บทที่ 2
แนวคิดและทฤษฎี

2.1 ความหมายของ Data Mining

Data Mining เป็นกระบวนการของการกลั่นกรองสารสนเทศ (Information) ที่ซ่อนอยู่ในฐานข้อมูลใหญ่ เพื่อทำนายแนวโน้มและพฤติกรรม โดยอาศัยข้อมูลในอดีต และเพื่อใช้สารสนเทศเหล่านี้ในการสนับสนุนการตัดสินใจทางธุรกิจ

วิวัฒนาการของ Data Mining

ปี 1960 Data Collection คือ การนำข้อมูลมาจัดเก็บอย่างเหมาะสมในอุปกรณ์ที่น่าเชื่อถือและป้องกันการสูญหายได้เป็นอย่างดี

ปี 1980 Data Access คือ การนำข้อมูลที่จัดเก็บมาสร้างความสัมพันธ์ต่อกันในข้อมูลเพื่อประโยชน์ในการนำไปวิเคราะห์ และการตัดสินใจอย่างมีคุณภาพ

ปี 1990 Data Warehouse & Decision Support คือ การรวบรวมข้อมูลมาจัดเก็บลงไปในฐานข้อมูลขนาดใหญ่โดยครอบคลุม ทุกแง่ทุกมุมขององค์กร เพื่อช่วยสนับสนุนการตัดสินใจ

ปี 2000 Data Mining คือ การนำข้อมูลจากฐานข้อมูลมาวิเคราะห์และประมวลผล โดยการสร้างแบบจำลอง และความสัมพันธ์ ทางสถิติ

จากคำจำกัดความ Data Mining หมายถึงการที่ผู้ใช้ดึงข้อมูลโดยการสังเคราะห์และตรวจสอบข้อมูลอย่างละเอียด โดยการสังเคราะห์ดังกล่าวอาจเป็นการเรียนรู้ข้อมูลในอดีตหรือข้อมูลในปัจจุบัน ผลลัพธ์ที่ได้มาต้องมีลักษณะของข้อมูลที่เป็นข้อมูลแบบ Unknown, ข้อมูลแบบ Valid, และข้อมูลแบบ Actionable มาจากฐานข้อมูลขนาดใหญ่ซึ่งมาจากรายการ Transaction, ฐานข้อมูลของฝ่ายขาย, E-Mail เพื่อนำข้อมูลดังกล่าวไปใช้เป็นพื้นฐานในการ ประกอบการตัดสินใจ ในเชิงธุรกิจ ทำให้เข้าใจแนวโน้มและรูปแบบของตลาด

ข้อมูลแบบ Unknown ข้อมูลที่ถูกใช้จะต้องเป็นข้อมูลผู้ใช้งานไม่รู้มาก่อนและไม่ชัดเจนจนไม่สามารถตั้งสมมติฐานได้ล่วงหน้าว่าควรจะเป็นแบบใด ตัวอย่างเช่น เจ้าของห้างสรรพสินค้าแห่งหนึ่งเพิ่งจะค้นพบว่าพฤติกรรมของผู้บริโภคใหม่ที่เป็นพ่อบ้าน มักจะซื้อสินค้าประเภทเบียร์และผ้าอ้อมในวันศุกร์ตอนเย็น ดังนั้นเพื่อเป็นสัญญาณให้เจ้าของกิจการควรจะเตรียมสินค้าไว้เพื่อจำหน่าย ในขณะเดียวกันห้างสรรพสินค้าคู่แข่งอาจจะไม่รู้เรื่องนี้ หรือตัวอย่างของเจ้าของร้านขายรถยนต์พบว่ารถขนาดใหญ่ราคาแพงมักจะมีผู้ซื้อเป็นผู้สูงอายุ ซึ่งเจ้าของร้านไม่รู้มาก่อน แต่ข้อมูลดังกล่าวไม่

เป็นลักษณะ Unknown ตามสมมติฐานดังกล่าวมีอยู่ เพราะคนที่มีอายุมักจะมีฐานะที่ดีขึ้นเมื่อเทียบกับคนในวัยที่อายุน้อยกว่า

ข้อมูลแบบ Valid เมื่อผู้ใช้ได้เริ่มใช้เทคนิค Data Mining จะค้นพบสิ่งที่น่าสนใจโดยต้องพิจารณา ด้วยว่าสิ่งนั้น Valid หรือไม่ เช่น ผู้ใช้มักจะพบว่ามีความสัมพันธ์ของการซื้อของ 2 สิ่งเสมอ เมื่อจำนวนความหลากหลายของสินค้ามีมากขึ้น แต่ไม่ได้หมายความว่าจะต้องให้ห้างสรรพสินค้าเก็บสินค้ามากขึ้น เพราะข้อมูลที่ได้อาจเกิดความคลาดเคลื่อน เพราะฉะนั้นจะต้องทำการ Validation และ Checking ความถูกต้องของข้อมูลและวิเคราะห์ความถูกต้องอีกครั้ง

ข้อมูลแบบ Actionable ข้อมูลจะต้องถูกแปลงออกมาและนำมาตัดสินใจให้เป็นความได้เปรียบเชิงธุรกิจ บางครั้ง ข้อมูลที่เราค้นพบเป็นสิ่งที่คู่แข่งได้ทำไปแล้ว จึงต้องมีวิจารณญาณในการใช้ด้วย บางทีข้อมูลดังกล่าว อาจจะไม่มีประโยชน์ก็ได้

คำว่า Data Mining นั้นมีความหมายแตกต่างกันใน 2 แง่มุม คือ ในมุมมองทางวิชาการและในมุมมองเชิงธุรกิจ ในมุมมองเชิงวิชาการนั้น นักวิจัยจะอ้างถึงกระบวนการทั้งหมดในการทำ Data Mining ว่า “Knowledge discovery in database (KDD)” และใช้คำว่า “Data Mining” แทนขั้นตอนขั้นหนึ่งของกระบวนการที่เกี่ยวข้องกับการค้นหารูปแบบ ความสัมพันธ์ของข้อมูลเท่านั้น อย่างไรก็ตาม ในแง่มุมเชิงธุรกิจแล้ว จะใช้คำว่า “Data Mining” แทนความหมายของ ขั้นตอนทั้งหมด เดิมงานค้นคว้าทางด้าน Data Mining นั้นมีการทำการค้นคว้ากันอยู่แล้วในหลาย ๆ สาขาวิชา แต่มีชื่อเรียก แตกต่างกันไปตามแต่ละด้าน นักวิจัยในด้านสถิติ (statistics) , ฐานข้อมูล (database) , neural networks , pattern recognition , machine learning , econometrics และอีกหลาย ๆ ด้าน ต่างก็มีการค้นคว้าเกี่ยวกับปัญหาในลักษณะเดียวกันนี้ แต่ยังไม่ค่อยมีการใช้ประโยชน์ของการค้นคว้าของอีกฝ่ายหนึ่ง คือ ต่างฝ่ายต่างทำการค้นคว้าของตนเอง ไม่ค่อยมีการแลกเปลี่ยนความรู้กัน ทำให้การค้นคว้าและการเผยแพร่ผลงานดำเนินไปอย่างไม่รวดเร็วเท่าที่ควร ต่อมาจึงมีการใช้ “Data Mining” เป็นชื่อรวม ของวิธีการแก้ปัญหาในลักษณะนี้ ซึ่งทำให้การเผยแพร่ความรู้ในการแก้ปัญหาลักษณะนี้ทำได้รวดเร็วและสามารถอ้างอิงได้ สะดวกขึ้น

หลักการทั่วไปของ Knowledge Discovery in Database (KDD) and Data Mining

KDD หมายถึงกระบวนการในการค้นหาลักษณะแฝงของข้อมูลที่อยู่ในกลุ่มข้อมูลจำนวนมาก ซึ่งมีขั้นตอนการทำ Data Mining เป็นกระบวนการที่สำคัญในการค้นหาลักษณะที่น่าสนใจของข้อมูลเหล่านี้ เช่น รูปแบบ ความสัมพันธ์ การเปลี่ยนแปลง โครงสร้างที่เด่นชัด หรือ ลักษณะที่ผิดปกติของข้อมูลจากข้อมูลจำนวนมากๆ ที่เก็บอยู่ในฐานข้อมูล หรือแหล่งที่เก็บข้อมูลอื่นๆ ซึ่งวิธีการต่างๆ ที่นำมาใช้ในการทำ mining นี้ก็มีวัตถุประสงค์ต่างๆกันขึ้นอยู่กับผลลัพธ์ของ กระบวน

การโดยรวมที่ต้องการ ดังนั้นจึงควรมีการนำเสนอวิธีการที่หลากหลายสำหรับเป้าหมายที่แตกต่างกัน เพื่อให้ได้ผลลัพธ์ ที่เหมาะสมตามที่ต้องการ หลังจากนำไปใช้งานแล้ว และเนื่องจากความแพร่หลายของการจัดเก็บข้อมูลในลักษณะที่เป็น รูปแบบทางอิเล็กทรอนิกส์ และความต้องการในการเปลี่ยนข้อมูลเหล่านั้นให้เป็นข้อมูลที่มีประโยชน์ต่อการนำไปประยุกต์ ใช้ในงานด้านต่างๆ เช่น การวิเคราะห์ด้านการตลาด การบริหารธุรกิจ รวมถึงระบบที่ช่วยสนับสนุนการตัดสินใจ เป็นต้น ดังนั้นจึงทำให้การนำ data mining มาใช้ได้รับความสนใจมากในช่วง 2-3 ปีที่ผ่านมา

จากที่ได้กล่าวแล้วว่า Data Mining เป็นขั้นตอนหนึ่งที่สำคัญในกระบวนการค้นหาลักษณะแฝงของข้อมูล ที่มีประโยชน์ในฐานข้อมูล (Knowledge Discovery in Database: KDD) ซึ่งโดยทั่วไปกระบวนการของ KDD นั้นประกอบด้วยขั้นตอนต่างๆ ดังนี้

1. การคัดเลือกข้อมูล (Data Selection) เป็นการระบุถึงแหล่งข้อมูลที่จะนำมาใช้ในการทำ mining รวมถึง การนำข้อมูลที่ต้องการออกมาจากฐานข้อมูลเพื่อทำการพิจารณาในเบื้องต้นต่อไป

2. การกรองข้อมูล (Data Cleaning) เป็นกระบวนการที่ทำให้เกิดความมั่นใจในคุณภาพของข้อมูลที่จะนำมาใช้วิเคราะห์ว่าถูกต้อง โดยการนำข้อมูลที่ไม่ถูกต้องออก

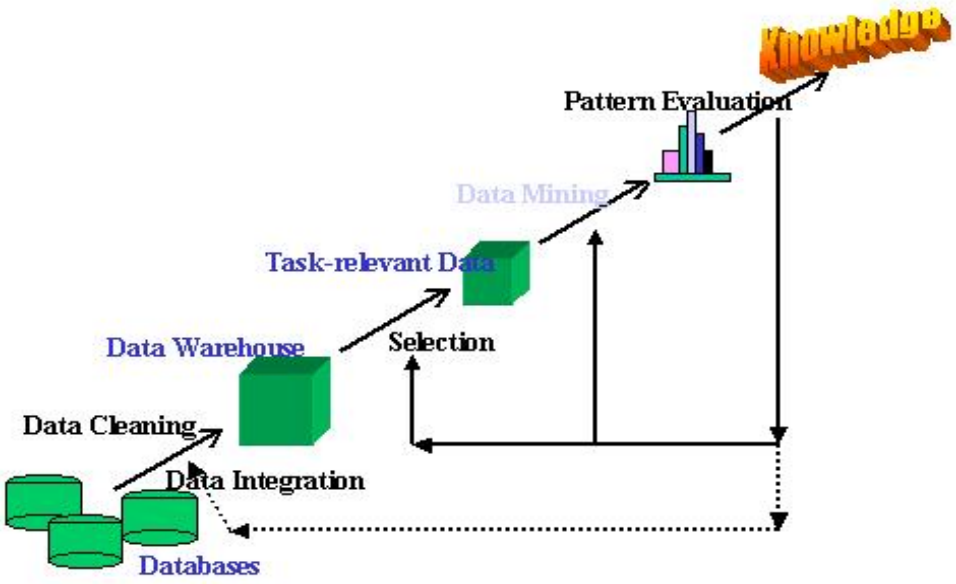
3. การแปลงรูปแบบข้อมูล (Data Transformation) เป็นการแปลงข้อมูลที่เลือกมาให้อยู่ในรูปแบบที่เหมาะสม สำหรับการนำไปใช้วิเคราะห์ตามอัลกอริทึม (Algorithm) และแบบจำลองที่ใช้ในการทำ data mining ต่อไป

4. การทำ Mining ข้อมูล (Data Mining) การใช้เทคนิคภายใน Data Mining เพื่อทำการ Mine ข้อมูล โดยทั่วไป ประเภทของงานตามลักษณะของแบบจำลองที่ใช้ในการทำ Data Mining นั้นสามารถแบ่งกลุ่มได้เป็น 2 ประเภทใหญ่ๆ คือ

4.1 Predictive Data Mining คือ เป็นการคาดคะเนลักษณะหรือประมาณค่าที่ชัดเจนของข้อมูลที่จะเกิดขึ้น โดยใช้พื้นฐานจากข้อมูลที่ผ่านมาในอดีต

4.2 Descriptive Data Mining คือ เป็นการหาแบบจำลองเพื่ออธิบายลักษณะบางอย่างของข้อมูลที่มีอยู่ ซึ่งโดยส่วนมากจะเป็นลักษณะการแบ่งกลุ่มให้กับข้อมูล

5. การวิเคราะห์และประเมินผลลัพธ์ที่ได้ (Result Analysis and Evaluation) เป็นขั้นตอนการแปลความหมายและการประเมินผลลัพธ์ที่ได้ว่ามีความเหมาะสมหรือตรงกับวัตถุประสงค์ที่ต้องการหรือไม่ โดยทั่วไปควรมีการแสดงผลในรูปแบบที่สามารถเข้าใจได้โดยง่าย



รูป 2.1 แสดงขั้นตอนต่างๆ ของกระบวนการ KDD

2.2 ขั้นตอนการทำงานของ Data Mining

1. การกำหนดวัตถุประสงค์ทางธุรกิจ (Problem formulation)

การกำหนดวัตถุประสงค์ทางธุรกิจ คือ จะต้องเข้าใจปัญหาและความต้องการทางธุรกิจ จะเป็นส่วนที่กำหนดว่าเมื่อไหร่ที่จะใช้ Data Mining ในการแก้ปัญหาซึ่งในส่วนนี้จะประกอบด้วยการวิเคราะห์ทางธุรกิจ และการวิเคราะห์เบื้องต้นว่าเรามีข้อมูลใดอยู่บ้าง และต้องการอะไรจากข้อมูล ซึ่งขั้นตอนนี้จะสามารถมองถึง อัลกอริทึม และฐานข้อมูลที่สัมพันธ์กับวัตถุประสงค์ทางธุรกิจได้

การใช้งาน Data Mining ให้ได้ประโยชน์สูงสุดจำเป็นต้องมีการกำหนดวัตถุประสงค์ที่ชัดเจน เช่น ต้องการ เพิ่มยอดการตอบรับการขายทางจดหมาย ขึ้นอยู่กับการระบุเป้าหมายว่า จะเพิ่มอัตราการตอบรับหรือเพิ่มมูลค่าการตอบรับซึ่ง จำเป็นที่จะต้องสร้าง Model ที่แตกต่างกัน วัตถุประสงค์ที่กำหนดขึ้นมาจะต้องมีการระบุวิธีการในการวัดผลลัพธ์ที่ได้จาก โครงการ รวมถึงต้นทุนที่สมเหตุสมผลด้วย

2. การคัดเลือกและการเตรียมข้อมูล (Data selection and preparation)

การเตรียมข้อมูล (Data Preparation)

เป็นหัวใจของขั้นตอนในการทำทั้งหมด เป็นช่วงที่ใช้เวลามากที่สุดในขั้นตอนโดยปกติและต้องการเวลาประมาณ 60% ของเวลาทั้งหมดในการเตรียมข้อมูล ในขั้นตอนนี้อาจสามารถแบ่งออกได้เป็นขั้นตอนย่อยดังต่อไปนี้

การเลือกข้อมูล (Data Selection)

จุดประสงค์คือการระบุแหล่งของข้อมูลที่มีและทำการดึงเอาข้อมูลออกมาใช้สำหรับการวิเคราะห์เบื้องต้นในการ เตรียมตัวสำหรับการที่จะทำการ Mining ในขั้นต่อ ๆ ไป การเลือกข้อมูลนั้นจะแตกต่างไปตามวัตถุประสงค์ของแต่ละธุรกิจ ที่ได้กำหนดไว้ตั้งแต่ต้น และการเลือกข้อมูลก็ยังถูกกำหนดโดยลักษณะงานที่จะถูกนำมาใช้อีกด้วย

ตัวแปรที่ถูกเลือกมาแต่ละตัวนั้นจะต้องถูกทำความเข้าใจว่าตัวแปรแต่ละตัวหมายความว่าอะไร ประกอบด้วยอะไร ไม่เพียงแต่คำจำกัดความทางธุรกิจเท่านั้น แต่จะต้องมีคำอธิบายอย่างชัดเจนเกี่ยวกับชนิดของข้อมูล, ค่าที่เป็นไปได้, แหล่งกำเนิดของข้อมูล, รูปแบบของข้อมูล และลักษณะอื่นๆ จะมีตัวแปร 2 ชนิดคือ

- ตัวแปรแบบ Categorical
    1. Nominal Variable กล่าวถึงชนิดนี้ของ Object ที่มันอ้างถึงแต่ไม่มีลำดับ ในค่าที่เป็นไปได้(Possible Value) ตัวอย่างเช่น สถานะการแต่งงาน (โสด, แต่งงาน, หย่า, ไม่ทราบ), เพศ (ชาย, หญิง), ระดับการศึกษา (ปริญญาโท, ปริญญาตรี, ม. ปลาย, ปวช)
    2. Ordinal Variable มีลำดับสำหรับค่าที่เป็นไปได้ ตัวอย่างเช่น ลำดับของ ลูกค้า (ดี, ปานกลาง, ไม่ดี)
- ตัวแปรแบบ Quantitative ซึ่งมีการวัดความแตกต่างระหว่างค่าที่เป็นไปได้
    1. Continuous (ค่าที่ต่อเนื่อง) เช่นรายได้, เฉลี่ยจำนวนครั้งที่ซื้อ, รายได้
    2. Discrete (ค่าเป็นจำนวนเต็ม) เช่นจำนวนพนักงาน, เวลาปี (เดือน, ฤดู, ไตรมาส)

ตัวแปรของข้อมูลมีหลายตัวมากแต่ตัวแปรที่ถูกเลือกสำหรับทำ Data Mining นั้นถูกเรียกว่า “Active Variable” เพราะว่ามันจะถูกใช้สร้างความแตกต่างของกลุ่มย่อยต่างๆ และสามารถถูกนำมาทำนายผลได้ เมื่อคุณทำการเลือกข้อมูลจะต้อง พิจารณาอายุของข้อมูลด้วย เพราะว่าสถานการณ์ภายนอกเปลี่ยนแปลงตลอดเวลาซึ่งจะทำให้ประสิทธิภาพของการทำ Mining ลดลง ตัวอย่าง รสนิยมการใช้ชีวิต การเปลี่ยนงาน

การกลั่นกรองข้อมูล (Data Preprocessing)

จุดประสงค์ก็เพื่อทำให้มั่นใจว่าคุณภาพของข้อมูลที่ถูกเลือกนั้นเหมาะสม ข้อมูลที่สมบูรณ์เป็นเครื่องประกันว่าการทำ Data Mining จะสำเร็จ ในขั้นตอนนี้เป็นขั้นตอนที่มีปัญหามากกว่า ในขั้นตอนของการเตรียมข้อมูล เพราะข้อมูลส่วนใหญ่ที่มีในองค์กร ไม่ได้ถูกเตรียมมาเพื่องาน Data Mining โดยเฉพาะ ข้อมูลจะถูกนำมาจากแหล่งต่าง ๆ ถูกจัดเก็บไม่ดี ข้อมูลที่ถูกนำมาจากภายนอก

