บทที่ 2
แนวคิด ทฤษฎี และผลงานวิจัยที่เกี่ยวข้อง

2.1 การทำเหมืองข้อมูล

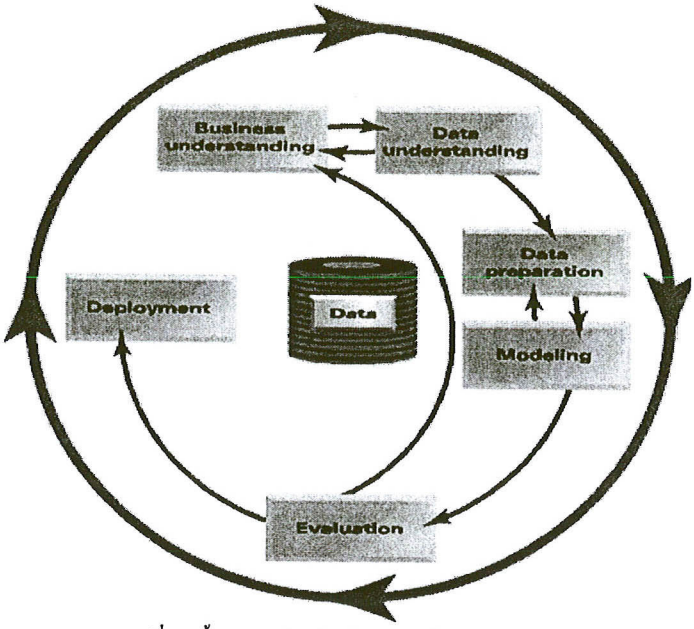
ความหมายของการทำเหมืองข้อมูล (Data Mining) มีหลายนิยามที่ให้คำจำกัดความของการทำเหมืองข้อมูลไว้ โดยสามารถสรุปหลักสำคัญได้ว่าการทำเหมืองข้อมูล คือกระบวนการค้นหาความรู้ที่เก็บอยู่ในฐานข้อมูลขนาดใหญ่ ซึ่งปัจจุบันองค์กรขนาดใหญ่มักจะนำข้อมูลต่าง ๆ ขององค์กร เช่น ข้อมูลการทำงาน ข้อมูลบุคลากร ฯลฯ เก็บไว้ในฐานข้อมูลขนาดใหญ่ เพื่อเป็นแหล่งสะสมความรู้ขององค์กร การทำเหมืองข้อมูลถูกนำไปประยุกต์ใช้กับงานหลายๆ ด้าน เช่น ด้านอุตสาหกรรม การแพทย์ การประกัน การธนาคาร การตลาด และ การเงิน ฯลฯ การทำเหมืองข้อมูลสามารถนำไปใช้เพื่อลดต้นทุน การเพิ่มยอดขาย ตลอดจนเป็นเครื่องมือในการทำวิจัย

การทำเหมืองข้อมูล เป็นขั้นตอนหนึ่งในกระบวนการค้นหาความรู้ในฐานข้อมูล (Knowledge Discovery in database: KDD) ซึ่งกระบวนการทั้งหมดจะทำการเปลี่ยนข้อมูลดิบไปเป็นความรู้ที่สามารถนำไปใช้ได้ โดยที่ข้อมูลนำเข้า (Input Data) เป็นข้อมูลที่ถูกเก็บอยู่ในรูปแบบต่าง ๆ ซึ่งอาจจะเป็น ไฟล์ข้อมูล หรือ ตารางจากฐานข้อมูล โดยที่ข้อมูลเหล่านี้อาจจะถูกเก็บอยู่ในที่เดียวกันหรืออาจอยู่ต่างที่กัน

ขั้นตอนการทำเหมืองข้อมูล ตามหลักของ CRISP-DM (Cross Reference Industry Standard for Data Mining) (Daniel T. L., 2005) ซึ่งเป็นกระบวนการมาตรฐานสำหรับการทำเหมืองข้อมูล ที่พัฒนาโดยบริษัท DaimlerChrysler, SPSS และ NCR ประกอบด้วย 6 ขั้นตอน คือ

1. การทำความเข้าใจธุรกิจ (Business Understanding) ขั้นตอนนี้เป็นขั้นตอนการทำความเข้าใจเพื่อระบุปัญหา ทำการแปลงโจทย์ที่ได้ให้อยู่ในรูปแบบที่เหมาะสมต่อการนำมาวิเคราะห์ข้อมูลทางเหมืองข้อมูล

