



รายงานการวิจัย

การวิเคราะห์ดิสคริมิแนนต์ด้วยเทคนิคการทำเหมืองความสัมพันธ์ สำหรับ
ระบบสนับสนุนการตัดสินใจด้านการแพทย์
(Discriminant Analysis with Association Mining Technique for
Medical Decision Support System)

มหาวิทยาลัยเทคโนโลยีสุรนารี

ได้รับทุนอุดหนุนการวิจัยจาก
มหาวิทยาลัยเทคโนโลยีสุรนารี

ผลงานวิจัยเป็นความรับผิดชอบของหัวหน้าโครงการวิจัยแต่เพียงผู้เดียว



รายงานการวิจัย

การวิเคราะห์ดิสคริมิแนนต์ด้วยเทคนิคการทำเหมืองความสัมพันธ์ สำหรับ
ระบบสนับสนุนการตัดสินใจด้านการแพทย์

(Discriminant Analysis with Association Mining Technique for
Medical Decision Support System)

ผู้วิจัย

รองศาสตราจารย์ ดร.นิตยา เกิดประสพ

รองศาสตราจารย์ ดร.กิตติศักดิ์ เกิดประสพ

สาขาวิชาวิศวกรรมคอมพิวเตอร์

สำนักวิชาวิศวกรรมศาสตร์

หัวหน้าโครงการ

ผู้วิจัยร่วม

ได้รับทุนอุดหนุนการวิจัยจากมหาวิทยาลัยเทคโนโลยีสุรนารี ปีงบประมาณ พ.ศ. 2557 2558 และ 2559

ผลงานวิจัยเป็นความรับผิดชอบของหัวหน้าโครงการวิจัยแต่เพียงผู้เดียว

กิตติกรรมประกาศ

คณะผู้วิจัยขอขอบคุณมหาวิทยาลัยเทคโนโลยีสุรนารี และสำนักงานคณะกรรมการวิจัยแห่งชาติ ที่สนับสนุนโครงการวิจัยนี้ด้วยการจัดสรรงบประมาณให้อย่างพอเพียงและต่อเนื่อง ตั้งแต่ปีงบประมาณ พ.ศ.2557 2558 และ 2559 รวมถึงขอขอบคุณผู้ทรงคุณวุฒิทั้งภายนอกและภายในมหาวิทยาลัย ที่ได้เสียสละเวลาทำหน้าที่ตรวจข้อเสนอโครงการวิจัยและร่างรายงานการวิจัยฉบับสมบูรณ์ ข้อเสนอแนะจากผู้ทรงคุณวุฒิทุกท่านเป็นประโยชน์อย่างมากต่อคณะผู้วิจัยในการปรับปรุงการออกแบบ และขั้นตอนการดำเนินงานของโครงการวิจัย งานวิจัยนี้สำเร็จได้อย่างดีด้วยการมีส่วนร่วมจากนักศึกษาทั้งในระดับปริญญาโทบัณฑิตและปริญญาตรีบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ ที่ได้ทำหน้าที่เป็นผู้ช่วยวิจัยในโครงการวิจัยนี้



บทคัดย่อภาษาไทย

การวิเคราะห์ดิสคริมิแนนต์เป็นวิธีทางสถิติที่รู้จักกันแพร่หลายสำหรับแก้ปัญหาการจำแนกข้อมูลหลายกลุ่มที่เป็นการวิเคราะห์แบบหลายตัวแปร โดยการวิเคราะห์เพื่อจำแนกกลุ่มใช้วิธีเรียนรู้ลักษณะเด่นในกลุ่มข้อมูลที่แตกต่างจากข้อมูลกลุ่มอื่นและใช้ลักษณะที่เรียนรู้ได้สำหรับจำแนกข้อมูลที่เกิดขึ้นใหม่ การวิเคราะห์ดิสคริมิแนนต์นี้มีศักยภาพสูงที่จะประยุกต์เป็นฐานความรู้หลักในระบบสนับสนุนการตัดสินใจด้านการแพทย์ แต่ข้อจำกัดที่สำคัญคือผลการวิเคราะห์ดิสคริมิแนนต์จะมีตัวแปรการทำนายที่มากเกินไปเนื่องจากข้อมูลจริงในทางการแพทย์มีการบันทึกรายละเอียดผู้ป่วยไว้จำนวนมากทั้งข้อมูลทางคลินิกและประวัติการรักษา โครงการวิจัยนี้จึงมีวัตถุประสงค์ที่จะนำเสนอเทคนิคอื่นที่จะช่วยให้สามารถวิเคราะห์ดิสคริมิแนนต์ได้โดยการประยุกต์ใช้วิธีการทำเหมืองความสัมพันธ์ ข้อกำหนดของเทคนิคที่นำเสนอใหม่คือจะต้องให้โมเดลข้อมูลที่มีความถูกต้องสูงและมีขนาดโมเดลที่เล็กเพื่อให้เหมาะสมที่จะแปลงเป็นฐานความรู้ งานวิจัยนี้จึงนำเสนอเทคนิคการผนวกรวมการทำเหมืองความสัมพันธ์ร่วมกับการทำเหมืองข้อมูลแบบจำแนก โดยวัตถุประสงค์หลักคือสร้างโมเดลเพื่อการจำแนกแต่ประยุกต์ใช้การทำเหมืองความสัมพันธ์ที่มีข้อเด่นด้านการเรียนรู้ความสัมพันธ์ของลักษณะข้อมูล งานวิจัยนี้ได้เพิ่มเติมเทคนิคฟัซซีเพื่อจัดการกับข้อมูลตัวเลขที่เป็นค่าต่อเนื่อง ในการทดสอบเพื่อประเมินประสิทธิภาพของวิธีการที่เสนอขึ้น ใช้การวัดประสิทธิภาพเปรียบเทียบกับเทคนิคอื่น ๆ ในกลุ่มวิธีการทำเหมืองแบบจำแนก กลุ่มการทำเหมืองความสัมพันธ์เพื่อการจำแนก และกลุ่มที่ใช้เทคนิคฟัซซีร่วมกับการทำเหมืองความสัมพันธ์เพื่อการจำแนก ผลการทดสอบเปรียบเทียบยืนยันว่าเทคนิคที่เสนอขึ้นใหม่นี้มีประสิทธิภาพสูง

บทคัดย่อภาษาอังกฤษ

Discriminant analysis is a well-known multivariate statistical analysis for solving classification problem, where two or more groups of populations are given for measuring common characteristics that can best discriminate the distinct groups and then apply the learned measurement to classify new observations. The discriminant analysis has great potential to be applied to the medical domain as a core knowledge base in the decision support system. However, applying discriminant analysis for the decision support system is hindered by the excessive amount of predictive variables when dealing with real-world data that record so many clinical measures and treatments on each patient. The objective of this research project is to propose a different technique of discriminant analysis by applying the association mining to learn the discriminative characteristics among different groups of populations. Major criteria of the proposed association mining based discriminant analysis are to obtain a high accurate model as well as a small set of model suitable for transferring to be a knowledge base. Therefore, this research proposes a combination of association rule mining and data classification rule induction techniques. Association rule mining is good at finding relationships among the whole data set and represents them as association rules. This research also uses fuzzy set technique to control a continuous data to enhance efficiency of the data classification. To evaluate the performance of the proposed method, this research compares accuracy of the classification rules and the number of rules obtained from different kinds of data classification algorithms. These algorithms include the traditional data classification algorithms, the associative classification algorithms, and the fuzzy association rule-based classifier algorithms. The experimental results confirm the efficiency of the proposed method.

สารบัญ

	หน้า
กิตติกรรมประกาศ	ก
บทคัดย่อภาษาไทย	ข
บทคัดย่อภาษาอังกฤษ	ค
สารบัญ	ง
สารบัญตาราง	ฉ
สารบัญภาพ	ช
บทที่ 1 บทนำ	
1.1 ความสำคัญและที่มาของปัญหาการวิจัย	1
1.2 วัตถุประสงค์ของโครงการวิจัย	7
1.3 ขอบเขตของการวิจัย	7
1.4 ประโยชน์ที่ได้รับ	7
บทที่ 2 เอกสารและงานวิจัยที่เกี่ยวข้อง	
2.1 การวิเคราะห์ทฤษฎีนิรนัย	8
2.2 งานวิจัยที่เกี่ยวข้อง	9
บทที่ 3 การออกแบบและพัฒนาโปรแกรม	
3.1 กรอบแนวคิดของงานวิจัย	12
3.2 การออกแบบโมดูลสร้างโมเดลเพื่อการวินิจฉัย (Knowledge Modeling)	13
3.3 อัลกอริทึมจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือที่กะทัดรัด ...	14
3.4 การแปลงโมเดลเป็นฐานความรู้	19
บทที่ 4 ผลการทดสอบโปรแกรม CCFAR	
4.1 ข้อมูลที่ใช้ในการทดสอบ	22
4.2 เกณฑ์ที่ใช้ในการทดสอบ	24
4.3 ผลการทดสอบประสิทธิภาพการจำแนกข้อมูลของโปรแกรม CCFAR	26
4.4 อภิปรายผลการทดสอบ	30
บทที่ 5 บทสรุป	
5.1 สรุปผลการวิจัย	31
5.2 ข้อจำกัดของโปรแกรมและข้อเสนอแนะ	33
บรรณานุกรม	34

สารบัญ(ต่อ)

	หน้า
ภาคผนวก ก ผลผลิตของงานวิจัย	38
ภาคผนวก ก บทความวิจัยตีพิมพ์ในวารสารและเอกสารการประชุมวิชาการ	39
<i>International Journal Article</i>	
1. P. Kongchai, N. Kerdprasop and K. Kerdprasop (2014). The fuzzy search for association rules with interestingness measures. <i>International Journal of Computer Theory and Engineering</i> , Vol.6, No.6, December, pp.490-494. (indexed in EI and INSPEC, ISSN: 1793-8201)	40
<i>Refereed International Conference Proceedings</i>	
2. P. Kongchai, K. Suksut, R. Sutamma, S. Phoemhansa, N. Kerdprasop, K. Kerdprasop (2015). The compact fuzzy association rules for data classification. <i>Proceedings of the 3rd International Conference on Industrial Application Engineering 2015 (ICIAE2015)</i> , Kitakyushu, Japan, 28-31 March, pp.30-37.	45
3. K. Suksut, P. Kongchai, S. Phoemhansa, R. Sutamma, K. Kerdprasop, N. Kerdprasop (2015). Single versus multiple measures for fuzzy association rule mining. <i>Proceedings of the 3rd International Conference on Industrial Application Engineering 2015 (ICIAE2015)</i> , Kitakyushu, Japan, 28-31 March, pp.273-279.	53
4. P. Kongchai, N. Kerdprasop, K. Kerdprasop (2013). Dissimilar rule mining and ranking technique for associative classification. <i>Proceedings of the 2013 IAENG International Conference on Data Mining and Applications</i> , Hong Kong, 13-15 March, pp.356-361.	60
ภาคผนวก ข ลิขสิทธิ์โปรแกรม	66
โปรแกรมจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือที่กะทัดรัด (Data classification program with compact fuzzy association rules)	
ประวัติผู้วิจัย	69

สารบัญตาราง

	หน้า
ตารางที่ 1.1 ข้อมูลการวินิจฉัยความเหมาะสมของการใช้คอนแทกเลนส์ในกลุ่มผู้มีปัญหา ด้านสายตา	2
ตารางที่ 2.2 รายละเอียดข้อมูลของคนไข้รายใหม่	9
ตารางที่ 3.1 ข้อมูลตัวอย่างประกอบการอธิบายอัลกอริทึม CCFAR	14
ตารางที่ 3.2 แอททริบิวต์ Age, Income, Balance และค่าความเป็นสมาชิกในแต่ละ ช่วงข้อมูล	15
ตารางที่ 3.3 ไอเท็มเซตแบบคลุมเครือที่ปรากฏบ่อยสูงกว่าเกณฑ์ขั้นต่ำ	17
ตารางที่ 3.4 กฎการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือ	18
ตารางที่ 3.5 กฎการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือที่มีค่าคะแนน สูงสุด	19
ตารางที่ 4.1 ข้อมูลที่ใช้ในการทดสอบประสิทธิภาพของโมเดลที่ได้จากโปรแกรม CCFAR	22
ตารางที่ 4.2 รายละเอียดแอททริบิวต์ของข้อมูล heart disease	23
ตารางที่ 4.3 รายละเอียดแอททริบิวต์ของข้อมูล Pima Indians diabetes	24
ตารางที่ 4.4 ผลการทดสอบประสิทธิภาพของโปรแกรมกับชุดข้อมูล heart disease	28
ตารางที่ 4.5 ผลการทดสอบประสิทธิภาพของโปรแกรมกับชุดข้อมูล Pima Indians diabetes	28
ตารางที่ 5.1 สรุปลำดับความสามารถของโปรแกรมตามเกณฑ์ค่าความเหมาะสมของกฎ	32



สารบัญภาพ

	หน้า
รูปที่ 1.1 เปรียบเทียบผลการวิเคราะห์ข้อมูลเพื่อวินิจฉัยความเหมาะสมในการใส่คอนแท็กเลนส์ ด้วยวิธี (a) การวิเคราะห์การถดถอยโลจิสติก และ (b) การวิเคราะห์ดิสคริมิแนนต์	4
รูปที่ 1.2 การกำหนดค่าพารามิเตอร์ของอัลกอริทึม ClassificationViaRegression	5
รูปที่ 1.3 แสดงกฎความสัมพันธ์ส่วนหนึ่งของข้อมูลคอนแท็กเลนส์ จากกฎทั้งหมด 64 กฎที่ ตรงตามเกณฑ์ support ≥ 0.1 และ confidence = 1.0	6
รูปที่ 2.1 โมเดลที่ได้จากการวิเคราะห์ดิสคริมิแนนต์เพื่อวินิจฉัยการใส่คอนแท็กเลนส์	8
รูปที่ 3.1 กรอบของงานพัฒนาระบบสนับสนุนการตัดสินใจด้านการแพทย์	12
รูปที่ 3.2 ขั้นตอนการทำงานของอัลกอริทึม CCFAR	14
รูปที่ 3.3 ฐานความรู้ที่ได้จากกฎการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือ	19
รูปที่ 3.4 การโต้ตอบกับฐานความรู้ของระบบสนับสนุนการตัดสินใจอย่างง่าย	20
รูปที่ 3.5 การสอบถามเหตุผลของการได้มาซึ่งคำตอบของระบบ	21
รูปที่ 3.6 คำตอบของระบบในกรณีที่ตัวเลือกของผู้ใช้ไม่ตรงกับข้อมูลที่มีในฐานความรู้	21
รูปที่ 4.1 กราฟเปรียบเทียบประสิทธิภาพของโปรแกรมกับชุดข้อมูล heart disease	29
รูปที่ 4.2 กราฟเปรียบเทียบประสิทธิภาพของโปรแกรมกับชุดข้อมูล Pima Indians diabetes	29

บทที่ 1

บทนำ

1.1 ความสำคัญและที่มาของปัญหาการวิจัย

การวิเคราะห์ข้อมูลในทางสถิติเมื่อข้อมูลที่ต้องการวิเคราะห์ประกอบขึ้นจากแอททริบิวต์หรือตัวแปรมากกว่าหนึ่งตัวแปร นักสถิติเรียกตัวแปรที่เป็นจุดสนใจหลักของการวิเคราะห์ว่าตัวแปรเป้าหมายหรือตัวแปรตาม (dependent variable) และเรียกตัวแปรอื่นที่เหลือว่าตัวแปรต้น หรือตัวแปรอิสระ (independent variable) วัตถุประสงค์ของการวิเคราะห์เป็นได้หลายประการ เช่น เพื่อพิจารณาความสัมพันธ์ระหว่างตัวแปร เพื่อสร้างโมเดลในลักษณะฟังก์ชันหรือสมการของตัวแปรที่สามารถจำแนกข้อมูลออกเป็นกลุ่มย่อย หรือเพื่อค้นหาตัวแปรหลักเพียงบางตัวแปรที่มีบทบาทสำคัญในการจำแนกกลุ่มข้อมูล ลักษณะงานเช่นนี้นักสถิติอาจเลือกใช้เทคนิคการวิเคราะห์จำแนกประเภทหรือการวิเคราะห์ดิสคริมิแนนต์ (discriminant analysis) การวิเคราะห์การถดถอย (regression analysis) และการวิเคราะห์ความแปรปรวน (analysis of variance -- ANOVA)

เทคนิคการวิเคราะห์ข้อมูลทั้งสามแบบนี้ให้ผลลัพธ์ในลักษณะเดียวกัน แต่มีความแตกต่างกันที่ประเภทหรือชนิดข้อมูลของตัวแปรตามที่เป็นเป้าหมายของการวิเคราะห์ การวิเคราะห์การถดถอยและการวิเคราะห์ความแปรปรวนมีข้อกำหนดที่สำคัญคือ ตัวแปรตามต้องเป็นข้อมูลตัวเลขในลักษณะค่าต่อเนื่อง (continuous value) เช่น ตัวเลขอุณหภูมิ ค่ารายได้ ดัชนีตลาดหลักทรัพย์ มูลค่าผลกำไร เป็นต้น วัตถุประสงค์หลักของการวิเคราะห์การถดถอยและการวิเคราะห์ความแปรปรวนมักจะเพื่อการอธิบายหรือการทำนายค่าของตัวแปรเป้าหมาย ในขณะที่การวิเคราะห์ดิสคริมิแนนต์ใช้กับกรณีที่ตัวแปรเป้าหมายเป็นค่าที่แจกแจงได้ หรือเรียกว่าค่ากลุ่ม (categorical value) ค่าที่แจกแจงได้นี้เป็นลักษณะสัญลักษณ์หรือข้อความที่ใช้จำแนกกลุ่มข้อมูล เช่น สูง-ปานกลาง-ต่ำ หรือ มาก-น้อย วัตถุประสงค์ของการวิเคราะห์ดิสคริมิแนนต์จะเป็นได้ทั้งเพื่อการทำนายกลุ่ม เพื่อการอธิบายลักษณะข้อมูลในกลุ่ม และเพื่อการค้นหาปัจจัยหลักที่มีบทบาทสำคัญในการแบ่งแยกกลุ่ม

ในงานการวิเคราะห์ข้อมูลด้านการแพทย์ ข้อมูลหนึ่งระเบียนหรือหนึ่งรายการมักจะประกอบด้วยข้อมูลหลายลักษณะปะปนกัน ทั้งข้อมูลที่เป็นค่าต่อเนื่องและข้อมูลกลุ่มที่เป็นค่าแจกแจงได้ นอกจากนี้ข้อมูลในแต่ละกลุ่มหรือแต่ละคลาสยังมีความสำคัญไม่เท่ากันเช่น ข้อมูลของผู้ป่วยโรคร้ายแรงมีความสำคัญมากกว่าข้อมูลของบุคคลปกติที่มารับการตรวจสุขภาพประจำปี งานวิจัยนี้จึงให้ความสำคัญกับการวิเคราะห์ดิสคริมิแนนต์มากกว่าการวิเคราะห์ด้วยเทคนิคอื่น เมื่อข้อมูลที่ต้องการจำแนกหรือทำนายคลาสเป็นค่าไม่ต่อเนื่องที่แจกแจงได้ นักสถิติอาจจะเลือกใช้เทคนิคการวิเคราะห์การถดถอยโลจิสติก (logistic regression) ที่ให้ผลลัพธ์เป็นสมการเชิงเส้นเช่นเดียวกับการวิเคราะห์

ดิสคริมิแนนต์ รูปแบบทั่วไปของสมการจำแนกจะเป็น $Y = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \alpha$ เมื่อ Y คือ ตัวแปรเป้าหมายที่ต้องการทำนายค่า X_1, \dots, X_k คือตัวแปรอิสระจำนวน k ตัวที่ทำหน้าที่ทำนายค่าของตัวแปรเป้าหมาย β_1, \dots, β_k คือสัมประสิทธิ์ที่บอกระดับความสำคัญของแต่ละตัวแปร และ α คือค่าคงที่

ถ้าข้อมูลในส่วนตัวแปรอิสระมีการกระจายแบบปกติ (normal distribution) และจำนวนข้อมูลมีมากเพียงพอ การวิเคราะห์ดิสคริมิแนนต์มักจะให้สมการเพื่อการทำนายที่แม่นยำกว่า เทคนิคการวิเคราะห์การถดถอยโลจิสติก (Tabachnick & Fidell, 1996) ข้อสังเกตนี้ยืนยันได้จากผลการวิเคราะห์ข้อมูล contact-lens (Cendrowska, 1987) ตามตารางที่ 1.1 ที่มีจำนวนข้อมูล 24 รายการ ตัวแปรอิสระมีจำนวน 4 ตัวแปร คือ age, spectacle-prescription, astigmatic และ tear-production-rate ตัวแปรทั้งหมดมีการกระจายแบบปกติ เป้าหมายของการวิเคราะห์ คือ การสร้างสมการเชิงเส้นเพื่อทำนายค่า wear-contact-lens ซึ่งเป็นการพิจารณาว่าผู้ที่มีปัญหาด้านสายตา รายนั้น ๆ สามารถใส่คอนแทกเลนส์ได้หรือไม่ (หมายเหตุ ข้อมูลนี้เดิมจำแนกประเภทการวินิจฉัยการใส่คอนแทกเลนส์เป็นสามประเภทคือ soft, hard, none ข้อมูลในตารางที่ 1.1 เปลี่ยนแปลงค่าของ soft และ hard เป็น yes และเปลี่ยนค่า none เป็น no เพื่อลดจำนวนการจำแนกเป็นเพียงสองกลุ่ม คือ yes/no)

ตารางที่ 1.1 ข้อมูลการวินิจฉัยความเหมาะสมของการใช้คอนแทกเลนส์ในกลุ่มผู้มีปัญหาด้านสายตา

Age of the patient	Spectacle prescription	Astigmatic	Tear production rate	Wear contact lenses
young	myope	no	reduced	no
young	myope	no	normal	yes
young	myope	yes	reduced	no
young	myope	yes	normal	yes
young	hypermetrope	no	reduced	no
young	hypermetrope	no	normal	yes
young	hypermetrope	yes	reduced	no
young	hypermetrope	yes	normal	yes
pre-presbyopic	myope	no	reduced	no
pre-presbyopic	myope	no	normal	yes
pre-presbyopic	myope	yes	reduced	no

ตารางที่ 1.1 ข้อมูลการวินิจฉัยความเหมาะสมของการใช้คอนแทกเลนส์ในกลุ่มผู้มีปัญหาทางด้านสายตา (ต่อ)

Age of the patient	Spectacle prescription	Astigmatic	Tear production rate	Wear contact lenses
pre-presbyopic	myope	yes	normal	Yes
pre-presbyopic	hypermetrope	no	reduced	no
pre-presbyopic	hypermetrope	no	normal	yes
pre-presbyopic	hypermetrope	yes	reduced	no
pre-presbyopic	hypermetrope	yes	normal	no
presbyopic	myope	no	reduced	no
presbyopic	myope	no	normal	no
presbyopic	myope	yes	reduced	no
presbyopic	myope	yes	normal	yes
presbyopic	hypermetrope	no	reduced	no
presbyopic	hypermetrope	no	normal	yes
presbyopic	hypermetrope	yes	reduced	no
presbyopic	hypermetrope	yes	normal	no

ผลการวิเคราะห์ข้อมูลในตารางที่ 1.1 ด้วยเทคนิคการวิเคราะห์การถดถอยโลจิสติก และเทคนิคการวิเคราะห์ดิสคริมิแนนต์ (ประมวลผลด้วยซอฟต์แวร์ WEKA version 3.6.5 (Hall *et al.*, 2009)) แสดงได้ดังรูปที่ 1.1(a) และ 1.1(b) ตามลำดับ ความถูกต้องของการจำแนกวัดผลด้วยวิธี 10-fold cross validation ซึ่งเป็นการวัดผลแบบไขว้ของข้อมูลฝึกและข้อมูลทดสอบจำนวนสิบครั้ง และแสดงค่าความถูกต้องของการจำแนกในลักษณะตารางที่เรียกว่า confusion matrix ซึ่งเป็นการแจกแจงผลการทำนายเป็นรายคลาส ระหว่างคลาสของข้อมูลที่แท้จริงเทียบกับคลาสที่ได้จากการทำนายด้วยโมเดล เมื่อพิจารณาค่าในเมตริกซ์พบว่าการวิเคราะห์การถดถอยโลจิสติกทำนายการวินิจฉัยได้ถูกต้อง 19 รายจากทั้งหมด 24 ราย คิดเป็นร้อยละของความถูกต้องเท่ากับ 79.17% ในขณะที่การวิเคราะห์ดิสคริมิแนนต์ทำนายการวินิจฉัยได้ถูกต้อง 20 รายจากทั้งหมด 24 ราย คิดเป็นร้อยละของความถูกต้องเท่ากับ 83.33%

```

Classifier output
=== Classifier model (full training set) ===
Logistic Regression with ridge parameter of 0.0
Coefficients...
Variable                Class
                        yes
=====
age=young                12.9696
age=pre-presbyopic      -5.8375
age=presbyopic          -7.1321
spectacle-prescrip     -1.2946
astigmatism             -1.2946
tear-prod-rate          52.1745
Intercept               -43.7479

Odds Ratios...
Variable                Class
                        yes
=====
age=young                429147.8108
age=pre-presbyopic      0.0029
age=presbyopic          0.0008
spectacle-prescrip     0.274
astigmatism             0.274
tear-prod-rate          4.561605452373136E22

=== Confusion Matrix ===
  a  b  <-- classified as
  7  2 | a = yes
  3 12 | b = no

Classifier output
=== Classifier model (full training set) ===
Classification via Regression
Classifier for class with index 0:
Linear Regression Model
contact-lenses =
    0.125 * age=pre-presbyopic,young +
    0.125 * age=young +
    0.0833 * spectacle-prescrip=myope +
    0.0833 * astigmatism=no +
    0.75 * tear-prod-rate=normal +
   -0.2083

Classifier for class with index 1:
Linear Regression Model
contact-lenses =
    0.125 * age=pre-presbyopic,presbyopic +
    0.125 * age=presbyopic +
    0.0833 * spectacle-prescrip=hypermetrope +
    0.0833 * astigmatism=yes +
    0.75 * tear-prod-rate=reduced +
    0.0417

=== Confusion Matrix ===
  a  b  <-- classified as
  8  1 | a = yes
  3 12 | b = no

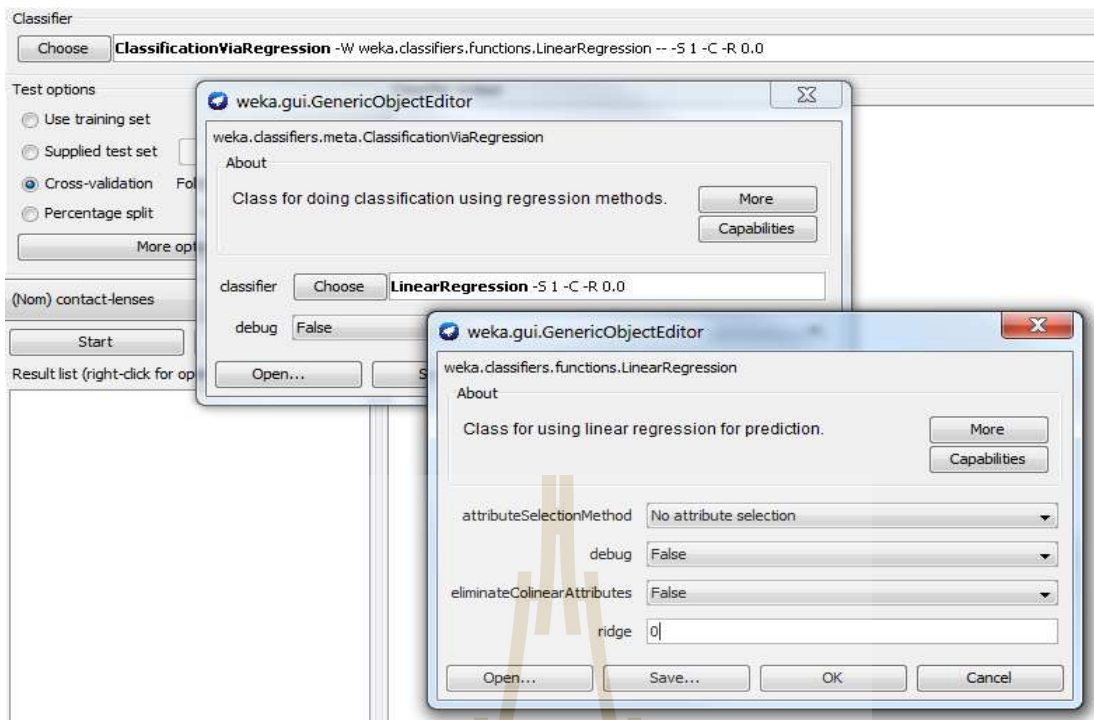
```

(a) Logistic regression

(b) Discriminant analysis

(meta.ClassificationViaRegression)

รูปที่ 1.1 เปรียบเทียบผลการวิเคราะห์ข้อมูลเพื่อวินิจฉัยความเหมาะสมในการใส่คอนแทกเลนส์ด้วยวิธี (a) การวิเคราะห์การถดถอยโลจิสติก และ (b) การวิเคราะห์ดิสคริมิแนนต์



