วิธีการทางสถิติสำหรับการทำเหมืองข้อมูล

Statistical Methods for Data Mining

พจนา แววสวัสดิ์ *

บทคัดย่อ

วิธีการทางสถิติเป็นเครื่องมือที่สำคัญสำหรับนำไปใช้ในการวิเคราะห์ข้อมูล และบ่อยครั้งถูกนำมาใช้ร่วมกับการทำเหมืองข้อมูล เพื่อให้ระบบการสนับสนุนการตัดสินใจมีความเชื่อถือมากยิ่งขึ้น สถิติมีความแตกต่างกับการทำเหมืองข้อมูล นั่นคือ การวิเคราะห์ทางสถิติผู้วิเคราะห์จะเป็นผู้กำหนดว่าจะศึกษาข้อมูลที่มีรูปแบบใด และจะใช้ข้อมูลหรือตัวแปรส่วนใดบ้าง แต่การทำเหมืองข้อมูลจะกระทำขั้นตอนต่างๆ เหล่านี้ให้โดยอัตโนมัติด้วยหลักการทางคอมพิวเตอร์ ในการประยุกต์ใช้การทำเหมืองข้อมูล สามารถจัดกลุ่มได้เป็น 2 กลุ่ม คือ กลุ่มที่ใช้การทำเหมืองข้อมูลเพื่อการทำนาย และกลุ่มที่ใช้เพื่อการอธิบาย ทั้งสองกลุ่มนี้ยังต้องอาศัยวิธีการทางสถิติในการวิเคราะห์เพื่อหาความรู้ใหม่ๆ สำหรับนำไปใช้ในการตัดสินใจ วิธีการทางสถิติที่ถูกนำไปใช้ในการทำเหมืองข้อมูล ได้แก่ การวิเคราะห์ความถดถอย การวิเคราะห์ปัจจัย การวิเคราะห์จำแนกประเภท การวิเคราะห์การจัดกลุ่ม การวิเคราะห์ด้วยวิธีแบบเบส์ เป็นต้น นอกจากนี้ในการทำเหมืองข้อมูลยังมีวิธีการสำหรับการวิเคราะห์ ได้แก่ ต้นไม้การตัดสินใจ กฎความสัมพันธ์ และโครงสร้างเส้นประสาท เป็นต้น

Abstract

Statistical methods is the significant tool for data analysis. They are often used with data mining to support the decision making system. Statistics analysis method is different from data mining method as the statistical methods need to specify data and variables. In contrast, data mining can be classified into two groups as data mining for forecasting and data mining for decision making even if there two groups still . The statistics method to analyze for findings and to make a decision. Statistical methods which used for data mining are regression analysis ,factor analysis, discriminant analysis, cluster analysis and bayesian analysis etc. Furthermore,data mining also used to analysis the decision tree, association rules and Neural Networks etc.

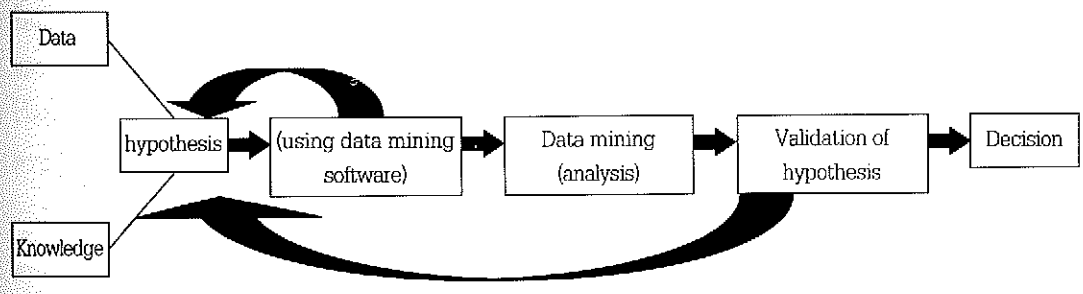
* อาจารย์ประจำสำนักวิจัย มหาวิทยาลัยศรีปทุม

บทนำ

ปัจจุบันข้อมูลส่วนใหญ่ถูกเก็บรวบรวมไว้ในฐานข้อมูลที่มีขนาดใหญ่ และมีการขยายขนาดของข้อมูลอย่างรวดเร็ว เช่น ข้อมูลทางด้านการสื่อสารโทรคมนาคม ข้อมูลทางด้านการบริการ ข้อมูลทางด้านการเงิน ข้อมูลทางด้านวิทยาศาสตร์ เป็นต้น ดังนั้นนักวิเคราะห์จึงมีความจำเป็นในการหาวิธีการวิเคราะห์ข้อมูลเพื่อหาความรู้ใหม่ๆ สำหรับนำไปสร้างระบบการสนับสนุนการตัดสินใจ วิธีการหนึ่งที่สามารถนำมาใช้ในการค้นหา และวิเคราะห์ข้อมูลในปัจจุบันคือ การทำเหมืองข้อมูล (data mining) การทำเหมืองข้อมูลเป็นเทคโนโลยีหนึ่งที่ถูกนำมาใช้ในการสร้างระบบสนับสนุนการตัดสินใจ เนื่องจากการทำเหมืองข้อมูลมีความสามารถในการดำเนินการหรือประยุกต์ใช้กับข้อมูลที่เก็บอยู่ในฐานข้อมูลที่มีขนาดใหญ่ได้ การประยุกต์ใช้ข้อมูลที่กล่าวถึงนี้มีได้หลายแนวทาง แต่โดยทั่วไปมักจะเป็นการสรุปภาพรวมของข้อมูลในฐานข้อมูล การวิเคราะห์แนวโน้มการเปลี่ยนแปลงของข้อมูล หรือการค้นหาความสัมพันธ์ที่ซ่อนอยู่ภายในกลุ่มของข้อมูล

ในทางสถิติการวิเคราะห์ข้อมูลด้วยโปรแกรมสำเร็จรูป เช่น SPSS SAS STATISTICA และ S-PLUS เป็นต้น นักวิเคราะห์จะเป็นผู้กำหนดว่าจะศึกษาข้อมูลที่มีรูปแบบใด และจะใช้ข้อมูลหรือตัวแปรส่วนใดบ้าง แต่การทำเหมืองข้อมูล (data mining) จะกระทำขั้นตอนต่างๆ เหล่านี้ให้โดยอัตโนมัติด้วยหลักการทางคอมพิวเตอร์ การทำเหมืองข้อมูลมีความสามารถในการค้นหาแนวโน้ม รูปแบบร่วม หรือลักษณะอื่นๆ ที่น่าสนใจ โดยไม่ต้องพึ่งพาการสั่งงานทุกขั้นตอนจากนักวิเคราะห์ข้อมูล และอาจจะค้นพบความรู้ใหม่ๆ ที่น่าสนใจสำหรับนำไปใช้ในการตัดสินใจจากข้อมูลที่นักวิเคราะห์ไม่ได้คาดหมายมาก่อน การประยุกต์ใช้การทำเหมืองข้อมูล (data mining) สามารถนำไปประยุกต์ใช้ได้ในหลายด้าน แต่สามารถจัดกลุ่มได้เป็นสองกลุ่ม คือ กลุ่มที่ใช้การทำเหมืองข้อมูล (data mining) เพื่อการทำนาย และกลุ่มที่ใช้เพื่อการอธิบาย การทำเหมืองข้อมูลเพื่อการทำนาย เป็นการนำความรู้ที่ได้มาจากข้อมูลที่มีอยู่ ไปใช้ประโยชน์ในการทำนายข้อมูลใหม่ที่จะเกิดขึ้นในอนาคต ส่วนการทำเหมืองข้อมูลเพื่อการอธิบาย เป็นการนำความรู้ที่ได้มาจากข้อมูล มาใช้อธิบายเรื่องที่เราสนใจ

การทำเหมืองข้อมูลเพื่อการอธิบาย เป็นการค้นหารูปแบบที่น่าสนใจจากกลุ่มข้อมูล รูปแบบนี้มักจะเป็นความสัมพันธ์ หรือลักษณะที่เชื่อมโยงกันของข้อมูล การทำเหมืองข้อมูลแบบนี้ต่างจากแบบแรกตรงที่ผู้ใช้ไม่ได้กำหนดล่วงหน้าว่าจะให้โปรแกรมการทำเหมืองข้อมูลค้นหารูปแบบหรือตัวแบบเป็นรูปแบบใด

