

เทคนิคเหมืองข้อมูล

DATA MINING TECHNIQUES

จรัสศรี รุ่งรัตนอุบล



สำนักพิมพ์มหาวิทยาลัยนเรศวร

Naresuan University Publishing House

www.nupress.grad.nu.ac.th



สำนักพิมพ์มหาวิทยาลัยนเรศวร Naresuan University Publishing House

บัณฑิตวิทยาลัย มหาวิทยาลัยนเรศวร 99 หมู่ 9 อาคารมหาธรรมราชา ชั้น 1 มหาวิทยาลัยนเรศวร
ตำบลท่าโพธิ์ อำเภอเมือง จังหวัดพิษณุโลก 65000 โทร. 0 5596 8833-8836 E-mail : nuph@nu.ac.th
 www.nupress.grad.nu.ac.th
 สำนักพิมพ์มหาวิทยาลัยนเรศวร
 @nupress

สงวนลิขสิทธิ์ หนังสือเล่มนี้ตามพระราชบัญญัติลิขสิทธิ์ (ฉบับเพิ่มเติม) พ.ศ. 2558 ห้ามคัดลอกเนื้อหา ภาพประกอบ รวมทั้งดัดแปลงเป็นฉบับอื่นที่ก่อกวนลิขสิทธิ์ การผลิต การลอกเลียนไม่ว่าส่วนใดส่วนหนึ่งของหนังสือเล่มนี้ หรือเผยแพร่ด้วยรูปแบบและวิธีการอื่นใด จะต้องได้รับอนุญาตเป็นลายลักษณ์อักษรจากสำนักพิมพ์มหาวิทยาลัยนเรศวร เท่านั้น

ข้อมูลทางบรรณานุกรมของหอสมุดแห่งชาติ

National Library of Thailand Cataloging in Publication Data

จรัสศรี รุ่งรัตนอุบล.

เทคนิคเหมืองข้อมูล Data Mining Techniques.--พิษณุโลก : สำนักพิมพ์มหาวิทยาลัยนเรศวร, 2566.

298 หน้า.

1. เหมืองข้อมูล. I. ชื่อเรื่อง.

006.312

ISBN 978-616-426-292-8

ISBN (e-book) 978-616-426-293-5

สพน. 118

ราคา 400 บาท

พิมพ์ครั้งที่ 1 มกราคม พ.ศ. 2566

จัดพิมพ์โดย สำนักพิมพ์มหาวิทยาลัยนเรศวร

วางจำหน่ายที่

1. ศูนย์หนังสือแห่งจุฬาลงกรณ์มหาวิทยาลัย
ถนนพญาไท แขวงวังใหม่ เขตปทุมวัน กรุงเทพฯ 10330 โทร. 0 2218 9812
2. ศูนย์หนังสือมหาวิทยาลัยเกษตรศาสตร์
ถนนงามวงศ์วาน แขวงลาดยาว เขตจตุจักร กรุงเทพฯ 10900 โทร. 0 2579 0113
3. ศูนย์หนังสือมหาวิทยาลัยธรรมศาสตร์
ถนนพระจันทร์ แขวงพระบรมมหาราชวัง เขตพระนคร กรุงเทพฯ 10200 โทร. 0 2613 3899
4. สำนักพิมพ์มหาวิทยาลัยนเรศวร
บัณฑิตวิทยาลัย มหาวิทยาลัยนเรศวร อาคารมหาธรรมราชา จังหวัดพิษณุโลก 65000 โทร. 0 5596 8833-8836

ประธานกองบรรณาธิการ รองศาสตราจารย์ ดร.กรรณกาญจน์ ขุทธิพงษ์ คณบดีบัณฑิตวิทยาลัย มหาวิทยาลัยนเรศวร

กองบรรณาธิการ

รองศาสตราจารย์ ดร.สุชาติ แย้มเม่น • รองศาสตราจารย์สุทัศน์ เขียมวัฒนา • รองศาสตราจารย์ ดร.ศักดา สมกุล •
รองศาสตราจารย์ ดร.เกตุจันทร์ จำปาไชยศรี • รองศาสตราจารย์ ดร. พญ.สุชาติ พงษ์เจริญ •
รองศาสตราจารย์ ดร. ภญ.กรรณก อิงคินันท์ • รองศาสตราจารย์ ดร.นิทรา กิจธีระวุฒิจ • รองศาสตราจารย์ ดร.สุทิสา ถาน้อย •
รองศาสตราจารย์ ดร.กิตติมา ขาววิชัย • รองศาสตราจารย์ ดร.รุจโรจน์ แก้วอุไร • รองศาสตราจารย์ นาวาโท ดร.วัฒนชัย หนัณย •
รองศาสตราจารย์ ดร.วัชรพล พุทธิรักษา • รองศาสตราจารย์ ดร.พงศ์พันธ์ กิจสนาโยธิน • ผู้ช่วยศาสตราจารย์ ดร.ยุวรงค์ จันทรวิจิตร •
ผู้ช่วยศาสตราจารย์จรรยาภรณ์ สุวพันธ์ • พิชรี ท่วมใจดี • นวพรรณ ดันติพลผล

