บทที่ 2
ความรู้พื้นฐานเกี่ยวกับดาต้าไมนิ่ง

2.1 ความหมายของดาต้าไมนิ่ง

ดาต้าไมนิ่ง(Data Mining) คือการค้นหาความสัมพันธ์และรูปแบบทั้งหมด ซึ่งมีอยู่จริงในฐานข้อมูล แต่ได้ถูกซ่อนไว้ภายในข้อมูลจำนวนมาก ดาต้าไมนิ่งจะทำการสำรวจละวิเคราะห์อย่างอัตโนมัติหรือกึ่งอัตโนมัติ ในปริมาณข้อมูลจำนวนมากให้อยู่ในรูปแบบที่เต็มไปด้วยความหมายและอยู่ในรูปของกฎ โดยความสัมพันธ์เหล่านี้แสดงให้เห็นถึงความรู้ต่างๆที่มีประโยชน์ในฐานข้อมูล

ในปัจจุบันองค์กรส่วนใหญ่จะเผชิญกับปัญหาของ “ข้อมูลดิบจำนวนมากแต่ข้อมูลที่ประยุกต์ใช้ได้นั้นมีน้อย” ดาต้าไมนิ่งจึงเป็นสาขาที่คาดว่าจะเป็นที่รู้จักและนำมาใช้ประยุกต์ใช้อย่างแพร่หลาย เนื่องจากดาต้าไมนิ่งสามารถดึคามรู้ออกมาจากข้อมูลจำนวนมากที่ถูกเก็บสะสมไว้

ในโลกของธุรกิจปัจจุบันบริษัทต่างๆจะพยายามหาเทคนิคที่สามารถนำความสำเร็จมาสู่บริษัท เช่น ในโลกธุรกิจขนาดย่อมจะสร้างความสัมพันธ์กับลูกค้า โดยสังเกตจากความต้องการ ความชอบและความสนใจของลูกค้า และอาจมีการเรียนรู้ได้จากผลสะท้อนในอดีตว่าจะทำอย่างไรให้การบริการลูกค้ามีประสิทธิภาพดีขึ้นในอนาคต หรือ บริษัทที่เป็นผู้ออกบัตรเครดิตและธนาคารต่างๆ จะมีขบวนการที่ใช้ดาต้าไมนิ่งให้เป็นประโยชน์ ในการตัดสินใจว่าลูกค้ากลุ่มใดเป็นกลุ่มที่ดี , ทำความเข้าใจลูกค้า , ช่วยในการแยกประเภทของลูกค้าและจะทำนายกลุ่มของประชากรที่คาดว่าจะมาเป็นลูกค้าในอนาคต เป็นต้น อย่างไรก็ตามการเรียนรู้นั้นต้องมากกว่าการเก็บสะสมข้อมูลอย่างตรงไปตรงมา ซึ่งจะทำให้การทำงานไม่เป็นประสิทธิภาพ

2.2 วัฏจักรขั้นตอนการทำงานของดาต้าไมนิ่ง

วัฏจักรขั้นตอนการทำงานของดาต้าไมนิ่งประกอบไปด้วย 4 ขั้นตอนหลักๆ ดังนี้

2.2.1 การระบุโอกาสทางธุรกิจหรือการระบุปัญหาที่เกิดขึ้นกับธุรกิจ

เป็นการระบุขอบเขตของข้อมูลที่จะนำมาทำการวิเคราะห์เพื่อหาความได้เปรียบทางการตลาดหรือเพื่อนำมาทำการแก้ไขปัญหา

2.2.2 ส่วนของดาต้าไมนิ่ง

เป็นการนำเทคนิคของดาต้าไมนิ่ง ไปใช้ถ่ายทอดหรือทำการเปลี่ยนแปลงข้อมูลดิบให้อยู่ในรูปของข้อมูลที่จะนำไปใช้ได้จริงในทางธุรกิจ