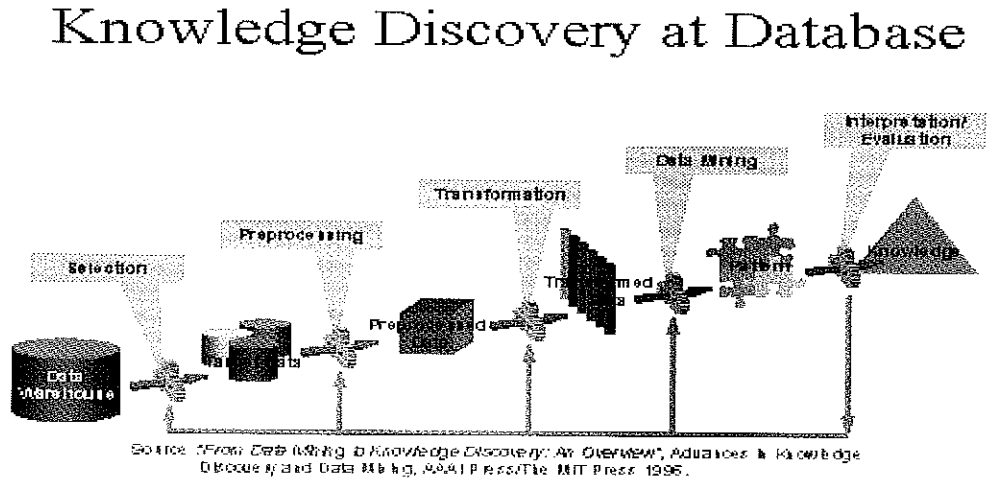


รูปที่ 1 ระบบสนับสนุนการตัดสินใจด้วยการใช้เทคนิคการทำเหมืองข้อมูล

แต่ให้ค้นหาทุกรูปแบบที่น่าสนใจจากข้อมูลจะเห็นได้ว่าการทำเหมืองข้อมูลมีได้หลากหลายรูปแบบ ทั้งนี้เนื่องจากความรู้ที่ต้องการได้จากข้อมูลมีได้หลายลักษณะ การทำเหมืองข้อมูลสามารถนำไปประยุกต์ใช้ในด้านต่างๆ เช่นในทางด้านธุรกิจ มีการนำเอาเทคนิคการทำเหมืองข้อมูลไปวิเคราะห์กลุ่มลูกค้า เพื่อช่วยในการแบ่งกลุ่มในการนำเสนอสินค้าให้ได้ตรงตามกลุ่มเป้าหมาย ใช้ในการวิเคราะห์การออกเงินกู้แก่ลูกค้า เพื่อตัดสินใจว่าควรจะให้เงินกู้แก่ลูกค้าคนใด ใช้แบ่งประเภทของลูกค้าว่าลูกค้าคนใดมีโอกาสเกิดหนี้สูญได้มากกว่ากัน รวมทั้งการใช้เทคนิคการทำเหมืองข้อมูลไปใช้เป็นเครื่องมือในการแก้ไขปัญหาในงานวิจัย ตัวอย่างเช่น ใช้ค้นหาผลข้างเคียงของการใช้ยา ใช้ในการวิเคราะห์หาความสัมพันธ์ของสารพันธุกรรม เป็นต้น

เนื่องจากทุกวันนี้เราสามารถหาข้อมูลที่มีขนาดใหญ่ได้ในรูปแบบอิเล็กทรอนิกส์ และความต้องการข้อมูลที่เร่งด่วน ซึ่งข้อมูลถูกนำมาใช้เป็นข้อสนเทศและความรู้ที่จะนำไปประยุกต์ใช้ในการวิเคราะห์ทางการตลาด การจัดการทางด้านธุรกิจ และสนับสนุนการตัดสินใจ การทำเหมืองข้อมูลเป็นเรื่องหนึ่งที่น่าสนใจอย่างยิ่งในอุตสาหกรรมเกี่ยวกับข้อสนเทศในไม่กี่ปีที่ผ่านมา การทำเหมืองข้อมูลเป็นที่นิยมในการจัดการที่เหมือนกับเรื่องการค้นหาความรู้ในฐานข้อมูล (Knowledge discovery in database) ซึ่งนักวิจัยหลายคนถือว่าการทำเหมืองข้อมูลเป็นขั้นตอนที่สำคัญของการค้นหาความรู้ (Knowledge discovery) โดยทั่วไปการค้นหาความรู้ จะประกอบด้วยขั้นตอนการดำเนินงานดังนี้

1. **Data cleaning** เป็นขั้นตอนในการพิจารณาความซับซ้อน ความคลาดเคลื่อน ความผิดพลาดหรือความแตกต่างของข้อมูล
2. **Data integration** เป็นการรวบรวมแหล่งข้อมูลที่แตกต่างกันเข้าไว้ด้วยกัน
3. **Data selection** เมื่อข้อมูลมีความแตกต่างกันเราจะทำการวิเคราะห์เพื่อดึงข้อมูลที่เราสนใจมาจากฐานข้อมูล
4. **Data transformation** การนำข้อมูลมาทำการแปลงหรือรวบรวมข้อมูลตามความเหมาะสมจากเหมืองข้อมูลก่อนการสรุปและรวบรวม



รูปที่ 2 โครงสร้างการค้นหาความรู้จากฐานข้อมูล

5. **Data mining** เป็นกระบวนการที่จำเป็นอย่างยิ่งที่นำไปประยุกต์ใช้กับการจัดรูปแบบของข้อมูล

6. **Pattern evaluation** เป็นการตรวจสอบรูปแบบที่เราสนใจ เพื่อจะนำเสนอเป็นความรู้บนพื้นฐานค่าที่ถูกวัด

7. **Knowledge presentation** เป็นเทคนิคในการนำเสนอภาพลักษณ์และความรู้

การทำเหมืองข้อมูล (Data mining)

เหตุผลสำคัญที่การทำเหมืองข้อมูลเป็นที่ให้ความสนใจอย่างมากในปัจจุบัน เนื่องมาจากความสามารถในการพิจารณาข้อมูลขนาดใหญ่และความต้องการที่เร่งด่วนของข้อมูล เพื่อนำข้อมูลมาใช้เป็นข้อสนเทศและข้อความรู้ เพื่อสนับสนุนระบบการตัดสินใจ ความได้เปรียบของข้อสนเทศและข้อความรู้สามารถนำมาประยุกต์ใช้ในการจัดการทางด้านธุรกิจ ควบคุมการผลิต และการวิเคราะห์ตลาด ไปจนถึงการออกแบบทางด้านวิศวกรรม และการค้นคว้าทางวิทยาศาสตร์ การทำเหมืองข้อมูลเป็นวิธีการที่อาศัยข้อมูลจำนวนมาก ด้วยหลักการที่ว่าข้อมูลยิ่งมากเท่าใด ผลการวิเคราะห์ก็น่าเชื่อถือมากยิ่งขึ้น

การทำเหมืองข้อมูล สามารถนำมาใช้เป็นผลจากวิวัฒนาการทางเทคโนโลยีข้อสนเทศ ในส่วนของการวิวัฒนาการจะเป็นหลักฐานสำหรับอุตสาหกรรมของข้อมูล การทำเหมืองข้อมูลมีนักวิชาการได้ให้ความหมายของการทำเหมืองข้อมูล โดยได้นิยามไว้ในสาขาต่างๆ ไว้อย่างคลุมเครือ โดยการให้คำนิยามนั้นขึ้นอยู่กับพื้นความรู้ของผู้กำหนดจากการศึกษา มีผู้ให้คำนิยามไว้ดังนี้

การทำเหมืองข้อมูล คือ กระบวนการในการตรวจสอบความถูกต้อง ความแปลกใหม่ ความเป็นไปได้ของการใช้ประโยชน์ และความเข้าใจในรูปแบบของข้อมูล - Fayyad

การทำเหมืองข้อมูล คือ กระบวนการในการขจัดความไม่รู้ การสรุปอย่างกว้างๆ และการดำเนินการเกี่ยวกับข้อสนเทศจากฐานข้อมูลขนาดใหญ่และนำไปใช้ในการตัดสินใจทางธุรกิจ - Zekolin