แล้วนำมาเพื่อให้เข้ากับข้อมูลภายในที่มีอยู่ ปัญหาหลักของ Data คือ คุณภาพและความถูกต้องของข้อมูล (Data Integrity)

ในขั้นตอนนี้ก่อนอื่นจะต้องทำการทบทวนโครงสร้างของข้อมูลใหม่ และวัดคุณภาพของมัน โดยวิธีทางสถิติ หรือสุ่มตัวอย่าง

เครื่องมือที่ใช้ในการทำการกลั่นกรองข้อมูลมีดังต่อไปนี้

1. ค่าตัวแปรเป็นแบบ Categorical การแบ่งความถี่ของค่าจะเป็นวิธีที่ทำให้เกิดความเข้าใจใน Data Content เครื่องมือทางด้านกราฟฟิคจะเป็นตัวช่วยให้เห็นและกำหนดค่าที่หายไปได้
2. ตัวแปรแบบ Quantitative ตัวแปรประเภทนี้มักมีการใช้การวัด ตัวอย่างเช่น ค่าสูงสุด ค่าต่ำสุด ค่าเฉลี่ย ค่ากลาง ค่ามัธยฐาน และค่าอื่น ๆ ทางสถิติ เมื่อนำค่าเหล่านี้มาเข้าสูตรคำนวณก็จะบอกถึงค่าที่ ไม่สมบูรณ์ หรือค่าที่มีปัญหา

เครื่องมือทางกราฟฟิคอื่น ๆ เช่น Scatter plots คือรูป 2 มิติซึ่งแสดงความสัมพันธ์ระหว่างตัวแปร 2 ตัวแปรขึ้นไป หรือมากกว่า จากกราฟตัวอย่างจะเห็นได้ว่ามีการเปรียบเทียบรายได้ กับอายุ จะเห็นได้ว่าจุดจะอยู่สูงขึ้นตามระดับของอายุ ทำให้เราพอที่จะทำนายได้ว่ารายได้ของ อาชีพนี้จะสูงขึ้นเมื่ออายุสูงขึ้น ส่วน Box plot ถูกใช้ให้เป็นประโยชน์สำหรับเปรียบเทียบศูนย์ กลาง (ค่าเฉลี่ย) หรือกระจาย (ค่าเบี่ยงเบน) ของตัวแปรตั้งแต่ 2 ตัวแปรขึ้นไป จากกราฟตัว อย่างตารางแสดง Data Element ของข้อมูล อธิบายถึงรายได้ของผู้ชายและผู้หญิง รูปสี่เหลี่ยมคือเรียกว่า Box และเส้นตั้ง 2 เส้นเรียกว่า Whisker จากความสูงของ Box พอจะสรุปได้ว่ารายได้ของผู้ชายสูงกว่าผู้หญิง

ระหว่างการทำขั้นตอนการกลั่นกรองข้อมูลจะมีปัญหาบ่อย ๆ ที่มักพบได้ ได้แก่

Noisy Data คือตัวแปรตัวหนึ่งหรือมากกว่ามีค่าซึ่งเกินกว่าค่าที่เราคาดไว้ ซึ่งอาจจะหมายถึงทั้งข้อดีและข้อเสียก็ได้ ในข้อดีก็คือ มันจะแสดงอย่างชัดเจนถึงโอกาสซึ่งเรากำลังมองหาอยู่ ในข้อเสีย คือมันอาจจะเป็นข้อมูลที่ไม่สมบูรณ์ สาเหตุที่เกิดขึ้นอาจจะมาจากความไม่รอบคอบของมนุษย์ ตัวอย่างเช่น Operator ใส่อายุให้คนเป็น 300 ปี หรือใส่ค่าของรายได้ เป็นติดลบ ค่าเหล่า นี้ควรจะถูกแก้ไข หรือเอาออกจากการวิเคราะห์ ควรมีขั้นตอนการตรวจสอบข้อมูลก่อนนำมาใช้

ค่าที่หายไป Missing Value คือค่าที่ไม่ได้แสดงในข้อมูลที่เราได้เลือกแล้ว หรือค่าที่ไม่สมบูรณ์ที่เราลบออกไป ระหว่างการทำ Noise Detection ค่าอาจจะหายไปเพราะเกิดจากความไม่รอบคอบของมนุษย์ เพราะว่าไม่มีข้อมูลนั้นระหว่างการทำ Input ข้อมูล การจัดการกับค่าที่หายไปนั้นสามารถจัดการได้ด้วยเทคนิคที่ต่างกัน

การสำรวจและตรวจสอบข้อมูล (Data Cleaning and exploration)

เมื่อทำการเก็บข้อมูลเรียบร้อยแล้ว ขั้นตอนต่อไปที่ควรกระทำก็ คือการตรวจสอบข้อมูล เหตุที่ต้องทำการตรวจสอบข้อมูลมี 2 ข้อ คือ ข้อแรก นักวิเคราะห์ควรมีความคุ้นเคยกับตัวข้อมูล ไม่ใช่รู้แต่ชื่อของ attribute และความหมายของมันเท่านั้น แต่ต้องรู้ถึงเนื้อหา (content) หรือความมุ่งหมายที่แท้จริงของข้อมูลด้วย ข้อสอง อาจมีความผิดพลาดของการเก็บสะสมข้อมูล เกิดขึ้นในขณะที่ทำการรวบรวมข้อมูลจากฐานข้อมูลหลายๆ แหล่งเข้ามาเป็นหนึ่งเดียวเพื่อใช้ในการวิเคราะห์ ซึ่งนักวิเคราะห์ ที่ดีจะต้องทำการตรวจสอบข้อมูลเหล่านี้ให้ถูกต้อง ตัวอย่างของความผิดพลาดที่เกิดขึ้น ได้แก่ ความผิดพลาดในการเก็บข้อมูล จาก attribute ที่ไม่ต้องการ ซึ่งเกิดจากความสับสนในการตั้งชื่อ attribute นั้น (mislabeling of field) เช่น เราต้องการเก็บค่าของระดับการศึกษาของผู้สมัครเข้าศึกษาต่อ ซึ่งในความเป็นจริงถูกเก็บไว้ใน attribute ที่ชื่อ “LEVEL_EDU” แต่ในฐานข้อมูลนั้นบังเอิญมี attribute อีกตัวหนึ่งชื่อ “EDUCATION” ซึ่งเก็บระดับการศึกษาที่ผู้สมัครต้องการเข้าศึกษา ซึ่งถ้าเราไม่ได้ตรวจสอบความสัมพันธ์และความมุ่งหมายที่แท้จริงของแต่ละ attribute แล้ว ก็อาจเกิดการสับสน โดยเก็บข้อมูลของ attribute “EDUCATION” ไปแทนก็ได้ ซึ่งเมื่อนำข้อมูลที่ได้ไปทำ Data Mining ผลลัพธ์ที่ได้ ก็จะผิดพลาดด้วย

การแปลงข้อมูล (Data Transformation)

ระหว่างขั้นตอนของการแปลงข้อมูล ข้อมูลที่ได้กลั่นกรองแล้วจะถูกแปลงให้เป็นรูปแบบของข้อมูลที่พร้อมจะถูกวิเคราะห์ รูปแบบของข้อมูลที่พร้อมจะถูกวิเคราะห์ คือรูปแบบของข้อมูลที่ไม่มีความขัดแย้ง ถูกจัดระเบียบมาอย่างเรียบร้อย กลั่นกรองมาจากแหล่งข้อมูลภายนอก และภายใน

ขั้นตอนนี้เป็นขั้นตอนที่สำคัญมากเนื่องจากความถูกต้อง และสมบูรณ์ของผลลัพธ์สุดท้ายซึ่งขึ้นอยู่กับว่า นักวิเคราะห์ ข้อมูลนั้นตัดสินใจกำหนดโครงสร้างและเสนอลักษณะของ Input อย่างไร ตัวอย่างเช่น หลักการรูปแบบของข้อมูลถูกกำหนด แล้ว ข้อมูลที่ถูกกลั่นกรองจะเหมาะสมกับรูปแบบเฉพาะสำหรับแต่ละ กรรมวิธีของ Data Mining ที่จะถูกใช้ การแปลงข้อมูลยัง รวมไปถึงการทำ Data Recording และ Data Format Conversion เช่นการแปลงวันที่ เป็นต้น

ทางสถิติการทำการแปลงข้อมูลยังมีเทคนิคของ Data Reduction จุดประสงค์เพื่อที่จะลดตัวแปรสำหรับการทำการ Process โดยการนำเอาตัวแปรตั้งแต่ 2 ตัวขึ้นไปมารวมกันแล้วทำการ Process ข้อดีก็คือลดจำนวนของตัวแปรลง และยัง สามารถจัดการได้ง่ายขึ้น

อีกเทคนิคเรียกว่า Discretization โดยการแปลงตัวแปรแบบ Quantitative ให้เป็นแบบ Categorical โดยการแบ่ง ค่าของตัวแปรที่จะเป็น Input ให้เป็นช่วง ๆ เช่นการแปลงเงินเดือน อายุ

อีกเทคนิคเรียกว่า One of N โดยการแปลงตัวแปรแบบ Categorical ให้เป็น Numeric ตัวอย่างเช่น ชนิดของรถ Ford, Lincoln, Nissan ให้เป็น 100, 010, 001 ปกติแบบนี้มักจะเป็น Input ของพวก Neural Network

การปรับแต่งข้อมูล (Data Engineering)

ขั้นตอนก่อนหน้านี้เป็นขั้นตอนของการสร้าง และการตรวจสอบความถูกต้องของข้อมูลที่จะนำมาใช้ แต่ในขั้นตอน นี้ที่เราต้องทำ คือการปรับแต่งฐานข้อมูล ซึ่งในขั้นตอนนี้จะมีปัญหาหลัก ๆ อยู่ 3 ข้อคือ หนึ่งฐานข้อมูลที่ได้อาจมี attributes จำนวนมากที่สามารถใช้ประโยชน์ได้แต่ถูกละเลย การเลือกกลุ่มของ attributes ที่จะใช้เป็นปัญหาที่สำคัญปัญหาหนึ่ง สอง ฐานข้อมูลที่ได้อาจมีจำนวนระเบียน (record) มากเกินไปกว่าที่จะสามารถทำการวิเคราะห์ให้เสร็จลงได้ในเวลาที่เหมาะสม ซึ่งในกรณีนี้เราต้องทำการสุ่มข้อมูลตัวอย่างขึ้นมาใช้แทน สาม ข้อมูลบางอย่างอาจใช้ให้เกิดประโยชน์ได้ โดยการนำเสนอ ในรูปแบบของการวิเคราะห์แบบเฉพาะเจาะจง การทำ Data engineering นั้นจะมีการทำซ้ำขึ้นมาหลาย ๆ ครั้ง เพื่อทดสอบ การใช้ attribute ที่แตกต่าง , ขนาดของกลุ่มตัวอย่างที่ต่างกัน เช่น เราจะทำนายอนาคตเมื่อเวลาผ่านไป 1, 2, 3, หรือ 4 เดือน เราอาจทำนายได้โดยใช้เพียง attribute เป็นตัวทำนายหรืออาจใช้ข้อมูลทุกอย่างที่เรามีเป็นตัวทำนายก็ได้ เป็นต้น

3. Visualization

เป็นการนำเสนอข้อมูลในรูปแบบกราฟฟิค การนำเสนอจะสามารถทำได้มากกว่า 2 มิติ ซึ่งจะสร้างความละเอียด ของการนำเสนอ และสร้างความเข้าใจให้มากขึ้น

4. Analysis

หลังจากเลือก algorithm ที่เหมาะสมกับลักษณะของปัญหาแล้ว เราก็จะนำ algorithm นั้นมาทำการวิเคราะห์ ข้อมูลในฐานข้อมูลที่เตรียมไว้ ซึ่งในบางครั้งขั้นตอนนี้จะถูกเรียกว่า “Data Mining” ในขณะที่จะเรียกกระบวนการทั้งหมดว่า “knowledge discovery in databases” ผลลัพธ์ที่ได้จากขั้นตอนนี้จะเป็นรูปแบบของความสัมพันธ์ของ ข้อมูลที่จะนำมาใช้ ในการพยากรณ์ (prediction) หรือวิเคราะห์ต่อไป

นำข้อมูลที่จัดเตรียมไว้มาทำ Data Mining ซึ่งมีการทำงานอยู่ 4 ชนิดด้วยกันคือ

- Data Segmentation เป็นกระบวนการแบ่ง Database ออกเป็นกลุ่มเพื่อให้ง่ายต่อกาวิเคราะห์ เช่นการแบ่งลูกค้าออก ตามอายุ เพศ รายได้ เป็นต้น
- Predictive Modeling เป็นการสร้างแบบจำลองพยากรณ์ แบ่งเป็น 2 ลักษณะ คือ

Classification เป็นการจัดกลุ่มให้กับแต่ละข้อมูลในฐานข้อมูล โดยมีการระบุค่า หรือลักษณะที่เป็นไปได้ของข้อมูล ภายในแต่ละกลุ่ม เช่น การจัดกลุ่มของผู้ป่วยตามผลของการใช้ยารักษา เพื่อระบุรูปแบบการรักษาให้กับผู้ป่วยใหม่ ที่เข้ารับการรักษา เป็นต้น

Value Prediction เป็นการพยากรณ์ค่าที่เป็นไปได้ หรือการกระจายของค่าที่เป็นไปได้ของตัวแปรใดๆ ในกลุ่มข้อมูล การทำนายค่าที่เป็นตัวเลข เช่น การทำนายภาษีที่จะเก็บได้ในปี เป็นต้น

- Link Analysis **(Associations)** เป็นการหาความสัมพันธ์ของข้อมูลภายในกลุ่มข้อมูล เพื่อใช้ลักษณะของข้อมูลหนึ่งๆ ในการบอกลักษณะที่จะเกิดขึ้นกับข้อมูลอีกตัวหนึ่ง ซึ่งอาจจะเป็นการหาความสัมพันธ์ของข้อมูลในกลุ่มเดียวกัน เช่น การระบุว่าในกลุ่มของลูกค้าที่ซื้อนมนั้น จะมีลูกค้า 64% ที่ซื้อขนมปังด้วย หรืออาจจะเป็นการหาความสัมพันธ์ของ ตัวแปรระหว่างกลุ่มข้อมูลก็ได้ เช่น ในทุกๆ ครั้งที่ดัชนีของตลาดหุ้นหนึ่งลดลง 5% ดัชนีของตลาดหุ้นอื่นจะเพิ่มขึ้น 13% ภายในช่วง 2-6 เดือนหลังจากนั้น เป็นต้น ซึ่งลักษณะของการหาความสัมพันธ์นั้นอาจแบ่งได้เป็น 3 กลุ่ม คือ การหาความสัมพันธ์ระหว่างข้อมูล (Association discovery) การหาความสัมพันธ์ในลักษณะที่เป็นลำดับของข้อมูล (Sequential Pattern discovery) และ การหาความสัมพันธ์ของข้อมูลกับช่วงเวลาใดๆ (Similar Time Sequence discovery)

- Deviation Detection เป็นเทคนิคที่ใช้ในการแสดงลักษณะของข้อมูลที่ผิดปกติ หรือผิดไปจากที่คาดไว้ โดยมีการแสดงผล อยู่ในลักษณะที่สามารถทำความเข้าใจและแปลความหมายได้ง่าย เช่น การใช้กราฟ เป็นต้น

5. Interprete

หลังจากที่การสร้าง Model แล้วจำเป็นต้องตรวจสอบผลลัพธ์และตีความหมาย ความถูกต้องที่ตรวจออกมาได้นั้น เป็นชุดตัวอย่างที่ส่งเข้าไป ดังนั้นผลลัพธ์ที่ได้อาจ มีความปรวนแปรหากมีการนำไปใช้กับข้อมูลอื่น ๆ

6. Presentation

เป็นการแสดงผลการวิเคราะห์โดยอาศัยเครื่องมือที่มีความสามารถและเข้าใจง่าย การแสดงผลอาจจะอยู่ในรูปแบบของ รายงาน ตาราง กราฟ แผนที่หลายมิติ เป็นต้น

2.3 Data Mining Tasks

1. Classification

ตัวอย่างนี้จะสร้างความเข้าใจใน Classification Study ซึ่งกรณีของตัวอย่างนี้พบได้ทั่วไปในวงการธุรกิจ นักวิเคราะห์ในองค์กรที่ทำธุรกิจเกี่ยวกับการสื่อสารแห่งหนึ่งต้องการเข้าใจว่าทำไม

ลูกค้าบางกลุ่มถึงยังคงซื่อสัตย์และมี Brand Loyalty สูงกับสินค้าขององค์กร แต่ในขณะเดียวกันลูกค้าอีกกลุ่มกลับไปหาคู่แข่งแทน ท้ายที่สุดนักวิเคราะห์จึงต้องการ จะทำนายลักษณะและนิสัยของลูกค้าที่องค์กรจะต้องเสียไปให้คู่แข่ง

เนื่องจากขณะนี้นักวิเคราะห์มีเป้าหมายในใจเรียบร้อยแล้ว ดังนั้นนักวิเคราะห์จึงสามารถสร้าง Model ที่ข้อมูลต่าง ๆ ได้มาจากข้อมูลในอดีตของลูกค้าที่มีความซื่อสัตย์ต่อองค์กรและกลุ่มลูกค้าที่ไม่มีความซื่อสัตย์ต่อองค์กรด้วย Model ที่สมบูรณ์ ถูกต้องจะสามารถทำให้องค์กรเข้าใจและทำนายลักษณะของธุรกิจที่จะเกิดขึ้นได้

จากตัวอย่างเหตุการณ์จะสามารถอธิบายขั้นตอนของการกำหนดการศึกษาได้ การศึกษาจะกำหนดขอบเขตของ กิจกรรมของ Data Mining ได้ นอกจากนี้การศึกษาจะสามารถกำหนดจุดประสงค์และข้อมูลที่ต้องการใช้ได้ทั้งหมด ด้วยการกำหนดปัญหาทางธุรกิจ นั่นก็เป็นสิ่งที่บอกให้นักวิเคราะห์ทราบได้เลยว่าขั้นตอนในการทำ Data Mining จะทำ อย่างไรและจุดประสงค์ของการทำคืออะไร

ในการศึกษาต้องการหัวข้อในการศึกษา หัวข้อในการศึกษาอาจหมายได้ถึง Data Element ของ Object ที่เรา ต้องการจะศึกษา เช่น เราต้องการจะศึกษาถึง Object “ลูกค้า” ซึ่งมี Data Element ที่เกี่ยวข้องคือ ชนิดของลูกค้า แนวโน้มการซื้อสินค้า ระยะเวลาที่เป็นลูกค้าขององค์กร และอื่น ๆ ซึ่ง Data Element จะเป็นตัวกำหนดลักษณะ และชนิดของลูกค้าการทำ Classification Studies นั้น เราสามารถกำหนดโครงร่างลักษณะเฉพาะหรืออุปนิสัยของลูกค้า ได้โดยดูได้จากตาราง

ชื่อคอลัมน์	ชนิดของข้อมูล	ค่าที่ได้	คำอธิบาย
เบอร์ลูกค้า	ตัวเลข	ค่าเฉพาะ	ตัวกำหนดเฉพาะ สำหรับลูกค้า
ระยะเวลา	ตัวเลข	จำนวนเต็ม	จำนวนที่ลูกค้าอยู่กับองค์กร
แนวโน้ม	ตัวอักษร	เพิ่มขึ้น , เหมือนเดิม , ลดลง	ตัวบ่งชี้แนวโน้มการใช้สินค้า 6 เดือนล่าสุด
สถานะ	ตัวอักษร	สูง , กลาง , ต่ำ , ไม่ทราบ	การสำรวจผลความพอใจของลูกค้า
ชนิดของลูกค้า	ตัวอักษร	ยังคงซื่อสัตย์ , ไม่ซื่อสัตย์	ลูกค้ายังคงอยู่กับองค์กรหรือเสียให้คู่แข่งไปแล้ว