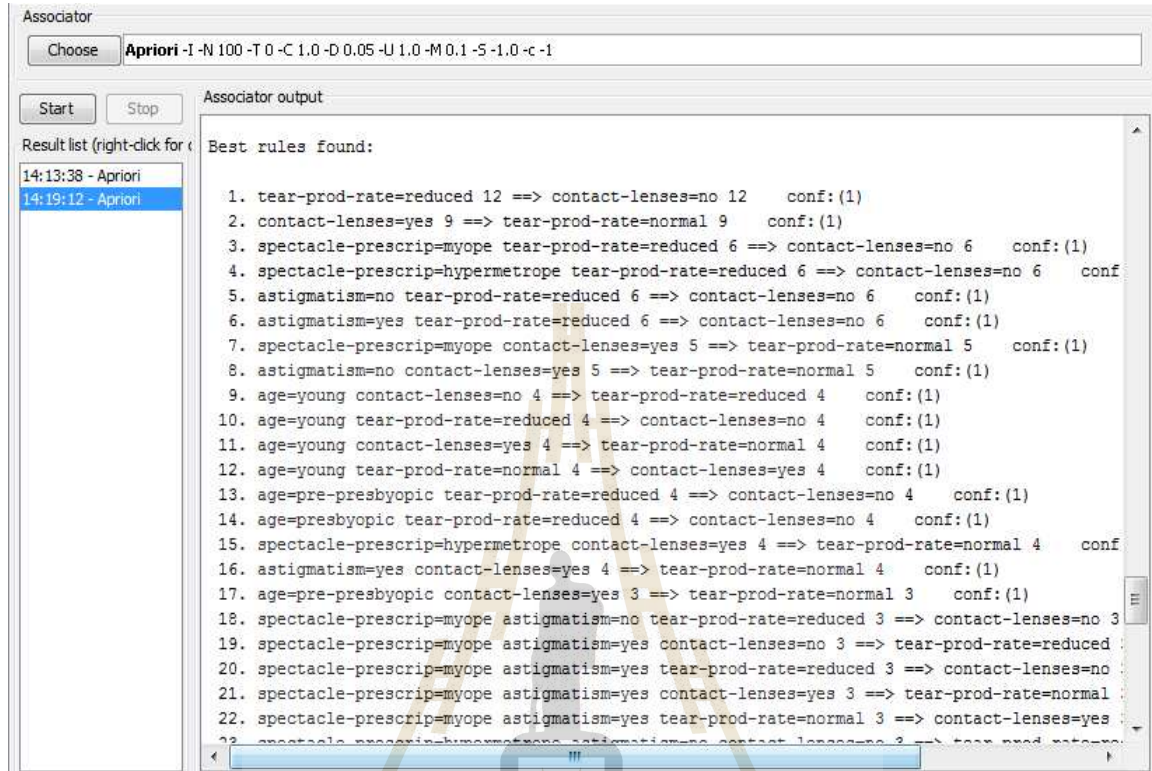
รูปที่ 1.2 การกำหนดค่าพารามิเตอร์ของอัลกอริทึม ClassificationViaRegression

ซอฟต์แวร์ WEKA ไม่มีอัลกอริทึมที่วิเคราะห์ดิสคริมิแนนต์ได้โดยตรง แต่สามารถใช้วิธีใกล้เคียงได้ด้วยการเลือกอัลกอริทึม ClassificationViaRegression จากกลุ่ม meta classifier และกำหนดพารามิเตอร์ในการเลือกอัลกอริทึมและกำหนดการเลือกแอททริบิวต์ตามรูปที่ 1.2 จากผลการวิเคราะห์ข้อมูลคอนแทกเลนส์ด้วยเทคนิคการวิเคราะห์การถดถอยโลจิสติก และการวิเคราะห์ดิสคริมิแนนต์ พบว่าโมเดลในลักษณะสมการเชิงเส้นสามารถใช้ทำนายข้อมูลใหม่ได้ว่า ผู้มีปัญหาด้านสายตาสมควรจะใช้คอนแทกเลนส์หรือไม่ แต่ลักษณะของโมเดลที่เป็นตัวแบบทางคณิตศาสตร์ไม่เอื้อต่อการอธิบายลักษณะการจำแนกในเชิงพรรณนาและอธิบายเหตุผล (reasoning) ของการทำนายได้ยาก

ข้อด้อยในด้านการไม่สามารถอธิบายเหตุผลของโมเดล เป็นปัญหาสำคัญที่ทำให้เทคนิคการวิเคราะห์ดิสคริมิแนนต์ด้วยหลักการสถิตินี้ไม่เหมาะสมที่จะประยุกต์ใช้ในกลไกอนุमानข้อแนะนำจากฐานความรู้ของระบบสนับสนุนการตัดสินใจด้านการแพทย์ งานวิจัยนี้จึงพิจารณาที่จะใช้เทคนิคการวิเคราะห์ความสัมพันธ์ในลักษณะของการทำเหมืองข้อมูลประเภทการค้นหารูปแบบที่ปรากฏบ่อยที่จะนำไปสู่การสังเคราะห์ความสัมพันธ์ของตัวแปรในรายการข้อมูล

การทำเหมืองข้อมูลประเภทการค้นหาคำความสัมพันธ์ (association mining) มีความสามารถในการค้นหาคำสัมพันธ์ระหว่างตัวแปร ที่หลากหลายมากกว่าเทคนิคการทำเหมืองจำแนกประเภท (classification mining) เทคนิคการค้นหาคำสัมพันธ์จึงมีความเหมาะสมมากกว่าในการประยุกต์ใช้ในงานการวิเคราะห์ดิสคริมิแนนต์ แต่การใช้เทคนิคการทำเหมืองความสัมพันธ์โดยตรงด้วยอัลกอริทึมที่เป็นที่นิยมทั่วไป เช่น Apriori (Agrawal & Srikant, 1994) จะให้ผลลัพธ์เป็น

กฎความสัมพันธ์ในปริมาณที่มากเกินไป และหลายกฎมีความซ้ำซ้อน ดังแสดงตัวอย่างผลการทำงานของอัลกอริทึม Apriori กับข้อมูลคอนแทกเลนส์ได้ดังรูปที่ 1.3



รูปที่ 1.3 แสดงกฎความสัมพันธ์ส่วนหนึ่งของข้อมูลคอนแทกเลนส์ จากกฎทั้งหมด 64 กฎที่ตรงตามเกณฑ์ $\text{support} \geq 0.1$ และ $\text{confidence} = 1.0$

ในงานด้านการแพทย์ที่มีการบันทึกข้อมูลประวัติผู้ป่วย ประวัติการรักษา รวมถึงผลการตรวจจากห้องปฏิบัติการในลักษณะของข้อมูลที่เป็นทั้งค่าต่อเนื่องและค่าไม่ต่อเนื่อง ทำให้การวิเคราะห์ดิสคริมิแนนต์เป็นเทคนิคที่สามารถนำมาช่วยวิเคราะห์ข้อมูลผู้ป่วย เพื่อสร้างโมเดลหรือตัวแบบข้อมูลสำหรับเก็บไว้ในฐานความรู้ของระบบสนับสนุนการตัดสินใจด้านการแพทย์ เพื่อใช้ระบบสนับสนุนนี้ให้เป็นประโยชน์ในการช่วยวินิจฉัยผู้ป่วยในอนาคต แต่จากปัญหาการตีความและอธิบายเหตุผลได้ยากของโมเดลที่ได้จากเทคนิคการวิเคราะห์ดิสคริมิแนนต์ด้วยหลักการทางสถิติ ทำให้ผู้วิจัยมีแนวคิดที่จะพัฒนาเทคนิคการวิเคราะห์ข้อมูลด้วยหลักการของการทำเหมืองความสัมพันธ์ที่สามารถอธิบายเหตุผลหรือที่มาของคำวินิจฉัยได้ แต่การประยุกต์ใช้อัลกอริทึมเพื่อการทำเหมืองความสัมพันธ์ที่มีอยู่ในปัจจุบันโดยตรงไม่เหมาะสมกับงานวิเคราะห์ข้อมูลเพื่อสร้างฐานความรู้ด้านการแพทย์ ทั้งนี้เนื่องจากอัลกอริทึมที่มีอยู่ใช้อยู่ในปัจจุบันจะค้นหาความสัมพันธ์ที่มากเกินไปจนความจำเป็น นอกจากนี้หลายกฎความสัมพันธ์ที่ได้มีลักษณะที่ซ้ำซ้อนกัน

โครงการวิจัยนี้จึงมีจุดมุ่งหมายในการคิดค้นและพัฒนาอัลกอริทึมใหม่ที่สามารถค้นหา รูปแบบความสัมพันธ์ที่เด่นชัดจากข้อมูล โดยมีข้อกำหนดที่จำนวนกฎความสัมพันธ์ที่ได้จะต้องมี ปริมาณเหมาะสม และกฎที่ได้จะต้องมีความถูกต้องสูงเพียงพอที่จะบันทึกไว้ในฐานความรู้ของระบบ สนับสนุนการตัดสินใจด้านการแพทย์

1.2 วัตถุประสงค์ของโครงการวิจัย

- ออกแบบและพัฒนาอัลกอริทึมเพื่อการวิเคราะห์ตัดสินใจแนบด้วยหลักการการทำ เหมือนความสัมพันธ์ ออกแบบฐานความรู้ กลไกการวิเคราะห์และจัดการความรู้ ส่วนจัดการแบบจำลองการตัดสินใจและส่วนอธิบายเหตุผลของระบบสนับสนุนการ ตัดสินใจด้านการแพทย์
- พัฒนาโปรแกรมการทำเหมือนความสัมพันธ์เพื่อการวิเคราะห์ตัดสินใจแนบกับข้อมูล ด้านการแพทย์เพื่อเป็นต้นแบบสำหรับระบบสนับสนุนการตัดสินใจด้านการแพทย์

1.3 ขอบเขตของการวิจัย

ข้อมูลที่ใช้ในการทดสอบอัลกอริทึม เป็นชุดข้อมูลมาตรฐานจาก UCI Machine Learning Repository (<http://www.ics.uci.edu/~mlearn/>) ที่ประกอบด้วย training data และ test data การตรวจสอบความถูกต้องของอัลกอริทึมวิเคราะห์ตัดสินใจแนบ และการทดสอบความ ถูกต้องของระบบสนับสนุนการตัดสินใจด้านการแพทย์ จะใช้การเทียบผลกับ test data

1.4 ประโยชน์ที่ได้รับ

งานวิจัยนี้เป็นการออกแบบและพัฒนาขั้นตอนวิธีการ เพื่อให้ได้องค์ความรู้ใหม่ในด้านการ ทำเหมือนความสัมพันธ์เพื่อการวิเคราะห์ตัดสินใจแนบ และมีการออกแบบโครงสร้างของระบบ สนับสนุนการตัดสินใจด้านการแพทย์ที่สามารถผนวกฐานความรู้จากโมเดลตัดสินใจแนบ ประโยชน์ที่ ได้รับโดยตรงคือเทคนิคและอัลกอริทึมใหม่ ที่สามารถตีพิมพ์ผลงานวิจัยในวารสารวิชาการได้จำนวน 1 บทความ และผลงานวิจัยที่นำเสนอและตีพิมพ์ใน International Conference Proceedings อีก จำนวน 3 บทความ

การทดสอบระบบที่ออกแบบและพัฒนาขึ้นในลักษณะ prototype ทำให้ได้โปรแกรม ต้นแบบที่สามารถจดลิขสิทธิ์ได้จำนวน 1 โปรแกรม ได้แก่ โปรแกรมจำแนกประเภทข้อมูลด้วยกฎ ความสัมพันธ์แบบคลุมเครือที่กะทัดรัด (data classification program with compact fuzzy association rules) นอกจากนี้ผู้ช่วยวิจัยที่เป็นนักศึกษาระดับปริญญาโทและปริญญาเอก ได้มีโอกาส มีส่วนร่วมในโครงการวิจัยนี้เพื่อพัฒนาความสามารถและทักษะในการทำงานวิจัยระดับสูง

บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

2.1 การวิเคราะห์ดิสคริมิแนนต์

การวิเคราะห์ดิสคริมิแนนต์ในลักษณะของการวิเคราะห์พหุตัวแปร (multivariate analysis) เพื่อพิจารณาน้ำหนักที่เหมาะสมของตัวแปรในกลุ่มหรือคลาสเดียวกัน โดยที่เมื่อนำตัวแปรเหล่านั้นมาพิจารณาประกอบกันแล้วจะสามารถทำนายค่าตัวแปรเป้าหมาย (หรือค่าของคลาสข้อมูล) ได้อย่างถูกต้องมากที่สุด การวิเคราะห์แบบนี้เป็นเทคนิคทางสถิติเทคนิคหนึ่งที่ยิยมใช้ในการทำเหมืองข้อมูลประเภทการค้นหาโมเดลเพื่อจำแนกประเภท (classification) ด้วยวัตถุประสงค์ของการใช้โมเดลทำนายคลาสของข้อมูลใหม่ (Fernandez, 2002) จากตัวอย่างโมเดลดิสคริมิแนนต์ในรูปที่ 2.1 ด้านซ้ายมือ สามารถแสดงเป็นฟังก์ชันหรือสมการเชิงเส้นได้ตั้งด้านขวามือของรูปที่ 2.1

<pre>Classifier output === Classifier model (full training set) === Classification via Regression Classifier for class with index 0: Linear Regression Model contact-lenses = 0.125 * age=pre-presbyopic,young + 0.125 * age=young + 0.0833 * spectacle-prescrip=myope + 0.0833 * astigmatism=no + 0.75 * tear-prod-rate=normal + -0.2083 Classifier for class with index 1: Linear Regression Model contact-lenses = 0.125 * age=pre-presbyopic,presbyopic + 0.125 * age=presbyopic + 0.0833 * spectacle-prescrip=hypermetrope + 0.0833 * astigmatism=yes + 0.75 * tear-prod-rate=reduced + 0.0417</pre>	<pre>contact-lenses (yes) = 0.125 * age=pre-presbyopic,young + 0.125 * age=young + 0.0833 * spectacle-prescrip=myope + 0.0833 * astigmatism=no + 0.75 * tear-prod-rate=normal + (- 0.2083) contact-lenses (no) = 0.125 * age=pre-presbyopic,presbyopic + 0.125 * age=presbyopic + 0.0833 * spectacle-prescrip=hypermetrope + 0.0833 * astigmatism=yes + 0.75 * tear-prod-rate=reduced + 0.0417</pre>
---	--

รูปที่ 2.1 โมเดลที่ได้จากการวิเคราะห์ดิสคริมิแนนต์เพื่อวินิจฉัยการใส่คอนแทกเลนส์

เมื่อต้องการใช้โมเดลดิสคริมิแนนต์ที่ได้ตามรูปที่ 2.1 ช่วยในการวินิจฉัยความเหมาะสมในการใส่คอนแทกเลนส์ของคนไข้รายใหม่ จะต้องคำนวณผลลัพธ์ของทั้งฟังก์ชันที่ให้ผลวินิจฉัยเป็น

yes และฟังก์ชันที่ให้ผลเป็น no เพื่อทำนายผลตามค่าที่สูงที่สุด ตัวอย่างเช่น ถ้าผู้มีปัญหาทางด้านสายตา รายใหม่มีลักษณะดังตารางที่ 2.1

ตารางที่ 2.1 รายละเอียดข้อมูลของคนใช้รายใหม่

Age of the patient	Spectacle prescription	Astigmatic	Tear production rate	Wear contact lenses
young	myope	no	normal	?

การคำนวณด้วยโมเดลตัดสินใจแบบต้นไม้ตามสมการในรูปที่ 2.1 ทั้งในกรณีผลวินิจฉัยเป็น yes และ no จะให้ผลลัพธ์ดังต่อไปนี้

$$\begin{aligned} \text{contact-lenses (yes)} &= 0.125 * (1) + 0.125 * (1) + 0.0833 * (1) + 0.0833 * (1) \\ &\quad + 0.75 * (1) - 0.2083 \\ &= 0.9583 \end{aligned}$$

$$\begin{aligned} \text{contact-lenses (no)} &= 0.125 * (0) + 0.125 * (0) + 0.0833 * (0) + 0.0833 * (0) \\ &\quad + 0.75 * (0) + 0.0417 \\ &= 0.0417 \end{aligned}$$

จากผลการคำนวณ เมื่อเปรียบเทียบค่าแล้ว 0.9583 มีค่าสูงกว่า 0.0417 ดังนั้นโมเดลตัดสินใจจะทำนายว่าผู้มีปัญหาทางด้านสายตารายใหม่นี้สามารถใส่คอนแทกเลนส์ได้

2.2 งานวิจัยที่เกี่ยวข้อง

ลักษณะของการสร้างโมเดลตัดสินใจแบบต้นไม้ (ดังแสดงด้วยตัวอย่างข้อมูลคอนแทกเลนส์) จะสังเกตได้ว่าเมื่อตัวแปรอิสระเป็นค่าแจกแจงได้ ฟังก์ชันตัดสินใจแบบต้นไม้ที่ได้จะมีจำนวนเทอมมาก และฟังก์ชันจะยาวมากขึ้นเมื่อเพิ่มจำนวนตัวแปรอิสระในชุดข้อมูลฝึก (training data) การที่ฟังก์ชันประกอบขึ้นจากเทอมต่าง ๆ ในจำนวนมาก จะเพิ่มเวลาการประมวลผลของเครื่องคอมพิวเตอร์ และอาจจะทำให้ได้โมเดลที่มีลักษณะเจาะจงกับชุดข้อมูลฝึกมากเกินไป (overfitting) นักวิจัยหลายคนจึงเสนอแนวทางลดความซับซ้อนของฟังก์ชันตัดสินใจแบบต้นไม้เพื่อหลีกเลี่ยงปัญหาโมเดลที่เจาะจงเกินไป

แนวทางลดความซับซ้อนของโมเดลตัดสินใจแบบต้นไม้ที่เห็นได้ชัดเจนมากที่สุด คือการลดจำนวนตัวแปรอิสระ ให้เหลือเฉพาะตัวแปรที่มีอำนาจจำแนกสูงเท่านั้น ในทางสถิตินิยมใช้เทคนิคการ

วิเคราะห์องค์ประกอบหลัก (principal component analysis -- PCA) การวิเคราะห์ปัจจัย (factor analysis) และการแปลงแกนหรือมิติข้อมูลด้วยหลักการแปลงทางคณิตศาสตร์ มาช่วยค้นหาและคัดเลือกเฉพาะตัวแปรสำคัญที่ควรนำมาใช้ในการสร้างโมเดลดิสคริมิแนนต์ งานวิจัยในแนวทางนี้ประกอบด้วย งานของ Belhumeur และคณะ (1997) ที่ใช้เทคนิคการวิเคราะห์องค์ประกอบหลัก งานของ Zhang และ Jia (2007) ใช้การแปลงมิติข้อมูลเพื่อลดความซับซ้อนของโมเดล งานของ Lin และ Chen (2009) ใช้แนวทางปัญญาประดิษฐ์มาผนวกกับการวิเคราะห์ดิสคริมิแนนต์ ในขณะที่งานของ Howland และ Park (2004) ใช้วิธีการปรับปรุงการคำนวณดิสคริมิแนนต์โดยตรง ให้สามารถทำงานกับชุดข้อมูลฝึกที่รายการข้อมูลมีปริมาณน้อยแต่มีจำนวนมิติหรือตัวแปรมาก

การประยุกต์ใช้เทคนิคการวิเคราะห์ดิสคริมิแนนต์ด้านการแพทย์และวิทยาศาสตร์สุขภาพ มีได้หลากหลายลักษณะ เช่น ในด้านการวินิจฉัยโรค มีการใช้ฟังก์ชันดิสคริมิแนนต์เพื่อลดมิติข้อมูลในระบบผู้เชี่ยวชาญเพื่อวินิจฉัยโรคที่เกี่ยวข้องกับลิ้นหัวใจ (Sengur, 2008) ทีมนักวิจัยจากบราซิล (Silva *et al.*, 2009) ได้ใช้เทคนิคการวิเคราะห์ดิสคริมิแนนต์เพื่อจำแนกผลการตรวจเนื้อเยื่อเต้านมว่าเป็นเนื้อเยื่อปกติ เนื้อเยื่ออก หรือเนื้อร้าย ทีมวิจัยของ Dogantekin (2009; 2010; 2011) ได้ประยุกต์ใช้เทคนิคดิสคริมิแนนต์ในขั้นตอนเตรียมข้อมูลเพื่อคัดเลือกเฉพาะตัวแปรสำคัญ ที่เหมาะสมจะนำไปสร้างโมเดลจำแนกประเภทในขั้นต่อไปด้วยเทคนิคนิวรอลเน็ตเวิร์ค เพื่อใช้ในระบบวินิจฉัยโรคเบาหวานและโรคไตเรื้อรัง

ในการศึกษาเกี่ยวกับยีนส์และโครโมโซม ในระยะแรกมีการประยุกต์ใช้เทคนิควิเคราะห์ดิสคริมิแนนต์ร่วมกับทฤษฎีความน่าจะเป็นแบบเบย์เพื่อจำแนกโครโมโซมมนุษย์ (Ledley *et al.*, 1980) ต่อมากลุ่มวิจัยของ Umene และคณะ (2007) ได้ใช้เทคนิคการวิเคราะห์ดิสคริมิแนนต์กับข้อมูลดีเอ็นเอเพื่อยืนยันสมมุติฐานเกี่ยวกับการเกิดใหม่และการเกิดซ้ำของเชื้อไวรัสโรคเรื้อรัง กลุ่มนักวิจัยจากประเทศจีนและเกาหลี (Li *et al.*, 2010) ได้เสนอแนวคิดการจำแนกรหัสพันธุกรรมโดยประยุกต์ใช้เทคนิควิเคราะห์ดิสคริมิแนนต์เพื่อลดจำนวนมิติข้อมูลก่อนที่จะเข้าสู่ขั้นตอนการจำแนก ทีมวิจัยนำโดย Cao และคณะ (2011) ได้เสนอปรับปรุงเทคนิคดิสคริมิแนนต์ของฟิชเชอร์ผนวกเข้ากับเทคนิคต้นไม้ตัดสินใจ เพื่อการวิเคราะห์ข้อมูลเมตาโบลีซึม ปีต่อมา Huang และคณะ (2012) ได้พัฒนาเทคนิคจากหลักการดิสคริมิแนนต์ของฟิชเชอร์เช่นเดียวกันแต่ใช้เพื่อวัตถุประสงค์การวิเคราะห์รหัสพันธุกรรม และในงานวิจัยล่าสุด Xu และคณะ (2012) ได้ใช้เทคนิคดิสคริมิแนนต์ร่วมกับการวิเคราะห์องค์ประกอบหลักเพื่อค้นหาไปโอมาร์คเกอร์หรือตัวบ่งชี้ทางชีวภาพสำหรับการทำนายเบาหวานชนิดที่สอง

จากการทบทวนวรรณกรรมที่เกี่ยวข้อง พบว่างานวิจัยที่เกี่ยวข้องกับการวิเคราะห์ดิสคริมิแนนต์ โดยเฉพาะในงานด้านการแพทย์และวิทยาศาสตร์สุขภาพ กลุ่มนักวิจัยจะเสนอแนวคิดเป็นสองแนวทาง แนวทางแรกคือการใช้เทคนิคการวิเคราะห์ดิสคริมิแนนต์ในขั้นตอนแรกของการ

วิเคราะห์ข้อมูลซึ่งเป็นขั้นเตรียมข้อมูล เพื่อลดมิติหรือจำนวนตัวแปรในข้อมูลก่อนที่จะส่งข้อมูลที่ผ่านการเตรียมแล้วไปยังอัลกอริทึมเพื่อการจำแนก เช่น นิวรอลเน็ตเวิร์ก ซัพพอร์ตเวกเตอร์แมชชีน แนวทางที่สองคือใช้เทคนิคการวิเคราะห์ดิสคริมิแนนต์เป็นอัลกอริทึมเพื่อการจำแนกโดยตรง แต่ปรับปรุงข้อดีด้วยการเสริมเทคนิคอื่น เช่น การวิเคราะห์องค์ประกอบหลัก

ทั้งสองแนวทางที่กล่าวมานั้นต่างก็ให้ผลลัพธ์เป็นโมเดลเพื่อการจำแนกประเภท เป็นฟังก์ชันคณิตศาสตร์ในลักษณะของกล่องดำ (black box) ที่ไม่สามารถอธิบายได้อย่างชัดเจนว่าโมเดลนั้นให้ผลการจำแนกโดยพิจารณาจากองค์ประกอบใด ด้วยสาเหตุใด หรือใช้เหตุผลใดในการจำแนกแต่ละข้อมูล ประเด็นของการไม่สามารถแสดงเหตุผลหรือขาดความสามารถในการแสดงคำอธิบาย ทำให้เป็นข้อดีของเทคนิคการวิเคราะห์ดิสคริมิแนนต์ด้วยวิธีการทางสถิติ ที่จะนำมาใช้ในระบบสนับสนุนการตัดสินใจ ดังนั้นโครงการวิจัยนี้จึงได้ถูกเสนอขึ้นด้วยวัตถุประสงค์ที่จะพัฒนาเทคนิคในการวิเคราะห์ดิสคริมิแนนต์ ด้วยหลักการของการทำเหมืองความสัมพันธ์ เพื่อที่จะให้ผลลัพธ์เป็นโมเดลที่เข้าใจได้ง่ายและสามารถอธิบายได้

บทที่ 3

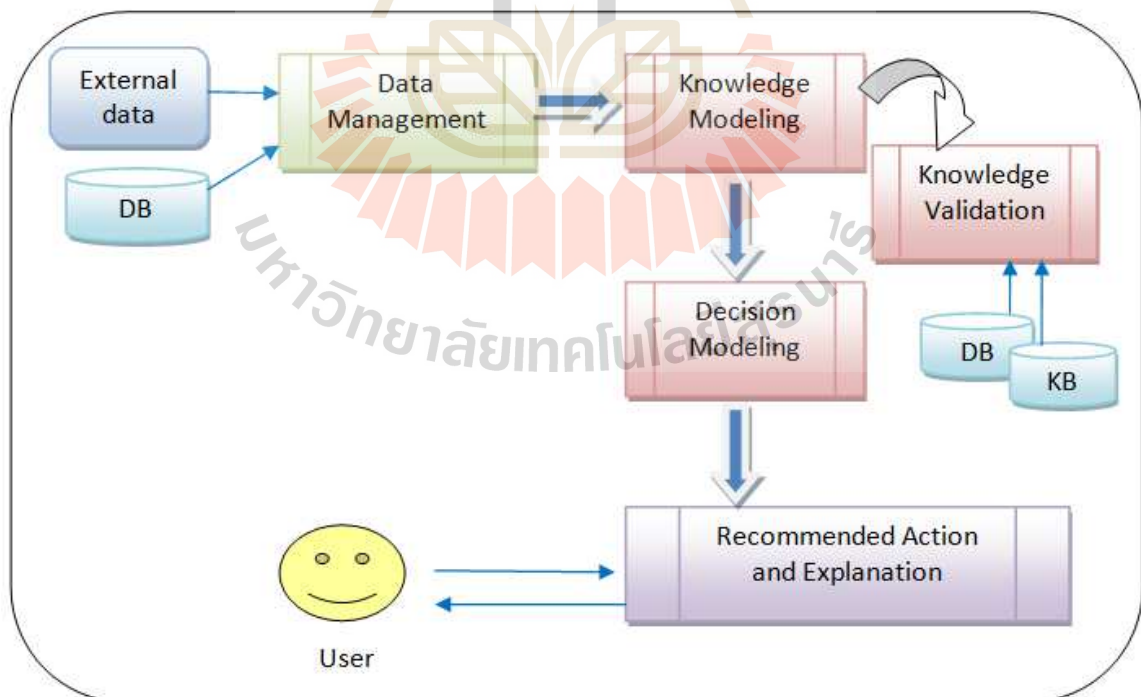
การออกแบบและพัฒนาโปรแกรม

3.1 กรอบแนวคิดของงานวิจัย

กรอบแนวคิดหลักของงานวิจัยนี้ คือ การพัฒนาเทคนิคการวิเคราะห์หัตถศริมิแนนต์เพื่อใช้ในระบบสนับสนุนการตัดสินใจด้านการแพทย์ โครงสร้างพื้นฐานของระบบสนับสนุนการตัดสินใจ โดยทั่วไปจะประกอบด้วย 3 ส่วนประกอบหลัก คือ

- (1) ส่วนของฐานข้อมูล และ/หรือ ฐานความรู้
- (2) ส่วนติดต่อกับผู้ใช้
- (3) ส่วนสร้างและจัดการแบบจำลองการตัดสินใจ

ระบบสนับสนุนการตัดสินใจด้านการแพทย์ของโครงการวิจัยนี้ ได้เพิ่มเติมความสามารถด้านการทำเหมืองข้อมูล เข้ามาช่วยในส่วนของการสร้างและจัดการแบบจำลอง โครงสร้างของระบบโดยรวม แสดงได้ดังรูปที่ 3.1



รูปที่ 3.1 กรอบของงานพัฒนาระบบสนับสนุนการตัดสินใจด้านการแพทย์

แต่ละส่วนประกอบของระบบอธิบายได้ดังนี้

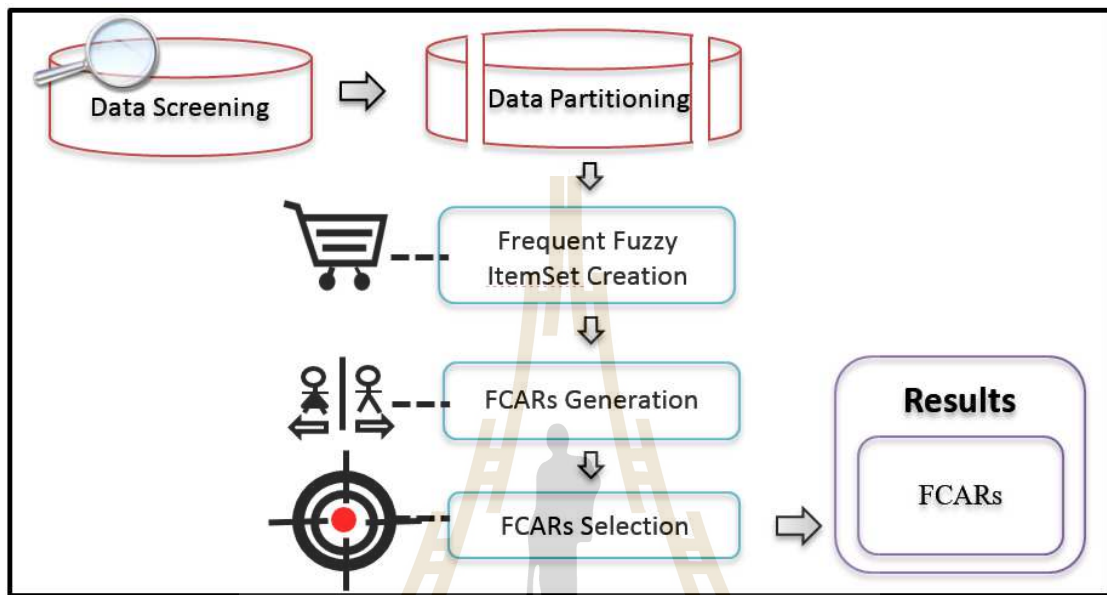
- โมดูล Data Management ทำหน้าที่รวบรวม คัดแยก ปรับปรุง และกลั่นกรองข้อมูล เพื่อนำไปใช้ในการวิเคราะห์โมเดลประกอบการตัดสินใจ
- โมดูล Knowledge Modeling เป็นส่วนประกอบหลักของโครงการวิจัยนี้ ทำหน้าที่วิเคราะห์ดิสคริมิแนนต์เพื่อสร้างโมเดลเพื่อการวินิจฉัย และโมเดลเพื่อการอธิบายลักษณะเด่นของข้อมูลผู้ป่วยในแต่ละกลุ่ม โมเดลที่ได้จะต้องผ่านการตรวจสอบยืนยันความถูกต้องด้วยโมดูล Knowledge Validation ในการพัฒนาระบบต้นแบบของโครงการวิจัยนี้จะใช้การเปรียบเทียบกับข้อมูลทดสอบ เป็นเครื่องมือในการยืนยันความถูกต้องของโมเดล
- โมดูล Decision Modeling ทำหน้าที่จัดรูปแบบของความรู้ที่ได้จากโมดูล Knowledge Modeling ให้อยู่ในรูปแบบที่เหมาะสมสำหรับการสื่อสารและแสดงคำแนะนำต่อผู้ใช้
- โมดูล Recommended Action and Explanation เป็นส่วนประกอบที่รวมส่วนติดต่อกับผู้ใช้ ทำหน้าที่รับคำถามจากผู้ใช้ แสดงข้อแนะนำ รวมถึงแสดงคำอธิบายประกอบกับข้อแนะนำในกรณีที่ผู้ใช้มีข้อสงสัย

