บทที่ 1 บทนำ

หัวข้อ

1.1 การทำเหมืองข้อมูลคืออะไร?

1.2 ความท้าทายที่เป็นแรงผลักดัน

1.3 ต้นกำเนิดของการทำเหมืองข้อมูล

1.4 ประเภทของการทำเหมืองข้อมูล

ความก้าวหน้าอย่างรวดเร็วของเทคโนโลยีการเก็บรวบรวมข้อมูล และเทคโนโลยีอุปกรณ์จัดเก็บข้อมูลทำให้องค์กรต่าง ๆ สามารถเก็บรวบรวมข้อมูลปริมาณมากมายได้ อย่างไรก็ตามการสกัดเอาสารสนเทศที่มีประโยชน์มาใช้งานนั้นเป็นสิ่งที่ท้าทายมาก เครื่องมือและเทคนิคการวิเคราะห์ข้อมูลแบบเดิมมักไม่สามารถนำมาใช้จัดการกับข้อมูลที่มีขนาดใหญ่มากได้ ในบางครั้งคุณสมบัติของข้อมูลที่แตกต่างจากไปจากรูปแบบที่เคยมีมาก่อนหน้า ก็ทำให้วิธีการวิเคราะห์ข้อมูลแบบดั้งเดิมไม่สามารถนำมาใช้กับข้อมูลเหล่านี้ได้แม้ว่ามันจะมีขนาดค่อนข้างเล็กก็ตาม ในบางสถาณการณ์โจทย์ของการวิเคราะห์ข้อมูลก็ไม่สามารถหาคำตอบได้ด้วยเทคนิคการวิเคราะห์ข้อมูลที่มีอยู่ ดังนั้นจึงจำเป็นต้องมีการพัฒนาเทคนิคและวิธีการใหม่ๆ เพื่อจัดการกับความท้าทายใหม่เหล่านี้

การทำเหมืองข้อมูล (data mining) คือเทคโนโลยีที่ผสานวิธีการวิเคราะห์ข้อมูลแบบดั้งเดิมเข้ากับอัลกอริทึมขั้นสูงสำหรับการประมวลผลข้อมูลปริมาณมหาศาล การทำเหมืองข้อมูลยังเปิดโอกาสให้เกิดการสำรวจและวิเคราะห์ข้อมูลชนิดใหม่และการวิเคราะห์ข้อมูลชนิดเดิมด้วยมุมมองและวิธีการแบบใหม่ ตัวอย่างการประยุกต์เทคนิคการทำเหมืองข้อมูล ในเชิงธุรกิจ การแพทย์ วิทยาศาสตร์ และวิศวกรรมศาสตร์ ได้แก่

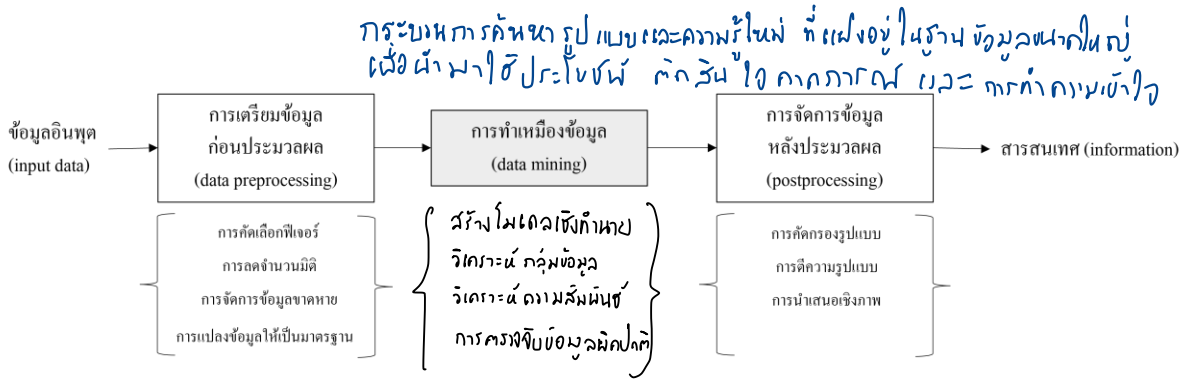
- การเก็บรวบรวมข้อมูล ณ จุดขาย (Point-of-Sale) ด้วยเครื่องสแกนบาร์โค้ด ตัวระบุคลื่นความถี่วิทยุ (radio frequency identification : RFID) และสมาร์ทการ์ด ทำให้ธุรกิจค้าปลีกมีข้อมูลธุรกรรมการซื้อสินค้าของลูกค้าปริมาณมหาศาล ธุรกิจค้าปลีกสามารถนำข้อมูลนี้ และข้อมูลจากแหล่งอื่น เช่น บันทึกการบริการลูกค้าจากศูนย์ call center, และบันทึกธุรกรรมบนเว็บไซต์พาณิชย์อิเล็กทรอนิกส์ (e-commerce Web sites) มาวิเคราะห์ด้วยเทคนิคการทำเหมืองข้อมูล เพื่อให้ธุรกิจเข้าใจความต้องการของลูกค้า และตัดสินใจเชิงธุรกิจได้ดียิ่งขึ้น เช่น การทำโปรไฟล์ลูกค้า การทำการตลาดแบบพุ่งเป้า การบริหารจัดการสายงาน การจัดวางสินค้าในร้านค้า และการตรวจจับการฉ้อโกง เป็นต้น นอกจากนี้ การทำเหมืองข้อมูลยังสามารถนำไปใช้ตอบคำถามเชิงธุรกิจ เช่น สินค้าชนิดใดที่ลูกค้ามักจะซื้อคู่กัน ลูกค้ากลุ่มใดที่สนใจและมีโอกาสจะซื้อผลิตภัณฑ์ชนิดใหม่สูง
- องค์การนาซ่าใช้ดาวเทียมในการสำรวจและเก็บข้อมูลเกี่ยวกับพื้นผิว มหาสมุทร และชั้นบรรยากาศของโลกอยู่ตลอดเวลา ซึ่งจำเป็นต้องใช้เทคนิคการวิเคราะห์ข้อมูลแบบใหม่ของการทำเหมืองข้อมูลในการวิเคราะห์เพื่อค้นหาคำตอบสำหรับคำถาม เช่น ความสัมพันธ์ระหว่างความถี่และความรุนแรงของภาวะโลกร้อน เช่น ภัยแล้ง น้ำท่วม กับการรุกล้ำทำลายระบบนิเวศน์ เป็นอย่างไร? อุณหภูมิของพื้นผิวมหาสมุทรมีผลกระทบอย่างไรต่ออุณหภูมิบนพื้นผิวดิน?

