การค้นหาและจัดอันดับกฎความสัมพันธ์เพื่อวิเคราะห์ปัจจัยเสี่ยงที่นำไปสู่การเกิดโรค

นายพงศกร ดุรงค์ดำรงชัย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์

มหาวิทยาลัยเทคโนโลยีสุรนารี

ปีการศึกษา 2557

ASSOCIATION RULE SEARCH AND RANKING FOR THE ANALYSIS OF FACTORS CONTRIBUTING TO DISEASES

PongsakornDurongdumrongchai

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 2014

การค้นหาและจัดอันดับกฎความสัมพันธ์เพื่อวิเคราะห์ปัจจัยเสี่ยงที่นำไปสู่การเกิดโรค

มหาวิทยาลัยเทคโนโลยีสุรนารี อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

คณะกรรมการสอบวิทยานิพนธ์

_____

(รศ.ดร.กิตติศักดิ์ เกิดประสพ)
ประธานกรรมการ

_____

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

_____

(ผศ.ดร.คะชาชาญศิลป์)
กรรมการ

_____

(ศ.ดร.ชูกิจ ลิมปิจำนงค์)
รองอธิการบดีฝ่ายวิชาการและนวัตกรรม

_____

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

พงศกร ดุรงค์ดำรงชัย : การค้นหาและจัดอันดับกฎความสัมพันธ์เพื่อวิเคราะห์ปัจจัยเสี่ยงที่นำไปสู่การเกิดโรค(ASSOCIATION RULE SEARCH AND RANKING FOR THE ANALYSIS OF FACTORS CONTRIBUTING TO DISEASES)อาจารย์ที่ปรึกษา: รองศาสตราจารย์ ดร.นิตยา เกิดประสพ, 121หน้า

งานวิจัยนี้ได้ศึกษาปัญหาการค้นหากฎความสัมพันธ์กับข้อมูลทางการแพทย์ที่มีลักษณะของข้อมูลผสมระหว่างข้อมูลที่เป็นตัวเลขและข้อมูลที่เป็นข้อความ การค้นหากฎความสัมพันธ์ในงานวิจัยนี้ใช้อัลกอริทึมเอไพรเออรี อัลกอริทึมพรีดิคชันเอไพรเออรี และอัลกอริทึมเทอเชียสแต่เนื่องจากการค้นหาความสัมพันธ์แบบปกตินั้นไม่สามารถทำได้กับข้อมูลที่อยู่ในลักษณะของตัวเลขฉะนั้นงานวิจัยนี้จึงได้ใช้การแบ่งช่วงข้อมูลเข้ามาช่วยเพื่อให้สามารถทำการค้นหากฎความสัมพันธ์ได้ ในกรณีข้อมูลสูญหายหรือไม่สามารถระบุค่า ผู้วัจัยใช้วิธีลบข้อมูลทรานแซคชันที่ไม่สามารถระบุข้อมูลได้ชัดเจนทิ้งซึ่งเป็นข้อมูลส่วนน้อย การแบ่งช่วงข้อมูลและการลบข้อมูลที่ไม่สามารถระบุข้อมูลได้ชัดเจนนั้นจัดอยู่ในขั้นตอนการเตรียมข้อมูลก่อนการค้นหากฎความสัมพันธ์ ในอดีตได้มีงานวิจัยจำนวนมากที่ได้เสนอเทคนิคในการค้นหากฎความสัมพันธ์ในรูปแบบต่าง ๆ และทำการวัดประสิทธิภาพของกฎความสัมพันธ์ในแง่ของค่าความเชื่อมั่นเพียงอย่างเดียว แต่งานวิจัยในด้านการค้นหากฎความสัมพันธ์ในรูปแบบของการรวมกฎความสัมพันธ์และจัดอันดับกฎความสัมพันธ์รวมไปถึงการลดจำนวนกฎความสัมพันธ์ที่มีความซ้ำซ้อนนั้นค่อนข้างมีความซับซ้อนในกระบวนการทำงาน จึงทำให้งานวิจัยทางด้านนี้มีปรากฏค่อนข้างน้อย ผู้วิจัยได้เห็นถึงความสำคัญในจุดนี้จึงได้เสนอเทคนิคการเพิ่มกระบวนการทั้งก่อนและหลังการค้นหากฎความสัมพันธ์กับข้อมูลทางการแพทย์ที่ต้องอาศัยทั้งความละเอียดและรอบคอบในการวิเคราะห์ข้อมูล อีกทั้งผู้วิจัยยังได้เสนอมาตรวัดประสิทธิภาพใหม่ที่เรียกว่า Confidence and Accuracy (CAA) ซึ่งเป็นมาตรวัดที่เป็นการผสมผสานกันระหว่าง ค่าความเชื่อมั่นและค่าความแม่นตรง เพื่อสร้างเกณฑ์ในการค้นหากฎความสัมพันธ์ที่มีความครอบคลุมและสามารถวิเคราะห์ปัจจัยเสี่ยงได้อย่างชัดเจน

สาขาวิชาวิศวกรรมคอมพิวเตอร์ ลายมือชื่อนักศึกษา_____

ปีการศึกษา 2557 ลายมือชื่ออาจารย์ที่ปรึกษา_____

PONGSAKORN DURONGDUMRONGCHAI: ASSOCIATION RULESEARCH AND RANKING FOR THE ANALYSIS OF FACTORS CONTRIBUTING TO DISEASES. THESIS ADVISOR: ASSOC. PROF. NITTAYA KERDPRASOP, Ph.D., 121 PP.

ASSOCIATION RULES/MERGED RULE/RANKING RULE / REMOVED REDUNDANT RULE

Thisresearchwasstudiedthe problemof association rulesearchfor medical datasets with appearancenumeric and nominal of data mixtures. Association rulesearch inthis research use 3 algorithmincludingApriori, Predictive Apriori,andTertius. However, normal association rulesearch algorithm cannot deal with numeric data. Therefore, this research applies discretization technique to enablethe association rulesearch.In the case of data with missing or unidentified values, we decideto remove such transactions,with rarely occur in our datasets.Researchin thepast offered severaltechniques forfinding association rules in various forms,andmeasured theperformanceof the relationship in terms of confidence, which is a singlemeasurement.However, there are a fewresearchinthe field of associationrulesearch intheform of mergedruleand ranking ruleas well as removed redundantassociationrules, because these process are rather complicate. Werealize theimportanceofthis point and thus propose the pre-processing and post-processingassociation rule mining techniques to work with the medical datasets that requiresdetailed andcareful in data analysis.

Moreover we also propose a new performance measurecalled Confidence and Accuracy(CAA) to integrate the confidence and accuracy metrics for the complete and accurate analysis of factors contributing to diseases.

School of Computer Engineering Student's Signature_____

Academic Year 2014 Advisor's Signature_____

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จลุล่วงด้วยดี ผู้วิจัยขอกราบขอบพระคุณบุคคลและกลุ่มบุคคลต่างๆที่ได้กรุณาให้คำปรึกษา แนะนำ ช่วยเหลืออย่างดียิ่ง ทั้งในด้านวิชาการ และด้านการดำเนินงานวิจัย ดังต่อไปนี้

รองศาสตราจารย์ดร.นิตยา เกิดประสพ อาจารย์ที่ปรึกษาวิทยานิพนธ์และรองศาสตราจารย์ ดร.กิตติศักดิ์ เกิดประสพ ที่ให้คำปรึกษาในการทำงานวิจัย การจัดการรูปแบบ และช่วยตรวจทานความถูกต้องของวิทยานิพนธ์

ผู้ช่วยศาสตราจารย์ ดร.ชาญวิทย์ แก้วกสิผู้ช่วยศาสตราจารย์ ดร.คะชา ชาญศิลป์ ผู้ช่วยศาสตราจารย์ สมพันธุ์ ชาญศิลป์ และผู้ช่วยศาสตราจารย์ดร.ปรเมศวร์ ห่อแก้ว อาจารย์ประจำสาขาวิชาวิศวกรรมคอมพิวเตอร์ สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี

คุณกัลญา พับโพธิ์ เลขานุการสาขาวิชาวิศวกรรมคอมพิวเตอร์ ที่ให้ความช่วยเหลือในการประสานงานด้านเอกสารระหว่างศึกษา

คุณภาสพิชญ์ ชูใจ คุณไพชยนต์ คงไชย คุณนันทวุฒิ คะอังกุคุณกีระชาติ สุขสุทธิ์ คุณกิตติพงศ์ ชมบุญ คุณศักย์ เพิ่มหรรษา คุณรติพร จันทร์กลั่น คุณพงศกร ธีรรัศมีและนักศึกษาบัณฑิตสาขาวิชาวิศวกรรมคอมพิวเตอร์ ทุกท่านที่ให้คำปรึกษาและช่วยเหลือด้วยดีมาโดยตลอด

นอกจากนี้ขอขอบคุณครู อาจารย์ทั้งในอดีตและปัจจุบันที่ให้ความรู้แก่ผู้วิจัยจนประสบความสำเร็จในชีวิต

ท้ายที่สุดที่จะลืมไม่ได้ ขออุทิศบุญกุศลแด่ คุณแม่พูลศิลป์ ดุรงค์ดำรงชัย มารดาผู้ล่วงลับของข้าพเจ้า และขอกราบขอบพระคุณ คุณพ่อยิ่งยง ดุรงค์ดำรงชัย รองผู้อำนวยการเชี่ยวชาญ บิดาของข้าพเจ้าที่ให้กำเนิด อบรม เลี้ยงดูด้วยความรัก และส่งเสริมการศึกษาเป็นอย่างดีโดยตลอด ทำให้ผู้วิจัยมีความรู้ ความสามารถ มีจิตใจที่เข้มแข็ง รวมทั้งเป็นกำลังใจที่ยิ่งใหญ่แก่ผู้วิจัย จนทำให้ผู้วิจัยประสบความสำเร็จในชีวิตเรื่อยมา

พงศกร ดุรงค์ดำรงชัย

สารบัญ

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญรูป

รูปที่ **หน้า**

สารบัญรูป (ต่อ)

รูปที่ **หน้า**

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของปัญหาการวิจัย

ปัจจุบันเทคโนโลยีทางการแพทย์ได้ให้ความสำคัญในการวิเคราะห์ปัจจัยเสี่ยงที่จะส่งผลต่อการเกิดโรคชนิดต่าง ๆ โรคชนิดหนึ่งที่ทางการแพทย์กำลังให้ความสนใจคือ โรคที่เกี่ยวข้องกับระบบทางเดินหายใจ เนื่องจากระบบทางเดินหายใจของมนุษย์จะนำอากาศที่มีออกซิเจนเข้าไปเลี้ยงส่วนต่าง ๆ ของสมองซึ่งระบบหายใจที่ผิดปกติอาจส่งผลร้ายต่อระบบสมอง และระบบทางเดินหายใจที่ผิดปกติซึ่งทางการแพทย์ปัจจุบันได้ให้ความสนใจคือ โรคหอบหืด(Asthma) จัดอยู่ในภาวะภูมิแพ้ ซึ่งในทางการแพทย์จะแบ่งโรคหอบหืดออกเป็น 3 ระดับได้แก่ระดับการเกิดหอบหืดที่ความถี่ไม่บ่อยครั้ง (Low asthma) ระดับการเกิดหอบหืดที่ความถี่ปานกลาง (Moderateasthma) และระดับการเกิดหอบหืดที่ความถี่บ่อยครั้ง (High asthma)

การวิเคราะห์ปัจจัยเสี่ยงของการเกิดโรคหอบหืด สามารถช่วยให้ทราบว่าปัจจัยใดที่จะเกิดเป็นปัจจัยเสี่ยงในอนาคตที่จะก่อให้เกิดโรคหอบหืด และสามารถช่วยป้องกันให้ผู้ป่วยที่เข้าข่ายจะเป็นโรคหอบหืดสามารถลดปัจจัยเสี่ยงที่จะไปกระตุ้นการเกิดโรคหอบหืดได้ ซึ่งจะส่งผลดีต่อทางการแพทย์ที่จะสามารถช่วยลดต้นทุนในการวิเคราะห์และวินิจฉัยโดยแพทย์ผู้เชี่ยวชาญทางด้านโรคที่เกี่ยวข้องกับระบบทางเดินหายใจ

ในการวิจัยนี้ได้นำข้อมูลผู้ป่วยที่เป็นโรคหอบหืดที่มี 3 ระดับ นำมาทำการวิเคราะห์ปัจจัยเสี่ยงที่ทำให้เกิดโรคหอบหืดในผู้ป่วยโดยมีตัวอย่างของข้อมูลดังตารางที่ 1.1

ตารางที่ 1.1 ตัวอย่างของข้อมูลผู้ป่วยที่เป็นโรคหอบหืด

Weight	Height	Waist	Body fat	PA_level
62.4	163	75.5	33	Moderate
61.1	156	76.5	36.1	High
47.5	153	65.5	27.6	Low
57.1	160	83	31.6	Moderate
57.7	150	87	38.2	Low

จากตารางที่ 1.1 จะสังเกตเห็นว่าข้อมูลเป็นลักษณะผสมผสานกันระหว่างข้อมูลลักษณะตัวเลข (Numeric) และข้อมูลลักษณะข้อความ (Nominal) เมื่อนำไปใช้ในการหากฎความสัมพันธ์ (Association rules) แบบปกติ จะไม่สามารถแสดงกฎความสัมพันธ์ออกมาได้ แต่จะถูกแสดงในลักษณะของการแจ้งเตือนข้อผิดพลาด (“Cannot handle numeric attributes!”) ดังในบรรทัดสุดท้ายของรูปที่ 1.1

รูปที่ 1.1 ลักษณะการแจ้งเตือนข้อผิดพลาดของการใช้ข้อมูลลักษณะตัวเลขเมื่อนำไปหากฎความสัมพันธ์

จากรูปที่ 1.1 จะเห็นว่าข้อความของการแจ้งเตือนระบุไว้อย่างชัดเจนว่า การหากฎความสัมพันธ์แบบปกติโดยใช้อัลกอริทึมเอไพรออรี(Apriori)ไม่สามารถจัดการคุณลักษณะของข้อมูลที่อยู่ในลักษณะของตัวเลขได้ เมื่อไม่สามารถหากฎความสัมพันธ์ออกมาได้จึงไม่สามารถสร้างโมเดลที่จะนำไปใช้ในการวิเคราะห์ปัจจัยเสี่ยงที่ทำให้ผู้ป่วยเป็นโรคหอบหืด

จากปัญหาที่กล่าวมาข้างต้นผู้วิจัยจึงได้เสนอแนวทางการเพิ่มกระบวนการจัดการข้อมูลที่เป็นลักษณะตัวเลขด้วยวิธีการแบ่งช่วงข้อมูลที่มีลักษณะเป็นตัวเลข(Discretization) อีกทั้งผู้วิจัยยังทำการรวมกฎความสัมพันธ์ (Merge rules)ที่ได้จากทั้ง 3 วิธีได้แก่ วิธีหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรีวิธีหากฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทีพเอไพรออรี(Predictive Apriori)และวิธีหากฎความสัมพันธ์ด้วยอัลกอริทึมเทอเชียส(Tertius) รวมไปถึงการจัดอันดับกฎความสัมพันธ์ (Ranking rules) และขจัดกฎความสัมพันธ์ที่ซ้ำซ้อน (Remove redundant rules) เพื่อให้ได้กฎความสัมพันธ์ที่ดีที่สุด (Best rules) จากกฎความสัมพันธ์ที่ได้จากกระบวนการข้างต้นจะถูกนำไปเปรียบเทียบประสิทธิภาพในเชิงของกฎความสัมพันธ์ เพื่อเพิ่มความสามารถในการวิเคราะห์ปัจจัยเสี่ยงให้ครอบคลุมมากยิ่งขึ้น และสามารถเพิ่มค่าความเชื่อมั่น(Confidence) ให้กับวิธีการหากฎความสัมพันธ์ที่ผู้วิจัยได้นำเสนอแนวทาง

1.2 วัตถุประสงค์ของการวิจัย

ผู้วิจัยได้ตั้งวัตถุประสงค์ในการวิจัยดังต่อไปนี้

1) เพื่อค้นหากฎความสัมพันธ์จากข้อมูลที่ผสมผสานกันระหว่างข้อมูลลักษณะตัวเลขและข้อมูลลักษณะข้อความ
2) เพื่อรวมกฎความสัมพันธ์ที่มีลักษณะเหมือนกันจากการหากฎความสัมพันธ์แบบปกติโดยใช้อัลกอริทึมทั้ง 3 วิธี ได้แก่ วิธีหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรีวิธีหากฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทีพเอไพรออรีและวิธีหากฎความสัมพันธ์ด้วยอัลกอริทึมเทอเชียส
3) เพื่อจัดอันดับกฎความสัมพันธ์และขจัดกฎความสัมพันธ์ที่ซ้ำซ้อน

1.3 ข้อตกลงเบื้องต้น

1) ข้อมูลจริงที่ใช้ในการทดลองได้แก่ ข้อมูลผู้ป่วยที่เป็นโรคหอบหืดที่มีระดับความรุนแรง 3 ระดับของโรงพยาบาลมหาราชจังหวัดนครราชสีมา เก็บข้อมูล ณ วันที่ 8 พฤศจิกายน พ.ศ.2557
2) ข้อมูลผู้ป่วยที่เป็นโรคหัวใจจากแหล่งข้อมูลมาตรฐาน (UC Irvine)ใช้ในการเปรียบเทียบกับการหากฎความสัมพันธ์แบบอื่น

1.4 ขอบเขตของการวิจัย

จากการศึกษาค้นคว้าข้อมูล ผู้วิจัยได้กำหนดขอบเขตของการวิจัยไว้ดังนี้

1) การวิจัยนี้จะทำเฉพาะในส่วนของการหากฎความสัมพันธ์ ซึ่งจะไม่ทำในส่วนของการพยากรณ์ผลของข้อมูลที่จะเกิดขึ้นในอนาคต (Classification)
2) ข้อมูลทางการแพทย์ในงานวิจัยนี้จะใช้ข้อมูลผู้ป่วยโรคหอบหืดและข้อมูลผู้ป่วยโรคหัวใจ(Heart disease)

1.5 ประโยชน์ที่คาดว่าจะได้รับ

ประโยชน์ที่เกิดขึ้นจากการวิจัยนี้ ประกอบด้วย

1) ได้กฎความสัมพันธ์สำหรับการวิเคราะห์ปัจจัยเสี่ยงของการเกิดโรคหอบหืดที่สามารถอธิบายในลักษณะช่วงของข้อมูลที่เป็นตัวเลข
2) ได้กฎความสัมพันธ์ที่มีความเชื่อมั่นสูงขึ้นกว่าการใช้วิธีการหากฎความสัมพันธ์แบบปกติ

บทที่ 2
ปริทัศน์วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงปริทัศน์วรรณกรรมและงานวิจัยที่เกี่ยวข้อง โดยมีรายละเอียดการค้นหากฎความสัมพันธ์ การเตรียมข้อมูลก่อนหากฎความสัมพันธ์ ประเภทของการค้นหากฎความสัมพันธ์ รูปแบบของการแสดงกฎความสัมพันธ์พื้นฐานทั้งสามแบบ การจัดอันดับของกฎความสัมพันธ์ การขจัดกฎความสัมพันธ์ที่ซ้ำซ้อนรูปแบบของการจัดอันดับและการลดความซ้ำซ้อนของกฎความสัมพันธ์มาตรวัดประสิทธิภาพกฎความสัมพันธ์ และงานวิจัยที่เกี่ยวข้อง

2.1 การค้นหากฎความสัมพันธ์

2.1.1 ความหมายของกฎความสัมพันธ์และการนำไปใช้

กฎความสัมพันธ์ประกอบขึ้นจากคำว่า “กฎ” ซึ่งหมายถึงหลักเกณฑ์หรือข้อบังคับที่มนุษย์ได้สร้างขึ้นเพื่อต้องการใช้เป็นข้อตกลงร่วมกันให้มีความเป็นอันหนึ่งอันเดียวกัน รวมกับคำว่า “ความสัมพันธ์” ซึ่งจะหมายถึงการเกี่ยวข้องซึ่งกันและกันไม่ว่าจะทางตรงหรือทางอ้อมก็ตาม ดังนั้นกฎความสัมพันธ์จึงหมายถึงหลักเกณฑ์หรือข้อบังคับของความเกี่ยวข้องกันที่ใช้การตกลงร่วมกันในที่นี้ขอยกตัวอย่างกฎความสัมพันธ์ที่ไม่ซับซ้อนหนึ่งกฎความสัมพันธ์สำหรับการนำไปใช้ในการวิเคราะห์ข้อมูล

IF[Sex=Female && Smoking=Never_smoker&& Exercise=No_exercise]

THEN[PA_level=Low]

จากกฎความสัมพันธ์ที่อยู่ในรูป ***IF…THEN…***เบื้องต้นสามารถตีความได้ว่า

“ถ้าผู้หญิงที่ไม่สูบบุหรี่และไม่ออกกำลังกายแล้วระดับของการเกิดหอบหืด (PA_level)จะจัดอยู่ในระดับต่ำ (Low)”

2.1.2 การค้นหากฎความสัมพันธ์ในแง่มุมทางการแพทย์

การค้นหากฎความสัมพันธ์ของข้อมูลทางด้านการแพทย์ทั้งในอดีตและปัจจุบันได้มีนักวิจัยหลายกลุ่มที่ได้ศึกษาค้นคว้าและพยายามที่จะสร้างอัลกอริทึมที่ดีที่สุด สำหรับการค้นหากฎความสัมพันธ์ของข้อมูลทางการแพทย์ให้ครอบคลุมมากที่สุด ซึ่งจะต้องสามารถตีความได้

กว้างขวางที่สุดรวมไปถึงค่าความเชื่อมั่นที่อยู่ในเกณฑ์ที่ดีถึงดีที่สุด ยกตัวอย่างเช่นงานวิจัยของ (Jesminet al., 2013) ได้นำข้อมูลผู้ป่วยโรคหัวใจของฐานข้อมูลมาตรฐาน(UCI, 2010)มาทำการค้นหากฎความสัมพันธ์ด้วยวิธีหากฎความสัมพันธ์ พื้นฐานทั้ง 3 วิธี ได้แก่ วิธีหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรีวิธีหากฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทีพเอไพรออรี และวิธีหากฎความสัมพันธ์ด้วยอัลกอริทึมเทอเชียส โดยจะทำการเลือกกฎความสัมพันธ์พื้นฐานที่ให้ค่าความเชื่อมั่นที่ดีที่สุดนำไปใช้ในการพิจารณาคุณสมบัติอื่นต่อไป และงานวิจัยของ (Meghana et al., 2013)ได้ใช้ข้อมูลการตรวจหาเนื้องอกในสมอง (Brain tumor detection) ซึ่งเป็นข้อมูลที่เกี่ยวกับทางด้านการแพทย์มาใช้ในการค้นหากฎความสัมพันธ์