ประสานงาน

ภาคีสัน เตตสิทิกุล

ฝ่ายขาย/การเงิน

พิมพ์ภรณ์ ดวงสำโรจน์ • วสันต์ มาสวัสดิ์

ออกแบบปก

สรุยา แสงเย็นพันธ์

ออกแบบรูปเล่ม

ธรรมบุญ กองกุล

พิมพ์ที่

บริษัท กู๊ดเอด พรินต์ติ้ง แอนด์ แพคเกจจิ้ง กรู๊ป จำกัด 6/1 นิคมอุตสาหกรรมบางชัน ซอยเสรีไทย 58 แขวงมีนบุรี เขตมีนบุรี กรุงเทพฯ 10510



สำนักพิมพ์นี้เป็นสมาชิกสมาคมผู้จัดพิมพ์
และผู้จำหน่ายหนังสือแห่งประเทศไทย
<https://pubat.or.th>



จับกับ
กระดาษคุณภาพ เพื่อผลงานคุณภาพ
กระดาษอนิเมะสายการบิน



 @nupress

 สำนักพิมพ์มหาวิทยาลัยนเรศวร

กรณีต้องการสั่งซื้อหนังสือปริมาณมาก หรือเข้าชั้นเรียนติดต่อได้ที่ฝ่ายจัดจำหน่ายสำนักพิมพ์มหาวิทยาลัยนเรศวร
โทร. 0 5596 8836 Email : nuph@nu.ac.th

คำนำ

เหมืองข้อมูลเป็นหนึ่งในความรู้สมัยใหม่ที่ได้รับความนิยมอย่างมากในปัจจุบัน โดยได้ถูกประยุกต์ใช้ในแทบทุกองค์กร ไม่ว่าจะเป็นภาครัฐและเอกชนเพื่อใช้ในการตัดสินใจ การวางแผนกลยุทธ์ การปรับปรุงการให้บริการและการดำเนินงานต่าง ๆ ในองค์กร เหมืองข้อมูลเป็นการทำงานที่เน้นการค้นหาสารสนเทศหรือองค์ความรู้จากข้อมูลขนาดใหญ่ เพื่อนำสิ่งที่ได้มาให้เป็นประโยชน์ โดยเหมืองข้อมูลเป็นการผสมผสานศาสตร์ทางสถิติ ปัญญาประดิษฐ์ การรู้จำและฐานข้อมูลเข้าด้วยกัน

หนังสือ “เทคนิคเหมืองข้อมูล” เล่มนี้ เป็นเนื้อหาการทำเหมืองข้อมูลโดยมีวัตถุประสงค์เพื่อให้หลักการ เทคนิค และขั้นตอนวิธีที่สำคัญของเหมืองข้อมูล โดยเน้นการนำเสนอแนวคิดและขั้นตอนวิธีของเทคนิคเหมืองข้อมูลต่าง ๆ เช่น เทคนิคค้นหาค่าที่ซ่อนอยู่ โครงข่ายประสาทเทียม การจัดกลุ่มด้วยเคมีน การวิเคราะห์ความสัมพันธ์ เป็นต้น โดยผู้เขียนได้ยกตัวอย่างงานวิจัยที่เกี่ยวกับการประยุกต์ใช้เหมืองข้อมูลของผู้เขียน ผู้ร่วมวิจัย และนักศึกษาได้จัดทำร่วมกันเพื่อเป็นแนวทางการประยุกต์ใช้ให้กับผู้อ่าน นอกจากนี้ผู้เขียนได้ใช้โปรแกรมเหมืองข้อมูล เวกา (Weka) ที่พัฒนาโดย University of Waikato ประเทศนิวซีแลนด์ เพื่อนำเสนอผลลัพธ์การทำงานของแต่ละเทคนิคเหมืองข้อมูล โดยโปรแกรมเวกามีรูปแบบการใช้งานง่าย เหมาะกับการใช้งานเพื่อศึกษาเทคนิคเหมืองข้อมูล

ผู้เขียนได้มีโอกาสศึกษา ทำวิจัยและสอนเกี่ยวกับเทคนิคเหมืองข้อมูลตั้งแต่ปี พ.ศ. 2553 และได้รวบรวมหลักการและเทคนิคที่สำคัญของเหมืองข้อมูลไว้ในหนังสือเล่มนี้ ปัจจุบันเทคนิคเหมืองข้อมูลเป็นเนื้อหาที่สอดแทรกไว้ในหลายวิชา เช่น เทคนิคการทำเหมืองข้อมูล (Data Mining Techniques) การทำเหมืองข้อมูลและข้อมูลขนาดใหญ่ (Data Mining and Big Data) การวิเคราะห์ข้อมูลขนาดใหญ่ (Big Data Analytics) เป็นต้น ทำให้มีผู้สนใจศึกษาเป็นวงกว้าง หนังสือเล่มนี้เหมาะกับผู้สนใจหลักการพื้นฐานของการทำเหมืองข้อมูลและการทำงานของเทคนิคเหมืองข้อมูลแบบต่าง ๆ โดยผู้เขียนหวังว่าผู้อ่านจะสามารถนำเทคนิคเหมืองข้อมูลไปประยุกต์ใช้งานได้อย่างถูกต้องและมีประสิทธิภาพ และสามารถตีความผลลัพธ์จากการทำเหมืองข้อมูลได้อย่างถูกต้อง โดยไม่มองว่าเหมืองข้อมูลเป็นเพียงโปรแกรมอัตโนมัติที่เมื่อใส่ข้อมูลแล้วจะได้ผลลัพธ์ออกมา