- นักวิจัยด้านชีวโมเลกุล ใช้เทคนิคการทำเหมืองข้อมูลวิเคราะห์ข้อมูลจีโนม ไมโครอาร์เรย์ ที่มีความซับซ้อนและขนาดใหญ่มาก เพื่อทำความเข้าใจโครงสร้างและฟังก์ชันการทำงานของยีน

1.1 การทำเหมืองข้อมูลคืออะไร?

การทำเหมืองข้อมูล (data mining) คือกระบวนการค้นหารูปแบบและความรู้ใหม่ ที่แฝงอยู่ในฐานข้อมูลขนาดใหญ่ เพื่อนำมาใช้ประโยชน์ในการตัดสินใจ การคาดการณ์ และการทำความเข้าใจปรากฏการณ์ต่าง ๆ

การทำเหมืองข้อมูล เป็นขั้นตอนหนึ่งของกระบวนการค้นหาความรู้จากฐานข้อมูล (knowledge discovery in databases : KDD) ซึ่งแปลงข้อมูลดิบไปเป็นความรู้หรือสารสนเทศที่สามารถนำไปใช้ประโยชน์ได้ ดังรูปที่ 1



รูปที่ 1 ขั้นตอนหลักของกระบวนการค้นหาความรู้จากฐานข้อมูล *(KDD)*

ข้อมูลอินพุตของกระบวนการ KDD อาจถูกจัดเก็บอยู่ในฐานข้อมูลกลางขององค์กร หรืออาจกระจายอยู่หลายแห่ง รูปแบบของข้อมูลอินพุตไม่จำกัดเฉพาะตารางในฐานข้อมูล แต่สามารถมีรูปแบบได้หลากหลาย เช่น plain text file, spreadsheet, log files, JSON files เป็นต้น

วัตถุประสงค์ของการเตรียมข้อมูลก่อนประมวลผล (preprocessing) คือการแปลงข้อมูลอินพุตให้อยู่ในรูปแบบที่เหมาะสมกับวิเคราะห์ในขั้นตอนถัดไป ขั้นตอนการเตรียมข้อมูลก่อนประมวลผล ได้แก่ การผสานรวมข้อมูลจากแหล่งข้อมูลต่าง ๆ เข้าด้วยกัน (data integration), การทำความสะอาดข้อมูล (data cleansing) เช่น การกำจัดข้อมูลซ้ำซ้อน (duplicate removal) การจัดการกับข้อมูลที่ขาดหาย (missing data handling), การคัดเลือกฟีเจอร์ (feature selection), การลดจำนวนมิติของข้อมูล (dimensionality reduction), และการแปลงข้อมูลให้เป็นมาตรฐาน (data standardization) ในทางปฏิบัติการประมวลผลข้อมูลก่อนการวิเคราะห์มักเป็นขั้นตอนที่ใช้แรงงานและเวลามากที่สุดในกระบวนการ KDD

เมื่อได้ผลลัพธ์จากการทำเหมืองข้อมูลแล้ว ก่อนนำผลที่ได้ไปใช้งานต้องมีการจัดการข้อมูลหลังประมวลผล (postprocessing) เพื่อตรวจสอบความถูกต้องและคัดเลือกผลลัพธ์ที่มีประโยชน์นำไปใช้งานจริงได้ โดยเทคนิคที่ใช้ในขั้นตอนนี้ ได้แก่ การตรวจสอบนัยสำคัญด้วยวิธีการทางสถิติ การนำเสนอข้อมูลเชิงภาพ (data visualization) การคัดกรองรูปแบบ (pattern filtering) และการตีความหมาย (pattern interpretation) ผลลัพธ์ที่ผ่านการคัดกรองของขั้นตอน postprocessing คือเอาต์พุตของกระบวนการ KDD เรียกว่า สารสนเทศ (information)