ตาราง 2.1 แสดง Data Element ของข้อมูล

จากตัวอย่างข้างต้นเรากำหนดให้ ชนิดของลูกค้าเป็น Output หรือ Dependent Variable ซึ่งถูกใช้เป็น พื้นฐานในการศึกษาว่าอะไรคือสาเหตุที่ทำให้ลูกค้าซื่อสัตย์กับองค์กรและทำลูกค้าถึงจากองค์กรไป และเราจะใช้ Data Element ตัวอื่น ๆ มาช่วยในการอธิบายสิ่งที่เกิดขึ้น เรากำหนดให้ชนิดของลูกค้าเป็น Training Data ถ้าเราเปลี่ยน Data Element ตัวอื่นมาเป็น Output จุดประสงค์ของการศึกษาก็จะเปลี่ยนไปด้วย

มีเทคนิคของ Data Mining จำนวนมากที่ใช้สำหรับปัญหาแบบ classification และ regression และแต่ละเทคนิคก็มี algorithm มากมาย แต่ละ algorithm ก็ให้ผลลัพธ์ที่แตกต่างกันไป สิ่งที่แยกปัญหา classification ออกจากแบบ regression คือ ปัญหา classification จะให้ผลลัพธ์เป็นค่าที่แน่นอน เช่น “ใช่”, “ไม่ใช่” หรือ “สูง”, “กลาง” และ “ต่ำ” เป็นต้น ตัวอย่างเช่น แบบจำลองอาจทำนายว่า “นาย A จะตอบรับข้อเสนอของทางบริษัท” ในขณะที่ผลลัพธ์ที่จะได้จากปัญหาแบบ regression เป็นค่าเฉพาะที่แน่นอน แต่ค่านี้จะไม่จำกัดคือ อาจจเป็นค่าอะไรก็ได้ ตัวอย่างเช่นจากแบบจำลองที่ได้จากการทำ Data Mining แบบ regression แบบจำลองอาจทำนายว่า “นาย A จะได้รับผลกำไร 500 บาท เป็นต้น”

โดยทั่วไปแล้ว ปัญหาในแบบ regression จะสามารถเปลี่ยนเป็นปัญหาแบบ classification ได้โดยการแบ่งค่า ที่ต้องการทำนายให้เป็นกลุ่มของค่าที่ไม่ต่อเนื่องกัน (discrete categories) และปัญหาแบบ classification ก็สามารถเปลี่ยน เป็นแบบ regression ได้ โดยการทำนายค่าหรือความน่าจะเป็นสำหรับแต่ละกลุ่ม และกำหนดค่าของช่วงของค่า หรือความน่า จะเป็นที่ทำนายได้

เทคนิคของ Data Mining ที่ใช้ในการแก้ปัญหาแบบ classification และ regression

เทคนิคที่ใช้ในการทำ Data Mining แบบ classification และ regression ที่ใช้กันในผลิตภัณฑ์ด้าน Data Mining ในปัจจุบัน ได้แก่

- Decision tree เป็นเทคนิคที่ให้ผลลัพธ์ในลักษณะของโครงสร้างต้นไม้

  โดยปกติมักประกอบด้วยกฎในรูปแบบ “ถ้า เงื่อนไข แล้ว ผลลัพธ์” เช่น

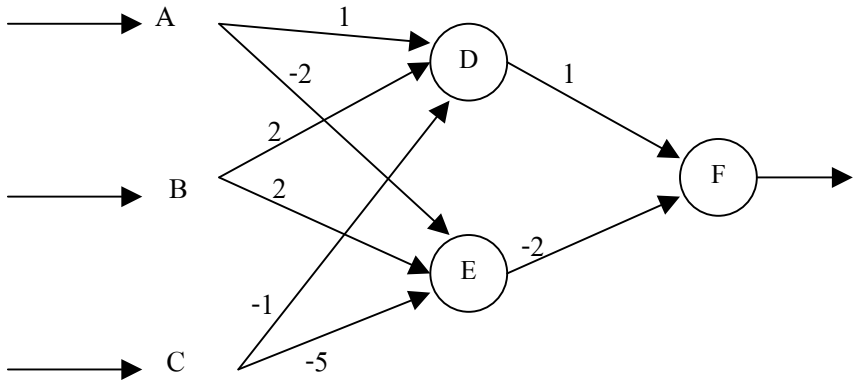
  “If Income = High and Married = No THEN Risk = Poor”

  “If Income = High and Married = Yes THEN Risk = Good”

Decision tree เป็นเทคนิคที่ค่อนข้างแพร่หลาย เนื่องจากผู้ใช้สามารถทำความเข้าใจผลลัพธ์ได้ง่าย เทคนิค Decision tree จะจำกัดข้อมูลที่เป็นตัวแปรตาม (dependent variable) 1 ตัวต่อ 1 แบบจำลอง ถ้าต้องการทำนายตัวแปรตามหลายๆ ตัว จะต้องสร้างแบบจำลองสำหรับตัวแปรตามแต่ละตัว algorithm ของเทคนิคแบบ Decision tree ส่วนใหญ่ไม่รองรับข้อมูลแบบต่อเนื่อง (continuous data) จะต้องมีการแบ่งให้เป็นข้อมูลแบบไม่ต่อเนื่อง (discrete data) เสียก่อน algorithm ที่เหล่านั้นได้ก่อน Chi-squared Automatic Interaction Detection (CHAID) , Classification and Regression

Trees (CART) , C4.5 และ C5.0 algorithm เหล่านี้ส่วนมากมักเหมาะกับปัญหาแบบ classification Algorithm บางตัวปรับให้ใช้ได้กับปัญหาแบบ regression เช่น Classification and Regression Trees (CART) ซึ่งรองรับทั้งปัญหาในแบบ Classification และ regression นอกจากนี้ยังรองรับข้อมูลในแบบที่ต่อเนื่องด้วย

- Neural networks มีพื้นฐานมาจากแบบจำลองการทำงานของสมองมนุษย์ และก็สามารถใช้ได้ดีกับปัญหา classification, regression และ clustering เทคนิคนี้มักถูกเรียกว่า “black box” เนื่องจากการทำงานของมันมีความซับซ้อนมากกว่าเทคนิคอื่น ๆ ค่อนข้างมาก ผลลัพธ์ที่ได้ก็ยากต่อการทำความเข้าใจ



รูป 2.2 แสดงผลลัพธ์ของการใช้เทคนิคแบบ Neural Networks

เช่น ในรูปแสดงผลลัพธ์ของการใช้เทคนิคแบบ Neural networks ในการวิเคราะห์ปัญหาความเสี่ยงของการให้กู้เงิน ซึ่งประกอบด้วยจุด 6 จุด A-F โดยที่ A, B, C เป็นจุที่เป็นข้อมูลเข้า ซึ่งแทนตัวแปรอิสระ หนี้สิน (debt), รายได้ (income) และสถานภาพสมรส (Married) ในขณะที่จุด F เป็นผลลัพธ์ของการวิเคราะห์ แทนตัวแปรตามคือ ความเสี่ยง (risk) และตัวเลขที่กำกับอยู่ตามเส้นลูกศรคือ ค่าถ่วงน้ำหนัก (weight) เป็นต้น

ถึงแม้ว่าเทคนิคนี้จะทำงานได้ดีกับปัญหา classification, regression และ clustering ก็ตาม แต่มันเป็นเทคนิคที่ค่อนข้างซับซ้อนกว่าเทคนิคอื่น ความซับซ้อนและการไม่สามารถอธิบายได้ของผลลัพธ์ มักทำให้ผู้ใช้หลีกเลี่ยงเทคนิคนี้ อย่างไรก็ตาม เทคนิคนี้ก็มีข้อดีที่สำคัญที่ไม่มีในเทคนิคอื่น ๆ ก็คือ เทคนิคนี้ไม่มีข้อจำกัดเกี่ยวกับชนิดของความสัมพันธ์ เช่น เทคนิคแบบ neural networks สามารถสร้างแบบจำลองความสัมพันธ์ระหว่างตัวแปรตามกับสัดส่วนของตัวแปรอิสระ 2 ตัวได้ ซึ่งทำได้ยาก ถ้าใช้เทคนิคแบบ Decision tree หรือ Naïve-Bayes นอกจากนี้ เทคนิคแบบ Neural

Networks ยังไม่มีปัญหากับความสัมพันธ์ที่เป็นแบบตรีโกณมิติ (trigonometric) หรือ logarithmic ด้วย ในการใช้งานจริงนั้น เทคนิคแบบ Decision tree หรือ Naïve-Bayes อาจให้ผลลัพธ์ที่ถูกต้องเพียงพอกับความต้องการ แต่ถ้าต้องการความแม่นยำมาก ๆ แล้ว เทคนิคแบบ Neural networks อาจเป็นหนทางที่ดีที่สุด ทางเดียวที่จะรู้ว่าควรใช้เทคนิคแบบ Neural networks หรือไม่ ก็คือ การเปรียบเทียบความเที่ยงตรงของแบบจำลองกับเทคนิคอื่น (Decision tree หรือ Naïve-Bayes) ถ้าไม่ได้ดีกว่ากันอย่างเห็นได้ชัด ก็ควรเลือกเทคนิคอื่น แต่ถ้าผลลัพธ์ที่ได้จากแบบจำลองของเทคนิค Neural networks มีความเที่ยงตรงกว่าอย่างเห็นได้ชัด นั่นอาจหมายถึง เราต้องทำการปรับปรุงแบบจำลองของเทคนิค Decision tree หรือ บางทีการใช้เทคนิคแบบ Neural networks อาจเหมาะสมสำหรับปัญหานี้มากที่สุดก็ได้

- Naïve-Bayes เป็นเทคนิคที่ถูกตั้งชื่อตาม Thomas Bayes (1702-1761) เทคนิคแบบ Naïve-

Bayes ใช้ทฤษฎี Bayes Theorem ในการคำนวณความน่าจะเป็นซึ่งถูกใช้ในการทำนายผล เมื่อทำการวิเคราะห์กรณีใหม่ การทำนายผลทำได้โดยการรวมผลของตัวแปรอิสระ (independent variable) ที่มีต่อตัวแปรตาม (dependent variable) Naïve-Bayes เป็นเทคนิคในการแก้ปัญหาแบบ classification ที่ทั้งสามารถคาดการณ์ผลลัพธ์ได้และสามารถอธิบายได้ด้วย มันจะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอิสระแต่ละตัวกับตัวแปรตามเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ ในทางทฤษฎีแล้วการทำนายผลของ Naïve-Bayes จะถูกต้องถ้าตัวแปรอิสระทั้งหมดเป็นอิสระต่อกัน ไม่ขึ้นกับตัวแปรอิสระตัวใดตัวหนึ่ง ซึ่งในความเป็นจริงแล้วมีไม่มากนักที่ตัวแปรอิสระทั้งหมดเป็นอิสระต่อกัน ตัวอย่างเช่น ข้อมูลเกี่ยวกับประวัติบุคคล ซึ่งมักประกอบด้วยรายละเอียดย่อยมากมาย อาทิ น้ำหนัก , การศึกษา , รายได้ เป็นต้น จะเห็นว่ารายละเอียดเหล่านี้มักขึ้นอยู่กับอายุ ในกรณีนี้การใช้ Naïve-Bayes จะต้องคำนึงถึงผลของอายุให้มาก ๆ นอกจากนี้ เทคนิคแบบ Naïve-Bayes ยังไม่รองรับข้อมูลที่เป็นข้อมูลต่อเนื่อง (continuous data) ด้วย ดังนั้น ตัวแปรอิสระหรือตัวแปรตามที่มีค่าเป็นค่าต่อเนื่องจะต้องถูกแบ่งเป็นช่วงเช่น ถ้ามีตัวแปรอิสระที่เป็นค่าของอายุก็อาจแปลงค่าเหล่านั้นให้เป็นช่วงแคบ ๆ อาทิ “ต่ำกว่า 20 ปี” , “20-40 ปี” , “40 ปีขึ้นไป” เป็นต้น ซึ่งการแบ่งช่วงนั้น ถ้าแบ่งไม่เหมาะสม ก็จะมีผลต่อคุณภาพของแบบจำลองที่สร้างขึ้น แต่ถ้าไม่คำนึงถึงข้อจำกัดนี้แล้ว เทคนิคแบบ Naïve-Bayes สามารถให้ผลลัพธ์ที่ดีและรวดเร็วได้ ความง่ายและความเร็วทำให้เทคนิคนี้เป็นเครื่องมือที่ดีในการสร้างแบบจำลองและหารูปแบบความสัมพันธ์ที่ไม่ซับซ้อน

- K-nearest neighbor (K-NN) เป็นเทคนิคที่เหมาะกับปัญหาแบบ classification เทคนิคนี้แตกต่างจากเทคนิคอื่นตรงที่มันไม่ได้ใช้ข้อมูลฝึกหัด (training data) ในการสร้างแบบจำลอง แต่จะ

ใช้ข้อมูลนั้นมาเป็นตัวแบบจำลองเลย ในการใช้งาน K-NN algorithm นั้นเราต้องระบุค่าตัวเลขจำนวนเต็มบวกให้กับ k ด้วย ซึ่งค่านี้จะเป็นตัวบอกจำนวนของกรณี (case) ที่จะต้องค้นหาในการทำนายกรณีใหม่ algorithm แบบ K-NN ได้แก่ 1-NN, 2-NN, 3-NN, ……. K-NN โดยที่ k แทนเลขจำนวนเต็มบวก เช่น 4-NN หมายถึง algorithm นี้จะค้นหา 4 กรณีที่มีลักษณะใกล้เคียงกับกรณีใหม่ (4 nearest cases) ในการทำนายกรณีใหม่

2. Estimation / Prediction

ลักษณะของ Classification นั้นคำนึงถึงผลกำหนดที่ออกมาชัดเจนว่าคุณสมบัติดังกล่าวจะอยู่ในชั้นใด แต่ **Estimation** เป็นการประเมินที่ไม่สามารถกำหนดค่าหรือคุณสมบัติดังกล่าวให้ชัดเจน เป็นการจัดการกับค่าที่มีผลในการวัดที่ต่อเนื่อง ตัวอย่างเช่น

- การประเมินรายได้ของครอบครัว
- การประเมินความสูงของบุคคลในครอบครัว
- การประเมินจำนวนของเด็กๆ ในครอบครัว

Prediction เหมือนกับ Classification และ Estimation ยกเว้นว่า Record ที่ถูกแยกจัดลำดับนั้นเกิดขึ้นตามการทำนาย พฤติกรรมในอนาคตหรือการทำนายค่าที่จะเกิดขึ้นในอนาคต ข้อมูลในอดีตจะถูกสร้างเป็น Model ขึ้นมาเพื่อทำนายหรืออธิบาย สิ่งที่จะเกิดขึ้นในอนาคต ตัวอย่างเช่น

- การทำนายว่าลูกค้ากลุ่มใดที่องค์กรจะสูญเสียไปภายใน 6 เดือนหน้า
- การทำนายว่ายอดซื้อของลูกค้าจะเป็นเท่าใดถ้าบริษัทลดราคาสินค้า 10 %

3. Segmentation / Clustering

Clustering คือวิธีของการรวมกลุ่มของข้อมูลที่มีลักษณะเหมือนกัน รูปแบบและแนวโน้มที่เหมือนกัน การศึกษาของ Clustering ไม่มี Output หรือ Independent Variable เหมือน Classification Studies และไม่มีการจัดเป็นลักษณะโครงร่างของ Object ใด ๆ ดังนั้นการศึกษาแบบนี้จึงถูกเรียกว่า Unsupervised Learning หรือ Segmentation การทำ Clustering เองสามารถทำบนพื้นฐานของข้อมูลในอดีตได้เหมือนกัน แต่ผลลัพธ์ที่ได้มาไม่ได้ออกจาก Training Data

ตัวอย่างของ Clustering เช่น องค์กรต้องการทราบความเหมือนที่มีในกลุ่มลูกค้าของตนเองเพื่อที่ว่าองค์กรจะสามารถเข้าใจลักษณะเฉพาะของกลุ่มลูกค้าเป้าหมายขององค์กร และสร้างกลุ่มของลูกค้าเพื่อองค์กรจะสามารถขายสินค้าได้ในอนาคต องค์กรจะทำการแยกกลุ่มของลูกค้าออกเป็นกลุ่ม ๆ

เทคนิคของ Clustering พยายามมองหาความเหมือนและความแตกต่างภายในกลุ่มของข้อมูลและแบ่งกลุ่มต่าง ๆ ออกเป็นส่วน ๆ

เทคนิคในการทำ Data Mining เพื่อแก้ปัญหาแบบ clustering

- Demographic Clustering แนวคิดพื้นฐานของ Demographic Clustering คือการสร้าง segment โดยการเปรียบเทียบข้อมูล แต่ละตัวกับทุก ๆ segment ที่สร้างขึ้นในขณะที่กำลังทำ Data Mining โดยการสร้างความแตกต่างระหว่างคะแนน ให้มากที่สุด algorithm จะใส่ข้อมูลลงในแต่ละ segment ซึ่ง segment ใหม่สามารถถูกสร้างขึ้นได้ตลอดเวลาที่ทำ Data Mining ข้อดีของเทคนิคนี้ คือ มันสามารถกำหนดจำนวนของ segment ที่ต้องสร้างขึ้นได้โดยอัตโนมัติและ ผลลัพธ์ของชุดข้อมูลขนาดใหญ่ที่ถูกแบ่งอย่างชัดเจน Demographic Clustering เหมาะกับข้อมูลที่มีลักษณะเป็นกลุ่ม โดยเฉพาะจำนวนของกลุ่มน้อย ๆ
- Neural Clustering เทคนิคนี้นำ Kohonen feature map neural network มาใช้ Kohonen feature map ใช้กระบวนการ ที่เรียกว่า self-organization ในการตั้งค่าหน่วยของผลลัพธ์เข้าสู่ topological map Feature map neural network ประกอบด้วยชั้นของหน่วยประมวลผล 2 ชั้น โดยชั้นของข้อมูลเข้า (input layer) จะเชื่อมต่อกับชั้นของผลลัพธ์ (output layer) อย่างสมบูรณ์ เมื่อรูปแบบของข้อมูลเข้าถูกแสดงสู่ feature map หน่วยต่าง ๆ ในชั้นของผลลัพธ์ จะแข่งขันกันเพื่อสิทธิ์ที่จะได้เป็นผู้ชนะ หน่วยผลลัพธ์ที่ชนะคือ หน่วยที่น้ำหนักการเชื่อมต่อใกล้เคียงกับรูปแบบข้อมูล เข้ามากที่สุด (ในความหมายของ Euclidean distance) Kohonen feature map สร้าง topological map โดยปรับแต่งไม่เพียงแต่ น้ำหนักของผู้ชนะเท่านั้น ยังปรับแต่งน้ำหนักของหน่วยผลลัพธ์ที่อยู่ประชิดกับผู้ชนะด้วย

4. Description / Visualization

Description จุดประสงค์ของการทำ **Data Mining** การหาคำอธิบายถึงสิ่งที่จะเกิดขึ้นโดยอาศัยข้อมูลจากฐานข้อมูล ตัวอย่างเช่น ผู้หญิงจะสนับสนุนพรรคการเมืองพรรคหนึ่งมากกว่าผู้ชาย

Visualization เป็นการนำเสนอข้อมูลในรูปแบบกราฟฟิค การนำเสนอจะสามารถทำได้มากกว่า 2 มิติ ซึ่งจะสร้างความละเอียดของการนำเสนอและสร้างความเข้าใจให้มากขึ้น ตัวอย่างเช่น องค์กรต้องการที่จะหาสถานที่ในการตั้งสาขาขององค์กรในเขตพื้นที่ภาคเหนือของประเทศ ดังนั้นองค์กรจึงใช้รูปแผนที่ที่มีการ Plot ที่ตั้งขององค์กรคู่แข่งที่มีสาขาตั้งอยู่ในเขตนั้น เพื่อพิจารณาสถานที่ตั้งที่เหมาะสมที่สุด

