ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)

ปริญญา

วิทยาการคอมพิวเตอร์	วิทยาการคอมพิวเตอร์
สาขา	**ภาควิชา**

เรื่อง การคัดเลือกลักษณะเด่นโดยใช้กฎความสัมพันธ์สำหรับการเรียนรู้ของต้นไม้ตัดสินใจ

Feature Selection Using Association Rules for Decision Tree Learning

นามผู้วิจัย นางสาวลลิตา ศรีชัยญา

ได้พิจารณาเห็นชอบโดย

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก .....

(ผู้ช่วยศาสตราจารย์นวลวรรณ สุนทรภิษัช, Ph.D.)

หัวหน้าภาควิชา .....

(ผู้ช่วยศาสตราจารย์ศิริกร จันทร์นวล, M.S.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

.....

(รองศาสตราจารย์กัญจนา ธีระกุล, D.Agr.)

คณบดีบัณฑิตวิทยาลัย

วันที่ ..... **เดือน** ..... **พ.ศ.** .....

วิทยานิพนธ์

เรื่อง

การคัดเลือกลักษณะเด่นโดยใช้กฎความสัมพันธ์สำหรับการเรียนรู้ของต้นไม้ตัดสินใจ

Feature Selection Using Association Rules for Decision Tree Learning

โดย

นางสาวลลิตา ศรีชัยญา

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

เพื่อความสมบูรณ์แห่งปริญญาวิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)

พ.ศ. 2553

ลลิตา ศรีชัยญา 2553: การคัดเลือกลักษณะเด่นโดยใช้กฎความสัมพันธ์สำหรับการเรียนรู้ของต้นไม้ตัดสินใจ ปริญญาวิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์) สาขาวิทยาการคอมพิวเตอร์ ภาควิชาวิทยาการคอมพิวเตอร์ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก: ผู้ช่วยศาสตราจารย์นวลวรรณ สุนทรภิษัช, Ph.D. 93 หน้า

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาขั้นตอนวิธีการคัดเลือกลักษณะเด่นของข้อมูลที่สามารถเพิ่มประสิทธิภาพการเรียนรู้ของต้นไม้ตัดสินใจ เพื่อให้ได้ต้นไม้ตัดสินใจที่มีประสิทธิภาพในการจำแนกข้อมูล โดยการนำเอาเทคนิคการหากฎความสัมพันธ์มาใช้ร่วมกับเทคนิคการจำแนกข้อมูล โดยอาศัยการเรียนรู้ของต้นไม้ตัดสินใจ ซึ่งขั้นตอนการดำเนินงานสามารถแบ่งออกเป็น 3 ขั้นตอนคือ ขั้นตอนการเตรียมข้อมูล ขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล และขั้นตอนการคัดกรองแอททริบิวท์ งานวิจัยนี้ได้ทำการทดลองกับชุดข้อมูลทั้งสิ้น 4 ชุด โดยนำมาจากฐานข้อมูล UCI Machine Learning Repository โดยผู้วิจัยได้ทำการวัดประสิทธิภาพเปรียบเทียบกับขั้นตอนวิธี J48 และขั้นตอนวิธี CBA

ผลการทดลองพบว่าประสิทธิภาพในการจำแนกข้อมูลโดยอาศัยการคัดเลือกลักษณะเด่นของข้อมูลด้วยกฎความสัมพันธ์ (AssoTree) ให้ประสิทธิภาพการจำแนกข้อมูลที่สูงกว่าขั้นตอนวิธี J48 และขั้นตอนวิธี CBA และส่งผลต่อการเรียนรู้ของต้นไม้ตัดสินใจ โดยทำให้ได้ต้นไม้ที่มีขนาดเล็กลง แต่ยังคงมีประสิทธิภาพที่เหนือกว่าการ ไม่ใช้วิธีการคัดเลือกลักษณะเด่นของข้อมูล

_____ _____ ___ / ___ / ___

ลายมือชื่อนิสิต ลายมือชื่ออาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

Lalita Srichaiya 2010: Feature Selection Using Association Rules for Decision Tree Learning. Master of Science (Computer Science), Major Field: Computer Science, Department of Computer Science. Thesis Advisor: Assistant Professor Nuanwan Soonthornphisaj, Ph.D. 93 pages.

The objectives of this research are to study a feature selection algorithm that can filter out a set of attributes in order to enhance the performance of decision tree learning. We proposed an algorithm called AssoTree that combines an association rule mining with the decision tree learning. The algorithm has 3 parts: data preprocessing, feature selection and attribute filtering, respectively.

The experiments are done on 4 benchmark data sets collected from UCI Machine Learning Repository. The performances of AssoTree are compared to two algorithms which are decision tree (J48) and Classification based on Association (CBA). The experimental results show that AssoTree outperforms other two algorithms since AssoTree can effectively filter out a set of attributes before supplying the dataset to the decision tree algorithm. AssoTree has the higher accuracy and gives the smaller tree size.

_____ _____ ___ / ___ / ___

Student's signature Thesis Advisor's signature

กิตติกรรมประกาศ

กราบขอบพระคุณ ผู้ช่วยศาสตราจารย์นวลวรรณ สุนทรภิษัช ประธานกรรมการที่ปรึกษา ที่ให้ความช่วยเหลือเป็นอย่างดี ให้คำปรึกษา คำแนะนำวิธีการและขั้นตอนในการดำเนินงาน การค้นคว้าวิจัย ชี้แนวทางการแก้ไขปัญหา ตลอดจนการตรวจแก้ไขวิทยานิพนธ์จนกระทั่งเสร็จสมบูรณ์ และกราบขอบพระคุณ ผศ.ดร.วรเศรษฐ สุวรรณิก ประธานการสอบ และอ.ดวงดาว วิชาดากุล ผู้ทรงคุณวุฒิ ที่ให้คำแนะนำ ความกรุณาตรวจแก้ไขวิทยานิพนธ์ให้สมบูรณ์ยิ่งขึ้น

กราบขอบพระคุณ คุณพ่อ คุณแม่ ผู้ซึ่งสร้างครอบครัวที่อบอุ่น ผู้เป็นความรัก กำลังใจ และแรงพลังที่ยิ่งใหญ่ให้แก่ข้าพเจ้าทั้งกาย และใจ และขอบคุณพี่ชาย ที่คอยถามไถ่ ห่วงใย และเอื้อเฟื้อข้าพเจ้าเสมอมา

สุดท้ายขอขอบคุณเพื่อน พี่ และน้อง ทุกคนที่ช่วยเหลือในการค้นคว้าหาความรู้เพิ่มเติมเพื่อนำมาใช้เป็นประโยชน์ในการทำวิจัย รวมถึงยังเป็นกำลังใจให้แก่ข้าพเจ้า และบุคคลอีกผู้หนึ่งซึ่งคอยสนับสนุน เป็นแรงพลัง และกำลังใจให้กับข้าพเจ้าตลอดมา ขอขอบคุณภาควิชาวิทยาการคอมพิวเตอร์ โครงการปริญญาโทที่ได้สนับสนุนด้านอุปกรณ์ และสถานที่ตลอดระเวลาในการทำวิทยานิพนธ์ และขอขอบคุณทุกสิ่งทุกอย่างที่ทำให้วิทยานิพนธ์นี้สำเร็จลุล่วงไปได้ด้วยดี

ลลิตา ศรีชัยญา

เมษายน 2553

สารบัญ

สารบัญตาราง

สารบัญตาราง (ต่อ)

สารบัญภาพ

การคัดเลือกลักษณะเด่นโดยใช้กฎความสัมพันธ์สำหรับการเรียนรู้ของต้นไม้ตัดสินใจ

Feature Selection Using Association Rules for Decision Tree Learning

คำนำ

การทำเหมืองข้อมูลเป็นกระบวนการที่สำคัญในการค้นหารูปแบบ แนวทาง และความสัมพันธ์ของข้อมูลที่ซ่อนอยู่ เทคนิคที่นำมาใช้ในการทำเหมืองข้อมูลส่วนใหญ่แบ่งเป็น 4 เทคนิค ได้แก่ เทคนิคการจำแนกข้อมูล (Classification) เทคนิคการหากฎความสัมพันธ์ (Association Rules) เทคนิคการแบ่งกลุ่มข้อมูล (Clustering) และเทคนิคจินตทัศน์ (Visualization)

เทคนิคการจำแนกข้อมูล เป็นการวิเคราะห์เพื่อจำแนกข้อมูลเป็นประเภทกลุ่มข้อมูล มีการกำหนดข้อมูลตัวอย่างของข้อมูลแต่ละประเภท เพื่อเป็นตัวแทนสำหรับการจำแนกข้อมูลทั้งหมด ลักษณะเช่นนี้เรียกว่า “Supervised Learning” เทคนิคที่ใช้ในการจำแนกข้อมูลมีมากมาย แต่มีเทคนิคหนึ่งเป็นที่นิยม และคุ้นเคยมาก คือ การเรียนรู้ของต้นไม้ตัดสินใจ (Decision Tree) โดยขั้นตอนการทำงานของการเรียนรู้ของต้นไม้ตัดสินใจ เกี่ยวข้องกับการเลือกแอททริบิวท์ที่เหมาะสมมาสร้างเป็นโหนดในต้นไม้ โดยแอททริบิวท์ที่จะนำมาสร้างเป็นโหนด ต้องสามารถแยกแยะข้อมูลเป็นประเภทที่มีการปะปนกันของข้อมูลน้อยที่สุด

เทคนิคการหากฎความสัมพันธ์ เป็นการค้นหากฎความสัมพันธ์ของข้อมูล ที่แสดงความสัมพันธ์ของเหตุการณ์หรือวัตถุที่เกิดขึ้นพร้อมกัน โดยจะทำการประเมินจากข้อมูลในตารางที่รวบรวมไว้ แล้วผลการวิเคราะห์ที่ได้จะเป็นคำตอบของปัญหาที่ได้เป็นความสัมพันธ์กันของข้อมูล ซึ่งความสัมพันธ์กันของข้อมูลจะใช้ค่าสนับสนุน (Support) และค่าความเชื่อมั่น (Confidence) บอกคุณภาพของกฎว่าดีหรือไม่เพียงใด โดยค่าสนับสนุน แสดงถึงความมีนัยสำคัญของความสัมพันธ์ของข้อมูล และค่าความเชื่อมั่น แสดงถึงความน่าเชื่อมั่นของกฎที่เกิดจากความสัมพันธ์ของข้อมูล

เทคนิคการจำแนกข้อมูลโดยใช้กฎความสัมพันธ์ (Associative Classification) เป็นเทคนิคหนึ่งในการจำแนกข้อมูลที่น่าสนใจ และให้ประสิทธิภาพในการทำนายสูง เนื่องจากมีการนำเทคนิคการหากฎความสัมพันธ์รวมกับเทคนิคการจำแนกข้อมูล เป้าหมายของเทคนิคการจำแนกข้อมูลโดยใช้กฎความสัมพันธ์ คือ เพื่อทำนายกลุ่มให้กับข้อมูลในอนาคตที่ไม่ทราบกลุ่ม โดยการทำงานของ

เทคนิคการจำแนกข้อมูลโดยใช้กฎความสัมพันธ์แบ่งเป็น 2 ส่วนคือ การสร้างกฎความสัมพันธ์ (Rule generator) และการสร้างโมเดลการทำนาย (Classifier builder) โดยขั้นตอนวิธีที่เป็นต้นแบบของเทคนิคดังกล่าว คือ CBA (Classification based on Association rules) (Liu. *et al.*, 1998) และภายหลังได้มีผู้นำมาพัฒนาต่อเป็นขั้นตอนวิธี CMAR (Classification based on Multiple class Association Rules) (Li. *et al.*, 2001) ซึ่งเป็นขั้นตอนวิธีแรกที่ใช้กฎความสัมพันธ์หลายกฎมาพิจารณา

งานวิจัยนี้มีการปรับปรุงเทคนิคการจำแนกข้อมูลโดยใช้กฎความสัมพันธ์ เพื่อให้มีประสิทธิภาพในการทำนายที่แม่นยำขึ้น โดยนำเสนอขั้นตอนวิธี AssoTree ประกอบด้วยขั้นตอน 3 ขั้นตอนในการดำเนินงาน คือ ขั้นตอนการเตรียมข้อมูล ขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล และวิธีการคัดกรองแอททริบิวท์ วัตถุประสงค์ของงานวิจัยนี้เพื่อศึกษาขั้นตอนวิธีการคัดเลือกลักษณะเด่นของข้อมูล และความสำคัญของแอททริบิวท์ที่มีผลต่อการเรียนรู้ของต้นไม้ตัดสินใจ เพื่อให้ได้แอททริบิวท์ที่สำคัญที่สุดในการสร้างโหนดต้นไม้ตัดสินใจที่ใช้สำหรับการจำแนกข้อมูล

เนื่องจากการจำแนกข้อมูลที่อาศัยเทคนิคการเรียนรู้ของต้นไม้ตัดสินใจ จะเกี่ยวข้องกับการเลือกแอททริบิวท์ที่เหมาะสมมาสร้างเป็นโหนดในต้นไม้ เพื่อใช้สำหรับแบ่งแยกข้อมูล ดังนั้นเพื่อให้ได้แอททริบิวท์ที่สำคัญที่สุดที่จะนำไปสร้างเป็นโหนดในต้นไม้ตัดสินใจที่ใช้สำหรับการจำแนกข้อมูล จึงนำเอาเทคนิคการหากฎความสัมพันธ์เข้ามาใช้ เพื่อหาแอททริบิวท์ที่มีความสัมพันธ์กันที่ช่วยในจำแนกข้อมูลที่ดีที่สุด

การทดลองได้ใช้ข้อมูลทั้งสิ้น 4 ชุดข้อมูล ได้มาจากฐานข้อมูล UCI Machine Learning Repository (Aha *et al.*, 1987) ได้ทำการเปรียบเทียบขั้นตอนวิธี AssoTree กับเทคนิคการจำแนกข้อมูลด้วยขั้นตอนวิธีที่เป็นมาตรฐาน และนิยมกัน คือ ขั้นตอนวิธี J48 และเทคนิคที่เป็นต้นแบบของเทคนิคการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ คือ ขั้นตอนวิธี CBA

วัตถุประสงค์

1. เพื่อศึกษาขั้นตอนวิธีการคัดเลือกลักษณะเด่นของข้อมูลฝึก ที่สามารถเพิ่มประสิทธิภาพการเรียนรู้ของต้นไม้ตัดสินใจได้

2. เพื่อศึกษาความสำคัญของแอททริบิวท์ที่มีผลต่อการเรียนรู้ของต้นไม้ตัดสินใจ

ประโยชน์ที่คาดว่าจะได้รับ

1. เพิ่มประสิทธิภาพของขั้นตอนวิธีการเรียนรู้ของต้นไม้ตัดสินใจ

2. ได้องค์ความรู้ใหม่จากคุณลักษณะเด่นที่สำคัญของชุดข้อมูลที่มีประโยชน์ต่อการทำนายข้อมูล

3. สามารถนำเอาขั้นตอนวิธีจากงานวิจัยไปประยุกต์ใช้สำหรับการทำเหมืองข้อมูลต่อไป

ขอบเขตและข้อจำกัด

1. ทดลองกับข้อมูลที่มีจำนวนคลาส 2 คลาส

2. ข้อมูลที่ทดลองต้องมีจำนวนแอททริบิวท์มาก

การตรวจเอกสาร

ทฤษฎีที่เกี่ยวข้อง

ส่วนของทฤษฎีที่เกี่ยวข้อง ประกอบด้วยการทำเหมืองข้อมูล การหากฎความสัมพันธ์ และการจำแนกข้อมูลด้วยต้นไม้ตัดสินใจ

1. การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล เป็นเทคนิคที่ใช้ในกระบวนการค้นหาองค์ความรู้ หรืออาจเรียกว่า การค้นหาความรู้ในฐานข้อมูล (Knowledge Discovery in Database: KDD) เป็นกระบวนการค้นหาความสัมพันธ์และรูปแบบทั้งหมด ซึ่งมีอยู่จริงในฐานข้อมูล แต่ถูกซ่อนไว้ภายในข้อมูลจำนวนมาก โดยการทำเหมืองข้อมูลจะทำการสำรวจและวิเคราะห์อัตโนมัติ ให้อยู่ในรูปแบบที่มีความหมายและในรูปของกฎ โดยความสัมพันธ์เหล่านี้แสดงให้เห็นถึงความรู้ต่างๆ ที่สามารถนำไปใช้ให้เกิดประโยชน์ได้

1.1 วัตถุประสงค์ของการทำเหมืองข้อมูล

1.1.1 เพื่อการค้นพบองค์ความรู้ใหม่ในฐานข้อมูล (Knowledge Discovery in Database) หมายถึงกระบวนการค้นพบความรู้แฝงของข้อมูลที่อยู่ในกลุ่มข้อมูลจำนวนมากที่เก็บอยู่ในฐานข้อมูล หรือแหล่งเก็บข้อมูล ซึ่งลักษณะที่น่าสนใจของข้อมูลเหล่านี้ เช่น รูปแบบความสัมพันธ์ การเปลี่ยนแปลง โครงสร้างที่เด่นชัด หรือลักษณะที่ผิดปกติของข้อมูล

1.1.2 เพื่อการสกัดองค์ความรู้ที่ซ่อนเร้นอยู่ (Knowledge Extraction) หมายถึงการแยกข้อมูลจากฐานข้อมูลขนาดใหญ่ เพื่อให้ได้ข้อมูลที่เกิดประโยชน์

1.1.3 เพื่อจัดการกับข้อมูลในอดีต (Data Archeology) เนื่องจากการทำเหมืองข้อมูลเป็นการสังเคราะห์ข้อมูลอย่างละเอียดจากฐานข้อมูลขนาดใหญ่ โดยเรียนรู้ข้อมูลจากอดีต

1.1.4 เพื่อสำรวจข้อมูล (Data Exploration)

1.1.5 เพื่อค้นหารูปแบบของข้อมูลที่ซ่อนอยู่ (Data Pattern Processing)

1.1.6 เพื่อใช้ขุดเจาะข้อมูล (Data Dredging)

1.1.7 เพื่อเก็บเกี่ยวผลประโยชน์ให้ได้มาซึ่งสารสนเทศที่มีประโยชน์

1.2 ขั้นตอนการทำเหมืองข้อมูล ประกอบด้วย 4 ขั้นตอนหลัก ดังนี้

1.2.1 ขั้นตอนการกำหนดวัตถุประสงค์ (Object Determination) เป็นขั้นตอนการกำหนดขอบเขต เป้าหมายของการทำเหมืองข้อมูล ซึ่งจะมีผลต่อทุกขั้นตอนในการทำเหมืองข้อมูล

1.2.2 ขั้นตอนการเตรียมข้อมูล (Data Preprocessing) เป็นขั้นตอนจัดการข้อมูลให้สามารถนำเข้าสู่ขั้นตอนวิธีของการทำเหมืองข้อมูลได้ ขั้นตอนการเตรียมข้อมูลสามารถแบ่งได้เป็น 3 ส่วนได้แก่

ก. การคัดเลือกข้อมูล (Data Selection) เป็นการเลือกข้อมูลที่จะนำมาใช้ในการทำเหมืองข้อมูลโดยการระบุแหล่งของข้อมูลและทำการดึงข้อมูลออกมาสำหรับการวิเคราะห์เบื้องต้น ซึ่งจะแตกต่างไปตามวัตถุประสงค์ที่ได้กำหนด และลักษณะงานที่จะนำไปใช้ด้วย

รูปแบบของข้อมูล แบ่งเป็น 2 ประเภท คือ

1. ข้อมูลเชิงคุณภาพ (Categorical) แบ่งเป็น 2 ชนิด ได้แก่

1.1 ข้อมูลชนิดโนมินัล (Nominal Attribute) เป็นข้อมูลที่ค่าเป็นไปได้ไม่มีลำดับ เช่น สถานะการแต่งงาน (โสด แต่งงาน หย่า) เพศ (ชาย หญิง) ระดับการศึกษา (ประถมศึกษา มัธยมศึกษา ปริญญาตรี ปริญญาโท) เป็นต้น

1.2 ข้อมูลชนิดลำดับ (Ordinal Attribute) เป็นข้อมูลที่ค่าเป็นไปได้มีลำดับ เช่น ลำดับของลูกค้า (ดี ปานกลาง ไม่ดี) ระดับอุณหภูมิ (ต่ำ กลาง สูง) เป็นต้น

2. ข้อมูลเชิงปริมาณ เป็นข้อมูลที่อยู่ในรูปตัวเลขที่แสดงถึงปริมาณของข้อมูล แบ่งเป็น 2 ชนิด ได้แก่

2.1 ข้อมูลชนิดตัวเลขต่อเนื่อง (Continuous) เช่น รายได้ ค่าเฉลี่ย เป็นต้น

2.2 ข้อมูลชนิดตัวเลขไม่ต่อเนื่อง (Discrete) เช่น จำนวนรถยนต์ จำนวนพนักงาน เวลาปี เป็นต้น

ข. การทำความสะอาดข้อมูล (Data Cleaning) เป็นขั้นตอนที่แยกข้อมูลที่ไม่เกี่ยวข้อง ซ้ำซ้อน ข้อมูลที่ไม่สอดคล้องกันออกไป ทำให้ได้ข้อมูลที่มีคุณภาพที่นำไปวิเคราะห์ได้

ค. การแปลงรูปข้อมูล (Data Transformation) เป็นขั้นตอนการลดรูปและจัดข้อมูลให้อยู่ในรูปแบบเดียวกันที่เหมาะสมสำหรับการนำไปใช้งานตามคุณสมบัติเฉพาะของแต่ละขั้นตอนวิธีที่ใช้ในการทำเหมืองข้อมูล

1.2.3 ขั้นตอนการทำเหมืองข้อมูล เป็นขั้นตอนที่อาศัยเทคนิคการทำเหมืองข้อมูล ในการดำเนินการจะมีการเลือกฟังก์ชันการทำเหมืองข้อมูล เช่น Summarization, Classification, Regression, Association และ Clustering เป็นต้น แต่ละฟังก์ชันจะมีขั้นตอนวิธีให้เลือกใช้แตกต่างกัน การทำเหมืองข้อมูล สามารถแบ่งประเภทของงานตามลักษณะแบบจำลอง ออกเป็น 2 ประเภทใหญ่ คือ

ก. แบบจำลองเชิงทำนาย (Predictive Model หรือ Supervised Model) เป็นการคาดคะเนลักษณะ หรือประมาณค่าที่ชัดเจนของข้อมูลที่เกิดขึ้น โดยใช้พื้นฐานจากข้อมูลที่ผ่านมาในอดีต ลักษณะของแบบจำลองเชิงทำนายทุกข้อมูลจะมีคำเฉลย เพื่อใช้ในการเรียนรู้ และทำนายผลของข้อมูล โดยขั้นตอนวิธีแบบนี้จะเน้นการจัดกลุ่มข้อมูลตามคุณสมบัติของคำเฉลย ซึ่งถ้าคลาสข้อมูลเป็นข้อมูลชนิดโนมินัล จะเรียกกระบวนการที่ใช้แบ่งแยกว่า การจำแนก (Classification) แต่ถ้าคลาสข้อมูลเป็นข้อมูลชนิดตัวเลขต่อเนื่อง จะเรียกกระบวนการที่ใช้แบ่งแยกว่า การถดถอย (Regression)

ข. แบบจำลองเชิงพรรณนา (Descriptive Model หรือ Unsupervised Model) เป็นการหาแบบจำลองเพื่ออธิบายลักษณะบางอย่างของข้อมูลที่มีอยู่ โดยการหาความสัมพันธ์ (Association) หรือการจัดกลุ่มข้อมูล (Clustering)

1.2.4 ขั้นตอนการสรุปผลลัพธ์ (Result Analysis and Evaluation) เป็นขั้นตอนการแปลความหมายและการประเมินผลลัพธ์ที่ได้นั้นเหมาะสมหรือตรงกับวัตถุประสงค์หรือไม่และนำเสนอในรูปแบบที่สามารถเข้าใจได้ง่าย ข้อมูลความรู้ที่ได้นำไปเป็นข้อมูลสารสนเทศที่ช่วยในการตัดสินใจต่อไป

1.3 เทคนิคที่นำมาใช้ในการทำเหมืองข้อมูล แบ่งเป็น 4 ประเภท ได้แก่

1.3.1 การหาความสัมพันธ์ของข้อมูล

การหาความสัมพันธ์ของข้อมูล เป็นรูปแบบที่แสดงความสัมพันธ์ของเหตุการณ์หรือวัตถุที่เกิดขึ้นพร้อมกัน เช่น ความสัมพันธ์ของการซื้อสินค้า พบว่า ถ้าลูกค้าซื้อไส้กรอกแล้ว จะต้องซื้อน้ำอัดลมตามมา

1.3.2 การจำแนกข้อมูล

การจำแนกข้อมูล เป็นรูปแบบที่แสดงในรูปของกฎ เพื่อระบุประเภทของวัตถุจากคุณสมบัติของวัตถุ เช่น ความสัมพันธ์ระหว่างผลการตรวจร่างกาย กับการเกิดโรค หรือในทางธุรกิจความสัมพันธ์ระหว่างผู้ก่อหนี้ดีหรือหนี้เสีย กับการอนุมัติเงินกู้

1.3.3 การแบ่งกลุ่มข้อมูล

การแบ่งกลุ่มข้อมูล เป็นรูปแบบที่แสดงการแบ่งกลุ่มข้อมูลใหม่ที่มีลักษณะคล้ายกันไว้ในกลุ่มเดียวกัน เช่น การวิเคราะห์หาสาเหตุของโรค โดยจัดกลุ่มผู้ป่วยตามลักษณะอาการที่คล้ายคลึงกัน

1.3.4 การแสดงผลข้อมูลในรูปจินตทัศน์

การแสดงผลข้อมูลในรูปจินตทัศน์ เป็นรูปแบบที่แสดงผลการนำเสนอข้อมูลในรูปแบบกราฟิก ที่ช่วยให้ผู้ใช้หรือผู้ตัดสินใจสามารถค้นพบรูปแบบได้จากการแสดงผลจากรูปภาพแทนการใช้ข้อความในการนำเสนอข้อมูล

2. ความหมายแอททริบิวท์

แอททริบิวท์ (Attribute) หมายถึง คุณลักษณะเฉพาะของแต่ละชุดข้อมูล และค่าในแต่ละแอททริบิวท์เป็นค่าที่เป็นไปได้ในแอททริบิวท์ใดแอททริบิวท์หนึ่งเท่านั้น (วิเชียร, 2546)

ตัวอย่างเช่น ชุดข้อมูล Car (Aha *et al*., 1987) ประกอบด้วยแอททริบิวท์ทั้งหมด 7 แอททริ-บิวท์ ได้แก่ buying, maint, doors, persons, LugBoot, safety และ class รวมถึงค่าที่เป็นไปได้ของแต่ละแอททริบิวท์ ตามตารางที่ 1 ดังนี้

ตารางที่ 1 ตัวอย่างแอททริบิวท์ของชุดข้อมูล Car

buying	maint	doors	persons	LugBoot	safety	class
vhigh	vhigh	2	2	small	low	unacc
vhigh	vhigh	2	2	small	med	unacc
vhigh	vhigh	3	4	big	high	unacc
vhigh	vhigh	5more	2	med	low	unacc
vhigh	low	4	more	small	high	acc

3. เทคนิคการลดรูปข้อมูลชนิดตัวเลขต่อเนื่อง

เทคนิคการการลดรูปข้อมูลชนิดตัวเลขต่อเนื่อง หรือเรียกว่า การทำดิสครีตไทเซชัน (Discretization) (Wasilewska, 2552) เป็นเทคนิคหนึ่งของการลดรูปข้อมูลที่ใช้ในขั้นตอนของการเตรียมข้อมูล โดยการลดรูปข้อมูลชนิดตัวเลขต่อเนื่อง เป็นข้อมูลแบบไม่ต่อเนื่อง หรือข้อมูลที่เรียกว่าข้อมูลชนิดโนมินัล ซึ่งการทำงานจะทำการแบ่งค่าข้อมูลดิบออกเป็นช่วง แล้วทำการแทนค่าแต่ละช่วง ด้วยการคำนวณหาค่าที่อาศัยวิธีการที่เหมาะสม โดยสามารถแบ่งประเภทวิธีการทำ Discretization เป็น 2 แบบ (Dougherty *et al.*, 1995) คือ

3.1 แบบซุปเปอร์ไวส์ (Supervised)

การทำ Discretization แบบ Supervised เป็นการลดรูปข้อมูลชนิดตัวเลขต่อเนื่อง เป็นข้อมูลแบบไม่ต่อเนื่อง โดยมีการนำข้อมูลคลาสมาพิจารณาในการคำนวณหาค่าที่เป็นตัวแทนของการแบ่งข้อมูลเป็นช่วง ซึ่งอาศัยเทคนิคพื้นฐานเอ็นโทรพี (Entropy-based) มาใช้ในการลดรูปข้อมูล วิธีการคือ ค่าแต่ละค่าในแอททริบิวท์นั้นจะถูกนำมาพิจารณา เพื่อเป็นตัวถูกเลือกที่นำไปใช้สำหรับการแบ่งข้อมูล โดยจะมีการแบ่งข้อมูลเป็น 2 ช่วง (Binary split) การแบ่งกลุ่มจะอยู่ในรูปแบบ ดังนี้

$A \leq T$ และ $A > T$

โดยที่ A คือ แอททริบิวท์ที่ข้อมูลเป็นชนิดตัวเลขต่อเนื่อง

T คือ ค่าตัวแทนที่เป็นตัวแบ่งข้อมูลออกเป็นช่วง

ความหมาย คือ $A \leq T$ หมายถึง ค่าของข้อมูลที่มีค่าน้อยกว่า หรือเท่ากับค่า T ซึ่งจะอยู่ทางด้านซ้าย และ $A > T$ หมายถึง ค่าของข้อมูลที่มีค่ามากกว่าค่า T ซึ่งจะอยู่ทางด้านขวา วิธีการหาค่าตัวแทนสำหรับแบ่งข้อมูล เริ่มจากเรียงลำดับข้อมูลจากน้อยไปมาก จากนั้นคำนวณหาค่ากลางระหว่างคู่ของข้อมูลที่เรียง ซึ่งจะได้เป็นค่าที่จะถูกเลือกไปเป็นจุดตัด (Cut point) สำหรับการแบ่งข้อมูล หลังจากนี้จะอาศัยข้อมูลคลาสมาพิจารณาการเลือกค่าตัวแทน โดยอาศัยจากการคำนวณค่าเอ็นโทรพีของคลาสข้อมูล ตามสมการ ดังนี้

$$Entropy(S) = - \sum_{i=1}^{k} P(C_i, S) \log(P(C_i, S)) \quad (1)$$

โดยที่ $Ent(S)$ คือ ค่าความไร้ระเบียบของข้อมูลคลาสของกลุ่มข้อมูลย่อย S

$P(C_i, S)$ คือ สัดส่วนของข้อมูล S ที่มี C_i เป็นคลาส

k คือ จำนวนคลาสข้อมูล

$$E(A,T,S) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2) \quad (2)$$

โดยที่ $E(A,T,S)$ คือ ค่าความไร้ระเบียบของชุดข้อมูล S ที่แอททริบิวท์ A ด้วยจุดตัด T