1.2 ความท้าทายที่เป็นแรงผลักดัน

ความท้าทายสำคัญที่เป็นแรงผลักดันให้เกิดการพัฒนาเทคนิคการทำเหมืองข้อมูล ได้แก่

การเพิ่มขึ้นของขนาดข้อมูลอินพุต (Scalability) ความก้าวหน้าของเทคโนโลยีการสร้างและเก็บข้อมูล ทำให้ขนาดข้อมูลในระดับกิกะไบต์ (gigabytes) เทอระไบต์ (terabytes) หรือแม้แต่เพตะไบต์ (petabytes) กลายเป็นสิ่งที่พบได้ทั่วไปในการทำเหมืองข้อมูล อัลกอริทึมการทำเหมืองข้อมูลต้องสามารถจัดการกับข้อมูลอินพุตได้โดยใช้เวลาและทรัพยากรในการประมวลผลที่มีอยู่อย่างจำกัดได้อย่างมีประสิทธิภาพ แม้ว่าข้อมูลจะมีขนาดใหญ่มาก (เรียกคุณสมบัตินี้ว่า scalability) ตัวอย่างของเทคนิคที่ถูกนำมาใช้เพื่อจัดการกับข้อมูลขนาดใหญ่มาก เช่น การสุ่มตัวอย่าง (sampling), อัลกอริทึมแบบขนานและแบบกระจาย (parallel and distributed algorithms), อัลกอริทึมแบบ out-of-core เป็นต้น

ข้อมูลที่มีมิติสูง (High Dimensionality) ในปัจจุบันข้อมูลที่มีแอทริบิวต์ (จำนวนมิติ) ในหลักร้อยหรือหลักพันแอทริบิวต์สามารถพบได้แพร่หลาย เช่น ข้อมูลไมโครอาร์เรย์ในสาขาวิชาชีวสารสนเทศ ข้อมูลเชิงเวลาและเชิงพื้นที่ที่ได้จากเซ็นเซอร์ชนิดต่าง ๆ เป็นต้น ความซับซ้อนของอัลกอริทึมการทำเหมืองข้อมูล มักจะแปรผันตามกับจำนวนมิติของข้อมูล ดังนั้นการพัฒนาอัลกอริทึมที่สามารถจัดการกับข้อมูลที่มีจำนวนมิติสูงจึงมีความสำคัญและจำเป็นมากขึ้นเรื่อย ๆ

ข้อมูลที่มีความหลากหลายและความซับซ้อน (Heterogeneous and Complex Data) วิธีการวิเคราะห์ข้อมูลแบบดั้งเดิมมักสามารถจัดการกับข้อมูลที่ประกอบด้วยแอทริบิวต์ชนิดเดียวกันทั้งหมด แต่ปัจจุบันข้อมูลมักประกอบด้วยแอทริบิวต์หลากหลายชนิด และมีความซับซ้อน ตัวอย่างเช่น ข้อมูลโครงสร้างการเชื่อมโยงระหว่างเว็บเพจ ข้อมูลอนุกรมเวลา ข้อมูลเท็กซ์ ข้อมูลกราฟ เป็นต้น

ความเป็นเจ้าของข้อมูลและข้อมูลที่กระจัดกระจายอยู่หลายแหล่ง (Data Ownership and Distribution) ข้อมูลสำหรับการทำเหมืองข้อมูลอาจถูกจัดเก็บอยู่ในแหล่งข้อมูลแตกต่างกันหลายที่ อาจมีทั้งข้อมูลที่เป็นขององค์กรเอง และข้อมูลที่เป็นของหน่วยงานหรือบุคคลอื่น อัลกอริทึมการทำเหมืองข้อมูลแบบกระจาย (distributed data mining algorithms) เป็นอัลกอริทึมที่สามารถประมวลผลข้อมูลที่กระจัดกระจายอยู่หลายแหล่งและไม่ได้เป็นของหน่วยงานใดหน่วยงานหนึ่ง โดยเฉพาะ ความท้าท้ายที่สำคัญของการทำเหมืองข้อมูลแบบกระจาย ได้แก่ การลดจำนวนการสื่อสารระหว่างแหล่งข้อมูล การรวมผลลัพธ์ที่ได้จากหลายแหล่งข้อมูลอย่างมีประสิทธิภาพ และการจัดการด้านความปลอดภัยและความเป็นส่วนตัว

