วิธีแบ่งช่วงข้อมูลสำหรับการหากฎความสัมพันธ์

นายนันทวุฒิ คะอังกุ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์
มหาวิทยาลัยเทคโนโลยีสุรนารี
ปีการศึกษา 2555

A DISCRETIZATION METHOD FOR ASSOCIATION RULE MINING

Nuntawut Kaoungku

A Thesis Submitted in Partial Fulfillment of the Requirements for the

Degree of Master of Engineering in Computer Engineering

Suranaree University of Technology

Academic Year 2012

วิธีแบ่งช่วงข้อมูลสำหรับการหากฎความสัมพันธ์

มหาวิทยาลัยเทคโนโลยีสุรนารี อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

คณะกรรมการสอบวิทยานิพนธ์

_____

(รศ. ดร.กิตติศักดิ์ เกิดประสพ)
ประธานกรรมการ

_____

(รศ. ดร.นิตยา เกิดประสพ)
กรรมการ (อาจารย์ที่ปรึกษาวิทยานิพนธ์)

_____

(อ. ดร.จิติมนต์ อั่งสกุล)
กรรมการ

_____

(ศ. ดร.ชูกิจ ลิมปิจำนงค์)
รองอธิการบดีฝ่ายวิชาการ

_____

(รศ. ร.อ. ดร.กนต์ธร ชำนิประศาสน์)
คณบดีสำนักวิชาวิศวกรรมศาสตร์

นันทวุฒิ คะอังกุ : วิธีแบ่งช่วงข้อมูลสำหรับการหากฎความสัมพันธ์
(A DISCRETIZATION METHOD FOR ASSOCIATION RULE MINING)
อาจารย์ที่ปรึกษา: รองศาสตราจารย์ ดร. นิตยา เกิดประสพ, 88 หน้า.

งานวิจัยนี้ได้ศึกษาปัญหาการแบ่งช่วงข้อมูลสำหรับการหากฎความสัมพันธ์ ซึ่งการแบ่งช่วงข้อมูลเป็นการจัดการกับข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยการจัดเป็นช่วงของค่า การแบ่งช่วงนี้เป็นขั้นตอนก่อนนำข้อมูลไปเข้ากระบวนการหากฎความสัมพันธ์เพื่อช่วยลดจำนวนกฎความสัมพันธ์ที่ได้ออกมาและช่วยให้ได้กฎที่มีค่าสนับสนุนที่สูงขึ้น ในอดีตได้มีหลากหลายงานวิจัยที่ได้เสนอเทคนิคในการแบ่งช่วงของข้อมูล แต่เทคนิคส่วนใหญ่ทำการแบ่งช่วงของข้อมูลสำหรับงานทำเหมืองข้อมูลประเภทการจำแนกข้อมูล ซึ่งวัดประสิทธิภาพของเทคนิคได้ชัดเจนจากการพิจารณาค่าความถูกต้องของโมเดล แต่การแบ่งช่วงข้อมูลเพื่อใช้กับการหากฎความสัมพันธ์จะวัดประสิทธิภาพได้ยากกว่า จึงทำให้มีงานวิจัยด้านนี้ปรากฏค่อนข้างน้อย ผู้วิจัยได้เห็นความสำคัญในจุดนี้จึงได้เสนอเทคนิคการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการหากฎความสัมพันธ์ เพื่อเพิ่มประสิทธิภาพในการหากฎความสัมพันธ์ให้ดียิ่งขึ้น โดยจะทำการเปรียบเทียบประสิทธิภาพการแบ่งช่วงข้อมูลที่ได้เสนอในการวิจัยนี้กับเทคนิคการแบ่งช่วงข้อมูลอื่นด้วยวิธีอื่น ๆ โดยใช้ไลบรารีภาษาอาร์ในส่วนที่เกี่ยวข้องกับการหากฎความสัมพันธ์ งานวิจัยนี้จะใช้ค่าความถูกต้อง และมาตรวัด 4 มาตรวัดในการประเมินประสิทธิภาพกฎความสัมพันธ์

สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนักศึกษา_____

ปีการศึกษา 2555 ลายมือชื่ออาจารย์ที่ปรึกษา_____

NUNTAWUT KAOUNGKU : A DISCRETIZATION METHOD FOR ASSOCIATION RULE MINING. THESIS ADVISOR : ASSOC. PROF. NITTAYA KERDPRASOP, Ph.D., 88 PP.

DISCRETIZATION/ASSOCIATION RULE MINING/CHI SQUARE

In this research, we study the problem of discretization for association rule mining. The discretization method is a handling technique to cope with numerical or continuous data. It is the pre-processing step of association rule mining to reduce the number of rules, and to increase the support value. In the past, many researchers have proposed techniques for discretization, but most of them perform a discretization for classification in which the performance measure is obviously the accuracy of the model. The discretization method for association rule mining, on the contrary, has no clear evaluation metric. We thus propose in this research the discretization technique for a specific task of association rule mining, and the rule assessment method. The implementation and experimentation have been done with the R language and its libraries. We comparatively experiment our proposed discretization method with the existing techniques. Assessment metrics are accuracy and the other four metrics.

School of Computer Engineering Student's Signature_____

Academic Year 2012 Advisor's Signature_____

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จลุล่วงด้วยดี ผู้วิจัยขอกราบขอบพระคุณ บุคคล และกลุ่มบุคคลต่างๆที่ได้กรุณาให้คำปรึกษา แนะนำ ช่วยเหลืออย่างดียิ่ง ทั้งในด้านวิชาการ และด้านการดำเนินงานวิจัย ดังต่อไปนี้

รองศาสตราจารย์ ดร.นิตยา เกิดประสพ อาจารย์ที่ปรึกษาวิทยานิพนธ์ และรองศาสตราจารย์ ดร.กิตติศักดิ์ เกิดประสพ ที่ให้คำปรึกษาในการทำงานวิจัย การจัดการรูปแบบ และช่วยตรวจทานความถูกต้องของวิทยานิพนธ์

ผู้ช่วยศาสตราจารย์ ดร.พิชโยทัย มหัทธนาภิวัฒน์ ผู้ช่วยศาสตราจารย์ ดร.คะชา ชาญศิลป์ ผู้ช่วยศาสตราจารย์ สมพันธ์ ชาญศิลป์ และผู้ช่วยศาสตราจารย์ ดร.ปรเมศวร์ ห่อแก้ว อาจารย์ประจำสาขาวิชาวิศวกรรมคอมพิวเตอร์ สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี

คุณกัลญา พับโพธิ์ เลขานุการสาขาวิชาวิศวกรรมคอมพิวเตอร์ ที่ให้ความช่วยเหลือในการประสานงานด้านเอกสารระหว่างศึกษา

คุณภาสพิชญ์ ชูใจ คุณไพชยนต์ คงไชย คุณพัชราวรรณ ชินไธสง คุณกิตติพงศ์ ชมบุญและนักศึกษาบัณฑิตสาขาวิชาวิศวกรรมคอมพิวเตอร์ ทุกท่านที่ให้คำปรึกษาและช่วยเหลือด้วยดีมาโดยตลอด

นอกจากนี้ขอขอบคุณครู อาจารย์ทั้งในอดีตและปัจจุบันที่ให้ความรู้แก่ผู้วิจัยจนประสบความสำเร็จในชีวิต

ท้ายที่สุดที่จะลืมไม่ได้ ขอกราบขอบพระคุณ บิดา มารดา ที่ให้กำเนิด อบรม เลี้ยงดูด้วยความรัก และส่งเสริมการศึกษาเป็นอย่างดีโดยตลอด ทำให้ผู้วิจัยมีความรู้ ความสามารถ มีจิตใจที่เข้มแข็ง รวมทั้งเป็นกำลังใจที่ยิ่งใหญ่แก่ผู้วิจัย จนทำให้ผู้วิจัยประสบความสำเร็จในชีวิตเรื่อยมา

นันทวุฒิ คะอังกุ

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญรูป

รูปที่ **หน้า**

สารบัญรูป (ต่อ)

รูปที่ **หน้า**

สารบัญรูป (ต่อ)

รูปที่ **หน้า**

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของปัญหาการวิจัย

ในปัจจุบันข้อมูลมีความสำคัญมาก เนื่องจากสมัยก่อนการเก็บข้อมูลนั้นอยู่ในรูปแบบของการจดบันทึกลงกระดาษมากกว่าการเก็บข้อมูลในรูปแบบของดิจิตอล แต่ในปัจจุบันคอมพิวเตอร์มีวิวัฒนาการที่ก้าวหน้ามากกว่าสมัยก่อนเป็นอย่างมาก ทำให้การเก็บข้อมูลในรูปแบบเดิมนั้นถูกแทนที่ด้วยการเก็บข้อมูลในรูปแบบของข้อมูลดิจิตอล เพราะนอกจากจะช่วยลดทรัพยากรในการเก็บข้อมูลแล้วยังช่วยเพิ่มความสะดวกในการบันทึกข้อมูลและการเรียกใช้ข้อมูล

ข้อมูลในโลกยุคปัจจุบันแทบจะไม่สามารถปฏิเสธได้ว่ามีบทบาทสำคัญเป็นอย่างมากในแต่ละองค์กร เพราะนับวันข้อมูลในแต่ละองค์กรมีปริมาณที่เพิ่มมากขึ้นและมีการจัดเก็บข้อมูลที่ซับซ้อนมากขึ้นตามไปด้วย ทำให้การจัดการข้อมูลเหล่านั้นมีความสำคัญเป็นอย่างมาก และเนื่องด้วยในโลกยุคปัจจุบันนี้เป็นการแข่งขันกันด้วยข้อมูล องค์กรใดมีการจัดการข้อมูลที่ดีกว่าก็มักจะได้เปรียบองค์กรอื่น ดังนั้นคอมพิวเตอร์จึงเป็นเครื่องมือที่ถูกนำมาใช้ในการจัดเก็บและจัดการกับข้อมูล คอมพิวเตอร์นั้นสามารถนำข้อมูลที่มีปริมาณมากมาวิเคราะห์ได้อย่างมีประสิทธิภาพมากกว่าการใช้มนุษย์มาวิเคราะห์ข้อมูลเพราะนอกจากจะช่วยลดเวลาแล้วยังช่วยลดความผิดพลาดที่เกิดจากความเหนื่อยล้าและความบกพร่องของมนุษย์

วิธีการนำข้อมูลมาวิเคราะห์ที่นิยมใช้ในปัจจุบันคือการทำเหมืองข้อมูล (Data mining) ซึ่งการทำเหมืองข้อมูลเป็นวิธีการสกัดความรู้จากฐานข้อมูลที่มีข้อมูลปริมาณมากโดยใช้หลักการทางสถิติ เพื่อต้องการที่จะหารูปแบบของข้อมูล แนวโน้มของข้อมูล หรือความสัมพันธ์ภายในกลุ่มข้อมูล การทำเหมืองข้อมูลนี้ใช้กันแพร่หลายในหลากหลายด้าน ไม่ว่าจะเป็นด้านการแพทย์ที่ใช้ในการวิเคราะห์ข้อมูลของผู้ป่วยเพื่อทำนายโรค ด้านธุรกิจที่ใช้ในการวิเคราะห์รูปแบบการซื้อสินค้าของลูกค้า หรือด้านการศึกษาที่ใช้ในการวิเคราะห์พฤติกรรมของผู้เรียนเพื่อที่จะวางแผนจัดการเรียนการสอนให้กับผู้เรียนได้อย่างมีประสิทธิภาพ เป็นต้น

เทคนิคในการทำเหมืองข้อมูลนั้นมีหลากหลายรูปแบบแล้วแต่การนำไปใช้งาน ซึ่งเทคนิคหนึ่งที่นิยมนำไปใช้กันอย่างแพร่หลายคือการหากฎความสัมพันธ์ (Association rule mining) เป็นการหาความสัมพันธ์จากเหตุการณ์หรือวัตถุที่เกิดขึ้นพร้อม ๆ กัน แล้วนำเหตุการณ์ที่เกิดพร้อมกันนั้นไปสร้างเป็นกฎความสัมพันธ์ แต่ปัญหาที่พบบ่อยในการหากฎความสัมพันธ์ก็คือข้อมูลที่มี

ลักษณะที่เป็นตัวเลขต่อเนื่อง (ในงานวิจัยนี้ตัวเลขต่อเนื่อง หมายถึง ตัวเลขที่เป็นค่า Continuous และตัวเลข Integer ที่มีค่าที่แตกต่างกันจำนวนมาก) จะทำให้กฎความสัมพันธ์ที่ได้ออกมานั้นมีจำนวนกฎที่มากเกินไป และแต่ละกฎให้ค่าสนับสนุน (Support) และค่าความเชื่อมั่น (Confidence) ที่ต่ำมากทำให้ไม่มีประสิทธิภาพมากเพียงพอต่อการนำไปใช้งาน (ดังแสดงด้วยตัวอย่างในรูปที่ 1.1 และ 1.2) ดังนั้นจึงต้องมีเทคนิคในการที่จะมาจัดการข้อมูลก่อนการนำไปเข้ากระบวนการหากฎความสัมพันธ์ เทคนิคดังกล่าวคือการแบ่งช่วงข้อมูล (Discretization)

Age	Sex	Buy
10	M	Y
26	M	N
14	F	Y
30	F	Y
18	M	N
35	M	Y

กฎความสัมพันธ์			
1. Buy=N	==> Sex=M	supp:(0.33)	conf:(1)
2. Sex=F	==> Buy=Y	supp:(0.33)	conf:(1)
3. Age=10	==> Sex=M	supp:(0.16)	conf:(1)
4. Age=10	==> Buy=Y	supp:(0.16)	conf:(1)
5. Age=14	==> Sex=F	supp:(0.16)	conf:(1)
6. Age=14	==> Buy=Y	supp:(0.16)	conf:(1)
7. Age=18	==> Sex=M	supp:(0.16)	conf:(1)

รูปที่ 1.1 ตัวอย่าง 7 กฎความสัมพันธ์จากทั้งหมด 66 กฎความสัมพันธ์ที่ได้จากข้อมูลที่เป็นตัวเลข

Age	Sex	Buy
range1[-∞-22)	M	Y
range2[22-∞]	M	N
range1[-∞-22)	F	Y
range2[22-∞]	F	Y
range1[-∞-22)	M	N
range2[22- ∞]	M	Y

กฎความสัมพันธ์			
1. Buy=N	==> Sex=M	supp:(0.33)	conf:(1)
2. Sex=F	==> Buy=Y	supp:(0.33)	conf:(1)
3. Age=range1[-∞-22)	==> Sex=M	supp:(0.33)	conf:(0.67)
4. Age=range1[-∞-22)	==> Buy=Y	supp:(0.33)	conf:(0.67)
5. Sex=M	==> Age=range1[-∞-22)	supp:(0.33)	conf:(0.5)
6. Buy=Y	==> Age=range1[-∞-22)	supp:(0.33)	conf:(0.5)
7. Buy=Y	==> Sex=M	supp:(0.33)	conf:(0.5)

รูปที่ 1.2 ตัวอย่าง 7 กฎความสัมพันธ์จากทั้งหมด 58 กฎความสัมพันธ์ที่ได้จากการปรับปรุงประสิทธิภาพกฎความสัมพันธ์ด้วยการแบ่งช่วงข้อมูล

การแบ่งช่วงข้อมูลเป็นการจัดการกับข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยการจัดเป็นช่วง (Interval) การแบ่งช่วงนี้เป็นขั้นตอนก่อนนำข้อมูลไปเข้ากระบวนการหากฎความสัมพันธ์เพื่อช่วยลดจำนวนกฎความสัมพันธ์ที่ได้ออกมาและช่วยให้ได้กฎที่ให้ค่าสนับสนุนที่สูงขึ้น ในอดีตนั้นได้มีหลากหลายงานวิจัยที่ได้นำเสนอเทคนิคในการแบ่งช่วงของข้อมูล แต่เทคนิคเหล่านั้นส่วนใหญ่แล้วจะเป็นการแบ่งช่วงของข้อมูลสำหรับการจำแนกข้อมูล (Classification) ซึ่งสามารถวัดประสิทธิภาพของเทคนิคได้ชัดเจนจากการพิจารณาค่าความถูกต้อง (Accuracy) ของโมเดล แต่การแบ่งช่วงข้อมูลเพื่อใช้กับการหากฎความสัมพันธ์จะวัดประสิทธิภาพได้ยากกว่า จึงส่งผลให้มีงานวิจัยด้านนี้ปรากฏค่อนข้างน้อย

จากที่กล่าวมาแล้วนั้นผู้วิจัยจึงได้เสนอเทคนิคการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการหากฎความสัมพันธ์ เพื่อเพิ่มประสิทธิภาพในการหากฎความสัมพันธ์ให้ดียิ่งขึ้น

1.2 วัตถุประสงค์ของการวิจัย

จากแนวคิดในการทำงานวิจัย ผู้วิจัยได้ตั้งวัตถุประสงค์ในการวิจัยไว้ดังนี้

1) เพื่อศึกษาและพัฒนาวิธีการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการหากฎความสัมพันธ์
2) เพื่อเพิ่มประสิทธิภาพในการหากฎความสัมพันธ์ให้ได้กฎที่มีจำนวนน้อยลง แต่ให้ค่าสนับสนุนที่สูงขึ้น โดยยังคงความเชื่อมั่นในระดับที่ยอมรับได้
3) เพื่อศึกษาและพัฒนาการทดสอบประสิทธิภาพของกฎความสัมพันธ์ ให้มีความชัดเจนมากยิ่งขึ้น

1.3 ข้อตกลงเบื้องต้น

1) ข้อมูลที่ใช้ทดสอบประสิทธิภาพของวิธีการแบ่งช่วงข้อมูลจะเป็นข้อมูลสังเคราะห์และข้อมูลจริงจาก UCI Machine Learning Repository (http://archive.ics.uci.edu/ml)
2) การทดสอบประสิทธิภาพของการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการหากฎความสัมพันธ์ จะใช้มาตรวัด 4 มาตรวัด ได้แก่ Support, Confidence, Lift และ Coverage
3) งานวิจัยนี้เลือกใช้ภาษา R ในการพัฒนาโปรแกรม

1.4 ขอบเขตของการวิจัย

จากการศึกษาค้นคว้าข้อมูล ผู้วิจัยได้กำหนดขอบเขตของการวิจัยไว้ดังนี้

1) ในการหากฎความสัมพันธ์จะเลือกใช้อัลกอริทึม Apriori
2) การเปรียบเทียบประสิทธิภาพการแบ่งช่วงข้อมูลจะเปรียบเทียบเทคนิคที่ได้เสนอในการวิจัยนี้กับเทคนิคการแบ่งช่วงข้อมูลอื่น ๆ โดยใช้ไลบรารีที่เกี่ยวข้องกับการหากฎความสัมพันธ์ซึ่งนิยมใช้ในหลายงานวิจัย
3) งานวิจัยนี้ใช้ภาษาอาร์ (R language) ทั้งในส่วนของการพัฒนาอัลกอริทึมการแบ่งช่วงข้อมูลและการทดสอบประสิทธิภาพ

1.5 ประโยชน์ที่ได้รับ

ประโยชน์ที่เกิดขึ้นจากงานวิจัยนี้ ประกอบด้วย

1) อัลกอริทึมการแบ่งช่วงข้อมูลที่พัฒนาขึ้นสามารถแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องได้และเหมาะสมกับการนำไปใช้ในขั้นตอนการหากฎความสัมพันธ์
2) อัลกอริทึมการแบ่งช่วงข้อมูลที่พัฒนาขึ้นสามารถช่วยเพิ่มประสิทธิภาพการหากฎความสัมพันธ์ให้ดียิ่งขึ้น
3) มาตรวัดความถูกต้องของกฎความสัมพันธ์ เป็นแนวทางใหม่ที่เสนอขึ้นในงานวิจัยนี้ ที่จะช่วยให้การประเมินประสิทธิภาพกฎความสัมพันธ์ มีความชัดเจนมากยิ่งขึ้น

บทที่ 2
ปริทัศน์วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงปริทัศน์วรรณกรรมและงานวิจัยที่เกี่ยวข้อง โดยมีรายละเอียดของการแบ่งช่วงข้อมูล (Discretization method) มาตรวัด (Measure) การหากฎความสัมพันธ์ (Association rule mining) การเขียนโปรแกรมด้วยภาษาอาร์ (R language) และงานวิจัยที่เกี่ยวข้อง

2.1 วิธีแบ่งช่วงข้อมูล (Discretization method)

ข้อมูลที่จะนำมาวิเคราะห์ด้วยการทำเหมืองข้อมูลสามารถมีลักษณะที่แตกต่างกันออกไปได้หลากหลายรูปแบบ เช่น ข้อมูลที่เป็นข้อความ ข้อมูลที่เป็นตัวเลข และข้อมูลที่เป็นทั้งข้อความและตัวเลขผสมกัน ซึ่งในส่วนของข้อมูลที่เป็นตัวเลขเมื่อนำมาผ่านกระบวนการทำเหมืองข้อมูลทั้งการทำเหมืองข้อมูลแบบจำแนก (Classification) และแบบหาความสัมพันธ์ (Association) นั้นผลลัพธ์ที่ได้ออกมานั้นไม่ดีเท่าที่ควร ดังนั้นจึงได้มีวิธีการจัดการข้อมูลที่เป็นตัวเลขต่อเนื่องเรียกว่าการแบ่งช่วงข้อมูล (Discretization) เพื่อจัดการแบ่งข้อมูลที่เป็นตัวเลขต่อเนื่องออกเป็นช่วงของข้อมูล ทำให้ข้อมูลถูกแปลงจากข้อมูลตัวเลข (Numeric) เป็นข้อมูลเชิงกลุ่ม (Categorical) และเมื่อนำไปผ่านกระบวนการทำเหมืองข้อมูลก็จะได้ประสิทธิภาพที่สูงขึ้น ซึ่งในปัจจุบันได้มีงานวิจัยเกี่ยวกับการแบ่งช่วงข้อมูลหลากหลายวิธีสามารถแยกออกเป็น 3 แบบ (Liu et al, 2002) ดังต่อไปนี้

2.1.1 อัลกอริทึมแบบล่างขึ้นบน (Bottom-up)

1. อัลกอริทึม Chi2

อัลกอริทึม Chi2 (Liu and Setiono, 1995) มีพื้นฐานมาจาก X^2 ที่นิยมใช้ในงานทางด้านสถิติถูกนำมาใช้ในการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่อง ซึ่งสามารถคำนวณหาค่า X^2 ได้ดังสมการที่ 2.1

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{k}\frac{(A_{ij} - E_{ij})^2}{E_{ij}} \tag{2.1}$$

โดยกำหนดให้

k	แทนจำนวนของคลาส
A_{ij}	แทนจำนวน pattern โดย i คือช่วง, j คือคลาส
Eij	แทนความถี่จาก $A_{ij} = R_i * C_j / N$
R_i	แทนจำนวน pattern โดยช่วง ith = $\sum_{j=1}^{k} A_{ij}$
C_j	แทนจำนวนคลาส โดยช่วง jth = $\sum_{i=1}^{2} A_{ij}$
N	แทนจำนวน pattern ทั้งหมด = $\sum_{i=1}^{2} R_i$

```
Phase 1:
 set sigLevel = .5;
 do while (InConsistency(data) < δ) {
   for each numeric attribute {
     Sort(attribute, data);
     chi-sq-initialization(attribute, data);
     do {
       chi-sq-calculation(attribute,data)
     } while (Merge(data))
   }
   sigLevel0 = sigLevel;
   sigLevel = decreSigLevel(sigLevel);
}
Phase 2:
 set all sigLvl[i] = sigLevel0 for attribute i;
 do until no-attribute-can-be-merged {
   for each attribute i that can be merged {
     Sort(attribute, data);
     chi-sq-initialization(attribute, data);
     do {
       chi-sq-calculation(attribute,data)
     } while (Merge(data))
     if (InConsistency(data) < δ)
       sigLvl[i] = decreSigLevel(sigLvl[i]);
     else
       attribute i cannot be merged;
   }
 }
```

รูปที่ 2.1 Pseudo code ของอัลกอริทึม Chi2 (Liu and Setiono, 1995)

จากรูปที่ 2.1 แสดง Pseudo code ของอัลกอริทึม Chi2 ซึ่งจะแบ่งการทำงานออกเป็นสองส่วนด้วยกัน ส่วนที่ 1 จะเริ่มต้นด้วยการกำหนดค่าระดับนัยสำคัญที่สูง คือ 0.5 (sigLevel = 0.5) ซึ่งจะใช้สำหรับข้อมูลทุกตัวที่เป็นตัวเลขต่อเนื่อง หลังจากนั้นจะทำการเรียงข้อมูลทุกตัวที่เป็นตัวเลขต่อเนื่อง แล้วจะทำการดำเนินการต่อไปนี้

- คำนวณหาค่า X^2 ตามสมการ
- รวมคู่ของช่วงที่ติดกันกับค่า X^2 ที่ต่ำ

ส่วนที่ 2 จะเป็นในส่วนปลีกย่อยของส่วนที่ 1 เริ่มจาก sigLevel0 ที่ได้กำหนดไว้ในส่วนที่ 1 แล้วทำการตรวจสอบความสอดคล้องหลังจากดำเนินการรวมแต่ละคอลัมน์ ถ้า inconsistency rate ไม่เกินค่าที่กำหนด ก็ทำการกำหนด sigLevel[i] สำหรับการรวมคอลัมน์ในรอบถัดไป ซึ่งกระบวนการนี้จะหยุดก็ต่อเมื่อไม่มีค่าในคอลัมน์ที่จะต้องนำมารวมในช่วง

2. อัลกอริทึม Modified Chi2

อัลกอริทึม Modified Chi2 (Shen and Tay, 2001) เป็นอัลกอริทึมที่ได้รับการปรับปรุงมาจากอัลกอริทึม Chi2 ซึ่งจะปรับปรุงให้อัลกอริทึม Chi2 นั้นเป็นการแบ่งช่วงแบบอัตโนมัติ โดยกระบวนการทำงานของอัลกอริทึม Chi2 จะแบ่งออกเป็น 2 ส่วน ในส่วนที่ 1 อัลกอริทึม Modified Chi2 ยังคงขั้นตอนกระบวนการของอัลกอริทึม Chi2 ไว้เหมือนเดิม แต่การทำงานในส่วนนี้จะเป็นแบบอัตโนมัติโดยจะทำการปรับปรุงการเพิ่มค่า significant level α ให้ทำการเพิ่มค่าโดยอัตโนมัติ และในส่วนของ consistency check ที่จะใช้เป็นเกณฑ์ในการหยุดกระบวนการ อัลกอริทึม Modified Chi2 จะกำหนดค่า x^2 ที่เหมาะสมโดยอัตโนมัติแต่ยังคงไว้ซึ่งข้อมูลเดิมเอาไว้

ในส่วนที่ 2 เป็นกระบวนการที่ดีกว่าในส่วนที่ 1 เริ่มต้นโดย significant level α_0 ที่กำหนดไว้ในส่วนที่ 1 ซึ่งแต่ละคอลัมน์ที่ i ที่มีความสัมพันธ์กับ sigLv[i] จะนำมาทำการรวมกันและ consistency check จะดำเนินการหลังจากในแต่ละคอลัมน์ทำการรวมกันแล้ว ซึ่งถ้าค่า inconsistency rate ไม่เกินอัตราที่กำหนดไว้ล่วงหน้าแล้ว sigLv[i] จะลดลงสำหรับคอลัมน์ที่ i ในรอบถัดไปของการรวม มิฉะนั้นคอลัมน์ที่ i จะไม่มีผลต่อการรวมอีก กระบวนการนี้จะทำซ้ำจนกว่าค่าในคอลัมน์จะไม่สามารถรวมกันได้

อัลกอริทึม Modified Chi2 จะทำการปรับปรุงในส่วน inconsistency checking (ConCheck (data)< δ) ที่อยู่ในอัลกอริทึม Chi2 และถูกแทนที่ด้วยการประมาณที่มีคุณภาพ L_c หลังจากแต่ละขั้นตอนของการแบ่งช่วงข้อมูล ($L_{c-Discretized} \leq L_{c-Original}$) โดยการประมาณคุณภาพ L_c ที่ได้จากทฤษฎี Rough set ซึ่งสามารถคำนวณหาค่า L_c ได้ดังสมการที่ 2.2

$$L_c = \frac{\sum card(\underline{B}X_i)}{card(U)} \tag{2.2}$$

โดยกำหนดให้

U	แทนเซตทั้งหมดของวัตถุในชุดข้อมูล
X	แทนการเป็นสับเซตของ *U*
BX	แทนการประมาณขั้นต่ำของ *X* ใน *B* ($B \subseteq A$)
A	แทนเซตของคอลัมน์

3. อัลกอริทึม Extended Chi2

อัลกอริทึม Extended Chi2 (Su and Hsu, 2005) เป็นการนำเอาทฤษฎี Rough Set มาใช้ในขั้นตอนของ Inconsistency check ในอัลกอริทึม Chi2 และอัลกอริทึม Modified Chi2 ใน อัลกอริทึม Extended Chi2 สามารถคำนวณค่า inconsistency rate ได้ดังสมการที่ 2.3

$$\varepsilon(C, D) = max(m_1, m_2) \tag{2.3}$$

โดยกำหนดให้

$$m_1 = 1 - min\{c(E, D) | E \in C^* \; and \; 0.5 < c(E, D)\},$$
$$m_2 = max\{c(E, D) | E \in C^* \; and \; c(E, D) < 0.5\}$$
$$c(E, D) = 1 - \frac{card(E \cap D)}{card(E)}$$

ขั้นตอนของอัลกอริทึม Extended Chi2 แสดงรายละเอียดดังนี้

1. กำหนดค่าระดับนัยสำคัญ α=0.5 จากนั้นคำนวณค่า inconsistency rate ε
2. คำนวณค่า chi-square จากคอลัมน์ที่เป็นตัวเลขทำการเรียงข้อมูลในคอลัมน์ และคำนวณค่า x^2
3. เปรียบเทียบระหว่างการคำนวณค่า x^2 และเกณฑ์ที่สอดคล้องกัน รวมช่วงที่ติดกันที่มีค่าความแตกต่างปกติและคำนวณค่า x^2 ที่น้อยกว่าเกณฑ์ที่สอดคล้องกัน ถ้าไม่มีทั้งสองช่วงที่อยู่ติดกันตามเงื่อนไขนี้แล้วข้ามไปที่ขั้นตอนที่ 5