A คือ แอททริบิวท์

T คือ จุดตัดสำหรับการแบ่งข้อมูล

S คือ ชุดข้อมูล

S_1 และ S_2 คือ กลุ่มข้อมูลย่อยของชุดข้อมูล S

การเลือกจุดตัดสำหรับการแบ่งข้อมูลนั้น จะเลือกจุดตัดที่ทำให้ได้ค่าความไร้ระเบียบน้อยที่สุด

3.2 แบบอันซุปเปอร์ไวส์ (Unsupervised)

การทำ Discretization แบบ Unsupervised เป็นการลดรูปข้อมูลชนิดตัวเลขต่อเนื่องเป็นข้อมูลแบบไม่ต่อเนื่อง โดยจะไม่นำข้อมูลคลาสมาพิจารณาในการคำนวณหาค่าที่เป็นตัวแทนของการแบ่งข้อมูลเป็นช่วง ซึ่งอาศัยเทคนิคบินนิงอย่างง่าย (Simple Binning) มาใช้ในการลดรูปข้อมูล วิธีการที่ใช้ คือ การแบ่งข้อมูลด้วยความกว้างที่เท่ากัน (Equal-width partitioning) โดยจะกำหนดจำนวนช่วงที่ต้องการ และความกว้างของแต่ละช่วงมีความกว้างเท่ากัน ซึ่งความกว้างของแต่ละช่วงสามารถคำนวณได้จากสมการ ดังนี้

$$W = (B - A) / N \tag{3}$$

โดยที่ A และ B คือ ค่าที่ต่ำที่สุด และค่าสูงที่สุด ตามลำดับ ของค่าในข้อมูลแอททริบิวท์
N คือ จำนวนช่วงที่ต้องการ
W คือ ความกว้างของแต่ละช่วง

ขั้นตอนการทำงาน เริ่มจากเรียงลำดับข้อมูลจากน้อยไปมาก แล้วแบ่งข้อมูลเป็นกลุ่มย่อยให้มีความกว้างเท่ากันทุกกลุ่มตามสมการ (3) จากนั้นทำการปรับข้อมูลภายในกลุ่มย่อยให้เป็นข้อมูลเดียวกัน โดยการเลือกปรับข้อมูลที่อาศัยค่าทางสถิติมาใช้ (วัฒนา, 2553) แบ่งได้เป็น 3 ค่า ได้แก่

3.2.1 ค่าฐานนิยม (Mode) คือ ค่าที่มาจากข้อมูลที่มีการซ้ำกันมากที่สุด

3.2.2 ค่าเฉลี่ย (Mean) คือ ค่าที่คำนวณจากผลบวกของข้อมูลทั้งหมด และหารด้วยจำนวนข้อมูลทั้งหมด ตามสมการ ดังนี้

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{3.1}$$

โดยที่ $\overline{X}$ คือ ค่าเฉลี่ย

$\sum_{i=1}^{n} X_i$ คือ ผลรวมข้อมูล n จำนวนจาก X_1 ถึง X_n

n คือ จำนวนข้อมูลทั้งหมด

3.2.3 ค่ามัธยฐานหรือค่ากลาง (Median) คือ ค่าที่ได้จากข้อมูลที่เรียงลำดับ แล้วเลือกค่าอยู่ตรงกลางระหว่างข้อมูลนั้น

4. การหากฎความสัมพันธ์ (Association Rules)

การหากฎความสัมพันธ์ เป็นเทคนิคที่อาศัยการเชื่อมโยงข้อมูล หรือกลุ่มข้อมูลที่เกิดขึ้นในเหตุการณ์เดียวกันเข้าด้วยกัน เช่น ใช้ในการวิเคราะห์ข้อมูลการสั่งซื้อสินค้าของลูกค้าที่เรียกว่า “Market – Basket Analysis” ซึ่งจะประเมินจากข้อมูลในตารางที่รวบรวมไว้ ผลการวิเคราะห์ที่ได้จะเป็นคำตอบของปัญหา ได้เป็นความสัมพันธ์ของข้อมูล

ในรูปแบบกฎความสัมพันธ์ที่เกิดขึ้น สามารถเขียนความสัมพันธ์ระหว่างข้อมูลออกมาได้ในรูปของ A $\rightarrow$ B โดยที่ A เป็นเหตุการณ์ที่เกิดขึ้นก่อน (Antecedent) หรือ LHS (Left – Hand Side) และ B เป็นผลของเหตุการณ์ (Consequent) หรือ RHS (Right – Hand Side) เช่น ในกฎของความสัมพันธ์ “ถ้ากบร้อง แล้วฝนจะตก” เหตุการณ์ที่เกิดขึ้นก่อนคือ กบร้อง เหตุการณ์ที่เกิดหลังคือ ฝนตก ซึ่งสามารถแสดงความสัมพันธ์ได้จากรูปแบบกฎความสัมพันธ์ ดังนี้

$$\text{LHS} \rightarrow \text{RHS}$$

โดยที่ LHS และ RHS เป็นชุดของข้อมูลที่เกิดขึ้นบ่อยในฐานข้อมูล

ตัวอย่างการนำเทคนิคการหากฎความสัมพันธ์ไปประยุกต์ใช้กับงานจริง ได้แก่ ระบบแนะนำหนังสือให้กับลูกค้าแบบอัตโนมัติบนเวบไซต์ซื้อหนังสือออนไลน์ ข้อมูลการสั่งซื้อทั้งหมดจะถูกนำมาประมวลผลเพื่อหาความสัมพันธ์ของข้อมูล คือ ลูกค้าที่ซื้อหนังสือเล่มหนึ่งๆ มักจะซื้อหนังสือเล่มใดพร้อมกันด้วยเสมอ ความสัมพันธ์ที่ได้จะสามารถนำไปใช้คาดเดาได้ว่า ควรแนะนำหนังสือเล่มใดเพิ่มเติมให้กับลูกค้า เพื่อเป็นการเพิ่มยอดขายให้บริษัทได้ เช่น buys (x, database) ->

buys (x, data mining) [80% , 60%] หมายความว่า เมื่อซื้อหนังสือ database แล้วจะมีโอกาสที่จะซื้อหนังสือ data mining ด้วย 60% และมีการซื้อทั้งหนังสือ database และหนังสือ data mining พร้อมๆ กัน 80%

วิธีการหากฎความสัมพันธ์

การค้นหากฎความสัมพันธ์ของข้อมูล จะอาศัยความสำคัญของกฎดูจากการบอกคุณภาพของกฎว่าดีหรือไม่ดี ขึ้นอยู่กับ 2 ปัจจัยที่สำคัญ ได้แก่

4.1 ค่าสนับสนุน (Support)

ค่าสนับสนุน เป็นค่าเปอร์เซ็นต์ของความถี่ที่วัดสัดส่วนของข้อมูลที่เกิดขึ้นในชุดข้อมูล เช่น ค่าสนับสนุนของ A หมายถึง จำนวนเรคคอร์ดที่ประกอบด้วยข้อมูล A เมื่อเทียบกับจำนวนเรคคอร์ดทั้งหมดในฐานข้อมูล สามารถแสดงสมการ ดังนี้

$$Support = \frac{\#\,Record(LHS, RHS)}{\#\,TotalRecord} \qquad (4)$$

โดยที่ *LHS* คือ ข้อมูลที่เกิดทางด้านซ้าย
RHS คือ ข้อมูลที่เกิดทางด้านขวา
Record (LHS,RHS) คือ จำนวนเรคคอร์ดที่ประกอบด้วยข้อมูล LHS และ RHS
TotalRecord คือ จำนวนเรคคอร์ดทั้งหมดในฐานข้อมูล

ค่าสนับสนุนที่ดี ควรมีค่าในระดับที่สูง เพราะแสดงให้เห็นถึงความมีนัยสำคัญของความสัมพันธ์ของข้อมูล และในทางตรงข้ามถ้าค่าสนับสนุนมีค่าในระดับต่ำ อาจจะแสดงให้ เห็นถึงความไม่มีนัยสำคัญของความสัมพันธ์

ค่าสนับสนุนขั้นต่ำ (Minimum Support) เป็นค่าสนับสนุนขั้นต่ำของชุดข้อมูลที่กำหนดโดยผู้ใช้งาน ซึ่งส่งผลต่อความสัมพันธ์ของข้อมูลที่จะแสดงเฉพาะความสัมพันธ์ข้อมูลที่มีค่าสนับสนุนมากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำ

4.2 ค่าความเชื่อมั่น (Confidence)

ค่าความเชื่อมั่น เป็นค่าเปอร์เซ็นต์ของความถี่ของเหตุการณ์อื่นที่เกิดร่วมกับข้อมูลนั้น หรือบอกว่ากฎหนึ่งมีความน่าเชื่อถือเพียงใด เช่น เมื่อมีข้อมูล A เกิดขึ้น จะมีข้อมูล B เกิดขึ้นด้วย หมายความว่า จำนวนเรคคอร์ดที่ประกอบด้วยข้อมูล A และมีข้อมูล B เกิดขึ้นด้วย เมื่อเทียบกับจำนวนเรคคอร์ที่ประกอบด้วยข้อมูล A ทั้งหมด สามารถแสดงสมการ ดังนี้

$$Confidence = \frac{\#\ Record(LHS, RHS)}{\#\ Record(LHS)} \qquad (5)$$

โดยที่ *LHS* คือ ข้อมูลที่เกิดทางด้านซ้าย

RHS คือ ข้อมูลที่เกิดทางด้านขวา

Record (LHS,RHS) คือ จำนวนเรคคอร์ดที่ประกอบด้วยข้อมูล LHS และ RHS

Record (LHS) คือ จำนวนเรคคอร์ดที่ประกอบด้วยข้อมูล LHS ทั้งหมด

ค่าความเชื่อมั่นที่ดี ควรจะมีค่าในระดับที่สูง เพราะแสดงให้เห็นถึงความน่าเชื่อมั่นของกฎว่า เมื่อเหตุการณ์หนึ่งเกิดขึ้น จะสัมพันธ์กับอีกเหตุการณ์หนึ่งมีความน่าเชื่อมั่นเพียงใด

ค่าความเชื่อมั่นขั้นต่ำ (Minimum Confidence) เป็นค่าความเชื่อมั่นขั้นต่ำของกฎ ที่กำหนดโดยผู้ใช้งาน ซึ่งจะส่งผลต่อการสร้างกฎของขั้นตอนวิธีที่จะแสดงผลเฉพาะกฎความสัมพันธ์ที่มีค่าความเชื่อมั่นที่มากกว่าหรือเท่ากับค่าความเชื่อมั่นขั้นต่ำ

5. ขั้นตอนวิธีเอพริออริ (Apriori Algorithm)

ขั้นตอนวิธีเอพริออริ (Agrawal *et al.*, 1994) เป็นขั้นตอนวิธีที่นิยมใช้อย่างแพร่หลายในการหากฎความสัมพันธ์ สำหรับข้อมูลที่เกิดขึ้นบ่อยในฐานข้อมูล โดยจะค้นหาแบบวงกว้างก่อนนับเรคคอร์ด ซึ่งจะสร้างและตรวจสอบเซตไอเท็มที่เกิดบ่อยทีละชั้น ขั้นตอนการทำงานของขั้นตอนวิธีเอพริออริ ดังภาพที่ 1 ดังนี้

ขั้นตอนวิธี เอพริออริ

Assign minimum support and minimum confidence

(1) Let F_k be the set of frequent itemsets of size k

(2) Let C_k be the set of candidate itemsets of size k

(3) Let F_1 be the set of large items

(4) Let k = 1

(5) For all items in frequent itemsets F_k repeat step 5.1 – 5.3

(5.1) Generate new candidates $C_k + 1$ from F_k

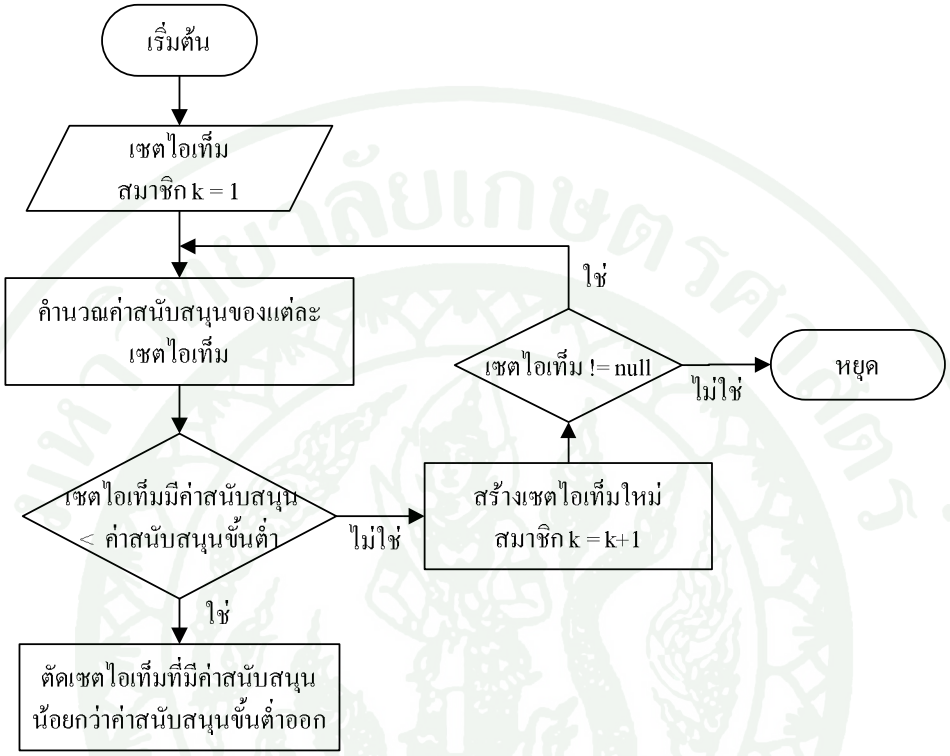
(5.2) For each transaction T in the database, increment the count of all candidates in $C_k + 1$ that are contained in T

(5.3) Generate frequent itemsets $F_k + 1$ of size k from candidates in $C_k + 1$ with minimum support

The final solution is $U_k F_k$

ภาพที่ 1 ขั้นตอนวิธีเอพริออริ

จากภาพที่ 1 (ดูภาพที่ 2 ประกอบ) ขั้นตอนการทำงานขั้นตอนวิธีเอพริออริ ข้อมูลนำเข้า คือ เซตไอเท็ม หมายถึง เซตของแอททริบิวท์ในข้อมูลตัวอย่าง จากนั้นกำหนดค่าสนับสนุนขั้นต่ำและค่าความเชื่อมั่นขั้นต่ำ ขั้นตอนวิธีเริ่มจากตรวจสอบไอเท็ม คือ แอททริบิวท์ ให้เซตไอเท็มเริ่มต้นที่มีจำนวนสมาชิกเท่ากับหนึ่ง แล้วคำนวณค่าสนับสนุนของแต่ละเซตไอเท็ม ถ้าเซตไอเท็มใดมีค่าสนับสนุนน้อยกว่าค่าสนับสนุนขั้นต่ำก็จะตัดไอเท็มนั้นออก ไม่นำไปสร้างเซตไอเท็มในชั้นถัดไป ในทางกลับกันถ้าเซตไอเท็มใดมีค่าสนับสนุนมากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำจะนำไปสร้างเป็นเซตไอเท็มในชั้นถัดไป โดยเพิ่มจำนวนสมาชิกในเซตไอเท็มอีกหนึ่ง แล้วคำนวณค่าสนับสนุน การทำงานของขั้นตอนวิธีจะวนไปเรื่อยๆ จนกระทั่งไม่เหลือเซตไอเท็มที่สร้างในชั้นถัดไปได้



ภาพที่ 2 ขั้นตอนการทำงานของขั้นตอนวิธีเอพริออริ

ตัวอย่างขั้นตอนการหากฎความสัมพันธ์ โดยใช้ขั้นตอนวิธีเอพริออริ จากตัวอย่างข้อมูลการขายสินค้าตามตารางที่ 2 ดังนี้

ตารางที่ 2 ตัวอย่างข้อมูลการขายสินค้า

ลูกค้า	สินค้าที่ซื้อ
2000	A, B, C
1000	A, C
4000	A, D
5000	B, E, F

ขั้นตอนการทำงานของขั้นตอนวิธีเอพริออริ มีดังนี้

1. ผู้ใช้กำหนดค่าเริ่มต้น 2 ค่า คือ คือสนับสนุนขั้นต่ำ และค่าความเชื่อมั่นขั้นต่ำ สมมติให้ค่าสนับสนุนขั้นต่ำและค่าความเชื่อมั่นขั้นต่ำอย่างละ 50% ตามลำดับ

2. เริ่มพิจารณาจากเซตไอเท็มที่มีจำนวนสมาชิก $k = 1$ คำนวณหาความถี่ของไอเท็มที่เกิดขึ้นบ่อยในฐานข้อมูล โดยคำนวณหาค่าสนับสนุนตามสมการ (4) ได้ผลตามตารางที่ 3

ตารางที่ 3 คำนวณค่าสนับสนุนของเซตไอเท็มที่มีจำนวนสมาชิก 1 ตัว

ข้อมูลไอเท็ม สมาชิก 1 ตัว	จำนวนที่นับได้	ค่าสนับสนุน (%)
{A}	3	(3/4)*100 = 75
{B}	2	(2/4)*100 = 50
{C}	2	(2/4)*100 = 50
{D}	1	(1/4)*100 = 25
{E}	1	(1/4)*100 = 25
{F}	1	(1/4)*100 = 25

3. เปรียบเทียบค่าสนับสนุน ถ้าค่าสนับสนุนมีค่าน้อยกว่าค่าสนับสนุนขั้นต่ำจะไม่นำไอเท็มนั้นมาพิจารณาในรอบถัดไป ตามตารางที่ 4

ตารางที่ 4 คัดเลือกไอเท็มจากการเปรียบเทียบค่าสนับสนุนรอบ $k = 1$

เซตไอเท็ม สมาชิก 1 ตัว	จำนวนที่นับได้	ค่าสนับสนุน (%)
{A}	3	(3/4)*100 = 75
{B}	2	(2/4)*100 = 50
{C}	2	(2/4)*100 = 50
~~{D}~~	~~1~~	~~(1/4)*100 = 25~~
~~{E}~~	~~1~~	~~(1/4)*100 = 25~~
~~{F}~~	~~1~~	~~(1/4)*100 = 25~~

4. นำไอเท็มที่ได้จากขั้นตอนที่ 3 มาพิจารณาเพื่อหาเซตไอเท็มรอบถัดไปที่มีจำนวนสมาชิก k = 2 แล้วทำตามขั้นตอนที่ 2 และ 3 ได้ผลตามตารางที่ 5 และตารางที่ 6 ตามลำดับ

ตารางที่ 5 คำนวณค่าสนับสนุนของเซตไอเท็มที่มีจำนวนสมาชิก 2 ตัว

เซตไอเท็ม สมาชิก 2 ตัว	จำนวนที่นับได้	ค่าสนับสนุน (%)
{A,B}	1	(1/4)*100 = 25
{A,C}	2	(2/4)*100 = 50
{B,C}	1	(1/4)*100 = 25

ตารางที่ 6 คัดเลือกไอเท็มจากการเปรียบเทียบค่าสนับสนุนรอบ $k = 2$

เซตไอเท็ม สมาชิก 2 ตัว	จำนวนที่นับได้	ค่าสนับสนุน (%)
~~{A,B}~~	~~1~~	~~(1/4)*100 = 25~~
{A,C}	2	(2/4)*100 = 50
~~{B,C}~~	~~1~~	~~(1/4)*100 = 25~~

5. นำไอเท็มที่ได้จากขั้นตอนที่ 4 มาพิจารณาเพื่อหาเซตไอเทมที่มีจำนวนสมาชิก $k = 3$ แต่เนื่องจากผลลัพธ์ที่ได้มีไอเท็มเพียง 2 ตัว ทำให้ไม่สามารถหาจำนวนสมาชิก 3 ตัวได้ จึงจบการทำงาน ดังนั้นเซตไอเท็มที่เกิดขึ้นบ่อยในฐานข้อมูล คือ {A,C}

6. การสร้างกฎความสัมพันธ์ โดยนำเอาเซตไอเท็มที่ได้จากขั้นตอนที่ 5 มาสร้างเป็นกฎความสัมพันธ์ ซึ่งจะได้กฎความสัมพันธ์ 2 กฎ ดังนี้

กฎข้อที่ 1 $A \rightarrow C$

กฎข้อที่ 2 $C \rightarrow A$

7. นำกฎที่ได้มาตรวจสอบหาค่าความเชื่อมั่น โดยคำนวณหาค่าความเชื่อมั่นตามสมการ (5) ได้ผลตามตารางที่ 7

ตารางที่ 7 คำนวณค่าความเชื่อมั่นของแต่ละกฎความสัมพันธ์

กฎความสัมพันธ์	จำนวนที่นับได้	ค่าความเชื่อมั่น (%)
$A \rightarrow C$	2	(2/3)*100 = 66.67
$C \rightarrow A$	0	(0/3)*100 = 0

8. เปรียบเทียบค่าความเชื่อมั่น ถ้าค่าความเชื่อมั่นมีค่าน้อยกว่าค่าความเชื่อมั่นขั้นต่ำจะไม่นำกฎนั้นมาใช้ ตามตารางที่ 8

ตารางที่ 8 คัดเลือกกฎความสัมพันธ์จากการเปรียบเทียบค่าความเชื่อมั่น

กฎความสัมพันธ์	จำนวนที่นับได้	ค่าความเชื่อมั่น (%)
$A \rightarrow C$	2	(2/3)*100 = 66.67
~~$C \rightarrow A$~~	~~0~~	~~(0/3)*100 = 0~~

ดังนั้นผลลัพธ์ที่ได้จากตารางที่ 8 คือ กฎความสัมพันธ์ $A \rightarrow C$ = (50%, 66.67%) อธิบายได้ว่า มีสินค้า 100 รายการ สามารถขายสินค้า A แล้วจะมีโอกาสขายสินค้า C ด้วย 50 รายการ (50%) และโอกาสการขายสินค้า A และสินค้า C พร้อมๆกัน 67 รายการ (66.67%) ซึ่งเหตุการณ์ A คือ เหตุการณ์ที่เกิดก่อน แล้วเหตุการณ์ C คือผลที่ตามมาหลังจากเกิดเหตุการณ์ A แล้ว

6. การจำแนกข้อมูล

การจำแนกข้อมูล เป็นเทคนิคที่จำแนกกลุ่มข้อมูลด้วยคุณสมบัติต่างๆ โดยทำการสำรวจรายการในฐานข้อมูล เพื่อแยกแยะให้อยู่ในหมวดที่เราได้กำหนดไว้ล่วงหน้า ขั้นตอนการจำแนกประเภทข้อมูลมีการทำงานอยู่ 3 ขั้นตอน ดังนี้

6.1 การสร้างโมเดล เป็นขั้นตอนการสร้างโมเดลโดยอาศัยการเรียนรู้จากข้อมูลฝึก (Training Data)

6.2 การประเมินผลความถูกต้อง เป็นขั้นตอนในการทดสอบความถูกต้องของโมเดลที่ถูกสร้างขึ้น โดยใช้ข้อมูลทดสอบ (Test Data) สำหรับการประเมินผลความถูกต้อง

6.3 การจำแนกประเภทข้อมูล เป็นขั้นตอนที่นำเอาตัวโมเดลที่ผ่านการตรวจสอบความถูกต้องมาใช้กับข้อมูลที่ไม่ทราบมาก่อน (Unknown Data)

7. การเรียนรู้ของต้นไม้ตัดสินใจ

การเรียนรู้ของต้นไม้ตัดสินใจ (Ross, 1993b) เป็นเทคนิคหนึ่งที่ใช้ในการจำแนกข้อมูล และเป็นที่นิยม คุ้นเคยกันมาก เป็นเครื่องมือที่ช่วยกำหนดขอบเขตของปัญหาและช่วยสร้างทางเลือกที่เป็นไปได้ในการแก้ปัญหา

7.1 คุณลักษณะของต้นไม้ตัดสินใจ ได้แก่

7.1.1 แสดงการเชื่อมโยงความสัมพันธ์ของปัญหาได้ชัดเจน

7.1.2 ช่วยจัดการกับสถานการณ์ที่ซับซ้อนต่างๆให้อยู่ในรูปแบบที่กระชับขึ้น เพื่อช่วยให้เห็นภาพของปัญหาได้ชัดเจนยิ่งขึ้น

7.1.3 มีโครงสร้างที่สามารถบอกผลลัพธ์ที่อาจเกิดขึ้นจากการคัดเลือกทางเลือกต่างๆ สำหรับการตัดสินใจ

7.1.4 ช่วยวิเคราะห์ลำดับการตัดสินใจแก้ไขปัญหาต่างๆ พร้อมทั้งวิเคราะห์ผลลัพธ์จากการตัดสินใจด้วยแนวทางต่างๆ

7.1.5 ช่วยจัดสมดุลด้านความเสี่ยงในการตัดสินใจคัดเลือกแนวทางแก้ไขปัญหาต่างๆ เหมาะกับปัญหาที่มีจำนวนทางเลือกไม่มากนัก เนื่องจากถ้าจำนวนทางเลือกในการแก้ปัญหามีมากอาจจะทำให้โครงสร้างต้นไม้ตัดสินใจดูซับซ้อน ยุ่งเหยิง

7.2 ขั้นตอนการสร้างต้นไม้ตัดสินใจ

ขั้นตอนพื้นฐานที่ง่ายที่สุดคือ ขั้นตอนวิธิเชิงละโมบ (Greedy Algorithm) โดยการสร้างต้นไม้จะเริ่มจากบนลงล่างแบบวนซ้ำ (Recursive) ด้วยวิธีการแบ่งปัญหาใหญ่เป็นปัญหาย่อย (Divide-and-Conquer) โดยต้นไม้ตัดสินใจนี้ถูกสร้างขึ้นด้วยการเรียนรู้จากข้อมูลฝึก (Training set)

ซึ่งอาศัยค่าความไร้ระเบียบของข้อมูล (Entropy) และค่าเกน (Gain) ในการคัดเลือกแอททริบิวท์ที่มีประสิทธิภาพในการจำแนกข้อมูลสูง สมการสำหรับการคำนวณค่าความไร้ระเบียบและค่าเกนตามลำดับ ดังต่อไปนี้

$$Entropy(S) = \frac{-P}{P+N} log_2 \frac{P}{P+N} - \frac{N}{P+N} log_2 \frac{N}{P+N} \quad (6)$$

โดยที่ S คือ ชุดข้อมูลฝึก

P คือ จำนวนข้อมูลฝึกระบบคลาสบวก (Positive)

N คือ จำนวนข้อมูลฝึกระบบคลาสลบ (Negative)

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (7)$$

โดยที่ A คือ ข้อมูลแอททริบิวท์ A

$|S_v|$ คือ จำนวนข้อมูลที่พบในแอททริบิวท์ A ที่มีข้อมูลเป็น v

$|S|$ คือ จำนวนข้อมูลทั้งหมดที่พบในชุดข้อมูลฝึก

$Entropy(S)$ คือ ค่าความไร้ระเบียบของชุดข้อมูลฝึก

$Entropy(S_v)$ คือ ค่าความไร้ระเบียบของแอททริบิวท์ A ที่มีข้อมูลเป็น v

ขั้นตอนการสร้างต้นไม้ตัดสินใจ ดังนี้

7.2.1 สร้างโหนดแรกจากข้อมูลฝึก โดยการเลือกแอททริบิวท์มีค่าแตกต่างที่ดีที่สุด

7.2.2 ทำการทดสอบข้อมูล ถ้าข้อมูลทั้งหมดเป็นข้อมูลประเภทเดียวกัน ให้โหนดนั้นเป็นโหนดใบ แต่ถ้าข้อมูลต่างประเภทกัน ให้คำนวณค่าเกนของแต่ละแอททริบิวท์ ตามสมการที่ (7)

7.2.3 เลือกแอททริบิวท์ที่มีค่าเกนสูงสุด เพื่อใช้แอททริบิวท์นั้นเป็นโหนดที่แบ่งแยกข้อมูล

7.2.4 สร้างกิ่งของต้นไม้ตามแต่ละค่าของแอททริบิวท์ และแบ่งข้อมูลออกเป็นส่วนๆ

7.2.5 ทำวนซ้ำจนกว่าจะไม่สามารถแยกความแตกต่างของข้อมูลได้อีกต่อไป

7.3 การตัดเล็มต้นไม้ (Pruning)

การตัดเล็มต้นไม้ เป็นขั้นตอนการตัดโหนดในต้นไม้ตัดสินใจ เพื่อลดความซ้ำซ้อนและการรู้จำจากข้อมูลฝึกมากเกินไป ทำให้สามารถจำแนกประเภทข้อมูลได้เร็วขึ้น รูปแบบของการตัดเล็มต้นไม้ตัดสินใจแบ่งเป็น 2 รูปแบบ ดังนี้

7.3.1 การตัดเล็มก่อนการสร้างต้นไม้เสร็จเต็มรูป (Pre-Pruning) เป็นขั้นตอนการหยุดสร้างโหนดเพิ่ม เมื่อพบว่า คลาสที่พบในข้อมูลนั้นเป็นคลาสประเภทเดียวกันหมด หรือพบว่าค่าของแอททริบิวท์นั้นเป็นค่าเดียวกันทั้งหมด