การทำเหมืองข้อมูล คือ วิธีการที่ถูกใช้ในการค้นหาความรู้ และนำไปจำแนกความสัมพันธ์และรูปแบบภายในข้อมูลที่เราไม่ทราบ - Ferruzza

การทำเหมืองข้อมูล คือ กระบวนการในการค้นหาจุดเด่นของข้อมูล - John

การทำเหมืองข้อมูล คือ กระบวนการในการสนับสนุนการตัดสินใจเมื่อข้อมูลมีขนาดใหญ่บนพื้นฐานของความไม่รู้และรูปแบบของข้อสนเทศ - Parsaye

ผู้เขียนสามารถสรุปได้ว่า *การทำเหมืองข้อมูล คือ กระบวนการในการค้นหาข้อมูลที่สนใจ เช่น รูปแบบความสัมพันธ์ การเปลี่ยนแปลง โครงสร้างที่น่าเชื่อถือได้ รวมทั้งความรู้ใหม่ๆ ของข้อมูลจากแหล่งข้อมูลที่มีขนาดใหญ่ เช่น ฐานข้อมูล หรือแหล่งเก็บรวบรวมข้อสนเทศอื่นๆ*

จากความหมายข้างต้นสามารถสรุปได้ว่า การทำเหมืองข้อมูล คือ กระบวนการในการค้นหาตัวแบบความสัมพันธ์ และการประมาณค่าจากข้อมูลที่มีอยู่ หรืออาจกล่าวได้ว่าการทำเหมืองข้อมูลเป็นกระบวนการในการวิเคราะห์เพื่อกำหนดลักษณะของข้อมูล โดยปกติแล้วจะเป็นข้อมูลที่มีขนาดใหญ่ เช่น ข้อมูลทางด้านธุรกิจ หรือ ข้อมูลทางด้านการตลาด เป็นต้น ดังนั้นการทำเหมืองข้อมูลเป็นการค้นหาความสอดคล้องหรือความสัมพันธ์ระหว่างตัวแปรในฐานข้อมูล เป้าหมายหลักของการทำเหมืองข้อมูลมีเป้าหมายที่สำคัญ 2 เรื่องด้วยกัน คือ การทำนาย(prediction) และการบรรยาย (Description) ในการทำนาย (prediction) เป็นการใช้ตัวแปรในฐาน

ข้อมูลมาใช้ในการทำนายเหตุการณ์ที่เราไม่ทราบ หรือนำมาใช้ในการหาค่าในตัวแปรที่เราสนใจในอนาคต สำหรับการบรรยาย (Description) จะเป็นการบรรยายลักษณะของข้อมูลเพื่อให้เข้าใจมากยิ่งขึ้น

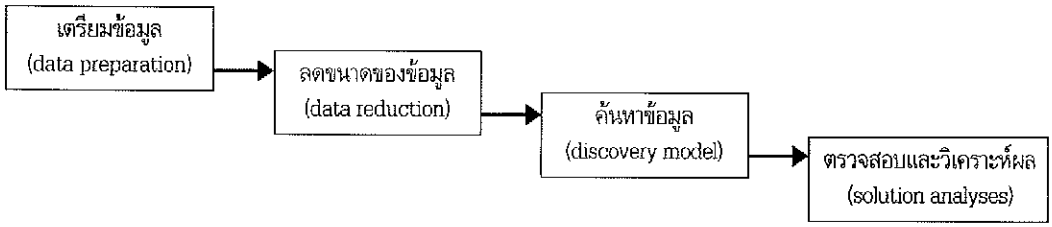
ในการทำเหมืองข้อมูลเป็นการค้นหาความรู้จากฐานข้อมูลที่เรามีอยู่ ซึ่งจะช่วยให้เราเข้าใจลักษณะของข้อมูล รวมทั้งสามารถนำข้อมูลมาใช้ในการทำนายหรือทราบลักษณะแนวโน้มที่จะเกิดขึ้นในอนาคต โดยทั่วไปแล้วการทำเหมืองข้อมูลประกอบด้วย 4 ขั้นตอน ด้วยกัน คือ

ขั้นตอนที่ 1 เตรียมข้อมูล (data preparation) ถ้าข้อมูลไม่อยู่ในรูปแบบที่ถูกต้องหรือเหมาะสม จะต้องมีการปรับข้อมูลให้อยู่ในรูปแบบที่โปรแกรมการทำเหมืองข้อมูล จะเรียกใช้งานได้ โดยทั่วไปข้อมูลที่ถูกนำมาใช้ในการทำเหมืองข้อมูลอยู่ในรูปของ Data warehouse หรืออยู่ในรูปของ Spreadsheet ข้อมูลที่นำมาใช้ต้องเป็นข้อมูลที่ไม่ได้มีการสรุปมาแล้ว ตัวแปรที่นำมาใช้ในการวิเคราะห์ผู้วิเคราะห์ต้องทราบมาตรวัดของตัวแปรแต่ละตัวว่าอยู่ในมาตรวัดใด ในการเตรียมข้อมูลสำหรับการทำเหมืองข้อมูลประกอบด้วย การตรวจสอบข้อมูล เพื่อดูว่าข้อมูลที่มีอยู่ควรแก้ไขหรือไม่ การบรรณาธิกรณ์ข้อมูล และการทำความสะอาดข้อมูล

ขั้นตอนที่ 2 ลดขนาดของข้อมูล (data reduction) การจะหาตัวแบบหรือรูปแบบที่ข้อมูลส่วนใหญ่แสดงลักษณะเหล่านั้นออกมาเหมือนกัน จำเป็นต้องใช้ข้อมูลตัวอย่างจำนวนมาก ถ้าข้อมูลน้อยเกินไปอาจจะหาลักษณะร่วมเหล่านั้นไม่พบ แต่ในทางตรงกันข้ามถ้าข้อมูลมีปริมาณมากเกินไป การค้นหาตัวแบบหรือรูปแบบจากกลุ่มข้อมูลขนาดใหญ่ต้องใช้เวลามาก ซึ่งถ้าลดจำนวนข้อมูลลงด้วยสัดส่วนที่ถูกต้อง โมเดลที่ได้ยังคงเป็นเช่นเดิมในขณะที่โปรแกรมใช้เวลาในการค้นหาโมเดลสั้นลง

การลดขนาดของข้อมูลทำได้ในสองลักษณะ คือ ลดจำนวนเรคคอร์ด และลดจำนวนตัวแปร ของแต่ละค่าของข้อมูล ข้อมูลที่ผ่านการลดขนาดแล้วจะถูกแบ่งออกเป็นสองส่วน ส่วนแรกใช้ในกระบวนการ ค้นหาแพทเทิร์น หรือความสัมพันธ์จากข้อมูล เรียกข้อมูลส่วนนี้ว่า training set (ส่วนที่ใช้ในการฝึกฝน) ส่วนที่สองใช้ตรวจสอบความถูกต้องของรูปแบบ เรียกข้อมูลส่วนนี้ว่า test set(ส่วนที่ใช้ในการทดสอบ)

ขั้นตอนที่ 3 ค้นหาโมเดลจากข้อมูล (data modeling/discovery) กระบวนการค้นหาโมเดลหรือความสัมพันธ์จะเริ่มจากข้อมูลเริ่มต้นจำนวนไม่มากนัก จากนั้นนำผลที่ได้จากกระบวนการค้นหา (learning process/method) ไปยืนยันกับข้อมูลทดสอบ ถ้าผลที่ได้ยังไม่น่าพอใจอาจจะต้องปรับค่าพารามิเตอร์บางตัวของ learning method และเริ่มกระบวนการค้นหาใหม่กับข้อมูลจำนวนมากขึ้น จนกว่าผลที่ได้มีความถูกต้องอยู่ในระดับที่ยอมรับได้ จึงจะจบกระบวนการค้นหา