4. ตรวจสอบค่า inconsistency rate และถ้า inconsistency rate ไม่สอดคล้องกับค่า inconsistency rate ที่กำหนดไว้แล้วจะไม่ทำการรวม ไปขั้นตอนที่ 5 มิฉะนั้นไปที่ขั้นตอนที่ 2
5. ลดระดับนัยสำคัญ $\alpha \rightarrow \alpha_0$
6. คำนวณค่า x^2 ที่ดีกว่าเดิมจากคอลัมน์ที่เป็นตัวเลขทำการเรียงข้อมูลในคอลัมน์และคำนวณค่า x^2
7. รวมช่วงที่ดีกว่าเดิม เปรียบเทียบระหว่างคำนวณค่า x^2 และเกณฑ์ที่สอดคล้องกัน รวมช่วงที่ติดกันที่มีค่าความแตกต่างปกติและคำนวณค่า x^2 ที่น้อยกว่าเกณฑ์ที่สอดคล้องกัน ถ้าไม่มีทั้งสองช่วงที่อยู่ติดกันตามเงื่อนไขนี้แล้วไปที่ขั้นตอนที่ 9
8. ตรวจสอบค่า inconsistency rate ที่ดีกว่าเดิม ตรวจสอบค่า inconsistency rate และถ้า inconsistency rate ไม่สอดคล้องกับค่า inconsistency rate ที่กำหนดไว้แล้วจะไม่ทำการรวม ไปขั้นตอนที่ 9 มิฉะนั้นไปที่ขั้นตอนที่ 6
9. ลดระดับนัยสำคัญที่ดีกว่าเดิมแล้วหยุดกระบวนการ

2.1.2 อัลกอริทึมแบบบนลงล่าง (Top-down)

1. อัลกอริทึม CAIM

อัลกอริทึม CAIM (Class-Attribute Interdependence Maximization) เป็นอัลกอริทึมแบ่งช่วงข้อมูลแบบมีผู้ฝึกสอน (Kurgan and Cios, 2004) โดยจุดประสงค์ของอัลกอริทึม CAIM คือ ทำการเพิ่มการพึ่งพาอาศัยกันของคลาสคอลัมน์และสร้างช่วงที่ได้จากการแบ่งช่วงให้น้อยที่สุด ซึ่งสามารถคำนวณหาค่า CAIM ดังสมการที่ 2.4

$$CAIM(C,D|F) = \frac{\sum_{r=1}^{n} \frac{max_r^2}{M_{+r}}}{n} \tag{2.4}$$

โดยกำหนดให้

C	แทนคลาส
D	แทนการแบ่งช่วงข้อมูล
F	แทนคอลัมน์
n	แทนจำนวนช่วงข้อมูล
$\max_r$	แทนค่าที่มากที่สุดในค่า q_{ir}
M_{+r}	แทนจำนวนคอลัมน์ทั้งหมดที่ข้อมูลเป็นตัวเลขต่อเนื่อง

กระบวนการทำงานของอัลกอริทึม CAIM จะแบ่งออกเป็น 2 ขั้นตอนหลักดังนี้

ขั้นที่ 1 มีกระบวนการทำงานดังต่อไปนี้

1.1 หาค่าที่มากที่สุด (d_n) และน้อยที่สุด (d_0) ในข้อมูลที่เป็นตัวเลขต่อเนื่อง

1.2 กำหนดรูปแบบของค่าที่แตกต่างจากข้อมูลที่เป็นตัวเลขต่อเนื่องโดยเรียงลำดับและกำหนดขอบเขตของช่วงทั้งหมดที่เป็นไปได้แทนด้วย B ตามค่าที่ต่ำสุดและค่าที่มากสุดและจุดกึ่งกลางทั้งหมดของทุกคู่ที่อยู่ติดกันในชุด

1.3 กำหนดแบบการแบ่งช่วงข้อมูล D:{[$d_0 d_n$]} และ GlobalCAIM=0

ขั้นที่ 2 มีกระบวนการทำงานดังต่อไปนี้

2.1 กำหนด k=1

2.2 ลองเพิ่มขอบเขตภายในที่ไม่ได้อยู่ใน D และ B แล้วคำนวณค่าที่สอดคล้องกับ CAIM

2.3 หลังจากขั้นตอนก่อนหน้านี้รับค่า CAIM ที่มากที่สุด

2.4 ถ้า (CAIM > GlobalCAIM or k<S) แล้วปรับปรุง D จากค่าที่ได้รับจาก 2.3 และกำหนด GlobalCAIM=CAIM ถ้าไม่เช่นนั้นก็หยุด

2.5 กำหนด k=k+1 แล้วกลับไปที่ 2.2

2. อัลกอริทึม CACC

อัลกอริทึม CACC (Class-Attribute Contingency Coefficient) เป็นอัลกอริทึมที่พัฒนามาจากอัลกอริทึม CAIM เพื่อใช้สำหรับแก้ปัญหาในกรณีที่เกิด overfitting มีลักษณะแบบ static, global, incremental, supervised and top-down โดยขึ้นอยู่กับการจัดแบ่งหมวดหมู่และความสัมพันธ์ระหว่างคลาสเป้าหมายกับคอลัมน์ (Tsai, Lee and Yang, 2008) ซึ่งสามารถคำนวณค่า CACC ได้ดังสมการที่ 2.5

$$CACC = \sqrt{\frac{y'}{y' + M'}} \tag{2.5}$$

โดยกำหนดให้

M' แทนด้วยจำนวนข้อมูลที่สุ่มมา

$$y' = M\left[\left(\sum_{i=1}^{S}\sum_{r=1}^{n}\frac{q_{ir}^2}{M_{i+}M_{+r}}\right)-1\right]/\log(n)$$

เมื่อ

M คือจำนวนข้อมูลที่สุ่มมา

n คือจำนวนของช่วงข้อมูล

q_{ir} คือจำนวนคลาสที่ i ที่ถูกสุ่มมา (i=1,2,...,S และ r=1,2,...,n) ในช่วง $(d_{r-1},d_r]$

M_{i+} คือจำนวนคลาสที่ i ถูกสุ่มมา

M_{+r} คือจำนวนช่วงที่ถูกสุ่มมา

```
1  Input:  Dataset with  i continuous attribute, M examples and S target classes;
2  Begin
3     For each continuous attribute A_i
4        Find the maximum d_n and the minimum d_0 values of  A_i;
5        Form a set of all distinct values of A in ascending order;
6        Initialize all possible interval boundaries B with the minimum and maximum
7        Calculate the midpoints of all the adjacent pairs in the set;
8        Set the initial discretization scheme as D: {[d_0,d_n]}and Globalcacc = 0;
9        Initialize k = 1;
10          For each  inner boundary  B which is not already in  scheme D,
11               Add it into D;
12               Calculate the corresponding cacc value;
13          Pick up the scheme D' with the highest cacc value;
14          If cacc  > Globalcacc or k < S  then
15               Replace D with D';
16               Globalcacc = cacc;
17               k = k + 1;
18               Goto Line 10;
18          Else
19             D' = D;
20          End If
21    Output the Discretization scheme D' with k intervals for continuous attribute A_i;
22 End
```

รูปที่ 2.2 Pseudo code ของอัลกอริทึม CACC (Tsai, Lee and Yang, 2008)

จากรูปที่ 2.2 แสดง Pseudo code ของอัลกอริทึม CACC โดยกำหนดให้ *i* คือคอลัมน์ที่เป็นตัวเลขต่อเนื่อง, *M* คือข้อมูลที่สุ่มมา และ *S* คือคลาสเป้าหมาย โดยมีขั้นตอนของอัลกอริทึม CACC เป็นดังนี้

1. เริ่มต้นดึงข้อมูลจากคอลัมน์ที่เป็นตัวเลขต่อเนื่องเก็บไว้ที่ A_i, หาค่า maximum (d_n) และ minimum (d_0)
2. กำหนดช่วง boundaries (B) จากค่า maximum และ minimum
3. คำนวณจุดกึ่งกลางของทุกขอบเขตที่อยู่ติดกัน
4. กำหนดรูปแบบการแบ่งช่วงโดย D:{[d_0,d_n]} และ Globalcacc = 0
5. กำหนด k=1
6. การวนซ้ำที่ *k*th, คำนวณทุกจุดตัดที่เป็นไปได้โดยหาจาก cacc ที่มีค่าสูงสุด แล้วแบ่งคอลัมน์ตามช่วง k+1
7. สร้างรูปแบบการแบ่งช่วงย่อยที่ดีที่สุดโดยใช้วิธีการเหมือนกันกับอัลกอริทึม CAIM
8. ถ้าสร้างค่า cacc รอบที่ k+1 น้อยกว่าค่า Globalacc ในรอบที่ k จะยุติกระบวนการและส่งรูปแบบการแบ่งช่วงข้อมูลออกเป็นผลลัพธ์

3. อัลกอริทึม Ameva

อัลกอริทึม Ameva (Abril et al, 2009) เป็นอัลกอริทึมการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องแบบมีผู้ฝึกสอน ซึ่งจะช่วยเพิ่มค่าการหาความสัมพันธ์ระหว่างตัวแปรที่อยู่ใน Chi-square ในทางสถิติและจะสร้างช่วงของข้อมูลให้น้อยที่สุด โดยอัลกอริทึม Ameva ไม่จำเป็นต้องกำหนดจำนวนของช่วงข้อมูลที่ต้องการจะแบ่ง ซึ่งสามารถคำนวณค่า Ameva ได้ดังสมการที่ 2.6

$$Ameva(k) = \frac{x^2(k)}{k(l-1)} \tag{2.6}$$

โดยกำหนดให้

k แทนด้วยจำนวนช่วงข้อมูล

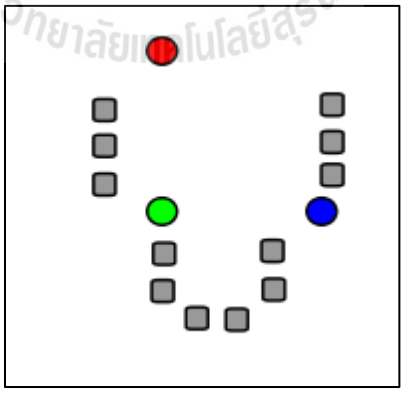
$$x^2(k) = N\left(-1 + \sum_{i=1}^{l}\sum_{j=1}^{k}\frac{n_{ij}^2}{n_i n_j}\right)$$

2.1.3 อัลกอริทึมการแบ่งกลุ่มข้อมูล (Clustering)

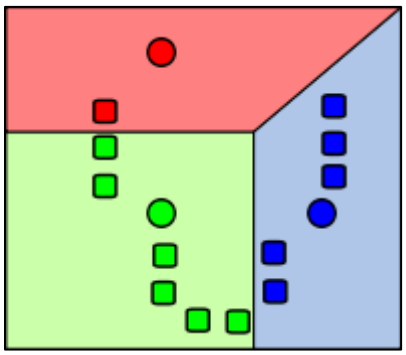
อัลกอริทึมการแบ่งกลุ่มข้อมูลเป็นกระบวนการวิเคราะห์ข้อมูลอย่างหนึ่งที่ใช้ในงานทางด้านการทำเหมืองข้อมูล โดยจะแบ่งข้อมูลที่มีลักษณะที่คล้ายกันไว้ในกลุ่มเดียวกัน (Cluster) โดยเกณฑ์ที่ใช้แบ่งกลุ่มนั้นจะวัดจากความเหมือน (Similarity) หรือความใกล้ชิด (Proximity) ซึ่งหาได้จากหลายวิธี เช่น การวัดระยะแบบยูคลิด(Euclidean distance) และการวัดระยะแบบแมนฮัตตัน (Manhattan distance) เป็นต้น (Wikipedia, 2012c)

อัลกอริทึมการแบ่งกลุ่มข้อมูลหนึ่งที่นิยมนำมาใช้กันอย่างแพร่หลาย คือ การจัดกลุ่มข้อมูลแบบ k-means clustering ซึ่งใช้วิธีแบ่งข้อมูลออกเป็นกลุ่มวัดจากค่าระยะทางที่น้อยที่สุดระหว่างข้อมูล และจุดศูนย์กลางของแต่ละกลุ่ม (Cluster centroid) โดยมีขั้นตอนของการจัดกลุ่มข้อมูลแบบ k-means clustering (Wikipedia, 2012d) เป็นดังนี้

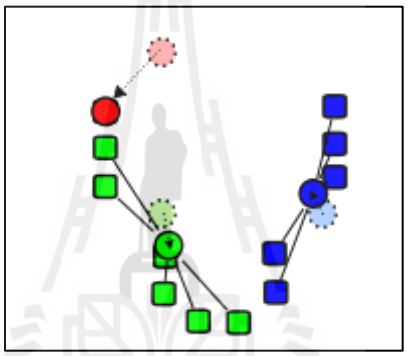
1. กำหนดจำนวนกลุ่ม K กลุ่ม และกำหนดจุดศูนย์กลางเริ่มต้น (Centroid) จำนวน K จุด (รูปที่ 2.3 เมื่อ K คือ 3 และแทนจุดศูนย์กลางด้วยภาพวงกลม)
2. นำข้อมูลทั้งหมด (แทนด้วยภาพสี่เหลี่ยม) จัดเข้ากลุ่มที่มีจุดศูนย์กลางที่อยู่ใกล้ข้อมูลนั้นมากที่สุด โดยคำนวณจากการวัดระยะห่างระหว่างข้อมูลกับจุดศูนย์กลาง (รูปที่ 2.4)
3. คำนวณจุดศูนย์กลาง K จุดใหม่ โดยหาจากค่าเฉลี่ยทุกข้อมูลที่อยู่ในกลุ่ม (รูปที่ 2.5)
4. ทำซ้ำในข้อ 2 จนกระทั่งจุดศูนย์กลางไม่เปลี่ยนแปลง (รูปที่ 2.6)



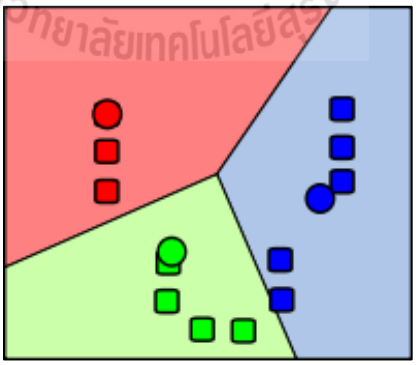
รูปที่ 2.3 กำหนดจำนวนกลุ่ม 3 กลุ่ม และกำหนดจุดศูนย์กลางเริ่มต้น (Centroid) จำนวน 3 จุด (Wikipedia, 2012d)



รูปที่ 2.4 นำข้อมูลทั้งหมดจัดเข้ากลุ่มที่มีจุดศูนย์กลางที่อยู่ใกล้ข้อมูลนั้นมากที่สุด (Wikipedia, 2012d)



รูปที่ 2.5 คำนวณจุดศูนย์กลาง 3 จุดใหม่ (Wikipedia, 2012d)



รูปที่ 2.6 ทำซ้ำในข้อ 2 จนกระทั่งจุดศูนย์กลางไม่เปลี่ยนแปลง (Wikipedia, 2012d)

2.2 ค่าสหสัมพันธ์ (Correlation)

สหสัมพันธ์ คือ การหาความสัมพันธ์เชิงเส้น (Linear relationship) ระหว่างตัวแปร 2 ตัวแปรขึ้นไป ซึ่งค่าสหสัมพันธ์ที่คำนวณออกมาได้ เรียกว่า ค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation coefficient) ใช้สัญลักษณ์ r แทนระดับความสัมพันธ์ (อิศรัฏฐ์ รินไธสง, 2012) ดังตารางที่ 2.1 โดยสามารถคำนวณค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปร x และ y ได้ดังสมการที่ 2.7

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{S_x}\right)\left(\frac{y_i - \bar{y}}{S_y}\right) \tag{2.7}$$

โดยกำหนดให้

n	แทนจำนวนข้อมูล
x_i	แทนข้อมูล x ตัวที่ i
$\bar{x}$	แทนค่าเฉลี่ยตัวแปร x
S_x	แทนค่าเบี่ยงเบนมาตรฐานตัวแปร x
y_i	แทนข้อมูล y ตัวที่ i
$\bar{y}$	แทนค่าเฉลี่ยตัวแปร y
S_y	แทนค่าเบี่ยงเบนมาตรฐานตัวแปร y

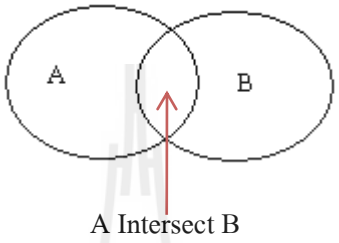
ตารางที่ 2.1 ระดับค่าสัมบูรณ์ของความสัมพันธ์เชิงเส้น (Hinkle et al, 1998)

|r|	**ระดับความสัมพันธ์**
0.00 - 0.30	มีความสัมพันธ์ในระดับต่ำมาก
0.30 - 0.50	มีความสัมพันธ์ในระดับต่ำ
0.50 - 0.70	มีความสัมพันธ์ในระดับปานกลาง
0.70 - 0.90	มีความสัมพันธ์ในระดับสูง
0.90 - 1.00	มีความสัมพันธ์ในระดับสูงมาก

2.3 มาตรวัด (Measure)

2.3.1 Support

มาตรวัดที่ใช้วัดความถี่ของเหตุการณ์สองเหตุการณ์ที่เกิดขึ้นร่วมกันหรือพร้อมกันว่ามีมากน้อยเพียงใด โดยนับความถี่ความสัมพันธ์ที่เกิดขึ้นภายในชุดข้อมูลนั้น เรียกว่า ค่าสนับสนุน (Support) แผนภาพในรูปที่ 2.7 (Wikipedia, 2012b) แสดงการเกิดขึ้นพร้อมกันของเหตุการณ์ A และ B ซึ่งหมายถึงการ intersect ของเซตของเหตุการณ์ A และเซตของเหตุการณ์ B



รูปที่ 2.7 การเกิดเหตุการณ์ A และ B

จากรูปที่ 2.7 การหาค่าสนับสนุนในส่วนที่เหตุการณ์ A และเหตุการณ์ B เกิดร่วมกัน นั้นสามารถคำนวณได้ดังสมการที่ 2.8

$$Support\{A \rightarrow B\} = P(A \wedge B) \tag{2.8}$$

ข้อมูลในตารางที่ 2.2 (Lai and Cerpa, 2001) แสดงตัวอย่างการหาค่าสนับสนุนของการเกิดเหตุการณ์ {A→B}, {B→C} และ {AB→C}

ตารางที่ 2.2 ตัวอย่างการหาค่าสนับสนุน

TID	Items	Support {X} = Occurrence {X} / Total Support
1	ABC	Total Support = 5 Support{A→B} = 2/5 = 40% Support{B→C}= 3/5 = 60% Support{AB→C}= 1/5 = 20%
2	ABD	
3	BC	
4	AC	
5	BCD	

2.3.2 Confidence

ค่าความเชื่อมั่น (Confidence) เป็นการดูความถี่ของเหตุการณ์ที่เกิดขึ้น ร่วมกับเหตุการณ์ที่เกิดอื่น ๆ ที่เกิดขึ้นร่วมกัน เช่น เมื่อเกิดเหตุการณ์ A แล้วบ่อยแค่ไหนที่จะเกิดเหตุการณ์ B (Wikipedia, 2012b) ซึ่งสามารถคำนวณหาค่าความเชื่อมั่นได้ดังสมการที่ 2.9

$$Confidence\,\{A \rightarrow B\} = \frac{\text{ความถี่ของ } A \text{ และ } B}{\text{ความถี่ของ } A} \tag{2.9}$$

จากตารางที่ 2.3 เป็นการแสดงตัวอย่างการคำนวณค่าความเชื่อมั่นของกฎความสัมพันธ์ {A→B}, {B→C} และ {AB→C}

ตารางที่ 2.3 ตัวอย่างการหาค่าความเชื่อมั่น (Lai and Cerpa, 2001)

TID	**Items**	**Given X → Y** **Confidence = Occurrence{X∧Y} / Occurrence{X}**
1	ABC	Confidence{A→B} = 2/3 = 66% Confidence{B→C} = 3/4 = 75% Confidence{AB→C} = 1/2 = 50%
2	ABD	
3	BC	
4	AC	
5	BCD	

2.3.3 Lift

Lift จะเป็นมาตรวัดที่ใช้วัดประสิทธิภาพกฎความสัมพันธ์ โดยจะวัดอิทธิพลของกฎความสัมพันธ์ที่เกิดขึ้น (Wikipedia, 2012a) ซึ่งสามารถคำนวณได้ดังสมการที่ 2.10

$$Lift\,\{A \rightarrow B\} = \frac{\text{ความเชื่อมั่นของ } A \rightarrow B}{P(B)} \tag{2.10}$$

ค่า Lift ที่ได้ออกมานั้นจะสามารถบอกความเป็นไปได้ว่า ถ้าเกิดเหตุการณ์ A แล้วจะเกิดเหตุการณ์ B ซึ่งทั้งสองเหตุการณ์นั้นจะต้องขึ้นต่อกัน ในกรณีที่ค่า Lift ที่ได้ออกมานั้นมีค่ามากหมายถึงกฎความสัมพันธ์ที่ได้นั้นมีความสำคัญมากพอที่จะนำไปใช้ในการทำนาย ตัวอย่างการคำนวณค่า Lift ของเหตุการณ์ {A→B}, {B→C} และ {AB→C} แสดงได้ดังตารางที่ 2.4

ตารางที่ 2.4 ตัวอย่างการหาค่าลิฟท์

TID	Items	Given X → Y Lift = Confidence{X→Y} / Support{Y}
1	ABC	Lift {A→B} = 0.66/0.80 = 83.33% Lift {B→C} = 0.75/0.80 = 93.75% Lift {AB→C} = 0.50/0.80 = 62.50%
2	ABD	
3	BC	
4	AC	
5	BCD	

2.3.4 Coverage

ค่าครอบคลุม (Coverage) เป็นมาตรวัดที่ใช้พิจารณาความถี่ของกฎความสัมพันธ์จาก A → B ที่เหมาะสมจากฐานข้อมูล (Michael Hahsler, 2011) ซึ่งสามารถคำนวณได้ดังสมการที่ 2.11

$$Coverage\{A \rightarrow B\} = Support\{A\} = P(A) \tag{2.11}$$

ตัวอย่างการคำนวณค่า Coverage ของเหตุการณ์ {A→B}, {B→C} และ {AB→C} แสดงได้ดังตารางที่ 2.5

ตารางที่ 2.5 ตัวอย่างการหาค่าครอบคลุม

TID	Items	Given X → Y Coverage = Support {X}
1	ABC	Coverage {A→B} = 3/5 = 60% Coverage {B→C} = 4/5 = 80% Coverage {AB→C} = 2/5 = 40%
2	ABD	
3	BC	
4	AC	
5	BCD	

2.4 การหากฎความสัมพันธ์ (Association rule mining)

การหากฎความสัมพันธ์ (Association rules mining) เป็นเทคนิคหนึ่งในการทำเหมืองข้อมูลที่ค้นหาความสัมพันธ์ของเหตุการณ์หรือวัตถุที่เกิดขึ้นร่วมกันหรือพร้อมกัน แล้วนำมาสร้างกฎขึ้นมาเพื่อที่จะทำนายเหตุการณ์หรือการเกิดขึ้นของวัตถุนั้น ๆ ในอนาคต ซึ่งการหากฎความสัมพันธ์นั้นสามารถนำไปใช้งานได้หลายรูปแบบ เช่น การวิเคราะห์พฤติกรรมซื้อสิ้นค้า ถ้าลูกค้าซื้อสินค้า A แล้วมักจะซื้อ B ตามไปด้วย เป็นต้น และได้มีงานวิจัยที่เกี่ยวข้องกับการเพิ่มประสิทธิภาพในการหากฎความสัมพันธ์ในรูปแบบต่าง ๆ เช่น การเพิ่มความเร็วในการหากฎความสัมพันธ์ การเพิ่มค่าความถูกต้อง (Accuracy) หรือการประยุกต์ใช้การหากฎความสัมพันธ์กับข้อมูลในรูปแบบต่าง ๆ เป็นต้น (Agrawal et al,1993) ในการหากฎความสัมพันธ์นั้นจะต้องมีการนับความถี่ของการเกิดเหตุการณ์ซึ่งจะอธิบายด้วยข้อมูลตัวอย่างตามตารางที่ 2.6

ตารางที่ 2.6 รายการซื้อสินค้าของลูกค้าทั้งหมด

รายการสินค้า	นม	น้ำ	ขนม	ไส้กรอก
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

จากตารางที่ 2.6 เป็นข้อมูลรายการซื้อสินค้าของลูกค้า แล้วนำไปผ่านการหาความถี่ของการซื้อสินค้าของลูกค้าในแต่ละชิ้นของสินค้า เพื่อหาความสัมพันธ์ของสินค้าแต่ละอย่างซึ่งจะแสดงได้ดังตารางที่ 2.7 หลังจากนั้นก็จะนำสินค้าที่มีความถี่สูงไปสร้างกฎความสัมพันธ์ ซึ่งอยู่ในรูปแบบของ If condition Then result โดยเกณฑ์ที่ใช้ในการหากฎนั้น มีดังนี้

- Support เป็นค่าที่บอกว่าความถี่ที่เกิดขึ้นบ่อยมากน้อยแค่ไหน
- Confidence เป็นค่าที่บอกว่าโอกาสที่จะเกิดขึ้น เช่น ถ้ามี condition เกิดขึ้น โอกาสที่จะเกิด result มีมากน้อยแค่ไหน

ตารางที่ 2.7 ความถี่ของการซื้อสินค้าของลูกค้า เพื่อหาความสัมพันธ์ของสินค้าแต่ละอย่าง

	นม	น้ำ	ขนม	ไส้กรอก
นม	2*	2	1	0
น้ำ	2	4*	2	0
ขนม	1	1	2*	0
ไส้กรอก	0	0	0	1*

จากข้อมูลตัวอย่างตามตารางที่ 2.6 และ 2.7 สามารถค้นหากฎความสัมพันธ์ที่ประกอบด้วยมาตรวัด Support, Confidence, Lift และ Coverage แสดงได้ดังตารางที่ 2.8

ตารางที่ 2.8 กฎความสัมพันธ์และมาตรวัด Support, Confidence, Lift และ Coverage

กฎความสัมพันธ์	Support	Confidence	Lift	Coverage
ขนม => นม	0.2	0.5	1.25	0.4
นม => ขนม	0.2	0.5	1.25	0.4
ขนม => น้ำ	0.4	1.0	1.25	0.4
น้ำ => ขนม	0.4	0.5	1.25	0.8
นม => น้ำ	0.4	1.0	1.25	0.4
น้ำ => นม	0.4	0.5	1.25	0.8
นม,ขนม => น้ำ	0.2	1.0	1.25	0.2
น้ำ,ขนม => นม	0.2	0.5	1.25	0.4
นม,น้ำ => ขนม	0.2	0.5	1.25	0.4

2.5 การเขียนโปรแกรมด้วยภาษาอาร์ (R language programming)

ภาษาอาร์เป็นภาษาที่ถูกพัฒนาขึ้นมาสำหรับใช้ในการคำนวณทางด้านสถิติและแสดงผลการคำนวณด้วยกราฟิก ซึ่งเป็นภาษาที่นิยมใช้กันอย่างแพร่หลาย เนื่องจากเป็นโอเพนซอร์ซ (Open source) (Paradis et al. 2004) ภาษาอาร์เป็นภาษาเชิงฟังก์ชัน ที่มีลักษณะคล้ายกับภาษา S เนื่องจากถูกพัฒนาขึ้นมาเพื่อใช้ทดแทนภาษา S โดยได้ถูกพัฒนาขึ้นในปี ค.ศ. 1995 จากภาควิชาสถิติ มหาวิทยาลัยโอคแลนด์ ประเทศนิวซีแลนด์ ผู้พัฒนาคือ Robert Gentleman และ Ross Ihaka (กิตติ ศักดิ์ เกิดประสพ, 2012)

2.5.1 การใช้งานภาษาอาร์ (R Tutorial)

ภาษาอาร์เป็นภาษาเชิงฟังก์ชันที่ผนวกกับแนวคิดเชิงวัตถุ (Object) ซึ่งจะมีการเก็บตัวแปร ข้อมูล และฟังก์ชันอยู่ในรูปแบบเชิงวัตถุ ดังแสดงในรูปที่ 2.8

```
> x <- 20
> y -> 30
> z = 40
```

รูปที่ 2.8 ตัวอย่างการกำหนดค่าให้กับตัวแปรในภาษาอาร์

ในการกำหนดค่าให้กับตัวแปรนั้นสามารถกำหนดได้หลายรูปแบบดังแสดงในตัวอย่าง (รูปที่ 2.8) ซึ่งทุกรูปแบบมีความหมายเหมือนกันแต่ส่วนมากมักใช้ x <- 20 เนื่องจากเป็นการสื่อถึงการอ้างถึงวัตถุ นอกจากนี้ภาษาอาร์ยังสามารถใส่นิพจน์ทางคณิตศาสตร์ในระหว่างการอ้างถึงวัตถุ (กิตติศักดิ์ เกิดประสพ, 2012) ดังแสดงในรูปที่ 2.9

```
> x <- 20
> y <- 20^2
```

รูปที่ 2.9 ตัวอย่างการใส่นิพจน์ทางคณิตศาสตร์ในระหว่างการอ้างถึงวัตถุ

2.5.2 เวกเตอร์ (Vector)

การเก็บข้อมูลในภาษาอาร์โครงสร้างข้อมูลมักจะอยู่ในรูปแบบของเวกเตอร์ ซึ่งสามารถเก็บข้อมูลภายในได้ครั้งละจำนวนมาก แต่ชนิดของข้อมูลภายในเวกเตอร์นั้นต้องเป็นชนิดเดียวกันเท่านั้น (กิตติศักดิ์ เกิดประสพ, 2012) ดังแสดงในรูปที่ 2.10

```
> c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
[1] 1 2 3 4 5 6 7 8 9 10
> c(1:100)
[1] 1 2 3 … 100
> c(“A”, “B”, “C”, “D”)
[1] A B C D
> c(TRUE, FALSE, FALSE, TRUE)
[1] TRUE FALSE FALSE TRUE
```

รูปที่ 2.10 ตัวอย่างการสร้างเวกเตอร์

ในส่วนของการเรียกใช้ข้อมูลเวกเตอร์นั้นสามารถใช้ [ ] ในการเรียกใช้ ดังแสดงในรูปที่ 2.11

```
> x <- c(11,12,13,14,15,16,17,18,19,20)
> x[1]
[1] 1
> x[c(1,2,3,4)]
[1] 11 12 13 14
```

รูปที่ 2.11 ตัวอย่างการเรียกใช้เวกเตอร์

2.5.3 เฟรมข้อมูล (Data frame)

เฟรมข้อมูลเป็นลักษณะของข้อมูลที่นิยมนำมาใช้กันเป็นส่วนมากเนื่องจากเมื่อข้อมูลที่จะนำมาใช้มีจำนวนมากแล้วการเก็บข้อมูลในรูปแบบของเฟรมข้อมูล นั้นจะสะดวกสำหรับการเรียกใช้ข้อมูล ซึ่งข้อมูลในแต่ละคอลัมน์ของเฟรมข้อมูล นั้นสามารถที่จะมีชนิดต่างกันได้ แต่ภายในคอลัมน์นั้น ๆ ต้องเป็นข้อมูลชนิดเดียวกัน เฟรมข้อมูลสามารถสร้างได้ด้วยคำสั่งง่าย ๆ ดังแสดงในรูปที่ 2.12

```
> data.frame(Age=c('20','21','18','19','15'),
                    + Sex =  c('M','F','F','M','M'))
> my.dataset
  Age    Sex
1  20      M
2  21      F
3  18      F
4  19      M
5  15      M
```