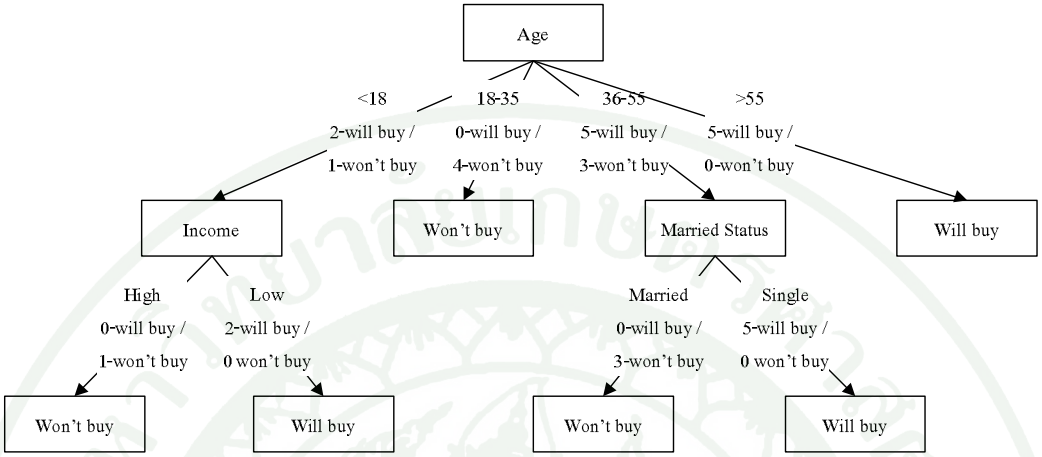
7.3.2 การตัดเล็มหลังการสร้างต้นไม้เสร็จเต็มรูป (Post-Pruning) เป็นขั้นตอนการตัดเล็มโหนดจากล่างขึ้นบน โดยเลือกต้นไม้ย่อยที่มีความลึกมากที่สุดแล้วทำการตัดต้นไม้ย่อยที่เลือกออก จากนั้นแทนต้นไม้ย่อยนั้นด้วยโหนดใบ ซึ่งการเลือกคลาสที่ใส่ในโหนดใบนั้นได้มาจากการหาคลาสที่ปรากฏอยู่ในต้นไม้ย่อยที่ตัดทิ้งไปที่มีค่ามากที่สุด (Majority Vote)

ต้นไม้ตัดสินใจจะแสดงผลลัพธ์ในรูปของโครงสร้างต้นไม้ ซึ่งสามารถแปลงเป็นกฎที่เข้าใจได้ง่าย โดยที่แต่ละโหนดของต้นไม้แสดงถึงแอททริบิวท์ แต่ละกิ่งแสดงถึงเงื่อนไขในการทดสอบ และโหนดใบแสดงถึงประเภทข้อมูลที่กำหนดไว้ โดยลักษณะของโครงสร้างต้นไม้ที่มีขนาดเล็ก ไม่ซับซ้อน ยุ่งเหยิง จะเป็นต้นไม้ที่มีคุณภาพที่ดีที่สุด ที่เหมาะสำหรับการนำไปใช้ในการตัดสินใจ

ตัวอย่างการเรียนรู้ของต้นไม้ตัดสินใจ จากตัวอย่างข้อมูลฝึก ตามตารางที่ 9 และนำข้อมูลฝึกที่ได้มาสร้างเป็นต้นไม้ตัดสินใจ ตามภาพที่ 3 ตามลำดับ ดังนี้

ตารางที่ 9 ตัวอย่างข้อมูลฝึกสำหรับการเรียนรู้ของต้นไม้ตัดสินใจ

อายุ (Age)	การศึกษา (Education)	รายได้ (Income)	สถานภาพ (Marital Status)	ซื้อสินค้าหรือไม่ (Purchase)
36-35	Master's	High	Single	Will buy
18-35	High school	Low	Single	Won't buy
36-55	Master's	Low	Single	Will buy
18-35	Bachelor's	High	Single	Won't buy
< 18	High school	Low	Single	Will buy
18-35	Bachelor's	High	Married	Won't buy
36-55	Bachelor's	Low	Married	Won't buy
> 55	Bachelor's	High	Single	Will buy
36-55	Master's	Low	Married	Wont' buy
> 55	Master's	Low	Married	Will buy
36-55	Master's	High	Single	Will buy
> 55	Master's	High	Single	Will buy
< 18	High school	High	Single	Won't buy
36-55	Master's	Low	Single	Will buy
36-55	High school	Low	Single	Will buy
< 18	High school	Low	Married	Will buy
18-35	Bachelor's	High	Married	Won't buy
> 55	High school	High	Married	Will buy
> 55	Bachelor's	Low	Single	Will buy
36-55	High school	High	Married	Won't buy



ภาพที่ 3 ต้นไม้ตัดสินใจที่ได้จากการเรียนรู้จากชุดข้อมูลฝึก

8. ขั้นตอนวิธี J48

ขั้นตอนวิธี J48 (Ross, 1993a) เป็นขั้นตอนวิธีที่รู้จักเป็นอย่างดี และมีประสิทธิภาพสูง สำหรับการสร้างต้นไม้ตัดสินใจ ซึ่งขั้นตอนการเรียนรู้ของต้นไม้ตัดสินใจจะเกี่ยวข้องกับการเลือกแอททริบิวท์ที่เหมาะสมมาสร้างเป็นโหนดต้นไม้ โดยแอททริบิวท์นั้นต้องสามารถแยกแยะข้อมูลออกเป็นประเภทที่มีการปะปนกันของข้อมูลน้อยที่สุด

วิธีการเลือกแอททริบิวท์มาสร้างโหนดต้นไม้ เป็นวิธีการเรียนรู้ที่อาศัยค่าความไร้ระเบียบของข้อมูล และค่าเกน ตามสมการที่ (6) และสมการที่ (7) ตามลำดับ หลักการการเลือกแอททริบิวท์ คือ การประเมินว่าเมื่อเลือกแอททริบิวท์ใดมาสร้างเป็นโหนดในต้นไม้แล้ว จะต้องทำให้ข้อมูลสามารถถูกจำแนกออกไปตามกิ่งของต้นไม้ได้ดีที่สุด คือ ข้อมูลที่ถูกจำแนกมีการปะปนกันของข้อมูลน้อยที่สุด

ขั้นตอนวิธี J48

input: training data

(1) Choose an attribute that best differentiates the output attribute values

(2) Create branch for each value of the chosen attribute

(3) Divide the instance into subgroup as to reflect the attribute values of the chosen node

(4) For each subgroup, terminate the attribute selection process if :

a. All members of a subgroup have the same value for the output attribute, terminate the attribute selection process for the current path and label the branch on the current path with the specified value

b. The subgroup contain a single node or no further distinguishing attribute can be determined. As in (a), label the branch with the output value seen by the majority of remaining instances

(5) For each subgroup created in (3) that has not been labeled as terminal, repeat the above process.

output: decision tree that can be used to classify each instance

ภาพที่ 4 ขั้นตอนวิธี J48

จากภาพที่ 4 การทำงานของขั้นตอนวิธี J48 เริ่มต้นจากข้อมูลฝึก และชุดของแอททริบิวท์ ขั้นตอนวิธีเริ่มจากตรวจสอบข้อมูลฝึกว่ามีคลาสเหมือนกันหรือไม่ ถ้าข้อมูลฝึกมีคลาสที่ปะปนกัน ดังนั้นจะทำการเลือกแอททริบิวท์ที่ดีที่สุด โดยการคำนวณค่าความเหมาะสม อาศัยค่าค่าความไร้ระเบียบของข้อมูล และค่าเกน การเลือกแอทริบิวพิจารณาเลือกแอทริบิวท์ที่มีค่าเกนสูงสุด จะนำแอททริบิวท์ที่ได้มาสร้างเป็นโหนดต้นไม้ โดยสร้างกิ่งเชื่อมออกจากโหนด โดยกิ่งแต่ละอันแทนค่าในแอททริบิวท์นั้นๆ และจะทำวนซ้ำจนไม่สามารถแยกความแตกต่างของคลาสภายในข้อมูลได้อีกต่อไป

9. การจำแนกข้อมูลโดยใช้กฎความสัมพันธ์ (Associative Classification)

การจำแนกข้อมูลโดยใช้กฎความสัมพันธ์ (Liu. *et al.*, 1998) เป็นการนำเทคนิคที่เกิดจากการรวมกันของเทคนิคหลักที่สำคัญ 2 เทคนิคที่ใช้สำหรับการทำเหมืองข้อมูล คือ การหากฎความสัมพันธ์ และการจำแนกข้อมูล เป้าหมายของเทคนิคการจำแนกข้อมูลโดยใช้กฎความสัมพันธ์ คือ เพื่อทำนายกลุ่มให้กับข้อมูลในอนาคตที่ไม่ทราบกลุ่ม และเป็นการเพิ่มประสิทธิภาพในการทำนาย ซึ่งให้ประสิทธิภาพในการทำนายสูง และเป็นที่นิยมในปัจจุบัน โดยขั้นตอนวิธีที่ใช้สำหรับการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ คือ CBA (Liu. *et al.*, 1998)

10. ขั้นตอนวิธี CBA (Classification based on Association rules)

ขั้นตอนวิธี CBA เป็นขั้นตอนวิธีที่อาศัยเทคนิคการหากฎความสัมพันธ์ และการจำแนกข้อมูลมาใช้สำหรับการทำนายข้อมูล โดยแบ่งขั้นตอนการทำงานเป็น 2 ส่วนหลัก คือ

10.1 การสร้างกฎความสัมพันธ์ (Rule Generator)

การสร้างกฎความสัมพันธ์ เป็นขั้นตอนที่นำเทคนิคการหากฎความสัมพันธ์เข้ามาใช้ ซึ่งอาศัยขั้นตอนวิธีเอพริออริ แต่มีข้อจำกัดในการสร้างกฎ คือ จะสนใจเฉพาะกฎที่ให้ค่าทางขวามือหรือผลของเหตุการณ์ เป็นข้อมูลคลาสเท่านั้น ซึ่งกฎแบบนี้เรียกว่า Class Association Rules (CARs) (Liu. *et al.*, 1998)

ขั้นตอนในการสร้างกฎความสัมพันธ์ มีดังนี้

10.1.1 ค้นหาเซตไอเท็มที่เกิดบ่อย โดยพิจารณาที่ค่าสนับสนุนสูงกว่าค่าสนับสนุนขั้นต่ำ ซึ่งขั้นตอนนี้จะเหมือนกับขั้นตอนวิธีเอพริออริ

10.1.2 สร้างกฎความสัมพันธ์จากเซตไอเท็มขั้นตอน 1.1 แต่พิจารณาเฉพาะกฎที่เป็น CARs และสำหรับทุกๆเซตไอเท็มที่มีเซตไอเท็มเหมือน จะเลือกเพียงกฎเดียวที่มีค่าความเชื่อมั่นสูงที่สุด

10.1.3 ทำการตัดเล็มต้นไม้ โดยการตัดเซตไอเท็มที่มีค่าความเชื่อมั่นน้อยกว่าค่าความเชื่อมั่นขั้นต่ำ

10.2 การสร้างโมเดลการทำนาย (Classifier Builder)

การสร้างโมเดลการทำนาย เป็นขั้นตอนที่อาศัยเทคนิคการเรียนรู้ของต้นไม้ตัดสินใจด้วยขั้นตอนวิธี J48 โดยนำเอาเซตของกฎ CARs จากขั้นตอนการสร้างกฎความสัมพันธ์มาจัดเรียงและสร้างโมเดลสำหรับทำนายข้อมูล

หลักการสำหรับการเรียงกฎ มีดังนี้

กำหนดให้ 2 กฎ คือ r_i และ r_j สามารถกำหนดว่า r_i มีศักย์สูงกว่า r_j เมื่อ

1. ค่าความเชื่อมั่นของ r_i มากกว่า r_j หรือ
2. ค่าความเชื่อมั่นเท่ากัน แต่ค่าสนับสนุนของ r_i มากกว่า r_j หรือ
3. ค่าความเชื่อมั่นและค่าสนับสนุนเท่ากัน แต่ r_i ถูกสร้างมาก่อน r_j

ขั้นตอนการสร้างโมเดลสำหรับทำนายข้อมูล มีดังนี้

10.2.1 เรียงลำดับกฎทั้งหมด โดยอาศัยหลักการสำหรับการเรียงกฎที่กล่าวมา

10.2.2 เรียนรู้กฎด้วยการวนลูปกฎทั้งหมด โดยนำกฎแต่ละกฎไปเช็คกับข้อมูลที่ทราบคำตอบ เรียกว่า ข้อมูลฝึก ถ้ากฎสามารถรองรับข้อมูลทั้งค่าด้านซ้ายและค่าด้านขวา จะนำกฎไปเก็บไว้เป็นเซตของกฎ แล้วหาค่าดีฟอลต์คลาส (Default Class) เพื่อใช้สำหรับการทำนายข้อมูลให้กับกรณีที่ไม่มีกฎไหนรองรับข้อมูล จากนั้นนำเซตของกฎมาคำนวณหาเปอร์เซ็นต์ข้อผิดพลาด

10.2.3 นำเซตของกฎทุกเซตมาเปรียบเทียบกัน แล้วเลือกเซตของกฎที่มีเปอร์เซ็นต์ความผิดพลาดน้อยที่สุด แล้วนำเซตของกฎนั้นมาสร้างเป็นโมเดลสำหรับการทำนายข้อมูลที่เราไม่ทราบกลุ่มต่อไป

งานวิจัยที่เกี่ยวข้อง

ในปัจจุบันเทคนิคการจำแนกข้อมูลโดยอาศัยกฎความสัมพันธ์ เป็นเทคนิคหนึ่งที่ใช้ในการจำแนกข้อมูลที่ได้รับความสนใจ และให้ความแม่นยำสูงสำหรับการทำนาย เนื่องจากเป็นการนำเอาเทคนิคการหากฎความสัมพันธ์เข้ามาร่วมใช้กับเทคนิคการจำแนกข้อมูล โดยได้มีงานวิจัยหลายงานที่เสนอเทคนิคดังกล่าว

งานวิจัยที่นำเสนอเทคนิคการจำแนกข้อมูลโดยใช้กฎความสัมพันธ์เป็นคนแรก ได้นำเสนอขั้นตอนวิธี CBA (Classification Based on Association Rules) (Liu. *et al.*, 1998) ที่มีหลักการในการพิจารณากฎความสัมพันธ์ที่มีค่าความเชื่อมั่นของกฎสูงที่สุด โดยการทำงานของขั้นตอนวิธี CBA ประกอบด้วย 2 ส่วนคือ ส่วนการสร้างกฎความสัมพันธ์ เรียกว่า CBA-RG (CBA- Rule Generator) ซึ่งจะอาศัยขั้นตอนวิธีเอพริออริ และส่วนการสร้างตัวจำแนกประเภทข้อมูล เรียกว่า CBA-CB (CBA-Classifier Builder) โดยผลการวิจัยได้ทำการเปรียบเทียบกับต้นไม้ตัดสินใจ ทำให้พบว่าตัวจำแนกประเภทข้อมูลด้วยวิธี CBA สามารถให้ผลที่มีประสิทธิภาพที่ดีกว่า

งานวิจัยถัดมาได้นำเอางานวิจัยของ Bing Liu. มาพัฒนาต่อ เนื่องจากในงานวิจัยของ Bing Liu. จะพิจารณากฎความสัมพันธ์ที่มีค่าความเชื่อมั่นของกฎสูงสุดเพียงกฎเดียว ซึ่งอาจจะทำให้เกิดข้อมูลผิดพลาดได้ ทำให้ส่งผลต่อความแม่นยำในการจำแนกข้อมูล จึงได้นำเสนอขั้นตอนวิธี CMAR (Classification Based on Multiple Class Association Rules) (Li. *et al.*, 2001) ซึ่งจะพิจารณากฎความสัมพันธ์หลายกฎพร้อมกัน โดยอาศัยเทคนิค FP-Growth, FP-Tree ในการหารูปแบบของข้อมูล และใช้เทคนิค CR-Tree ในการเก็บและคืนกฎที่มีประสิทธิภาพ และทำการตัดเล็มกฎโดยอาศัยค่าความเชื่อมั่น ค่าความสัมพันธ์ (Correlation) และค่าความครอบคลุมในชุดข้อมูล (Database Coverage) ส่วนการหารูปแบบความสัมพันธ์จะใช้ค่าไค สแควร์ (Chi Square) ในการคำนวณหาความสัมพันธ์ของกฎ ในการทดลองได้ทดลองกับข้อมูลจำนวน 26 ชุดข้อมูล ผลการทดลองได้ทำการเปรียบเทียบกับเทคนิคต้นไม้ตัดสินใจ และขั้นตอนวิธี CBA ซึ่งผลของงานวิจัยพบว่า CMAR มีประสิทธิภาพในการจำแนกข้อมูลที่พิจารณาจากกฎความสัมพันธ์หลายกฎให้ประสิทธิภาพในการจำแนกข้อมูลที่สูงกว่า 2 เทคนิคที่นำมาเปรียบเทียบ

งานวิจัยถัดมานำเสนอวิธีการเพิ่มประสิทธิภาพในการจำแนกข้อมูล โดยนำเสนอขั้นตอนวิธี CBMAR (Classification Based on Multiple Level Classes Association Rules) (วีระพล และ

คณะ, 2548) ที่พิจารณากฎความสัมพันธ์หลายกฎพร้อมกันในการทำนาย และเพิ่มเติมในส่วนของจะให้ความสำคัญกับกฎความสัมพันธ์ที่มีความยาวมากที่สุดก่อน โดยเหตุผลที่เลือกพิจารณากฎความสัมพันธ์ที่มีความยาวมากที่สุด เพราะว่ากฎที่มีความยาวที่สุดนั้นจะเป็นกฎที่มีความใกล้เคียงหรือคล้ายคลึงกับข้อมูลทดสอบมากที่สุด โดยได้นำเสนอสูตรในการคำนวณเพื่อหาความสัมพันธ์ของกฎ 4 วิธี คือ หาคลาสที่มีกฎความสัมพันธ์อยู่มากที่สุด หาคลาสที่มีค่าความเชื่อมั่นเฉลี่ยมากที่สุด หาคลาสที่มีค่าความเชื่อมั่นถ่วงน้ำหนักมากที่สุด และหาคลาสที่มีผลรวมของทุกๆระดับของค่าความเชื่อมั่นถ่วงน้ำหนักมากที่สุด ซึ่งได้ทำการทดลองกับข้อมูลจำนวน 2 ชุดข้อมูล คือ ข้อมูล Zoo กับข้อมูล Glass โดยได้ทำการเปรียบเทียบกับ 2 เทคนิคคือ CBA และ CMAR ซึ่งผลการวิจัยพบว่า CBMAR มีประสิทธิภาพการทำนายข้อมูลแม่นยำกว่าอีก 2 เทคนิค

งานวิทยานิพนธ์นี้ ผู้วิจัยได้นำเสนอขั้นตอนวิธี AssoTree ประกอบด้วยขั้นตอน 3 ขั้นตอนในการดำเนินงานคือ ขั้นตอนการเตรียมข้อมูล ขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล และขั้นตอนการคัดกรองแอททริบิวท์ ซึ่งพื้นฐานของการทดลองจะมีขั้นตอนเช่นเดียวกับงานของ Bing Liu. แต่มีส่วนที่แตกต่างกันนั่นคือ ในขั้นตอนการของสร้างกฎความสัมพันธ์ ได้มีการพิจารณาข้อมูลแอททริบิวท์ที่เกิดขึ้นจากกฎความสัมพันธ์ที่ได้ และนำข้อมูลแอททริบิวท์นั้นมาใช้พิจารณาในการสร้างตัวจำแนกข้อมูล สำหรับการสร้างโมเดลการเรียนรู้ ซึ่งได้ทำการเปรียบเทียบขั้นตอนวิธี AssoTree กับขั้นตอนวิธี J48 และขั้นตอนวิธี CBA

อุปกรณ์และวิธีการ

อุปกรณ์

1. ฮาร์ดแวร์ (Hardware)

เครื่องคอมพิวเตอร์ส่วนบุคคล CPU Intel Core i5 2.27 GHz, 4 GB DDR3, HD 500 GB

2. ซอฟต์แวร์ (Software)

2.1 โปรแกรม Weka เวอร์ชั่น 3.6.2

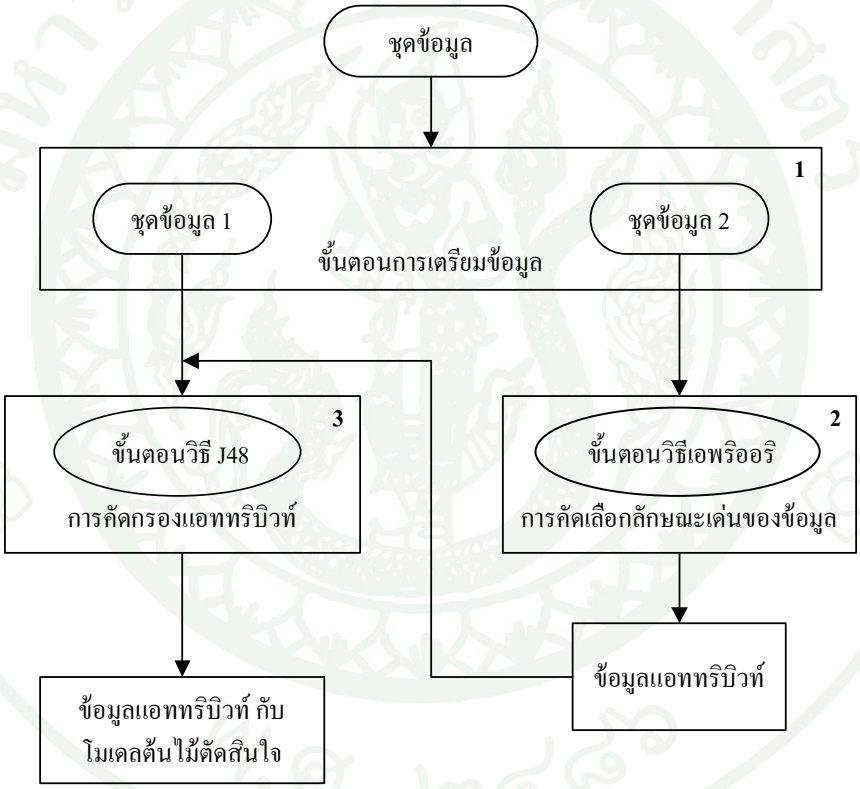
2.2 โปรแกรม Eclipse

3. ระบบปฏิบัติการ (Operating System)

Microsoft Windows 7

วิธีการ

เนื้อหาส่วนนี้จะกล่าวถึงข้อมูลที่ใช้สำหรับการทดลอง วิธีการในการดำเนินการวิจัย โดยขั้นตอนการทำงานในงานวิจัยประกอบด้วย 3 ขั้นตอน คือ ขั้นตอนการเตรียมข้อมูล ขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล และขั้นตอนการคัดกรองแอททริบิวท์ ดังภาพที่ 5



ภาพที่ 5 วิธีการในการดำเนินการวิจัย

1. ข้อมูลที่ใช้สำหรับการทดลอง (Data set)

ข้อมูลที่ใช้สำหรับการทดลองมีทั้งสิ้น 4 ชุดข้อมูล ซึ่งได้มาจากฐานข้อมูล UCI Machine Learning Repository (Aha *et al.*, 1987) ตามตารางที่ 11 ดังนี้

ตารางที่ 11 ชุดข้อมูลที่ใช้สำหรับการทดลอง

ชุดข้อมูล	จำนวนแอททริบิวท์	จำนวนคลาสข้อมูล	จำนวนเรคคอร์ด
Cleve	15	2	303
Crx	16	2	690
Hepatitis	20	2	155
Horse	28	2	368

ข้อมูลสำหรับการทดลองทั้ง 4 ชุดข้อมูล มีรายละเอียดของข้อมูล ดังต่อไปนี้

1.1 ชุดข้อมูล Cleve เป็นข้อมูลผู้ป่วยโรคหัวใจ เนื่องจากเป็นข้อมูลเกี่ยวกับทางการแพทย์ ดังนั้นชื่อแอททริบิวท์บางแอททริบิวท์ และค่าบางค่าได้ถูกเปลี่ยนชื่อไม่ให้มีความหมาย ประกอบด้วยรายละเอียด ดังนี้

1.1.1 จำนวนแอททริบิวท์ทั้งสิ้น 15 แอททริบิวท์ เป็นข้อมูลชนิดโนมินัล 9 แอททริ-บิวท์ และข้อมูลชนิดตัวเลขต่อเนื่อง 6 แอททริบิวท์

1.1.2 จำนวนคลาสข้อมูลมี 2 คลาสข้อมูล คือ คลาส buff และ คลาส sick

1.1.3 จำนวนเรคคอร์ดทั้งสิ้น 303 เรคคอร์ด

1.1.4 ข้อมูลขาดหายไป คิดเป็น 0.15%

1.2 ชุดข้อมูล Crx เป็นข้อมูลการอนุมัติบัตรเครดิต เนื่องจากข้อมูลเป็นความลับ ดังนั้นชื่อแอททริบิวท์ และค่าของแอททริบิวท์ ได้ถูกเปลี่ยนเป็นชื่อที่ไม่มีความหมาย ประกอบด้วยรายละเอียด ดังนี้

1.2.1 จำนวนแอททริบิวท์ทั้งสิ้น 16 แอททริบิวท์ เป็นข้อมูลชนิดโนมินัล 10 แอททริ-บิวท์ และข้อมูลชนิดตัวเลขต่อเนื่อง 6 แอททริบิวท์

1.2.2 จำนวนคลาสข้อมูลมี 2 คลาสข้อมูล คือ คลาส positive และคลาส negative

1.2.3 จำนวนเรคคอร์ดทั้งสิ้น 690 เรคคอร์ด

1.2.4 ข้อมูลขาดหายไป คิดเป็น 0.61%

1.3 ชุดข้อมูล Hepatitis เป็นข้อมูลโรคตับอักเสบ ประกอบด้วยรายละเอียด ดังนี้

1.3.1 จำนวนแอททริบิวท์ทั้งสิ้น 20 แอททริบิวท์ เป็นข้อมูลชนิดโนมินัล 14 แอททริ-บิวท์ และข้อมูลชนิดตัวเลขต่อเนื่อง 6 แอททริบิวท์

1.3.2 จำนวนคลาสข้อมูลมี 2 คลาสข้อมูล คือ คลาส die และ คลาส live

1.3.3 จำนวนเรคคอร์ทั้งสิ้น 155 เรคคอร์ด

1.3.4 ข้อมูลขาดหายไปคิดเป็น 5.39%

1.4 ชุดข้อมูล Horse เป็นข้อมูลอาการโคลิกในม้า ประกอบด้วยรายละเอียด ดังนี้

1.4.1 จำนวนแอททริบิวท์ทั้งสิ้น 28 แอททริบิวท์ เป็นข้อมูลชนิดโนมินัล 17 แอททริ-บิวท์ และข้อมูลชนิดตัวเลขต่อเนื่อง 11 แอททริบิวท์

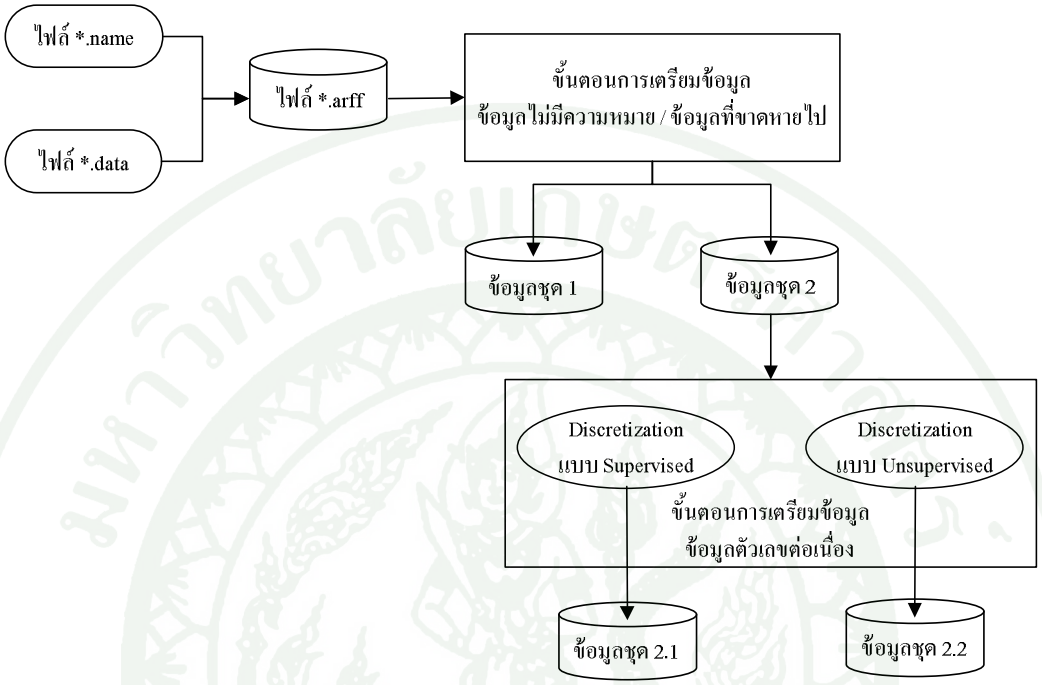
1.4.2 จำนวนคลาสข้อมูลมี 2 คลาสข้อมูล คือ คลาส no และ คลาส yes

1.4.3 จำนวนเรคคอร์ดทั้งสิ้น 368 เรคคอร์ด

1.4.4 ข้อมูลขาดหายไปคิดเป็น 18.70%

2. ขั้นตอนการเตรียมข้อมูล

ขั้นตอนการเตรียมข้อมูล เป็นขั้นตอนของการจัดการกับข้อมูลที่ไม่สมบูรณ์ เพื่อให้เหมาะสมกับการนำไปใช้กับขั้นตอนวิธีในแต่ละขั้นตอนของการทดลอง ดังภาพที่ 6 ดังนี้



ภาพที่ 6 ขั้นตอนการเตรียมข้อมูล

2.1 หลักเกณฑ์การพิจารณาการเตรียมข้อมูล ดังต่อไปนี้

2.1.1 กรณีข้อมูลมีแอททริบิวท์ที่ไม่มีความหมาย หรือไม่จำเป็นต่อการประมวลผล เช่น ลำดับของข้อมูล รหัสโรงพยาบาล จะทำการตัดแอททริบิวท์นั้นออกจากฐานข้อมูล

2.1.2 กรณีข้อมูลมีค่าบางส่วนขาดไปหาย หรือเป็นช่องว่าง โดยค่าข้อมูลที่หายไปเป็นค่าโนมินัล หรือค่าตัวเลขจะทำการเติมค่าข้อมูลที่เป็นไปได้ ด้วยการคำนวณจากค่าฐานนิยม (Mode) และค่าเฉลี่ย (Mean) โดยอาศัยขั้นตอนวิธี ReplaceMissingValue ในโปรแกรมเวกา แบ่งเป็น 2 แบบ ได้แก่

ก. ถ้าข้อมูลเป็นข้อมูลชนิดโนมินัล จะเติมค่าข้อมูลด้วยค่าฐานนิยม

ข. ถ้าข้อมูลเป็นข้อมูลชนิดตัวเลข จะเติมค่าข้อมูลด้วยค่าเฉลี่ย

2.1.3 กรณีข้อมูลเป็นข้อมูลชนิดตัวเลขต่อเนื่อง จะทำการลดรูปข้อมูลชนิดตัวเลขต่อเนื่องให้เป็นข้อมูลชนิดโนมินัล โดยอาศัยขั้นตอนวิธีในโปรแกรมเวกา แบ่งเป็น 2 แบบ ได้แก่