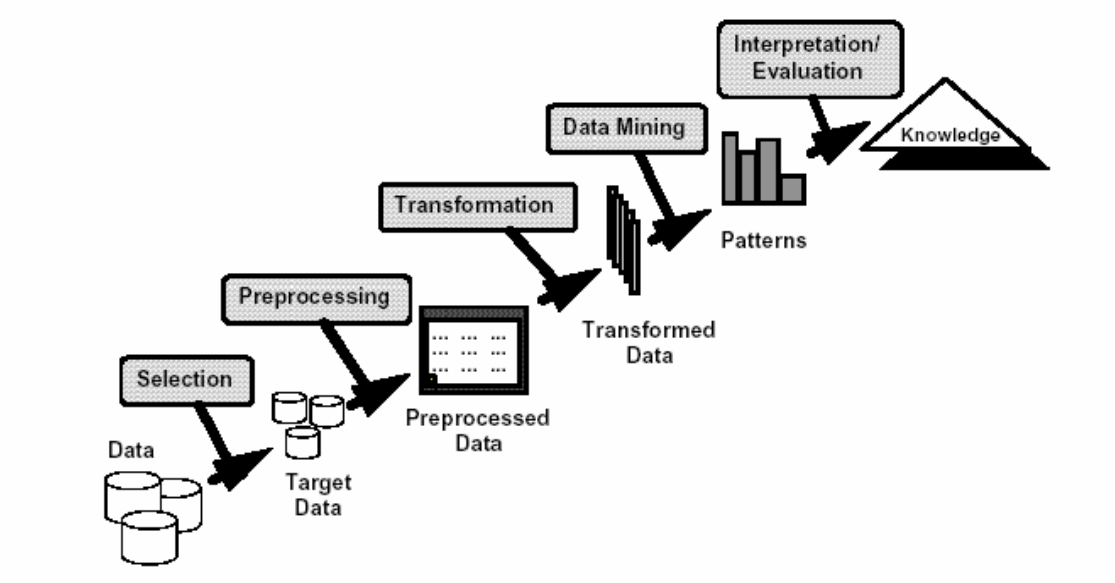
2.2.3 การปฏิบัติตามข้อมูล

คือการนำเอาข้อมูลที่เป็นผลลัพธ์ของส่วนดาต้าไมนิ่งมาลองปฏิบัติจริงกับธุรกิจ

2.2.4 การวัดประสิทธิภาพจากผลลัพธ์

การวัดประสิทธิภาพของเทคนิคของดาต้าไมนิ่งที่จะนำมาใช้จากผลลัพธ์ ซึ่งสามารถตรวจสอบได้หลายทาง เช่น วัดจากส่วนแบ่งของตลาด , วัดจากปริมาณลูกค้า หรือ วัดจากกำไรสุทธิ เป็นต้น

จากทั้ง 4 ขั้นตอนที่กล่าวมาข้างต้นคือการนำเอาดาต้าไมนิ่งไปใช้กับระบบทางธุรกิจ โดยแต่ละขั้นตอนจะพึ่งพาอาศัยกัน ผลลัพธ์จากขั้นตอนหนึ่งจะกลายมาเป็นอินพุทจากอีกขั้นตอนต่อไป ซึ่งดาต้าไมนิ่งจะเปลี่ยนข้อมูลดิบให้เป็นข้อมูลประยุกต์ ดังนั้นการระบุแหล่งข้อมูลที่ถูกต้องจึงเป็นสิ่งที่สำคัญอย่างยิ่งต่อผลลัพธ์ที่ได้จากการวิเคราะห์



2.3 งานของดาต้าไมนิ่ง (Task of data mining)

ในทางปฏิบัติจริงดาต้าไมนิ่งจะประสบความสำเร็จกับงานบางกลุ่มเท่านั้น และต้องอยู่ภายใต้ภาวะที่จัด ปัญหาเหมาะสมกับการใช้เทคนิคดาต้าไมนิ่งจะเป็นปัญหาที่ต้องใช้เหตุผลในการแก้, ปัญหาที่เกียวข้องกับเศรษฐศาสตร์และการเงิน ซึ่งจะสามารถจัดรูปแบบของธุรกิจให้อยู่ในรูปแบบของงานทั้ง 6 งานได้ ดังนี้

1. การจัดหมวดหมู่ (Classification)
2. การประเมินค่า (Estimation)
3. การทำนายล่วงหน้า (Prediction)
4. การจัดกลุ่มโดยอาศัยความใกล้ชิด (Affinity Group)
5. การรวมตัว (Clustering)
6. การบรรยาย (Description)