โมดูลหลักที่เป็นองค์ความรู้ใหม่ของโครงการวิจัยนี้ได้แก่ โมดูล Knowledge Modeling ทำหน้าที่วิเคราะห์ดิสคริมิแนนต์เพื่อสร้างโมเดลเพื่อการวินิจฉัย และโมเดลเพื่อการอธิบายลักษณะเด่นของข้อมูลผู้ป่วยในแต่ละกลุ่ม การพัฒนาอัลกอริทึมของโมดูลนี้ใช้วิธีการสร้างกฎการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือด้วยแนวคิดของฟัซซีเซต ข้อกำหนดสำคัญของอัลกอริทึมคือจะต้องสร้างเฉพาะกฎที่สำคัญและมีจำนวนไม่มาก นอกจากนี้กฎที่ได้จะต้องสามารถตีความได้ง่ายและมีประสิทธิภาพในการจำแนกประเภทข้อมูลได้ด้วยความถูกต้องสูง

3.2 การออกแบบโมดูลสร้างโมเดลเพื่อการวินิจฉัย (Knowledge Modeling)

- อัลกอริทึมในส่วน Knowledge Modeling ประกอบด้วยขั้นตอนการทำงาน 5 ส่วน คือ
- (1) การตรวจสอบข้อมูลก่อนการประมวลผล
 - (2) การแบ่งแยกข้อมูล
 - (3) การสร้างไอเท็มเซตปรากฏบ่อยแบบคลุมเครือ
 - (4) การสร้างกฎการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือ (Fuzzy Classification with Association Rules -- FCARs)
 - (5) การเลือกกฎ FCARs เพื่อนำไปใช้ในการวินิจฉัยข้อมูลใหม่

โดยที่ส่วนที่ (2) ถึงส่วนที่ (4) ได้ใช้อัลกอริทึมของ Pach และคณะ (2008) เป็นพื้นฐานในการพัฒนา ขั้นตอนทั้งห้าขั้นตอนแสดงเป็นแผนภาพได้ดังรูปที่ 3.2 อัลกอริทึมที่พัฒนาขึ้นนี้ให้ชื่อว่า อัลกอริทึมจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือที่กะทัดรัด (Classification with Compact Fuzzy Association Rules -- CCFAR)



รูปที่ 3.2 ขั้นตอนการทำงานของอัลกอริทึม CCFAR

3.3 อัลกอริทึมจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือที่กะทัดรัด

การทำงานของอัลกอริทึม CCFAR ตามขั้นตอนหลักที่แสดงดังรูปที่ 3.2 สามารถอธิบายรายละเอียดการทำงานในแต่ละขั้นด้วยข้อมูลตัวอย่างตามตารางที่ 3.1

ตารางที่ 3.1 ข้อมูลตัวอย่างประกอบการอธิบายอัลกอริทึม CCFAR

Id	Age	Health Expense	Health Insurance	Class
1	18	10000	4000	2
2	20	18000	20000	1
3	19	17000	5000	2
4	24	20000	10000	1
5	25	22000	9000	1

ขั้นตอนตรวจสอบข้อมูลก่อนการประมวลผล (Data Screening)

ในกรณีที่ข้อมูลเข้าเป็นข้อมูลตัวเลข ขั้นตอนนี้จะแปลงข้อมูลให้อยู่ในรูปแบบของฟัซซีเซต โดยจะแปลงข้อมูลตัวเลขให้เป็นระดับของความเป็นสมาชิกของเซตนั้น ๆ และการแปลงเป็นระดับความเป็นสมาชิกจะพิจารณาจากแอททริบิวต์เป้าหมาย (Class Target) โดยตามตัวอย่างจะเป็นข้อมูลคอลัมน์สุดท้ายของตาราง ในการประมวลผลจะใช้ข้อมูล 4 แอททริบิวต์ คือ Age, Income, Balance และ Class ซึ่งเป็นแอททริบิวต์เป้าหมาย แต่จะไม่ใช้แอททริบิวต์ Id เนื่องจากเป็นเพียงข้อมูลที่บ่งบอกถึงหมายเลขแถวเท่านั้น

ขั้นตอนการแบ่งแยกข้อมูล (Data Partitioning)

การแบ่งแยกข้อมูลเป็นการทำให้ข้อมูลที่น่าเข้ามาประมวลผลนั้น มีลักษณะเป็นข้อมูลแบบฟัซซีเซต โดยงานวิจัยนี้ได้เลือกอัลกอริทึม fuzzy c-means (Bezdek et al., 1984) มาใช้เพื่อการแบ่งกลุ่มหรือแบ่งแยกข้อมูลให้มีลักษณะเป็นระดับของความเป็นสมาชิกในเซตต่าง ๆ ซึ่งจะช่วยแก้ปัญหาข้อมูลที่มีลักษณะเป็นตัวเลขค่าต่อเนื่องได้ ข้อมูลในแอททริบิวต์ Age, Income, Balance เมื่อกำหนดการแบ่งกลุ่มเป็น 3 ระดับคือ Low, Medium และ High จะได้ช่วงของข้อมูลดังตารางที่ 3.2 ข้อมูลแต่ละรายการมีช่วงของค่าดังนี้

ข้อมูลรายการที่ 1 มีช่วงของข้อมูลเป็น Age = Low, Exps = Low, Insur = Low

ข้อมูลรายการที่ 2 มีช่วงของข้อมูลเป็น Age = Medium, Exps = Medium, Insur = High

ข้อมูลรายการที่ 3 มีช่วงของข้อมูลเป็น Age = Low, Exps = Medium, Insur = Low

ข้อมูลรายการที่ 4 มีช่วงของข้อมูลเป็น Age = High, Exps = High, Insur = Medium

ข้อมูลรายการที่ 5 มีช่วงของข้อมูลเป็น Age = High, Exps = High, Insur = Medium

ตารางที่ 3.2 แอททริบิวต์ Age, Income, Balance และค่าความเป็นสมาชิกในแต่ละช่วงข้อมูล

Id	Age			Health Expense (Exps)			Health Insurance (Insur)		
	Low	Medium	High	Low	Medium	High	Low	Medium	High
1	0.9863	0.0127	0.0010	1	0	0	0.9909	0.0081	0.0010
2	0.0119	0.9862	0.0019	0.0031	0.9769	0.0199	0	0	1
3	0.4996	0.4900	0.0104	0.0060	0.9773	0.0167	0.9866	0.0123	0.0011
4	0.0074	0.0141	0.9785	0.0111	0.1849	0.8040	0.0081	0.9894	0.0025
5	0.0053	0.0090	0.9857	0.0045	0.0329	0.9626	0.0122	0.9857	0.002

ขั้นตอนการสร้างไอเท็มเซตปรากฏบ่อยแบบคลุมเครือ (Frequent Fuzzy Itemset Creation)

ขั้นตอนนี้ประกอบด้วย 2 ขั้นตอนย่อย คือ การหาค่าสนับสนุนแบบคลุมเครือขั้นต่ำ และการค้นหาไอเท็มเซตปรากฏบ่อยแบบคลุมเครือ โดยไอเท็มเซตปรากฏบ่อยจะต้องมีค่าสนับสนุนแบบคลุมเครือมากกว่าค่าขั้นต่ำ จากตัวอย่างข้อมูลตามตารางที่ 3.1 เมื่อจำนวนข้อมูลทั้งหมด (N) คือ 5 และ min(Class) หมายถึงจำนวนครั้งที่ปรากฏในชุดข้อมูลของคลาสข้อมูลส่วนน้อย ในตัวอย่างตามตาราง 3.1 คลาสข้อมูลส่วนน้อยคือคลาส 2 ซึ่งปรากฏด้วยความถี่เป็น 2 วิธีคำนวณเกณฑ์ค่าสนับสนุนแบบคลุมเครือขั้นต่ำ (minimum fuzzy support -- γ) แสดงได้ดังสมการที่ 2-1 (Pach et al., 2008) จะได้ค่าสนับสนุนแบบคลุมเครือขั้นต่ำ คือ 0.2

$$\gamma = (\min(\text{Class})/N) / 2 = (2/5) / 2 = 0.2 \quad (3-1)$$

จากนั้นสร้างไอเท็มเซตแบบคลุมเครือที่มีค่าสนับสนุนแบบคลุมเครือสูงกว่าเกณฑ์ขั้นต่ำที่ 0.2 และเป็นเซตที่มีขนาด 1 ไอเท็ม 2 ไอเท็ม 3 ไอเท็ม ไปตามลำดับ วิธีคำนวณค่าสนับสนุนแบบคลุมเครือ (fuzzy support -- FS) แสดงได้ดังสมการที่ 3-2

$$FS = \sum(\text{Membership-of-itemsets}) / \text{Number-of-transactions} \quad (3-2)$$

ตัวอย่างการคำนวณค่า FS ของไอเท็มเซต Age = Low ตามระดับค่าความเป็นสมาชิกในตารางที่ 3.2 แสดงได้ดังนี้

$$\begin{aligned} FS(\text{Age}=\text{Low}) &= (0.9863 + 0.0119 + 0.4996 + 0.0074 + 0.0053) / 5 \\ &= 0.3021 \end{aligned}$$

ในกรณีไอเท็มเซตมีจำนวนไอเท็มมากกว่าหนึ่งไอเท็ม เช่น Age=Low & Exps=Low การคำนวณค่า Membership-of-itemsets จะเป็นผลคูณของค่าความเป็นสมาชิกของแต่ละไอเท็มในเซต ซึ่งตามตัวอย่างนี้ คือผลคูณของค่าความเป็นสมาชิกของไอเท็ม Age=Low และไอเท็ม Exps=Low ของแต่ละทรานแซคชัน แสดงตัวอย่างการคำนวณค่า FS ได้ดังนี้

$$\begin{aligned} FS(\text{Age}=\text{Low} \ \& \ \text{Exps}=\text{Low}) &= (0.9863*1) + (0.0119*0.0031) + \\ & \quad (0.4996*0.006) + (0.0074*0.0111) + \\ & \quad (0.0053*0.0045) / 5 \\ &= 0.1979 \end{aligned}$$

จำนวนไอเท็มเซตทั้งหมดตั้งแต่ขนาด 1 ไอเท็ม 2 ไอเท็ม 3 ไอเท็ม ไปตามลำดับ ที่ตรงตามเกณฑ์คัดเลือกค่าขั้นต่ำพร้อมทั้งค่าสนับสนุนแบบคลุมเครือสรุปได้ดังตารางที่ 3.3 ข้อมูลตามตัวอย่างนี้สามารถสร้างไอเท็มเซตได้สูงสุด 3 ไอเท็ม

ตารางที่ 3.3 ไอเท็มเซตแบบคลุมเครือที่ปรากฏบ่อยสูงกว่าเกณฑ์ขั้นต่ำ

1-Itemsets	FS	2-Itemsets	FS	3-Itemsets	FS
Age=low	0.3021	Age=medium & Exps=medium	0.2891	Age=high & Exps=high & Insur=medium	0.3427
Age=medium	0.3024	Age=high & Exps=high	0.3472		
Age=high	0.3955	Age=low & Insur=low	0.2941		
Exps=low	0.2050	Age=high & Insur=medium	0.3880		
Exps=medium	0.4344	Exps=high & Insur=medium	0.3489		
Exps=high	0.3606				
Insur=low	0.3996				
Insur=medium	0.3991				
Insur=high	0.2013				

ขั้นตอนการสร้างกฎการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือ (FCARs Generation)

FCARs หรือกฎการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือ (Fuzzy Classification with Association Rules) จะถูกสร้างจากไอเท็มเซตแบบคลุมเครือที่ปรากฏบ่อยสูงกว่าเกณฑ์ขั้นต่ำ กฎที่สร้างขึ้นจะมีค่าของคลาสเป็นส่วนเป้าหมายที่ปรากฏด้านขวามือของกฎ การกำหนดค่าของคลาสเป้าหมายจะอ้างอิงตามข้อมูลตั้งต้น โดยใช้ค่าของคลาสตามปรากฏการณ์ในรายการข้อมูลเป็นส่วนใหญ่ นอกจากนี้กฎที่ได้จะต้องมีการกำหนดคะแนนของกฎ คะแนนที่สูงจะหมายถึงกฎที่น่าเชื่อถือ การคำนวณคะแนน (Score) จะใช้วิธีการตามสมการที่ 3-3

$$\text{Score (rule: a} \rightarrow \text{b)} = \text{FC} \times \text{FCORR} \times \text{Firing_strength} \quad (3-3)$$

โดย

FC หมายถึง fuzzy confidence ของกฎ คำนวณจากค่า fuzzy support ของ a & b หารด้วยค่า fuzzy support ของ a

FCORR หมายถึง fuzzy correlation เป็นเกณฑ์ที่ใช้บอกถึงระดับความสัมพันธ์ระหว่าง แอททริบิวต์ a และแอททริบิวต์ b ความสัมพันธ์มีค่าอยู่ในช่วง [-1, 1] โดยค่าบวกหมายถึงมีความสัมพันธ์ไปในทิศทางเดียวกัน ค่าลบหมายถึงมีความสัมพันธ์ตรงกันข้าม และศูนย์หมายถึงไม่มีความสัมพันธ์กัน ค่า FCORR ของกฎ a-> b คำนวณได้ตามสมการ 3-4

$$\frac{FS(a \cup b) - FS(a) \times FS(b)}{\sqrt{FS(a) \times (1 - FS(a)) \times FS(b) \times (1 - FS(b))}} \quad (3-4)$$

Firing_strength หมายถึง ค่าสนับสนุนของกฎ คำนวณเช่นเดียวกับค่า fuzzy support

ในขั้นตอนการสร้างกฎ FCARs จะคัดเลือกเฉพาะกฎที่มีค่าคะแนนสูงกว่าศูนย์ กฎทั้งหมดที่สร้างและคัดเลือกด้วยเกณฑ์คะแนนขั้นต่ำสรุปได้ดังตารางที่ 3.4

ตารางที่ 3.4 กฎการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือ

FCARs (Fuzzy Classification with Association Rules)		Score
Age=low	-> 2	1.1459
Age=medium	-> 1	0.0611
Age=high	-> 1	1.2725
Health expense=low	-> 2	0.5953
Health expense=medium	-> 2	0.0393
Health expense=high	-> 1	1.0602
Health insurance=low	-> 2	1.9225
Health insurance=medium	-> 1	1.2677
Health insurance=high	-> 1	0.4088
Age=me & Health expense=medium	-> 1	0.0577
Age=high & Health expense=high	-> 1	1.0330
Age=low & Health insurance=low	-> 2	1.1619
Age=high & Health insurance=medium	-> 1	1.2608
Health expense=high & Health insurance=medium	-> 1	1.0421
Age=high & Health expense=high & Health insurance=medium	-> 1	1.0104

ขั้นตอนการเลือกกฎการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือ (FCARs Selection)

กฎการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือ (FCARs) ที่ได้จากขั้นตอนก่อนหน้าจะถูกลดจำนวนลงโดยกฎแต่ละขนาดของไอเท็มและแต่ละคลาส จะถูกคัดเลือกไว้เฉพาะกฎที่มีค่าคะแนนสูงสุด (Fuzzy Classification with Association Rules) ผลลัพธ์ที่ได้จะเป็นตามตารางที่ 3.5

ตารางที่ 3.5 กฎการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือที่มีค่าคะแนนสูงสุด

FCARs (Fuzzy Classification with Association Rules)		Score
Age=high	-> 1	1.2725
Health insurance=low	-> 2	1.9225
Age=low & Health insurance=low	-> 2	1.1619
Age=high & Health insurance=medium	-> 1	1.2608
Age=high & Health expense=high & Health insurance=medium	-> 1	1.0104

3.4 การแปลงโมเดลเป็นฐานความรู้

โมเดลในลักษณะของกฎการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือ จะถูกแปลงเป็นฐานความรู้ในลักษณะของข้อความเชิงตรรกะดังรูปที่ 3.3

```

health.kb - Notepad
File Edit Format View Help
% Knowledge base automatically created for the simple decision support system.

% top_goal is where the inference starts.

top_goal(X,V) :- type(X,V).

% Generated rules:

type(class_1, 1.2725):-age(high).
type(class_2, 1.9225):-health_insurance(low).
type(class_2, 1.1619):-age(low), health_insurance(low).
type(class_1, 1.2608):-age(high), health_insurance(medium).
type(class_1, 1.0104):-age(high), health_expense(high), health_insurance(medium).

% Generated menu:

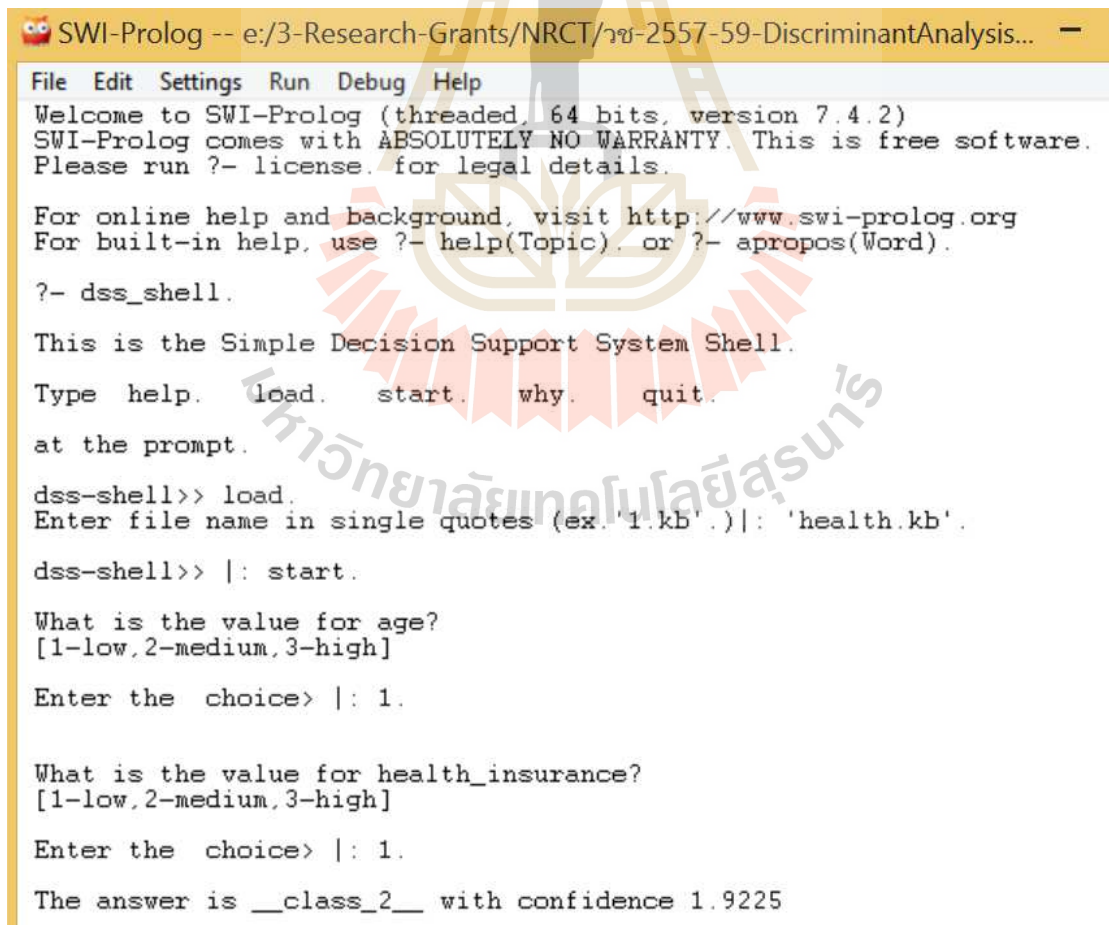
age(X):-menuask(age,X,[low,medium,high]).
health_expense(X):-menuask(health_expense,X,[low,medium,high]).
health_insurance(X):-menuask(health_insurance,X,[low,medium,high]).
class(X):-menuask(class,X,[1,2]).

% end of automatic KB creation
  
```

รูปที่ 3.3 ฐานความรู้ที่ได้จากกฎการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือ

การใช้งานฐานความรู้จะกระทำผ่านกลไกการอุปนัยของระบบสนับสนุนการตัดสินใจ ในงานวิจัยนี้พัฒนาระบบสนับสนุนการตัดสินใจอย่างง่ายด้วยภาษาโปรล็อก (Prolog) ซึ่งเป็นโปรแกรมเชิงตรรกะ ผู้ใช้เริ่มใช้งานโปรแกรมด้วยคำสั่ง `dss_shell` ซึ่งหมายถึงเปลี่ยนนอกของระบบสนับสนุนการตัดสินใจที่จะทำหน้าที่ติดต่อกับฐานความรู้ และทำหน้าที่อุปนัยเพื่อค้นหาคำตอบให้กับผู้ใช้ เปลี่ยนของระบบนี้สามารถใช้งานกับฐานความรู้ได้หลากหลาย ผู้ใช้จึงต้องระบุชื่อฐานความรู้ในตัวอย่างนี้บันทึกฐานความรู้ไว้ในชื่อ `health.kb`

เมื่อระบุฐานความรู้แล้ว เริ่มใช้งานโปรแกรมด้วยคำสั่ง `start` การใช้งานโปรแกรมจะเป็นลักษณะโต้ตอบ โดยโปรแกรมจะตั้งคำถามให้ผู้ใช้ตอบ คำตอบจะเป็นทางเลือกให้ผู้ใช้เลือกคำตอบด้วยตัวเลขอย่างง่าย เช่น 1/2/3 แทนความหมาย low/medium/high ตัวอย่างการถามตอบเกี่ยวกับการจำแนกประเภทผู้ป่วยเป็น class 1 หรือ class 2 แสดงดังรูปที่ 3.4 ผู้ใช้สามารถสอบถามเหตุผลที่ระบบให้คำตอบดังที่ปรากฏได้ด้วยการใช้คำสั่ง `why` ระบบจะแสดงผลของการได้มาซึ่งคำตอบดังแสดงในรูปที่ 3.5 และในกรณีที่ตัวเลือกในคำตอบของผู้ใช้ไม่ปรากฏผลสรุปในฐานความรู้ ระบบจะให้คำตอบดังแสดงในรูปที่ 3.6



```

SWI-Prolog -- e:/3-Research-Grants/NRCT/ฯ-2557-59-DiscriminantAnalysis...
File Edit Settings Run Debug Help
Welcome to SWI-Prolog (threaded, 64 bits, version 7.4.2)
SWI-Prolog comes with ABSOLUTELY NO WARRANTY. This is free software.
Please run ?- license. for legal details.

For online help and background, visit http://www.swi-prolog.org
For built-in help, use ?- help(Topic). or ?- apropos(Word).

?- dss_shell.

This is the Simple Decision Support System Shell.
Type help. load. start. why. quit.
at the prompt.

dss-shell>> load.
Enter file name in single quotes (ex. '1.kb'.)|: 'health.kb'.

dss-shell>> |: start.

What is the value for age?
[1-low,2-medium,3-high]

Enter the choice> |: 1.

What is the value for health_insurance?
[1-low,2-medium,3-high]

Enter the choice> |: 1.

The answer is __class_2__ with confidence 1.9225

```

รูปที่ 3.4 การโต้ตอบกับฐานความรู้ของระบบสนับสนุนการตัดสินใจอย่างง่าย

```

dss-shell>> |: start.

What is the value for age?
[1-low,2-medium,3-high]

Enter the choice> |: 1.

What is the value for health_insurance?
[1-low,2-medium,3-high]

Enter the choice> |: 1.

The answer is __class_2__ with confidence 1.9225

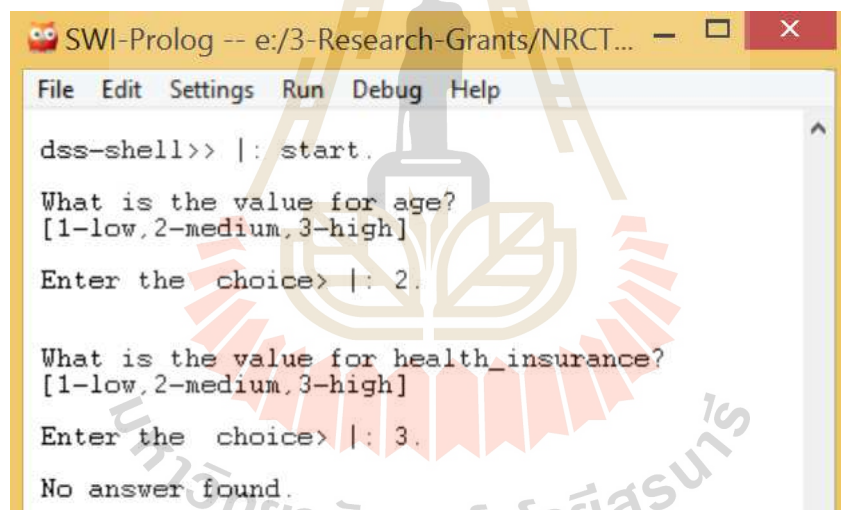
dss-shell>> |: why.

The answer is ...class_2... with confidence = 1.9225.
Because the known storage are:

[health_insurance(low),age(low)]

```

รูปที่ 3.5 การสอบถามเหตุผลของการได้มาซึ่งคำตอบของระบบ



The screenshot shows a window titled "SWI-Prolog -- e:/3-Research-Grants/NRCT...". The window contains a text-based interface for a Prolog shell. The text displayed is as follows:

```

File Edit Settings Run Debug Help

dss-shell>> |: start.

What is the value for age?
[1-low,2-medium,3-high]

Enter the choice> |: 2.

What is the value for health_insurance?
[1-low,2-medium,3-high]

Enter the choice> |: 3.

No answer found.