ก. แบบ Supervised (Fayyad *et al.*, 1993) จะนำเอาคลาสข้อมูลมาพิจารณาในการลดรูปข้อมูลอาศัยเทคนิคเอ็นโทรพีเบส (Entropy-based)

ข. แบบ Unsupervised (Dougherty *et al.*, 1995) จะไม่นำเอาคลาสข้อมูลมาใช้พิจารณาอาศัยเทคนิคบินนิง (Binning)

2.2 ขั้นตอนการเตรียมข้อมูล (ดูภาพที่ 6 ประกอบ) มีดังนี้

2.2.1 จัดรูปแบบไฟล์ชุดข้อมูลที่ได้จากฐานข้อมูลให้เป็นรูปแบบไฟล์ของโปรแกรมเวกา คือ ไฟล์ ARFF (Attribute-Relation File Format) ในแต่ละชุดข้อมูลที่ได้จากฐานข้อมูล จะประกอบด้วยไฟล์ 2 ชนิด คือ

ก. ไฟล์ชื่อ (*.name) เป็นไฟล์ที่อธิบายรายละเอียดต่างๆของชุดข้อมูล ได้แก่ ชื่อชุดข้อมูล แหล่งที่มา ความสัมพันธ์ ชื่อและจำนวนแอททริบิวท์ ประเภทข้อมูลของแต่ละแอท-ทริบิวท์ ค่าความหมายข้อมูลของแต่ละแอททริบิวท์ข้อมูลคลาส เป็นต้น

ข. ไฟล์ข้อมูล (*.data) เป็นไฟล์ที่เก็บเฉพาะข้อมูลดิบ

ไฟล์ ARFF เป็นไฟล์ที่แสดงชื่อชุดข้อมูล รายการของแอททริบิวท์ และรายละเอียดของข้อมูล โดยรูปแบบของไฟล์ ARFF ดังตัวอย่างต่อไปนี้

```
% 1. Title: Iris Plants Database                                    ┐
%                                                                   │
% 2. Sources:                                                       │
%      (a) Creator: R.A. Fisher                                     ├ ส่วนคอมเมนต์
%      (b) Donor: Michael Marshall                                  │
%      (c) Date: July, 1988                                         │
%                                                                   ┘
@RELATION iris

@ATTRIBUTE sepallength       REAL                                   ┐
@ATTRIBUTE sepalwidth        REAL                                   ├ ส่วนประกาศแอททริบิวท์
@ATTRIBUTE petallength       REAL                                   │
@ATTRIBUTE petalwidth        REAL                                   │
@ATTRIBUTE class     {Iris-setosa,Iris-versicolor,Iris-virginica}   ┘
@DATA                                                               ┐
5.1,3.5,1.4,0.2,Iris-setosa                                         │
4.9,3.0,1.4,0.2,Iris-setosa                                         ├ ส่วนข้อมูล
4.7,3.2,1.3,0.2,Iris-setosa                                         │
4.6,3.1,1.5,0.2,Iris-setosa                                         ┘
```

2.2.2 จัดการกับข้อมูลที่ไม่มีความหมาย และข้อมูลที่ขาดหายไปตามวิธีการพิจารณาการเตรียมข้อมูลข้อที่ 2.1.1 และ 2.1.2 ตามลำดับ

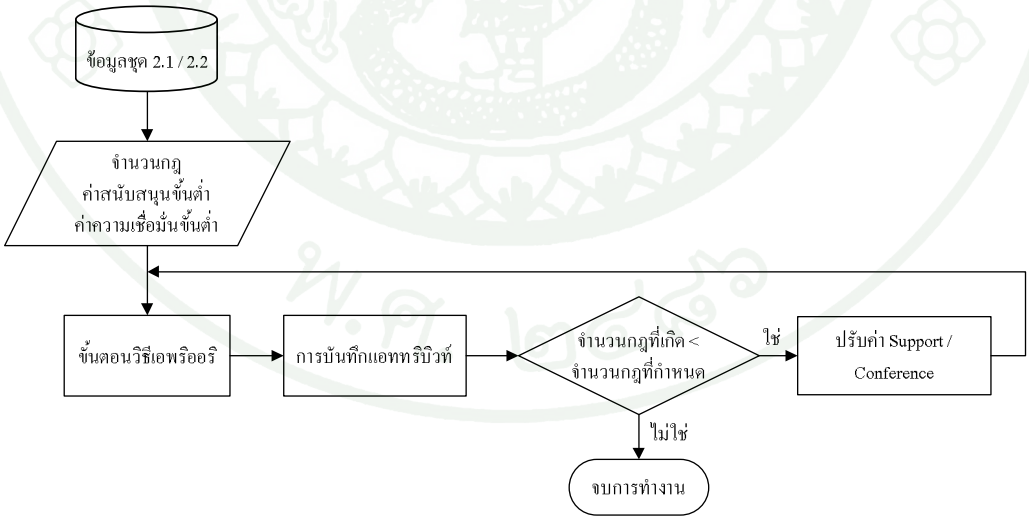
2.2.3 คัดลอกข้อมูลเป็น 2 ชุด ดังนี้

ก. ข้อมูลชุดที่ 1 เป็นชุดข้อมูลดั้งเดิม สำหรับการทดลองขั้นตอนการคัดกรองแอททริบิวท์

ข. ข้อมูลชุดที่ 2 เป็นชุดข้อมูลสำหรับการทดลองขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล เนื่องจากข้อมูลยังมีข้อมูลทั้งชนิดโนมินัล และชนิดตัวเลขต่อเนื่อง และในขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล มีข้อจำกัดที่สามารถใช้ได้กับข้อมูลที่เป็นชนิดโนมินัลเท่านั้น ดังนั้นต้องทำการลดรูปข้อมูลชนิดตัวเลขต่อเนื่อง ให้เป็นข้อมูลโนมินัลก่อนนำไปใช้ ตามวิธีการพิจารณาการเตรียมข้อมูลข้อที่ 2.1.3

3. ขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล (Feature Selection)

ขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล เป็นขั้นตอนในการหาแอททริบิวท์ที่มีความสำคัญต่อการเรียนรู้สำหรับการทำนายข้อมูลให้ได้มากที่สุด เพื่อไม่ให้ส่งผลต่อประสิทธิภาพการทำนายข้อมูล โดยอาศัยเทคนิคการหากฎความสัมพันธ์ด้วยขั้นตอนวิธีเอพริออริเข้ามาใช้สำหรับการทดลอง ซึ่งจะทำการหาความสัมพันธ์กันของข้อมูลแอททริบิวท์ผ่านกฎความสัมพันธ์ และจะให้ความสนใจกับแอททริบิวท์ที่เกิด และจะเก็บแอททริบิวท์ที่เกิดขึ้นทั้งหมดจากกฎความสัมพันธ์ไว้ แต่จะยกเว้นแอททริบิวท์ที่เป็นคลาสข้อมูล ดังภาพที่ 7 ดังนี้



ภาพที่ 7 ขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล

ขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล (ดูภาพที่ 7 ประกอบ) ดังนี้

3.1 ข้อมูลนำเข้าคือ ชุดข้อมูลที่ 2 แบ่งเป็น ข้อมูลชุด 2.1 และ ข้อมูลชุด 2.2 ที่ได้จากขั้นตอนการเตรียมข้อมูล

3.2 กำหนดค่าเริ่มต้นจำนวน 3 ค่า ได้แก่จำนวนกฎความสัมพันธ์ ค่าสนับสนุนขั้นต่ำ และค่าความเชื่อมั่นขั้นต่ำ ดังนี้

3.2.1 จำนวนกฎความสัมพันธ์ที่ต้องการ เท่ากับ 100

3.2.2 ค่าสนับสนุนขั้นต่ำ เท่ากับ 50%

3.2.3 ค่าความเชื่อมั่นขั้นต่ำ เท่ากับ 90%

3.3 คัดเลือกแอททริบิวท์จากกฎความสัมพันธ์ โดยทำผ่านขั้นตอนวิธีเอพริออริ

3.4 เก็บแอททริบิวท์ที่เกิดขึ้น โดยเรียงลำดับการเกิดก่อนและหลัง

3.5 ตรวจสอบจำนวนกฎความสัมพันธ์ที่เกิด โดยมีเงื่อนไขการทำงาน ดังนี้

3.5.1 ถ้าจำนวนกฎที่เกิดขึ้น มีค่าเท่ากับ จำนวนกฎที่กำหนด ไปทำขั้นตอนที่ 3.6

3.5.2 ถ้าจำนวนกฎที่เกิดขึ้น มีค่าน้อยกว่า จำนวนกฎที่กำหนด ไปทำขั้นตอนที่ 3.5.3

3.5.3 ตรวจสอบค่าสนับสนุนขั้นต่ำ

ก. ถ้าค่าสนับสนุนขั้นต่ำมากกว่า 0.1 ให้ลดค่าสนับสนุนขั้นต่ำลง 0.05 แล้วกลับไปทำขั้นตอนที่ 3.3

ข. ถ้าค่าสนับสนุนขั้นต่ำน้อยกว่า 0.1 หรือเท่ากับ 0.1 ไปทำขั้นตอนที่ 3.5.4

3.5.4 ตรวจสอบค่าความเชื่อมั่นขั้นต่ำ

ก. ถ้าค่าความเชื่อมั่นขั้นต่ำมากกว่า 0.8 ให้ลดค่าความเชื่อมั่นขั้นต่ำลง 0.05 แล้วกลับไปทำขั้นตอนที่ 3.3

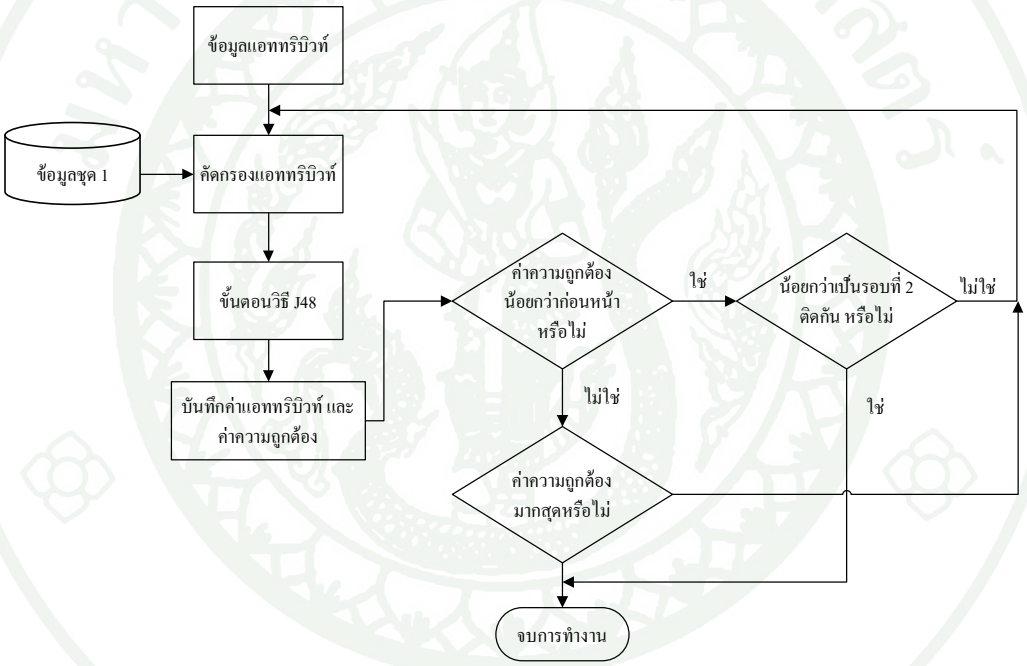
ข. ถ้าค่าความเชื่อมั่นขั้นต่ำน้อยกว่า 0.8 หรือเท่ากับ 0.8 ไปทำขั้นตอนที่ 3.6

3.6 สิ้นสุดการหากฎความสัมพันธ์

ผลลัพธ์ที่ได้จากขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล คือ แอททริบิวท์ที่เกิดขึ้นในกฎความสัมพันธ์ ซึ่งจะถือว่าแอททริบิวท์เหล่านี้มีความสำคัญต่อการนำไปใช้สร้างโมเดลจำแนกข้อมูล

4. ขั้นตอนการคัดกรองแอททริบิวท์

ขั้นตอนการคัดกรองแอททริบิวท์ เป็นขั้นตอนการหาแอททริบิวท์ที่ดีที่สุด สำหรับการสร้างโมเดลจำแนกข้อมูล โดยนำเทคนิคการเรียนรู้ของต้นไม้ตัดสินใจด้วยขั้นตอนวิธี J48 มาใช้สำหรับการทดลอง โดยในขั้นตอนนี้อาศัยข้อมูลแอททริบิวท์ที่ได้จากขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูลมาพิจารณา ดังภาพที่ 8 ดังนี้



ภาพที่ 8 ขั้นตอนการคัดกรองแอททริบิวท์

ขั้นตอนการคัดกรองแอททริบิวท์ (ดูภาพที่ 6 ประกอบ) ดังนี้

4.1 ข้อมูลนำเข้า คือ ชุดข้อมูลที่ 1 ที่ได้จากขั้นตอนการเตรียมข้อมูล

4.2 นำข้อมูลแอททริบิวท์ในชุดข้อมูลที่ 1 มาเปรียบเทียบกับชุดข้อมูลแอททริบิวท์ที่ได้จากขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล ถ้าไม่พบแอททริบิวท์ใดให้ตัดแอททริบิวท์นั้นออกจากชุดข้อมูล

4.3 คัดกรองแอททริบิวท์ ตัดแอททริบิวท์ในชุดข้อมูลออก 1 ตัว ซึ่งอ้างอิงจากข้อมูลแอท-ทริบิวท์ที่ได้จากขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล มีหลักเกณฑ์การตัดแอททริบิวท์ ดังนี้

4.3.1 กรณีครั้งแรกของการทำงาน ให้เริ่มจากข้อมูลแอททริบิวท์ตัวสุดท้าย

4.3.2 กรณีไม่ใช่ครั้งแรกของการทำงาน เป็นข้อมูลแอททริบิวท์ตัวถัดมา

4.4 นำชุดข้อมูลที่ได้จากขั้นตอน 4.3 มาผ่านเทคนิคการเรียนรู้ของต้นไม้ตัดสินใจ อาศัยขั้นตอนวิธี J48

4.5 ตรวจสอบค่าความถูกต้อง มีเงื่อนไขการทำงาน ดังนี้

4.5.1 กรณีเป็นรอบแรกของการทำงาน ให้เก็บข้อมูลแอททริบิวท์ในชุดข้อมูล และเก็บค่าความถูกต้องที่ได้ไว้ ไปทำขั้นตอนที่ 4.3

4.5.2 กรณีไม่ใช่รอบแรกของการทำงาน มีเงื่อนไข ดังนี้

ก. ถ้าค่าความถูกต้องมากกว่าค่าความถูกต้องครั้งก่อนหน้า และข้อมูลแอททริบิวท์เหลือมากกว่า 1 แอททริบิวท์ ให้เก็บข้อมูลแอททริบิวท์ในชุดข้อมูล และค่าความถูกต้องที่ได้ไว้ ไปทำขั้นตอนที่ 4.3 แต่ถ้าข้อมูลแอททริบิวท์เหลือเท่ากับ 1 แอททริบิวท์ ไปทำขั้นตอนที่ 4.6

ข. ถ้าค่าความถูกต้องน้อยกว่าค่าความถูกต้องครั้งก่อนหน้า กรณีเกิดครั้งแรกให้เก็บข้อมูลแอททริบิวท์ในชุดข้อมูล และค่าความถูกต้องที่ได้ไว้ ไปทำขั้นตอน 4.3 กรณีเกิดเป็นครั้งที่ 2 ให้ถือเอาค่าความถูกต้องที่ได้เมื่อ 2 ครั้งก่อนหน้าเป็นค่าที่ดีที่สุด ไปทำขั้นตอนที่ 4.6

4.6 สิ้นสุดการคัดกรองแอททริบิวท์

ผลลัพธ์สุดท้ายที่ได้ คือ โมเดลต้นไม้ที่ให้ค่าความถูกต้องที่ดีที่สุด ที่ประกอบด้วยแอททริบิวท์ที่มีความสำคัญที่สุดต่อการเรียนรู้ในการจำแนกข้อมูล

ผลและวิจารณ์

เนื้อหาส่วนนี้จะกล่าวถึง ผลการทดลอง และวิจารณ์ปัญหาที่เกิดขึ้นในการทดลอง

ผล

ส่วนของผลการทดลอง จะประกอบไปด้วย ผลการทดลองของแต่ละขั้นตอนการทำงาน และผลการเปรียบเทียบประสิทธิภาพในการจำแนกประเภทข้อมูล โดยนำผลการทดลองที่ได้จากขั้นตอนวิธี AssoTree เปรียบเทียบกับเทคนิคมาตรฐานการเรียนรู้ของต้นไม้ตัดสินใจ คือ J48 และเทคนิคการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ คือ CBA

1. ผลการทดลองของขั้นตอนการเตรียมข้อมูล

ผลการทดลองของขั้นตอนการเตรียมข้อมูล หลังจากผ่านการจัดการกับข้อมูลที่ไม่มีความหมาย และข้อมูลที่ขาดหายไปตามวิธีการพิจารณาการเตรียมข้อมูลข้อที่ 2.1.1 และ 2.1.2 ตามลำดับ ได้ผลตามตารางที่ 12 ดังนี้

ตารางที่ 12 ชุดข้อมูลหลังจากจัดการข้อมูลที่ไม่มีความหมาย และข้อมูลที่ขาดหายไป

ชุดข้อมูล	จำนวนแอททริบิวท์	จำนวนคลาสข้อมูล	จำนวนเรคคอร์ด
Cleve	14	2	303
Crx	16	2	690
Hepatitis	20	2	155
Horse	24	2	368

จากตารางที่ 12 จำนวนแอททริบิวท์ของชุดข้อมูลจะลดลงเมื่อเทียบกับจำนวนแอททริบิวท์ของชุดข้อมูลดังเดิมในตารางที่ 11 โดยชุดข้อมูล Cleve เหลือแอททริบิวท์ทั้งสิ้น 14 แอททริบิวท์ เนื่องจากมีแอททริบิวท์ 1 แอททริบิวท์ที่ไม่สามารถหาความหมายได้ จึงได้ตัดแอททริบิวท์นั้นทิ้ง และชุดข้อมูล Horse มีแอททริบิวท์ที่เป็นรหัสโรงพยาบาลซึ่งไม่มีความหมายต่อการทดลอง และ

แอททริบิวท์อีก 3 แอททริบิวท์เป็นข้อมูลชนิดตัวเลขต่อเนื่อง แต่ค่าของแอททริบิวท์มีค่าเป็น 00000 ทั้งหมด ซึ่งไม่สามารถหาความหมายที่มีผลต่อการทดลองได้ จึงตัดแอททริบิวท์ทั้ง 4 แอททริบิวท์ทิ้ง ทำให้เหลือแอททริบิวท์ทั้งสิ้น 24 แอททริบิวท์ สำหรับข้อมูลที่ขาดหายไปได้เติมค่าที่เป็นไปได้ ทำให้ข้อมูลที่ขาดหายไปคิดเป็น 0%

ส่วนผลการทดลองการจัดการกับข้อมูลชนิดตัวเลขต่อเนื่อง ดูที่ภาคผนวก ก

2. ผลการทดลองของขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล

ผลลัพธ์ของขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล คือข้อมูลแอททริบิวท์ที่เกิดขึ้นในกฎความสัมพันธ์ ซึ่งจะถือว่าแอททริบิวท์เหล่านี้มีความสำคัญต่อการนำไปใช้สร้างโมเดลจำแนกข้อมูล โดยแสดงข้อมูลแอททริบิวท์ตามเทคนิคการลดรูปข้อมูลแบบ Supervised และแบบ Unsupervised มีรายละเอียด ดังต่อไปนี้

ตารางที่ 13 ข้อมูลแอททริบิวท์ของชุดข้อมูล Cleve ของขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล

ลำดับ	ข้อมูลแอททริบิวท์ดั้งเดิม	แบบ Supervised	แบบ Unsupervised
1	age	cholesteral	oldpeak
2	sex	trestbps	numberOfVessle
3	chest-pain-type	fast-blood-sugar	fast-blood-sugar
4	trestbs	oldpeak	thal
5	cholesteral	sex	exercise
6	fast-blood-sugar	exercise	rest-ecg
7	rest-ecg	numberOfVessle	slope
8	max-heart-rate	max-heart-rate	sex
9	exercise	thal	chest-pain-type
10	oldpeak	age	cholesteral
11	slope	rest-ecg	-
12	numberOfVessle	-	-
13	thal	-	-
14	class	-	-

จากตารางที่ 13 ชุดข้อมูล Cleve มีแอททริบิวท์ที่เกิดขึ้นในกฎความสัมพันธ์ ดังนี้ ชุดข้อมูลที่ใช้เทคนิคการลดรูปข้อมูลแบบ Supervised มีจำนวนแอททริบิวท์ 11 แอททริบิวท์ และแบบ Unsupervised มีจำนวนแอททริบิวท์ 10 แอททริบิวท์ จากจำนวนแอททริบิวท์ดั้งเดิมมีจำนวนแอททริบิวท์ทั้งสิ้น 13 แอททริบิวท์ ไม่รวมแอททริบิวท์ที่เป็นคลาสแอททริบิวท์ แสดงให้เห็นว่ามีแอททริบิวท์ที่ไม่เกิดขึ้นในกฎความสัมพันธ์จำนวน 2 และ 3 แอททริบิวท์ตามลำดับ และการเกิดของแอททริบิวท์ รวมถึงลำดับของการเกิดก่อนและหลังของแอททริบิวท์ระหว่างชุดข้อมูลแบบ Supervised และแบบ Unsupervised ให้ผลที่ต่างกัน

ตารางที่ 14 ข้อมูลแอททริบิวท์ของชุดข้อมูล Crx ของขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล

ลำดับ	ข้อมูลแอททริบิวท์ดั้งเดิม	แบบ Supervised	แบบ Unsupervised
1	A1	A5	A13
2	A2	A4	A15
3	A3	A13	A11
4	A4	A2	A5
5	A5	A11	A4
6	A6	A10	A8
7	A7	A15	A1
8	A8	A1	A14
9	A9	A14	A10
10	A10	A7	A7
11	A11	A9	A12
12	A12	A12	-
13	A13	A3	-
14	A14	A8	-
15	A15	-	-
16	class	-	-

จากตารางที่ 14 ชุดข้อมูล Crx มีแอททริบิวท์ที่เกิดขึ้นในกฎความสัมพันธ์ ดังนี้ ชุดข้อมูลที่ใช้เทคนิคการลดรูปข้อมูลแบบ Supervised มีจำนวนแอททริบิวท์ 14 แอททริบิวท์ และแบบ Unsupervised มีจำนวนแอททริบิวท์ 11 แอททริบิวท์ จากจำนวนแอททริบิวท์ดั้งเดิมมีจำนวนแอททริบิวท์ทั้งสิ้น 15 แอททริบิวท์ ไม่รวมแอททริบิวท์ที่เป็นคลาสแอททริบิวท์ แสดงให้เห็นว่ามีแอททริบิวท์ที่ไม่เกิดขึ้นในกฎความสัมพันธ์จำนวน 1 และ 4 แอททริบิวท์ตามลำดับ และการเกิดของแอททริบิวท์ รวมถึงลำดับของการเกิดก่อนและหลังของแอททริบิวท์ระหว่างชุดข้อมูลแบบ Supervised และแบบ Unsupervised ให้ผลที่ต่างกัน

ตารางที่ 15 ข้อมูลแอททริบิวท์ของชุดข้อมูล Hepatitis ของขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล

ลำดับ	ข้อมูลแอททริบิวท์ดั้งเดิม	แบบ Supervised	แบบ Unsupervised
1	age	ALKphosphate	ascites
2	sex	age	varices
3	steroid	sgot	antivirals
4	antivirals	sex	sex
5	fatigue	varices	liverbig
6	malaise	ascites	spleenPalpable
7	anorexia	protime	anorexia
8	liverbig	antivirals	spiders
9	liverfirm	-	malaise
10	spleenPalpable	-	liverfirm
11	spiders	-	bilirubin
12	ascites	-	-
13	varices	-	-
14	bilirubin	-	-
15	ALKphosphate	-	-
16	sgot	-	-
17	albumin	-	-
18	protime	-	-
19	histology	-	-
20	class	-	-

จากตารางที่ 15 ชุดข้อมูล Hepatitis มีแอททริบิวท์ที่เกิดขึ้นในกฎความสัมพันธ์ ดังนี้ ชุดข้อมูลที่ใช้เทคนิคการลดรูปข้อมูลแบบ Supervised มีจำนวนแอททริบิวท์ 8 แอททริบิวท์ และแบบ Unsupervised มีจำนวนแอททริบิวท์ 11 แอททริบิวท์ จากจำนวนแอททริบิวท์ดั้งเดิมมีจำนวนแอททริบิวท์ทั้งสิ้น 19 แอททริบิวท์ ไม่รวมแอททริบิวท์ที่เป็นคลาสแอททริบิวท์ แสดงให้เห็นว่ามีแอททริบิวท์ที่ไม่เกิดขึ้นในกฎความสัมพันธ์จำนวน 11 และ 8 แอททริบิวท์ตามลำดับ และการเกิดของ

แอททริบิวท์ รวมถึงลำดับของการเกิดก่อนและหลังของแอททริบิวท์ระหว่างชุดข้อมูลแบบ Supervised และแบบ Unsupervised ให้ผลที่ต่างกัน

ตารางที่ 16 ข้อมูลแอททริบิวท์ของชุดข้อมูล Horse ของขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล

ลำดับ	ข้อมูลแอททริบิวท์ดั้งเดิม	แบบ Supervised	แบบ Unsupervised
1	surgery	pulse	nasogastric-reflux-PH
2	age	rectal-temp	Age
3	rectal-temp	respiratory-rate	nasogastric-reflux
4	pulse	nasogastric-reflux-PH	abdo-appearance
5	respiratory-rate	packed-cell-volume	capillary-refill-time
6	extremities-temp	-	nasogastric-tube
7	peripheral-pluse	-	abdo-total-protein
8	mucous-membranes	-	Outcome
9	capillary-refill-time	-	peripheral-pulse
10	pain	-	Feces
11	peristalsis	-	surgical-lesion
12	abdominal-distension	-	total-protein
13	nasogastric-tube	-	extremities-temp
14	nasogastric-reflux	-	peristalsis
15	nasogastric-reflux-PH	-	surgery
16	feces	-	abdomen
17	abdomen	-	abdominal-distension
18	packed-cell-volume	-	-
19	total-protein	-	-
20	abdo-appearance	-	-
21	abdo-total-protein	-	-

ตารางที่ 16 (ต่อ)

ลำดับ	ข้อมูลแอททริบิวท์ดั้งเดิม	แบบ Supervised	แบบ Unsupervised
22	outcome	-	-
23	surgical-lesion	-	-
24	class	-	-

จากตารางที่ 16 ชุดข้อมูล Horse มีแอททริบิวท์ที่เกิดขึ้นในกฎความสัมพันธ์ ดังนี้ ชุดข้อมูลที่ใช้เทคนิคการลดรูปข้อมูลแบบ Supervised มีจำนวนแอททริบิวท์ 5 แอททริบิวท์ และแบบ Unsupervised มีจำนวนแอททริบิวท์ 17 แอททริบิวท์ จากจำนวนแอททริบิวท์ดั้งเดิมมีจำนวนแอททริบิวท์ทั้งสิ้น 23 แอททริบิวท์ ไม่รวมแอททริบิวท์ที่เป็นคลาสแอททริบิวท์ แสดงให้เห็นว่ามีแอททริบิวท์ที่ไม่เกิดขึ้นในกฎความสัมพันธ์จำนวน 18 และ 6 แอททริบิวท์ตามลำดับ และการเกิดของแอททริบิวท์ รวมถึงลำดับของการเกิดก่อนและหลังของแอททริบิวท์ระหว่างชุดข้อมูลแบบ Supervised และแบบ Unsupervised ให้ผลที่ต่างกัน

3. ผลการทดลองของขั้นตอนการคัดกรองแอททริบิวท์

ผลลัพธ์ของขั้นตอนการคัดกรองแอททริบิวท์ คือ ข้อมูลแอททริบิวท์ที่สำคัญที่สุด ซึ่งแอททริบิวท์ที่ได้เป็นแอททริบิวท์ที่ให้โมเดลต้นไม้ที่มีค่าความถูกต้องดีที่สุด โดยแสดงข้อมูลแอททริบิวท์ตามเทคนิคการลดรูปข้อมูลแบบ Supervised และ แบบ Unsupervised มีรายละเอียดดังต่อไปนี้

ตารางที่ 17 ข้อมูลแอททริบิวท์ของชุดข้อมูล Cleve ของขั้นตอนการคัดกรองแอททริบิวท์

ลำดับ	แบบ Supervised	แบบ Unsupervised
1	cholesteral	oldpeak
2	trestbps	numberOfVessle
3	fast-blood-sugar	fast-blood-sugar
4	oldpeak	thal
5	sex	exercise
6	exercise	rest-ecg
7	numberOfVessle	slope
8	max-heart-rate	sex

จากตารางที่ 17 ชุดข้อมูล Cleve แอททริบิวท์ที่สำคัญที่สุดที่ได้หลังจากผ่านการคัดกรองแอททริบิวท์แล้ว ทั้งชุดข้อมูลที่ใช้เทคนิคการลดรูปข้อมูลแบบ Supervised และแบบ Unsupervised จำนวนแอททริบิวท์ที่ได้เท่ากัน คือมีจำนวนแอททริบิวท์ทั้งสิ้น 8 แอททริบิวท์ จากจำนวนแอททริบิวท์ที่ได้จากขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล มีจำนวนแอททริบิวท์ 11 และ 10 แอททริบิวท์ ตามลำดับ ไม่รวมแอททริบิวท์ที่เป็นคลาสแอททริบิวท์ แสดงให้เห็นว่ามีแอททริบิวท์ที่ไม่สำคัญต่อการนำไปใช้สร้างโมเดลจำแนกข้อมูล จำนวน 3 และ 2 แอททริบิวท์ตามลำดับ

ตารางที่ 18 ข้อมูลแอททริบิวท์ของชุดข้อมูล Crx ของขั้นตอนการคัดกรองแอททริบิวท์

ลำดับ	แบบ Supervised	แบบ Unsupervised
1	A5	A13
2	A4	A15
3	A13	A11
4	A2	A5
5	A11	A4
6	A10	A8
7	A15	A1
8	A1	A14
9	A14	A10
10	A7	A7
11	A9	-