ผู้เขียนขอขอบคุณบิดา มารดา และครูอาจารย์ที่อบรมสั่งสอนให้ความรู้ แง่คิดที่ดีในการทำงานและใช้ชีวิต ตลอดจนกัลยาณมิตรทุกคนที่เป็นกำลังใจให้ผู้เขียนจัดทำหนังสือเล่มนี้จนเสร็จสมบูรณ์

สุดท้ายนี้ ผู้เขียนหวังเป็นอย่างยิ่งว่าหนังสือเล่มนี้จะเป็นประโยชน์ต่อผู้อ่านที่สนใจ หากพบข้อบกพร่องและมีข้อคิดเห็นหรือข้อเสนอแนะประการใด ติดต่อผู้เขียนได้ที่ e-mail : jaratsrir@nu.ac.th ผู้เขียนขออภัยในความผิดพลาดต่าง ๆ เพื่อนำมาเป็นแนวทางในการปรับปรุงเนื้อหาให้มีความชัดเจนและดียิ่งขึ้นในโอกาสต่อไป

ผู้ช่วยศาสตราจารย์ ดร.จรัสศรี รุ่งรัตนอุบล

สารบัญ

บทที่ 1 แนะนำการทำเหมืองข้อมูล (Introduction to Data Mining).....	1
ลักษณะการจัดเก็บข้อมูลและปริมาณข้อมูลในปัจจุบัน.....	2
เหมืองข้อมูลคืออะไร.....	7
ทำไมต้องใช้เหมืองข้อมูล.....	10
กระบวนการทำงานของเหมืองข้อมูล.....	11
ตัวอย่างการประยุกต์ใช้เหมืองข้อมูล.....	20
เครื่องมือและโปรแกรมเหมืองข้อมูล.....	25
บทสรุป.....	26
แบบฝึกหัดท้ายบท.....	27
บทที่ 2 การเตรียมข้อมูล (Data Preprocessing).....	29
ข้อมูลคืออะไร.....	30
ประเภทและลักษณะข้อมูล.....	31
การเตรียมข้อมูล.....	37
มาตรวัดความคล้ายและความต่างของวัตถุ.....	45
บทสรุป.....	53
แบบฝึกหัดท้ายบท.....	54
บทที่ 3 เทคนิคการจำแนก (Classification).....	55
ขั้นตอนการพัฒนาตัวจำแนก.....	56
ขั้นตอนวิธีค้นหาเพื่อนบ้านใกล้ที่สุด k ตัว (K-nearest Neighbor Algorithm).....	59
1. การแปลงข้อมูลเป็นรูปแบบมาตรฐานกับขั้นตอนวิธีค้นหาเพื่อนบ้านใกล้ที่สุด k ตัว.....	64
2. การใช้ส่วนกลับของระยะห่างกับขั้นตอนวิธีค้นหาเพื่อนบ้านใกล้ที่สุด k ตัว.....	65
3. การใช้โปรแกรมเวกากับขั้นตอนวิธีค้นหาเพื่อนบ้านใกล้ที่สุด k ตัว.....	69
เทคนิคต้นไม้ตัดสินใจ (Decision Tree).....	73
1. หลักการพัฒนาแบบจำลองการพยากรณ์เพื่อการจำแนกหรือตัวจำแนก.....	73
2. กระบวนการสร้างต้นไม้ตัดสินใจ.....	74
3. การประเมินประสิทธิภาพของโมเดลเพื่อพยากรณ์หรือตัวจำแนก.....	92

4. การแตกกิ่งของลักษณะประจำปจจัยเข้าแบบต่าง ๆ	96
5. การใช้โปรแกรมเวกากับขั้นตอนวิธีต้นไม้ตัดสินใจ	101
5.1 การใช้โปรแกรมเวกาเพื่อพัฒนาตัวจำแนกแบบต้นไม้ตัดสินใจกับข้อมูลที่ค่า ลักษณะประจำเต็มหน่วย	101
5.2 การใช้โปรแกรมเวกาเพื่อพัฒนาตัวจำแนกแบบต้นไม้ตัดสินใจกับข้อมูลที่ค่าลักษณะ ประจำต่อเนื่อง	108
การสร้างกฎเพื่อการจำแนก (Rule Based Classifier)	110
1. การคำนวณค่าความครอบคลุมและความถูกต้องของกฎ.....	111
2. หลักการสร้างกฎเพื่อการจำแนก.....	115
3. การใช้โปรแกรมเวกาเพื่อสร้างกฎเพื่อการจำแนก	118
ตัวจำแนกเบย์อย่างง่าย (Naïve Bayes Classifier).....	123
1. แนวคิดของตัวจำแนกแบบเบย์.....	124
2. การใช้โปรแกรมเวกากับตัวจำแนกแบบเบย์	132
โครงข่ายประสาทเทียม (Artificial Neural Network).....	135
1. สถาปัตยกรรมของโครงข่ายประสาทเทียม.....	135
2. การใช้โปรแกรมเวกาเพื่อพัฒนาตัวจำแนกด้วยโครงข่ายประสาทเทียม.....	137
บทสรุป.....	143
แบบฝึกหัดท้ายบท	144