2. การทำความเข้าใจข้อมูล (Data Understanding) เป็นการพิจารณาชุดข้อมูลที่จะนำมาใช้และการตั้งสมมุติฐานที่จะใช้ในการแก้ปัญหา รวมถึงการรวบรวมข้อมูลที่เกี่ยวข้อง โดยข้อมูลที่จะนำมาใช้ต้องเป็นข้อมูลที่น่าเชื่อถือ มีปริมาณที่เหมาะสม และมีรายละเอียดเพียงพอต่อการนำไปใช้ในการวิเคราะห์



รูปที่ 2.1 ขั้นตอนการทำเหมืองข้อมูลตามหลักของ CRISP-DM

ที่มา : Daniel T. L., 2005

3. การเตรียมข้อมูล (Data Preprocessing) เป้าหมายของขั้นตอนนี้ คือการเปลี่ยนข้อมูลนำเข้าที่เป็นข้อมูลดิบให้อยู่ในรูปที่สามารถนำไปวิเคราะห์ได้ ซึ่งถือว่าเป็นขั้นตอนที่ใช้เวลานานที่สุด เนื่องจากแบบจำลองที่ได้จากการทำเหมืองข้อมูลจะให้ผลลัพธ์ที่ถูกต้องหรือไม่นั้นขึ้นอยู่กับคุณภาพของข้อมูลที่ใช้ โดยแบ่งออกเป็น 3 ขั้นตอนย่อยคือ

- การคัดเลือกข้อมูล (Data Selection) เป็นขั้นตอนในการกำหนดเป้าหมายของการวิเคราะห์ข้อมูล เพื่อคัดเลือกเฉพาะข้อมูลที่เกี่ยวข้องกับเป้าหมายที่กำหนดไว้
- การกลั่นกรองข้อมูล (Data Cleaning) เป็นขั้นตอนการกรองข้อมูลที่ไม่ถูกต้อง ซ้ำซ้อน หรือที่ไม่จำเป็นต้องใช้ออกไป และการจัดการข้อมูลที่ไม่ครบถ้วน (Missing Data Handling)
- การแปลงรูปข้อมูล (Data Transformation) เป็นขั้นตอนการเตรียมข้อมูลให้อยู่ในรูปแบบที่พร้อมสำหรับนำไปใช้ในการวิเคราะห์ตามอัลกอริทึมของการทำเหมืองข้อมูลที่เลือกใช้

4. การพัฒนาตัวแบบ (Modeling) เป็นขั้นตอนการเลือกและประยุกต์เทคนิคที่จะใช้ทำเหมืองข้อมูล โดยขั้นตอนนี้มีการนำเทคนิคต่างๆ มาช่วยในการดึงรูปแบบที่ซ่อนอยู่ออกมา การทำเหมืองข้อมูลสามารถแบ่งอัลกอริทึมในการทำงานได้เป็น 2 ประเภทหลัก คือการสร้างแบบจำลองแบบทำนาย เน้นการจัดกลุ่มโดยอาศัยผลเฉลยที่มีอยู่ซึ่ง เช่น การจำแนกและการวิเคราะห์การถดถอย ส่วนประเภทที่สองคือ การสร้างแบบจำลองเชิงพรรณนา เป็นการหาความสัมพันธ์โดยไม่มีผลเฉลย

5. การแปลผลและการประเมินผล (Evaluation) ขั้นตอนนี้เป็นการประเมินประสิทธิภาพของผลลัพธ์จากโมเดลวิเคราะห์ข้อมูลว่าครอบคลุมและสามารถตอบโจทย์ทางธุรกิจที่ตั้งไว้ในขั้นตอนแรกหรือไม่

6. การนำไปใช้ (Deployment) เป็นการวางแผนและการนำตัวแบบเหมืองข้อมูลไปใช้ประโยชน์

2.2 การจำแนกประเภทข้อมูล (Data Classification)