จากตารางที่ 18 ชุดข้อมูล Crx แอททริบิวท์ที่สำคัญที่สุดที่ได้หลังจากผ่านการคัดกรองแอททริบิวท์แล้ว ดังนี้ ชุดข้อมูลที่ใช้เทคนิคการลดรูปข้อมูลแบบ Supervised มีจำนวนแอททริบิวท์ทั้งสิ้น 11 แอททริบิวท์ และแบบ Unsupervised มีจำนวนแอททริบิวท์ 10 แอททริบิวท์ จากจำนวนแอททริบิวท์ที่ได้จากขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล มีจำนวนแอททริบิวท์ 14 และ 11 แอททริบิวท์ ตามลำดับ ไม่รวมแอททริบิวท์ที่เป็นคลาสแอททริบิวท์ แสดงให้เห็นว่ามีแอททริบิวท์ที่ไม่สำคัญต่อการนำไปใช้สร้างโมเดลจำแนกข้อมูล จำนวน 3 และ 1 แอททริบิวท์ตามลำดับ

ตารางที่ 19 ข้อมูลแอททริบิวท์ของชุดข้อมูล Hepatitis ของขั้นตอนการคัดกรองแอททริบิวท์

ลำดับ	แบบ Supervised	แบบ Unsupervised
1	ALKphosphate	ascites
2	age	varices
3	sgot	antivirals
4	sex	sex
5	varices	liverbig
6	-	spleenPalpable
7	-	anorexia
8	-	spiders
9	-	malaise

จากตารางที่ 19 ชุดข้อมูล Hepatitis แอททริบิวท์ที่สำคัญที่สุดที่ได้หลังจากผ่านการคัดกรองแอททริบิวท์แล้ว ดังนี้ ชุดข้อมูลที่ใช้เทคนิคการลดรูปข้อมูลแบบ Supervised มีจำนวนแอททริบิวท์ทั้งสิ้น 5 แอททริบิวท์ และ แบบ Unsupervised มีจำนวนแอททริบิวท์ 9 แอททริบิวท์ จากจำนวนแอททริบิวท์ที่ได้จากขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล มีจำนวนแอททริบิวท์ 8 และ 11 แอททริบิวท์ ตามลำดับ ไม่รวมแอททริบิวท์ที่เป็นคลาสแอททริบิวท์ แสดงให้เห็นว่ามีแอททริบิวท์ที่ไม่สำคัญต่อการนำไปใช้สร้างโมเดลจำแนกข้อมูล จำนวน 3 และ 2 แอททริบิวท์ตามลำดับ

ตารางที่ 20 ข้อมูลแอททริบิวท์ของชุดข้อมูล Horse ของขั้นตอนการคัดกรองแอททริบิวท์

ลำดับ	แบบ Supervised	แบบ Unsupervised
1	pulse	nasogastric-reflux-PH
2	rectal-temp	Age
3	respiratory-rate	nasogastric-reflux
4	nasogastric-reflux-PH	abdo-appearance
5	-	capillary-refill-time
6	-	nasogastric-tube
7	-	abdo-total-protein
8	-	Outcome
9	-	peripheral-pulse
10	-	Feces
11	-	surgical-lesion
12	-	total-protein

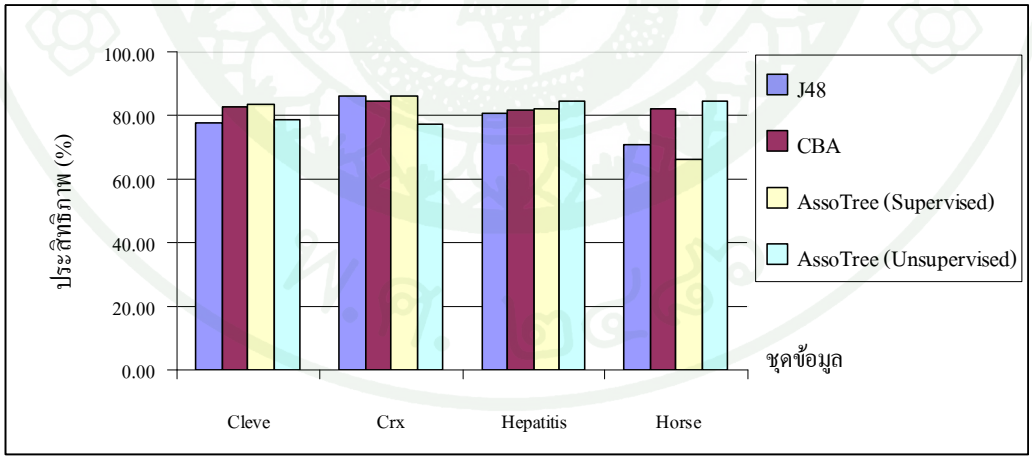
จากตารางที่ 20 ชุดข้อมูล Horse แอททริบิวท์ที่สำคัญที่สุดที่ได้หลังจากผ่านการคัดกรองแอททริบิวท์แล้ว ดังนี้ ชุดข้อมูลที่ใช้เทคนิคการลดรูปข้อมูลแบบ Supervised มีจำนวนแอททริบิวท์ทั้งสิ้น 4 แอททริบิวท์ และแบบ Unsupervised มีจำนวนแอททริบิวท์ 12 แอททริบิวท์ จากจำนวนแอททริบิวท์ที่ได้จากขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล มีจำนวนแอททริบิวท์ 5 และ 17 แอททริบิวท์ ตามลำดับ ไม่รวมแอททริบิวท์ที่เป็นคลาสแอททริบิวท์ แสดงให้เห็นว่ามีแอททริบิวท์ที่ไม่สำคัญต่อการนำไปใช้สร้างโมเดลจำแนกข้อมูล จำนวน 1 และ 5 แอททริบิวท์ตามลำดับ

4. ผลการเปรียบเทียบประสิทธิภาพระหว่างขั้นตอนวิธี J48, CBA และ AssoTree

ในการทดลองการเปรียบเทียบประสิทธิภาพนี้ ค่าพารามิเตอร์ที่ใช้จะเป็นค่ามาตรฐานที่ได้มีการกำหนดไว้แล้ว โดยทำการทดลองแบบ 10 โฟลดครอสวาดิเดชัน (10 folds cross-validation) มีรายละเอียด ดังต่อไปนี้

ตารางที่ 21 เปรียบเทียบประสิทธิภาพระหว่าง J48, CBA และ AssoTree

ชุดข้อมูล	J48	CBA	AssoTree (Supervised)	AssoTree (Unsupervised)
Cleve	77.56	82.80	83.50	78.88
Crx	85.94	84.70	86.09	77.25
Hepatitis	80.65	81.80	81.94	84.52
Horse	70.92	82.10	66.30	84.51



ภาพที่ 9 ผลการเปรียบเทียบประสิทธิภาพระหว่างขั้นตอนวิธี J48, CBA และ AssoTree

จากผลการทดลองในตารางที่ 21 และภาพที่ 9 เป็นชุดข้อมูลที่ใช้เทคนิคการลดรูปข้อมูลทั้งแบบ Supervised และแบบ Unsupervised พบว่า ขั้นตอนวิธี AssoTree แบบ Supervised ให้ประสิทธิภาพการทำนายที่ดีกว่าทั้งขั้นตอนวิธี J48 และขั้นตอนวิธี CBA จำนวน 3 ชุดข้อมูลคือ ชุด

ข้อมูล Cleve, Crx และ Hepatitis เท่ากับ 83.50%, 86.09% และ 81.94% ตามลำดับ แต่ให้ประสิทธิภาพที่ต่ำกว่าทั้งขั้นตอนวิธี J48 และขั้นตอนวิธี CBA ในชุดข้อมูล Horse เท่ากับ 66.30% และขั้นตอนวิธี AssoTree แบบ Unsupervised ให้ประสิทธิภาพการทำนายที่ดีกว่าทั้งขั้นตอนวิธี J48 และขั้นตอนวิธี CBA จำนวน 2 ชุดข้อมูลคือ ชุดข้อมูล Hepatitis และ Horse เท่ากับ 84.52% และ 84.51% ตามลำดับ และให้ประสิทธิภาพที่ดีกว่าขั้นตอนวิธี J48 แต่ให้ประสิทธิภาพที่ต่ำกว่าขั้นตอนวิธี CBA ในชุดข้อมูล Cleve เท่ากับ 78.88% และให้ประสิทธิภาพที่ต่ำกว่าทั้งขั้นตอนวิธี J48 และขั้นตอนวิธี CBA ในชุดข้อมูล Crx เท่ากับ 77.25%

วิจารณ์

งานวิจัยที่นำเสนอเป็นการนำเอาประโยชน์จากเทคนิคการหากฎความสัมพันธ์มาประยุกต์ใช้กับการจำแนกข้อมูลด้วยการเรียนรู้ของต้นไม้ตัดสินใจ เพื่อใช้ศึกษาความสำคัญของแอททริบิวท์ที่มีผลต่อการเรียนรู้ของต้นไม้ตัดสินใจ ในขั้นตอนการทดลอง ข้อจำกัดของการทำการทดลองคือ ทดลองกับข้อมูลที่มีจำนวนคลาส 2 คลาสเท่านั้น และข้อมูลต้องมีจำนวนแอททริบิวท์มาก ดังนั้นชุดข้อมูลที่ใช้ในการทดลองมีจำนวน 4 ชุดข้อมูลที่ได้มาจากฐานข้อมูล UCI Machine Learning Repository โดยมีขั้นตอนการทำงานแบ่งเป็น 3 ขั้นตอนคือ ขั้นตอนการเตรียมข้อมูล ขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล และขั้นตอนการคัดกรองแอททริบิวท์

1. ขั้นตอนการเตรียมข้อมูล

ขั้นตอนการเตรียมข้อมูล เป็นขั้นตอนที่จัดการกับชุดข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกับขั้นตอนวิธีที่จะนำไปใช้ โดยได้มีการพิจารณาตามแต่กรณีของชนิดข้อมูลในชุดข้อมูล ถ้าข้อมูลไม่มีความหมายต่อการเรียนรู้ จะตัดข้อมูลนั้นออกจากชุดข้อมูล ถ้าข้อมูลมีการขาดหายไปไม่ครบถ้วน จะอาศัยขั้นตอนวิธี ReplaceMissingValue ในโปรแกรมเวกามาเติมค่าที่ขาดหายไป โดยการคำนวณค่าฐานนิยม (Mode) และค่าเฉลี่ย (Mean) จากข้อมูลที่มีอยู่ และถ้าข้อมูลเป็นชนิดตัวเลขต่อเนื่องจะมีการลดรูปข้อมูลชนิดตัวเลขต่อเนื่อง เป็นข้อมูลโนมินัล เนื่องจากในขั้นตอนของการคัดเลือกลักษณะเด่นของข้อมูล จะใช้ได้กับข้อมูลที่เป็นชนิดโนมินัลเท่านั้น ดังนั้นอาศัยขั้นตอนวิธีการลดรูปข้อมูลชนิดตัวเลขต่อเนื่อง (Discretization) ในโปรแกรมเวกา โดยได้แบ่งเป็น 2 แบบคือ แบบ Supervised และแบบ Unsupervised แบบ Supervised จะนำเอาคลาสข้อมูลมาพิจารณาในการลดรูปข้อมูล โดยอาศัยเทคนิคเอ็นโทรพีเบส (Entropy-based) และแบบ Unsupervised จะไม่นำเอาคลาส

ข้อมูลไปใช้พิจารณา โดยอาศัยเทคนิคบินนิง (Binning) เหตุผลที่ต้องใช้การลดรูปข้อมูลชนิดตัวเลขต่อเนื่องทั้ง 2 แบบเพื่อศึกษาว่าผลลัพธ์ที่ได้จากการลดรูปข้อมูลชนิดตัวเลขต่อเนื่องที่อาศัยเทคนิคทั้ง 2 แบบจะส่งผลต่อประสิทธิภาพของการจำแนกข้อมูลหรือไม่ และให้ประสิทธิภาพอย่างไร ผลการทดลองที่ได้ ดังตารางที่ 22 ดังนี้

ตารางที่ 22 เปรียบเทียบประสิทธิภาพของขั้นตอนวิธี AssoTree ด้วยชุดข้อมูลที่อาศัยเทคนิคการลดรูปข้อมูลแบบ Supervised และแบบ Unsupervised

ชุดข้อมูล	เทคนิคการลดรูปข้อมูล	AssoTree
Cleve	Supervised	83.50
	Unsupervised	78.88
Crx	Supervised	86.09
	Unsupervised	77.25
Hepatitis	Supervised	81.94
	Unsupervised	84.52
Horse	Supervised	66.30
	Unsupervised	84.51

จากตารางที่ 22 โดยภาพรวมแสดงให้เห็นว่า ประสิทธิภาพที่ได้เมื่อเปรียบเทียบระหว่างเทคนิคการลดรูปข้อมูลแบบ Supervised และแบบ Unsupervised ของแต่ละชุดข้อมูล จะให้ประสิทธิภาพที่แตกต่างกัน โดยชุดข้อมูล Cleve และ Crx ประสิทธิภาพของการใช้เทคนิคการลดรูปข้อมูลแบบ Supervised จะให้ประสิทธิภาพที่ดีกว่าแบบ Unsupervised แต่ในชุดข้อมูล Hepatitis และ Horse เทคนิคการลดรูปข้อมูลแบบ Unsupervised จะให้ประสิทธิภาพดีกว่าแบบ Supervised ความแตกต่างที่เป็นเกิดขึ้น เนื่องจากค่าข้อมูลภายในแต่ละแอททริบิวท์ของแต่ละชุดข้อมูลที่เป็นตัวเลขต่อเนื่องก่อนการทำการลดรูปข้อมูลเป็นแบบโนมินัล มีขอบเขตของค่าตัวเลขที่แตกต่างระหว่างตัวเลขที่น้อยที่สุดกับมากที่สุดแตกต่างกัน (range) ทำให้ส่งผลต่อการคำนวณแบ่งช่วงข้อมูลเป็นข้อมูลชนิดโนมินัลได้ ดูที่ภาคผนวก ก การวิจารณ์ขั้นตอนการเตรียมข้อมูล แสดงได้ว่าข้อมูลที่ทำมาใช้ในการทดลองมีลักษณะของข้อมูลที่หลากหลาย ข้อมูลอาจจะมีการขาดหายไปไม่ครบถ้วน หรือชนิดของข้อมูลแตกต่างกัน มีทั้งข้อมูลชนิดโนมินัล และข้อมูลชนิดตัวเลขต่อเนื่อง ดังนั้นเพื่อไม่ให้ส่งผลต่อขั้นตอนการทดลองต่อไป และประสิทธิภาพการจำแนกข้อมูล การเลือก

เทคนิคมาใช้จัดการกับข้อมูล ต้องพิจารณาให้เหมาะสมตามแต่ละชนิดของข้อมูล และขั้นตอนวิธีที่จะนำเอาข้อมูลนั้นไปใช้ต่อไป

2. ขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล

ขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล เป็นขั้นตอนที่นำข้อมูลฝึกมาหาความสัมพันธ์ของข้อมูล เพื่อให้ได้แอททริบิวท์ที่มีความสำคัญ ในขั้นตอนการทดลองอาศัยเทคนิคการหากฎความสัมพันธ์ด้วยขั้นตอนวิธีเอพริออริ ซึ่งข้อจำกัดของขั้นตอนวิธีเอพริออริ คือ ใช้ได้กับข้อมูลที่เป็นข้อมูลชนิดโนมินัลเท่านั้น จึงอาศัยเทคนิคการลดรูปข้อมูลแบบ Supervised และแบบ Unsupervised มาใช้ จากนั้นกำหนดค่าเริ่มต้นจำนวน 3 ค่า คือ จำนวนกฎความสัมพันธ์ ค่าสนับสนุนขั้นต่ำ และค่าความเชื่อมั่นขั้นต่ำ โดยกำหนดจำนวนกฎความสัมพันธ์เท่ากับ 100 หมายถึงจำนวนกฎที่ต้องการหาความสัมพันธ์ เพื่อต้องการหาแอททริบิวท์ที่มีความสำคัญให้ได้มากที่สุด ให้เพียงพอต่อการนำไปใช้ในการจำแนกข้อมูล ซึ่งจำนวนกฎความสัมพันธ์จะขึ้นอยู่กับค่าสนับสนุนขั้นต่ำ และค่าความเชื่อมั่นขั้นต่ำ ได้กำหนดค่าสนับสนุนขั้นต่ำเท่ากับ 50% และค่าความเชื่อมั่น 90% ซึ่งค่าสนับสนุนขั้นต่ำจะส่งผลต่อการคัดเลือกแอททริบิวท์ที่จะนำมาสร้างเป็นกฎความสัมพันธ์ โดยถ้าค่าสนับสนุนขั้นต่ำสูงเกินไป จะทำให้ไม่ได้แอททริบิวท์ที่มากพอต่อการนำไปสร้างกฎความสัมพันธ์ แต่ถ้าค่าสนับสนุนขั้นต่ำน้อยเกินไป ก็จะทำให้แอททริบิวท์ที่ได้ไม่มีความหมายต่อกฎความสัมพันธ์ ส่วนค่าความเชื่อมั่นขั้นต่ำจะส่งผลต่อกฎความสัมพันธ์ที่ได้ โดยถ้าค่าความเชื่อมั่นขั้นต่ำมีค่าต่ำ กฎความสัมพันธ์ก็จะไม่มีความน่าเชื่อถือ แต่ถ้าค่าความเชื่อมั่นขั้นต่ำมีค่าสูง กฎความสัมพันธ์ที่ได้จะมีความน่าเชื่อถือ ความหมายคือแอททริบิวท์ที่ได้จากกความสัมพันธ์จะมีความน่าเชื่อถือไปด้วย และยังส่งผลต่อความสำคัญของลำดับการเกิดของแอททริบิวท์ โดยถือว่าแอททริบิวท์ในลำดับต้นๆ ที่เกิดจากกฎความสัมพันธ์ที่มีค่าความเชื่อมั่นสูงนั้นมีความสำคัญต่อการจำแนกข้อมูล แต่เนื่องจากวัตถุประสงค์ของการทดลองในขั้นตอนนี้ คือต้องการหาแอททริบิวท์ให้ได้มากที่สุดให้เพียงพอต่อการจำแนกข้อมูล ดังนั้นจึงได้มีการปรับค่าสนับสนุนขั้นต่ำ และค่าความเชื่อมั่นขั้นต่ำภายในการทดลอง โดยสามารถปรับค่าสนับสนุนขั้นต่ำลงถึงค่าเท่ากับ 10% และค่าความเชื่อมั่นขั้นต่ำเท่ากับ 80% จากผลการทดลองในตารางที่ 13 ถึงตารางที่ 16 แสดงข้อมูลแอททริบิวท์ของแต่ละชุดข้อมูล โดยแอททริบิวท์ที่ได้เป็นเพียงบางแอททริบิวท์เท่านั้นไม่ได้เกิดแอททริบิวท์ทั้งหมดจากแอททริบิวท์ทั้งหมดในชุดข้อมูลดิบ และจากการเปรียบเทียบข้อมูลแอททริบิวท์ที่ได้ระหว่างเทคนิคการลดรูปข้อมูลแบบ Supervised กับแบบ Unsupervised พบว่าลำดับการเกิดของแอททริบิวท์ที่ได้จะไม่เหมือนกัน เนื่องมาจากค่าข้อมูลของแอททริบิวท์ที่ผ่านการลดรูปข้อมูลด้วย

เทคนิคการลดรูปข้อมูลแบบ Supervised และแบบ Unsupervised ส่งผลต่อแอททริบิวท์ที่เกิดขึ้นในกฎความสัมพันธ์และลำดับของการเกิดขึ้นก่อนและหลังของแอททริบิวท์

3. ขั้นตอนการคัดกรองแอททริบิวท์

ขั้นตอนการคัดกรองแอททริบิวท์ เป็นขั้นตอนที่นำเอาข้อมูลแอททริบิวท์ที่หามาได้ มาใช้พิจารณาเพื่อหาแอททริบิวท์ที่มีความจำเป็นที่สุดต่อการจำแนกข้อมูล เนื่องจากขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูลได้มีการกำหนดค่าของจำนวนกฎความสัมพันธ์ ค่าสนับสนุนขั้นต่ำ และค่าความเชื่อมั่นขั้นต่ำ เพื่อให้ได้แอททริบิวท์มากที่สุด ทำให้ไม่สามารถมั่นใจได้ว่าทุกแอททริบิวท์ที่เกิดขึ้นจะสำคัญต่อการจำแนกข้อมูลทั้งหมด ดังนั้นจึงต้องนำเอาข้อมูลแอททริบิวท์ที่ได้มาผ่านขั้นตอนการคัดกรองแอททริบิวท์ โดยอาศัยเทคนิคการเรียนรู้ด้วยต้นไม้ตัดสินใจด้วยขั้นตอนวิธี J48 ผลการทดลองที่ได้แบ่งเป็น 2 ส่วน คือ การเปรียบเทียบจำนวนแอททริบิวท์ที่เหลือหลังจากผ่านขั้นตอนวิธี AssoTree และการเปรียบเทียบประสิทธิภาพของขั้นตอนวิธี AssoTree กับขั้นตอนวิธี J48 และขั้นตอนวิธี CBA ตามตารางที่ 23 และตารางที่ 24 ดังนี้

ตารางที่ 23 เปรียบเทียบจำนวนแอททริบิวท์ที่ได้หลังจากผ่านขั้นตอนวิธี AssoTree

ชุดข้อมูล	จำนวนแอททริบิวท์ดั้งเดิม	จำนวนแอททริบิวท์ที่ได้ในขั้นตอนวิธี AssoTree		ประสิทธิภาพของ AssoTree (Supervised)	ประสิทธิภาพของ AssoTree (Unsupervised)
		Supervised	Unsupervised		
Cleve	14	8	8	83.50	78.88
Crx	16	11	10	86.09	77.25
Hepatitis	20	5	9	81.94	84.52
Horse	24	4	12	66.30	84.51

จากตารางที่ 23 พบว่าประสิทธิภาพที่ได้ของขั้นตอนวิธี AssoTree จะสัมพันธ์กับแอททริบิวท์ ดูได้จากจำนวนแอททริบิวท์ที่ได้ภายหลังที่ผ่านขั้นตอนวิธี AssoTree เมื่อเทียบกับจำนวนแอททริบิวท์ดั้งเดิมทั้งหมดที่ได้มา แสดงให้เห็นว่าเมื่อเปรียบเทียบจำนวนแอททริบิวท์ระหว่างเทคนิคการลดรูปข้อมูลแบบ Supervised และแบบ Unsupervised ถ้าจำนวนแอททริบิวท์ที่ได้มีจำนวนน้อย จะส่งผลต่อประสิทธิภาพ ทำให้ได้ประสิทธิภาพต่ำ เพราะว่าถ้ามีแอททริบิวท์น้อย

ก็จะไม่เพียงพอต่อการเรียนรู้ในการจำแนกข้อมูล แต่ในชุดข้อมูล Cleve จำนวนแอททริบิวท์ที่ได้มีจำนวนเท่ากัน ไม่สามารถหมายความได้ว่า จำนวนแอททริบิวท์เท่ากันแล้ว ประสิทธิภาพที่ได้จะเท่ากัน เพราะว่าค่าของแอททริบิวท์ที่ได้ระหว่างเทคนิคการลดรูปข้อมูลแบบ Supervised และแบบ Unsupervised ไม่เหมือนกัน เนื่องมาจากใช้วิธีการลดรูปข้อมูลที่ไม่เหมือนกัน ทำให้ค่าของแอททริบิวท์ที่ได้ต่างกัน และอีกสาเหตุหนึ่ง คือ ลำดับการเกิดของแอททริบิวท์ และการพิจารณาตัดแอททริบิวท์ในขั้นตอนการคัดกรองแอททริบิวท์ โดยแอททริบิวท์ที่มีความสำคัญต่อการจำแนกข้อมูล อาจจะถูกพิจารณาตัดออกไปได้ ซึ่งจะทำให้ได้ประสิทธิภาพต่ำเช่นกัน

ตารางที่ 24 เปรียบเทียบภาพรวมของประสิทธิภาพระหว่างขั้นตอนวิธี J48, CBA และ AssoTree

ชุดข้อมูล	เทคนิคการแปลงข้อมูล	จำนวนแอททริบิวท์ดั้งเดิม	จำนวนแอททริบิวท์ที่ได้ในขั้นตอนวิธี AssoTree	J48	CBA	AssoTree
Cleve	Supervised	14	8	77.56	82.80	83.50
	Unsupervised		8			78.88
Crx	Supervised	16	11	85.94	84.70	86.09
	Unsupervised		10			77.25
Hepatitis	Supervised	20	5	80.65	81.80	81.94
	Unsupervised		9			84.52
Horse	Supervised	24	4	70.92	82.10	66.30
	Unsupervised		12			84.51

จากตารางที่ 24 พบว่าชุดข้อมูล Hepatitis ที่อาศัยเทคนิคการลดรูปข้อมูลทั้งแบบ Supervised และแบบ Unsupervised ให้ประสิทธิภาพที่สูงกว่าทั้งขั้นตอนวิธี J48 และขั้นตอนวิธี CBA แม้ว่าจำนวนแอททริบิวท์ที่ได้จะมีจำนวนน้อยที่มีจำนวนแอททริบิวท์เพียง 5 แอททริบิวท์ในเทคนิคการลดรูปข้อมูลแบบ Supervised และมีจำนวน 9 แอททริบิวท์ในเทคนิคการลดรูปข้อมูลแบบ Unsupervised เมื่อเทียบกับจำนวนแอททริบิวท์ดั้งเดิมที่มีทั้งสิ้น 20 แอททริบิวท์ แสดงให้เห็นว่าแอททริบิวท์ที่ได้จากขั้นตอนวิธี AssoTree เป็นแอททริบิวท์ที่มีความสำคัญที่สุดต่อการจำแนกข้อมูล และในชุดข้อมูล Crx ที่อาศัยเทคนิคการลดรูปข้อมูลแบบ Unsupervised กับชุดข้อมูล Horse ที่อาศัยเทคนิคการลดรูปข้อมูลแบบ Supervised ให้ประสิทธิภาพที่ต่ำกว่าทั้งขั้นตอนวิธี J48 และ

ขั้นตอนวิธี CBA เนื่องมาจากแอททริบิวท์ที่ได้มีจำนวนน้อยในชุดข้อมูล Horse ที่มีเพียง 4 แอททริ-บิวท์ ทำให้ไม่เพียงพอต่อการจำแนกข้อมูล และในชุดข้อมูล Crx จำนวนแอททริบิวท์ระหว่างเทคนิคการลดรูปข้อมูลแบบ Supervised และแบบ Unsupervised จะมีจำนวนใกล้เคียงกันคือ 11 แอททริบิวท์ และ 10 แอททริบิวท์ ตามลำดับ แต่ประสิทธิภาพที่ได้แตกต่างกันมากมีค่าเท่ากับ 86.09% และ 77.25% ตามลำดับ เพราะว่าเมื่อพิจารณาถึงข้อมูลแอททริบิวท์ และจากการลำดับของการเกิดแอททริบิวท์ ตามตารางที่ 14 พบว่าแอททริบิวท์ที่ได้ต่างกันคือ เทคนิคการลดรูปข้อมูลแบบ Supervised ได้แอททริบิวท์จำนวน 14 แอททริบิวท์ แต่แบบ Unsupervised ได้จำนวนแอททริบิวท์เพียง 11 แอททริบิวท์ ซึ่งแอททริ-บิวที่ไม่ปรากฏในแบบ Unsupervised แต่ปรากฏในแบบ Supervised คือ A7, A3 และ A8 อาจจะเป็นแอททริบิวท์ที่สำคัญต่อการจำแนกข้อมูลเช่นกัน รวมถึงการพิจารณาการตัดแอททริบิวท์ภายในขั้นตอนวิธี AssoTree จึงส่งผลต่อประสิทธิภาพทำให้ได้ประสิทธิภาพที่ต่ำ

การพิจารณาลำดับการเกิดก่อนและหลังของแอททริบิวท์ โดยทำการทดลองและพิจารณาลำดับของแอททริบิวท์ในขั้นตอนการคัดกรองแอททริบิวท์ ดังนี้ ทุกครั้งที่มีการตัดแอททริบิวท์ออกในแต่ละรอบของการทดลอง จะนำเอาชุดข้อมูลที่ประกอบด้วยแอททริบิวท์ที่เหลือนั้น ไปผ่านขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล เพื่อพิจารณาลำดับการเกิดก่อนและหลังของแอททริ-บิวท์ว่ามีการเปลี่ยนแปลงจากผลลัพธ์ในขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูลที่ได้ครั้งแรกหรือไม่ ซึ่งผลการทดลองแสดงให้เห็นว่า ลำดับของแอททริบิวท์ที่ได้ไม่เปลี่ยนแปลง แสดงให้เห็นว่าการตัดแอททริบิวท์ออกไป จะไม่ส่งผลต่อลำดับการเกิดก่อนและหลังของแอททริบิวท์จากขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล

สรุปและข้อเสนอแนะ

สรุป

วัตถุประสงค์ของงานวิจัยนี้คือ ศึกษาขั้นตอนวิธีการคัดเลือกลักษณะเด่นของข้อมูลฝึก และความสำคัญของแอททริบิวท์ที่มีผลต่อการเรียนรู้ของต้นไม้ตัดสินใจ เพื่อเพิ่มประสิทธิภาพความสามารถการจำแนกข้อมูล จึงได้นำเสนอขั้นตอนวิธี AssoTree โดยนำเทคนิคการหากฎความสัมพันธ์อาศัยขั้นตอนวิธีเอพริออริมาใช้ร่วมกับเทคนิคการจำแนกข้อมูลโดยอาศัยเรียนรู้ของต้นไม้ตัดสินใจด้วยขั้นตอนวิธี J48 ในงานวิจัยได้ทำการทดลองกับข้อมูล 4 ชุดข้อมูล คือ ชุดข้อมูล Cleve, Crx, Hepatitis และ Horse ซึ่งได้มาจากฐานข้อมูล UCI Machine Learning Repository โดยมีขั้นตอนการทำงาน 3 ขั้นตอน ได้แก่ ขั้นตอนการเตรียมข้อมูล ขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล และขั้นตอนการคัดกรองแอททริบิวท์

ผลการทดลองได้เปรียบเทียบประสิทธิภาพขั้นตอนวิธี AssoTree กับขั้นตอนวิธี J48 และขั้นตอนวิธี CBA พบว่าชุดข้อมูล Hepatitis ที่ใช้ทั้งเทคนิคการลดรูปข้อมูลแบบ Supervised และแบบ Unsupervised ให้ประสิทธิภาพที่สูงกว่าขั้นตอนวิธีทั้งสองประเภทที่นำมาเปรียบเทียบมีค่าเท่ากับ 81.94% และ 84.52% ตามลำดับ โดยแอททริบิวท์ที่ใช้ในจำแนกข้อมูลจะมีจำนวนน้อยเพียง 5 แอททริบิวท์สำหรับแบบ Supervised และจำนวน 9 แอททริบิวท์สำหรับแบบ Unsupervised เมื่อเทียบกับจำนวนแอททริบิวท์ดั้งเดิมที่มีทั้งสิ้น 20 แอททริบิวท์ และในชุดข้อมูล Cleve, Crx และ Horse ที่ใช้เทคนิคการลดรูปข้อมูลแบบ Supervised หรือแบบ Unsupervised