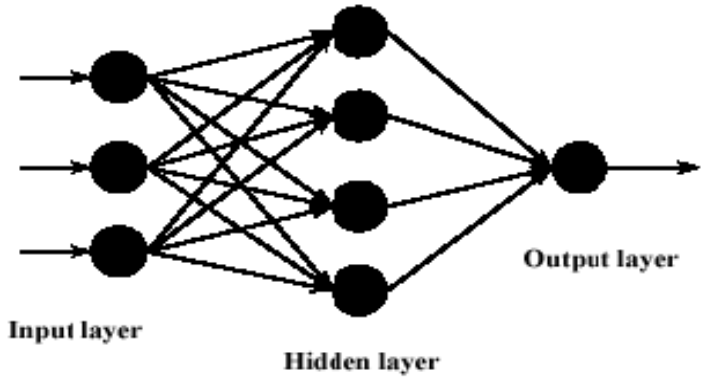
Data Visualization จะใช้มากกับ Data Mining Tools สิ่งที่สำคัญของ Visualization ก็คือ ไม่สามารถเน้นการวิเคราะห์ข้อมูลที่มีประสิทธิภาพ ในขณะที่แบบแผนทางสถิติและ Confirmatory Analysis เป็นการสร้างการวิเคราะห์ข้อมูลที่แท้จริง

2.4 Data Mining Tools and Technologies

1. Neural Network

เป็นการเลียนแบบการทำงานของระบบประสาทเทียมซึ่งเลียนแบบการทำงานของระบบประสาทในสมองของมนุษย์ การทำงานของ Neural Network แต่ละ Process จะรับ Input เข้าไปคำนวณ และสร้าง Output ออกมาในลักษณะที่ไม่ใช่เป็นการทำงานแบบเชิงเส้นตรง เพราะว่า Input แต่ละตัวจะถูกให้ลำดับความสำคัญของค่าไม่เท่ากัน ค่าของ Output ที่ได้จากการเชื่อมโยงกันนี้จะถูกนำมาเปรียบเทียบกับ Output ที่ได้ตั้งเอาไว้ ถ้าค่าที่ออกมาเกิดความคลาดเคลื่อน ก็จะนำไปสู่การปรับค่าหรือน้ำหนักของค่าที่ใส่ไว้ให้แต่ละ Input

Neural Network เป็นการสร้างแบบจำลองที่เลียนแบบการทำงานของสมองมนุษย์ มีโครงสร้างเป็นกลุ่มของ Node ที่เชื่อมโยงถึงกันในแต่ละ Layer คือ Input Layer, Hidden Layer และ Output Layer



รูป 2.3 แสดงตัวอย่างของ Neural Network

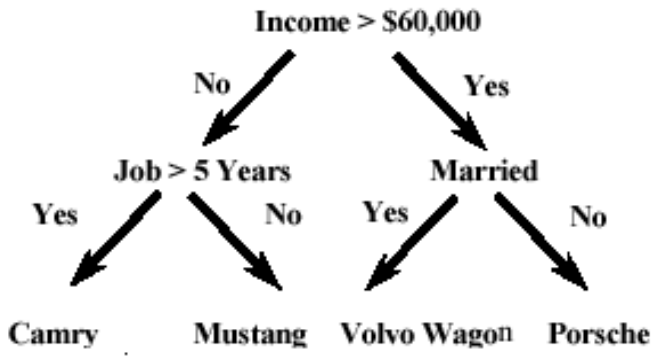
ข้อจำกัดของ Neural Network

- Neural Network รับข้อมูลได้เฉพาะ ข้อมูลตัวเลขที่อยู่ในช่วง 0 – 1 เท่านั้น กรณีที่ข้อมูลนำเข้ามีค่า มากกว่า นี้ต้องทำการปรับลดค่าลง หรือในกรณีที่เป็นข้อมูลอื่นที่ไม่ใช่ตัวเลขต้อง ทำการแปลงค่าก่อน
- การสร้างแบบจำลองด้วย Neural Network นั้นจะไม่สามารถอธิบายได้ว่าผลลัพธ์ที่ได้นั้น มาจากไหน
- เนื่องจากการที่ไม่สามารถอธิบายผลลัพธ์ที่ได้มาได้ ดังนั้นการสร้างแบบจำลองด้วย Neural Network จะไม่สามารถรับรองได้เลยว่าเป็นแบบจำลองที่ดีหรือไม่จนกว่าจะได้ทำการทดสอบกับข้อมูลทดสอบก่อนจน แน่ใจก่อน

2. Decision Trees

เป็นการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ในรูปของ Decision Tree ซึ่ง Decision Tree นั้นมีการทำงานแบบ Supervised Learning คือ สามารถสร้างแบบจำลองการจัดหมวดหมู่ได้จากกลุ่มตัวอย่างของข้อมูลที่ได้กำหนดได้ก่อนล่วงหน้า ที่เรียกว่า Training Set ได้อัตโนมัติ และสามารถพยากรณ์กลุ่มของรายการที่ยังไม่เคยนำมาจัดหมวดหมู่ได้ด้วย

รูปแบบของ Tree จะประกอบด้วย Node แรกสุดที่เรียกว่า Root Node จาก Root Node ก็จะแตกออกเป็น Node ลูก และที่ Node ลูกก็จะมีลูกของตัวเองซึ่ง Node ที่ระดับสุดท้ายจะเรียกว่า Leaf Node



รูป 2.4 แสดงตัวอย่างของ Decision Tree

จะเห็นว่า จาก Root Node จนถึง Leaf Node จะมีเพียงเส้นทางเดียวเท่านั้น ซึ่งเส้นทางนี้จะอธิบาย ถึงกฎที่ใช้สำหรับการจัดหมวดหมู่ของแต่ละกลุ่ม ซึ่งในแต่ละ Leaf Node นั้นอาจเป็นกลุ่มเดียวกัน ซึ่งเกิดจากเหตุผล ที่แตกต่างกันได้

วิธีการที่ใช้สร้าง Decision Tree การนำข้อมูลมาสร้าง Tree มีขั้นตอนดังนี้

- หา Attribute ที่สำคัญที่สุดมาแบ่งข้อมูลโดย Attribute นี้จะถูกนำมาสร้างเป็น Root Node โดยจะมี Target Attribute เป็นผลลัพธ์ซึ่งเป็น Leaf Node ถูกกำหนดไว้ก่อน
- นำค่าที่เป็นไปได้ใน Attribute ที่ถูกเลือกมาแตกออกเป็นกลุ่มของตัวเอง
- แบ่งข้อมูลทั้งหมดตามกลุ่มที่แตกออกจาก Root Node
- วนกลับไปทำที่ขั้นตอนแรก คือ หา Attribute ที่สำคัญที่สุดจากข้อมูลที่เข้ามาเพื่อหาตัวแบ่งต่อไป

ข้อจำกัดของ Decision Tree

- การแบ่งกลุ่มแบบ Decision Tree กรณีเป็นข้อมูลที่มีค่าต่อเนื่อง เช่น ข้อมูลรายได้ ข้อมูลราคา ต้องทำการแปลงให้อยู่ในช่วงหรือตัดเป็นกลุ่มก่อน
- เมื่อ Algorithm เลือกว่าจะใช้ค่าไหนเป็นตัวแบ่งกลุ่มแล้วก็จะไม่สนใจค่าอื่นที่อาจมีความสำคัญเช่นเดียวกัน
- การจัดการกับข้อมูลที่ไม่ทราบค่า อาจมีผลกระทบกับผลลัพธ์ของ Decision Tree
- Tree ที่มีระดับชั้นมากเกินไป จะทำให้ข้อมูลที่ผ่าน Node แตกออกเป็นชิ้นเล็กชิ้นน้อย ซึ่งข้อมูลเหล่านั้น จะไม่มีประโยชน์ในการนำมาใช้ทำการวิเคราะห์
- ปัญหาเรื่อง Overfitting / Overtaining เกิดจากการที่แบบจำลองได้เรียนรู้เข้าไปถึงรายละเอียดของข้อมูล มากเกินไปจะทำให้เกิด Node ที่เป็นส่วนเฉพาะเจาะจงกับกลุ่มข้อมูลที่ใช้ในการเรียนรู้ ซึ่งจะต้องหาวิธี การในการตัดกิ่งนี้ออกไป

3. Memory Based Reasoning (MBR)

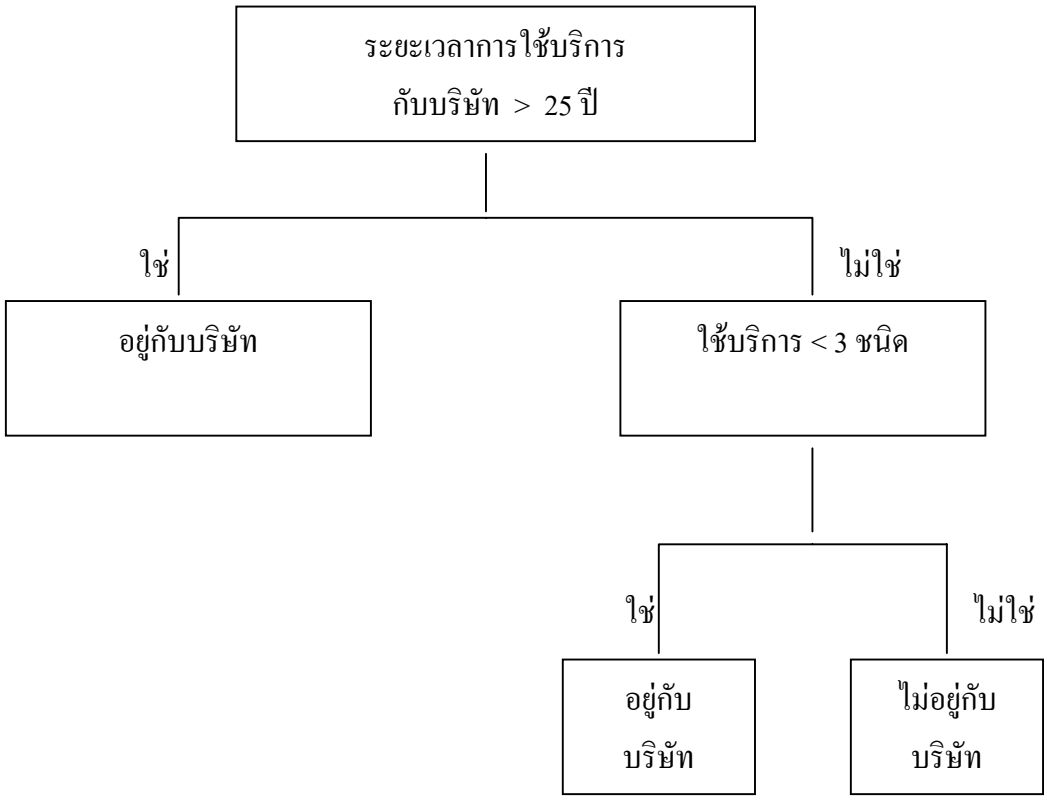
เปรียบเหมือนกับประสบการณ์การเรียนรู้ของมนุษย์ ซึ่งอาศัยการสังเกตการณ์ที่เกิดขึ้นแล้วสร้างรูปแบบของสิ่งนั้น ขึ้นมา ใน Data Mining เราใช้ MBR เพื่อทำการวิเคราะห์ฐานข้อมูลที่มีอยู่และกำหนดลักษณะพิเศษของข้อมูลที่อยู่ในนั้น แน่นอนข้อมูลจะต้องมีลักษณะสมบูรณ์ , การทำการสังเกตอย่างสมบูรณ์จะช่วยสร้างการทำนายอย่างละเอียดแม่นยำยิ่งขึ้น Model จะถูกบอกคำตอบที่ถูกต้องจากกรณีศึกษาที่ได้แก้ปัญหาไว้ก่อนหน้าแล้ว การทำงานแบบนี้วิธีนี้ถูกเรียกว่า “Supervised Learning”

ตัวอย่างของนักวิเคราะห์ต้องการเข้าใจว่าทำไมลูกค้าบางกลุ่มซึ่งซื่อสัตย์แต่อีกกลุ่มบริษัทกลับเสียลูกค้าไป และนักวิเคราะห์ จะทำนายว่าลูกค้าคนใดที่บริษัทกำลังจะเสียไปให้คู่แข่ง นักวิเคราะห์สามารถสร้าง Model จากข้อมูลในอดีต Model ที่ดีก็จะทำให้เรารู้ว่าลูกค้าคนใดจะอยู่กับบริษัทและลูกค้าคนใดจะเสียไป ตัวอย่างนี้เป็นขั้นตอนของการกำหนด “การศึกษา Study” การศึกษาจะเป็นตัวกำหนดขอบเขตของกิจกรรม การศึกษาจะกำหนดจุดประสงค์ให้ทั้งหมดและข้อมูลที่จะถูกใช้อาจจะไม่ต้อง กำหนด จุดประสงค์ไว้ล่วงหน้า

จุดประสงค์ของการศึกษา คือ ต้องการเข้าใจว่าอะไรทำให้ลูกค้าอยู่กับบริษัทและจากบริษัทไป จุดประสงค์นี้แตกต่าง จากการถามคำถามเฉพาะ เพราะเราไม่ได้กำหนดความสัมพันธ์เอาไว้ เทคนิคในการทำ MBR จะมีจุดประสงค์หลัก คือการทำการคาดเดาอย่างมีหลักการเกี่ยวกับตัวแปรที่สนใจ โดยมักจะใช้เทคนิคของ Neural Network และ Decision Tree

อธิบายรูปแบบของการทำ MBR จากตัวอย่างในรูป คือบริษัทประกันภัยซึ่งมีความสนใจที่ทราบสาเหตุของ การลดลงของจำนวนลูกค้าว่า โดย MBR กำหนด 2 ตัวแปรที่สนใจคือระยะเวลาที่ลูกค้าอยู่กับบริษัท(ถือกรมธรรม์) และจำนวนของบริการของบริษัทที่ลูกค้าใช้บริการอยู่ จะเห็นได้

ชัดว่าลูกค้าที่อยู่กับบริษัทน้อยกว่า 2 ปีครึ่ง และใช้บริการน้อยกว่า 3 บริการมักจะหนีไปใช้บริการของบริษัทอื่นๆ



รูป 2.5 แสดงแบบการตัดสินใจของบริษัทประกันภัย (Cabena et al., 1997)

การทำงานของ MBR ตั้งอยู่บนรากฐาน 2 ประการคือ การทำ Classification และการทำ Value Prediction ตัวอย่างของ Classification เช่นบริษัทที่ต้องการทำ Sales Promotion ซึ่งจะใช้ Mailing List จากฐานข้อมูลการซื้อของลูกค้า Mailing List ซึ่งมีการตอบรับกลับมาจากการส่ง Mail ไปครั้งก่อนหน้าจะมีการกำหนดเป็น Classification (Classification ถูกใช้กำหนด ชั้นของแต่ละ Record ในฐานข้อมูล จากตัวอย่างคือ การอยู่กับบริษัท และการไม่อยู่กับบริษัท) หรือ Profile Classification ดังกล่าวจะถูกเก็บรวบรวมและพัฒนาเพื่อที่จะบอกลักษณะของผู้ที่ตอบรับ Mail เพื่อกำหนดเป็นตัวทำนาย ที่จะตอบรับ และจะนำเอารายชื่อดังกล่าวไปส่ง

ส่วน Value Prediction จุดประสงค์เพื่อกำหนดความต่อเนื่องของมูลค่าซึ่งมีความเกี่ยวข้องกับ Record ในฐานข้อมูล ตัวอย่างเช่นการศึกษา Lifetime Customer การทำ Mining ก็จะศึกษาถึงข้อ

มูลที่ผ่านมาในอดีตของลูกค้าร่วมไปถึงสถานะ ทางการเงินของลูกค้านั้นด้วยนอกจากนี้ยังมีตัวแปรอื่นๆ อีก เช่นจำนวนครอบครัว รายได้ ประวัติการใช้รถ MBR เป็นรูปแบบที่มักจะถูกใช้อย่างกว้างๆ ในอุตสาหกรรมทั่วๆไป ทางธุรกิจนั้นมักจะใช้กับ Customer Retention Management, Credit Approval, Cross Selling และ Target Marketing

4. Cluster Detection

จุดประสงค์ของ Cluster Detection คือการแบ่งฐานข้อมูลออกเป็นส่วนๆ หรือเราเรียกว่า Segment คือกลุ่มของ Record ที่มีความเหมือนและลักษณะที่คล้ายกัน หรือเรียกว่า “Homogeneity” ส่วน Record ที่อยู่ใน Segment อื่นๆ ก็จะมีความแตกต่างกัน หรือเรียกกลุ่มที่อยู่นอก Segment ว่า “Herterogeneity” Cluster Detection ถูกใช้เพื่อค้นหา Sub Group ที่เหมือนๆ กันในฐานข้อมูลเพื่อที่จะเพิ่มความถูกต้องในการวิเคราะห์ และสามารถมุ่งไปยังกลุ่มเป้าหมายได้ถูกต้อง

เรานำกราฟมาอธิบายกลุ่มของประชากรโดยเปรียบเทียบรายได้และอายุ ในรูปจะเห็นว่ากลุ่มหนึ่งเป็นกลุ่มที่มีอายุ และมีรายได้สูง ส่วนอีกกลุ่มหนึ่งอายุน้อยรายได้ปานกลาง มีการศึกษา มีการแบ่งข้อมูลออกเป็น 2 Segment

เทคนิค Cluster Detection เป็นวิธีของการรวมกลุ่มของแถวของข้อมูลซึ่งมีส่วนร่วมที่คล้ายกันแนวโน้มและรูปแบบ Clustering Studies ไม่มี Dependent Variable ดังนั้นจึงไม่สามารถศึกษาได้ลงไปอย่างเฉพาะเจาะจงไม่สามรถทำให้เกิดผลที่แน่นอน เราจึงเรียกการศึกษาแบบนี้ว่าเป็น “Unsupervised Learning” ตัวอย่างเช่น เราต้องการทราบว่าอะไรที่เหมือนกันในกลุ่มฐานลูกค้าของบริษัท เทคนิค Clustering ก็จะทำการจำแนกแยกกลุ่มให้

Cluster Detection แตกต่างจาก Data Mining เทคนิคอื่นๆ คือจุดประสงค์ค่อนข้างคลุมเครือเมื่อเทียบกับเทคนิคของ Data Mining ตัวอื่นๆ

5. Link Analysis

Link Analysis มุ่งเน้นทำงานบน Record คือความสัมพันธ์ หรือความเกี่ยวโยงกันระหว่าง Record หรือกลุ่มของ Record ความสัมพันธ์ดังกล่าวเรียกว่า “Association” เทคนิคนี้มุ่งมองไปที่รูปแบบการซื้อหรือเหตุการณ์ที่เกิดขึ้นเป็นลำดับ โดยมีเทคนิคที่ใช้บน Link Analysis อยู่ 3 อย่าง

- Association Discovery ใช้วิเคราะห์การซื้อสินค้าภายในรายการเดียวกัน ศึกษาถึงความสัมพันธ์อย่างใกล้ชิดที่ถูก ปิดซ่อนอยู่ของสินค้า ซึ่งสินค้าเหล่านั้นมักมีแนวโน้มที่จะถูกซื้อควบคู่กันไป การวิเคราะห์แบบนี้เรียกว่า “Market Basket Analysis” คือรายการทั้งหมดที่ลูกค้าซื้อต่อครั้งที่ซุปเปอร์มาร์เก็ต สามารถใช้ Input Device โดยใช้ Bar Code Scanner มีหลายงานด้วยกัน เช่น ซุปเปอร์มาร์เก็ต การเตรียม Inventory การวางแผนการเรียง Shelf การทำ Mailing List สำหรับ Direct Mail และการวางแผนเพื่อจัด Promotion สนับสนุน

การขาย ตัวอย่างของ Association เช่น อาจพบว่า 75 % ของผู้ซื้อน้ำอัดลมจะซื้อข้าวโพดคั่วด้วย

