ระบบเหมืองข้อมูลสำหรับวิธีการแบ่งประเภทของข้อมูลแบบเครือข่ายความเชื่อเบย์

DATA MINING SYSTEM FOR CLASSIFICATION USING BAYESIAN BELIEF NETWORKS

โดย

สมรศม รุ่งฟ้า

SAMORNSOM ROONGFAR

อาจารย์ที่ปรึกษา

รศ.ดร.วรพจน์ กรีสุระเดช

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ภาคเรียนที่ 2 ปีการศึกษา 2551

DATA MINING SYSTEM FOR CLASSIFICATION USING BAYESIAN BELIEF NETWORKS

SAMORNSOM ROONGFAR

A SYSTEM DEVELOPMENT PROJECT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE PROGRAM IN INFORMATION TECHNOLOGY

FACULTY OF INFORMATION TECNOLOGY

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2/ 2008

หัวข้อ	**ระบบเหมืองข้อมูลสำหรับวิธีการการแบ่งประเภทของข้อมูลแบบเครือข่ายความเชื่อเบย์**
นักศึกษา	**นางสาวสมรศม รุ่งฟ้า**
รหัสนักศึกษา	**49066802**
ปริญญา	**วิทยาศาสตรมหาบัณฑิต**
สาขาวิชา	**เทคโนโลยีสารสนเทศ**
แขนงวิชา	**วิทยาการสารสนเทศ**
ปีการศึกษา	**2551**
อาจารย์ที่ปรึกษา	**รศ.ดร.วรพจน์ กรีสุระเดช**

บทคัดย่อ

ในการทำเหมืองข้อมูลนั้น เป็นการดึงข้อมูลจากข้อมูลความรู้ที่เก็บไว้ในฐานข้อมูลขนาดใหญ่ , คลังสินค้า หรือที่เก็บข้อมูลอื่นๆที่มีขนาดใหญ่ แล้วนำข้อมูลนั้นมาใช้งานให้เกิดประโยชน์สูงสุด เพื่อค้นหาแนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในฐานข้อมูลที่มีขนาดใหญ่เหล่านั้น โดยอาศัยหลักการทางสถิติ การเรียนรู้ของเครื่อง และหลักการทางคณิตศาสตร์ เมื่อเราได้ข้อมูลที่น่าสนใจจากการทำเหมืองข้อมูลแล้ว เราสามารถนำข้อมูลที่เรามีอยู่ไปสร้างแบบจำลองในการทำนายได้

โครงการนี้เป็นการพัฒนาระบบเหมืองข้อมูลโดยใช้วิธีการแบ่งประเภทของข้อมูลแบบเครือข่ายความเชื่อเบย์ซึ่งเป็นเทคนิคหนึ่งของการทำเหมืองข้อมูลมาอธิบายความไม่ขึ้นต่อกันอย่างมีเงื่อนไขระหว่างตัวแปรได้ และระบบนี้สามารถสร้างแบบจำลองในการทำนายเพื่อช่วยในการวิเคราะห์และทำนายแนวโน้มของข้อมูลในอนาคตอย่างมีประสิทธิภาพได้

Title	Data Mining System for Classification using Bayesian Belief Networks
Student	Miss. Samornsom Roongfar
Student ID.	49066802
Degree	Master of Science
Programme	Information Science
Academic Year	2008
Advisor	Assoc. Prof. Dr. Worapoj Kreesuradej

ABSTRACT

Data mining is the automated or convenient extraction of patterns representing knowledge implicitly stored in large database , data warehouse and other massive information repositories . Data mining is a multidisciplinary field and we focus on issues relating to the feasibility , usefulness , efficiency and scalability of techniques for discovery of pattern hidden in large database by using statistics , machine learning, visualization , mathematics and information science . When we mined the data , we can use the mining data for creating model and the interesting model is classification using Bayesian Belief Network.

This project is made to develop data mining system for classification using Bayesian Belief Networks , data mining's technique. It uses condition independence between variables for creating the most efficiency model. And this data mining system can create model for analysis and predicate direction of the data in the future.

กิตติกรรมประกาศ

ในการพัฒนาระบบเหมืองข้อมูลสำหรับวิธีการแบ่งประเภทของข้อมูลแบบเครือข่ายความเชื่อเบย์สามารถดำเนินการและสำเร็จไปได้ด้วยดี ข้าพเจ้าขอขอบพระคุณ รศ.ดร. วรพจน์ กรีสุระเดช ผู้เป็นอาจารย์ที่ปรึกษาที่กรุณาให้คำปรึกษาและคำแนะนำ ข้อเสนอแนะ รวมทั้งวิธีการแก้ปัญหาต่างๆในการพัฒนาระบบ

นอกจากนี้ข้าพเจ้าขอขอบพระคุณ บิดา มารดา และครอบครัวของข้าพเจ้าที่ได้ให้ความสนับสนุนทางด้านกำลังใจและทางทุนทรัพย์จนทำให้โครงงานนี้สำเร็จไปได้ด้วยดี

ข้าพเจ้าหวังเป็นอย่างยิ่งว่าบทความนี้จะเป็นแนวคิดในการปฏิบัติงานเพื่อสามารถนำไปใช้ประยุกต์กับงานด้านอื่นๆได้เป็นอย่างดี

สมรศม รุ่งฟ้า

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

สารบัญรูป

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันการดำเนินงานต่างๆในแต่ละองค์กรมีการใช้และเก็บข้อมูลเป็นจำนวนมากในระบบฐานข้อมูล ดังนั้นจึงได้มีการนำเอาหลักการของการทำเหมืองข้อมูลมาประยุกต์ใช้ เพื่อช่วยในการวิเคราะห์และหาความเป็นไปได้ของข้อมูลจากระบบฐานข้อมูลที่มีปริมาณมากออกมาใช้ให้เกิดประโยชน์สูงสุดสำหรับแต่ละองค์กร และนำผลลัพธ์ที่ได้จากการทำเหมืองข้อมูลไปใช้งานให้เหมาะกับวัตถุประสงค์ต่างๆในอนาคตได้

เครือข่ายความเชื่อเบย์ใช้หลักการการแบ่งประเภทของข้อมูล และเป็นเทคนิคหนึ่งของการทำเหมืองข้อมูลที่ใช้ในการจำแนกกลุ่มข้อมูลด้วยลักษณะต่างๆที่ได้มีการกำหนดไว้แล้ว และสร้างแบบจำลองเพื่อการทำนายค่าข้อมูลในอนาคตได้ เช่น จัดกลุ่มนักเรียนว่า ดีมาก ดี ปานกลาง ไม่ดี โดยพิจารณาจากประวัติและผลการเรียน หรือแบ่งประเภทของลูกค้าว่าเชื่อถือได้หรือไม่ โดยพิจารณาจากข้อมูลที่มีอยู่

1.2 วัตถุประสงค์

1. เพื่อศึกษาหลักการและวิธีการทำเหมืองข้อมูล รวมทั้งเทคนิคการแบ่งประเภทของข้อมูลซึ่งเป็นเทคนิคหนึ่งของการทำเหมืองข้อมูล
2. เพื่อศึกษาหลักการทำงาน กระบวนการเรียนรู้ และการสร้างเครือข่ายของเครือข่ายความเชื่อเบย์
3. เพื่อพัฒนาระบบเหมืองข้อมูล โดยใช้เครือข่ายความเชื่อเบย์ในการแบ่งประเภทของข้อมูล เพื่อสร้างแบบจำลองที่มีประสิทธิภาพและสามารถทำนายค่าข้อมูลในอนาคตได้

1.3 ขอบเขตของการพัฒนา

1. ระบบเหมืองข้อมูลสามารถใช้เทคนิคการแบ่งประเภทของข้อมูลแบบเครือข่ายความเชื่อเบย์ได้
2. ระบบเหมืองข้อมูลสามารถทำงานร่วมกับ Microsoft SQL Server 2000 เพื่อเป็นฐานข้อมูลได้

3. ระบบเหมืองข้อมูลสามารถเลือกชุดข้อมูลจากระบบฐานข้อมูลตามต้องการ เพื่อใช้ในการแบ่งประเภทของข้อมูล

4. ระบบเหมืองข้อมูลสามารถแปลงรูปแบบข้อมูล เพื่อลดรูปและจัดข้อมูลให้อยู่ในรูปแบบเดียวกันได้

5. ระบบเหมืองข้อมูลสามารถสร้างแบบจำลองตามหลักการของเครือข่ายความเชื่อเบย์ได้อย่างมีประสิทธิภาพได้

1.4 ขั้นตอนและวิธีการดำเนินงาน

1. ศึกษาหลักการและวิธีการในการทำเหมืองข้อมูล รวมทั้งเทคนิคการแบ่งประเภทของข้อมูล

2. ศึกษาการหลักทำงาน และกระบวนการเรียนรู้ของเครือข่ายความเชื่อเบย์

3. ศึกษาวิธีการสร้างเครือข่ายและการคำนวณหาความน่าจะเป็นของแต่ละโหนดในเครือข่ายความเชื่อเบย์

4. ศึกษาวิธีการพัฒนาระบบเหมืองข้อมูลที่ใช้ในการแบ่งประเภทของข้อมูล

5. ศึกษาวิธีการทำงานของระบบเหมืองข้อมูลร่วมกับ Microsoft SQL Server 2000 ในส่วนของระบบฐานข้อมูล

6. ออกแบบและพัฒนาระบบเหมืองข้อมูลตามหลักการทำงานของเครือข่ายความเชื่อเบย์

7. ทดสอบและตรวจสอบข้อผิดพลาดต่างๆของระบบเหมืองข้อมูล เพื่อทำการปรับปรุงและแก้ไขให้สมบูรณ์

8. สรุปผลการทำงานของระบบเหมืองข้อมูล

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. สามารถเข้าใจหลักการและวิธีการของการทำเหมืองข้อมูล รวมทั้งเทคนิคการแบ่งประเภทของข้อมูล

2. สามารถเข้าใจหลักการทำงาน และกระบวนการเรียนรู้ของเครือข่ายความเชื่อเบย์

3. สามารถนำหลักการในการสร้างเครือข่ายและการคำนวณหาความน่าจะเป็นไปใช้ในระบบเหมืองข้อมูลการแบ่งประเภทของข้อมูลแบบเครือข่ายความเชื่อเบย์ได้

4. สามารถนำหลักการและวิธีการต่างๆที่ได้ศึกษามาประยุกต์ใช้ในการออกแบบและพัฒนาระบบเหมืองข้อมูลได้

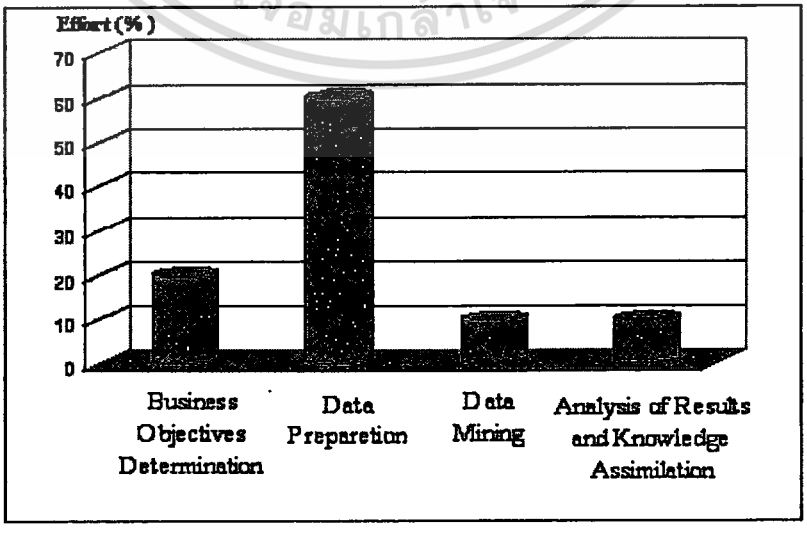
5. สามารถสร้างแบบจำลองและทำนายค่าข้อมูลในอนาคตที่มีประสิทธิภาพได้

บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

ปัจจุบันระบบฐานข้อมูลมีความสำคัญอย่างมาก จึงมีการหากลยุทธ์วิธีการต่างๆ เพื่อดึงเอาข้อมูลที่มีประโยชน์สูงสุดออกจากฐานข้อมูลมาใช้ ดังนั้นจึงเห็นความสำคัญของการทำเหมืองข้อมูล เพราะจะช่วยให้ค้นพบองค์ความรู้ใหม่ที่มีประโยชน์สูงสุดที่ซ่อนอยู่ในข้อมูลดิบในอดีตและสามารถทำนายสถานการณ์ในอนาคตออกมาได้ เพื่อช่วยในการตัดสินใจในการทำงานในแต่ละด้านให้เกิดประโยชน์สูงสุด โดยหัวข้อนี้จะกล่าวถึงทฤษฎีการทำเหมืองข้อมูลที่เกี่ยวข้องกับการพัฒนาระบบเหมืองข้อมูล ซึ่งเนื้อหาในบทนี้จะกล่าวถึง ความหมายของการทำเหมืองข้อมูล ขั้นตอนการทำเหมืองข้อมูล แบบจำลองการทำเหมืองข้อมูล และการสร้างแบบจำลองเพื่อการทำนาย โดยเนื้อหาทั้งหมดนี้ยังมีความสำคัญและจำเป็นต่อการศึกษาและการออกแบบพัฒนาระบบเหมืองข้อมูลอีกด้วย

2.1 ความหมายของการทำเหมืองข้อมูล

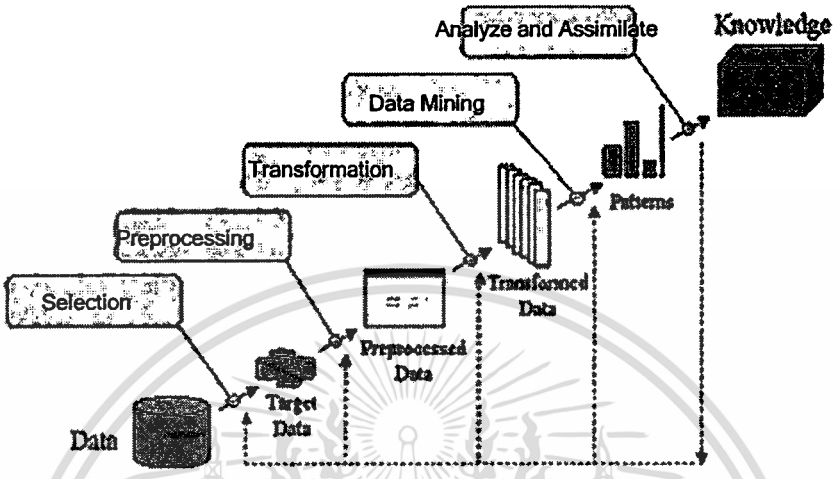
การทำเหมืองข้อมูล หมายถึง การดึงข้อมูลจากฐานข้อมูลที่มีขนาดใหญ่เพื่อนำข้อมูลนั้นมาใช้งานให้เกิดประโยชน์สูงสุด เพื่อค้นหารูปแบบ แนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในฐานข้อมูลที่มีขนาดใหญ่เหล่านั้น โดยอาศัยหลักการทางสถิติ การรู้จำ การเรียนรู้ของเครื่อง และหลักการทางคณิตศาสตร์ เพื่อช่วยในการตัดสินใจในการทำธุรกิจ และการทำนายค่าข้อมูลในอนาคตได้



รูปที่ 2.1 ขั้นตอนการทำเหมืองข้อมูล

2.2 ขั้นตอนการทำเหมืองข้อมูล

ในการทำเหมืองข้อมูลนั้นประกอบด้วย 6 ขั้นตอนและสามารถสรุปได้ดังต่อไปนี้ คือ



รูปที่ 2.2 ขั้นตอนการทำเหมืองข้อมูล

2.2.1 การกำหนดวัตถุประสงค์ทางธุรกิจ (Business Objective Determination)

การกำหนดวัตถุประสงค์ทางธุรกิจเป็นขั้นตอนที่ต้องทำเป็นอันดับแรก เพราะเป็นขั้นตอนที่กำหนดถึงผลลัพธ์หรือเทคนิคที่จะใช้ในการทำเหมืองข้อมูล โดยสิ่งที่ต้องทำคือ การกำหนดปัญหาและวัตถุประสงค์ทางธุรกิจให้ชัดเจน รวมทั้งการวิเคราะห์ข้อมูลทางธุรกิจเบื้องต้นว่ามีข้อมูลอะไรบ้างที่น่าสนใจ แล้วต้องการอะไรจากข้อมูลเหล่านั้น และเวลาไหนควรจะทำเหมืองข้อมูลในการแก้ปัญหาทางธุรกิจ

2.2.2 การเลือกและเตรียมข้อมูล (Data Selection and Preparation)

การเลือกและเตรียมข้อมูล เป็นขั้นตอนในการเลือกและจัดการข้อมูลให้ครบถ้วนสมบูรณ์เพื่อให้สามารถนำข้อมูลเหล่านั้นเข้าสู่อัลกอริทึมในการทำเหมืองข้อมูลได้ ซึ่งประกอบไปด้วย 3 ขั้นตอนย่อยดังนี้

- **การเลือกข้อมูล (Data Selection)**

การเลือกข้อมูล เป็นการเลือกชุดข้อมูลที่มีประโยชน์ต่อการทำเหมืองข้อมูล และตัดข้อมูลที่ไม่มีผลต่อการทำเหมืองข้อมูลออกไป โดยการเลือกข้อมูลนั้นจะแตกต่างกันไปตามวัตถุประสงค์ทางธุรกิจที่ได้กำหนดขึ้นไปแล้วตั้งแต่ต้นและขึ้นกับลักษณะของงานที่จะนำมาใช้งานอีกด้วย ซึ่งตัวแปรของข้อมูลที่ถูกเลือกมาใช้ในการทำเหมืองข้อมูลจะต้องมีการทำความเข้าใจว่าหมายถึงอะไร มีรูปแบบและลักษณะของข้อมูลเป็นอะไร ค่าที่เป็นได้ทั้งหมดของตัวแปรมีอะไรบ้าง