รูปที่ 3 ขั้นตอนการทำเหมืองข้อมูล

ขั้นตอนที่ 4 ตรวจสอบและวิเคราะห์ผล (solution analyses) ตัวแบบหรือความสัมพันธ์ที่หามาได้ในขั้นตอนที่ 3 จะต้องถูกนำมาทดสอบอัตราความผิดพลาดและวิเคราะห์ความซับซ้อนของรูปแบบโมเดล ถ้าอัตราความผิดพลาดยังสูงเกินไป อาจจะต้องย้อนกลับไปที่ขั้นตอนที่ 3 อีกครั้ง เพื่อปรับปรุงตัวแบบให้ถูกต้องยิ่งขึ้น ในทำนองเดียวกัน ถ้าตัวแบบที่หามาได้มีรูปแบบที่ซับซ้อนเกินไปจนยากต่อการทำความเข้าใจ อาจจะต้องย้อนกระบวนการกลับไปที่ขั้นตอนที่ 3 เพื่อให้หาตัวแบบใหม่ที่มีความถูกต้องเท่าเดิมแต่มีรูปแบบที่ซับซ้อนน้อยลง

สถิติสำหรับการทำเหมืองข้อมูล

สถิติเป็นสาขาหนึ่งของคณิตศาสตร์ ที่ถูกนำไปใช้ในเรื่องการเก็บรวบรวมข้อมูลและใช้อธิบายคุณลักษณะของข้อมูล โดยปกติแล้วสถิตินิยมนำมาใช้ในการวิเคราะห์ข้อมูล ในหลายๆ เรื่อง สถิตินับว่าเป็นเครื่องมือที่มีความจำเป็นอย่างยิ่งต่อการนำไปใช้ในการตัดสินใจทางด้านธุรกิจ เพื่อช่วยลดความเสี่ยงและความไม่แน่นอนที่เกิดขึ้นก่อนการตัดสินใจ ซึ่งในกรณีข้อมูลที่เก็บรวบรวมอยู่มีขนาดใหญ่การตัดสินใจในทางธุรกิจก็ยังคงอยู่บนพื้นฐานของการคาดคะเน ดังนั้นถ้าเราเก็บรวบรวมข้อมูลมาอย่างถูกต้องและมีปริมาณมากพอทำให้การคำนวณหรือการวิเคราะห์ทางสถิติมีความแม่นยำและถูกต้องมากยิ่งขึ้น

คำว่า **"สถิติ" (statistics)** หรือเทคนิคทางสถิติไม่ใช่เป็นการทำเหมืองข้อมูล เพียงแต่สถิติถูกนำมาใช้ก่อนการทำเหมืองข้อมูล อย่างไรก็ตามเทคนิคทางสถิติจะถูกนำมาใช้ในการดำเนินการเกี่ยวกับการวิเคราะห์ข้อมูลสำหรับการทำเหมืองข้อมูล โดยจะถูกนำมาใช้ในการค้นหารูปแบบและสร้างตัวแบบสำหรับใช้ในการคาดคะเน

วิธีการทางสถิติที่ใช้ในการทำเหมืองข้อมูล

การหาความน่าจะเป็นในทางสถิติเป็นการพิจารณาบนพื้นฐานของข้อมูลที่มีอยู่ แต่ในบางครั้งอาจจะพิจารณาจากประสบการณ์ของผู้วิเคราะห์ เนื่องจากปัจจุบันเรามีเครื่องคอมพิวเตอร์ที่ถูกนำมาใช้เป็นเครื่องมือในการคำนวณ รวมทั้งข้อมูลที่มีอยู่ในปัจจุบันถูกเก็บรวบรวมไว้ในรูปของฐานข้อมูลขนาดใหญ่ ดังนั้นข้อมูลในฐานข้อมูลไม่ได้มาจากการสุ่มหรือมาจากความไม่แน่นอน ผู้วิเคราะห์อาจจะไม่จำเป็นต้องการอนุมานในกระบวนการการวิเคราะห์ ยกเว้นถ้าข้อมูลที่เราเก็บรวบรวมมามีการสุ่มขึ้นมาทำการวิเคราะห์เพียงบางส่วน เราจึงอาศัยการอนุมานมาช่วยในการวิเคราะเพื่อหาคำตอบที่เราสนใจ สถิติที่นำมาใช้ในการทำเหมืองข้อมูลส่วนใหญ่ถูกนำมาใช้ในการประมาณค่าเพื่อหาคำตอบในสิ่งที่ต้องการทราบ ภายใต้ความสามารถของเครื่องคอมพิวเตอร์ ในบางครั้งการใช้ข้อมูลทั้งหมดทำให้ผู้วิเคราะห์ ไม่สามารถหาคำตอบได้ ดังนั้นจึงจำเป็นต้องมีการสุ่มตัวอย่างและใช้วิธีการประมาณในการหาคำตอบ

การวิเคราะห์ข้อมูลด้วยวิธีการทางสถิติเป็นที่ทราบกันดีว่าสถิติเป็นกระบวนการหนึ่งในการทำเหมืองข้อมูล ซึ่งในอดีตการประยุกต์ใช้คอมพิวเตอร์สำหรับการวิเคราะห์ข้อมูลถูกพัฒนาโดยนักสถิติ สถิติจึงเป็นเครื่องมือที่ถูกนำไปใช้สำหรับการทำเหมืองข้อมูลที่สำคัญและเป็นตัวสนับสนุนกระบวนการในการทำเหมืองข้อมูล แต่อาจจะไม่ได้นำมาใช้ในเรื่องต่างๆ ทั้งหมด ในเทคนิคการทำเหมืองข้อมูลส่วนใหญ่จะเป็นเทคนิคที่สร้างขึ้นมาเพื่อวิเคราะห์กับข้อมูลที่มีขนาดใหญ่ มีความซับซ้อนและมีหลายมิติ ดังนั้นเพื่อให้การวิเคราะห์มีความน่าเชื่อถือ จึงจำเป็นต้องอาศัยวิธีการทางสถิติมาช่วยในการวิเคราะห์ข้อมูลเชิงตัวเลข โดยเทคนิคเหล่านี้จะถูกนำไปใช้ในการวิเคราะห์ข้อมูลเชิงวิทยาศาสตร์ เช่น ข้อมูลที่ได้จากการทดลอง ข้อมูลทางด้านเศรษฐศาสตร์ และ

ข้อมูลทางด้านสังคมศาสตร์ เทคนิคทางสถิติที่ถูกนำมาใช้ในการทำเหมืองข้อมูลมีดังต่อไปนี้

การวิเคราะห์ข้อมูลเบื้องต้น โดยทั่วไปจะใช้การวัดแนวโน้มเข้าสู่ส่วนกลาง เช่น ค่าเฉลี่ย ค่ามัธยฐาน และค่าฐานนิยม ในการวัดตำแหน่งกลางของข้อมูล รวมทั้งการวัดการกระจาย เช่น ความแปรปรวน และค่าเบี่ยงเบนมาตรฐาน รวมทั้งการตรวจสอบการกระจายด้วยรูปภาพในลักษณะต่างๆ ส่วนใหญ่ข้อมูลที่นำมาพิจารณาจะเป็นข้อมูลเชิงตัวเลข

การวิเคราะห์ความถดถอย (Regression Analysis) เป็นวิธีการที่นำมาใช้ในการพยากรณ์ค่าของตัวแปรตามจากตัวแปรอิสระที่นำมาใช้ในการวิเคราะห์ ตัวแปรที่นำมาใช้ต้องเป็นตัวเลข ตัวแบบของการวิเคราะห์ความถดถอยมีอยู่หลายแบบ เนื่องจากการวิเคราะห์นั้นต้องเลือกตัวแบบที่เหมาะสมมาใช้ในการวิเคราะห์ เช่น การวิเคราะห์ความถดถอยเชิงเส้น การวิเคราะห์ความถดถอยเชิงซ้อน การวิเคราะห์ความถดถอยแบบสองขั้น การวิเคราะห์ความถดถอยแบบถ่วงน้ำหนัก การวิเคราะห์ความถดถอยแบบพหุนาม และการวิเคราะห์ความถดถอยโลจิสติค เป็นต้น

การวิเคราะห์ความถดถอยต้นไม้ (Regression Trees Analysis) เป็นวิธีการที่นำไปใช้ในการจำแนกหรือ พยากรณ์ในเรื่องที่เราสนใจ โครงสร้างของต้นไม้เป็นข้อมูลที่อยู่ในลักษณะเชิงกลุ่ม การวิเคราะห์ความถดถอยต้นไม้ เหมือนกับการตัดสินใจแผนผังต้นไม้ โดยความแตกต่างของข้อมูลในแต่ละกลุ่มจะบอกด้วยระดับของใบ