```

รูปที่ 3.6 คำตอบของระบบในกรณีที่ตัวเลือกของผู้ใช้ไม่ตรงกับข้อมูลที่มีในฐานความรู้

บทที่ 4

ผลการทดสอบโปรแกรม CCFAR

โครงการวิจัยนี้มีวัตถุประสงค์หลักเพื่อพัฒนาวิธีการวิเคราะห์ตัดสินใจในรูปแบบใหม่ที่ใช้เทคนิคเหมืองข้อมูลประเภทการค้นหาความสัมพันธ์ อัลกอริทึมที่พัฒนาขึ้นใหม่มีชื่อว่า อัลกอริทึมจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือที่กะทัดรัด (Classification with Compact Fuzzy Association Rules -- CCFAR) อัลกอริทึม CCFAR จะถูกใช้ทำหน้าที่ค้นหาโมเดลข้อมูลซึ่งจะเป็นส่วนประกอบสำคัญในการสร้างฐานความรู้สำหรับระบบสนับสนุนการตัดสินใจด้านการแพทย์ โมเดลที่ได้จะอยู่ในรูปแบบของกฎซึ่งสามารถประยุกต์ใช้ในการวินิจฉัยด้านการแพทย์ จุดเด่นของอัลกอริทึม CCFAR คือสามารถค้นหาความสัมพันธ์ที่แสดงรูปแบบที่เด่นชัดจากข้อมูลภายใต้ข้อกำหนดที่จำนวนกฎความสัมพันธ์ที่ได้จะต้องมีปริมาณเหมาะสม และกฎที่ได้จะต้องมีความถูกต้องสูงเพียงพอที่จะบันทึกไว้ในฐานความรู้ของระบบสนับสนุนการตัดสินใจด้านการแพทย์ ดังนั้น การทดสอบประสิทธิภาพจึงเป็นการทดสอบความสามารถของโปรแกรม CCFAR ทั้งในด้านความถูกต้องของโมเดลในการวินิจฉัยอาการที่นำไปสู่การเกิดโรค รวมถึงวิเคราะห์ประสิทธิภาพของโมเดลที่ได้ในด้านจำนวนกฎและความกะทัดรัดของกฎ การทดสอบจะใช้ข้อมูลที่เกี่ยวข้องกับการแพทย์

4.1 ข้อมูลที่ใช้ในการทดสอบ

ในการทดสอบความถูกต้องและประสิทธิภาพของโมเดลที่ได้จากโปรแกรม CCFAR จะใช้ข้อมูล 2 ชุด โดยเป็นข้อมูลเกี่ยวกับคนไข้ที่มีความเสี่ยงเป็นโรคหัวใจ (heart disease) และข้อมูลเกี่ยวกับคนไข้ที่มีความเสี่ยงเป็นโรคเบาหวาน (Pima Indians diabetes) ลักษณะโดยสรุปของข้อมูลทั้งสองชุดสรุปได้ดังตารางที่ 4.1

ตารางที่ 4.1 ข้อมูลที่ใช้ในการทดสอบประสิทธิภาพของโมเดลที่ได้จากโปรแกรม CCFAR

ชื่อชุดข้อมูล	จำนวน instances	จำนวน attributes	จำนวน class
heart disease	270	14	2
Pima Indians diabetes	768	9	2

ข้อมูล heart disease เป็นข้อมูลเกี่ยวกับคนไข้ที่มีความเสี่ยงเป็นโรคหัวใจ ข้อมูลนี้เผยแพร่เป็นสาธารณะ (<https://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29>) จำนวนแอททริบิวต์หรือตัวแปรต้นที่ใช้ประกอบการวินิจฉัยมี 13 แอททริบิวต์ แอททริบิวต์เป้าหมายหรือตัวแปรตามมี 1 แอททริบิวต์ คือ แอททริบิวต์ Class ซึ่งมีค่าที่เป็นไปได้สองค่าคือ 1 แทนความหมายว่าคนไข้รายนั้น ๆ ไม่มีความเสี่ยงของการเป็นโรคหัวใจ และ 2 แทนความหมายว่าคนไข้รายนั้น ๆ มีอาการของโรคหัวใจ ความหมายและค่าที่เป็นไปได้ของแอททริบิวต์ทั้งหมดแสดงได้ดังตารางที่ 4.2

ตารางที่ 4.2 รายละเอียดแอททริบิวต์ของข้อมูล heart disease

ชื่อแอททริบิวต์	ความหมายของแอททริบิวต์	ประเภทของข้อมูล	ช่วงของค่าข้อมูล
<i>Age</i>	Age of patient	real	[29, 77]
<i>Sex</i>	Sex of patient	binary	[0, 1]
<i>ChestPainType</i>	Chest pain type (4 types)	nominal	[1, 4]
<i>RestBloodPressure</i>	Resting blood pressure	real	[94, 200]
<i>SerumCholesterol</i>	Serum cholesterol in mg/dl	real	[126, 564]
<i>FastingBloodSugar</i>	Fasting blood sugar > 120 mg/dl	binary	[0, 1]
<i>ResElectrocardio</i>	Resting electrocardiographic results	nominal	[0, 2]
<i>MaxHeartRate</i>	Maximum heart rate achieved	real	[71, 202]
<i>ExInducedAngina</i>	Exercise induced angina	binary	[0, 1]
<i>Oldpeak</i>	ST depression induced by exercise relative to rest	real	[0.0, 62.0]
<i>Slope</i>	The slope of the peak exercise ST segment	order	[1, 3]
<i>NumMajorVessels</i>	Number of major vessels (0-3) colored by fluoroscopy	real	[0, 3]
<i>Thal</i>	Thalassemia (3 = normal; 6 = fixed defect; 7 = reversible defect)	nominal	[3, 7]
<i>Class</i>	1 = absence; 2= presence of heart disease	nominal	[1, 2]

ข้อมูล Pima Indians diabetes เป็นข้อมูลเกี่ยวกับการวินิจฉัยว่าคนไข้มีความเสี่ยงที่จะเป็นโรคเบาหวานหรือไม่ โดยพิจารณาจากอายุ ผลการตรวจระดับอินซูลิน ความดันโลหิต จำนวนครั้งของการตั้งครรภ์ และแอททริบิวต์อื่น ๆ ตามที่แสดงในตารางที่ 4.3 ข้อมูลนี้รวบรวมจากกลุ่มตัวอย่างเพศหญิงจำนวน 768 คน ทั้งหมดเป็นคนเชื้อสายอินเดียนแดงเผ่าพีม่าซึ่งมีหลักแหล่งอาศัยอยู่บริเวณตอนกลางและตอนใต้ของรัฐออริโซนา ประเทศสหรัฐอเมริกา (Smith et al., 1988) ข้อมูลชุดนี้สามารถดาวน์โหลดได้จากฐานข้อมูล KEEL – Knowledge Extraction based on Evolutionary Learning (<http://sci2s.ugr.es/keel/dataset.php?cod=21>)

ตารางที่ 4.3 รายละเอียดแอททริบิวต์ของข้อมูล Pima Indians diabetes

ชื่อแอททริบิวต์	ความหมายของแอททริบิวต์	ประเภทของข้อมูล	ช่วงของค่าข้อมูล
<i>Pregnancies</i>	Number of times pregnant	real	[0, 17]
<i>Glucose</i>	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	real	[0.0, 199.0]
<i>BloodPressure</i>	Diastolic blood pressure (mm Hg)	real	[0.0, 122.0]
<i>SkinThickness</i>	Triceps skin fold thickness (mm)	real	[0.0, 99.0]
<i>Insulin</i>	2-Hour serum insulin (mu U/ml)	real	[0.0, 846.0]
<i>BMI</i>	Body mass index (weight in kg/(height in m) ²)	real	[0.0, 67.1]
<i>DiabetesPedFunc</i>	Diabetes pedigree function	real	[0.078, 2.42]
<i>Age</i>	Age of patient	real	[21, 81]
<i>Class</i>	0 = tested negative (no diabetes); 1 = tested positive (has diabetes)	nominal	[0, 1]

4.2 เกณฑ์ที่ใช้ในการทดสอบ

การทดสอบประสิทธิภาพการจำแนกข้อมูลของโปรแกรม CCFAR จะใช้เกณฑ์การทดสอบสองเกณฑ์คือค่าความถูกต้องโดยรวมของการจำแนก (Accuracy) ซึ่งหมายถึงความถูกต้องของโมเดลในการจำแนกข้อมูลทดสอบในทุกคลาสข้อมูล และเกณฑ์ค่าความเหมาะสมของกฎ

(Suitability of rule set) เพื่อทดสอบปริมาณและความกะทัดรัดของกฎที่เหมาะสมกับการนำไปใช้ในการสร้างฐานความรู้

การประเมินค่าความถูกต้องโดยรวมของกฎ คำนวณได้ตามวิธีที่แสดงในสมการที่ 4-1 ค่า Accuracy จะมีค่าอยู่ในช่วง [0, 1] ค่า 0 หมายถึงโมเดลมีความสามารถในการทำนายต่ำที่สุด ค่า 1 หรือค่าที่เข้าใกล้ 1 หมายถึงโมเดลมีความสามารถในการทำนายข้อมูลได้ถูกต้องสูงมาก

$$\text{Accuracy} = \frac{\text{True positive} + \text{True negative}}{\text{All data instances}} \quad (4-1)$$

โดยค่า True positive หมายถึง จำนวนข้อมูลที่อยู่ในคลาสที่กำหนดให้เป็น positive และโมเดล

ทำนายถูกต้องว่าข้อมูลนั้นเป็นคลาส positive

True negative หมายถึง จำนวนข้อมูลที่อยู่ในคลาสที่กำหนดให้เป็น negative และโมเดล

ทำนายถูกต้องว่าข้อมูลนั้นเป็นคลาส negative

All data instances หมายถึง จำนวนข้อมูลทั้งหมดที่ใช้ทดสอบโมเดล

การประเมินค่าความเหมาะสมของกฎด้วยเกณฑ์ Suitability of rule set เป็นการประเมินระดับความเหมาะสมของโมเดลที่ได้ว่ามีความถูกต้องและจำนวนกฎที่ได้มีความเหมาะสมซึ่งหมายถึงมีความกะทัดรัดเพียงใด โดยเกณฑ์ Suitability of rule set มีค่าอยู่ในช่วง [0, 1] ถ้าเป็นค่า 0 แสดงว่าโมเดลที่ได้ไม่เหมาะสมเนื่องจากมีค่าความถูกต้องต่ำและมีจำนวนกฎที่มากซึ่งทำให้กฎขาดความกะทัดรัด แต่ถ้า Suitability of rule set มีค่าเป็น 1 หรือมีค่าเข้าใกล้ 1 แสดงว่าโมเดลที่ได้เหมาะสมมาก นั่นคือมีค่าความถูกต้องสูงและมีจำนวนกฎที่น้อยมากทำให้ได้ชุดของกฎที่กะทัดรัด

การคำนวณค่า Suitability of rule set ที่นำเสนอขึ้นใหม่ในโครงการวิจัยนี้ใช้วิธีเฉลี่ยค่าระหว่างความถูกต้องและความกะทัดรัดของกฎในลักษณะมัชฌิมฮาร์โมนิก (harmonic mean) ตามวิธีที่แสดงในสมการที่ 4-2 โดยที่ค่าความกะทัดรัดของกฎ (Rule compactness) ใช้วิธีทำให้เป็นบรรทัดฐาน (normalize) เพื่อให้ค่าเฉลี่ยของจำนวนกฎที่ได้อยู่ในรูปแบบปกติที่มีค่าอยู่ในช่วง [0, 1] คำนวณได้ตามสมการที่ 4-3

$$\text{Suitability of rule set} = 2 * \frac{\text{Accuracy} * \text{Rule compactness}}{\text{Accuracy} + \text{Rule compactness}} \quad (4-2)$$

$$\text{Rule compactness} = \frac{\text{RuleCompact}_{\max} - \text{RuleCompact}_{\text{algorithm}}}{\text{RuleCompact}_{\max} - \text{RuleCompact}_{\min}} \quad (4-3)$$

โดยค่า Accuracy หมายถึง ค่าความแม่นยำในการทำนายของโมเดลที่คำนวณเฉลี่ยจากการทดสอบหลายครั้ง โดยในงานวิจัยนี้ใช้การเฉลี่ยจากการทดสอบโมเดล 10 ครั้ง หรือ 10-fold cross validation

Rule compactness หมายถึง ความกะทัดรัดของกฎที่พิจารณาจากจำนวนกฎที่ได้จากโมเดล โดยเฉลี่ยจากการทดสอบ 10 ครั้ง และทำการนอร์มัลไลซ์ให้อยู่ในช่วง $[0, 1]$ ด้วยการคำนวณตามสมการ 4-3

RuleCompact_{max} หมายถึง จำนวนกฎสูงสุดที่ได้จากโมเดล ในกรณีที่ทดสอบหลายโมเดล หรือทดสอบโมเดลเดียวแต่ประมวลผลหลายครั้ง จะคัดเลือกจากครั้งที่ให้จำนวนกฎมากที่สุด

RuleCompact_{min} หมายถึง จำนวนกฎน้อยที่สุดที่ได้จากโมเดล ในกรณีที่ทดสอบหลายโมเดล หรือทดสอบโมเดลเดียวแต่ประมวลผลหลายครั้ง จะคัดเลือกจากครั้งที่ให้จำนวนกฎน้อยที่สุด

RuleCompact_{algorithm} หมายถึง จำนวนกฎที่ได้จากโมเดลซึ่งสร้างจากอัลกอริทึมที่กำลังพิจารณา เพื่อต้องการเปรียบเทียบประสิทธิภาพด้านความกะทัดรัดของกฎเทียบกับอัลกอริทึมอื่น ๆ

4.3 ผลการทดสอบประสิทธิภาพการจำแนกข้อมูลของโปรแกรม CCFAR

ในการทดสอบประสิทธิภาพการจำแนกข้อมูลจะพิจารณาจากเกณฑ์ Accuracy และ Suitability of rule set โดยเกณฑ์ Accuracy ใช้ประเมินความถูกต้องของโมเดล และเกณฑ์ Suitability of rule set ใช้ประเมินความเหมาะสมของกฎ ซึ่งใช้ค่าความถูกต้องและจำนวนกฎที่ได้มาพิจารณาร่วมกัน

การพิจารณาความสามารถของโปรแกรม CCFAR ตามเกณฑ์ Accuracy และ Suitability of rule set จะเปรียบเทียบกับอัลกอริทึมอื่นที่นิยมใช้ในงานทำเหมืองข้อมูลแบบจำแนกที่สามารถแสดงผลลัพธ์ในรูปแบบของกฎ ได้แก่ อัลกอริทึม C4.5 รวมถึงเปรียบเทียบกับอัลกอริทึมงานทำเหมืองข้อมูลแบบจำแนกที่อิงแนวทางการค้นหาความสัมพันธ์ที่มีแนวคิดคล้ายคลึงกับงานวิจัยนี้อีก 4 อัลกอริทึม ได้แก่ อัลกอริทึม CBA, OAC, FURIA, CFARC

อัลกอริทึม C4.5 (Quinlan, 1992) เป็นอัลกอริทึมที่รู้จักกันดีและนิยมใช้ในการจำแนกประเภทข้อมูล อัลกอริทึมนี้ปรับปรุงเพิ่มเติมจากอัลกอริทึม Id3 (Quinlan, 1986) ให้สามารถจัดการกับข้อมูลที่เป็นค่าต่อเนื่องได้โดยใช้หลักการค้นหาแบบฮิวริสติกเพื่อสร้างต้นไม้ตัดสินใจ และใช้เทคนิคในการตัดกิ่งต้นไม้เพื่อจัดการกับปัญหา overfitting และทำให้อัลกอริทึมนี้สามารถประมวลผลได้รวดเร็ว ข้อเด่นของอัลกอริทึมนี้คือกฎที่ได้รับนั้นผู้ใช้สามารถทำความเข้าใจได้ง่าย

อัลกอริทึม CBA – Classification Based on Associations (Liu *et al.*, 1998) เป็นอัลกอริทึมที่ใช้ในการจำแนกข้อมูลที่ได้รับการทดสอบว่ามีความแม่นยำในการจำแนกที่สูงกว่า C4.5 โดยอัลกอริทึม CBA เป็นอัลกอริทึมแรกที่น่าแนวคิดของการทำเหมืองข้อมูลแบบกฎความสัมพันธ์และการทำเหมืองข้อมูลแบบจำแนกประเภทมาทำงานร่วมกัน โดยขั้นตอนการทำงานของ CBA จะประกอบด้วย 2 ส่วน คือ ส่วนที่สร้างกฎความสัมพันธ์ซึ่งเรียกว่า CBA-RG และส่วนที่สร้างกฎการจำแนกจากกฎความสัมพันธ์ เรียกว่า CBA-CB ซึ่งแนวคิดหลักของส่วนสร้างกฎความสัมพันธ์ CBA-RG คือ สร้างกฎความสัมพันธ์ที่มีค่าสนับสนุนมากกว่าค่าสนับสนุนขั้นต่ำ และกฎความสัมพันธ์ที่ได้จะอยู่ในรูปแบบ $\langle \text{condset}, y \rangle$ ซึ่ง condset หมายถึง เซตของไอเท็ม และ y หมายถึงคลาสเป้าหมาย ส่วนแนวคิดของส่วน CBA-CB คือ ทำการคัดเลือกกฎความสัมพันธ์ที่มีค่าความผิดพลาดน้อยที่สุด

อัลกอริทึม OAC -- Optimal Association Classifier (Hu and Li, 2005) เป็นอัลกอริทึมที่มีประสิทธิภาพในการจำแนกข้อมูลที่ดี โดยใช้หลักการสร้างกฎความสัมพันธ์ด้วยอัลกอริทึม Apriori และคัดเลือกกฎความสัมพันธ์ด้วยการใช้เงื่อนไขบังคับ (Bayardo *et al.*, 1999) เพื่อเลือกกฎความสัมพันธ์ที่มีคลาสเป้าหมายตามต้องการ จากนั้นเลือกกฎที่เหมาะสมสำหรับการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์ด้วยการเลือกกฎที่ให้ความแม่นยำสูงที่สุด (Li *et al.*, 2002)

อัลกอริทึม FURIA -- An Algorithm For Unordered Fuzzy Rule (Hühn and Hüllermeier, 2009) เป็นอัลกอริทึมที่มีพื้นฐานมาจากอัลกอริทึม RIPPER แต่ในอัลกอริทึมนี้จะใช้วิธีการไม่เรียงลำดับกฎ (Unordered Rule Set) เพื่อลดการลำเอียงของคลาสเป้าหมายหลัก แต่ทำให้เกิดปัญหาใหม่คือกฎมีความขัดแย้งกันและกฎที่ได้รับมีลักษณะไม่ครอบคลุม ซึ่ง FURIA แก้ปัญหาด้วยวิธีการ Stretching เป็นวิธีการเลือกกฎที่มีค่าการครอบคลุมมากที่สุดและทำให้กฎที่ได้มีความเจาะจงน้อยลง

อัลกอริทึม CFARC -- Compact Fuzzy Association Rule-Based Classifier (Pach *et al.*, 2008) เป็นอัลกอริทึมสำหรับจัดการกับปัญหา กฎการจำแนกประเภทข้อมูลที่มีจำนวนมากเกินไป ปัญหาที่ต้องกำหนดค่าสนับสนุนขั้นต่ำและค่าความเชื่อมั่นขั้นต่ำที่เหมาะสมในการสร้างกฎความสัมพันธ์ และปัญหาการจัดการกับข้อมูลตัวเลขต่อเนื่อง ซึ่งวิธีการที่ใช้ในการแก้ปัญหาลดจำนวนกฎการจำแนกประเภทข้อมูลคือการทำการตัดกิ่งกฎที่มีค่า FCORR ที่มีค่าติดลบทิ้ง แล้วเลือกกฎที่มีค่าคะแนนสูงมาใช้ในการทำนายข้อมูล

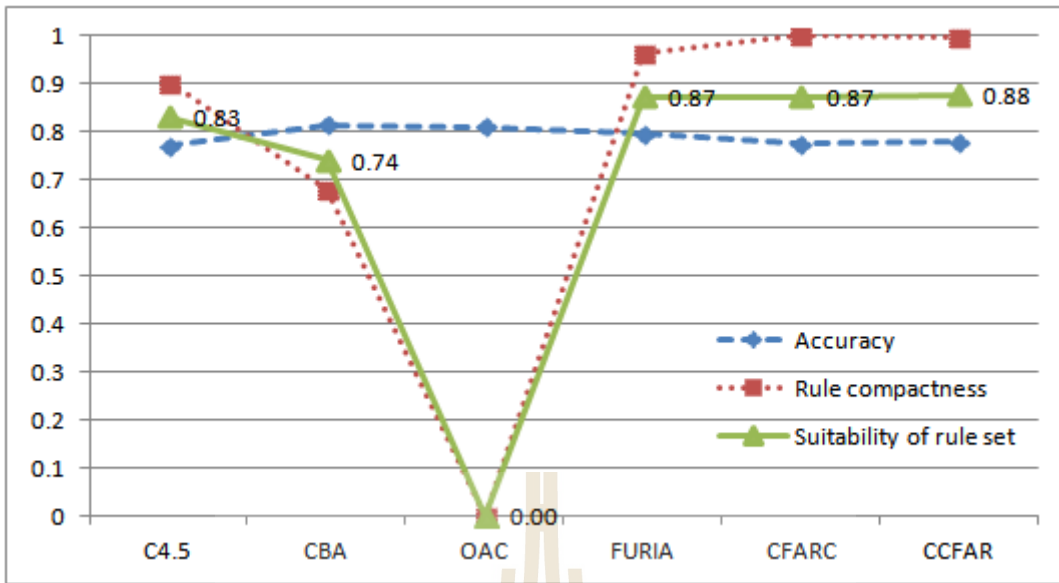
ผลการทดสอบประสิทธิภาพของโมเดลที่สร้างจากโปรแกรม CCFAR ของงานวิจัยนี้ เปรียบเทียบกับอีก 5 อัลกอริทึมคือ C4.5, CBA, OAC, FURIA, CFARC ด้วยชุดข้อมูลทดสอบโรคหัวใจ (heart disease) และโรคเบาหวาน (Pima Indians diabetes) แสดงได้ดังตารางที่ 4.4 และ 4.5 กราฟการเปรียบเทียบประสิทธิภาพแสดงได้ดังรูปที่ 4.1 และ 4.2

ตารางที่ 4.4 ผลการทดสอบประสิทธิภาพของโปรแกรมกับชุดข้อมูล heart disease

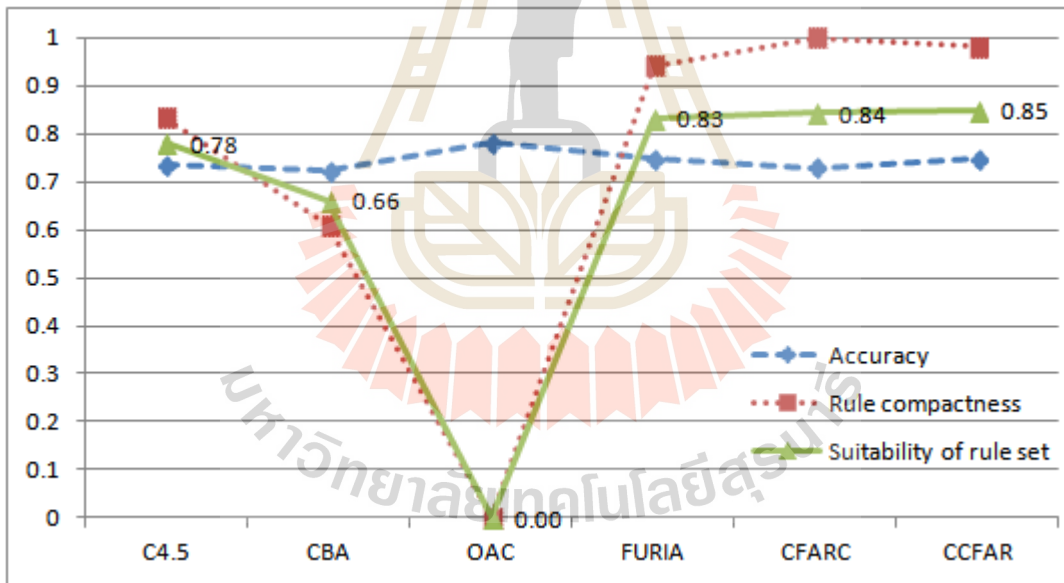
อัลกอริทึม	มาตรวัด			
	Accuracy	จำนวนกฎ (เฉลี่ย)	Rule compactness	Suitability of rule set
C4.5 (Quinlan, 1992)	0.770	17.4	0.904	0.831
CBA (Liu, Hsu, and Ma, 1998)	0.815	52.0	0.680	0.741
OAC (Hu and Li, 2005)	0.811	157.0	0.000	0.000
FURIA (Hühn and Hüllermeier, 2009)	0.797	8.4	0.963	0.872
CFARC (Pach <i>et al.</i> , 2008)	0.774	2.7	1.000	0.872
CCFAR	0.781	3.0	0.998	0.876

ตารางที่ 4.5 ผลการทดสอบประสิทธิภาพของโปรแกรมกับชุดข้อมูล Pima Indians diabetes

อัลกอริทึม	มาตรวัด			
	Accuracy	จำนวนกฎ (เฉลี่ย)	Rule compactness	Suitability of rule set
C4.5 (Quinlan, 1992)	0.734	20.3	0.833	0.780
CBA (Liu, Hsu, and Ma, 1998)	0.724	45.0	0.609	0.661
OAC (Hu and Li, 2005)	0.781	112.0	0.000	0.000
FURIA (Hühn and Hüllermeier, 2009)	0.747	8.5	0.940	0.832
CFARC (Pach <i>et al.</i> , 2008)	0.729	2.0	1.000	0.843
CCFAR	0.747	4.0	0.981	0.848



รูปที่ 4.1 กราฟเปรียบเทียบประสิทธิภาพของโปรแกรมกับชุดข้อมูล heart disease



รูปที่ 4.2 กราฟเปรียบเทียบประสิทธิภาพของโปรแกรมกับชุดข้อมูล Pima Indians diabetes

4.4 อภิปรายผลการทดสอบ

ในการทดสอบความสามารถของโปรแกรมกับชุดข้อมูลทดสอบ heart disease พบว่าโปรแกรม CBA ให้ค่าความแม่นยำในการทำนายคลาสของข้อมูลได้ถูกต้องมากที่สุด คิดเป็น 81.50% ในขณะที่ C4.5 ให้ค่าความแม่นยำต่ำที่สุดที่ 77.00% และโปรแกรม CCFAR ของโครงการวิจัยนี้จัดอยู่ในลำดับที่ 4 ที่ค่าความแม่นยำในการจำแนกข้อมูลคิดเป็น 78.10% โดยมีความแม่นยำต่ำกว่าโปรแกรมในลำดับที่หนึ่งประมาณ 3.4%

เมื่อพิจารณาในประเด็นจำนวนกฎที่ได้รับจากโปรแกรมบนชุดข้อมูลทดสอบ heart disease พบว่าโปรแกรม OAC ให้จำนวนกฎที่สูงมากที่สุดที่ประมาณ 157 กฎ ในขณะที่โปรแกรม CFARC และ CCFAR ให้จำนวนกฎเฉลี่ยอยู่ที่ 2.7 และ 3.0 โดยลำดับ การที่โปรแกรม OAC ให้ผลลัพธ์เป็นจำนวนกฎปริมาณสูงมากส่งผลให้การคำนวณค่า Rule compactness มีค่าเป็น 0 และเมื่อนำไปคำนวณร่วมกับค่าความแม่นยำด้วยวิธีคำนวณแบบ harmonic mean ส่งผลให้ค่า Suitability of rule set ของโปรแกรมนี้นี้มีค่าเป็นศูนย์ โดยโปรแกรมที่ให้ค่า Suitability of rule set สูงที่สุดคือโปรแกรม CCFAR ของงานวิจัยนี้ ลำดับที่สองคือโปรแกรม FURIA และ CFARC ซึ่งมีค่า Suitability of rule set เท่ากัน

ในกรณีที่ทดสอบกับชุดข้อมูล Pima Indian diabetes โปรแกรม OAC มีค่าความแม่นยำสูงที่สุดที่ 78.10% ในขณะที่โปรแกรม CCFAR และ FURIA อยู่ในลำดับที่สองที่ค่าความแม่นยำเท่ากันคือ 74.70% แต่เนื่องจากโปรแกรม OAC มีผลลัพธ์เป็นกฎการจำแนกที่มีปริมาณมากถึง 112 กฎ ทำให้ค่า Rule compactness มีค่าเป็นศูนย์และส่งผลต่อเนื่องให้ค่า Suitability of rule set มีค่าเป็นศูนย์เช่นเดียวกัน ในขณะที่โปรแกรม CCFAR ของโครงการวิจัยนี้มีผลลัพธ์เป็นจำนวนกฎที่น้อย (ประมาณ 4 กฎ) ทำให้ค่า Rule compactness สูง และเมื่อนำมาคำนวณร่วมกับค่าความแม่นยำในการจำแนกจะให้ค่า Suitability of rule set ที่สูงที่สุดเมื่อเปรียบเทียบกับโปรแกรมอื่นอีก 5 โปรแกรม

บทที่ 5

บทสรุป

5.1 สรุปผลการวิจัย

งานวิจัยนี้นำเสนอการออกแบบและพัฒนาขั้นตอนวิธีการ เพื่อให้ได้องค์ความรู้ใหม่ในด้านการทำเหมืองความสัมพันธ์เพื่อการวิเคราะห์ดิสคริมิแนนต์ และมีการออกแบบโครงสร้างของระบบสนับสนุนการตัดสินใจด้านการแพทย์ที่สามารถผนวกฐานความรู้จากโมเดลดิสคริมิแนนต์ข้อกำหนดหลักของงานวิจัยนี้คือโมเดลดิสคริมิแนนต์ (หรือในสาขาการทำเหมืองข้อมูลเรียกว่าโมเดลการจำแนก) จะต้องอยู่ในรูปแบบของกฎเพื่อให้สามารถแปลงเป็นฐานความรู้ได้สะดวก โมเดลจะต้องมีความถูกต้องสูงและจะต้องมีขนาดของโมเดลไม่ใหญ่เกินไปซึ่งจะทำให้โมเดลซับซ้อน และมีโอกาสที่จะทำให้เกิดการเจาะจงกับข้อมูลฝึกสอนมากเกินไป (หรือเรียกว่า overfitting) อันจะส่งผลเสียให้โมเดลขาดความแม่นยำเมื่อนำไปใช้กับข้อมูลอื่นในอนาคต

ดังนั้นงานวิจัยนี้จึงมุ่งเน้นในการพัฒนาวิธีการเพื่อค้นหาและเลือกกฎที่มีประสิทธิภาพในการจำแนกที่สูงและกฎที่ได้รับมีจำนวนน้อย ซึ่งอัลกอริทึมที่ผู้วิจัยพัฒนามีชื่อว่า Classification with Compact Fuzzy Association Rules หรือ CCFAR โดยเป็นอัลกอริทึมที่นำเทคนิคการหาความสัมพันธ์มาช่วยในการวิเคราะห์หาความสัมพันธ์กันของข้อมูล และใช้เทคนิคฟuzzyเซตมาช่วยในการแก้ปัญหาข้อมูลตัวเลขนำเข้าที่มีลักษณะเป็นค่าต่อเนื่อง อีกทั้งเทคนิคดังกล่าวยังเพิ่มความถูกต้องในการจำแนกข้อมูลที่มีการซ้อนทับกันมากอีกด้วย ซึ่งขั้นตอนวิธีการของงานวิจัยนี้ได้ใช้อัลกอริทึม CFARC (Pach et al., 2008) เป็นพื้นฐานในการพัฒนา โดยงานวิจัยนี้ได้เพิ่มเติมขั้นตอน การตรวจสอบข้อมูลก่อนการประมวลผล (ขั้นตอนที่ 1) และขั้นตอนการคัดเลือกกฎ (ขั้นตอนที่ 5) นอกจากนี้ยังได้ปรับเปลี่ยนเทคนิคการคำนวณค่าคะแนนของกฎ FCARs ในขั้นตอนที่ 4 โดยสรุปแล้วอัลกอริทึม CCFAR ที่พัฒนาขึ้นประกอบด้วยขั้นตอนดังต่อไปนี้

- (1) ขั้นตอนการตรวจสอบข้อมูลก่อนการประมวลผล
- (2) ขั้นตอนการแปลงข้อมูลตัวเลขค่าต่อเนื่องให้เป็นข้อมูลแบบฟuzzy
- (3) ขั้นตอนการสร้างไอเท็มเซตที่ปรากฏบ่อยแบบคลุ่มเครือ
- (4) ขั้นตอนการนำไอเท็มเซตที่ปรากฏบ่อยแบบคลุ่มเครือไปสร้างกฎความสัมพันธ์แบบคลุ่มเครือ และทำการเลือกกฎ FCARs ที่มีค่าคะแนนเป็นบวก โดยงานวิจัยนี้พัฒนาสมการที่ใช้คำนวณค่าคะแนนของกฎ FCARs ขึ้นมาใหม่
- (5) ขั้นตอนการคัดเลือกกฎไปใช้งาน ซึ่งประกอบด้วย 4 ขั้นตอนย่อยคือ
 - (5.1) การเลือกกฎ FCARs ที่มีค่าคะแนนมากที่สุดของแต่ละคลาสและแต่ละขนาด

