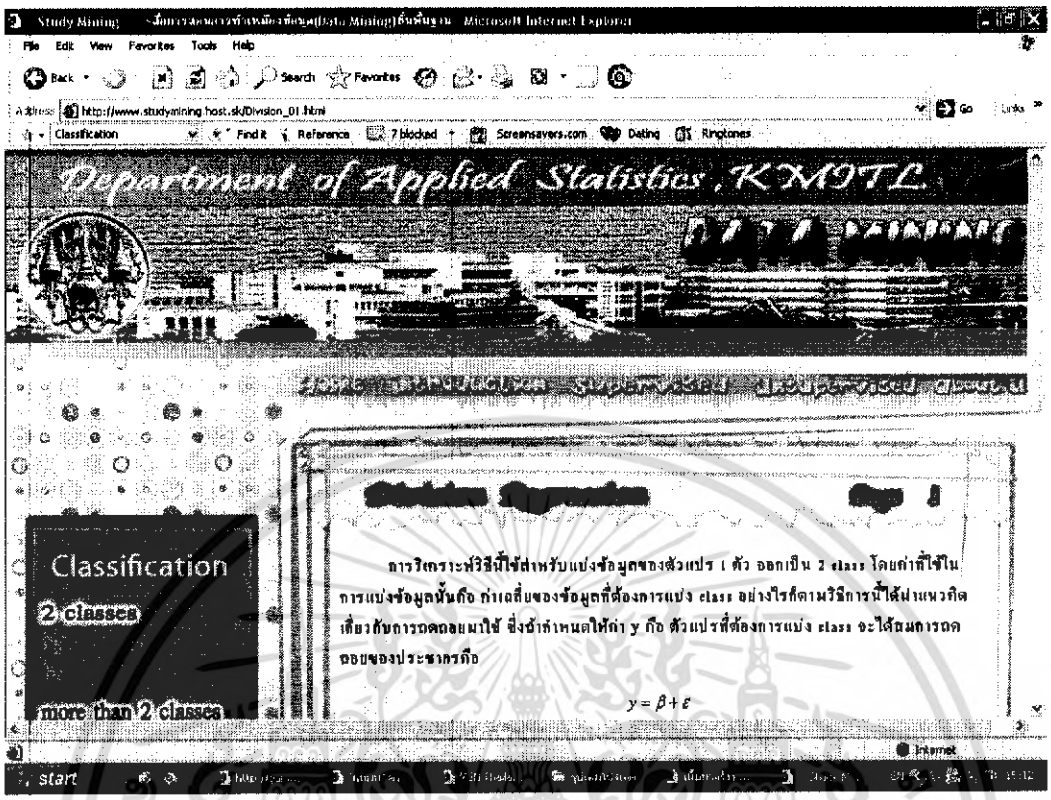
รูปที่ 4-10 หน้าต่าง Classification เมื่อเลือกเมนู Classification จากปุ่ม Supervised

โดยผู้ศึกษาสามารถเลือกที่จะเข้าศึกษาในเทคนิคการจัดแบ่งหมวดหมู่(Classification) ต่างๆได้จากลิงค์ภายในเนื้อหาและแถบเมนูด้านข้าง ซึ่งแบ่งเทคนิคการจัดแบ่งหมวดหมู่ออกเป็น 2 ส่วนคือ ผลที่ได้จากการแบ่งเป็น 2 classes และ more than 2 classes ที่ปรากฏอยู่ในทุกหน้าต่างของเทคนิคการจัดแบ่งหมวดหมู่ ดังรูปที่ 4-11

รูปที่ 4-11 แถบเมนูที่เชื่อมเทคนิคต่างๆใน Classification

4.3.1 การเข้าสู่เทคนิคการจัดแบ่งหมวดหมู่แบบ Division Regression

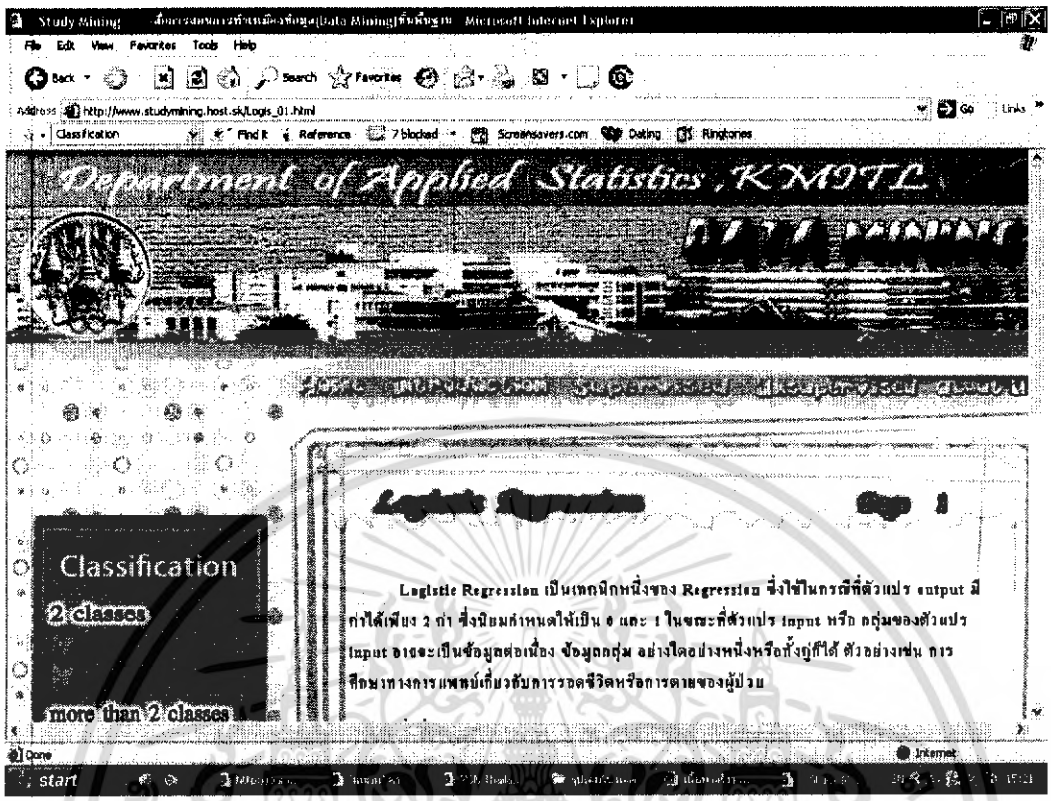
ผู้ศึกษาจะเข้าสู่เทคนิคการจัดหมวดหมู่แบบ Division Regression ได้เมื่อคลิกเลือก [redacted] จากแถบเมนูด้านข้าง และเมื่อได้เข้าสู่เว็บเพจแล้วจะประกอบไปด้วยการอธิบายถึงขั้นตอนการจัดแบ่งหมวดหมู่แบบ Division Regression พร้อมทั้งยกตัวอย่างประกอบ



รูปที่ 4-12 หน้าต่าง Classification โดยวิธี Division Regression

4.3.2 การเข้าสู่เทคนิคการจัดแบ่งหมวดหมู่แบบ Logistic Regression

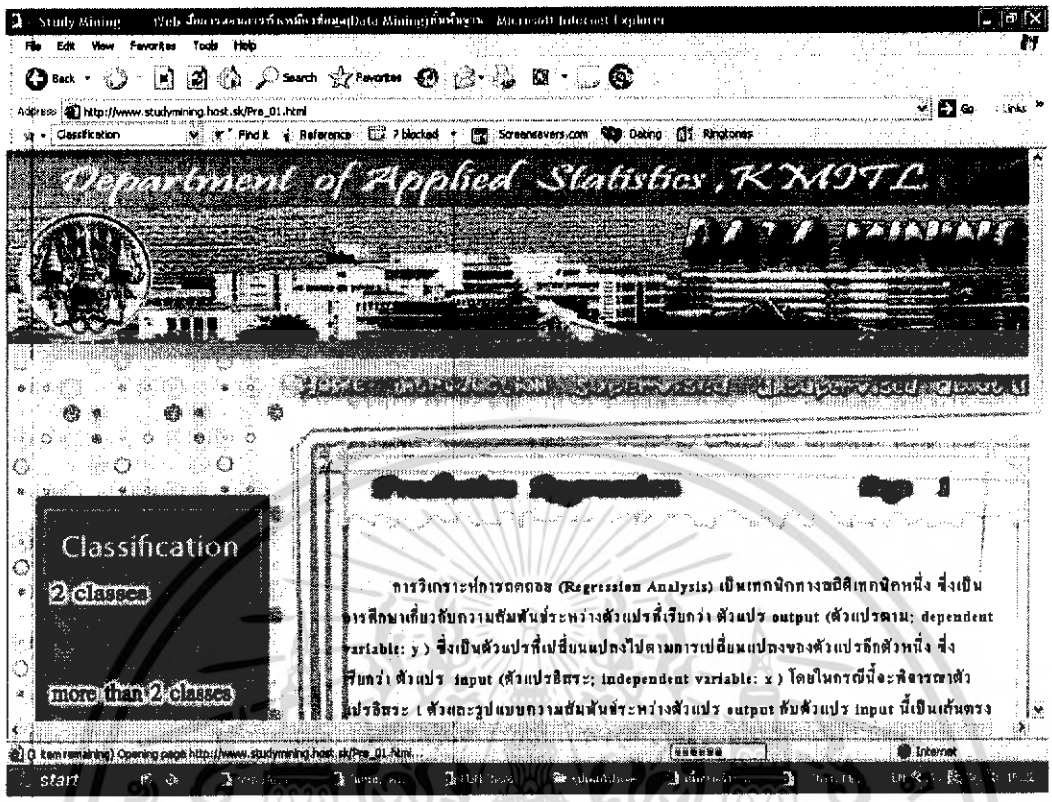
ผู้ศึกษาจะเข้าสู่เทคนิคการจัดหมวดหมู่แบบ Logistic Regression ได้เมื่อคลิกเลือก [redacted] จากแถบเมนูด้านข้าง และเมื่อได้เข้าสู่เว็บเพจแล้วจะประกอบไปด้วยการอธิบายถึงขั้นตอนการจัดแบ่งหมวดหมู่แบบ Logistic Regression พร้อมทั้งยกตัวอย่างประกอบ



รูปที่ 4-13 หน้าต่าง Classification โดยวิธี Logistic Regression

4.3.3 การเข้าสู่เทคนิคการจัดแบ่งหมวดหมู่แบบ Predictive Regression

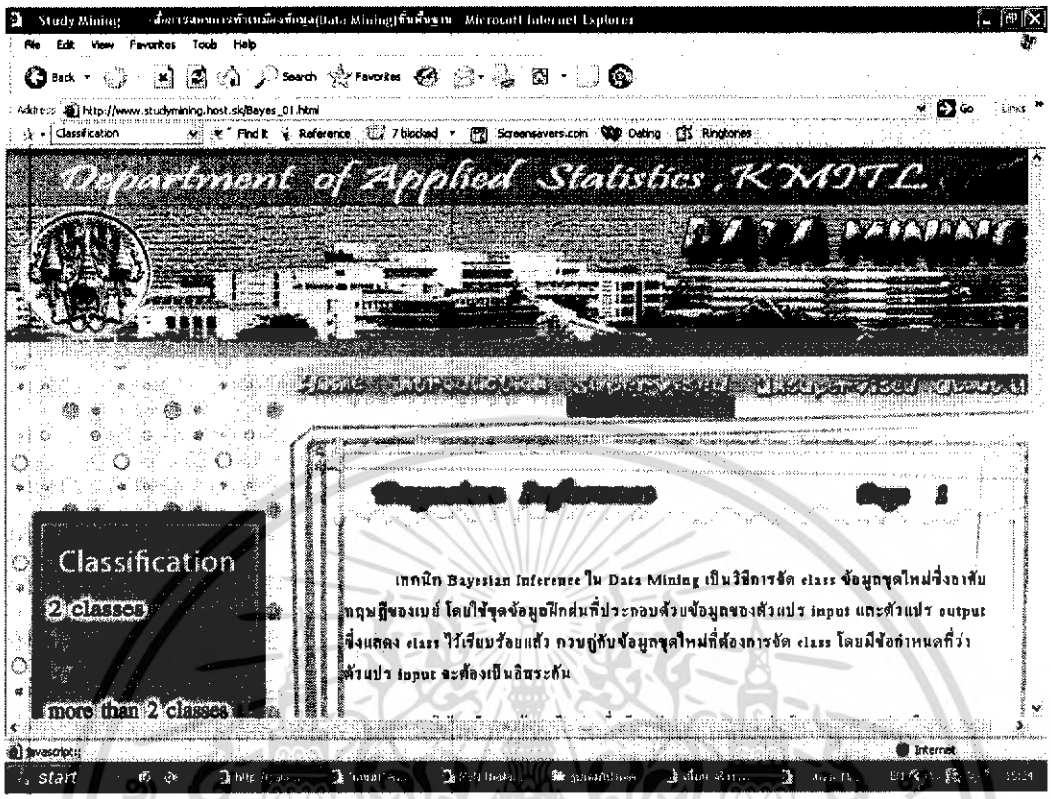
ผู้ศึกษาจะเข้าสู่เทคนิคการจัดหมวดหมู่แบบ Predictive Regression ได้เมื่อคลิกเลือก [redacted] จากแถบเมนูด้านข้าง และเมื่อได้เข้าสู่เว็บเพจแล้วจะประกอบไปด้วยการอธิบายถึงขั้นตอนการจัดแบ่งหมวดหมู่แบบ Predictive Regression พร้อมทั้งยกตัวอย่างประกอบ



รูปที่ 4-14 หน้าต่าง Classification โดยวิธี Predictive Regression

4.3.4 การเข้าสู่เทคนิคการจัดแบ่งหมวดหมู่แบบ Bayesian Inference

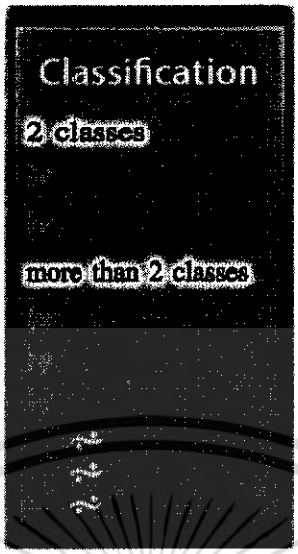
ผู้ศึกษาจะเข้าสู่เทคนิคการจัดหมวดหมู่แบบ Bayesian Inference ได้เมื่อคลิกเลือก [redacted] จากแถบเมนูด้านข้าง และเมื่อได้เข้าสู่เว็บเพจแล้วจะประกอบไปด้วยการอธิบายถึงขั้นตอนการจัดแบ่งหมวดหมู่แบบ Bayesian Inference พร้อมทั้งยกตัวอย่างประกอบ



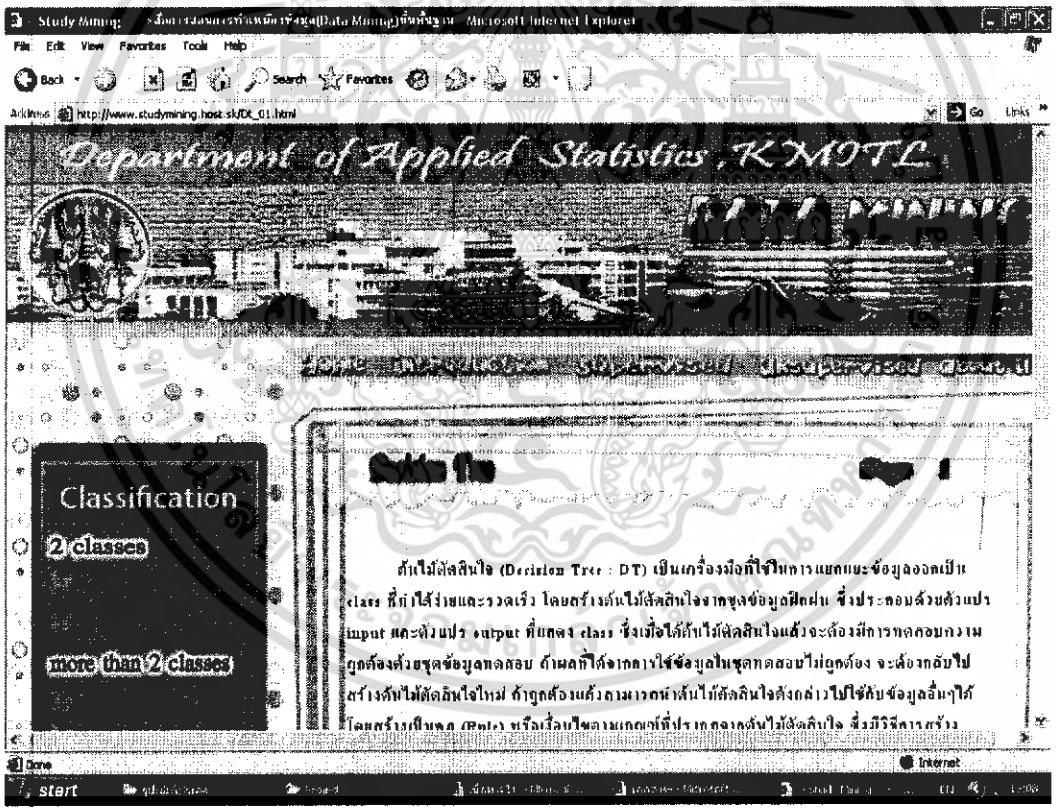
รูปที่ 4-15 หน้าต่าง Classification โดยวิธี Bayesian Inference

4.3.5 การเข้าสู่เทคนิคการจัดแบ่งหมวดหมู่แบบต้นไม้การตัดสินใจ (Decision Tree)

ผู้ศึกษาจะเข้าสู่บทนำการจัดหมวดหมู่แบบต้นไม้การตัดสินใจ (Decision Tree) เมื่อคลิกเลือก [redacted] จากแถบเมนูด้านข้างและเมื่อได้เข้าสู่เว็บเพจแล้ว ภายในหน้าบทนำการจัดแบ่งหมวดหมู่แบบต้นไม้การตัดสินใจ (Decision Tree) จะกล่าวแนะนำเกี่ยวกับการจัดแบ่งหมวดหมู่แบบต้นไม้ตัดสินใจและวิธีการที่ใช้ในการจัดหมวดหมู่แบบดังกล่าว นอกจากนี้แถบเมนูด้านข้างจะเปลี่ยนไปโดยจะมีเมนูย่อยเพิ่มขึ้นมา เพื่อเชื่อมไปยังวิธีต่างที่ใช้ในต้นไม้การตัดสินใจ (Decision Tree) ดังรูปที่ 4-16



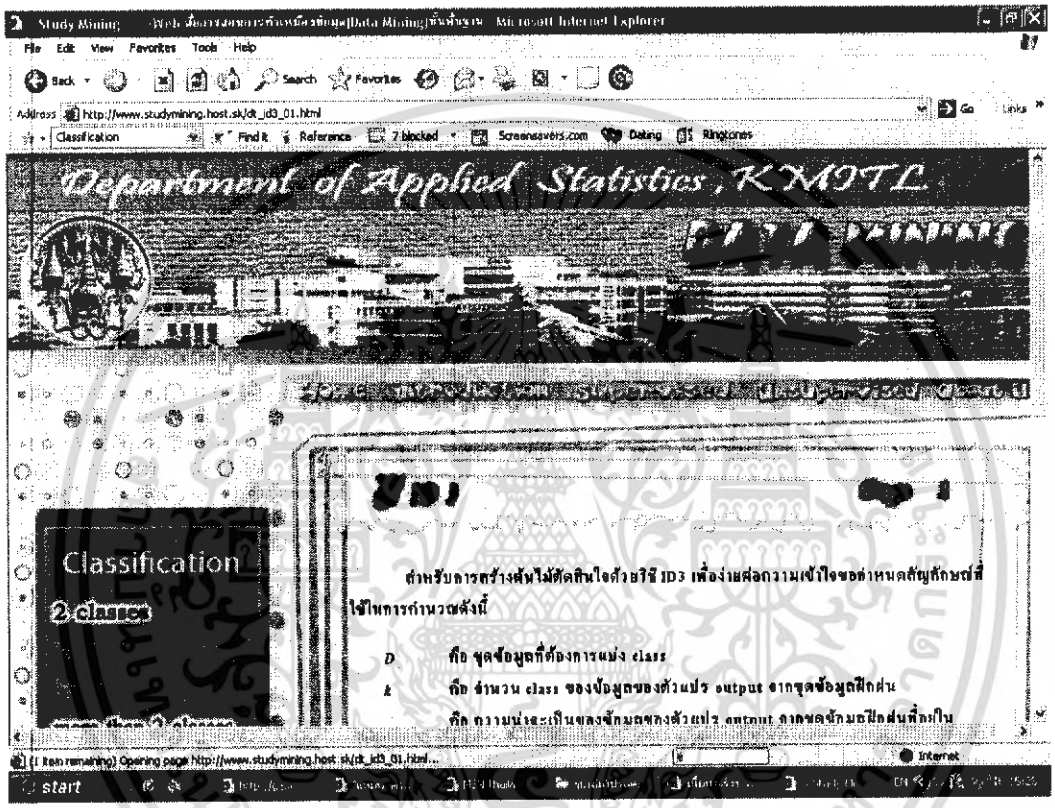
รูปที่ 4-16 แถบข้างที่แสดงในหน้าต่าง Classification โดยวิธี Decision Tree



รูปที่ 4-17 หน้าต่างบทนำ Classification โดยวิธี Decision Tree

4.3.5.I การเข้าสู่การจัดแบ่งหมวดหมู่ต้นไม้ตัดสินใจด้วยวิธี ID 3

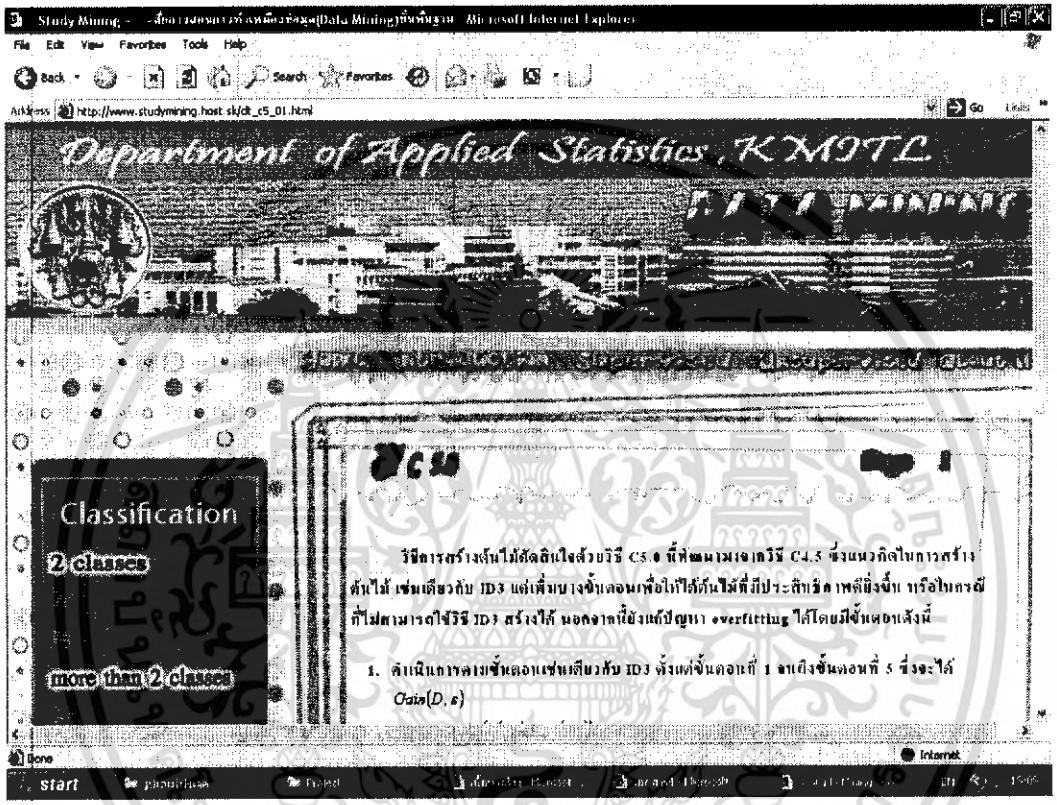
ผู้ศึกษาจะเข้าสู่กระบวนการจัดหมวดหมู่ต้นไม้ตัดสินใจด้วยวิธี ID 3 ได้เมื่อคลิกเลือก [redacted] จากแถบเมนูด้านข้าง เมื่อเข้าสู่เว็บเพจแล้วภายในหน้าการจัดหมวดหมู่ต้นไม้ตัดสินใจด้วยวิธี ID 3 จะกล่าวถึงขั้นตอนการจัดหมวดหมู่แบบต้นไม้ตัดสินใจด้วยวิธี ID 3 พร้อมทั้งยกตัวอย่างประกอบ



รูปที่ 4-18 หน้าต่างการจัดหมวดหมู่ต้นไม้ตัดสินใจด้วยวิธี ID 3

4.3.5.2 การเข้าสู่การจัดแบ่งหมวดหมู่ต้นไม้ตัดสินใจด้วยวิธี C5.0

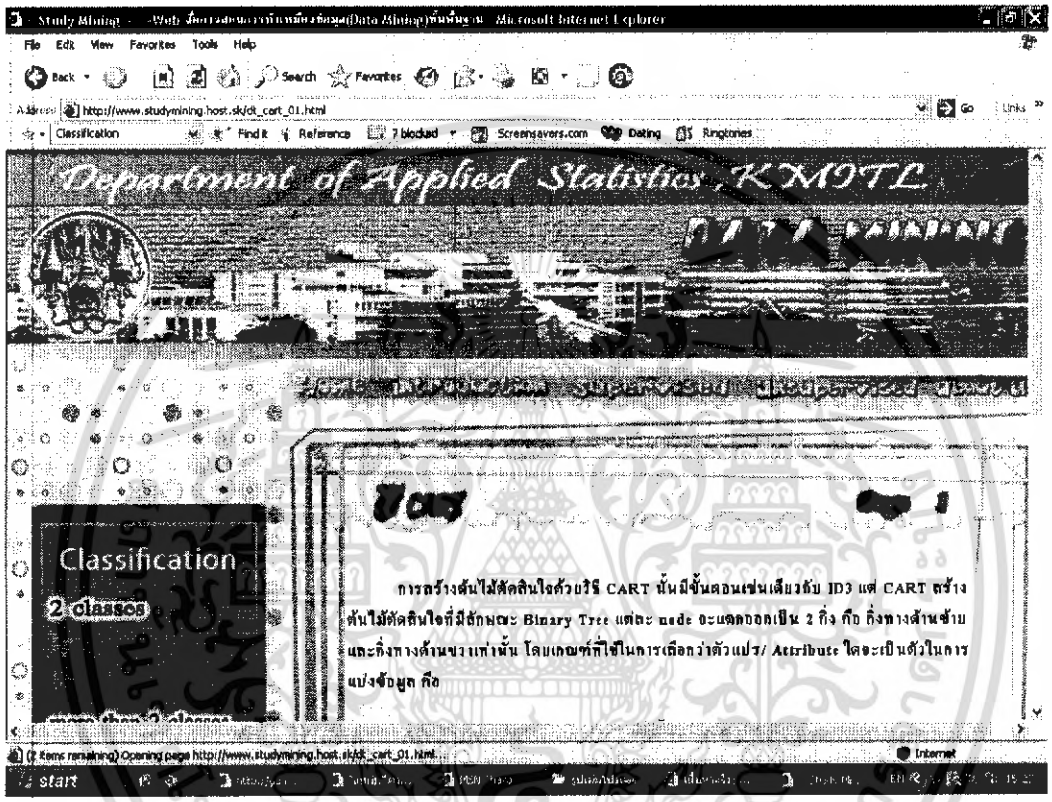
ผู้ศึกษาจะเข้าสู่กระบวนการจัดหมวดหมู่ต้นไม้ตัดสินใจด้วยวิธี C5.0 ได้เมื่อคลิกเลือก [redacted] จากแถบเมนูด้านข้าง เมื่อเข้าสู่เว็บเพจแล้วภายในหน้าการจัดหมวดหมู่ต้นไม้ตัดสินใจด้วยวิธี C5.0 จะกล่าวถึงขั้นตอนการจัดหมวดหมู่แบบต้นไม้ตัดสินใจด้วยวิธี C5.0 พร้อมทั้งยกตัวอย่างประกอบ