และควรมีคำอธิบายอย่างชัดเจน เกี่ยวกับชนิดของข้อมูล เป็นต้น

- **การประมวลผลข้อมูลเบื้องต้น (Data Preprocessing)**

จากขั้นตอนการเลือกข้อมูลนั้น ข้อมูลที่ถูกเลือกอาจยังไม่อยู่ในรูปแบบที่สามารถนำไปใช้งานได้เลย จึงต้องมีการประมวลผลข้อมูลเบื้องต้นก่อน เพื่อให้ข้อมูลนั้นมีคุณภาพ ถูกต้องและสมบูรณ์ โดยการประมวลผลข้อมูลเบื้องต้นสามารถทำได้ 3 วิธี

- การแก้ไขข้อมูล (Data Cleaning) โดยการแก้ไขข้อมูลสามารถทำได้ 2 ประเภท คือ การแก้ปัญหาค่าหรือข้อมูลที่เป็นค่าว่าง (Missing Value) หมายถึง การที่ข้อมูลขาดหายไปในบางช่วงของข้อมูลที่นำมาใช้งาน และการแก้ปัญหาค่าหรือข้อมูลที่มีความคลาดเคลื่อน (Noisy Data) หมายถึง การที่ข้อมูลแตกต่างไปจากค่าข้อมูลที่เป็นไปได้หรือที่คาดการณ์ไว้

- การรวมข้อมูล (Data Integration) เป็นการนำเอาข้อมูลจากแหล่งข้อมูลต่างๆที่มีความหมายเหมือนกันมารวมกัน เพื่อทำให้จำนวนข้อมูลลดลงและไม่เปลืองทรัพยากรในการทำงาน

- การลดข้อมูล (Data Reduction) เป็นการลดข้อมูลเพื่อประหยัดเวลาและทรัพยากร โดยสามารถทำได้ 2 วิธี คือ เทคนิคการลดมิติ (Dimensionality Reduction) หมายถึง การลดตัวแปรที่ไม่ส่งผลต่อการทำเหมืองข้อมูล และเทคนิคการลดขนาดของข้อมูล (Size Reduction) หมายถึง การลดขนาดของข้อมูลโดยใช้วิธีการสุ่มตัวอย่าง โดยเลือกชุดข้อมูลเฉพาะชุดข้อมูลที่จะนำมาเป็นกลุ่มตัวอย่าง

- **การแปลงค่าข้อมูล (Data Transformation)**

การแปลงค่าข้อมูล เป็นการแปลงข้อมูลที่ไม่เหมาะสมหรือยังไม่อยู่ในรูปแบบที่จะนำไปใช้งานได้ให้อยู่ในรูปแบบที่พร้อมจะทำเหมืองข้อมูล โดยปรับขอบเขตของข้อมูลให้อยู่ในช่วงที่เหมาะสมต่อการใช้งานในการสร้างแบบจำลอง และปรับให้อยู่ในลักษณะที่สามารถทำงานร่วมกับอัลกอริทึมในการทำเหมืองข้อมูลได้อีกด้วย

2.2.3 การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล เป็นการเลือกอัลกอริทึมและแบบจำลองที่จะใช้ในการทำเหมืองข้อมูลเพื่อให้ได้ผลลัพธ์ที่ตรงกับวัตถุประสงค์ทางธุรกิจหรือปัญหาที่ต้องการแก้ไข ซึ่งในแต่ละปัญหาอาจจะมีหลายวิธี ดังนั้นจึงต้องมีการเลือกอัลกอริทึมและแบบจำลองที่เหมาะสมด้วย เพื่อจะได้ผลลัพธ์ตามต้องการ

2.2.4 การวิเคราะห์ผลลัพธ์ (Analysis of Results)

การวิเคราะห์ผลลัพธ์ เป็นการวิเคราะห์ผลการประมวลผล ซึ่งจะทำการแปลความหมายและประเมินผลลัพธ์ที่ได้จากการทำเหมืองข้อมูล ว่าสามารถนำแบบจำลองไปใช้แล้วเหมาะสมและตรงกับวัตถุประสงค์ตามที่ได้กำหนดไว้หรือไม่ รวมทั้งการประเมินความถูกต้องของการทำเหมือง

ข้อมูลอีกด้วย เพราะบางครั้งผลที่ได้จากการทำเหมืองข้อมูลนั้นอาจมีความผิดพลาดเกิดขึ้น และนำไปสู่การแก้ไขปรับปรุงในขั้นตอนที่ผ่านมา ก่อนนำไปใช้งานจริง

2.2.5 การนำความรู้ที่ได้ไปประยุกต์ใช้ (Assimilation of Knowledge)

การนำความรู้ที่ได้ไปประยุกต์ใช้ เป็นการนำผลลัพธ์ที่ได้จากขั้นตอนการวิเคราะห์ผลลัพธ์มาประยุกต์ใช้กับงานต่างๆให้เกิดประโยชน์สูงสุด

2.3 แบบจำลองการทำเหมืองข้อมูล

เราสามารถจำแนกแบบจำลองที่ใช้ในการทำเหมืองข้อมูลออกเป็น 2 ประเภท คือ

2.3.1 การสร้างแบบจำลองในการทำนาย (Predictive Modeling)

แบบจำลองนี้เป็นการคาดคะเนลักษณะ หรือการประมาณค่าที่ชัดเจนของข้อมูลที่จะเกิดขึ้นโดยใช้พื้นฐานจากข้อมูลที่ผ่านมาในอดีต และจะมุ่งเน้นในเรื่องของการแบ่งแยกข้อมูลออกเป็นกลุ่มตามคุณสมบัติ

ตัวอย่างเทคนิคการทำเหมืองข้อมูลที่ใช้แบบจำลองนี้ เช่น การแบ่งประเภทของข้อมูล (Classification) ซึ่งเป็นการจัดกลุ่มให้กับแต่ละข้อมูลในฐานข้อมูล โดยมีการระบุค่าหรือลักษณะที่เป็นไปได้ของข้อมูลภายในแต่ละกลุ่ม และการถดถอย (Regression) เป็นต้น

2.3.2 การสร้างแบบจำลองในการบรรยาย(Descriptive Modeling)

แบบจำลองนี้เป็นการอธิบายลักษณะบางอย่างของข้อมูลที่มีอยู่ ซึ่งโดยส่วนมากเป็นลักษณะการแบ่งกลุ่มให้กับข้อมูลซึ่งไม่ได้มีจุดมุ่งหมายเพื่อการทำนาย

ตัวอย่างเทคนิคการทำเหมืองข้อมูลที่ใช้แบบจำลองนี้ เช่น การหาความสัมพันธ์ระหว่างข้อมูล (Association) ซึ่งเป็นการวิเคราะห์ข้อมูลที่เกิดขึ้นพร้อมกันภายในกลุ่มข้อมูลเดียวกัน และการจัดกลุ่มข้อมูล (Clustering) เป็นต้น

2.4 ขั้นตอนในการสร้างแบบจำลอง

แบบจำลองเพื่อการทำนายนั้น เป็นการคาดคะเนถึงความเป็นไปได้ โดยการสังเกตกลุ่มของข้อมูลที่มีอยู่ แล้วจึงนำกลุ่มข้อมูลที่ต้องการไปวิเคราะห์ ในการสร้างแบบจำลองนั้นแบ่งออกเป็น 2 ระยะ คือ

2.4.1 ระยะการเรียนรู้ (Training Phase)

เป็นระยะที่ทำการสร้างแบบจำลองขึ้นมาใหม่ โดยใช้ข้อมูลที่ผ่านมาในอดีต

2.4.2 ระยะการทดสอบ (Testing Phase)

เป็นระยะที่ทำการทดสอบแบบจำลองที่สร้างขึ้นมา โดยใช้ข้อมูลใหม่ที่ไม่เคยเห็นมาก่อนเพื่อตรวจสอบความแม่นยำของแบบจำลอง

2.5 แนวโน้มและการประยุกต์การใช้งานการทำเหมืองข้อมูล

เนื่องจากในปัจจุบันมีการนำหลักการและเทคนิคของการทำเหมืองข้อมูลมาใช้กันอย่างแพร่หลาย ดังนั้น จึงมีการค้นคว้าวิจัยและพัฒนาเพื่อประยุกต์ใช้กับงานหลายๆด้าน โดยตัวอย่างการประยุกต์ใช้งานที่น่าสนใจในปัจจุบัน เช่น

2.5.1 การใช้งานด้านการแพทย์ (Biomedical and DNA Data Analysis)

ส่วนมากเป็นการนำไปใช้ในการวิเคราะห์รูปแบบการจัดเรียงตัวของหน่วยพันธุกรรม เพื่อหาสาเหตุของความผิดปกติที่ทำให้เกิดโรค ความสัมพันธ์ของรูปแบบการจัดเรียงตัวของหน่วยพันธุกรรมกับระดับความรุนแรงของโรค รวมถึงการใช้ในด้านการวินิจฉัยโรค การป้องกัน และการรักษาด้วย

2.5.2 การใช้งานเพื่อการวิเคราะห์ด้านการเงิน (Financial Analysis)

เป็นงานที่เกี่ยวข้องกับบริษัทเงินทุนหรือธนาคารต่างๆ เช่น การวิเคราะห์การให้สินเชื่อ การทำนายอัตราการจ่ายเงินกู้ การแบ่งกลุ่มลูกค้าเพื่อหาเป้าหมายทางการตลาด

2.5.3 การใช้งานด้านการขาย (Retail Industry)

เป็นงานที่ใช้ในการหากลยุทธ์ที่ทำให้เกิดการได้เปรียบคู่แข่งทางการค้า เช่น การหาลักษณะการซื้อของลูกค้า ความสัมพันธ์ระหว่างการซื้อและช่วงเวลา ความสัมพันธ์ระหว่างตัวสินค้าและการวิเคราะห์ประสิทธิภาพของการโฆษณาซึ่งสิ่งต่างๆเหล่านี้ช่วยให้สามารถหาวิธีการตอบสนองความต้องการของลูกค้าได้มากที่สุด และอาจหมายถึงส่วนแบ่งทางการตลาดที่เพิ่มขึ้นด้วย

2.5.4 การใช้งานด้านโทรคมนาคม (Telecommunication Industry)

เพื่อสนับสนุนการให้บริการด้านการติดต่อสื่อสารของลูกค้า เช่น การวิเคราะห์ลักษณะการใช้บริการด้านการติดต่อสื่อสาร การหาความสัมพันธ์ของการใช้บริการกับช่วงเวลา หรือ การตรวจจับรูปแบบที่ผิดปกติในระบบการติดต่อสื่อสาร เป็นต้น

2.6 เทคนิคการทำเหมืองข้อมูล : เครือข่ายความเชื่อเบย์ (Bayesian Belief Networks)

- **ความหมายของเครือข่ายความเชื่อเบย์**

เครือข่ายความเชื่อเบย์ หรือเรียกโดยย่อว่า เครือข่ายเบย์(Bayes Net) เป็นวิธีการเรียนรู้ที่ลดข้อจำกัดของการเรียนรู้เบย์อย่างง่าย(Naïve Bayes) ในสมมติฐานของความไม่ขึ้นต่อกันระหว่างคุณสมบัติ ซึ่งในความเป็นจริงแล้วเราพบว่าคุณสมบัติบางตัวจะขึ้นต่อกันบ้าง และควรที่จะนำความขึ้นต่อกันนี้เข้ามาใส่ไว้ในแบบจำลองด้วย เราจึงใช้เครือข่ายความเชื่อเบย์ในการอธิบายความไม่ขึ้นต่อกันอย่างมีเงื่อนไข (condition independent) ระหว่างตัวแปร (ในบริบทของเครือข่ายความเชื่อเบย์นิยมใช้คำว่า "ตัวแปร"(variables) แทนคำว่า "คุณสมบัติ")

ในแบบจำลองนี้เราสามารถใช้

1. ความรู้ก่อนหน้า(prior knowledge) เกี่ยวกับความ(ไม่)ขึ้นต่อกันระหว่างตัวแปร ร่วมกับ

2. ตัวอย่างสอน เพื่อทำให้กระบวนการเรียนรู้มีประสิทธิภาพ โดยเราสามารถใส่ความรู้ก่อนหน้าในเครือข่ายความเชื่อเบย์ให้อยู่ในรูปของโครงสร้างเครือข่าย และตารางความน่าจะเป็นแบบมีเงื่อนไข ดังจะกล่าวต่อไป

- **นิยามความไม่ขึ้นต่อกันอย่างมีเงื่อนไข**

X ไม่ขึ้นกับ Y อย่างมีเงื่อนไขเมื่อรู้ Z ถ้าความน่าจะเป็นของ X ไม่ขึ้นกับค่าของ Y เมื่อรู้ค่า Z นั่นคือ

$$(\forall x_i , y_j , z_k)\ P(X=x_i|Y=y_j ,Z=z_k) = P(X=x_i|Z=z_k) \qquad (2.1)$$

หรือในรูปง่าย

$$P(X|Y,Z) = P(X|Z) \qquad (2.2)$$

นิยามจากสมการ (2.1) หมายความว่า สำหรับ x_i ,y_j ,z_k ใดๆ ความน่าจะเป็นที่ X จะมีค่าเป็น x_i เมื่อรู้ว่า Y มีค่าเป็น y_j และ Z มีค่าเป็น z_k จะมีค่าเท่ากับความน่าจะเป็นของ X จะมีค่าเป็น x_i เมื่อรู้ว่า Z มีค่าเป็น z_k ในกรณีที่ความน่าจะเป็นทั้งสองเท่ากันเช่นนี้ เราเรียกว่าค่าของ X ไม่ขึ้นกับค่าของ Y อย่างมีเงื่อนไข เมื่อรู้ค่าของ Z เราจึงสามารถตัด Y ทิ้งไปได้

เช่น ฟ้าร้องไม่ขึ้นกับฝนตก ถ้ารู้ว่าฟ้าแลบ

$$P(\text{Thunder}|\text{Rain} ,\text{Lighting}) = P(\text{Thunder}|\text{Lighting})$$

ดังนั้น ถ้ามีฟ้าแลบสามารถบอกได้เลยว่า จะต้องได้ยินเสียงฟ้าร้องด้วยความน่าจะเป็นเท่าไร โดยไม่ต้องสนใจว่าเกิดฝนตกหรือไม่

- **ส่วนประกอบของเครือข่ายความเชื่อเบย์**

เครือข่ายความเชื่อเบย์ประกอบด้วย 2 ส่วน คือ

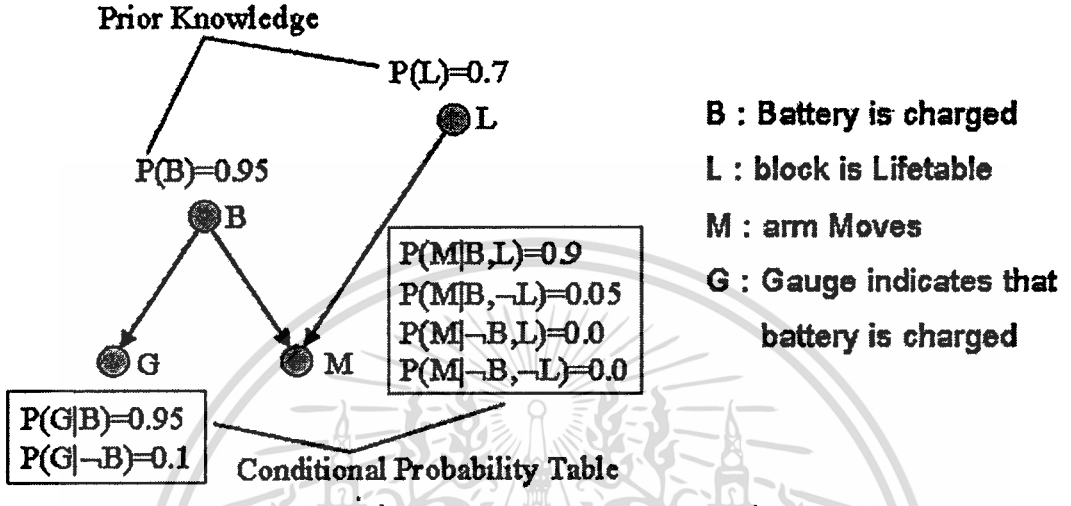
1) กราฟอวัฏจักรระบุทิศทาง (Directed Acyclic Graph)

เป็นเครือข่ายที่ประกอบด้วยโหนดหลายโหนด โดยแต่ละโหนดหมายถึงคุณสมบัติของข้อมูลหรือตัวแปรที่ไม่ขึ้นต่อกันอย่างมีเงื่อนไข(conditional independenct)กับโหนดอื่น เมื่อรู้โหนดพ่อแม่โดยตรง(immediate predecessors)

โดยเครือข่ายความเชื่อเบย์จะแสดงเครือข่ายในรูปของกราฟแบบมีทิศทางซึ่งสามารถบอกได้ว่า มีตัวแปรใดบ้างที่ขึ้นกับตัวแปรอื่น และตัวแปรใดบ้างที่ไม่ขึ้นกับตัวอื่น

2) ตารางความน่าจะเป็นแบบมีเงื่อนไข (Conditional Probability Table : CPT)

แต่ละโหนดจะมีตารางความน่าจะเป็นแบบมีเงื่อนไข โดยการเขียนแทนค่าของตัวแปรอย่างง่าย โดยใช้ตัวแปรนั้นแทนค่าจริง และใส่เครื่องหมาย ¬ แทนค่าเท็จ เช่น G แทนค่าตัวแปร G เป็นจริง ส่วน ¬G แทนค่าตัวแปรG เป็นเท็จ



รูปที่ 2.3 ตัวอย่างของเครือข่ายความเชื่อเบย์

จากรูป 2.3 จะเห็นได้ว่าเครือข่ายความเชื่อเบย์ประกอบด้วยโหนดหลายโหนด แต่ละโหนดหมายถึงคุณสมบัติของข้อมูลหรือตัวแปร และโหนดแต่ละโหนดจะมีตารางความน่าจะเป็นแบบมีเงื่อนไข – ซีพีที (Conditional Probability Table - CPT) ติดอยู่ด้วย เครือข่ายความเชื่อเบย์นี้แสดงในรูปของกราฟซึ่งสามารถบอกได้ว่า มีตัวแปรใดบ้างที่ขึ้นกับตัวแปรอื่น และตัวแปรใดบ้างที่ไม่ขึ้นกับตัวอื่น