- (5.2) การนับความถี่ของแต่ละแอททริบิวต์จากกฎ FCARs ในขั้นตอนย่อยที่ 5.1 และทำการเลือกแอททริบิวต์ที่มีความถี่มากที่สุด
- (5.3) การคัดเลือกกฎ FCARs ที่ดีที่สุดด้วยแอททริบิวต์ที่มีความถี่มากที่สุด
- (5.4) การลบกฎที่มีลักษณะซ้ำซ้อน

การทดสอบประสิทธิภาพของอัลกอริทึม CCFAR เป็นการทดสอบด้วยชุดข้อมูลด้านการแพทย์สองชุดข้อมูล คือ heart disease และ Pima Indian diabetes เพื่อเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลและจำนวนกฎที่ได้รับของอัลกอริทึม CCFAR โดยเทียบกับอัลกอริทึมอื่นอีก 5 อัลกอริทึม ประกอบด้วย

- อัลกอริทึม C4.5 (Quinlan, 1992)
- อัลกอริทึม CBA – Classification Based on Associations (Liu *et al.*, 1998)
- อัลกอริทึม OAC -- Optimal Association Classifier (Hu and Li, 2005)
- อัลกอริทึม FURIA -- An Algorithm For Unordered Fuzzy Rule (Hühn and Hüllermeier, 2009)
- อัลกอริทึม CFARC -- Compact Fuzzy Association Rule-Based Classifier (Pach *et al.*, 2008)

เกณฑ์ที่ใช้ในการทดสอบประสิทธิภาพ คือ ค่าความถูกต้องในการจำแนกข้อมูล จำนวนกฎที่ได้รับ และค่าความเหมาะสมของกฎเมื่อประเมินความถูกต้องร่วมกับจำนวนกฎที่ได้รับ ผลการทดสอบกับทั้งสองชุดข้อมูลพบว่าโปรแกรมที่พัฒนาจากอัลกอริทึม CCFAR ให้ผลลัพธ์ที่ดีที่สุดเมื่อประเมินด้วยเกณฑ์ค่าความเหมาะสมของกฎ โดยสรุปลำดับความสามารถของอัลกอริทึมได้ดังตารางที่ 5.1 โดยค่าที่ปรากฏในวงเล็บคือค่าความเหมาะสมของกฎ

ตารางที่ 5.1 สรุปลำดับความสามารถของโปรแกรมตามเกณฑ์ค่าความเหมาะสมของกฎ

ชุดข้อมูล	ลำดับที่ 1	ลำดับที่ 2	ลำดับที่ 3	ลำดับที่ 4	ลำดับที่ 5	ลำดับที่ 6
Heart disease	CCFAR (0.876)	CFARC (0.872)	FURIA (0.872)	C4.5 (0.831)	CBA (0.741)	OAC (0.0)
Pima Indian diabetes	CCFAR (0.848)	CFARC (0.843)	FURIA (0.832)	C4.5 (0.780)	CBA (0.661)	OAC (0.0)

5.2 ข้อจำกัดของโปรแกรมและข้อเสนอแนะ

โปรแกรมที่งานวิจัยนี้พัฒนาขึ้นยังมีข้อจำกัดในประเด็นกฎ FCARs ที่ได้รับจาก อัลกอริทึม CCFAR จะอยู่ในรูปแบบของพีชซีเซต เพราะฉะนั้นในการนำไปประยุกต์ใช้งานจะต้องแปลงข้อมูลดังกล่าวให้อยู่ในรูปแบบของช่วงข้อมูลเสียก่อน และในงานวิจัยนี้ได้กำหนดระดับความเป็นสมาชิกของพีชซีเซตให้เป็น 3 ระดับ ได้แก่ ระดับ Low ระดับ Medium และระดับ High เพราะสามารถตีความได้ง่ายและสามารถนำไปเปรียบเทียบกับอัลกอริทึมอื่นได้ ซึ่งงานวิจัยในอนาคตอาจจะต้องเพิ่มความสามารถในส่วนนี้ที่ทำให้อัลกอริทึมมีความยืดหยุ่นมากขึ้น สามารถกำหนดระดับความเป็นสมาชิกของพีชซีเซตได้หลายระดับมากขึ้น

นอกจากนี้การทดสอบความสามารถของโปรแกรม CCFAR เพื่อเปรียบเทียบกับ อัลกอริทึมอื่นยังคงค่อนข้างจำกัดด้วยชุดข้อมูลเพียงสองชุด ซึ่งในอนาคตอาจพิจารณาเพิ่มชุดข้อมูลในการทดสอบมากขึ้นเพื่อยืนยันความสามารถของโปรแกรม และในส่วนของการใช้ประโยชน์ฐานความรู้ อาจต้องใช้ในการสอบทานความถูกต้องและความเป็นไปได้ในการนำไปใช้งานจริงกับแพทย์ผู้เชี่ยวชาญ



บรรณานุกรม

- R. Agrawal and R. Srikant (1994), Fast algorithms for mining association rules in large databases, *Proceedings of the 20th International Conference on Very Large Data Bases*, pp.487-499.
- R.J. Bayardo, R. Agrawal, and D. Gunopulos (1999), Constraint-based rule mining in large, dense database, *Proceedings of the 15th International Conference on Data Engineering*, pp.188-197.
- P.N. Belhumeur, J. Hespanda, and D. Kriegeman (1997), Eigenfaces vs fisherfaces: recognition using class specific linear projection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.19, No.7, pp.711-720.
- J.C. Bezdek, R. Ehrlich, and W. Full (1984), FCM: The fuzzy c-means clustering algorithm, *Computers and Geosciences*, Vol.10, No.2, pp.191-203.
- D.-S. Cao, M.-M. Zeng, L.-Z. Yi, B. Wang, Q.-S. Xu, Q.-N. Hu, L.-X. Zhang, H.-M. Lu, and Y.-Z. Liang (2011), A novel kernel Fisher discriminant analysis: constructing informative kernel by decision tree ensemble for metabolomics data analysis, *Analytica Chimica Acta*, Vol.706, pp.97-104.
- J. Cendrowska (1987), PRISM: An algorithm for inducing modular rules, *International Journal of Man-Machine Studies*, Vol.27, pp.349-370.
- Z. Chen and G. Chen (2008), Building an associative classifier based on fuzzy association rules, *International Journal of Computational Intelligence Systems*, Vol.1, No.3, pp.262-273.
- E. Dogantekin, A. Dogantekin, and D. Avci (2009), Automatic hepatitis diagnosis system based on linear discriminant analysis and adaptive network based on fuzzy inference system, *Expert Systems with Applications*, Vol.36, No.8, pp.11282-11286.
- E. Dogantekin, A. Dogantekin, D. Avci, and L. Avci (2010), An intelligent diagnosis system for diabetes on linear discriminant analysis and adaptive network

- based fuzzy inference system: LDA-ANFIS, *Digital Signal Processing*, Vol.20, No.4, pp.1248-1255.
- E. Dogantekin, A. Dogantekin, and D. Avci (2011), An expert system based on generalized discriminant analysis and wavelet support vector machine for diagnosis of thyroid diseases, *Expert Systems with Applications*, Vol.38, No.1, pp.146-150.
- G.C.J. Fernandez (2002), Discriminant analysis: a powerful classification technique in data mining, *Proceedings of the SAS Users International Conference*, pp.247-256.
- M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten (2009), The WEKA data mining software: an update, *SIGKDD Explorations*, Vol.11, No.1, pp.10-18.
- P. Howland and H. Park (2004), Generalized discriminant analysis using the generalized singular value decomposition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.26, No.8, pp.995-1006.
- H. Hu and J. Li (2005), Using association rules to make rule-based classifiers robust, *Proceedings of the 16th Australasian Database Conference*, pp.47-54.
- H. Huang, J. Li, and J. Liu (2012), Gene expression data classification based on improved semi-supervised local Fisher discriminant analysis, *Expert Systems with Applications*, Vol.39, No.3, pp.2314-2320.
- J. Hühn and E. Hüllermeier (2009), FURIA: an algorithm for unordered fuzzy rule induction, *Data Mining and Knowledge Discovery*, Vol.19, No.3, pp.293-319.
- J.R. Quinlan (1986), Induction of decision trees, *Machine Learning*, Vol.1, pp.81-106.
- J.R. Quinlan (1993), *C4.5: Programs for Machine Learning*, Morgan Kaufmann.
- R.S. Ledley, P.S. Ing, and H.A. Lubs (1980), Human chromosome classification using discriminant analysis and Bayesian probability, *Computers in Biology and Medicine*, Vol.10, No.4, pp.209-218.
- J. Li, H. Shen, and R. Topor (2002), Mining the optimal class association rule set, *Knowledge Based Systems*, Vol.15, No.7, pp.399-405.

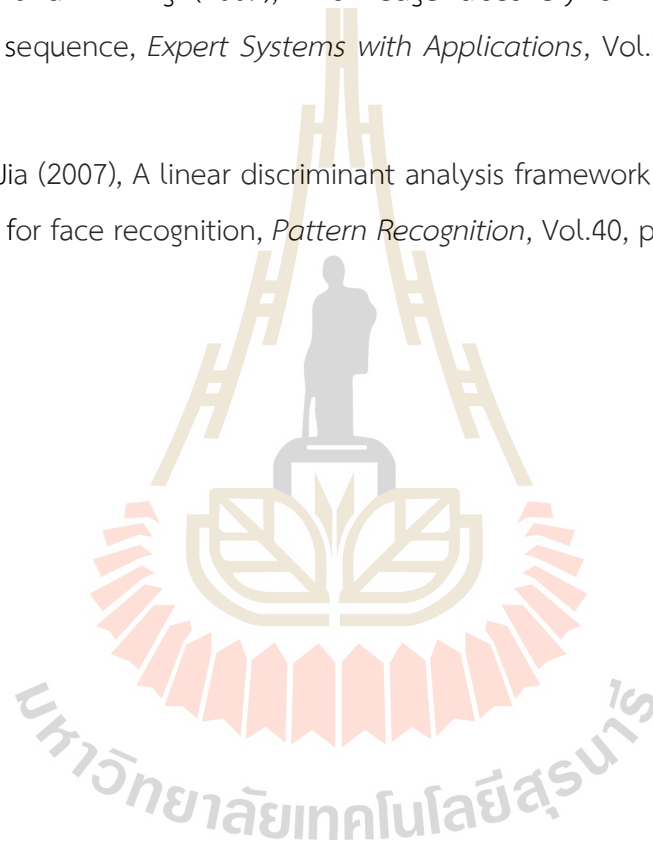
- B. Li, C.-H. Zheng, D.-S. Huang, L. Zhang, and K. Han (2010), Gene expression data classification using locally linear discriminant embedding, *Computers in Biology and Medicine*, Vol.40, No.10, pp.802-810.
- S.-W. Lin and S.-C. Chen (2009), PSOLDA: a particle swarm optimization approach for enhancing classification accuracy rate of linear discriminant analysis, *Applied Soft Computing*, Vol.9, pp.1008-1015.
- B. Liu, W. Hsu, and Y. Ma (1998), Integrating classification and association rule mining, *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, pp.80-86.
- F.P. Pach, A. Gyenesei, and J. Abonyi (2008), Compact fuzzy association rule-based classifier, *Expert Systems with Applications*, Vol.34, No.4, pp.2406-2416.
- A. Sengur (2008), An expert system based on linear discriminant analysis and adaptive neuro-fuzzy inference system to diagnosis heart valve diseases, *Expert Systems with Applications*, Vol.35, pp.214-222.
- M.P. Silva, O. Zucchi, A. Ribeiro-Silva, and M.E. Poletti (2009), Discriminant analysis of trace elements in normal, benign and malignant breast tissues measured by total reflection x-ray fluorescence, *Spectrochimica Acta Part B*, Vol.64, pp.587-592.
- J.W. Smith, J.E. Everhart, W.C. Dickson, W.C. Knowler, and R.S. Johannes (1988), Using the ADAP learning algorithm to forecast the onset of diabetes mellitus, *Proceedings of the Symposium on Computer Applications and Medical Care*, pp.261-265.
- B.G. Tabachnick and L.S. Fidell (1996), *Using Multivariate Statistics*, NY: HarperCollins.
- P.D. Turney (1995), Cost-sensitive classification: empirical evaluation of a hybrid genetic decision tree induction algorithm, *Journal of Artificial Intelligence Research*, Vol.2, pp.369-409.
- K. Umene, C. Koga, and T. Kameyama (2007), Discriminant analysis of DNA polymorphisms in herpes simplex virus type 1 strains involved in primary

compared to recurrent infections, *Journal of Virological Methods*, Vol.139, No.2, pp.159-165.

W. Xu, L. Zhang, Y. Huang, Q. Yang, H. Xiao, and D. Zhang (2012), Fatty acid metabolic profiles and biomarker discovery for type 2 diabetes mellitus using graphical index of separation combined with principal component analysis and partial least squares-discriminant analysis, *Chemometrics and Intelligent Laboratory Systems*, doi:10.1016/j.chemolab.2012.06.008

I. Yeh, K. Yang, and T. Ting (2009), Knowledge discovery on RFM model using Bernoulli sequence, *Expert Systems with Applications*, Vol.36, No.3, pp.5866-5871.

X. Zhang and Y. Jia (2007), A linear discriminant analysis framework based on random subspace for face recognition, *Pattern Recognition*, Vol.40, pp.2585-2591.



ภาคผนวก

ผลผลิตของงานวิจัย



ภาคผนวก ก

บทความวิจัยตีพิมพ์ในวารสารและเอกสารการประชุมวิชาการ

1. P. Kongchai, N. Kerdprasop and K. Kerdprasop (2014). The fuzzy search for association rules with interestingness measures. *International Journal of Computer Theory and Engineering*, Vol.6, No.6, December, pp.490-494. (indexed in EI and INSPEC, ISSN: 1793-8201)
2. P. Kongchai, K. Suksut, R. Sutamma, S. Phoemhansa, N. Kerdprasop, K. Kerdprasop (2015). The compact fuzzy association rules for data classification. *Proceedings of the 3rd International Conference on Industrial Application Engineering 2015 (ICIAE2015)*, Kitakyushu, Japan, 28-31 March, pp.30-37.
3. K. Suksut, P. Kongchai, S. Phoemhansa, R. Sutamma, K. Kerdprasop, N. Kerdprasop (2015). Single versus multiple measures for fuzzy association rule mining. *Proceedings of the 3rd International Conference on Industrial Application Engineering 2015 (ICIAE2015)*, Kitakyushu, Japan, 28-31 March, pp.273-279.
4. P. Kongchai, N. Kerdprasop, K. Kerdprasop (2013). Dissimilar rule mining and ranking technique for associative classification. *Proceedings of the 2013 IAENG International Conference on Data Mining and Applications*, Hong Kong, 13-15 March, pp.356-361.

The Fuzzy Search for Association Rules with Interestingness Measure

Phaichayon Kongchai, Nittaya Kerdprasop, and Kittisak Kerdprasop

Abstract—Association rule are important to retailers as a source of knowledge to manage shelf, to plan an effective promotion, and so on. However, when we are mining with association rule discovery technique, we normally obtain a large number of rules. To select only good rule is difficult. Therefore, in this paper we propose the fuzzy search technique to discover interesting association rule. The comparative result of fuzzy versus non-fuzzy searches are presented in the experiment section. We found that fuzzy search is more flexible than the non-fuzzy one in finding highly constrained rules.

Index Terms—Fuzzy set, fuzzy search, membership function, association rule mining.

I. INTRODUCTION

Association rule mining is a method to discover the patterns of information, such as the pattern to reveal that there are many people coming in the supermarket to buy some specific set of products. Therefore, the owner would like to know the buying patterns of customers. The owner should perform by 2 steps, first step, to records the purchase of individual customers in the tables. Second step, when get enough information then bring it to association rule mining and then the results are association rules. This method is called "Market Basket Analysis" and this association rules are usually used in business [1]. Therefore, to select the appropriate association rules to apply, it is necessarily very much and the researcher [2] proposes an algorithm to search with many constraints and can be reduced the search space.

But the most researchers continue to straightforward search association rules with normal constraint, For example if the user required support value to be equal 1.0 and items in the 'then' must be X (The variable X means items that the user wants.) [3]-[6] this searching technique is less efficient than fuzzy searching technique because the results must be support value as 1.0 only (Which makes it does not received the close results, such as support 0.99 but the fuzzy searching technique may be received the results with that support 0.99.). The Fuzzy set is very popular in a variety of major because it can indicate the level of what is uncertain. There are many researchers used it during processing, such as a data support system for sales promotion analysis using fuzzy query [7] this work used technical fuzzy and used probability to search sales information by SQL language. Which the

original searching was not able to searched some information but fuzzy searching and weight with probability they can do it. There are many tasks related to using SQL and fuzzy in that searching [8], [9].

This research proposed the method to search association rules by applying fuzzy set and search them from the measure performance of association rule (Support and Confident). We also proposed technique to select and to rank the association rules with the scores, therefore the results are very satisfactory in some case.

II. RELATED THEORIES AND STUDIES

A. Basic of Association Rule

Association rule is a data mining to discover the patterns or relationships of items from the large database [10]. Which is can be using association rules to predict something happen in the future.

Example 1

Beer, Coke => Diaper
or
If Beer, Coke then Diaper

It means if people who buy beer and coke then buy diaper together. Therefore the association rule is important to business or something that need to find relationships. To create the association rules that have two steps.

Step 1, Find all frequent itemsets meaning itemsets whose greater than or equal minimum support, and then we can find the support with equation (1).

Support(A) = all transaction that contain A/all transaction (1)

Step 2, Generate association rules with the frequent itemsets whose greater than or equal 2-itemsets and the association rules must be greater than or equal minimum confidence. We can find the confidence of the association rules by an equation (2).

Confidence(A=>B) = support(A and B)/support(A) (2)

B. Fuzzy Set

Fuzzy sets are two words that come from the word "fuzzy" that mean something is not clear, such as, The feeling of people are differently to discriminate, such as someone open air at 15 degrees, that is cool, but other people said that be very cold, and another word is "set" in this case mean a set of

Manuscript received December 13, 2013; revised March 1, 2014. This work was supported in part by grant from Suranaree University of Technology through the funding of Data Engineering Research Unit.

The authors are with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand (e-mail: Zaguraba_ii@hotmail.com).

mathematical sets that are composed a different member of the set, for instance, that consists of people, animals and objects. Therefore the term of “fuzzy sets” mean the sets that are composed a fuzzy member. This uncertainty, we cannot say whether it is true or not true, but we can tell level of the fact by membership functions.

In detail of how to compute the membership functions we explained in the next section. However, we can computed the relationship of fuzzy sets by 3 fuzzy set operations [11].

- 1) Fuzzy Complements may be called a complement of any set, which is a set has a relationship with another set, and then we can compute this relationship by equation as follows (3).

$$\mu_{A^c}(X) = 1 - \mu_A(X) \quad (3)$$

- 2) Fuzzy Intersections is to extract the duplicate memberships of two sets or more, and then we can compute this relationship by equation as follows (4).

$$\mu_{A \cap B}(X) = \min[\mu_A(X), \mu_B(X)] \quad (4)$$

- 3) Fuzzy Unions are to include the fuzzy members of sets with the relationship. By equation as follows (5).

$$\mu_{A \cup B}(X) = \max[\mu_A(X), \mu_B(X)] \quad (5)$$

In this paper we used two operators to be computed the relationship of fuzzy sets are fuzzy intersections and fuzzy unions.

III. THE FUZZY SEARCHING ASSOCIATION RULES TECHNIQUE

The methodology of this research to search the association rules with fuzzy set technique is composed of three steps: (Fig. 1) first step user is defining user-constraint to search association rules from measure support and confidence, The second step computes a membership value of the association rules from membership function, and the last step the results were to calculate the score for select appropriate association rules and ranking, details of all the methods to describe as follows.

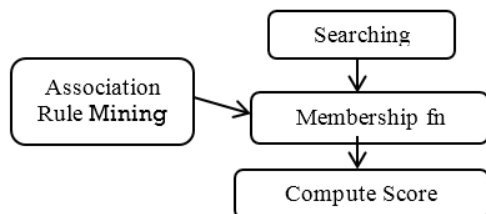


Fig. 1. The process of fuzzy search for association rules.

A. Searching

Association rule discovery based on user-constraint, the user can be defined the support and confidence value of the association rules by using the word “approximately”, “approximately more than”, “approximately less than”,

“and” and “or”.

Example 2

User require association rules with the support *approximately* 0.7 and The confidence *approximately* 0.8

Example 3

User require association rules with the support *approximately less than* 0.7 or The confidence *approximately* 0.8

Example 4

User require association rules with the support *approximately* 0.7 and The confidence *approximately* more than 0.8

From examples, you can be seen that the words (the words use Italics) used are different from normal search (Example, equal, less than, more than) because the user can define the conditions in the word form of approximately (Notice: the numbers are defined as 0.7 and 0.8 because we make it easier to explain, but you can change the value that you want.).

B. Association Rule

This research used association rules are the input to search the approximately rules. By association rules will be formatted as if – then. And the “if” and “then” contain information are call item, such as if A and B then C, if A and B then C and D. In addition, each rule has relation values that indicate the quality of association rules. Typically, most people used the support and confidence as the criterion to selected association rules, the support measure will indicate a number of transactions or the number of rows that support association rules, and the confidence measure can indicate the validity of the rules, which these measures will be important in determining selection association rules.

TABLE I: THE EXAMPLE OF ASSOCIATION RULES.

NO.	Rules	Support	Confidence
1	G then F	0.1	0.4
2	B then C	0.2	0.44
3	F then E	0.3	0.45
4	A or B then C	0.48	0.56
5	A and B then E	0.5	0.58
6	G and B then E	0.6	0.59
7	F and T and B then C	0.79	0.62
8	A and B and G then F	0.83	0.75
9	G and R then F and B	0.9	0.76
10	A and F then C and B	1	0.8

C. Membership Functions

The Membership functions are intended to indicate the degree of the fuzzy sets. In this research we choose three functions are Triangular membership function (Fig. 2 a), R membership function (Fig. 2 b) and L membership function (Fig. 2 c). Which we will choose the triangulation membership function for conditions with the word “approximately” because the center point of the function

(Fig. 2 a, point b), with a maximum membership value is 1 which corresponds a user-defined the word about it, the R membership function is selected when the condition has the word "approximately less than" because of the value of most preferred users, they must be less than or equal to the value of user-defined then they have the membership value is 1, and the values which greater than the value of user-defined that will be had the membership value is reducing, and the L membership function be selected when the condition has the word "approximately more than" because of the value of most preferred users, they must be more than or equal to the number of user-defined then they have the membership value is 1, and the values whose less than the value of user-defined that will be had the membership value is reducing. From the Example 4 the user requires the association rules with the support approximately 0.8 and the confidence rather than 0.7. In the process of membership can be achieved by the following.

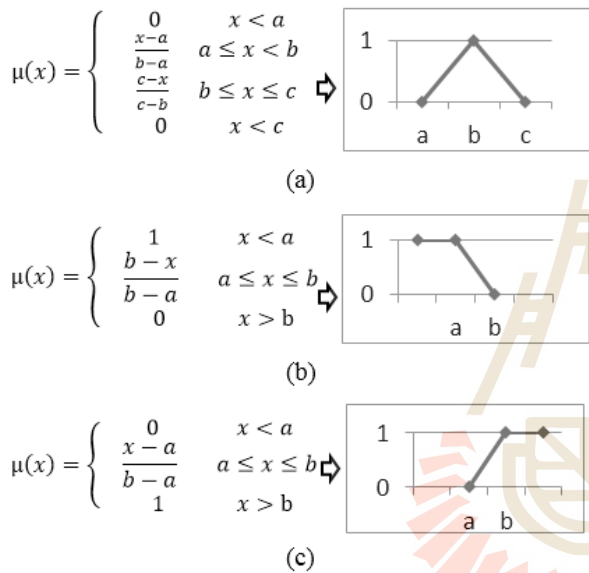


Fig. 2. Three membership functions.

The first condition is the user-defined support approximately 0.8, the word "approximately" (approximately only), it means the function must be a triangulation membership function (For example we used information from the Table I). The values of the variables a , b and c from the triangulation membership function they are meaning, the variable a is the minimum value of data, it is 0.1 but to increase the flexibility of the results, in this research will be decreasing the minimum value with minus 0.05 for increase the chances to discover the association rule with the minimum support, therefore the new minimum value is 0.05, the value of the variable b is 0.8 because user-defined and the neighboring of this value is the most important than the other values, the variable c is the maximum value of data, but to increase the flexibility of the results, in this research will be increasing of the maximum value to add 0.05 to increase the chances to discover the association rule with the maximum support, the new value of variable c is 1.05. And then we concluded the value of variables that showed in the Table II and instead support value of each association rule into an equation (Fig. 2 a), after that the results showed in the Table

IV (notice: $\mu_A(X)$ From the Table IV if there have variables which has value more than 1 then we will be decreased it to 1, such as there has one variable it has membership value is 1.05 that is more than 1 then we will be decreased it to 1.00).

The second condition is the user-defined confidence approximately more than 0.7, it means the function must be an L membership function because the word "approximately more than", and this function composed two variables a and b are meaning, the variable a is the minimum value of data, the value of the variable b is 0.7 because same the reason variable b in triangulation membership function, And then We are concluding the value of variables in the Table III and instead confidence of each association rule into an equation (Fig. 2 c), the results showed in the Table IV. In this research does not explain the R membership function because its method is similar to the L membership function.

TABLE II: THE VALUES OF TRIANGULATION MEMBERSHIP FUNCTION

Variable	Support
a	0.05
b	0.8
c	1.05

TABLE III: THE VALUES OF L MEMBERSHIP FUNCTION

Variable	Confidence
a	0.4
b	0.7

TABLE IV: THE RESULTS OF L MEMBERSHIP FUNCTION AND TRIANGULATION MEMBERSHIP FUNCTION

Support	$\mu_A(X)$	Confidence	$\mu_B(X)$
0.1	0.06	0.4	0
0.2	0.2	0.44	0.13
0.3	0.33	0.45	0.16
0.48	0.57	0.56	0.53
0.5	0.6	0.58	0.6
0.6	0.73	0.59	0.63
0.79	1.0	0.62	0.73
0.83	0.88	0.75	1
0.9	0.6	0.76	1
1	0.2	0.8	1

TABLE V: THE SCORES OF EACH ASSOCIATION RULE.

NO.	$\mu_{B(AorB)}(X)$
1	0
2	0.13
3	0.16
4	0.53
5	0.6
6	0.63
7	0.73
8	0.88
9	0.6
10	0.2

D. Computation Score to Rank

This step is the final step to compute the scoring of

association rules for select and rank them. Which association rules are selected there score of them must be more than a scoring of user-defined, and in this research we defined the scoring is 0.6. And then we rank the association rules which any association rules have a score more than other rules, they will be ranked in the first order. The computation score with this method, if the user selects the term connected by "and" is used an equation (6), but if the user selects the term connected by "or" is used in equation (7). And then sort the association rules with the scores.

$$\mu_{(A \text{ and } B)}(X) = \min[\mu_A(X), \mu_B(X)] \quad (6)$$

$$\mu_{(A \text{ or } B)}(X) = \max[\mu_A(X), \mu_B(X)] \quad (7)$$

From the example 4 with the condition is "and", which can be calculated from the Table IV instead into Equation (6), the results are shown in the Table V and then we selected the association rules with the scoring more than 0.6 the results are No.6, 7, 8 and then we ranked them which any association rules that have a score more than other rules, they will be ranked in the first order, therefore the results are No.8, 7, 6 respectively.

IV. EXPERIMENT

This research used data from a random support and confidence (the values are 0.01 to 1.00) to 10k, 50k, 100k and 150k records to test performance of fuzzy searching and to compare non-fuzzy (normal) searching with the various conditions. And the scoring of user-defined in fuzzy searching association rules we defined it to be greater than 0.8. The results are shown in the Table VI.

It can be noticed from the results (Table VI) that the condition can be reduced the number of association rules very much. In particular, the conditions "s = 0.52 and c = 0.52" because condition "and" to find the minimum value of membership functions, which most association rules do not a predetermined threshold. Therefore, the association rules that have more quality. But the condition "s <= 0.52" and condition "s >= 0.52" to give similar results because the association rule where the member is 1 will be support begin at the 0.52 (Middle), lead to a similar number of association rules but the most association rules are differently.

TABLE VI: THE RESULTS FROM 5 CONDITIONS OF THE FUZZY SEARCHING

Constraint	10k	50k	100k	150k
s = 0.52	2,175	11,042	22,002	32,956
s <= 0.52	6,132	30,613	61,002	91,516
s >= 0.52	6,043	30,431	61,231	91,440
s = 0.52 and c = 0.52	468	2,422	4,845	7,225
s = 0.52 or c = 0.52	3,876	19,596	39,200	58,743

By the words from Table VI. Represented by these symbols.

"Approximately" represented by "≈"

"Approximately less than" represented by "≈<"

"Approximately more than" represented by "≈>"

"Support" represented by "s"

"Confidence" represented by "c"

TABLE VII: THE RESULTS FROM 5 CONDITIONS OF THE NON-FUZZY SEARCHING

Constraint	10k	50k	100k	150k
s = 0.52	103	490	144	1,485
s <= 0.52	4,906	25,011	75,332	75,016
s >= 0.52	5,197	25,479	76,113	76,469
s = 0.52 and c = 0.52	0	0	3	5
s = 0.52 or c = 0.52	209	992	285	2,988

DATA=10K

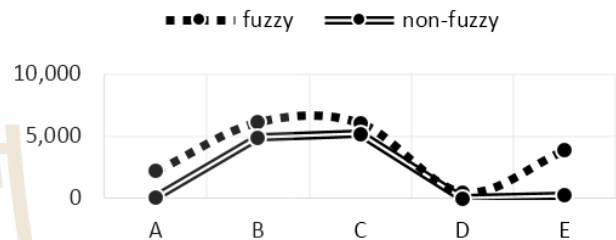


Fig. 3. Fuzzy searching Vs. non-fuzzy (10K).