การวิเคราะห์ความแปรปรวน (ANOVA) เป็นวิธีการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเชิงปริมาณกับตัวแปรเชิงกลุ่ม โดยตัวแปรตามเป็นตัวแปรเชิงปริมาณ ส่วนตัวแปรอิสระเป็นตัวแปรเชิงกลุ่ม เพื่อใช้ในการทดสอบความแตกต่างค่าเฉลี่ยของตัวแปรตามแยกตามค่าตัวแปรอิสระ เช่น การทดสอบความแตกต่าง ยอดขายของโทรศัพท์มือถือยี่ห้อต่างๆ ซึ่งในที่นี้ตัวแปรตามคือยอดขาย ส่วนตัวแปรอิสระคือยี่ห้อของโทรศัพท์มือถือ โดยในการวิเคราะห์ความแปรปรวนแบ่งเป็น 2 กลุ่ม คือ การวิเคราะห์ความแปรปรวนของตัวแปรตาม 1 ตัว และ การวิเคราะห์ความแปรปรวนของตัวแปรตามหลายตัว

การวิเคราะห์ปัจจัย (Factor Analysis) เป็นเทคนิคที่ใช้ในการจัดกลุ่มหรือรวมตัวแปรที่มีความสัมพันธ์กันไว้ในกลุ่มเดียวกัน หรือ ปัจจัยเดียวกัน ตัวแปรที่อยู่ในปัจจัยเดียวกันจะมีความสัมพันธ์กันมาก โดยความสัมพันธ์นั้นอาจเป็นได้ทั้งทิศทางบวก และทิศทางลบ นอกจากนี้การวิเคราะห์ปัจจัยยังช่วยลดจำนวนตัวแปรหลายตัวให้เหลือปัจจัยที่สำคัญ โดยที่จำนวนปัจจัยจะมีจำนวนน้อยกว่าจำนวนตัวแปร ในการวิเคราะห์ข้อมูลบางเรื่องใช้การวิเคราะห์ปัจจัย ทำการตรวจสอบความถูกต้อง สำหรับกำหนดความสำคัญหรือน้ำหนักให้กับตัวแปร รวมทั้งใช้การวิเคราะห์ปัจจัยแก้ไขปัญหาการที่ตัวแปรอิสระในการวิเคราะความถดถอยมีความสัมพันธ์กัน

การวิเคราะห์จำแนกประเภท (Discriminant Analysis) เป็นเทคนิคการวิเคราะห์ที่ใช้ในการแบ่งกลุ่มในเรื่องที่เราสนใจ โดยทำการแบ่งเป็นกลุ่มย่อยๆ ตั้งแต่ 2 กลุ่มขึ้นไป โดยตัวแปรตามที่นำมาใช้ในการวิเคราะห์ข้อมูลเป็นข้อมูลเชิงกลุ่ม ซึ่งอาจมีค่าเป็น 1 และ 0 หรือมีค่าเป็น 1 2 3 หรือมากกว่าได้ และตัวแปรอิสระเป็นตัวแปรที่มีการวัดเชิงปริมาณ เทคนิคนี้จะถูกนำมาใช้เมื่อต้องการทราบว่าตัวแปรอิสระต่างๆ ที่มีอยู่คาดว่ามีความสัมพันธ์กับตัวแปรตาม จะมีความสามารถนำมาใช้เป็นตัวแบ่งกลุ่มของตัวแปรตามได้ถูกต้องหรือไม่

การวิเคราะห์จัดกลุ่ม (Cluster Analysis) เป็นเทคนิคที่ต้องการจัดกลุ่มซึ่งอาจจะเป็นการจัดกลุ่ม (ค่าสังเกต) โดยพยายามแยกให้เห็นว่าค่าสังเกตได้ควรจะรวมอยู่ในการวิเคราะห์ใด แนวคิดหลักของเทคนิคนี้จะจัดสิ่งที่มีลักษณะคล้ายคลึงกันอยู่ในกลุ่มเดียวกัน ลักษณะความคล้ายกันนี้อาจจะคล้ายกันในบางลักษณะเท่านั้น ในการจัดกลุ่มผู้วิเคราะห์จะไม่ทราบว่าค่าของข้อมูลหรือค่าสังเกตอยู่ในกลุ่มใดมาก่อน ซึ่งการจะจัดกลุ่มได้ถูกต้องนั้น ผู้วิเคราะห์จะต้องเก็บรวบรวมตัวแปรที่มีความสำคัญและเกี่ยวข้องกับเรื่องที่เราในใจได้ครบ

การวิเคราะห์อนุกรมเวลา (Time Series Analysis) การประมาณหรือการคาดเดาเหตุการณ์ที่จะเกิดขึ้นในอนาคต เช่น การประมาณความต้องการของสินค้าหรือบริการ ความต้องการด้านแรงงานในอนาคต เป็นต้น ในการตัดสินใจทางธุรกิจนั้นมักจะเกี่ยวข้องกับการเลือกทางเลือกที่จะนำไปปฏิบัติ โดยการประเมินค่าผลลัพธ์ที่ได้จากทางเลือกนั้นๆ คุณภาพของการตัดสินใจส่วนใหญ่ขึ้นกับคุณภาพในการพยากรณ์ การพยากรณ์จึงมักเป็นส่วนหนึ่งในระบบสนับสนุนการตัดสินใจ เพื่อใช้ในการพยากรณ์ค่าของตัวแปรในอนาคต โดยการพยากรณ์เชิงปริมาณจะเหมาะสมกับสถานการณ์ที่มีข้อมูลในอดีตอยู่แล้ว เป็นข้อมูลที่สามารถทำให้อยู่ในรูปของตัวเลขได้ และเป็นเหตุการณ์ที่สามารถ สมมติได้ว่า แบบแผนในอดีตบางอย่างยังจะดำเนินต่อไปในอนาคต วิธีการพยากรณ์มีทั้งแบบเป็นทางการ และแบบไม่เป็นทางการในที่นี้จะกล่าวถึงเฉพาะแบบเป็นทางการเท่านั้น ได้แก่

1. วิธีใช้วิจารณญาณ (Judgment Methods) เป็นวิธีการพยากรณ์โดยการประมาณที่เป็นนามธรรม และพยากรณ์โดยใช้ความคิดเห็นของผู้เชี่ยวชาญใช้สำหรับการพยากรณ์ในระยะยาว โดยเฉพาะอย่างยิ่งในกรณีที่มีปัจจัยภายนอกเข้ามามีบทบาทเกี่ยวข้องมากๆ และใช้ในกรณีที่มีข้อมูลเดิมค่อนข้างจำกัด หรือไม่มีข้อมูลเดิมอยู่เลย

2. วิธีการนับ (Counting Methods) เป็นวิธีการพยากรณ์ที่เกี่ยวข้องกับการทำการทดลองหรือการสำรวจเพื่อสุ่มตัวอย่างขึ้นมาเพื่อใช้เป็นตัวแทนของกลุ่มข้อมูลทางการทางตลาดทั้งหมด เป็นวิธีที่ใช้สำหรับการพยากรณ์ความต้องการของผลิตภัณฑ์และบริการ โดยใช้ข้อมูลที่มีอยู่จริง และข้อมูลที่มีอยู่แล้วในอดีต

3. การวิเคราะห์อนุกรมเวลา (Time-series Analysis) อนุกรมเวลา (Time Series) เป็นเซตของค่าตัวแปรทางธุรกิจหรือทางเศรษฐกิจซึ่งวัดในช่วงเวลาที่ประสบผลสำเร็จอย่างต่อเนื่อง การใช้วิธีการพยากรณ์แบบนี้ในการตัดสินใจเพราะเชื่อว่าความรู้จากพฤติกรรมในอดีต อาจช่วยให้เข้าใจพฤติกรรมที่จะเกิดในอนาคตได้

4. วิธีการหาความสัมพันธ์ หรือวิธีการอย่างไม่เป็นทางการ (Association or Causal Methods) เป็นการพยากรณ์โดยทำการวิเคราะห์ข้อมูลเพื่อหาความสัมพันธ์ของข้อมูล และความสัมพันธ์ระหว่างสาเหตุและผลลัพธ์ที่เกิดขึ้น โดยวิธีนี้มีประสิทธิภาพและซับซ้อนกว่าแบบวิธีการวิเคราะห์ชุดเวลา เพราะมีตัวแปรเข้ามาเกี่ยวข้องมากกว่า ใช้วิธีการทางสถิติในการแยกแยะประเภทของตัวแปร และเป็นวิธีที่เหมาะกับการพยากรณ์ในช่วงระยะเวลาปานกลาง