บทที่ 4 การวิเคราะห์การจัดกลุ่ม (Cluster Analysis).....147

การวิเคราะห์การจัดกลุ่ม.....	148
การจัดกลุ่มแบบลำดับขั้น (Hierarchical Clustering).....	150
การใช้โปรแกรมเวกาเพื่อการจัดกลุ่มแบบลำดับขั้น.....	167
การจัดกลุ่มข้อมูลแบบเคมีน (K-mean Clustering)	170
การใช้โปรแกรมเวกาเพื่อการจัดกลุ่มแบบเคมีน	179
บทสรุป.....	182
แบบฝึกหัดท้ายบท	183

บทที่ 5 การวิเคราะห์ความสัมพันธ์ (Association Analysis).....185

กฎความสัมพันธ์ (Association Rules).....	186
---	-----

ขั้นตอนวิธีที่ใช้ในการสร้างกฎความสัมพันธ์	187
ขั้นตอนวิธีอะไพอรี (A priori algorithm).....	198
การเตรียมข้อมูลสำหรับการหากฎความสัมพันธ์.....	202
การใช้โปรแกรมเวก้าเพื่อค้นหากฎความสัมพันธ์.....	203
การตีความหมายของกฎความสัมพันธ์	210
สรุปการใช้งานกฎความสัมพันธ์	211
บทสรุป.....	211
แบบฝึกหัดท้ายบท	212
 บทที่ 6 การพยากรณ์ (Prediction).....	215
การวิเคราะห์การถดถอย (Regression Analysis).....	216
1. การวิเคราะห์การถดถอยอย่างง่าย	217
2. การวิเคราะห์การถดถอยเชิงพหุ	236
3. เกณฑ์การวัดค่าประสิทธิภาพของตัวแบบพยากรณ์	249
4. การใช้โปรแกรมเวก้ากับการวิเคราะห์การถดถอย	250
โครงข่ายประสาทเทียม (Artificial Neural Networks: ANN).....	254
1. โครงสร้างของโครงข่ายประสาทเทียมแบบง่าย	256
2. โครงสร้างและองค์ประกอบของโครงข่ายประสาทเทียม	258
3. การแปลงข้อมูลเข้าและการแปลงข้อมูลออกหรือผลลัพธ์.....	263
4. หลักการคำนวณค่าน้ำหนักแบบวิธีแพร่ย้อนกลับ (Back Propagation)	265
5. การใช้โปรแกรมเวก้ากับโครงข่ายประสาทเทียมแบบ MultilayerPerceptron.....	275
บทสรุป.....	281
แบบฝึกหัดท้ายบท	282
 บรรณานุกรม.....	284
 ดัชนี	287

บทที่ 1

แนะนำการทำเหมืองข้อมูล





ในชีวิตประจำวันของเราทุกคนจะต้องข้องเกี่ยวกับข้อมูลต่าง ๆ มากมายที่เราจำเป็นต้องจดจำและจดบันทึกลงบนกระดาษหรือบนอุปกรณ์ช่วยจำ ตั้งแต่อดีตจนถึงปัจจุบันมนุษยชาติมีการบันทึกข้อมูลเรื่องราวต่าง ๆ อย่างต่อเนื่องเพื่อเก็บไว้เป็นข้อมูลทางสถิติหรือข้อมูลทางประวัติศาสตร์ เพื่อนำข้อมูลเหล่านี้มาใช้ให้เกิดประโยชน์ต่อการวางแผนการทำงาน การกำหนดทิศทางการดำเนินงาน หรือเพื่อสนับสนุนการตัดสินใจในเรื่องต่าง ๆ เช่น การทำนายผลประกอบการของบริษัท การวางแผนงานเชิงรุกของบริษัท เป็นต้น

ถ้าเราลองพิจารณาถึงข้อมูลส่วนบุคคลต่าง ๆ ที่เราต้องจัดเก็บตั้งแต่เกิด จะประกอบด้วยข้อมูลมากมาย เช่น วันเกิด น้ำหนักแรกเกิด ความสูง น้ำหนัก โรคภัย วุฒิการศึกษา ประวัติการทำงาน อายุ เงินเดือน วันแต่งงาน บันทึกค่าใช้จ่าย วันตาย เป็นต้น ข้อมูลเหล่านี้เป็นเพียงตัวอย่างอันเล็กน้อยของข้อมูลที่มีการจดบันทึกและจัดเก็บจริงของคนคนเดียว แต่ถ้าลองคิดดู คนบนโลกใบนี้มีจำนวนกว่าหมื่นล้านคนจะมีปริมาณข้อมูลจำนวนมากมายมหาศาลเพียงใด และนอกเหนือจากข้อมูลส่วนบุคคลแล้ว ยังมีข้อมูลแวดล้อมอื่น ๆ อีกมากมายที่อยู่รอบตัวเรา เช่น ราคาอาหาร ราคาน้ำมัน ราคาทอง ปริมาณน้ำฝนและอุณหภูมิจากสถานีวัด ภาพถ่ายจากดาวเทียม ข่าวสารในแต่ละวัน เป็นต้น