ไม่มีเทคนิคหรือเครื่องมือเพียงชนิดเดียวของดาต้าไมนิ่งที่เหมาะสมกับงานทุกชนิด ง่านในแต่ล่ะชนิดก็จะมีเทคนิคของดาต้าไมนิ่งที่แตกต่างกันไปขึ้นอยู่กับชนิดของงาน

2.3.1 การจัดหมวดหมู่

การจัดหมวดหมู่ถือว่าเป็นงานธรรมดาทั่วไปของดาต้าไมนิ่ง เพราะการทำความเข้าใจและการติดต่อสื่อสารต่างๆ ก็เกี่ยวข้องกับการแบ่งเป็นหมวดหมู่ , การจัดแยกประเภท และการแบ่งแยกชนิด

โดยการจัดหมวดหมู่ประกอบด้วยการสำรวจจุดเด่นของวัตถุที่ปรากฏออกมา และทำการกำหนดจุดเด่นนั้นๆ เป็นตัวที่ใช้แบ่งหมวดหมู่ งานในการแบ่งหมวดหมู่คือการบ่งบอกลักษณะ โดยการอธิบายจุดเด่นที่เป็นที่รู้จักดีในหมวดหมู่นั้น และเทรนนิ่งเซต (Training Set) ของตัวอย่างในแต่ละหมวดหมู่ ซึ่งมีภาระหน้าที่ในการสร้างโมเดลของบางชนิดที่ไม่สามารถจะจัดหมวดหมู่ของข้อมูลได้ ให้สามารถจัดเป็นหมวดหมู่ได้ ตัวอย่างของการจัดหมวดหมู่ เช่น การจัดหมวดหมู่ของผู้ยื่นขอเครดิต เป็นระดับต่ำ, ระดับกลาง และระดับสูง ของความเสี่ยงที่จะได้รับ เป็นต้น

2.3.2 การประเมินค่า

การประเมินค่าทางธุรกิจอย่างต่อเนื่องจะก่อให้เกิดผลลัพธ์ที่มีประโยชน์กับธุรกิจ การป้อนข้อมูลที่เรามีอยู่เข้าไป เพื่อใช้ในการประเมินสิ่งต่างๆ ที่จะก่อให้เกิดประโยชน์ หรือสำหรับตัวแปรที่เราไม่รู้ค่าแน่นอนเช่น รายได้จากการค้า ,จุดสูงสุดทางธุรกิจ หรือดุลยภาพของบัตรเครดิต ในทางปฏิบัติการประเมินค่าจะถูกใช้ในการทำงานการจัดหมวดหมู่ ตัวอย่างของการประเมินค่าเช่น การประเมินรายได้รวมของครอบครัว หรือการประเมินจำนวนบุตรในครอบครัว

2.3.3 การทำนายล่วงหน้า

การทำนายล่วงหน้าก็เป็นงานที่มีลักษณะคล้ายกับการจัดหมวดหมู่หรือการประเมินค่า ยกเว้นเพียงแต่จะใช้สถิติการบันทึกของการจัดหมวดหมู่ในการทำนายอนาคตของพฤติกรรมหรือการประเมินค่าที่จะเกิดขึ้นในอนาคต ตัวอย่างของงานการทำนายล่วงหน้า เช่น การทำนายการเปลี่ยนแปลงพฤติกรรมของตลาด หรือการทำนายจำนวนลูกค้าที่จะออกจากธุรกิจของเราใน 6 เดือนข้างหน้า เป็นต้น

2.3.4 การจัดกลุ่มโดยอาศัยความใกล้ชิดกันหรือการวิเคราะห์ของตลาด

งานในการจัดกลุ่มหรือการวิเคราะห์ตลาด คือการตัดสินใจรวมสิ่งที่สามารถไปด้วยกันเข้าไว้ในกลุ่มเดียวกัน ตัวอย่างของการจัดกลุ่มโดยอาศัยความใกล้ชิดกันหรือการวิเคราะห์ของตลาด เช่น การตัดสินใจว่าสิ่งใดบ้างที่จะไปอยู่ด้วยกันอย่างสม่ำเสมอในรถเข็นในซุปเปอร์มาร์เกต