1.3 ต้นกำเนิดของการทำเหมืองข้อมูล

ความท้าทายของการวิเคราะห์ข้อมูลดังอธิบายในหัวข้อ 1.2 เป็นแรงผลักดันให้นักวิจัยหลากหลายสาขาคิดค้นและพัฒนาเครื่องมือการประมวลผลข้อมูลที่มีประสิทธิภาพจัดการกับข้อมูลที่มีความหลากหลายและขนาดใหญ่ขึ้นได้ โดยการประยุกต์ใช้แนวคิดและเทคนิควิธีการจากหลายสาขาวิชา ดังนี้คือ

1) เทคนิคการสุ่ม การทดสอบสมมติฐานจากสถิติ (statistics) การสุ่มข้อมูลต้องคงลักษณะเดิมเอาไว้

2) เทคนิคการค้นหา การสร้างโมเดล และการเรียนรู้จากปัญญาประดิษฐ์ (artificial intelligence), การรู้จำรูปแบบ (pattern recognition), และการเรียนรู้ของเครื่อง (machine learning)

3) เทคนิคการบริหารจัดการข้อมูล การทำดัชนี และการประมวลผลคิวรี จากเทคโนโลยีฐานข้อมูล (database technology)

4) เทคนิคการประมวลผลข้อมูลแบบขนานและแบบกระจาย จากการคำนวณแบบขนานและแบบกระจาย (parallel computing and distributed computing)

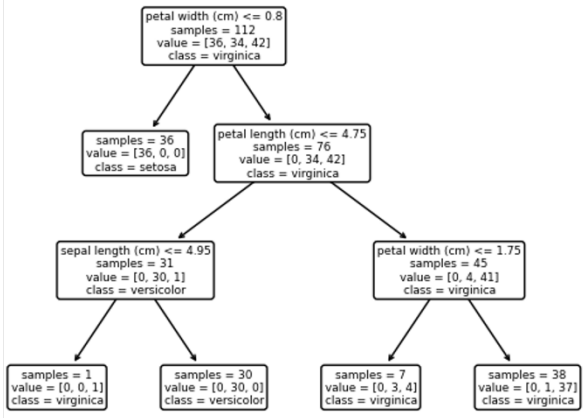
1.4 ประเภทของการทำเหมืองข้อมูล

รูปแบบของการทำเหมืองข้อมูลสามารถแบ่งออกเป็น 4 ประเภทหลัก คือ **การสร้างโมเดลเชิงทำนาย** (predictive modeling) **การวิเคราะห์ความสัมพันธ์** (association analysis) **การวิเคราะห์กลุ่มข้อมูล** (clustering analysis) และ**การตรวจจับข้อมูลผิดปกติ** (anomaly detection)

การสร้างโมเดลเชิงทำนาย (Predictive modeling) คือการสร้างโมเดลเพื่อทำนายค่า (regression) หรือจำแนกประเภท (classification) จากค่าอินพุต ในมุมมองทางคณิตศาสตร์โมเดลที่สร้างขึ้นเปรียบได้กับฟังก์ชันที่ใช้คำนวณค่าเอาต์พุตซึ่งเป็นคำทำนาย จากค่าของอินพุตที่ป้อนให้กับโมเดล

ตัวอย่างที่ 1.1 การทำนายสายพันธุ์ของดอกไอริส. ชุดข้อมูล Iris (https://archive.ics.uci.edu/ml/datasets/iris) ดังตัวอย่างในรูปที่ 2 เป็นชุดข้อมูลที่ประกอบด้วยคุณลักษณะของดอกไอริสจำนวน 150 ดอก โดยข้อมูลเกี่ยวกับดอกไอริสแต่ละดอกในชุดข้อมูลได้แก่ ความยาวกลีบเลี้ยง (sepal length (cm)) ความกว้างกลีบเลี้ยง (sepal width (cm)) ความยาวกลีบดอก (petal length (cm)) ความกว้างกลีบดอก (petal width (cm)) และสายพันธุ์ (class : มีค่าที่เป็นไปได้ 3 ค่า คือ Iris Setosa, Iris Versicolor, และ Iris Virginica) เราสามารถสร้างโมเดลที่ทำนายสายพันธุ์ของดอกไอริสจากข้อมูลขนาดของกลีบเลี้ยงและกลีบดอก ได้ โดยใช้อัลกอริทึม เช่น ต้นไม้ตัดสินใจ (decision tree) และ k-Nearest Neighbor classifier เป็นต้น

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
5	7.0	3.2	4.7	1.4	versicolor
6	6.4	3.2	4.5	1.5	versicolor
7	6.9	3.1	4.9	1.5	versicolor
8	5.5	2.3	4.0	1.3	versicolor
9	6.5	2.8	4.6	1.5	versicolor
10	6.3	3.3	6.0	2.5	virginica
11	5.8	2.7	5.1	1.9	virginica
12	7.1	3.0	5.9	2.1	virginica
13	6.3	2.9	5.6	1.8	virginica
14	6.5	3.0	5.8	2.2	virginica



รูปที่ 2 ตัวอย่างชุดข้อมูล *Iris dataset*

การวิเคราะห์ความสัมพันธ์ (association analysis) คือเทคนิคการทำเหมืองข้อมูลที่ใช้สำหรับค้นหาฟีเจอร์ที่มีความสัมพันธ์กันอย่างมีนัยสำคัญในรูปแบบของ กฎการอนุมาน (implication rules) หรือซับเซตของฟีเจอร์