DATA=50K

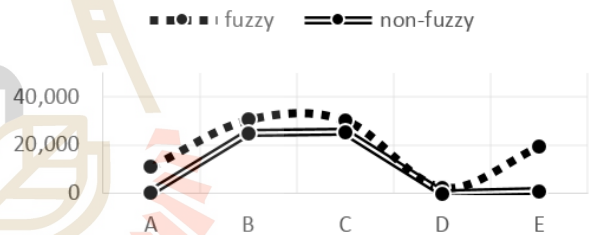


Fig. 4. Fuzzy searching Vs. non-fuzzy (50K).

DATA=100K

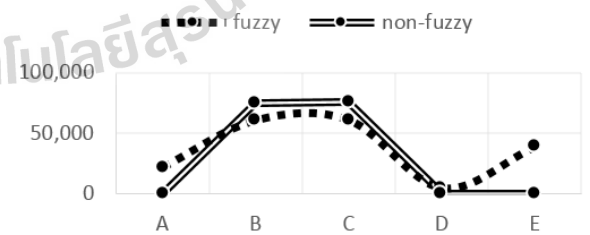


Fig. 5. Fuzzy searching Vs. non-fuzzy (100K).

DATA=150K

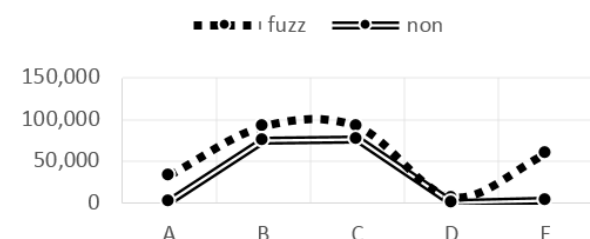


Fig. 6. Fuzzy searching Vs. non-fuzzy (150K).

This results of non-fuzzy (Table VII) is decreased more than fuzzy searching in all cases (Fig. 3 to Fig. 6 Note: the symbols in graphs they mean A: $s = 0.52$, B: $s \geq 0.52$, C: $s = 0.52$ and $c = 0.52$, D: $s = 0.52$ or $c = 0.52$), in condition " $s \leq 0.52$ " and condition " $s \geq 0.52$ " to give a little different form fuzzy searching, and conditions " $s = 0.52$ and $c = 0.52$ " and conditions " $s = 0.52$ or $c = 0.52$ " have decreased by almost of 10x. From fuzzy searching, but the condition " $s = 0.52$ and $c = 0.52$ " do not give the results in data 10k and 50k because don't have association rules are supported = 0.52 and confidence = 0.52, but the fuzzy searching can give the results, such as the association rules have support = 0.51 and confidence = 0.53, support = 0.52 and confidence = 0.53, etc.

Therefore to search the association rule for match the users-requirement should be using non-fuzzy searching, but in some cases this approach cannot provide an answer. However, from that problem should be using the method are flexible and able to give an answer that is close to the most users-requirement this method is fuzzy searching.

V. CONCLUSION AND FUTURE WORK

This research proposed methods to search association rules with fuzzy technique from interestingness measures are the support and the confidence. The fuzzy searching are flexible and able to give an answer in some case that is close to the most users-requirement while the non-fuzzy searching do not. In particular, the use of condition "and" can find the rules on stricter conditions than the other conditions then the results are quite similar to the requirements. The conditions "Approximately less than" and "Approximately more than" they give the more results because they are quite flexible conditions. And other conditions the results will be good quality. For future work, we will use this technique to incorporate weight in the association rule discovery with constraint logic to optimize the number of association rules.

REFERENCES

[1] J. Ha and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2000.
 [2] G. I. Webb, "Efficient search for association rules," in *Proc. the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000, pp. 99-107.
 [3] B. Jeudy and J. Boulicaut, "Constraint-Based discovery and inductive query: application to association rule mining," *Pattern Detection and Discovery*, pp. 110-124, 2002.
 [4] R. T. Ng, L. V. S. Lakshmanan, J. Han, and A. Pang, "Exploratory mining and pruning optimizations of constrained association rules," in

Proc. 1998 ACM SIGMOD Int. Conf. Management of Data, 1988, pp. 13-24.
 [5] R. Srikant, Q. Vu, and R. Agrawal, "Mining association rules with item constraints," in *Proc. the 1997 ACM KDD*, 1997, pp. 67-73.
 [6] T. Trifonov and T. Georgieva, "Application for discovering the constraint-based association rules in an archive for unique bulgarian bells," *European Journal of Scientific*, vol. 31, pp. 366-371, 2009.
 [7] M. Kawsana and S. Nitsuwat, "Data support system for sales promotion analysis using fuzzy query technique," in *Proc. the 5th National Conference on Computer and Information Technology (NCCIT2009)*, 2009, pp. 684-689.
 [8] G. Bordogna and G. Psaila, "Extending SQL with customizable soft selection conditions," in *Proc. the 2005 ACM symposium on Applied computing*, 2005, pp. 1107-1111.
 [9] M. Hudec, "Fuzzy improvement of the SQL," in *Proc. the BALTICOR 2007 8th Balkan Conference on Operational Research*, 2007, pp. 257-267.
 [10] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proc. the 20th International Conference on Very Large Data Bases*, Santiago, Chile, 1994, pp. 487-499.
 [11] W. Siler and J. J. Buckley, *Fuzzy Expert Systems and Fuzzy Reasoning*. Wiley, 2005, ch. 3, pp. 29-54.



Phaichayon Kongchai is currently a doctoral student with the School of Computer Engineering, Suranaree University of Technology, Thailand. He received his bachelor degree in computer engineering from Suranaree University of Technology (SUT), Thailand, in 2010, and master degree in computer engineering from SUT in 2012. His current research includes constraint data mining, association mining, functional and logic programming languages, statistical machine learning.



Nittaya Kerdprasop is an associate professor at the School of Computer Engineering, Suranaree University of Technology, Thailand. She received her bachelor degree in radiation techniques from Mahidol University, Thailand, in 1985, master degree in computer science from the Prince of Songkla University, Thailand, in 1991 and doctoral degree in computer science from Nova Southeastern University, U.S.A, in 1999. She is a member of ACM and IEEE Computer Society. Her research of interest includes knowledge discovery in databases, artificial intelligence, logic programming, and intelligent databases.



Kittisak Kerdprasop is an associate professor and chair of the School of Computer Engineering, Suranaree University of Technology, Thailand. He received his bachelor degree in mathematics from Srinakarinwirot University, Thailand in 1986, master degree in computer science from the Prince of Songkla University, Thailand, in 1991 and doctoral degree in computer science from Nova Southeastern University, U.S.A. in 1999. His current research includes data mining, artificial intelligence, functional and logic programming languages, computational statistics.

The Compact Fuzzy Association Rules For Data Classification

Phaichayon Kongchai*, Keerachart Suksut, Rattaphong Sutamma,
Sak Phoemhansa, Nittaya Kerdprasop and Kittisak Kerdprasop

Data Engineering Research Unit, School of Computer Engineering, Suranaree University of Technology
111 University Avenue, Nakhon Ratchasima 30000 THAILAND

*Corresponding Author: Zaguraba_ii@hotmail.com

Abstract

Data classification mining is a method to find data generalization in a form of rules then used these rules to predict some unknown value in the future data. But in actual applications, the rules may be of low accuracy and the number of rules may be so overwhelmed that users could not efficiently apply them. Therefore, this research proposes the development of data classification algorithm with compact fuzzy association rules to optimize accuracy and interpretability of the model. To evaluate the performance of the proposed method, this research will compare accuracy of the classification model and the number of rules against 9 different data classification algorithms. The results showed that our CCFAR algorithm is comparable in terms of accuracy. When considering both accuracy and size of model, our algorithm is the best one.

Keywords: Data Classification, Associative Classification, Fuzzy Classification Association Rule, Fuzzy Set.

1. Introduction

Data classification technique is one of a data mining widely used to predict the future data by inducing the model from sophisticated data. For instance, in a medical science the model is used to predict a patient who is risky of having a breast cancer⁽¹⁾. In a commercial bank, the model is used to screen credit requests and evaluate the credit rating of consumer. For its potential benefits, many researches have long been concentrating on improving the efficiency of data classification by proposing algorithms like C4.5⁽²⁾ and OneR⁽³⁾.

But these algorithms have a low predictive accuracy. Therefore, Liu and Ma⁽⁴⁾ proposed a new method call the "associative classification" which is a combination between the association rule mining technique and the data classification technique. Their proposed algorithm was called the Classification based on Association Rules (CBA). The CBA has been tested its predictive performance by comparing with the C4.5, and it turns out that has high accuracy than C4.5 algorithm. This CBA has drawn attention from many researches to develop algorithm based on associative classification technique, such as GARC⁽⁵⁾ algorithm, OAC⁽⁶⁾ algorithm, and many more. However, the inherent problem of associative classification is that the association process can handle only symbolic and binary values. It cannot process continuous values. To solve this problem, we propose to use a fuzzy set to transform the continuous value to the degree of membership⁽⁷⁻⁹⁾.

In this paper, we present the idea and the development of an algorithm called data classification with compact fuzzy association rules (CCFAR). Our main focus is to optimize both an accuracy and interpretability of the model. In addition, we applied the concept of OneR⁽³⁾ algorithm to select the best rules and reduce the number of rules in the final result.

2. Related Works

Classification task is the mainstream of many researches are concentrating on developing algorithms for increasing the accuracy and reducing the number of rules. We can summarize the researches that are related to our works into 3 groups, that are group of data classification, group of associative classification, and group of fuzzy

associative classification. The details of 3 groups are as follows:

Firstly, the group of data classification is the initiative concept of data mining to classify target variable with some related features. Examples of this group are as follows:

- Quinlan⁽²⁾ proposed C4.5 algorithm which is well known in the classification because the algorithm is able to process quickly and the model is easily understandable. The main concept of this algorithm is the use of heuristic search to construct a decision tree and prune tree.

- Cohen⁽⁹⁾ proposed an algorithm called RIPPER (Repeated Incremental Pruning to Produce Error Reduction) developed from the IREP* algorithm by the principle of growing and pruning techniques to select the low error rate rules.

- Holte⁽³⁾ proposed an algorithm named OneR or One-Attribute Rule, which is an algorithm that easy-to-understand algorithm and the model is a compact set of rules. OneR is simple because it select a single attribute with the fewest error to build the model.

Secondly, the group of associative classification uses the association mining to build the association rules, then using many techniques of data classification to make association rules appropriate for classification. Examples of this group are as follows:

- Liu and Ma⁽⁴⁾ proposed CBA algorithm, which produced is a high accuracy for the data classification. The CBA composed of 2 steps: to build the model with CBA-RG (to create the association rules), and CBA-CB (to make association rules for classification by selecting the low error rate rules).

- Chen and Zhang⁽⁵⁾ proposed an algorithm called GARC (Gain Based Association Rule Classification) with a efficacy in classifying data used the model is a compact. Because of the GARC was using the information gain threshold to create frequent item-sets and then it used redundant and conflictive techniques to build the class association rules.

- Hu and Li⁽⁶⁾ proposed an algorithm named OAC (Optimal Association Classifier) with a good efficacy in classifying data. The principle of OAC was to generate association rules by using the constraint of apriori⁽¹⁰⁾ algorithm, then it selected the class association rules by the method call OCARM (Optimal Class Association Rule Mining⁽¹¹⁾).

Finally, the group of fuzzy associative classification uses the fuzzy set to controll the continuous value in the association rule mining processing. Examples of this group

are as follows:

- Pach et al. ⁽⁷⁾ proposed an algorithm called CFARC (Compact Fuzzy Association Rule-Based Classifier) to resolve the problem of how to define the minimum support and minimum confidence in association rule mining by applying fuzzy correlation threshold. In their experimental results, they showed a good accuracy and a compact set of model.

- Chen⁽⁸⁾ proposed an algorithm named CFAR (Classification with Fuzzy Association Rules). It generated rules by using the apriori algorithm, then it selected the rules that have highest confidence value and deleted remaining rules with lowest confidence value.

- Hühn and Hüllermeier⁽¹²⁾ proposed an algorithm called FURIA (An Algorithm For Unordered Fuzzy Rule) based on the Ripper algorithm. But the FURIA used unordered rule set to reduce bias of the target class and it used stretching technique to obtain the generalized rules.

3. Preliminaries

In this section, we introduce the basic definitions of fuzzy sets, fuzzy association rules and fuzzy associative classification rules.

3.1 Fuzzy Sets

Fuzzy sets are sets that cannot explain something clearly. For example, an explanation of the people who are very tall, someone say that the people who are 180 cm. high are very tall, but others may say that the people who are 185 cm. high are very tall. So, we can see from this example that it is impossible to tell exactly regarding who is very tall. Therefore, to solve such problem Zadeh⁽¹³⁾ proposed the concepts of fuzzy sets to explain something that is not clear with the degree of membership function (rang value of the degree is [0, 1]). For instance, the people who are 180 cm. high; they are medium tall at the degree of 0.7 and very tall at the degree of 0.3. The people who are 185 cm high, they are medium tall at the degree of 0.1 and very tall at the degree of 0.9. From this example, we can see the people who are 180 cm. high and 185 cm. high would be moderately tall and very tall, respectively with the different degrees.

Therefore, this research applied the concept of fuzzy sets to explain the continuous value of numeric data. We used fuzzy partitions with fuzzy c-means (FCM⁽¹⁴⁾) algorithm.

3.1.1 Fuzzy partitions

In this section, we present the important step of our CCFAR algorithming for transform the continuous value (Fig. 1 a) to partitions with the FCM algorithm. The FCM is used to indicate the degree of membership in a set with the value ranging from 0 to 1 (Fig. 1 c), which is different from the k-means algorithm that indicates the degree of membership in a set with the discrete value (Fig. 1 b).

The processes of FCM are composed four main steps.

- 1) Initialize C and μ_{ij}
- 2) Calculate the center vectors by Eq.(1)

$$c_j^{(t)} = \frac{\sum_{i=1}^N \mu_{ij}^{(t)m} x_i}{\sum_{i=1}^N \mu_{ij}^{(t)m}} \quad (1)$$

- 3) Compute and update membership of data by Eq.(2)

$$\mu_{ij}^{(t+1)m} = \frac{1}{\sum_{k=1}^C \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \quad (2)$$

- 4) If $\|\mu_{ij}^{(t+1)m} - \mu_{ij}^{(t)m}\| < \varepsilon$ then STOP; otherwise repeat step 2.

Where x_i is data vector, m is fuzziness that can be any real number equals or greater than 1, μ_{ij} is the degree of membership of x_i to be in the cluster j , N is data for clustering, C is number of clusters and c_j is center of cluster j .

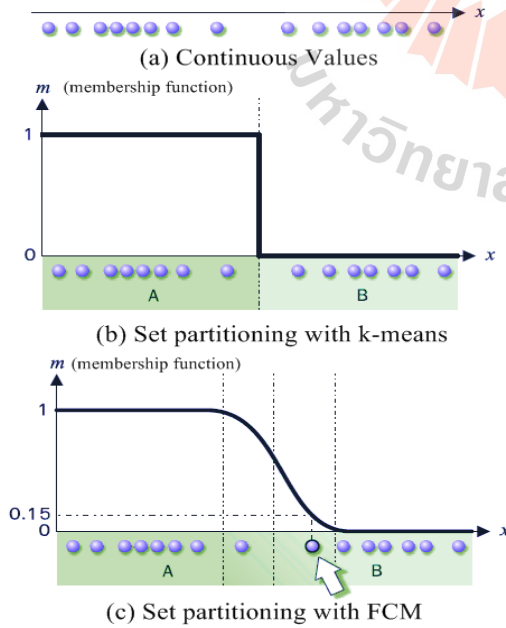


Fig. 1. The membership functions of k-means and FCM.

3.2 Fuzzy Association Rule

Association rule is originally an analysis of customer purchases (also called Market Basket Analysis) by storing the items in the basket of the customer into transaction. Then, it analyzed these transactions to discover the association rules. The association rules are expressions of the type $X \rightarrow Y$, where X and Y are sets of items. This means that if customers buy item X , then customers buy item Y together. But the processing of the association rule mining is unable to handle the continuous value (Table 1). So, we used fuzzy set to convert the continuous value into a degree of belonging to a set. For example, the Table 2 shows the degree of attribute age associated with linguistic values low, medium and high.

In recent years, many researches have proposed methods to mining fuzzy association rules from continuous value^(15,16). The processing of fuzzy association rules are composed of 2 steps. First step, create frequent Item-sets by counting item-sets that have fuzzy support (Eq.3) higher than the minimum fuzzy support. Second step, create fuzzy association rules from frequent item-sets that have fuzzy confidence (Eq.4) higher than the minimum fuzzy confidence.

$$FS(\langle Z:A \rangle) = \frac{\sum_{k=1}^N \prod_{\langle z_i:A_{i,j} \rangle \in \langle Z:A \rangle} t_k(z_i)}{N} \quad (3)$$

$$FC(\langle X:A \rangle \rightarrow \langle Y:B \rangle) = \frac{FS(\langle X:A \rangle \cup \langle Y:B \rangle)}{FS(\langle X:A \rangle)} \quad (4)$$

Let $\langle Z:A \rangle$ be a fuzzy item-set denoted as $\langle Z:A \rangle = [\langle z_{i1}:A_{i,j} \rangle \cup \langle z_{i2}:A_{i,j} \rangle \cup \dots \cup \langle z_{iq}:A_{i,q,j} \rangle]$, where $q \leq n + 1$, z_i is an attribute, such as Age, Income and Balance, $A_{i,j}$ is a fuzzy interval, such as Low, Medium and High, t_k is a transaction, and N is the number of transactions.

Table 1. The example of data.

Id	Age
1	18
2	20
3	19
4	24
5	25

Table 2. The degrees of attribute age with FCM algorithm.

Id	Age		
	Age = Low	Age = Medium	Age = High
1	0.9863	0.0127	0.0010
2	0.0119	0.9862	0.0019
3	0.4996	0.4900	0.0104
4	0.0074	0.0141	0.9785
5	0.0053	0.0090	0.9857

3.3 Fuzzy Associative Classification Rule

Fuzzy associative classification rule and Fuzzy association rule are similar techniques in terms of processing steps, but they differ at consequent part of rule. The fuzzy associative classification rule must contain only one class label at consequent part of rule ($C = \{C_1, \dots, C_k\}$). Therefore, the fuzzy confidence of fuzzy associative classification rule can be defined as follows:

$$FS(\langle Z:A \rangle) = \frac{\sum_{k=1}^N \prod_{\langle z_i, C_k : A_i, C_k, j \rangle \in \langle Z:A \rangle} t_k(z_i, C_k)}{N} \quad (5)$$

$$FC(\langle X:A \rangle \rightarrow \langle Y:C \rangle) = \frac{FS(\langle X:A \rangle \cup \langle Y:C \rangle)}{FS(\langle X:A \rangle)} \quad (6)$$

4. Our Proposed Methodology: CCFAR

In this section, we described our algorithm that has been named data classification algorithm with compact fuzzy association rules (CCFAR) to obtain the fuzzy associative classification rules with optimum combination of accuracy and interpretability of the model. Our methodology is composed of five steps (Fig 2).

1) Data Screening: The data characteristic of CCFAR must be numeric data because of the transformation of fuzzy set that will convert numeric data only. At current stage, our algorithm assumes that there is no missing value in the data.

2) Data Partitioning: This step is transforming the numeric data to fuzzy intervals or fuzzy sets. In the CCFAR algorithm, we used FCM to make the transformation because of its easy-to-understand property and high performance. As an example, we used the data from Table 1 and we defined fuzzy intervals to be 3 partitions; the results were represented in table 2.

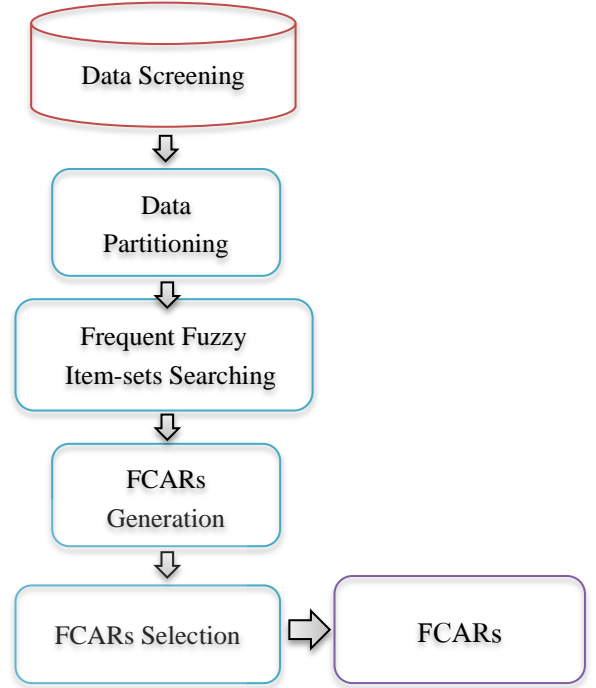


Fig. 2. The development of data classification algorithm with compact fuzzy association rules.

3) Frequent Fuzzy Item-Sets Searching: The fuzzy frequent item-sets are fuzzy item-sets that have fuzzy support more than minimum fuzzy support (γ). At this step, the process are the same as CFARC⁽⁷⁾ algorithm (Fig. 3).

4) Fuzzy classification association rule (FCARs) generation: This step is similar to the CFARC⁽⁷⁾ algorithm (Fig. 4), but it differed in the computation of score values of all rules. In this research we computed score values of all rules by using our new metric as shown in Eq.9 fuzzy correlation (Eq.7), fuzzy confidence (Eq.6) and firing strength(β) (Eq.8). The intuitive idea is that we need a high correlated, high confidence, as well as high support rules.

$$FCORR(\langle X:A \rangle \rightarrow \langle Y:C \rangle) = \frac{FS(\langle X:A \rangle \cup \langle Y:C \rangle) - FS(\langle X:A \rangle) \times FS(\langle Y:C \rangle)}{\sqrt{FS(\langle X:A \rangle) \times (1 - FS(\langle X:A \rangle)) \times FS(\langle Y:C \rangle) \times (1 - FS(\langle Y:C \rangle))}} \quad (7)$$

$$\beta(\langle X:A \rangle \rightarrow \langle Y:C \rangle) = \sum_{k=1}^N \prod_{\langle z_i: A_{i,j} \rangle \in \langle Z:A \rangle} t_k(z_i) \quad (8)$$

$$SCORE = FCORR \times FC \times \text{Firing_strength}(\beta) \quad (9)$$

Frequent fuzzy item set searching (an Apriori fuzzy implementation)

Input: DF fuzzy data

Output: the set of frequent fuzzy item set

Method:

1. Determine the supports of the classes by the distribution of classes;
2. Set the minimal fuzzy support (γ) to the half of the minimum frequency of classes;
3. Generate the 1- candidate fuzzy items;
4. Calculate FS values then select the frequent fuzzy items from the 1- candidate which has $FS > \gamma$, and $n = 2$;
5. While there exist some $n-1$ size frequent item sets: Generate the n -size candidate sets from $n-1$ size frequents (and 1-size frequents);

Fig. 3. The frequent fuzzy item set searching⁽⁷⁾.

Fuzzy classification association rule generation

Input: a set of frequent fuzzy item sets

Output: positive correlated FCARs separated by size

Method:

1. Generate association rules with class label consequent from all the frequent item sets to consider the size of item sets
2. Calculate the Score values of all the rules;
3. Select rules with positive SCORE value for all size;

Fig. 4. The fuzzy classification association rule generation⁽⁷⁾

5) Fuzzy Classification Association Rule Selection:

This step is selecting the FCARs to be used in the data prediction. It's composing of 4 parts as follows (Fig.5):

- Frist part: The rules that have the same class label and the same size will be grouped together. After that, this algorithm will select rule with the highest SCORE from each group.

- Second part: The rules with the highest SCORE in each group will be counted the frequency of attributes in antecedent of rule. For example:

From Table 3, the frequency of attribute Age is 5 (No. 1, 2, 3, 4 and 5), attribute Inc is 3 (No. 3, 5 and 6) and attribute Bal is 1 (No. 2). Thus, the highest frequent attribute is the attribute Age. *If frequencies of attributes are equal, attributes will be randomly selected.

Fuzzy classification association rule selection

Input: a set of FCARs

Output: the Compact FCARs

Method:

1. Select rules with the highest SCORE for each class in each size of the rules;
2. Find maximum frequent attribute from step 1 then select OneR;
3. Select highest SCORE and shortest rule from a set of FCARs with OneR. (If a set of FCARs does not have a set of OneR then create new rule);
4. Remove redundancy rules by select the shortest rule;

Fig. 5. The fuzzy classification association rule selection.

- Third part: A set of FCARs with the highest SCORE and shortest rule from each fuzzy interval is selected. For examples, the rules from Table 3 that contain attribute Age (the highest frequency) are No.1, 2, 3, 4 and 5. After that, it selects the FCARs from the rules No. 1-5 with the highest SCORE and the shortest rule, which are rule No. 1 (Shortest rule of Age = high), 2 (Highest SCORE of Age = high), and 4 (highest SCORE and shortest rule of Age = medium). But this selection is incomplete because the rule that contains Age = low is missing. Therefore, this algorithm will create new rule that contains Age = low with don't cared support and SCORE values, such as:

Rule 1: Age = low \rightarrow yes, SCORE = -0.12

Rule 2: Age = low \rightarrow no, SCORE = 0.12

In this algorithm, the Rule 2 will be selected because it has positive SCORE value. Thus, we can conclude the results as follows:

1. Age = high \rightarrow yes
2. Age = high and Bal = low \rightarrow no
3. Age = medium \rightarrow no
4. Age = low \rightarrow no

- Fourth part: we will remove rules that are superset in antecedent part (left-hand-side) with the same class as other rules. For instance,

Rule 1: Age = low \rightarrow yes,

Rule 2: Age = low and Bal = high \rightarrow yes

From example, we removed the Rule 2 because of it is superset in antecedent part and has same class of Rule 1.

Table 3. The FCARs with SCORE.

No.	FCARs	SCORE
1	Age = high → yes	1.2725
2	Age = high and Bal = low → no	1.9225
3	Age = high and Inc = me → no	1.1619
4	Age = medium → no	1.2608
5	Age = medium, Inc = high → no	0.8432
6	Inc = high → yes	1.1232

Table 4. Data sets.

Data Set	Size	# Attribute	# Class
Iris	150	4	3
Heart	270	13	2
Pima	768	8	2

5. Experimentation and Results

In this section, we evaluated our CCFAR algorithm using three measurements accuracy: Acc (Eq.10), number of the models (Compact Value: CV) that we normalized it to be the Normalization of Compact Value: NCV (Eq.11), and the combination of Acc and NCV in which we call Suitable Rule: SR (Eq.12). Our algorithm was tested on three different data sets and compared with nine different

algorithms. These algorithms are grouped as: data classification, associative classification and fuzzy associative classification. The nine algorithms are namely C4.5, RIPPER, OneR, CBA, GARC, OAC, FURIA, CFAR, and CFARC. The data are taken from the UC Irvine Machine Learning Repository, namely Iris, Heart, and Pima (details shown in Table 4). The classification performances of all algorithms were measured by ten-fold cross validation.

$$Acc = \frac{TP+TN}{TP + TN + FP + FN} \quad (10)$$

Where TP is the number of true positive examples, FP is the number of false positive examples, TN is the number of true negative examples and FN is the number of false negative examples.

$$NCV_{AL} = \frac{Avg(CV)_{max} - Avg(CV)_{AL}}{Avg(CV)_{max} - Avg(CV)_{min}} \quad (11)$$

$$SR_{AL} = \frac{Avg(Acc)_{AL} + NCV_{AL}}{2} \quad (12)$$

Let Avg(CV) is an average of compact value, AL is an Algorithm.

Table 5. The Accuracy (Acc) of classification of CCFAR algorithm and other algorithms.

Data Set	CLASS			AC			FCAR			
	C4.5	RIPPER	OneR	CBA	GARC	OAC	FURIA	CFAR	CFARC	CCFAR*
Iris	0.933	0.933	0.940	0.929	0.960	0.940	0.947	0.913	0.959	0.960
Heart	0.770	0.822	0.729	0.815	0.880	0.811	0.797	-	0.774	0.781
Pima	0.734	0.747	0.724	0.724	0.762	0.781	0.747	0.651	0.729	0.747
Avg(Acc)	0.812	0.834	0.797	0.822	0.867	0.844	0.830	0.782	0.820	0.828
Rank.	8	3	9	6	1	2	5	10	7	4

Table 6. The number of classification rules (Compact Value) of CCFAR algorithm and other algorithms.

Data Set	CLASS			AC			FCAR			
	C4.5	RIPPER	OneR	CBA	GARC	OAC	FURIA	CFAR	CFARC	CCFAR*
Iris	4.9	3.6	3.0	5.0	7.0	9.0	4.4	9.1	3	3
Heart	17.4	4.1	2.0	52.0	12.0	157.0	8.4	-	2.7	3
Pima	20.3	4.1	7.8	45.0	6.0	112.0	8.5	2.0	2	4
Avg(CV)	14.2	3.9	4.2	34	8.3	92.6	7.1	5.5	2.6	3.3
NCV	0.86	0.983	0.980	0.64	0.93	0	0.94	0.96	1	0.99
Rank.	8	3	4	9	7	10	6	5	1	2

Table 7. The complete rules of CCFAR algorithm and other algorithms.

Data Set	CLASS			AC			FCAR			
	C4.5	RIP PER	OneR	CBA	GARC	OAC	FU RIA	CFAR	CFA RC	CCFAR*
SR	0.840	0.908	0.888	0.735	0.900	0.422	0.889	0.873	0.9100	0.9104
Rank.	8	3	6	9	4	10	5	7	2	1

The results in Tables 5-7 were summarized and discussed as follows:

- Table 5 shows the accuracy of classification of CCFAR algorithm compared to the 9 algorithms. The symbol “-” means no available published result. GARC algorithm shows the highest accuracy (0.867) in all dataset and it has been ranked number 1 in the comparison among classifier data from 10 algorithms. Our proposed CCFAR algorithm was ranked number 4 (0.828) in classification accuracy.