ลักษณะการจัดเก็บข้อมูลและปริมาณข้อมูลในปัจจุบัน

ข้อมูลที่ได้กล่าวมาแล้วนั้นเราอาจจะเรียกว่าเป็น “ข้อมูลดิบ” เพราะเป็นข้อมูลที่เกิดจากการบันทึก โดยที่ยังไม่ได้ผ่านการกลั่นกรองหรือประมวลผลเพื่อให้ได้ข้อมูลที่มีลักษณะเป็น “ข้อสรุป” หรือ “สารสนเทศ” โดยในทางปฏิบัติแล้ว เรามักจะนำข้อมูลดิบเหล่านี้มาจัดทำข้อสรุปหรือสารสนเทศ โดยวิธีการประมวลผลด้วยหลักสถิติพื้นฐานและสถิติขั้นสูง เช่น ค่าเฉลี่ยของปริมาณน้ำฝน ค่าเฉลี่ยช่วงอายุคน รายได้ประชาชาติของคนไทย การอนุมานทางสถิติสำหรับพยากรณ์ราคาน้ำมันในตลาดโลก เป็นต้น

ในอดีตก่อนที่จะมีการใช้เทคโนโลยีสารสนเทศและการสื่อสาร ข้อมูลมักจะถูกจัดเก็บด้วยคนโดยการจดบันทึก และการประมวลผลก็จะถูกประมวลโดยนักวิเคราะห์ข้อมูลหรือนักสถิติ ต่อมาเมื่อเริ่มมีการนำคอมพิวเตอร์มาใช้ ข้อมูลเหล่านี้จะถูกจัดเก็บลงคอมพิวเตอร์โดยการคีย์เข้าระบบโดยคนและการประมวลผลก็จะทำด้วยโปรแกรมคอมพิวเตอร์ จนมาถึงยุคปัจจุบันที่คอมพิวเตอร์และเทคโนโลยีสารสนเทศและการสื่อสารได้พัฒนาอย่างต่อเนื่องแบบก้าวกระโดด ข้อมูลต่าง ๆ ถูกจัดเก็บผ่านอุปกรณ์เชื่อมต่อต่าง ๆ ไม่ว่าจะเป็นอุปกรณ์รับส่งสัญญาณ (Sensor)

บทที่ 2

การเตรียมข้อมูล (Data Preprocessing)





ในแต่ละวันเราจะได้รับข้อมูลและสารสนเทศมากมาย โดยข้อมูลเหล่านี้อาจจะเป็นข้อมูลที่ผ่านมาและผ่านไปโดยที่เราไม่ได้สนใจ หรือบางทีอาจเป็นข้อมูลที่มีความสำคัญที่เราจะต้องจดจำและรับทราบเอาไว้ หรือเป็นข้อมูลที่เราต้องเก็บมาวิเคราะห์ สังเคราะห์ เพื่อนำไปใช้ให้เกิดประโยชน์ต่อไป

ข้อมูลคืออะไร

ข้อมูล (Data) คือ ข้อเท็จจริงเกี่ยวกับเรื่องที่เราสนใจ ซึ่งอาจเป็นการจัดเก็บแบบจัดบันทึกรายวัน หรือเป็นการจัดเก็บอย่างมีระบบระเบียบในลักษณะของฐานข้อมูล ซึ่งในที่นี้จะอธิบายข้อมูลในมุมมองของกลุ่มของค่าของข้อมูลที่อยู่รวมกัน ซึ่งจะเรียกว่า **ลักษณะประจำ (Attributes)** หรือตัวแปร (**Variable**)

โดยความหมาย **ลักษณะประจำ (Attributes)** คือ คุณสมบัติหรือลักษณะประจำของข้อมูลหรือวัตถุหรือสิ่งที่เราสนใจ เช่น ลักษณะประจำอายุ ลักษณะประจำเพศ ลักษณะประจำสีตา เป็นต้น ซึ่งจะมีลักษณะและค่าแตกต่างกันไป

ค่าของข้อมูลตามลักษณะประจำเหล่านี้ อาจได้มาจากการวัด การนับ หรือการบันทึกค่าโดยคน (**Human**) หรือโดยอุปกรณ์รับส่งสัญญาณ (**Sensor**) หรือโดยคอมพิวเตอร์ (**Computer**) ข้อมูลสำหรับการทำเหมืองข้อมูลส่วนใหญ่จะมาในรูปแบบของชุดข้อมูลซึ่งจะมีลักษณะแตกต่างกันไป ในที่นี้ขอกล่าวถึงข้อมูลที่มีโครงสร้างตารางที่คล้ายคลึงกับตารางในฐานข้อมูลมาเป็นตัวอย่าง ชุดข้อมูล (**Dataset**) สำหรับการทำเหมืองข้อมูล โดยในตารางจะประกอบด้วย **แถวสดมภ์ (Column)** หรือ **ลักษณะประจำ** ที่จัดเก็บ **ค่า** ที่เป็นไปตามลักษณะประจำ ในที่นี้เราจะเรียกว่า **ค่าลักษณะประจำ (Attribute values)** เช่น ลักษณะประจำอายุจะมีค่าลักษณะประจำเป็นตัวเลขบวก ลักษณะประจำเพศมีค่าลักษณะประจำเป็นค่า “หญิง” หรือ “ชาย” เป็นต้น ค่าลักษณะประจำจะเป็นไปตามลักษณะของลักษณะประจ้านั้น ๆ ซึ่งการเข้าใจลักษณะประจำและค่าลักษณะประจ้านั้นมีความสำคัญอย่างมากในกระบวนการคัดกรองและทำความสะอาดข้อมูล

แต่ละแถวในตารางหรือที่เรียกว่า “ระเบียน” ในฐานข้อมูลจัดเป็น **ค่าข้อมูล (Data Value)** ซึ่งประกอบด้วยค่าต่าง ๆ ตามลักษณะประจำในตาราง ซึ่งสำหรับหนังสือเล่มนี้จะเรียกว่า **วัตถุ (Object)** ดังแสดงในภาพ 2.1