ตัวอย่างที่ 1.2 การวิเคราะห์ตะกร้าสินค้า. การวิเคราะห์ข้อมูลความสัมพันธ์สามารถนำไปใช้กับข้อมูล ทรานแซคชัน ณ จุดขาย (point-of-sale transaction data) เพื่อค้นหาสินค้าที่มักถูกซื้อพร้อมกันในรูปของกฎความสัมพันธ์ เช่น {PC} → {Printer, A4Paper} ซึ่งบ่งบอกว่าลูกค้าที่ซื้อเครื่องคอมพิวเตอร์พีซี มักจะซื้อเครื่องพิมพ์และหมึกด้วย เป็นต้น กฎความสัมพันธ์ที่ค้นพบนี้สามารถนำไปใช้ในการทำ cross-selling (การขายสินค้าที่มีความเกี่ยวเนื่องกับสินค้าหลัก) ได้

Transaction ID	Items
1	{PC, Printer, A4Paper, Monitor}
2	{RAM, Router, Solid State Disk, UPS}
3	{PC, Printer, RAM, Monitor, Mouse, Keyboard}
4	{PC, Printer, UPS, Keyboard}
5	{Mouse, Router, RAM}
6	{PC, A4Paper, Monitor, RAM, Solid State Disk}
7	{Mouse, Keyboard, Microphone, Headphone}
8	{PC, Printer, Ink A4Paper}

รูปที่ 3 ตัวอย่างข้อมูลทรานแซคชัน ณ จุดขาย *(point-of-sale transaction data)*

การวิเคราะห์กลุ่มข้อมูล (Clustering analysis) คือการค้นหากลุ่มข้อมูล (data clusters) ที่ประกอบด้วยจุดข้อมูล (data items or observations) ที่มีความเกี่ยวข้องกันอย่างใกล้ชิด โดยที่จุดข้อมูลที่ถูกจัดอยู่ในกลุ่มหรือคลัสเตอร์เดียวกันจะมีความคล้ายคลึงกันมากกว่าจุดข้อมูลที่อยู่คนละคลัสเตอร์

ตัวอย่างที่ 1.3 การจัดกลุ่มเอกสาร. รูปที่ 4 แสดงชุดข้อมูลที่ประกอบด้วยบทความข่าว ซึ่งถูกแสดงในรูปแบบของรายการคำศัพท์และจำนวนครั้งที่ปรากฏในบทความ เช่น บทความที่ 1 ประกอบด้วยคำศัพท์คือ *เครื่องจักร* จำนวน 3 คำ , *แรงงาน* จำนวน 2 คำ, *ตลาด* จำนวน 4 คำ, *อุตสาหกรรม* จำนวน 2 คำ, *ประเทศไทย* จำนวน 2 คำ, และ*เงินบาท* จำนวน 1 คำ จากชนิดของคำศัพท์ของบทความในชุดข้อมูลตัวอย่างนี้จะเห็นได้ว่า บทความข่าวชุดนี้สามารถแบ่งออกได้เป็น 2 คลัสเตอร์ ได้แก่ บทความที่ 1 – 4 เป็นคลัสเตอร์ที่ประกอบด้วยบทความข่าวเศรษฐกิจ ส่วนบทความที่ 5-8 เป็นคลัสเตอร์ของบทความข่าวกีฬา อัลกอริทึมการจัดกลุ่มที่ดีจะต้องสามารถระบุคลัสเตอร์ทั้งสองดังกล่าวได้โดยการคำนวณหาความคล้ายคลึงกันระหว่างคำที่ปรากฏในบทความข่าว

บทความที่	คำที่ปรากฏในบทความ: จำนวนครั้ง
1	เครื่องจักร: 3, แรงงาน: 2, ตลาด: 4, อุตสาหกรรม: 2, ประเทศไทย: 2, เงินบาท: 1
2	เงินบาท: 1, อุตสาหกรรม: 2, การทำงาน: 2, เครื่องจักร: 1, รัฐบาล: 2, การกู้ยืม: 1
3	งาน: 5, เงินเฟ้อ: 3, ราคาสินค้า: 2, การว่างงาน: 3, ตลาด: 1, ดัชนีราคาสินค้า: 2
4	ท้องถิ่น: 3, การคาดการณ์: 2, ดุลการค้า: 1, ตลาด: 2, สินค้า: 3, ผู้บริโภค: 2
5	สโมสร: 2, ไทยแลนด์ลีก: 1, ถ้วยรางวัล: 1, นักเตะ: 3, ค้าแข่ง: 1, การแข่งขัน: 2
6	ลิเวอร์พูล: 1, เชลซี: 2, หงส์แดง: 1, ฟุตบอล: 2, พรีเมียร์ลีก: 2, ผู้จัดการทีม: 1
7	ฟุตบอล: 2, พรีเมียร์ลีก: 2, ผู้จัดการทีม: 1, โรนัลโด้ซ: 1, นักเตะ: 2, การแข่งขัน: 1
8	สโมสร: 2, ไทยแลนด์ลีก: 3, ตารางการแข่งขัน: 1, ผู้จัดการทีม: 3, ทีมชาติไทย: 1