2.3.5 การรวมตัว

การรวมตัวคืองานที่ทำการรวมส่วนต่างๆ ในแต่ละส่วนที่ต่างชนิดกันให้อยู่ในรวมกันเป็นกลุ่มย่อย หรือคลัสเตอร์ โดยในแต่ละคลัสเตอร์อาจจะประกอบด้วยส่วนต่างๆที่ต่างชนิดกัน ซึ่งความแตกต่างของการรวมตัวจากการจัดหมวดหมู่คือ การรวมตัวจะไม่พึ่งพาอาศัยการกำหนดหมวดหมู่ล่วงหน้า และไม่ใช้ตัวอย่าง ข้อมูลจะรวมตัวกันบนพื้นฐานของความคล้ายในตัวเอง

2.3.6 การบรรยาย

ในบางครั้งวัตถุประสงค์ของดาต้าไมนิ่ง คือต้องการอธิบายความสับสนของฐานข้อมูลในทางที่จะเพิ่มความเข้าใจในส่วนของประชากร , ผลิตภัณฑ์ , หรือขบวนการให้มากขึ้น

เทคนิคดาต้าไมนิ่งส่วนใหญ่ต้องการเทรนนิ่งข้อมูลจำนวนมากที่ประกอบด้วยหลายๆ ตัวอย่าง เพื่อจะสร้างกฎที่ใช้ในการจัดหมวดหมู่ , กฎของความสัมพันธ์, คลัสเตอร์ ,การทำนายล่วงหน้า ดังนั้นชุดของข้อมูลขนาดเล็กจะนำไปสู่ความไม่น่าไว้วางใจของผลสรุปทีได้

ไม่มีเทคนิคใดเลยที่จะสามารถแก้ปัญหาของดาต้าไมนิ่งได้ทุกปัญหา ดังน้นความหลากหลายของเทคนิคจึงเป็นสิ่งที่จำเป็นในการไปสู่วิธีการแก้ปัญหาของดาต้าไมนิ่งได้ดีที่สุด

2.4 เทคนิคของดาต้าไมนิ่ง

การแก้ปัญหาของงานชนิดต่างๆ โดยใช้วิธีดาต้าไมนิ่ง ในแต่ละงานก็จะมีเทคนิคของดาต้าไมนิ่งที่จะนำมาใช้ได้อย่างเหมาะสม โดยเทคนิคของดาต้าไมนิ่งนั้นมีมากมาย ซึ่งจะขอยกตัวอย่างของเทคนิคที่ถูกใช้กันค่อนข้างแพร่หลาย

2.4.1 นิวรอนเน็ตเวิร์ก

นิวรอนเน็ตเวิร์ก คือระบบทีมีการประมวลผลข้อมูลซึ่งรวมคุณสมบัติของไบโอลอจิกคอล นิวรอนเน็ตเวิร์ก ถูกพัฒนาขึ้นโดยโมเดลทางคณิตศาสตร์ของกระบวนการเรียนรู้ของมนุษย์ และจะเรียนรู้จากชุดข้อมูลของชุดความรู้ (Training Set)

นิวรอนเน็ตเวิร์ก ประกอบด้วยหน่วยความจำจำนวนมากเรียกว่า นิวรอน , เซลหรือโหนด แต่ละนิวรอนต่อกันโดยคอนเน็กชั้นลิ่ง (Connection Link) ที่ค่าน้ำหนัก (Weight) ของมันอยู่ โดยค่าน้ำหนักจะแสดงรายละเอียดที่เน็ตเวิร์กใช้ในการแก้ปัญหา โดยนิวรอนเนตเวิร์กถูกใช้ในการแก้ปัญหาอย่างกว้างขวาง เช่น การเก็บและการเรียกข้อมูล, การแยกประเภทของข้อมูล, การเปลี่ยนจากรูปแบบของอินพุทให้อยู่ในรูปแบบของเอาท์พุท, ความสามารถในการตรวจสอบรูปแบบของข้อมูลที่คล้ายคลึงกับความคิดของมนุษย์ เป็นต้น ถึงแม้ว่านิวรอนเน็ตเวิร์ก สามารถนำไปประยุกต์ใช้กับงานหลายๆ ชนิดได้อย่างมีประสิทธิภาพ แต่นิวรอนเน็ตเวิร์ก ก็ยังมีข้อเสียอยู่บ้าง ดังนี้