บทที่ 3

เทคนิคการจำแนก (Classification)





เทคนิคการจำแนกเป็นเทคนิคหนึ่งในการทำเหมืองข้อมูลที่ใช้เพื่อทำนายคำตอบที่เป็นค่าเชิงคุณภาพ (Qualitative Value) หรือค่าเต็มหน่วย (Discrete Value) หรือค่าแบบแค็ตตาล็อก (Catalogue Value) เช่น ใช่/ไม่ใช่ ซื้อ/ไม่ซื้อ คำตอบ ก/ข/ค/ง ระดับความพึงพอใจ ดีมาก/ดี/พอใช้ เป็นต้น โดยใช้หลักการการนำชุดข้อมูลที่มีอยู่มาพัฒนาโมเดลเพื่อการจำแนก และประยุกต์ใช้หาคำตอบหรือทำนายคำตอบของข้อมูลชุดใหม่ (Unseen Objects) ที่เข้ามา

โดยเทคนิคนี้ได้รับความนิยมอย่างมาก และถูกนำมาประยุกต์ใช้เพื่อสนับสนุนการตัดสินใจทางธุรกิจและทางวิทยาศาสตร์ เพราะการพยากรณ์เพื่อจำแนกข้อมูลใหม่ที่จะเข้ามาควรจะถูกจัดหรือจำแนกให้เป็นหมวดใดเป็นสิ่งที่นำมาใช้เพื่อการวางแผนและการตัดสินใจในการดำเนินการต่าง ๆ ได้ ตัวอย่างของการประยุกต์ใช้การจำแนก ดังเช่น

1. การจำแนกลักษณะของเซลล์ว่าเป็นเซลล์ผิดปกติประเภท เนื้องอกหรือมะเร็ง
2. การตรวจสอบรายการธุรกรรมทางบัตรเครดิตว่าเป็น แบบปกติหรือปลอมแปลง
3. การจำแนกเพื่อระบุว่าโครงสร้างโปรตีนเป็นแบบใดใน 3 แบบนี้ alpha-helix beta-sheet หรือ random-coil
4. การจำแนกข่าวด้วยการพิจารณาเนื้อหาภายในเพื่อจำแนกว่าควรจะเป็นข่าวประเภทใดในประเภทต่อไปนี้ ข่าวการเงิน (Finance) ข่าวกีฬา (Sports) ข่าวบันเทิง (Entertainment) หรือข่าวอาชญากรรม (Crime)

โดยการพัฒนาโมเดลเพื่อการจำแนก (Classification Model) หรือตัวจำแนก (Classifier) จะมีหลักในการพัฒนาและอัลกอริทึมที่เกี่ยวข้องหลายตัวที่นิยมใช้ในปัจจุบัน โดยในที่นี้จะกล่าวถึงขั้นตอนวิธีการค้นหาเพื่อนบ้านใกล้ที่สุด k ตัว (K-nearest Neighbor Algorithm) วิธีต้นไม้ตัดสินใจ (Decision Tree) การสร้างกฎ (Rule-based Classifier) วิธีเบย์อย่างง่าย (Naïve Bayes Classifier) และโครงข่ายประสาทเทียม (Artificial Neural Network)

ขั้นตอนการพัฒนาตัวจำแนก

ดังที่ได้เกริ่นไว้ก่อนหน้านี้ว่าการพัฒนาตัวจำแนกเป็นการทำงานที่ต้องการได้ตัวจำแนกที่มีประสิทธิภาพโดยการนำชุดข้อมูลที่มีมาพัฒนาตัวโมเดลเพื่อการจำแนก โดยมีจุดประสงค์ที่ต้องการใช้ตัวจำแนกที่ได้นี้มาทำนายหรือจำแนกข้อมูลชุดใหม่ ดังนั้นกลไกการพัฒนาตัวจำแนกนั้นจะประกอบด้วยขั้นตอนดังที่จะอธิบายดังต่อไปนี้

บทที่ 4

การวิเคราะห์การจัดกลุ่ม (Cluster Analysis)





การวิเคราะห์การจัดกลุ่ม (Cluster Analysis) เป็นอีกหนึ่งเทคนิคของเหมืองข้อมูล ที่ได้รับความนิยมใช้ในงานด้านต่าง ๆ อย่างแพร่หลาย เช่น การจัดกลุ่มลูกค้าของบริษัท การจัดกลุ่มเอกสาร การจัดกลุ่มผู้ป่วย เป็นต้น การจัดกลุ่มข้อมูลเป็นเทคนิคที่อยู่ในกลุ่มของการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) ที่เน้นการบรรยายลักษณะข้อมูลมากกว่าการทำนายหรือพยากรณ์ ที่จัดเป็นการเรียนรู้แบบมีผู้สอน (Supervised Learning) ส่วนใหญ่งานด้านนี้มีไว้เพื่อลดขนาดหรือมิติของข้อมูลให้เป็นกลุ่มหรือคลัสเตอร์ โดยมีจุดประสงค์เพื่อรวมกลุ่มของสิ่งที่มีความคล้ายกันให้อยู่กลุ่มเดียวกัน เพื่อจะได้ทำให้ง่ายต่อการดำเนินการทางการทำธุรกิจ หรือการวิเคราะห์ปัจจัยได้เจาะจงยิ่งขึ้น เช่น การสร้างโปรไฟล์การตลาดท่องเที่ยวด้วยการวิเคราะห์การจัดกลุ่ม การวิเคราะห์การจัดกลุ่มของลูกค้าที่มีลักษณะหรือพฤติกรรมการบริโภคที่คล้ายคลึงกัน การจัดกลุ่มเอกสารที่มีสาระหลักหรือสาระสำคัญที่คล้ายคลึงกัน เป็นต้น