รูปที่ 4 ชุดข้อมูลบทความข่าว

การตรวจจับข้อมูลผิดปกติ (Anomaly detection) คือการระบุจุดข้อมูลที่มีคุณลักษณะแตกต่างไปจากจุดข้อมูลอื่นอย่างมีนัยสำคัญ เราเรียกจุดข้อมูลดังกล่าวว่า ค่าผิดปกติ (anomalies หรือ outliers) ตัวอย่างเช่น การระบุการบุกรุกเครือข่าย การตรวจสอบการใช้บัตรเครดิตโดยทุจริต เป็นต้น

ตัวอย่างที่ 1.4 การตรวจจับการฉ้อโกงบัตรเครดิต. บริษัทผู้ให้บริการบัตรเครดิตจะมีการบันทึกรายการซื้อสินค้าด้วยบัตรเครดิตทุกรายการของผู้ถือบัตร รวมทั้งข้อมูลส่วนตัว เช่น อายุ รายได้ต่อปี และที่อยู่ เนื่องจากจำนวนทรานแซคชันบัตรเครดิตที่มีการฉ้อโกงมีจำนวนน้อยกว่าจำนวนทรานแซคชันที่ถูกกฎหมายมาก เทคนิคการตรวจจับข้อมูลผิดปกติจึงถูกนำมาใช้เพื่อสร้างโปรไฟล์หรือคุณลักษณะของทรานแซคชันที่ถูกกฎหมายสำหรับผู้ถือบัตรแต่ละคน เมื่อมีรายการทรานแซคชันบัตรเครดิตใหม่เกิดขึ้น บริษัทบัตรเครดิตฯ จะเปรียบเทียบคุณลักษณะของทรานแซคชันใหม่กับโปรไฟล์ของผู้ถือบัตร ถ้าคุณลักษณะของทรานแซคชันใหม่มีความแตกต่างไปจากโปรไฟล์ของผู้ถือบัตรอย่างมีนัยสำคัญ ทรานแซคชันใหม่ดังกล่าวจะถูกระบุว่าเป็นทรานแซคชันที่มีความเสี่ยงต่อการฉ้อโกงสูง และแจ้งให้กับผู้ถือบัตรที่แท้จริงทราบ

สรุป

- การทำเหมืองข้อมูลคือเทคโนโลยีที่ผสานรวมการวิเคราะห์ข้อมูลแบบดั้งเดิม กับอัลกอริทึมการประมวลผลข้อมูลที่ซับซ้อนและชาญฉลาด เพื่อสกัดรูปแบบและความรู้ที่เป็นประโยชน์ซึ่งแฝงอยู่ในข้อมูลปริมาณมหาศาล
- การทำเหมืองข้อมูลเป็นขั้นตอนหนึ่งของกระบวนการค้นหาความรู้จากฐานข้อมูลขนาดใหญ่ (KDD: Knowledge Discovery from Database) ซึ่งประกอบด้วยขั้นตอนหลัก คือ การเตรียมข้อมูล การทำเหมืองข้อมูล และการนำข้อมูลไปใช้งาน
- การทำเหมืองข้อมูลนำเทคนิคจากหลากหลายสาขาวิชา เช่น สถิติ ปัญญาประดิษฐ์ ระบบฐานข้อมูล และการคำนวณประสิทธิภาพสูง มาประยุกต์ใช้เพื่อจัดการกับความท้าทายในการประมวลผลข้อมูลขนาดใหญ่
- งานการทำเหมืองข้อมูลสามารถแบ่งได้เป็น 4 ประเภทหลัก คือ การสร้างโมเดลเชิงทำนาย (predictive modeling) การวิเคราะห์ความสัมพันธ์ (association analysis) การวิเคราะห์กลุ่มข้อมูล (clustering analysis) และการตรวจจับข้อมูลผิดปกติ (anomaly detection)

แบบฝึกหัด

1. จงพิจารณาว่ากิจกรรมใดต่อไปนี้จัดว่าเป็นการทำเหมืองข้อมูล

ก. การแบ่งข้อมูลลูกค้าออกเป็นกลุ่มโดยใช้ระดับรายได้ (มันง่ายไป ทำ if ก็ได้)

ข. การแบ่งข้อมูลลูกค้าออกเป็นกลุ่มโดยใช้วุฒิการศึกษา (มันง่ายไป ทำ if ก็ได้)

ค. การเรียงลำดับข้อมูลนิสิตตามอายุจากมากไปน้อย

predictive — ง. การทำนายราคาหุ้นในอีกหนึ่งอาทิตย์ข้างหน้าของบริษัทแห่งหนึ่งโดยใช้ประวัติราคาหุ้นของบริษัท