ภาพรวมแสดงให้เห็นว่าประสิทธิภาพขั้นตอนวิธี AssoTree ให้ประสิทธิภาพที่สูงกว่าขั้นตอนวิธี J48 และขั้นตอนวิธี CBA แต่ในชุดข้อมูล Crx ที่ใช้เทคนิคการลดรูปข้อมูลแบบ Unsupervised และชุดข้อมูล Horse ที่ใช้เทคนิคการลดรูปข้อมูลแบบ Supervised ประสิทธิภาพที่ได้จากขั้นตอนวิธี AssoTree จะให้ประสิทธิภาพที่ต่ำกว่าทั้งขั้นตอนวิธี J48 และขั้นตอนวิธี CBA มีค่าเท่ากับ 77.25% และ 66.30% ตามลำดับ เนื่องจากแอททริบิวท์ที่ได้ของชุดข้อมูล Horse มีจำนวนเพียง 4 แอททริบิวท์ เมื่อเทียบกับจำนวนแอททริบิวท์ดั้งเดิมที่ได้มาที่มีถึง 20 แอททริบิวท์ จึงทำให้ไม่เพียงพอต่อการจำแนกข้อมูล และในชุดข้อมูล Crx แอททริบิวท์ที่ปรากฏในขั้นตอนของการคัดเลือกลักษณะเด่นของข้อมูลของเทคนิคแบบ Supervised และแบบ Unsupervised ได้แอททริบิวท์ที่แตกต่างกัน โดยมีแอททริบิวท์ 3 แอททริบิวท์ที่ปรากฏในแบบ Supervised แต่ไม่ปรากฏในแบบ

Unsupervised คือ A7, A3 และ A8 ซึ่งทำให้มีผลต่อประสิทธิภาพที่ทำให้มีค่าต่ำกว่าขั้นตอนวิธี J48 และขั้นตอนวิธี CBA

ดังนั้นแสดงให้เห็นว่าการคัดกรองแอททริบิวท์ยังมีผลต่อประสิทธิภาพการจำแนกข้อมูล โดยแอททริบิวท์ที่มีทั้งหมดจากชุดข้อมูลไม่จำเป็นว่าทุกแอททริบิวท์จะมีความสำคัญต่อการจำแนกข้อมูล มีเพียงบางแอททริบิวท์เท่านั้นก็เพียงพอต่อการจำแนกข้อมูล และส่งผลต่อการเรียนรู้ด้วยต้นไม้ตัดสินใจโดยทำให้ได้ต้นไม้ที่มีขนาดเล็กลง ซึ่งต้นไม้ขนาดเล็กจะไม่ซับซ้อน และให้ประสิทธิภาพในการทำนายที่ดีกว่าต้นไม้ขนาดใหญ่

จุดเด่นของงานวิจัยนี้ คือ การหาแอททริบิวท์ที่มีความสัมพันธ์กับชุดข้อมูล และนำเอาแอททริบิวที่ได้ไปพิจารณาเพื่อให้ได้แอททริบิวท์ที่มีความสำคัญต่อการจำแนกข้อมูล

จุดด้อยของงานวิจัยนี้ คือ ใช้กับข้อมูลที่มีจำนวนคลาส 2 คลาส และข้อจำกัดที่เลือกใช้ข้อมูลที่มีจำนวนแอททริบิวท์มาก เพื่อให้เพียงพอต่อการเรียนรู้ รวมถึงเรื่องเวลาที่ใช้ในการทดลอง โดยระยะเวลาที่ใช้ในการทดลองจะนานขึ้น แต่ถ้าเทียบกับผลที่ได้รับที่สามารถให้ประสิทธิภาพที่ดีกว่าก็จะคุ้มค่าต่อเวลาที่เพิ่มขึ้น รวมถึงในปัจุบันเทคโนโลยีด้านคอมพิวเตอร์ได้มีการพัฒนาไปมาก ทำให้ประสิทธิภาพการทำงานของเครื่องคอมพิวเตอร์สามารถรองรับการทดลองที่ซับซ้อนได้ ดังนั้นจึงไม่ส่งผลในเรื่องของเวลามากนัก

ข้อเสนอแนะ

สำหรับแนวทางการพัฒนาต่อไป ควรทดสอบกับชุดข้อมูลที่มีจำนวนคลาสมากกว่า 2 คลาสขึ้นไป เพื่อสามารถนำไปประยุกต์ใช้กับข้อมูลที่มีความหลากหลายมากยิ่งขึ้น นอกจากนี้ ควรมีการประยุกต์โดยการนำเทคนิคการหากฎความสัมพันธ์ไปทำงานร่วมกับเทคนิคการจำแนกข้อมูลอื่นๆ เพื่อเปรียบเทียบประสิทธิภาพของการจำแนกประเภทข้อมูลให้ชัดเจนขึ้น

เอกสารและสิ่งอ้างอิง

วัฒนา สุนทรธัย. **สถิติพื้นฐาน.** แหล่งที่มา: http://tulip.bu.ac.th/~wathna.s/fundstat.htm, 7 เมษายน 2553.

วีระพล หาญโชติช่วง, ธนาวินท์ รักธรรมานนท์ และ กฤษณะ ไวยมัย. 2548. การเพิ่มประสิทธิภาพสำหรับเทคนิคการจำแนกข้อมูลโดยใช้กฎความสัมพันธ์, *ใน* **การประชุมทางวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศครั้งที่ 1**.

วิเชียร เปรมชัยสวัสดิ์. 2546. **ระบบฐานข้อมูล.** พิมพ์ครั้งที่ 1 สมาคมส่งเสริมเทคโนโลยี (ไทย-ญี่ปุ่น), กรุงเทพฯ.

Aha, David, Patrick Murphy, Christopher Merz, Eamonn Keogh, Cathy Blake and Seth Hettich. 1987. **UCI Repository of Machine Learning Database.** Available Source: http://archive.ics.uci.edu/ml/about.html, November 12, 2006.

Wasilewska, Anita. **Preprocessing Lecture Notes (Chapter 3).** Available Source: http://www.cs.sunysb.edu/~cse634/lecture_notes/07preprocessing.pdf, December 25, 2552.

Agrawal, Rakesh and Ramakrishnan Srikant. 1994. Fast Algorithm for Mining Association Rules, pp. 487 - 499. *In* **Proceedings of 20th International Conference on Very Large Data Bases (VLDB'94) 20.** Chile.

Dougherty, James, Ron Kohavi and Mehran Sahami. 1995. Supervised and Unsupervised Discretization of Continuous Features, *In* **Proceedings of the Twelfth International Conference on Armand Prieditis & Stuart Russell Machine Learning.** Morgan Kaufmann Publishers, San Francisco, CA.

Li., Wenmin, Jiawei Han and Jian Pei. 2001. CMAR: Accurate and Efficient Classification base on Multiple Class Association Rules, pp. 369-376. *In* **ICDM 01.**

Liu., Bing, Wynne Hsu and Yiming Ma. 1998. Integrating Classification and Asoociation Rule Mining, pp. 80-86. *In* **International Conference on knowledge Discovery and Data Mining (Kdd-98).**

Ross, Quinlan J.. 1993a. **C4.5: Programs for Machine Learning.**

_____. 1993b. Induction of Decision Tree Machine Learning 1(1): 81-106.

Fayyad, Usama M. and Irani Keki B. 1993. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning, pp. 1022-1027. *In* **Proceedings of the 13th Int. Joint Conference on Artificial Intelligence. ed.**Morgan Kaufmann.

ภาคผนวก

ภาคผนวก ก

ผลการทดลองเพิ่มเติม

ตารางผนวกที่ ก1 เปรียบเทียบช่วงการลดรูปข้อมูลสำหรับข้อมูลดั้งเดิม

ชุดข้อมูล		Cleve	Crx	Hepatitis	Horse
จำนวนแอททริบิวท์ที่ทำการลดรูป		6	6	6	7
1st attr.	Distinct	41	350	49	41
	Unique	4	170	15	9
	# range	10	14	9	17
2nd attr.	Distinct	49	215	35	55
	Unique	16	106	16	15
	# range	12	14	10	12
3rt attr.	Distinct	152	132	84	41
	Unique	62	56	64	11
	# range	16	22	9	11
4th attr.	Distinct	91	23	85	25
	Unique	28	5	49	11
	# range	11	35	11	18
5th attr.	Distinct	40	171	30	55
	Unique	10	104	7	14
	# range	10	29	11	14
6th attr.	Distinct	5	240	45	85
	Unique	0	186	23	25
	# range	6	49	9	7
7th attr.	Distinct				45
	Unique				19
	# range				18

ตารางผนวกที่ ก2 เปรียบเทียบช่วงการลดรูปข้อมูลสำหรับข้อมูลที่ลดรูปแบบ Supervised

ชุดข้อมูล		Cleve	Crx	Hepatitis	Horse
1st attr.	Distinct	2	2	1	1
	Unique	0	0	0	0
	# range	2	2	1	1
2nd attr.	Distinct	1	2	2	1
	Unique	0	0	0	0
	# range	1	2	2	1
3rt attr.	Distinct	1	2	1	1
	Unique	0	0	0	0
	# range	1	2	1	1
4th attr.	Distinct	2	3	1	1
	Unique	0	0	0	0
	# range	2	3	1	1
5th attr.	Distinct	2	2	3	1
	Unique	0	0	0	0
	# range	2	2	3	1
6th attr.	Distinct	2	2	2	2
	Unique	0	0	0	0
	# range	2	2	2	2
7th attr.	Distinct				2
	Unique				0
	# range				2

ตารางผนวกที่ ก3 เปรียบเทียบช่วงการลดรูปข้อมูลสำหรับข้อมูลที่ลดรูปแบบ Unsupervised

ชุดข้อมูล		Cleve	Crx	Hepatitis	Horse
1st attr.	Distinct	10	10	10	10
	Unique	1	0	1	1
	# range	10	10	10	10
2nd attr.	Distinct	10	10	7	10
	Unique	0	0	0	1
	# range	10	10	10	10
3rt attr.	Distinct	8	9	10	10
	Unique	1	1	0	0
	# range	10	10	10	10
4th attr.	Distinct	10	6	8	10
	Unique	1	3	3	0
	# range	10	10	10	10
5th attr.	Distinct	9	7	9	9
	Unique	1	2	2	0
	# range	10	10	10	10
6th attr.	Distinct	5	7	9	10
	Unique	0	5	1	3
	# range	10	10	10	10
7th attr.	Distinct				9
	Unique				0
	# range				10

ภาคผนวก ข

ผลงานตีพิมพ์

ขั้นตอนวิธีการเพิ่มประสิทธิภาพการเรียนรู้ของต้นไม้ตัดสินใจโดยใช้กฎความสัมพันธ์

Efficiency Improvement of Decision Tree Learning Module using Association Rule Technique

(NCIT2008) วันที่ 6 - 7 พฤศจิกายน 2551

ขั้นตอนวิธีการเพิ่มประสิทธิภาพการเรียนรู้ของต้นไม้ตัดสินใจโดยใช้กฎความสัมพันธ์

EFFICIENCY IMPROVEMENT OF DECISION TREE LEARNING MODULE USING ASSOCIATION RULE TECHNIQUE

ลลิตา ศรีชัยญา และ นวลวรรณ สุนทรภิษัช
ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์
50 ถนนพหลโยธิน เขตจตุจักร กรุงเทพฯ 10900
อีเมล์: g4964160@ku.ac.th, fscinws@ku.ac.th

บทคัดย่อ

งานวิจัยนี้เป็นการนำเสนอ วิธีการคัดเลือกลักษณะเด่นของข้อมูล และทำการคัดกรองข้อมูลให้ได้คุณลักษณะของข้อมูลฝึกที่เหมาะสมที่สุดสำหรับการเรียนรู้ โดยใช้ต้นไม้ตัดสินใจ เพื่อให้ได้ต้นไม้ตัดสินใจ ที่มีประสิทธิภาพในการจำแนกข้อมูล โดยการนำเอาเทคนิคการหากฎความสัมพันธ์ (Association Rule) เข้ามาใช้ร่วมกับต้นไม้ตัดสินใจ

ผู้วิจัยได้ทำการทดลอง โดยใช้ชุดข้อมูลมาตรฐานจาก UCI Machine Learning Repository จำนวน 16 ชุดข้อมูล โดยทำการทดลองวัดประสิทธิภาพเปรียบเทียบกับอัลกอริทึม C4.5 ซึ่งผลการทดลองพบว่าอัลกอริทึมที่ผู้วิจัยนำเสนอมีประสิทธิภาพดีกว่า C4.5 ใน 13 ชุดข้อมูล

คำสำคัญ – การทำเหมืองข้อมูล การคัดเลือกลักษณะเด่น การหากฎความสัมพันธ์ การจำแนกประเภทข้อมูล

Abstract

This research proposes a feature selection algorithm that can filter out useless attributes in order to get a set of information attributes. The decision tree learning algorithm is used as a main mechanism to construct a tree for the classification problem. We apply an association rule technique to select the most suitable set of feature.

The experiments are done on a set of benchmark dataset collected from UCI Machine Learning Repository. We compare the performance between C4.5 and our proposed algorithm on 16 datasets. The experimental results show that our proposed algorithm outperforms the traditional decision tree algorithm in 13 datasets.

Key words — data mining, feature selection, association rules, classification

1. บทนำ

ปัจจุบันนี้ข้อมูลที่เก็บในฐานข้อมูลมีจำนวนมาก โดยข้อมูลเหล่านั้นจะประกอบด้วยข้อมูลต่างๆ ซึ่งข้อมูลเหล่านี้จะเป็นประโยชน์ก็ต่อเมื่อเรานำข้อมูลมาผ่านการประมวลผล และจัดการให้มีความถูกต้องทันสมัย และสามารถนำมาใช้งานได้ตามที่ต้องการ ข้อมูลนี้เรียกว่า ข้อมูลสารสนเทศ ซึ่งการจะได้มาของข้อมูลสารสนเทศนี้จะต้องอาศัยกระบวนการค้นหาความรู้ในฐานข้อมูล (Knowledge Discovery in Database: KDD) โดยกระบวนการการค้นหาองค์ความรู้นี้ จะนำเอาเทคนิคที่เรียกว่า การทำเหมืองข้อมูล (Data Mining) มาประยุกต์ใช้

การทำเหมืองข้อมูล เป็นกระบวนการที่สำคัญในการค้นหารูปแบบ แนวทาง และความสัมพันธ์ของข้อมูลที่ซ่อนอยู่ เทคนิคที่นำมาใช้ในการการทำเหมืองข้อมูลส่วนใหญ่แบ่งเป็น 4 ประเภท ได้แก่ เทคนิคการจำแนกประเภทข้อมูล (Classification) เทคนิคการหากฎความสัมพันธ์ (Association Rule) เทคนิคการแบ่งกลุ่มข้อมูล (Clustering) และเทคนิคจิตทัศน์ (Visualization)

เทคนิคการจำแนกประเภทข้อมูล (Classification) เป็นเทคนิคหนึ่งที่ได้รับความสนใจอย่างมาก เนื่องจากเป็นเทคนิคในการจำแนกกลุ่มข้อมูลด้วยคุณลักษณะต่างๆ เทคนิคที่ใช้ในการจำแนกประเภทข้อมูลมีมากมาย แต่มีเทคนิคหนึ่งเป็นที่นิยม และคุ้นเคยกันมาก คือ เทคนิคการใช้ต้นไม้ตัดสินใจ (Decision Tree) โดยขั้นตอนการทำงานของเทคนิคการใช้ต้นไม้ตัดสินใจเกี่ยวข้องกับการเลือกแอททริบิวที่เหมาะสมมาสร้างเป็นโหนดในต้นไม้ โดยแอททริบิวที่จะนำมาสร้างเป็นโหนด ต้องสามารถแยกแยะข้อมูลออกเป็นประเภทที่มีการปะปนกันของข้อมูลน้อยที่สุด ซึ่งมีผู้วิจัยหลายคนทำการศึกษาหลักเกณฑ์ หรือสูตรที่ใช้สำหรับคำนวณเพื่อเปรียบเทียบว่า แอททริบิวใดควรถูกเลือกมาสร้างเป็นโหนดในต้นไม้มากที่สุด

งานวิจัยนี้มีวัตถุประสงค์ เพื่อศึกษาขั้นตอนวิธีการสกัดลักษณะเด่นของข้อมูลฝึกที่สามารถเพิ่มประสิทธิภาพการเรียนรู้ของต้นไม้ตัดสินใจได้ โดยได้นำเสนอวิธีการคัดเลือกลักษณะเด่นของข้อมูล และวิธีการคัดกรองข้อมูลให้ได้ข้อมูลที่สำคัญที่สุด เพื่อนำเอามาสร้างเป็นโหนดในต้นไม้ ซึ่งนำเอาเทคนิคการหากฎความสัมพันธ์ (Association Rule) เข้ามาใช้ร่วมกับต้นไม้ตัดสินใจ

การทดลองได้นำข้อมูลมาจากฐานข้อมูล UCI Machine Learning Repository ซึ่งเป็นฐานข้อมูลมาตรฐานที่มีผู้วิจัยหลายคนนำไปใช้ในการทดลอง ซึ่งผลการทดลองได้เปรียบเทียบกับอัลกอริทึมที่ใช้สำหรับการสร้างต้นไม้ตัดสินใจที่นิยมที่สุด คือ C4.5 [6] พบว่าวิธีการที่ผู้วิจัยนำเสนอให้

ประสิทธิภาพในการจำแนกประเภทข้อมูลที่ดีกว่า

ในหัวข้อต่อไปคือหัวข้อที่ 2 จะกล่าวถึงทฤษฎีที่เกี่ยวข้องกับการทำงานวิจัย หัวข้อที่ 3 กล่าวถึงวิธีการในการทดลอง โดยผลของทดลองจะอยู่ในหัวข้อที่ 4 และหัวข้อสุดท้ายเป็นการสรุปผลการทดลอง

3. ทฤษฎีที่เกี่ยวข้อง

3.1 การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล หรืออาจเรียกว่าการค้นหาความรู้ในฐานข้อมูล (Knowledge Discovery in Database: KDD) เป็นกระบวนการในการค้นพบความรู้แฝงของข้อมูลที่อยู่ฐานข้อมูลขนาดใหญ่ ซึ่งมีขั้นตอนการทำเหมืองข้อมูลเป็นกระบวนการที่สำคัญในการค้นหารูปแบบ แนวทาง และความสัมพันธ์ของข้อมูลที่ซ่อนอยู่ โดยใช้ขั้นตอนวิธีจากวิชาสถิติ การเรียนรู้ของเครื่อง การรู้จำรูปแบบ และหลักคณิตศาสตร์

ขั้นตอนการทำเหมืองข้อมูล ประกอบด้วย 4 ขั้นตอนหลัก ดังนี้

1. ขั้นตอนการเตรียมข้อมูล (Data Preparation) เป็นขั้นตอนในการจัดการข้อมูลให้สามารถนำเข้าสู่อัลกอลิทึมของการทำเหมืองข้อมูลได้ เช่น การทำ Data Cleaning, Data Integration, Data Reduction เป็นต้น ซึ่ง Data Preparation สามารถแบ่งได้ออกเป็น 3 ส่วน ได้แก่ Data Selection, Data Preprocessing และ Data Transformation
2. ขั้นตอนการทำเหมืองข้อมูล เป็นขั้นตอนที่มีการดำเนินการหลายแบบ เช่น Database Segmentation, Predictive Modeling, Link Analysis เป็นต้น ในแต่ละการดำเนินการจะมีอัลกอริทึมให้เลือกใช้ เช่น การทำ Database Segmentation อาจใช้ K-Mean Algorithms หรือใช้ Unsupervised Learning Neural Networks ถ้าการทำ Predictive Modeling อาจใช้ CART (Classification and Regression Tree) หรือใช้ Supervised Learning Neural Network เช่น Back propagation Neural Network หรือถ้าการทำ Link Analysis ซึ่งมีการทำอยู่ 2 ลักษณะ คือ Association Rule และ Sequential Pattern Discovery
3. ขั้นตอนการสรุปผลลัพธ์ (Analysis of Results and Knowledge Presentation) เป็นขั้นตอนสำหรับนักวิเคราะห์ข้อมูลที่ต้องเก็บผลลัพธ์ของการทำเหมืองข้อมูล สรุปความหมายของผลลัพธ์ที่ได้ ซึ่งจะเป็นข้อมูลความรู้ที่นำไปเป็นข้อมูลสารสนเทศที่ช่วยในการตัดสินใจ

ความรู้ที่ได้จากการทำเหมืองข้อมูลมีหลายรูปแบบ ได้แก่

1. การหากฎความสัมพันธ์ (Association Rule) เป็นความรู้ที่แสดงความสัมพันธ์ของเหตุการณ์หรือวัตถุที่เกิดที่เกิดขึ้นพร้อมกัน ตัวอย่างของการนำการหากฎความสัมพันธ์

ไปประยุกต์ใช้ เช่น การวิเคราะห์ข้อมูลการขายสินค้า โดยการเก็บข้อมูลจากระบบ ณ จุดขาย (POS) หรือร้านค้าออนไลน์ แล้วจะพิจารณาสินค้าที่ผู้ซื้อมักจะซื้อพร้อมกัน จากนั้นจะนำผลที่ได้ไปใช้ประโยชน์สำหรับกลยุทธ์ทางธุรกิจ เช่นร้านค้าอาจจะจัดร้านให้สินค้าสองอย่างอยู่ใกล้กัน เพื่อเพิ่มยอดขาย

2. การจำแนกประเภทข้อมูล (Data Classification) เป็นความรู้ที่แสดงในรูปของกฎ เพื่อระบุประเภทของวัตถุจากคุณสมบัติของวัตถุ เทคนิคประเภทนี้เหมาะกับการสร้างแบบจำลองเพื่อการพยากรณ์ค่าข้อมูล (Predictive Modeling) ในอนาคต ซึ่งลักษณะเช่นนี้เรียกว่า “Supervised Learning” เช่น การหาความสัมพันธ์ระหว่างผลการตรวจร่างกายต่างๆ กับการเกิดโรค โดยใช้ข้อมูลผู้ป่วยและการวินิจฉัยของแพทย์ที่เก็บไว้ เพื่อนำมาช่วยในการวินิจฉัยโรคของผู้ป่าว หรือในทางธุรกิจใช้เพื่อดูคุณสมบัติของผู้ก่อหนี้ดีหรือหนี้เสีย เพื่อใช้ประกอบในการพิจารณาตัดสินใจการอนุมัติเงินกู้

3. การแบ่งกลุ่มข้อมูล (Data Clustering) เป็นความรู้ที่แสดงการจำแนกกลุ่มข้อมูลใหม่ที่มีลักษณะคล้ายกันไว้ในกลุ่มเดียวกัน ลักษณะเช่นนี้เรียกว่า “Unsupervised Learning” เช่น การแบ่งกลุ่มผู้ป่วยที่เป็นโรคเดียวกันตามลักษณะอาการ เพื่อนำไปใช้ประโยชน์ในการวิเคราะห์หาสาเหตุของโรค โดยจะพิจารณาจากผู้ป่วยที่มีอาการคล้ายคลึงกัน

4. จินตทัศน์ (Visualization) เป็นความรู้ที่แสดงผลนำเสนอข้อมูลในรูปแบบกราฟิกที่จะทำให้ผู้ใช้หรือผู้ตัดสินใจสามารถค้นพบองค์ความรู้ได้จากการแสดงผล แทนการใช้ข้อความในการนำเสนอข้อมูล

3.2 การคัดเลือกลักษณะเด่นของข้อมูล (Feature Selection)

การคัดเลือกลักษณะเด่นของข้อมูล (Feature Selection) อาจรู้จักในอีกความหมายคือ การคัดเลือกตัวแปร (Variable Selection) การคัดเลือกแอททริบิว (Attribute Selection) การลดลักษณะเด่นของข้อมูล (Feature Reduction) เป็นเทคนิคที่ใช้ทั่วไปสำหรับการเรียนรู้ของเครื่อง (Machine Learning) โดยจะทำการคัดเลือกลักษณะข้อมูลที่ไม่มีความสัมพันธ์กัน และมีความซ้ำซ้อนกันออจากชุดข้อมูล ได้มีนักวิจัยนำเสนออัลกอริทึมสำหรับการคัดเลือกลักษณะเด่นของข้อมูล เช่นงานวิจัยของ Kittisak Kerdprasop และคณะ [4] วัตถุประสงค์หลักของการลดลักษณะเด่นของข้อมูล คือ การเพิ่มประสิทธิภาพการทำนาย เพื่อการสังเคราะห์โมเดลได้อย่างรวดเร็ว และเพื่อลดความซับซ้อนของรูปแบบ

3.3 การหากฎความสัมพันธ์ (Association Rule)

การหากฎความสัมพันธ์ (Association Rule) เป็นเทคนิคที่อาศัยการเชื่อมโยงข้อมูล หรือกลุ่มข้อมูลที่เกิดขึ้นในเหตุการณ์เดียวกันเข้าด้วยกัน เช่นการใช้ Market – Basket Analysis เพื่อวิเคราะห์ข้อมูลการสั่งซื้อสินค้า

กฎความสัมพันธ์สามารถเขียนความสัมพันธ์ที่เกิดขึ้นระหว่างข้อมูลออกมาได้ในรูป A $\rightarrow$ B โดยที่ A เป็นเหตุการณ์ที่เกิดขึ้นก่อน (Antecedent) หรือ LHS (Left – Hand Side) และเรียก B เป็นผลของเหตุการณ์ (Consequent) หรือ RHS (Right – Hand Side) สามารถแสดงความสัมพันธ์ได้จากรูปแบบกฎความสัมพันธ์ ตาม (1)

$$LHS \rightarrow RHS \qquad (1)$$

โดยที่ LHS และ RHS เป็นชุดของข้อมูลที่เกิดขึ้นบ่อยในฐานข้อมูล

ปัจจัยที่สำคัญที่ใช้ในการบอกคุณภาพของกฎว่าดีหรือไม่ดี ขึ้นอยู่กับ 2 ปัจจัยที่สำคัญ คือ

1. ค่าสนับสนุน (Support) เป็นค่าความถี่ที่ใช้วัดสัดส่วนของข้อมูลที่เกิดขึ้นในทรานแซคชัน เช่น ค่าสนับสนุนของชุดข้อมูล A หมายถึง จำนวนทรานแซคชันที่ประกอบด้วยชุดข้อมูล A เมื่อเทียบกับจำนวนทรานแซคชันทั้งหมดในฐานข้อมูล สามารถแสดงสมการ ดังนี้

$$Support = \frac{\#Transaction(LHS, RHS)}{\#TotalTransactions} \qquad (2)$$

ค่าสนับสนุนที่ดีควรจะมีค่าในระดับที่สูง เพราะแสดงให้เห็นถึงความมีนัยสำคัญของความสัมพันธ์ของข้อมูล และในทางตรงกันข้ามถ้าค่าสนับสนุนมีค่าในระดับต่ำอาจจะแสดงให้เห็นถึงความไม่มีนัยสำคัญของความสัมพันธ์นั้นได้

ค่าสนับสนุนขั้นต่ำ (Minimum Support) เป็นค่าสนับสนุนขั้นต่ำของชุดข้อมูล ซึ่งจะถูกกำหนดโดยผู้ใช้งาน

2. ค่าความเชื่อมั่น (Confidence) คือค่าความถี่ของเหตุการณ์อื่นที่เกิดร่วมกับข้อมูลนั้น หรือเป็นค่าที่บ่งบอกว่ากฎหนึ่งมีความน่าเชื่อถือเพียงใด เช่น ในการหากฎที่มีระดับนัยสำคัญ เมื่อมีเหตุการณ์ A เกิดขึ้นจำนวนหนึ่ง จะมีเหตุการณ์ B เกิดขึ้นด้วย หมายความว่า ต้องหาเงื่อนไขที่จะทำนายเหตุการณ์ B ที่เกิดขึ้นเนื่องจากเหตุการณ์ A เกิดขึ้น สามารถแสดงสมการ ดังนี้

$$Confidence = \frac{\#Transaction(LHS, RHS)}{\#Transaction(LHS)} \qquad (3)$$

ค่าความเชื่อมั่นที่ดีควรจะมีค่าในระดับที่สูง เพราะแสดงให้เห็นถึงความน่าเชื่อมั่นของกฎว่าเหตุการณ์หนึ่ง จะสัมพันธ์กับอีกเหตุการณ์หนึ่งมีความน่าเชื่อมั่นเพียงใด

ค่าความเชื่อมั่นขั้นต่ำ (Minimum Confidence) เป็นค่าความเชื่อมั่นขั้นต่ำของกฎซึ่งจะถูกกำหนดโดยผู้ใช้งาน

3.4 อัลกอริทึมเอพริออริ (Apriori Algorithm)

อัลกอริทึมเอพริออริ เป็นอัลกอริทึมที่ได้รับการนิยมใช้อย่างแพร่หลายสำหรับการหากฎความสัมพันธ์ที่เป็นที่นิยมสำหรับชุดข้อมูลที่เกิดขึ้นบ่อยในฐานข้อมูล (Frequent Itemset) คือ อัลกอริทึมเอพริออริ (Apriori Algorithm) มีขั้นตอนการทำงานดังนี้คือ สร้าง Candidate Itemset ขึ้นมาก่อน โดยเริ่มสร้างจาก Candidate Itemset ที่มีขนาดเท่ากับ k-Itemset โดยที่ k= {1, 2, 3, …, n} ซึ่ง Candidate Itemset ในระดับที่ K จะได้มาจากการ join หรือ union กันระหว่าง Frequent Itemset ในระดับที่ k-1 เช่น Itemset(ABC) = Itemset(AB) ∪ Itemset(AC) จากนั้นทำการหา Frequent Itemset ในระดับที่ k โดยจะทำการอ่านข้อมูลจากฐานข้อมูลเพื่อบันทึกค่าความถี่ของ Candidate Itemset แต่ละตัวในระดับที่ k จากนั้นทำการตรวจสอบว่า Candidate Itemset ใดมีค่าสนับสนุน (Support) มากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำ (Minimum Support) และจะถือว่า Candidate Itemset นั้นมีคุณสมบัติเป็น Frequent Itemset ในระดับที่ k จากนั้นจะทำการสร้าง Candidate Itemset หา Frequent Itemset ในระดับที่ k+1 ต่อไป โดยทำเช่นนี้ไปเรื่อยๆ จนกว่าจะไม่สามารถทำการสร้าง Candidate Itemset ในระดับที่ k+1 ได้ โดยอัลกอริทึมเอพริออริ