การจำแนกประเภทข้อมูล (Han and Kamber, 2000) เป็นวิธีการหนึ่งที่นิยมในการทำเหมืองข้อมูล ซึ่งเป็นเทคนิคในการจำแนกข้อมูลให้อยู่เป็นกลุ่มๆ ตามที่กำหนด โดยจะนำข้อมูลส่วนหนึ่งมาสอนให้ระบบเรียนรู้ซึ่งเรียกว่าข้อมูลฝึก เพื่อจำแนกข้อมูลออกเป็นกลุ่มตามที่กำหนดไว้ ผลลัพธ์ที่ได้คือแบบจำลองการจำแนกประเภทข้อมูล (Classification Model) และนำข้อมูลในส่วนที่เหลือเป็นข้อมูลทดสอบ เพื่อทดสอบแบบจำลองโดยเปรียบเทียบผลจากกลุ่มที่ใช้ในการเรียนรู้และกลุ่มที่ใช้ในการทดสอบ หากผลการทดสอบยังไม่เป็นที่น่าพอใจก็สามารถปรับปรุงแบบจำลองจนกว่าจะได้ค่าความถูกต้องในระดับที่น่าพอใจ จากนั้นเมื่อมีข้อมูลใหม่ที่ไม่ทราบประเภทก็สามารถนำข้อมูลรายการใหม่นั้นมาผ่านแบบจำลอง และจำแนกประเภทของข้อมูลใหม่นั้นได้

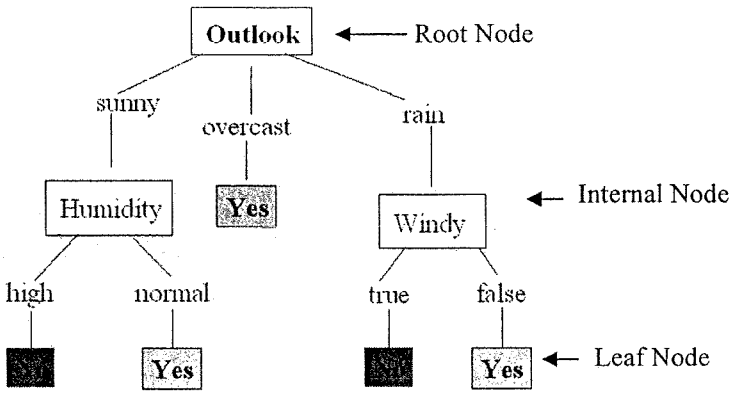
2.3 ทฤษฎีแบบจำลองต้นไม้การตัดสินใจ (Decision Tree)

ต้นไม้การตัดสินใจ เป็นเทคนิคหนึ่งของการจำแนกข้อมูล ถูกนำมาใช้ในการจำแนกข้อมูลเนื่องจากมีโครงสร้างคล้ายกับต้นไม้ ทำให้เข้าใจได้ง่าย ประกอบไปด้วย

- โหนด (Node) แสดงแอตทริบิวต์ (Attribute) หรือเงื่อนไขที่ใช้ในการตัดสินใจ โหนดแบ่งเป็น 3 ประเภท คือ
  - รูทโหนด (Root Node) เป็นโหนดบนสุดซึ่งจะไม่มีกิ่ง (Branch) เข้ามาหา
  - อินเตอร์นอลโหนด (Internal Node) เป็นโหนดภายในที่มีกิ่งเข้ามาและออกไป
  - โหนดใบ (Leaf Node) แสดงคลาสที่กำหนดไว้

- กิ่ง (Branch) คือเส้นเชื่อมความสัมพันธ์ระหว่างโหนด ซึ่งใช้แสดงผลในการทดสอบ

การสร้างต้นไม้การตัดสินใจทำโดยการเลือกชุดของแอตทริบิวต์ ที่มีอิทธิพลในการจำแนกสูงสุด มาสร้างเป็นรูทโหนดก่อนในขั้นตอนแรก จากนั้นก็จะหาแอตทริบิวต์ ที่มีอิทธิพลในการจำแนกสูงสุดตัวถัดไป (ซึ่งอาจเป็นตัวเดิมก็ได้) มาสร้างเป็นโหนดและกิ่งไปเรื่อย ๆ จนกระทั่งเงื่อนไขในการเติบโตของต้นไม้ตัดสินใจจะเป็นจริง



รูปที่ 2.2 แสดงตัวอย่างของต้นไม้การตัดสินใจ

ขั้นตอนการสร้างต้นไม้การตัดสินใจ

1) เลือกแอตทริบิวต์ ที่ดีที่สุดสำหรับใช้ในการตัดสินใจ
2) กำหนดค่าให้กับแอตทริบิวต์ที่เลือกเพื่อสร้างโหนดปัจจุบัน
3) สร้างโหนดลูกของโหนดปัจจุบัน โดยเลือกจากแอตทริบิวต์ที่เหลืออยู่
4) จัดเรียงข้อมูลฝึก และจำแนกไปยังโหนดที่สร้างขึ้นในข้อ 3)
5) ถ้าข้อมูลจำแนกได้สมบูรณ์ให้หยุดทำงาน ถ้ายังไม่สมบูรณ์ให้กลับไปทำข้อ 1)