- Sequential Pattern Discovery ถูกใช้ระบุความเกี่ยวเนื่องกันของการซื้อสินค้าของลูกค้ามันมีจุดมุ่งหมายที่จะเข้าใจ พฤติกรรมการซื้อสินค้าของลูกค้าในลักษณะ Long Term เช่นผู้ขายอาจพบว่าลูกค้าที่ซื้อทีวีมีแนวโน้มที่จะซื้อวีดีโอในเวลา ต่อมา
- Similar Time Sequence Discovery ใช้ค้นหาความเกี่ยวเนื่องกันระหว่างกลุ่มของข้อมูล 2 กลุ่ม ซึ่งการขึ้นต่อกัน ทางด้านเวลา โดยมีรูปแบบการเคลื่อนที่เหมือนกัน ผู้ขายสินค้ามักจะใช้เพื่อดูแนวโน้มเพื่อเตรียมสต๊อก เช่นเมื่อไรก็ตามที่ ยอดขาย สินค้าน้ำอัดลมสูงขึ้น ยอดขายมันฝรั่งจะสูงขึ้นตาม

6. Genetic Algorithm

เปรียบเสมือนเป็นการสร้างพันธุกรรมที่ดีสุดบนขั้นตอนของวิวัฒนาการทางชีวภาพ แนวคิดหลักก็คือเมื่อเวลาผ่านไป วิวัฒนาการของเซลล์ชีวิตจะเลือกสายพันธ์ที่ดีที่สุด “Fittest Species” Genetic Algorithm จะมีความสามารถในการทำงานแบบ รวมกลุ่มข้อมูลเข้าด้วยกัน เช่น อาจมีความต้องการที่จะแบ่งกลุ่มหรือจับรวมกลุ่มของข้อมูลเป็น 3 ชุด ขั้นตอนการทำงานของ Genetic Algorithm ก็จะเริ่มด้วยการจับกลุ่มข้อมูลเป็นกลุ่มๆ ด้วยการเดาสุ่ม เปรียบเหมือนกลุ่ม 3 กลุ่มนี้เป็นเซลล์ของสิ่งมีชีวิต Genetic Algorithm จะมี “Fittest Function” ที่จะบอกว่ากลุ่มข้อมูลใดเหมาะกับกลุ่มๆ ไหน โดย Fittest Function จะเป็นตัวบ่งชี้ว่าข้อมูลเหมาะกับกลุ่มมากกว่าข้อมูลอื่นๆ นอกจากนี้ในขั้นตอนต่อมา Genetic Algorithm จะมี “Operator” ซึ่งยอมให้มีการเลียนแบบและแก้ไขลักษณะของกลุ่มของข้อมูล Operator จะจำลองหน้าที่ของชีวิตที่ถูกพบในธรรมชาติ คือชีวิตมีการแพร่พันธุ์ จับคู่ผสมพันธุ์ และเปลี่ยนรูปร่างตามต้นแบบของพันธุ์ เปรียบกับข้อมูลถ้ามีข้อมูลใดในกลุ่ม ของข้อมูล ถูกพบว่าตรงกับคุณสมบัติของ Fittest Function แล้ว มันจะคงอยู่และถูกถ่ายเข้าไปในกลุ่มนั้น แต่ถ้าไม่ตรงกับคุณสมบัติ ก็ยังมีโอกาสที่จะถ่ายข้ามไปยังกลุ่มอื่นได้

7. Rule Induction

Rule Induction เป็นวิธีสำหรับการดึงเอาชุดของกฎเกณฑ์ต่างๆ มาเพื่อจัดแบ่งเงื่อนไขหรือกรณี ดังที่กล่าวข้างต้น โครงสร้างต้นไม้สามารถสร้างชุดของกฎต่างๆ และขณะที่บางครั้งเรียกวิธีการแบบนี้ว่า การสร้างกฎใหม่จากตัวอย่าง แต่วิธีการ หลังก็ยังมีความหมายที่แตกต่างกัน เนื่องจากวิธีการ Rule Induction จะสร้างชุดของกฎที่เป็นอิสระซึ่งไม่จำเป็นต้อง อยู่ในรูปโครงสร้างของต้นไม้ เพราะตัวสร้างกฎ (Rule Inducer) ไม่ได้บังคับการแตกข้อมูลเป็นแต่ละระดับ แต่อาจจะสามารถค้นหา Pattern ที่แตกต่างกันได้และบางครั้งอาจดีกว่าสำหรับการจัดแบ่ง Class ของผลลัพธ์

8. K-nearest neighbor

มนุษย์เมื่อต้องลองแก้ปัญหาใหม่ โดยทั่วไปมักจะมองที่ทางแก้ปัญหาอย่างง่ายซึ่งพวกเขาเคยใช้แก้อย่างได้ผลมาก่อน เทคนิคของ K-nearest neighbor (K-NN) ก็ใช้วิธีการเดียวกันในการจัดแบ่งคลาสนั่นเอง เทคนิคนี้จะตัดสินใจ ว่าคลาสไหนที่จะแทนเงื่อนไขหรือกรณีใหม่ๆได้บ้าง โดยการตรวจสอบจำนวนบางจำนวน (“K” ใน K-nearest neighbor) ของกรณีหรือเงื่อนไขที่เหมือนกันหรือใกล้เคียงกันมากที่สุด โดยจะหาผลรวม (Count Up) ของจำนวนเงื่อนไข หรือกรณีต่างๆสำหรับแต่ละคลาส และกำหนดเงื่อนไขใหม่ๆ ให้คลาสที่เหมือนกันกับคลาสที่ใกล้เคียงกับมันมากที่สุด

สิ่งแรกที่เราต้องทำในการนำเทคนิคของ K-NN ไปใช้ในตัวอย่างนี้คือ หาวิธีการวัดระยะห่าง (Distance) ระหว่างแต่ละ Attribute ในข้อมูลให้ได้ และจากนั้นคำนวณค่าออกมา ซึ่งวิธีนี้จะเหมาะสำหรับข้อมูลแบบตัวเลข (ต่างกับ Decision Tree) แต่ตัวแปรที่เป็นค่าแบบไม่ต่อเนื่องนั้นก็สามารถทำได้ เพียงแต่ต้องการการจัดการแบบพิเศษเพิ่มขึ้น อย่างเช่น ถ้าเป็นเรื่องของสี เราจะใช้อะไรวัดความแตกต่างระหว่างสีน้ำเงินกับสีเขียว ต่อจากนั้นเราต้องมีวิธีในการรวมค่าระยะห่างของ Attribute ทุกค่าที่วัดมาได้ เมื่อเราสามารถคำนวณระยะห่างระหว่างเงื่อนไขหรือกรณีต่างๆ ได้จากนั้นเราเลือกชุดของเงื่อนไข ที่ใช้จัดคลาสมาเป็นฐานสำหรับการจัดคลาสในเงื่อนไขใหม่ๆ ได้แล้ว เราจะตัดสินได้ว่าขอบเขตของจุดข้างเคียงที่ควรเป็นนั้น ควรมีขนาดใหญ่เท่าไร และอาจตัดสินใจได้ด้วยว่าจะนับจำนวนจุดข้างเคียงตัวมันได้อย่างไร (โดยอาจจะให้น้ำหนักกับ จุดข้างเคียงที่ใกล้ตัวมันมากที่สุดกว่าจุดที่ไกลห่างออกไป ก็ทำให้เราเลือกได้)

K-NN ค่อนข้างใช้ปริมาณงานในการคำนวณสูงมากบนคอมพิวเตอร์ เพราะเวลาที่ใช้สำหรับการคำนวณจะเพิ่มขึ้นแบบแฟคทอเรียลตามจำนวนจุดทั้งหมด ขณะที่ Decision Tree หรือ Neural Network จะประมวลผลเพื่อสร้างเงื่อนไข หรือกรณีใหม่ได้รวดเร็วกว่า เพราะเทคนิคของ K-NN ต้องการให้มีการคำนวณเกิดขึ้นทุกครั้งที่มีกรณีใหม่ๆ เกิดขึ้น ดังนั้นเพื่อจะเพิ่มความรวดเร็วสำหรับเทคนิค K-NN ให้มากขึ้น ข้อมูลทั้งหมดที่ใช้บ่อยจะต้องถูกเก็บไว้ในหน่วยความจำ (Memory) วิธีนี้จะมีชื่อว่า Memory-Based Reasoning ซึ่งจะเป็นวิธีที่นำมาอ้างถึงเป็นประจำในการจัดเก็บกลุ่มคลาสของ K-NN ในหน่วยความจำ

ถ้าข้อมูลที่ต้องการหาคำตอบมีตัวแปรอิสระเพียงไม่กี่ตัวแล้ว จะทำให้เราสามารถเข้าใจ Model K-NN ได้ง่ายขึ้น ตัวแปรเหล่านี้ยังมีประโยชน์ด้วยสำหรับนำมาสร้าง Model ต่างๆ ที่เกี่ยวข้องกับชนิดของข้อมูลที่ไม่เป็นมาตราฐาน เช่น Text เพียงแต่อาจต้องมีมาตราฐานการวัดค่าสำหรับชนิดของข้อมูลดังกล่าวที่เหมาะสมด้วย

9. Association and Sequence Detection

Association Discovery ใช้ในการหากฎความสัมพันธ์ที่เกิดขึ้นระหว่าง Item ต่างๆ เช่นการใช้ Market-basket analysis เพื่อวิเคราะห์ข้อมูลการสั่งซื้อสินค้า Sequence Detection ก็เป็นวิธีการในทำนองเดียวกัน แต่จะใช้ลำดับของเหตุการณ์ ที่เกิดขึ้นเข้ามาเกี่ยวข้องด้วย

เราจะเขียนความสัมพันธ์ออกมาในรูปของ A → B เรียกว่า A ว่าเป็นเหตุการณ์ที่เกิดขึ้นก่อน (Antecedent) หรือ LHS (Left - Hand Side) และเรียก B ว่าผลของเหตุการณ์ (Consequent) หรือ RHS (Right – Hand Side) เช่นในกฎของความสัมพันธ์ “ถ้าคนซื้อค้อน แล้วจะซื้อตะปู” เหตุการณ์ที่เกิดก่อนก็คือ “คนซื้อค้อน” และผลที่ตามมาก็คือ “ซื้อตะปู”

วิธีการที่ง่ายที่สุดในการวัดสัดส่วนของ Item ที่เกิดขึ้นใน Transaction ก็คือใช้ในการนับ เราจะเรียกความถี่ของความสัมพันธ์ที่เกิดขึ้นซึ่งปรากฏอยู่ในฐานข้อมูลว่า Support หรือ Prevalence เช่น จากตัวอย่างความสัมพันธ์ของค้อนและตะปู ถ้าความสัมพันธ์ของค้อนและตะปูจำนวน 15 Transaction จากจำนวนทั้งหมด 1,000 Transaction เราก็จะได้ค่า Support ของความสัมพันธ์นี้ 1.5% ค่า Support ที่มีค่าในระดับต่ำ เช่นหนึ่งในล้าน อาจแสดงให้เห็นถึงความไม่มีนัยสำคัญของความสัมพันธ์นั้นก็ได้นอกจากเราจะดูความถี่ที่เกี่ยวข้องกับเหตุการณ์ที่เกิดขึ้นของ Item นั้นๆแล้ว เราจะต้องดูความถี่ของเหตุการณ์อื่นๆ ที่เกิดขึ้นร่วมกับ Item นั้นด้วยในการหากฎที่มีระดับนัยสำคัญ หากเราตั้งโจทย์ว่า เมื่อมีเหตุการณ์ A (Antecedent) เกิดขึ้นเป็นจำนวนหนึ่ง จะมีเหตุการณ์ B (Consequent) เกิดขึ้นเป็นจำนวนเท่าใด หมายความว่า เราต้องหาเงื่อนไขที่จะทำนายเหตุการณ์ B ที่เกิดขึ้นเนื่องจาก A เมื่อเปรียบเทียบกับปัญหาในข้างต้นจะได้ว่า “เมื่อผู้คนซื้อค้อนไปแล้ว บ่อยแค่ไหนที่เขาจะซื้อตะปูไปด้วย” เราเรียกการทำนายผลอย่างมีเงื่อนไขนี้ว่าความเชื่อมั่น (Confidence) เราจะคำนวณความเชื่อมั่นออกมาในรูปของอัตราส่วน (ความถี่ของ A และ B)/(ความถี่ของ A)

Lift เป็นเครื่องมืออีกอย่างหนึ่งที่ใช้ในการวัดอิทธิพลที่มีความสัมพันธ์ที่เกิดขึ้น ค่า Lift ที่มากแสดงว่ามีความ เป็นไปได้มากที่เมื่อเกิดเหตุการณ์ A ขึ้นแล้ว จะมีเหตุการณ์ B จะเกิดขึ้นตามมา Lift จะคำนวณออกมาในรูปอัตราส่วนของ (ความเชื่อมั่นของ A→ B)/(ความถี่ของ B)

ผู้ค้าปลีกวัสดุภัณฑ์อาจแปลความหมายของตัวเลขเหล่านี้ได้ว่า การขายค้อนและตะปูสามารถนำมาเป็นตัวพยากรณ์ การขายไม้แปรรูปได้ดีกว่าจะนำการขายค้อนมาพยากรณ์การขายตะปู หากไม้แปรรูปเป็นสินค้าได้กำไรดีในกิจการ เราก็สามารถ นำข้อมูลที่ได้จากการวิเคราะห์มาวางแผนกลยุทธ์ทางการตลาด

คุณลักษณะอีกอย่างหนึ่งของตัวสร้างกฎความสัมพันธ์ก็คือ มีความสามารถในการระบุลำดับขั้นของ Item จากตัวอย่าง เราจะมองถึงข้อมูลของค้อนและตะปูโดยรวม ไม่ได้มองลงไปในสินค้าแต่ละตัวเราจึงต้องเลือกข้อมูลสรุปที่ได้มาใช้ด้วยความ ระมัดระวัง มิเช่นนั้นเราอาจไม่ได้ข้อ

มูลที่ต้องการจริงๆก็ได้ โครงสร้างตามลำดับขั้นของ Item จะทำให้เราสามารถควบคุม ระดับของข้อมูลสรุปที่ได้และสามารถทำการทดลองหาข้อมูลสรุปในระดับต่างๆ

Sequence Detection จะเป็นการเพิ่มตัวแปรด้านเวลาเข้าไป ทำให้สามารถติดตามลำดับเหตุการณ์ที่เกิดขึ้น เพื่อนำมาใช้ในการวิเคราะห์พฤติกรรมของข้อมูล

บ่อยครั้งที่ยากในการตัดสินใจว่าเราจะทำอย่างไรกับกฎความสัมพันธ์ที่ได้ ตัวอย่างในเรื่องแผนการวางผลิตภัณฑ์เพื่อ จัดจำหน่ายการวางผลิตภัณฑ์ที่มีความใกล้เคียงกันไว้ด้วยกันอาจเป็นการลดรายได้รวมทางการตลาดลงไป เนื่องจากลูกค้าจะ เลือกหยิบสินค้าที่ต้องการเพียงอย่างเดียวแทนที่จะเดินเลือกซื้อสินค้าที่ต้องการไปเรื่อยๆ นี่แสดงให้เห็นว่าการวิเคราะห์และ การทดลองมักมีความจำเป็นต้องใช้ร่วมกับกฎความสัมพันธ์ที่ได้จากการวิเคราะห์ เพื่อให้ได้ประโยชน์สูงสุด

10. Logistic Regression

Logistic Regression เป็นการวิเคราะห์ความถดถอยแบบเส้นตรงทั่วไป ที่ใช้ในการพยากรณ์ผลลัพธ์ของ สองตัวแปรเช่น Yes/No หรือ 0/1 แต่เนื่องจากตัวแปรตาม (Dependent Variable) มีค่าเพียงสองอย่างเท่านั้น เราจึงไม่สามารถสร้างแบบจำลองได้ด้วยวิธีการวิเคราะห์ความถดถอยแบบเส้นตรง

ดังนั้น แทนที่จะทำการพยากรณ์ผลลัพธ์โดยอาศัยเพียงค่าของตัวแปรตามที่ได้ เราจะสร้างแบบจำลองโดยอาศัย Algorithm ของความน่าจะเป็นของการเกิดเหตุการณ์ เราเรียกอัลกอริทึมที่สร้างขึ้นมานี้ว่า Log Odds หรือ logic Transformation

$$\text{อัตราส่วนความน่าจะเป็น :} \quad \frac{\text{ความน่าจะเป็นที่เหตุการณ์จะเกิด}}{\text{ความน่าจะเป็นที่เหตุการณ์ไม่เกิด}}$$

สามารถแปลความหมายได้เช่นเดียวกันกับความน่าจะเป็นในเกมการแข่งขัน หรือในทางกีฬา เช่น เมื่อเราบอกว่า ความน่าจะเป็นที่ทีมใดทีมหนึ่งจะชนะการแข่งขันฟุตบอลคือ 3 ต่อ 1 หมายความว่าความน่าจะเป็นที่ทีมนี้จะชนะ เป็น 3 เท่าของโอกาสที่นะแพ้ หรือมีโอกาสชนะ 75% และมีโอกาสแพ้ 25% วิธีการเช่นนี้สามารถนำมาใช้กับกลุ่มลูกค้า ที่จะวิเคราะห์ได้เช่นกัน ตัวอย่างการส่งจดหมายให้กลุ่มลูกค้า หากเราบอกว่าโอกาสที่ลูกค้าจะตอบสนองเป็น 3 ต่อ 1 นั้นหมายความว่าลูกค้าที่ตอบจดหมายมีค่าเป็น 3 เท่าของลูกค้าที่ไม่ตอบจดหมาย

Neural Network จะใช้ Logistic Regression เป็นเครื่องมือที่ช่วยจำแนกประเภทของตัวแปรประเภทของลูกค้าที่จะซื้อหรือไม่ซื้อสินค้า และใช้การวิเคราะห์ความถดถอยในการวิเคราะห์ตัวแปรต่อเนื่อง เช่นความเป็น ไปได้ในการซื้อสินค้าของลูกค้า เป็นต้น

แม้ว่า Logistic Regression เป็นวิธีการที่มีประสิทธิภาพ แต่ก็มีข้อจำกัดในเรื่องความเป็นไปได้ของตัวแปรตาม (Dependent Variable) เนื่องจากตัวแปรตามเหล่านั้นอาจไม่เป็นอิสระกันก็ได้ นอกจากนั้นผู้ที่ทำการวิเคราะห์แบบจำลอง จะต้องอาศัยประสบการณ์ของตนเองในการวิเคราะห์และต้องทำการเลือกข้อมูลที่จะนำมาวิเคราะห์ได้อย่างถูกต้อง จากตัวอย่างที่ผ่านมา ผู้วิเคราะห์จะต้องเลือกระหว่าง รายได้ ค่า Square ของรายได้ หรือค่า Algorithm ของรายได้ จะเลือกตัวแปรใดมาทำการวิเคราะห์และพยากรณ์ จะเห็นได้ว่าการวิเคราะห์ส่วนใหญ่จะขึ้นอยู่กับประสบการณ์ของผู้ทำการวิเคราะห์ ซึ่งต้องเลือกตัวแปรและวิธีการที่เหมาะสม จึงจะทำให้ได้ผลการวิเคราะห์ที่ถูกต้อง

Neural Network จะใช้ Hidden Layer ในการประมาณรูปแบบการวิเคราะห์ที่ไม่ใช่เส้นตรง (Non - Linear) และทำการวิเคราะห์แบบกึ่งอัตโนมัติ ผู้ใช้จำเป็นต้องใช้ความชำนาญเฉพาะตัวกับระบบ Neural Network ตัวอย่างเช่น พฤติกรรมการเลือกฟังก์ชัน จะมีผลกระทบกับความสามารถในการเรียนรู้ของระบบ Neural Network ด้วยเป็นที่น่าสังเกตว่า Logic Transformation มีผลกระทบต่อ Logistic Regression เช่นเดียวกับที่พฤติกรรมการเลือกฟังก์ชันมีผลกระทบกับ Neural Network และนั่นก็เป็นเหตุผลหลักที่ไม่มี Hidden Layer ใดใน Neural Network เป็น Logistic Regression