รูปที่ 4-19 หน้าต่างการจัดหมวดหมู่ต้นไม้ตัดสินใจด้วยวิธี C5.0

4.3.5.3 การเข้าสู่การจัดแบ่งหมวดหมู่ต้นไม้ตัดสินใจด้วยวิธี CART

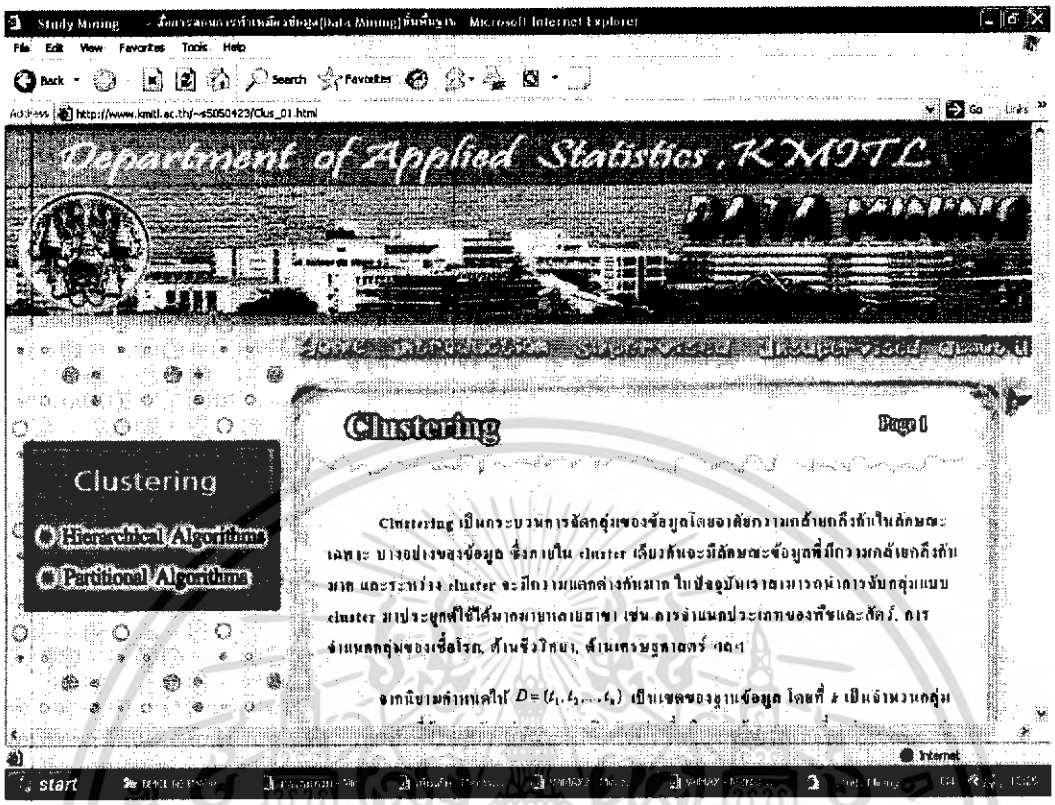
ผู้ศึกษาจะเข้าสู่กระบวนการจัดหมวดหมู่ต้นไม้ตัดสินใจด้วยวิธี CARTได้เมื่อคลิกเลือก [ ] จากแถบเมนูด้านข้าง เมื่อเข้าสู่เว็บเพจแล้วภายในหน้าการจัดหมวดหมู่ต้นไม้ตัดสินใจด้วยวิธี CART จะกล่าวถึงขั้นตอนการจัดหมวดหมู่แบบต้นไม้ตัดสินใจด้วยวิธี CART พร้อมทั้งยกตัวอย่างประกอบ



รูปที่ 4-20 หน้าต่างการจัดหมวดหมู่ต้นไม้ตัดสินใจด้วยวิธี CART

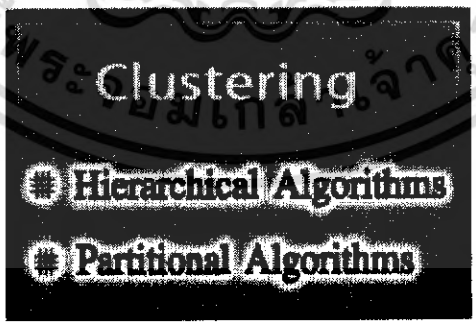
4.4 การเข้าสู่หน้าการจัด Cluster (Clustering)

ผู้ศึกษาจะเข้าสู่หน้า Clustering ได้ก็ต่อเมื่อเลื่อนเมาส์ไปที่ปุ่ม [ ] แล้วทำการเลือกหัวข้อ Clustering บน Popup Menu เมื่อได้เข้าสู่เว็บเพจแล้ว ภายในหน้า Clustering จะกล่าวแนะนำเกี่ยวกับวิธีการจัด cluster ต่างๆ ซึ่งมีหน้าต่างดังรูปที่ 4-21



รูปที่ 4-21 หน้าต่าง Clustering เมื่อเลือกเมนู Clustering บนปุ่ม

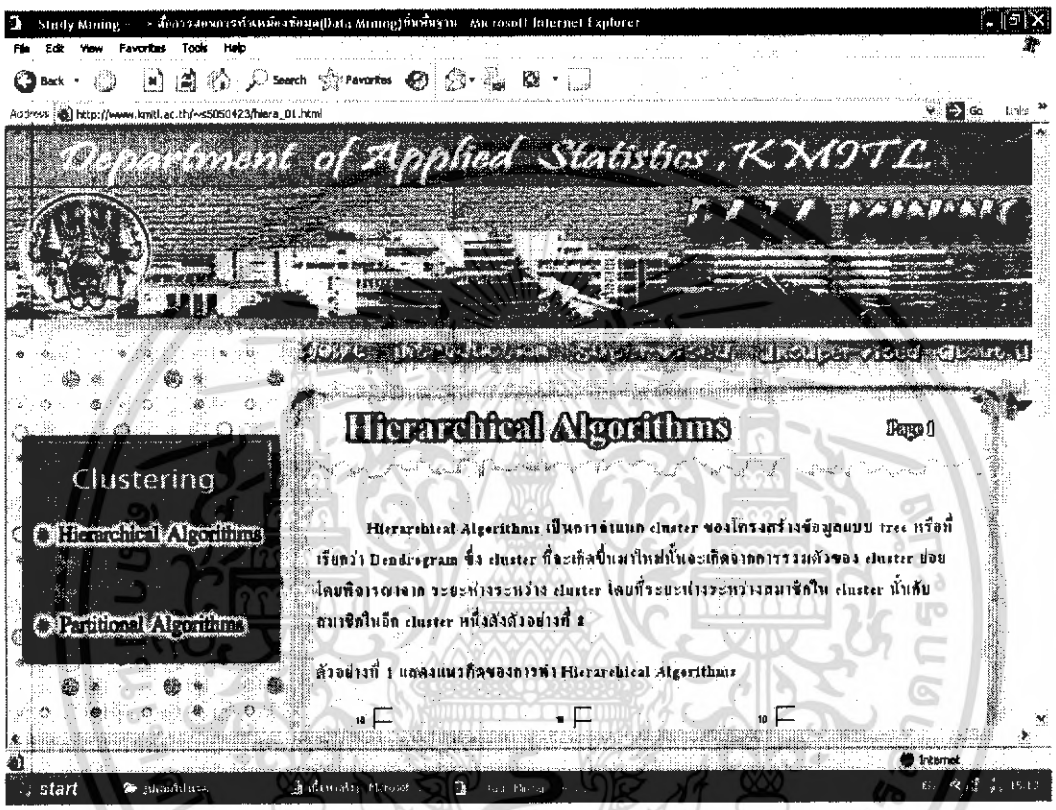
โดยผู้ศึกษาสามารถเลือกที่จะเข้าศึกษาในเทคนิคการจัด cluster (Clustering) ต่างๆได้จากลิงค์ภายในเนื้อหาและแถบเมนูด้านข้าง ซึ่งแบ่งวิธีการจัดแบ่ง cluster ออกเป็น 2 วิธีหลักๆ คือ Hierarchical Algorithms และ Partitional Algorithms ที่ปรากฏอยู่ในทุกหน้าต่างของเทคนิคการจัด cluster ดังรูปที่ 4-22



รูปที่ 4-22 แถบข้างที่แสดงในหน้าต่างการจัด cluster (Clustering)

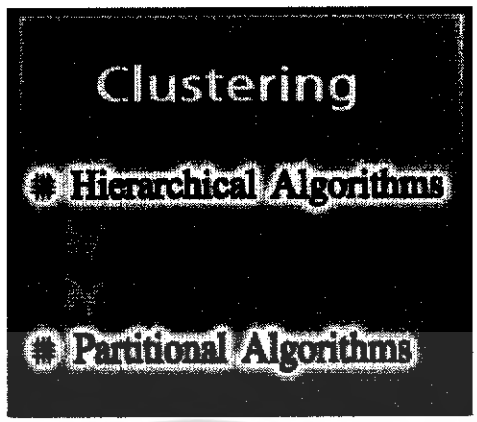
4.4.1 การเข้าสู่เทคนิคการจัด cluster แบบ Hierarchical Algorithms

ผู้ศึกษาจะเข้าสู่หน้าบทนำการจัด cluster แบบ Hierarchical Algorithms ได้ก็ต่อเมื่อคลิกที่ปุ่ม Hierarchical Algorithms เมื่อได้เข้าสู่เว็บเพจแล้วภายในการจัด cluster จะกล่าวถึง วิธีการจัด cluster แบบ Hierarchical Algorithms ซึ่งมีหน้าต่างดังรูปที่ 4-23



รูปที่ 4-23 หน้าต่าง Clustering เมื่อเลือกเมนู Hierarchical Algorithms

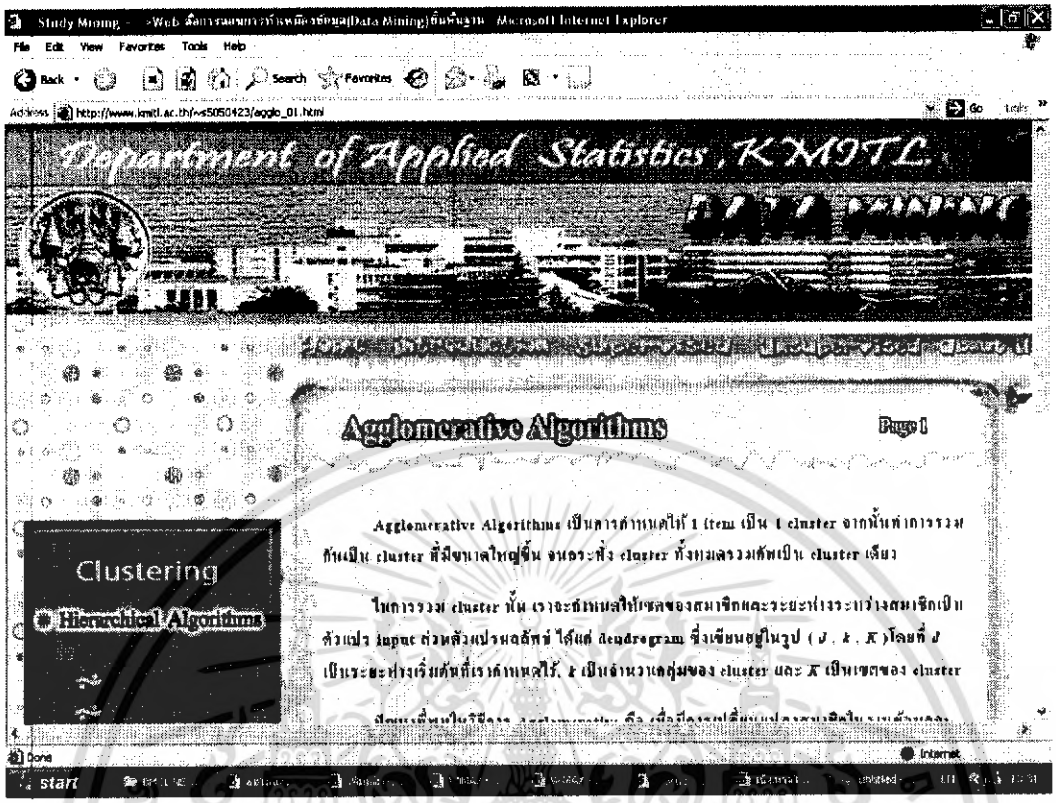
โดยผู้ศึกษาสามารถเลือกที่จะเข้าศึกษาในเทคนิคการจัด cluster แบบ Hierarchical Algorithms ได้จากลิงค์ภายในเนื้อหาและแถบเมนูด้านข้าง ซึ่งแบ่งเทคนิคการจัด cluster แบบ Hierarchical Algorithms เป็น 2 เทคนิค คือ Agglomerative Algorithms และ Divisive Clustering ที่ปรากฏอยู่ในทุกหน้าต่างของเทคนิคการจัด cluster ดังรูปที่ 4-24



รูปที่ 4-24 แถบข้างที่แสดงในหน้าต่างการจัด cluster แบบ Hierarchical Algorithms

4.4.1.1 การเข้าสู่เทคนิคการจัด cluster แบบ Hierarchical Algorithms ด้วยวิธีการ Agglomerative Algorithms

ผู้ศึกษาสามารถเข้าสู่เทคนิคการจัด cluster แบบ Hierarchical Algorithms ด้วยวิธีการ Agglomerative Algorithms โดยการคลิกเลือก [redacted] จากแถบเมนูด้านข้าง เมื่อเข้าสู่เว็บเพจแล้ว ภายในหน้า Agglomerative Algorithms จะประกอบไปด้วยการแนะนำถึงรายละเอียดและเทคนิคที่นิยมใช้ใน Agglomerative Algorithms ดังรูปที่ 4-25 นอกจากนี้แถบเมนูด้านข้างจะเปลี่ยนไปโดยจะมีเมนูย่อยเพิ่มขึ้นมาเพื่อเชื่อมไปยังเทคนิคต่างที่ใช้ในการจัด cluster แบบ Hierarchical Algorithms ด้วยวิธีการ Agglomerative Algorithms คือ Single Link Technique, Complete Link Technique และ Average Link Technique ดังรูปที่ 4-26

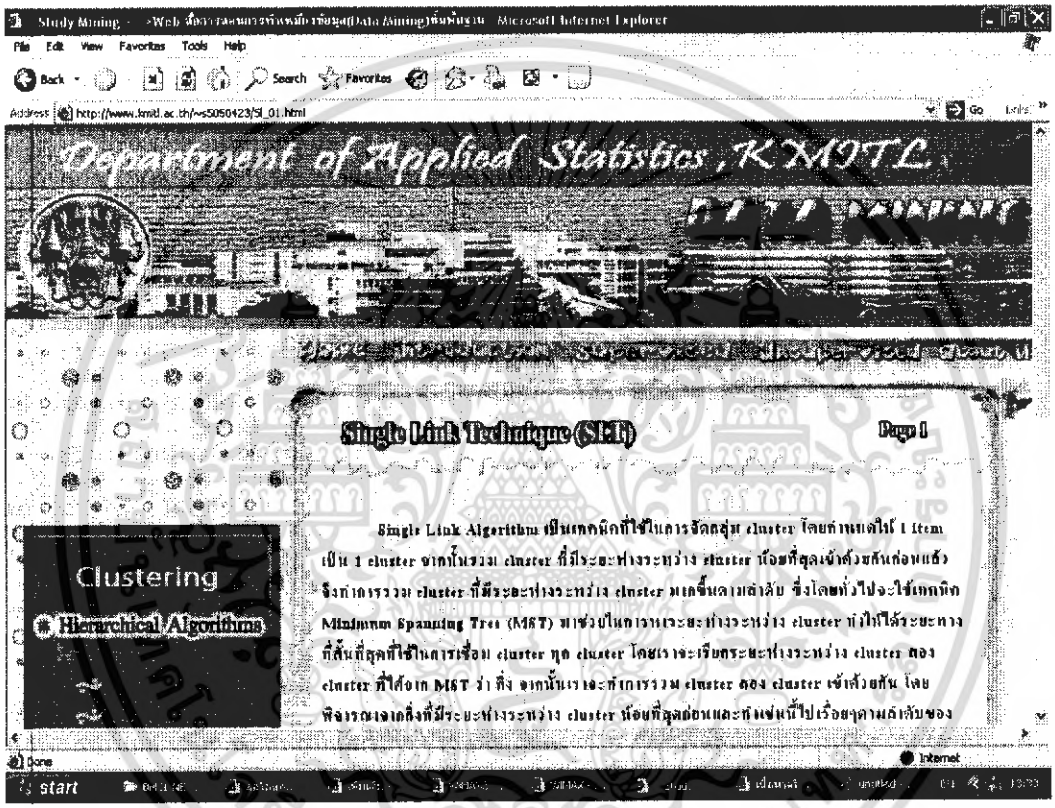


รูปที่ 4-25 หน้าต่างบทนำการจัด cluster แบบ Hierarchical Algorithms ด้วยวิธีการ Agglomerative Algorithms

รูปที่ 4-26 แถบข้างที่แสดงในหน้าต่างการจัด cluster แบบ Hierarchical Algorithms ด้วยวิธีการ Agglomerative Algorithms

4.4.1.1.1 การเข้าสู่ Single link technique (SLT)

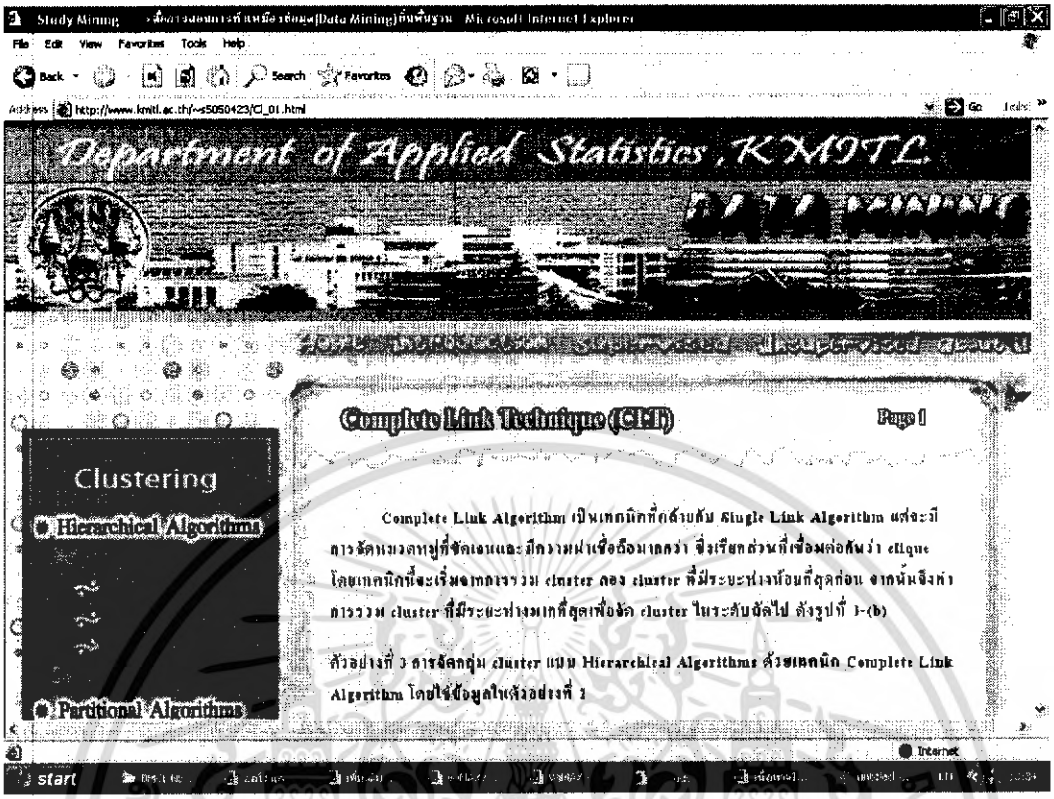
ผู้ศึกษาจะเข้าสู่ Single Link Technique (SLT) ได้ก็ต่อเมื่อคลิกเลือก [redacted] จากแถบเมนูด้านข้าง เมื่อเข้าสู่เว็บเพจแล้วภายในหน้าการจัด cluster แบบ Hierarchical Algorithms ด้วยวิธี Agglomerative Algorithms โดยอาศัย Single Link Technique (SLT) จะกล่าวถึง ขั้นตอนการจัด cluster โดย Single Link Technique (SLT) พร้อมทั้งยกตัวอย่างประกอบ



รูปที่ 4-27 หน้าต่างการจัด cluster โดยอาศัย Single Link Technique (SLT)

4.4.1.1.2 การเข้าสู่ Complete link technique (CLT)

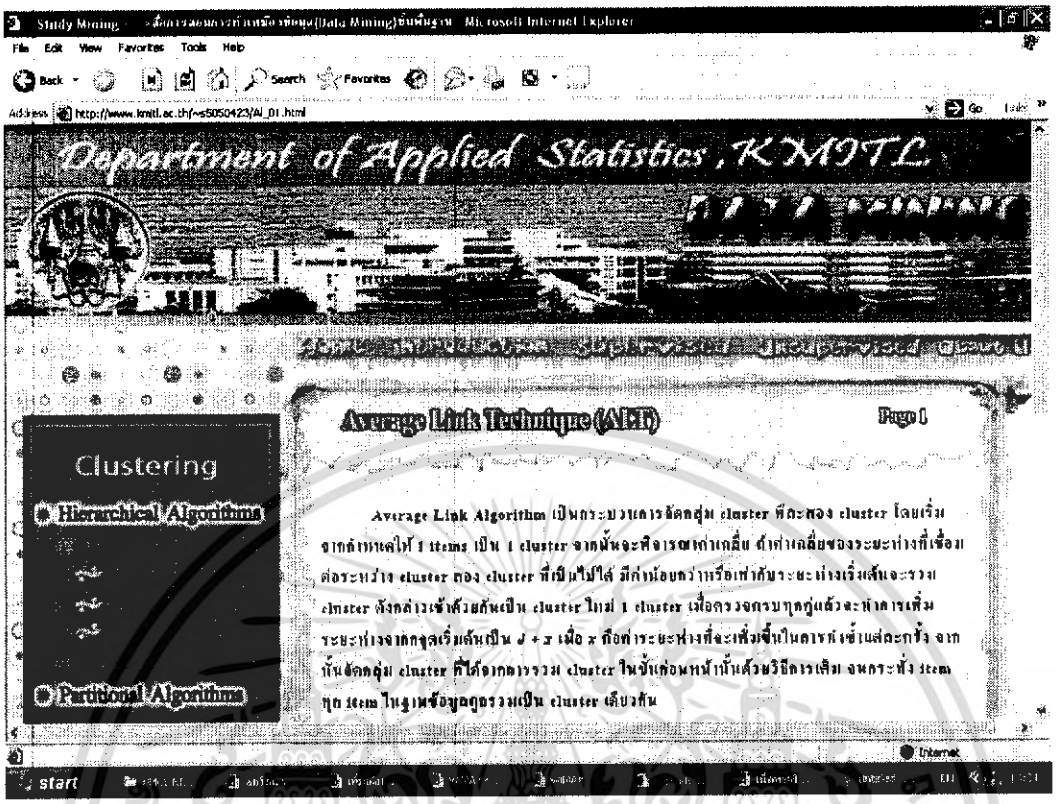
ผู้ศึกษาจะเข้าสู่ Complete Link Technique (CLT) ได้ก็ต่อเมื่อคลิกเลือก [redacted] จากแถบเมนูด้านข้าง เมื่อเข้าสู่เว็บเพจแล้วภายในหน้าการจัด cluster แบบ Hierarchical Algorithms ด้วยวิธี Agglomerative Algorithms โดยอาศัย Complete Link Technique (CLT) จะกล่าวถึง ขั้นตอนการจัด cluster โดย Complete Link Technique (CLT) พร้อมทั้งยกตัวอย่างประกอบ



รูปที่ 4-28 หน้าต่างการจัด cluster โดยอาศัย Complete Link Technique (CLT)

4.4.1.1.3 การเข้าสู่ Average link technique (ALT)

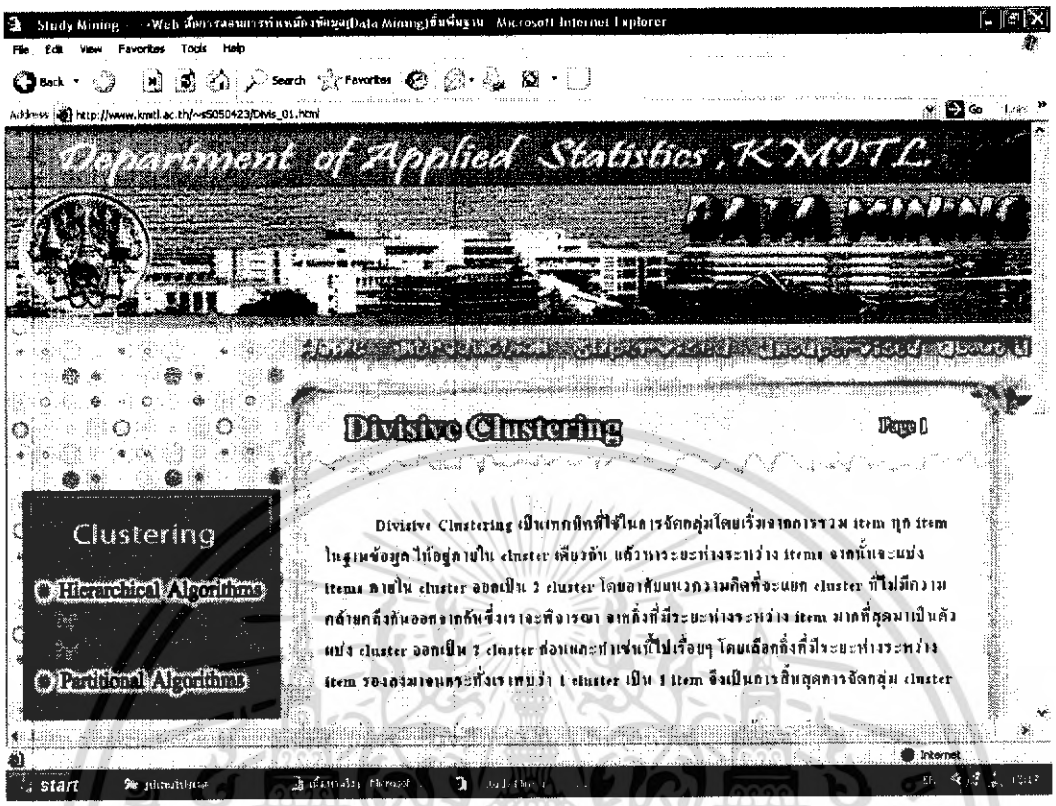
ผู้ศึกษาจะเข้าสู่ Average Link Technique (ALT) ได้ก็ต่อเมื่อคลิกเลือก [redacted] จากแถบเมนูด้านข้าง เมื่อเข้าสู่เว็บเพจแล้วภายในหน้าการจัด cluster แบบ Hierarchical Algorithms ด้วยวิธี Agglomerative Algorithms โดยอาศัย Average Link Technique (ALT) จะกล่าวถึง ขั้นตอนการจัด cluster โดย Average Link Technique (ALT) พร้อมทั้งยกตัวอย่างประกอบ



รูปที่ 4-29 หน้าต่างการจัด cluster โดยอาศัย Average Link Technique (ALT)

4.4.1.2 การเข้าสู่เทคนิคการจัด cluster แบบ Hierarchical Algorithms ด้วยวิธีการ Divisive Clustering

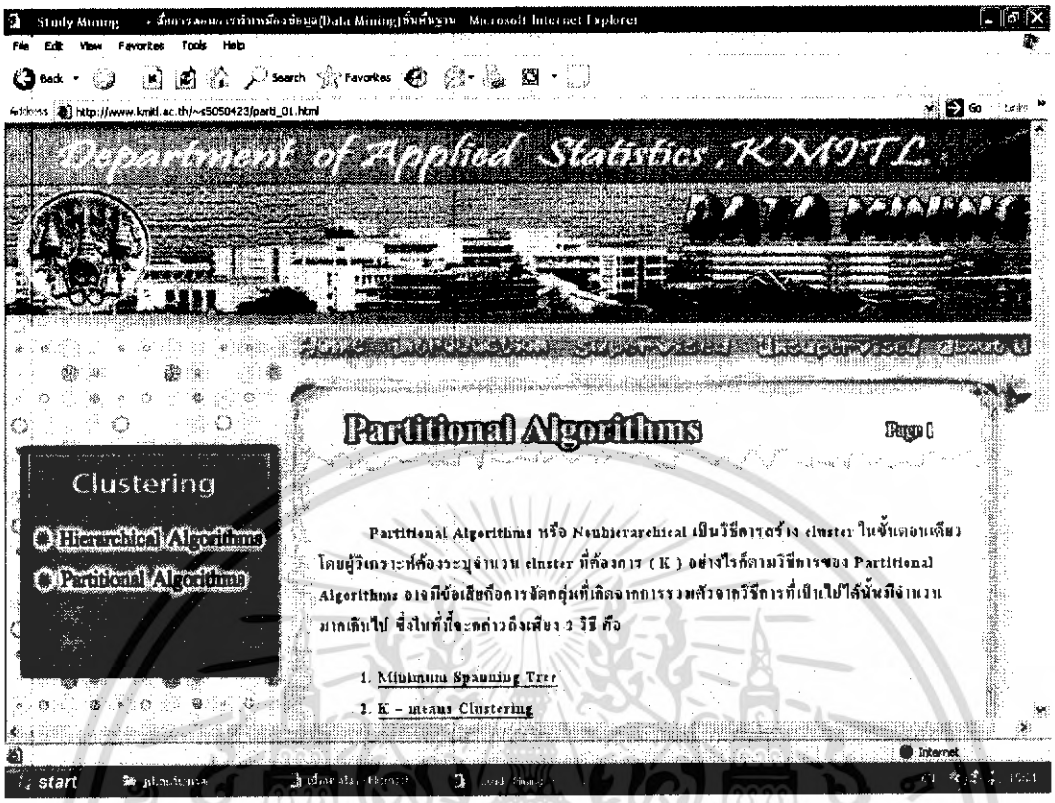
ผู้ศึกษาสามารถเข้าสู่เทคนิคการจัด cluster แบบ Hierarchical Algorithms ด้วยวิธีการ Divisive Clustering โดยการคลิกเลือก [redacted] จากแถบเมนูด้านข้าง เมื่อเข้าสู่เว็บเพจแล้ว ภายในหน้า Divisive Clustering จะกล่าวถึงวิธีการจัด cluster โดยอาศัยวิธีการ Divisive Clustering พร้อมทั้งยกตัวอย่างประกอบ



รูปที่ 4-30 หน้าต่างบทนำการจัด cluster แบบ Hierarchical Algorithms ด้วยวิธีการ Divisive Clustering

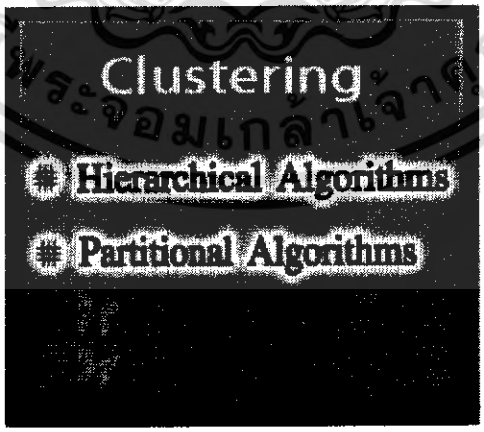
4.4.2 การเข้าสู่เทคนิคการจัด cluster แบบ Partitional Algorithms

ผู้ศึกษาจะเข้าสู่หน้าบทนำการจัด cluster แบบ Partitional Algorithms ได้ก็ต่อเมื่อคลิกที่ปุ่ม Partitional Algorithms เมื่อได้เข้าสู่เว็บเพจแล้วภายในการจัด cluster จะกล่าวถึง วิธีการจัด cluster แบบ Partitional Algorithms ซึ่งมีหน้าต่างดังรูปที่ 4-31