1. นิวรอนเน็ตเวิร์กเป็นวิธีที่ยากต่อการทำความเข้าใจในโมเดลที่ถูกผลิตออกมา
2. นิวรอนเน็ตเวิร์กมีคุณสมบัติที่ไวต่อรูปแบบของอินพุท ถ้าเราแทนข้อมูลด้วยรูปแบบที่แตกต่างกันก็จะสามารถผลิตผลลัพธ์ที่แตกต่างกันออกมา ดังนั้นการกำหนดค่าเริ่มต้นให้กับข้อมูลจึงเป็นส่วนที่มีความสำคัญส่วนหนึ่ง

2.4.2 จีเนติก อัลกอริทึ่ม (Genetic Algorithms : GA)

จีเนติก อัลกอริทึ่ม เป็นทฤษฎีที่จำลองกระบวนการวิวัฒนาการทางธรรมชาติ คือการคัดเลือกทางธรรมชาติ ละอาศัยพื้นฐานความคิดทางพันธุกรรมในการถ่ายทอดลักษณะต่างๆ ไปยังรุ่นถัดไป ที่สามารถนำมาพัฒนาใช้ในการหาคำตอบที่เหมาะสมที่สุดของแต่ละปัญหา

จีเนติก อัลกอริทึ่มเป็นวิธีการหาคำตอบโดยการพิจารณา และดำเนินการจากกลุ่มของคำตอบของปัญหาที่ถูกสร้างขึ้นมาโดยการเข้ารหัส คือการแปลงค่าตัวแปรหรือพารามิเตอร์ของปัญหาให้อยู่ในรูปโครงสร้างของโครโมโซมที่กำหนด เพื่อคัดเลือกโครโมโซมคำตอบที่เหมาะสมสำหรับสร้างวิวัฒนาการของคำตอบให้ดีขึ้นตามกระบวนการทางพันธุศาสตร์ โดยการแลกเปลี่ยนค่าพารามิเตอร์ต่างๆ ระหว่างโครโมโซมที่ถูกคัดเลือกอันจะทำให้คำตอบของปัญหาถูกปรับปรุงให้ดีขึ้น

จีเนติก อัลกอริทึ่มใช้กระบวนการหลักๆ 3 กระบวนการในการหาคำตอบที่ใกล้เคียงหรือดีที่สุดของปัญหาดังนี้

1. การคัดเลือก (Selection)
2. การครอสโอเวอร์ (Crossover)
3. การมิวเตชั่น (Mutation)

ถึงแม้ว่าในปัจจุบันจีเนติก อัลกอริทึ่ม ยังเป็นวิธีการที่ไม่ได้แพร่หลายนัก แต่สาขาวิชาทางด้านจีเนติก อัลกอริทึ่มก็นับว่าเป็นอีกสาขาวิชาหนึ่งที่สนใจและน่าจะเป็นวิธีที่ได้รับความนิยมในอีกไม่กี่ปีข้างหน้า เนื่องมาจากสามารถนำมาประยุกต์ใช้ได้กับหลายๆ ปัญหารวมทั้งปัญหาทางดาต้าไมนิ่งอีกด้วย

2.5 บทสรุป

ดาต้าไมนิ่งคือการค้นหาความสัมพันธ์และรูปแบบทั้งหมด ซึ่งมีอยู่จริงในฐานข้อมูล แต่ได้ถกซ่อนไว้ภายในข้อมูลจำนวนมากอย่างอัตโนมัติหรือกึ่งอัตโนมัติ โดยดาต้าไมนิ่งจะเหมาะกับการแก้ปัญหาบางชนิดเท่านั้น เช่น ปัญหาที่ต้องใช้เหตุผลในการแก้ หรือปัญหาที่เกี่ยวข้องกับเศรษฐศาสตร์และการเงินเป็นต้น

ดาต้าไมนิ่งมีเทคนิคต่างๆที่ใช้ในการแก้ปัญหาอยู่หลายเทคนิค ซึ่งจะไม่มีเทคนิคใดเลยที่สามารถแก้ปัญหาของดาต้าไมนิ่งได้ทุกปัญหา ดังนั้นความหลากหลายของเทคนิคเป็นสิ่งที่จำเป็นที่จะนำไปสู่วิธีการแก้ปัญหาของดาต้าไมนิ่งได้ดีที่สุด