11. Discriminant analysis

Discriminant analysis เป็นวิธีการทางคณิตศาสตร์ที่เก่าแก่วิธีหนึ่งซึ่งใช้ในการจำแนก และวิเคราะห์วิธีนี้ได้รับการ เผยแพร่ครั้งแรกในปี 1936 โดย R. A. Fisher เพื่อแยกต้น Iris ออกเป็น 3 พันธุ์ วิธีการนี้ทำให้ค้นพบพันธุ์ ของต้นไม้ประเภทอื่นๆ อีกมาก ผลลัพธ์ที่ได้จากแบบจำลองชนิดนี้ง่ายต่อการทำความเข้าใจ เพราะผู้ใช้งานทั่วๆ ไปก็สามารถ พิจารณาได้ว่าผลลัพธ์จะอยู่ทางด้านใดของเส้นทางในแบบจำลอง การเรียนรู้สามารถทำได้ง่าย วิธีการที่ใช้มีความไวต่อรูปแบบ ของข้อมูล วิธีนี้ถูกนำมาใช้มากในทางการแพทย์ สังคมวิทยา และชีววิทยา

Discriminant analysis ไม่เป็นที่นิยมในการทำ Data Mining เนื่องจากเหตุผล 3 ประการคือ

1. ตัวแปรที่ใช้ในการวิเคราะห์ต้องตั้งอยู่บนสมมุติฐานว่า ข้อมูลมีการกระจายแบบปกติ รูประฆังคว่ำ (Normally distributed) ซึ่งอาจเป็นไปไม่ได้
2. ตัวแปรต่างๆ ที่ยังไม่ได้รับการจัดลำดับ และไม่เป็นอิสระกัน ไม่สามารถใช้กับวิธีการนี้ได้
3. ขอบเขตข้อมูลที่ใช้ในการแบ่งแยกประเภท ต้องอยู่ในรูปแบบเส้นตรง (Linear form) แต่บางครั้งเราไม่สามารถแบ่งแยกข้อมูลบางอย่างได้

Discriminant analysis ที่มีการปรับปรุงต่อมาในระยะหลัง ได้แก้ปัญหาบางอย่างที่เกิดขึ้นในการวิเคราะห์ เช่น สามารถใช้ฟังก์ชัน Quadratic ได้ แทนที่จะต้องเป็นฟังก์ชั่นเส้นตรงเพียงอย่างเดียว นอกจากนี้ยังสามารถใช้กระจาย แบบปกติของข้อมูลโดยประมาณในการวิเคราะห์

12. Generalized Additive Models (GAM)

เป็น Model ที่ขยายความสามารถของ Linear Regression และ Logistic Regression ว่า Additive ก็เพราะว่ามีการตั้งสมมติฐานว่า Model สามารถเขียนออกมาได้ในรูปของผลรวมของ Possibly Non-Linear Function ซึ่ง GAM สามารถใช้งานได้ทั้งแบบ Regression และ Classification คุณสมบัติหลักที่เพิ่มเติมเข้าไปก็คือการหาค่า Lift ตัวแปรผลลัพธ์จะเกิดขึ้นจากฟังก์ชันใดของตัวแปรที่ใช้ในการพยากรณ์ก็ได้ ตราบใดที่ไม่มีการก้าวกระโดดที่ไม่ต่อเนื่อง ตัวอย่าง เช่น สมมติว่าการขาดการชำระเงินเป็นฟังก์ชันที่ซับซ้อนของตัวแปรรายได้ ซึ่งความน่าจะเป็นของการขาด การชำระเงินจะลดลงตามรายได้ที่เพิ่มขึ้น และความน่าจะเป็นของการขาดการชำระเงินจะเพิ่มขึ้นอีกครั้งในกลุ่มผู้มีรายได้ ปานกลาง ในที่สุดจะขึ้นสูงสุดก่อนที่จะตกลงอีกครั้งในกลุ่มมีรายได้สูง ในกรณีนี้ Linear Model จะไม่สามารถ แสดงให้เห็นถึงความสัมพันธ์ระหว่างรายได้กับความน่าจะเป็นของการขาดการชำระเงิน ซึ่งมีลักษณะเป็น Non-Linear ได้

GAM จะใช้ความสามารถของคอมพิวเตอร์ในการค้นหารูปแบบของฟังก์ชันที่ให้ Curve ที่เหมาะสม ทำการรวม ค่าความสัมพันธ์ต่างๆ เข้าด้วยกัน ดังที่อธิบายมาแล้วข้างต้น แทนที่จะมีการใช้ Parameter จำนวนมาก เหมือนที่ Neural Network ใช้ GAM ก้าวไปเหนือกว่านั้นอีกขั้นหนึ่งและประเมินค่าของ Output ในแต่ละ Input และเช่นเดียวกันกับ Neural Network GAM จะสร้างเส้นโค้งขึ้นมาอย่างอัตโนมัติ โดยอาศัยข้อมูลที่มี

13. Multivariate Adaptive Regression Splits (MARS)

ในกลางทศวรรษที่ 80 Jerome H. Friedman หนึ่งในผู้ที่คิดค้น CART ได้พัฒนาวิธีการใหม่ขึ้นมา โดยต้องการจะกำจัดข้อเสียต่อไปนี้ออกไป

- Discontinuous predictions (Or hard splits)
- Dependence of all splits on previous ones
- Reduced interpretability due to interactions, especially high-order interaction

โดยการคิด MARS Algorithm โดยความคิดพื้นฐานง่ายๆ เพื่อที่จะจำกัดข้อเสียดังกล่าวโดย

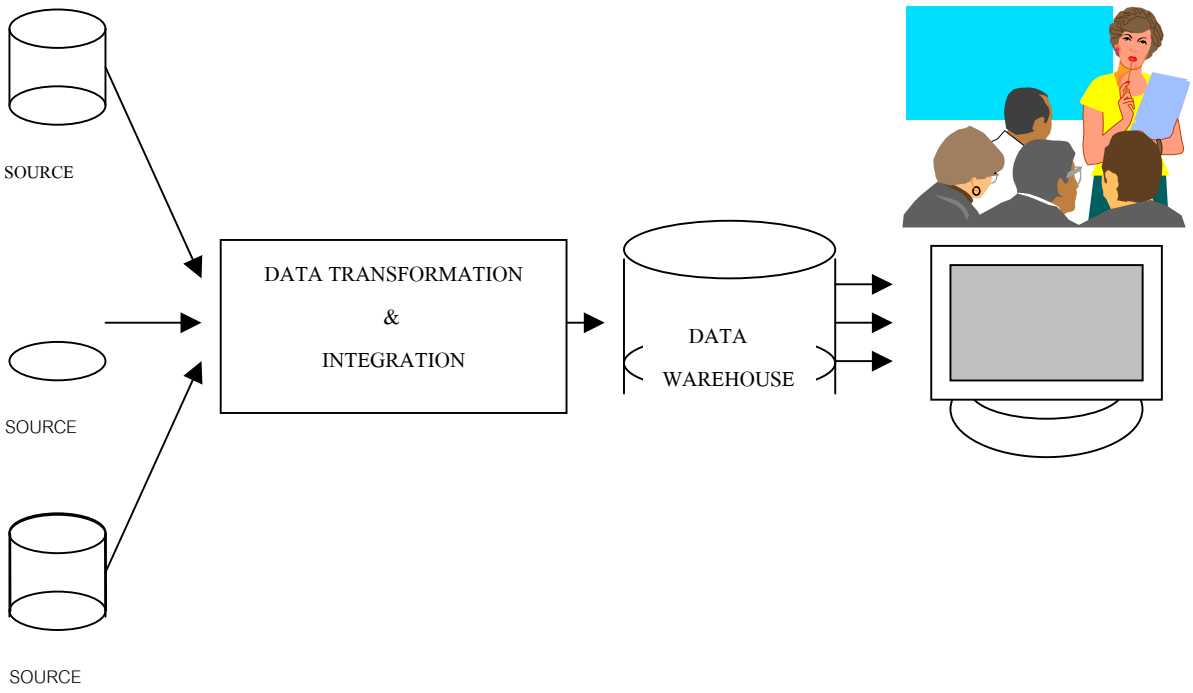
- แทนที่ Discontinuous Branching ที่ Node ด้วยเส้นตรง 1 คู่ และในขั้นตอนสุดท้ายของกระบวนการ สร้าง Model เส้นตรงดังกล่าวจะถูกแทนที่ด้วย Smooth Function เรียกว่า Splits
- ไม่จำเป็นที่ว่าการแบ่งแยกในครั้งใหม่ จะต้องขึ้นอยู่กับครั้งก่อน ทำให้ MARS สูญเสียโครง สร้างแบบ Tree ใน CART และไม่สามารถที่สร้างออกมาเป็นกฎได้ หรือกล่าวอีกนัยหนึ่ง MARS สามารถที่จะค้นหาและแสดงรายการตัวแปรอิสระที่มีความสำคัญสูงสุดเช่นเดียวกับปฏิสัมพันธ์ระหว่างตัวแปรอิสระ อีกทั้ง MARS

สามารถ Plot จุดแสดงความเป็นอิสระของแต่ละตัวแปรอิสระออกมาได้ ผลลัพธ์ที่ได้ก็คือ Non-linear step-wise regression tools

2.5 Data warehouse , Data mining และ Data Mart

Data warehouse

จากรูปแสดงขั้นตอนของการทำ Data Warehouse



รูป 2.6 แสดง ตำแหน่งของ DATA WAREHOUSE

ขั้นตอนแรกก่อนที่จะทำ Data Mining Process คือการจัดขนาดของข้อมูลใหญ่ๆ ให้อยู่ในรูปแบบที่ง่ายต่อการเข้าถึง การเข้าไปใช้งานและการ Sort โดยผู้ใช้ การรวบรวมข้อมูลใน Process ของ Data Mining อาจจะยุ่งยาก เพราะข้อมูลไม่อยู่ในรูปแบบที่เหมาะสมที่จะเข้าไปใช้งานได้

จุดประสงค์ของ Data Warehouse คือช่วยปรับปรุงประสิทธิภาพในการตัดสินใจเกี่ยวกับธุรกิจที่เกี่ยวข้องกับตัวเลขจำนวนมาก พื้นฐานดังกล่าวตั้งอยู่บนหลักของ Informational Data (ข้อมูลที่ใช้จัดการองค์กรซึ่งเป็นข้อมูล สรุปเพื่อการตัดสินใจ) แทนที่จะเป็น Operational Data (ข้อมูลที่ใช้ดำเนินกิจกรรมขององค์กร เช่น ข้อมูลของพวก Transaction ต่างๆ)

คำจำกัดความของ Data Warehouse คือการรวบรวมของข้อมูลเพื่อสนับสนุนการตัดสินใจของฝ่ายบริหาร ข้อมูล ดังกล่าว ถูกแบ่งเป็นระดับๆ หลายระดับ เพื่อให้เกิดความสามารถในการเข้าถึงข้อมูลได้อย่างรวดเร็ว

Subjected Oriented ข้อมูลใน Warehouse ถูกกำหนดในลักษณะ Business Term เช่น ลูกค้า , สินค้า, รายงานวิเคราะห์ยอดขาย

Integrated Term ที่ใช้ใน Data Warehouse จะต้องถูกกำหนดให้สมบูรณ์เหมือนกันทั้งองค์กร และจะต้อง ถูกต้องกับแหล่งข้อมูลทั้งภายในและภายนอก

Time Variant ข้อมูลใน Data Warehouse เป็นลักษณะ Time Stamp ณ เวลาที่ข้อมูลถูกใส่เข้ามาหรือข้อมูลถูกสรุป ดังนั้นจะเป็นการบันทึกในลักษณะต่อเนื่องและมีประวัติและแนวโน้มการวิเคราะห์ที่เป็นไปได้

Non Volatile เมื่อถูกใส่เข้ามาใน Data Warehouse แล้วข้อมูลจะไม่ถูก Update อีก ดังนั้นจึงเป็น แหล่งที่มีรายงานถูกต้องและใช้วิเคราะห์เชิงเปรียบเทียบ

โดยมีเครื่องมือ 2 ตัวในการจัดการทำ Data Warehouse คือ (Data Transformation, Data Cleaning) และ End User Data Access เครื่องมือเหล่านี้จะทำให้มั่นใจว่า Data Warehouse จะมีความถูกต้องของข้อมูล แม่นยำ มีประสิทธิภาพและมีต้นทุนในการบริหารต่ำ

ขั้นตอนของการทำ Data Warehouse จะเริ่มจากขั้นตอนต่อไปนี้

- Data Extraction จะช่วยสังเคราะห์เอาข้อมูลที่เป็นประโยชน์สำหรับ Data Mining เท่านั้น
- Sampling and Selecting จะเป็นตัวกำหนดขนาดของข้อมูล
- Aggregation จะเป็นตัวรวบรวมข้อมูลที่เกี่ยวข้องเข้าไว้ด้วยกัน
- Data Cleaning จะเป็นตัวสร้างความมั่นใจว่าข้อมูลจะสมบูรณ์ และลดความซ้ำซ้อนของข้อมูล
- Normalization จะเป็นตัวช่วยลดความซ้ำซ้อนของข้อมูล
- Overlay เช่น Demographic จะเป็นตัวช่วยทำให้เร่ง Data Access ได้เร็วขึ้น

ปัญหาหลักๆ ของ Data Warehouse ก็คือคุณภาพของข้อมูล เพื่อหลีกเลี่ยงปัญหา GIGO (Garbage In Garbage Out) ข้อมูลควรมี Missing Value น้อยที่สุด เพราะอาจจะมีผลกระทบต่อการวิเคราะห์ข้อมูลของ Data Mining ได้

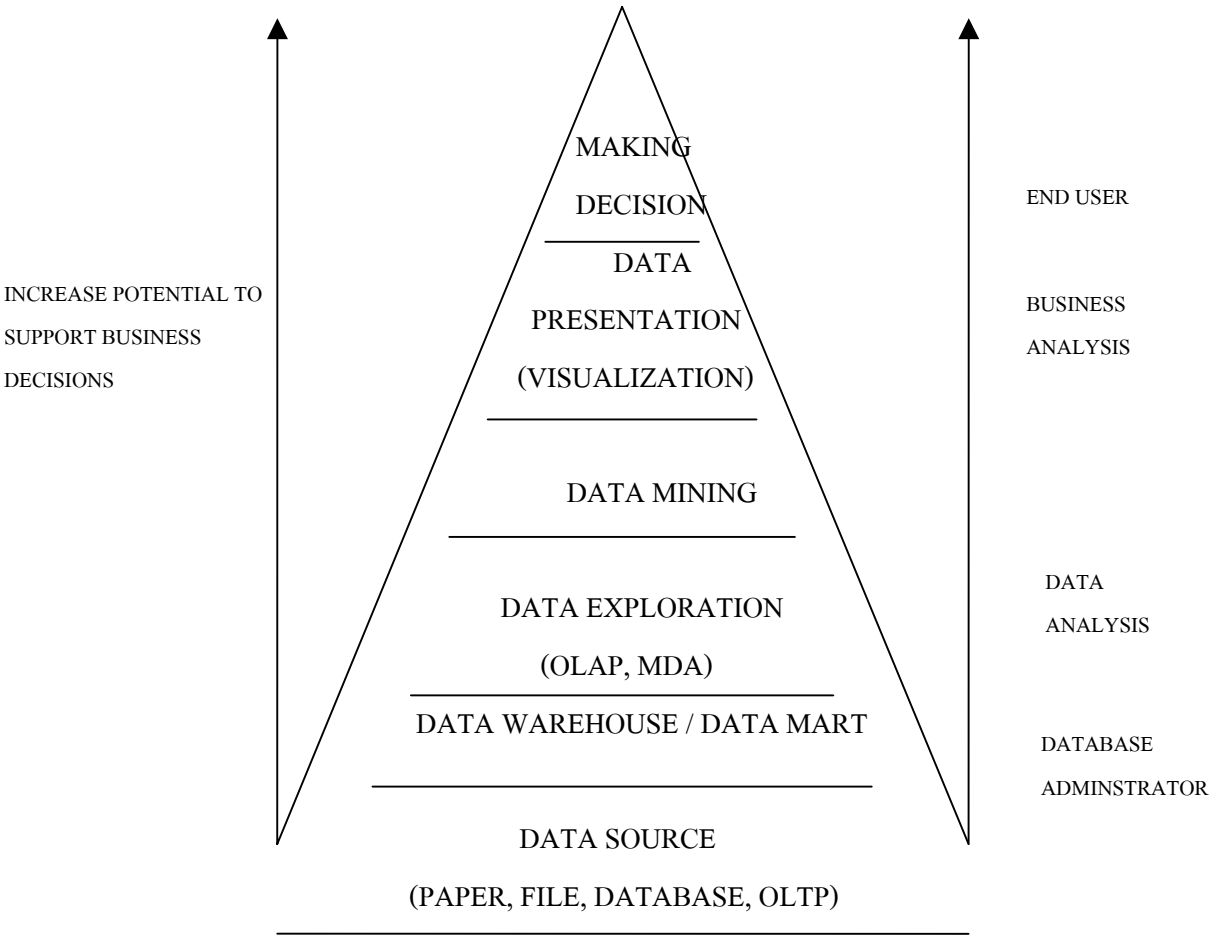
Data Mart

ปัจจุบันหลายองค์กรเริ่มหันไปหา Data Mart ซึ่งมีความเฉพาะเจาะจงมากกว่า และเข้าถึงได้มากกว่า แต่ขนาดเล็กกว่า Data Warehouse มาก Data Mart เป็นการแยกเก็บข้อมูลจาก Data Warehouse เพื่อเก็บข้อมูลให้กับแผน เฉพาะที่มีการเรียกใช้ข้อมูลนั้นๆ บ่อยเพื่อไม่ให้เกิดความซ้ำซ้อนและลดปริมาณข้อมูลที่ไม่เกี่ยวข้องทำให้การ Process ข้อมูลเร็วขึ้น

Data Mining

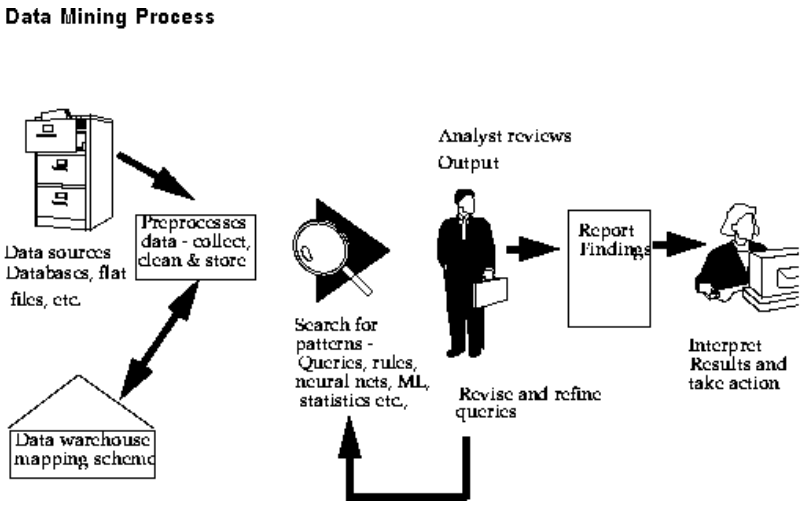
ถือได้ว่าเป็นระดับการนำข้อมูลไปใช้ที่สูงกว่า Data Warehouse และ Data Mart Data Mining เป็นวิธีคิดที่จะนำเอาข้อมูลมาใช้เพื่อการวิเคราะห์ให้เกิดประโยชน์สูงสุด โดยเฉพาะอย่างยิ่งการตัดสินใจของฝ่ายบริหาร ซึ่งระบบนี้เป็นขั้นตอนต่อไปของ Data Warehouse มีระบบการทำงานอัตโนมัติ สามารถตัดสินใจแทนผู้ใช้ได้ โดยอาศัยกฎเกณฑ์ต่างๆ ที่กำหนดขึ้นมาแล้วป้อนให้คอมพิวเตอร์คิด เครื่องมือทางธุรกิจ,เทคนิคต่างๆที่เราใช้เพื่อสนับสนุนการตัดสินใจทางธุรกิจนั้นมีพื้นฐานมาจาก เทคโนโลยีสารสนเทศ