จากเครือข่ายความเชื่อเบย์ข้างต้น สมมติว่าเรากำลังเขียนเครือข่ายที่อธิบายหุ่นยนต์ตัวหนึ่งที่กำลังจะย้ายของในโดเมนโลกของบล็อก หุ่นยนต์ตัวนี้จะชาร์จแบตเตอรีและมีเกจ (G) คอยวัดว่าขณะนี้แบตเตอรีเหลืออยู่หรือไม่ หุ่นยนต์ทำงานด้วยการเคลื่อนแขนไปยกบล็อก เมื่อเราจำลองเหตุการณ์นี้ในเครือข่ายความเชื่อเบย์ จะได้ว่าแบตเตอรี (B) จะส่งผลต่อเกจ (G) นอกจากนั้นยังส่งผลต่อการเคลื่อนแขนของหุ่นยนต์ (M) และเราได้ใส่ความรู้ก่อนหน้าเข้าไปในรูปของตารางความน่าจะเป็นแบบมีเงื่อนไขว่า 70% ของบล็อกทั้งหมดสามารถยกได้ (P(L)=0.7) และในเวลา 1 ชั่วโมง มี 95 ชั่วโมงที่แบตเตอรีมีไฟ (P(B)=0.95)

เมื่อดูซีพีทีของโหนด G พบว่า ถ้าแบตเตอรีมีไฟ เกจซึ่งมีความบกพร่องอยู่บ้างนี้จะแสดงผลว่ามีไฟด้วยความน่าจะเป็นเท่ากับ 0.95 (P(G|B)=0.95) และถ้าไฟหมดแต่เกจยังแสดงว่ามีไฟด้วยความน่าจะเป็นเท่ากับ 0.1 (P(G|¬ B)=0.1)

เมื่อดูซีพีทีของโหนด M จะเห็นได้ว่า P(M|B,L)=0.9 หมายความว่า หุ่นยนต์จะเคลื่อนแขนถ้าแบตเตอรีมีไฟและบล็อกสามารถยกได้ และถ้ามีไฟแต่บล็อกไม่สามารถยกได้ แขนจะเคลื่อนด้วย

ความน่าจะเป็น 0.05 (P(M|B, ¬L)=0.9) และถ้าไม่มีไฟและบล็อกสามารถยกได้ หุ่นยนต์จะไม่เคลื่อนแขน (P(M|¬B, L)=0.0) และสุดท้ายถ้าบล็อกยกไม่ได้และไฟไม่มี แขนจะไม่เคลื่อนเช่นกัน (P(M|¬B, ¬L)=0.0)

ทั้งหมดนี้คือความน่าจะเป็นทั้งหมดที่เราป้อนให้กับระบบในรูปของซีพีที ผู้ที่ป้อนข้อมูลคือผู้เชี่ยวชาญที่ทำงานเกี่ยวกับหุ่นยนต์ เมื่อเราทราบค่าต่างๆทั้งหมดเราก็สามารถที่จะคำนวณความน่าจะเป็นต่างๆที่เกิดขึ้นภายในระบบได้ เช่น ถ้าต้องการคำนวณหาว่า ความน่าจะเป็นที่แบตเตอรีมีไฟ บล็อกสามารถยกได้ เกจขึ้นและหุ่นยนต์เคลื่อนแขน ทั้งสี่เหตุการณ์เกิดขึ้นพร้อมกันว่ามีค่าเท่าไรก็สามารถคำนวณได้จากเครือข่ายความเชื่อเบย์นี้

- **ความน่าจะเป็นร่วม (Joint Probability)**

ความน่าจะเป็นร่วมระหว่างตัวแปร คือ ความน่าจะเป็นที่ตัวแปรหลายตัวจะมีค่าตามที่กำหนด เช่น P(Battery, Liftable, Gauge, Move) เราสามารถเขียนความน่าจะเป็นร่วมให้อยู่ในรูปทั่วไปได้ดังนี้

$$P(y_1,\ldots,y_n) = \prod_{i=1}^{n} P(y_i|Parents(Y_i)) \qquad (2.3)$$

จากสมการ 2.3 $Parents(Y_i)$ หมายถึง โหนดพ่อแม่โดยตรงของโหนด Y_i ถ้าเราต้องการจะหาความน่าจะเป็นที่ $y_1,\ldots,y_n$ เกิดขึ้นพร้อมกันสามารถคำนวณได้จากความน่าจะเป็นของ y_1 คูณกับความน่าจะเป็นของ y_2 คูณไปเรื่อยๆจนถึง y_n แต่ต้องดูว่าโหนดแต่ละโหนดขึ้นกับโหนดพ่อแม่ตัวใดบ้าง เช่น y_1 ขึ้นกับโหนดใด , y_2 ขึ้นกับโหนดใด เป็นต้น จากตัวอย่างจะได้

$$\begin{aligned} P(G,M,B,L) &= P(G|B,M,L)P(M|B,L)P(B|L)P(L) \\ &= P(G|B)P(M|B,L)P(B)P(L) \\ &= (0.95)(0.9)(0.95)(0.7) \\ &= 0.57 \end{aligned}$$

สังเกตได้ว่าบรรทัดแรกใช้กฎลูกโซ่กระจาย P(G,M,B,L) ออกมาเป็นด้านขวามือ และเมื่อกระจายแล้วจะเห็นว่าตัวแปรบางตัวไม่ขึ้นกับตัวอื่น ดังนั้นจึงลดรูปลงมาเหลือแค่ตัวแปรที่ขึ้นต่อกันเท่านั้น สังเกตได้ว่า เมื่อลดรูปลงมาแล้วโหนดที่เราสนใจจะขึ้นกับพ่อแม่ของมันเท่านั้น เช่น P(G|B,M,L) ลดรูปลงเหลือ P(G|B) หรือหาความน่าจะเป็นของ G เมื่อรู้ B กรณีตัวอย่างที่ยกมาเป็นกรณีง่ายๆ เพราะเรารู้ค่าความน่าจะเป็นครบทั้งสี่ตัวแล้ว แต่ในบางกรณีเราอาจทราบค่าของตัวแปรเพียงแค่ 2 ตัวหรือ 3 ตัวไม่ใช่ทั้งหมดก็สามารถใช้เทคนิคในการอนุมานของเครือข่ายความเชื่อเบย์เพื่อหาความน่าจะเป็นร่วมได้เช่นกัน

- **ทฤษฎีของเครือข่ายความเชื่อเบย์**

ในกรณีของการเรียนรู้ของเครื่องนั้น สิ่งที่เราสนใจคือ เมื่อเรามีชุดข้อมูลหรือเซตของตัวอย่างสอน D เราต้องการหาความน่าจะเป็นที่สมมติฐาน(h) ที่เราสนใจมีโอกาสจะเกิดขึ้นเท่าไร เราก็สามารถใช้ทฤษฎีของเบย์ในการคำนวณได้ดังนี้

$$P(h|D) = \frac{P(D|h)\,P(h)}{P(D)} \qquad (2.4)$$

โดยที่ P(h|D)คือ ความน่าจะเป็นของ h เมื่อรู้ D

P(D|h) คือ ความน่าจะเป็นของ D เมื่อรู้ h

P(h) คือ ความน่าจะเป็นก่อนหน้าของสมมติฐาน h

P(D) คือ ความน่าจะเป็นก่อนหน้าของเซตตัวอย่าง D

เราเรียก P(h) ว่าเป็นความน่าจะเป็นก่อน(prior probability)ซึ่งเป็นความน่าจะเป็นที่สมมติฐาน h จะเป็นจริงโดยที่เรายังไม่ได้ดูข้อมูลตัวอย่างสอน

เราเรียก P(h|D) ว่าเป็นความน่าจะเป็นภายหลัง(post probability)ซึ่งเป็นความน่าจะเป็นที่สมมติฐานh จะเป็นจริงโดยมีเงื่อนไขว่า D เป็นจริง(เราเห็นข้อมูลตัวอย่างสอน D แล้ว) โดยความน่าจะเป็นก่อนเป็นค่าที่ได้จากข้อมูลเบื้องต้น ส่วนความน่าจะเป็นภายหลังเป็นค่าความน่าจะเป็นก่อนที่ถูกปรับด้วยข้อมูลที่เพิ่มขึ้น

จะเห็นได้ว่าการใช้ทฤษฎีของเบย์สามารถใช้คำนวณความน่าจะเป็นของสมมติฐานแต่ละตัวได้ เมื่อรู้เซตตัวอย่างสอนเป็นจริงซึ่งจะช่วยให้เราเลือกสมมติฐานที่ดีที่สุดได้ โดยเราเรียกสมมติฐานที่ดีที่สุดว่า สมมติฐานภายหลังมากสุด – เอ็มเอพี(Maximum A Posterior hypothesis - MAP) มีนิยามดังนี้

$$h_{MAP} = \underset{h \in H}{\arg\max}\, P(h|D) \qquad (2.5)$$

จากทฤษฎีของเบย์จะได้

$$h_{MAP} = \underset{h \in H}{\arg\max}\, \frac{P(D|h)\,P(h)}{P(D)} \qquad (2.6)$$

เนื่องจากสำหรับ $h \in H$ ทุกตัวมีค่า P(D) เท่ากันหมด ดังนั้น เราจึงสามารถละ P(D)ได้ ดังนั้น h ที่ดีที่สุดตามเอ็มพีเอ คือ h ที่ทำให้ค่า P(D|h) P(h) มีค่าสูงสุด แต่เทคนิคการเรียนรู้ของเครื่องหลายวิธีไม่ได้หาค่า h_{MAP} แต่มักหาค่า h_{ML}(Maximum Likelihood Hypothesis) ซึ่งหมายถึง

สมมติฐานที่ตรงหรือสอดคล้องกับข้อมูลสอนมากสุดจะเป็นสมมติฐานที่ดีที่สุดโดยไม่ได้พิจารณาความน่าจะเป็นก่อน ดังสมการนี้

$$h_{ML} = \underset{h \in H}{\arg\max}\, P(D|h) \qquad (2.7)$$

ในกรณีที่กำหนดให้เราใช้สมมติฐานได้เพียงข้อเดียว ในการจำแนกประเภทของตัวอย่าง จะได้ว่า h_{MAP} เป็นสมมติฐานที่ดีที่สุด แต่การจำแนกประเภทของตัวอย่างด้วย h_{MAP} ไม่ใช่การจำแนกประเภทที่น่าจะเป็นที่สุด(most probable classification) สำหรับตัวอย่างนั้น ในบางกรณีที่เราใช้สมมติฐานหลายข้อ เราสามารถจำแนกประเภทของตัวอย่างได้ดีกว่าการใช้ h_{MAP} เพียงตัวเดียว เช่น สมมติว่าเรามีสมมติฐาน 3 ข้อ แต่ละข้อมีค่าความน่าจะเป็นภายหลัง ดังต่อไปนี้

$P(h_1|D) = 0.4 \quad P(h_2|D) = 0.3 \quad P(h_3|D) = 0.3$

และเมื่อให้ตัวอย่าง x ผลการจำแนกประเภทของสมมติฐานเป็นดังนี้

$$h_1(x) = + \qquad h_2(x) = - \qquad h_3(x) = -$$

ในกรณีนี้เราควรจะจำแนกประเภทของ x เป็นบวกหรือลบ ซึ่งถ้าใช้ h_{MAP} ก็จะได้ว่า h_1 เป็นสมมติฐานที่ดีที่สุด เนื่องจาก h_1 มีค่าความน่าจะเป็นภายหลังมากที่สุด แต่เมื่อพิจารณาสมมติฐานอื่นในปริภูมิของสมมติฐาน เราพบว่า h_{MAP} ให้คำตอบเป็น + เพียงตัวเดียว แต่สมมติฐานอีกสองตัวให้คำตอบเป็น – เราจะได้การจำแนกประเภทที่น่าจะเป็นที่สุดในแบบของเบย์มีสูตรการคำนวณดังนี้

$$\underset{v_i \in V}{\arg\max} \sum_{h_i \in H} P(v_i|h_i)\, P(h_i|D) \qquad (2.8)$$

โดยที่ v เป็นเซตของค่า(ประเภท)ของตัวอย่าง

H เป็นปริภูมิของสมมติฐาน

จากตัวอย่างด้านบนเราจะได้ว่า

$P(h_1|D) = 0.4 \quad P(-|h_1) = 0.0 \quad P(+|h_1) = 1.0$

$P(h_2|D) = 0.4 \quad P(-|h_1) = 1.0 \quad P(+|h_1) = 0.0$

$P(h_3|D) = 0.4 \quad P(-|h_1) = 1.0 \quad P(+|h_1) = 0.0$

ทำให้ได้ค่าความน่าจะเป็นของประเภท + และ –

$$\sum_{h_i \in H} P(+|h_i)\, P(h_i|D) = 0.4$$

$$\sum_{h_i \in H} P(+|h_i)\, P(h_i|D) = 0.6$$

ดังนั้น

$$\underset{v_i \in V}{\arg\max} \sum_{h_i \in H} P(v_i|h_i)\, P(h_i|D) = -$$

- **เทคนิคการอนุมานสำหรับเครือข่ายความเชื่อเบย์**

1) การอนุมานจากเหตุ (Causal Reasoning)

เมื่อเราทราบเหตุ เราสามารถหาได้ว่าผลจะเกิดขึ้นด้วยความน่าจะเป็นเท่าไร

2) การอนุมานจากผล (Diagnosis Reasoning)

เทคนิคนี้จะตรงข้ามกับเทคนิคแรก กล่าวคือ เราทราบผลแล้ว แต่อยากทราบว่าสาเหตุจะเกิดขึ้นด้วยความน่าจะเป็นเท่าไร

3) การอธิบายลดความเป็นไปได้ (Explaining Away)

เป็นการทำอนุมานจากเหตุ(Causal Reasoning)ภายในการอนุมานจากผล (Diagnosis Reasoning) เป็นการผสมระหว่างวิธีการทั้งสองแบบข้างต้น

- **การเรียนรู้เครือข่ายความเชื่อเบย์ (Bayesian Belief Network Learning)**

การเรียนรู้แบบเบย์(Bayesian Learning) เป็นวิธีการเรียนรู้ที่ใช้ทฤษฎีความน่าจะเป็นซึ่งมีพื้นฐานมาจากทฤษฎีของเบย์ (Bayes Theorem) เข้ามาช่วยในการเรียนรู้ จุดมุ่งหมายก็เพื่อต้องการสร้างแบบจำลองที่อยู่ในรูปของความน่าจะเป็น ซึ่งเป็นค่าที่บันทึกได้จากการสังเกต จากนั้นนำแบบจำลองมาหาว่าสมมติฐานใดถูกต้องที่สุดโดยใช้ความน่าจะเป็นเข้ามาช่วย โดยเราจะใช้ความรู้ก่อนหน้า(prior knowledge) ร่วมกับข้อมูล(data) มาช่วยในการเรียนรู้เพื่อหาสมมติฐานที่ดีที่สุด

ความรู้ก่อนหน้า หมายถึง ความรู้ที่เรามีเกี่ยวกับสมมติฐานแต่ละตัวก่อนที่เราจะเก็บข้อมูล เมื่อใช้งานเราจะนำความน่าจะเป็นของข้อมูลที่เก็บได้มาปรับสมมติฐานซ้ำอีกครั้ง ซึ่งพบว่าวิธีนี้ให้ประสิทธิภาพในการเรียนรู้ได้ดีไม่ด้อยกว่าวิธีการเรียนรู้ประเภทอื่น โดยมีเทคนิคที่ใช้งานจริงได้ดี เช่น Naïve Bayes Learning และ Bayesian Belief Network Learning เป็นต้น

การเรียนรู้เครือข่ายความเชื่อเบย์ มีขั้นตอนในการทำงานดังนี้ คือ การหาโครงสร้างเครือข่ายและ/หรือตารางความน่าจะเป็นแบบมีเงื่อนไข (CPT) ที่สอดคล้องกับตัวอย่างสอนมากที่สุด โดยเครือข่ายความเชื่อเบย์มีปัญหาในการเรียนรู้ดังนี้ คือ

1) โครงสร้างไม่รู้ (Structure Unknown)

เป็นกรณียากที่สุด เพราะเราไม่รู้ว่าโครงสร้างของเครือข่ายความเชื่อเบย์มีรูปร่างเป็นอย่างไร มีการเชื่อมต่อระหว่างโหนดอย่างไร และเราไม่รู้ค่าในตารางความน่าจะเป็นแบบมีเงื่อนไขอีกด้วย ดังนั้น การเรียนรู้ต้องคำนวณหาทั้งโครงสร้างเครือข่ายและตารางความน่าจะเป็นแบบมีเงื่อนไข โดยอัลกอริทึมที่ใช้ในการสร้างเครือข่ายความเชื่อเบย์มีหลายวิธี แต่วิธีที่จะนำมาใช้คือ อัลกอริทึมเคสอง การสร้างเครือข่ายความเชื่อเบย์นั้น

- **อัลกอริทึมเคสอง (K2 Algorithm)**

อัลกอริทึมเคสองนั้นเป็นอัลกอริทึมที่ใช้ในการสร้างเครือข่ายความเชื่อเบย์แบบไม่รู้โครงสร้าง ซึ่งอัลกริทึมนี้เป็นวิธีการหนึ่งของการค้นหาแบบฮิวริสติก (Heuristic-Search Method) โดยจะคำนวณหาค่าความน่าจะเป็นของแต่ละโหนดทุกๆโหนด หากโหนดไหนของโหนดที่เรา

สนใจมีค่ามากที่สุดจะเป็นเป็นโหนดแม่ และโหนดที่เราสนใจอยู่เป็นโหนดลูก แสดงว่าโหนดลูกนั้นขึ้นอยู่กับโหนดแม่โหนดนั้น

ตารางที่ 2.1 อัลกอริทึมเคสองที่ใช้ในการสร้างเครือข่ายความเชื่อเบย์

```
1. procedure K2;
2. {Input: A set of n nodes, an ordering on the nodes, an upper bound u on the
3.       number of parents a node may have, and a database D containing m cases.}
4. {Output: For each node, a printout of the parents of the node.}
5. for i:= 1 to n do
6.       π_i := Ø;
7.       P_old := f(i, π_i);  {This function is computed using Equation 2.9}
8.       OKToProceed := true;
9.       While OKToProceed and |π_i| < u do
10.            let z be the node in Pred(x_i) - π_i that maximizes f(i, π_i ∪ {z});
11.            P_new := f (i, π_i ∪ {z});
12.            if P_new > P_old then
13.                 P_old := P_new;
14.                 π_i := π_i ∪ {z};
15.            else OKToProceed := false;
16.      end {while};
17.      write('Node: ', x_i, ' Parent of x_i : ', π_i);
18. end {for};
19. end {K2};
```