2.1.3 ประโยชน์ของการนำกฎความสัมพันธ์มาใช้ในทางการแพทย์

อย่างที่ทราบกันดีว่าสำหรับทางการแพทย์แล้วนั้นการวินิจฉัยหรือการตรวจสอบโรคของผู้ป่วยเป็นเรื่องที่มีความละเอียดอ่อนและค่อนข้างซับซ้อนมากรวมไปถึงยังต้องใช้ผู้เชี่ยวชาญทางด้านการแพทย์มาวิเคราะห์หรือวินิจฉัยจากอาการของผู้ป่วยที่เป็นโรคต่าง ๆ ดังนั้นการสร้างกฎความสัมพันธ์ที่ดีที่สุดและครอบคลุมการวินิจฉัยให้มากที่สุดรวมถึงให้ค่าความเชื่อมั่นที่ดีถึงดีที่สุดเป็นสิ่งสำคัญและจำเป็นต่อวงการทางการแพทย์ในปัจจุบัน ยกตัวอย่างเช่น การวิเคราะห์และวินิจฉัยโรคโดยไม่พึ่งพากฎความสัมพันธ์ ซึ่งเป็นการวินิจฉัยตามอาการโดยไม่พึ่งกฎความสัมพันธ์ที่บันทึกอาการต่าง ๆ ของผู้ป่วยแต่ละราย อาจจะส่งผลกระทบทำให้การวินิจฉัยผิดพลาดและส่งผลร้ายทำให้ผู้ป่วยไม่ได้ถูกรักษาตามอาการป่วยที่แท้จริง นอกจากจะทำให้ผู้ป่วยไม่หายเป็นปกติแล้วยังส่งผลให้มีความสิ้นเปลืองในด้านอุปกรณ์ทางการแพทย์รวมไปถึงค่ารักษาพยาบาลของผู้ป่วยซึ่งจะส่งผลเสียต่อด้านทรัพยากรอีกด้วย

2.2 การเตรียมข้อมูลก่อนการหากฎความสัมพันธ์

2.2.1 การลบข้อมูลที่ไม่สามารถบ่งชี้ข้อมูลได้ชัดเจน

ข้อมูลที่ไม่สามารถบ่งชี้ได้ชัดเจนหรือข้อมูลสูญหาย(Missing data) คือข้อมูลที่ไม่ได้มีการระบุไว้ในทรานแซคชัน (Transaction) ใด ๆ ซึ่งข้อมูลดังกล่าวส่วนใหญ่จะอยู่ในลักษณะของข้อมูลว่างเปล่า (Blank data) และการมีข้อมูลที่ไม่สามารถบ่งชี้ได้ชัดเจนจะส่งผลเสียในการวิเคราะห์เพื่อค้นหากฎความสัมพันธ์ จึงได้เกิดกระบวนการหนึ่งซึ่งเรียกว่าการลบข้อมูลที่ไม่สามารถบ่งชี้ข้อมูลได้ชัดเจน(Remove transaction)ซึ่งการลบข้อมูลที่ไม่สามารถบ่งชี้ข้อมูลได้ชัดเจนจัดอยู่ในกระบวนการเตรียมข้อมูล (Pre-process)ข้อดีของการลบข้อมูลที่ไม่สามารถบ่งชี้ได้ชัดเจนนั้น สามารถช่วยให้การตีความของกฎความสัมพันธ์ไม่เกิดความผิดเพี้ยนไปจากความเป็น

จริง ดังนั้นจึงมีความจำเป็นอย่างมากที่จะต้องทำการลบข้อมูลที่ไม่สามารถบ่งชี้ได้ชัดเจน ยกตัวอย่างเช่น งานวิจัยของ (Rahman and Zahidul, 2013)ได้มีการระบุไว้ชัดเจนว่าได้ทำการลบทรานแซคชันที่ข้อมูลไม่สามารถบ่งชี้ได้อย่างชัดเจนดังตัวอย่างแสดงในรูปที่2.1

ข้อมูลก่อนการลบทรานแซคชัน

Sex	Weight	Bodyfat	PA_level
Male	101	31.4	High
Female		26.4	Moderate
Male	106.3	32.4	High
Male	113.3	36.4	Low

ข้อมูลหลังการลบทรานแซคชัน

Sex	Weight	Bodyfat	PA_level
Male	101	31.4	High
Male	106.3	32.4	High
Male	113.3	36.4	Low

รูปที่ 2.1 ตัวอย่างการลบข้อมูลที่ไม่สามารถบ่งชี้ได้ชัดเจน

จากรูปที่ 2.1 อธิบายได้ว่าจากตารางด้านซ้ายมือพบว่าทรานแซคชันลำดับที่ 2 คอลัมน์ (Column) ของ Weightที่ใช้อธิบายน้ำหนักของบุคคลขาดหายไป ดังนั้นเมื่อข้อมูลมีการขาดหายไปบางส่วนของทรานแซคชันลำดับที่ 2 จะส่งผลให้การสร้างกฎความสัมพันธ์มีความผิดเพี้ยนไปจากความเป็นจริง จึงมีความจำเป็นที่จะต้องทำการลบทรานแซคชันลำดับที่ 2 ดังนั้นข้อมูลใหม่จึงมีลักษณะเป็นดังตารางด้านขวามือจะเหลือข้อมูลเพียง 3 ทรานแซคชัน

2.2.2 การแบ่งช่วงข้อมูลที่มีลักษณะเป็นตัวเลข

การแบ่งช่วงข้อมูลที่มีลักษณะเป็นตัวเลขเป็นหนึ่งในวิธีการแปลงข้อมูลที่อยู่ในลักษณะของตัวเลขให้เป็นช่วงของข้อมูลซึ่งจะอยู่ในลักษณะของข้อมูลที่เป็นข้อความการแบ่งช่วงข้อมูลที่มีลักษณะเป็นตัวเลขจะสามารถช่วยแก้ไขปัญหาในการหากฎความสัมพันธ์ที่มีข้อมูลอยู่ในลักษณะผสมผสานกันระหว่างข้อมูลที่เป็นลักษณะตัวเลขและข้อมูลที่เป็นลักษณะข้อความ ซึ่งมีงานวิจัยจำนวนมากที่ใช้วิธีการแบ่งช่วงข้อมูลหนึ่งในงานวิจัยนั้นได้แก่งานวิจัยของ (Yan et al., 2011)ซึ่งใช้วิธีการแบ่งช่วงข้อมูลในแง่ของการสร้างข้อมูลชุดฝึกหัด (Training set) ชุดใหม่ และสร้างข้อมูลชุดทดสอบ (Test set) ชุดใหม่ตัวอย่างการแบ่งช่วงข้อมูลที่มีลักษณะเป็นตัวเลขแสดงได้ดังรูปที่ 2.2

ก่อนการจัดช่วงของข้อมูล

Sex	Weight	Bodyfat	PA_level
Male	96.7	31.4	Low
Male	110.8	39.4	Low
Male	113.3	36.4	Low

หลังการจัดช่วงของข้อมูล

Sex	Weight	Bodyfat	PA_level
Male	95.0-100.0	30.0-35.0	Low
Male	110.0-115.0	35.0-40.0	Low
Male	110.0-115.0	35.0-40.0	Low

รูปที่ 2.2 ตัวอย่างการแบ่งช่วงข้อมูลที่มีลักษณะเป็นตัวเลข

จากรูปที่ 2.2 อธิบายได้ว่าจากตารางด้านซ้ายมือพบว่าคุณลักษณะ (Attribute) ในส่วนของ Weight ที่ใช้อธิบายน้ำหนักของบุคคลและ Bodyfat ที่ใช้อธิบายดัชนีมวลรวมของบุคคล ลักษณะของข้อมูลอยู่ในลักษณะของตัวเลข เมื่อข้อมูลอยู่ในลักษณะนี้จะยังไม่สามารถทำการค้นหากฎความสัมพันธ์ได้ เนื่องจากข้อมูลทั้งหมดยังไม่ได้อยู่ในรูปของลักษณะข้อความ ดังนั้นจึงมีความจำเป็นที่ต้องทำการจัดช่วงของข้อมูลที่เป็นตัวเลข ซึ่งจะทำให้ข้อมูลถูกแปลงให้อยู่ในรูปของลักษณะข้อความ ข้อมูลใหม่จึงมีลักษณะเป็นดังตารางด้านขวามือ

2.3 ประเภทของการค้นหากฎความสัมพันธ์

2.3.1 การหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรี

อัลกอริทึมค้นหากฎความสัมพันธ์เอไพรออรีเป็นหนึ่งในวิธีการหากฎความสัมพันธ์แบบพื้นฐานถูกนำเสนอโดยAgrawal และ Srikant (1994)

2.3.1.1 แนวคิดเบื้องต้นของวิธีหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรี

แนวคิดเริ่มด้วยมีชุดข้อมูลใด ๆ ที่บรรจุทรานแซคชันต่าง ๆ ไว้ภายในเพื่อที่จะสร้างชุดของไอเท็มเซต(Item set)จากชุดข้อมูลใด ๆและจะนำมาพิจารณาใช้ในการสร้างกฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรี ซึ่งเกณฑ์การวัดประสิทธิภาพนั้นงานวิจัยส่วนมากที่ใช้วิธีหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรี มักจะนำเสนอในรูปแบบของค่าความเชื่อมั่น พร้อมทั้งแสดงกฎความสัมพันธ์โดยเรียงลำดับตามค่าความเชื่อมั่นจากความเชื่อมั่นสูงสุดลดระดับลงไปจนถึงความเชื่อมั่นที่สามารถยอมรับได้ เช่นงานวิจัยของ (Mu, 2007) ได้ใช้วิธีหากฎความสัมพันธ์

ด้วยอัลกอริทึมเอไพรออรีในแง่ของการค้นหากฎความสัมพันธ์ที่ซ่อนอยู่ภายในข้อมูล ซึ่งวิธีหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรีมีอัลกอริทึมดังรูปที่ 2.3

2.3.1.2 การหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรี

```
APRIORI (D, I, minsup):
1  F ← ∅
2  C(1) ← {∅} // Initial prefix tree with single items
3  foreach i ∈ I do Add i as child of ∅ in C(1) with sup(i) ← 0
4  k ← 1 // k denotes the level
5  while C(k) ≠ ∅ do
6      COMPUTESUPPORT (C(k), D)
7      foreach leaf X ∈ C(k) do
8          if sup(X) ≥ minsup then F ← F ∪ {(X, sup(X))}
9          else remove X from C(k)
10     C(k+1) ← EXTENDPREFIXTREE (C(k))
11     k ← k + 1
12 return F(k)

COMPUTESUPPORT (C(k), D):
13 foreach ⟨t, i(t)⟩ ∈ D do
14     foreach k-subset X ⊆ i(t) do
15         if X ∈ C(k) then sup(X) ← sup(X) + 1

EXTENDPREFIXTREE (C(k)):
16 foreach leaf Xa ∈ C(k) do
17     foreach leaf Xb ∈ SIBLING(Xa), such that b > a do
18         Xab ← Xa ∪ Xb
           // prune candidate if there are any infrequent subsets
19         if Xj ∈ C(k), for all Xj ⊂ Xab, such that |Xj| = |Xab| − 1 then
20             Add Xab as child of Xa with sup(Xab) ← 0
21     if no extensions from Xa then remove Xa from C(k)
22 return C(k)
```

รูปที่ 2.3 อัลกอริทึมค้นหากฎความสัมพันธ์เอไพรออรี(Mohammed and Wagner, 2014)

จากรูปที่ 2.3 เป็นวิธีการค้นหากฎความสัมพันธ์พื้นฐานด้วยอัลกอริทึมเอไพรออรีซึ่งมีกระบวนการทำงานหลัก 4 ส่วนดังนี้

1) กำหนดค่าตัวแปรเริ่มต้นของการทำงาน
2) พิจารณาค่าสนับสนุนของกฎความสัมพันธ์
3) ตรวจสอบค่าสนับสนุนของกฎความสัมพันธ์ว่าอยู่เกินช่วงค่าสนับสนุนต่ำที่สุดหรือไม่
4) สร้างกฎความสัมพันธ์โดยอาศัยหลักการของ Prefix Treeซึ่งจะได้กฎความสัมพันธ์ที่ดีที่สุดอยู่ในมาตรวัดของค่าความเชื่อมั่นอัลกอริทึมการหากฎความสัมพันธ์เอไพรออรี มีกระบวนการทำงานดังนี้

ฟังก์ชัน เอไพรเออรี

รับค่าจากตัวแปร D ที่แทน Databaseหรือฐานข้อมูล รับค่าจากตัวแปร I ที่แทน Itemsets และรับค่า minsup ที่แทน minimum support

บรรทัดที่ 1กำหนดค่าให้ Frequentitemsetมีค่าเป็น เซตว่าง

บรรทัดที่ 2กำหนดค่าให้ Candidate K-itemsetsที่แทนด้วย$C^{(k)}$มีค่าเป็นเซตว่าง

บรรทัดที่ 3-12เป็นการทำงานใน LoopForeachโดยจะทำการวนLoopทั้งสิ้น I รอบโดยกำหนดค่า support ของI เป็น0 ค่าของK มีค่าเป็น1 ภายในมีการทำงานดังต่อไปนี้

- สำหรับ $C^{(k)}$ ที่ไม่เป็นเซตว่าง ให้ทำดังต่อไปนี้
  - เรียกใช้ฟังก์ชัน ComputeSupport โดยส่งค่า($C^{(k)}$, D)
  - บรรทัดที่ 7-9เป็นการทำงานใน LoopForeachโดยจะทำการวน Loopทั้งสิ้นจำนวน kรอบ จะได้leafของ X
    - ถ้าค่า supportของ Xมากกว่าหรือเท่ากับ ค่า minsupให้เก็บค่าXและค่า supportของ Xอยู่ในรูปของเซต ไว้ที่ตัวแปร F
    - ถ้าค่า supportของ Xต่ำกว่า minsupให้ลบ Xจาก $C^{(k)}$
  - นำค่า$C^{(k)}$ที่ได้จากฟังก์ชัน ExtendPrefixTreeเก็บไว้ที่$C^{(k+1)}$
  - เพิ่มค่าK ขึ้น1 ค่า

ฟังก์ชัน ComputeSupport

รับค่าจากตัวแปร $C^{(k)}$และรับค่าจากตัวแปร D

บรรทัดที่ 13-15เป็นการทำงานใน LoopForeachโดยจะทำการวน Loopทั้งสิ้นจำนวนtรอบ โดยจะใช้เซตของitemใน transactionมาทำงานในบรรทัดที่ 14

บรรทัดที่ 14-15เป็นการทำงานใน LoopForeachโดยจะทำการวน Loopทั้งสิ้นจำนวนkรอบ โดยจะใช้ เซตของ X

- ถ้า X เป็นสมาชิกของ $C^{(k)}$แล้วจะทำการบวกค่า 1ให้กับ ค่าsupportของ X

ฟังก์ชัน ExtendPrefixTree
รับค่าจากตัวแปร $C^{(k)}$
บรรทัดที่ 16-21เป็นการทำงานใน LoopForeachโดยจะทำการวน Loopทั้งสิ้นจำนวน kรอบ โดยจะทำการสร้าง Treeโดยเริ่มขยาย Treeจาก Rootnodeแล้วสร้างโหนดลูกจากทางซ้ายไปขวา

2.3.2 การหากฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทีพเอไพรออรี

การหากฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทีพเอไพรออรีเป็นวิธีการหากฎความสัมพันธ์ที่ได้รับแรงบันดาลใจมาจากการหากฎความสัมพันธ์แบบเอไพรออรี ถูกนำเสนอโดย Scheffer (2001)

2.3.2.1 แนวคิดเบื้องต้นวิธีหากฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทีพเอไพรออรี

แนวคิดเริ่มต้นคือต้องการที่จะสร้างมาตรวัดความแม่นตรง (Accuracy)ซึ่งมาจากการรวมกันระหว่างค่าสนับสนุน (Support)ของกฎความสัมพันธ์ และค่าความเชื่อมั่นของการค้นหากฎความสัมพันธ์และนำความแม่นตรงมาใช้สำหรับการจัดลำดับของกฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทีพเอไพรออรี ซึ่งตรงกันข้ามกับแนวคิดของอัลกอริทึมเอไพรออรีที่มุ่งเน้นในการหาค่าความเชื่อมั่นจากกฎความสัมพันธ์ของข้อมูลเท่านั้น ดังนั้นทำให้การประมวลผลของอัลกอริทึมพรีดิคทีพเอไรออรีมีข้อด้อยตรงที่จะต้องใช้หน่วยความจำมากกว่าอัลกอริทึมเอไพรออรีเมื่อทำการพิจารณาข้อมูลที่เหมือนกัน อย่างไรก็ตามการค้นหากฎความสัมพันธ์แบบด้วยอัลกอริทึมพรีดิคทีพเอไรออรียังคงถูกนำมาใช้ในงานวิจัยจำนวนมาก เช่นงานวิจัยของ(Divya andLekha, 2013)ได้ใช้วิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทีพไอไพรออรีมาใช้ในการค้นหากฎความสัมพันธ์ เพื่อใช้ในการเปรียบเทียบประสิทธิภาพของการสร้างกฎความสัมพันธ์ที่สามารถอธิบายได้ครอบคลุมและเปรียบเทียบเวลาที่ใช้ในการประมวลผลกับการค้นหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรี ซึ่งวิธีหากฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทีพไอไพรออรีแสดงได้ดังรูปที่ 2.4

2.3.2.2 การหากฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทีพเอไพรออรี

Table 1. Algorithm PredictiveApriori: discovery of n most predictive association rules.

1. **Input:** n (desired number of association rules), database with items $a_1, \ldots, a_k$.
2. **Let** $\tau = 1$.
3. **For** $i = 1 \ldots k$ **Do:** Draw a number of association rules $[x \Rightarrow y]$ with i items at random. Measure their confidence (provided $s(x) > 0$). Let $\pi_i(c)$ be the distribution of confidences.
4. **For** all c, **Let** $\pi(c) = \frac{\sum_{i=1}^{k} \pi_i(c)\binom{k}{i}(2^i - 1)}{\sum_{i=1}^{k}\binom{k}{i}(2^i-1)}$.
5. **Let** $X_0 = \{\emptyset\}$; **Let** $X_1 = \{\{a_1\}, \ldots, \{a_k\}\}$ be all item sets with one single element.
6. **For** $i = 1 \ldots k-1$ **While** ($i = 1$ or $X_{i-1} \neq \emptyset$).
   (a) **If** $i > 1$ **Then** determine the set of candidate item sets of length i as $X_i = \{x \cup x' | x, x' \in X_{i-1}, |x \cup x'| = i\}$. Generation of X_i can be optimized by considering only item sets x and $x' \in X_{i-1}$ that differ only in the element with highest item index. Eliminate double occurrences of item sets in X_i.
   (b) Run a database pass and determine the support of the generated item sets. Eliminate item sets with support less than τ from X_i.
   (c) **For** all $x \in X_i$ **Call RuleGen**(x).
   (d) **If** *best* has been changed, **Then Increase** τ to be the smallest number such that $E(c|1, \tau) > E(c(best[n])|\hat{c}(best[n], s(best[n]))$ (refer to Equation 6). **If** $\tau >$ database size, **Then Exit.**
   (e) **If** τ has been increased in the last step, **Then** eliminate all item sets from X_i which have support below τ.
7. **Output** $best[1] \ldots best[n]$, the list of the n best association rules.

Algorithm RuleGen(x) (generate all rules with body x)

10. **Let** γ be the smallest number such that $E(c|\gamma/s(x), s(x)) > E(c(best[n])|\hat{c}(best[n], s(best[n]))$.
11. **For** $i = 1 \ldots k$ **With** $a_i \notin x$ **Do** (for all items not in x)
    (a) **If** $i = 1$ **Then Let** $Y_1 = \{\{a_i\} | a_i \notin x\}$ (item sets with one element not in x).
    (b) **Else Let** $Y_i = \{y \cup y' | y, y' \in Y_{i-1}, |y \cup y'| = i\}$ analogous to the generation of candidates in step 6a.
    (c) **For** all $y \in Y_i$ **Do**
        i. Measure the support $s(x \cup y)$. **If** $s(x \cup y) \leq \gamma$, **Then** eliminate y from Y_i and **Continue** the for loop with the next y.
        ii. Equation 6 gives the predictive accuracy $E(c([x \Rightarrow y])|s(x \cup y)/s(x), s(x))$.
        iii. **If** the predictive accuracy is among the n best found so far (recorded in *best*), **Then** update *best*, remove rules in *best* that are subsumed by other rules, and **Increase** γ to be the smallest number such that $E(c|\gamma/s(x), s(x)) \geq E(c(best[n])|\hat{c}(best[n], s(best[n]))$.
12. **If** any subsumed rule has been erased in step 11(c)iii, **Then** recur from step 10.