- Table 6 represents the number of classification rules of CCFAR algorithm and the other 9 algorithms. The measure NCV is a normalization of the compact values computed as in Eq.8. The CFARC algorithm gives the smallest number of rules (2.6) comparing from 10 algorithms. Our CCFAR algorithm was ranked number 2 with average number of rules equals (3.3).

- From table 7, we showed the combination of accuracy and normalization of compact value which is called the SR values of CCFAR algorithm and other nine algorithms. We can see that the best SR is our presented method CCFAR (0.9104). It means our algorithm is a good classifier in terms of both accuracy and good compact of fuzzy associative classification rules considered all together.

6. Conclusions

This paper proposes the development of data classification algorithm with compact fuzzy association rules called (CCFAR) to optimize both an accuracy and interpretability of the classification model. To evaluate the performance of the proposed method, our algorithm was tested on three different data sets and compared the results with the other nine different algorithms. The results showed that our proposed CCFAR algorithm was ranked number 3 based on accuracy of classification, and it was ranked number 2 based on the smallest number of rules. CCFAR is the best algorithm when we combine accuracy and compact value together.

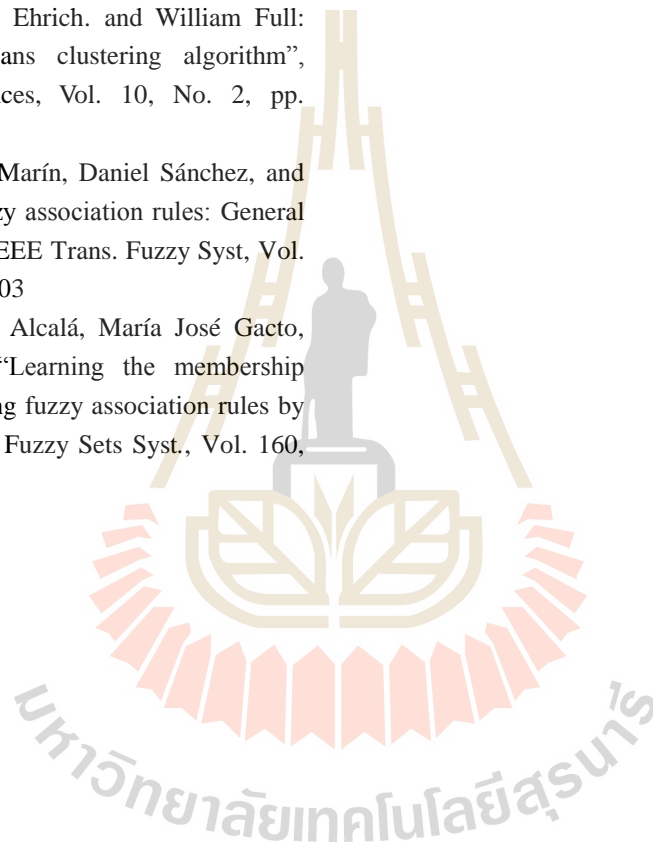
Acknowledgment

The first author has been funded by scholarship from the Suranaree University of Technology.

References

- (1) Bellaachia Abdelghani, and Guven Erhan: “Predicting breast cancer survivability using data mining techniques”, In Proceedings of 2nd International Conference on Software Technology and Engineering (ICSTE), Vol. 58, No. 13, pp. 10-110, 2006
- (2) Quinlan J. Ross: “C4.5: programs for machine learning”, Morgan Kaufmann, Vol. 1, 1992
- (3) Holte Robert C: “Very simple classification rules perform well on most commonly used datasets”, Machine learning, Vol. 11, No. 1, pp. 63-90, 1993
- (4) Bing Liu, Wynne Hsu, and Yiming Ma: “Integrating classification and association rule mining”, In Proceedings of the 4th American Association for Artificial Intelligence, 1998
- (5) Guoqing Chen, Hongyan Liu, Lan Yu, Qiang Wei, and Xing Zhang: “A new approach to classification based on association rule mining”, Decision Support Systems. 42(2), pp. 674-689, 2006
- (6) Hu Hong, and Li Jiuyong: “Using association rules to make rule-based classifiers robust”, In Proceedings of the 16th Australasian database conference, Vol. 39, pp. 47-54, 2005
- (7) Ferenc Péter Pach, Attila Gyenesei b, and Janos Abonyi: “Compact fuzzy association rule-based classifier”, Expert systems with applications, Vol. 34, No. 4, pp. 2406-2416, 2008
- (8) Chen Zuoliang. and Guoqing Chen: “Building an associative classifier based on fuzzy association rules”, International Journal of Computational Intelligence Systems, 1(3), pp. 262-273, 2008
- (9) Cohen W. William: “Fast effective rule induction”, In Proceedings of the Twelfth International Conference on Machine Learning, pp. 1995

- (10) Bayardo Jr. Roberto, Rakesh Agrawal, and Dimitrios Gunopulo: "Constraint-based rule mining in large, dense databases", In Proceedings of 15th International Conference on Data Engineering, pp. 188-197, 1999
- (11) Li Jiuyong Hong Shen, and Rodney Topor: "Mining the optimal class association rule set", Knowledge-Based Systems. 15(7), pp. 399-405, 2002
- (12) Hühn Jens, and Eyke Hüllermeier: "FURIA: an algorithm for unordered fuzzy rule induction", Data Mining and Knowledge Discovery, 19(3), pp. 293-319, 2009
- (13) Lotfali Askar Zadeh: Fuzzy sets. Information and control, Vol. 8, No. 3: pp. 338-353, 1965
- (14) James C. Bezdek, Robert Ehrlich. and William Full: "FCM: The fuzzy c-means clustering algorithm", Computers and Geosciences, Vol. 10, No. 2, pp. 191-203, 1984
- (15) Miguel Delgado, Nicolás Marín, Daniel Sánchez, and María-Amparo Vila: "Fuzzy association rules: General model and applications," IEEE Trans. Fuzzy Syst, Vol. 11, No. 2, pp. 214–225, 2003
- (16) Jesús Alcalá-Fdez, Rafael Alcalá, María José Gacto, and Francisco Herrera: "Learning the membership function contexts for mining fuzzy association rules by using genetic algorithms," Fuzzy Sets Syst., Vol. 160, No. 7, pp. 905–921, 2009.



Single Versus Multiple Measures for Fuzzy Association Rule Mining

Keerachart Suksut*, Phaichayon Kongchai, Sak Phoemhansa,
Rattaphong Sutamma, Kittisak Kerdprasop, Nittaya Kerdprasop

Data Engineering Research Unit, School of Computer Engineering, Suranaree University of Technology
111 University Avenue, Nakhon Ratchasima 30000 THAILAND.

*Corresponding Author: mikaiterng@gmail.com

Abstract

Data mining can identify patterns of data, find relationships within the data to predict the outcome or predict future data trends. Association rule mining is part of data mining and it has been applied in many fields. The performance of association rule mining depends on the support and confidence measures. In this research, we perform a comparative of number of rules that is affected from different kinds of measures. We propose the idea of using a combination of measures (such as support*confidence, confidence*lift) instead of considering only support value or confidence. The main objective is to reduce the number of rules in the process of fuzzy association rule mining, in which a compact rule set is important to the real application of the fuzzy system. The experimental results reveal that our idea of combining multiple measures for fuzzy association rule mining can significantly reduce the number of association rules.

Keywords: Data mining, Association Rule, Fuzzy Association rule mining, Measure for Association Rules.

1. Introduction

At present, data mining has been very popular because it helps extract, search, predict, and many other knowledge-intensive tasks. Usefulness is from employing existing data to predict future trends in data, extracting patterns of the data, and finding the relationship within the data group. These benefits make data mining been widely applied. In medicine, it is used in the data analysis of the patient to see a trend whether or not the patient is likely to be ill from a disease. The business side also uses data mining to analyze the buying patterns of consumers.

Data mining techniques can be practiced in various ways. The technology that has been adopted widely is association rule mining in relation to the incident and bring those things that occur together to create the association rules.

Finding association rules was proposed by Agrawal⁽¹⁾ to be used in data mining. To find efficiently association rules a criterion for calculating the support of rule is necessary. A confidence criterion is also needed to reduce the number of rules.

The most relevant research to ours is to reduce the number of rules in different ways, such as the work of F. P Pach et al.⁽²⁾. They proposed the concept to reduce the number of fuzzy association rules with pruning technique, which makes the association rules smaller in their size. The research of M. M. Ballesteros et al.⁽³⁾ aimed to improve the algorithm and they compared the fitness function obtained from the various developed algorithms as well as there is quite a few the traditional algorithm. However, research work that compares the number of rules based on different measures, especially a multi-criteria measure. Therefore, we propose this research that was performed to compare the number of rules for different measures and also apply more than one measures to reduce the number of discovered association rules.

2. Background

2.1 Association Rule Mining

Association rule mining is a process that gained much popularity in data mining. The process to find relationships that hidden in a dataset can be explained with the following example.

From Table 1 which is a set of ten transactions, the

frequency of the purchases for each product type in association to other products are shown in table 2.

From Table 2, one can create an instance of the relationships that can be represented as “if antecedent then consequence” (or “antecedent \Rightarrow consequence”) as follows:

- If customer buy milk, then they also buy bread.
- If customer buy beer, then they also buy potatoes.
- If customer buy bean, then they also buy potatoes.

Table 1 Transactional database of customer’s purchases.

Transaction ID	Item
1	Milk, Bread, Butter, Water
2	Milk, Bread, Cola
3	Beer, Bean, Potatoes
4	Milk, Bread, Water
5	Water, Cola, Beer
6	Milk, Butter, Potatoes
7	Cola, Potatoes, Bean
8	Milk, Bread
9	Milk, Bread, Potatoes
10	Beer, Bean, Potatoes

Table 2 Frequency of each combination of two product types.

	Mil k	Brea d	Butte r	Wate r	Col a	Bee r	Bea n	Potatoe s
Milk	6*	5	2	2	0	0	0	2
Bread	5	5*	1	2	1	0	0	0
Butter	2	1	1*	1	0	0	0	1
Water	2	2	1	3*	1	1	0	0
Cola	0	1	0	1	3*	1	1	1
Beer	0	0	0	1	1	3*	2	2
Bean	0	0	0	0	1	2	3*	3
Potatoe s	2	0	1	0	1	2	3	4*

*Purchases counted as one-item frequency.

Table 3 Measures for finding support of rules.

Measure	Equation	Range
support(X)	$n(X) / N$	[0,1]
support(X \Rightarrow Y)	$n(X \cap Y) / N$	[0,1]
confidence(X \Rightarrow Y)	$support(X \Rightarrow Y) / support(X)$	[0,1]
lift(X \Rightarrow Y) ⁽⁴⁾	$support(X \Rightarrow Y) / (support(X) * support(Y))$	[0,∞]
conviction(X \Rightarrow Y) ⁽⁵⁾	$(1 - support(Y)) / (1 - confidence(X \Rightarrow Y))$	[0,∞]
gain(X \Rightarrow Y) ⁽⁶⁾	$confidence(X \Rightarrow Y) - support(Y)$	[-0.5,1]
leverage(X \Rightarrow Y) ⁽⁷⁾	$support(X \Rightarrow Y) - (support(X) * support(Y))$	[-0.25,0.25]

2.2 Different Measures of Association Rule Mining

(a) Support measure

Support measure is frequency of each purchased item. This measure supports usefulness of association rules. The higher the frequent that association occur. The boundary of support measure ranges between 0 to 1 and can be computed as in the following equation:

$$support(X \Rightarrow Y) = \frac{n(X \cap Y)}{N} \quad (1)$$

where:

$n(X \cap Y)$ is the number of time that the item X happened together with Y in the same transaction.

N is the total number of transaction.

Example 1: finding the value of support of rule with the support measure.

From Table 1 when using the support measure to find the value of support of rule that customers buy milk and also buy bread, the computation is as follows:

$$support(milk \Rightarrow bread) = \frac{5}{10} = 0.5$$

That means the value of support of rule “if milk then bread” is 0.5, or 50% of the whole transactions.

(b) Confidence measure

Confidence measure is the measure to indicate the reliability of the rule. The boundary of confidence measure is between 0 to 1 and can be computed as follows:

$$confidence(X \Rightarrow Y) = \frac{support(X \Rightarrow Y)}{support(X)} \quad (2)$$

where:

support(X \Rightarrow Y) is the frequency that both items X and Y occur in the same transaction,

support(X) is the support that X has happened counted from all transactions.

Example 2: finding the value of confidence of rule with the confidence measure.

From Table 1 when using the confidence measure to find the value of confidence of rule that customers buy milk and also buy bread, the computation is as follows:

$$confidence(milk \Rightarrow bread) = \frac{0.5}{0.6} = 0.833$$

That means the value of confidence of rule that if milk is bought then bread also be bought is 0.833, or 83% of confidence.

(c) Lift measure

Lift measure has its boundary of value between 0 to infinity. If the value is less than 1, it means that X and Y are related in a negative way. If a value equal 1, it means that X and Y are independent. If a value is greater than 1, it indicates that X and Y are correlated in a positive way. The computation is shown in the following equation:

$$lift(X \Rightarrow Y) = \frac{support(X \Rightarrow Y)}{support(X) * support(Y)} \quad (3)$$

where:

- support(X⇒Y) is the frequency that X and Y occur in the same transaction,
- support(X) is the support that X happened against all transactions,
- support(Y) is the support that Y happened against all transactions.

Example 3: finding the value of lift of rule with the lift measure.

From Table 1 when using the lift measure to find the lifted confidence that customers who buy milk also buy bread, the lift value computation can be illustrated as follows:

$$lift(milk \Rightarrow bread) = \frac{0.5}{0.6 * 0.5} = 1.667$$

That means the value of lift of rule milk ⇒ bread is 1.667; that is, milk and bread have a positive relationship.

(d) Conviction measure

Conviction measure has its boundary value between 0 to infinity. If the value is less than 1, it means that X and Y are related in a negative way. If a value equal 1, it means that X and Y are independent. If a value is greater than 1, it indicates that X and Y are correlated in a positive way. The conviction measure can be computed as in the following equation:

$$conviction(X \Rightarrow Y) = \frac{(1 - support(Y))}{(1 - confidence(X \Rightarrow Y))} \quad (4)$$

where:

- support(Y) is the support value that Y happened against all transactions,
- confidence(X⇒Y) is the reliability of the X and Y co-occurring in the same transaction.

Example 4: finding the value of conviction of rule with the conviction measure.

From Table 1 when using the conviction measure to find the conviction of rules that customers who buy milk also buy bread, the computation is as follows:

$$conviction(milk \Rightarrow bread) = \frac{(1 - 0.5)}{(1 - 0.833)} = 2.99$$

That means the conviction of rule that if milk is bought then bread is also bought is 2.99. That means milk and bread have a positive relationship.

(e) Gain measure

Gain measure is a measure that is used to compute added value or change of rules. The boundary of gain measure is between -0.5 to 1. Its computation is as follows:

$$gain(X \Rightarrow Y) = confidence(X \Rightarrow Y) - support(Y) \quad (5)$$

where:

- support(Y) is the support that Y happened against all transactions,
- confidence(X⇒Y) is the reliability of the X and Y co-occurring in the same transaction.

Example 5: finding the gain value of rule with the gain measure.

From Table 1 when using the gain measure to compute usefulness of rules that customers who buy milk also buy bread, the computation is as follows:

$$gain(milk \Rightarrow bread) = 0.833 - 0.5 = 0.333$$

That means the gain value of rule that if milk is bought then bread is also bought is 0.333.

(f) Leverage measure

Leverage measure is a measure that indicates the strength of the rule. If the leverage is less than 0, it means that X and Y are related in a negative way. If the value is equal to 0, it means that X and Y are independent, and any value greater than 0 indicates that X and Y are positively dependent. The computation is as follows:

$$leverage(X \Rightarrow Y) = support(X \Rightarrow Y) - (support(X) * support(Y)) \quad (6)$$

where:

- support(X⇒Y) is the support that X and Y co-occurring in the same transaction,
- support(X) is the support that X happened against all transactions,
- support(Y) is the support that Y happened against all transactions.

Example 6: finding the value of leverage of rule with the leverage measure.

From Table 1 when using the leverage measure to evaluate the rule $milk \Rightarrow bread$, the computation is as follows:

$$leverage(milk \Rightarrow bread) = 0.5 - (0.6 * 0.5) = 0.2$$

That means the value of leverage of rule that if milk is bought then bread is also bought is 0.2, which implies milk and bread have a positive relationship.

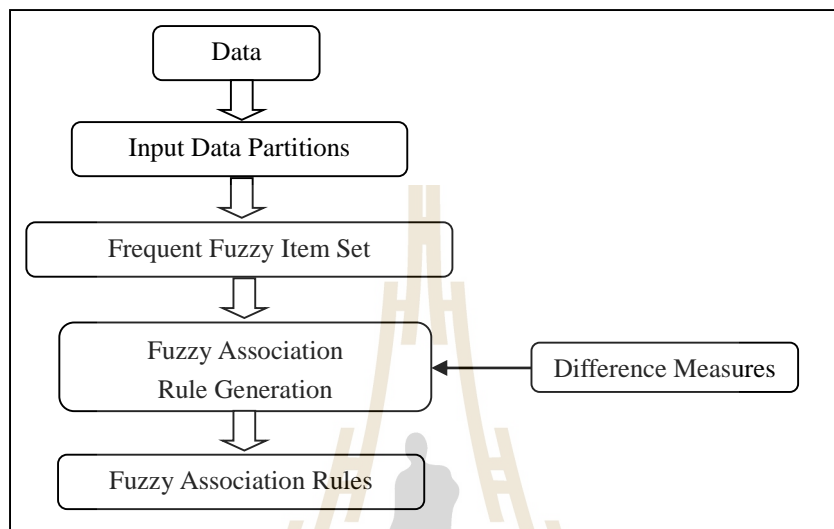


Figure 1 A research framework for comparing the effect of different measure for fuzzy

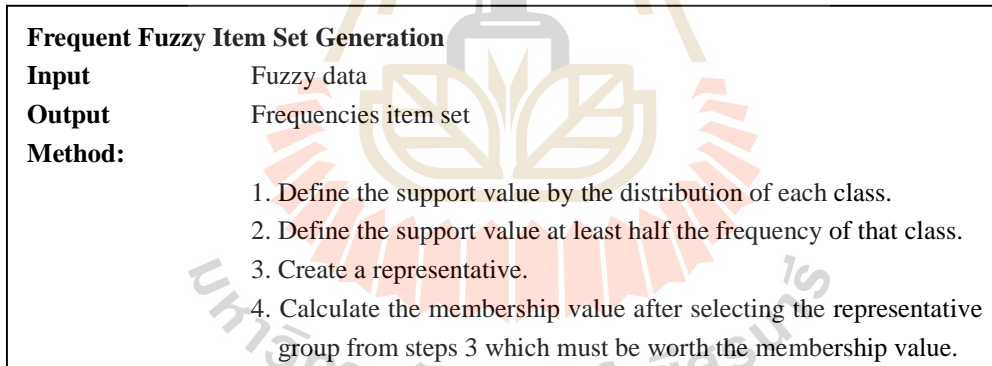


Figure 2 Frequent fuzzy item set algorithm⁽²⁾

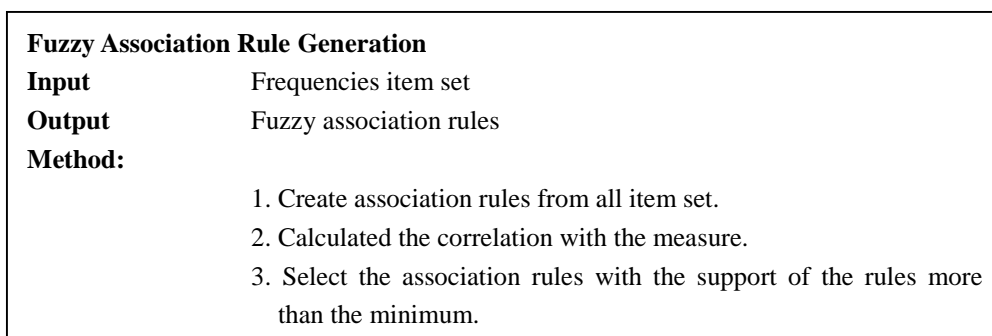


Figure 3 Fuzzy association rule-base generation algorithm⁽²⁾

3. Methodology

Researchers have designed the process and findings of the comparison of single measure and multi-measure for the discovery of fuzzy association rules as show in Figure 1.

From figure 3 can explain the process of the research are as follows:

- Input data partitions

In the process of the associative-based classification is the partitioning of an attribute whose value is continuous. It is a transformation of continuous values to be discrete intervals based on the concept of fuzzy sets.

- Frequent fuzzy item set generation

At this process, we will search and count the frequency of items and combination of items. These items are to be used in the creation of rules, regardless of the support of the rule. The procedure is shown in the Figure 2.

- Fuzzy association rule generation

In this process, we will create the rules that are screened by different measures and will vary various minimum support thresholds in order to create various final rule sets. The procedure works as shown in figure 3.

4. Experimental Results

This study used Breast Cancer Wisconsin data that are obtained from the UCI Machine Learning Repository. This dataset has 699 instances with 10 attributes and a target attribute of two classes. We used this data to find fuzzy association rules with various single measures and different combinations of two measures. The support of the rules has been set to be 90%, 80%, 70% 60%, and 50%. The setting of each measuring value is summarized and shown in Table 4. The experimental results are shown in Tables 5-7.

The results shown in Table 5 are the comparison of number of rules obtained from using different single measure with the support of rules ranging from 90%, 80%, 70%, 60%, and 50%. It can be seen that the each single measure gives a number of fuzzy association rules different from the value in each criteria. There are some measures of screening fuzzy association rules that appeared to produce rules less than the minimum support of the rule. We noticed that the number of rules in each measure is not equal.

Table 4 Percentage value for each measure used in the experimentation compared with the boundary value for each measure.

	50%	60%	70%	80%	90%	100%
support	0.5	0.6	0.7	0.8	0.9	1.0
confidence	0.5	0.6	0.7	0.8	0.9	1.0
lift	0.5	0.6	0.7	0.8	0.9	1.0
conviction	0.5	0.6	0.7	0.8	0.9	1.0
gain	0.475	0.58	0.685	0.79	0.895	1.0
leverage	0	0.05	0.1	0.15	0.2	0.25

Table 5 Number of rules and average values obtained of each single measure (SOR is minimum support of rule).

		Measure					
		Support	Confidence	Gain	Leverage	Lift	Conviction
SOR = 90%	Number of Rules	0	105	0	0	252	21
	Average	0	0.9392	0	0	1.327	2.0459
SOR = 80%	Number of Rules	0	105	0	84	252	21
	Average	0	0.9392	0	0.1657	1.327	2.0459
SOR = 70%	Number of Rules	0	238	0	210	252	21
	Average	0	0.8801	0	0.1487	1.327	2.0459
SOR = 60%	Number of Rules	22	252	0	252	252	21
	Average	0.6293	0.8680	0	0.1351	1.327	2.0459
SOR = 50%	Number of Rules	252	252	0	252	252	21
	Average	0.5651	0.8680	0	0.1351	1.327	2.0459

Table 6 Number of rules and average values obtained from a combination of support and other five criteria (SOR is minimum support of rule).

		Measure				
		Support* Confidence	Support* Gain	Support* Leverage	Support* Lift	Support* Conviction
SOR = 70%	Number of Rules	0	0	8	216	21
	Average	0	0	0.1009	0.7630	1.2289
SOR = 60%	Number of Rules	4	0	210	252	21
	Average	0.625	0	0.0831	0.7479	1.2289
SOR = 50%	Number of Rules	108	0	252	252	21
	Average	0.5356	0	0.0759	0.7479	1.2289

Table 7 Number of rules and average values obtained from a combination of confidence and other criteria (SOR is minimum support of rule).

		Measure				
		Confidence * Support	Confidence * Gain	Confidence * Leverage	Confidence * Lift	Confidence * Conviction
SOR = 70%	Number of Rules	0	0	208	252	21
	Average	0	0	0.1307	1.1537	1.4265
SOR = 60%	Number of Rules	4	0	234	252	21
	Average	0.6250	0	0.1236	1.1537	1.4265
SOR = 50%	Number of Rules	108	0	208	252	21
	Average	0.5356	0	0.1180	1.1537	1.4265

When using more than one measures for determining the number of fuzzy association rules, we set the support of the rule from 70% were not able to give any rule.

It can be seen from Table 6 that a comparison in terms of of number of rules obtained from using different 2 measures by using support measure as the main measure and varying support of the rule from 70% down to 60% and 50% can produce a small set of fuzzy association rules. Using more than a single measure as the condition in the selection of rules will result in less fuzzy association rules because the support of rules when combined with other calculation criteria results in the decrease of the rules that satisfy the threshold.

When we use more than a single measure to screen rules using the confidence as the main measure with the support of the rule ranging from 70%, 60%, to 50%, the results are shown as in Table 7.

It can be seen that the number of fuzzy association rules by using confidence measure as the main measure is different from using support measure as the main measure. Using the confidence as the main measure can provide a slightly larger number of fuzzy association rules than the use of support measure as the main measure. This may be can see from the

fact that confidence is a more relax criteria than the support measure.

5. Conclusions

Association rule mining is one technique of data mining that has been widely used in a number of applications. For some numeric and continuous values, mining for association rules cannot be applied directly because of the enormous amount of possible values. Fuzzy set concept can be applied to handle this situation and gives rise to the new sub-area called fuzzy association rule mining. A small number of rules given by the fuzzy association rule mining process is important to the performance justification of the process. In the work we propose the findings from our experimentation that a combination of measures to select the final rule set gives a better result than a single measure. Moreover, we found that the two-measures such as support*lift and confidence*leverage yield a reasonable set of fuzzy association rules.

References

- (1) Rakesh Agrawal, Tomasz Imielinski and Arun Swami. Mining Association Rules between Sets of Items in Large Database. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207-216, 1993.
- (2) Ferenc Peter Pach, Attila Gyenesei and Janos Abonyi. Compact Fuzzy Association Rule-Based Classifier. Expert System with Applications. Vol. 34, Issue 4, pp. 2406-2416, 2008.
- (3) Maria Martinez Ballesteros, Francisco Martinez Alvarez, Alicia Troncoso and Jose C. Riquelme. Selecting the Best Measures to Discover Quantitative Association Rules. Neurocomputing. Vol. 126. pp. 3-14, 2014.
- (4) Sergey Brin, Rajeev Motwani Motwani and Craig Silverstein. Beyond market baskets: generalizing association rules to correlations. Proceedings of the ACM SIGMOD. Vol. 26. pp. 265–276, 1997.
- (5) Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. Proceedings of the ACM SIGMOD. pp. 265–276, 1997.
- (6) Liqiang Geng and Howard J. Hamilton. Interestingness measures for data mining: a survey. ACM Computing Surveys. Vol. 38. Issue 3. pp. 1–42, 2006.
- (7) G Piatetsky Shapiro. Discovery. Analysis and presentation of strong rules. Knowledge Discovery in Databases. pp. 229–248, 1991.

Dissimilar Rule Mining and Ranking Technique for Associative Classification

Phaichayon Kongchai*, Nittaya Kerdprasop, and Kittisak Kerdprasop

Abstract—This research presents an associative classification with dissimilar rules (ACDR) algorithm to discover association rules with the highest priority and the top frequency. The proposed algorithm has the ability to reduce redundant rules and to sort rules in decreasing order by their priorities. The results are dissimilar rules that can be used to predict information in the future. This algorithm can be applied as an associative classification technique and then sorted the results by interestingness measures. We develop the program with Rstudio, which is a very popular software package in statistical analysis and data mining. In the experimentation, we used the post-operative patients dataset to evaluate efficiency of the algorithm. The results confirm effectiveness of the ACDR algorithm by discovering a minimal but powerful set of association rules.

Index Terms—R Language, Association Rule, Algorithm Apriori, Associative Classification

I. INTRODUCTION

Association rule mining is to find the relations among data items from large database. The results can be used to predict future information or explain current relation. Apriori algorithm [1] is a popular method for association rule mining. This algorithm was developed based on AIS algorithm and focused on the pruning infrequent item sets. Many open-sources software can be used to discover the frequent patterns such as WEKA, which is software that can import data into the program and the final results are association rules, RapidMiner that has many tools for data mining and users can use operator chaining technique for mining with many algorithms in a single execution. But in this research we select the Rstudio for mining association rules because with this software, users can implement and extend algorithm easier than WEKA and RapidMiner that are Java implementation.

Rstudio is a suite of program environment to run the R language program, which is commonly language used to compute the statistics applications. This program environment provides several types of graphical display and has many libraries for discovering classification and association rules. In this research, we use the library arules because it can find the patterns with only a few lines of code. Moreover, this library was designed to allow users to

specify the mining for association rules with the constraints. With the constraint mining feature, it was thus easier and faster to find associative patterns with the proposed ACDR (Associative Classification with Dissimilar Rules) algorithm.

The main contribution of this research is proposing the ACDR algorithm. It can be used to discover dissimilar rules for classification. The algorithm has 5 main steps: searching for association rules, categorizing rules into target association rules and general association rules, classifying rules into groups by their right-hand-side item (RHS), analyzing with selected agent of each group, and sorting rules.

The proposed algorithm works with any dataset, but for the demonstration purpose, we apply the algorithm to the post-operative patients dataset.

II. RELATED WORK

This research aims to reduce the number of association rules that are redundant and retain the remaining rules that are important for predicting the future events. Kannan and Bhaskaran [4] proposed algorithm for reducing redundant rules by clustering association rules into many groups then cut redundant rules by interestingness measures. Mutter et al. [5] used CBA (Confidence-Based Association Rule Mining) algorithm to reduce the number of association rules. They ranked rules by confidence values then output rules for top hundred association patterns. Our work presented in this paper is different from others in that we used associative classification technique to rank and reduce association rules.

Associative classification technique is an integrated of classification rules and association rules. The goal of this technique is to search for the results having the format “If one item or more items have occurred, then another item must occur”. It is like the classification rules. Hanchotchuang et al. [3] used associative classification technique for predicting unknown class label by guessing the class label with association rules then the results will be classified with classification rules. Tang and Liao [7] proposed a new Class Based Associative Classification algorithm (CACA). Their algorithm tried to reduce the searching space and results are better accuracy of classification models.