ในงานวิจัยนี้ผู้วิจัยใช้อัลกอริทึม C4.5 (Quinlan,1993) ในการสร้างต้นไม้การตัดสินใจ ซึ่งอัลกอริทึมนี้ใช้ค่าอัตราส่วนเกนในการพิจารณาเลือกแอตทริบิวต์สำหรับการแตกโหนด โดยค่าอัตราส่วนเกนสามารถคำนวณได้ดังสมการ (1)

$$\text{GainRATIO}_{\text{split}} = \frac{\text{Gain}_{\text{split}}}{\text{SplitINFO}} \quad (1)$$

ซึ่งค่าอัตราส่วนเกนเป็นการแก้ปัญหาของ Information Gain ด้วยการทำ Normalize ค่า Information Gain โดยถึงแม้ว่าแอตทริบิวต์ตัวใดตัวหนึ่งทำให้มีโหนดลูกจำนวนมากก็จะถูกปรับค่าด้วย SplitINFO ซึ่งสามารถคำนวณได้ดังสมการ (2)

$$\text{SpiltINFO} = -\sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n} \quad (2)$$

สมมติว่าโหนดแม่ p ถูกแตกเป็นโหนดลูกจำนวน k โหนด

n_i คือจำนวนตัวอย่างในโหนดลูกแต่ละโหนด

n คือจำนวนตัวอย่างที่โหนดแม่ p

ปัญหาในการสร้างแบบจำลองต้นไม้การตัดสินใจมี 2 ปัญหาได้แก่ อันเดอร์ฟิตติ้ง (Underfitting) และโอเวอร์ฟิตติ้ง (Overfitting) อันเดอร์ฟิตติ้งเป็นปัญหาที่เกิดจากแบบจำลองง่ายจนเกินไปซึ่งเกิดจากการเรียนรู้จากข้อมูลที่น้อยเกินไป ทำให้แบบจำลองที่เกิดขึ้นไม่สมบูรณ์พอที่จะนำไปใช้กับข้อมูลที่ไม่เคยพบได้ ส่วนปัญหาโอเวอร์ฟิตติ้ง คือการที่แบบจำลองที่ได้จากการใช้ข้อมูลฝึก มีค่าความถูกต้องในการบ่งบอกคลาสเป้าหมายสูง แต่เมื่อนำไปใช้กับข้อมูลทดสอบได้ค่าความถูกต้องต่ำ หรืออีกนัยหนึ่งคือแบบจำลองที่ได้เป็นการเรียนรู้ข้อมูลที่ใช้สอนหรือข้อมูลฝึกดีมาก แต่ไม่สามารถนำไปใช้กับข้อมูลที่ไม่เคยพบมาก่อนได้ดี ต้นไม้ที่แสดงลักษณะปัญหาโอเวอร์ฟิตติ้งมักเป็นต้นไม้ที่มีการแตกกิ่งมากมาย แต่ละกิ่งมีจำนวนข้อมูลอยู่ปริมาณน้อย วิธีการหลีกเลี่ยงปัญหานี้ทำได้ 2 ลักษณะ

- การตัดเล็มขณะที่เรียนรู้ (Pre-Pruning) หยุดการสร้างเมื่อตัววัดการแตกกิ่งมีค่าต่ำกว่าเกณฑ์ที่กำหนด เช่นข้อมูลในจุดยอดมีน้อยเกินไป
- การตัดเล็มหลังการเรียนรู้ (Post-Pruning) การลดจำนวนกิ่ง กำจัดกิ่งโดยยังคงความถูกต้องของการจำแนกประเภทข้อมูล (Classification) ในระดับที่ยอมรับได้

ข้อดีของการสร้างแบบจำลองด้วยต้นไม้การตัดสินใจ

- ง่ายต่อการทำความเข้าใจ เนื่องจากแบบจำลองถูกแสดงในรูปของต้นไม้ทำให้สามารถสร้างกฎ (Rule) แสดงความสัมพันธ์ได้ง่าย
- สามารถทำนายหรือจำแนกประเภทของข้อมูลได้ทั้งข้อมูลที่เป็นตัวเลข (Numeric) และเป็นข้อมูลที่เป็นชนิดของกลุ่ม (Category)