รูปที่ 2.4 อัลกอริทึมค้นหากฎความสัมพันธ์พรีดิคทีพเอไพรออรี(Scheffer, 2001)

จากรูปที่ 2.4เป็นวิธีการค้นหากฎความสัมพันธ์พื้นฐานด้วยอัลกอริทึมพรีดิคทีพเอไพรออรีซึ่งมีกระบวนการทำงานหลัก 6ส่วนดังนี้

1) กำหนดค่าตัวแปรเริ่มต้นของการทำงาน
2) กำหนดให้มีการนำค่าความเชื่อมั่นมาใช้ในการคำนวณหาค่าความแม่นตรง
3) พิจารณาค่าสนับสนุนของกฎความสัมพันธ์
4) ตรวจสอบค่าสนับสนุนของกฎความสัมพันธ์ว่าอยู่เกินช่วงค่าสนับสนุนต่ำที่สุดหรือไม่
5) เรียกใช้ฟังก์ชัน RuleGen สำหรับการสร้างกฎความสัมพันธ์โดยมีการทำงานภายในดังนี้
   - นำค่าสนับสนุนและค่าความเชื่อมั่นมาคำนวณหาค่าความแม่นตรง
   - เมื่อได้ค่าความแม่นตรงแล้วนำมาตรวจสอบว่าอยู่เกินช่วงค่าความแม่นตรงที่ต่ำที่สุดหรือไม่
6) สร้างกฎความสัมพันธ์โดยจะได้กฎความสัมพันธ์ที่ดีที่สุดอยู่ในมาตรวัดของค่าความแม่นตรง

2.3.3 การหากฎความสัมพันธ์ด้วยอัลกอริทึมเทอเชียส

กฎความสัมพันธ์เทอเชียสถูกนำเสนอโดย Peterและ Nicolas(2001) เป็นการหากฎความสัมพันธ์แบบหนึ่งที่พัฒนามาจากการเขียนโปรแกรมเชิงตรรกะ (Logicprogramming) ที่ใช้สำหรับหาค่าสูงสุดของฟังก์ชันการประเมินผล (Evaluation)

2.3.3.1 แนวคิดเบื้องต้นของวิธีหากฎความสัมพันธ์ด้วยอัลกอริทึมเทอเชียส

แนวคิดของการค้นหากฎความสัมพันธ์ด้วยอัลกอริทึมเทอเชียสคือการค้นหากฎความสัมพันธ์ลำดับที่หนึ่งที่ดีที่สุดโดยจะมีค่าความเชื่อมั่นสูงที่สุด ซึ่งเป็นข้อดีในด้านของความเร็วในการประมวลผลที่ดีกว่าการหากฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทีพเอไพรออรีส่วนกฎความสัมพันธ์ในลำดับถัดมาจะให้ค่าความเชื่อมั่นที่น้อยกว่ากฎความสัมพันธ์ลำดับที่หนึ่งอยู่มาก ซึ่งในจุดนี้จึงเป็นข้อด้อยของการค้นหากฎความสัมพันธ์แบบเทอเชียสซึ่งงานวิจัยส่วนมากที่ใช้การค้นหากฎความสัมพันธ์ด้วยอัลกอริทึมเทอเชียสจะแสดงมาตรวัดประสิทธิภาพอยู่ในรูปค่าความเชื่อมั่น แม้ว่าการค้นหากฎความสัมพันธ์ด้วยอัลกอริทึมเทอเชียสจะมีข้อด้อยในส่วนของการแสดงผลค่าความเชื่อมั่นไม่มีสมดุลของกฎความสัมพันธ์แต่ก็ยังคงถูกนำมาใช้ในงานวิจัยจำนวนมาก เช่นงานวิจัยของ(Loboand Sunita, 2012)ได้ใช้วิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมเทอเชียสในการค้นหากฎความสัมพันธ์ที่มีกระบวนการแตกต่างกันสองแบบ แบบที่หนึ่งคือไม่ใช้ฟิลเตอร์ (Filter) และแบบที่สองใช้ฟิลเตอร์ เพื่อใช้เปรียบเทียบประสิทธิภาพของการสร้างกฎ

ความสัมพันธ์ที่ดีที่สุด กับการค้นหาความสัมพันธ์ทั้งสองอัลกอริทึมได้แก่ วิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรีและวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทีพเอไพรออรีซึ่งวิธีหากฎความสัมพันธ์ด้วยอัลกอริทึมเทอเชียสมีแนวคิดพื้นฐานแสดงได้ดังรูปที่ 2.5 2.6 และ 2.7

2.3.3.2 ขั้นตอนพื้นฐานของอัลกอริทึมเทอเชียส

```
For i = 1 to number_of_rules - 1 do
    Cur = C_i
    For j = i + 1 to number_of_rules do
        If Cur and C_j can be resolved then
            Cur = resolve (Cur, C_j)
            Output Cur
```

รูปที่ 2.5 ขั้นตอนการเตรียมข้อมูลก่อนการค้นหากฎความสัมพันธ์ (Peter et al., 2006)

จากรูปที่ 2.5 เป็นวิธีการเตรียมข้อมูลก่อนการค้นหากฎความสัมพันธ์ซึ่งมีหลักแนวคิดคือ ถ้าเมื่อใดก็ตามมีกฎความสัมพันธ์ใหม่เพิ่มเข้ามาในกระบวนการค้นหากฎความสัมพันธ์ทรานแซคชันในตารางเก่าจะเปลี่ยนข้อมูลเป็นข้อมูลเริ่มต้นทั้งหมด

```
For every cell of the contingency table, generate the corresponding clause D.
    For every rule C of the background knowledge
        verify if C implies D using the simplified subsumption test.
        If so, set to 0 the corresponding cell of the contingency table and exit from the
        inner cycle
generate clause D' by removing the signs of literals from clause D.
For each clause C in the background knowledge
    generate clause C' by removing the signs of literals from C.
    use the simplified subsumption test to see whether C' implies D'.
    If so, set to 0 the corresponding cell of the contingency table
```

รูปที่ 2.6 แนวคิดพื้นฐานของอัลกอริทึมเทอเชียส (Perter et al., 2006)

จากรูปที่ 2.6 แนวคิดพื้นฐานของอัลกอริทึมเทอเชียส ที่กล่าวถึงระบบการคิดแบบพิจารณาที่ข้อมูลมีสองมิติ

```
If for every literal of C there is a literal in D with the same predicate symbol and sign then.
    For all couples (L1, L2) of literals in C
        Consider the couple (M1, M2) formed by the literals of D that correspond to (L1, L2).
        For all couples of arguments of (L1, L2) that are identical.
            If the corresponding arguments of (M1, M2)are different then.
                Return failure.
    Return success.
Else return failure.
```

รูปที่ 2.7การทดสอบประสิทธิภาพแนวคิดพื้นฐานของอัลกอรึทึมเทอเชียส (Peter et al., 2006)

จากรูปที่ 2.7การทดสอบประสิทธิภาพแนวคิดพื้นฐานของอัลกอริทึมเทอเชียสจะทำการทดสอบประสิทธิภาพของข้อกฎความสัมพันธ์โดยจะดูกฎความสัมพันธ์ความสัมพันธ์เป็นระบบคู่ (Couple)

2.4 รูปแบบของการแสดงกฎความสัมพันธ์พื้นฐานทั้งสามอัลกอริทึม

ผลการค้นหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรี พรีดิคทีพเอไพรออรี และเทอเชียส แสดงได้โดยใช้ข้อมูลตัวอย่างขนาดเล็กจำนวน 6 ทรานแซคชันดังตารางที่ 2.1

ตารางที่ 2.1ตัวอย่างข้อมูลโรคหอบหืดใช้ในการค้นหากฎความสัมพันธ์

Age	Sex	Religion	Smoking	Exercise	Weight	Bodyfat	PA_level
38	Male	Buddhist	Never_smoker	Exercise	85.2	24.7	High
54	Male	Buddhist	Currents_smoker	No_exercise	57	24.5	Low
35	Male	Buddhist	Currents_smoker	Exercise	59.6	14.1	Moderate
45	Female	Buddhist	Never_smoker	Exercise	57	31.9	Moderate
43	Female	Buddhist	Never_smoker	No_exercise	49.8	31.5	Low
42	Male	Buddhist	Previous_smoker	Exercise	76.1	27.4	High

2.4.1 รูปแบบของการแสดงกฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรี

IF [Sex=Female&& Smoking=Never_smoker && Exercise=No_exercise]

THEN[PA_level=Low] **conf: (1)**

2.4.2 รูปแบบของการแสดงกฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทีพเอไพรออรี

IF [Exercise=No_exercise**]**

THEN[PA_level=Low] **acc: (0.89212)**

2.4.3 รูปแบบของการแสดงกฎความสัมพันธ์ด้วยอัลกอริทึมเทอเชียส

IF [Sex=Male&&Exercise=Exercise**]**

THEN[PA_level=High] **conf: (0.683)**

ซึ่งสามารถสรุปได้ว่า การหากฎความสัมพันธ์ด้วยทั้งสามอัลกอริทึมให้รูปแบบของกฎความสัมพันธ์อยู่ในรูปของ ***IF…THEN…***เหมือนกัน แต่จะแตกต่างกันในส่วนเงื่อนไขของกฎและส่วนมาตรวัดประสิทธิภาพที่วิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรี และวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมเทอเชียสจะใช้ค่าความเชื่อมั่นเป็นมาตรวัดประสิทธิภาพ ส่วนวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทีพเอไพรออรีจะใช้ค่าความแม่นตรงเป็นมาตรวัดประสิทธิภาพ

2.5 การจัดอันดับของกฎความสัมพันธ์

การจัดอันดับของกฎความสัมพันธ์ หมายถึงการเรียงลำดับกฎความสัมพันธ์ตามค่าดัชนีชี้วัดความเหมาะสมโดยเรียงลำดับจากดัชนีชี้วัดที่มีค่าสูงไปจนถึงดัชนีชี้วัดที่มีค่าต่ำ จัดเป็นหนึ่งในกระบวนการคัดเลือกกฎที่มีคุณประโยชน์สำหรับการนำมาวิเคราะห์ความสัมพันธ์ของกฎนั้น ๆ สำหรับการจัดอันดับของกฎความสัมพันธ์จะทำให้ทราบว่ากฎความสัมพันธ์ใดที่สามารถบ่งชี้ได้ว่ากฎความสัมพันธ์ใดสามารถอธิบายคุณลักษณะของข้อมูลได้ดีที่สุด และมีงานวิจัยจำนวนไม่น้อยที่ได้นำวิธีการจัดอันดับของกฎความสัมพันธ์มาใช้ เช่นงานวิจัยของ (Mehdiet al., 2009) นำมาใช้ในส่วนของการจัดลำดับของกฎความสัมพันธ์ที่ได้จากData Envelopment Analysis(DEA)ที่ถูกพัฒนาโดย (Cook and Kress, 1990)โดยใช้วิธีที่เรียกว่าChen's method ซึ่งปรากฏในงานวิจัยของ (Mu, 2007) มาจัดอันดับกฎความสัมพันธ์

2.5.1 หลักการจัดอันดับของกฎความสัมพันธ์

หลักการจัดอันดับของกฎความสัมพันธ์คือ กฎความสัมพันธ์ที่จะมีค่าดัชนีชี้วัดที่สูงนั้นจะมีเกณฑ์ชี้วัดอยู่สองเกณฑ์ เกณฑ์ที่หนึ่งพิจารณาจากกฎความสัมพันธ์นั้นแล้วสามารถนำไปใช้ประโยชน์ได้จริง ซึ่งหมายถึงจะต้องเป็นกฎความสัมพันธ์ที่สามารถอธิบายคุณลักษณะของข้อมูลได้มากที่สุด และเกณฑ์ที่สองพิจารณาจากคลาสของกฎความสัมพันธ์ที่มีความน่าสนใจ ซึ่งนั่นหมายถึงจะต้องเป็นคลาสของกฎความสัมพันธ์ที่พบมากที่สุด และทั้งสองการพิจารณานี้จะต้องผ่านเกณฑ์ค่าความเชื่อมั่นและค่าความแม่นตรงที่ยอมรับได้ก่อน แล้วจึงนำมาพิจารณาการจัดอันดับของกฎความสัมพันธ์ วิธีพิจารณาการจัดอันดับได้ถูกนำมาใช้ในงานวิจัยจำนวนไม่น้อย ยกตัวอย่างเช่นงานวิจัยของ (Ramarajand Rameshkumar, 2009) ได้ให้ความสำคัญอย่างมากกับเกณฑ์การพิจารณาอันดับของกฎความสัมพันธ์โดยใช้อัลกอริทึมrankRule เป็นเกณฑ์ในการพิจารณาอันดับของกฎความสัมพันธ์

2.5.2 ประโยชน์ของการจัดอันดับกฎความสัมพันธ์

การจัดอันดับความสัมพันธ์ จะมีประโยชน์ในการอธิบายคุณลักษณะของกฎความสัมพันธ์ ซึ่งกฎความสัมพันธ์ที่ได้มีการจัดอันดับจะทำให้ง่ายต่อการวิเคราะห์ และบางครั้งอาจช่วยให้ทราบถึงกฎความสัมพันธ์ที่ซ่อนอยู่ อีกทั้งเมื่อนำมารวมกับเกณฑ์การพิจารณาในการปรากฏขึ้นของกฎความสัมพันธ์ใด ๆ อยู่บ่อยครั้ง จะทำให้สามารถช่วยยืนยันได้ชัดเจนขึ้นอีกว่า กฎความสัมพันธ์ดังกล่าว เหมาะที่จะนำมาวิเคราะห์กฎความสัมพันธ์ ในทางการแพทย์การอธิบายคุณลักษณะถ้าสามารถอธิบายได้ครอบคลุมและรัดกุมจะเป็นประโยชน์ต่อการวิเคราะห์อย่างมาก

2.6 การขจัดกฎความสัมพันธ์ที่ซ้ำซ้อน

การขจัดกฎความสัมพันธ์ที่ซ้ำซ้อน หมายถึงการลดจำนวนกฎความสัมพันธ์ที่สามารถอธิบายได้ด้วยกฎความสัมพันธ์อื่นที่ดีกว่า จัดเป็นหนึ่งในกระบวนการคัดเลือกกฎที่มีประโยชน์สำหรับการนำมาวิเคราะห์ความสัมพันธ์ของกฎนั้น ๆ สำหรับการขจัดกฎความสัมพันธ์ที่ซ้ำซ้อนจะทำให้ทราบว่ากฎความสัมพันธ์ใดที่สามารถอธิบายถึงคุณลักษณะได้ดีที่สุด และจะทำให้กฎความสัมพันธ์ที่ได้จากการค้นหากฎความสัมพันธ์มีความกระชับ (Compact) มากยิ่งขึ้น ที่ผ่านมามีงานวิจัยจำนวนไม่น้อยที่ได้นำวิธีการขจัดกฎความสัมพันธ์ที่ซ้ำซ้อนมาใช้ ยกตัวอย่างเช่น งานวิจัยของ (Davidet al., 2009) ได้นำมาใช้ในส่วนของการลดกฎความสัมพันธ์ที่ซ้ำซ้อน เพื่อให้กฎความสัมพันธ์ที่ได้นั้นมีความกระชับมากยิ่งขึ้น

2.6.1 หลักการขจัดกฎความสัมพันธ์ที่ซ้ำซ้อน

หลักการขจัดกฎความสัมพันธ์ที่ซ้ำซ้อนให้อยู่ในรูปความสัมพันธ์ที่ไม่ซ้ำซ้อนมีแนวคิดคือ กฎความสัมพันธ์ที่ซ้ำซ้อนนั้นจะเป็นกฎความสัมพันธ์ที่เป็นกฎความสัมพันธ์ย่อย(Sub set rule) ของกฎความสัมพันธ์ที่ใหญ่กว่า(Supper set rule) ดังนั้นกฎความสัมพันธ์ที่เป็นกฎความสัมพันธ์ย่อยจะต้องถูกตัดทิ้ง เนื่องมาจากกฎความสัมพันธ์ที่ใหญ่กว่านั้นสามารถอธิบายคุณลักษณะของข้อมูลได้ดีกว่า ในขณะที่มีค่าความเชื่อมั่นที่เท่า ๆ กัน งานวิจัยทางด้านขจัดกฎความสัมพันธ์ที่ซ้ำซ้อนยังไม่ปรากฏอยู่มาก เนื่องจากเป็นวิธีที่ค่อนข้างมีความละเอียดอ่อนในเรื่องของการพิจารณา อย่างไรก็ตามยังมีงานวิจัยของ (Huawenet al., 2011) ให้ความสำคัญในการขจัดกฎที่ซ้ำซ้อนโดยใช้อัลกอริทึม Closed rule-set เป็นเกณฑ์ในการพิจาณากฎความสัมพันธ์ที่ซ้ำซ้อน

2.6.2 ประโยชน์ของการขจัดกฎความสัมพันธ์ที่ซ้ำซ้อน

การขจัดกฎความสัมพันธ์ที่ซ้ำซ้อน จะทำให้ทราบว่ากฎความสัมพันธ์ใดสามารถอธิบายถึงคุณลักษณะได้ดีที่สุด ครอบคลุมที่สุด ส่งผลให้กฎความสัมพันธ์มีความกระชับยิ่งขึ้น ซึ่งอาจเกิดประโยชน์ในการนำกฎความสัมพันธ์ที่ได้มาทำการวิเคราะห์ต่าง ๆ ได้หลากหลายแนวทางและหลากหลายแง่คิด อย่างไรก็ตามการขจัดกฎความสัมพันธ์ที่ซ้ำซ้อนนั้น จะมีประโยชน์ต่อทางการแพทย์อย่างมากในด้านของการทราบถึงกฎความสัมพันธ์ที่เป็นสับเซต จะช่วยลดระยะเวลาในการวิเคราะห์กฎความสัมพันธ์ที่ไม่จำเป็นต้องนำมาวิเคราะห์ ซึ่งทำให้มีความรวดเร็วในการประมวลผล

2.7 รูปแบบของการจัดอันดับและการลดความซ้ำซ้อนของกฎความสัมพันธ์

ตารางที่ 2.2 ตัวอย่างข้อมูลโรคหอบหืดที่ใช้ในการจัดอันดับและขจัดความซ้ำซ้อนของกฎความสัมพันธ์

Age	Sex	Religion	Smoking	Exercise	Weight	Bodyfat	PA_level
38	Male	Buddhist	Never_smoker	Exercise	85.2	24.7	High
54	Male	Buddhist	Currents_smoker	No_exercise	57	24.5	Low
35	Male	Buddhist	Currents_smoker	Exercise	59.6	14.1	Moderate
45	Female	Buddhist	Never_smoker	Exercise	57	31.9	Moderate
43	Female	Buddhist	Never_smoker	No_exercise	49.8	31.5	Low
42	Male	Buddhist	Previous_smoker	Exercise	76.1	27.4	High
45	Female	Buddhist	Never_smoker	No_exercise	68.8	39	Low
44	Female	Buddhist	Never_smoker	Exercise	54.6	38.3	Moderate
41	Female	Islam	Never_smoker	Exercise	52.6	30.9	High
57	Male	Buddhist	Never_smoker	No_exercise	70.8	24.5	Low
41	Female	Buddhist	Never_smoker	Exercise	43	28.9	Moderate
47	Male	Buddhist	Never_smoker	Exercise	55.9	16.2	High

2.7.1 รูปแบบของการจัดอันดับกฎความสัมพันธ์

ข้อมูลจากตารางที่ 2.2เมื่อนำไปหากฎความสัมพันธ์แบบปกติได้กฎความสัมพันธ์ดังรูปที่ 2.8จากนั้นนำกฎความสัมพันธ์ที่ได้จากรูปที่ 2.8 มาทำการจัดอันดับของกฎความสัมพันธ์ ได้ดังรูปที่ 2.9

IF [Exercise=No_exercise]

THEN [PA_level=Low] **conf:(0.98)**

IF [Religion=Buddhist&& Exercise=No_exercise]

THEN [PA_level=Low] **conf:(0.98)**

IF [Smoking=Never_smoker&& Exercise=No_exercise]

THEN[PA_level=Low] **conf:(0.975)**

IF [Sex=Female&& Religion=Buddhist&& Exercise=Exercise]

THEN [PA_level=Moderate] **conf:(0.96)**

IF [Religion=Buddhist &&Smoking=Never_smoker&&Exercise=No_exercise]

THEN[PA_level=Low] **conf:(0.97)**

รูปที่ 2.8 ผลลัพธ์ของการค้นหากฎความสัมพันธ์พื้นฐานด้วยอัลกอริทึมเอไพรออรี

IF [Exercise=No_exercise]

THEN [PA_level=Low] **conf:(0.98)**

IF [Religion=Buddhist&& Exercise=No_exercise]

THEN [PA_level=Low] **conf:(0.98)**

IF [Smoking=Never_smoker&& Exercise=No_exercise]

THEN[PA_level=Low] **conf:(0.975)**

IF [Religion=Buddhist && Smoking=Never_smoker&&Exercise=No_exercise]

THEN[PA_level=Low] **conf:(0.97)**

IF [Sex=Female&& Religion=Buddhist&& Exercise=Exercise]

THEN [PA_level=Moderate] **conf:(0.96)**

รูปที่ 2.9จัดอันดับของกฎความสัมพันธ์ที่ได้จากอัลกอริทึมเอไพรออรี

2.7.2 รูปแบบของการขจัดกฎความสัมพันธ์ที่ซ้ำซ้อน

จากรูปที่ 2.9นำกฎความสัมพันธ์ที่ได้มาขจัดกฎความสัมพันธ์ที่ซ้ำซ้อน ได้ดังรูปที่ 2.10

IF [Exercise=No_exercise]

THEN [PA_level=Low] **conf:(0.98)**

IF [Sex=Female&& Religion=Buddhist&& Exercise=Exercise]

THEN [PA_level=Moderate] **conf:(0.96)**

รูปที่ 2.10จัดกฎความสัมพันธ์ให้อยู่ในรูปกฎความสัมพันธ์ที่ไม่ซ้ำซ้อน

จากรูปที่ 2.10จะได้กฎความสัมพันธ์แบบใหม่โดยกฎความสัมพันธ์จะเหลือเพียงสองอันดับ เนื่องจากกฎความสัมพันธ์ที่ถูกตัดไป เป็นกฎความสัมพันธ์ที่เป็นกฎความสัมพันธ์ย่อยของกฎความสัมพันธ์ทั้งสอง

2.8 มาตรวัดประสิทธิภาพกฎความสัมพันธ์

2.8.1 มาตรวัดค่าสนับสนุน

ค่าสนับสนุนของกฎ X แล้ว Y (X → Y)หมายถึงจำนวนทรานแซคชัน ที่มี X และ Y เกิดขึ้นร่วมกันเป็นส่วนย่อยของทรานแซคชันนั้น ๆมาตรวัดค่าสนับสนุนนี้ แสดงได้ดังสมการที่ 2-1 (Mohammed and Wagner, 2014)

$$\text{Support}(X \rightarrow Y) = \text{Support}(XY) \quad (2\text{-}1)$$

2.8.2 มาตรวัดประสิทธิภาพค่าความเชื่อมั่น

ค่าความเชื่อมั่นของกฎ (X → Y)หมายถึงจำนวนทรานแซคชัน ที่มี X และ Y เกิดขึ้นร่วมกันเป็นส่วนย่อยของทรานแซคชันนั้น ๆ หารด้วยค่าสนับสนุนของ X โดยผลลัพธ์จะมีค่าอยู่ในช่วงของ 0 ถึง 1มาตรวัดค่าความเชื่อมั่นนี้ แสดงได้ดังสมการที่ 2-2(Mohammed and Wagner, 2014)

$$\text{Conf}(X \rightarrow Y) = \frac{\text{Support}(XY)}{\text{Support}(X)} \quad (2\text{-}2)$$

2.8.3 มาตรวัดค่าความแม่นตรง

ค่าความแม่นตรง ของกฎ (X → Y) จะเป็นการรวมกันระหว่างค่าความเชื่อมั่นกับค่าสนับสนุนมาตรวัดค่าความแม่นตรงนี้ แสดงได้ดังสมการที่ 2-3(Scheffer, 2001)

$$\text{Accuracy} = \frac{\text{P}(\text{Conf}(\text{X} \rightarrow \text{Y})|\text{Support}(\text{X} \rightarrow \text{Y}))}{\text{Support}(\text{X})} \qquad (2\text{-}3)$$

2.8.4 ตัวอย่างการหาค่าสนับสนุน ค่าความเชื่อมั่น และค่าความแม่นตรง

ตารางที่ 2.3ตัวอย่างข้อมูลผู้ป่วยโรคหอบหืดที่นำมาใช้ในการวัดประสิทธิภาพ

Age	Sex	Religion	Smoking	Exercise	Weight	Bodyfat	PA_level
45	Female	Buddhist	Never_smoker	Exercise	57	31.9	Moderate
43	Female	Buddhist	Never_smoker	No_exercise	49.8	31.5	Low
42	Male	Buddhist	Previous_smoker	Exercise	76.1	27.4	High
38	Male	Buddhist	Never_smoker	Exercise	85.2	24.7	High
54	Male	Buddhist	Currents_smoker	No_exercise	57	24.5	Low
35	Male	Buddhist	Currents_smoker	Exercise	59.6	14.1	Moderate
45	Female	Buddhist	Never_smoker	No_exercise	68.8	39	Low
44	Female	Buddhist	Never_smoker	Exercise	54.6	38.3	Moderate
41	Female	Islam	Never_smoker	Exercise	52.6	30.9	High