จากตาราง 2.1 เป็นอัลกอริทึมของเคสองที่ใช้ในการสร้างเครือข่ายความเชื่อเบย์ โดยข้อมูลที่ใส่เข้าไปในตอนแรก คือ จำนวนโหนดทั้งหมด (n) , จำนวนโหนดแม่ที่แต่ละโหนดสามารถมีได้ (μ) และข้อมูลในฐานข้อมูล D ที่มีทั้งหมด m กรณี (case) ส่วนผลลัพธ์ที่ได้จะสามารถบอกได้ว่า แต่ละโหนดมีโหนดแม่อะไรบ้างที่ขึ้นต่อกัน

จากตาราง 2.1 ในบรรทัดที่ 7 เป็นการหาค่าความน่าจะเป็นของแต่ละโหนด โดยที่ฟังก์ชันที่ใช้ในการคำนวณ คือ

$$f(\mathrm{i}, \pi_\mathrm{i}) = \prod_{\mathrm{j}=1}^{q_\mathrm{i}} \frac{(r_i - 1)!}{(N_{\mathrm{ij}} + r_\mathrm{i} - 1)!} \prod_{\mathrm{k}=1}^{r_\mathrm{i}} \alpha_{\mathrm{ijk}}! \qquad (2.9)$$

จากสมการ 2.9 มีตัวแปรที่สำคัญ ดังนี้

$f(\mathrm{i}, \pi_\mathrm{i})$ เป็นค่าความน่าจะเป็นของฐานข้อมูล D โดย π_i เป็นโหนดแม่ของ x_i

π_i = เซตของโหนดแม่ของโหนด x_i

q_i = $|\square_i|$

$\square_i$ = โหนดแม่ของ x_i ที่เป็นไปได้ทั้งหมดในฐานข้อมูล D กล่าวคือ ถ้า p_1 ,..., p_s เป็นโหนดแม่ของ x_iแล้ว $\square_i$ เป็นผลคูณของ $\{V_1^{p1},...,V_{rp1}^{p1}\}$ x ... x $\{V_1^{ps},...,V_{rps}^{ps}\}$ของค่าแอททริบิวต์ ที่เป็นไปได้ p_1 จนถึง p_s

r_i = $|V_i|$

V_i = ค่าแอททริบิวต์ที่เป็นไปได้ทั้งหมดของโหนด x_i

α_{ijk} = จำนวนกรณี (case) ที่เกิดขึ้นในฐานข้อมูล D ที่แอททริบิวต์ x_i ถูกแทนที่ด้วยค่าของมันที่ตำแหน่ง k และโหนดแม่ของ x_i ที่อยู่ใน π_i ถูกแทนที่ด้วยตำแหน่ง j ซึ่งอยู่ในตัวแปร $\square_i$

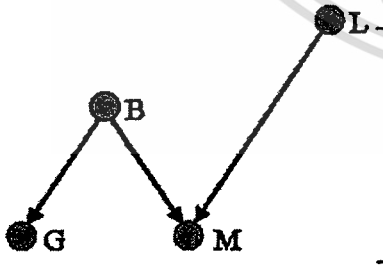
N_{ij} = $\sum_{k=1}^{r_i} \alpha_{ijk}$, นั่นคือ จำนวนของกรณีที่เป็นไปได้ทั้งหมดในฐานข้อมูล ซึ่งโหนดแม่ของ x_i ที่อยู่ใน π_i เป็นค่าที่อยู่ในตำแหน่ง j ซึ่งอยู่ในตัวแปร $\square_i$

จากอัลกอริทึมเคสองนี้ ทำให้เราทราบความสัมพันธ์ของแต่ละโหนดในฐานข้อมูลได้ว่าแต่ละโหนด มีโหนดแม่อะไรบ้าง ทำให้สามารถสร้างเครือข่ายความเชื่อเบย์ได้ว่าลักษณะแบบใดและคำนวณหาตารางความน่าจะเป็นแบบมีเงื่อนไขของแต่ละโหนดต่อไปได้

2) โครงสร้างรู้ (Structure Known)

เป็นกรณีที่รู้โครงสร้างแล้ว ซึ่งผู้เขียนเครือข่ายความเชื่อเบย์เป็นผู้เชี่ยวชาญในปัญหานั้นๆ สามารถบอกโครงสร้างได้อย่างชัดเจน รู้ความสัมพันธ์ระหว่างโหนดในปัญหานั้นๆ แต่อาจไม่รู้ค่าที่ถูกต้องและแม่นยำในตารางความน่าจะเป็นแบบมีเงื่อนไข ดังนั้นกรณีนี้การเรียนรู้เป็นการหาค่าในตารางความน่าจะเป็นแบบมีเงื่อนไขโดยอาศัยตัวอย่างสอน โดยมี 2 วิธี ดังนี้

วิธีที่ 1 การเรียนรู้เครือข่ายความเชื่อเบย์ในกรณีที่โครงสร้างรู้และข้อมูลครบ



G	M	B	L	จำนวนตัวอย่าง
True	True	True	True	54
True	True	True	False	1
True	False	True	True	7
True	False	True	False	27
False	True	True	True	3
False	False	True	False	2
False	False	False	True	4
False	False	False	False	2
				100

รูปที่ 2.4 ตัวอย่างสอนสำหรับการเรียนรู้ซีพีทีในกรณีข้อมูลครบ

เป็นกรณีที่ง่ายที่สุด สามารถทำการเรียนรู้ได้ในลักษณะเดียวกับการเรียนรู้ของตัวจำแนกประเภทเบย์อย่างง่าย (Naïve Bayes) โดยนับจำนวนครั้งที่เกิดขึ้นของข้อมูลเพื่อไปคำนวณหาตารางความน่าจะเป็นแบบมีเงื่อนไขของแต่ละโหนดว่ามีค่าเป็นเท่าไร โดยใช้หลักการดังนี้

$$P(Vi=vi \mid Parents(Vi)=Pi) = \frac{\text{จำนวนตัวอย่างที่มี } Vi=vi}{\text{จำนวนตัวอย่างที่มี } Parents(Vi)=Pi} \quad (2.10)$$

จากสมการ 2.10 สามารถหาค่าความน่าจะเป็นจากตัวอย่างสอนในรูป 2.4 ได้ดังนี้

P(B=true) = (54+1+7+27+3+2)/100 = 0.94

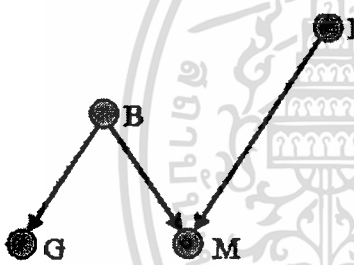
คือนับจำนวนตัวอย่างที่ B เป็นจริงในตารางแล้วหารด้วยจำนวนตัวอย่างทั้งหมด ค่าความน่าจะเป็นอื่นๆก็คำนวณในทำนองเดียวกัน

P(L=true) = (54+7+3+4)/100 = 0.68

P(M|B, ¬L) เท่ากับอัตราส่วนที่ M =true เมื่อ B=true ,L=false เท่ากับ 1/(1+27+2) = 0.03

ดังนั้นเราสามารถคำนวณหาความน่าจะเป็นของโหนด G ได้เช่นเดียวกัน

วิธีที่ 2 การเรียนรู้เครือข่ายความเชื่อเบย์ในกรณีที่โครงสร้างรู้และข้อมูลมีค่าหาย



G	M	B	L	จำนวนตัวอย่าง
True	True	True	True	54
True	True	True	False	1
*	*	True	True	7
True	False	True	False	27
False	True	*	True	3
False	False	True	False	2
False	False	False	True	4
False	False	False	False	2
				100

รูปที่ 2.5 ตัวอย่างสอนสำหรับการเรียนรู้ซีพีทีในกรณีข้อมูลมีค่าหาย

เป็นกรณีที่ข้อมูลบางตัวมีค่าบางค่าหายไป โดยเราสามารถหาค่าน้ำหนักได้ ถ้าเรารู้ค่าความน่าจะเป็นในตารางความน่าจะเป็นแบบมีเงื่อนไขก่อน ซึ่งเราไม่รู้

จากรูป 2.5 จะเห็นได้ว่า '*' ในตาราง หมายถึง ค่าที่หายไป ซึ่งมีอยู่ 2 แถวคือแถวที่ 5 มี 3 ตัวอย่างและแถวที่ 3 มี 7 ตัวอย่าง

พิจารณาแถวที่ 5 ของข้อมูล เป็นกรณีของตัวอย่าง 3 ตัว ที่มีค่า G=False , M=True , L=True ในกรณีนี้เราไม่รู้ค่าของ B แต่อาจคำนวณหา P(B|¬G,M,L) หรือ P(¬B|¬G,M,L) ได้ถ้าหากเรารู้ซีพีที สมมติว่าเรารู้ซีพีที จะทำให้เราสามารถหาความน่าจะเป็นที่ B เป็นจริงหรือเท็จของตัวอย่างทั้งสามตัวได้ จากนั้นเราจะแทนที่ตัวอย่างทั้งสามนี้ด้วยตัวอย่างมีน้ำหนัก (weighted example) 2 ตัวดังนี้

ตัวอย่างที่ 1 ที่ B= True มีน้ำหนักเท่ากับ P(B|¬G,M,L)

ตัวอย่างที่ 2 ที่ B= False มีน้ำหนักเท่ากับ P(¬B|¬G,M,L)

พิจารณาแถวที่ 3 ของข้อมูล เป็นกรณีของตัวอย่าง 7 ตัว ที่มีค่า B=True , L=True ส่วน G และ M ไม่รู้ค่า เราสามารถแทนที่ตัวอย่างทั้งเจ็ดตัวด้วยตัวอย่างมีน้ำหนัก 4 ตัวดังนี้

ตัวอย่างที่ 1 ที่ G= True , M=True มีน้ำหนักเท่ากับ P(G,M|B,L)

ตัวอย่างที่ 2 ที่ G= True , M=False มีน้ำหนักเท่ากับ P(G, ¬M|B,L)

ตัวอย่างที่ 3 ที่ G= False , M=True มีน้ำหนักเท่ากับ P(¬G,M|B,L)

ตัวอย่างที่ 4 ที่ G= False , M=False มีน้ำหนักเท่ากับ P(¬G, ¬M|B,L)

วิธีการทำคือ เราจะสมมติค่าความน่าเป็นในตารางความน่าจะเป็นแบบมีเงื่อนไขโดยสุ่มค่าเริ่มต้นเข้าไปในตารางความน่าจะเป็นแบบมีเงื่อนไขเสมือนว่าเรามีค่าในตารางแล้ว และเราก็สามารถหาน้ำหนักของตัวอย่างไม่ทราบค่าได้ทุกตัว ก็จะทำให้เซตตัวอย่างไม่รู้ค่าเป็นเซตตัวอย่างที่เรารู้ค่าทุกตัว การเรียนรู้จะเหมือนกับกรณีที่ตัวอย่างมีข้อมูลครบ แต่การคำนวณค่าน้ำหนักจะไม่ได้ค่าที่ถูกต้องเพราะว่าเราสุ่มค่าเริ่มต้นในตารางความน่าจะเป็นแบบมีเงื่อนไข ซึ่งไม่ใช่ค่าที่ถูก แต่เนื่องจากว่าเมื่อเราได้น้ำหนักแล้วนำตัวอย่างไปรวมกับตัวอย่างที่เหลือที่เป็นตัวอย่างที่มีข้อมูลครบก็จะทำให้การประมาณค่าในตารางความน่าจะเป็นแบบมีเงื่อนไขครั้งใหม่มีความถูกต้องเพิ่มขึ้นกว่าตารางความน่าจะเป็นแบบมีเงื่อนไขเริ่มต้น เพราะว่าตัวอย่างส่วนใหญ่ของเราเป็นตัวอย่างที่ถูกต้องจะมีแค่ตัวอย่างมีน้ำหนักเท่านั้นที่ไม่ถูกต้องสมบูรณ์ แสดงว่าการปรับค่าตารางความน่าจะเป็นแบบมีเงื่อนไขทำให้ได้ตารางความน่าจะเป็นแบบมีเงื่อนไขใหม่ที่ดีขึ้น และถ้าเราทำซ้ำกระบวนการเดิมด้วยตารางความน่าจะเป็นแบบมีเงื่อนไขที่ดีขึ้นก็จะทำให้การหาค่าน้ำหนักมีความแม่นยำยิ่งขึ้น และส่งผลให้การปรับตารางความน่าจะเป็นแบบมีเงื่อนไขในรอบถัดไปดีขึ้นอีก เมื่อวนซ้ำไปเรื่อยๆก็จะได้ตารางความน่าจะเป็นแบบมีเงื่อนไขที่ดีขึ้นเรื่อยๆ จนกระทั่งตารางความน่าจะเป็นแบบมีเงื่อนไขไม่เปลี่ยนแปลง เราก็หยุดกระบวนการเรียนรู้ได้ อัลกอริทึมการเรียนรู้แบบนี้เรียกว่า **อัลกอริทึมการเรียนรู้แบบอีเอ็ม(EM – Expectation Maximization Algorithm)**

บทที่ 3
การวิเคราะห์และออกแบบระบบ

ในการวิเคราะห์และออกแบบระบบเหมืองข้อมูลสำหรับวิธีการการแบ่งประเภทของข้อมูลแบบเครือข่ายความเชื่อเบย์ สามารถแบ่งการทำงานออกเป็น 3 ส่วน คือ ส่วนการสร้างแบบจำลองใหม่ให้กับระบบ ส่วนการทดสอบแบบจำลอง และส่วนการแบ่งประเภทของข้อมูล

3.1 รายละเอียดของระบบ

ระบบเหมืองข้อมูลสำหรับวิธีการการแบ่งประเภทของข้อมูลแบบเครือข่ายความเชื่อเบย์ มีกระบวนการทำงานหลักๆในแต่ละส่วน ดังนี้

3.1.1 ส่วนการเชื่อมต่อฐานข้อมูล (Connect Database)

การเชื่อมต่อฐานข้อมูล เป็นการติดต่อกับฐานข้อมูล SQL Server 2000 ซึ่งเป็นฐานข้อมูลที่ใช้กับระบบ

3.1.2 ระบบการประมวลผลข้อมูลเบื้องต้น (Preprocess Data System)

หลังจากทำการเชื่อมต่อกับฐานข้อมูลแล้ว ระบบนี้จะให้ผู้ใช้เลือกชุดข้อมูลที่จะนำมาใช้ในการสร้างแบบจำลอง โดยมีการทำงานหลักๆอยู่ 2 ส่วน คือ ส่วนการแก้ไขข้อมูล (Data Cleaning) เพื่อจัดการกับข้อมูลที่เป็นค่าว่าง (Missing Value) หากพบว่าแอททริบิวต์มีข้อมูลที่เป็นค่าว่าง สามารถจัดการได้โดยลบแถวของแอททริบิวต์ที่มีค่าว่างออก หรือใส่ค่าข้อมูลในแอททริบิวต์ที่มีค่าว่างเป็นค่าใหม่ หรือใส่ค่าข้อมูลในแอททริบิวต์ที่มีค่าว่างว่าไม่ทราบค่าก็ได้ และส่วนการแปลงค่าข้อมูล (Data Transformation) เพื่อจัดการกับค่าหรือประเภทของข้อมูลใหม่ โดยอาจทำการสร้างแอททริบิวต์ขึ้นมาใหม่ หรือ แปลงค่าข้อมูลจากตัวเลขเป็นตัวอักษร หรือ แปลงค่าข้อมูลจากตัวอักษรเป็นตัวเลข หรือรวมกลุ่มค่าที่เป็นไปได้ของข้อมูลเป็นกลุ่มข้อมูลเพื่อปรับขอบเขตของข้อมูลให้เป็นช่วงที่เหมาะสม โดยทั้ง 2 ส่วนนี้ทำให้ได้ชุดข้อมูลที่เหมาะสมที่จะทำการสร้างแบบจำลองในขั้นตอนถัดไปของระบบได้

3.1.3 ส่วนการเลือกข้อมูล (Select Data Set)

การเลือกข้อมูล เป็นการเลือกชุดข้อมูลเข้าสู่ระบบ และระบุแอททริบิวต์ที่จำเป็นต่อการสร้างแบบจำลองให้กับระบบ

3.1.4 ส่วนการสร้างแบบจำลอง (Create Model)

เมื่อได้ชุดข้อมูลที่มีแอททริบิวต์ที่เหมาะสมและพร้อมที่จะสร้างแบบจำลองแล้ว ให้นำชุดข้อมูลนั้นมาสร้างแบบจำลอง โดยทำการเรียนรู้เครือข่ายความเชื่อเบย์ตามหลักการของอัลกอริทึมเคสอง เพื่อหาโครงสร้างเครือข่ายความเชื่อเบย์และคำนวณหาค่าความน่าจะเป็นแบบมีเงื่อนไข

เพื่อที่จะได้ตารางความน่าจะเป็นแบบมีเงื่อนไข (CPT) ที่สอดคล้องกับตัวอย่างสอนมากที่สุดในแต่ละโหนด และสามารถเก็บแบบจำลองนี้ไว้ใช้ในการแบ่งประเภทของข้อมูลได้ในอนาคต

3.1.5 ส่วนการทดสอบแบบจำลอง (Test Model)

หลังจากได้แบบจำลองที่เหมาะสมแล้ว ให้นำชุดข้อมูลที่ต้องการทดสอบเข้าไปทดสอบแบบจำลองที่ผ่านการเรียนรู้มาแล้ว และทำการทดสอบแบบจำลองเพื่อเป็นการประเมินแบบจำลองและเพื่อให้ผู้ใช้มั่นใจในความถูกต้องของแบบจำลองอีกด้วยว่า มีความถูกต้องและความผิดพลาดเท่าไหร่