จากรูป เริ่มต้นตั้งแต่ ตารางข้อมูลธรรมดาไปจนถึงการตัดสินใจระดับสูง เราจะเห็นได้ว่า Data Mining เป็นส่วนประกอบอันใหม่ที่มีความสำคัญของเครื่องมือทางธุรกิจอย่างหนึ่งคุณค่าของข้อมูลที่ใช้สนับสนุนการตัดสินใจจะเพิ่มขึ้นจากล่างไปบนสุดของรูปปิรามิด จำนวนของข้อมูลและขนาด และระดับการตัดสินใจในข้อมูลที่ลักษณะที่ต่างๆ กัน จึงมีระดับของผู้ตัดสินใจต่างกัน Database administrator จะตัดสินใจบนระดับของ Data Warehouse และแหล่งข้อมูลเท่านั้น ส่วนนักวิเคราะห์ธุรกิจและผู้บริหารจะตัดสินใจบนเหนือของปิรามิด



รูป 2.7 แสดง Data Mining และเครื่องมือทางธุรกิจต่าง ๆ (Cabena et al., 1997)

การนำข้อมูลของ Data Warehouse ที่รวบรวมข้อมูลจากหลายๆ ที่และดึงข้อมูลเหล่านั้นเข้าไปในฐานข้อมูล ที่มีขนาดใหญ่ โดย Data Mining จะนำข้อมูลมาสร้างแบบจำลองทางสถิติ ในการหารูปแบบความสัมพันธ์ของฐานข้อมูลที่มีอยู่ ในการช่วยวิเคราะห์การตัดสินใจในธุรกิจหรือกิจการอื่นๆ ตามต้องการ



รูป 2.8 แสดงกระบวนการจัดการ Data Mining

ความสัมพันธ์ระหว่างData Warehouse กับ Data Mining

1. ระบบคลังข้อมูล (Data Warehouse)

คือระบบคลังข้อมูลเพื่อการบริหารได้ถูกออกแบบมาเพื่อใช้เก็บข้อมูลขนาดใหญ่ในรูปแบบ RDBMS (Relational Database Management Systems) ที่มีประสิทธิภาพสูง ในระบบคลังข้อมูล ข้อมูลที่ซับซ้อนจะถูกรวบรวม หรือเปลี่ยนแปลงให้ง่ายต่อการจัดเก็บและสามารถเรียกกลับมาใช้ได้อย่างรวดเร็ว ถูกต้อง โดยข้อมูลต่างๆ เหล่านี้จะถูกนำมาใช้ สำหรับการวิเคราะห์และช่วยในเรื่องการตัดสินใจ (DSS) โดยอาศัยเครื่องมือ (Tool) ต่างๆ มาใช้ในการจัดการทำรายงาน และเพิ่มประสิทธิภาพสำหรับการตัดสินใจให้รวดเร็วยิ่งขึ้น โดยผู้บริหาร นักวางแผนงาน และนักวิเคราะห์ข้อมูลสามารถ เรียกหาข้อมูล หรือ Query เพื่อให้ได้รับคำตอบในรูปแบบตารางรายงาน หรือ รายงาน กราฟ ซึ่งเครื่องมือนี้ ถือได้ว่าเป็น สิ่งสำคัญในอันที่จะนำองค์กรไปสู่ความสำเร็จในกระบวนการการตัดสินใจ

กุญแจสำคัญสำหรับองค์ประกอบของระบบคลังข้อมูล

The integration environment การรวบรวมข้อมูลจากแหล่งต่างๆ

The data warehouse environment การจัดการข้อมูลให้อยู่บนมาตรฐานเดียวกัน (Homogeneous model)

The decision support environment เป็นกระบวนการสนับสนุนการตัดสินใจโดยใช้เครื่องมือต่างๆ เช่น Ad-hoc querying, What-if analysis, Analyzing or OLAP and Data mining เพื่อช่วยในการวิเคราะห์โอกาสทางธุรกิจ และการวางแผนเชิงกลยุทธ์

2. ระบบการวิเคราะห์ข้อมูลและช่วยในการตัดสินใจ (Data Mining)

องค์กรธุรกิจโดยเฉพาะธุรกิจให้บริการด้านโทรคมนาคมต่างพยายามศึกษาข้อมูลจากการให้บริการเพื่อสร้างความพึงพอใจของลูกค้าหรือผู้ใช้บริการ และหาวิธีการบริหารข้อมูลและนำข้อมูลที่เป็นประโยชน์เหล่านั้นมาใช้ให้มีประสิทธิภาพ และได้ประสิทธิผลสูงสุด Data mining อาจเป็นกุญแจสำคัญที่จะนำองค์กรไปสู่ผู้นำในตลาดได้ ซึ่งเทคโนโลยี data mining ได้ใช้ความก้าวหน้าทางการวิเคราะห์ทางสถิติและเทคนิคแบบจำลอง ในการหารูปแบบและความสัมพันธ์ของฐานข้อมูล (database) หรือข้อมูลบนระบบคลังข้อมูล (Data Warehouse) ขององค์กรที่ซ่อนอยู่ ซึ่งการใช้วิธีธรรมดา อาจไม่สามารถมองเห็น

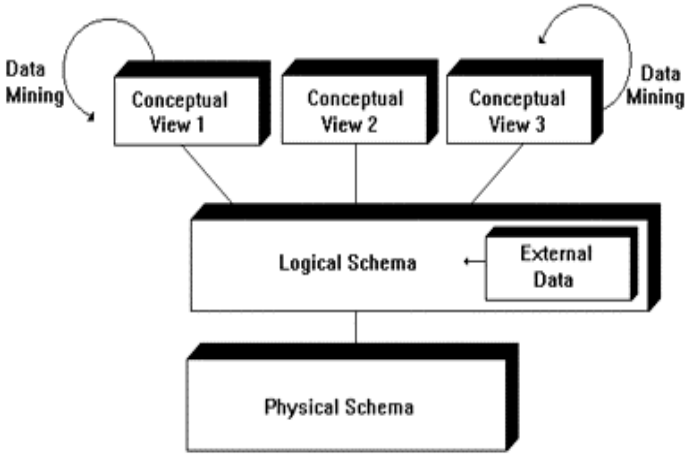
3. ความสัมพันธ์ระหว่าง Data mining และ Data Warehouse

ความเกี่ยวข้องสัมพันธ์กันระหว่าง Data Warehouse กับ Data Mining อยู่ที่ Data Mining คือเทคนิค อันล้ำยุคในการค้นหารูปแบบ (Pattern) ของข้อมูลซึ่ง Tool ที่ใช้ทำ Data Mining แตกต่างจาก Tool ที่ใช้ใน การค้นหาและรายงานโดยทั่วไป โดยได้ถูกรวบรวมเอาไว้เป็น Package ใน Software Tools บริษัทที่ดำเนินธุรกิจสามารถซื้อ Tool ตัวนี้ได้จากร้านค้าคอมพิวเตอร์ ด้วยเทคนิคของ Data Mining อย่างเช่น Neural Networks, Decision Tree, Statistical Processing และ Data Visualization จะสามารถช่วยให้การสำรวจรูปแบบข้อมูลและวิเคราะห์ข้อมูลใน Data Warehouse ทำได้ดีขึ้น แนวโน้มที่ไกลตัวที่สุดซึ่งกำลังพัฒนาตัวเองอยู่ เรามักจะได้ยินชื่อว่า Warehouse Enabled OLTP คือ Application ที่รวบรวมเอาการสนับสนุนการตัดสินใจจากการทำData Warehouse และการประมวลผลแบบ Online Transaction Processing : OLTP

รูปแบบของการสร้างระบบ Data Mining สามารถแยกออกจากส่วนของ Data Warehouse ได้เป็นลักษณะดังนี้

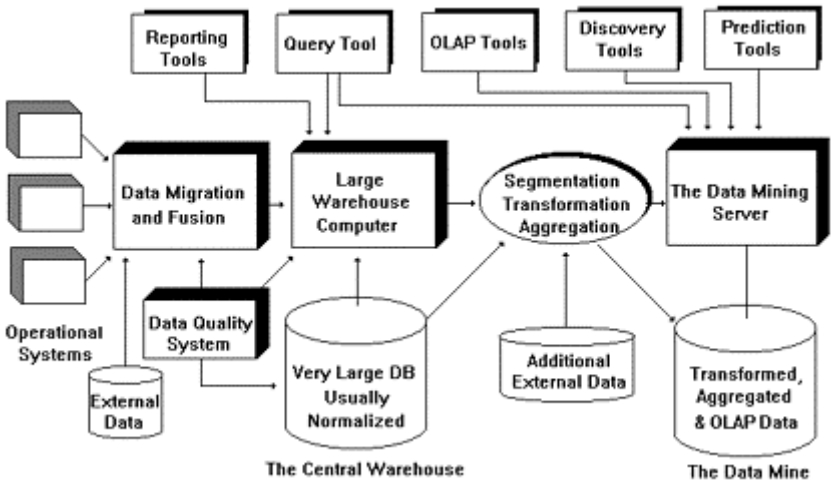
1. Data Mining Above the Warehouse

เหมาะสำหรับการวิเคราะห์ข้อมูลประกอบ ที่ไม่ใช่เป้าหมายหลักขององค์กร (Key objective) หรือข้อมูลจำนวนไม่มากไม่สลับซับซ้อน มีลักษณะดังภาพ



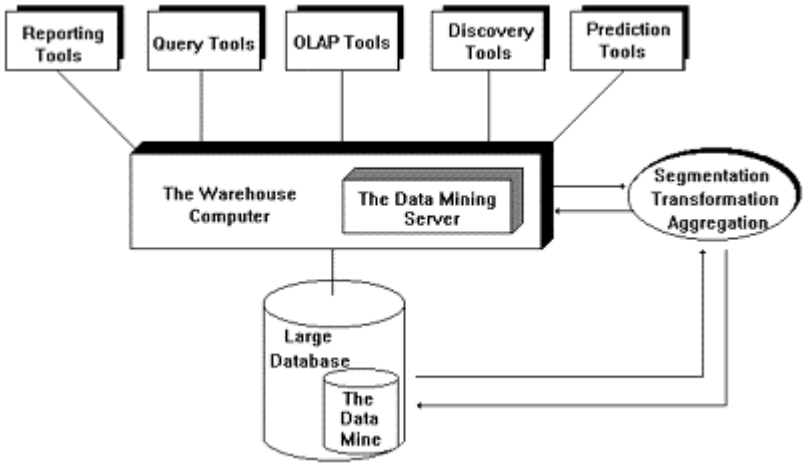
รูป 2.9 แสดง Data Mining Above the Warehouse

2. Data Mining Beside the Warehouse



รูป 2.10 แสดง Data Mining Beside the Warehouse

3. Data Mining Within the Warehouse



รูป 2.11 แสดง Data Mining Within The Warehouse

4. Stand-alone Data Mining

แต่สำหรับองค์กรที่มีการใช้ประโยชน์จากข้อมูลจำนวนมาก แม้ว่าจะมีฐานข้อมูลย่อยของแต่ละแผนกอยู่ โดยมีระบบการจัดการข้อมูลที่มีมาตรฐาน มีประสิทธิภาพดีอยู่แล้ว การติดตั้งระบบคลังข้อมูล(Data Warehouse) ก็ไม่จำเป็นสำหรับระบบการจัดการวิเคราะห์ข้อมูลและช่วยการตัดสินใจ(Data mining) ก็ได้

2.6 Algorithm สำหรับ Data Mining

เป็นขั้นตอนในการเลือกใช้ Algorithm ที่เหมาะสมกับปัญหาที่ต้องการทำ Data Mining ซึ่งขึ้นอยู่กับลักษณะ ของปัญหาและลักษณะของข้อมูล เช่น ถ้าปัญหาคือ “ทำไมลูกค้าเปลี่ยนใจไปใช้ผลิตภัณฑ์ของบริษัทคู่แข่ง” ซึ่งเรามีข้อมูล 2 ส่วนคือ ข้อมูลของลูกค้าที่เปลี่ยนใจไปใช้ผลิตภัณฑ์ของบริษัทคู่แข่งและลูกค้าที่ยังคงใช้ผลิตภัณฑ์ของบริษัท โดยสิ่งที่เรา ต้องการคือ รูปแบบของความสัมพันธ์บางอย่างของลูกค้าที่ทำให้ลูกค้ารายนั้นมีแนวโน้มที่จะเปลี่ยนใจไปใช้ผลิตภัณฑ์ของคู่แข่ง ซึ่ง Algorithm ที่เหมาะสมกับปัญหาลักษณะนี้ได้แก่ Classification Tree Algorithm เป็นต้น การเลือก Algorithm นั้นอาจเลือกใช้มากกว่า 1 Algorithm เพื่อใช้ในการเปรียบเทียบผลลัพธ์

อัลกอลิทึม ในการทำ Data Mining มีอยู่มากมาย ทั้งนี้เพราะ Data Mining ครอบคุมเนื้อที่กว้างมากนั่นเอง ยกตัวอย่างอัลกอลิทึมที่สามารถนำไปประยุกต์ใช้กับงานทั่วๆไปได้

1. อัลกอลิทึม การนับความถี่ของรายการ

การนับจำนวนรายการที่เกิดขึ้นภายใต้เงื่อนไข เฉพาะ อัลกอลิทึมนี้มีที่มาจากการวิเคราะห์การซื้อของ เรียกว่า Market basket analysis กล่าวคือในการซื้อสินค้าของลูกค้า 1 ครั้ง โดยไม่ต้องจำกัดว่าจะซื้อสินค้าในห้างร้าน หรือสั่งผ่านไปรษณีย์ หรือการสั่งซื้อสินค้าจาก visual store บนเว็บ โดยปกติเราต้องจะต้องการทราบว่าสินค้าใดบ้างที่ลูกค้ามักจะซื้อด้วยกัน เพื่อจะนำไปพิจารณาปรับปรุงการจัดวางสินค้าในร้าน หรือใช้เพื่อหาวิธีวางรูปคู่กันในใบโฆษณาสินค้า ก่อนอื่นกำหนดคำว่ากลุ่มรายการ (itemset) ก่อน หมายถึง กลุ่มสินค้าที่จะ ปรากฏร่วมกัน เช่น (รองเท้า,ถุงเท้า), (ปากกา, หมึก)หรือ(นม,น้ำผลไม้) โดยกลุ่มรายการดังกล่าวนี้ อาจจะจับคู่กลุ่มลูกค้ากับสินค้า ก็ได้เช่น วิเคราะห์หา ”ลูกค้าที่ซื้อสินค้าบางชนิดซ้ำๆกันอย่างน้อย 5 ครั้งแล้ว” กรณีนี้ฐานข้อมูลเรามีการเก็บรายการซื้อขายเป็นจำนวนมาก และคำถามข้างต้น (query) นี้จำเป็นต้องค้นหา ทุกๆคู่ของลูกค้ากับสินค้า เช่น {นาย ก, สินค้า A},{นาย ก, สินค้า B},{นาย ก, สินค้า C},{นาย ข, สินค้า B} เป็นต้น นับเป็นงานที่หนักพอควรสำหรับ DBMS และถ้าจะเขียน query ข้างต้นเป็น SQL จะได้ว่า

```
SELECT P.custid,P.item,SUM(P.qty)
FROM  Purchases P
GROUP BY P.custid,P.item
HAVING SUM(P.qty) > 5
```

หลังจากที่ DBMS ประมวลผล SQL นี้อย่างหนัก เนื่องจากมีข้อมูลที่จะต้องตรวจสอบมากมายหลายคู่ และแต่ละคู่ต้องค้นหามาจากทั้งฐานข้อมูลเลย แต่ผลลัพธ์ของ query ชนิดนี้ว่าเป็น “iceberg query” ซึ่งเปรียบเทียบกับ สำนวนไทยก็คือ งมเข็มในมหาสมุทรนั่นเอง

แสดง อัลกอลิทึม ในการค้นหากลุ่มรายการทั้งหมดจากฐานข้อมูล ภายใต้เงื่อนไขที่กำหนดขึ้น

```
for each item  // level 1
// นั่นคือปรากฏในจำนวนรายการที่มากกว่าที่กำหนด
check if it is a frequent itemset
K = 1
Repeat              // ทำซ้ำเพื่อหา frequent itemsets
        // level k +1
        for each new frequent itemset IK with K items
        generate all items IK + 1 with k +1 items,
        Ik is a subset of IK + 1
```

Scan all transactions once and check if the generated

K + 1 – itemsets are frequent.

k = k + 1

until no new frequent itemsets are identified

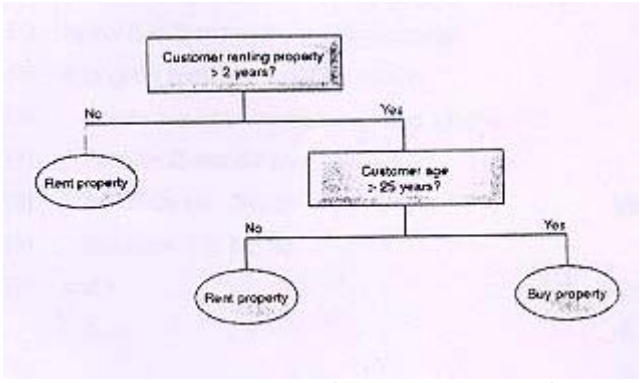
ผลลัพธ์ของ อัลกอลิทึมนี้ จะใช้แสดงกลุ่มของรายการที่ปรากฏบ่อยครั้ง ดังที่เรากำหนด

2. อัลกอลิทึม เพื่อการจัดหมวดหมู่ (Classification)

การจัดหมวดหมู่ของข้อมูลคือการสำรวจรายการในฐานข้อมูล เพื่อแยกแยะให้อยู่ในหมวดที่เราได้กำหนดไว้ล่วงหน้า แล้ว เช่น การแบ่งกลุ่มสินค้าเป็นกลุ่มเครื่องใช้ กลุ่มอาหารสด กลุ่มอาหารแห้ง เป็นต้น อัลกอลิทึม ที่ใช้ในการจัดหมวดหมู่ ออกเป็น 2 แบบ หลักๆคือ

- แบบต้นไม้ (Decision tree)
- แบบนิวรอลเน็ต (Neural network)

โครงสร้างแบบต้นไม้ เป็นที่นิยมกันมาก เป็นลักษณะที่คนจำนวนมากคุ้นเคย ทำให้เข้าใจได้ง่าย มีลักษณะเหมือนแผนภูมิองค์กร จากรูปแสดงให้เห็นถึง Decision tree สำหรับวิเคราะห์ว่าลูกค้าบ้านเช่ามีความสนใจ ที่จะซื้อบ้านเป็นของตนเองหรือไม่ โดยใช้ปัจจัยในการวิเคราะห์ คือ ระยะเวลาที่ลูกค้าได้เช่าบ้านมา และอายุของลูกค้า

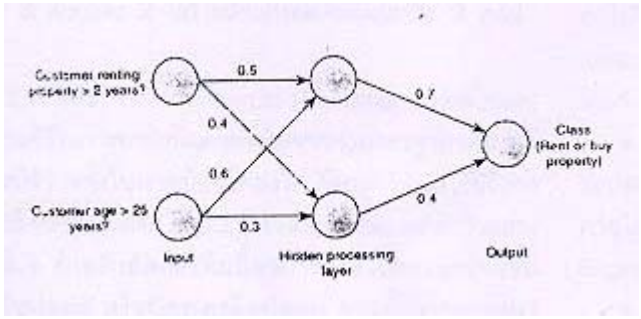


รูป 2.12 แสดงตัวอย่างของ Decision tree เพื่อวิเคราะห์โอกาสที่ลูกค้าบ้านเช่าจะซื้อบ้าน