รูปที่ 4-31 หน้าต่าง Clustering เมื่อเราเลือกเมนู Partitional Algorithms

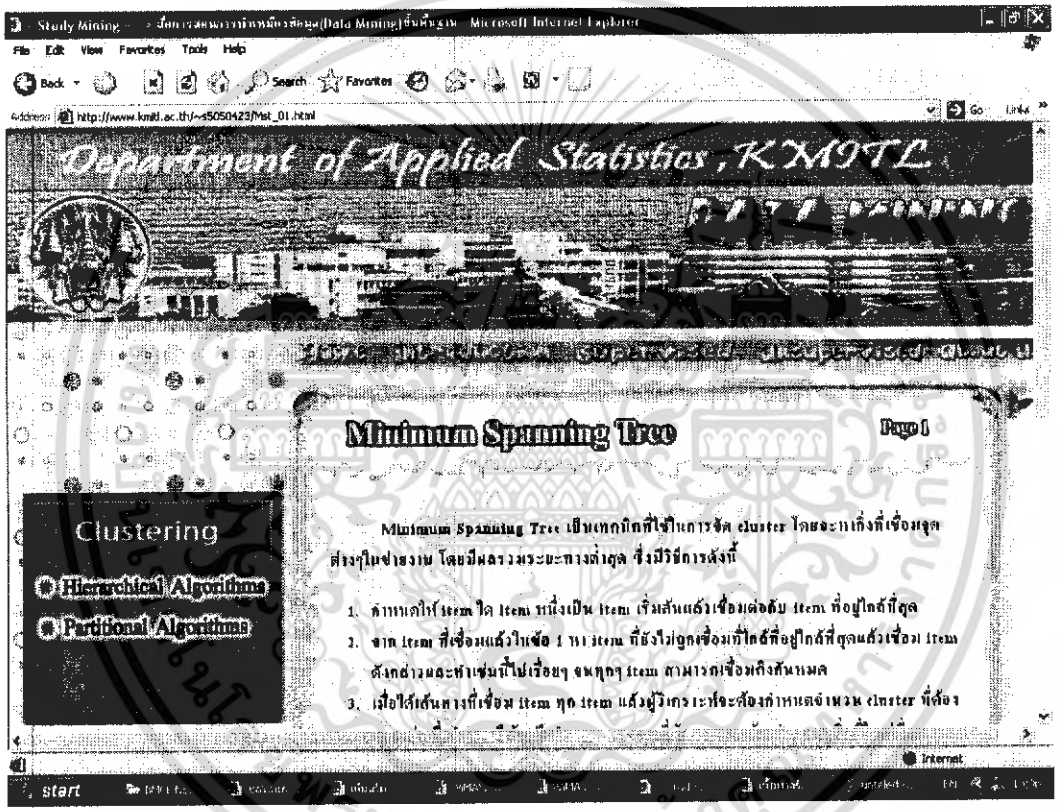
โดยผู้ศึกษาสามารถเลือกที่จะเข้าศึกษาในเทคนิคการจัด cluster แบบ Partitional Algorithmsได้จากลิงค์ภายในเนื้อหาและแถบเมนูด้านข้าง ซึ่งแบ่งเทคนิคการจัด cluster แบบ Partitional Algorithms เป็น 2 เทคนิค คือ Minimum Spanning Tree และ K-means Clustering ที่ปรากฏอยู่ในทุกหน้าต่างของเทคนิคการจัด cluster ดังรูปที่ 4-32



รูปที่ 4-32 แถบข้างที่แสดงในหน้าต่างการจัด cluster แบบ Partitional Algorithms

4.4.2.1 การเข้าสู่เทคนิคการจัด cluster แบบ Partitional Algorithms ด้วยเทคนิค Minimum Spanning Tree (MST)

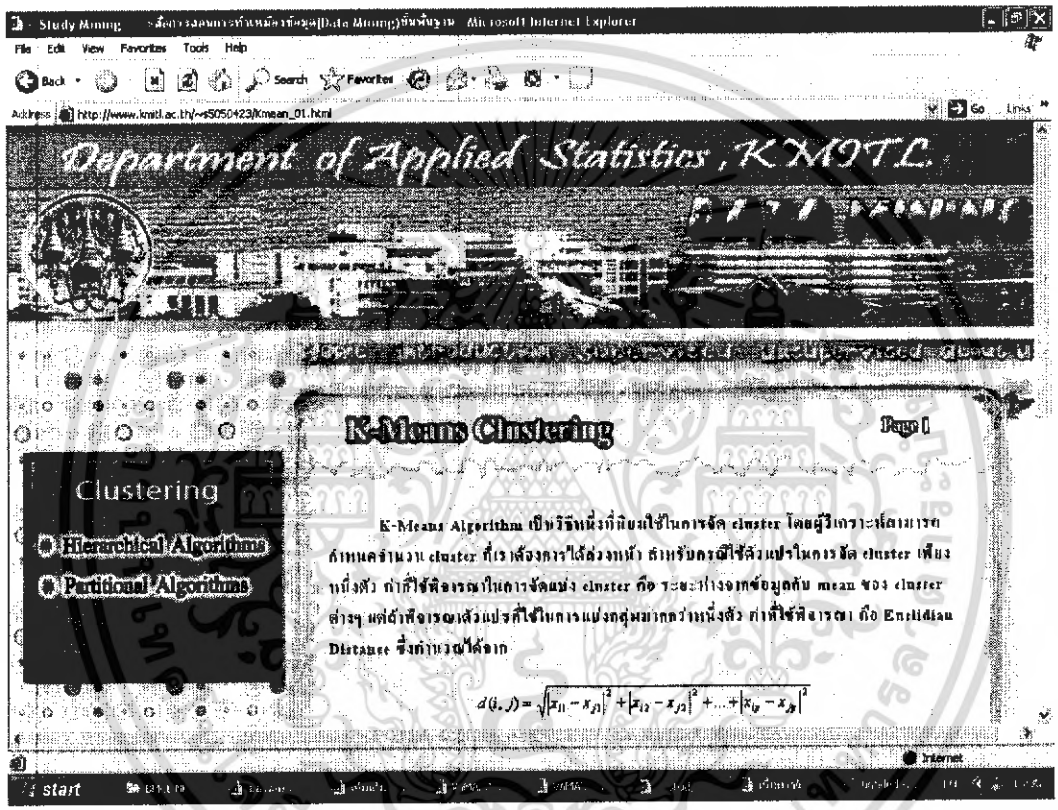
ผู้ศึกษาสามารถเข้าสู่เทคนิคการจัด cluster แบบ Partitional Algorithms ด้วยวิธีการ Minimum Spanning Tree (MST) โดยการคลิกเลือก [redacted] จากแถบเมนูด้านข้าง เมื่อเข้าสู่เว็บเพจแล้ว ภายในหน้า Minimum Spanning Tree (MST) จะกล่าวถึงขั้นตอนการจัดcluster โดยอาศัยเทคนิค Minimum Spanning Tree (MST) พร้อมทั้งยกตัวอย่างประกอบ



รูปที่ 4-33 หน้าต่างบทนำการจัด cluster แบบ Partitional Algorithms โดยอาศัยเทคนิค Minimum Spanning Tree (MST)

4.4.2.2 การเข้าสู่เทคนิคการจัดcluster แบบ Partitional Algorithms ด้วยเทคนิค K – Means Clustering

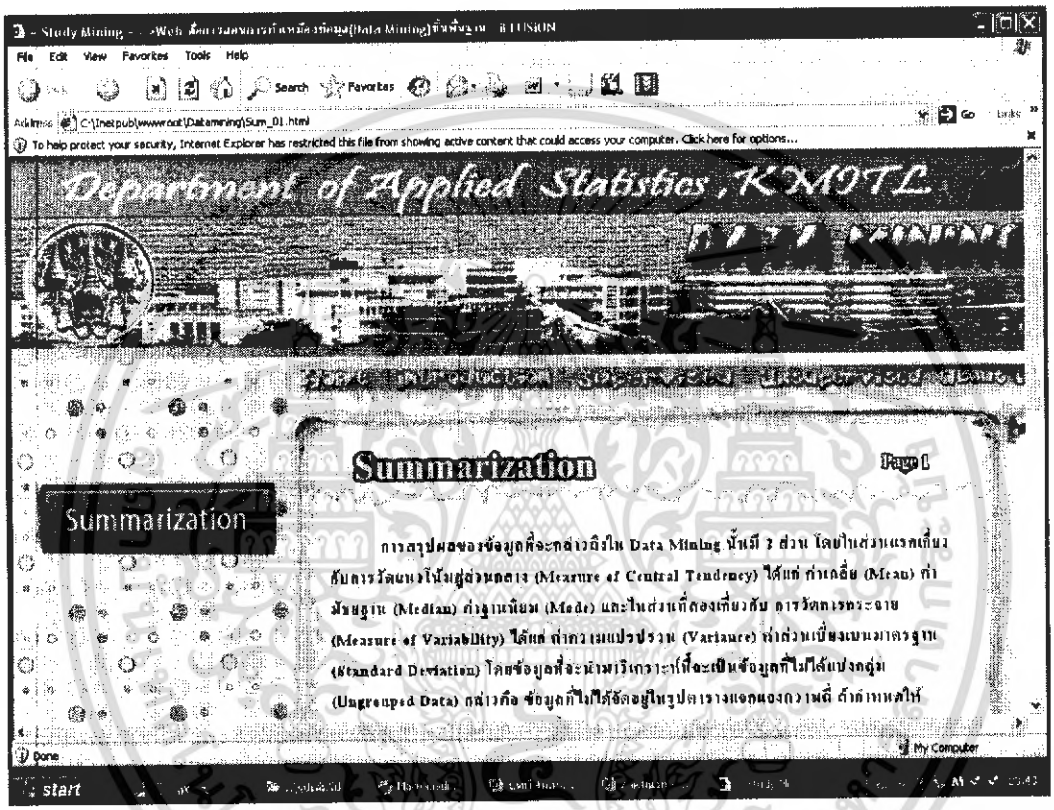
ผู้ศึกษาสามารถเข้าสู่เทคนิคการจัด cluster แบบ Partitional Algorithms ด้วยวิธีการ K – Means Clustering โดยการคลิกเลือก [redacted] จากแถบเมนูด้านข้าง เมื่อเข้าสู่เว็บเพจแล้ว ภายในหน้า K – Means Clustering จะกล่าวถึงขั้นตอนการจัด cluster โดยอาศัยเทคนิค K – Means Clustering พร้อมทั้งยกตัวอย่างประกอบ



รูปที่ 4-34 หน้าต่างบทนำการจัด cluster แบบ Partitional Algorithms โดยอาศัยเทคนิค K – Means Clustering

4.5 การเข้าสู่หน้า Summarization

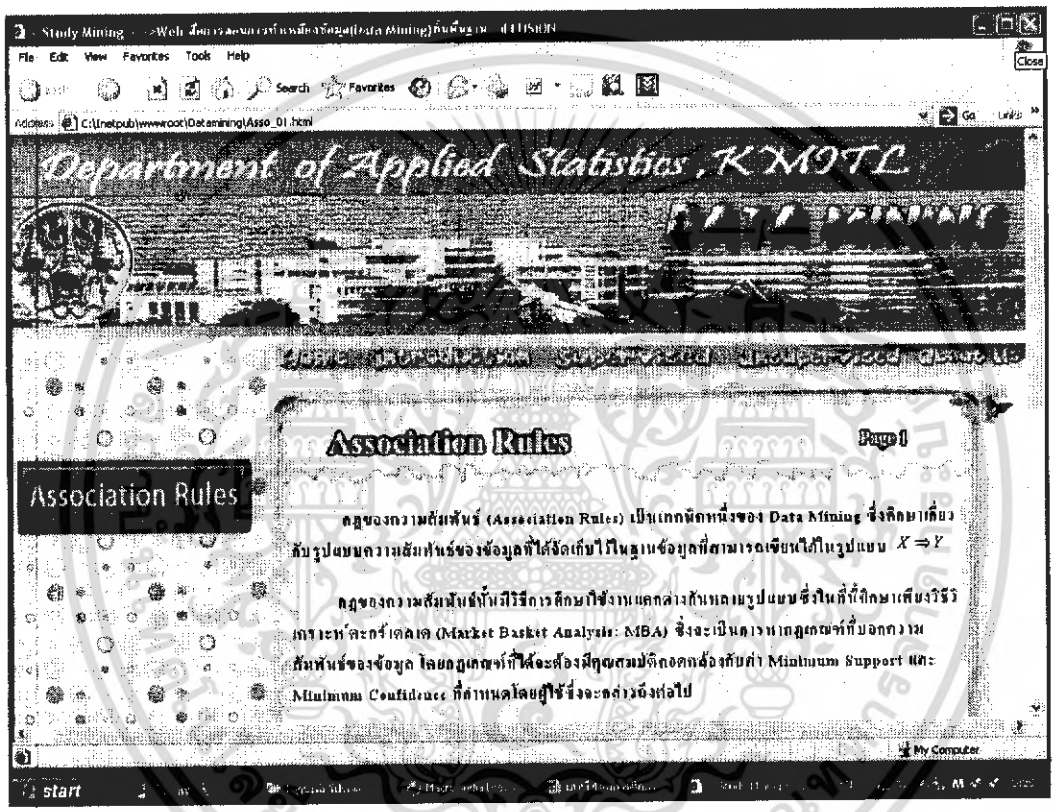
ผู้ศึกษาจะเข้าสู่หน้า Summarization ได้ก็ต่อเมื่อเลื่อนเมาส์ไปที่ปุ่ม แล้วทำการเลือกหัวข้อ Summarization บน Popup Menu เมื่อได้เข้าสู่เว็บเพจแล้วภายในหน้า Summarization จะกล่าวถึง ค่าที่ใช้ในการทำ Summarization เช่น ค่าเฉลี่ยเลขคณิต ค่ามัธยฐาน ค่าส่วนเบี่ยงเบนมาตรฐาน ฯลฯ และวิธีการคำนวณหาค่าดังกล่าว



รูปที่ 4-35 หน้าต่าง Summarization เมื่อเลือกเมนู Summarization

4.6 การเข้าสู่หน้ากฎของความสัมพันธ์ (Association Rules)

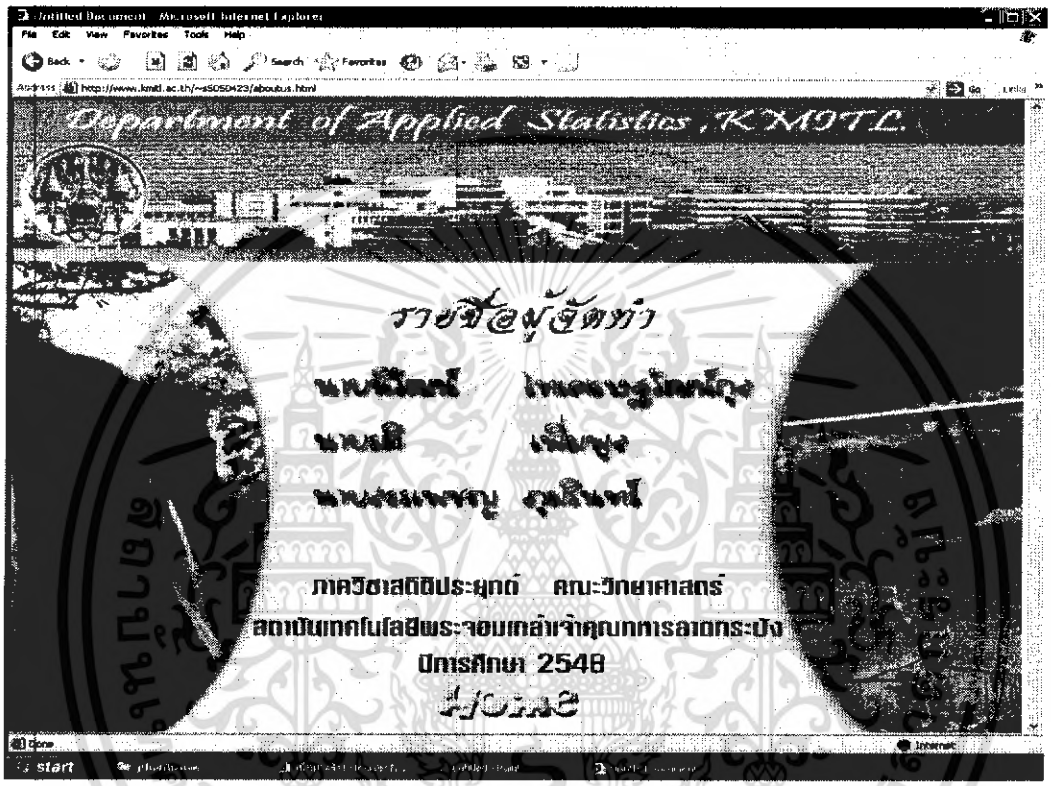
ผู้ศึกษาจะเข้าสู่หน้า Association Rules ได้ก็ต่อเมื่อเลื่อนเมาส์ไปที่ปุ่ม แล้วทำการเลือกหัวข้อ Association Rules บน Popup Menu เมื่อได้เข้าสู่เว็บเพจแล้วภายในหน้า Association Rules จะกล่าวถึงความหมายของกฎความสัมพันธ์ (Association Rules) วิธีการหากฎความสัมพันธ์ โดยการวิเคราะห์ตะกร้าตลาด (Market Basket Analysis) รวมทั้งการอธิบายถึงขั้นตอนในการวิเคราะห์หากฎของความสัมพันธ์โดยใช้กฎต่างในการวิเคราะห์



รูปที่ 4-36 หน้าต่าง Association Rules เมื่อเลือกเมนู Association Rules

4.7 การเข้าสู่หน้า About us

ถ้าผู้ศึกษาต้องการทราบถึงรายละเอียดเกี่ยวกับผู้จัดทำสื่อการสอนผ่านระบบเครือข่ายอินเทอร์เน็ตต้องคลิกที่ปุ่ม  จากแถบเมนูด้านบนเพื่อเข้าสู่หน้าเว็บเพจดังกล่าว และถ้าต้องการกลับไปหน้าแรกให้คลิกที่ปุ่ม



รูปที่ 4-37 หน้าต่างหน้า About us เมื่อเลือกเมนู 

บทที่ 5
สรุปผลการศึกษาและข้อเสนอแนะ

5.1 ผลสรุป

การศึกษาครั้งนี้เป็น การสร้างสื่อการสอนเรื่อง การทำเหมืองข้อมูลขั้นพื้นฐาน ผ่านระบบเครือข่ายอินเทอร์เน็ต ที่เว็บไซต์ http://www.studymining.host.sk ซึ่งบรรลุตามวัตถุประสงค์ที่ตั้งไว้ กล่าวคือ สามารถเป็นแหล่งข้อมูลทางวิชาการสำหรับผู้ที่สนใจได้ทำความเข้าใจเกี่ยวกับเนื้อหาเรื่อง Data Mining ขั้นพื้นฐาน และผู้ศึกษาสามารถใช้ความรู้ดังกล่าวมาช่วยในการวิเคราะห์ข้อมูลด้วยตนเอง โดยในแต่ละเรื่องที่ได้กล่าวถึงในสื่อการสอนได้มีการจัดทำเนื้อหาอย่างละเอียด และได้ทำการเรียบเรียงเนื้อหาเพื่อให้ง่ายแก่ความเข้าใจ มีการบอกวิธีการวิเคราะห์ด้วยเทคนิคต่างๆอย่างเป็นขั้นตอนโดยละเอียด พร้อมทั้งได้ยกตัวอย่างประกอบเพื่อเสริมสร้างความเข้าใจในการศึกษาเรียนรู้ให้มากขึ้น ดังนั้นจึงคาดว่าสื่อการสอนดังกล่าวจะเป็นประโยชน์สำหรับผู้ที่สนใจศึกษาหาความรู้ด้วยตนเอง

5.2 ข้อเสนอแนะ

เนื่องจากระยะเวลาในการสร้างสื่อการสอนนี้มีระยะเวลาจำกัด ซึ่งส่งผลให้ผู้จัดทำสามารถสร้างสื่อการสอนเกี่ยวกับเทคนิคต่างๆในการทำเหมืองข้อมูลเพียงบางเทคนิคเท่านั้น ดังนั้นจึงน่าจะเป็นประโยชน์ที่จะศึกษาเพิ่มเติมสำหรับเทคนิคอื่นๆต่อไป

นอกจากนี้ขั้นตอนการดำเนินงานในเทคนิคต่างๆ เป็นลักษณะวนลูปทำซ้ำ และมีการคำนวณที่ยุ่งยากซับซ้อน โดยเฉพาะกรณีที่มีข้อมูลมีจำนวนมากหรือตัวแปรที่เกี่ยวข้องมีจำนวนมาก ดังนั้นการสร้างโปรแกรมที่จะนำมาใช้ในการวิเคราะห์ข้อมูลสำหรับแต่ละเทคนิคจึงน่าจะเป็นประโยชน์และทำให้ผู้ที่สนใจสามารถวิเคราะห์ข้อมูลได้ง่ายยิ่งขึ้นและส่งผลให้เทคนิคต่างๆใน Data Mining เป็นที่นิยมใช้กันอย่างแพร่หลายในอนาคต

บรรณานุกรม

สถาบันบัณฑิตพัฒนบริหารศาสตร์. 2547. เอกสารประกอบการสัมมนาทางวิชาการเรื่อง ความรู้พื้นฐานทางการทำเหมืองข้อมูล, กรุงเทพฯ.

ณัฐพงษ์ สววิบูลย์ และ อินทกะ พิริยะกุล. 2546. ปัญหาพิเศษเรื่อง การประยุกต์ใช้ดาต้าไมนิ่งในทางธุรกิจ. ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง, กรุงเทพฯ.

ดวงพร เกี๋ยงคำ และ วงศ์ประชา จันทร์สมวงศ์. 2548. อินไซท์ Dreamweaver MX 2004: สำนักพิมพ์ Provision, กรุงเทพฯ.

ธนพงษ์ นิติการุญ และ ปราการ อัศวชีวันธรกุล. 2545. ปัญหาพิเศษเรื่อง ดาต้าไมนิ่ง. ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง, กรุงเทพฯ.

นุชชิดา พิริยะพงศธร. 2548. การใช้ Clustering โดย SQL Server Clustering using SQL Server. คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง, กรุงเทพฯ.

พัชร ทิพย์เกตุ และคณะ. 2547. ปัญหาพิเศษเรื่อง โปรแกรมวิเคราะห์และพยากรณ์ข้อมูลอนุกรมเวลาผ่านเครือข่ายคอมพิวเตอร์. ภาควิชาสถิติประยุกต์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง, กรุงเทพฯ.

พนิดา พานิชกุล และ สุรเชษฐ์ วงศ์ชัยพรพงษ์. 2546. คัมภีร์ Dreamweaver MX 2004.: สำนักพิมพ์ เคทีพี, กรุงเทพฯ.

พยูน พาณิชย์กุล. 2548. การใช้ Decision Tree โดย SQL Server Decision Tree using SQL Server. คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง, กรุงเทพฯ.

เอกเทพ ภักดีศิริมงคล. 2548. ทีเด็ดตกแต่งภาพ Photoshop CS.: สำนักพิมพ์ สวัสดีไอที, กรุงเทพฯ.

เอกลักษณ์ รัตนเจริญพงศ์ และ วิฑูรย์ พิทักษ์วีระกุล. 2546. ปัญหาพิเศษเรื่อง การทำเหมืองภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง, กรุงเทพฯ.

บรรณานุกรม (ต่อ)

Margaret H. Dunhan. 2003. Data Mining Introductory And Advance Topics. Pearson Education INC., United States of America.

Mehmed Kantardzic. 2003. Data Mining Concepts, Models And Algorithms. A JOHN WILEY & SONS INC., United States of America.

ภาคผนวก

เทคนิคต่างๆ ใน Data Mining

1. บทนำ (Introduction)

ในปัจจุบันองค์การส่วนใหญ่จะประสบกับปัญหาของการที่มีข้อมูลดิบเป็นจำนวนมากแต่สามารถนำสารสนเทศมาใช้ประโยชน์ได้น้อย เนื่องจากยังขาดความรู้ ความเข้าใจในการนำข้อมูลดังกล่าวมาวิเคราะห์ ทั้งนี้วิธีการวิเคราะห์ข้อมูลทางสถิติเป็นวิธีการที่ยุ่งยาก ซับซ้อนผู้วิเคราะห์ต้องมีความรู้ ความเข้าใจเป็นอย่างดี อีกทั้งวิธีการวิเคราะห์ทางสถิติยังมีข้อจำกัดบางประการ เช่น ข้อสมมติเกี่ยวกับข้อมูลทำให้เกิดวิธีการข้อมูลแนวใหม่เรียกว่า การทำเหมืองข้อมูล (Data Mining) ซึ่งในปัจจุบันมีการนำแนวคิดนี้มาประยุกต์ใช้ในทางธุรกิจและอื่นๆอีกมากมาย เช่น การจัดการความสัมพันธ์กับลูกค้า (CRM), การวิเคราะห์ข้อมูลในชีวสารสนเทศ (Bioinformatics) เป็นต้น

Data Mining หรือ การทำเหมืองข้อมูลเป็นกระบวนการค้นหาความรู้ ความสัมพันธ์ และรูปแบบของข้อมูลที่มีอยู่ในฐานข้อมูลขนาดใหญ่โดยอาศัยการผสมผสานแนวคิดทางสารสนเทศและวิธีทางสถิติ ซึ่งประเภทของการเรียนรู้ (Type of Learning Method) สามารถแบ่งได้เป็น 2 ประเภท คือ Supervised Learning และ Unsupervised Learning ดังนี้

1. **Supervised Learning**

Supervised Learning เป็นกระบวนการเรียนรู้เพื่อหาความสัมพันธ์ของข้อมูลจากข้อมูล input – output ที่ทราบค่าซึ่งเรียกว่า ชุดข้อมูลฝึกฝน (Training Data) ทำหน้าที่เสมือนเป็นผู้สอนและสร้างเป็นระบบการเรียนรู้ (Learning System) เพื่อใช้ในการประมาณค่า output จากข้อมูล input ชุดใหม่ โดยงานที่เกี่ยวข้องกับการเรียนรู้ประเภทนี้คือ

1.1. การจัดแบ่งหมวดหมู่ (Classification)

การจัดแบ่งหมวดหมู่หรือการแยกประเภท ซึ่งใช้ในการจำแนกประเภทของข้อมูล Output ที่เราต้องการโดยอาศัยเทคนิคต่างๆใน Data Mining ดังต่อไปนี้

1.1.1 การแบ่งหมวดหมู่เป็น 2 class (Classification into 2 classes) มี 2 วิธีดังนี้

1.1.1.1 Division Regression

วิธีการนี้ได้นำแนวคิดของ การถดถอย ซึ่งเป็นเทคนิคในทางสถิติในการหาจุดแบ่ง หรือเกณฑ์เพื่อแบ่งข้อมูลออกเป็น 2 class

1.1.1.2 Logistic Regression

วิธีการนี้ได้นำแนวคิดของ Logistic Regression ซึ่งเป็นเทคนิคในทางสถิติมาใช้ในการประมาณความน่าจะเป็นที่จะเกิดความสำเร็จของตัวแปร output จากตัวแปร input เพื่อนำความน่าจะเป็นดังกล่าวมาใช้ในการจัดแบ่งข้อมูลออกเป็น 2 class

1.1.2 การแบ่งหมวดหมู่มากกว่า 2 class (Classification into more than 2 classes) มี 3 วิธีดังนี้

1.1.2.1 Predictive Regression

วิธีการนี้ได้นำแนวคิดของ การถดถอย ซึ่งเป็นเทคนิคในทางสถิติในการหาตัวแบบที่ดีที่สุดเพื่อนำมาใช้ในการพยากรณ์ class จากข้อมูล input ที่กำหนดให้

1.1.2.2 Bayesian Inference

วิธีการนี้ได้นำแนวคิดของ เบย์ ซึ่งเป็นเทคนิคทางสถิติมาใช้ในการหาความน่าจะเป็น เพื่อนำความน่าจะเป็นดังกล่าวไปจัดข้อมูลว่าควรจะอยู่ใน class ใด

1.1.2.3 ต้นไม้การตัดสินใจ (Decision Tree)

วิธีการนี้ได้นำแนวคิดของ ต้นไม้การตัดสินใจ ซึ่งเป็นเทคนิคทางสถิติมาใช้ในการสร้างกฎเกณฑ์เพื่อใช้ในการแบ่ง class ของข้อมูล

2. Unsupervised Learning

Unsupervised Learning เป็นกระบวนการเรียนรู้เพื่อหารูปแบบที่เหมาะสมว่า output ควรเป็นอย่างไร จากลักษณะของข้อมูลในชุดข้อมูลฝึกฝนซึ่งมีเพียงข้อมูล input เท่านั้น โดยงานที่เกี่ยวกับการเรียนรู้ประเภทนี้ คือ

2.1 การจัด Cluster (Clustering)

การจัด Cluster หรือ การแบ่ง Cluster ซึ่งใช้ในการจัดกลุ่มของข้อมูล ซึ่งศึกษาจากข้อมูล inputโดยภายในกลุ่ม (cluster) เดียวกันจะมีความคล้ายคลึงกันมากที่สุดและระหว่าง cluster ต่างกันจะมีความแตกต่างกันมากที่สุด ซึ่งการจัดกลุ่มจะแตกต่างจากการจัดแบ่งหมวดหมู่ (Classification) คือ การจัดกลุ่มพยายามหาความคล้ายคลึงของข้อมูล input เพื่อสร้างกลุ่ม

โดยไม่มีการกำหนด class ของข้อมูลเอาไว้ก่อน ซึ่ง Algorithms ในการจัด Cluster ของข้อมูล โดยทั่วไปแบ่งเป็น 2 ประเภท คือ

2.1.1 การจัด Cluster แบบลำดับชั้น (Hierarchical Algorithms)

การจัด Cluster แบบลำดับชั้นนี้จะกำหนดให้ข้อมูลแต่ละค่าเป็น 1 cluster แล้วจัด cluster ที่อยู่ใกล้กันเข้าไว้ด้วยกันเป็น cluster ใหม่ 1 cluster แล้วทำเช่นนี้ต่อไป ซึ่งจะกล่าวถึง 2 วิธีดังนี้