3.5 การจำแนกประเภทข้อมูล (Classification)

การจำแนกประเภทข้อมูล เป็นเทคนิคในการจำแนกกลุ่มข้อมูลด้วยคุณลักษณะต่างๆ โดยทำการสำรวจรายการในฐานข้อมูล เพื่อแยกแยะให้อยู่ในหมวดที่เราได้กำหนดไว้ล่วงหน้า เทคนิคที่ใช้ในการจำแนกประเภทข้อมูลมีมากมาย แต่มีเทคนิคหนึ่งเป็นที่นิยมและคุ้นเคยกันมาก คือ เทคนิคการใช้ต้นไม้ตัดสินใจ

เทคนิคการใช้ต้นไม้ตัดสินใจ เป็นเครื่องมือที่ช่วยกำหนดของเขตของปัญหาและช่วยสร้างทางเลือกที่เป็นไปได้ในการแก้ปัญหา

คุณลักษณะของต้นไม้ตัดสินใจ ได้แก่

1. แสดงการเชื่อมโยงความสัมพันธ์ของปัญหาได้ชัดเจน โดยอาศัยแนวทางกราฟิก
2. ช่วยจัดการกับสถานการณ์ที่ซับซ้อน ให้อยู่ในรูปแบบที่กระชับขึ้น เพื่อช่วยให้เห็นภาพของปัญหาได้ชัดเจนขึ้น
3. มีโครงสร้างที่สามารถบอกผลลัพธ์ที่อาจเกิดขึ้นจากการคัดเลือกทางเลือกต่างๆ สำหรับการตัดสินใจ
4. ช่วยวิเคราะห์ลำดับการตัดสินใจแก้ไขปัญหาต่างๆ พร้อมทั้งวิเคราะห์ผลลัพธ์จากการตัดสินใจด้วยแนวทางต่างๆ
5. ช่วยจัดสมดุลด้านความเสี่ยงในการตัดสินใจคัดเลือกแนวทางแก้ไขปัญหาต่างๆ
6. เหมาะกับปัญหาที่มีจำนวนทางเลือกไม่มากนัก เนื่องจากถ้าจำนวน

ทางเลือกในการแก้ปัญหามีมาก อาจจะทำให้โครงสร้างต้นไม้ตัดสินใจดูซับซ้อน ยุ่งเหยิง

อัลกอริทึมที่ใช้ในการสร้างต้นไม้ตัดสินใจมีมากมาย ที่เป็นที่นิยม เช่น C4.5 [6], CART [3], ID3 [7] เป็นต้น ซึ่งแต่ละอัลกอริทึมจะมีลักษณะที่แตกต่างกัน และต้นไม้ที่ได้ก็จะมีลักษณะที่แตกต่าง โดยลักษณะของโครงสร้างต้นไม้ที่มีขนาดเล็กไม่ซับซ้อน ยุ่งเหยิง จะเป็นต้นไม้ที่มีคุณภาพที่ดีที่สุด ที่เหมาะสำหรับนำไปใช้การตัดสินใจ และจากงานวิจัยอัลกอริทึมที่ให้ประสิทธิภาพในการสร้างต้นไม้ตัดสินใจที่ดีที่สุด คือ อัลกอริทึม C4.5 หรือ J48 ในโปรแกรมเวกา

3.6 อัลกอริทึม C4.5

อัลกอริทึม C4.5 เป็นอัลกอริทึมที่ได้รับความนิยม และมีประสิทธิภาพสูงสำหรับการสร้างต้นไม้ตัดสินใจ โดยขั้นตอนวิธีการเรียนรู้ด้วยต้นไม้ตัดสินใจ จะเกี่ยวข้องกับการเลือกแอททริบิวที่เหมาะสมมาสร้างเป็นโหนดในต้นไม้ โดยแอททริบิวนั้น ต้องสามารถแยกแยะข้อมูลออกเป็นประเภทที่ปะปนกันได้น้อยที่สุด

วิธีการเลือกแอททริบิวมาสร้างโหนดในต้นไม้ เป็นวิธีการเรียนรู้ที่อาศัยค่าความไร้ระเบียบของข้อมูล (Entropy) และค่าเกน (Gain) เป็นเครื่องมือในการเลือกแอททริบิวที่มีอำนาจในการจำแนกข้อมูลสูง หลักการคือ การประเมินว่าเมื่อเลือกแอททริบิวนั้นมาสร้างเป็นโหนดในต้นไม้แล้ว จะต้องทำให้ข้อมูลสามารถถูกจำแนกออกไปตามกิ่งของต้นไม้ได้ดีที่สุด คือ ข้อมูลที่ถูกจำแนกมีการปะปนกันของข้อมูลน้อยที่สุด

$$Entropy(S) = \frac{-P}{P+N} log_2 \frac{P}{P+N} - \frac{N}{P+N} log_2 \frac{N}{P+N} \quad (4)$$

โดย S คือ ชุดข้อมูลฝึกก่อนการแยกแยะ

P คือ จำนวนข้อมูลฝึกระบบคลาสบวก

N คือ จำนวนข้อมูลฝึกระบบคลาสลบ

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (5)$$

โดย A คือ ข้อมูลแอททริบิว A

$|S_v|$ คือ จำนวนคลาสที่พบในแอททริบิว A ที่มีข้อมูลเป็น v

$|S|$ คือ จำนวนคลาสทั้งหมดที่พบในชุดข้อมูลฝึกระบบ S

Entropy(S) คือค่าความไร้ระเบียบของชุดข้อมูลฝึกระบบ S

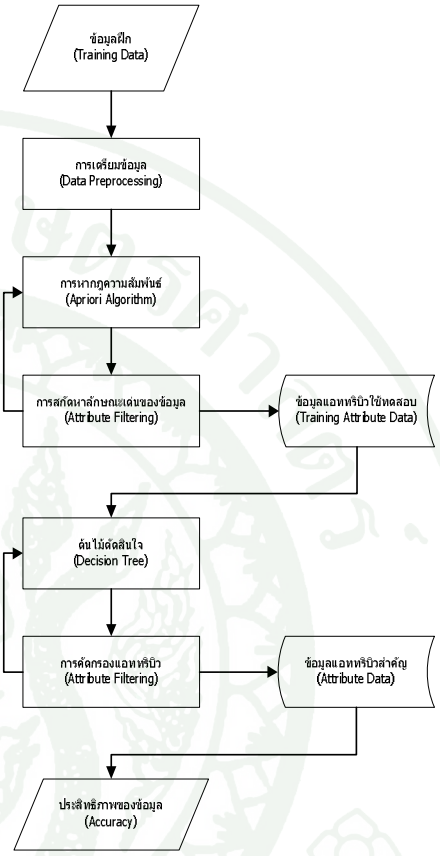
ขั้นตอนการสร้างต้นไม้ตัดสินใจ มีพื้นฐานที่นำใช้ในการสร้างต้นไม้ตัดสินใจ คือ ขั้นตอนวิธีเชิงละโมบ (Greedy Algorithm) โดยการสร้างต้นไม้จะทำจากบนลงล่างแบบวนซ้ำ (Recursive) ด้วยวิธีการแบ่งปัญหาใหญ่เป็นปัญหาย่อย (Divide-and-

Conquer) วิธีการทำงาน คือ ขั้นแรกต้นไม้ทำการสร้างโหนดแรกขึ้นมา จากนั้นทำการทดสอบข้อมูล ถ้าข้อมูลเป็นประเภทเดียวกันให้โหนดนั้นเป็นโหนดใบ ถ้าไม่ใช่จะมีการคำนวณค่าเกนของแต่ละแอททริบิว ขั้นต่อมาขั้นตอนวิธีนี้จะทำการเลือกแอททริบิวที่มีค่าเกนสูงสุด เพื่อใช้แอททริบิวนั้นเป็นโหนดที่แบ่งแยกข้อมูล จากนั้นทำการสร้างกิ่งของต้นไม้และแบ่งข้อมูลออกเป็นส่วน แล้ววนทำงานอย่างนี้ซ้ำไปเรื่อยๆ จนกว่าจะไม่สามารถแยกความแตกต่างของข้อมูลได้อีก

ต้นไม้ตัดสินใจจะแสดงผลลัพธ์ในรูปของโครงสร้างต้นไม้ ซึ่งสามารถแปลงเป็นกฎที่เข้าใจได้ง่าย โดยที่แต่ละโหนดของต้นไม้แสดงถึงแอททริบิว แต่ละกิ่งแสดงถึงเงื่อนไขในการทดสอบ และโหนดปลาย (Leaf Node) แสดงถึงประเภทข้อมูลที่กำหนดไว้

4. วิธีการทดลอง

ส่วนนี้จะกล่าวถึงวิธีการในการดำเนินการวิจัย ซึ่งผู้วิจัยได้นำเสนออัลกอริทึมชื่อ Association Tree โดยขั้นตอนการทำงานประกอบด้วย 3 ขั้นตอน คือ ขั้นตอนการเตรียมข้อมูล (Data Preprocessing) ขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล (Feature Selection) และขั้นตอนการคัดกรองแอททริบิว (Attribute Filtering)



ภาพที่ 1 แสดงภาพรวมวิธีการในการดำเนินการวิจัย

4.1 ขั้นตอนการเตรียมข้อมูล (Data Preprocessing)

ข้อมูลที่ใช้ในการทดลองมีทั้งสิ้น 16 ชุดข้อมูล ซึ่งได้มาจากฐานข้อมูล UCI Machine Learning Repository [5] เนื่องจากลักษณะของข้อมูลที่ทำการทดลองมีความหลากหลาย และแตกต่างกัน บางข้อมูลมีค่าขาดหายไป (Missing Value) บางข้อมูลชนิดของการจัดเก็บแตกต่างกัน มีทั้งข้อมูลชนิดโนมินัล (Nominal Attribute) ข้อมูลชนิด

ลำดับ (Ordinal Attribute) ข้อมูลชนิดต่อเนื่อง (Continuous Attribute) ดังนั้นจึงต้องมีการจัดการกับปัญหาต่างๆ ของข้อมูล โดยพิจารณาตามลักษณะของข้อมูลดังต่อไปนี้

1. กรณีข้อมูลมีแอททริบิวที่ไม่จำเป็นต่อการประมวลผล เช่น ลำดับของข้อมูลรหัสโรงพยาบาล จะทำการตัดแอททริบิวนั้นออกจากฐานข้อมูล
2. กรณีข้อมูลมีค่าบางส่วนขาดหายไป (Missing Value) จะทำการเติมค่าข้อมูลที่เป็นไปได้ ด้วยการคำนวณจากค่าที่มีมากสุด กับค่ากลาง ที่ได้จากข้อมูลที่เหลือในข้อมูลฝึก
3. กรณีข้อมูลเป็นข้อมูลชนิดต่อเนื่อง (Continuous Attribute) จะทำการแปลงข้อมูลเป็นข้อมูลชนิดประเภทข้อมูล (Nominal Attribute) อาศัยเทคนิคการ binning เป็นการปรับข้อมูล โดยดูจากค่าของข้อมูลที่อยู่ใกล้ๆ กัน ซึ่งเหตุผลที่ต้องทำการแปลงข้อมูลชนิดต่อเนื่อง ให้เป็นข้อมูลชนิดประเภทข้อมูล เนื่องจากในขั้นตอนของการคัดเลือกลักษณะเด่นของข้อมูล ที่อาศัยเทคนิคการหากฎความสัมพันธ์นั้น ไม่รองรับข้อมูลที่เป็นข้อมูลชนิดต่อเนื่อง

4.2 ขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล (Feature Selection)

ขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล มีหลักการในการทำงาน คือ จะใช้เทคนิคการหากฎความสัมพันธ์ โดยใช้อัลกอริทึมที่รู้จักเป็นอย่างดี คือ อัลกอริทึมเอพริออริ [1] ซึ่งขั้นแรกของการทำการทดลองจะต้องกำหนดค่าเริ่มต้นจำนวน 3 ค่า คือ จำนวนกฎ ค่าสนับสนุนขั้นต่ำ และค่าความเชื่อมั่นขั้นต่ำ โดยกำหนดค่าสนับสนุนขั้นต่ำเท่ากับ 50% และค่าความเชื่อมั่นเท่ากับ 90% เพราะว่าค่าสนับสนุนและค่าความเชื่อมั่นสูงส่งผลต่อประสิทธิภาพของกฎ ซึ่งจะผลต่อแอททริบิวที่ถูกคัดเลือกมาสร้างเป็นกฎด้วย จากนั้นทำการหากฎความสัมพันธ์กับข้อมูลฝึก ซึ่งจะทำการสังเกตจำนวนกฎที่ได้ ถ้าจำนวนกฎที่ได้น้อยกว่า จำนวนกฎที่กำหนด จะทำการลดค่าสนับสนุนขั้นต่ำลง และถ้าลดค่าสนับสนุนขั้นต่ำลงจนมีค่าเท่ากับ 0.1 แล้วแต่ยังได้จำนวนกฎน้อยกว่าจำนวนกฎที่กำหนด จะทำการลดค่าความเชื่อมั่นลดด้วย เนื่องจากต้องการหาแอททริบิวที่สำคัญต่อการทำนายให้ได้มากที่สุด เพื่อไม่ให้ส่งผลต่อประสิทธิภาพของการทำนาย ซึ่งในทุกรอบของการหากฎความสัมพันธ์ จะทำการเก็บแอททริบิวที่เกิดขึ้นทั้งหมด จากนั้นเมื่อการทำงานสิ้นสุดจะได้แอททริบิวที่สำคัญต่อการทดลองขั้นต่อไป โดยขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล มีอัลกอริทึมของการทำงาน ตามตารางที่ 1

ตารางที่ 1 อัลกอริทึมการคัดเลือกลักษณะเด่นของข้อมูล

อัลกอริทึม การคัดเลือกลักษณะเด่นของข้อมูล

```
Initialize NumRule =10, MinSup = 0.5,
MinConf = 0.9
While (NumRule <= 100)
        SetofRule =
        Apriori(NumRule,MinSup,MinConf)
        N = count(SetofRule)
        If (N = NumRule) then
          NumRule = NumRule + 10
        else if ( N < NumRule) and
                (MinSup > 0.1) then
                MinSup = MinSup – 0,05
                NumRule = 10
        else if(N < NumRule) and (MinSup
                = 0.1) and (MinConf > 0.8)
                then MinConf = MinConf –
                0.05
                NumRule = 10
        else
                Break
        End if
End While
        A_t = Set of Original Attribute
        A_R = ExtractRuleAtr(SetofRule)
        For (i = 1; i < N_At; i++)
                If A_ti is in A_R
                  A_temp = A_temp ∪ A_R
        End For
```

ผลลัพธ์จากการประมวลผลของขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูลข้างต้น จะได้ผลลัพธ์เป็นเซตของแอททริบิวที่ถือว่ามีสำคัญต่อการจำแนกประเภทข้อมูล

4.3 ขั้นตอนการคัดกรองแอททริบิว (Attribute Filtering)

ขั้นตอนการคัดกรองแอททริบิว ภายหลังจากการทำการคัดเลือกลักษณะเด่นของข้อมูลแล้ว จะเข้าสู่ขั้นตอนการคัดกรองแอททริบิว เนื่องจากในขั้นตอนการคัดเลือกได้กำหนดค่าเพื่อหาแอททริบิวที่เด่นที่ถือว่ามีความสำคัญต่อการจำแนกประเภทข้อมูล แต่ก็ไม่สามารถมั่นใจได้ว่า ทุกแอททริบิวจะสำคัญต่อการจำแนกประเภทข้อมูลทั้งหมด ดังนั้นจึงต้องมีการคัดกรองแอททริบิวเพื่อให้ได้แอททริบิวที่จำเป็นที่สุดสำหรับการจำแนกประเภทข้อมูล มีหลักการในการทำงาน คือ จะใช้ต้นไม้ตัดสินใจ โดยใช้อัลกอริทึม C4.5 หรือ J48 ในโปรแกรมเวกา (Weka Program) ซึ่งขั้นแรกจะนำเซทของแอททริบิวที่ได้จากขั้นตอนข้างต้นมาสร้างเป็นต้นไม้ตัดสินใจ แล้วเก็บค่าความถูกต้อง (Accuracy) ที่ได้ไว้ จากนั้นให้ทำการตัดแอททริบิวที่มี โดยมีหลักการในการเลือกตัด คือ ให้ตัดแอททริบิวตัวที่พบเป็นตัวสุดท้ายจากการทดลองในขั้นตอนแรกออก เนื่องจากสมมติฐานที่ว่า แอททริบิวที่พบในภายหลังอาจจะไม่มีความเป็นนัยสำคัญต่อความสัมพันธ์ของกฎ แต่ได้มาเพราะว่ามีการ

กำหนดค่าจึงส่งผลต่อการคัดเลือกแอททริบิวมาสร้างเป็นกฎความสัมพันธ์ จากนั้นนำเซตแอททริบิวที่เหลือไปสร้างต้นไม้สินใจ แล้วทำการเปรียบเทียบค่าความถูกต้องกับที่เก็บไว้ ถ้าค่าความถูกต้องที่ได้มีค่ามากกว่า จะทำบันทึกแอททริบิวที่ตัดออกไปเก็บไว้ แล้วทำการตัดแอททริบิวตัวถัดไป วนทำการทดลองใหม่จนกระทั่งแอททริบิวทุกตัวถูกพิจารณา ท้ายที่สุดจะได้แอททริบิวที่ต้องตัดออกเหลือเฉพาะแอททริบิวที่สำคัญที่สุดต่อการจำแนกประเภทข้อมูล โดยขั้นตอนการคัดกรองแอททริบิว มีอัลกอริทึมของการทำงาน ตามตารางที่ 2

ตารางที่ 2 อัลกอริทึมการคัดกรองแอททริบิว

อัลกอริทึม การคัดกรองแอททริบิว

A_{temp} = Set of Selectiong Attribute
TreePerf1 = DecisionTree(TrainData, A_{temp})
MaxPerf = TreePerf1
For (I = n; I > 0; i- -)
$A_{remove} = A_{temp}$
$A_{temp} = A_{temp} - A_{temp,I}$
TreePerf2 = DecisionTree(TrainData, A_{temp})
If (TreePerf2 >= TreePerf1) and (TreePerf2 >= MaxPerf) then
MaxPerf = TreePerf2
End If
End For

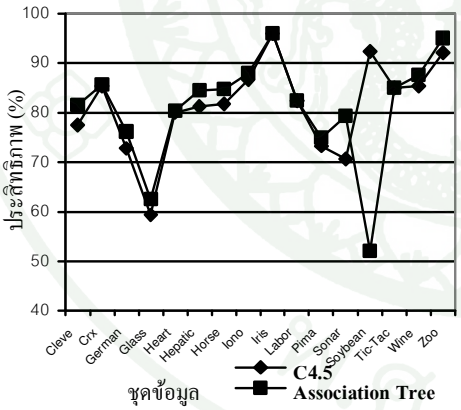
5. ผลการทดลอง

ในหัวข้อนี้จะกล่าวถึง การเปรียบเทียบประสิทธิภาพในการจำแนกประเภทข้อมูล โดยนำผลการทดลองที่ได้จากการทดลองด้วยอัลกอริทึม Association Tree เปรียบเทียบกับเทคนิคต้นไม้ตัดสินใจ ได้ทำการทดลองกับข้อมูลจำนวนทั้งสิ้น 16 ชุดข้อมูล ที่ได้มาจากฐานข้อมูล UCI Machine Learning Repository ในการเปรียบเทียบประสิทธิภาพนี้ ค่าพารามิเตอร์ที่ใช้จะเป็นค่ามาตรฐานที่ได้มีการกำหนดไว้แล้ว ได้ทำการทดลองโดยใช้ค่าสนับสนุน และการทดลองแบบ 10 โฟลดครอสวาดิเดชัน (10 fold cross-validation) ซึ่งผลของการเปรียบเทียบประสิทธิภาพตามตารางที่ 3 และแสดงการเปรียบเทียบในรูปแบบกราฟตามภาพที่ 2 ตามลำดับ

ตาราง 3 ตารางเปรียบเทียบประสิทธิภาพ

ข้อมูล	จ.น.แอททริบิว	จ.น.คลาส	จ.น.ตัวอย่าง	C4.5	Association Tree
Cleve	13	2	303	77.56	81.52
Crx	15	2	690	85.36	85.65
German	20	2	1000	72.80	76.20
Glass	9	7	214	59.35	62.61
Heart	13	2	270	80.00	80.37
Hepatic	19	2	155	81.29	84.52
Horse	22	2	368	81.79	84.78

ข้อมูล	จ.น.แอททริบิว	จ.น.คลาส	จ.น.ตัวอย่าง	C4.5	Association Tree
Iono	34	2	351	86.61	88.03
Iris	4	3	150	96.00	96.00
Labor	16	2	57	82.46	82.46
Pima	8	2	768	73.31	75.00
Sonar	60	2	208	70.67	79.33
Soybean	35	19	368	92.39	52.12
Tic-tac	9	2	958	85.07	85.07
Wine	13	3	178	85.39	87.64
Zoo	16	7	101	92.08	95.05



ภาพที่ 4กราฟเปรียบเทียบประสิทธิภาพในการจำแนกประเภทข้อมูล

6. สรุปผลการทดลอง

ในงานวิจัยนี้แสดงให้เห็นว่า แอททริบิวที่ใช้ในการจำแนกประเภทข้อมูลมีส่วนสำคัญต่อประสิทธิภาพของการทำนายข้อมูล โดยแอททริบิวที่มีทั้งหมดจากชุดข้อมูลไม่จำเป็นว่าทุกแอททริบิวจะมีสำคัญต่อการทำนายผลทั้งหมด มีเพียงแค่บางแอททริบิวเท่านั้นก็เพียงพอต่อการทำนายผล และยังสามารถทำให้ต้นไม้ตัดสินใจที่ได้มีขนาดที่เล็ก ซึ่งต้นไม้ที่มีขนาดเล็ก จะไม่ซับซ้อนและให้ประสิทธิภาพในการทำนายที่ดีกว่าต้นไม้ขนาดใหญ่ จึงได้นำเสนอเทคนิคการคัดเลือกลักษณะเด่นของข้อมูล โดยใช้เทคนิคการหากฎความสัมพันธ์ และเทคนิคการคัดกรองแอททริบิว โดยใช้เทคนิคต้นไม้ตัดสินใจ

จากผลการทดลองพบว่า อัลกอริทึม Association Treeสามารถเพิ่มประสิทธิภาพการจำแนกประเภทข้อมูล ที่สามารถใช้ได้กับข้อมูลที่หลากหลาย และได้ประสิทธิภาพของการจำแนกประเภทข้อมูลที่ดีกว่าการจำแนกประเภทข้อมูลโดยใช้ C4.5 จากตารางที่ 3 แสดงประสิทธิภาพของอัลกอริทึม Association Tree เปรียบเทียบกับ C4.5 แบบดั้งเดิม พบว่าเทคนิคของผู้วิจัยให้ประสิทธิภาพการจำแนกประเภทข้อมูลที่ดีที่สุดทั้งหมด 12 ชุดข้อมูลจากชุดข้อมูลทั้งสิ้น 16 ชุดข้อมูล และให้ประสิทธิภาพการทำนายที่เท่ากันจำนวน 3 ชุดข้อมูล แต่ให้ประสิทธิภาพที่ต่ำกว่าอยู่ 1 ชุดข้อมูล

จากการวิเคราะห์ชุดข้อมูลที่ให้ประสิทธิภาพที่ต่ำกว่า คือชุดข้อมูล Soybean โดยอัลกอริทึม C4.5 ให้ประสิทธิภาพเท่ากับ

92.39% และอัลกอริทึม Association Tree ให้ประสิทธิภาพเท่ากับ 52.12% สาเหตุเนื่องจากจำนวนตัวอย่างของชุดข้อมูลนี้ที่มีเพียง 368 ตัวอย่าง เมื่อเทียบกับจำนวนคลาสข้อมูลที่ชุดข้อมูลนี้มีมากถึง 19 คลาสข้อมูล โดยถ้าเทียบกับชุดข้อมูล Horse ที่มีจำนวนตัวอย่างเท่ากันแต่มีจำนวนคลาสเพียง 2 คลาสข้อมูล ซึ่งถือว่าข้อมูลมีเพียงพอต่อการเรียนรู้และจำแนกประเภทข้อมูล จึงเป็นสาเหตุผลการทดลองบนชุดข้อมูล Soybean ได้ประสิทธิภาพการจำแนกประเภทข้อมูลที่เปรียบเทียบกับ C4.5 แล้วได้ประสิทธิภาพที่ต่ำกว่า

7. เอกสารอ้างอิง

[1] Agrawal R., Srikant R., "Fast Algorithms for Mining Association Rules," VLDB, Chile, pp. 487-499, Sep 12-15 1994.

[2] Bing Liu., Wynne Hsu., Yiming Ma. 1998. "Integrating Classification and Association Rule Mining," In Proc Of the Fourth International Conference on Knowledge discovery and Data Mining (Ddd-98), pp. 80-86.

[3] Breiman L., Friedman J.H., Olshen R.A., and Stone C.J. "Classification and Regression Trees," Chapman & Hall, Inc., 1984.

[4] Kittisak Kerdprasop, Nittaya Kerdprasop, Naris Mingmora and Nurupon Wongprachanukul. "Wrapper and Filter Approaches to Feature Selection in Data Mining," Proceedings of 30th Congress on Science and Technology of Thailand, Impact Exhibition and Convention Center, Bangkok, Thailand, Oct. 19-21,2004.

[5] Merz, C. J, and Murphy, P. 1996. UCI repository of Machine Learning Database. [http://archive.ics.uci.edu/ml/index.html]

[6] Quinlan, J.Ross., "C4.5: Programs for machine learning," 1993.

[7] Quinlan, J.Ross., "Induction of Decision Trees. Machine Learning 1(1)," pp. 81-106, 1993.

[8] Vearapon Hanchodchung, Thanawin Rakthanmanon, Kitsana Waiyamai "Using Maximal Frequent Itemsets for Improving Associative Classification," NCCIT, 2005. Wenmin Li., Jiawei Han., and Jiawei Han., Jian Pei., "CMAR: Accurate and efficient classification base on multiple class association rules," In ICDM'01, pp.369-376, 2001.

เทคนิคการจำแนกข้อมูลด้วยต้นไม้เชิงสัมพันธ์

Data Classification Technique using Association Tree

(JCSSE2010) วันที่ 12 - 14 พฤษภาคม 2553

เทคนิคการจำแนกข้อมูลด้วยต้นไม้เชิงสัมพันธ์

Data Classification Technique using Association Tree

ลลิตา ศรีชัยญา และ นวลวรรณ สุนทรภิษัช
ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์
อีเมล์: g4964160@ku.ac.th, fscinws@ku.ac.th

บทคัดย่อ

การทำเหมืองข้อมูลเป็นกระบวนการขุดค้นความรู้ที่ซ่อนอยู่อยู่ในฐานข้อมูลขนาดใหญ่ ซึ่งมีเทคนิคหลักที่สำคัญ 2 เทคนิคที่ใช้สำหรับการทำเหมืองข้อมูล คือ การหากฎความสัมพันธ์ (Association Rules) และ การจำแนกประเภทข้อมูล (Data Classification) ซึ่งต่อมาได้มีนักวิจัยนำเทคนิคทั้ง 2 เทคนิคดังกล่าวมารวมกันเกิดเป็นเทคนิคหนึ่งที่ใช้ในการจำแนกประเภทข้อมูลที่สำคัญ ให้ประสิทธิภาพในการทำนายสูง และเป็นที่นิยมในปัจจุบัน นั่นคือ เทคนิคการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ (Associative Classification)

งานวิจัยนี้เป็นการนำเสนอ วิธีการคัดเลือกลักษณะเด่นของข้อมูลเพื่อให้ได้คุณลักษณะของข้อมูลฝึกที่เหมาะสมที่สุดสำหรับการเรียนรู้โดยใช้ต้นไม้ตัดสินใจ เพื่อให้ได้ต้นไม้ตัดสินใจที่มีประสิทธิภาพในการจำแนกข้อมูล โดยการนำเอาเทคนิคการหากฎความสัมพันธ์เข้ามาใช้ร่วมกับต้นไม้ตัดสินใจ ผู้วิจัยได้ทำการทดลองกับชุดข้อมูล 3 ชุดข้อมูล ที่ได้จากฐานข้อมูลมาตรฐาน UCI Machine Learning Repository โดยทำการทดลองวัดประสิทธิภาพเปรียบเทียบกับอัลกอริทึม C4.5, CBA และ CMAR ซึ่งผลการทดลองพบว่าอัลกอริทึมที่ผู้วิจัยนำเสนอมีประสิทธิภาพที่ดีกว่าใน 2 ชุดข้อมูลจาก 3 ชุดข้อมูล

คำสำคัญ: การทำเหมืองข้อมูล การคัดเลือกลักษณะเด่น การหากฎความสัมพันธ์ การจำแนกประเภทข้อมูล

Abstract

Data Mining is a knowledge discovery technique. There are two main techniques which are Association Rules and Data Classification. Researcher combined both association rules and data classification techniques called Associative Classification in order to improve classification performance.

This research proposes an itemsets feature selection algorithm that can filter out a set of information attributes for decision tree learning to construct a tree with high accuracy. We apply an association rule technique with classification technique. The experiments are done on three benchmark dataset collected from UCI Machine Learning Repository. We compare the performance with C4.5, CBA, CMAR and our proposed algorithm. The experimental results show that our proposed algorithm outperforms other algorithms in 2 datasets from 3 datasets.