การวิเคราะห์ด้วยวิธีการแบบเบส์ (Bayesian Analysis) วิธีการนี้เป็นวิธีการหนึ่งที่ถูกนำไปใช้ร่วมกับข้อสนเทศ (Information) ของข้อมูลที่นำมาใช้ในการวิเคราะห์ ซึ่งการวิเคราะห์ด้วยวิธีการแบบเบส์จะเริ่มจากการกำหนดการแจกแจงความน่าจะเป็นสำหรับการวิเคราะห์ การแจกแจงที่เรากำหนดให้กับข้อมูลก่อนหน้านี้เราเรียกว่า การแจกแจงก่อนปรับ (prior distribution) และข้อมูลชุดใหม่ที่ถูกปรับด้วย การแจกแจงก่อนปรับ (prior distribution) เราเรียกว่า การแจกแจงหลังปรับ (posterior distribution) การแจกแจงข้างต้นถูกนำมาใช้ในการจัดกลุ่มของข้อมูล

วิธีการทางด้านคอมพิวเตอร์ สำหรับการทำเหมืองข้อมูล

ต้นไม้การตัดสินใจ (Decision trees) ต้นไม้การตัดสินใจเป็นวิธีการหนึ่งของการทำเหมืองข้อมูลที่ถูกนำไปใช้เป็นเครื่องมือในการจำแนกข้อมูล การสร้างต้นไม้การตัดสินใจ สามารถทำได้ 2 ขั้นตอน คือ

1. การสร้างต้นไม้การตัดสินใจจากข้อมูลฝึกฝน (Training Data) และทดสอบความถูกต้องของต้นไม้ด้วยข้อมูลทดสอบ (Testing Data)

2. ใช้ต้นไม้การตัดสินใจในการแยกกลุ่มของข้อมูล ข้อดีของการใช้ต้นไม้ในการตัดสินใจนั้น ผู้วิเคราะห์สามารถใช้ได้ง่ายและรวดเร็ว ซึ่งกฎที่สร้างขึ้นจากต้นไม้การตัดสินใจง่ายต่อการเข้าใจ ความเร็วในการจำแนกข้อมูลไม่ได้ขึ้นกับขนาดของฐานข้อมูล ส่วนข้อเสียของวิธีการนี้ พบว่า ไม่สามารถใช้ได้กับข้อมูลที่มีค่าอยู่ในลักษณะต่อเนื่อง ดังนั้นค่าของตัวแปรต้องถูกนำมาแบ่งออกเป็นช่วง หรือเป็นกลุ่ม กรณีที่มีข้อมูลสูญหาย จะส่งผลให้วิธีการนี้ไม่สามารถจำแนกกลุ่มข้อมูลได้ถูกต้อง รวมทั้งวิธีการนี้จะไม่คำนึงถึงความสัมพันธ์ระหว่างตัวแปร การสร้างต้นไม้การตัดสินใจมีหลายแบบ เช่น C4.5 C5.0 และ CART ทั้งนี้ขึ้นอยู่กับลักษณะของข้อมูล

กฎความสัมพันธ์ (Association Rules) กฎความสัมพันธ์เป็นวิธีการหนึ่งของการทำเหมืองข้อมูลที่สำคัญ และสามารถนำไปประยุกต์ใช้ได้จริงกับงานต่างๆ หลักการทำงานของวิธีนี้ คือ การค้นหาความสัมพันธ์ของข้อมูลจากข้อมูลขนาดใหญ่ที่มีอยู่เพื่อนำไปใช้ในการวิเคราะห์ หรือทำนายปรากฏการณ์ต่างๆ หรือมาจากการวิเคราะห์การซื้อสินค้าของลูกค้าเรียกว่า **"Market Basket Analysis"** ซึ่งประเมินจากข้อมูลในตารางที่รวบรวมไว้ ผลการวิเคราะห์ที่ได้จะเป็นคำตอบของปัญหา ซึ่งการวิเคราะห์แบบนี้เป็นการใช้ **"กฎความสัมพันธ์"** เพื่อหาความสัมพันธ์ของข้อมูล

ตัวอย่างการนำเทคนิคนี้ไปประยุกต์ใช้กับงานจริง ได้แก่ การพัฒนาระบบแนะนำหนังสือให้กับลูกค้าแบบอัตโนมัติ ของ Amazon ข้อมูลการสั่งซื้อทั้งหมดของ Amazon ซึ่งมีขนาดใหญ่มากจะถูกนำมาประมวลผลเพื่อหาความสัมพันธ์ของข้อมูล นั่นคือ ลูกค้าที่ซื้อหนังสือเล่มหนึ่งๆ มักจะซื้อหนังสือเล่มใดพร้อมกันด้วยเสมอ ความสัมพันธ์ที่ได้จากกระบวนการนี้จะสามารถนำไปใช้คาดการณ์ได้ว่าควรแนะนำหนังสือเล่มใดเพิ่มเติมให้กับลูกค้าที่เพิ่งซื้อหนังสือจากร้าน ตัวอย่างเช่น เมื่อลูกค้าซื้อหนังสือฐานข้อมูล (database) แล้วมีโอกาสที่จะซื้อหนังสือการทำเหมืองข้อมูล (data mining) ด้วย และมีการซื้อทั้งหนังสือฐานข้อมูล (database) และหนังสือการทำเหมืองข้อมูล (data mining) พร้อมๆ กัน

นอกจากนี้การทำเหมืองข้อมูลยังถูกนำไปประยุกต์ใช้ในการซื้อสินค้าของลูกค้า 1 ครั้ง โดยไม่ต้องจำกัดว่าจะซื้อสินค้าในห้างร้าน หรือสั่งผ่านทางไปรษณีย์ หรือการซื้อสินค้าจากร้านค้าเสมือนจริงบน web โดยปกติเราจะต้องการทราบว่าสินค้าใดบ้างที่ลูกค้ามักซื้อด้วยกัน เพื่อนำไปพิจารณาปรับปรุงการจัดวางสินค้าในร้าน หรือใช้เพื่อหาวิธีวางรูปคู่กันในใบโฆษณาสินค้า ก่อนอื่นขอกำหนดคำว่า กลุ่มรายการ (itemset) หมายถึงกลุ่มสินค้าที่ปรากฏร่วมกัน เช่น {รองเท้า ถุงเท้า} {ปากกา หมึก} หรือ {นม น้ำผลไม้} โดยกลุ่มรายการดังกล่าวนี้ อาจจะจับคู่กลุ่มลูกค้ากับสินค้าก็ได้เช่น วิเคราะห์หา "ลูกค้าที่ซื้อสินค้าบางชนิดซ้ำๆ กัน อย่างน้อย 5 ครั้งแล้ว" กรณีนี้ฐานข้อมูลเรามีการเก็บรายการซื้อขายเป็นจำนวนมาก และคำถามข้างต้นนี้จำเป็นต้องค้นหาทุกๆ คู่ของลูกค้ากับสินค้า เช่น {คุณ ก สินค้า A} { คุณ ก สินค้าB} {คุณ ก สินค้า C} {คุณ ข สินค้า B} เป็นต้น

โครงสร้างเส้นประสาท (Neural Networks) เป็นวิธีการที่จำลองแบบมาจากโครงสร้างเส้นประสาทใน

สมองของมนุษย์ และถือได้ว่าเป็นเทคโนโลยีที่มีที่มาจากงานวิจัยด้านปัญญาประดิษฐ์ (Artificial Intelligence (AI)) เพื่อใช้ในการคำนวณค่าฟังก์ชันจากกลุ่มข้อมูล วิธีการของโครงสร้างเส้นประสาท (Neural Networks) ซึ่งเป็นวิธีการที่ให้เครื่องเรียนรู้จากตัวอย่างต้นแบบแล้วฝึก (train) ให้ระบบได้รู้จักที่จะคิดแก้ปัญหาที่กว้างขึ้นได้ ในโครงสร้างของโครงสร้างเส้นประสาทจะประกอบด้วยโหนด (node) สำหรับ Input-Output และการประมวลผลกระจายอยู่ในโครงสร้างเป็นชั้นๆ ได้แก่ input layer output layer และ hidden layers การประมวลผลของโครงสร้างเส้นประสาท จะอาศัยการส่งการทำงานผ่านโหนดต่างๆ ใน layer เหล่านี้