รูปที่ 2.12 ตัวอย่างการสร้างเฟรมข้อมูล

การเรียกใช้เฟรมข้อมูลนั้นสามารถเรียกใช้ได้หลายแบบ ดังแสดงในรูปที่ 2.13

```
> my.dataset$Age
[1]  20   21   18   19    15
> attach ( my.dataset )
> Age
[1] 20   21   18   19    15
> detach (my.dataset)
```

รูปที่ 2.13 ตัวอย่างการเรียกใช้เฟรมข้อมูล

2.5.4 ฟังก์ชันในภาษาอาร์

ภาษาอาร์เป็นภาษาเชิงฟังก์ชันดังนั้นผู้ใช้สามารถสร้างฟังก์ชันเพื่อเก็บชุดคำสั่งไว้เรียกใช้งานเมื่อมีความต้องการ ซึ่งฟังก์ชันที่กำหนดขึ้นมานั้นจะมีลักษณะเชิงวัตถุ โดยมีรูปแบบการเขียนฟังก์ชันดังแสดงในรูปที่ 2.14

```
ชื่อฟังก์ชัน <- function(ชุดของพารามิเตอร์) {
                ชุดของคำสั่ง
            }
```

รูปที่ 2.14 รูปแบบการเขียนฟังก์ชันด้วยภาษาอาร์

การเขียนฟังก์ชันในภาษาต้องเขียน 1 คำสั่งต่อ 1 บรรทัดโดยถ้ามีการส่งค่ากลับ ให้ใช้คำสั่ง return() แต่ถ้าไม่มีคำสั่ง return() จะใช้ผลลัพธ์ในคำสั่งสุดท้ายเป็นการส่งค่ากลับ ดังแสดงในรูปที่ 2.15

```
>  fr <- function(data){
                x<-length(data)
                return(x)
        }
> fr(c(10:100))
[1] 90
```

รูปที่ 2.15 ตัวอย่างการเขียนฟังก์ชันด้วยภาษาอาร์

2.6 งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้องกับการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องนั้นมีมากมาย ซึ่งจะประกอบด้วยงานในส่วนของการนำเสนอแนวคิดการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่อง การปรับปรุงอัลกอริทึมการแบ่งช่วงข้อมูล การใช้อัลกอริทึมการแบ่งช่วงข้อมูลสำหรับการหากฎความสัมพันธ์ และผู้วิจัยได้ทำการศึกษาค้นคว้างานวิจัยที่มีความเกี่ยวข้องกับงานวิจัยที่จะทำโดยมีรายละเอียดโดยสรุปดังนี้

María N. Moreno, Saddys Segrera, Vivian F. López และ José M. Polo (2006) ได้เสนอวิธีการหากฎความสัมพันธ์สำหรับข้อมูลที่มีลักษณะเชิงปริมาณเพื่อเพิ่มค่าความเชื่อมั่นให้มีค่าที่สูง โดยนำไปใช้กับการจัดการซอฟต์แวร์เมตริกซ์ ซึ่งเทคนิคที่ได้นำเสนอเป็นแบบมีผู้ฝึกสอน คือใช้อัลกอริทึม k-means ในการแบ่งช่วงข้อมูลหลายคอลัมน์โดยใช้การหาระยะทางแบบยูคลิด การทดลองจะใช้ข้อมูลจำลองแบบไดนามิกที่พัฒนาโดย Ramos และคณะ โดยการทดสอบ

ประสิทธิภาพจะใช้มาตรวัด 2 มาตรวัด ได้แก่ ค่าความเชื่อมั่น และค่าสนับสนุน ผลการทดลองปรากฏว่า ได้กฎความสัมพันธ์และช่วงของข้อมูลที่มีจำนวนลดลง และกฎความสัมพันธ์ที่ได้นั้นมีค่าความเชื่อมั่นและค่าสนับสนุนที่สูง

Yiping Ke, James Cheng และ Wilfred Ng (2008) ได้เสนอโครงสร้าง MIC (Mutual Information and Clique) เพื่อต้องการแก้ปัญหาข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการหากฎความสัมพันธ์ โดยใช้เทคนิคการแบ่งช่วงข้อมูลซึ่งแบ่งกระบวนการทำงานออกเป็น 3 ส่วน ได้แก่ ส่วนที่ 1 แบ่งช่วงข้อมูลในคอลัมน์ที่เป็นตัวเลขต่อเนื่อง ส่วนที่ 2 นำข้อมูลที่ได้จากส่วนที่ 1 ไปสร้าง MI Graph และส่วนที่ 3 นำผลที่ได้จากส่วนที่ 2 มาใช้ในการหาความถี่ของ itemset การทดลองใช้ชุดข้อมูลจำนวน 6 ชุด ได้แก่ synthetic, covtype, letter-recognition, ann-thyroid และ yeast ซึ่งจะทดลองในส่วนของ เวลาในการสร้างกฎความสัมพันธ์ จำนวนกฎความสัมพันธ์ และมาตรวัดต่าง ๆ

Hantian Wei (2009) ได้เสนออัลกอริทึม MVD-CG (multivariate discretization based on density-based clustering and genetic algorithm) เป็นการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องหลายตัวแปรสำหรับการหากฎความสัมพันธ์โดยใช้เทคนิคการจัดกลุ่มและเทคนิคเชิงพันธุกรรม ซึ่งปรับปรุงมาจากอัลกอริทึม MVD การทดลองจะใช้ชุดข้อมูลจริงคือ IPUMS โดยจะทำการเปรียบเทียบจำนวนกฎความสัมพันธ์และคุณภาพของกฎความสัมพันธ์ระหว่างอัลกอริทึม MVD-CG และอัลกอริทึม MVD ผลการทดลองปรากฏว่าอัลกอริทึม MVD-CG มีประสิทธิภาพที่ดีกว่าอัลกอริทึม MVD ในการหากฎความสัมพันธ์โดยมีค่าความเชื่อมั่นที่สูงกว่า

Attila Gyenesei (2001) ได้เสนออัลกอริทึมใหม่โดยใช้ Fuzzy set ในการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องเพื่อลดเวลาการหากฎความสัมพันธ์และลดจำนวนกฎความสัมพันธ์ที่ไม่มีประโยชน์ โดยจะสนใจในส่วนของ non-sharp ในเทคนิค Fuzzy set ซึ่งการทดลองจะเปรียบเทียบระหว่างการแบ่งช่วงแบบธรรมดา การแบ่งช่วงด้วย Fuzzy set โดยไม่ได้ทำการ normalization และการแบ่งช่วงด้วย Fuzzy set โดยทำการ normalization ผลการทดลองปรากฏว่า การแบ่งช่วงด้วย Fuzzy set โดยทำการ normalization ให้ประสิทธิภาพที่ดีที่สุดโดยดูจากค่า support, confidence และเวลาที่ลดลงในการหากฎความสัมพันธ์

Hyontai Sug (2011) ได้เสนอการหากฎความสัมพันธ์แบบหลายมิติซึ่งจะแตกต่างจากการหากฎความสัมพันธ์ทั่วไปตรงที่การหากฎความสัมพันธ์แบบหลายมิตินั้นจะมีคอลัมน์ที่แตกต่าง โดยในการแบ่งช่วงข้อมูลจะจัดกลุ่มย่อยของคลาส เป้าหมายและเลือกค่าอินสแตนซ์จากค่าอินสแตนซ์ทั้งหมดในคลาส ซึ่งวิธีนี้สามารถช่วยลดขนาดของชุดข้อมูลและช่วยลดเวลาในการหากฎความสัมพันธ์ การทดลองใช้ชุดข้อมูล Statlog ซึ่งเป็นชุดข้อมูลจาก UCI ผลการทดลองที่ได้คือ

สามารถสร้างตารางกฎความสัมพันธ์แบบหลายมิติที่มีขนาดเล็กลง และช่วยลดจำนวนกฎความสัมพันธ์

จากการศึกษางานวิจัยที่เกี่ยวข้องพบว่าอัลกอริทึมที่นำมาใช้แบ่งช่วงข้อมูลในแต่ละงานวิจัยนั้นมีทั้งอัลกอริทึมแบบมีผู้ฝึกสอน และไม่มีผู้ฝึกสอนซึ่งจะเหมาะสมกับลักษณะข้อมูลที่แตกต่างกันออกไป โดยงานวิจัยส่วนมากจะเสนออัลกอริทึมแบบมีผู้ฝึกสอนโดยที่มุ่งเน้นไปในส่วนการพัฒนาอัลกอริทึมเพื่อเพิ่มประสิทธิภาพของกฎความสัมพันธ์ ในงานวิจัยส่วนมากจะวัดประสิทธิของกฎความสัมพันธ์ที่ได้จากข้อมูลที่แบ่งช่วงข้อมูลแล้วจากมาตรวัดที่มาพร้อมกับกฎความสัมพันธ์ เช่น Support, Confidence เป็นต้น แต่มีงานวิจัยส่วนน้อยที่จะนำอัลกอริทึมการแบ่งช่วงข้อมูลที่ตนเองเสนอไปเปรียบเทียบกับอัลกอริทึมอื่นอย่างจริงจัง เนื่องจากการเปรียบเทียบประสิทธิภาพของกฎความสัมพันธ์โดยตรงนั้นกระทำได้ยาก ในงานวิจัยนี้ได้พัฒนาโปรแกรมแบ่งช่วงข้อมูลที่สามารถใช้กับข้อมูลที่มีลักษณะข้อมูลที่แตกต่างกันออกไปได้ และเสนอการวัดประสิทธิภาพวิธีแบ่งช่วงข้อมูลสำหรับการหากฎความสัมพันธ์จากค่าความถูกต้อง โดยสาระสำคัญในงานวิจัยนี้เมื่อเปรียบเทียบกับงานวิจัยอื่นสรุปได้ดังตารางที่ 2.9

ตารางที่ 2.9 สรุปเปรียบเทียบงานวิจัยที่เกี่ยวข้องกับการแบ่งช่วงข้อมูลสำหรับการหากฎความสัมพันธ์

กระบวนการทำงาน	งานวิจัยที่เกี่ยวข้อง					
	ก	ข	ค	ง	จ	ฉ
อัลกอริทึมการแบ่งช่วงข้อมูลสำหรับการหากฎความสัมพันธ์						
Bottom-up		✓				✓
Top-down	✓		✓	✓	✓	
Supervised	✓		✓		✓	✓
Unsupervised		✓		✓		✓
k-means	✓	✓	✓			✓
FP-growth					✓	
Fuzzy set				✓		
เกณฑ์การประเมินประสิทธิภาพกฎความสัมพันธ์						
Support	✓	✓	✓	✓	✓	✓
Confidence	✓	✓	✓	✓	✓	✓
Lift						✓
Coverage						✓
Accuracy						✓
ขอบเขตของการวิจัย						
วิจัยเพื่อทดสอบประสิทธิภาพ	✓	✓	✓	✓	✓	✓
วิจัยเพื่อเสนอแนวคิดใหม่	✓	✓	✓	✓	✓	✓
มีการประยุกต์ใช้กับข้อมูลจริง	✓	✓	✓	✓	✓	✓

หมายเหตุ งานวิจัยที่เกี่ยวข้อง ประกอบด้วย

ก แทนงานวิจัยของ María N. Moreno และคณะ (2006)

ข แทนงานวิจัยของ Yiping Ke และคณะ (2006)

ค แทนงานวิจัยของ Hantian Wei (2009)

ง แทนงานวิจัยของ Attila Gyenesei (2001)

จ แทนงานวิจัยของ Hyontai Sug (2011)

ฉ แทนงานวิจัยของ วิธีแบ่งช่วงข้อมูลสำหรับการหากฎความสัมพันธ์ (งานวิจัยของวิทยานิพนธ์ฉบับนี้)

บทที่ 3
วิธีดำเนินการวิจัย

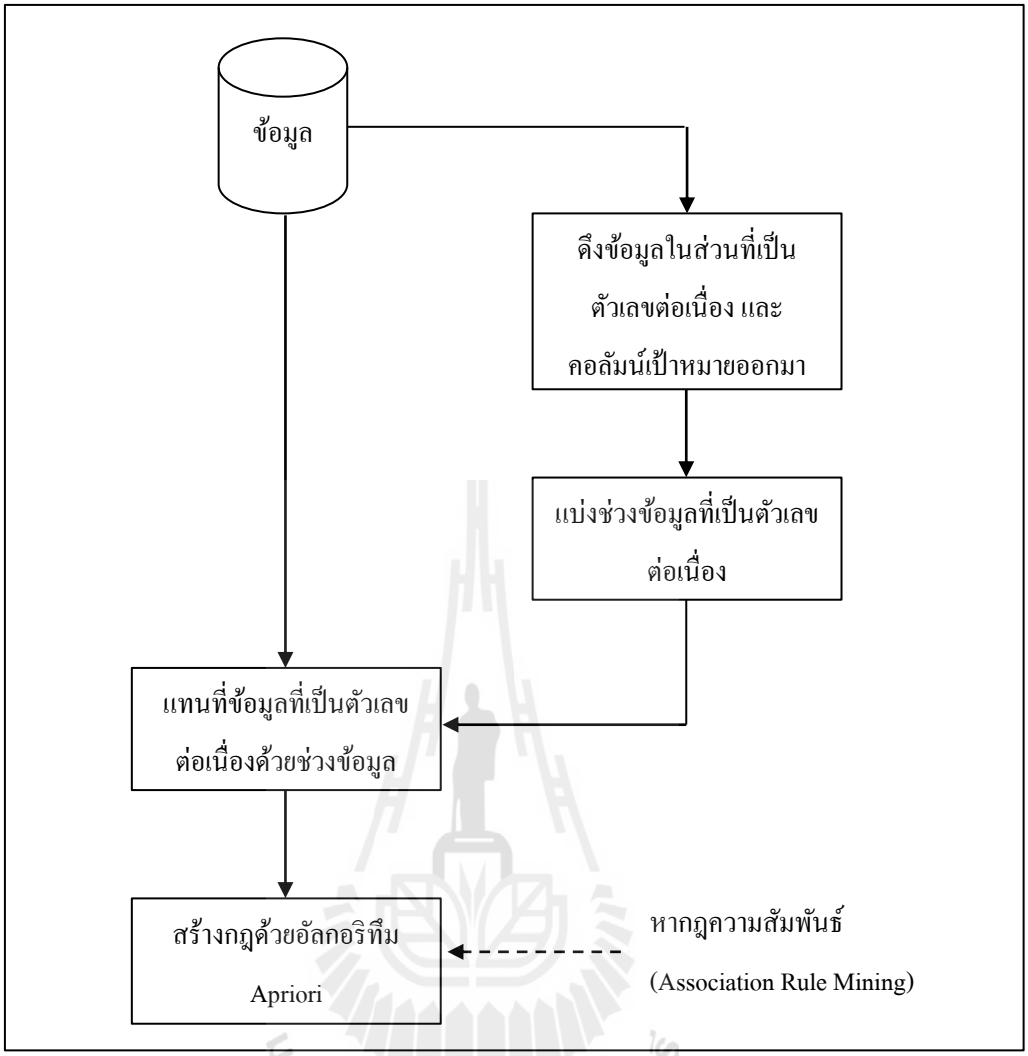
งานวิจัยนี้มีวัตถุประสงค์เพื่อเสนอวิธีแบ่งช่วงข้อมูลสำหรับการหากฎความสัมพันธ์ ในบทนี้จะกล่าวถึง วิธีการวิจัย เครื่องมือที่ใช้ในการวิจัย และกระบวนการต่าง ๆ ของการวิจัย โดยมีรายละเอียดดังนี้

3.1 กรอบแนวคิดของการวิจัย

แนวคิดหลักของงานวิจัยนี้คือ วิธีแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการหากฎความสัมพันธ์ โดยสามารถแบ่งกรอบแนวคิดของงานวิจัยนี้ออกเป็น 2 ส่วน ได้แก่ กรอบแนวคิดที่ 1 วิธีแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการหากฎความสัมพันธ์ และกรอบแนวคิดที่ 2 การทดสอบประสิทธิภาพกฎความสัมพันธ์จากข้อมูลที่ได้จากวิธีแบ่งช่วงข้อมูล

3.1.1 กรอบแนวคิดที่ 1: วิธีแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการหากฎความสัมพันธ์

กรอบแนวคิดที่ 1 จะเป็นการหากฎความสัมพันธ์ที่ได้หลังจากข้อมูลผ่านขั้นตอนวิธีแบ่งช่วงข้อมูล โดยจะนำข้อมูลที่เป็นตัวเลขต่อเนื่องไปทำการแบ่งช่วงข้อมูล แล้วนำข้อมูลที่แบ่งช่วงแล้วนั้นมาแทนที่ข้อมูลเดิมที่เป็นตัวเลขต่อเนื่อง จะได้ข้อมูลชุดใหม่ที่ถูกแทนที่ข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยช่วงข้อมูล และจะนำข้อมูลชุดใหม่ที่ได้จากวิธีแบ่งช่วงข้อมูลไปหากฎความสัมพันธ์ โดยมีกรอบแนวคิด ดังรูปที่ 3.1



รูปที่ 3.1 กรอบแนวคิดวิธีแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการหากฎความสัมพันธ์

จากกรอบแนวคิดวิธีแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการหากฎความสัมพันธ์จะประกอบไปด้วย 4 ขั้นตอน คือ

1) ดึงข้อมูลในส่วนที่เป็นตัวเลขต่อเนื่อง คือ จะดึงข้อมูลเฉพาะคอลัมน์ที่เป็นตัวเลขต่อเนื่องและคอลัมน์เป้าหมายออกจากข้อมูลชุดเดิม

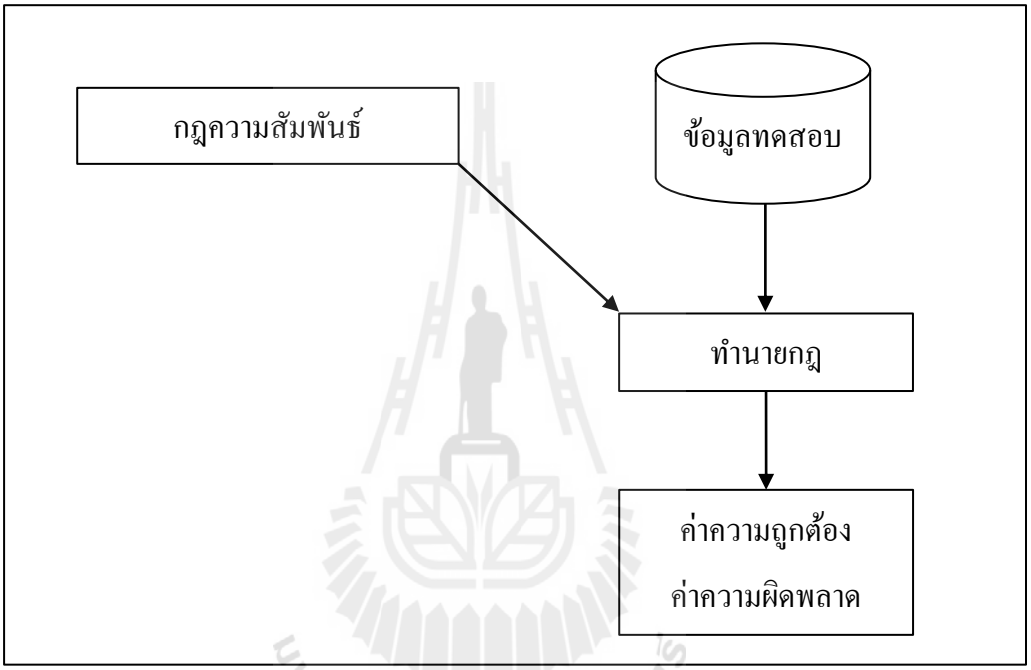
2) แบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่อง คือ จะนำคอลัมน์ที่เป็นตัวเลขต่อเนื่องและคอลัมน์เป้าหมายมาทำการแบ่งช่วงข้อมูลด้วยอัลกอริทึมที่ได้ออกแบบไว้

3) แทนที่ข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยช่วงข้อมูล คือ จะนำคอลัมน์ที่แบ่งช่วงข้อมูลเสร็จเรียบร้อยแล้ว ไปแทนที่คอลัมน์ที่เป็นตัวเลขต่อเนื่องจากข้อมูลชุดเดิม

4) หากฎความสัมพันธ์ คือ เมื่อข้อมูลที่เป็นตัวเลขต่อเนื่องผ่านกระบวนการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่อง ก็จะได้ข้อมูลชุดใหม่ที่พร้อมสำหรับนำไปหากฎความสัมพันธ์

3.1.2 กรอบแนวคิดที่ 2: การทดสอบประสิทธิภาพกฎความสัมพันธ์จากข้อมูลที่ได้จากวิธีแบ่งช่วงข้อมูล

กรอบแนวคิดที่ 2 จะเป็นการทดสอบประสิทธิภาพกฎความสัมพันธ์ของข้อมูลที่ได้จากวิธีแบ่งช่วงข้อมูล โดยจะนำชุดข้อมูลทดสอบมาใช้ในการทำนายในแต่ละกฎความสัมพันธ์ ซึ่งจะได้ออกมาเป็นค่าความถูกต้องและค่าความผิดพลาดของแต่ละกฎความสัมพันธ์ โดยมีกรอบแนวคิด ดังรูปที่ 3.2

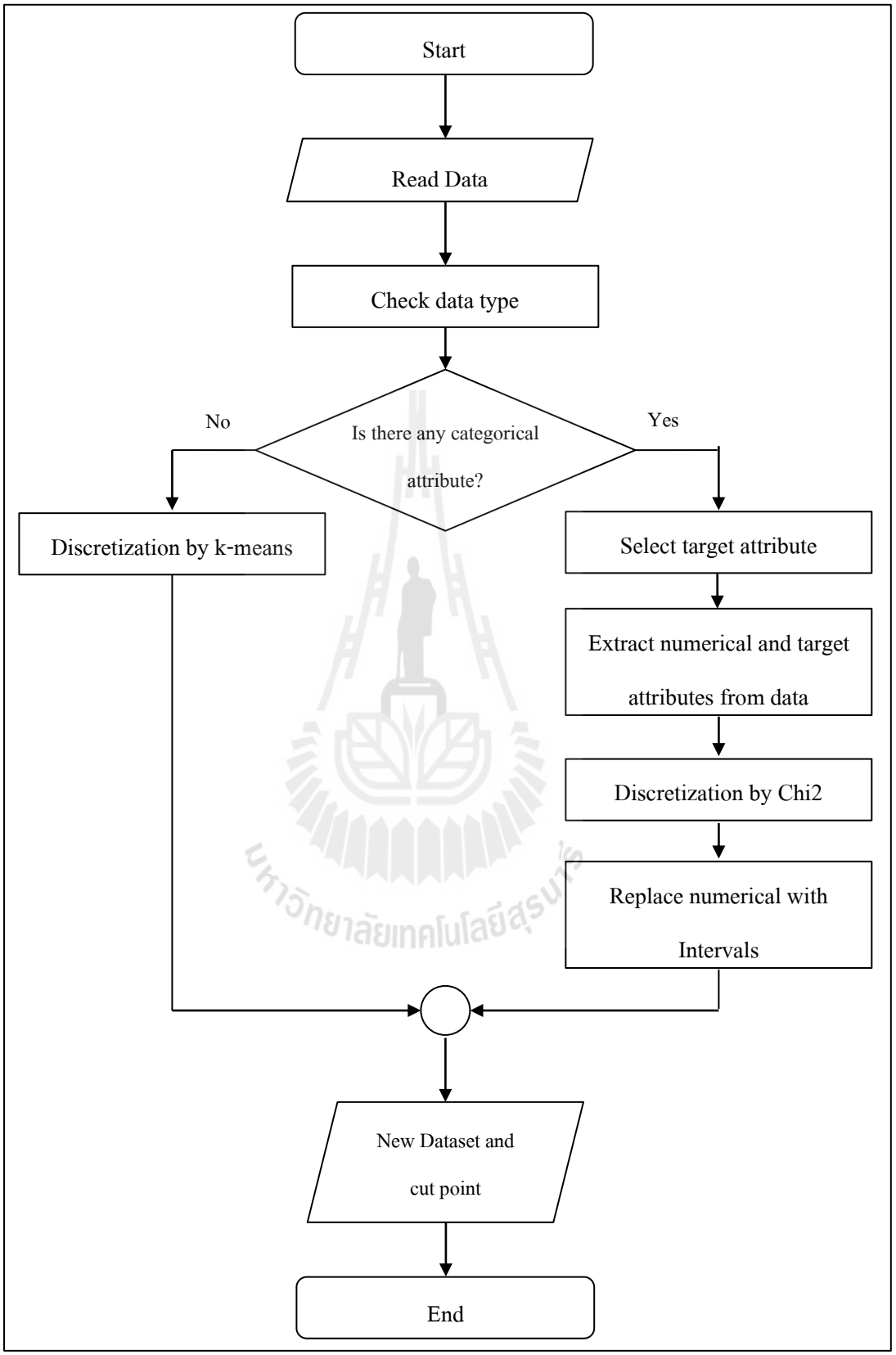


รูปที่ 3.2 กรอบแนวคิดการทดสอบประสิทธิภาพกฎความสัมพันธ์จากข้อมูลที่ผ่านวิธีแบ่งช่วงข้อมูล

3.2 การออกแบบอัลกอริทึม

3.2.1 ออกแบบอัลกอริทึมการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการหากฎความสัมพันธ์

การแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องนั้นเป็นส่วนของขั้นตอนก่อนการประมวลผล (Pre-processing) สำหรับงานทำเหมืองข้อมูล ซึ่งในงานวิจัยนี้ได้เสนอขั้นตอนวิธีแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องก่อนนำไปผ่านกระบวนการหากฎความสัมพันธ์ ซึ่งจะทำการนำข้อมูลที่เป็นตัวเลขต่อเนื่องมาผ่านกระบวนการแบ่งช่วงข้อมูลตามอัลกอริทึมที่ได้ออกแบบไว้ เมื่อได้ข้อมูลที่ทำการแบ่งช่วงเสร็จเรียบร้อยแล้วจะนำไปแทนที่ข้อมูลเดิมเพื่อนำไปเข้าสู่กระบวนการหากฎความสัมพันธ์ ขั้นตอนการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องแสดงได้ดังรูปที่ 3.3 และ 3.4



รูปที่ 3.3 ผังงานแสดงขั้นตอนการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่อง

Algorithm Chi2+Select target

```
//Input: Dataset D.
//Output: New discretized data, Cut point.

1)  categorical_type = check_type(D);
2)  if categorical_type is true {
3)          target = minimum_correlation(D);
4)          {numData, list} = extract_numeric(D, target);
5)          {discData, cutp}= chi_sq(numData);
6)          disc_data = replace_data(D, discData, list);
7)  } else{
8)          {disc_data, cutp} = kmeans(D);
9)  }
10) as_factor(disc_data);
11) return {disc_data, cutp};
```

รูปที่ 3.4 คำสั่งเทียมของอัลกอริทึมการแบ่งช่วงข้อมูล

จากรูปที่ 3.3 และ 3.4 เป็นอัลกอริทึมหลักในการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่อง การทำงานในบรรทัดแรกจะเป็นการตรวจสอบชุดข้อมูลว่าเป็นข้อมูลชนิดใด ซึ่งถ้าชนิดข้อมูลเป็นตัวเลขอย่างเดียวจะใช้อัลกอริทึม k-means ในการแบ่งช่วงข้อมูล และถ้าชนิดข้อมูลเป็นตัวเลขต่อเนื่อง และข้อมูลเชิงกลุ่มปะปนกัน จะใช้อัลกอริทึม Chi2 ในการแบ่งช่วงข้อมูล โดยอัลกอริทึมในส่วนการทำงานขั้นตอนต่าง ๆ ของการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องจะประกอบไปด้วย 6 ขั้นตอน คือ

ขั้นตอนที่ 1 ตรวจสอบชนิดข้อมูล คือ ชุดข้อมูลที่เป็นตัวเลขต่อเนื่องที่จะนำไปแบ่งช่วงข้อมูลนั้นในงานวิจัยนี้จะแบ่งได้ออกเป็น 2 ลักษณะ คือ ข้อมูลทุกคอลัมน์เป็นตัวเลขต่อเนื่อง และข้อมูลบางคอลัมน์เป็นตัวเลขต่อเนื่อง ซึ่งจากข้อมูล 2 ลักษณะนี้จะใช้เทคนิคการแบ่งช่วงข้อมูลที่แตกต่างกัน โดยข้อมูลทุกคอลัมน์เป็นตัวเลขต่อเนื่องจะใช้อัลกอริทึม k-means ในการแบ่งช่วงข้อมูล และข้อมูลบางคอลัมน์เป็นตัวเลขต่อเนื่องจะใช้อัลกอริทึม Chi2 ในการแบ่งช่วงข้อมูล ขั้นตอนการตรวจสอบชนิดข้อมูลแสดงได้ดังรูปที่ 3.5

```
Procedure check_type (data)
    1) Set type to F;
    2) For each attribute i {
    3)         If data[i] is categorical {
    4)                 Set type to True;
    5)                 Break;
    6)         }
    7) }
    8) return type;
```