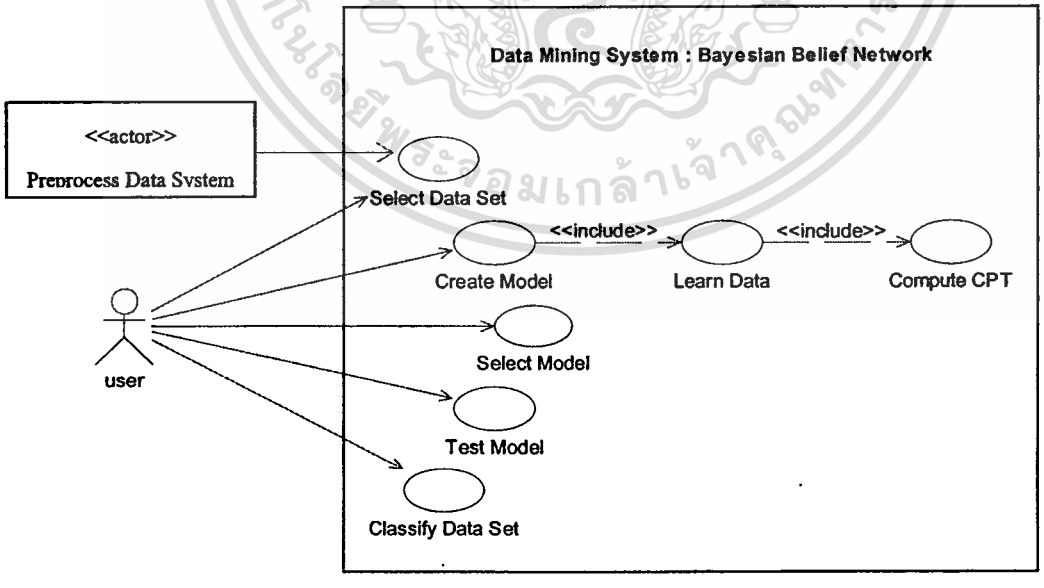
3.1.6 ส่วนการแบ่งประเภทของข้อมูล (Classify Data)

หลังจากทดสอบแบบจำลองจนกระทั่งผู้ใช้มั่นใจในความถูกต้องแล้ว และต้องการแบ่งประเภทของข้อมูลจากชุดข้อมูลใหม่ ให้นำชุดข้อมูลใหม่เข้าสู่ขั้นตอนการแบ่งประเภทของข้อมูล และเลือกแบบจำลองที่ใช้ในการแบ่งประเภทของข้อมูล หลังจากนั้นจะได้ผลลัพธ์ของการแบ่งประเภทของข้อมูลของชุดข้อมูลใหม่ทั้งหมด

3.2 การจำลองการทำงานของระบบด้วยยูเอ็มแอล

ในการจำลองการทำงานของระบบการทำเหมืองข้อมูลในการแบ่งประเภทของข้อมูล โดยใช้เครือข่ายความเชื่อเบย์นั้น จะจำลองการทำงานด้วยยูเอ็มแอล (UML) เป็นหลัก ดังต่อไปนี้

3.2.1 การจำลองการทำงานด้วยยูสเคสไดอะแกรม



รูปที่ 3.1 ยูสเคสไดอะแกรมของระบบทั้งหมด

ตารางที่ 3.1 คำอธิบายยูสเคสไดอะแกรมของ Select Data Set

ยูสเคส	Select Data Set
วัตถุประสงค์	เพื่อเลือกชุดข้อมูลเข้าสู่ระบบและนำชุดข้อมูลดังกล่าวไปประมวลผลข้อมูลเบื้องต้นในขั้นตอนถัดไป
เมื่อทำงานเสร็จ	นำชุดข้อมูลที่เลือกไปประมวลผลข้อมูลเบื้องต้น
แอคเตอร์ที่เกี่ยวข้อง	Preprocess Data System และ User
อินพุต	ชุดข้อมูลที่ต้องการจะทำเหมืองข้อมูล
เอาท์พุต	ชุดข้อมูลที่มีเฉพาะแอททริบิวต์ที่ต้องการ และพร้อมที่จะนำไปใช้ในการประมวลผลข้อมูลเบื้องต้นในขั้นตอนถัดไป
รายละเอียด	1.ผู้ใช้เลือกชุดข้อมูลที่ต้องการเข้าสู่ระบบ 2.เมื่อได้ชุดข้อมูลที่ต้องการแล้ว ทำการเลือกแอททริบิวต์ที่ต้องการทำเหมืองข้อมูลเท่านั้น เพื่อที่จะได้นำชุดข้อมูลดังกล่าวไปประมวลผลข้อมูลเบื้องต้นในขั้นตอนถัดไป

ตารางที่ 3.2 คำอธิบายยูสเคสไดอะแกรมของ Create Model

ยูสเคส	Create Model
วัตถุประสงค์	เพื่อสร้างแบบจำลองในการแบ่งประเภทของข้อมูลแบบเครือข่ายความเชื่อเบย์ที่เหมาะสมได้
เมื่อทำงานเสร็จ	นำแบบจำลองที่ได้ไปทดสอบเพื่อประเมินความถูกต้องอีกครั้ง
แอคเตอร์ที่เกี่ยวข้อง	User
อินพุต	ชุดข้อมูลที่ผ่านการะประมวลผลข้อมูลเบื้องต้นแล้ว และต้องการจะสร้างแบบจำลอง
เอาท์พุต	แบบจำลองที่ใช้ในการแบ่งประเภทของข้อมูลแบบเครือข่ายความเชื่อเบย์ตามต้องการ
รายละเอียด	1.นำเซตข้อมูลที่ต้องการสร้างแบบจำลองเข้าสู่ขั้นตอนการเรียนรู้เพื่อหาความขึ้นต่อกันระหว่างแอททริบิวต์(โหนดแม่ และโหนดลูก) และสามารถนำไปสร้างเครือข่ายความเชื่อเบย์ได้ 2.เมื่อทราบความขึ้นต่อกันระหว่างแอททริบิวต์แล้ว ให้ไปคำนวณหาค่าในตารางความน่าจะเป็นแบบมีเงื่อนไขของแต่ละแอททริบิวต์ และนำค่าความน่าจะเป็นที่คำนวณได้ไปทำการแบ่งประเภทข้อมูลในขั้นตอนถัดไป

ตารางที่ 3.3 คำอธิบายยูสเคสไดอะแกรมของ Learn Data

ยูสเคส	Learn Data
วัตถุประสงค์	เพื่อทำการเรียนรู้ชุดข้อมูลที่ต้องการสร้างแบบจำลอง
เมื่อทำงานเสร็จ	นำความขึ้นต่อกันระหว่างแอททริบิวต์ไปคำนวณหาค่าในตารางความน่าจะเป็นแบบมีเงื่อนไขซึ่งอยู่ในขั้นตอนถัดไป
แอคเตอร์ที่เกี่ยวข้อง	
อินพุต	ชุดข้อมูลที่ผ่านการะประมวลผลข้อมูลเบื้องต้นแล้ว และต้องการสร้างแบบจำลอง
เอาท์พุต	ความขึ้นต่อกันระหว่างแอททริบิวต์เพื่อที่จะนำไปคำนวณหาค่าในตารางความน่าจะเป็นแบบมีเงื่อนไขของแต่ละแอททริบิวต์ได้ในขั้นตอนถัดไป
รายละเอียด	1.นำเซตข้อมูลที่ต้องการสร้างแบบจำลองเข้าสู่ขั้นตอนการเรียนรู้เพื่อหาความขึ้นต่อกันระหว่างแอททริบิวต์(โหนดแม่ และโหนดลูก) ตามหลักอัลกอริทึมเคสอง และสามารถนำไปสร้างเครือข่ายความเชื่อเบย์ได้

ตารางที่ 3.4 คำอธิบายยูสเคสไดอะแกรมของ Compute CPT

ยูสเคส	Compute CPT
วัตถุประสงค์	เพื่อทำการคำนวณหาค่าในตารางความน่าจะเป็นแบบมีเงื่อนไข
เมื่อทำงานเสร็จ	นำค่าในตารางความน่าจะเป็นแบบมีเงื่อนไขของแต่ละแอททริบิวต์ไปทำการแบ่งประเภทข้อมูลในขั้นตอนถัดไปได้
แอคเตอร์ที่เกี่ยวข้อง	
อินพุต	ความขึ้นต่อกันระหว่างแอททริบิวต์ที่ผ่านขั้นตอนการเรียนรู้
เอาท์พุต	ตารางความน่าจะเป็นแบบมีเงื่อนไขของแต่ละแอททริบิวต์
รายละเอียด	1.เมื่อทราบความขึ้นต่อกันระหว่างแอททริบิวต์แล้ว ให้คำนวณหาค่าในตารางความน่าจะเป็นแบบมีเงื่อนไขของแต่ละแอททริบิวต์ และนำค่าความน่าจะเป็นที่คำนวณได้ไปทำการแบ่งประเภทข้อมูลในขั้นตอนถัดไป

ตารางที่ 3.5 คำอธิบายยูสเคสไดอะแกรมของ Select Model

ยูสเคส	Select Model
วัตถุประสงค์	เพื่อเลือกแบบจำลองที่ผ่านการเรียนรู้ไปทดสอบความถูกต้องของแบบจำลอง
เมื่อทำงานเสร็จ	นำแบบจำลองที่เลือกไปทดสอบความถูกต้องของแบบจำลอง
แอคเตอร์ที่เกี่ยวข้อง	User
อินพุต	แบบจำลองที่ผ่านการเรียนรู้แล้ว และต้องการจะทดสอบความถูกต้องของแบบจำลอง
เอาท์พุต	แบบจำลองที่ต้องการทดสอบความถูกต้องของแบบจำลอง
รายละเอียด	1.ผู้ใช้ทำการเลือกแบบจำลองที่ผ่านการเรียนรู้แล้ว ไปทดสอบความถูกต้องของแบบจำลองในขั้นต้อนถัดไป

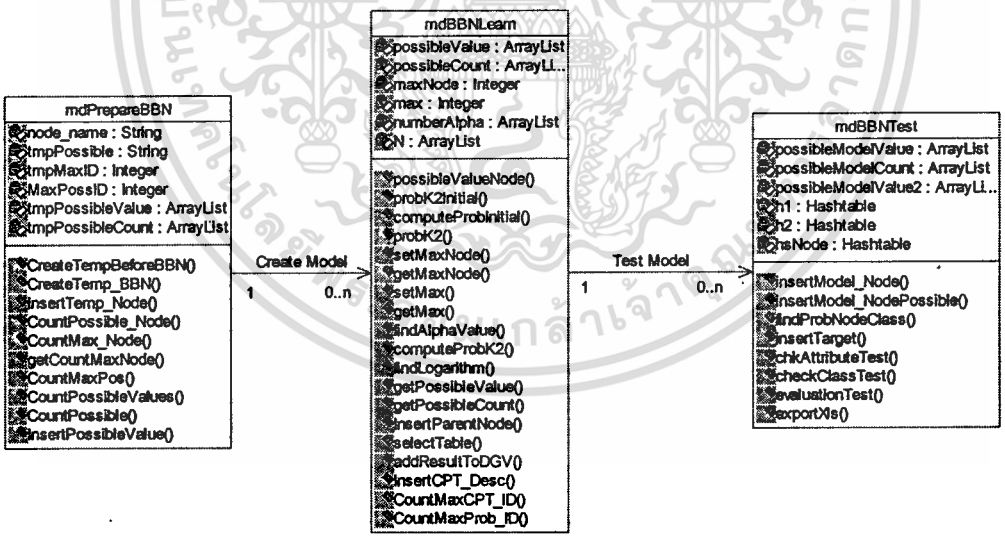
ตารางที่ 3.6 คำอธิบายยูสเคสไดอะแกรมของ Test Model

ยูสเคส	Test Model
วัตถุประสงค์	เพื่อทดสอบความถูกต้องของแบบจำลอง ซึ่งทำให้ผู้ใช้มีความมั่นใจมากขึ้นว่าแบบจำลองนั้นมีความถูกต้องและน่าเชื่อถือ
เมื่อทำงานเสร็จ	สามารถนำแบบจำลองที่น่าเชื่อถือนี้ไปแบ่งประเภทของข้อมูลได้
แอคเตอร์ที่เกี่ยวข้อง	User
อินพุต อินพุต (ต่อ)	แบบจำลองที่ผ่านการเรียนรู้ และต้องการทดสอบความถูกต้องของแบบจำลอง
เอาท์พุต	ค่าประเมินความถูกต้องของแบบจำลอง
รายละเอียด	1.ผู้ใช้เลือกแบบจำลองที่ต้องการทดสอบความถูกต้อง 2.ผู้ใช้เลือกชุดข้อมูลชุดใหม่ที่ทราบประเภทของข้อมูลแล้วเข้าสู่ระบบและทำการทดสอบความถูกต้องของแบบจำลอง เพื่อดูว่าการทำงานของแบบจำลองยังได้ผลลัพธ์คงเดิมหรือไม่ 3.ตรวจสอบความถูกต้องของแบบจำลอง โดยประเมินความถูกต้อง และความคลาดเคลื่อนของแบบจำลอง ระหว่างประเภทของข้อมูลที่มีอยู่และกับประเภทของข้อมูลของระบบว่าเหมือนกันหรือต่างกันอย่างไร

ตารางที่ 3.7 คำอธิบายยูสเคสไดอะแกรมของ Classify Data Set

ยูสเคส	Classify Data Set
วัตถุประสงค์	เพื่อแบ่งประเภทของข้อมูลจากชุดข้อมูลที่ผู้ใช้ต้องการ
เมื่อทำงานเสร็จ	ประเภทของข้อมูลของแต่ละข้อมูลในชุดข้อมูล
แอคเตอร์ที่เกี่ยวข้อง	User
อินพุต	แบบจำลองและชุดข้อมูลที่ต้องการแบ่งประเภทของข้อมูล
เอาท์พุต	ประเภทของข้อมูลของแต่ละข้อมูลในชุดข้อมูล
รายละเอียด	1.ผู้ใช้เลือกแบบจำลองที่ใช้ในการแบ่งประเภทของข้อมูล และเลือกชุดข้อมูลที่ต้องการแบ่งประเภทของข้อมูลเข้าสู่ระบบ แล้วระบบจะทำการคำนวณหาค่าความน่าจะเป็นจากเครือข่ายความเชื่อเบย์ตามแบบจำลองที่เลือกไว้ แล้วจะได้ประเภทของข้อมูลในแต่ละข้อมูล

3.2.2 การออกแบบระบบด้วยคลาสไดอะแกรม



รูปที่ 3.2 คลาสไดอะแกรมของระบบ

จากรูปที่ 3.2 เป็นการออกแบบระบบด้วยคลาสไดอะแกรมซึ่งอธิบายรายละเอียดการทำงานของออบเจ็กต์ต่างๆที่มีอยู่ในระบบเหมืองข้อมูลสำหรับการแบ่งประเภทของข้อมูลโดยวิธีเครือข่ายความเชื่อเบย์ และแต่ละคลาสมีความหมายดังนี้

- **คลาส mdPrepareBBN**

เป็นคลาสที่นำข้อมูลที่ผ่านระบบการประมวลผลมาใช้ในการเตรียมข้อมูลก่อนที่จะทำการสร้างแบบจำลอง โดยทำการสร้างตารางใหม่ที่มีแอททริบิวต์ตามที่ผู้ใช้เลือก และเก็บค่าที่เป็นไปได้ทั้งหมดของแต่ละแอททริบิวต์ เพื่อไว้ใช้ในการสร้างแบบจำลองในขั้นตอนถัดไป

- **คลาส mdBBNLearn**

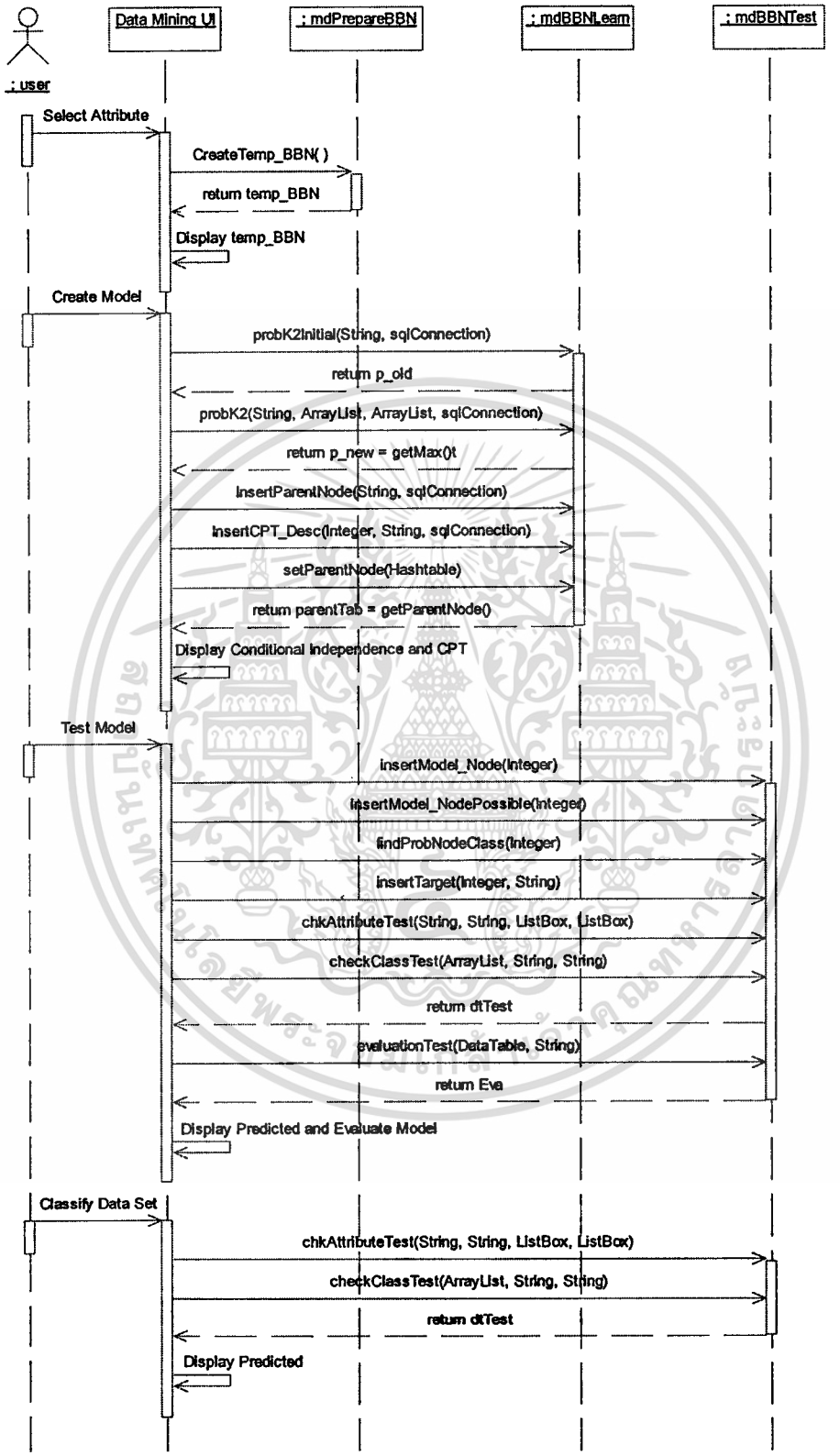
เป็นคลาสที่นำข้อมูลที่เตรียมไว้มาสร้างแบบจำลอง ซึ่งการสร้างแบบจำลองนี้จะต้องผ่านการเรียนรู้เครือข่ายความเชื่อเบย์โดยมีอยู่ 2 ขั้นตอน คือ ขั้นตอนแรก ทำการเรียนรู้เพื่อหาความขึ้นต่อกันและไม่ขึ้นต่อกันระหว่างแอททริบิวต์ตามหลักอัลกอริทึมเคสอง เพื่อให้ทราบว่าแอททริบิวต์ไหนเป็นโหนดแม่ โหนดลูก แล้วจะได้เครือข่ายความเชื่อเบย์ออกมา ขั้นตอนที่สองคำนวณหาค่าความน่าจะเป็นแบบมีเงื่อนไขตามหลักทฤษฎีของเครือข่ายความเชื่อเบย์ เพื่อไว้ใช้ในทดสอบแบบจำลอง และการแบ่งประเภทของข้อมูลในขั้นตอนถัดไปได้