Clustering — จ. การทำ market segmentation โดยการแบ่งกลุ่มข้อมูลลูกค้าของบริษัทออกเป็นกลุ่มจำนวน 5 กลุ่ม

association — ฉ. การสร้างรายการเพลงสำหรับผู้ใช้งานโดยอัตโนมัติจากโปรไฟล์และพฤติกรรมการฟังของผู้ใช้งาน

association — ช. การค้นหารายการสินค้าที่ลูกค้ามักซื้อพร้อมกัน

ซ. การแปลงสีของรูปภาพจากภาพสีเป็นขาวดำ

ฌ. การปรับค่ากลางของข้อมูลให้อยู่ที่ค่าเฉลี่ย (mean centering)

ญ. การกำจัดข้อมูลที่ซ้ำซ้อนกันออกจากฐานข้อมูล

2. เทคนิคใดต่อไปนี้ ได้ถูกนำมาใช้ในการทำเหมืองข้อมูลที่มีขนาดใหญ่มาก

ก. การสุ่มตัวอย่าง (data sampling)

ข. การค้นคืนสารสนเทศ (information retrieval) — search engine

ค. การเข้ารหัสข้อมูล (data encryption)

ง. การประมวลผลแบบขนานและแบบกระจาย (parallel and distributed processing)

จ. อัลกอริทึมการจัดอันดับ (ranking algorithms)

เอกสารอ้างอิง

[1] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37.

[2] Ian H. Witten, Eibe Frank, Mark A. Hall, Christopher J. Pal. “Data Mining: Practical Machine Learning Tools and Techniques”, Morgan Kaufmann, 4th edition, 2016.

[3] Jiawei Han, Micheline Kamber, Jian Pei. “Data Mining: Concepts and Techniques”, Morgan Kaufmann, 3rd edition, 2011.

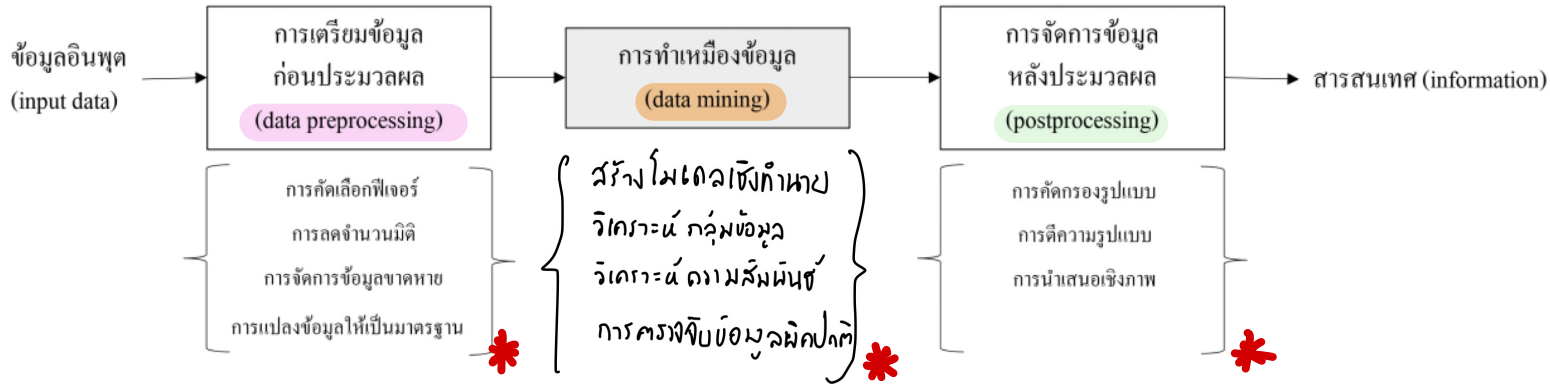
[4] Ming-Syan Chen, Jiawei Han and P. S. Yu, "Data mining: an overview from a database perspective" in IEEE Transactions on Knowledge and Data Engineering, vol. 8, no. 6, pp. 866-883, 1996.

[5] Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar. “Introduction to Data Mining”. Pearson, 2nd edition, 2018.

สรุป

การทำ Data mining คือ กระบวนการค้นหารูปแบบและความรู้ใหม่ ที่แฝงอยู่ในฐานข้อมูลขนาดใหญ่ เพื่อนำมาใช้ประโยชน์ ตัดสินใจ คาดการณ์ และการทำความเข้าใจ

ขั้นตอนการแปลงข้อมูลดิบเป็นสารสนเทศ



รูปที่ 1 ขั้นตอนหลักของกระบวนการค้นหาความรู้จากฐานข้อมูล (KDD)

data preprocessing เป็นการแปลง input ให้เหมาะสมกับการวิเคราะห์ในขั้นถัดไป เช่น การคัดเลือกฟีเจอร์
- การลดจำนวนมิติ
- การแปลงข้อมูลให้เป็นมาตรฐาน

data mining ในประเภทที่ใช้จะมีหลายอันหลัก
1. Predictive modeling
2. Clustering analysis
3. Association analysis
4. Anomaly detection

postprocessing หลังการประมวลผล ก่อนจะทำข้อมูลไปใช้ ต้องมีการจัดข้อมูลหลังประมวลผล เพื่อตรวจสอบความถูกต้องและคัดเลือกส่วนที่มีประโยชน์ไปใช้งานจริง