ความแตกต่างระหว่างสถิติกับการทำเหมืองข้อมูล (different between statistics and data mining)

เทคนิคที่ใช้ในการทำเหมืองข้อมูลที่ถูกต้องและมีเหตุและผลบ่อยครั้งจะอาศัยเทคนิคทางด้านสถิติที่มีความจำเป็น เช่น การตรวจสอบข้อมูล (clean data) การกำหนดกลุ่มเป้าหมายที่ใช้ในการคาดคะเน (defined target to predict) สำหรับบางเทคนิคถูกกำหนดว่าเป็นการทำเหมืองข้อมูล เช่น CART และ CHAID ที่ถูกพัฒนาโดยนักสถิติ

สิ่งที่ทำให้สถิติกับการทำเหมืองข้อมูลแตกต่างกันมีอยู่หลายเหตุผลด้วยกัน สิ่งแรกเทคนิคการทำเหมืองข้อมูล เช่น CART neural networks และเทคนิคอื่นๆ จะเป็นการวิเคราะห์บนพื้นฐานของผู้วิเคราะห์ ซึ่งในบางเรื่องอาจจะไม่มีเหตุผลเพียงพอในการหาข้อสรุป โดยเหตุผลที่ทำมาใช้บางครั้งอาจจะถูก บางครั้งอาจจะผิด เพราะการใช้คอมพิวเตอร์ในข้อมูลทางธุรกิจหรือการกำหนดคุณลักษณะของข้อมูลจะถูกกำหนดโดยผู้ใช้

โปรแกรมสำหรับการทำเหมืองข้อมูล

ปัจจุบันมีโปรแกรมสำหรับใช้เป็นเครื่องมือที่ใช้ในการวิเคราะห์ข้อมูลมากมาย โปรแกรมที่สร้างมาสำหรับการทำเหมืองข้อมูลนอกจากจะมีหน้าที่ในการเก็บรวบรวมข้อมูลแล้ว ยังมีฟังก์ชันพื้นฐานสำหรับวิเคราะห์ข้อมูลอีกด้วย บทความนี้ผู้เขียนได้นำเสนอโปรแกรมสำเร็จรูปที่นิยมนำมาใช้ในการทำเหมืองข้อมูล พร้อมทั้งบอกรายละเอียดเกี่ยวกับเทคนิคในโปรแกรมสำเร็จรูป โดยเครื่องหมาย X ในตารางแสดงถึงเทคนิคที่มีอยู่ในโปรแกรมสำเร็จรูป (ดังตารางที่ 1 และตารางที่ 2)

ตัวอย่าง การทำเหมืองข้อมูลด้วยโปรแกรม SAS

การทำเหมืองข้อมูลด้วยโปรแกรม SAS ผู้วิเคราะห์สามารถเลือกใช้ฟังก์ชัน Enterprise Miner สำหรับการทำเหมืองข้อมูล โดยหลักการทำเหมืองข้อมูลของ Enterprise Miner จะอาศัยหลักการ Graphical User Interface มาใช้ในการวิเคราะห์สำหรับกระบวนการใน การทำเหมืองข้อมูลของโปรแกรม SAS ด้วยฟังก์ชัน Enterprise Miner จะใช้วิธีการ SAMMA ซึ่งมีกระบวนการดังต่อไปนี้

1. **Sampling** เป็นกระบวนการสุ่มข้อมูลจากฐานข้อมูลที่มีขนาดใหญ่ ซึ่งจะช่วยลดขนาดของข้อมูลเพื่อความรวดเร็วของการประมวลผล โดยอาศัยการสุ่มตัวอย่างด้วยหลักการสุ่มตัวอย่างทางสถิติ นอกจากนี้ยังมีคำสั่งการแบ่งข้อมูล ซึ่งโดยทั่วไปจะแบ่งข้อมูลออกเป็น 2 ส่วน ส่วนที่หนึ่งเป็นข้อมูลที่ใช้ฝึกฝน (training) ข้อมูลอีกส่วนหนึ่งใช้ในการทดสอบ(testing)

2. **Exploration** เป็นกระบวนการตรวจสอบความผิดปกติของข้อมูล และแก้ไขในกรณีที่เกิดข้อมูลสูญหาย โดยทั่วไปจะอาศัยการสร้างแผนภาพตรวจสอบลักษณะการกระจายของข้อมูล เพื่อหาข้อมูลที่ผิดปกติ

ตารางที่ 1 โปรแกรมสำเร็จรูปสำหรับการจำแนกข้อมูล (Classification Software)

Software : Features	Bayes Knowledge Discoverer (Unix)	Clementine (NT)	Darwin (NT)	IBM Intelligent Miner (NT)	Microsoft Bayes Network (NT)	MLC++ (Unix)	Model 1 (NT)	SAS Enterprise Miner (NT)	SGI Mineset (NT)	SNNS (Unix)	Tetrad (Unix)
Simple Bayes						X	X				
Decision Trees		X[a]	X[b]	X		X[c]	X[d]	X[d]	X[e]		
Logistic Regression							X	X			
Linear Regression		X				X[f]	X	X	X[f]		
Neural Networks		X	X	X		X	X	X		X	
Rule Builders		X				X[g]					
Association Rule Builder		X		X				X			
Decision Table						X			X		
Radial Basis Functions		X		X				X			
Instance Based						X[h]					
Linear Discriminators						X[i]					
Memory-based			X					X			
Bayesian Networks	X				X						X
Time Series				X				X			

a. C5.0 with boosting b. CART c. ID3, MC4, C4.5, T2 d. CHAID, CART e. Option Trees f. Regression Trees g. OneR h. IB i. Perceptron, Window

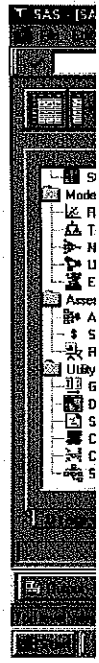
ตารางที่ 2 โปรแกรมสำเร็จรูปสำหรับการจัดกลุ่มข้อมูล (Clustering Software)

Software : Features	Autoclass III	Clementine (NT)	Darwin (NT)	IBM Intelligent Miner (NT)	Model 1 (NT)	SAS Enterprise Miner (NT)	SGI Mineset (NT)
Miscellaneous						X	
Kohonen Networks		X			X	X	
Kmeans clustering		X	X		X	X	X[a]
Bayesian	X						

a. single and iterative

3. M
จะต้องปรับข้อมู
สถิติ ดังนั้น เ
ผู้วิเคราะห์จะทำ
มาช่วยในการพิจา
จะต้องมีหลักการ
ผู้วิเคราะห์อาจจะ
คลึงกันเข้าไว้ด้วย
ตัวแบบ