กำหนดให้ X เป็น {Female,Never_smoker} และกำหนดให้ Y เป็น Moderate

จากสมการที่ 2-1 แทนค่าในสมการได้ดังนี้

Support({Female,Never_smoker} → {Moderate})

Support = 2

ฉะนั้นค่าสนับสนุนจะมีค่าเป็น 2

จากสมการที่ 2-2แทนค่าในสมการได้ดังนี้

Conf({Female,Never_smoker} → {Moderate})

Conf = 0.4

ฉะนั้นค่าความเชื่อมั่นจะมีค่าเป็น 0.4

จากสมการที่ 2-3 แทนค่าในสมการได้ดังนี้

$$\text{Accuracy} = \frac{P(\text{Conf}(\{\text{Female,Nerver\_smoker}\} \rightarrow \{\text{Moderate}\}) | \text{Support}(\{\text{Female,Nerver\_smoker}\} \rightarrow \{\text{Moderate}\}))}{\text{Support}(\{\text{Female,Nerver\_smoker}\})}$$

Accuracy = 0.16

ฉะนั้นค่าความแม่นตรงจะมีค่าเป็น 0.16

2.9 งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้องกับการค้นหาและจัดอันดับกฎความสัมพันธ์ได้มีการนำเสนออยู่จำนวนมากซึ่งประกอบไปด้วยงานวิจัยที่เสนอแนวคิดในการปรับปรุงกระบวนการด้วยเทคนิคต่าง ๆ เพื่อเพิ่มประสิทธิภาพในการค้นหากฎความสัมพันธ์ให้ได้กฎความสัมพันธ์ที่ดีที่สุด ผู้วิจัยได้ทำการศึกษาค้นคว้างานวิจัยที่มีความเกี่ยวข้องกับวิทยานิพนธ์นี้โดยสรุปรายละเอียดได้ดังต่อไปนี้

Jesminและคณะ (2013) ได้ใช้อัลกอริทึมการหากฎความสัมพันธ์พื้นฐานได้แก่ วิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรีวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทีพเอไพรออรีและวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมเทอเชียสมาประยุกต์ใช้ในงานวิจัยเพื่อค้นหากฎความสัมพันธ์ที่ดีที่สุด ซึ่งจะใช้ในการวิเคราะห์ข้อมูลผู้ป่วยโรคหัวใจเมื่อได้กฎความสัมพันธ์ของข้อมูลผู้ป่วยโรคหัวใจจากอัลกอริทึมการหากฎความสัมพันธ์พื้นฐานทั้ง 3 อัลกอริทึม นำมาเปรียบเทียบใน 2 ส่วนด้วยกัน ได้แก่ ประสิทธิภาพค่าความเชื่อมั่นและค่าแม่นตรงที่ได้จากกฎความสัมพันธ์ รวมไปถึงเวลาในการประมวลผลของอัลกอริทึมการค้นหากฎความสัมพันธ์พื้นฐานแต่ละอัลกอริทึมผลการทดลองปรากฏว่าวิธีการหากฎความสัมพันธ์พื้นฐานด้วยอัลกอริทึมเอไพรออรีสามารถหากฎความสัมพันธ์ที่มีประสิทธิภาพค่าความเชื่อมั่นได้มากกว่า 90% และเวลาที่ใช้ในการประมวลผลของอัลกอริทึมเอไพรออรีรวดเร็วที่สุดวิธีการหากฎความสัมพันธ์พื้นฐานด้วยอัลกอริทึมพรีดิคทีพเอไพรออรีสามารถหากฎความสัมพันธ์ที่มีประสิทธิภาพค่าความแม่นตรงได้มากกว่า 99% และเวลาที่ใช้ในการประมวลผลของอัลกอริทึมพรีดิคทีพเอไพรออรีช้าที่สุดและวิธีการหากฎความสัมพันธ์พื้นฐานด้วยอัลกอริทึมเทอเชียสสามารถหากฎความสัมพันธ์ที่มีประสิทธิภาพค่าความเชื่อมั่นได้มากกว่า 79%และเวลาที่ใช้ในการประมวลผลของอัลกอริทึมเทอเชียสดีกว่าวิธีการหากฎความสัมพันธ์พื้นฐานด้วยอัลกอริทึมพรีดิคทีพเอไพรออรี

Meghanaและคณะ(2013) ได้ใช้วิธีการหากฎความสัมพันธ์สองแบบ แบบที่หนึ่งเป็นวิธีการหากฎความสัมพันธ์พื้นฐานได้แก่ วิธีการหากฎความสัมพันธ์พื้นฐานด้วยอัลกอริทึมเอไพรออรีและวิธีการหากฎความสัมพันธ์พื้นฐานด้วยอัลกอริทึมพรีดิคทีพเอไพรออรี และแบบที่สองเป็นวิธีการหากฎความสัมพันธ์โดยใช้ฟิลเตอร์ในการคัดกรองข้อมูล และในงานวิจัยได้มีการนำฟิลเตอร์ Replace missing value มาใช้ร่วมกับวิธีการหากฎความสัมพันธ์พื้นฐานด้วยอัลกอริทึมเอไพรออรี

ซึ่งจะใช้ในการวิเคราะห์ข้อมูลการตรวจหาเนื้องอกในสมองมาประยุกต์ใช้ในงานวิจัยเพื่อค้นหากฎความสัมพันธ์ที่ดีที่สุด เมื่อได้กฎความสัมพันธ์ที่ทำการวิจัยจากข้อมูลการตรวจหาเนื้องอกในสมองจากวิธีการหากฎความสัมพันธ์ทั้ง 2 แบบจาก 3 อัลกอริทึมที่ใช้หากฎความสัมพันธ์ พบว่าทั้ง 3 อัลกอริทึมสามารถหากฎความสัมพันธ์ที่ดีที่สุดสำหรับข้อมูลการตรวจหาเนื้องอกในสมองซึ่งเป็นฐานข้อมูลขนาดใหญ่ได้ง่าย

LoboและSunita(2012) ได้ใช้วิธีการหากฎความสัมพันธ์สี่อัลกอริทึม ได้แก่ วิธีการหากฎความสัมพันธ์พื้นฐานด้วยอัลกอริทึมเอไพรออรี วิธีการหากฎความสัมพันธ์พื้นฐานด้วยอัลกอริทึมพรีดิคทีพเอไพรออรีวิธีการหากฎความสัมพันธ์พื้นฐานด้วยอัลกอริทึมเทอเชียสและอัลกอริทึม Filtered Associator เพื่อใช้ในการค้นหาวิธีการหากฎความสัมพันธ์ที่ดีที่สุดซึ่งจะทำการเปรียบเทียบประสิทธิภาพของวิธีการหากฎความสัมพันธ์ทั้งสี่อัลกอริทึม และนำเสนอผลของประสิทธิภาพ ในงานวิจัยนี้ใช้ข้อมูลจริงจากหลักสูตร Moodle ของวิทยาลัยสำหรับการทำนายการเลือกหลักสูตรการเรียนของนักเรียน (Course selection by student) จากผลการทดลองพบว่าวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรีให้ประสิทธิภาพที่ดีกว่าวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทีพเอไรออรีวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมเทอเชียสและอัลกอริทึม Filtered Associator สำหรับการทำนายการเลือกหลักสูตรการเรียนของนักเรียน

DivyaและLekha(2013) ได้ใช้วิธีการหากฎความสัมพันธ์พื้นฐานด้วยอัลกอริทึมเอไพรออรีเพื่อใช้ในการค้นหาวิธีการหากฎความสัมพันธ์ของชุดข้อมูลที่เกี่ยวกับการก่ออาชญากรรมของผู้หญิงซึ่งเป็นข้อมูลที่ได้มาจากฐานข้อมูลมาตรฐานมีการเก็บรวบรวมข้อมูลจริงจากศาลของ Sirsa ที่เก็บรวบรวมข้อมูลเกี่ยวกับอาชญากรรมบนการละลายพฤติกรรมทางจิตใจของผู้หญิงซึ่งจะทำการค้นหากฎความสัมพันธ์ที่ซ่อนอยู่ที่เป็นปัจจัยของการเกิดอาชญากรรม และหาการกระทำผิดจริงที่ซ่อนอยู่ ในการเปรียบเทียบประสิทธิภาพของวิธีการหากฎความสัมพันธ์จะเปรียบเทียบระหว่างวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรีและวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทีพเอไพรออรีซึ่งได้ผลสรุปว่าวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรีจะมีประสิทธิภาพที่ดีกว่าในด้านของความเชื่อมั่นและในด้านของเวลาในการประมวลผลที่รวดเร็วกว่าวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทีพเอไพรออรี

Ramaraj และ Rameskumar (2009) ได้ใช้วิธีการหากฎความสัมพันธ์ที่แตกต่างกันสามอัลกอริทึม ได้แก่วิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรีวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมเอฟพีโกรท(FP-growth)และวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมอีแคลท(Eclat)เพื่อใช้ในการค้นหากฎความสัมพันธ์ของชุดข้อมูลการเล่นหมากรุก (Chess) ซึ่งมีลักษณะ

ของข้อมูลเป็นข้อมูลขนาดเล็ก(Small dataset) ซึ่งแบ่งขั้นตอนการทำงานเป็นสองส่วน ส่วนที่หนึ่งทำการค้นหากฎความสัมพันธ์จากทรานแซคชันที่อยู่ภายในข้อมูลการเล่นหมากรุก ส่วนที่สองเมื่อได้รับผลลัพธ์จากส่วนที่หนึ่งนำมาพิจารณาความเข้มข้น (Concentrate)ของกฎความสัมพันธ์ที่ได้จากการค้นหากฎความสัมพันธ์ภายในข้อมูลหมากรุกและในงานวิจัยนี้ได้เสนอมาตรวัดใหม่ที่จะใช้สำหรับการจัดอันดับ (Rank) ของกฎความสัมพันธ์ซึ่งเรียกว่า Normalized discounted comulative gain (nDCG) การจัดอันดับกฎความสัมพันธ์จะช่วยให้แบ่งเบาภาระในการตัดสินใจเพื่อเลือกกฎความสัมพันธ์ใด ๆ มาใช้ประโยชน์ท้ายที่สุดงานวิจัยประสบความสำเร็จ ในแง่มุมของประสิทธิภาพสำหรับการพิจารณาการใช้งานกฎความสัมพันธ์จากวิธีการจัดอันดับความสัมพันธ์ที่ได้มีการเสนอขึ้นมา

จากการศึกษางานวิจัยที่เกี่ยวข้องพบว่าวิธีการค้นหากฎความสัมพันธ์แบบพื้นฐานได้แก่ วิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรีวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทีพเอไพรออรีและวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมเทอเชียสเป็นที่นิยมในการนำมาใช้หากฎความสัมพันธ์กับข้อมูลที่มีความสำคัญ เพื่อใช้ในการวิเคราะห์หรือตีความ เช่น ข้อมูลทางการแพทย์ ข้อมูลทางอาชญากรรม และข้อมูลด้านการศึกษา เป็นต้น ในแต่ละงานวิจัยก็มีแนวทางและเทคนิคเฉพาะของแต่ละงานวิจัย ซึ่งมีจุดมุ่งหมายเดียวกันนั่นคือการค้นหากฎความสัมพันธ์ที่ดีที่สุดด้วยเทคนิคของแต่ละงานวิจัย ซึ่งจะมีความเหมาะสมตามลักษณะของข้อมูลที่แตกต่างกันออกไปทั้งนี้ขึ้นอยู่กับการเลือกใช้แต่ละเทคนิคให้มีความเหมาะสม โดยงานวิจัยส่วนใหญ่จะเน้นการเปรียบเทียบประสิทธิภาพในแต่ละเทคนิคของวิธีการค้นหากฎความสัมพันธ์ และทำการวัดประสิทธิภาพกฎความสัมพันธ์จากค่าความความเชื่อมั่นและค่าความแม่นตรง แต่มีงานวิจัยส่วนน้อยที่มีการใช้เทคนิคการรวมกฎความสัมพันธ์ และการจัดอันดับกฎความสัมพันธ์ ผสมผสานกับการขจัดกฎความสัมพันธ์ที่ซ้ำซ้อนสำหรับการนำมาใช้กับข้อมูลทางด้านการแพทย์ เนื่องจากกระบวนการมีความซับซ้อนและต้องมีการวิเคราะห์ในเชิงลึก ทำให้งานวิจัยนี้ได้เสนอเทคนิคการค้นหากฎความสัมพันธ์โดยใช้การรวมกฎความสัมพันธ์ การจัดอันดับกฎความสัมพันธ์ และการขจัดกฎความสัมพันธ์ที่ซ้ำซ้อนเพื่อใช้กับข้อมูลทางการแพทย์ที่มีลักษณะของข้อมูลผสมผสานกันระหว่างข้อมูลที่เป็นตัวเลขกับข้อมูลที่เป็นข้อความงานวิจัยนี้ได้เสนอมาตรวัดประสิทธิภาพใหม่คือ CAA (Confidence and Accuracy) พร้อมทั้งเสนอมาตรวัดค่าดัชนีความถี่ของการรวมกฎ FMR (Frequency of merged rules)ในการวัดประสิทธิภาพของข้อมูลทางการแพทย์ สาระสำคัญในงานวิจัยนี้เมื่อเปรียบเทียบกับงานวิจัยอื่นสรุปได้ดังตารางที่ 2.4

ตารางที่ 2.4สรุปเปรียบเทียบงานวิจัยที่เกี่ยวข้องกับการค้นหาและจัดอันดับกฎความสัมพันธ์เพื่อวิเคราะห์ปัจจัยเสี่ยงที่นำไปสู่การเกิดโรค

กระบวนการทำงาน	งานวิจัยที่เกี่ยวข้อง					
	ก	ข	ค	ง	จ	ฉ*
อัลกอริทึมการค้นหากฎความสัมพันธ์แบบพื้นฐาน						
Apriori	✓	✓	✓	✓	✓	✓
Predictive Apriori		✓	✓	✓	✓	✓
Tertius		✓			✓	✓
FP-growth	✓					
Eclat	✓					
วิธีการเตรียมข้อมูลก่อนการหากฎความสัมพันธ์						
Replace missing value			✓			
Filtered associator		✓				
Remove transaction						✓
Discretization						✓
วิธีการจัดการกฎความสัมพันธ์						
Merge rules						✓
Ranking of association rules	✓					✓
Remove redundant rules						✓
มาตรวัดประสิทธิภาพกฎความสัมพันธ์						
Support	✓	✓	✓	✓	✓	✓
Confidence	✓	✓	✓	✓	✓	✓
Lift	✓					
Accuracy		✓	✓	✓	✓	✓
Confidence and Accuracy (CAA)						✓
Frequency of merged rules (FMR)						✓

ตารางที่ 2.4สรุปเปรียบเทียบงานวิจัยที่เกี่ยวข้องกับการค้นหาและจัดอันดับกฎความสัมพันธ์เพื่อวิเคราะห์ปัจจัยเสี่ยงที่นำไปสู่การเกิดโรค(ต่อ)

กระบวนการทำงาน	งานวิจัยที่เกี่ยวข้อง					
	ก	ข	ค	ง	จ	ฉ*
ข้อมูลที่ใช้ในงานวิจัย						
Medical dataset			✓		✓	✓
Other dataset	✓	✓		✓		

หมายเหตุ งานวิจัยที่เกี่ยวข้อง ประกอบด้วย

ก แทนงานวิจัยของ Ramaraj และ Rameskumar (2009)

ข แทนงานวิจัยของ Lobo และSunita(2012)

ค แทนงานวิจัยของ Meghanaและคณะ(2013)

ง แทนงานวิจัยของ DivyaและLekha(2013)

จแทนงานวิจัยของ Jesminและคณะ (2013)

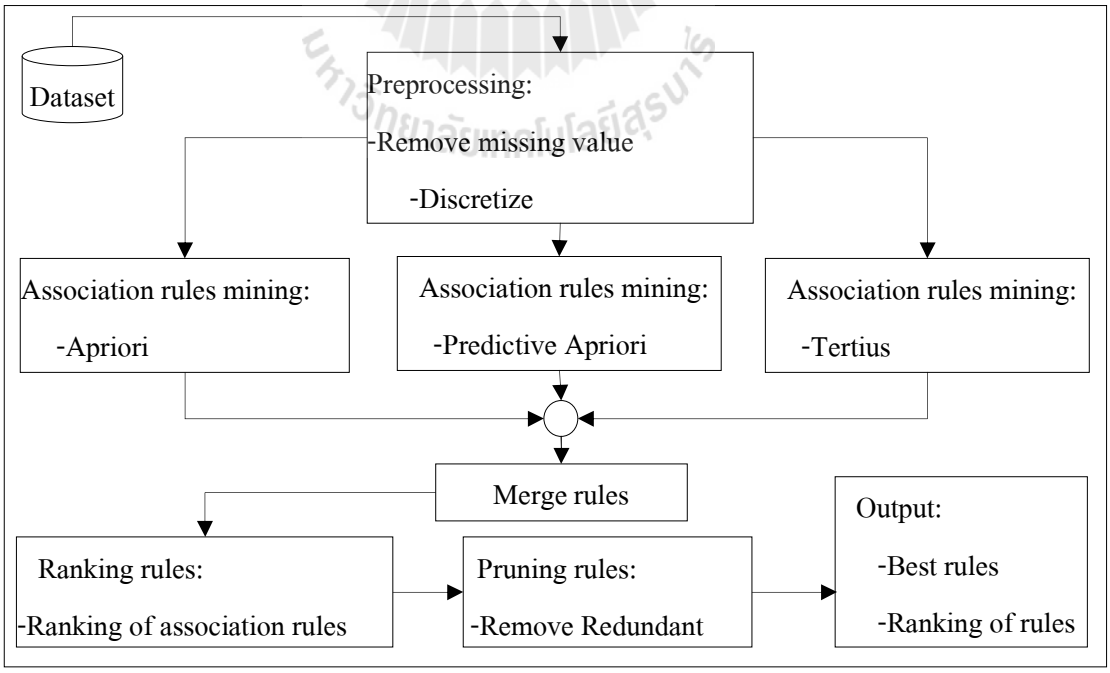
ฉ*แทนงานวิจัยของ วิทยานิพนธ์ฉบับนี้

บทที่ 3
วิธีดำเนินการวิจัย

การวิจัยนี้มีวัตถุประสงค์เพื่อเสนอแนวทางการเพิ่มกระบวนการจัดการข้อมูลที่เป็นลักษณะตัวเลขด้วยวิธีการแบ่งช่วงข้อมูลการรวมกฎความสัมพันธ์ รวมไปถึงการจัดอันดับของความสัมพันธ์และขจัดกฎความสัมพันธ์ที่ซ้ำซ้อน เพื่อให้ได้กฎความสัมพันธ์ที่ดีที่สุด ในบทนี้จะกล่าวถึง วิธีการวิจัย กระบวนการต่าง ๆ ของการวิจัย และเครื่องมือที่ใช้ในการวิจัย โดยมีรายละเอียดดังนี้

3.1 กรอบแนวคิดของการวิจัย

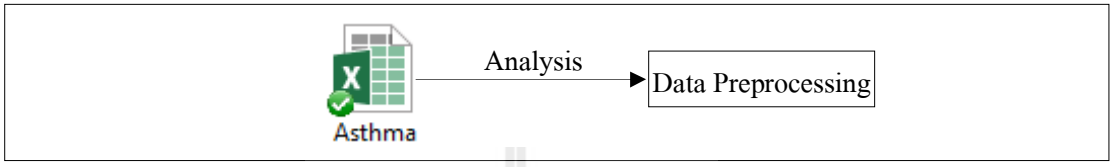
กรอบแนวคิดหลักของการวิจัยคือ เสนอแนวทางในการเพิ่มกระบวนการจัดการข้อมูลที่เป็นลักษณะตัวเลขด้วยวิธีการแบ่งช่วงข้อมูลอีกทั้งทำการรวมกฎความสัมพันธ์และรวมถึงการจัดอันดับของความสัมพันธ์พร้อมทั้งขจัดกฎความสัมพันธ์ที่ซ้ำซ้อนเพื่อให้ได้กฎความสัมพันธ์ที่ดีที่สุด โดยกรอบแนวคิดของการวิจัยแสดงดังรูปที่3.1



รูปที่3.1กรอบแนวคิดของการวิจัย

จากรูปที่ 3.1แสดงกรอบแนวคิดของการวิจัยสามารถอธิบายรายละเอียดด้วยตัวอย่างข้อมูลขนาดเล็ก โดยมีขั้นตอนดังต่อไปนี้

ขั้นตอนที่ 1 นำข้อมูลจากฐานข้อมูลเข้ามาวิเคราะห์ในกระบวนการจัดการข้อมูลก่อนการหากฎความสัมพันธ์ โดยจะใช้ตัวอย่างของข้อมูลผู้ป่วยโรคหอบหืด ดังรูปที่ 3.2



รูปที่ 3.2 นำเข้าข้อมูลจากฐานข้อมูลโรคหอบหืด

ขั้นตอนที่ 2 ขจัดข้อมูลที่ไม่สามารถบ่งชี้ได้ชัดเจน ด้วยวิธีการลบทรานแซคชันที่มีการระบุข้อมูลที่ไม่แน่ชัดแสดงตัวอย่างดังรูปที่ 3.3

ตารางข้อมูลก่อนการลบทรานแซคชันตารางข้อมูลหลังการลบทรานแซคชัน

Marital_status	Religion	PA_level
Married	buddhist	Moderate
Divorced		High
Married	buddhist	Low

Marital_status	Religion	PA_level
Married	Buddhist	Moderate
Married	Buddhist	Low

รูปที่ 3.3 ตัวอย่างข้อมูลโรคหอบหืดก่อนและหลังการลบทรานแซคชันที่มีข้อมูลไม่แน่ชัด

ขั้นตอนที่ 3 แบ่งช่วงข้อมูลด้วยวิธีการจัดช่วงข้อมูลที่เป็นตัวเลข เฉพาะแอททริบิวต์ที่มีลักษณะเป็นตัวเลขแสดงตัวอย่างดังรูปที่ 3.4