Key Words: Data Mining, Feature Selection, Association Rules, Data Classification

1. บทนำ

เทคนิคการจำแนกประเภทข้อมูล (Data Classification) [5] และเทคนิคการหากฎความสัมพันธ์ (Association Rules) [1] เป็นเทคนิคหลักที่สำคัญในการทำเหมืองข้อมูล (Data Mining) ซึ่งเทคนิคการจำแนกประเภทข้อมูล เป็นการวิเคราะห์เพื่อจำแนกกลุ่มข้อมูลด้วยคุณสมบัติต่างๆ ซึ่งมีการกำหนดข้อมูลตัวอย่างของข้อมูลแต่ละประเภทเพื่อเป็นตัวแทนสำหรับการจำแนกประเภทข้อมูลทั้งหมด ลักษณะเช่นนี้เรียกว่า “Supervised Learning” เทคนิคการหากฎความสัมพันธ์เป็นการค้นหากฎความสัมพันธ์ของข้อมูล ความสำคัญของกฎใช้ค่าสนับสนุน (Support) และค่าความเชื่อมั่น (Confidence) บอกคุณภาพของกฎว่าดีหรือไม่เพียงใด

งานวิจัยที่มีการนำเทคนิคการหากฎความสัมพันธ์ร่วมกับเทคนิคการจำแนกประเภทข้อมูล เรียกว่าเทคนิคการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ (Associative Classification) เพื่อทำนายกลุ่มให้กับข้อมูลในอนาคตที่ไม่ทราบกลุ่มข้อมูล โดยการทำงานแบ่งเป็น 2 ส่วนคือ การสร้างกฎความสัมพันธ์ และการสร้างโมเดลการทำนาย โดยอัลกอริทึมที่เป็นต้นแบบของเทคนิคดังกล่าวคือ CBA [2] และต่อมาได้มีผู้นำมาพัฒนาต่อ ชื่อว่า CMAR [8]

งานวิจัยนี้มีการปรับปรุงเทคนิคการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ เพื่อให้มีประสิทธิภาพในการทำนายที่แม่นยำขึ้น โดยนำเสนอวิธีการสกัดลักษณะเด่นของข้อมูลฝึก และวิธีการคัดกรองข้อมูล เพื่อให้ได้ข้อมูลที่สำคัญที่สุดในการสร้างโหนดต้นไม้ตัดสินใจ การทดลองได้ใช้ข้อมูลจากฐานข้อมูล UCI Machine Learning Repository [4] ซึ่งเป็นฐานข้อมูลมาตรฐานที่มีผู้วิจัยหลายคนนำไปใช้ในการทดลอง ซึ่งผลการทดลองได้เปรียบเทียบกับอัลกอริทึมที่ใช้สำหรับการสร้างต้นไม้ตัดสินใจที่เป็นที่นิยมคือ C4.5 [5] และเปรียบเทียบกับเทคนิคการจำแนกประเภท

ข้อมูลโดยใช้กฎความสัมพันธ์ตัวอื่นคือ CBA [2] และ CMAR [8]

ในหัวข้อต่อไปคือ หัวข้อที่ 2 จะกล่าวถึงทฤษฎีที่เกี่ยวข้อง หัวข้อที่ 3 กล่าวถึงวิธีการทดลอง ซึ่งผลการทดลองจะอยู่ในหัวข้อที่ 4 และหัวข้อสุดท้ายเป็นการสรุปผลการทดลอง

2. ทฤษฎีที่เกี่ยวข้อง

2.1 *การหากฎความสัมพันธ์ (Association Rules)*

การหากฎความสัมพันธ์ เป็นเทคนิคที่อาศัยการเชื่อมโยงข้อมูล หรือกลุ่มข้อมูลที่เกิดขึ้นในเหตุการณ์เดียวกันเข้าด้วยกัน เช่น Market-Basket Analysis เพื่อวิเคราะห์ข้อมูลการสั่งซื้อสินค้า กฎความสัมพันธ์สามารถเขียนความสัมพันธ์ที่เกิดขึ้นระหว่างข้อมูลออกมาได้ในรูป $A \rightarrow B$ โดยที่ A เป็นเหตุของเหตุการณ์ (Antecedent) หรือ LHS (Left-Hand Side) และเรียก B เป็นผลของเหตุการณ์ (Consequent) หรือ RHS (Right-Hand Side) สามารถแสดงความสัมพันธ์ในรูปแบบของกฎความสัมพันธ์ตามสมการ (1)

$$LHS \rightarrow RHS \qquad (1)$$

โดยที่ LHS และ RHS เป็นชุดข้อมูลที่เกิดขึ้นบ่อยในฐานข้อมูล

ปัจจัยที่สำคัญที่ใช้ในการบอกคุณภาพของกฎว่าดีหรือไม่ดี ขึ้นอยู่กับปัจจัยที่สำคัญ 2 ประการ คือ

1. ค่าสนับสนุน (Support) เป็นค่าความถี่ที่ใช้วัดสัดส่วนข้อมูลที่เกิดขึ้นในทรานแซคชัน เช่น ค่าสนับสนุนของชุดข้อมูล A หมายถึง จำนวนทรานแซคชันที่ประกอบด้วยชุดข้อมูล A เมื่อเทียบกับจำนวนทรานแซคชันทั้งหมดในฐานข้อมูล

ค่าสนับสนุนที่ดีควรจะมีค่าในระดับที่สูง เพราะแสดงถึงความมีนัยสำคัญของความสัมพันธ์ของข้อมูล ในทางตรงข้ามถ้าค่าสนับสนุนมีค่าในระดับต่ำ แสดงถึงความไม่มีนัยสำคัญของความสัมพันธ์นั้น ค่าสนับสนุนขั้นต่ำ (Minimum Support) เป็นค่าสนับสนุนขั้นต่ำของชุดข้อมูล ซึ่งกำหนดโดยผู้ใช้งาน

2. ค่าความเชื่อมั่น (Confidence) เป็นค่าความถี่ของเหตุการณ์อื่นที่เกิดร่วมกับข้อมูลนั้น หรือเป็นค่าที่บ่งบอกว่ากฎหนึ่งมีความน่าเชื่อถือเพียงใด เช่น การหากฎที่มีระดับนัยสำคัญเมื่อมีเหตุการณ์ A เกิดขึ้นจำนวนหนึ่ง จะมีเหตุการณ์ B เกิดขึ้นด้วย หมายความว่า ต้องหาเงื่อนไขที่จะทำนายเหตุการณ์ B ที่เกิดขึ้น เนื่องจากเหตุการณ์ A

ค่าความเชื่อมั่นที่ดีควรจะมีค่าในระดับที่สูง เพราะแสดงให้เห็นถึงความน่าเชื่อมั่นของกฎว่าเหตุการณ์หนึ่ง จะสัมพันธ์กับอีกเหตุการณ์หนึ่งมีความน่าเชื่อมั่นเพียงใด ค่าความเชื่อมั่นขั้นต่ำ (Minimum Confidence) เป็นค่าความเชื่อมั่นขั้นต่ำของกฎ ซึ่งกำหนดโดยผู้ใช้งาน

2.2 *อัลกอริทึมเอพริออริ (Apriori Algorithm)*

อัลกอริทึมเอพริออริ เป็นอัลกอริทึมที่นิยมใช้อย่างแพร่หลายสำหรับการหากฎความสัมพันธ์สำหรับชุดข้อมูลที่เกิดขึ้นบ่อยในฐานข้อมูล (Frequent Itemset) ขั้นตอนการทำงานของอัลกอริทึมเอพริออริ เริ่มจากสร้าง Candidate Itemset ที่มีขนาดเท่ากับ k-Itemset โดยที่ k={1,2,3,…,n} ซึ่ง Candidate Itemset ในระดับที่ k มาจากการ join หรือ union กันระหว่าง Frequent Itemset ในระดับที่ k-1 เช่น Itemset(ABC) = Itemset(AB) ∪ Itemset(AC) จากนั้นหา Frequent Itemset ในระดับที่ k โดยการอ่านข้อมูลจากฐานข้อมูลเพื่อบันทึกค่าความถี่ของ Candidate Itemset แต่ละตัว แล้วทำการตรวจสอบว่า Candidate Itemset ใดมีคุณสมบัติเป็น Frequent Itemset ในระดับที่ k ดูจากค่าสนับสนุนมากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำ จากนั้นทำการสร้าง Candidate Itemset หา Frequent Itemset ในระดับที่ k+1 ต่อไป โดยทำเช่นนี้ไปเรื่อยๆ จนกว่าจะไม่สามารถทำการสร้าง Candidate Itemset ในระดับที่ k+1 ได้

2.3 *การจำแนกประเภทข้อมูล (Data Classification)*

การจำแนกประเภทข้อมูล เป็นเทคนิคการจำแนกกลุ่มข้อมูลด้วยคุณลักษณะต่างๆ โดยการสำรวจรายการในฐานข้อมูล เพื่อแยกแยะให้อยู่ในหมวดที่เราได้กำหนดไว้ล่วงหน้า เทคนิคที่ใช้ในการจำแนกประเภทข้อมูลมีมากมาย แต่มีเทคนิคหนึ่งเป็นที่นิยมและคุ้นเคยกันมาก คือ เทคนิคการใช้ต้นไม้ตัดสินใจ

อัลกอริทึมที่ใช้ในการสร้างต้นไม้ตัดสินใจที่นิยม เช่น C4.5 [5], CART [3], ID3 [6] เป็นต้น แต่ละอัลกอริทึมมีลักษณะแตกต่างกัน และต้นไม้ที่ได้จะมีลักษณะที่แตกต่างกันด้วย โดยลักษณะของโครงสร้างต้นไม้ที่มีขนาดเล็ก ไม่ซับซ้อน ยุ่งเหยิง จะเป็นต้นไม้ที่มีคุณภาพที่ดีที่สุด เหมาะสำหรับนำไปใช้ในการตัดสินใจ และจากงานวิจัยอัลกอริทึมที่ให้ประสิทธิภาพในการสร้างต้นไม้ตัดสินใจที่ดีที่สุด คือ อัลกอริทึม C4.5 หรือ J48 ในโปรแกรมเวกา [9]

2.4 *อัลกอริทึม C4.5*

อัลกอริทึม C4.5 เป็นอัลกอริทึมที่ได้รับความนิยม และมีประสิทธิภาพสูงสำหรับการสร้างต้นไม้ตัดสินใจ โดยขั้นตอนการเรียนรู้ด้วยต้นไม้ตัดสินใจจะเกี่ยวข้องกับการเลือกแอททริบิวที่เหมาะสมมาสร้างเป็นโหนดในต้นไม้ โดยแอททริบิวนั้นต้องสามารถแยกแยะข้อมูลออกเป็นประเภทที่ปะปนกันได้น้อยที่สุด วิธีการเลือกแอททริบิวมาสร้างโหนดต้นไม้ เป็นวิธีการเรียนรู้ที่อาศัยค่าความไร้ระเบียบของข้อมูล (Entropy) และค่าเกน (Gain) เป็นเครื่องมือในการเลือกแอท-ทริบิว หลักการคือ การประเมินว่าเมื่อเลือก

แอททริบิวนั้นมาสร้างเป็นโหนดต้นไม้ แล้วทำให้ข้อมูลสามารถจำแนกออกไปตามกิ่งของต้นไม้ได้ดีที่สุด คือ ข้อมูลมีการปะปนกันของข้อมูลน้อยที่สุด

ต้นไม้ตัดสินใจแสดงผลลัพธ์ในรูปของโครงสร้างต้นไม้ สามารถแปลงเป็นกฎที่เข้าใจได้ง่าย โดยแต่ละโหนดของต้นไม้แสดงถึงแอททริบิว แต่ละกิ่งแสดงถึงเงื่อนไขในการทดสอบ และโหนดปลายแสดงถึงประเภทข้อมูลที่กำหนดไว้

2.5 *หลักการทำงาน Associative Classification*

Associative Classification เป็นเทคนิคที่เกิดจากการรวมกันของ 2 เทคนิค คือ การหากฎความสัมพันธ์ และ การจำแนกประเภทข้อมูล โดยอัลกอริทึมที่รู้จักคือ CBA (Classification based on association rules) [2]

การทำงานของ Associative Classification แบ่งการทำงานเป็น 2 ส่วน คือ การสร้างกฎความสัมพันธ์ (Rule generator) และ การสร้างโมเดลการทำนาย (Classifier builder)

2.5.1 การสร้างกฎความสัมพันธ์ (Rule generator)

ส่วนการสร้างกฎความสัมพันธ์ นำเทคนิคการหากฎความสัมพันธ์เข้ามาใช้ ซึ่งอาศัยอัลกอริทึมที่รู้จัก คือ Apriori แต่จะมีข้อจำกัดในการสร้างกฎที่ จะสนใจเฉพาะกฎที่ค่าทางขวาเป็นข้อมูลคลาสเท่านั้น ซึ่งกฎแบบนี้เรียกว่า Class association rules (CARs) โดยขั้นตอนในการสร้างกฎความสัมพันธ์ มีดังนี้

1. ค้นหา frequent rule items ที่มีค่า support สูงกว่า minimum support ซึ่งขั้นตอนนี้จะเหมือนกับ Apriori ผลลัพธ์จะเลือกเฉพาะกฎที่เป็น CARs
2. สร้าง Class Association Rule (CARs) จาก frequent ruleitems และสำหรับทุกๆ ruleitems ที่มีเซตของ item เหมือนกัน เราจะเลือกเพียงกฎเดียวที่มีค่าความเชื่อมั่นสูงที่สุด
3. ทำการ Prune rule items ที่มีค่าความเชื่อมั่นน้อยกว่าค่า minimum confidence

2.5.2 การสร้างโมเดลในการทำนาย (Classifier builder)

หลังจากได้เซตของกฎ CARs แล้ว ส่วนนี้จะทำการเรียงกฎ และการสร้างโมเดลที่ใช้ในการทำนายข้อมูล โดยหลักการในการเรียงกฎมีดังนี้

กำหนดให้ 2 กฎ คือ r_i และ r_j สามารถกำหนดว่า r_i มีศักย์สูงกว่า r_j เมื่อ

1. ค่าความเชื่อมั่นของ r_i มากกว่า r_j หรือ
2. ค่าความเชื่อมั่นเท่ากัน แต่ค่าสนับสนุนของ r_i มากกว่า r_j หรือ
3. ทั้งค่าความเชื่อมั่นและค่าสนับสนุนเท่ากัน แต่ r_i ถูกสร้างมาก่อน r_j

ขั้นตอนการสร้างโมเดลในการทำนาย มีดังนี้

1. เรียงลำดับกฎทั้งหมด โดยใช้หลักการข้างต้น

2. วนลูปกฎ โดยนำกฎแต่ละกฎไปเช็คกับตัวข้อมูลที่เราทราบกลุ่มแล้ว เรียกว่า training data จากนั้น ถ้ากฎสามารถรองรับตัวข้อมูลทั้งค่าด้านซ้ายและค่าด้านขวา จะนำกฎไปเก็บไว้ในเซตของกฎ แล้วทำการหา default class เพื่อใช้ในการทำนายข้อมูล เพื่อใช้ในกรณีที่ไม่มีกฎไหนรองรับข้อมูล จากนั้นนำเซตของกฎมาคำนวณหาเปอร์เซนต์ข้อผิดพลาด

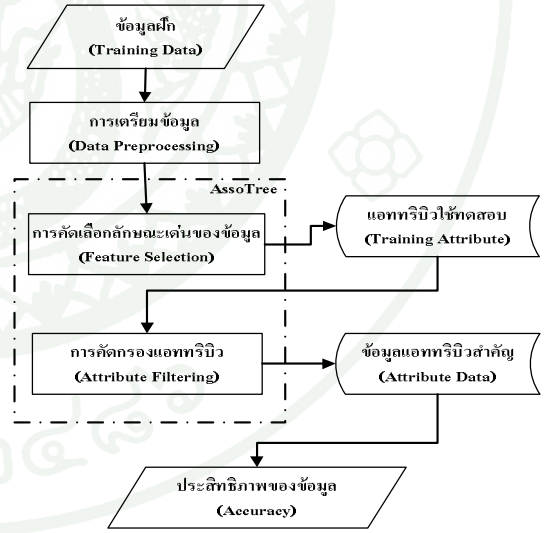
3. นำเซตของกฎทุกเซตมาเปรียบเทียบกัน แล้วเลือกเซตของกฎที่มีเปอร์เซนต์ความผิดพลาดน้อยที่สุด เพื่อนำเซตของกฎนั้นมาเป็นโมเดล เพื่อใช้ในการทำนายข้อมูลที่เราไม่ทราบกลุ่ม

2.6 *อัลกอริทึม CMAR*

อัลกอริทึม CMAR (Classification based on multiple class association rules) เป็นอัลกอริทึมที่พัฒนาต่อจากอัลกอริทึม CBA มีขั้นตอนแบบเดียวกับอัลกอริทึม CBA แต่ต่างกันที่การพิจารณากฎความสัมพันธ์ ซึ่งอัลกอริทึม CMAR จะพิจารณากฎความสัมพันธ์หลายๆกฎพร้อมกัน โดยนำเอาโครงสร้าง CR-tree มาใช้สำหรับเก็บและหาเซตของกฎที่ได้ ซึ่งขึ้นกับค่าความเชื่อมั่นและค่าสนับสนุน และส่วนของการสร้างโมเดล อาศัยค่า chi square (X^2) เป็นค่าน้ำหนักสำหรับแต่ละเซตของกฎ เพื่อเลือกเซตของกฎที่เหมาะสมที่จะนำมาสร้างเป็นโมเดลในการทำนายข้อมูล

3. วิธีการทดลอง

ส่วนนี้กล่าวถึงวิธีการในการดำเนินการวิจัย ผู้วิจัยได้นำเสนออัลกอริทึมชื่อ AssoTree ซึ่งมีขั้นตอนการทำงาน 3 ขั้นตอน คือ ขั้นตอนการเตรียมข้อมูล (Data Preprocessing) ขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล (Feature Selection) และขั้นตอนการคัดกรองแอททริบิว (Attribute Filtering)



รูปภาพ 1 ภาพรวมในการดำเนินการวิจัย

3.1 ขั้นตอนการเตรียมข้อมูล (Data Preprocessing)

ข้อมูลที่ใช้ในการทดลองมี 3 ชุดข้อมูล ได้มาจากฐานข้อมูล UCI Machine Learning

Repository [5] ลักษณะข้อมูลที่ทำการทดลองมีความหลากหลาย และแตกต่างกัน บางชุดข้อมูลมีค่าขาดหายไป (Missing Value) บางชุดข้อมูลลักษณะการจัดเก็บแตกต่างกัน มีทั้งข้อมูลชนิดโนมินัล (Nominal Attribute) ข้อมูลชนิดลำดับ (Ordinal Attribute) ข้อมูลชนิดต่อเนื่อง (Continuous Attribute) ดังนั้นจึงต้องมีการจัดการกับข้อมูล โดยพิจารณาตามลักษณะของข้อมูล ดังนี้

1. กรณีข้อมูลมีแอททริบิวไม่จำเป็นต่อการประมวลผล เช่น ลำดับของข้อมูล รหัสโรงพยาบาล จะทำการตัดแอททริบิวนั้นออกจากฐานข้อมูล
2. กรณีข้อมูลมีค่าบางส่วนขาดหายไป ทำการเติมค่าข้อมูลที่เป็นไปได้ ด้วยการคำนวณจากค่าที่มีมากสุด กับค่ากลางที่ได้จากข้อมูลที่เหลือในข้อมูลฝึก
3. กรณีข้อมูลเป็นชนิดต่อเนื่อง (Continuous Attribute) ทำการแปลงข้อมูลเป็นข้อมูลประเภทโนมินัล อาศัยเทคนิค binning ในการปรับข้อมูล โดยดูจากค่าของข้อมูลที่อยู่ใกล้กันวัตถุประสงค์ที่ต้องแปลงข้อมูล เนื่องจากในขั้นตอนของการคัดเลือกลักษณะเด่นของข้อมูล ที่อาศัยเทคนิคการหากฎความสัมพันธ์นั้น ไม่รองรับข้อมูลชนิดต่อเนื่อง

3.2 อัลกอริทึม AssociationTree

ในส่วนนี้จะกล่าวถึงขั้นตอนหลักที่ใช้ในการเพิ่มประสิทธิภาพให้กับเทคนิคการเรียนรู้ของต้นไม้ตัดสินใจโดยใช้กฎความสัมพันธ์ โดยแบ่งขั้นตอนการทำงานเป็น 2 ขั้นตอนหลัก ดังนี้

3.2.1 ขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล (Feature Selection)

ขั้นตอนการคัดเลือกลักษณะเด่นของข้อมูล หลักการการทำงาน คือ ใช้เทคนิคการหากฎความสัมพันธ์ โดยใช้อัลกอริทึมเอพริออริ [1] มีการกำหนดค่าเริ่มต้น 3 ค่า คือ จำนวนกฎ 100 กฎ ค่าสนับสนุนขั้นต่ำ 50% และค่าความเชื่อมั่น 90% เนื่องจากต้องการหาแอททริบิวที่สำคัญต่อการทำนายให้ได้มากที่สุด เพื่อไม่ให้ส่งผลต่อประสิทธิภาพของการทำนาย ในทุกรอบของการหากฎความสัมพันธ์ ถ้าจำนวนกฎที่ได้น้อยกว่าจำนวนกฎที่กำหนด จะลดค่าสนับสนุนขั้นต่ำลง ถ้าลดลงจนมีค่าเท่ากับ 0.1 แล้วยังได้จำนวนกฎน้อยกว่าที่กำหนด จะทำการลดค่าความเชื่อมั่นลงด้วย แต่ละรอบของการหากฎความสัมพันธ์จะเก็บ
แอททริบิวที่เกิดขึ้นทั้งหมด ผลลัพธ์ที่ได้จะเป็นข้อมูลแอททริ-บิวที่สำคัญต่อการทดลองขั้นต่อไป

3.2.2 ขั้นตอนการคัดกรองแอททริบิว (Attribute Filtering)

เนื่องจากขั้นตอนการคัดเลือกแอททริบิวได้กำหนดค่าเพื่อหาแอททริบิวที่เด่นที่มีความสำคัญต่อการจำแนกประเภทข้อมูล แต่ไม่สามารถมั่นใจได้ว่า ทุกแอททริบิวจะ

สำคัญต่อการจำแนกข้อมูลทั้งหมด ดังนั้นจึงมีการคัดกรองแอททริบิวเพื่อให้ได้แอททริบิวที่จำเป็นที่สุดสำหรับการจำแนกประเภทข้อมูล หลักการทำงานคือ ใช้ต้นไม้ตัดสินใจ โดยใช้อัลกอริทึม C4.5 หรือ J48 ในโปรแกรมเวกา (Weka Program) ขั้นแรกจะเตรียมชุดข้อมูลใหม่ที่มีเฉพาะแอททริบิวที่ได้จากขั้นตอนที่ผ่านมา แล้วทำการตัดแอททริบิวที่มี โดยหลักการในการเลือกตัดแอททริบิว คือ ให้ตัดแอททริบิวที่พบเป็นตัวสุดท้ายจากการทดลองในขั้นตอนที่ผ่านมาออกก่อน เนื่องจากสมมติฐานที่ว่า แอททริบิวที่พบในภายหลังอาจจะไม่มีความเป็นนัยสำคัญต่อความสัมพันธ์ของกฎ แต่ได้มาเพราะว่ามีการกำหนดค่าจึงส่งผลต่อการคัดเลือกแอททริบิวมาสร้างเป็นกฎความสัมพันธ์ จากนั้นนำชุดข้อมูลที่ประกอบด้วยเซตแอททริบิวที่เหลือไปสร้างต้นไม้ตัดสินใจ แล้วเก็บค่าความถูกต้องไว้ จากนั้นทำการตัดแอททริบิวตัวถัดไป และทำการเปรียบเทียบค่าความถูกต้องที่ได้ ถ้าค่าความถูกต้องมีค่ามากกว่าที่บันทึกไว้ก่อนหน้า จะบันทึกค่าที่มากกว่าไว้ แล้ววนตัดแอททริบิวตัวถัดไป แต่ถ้าค่าความถูกต้องมีค่าน้อยกว่า จะพิจารณาแอททริบิวตัวถัดไปอีก 2 แอททริบิวเท่านั้น เนื่องจากถ้าตัดแอททริบิวออกมากจนเหลือแอททริบิวที่ใช้ในการสร้างต้นไม้ตัดสินใจน้อยเกินไป จะส่งผลต่อประสิทธิภาพการทำนายข้อมูลของต้นไม้ตัดสินใจ ท้ายที่สุดของขั้นตอนการคัดกรองแอททริบิวจะได้ชุดข้อมูลที่ประกอบด้วยเซตแอททริบิวที่สำคัญที่สุดต่อการจำแนกประเภทข้อมูล และมีประสิทธิภาพในการทำนายข้อมูล

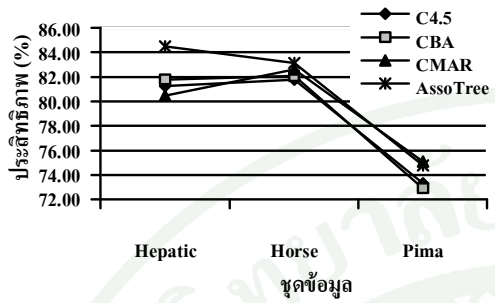
4. ผลการทดลอง

ในหัวข้อนี้จะกล่าวถึง การเปรียบเทียบประสิทธิภาพในการจำแนกประเภทข้อมูล โดยนำผลการทดลองที่ได้จากการทดลองด้วยอัลกอริทึม AssoTree เปรียบเทียบกับเทคนิคต้นไม้ตัดสินใจ คือ C4.5 [5] และเทคนิคการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ 2 เทคนิคคือ CBA [2] และ CMAR [8] ได้ทำการทดลองกับชุดข้อมูล 3 ชุดข้อมูล ได้แก่ Hepatic, Horse และ Pima ได้มาจากฐานข้อมูล UCI Machine Learning Repository [4]

ในการทดลองได้ใช้ค่าสนับสนุน ค่าความเชื่อมั่น และการทดลองแบบ 10 โฟลด-ครอสวาดิเดชัน (10 fold cross-validation) ซึ่งผลการเปรียบเทียบประสิทธิภาพตามตารางที่ 1 และแสดงการเปรียบเทียบในรูปแบบกราฟตามภาพที่ 2

ตารางที่ 1 ข้อมูลเปรียบเทียบประสิทธิภาพ

ชุดข้อมูล	C4.5	CBA	CMAR	AssoTree
Hepatic	81.29	81.80	80.50	84.52
Horse	81.79	82.10	82.60	83.15
Pima	73.31	72.90	75.10	74.74



รูปภาพ 2. กราฟเปรียบเทียบประสิทธิภาพ

จากการทดลองพบว่า อัลกอริทึม AssoTree ของผู้วิจัยให้ประสิทธิภาพการจำแนกประเภทข้อมูลที่ดีที่สุด ในชุดข้อมูล Hepatic และ Horse แต่ให้ประสิทธิภาพที่ต่ำกว่าชุดข้อมูล Pima ตามตารางที่ 2 จากการวิเคราะห์พบว่า อัลกอริทึม AssoTree สามารถทำงานให้ประสิทธิภาพที่ดีกับข้อมูลที่มีจำนวนแอททริ-บิวมาก เนื่องจากขั้นตอนของอัลกอริทึมจะอาศัยแอททริบิวในการพิจารณา ดังนั้นถ้าจำนวนแอททริบิวมากก็จะเพียงพอต่อการเรียนรู้และจำแนกข้อมูล จึงเป็นสาเหตุทำให้ชุดข้อมูล Pima ให้ประสิทธิ-ภาพที่ต่ำกว่า เนื่องจากมีจำนวนแอททริ-บิวเพียง 9 แอททริบิว แม้จะมีจำนวนตัวอย่างมากที่สุดเมื่อเทียบกับอีก 2 ชุดข้อมูลที่เปรียบเทียบด้วยก็ตาม

5. สรุปผลการทดลอง

ในงานวิจัยนี้แสดงให้เห็นว่า แอททริบิวที่ใช้ในการจำแนกประเภทข้อมูลมีส่วนสำคัญต่อประสิทธิภาพของการทำนายข้อมูล เพราะแอททริบิวเป็นสิ่งที่นำไปสร้างเป็นต้นไม้ตัดสินใจ โดยแอททริบิวที่มีทั้งหมดจากชุดข้อมูลไม่จำเป็นว่าทุกแอททริบิวจะมีความสำคัญต่อการทำนายทั้งหมด มีแค่บางแอท-ทริบิวเท่านั้นก็เพียงพอต่อการทำนายผล และยังสามารถทำให้ต้นไม้ตัดสินใจมีขนาดเล็ก ซึ่งต้นไม้ที่มีขนาดเล็กจะไม่ซับซ้อน และให้ประสิทธิภาพในการทำนายที่ดีกว่าต้นไม้ขนาดใหญ่ ผู้วิจัยจึงได้นำเสนออัลกอริทึมการคัดเลือกลักษณะเด่นของข้อมูล โดยใช้เทคนิคการหากฎความสัมพันธ์ร่วมกับเทคนิคต้นไม้ตัดสินใจโดยขั้นตอนวิธี AssoTree เหมาะสำหรับชุดข้อมูลที่มีจำนวนแอททริบิวมาก

6. เอกสารอ้างอิง

[1] Agrawal R., Srikant R., “Fast Algorithms for Mining Association Rules,” VLDB, Chile, pp. 487-499, Sep 12-15 1994.

[2] Bing Liu., Wynne Hsu., Yiming Ma, “Integrating Classification and Association Rule Mining,” Proceeding of the Fourth International Conference on Knowledge Discovery and Data Mining (Kdd-98), New York, USA, pp. 80-86, 1998. (The CBA system can be downloaded from http://www.comp.nus.edu.sg/~dm2)

[3] Breiman L., Friedman J.H., Olshen R.A., and Stone C.J., “Classification

and Regression Trees," Chapman &
Hall, Inc., 1984.

[4] Merz, C. J., and Murphy P., 1996, UCI Repository of Machine Learning Database. http://archive.ics.uci.edu/ml/index.html

[5] Quinlan, J. Ross., "C4.5: Programs for Machine Learning," 1993.

[6] Quinlan, J. Ross., "Induction of Decision Trees Machine Learning 1(1)," pp. 81-106, 1993.

[7] Vearapon Hanchodchung, Thanawin Rakthanmanon, Kitsana Waiyamai, "Using Maximal Frequent Itemsets for Improving Associative Classification," NCCIT, 2005.

[8] Wenmin Li., Jiawei Han., and Jian Pei., "CMAR: Accurate and Efficient Classification Base on Multiple Class Association Rules," In ICDM'01, pp. 369-376, 2001.

[9] Weka 3, http://www.cs.waikato.ac.nz/~ml/weka/

ประวัติการศึกษา และการทำงาน

ชื่อ –นามสกุล	นางสาวลลิตา ศรีชัยญา
วัน เดือน ปี ที่เกิด	วันที่ 14 เมษายน 2525
สถานที่เกิด	กรุงเทพมหานคร
ประวัติการศึกษา	ระดับปริญญาตรี คณะวิทยาศาสตร์ วิชาเอกวิทยาการคอมพิวเตอร์ มหาวิทยาลัยเกษตรศาสตร์
ผลงานดีเด่นและรางวัลทางวิชาการ	1. นำเสนอแบบปากเปล่าและตีพิมพ์บทความวิจัยเรื่อง "ขั้นตอนวิธีการเพิ่มประสิทธิภาพการเรียนรู้ของต้นไม้ตัดสินใจโดยใช้กฎความสัมพันธ์" ในการประชุมวิชาการ The National Conference on Information Technology 2008 (NCIT 2008) ณ วันที่ 6-7 พฤศจิกายน 2551 2. นำเสนอแบบปากเปล่าและตีพิมพ์บทความวิจัยเรื่อง "เทคนิคการจำแนกข้อมูลด้วยต้นไม้เชิงสัมพันธ์" ในการประชุมวิชาการ The 7th International Joint Conference on Computer Science and Software Engineering (JCSSE 2010) ณ วันที่ 12 - 14 พฤษภาคม 2553