เครื่องมือที่ช่วยให้ data mining มีประสิทธิภาพจากเทคนิคดังต่อไปนี้
1. สถิติ การสุ่ม
2. ปัญญาประดิษฐ์ การค้นหา
3. ระบบฐานข้อมูล
4. เทคนิคการประมวลผลขนานกับกระจาย

data mining คือ เหมืองข้อมูล เปรียบเสมือน" การขุดเหมืองเพื่อให้ได้แร่ธาตุมา"
นั่นคือข้อมูลดิบ ขุดคือการจัดการ และแร่คือผลลัพธ์ที่เราจะนำไปใช้ทำอะไร

แบบฝึกหัด

1. จงพิจารณาว่ากิจกรรมใดต่อไปนี้จัดว่าเป็นการทำเหมืองข้อมูล

ก. การแบ่งข้อมูลลูกค้าออกเป็นกลุ่มโดยใช้ระดับรายได้ { มีน่าย ไป ต้อง if ก็ได้

ข. การแบ่งข้อมูลลูกค้าออกเป็นกลุ่มโดยใช้วุฒิการศึกษา

ค. การเรียงลำดับข้อมูลนิสิตตามอายุจากมากไปน้อย

predictive ง. การทำนายราคาหุ้นในอีกหนึ่งอาทิตย์ข้างหน้าของบริษัทแห่งหนึ่งโดยใช้ประวัติราคาหุ้นของบริษัท

Clustering จ. การทำ market segmentation โดยการแบ่งกลุ่มข้อมูลลูกค้าของบริษัทออกเป็นกลุ่มจำนวน 5 กลุ่ม

association ฉ. การสร้างรายการเพลงสำหรับผู้ใช้งานโดยอัตโนมัติจากโปรไฟล์และพฤติกรรมการฟังของผู้ใช้งาน

association ช. การค้นหารายการสินค้าที่ลูกค้ามักซื้อพร้อมกัน

ซ. การแปลงสีของรูปภาพจากภาพสีเป็นขาวดำ

ฌ. การปรับค่ากลางของข้อมูลให้อยู่ที่ค่าเฉลี่ย (mean centering)

ญ. การกำจัดข้อมูลที่ซ้ำซ้อนกันออกจากฐานข้อมูล

ก. ไม่เป็น เพราะ มีน่าย เกินไป เหมือน ตั้งแต่ if else

ข. ไม่เป็น เพราะ มีน่าย เกินไป เหมือน ตั้งแต่ if else

ค. ไม่เป็น เพราะ ตั้งแต่ so

ง. เป็น เพราะ ในการ ทำ นาย เรา สามารถ ใช้ predictive ในการทำนายได้

จ. เป็น เพราะ ในการแบ่งกลุ่ม เรา สามารถใช้ clustering ในการวิเคราะห์กลุ่มข้อมูล

ฉ. เป็น เพราะ ในการหา รายการเพลง เรา สามารถใช้ association ในการหา ความสัมพันธ์ของเพลงกับผู้ใช้งาน

ช. เป็น เพราะ ในการหา คู่สินค้าที่ ซื้อ พร้อมกัน เราสามารถใช้ association ในการหา ความสัมพันธ์การซื้อคู่สินค้าได้

ซ. ไม่เป็น เพราะการเปลี่ยนสีรูปภาพ เราใช้ image processing

ฌ. ไม่เป็น เพราะ การปรับค่ากลางของข้อมูล มีน่าย เกินไป

ญ. ไม่เป็น เพราะการจัด การ ข้อมูลที่ ซ้ำซ้อน อยู่ใน ขั้นตอน การ เตรียมข้อมูลก่อนประมวลผล

2. เทคนิคใดต่อไปนี้ ได้ถูกนำมาใช้ในการทำเหมืองข้อมูลที่มีขนาดใหญ่มาก

ก. การสุ่มตัวอย่าง (data sampling)

ข. การค้นคืนสารสนเทศ (information retrieval) — search engine

ค. การเข้ารหัสข้อมูล (data encryption)

ง. การประมวลผลแบบขนานและแบบกระจาย (parallel and distributed processing)

จ. อัลกอริทึมการจัดอันดับ (ranking algorithms)

ใช้ก็ได้

ก. เป็นเพราะ ใช้ predictive ในการสร้างโมเดลเชิงทำนายได้

ข. ไม่เป็น เพราะ เป็นการใช้ search engine เพื่อค้นหาข้อมูล ฯลฯ

ค. การเข้ารหัสข้อมูลในบริษัทใหญ่ คิดว่าอาจจะใช้ เพราะ น่าจะเข้าใช้ข้อมูลลูกค้าได้เร็วขึ้น

ง. เป็นเพราะ ใช้เทคนิค parallel and distributed processing ในการประมวลผลข้อมูล

จ. ไม่เป็น เพราะ แค่เป็นการจัดลำดับ