- **คลาส mdBBNTest**

เป็นคลาสที่นำแบบจำลองที่ผ่านการเรียนรู้มาทดสอบแบบจำลองว่าน่าเชื่อถือได้หรือไม่ โดยให้ผู้ใช้เลือกแบบจำลองที่ได้สร้างไว้แล้ว และชุดข้อมูลใหม่ที่ทราบผลลัพธ์ในการแบ่งประเภทของข้อมูล และทำการแบ่งประเภทของข้อมูล แล้วนำมาเปรียบเทียบกันระหว่างประเภทข้อมูลเก่าและประเภทข้อมูลใหม่ที่ได้จากระบบ ว่ามีความเหมือนกันหรือต่างกันอย่างไร เพื่อบอกให้ผู้ใช้ทราบว่าแบบจำลองนี้น่าเชื่อถือหรือไม่ และในส่วนของการแบ่งประเภทของข้อมูล สามารถใช้ร่วมกันได้กับการทดสอบแบบจำลองของระบบ โดยผู้ใช้จะต้องเลือกแบบจำลอง และชุดข้อมูลใหม่ที่ต้องการแบ่งประเภทของข้อมูลเข้าสู่ระบบ และระบบจะทำการคำนวณหาค่าความน่าจะเป็นที่มีค่ามากที่สุดในแต่ละเงื่อนไขออกมา เพื่อให้ได้ประเภทของข้อมูลของชุดข้อมูลนั้นออกมา

3.2.3 การจำลองการทำงานของระบบด้วยซีเควนซ์ไดอะแกรม

ซีเควนซ์ไดอะแกรมจะแสดงลำดับขั้นตอนการทำงานของระบบการทำเหมืองข้อมูลตามลำดับเหตุการณ์เกิดก่อนเกิดหลัง เพื่อให้สามารถเข้าใจขั้นตอนการทำงานของระบบได้ชัดเจนยิ่งขึ้น ซึ่งจากการออกแบบยูสเคสไดอะแกรมจะสามารถแสดงถึงกิจกรรมต่างที่เกิดขึ้นได้โดยซีเควนซ์ไดอะแกรมดังนี้



รูปที่ 3.3 ซีเควนซ์ไดอะแกรมของระบบเหมืองข้อมูลสำหรับวิธีการการแบ่งประเภทของข้อมูล

Actor: User

Description: ผู้ใช้งานจะทำการสร้างแบบจำลอง

Basic Flow:

1. ผู้ใช้งานเลือกแอททริบิวต์จากชุดข้อมูลที่จะใช้ในการสร้างแบบจำลอง

2. ระบบจะทำสร้างชุดข้อมูลใหม่จากชุดข้อมูลที่ผ่านระบบการประมวลผลข้อมูลเบื้องต้นมาแล้ว เพื่อใช้ในการสร้างแบบจำลองในขั้นตอนถัดไป

3. เมื่อผู้ใช้งานต้องการสร้างแบบจำลอง ระบบจะทำการเรียนรู้ตามอัลกอริทึมเคสอง เพื่อหาความขึ้นต่อกันระหว่างแอททริบิวต์ เพื่อที่จะสามารถสร้างเครือข่ายความชื่อเบย์ได้ และคำนวณหาค่าความน่าจะเป็นแบบมีเงื่อนไขเพื่อนำไปใช้ในการแบ่งประเทของข้อมูลในขั้นตอนถัดไปได้

4. เมื่อระบบทำการเรียนรู้เสร็จสมบูรณ์แล้ว ระบบจะแสดงความสัมพันธ์ของแต่ละแอททริบิวต์ในเครือข่ายความเชื่อเบย์ และตารางความน่าจะเป็นแบบมีเงื่อนไขให้ผู้ใช้งานได้ทราบ

5. เมื่อผู้ใช้งานต้องการทดสอบแบบจำลองที่ผ่านการเรียนรู้มาแล้ว สามารถทำได้โดยการเลือกแบบจำลอง และชุดข้อมูลสำหรับทดสอบ

6. ระบบจะทำการทดสอบแบบจำลองโดยจะใช้ข้อมูลสำหรับทดสอบ มาแบ่งประเภทของข้อมูล เพื่อมาเปรียบเทียบกับค่าเดิม ว่าเหมือนกันหรือต่างกันอย่างไร เพื่อเป็นการประเมินความน่าเชื่อถือของแบบจำลอง

7. เมื่อทำการทดสอบแบบจำลองเสร็จระบบจะแสดงผลลัพธ์ของการทดสอบระบบออกมาว่า มีความถูกต้อง และ ผิดพลาดอยู่เท่าไร

8. เมื่อผู้ใช้ได้แบบจำลองที่เหมาะสมแล้ว สามารถนำแบบจำลองไปหาประเภทของข้อมูลตามเงื่อนไขของผู้ใช้งานได้ในแต่ละชุดข้อมูล และจะได้ประเภทของข้อมูลทั้งหมดออกมา

3.3 การออกแบบฐานข้อมูลที่ใช้ในระบบ

ระบบเหมืองข้อมูลมีตารางที่ใช้ในการทำเหมืองข้อมูลแบบเครือข่ายความเชื่อเบย์ทั้งหมด 7 ตาราง ดังนี้

3.3.1 ตาราง NODE

เป็นตารางชั่วคราวที่เก็บข้อมูลของแต่ละแอททริบิวต์ในชุดข้อมูลที่ทำการสร้างแบบจำลอง และหากทำการสร้างแบบจำลองเสร็จแล้ว จะทำการลบข้อมูลออกทั้งหมดเพื่อใช้ในการสร้างแบบจำลองครั้งถัดไป

ตารางที่ 3.8 รายละเอียดของตาราง NODE

Column Name	Column Description	Data Type	Length	Null	Relationship
node_id	รหัสของโหนด	int	4	No	PK
model_id	รหัสของแบบจำลอง	int	4	Yes	
node_name	ชื่อของโหนด	varchar	50	No	
possible_count	จำนวนค่าที่เป็นไปได้	int	4	Yes	
parent_count	จำนวนของโหนดแม่	int	4	Yes	
parent_name	ชื่อของโหนดแม่	varchar	50	Yes	

3.3.2 ตาราง NODE_POSSIBLE

เป็นตารางชั่วคราวที่เก็บค่าของแต่ละแอททริบิวต์ในชุดข้อมูลที่ทำการสร้างแบบจำลอง และหากทำการสร้างแบบจำลองเสร็จแล้ว จะทำการลบข้อมูลออกทั้งหมดเพื่อใช้ในการสร้างแบบจำลองครั้งถัดไป

ตารางที่ 3.9 รายละเอียดของตาราง NODE_POSSIBLE

Column Name	Column Description	Data Type	Length	Null	Relationship
possible_id	รหัสค่าความเป็นไปได้	int	4	No	PK
model_id	รหัสของแบบจำลอง	int	4	Yes	FK
node_id	รหัสของโหนด	int	4	Yes	FK
possible_value	ค่าที่เป็นไปได้ของโหนด	varchar	50	Yes	
possible_count	จำนวนแถวที่มีค่าที่เป็นได้ของโหนด	int	4	Yes	

3.3.3 ตาราง TABLE_DESC

เป็นตารางที่เก็บชื่อของชุดข้อมูลทั้งหมดที่มีอยู่ในระบบ

ตารางที่ 3.10 รายละเอียดของตาราง TABLE_DESC

Column Name	Column Description	Data Type	Length	Null	Relationship
table_id	รหัสของตาราง	int	4	No	PK
table_name	ชื่อของตาราง	varchar	50	Yes	

3.3.4 ตาราง CPT

เป็นตารางที่เก็บความขึ้นต่อกันระหว่างแอททริบิวต์และค่าความน่าจะเป็นแบบมีเงื่อนไขของแต่ละแอททริบิวต์ในแต่ละชุดข้อมูล

ตารางที่ 3.11 รายละเอียดของตาราง CPT

Column Name	Column Description	Data Type	Length	Null	Relationship
prob_id	รหัสของตารางความน่าจะเป็น	int	4	No	PK
prob_name	ชื่อของตารางความน่าจะเป็น	varchar	50	No	
cpt_id	รหัสของตารางซีพีที	int	4	No	
child_node	โหนดลูก	varchar	50	No	
child_value	ค่าที่เป็นไปได้ของโหนดลูก	varchar	50	No	
parent_node	โหนดแม่	varchar	50	Yes	
parent_value	ค่าที่เป็นไปได้ของโหนดแม่	varchar	50	Yes	
prob_value	ค่าความน่าจะเป็น	float	8	No	

3.3.5 ตาราง MODEL_DESC

เป็นตารางที่เก็บรายละเอียดของแบบจำลองที่ผ่านการเรียนรู้จากระบบ

ตารางที่ 3.12 รายละเอียดของตาราง MODEL_DESC

Column Name	Column Description	Data Type	Length	Null	Relationship
model_id	รหัสของแบบจำลอง	int	4	No	PK
model_name	ชื่อของแบบจำลอง	varchar	50	No	
table_name	ชื่อของชุดข้อมูล	varchar	50	Yes	
class_name	ประเภทของชุดข้อมูล	varchar	50	Yes	
cpt_id	รหัสของตารางซีพีที	int	4	No	FK
sum_node	จำนวนโหนดทั้งหมด	int	4	Yes	
create_dt	วันที่สร้างแบบจำลอง	datetime	-	Yes	

ตารางที่ 3.12 (ต่อ)

Column Name	Column Description	Data Type	Length	Null	Relationship
evaluate_good	ค่าประเมินความถูกต้องจากการทดสอบ	varchar	50	Yes	
evaluate_bad	ค่าประเมินความผิดพลาดจากการทดสอบ	varchar	50	Yes	

3.3.6 ตาราง MODEL_NODE

เป็นตารางที่เก็บข้อมูลของแต่ละแอททริบิวต์ในแต่ละแบบจำลอง เพื่อไว้ใช้สำหรับทดสอบแบบจำลองและแบ่งประเภทของข้อมูล

ตารางที่ 3.13 รายละเอียดของตาราง MODEL_NODE

Column Name	Column Description	Data Type	Length	Null	Relationship
mn_id	รหัสของโหนดในแบบจำลอง	int	4	No	PK
model_id	รหัสของแบบจำลอง	int	4	Yes	FK
node_id	รหัสของโหนด	int	4	No	
node_name	ชื่อของโหนด	varchar	50	No	
possible_count	จำนวนค่าที่เป็นไปได้	int	4	Yes	
parent_count	จำนวนของโหนดแม่	int	4	Yes	
parent_name	ชื่อของโหนดแม่	varchar	50	Yes	

3.3.7 ตาราง MODEL_NODE_POSSIBLE

เป็นตารางที่เก็บค่าของแต่ละแอททริบิวต์ในแต่ละแบบจำลอง เพื่อไว้ใช้สำหรับทดสอบแบบจำลองและแบ่งประเภทของข้อมูล

ตารางที่ 3.14 รายละเอียดของตาราง MODEL_NODE_POSSIBLE

Column Name	Column Description	Data Type	Length	Null	Relationship
mnp_id	รหัสของค่าที่เป็นไปได้ในแบบจำลอง	int	4	No	PK

ตารางที่ 3.14 (ต่อ)

Column Name	Column Description	Data Type	Length	Null	Relationship
model_id	รหัสของแบบจำลอง	int	4	No	FK
node_id	รหัสของโหนด	int	4	No	FK
possible_value	ค่าที่เป็นไปได้ของโหนด	varchar	50	No	
possible_count	จำนวนแถวที่มีค่าที่เป็นไปได้ของโหนด	int	4	No	
prob_node	ค่าความน่าจะเป็นของโหนด	float	8	Yes	

บทที่ 4

การพัฒนาระบบ

ในการพัฒนาระบบเหมืองข้อมูลสำหรับวิธีการการแบ่งประเภทของข้อมูลแบบเครือข่ายความเชื่อเบย์ จะกล่าวถึงเครื่องมือที่ใช้ในการพัฒนาระบบ การทำงานหลักของระบบทั้งหมด รวมทั้งทฤษฎีที่ใช้ในการพัฒนาระบบของเครือข่ายความเชื่อเบย์ แหล่งข้อมูลที่ใช้ในการพัฒนาระบบ และตัวอย่างการใช้งานระบบ ซึ่งจะกล่าวดังต่อไปนี้

4.1 เครื่องมือที่ใช้ในการพัฒนาระบบ

4.1.1 ฮาร์ดแวร์

ในการพัฒนาระบบใช้เครื่องคอมพิวเตอร์ที่มีคุณสมบัติดังนี้

- CPU : Intel® Core ™ 2 Duo CPU T7300 @ 2.00 GHz 2.00 GHz
- RAM : 1.00 GB
- System Type : 32 – bit Operating System
- Hard Disk : 160 GB

4.1.2 ซอฟท์แวร์

ในการพัฒนาระบบใช้ซอฟท์แวร์ดังนี้

- Windows Vista ™ Home Premium
- Microsoft Visual Studio 2005
- Microsoft SQL Server 2000

4.2 การทำงานหลักของระบบ

ในการพัฒนาระบบเหมืองข้อมูลสำหรับวิธีการการแบ่งประเภทของข้อมูลแบบเครือข่ายความเชื่อเบย์ สามารถแบ่งการทำงานหลักของระบบออกเป็น 3 ส่วน ดังนี้ คือ ส่วนการสร้างแบบจำลองใหม่ให้กับระบบ ส่วนการทดสอบแบบจำลอง และส่วนการแบ่งประเภทของข้อมูล

4.2.1 ส่วนการสร้างแบบจำลอง

ในการสร้างแบบจำลองมีขั้นตอนในการทำงานดังนี้

1) รับชุดข้อมูลจากระบบประมวลผลข้อมูลเบื้องต้นที่ได้รับการแก้ไขข้อมูล (Data Cleaning) และการแปลงค่าของข้อมูล(Data Transformation)

2) เลือกแอททริบิวต์ที่จำเป็นต่อการสร้างแบบจำลองเครือข่ายความเชื่อเบย์ทั้งหมด

3) เลือกแอททริบิวต์ที่บ่งบอกว่าเป็นประเภทของข้อมูล เพื่อให้ระบบได้นำไปใช้ในการเรียนรู้

4) หลังจากเลือกแอททริบิวต์ทั้งหมดแล้ว ระบบจะทำการเรียนรู้เครือข่ายความเชื่อเบย์ตามหลักอัลกอริทึมเคสอง แล้วระบบจะแสดงความขึ้นต่อกันระหว่างแอททริบิวต์ จะทำให้ทราบว่าแอททริบิวต์ไหนเป็นโหนดแม่ และโหนดลูก

5) เมื่อทราบความขึ้นต่อกันระหว่างแอททริบิวต์แล้ว ระบบจะทำการคำนวณหาค่าความน่าจะเป็นในตารางความน่าจะเป็นแบบมีเงื่อนไขตามอัลกอริทึมของเครือข่ายความเชื่อเบย์ และแสดงตารางความน่าจะเป็นแบบมีเงื่อนไขให้ผู้ใช้ทราบ

6) ผู้ใช้สามารถเก็บแบบจำลองที่สร้างขึ้นนี้ไว้เพื่อใช้ในการทดสอบแบบจำลอง หรือ ใช้ในการแบ่งประเภทของข้อมูล

4.2.2 ส่วนการทดสอบแบบจำลอง

ในการทดสอบแบบจำลองมีขั้นตอนในการทำงานดังนี้

1) เลือกแบบจำลองที่ต้องการทดสอบแบบจำลองที่เก็บไว้ในระบบ

2) เลือกชุดข้อมูลที่ทราบประเภทของข้อมูลมาใช้ในการทดสอบแบบจำลอง เพื่อที่ระบบจะได้ประเมินความถูกต้องและความผิดพลาดของแบบจำลองได้

3) ระบุแอททริบิวต์ที่เป็นประเภทของข้อมูลสำหรับชุดข้อมูลที่นำมาทดสอบ

4) หลังจากเตรียมแบบจำลอง และชุดข้อมูลแล้ว ระบบจะทำการแบ่งประเภทของข้อมูลตามอัลกอริทึมเครือข่ายความเชื่อเบย์ โดยจะเพิ่มแอททริบิวต์อีกหนึ่งแอททริบิวต์ เพื่อบอกประเภทของข้อมูลที่ได้จากระบบ

5) หลังจากแบ่งประเภทของข้อมูลแล้ว ระบบจะทำการประเมินความถูกต้องของแบบจำลอง โดยเปรียบเทียบประเภทของข้อมูลเดิม และประเภทของข้อมูลใหม่ที่ได้จากระบบ และระบบจะทำการแสดงค่าประเมินความถูกต้องที่ได้แสดงให้ผู้ใช้ทราบ เพื่อให้ผู้ใช้เกิดความมั่นใจในแบบจำลองมากขึ้น

4.2.3 ส่วนการแบ่งประเภทของข้อมูล

1) เมื่อผู้ใช้มั่นใจในแบบจำลองจากการทดสอบแบบจำลองแล้ว และต้องการทำเหมืองข้อมูลเพื่อไว้ใช้ในการแบ่งประเภทของข้อมูลแบบเครือข่ายความเชื่อเบย์

2) เลือกแบบจำลองที่ใช้ในการแบ่งประเภทของข้อมูล และชุดข้อมูลใหม่ที่ต้องการแบ่งประเภทของข้อมูล โดยระบบจะทำการเพิ่มแอททริบิวต์อีกหนึ่งแอททริบิวต์ เพื่อบอกประเภทของข้อมูลที่ได้จากระบบ

3) แสดงผลการแบ่งประเภทของข้อมูลที่ได้จากระบบ

4.3 ทฤษฎีที่ใช้ในการพัฒนาระบบ

4.3.1 อัลกอริทึมเคสอง

อัลกอริทึมเคสองใช้ในการหาความขึ้นต่อกันระหว่างแอททริบิวต์ เพื่อที่ระบบจะได้สามารถสร้างเครือข่ายความเชื่อเบย์ได้ .