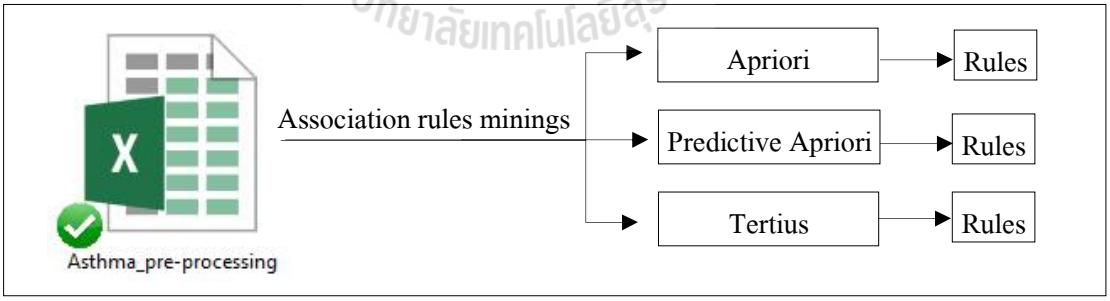
ตารางข้อมูลก่อนการจัดช่วงข้อมูลตารางข้อมูลหลังการจัดช่วงข้อมูล

Weight	Bodyfat	PA_level
62.4	33	Moderate
61.1	36.1	High
47.5	27.6	Low
57.1	31.6	Moderate

Weight	Bodyfat	PA_level
60.0-65.0	30.0-35.0	moderate
60.0-65.0	35.0-40.0	High
45.0-50.0	25.0-30.0	Low
55.0-60.0	30.0-35.0	moderate

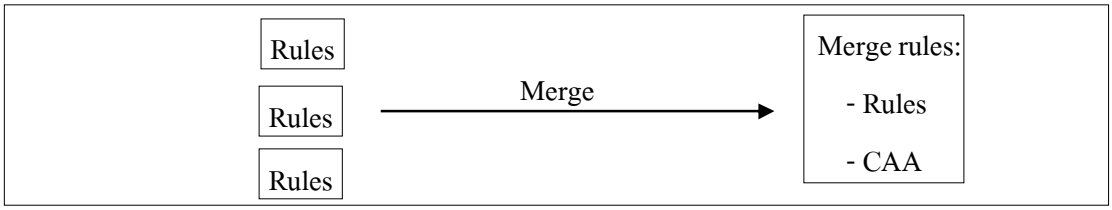
รูปที่ 3.4 ข้อมูลโรคหอบหืดก่อนและหลังการแบ่งช่วงข้อมูลที่มีลักษณะเป็นข้อมูลตัวเลข

ขั้นตอนที่ 4 นำข้อมูลโรคหอบหืดจากการวิเคราะห์ในกระบวนการจัดการข้อมูลมาหากฎความสัมพันธ์ของข้อมูลโรคหอบหืดด้วยวิธีการหากฎความสัมพันธ์ทั้งสามอัลกอริทึมได้แก่ วิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรีวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทีพเอไพรออรีและวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมเทอเชียส แสดงดังรูปที่ 3.5



รูปที่ 3.5 นำข้อมูลโรคหอบหืดมาหากฎความสัมพันธ์ทั้งสามอัลกอริทึม

ขั้นตอนที่ 5จากขั้นตอนที่ 4จะได้ชุดของกฎความสัมพันธ์พื้นฐานจากทั้ง 3 วิธีการหากฎความสัมพันธ์ จากนั้นทำการรวมกฎความสัมพันธ์จากชุดของกฎความสัมพันธ์ทั้ง 3 พร้อมทั้งกำหนดมาตรวัดใหม่ที่เรียกว่า CAA แนวคิดแสดงดังรูปที่ 3.6



รูปที่ 3.6รวมกฎความสัมพันธ์จากทั้งสามอัลกอริทึม

มาตรวัด CAAสามารถคำนวณได้จากสมการที่ 3-1

$$CAA = \left| \frac{Conf + Acc}{n_{rule}} \right| \qquad (3\text{-}1)$$

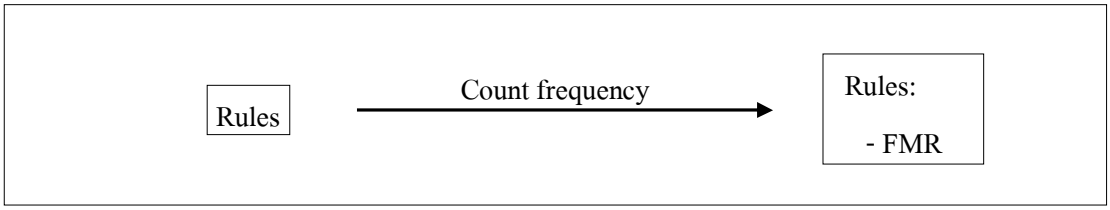
CAA คือ มาตรวัดใหม่ของการรวมกฎความสัมพันธ์

Conf คือ ผลรวมค่าความเชื่อมั่นของกฎความสัมพันธ์ทุกกฎที่นำมารวม

Acc คือ ผลรวมค่าความแม่นตรงของกฎความสัมพันธ์ทุกกฎที่นำมารวม

n_{rule} คือ จำนวนกฎที่นำมารวมกฎความสัมพันธ์

ขั้นตอนที่ 6จากขั้นตอนที่ 5จะได้ชุดของกฎความสัมพันธ์ที่ผ่านการรวมกฎความสัมพันธ์พร้อมกับมาตรวัด CAA จากนั้นทำการคำนวณเพื่อทำการหาค่าดัชนีความถี่ของการรวมกฎเรียกว่ามาตรวัด Frequency of Merged Rules หรือFMRแสดงแนวคิดดังรูปที่ 3.7



รูปที่ 3.7การหาค่าดัชนีความถี่ของการรวมกฎ

มาตรวัด FMR สามารถคำนวณได้จากสมการที่ 3-2

$$FMR = Count(n_{rule}) \quad (3\text{-}2)$$

FMR คือ มาตรวัดใหม่ของการหาค่าดัชนีความถี่ของการรวมกฎ

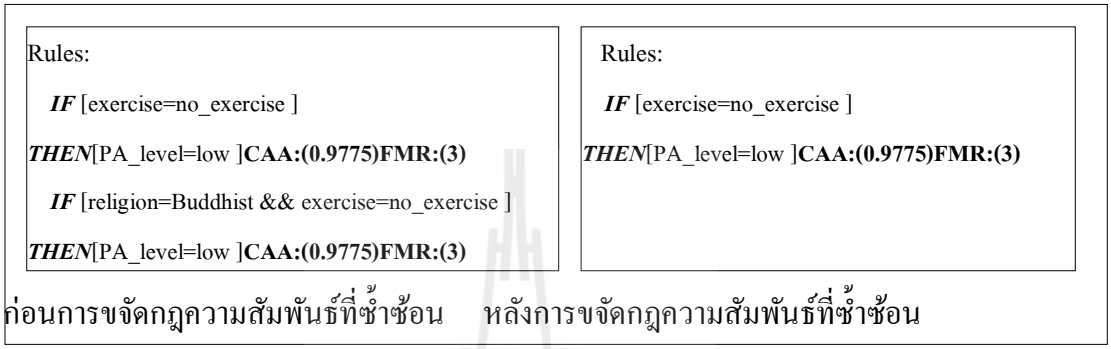
n_{rule} คือ จำนวนกฎที่นำมารวมกฎความสัมพันธ์

ขั้นตอนที่ 7จากขั้นตอนที่ 6จะได้กฎความสัมพันธ์ที่ผ่านการรวมกฎความสัมพันธ์พร้อมกับมาตรวัด CAA และ FMRจากนั้นทำการจัดอันดับกฎความสัมพันธ์จากกฎความสัมพันธ์ทั้งหมด โดยพิจารณาการจัดอันดับจาก 2 เงื่อนไข เงื่อนไขหลักคือ เป็นกฎความสัมพันธ์ที่อธิบายคุณลักษณะมากที่สุดซึ่งจะต้องมีค่ามาตรวัด CAAมากที่สุด และ FMR สูงที่สุดด้วย เงื่อนไขรอง เมื่อเงื่อนไขแรกมีมากกว่าหนึ่งตัวที่อยู่ลำดับเดียวกัน จะพิจารณาจากจำนวนของคลาสที่ปรากฏในกฎความสัมพันธ์ทั้งหมด จะใช้จำนวนคลาสเป็นน้ำหนักในการพิจารณาตัวอย่างดังรูปที่ 3.8 โดยพิจารณาจากกฎความสัมพันธ์ 2 กฎ

ก่อนการจัดอันดับกฎความสัมพันธ์	หลังการจัดอันดับกฎความสัมพันธ์
Rules: ***IF*** [exercise=no_exercise] ***THEN***[PA_level=low]**CAA:(0.9775)FMR:(2)** ***IF***[sex=female religion=buddhist exercise=exercise] ***THEN***[PA_level=moderate] **CAA:(0.9775)FMR:(3)**	Rules: ***IF***[sex=female religion=buddhist exercise=exercise] ***THEN***[PA_level=moderate]**CAA:(0.9775)FMR:(3)** ***IF***[exercise=no_exercise] ***THEN***[PA_level=low] **CAA:(0.9775)FMR:(2)**

รูปที่ 3.8 รูปแบบการจัดอันดับกฎความสัมพันธ์

ขั้นตอนที่ 8จากขั้นตอนที่ 7จะได้กฎความสัมพันธ์ที่ผ่านการจัดอันดับกฎความสัมพันธ์ จากนั้นทำการขจัดกฎความสัมพันธ์ที่ซ้ำซ้อน โดยพิจารณาการจัดอันดับจากกฎความสัมพันธ์ ถ้ากฎความสัมพันธ์ A เป็นสับเซตของกฎความสัมพันธ์ B โดยที่กฎความสัมพันธ์ทั้งสองจะต้องมีคลาสที่เหมือนกันและมีค่า CAAเท่ากัน รวมถึง FMR ของกฎความสัมพันธ์สับเซตจะต้องมีค่าน้อยกว่าหรือเท่ากับ กฎความสัมพันธ์ซุปเปอร์เซตดังนั้นกฎความสัมพันธ์ A จะถูกขจัด ดังรูปที่ 3.9



รูปที่ 3.9 รูปแบบการขจัดกฎความสัมพันธ์ที่ซ้ำซ้อน

3.2 การออกแบบอัลกอริทึม

3.2.1 อัลกอริทึมลบข้อมูลที่ไม่สามารถระบุได้ชัดเจน

ขั้นตอนการทำงานของอัลกอริทึมการหากฎความสัมพันธ์ในแง่มุมของข้อมูลที่ไม่สามารถระบุได้ชัดเจนแสดงดังรูปที่ 3.10

```
Algorithm Remove_transaction
Input: Dataset, D.
Output: NewDataset ,NewD.
1)      Read_data = read(D)
2)      Initial i = 1
3)      While (i<=Read_data.Transaction(size))
4)          Data[i] = Transaction(Read_data[i])
5)          If (Data[i].Empty() == True)
6)              Remove(Transaction[i])
7)              NewD = Data[i].Transaction[i]
8)              i = i+1
9)      END While Loop
10)     Return (NewD)
11)     END
```

รูปที่ 3.10รหัสเทียมของอัลกอริทึมการจัดการข้อมูลที่ไม่สามารถระบุได้ชัดเจน

จากรูปที่ 3.10สามารถอธิบายขั้นตอนการทำงานของอัลกอริทึมได้ดังนี้

อัลกอริทึม Remove_transaction

อินพุต ตัวแปรD แทน Dataset ของข้อมูลใด ๆ

เอ้าพุตท์ ตัวแปร NewDแทน NewDatasetซึ่งเป็นข้อมูลใหม่

ขั้นที่ 1 อ่านข้อมูลจากฐานข้อมูลนำมาเก็บไว้ที่ตัวแปร Read_data

ขั้นที่ 2กำหนดค่าให้ตัวแปร iมีค่าเป็น 1

ขั้นที่ 3-9 พิจารณา i เมื่อ iมีค่าน้อยกว่าขนาดของทรานแซคชันในฐานข้อมูลที่พิจารณา

การทำงานใน While Loop

- อ่านข้อมูลทีละ Transactionเริ่มอ่านข้อมูลที่ transaction ลำดับที่ iเก็บไว้ที่ตัวแปร Data[i] วนการทำงานตามจำนวน transaction

-ถ้า Data[i] มีค่าว่างให้ทำการ ลบ transaction ลำดับที่ i

-เก็บข้อมูลของฐานข้อมูลใหม่ลงตัวแปร NewD

-ให้เพิ่มค่า iขึ้น 1 ค่า

จบการทำการ While Loop

ขั้นที่ 10คืนค่า NewD

จบการทำงาน

3.2.2 อัลกอริทึมแบ่งช่วงข้อมูลที่เป็นตัวเลข

ขั้นตอนการทำงานของอัลกอริทึมการจัดการข้อมูลก่อนการหากฎความสัมพันธ์ในแง่มุมของข้อมูลที่อยู่ในลักษณะตัวเลข(Numeric data type)แสดงดังรูปที่ 3.11

```
Algorithm Discretize
Input: Dataset, D.
Output: NewDatasetDiscretize,NewDataDisc.
1)  Read_data = read(D)
2)  Initial i = 1
3)  While (i<=Read_data.Attribute(size))
4)          Type[i] = AttributeCheckType(Read_data[i])
5)          If (Type[i] == numeric)
6)                  Attribute[i].Discretize()
7)                  NewDataDisc = Data[i].Attribute[i]
8)                  i = i+1
9)  END While Loop
10) Return(NewDataDisc)
11) END
```

รูปที่ 3.11 รหัสเทียมของอัลกอริทึมการจัดการข้อมูลที่มีลักษณะเป็นตัวเลข

จากรูปที่ 3.11สามารถอธิบายขั้นตอนการทำงานของอัลกอริทึมได้ดังนี้

อัลกอริทึม Discretize

อินพุต ตัวแปร D แทน Dataset ของข้อมูลใด ๆ

เอ้าพุตท์ ตัวแปร NewDataDiscแทน NewDatasetDiscretizeซึ่งเป็นข้อมูลใหม่

ขั้นที่ 1 อ่านข้อมูลจากฐานข้อมูลนำมาเก็บไว้ที่ตัวแปร Read_data

ขั้นที่ 2 กำหนดค่าให้ตัวแปร iมีค่าเป็น 1

ขั้นที่ 3-9 พิจารณา i เมื่อ iมีค่าน้อยกว่าขนาดของแอทริบิวต์ในฐานข้อมูลที่พิจารณา

การทำงานใน While Loop

- อ่านข้อมูลของAttributeเริ่มอ่านข้อมูลที่ Attribute ลำดับที่ iเก็บType ของ Attributeเก็บไว้ที่ตัวแปรType[i]วนการทำงานตามจำนวน Attribute

-ถ้า Type[i]มีค่าเป็น numericทำการจัดช่วงข้อมูลของ attribute ลำดับที่ i

-เก็บข้อมูลของฐานข้อมูลใหม่ลงตัวแปร NewDataDisc

-ให้เพิ่มค่า iขึ้น 1 ค่า

จบการทำการ While Loop

ขั้นที่ 10คืนค่า NewDataDisc

จบการทำงาน

3.2.3 อัลกอริทึมรวมกฎความสัมพันธ์

ขั้นตอนการทำงานของอัลกอริทึมการจัดการข้อมูลหลังการหากฎความสัมพันธ์ในแง่มุมของการรวมกฎความสัมพันธ์แสดงดังรูปที่ 3.12

Algorithm MergeRules

Input: Rules, R.

Output: NewRules ,NewR.

1) Read_rules = read(R)
2) Initial FMR = 1 , i =1
3) While (i<=Read_rules.size))
4) Case 1 ((Ruleof(APRIORI) equal (Rueof(PREDICTIVE_APRIORI))

   FMR = FMR + 1, NumofRule = rules.consider()

   $\text{CAA} = \frac{Confidence(APRIORI) + Accuracy(\text{PREDICTIVE\_APRIORI})}{\text{NumofRule}}$

   NewR = string.concat (SplitRuleof(APRIORI), FMR, CAA) , i = i+1
5) Case 2 ((Ruleof(APRIORI) equal (Rueof(TERTIUS))

   FMR = FMR + 1, NumofRule = rules.consider()

   $\text{CAA} = \frac{Confidence(APRIORI) + Confidence(\text{TERTIUS})}{\text{NumofRule}}$

   NewR = string.concat (SplitRuleof(TERTIUS), FMR, CAA) , i = i+1
6) Case 3 ((Ruleof(PREDICTIVE_APRIORI) equal (Rueof(TERTIUS))

   FMR = FMR + 1, NumofRule = rules.consider()

   $\text{CAA} = \frac{Accuracy(PREDICTIVE\_APRIORI) + Confidence(\text{TERTIUS})}{\text{NumofRule}}$

   NewR = string.concat (SplitRuleof(PREDICTIVE_APRIORI), FMR, CAA) , i = i+1
7) Case 4 ((Ruleof(APRIORI) equal (Ruleof(PREDICTIVE_APRIORI) equal (Rueof(TERTIUS))

   FMR = FMR + 2, NumofRule = rules.consider()

   $\text{CAA} = \frac{Confidence(APRIORI) + Accuracy(\text{PREDICTIVE\_APRIORI})(CAA*2) + Confidence(TERTIUS)}{\text{NumofRule}}$

   NewR = string.concat(SplitRuleof(PREDICTIVE_APRIORI), FMR, CAA) , i = i+1
8) Other Case FMR = FMR,CAA = Confidence(APRIORI)

   NewR = string.concat (SplitRuleof(APRIORI), FMR, CAA) , i = i+1
9) END While Loop
10) Return(NewR)
11) END

รูปที่ 3.12รหัสเทียมของอัลกอริทึมการจัดการข้อมูลด้วยการรวมกฎความสัมพันธ์

จากรูปที่ 3.12สามารถอธิบายขั้นตอนการทำงานของอัลกอริทึมได้ดังนี้

อัลกอริทึม MergeRules

อินพุต ตัวแปร Rแทน Rulesของกฎความสัมพันธ์

เอ้าพุตท์ ตัวแปร NewRulesแทน NewRซึ่งเป็นกฎความสัมพันธ์ใหม่

ขั้นที่ 1 อ่านกฎความสัมพันธ์นำมาเก็บไว้ที่ตัวแปร Read_rules

ขั้นที่ 2 กำหนดค่าให้ตัวแปร FMRมีค่าเป็น 1 และกำหนดค่าให้ตัวแปร iมีค่าเป็น 1

ขั้นที่ 3-9การทำงานใน While Loopวนการทำงาน ตามจำนวนของกฎความสัมพันธ์

Case1 -ถ้ากฎความสัมพันธ์ Aprioriเหมือนกฎความสัมพันธ์ Predictive Apriori

FMR เพิ่มขึ้น 1 ค่า, คำนวณค่า NumOfRuleและ คำนวณค่า CAA

เก็บกฎความสัมพันธ์ใหม่ลง NewRและเพิ่มค่าi ขึ้น 1 ค่า

Case 2 -ถ้ากฎความสัมพันธ์ Apriori เหมือนกฎความสัมพันธ์ Tertius

FMR เพิ่มขึ้น 1 ค่า, คำนวณค่า NumOfRuleและคำนวณค่า CAA

เก็บกฎความสัมพันธ์ใหม่ลง NewRและเพิ่มค่าi ขึ้น 1 ค่า

Case 3 -ถ้ากฎความสัมพันธ์ Predictive Aprioriเหมือนกฎความสัมพันธ์ Tertius

FMR เพิ่มขึ้น 1 ค่า, คำนวณค่า NumOfRuleและคำนวณค่า CAA

เก็บกฎความสัมพันธ์ใหม่ลง NewRและเพิ่มค่าiขึ้น 1 ค่า

Case 4 -ถ้ากฎความสัมพันธ์ Apriori เหมือนกฎความสัมพันธ์ Predictive Aprioriและเหมือน กฎความสัมพันธ์ Tertius

FMR เพิ่มขึ้น 2ค่า, คำนวณค่า NumOfRuleและคำนวณค่า CAA

เก็บกฎความสัมพันธ์ใหม่ลง NewRและเพิ่มค่าiขึ้น 1 ค่า

กรณีอื่น ๆ

FMR = FMR และคำนวณค่า CAA

เก็บกฎความสัมพันธ์ใหม่ลง NewRและเพิ่มค่าiขึ้น 1 ค่า

จบการทำการ While Loop

ขั้นที่ 10คืนค่า NewR

จบการทำงาน

3.2.4 อัลกอริทึมจัดอันดับกฎความสัมพันธ์

ขั้นตอนการทำงานของอัลกอริทึมการจัดการข้อมูลหลังการหากฎความสัมพันธ์แง่มุมของการจัดอันดับกฎความสัมพันธ์แสดงดังรูปที่ 3.13

```
Algorithm RankingRules
Input: Rules, R.
Output: NewRank, NewRk.
1)  Read_rules = read(R)
2)  Initial i = 1
3)    While (i<=Read_rules.size))
4)          If (Read_rules.split(FMR) == 3)
5)                If ((Read_rules.split(CAA) == (Read_rules.split.Maximum(CAA))
6)                      NewRank = PushRule() , i = i + 1
7)                Else   PopRule()
8)          Else If ((Read_rules.split(FMR) == 3).Empty())
9)                If (Read_rules.split(FMR) == 2)
10)                     If ((Read_rules.split(CAA) == (Read_rules.split.Maximum(CAA))
11)                     NewRank = PushRule() , i = i + 1
12)               Else   PopRule()
13)         Else If ((Read_rules.split(FMR) == 2).Empty())
14)               If (Read_rules.split(FMR) == 1)
15)                     If ((Read_rules.split(CAA) == (Read_rules.split.Maximum(CAA))
16)                           NewRank = PushRule() , i = i + 1
17)                     Else   PopRule()
18) END While Loop
19) Return(NewRk)
20) END
```