4. M
กำหนดตัวแบบ
ตัวแบบที่มีความ
ถูกต้อง โดยอาศัย
ค่าเฉลี่ยกำลังสอง



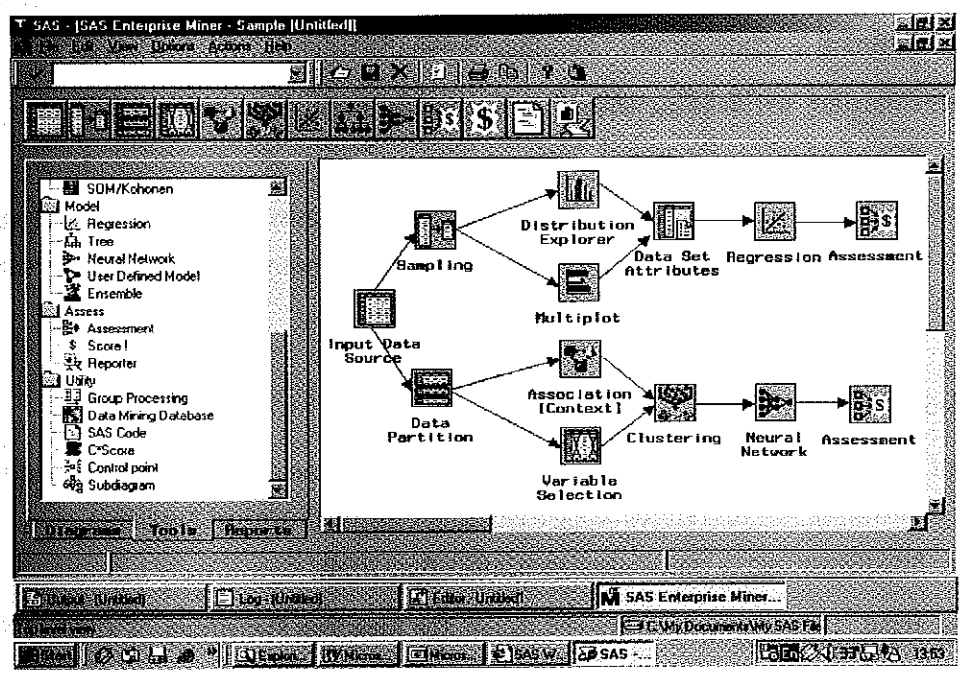
3. Modifying เป็นกระบวนการที่ผู้วิเคราะห์จะต้องปรับข้อมูลให้เป็นไปตามข้อตกลงเบื้องต้นทางสถิติ ดังนั้น เมื่อพบความผิดปกติในขั้นตอนที่ 2 ผู้วิเคราะห์จะทำการแปลงข้อมูลโดยอาศัยหลักทางสถิติมาช่วยในการพิจารณา หรือบางครั้งอาจพบข้อมูลสูญหายจะต้องมีหลักการในการหาค่ามาแทน ในกระบวนการนี้ผู้วิเคราะห์อาจจะทำการคัดเลือกข้อมูลที่มีลักษณะคล้ายคลึงกันเข้าไว้ด้วยกัน เพื่อง่ายต่อการวิเคราะห์และสร้างตัวแบบ

4. Modeling เป็นกระบวนการสร้างหรือกำหนดตัวแบบตามข้อตกลงเบื้องต้นทางสถิติ เพื่อหาตัวแบบที่มีความสามารถในการพยากรณ์ที่แม่นยำและถูกต้อง โดยอาศัยเกณฑ์การคัดเลือกตัวแบบทางสถิติ เช่น ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อน เป็นต้น

5. Assessment เป็นกระบวนการตรวจสอบความสามารถในการพยากรณ์ของตัวแบบว่ามีความแม่นยำเพียงใด ในการนำไปใช้ในการพยากรณ์ โดยอาศัยการเปรียบเทียบตัวแบบที่ผู้วิเคราะห์สร้างขึ้นมาว่าตัวแบบใดมีความสามารถในการพยากรณ์ได้ถูกต้องมากกว่ากัน

บทสรุป

ปัจจุบันหลายหน่วยงานได้เก็บรวบรวมข้อมูลไว้ในฐานข้อมูลที่มีขนาดใหญ่ และมีการขยายขนาดของข้อมูลอย่างรวดเร็ว ทำให้ระบบการประมวลผลข้อมูลและการค้นหาความรู้ใหม่ๆ จากฐานข้อมูลมีความซับซ้อนมากยิ่งขึ้น วิธีการหนึ่งที่ถูกนำมาใช้ในการประมวลผลข้อมูลและการค้นหาความรู้ใหม่ๆ คือ การทำเหมืองข้อมูล



รูปที่ 4 การทำเหมืองข้อมูลด้วยโปรแกรม SAS โดยใช้ฟังก์ชัน Enterprise Miner

(Data mining) การทำเหมืองข้อมูล (data mining) ถูกนำไปประยุกต์ใช้ได้ในหลายด้าน แต่สามารถจัดกลุ่มได้เป็นสองกลุ่ม คือ กลุ่มที่ใช้การทำเหมืองข้อมูลเพื่อการทำนาย และกลุ่มที่ใช้เพื่อการอธิบาย ในการนำเทคนิคการทำเหมืองข้อมูลไปประยุกต์ใช้มีขั้นตอนที่สำคัญ 4 ขั้นตอน คือ 1 เตรียมข้อมูล (data preparation) 2 ลดขนาดของข้อมูล (data reduction) 3 ค้นหาโมเดลจากข้อมูล (data modeling/discovery) และ 4 ตรวจสอบและวิเคราะห์ผล (solution analyses) ด้วยข้อจำกัดของระบบคอมพิวเตอร์ ฐานข้อมูลที่มีขนาดใหญ่มาก รวมทั้งความเสี่ยงและความไม่แน่นอนของการตัดสินใจ ดังนั้นการทำเหมืองข้อมูลจึงต้องอาศัยวิธีการทางสถิติมาช่วยในการประมวลผล เพื่อลดความเสี่ยงและความไม่แน่นอนของการตัดสินใจเกี่ยวกับสถานการณ์ที่เกิดขึ้นในอนาคต

บรรณานุกรม

ภาษาไทย

กัลยา วานิชย์บัญชา. **หลักสถิติ.** พิมพ์ครั้งที่ 7. กรุงเทพมหานคร : โรงพิมพ์จุฬาลงกรณ์มหาวิทยาลัย, 2545.

กัลยา วานิชย์บัญชา. **การวิเคราะห์ตัวแปรหลายตัวด้วย SPSS for Windows**. พิมพ์ครั้งที่ 2. กรุงเทพมหานคร : โรงพิมพ์จุฬาลงกรณ์มหาวิทยาลัย, 2544.

สุชาติ ประสิทธรัฐสินธุ์. **เทคนิคการวิเคราะห์ตัวแปรหลายตัว สำหรับการวิจัยทางสังคมศาสตร์และพฤติกรรมศาสตร์.** พิมพ์ครั้งที่ 3. กรุงเทพมหานคร : สถาบันบัณฑิตพัฒนบริหารศาสตร์, 2537.

สุชาติ ประสิทธรัฐสินธุ์. **การใช้สถิติในงานวิจัยอย่างถูกต้องและได้มาตรฐานสากล.** กรุงเทพมหานคร : สถาบันบัณฑิตพัฒนบริหารศาสตร์, 2545.

บัณฑิตพัฒนบริหารศาสตร์,สถาบัน. **เอกสารประกอบการบรรยายโครงการอบรมการศึกษาต่อเนื่อง หลักสูตรความรู้พื้นฐานทางการทำเหมืองข้อมูล รุ่นที่ 1 คณะสถิติประยุกต์.** กรุงเทพมหานคร : สถาบันบัณฑิตพัฒนบริหารศาสตร์, 2546.

ภาษาอังกฤษ

Center for Automated Learning and Discovery. **Lab Software**. Carnegie Mellon University Pittsburgh, 4609 Wean Hall 5000 Forbes Avenue Pittsburgh, PA 15213.[Online] Available : http://www.cald.cs.cmu.edu/software.html(9 April 2004).

David J. Hand. "Data mining: statistics and more?" **The American Statistician**. 52, (1998) : 112-118.

Glymour C., Madigan D., Pregibon D., and Smyth P. "Statistical inference and data mining," **Communications of the ACM**. 39, (1996) : 35-41.

Jiawei Han and Micheline Kamber. **Data Mining : Concepts and Techniques.** San Francisco : Morgan Kaufman Publisher, 2001.

J.H.Friedman. **Data mining and statistics : what's the connection?**. [Online] Available :http://www-stat.stanford.edu/~jhf/ftp/dm-stat.ps. (10 February 2003).

Kurt Thearling .**Thearling.com Information about data mining and analytic technologies**. [Online] Available :http://www.thearling.com/index.htm(15 January 2004).

Mehmed Kantardzic. **Data Mining :Concepts,Model,Methods,and Algorithms**. Hoboken : IEEE press A John Wiley & Sun, 2003.

Nittaya Kerdprasop. **A Proper Algorithm and Technique for Mining the Medical Diagnosis Data Sets**. [Online] Available : http://www.sut.ac.th/engineering/Computer/faculty/nittaya /myweb/ ongoing_research.htm(6 January 2003).

Paolo Giudici. **Applied Data Mining : Statistical Methods for Business and Industry**. West Sussex : John Wiley & Sun, 2003.

SAS Institute. **SAS Enterprise Miner Reference Manual**. Campus Drive Cary : SAS Institute, 2001.

Zhihua Xiao. **Statistics and Data mining**. Kent Ridge Crescent : Department of Information System and Computer Science. National University of Singapore, 2003.