2.1.1.1 การจัด Cluster แบบลำดับชั้นจากล่างขึ้นบน (Agglomerative Algorithms)

2.1.1.2 การจัด Cluster แบบลำดับชั้นจากบนลงล่าง (Divisive Clustering)

2.1.2 การจัด Cluster แบบแบ่งเป็นส่วน (Partitional Algorithms)

การจัด Cluster แบบแบ่งเป็นส่วนเริ่มจากให้ข้อมูลทั้งหมดเป็น cluster เดียว แล้วจึงแบ่งออกเป็น cluster ตามที่ต้องการด้วยเทคนิคต่างๆ

2.2 Summarization

Summarization เป็นกระบวนการที่ใช้ในการวิเคราะห์หาลักษณะเบื้องต้นของข้อมูลในฐานข้อมูลเช่น ค่าเฉลี่ย ค่ามัธยฐาน ค่าความแปรปรวน และส่วนเบี่ยงเบนมาตรฐาน

2.3 กฎของความสัมพันธ์ (Association Rules)

การวิเคราะห์หาความสัมพันธ์ของข้อมูลในฐานข้อมูลเพื่อหากฎความสัมพันธ์หรือความน่าจะเป็นของข้อมูลที่เราสนใจจากฐานข้อมูลที่จะเกิดขึ้นพร้อมกัน

จากที่กล่าวมานี้จะเห็นได้ว่า เทคนิคการวิเคราะห์ข้อมูลของ Data Mining เป็นอีกทางเลือกที่ผู้ใช้สามารถนำมาวิเคราะห์ข้อมูล โดยผู้ใช้จะต้องเลือกให้เหมาะสมกับชนิดของข้อมูลและสิ่งที่มีอยู่ อย่างไรก็ตาม Data Mining ยังมีเทคนิคในการวิเคราะห์ข้อมูล ซึ่งจะเป็นประโยชน์ในการศึกษาต่อไป

2. Supervised Learning

Supervised Learning เป็นกระบวนการเรียนรู้เพื่อหาความสัมพันธ์ของข้อมูลจากข้อมูล input – output ที่ทราบค่าซึ่งเรียกว่า ชุดข้อมูลฝึกฝน (Training Data) ทำหน้าที่เสมือนเป็นผู้สอนและสร้างเป็นระบบการเรียนรู้ (Learning System) เพื่อใช้ในการประมาณค่า output จากข้อมูล input ชุดใหม่

2.1 Division Regression

การวิเคราะห์วิธีนี้ใช้สำหรับแบ่งข้อมูลของตัวแปร 1 ตัว ออกเป็น 2 class โดยค่าที่ใช้ในการแบ่งข้อมูลนั้นคือ ค่าเฉลี่ยของข้อมูลที่ต้องการแบ่ง class อย่างไรก็ตามวิธีการนี้ได้นำแนวคิดเกี่ยวกับการถดถอยมาใช้ ซึ่งถ้ากำหนดให้ค่า y คือ ตัวแปรที่ต้องการแบ่ง class จะได้สมการถดถอยของประชากรคือ

$$y = \beta + \varepsilon$$

เมื่อ β คือ จุดแบ่ง

และสมการถดถอยของตัวอย่าง ที่ได้จากการนำข้อมูลตัวอย่างซึ่งเป็นชุดข้อมูลฝึกฝนมาวิเคราะห์มีรูปแบบดังนี้

$$\hat{y} = b$$

เมื่อ $\hat{y}$ คือ ค่าประมาณของ y

b คือ ค่าประมาณของ β

โดยใช้วิธีกำลังสองน้อยที่สุด (Least Square Method) ในการหาค่า b จะได้

$$b = \frac{\sum_{i=1}^{n} y_i}{n}$$

ซึ่งคือ ค่าเฉลี่ยของข้อมูลที่ต้องการแบ่ง class

ตัวอย่างที่ 1 การแบ่งข้อมูลออกเป็น 2 class โดยใช้ข้อมูลความสูงของตัวอย่าง 12 คน ดังนี้

{1.6, 1.9, 1.88, 1.7, 1.85, 1.6, 1.7, 1.8, 1.95, 1.9, 1.8, 1.75}

วิธีทำ ให้ y เป็นข้อมูลตัวอย่างซึ่งต้องการแบ่งออกเป็น 2 class คือ short และ medium จะได้จุดแบ่งคือ

$$b = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$= \frac{21.43}{12}$$

$$= 1.786$$

กล่าวคือ ถ้าความสูงน้อยกว่า 1.786 m จะถูกจัดอยู่ใน class short และถ้าความสูงมากกว่าหรือเท่ากับ 1.786 m จะถูกจัดอยู่ใน class medium ซึ่งแสดงในรูปที่ 1

(a) ความสูงของข้อมูล (b)การแบ่ง class

รูปที่ 1 ข้อมูลความสูง

2.2 Logistic Regression

Logistic Regression เป็นเทคนิคหนึ่งของ Regression ซึ่งใช้ในกรณีที่ตัวแปร output มีค่าได้เพียง 2 ค่า ซึ่งนิยมกำหนดให้เป็น 0 และ 1 ในขณะที่ตัวแปร input หรือ กลุ่มของตัวแปร input อาจจะเป็นข้อมูลต่อเนื่อง ข้อมูลกลุ่ม อย่างใดอย่างหนึ่งหรือทั้งคู่ก็ได้ ตัวอย่างเช่น การศึกษาทางการแพทย์เกี่ยวกับการรอดชีวิตหรือการตายของผู้ป่วย

สิ่งที่ Logistic Regression แตกต่างจาก Linear Regression คือ Logistic Regression มิได้กำหนดว่าความสัมพันธ์ระหว่างตัวแปร output กับตัวแปร input ต่างๆ จะต้องเป็นเชิงเส้น นอกจากนี้ตัวแปร output และค่าคลาดเคลื่อน ไม่ได้กำหนดว่าจะต้องมีการแจกแจงแบบปกติ

ถ้ากำหนดให้ y เป็นตัวแปร output และ x_i เป็นตัวแปร input โดยที่ $i = 1,...,k$ และ p คือความน่าจะเป็นที่ตัวแปร y มีค่าเป็น 1 $(y = 1)$ จะสามารถเขียนตัวแบบของประชากร ได้ดังนี้

$$\log it(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

เมื่อ β_i คือ ค่าสัมประสิทธิ์การถดถอย

ε คือ ค่าความคลาดเคลื่อน

และตัวแบบของตัวอย่างที่ได้จากการนำข้อมูลตัวอย่างซึ่งเป็นชุดข้อมูลฝึกฝนมาวิเคราะห์มีรูปแบบดังนี้

$$\log it(\hat{p}) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

หรือ $\hat{p} = \dfrac{e^{\log it(\hat{p})}}{1 + e^{\log it(\hat{p})}}$

เมื่อ $\hat{p}$ คือ ค่าประมาณของ p

b_i คือ ค่าประมาณของสัมประสิทธิ์การถดถอย

โดยใช้วิธีกำลังสองน้อยที่สุด (Least Square Method) ในการหาค่า b_i

เมื่อเราได้ตัวแบบซึ่งหาจากชุดข้อมูลฝึกฝนแล้วนั้น ถ้าเราใส่ค่าของตัวแปร input ในตัวแบบ ผลที่ได้คือ $\log it(\hat{p})$ ซึ่งสามารถนำไปหาความน่าจะเป็น $(\hat{p})$ ตัวแปร output มีค่าเป็น 1 $(y = 1)$

เทคนิค Logistic Regression ใน Data Mining นั้นถูกนำมาใช้เพื่อแบ่งข้อมูลของตัวแปร output ออกเป็น 2 classโดยมีจุดแบ่งคือ ความน่าจะเป็นที่ได้จากตัวแบบซึ่งมีค่าเท่ากับ 0.5 กล่าวคือ ถ้าข้อมูลของตัวแปร input ชุดใดให้ความน่าจะเป็นน้อยกว่า 0.5 จะถูกจัดอยู่ใน class ที่ 1 ($y = 0$) แต่ถ้าความน่าจะเป็นมากกว่าหรือเท่ากับ 0.5 จะถูกจัดอยู่ใน class ที่ 2 ($y = 1$) โดยมีขั้นตอนดังนี้

1. สร้างตัวแบบของตัวอย่างจากชุดข้อมูลฝึกฝน
2. นำค่า Input ที่ต้องการจัดแบ่ง class มาแทนค่าในตัวแบบของตัวอย่างเพื่อคำนวณหาค่า $\log it(\hat{p})$
3. หาค่าความน่าจะเป็นที่ค่าตัวแปร output จะมีค่าเป็น 1 $(y = 1)$ จาก

$$\hat{p} = \frac{e^{\log it(\hat{p})}}{1 + e^{\log it(\hat{p})}}$$

4. สรุป class ของข้อมูลดังกล่าว โดยถ้า $\hat{p} < 0.5$ (เกณฑ์ในการแบ่ง class)จะได้ว่าข้อมูล input ดังกล่าวอยู่ใน class ที่ 1 $(y = 0)$ แต่ถ้า $\hat{p} \geq 0.5$ จะจัดข้อมูล input ดังกล่าวอยู่ใน class ที่ 2 $(y = 1)$

ตัวอย่างที่ 2 ถ้าตัวแบบของตัวอย่างที่คำนวณได้จากชุดข้อมูลฝึกฝนคือ

$$\log it(\hat{p}) = 1.5 - 0.6x_1 + 0.4x_2 - 0.3x_3$$

อยากทราบว่าชุดตัวแปร input $\{x_1, x_2, x_3\} = \{1, 0, 1\}$ จะถูกจัดอยู่ใน class ใด

วิธีทำ เมื่อนำค่า input ที่กำหนดมาแทนค่าในตัวแบบของตัวอย่างจะได้ว่า

$$\log it(\hat{p}) = 1.5 - 0.6(1) + 0.4(0) - 0.3(1)$$

$$= -0.6$$

คำนวณหา $\hat{p}$ จาก

$$\hat{p} = \frac{e^{\log it(\hat{p})}}{1 + e^{\log it(\hat{p})}}$$

$$= \frac{e^{-0.6}}{1 + e^{-0.6}}$$

$$= 0.35$$

จากข้อมูลชุดนี้ได้ความน่าจะเป็นที่ตัวแปร output มีค่าเป็น 1 เท่ากับ 0.35 ซึ่งน้อยกว่า 0.5 (เกณฑ์ในการแบ่ง class ที่ตั้งไว้) จะได้ว่าตัวแปร input = {1, 0, 1} จะถูกจัดอยู่ใน class ที่ 1 $(y = 0)$

2.3 Predictive Regression

การวิเคราะห์การถดถอย (Regression Analysis) เป็นเทคนิคทางสถิติเทคนิคหนึ่ง ซึ่งเป็นการศึกษาเกี่ยวกับความสัมพันธ์ระหว่างตัวแปรที่เรียกว่า ตัวแปร output (ตัวแปรตาม; dependent variable: y) ซึ่งเป็นตัวแปรที่เปลี่ยนแปลงไปตามการเปลี่ยนแปลงของตัวแปรอีกตัวหนึ่ง ซึ่งเรียกว่า ตัวแปร input (ตัวแปรอิสระ; independent variable: x) โดยในกรณีนี้จะพิจารณาตัวแปรอิสระ 1 ตัว และรูปแบบความสัมพันธ์ระหว่างตัวแปร output กับตัวแปร input นี้เป็นเส้นตรง ซึ่งเรียกเส้นตรงที่แทนความสัมพันธ์ระหว่างตัวแปรทั้งสองว่า เส้นถดถอย (Regression Line) และเส้นการถดถอยนี้นำไปใช้ในการพยากรณ์ค่าของตัวแปร output ต่อไป

สำหรับ Data Mining เมื่อได้ค่าพยากรณ์ของตัวแปร output แล้วจะต้องนำค่าดังกล่าวมาเปรียบเทียบกับเกณฑ์ของการแบ่ง class ที่ตั้งไว้แล้ว ว่าค่าพยากรณ์ดังกล่าวจะอยู่ใน class ใด โดยการวิเคราะห์ด้วยวิธีนี้สามารถใช้ได้ไม่ว่าจะมีจำนวน class เท่าใดก็ตาม

ถ้ากำหนดให้ y เป็นตัวแปร output และ x เป็นตัวแทน input จะสามารถเขียนรูปแบบสมการถดถอยของประชากร ได้ดังนี้

$$y = \beta_0 + \beta_1 x + \varepsilon$$

เมื่อ β_i คือ ค่าสัมประสิทธิ์ของการถดถอย

ε คือ ค่าความคลาดเคลื่อน โดยที่ $\varepsilon \sim NID(0, \sigma_\varepsilon^2)$

และสมการถดถอยของตัวอย่างที่ได้จากการนำข้อมูลซึ่งเป็นชุดข้อมูลฝึกฝนมาวิเคราะห์ มีรูปแบบดังนี้

$$\hat{y} = b_0 + b_1 x$$

เมื่อ $\hat{y}$ คือ ค่าประมาณของ y

b_i คือ ค่าประมาณของค่าสัมประสิทธิ์การถดถอย

ซึ่งใช้วิธีกำลังสองน้อยที่สุด (Least Square Method) ในการหาค่าประมาณของสัมประสิทธิ์การถดถอย (b_i)

ถ้าให้ค่าคลาดเคลื่อนของการพยากรณ์ คือ $\varepsilon_i = y_i - \hat{y}_i$ สำหรับ $i = 1,...,n$ และด้วยวิธีกำลังสองน้อยที่สุด ค่า b_0 และ b_1 ที่เหมาะสมคือ ค่าที่ทำให้ $\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ มีค่าน้อยที่สุด ซึ่งจะได้ว่า

$$b_1 = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

และ

$$b_0 = \bar{y} - b_1 \bar{x}$$

เมื่อ $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ และ $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

ข้อสังเกต จากสมการถดถอยของตัวอย่าง

$$\hat{y} = b_0 + b_1 x$$

ซึ่งโดยทั่วไปเราจะหาค่าพยากรณ์ของ output $(\hat{y})$ แล้วนำค่านี้ไปเทียบกับเกณฑ์ของตัวแปร output ที่เราตั้งไว้เพื่อใช้แบ่ง class กล่าวคือทุกครั้งที่มี x ค่าใหม่จะต้องคำนวณหา $\hat{y}$ ทุกครั้ง อย่างไรก็ตามมีวิธีที่ง่ายกว่าคือ นำค่า $\hat{y}$ ที่เป็นเกณฑ์ในการแบ่ง class มาแทนค่าในสมการถดถอยของตัวอย่างเพื่อหาค่า x และใช้ x ที่ได้นี้เป็นเกณฑ์ในการแบ่ง class ได้เช่นกัน

ตัวอย่างที่ 3 การสร้างสมการถดถอยโดยใช้ข้อมูลความสูงของตัวอย่าง 12 คน ซึ่งถูกจัดอยู่ใน class ต่างๆ ดังนี้

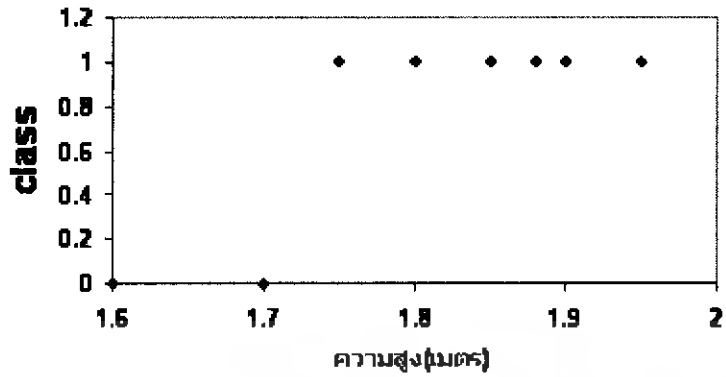
no.	height(m)	class
1	1.6	short
2	1.9	medium
3	1.88	medium
4	1.7	short
5	1.85	medium
6	1.6	short
7	1.7	short
8	1.8	medium
9	1.95	medium
10	1.9	medium
11	1.8	medium
12	1.75	medium

วิธีทำ จากข้อมูลพบว่าตัวแปร input คือ ตัวแปร height และตัวแปร output คือ ตัวแปร class ซึ่งตัวแปร class ประกอบด้วย 2 class คือ short และ medium ถ้าให้ short มีค่าเป็น 0 และ medium มีค่าเป็น 1

ให้ y เป็นตัวแปร class และ x เป็นตัวแปร height จะได้

x : height	1.6	1.9	1.88	1.7	1.85	1.6	1.7	1.8	1.95	1.9	1.8	1.75
y :class	0	1	1	0	1	0	0	1	1	1	1	1

ซึ่งสามารถแสดงข้อมูลความสูงและ class ได้ดังรูปที่ 2



รูปที่ 2 ข้อมูลความสูงและ class

จากข้อมูลตัวอย่างหาสมการถดถอยของตัวอย่าง ซึ่งมีรูปแบบดังนี้

$$\hat{y} = b_0 + b_1 x$$

การคำนวณค่าต่างๆ

x_i	y_i	$x_i y_i$	x_i^2
1.6	0	0	2.56
1.9	1	1.9	3.61
1.88	1	1.88	3.534
1.7	0	0	2.89
1.85	1	1.85	3.4225
1.6	0	0	2.56
1.7	0	0	2.89
1.8	1	1.8	3.24
1.95	1	1.95	3.803
1.9	1	1.9	3.61
1.8	1	1.8	3.24
1.75	1	1.75	3.063
$\sum_{i=1}^{12} x_i = 21.43$	$\sum_{i=1}^{12} y_i = 8$	$\sum_{i=1}^{12} x_i y_i = 14.83$	$\sum_{i=1}^{12} x_i^2 = 38.422$

แทนค่าที่คำนวณได้เพื่อหา b_0 และ b_1 ดังนี้

$$b_1 = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$= \frac{12(14.83) - (21.43)(8)}{12(38.422) - (21.43)^2}$$

$$= 3.584$$

$$\bar{y} = \frac{\sum_{i=1}^{12} y_i}{12}$$

$$= \frac{8}{12}$$

$$= 0.667$$

$$\bar{x} = \frac{\sum_{i=1}^{12} x_i}{12}$$

$$= \frac{21.43}{12}$$

$$= 1.786$$

$$b_0 = \bar{y} - b_1\bar{x}$$

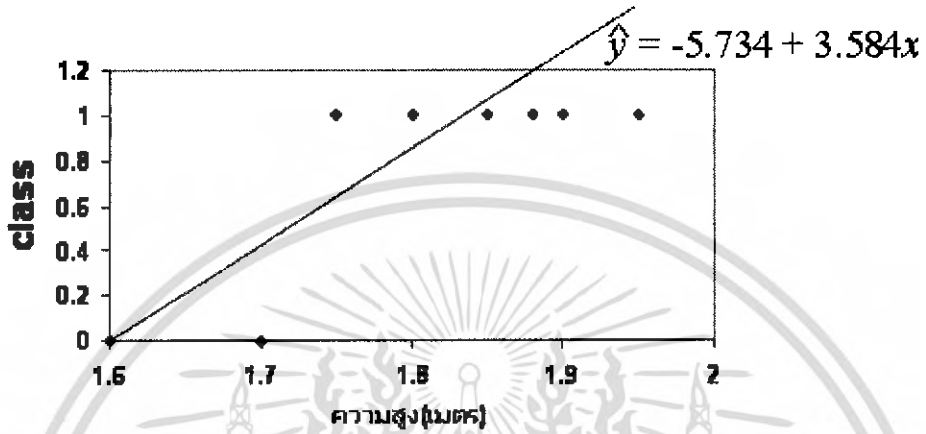
$$= 0.667 - (3.584 \times 1.786)$$

$$= -5.734$$

ดังนั้น สมการถดถอยของตัวอย่างคือ

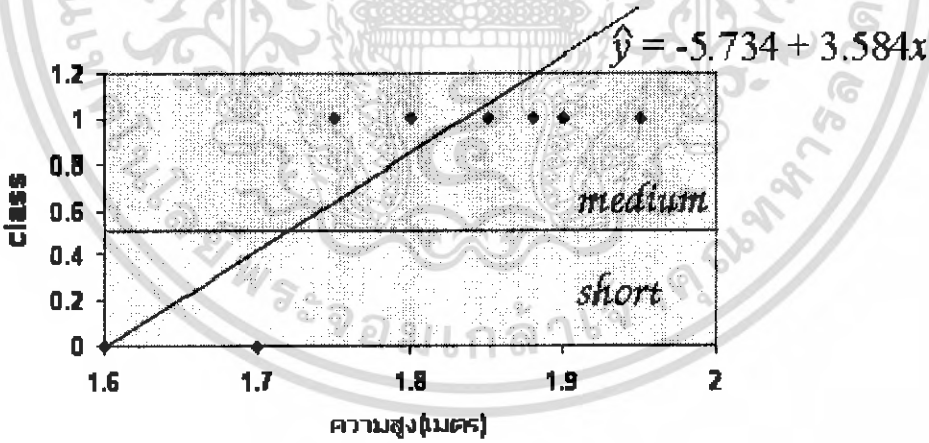
$$\hat{y} = -5.734 + 3.584x$$

ซึ่งสามารถแสดงสมการถดถอยของตัวอย่าง ได้ดังรูปที่ 3



รูปที่ 3 เส้นถดถอยของข้อมูล height กับ output

ถ้าเกณฑ์ในการแบ่ง class เป็นดังนี้ ถ้า $\hat{y} < 0.5$ จะให้ตัวแปร class เป็น short และถ้า $\hat{y} \geq 0.5$ จะให้ตัวแปร class เป็น medium ซึ่งแสดงในรูปที่ 4



รูปที่ 4 เส้นถดถอยและการพยากรณ์

ในการหาค่าความสูงเพื่อใช้เป็นเกณฑ์ในการแบ่ง class จากเกณฑ์ของตัวแปร class ซึ่งคือ $\hat{y} = 0.5$ มาคำนวณจะได้

$$0.5 = -5.734 + 3.584x$$

$$x = 1.74$$

กล่าวคือ ถ้าผู้ใดมีความสูงน้อยกว่า 1.74 m จะให้ตัวแปร class เป็น short และถ้าผู้ใดมีความสูงมากกว่าหรือเท่ากับ 1.74 m จะให้ตัวแปร class เป็น medium

2.4 Bayesian Inference

เทคนิค Bayesian Inference ใน Data Mining เป็นวิธีการจัด class ข้อมูลชุดใหม่ซึ่งอาศัยทฤษฎีของเบย์ โดยใช้ชุดข้อมูลฝึกฝนที่ประกอบด้วยข้อมูลของตัวแปร input และตัวแปร output ซึ่งแสดง class ไว้เรียบร้อยแล้ว ควบคู่กับข้อมูลชุดใหม่ที่ต้องการจัด class โดยมีข้อกำหนดที่ว่าตัวแปร input จะต้องเป็นอิสระกัน

กำหนดให้ S คือ ชุดข้อมูลฝึกฝน ซึ่งมี n ตัวอย่างและถูกแบ่งเป็น m class กล่าวคือ $S = \{S_1,...,S_n\}$ โดยที่ข้อมูลแต่ละตัวอย่างจะประกอบด้วย ข้อมูลของตัวแปร input (A_j) k ตัวแปร และข้อมูล output ซึ่งแสดง class (C_i) กล่าวคือ $S_h = \{A_1,..., A_k, C_i\}$ สำหรับ $h = 1, 2,...,n$

ถ้า X คือข้อมูลชุดใหม่ ซึ่งประกอบไปด้วยค่าของตัวแปร input k ตัวแปร กล่าวคือ $X = \{a_1, a_2,..., a_k\}$

ขั้นตอนในการคำนวณเพื่อจัด class ของข้อมูลชุดใหม่มีดังนี้

1. คำนวณหาความน่าจะเป็นของ class ที่ i จากชุดข้อมูลฝึกฝน

$$P(C_i) = \frac{\text{จำนวนของข้อมูล input ที่มีอยู่ใน class ที่ } i}{\text{จำนวนข้อมูลทั้งหมด}}$$

2. คำนวณหา $P(A_j \mid C_i)$ สำหรับทุก i และ j
3. คำนวณหา $P(X \mid C_i)$ โดยที่

$$P(X \mid C_i) = \prod_{j=1}^{k} P(A_j \mid C_i) \text{ สำหรับทุก } i$$

4. คำนวณหา $P(C_i \mid X) = P(X \mid C_i)P(C_i)$ สำหรับทุก i
5. จัดชุดข้อมูล X ให้อยู่ใน class ที่ i ซึ่งมี $P(C_i \mid X)$ สูงที่สุด

ตัวอย่างที่ 4 การจัด class ของข้อมูลชุดใหม่ {1, 2, 2} ด้วยเทคนิค Bayesian Inference โดยใช้ชุดข้อมูลฝึกฝนต่อไปนี้

ตัวอย่างที่	ตัวแปร A_1	ตัวแปร A_2	ตัวแปร A_3	class
1	1	2	1	1
2	0	0	1	1
3	2	1	2	2
4	1	2	1	2
5	0	1	2	1
6	2	2	2	2
7	1	0	1	1

วิธีทำ จากชุดข้อมูลฝึกฝนซึ่งมีทั้งหมด 7 ตัวอย่าง ประกอบด้วย ตัวแปร input 3 ตัวแปร คือตัวแปร A_1, A_2, A_3 $(k = 3)$ และตัวแปร output คือ ตัวแปร class ซึ่งแบ่งได้ 2 class $(i = 1, 2)$ และชุดข้อมูลใหม่ $X = \{1, 2, 2\}$ $a_1 = 1, a_2 = 2, a_3 = 2$ นั่นคือจะได้

1. คำนวณหาความน่าจะเป็นของ class ที่ i จากชุดข้อมูลฝึกฝน

$$P(C_1) = \frac{\text{จำนวนชุดของข้อมูล input ใน class ที่ 1}}{\text{m}}$$

$$= \frac{4}{7}$$

$$= 0.5714$$

$$P(C_2) = \frac{\text{จำนวนชุดของข้อมูล input ใน class ที่ 2}}{\text{m}}$$

$$= \frac{3}{7}$$

$$= 0.4286$$

2. คำนวณหา $P(A_j \mid C_i)$ สำหรับทุก i และ j

จาก $X = \{1, 2, 2\}$ จะได้

$$P(A_1 = 1 \mid C_1) = \frac{2}{4} = 0.5$$

$$P(A_1 = 1 \mid C_2) = \frac{1}{3} = 0.33$$

$$P(A_2 = 2 \mid C_1) = \frac{1}{4} = 0.25$$

$$P(A_2 = 2 \mid C_2) = \frac{2}{3} = 0.66$$

$$P(A_3 = 2 \mid C_1) = \frac{1}{4} = 0.25$$

$$P(A_3 = 2 \mid C_2) = \frac{2}{3} = 0.66$$

3. คำนวณหาค่า $P(X \mid C_i)$ จาก $P(X \mid C_i) = \prod_{j=1}^{k} P(A_j \mid C_i)$ สำหรับทุก i ดังนี้

$$P(X \mid C_1) = P(A_1 = 1 \mid C_1) \times P(A_2 = 2 \mid C_1) \times P(A_3 = 2 \mid C_1)$$

$$= 0.5 \times 0.25 \times 0.25$$

$$= 0.03125$$

$$P(X \mid C_2) = P(A_1 = 1 \mid C_2) \times P(A_2 = 2 \mid C_2) \times P(A_3 = 2 \mid C_2)$$

$$= 0.33 \times 0.66 \times 0.66$$

$$= 0.14375$$

4. คำนวณหา $P(C_i \mid X) = P(X \mid C_i)P(C_i)$ สำหรับทุก i

$$P(C_1 \mid X) = P(X \mid C_1) \times P(C_1) = 0.03125 \times 0.5714 = 0.0179$$

$$P(C_2 \mid X) = P(X \mid C_2) \times P(C_2) = 0.14375 \times 0.4286 = 0.0616$$

5. จัดชุดข้อมูล X ให้อยู่ใน class ที่ 2 เนื่องจาก $P(C_2 \mid X)$ มีค่าสูงที่สุด

2.5 ต้นไม้ตัดสินใจ (Decision Tree)

ต้นไม้ตัดสินใจ (Decision Tree : DT) เป็นเครื่องมือที่ใช้ในการแยกแยะข้อมูลออกเป็น class ที่ทำได้ง่ายและรวดเร็ว โดยสร้างต้นไม้ตัดสินใจจากชุดข้อมูลฝึกฝน ซึ่งประกอบด้วยตัวแปร input และตัวแปร output ที่แสดง class ซึ่งเมื่อได้ต้นไม้ตัดสินใจแล้วจะต้องมีการทดสอบความถูกต้องด้วยชุดข้อมูลทดสอบ ถ้าผลที่ได้จากการใช้ข้อมูลในชุดทดสอบไม่ถูกต้อง จะต้องกลับไปสร้างต้นไม้ตัดสินใจใหม่

ถ้าถูกต้องแล้วสามารถนำต้นไม้ตัดสินใจดังกล่าวไปใช้กับข้อมูลอื่นๆได้ โดยสร้างเป็นกฎ (Rule) หรือเงื่อนไขตามเกณฑ์ที่ปรากฏจากต้นไม้ตัดสินใจ ซึ่งมีวิธีการสร้างต้นไม้ตัดสินใจ 3 วิธี คือ ID3, C5.0, CART

2.5.1. การสร้างต้นไม้ตัดสินใจด้วยวิธี ID3

สำหรับการสร้างต้นไม้ตัดสินใจด้วยวิธี ID3 เพื่อง่ายต่อความเข้าใจขอกำหนดสัญลักษณ์ที่ใช้ในการคำนวณดังนี้

D คือ ชุดข้อมูลที่ต้องการแบ่ง class

k คือ จำนวน class ของข้อมูลของตัวแปร output จากชุดข้อมูลฝึกฝน

p_i คือ ความน่าจะเป็นของข้อมูลของตัวแปร output จากชุดข้อมูลฝึกฝนที่อยู่ใน class ที่ i

s คือ จำนวนช่วงของข้อมูลที่แบ่งจาก ตัวแปร/ Attribute ที่เลือกใช้ ซึ่งจะได้ $D_1, D_2, ..., D_s$