การวิเคราะห์การจัดกลุ่ม

การวิเคราะห์การจัดกลุ่ม เป็นเทคนิคที่ใช้ในการจัดกลุ่มข้อมูลโดยมีหลักการง่าย ๆ คือ ข้อมูลที่อยู่ในกลุ่มเดียวกันจะมีความคล้ายคลึงกัน แต่จะมีความแตกต่างกับข้อมูลที่อยู่คนละกลุ่ม โดยกระบวนการของการจัดกลุ่มจะนำข้อมูลทั้งหมดที่มีมาวัดค่าความคล้าย โดยถ้าข้อมูลมีความคล้ายกันมากพอก็จะถูกจัดให้อยู่กลุ่มเดียวกัน ถ้ามีความคล้ายกันน้อยก็จะถูกจัดให้อยู่คนละกลุ่ม กล่าวโดยสรุปคือ ให้นำสิ่งที่มีลักษณะคล้ายกันมาไว้กลุ่มเดียวกัน ส่วนข้อมูลที่มีลักษณะต่างกัน ก็ให้อยู่คนละกลุ่ม

การใช้งานการวิเคราะห์การจัดกลุ่มนั้นมีวัตถุประสงค์หลักดังต่อไปนี้ คือ เพื่อทำความเข้าใจพฤติกรรมบางอย่าง (Understanding) เพื่อทำการสรุปหรือลดจำนวนปัจจัยให้น้อยลง โดยการจัดทำเป็นข้อสรุปตามกลุ่ม (Summarization) และเพื่อทำการจัดให้เป็นส่วน ๆ (Segmentation) ดังตัวอย่างการใช้งานของการวิเคราะห์การจัดกลุ่มต่อไปนี้

1. การจัดกลุ่มเพื่อเข้าใจพฤติกรรมการซื้อขาย (Understanding buyers' behaviors) โดยการจัดกลุ่มลูกค้าที่มีพฤติกรรมการซื้อที่คล้ายคลึงกันแล้วศึกษาว่าพฤติกรรมเด่น ๆ ของกลุ่มคืออะไร
2. การจัดกลุ่มของหุ้นหรือสต็อก (Stocks) ที่มีลักษณะหรือพฤติกรรมของราคาขึ้นและลงที่คล้ายคลึงกัน

บทที่ 5

การวิเคราะห์ความสัมพันธ์ (Association Analysis)





กฎความสัมพันธ์ (Association Rules)

การวิเคราะห์กฎความสัมพันธ์เป็นการศึกษาหาลักษณะบางอย่างที่ไปในทิศทางเดียวกัน หรือมีความเกี่ยวข้องกัน (Affinity) โดยมีจุดเริ่มต้นจากการวิเคราะห์ข้อมูลการซื้อสินค้า หรือที่รู้จักกันดีในชื่อการวิเคราะห์ตะกร้าซื้อสินค้า (Market basket analysis) ซึ่งคือการวิเคราะห์รายการทั้งหมดที่ลูกค้าซื้อสินค้าต่อครั้ง

การวิเคราะห์กฎความสัมพันธ์เป็นการค้นหาความสัมพันธ์เชิงปริมาณระหว่างลักษณะประจำตั้งแต่ 2 ตัวเป็นต้นไป โดยลักษณะของกฎความสัมพันธ์ที่ได้จะมาในรูปของกฎดังนี้

“If antecedent, then consequent”

หรือใช้สัญลักษณ์

Antecedent --> Consequent

โดย antecedent หมายถึง สิ่งที่มาก่อน และ consequent หมายถึง ผลที่จะเกิดตามมา โดยการที่จะได้กฎความสัมพันธ์จากชุดข้อมูล ซึ่งโดยมากจะเป็นข้อมูลรายการเปลี่ยนแปลง (Transaction Data) โดยใช้เครื่องวัดหรือเกณฑ์การวัดที่เรียกว่า **ค่าสนับสนุน** (Support) และ **ค่าความเชื่อมั่น** (Confidence) ในที่นี้จะยกตัวอย่างการคำนวณค่าทั้งสองแบบคร่าว ๆ ผ่านตัวอย่าง 5.1

ตัวอย่าง 5.1

ถ้าสมมติว่าร้านสะดวกซื้อแห่งหนึ่งมีรายการซื้อสินค้า 1,000 รายการ โดยมีรายการที่มีการซื้อ **ขนมปัง** ทั้งหมด 200 รายการ โดยใน 200 รายการนี้มีการซื้อ **นมข้นหวาน** ทั้งหมด 50 รายการ ถ้ากฎความสัมพันธ์ที่ได้ คือ ถ้าซื้อ **ขนมปัง** แล้วจะซื้อ **นมข้นหวาน** เขียนได้ดังนี้

“If buy **ขนมปัง**, then buy **นมข้นหวาน**”

จงหาค่าสนับสนุนและค่าความเชื่อมั่นของกฎนี้

ค่าสนับสนุนจะมีค่าเท่ากับ $50/1000 = 0.05$ และค่าความเชื่อมั่นจะมีค่าเท่ากับ $50/200 = 0.25$ โดยวิธีการคำนวณของ 2 ค่านี้ ผู้เขียนจะอธิบายต่อไปในภายหลัง

บทที่ 6

การพยากรณ์ (Prediction)





การพยากรณ์ (Prediction) เป็นการนำข้อมูลมาทำนายค่าตอบเช่นเดียวกับการจำแนกที่อธิบายไว้ในบทที่ 2 เพียงแต่ค่าของการพยากรณ์หรือการทำนายจะเป็นค่าแบบต่อเนื่อง (Continuous Value) ซึ่งแตกต่างจากเทคนิคการจำแนกที่ค่าตอบของการทำนายจะเป็นค่าเต็มหน่วย (Discrete Value) หรือที่เรียกว่า คลาส (Class) ที่เป็นการสื่อถึงค่าตอบแบบเต็มหน่วย ขั้นตอนการพัฒนาตัวพยากรณ์จะมีความคล้ายคลึงกับการพัฒนาตัวจำแนก โดยจะมีการแบ่งข้อมูลเป็นข้อมูลฝึกสอนและข้อมูลทดสอบเหมือนกัน แต่สิ่งที่แตกต่างกันคือการวัดประสิทธิภาพของการพยากรณ์หรือความแม่นยำในการพยากรณ์ (Predicted accuracy) ซึ่งจะใช้เกณฑ์การวัดค่าความแม่นยำอีกลักษณะหนึ่งที่ไม่ใช่การวัดร้อยละการจำแนกที่ถูกต้องและเมตริกซ์สับสนเหมือนเทคนิคการจำแนก โดยเกณฑ์การวัดประสิทธิภาพที่นิยมใช้กัน เช่น รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Root mean squared error: RMSE) ความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean absolute error: MAE) เป็นต้น

ตัวอย่างการประยุกต์ใช้เทคนิคการพยากรณ์ เช่น การพยากรณ์จำนวนจุลินทรีย์ที่ระดับอุณหภูมิต่าง ๆ การพยากรณ์เกรดเฉลี่ยของนักศึกษา การพยากรณ์รายรับของบริษัทในปีหน้า การพยากรณ์ราคาทองแท่ง การพยากรณ์ค่าดัชนีราคาตลาดหุ้น เป็นต้น

โดยเทคนิคการสร้างตัวแบบการพยากรณ์นั้นมีด้วยกันหลายวิธี โดยวิธีที่ได้รับการนิยมน้อยมาก ได้แก่ การวิเคราะห์การถดถอย (Regression analysis) ซึ่งเป็นเทคนิคทางสถิติที่มีการใช้กันมาอย่างยาวนาน (Larose, 2006) จนทางวิชาการบางที่จะเรียกงานด้านการพยากรณ์ค่าต่อเนื่องว่า “regression” นอกจากนี้โครงข่ายประสาทเทียม (Artificial Neural Network) เป็นอีกหนึ่งวิธีที่ได้รับความนิยมใช้ในการสร้างตัวแบบการพยากรณ์ โดยเป็นเทคนิคในกลุ่มของการเรียนรู้ด้วยเครื่อง (Machine learning) ในบทนี้จะกล่าวถึงเทคนิคการพยากรณ์ 2 วิธีนี้

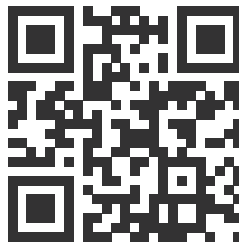
การวิเคราะห์การถดถอย (Regression Analysis)

การวิเคราะห์การถดถอย (Regression analysis) เป็นการศึกษาเพื่อหารูปแบบความสัมพันธ์หรือฟังก์ชันเพื่อใช้ทำนายค่าตัวแปรที่ต้องการศึกษาซึ่งจะเรียกว่าเป็น ตัวแปรตาม (Dependent Variable) หรือตัวแปรตอบสนอง (Output Response) หรือค่าที่เป็นผลลัพธ์ของการทำนาย (Predicted Value) มักแทนด้วยตัวแปร Y โดยอาศัยความรู้เกี่ยวกับค่าของตัวแปรที่เกี่ยวข้องหนึ่งตัวหรือมากกว่าซึ่งจะเรียกว่า ตัวแปรอิสระ (Independent Variable) มักแทนด้วยตัวแปร X



สำนักพิมพ์
มหาวิทยาลัยนเรศวร

สั่งซื้อหนังสือออนไลน์ จัดส่งถึงบ้านสะดวกรวดเร็ว



สั่งซื้อทันที

กรณีต้องการสั่งซื้อหนังสือปริมาณมาก หรือเข้าชั้นเรียนติดต่อได้ที่
ฝ่ายจัดจำหน่ายสำนักพิมพ์มหาวิทยาลัยนเรศวร

✉ nuph@nu.ac.th

📘 สำนักพิมพ์มหาวิทยาลัยนเรศวร

☎ 0 5596 8833-8836

🐦 [nu_publishing](https://www.facebook.com/nu_publishing)



NUPH
online store

www.nupress.grad.nu.ac.th