โครงสร้างอีกแบบหนึ่ง ของ อัลกอลิทึมนี้คือ โครงสร้างนิวรอลเน็ตเวิร์ก

นิวรอลเน็ต หรือ นิวรอลเน็ตเวิร์ก เป็นเทคโนโลยี่ที่มีมาจากงานวิจัยด้านปัญญาประดิษฐ์ (Artificial Intelligence : AI) เพื่อใช้ในการคำนวณ ค่าฟังชันจากกลุ่มข้อมูล วิธีการของนิวรอลเน็ต (แท้ที่จริงต้องเรียกให้เต็มว่า Artificial Neural Networks หรือ ANN) เป็นวิธีการที่ให้เครื่องเรียนรู้

จากตัวอย่างต้นแบบ แล้วฝึก (train) ให้ระบบรู้จักที่จะคิดแก้ปัญหาที่กว้างขึ้นได้ ในโครงสร้างของนิวรอลเน็ต จะประกอบด้วยโหนด (node) สำหรับ อินพุต-เอาต์พุต และการประมวลผล กระจายอยู่ในโครงสร้างเป็นชั้นๆ ได้แก่ input layer ,output layer และ hidden layer การประมวลผลของนิวรอลเน็ต จะอาศัยการส่งการทำงานผ่านโหนดต่างๆ ใน layer เหล่านี้ ตัวอย่างรูปเป็นการวิเคราะห์แบบเดียวกับรูปข้างบน ในโครงสร้างแบบ นิวรอลเน็ต



รูป 2.13 แสดง นิวรอลเน็ต เพื่อวิเคราะห์การเช่าและซื้อบ้านของลูกค้า

ตารางแสดง Business_info แสดงถึงรายการทั้งหมด เกี่ยวกับลูกค้าบ้านเช่าของบริษัท โดยมีรายละเอียดเกี่ยวกับอายุ และระยะเวลาการเช่า รวมทั้งการซื้อบ้านของลูกค้าแต่ละราย ดังนี้

ตาราง 2.2 Business_info

Age	**Rent_period**	**Buy**
23	3	No
36	1.5	No
20	1.5	No
27	2	Yes
20	1	No
50	2.5	Yes
36	1	No
36	2	Yes
22	2.5	No

SQL สำหรับ Decision tree ของตัวอย่างนี้แบ่งเป็น 2 ชุด สำหรับปัจจัยแต่ละอย่าง

1. SQL สำหรับ root node ดังนี้

```
SELECT B.rent_Period , B.Buy ,COUNT(*)
FROM Business_info B
WHERE B.Rent_Period > 2
GROUP BY        B.Rent_Period,B.Buy
```

ตาราง 2.3 ผลลัพธ์ของ SQL สำหรับ root node

Rent_Period	**Buy**	**Yes**	**No**
1	0	2	
1.5	0	2	
2	2	0	
2.5	1	1	
3	0	1	

2. SQL สำหรับ node ที่เป็น child ทางขวาของ root คือ

```
SELECT B.Age , B.Buy ,COUNT(*)
FROM Business_info B
WHERE B.Age > 25
GROUP BY        B.Age,B.Buy
```

ตาราง 2.4 แสดงผลลัพธ์ของ SQL สำหรับ node ที่เป็น node ทางขวาของ root

Rent_period	**Buy**	**Yes**	**No**
20	0	2	
22	0	1	
23	2	1	
27	1	0	
36	1	2	
50	1	0	

ผลลัพธ์ที่ได้แต่ละโหนดของ Decision tree เรียกว่า AVC sets (Attribute value , Class label) จากตัวอย่างข้างต้นจะเห็นได้ว่ามี 2 AVC sets เพื่อใช้ในการจัดกลุ่มลูกค้า

แสดงวิธีการสร้าง Decision tree ในหน่วยความจำ

Top-Down Decision tree Induction schema :

BuildTree (Node n. data partition D ,split selection method S)

(1) Apply S to D to find the splitting criterion

(2) If (a good splitting criterion is found)

(3) Create two children nodes n1 and n2 of n

(4) Partition D into D1 and D2

(5) BuildTree (n1 ,D1 ,S)

(6) BuildTree(n2 , D2,S)

(7) End if

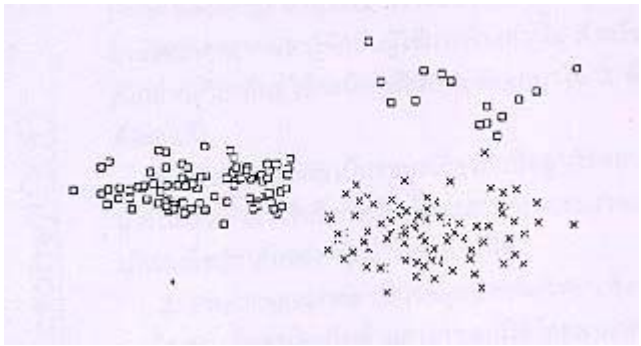
3. อัลกอลิทึม อื่นๆ

นอกจากอัลกอลิทึมข้างต้นแล้ว Data Mining ยังมี อัลกอลิทึมอื่นๆอีกจำนวนวนมาก เช่น

- Database Clustering หรือ Segmentation ได้แก่การแบ่งข้อมูลเป็นแบบกลุ่มๆ โดยที่ไม่รู้ล่วง

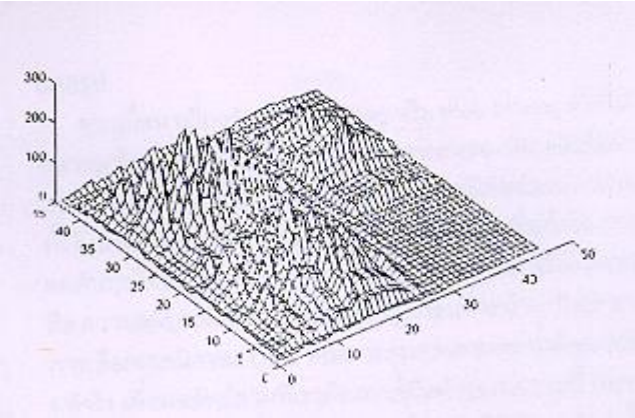
หน้าว่าจะมีทั้งหมดกี่กลุ่ม โดยการจัดกลุ่มข้อมูลดังกล่าวได้จากการพิจารณาคุณสมบัติในหลายๆมิติของข้อมูล ถ้ารายการในข้อมูลมีลักษณะ คล้ายคลึงกันเป็นกลุ่มเดียวกันได้ ก็จะรวมเข้าด้วยกัน รูปแสดงกลุ่มของข้อมูลที่พิจารณาจากคุณสมบัติเพียง 2 มิติ (ข้อมูลอาจจะมีหลายมิติก็ได ซึ่งมักจะไม่แสดงเป็น

รูปภาพได้)



รูป 2.14 แสดงข้อมูลใน 2 มิติ แสดงการแบ่งข้อมูลเป็น 3 กลุ่ม

- การตรวจหาค่าความเบี่ยงเบน (Deviation Detection) เป็นกรรมวิธีในการหาค่าที่แตกต่างไปจากมาตรฐาน หรือค่าที่คาดคิดไว้ว่า มีความแตกต่างมากน้อยเพียงใด โดยทั่วไปมักใช้วิธีการทาง สถิติหรือการแสดงให้เห็นภาพ(Visualization) ดังตัวอย่างในรูป สำหรับ อัลกอลิทึม นี้สามารถใช้ในการตรวจสอบ ลายเซ็นปลอม หรือ บัตรเครดิตปลอม รวมทั้งการตรวจหา จุดบกพร่อง ชิ้นงานในโรงงานอุตสาหกรรม



รูป 2.15 Visualization แสดงค่าเบี่ยงเบน

2.7 ข้อเสียของ Data Mining

จากที่เคยกล่าวไว้ข้างต้นว่า Data Mining เป็นเพียงเครื่องมือที่ใช้ในการวิเคราะห์เท่านั้น ไม่สามารถเข้าใจธุรกิจ หรือเข้าใจข้อมูลได้ดีเท่าคน ดังนั้นผู้ใช้ Data Mining จึงจำเป็นต้องมีความรู้ความเข้าใจในข้อมูลธุรกิจ เครื่องมือ และ อัลกอลิทึมได้เป็นอย่างดี

อย่างไรก็ตาม Data Mining จะช่วยหารูปแบบและความสัมพันธ์ของข้อมูล แต่ไม่ระบุว่าค่าของข้อมูลจริง หรือค่าที่แสดงความสัมพันธ์จริง เป็นเพียงแค่ทำนายเท่านั้น ผู้ใช้ต้องทำการตัดสินใจอีกครั้ง

เป็นความเข้าใจผิดที่ว่า Data Mining จะช่วยค้นหาคำตอบโดยที่ไม่ต้องถามคำถามใดๆ อันที่จริงแล้ว Data Mining ยังต้องการให้ผู้ใช้บอกรูปแบบของการค้นหาคำตอบด้วย

อนึ่ง Data Mining ไม่ได้เข้ามาแทนที่ความชำนาญของนักวิเคราะห์ แต่จะเป็นเครื่องมือที่จะช่วยให้นักวิเคราะห์ หรือนักบริหารในการต่อกรกับคู่แข่งได้เป็นอย่างดี

2.8 ประโยชน์ของการใช้ Data Mining

Data Mining ถูกนำมาใช้สนับสนุนการตัดสินใจ โดยกานสร้างมูลค่าเพิ่มให้กับข้อมูลที่มีอยู่ ประโยชน์ ที่แต่ละ องค์กรได้รับจากการใช้ Data Mining สรุปได้ ดังนี้

- การเอาชนะคู่แข่งขัน ลูกค้าที่ดีมักจะเป็นที่ชื่นชอบของบริษัทคู่แข่งเช่นกัน บริษัทเหล่านั้นจะพยายามแย่ง ส่วน แบ่งจาก Segment ที่สามารถสร้างผลกำไรให้กับบริษัทได้ และก็พยายามที่จะแย่งชิงส่วนแบ่งตลาดนั้นมา Data Mining สามารถนำมาใช้ประโยชน์ได้ทั้งการแย่งส่วนแบ่งตลาด และในแง่การป้องกันมิให้เกิดการเสีย ส่วนแบ่งตลาด
- ทำให้เกิดความรู้ที่สามารถนำมาใช้ หรือประกอบการตัดสินใจได้ เนื่องจาก Data Mining จะใช้เทคนิคที่ซับ ซ้อน และมีลักษณะเป็น Artificial Intelligence ในการสร้างโมเดลที่อิงกับข้อมูล ซึ่งรวบรวมจากแหล่งต่าง ๆ เช่น รายการทางธุรกิจ ข้อมูลประวัติลูกค้า และข้อมูลอื่น ๆ จากแหล่ภายนอก ความรู้ที่ได้จึงช่วยให้องค์กร สามารถคาดการณ์อนาคต และสามารถเจาะกลุ่มตลาดได้ถูกต้องมากขึ้น
- ใช้ในการหาข้อผิดพลาดของการปฏิบัติงาน หรือการให้บริการได้ (Fraud Detection)
- ช่วยประหยัดค่าใช้จ่าย โดยการทำให้ขั้นตอนการทำงานมีประสิทธิภาพมากขึ้น (Save Money)
- การกำหนดเป้าหมายกลุ่มลูกค้าได้อย่างมีประสิทธิภาพมากขึ้น ทำให้ยอดขายเพิ่มขึ้น เพิ่มจำนวนลูกค้า และ ลดโอกาสของความเสี่ยงต่าง ๆ

2.9 แนวโน้มและการประยุกต์ใช้งานData Mining (Data Mining Trend and Application)

เนื่องจากในปัจจุบันมีการนำหลักการและเทคนิคของ data mining มาใช้กันอย่างแพร่หลาย ดังนั้นจึงมีการค้นคว้าวิจัย และพัฒนาเพื่อประยุกต์ใช้กับงานในหลายๆ ด้าน โดยตัวอย่างการประยุกต์ใช้งานที่น่าสนใจในปัจจุบัน ได้แก่

- **การใช้งานด้านการแพทย์** (Biomedical and DNA Data Analysis)

ส่วนมากเป็นการนำไปในการวิเคราะห์รูปแบบการจัดเรียงตัวของหน่วยพันธุกรรมเพื่อหาสาเหตุของความผิดปกติที่ ทำให้เกิดโรค ความสัมพันธ์ของรูปแบบการจัดเรียงตัวของหน่วยพันธุกรรมกับระดับความรุนแรงของโรค รวมถึงการใช้ใน ด้าน การวินิจฉัยโรค การป้องกัน และการรักษาด้วย

- **การใช้งานเพื่อการวิเคราะห์ด้านการเงิน** (Financial Analysis)

เป็นงานที่เกี่ยวกับบริษัทเงินทุน หรือธนาคารต่างๆ เช่น การวิเคราะห์การให้สินเชื่อ การทำนายอัตราการจ่ายเงินกู้ การแบ่งกลุ่มลูกค้าเพื่อหาเป้าหมายทางการตลาด เป็นต้น

- **การใช้งานด้านการขาย** (Retail Industry)

เป็นงานที่มีการเก็บรวบรวมข้อมูลจำนวนมาก จึงมีการนำ Data Mining มาประยุกต์ใช้กับข้อมูลเหล่านี้ เพื่อหากลยุทธ์ ที่ทำให้เกิดการได้เปรียบคู่แข่งทางการค้า เช่น การหาลักษณะการซื้อของลูกค้า ความสัมพันธ์ของการซื้อกับช่วงเวลา
ความสัมพันธ์ระหว่างตัวสินค้า และ การวิเคราะห์ประสิทธิภาพของการโฆษณา เป็นต้น ซึ่งช่วยให้สามารถหาวิธีการตอบสนอง ความต้องการของลูกค้าได้มากที่สุด และอาจหมายถึงส่วนแบ่งทางการตลาดที่เพิ่มขึ้นนั่นเอง

- **การใช้งานด้านโทรคมนาคม (Telecommunication Industry)**

เพื่อสนับสนุนการให้บริการด้านการติดต่อสื่อสารของลูกค้า เช่น การวิเคราะห์ลักษณะการใช้บริการด้านการติดต่อสื่อสาร การหาความสัมพันธ์ของการใช้บริการกับช่วงเวลา หรือการตรวจจับรูปแบบที่ผิดปกติในระบบการติดต่อสื่อสาร เป็นต้น

จากลักษณะการนำไปใช้งานข้างต้นในปัจจุบันเนื่องจากเทคนิค หรือวิธีการที่นำมาใช้นั้นยังมีข้อจำกัดสำหรับการใช้ กับงาน หรือข้อมูลในบางประเภท ดังนั้นจึงมีแนวโน้มในการวิจัยพัฒนา และประยุกต์ใช้อย่างต่อเนื่อง เพื่อหาวิธีการที่เหมาะสมที่สุด ซึ่งแนวโน้มของการใช้งานที่ได้รับความสนใจในการศึกษาต่อไปในอนาคต ได้แก่

- **การประยุกต์ใช้งานแนวใหม่ๆ** (Application Exploration)

เป็นการนำเทคนิคของ Data Mining เข้ามาใช้กับงานในด้านอื่นๆ นอกเหนือจากการนิยมใช้กับงานเพื่อการ แข่งขันกันในเชิงธุรกิจดังเช่นในช่วงที่ผ่านมา เช่น การใช้งานในเชิงการแพทย์ การวิเคราะห์ทางด้านการเงินหรือการใช้งาน ในด้านโทรคมนาคม เป็นต้น โดยจะมีการพัฒนาเพื่อเป็นระบบที่ใช้งานเฉพาะทางเพิ่มมากขึ้น

- **การพัฒนาวิธีการเพื่อใช้กับฐานข้อมูลขนาดใหญ่** (Scalable Data Mining)

เป็นการพัฒนาเพื่อให้ระบบสามารถใช้งานกับฐานข้อมูลขนาดใหญ่ได้อย่างมีประสิทธิภาพ ซึ่งวิธีการหนึ่งที่ได้รับ การพัฒนา คือการทำ mining ในลักษณะที่มีเงื่อนไข (Constraint-Based Mining) โดยเปิดโอกาสให้ผู้ใช้สามารถ ใส่เงื่อนไขเฉพาะบางอย่างให้กับระบบ เพื่อเป็นแนวทางให้ระบบสามารถค้นหาคำตอบได้ใกล้เคียงกับวัตถุประสงค์ของผู้ใช้ มากขึ้น

- **การรวมงานของ data mining เข้าเป็นส่วนหนึ่งของระบบฐานข้อมูล ระบบคลังข้อมูล รวมถึงระบบฐานข้อมูลบน web** (Integration of Data Mining with Database System ,Data Warehouse System ,and Web Database System)

เป็นการพัฒนาให้ Data Mining กลายเป็นส่วนหนึ่งของระบบฐานข้อมูล ระบบคลังข้อมูล รวมทั้งระบบฐานข้อมูลบน web ด้วย เนื่องจากเป็นระบบหลักที่ใช้ในการเก็บข้อมูลอยู่แล้ว ทำให้งานวิเคราะห์ข้อมูลในระบบนั้นสามารถทำงานร่วมกับระบบจัดเก็บข้อมูลได้ง่าย และมีประสิทธิภาพเพิ่มมากขึ้น

- **การสร้างมาตรฐานให้กับภาษาในการทำ Data Mining** (Standardization of Data Mining Language)

เป็นการพัฒนาให้เกิดภาษาเฉพาะสำหรับกระบวนการทำ Data Mining เพื่อให้เกิดความสะดวกและง่ายต่อการประยุกต์ใช้ รวมทั้งเป็นการเพิ่มความสามารถในการติดต่อกันระหว่างระบบด้วย

- **การสร้าง Data Mining เพื่อให้หาความหมายข้อมูลได้ง่ายขึ้น** (Visual Data Mining)

เนื่องจากการใช้งานในลักษณะนี้ เป็นวิธีการสำคัญที่มีประสิทธิภาพในการค้นหาลักษณะแฝงบางประการจากข้อมูลจำนวณมากๆ ดังนั้นการศึกษา และการพัฒนาในแนวทางนี้จะเป็นการหาเทคนิคใหม่ๆ เพื่อความสะดวกต่อการใช้งาน และง่ายต่อการเข้าใจ สามารถใช้ในการวิเคราะห์ข้อมูลได้อย่างมีประสิทธิภาพ

- **การหาวิธีการใช้งานกับข้อมูลที่มีความซับซ้อน** (New Methods for Mining Complex Types of Data)

เช่นข้อมูลลักษณะเชิงภูมิศาสตร์ มัลติมีเดีย หรือข้อมูลในลักษณะตัวอักษร เป็นต้น ซึ่งมีการใช้งานค่อนข้างมาก ในปัจจุบัน ดังนั้นจึงมีการค้นหาวิธีใหม่ๆ รวมทั้งมีการรวมวิธีการที่มีอยู่เพื่อประยุกต์ใช้กับการวิเคราะห์ข้อมูลประเภทนี้ได้ อย่างเหมาะสม

การใช้ Mining กับข้อมูลบน Web (Web Mining)

เป็นการประยุกต์ใช้งานกับข้อมูลบนอินเตอร์เน็ต เนื่องจากอินเตอร์เน็ตเป็นแหล่งข้อมูลขนาดใหญ่ และมีผู้ใช้งาน จำนวนมาก ดังนั้นจึงมีการนำข้อมูลต่างๆบน web ซึ่งได้แก่ web content , web log รวมถึงการให้บริการต่างๆบนอินเตอร์เน็ตมาใช้ทำ mining เพื่อหาแนวทางในการตอบสนองผู้ใช้งานให้ได้มากที่สุด

- **การรักษาความปลอดภัยของข้อมูล** (Information Security in Data Mining)

เป็นการพัฒนาวิธีการเพื่อสร้างความเชื่อมั่นในเรื่องความปลอดภัยของข้อมูลในขณะที่มีการพัฒนาวิธีการเข้าถึงข้อมูล และการ Mining ให้สะดวกต่อการใช้งานมากขึ้น