$H(D)$ คือ ค่า Entropy ของข้อมูลที่ยังไม่ได้แบ่งช่วง

$H(D_i)$ คือ ค่า Entropy ของข้อมูลในช่วงที่ i

$p(D_i)$ คือ ความน่าจะเป็นของข้อมูลของตัวแปร input ที่ใช้ในการแบ่งข้อมูลจากชุดข้อมูลฝึกฝนที่อยู่ในช่วงที่ i

ขั้นตอนการสร้างต้นไม้การตัดสินใจ มีดังนี้

1. คำนวณหา Entropy ของข้อมูล โดยที่ Entropy $H(p_1, ..., p_k) = \sum_{i=1}^{k} p_i \log\left(\frac{1}{p_i}\right)$

2. เลือกตัวแปร/ Attribute จากตัวแปร input ที่จะนำมาใช้ในการแบ่งข้อมูล ซึ่งถือเป็น root node

3. แบ่งข้อมูลของตัวแปร/ Attribute ออกเป็นช่วงๆ
4. คำนวณหา Entropy ของข้อมูลแต่ละช่วง
5. คำนวณหา Gain ของ Entropy จาก

$$Gain\,(D, S) = H(D) - \sum_{i=1}^{S} p(D_i)H(D_i)$$

6. ทำขั้นตอน 1-4 สำหรับตัวแปร/ Attribute อื่นจนครบ ตัวแปร/ Attribute ที่ให้ Gain สูงสุด จะเป็นตัวแปรที่ดีที่สุดที่ใช้ในการแบ่งแยกข้อมูล
7. ตรวจสอบว่าจะต้องมีการแบ่ง class ต่อหรือไม่โดยพิจารณาช่วงของข้อมูลของตัวแปรที่ดีที่สุด ถ้ายังพบว่ามีช่วงใดช่วงหนึ่งของข้อมูลที่ยังมีข้อมูล output ต่างกัน จะต้องดำเนินการแบ่งข้อมูลในช่วงดังกล่าวย่อยลงไปอีกโดยทำซ้ำตั้งแต่ขั้นตอนที่ 2 จนกระทั่งไม่มีข้อมูลในช่วงใดที่ข้อมูล output ต่างกัน

ข้อสังเกต

ผลที่ได้จากการสร้างต้นไม้ตัดสินใจด้วยวิธี ID3 นี้ ตัวแปร/ Attribute ที่จะถูกนำมาใช้ในการตัดสินใจ มักเป็นตัวแปร/ Attribute ที่มีค่าหลายๆค่าซึ่งมักทำให้เกิดปัญหา overfitting

ตัวอย่างที่ 5 การสร้างต้นไม้ตัดสินใจด้วยวิธี ID3 ด้วยชุดข้อมูลฝึกฝนจำนวน 15 ตัวอย่าง ดังนี้

no.	gender	height(m)	class
1	F	1.6	short
2	M	2	tall
3	F	1.9	medium
4	F	1.88	medium
5	F	1.7	short
6	M	1.85	medium
7	F	1.6	short
8	M	1.7	short
9	M	2.2	tall
10	M	2.1	tall
11	F	1.8	medium
12	M	1.95	medium
13	F	1.9	medium
14	F	1.8	medium
15	F	1.75	medium

วิธีทำ จากชุดข้อมูลฝึกฝนซึ่งประกอบด้วยตัวแปร input 2 ตัว คือ ตัวแปร gender และตัวแปร height ส่วนตัวแปร output คือ ตัวแปร class ซึ่งประกอบด้วย 3 class คือ short medium และ tall จะได้ $k = 3$

ก่อนการแบ่งข้อมูลจะได้ว่า

1. คำนวณ Entropy ของข้อมูลเริ่มต้นก่อนการแบ่ง

   จากข้อมูลถ้าให้ class ที่ 1 คือ short

   class ที่ 2 คือ medium

   class ที่ 3 คือ tall

   จะได้

$$p_1 = \frac{\text{จำนวนของตัวอย่างที่อยู่ใน class ที่ 1}}{\text{จำนวนตัวอย่างทั้งหมด}} = \frac{4}{15}$$

$$p_2 = \frac{\text{จำนวนของตัวอย่างที่อยู่ใน class ที่ 2}}{\text{จำนวนตัวอย่างทั้งหมด}} = \frac{8}{15}$$

$$p_3 = \frac{\text{จำนวนของตัวอย่างที่อยู่ใน class ที่ 3}}{\text{จำนวนตัวอย่างทั้งหมด}} = \frac{3}{15}$$

ดังนั้น ค่า Entropy ของข้อมูลเริ่มต้นก่อนการแบ่ง

$$H(D) = H(p_1, p_2, p_3)$$

$$= \frac{4}{15}\log\left(\frac{15}{4}\right) + \frac{8}{15}\log\left(\frac{15}{8}\right) + \frac{3}{15}\log\left(\frac{15}{3}\right)$$

$$= 0.4385$$

2. เลือกตัวแปร/ Attribute จากตัวแปร input ที่นำมาใช้ในการแบ่งข้อมูล

กรณีที่ เลือกตัวแปร gender เป็นตัวแปร/ Attribute ที่ใช้ในการแบ่งข้อมูล แบ่งตัวแปร gender ออกเป็น 2 กลุ่ม $(s = 2)$ คือ female และ male แล้วหา Entropy สำหรับกลุ่ม female และ Entropy สำหรับกลุ่ม male ดังนี้

กลุ่ม female

$$p_1 = \frac{\text{จำนวนเพศหญิงใน class ที่ 1}}{\text{จำนวนเพศหญิงทั้งหมด}} = \frac{3}{9}$$

$$p_2 = \frac{\text{จำนวนเพศหญิงใน class ที่ 2}}{\text{จำนวนเพศหญิงทั้งหมด}} = \frac{6}{9}$$

$$p_3 = \frac{\text{จำนวนเพศหญิงใน class ที่ 3}}{\text{จำนวนเพศหญิงทั้งหมด}} = 0$$

Entropy สำหรับกลุ่ม female คือ

$$H(D_1) = \frac{3}{9}\log\left(\frac{9}{3}\right) + \frac{6}{9}\log\left(\frac{9}{6}\right)$$

$$= 0.2764$$

กลุ่ม male

$$p_1 = \frac{\text{จำนวนเพศชายใน class ที่ 1}}{\text{จำนวนเพศชายทั้งหมด}} = \frac{1}{6}$$

$$p_2 = \frac{\text{จำนวนเพศชายใน class ที่ 2}}{\text{จำนวนเพศชายทั้งหมด}} = \frac{2}{6}$$

$$p_3 = \frac{\text{จำนวนเพศชายใน class ที่ 3}}{\text{จำนวนเพศชายทั้งหมด}} = \frac{3}{6}$$

Entropy สำหรับกลุ่ม male คือ

$$H(D_2) = \frac{1}{6}\log\left(\frac{6}{1}\right) + \frac{2}{6}\log\left(\frac{6}{2}\right) + \frac{3}{6}\log\left(\frac{6}{3}\right)$$

$$= 0.4392$$

คำนวณหา $p(D_i)$ จาก

$$p(D_1) = \frac{\text{จำนวนเพศหญิง}}{\text{จำนวนตัวอย่างทั้งหมด}} = \frac{9}{15}$$

$$p(D_2) = \frac{\text{จำนวนเพศชาย}}{\text{จำนวนตัวอย่างทั้งหมด}} = \frac{6}{15}$$

ดังนั้นผลรวมของ Entropy จากการแบ่งข้อมูลด้วยตัวแปร gender คือ

$$\sum_{i=1}^{2} p(D_i)H(D_i) = \left(\frac{9}{15} \times 0.2764\right) + \left(\frac{6}{15} \times 0.4392\right)$$

$$= 0.3415$$

ซึ่งจะได้ Gain ของ Entropy คือ

$$Gain(D, 2) = H(D) - \sum_{i=1}^{2} p(D_2)H(D_i)$$

$$= 0.4385 - 0.3415$$

$$= 0.097$$

กรณีที่ เลือกตัวแปร height เป็นตัวแปร/ Attribute ในการแบ่งข้อมูล

เนื่องจากข้อมูลของตัวแปร height มีลักษณะเป็นค่าต่อเนื่อง จึงจำเป็นต้องแบ่งค่าความสูงออกเป็นช่วง

ถ้าแบ่งข้อมูลของตัวแปร height ออกเป็น 6 ช่วง $(s = 6)$ ดังนี้

$$(0, 1.6], (1.6, 1.7], (1.7, 1.8], (1.8, 1.9], (1.9, 2.0], (2.0, \infty]$$

คำนวณหา Entropy ของข้อมูลแต่ละช่วง สำหรับ $j = 1,..., 6$ และ $i = 1, 2, 3$ โดยที่

$$p(i) = \frac{\text{จำนวนข้อมูลช่วงที่ j ใน class ที่ i}}{\text{จำนวนข้อมูลทั้งหมดในช่วงที่ j}}$$

จะได้

Entropy ของข้อมูลช่วงที่ 1 คือ

$$H(D_1) = \frac{2}{2}\log\frac{2}{2} + 0 + 0$$

$$= 0$$

Entropy ของข้อมูลช่วงที่ 2 คือ

$$H(D_2) = \frac{2}{2}\log\frac{2}{2} + 0 + 0$$

$$= 0$$

Entropy ของข้อมูลช่วงที่ 3 คือ

$$H(D_3) = 0 + \frac{3}{3}\log\frac{3}{3} + 0$$

$$= 0$$

Entropy ของข้อมูลช่วงที่ 4 คือ

$$H(D_4) = 0 + \frac{4}{4}\log\frac{4}{4} + 0$$

$$= 0$$

Entropy ของข้อมูลช่วงที่ 5 คือ

$$H(D_5) = 0 + \frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}$$

$$= 0 + \left(\frac{1}{2}\times 0.301\right) + \left(\frac{1}{2}\times 0.301\right)$$

$$= 0.301$$

Entropy ของข้อมูลช่วงที่ 6 คือ

$$H(D_6) = 0 + 0 + \frac{2}{2}\log\frac{2}{2}$$

$$= 0$$

คำนวณหา $p(D_i)$ จาก

$$p(D_i) = \frac{\text{จำนวนข้อมูลในช่วงที่ i}}{\text{จำนวนข้อมูลทั้งหมด}}$$

จะได้

$$p(D_1) = \frac{2}{15}$$

$$p(D_2) = \frac{2}{15}$$

$$p(D_3) = \frac{3}{15}$$

$$p(D_4) = \frac{4}{15}$$

$$p(D_5) = \frac{2}{15}$$

$$p(D_6) = \frac{2}{15}$$

ดังนั้นผลรวมของ Entropy จากการแบ่งข้อมูลด้วยตัวแปร height คือ

$$\sum_{i=1}^{6} p(D_i)H(D_i) = \left(\frac{2}{15}\times 0\right)+\left(\frac{2}{15}\times 0\right)+\left(\frac{3}{15}\times 0\right)+\left(\frac{4}{15}\times 0\right)+\left(\frac{2}{15}\times 0.301\right)+\left(\frac{2}{15}\times 0\right)$$

$$= 0.04$$

ซึ่งจะได้ Gain ของ Entropy คือ

$$Gain(D, 6) = H(D) - \sum_{i=1}^{6} p(D_i)H(D_i)$$

$$= 0.4385 - 0.04$$

$$= 0.3985$$

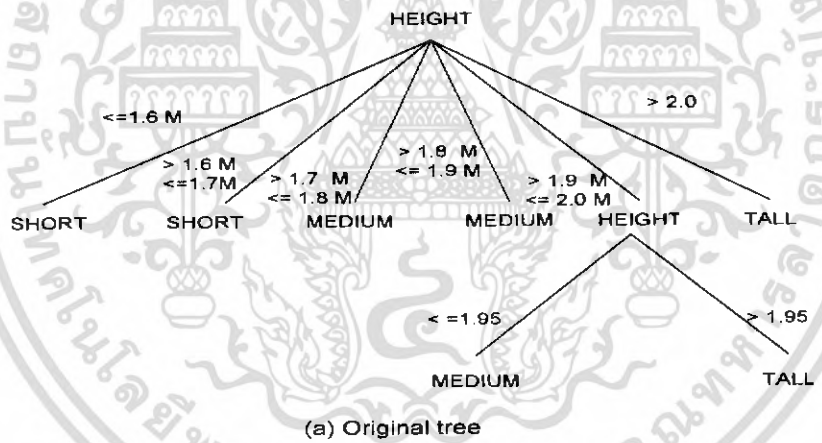
3. จาก Gain ของ Entropy ที่ได้พบว่า การใช้ตัวแปร height ในการแบ่งข้อมูล ให้ค่า Gain ของ Entropy สูงกว่า ดังนั้นในรอบที่ 1 นี้จะใช้ตัวแปร height ในการแบ่งข้อมูล
4. พิจารณาข้อมูลแต่ละช่วงของตัวแปร height ซึ่งถูกแบ่งเป็น

$$(0, 1.6], (1.6, 1.7], (1.7, 1.8], (1.8, 1.9], (1.9, 2.0], (2.0, \infty]$$

แล้วพบว่า ในช่วงที่ 5 คือช่วง (1.9, 2.0] นั้นพบว่า มี 2 ตัวอย่างที่อยู่ในช่วงนี้ ซึ่งค่าของตัวแปร class ต่างกัน คือ มีทั้ง medium และ tall ดังนี้

no.	gender	height(m)	class
2	M	2	tall
12	M	1.95	medium

ดังนั้น ในช่วง (1.9, 2.0] นี้จะต้องมีการแบ่งข้อมูลด้วยวิธีการเดิมต่อไป ซึ่งสุดท้ายจะได้



รูปที่ 5 ผลลัพธ์ที่ได้จากการจัดหมวดหมู่ด้วยวิธี ID3

2.5.2 การสร้างต้นไม้ตัดสินใจด้วยวิธี C5.0

วิธีการสร้างต้นไม้ตัดสินใจด้วยวิธี C5.0 นี้พัฒนามาจากวิธี C4.5 ซึ่งแนวคิดในการสร้างต้นไม้ เช่นเดียวกับ ID3 แต่เพิ่มบางขั้นตอนเพื่อให้ได้ต้นไม้ที่มีประสิทธิภาพดียิ่งขึ้น หรือในกรณีที่ไม่สามารถใช้วิธี ID3 สร้างได้ นอกจากนี้ยังแก้ปัญหา overfitting ได้โดยมีขั้นตอนดังนี้

1. ดำเนินการตามขั้นตอนเช่นเดียวกับ ID3 ตั้งแต่ขั้นตอนที่ 1 จนถึงขั้นตอนที่ 5 ซึ่งจะได้ $Gain(D, s)$
2. คำนวณหา $H(p(D_1),..., p(D_s))$
3. คำนวณหา Gain Ratio

$$Gain\ Ratio\ (D, s) = \frac{Gain\ (D, s)}{H(p(D_1),..., p(D_s))}$$

4. เลือก ตัวแปร/ Attribute ที่ให้ Gain Ratio สูงสุด เป็นตัวแปรที่ใช้ในการแบ่งข้อมูล
5. ตรวจสอบว่าควรจะแบ่งข้อมูลต่อไปหรือไม่ เช่นเดียวกับวิธี ID3

กล่าวโดยสรุปการสร้างต้นไม้ด้วยวิธี C5.0 นี้ มีวิธีการเช่นเดียวกับ ID3 แต่ใช้เกณฑ์ในการเลือกตัวแปร /Attribute ที่ใช้แบ่งข้อมูลต่างกันโดยวิธี C5.0 ใช้ $Gain\ Ratio\ (D, s)$ แต่วิธี ID3 ใช้ $Gain\ (D, s)$

ตัวอย่างที่ 6 การสร้างต้นไม้ตัดสินใจด้วยวิธี C5.0 โดยใช้ข้อมูลตัวอย่างที่ 5 โดยจะแสดงการหา Gain Ratio ถ้าใช้ตัวแปร gender ในการแบ่งข้อมูล

วิธีทำ จากตัวอย่างที่ 5 ถ้าใช้ตัวแปร gender ในการแบ่งข้อมูลจะได้

$$Gain\ (D, 2) = 0.097$$

คำนวณหา $H(p(D_1),..., p(D_s))$ ดังนี้

ในการใช้ตัวแปร gender ในการแบ่งข้อมูล $s = 2$ จะได้

$$H(p(D_1), p(D_2)) = H\left(\frac{9}{15}, \frac{6}{15}\right)$$

$$= \frac{9}{15}\log\left(\frac{15}{9}\right) + \frac{6}{15}\log\left(\frac{15}{6}\right)$$

$$= 0.292$$

คำนวณหา *Gain Ratio* ซึ่ง

$$Gain\ Ratio\ (D, 2) = \frac{Gain\ (D, 2)}{H(p(D_1), p(D_2))}$$

$$= \frac{0.097}{0.292}$$

$$= 0.332$$

2.5.3. การสร้างต้นไม้ตัดสินใจด้วยวิธี CART (Classification and Regression Tree)

การสร้างต้นไม้ตัดสินใจด้วยวิธี CART นั้นมีขั้นตอนเช่นเดียวกับ ID3 แต่ CART สร้างต้นไม้ตัดสินใจที่มีลักษณะ Binary Tree แต่ละ node จะแตกออกเป็น 2 กิ่ง คือ กิ่งทางด้านซ้าย และกิ่งทางด้านขวาเท่านั้น โดยเกณฑ์ที่ใช้ในการเลือกว่าตัวแปร/ Attribute ใดจะเป็นตัวในการแบ่งข้อมูล คือ

$$\phi(s|t) = 2p_L p_R \sum_{j=1}^{m} \left| p(c_j | t_L) - p(c_j | t_R) \right|$$

กำหนดให้

m คือ จำนวน class ของตัวแปร output จากชุดข้อมูลฝึกฝน

s คือ เกณฑ์ที่ใช้ในการแบ่งข้อมูล

t คือ node ปัจจุบัน

p_L, p_R คือ ความน่าจะเป็นของข้อมูลที่อยู่ทางซ้ายและขวาตามลำดับ

$p(c_j | t_L), p(c_j | t_R)$ คือ ความน่าจะเป็นของข้อมูลในclass c_j ที่อยู่ทางซ้ายและขวาตามลำดับ

เลือกตัวแปร/ Attribute ที่ให้ $\phi(s|t)$ สูงสุด เป็นตัวแปรในการแบ่งข้อมูล

ตัวอย่าง ที่ 7 การสร้างต้นไม้ตัดสินใจด้วยวิธี CART โดยใช้ข้อมูลตัวอย่างที่ 5

วิธีทำ จากตัวอย่างที่ 5

กรณีที่ 1 ถ้าเลือกตัวแปร height ในการแบ่งข้อมูล โดยกำหนดค่าที่ใช้ในการแบ่ง คือ 1.6, 1.7, 1.8, 1.9, 2.0 สามารถหาค่า $\phi(s \mid t)$ ได้ดังนี้

1. ใช้ค่า 1.6 เป็นจุดแบ่ง จะได้ข้อมูลดังนี้

class	height $<$ 1.6(m)	height $\geq$1.6(m)
short	0	4
medium	0	8
tall	0	3

จะได้ $p_L = 0, p_R = 1$

จาก $\phi(s \mid t) = 2p_L p_R \sum_{j=1}^{m} \left| p(c_j \mid t_L) - p(c_j \mid t_R) \right|$

ดังนั้น $\phi(1.6) = 0$

2. ใช้ค่า 1.7 เป็นจุดแบ่ง จะได้ข้อมูลดังนี้

class	height $<$ 1.7(m)	height $\geq$1.7(m)
short	2	2
medium	0	8
tall	0	3

จะได้ $p_L = \frac{2}{15}, p_R = \frac{13}{15}$

$$\sum_{j=1}^{3}\left| p(c_j \mid t_L) - p(c_j \mid t_P)\right|$$

$$= \left|\frac{2}{15} - \frac{2}{15}\right| + \left|0 - \frac{8}{15}\right| + \left|0 - \frac{3}{15}\right|$$

$$= 0 + \frac{8}{15} + \frac{3}{15}$$

$$= \frac{11}{15}$$

จาก $\phi(s \mid t) = 2p_L p_R \sum_{j=1}^{m}\left| p(c_j \mid t_L) - p(c_j \mid t_R)\right|$

ดังนั้น

$$\phi(1.7) = 2\left(\frac{2}{15}\right)\left(\frac{13}{15}\right)\left(\frac{11}{15}\right)$$

$$= 0.169$$

3. ใช้ค่า 1.8 เป็นจุดแบ่ง จะได้ข้อมูลดังนี้

class	height $<$ 1.8(m)	height $\geq$ 1.8(m)
short	4	0
medium	1	7
tall	0	3

จะได้ $p_L = \frac{5}{15}, p_R = \frac{10}{15}$

$$\sum_{j=1}^{3}\left| p(c_j \mid t_L) - p(c_j \mid t_P)\right|$$

$$= \left|\frac{4}{15} - 0\right| + \left|\frac{1}{15} - \frac{7}{15}\right| + \left|0 - \frac{3}{15}\right|$$

$$= \frac{4}{15} + \frac{6}{15} + \frac{3}{15}$$

$$= \frac{13}{15}$$

จาก $\phi(s \mid t) = 2p_L p_R \sum_{j=1}^{m}\left| p(c_j \mid t_L) - p(c_j \mid t_R)\right|$

ดังนั้น

$$\phi(1.8) = 2\left(\frac{5}{15}\right)\left(\frac{10}{15}\right)\left(\frac{13}{15}\right)$$

$$= 0.385$$

4. ใช้ค่า 1.9 เป็นจุดแบ่ง จะได้ข้อมูลดังนี้

class	height $<$ 1.9(m)	height $\geq$1.9(m)
short	4	0
medium	5	3
tall	0	3

จะได้ $p_L = \frac{9}{15}, p_R = \frac{6}{15}$

$$\sum_{j=1}^{3}\left| p(c_j \mid t_L) - p(c_j \mid t_P)\right|$$

$$= \left|\frac{4}{15} - 0\right| + \left|\frac{5}{15} - \frac{3}{15}\right| + \left|0 - \frac{3}{15}\right|$$

$$= \frac{4}{15} + \frac{2}{15} + \frac{3}{15}$$

$$= \frac{9}{15}$$

จาก $\phi(s \mid t) = 2p_L p_R \sum_{j=1}^{m}\left| p(c_j \mid t_L) - p(c_j \mid t_R)\right|$

ดังนั้น

$$\phi(1.9) = 2\left(\frac{9}{15}\right)\left(\frac{6}{15}\right)\left(\frac{9}{15}\right)$$

$$= 0.256$$

5. ใช้ค่า 2.0 เป็นจุดแบ่ง จะได้ข้อมูลดังนี้

class	height $<$ 2.0(m)	height $\geq$2.0(m)
short	4	0
medium	8	0
tall	0	3

จะได้ $p_L = \frac{12}{15}, p_R = \frac{3}{15}$

$$\sum_{j=1}^{3}\left| p(c_j \mid t_L) - p(c_j \mid t_P)\right|$$

$$= \left|\frac{4}{15} - \frac{0}{15}\right| + \left|\frac{8}{15} - 0\right| + \left|0 - \frac{3}{15}\right|$$

$$= \frac{4}{15} + \frac{8}{15} + \frac{3}{15}$$

$$= 1$$

จาก $\phi(s \mid t) = 2p_L p_R \sum_{j=1}^{m}\left| p(c_j \mid t_L) - p(c_j \mid t_R)\right|$

ดังนั้น

$$\phi(2.0) = 2\left(\frac{12}{15}\right)\left(\frac{3}{15}\right)(1)$$

$$= 0.32$$

กรณีที่ 2 ถ้าเลือกตัวแปร gender ในการแบ่งข้อมูล จะได้

class	female (คน)	male (คน)
short	3	1
medium	6	2
tall	0	3

จะได้ $p_L = \frac{9}{15}, p_R = \frac{6}{15}$

$$\sum_{j=1}^{3}\left| p(c_j \mid t_L) - p(c_j \mid t_P)\right|$$

$$= \left|\frac{3}{15}-\frac{1}{15}\right|+\left|\frac{6}{15}-\frac{2}{15}\right|+\left|0-\frac{3}{15}\right|$$

$$= \frac{2}{15}+\frac{4}{15}+\frac{3}{15}$$

$$= \frac{9}{15}$$

จาก $\phi(s \mid t) = 2p_L p_R \sum_{j=1}^{m}\left| p(c_j \mid t_L) - p(c_j \mid t_R)\right|$

ดังนั้น

$$\phi(gender) = 2\left(\frac{9}{15}\right)\left(\frac{6}{15}\right)\left(\frac{9}{15}\right)$$

$$= 0.224$$

จากเกณฑ์ที่คำนวณได้ทั้งหมดพบว่า $\phi(1.8)$ มีค่าสูงสุดคือ 0.385 ดังนั้นจะใช้ 1.8 เป็น root node แยกออกเป็นกิ่งซ้ายและกิ่งขวา และดำเนินการต่อไปจนได้ต้นไม้ที่สมบูรณ์

3. Unsupervised Learning

Unsupervised Learning เป็นกระบวนการเรียนรู้เพื่อหารูปแบบที่เหมาะสมว่า output ควรเป็นอย่างไร จากลักษณะของข้อมูลในชุดข้อมูลฝึกฝนซึ่งมีเพียงข้อมูล input เท่านั้น

3.1 การจัดกลุ่ม (Clustering)

Clustering เป็นกระบวนการจัดกลุ่มของข้อมูลโดยอาศัยความคล้ายคลึงกันในลักษณะเฉพาะบางอย่างของข้อมูลซึ่งภายใน cluster เดียวกันจะมีลักษณะข้อมูลที่มีความคล้ายคลึงกันมาก และระหว่าง cluster จะมีความแตกต่างกันมาก ในปัจจุบันเราสามารถนำการจับกลุ่มแบบ cluster มาประยุกต์ใช้ได้มากมายหลายสาขา เช่น การจำแนกประเภทของพืชและสัตว์, การจำแนกกลุ่มของเชื้อโรค, ด้านชีววิทยา, ด้านเศรษฐศาสตร์ ฯลฯ

จากนิยามกำหนดให้ $D=\{t_1,t_2,...,t_n\}$ เป็นเซตของฐานข้อมูล โดยที่ k เป็นจำนวนกลุ่มของ cluster ที่ต้องการจัดกลุ่ม และ t_i เป็นตัวอย่างที่ i ในฐานข้อมูล (D) ซึ่งแต่ละ t_j จะถูกจัดให้อยู่ใน cluster ใดๆ (k_j) ด้วยเทคนิคต่างๆทาง cluster

จากที่กล่าวมาข้างต้นเราจะกำหนดให้ ค่าที่ใช้ในการวัดความคล้ายคลึงกันคือ $sim(t_i,t_j)$ ซึ่งสิ่งที่ใช้ในการบอกถึงความคล้ายคลึงกันของตัวอย่าง 2 ตัวอย่าง หาได้จากการวัดระยะห่าง $dis(t_i,t_j)$ โดยที่ t_i จะเป็นสมาชิกของ (k_j) เมื่อ $dis(t_{jl},t_i)\le dis(t_{jl},t_{jm})$ เมื่อ t_{jL},t_{jm} เป็นสมาชิกของ cluster k_j

ในการวัดระยะทาง $(dis(t_i,t_j))$ นั้นเราอาจจะใช้จุดศูนย์กลาง, รัศมี หรือเส้นผ่านศูนย์กลางเป็นสิ่งที่บ่งบอกถึงลักษณะเฉพาะในแต่ละ cluster สำหรับระยะห่างระหว่าง ตัวอย่าง 2 ตัวอย่างสามารถคำนวณได้ดังนี้

$$\text{จุดศูนย์กลาง } (C_m) = \frac{\sum_{i=1}^{N} t_{mi}}{N}$$

$$\text{รัศมี } (R_m) = \sqrt{\frac{\sum_{i=1}^{N}(t_{mi}-C_m)^2}{N}}$$

$$\text{เส้นผ่านศูนย์กลาง } (D_m) = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(t_{mi}-t_{mj})^2}{N(N+1)}}$$

โดยค่าจุดศูนย์กลาง (centroid) คือ จุดกึ่งกลางของ cluster เป็นค่าที่ไม่จำเป็นต้องมีอยู่จริงใน cluster นั้น แต่เราสามารถนำมาใช้ในการคำนวณระยะห่างระหว่าง cluster สอง cluster โดยที่ $dis(K_i, K_j) = dis(C_i, C_j)$ จาก Clustering Algorithms

Outlier คือ ค่าที่แตกต่างจากค่าที่อยู่ในกลุ่มข้อมูลและอาจเป็นความผิดพลาดในข้อมูล หรือ Clustering Algorithms อาจค้นพบและนำออกไป เพื่อที่จะทำให้ได้ผลที่ดีมากขึ้น ซึ่งการตัดข้อมูลนั้นออกไป ทำให้เกิดความคลาดเคลื่อนของข้อมูลซึ่งอาจส่งผลให้มีประสิทธิภาพลดลง ดังนั้นจึงควรนำเทคนิคทางสถิติมาวิเคราะห์ว่าจะเก็บ outlier ไว้หรือจะตัดทิ้ง เช่น เทคนิคการวัดระยะทาง (Distance Measure)

ข้อแตกต่างระหว่างการจำแนกประเภท (Classification) กับ Clustering

1. ไม่ทราบจำนวน cluster ที่ดีที่สุดที่ใช้ในการจัดกลุ่ม cluster
2. ไม่มีการกำหนด class ก่อนที่จะทำการจัดกลุ่ม cluster
3. ผลของ cluster จะมีการเปลี่ยนแปลงเมื่อข้อมูลในฐานข้อมูลมีการเคลื่อนไหว

ปัญหาของ Clustering

1. มีข้อมูลที่ไม่สามารถจัดกลุ่ม cluster ได้ต้องอยู่อย่างโดดเดี่ยว
2. เมื่อข้อมูลมีความเคลื่อนไหวอาจทำให้สมาชิกใน cluster เปลี่ยนแปลงได้
3. การตีความหมายของแต่ละ cluster อาจทำได้ยากเพราะผลลัพธ์ที่ได้จะไม่มีความชัดเจน จึงจำเป็นต้องขอคำปรึกษาจากผู้เชี่ยวชาญ
4. ผลลัพธ์ที่ได้อาจมีผลลัพธ์ที่ถูกต้องมากกว่า 1 ผลลัพธ์จึงยากที่จะกำหนดจำนวนของ cluster ได้
5. ไม่ทราบล่วงหน้า ถึง class ที่จะนำมาทำ cluster