รูปที่ 3.13รหัสเทียมของอัลกอริทึมการจัดการข้อมูลด้วยการจัดอันดับกฎความสัมพันธ์

จากรูปที่ 3.13สามารถอธิบายขั้นตอนการทำงานของอัลกอริทึมได้ดังนี้

อัลกอริทึม RankingRules

อินพุต ตัวแปร Rแทน Rulesของกฎความสัมพันธ์ที่ผ่านการรวมกฎความสัมพันธ์

เอ้าพุตท์ ตัวแปร NewRankแทน NewRkซึ่งเป็นกฎความสัมพันธ์ใหม่

ขั้นที่ 1 อ่านข้อมูลกฎความสัมพันธ์นำมาเก็บไว้ที่ตัวแปร Read_rules

ขั้นที่ 2กำหนดค่าให้ตัวแปร iมีค่าเป็น 1

ขั้นที่ 3-17 การทำงานใน While Loop

-วนการทำงาน ตามจำนวนของกฎความสัมพันธ์ที่ผ่านการรวมกฎความสัมพันธ์

-ถ้าค่า FMR ที่ดึงมาจากกฎความสัมพันธ์ มีค่าเท่ากับ 3จริง

ถ้าค่า CAA ที่ดึงมาจากกฎความสัมพันธ์มีค่าสูงที่สุด

จริงเก็บกฎความสัมพันธ์ลง NewRkและเพิ่มค่า iขึ้น 1 ค่า

ถ้าค่า CAA ที่ดึงมาจากกฎความสัมพันธ์มีค่าสูงที่สุด

ไม่จริงพิจารณากฎความสัมพันธ์ลำดับถัดไป

-ถ้าค่า FMR ที่ดึงมาจากกฎความสัมพันธ์ที่เท่ากับ 3 ไม่มีอยู่ จริง

ถ้าค่า FMR ที่ดึงมาจากกฎความสัมพันธ์ มีค่าเท่ากับ 2 จริง

ถ้าค่า CAA ที่ดึงมาจากกฎความสัมพันธ์มีค่าสูงที่สุด

จริงเก็บกฎความสัมพันธ์ลง NewRkและเพิ่มค่า iขึ้น 1 ค่า

ถ้าค่า CAA ที่ดึงมาจากกฎความสัมพันธ์มีค่าสูงที่สุด

ไม่จริงพิจารณากฎความสัมพันธ์ลำดับถัดไป

-ถ้าค่า FMR ที่ดึงมาจากกฎความสัมพันธ์ที่เท่ากับ 2 ไม่มีอยู่ จริง

ถ้าค่า FMR ที่ดึงมาจากกฎความสัมพันธ์ มีค่าเท่ากับ 1 จริง

ถ้าค่า CAA ที่ดึงมาจากกฎความสัมพันธ์มีค่าสูงที่สุด

จริงเก็บกฎความสัมพันธ์ลง NewRkและเพิ่มค่า iขึ้น 1 ค่า

ถ้าค่า CAA ที่ดึงมาจากกฎความสัมพันธ์มีค่าสูงที่สุด

ไม่จริงพิจารณากฎความสัมพันธ์ลำดับถัดไป

จบการทำการ While Loop

ขั้นที่ 19คืนค่า NewRk

จบการทำงาน

3.2.5 อัลกอริทึมขจัดกฎความสัมพันธ์ที่ซ้ำซ้อน

ขั้นตอนการทำงานมของอัลกอริทึมการจัดการข้อมูลหลังการหากฎความสัมพันธ์แง่มุมของการขจัดกฎความสัมพันธ์ที่ซ้ำซ้อนแสดงดังรูปที่ 3.14

```
Algorithm RemoveRedundantRules
Input: Rules, R.
Output: NewRules, NewR.
1) Read_rules = read(R)
2) Initial j = 1
3) While (j<=Read_rules.size))
4)          Attribute_Rule_before[j] = Read_rules.split(Attribute)[j]
5)          FMR_Rule_before[j] = Read_rules.split(FMR)[j]
6)          CAA_Rule_before[j] = Read_rules.split(CAA)[j] , j = j + 1
7)          Attribute_Rule_after[j] = Read_rules.split(Attribute)[j]
8)          FMR_Rule _after[j] = Read_rules.split(FMR)[j]
9)          CAA_Rule_after[j] = Read_rules.split(CAA)[j]
10)         If (Attribute_Rule_before[j].IsSupperset().Attribute_Rule_after[j] == True)
11)               If (FMR_Rule_before[j] >FMR_Rule _after[j])
                        Read_rules[j].Remove() , j = j + 1
12)               Else If (FMR_Rule_before[j] == FMR_Rule _after[j])
13)                     If (CAA_Rule_before[j] >= CAA_Rule _after[j])
                              Read_rules[j].Remove() , j = j + 1
14)                     ElseNewR = Write.Rules[j] ,j = j + 1
15)               Else If (FMR_Rule_before[j] <FMR_Rule _after[j])
                        NewR = Write.Rules[j] ,j = j + 1
16)         Else If (Attribute_Rule_before[j].IsSupperset().Attribute_Rule_after[j] == False)
                NewR = Write.Rules[j] ,j = j + 1
17) END While Loop
18) Return NewR
19) END
```

รูปที่ 3.14รหัสเทียมของอัลกอริทึมการจัดการข้อมูลด้วยการขจัดกฎความสัมพันธ์ที่ซ้ำซ้อน

จากรูปที่ 3.14สามารถอธิบายขั้นตอนการทำงานของอัลกอริทึมได้ดังนี้

อัลกอริทึม RemoveRedundantRules

อินพุต ตัวแปร Rแทน Rules ของกฎความสัมพันธ์ที่ผ่านการรวมกฎความสัมพันธ์

เอ้าพุตท์ ตัวแปร NewRulesแทน NewRซึ่งเป็นกฎความสัมพันธ์ใหม่

ขั้นที่ 1 อ่านข้อมูลกฎความสัมพันธ์นำมาเก็บไว้ที่ตัวแปร Read_rules

ขั้นที่ 2กำหนดค่าให้ตัวแปร jมีค่าเป็น 1

ขั้นที่ 3-16 การทำงานใน While Loop

- วนการทำงาน ตามจำนวนของกฎความสัมพันธ์ที่ผ่านการจัดอันดับกฎความสัมพันธ์

ดึงค่า Attribute ทั้งหมดที่อยู่ในกฎความสัมพันธ์ลำดับที่ j

เก็บไว้ที่ Attribute_rule_beforeลำดับที่ j

ดึงค่า FMR ที่อยู่ในกฎความสัมพันธ์ลำดับที่ j

เก็บไว้ที่ FMR_rule_beforeลำดับที่ j

ดึงค่า CAA ที่อยู่ในกฎความสัมพันธ์ลำดับที่ j

เก็บไว้ที่ CAA_rule_beforeลำดับที่ j

เพิ่มค่าj ขึ้น 1 ค่า

ดึงค่า Attribute ทั้งหมดที่อยู่ในกฎความสัมพันธ์ลำดับที่ j

เก็บไว้ที่ Attribute_rule_afterลำดับที่ j

ดึงค่า FMR ที่อยู่ในกฎความสัมพันธ์ลำดับที่ j

เก็บไว้ที่ FMR_rule_afterลำดับที่ j

ดึงค่า CAA ที่อยู่ในกฎความสัมพันธ์ลำดับที่ j

เก็บไว้ที่ CAA_rule_afterลำดับที่ j

-ถ้า Attribute_rule_beforeลำดับที่ j เป็น Supersetของ Attribute_rule_afterลำดับที่ j จริง

ถ้า FMR_rule_beforeลำดับที่ j มีค่ามากกว่า FMR_rule_beforeลำดับที่ j

<u>จริง</u> ลบกฎความสัมพันธ์ลำดับที่ j และเพิ่มค่าj อีก 1 ค่า

ถ้า FMR_rule_beforeลำดับที่ j มีค่าเท่ากับ FMR_rule_beforeลำดับที่ j

และ ถ้า CAA_rule_beforeลำดับที่ j มีค่ามากกว่าหรือเท่ากับ

อัลกอริทึม RemoveRedundantRules (ต่อ)

CAA_rule_beforeลำดับที่ jจริง

ลบกฎความสัมพันธ์ลำดับที่ j และเพิ่มค่า j อีก 1 ค่า

ถ้า FMR_rule_beforeลำดับที่ j มีค่าน้อยกว่า จริง

FMR_rule_beforeลำดับที่ j

เขียนกฎความสัมพันธ์ลำดับที่ j ลงที่ NewR

และเพิ่มค่า j อีก 1 ค่า

-ถ้า Attribute_rule_beforeลำดับที่ j เป็น Supperซองของ

Attribute_rule_afterลำดับที่ j ไม่จริง

เขียนกฎความสัมพันธ์ลำดับที่ j ลงที่ NewR

และเพิ่มค่า j อีก 1 ค่า

จบการทำการ While Loop

ขั้นที่ 18คืนค่า NewR

จบการทำงาน

3.3 เครื่องมือที่ใช้ในการวิจัย

เครื่องมือที่ใช้ในงานวิจัยนี้ ประกอบด้วยฮาร์ดแวร์และซอฟต์แวร์ ดังนี้

1) เครื่องคอมพิวเตอร์ โดยมีรายละเอียดดังนี้
   - หน่วยประมวลผลกลาง : Intel® Core i7-4500 CPU @ 1.80 GHz
   - หน่วยความจำสำรอง : 1 TB
   - หน่วยความจำหลัก : 8 GB
   - อุปกรณ์เสริมอื่นๆ เช่น เมาส์ แป้นพิมพ์ เป็นต้น
2) ซอฟต์แวร์ฏิบัติการและโปรแกรมประยุกต์สำหรับการทำงานเบื้องต้นประกอบไปด้วย
   - ระบบปฏิบัติการ : Windows 8Enterprise 64-bit Operating System

บทที่ 4
การทดสอบและอภิปรายผล

การทดสอบประสิทธิภาพของระบบนั้นจะทดสอบประสิทธิภาพโดยใช้เกณฑ์วัดค่าความถูกต้องที่นำเสนอขึ้นมาใหม่นั้นคือ CAAซึ่งทำการทดสอบกับข้อมูลทางการแพทย์ 2 ข้อมูล ได้แก่ ข้อมูลโรคหอบหืด และข้อมูลโรคหัวใจ การทดสอบประสิทธิภาพนั้นจะทดสอบในแง่มุมของการทดสอบประสิทธิภาพตามวิธีการที่งานวิจัยนี้ได้เสนอโดยใช้ข้อมูลโรคหอบหืดในการเปรียบเทียบกับงานวิจัยอื่นจะใช้ข้อมูลโรคหัวใจ

4.1 ข้อมูลที่ใช้ในการทดสอบ

สำหรับการค้นหากฎความสัมพันธ์จะใช้ข้อมูลที่ได้จากการเก็บข้อมูลจริง ซึ่งเป็นข้อมูลโรคหอบหืดที่มีระดับความรุนแรง 3 ระดับ โดยมีข้อมูลทั้งหมด 698 แถว ประกอบไปด้วยคอลัมน์ 12 คอลัมน์ ในการเปรียบเทียบประสิทธิภาพกับงานวิจัยอื่นจะใช้ข้อมูลผู้ป่วยที่เป็นโรคหัวใจจากแหล่งข้อมูลมาตรฐานซึ่งมีการแบ่งเป็น 2 ประเภท ได้แก่ ผู้ป่วยที่เป็นโรคหัวใจ และผู้ป่วยที่ไม่เป็นโรคหัวใจ โดยมีข้อมูลทั้งหมด 270แถว ประกอบไปด้วยคอลัมน์ 14 คอลัมน์ สามารถแสดงตัวอย่างของรายละเอียดข้อมูลทั้ง สอง ที่ใช้ในการวิจัย ดังตารางที่ 4.1 และ 4.2 ตามลำดับความหมายของแต่ละคอลลัมน์อธิบายไว้ในตารางที่ 4.3 และ 4.4

ตารางที่ 4.1 ตัวอย่างเบื้องต้นของข้อมูลโรคหอบหืด

Age	Sex	Education	Marital_status	Religion	Smoking	Exercise	Weight	Height	Waist	Bodyfat	PA_level
54	Female	Bachelor_degree	Married	Buddhist	Never_smoker	Exercise	62.4	163	75.5	33	Moderate
48	Female	Bachelor_degree	Divorced		Never_smoker	Exercise	61.1	156	76.5	36.1	High
52	Female	Master_degree	Married	Buddhist	Previous_smoker	No_exercise	47.5	153	65.5	27.6	Low
51	Female	Master_degree	Married	Buddhist	Never_smoker	Exercise	57.1	160	83	31.6	Moderate
55	Female	Master_degree	Single	Buddhist	Never_smoker	No_exercise	57.7	150	87	38.2	Low
57	Female	Bachelor_degree	Married	Buddhist	Never_smoker	No_exercise	53.4	158	76	32.2	Low
37	Female	Bachelor_degree	Married	Buddhist	Never_smoker	No_exercise	44.4	152	63.5	24.5	Low
54	Female		Married	Buddhist	Never_smoker	No_exercise	59.1	150	79	36.3	Low
38	Male	Lessthan_bachelor_degree	Married	Buddhist	Never_smoker	Exercise	85.2	179	91.5	24.7	High
54	Male	Lessthan_bachelor_degree	Widowed	Buddhist	Currents_smoker	No_exercise	57	156	84	24.5	Low
35	Male	Lessthan_bachelor_degree	Single	Buddhist	Currents_smoker	Exercise	59.6	169	67.8	14.1	Moderate
35	Female	Master_degree	Single	Buddhist	Never_smoker	Exercise	50.3	165	67.5	25	Moderate
58	Female	Bachelor_degree	Single	Buddhist	Never_smoker	Exercise	45.2	158	65.5	26.6	High
42	Female	Bachelor_degree	Divorced	Buddhist	Never_smoker	Exercise	52.8	159	72	30.9	Moderate
39	Male	Master_degree	Married	Buddhist	Never_smoker	Exercise	69.6	169	84	23	Moderate

ตารางที่ 4.2 ตัวอย่างเบื้องต้นของข้อมูลโรคหัวใจ

Age	Sex	Cpt	Rbp	Chol	Fbs	Restecg	MaxHr	Exang	Oldpeak	Slope	Ca	Thal	Class
70	Female	Asympt	130	322	F	Hyp	109	No	2.4	Flat	3	Normal	Presence
67	Male	Notang	115	564	F	Hyp	160	No	1.6	Flat	0	Reversible_defect	Absence
57	Female	Abnang	124	261	F	Norm	141	No	0.3	Up	0	Reversible_defect	Presence
64	Female	Asympt	128	263	F	Norm	105	Yes	0.2	Flat	1	Reversible_defect	Absence
74	Male	Abnang	120	269	F	Hyp	121	Yes	0.2	Up	1	Normal	Absence
65	Female	Asympt	120	177	F	Norm	140	No	0.4	Up	0	Reversible_defect	Absence
56	Female	Notang	130	256	T	Hyp	142	Yes	0.6	Flat	1	Fixed_defect	Presence
59	Female	Asympt	110	239	F	Hyp	142	Yes	1.2	Flat	1	Reversible_defect	Presence
60	Female	Asympt	140	293	F	Hyp	170	No	1.2	Flat	2	Reversible_defect	Presence
63	Male	Asympt	150	407	F	Hyp	154	No	4	Flat	3	Reversible_defect	Presence
59	Female	Asympt	135	234	F	Norm	161	No	0.5	Flat	0	Reversible_defect	Absence
53	Female	Asympt	142	226	F	Hyp	111	Yes	0	Up	0	Reversible_defect	Absence
44	Female	Notang	140	235	F	Hyp	180	No	0	Up	0	Normal	Absence
61	Female	Angina	134	234	F	Norm	145	No	2.6	Flat	2	Normal	Presence
57	Male	Asympt	128	303	F	Hyp	159	No	0	Up	1	Normal	Absence

ตารางที่ 4.3 คุณลักษณะของข้อมูลโรคหอบหืด

คุณลักษณะ	**อธิบายคุณลักษณะ**	**ค่าที่เป็นไปได้**
Age	อายุของผู้ป่วย	35-64
Sex	เพศของผู้ป่วย	เพศชาย เพศหญิง
Education	ระดับการศึกษาของผู้ป่วย	ต่ำกว่าปริญญาตรี ปริญญาตรี ปริญญาโท ปริญญาเอก
Marital_status	สถานภาพสมรสของผู้ป่วย	โสด แต่งงาน แยกกัน หย่าร้าง ม่าย
Religion	ศาสนาของผู้ป่วย	พุทธ คริสต์ อิสลาม
Smoking	ประวัติการสูบบุหรี่ของผู้ป่วย	ไม่เคยสูบบุหรี่ สูบบุหรี่ก่อนหน้า ปัจจุบันสูบบุหรี่
Exercise	ประวัติการออกกำลังกายของผู้ป่วย	ไม่ออกกำลังกาย ออกกำลังกาย
Weight	น้ำหนักของผู้ป่วย	37.2-113.3
Height	ความสูงของผู้ป่วย	141-192
Waist	รอบเอวของผู้ป่วย	57-119
Bodyfat	ดัชนีมวลกายของผู้ป่วย	11.7-47.6
PA_level	ระดับความรุนแรงโรคหอบหืดของผู้ป่วย	ต่ำ ปานกลาง สูง

ตารางที่ 4.4 คุณลักษณะของข้อมูลโรคหัวใจ

คุณลักษณะ	**อธิบายคุณลักษณะ**	**ค่าที่เป็นไปได้**
Age	อายุของผู้ป่วย	29-77
Sex	เพศของผู้ป่วย	เพศชาย เพศหญิง
Cpt = Chest pain type	รูปแบบของการปวดหน้าอกในผู้ป่วย	หลอดเลือดหัวใจตีบ หลอดเลือดหัวใจตีบผิดปกติ ไม่เป็นหลอดเลือดหัวใจตีบ ไม่มีอาการ
Rbp = Resting blood pressure	ช่วงการคลายตัวของหัวใจ	94-200
Chol = Serum cholesterol in mg/dl	ปริมาณคอเลสเตอรอลในเลือดของผู้ป่วย	126-564
Fbs = fasting blood sugar >120 mg/dl	ปริมาณน้ำตาลในเลือดของผู้ป่วย	จริง เท็จ
Restecg = resting electrocardiographic results	ผลลัพธ์จากการประมวลผลการเต้นของหัวใจด้วยคลื่นไฟฟ้า	ปกติ ผิดปกติ โค้งแบบผิดปกติ
MaxHr = maximum heart rate achieved	อัตราการเต้นของหัวใจสูงสุดที่วัดได้ของผู้ป่วย	71-202
Exang = exercise induced angina	การออกกำลังกายที่เหนี่ยวนำให้เกิดโรคหลอดเลือดหัวใจตีบของผู้ป่วย	ใช่ ไม่ใช่
Oldpeak = ST depression induced by exercise relative to rest	ระดับภาวะซึมเศร้าจากการออกกำลังกายเทียบกับสภาพปกติขณะพัก	0-6.2

ตารางที่ 4.4 คุณลักษณะของข้อมูลโรคหัวใจ (ต่อ)

คุณลักษณะ	**อธิบายคุณลักษณะ**	**ค่าที่เป็นไปได้**
Slope = the slope characteristics of the peak exercise ST segment	ระดับความลาดชันสูงสุดของส่วนการออกกำลังกาย	ลาดชันสูง แบนราบ ลาดชันต่ำ
Ca = number of fluoroscopy colored major vessels	ระดับความเข้มสีในเลือดของผู้ป่วย	0-3
Thal = The heart rate status	สถานะอัตราการเต้นของหัวใจในผู้ป่วย	ปกติ บกพร่องคงที่ บกพร่องย้อนกลับได้
Class = absence or presence of heart disease	ผลสรุปของผู้ป่วยที่แยกระหว่าง ผู้ป่วยที่เป็นโรคหัวใจ หรือผู้ป่วยที่เป็นโรคหัวใจ	เป็นโรคหัวใจ ไม่เป็นโรคหัวใจ

4.2 การทดสอบประสิทธิภาพกับข้อมูลโรคหอบหืดตามวิธีที่งานวิจัยนี้ได้เสนอ

ตามที่อธิบายไว้ในบทที่ 3 ผู้วิจัยได้ออกแบบขั้นตอนวิธีไว้ทั้งสิ้น 5 ขั้นตอนวิธีในการทดสอบประสิทธิภาพนี้ผู้วิจัยจะทำการทดสอบประสิทธิภาพทีละขั้นตอนตามขั้นตอนวิธีทั้ง 5 ตามลำดับ

4.2.1 การทดสอบตามขั้นตอนวิธี

ข้อมูลโรคหอบหืดประกอบด้วยข้อมูล 698 แถว ซึ่งมี 12 คอลัมน์ 3 คลาส สามารถนำมาทดสอบประสิทธิภาพตามวิธีที่งานวิจัยได้เสนอ

ขั้นตอนวิธีที่ 1 ลบข้อมูลที่ไม่สามารถบ่งชี้ได้ชัดเจน

เมื่อทำการหากฎความสัมพันธ์โดยข้อมูลที่พิจารณานั้นประกอบด้วยข้อมูลที่ไม่สามารถบ่งชี้ได้ชันเจน เมื่อนำมาทำการหากฎความสัมพันธ์จะได้กฎความสัมพันธ์ที่ผิดเพี้ยนไป ดังนั้นผู้วัจัยจึงเสนอวิธีการลบข้อมูลที่ไม่สามารถบ่งชี้ได้ชัดเจน สามารถตรวจพบข้อมูลที่ไม่สามารถบ่งชี้ได้ชัดเจนทั้งสิ้น 19 ทรานแซคชัน ดังตารางที่ 4.5เมื่อทำการลบข้อมูลแล้ว ส่งผลให้เหลือจำนวนข้อมูลทั้งสิ้น 678 ทรานแซคชัน จะนำข้อมูลที่เหลือไปใช้ในขั้นตอนวิธีที่2

ตารางที่ 4.5 ข้อมูลโรคหอบหืดที่ไม่สามารถระบุได้ชัดเจน

Age	Sex	Education	Marital_status	Religion	Smoking	Exercise	Weight	Height	Waist	Bodyfat	PA_level
62		Master_degree	Married	Buddhist	Never_smoker	Exercise	51.8	149	71	33.7	High
52	Female		Married	Buddhist	Never_smoker	Exercise	55.8	155	78	30.6	Moderate
54	Female		Married	Buddhist	Never_smoker	No_exercise	59.1	150	79	36.3	Low
54	Female				Currents_smoker	Exercise	65		86	37.7	Moderate
51	Male		Married	Buddhist	Never_smoker	No_exercise	80.8	170	94	27.1	Low
36	Female	Bachelor_degree		Buddhist	Previous_smoker	No_exercise	59.8	156	77	34	Low
39	Male	Master_degree		Buddhist	Previous_smoker	Exercise	69.6	164	85	21.4	Moderate
48	Female	Bachelor_degree	Divorced		Never_smoker	Exercise	61.1	156	76	36.1	High
54	Male	Bachelor_degree	Divorced		Currents_smoker	Exercise	63.9		79	20.7	Moderate
49	Female	Bachelor_degree	Married		Never_smoker	No_exercise	57.2	151	82	36.6	Low
51	Female	Bachelor_degree	Married		Never_smoker	Exercise	42.6	146	70	26.8	Moderate
54	Female	Bachelor_degree	Married		Never_smoker	No_exercise	51.7	150	78	33.5	Low
59	Female	Bachelor_degree	Married		Never_smoker	Exercise	55.9	152	75	33.9	Moderate
55	Male	Bachelor_degree	Married		Never_smoker	No_exercise	68.2	167	87	20.9	Low
56	Male	Bachelor_degree	Married		Previous_smoker	Exercise	74.9	167	91	25.8	Moderate
55	Female	Bachelor_degree	Single		Never_smoker	Exercise	57.6	158	76	33.1	Moderate
38	Female	Bachelor_degree	Single		Never_smoker	Exercise		160	62	26.4	Moderate
38	Male	Bachelor_degree	Single	Buddhist	Never_smoker	Exercise	85.7		88	24.9	Moderate
52	Female	Bachelor_degree	Single	Buddhist	Never_smoker	Exercise	64.2	158		37.2	High