รูปที่ 3.5 คำสั่งเทียมขั้นตอนการตรวจสอบชนิดข้อมูล

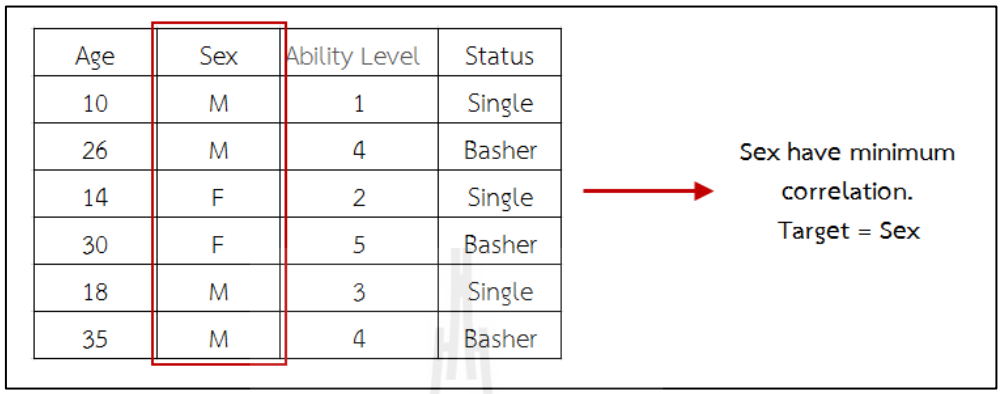
ขั้นตอนที่ 2 การหาคอลัมน์เป้าหมายที่มีค่าสหสัมพันธ์น้อยที่สุด เนื่องจากอัลกอริทึม Chi2 เป็นอัลกอริทึมแบบมีผู้ฝึกสอน (Supervised) จึงต้องมีการกำหนดคอลัมน์เป้าหมายให้กับอัลกอริทึมด้วย แต่เนื่องจากในข้อมูล 1 ชุด มีคอลัมน์ที่สามารถเป็นคอลัมน์เป้าหมายได้มากกว่า 1 คอลัมน์ ดังนั้นผู้วิจัยจึงตั้งสมมติฐานไว้ว่าคอลัมน์ที่มีค่าสหสัมพันธ์ (Correlation) น้อยที่สุด เมื่อนำไปแบ่งช่วงข้อมูลสำหรับการหากฎความสัมพันธ์แล้วจะให้ประสิทธิภาพที่ดีที่สุด ขั้นตอนการตรวจสอบชนิดข้อมูลแสดงได้ดังรูปที่ 3.6

```
Procedure minimum_correlation (data)
1)  Set minCor, point to 0 and count to 1;
2)  For each attribute i {
3)          If data is numeric {
4)                  list[count] = i;
5)                  count = count+1;
6)          }
7)  }
8)  For each attribute i {
9)          If data[i] is categorical {
10)                 cor = correlation(data[,list], data[,i]);
11)                 meanCor = abs(sum(cor)/length(list));
12)                 If point == 0 {
13)                         minCor = meanCor;
14)                         point = i
15)                 } else {
16)                         If min > meanCor {
17)                                 minCor = meanCor;
18)                                 point = i
19)                         }
20)                 }
21)         }
22) }
23) return {minCor, point};
```

จากรูปที่ 3.6 คำสั่งเทียมขั้นตอนการหาคอลัมน์เป้าหมายที่มีค่าสหสัมพันธ์น้อยที่สุด

การทำงานในบรรทัดแรกจะกำหนดค่าสหสัมพันธ์ของคอลัมน์เป้าหมายและตำแหน่งของคอลัมน์เป้าหมายในชุดข้อมูลเป็นศูนย์ หลังจากนั้นจะหาค่าสหสัมพันธ์ในแต่ละคอลัมน์เป้าหมายกับทุกคอลัมน์ที่เป็นตัวเลขเพื่อมาหาตำแหน่งคอลัมน์เป้าหมายที่มีค่าเฉลี่ยสหสัมพันธ์ที่น้อยที่สุด จากรูปที่ 3.7 แสดงการหาคอลัมน์เป้าหมายที่มีค่าสหสัมพันธ์น้อยที่สุด โดยค่าสหสัมพันธ์ที่นำมา

วิเคราะห์ได้จากคอลัมน์ Age ซึ่งคอลัมน์ที่เป็นตัวเลขเนื่อง กับคอลัมน์อื่น ได้แก่ {Sex, Age} = 0.054, {Ability Level, Age} = 0.27, {Status, Age} = 0.29 จะเห็นได้ว่าค่าสหสัมพันธ์ระหว่างคอลัมน์ Sex และ Age น้อยที่สุด ดังนั้นจึงเลือกคอลัมน์ Sex เป็นคอลัมน์เป้าหมาย



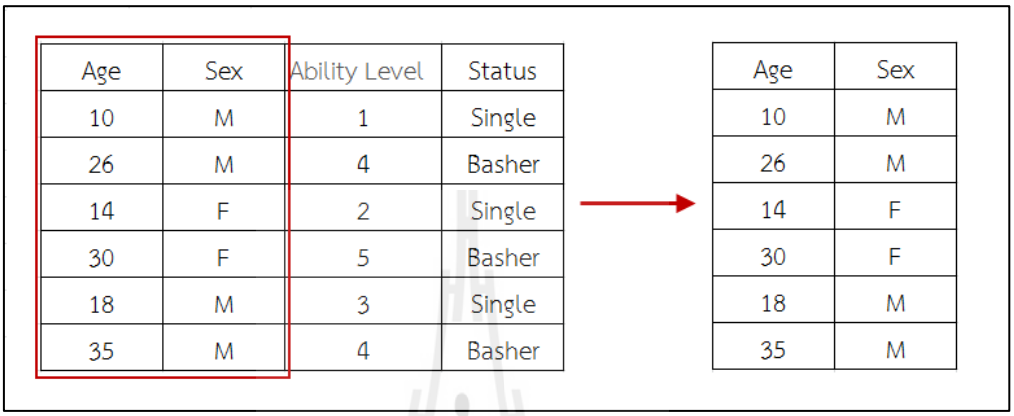
รูปที่ 3.7 ตัวอย่างขั้นตอนการหาคอลัมน์เป้าหมายที่มีค่าสหสัมพันธ์น้อยที่สุด

ขั้นตอนที่ 3 การดึงคอลัมน์ที่เป็นตัวเลขต่อเนื่องและคอลัมน์เป้าหมาย วิธีแบ่งช่วงข้อมูลด้วยอัลกอริทึม Chi2 ข้อมูลที่จะนำมาใช้ทุกคอลัมน์จำเป็นต้องตัวเลขต่อเนื่องยกเว้นคอลัมน์เป้าหมาย ดังนั้นจึงต้องมีการเตรียมข้อมูลก่อน ด้วยการดึงคอลัมน์ที่เป็นตัวเลขต่อเนื่องและคอลัมน์เป้าหมายออกจากชุดข้อมูลเดิมก่อน แสดงรายละเอียดขั้นตอนนี้ดังรูปที่ 3.8

```
Procedure extract_numeric (data, class)
1)  Set count to 1;
2)  For each attribute i {
3)          If data[i] is numeric {
4)                  list[count] = i;
5)                  count++;
6)          }
7)  }
8)  list[count] = class
9)  return {data[list], list};
```

รูปที่ 3.8 คำสั่งเทียมขั้นตอนการดึงคอลัมน์ที่เป็นตัวเลขต่อเนื่องและคอลัมน์เป้าหมาย

จากรูปที่ 3.8 เป็นขั้นตอนวิธีการดึงคอลัมน์ที่เป็นตัวเลขต่อเนื่องและคอลัมน์เป้าหมาย การทำงานคือจะตรวจสอบชนิดข้อมูลในแต่ละคอลัมน์ ถ้าชนิดข้อมูลเป็นตัวเลขก็จะเก็บตำแหน่งของคอลัมน์นั้น ๆ ไว้เป็นชุดของตำแหน่งคอลัมน์ที่มีชนิดข้อมูลเป็นตัวเลข และคอลัมน์เป้าหมาย จากรูปที่ 3.9 แสดงการดึงคอลัมน์ที่เป็นตัวเลขต่อเนื่องและคอลัมน์เป้าหมาย



Age	Sex	Ability Level	Status
10	M	1	Single
26	M	4	Basher
14	F	2	Single
30	F	5	Basher
18	M	3	Single
35	M	4	Basher

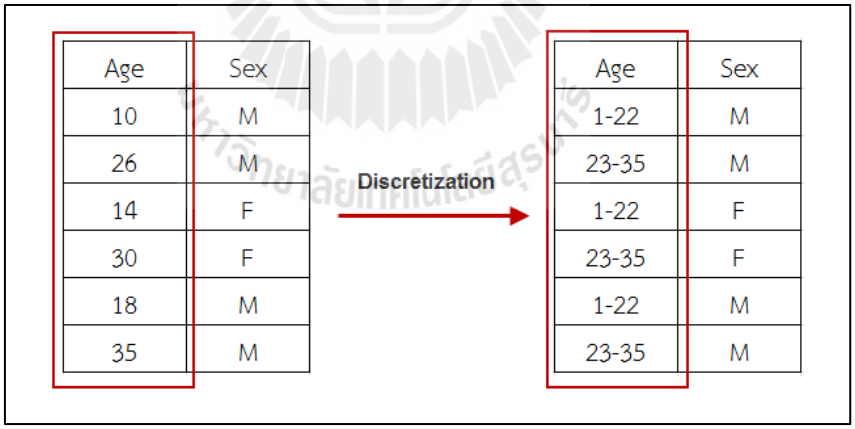
Age	Sex
10	M
26	M
14	F
30	F
18	M
35	M

รูปที่ 3.9 ตัวอย่างการดึงคอลัมน์ที่เป็นตัวเลขต่อเนื่องและคอลัมน์เป้าหมาย

ขั้นตอนที่ 4 การแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยอัลกอริทึม Chi2 เป็นการใช้เทคนิคการวิเคราะห์ข้อมูล Chi2 ทางด้านสถิติมาใช้ในการแบ่งช่วงข้อมูลเพื่อทดสอบความเป็นอิสระต่อกันในแต่ละช่วงข้อมูล ซึ่งได้อธิบายไว้ในบทที่ 2 โดยรูปที่ 3.10 และ 3.11 แสดงคำสั่งเทียมและตัวอย่างการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยอัลกอริทึม Chi2

```
Phase 1:
  set sigLevel = .5;
  do while (InConsistency(data) < δ) {
    for each numeric attribute {
      Sort(attribute, data);
      chi-sq-initialization(attribute, data);
      do {
        chi-sq-calculation(attribute,data)
      } while (Merge(data))
    }
    sigLevel0 = sigLevel;
    sigLevel = decreSigLevel(sigLevel);
  }
Phase 2:
  set all sigLvl[i] = sigLevel0 for attribute i;
  do until no-attribute-can-be-merged {
    for each attribute i that can be merged {
      Sort(attribute, data);
      chi-sq-initialization(attribute, data);
      do {
        chi-sq-calculation(attribute,data)
      } while (Merge(data))
      if (InConsistency(data) < δ)
        sigLvl[i] = decreSigLevel(sigLvl[i]);
      else
        attribute i cannot be merged;
    }
  }
```

รูปที่ 3.10 คำสั่งเทียมของอัลกอริทึม Chi2 (Liu and Setiono, 1995)



รูปที่ 3.11 ตัวอย่างการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยอัลกอริทึม Chi2

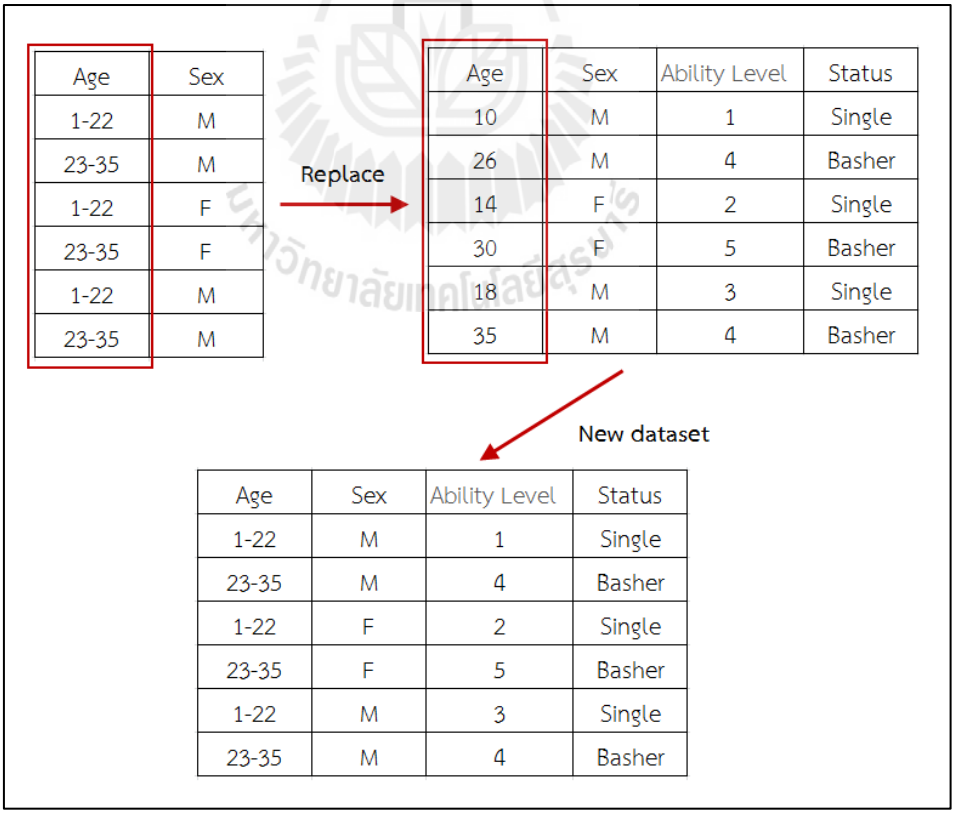
ขั้นตอนที่ 5 การแทนที่ข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยช่วงข้อมูล คือ เมื่อได้ข้อมูลที่ทำการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยอัลกอริทึม Chi2 เสร็จเรียบร้อยแล้ว ก็จะนำไปแทนที่คอลัมน์ที่เป็นตัวเลขต่อเนื่องของข้อมูลชุดเดิมด้วยช่วงข้อมูล ซึ่งจะได้ข้อมูลชุดใหม่ที่พร้อมนำไปเข้าสู่กระบวนการหากฎความสัมพันธ์ แสดงได้ดังรูปที่ 3.12

Procedure replace_data (data, discData, list)

1) For each numeric attribute i {
2) data[list[i]] = discData[i];
3) }
4) return data;

รูปที่ 3.12 คำสั่งเทียมขั้นตอนการแทนที่ข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยช่วงข้อมูล

จากรูปที่ 3.12 เป็นขั้นตอนวิธีการแทนที่ข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยช่วงข้อมูล การทำงานคือจากตำแหน่งของคอลัมน์ที่เป็นตัวเลขต่อเนื่องที่ได้จากขั้นตอนการดึงคอลัมน์ที่เป็นตัวเลขต่อเนื่องและคอลัมน์เป้าหมาย ก็จะนำช่วงข้อมูลไปแทนที่ในคอลัมน์ที่เป็นตัวเลขต่อเนื่องตำแหน่งเดิม จากรูปที่ 3.13 แสดงการแทนที่ข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยช่วงข้อมูล



รูปที่ 3.13 ตัวอย่างการแทนที่ข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยช่วงข้อมูล

ขั้นตอนที่ 6 การแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยอัลกอริทึม k-means เมื่อข้อมูลมีชนิดข้อมูลในทุกคอลัมน์เป็นแบบตัวเลขต่อเนื่องทำให้ไม่สามารถแบ่งช่วงข้อมูลด้วยอัลกอริทึม Chi2 ได้ ดังนั้นจึงใช้อัลกอริทึม k-means ในการแบ่งช่วงข้อมูล เนื่องจากอัลกอริทึม k-means เป็นอัลกอริทึมที่ทำงานกับกรณีข้อมูลเป็นตัวเลขทั้งหมดได้ดี แสดงได้ดังรูปที่ 3.14

```
Procedure K_means (data)
1) Set km and disCutp to list;
2) For each attribute i {
3)         km[i] = kmeans(data[i]);
4) }
5) For each attribute j {
6)         data[j] = km[j]$cluster;
7)         disCutp[j] = km[j]$centers;
8) }
9) Return {data, disCutp};
```

รูปที่ 3.14 คำสั่งเทียมขั้นตอนการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยอัลกอริทึม k-means

จากรูปที่ 3.14 เป็นขั้นตอนวิธีการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยอัลกอริทึม k-means การทำงานคือจะใช้อัลกอริทึม k-means ในการจัดกลุ่มข้อมูลในแต่ละคอลัมน์ หลังจากนั้นจะแทนที่ชุดข้อมูลเดิมด้วยกลุ่มข้อมูลที่ได้จากการจัดกลุ่มด้วยอัลกอริทึม k-means จากรูปที่ 3.15 แสดงตัวอย่างการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยอัลกอริทึม k-means

Age	Price	Weight	Height
10	150	60	160
26	300	70	170
14	420	65	166
30	100	55	150
18	130	80	175
35	148	40	149

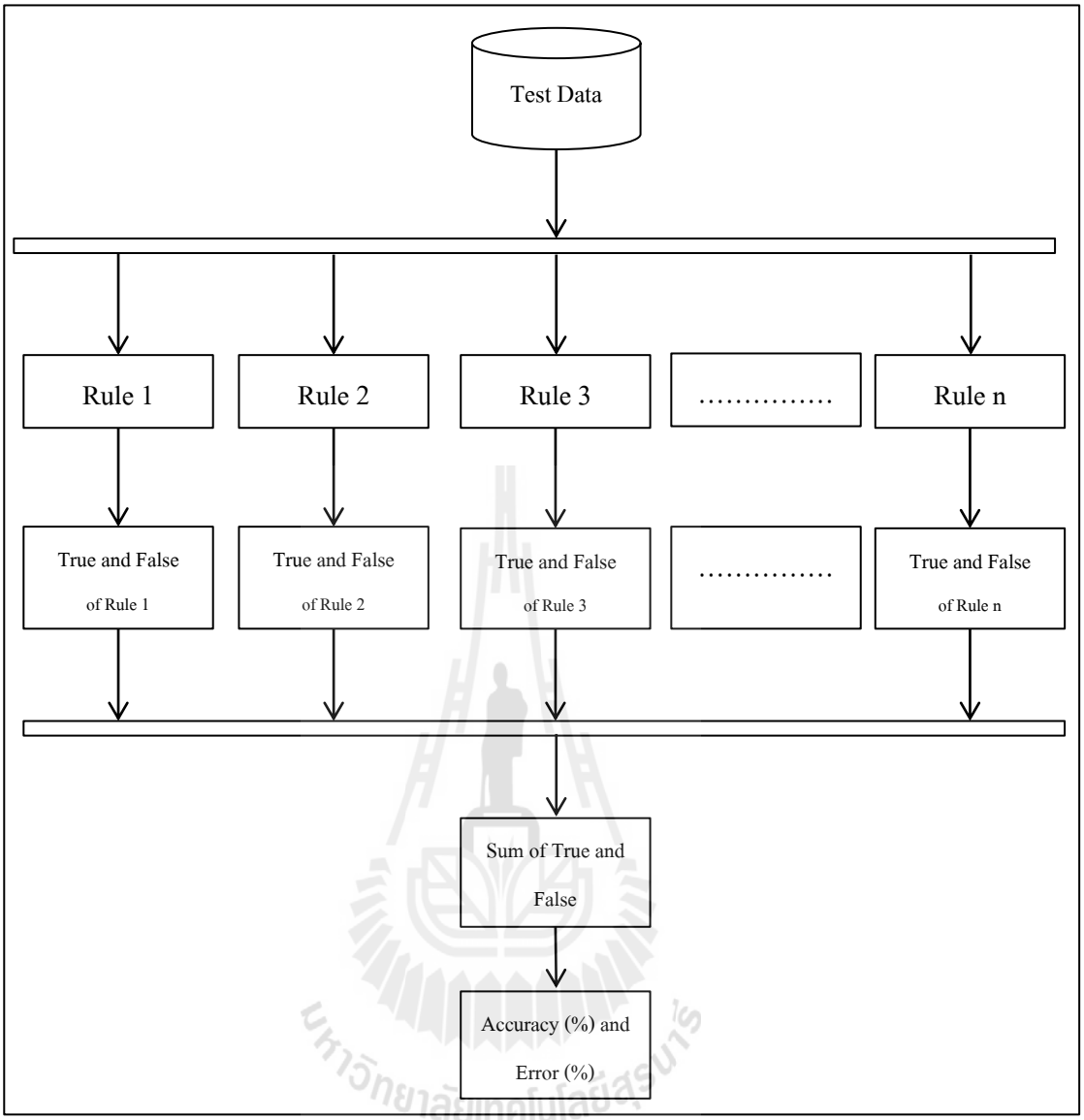
K-means →

Age	Price	Weight	Height
1-15	1-150	56-70	1-160
16-20	156-250	56-70	161-175
1-15	350-400	56-70	1-160
20-35	1-150	1-55	1-160
16-20	1-150	71-80	161-175
20-35	1-150	1-55	1-160

รูปที่ 3.15 ตัวอย่างการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยอัลกอริทึม k-means

3.2.2 ออกแบบการทดสอบประสิทธิภาพวิธีแบ่งช่วงข้อมูลสำหรับการหากฎความสัมพันธ์

การทดสอบประสิทธิภาพกฎความสัมพันธ์ส่วนใหญ่จะวัดประสิทธิภาพด้วยมาตรวัดมาตรฐาน เช่น support, confidence เป็นต้น การเปรียบเทียบกฎความสัมพันธ์ด้วยมาตรวัดมาตรฐานเหล่านี้ มีข้อจำกัดที่กฎความสัมพันธ์จะต้องเป็นกฎชุดเดียวกัน จึงจะเปรียบเทียบได้ว่าวิธีการเตรียมข้อมูลแบบใดให้กฎที่มีค่า support และค่า confidence ที่สูงกว่า การนำมาตรวัดเหล่านี้ ไปวัดประสิทธิภาพของอัลกอริทึมการแบ่งช่วงข้อมูลต่าง ๆ นั้นเป็นสิ่งที่กระทำได้ยาก เนื่องจากในการหากฎความสัมพันธ์ในแต่ละอัลกอริทึมที่ได้ออกมานั้น จะได้กฎความสัมพันธ์ที่ไม่เหมือนเดิม ดังนั้นงานวิจัยนี้จึงได้เสนอการทดสอบประสิทธิภาพการแบ่งช่วงข้อมูลสำหรับการหากฎความสัมพันธ์ด้วยมาตรวัด accuracy โดยวิธีพิจารณา accuracy จะแตกต่างจากกรณีการวัดความถูกต้องของ classification model เล็กน้อยดังแสดงในรูปที่ 3.16



รูปที่ 3.16 แนวคิดของวิธีการทดสอบประสิทธิภาพของกฎความสัมพันธ์

จากรูปที่ 3.16 แสดงวิธีการทดสอบประสิทธิภาพวิธีของกฎความสัมพันธ์ โดยตั้งแต่ขั้นตอนการแบ่งช่วงข้อมูลจะทำการแบ่งข้อมูลออกเป็น training set และ test set ในส่วนของการทดสอบประสิทธิภาพจะใช้ข้อมูลใน test set ทุกแถวมาใช้ประเมินผลการทำนายด้วยแต่ละกฎความสัมพันธ์ที่ได้จากวิธีแบ่งช่วงข้อมูลสำหรับการหากฎความสัมพันธ์ และจะได้ออกมาเป็นค่าความถูกต้องและค่าความผิดพลาด ดังแสดงตัวอย่างในรูปที่ 3.17

Training set

Age	Sex	Ability Level	Status
1-22	M	1	Single
23-35	M	4	Basher
1-22	F	2	Single
23-35	F	5	Basher
1-22	M	3	Single
23-35	M	4	Basher

หากฎความสัมพันธ์

Rule
Status=Basher ==> Age=[22 - 35]
Age=[22 - 35] ==> Status=Basher
Ability Level=5==> Age= [22 - 35]
Ability Level=2==> Sex=F
Ability Level=5==> Sex=F
Ability Level=5==> Status=Basher
Sex=F Ability Level=5==> Age= [22 - 35]
Age=[22 - 35], Ability Level=5==> Sex=F
Age= [22 - 35], Sex=F ==> Ability Level=5
Ability Level=5==> Age= [22 - 35], Sex=F

ทำนายด้วย กฎความสัมพันธ์

Age	Sex	Ability Level	Status
23-35	F	5	Basher
23-35	M	5	Single
1-22	M	1	Single
1-22	F	4	Basher
23-35	M	2	Single
1-22	F	3	Single

Test set

วัดประสิทธิภาพการทำนาย

	Rule	Test Data						Prediction	
		Record1	Record2	Record3	Record4	Record5	Record6	True	False
1	Status=Basher ==> Age=[22 - 35]	1	null	null	0	null	null	1	1
2	Age=[22 - 35] ==> Status=Basher	1	0	null	null	0	null	1	2
3	Ability Level=5==> Age= [22 - 35]	1	1	null	null	null	null	2	0
4	Ability Level=2==> Sex=F	null	null	null	null	0	null	0	1
5	Ability Level=5==> Sex=F	0	0	null	null	null	null	0	2
6	Ability Level=5==> Status=Basher	1	0	null	null	null	null	1	1
7	Sex=F Ability Level=5==> Age= [22 - 35]	1	null	null	null	null	null	1	0
8	Age=[22 - 35], Ability Level=5==> Sex=F	1	0	null	null	null	null	1	1
9	Age= [22 - 35], Sex=F ==> Ability Level=5	1	null	null	null	null	null	1	0
10	Ability Level=5==> Age= [22 - 35], Sex=F	1	0	null	null	null	null	1	1
							Sum	9	9

Accuracy (%) = (True*100) / (True+False)
= (9*100) / (9+9)
= 900 / 18
= 50 %

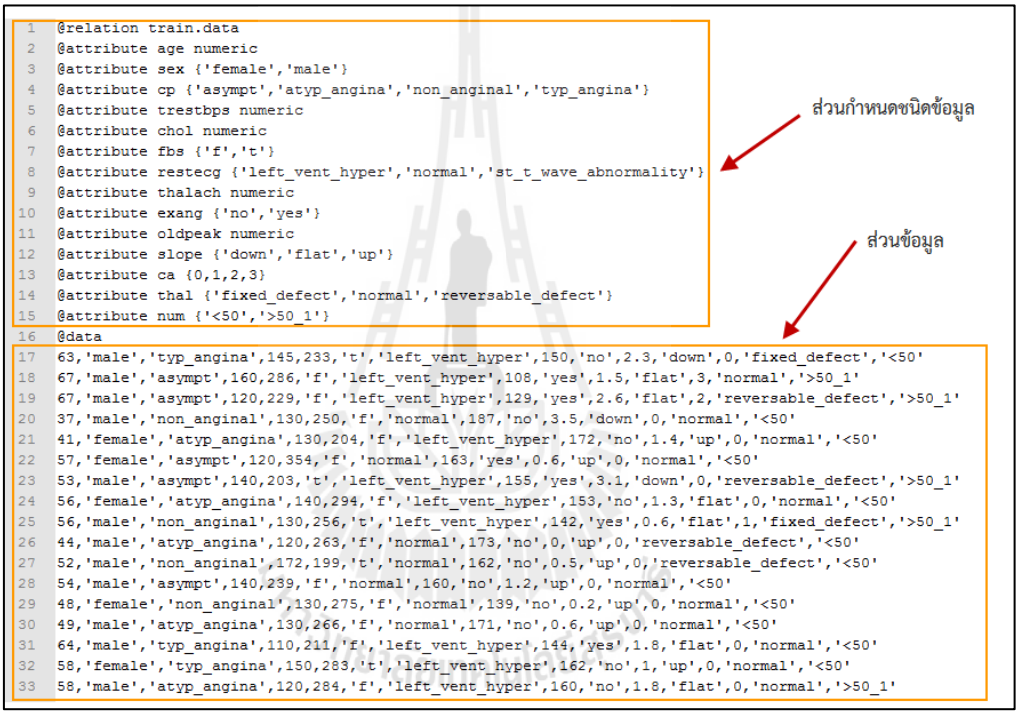
รูปที่ 3.17 ตัวอย่างการทดสอบประสิทธิภาพการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการหากฎความสัมพันธ์

3.3 การใช้งานโปรแกรม

เนื้อหาในส่วนนี้จะอธิบายการใช้งานของโปรแกรมการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับหากฎความสัมพันธ์ โดยจะแบ่งการทำงานของโปรแกรมออกเป็นขั้นตอนดังนี้

3.3.1 การเตรียมข้อมูล

การใช้งานโปรแกรมการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับหากฎความสัมพันธ์ ข้อมูลที่นำมาใช้ต้องมีการกำหนดชนิดข้อมูลในแต่ละคอลัมน์ ดังนั้นการวิจัยนี้จึงกำหนดให้ข้อมูลต้องอยู่ในรูปแบบ .arff



รูปที่ 3.18 รูปแบบของข้อมูลที่จะนำมาแบ่งช่วงข้อมูล

จารูปที่ 3.18 ข้อมูลจะประกอบไปด้วย 2 ส่วน คือ

1) ส่วนกำหนดชนิดข้อมูล เป็นส่วนคำอธิบายชนิดของข้อมูลในแต่ละคอลัมน์
2) ส่วนข้อมูล เป็นรายละเอียดข้อมูลในแต่ละเรคคอร์ด

3.3.2 การแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่อง

การเรียกใช้โปรแกรมในส่วนของการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องนั้น โปรแกรมจะตรวจสอบชนิดข้อมูลว่าถ้าข้อมูลมีเฉพาะตัวเลขจะใช้อัลกอริทึม k-means ในการแบ่ง

ช่วงข้อมูล แต่ถ้าชนิดข้อมูลมีทั้งตัวเลข และข้อมูลเชิงกลุ่มปะปนกัน จะใช้อัลกอริทึม Chi2 ในการแบ่งช่วงข้อมูล โดยมีรูปแบบและตัวอย่างการเรียกใช้โปรแกรมในส่วนของการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่อง ดังรูปที่ 3.19 และ 3.20 ตามลำดับ

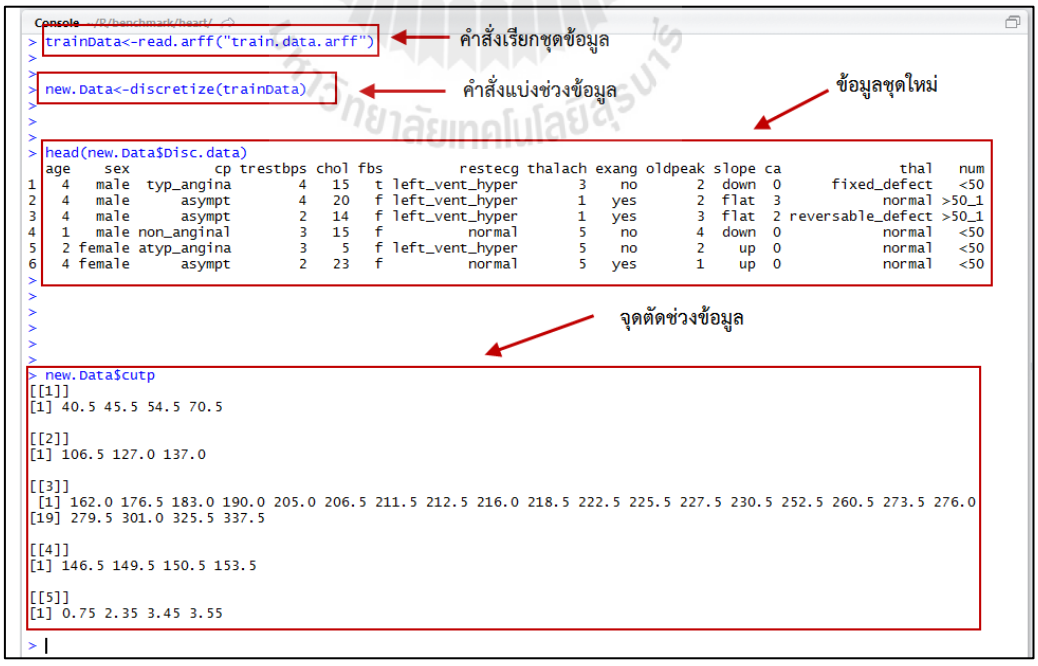
```
Data.disc, cutp <- discretize (data)
```

รูปที่ 3.19 รูปแบบการเรียกใช้โปรแกรมในส่วนของการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่อง

จากรูปที่ 3.19 คือ รูปแบบการเรียกใช้โปรแกรมในส่วนของการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่อง โดยมีตัวแปรดังนี้

1) data คือ ชุดข้อมูลที่เป็นตัวเลขต่อเนื่อง
2) Data.disc คือ ข้อมูลชุดใหม่ที่ถูกแทนที่ตัวเลขต่อเนื่องด้วยช่วงข้อมูล
3) cutp คือ จุดตัดของช่วงข้อมูลที่ใช้สำหรับแบ่งข้อมูลในแต่ละช่วง

จากรูปที่ 3.20 แสดงตัวอย่างและผลลัพธ์การเรียกใช้โปรแกรมแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่อง โดยคอลัมน์ที่เป็นตัวเลขจะถูกแทนที่ด้วยช่วงข้อมูล ตัวอย่างเช่น คอลัมน์ age ช่วงข้อมูลที่ได้ คือ 1 = <40.5, 2 = <45.5, 3 = <54.5, 4 = <70.5, 5 = >=70.5




```
> trainData<-read.arff("train.data.arff")
> new.Data<-discretize(trainData)
> head(new.Data$Disc.data)
  age    sex          cp trestbps chol fbs          restecg thalach exang oldpeak slope ca             thal   num
1   4   male  typ_angina        4   15   t left_vent_hyper       3    no       2  down  0     fixed_defect   <50
2   4   male      asympt        4   20   f left_vent_hyper       1   yes       2  flat  3           normal >50_1
3   4   male      asympt        2   14   f left_vent_hyper       1   yes       3  flat  2 reversable_defect >50_1
4   1   male non_anginal        3   15   f           normal       5    no       4  down  0           normal   <50
5   2 female atyp_angina        3    5   f left_vent_hyper       5    no       2    up  0           normal   <50
6   4 female      asympt        2   23   f           normal       5   yes       1    up  0           normal   <50
> new.Data$cutp
[[1]]
[1] 40.5 45.5 54.5 70.5

[[2]]
[1] 106.5 127.0 137.0

[[3]]
 [1] 162.0 176.5 183.0 190.0 205.0 206.5 211.5 212.5 216.0 218.5 222.5 225.5 227.5 230.5 252.5 260.5 273.5 276.0
[19] 279.5 301.0 325.5 337.5

[[4]]
[1] 146.5 149.5 150.5 153.5

[[5]]
[1] 0.75 2.35 3.45 3.55
```

รูปที่ 3.20 ตัวอย่างและผลลัพธ์การเรียกใช้โปรแกรมแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่อง

3.3.3 การหากฎความสัมพันธ์

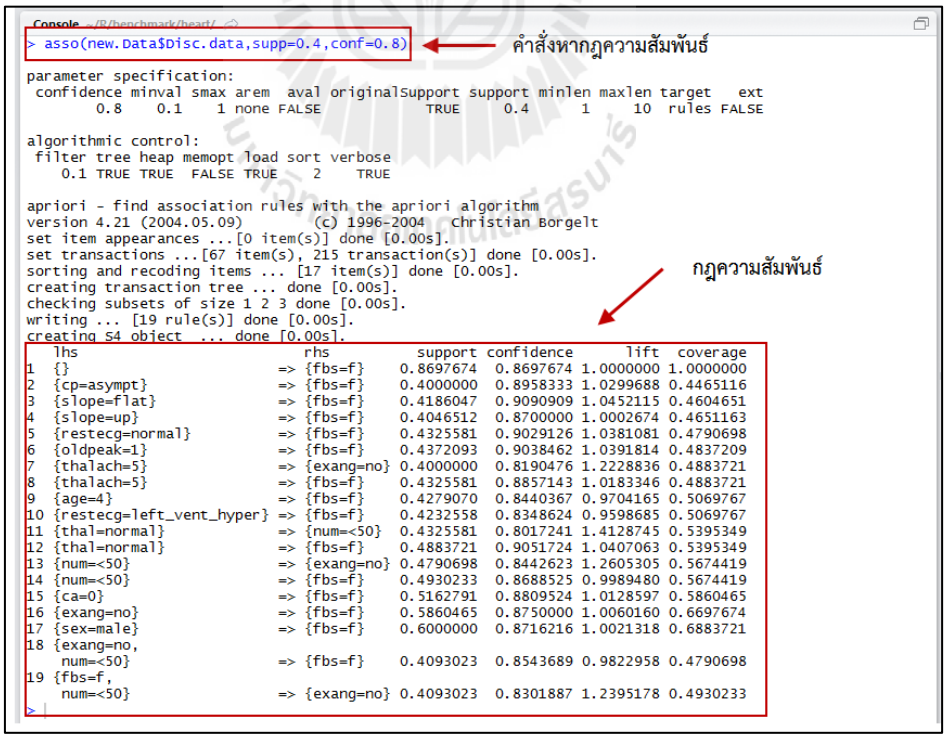
การเรียกใช้โปรแกรมในส่วนของการหากฎความสัมพันธ์ มีรูปแบบและตัวอย่างการเรียกใช้โปรแกรมในส่วนของการหากฎความสัมพันธ์ ดังรูปที่ 3.21 และ 3.22 ตามลำดับ

```
rules <- asso (data, supp, conf)
```

รูปที่ 3.21 รูปแบบการเรียกใช้โปรแกรมในส่วนของการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่อง

จากรูปที่ 3.21 แสดงรูปแบบการเรียกใช้โปรแกรมในส่วนของการหากฎความสัมพันธ์ โดยมีตัวแปรดังนี้

1) data คือ ชุดข้อมูลใหม่ที่แทนที่ข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยช่วงข้อมูล
2) supp คือ ค่า Minimum support สำหรับการหากฎความสัมพันธ์
3) conf คือ ค่า Minimum confidence สำหรับการหากฎความสัมพันธ์
4) rules คือ กฎความสัมพันธ์ที่ได้จากการเรียกใช้ฟังก์ชัน asso




```
> asso(new.Data$Disc.data,supp=0.4,conf=0.8)

parameter specification:
 confidence minval smax arem  aval originalSupport support minlen maxlen target   ext
        0.8    0.1    1 none FALSE            TRUE     0.4      1     10  rules FALSE

algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

apriori - find association rules with the apriori algorithm
version 4.21 (2004.05.09)        (c) 1996-2004   Christian Borgelt
set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[67 item(s), 215 transaction(s)] done [0.00s].
sorting and recoding items ... [17 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 done [0.00s].
writing ... [19 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
   lhs                          rhs          support confidence      lift  coverage
1  {}                        => {fbs=f}     0.8697674  0.8697674 1.0000000 1.0000000
2  {cp=asympt}               => {fbs=f}     0.4000000  0.8958333 1.0299688 0.4465116
3  {slope=flat}              => {fbs=f}     0.4186047  0.9090909 1.0452115 0.4604651
4  {slope=up}                => {fbs=f}     0.4046512  0.8700000 1.0002674 0.4651163
5  {restecg=normal}          => {fbs=f}     0.4325581  0.9029126 1.0381081 0.4790698
6  {oldpeak=1}               => {fbs=f}     0.4372093  0.9038462 1.0391814 0.4837209
7  {thalach=5}               => {exang=no}  0.4000000  0.8190476 1.2228836 0.4883721
8  {thalach=5}               => {fbs=f}     0.4325581  0.8857143 1.0183346 0.4883721
9  {age=4}                   => {fbs=f}     0.4279070  0.8440367 0.9704165 0.5069767
10 {restecg=left_vent_hyper} => {fbs=f}     0.4232558  0.8348624 0.9598685 0.5069767
11 {thal=normal}             => {num=<50}   0.4325581  0.8017241 1.4128745 0.5395349
12 {thal=normal}             => {fbs=f}     0.4883721  0.9051724 1.0407063 0.5395349
13 {num=<50}                 => {exang=no}  0.4790698  0.8442623 1.2605305 0.5674419
14 {num=<50}                 => {fbs=f}     0.4930233  0.8688525 0.9989480 0.5674419
15 {ca=0}                    => {fbs=f}     0.5162791  0.8809524 1.0128597 0.5860465
16 {exang=no}                => {fbs=f}     0.5860465  0.8750000 1.0060160 0.6697674
17 {sex=male}                => {fbs=f}     0.6000000  0.8716216 1.0021318 0.6883721
18 {exang=no,
    num=<50}                 => {fbs=f}     0.4093023  0.8543689 0.9822958 0.4790698
19 {fbs=f,
    num=<50}                 => {exang=no}  0.4093023  0.8301887 1.2395178 0.4930233
>
```

รูปที่ 3.22 ตัวอย่างและผลลัพธ์การเรียกใช้โปรแกรมในส่วนของการหากฎความสัมพันธ์

3.4 เครื่องมือที่ใช้ในการวิจัย

เครื่องมือที่ใช้ในงานวิจัยนี้ ประกอบด้วยฮาร์ดแวร์และซอฟต์แวร์ ดังนี้

1. เครื่องคอมพิวเตอร์สำหรับพัฒนาการแบ่งช่วงข้อมูลสำหรับการหากฎความสัมพันธ์ โดยมีรายละเอียดดังนี้

- หน่วยประมวลผลกลาง : Intel® Core 2 Duo 2.0 GHz
- หน่วยความจำสำรอง : 320 GB
- หน่วยความจำหลัก : 2 GB
- อุปกรณ์เสริมอื่น ๆ เช่น เมาส์ แป้นพิมพ์ เป็นต้น

2. ระบบปฏิบัติการและโปรแกรมประยุกต์สำหรับการพัฒนาการแบ่งช่วงข้อมูลสำหรับการหากฎความสัมพันธ์ ประกอบไปด้วย

- ระบบปฏิบัติการ : Windows 7 Ultimate 32-bit Operating System
- เครื่องมือในการพัฒนาโปรแกรม : R Studio 0.96.331

บทที่ 4
การทดสอบและอภิปรายผล

การทดสอบประสิทธิภาพของระบบนั้น จะทดสอบประสิทธิภาพวิธีแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการหากฎความสัมพันธ์ด้วยอัลกอริทึม Chi2 โดยการเปรียบเทียบประสิทธิภาพวิธีแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องนั้น จะเปรียบเทียบกับอีกสองอัลกอริทึม คือ อัลกอริทึม CAIM (Class-Attribute Interdependence Maximization) และอัลกอริทึม k-means ซึ่งจะเปรียบเทียบจากค่าความถูกต้อง (Accuracy) ที่ได้จากการทำนายในแต่ละกฎความสัมพันธ์ และมาตรวัดมาตรฐานอีก 4 มาตรวัด คือ Support, Coverage, Confidence และ Lift ซึ่งเป็นมาตรวัดที่แสดงพร้อมกับกฎความสัมพันธ์

4.1 ข้อมูลที่ใช้ในการทดสอบ

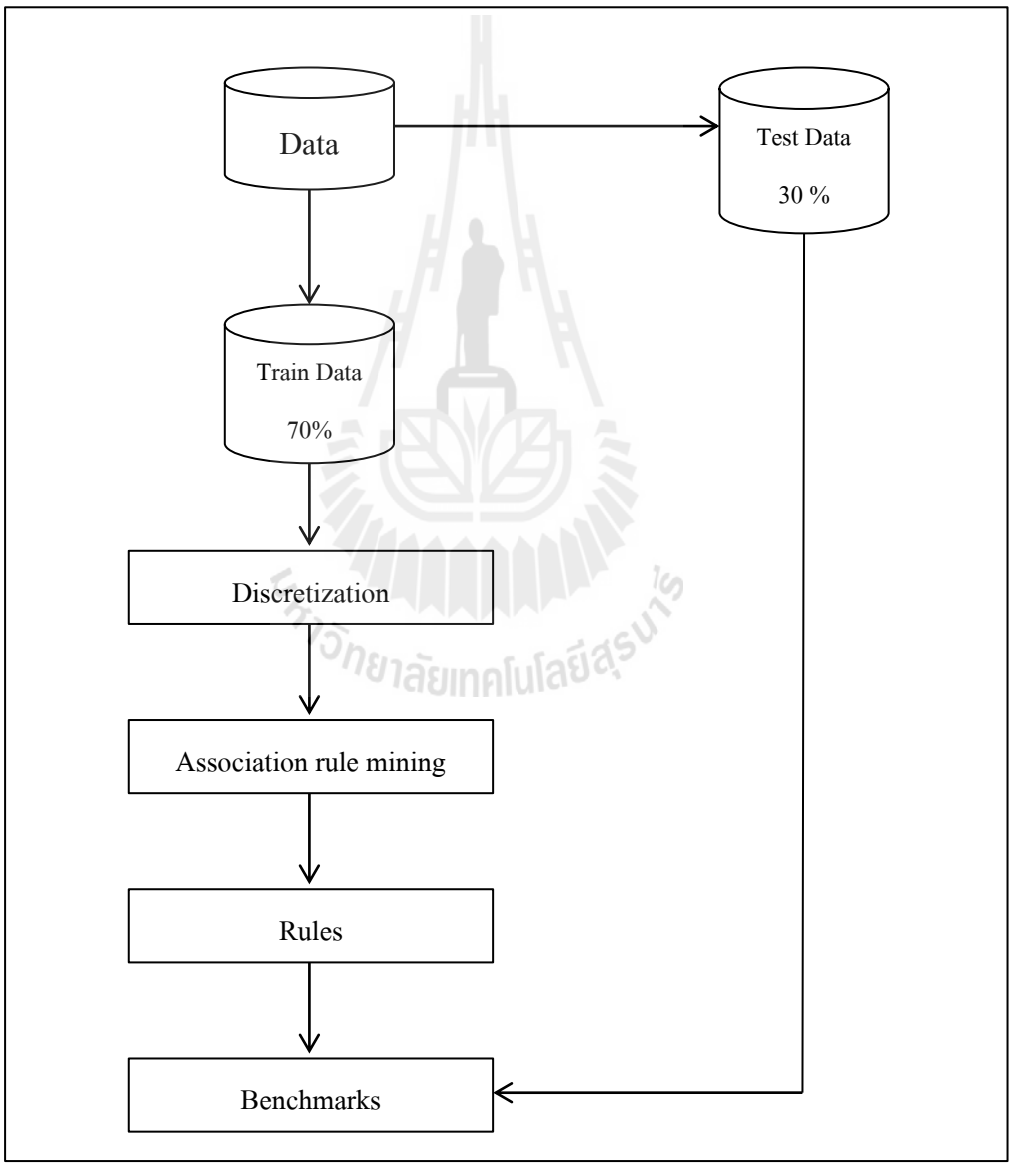
การทดสอบวิธีแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการหากฎความสัมพันธ์จะใช้ข้อมูลมาตรฐานจาก UCI Machine Learning Repository ซึ่งเป็นข้อมูลเกี่ยวกับโรคหัวใจ (Heart disease) สามารถดาวน์โหลดได้ที่ http://repository.seasr.org/Datasets/UCI/arff/ โดยมีข้อมูลทั้งหมด 303 แถว ประกอบไปด้วยคอลัมน์ 14 คอลัมน์ สามารถแบ่งเป็นคอลัมน์ที่เป็นตัวเลขต่อเนื่อง 5 คอลัมน์ และข้อมูลจะอยู่ในรูปแบบของไฟล์ .arff เพื่อให้มีความเหมาะสมกับการทำงานของโปรแกรม โดยมีรายละเอียดตัวอย่างข้อมูลดังรูปที่ 4.1

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
56	male	atyp_angina	120	236	f	normal	178	no	0.8	up	0	normal	<50
62	female	asympt	140	268	f	left_vent_hyper	160	no	3.6	down	2	normal	>50_1
63	male	asympt	130	254	f	left_vent_hyper	147	no	1.4	flat	1	reversable_defect	>50_1
57	male	asympt	140	192	f	normal	148	no	0.4	flat	0	fixed_defect	<50
57	male	non_anginal	150	168	f	normal	174	no	1.6	up	0	normal	<50
48	male	atyp_angina	110	229	f	normal	168	no	1	down	0	reversable_defect	>50_1
43	male	asympt	150	247	f	normal	171	no	1.5	up	0	normal	<50
60	male	asympt	117	230	t	normal	160	yes	1.4	up	2	reversable_defect	>50_1
43	male	asympt	120	177	f	left_vent_hyper	120	yes	2.5	flat	0	reversable_defect	>50_1
57	male	asympt	150	276	f	left_vent_hyper	112	yes	0.6	flat	1	fixed_defect	>50_1
55	male	asympt	132	353	f	normal	132	yes	1.2	flat	1	reversable_defect	>50_1
61	male	non_anginal	150	243	t	normal	137	yes	1	flat	0	normal	<50
65	female	asympt	150	225	f	left_vent_hyper	114	no	1	flat	3	reversable_defect	>50_1
61	female	asympt	130	330	f	left_vent_hyper	169	no	0	up	0	normal	>50_1
51	male	non_anginal	110	175	f	normal	123	no	0.6	up	0	normal	<50
44	male	asympt	112	290	f	left_vent_hyper	153	no	0	up	1	normal	>50_1
54	male	asympt	124	266	f	left_vent_hyper	109	yes	2.2	flat	1	reversable_defect	>50_1
51	female	asympt	130	305	f	normal	142	yes	1.2	flat	0	reversable_defect	>50_1
46	female	non_anginal	142	177	f	left_vent_hyper	160	yes	1.4	down	0	normal	<50
58	male	asympt	128	216	f	left_vent_hyper	131	yes	2.2	flat	3	reversable_defect	>50_1
54	female	non_anginal	135	304	t	normal	170	no	0	up	0	normal	<50
60	male	asympt	145	282	f	left_vent_hyper	142	yes	2.8	flat	2	reversable_defect	>50_1
60	male	non_anginal	140	185	f	left_vent_hyper	155	no	3	flat	0	normal	>50_1

รูปที่ 4.1 ตัวอย่างข้อมูลโรคหัวใจ (Heart disease)

4.2 การทดสอบประสิทธิภาพวิธีแบ่งช่วงข้อมูลด้วยอัลกอริทึมต่าง ๆ

วิธีทดสอบประสิทธิภาพการแบ่งช่วงข้อมูลสำหรับหากฎความสัมพันธ์จะใช้ค่าความถูกต้องเป็นตัวชี้วัดหลัก โดยแบ่งข้อมูลทดลองออกเป็น 2 ชุด คือ ข้อมูลในการเรียนรู้จำนวน 70% และข้อมูลทดสอบจำนวน 30% ซึ่งข้อมูลในการเรียนรู้จะถูกนำไปผ่านขั้นตอนการแบ่งช่วงข้อมูลและหากฎความสัมพันธ์ตามลำดับ เมื่อได้กฎความสัมพันธ์ออกมาก็จะนำข้อมูลทดสอบไปใช้ทำนายแต่ละกฎความสัมพันธ์เพื่อหาเป็นค่าความถูกต้องออกมา ดังรูปที่ 4.2 ที่แสดงแผนภาพวิธีการทดสอบประสิทธิภาพการแบ่งช่วงข้อมูลสำหรับหากฎความสัมพันธ์



รูปที่ 4.2 วิธีการทดสอบประสิทธิภาพการแบ่งช่วงข้อมูลสำหรับหากฎความสัมพันธ์

4.2.1 ผลของวิธีแบ่งช่วงข้อมูลด้วยอัลกอริทึม CAIM

ในการทดสอบวิธีแบ่งช่วงข้อมูลสำหรับการหากฎความสัมพันธ์ด้วยอัลกอริทึม CAIM ซึ่งเป็นอัลกอริทึมแบบมีผู้ฝึกสอน (Supervised) ดังนั้นจึงใช้คอลัมน์ num ที่มีจำนวนคลาสเท่ากับ 4 เป็นคอลัมน์เป้าหมาย เพราะชุดข้อมูลโรคหัวใจกำหนดให้คอลัมน์ num เป็นคอลัมน์เป้าหมายอยู่ก่อนแล้ว และการหากฎความสัมพันธ์กำหนดให้ Minimum Support = 0.4 และ Minimum Confidence = 0.8 ซึ่งผลการทดสอบที่ได้ดังรูปที่ 4.3 คือจุดตัดของช่วงข้อมูลในแต่ละคอลัมน์ที่เป็นตัวเลขต่อเนื่อง รูปที่ 4.4 แสดงตัวอย่าง 12 กฎความสัมพันธ์จากทั้งหมด 51 กฎความสัมพันธ์พร้อมมาตรวัด 4 มาตรวัด และรูปที่ 4.5 แสดงตัวอย่าง 12 กฎความสัมพันธ์จาก 51 กฎที่ใช้ข้อมูลทดสอบทำนายค่าความถูกต้องในแต่ละกฎความสัมพันธ์ ซึ่งได้ค่าความถูกต้อง 41.61%

Attributes	Cut Point
Age	0-29 = 1, 30-54.5 = 2, 54.6-76 = 3
Trestbps	0-94 = 1, 95-143 = 2, 144-200 = 3
Chol	0-149 = 1, 150-273.5 = 2, 273.6-417 = 3
Thalach	0-71 = 1, 72-146.5 = 2, 146.6-202 = 3
Oldpeak	<0 = 1, 0.1-0.75 = 2, 0.76-6.20 = 3

รูปที่ 4.3 ช่วงข้อมูลในแต่ละคอลัมน์ที่เป็นตัวเลขต่อเนื่องที่ได้จากอัลกอริทึม CAIM

Rules	support	confidence	lift	coverage
{cp=asympt} => {fbs=f}	0.40	0.90	1.03	0.45
{slope=flat} => {fbs=f}	0.42	0.91	1.05	0.46
{slope=up} => {fbs=f}	0.40	0.87	1.00	0.47
{age=1} => {trestbps=1}	0.43	0.91	1.17	0.47
{age=1} => {fbs=f}	0.43	0.90	1.04	0.47
{restecg=normal} => {fbs=f}	0.43	0.90	1.04	0.48
{oldpeak=1} => {trestbps=1}	0.42	0.87	1.11	0.48
{oldpeak=1} => {fbs=f}	0.44	0.90	1.04	0.48
{restecg=left_vent_hyper} => {fbs=f}	0.42	0.83	0.96	0.51
{oldpeak=2} => {fbs=f}	0.43	0.84	0.96	0.52
{age=2} => {fbs=f}	0.44	0.84	0.97	0.53
{thal=normal} => {num=<50}	0.43	0.80	1.41	0.54

รูปที่ 4.4 ตัวอย่าง 12 กฎความสัมพันธ์จากทั้งหมด 51 กฎความสัมพันธ์พร้อมมาตรวัด 4 มาตรวัด

Rules	True	False
{cp=asympt} => {fbs=f}	39	7
{slope=flat} => {fbs=f}	30	9
{slope=up} => {fbs=f}	33	7
{age=1} => {trestbps=1}	0	0
{age=1} => {fbs=f}	0	0
{restecg=normal} => {fbs=f}	37	9
{oldpeak=1} => {trestbps=1}	0	27
{oldpeak=1} => {fbs=f}	22	5
{restecg=left_vent_hyper} => {fbs=f}	30	7
{oldpeak=2} => {fbs=f}	12	3
{age=2} => {fbs=f}	33	6
{thal=normal} => {num=<50}	35	13

รูปที่ 4.5 ตัวอย่าง 12 กฎความสัมพันธ์จาก 51 กฎที่ใช้ข้อมูลทดสอบทำนายค่าความถูกต้องในแต่ละกฎความสัมพันธ์ ซึ่งได้ค่าความถูกต้อง 41.61%

4.2.2 ผลของวิธีแบ่งช่วงข้อมูลด้วยอัลกอริทึม k-means

ในการทดสอบวิธีแบ่งช่วงข้อมูลสำหรับการหากฎความสัมพันธ์ด้วยอัลกอริทึม k-means โดยอัลกอริทึม k-means เป็นอัลกอริทึมแบบไม่มีผู้ฝึกสอน (Unsupervised) ดังนั้นจึงไม่จำเป็นต้องใช้คอลัมน์เป้าหมายในการแบ่งช่วงข้อมูล และการหากฎความสัมพันธ์กำหนดให้ Minimum Support = 0.4 และ Minimum Confidence = 0.8 ซึ่งผลการทดสอบที่ได้ดังรูปที่ 4.6 คือ จุดตัดของช่วงข้อมูลในแต่ละคอลัมน์ที่เป็นตัวเลขต่อเนื่อง รูปที่ 4.7 จะเป็นตัวอย่าง 12 กฎความสัมพันธ์จากทั้งหมด 18 กฎความสัมพันธ์พร้อมมาตรวัด 4 มาตรวัด และรูปที่ 4.8 แสดงตัวอย่าง 12 กฎความสัมพันธ์จาก 18 กฎที่ใช้ข้อมูลทดสอบทำนายค่าความถูกต้องในแต่ละกฎความสัมพันธ์ ซึ่งได้ค่าความถูกต้อง 81.33%

Attributes	Cut Point
Age	65.38, 42.64, 55.54
Trestbps	115.51, 137.57, 168.71
Chol	310.98, 198.43, 248.80
Thalach	116.22, 172.23, 148.99
Oldpeak	3.69, 0.27, 1.77

รูปที่ 4.6 ช่วงข้อมูลในแต่ละคอลัมน์ที่เป็นตัวเลขต่อเนื่องที่ได้จากอัลกอริทึม k-means

Rules	support	confidence	lift	coverage
{cp=asympt} => {fbs=f}	0.40	0.90	1.03	0.45
{slope=flat} => {fbs=f}	0.42	0.91	1.05	0.46
{slope=up} => {fbs=f}	0.40	0.87	1.00	0.47
{trestbps=2} => {fbs=f}	0.41	0.85	0.98	0.48
{restecg=normal} => {fbs=f}	0.43	0.90	1.04	0.48
{restecg=left_vent_hyper} => {fbs=f}	0.42	0.83	0.96	0.51
{thal=normal} => {num=<50}	0.43	0.80	1.41	0.54
{thal=normal} => {fbs=f}	0.49	0.91	1.04	0.54
{num=<50} => {exang=no}	0.48	0.84	1.26	0.57
{num=<50} => {fbs=f}	0.49	0.87	1.00	0.57
{ca=0} => {fbs=f}	0.52	0.88	1.01	0.59
{oldpeak=2} => {fbs=f}	0.52	0.88	1.01	0.59

รูปที่ 4.7 ตัวอย่าง 12 กฎความสัมพันธ์จากทั้งหมด 18 กฎความสัมพันธ์พร้อมมาตรวัด 4 มาตรวัด

Rules	True	False
{cp=asympt} => {fbs=f}	39	7
{slope=flat} => {fbs=f}	30	9
{slope=up} => {fbs=f}	33	7
{trestbps=2} => {fbs=f}	36	10
{restecg=normal} => {fbs=f}	37	9
{restecg=left_vent_hyper} => {fbs=f}	30	7
{thal=normal} => {num=<50}	35	13
{thal=normal} => {fbs=f}	39	9
{num=<50} => {exang=no}	36	4
{num=<50} => {fbs=f}	33	7
{ca=0} => {fbs=f}	44	5
{oldpeak=2} => {fbs=f}	39	11

รูปที่ 4.8 ตัวอย่าง 12 กฎความสัมพันธ์จาก 18 กฎที่ใช้ข้อมูลทดสอบทำนายค่าความถูกต้องในแต่ละกฎความสัมพันธ์ ซึ่งได้ค่าความถูกต้อง 81.33%

4.2.3 ผลของวิธีแบ่งช่วงข้อมูลด้วยอัลกอริทึม Chi2

การทดสอบวิธีแบ่งช่วงข้อมูลสำหรับการหากฎความสัมพันธ์ด้วยอัลกอริทึม Chi2 โดยอัลกอริทึม Chi2 เป็นอัลกอริทึมแบบมีผู้ฝึกสอน (Supervised) ดังนั้นจึงใช้คอลัมน์ num ซึ่งมีจำนวนคลาสเท่ากับ 4 เป็นคอลัมน์เป้าหมาย และการหากฎความสัมพันธ์กำหนดให้ Minimum Support = 0.4 และ Minimum Confidence = 0.8 ซึ่งผลการทดสอบที่ได้ดังรูปที่ 4.9 คือจุดตัดของช่วงข้อมูลในแต่ละคอลัมน์ที่เป็นตัวเลขต่อเนื่อง รูปที่ 4.10 จะเป็นตัวอย่าง 12 กฎความสัมพันธ์จากทั้งหมด 19 กฎความสัมพันธ์พร้อมมาตรวัด 4 มาตรวัด และรูปที่ 4.11 แสดงตัวอย่าง 12 กฎความสัมพันธ์จาก 19 กฎที่ใช้ข้อมูลทดสอบทำนายค่าความถูกต้องในแต่ละกฎความสัมพันธ์ ซึ่งได้ค่าความถูกต้อง 81.86%

Attributes	Cut Point
Age	40.5, 45.5, 54.5, 70.5
Trestbps	106.5, 127, 137
Chol	162, 176.5, 183, 190, 205, 206.5, 211.5, 212.5, 216, 218.5, 222.5, 225.5, 227.5, 230.5, 252.5, 260.5, 273.5, 276, 279.5, 301, 325.5, 337.5
Thalach	146.5, 149.5, 150.5, 153.5
Oldpeak	0.75, 2.35, 3.45, 3.55

รูปที่ 4.9 ช่วงข้อมูลในแต่ละคอลัมน์ที่เป็นตัวเลขต่อเนื่องที่ได้จากอัลกอริทึม Chi2

Rules	support	confidence	lift	coverage
{cp=asympt} => {fbs=f}	0.40	0.90	1.03	0.45
{slope=flat} => {fbs=f}	0.42	0.91	1.05	0.46
{slope=up} => {fbs=f}	0.40	0.87	1.00	0.47
{restecg=normal} => {fbs=f}	0.43	0.90	1.04	0.48
{oldpeak=1} => {fbs=f}	0.44	0.90	1.04	0.48
{thalach=5} => {exang=no}	0.40	0.82	1.22	0.49
{thalach=5} => {fbs=f}	0.43	0.89	1.02	0.49
{age=4} => {fbs=f}	0.43	0.84	0.97	0.51
{restecg=left_vent_hyper} => {fbs=f}	0.42	0.83	0.96	0.51
{thal=normal} => {num=<50}	0.43	0.80	1.41	0.54
{thal=normal} => {fbs=f}	0.49	0.91	1.04	0.54
{num=<50} => {exang=no}	0.48	0.84	1.26	0.57

รูปที่ 4.10 ตัวอย่าง 12 กฎความสัมพันธ์จากทั้งหมด 19 กฎความสัมพันธ์พร้อมมาตรวัด 4 มาตรวัด

Rules	True	False
{cp=asympt} => {fbs=f}	39	7
{slope=flat} => {fbs=f}	30	9
{slope=up} => {fbs=f}	33	7
{restecg=normal} => {fbs=f}	37	9
{oldpeak=1} => {fbs=f}	34	8
{thalach=5} => {exang=no}	37	5
{thalach=5} => {fbs=f}	33	9
{age=4} => {fbs=f}	33	10
{restecg=left_vent_hyper} => {fbs=f}	30	7
{thal=normal} => {num=<50}	35	13
{thal=normal} => {fbs=f}	39	9
{num=<50} => {exang=no}	36	4

รูปที่ 4.11 ตัวอย่าง 12 กฎความสัมพันธ์จาก 19 กฎที่ใช้ข้อมูลทดสอบทำนายค่าความถูกต้องในแต่ละกฎความสัมพันธ์ ซึ่งได้ค่าความถูกต้อง 81.86%

4.2.4 ผลของวิธีแบ่งช่วงข้อมูลด้วยอัลกอริทึม Chi2+Select target

การทดสอบในแบบสุดท้าย เป็นการทดสอบวิธีแบ่งช่วงข้อมูลสำหรับการหากฎความสัมพันธ์ด้วยอัลกอริทึม Chi2+Select target ซึ่งเป็นอัลกอริทึมใหม่ที่เสนอในงานวิจัยฉบับนี้ Chi2+Select target จะทำการเลือกคอลัมน์เป้าหมายที่มีความสัมพันธ์น้อยที่สุดซึ่งในข้อมูลโรคหัวใจ คือคอลัมน์ CP และการหากฎความสัมพันธ์กำหนดให้ Minimum Support = 0.4 และ Minimum Confidence = 0.8 ซึ่งผลการทดสอบที่ได้ดังรูปที่ 4.12 คือจุดตัดของช่วงข้อมูลในแต่ละคอลัมน์ที่เป็นตัวเลขต่อเนื่อง รูปที่ 4.13 จะเป็นตัวอย่าง 12 กฎความสัมพันธ์จากทั้งหมด 63 กฎความสัมพันธ์พร้อมมาตรวัด 4 มาตรวัด และรูปที่ 4.14 แสดงตัวอย่าง 12 กฎความสัมพันธ์จาก 63 กฎที่ใช้ข้อมูลทดสอบทำนายค่าความถูกต้องในแต่ละกฎความสัมพันธ์ ซึ่งได้ค่าความถูกต้อง 85.54%

Attributes	Cut Point
Age	57.5
Trestbps	135.5
Chol	162, 177.5
Thalach	145.5, 152.5, 173.5, 178.5
Oldpeak	0.05, 0.85, 1.25, 1.45

รูปที่ 4.12 ช่วงข้อมูลในแต่ละคอลัมน์ที่เป็นตัวเลขต่อเนื่องที่ได้จากอัลกอริทึม Chi2+Select target

Rules	support	confidence	lift	coverage
{cp=asympt} => {fbs=f}	0.40	0.90	1.03	0.45
{cp=asympt} => {chol=3}	0.40	0.91	0.96	0.45
{slope=flat} => {fbs=f}	0.42	0.91	1.05	0.46
{slope=flat} => {chol=3}	0.44	0.95	1.01	0.46
{slope=up} => {fbs=f}	0.40	0.87	1.00	0.47
{slope=up} => {chol=3}	0.44	0.95	1.01	0.47
{restecg=normal} => {fbs=f}	0.43	0.90	1.04	0.48
{restecg=normal} => {chol=3}	0.45	0.94	1.00	0.48
{restecg=left_vent_hyper} => {fbs=f}	0.42	0.83	0.96	0.51
{restecg=left_vent_hyper} => {chol=3}	0.48	0.94	1.00	0.51
{thal=normal} => {num=<50}	0.43	0.80	1.41	0.54
{thal=normal} => {fbs=f}	0.49	0.91	1.04	0.54

รูปที่ 4.13 ตัวอย่าง 12 กฎความสัมพันธ์จากทั้งหมด 63 กฎความสัมพันธ์พร้อมมาตรวัด 4 มาตรวัด

Rules	True	False
{cp=asympt} => {fbs=f}	39	7
{cp=asympt} => {chol=3}	43	3
{slope=flat} => {fbs=f}	30	9
{slope=flat} => {chol=3}	36	3
{slope=up} => {fbs=f}	33	7
{slope=up} => {chol=3}	37	3
{restecg=normal} => {fbs=f}	37	9
{restecg=normal} => {chol=3}	40	6
{restecg=left_vent_hyper} => {fbs=f}	30	7
{restecg=left_vent_hyper} => {chol=3}	35	2
{thal=normal} => {num=<50}	35	13
{thal=normal} => {fbs=f}	39	9

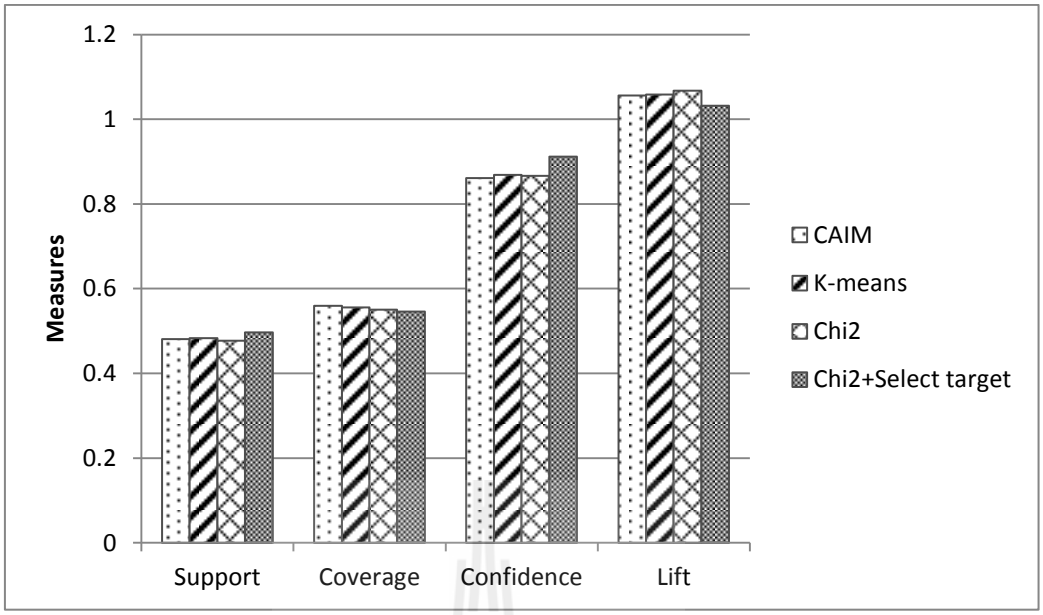
รูปที่ 4.14 ตัวอย่าง 12 กฎความสัมพันธ์จาก 63 กฎที่ใช้ข้อมูลทดสอบทำนายค่าความถูกต้องในแต่ละกฎความสัมพันธ์ ซึ่งได้ค่าความถูกต้อง 85.54%

4.3 เปรียบเทียบผลการทดลองวิธีแบ่งช่วงข้อมูลด้วยอัลกอริทึมต่าง ๆ

วิธีแบ่งช่วงข้อมูลสำหรับการหากฎความสัมพันธ์ทำให้กฎความสัมพันธ์ที่ได้ออกมานั้นมีประสิทธิภาพที่ดี โดยในงานวิจัยนี้ได้ใช้ข้อมูลโรคหัวใจในการทดสอบประสิทธิภาพ ซึ่งได้เปรียบเทียบทั้งหมดสี่อัลกอริทึม คือ CAIM, k-means, Chi2 และอัลกอริทึมที่ผู้วิจัยเสนอขึ้นใหม่ชื่อ Chi2+Select target การเปรียบเทียบจะใช้ตัวชี้วัด คือ ค่าความถูกต้องในการทำนายแต่ละกฎความสัมพันธ์ของแต่ละอัลกอริทึม และค่าเฉลี่ย Support, Coverage, Confidence และ Lift ในแต่ละอัลกอริทึม โดยจะแสดงการเปรียบเทียบในตารางที่ 4.1 และแสดงการเปรียบเทียบค่าเฉลี่ยสหสัมพันธ์ และค่าความถูกต้องในแต่ละคอลัมน์เป้าหมายในตารางที่ 4.3 การเปรียบเทียบค่าความถูกต้องในการทำนายแต่ละกฎความสัมพันธ์ของแต่ละอัลกอริทึมแสดงในตารางที่ 4.3 การเปรียบเทียบผลในตารางที่ 4.1 4.2 และ 4.3 แสดงเป็นภาพกราฟได้ดังรูปที่ 4.15 4.16 และ 4.17 ตามลำดับ

ตารางที่ 4.1 ค่าเฉลี่ย Support, Coverage, Confidence และ Lift ในแต่ละอัลกอริทึม

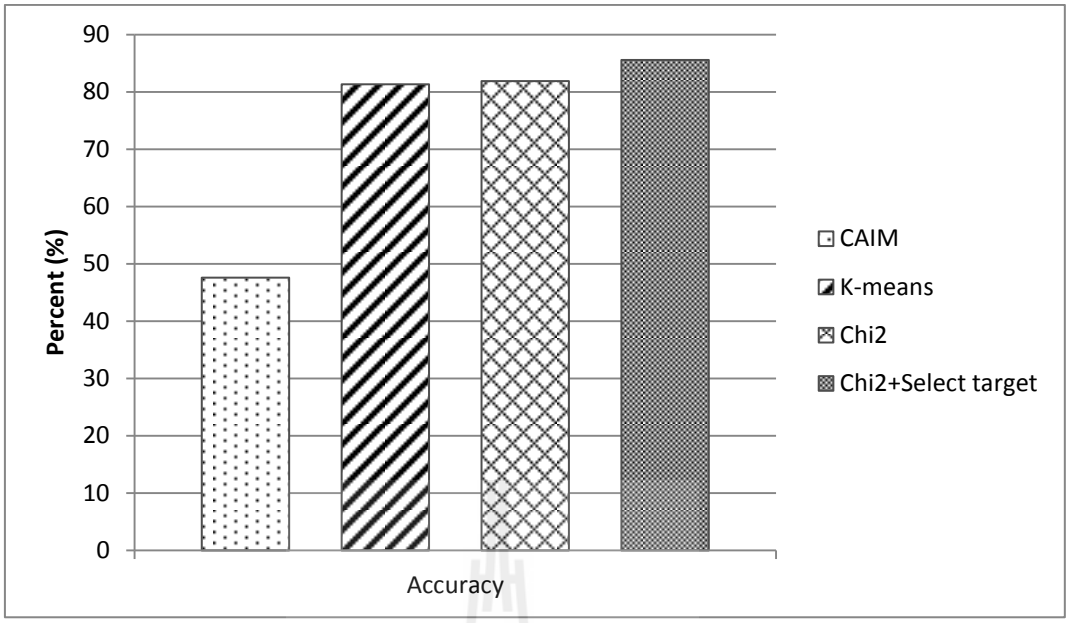
Algorithms	Support	Confidence	Lift	Coverage
CAIM	0.48	0.86	1.06	0.56
k-means	0.48	0.87	1.06	0.56
Chi2	0.48	0.87	1.07	0.55
Chi2+Select target	0.50	0.91	1.03	0.55



รูปที่ 4.15 แผนภูมิแสดงการเปรียบเทียบค่าเฉลี่ย Support, Coverage, Confidence และ Lift จากวิธีการแบ่งช่วงข้อมูลสำหรับหากฎความสัมพันธ์ด้วยอัลกอริทึมต่าง ๆ

ตารางที่ 4.2 ค่าความถูกต้องในการทำนายแต่ละกฎความสัมพันธ์ของแต่ละอัลกอริทึม

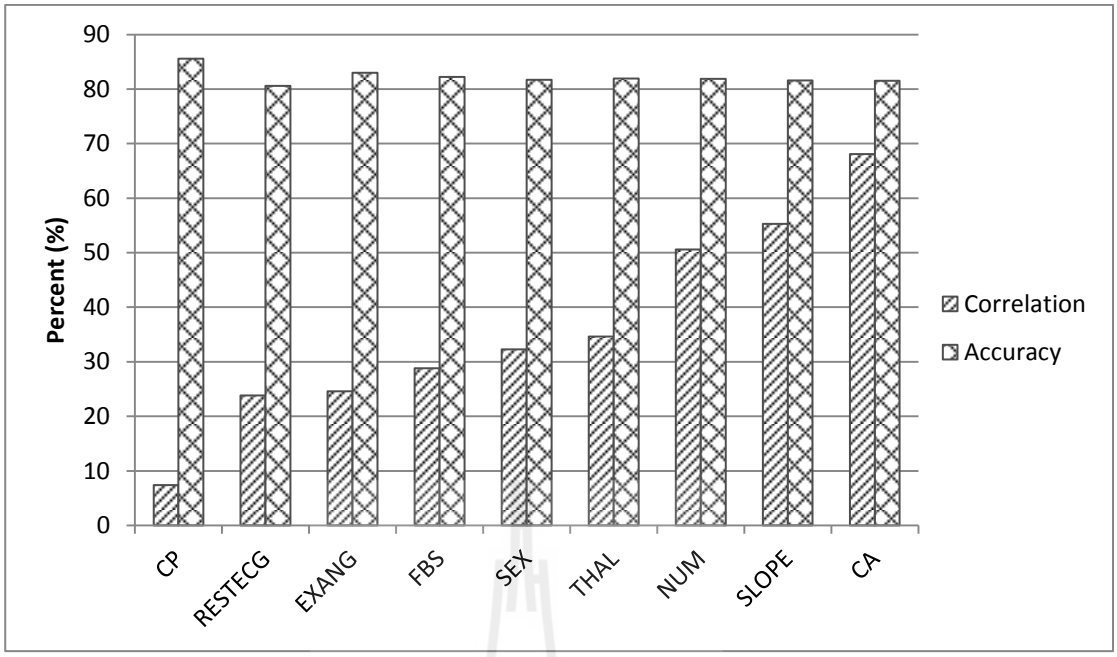
Algorithms	**True**	**False**	**Accuracy (%)**
CAIM	637	701	47.61
k-means	610	140	81.33
Chi2	641	142	81.86
Chi2+Select target	2325	393	85.54



รูปที่ 4.16 แผนภูมิแสดงการเปรียบเทียบค่าความถูกต้องวิธีการแบ่งช่วงข้อมูลสำหรับหากฎความสัมพันธ์ด้วยอัลกอริทึมต่าง ๆ

ตารางที่ 4.3 ค่าเฉลี่ยสหสัมพันธ์ และค่าความถูกต้องในแต่ละคอลัมน์เป้าหมาย

Attributes	Correlation	Accuracy (%)
CP	0.0740	85.54
RESTECG	0.2380	80.58
EXANG	0.2457	83.00
FBS	0.2883	82.23
SEX	0.3227	81.67
THAL	0.3464	81.95
NUM	0.5058	81.86
SLOPE	0.5526	81.60
CA	0.6806	81.53



รูปที่ 4.17 แผนภูมิแสดงการเปรียบเทียบค่าเฉลี่ยสหสัมพันธ์ และค่าความถูกต้องในแต่ละคอลัมน์เป้าหมาย

จากตารางที่ 4.3 แสดงการเปรียบเทียบค่าเฉลี่ยสหสัมพันธ์ และค่าความถูกต้องในแต่ละคอลัมน์เป้าหมายที่นำมาใช้ในขั้นตอนการแบ่งช่วงข้อมูล เนื่องจากในงานวิจัยนี้ได้เสนอเทคนิคการเลือกคอลัมน์เป้าหมายสำหรับการแบ่งช่วงข้อมูล โดยพิจารณาจากค่าสหสัมพันธ์ ดังนั้นจึงต้องการเปรียบเทียบในแต่ละคอลัมน์เป้าหมายที่แตกต่างกันจะให้ค่าสหสัมพันธ์และค่าความถูกต้องเป็นอย่างไร โดยแสดงเป็นภาพกราฟได้ดังรูปที่ 4.17

4.4 อภิปรายผล

จากผลการทดสอบประสิทธิภาพวิธีแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยอัลกอริทึมที่มีผู้ฝึกสอนคือ CAIM, Chi2, Chi2+Select target และอัลกอริทึมที่ไม่มีผู้ฝึกสอนคือ k-means สำหรับการหากฎความสัมพันธ์ด้วยข้อมูลโรคหัวใจจำนวน 303 รายการ สามารถสรุปผลการทดสอบเปรียบเทียบได้ดังนี้

1) การเปรียบเทียบโดยใช้ตัวชี้วัดคือค่าเฉลี่ยของ Support, Coverage, Confidence และ Lift ในแต่ละอัลกอริทึม เมื่อนำมาใช้เปรียบเทียบประสิทธิภาพวิธีแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยอัลกอริทึมต่าง ๆ จากตารางที่ 4.1 จะเห็นได้ว่าวิธีแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยอัลกอริทึม Chi2+Select target มีค่า Support และ Confidence มากที่สุด แต่จากรูปที่ 4.15 จะ

เห็นได้ว่าในแต่ละอัลกอริทึมมีค่าเฉลี่ย Support, Coverage, Confidence และ Lift ที่ไม่แตกต่างกันมาก ทำให้ไม่สามารถเจาะจงได้ว่าอัลกอริทึมใดมีประสิทธิภาพการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการหากฎความสัมพันธ์ที่ดีที่สุด

2) การเปรียบเทียบโดยใช้ตัวชี้วัดคือค่าความถูกต้องที่ได้จากการทำนายในแต่ละกฎความสัมพันธ์ เมื่อนำมาใช้เปรียบเทียบประสิทธิภาพวิธีแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยอัลกอริทึมต่าง ๆ จากรูปที่ 4.16 จะเห็นได้ว่าวิธีแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยอัลกอริทึม Chi2+Select target มีค่าความถูกต้องมากที่สุด เพราะว่าวิธีแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยอัลกอริทึม CAIM, Chi2 และ Chi2+Select target เป็นอัลกอริทึมแบบมีผู้ฝึกสอน ดังนั้นการเลือกคอลัมน์เป้าหมายจะมีผลต่อการแบ่งช่วงข้อมูล และจากตารางที่ 4.3 จะเห็นได้ว่าจำนวนข้อมูลทดสอบที่เข้าเงื่อนไขกฎความสัมพันธ์ที่ได้จากวิธีแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยอัลกอริทึม Chi2+Select target นั้นมีจำนวนมากที่สุดซึ่งสอดคล้องกับตัวชี้วัดก่อนหน้านี้ คือวิธีแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยอัลกอริทึม Chi2+Select target มีค่า Support และ Coverage มากที่สุด

จากผลการทดสอบประสิทธิภาพวิธีแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยอัลกอริทึม CAIM, k-means, Chi2 และ Chi2+Select target ผลการทดสอบสรุปได้ว่าการเปรียบเทียบโดยใช้ตัวชี้วัดที่เป็นค่าเฉลี่ยของ Support, Coverage, Confidence และ Lift ในแต่ละอัลกอริทึม ผลที่ได้คือวิธีแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยอัลกอริทึม Chi2 และ Chi2+Select target ให้ค่าดีที่สุด แต่อย่างไรก็ตามอัลกอริทึมต่าง ๆ มีค่าเฉลี่ยของ Support, Coverage, Confidence และ Lift ที่ไม่แตกต่างกันมาก เมื่อเปรียบเทียบโดยใช้ตัวชี้วัดคือค่าความถูกต้องที่ได้จากการทำนายในแต่ละกฎความสัมพันธ์ ให้ผลแตกต่างที่ชัดเจนมากกว่าคือวิธีแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยอัลกอริทึม Chi2+Select target มีค่าความถูกต้องมากที่สุด คือ 85.54% โดย Chi2, k-means และ CAIM มีค่าความถูกต้องที่ลดหลั่นลงไปตามลำดับคือ 81.86% 81.33% และ 47.61% ความถูกต้องที่สูงที่สุดของอัลกอริทึม Chi2+Select target นี้เนื่องจากอัลกอริทึม Chi2+Select target เป็นอัลกอริทึมแบบมีผู้ฝึกสอนจึงจำเป็นต้องใช้คอลัมน์เป้าหมายในการแบ่งช่วงข้อมูล จากรูปที่ 4.17 จะเห็นได้เทคนิคการเลือกคอลัมน์เป้าหมายที่มีค่าสหสัมพันธ์น้อยที่สุดมีผลสำหรับการแบ่งช่วงข้อมูล ซึ่งเป็นเหตุผลสำคัญที่ให้อัลกอริทึม Chi2+Select target มีค่าความถูกต้องมากกว่าอัลกอริทึมอื่น

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

ปัจจุบันเทคนิคในงานทางด้านการทำเหมืองข้อมูลเพื่อให้ได้ความรู้นั้นมีหลากหลายเทคนิค ซึ่งเทคนิคหนึ่งที่ถูกนำไปใช้กันอย่างแพร่หลายคือการหากฎความสัมพันธ์ แต่การหากฎความสัมพันธ์ยังคงมีปัญหาในส่วนของชนิดข้อมูลที่นำมาใช้หากฎความสัมพันธ์ นั่นก็คือข้อมูลที่มีลักษณะเป็นตัวเลขต่อเนื่อง เพราะว่าข้อมูลที่มีลักษณะเป็นตัวเลขต่อเนื่องเมื่อนำไปหากฎความสัมพันธ์จะทำให้ได้กฎความสัมพันธ์ที่จำนวนมากและไม่สามารถนำกฎความสัมพันธ์ที่ได้ไปใช้ในการทำนายข้อมูลได้อย่างถูกต้อง ดังนั้นจึงต้องมีเทคนิคที่จะนำมาใช้จัดการข้อมูลก่อนการนำไปหากฎความสัมพันธ์ ซึ่งเป็นขั้นตอนก่อนการประมวลผล เรียกว่า การแบ่งช่วงข้อมูล (Discretization)

ในงานวิจัยนี้มุ่งเน้นในกระบวนการออกแบบอัลกอริทึมและพัฒนาโปรแกรมเพื่อแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการหากฎความสัมพันธ์ รวมถึงกระบวนการออกแบบอัลกอริทึมสำหรับการทดสอบประสิทธิภาพกฎความสัมพันธ์ของข้อมูลที่ได้จากวิธีแบ่งช่วงข้อมูล ซึ่งจะทำให้ได้กฎความสัมพันธ์จากข้อมูลที่มีลักษณะเป็นตัวเลขต่อเนื่องมีประสิทธิภาพเพียงพอที่จะสามารถนำไปใช้ทำนายข้อมูลได้ โดยอัลกอริทึมและโปรแกรมสำหรับการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการหากฎความสัมพันธ์ จะเป็นการผสมผสานระหว่างอัลกอริทึมการแบ่งช่วงข้อมูลแบบมีผู้ฝึกสอน (Supervised) และแบบไม่มีผู้ฝึกสอน (Unsupervised) ซึ่งในส่วนอัลกอริทึมการแบ่งช่วงข้อมูลแบบมีผู้ฝึกสอนจะเพิ่มเติมเทคนิคในการเลือกคอลัมน์เป้าหมายที่ให้กฎความสัมพันธ์ที่มีประสิทธิภาพดีที่สุด

5.1 ขั้นตอนการดำเนินงานวิจัย

งานวิจัยนี้ได้พัฒนาอัลกอริทึมเพื่อการแบ่งช่วงข้อมูล ชื่อ Chi2+Select target ขั้นตอนของงานวิจัยนี้แบ่งออกเป็น

1) การศึกษาการทำงานของอัลกอริทึมไคสแควร์ (Chi-square) และการศึกษาการเขียนโปรแกรมด้วยภาษาอาร์ (R language) ซึ่งเป็นภาษาที่เหมาะสำหรับชุดข้อมูลขนาดใหญ่ และมีชุดคำสั่งในงานทางด้านทำเหมืองข้อมูล ซึ่งทำให้ง่ายในการนำไปใช้พัฒนาอัลกอริทึม

2) การออกแบบอัลกอริทึมการแบ่งช่วงข้อมูลที่เพิ่มเติมจากอัลกอริทึม Chi2 ทั้งนี้เนื่องจากอัลกอริทึม Chi2 เป็นอัลกอริทึมแบบมีผู้ฝึกสอนดังนั้นจำเป็นต้องมีคอลัมน์เป้าหมายในขั้นตอนการแบ่งช่วงข้อมูล แต่เนื่องจากชุดข้อมูลหนึ่งชุดอาจมีคอลัมน์เป้าหมายได้มากกว่า 1 คอลัมน์ ดังนั้นจึงเสนอเทคนิคการเลือกคอลัมน์เป้าหมายที่มีค่าสหสัมพันธ์ (Correlation) ระหว่างคอลัมน์เป้าหมายกับคอลัมน์ที่เป็นตัวเลขต่อเนื่องที่น้อยสุด เหตุผลสนับสนุนแนวคิดนี้คืออัลกอริทึม Chi2 จะให้ประสิทธิภาพที่ดีเมื่อตัวแปรต้นกับตัวแปรตามต้องเป็นอิสระต่อกัน

3) การออกแบบอัลกอริทึมการแบ่งช่วงข้อมูลจะแบ่งออกเป็นอัลกอริทึมแบบมีผู้ฝึกสอนและอัลกอริทึมแบบไม่มีผู้ฝึกสอนซึ่งจะใช้กับชุดข้อมูลที่แตกต่างกัน โดยถ้าข้อมูลทุกคอลัมน์เป็นตัวเลขต่อเนื่องจะใช้อัลกอริทึม k-means แบ่งช่วงข้อมูล แต่ถ้าข้อมูลมีบางคอลัมน์ที่ไม่ใช่ตัวเลขและข้อมูลบางคอลัมน์เป็นตัวเลขต่อเนื่องจะใช้อัลกอริทึม Chi2 แบ่งช่วงข้อมูล

4) กระบวนการออกแบบอัลกอริทึมสำหรับการทดสอบประสิทธิภาพกฎความสัมพันธ์ของข้อมูลที่ได้จากวิธีแบ่งช่วงข้อมูล โดยจะนำชุดข้อมูลทดสอบมาใช้ในการประเมินผลการทำนายในแต่ละกฎความสัมพันธ์ ซึ่งจะได้ผลการประเมินออกมาเป็นค่าความถูกต้องและค่าความผิดพลาดของแต่ละกฎความสัมพันธ์

การทดสอบประสิทธิภาพกฎความสัมพันธ์ของข้อมูลที่ได้จากวิธีแบ่งช่วงข้อมูล จะทดสอบกับข้อมูลโรคหัวใจ (Heart disease) ซึ่งเป็นชุดข้อมูลมาตรฐานโดยการเปรียบเทียบประสิทธิภาพการแบ่งช่วงข้อมูลทั้งสี่อัลกอริทึม (CAIM, k-means, Chi2 และ Chi2+Select target) จะทำการเปรียบเทียบจากการทดสอบประสิทธิภาพกฎความสัมพันธ์ของข้อมูลที่ได้จากวิธีแบ่งช่วงข้อมูลแต่ละวิธี และใช้มาตรวัดมาตรฐาน 4 มาตรวัด คือ Support, Coverage, Confidence และ Lift นอกจากนี้ยังได้เพิ่มมาตรวัดความถูกต้อง (Accuracy) ของกฎความสัมพันธ์ที่เสนอขึ้นใหม่ในงานวิจัยฉบับนี้เพื่อแสดงให้เห็นถึงประสิทธิภาพกฎความสัมพันธ์ของข้อมูลที่ได้จากวิธีแบ่งช่วงข้อมูลด้วยอัลกอริทึมต่าง ๆ

5.2 สรุปผลการวิจัย

ผลการทดสอบประสิทธิภาพวิธีแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยอัลกอริทึม CAIM, K-means, Chi2 และ Chi2+Select target ผลการทดสอบโดยใช้ตัวชี้วัดค่าเฉลี่ยของ Support, Coverage, Confidence และ Lift ในแต่ละอัลกอริทึมมีค่าที่แทบจะไม่แตกต่างกัน และในแต่ละตัวชี้วัดให้ค่าในแต่ละอัลกอริทึมไม่ไปในทิศทางเดียวกัน ซึ่งทำให้ไม่สามารถเจาะจงได้ว่าอัลกอริทึมใดที่ให้ประสิทธิวิธีแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องที่ดีที่สุด แต่เมื่อเปรียบเทียบโดยใช้ตัวชี้วัดที่ได้เสนอขึ้นใหม่ คือค่าความถูกต้อง (Accuracy) พบว่าทำให้สามารถเจาะจงได้ว่า

อัลกอริทึมใดให้ประสิทธิภาพดีที่สุด โดยอัลกอริทึม Chi2+Select target ให้ประสิทธิภาพดีที่สุด เนื่องจากใช้แนวคิดพื้นฐานที่ว่าอัลกอริทึม Chi2 เป็นอัลกอริทึมแบบมีผู้ฝึกสอนจึงจำเป็นต้องใช้คอลัมน์เป้าหมายในการแบ่งช่วงข้อมูล ทำให้คอลัมน์เป้าหมายที่เลือกนำมาใช้มีผลสำหรับประสิทธิภาพการแบ่งช่วงข้อมูล ซึ่งเป็นเหตุผลที่ให้อัลกอริทึม Chi2+Select target มีประสิทธิภาพการแบ่งช่วงข้อมูลที่ดีที่สุดเมื่อเทียบกับอัลกอริทึมอื่น

5.3 ปัญหาและข้อเสนอแนะ

ขั้นตอนเลือกคอลัมน์เป้าหมายและการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่อง ถ้าข้อมูลที่ใช้มีข้อมูลที่ผิดปกติ (Outlier) จะทำให้ขั้นตอนเลือกคอลัมน์เป้าหมายและการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องนั้นให้ประสิทธิภาพที่ไม่ดี และในขั้นตอนการเลือกคอลัมน์เป้าหมายที่มีค่าความสัมพันธ์ระหว่างคอลัมน์เป้าหมายกับคอลัมน์ที่เป็นตัวเลขต่อเนื่องที่น้อยสุด ถ้าชุดข้อมูลมีขนาดใหญ่อาจทำให้ใช้เวลานานในการหาค่าความสัมพันธ์หรือโปรแกรมอาจจะหยุดระหว่างการหาค่าความสัมพันธ์ระหว่างคอลัมน์เป้าหมายกับคอลัมน์ที่เป็นตัวเลขต่อเนื่อง

ในอนาคตถ้านำจุดเด่นในส่วนของการเลือกคอลัมน์เป้าหมายสำหรับอัลกอริทึมแบบมีผู้ฝึกสอนที่จำเป็นต้องใช้คอลัมน์เป้าหมายในการแบ่งช่วงข้อมูล นำไปประยุกต์ใช้กับอัลกอริทึมอื่นที่เป็นอัลกอริทึมแบบมีผู้ฝึกสอนด้วยกัน อาจทำให้อัลกอริทึมใหม่ที่ได้มีประสิทธิภาพสูงขึ้นกว่าเดิม การเลือกคอลัมน์เป้าหมายในการวิเคราะห์ค่าสหสัมพันธ์ในงานวิจัยนี้สามารถนำไปปรับปรุงให้เลือกคอลัมน์ที่มีค่าสหสัมพันธ์น้อยที่สุดแทนที่จะใช้ค่าเฉลี่ยสหสัมพันธ์ซึ่งอาจทำให้ได้ค่าที่ไม่เป็นจริง และในขั้นตอนก่อนการแบ่งช่วงข้อมูลในงานวิจัยนี้สามารถนำไปปรับปรุงให้จัดการกับข้อมูลที่ผิดปกติได้ ซึ่งจะทำให้อัลกอริทึมนั้นมีความทนทานต่อข้อมูลที่นำมาใช้มากยิ่งขึ้น นอกจากนี้ยังสามารถนำวิธีวัดความถูกต้อง (Accuracy) สำหรับการทดสอบประสิทธิภาพกฎความสัมพันธ์ของข้อมูลที่ได้จากวิธีแบ่งช่วงข้อมูลที่ได้เสนอขึ้นใหม่ในงานวิจัยนี้ไปปรับปรุงให้สามารถวัดประสิทธิภาพกฎความสัมพันธ์ให้เป็นอัตโนมัติมากขึ้น

รายการอ้างอิง

กิตติศักดิ์ เกิดประสพ. (2012). Data Mining With R [ออนไลน์]. ได้จาก **https://sites.google.com /site/kittisakthailand55/home/datamining2 -55**

อิศรัฏฐ์ รินไธสง. (2012). สถิติสำหรับการวิจัยทางการศึกษา [ออนไลน์]. ได้จาก **http://www.edu.tsu.ac.th/major/administration/data/FE511/บทที่ 18 การวิเคราะห์ค่าสหสัมพันธ์.doc**

Luis Gonzalez Abril, Francisco Javier Cuberos, Francisco Velasco, and Juan Antonio Ortega (2009). **Ameva: An Autonomous Discretization Algorithm.** Expert Systems with Applications, vol.36, no.3, pp.5327-5332

Rakesh Agrawal, Tomasz Imielinski, and Arun Swami (1993). **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

Attila Gyenesei (2001). **A Fuzzy Approach for Mining Quantitative Association Rules**. Proceedings of the Acta Cybernetica, pp.305-320

Michael Hahsler (2011). **A Comparison of Commonly Used Interest Measures for Association Rules [Online].** Available URL: http://michael.hahsler.net/research/association_rules/measures.html

Haiyang Hua and Huaici Zhao (2009). **A Discretization Algorithm of Continuous Attributes Based on Supervised Clustering**. Proceedings of the Chinese Conference on Pattern Recognition, pp.1-5

Yiping Ke, James Cheng, and Wilfred Ng (2008). **MIC Framework: An Information-Theoretic Approach to Quantitative Association Rule Mining.** Knowledge and Information Systems, vol.16, no.2 , pp.213-244

Lukasz Kurgan and Krzysztof Cios (2004). **CAIM Discretization Algorithm**. IEEE Transactions on Knowledge and Data Engineering, vol.16, no.2, pp.145-153

Kenneth Lai and Narciso Cerpa (2001). **Support vs Confidence in Association Rule Algorithms**. Proceedings of the OPTIMA Conference, pp.1-14

Huan Liu, Farhad Hussain, Chew Lim Tan, and Manoranjan Dash (2002). **Discretization: An Enabling Technique**. Data Mining and Knowledge Discovery, vol.6, no.4, pp.393-423

Huan Liu and Rudy Setiono (1995). **Chi2: Feature Selection and Discretization of Numeric Attributes**. Proceedings of the Seventh International Conference on Tools with Artificial Intelligence, pp.372-390

María N. Moreno, Saddys Segrera, Vivian F. López, and José M. Polo (2006). **A Method for Mining Quantitative Association Rules.** Proceedings of the WSEAS International Conference on Simulation, pp.173-178

Emmanuel Paradis, Julien Claude, and Korbinian Strimmer (2004). **APE: Analyses of Phylogenetics and Evolution in R language**. Bioinformatics, vol.20, no.2, pp.289-290

Cheng-Jung Tsai, Chien-I. Lee, and Wei-Pang Yang (2008). **A Discretization Algorithm Based on Class-Attribute Contingency Coefficient.** Information Sciences: an International Journal, vol.178, no.3, pp.714-731

Chao-Ton Su and Jyh-Hwa Hsu (2005). **An Extended Chi2 Algorithm for Discretization of Real Value Attributes**. IEEE Transactions on Knowledge and Data Engineering, vol.17, no.3, pp.437-441

Hyontai Sug (2011). **Discovery of Multidimensional Association Rules Focusing on Instances in Specific Class.** International Journal of Mathematics and Computers in Simulation, vol.5, no.3, pp.250-257

Lixiang Shen and Francis E. H. Tay (2001). **A Discretization Method for Rough Sets Theory**. IEEE Transactions on Knowledge and Data Engineering, vol.14, no.3, pp.666-670

Hantian Wei (2009). **A Novel Multivariate Discretization Method for Mining Association Rules.** Proceedings of the Information Processing, pp.378-381

Dennis E. Hinkle, William Wiersma, and Stephen G. Jurs. (1998**). Applied Statistic for the Behavioral Sciences 5 th edition**, p.118

Wikipedia, The Free Encyclopedia (2012a). **Lift (data mining) [Online].** Available URL: http://en.wikipedia.org/wiki/Lift_(data_mining)

Wikipedia, The Free Encyclopedia (2012b). **Association rule learning [Online].** Available URL: http://en.wikipedia.org/wiki/Association_rule_learning

Wikipedia, The Free Encyclopedia (2012c). **Cluster analysis [Online].** Available URL: http://en.wikipedia.org/wiki/Cluster_analysis

Wikipedia, The Free Encyclopedia (2012d). **K-means clustering [Online].** Available URL: http://en.wikipedia.org/wiki/K-means_clustering

ภาคผนวก ก

รหัสต้นฉบับของโปรแกรม

โปรแกรมแบ่งช่วงข้อมูลสำหรับหากฎความสัมพันธ์

```
#-----Cut missing value---------
cutMissing<- function(data){
                    cutMissing <-na.omit(data)
                    return(cutMissing)
}
#-----Check class---------
check_type<- function(data){
   count<-1
   for(i in seq(ncol(data))){
          if(!is.numeric(data[,i]))
                    count<-count+1
   }
   if(count==1)
          class<-F
   else
          class<-T

   return(class)
}
#-----extract numerical and target from data---------
extract_numeric<-function(data,class){
          list<-c()
          count<-1
          for(i in seq(ncol(data))){
                    if(is.numeric(data[,i])){
                              list[count]<-i
                              count<-count+1
                    }
```

```
   }
        list[count]<-class
return(list(data=data[,list],ls=list))
}
#------Replace numeric with intervals----------
replace_data<-function(data,discData,list){
        for(i in seq(ncol(discData))){
                data[list[i]]<-discData[i]
        }
        return(data)
}
#------Find Minimum Correlation----------------------
min_cor<-function(data){
        minCor<-0
        point<-0
        count<-1
        for(i in seq(ncol(data))){
                if(is.numeric(data[,i])){
                        list[count]<-i
                        count<-count+1
                }
        }
        for(i in seq(ncol(data))){
                if(!is.numeric(data[,i])){
                        correlation<-cor(data[,as.numeric(list)],as.numeric(data[,i]))
                        meanCor<-abs(sum(correlation)/length(list))
                        if(point==0){
                                minCor<-meanCor
                                point<-i
                        }else{
```

```
                                        if(minCor>meanCor){
                                                minCor<-meanCor
                                                point<-i
                                        }
                                }
                        }
                }
                return(list(min=minCor, point=point))
}
#------k-means--------------
k_means<-function(data){
   trainData<-data
   km<-list()
   disCutp<list()
   for(i in seq(ncol(data))){
           km[[i]] <- kmeans( trainData[,c(i)], 3)
   }
   for(i in seq(ncol(data))){
           data[,i] <- km[[i]]$cluster
           disCutp[[i]]<-km[[i]]$centers
   }
   new.data<-cFactor(data)
   return(list(cutp=disCutp,Disc.data=new.data))
}
#------Discretization--------------
discretize<-function(data,class=ncol(data)){
           cm.data<-cutMissing(data)
           if(check_type(cm.data)){
                   library("discretization")
                   class<- min_cor(data)
```

```
                num.data<- extract_numeric(cm.data,class$point)
                disc<-chiM(num.data$data,alpha=0.05)
                chiData<-disc$Disc.data
                disCutp<-disc$cutp
                new.data<-replace_data(cm.data,chiData,num.data$ls)
        } else {
                disc<-k_means(data)
                new.data<- disc$Disc.data
                cutp<- disc$ cutp
        }
        new.data<-cFactor(new.data)
        return(list(cutp=disCutp,Disc.data=new.data))
  }
#------As factor--------------
cFactor<- function(data){
   for(i in seq(ncol(data))){
        data[[i]]<-as.factor(data[[i]])
   }
   return(data)
}
#------association--------------
asso<-function(data,support=0.3){
        library("arules")
        tr <- as(data, "transactions")
        rules <- apriori(tr, parameter= list(supp=support))
        quality(rules) <- cbind(quality(rules), coverage = interestMeasure(rules, method =
"coverage", tr))
        WRITE(rules, file = "rule.csv", quote=TRUE, sep = ",", col.names = NA)
        inspect(rules)
}
```

ภาคผนวก ข

บทความวิจัยที่ได้รับการตีพิมพ์เผยแพร่

รายชื่อบทความวิจัยที่ได้รับการตีพิมพ์เผยแพร่

นันทวุฒิ คะอังกุ, กิตติศักดิ์ เกิดประสพ, นิตยา เกิดประสพ. 2555. **การแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการวิเคราะห์กฎความสัมพันธ์.** ในงานประชุมวิชาการระดับชาติเพื่อการพัฒนาด้านวิจัยอย่างยั่งยืน. มหาวิทยาลัยศรีนครินทรวิโรฒ. 25 - 26 ธันวาคม 2555

Nuntawut Kaoungku, Phatcharawan Chinthaisong, Kittisak Kerdprasop, Nittaya Kerdprasop. 2013. **Discretization and Imputation Techniques for Quantitative Data Mining.** The 2013 IAENG International Conference on Data Mining and Applications. Hong Kong. 13 - 15 March 2556

การแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการวิเคราะห์กฎความสัมพันธ์

A DISCRETIZATION METHODS FOR NUMERIC ATTRIBUTES IN ASSOCIATION RULE ANALYSIS

นันทวุฒิ คะอังกุ[*], กิตติศักดิ์ เกิดประสพ, นิตยา เกิดประสพ

Nuntawut Kaoungku[*], Kittisak Kerdprasop, Nittaya Kerdprasop

สาขาวิชาวิศวกรรมคอมพิวเตอร์ สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี

Department of Computer Engineering, Faculty of Engineering, Suranaree University of Technology, Thailand.

[*]**Corresponding author, E-mail:** b5111299@gmail.com

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการวิเคราะห์กฎความสัมพันธ์ การแบ่งช่วงข้อมูลที่เป็นตัวเลขนั้นมีหลากหลายวิธี แต่ในส่วนมากมักจะนำไปใช้ในการทำเหมืองข้อมูลประเภทอื่น ดังนั้นผู้วิจัยจึงทำการศึกษาการแบ่งช่วงข้อมูลที่เป็นตัวเลขสำหรับการทำเหมืองข้อมูลประเภทการหากฎความสัมพันธ์ โดยการวิเคราะห์ผลจะเป็นการเปรียบเทียบประสิทธิภาพการแบ่งช่วงข้อมูลที่เป็นตัวเลขสำหรับการทำเหมืองข้อมูลประเภทการหากฎความสัมพันธ์ระหว่างอัลกอริทึม Chi2 และ อัลกอริทึมแบบ Top-down โดยจะวัดจากมาตรวัด 4 มาตรวัด คือ ค่าสนับสนุน ค่าความเชื่อมั่น ค่าลิฟต์ และค่าความครอบคลุม

คำสำคัญ การทำเหมืองข้อมูล การแบ่งช่วงข้อมูล การวิเคราะห์กฎความสัมพันธ์ ภาษา R

Abstract

This research aims at studying the discretization methods for numeric attributes in association rule analysis with R language. There exist many discretization methods for numeric attributes, but they are most often used in other data mining tasks. We thus study the discretization methods for numeric attributes in association rule mining. We comparatively experiment with two discretization methods for grouping numeric attributes. The two methods are Chi2 and Top-down algorithm. The discovered association rules are then analyzed with the four measurements, that is, support, confidence, lift, and coverage.

Keywords: Data mining, Discretization, Association rule analysis, R language

บทนำ

ในปัจจุบันการที่จะนำข้อมูลมาทำการวิเคราะห์นั้นทำได้ง่ายมากกว่าสมัยก่อนที่ข้อมูลอยู่ในรูปแบบของการจดบันทึกลงกระดาษ ซึ่งเป็นการยากที่จะนำข้อมูลเหล่านั้นมาทำการวิเคราะห์และเสี่ยงต่อข้อมูลสูญหาย แต่ในยุคปัจจุบันนั้นมีขนาดของข้อมูลที่ใหญ่ขึ้นจนทำให้ถูกเก็บอยู่ในรูปแบบของข้อมูลดิจิตอลเพื่อช่วยลดความเสี่ยงกับ

ข้อมูลสูญหายในการเก็บของข้อมูลแบบเดิม จากข้อมูลที่อยู่ในรูปแบบดิจิตอลนี้นี่เองทำให้สามารถนำข้อมูลมาผ่านกระบวนการทางคอมพิวเตอร์เพื่อช่วยในการวิเคราะห์ข้อมูลซึ่งก็คือการทำเหมืองข้อมูล (Data mining) ที่ได้มีการนำไปใช้ในหลากหลายด้าน ตัวอย่างเช่น นำไปใช้งานทางด้านการค้าช่วยเพิ่มยอดขายให้กับสินค้า นำไปใช้งานทางด้านการศึกษาช่วยวิเคราะห์พฤติกรรมการเรียนรู้ของผู้เรียน เป็นต้น

การหากฎความสัมพันธ์ (Association Rules Mining) เป็นการทำเหมืองข้อมูลประเภทหนึ่ง ซึ่งจะเป็นการหาความสัมพันธ์ของเหตุการณ์หรือวัตถุ แล้วนำมาสร้างกฎเพื่อที่จะทำนายเหตุการณ์หรือการเกิดขึ้นของวัตถุนั้น ๆ ในอนาคต โดยการหากฎความสัมพันธ์นั้นสามารถนำไปใช้งานได้หลายหลายรูปแบบ [2,7]

ข้อมูลที่มีอยู่ในปัจจุบันนั้นมีหลากหลายรูปแบบ เช่น ข้อมูลที่เป็นตัวเลข ข้อมูลที่เป็นข้อความ ข้อมูลที่เป็นตัวเลขและข้อความ แต่การทำเหมืองข้อมูลในบางลักษณะ ไม่สามารถทำได้กับข้อมูลบางประเภท ซึ่งการหากฎความสัมพันธ์นั้นไม่สามารถที่จะหากฎความสัมพันธ์ได้จากข้อมูลที่มีลักษณะที่เป็นตัวเลขที่ต่อเนื่องได้ดีเท่าที่ควร ดังนั้นจึงได้มีวิธีการสำหรับจัดการข้อมูลที่เป็นตัวเลขต่อเนื่องเรียกว่าการแบ่งช่วงข้อมูล (Discretization) เพื่อให้ได้กฎความสัมพันธ์ที่มีจำนวนลดลงและให้ได้ความถูกต้องที่เพิ่มขึ้น ซึ่งในปัจจุบันได้มีงานวิจัยเกี่ยวกับวิธีการแบ่งช่วงข้อมูลด้วยหลากหลายวิธี [3] เช่น Chi2 เป็นอัลกอริทึมของการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องที่ใช้กันอย่างแพร่หลาย [8]

กฎความสัมพันธ์ที่ได้จากการหากฎความสัมพันธ์นั้นมีจำนวนมาก ทำให้มีความจำเป็นในการที่จะหามาตรวัดมาทำการวัดประสิทธิภาพของกฎความสัมพันธ์ที่ได้จากการหากฎความสัมพันธ์ เพื่อต้องการที่จะเลือกกฎความสัมพันธ์ที่มีประโยชน์มาใช้ในการทำนาย ซึ่งมาตรวัดที่จะนำมาใช้วัดประสิทธิภาพนั้นมีหลากหลายมาตรวัด [8] เช่น ค่าสนับสนุน (Support) เป็นมาตรวัดที่ใช้วัดความถี่ของเหตุการณ์ที่เกิดขึ้นว่ามีมากน้อยเพียงใด โดยนับความถี่ของความสัมพันธ์ที่เกิดขึ้นภายในชุดข้อมูลนั้น ค่าความเชื่อมั่น (Confidence) เป็นการดูความถี่ของเหตุการณ์ที่เกิดขึ้น ร่วมกับเหตุการณ์อื่น ๆ ที่เกิดขึ้นร่วมกัน เป็นต้น [7]

จากที่กล่าวมาข้างต้นจะเห็นได้ว่าการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องนั้นมีความสำคัญกับการทำเหมืองข้อมูลประเภทการหากฎความสัมพันธ์ โดยมีงานวิจัยที่ต้องการจะแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องแล้วนำไปทดสอบกันการทำเหมืองข้อมูลประเภทอื่น ๆ เช่น การจำแนกข้อมูล (Classification) ที่เมื่อทำออกมาแล้วสามารถวัดประสิทธิภาพได้จากความค่าถูกต้อง (Accuracy) แต่การแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการทำเหมืองข้อมูลประเภทการหากฎความสัมพันธ์จะใช้มาตรวัดที่ต่างออกไปในการวัดประสิทธิภาพในแต่ละอัลกอริทึมของการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่อง

การวิจัยนี้จึงสนใจการเปรียบเทียบประสิทธภาพการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการวิเคราะห์กฎความสัมพันธ์ ระหว่างอัลกอริทึม Chi2 และ อัลกอริทึมแบบ Top-down โดยจะวัดประสิทธิภาพของกฎความสัมพันธ์แต่ละกฎด้วยมาตรวัด 4 มาตรวัด ซึ่งได้แก่ Support, Confidence, Lift และ Coverage

วัตถุประสงค์ของการวิจัย

การวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการวิเคราะห์กฎความสัมพันธ์ โดยมุ่งเน้นการวัดประสิทธิภาพอัลกอริทึมในการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการทำเหมืองข้อมูลประเภทการหากฎความสัมพันธ์โดยใช้มาตรวัด 4 มาตรวัดในการวัดประสิทธิภาพและนำค่าที่ได้จากมาตรวัดมาสรุปผลของแต่ละอัลกอริทึมของการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่อง

วิธีดำเนินการวิจัย

งานวิจัยนี้เป็นการศึกษาการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการวิเคราะห์กฎความสัมพันธ์ ซึ่งจะแบ่งส่วนของการดำเนินงานออกเป็น 4 ส่วน คือ ศึกษาการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่อง, ศึกษาภาษา R, ศึกษามาตรวัด (Measure) และวัดประสิทธิภาพการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการหากฎความสัมพันธ์ ระหว่างอัลกอริทึม Chi2 และ อัลกอริทึมแบบ Top-down

1) ศึกษาการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่อง

1.1) อัลกอริทึม Chi2

อัลกอริทึม Chi2 มีพื้นฐานมาจาก X^2 ที่นิยมใช้ในงานทางด้านสถิติถูกนำมาใช้ในการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่อง [4] ซึ่งสมการในการหา X^2 มีดังนี้

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}}$$

โดยกำหนดให้

k	แทนจำนวนของคลาส
A_{ij}	แทนจำนวน pattern โดย *i* คือช่วง, *j* คือคลาส
R_i	แทนจำนวน pattern โดยช่วง *i*th = $\sum_{j=1}^{k} A_{ij}$
C_j	แทนจำนวนคลาส โดยช่วง *j*th = $\sum_{i=1}^{2} A_{ij}$
N	แทนจำนวน pattern ทั้งหมด = $\sum_{i=1}^{2} R_i$
Eij	แทนความถี่จาก $A_{ij} = R_i * C_j / N$

โดยอัลกอริทึม Chi2 จะแบ่งออกเป็นสองส่วนด้วยกัน ส่วนที่ 1 จะเริ่มต้นด้วยระดับนัยสำคัญที่สูง คือ 0.5 (sigLevel = 0.5) สำหรับข้อมูลทุกตัวที่เป็นตัวเลขต่อเนื่อง หลังจากนั้นจะทำการเรียงข้อมูลทุกตัวที่เป็นตัวเลขต่อเนื่อง แล้วจะทำการดำเนินการต่อไปนี้

- คำนวณหาค่า X^2 ตามสมการ
- รวมคู่ของช่วงที่ติดกันกับค่า X^2 ที่ต่ำ

ส่วนที่ 2 จะเป็นในส่วนปลีกย่อยของส่วนที่ 1 เริ่มจาก sigLevel0 ที่ได้กำหนดไว้ในส่วนที่ 1 แล้วทำการตรวจสอบความสอดคล้องหลังจากดำเนินการรวมแต่ละแอตทริบิวต์ ถ้า inconsistency rate ไม่เกิน ก็ทำการกำหนด sigLevel[i] สำหรับการรวมแอตทริบิวต์ในรอบถัดไป ซึ่งกระบวนการนี้จะหยุดก็ต่อเมื่อไม่มีค่าในแอตทริบิวต์

1.2) อัลกอริทึม Top-down

อัลกอริทึมแบบ Top-down [5] ในงานวิจัยนี้สนใจศึกษาในส่วนของอัลกอริทึม CAIM (Class-Attribute Interdependence Maximization) ซึ่งอัลกอริทึมนี้เป็นแบบ Supervised โดยจุดประสงค์ของอัลกอริทึม

CAIM คือ ทำการเพิ่มการพึ่งพาอาศัยกันของคลาสแอตทริบิวต์และสร้างช่วงที่ได้จากการแบ่งช่วงให้น้อยที่สุด สมการในการหาค่า CAIM มีดังนี้

$$CAIM(C,D \mid F) = \frac{\sum_{r=1}^{n} \frac{\max_r^2}{M_{+r}}}{n}$$

โดยกำหนดให้

C	แทนคลาส
D	แทนการแบ่งช่วงข้อมูล
F	แทนแอตทริบิวต์
n	แทนจำนวนช่วงข้อมูล
$\max_r$	แทนค่าที่มากที่สุดในค่า q_{ir}
M_{+r}	แทนจำนวนทั้งหมดของข้อมูลที่เป็นตัวเลขต่อเนื่องจากแอตทริบิวต์

กระบวนการทำงานของอัลกอริทึม CAIM จะแบ่งออกเป็น 2 ส่วนดังนี้

ขั้นที่ 1

1.1 หาค่าที่มากที่สุด(d_n) และน้อยที่สุด (d_0) ในข้อมูลที่เป็นตัวเลขต่อเนื่อง

1.2 กำหนดรูปแบบของค่าที่แตกต่างจากข้อมูลที่เป็นตัวเลขต่อเนื่องโดยเรียงลำดับและกำหนดขอบเขตของช่วงทั้งหมดที่เป็นไปได้แทนด้วย B ตามค่าที่ต่ำสุดและค่าที่มากสุดและจุดกึ่งกลางทั้งหมดของทุกคู่ที่อยู่ติดกันในชุด

1.3 กำหนดแบบการแบ่งช่วงข้อมูล D:{[$d_0 d_n$]} และ GlobalCAIM=0

ขั้นที่ 2

2.1 กำหนด k=1

2.2 ลองเพิ่มขอบเขตภายในที่ไม่ได้อยู่ใน D และ B แล้วคำนวณค่าที่สอดคล้องกับ CAIM

2.3 หลังจากขั้นตอนก่อนหน้านี้รับค่า CAIM ที่มากที่สุด

2.4 if (CAIM > GlobalCAIM or k<S) แล้วปรับปรุง D ที่จากค่าที่ได้รับจาก 2.3 และกำหนด GlobalCAIM=CAIM ถ้าไม่เช่นนั้นก็หยุด

2.5 กำหนด k=k+1 แล้วกลับไปที่ 2.2

2) ศึกษาภาษา R

ภาษา R เป็นภาษาเชิงฟังก์ชัน (Function language) ที่มีลักษณะคล้ายภาษา S ซึ่งเป็นภาษาที่นิยมใช้กันอย่างแพร่หลาย เนื่องจากเป็น Open source สามารถนำเข้าชุดข้อมูลได้หลากหลายรูปแบบ เช่น text file หรือ Database และสามารถแสดงผลออกมาในรูปแบบของกราฟฟิก เหมาะสำหรับการทำงานทางด้านสถิติ แต่เนื่องจากภาษา R เป็นภาษาเชิงฟังก์ชันและเชิงวัตถุ ทำให้ผู้ที่จะใช้ภาษา R นั้นต้องใช้เวลาในการศึกษาและทำ

ความเข้าใจภาษา R [1] ในภาษา R มีฟังก์ชันในส่วนของการทำเหมืองข้อมูล ซึ่งในงานวิจัยนี้จะศึกษาในส่วนของการแบ่งช่วงข้อมูลที่เป็นตัวเลขสำหรับการทำเหมืองข้อมูลประเภทการหากฎความสัมพันธ์ โดยจะมีการเรียกใช้คำสั่งสำคัญ ๆ ดังนี้

- new.dataset<-chi2(iris,0.5,0.05)$Disc.data เป็นคำสั่งในการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องโดยใช้อัลกอริทึม Chi2
- new.dataset<-disc.Topdown(iris, method=1)$Disc.data เป็นคำสั่งในการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องโดยใช้อัลกอริทึมแบบ Top-down
- rules <- apriori(tr, parameter= list(supp=0.3, conf=0.5)) เป็นคำสั่งในการหากฎความสัมพันธ์ (Association Rules Mining) ด้วยอัลกอริทึม Apriori
- interestMeasure(rules, c("support", "confidence", "lift", "coverage"), **tr)** เป็นคำสั่งในการเรียกใช้มาตรวัดต่าง ๆ ในการวัดประสิทธิภาพของกฎความสัมพันธ์ที่ได้จากการหากฎความสัมพันธ์

3) ศึกษามาตรวัด (Measure)

3.1) Support

มาตรวัดที่ใช้วัดความถี่ของเหตุการณที่เกิดขึ้นว่ามีมากน้อยเพียงใด โดยนับความถี่ความสัมพันธ์ที่เกิดขึ้นภายในชุดข้อมูลนั้น เรียกว่า ค่าสนับสนุน (Support) โดยมีสมการในการคำนวณดังนี้

$$Support(A \rightarrow B) = P(A \wedge B)$$

3.2) Confidence

ค่าความเชื่อมั่น (Confidence) เป็นการดูความถี่ของเหตุการณ์ที่เกิดขึ้น ร่วมกับเหตุการณ์ที่เกิดอื่น ๆ ที่เกิดขึ้นร่วมกัน เช่น เมื่อเกิดเหตุการณ์ A แล้วบ่อยแค่ไหนที่จะเกิดเหตุการณ์ B ซึ่งสามารถคำนวณค่าความเชื่อมั่นได้จากอัตราส่วนของ

$$Confidence\ (A \rightarrow B) = \frac{\text{ความถี่ของ } A \text{ และ } B}{\text{ความถี่ของ } A}$$

3.3) Lift

Lift จะเป็นมาตรวัดที่ใช้วัดประสิทธิภาพกฎความสัมพันธ์ โดยจะวัดอิทธิพลของกฎความสัมพันธ์ที่เกิดขึ้น ในการคำนวณหาค่า Lift นั้นสามารถหาได้จากอัตราส่วนของ

$$Lift\ (A \rightarrow B) = \frac{\text{ความเชื่อมั่นของ } A \rightarrow B}{\text{ความถี่ของ } B}$$

ค่า Lift ที่ได้ออกมานั้นจะสามารถบอกความเป็นไปได้ ถ้าเกิดเหตุการณ์ A แล้วจะเกิดเหตุการณ์ B ซึ่งทั้งสองเหตุการณ์จะขึ้นต่อกัน โดยถ้าได้ค่า Lift ที่ได้ออกมานั้นมากหมายถึงกฎความสัมพันธ์ที่ได้นั้นมีความสำคัญมากพอที่จะนำไปใช้ในการทำนาย

4) วัดประสิทธิภาพการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องสำหรับการหากฎความสัมพันธ์ ระหว่างอัลกอริทึม Chi2 และ อัลกอริทึมแบบ Top-down

การแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่อง ผู้วิจัยได้เลือกใช้อัลกอริทึม Chi2 และ อัลกอริทึมแบบ Top-down ในการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่อง แล้วนำข้อมูลที่แบ่งช่วงข้อมูลแล้วไปผ่านกระบวนการหากฎความสัมพันธ์ ซึ่งจะใช้ข้อมูล Iris จาก UCI Machine Learning Repository ข้อมูลนี้มี จำนวน 150 เรคคอร์ด จากตารางที่ 1 เป็นตัวอย่างข้อมูล Iris จำนวน 5 เรคคอร์ด โดยจะใช้มาตรวัดในการวัดประสิทธิภาพทั้งหมด 4 มาตรวัด ได้แก่ Support, Confidence, Lift และ Coverage ในงานวิจัยนี้

ตารางที่ 1 ตัวอย่างข้อมูล Iris

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.3	3.3	6.0	2.5	virginica

เมื่อนำข้อมูลที่เป็นตัวเลขต่อเนื่องไปผ่านกระบวนการแบ่งช่วงข้อมูล ผลลัพธ์ที่ได้ออกมาก็คือช่วงของข้อมูล จากภาพที่ 1 จะเป็นตัวอย่างช่วงข้อมูลที่ได้จากการแบ่งช่วงข้อมูลด้วยอัลกอริทึม Chi2 และ Top-down

if Sepal.Length<3.5 then Sepal.Length=SL1	if Sepal.Length>=4.30&&Sepal.Length<5.55 then Sepal.Length=SL1
if Sepal.Length>=3.5&&Sepal.Length<4.5 then Sepal.Length=SL2	if Sepal.Length>=5.55&&Sepal.Length<6.25 then Sepal.Length=SL2
if Sepal.Length>=4.5&&Sepal.Length<6.5 then Sepal.Length=SL3	if Sepal.Length>=6.25&&Sepal.Length<7.90 then Sepal.Length=SL3
if Sepal.Length>=6.5 then Sepal.Length=SL4	if Sepal.Width>=2.00&&Sepal.Width<2.95 then Sepal.Length=SW1
if Sepal.Width<3.5 then Sepal.Length=SW1	if Sepal.Width>=2.95&&Sepal.Width<3.05 then Sepal.Length=SW2
if Sepal.Width>=3.5&&Sepal.Width<4.5 then Sepal.Length=SW2	if Sepal.Width>=3.05&&Sepal.Width<4.40 then Sepal.Length=SW3
if Sepal.Width>=4.5 then Sepal.Length=SW3	if Petal.Length>=1.00&&Petal.Length<2.45 then Petal.Length=PL1
if Petal.Length<1.5 then Petal.Length=PL1	if Petal.Length>=2.45&&Petal.Length<4.75 then Petal.Length=PL2
if Petal.Length>=1.5&&Petal.Length<2.5 then Petal.Length=PL2	if Petal.Length>=4.75&&Petal.Length<6.90 then Petal.Length=PL3
if Petal.Length>=2.5&&Petal.Length<3.5 then Petal.Length=PL3	if Petal.Width>=0.10&&Petal.Width<0.80 then Petal.Width=PW1
if Petal.Length>=3.5 then Petal.Length=PL4	if Petal.Width>=0.80&&Petal.Width<1.75 then Petal.Width=PW2
if Petal.Width<1.5 then Petal.Width=PW1	if Petal.Width>=1.75&&Petal.Width<2.50 then Petal.Width=PW3
if Petal.Width>=1.5&&Petal.Width<3.5 then Petal.Width=PW2	
if Petal.Width>=3.5 then Petal.Width=PW3	

ภาพที่ 1 ตัวอย่างช่วงข้อมูลที่ได้จากการแบ่งช่วงข้อมูลด้วยอัลกอริทึม Chi2 (ซ้าย) และ Top-down (ขวา)

ตารางที่ 2 ตัวอย่างกฎความสัมพันธ์ที่ได้จากการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยอัลกอริทึม Chi2

Rules	Support	Confidence	Lift	Coverage
{Petal.Length=PL2} => {Petal.Width=PW2}	0.3	1	2.77	0.3
{Petal.Width=PW2} => {Petal.Length=PL2}	0.3	0.83	2.77	0.36
{Petal.Width=PW3} => {Species=virginica}	0.3	0.97	2.93	0.30
{Species=virginica} => {Petal.Width=PW3}	0.3	0.9	2.93	0.33
{Petal.Length=PL1} => {Petal.Width=PW1}	0.33	1	3	0.33

จากตารางที่ 2 เป็นตัวอย่างกฎความสัมพันธ์ที่ได้จากการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยอัลกอริทึม Chi2 ซึ่งนำมาแสดง 5 กฎ และประกอบไปด้วยมาตรวัดประสิทธิภาพของแต่ละกฎความสัมพันธ์

ตารางที่ 3 ตัวอย่างกฎความสัมพันธ์ที่ได้จากการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยอัลกอริทึม Top-down

Rules	Support	Confidence	Lift	Coverage
{Petal.Width=PW3} => {Species=virginica}	0.30	1	3	0.30
{Species=virginica} => {Petal.Width=PW3}	0.30	0.9	3	0.33
{Petal.Width=PW3} => {Petal.Length=PL3}	0.30	1	2.77	0.30
{Petal.Length=PL3} => {Petal.Width=PW3}	0.30	0.83	2.77	0.36
{Petal.Length=PL2} => {Species=versicolor}	0.30	0.97	2.93	0.31

จากตารางที่ 3 เป็นตัวอย่างกฎความสัมพันธ์ที่ได้จากการแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องด้วยอัลกอริทึม Top-down ซึ่งนำมาแสดง 5 กฎ ร่วมกับมาตรวัดประสิทธิภาพของแต่ละกฎความสัมพันธ์

ผลการวิจัย

งานวิจัยนี้ได้ทำการทดลองเปรียบเทียบประสิทธิภาพการแบ่งช่วงข้อมูลที่เป็นตัวเลขสำหรับการทำเหมืองข้อมูลประเภทการหากฎความสัมพันธ์ด้วยภาษา R ระหว่างอัลกอริทึม Chi2 และ อัลกอริทึมแบบ Top-down ซึ่งจะนำกฎความสัมพันธ์ได้ออกมานั้นมาทำการเปรียบเทียบประสิทธิภาพ จากการทดลองจะเห็นได้ว่าค่า Support กับ Coverage มีค่าไปในทิศทางเดียว และค่า Confidence กับ Lift ก็มีค่าไปในทิศทางเดียวกัน ดังนั้นผู้วิจัยจึงแยกการเปรียบเทียบประสิทธิภาพออกเป็น 2 กลุ่ม คือ Support กับ Coverage และ Confidence กับ Lift

ตารางที่ 4 ตารางแสดงผลการเปรียบเทียบค่า Support และ Coverage ของกฎความสัมพันธ์ 10 กฎแรก

	Support		Coverage	
Rules	Chi2	Top-down	Chi2	Top-down
{Species=setosa} => {Petal.Length=PL1}	0.33	0.33	0.33	0.33
{Species=setosa} => {Petal.Width=PW1}	0.33	0.33	0.33	0.33
{Petal.Width=PW2} => {Species=versicolor}	0.32	0.33	0.36	0.36

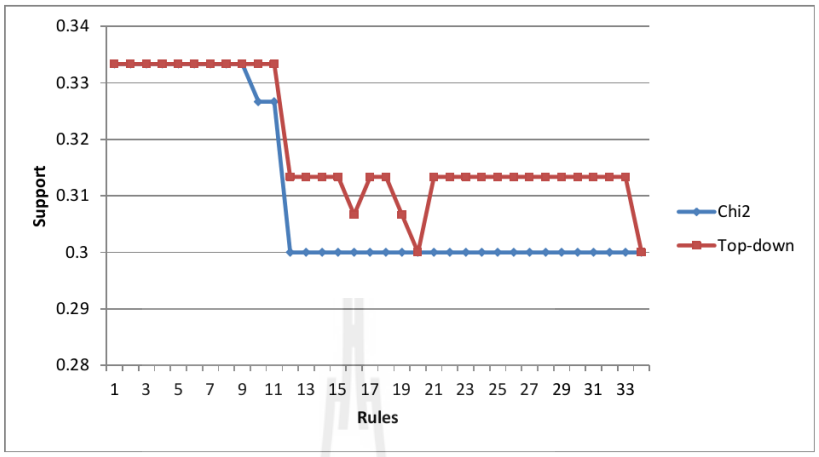
{Species=versicolor} => {Petal.Width=PW2}	0.32	0.33	0.33	0.33
{Petal.Width=PW1} => {Sepal.Length=SL1}	0.3	0.31	0.33	0.33
{Petal.Length=PL1,Petal.Width=PW1} => {Sepal.Length=SL1}	0.3	0.31	0.33	0.33
{Petal.Length=PL1,Species=setosa} => {Sepal.Length=SL1}	0.3	0.31	0.33	0.33
{Petal.Length=PL1} => {Sepal.Length=SL1}	0.3	0.31	0.33	0.33
{Petal.Length=PL2} => {Petal.Width=PW2}	0.3	0.30	0.3	0.30
{Petal.Width=PW1,Species=setosa} => {Sepal.Length=SL1}	0.3	0.31	0.33	0.33

จากตารางที่ 4 ที่แสดงผลการเปรียบเทียบค่า Support และ Coverage ของกฎความสัมพันธ์สิบกฎแรกจะเห็นได้ว่าระหว่างอัลกอริทึม Chi2 และ Top-down อัลกอริทึมแบบ Top-down มีค่าในส่วนของ Support ที่มีค่ามากกว่าอัลกอริทึม Chi2 เล็กน้อย แต่ในส่วนค่า Coverage ไม่มีความแตกต่างกัน

ตารางที่ 5 ตารางแสดงผลการเปรียบเทียบค่า Confidence และ Lift ของกฎความสัมพันธ์ 10 กฎแรก

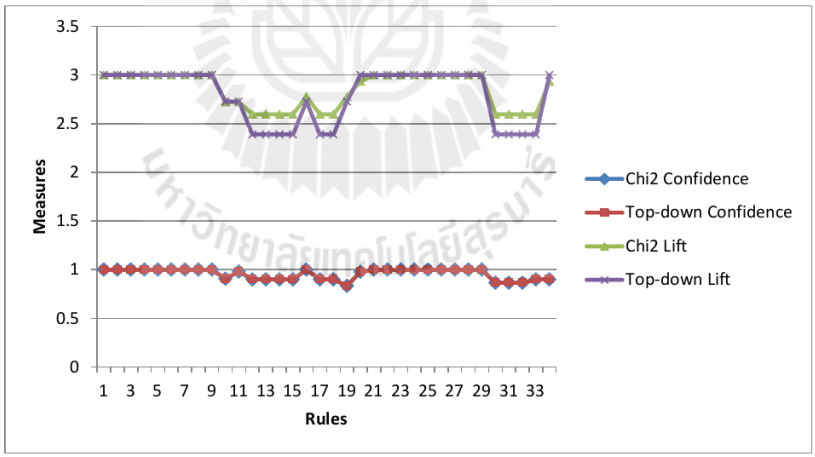
Rules	Confidence		Lift	
	Chi2	Top-down	Chi2	Top-down
{Species=setosa} => {Petal.Length=PL1}	1	1	3	3
{Species=setosa} => {Petal.Width=PW1}	1	1	3	3
{Petal.Width=PW2} => {Species=versicolor}	0.90	0.90	2.7	2.72
{Species=versicolor} => {Petal.Width=PW2}	0.98	1	2.7	2.72
{Petal.Width=PW1} => {Sepal.Length=SL1}	0.9	0.94	2.59	2.38
{Petal.Length=PL1,Petal.Width=PW1} => {Sepal.Length=SL1}	0.9	0.94	2.59	2.38
{Petal.Length=PL1,Species=setosa} => {Sepal.Length=SL1}	0.9	0.94	2.59	2.38
{Petal.Length=PL1} => {Sepal.Length=SL1}	0.9	0.94	2.59	2.38
{Petal.Length=PL2} => {Petal.Width=PW2}	1	1	2.77	2.72
{Petal.Width=PW1,Species=setosa} => {Sepal.Length=SL1}	0.9	0.94	2.59	2.38

จากตารางที่ 5 ที่แสดงผลการเปรียบเทียบค่า Confidence และ Lift ของกฎความสัมพันธ์สิบกฎแรกจะเห็นได้ว่าระหว่างอัลกอริทึม Chi2 และ Top-down อัลกอริทึมแบบ Top-down มีค่าในส่วนของ Confidence ที่มีค่ามากกว่าอัลกอริทึม Chi2 แต่ในส่วนของค่า Lift อัลกอริทึม Chi2 ให้ค่าที่มากกว่าอัลกอริทึมแบบ Top-down ซึ่งจะเห็นได้อย่างชัดเจน



ภาพที่ 2 กราฟแสดงผลการเปรียบเทียบความแตกต่างของค่า Support ในแต่ละกฎความสัมพันธ์ระหว่างอัลกอริทึม Chi2 และ Top-down

จากภาพที่ 2 เป็นกราฟแสดงผลการเปรียบเทียบความแตกต่างของค่า Support ในแต่ละกฎความสัมพันธ์ระหว่างอัลกอริทึม Chi2 และ Top-down ซึ่งจะเห็นได้ว่าค่า Support มีความแตกต่างกันน้อย โดยอัลกอริทึมแบบ Top-down จะให้ค่าที่มากกว่าเล็กน้อย



ภาพที่ 3 กราฟแสดงผลการเปรียบเทียบความแตกต่างของค่า Confidence และ Lift ในแต่ละกฎความสัมพันธ์ระหว่างอัลกอริทึม Chi2 และ Top-down

จากภาพที่ 3 เป็นกราฟแสดงผลการเปรียบเทียบความแตกต่างของค่า Confidence และ Lift ในแต่ละกฎความสัมพันธ์ระหว่างอัลกอริทึม Chi2 และ Top-down ซึ่งจะเห็นได้ว่าค่า Confidence มีความแตกต่างกันเล็กน้อย โดยอัลกอริทึมแบบ Top-down จะให้ค่าที่มากกว่าเล็กน้อย แต่ค่า Lift จะเห็นได้อย่างชัดเจนว่าอัลกอริทึม Chi2 นั้นให้ค่า Lift ที่มากกว่าอัลกอริทึมแบบ Top-down

สรุปและอภิปรายผล

การแบ่งช่วงข้อมูลที่เป็นตัวเลขต่อเนื่องนั้นมีความสำคัญในการหากฎความสัมพันธ์ เนื่องจากข้อมูลที่เป็นตัวเลขต่อเนื่องที่ไม่ได้ทำการแบ่งช่วงข้อมูลเมื่อนำไปหากฎความสัมพันธ์จะให้ได้กฎความสัมพันธ์ที่มากและไม่มีความสำคัญพอที่จะนำไปทำนายได้ แต่ข้อมูลที่เป็นตัวเลขต่อเนื่องที่ได้ทำการแบ่งช่วงข้อมูลเมื่อนำไปหากฎความสัมพันธ์จะทำให้ได้กฎความสัมพันธ์ที่น้อยลงและมีความสำคัญที่จะนำไปทำนายได้

จากการทดลองดังกล่าวจะเห็นได้ว่ามาตรวัดที่นำมาใช้วัดประสิทธิภาพการแบ่งช่วงข้อมูลที่เป็นตัวเลขสำหรับการทำเหมืองข้อมูลประเภทการหากฎความสัมพันธ์นั้น ค่า Support กับ Coverage มีค่าไปในทิศทางเดียวกัน และค่า Confidence กับ Lift ก็มีค่าไปในทิศทางเดียวกัน โดยในอัลกอริทึม Top-down ให้ค่า Support, Coverage และ Confidence ที่มากกว่าอัลกอริทึม Chi2 เล็กน้อย แต่ในส่วนของค่า Lift อัลกอริทึม Chi2 ให้ค่าที่มากกว่าอัลกอริทึม Top-down อย่างชัดเจน

เอกสารอ้างอิง

[1] รองศาสตราจารย์ ดร. กิตติศักดิ์ เกิดประสพ (2012). Data Mining With R. Retrieved October 22, 2012, from https://sites.google.com/site/kittisakthailand55/home/datamining2-55

[2] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami (1993). 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

[3] Huan Liu, Farhad Hussain, Chew Lim Tan, and Manoranjan Dash (2002). Discretization: An Enabling Technique. *Data Mining and Knowledge Discovery*, vol.6, no.4, pp. 393-423

[4] Huan Liu and Rudy Setiono (1995). Chi2: Feature Selection and Discretization of Numeric Attributes. *Proceedings of the Seventh International Conference on Tools with Artificial Intelligence*, pp. 372-390

[5] Kurgan, L. A. and Cios, K. J. (2004). CAIM Discretization Algorithm, *IEEE Transactions on knowledge and data engineering,* vol.16,no.2 , pp.145-153

[6] Chao-Ton Su and Jyh-Hwa Hsu (2005). An Extended Chi2 Algorithm for Discretization of Real Value Attributes. *IEEE Transactions on Knowledge and Data Engineering*, vol.17, no.3, pp. 437-441

[7] Wikipedia, The Free Encyclopedia (2012). Association rule learning. Retrieved October 22, 2012, from http://en.wikipedia.org/wiki/Association_rule_learning

[8] Zaki M.J. (2002). Scalable Algorithms for Association Mining. *IEEE Transaction on Knowledge and Data Engineering*, vol.12, no.3, pp. 372-390

Discretization and Imputation Techniques for Quantitative Data Mining

Nuntawut Kaoungku*, Phatcharawan Chinthaisong, Kittisak Kerdprasop, and Nittaya Kerdprasop

***Abstract*—Association rule mining from numeric datasets are known inefficient because number of discovered rules is superfluous and sometimes inapplicable. In this paper, we propose the discretization technique based on the Chi2 algorithm to categorize numeric values. We also handle missing values in the dataset with statistical methods. The discovered association rules are then evaluated with the four measurement metrics, that is, confidence, support, lift, and coverage. The dataset imputes with various missing value handling techniques has also been evaluated with data classification method to compare predictive accuracy.**

***Index Terms*—Association rule analysis, Data mining, Discretization, Missing value imputation**

I. Introduction

Current adoption of data mining technology can be seen in various fields such as economics, education, and medical. The model learning from datasets can facilitate future event prediction and explain current relations. Model built from datasets of data missing can cause errors in the prediction. Efficiency predictive model building requires the imputation of missing data.

Association rule mining is a type of data mining that will find the association of the objects and create a rule to predict the occurrence of the object in the future.

The decision tree is data mining classification. The rules or terms of learning from datasets with the data type specified already exist. The modeling helps to decide classification data that will occur in the future.

The discretization methods for numeric attributes and missing value imputation are important to the task of data mining for association rules. The association rule mining will use the gauge that differs from the efficiency measurement in each of the algorithms of discretization methods for numeric attributes.

Manuscript received December 8, 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.

N. Kaoungku is a master student with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand (e-mail: b5111299@gmail.com).

P. Chinthaisong is a master student with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand (e-mail: killuakaara@gmail.com).

K. Kerdprasop is an associate professor with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand.

N. Kerdprasop is an associate professor with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand.

This research solves the problem by preparing dataset appropriately before association and classification of discretization methods for numeric in association rule and predicts of data missing that is closest to the most possible value.

II. Preliminaries and Related Work

Data can be variety of formats. For example, numeric data, nominal data, and a mixed type of numeric and nominal data. But data mining in some categories is not possible. For instance, to find the association rules from dataset with numeric attributes is impossible for some algorithms. Therefore, methods for managing numeric attribute data is essential. The called a classified data (Discretization) to obtain the association rules to amount of reduction, and increased accuracy. Current research is on how to divide the discretization in a variety of ways. [3] For example, Chi 2 algorithm of discretization methods for numeric attributes that is widely used. [7] The discretization methods for numeric attributes in association rule analysis with R language [3]. The algorithms used to discretization are, The Chi2 algorithm formed by x^2 they are often used in statistics and discretization methods for numeric attributes.

The predictive value of the data missing is another important problem. We comparatively study the value of data missing technique, lost out in praise. The average value in that column if data missing is disrupted data or skew data. We are used median value. If the data in column aren't numeric. We used value that appears most often in the column. And how to use the value of the association between a column that has a data missing value with another column that is associated with the most. Other research also has using Rough Set theory: [5] Include: is used to determine the association between each column is set to create a rule that allows predicting. Datasets were used in this study was a series of patients that most data is dispersed across numeric data. With the numerical data will be grouped into ranges (Discretized) are so easy to do the research to find the value of the data missing. And Jianhua's research [1] propose a technique to fill up the missing data by using Rough Sets theoretical and add a technique to compare the 3 methods: how to cut data rows that contain data values that are missing out, and data mining. How to select values that are come to missing data from data that contains values that appear most frequently in the column, and how to convert the entire datasets as a Discernibility matrix and create a rule for predicting the missing value. By using a series of six sets of data to compare efficiency, how to find the value of the

data that is missing all three methods and data sets through the technique of value for the information that is missing, and the range, and then create a decision tree to test the data prepared for the test of efficiency technique to predict the best. The rules for an association with the four measurements the effectiveness and value of the gauge is a summary of each of the algorithm for discretization methods for numeric attributes.

III. METHODOLOGY

A. Framework

This research proposed discretization and imputation techniques for quantitative data mining. Figure 1 shows conceptual framework of the research. First, the missing value imputation has been applied. Second, the discretization has been performed on numeric attributes. Third, apply the association rule mining. Finally, the benchmarks on association rule mining result are to be evaluated.

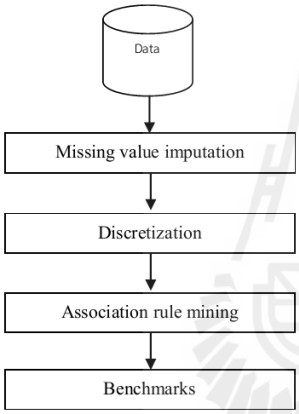


Fig. 1 Conceptual framework of the research

B. Predict the missing value

Techniques to handle missing values in our study are as follows:

1) Remove record that some values are missing.
2) Impute missing values with the average value of the attribute, if the data is normally distributed.
3) Use the correlation of column with missing values to another column, and impute with that column's value.

C. Algorithm Chi2

Chi2 algorithm that is based on the x^2 statistics was used to perform discretization the numerical data [4]. The computation for x2 is as follows.

$$x^2 = \sum_{i=1}^{2}\sum_{j=1}^{k}\left(A_{ij} - E_{ij}\right)^2 / E_{ij} \qquad (1)$$

where:

k = number of classes,

A_{ij} = number of patterns in the ith interval, jth class,

Eij = expected frequency of $A_{ij} = R_i * C_j / N$

R_i = number of patterns in the ith interval = $\sum_{j=1}^{k} A_{ij}$

C_j = number of patterns in the ith class = $\sum_{i=1}^{2} A_{ij}$

N = total number of patterns = $\sum_{i=1}^{2} R_i$

The Chi2 algorithm is divided into two parts. The first part starts with a high level of significance, that is 0.5 (sigLevel = 0.5), for all numerical data. After that, it will sort all the numbers continuously.

Part 2 will be on the sideline of the first start of sigLevel0 as set forth in Part 1, then the consistency check after performing an individual attribute the inconsistency rate cannot exceed the assigned sigLevel [i] for inclusion attributes in the next round. This process stops when there is no value left in the attribute.

D. Benchmarks

The benchmarks in this study are the four measurements: support, confidence, lift, and coverage.

1) Support is the frequency of the event occurring, Compute support of equation (2).

$$Support(A \rightarrow B) = P(A \wedge B) \qquad (2)$$

2) Confidence is the frequency of the incident with other events occurring together, Compute confidence of equation (3).

$$Confidence(A \rightarrow B) = Supp(A \rightarrow B) / Supp(A) \qquad (3)$$

3) Lift is the influence of the association rule mining, Compute lift of equation (4).

$$Lift(A \rightarrow B) = Conf(A \rightarrow B) / Supp(A) \qquad (4)$$

4) Coverage is considered the frequency of the association rules mining, Compute coverage of equation (5).

$$Coverage(A \rightarrow B) = Supp(A) = P(A) \qquad (5)$$

IV. EXPERIMENTAL RESULTS

This research experimentation used Hepatitis dataset from the UCI Machine Learning Repository [7]. Hepatitis dataset has 20 attributes and 103 data instances.

For discretization and imputation techniques for quantitative data mining, we used classification and association mining for experimental result assessment. Table 1 and Fig.2 show comparative accuracy of classification both algorithm missing value and missing value + discretization of three models. Model 1 is removing records that contain missing values. Model 2 is missing value imputation with the attribute mean. Model 3 is missing value imputation with correlated value.

TABLE 1
COMPARATIVE RESULTS OF CLASSIFICATION ACCURACY

Algorithm	Model 1	Model 2	Model 3
Missing value	65.95%	74.46%	80.85%
Missing value + Discretize	85.13%	89.36%	87.23%

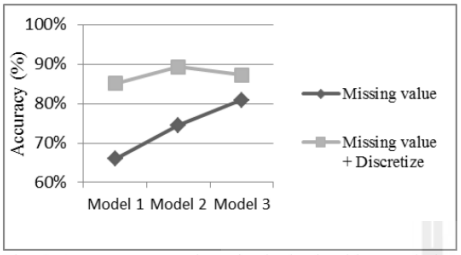


Fig. 2 Accuracy comparison for both algorithms: missing value and missing value + discretization

Table 2 show comparative results of association rule mining using the average of support, confidence, lift, and coverage values to measure performance.

TABLE 2
COMPARATIVE RESULTS OF ASSOCIATION RULE MINING

Models	The average of support	The average of confidence	The average of lift	The average of coverage
Model 1	60.99%	97.66%	103.63%	62.65%
Model 2	62.56%	98.37%	102.94%	63.78%
Model 3	62.02%	98.33%	103.07%	63.27%

Fig. 3 compares the average of confidence and lift for three models. It can be seen from the result that model 3 is the highest compared to the other models.

Fig. 4 compares the average of support and coverage values for three models. It can be seen from the result that model 2 is the highest compared to the other models.

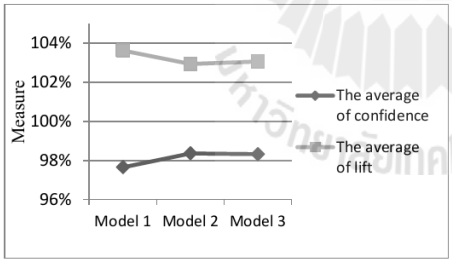


Fig. 3 Comparative the average of confidence and lift both three models

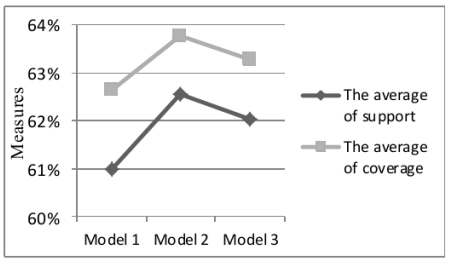


Fig. 4 Comparative the average of support and coverage both three models

V. CONCLUSION

This research aims to study discretization and imputation techniques for quantitative data mining. The results show that the best model of classification is model 2 that used missing value imputation with the average value if the data is normally distributed and used chi2 for discretization. The results also show that the best model of association rule mining is model 2. Therefore, it can be concluded that the model 2 that imputes missing values by attributes means gives the best result.

REFERENCES

[1] Jianhua Dai, Qing Xu, Wentao Wang (2011), "A Comparative Study on Strategies of Rule Induction for Incomplete Data Based on Rough Set Approach," *International Journal*, vol 3, no. 3

[2] Kittisak Kerdprasop (2012). "Data Mining Methodology and Development," Retrieved November 1, 2012, from https://sites.google.com/site/kittisakthailand55/home/datamining2-55

[3] Huan Liu, Farhad Hussain, Chew Lim Tan, and Manoranjan Dash (2002). "Discretization: An Enabling Technique," *Data Mining and Knowledge Discovery*, vol.6, no.4, pp. 393-423

[4] Huan Liu and Rudy Setiono (1995). "Chi2: Feature Selection and Discretization of Numeric Attributes," *Proceedings of the Seventh International Conference on Tools with Artificial Intelligence*, pp. 372-390

[5] Fulufhelo V. Nelwamondo and Tshilidzi Marwala. (2007). "Rough Sets Computations to Impute Missing Data," CoRRabs/0704.3635

[6] UC Irvine Machine Learning Repository, (1988) Hepatitis Data Set. Retrieved October 5, 2012 from http://archive.ics.uci.edu/ml/datasets/Hepatitis

[7] M.J. Zaki (2002). "Scalable Algorithms for Association Mining," *IEEE Transaction on Knowledge and Data Engineering*, vol.12, no.3, pp. 372-390

APPENDIX

Source code in R language to perform missing value imputation and discretization.

```
# Missing value imputation
  hepatitis<-read.csv("hepatitis.csv", fill = TRUE)
  hepatitis <- hepatitis [-c(62,199) , ]
  predict1<- function(dataM){
     cutMissing <-na.omit(dataM)
     return(cutMissing)
  }
```

```
predict2<- function(dataM,colM,more=F){
    if(more){
       dataM[is.na(dataM[[colM]]),colM]<-
    mean(dataM[[colM]],na.rm=T)
    }else{
       dataM[is.na(dataM[[colM]]),colM]<-
   median(dataM[[colM]],na.rm=T)
    }
    return(dataM)
}

lookCor<- function(crn){
    gg<-(cor(crn,use='complete.obs') )
    gp<-symnum(gg)
    return(gp)
}
creXY<- function(colM,dataM,NN){
    mM<-lm(colM,data=dataM)$coefficients[NN]
    mN<-mM[1][[1]]
    return(mN)
}
inputf<- function(oP){
    if( is.na(oP) ) return(NA)
    else return ( (oP+(-mY))/mX )
}
cor.input<- function(colA,colB,dataM){
    dataM[ is.na ( dataM[[colA]] ),colA ] <-
    sapply ( dataM[ is.na (dataM[[colA]]),colB],inputf)
    return(dataM)
}

dataset1<-predict1(hepatitis)
dataset2<-predict2(hepatitis,"Chla",T)
dataset2<-predict2(dataset2,"Cl",T)
dataset2<-predict2(dataset2,"PO4",F)

mX<-creXY(oPO4~PO4,hepatitis,2)
mY<-creXY(oPO4~PO4,hepatitis,1)
dataset3<-predict3("PO4","oPO4",hepatitis)
dataset3<-predict3("Chla","oPO4",dataset3)

library(rpart)

rt.a1<-rpart(a1~.,data=dataset1[,1:12])
plot(rt.a1,uniform=T,branch=1, margin=0.1, cex=0.9)
text(rt.a1,cex=0.75)

rt.a2<-rpart(a1~.,data=dataset2[,1:12])
plot(rt.a2,uniform=T,branch=1, margin=0.1, cex=0.9)
text(rt.a2,cex=0.75)

rt.a3<-rpart(a1~.,data=dataset3[,1:12])
plot(rt.a3,uniform=T,branch=1, margin=0.1, cex=0.9)
text(rt.a3,cex=0.75)

testPred <- predict(rt.a1, newdata = test.hepatitis)
print(testPred)
table(testPred, test.hepatitis$a1)
```

```
#Discretization
   hepatitisM<-read.csv("hepatitis.csv", fill = TRUE)
   new.dataset<-chi2(hepatitisM,0.5,0.05)$Disc.data

#Association rules mining
   rules <- apriori(new.dataset, parameter= list(supp=0.5,
   conf=0.8))

# Benchmarks
   quality(rules) <- cbind(quality(rules), coverage =
   interestMeasure(rules, method = "coverage", tr))
   WRITE(rules, file = "data_disc.csv", quote=TRUE, sep =
   ",", col.names = NA)
```

ประวัติผู้เขียน

นายนันทวุฒิ คะอังกุ เกิดเมื่อวันที่ 24 มีนาคม พ.ศ. 2532 ที่ อำเภอเมือง จังหวัดสกลนคร เริ่มเข้าศึกษาระดับชั้นอนุบาล 1 ถึงชั้นประถมศึกษาปีที่ 6 ที่โรงเรียนบ้านโพนงามคุรุราษฎร์วิทยา อำเภอกุดบาก จังหวัดสกลนคร จากนั้นได้เข้าศึกษาต่อในระดับมัธยมศึกษาตอนต้นและตอนปลาย ที่โรงเรียนกุดบากพัฒนาศึกษา อำเภอกุดบาก จังหวัดสกลนคร ปีการศึกษา 2551 ได้เข้าศึกษาต่อระดับปริญญาตรีในสาขาวิชาวิศวกรรมคอมพิวเตอร์ สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี และสำเร็จการศึกษาเมื่อปี พ.ศ. 2554 ภายหลังสำเร็จการศึกษาในระดับปริญญาตรี ได้เข้าศึกษาในระดับปริญญาโท สาขาวิชาวิศวกรรมคอมพิวเตอร์ สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี ในปี 2555

ในระหว่างการศึกษาได้รับความอนุเคราะห์อย่างดียิ่งจากอาจารย์ประจำวิชา Database System ได้รับความไว้วางใจให้เป็นผู้ช่วยสอนปฏิบัติการ ได้รับการตีพิมพ์เผยแพร่บทความวิชาการ ซึ่งรายละเอียดสามารถดูได้ที่ภาคผนวก ข