รูปแบบของ Clustering Algorithms

รูปแบบการจำแนกกลุ่มโดย Clustering Algorithm สามารถแบ่งได้เป็น 2 แบบ ดังนี้

1. Hierarchical Algorithms

Hierarchical Algorithms เป็นการจำแนก cluster โดยกำหนดให้ชั้นล่างสุดของ tree จะมี 1 item เป็น 1 cluster ส่วนชั้นบนสุดของ tree จะมี item ทั้งหมดรวมอยู่ใน cluster เดียวกัน โดยที่ Algorithms นี้จะไม่มีการกำหนดจำนวนของ cluster ที่ต้องการ ซึ่งจะส่งผลให้มีจำนวน cluster มากกว่า 1 รูปแบบ

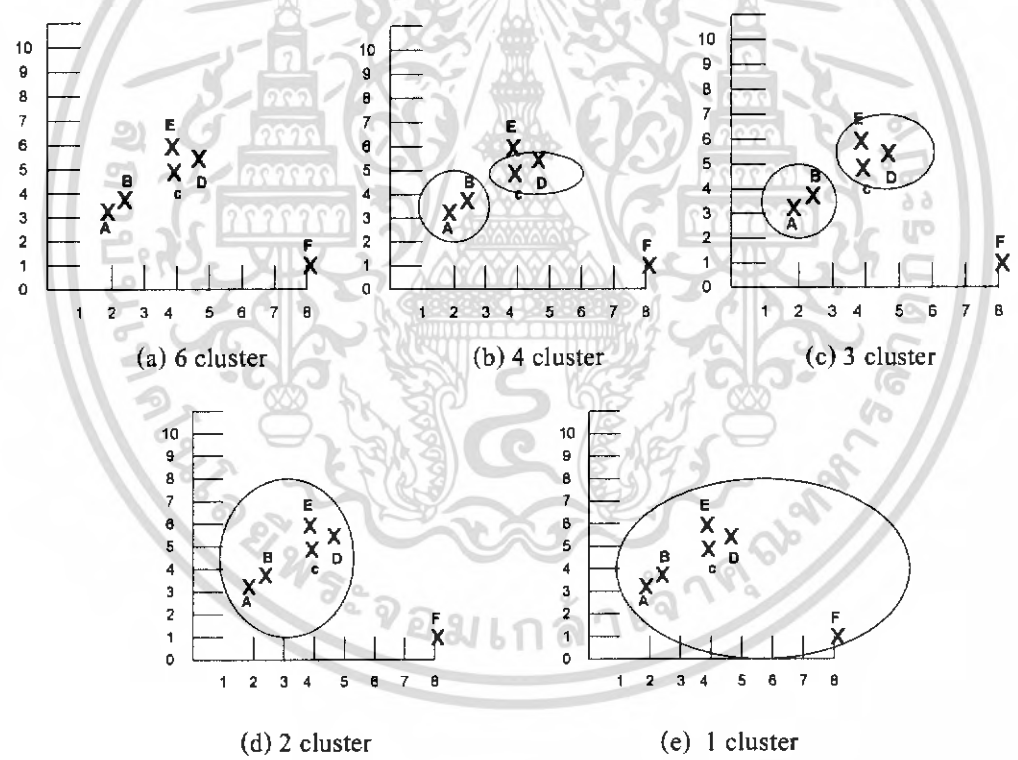
2. Partitional Algorithms

Partitional Algorithms เป็นการจำแนก cluster ด้วยเทคนิคต่างๆ เพื่อให้ได้ cluster ด้วยเทคนิคต่างๆ เพื่อให้ได้ cluster เท่ากับจำนวน cluster ที่เรากำหนดไว้ ซึ่งเป็นการกำหนดให้จำนวนรูปแบบของ cluster มีเพียง 1 รูปแบบเท่านั้น

3.1.1 Hierarchical Algorithms

Hierarchical Algorithms เป็นการจำแนก cluster ของโครงสร้างข้อมูลแบบ tree หรือที่เรียกว่า Dendrogram ซึ่ง cluster ที่จะเกิดขึ้นมาใหม่นั้นจะเกิดจากการรวมตัวของ cluster ย่อย โดยพิจารณาจาก ระยะห่างระหว่าง cluster โดยที่ระยะห่างระหว่างสมาชิกใน cluster นั้นกับสมาชิกในอีก cluster หนึ่งดังตัวอย่างที่ 8

ตัวอย่างที่ 8 แสดงแนวคิดของการทำ Hierarchical Algorithms



รูปที่ 6 แนวคิดของการทำ Hierarchical Algorithms

จากรูปที่ 6 แสดงให้เห็นว่าในฐานข้อมูลมีสมาชิก 6 ตัว คือ $\{A, B, C, D, E, F\}$ ซึ่งแสดงถึงการจำแนก cluster แบบ Hierarchical Algorithms ดังต่อไปนี้

1. กำหนดให้ cluster แต่ละตัวประกอบด้วยสมาชิกเพียงตัวเดียวซึ่งจะได้ cluster ทั้งหมด 6 cluster ดังรูป 6-(a)
2. จัด cluster โดยมีเซตของ cluster ที่มีสมาชิก 2 ตัวจำนวน 2 เซต เนื่องจากสมาชิก 2 ตัวนี้ ใกล้ชิดกันมากกว่าสมาชิกตัวอื่นๆ ดังรูป 6-(b)
3. จัด cluster อีกครั้งจะได้ cluster ใหม่ที่เกิดจากการรวมสมาชิกที่ใกล้กับ cluster หนึ่งที่มีสมาชิก 2 ตัว ดังรูป 6-(c)
4. จัด cluster ที่มีสมาชิก 2 ตัวและ cluster ที่มีสมาชิก 3 ตัวเพื่อทำให้เกิด cluster ที่มีสมาชิก 5 ตัว เนื่องจากข้อมูลมีความใกล้เคียงกันมากกว่าที่จะใกล้ชิดกับ cluster อื่นๆ ดังรูป 6-(d)
5. รวม cluster ทั้งหมดเข้าไว้ด้วยกันทำให้ cluster ดังกล่าวมีสมาชิกทั้งหมด 6 ตัว ดังรูป 6-(e)

โดยทั่วไปการจำแนก cluster โดยใช้เทคนิคแบบ Hierarchical Algorithms จะสามารถแสดงความสัมพันธ์กันระหว่าง cluster ซึ่งสามารถแบ่ง Hierarchical Algorithms ออกได้เป็น 2 เทคนิคย่อยๆ คือ

1. Agglomerative Algorithms

Agglomerative Algorithms เป็นการกำหนดให้ 1 item เป็น 1 cluster จากนั้นทำการรวมกันเป็น cluster ที่มีขนาดใหญ่ขึ้น จนกระทั่ง cluster ทั้งหมดรวมกันเป็น cluster เดียว

ในการรวม cluster นั้น เราจะกำหนดให้เซตของสมาชิกและระยะห่างระหว่างสมาชิกเป็นตัวแปร input ส่วนตัวแปรผลลัพธ์ ได้แก่ dendrogram ซึ่งเขียนอยู่ในรูป (d, k, K) โดยที่ d เป็นระยะห่างเริ่มต้นที่เรากำหนดไว้, k เป็นจำนวนกลุ่มของ cluster และ K เป็นเซตของ cluster

ปัญหาที่พบในวิธีการ Agglomerative คือ เมื่อมีการเปลี่ยนแปลงสมาชิกในฐานข้อมูลจะต้องดำเนินการจัดกลุ่ม cluster ใหม่

เทคนิคที่นิยมใน Agglomerative Algorithm มีอยู่ 3 เทคนิค คือ

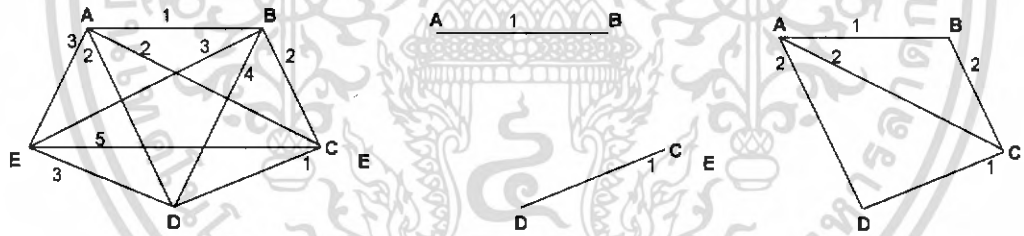
1.1 Single Link Technique (SLT)

Single Link Algorithm เป็นเทคนิคที่ใช้ในการจัดกลุ่ม cluster โดยกำหนดให้ 1 item เป็น 1 cluster จากนั้นรวม cluster ที่มีระยะห่างระหว่าง cluster น้อยที่สุดเข้าด้วยกันก่อนแล้วจึงทำการรวม cluster ที่มีระยะห่างระหว่าง cluster มากขึ้นตามลำดับ ซึ่งโดยทั่วไปจะใช้เทคนิคMinimum

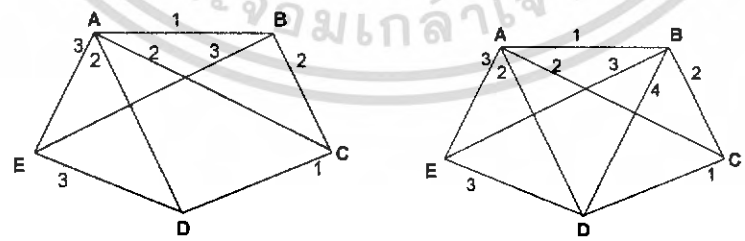
Spanning Tree (MST) มาช่วยในการหาระยะห่างระหว่าง cluster ทำให้ได้ระยะทางที่สั้นที่สุดที่ใช้ในการเชื่อม cluster ทุก cluster โดยเราจะเรียกระยะห่างระหว่าง cluster สอง cluster ที่ได้จาก MST ว่า กิ่ง จากนั้นเราจะทำการรวม cluster สอง cluster เข้าด้วยกัน โดยพิจารณาจากกิ่งที่มีระยะห่างระหว่าง cluster น้อยที่สุดก่อนและทำเช่นนี้ไปเรื่อยๆตามลำดับของระยะห่างระหว่าง cluster จนกระทั่งทุก items ในฐานข้อมูลถูกจัดอยู่ภายใน cluster เดียวกัน และระยะห่างจากจุดเริ่มต้นในการรวม cluster แต่ละครั้งมีค่าเท่ากับระยะห่างระหว่าง cluster ที่เรานำมารวมกัน

ตัวอย่างที่ 9 การจัดกลุ่ม cluster แบบ Hierarchical Algorithms ด้วยเทคนิค Single Link Technique จากข้อมูลซึ่งแสดงระยะห่างระหว่างสมาชิกต่อไปนี้

item	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0

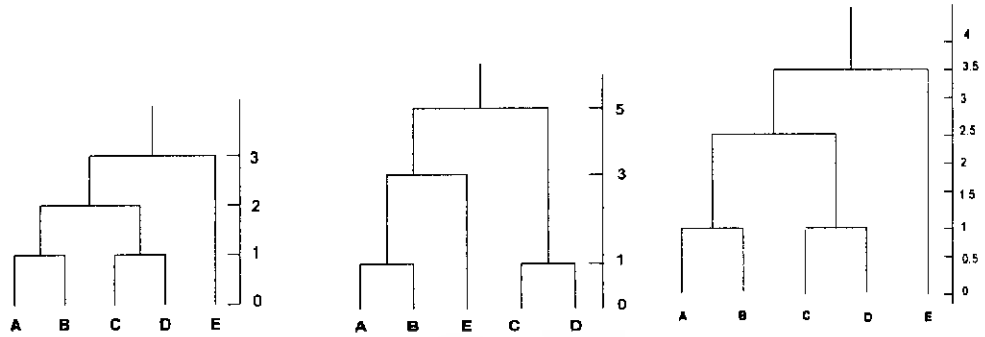


(a) กราฟกับระยะทางทั้งหมด (b) ระยะห่างระหว่าง cluster ≤ 1 (c)ระยะห่างระหว่าง cluster ≤ 2



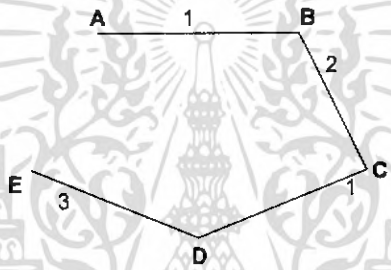
(d) ระยะห่างระหว่าง cluster ≤ 3 (e) ระยะห่างระหว่าง cluster ≤ 4

รูปที่ 7 กราฟระยะห่างระหว่าง cluster



(a) Single link (b) Complete link (c) Average link

รูปที่ 8 tree หรือ dendrogram

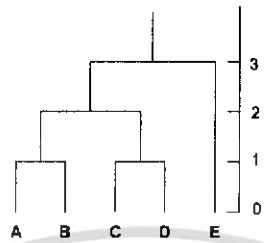


รูปที่ 9 ระยะห่างระหว่าง cluster ที่ได้จากเทคนิค MST

วิธีทำ

1. กำหนดให้ 1 item เป็น 1 cluster แล้วใช้เทคนิค Minimum Spanning Tree ในการหาเส้นทางที่สั้นที่สุดในการเชื่อมต่อ cluster ทุก cluster เพื่อหากิ่งที่มีระยะห่างระหว่าง cluster น้อยที่สุดในการเชื่อม cluster สอง cluster ดังรูปที่ 9 จากนั้นรวม cluster ที่มีระยะห่างระหว่าง cluster ทั้งสองน้อยที่สุด จากรูปที่ 9 เราจะเห็นได้ว่า ระยะห่างระหว่าง A กับ B และ C กับ D มีค่าน้อยที่สุดจึงรวม cluster ดังกล่าวทำให้เกิด cluster 3 cluster ดังนี้ $\{A, B\}, \{C, D\}, \{E\}$ และมีระยะห่างจากจุดเริ่มต้นเป็น 1 $(d = 1)$
2. เลือกกิ่งที่มีระยะห่างระหว่าง cluster มากขึ้นเป็นลำดับถัดมา ซึ่งเป็นกิ่งระหว่าง cluster $\{A, B\}$ และ $\{C, D\}$ จึงรวม cluster ทั้งสองเข้าด้วยกันซึ่งทำให้เกิด cluster ใหม่ทั้งหมด 2 cluster คือ $\{A, B, C, D\}, \{E\}$ และมีระยะห่างจากจุดเริ่มต้นเป็น 2 $(d = 2)$

3. เลือกกิ่งที่มีระยะห่างมากขึ้นเป็นลำดับถัดมาซึ่ง เป็นกิ่งระหว่าง cluster $\{A, B, C, D\}$ กับ $\{E\}$ จึงรวม cluster ทั้งสองเข้าด้วยกันเป็น 1 cluster ซึ่งมีระยะห่างจากจุดเริ่มต้นเป็น 3 $(d = 3)$ และสามารถเขียน dendrogram ได้ดังรูปที่ 10



รูปที่ 10 Dendrogram การจัด cluster ด้วยวิธี Single Link Technique

จาก cluster ที่ได้ในการทำขั้นสุดท้ายจะเห็นได้ว่าทุก item ในฐานข้อมูลถูกจัดรวมอยู่ใน cluster เดียวจึงเป็นการสิ้นสุดการจัดกลุ่ม cluster แบบ Hierarchical Algorithms ด้วยวิธี Single Link Technique

ถึงแม้ว่า Single Link Algorithm จะเป็นวิธีที่ง่ายแต่ก็มีปัญหาคือ ไม่ค่อยมีประสิทธิภาพและเกิด cluster ที่มีห่วงโซ่ยาวเกินไป

1.2 Complete Link Technique (CLT)

Complete Link Algorithm เป็นเทคนิคที่คล้ายกับ Single Link Algorithm แต่จะมีการจัดหมวดหมู่ที่ชัดเจนและมีความน่าเชื่อถือมากกว่า ซึ่งเรียกส่วนที่เชื่อมต่อกันว่า clique โดยเทคนิคนี้จะเริ่มจากการรวม cluster สอง cluster ที่มีระยะห่างน้อยที่สุดก่อน จากนั้นจึงทำการรวม cluster ที่มีระยะห่างมากที่สุดเพื่อจัด cluster ในระดับถัดไป

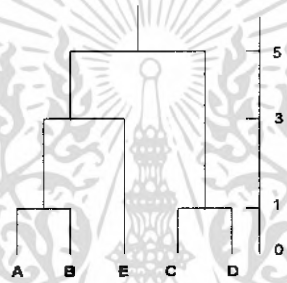
ตัวอย่างที่ 10 การจัดกลุ่ม cluster แบบ Hierarchical Algorithms ด้วยเทคนิค Complete Link Algorithm โดยใช้ข้อมูลในตัวอย่างที่ 9

วิธีทำ กำหนดให้ 1 item เป็น 1 cluster เช่นเดียวกับการใช้เทคนิค Single Link Technique แล้วใช้เทคนิค Minimum Spanning Tree (MST) ในการหาเส้นทางที่สั้นที่สุดที่เชื่อมต่อ cluster ทุก cluster จากนั้นทำการจัด cluster ดังต่อไปนี้

1. เลือก cluster ที่มีระยะห่างระหว่าง cluster ที่ได้จากเทคนิค Minimum Spanning Tree ที่มีค่าน้อยที่สุด รวมเข้าเป็น cluster เดียวกัน จากรูปที่ 4 เราจะเห็นได้ว่า ระยะห่างระหว่าง cluster A กับ B และ cluster C กับ D มีค่าน้อยที่สุดจึงทำการรวม cluster

ดังกล่าว ทำให้เกิด cluster 3 cluster ดังนี้ { A , B} , { C , D} , { E } และมีระยะห่างจากจุดเริ่มต้นเป็น1 (d = 1)

2. เลือก cluster ที่ได้จากข้อที่ 1ที่มีระยะห่างระหว่าง cluster มากที่สุดมารวมเข้าด้วยกัน จะเห็นได้ว่า cluster {A , B} กับ cluster {E} มีระยะห่างระหว่าง cluster มากที่สุด จึงทำการรวม cluster ทั้ง 2 เข้าด้วยกันจะได้ cluster ทั้งหมด 2 cluster คือ {A , B , E } และ { C , D } ซึ่งมีระยะห่างระหว่าง cluster เป็น 3
3. เลือก cluster ที่ได้จากข้อที่ 2 ที่มีระยะห่างระหว่าง cluster มากที่สุดคือ cluster {A , B , E } และ { C , D } แล้วรวม cluster ทั้ง 2 เข้าด้วยกัน จะได้ cluster เพียง cluster เดียว คือ{ A , B , E , C , D } ซึ่งมีระยะห่างระหว่าง cluster เป็น 5 และสามารถเขียน Dendrogram ได้ดังรูปที่ 11



รูปที่ 11 Dendrogram การจัด cluster ด้วยวิธี Complete Link Technique

จากในข้อที่ 3 จะเห็นได้ว่าทุก item ในฐานข้อมูลจะรวมอยู่ภายใน cluster เดียวกัน จึงเป็นการสิ้นสุดการจัดกลุ่มแบบ Hierarchical Algorithms ด้วยวิธี Complete Link Technique

1.3 Average Link Technique (ALT)

Average Link Algorithm เป็นกระบวนการจัดกลุ่ม cluster ทีละสอง cluster โดยเริ่มจากกำหนดให้ 1 items เป็น 1 cluster จากนั้นจะพิจารณาค่าเฉลี่ย ถ้าค่าเฉลี่ยของระยะห่างที่เชื่อมต่อระหว่าง cluster สอง cluster ที่เป็นไปได้ มีค่าน้อยกว่าหรือเท่ากับระยะห่างเริ่มต้นจะรวม cluster ดังกล่าวเข้าด้วยกันเป็น cluster ใหม่ 1 cluster เมื่อตรวจครบทุกคู่แล้วจะทำการเพิ่มระยะห่างจากจุดเริ่มต้นเป็น $d + x$ เมื่อ x คือค่าระยะห่างที่จะเพิ่มขึ้นในการทำซ้ำแต่ละครั้ง จากนั้นจัดกลุ่ม cluster ที่ได้จากการรวม cluster ในขั้นก่อนหน้านั้นด้วยวิธีการเดิม จนกระทั่ง item ทุก itemในฐานข้อมูลถูกรวมเป็น cluster เดียวกัน

ตัวอย่างที่ 11 การจัดกลุ่ม cluster แบบ Hierarchical Algorithms ด้วยวิธี Average Link Algorithm โดยใช้ข้อมูลในตัวอย่างที่ 9 โดยในการทำซ้ำแต่ละครั้งกำหนดให้ระยะห่างจากจุดเริ่มต้นเพิ่มขึ้นครั้งละ 0.5

วิธีทำ

ในครั้งที่ 1 กำหนดให้ 1 item เป็น 1 cluster จะได้ว่า

$K_1 = \{A\}, K_2 = \{B\}, K_3 = \{C\}, K_4 = \{D\}, K_5 = \{E\}$ แล้วจัดกลุ่ม cluster ดังต่อไปนี้

รอบที่ 1 :

ขั้นที่ 1 จับคู่ cluster ทุก cluster ที่เป็นไปได้ แล้วทำการหาค่าเฉลี่ยของระยะห่างระหว่าง 2 cluster จากระยะห่างทั้งหมดที่เชื่อม cluster สอง cluster ดังตารางที่ 1

ขั้นที่ 2 เปรียบเทียบค่าเฉลี่ยระยะห่างของ cluster แต่ละคู่ว่า มีค่าน้อยกว่าหรือเท่ากับระยะห่างจากจุดเริ่มต้น คือ 0.5 $(d = 0.5)$ หรือไม่ ถ้ามีค่าน้อยกว่าก็จะรวม cluster ทั้งสองเข้าด้วยกัน

<u>ตารางที่ 1</u> การจับคู่ cluster ทุกคู่ที่เป็นไปได้ในการทำซ้ำครั้งที่ 1, 2 และค่าเฉลี่ยระยะห่างระหว่าง cluster แต่ละคู่

cluster แต่ละคู่	Average Link Algorithm
K_1, K_2	$1/1 = 1$
K_1, K_3	$2/1 = 2$
K_1, K_4	$2/1 = 2$
K_1, K_5	$3/1 = 3$
K_2, K_3	$2/1 = 2$
K_2, K_4	$4/1 = 4$
K_2, K_5	$3/1 = 3$
K_3, K_4	$1/1 = 1$
K_3, K_5	$5/1 = 5$
K_4, K_5	$3/1 = 3$

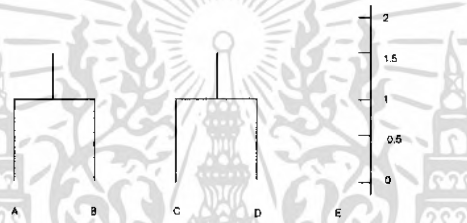
จากตารางที่ 1 จะเห็นได้ว่าไม่มี cluster คู่ใดที่มีค่าเฉลี่ยระยะห่างระหว่าง cluster น้อยกว่าหรือเท่ากับ 0.5 จึงไม่มีการรวม cluster เกิดขึ้น จากนั้นเราจึงเพิ่มระยะห่างจากจุดเริ่มต้นอีก 0.5 จะได้ค่าระยะห่างระหว่างจากจุดเริ่มต้นเป็น 1.0 $(d = 1.0)$ ในการทำซ้ำครั้งที่ 2

รอบที่ 2 :

ขั้นที่ 1 เนื่องจากในการทำซ้ำครั้งที่ 1 ไม่เกิดการรวมกันของ cluster จึงส่งผลให้กระบวนการในขั้นที่ 1 ได้ผลเช่นเดียวกับตารางที่ 1

ขั้นที่ 2 เปรียบเทียบระยะห่างเฉลี่ยระหว่าง cluster แต่ละคู่ว่ามีค่าน้อยกว่าหรือเท่ากับระยะห่างจากจุดเริ่มต้น $(d = 1.0)$ หรือไม่ ถ้ามีค่าน้อยกว่าก็จะทำการรวม cluster ทั้งสองเข้าด้วยกัน

จากตารางที่ 1 จะเห็นได้ว่ามีค่าเฉลี่ยระยะห่างระหว่าง cluster สอง cluster จำนวน 2 คู่ ที่มีค่าน้อยกว่าหรือเท่ากับ 1.0 คือ K_1 กับ K_2 และ K_3 กับ K_4 จึงรวม cluster ทั้งสอง ในแต่ละคู่เข้าด้วยกันจะได้ cluster ใหม่ 3 cluster คือ $K_1 = \{A, B\}, K_2 = \{C, D\}, K_3 = \{E\}$ ดังรูปที่ 12 จากนั้นจึงเพิ่มระยะห่าง จากจุดเริ่มต้นอีก 0.5 ซึ่งจะมีค่าเป็น 1.5 ในการทำซ้ำครั้งที่ 3



รูปที่ 12 การรวม cluster ในการทำซ้ำครั้งที่ 2

รอบที่ 3 :

ขั้นที่ 1 จับคู่ cluster ทุก cluster ที่เป็นไปได้จาก cluster ใหม่ที่ได้จากการจัดกลุ่มในการทำซ้ำครั้งที่ 2 แล้วหาค่าเฉลี่ยระยะห่างระหว่างสอง cluster จากระยะห่างทั้งหมดที่เชื่อมต่อ cluster สอง cluster ดังตารางที่ 2

ขั้นที่ 2 เปรียบเทียบค่าเฉลี่ยระยะห่างระหว่าง cluster แต่ละคู่ว่ามีค่าน้อยกว่าหรือเท่ากับระยะห่างจากจุดเริ่มต้น $(d = 1.5)$ หรือไม่ ถ้ามีค่าน้อยกว่าหรือเท่ากับ ก็จะต้องรวม cluster ทั้งสองเข้าด้วยกัน

ตารางที่ 2 การจับคู่ cluster ทุกคู่ที่เป็นไปได้ ในการทำซ้ำครั้งที่ 3 , 4 , 5 และ ค่าเฉลี่ยระหว่าง cluster แต่ละคู่

cluster แต่ละคู่	Average Link Algorithm
K_1, K_2	$\frac{2+2+2+4}{4} = \frac{10}{4} = 2.5$
K_1, K_3	$\frac{3+3}{2} = \frac{6}{2} = 3$
K_2, K_3	$\frac{5+3}{2} = \frac{8}{2} = 4$

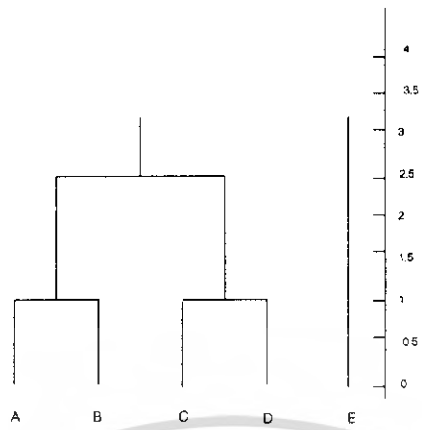
จากตารางที่ 2 จะเห็นได้ว่าไม่มี cluster คู่ใดที่มีค่าเฉลี่ยระยะห่างระหว่าง cluster น้อยกว่าหรือกับระยะห่างจากจุดเริ่มต้น $(d = 1.5)$ จึงไม่มีการรวม cluster เกิดขึ้น จากนั้นจึงทำการเพิ่มระยะห่างจากจุดเริ่มต้นเป็น 2.0 ในการทำซ้ำครั้งที่ 4 แต่ก็ยังไม่มี cluster คู่ใด ที่มีค่าเฉลี่ยระยะห่างระหว่าง cluster น้อยกว่าหรือเท่ากับ 2.0 เราจึงทำการเพิ่มระยะห่างจากจุดเริ่มเป็น 2.5 ในการทำซ้ำครั้งที่ 5

รอบที่ 5 :

ขั้นที่ 1 เนื่องจากในการทำซ้ำครั้งที่ 3, 4 ไม่เกิดการรวมกันของ cluster จึงส่งผลให้กระบวนการในขั้นที่ 1 ได้ผลตามตารางที่ 2 ดังเดิม

ขั้นที่ 2 เปรียบเทียบระยะห่างเฉลี่ยระหว่าง cluster แต่ละคู่มีค่าน้อยกว่าหรือเท่ากับระยะห่างจากจุดเริ่มต้น $(d = 2.5)$ หรือไม่น้อยกว่าหรือเท่ากับ ก็จะรวม cluster ดังกล่าวเข้าด้วยกัน

จากตารางที่ 2 จะเห็นได้ว่า มีค่าเฉลี่ยระยะห่างระหว่าง cluster สอง cluster จำนวน 1 คู่ ที่มีค่าน้อยกว่าหรือเท่ากับ 2.5 คือ K_1, K_2 จึงทำการรวม cluster และได้ cluster ใหม่ 3 cluster คือ $K_1 = \{A, B, C, D\}, K_2 = \{E\}$ ดังรูปที่ 13 จากนั้นเพิ่มระยะห่างจากจุดเริ่มต้นเป็น 3.0 ในการทำซ้ำ ครั้งที่ 6



รูปที่ 13 การรวม cluster ในการทำซ้ำครั้งที่ 6

รอบที่ 6 :

ขั้นที่ 1 จับคู่ cluster ทุก cluster ที่เป็นไปได้ จาก cluster ใหม่ที่ได้ จากการจัดกลุ่มในการทำซ้ำครั้งที่ 5 แล้วหาค่าเฉลี่ยระยะห่างระหว่าง 2 cluster จากระยะห่างทั้งหมดที่เชื่อมต่อ cluster สอง cluster ดังตารางที่ 3

ขั้นที่ 2 เปรียบเทียบค่าเฉลี่ย ระยะห่างระหว่าง cluster แต่ละคู่ว่ามีค่าน้อยกว่าหรือเท่ากับระยะห่างจากจุดเริ่มต้น $(d = 3.0)$ หรือไม่ ถ้ามีค่าน้อยกว่าหรือเท่ากับ ก็จะรวม cluster ทั้งสอง เข้าด้วยกัน