Further this research also does the top ranking after the discovery of important association rules. The ranking technique is to sort rules in decreasing order by their priorities. There are many researches which focus on sorting rules [2], [6], [9], [10]. In this research, we use four criteria to rank priorities of the association rules. The four criteria are the size of the association rules, confidence, support and target rules.

Manuscript received November 30, 2012; revised January 10, 2013. This work was supported in part by grant from Suranaree University of Technology through the funding of Data Engineering Research Unit.

P. Kongchai is a doctoral student with the School of Computer Engineering, Suranaree University of Technology, Thailand, (email: zaguraba_ji@hotmail.com).

N. Kerdprasop is an associate professor with the School of Computer Engineering, Suranaree University of Technology, Thailand.

K. Kerdprasop is an associate professor with the School of Computer Engineering, Suranaree University of Technology, Thailand.

III. METHODOLOGY

In this section we present ACDR algorithm for discovering association rules with the highest priority and the top frequency in descending order. The process of ACDR of two main parts, (1) to mine for association rules, and (2) to analyze association rules for finding important rules. We do the ranking priorities of the association rules with RStudio program. The details of ACDR algorithm, are shown in Fig. 1. Its diagrammatic flow is presented in Fig.2. Each subsection, A to E, is explanation of ACDR algorithm through the simple running example.

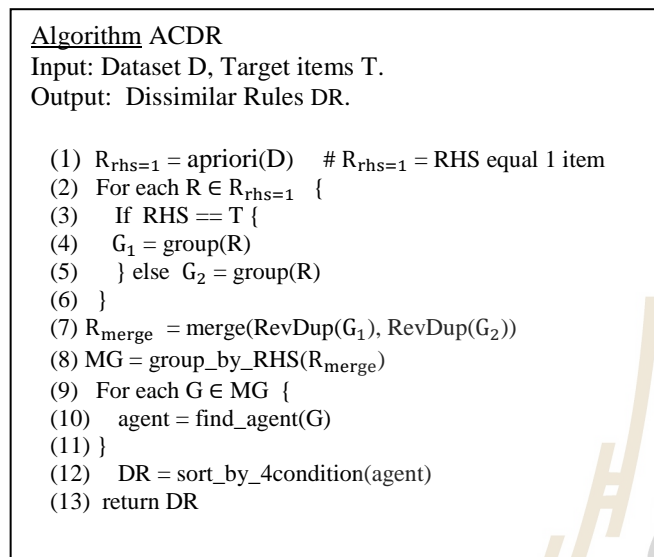


Fig. 1 ACDR (Associative Classification with Dissimilar Rules) algorithm.

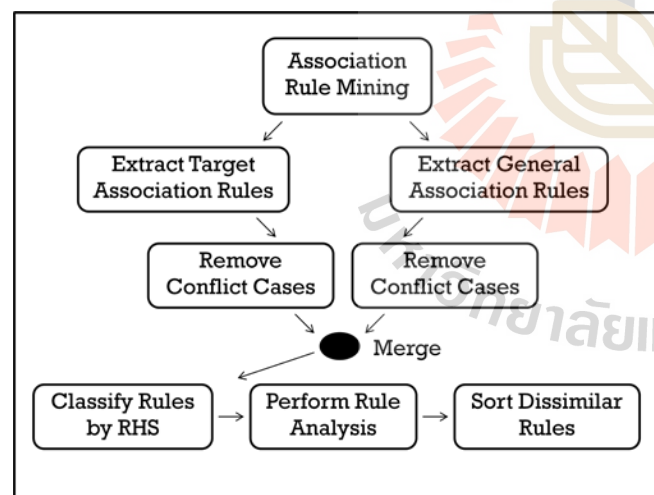


Fig. 2 The process of ACDR algorithm.

A. Association Rules Mining

This research uses apriori algorithm [1] as a basis for further extension because its association rule mining steps are simple but highly efficient pruning strategy to remove infrequent item sets with minimum support measure (eq1). Support measure of item A is proportion of number of transactions that contain A to the total number of transactions in the database.

$$\text{support}(A) = \frac{|A|}{|\text{transactions}|} \quad (1)$$

The results are frequent item sets that can be used further association rules constrained by the minimum confidence measure (eq2).

$$\text{confidence}(A \rightarrow B) = \frac{\text{support}(A \cap B)}{\text{support}(A)} \quad (2)$$

To implement the proposed methodology we are developed a program with R language which is suitable for data mining and the R system has many libraries for discovering association rules. For example to find association rules, the R code is as simple as the one show in Fig.3.

```

library(arules) # call library arules
Tr <- read.transactions("test.txt",format="basket",
,sep=",")
# read file and storing data in format transaction.
rules <- apriori(Tr, parameter= list(supp=0.1,
conf=0.6, minlen = 2))
# association rule mining by apriori algorithm and
set parameter with minimum support as 0.1, minimum
confidence as 0.6 and the size of rules to contain at
least 2 items.
inspect(rules)
# show all rules
    
```

Fig. 3 The R code for association rule mining.

From the commands in Fig.3 and the data as shown in Table 1, the result of program execution displayed in Table 2. With the simple six transactions given as the input, the output is a set of 17 association rules displayed in Table 2. These association rules have been constrained to contain exactly 1 item in the consequent part (or right-hand-side, RHS). This constraint is for later pruning the association rules.

B. Extract Target Association Rules and General Association Rules

This ACDR algorithm aims to predict or make decision on data that may occur in the future. Therefore, we applied a technique to include classification rules and association rules and call this technique is an associative classification. For our associative classification technique, we divided association rules into two groups, The first group is the rules to be defined by users contain target items (called Target Rules), and the second group is the rules not defined to contain target items (called General Rules). Target and general rule extraction is the step between lines 2-6 in Fig.1. Suppose the target item defined by users are item C and D, then the extracted target association rules are those illustrated in Table 3, whereas the rest (Table 4) is a set of general association rules.

TABLE 1
EXAMPLE TRANSACTION DATABASE

ID	Item
1	A, B, C
2	B, C
3	A, B, D
4	A, B, C, D
5	A
6	B

TABLE 2
ASSOCIATION RULES WITH A SINGLE ITEM IN THEIR RHS

NO.	Rules	Support	Confidence
1	{D} => {A}	0.333	1
2	{D} => {B}	0.333	1
3	{C} => {A}	0.333	0.667
4	{C} => {B}	0.5	1
5	{B} => {C}	0.5	0.6
6	{A} => {B}	0.5	0.75
7	{B} => {A}	0.5	0.6
8	{C,D} => {A}	0.167	1
9	{C,D} => {B}	0.167	1
10	{A,D} => {B}	0.333	1
11	{B,D} => {A}	0.333	1
12	{A,B} => {D}	0.333	0.667
13	{A,C} => {B}	0.333	1
14	{B,C} => {A}	0.333	0.667
15	{A,B} => {C}	0.333	0.667
16	{A,C,D} => {B}	0.167	1
17	{B,C,D} => {A}	0.167	1

TABLE 3
TARGET ASSOCIATION RULES THAT CONTAIN
THE TARGET ITEMS C AND D IN THE RHS

NO.	Rules	Support	Confidence
5	{B} => {C}	0.5	0.6
12	{A,B} => {D}	0.333	0.667
15	{A,B} => {C}	0.333	0.667

The rules in Tables 3 and 4 may contain conflicting cases such as rule number 12 and 15 have exactly the same antecedent parts, but they predict different consequences. We call such case a conflict. At line 7 of the ACDR algorithm (Fig.1), we remove conflicting cases from both the target and general association rules. The remaining rules are shown in Tables 5 and 6.

TABLE 4
GENERAL ASSOCIATION RULES

NO.	Rules	Support	Confidence
1	{D} => {A}	0.333	1
2	{D} => {B}	0.333	1
3	{C} => {A}	0.333	0.667
4	{C} => {B}	0.5	1
6	{A} => {B}	0.5	0.75
7	{B} => {A}	0.5	0.6
8	{C,D} => {A}	0.167	1
9	{C,D} => {B}	0.167	1
10	{A,D} => {B}	0.333	1
11	{B,D} => {A}	0.333	1
13	{A,C} => {B}	0.333	1
14	{B,C} => {A}	0.333	0.667
16	{A,C,D} => {B}	0.167	1
17	{B,C,D} => {A}	0.167	1

TABLE 5
TARGET RULES AFTER REMOVING CONFLICT CASES

NO.	Rules	Support	Confidence
5	{B} => {C}	0.5	0.6

TABLE 6
GENERAL RULES AFTER REMOVING CONFLICT CASES

NO.	Rules	Support	Confidence
6	{A} => {B}	0.5	0.75
7	{B} => {A}	0.5	0.6
10	{A,D} => {B}	0.333	1
11	{B,D} => {A}	0.333	1
13	{A,C} => {B}	0.333	1
14	{B,C} => {A}	0.333	0.667
16	{A,C,D} => {B}	0.167	1
17	{B,C,D} => {A}	0.167	1

C. Classify Rules by RHS (Right-Hand-Side) Item

The step at line 8 of the ACDR algorithm is to allocate rules into groups according to the items appeared in the RHS of the rules. The rule classifying strategy is as follow:

1. All items on the right hand side of association rules must be the same items. For example, from Table 6 rules 6, 10, 13 and 16 have the same item on their right hand side, which is B. Therefore, they are allocated as the same group.

2. Items on the right hand side are not the same, they will be allocated to the different groups.

From Tables 5 and 6, target and general rules are then classified into groups and the results are three groups as shown in Tables 7-9.

TABLE 7
GROUP C OF ASSOCIATION RULES

NO.	Rules	Support	Confidence
5	{B} => {C}	0.5	0.6

TABLE 8
GROUP B OF ASSOCIATION RULES

NO.	Rules	Support	Confidence
6	{A} => {B}	0.5	0.75
10	{A,D} => {B}	0.333	1
13	{A,C} => {B}	0.333	1
16	{A,C,D} => {B}	0.167	1

TABLE 9
GROUP A OF ASSOCIATION RULES

NO.	Rules	Support	Confidence
7	{B} => {A}	0.5	0.6
11	{B,D} => {A}	0.333	1
14	{B,C} => {A}	0.333	0.667
17	{B,C,D} => {A}	0.167	1

D. Rule Analysis

After classifying rules into groups, the next step is to select agent of each group (Fig. 1 line 9-11). These agents are for rule ranking and selecting. The criteria for rule selection are:

1. Select association rules with the longest size. The reason is that they can describe the complex conditions. For example, the patient who had a first degree of the tumor, had irradiated, had surgery and a healthy body then decision is that the patient is recovered from cancer.

2. Select association rules with the shortest size for describing the causes that may incur the damage. For Example, the patient who had the tumor and is in the final stage then the patient is cancerous.

From the rules in Tables 7-9, after analyzing rules with two criteria, we obtain the results as shown in Tables 10 and 11. Note that a single rule in group C remains the same one as shown in Table 7.

TABLE 10
GROUP B AFTER RULE SELECTION

NO.	Rules	Support	Confidence
6	{A} => {B}	0.5	0.75
16	{A,C,D} => {B}	0.167	1

TABLE 11
GROUP A AFTER RULE SELECTION

NO.	Rules	Support	Confidence
7	{B} => {A}	0.5	0.6
17	{B,C,D} => {A}	0.167	1

E. Sort Dissimilar Rules

The final process is to combine the three groups into one group and then sort the rules by the following criteria (Fig. 1 line 12).

1. If association rule was the shortest size, it will then be in the first order. If the rules are the same size, they will be considered by the next criterium.

2. If association rule is defined target item, it will be in the first order.

3. If association rule has the maximum confidence value, it will be in the first order. But if the rules have the same confidence value, they will be ranked by the next criterium.

4. If association rule has the maximum support value, it will be in the first order. But If the rules have the same support value, they will be ranked by order number.

The rules in Tables 7, 10 and 11 will be merged and then sorted with the four criteria. The results are shown in Table 12.

TABLE 12
ASSOCIATION RULES AFTER SORTING

NO.	Rules	Support	Confidence
5	{B} => {C}	0.5	0.6
6	{A} => {B}	0.5	0.75
7	{B} => {A}	0.5	0.6
16	{A,C,D} => {B}	0.167	1
17	{B,C,D} => {A}	0.167	1

From Table 12 association rules NO. 5 contains defined items by user (item C and D), thus it is ranked first. Association rules NO. 6 and 7 are rules of the same size, they must be ranked by confidence value. Rule NO. 6 has higher confidence value than rule NO. 7, it is therefore ranked preceding rule No.7. Association rules NO. 16 and 17 are the same size and also the same confidence value and support value, they will be ranked according to the order number. The result is that NO. 16 has been ranked preceding rule NO. 17.

IV. EXPERIMENT

This research experimented with the post-operative patients dataset obtained from the UCI Machine Learning Repository [8]. The dataset has 8 attributes (explained in Table 13) and 90 transactions.

To perform the experiment, we developed a program using Rstudio environment and coding with R language for discovery association rules by apriori algorithm. We set minimum support and minimum confidence to be 0.01 and we define target items as ADM-DECS=I, ADM-DECS=S and ADM-DECS=A.

The objectives of this experiment are to observe a decrease in the number of rules in each step of pruning associative classification rules and the efficiency of ranking important rules process (Fig. 4 and Table 14).

TABLE 13
DESCRIPTION OF POST-OPERATIVE PATIENTS' DATASET

Attribute	Description
L-CORE	patient's internal temperature in degree celsius: high (> 37), mid (>= 36 and <= 37), low (< 36)
L-SURF	patient's surface temperature in degree celsius : high (> 36.5), mid (>= 36.5 and <= 35), low (< 35)
L-O2	oxygen saturation in % excellent (>= 98), good (>= 90 and < 98), fair (>= 80 and < 90), poor (< 80)
L-BP	last measurement of blood pressure high (> 130/90), mid (<= 130/90 and >= 90/70), low (< 90/70)
SURF-STBL	stability of patient's surface temperature : stable, mod-stable, unstable
CORE-STBL	stability of patient's core temperature : stable, mod-stable, unstable
BP-STBL	patient's perceived comfort at discharge, measured as an integer between 0-10 and 11-20
ADM-DECS	discharge decision : I (patient sent to Intensive Care Unit), S (patient prepared to go home), A (patient sent to general hospital floor)

TABLE 14
THE PROCESS OF ACDR ALGORITHM AND NUMBER OF RULES AFTER PERFORMING EACH PROCESS

Process	Number of rules (Rules)
1. Association Rule Mining	88,423
2. Extracting Target Association Rules and General Association Rules	5,231
3. Classifying Rules by RHS items	5,231
4. Performing Rule Analysis	1,048
5. Sorting Dissimilar rules	1,048

The results from Table 14 are important rules discovery with five sub-processes. The first sub-process is association rule with the consequent part containing 1 item and the result contains 88,423 rules. The second sub-process is to find target rules and general rules and also removing conflicting cases. The result contains 5,231 rules. The third sub-process is classifying rules by their RHS, the result is the same set of rules because this step classifies rules then inserts into group but does not remove any rules. The fourth sub-process is analyzing and selecting association rules by their sizes. The results are 1,048 rules. The last sub-process is sorting association rules, and the results are 1,048 rules.

1. {L-BP=low} => {ADM-DECS=A}
2. {CORE-STBL=mod-stable} => {ADM-DECS=A}
3. {BP-STBL=stable,CORE-STBL=unstable} => {ADM-DECS=S}
4. {BP-STBL=stable,L-CORE=high} => {ADM-DECS=S}
5. {COMFORT=?,L-CORE=low} => {ADM-DECS=I}
6. {BP-STBL=stable,COMFORT=?} => {ADM-DECS=I}
7. {COMFORT=?,L-O2=good} => {ADM-DECS=I}
8. {COMFORT=?,L-SURF=mid} => {ADM-DECS=I}
9. {COMFORT=[11 - 20],CORE-STBL=unstable} => {ADM-DECS=S}
10. {CORE-STBL=unstable,L-BP=high} => {ADM-DECS=S}
11. {CORE-STBL=unstable,L-O2=good} => {ADM-DECS=S}
12. {BP-STBL=stable,COMFORT=[0 - 10],CORE-STBL=stable,L-BP=mid,L-CORE=mid,L-O2=excellent,L-SURF=mid,SURF-STBL=unstable} => {ADM-DECS=A}
13. {BP-STBL=stable,COMFORT=[0 - 10],CORE-STBL=stable,L-BP=high,L-CORE=mid,L-O2=excellent,L-SURF=mid,SURF-STBL=stable} => {ADM-DECS=A}
14. {BP-STBL=mod-stable,COMFORT=[0 - 10],CORE-STBL=stable,L-BP=high,L-CORE=low,L-O2=excellent,L-SURF=mid,SURF-STBL=stable} => {ADM-DECS=A}
15. {BP-STBL=stable,COMFORT=[0 - 10],CORE-STBL=stable,L-BP=mid,L-CORE=low,L-O2=excellent,L-SURF=low,SURF-STBL=stable} => {ADM-DECS=A}
...
77. {BP-STBL=stable,COMFORT=[0 - 10],CORE-STBL=stable,L-BP=mid,L-O2=excellent,L-SURF=mid,SURF-STBL=unstable} => {L-CORE=mid}
78. {BP-STBL=stable,COMFORT=[0 - 10],CORE-STBL=stable,L-CORE=mid,L-O2=excellent,L-SURF=mid,SURF-STBL=unstable} => {L-BP=mid}
79. {BP-STBL=stable,COMFORT=[11 - 20],L-BP=mid,L-CORE=mid,L-O2=good,L-SURF=mid,SURF-STBL=unstable} => {CORE-STBL=stable}
80. {BP-STBL=stable,COMFORT=[11 - 20],CORE-STBL=stable,L-BP=mid,L-O2=good,L-SURF=mid,SURF-STBL=unstable} => {L-CORE=mid}

Fig. 4 The results from ACDR algorithm.

The authors proposed an associative classification with dissimilar rules algorithm to discover association rules with the highest priority and the top frequency. The experimental results are composing of target rules and general rules. Target rules are the rules number 1-76 and general rules are the rules number 77-1,048. The rule number 1 can be interpreted as “if last measurement of blood pressure is low then discharge decision is to send the patient to general hospital floor”. The symbol “?” in rule number 5 means that the attribute comfort has some effect to the decision ADM-DECS=I but we do not know the value.

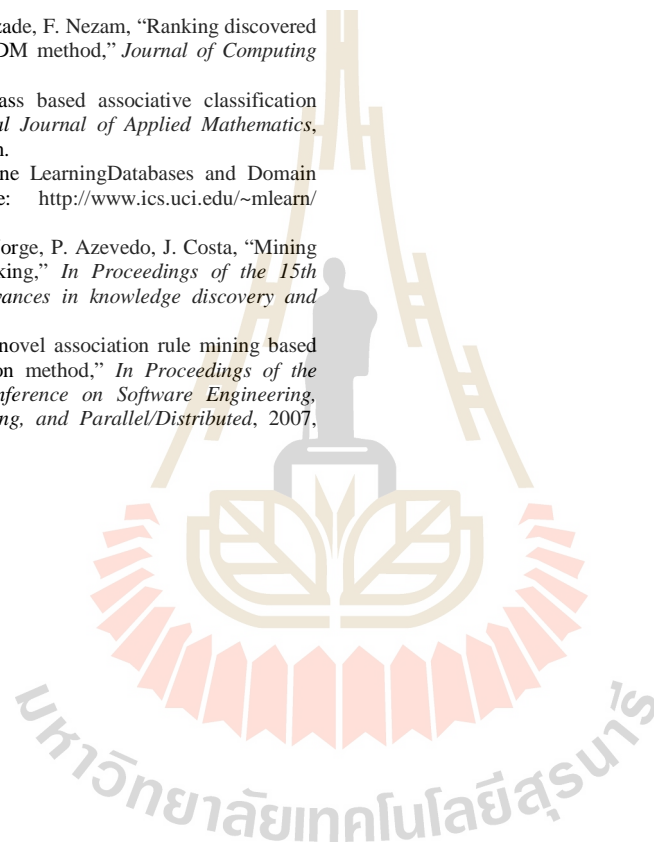
V. CONCLUSION

This research introduces a design approach called ACDR (Associative Classification with Dissimilar Rules) to reduce redundant target and general association rules, then the results can be used to predict information as most classification rules. The ACDR algorithm consists of 5 main steps which are (1) finding association rules, (2) clustering target association rules and general association rules into two groups then removing redundant rules, (3) classifying rules into groups by their RHS item, (4) performing rule analysis with selected agent of each group, and (5) sorting rules according to proposed criteria. The dataset for algorithm evaluation is the post-operative patients dataset. The final result after processing the dataset through the five main steps of the ACDR algorithm is a minimal rule set

containing 1,048 rules, which are significantly decreased from the original 88,423 rules.

REFERENCES

- [1] R. Agrawal, R. Srikant, "Fast algorithms for mining association rules," *In Proceedings of the International Conference on Very Large Data Bases*, 1994, pp. 487-499.
- [2] S. Bouker, R. Saidi, S. B. Yahia, M. E. Nguifo. "Ranking and selecting association rules based on dominance relationship," *In Proceedings of the 24th IEEE International Conference on Tools with Artificial Intelligence*, 2012.
- [3] W. Hranchothuang, T. Rakthanmanon, K. Waiyamai, "Using maximal frequent itemsets for improving associative classification," *In Proceedings of the 1st National Conference on Computing and Information Technology*, 2005, pp 24-25.
- [4] S. Kannan; R. Bhaskaran, "Association Rule Pruning based on interestingness measures with clustering," *IJCSI International Journal of Computer Science Issues*, V.6, 2009, pp. 35-43.
- [5] S. Mutter, M. Hall, E. Frank, "Using classification to evaluate the output of confidence-based association rule mining," *In Proceedings of Australian Conference on Artificial Intelligence*, 2004, pp. 538-549.
- [6] G. Peyman, M. R. Sepehri, B. Azade, F. Nezam, "Ranking discovered rules from data mining by MADM method," *Journal of Computing Issue 11*. V.3. 2011, pp. 64.
- [7] Z. Tang, Q. Liao, "A new class based associative classification algorithm," *IAENG International Journal of Applied Mathematics*, 2007, Advance online publication.
- [8] The UCI Repository Of Machine Learning Databases and Domain Theories [Online]. Available: <http://www.ics.uci.edu/~mllearn/MLRepository.html>
- [9] R. Sá. De. C. C. Soares, M. A. Jorge, P. Azevedo, J. Costa, "Mining association rules for label ranking," *In Proceedings of the 15th Pacific-Asia conference on Advances in knowledge discovery and data mining*, 2011, pp. 432-443.
- [10] J. Wu, Q. Song; J. Shen, "An novel association rule mining based missing nominal data imputation method," *In Proceedings of the Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed*, 2007, V.3. pp. 244-249.



ภาคผนวก ข

ลิขสิทธิ์โปรแกรม

โปรแกรมจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือที่กะทัดรัด
(Data classification program with compact fuzzy association rules)



มหาวิทยาลัยเทคโนโลยีสุรนารี

ชื่อภาษาไทย	โปรแกรมจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือที่กะทัดรัด
ชื่อภาษาอังกฤษ	Data classification program with compact fuzzy association rules
ทะเบียนข้อมูลเลขที่	ว1. 5344
ให้ไว้ ณ วันที่	8 เมษายน พ.ศ. 2558
คำอธิบายโปรแกรมโดยย่อ	<p>โปรแกรมที่พัฒนาขึ้นนี้ให้ชื่อว่า CCFAR (classification with compact fuzzy association rules) การทำเหมืองข้อมูลด้วยวิธีการจำแนกประเภทข้อมูล มีจุดประสงค์เพื่อนำโมเดลที่ได้จากกระบวนการเรียนรู้มาใช้ในการทำนายข้อมูลในอนาคต ซึ่งในปัจจุบันมีผู้วิจัยจำนวนมากให้ความสนใจที่จะพัฒนาประสิทธิภาพขั้นตอนวิธีการจำแนกประเภทข้อมูล เพื่อให้มีความแม่นยำในการจำแนกมากขึ้น และโมเดลที่ได้สามารถตีความหมายได้ง่าย แต่การที่จะเพิ่มประสิทธิภาพทั้งสองอย่างควบคู่กันไปในั้นยังไม่สามารถพัฒนาได้อย่างสมบูรณ์ ดังนั้นโปรแกรมนี้นี้จึงได้เสนอวิธีการพัฒนาขั้นตอนวิธีเพื่อจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือที่กะทัดรัด เพื่อเพิ่มประสิทธิภาพการจำแนกประเภทข้อมูลในสองด้านคือ เพื่อให้โมเดลที่ได้รับมีความแม่นยำอยู่ในเกณฑ์ดีและสามารถตีความหมายได้ดีด้วย โดยได้นำเทคนิคการทำเหมืองข้อมูลเพื่อหากฎความสัมพันธ์มาผสมผสานกับเทคนิคการจำแนกประเภทข้อมูล</p> <p>โปรแกรมจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือที่กะทัดรัดนี้พัฒนาด้วยภาษา MATLAB แนวคิดหลักของโปรแกรมนี้นี้ คือ การสร้างกฎการจำแนกประเภทข้อมูลด้วยกฎที่มีจำนวนไม่มาก สามารถตีความได้ง่าย และมีประสิทธิภาพในการจำแนกประเภทข้อมูลสูง โปรแกรมประกอบด้วยขั้นตอนการทำงาน 5 ส่วน คือ 1. การตรวจสอบข้อมูลก่อนการประมวลผล 2. การแบ่งแยกข้อมูล 3. การสร้างไอเท็มเซตปรากฏบ่อยแบบคลุมเครือ 4. การสร้างกฎ FCARs (fuzzy classification based on association rules) และ 5. การเลือกกฎ FCARs เพื่อนำไปใช้ทำนายข้อมูล</p>



รลข.01

ทะเบียนข้อมูลเลขที่ ว1. 5344

หนังสือรับรองการแจ้งข้อมูล
ลิขสิทธิ์
ออกให้เพื่อแสดงว่า
มหาวิทยาลัยเทคโนโลยีสุรนารี

ได้แจ้งข้อมูลลิขสิทธิ์ ประเภทงาน วรรณกรรม

ลักษณะงาน โปรแกรมคอมพิวเตอร์

ชื่อผลงาน โปรแกรมจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือที่กะทัดรัด

ไว้ต่อกรมทรัพย์สินทางปัญญา ตามคำขอแจ้งข้อมูลลิขสิทธิ์ เลขที่ 322069

เมื่อวันที่ 1 เดือน เมษายน พ.ศ. 2558

ให้ไว้ ณ วันที่ 8 เดือน เมษายน พ.ศ. 2558

ลงชื่อ.....

นางสาวศิริวรรณ นพริก

นักวิชาการพาณิชย์ปฏิบัติการ

ปฏิบัติราชการแทนผู้อำนวยการสำนักลิขสิทธิ์

- หมายเหตุ**
1. เอกสารนี้มิได้รับรองความเป็นเจ้าของลิขสิทธิ์
 2. การเปลี่ยนแปลงรายการข้างต้น ให้ดูด้านหลัง

ประวัติผู้วิจัย

รองศาสตราจารย์ ดร.นิตยา เกิดประสพ สำเร็จการศึกษาในระดับปริญญาเอกสาขา Computer Science จาก Nova Southeastern University เมือง Fort Lauderdale รัฐฟลอริดา สหรัฐอเมริกา เมื่อปีพุทธศักราช 2542 (ค.ศ. 1999) ด้วยทุนการศึกษาของกระทรวงวิทยาศาสตร์และเทคโนโลยี โดยทำวิทยานิพนธ์ระดับปริญญาเอกในหัวข้อเรื่อง "The application of inductive logic programming to support semantic query optimization" หลังสำเร็จการศึกษาได้ปฏิบัติราชการในตำแหน่งอาจารย์ ประจำสาขาคอมพิวเตอร์ ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ต่อมาในปีพุทธศักราช 2543 ได้มาปฏิบัติงานในตำแหน่งอาจารย์ประจำ สาขาวิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี จนถึงปัจจุบัน งานวิจัยที่ทำในขณะนี้ คือการประยุกต์เทคโนโลยีเหมืองข้อมูลกับงานด้านการแพทย์ การสาธารณสุขและสิ่งแวดล้อม รวมถึงการพัฒนาเทคนิคเพื่อเพิ่มความสามารถในการจัดการความรู้ของระบบเหมืองข้อมูล

รองศาสตราจารย์ ดร.กิตติศักดิ์ เกิดประสพ สำเร็จการศึกษาในระดับปริญญาเอกสาขา Computer Science จาก Nova Southeastern University เมือง Fort Lauderdale รัฐฟลอริดา สหรัฐอเมริกา เมื่อปีพุทธศักราช 2542 (ค.ศ. 1999) ด้วยทุนการศึกษาของทบวงมหาวิทยาลัย (หรือสำนักงานคณะกรรมการอุดมศึกษาในปัจจุบัน) โดยทำวิทยานิพนธ์ระดับปริญญาเอกในหัวข้อเรื่อง "Active database rule set reduction by knowledge discovery" หลังสำเร็จการศึกษาได้ปฏิบัติงานในตำแหน่งอาจารย์ ประจำสาขาวิชาวิศวกรรมคอมพิวเตอร์ สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี ปัจจุบันดำรงตำแหน่งหัวหน้าหน่วยวิจัยวิศวกรรมความรู้ เน้นการวิจัยเกี่ยวกับการพัฒนาระบบเหมืองข้อมูลประสิทธิภาพสูง การประยุกต์เหมืองข้อมูลกับงานวิศวกรรม และการวิเคราะห์ข้อมูลเชิงสถิติ รวมถึงการวิจัยพื้นฐานเกี่ยวกับเทคนิคการวิเคราะห์ข้อมูลโดยวิธีอัตโนมัติ โดยมีผลงานวิจัยตีพิมพ์ในวารสารวิชาการและเอกสารการประชุมวิชาการจำนวนมากกว่า 290 เรื่องในด้านฐานข้อมูล การวิเคราะห์ข้อมูล การทำเหมืองข้อมูลและการค้นหาความรู้