ตารางที่ 4.1 อัลกอริทึมเคสองที่ใช้ในการสร้างเครือข่ายความเชื่อเบย์

```
1. procedure K2;
2. {Input: A set of n nodes, an ordering on the nodes, an upper bound u on the
3.       number of parents a node may have, and a database D containing m cases.}
4. {Output: For each node, a printout of the parents of the node.}
5. for i:= 1 to n do
6.       π_i := Ø;
7.       P_old := f(i, π_i);  {This function is computed using Equation 2.9}
8.       OKToProceed := true;
9.       While OKToProceed and |π_i| < u do
10.            let z be the node in Pred(x_i) - π_i that maximizes f(i, π_i ∪ {z});
11.            P_new := f (i, π_i ∪ {z});
12.            if P_new > P_old then
13.                 P_old := P_new;
14.                 π_i := π_i ∪ {z};
15.            else OKToProceed := false;
16.      end {while};
17.      write('Node: ', x_i, ' Parent of x_i : ', π_i);
18. end {for};
19. end {K2};
```

4.3.3 ทฤษฎีของเครือข่ายความเชื่อเบย์

สมการการคำนวณหาค่าความน่าจะเป็นในตารางความน่าจะเป็นแบบมีเงื่อนไข สำหรับเครือข่ายความเชื่อเบย์

$$P(Vi=vi \mid Parents(Vi)=Pi) = \frac{\text{จำนวนตัวอย่างที่มี } Vi=vi}{\text{จำนวนตัวอย่างที่มี } Parents(Vi)=Pi} \quad (4.1)$$

สมการที่ใช้ในการแบ่งประเภทของข้อมูลตามหลักทฤษฎีเครือข่ายความเชื่อเบย์ โดยความน่าจะเป็นของประเภทไหนมีค่ามากสุด แสดงว่าเงื่อนไขนั้นมีประเภทของข้อมูลนั้น

$$h_{MAP} = \arg\max_{h \in H} P(D|h)\,P(h) \quad (4.2)$$

4.4 แหล่งข้อมูลที่ใช้ในการพัฒนาระบบ

แหล่งข้อมูลที่นำมาใช้ในการทดสอบระบบเป็นชุดข้อมูลที่ได้มาจาก UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/) ซึ่งมีหลายประเภท หลายรูปแบบ และเป็นชุดข้อมูลที่เหมาะกับการแบ่งประเภทของข้อมูล โดยชุดข้อมูลแต่ละชุดจะทราบประเภทของข้อมูลมาก่อนแล้ว ทำให้ง่ายต่อการทดสอบระบบ ว่ามีความถูกต้องเพียงไร

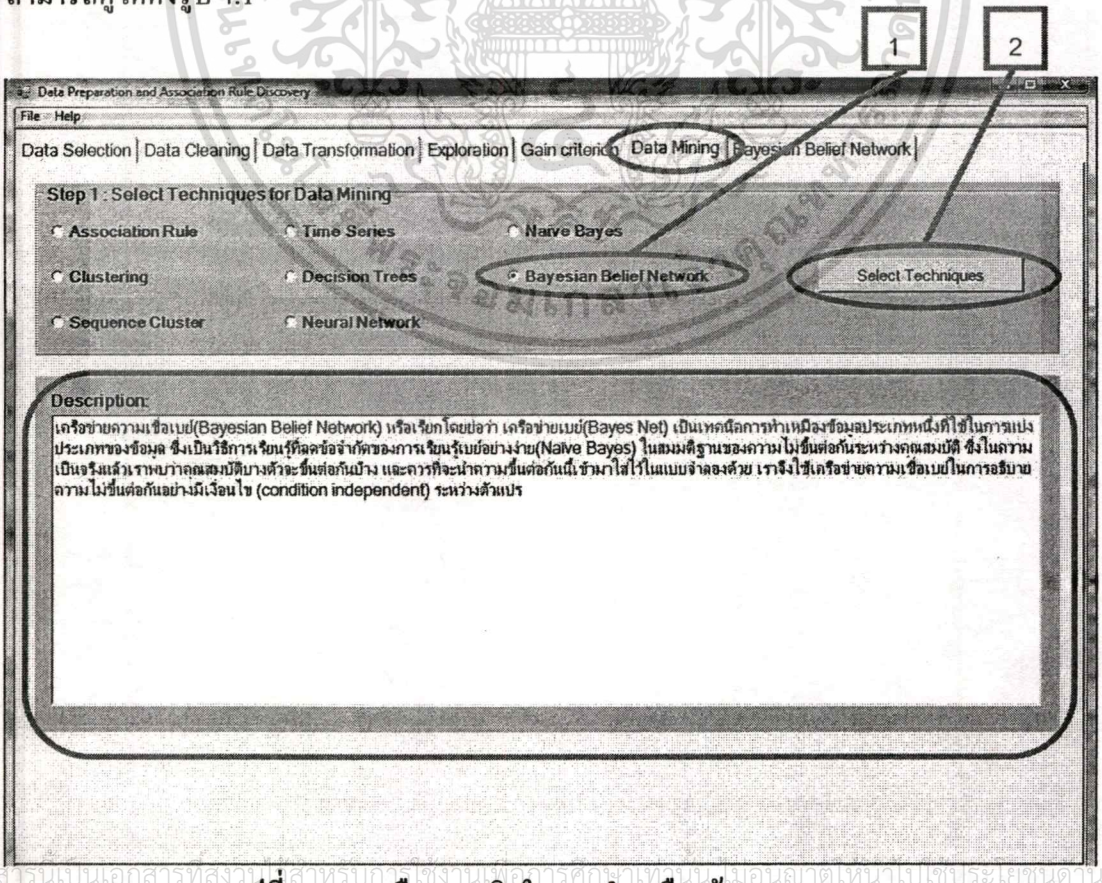
4.5 ตัวอย่างการใช้งานระบบ

สำหรับขั้นตอนการใช้งานระบบสามารถสรุปได้ดังนี้

- เมื่อระบบผ่านขั้นตอนทั้งหมดของระบบการประมวลผลข้อมูลเบื้องต้นแล้ว ซึ่งหมายถึงขั้นตอนการคัดเลือกข้อมูล (Data Selection), การแก้ไขข้อมูล (Data Cleaninig)และการแปลงค่าข้อมูล (Data Tansformation) แล้วจะได้ชุดข้อมูลที่พร้อมที่จะทำการสร้างแบบจำลองได้
- ระบบจะเข้าสู่ขั้นตอนการทำเหมืองข้อมูล (Data Mining) โดยมี 2 ขั้นตอน คือ

1) เลือกเทคนิคการทำเหมืองข้อมูล ซึ่งระบบนี้จะเลือกเทคนิคเครือข่ายความเชื่อเบย์ หรือ Bayesian Belief Network

2) เมื่อเลือกเทคนิคนี้แล้ว ระบบจะแสดงความหมายของแต่ละเทคนิคว่าคืออะไร ใช้ทำอะไร สามารถดูได้ดังรูป 4.1



รูปที่ 4.1 การเลือกเทคนิคในการทำเหมืองข้อมูลของระบบ

- เมื่อเลือกเทคนิคในการทำเหมืองข้อมูลแล้ว จะเข้าสู่ขั้นตอนการสร้างแบบจำลอง ซึ่งมีอยู่ 5 ขั้นตอน คือ

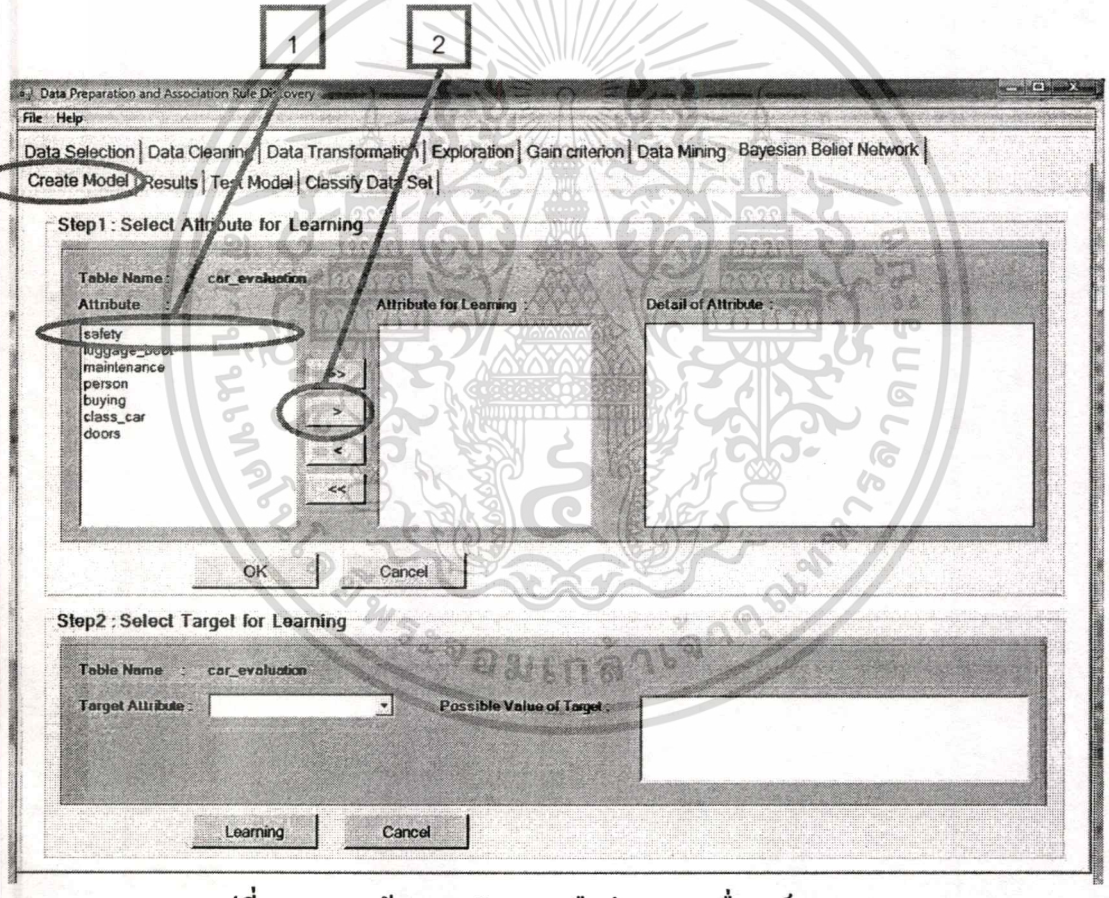
1) เลือกแอททริบิวต์ที่จำเป็นต่อการสร้างแบบจำลอง

2) นำแอททริบิวต์ที่เลือกไปใช้ต่อในการเรียนรู้เครือข่ายความเชื่อเบย์

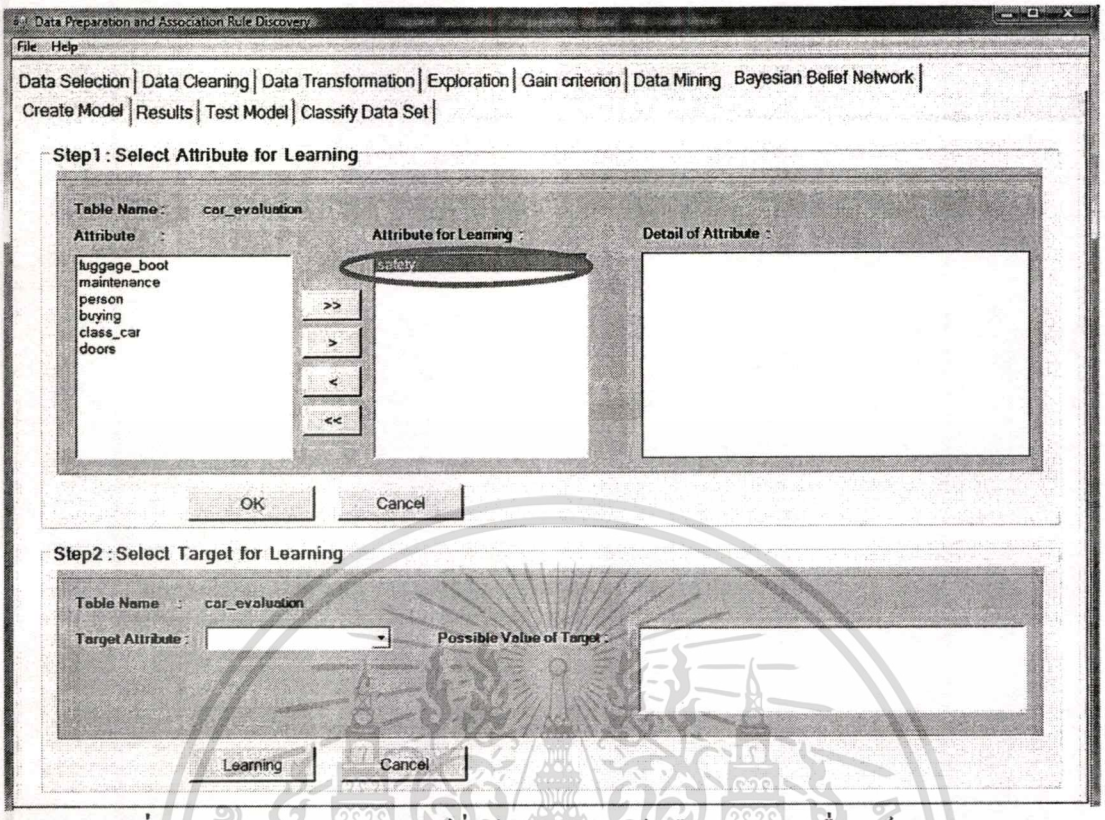
3) เลือก OK เพื่อตกลงสำหรับการเลือกแอททริบิวต์เพื่อใช้ในการเรียนรู้เครือข่ายความเชื่อเบย์

4) เลือกแอททริบิวต์ที่เป็นประเภทของข้อมูลเพื่อที่จะนำไปใช้ในการเรียนรู้เครือข่ายความเชื่อเบย์ เมื่อเลือกแอททริบิวต์แล้วระบบจะแสดงค่าที่เป็นไปได้ของแอททริบิวต์นั้นให้ผู้ใช้ทราบ

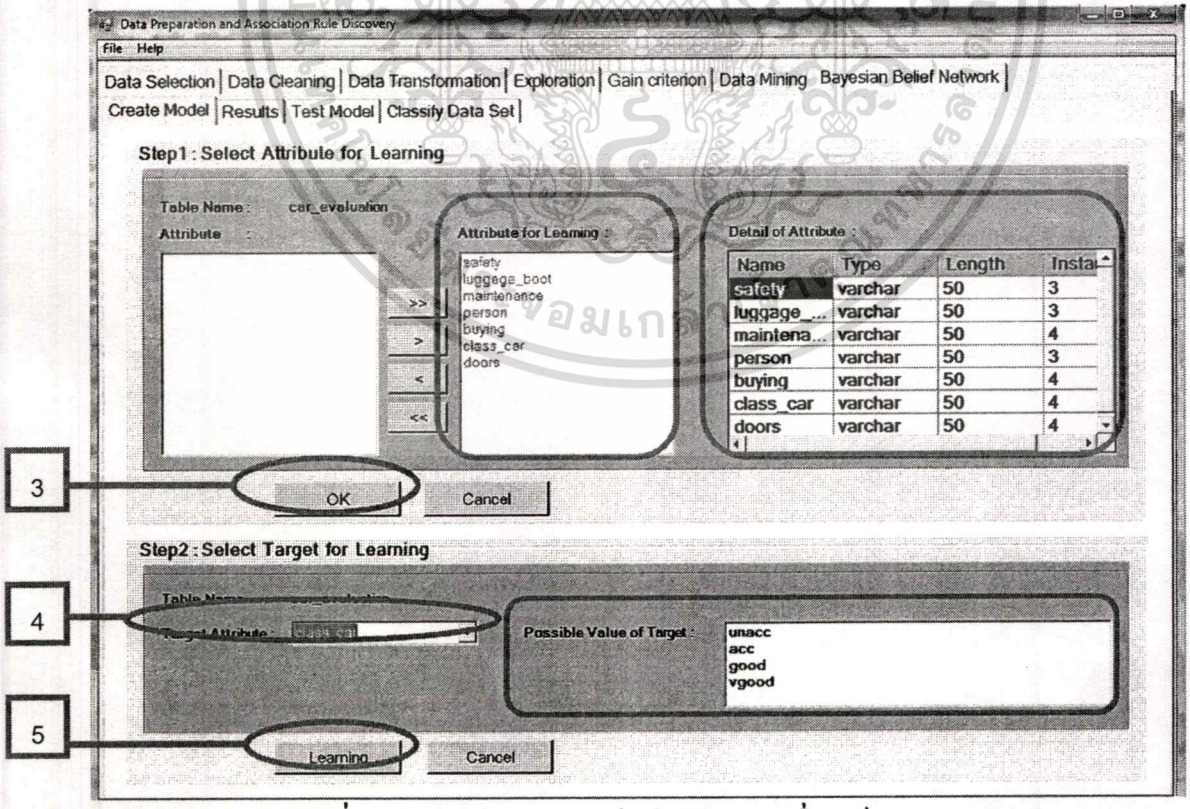
5) เลือก Learning เพื่อทำการเรียนรู้เครือข่ายความเชื่อเบย์