ตารางที่ 3 การจับคู่ cluster ทุกคู่ที่เป็นไปได้ ในการทำซ้ำครั้งที่ 6, 7 และค่าเฉลี่ยระหว่าง cluster แต่ละคู่

cluster แต่ละคู่	Average Link Algorithm
K_1, K_2	$\frac{3+3+5+3}{4} = 3.5$

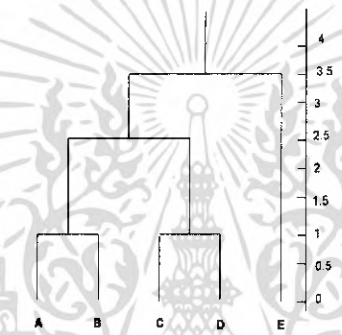
จากตารางที่ 3 จะเห็นได้ว่าไม่มี cluster คู่ใดที่มีค่าน้อยกว่าหรือเท่ากับ 3.0 จึงไม่เกิดการรวมกันของ cluster ดังนั้น เราจึงเพิ่มระยะห่างจากจุดเริ่มต้นเป็น 3.5 $(d = 3.5)$ ในการทำซ้ำครั้งที่ 7

รอบที่ 7 :

ขั้นที่ 1 เนื่องจากในการทำซ้ำครั้งที่ 6 ไม่เกิดการรวมกันของ cluster จึงส่งผลให้กระบวนการในขั้นที่ 1 ได้ผล ตามตารางที่ 3 ดังเดิม

ขั้นที่ 2 เปรียบเทียบระยะห่างเฉลี่ยระหว่าง cluster แต่ละคู่ว่ามีค่า น้อยกว่าหรือเท่ากับระยะห่างจากจุดเริ่มต้น $(d = 3.5)$ หรือไม่ถ้าน้อยกว่าหรือเท่ากับ ก็จะรวม cluster ทั้งสองเข้าด้วยกัน

จากตารางที่ 3 จะเห็นได้ว่า ค่าเฉลี่ยระยะห่างระหว่าง cluster สอง cluster จำนวน 1 คู่ ที่มีค่าน้อยกว่าหรือเท่ากับ 3.5 คือ cluster K_1 กับ K_2 จึงทำรวม cluster ทั้งสองเข้าด้วยกัน จะได้ cluster ใหม่ 1 cluster คือ $K_1 = \{A, B, C, D, E\}$ ดังรูปที่ 14



รูปที่ 14 Dendrogram การจัดcluster ด้วยวิธี Average Link Technique

เนื่องจากทุก item ในฐานข้อมูลจะรวมอยู่ภายใน cluster เดียวกันจึงเป็นการสิ้นสุดการจัดกลุ่มแบบ Hierarchical Algorithms ด้วยวิธี Average Link Technique

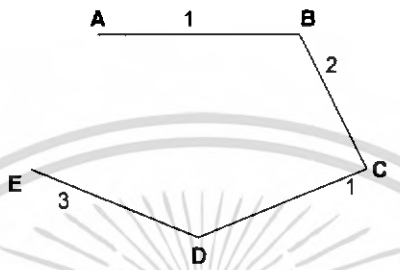
2. Divisive Clustering

Divisive Clustering เป็นเทคนิคที่ใช้ในการจัดกลุ่มโดยเริ่มจากการรวม item ทุก item ในฐานข้อมูล ให้อยู่ภายใน cluster เดียวกัน แล้วหาระยะห่างระหว่าง items จากนั้นจะแบ่ง items ภายใน cluster ออกเป็น 2 cluster โดยอาศัยแนวความคิดที่จะแยก cluster ที่ไม่มีความคล้ายคลึงกันออกจากกันซึ่งเราจะพิจารณา จากกิ่งที่มีระยะห่างระหว่าง item มากที่สุดมาเป็นตัวแบ่ง cluster ออกเป็น 2 cluster ก่อนและทำเช่นนี้ไปเรื่อยๆ โดยเลือกกิ่งที่มีระยะห่างระหว่าง item รองลงมาจนกระทั่งเราพบว่า 1 cluster เป็น 1 item จึงเป็นการสิ้นสุดการจัดกลุ่ม cluster

ตัวอย่างที่ 12 การจัดกลุ่ม cluster แบบ Hierarchical Algorithms ด้วยเทคนิค Divisive Clustering จากข้อมูลในตัวอย่างที่ 9

วิธีทำ

1. รวม item ทุก item ในฐานข้อมูลให้เป็น 1 cluster คือ $K_1 = \{A, B, C, D, E\}$
2. หาระยะห่างระหว่าง item โดยวิธี Minimum Spanning Tree ดังรูปที่ 15



รูปที่ 15 ระยะห่างที่ได้จากวิธี MST

3. จากนั้นเราทำการแบ่ง cluster ด้วยวิธีดังต่อไปนี้

รอบที่ 1 แบ่งกิ่งที่ได้จากวิธี MST โดยพิจารณาจากกิ่งที่มีระยะห่างระหว่าง item มากที่สุดก่อน ในที่นี้คือ item D และ E ซึ่งจะ ได้ cluster เป็น 2 cluster คือ

$K_1 = \{A, B, C, D\}$ และ $K_2 = \{E\}$

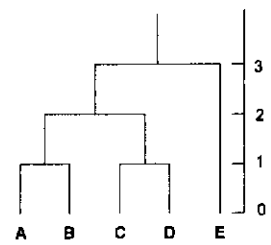
รอบที่ 2 แบ่งกิ่งที่มีระยะห่างระหว่าง item ที่ใหญ่รองลงมา คือกิ่งระหว่าง items B และ C จะได้ cluster ทั้งหมด 3 cluster คือ

$K_1 = \{A, B\}, K_2 = \{C, D\}, K_3 = \{E\}$

รอบที่ 3 แบ่งกิ่งที่มีระยะห่าง item มากเป็นอันดับ 3 คือ กิ่งระหว่าง item A กับ B และ C กับ D จะได้ cluster ทั้งหมด 5 cluster คือ

$K_1 = \{A\}, K_2 = \{B\}, K_3 = \{C\}, K_4 = \{D\}, K_5 = \{E\}$

ผลจากการทำรอบที่ 3 ได้ว่าในแต่ละ cluster มีเพียง 1 item เท่านั้น จึงเป็นการสิ้นสุดการจัดกลุ่ม cluster แบบ Hierarchical Algorithms ด้วยเทคนิค Divisive Clustering ซึ่งจะได้ผลลัพธ์เช่นเดียวกับวิธี Single Link Algorithm ดังรูปที่ 16



รูปที่ 16 Dendrogram การจัด cluster ด้วยวิธี Single Link Technique

3.1.2 Partitional Algorithms

Partitional Algorithms หรือ Nonhierarchical เป็นวิธีการสร้าง cluster ในขั้นตอนเดียวโดยผู้วิเคราะห์ต้องระบุจำนวน cluster ที่ต้องการ (k) อย่างไรก็ตามวิธีการของ Partitional Algorithms อาจมีข้อเสียคือการจัดกลุ่มที่เกิดจากการรวมตัวจากวิธีการที่เป็นไปได้นั้นมีจำนวนมากเกินไป ซึ่งในที่นี้จะกล่าวถึงเพียง 2 วิธี คือ

1. Minimum Spanning Tree
2. K – means Clustering

1. Minimum Spanning Tree (MST)

Minimum Spanning Tree เป็นเทคนิคที่ใช้ในการจัด cluster โดยจะหากิ่งที่เชื่อมจุดต่างๆ ในข่ายงาน โดยมีผลรวมระยะทางต่ำสุด ซึ่งมีวิธีการดังนี้

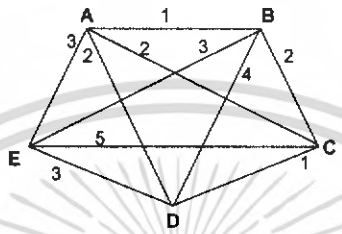
1. กำหนดให้ item ใด item หนึ่งเป็น item เริ่มต้นแล้วเชื่อมต่อกับ item ที่อยู่ใกล้ที่สุด
2. จาก item ที่เชื่อมแล้วในข้อ 1 หา item ที่ยังไม่ถูกเชื่อมที่ใกล้ที่อยู่ใกล้ที่สุดแล้วเชื่อม item ดังกล่าวและทำเช่นนี้ไปเรื่อยๆ จนทุกๆ item สามารถเชื่อมถึงกันหมด
3. เมื่อได้เส้นทางที่เชื่อม item ทุก item แล้วผู้วิเคราะห์จะต้องกำหนดจำนวน cluster ที่ต้องการแบ่ง ซึ่งกำหนดให้ k คือจำนวน cluster ที่ต้องการ จะต้องตัด $k-1$ กิ่งที่ใหญ่ที่สุดออกแล้วจับกลุ่ม cluster ตามกลุ่ม items ที่เชื่อมกันอยู่ ซึ่งจะได้ cluster ทั้งหมด k cluster ดังตัวอย่างต่อไปนี้

ตัวอย่างที่ 13 การจัดกลุ่ม Cluster 3 กลุ่มแบบ Partitional Algorithms ด้วยวิธี Minimum Spanning Tree จากข้อมูลในตัวอย่างที่ 9

วิธีทำ

ขั้นที่ 1 หากิ่งต่างๆที่เชื่อมจุดต่างๆในข่ายงานโดยวิธี MST ดังนี้

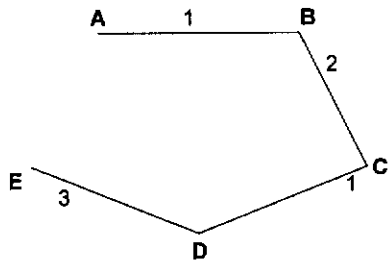
1. หาข้อมูลจากตัวอย่างที่ 9 มาวาดข่ายงานที่เชื่อมระหว่าง item ทั้งหมดดังรูปที่ 17



รูปที่ 17 ข่ายงานที่เชื่อม items

2. กำหนดให้ item A เป็น item เริ่มต้นแล้วเชื่อมต่อกับ item ที่อยู่ใกล้คือ item B ดังนั้น item ทั้งสองจึงถูกเชื่อมด้วยกิ่ง AB
3. item ที่ยังไม่ถูกเชื่อมที่อยู่ใกล้กิ่ง AB มากที่สุดคือ C (อาจเป็น D ก็ได้ เพราะมีระยะห่างจาก item ที่เชื่อมแล้วเป็น 2 เท่ากัน) ซึ่งอยู่ใกล้ B จึงเชื่อม item B ด้วยกิ่ง BC
4. item ที่ยังไม่ถูกเชื่อมที่อยู่ใกล้ item A , B , C มากที่สุดคือ D (ใกล้ C) จึงเชื่อม item D และ C ด้วยกิ่ง DC
5. item ที่ยังไม่ถูกเชื่อมที่อยู่ใกล้ item A , B , C , D มากที่สุดคือ E (ใกล้ D) จึงเชื่อม item E และ D ด้วยกิ่ง ED

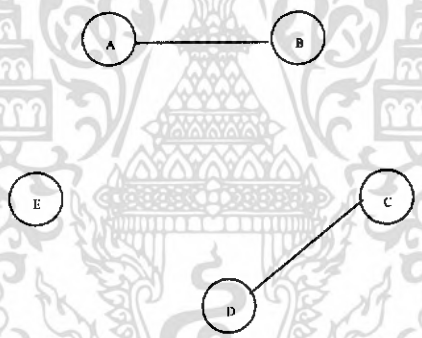
เนื่องจากทุก item ในฐานข้อมูลได้ถูกเชื่อมถึงกันหมดแล้ว จึงเป็นการสิ้นสุดการหากิ่งที่เชื่อม item แต่ละ item ในวิธี MST ดังรูปที่ 18 ซึ่งจะนำมาใช้ในการจัด cluster ในขั้นถัดไป



รูปที่ 18 ข่ายงานที่ได้จากเทคนิค MST

ขั้นที่ 2 การจัด cluster เป็น 3 cluster จะต้องทำการตัดกิ่งที่ใหญ่ที่สุดออก 2 กิ่ง นั่นคือ กิ่ง ED และ กิ่ง BC ดังรูปที่ 19 จากนั้นดำเนินการจัด cluster ตามกลุ่ม items ที่ยัง เชื่อมต่อกันอยู่จะได้ cluster ทั้งหมด 3 cluster ดังต่อไปนี้

$$K_1 = \{A , B\}$$
$$K_2 = \{C , D\}$$
$$K_3 = \{E\}$$



รูปที่ 19 แสดงการตัดกิ่ง Cluster ที่ใหญ่ที่สุดออก 2 กิ่ง

2. K-Means Clustering

K-Means Algorithm เป็นวิธีหนึ่งที่นิยมใช้ในการจัด cluster โดยผู้วิเคราะห์สามารถกำหนดจำนวน cluster ที่เราต้องการได้ล่วงหน้า สำหรับกรณีใช้ตัวแปรในการจัด cluster เพียงหนึ่งตัว ค่าที่ใช้พิจารณาในการจัดแบ่ง cluster คือ ระยะห่างจากข้อมูลกับ mean ของ cluster ต่างๆ แต่ถ้าพิจารณาตัวแปรที่ใช้ในการแบ่งกลุ่มมากกว่าหนึ่งตัว ค่าที่ใช้พิจารณา คือ Euclidian Distance ซึ่งคำนวณได้จาก

$$d(i,j) = \sqrt{\left|x_{i1} - x_{j1}\right|^2 + \left|x_{i2} - x_{j2}\right|^2 + \cdots + \left|x_{ip} - x_{jp}\right|^2}$$

ขั้นตอนในการวิเคราะห์มีดังนี้

1. กำหนดจำนวน cluster ที่ต้องการจะจัด cluster ซึ่งถ้ากำหนดให้ k คือ จำนวน cluster ที่ต้องการ
2. กำหนดค่า mean เริ่มต้นที่จะใช้ในการแบ่ง item จำนวน k ค่าให้กับ cluster จำนวน k cluster ซึ่งค่าดังกล่าวอาจได้จากการสุ่ม หรือข้อมูล k ตัวแรกที่จะนำมาใช้ในการจัด cluster
3. จัด item ในฐานข้อมูลที่มีค่าระยะห่างใกล้เคียงกับค่า mean ของ cluster เริ่มต้นที่กำหนดไว้ ให้อยู่ในกลุ่ม cluster ที่มีค่า mean ดังกล่าว หรือตัวที่กำหนด item เข้าไปไว้ใน cluster แล้ว คำนวณหาค่า mean ของแต่ละ cluster ใหม่เพื่อนำไปใช้ในการจัด cluster ครั้งต่อไป
4. คำนวณค่าระยะห่าง โดยนำมาเปรียบเทียบกับค่า mean ที่คำนวณได้ในแต่ละ cluster ถ้า item นั้นมีค่าระยะห่างใกล้เคียงกับค่า mean ใน cluster อื่นมากกว่า จะย้าย item ดังกล่าวไปยัง cluster นั้น และปรับค่า mean ใน cluster สำหรับ cluster ที่มีการเปลี่ยนแปลงข้อมูลอีกครั้ง
5. ทำซ้ำในขั้นตอนที่ 4 จนกระทั่งไม่มีการเคลื่อนย้าย item หรือไม่มีการเปลี่ยนแปลงค่า mean ในแต่ละ cluster

ในการจัดกลุ่ม cluster โดยวิธี K-means Algorithm นี้ จะทำให้สมาชิกใน cluster เดียวกันมีลักษณะเหมือนกันมากที่สุด และก็จะมีความแตกต่างจาก cluster อื่นมากเช่นกัน

ตัวอย่างที่ 14 จัดกลุ่ม cluster จำนวน สอง cluster ด้วยวิธี K-means Algorithm โดยตัวแปรที่พิจารณามีเพียงหนึ่งตัว โดยใช้ item ดังต่อไปนี้

$$\{2, 4, 10, 12, 3, 20, 30, 11, 25\}$$

วิธีทำ

ในกรณีนี้ตัวแปรที่พิจารณามีเพียง 1 ตัว ดังนั้นในการจัด cluster จะพิจารณาระยะห่างจาก mean กำหนดค่า mean เริ่มต้น 2 ค่าให้กับ cluster สอง cluster คือ $m_1 = 2$ และ $m_2 = 4$ ตามลำดับ และเนื่องจากใช้ตัวแปรเดียว จึงกำหนดให้ค่าของ item เป็นค่าระยะห่างที่นำมาใช้ในการจัดกลุ่ม ดังต่อไปนี้

รอบที่ 1

1. จัดแบ่ง item ออกเป็น 2 กลุ่มโดยถ้าค่า item ใดมีค่าระยะห่างใกล้เคียงกับค่า mean เริ่มต้นของ cluster ใดมากที่สุด ก็จัด item ดังกล่าวเข้าไปอยู่ใน cluster นั้น ซึ่งพบว่า item ที่อยู่ใกล้เคียงกับ $m_1 = 2$ คือ $k_1 = \{2, 3\}$ และ item ที่มีค่าระยะห่างใกล้เคียงกับ $m_2 = 4$ คือ $k_2 = \{4, 10, 11, 12, 20, 25, 30\}$

2. จาก cluster ที่ได้ในข้อ 1 หาค่า mean ของ cluster ทั้งสองใหม่ หาค่า item ที่อยู่ใน cluster ดังกล่าวจะได้ค่า mean เริ่มต้นเป็น

$$m_1 = \frac{2+3}{2} = 2.5 \text{ และ}$$

$$m_2 = \frac{4+10+11+12+20+25+30}{7} = 16$$

แล้วนำค่า mean ดังกล่าวมาใช้ในการจัดกลุ่มรอบถัดไป

รอบที่ 2

1. นำค่าระยะห่างของ item ในแต่ละ cluster เปรียบเทียบกับค่า mean เริ่มต้นที่คำนวณได้ใหม่ คือ $m_1 = 2.5$ และ $m_2 = 16$ จะพบว่า item 4 มีค่าระยะห่างใกล้เคียงกับค่า m_1 มากกว่าการย้าย item 4 มาไว้ใน cluster1 จะได้ผลลัพธ์ดังต่อไปนี้

$k_1 = \{2, 3, 4\}$ และ $k_2 = \{10, 11, 12, 20, 25, 30\}$

2. จากนั้น คำนวณค่า mean เริ่มต้นใหม่ของแต่ละ cluster จะได้ค่า mean เริ่มต้นเป็น

$$m_1 = \frac{2+3+4}{3} = 3 \text{ และ}$$

$$m_2 = \frac{10+11+12+20+25+30}{6} = 18$$

แล้วนำค่า mean ดังกล่าวมาใช้ในการจัดกลุ่มรอบถัดไป

รอบที่ 3

1. นำค่าระยะห่างของ item ในแต่ละ cluster ที่เปรียบเทียบกับค่า mean เริ่มต้นที่คำนวณได้ใหม่ คือ $m_1 = 3$ และ $m_2 = 18$ จะพบว่า item 10 มีค่าระยะห่างใกล้เคียงกับค่า m_1 มากกว่า ซึ่งจะย้าย item ดังกล่าวมาไว้ใน cluster1 ได้ผลลัพธ์ดังต่อไปนี้

$k_1 = \{2, 3, 4, 10\}$ และ $k_2 = \{11, 12, 20, 25, 30\}$

2. จากนั้น คำนวณค่า mean เริ่มต้นใหม่ของแต่ละ cluster จะได้ค่า mean เริ่มต้นเป็น

$$m_1 = \frac{2+3+4+10}{4} = 4.75 \text{ และ}$$

$$m_2 = \frac{11+12+20+25+30}{5} = 19.6$$

แล้วนำค่า mean ดังกล่าวมาใช้ในการจัดกลุ่มรอบถัดไป

รอบที่ 4

1. นำค่าระยะห่างของ item ในแต่ละ cluster ในรอบที่ 3 มาเปรียบเทียบกับค่า mean เริ่มต้นใหม่ของแต่ละ cluster จากรอบที่ 3 พบว่า item มีค่าใกล้เคียงกับค่า mean ใน cluster แรก $(m_1 = 4.75)$ มากกว่า cluster ที่ 2 $(m_2 = 19.6)$ จึงย้าย item ดังกล่าวมาไว้ใน cluster แรก จะได้ผลดังต่อไปนี้

$k_1 = \{2, 3, 4, 10, 11, 12\}$ และ $k_2 = \{20, 25, 30\}$

2. จากนั้น คำนวณค่า mean เริ่มต้นใหม่ของแต่ละ cluster จะได้ค่า mean เริ่มต้นเป็น

$$m_1 = \frac{2+3+4+10+11+12}{6} = 7$$

และ

$$m_2 = \frac{20+25+30}{3} = 25$$

แล้วนำค่า mean ดังกล่าวมาใช้ในการจัดกลุ่มรอบถัดไป

รอบที่ 5

นำค่าระยะห่างของ item ในแต่ละ cluster ในรอบที่ 4 มาเปรียบเทียบกับค่า mean ที่คำนวณได้ใหม่จากรอบที่แล้ว ซึ่งพบว่าไม่มีการเคลื่อนไหวของข้อมูลระหว่างสอง cluster จึงเป็นการสิ้นสุดการจัดกลุ่ม cluster ซึ่งจะได้ cluster ที่มีลักษณะดังนี้ $k_1 = \{2, 3, 4, 10, 11, 12\}$ และ $k_2 = \{20, 25, 30\}$

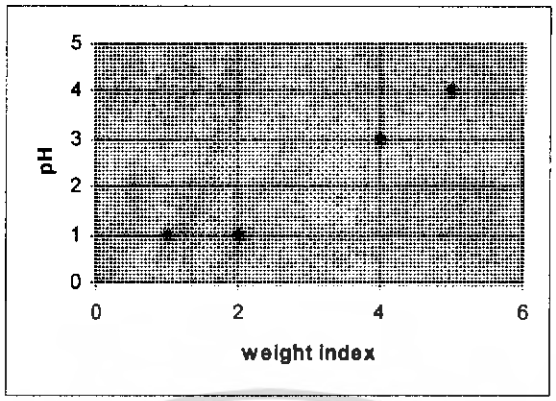
ตัวอย่างที่ 15 การจัดกลุ่มยาเป็น 2 กลุ่ม $(k = 2)$ โดยใช้ค่า pH และ Weight Index ในการจัดกลุ่มซึ่งมีข้อมูลดังนี้

ตารางที่ 4 ตัวอย่างของข้อมูล Medicine ที่จะใช้ในการจัดกลุ่ม

object	weight index	pH
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4

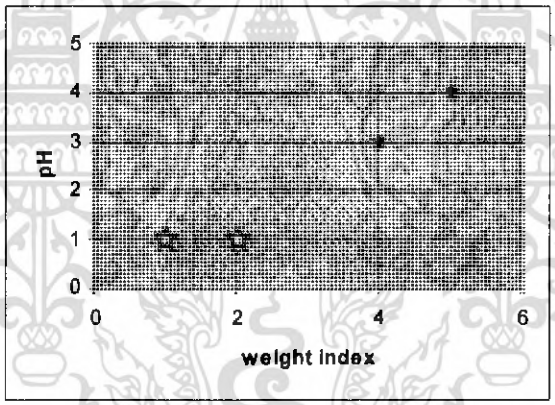
วิธีทำ

ในกรณีนี้ ตัวแปรที่พิจารณามี 2 ตัว ดังนั้นในการจัด cluster จะพิจารณาค่า Euclidian Distance ซึ่งกำหนดให้ Medicine แต่ละตัวแทนด้วยจุด 1 จุด ในกราฟ ซึ่งประกอบด้วยค่าตัวแปร 2 ตัว ซึ่งแสดงเป็นพิกัดได้ดังรูปที่ 20 แล้วทำการจัดกลุ่ม cluster ตามขั้นตอนดังต่อไปนี้



รูปที่ 20 แสดงพิกัดของข้อมูล Medicine

1. กำหนดค่า centroid (ค่ากลาง) เริ่มต้นเป็น Medicine A และ Medicine B ซึ่งสามารถชดเชยแทนด้วยพิกัด centroid ดังนี้ $c_1 = (1,1)$ และ $c_2 = (2,1)$



รูปที่ 21 แสดงการกำหนดพิกัดของ centroid เริ่มต้น

2. เป็นการหาค่า distance หรือระยะระหว่าง object แต่ละตัวเปรียบเทียบกับค่า centroids ในแต่ละกลุ่ม ซึ่งเราจะใช้ Euclidean distance ในการหา ดังนั้นเราจะได้ distance matrix ที่ iteration 0 เป็น

$$\mathbf{D}^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (1,1) \text{ group - 1} \\ \mathbf{c}_2 = (2,1) \text{ group - 2} \end{matrix}$$

$$\begin{bmatrix} A & B & C & D \\ 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} \\ X \\ Y \end{matrix}$$

แต่ละ column ใน distance matrix แสดงถึง object แถวแรกของ distance matrix เป็นระยะระหว่าง object แต่ละ object เปรียบเทียบกับ centroid ตัวแรก และแถวที่สองก็คือระยะระหว่างแต่ละ object เปรียบเทียบกับ centroid ตัวที่ 2 ตัวอย่างเช่น ระยะระหว่าง Medicine C $= (4, 3)$ กับ centroid ตัวแรก $(c_1 = (1, 1))$ คือ $\sqrt{(4-1)^2 + (3-1)^2} = 3.61$ และระยะระหว่าง centroid ตัวที่ 2 $(c_2 = (2, 1))$ คือ $\sqrt{(4-2)^2 + (3-1)^2} = 2.83$ เป็นต้น

3. เป็นการจัดกลุ่ม objects คือเรากำหนดแต่ละ object ไปไว้ในกลุ่มที่มี distance ที่น้อยที่สุด ดังนั้น Medicine A จะถูกกำหนดให้อยู่ในกลุ่มที่ 1 , Medicine B อยู่ในกลุ่ม 2, Medicine C อยู่ในกลุ่ม 2 และ Medicine D ก็อยู่ในกลุ่ม 2 เช่นเดียวกัน ซึ่ง object ที่อยู่ในกลุ่มใดนั้นแทนด้วย 1 ดังแสดงใน Group matrix ด้านล่างนี้

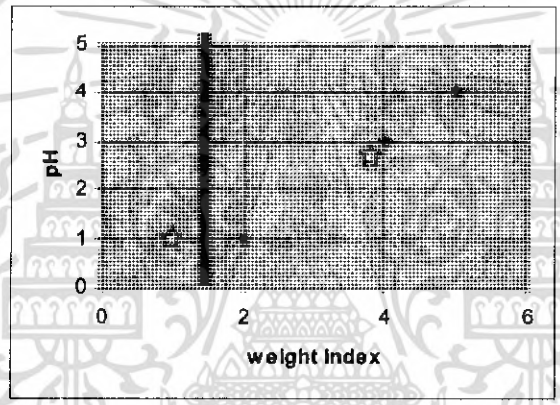
$$\mathbf{G}^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \quad \begin{matrix} \text{group - 1} \\ \text{group – 2} \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

4. ในรอบที่ 1 นี้จะเป็นการหาค่า centroids ใหม่ ซึ่งเมื่อเราทราบสมาชิกในแต่ละกลุ่มแล้ว ในขั้นตอนนี้เราจะทำการคำนวณหาค่า centroid ของแต่ละกลุ่มใหม่ ซึ่งในกลุ่มที่ 1 นั้นมีสมาชิกเพียงหนึ่งตัว ดังนั้นค่า centroid ก็ยังคงเป็นตัวเดิมคือ $(c_1 = (1,1))$ ในกลุ่มที่ 2 ซึ่งมีสมาชิดอยู่ 3 ตัว ดังนั้นค่า centroid คือการหาค่าเฉลี่ยของพิกัดระหว่างสมาชิก 3 ตัว เพราะฉะนั้นจะได้

$$c_2 = \left(\frac{2+4+5}{3}, \frac{1+3+4}{3}\right)$$

$$= \left(\frac{11}{3}, \frac{8}{3}\right)$$



รูปที่ 22 พิกัดของ centroids ที่ได้จากการคำนวณในรอบที่ 1

5. ในรอบที่ 1 จะเป็นการคำนวณหา distance ของทุกๆ object เปรียบเทียบกับ centroids ใหม่ คล้ายในขั้นตอนที่ 2 ซึ่ง distance matrix ใน รอบที่ 1 คือ

$$\mathbf{D}^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \begin{matrix} c_1 = (1,1) \text{ group - 1} \\ c_2 = (11/3, 8/3) \text{ group - 2} \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \begin{matrix} X \\ Y \end{matrix}$$

6. รอบที่ 1 จะเป็นการจัดกลุ่มของ object เหมือนในขั้นตอนที่ 3 ซึ่งเราจะทำการกำหนดแต่ละ object เข้าไปในกลุ่มที่มี distance น้อยสุด จาก distance matrix ที่ได้จากการคำนวณใหม่นี้ เราจะย้าย Medicine B ไปไว้ในกลุ่ม 1 ในขณะที่ object ตัวอื่นๆ ก็ยังอยู่ในกลุ่มเหมือนเดิม ซึ่งจะได้ Group matrix ดังแสดงข้างล่างนี้

$$G^1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad \begin{matrix} \text{group - 1} \\ \text{group - 2} \end{matrix}$$

$$\quad A \quad B \quad C \quad D$$

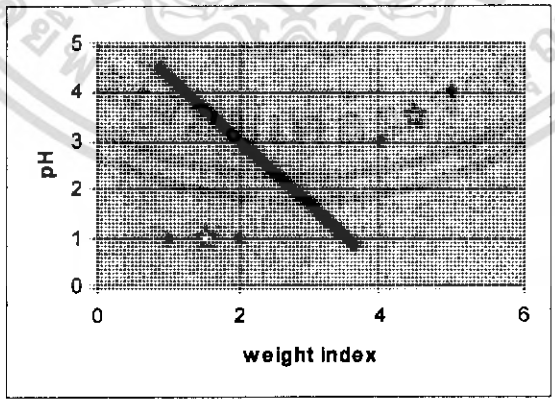
7. รอบที่ 2 เป็นการหาค่า centroids ใหม่ ซึ่งเราจะทำซ้ำขั้นตอนที่ 4 เพื่อคำนวณหาพิกัดของ centroids ใหม่ โดยจะพิจารณาจากการจัดกลุ่มในรอบที่แล้ว ซึ่งทั้งในกลุ่มที่ 1 และ 2 ต่างก็มีสมาชิก 2 ตัว ดังนั้น centroids ใหม่ที่ได้คือ

$$c_1 = \left(\frac{1+2}{2}, \frac{1+1}{2}\right)$$

$$= \left(1\frac{1}{2}, 1\right)$$

$$c_2 = \left(\frac{4+5}{2}, \frac{3+4}{2}\right)$$

$$= \left(4\frac{1}{2}, 3\frac{1}{2}\right)$$



รูปที่ 23 พิกัดของ centroids ที่ได้จากการคำนวณในรอบที่ 2

8. รอบที่ 2 ทำซ้ำใน step 2 อีกครั้ง เราจะได้ distance matrix ใหม่ใน รอบที่ 2 เป็นดังนี้

$$\mathbf{D}^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (3/2,1) \text{ group - 1} \\ \mathbf{c}_2 = (9/2,7/2) \text{ group - 2} \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