ขั้นตอนวิธีที่ 2 การจัดช่วงข้อมูลที่มีลักษณะเป็นตัวเลข

เมื่อพิจารณาข้อมูลโรคหอบหืดนั้นพบว่า มีข้อมูลผสมกันระหว่าง ข้อมูลที่เป็นตัวเลขและข้อมูลที่เป็นตัวอักษร ในขั้นตอนนี้จะทำการแบ่งช่วงของข้อมูลที่มีลักษณะเป็นตัวเลขซึ่งคอลัมน์ที่มีลักษณะเป็นตัวเลขมีทั้งสิ้น 5 คอลัมน์ ฉะนั้นขั้นตอนนี้จะทำการแบ่งช่วงของข้อมูลเฉพาะ 5 คอลัมน์ที่มีลักษณะเป็นตัวเลข ได้แก่ คอลัมน์ Age, Weight, Height, Waist และ Bodyfat ดังตารางที่ 4.6

ตารางที่ 4.6 ตัวอย่างข้อมูลโรคหอบหืดที่แบ่งช่วงของข้อมูลที่มีลักษณะเป็นตัวเลข

No.	Age	Weight	Height	Waist	Bodyfat
1	61.1-64	44.81-52.42	146.1-151.2	69.4-75.6	33.24-36.83
2	61.1-64	44.81-52.42	151.2-156.3	69.4-75.6	29.65-33.24
3	61.1-64	67.64-75.25	176.7-181.8	75.6-81.8	18.88-22.47
4	58.2-61.1	75.25-82.86	156.3-161.4	94.2-100.4	44.01-47.6
5	58.2-61.1	60.03-67.64	0-146.1	88-94.2	40.42-44.01
6	58.2-61.1	75.25-82.86	161.4-166.5	88-94.2	40.42-44.01
7	55.3-58.2	75.25-82.86	156.3-161.4	81.8-88	44.01-47.6
8	55.3-58.2	67.64-75.25	151.2-156.3	88-94.2	44.01-47.6
9	55.3-58.2	67.64-75.25	156.3-161.4	81.8-88	44.01-47.6
10	52.4-55.3	67.64-75.25	0-146.1	81.8-88	44.01-47.6
11	52.4-55.3	67.64-75.25	146.1-151.2	88-94.2	44.01-47.6
12	52.4-55.3	67.64-75.25	146.1-151.2	94.2-100.4	44.01-47.6
13	49.5-52.4	82.86-90.47	156.3-161.4	100.4-106.6	44.01-47.6
14	49.5-52.4	67.64-75.25	146.1-151.2	88-94.2	44.01-47.6
15	49.5-52.4	67.64-75.25	151.2-156.3	75.6-81.8	40.42-44.01
16	46.6-49.5	67.64-75.25	146.1-151.2	94.2-100.4	44.01-47.6
17	46.6-49.5	60.03-67.64	146.1-151.2	81.8-88	40.42-44.01
18	46.6-49.5	75.25-82.86	156.3-161.4	88-94.2	40.42-44.01
19	43.7-46.6	75.25-82.86	151.2-156.3	100.4-106.6	44.01-47.6
20	43.7-46.6	75.25-82.86	156.3-161.4	81.8-88	44.01-47.6

ตารางที่ 4.6 ตัวอย่างข้อมูลโรคหอบหืดที่แบ่งช่วงของข้อมูลที่มีลักษณะเป็นตัวเลข(ต่อ)

No.	Age	Weight	Height	Waist	Bodyfat
21	43.7-46.6	75.25-82.86	156.3-161.4	88-94.2	44.01-47.6
22	40.8-43.7	67.64-75.25	146.1-151.2	88-94.2	44.01-47.6
23	40.8-43.7	75.25-82.86	151.2-156.3	88-94.2	44.01-47.6
24	40.8-43.7	67.64-75.25	151.2-156.3	75.6-81.8	40.42-44.01
25	37.9-40.8	82.86-90.47	161.4-166.5	81.8-88	44.01-47.6
26	37.9-40.8	67.64-75.25	151.2-156.3	81.8-88	40.42-44.01
27	37.9-40.8	67.64-75.25	151.2-156.3	81.8-88	40.42-44.01
28	0-37.9	82.86-90.47	156.3-161.4	106.6-112.8	44.01-47.6
29	0-37.9	75.25-82.86	151.2-156.3	88-94.2	44.01-47.6
30	0-37.9	67.64-75.25	146.1-151.2	75.6-81.8	40.42-44.01

จากขั้นตอนที่ 2 ได้ทำการแบ่งช่วงข้อมูลที่มีลักษณะเป็นตัวเลข ซึ่งกระทำทั้งสิ้น 5 คอลัมน์ โดยทำการแบ่งช่วงของข้อมูล 10 ช่วงข้อมูล ทุก ๆ คอลัมน์ ดังตารางที่ 4.7

ตารางที่ 4.7 แปลงช่วงข้อมูลให้อยู่ในลักษณะตัวเลขจำเพาะ

No.	Age	Weight	Height	Waist	Bodyfat
1	0-37.9 = 1	0-44.81 = 1	0-146.1= 1	0-63.2 = 1	0-15.29 = 1
2	38-40.8 = 2	44.82-52.42 = 2	146.2-151.2= 2	63.3-69.4 = 2	15.3-18.88 = 2
3	40.9-43.6 = 3	52.43-60.03 = 3	151.3-156.3 = 3	69.5-75.6 = 3	18.89-22.47 = 3
4	43.8-46.6= 4	60.04-67.64 = 4	156.4-161.4= 4	75.7-81.8 = 4	22.48-26.06 = 4
5	46.7-49.5 = 5	67.65-75.25 = 5	161.5-166.5= 5	81.9-88= 5	26.07-29.65 = 5
6	49.6-52.4 = 6	75.26-82.86= 6	166.6-171.6 = 6	88.1-94.2= 6	29.66-33.24 = 6
7	52.5-55.3= 7	82.87-90.47= 7	171.7-176.7 = 7	94.3-100.4 = 7	33.25-36.83 = 7
8	55.4-58.2= 8	90.48-98.08 = 8	176.8-181.8 = 8	100.5-106.6 = 8	36.84-40.42 = 8
9	58.3-61.1= 9	98.09-105.69 = 9	181.9-186.9 = 9	106.7-112.8 = 9	40.43-44.01= 9
10	61.2-64 = 10	105.7-113.3 = 10	187-192 = 10	112.9-119 = 10	47.2-51.19 = 10

ขั้นตอนวิธีที่ 3 รวมกฎความสัมพันธ์

จากกฎความสัมพันธ์ที่ได้จากการค้นหากฎความสัมพันธ์ทั้ง 3 แบบ จะถูกนำมากระทำในข้อตอนที่ 3 การรวมกฎความสัมพันธ์ กฎความสัมพันธ์ที่ได้จากทั้ง 3 แบบ เป็นไปตามตารางที่ 4.8 4.9 และ 4.10ถัดมาจะเป็นในส่วนของกฎความสัมพันธ์ที่ผ่านการรวมกฎความสัมพันธ์ซึ่งแสดงค่า CAA และ ค่า FMR ดังตารางที่ 4.11

ตารางที่ 4.8กฎความสัมพันธ์จากการหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรี

No.	Rules:	Confidence
1.	IF { exercise=no_exercise && weight=3 } THEN { PA_level=low }	1
2.	IF { religion=buddhist && exercise=no_exercise && weight=3 } THEN { PA_level=low }	1
3.	IF { exercise=no_exercise && height=4 } THEN { PA_level=low }	1
4.	IF { smoking=never_smoker && exercise=no_exercise && weight=3 } THEN { PA_level=low }	1
5.	IF { sex=female && exercise=no_exercise && weight=3 } THEN { PA_level=low }	1
6.	IF { religion=buddhist && exercise=no_exercise && height=4 } THEN { PA_level=low }	1
7.	IF { religion=buddhist && smoking=never_smoker && exercise=no_exercise && weight=3 } THEN PA_level=low	1
8.	IF { sex=female && religion=buddhist && exercise=no_exercise&& weight=3 } THEN { PA_level=low }	1
9.	IF { smoking=never_smoker && exercise=no_exercise && height=4 } THEN { PA_level=low }	1

ตารางที่ 4.8 กฎความสัมพันธ์จากการหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรี (ต่อ)

No.	Rules:	Confidence
10.	IF { sex=female && smoking=never_smoker && exercise=no_exercise && weight=3 } THEN { PA_level=low }	1
11.	IF { religion=buddhist && smoking=never_smoker && exercise=no_exercise && height=4 } THEN { PA_level=low }	1
12.	IF { sex=female && religion=buddhist && smoking=never_smoker && exercise=no_exercise &&weight=3 } THEN { PA_level=low }	1
13.	IF { sex=female && exercise=no_exercise && height=4 } THEN { PA_level=low }	1
14.	IF { exercise=no_exercise && height=3 } THEN { PA_level=low }	1
15.	IF { sex=female && religion=buddhist && exercise=no_exercise && height=4} THEN { PA_level=low }	1
16.	IF { sex=female && smoking=never_smoker && exercise=no_exercise && height=4 } THEN { PA_level=low }	1
17.	IF { religion=buddhist && exercise=no_exercise && height=3 } THEN { PA_level=low }	1
18.	IF { sex=female && exercise=no_exercise && height=3 } THEN { PA_level=low }	1
19.	IF { sex=female && religion=buddhist && smoking=never_smoker && exercise=no_exercise && height=4 } THEN { PA_level=low }	1

ตารางที่ 4.8 กฎความสัมพันธ์จากการหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรี (ต่อ)

No.	Rules:	Confidence
20.	IF { exercise=no_exercise && waist=3 } THEN { PA_level=low }	1
21.	IF { smoking=never_smoker && exercise=no_exercise && height=3 THEN { PA_level=low }	1
22.	IF { sex=female && religion=buddhist && exercise=no_exercise && height=3} THEN {PA_level=low }	1
23.	IF { sex=female && exercise=no_exercise && waist=3 } THEN { PA_level=low }	1
24.	IF { smoking=never_smoker&& exercise=no_exercise&& waist=3 } THEN PA_level=low }	1
25.	IF{ religion=buddhist && exercise=no_exercise&& waist=3 } THEN { PA_level=low }	1
26.	IF { sex=female&& smoking=never_smoker&& exercise=no_exercise&& height=3 } THEN { PA_level=low }	1
27.	IF{ sex=female&& smoking=never_smoker&& exercise=no_exercise&& waist=3 } THEN { PA_level=low }	1
28.	IF { religion=buddhist && smoking=never_smoker&&exercise=no_exercise&& height=3 } THEN { PA_level=low }	1
29.	IF { sex=female&& religion=buddhist && exercise=no_exercise&& waist=3} THEN { PA_level=low }	1
30.	IF { religion=buddhist && smoking=never_smoker&& exercise=no_exercise&& waist=3 } THEN { PA_level=low }	1

ตารางที่ 4.9กฎความสัมพันธ์จากการหากฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทิพเอไพรออรี

No.	Rules:	Accuracy
1.	IF { exercise=no_exercise&& weight=3 } THEN { PA_level=low }	0.99494
2.	IF { exercise=no_exercise&& height=4 } THEN { PA_level=low }	0.99493
3.	IF { exercise=no_exercise&& height=3 } THEN { PA_level=low }	0.99491
4.	IF { exercise=no_exercise&& waist=3 } THEN { PA_level=low }	0.9949
5.	IF { exercise=no_exercise&& bodyfat=6 } THEN { PA_level=low }	0.99487
6.	IF { exercise=no_exercise&& bodyfat=5 } THEN { PA_level=low }	0.99486
7.	IF { exercise=no_exercise&& height=5 } THEN { PA_level=low }	0.99486
8.	IF { exercise=no_exercise&& waist=5 } THEN { PA_level=low }	0.99486
9.	IF { exercise=no_exercise&& weight=2 } THEN { PA_level=low 62 }	0.99485
10.	IF { exercise=no_exercise&& waist=4 } THEN { PA_level=low 62 }	0.99485
11.	IF { exercise=no_exercise&& bodyfat=7 } THEN { PA_level=low }	0.99483
12.	IF { exercise=no_exercise&& weight=4 } THEN { PA_level=low }	0.99482
13.	IF { exercise=no_exercise&& weight=5 } THEN { PA_level=low }	0.99481

ตารางที่ 4.9กฎความสัมพันธ์จากการหากฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทิพเอไพรออรี(ต่อ)

No.	Rules:	Accuracy
14.	IF { age=4&& exercise=no_exercise } THEN { PA_level=low }	0.99476
15.	IF { exercise=no_exercise&& bodyfat=4 } THEN { PA_level=low }	0.99473
16.	IF { age=3&& exercise=no_exercise } THEN { PA_level=low }	0.9947
17.	IF { age=5&& exercise=no_exercise } THEN { PA_level=low }	0.9947
18.	IF { exercise=no_exercise&& waist=2 } THEN { PA_level=low }	0.9947
19.	IF { age=6&& exercise=no_exercise } THEN { PA_level=low }	0.99466
20.	IF { exercise=no_exercise&& height=2 } THEN { PA_level=low }	0.99462
21.	IF { exercise=no_exercise&& waist=6 } THEN { PA_level=low }	0.99462
22.	IF { exercise=no_exercise&& bodyfat=8 } THEN { PA_level=low }	0.99462
23.	IF { age=7&& exercise=no_exercise } THEN { PA_level=low }	0.99458
24.	IF { age=8 &&exercise=no_exercise } THEN { PA_level=low }	0.99454
25.	IF{ age=1&& exercise=no_exercise } THEN { PA_level=low }	0.99451
26.	IF{ exercise=no_exercise&& weight=6 } THEN { PA_level=low }	0.99441

ตารางที่ 4.9กฎความสัมพันธ์จากการหากฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทิพเอไพรออรี(ต่อ)

No.	Rules:	Accuracy
27.	IF{ age=2&& exercise=no_exercise } THEN { PA_level=low }	0.99433
28.	IF{ exercise=no_exercise&& height=6 } THEN { PA_level=low }	0.99429
29.	IF{ exercise=no_exercise && bodyfat=3 } THEN { PA_level=low }	0.99411
30.	IF{ exercise=no_exercise&& waist=7 } THEN { PA_level=low }	0.99364

ตารางที่ 4.10 กฎความสัมพันธ์จากการหากฎความสัมพันธ์ด้วยอัลกอริทึมเทอเชียส

No.	Rules:	Confidence
1.	IF { exercise=no_exercise && weight=3 } THEN { PA_level=low }	0.9725
2.	IF { exercise=no_exercise&& height=4} THEN { PA_level=low }	0.972
3.	IF { exercise=no_exercise&& waist=3 } THEN { PA_level=low }	0.971
4.	IF { religion = buddhist && exercise = exercise && height = 3} THEN { PA_level = moderate }	0.235903
5.	IF { sex = female && exercise = exercise && height = 3 } THEN { PA_level = moderate }	0.235903
6.	IF { sex = female && religion = buddhist && exercise = exercise && height = 3 } THEN { PA_level = moderate }	0.235012
7.	IF { exercise = exercise && bodyfat = 6 } THEN { PA_level = moderate }	0.229250

ตารางที่ 4.10 กฎความสัมพันธ์จากการหากฎความสัมพันธ์ด้วยอัลกอริทึมเทอเชียส (ต่อ)

No.	Rules:	Confidence
8.	IF { religion = buddhist && exercise = exercise && bodyfat = 6 } THEN { PA_level = moderate }	0.228738
9.	IF { sex = female && exercise = exercise && bodyfat = 6 } THEN { PA_level = moderate }	0.227570
10.	IF { sex = female && religion = buddhist && exercise = exercise && bodyfat = 6 } THEN { PA_level = moderate }	0.227318
11.	IF { sex = female && smoking = never_smoker && exercise = exercise && bodyfat = 6 } THEN { PA_level = moderate }	0.223198
12.	IF { exercise = exercise && weight = 3 } THEN { PA_level = moderate }	0.221723
13.	IF { religion = buddhist && exercise = exercise && weight = 3 } THEN { PA_level = moderate }	0.219779
14.	IF { smoking = never_smoker && exercise = exercise && bodyfat = 6 } THEN { PA_level = moderate }	0.219317
15.	IF { religion = buddhist && smoking = never_smoker && exercise = exercise && bodyfat = 6 } THEN { PA_level = moderate }	0.219052
16.	IF { exercise = exercise && height = 4 } THEN { PA_level = moderate }	0.215500
17.	IF { religion = buddhist && exercise = exercise && height = 4} THEN { PA_level = moderate }	0.215152
18.	IF { sex = female && exercise = exercise && height = 4 } THEN { PA_level = moderate }	0.214898
19.	IF { smoking = never_smoker && exercise = exercise && weight = 3 } THEN { PA_level = moderate }	0.213863

ตารางที่ 4.10 กฎความสัมพันธ์จากการหากฎความสัมพันธ์ด้วยอัลกอริทึมเทอเชียส (ต่อ)

No.	Rules:	Confidence
20.	IF { religion = buddhist && smoking = never_smoker && exercise = exercise && weight = 3 } THEN { PA_level = moderate }	0.212156
21.	IF { religion = buddhist && exercise = exercise && weight = 3 && bodyfat = 6 } THEN { PA_level = moderate }	0.199310
22.	IF { exercise = exercise && weight = 3 && bodyfat = 6 } THEN { PA_level = moderate }	0.198611
23.	IF { religion = buddhist && smoking = never_smoker && exercise = exercise && height = 4 } THEN { PA_level = moderate }	0.193302
24.	IF { religion = buddhist && exercise = exercise && weight = 4} THEN { PA_level = moderate }	0.189896
25.	IF { smoking = never_smoker && exercise = exercise && height = 4 } THEN { PA_level = moderate }	0.189183
26.	IF { exercise = exercise && weight = 4 } THEN { PA_level = moderate }	0.188025
27.	IF{ marital_status = married &&religion = buddhist && exercise = exercise&& weight = 4 } THEN { PA_level = moderate }	0.183677
28.	IF { exercise = exercise && waist = 4 } THEN { PA_level = moderate }	0.177064
29.	IF{ religion = buddhist && exercise = exercise && waist = 4 } THEN { PA_level = moderate }	0.176023
30.	IF{ exercise = exercise && waist = 5 } THEN { PA_level = moderate }	0.170662

ตารางที่ 4.11 กฎความสัมพันธ์ที่ผ่านการรวมกฎความสัมพันธ์

No.	Rules:	FMR	CAA
1.	IF { exercise=no_exercise && weight=3 } THEN { PA_level=low }	3	0.98914
2.	IF { religion=buddhist && exercise=no_exercise &&weight=3 } THEN { PA_level=low }	1	1
3.	IF { exercise=no_exercise && height=4 } THEN { PA_level=low }	3	0.98897
4.	IF { smoking=never_smoker && exercise=no_exercise && weight=3 } THEN { PA_level=low }	1	1
5.	IF { sex=female && exercise=no_exercise && weight=3 } THEN { PA_level=low }	1	1
6.	IF { religion=buddhist && exercise=no_exercise && height=4 } THEN { PA_level=low }	1	1
7.	IF { religion=buddhist && smoking=never_smoker && exercise=no_exercise && weight=3 } THEN PA_level=low	1	1
8.	IF { exercise=no_exercise && height=3 } THEN { PA_level=low }	2	0.997455
9.	IF { sex=female && religion=buddhist &&exercise=no_exercise&& weight=3 } THEN { PA_level=low }	1	1
10.	IF { smoking=never_smoker && exercise=no_exercise && height=4 } THEN { PA_level=low }	1	1
11.	IF { sex=female && smoking=never_smoker && exercise=no_exercise && weight=3 } THEN { PA_level=low }	1	1

ตารางที่ 4.11 กฎความสัมพันธ์ที่ผ่านการรวมกฎความสัมพันธ์ (ต่อ)

No.	Rules:	FMR	CAA
12.	IF { religion=buddhist && smoking=never_smoker &&exercise=no_exercise && height=4 } THEN { PA_level=low }	1	1
13.	IF { sex=female && religion=buddhist && smoking=never_smoker && exercise=no_exercise && weight=3} THEN { PA_level=low }	1	1
14.	IF { sex=female && exercise=no_exercise && height=4 } THEN { PA_level=low }	1	1
15.	IF { sex=female && religion=buddhist &&exercise=no_exercise && height=4 } THEN { PA_level=low }	1	1
16.	IF { sex=female && smoking=never_smoker && exercise=no_exercise && height=4 } THEN { PA_level=low }	1	1
17.	IF { religion=buddhist && exercise=no_exercise&& height=3 } THEN { PA_level=low }	1	1
18.	IF { sex=female&&exercise=no_exercise && height=3} THEN { PA_level=low }	1	1
19.	IF { sex=female&& religion=buddhist &&smoking=never_smoker&&exercise=no_exercise&& height=4 } THEN { PA_level=low }	1	1
20.	IF { exercise=no_exercise&& waist=3 } THEN { PA_level=low }	3	0.98683
21.	IF { smoking=never_smoker&&exercise=no_exercise&& height=3} THEN { PA_level=low }	1	1

ตารางที่ 4.11 กฎความสัมพันธ์ที่ผ่านการรวมกฎความสัมพันธ์ (ต่อ)

No.	Rules:	FMR	CAA
22.	IF { sex=female&& religion=buddhist && exercise=no_exercise&& height=3 } THEN {PA_level=low }	1	1
23.	IF { sex=female&& exercise=no_exercise&& waist=3 } THEN { PA_level=low }	1	1
24.	IF { smoking=never_smoker&&exercise=no_exercise&& waist=3 } THEN PA_level=low }	1	1
25.	IF{ religion=buddhist && exercise=no_exercise&& waist=3 } THEN { PA_level=low }	1	1
26.	IF { sex=female&& smoking=never_smoker&& exercise=no_exercise&& height=3 } THEN { PA_level=low }	1	1
27.	IF{ sex=female&& smoking=never_smoker&& exercise=no_exercise&& waist=3 } THEN { PA_level=low }	1	1
28.	IF { religion=buddhist &&smoking=never_smoker&& exercise=no_exercise&& height=3 } THEN { PA_level=low }	1	1
29.	IF { sex=female&& religion=buddhist && exercise=no_exercise&& waist=3 } THEN { PA_level=low }	1	1
30.	IF { religion=buddhist && smoking=never_smoker&& exercise=no_exercise&& waist=3 } THEN { PA_level=low }	1	1
31.	IF { exercise=no_exercise&& bodyfat=6 }	1	0.99487

	THEN { PA_level=low }		

ตารางที่ 4.11 กฎความสัมพันธ์ที่ผ่านการรวมกฎความสัมพันธ์ (ต่อ)