รูปที่ 4.2 การสร้างแบบจำลองเครือข่ายความเชื่อเบย์ของระบบ



รูปที่ 4.3 การเลือกแอททริบิวต์ที่ใช้ในการเรียนรู้เครือข่ายความเชื่อเบย์ของระบบ

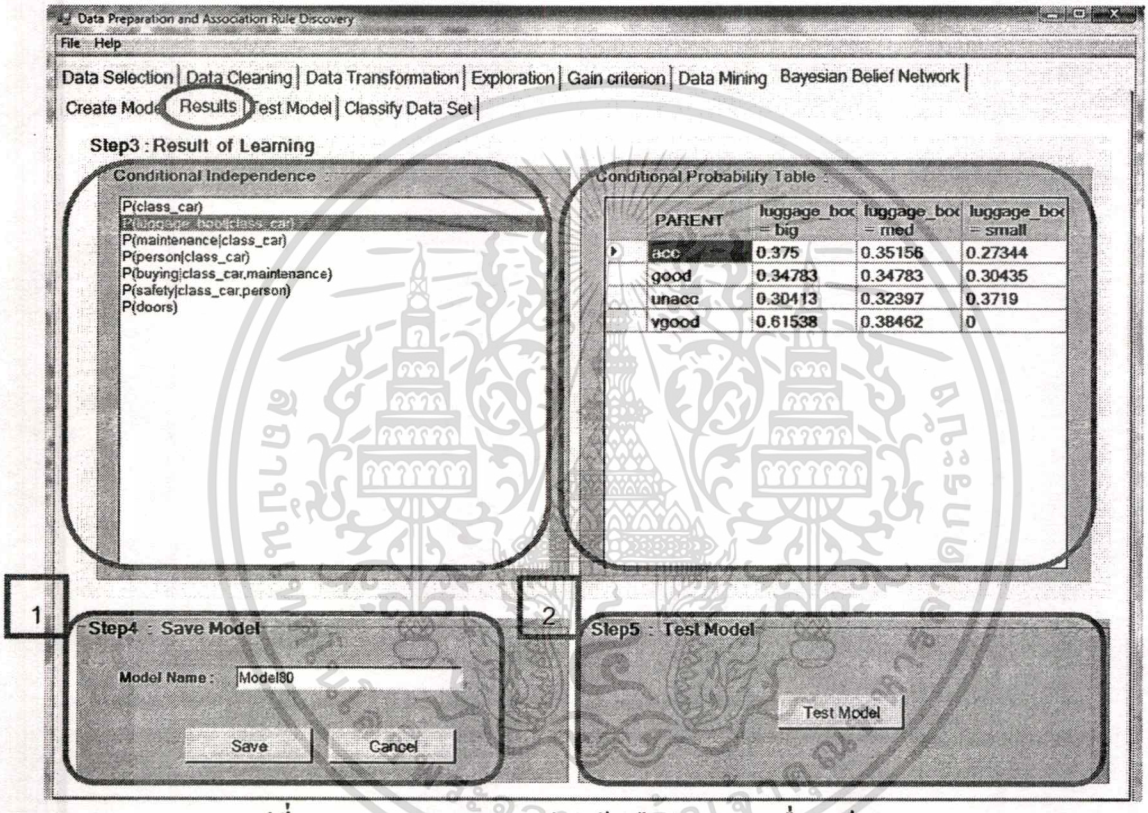


รูปที่ 4.4 การเลือกการเรียนรู้เครือข่ายความเชื่อเบย์ของระบบ

- เมื่อทำการเรียนรู้เสร็จแล้ว ระบบจะแสดงผลที่ได้จากการเรียนรู้ คือความขึ้นต่อกันระหว่างแอททริบิวต์ (Conditional Independence) และตารางความน่าจะเป็นแบบมีเงื่อนไข (Conditional Probability Table : CPT)

1) หลังจากนั้นผู้ใช้สามารถเก็บแบบจำลองไว้ใช้งานต่อไปได้ โดยใส่ชื่อแบบจำลองและเลือก Save เพื่อเก็บแบบจำลองไว้ใช้ในการทดสอบแบบจำลอง และการแบ่งประเภทของข้อมูล

2) เมื่อเก็บแบบจำลองแล้ว สามารถทดสอบแบบจำลองได้เลย โดยเลือก Test Model



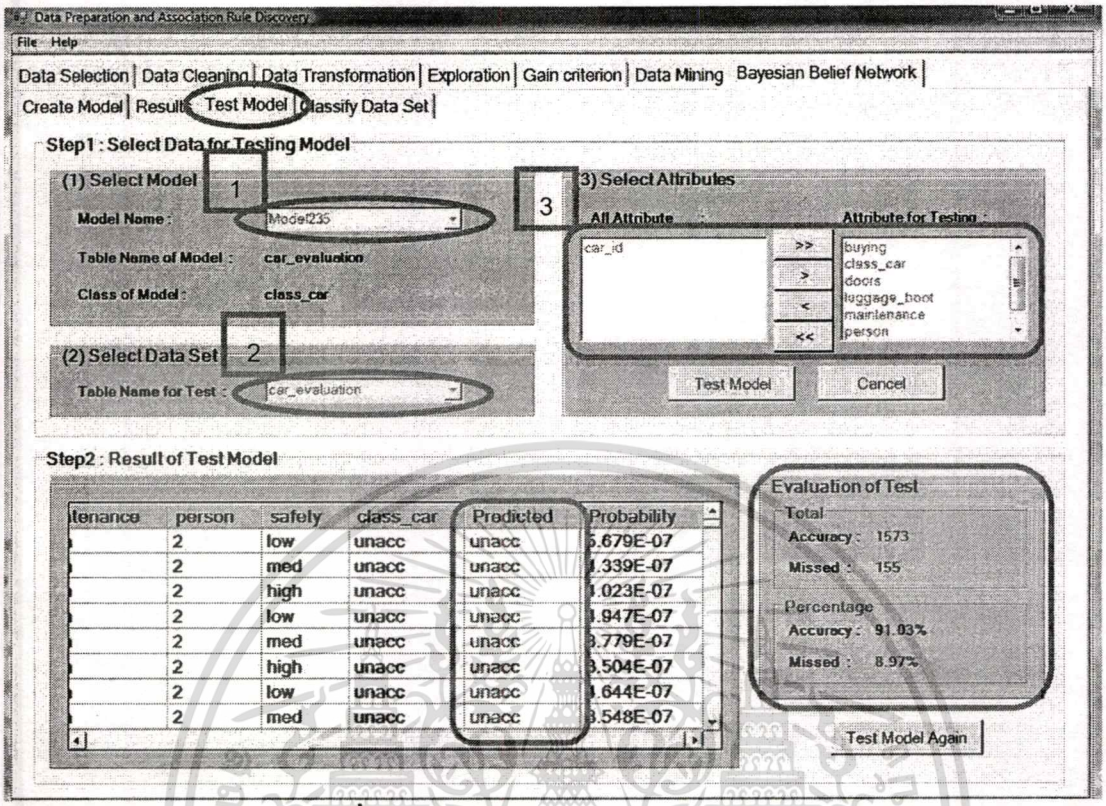
รูปที่ 4.5 การแสดงผลการเรียนรู้เครือข่ายความเชื่อเบย์ของระบบ

- เมื่อเลือกทดสอบแบบจำลองแล้ว (Test Model) ระบบจะทำการแบ่งประเภทของข้อมูลแบบเครือข่ายความเชื่อเบย์ เพื่อให้ได้ประเภทของข้อมูลใหม่ และระบบจะแสดงการประเมินแบบจำลองว่ามีความถูกต้อง และผิดพลาดเท่าไร โดยมีขั้นตอนดังนี้

1) เลือกแบบจำลองที่ต้องการทดสอบแบบจำลอง

2) เลือกชุดข้อมูลใหม่ที่ทราบประเภทของข้อมูลเข้าสู่ระบบเพื่อทดสอบแบบจำลอง

3) เลือกแอททริบิวต์ที่จำเป็นต่อการทดสอบแบบจำลอง จากนั้นระบบจะทำการแบ่งประเภทของข้อมูลใหม่ และนำประเภทของข้อมูลใหม่เปรียบเทียบกับประเภทของข้อมูลเก่า เพื่อทำการประเมินความถูกต้องของแบบจำลอง แล้วระบบจะแสดงผลการประเมินแบบจำลอง



รูปที่ 4.6 การแสดงผลการทดสอบของระบบ

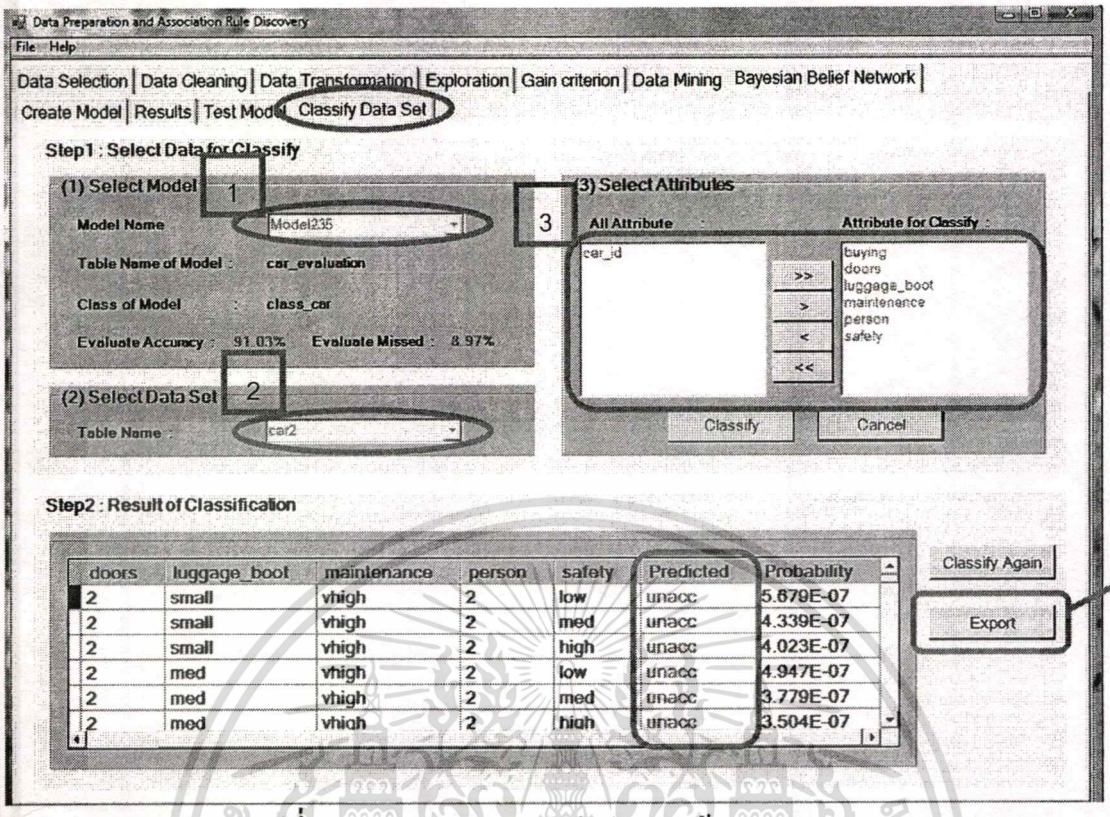
- เมื่อทดสอบแบบจำลองจนเป็นที่ยอมรับแล้ว ผู้ใช้สามารถเลือกการแบ่งประเภทของข้อมูล (Classify Data Set) ได้ โดยมีขั้นตอนดังนี้

1) เลือกแบบจำลองที่ใช้ในการแบ่งประเภทของข้อมูล

2) เลือกชุดข้อมูลใหม่เข้าสู่ระบบเพื่อทำการแบ่งประเภทของข้อมูล

3) เลือกแอททริบิวต์ที่จำเป็นต่อการแบ่งประเภทของข้อมูล จากนั้นระบบจะแสดงผลการแบ่งประเภทของข้อมูล

4) หากต้องการนำผลลัพธ์ออกจากระบบ ให้กดปุ่ม Export โดยสามารถ save ไฟล์ในรูปแบบ Excel หรือ Text หรือ CSV ก็ได้ตามต้องการ



รูปที่ 4.7 การแสดงผลการแบ่งประเภทข้อมูลของระบบ

บทที่ 5
สรุปผลการพัฒนาระบบ

ระบบเหมืองข้อมูลสำหรับวิธีการการแบ่งประเภทของข้อมูลแบบเครือข่ายความเชื่อเบย์นี้จัดทำขึ้นเพื่อแบ่งประเภทของข้อมูลจากชุดข้อมูล โดยแต่ละข้อมูลจะมีเงื่อนไขที่แตกต่างกัน ดังนั้นระบบจะทำการเรียนรู้ และคำนวณหาค่าความน่าจะเป็นในตารางความน่าจะเป็นแบบมีเงื่อนไขเพื่อนำไปใช้ในการแบ่งประเภทของข้อมูล โดยจะหาค่าความน่าจะเป็นที่มีค่ามากที่สุดตามทฤษฎีของเครือข่ายความเชื่อเบย์ และความน่าจะเป็นที่มีค่ามากที่สุดจะเป็นประเภทของข้อมูลของข้อมูลนั้นๆ

5.1 การทดสอบระบบ

- **แหล่งข้อมูล**

แหล่งข้อมูลที่นำมาใช้ในการทดสอบเป็นชุดข้อมูลที่ได้มาจาก UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/) ซึ่งมีหลายประเภท และเป็นชุดข้อมูลที่เหมาะกับการแบ่งประเภทของข้อมูล ซึ่งชุดข้อมูลแต่ละชุดจะทราบประเภทของข้อมูลมาก่อนแล้ว ทำให้ง่ายต่อการทดสอบระบบ ว่ามีความถูกต้องเพียงไร

5.2 สรุปผลการพัฒนาระบบ

การทำเหมืองข้อมูล เป็นการดึงข้อมูลจากฐานข้อมูลที่มีขนาดใหญ่เพื่อนำข้อมูลนั้นมาใช้งานให้เกิดประโยชน์สูงสุด และสามารถเข้ามาช่วยในการตัดสินใจ เพื่อค้นหารูปแบบ แนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในฐานข้อมูลที่มีขนาดใหญ่เหล่านั้น โดยอาศัยหลักการทางสถิติ การรู้จำ การเรียนรู้ของเครื่อง และหลักการทางคณิตศาสตร์ เทคนิคในการทำเหมืองข้อมูลที่น่าสนใจเทคนิคหนึ่ง คือ เครือข่ายความเชื่อเบย์ ซึ่งอธิบายความไม่ขึ้นต่อกันอย่างมีเงื่อนไข (Condition Independent) ระหว่างตัวแปร และแบบจำลองจะอยู่ในรูปของความน่าจะเป็นแบบมีเงื่อนไข โดยผลของการแบ่งประเภทของข้อมูลจะขึ้นอยู่กับชุดข้อมูลที่นำมาใช้ในการสร้างแบบจำลอง หากเป็นชุดข้อมูลที่เกิดจากกลุ่มตัวอย่างหลายๆกลุ่มแล้วจะได้แบบจำลองที่ดีได้ แต่หากเป็นชุดข้อมูลที่เกิดจากกลุ่มตัวอย่างเพียงกลุ่มเดียวที่คล้ายๆกัน จะได้แบบจำลองที่ไม่ที่น่าเชื่อถือได้ เพราะจะได้ชุดข้อมูลที่ไม่หลากหลาย และขึ้นอยู่กับแอททริบิวต์ที่นำมาใช้ในการสร้างแบบจำลอง หากเลือกแอททริบิวต์ที่ไม่เหมาะสม จะทำให้ได้แบบจำลองที่คลาดเคลื่อน และนำไปสู่การแบ่งประเภทของข้อมูลที่ไม่ถูกต้องอีกด้วย

ดังนั้น ในการทำเหมืองข้อมูลจำเป็นจะต้องเข้าใจลักษณะปัญหาที่แท้จริงก่อน เพื่อจะได้กำหนดปัญหาและขอบเขตของปัญหาได้ ถ้ากำหนดปัญหาได้ถูกต้องก็จะสามารถนำเทคนิคของการทำเหมืองข้อมูลที่ถูกต้องมาใช้และนำไปสู่ความรู้ที่เราต้องการ หรืออาจจะทำให้เราได้รับความรู้ใหม่ๆที่เก็บซ่อนอยู่ในฐานข้อมูลได้อีกด้วย

บรรณานุกรม

ผศ.ดร.บุญเสริม กิจศิริกุล. 1998 **Artificail Intelligence**. [Online]. Available : http://www.ti.com/sc/does/msp/speech/index.htm.

Charles River Analytics, Inc. 2004 **B Net Builder About Bayesian Belief Networks.** [Online]. Available : http://www.cra.com/pdf/BNetBuilderBackground.pdf

David Heckerman. 1996 **A Tutorial on Learning With Bayesian Networks."** [Online]. Available : http://research.microsoft.com/research/pubs/view.aspx?msr_tr_id=MSR-TR-95-06

Gregory F. Cooper , Edward Herskovits. 1992 **A Bayesain Method for the Induction of Probabilistic Networks from Data.** [Online].Available : http://smiweb. stanford.edu/ pubs/SMI Abstracts/SMI-91-0355.html.

Jie Cheng., Russell Greiner. 2004 **Learning Bayesian Belief Network Classifiers:Algorithms and System.**

Jie Cheng. 2004 **Belief Network (BN) PowerPredictor.** [Online]. Available : http://www.cs.ualberta.ca/~jcheng/bnpp.html

Jie Cheng. 2004 **J Cheng's Bayesian Belief Network Software.** [Online]. Available : http://www.cs.ualberta.ca/~jcheng/bnsoft.html

Murray Cumming. 1998 **Bayesian Belief Networks.** [Online]. Available : http://www.murrayc.com/learning/AI/bbn.shtml#BBNs

Prof. Carolina Ruiz ,Department of Computer Science,WPI. 1993 **Illustration of the K2 Algorithm for Learning Bayes Net Structures.** [Online].

ประวัติผู้เขียน

ชื่อผู้เขียน	นางสาวสมรศม รุ่งฟ้า
วัน เดือน ปีเกิด	16 สิงหาคม 2525
สถานที่เกิด	จังหวัดกรุงเทพมหานคร
วุฒิการศึกษาระดับปริญญาตรี	วิทยาศาสตรบัณฑิต
สถานที่สำเร็จการศึกษา	คณะวิทยาศาสตร์ สาขาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยศิลปากร
ปีการศึกษาที่สำเร็จการศึกษา	2547
ตำแหน่งหน้าที่	Programmer
สถานที่ทำงาน	บริษัท Systems Advisers Group Co., LTD จำกัด