2.4 งานวิจัยที่เกี่ยวข้อง

กฤษณะ ไวยมัย, ชิดชนก ส่งศิริ, และธนาวินท์ รักธรรมานนท์ (2001) ได้ใช้เทคนิคการจำแนกข้อมูลในรูปของต้นไม้การตัดสินใจ มาประยุกต์ใช้ในการช่วยให้นักศึกษาเลือกสาขาวิชาที่เหมาะสม โดยนำข้อมูลของนิสิตที่เรียนดีทุกวิชามาสร้างแบบจำลองกลางการจำแนกประเภทข้อมูล ซึ่งแต่ละโหนดของต้นไม้จะเป็นลักษณะและผลการเรียนในรายวิชาต่างๆของนิสิต เพื่อทำนายว่าลักษณะของนิสิตแต่ละคนจะมีความคล้ายคลึงกับนิสิตที่เรียนดีในสาขาวิชาใดมากที่สุด แต่ผลที่ได้ยังมีความถูกต้องต่ำ คือประมาณร้อยละ 50 จึง

ได้สร้างแบบจำลองการจำแนกประเภทข้อมูลสำหรับแต่ละสาขา โดยพิจารณาว่านิสิตเหมาะสมกับสาขานั้นหรือไม่ ซึ่งนิสิตที่มีผลการเรียนที่ดีอาจจะมีหลายสาขาวิชาที่เหมาะสมให้เลือก นิสิตก็สามารถเลือกสาขาวิชาที่ดีที่สุดโดยพิจารณาจากสัดส่วนของคลาสปลายทางที่ประกอบด้วย GOOD และ BAD ซึ่งผลการทดสอบมีความถูกต้องคิดเป็นร้อยละ 84.58 ในทุกแบบจำลอง

ชลนิศา สาระ (2550) เสนอวิธีการจำแนกกลุ่มสถานภาพการสำเร็จการศึกษาโดยแบบจำลองต้นไม้การตัดสินใจ ซึ่งใช้อัลกอริทึม C4.5 ทำให้เข้าใจถึงกระบวนการ พารามิเตอร์ที่ใช้ และการวัดประสิทธิภาพในการสร้างแบบจำลองต้นไม้สำหรับอัลกอริทึม C4.5

บุณยวีร์ บุณยมานพ (2551) ใช้เทคนิคการทำเหมืองข้อมูล เพื่อสร้างแบบจำลองการจำแนกประเภทข้อมูลของค่าใช้จ่ายเนื่องจากการใช้อินเทอร์เน็ต เพื่อจุดประสงค์ส่วนบุคคลของพนักงานในองค์กรจาก Web log โดยการเปรียบเทียบอัลกอริทึมต้นไม้การตัดสินใจ C4.5 และ Multilayer perceptron

วิรัช ศรเลิศล้ำวาณิช (2000) ใช้เทคนิคการทำเหมืองข้อมูลในทำการสกัดคำภาษาไทย โดยใช้ต้นไม้การตัดสินใจอัลกอริทึม C 4.5 ในการแยกข้อมูลที่เป็นคำและไม่ใช่คำ ที่ไม่กำกวมออกจากกัน จากข้อความที่มีอยู่ทั้งหมดใน 75 บทความ ซึ่งเป็นบทความจากหลากหลายแขนงวิชา ผลการทดสอบได้ค่า Precision คิดเป็นร้อยละ 85 และค่า Recall คิดเป็นร้อยละ 56 โดยที่ค่า Precision และค่า Recall ของทั้งข้อมูลฝึก และข้อมูลทดสอบใกล้เคียงกัน แสดงว่าต้นไม้การตัดสินใจนี้สามารถนำไปประยุกต์ใช้กับข้อมูลที่ไม่เคยพบได้ดี

Hyontai Sug (2006) เสนอวิธีหาความแม่นยำในการตัดสินใจจากเงื่อนไขของแอตทริบิวต์ เพื่อวัดความสามารถของแบบจำลองการจัดกลุ่มหรือการจำแนกประเภท โดยทดลองกับ Conflict Dataset จากการสุ่มข้อมูล และทดสอบกับค่า Confidence Factor 5 ค่า คือ 0.001, 0.01, 0.15, 0.25, และ 0.5 แล้วทำการเปรียบเทียบขนาด และอัตราการเกิดข้อผิดพลาด (Error Rate) ผลการทดลองจากการจำแนกประเภท พบว่าแอตทริบิวต์ ที่ใช้มีผลต่อการสร้างแบบจำลองที่มีความถูกต้องและความแม่นยำ