9. รอบที่2 เป็นการกำหนดแต่ละ object เข้าไปอยู่ในกลุ่มที่มี distance น้อยที่สุด

$$\mathbf{G}^2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad \begin{matrix} \text{group - 1} \\ \text{group – 2} \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

ผลลัพธ์ที่ได้ คือ $G^2 = G^1$ ซึ่งการเปรียบเทียบการจัดกลุ่มของรอบสุดท้ายและในรอบนี้แสดงให้เห็นว่า object ไม่ได้มีการเปลี่ยนกลุ่ม ด้วยเหตุนี้การทำงานของ K-means Algorithm ได้มาถึงจุดที่ไม่มีการเปลี่ยนแปลงแล้ว ดังนั้นเราจะเอากลุ่มที่ได้จากการจัดกลุ่มครั้งสุดท้ายมาเป็นผลลัพธ์ ดังแสดงในตารางด้านล่างนี้

ตารางที่ 5 ผลลัพธ์ที่ได้จากการจัดกลุ่ม

Object	weight index	pH	Group (result)
Medicine A	1	1	1
Medicine B	2	1	1
Medicine C	4	3	2
Medicine D	5	4	2

ข้อเสียของ K-means Algorithm

1. ใช้ได้เพียงข้อมูลที่มีรูปแบบเป็นตัวเลขหรือมีการแปลงเป็นตัวเลขแล้วเท่านั้น
2. จำเป็นต้องระบุจำนวนของกลุ่ม (k) ที่ต้องการจะจัดกลุ่มก่อนการจัดกลุ่มเสมอ
3. ทำงานผิดพลาดกับข้อมูลที่เป็น noisy และ outliers ได้เพื่อกำจัดปัญหาของ outliers เราสามารถใช้ค่า median (ค่ามัธยฐาน) แทนค่า mean (ค่าเฉลี่ย) ได้

2.2 Summarization

การสรุปผลของข้อมูลที่จะกล่าวถึงใน Data Mining นั้นมี 2 ส่วน โดยในส่วนแรกเกี่ยวกับการวัดแนวโน้มสู่ส่วนกลาง (Measure of Central Tendency) ได้แก่ ค่าเฉลี่ย (Mean) ค่ามัธยฐาน (Median) ค่าฐานนิยม (Mode) และในส่วนที่สองเกี่ยวกับ การวัดการกระจาย (Measure of Variability) ได้แก่ ค่าความแปรปรวน (Variance) ค่าส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation) โดยข้อมูลที่จะนำมาวิเคราะห์นี้จะเป็นข้อมูลที่ไม่ได้แบ่งกลุ่ม (Ungrouped Data) กล่าวคือ ข้อมูลที่ไม่ได้จัดอยู่ในรูปตารางแจกแจงความถี่ ถ้ากำหนดให้ $x_1,...,x_n$ คือค่าของข้อมูลขนาด n ที่เราต้องการวิเคราะห์

1. **การวัดแนวโน้มสู่ส่วนกลาง (Measure of Central Tendency)**

   สำหรับการวัดแนวโน้มสู่ส่วนกลาง จะกล่าวถึง 3 วิธี ดังนี้

   1. ค่าเฉลี่ยเลขคณิต (Arithmetic Mean) ซึ่งนิยมเรียกสั้นๆว่า ค่าเฉลี่ย (Mean) เป็นการหาค่ากลางของข้อมูล โดยนำข้อมูลทุกตัวมารวมกันแล้วหารด้วยจำนวนข้อมูลทั้งหมด

      $$Mean = \frac{\sum_{i=1}^{n} x_i}{n}$$

   2. ค่ามัธยฐาน (Median) คือ ค่าที่อยู่ ณ ตำแหน่งตรงกลางของชุดข้อมูล เมื่อได้มีการเรียงลำดับข้อมูลจากค่าน้อยไปหาค่ามากเรียบร้อยแล้ว ซึ่งค่ามัธยฐานนี้จะเป็นค่าที่บอกให้ทราบว่าข้อมูลจำนวนครึ่งหนึ่งที่มีค่าน้อยกว่าหรือเท่ากับค่านี้ และมีข้อมูลอีกครึ่งหนึ่งที่มีค่ามากกว่าค่านี้ สำหรับข้อมูลตัวอย่าง n จำนวน ถ้า n เป็นเลขคี่

      $$Median = \text{ค่าของข้อมูลที่อยู่ตำแหน่ง } \frac{n+1}{2}$$

      แต่ถ้า n เป็นเลขคู่

      $$Median = \text{ค่าเฉลี่ยของข้อมูลที่อยู่ตำแหน่ง } \frac{n}{2} \text{ และ } \frac{n+2}{2}$$

   3. ฐานนิยม (Mode) คือ ค่าของข้อมูลที่มีความถี่ (ความซ้ำ) มากที่สุด ซึ่งสามารถใช้ได้ดีเมื่อข้อมูลที่สนใจศึกษามีค่าที่เกิดขึ้นซ้ำๆ กัน

2. การวัดการกระจาย (Measure of Variability)

สำหรับการวัดการกระจายจะกล่าวถึง 2 วิธี ดังนี้

1. ค่าความแปรปรวน (Variance) ซึ่งพิจารณาว่าค่าของข้อมูลแต่ละตัว มีค่าแตกต่างจากค่าเฉลี่ยมากน้อยเพียงใด ซึ่งถ้าข้อมูลชุดใดมีการกระจายน้อย แสดงว่าข้อมูลแต่ละค่าแตกต่างจากค่าเฉลี่ยน้อย แต่ถ้าข้อมูลชุดใดมีการกระจายมาก แสดงว่าข้อมูลแต่ละค่าแตกต่างจากค่าเฉลี่ยมาก ซึ่งคำนวณได้ดังนี้

$$\text{ความแปรปรวน } (s^2) \quad = \quad \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

เมื่อ $\bar{x}$ คือค่าเฉลี่ยของชุดข้อมูลนี้

2. ค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) ซึ่งคือ รากที่สองของความแปรปรวน เฉพาะรากที่เป็นบวกเท่านั้น ซึ่งคำนวณได้ดังนี้

$$\text{ค่าเบี่ยงเบนมาตรฐาน } (s) \quad = \quad \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

ตัวอย่างที่ 16 การหาค่าจุดกึ่งกลางของข้อมูลและค่าการกระจายของชุดข้อมูล T ซึ่งมีค่าดังนี้

$$T = \{3, 5, 2, 9, 0, 7, 3, 6\}$$

วิธีทำ

ให้ x_i แทนข้อมูลที่ต้องการวิเคราะห์ เรียงลำดับข้อมูลในชุดข้อมูล T จากน้อยไปมาก จะได้ค่าดังนี้

ลำดับ	1	2	3	4	5	6	7	8
ข้อมูล	0	2	3	3	5	6	7	9

ค่ากึ่งกลางของข้อมูล

$$Mean = \frac{\sum_{i=1}^{8} x_i}{8}$$

$$= \frac{(0+2+3+3+5+6+7+9)}{8}$$

$$= 4.375$$

เนื่องจากข้อมูลชุดนี้มี 8 ค่า กล่าวคือ n เป็นเลขคู่ ดังนั้น

ค่ามัธยฐาน = ค่าเฉลี่ยของค่าที่ตำแหน่ง $\frac{8}{2}$ และ $\frac{10}{2}$

= ค่าเฉลี่ยของค่าที่ตำแหน่ง 4 และ 5

= ค่าเฉลี่ยของค่า 3 และ 5

$$= \frac{3+5}{2}$$

$$= 4$$

ค่าฐานนิยม คือ 3 (3 เกิดขึ้นมากที่สุด คือ 2 ครั้ง)

ค่าการกระจายของข้อมูล

ค่าความแปรปรวน (s^2)

$$= \frac{\sum_{i=1}^{8} (x_i - 4.375)^2}{7}$$

$$= 8.5532$$

ค่าส่วนเบี่ยงเบนมาตรฐาน (s)

$$= \sqrt{8.5532}$$

$$= 2.9246$$

2.3 กฎของความสัมพันธ์ (Association Rules)

กฎของความสัมพันธ์ (Association Rules) เป็นเทคนิคหนึ่งของ Data Mining ซึ่งศึกษาเกี่ยวกับรูปแบบความสัมพันธ์ของข้อมูลที่ได้จัดเก็บไว้ในฐานข้อมูลที่สามารถเขียนได้ในรูปแบบ $X \Rightarrow Y$

กฎของความสัมพันธ์นั้นมีวิธีการศึกษาใช้งานแตกต่างกันหลายรูปแบบซึ่งในที่นี้ศึกษาเพียงวิธีวิเคราะห์ตะกร้าตลาด(Market Basket Analysis: MBA) ซึ่งจะเป็นการหากฎเกณฑ์ที่บอกความสัมพันธ์ของข้อมูลโดยกฎเกณฑ์ที่ได้จะต้องมีคุณสมบัติสอดคล้องกับค่า Minimum Support และ Minimum Confidence ที่กำหนดโดยผู้ใช้ซึ่งจะกล่าวถึงต่อไป

1. การวิเคราะห์ตะกร้าตลาด (Market Basket Analysis: MBA)

การวิเคราะห์ตะกร้าตลาด (MBA) เป็นวิธีการพื้นฐานที่ใช้ในการวิเคราะห์พฤติกรรมการซื้อของผู้บริโภคโดยเป็นการศึกษาว่าหากลูกค้าซื้อสินค้าชนิดใดแล้วจะมีผลต่อการที่ลูกค้าจะซื้อสินค้าชนิดอื่นๆอีกหรือไม่ กล่าวคือเป็นการศึกษาความเชื่อมโยงของการซื้อสินค้าชนิดต่างๆซึ่งจัดรวมเป็นเซตสำหรับลูกค้าแต่ละคนที่ซื้อในแต่ละครั้งเพื่อนำไปใช้ปรับเปลี่ยนกลยุทธ์ในการปรับปรุงและออกแบบการวางสินค้าไว้ในร้านหรือในหน้าเว็บเพจ (Webpage)

โดยกลยุทธ์หนึ่งในการจัดวางสินค้า คือ การนำสินค้าที่เราพบว่าลูกค้าจะซื้อพร้อมกันบ่อยๆนำมาจัดวางใกล้ๆกันเพื่อที่จะส่งเสริมการขายได้หรือในทางกลับกันเราอาจจะวางสินค้าดังกล่าวไว้ไกลๆคนละข้างของร้านเพื่อเป็นการชักนำให้ลูกค้าเดินชมสินค้าซึ่งอาจมีการตัดสินใจซื้อสินค้าอื่นๆระหว่างทางได้ นอกจากนี้ผลการวิเคราะห์ยังสามารถนำมาใช้ในการวางแผนการขายอื่นๆ เช่นการกำหนดราคาสินค้าได้อีกด้วย

2. แนวความคิดพื้นฐานในการวิเคราะห์ตะกร้าตลาด (MBA)

จากนิยามกำหนดให้

$I = \{I_1, I_2, ..., I_m\}$ เป็นเซตของสินค้าหรือ item

$D = \{T_1, T_2, ..., T_m\}$ เป็นเซตของฐานข้อมูลทางธุรกิจ

เมื่อธุรกรรม T แต่ละตัวนั้นเป็นเซตของ item นั่นคือ $T \subseteq I$ โดยในแต่ละธุรกรรมนั้นจะมีตัวที่ใช้ในการแสดงความแตกต่างระหว่างเซตของ item แต่ละเซตออกจากกันโดยเอกลักษณ์ทางธุรกิจ (Transaction Identifier) หรือที่เรียกว่า TID

จากกฎของความสัมพันธ์ในธุรกิจที่เขียนในรูปแบบ $X \Rightarrow Y$ เมื่อ $X, Y \in I$ และ $X \cap Y = \phi$ โดยที่

ค่า Support (S) ในเงื่อนไข $X \Rightarrow Y$ คือ ร้อยละของจำนวนธุรกรรมใน D ที่ครอบคลุม $X \cup Y$ หรือค่าความน่าจะเป็นที่ลูกค้าจะซื้อเซตของสินค้า X และ Y พร้อมๆกัน

ค่า Confidence (C) คือ ค่าร้อยละของจำนวนธุรกรรมใน D ที่ครอบคลุมทั้ง X และ Y $(X \cap Y)$ หรือค่าความน่าจะเป็นเมื่อลูกค้าซื้อเซตของสินค้า X แล้วจะซื้อเซตของสินค้า Y ด้วย

ถ้ากฎที่ใช้นั้นถูกคัดเลือกมาจากค่า Minimum Support และ Minimum Confidence ที่ผู้ใช้กำหนด เราจะเรียกกฎดังกล่าวว่ากฎความสัมพันธ์แบบเข้มงวด (Strong Association Rules) โดยค่าดังกล่าวนั้นนิยมกำหนดให้อยู่ในรูปของเปอร์เซ็นต์ที่มีค่าตั้งแต่ 0% ถึง 100 % เช่น

$$\text{Computer} \Rightarrow \text{Financial\_management\_Software}$$
$$[\text{Support} = 2\%, \text{Confidence} = 60\%]$$

Support 2% หมายถึง ปริมาณของจำนวนธุรกรรมที่เรากำลังสนใจศึกษาอยู่ ส่วนค่า Confidence 60% หมายถึง 60% ของลูกค้าที่ทำการสั่งซื้อคอมพิวเตอร์แล้วจะสั่งซื้อ Software การจัดการทางด้านการเงินด้วย

โดยทั่วไปแล้วการวัดความน่าสนใจของข้อมูลที่ต้องการจะทำการวิเคราะห์ เราจะใช้ค่า Support และ ค่า Confidence ของข้อมูลดังกล่าวมาเปรียบเทียบกับค่า Minimum Support และ Minimum Confidence ที่กำหนดแล้วมีค่ามากกว่าแสดงว่าข้อมูลดังกล่าวมีความน่าสนใจ

ถ้ากำหนดให้ itemset คือ เซตของสินค้า(item)ที่ซื้อ

k – itemset คือ เซตที่ประกอบด้วนสินค้า k อย่าง

เช่น {Computer,Financial_Management_Software} เราจะเรียกว่า 2 – itemset โดยมีแนวคิดในการสร้างกฎ (Rules) จากฐานข้อมูลขนาดใหญ่ดังนี้

1. หา Large itemset โดย itemset นั้นจะต้องมีค่า Support ไม่น้อยกว่าค่า Minimum Support ที่กำหนดไว้
2. หากฎของความสัมพันธ์แบบเข้มงวด (Strong Association Rules) โดยกฎที่ได้จากค่านี้ต้องมี Support และ Confidence มากกว่า Minimum Support และ Minimum Confidence ที่กำหนดไว้ ซึ่งเราสามารถคำนวณหาค่า Confidence ได้จาก ค่า Support ใน Large itemset ที่เกี่ยวข้อง

อย่างไรก็ตามฐานข้อมูลในปัจจุบันมีขนาดใหญ่ซึ่งมีความยุ่งยากมากในการหากฎ (Rules) ดังกล่าว จำเป็นต้องใช้ในการวิเคราะห์สำหรับฐานข้อมูลขนาดใหญ่ โดยเทคนิคที่นิยมใช้กันในปัจจุบันคือ Apriori Algorithm ซึ่งจะกล่าวถึงต่อไป

3. การหากฎของความสัมพันธ์โดยใช้ Apriori Algorithms

Apriori Algorithm เป็นกระบวนการสร้างกฎของความสัมพันธ์จากค่าของความถี่ itemset (Frequent itemset) ในฐานข้อมูลที่สอดคล้องกับกฎการ Support เพื่อให้ได้ Large itemset โดยการทำซ้ำหลายๆครั้งซึ่งในการทำซ้ำแต่ละครั้งจะประกอบด้วย 3 ขั้นตอนที่สำคัญคือ

1. สร้าง Candidate itemset โดยการรวมกันของ Large (i -1) itemset $\left(L_{i-1} \times L_{i-1}\right)$ โดยอาศัยหลักการที่ว่า

$$L_k \times L_k = \{X \cup Y \text{ เมื่อ } X,Y \in L_k, \text{ จำนวนข้อมูล } X \cap Y = k-1 \}$$

เมื่อ i คือจำนวนครั้งของการทำซ้ำ และสมาชิกใน itemset มีจำนวน i ตัว โดยกำหนดให้ในการทำซ้ำครั้งแรก 1 item เป็น 1 Candidate itemset

2. นับจำนวนความถี่ที่เกิดขึ้นของแต่ละ Candidate itemset แล้วทำการคำนวณหาค่า Support ของแต่ละ Candidate itemset

3. เลือกข้อมูลโดยทำการคัดเลือก Candidate itemset ที่มีค่า Support สูงกว่าค่า Minimum Support ที่ผู้ใช้กำหนดไว้ ซึ่งเป็นค่าที่คาดว่าน่าจะยอมรับได้โดย item ที่ถูกคัดเลือกจะกลายเป็น Large i-items เพื่อนำไปสร้าง Candidate ในการทำซ้ำครั้งถัดไป

ตัวอย่างที่ 17 การสร้างกฎ (Rules)โดยใช้ Apriori Algorithms จากข้อมูลในฐานข้อมูลดังนี้

TID	items
001	A , C , D
002	B , C , E
003	A , B , C , E
004	B , E

วิธีทำ

กำหนดให้ค่า Minimum Support (S) = 50% ซึ่งจากข้อมูลในตารางจะเห็นได้ว่ามีข้อมูล 4 itemset อยู่ในฐานข้อมูลโดยใช้ Apriori Algorithms จะได้

รอบที่ 1 สร้าง 1– itemset

ขั้นที่ 1 สำหรับรอบที่ 1 เราจะกำหนด item แต่ละ item ในฐานข้อมูลเป็น Candidate itemset ดังตารางที่ 6-a

ขั้นที่ 2 ทำการ Scan ข้อมูลในฐานข้อมูลเพื่อหาความถี่ของข้อมูลในแต่ละ Candidate ที่เกิดขึ้นจริง แล้วหาค่า Support จากอัตราส่วนของความถี่แต่ละ Candidate itemset กับ จำนวน itemset ทั้งหมดในฐานข้อมูลดังตาราง 6-b

ขั้นที่ 3 เลือก Candidate itemset จาก C_1 ที่มีค่า Support สูงกว่าค่า Minimum Support ที่กำหนดไว้ คือ 50% ไปสู่ Large 1– itemset ดังตาราง 6-c

ตารางที่ 6 ผลการวิเคราะห์รอบที่ 1 โดยใช้ Apriori Algorithms

C_1

1-itemset
{A}
{B}
{C}
{D}
{E}

6-a)Candidate 1- itemset

C_1

1-itemset	Frequency	S[%]
{A}	2	50
{B}	3	75
{C}	3	75
{D}	1	25
{E}	3	75

6-b) ความถี่ และค่า Support ของ 1- itemset

L_1

1-itemset	Frequency	S[%]
{A}	2	50
{B}	3	75
{C}	3	75
{E}	3	75

6-c) แสดง L_1 1- itemset ที่ได้จากการเลือก(Prune Phase)

รอบที่ 2 สร้าง 2 – itemset จากค่าที่ผ่านเกณฑ์จากรอบที่ 1

ขั้นที่ 1 สร้าง Candidate itemset โดยใช้ข้อมูลจาก Large 1 - itemset(L_1) โดยการรวม item ทั้งหมดที่เป็นไปได้และแต่ละ item มีสมาชิกไม่เกิน 2 ตัว$(L_1 \times L_1)$ดังตารางที่ 7-a

ขั้นที่ 2 ทำการ Scan ข้อมูลในฐานข้อมูลเพื่อหาความถี่ของข้อมูลในแต่ละ Candidate แล้ว หาค่า Support ดังตารางที่ 7-b

ขั้นที่ 3 ทำการเลือก Candidate itemset จาก C_2 ที่มีค่าSupport สูงกว่า 50% ไปสู่ Large 2 - itemset(L_2) ดังตาราง 7-c

ตารางที่ 7 ผลการวิเคราะห์รอบที่ 2 โดยใช้ Apriori Algorithms

C_2

2 - itemset
{A , B}
{A , C}
{A , E}
{B , C}
{B , E}
{C , E}

7-a) Candidate 2 - itemset

C_2

2- itemset	Frequency	S[%]
{A , B}	1	25
{A , C}	2	50
{A , E}	1	25
{B , C}	2	50
{B , E}	3	75
{C , E}	2	50

7-b) ความถี่ และค่า Support ของ 2 - itemset

L_2

2- itemset	Frequency	S[%]
{A , C}	2	50
{B , C}	2	50
{B , E}	3	75
{C , E}	2	50

7-c) แสดง L_2 2 - itemset ที่ได้จากการเลือก (Prune Phase)

รอบที่ 3 สร้าง 3 – itemset จากค่าที่ผ่านเกณฑ์จากรอบที่ 2

ขั้นที่ 1 สร้าง Candidate itemset โดยใช้ข้อมูลจาก Large 2 itemset (L_2) โดยในการรวมกันนี้จะทำการรวม Large 1 itemset ที่มี item แรกเหมือนกันดังเช่น {B , C} กับ {B , E} จะได้ Candidate itemset ดังตารางที่ 8-a

ขั้นที่ 2 ทำการ Scan ข้อมูลในฐานข้อมูลเพื่อหาความถี่ของข้อมูลในแต่ละ Candidate แล้ว หาค่า Support ดังตารางที่ 8-b

ขั้นที่ 3 ทำการเลือก Candidate itemset จาก C_3 ที่มีค่าSupport สูงกว่า 50% ไปสู่ Large 3 - itemset (L_3) ดังตาราง 8-c

ตารางที่ 8 ผลการวิเคราะห์รอบที่3 โดยใช้ Apriori Algorithms

C_3

3- itemset
{B , C , E}

8-a)Candidate 3 - itemset

C_3

3- itemset	Frequency	S[%]
{B , C , E}	2	50

8-b) ความถี่และค่า Support ของ 3- itemset

L_3

3- itemset	Frequency	S[%]
{B , C , E}	2	50

8-c) แสดง L_3 3 - itemset ที่ได้จากการเลือก(Prune Phase)

เมื่อได้ Large itemset แล้วเราสามารถหากฎความสัมพันธ์แบบเข้มงวด (Strong Association Rules) จากขั้นตอนต่างๆ ดังนี้

ขั้นที่ 1 หาค่า Support ที่สอดคล้องกับกฎของความสัมพันธ์ (Association Rules)

ขั้นที่ 2 หาค่า Confidence ของกฎของความสัมพันธ์จากอัตราส่วนของค่า Support ที่มีความสัมพันธ์กับกฎดังกล่าว

ขั้นที่ 3 ถ้าค่า Confidence ที่คำนวณได้มากกว่าค่า Minimum Confidence แสดงว่ากฎ (Rules) ดังกล่าวเป็นกฎเป็นกฎของความสัมพันธ์แบบเข้มงวด (Strong Association Rules) นั่นคือ มีความเป็นไปได้สูงที่จะเกิดกฎ (Rules) ดังกล่าว

ตัวอย่างที่ 18 ตรวจสอบว่า $\{B,C\} \Rightarrow E$ ที่ได้โดยใช้ตัวอย่างที่ 1 เป็นกฎของความสัมพันธ์แบบเข้มงวด (Strong Association Rules) หรือไม่ โดยกำหนดให้ Minimum Confidence = 0.8

วิธีทำ

จากตัวอย่างที่ 1 กฎที่ได้คือ $\{B,C\} \Rightarrow E$ จะได้

$$C(\{B,C\}) \Rightarrow E = \frac{S(\{B,C,E\})}{S(\{B,C\})}$$

$$= \frac{2}{2} = 1$$

$$= 100\%$$

จะเห็นได้ว่าค่าที่ได้สูงกว่า 0.8 หรือ 80% แสดงว่า $\{B,C\} \Rightarrow E$ เป็นกฎของความสัมพันธ์แบบเข้มงวด (Strong Association Rules) นั่นคือมีความเป็นไปได้สูงที่เมื่อลูกค้าซื้อสินค้า B, C แล้วจะต้องซื้อสินค้า E

ประวัติคณะผู้จัดทำ

ชื่อ – นามสกุล	นายนิวัฒน์ ไทเศรษฐวัฒน์กุล
วัน/เดือน/ปี เกิด	21 พฤศจิกายน 2526
สถานที่เกิด	ฉะเชิงเทรา
การศึกษาระดับมัธยมต้น	โรงเรียนเบญจมราชรังสฤษฎิ์ ฉะเชิงเทรา
การศึกษาระดับมัธยมปลาย	โรงเรียนเบญจมราชรังสฤษฎิ์ ฉะเชิงเทรา

ชื่อ – นามสกุล	นายรติ เพิ่มพูล
วัน/เดือน/ปี เกิด	13 กรกฎาคม 2528
สถานที่เกิด	ฉะเชิงเทรา
การศึกษาระดับมัธยมต้น	โรงเรียนเบญจมราชรังสฤษฎิ์ ฉะเชิงเทรา
การศึกษาระดับมัธยมปลาย	โรงเรียนเบญจมราชรังสฤษฎิ์ ฉะเชิงเทรา

ชื่อ – นามสกุล	นายสรรเพชญ ภูมรินทร์
วัน/เดือน/ปี เกิด	20 กรกฎาคม 2527
สถานที่เกิด	กรุงเทพมหานคร
การศึกษาระดับมัธยมต้น	โรงเรียนหอวัง
การศึกษาระดับมัธยมปลาย	โรงเรียนหอวัง

ภาคผนวก

Source Code

```
#include <8051IO.h>
#include <8051REG.h>

#define str1 "ON"
#define str2 "OFF"

register char ch,menu,a,b;

main()
{
        serinit(9600);
        delay (50);

        while(!(SCON & 0x01));

        while(1)
        {
                printf("\t\t\t@@Intenet Remote Controller@@\n\n");
                printf("\t\t\t ==========\n\n");
                printf("\n1.)Read   Switch StatusOn-Off ");
                printf("\n2.)Change Switch StatusOn-Off ");
                printf("\n\nPress Enter Number 1-2 :...");
                mainmenu();

        }
}
checkbit()
{
                printf("\nSwitch 1 ==> ");
                if(P1 & 0x04)
                        {
                                putstr(str2);
```

```
        }
else
        {
                putstr(str1);
        }
printf("\nSwitch 2 ==> ");
if(P1 & 0x08)
        {
                putstr(str2);
        }
else
        {
                putstr(str1);
        }
printf("\nSwitch 3 ==> ");
if(P1 & 0x10)
        {
                putstr(str2);
        }
else
        {
                putstr(str1);
        }
printf("\nSwitch 4 ==> ");
if(P1 & 0x20)
        {
                putstr(str2);
        }
else
        {
                putstr(str1);
        }
printf("\nSwitch 5 ==> ");
if(P1 & 0x40)
```

```
                {
                        putstr(str2);
                }
        else
                {
                        putstr(str1);
                }
        printf("\nSwitch 6 ==> ");
        if(P1 & 0x80)
                {
                        putstr(str2);
                }
        else
                {
                        putstr(str1);
                }

        printf("\n\n\nPress Anykey To Main Menu.... ");
        getch();
        printf("\n\n");

}
checkkey()
{
        ch = getch();
switch (ch)
        {
case '1':
        {
        printf("\n\n\n\t\t========== !!! OK,Switch 1 's Change!!! ==========\n\n\n\n");
        P1 ^= 0x04;
        break;
        }
case '2':
```

```
        {
        printf("\n\n\n\t\t========== !!! OK,Switch 2 's Change!!! ==========\n\n\n\n");
        P1 ^= 0x08;
        break;
        }
case '3':
        {
        printf("\n\n\n\t\t========== !!! OK,Switch 3 's Change!!! ==========\n\n\n\n");
        P1 ^= 0x10;
        break;
        }
case '4':
        {
        printf("\n\n\n\t\t========== !!! OK,Switch 4 's Change!!! ==========\n\n\n\n");
        P1 ^= 0x20;
        break;
        }
case '5':
        {
        printf("\n\n\n\t\t========== !!! OK,Switch 5 's Change!!! ==========\n\n\n\n");
        P1 ^= 0x40;
        break;
        }
case '6':
        {
        printf("\n\n\n\t\t========== !!! OK,Switch 6 's Change!!! ==========\n\n\n\n");
        P1 ^= 0x80;
        break;
        }
        default:
                {
                        printf("\nPress Enter Number 1-6 : ...");
                        checkkey();
                }
```

```
        }
}
mainmenu()
{
        menu = getch();
switch (menu)
        {
case '1':
        {
        printf("\n\n\n\t\t\t Read Switch StatusOn-Off");
        printf("\n\t\t\t ==========\n\n");
        checkbit();
        break;
        }
case '2':
        {
        printf("\n\n\n\t\t\tChange Switch StatusOn-Off \n");
        printf("\t\t\t ==========\n\n");
        printf("\nChoose Switch To Change Status -: \n");
        printf("\nSwitch 1 Press Number 1");
        printf("\nSwitch 2 Press Number 2");
        printf("\nSwitch 3 Press Number 3");
        printf("\nSwitch 4 Press Number 4");
        printf("\nSwitch 5 Press Number 5");
        printf("\nSwitch 6 Press Number 6");
        printf("\n\nPress Enter Number 1-6 : ...");
        checkkey();
        break;
        }
        default:
        {
        printf("\nPress Enter Number 1-2 :...");
        mainmenu();
        }
```

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง