เพิ่มมูลค่าข้อมูลทางธุรกิจด้วย Data mining

วิชญ์คุทธ์ เมาระพงษ์
ที่ปรึกษาโครงการสารสนเทศของหน่วยงานภาครัฐ
สังกัดสถาบันวิจัยและให้คำปรึกษาแห่งมหาวิทยาลัยธรรมศาสตร์

ต่อ จากฉบับที่แล้ว

2. Classification & Prediction

● Classification เป็นกระบวนการสร้าง Model จัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ ตัวอย่างเช่น จัดกลุ่มนักเรียนว่าดีมาก ดี ปานกลาง ไม่ดี โดยพิจารณาจากประวัติและผลการเรียน หรือแบ่งประเภทของลูกค้าว่าเชื่อถือได้หรือไม่โดยพิจารณาจากข้อมูลที่มีอยู่ กระบวนการ Classification นี้แบ่งออกเป็น 3 ขั้นตอน ได้แก่

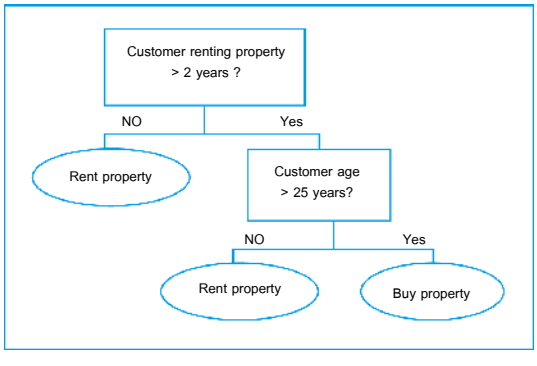
➠ *Model Construction (Learning)* เป็นขั้นการสร้าง Model โดยการเรียนรู้จากข้อมูลที่ได้กำหนดกลุ่มไว้เรียบร้อยแล้ว (Training data) ซึ่ง Model ที่ได้อาจแสดงในรูปของ

1. แบบต้นไม้ (Decision Tree)
2. แบบนิวรอลเน็ต (Neural Net)

1) โครงสร้างแบบต้นไม้ของ Decision Tree เป็นที่นิยมกันมากเนื่องจากเป็นลักษณะที่คนจำนวนมากคุ้นเคย ทำให้เข้าใจได้ง่าย มีลักษณะเหมือนแผนภูมิองค์การ

สมมติว่าองค์การขนาดใหญ่แห่งหนึ่ง ทำธุรกิจอสังหาริมทรัพย์ มีสำนักงานสาขาอยู่ประมาณ 50 แห่ง แต่ละสาขามีพนักงานประจำเป็นผู้จัดการและพนักงานขาย พนักงานเหล่านี้แต่ละคนจะดูแลอาคารต่างๆ หลายแห่ง รวมทั้งลูกค้าจำนวนมาก องค์การจำเป็นต้องใช้ระบบฐานข้อมูลที่กำหนดความสัมพันธ์ระหว่างองค์ประกอบเหล่านี้ เมื่อรวบรวมข้อมูลแบ่งเป็นตารางพื้นฐานต่างๆ เช่น ข้อมูลสำนักงานสาขา (Branch) ข้อมูลพนักงาน (Staff) ข้อมูลทรัพย์สิน (Property) และข้อมูลลูกค้า (Client) พร้อมทั้งกำหนดความสัมพันธ์ (Relationship) ของข้อมูลเหล่านี้ เช่น ประวัติการเช่าบ้านของลูกค้า (Customer rental) รายการให้เช่า (Rentals) รายการขายสินทรัพย์ (Sales) เป็นต้น ต่อมาเมื่อมีประชุมกรรมการผู้บริหารขององค์การส่วนหนึ่งของรายงานจากฐานข้อมูลสรุปว่า

“40% ของลูกค้าที่เช่าบ้านนานกว่าสองปีและมีอายุเกิน 25 ปี จะซื้อบ้านเป็นของตนเอง โดยกรณีเช่นนี้เกิดขึ้น 35% ของลูกค้าผู้เช่าบ้านขององค์การ”

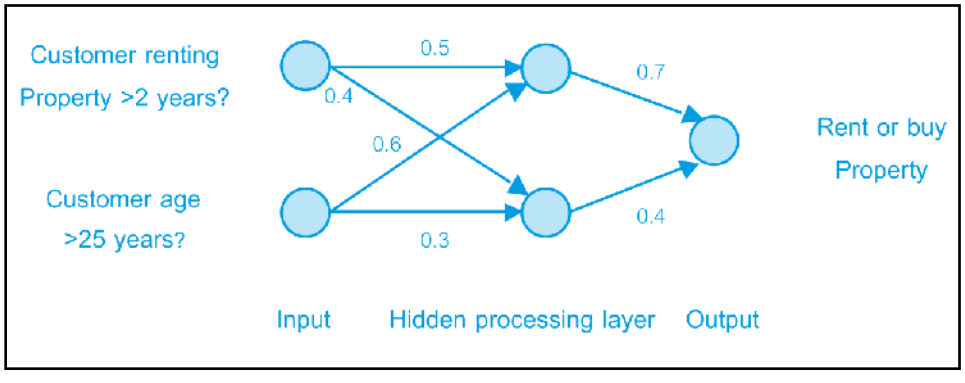


ตัวอย่างของ Decision Tree เพื่อวิเคราะห์โอกาสที่ลูกค้าบ้านเช่าจะซื้อบ้าน

จากภาพแสดงให้เห็นถึง Decision Tree สำหรับการวิเคราะห์ว่าลูกค้าบ้านเช่าจะมีความสนใจที่จะซื้อบ้านเป็นของตนเองหรือไม่ โดยใช้ปัจจัยในการวิเคราะห์ คือ ระยะเวลาที่ลูกค้าได้เช่าบ้านมา และอายุของลูกค้า โดยผลลัพธ์ที่ได้จากแต่ละโหมดของ Decision Tree เรียกว่า AVC set ใช้ในการจัดกลุ่มลูกค้า

2) นิวรอลเน็ต หรือนิวรอลเน็ตเวิร์ก (Neural Net)

เป็นเทคโนโลยีที่มีที่มาจากงานวิจัยด้านปัญญาประดิษฐ์ Artificial Intelligence: AI เพื่อใช้ในการคำนวณค่าฟังก์ชันจากกลุ่มข้อมูล วิธีการของนิวรอลเน็ต (Artificial Neural Networks หรือ ANN) เป็นวิธีการที่ให้ระบบเรียนรู้จากตัวอย่างต้นแบบ แล้วฝึกให้ระบบได้รู้จักที่จะคิดแก้ปัญหาที่กว้างขึ้นได้ ในโครงสร้างของนิวรอลเน็ตจะประกอบด้วยโหมด สำหรับ Input-Output และการประมวลผล กระจายอยู่ในโครงสร้างเป็นชั้นๆ ได้แก่ input layer, output layer และ hidden layers การประมวลผลของนิวรอลเน็ตจะอาศัยการส่งการทำงานผ่านโหมดต่างๆ ใน layer เหล่านี้



ตัวอย่าง นิวรอลเน็ตเพื่อวิเคราะห์พฤติกรรมการเช่าและซื้อบ้านของลูกค้า

➠ *Model Evaluation (Accuracy)* เป็นการประมาณความถูกต้องโดยอาศัยข้อมูลที่ใช้ทดสอบ (testing data) ซึ่งกลุ่มที่แท้จริงของข้อมูลที่ใช้ทดสอบนี้จะถูกนำมาเปรียบเทียบกับกลุ่มของข้อมูลที่หามาได้จาก model เพื่อทดสอบความถูกต้อง

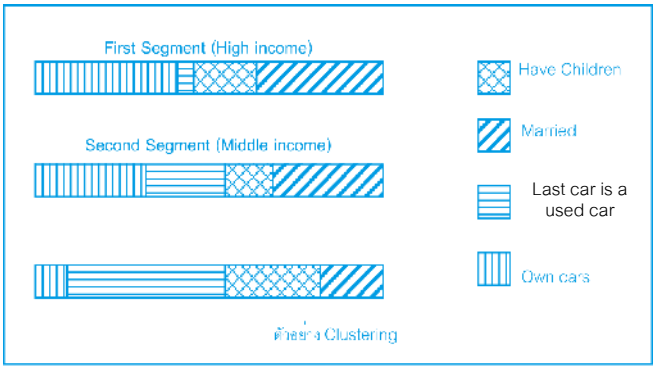
➠ *Model Usage (Classification)* เป็น Model สำหรับใช้กับข้อมูลที่ไม่เคยเห็นมาก่อน (unseen data) โดยจะทำการกำหนดกลุ่มให้กับข้อมูลใหม่ที่ได้มา หรือเป็นการทำนายค่าออกมาตามที่ต้องการ

● **Prediction** เป็นการทำนายหาค่าที่ต้องการจากข้อมูลที่มีอยู่ ตัวอย่างเช่น หายอดขายของเดือนถัดไปจากข้อมูลที่มีอยู่ หรือทำนายโรคจากอาการของคนไข้ในอดีต เป็นต้น

3. Database clustering หรือ Segmentation

เป็นเทคนิคการลดขนาดของข้อมูลด้วยการรวมกลุ่มตัวแปรที่มีลักษณะเดียวกันไว้ด้วยกัน ตัวอย่างเช่น องค์การจำหน่ายรถยนต์ได้แยกกลุ่มลูกค้าออกเป็น 3 กลุ่ม คือ

1. กลุ่มผู้มีรายได้สูง (>70,000)
2. กลุ่มผู้มีรายได้ปานกลาง (20,000 to 70,000)
3. กลุ่มผู้มีรายได้ต่ำ (Less than 20,000) และภายในแต่ละกลุ่มยังแยกออกเป็น
   - ➢ Have Children (มีบุตร)
   - ➢ Married (สมรสแล้ว)
   - ➢ Last car is a used car (ใช้รถมือสอง)
   - ➢ Own cars (มีรถอยู่แล้ว)



จากข้อมูลข้างต้นทำให้ทางองค์การรู้ว่าเมื่อมีลูกค้าเข้ามาที่องค์การ ควรจะเสนอขายรถประเภทใด เช่น ถ้าเป็นกลุ่มผู้มีรายได้สูงควรจะเสนอรถใหม่ เป็นรถครอบครัวขนาดใหญ่พอสมควร แต่ถ้าเป็นผู้มีรายได้ค่อนข้างต่ำควรเสนอรถมือสอง ขนาดค่อนข้างเล็ก

4. Deviation Detection

เป็นกรรมวิธีในการหาค่าที่แตกต่างไปจากค่ามาตรฐาน (ทางสถิติ) หรือค่าที่คาดคิดไว้ว่าต่างไปมากน้อยเพียงใด โดยทั่วไปมักใช้วิธีการทางสถิติ หรือการแสดงให้เห็นภาพ (Visualization) สำหรับเทคนิคนี้ใช้ในการตรวจสอบลายเซ็นปลอม หรือบัตรเครดิตปลอม รวมทั้งการตรวจหาจุดบกพร่องของชิ้นงานในโรงงานอุตสาหกรรม

5. Link Analysis

จุดมุ่งหมายของ Link Analysis คือ การสร้าง link ที่เรียกว่า “associations” ระหว่าง recode เดียว หรือ กลุ่มของ recode ในฐานข้อมูล link analysis สามารถแบ่งออกเป็น 3 ชนิด คือ

- ➢ associations discovery
- ➢ sequential pattern discovery
- ➢ similar time sequence discovery

อ่านต่อฉบับหน้า