No.	Rules:	FMR	CAA
32.	IF { exercise=no_exercise&& bodyfat=5 } THEN { PA_level=low }	1	0.99486
33.	IF { exercise=no_exercise&& height=5 } THEN { PA_level=low }	1	0.99486
34.	IF { exercise=no_exercise&& waist=5 } THEN { PA_level=low }	1	0.99486
35.	IF { exercise=no_exercise&& weight=2 } THEN { PA_level=low }	1	0.99485
36.	IF { exercise=no_exercise&& waist=4 } THEN { PA_level=low }	1	0.99485
37.	IF { exercise=no_exercise&& bodyfat=7 } THEN { PA_level=low }	1	0.99483
38.	IF { exercise=no_exercise&& weight=4 } THEN { PA_level=low }	1	0.99482
39.	IF { exercise=no_exercise&& weight=5 } THEN { PA_level=low }	1	0.99481
40.	IF { age=4&& exercise=no_exercise } THEN { PA_level=low }	1	0.99476
41.	IF { exercise=no_exercise&& bodyfat=4 } THEN { PA_level=low }	1	0.99473
42.	IF { age=3&& exercise=no_exercise } THEN { PA_level=low }	1	0.9947
43.	IF { age=5&& exercise=no_exercise } THEN { PA_level=low }	1	0.9947
44.	IF { exercise=no_exercise&& waist=2 } THEN { PA_level=low }	1	0.9947

ตารางที่ 4.11 กฎความสัมพันธ์ที่ผ่านการรวมกฎความสัมพันธ์ (ต่อ)

No.	Rules:	FMR	CAA
45.	IF { age=6&& exercise=no_exercise } THEN { PA_level=low }	1	0.99466
46.	IF { exercise=no_exercise&& height=2 } THEN { PA_level=low }	1	0.99462
47.	IF { exercise=no_exercise&& waist=6 } THEN { PA_level=low }	1	0.99462
48.	IF { exercise=no_exercise&& bodyfat=8 } THEN { PA_level=low }	1	0.99462
49.	IF { age=7&& exercise=no_exercise } THEN { PA_level=low }	1	0.99458
50.	IF { age=8 &&exercise=no_exercise } THEN { PA_level=low }	1	0.99454
51.	IF{ age=1&& exercise=no_exercise } THEN { PA_level=low }	1	0.99451
52.	IF{ exercise=no_exercise&& weight=6 } THEN { PA_level=low }	1	0.99441
53.	IF{ age=2&& exercise=no_exercise } THEN { PA_level=low }	1	0.99433
54.	IF{ exercise=no_exercise&& height=6 } THEN { PA_level=low }	1	0.99429
55.	IF{ exercise=no_exercise && bodyfat=3 } THEN { PA_level=low }	1	0.99411
56.	IF{ exercise=no_exercise&& waist=7 } THEN { PA_level=low }	1	0.99364
57.	IF { religion = buddhist && exercise = exercise && height = 3} THEN { PA_level = moderate }	1	0.235903

ตารางที่ 4.11 กฎความสัมพันธ์ที่ผ่านการรวมกฎความสัมพันธ์ (ต่อ)

No.	Rules:	FMR	CAA
58.	IF { sex = female && exercise = exercise && height = 3} THEN { PA_level = moderate }	1	0.235903
59.	IF { sex = female && religion = buddhist && exercise = exercise && height = 3 } THEN { PA_level = moderate }	1	0.235012
60.	IF { exercise = exercise && bodyfat = 6 } THEN { PA_level = moderate }	1	0.229250
61.	IF { religion = buddhist && exercise = exercise && bodyfat = 6} THEN { PA_level = moderate }	1	0.228738
62.	IF { sex = female && exercise = exercise && bodyfat = 6 } THEN { PA_level = moderate }	1	0.227570
63.	IF { sex = female && religion = buddhist && exercise = exercise && bodyfat = 6 } THEN { PA_level = moderate }	1	0.227318
64.	IF { sex = female && smoking = never_smoker && exercise = exercise && bodyfat = 6 } THEN { PA_level = moderate }	1	0.223198
65.	IF { exercise = exercise && weight = 3 } THEN { PA_level = moderate }	1	0.221723
66.	IF { religion = buddhist && exercise = exercise && weight = 3 } THEN { PA_level = moderate }	1	0.219779
67.	IF { smoking = never_smoker && exercise = exercise && bodyfat = 6 } THEN { PA_level = moderate }	1	0.219317
68.	IF { religion = buddhist && smoking = never_smoker && exercise = exercise && bodyfat = 6 }	1	0.219052

	THEN { PA_level = moderate }		

ตารางที่ 4.11 กฎความสัมพันธ์ที่ผ่านการรวมกฎความสัมพันธ์ (ต่อ)

No.	Rules:	FMR	CAA
69.	IF { exercise = exercise && height = 4 } THEN { PA_level = moderate }	1	0.215500
70.	IF { religion = buddhist && exercise = exercise && height = 4} THEN { PA_level = moderate }	1	0.215152
71.	IF { sex = female && exercise = exercise && height = 4} THEN { PA_level = moderate }	1	0.214898
72.	IF { smoking = never_smoker && exercise = exercise&&weight = 3 } THEN { PA_level = moderate }	1	0.213863
73.	IF { religion = buddhist && smoking = never_smoker && exercise = exercise && weight = 3 } THEN { PA_level = moderate }	1	0.212156
74.	IF { religion = buddhist && exercise = exercise && weight = 3 && bodyfat = 6 } THEN { PA_level = moderate }	1	0.199310
75.	IF { exercise = exercise && weight = 3 && bodyfat = 6 } THEN { PA_level = moderate }	1	0.198611
76.	IF { religion = buddhist && smoking = never_smoker && exercise = exercise && height = 4 } THEN { PA_level = moderate }	1	0.193302
77.	IF { religion = buddhist && exercise = exercise && weight = 4} THEN { PA_level = moderate }	1	0.189896
78.	IF { smoking = never_smoker && exercise = exercise && height = 4} THEN { PA_level = moderate }	1	0.189183
79.	IF { exercise = exercise && weight = 4 }	1	0.188025

	THEN { PA_level = moderate }		

ตารางที่ 4.11 กฎความสัมพันธ์ที่ผ่านการรวมกฎความสัมพันธ์ (ต่อ)

No.	Rules:	FMR	CAA
80.	IF{ marital_status = married &&religion = buddhist && exercise = exercise&& weight = 4 } THEN { PA_level = moderate }	1	0.183677
81.	IF { exercise = exercise && waist = 4 } THEN { PA_level = moderate }	1	0.177064
82.	IF{ religion = buddhist && exercise = exercise && waist = 4 } THEN { PA_level = moderate }	1	0.176023
83.	IF{ exercise = exercise && waist = 5 } THEN { PA_level = moderate }	1	0.170662

ขั้นตอนวิธีที่ 4 การจัดอันดับกฎความสัมพันธ์

จากกฎความสัมพันธ์ที่ได้จากรวมกฎความสัมพันธ์จากอัลกอริทึมการค้นหากฎความสัมพันธ์ทั้ง 3 แบบ จะถูกนำมาจัดอันดับกฎความสัมพันธ์แสดงการเรียงลำดับได้ดังตารางที่ 4.12

ตารางที่ 4.12 กฎความสัมพันธ์ที่ผ่านการจัดอันดับกฎความสัมพันธ์

Rank	Rules:	FMR	CAA
1.	IF { exercise=no_exercise && weight=3 } THEN { PA_level=low }	3	0.98914
2.	IF { exercise=no_exercise && height=4 } THEN { PA_level=low }	3	0.98897
3.	IF { exercise=no_exercise && waist=3 } THEN { PA_level=low }	3	0.98683
4.	IF { exercise=no_exercise && height=3 } THEN { PA_level=low }	2	0.997455
5.	IF { religion=buddhist && exercise=no_exercise &&weight=3 } THEN { PA_level=low }	1	1

ตารางที่ 4.12กฎความสัมพันธ์ที่ผ่านการจัดอันดับกฎความสัมพันธ์ (ต่อ)

Rank	Rules:	FMR	CAA
6.	IF { smoking=never_smoker && exercise=no_exercise && weight=3 } THEN { PA_level=low }	1	1
7.	IF { sex=female && exercise=no_exercise && weight=3 } THEN { PA_level=low }	1	1
8.	IF { religion=buddhist && exercise=no_exercise && height=4 } THEN { PA_level=low }	1	1
9.	IF { religion=buddhist && smoking=never_smoker && exercise=no_exercise && weight=3 } THEN PA_level=low	1	1
10.	IF { sex=female && religion=buddhist &&exercise=no_exercise&& weight=3 } THEN { PA_level=low }	1	1
11.	IF { religion=buddhist && smoking=never_smoker && exercise=no_exercise && height=4 } THEN { PA_level=low }	1	1
12.	IF { sex=female && religion=buddhist && smoking=never_smoker && exercise=no_exercise && weight=3} THEN { PA_level=low }	1	1
13.	IF { sex=female && exercise=no_exercise && height=4 } THEN { PA_level=low }	1	1
14.	IF { sex=female && religion=buddhist &&exercise=no_exercise && height=4 } THEN { PA_level=low }	1	1
15.	IF { sex=female && smoking=never_smoker && exercise=no_exercise && height=4 }	1	1

	THEN { PA_level=low }		

ตารางที่ 4.12กฎความสัมพันธ์ที่ผ่านการจัดอันดับกฎความสัมพันธ์ (ต่อ)

Rank	**Rules:**	**FMR**	**CAA**
16.	IF { religion=buddhist && exercise=no_exercise&& height=3 } THEN { PA_level=low }	1	1
17.	IF { sex=female&&exercise=no_exercise && height=3} THEN { PA_level=low }	1	1
18.	IF { sex=female&& religion=buddhist &&smoking=never_smoker&&exercise=no_exercise&& height=4 } THEN { PA_level=low }	1	1
19.	IF { smoking=never_smoker&&exercise=no_exercise&& height=3 } THEN { PA_level=low }	1	1
20.	IF { sex=female&& religion=buddhist && exercise=no_exercise&& height=3 } THEN {PA_level=low }	1	1
21.	IF { sex=female&& exercise=no_exercise&& waist=3} THEN { PA_level=low }	1	1
22.	IF { smoking=never_smoker&&exercise=no_exercise&& waist=3} THEN PA_level=low }	1	1
23.	IF{ religion=buddhist && exercise=no_exercise&& waist=3 } THEN { PA_level=low }	1	1
24.	IF { sex=female&& smoking=never_smoker&& exercise=no_exercise&& height=3 } THEN { PA_level=low }	1	1
25.	IF{ sex=female&& smoking=never_smoker&& exercise=no_exercise&& waist=3 }	1	1

	THEN { PA_level=low }		

ตารางที่ 4.12กฎความสัมพันธ์ที่ผ่านการจัดอันดับรวมกฎความสัมพันธ์ (ต่อ)

Rank	**Rules:**	**FMR**	**CAA**
26.	IF { religion=buddhist &&smoking=never_smoker&& exercise=no_exercise&& height=3 } THEN { PA_level=low }	1	1
27.	IF { sex=female&& religion=buddhist && exercise=no_exercise&& waist=3 } THEn { PA_level=low }	1	1
28.	IF { religion=buddhist && smoking=never_smoker&& exercise=no_exercise&& waist=3 } THEN { PA_level=low }	1	1
29.	IF { exercise=no_exercise&& bodyfat=6 } THEN { PA_level=low }	1	0.99487
30.	IF { exercise=no_exercise&& bodyfat=5 } THEN { PA_level=low }	1	0.99486
31.	IF { exercise=no_exercise&& height=5 } THEN { PA_level=low }	1	0.99486
32.	IF { exercise=no_exercise&& waist=5 } THEN { PA_level=low }	1	0.99486
33.	IF { exercise=no_exercise&& weight=2 } THEN { PA_level=low }	1	0.99485
34.	IF { exercise=no_exercise&& waist=4 } THEN { PA_level=low }	1	0.99485
35.	IF { exercise=no_exercise&& bodyfat=7 } THEN { PA_level=low }	1	0.99483
36.	IF { exercise=no_exercise&& weight=4 } THEN { PA_level=low }	1	0.99482

37.	IF { exercise=no_exercise&& weight=5 } THEN { PA_level=low }	1	0.99481

ตารางที่ 4.12กฎความสัมพันธ์ที่ผ่านการจัดอันดับกฎความสัมพันธ์ (ต่อ)

Rank	**Rules:**	**FMR**	**CAA**
38.	IF { age=4&& exercise=no_exercise } THEN { PA_level=low }	1	0.99476
39.	IF { exercise=no_exercise&& bodyfat=4 } THEN { PA_level=low }	1	0.99473
40.	IF { age=3&& exercise=no_exercise } THEN { PA_level=low }	1	0.9947
41.	IF { age=5&& exercise=no_exercise } THEN { PA_level=low }	1	0.9947
42.	IF { exercise=no_exercise&& waist=2 } THEN { PA_level=low }	1	0.9947
43.	IF { age=6&& exercise=no_exercise } THEN { PA_level=low }	1	0.99466
44.	IF { exercise=no_exercise&& height=2 } THEN { PA_level=low }	1	0.99462
45.	IF { exercise=no_exercise&& waist=6 } THEN { PA_level=low }	1	0.99462
46.	IF { exercise=no_exercise&& bodyfat=8 } THEN { PA_level=low }	1	0.99462
47.	IF { age=7&& exercise=no_exercise } THEN { PA_level=low }	1	0.99458
48.	IF { age=8 &&exercise=no_exercise } THEN { PA_level=low }	1	0.99454
49.	IF{ age=1&& exercise=no_exercise } THEN { PA_level=low }	1	0.99451
50.	IF{ exercise=no_exercise&& weight=6 }	1	0.99441

	THEN { PA_level=low }		

ตารางที่ 4.12กฎความสัมพันธ์ที่ผ่านการจัดอันดับกฎความสัมพันธ์ (ต่อ)

Rank	**Rules:**	**FMR**	**CAA**
51.	IF{ age=2&& exercise=no_exercise } THEN { PA_level=low }	1	0.99433
52.	IF{ exercise=no_exercise&& height=6 } THEN { PA_level=low }	1	0.99429
53.	IF{ exercise=no_exercise && bodyfat=3 } THEN { PA_level=low }	1	0.99411
54.	IF{ exercise=no_exercise&& waist=7 } THEN { PA_level=low }	1	0.99364
55.	IF { religion = buddhist && exercise = exercise && height = 3} THEN { PA_level = moderate }	1	0.235903
56.	IF { sex = female && exercise = exercise &&height = 3} THEN { PA_level = moderate }	1	0.235903
57.	IF { sex = female && religion = buddhist && exercise = exercise && height = 3 } THEN { PA_level = moderate }	1	0.235012
58.	IF { exercise = exercise && bodyfat = 6 } THEN { PA_level = moderate }	1	0.229250
59.	IF { religion = buddhist && exercise = exercise &&bodyfat = 6 } THEN { PA_level = moderate }	1	0.228738
60.	IF { sex = female && exercise = exercise && bodyfat = 6 } THEN { PA_level = moderate }	1	0.227570
61.	IF { sex = female && religion = buddhist && exercise = exercise && bodyfat = 6 } THEN { PA_level = moderate }	1	0.227318

ตารางที่ 4.12กฎความสัมพันธ์ที่ผ่านการจัดอันดับกฎความสัมพันธ์ (ต่อ)

Rank	Rules:	FMR	CAA
62.	IF { sex = female && smoking = never_smoker && exercise = exercise && bodyfat = 6 } THEN { PA_level = moderate }	1	0.223198
63.	IF { exercise = exercise && weight = 3 } THEN { PA_level = moderate }	1	0.221723
64.	IF { religion = buddhist && exercise = exercise &&weight = 3 } THEN { PA_level = moderate }	1	0.219779
65.	IF { smoking = never_smoker && exercise = exercise && bodyfat = 6 } THEN { PA_level = moderate }	1	0.219317
66.	IF { religion = buddhist && smoking = never_smoker && exercise = exercise && bodyfat = 6 } THEN { PA_level = moderate }	1	0.219052
67.	IF { exercise = exercise && height = 4 } THEN { PA_level = moderate }	1	0.215500
68.	IF { religion = buddhist && exercise = exercise && height = 4} THEN { PA_level = moderate }	1	0.215152
69.	IF { sex = female && exercise = exercise && height = 4} THEN { PA_level = moderate }	1	0.214898
70.	IF { smoking = never_smoker && exercise = exercise && weight = 3 } THEN { PA_level = moderate }	1	0.213863
71.	IF { religion = buddhist && smoking = never_smoker && exercise = exercise && weight = 3 } THEN { PA_level = moderate }	1	0.212156

ตารางที่ 4.12 กฎความสัมพันธ์ที่ผ่านการจัดอันดับกฎความสัมพันธ์ (ต่อ)

Rank	Rules:	FMR	CAA
72.	IF { religion = buddhist && exercise = exercise && weight = 3 && bodyfat = 6 } THEN { PA_level = moderate }	1	0.199310
73.	IF { exercise = exercise && weight = 3 && bodyfat = 6 } THEN { PA_level = moderate }	1	0.198611
74.	IF { religion = buddhist && smoking = never_smoker && exercise = exercise && height = 4 } THEN { PA_level = moderate }	1	0.193302
75.	IF { religion = buddhist && exercise = exercise && weight = 4} THEN { PA_level = moderate }	1	0.189896
76.	IF {sex = male &&smoking = never_smoker && exercise = exercise && height = 4 } THEN { PA_level = moderate }	1	0.18919
77.	IF { smoking = never_smoker && exercise = exercise && height = 4} THEN { PA_level = moderate }	1	0.189183
78.	IF {sex = female &&smoking = never_smoker && exercise = exercise && height = 4 } THEN { PA_level = moderate }	1	0.18918
79.	IF { age = 3 && sex = female && smoking = never_smoker && exercise = exercise && height = 4 } THEN { PA_level = moderate }	1	0.189175
80.	IF { exercise = exercise && weight = 4 } THEN { PA_level = moderate }	1	0.188025
81.	IF { exercise = exercise && waist = 4 }	1	0.177064

	THEN { PA_level = moderate }		

ตารางที่ 4.12 กฎความสัมพันธ์ที่ผ่านการจัดอันดับกฎความสัมพันธ์ (ต่อ)

Rank	Rules:	FMR	CAA
82.	IF{ religion = buddhist && exercise = exercise && waist = 4 } THEN { PA_level = moderate }	1	0.176023
83.	IF{ exercise = exercise && waist = 5 } THEN { PA_level = moderate }	1	0.170662

ขั้นตอนวิธีที่ 5 การขจัดกฎความสัมพันธ์ที่ซ้ำซ้อน

จากกฎความสัมพันธ์ที่ผ่านการจัดอับดับ นำมาขจัดกฎความสัมพันธ์ที่ซ้ำซ้อนทำให้จำนวนกฎ 83 กฎ ลดลงเหลือ 36 กฎ แสดงได้ดังตารางที่ 4.13

ตารางที่ 4.13กฎความสัมพันธ์ที่ผ่านการขจัดกฎความสัมพันธ์ที่ซ้ำซ้อน

Rank	Rules:	FMR	CAA
1.	IF { exercise=no_exercise&& weight=3 } THEN { PA_level=low }	3	0.98914
2.	IF { exercise=no_exercise&& height=4} THEN { PA_level=low }	3	0.98897
3.	IF { exercise=no_exercise&& waist=3 } THEN { PA_level=low }	3	0.98683
4.	IF { exercise=no_exercise && height=3 } THEN { PA_level=low }	2	0.997455
5.	IF { exercise=no_exercise&& bodyfat=6 } THEN { PA_level=low }	1	0.99487
6.	IF { exercise=no_exercise&& bodyfat=5 } THEN { PA_level=low }	1	0.99486
7.	IF { exercise=no_exercise&& height=5 } THEN { PA_level=low }	1	0.99486

ตารางที่ 4.13 กฎความสัมพันธ์ที่ผ่านการขจัดกฎความสัมพันธ์ที่ซ้ำซ้อน (ต่อ)

Rank	Rules:	FMR	CAA
8.	IF { exercise=no_exercise&& waist=5 } THEN { PA_level=low }	1	0.99486
9.	IF { exercise=no_exercise&& weight=2 } THEN { PA_level=low }	1	0.99485
10.	IF { exercise=no_exercise&& waist=4 } THEN { PA_level=low }	1	0.99485
11.	IF { exercise=no_exercise&& bodyfat=7 } THEN { PA_level=low }	1	0.99483
12.	IF { exercise=no_exercise&& weight=4 } THEN { PA_level=low }	1	0.99482
13.	IF { exercise=no_exercise&& weight=5 } THEN { PA_level=low }	1	0.99481
14.	IF { age=4&& exercise=no_exercise } THEN { PA_level=low }	1	0.99476
15.	IF { exercise=no_exercise&& bodyfat=4 } THEN { PA_level=low }	1	0.99473
16.	IF { age=3&& exercise=no_exercise } THEN { PA_level=low }	1	0.9947
17.	IF { age=5&& exercise=no_exercise } THEN { PA_level=low }	1	0.9947
18.	IF { exercise=no_exercise&& waist=2 } THEN { PA_level=low }	1	0.9947
19.	IF { age=6&& exercise=no_exercise } THEN { PA_level=low }	1	0.99466
20.	IF { exercise=no_exercise&& height=2 }	1	0.99462

	THEN { PA_level=low }		

ตารางที่ 4.13 กฎความสัมพันธ์ที่ผ่านการขจัดกฎความสัมพันธ์ที่ซ้ำซ้อน (ต่อ)

Rank	**Rules:**	**FMR**	**CAA**
21.	IF { exercise=no_exercise&& waist=6 } THEN { PA_level=low }	1	0.99462
22.	IF { exercise=no_exercise&& bodyfat=8 } THEN { PA_level=low }	1	0.99462
23.	IF { age=7&& exercise=no_exercise } THEN { PA_level=low }	1	0.99458
24.	IF { age=8 &&exercise=no_exercise } THEN { PA_level=low }	1	0.99454
25.	IF{ age=1&& exercise=no_exercise } THEN { PA_level=low }	1	0.99451
26.	IF{ exercise=no_exercise&& weight=6 } THEN { PA_level=low }	1	0.99441
27.	IF{ age=2&& exercise=no_exercise } THEN { PA_level=low }	1	0.99433
28.	IF{ exercise=no_exercise&& height=6 } THEN { PA_level=low }	1	0.99429
29.	IF{ exercise=no_exercise && bodyfat=3 } THEN { PA_level=low }	1	0.99411
30.	IF{ exercise=no_exercise&& waist=7 } THEN { PA_level=low }	1	0.99364
31.	IF { exercise = exercise && bodyfat = 6 } THEN { PA_level = moderate }	1	0.229250
32.	IF { exercise = exercise && weight = 3 } THEN { PA_level = moderate }	1	0.221723
33.	IF { exercise = exercise && height = 4 }	1	0.215500

	THEN { PA_level = moderate }		

ตารางที่ 4.13 กฎความสัมพันธ์ที่ผ่านการขจัดกฎความสัมพันธ์ที่ซ้ำซ้อน (ต่อ)

Rank	**Rules:**	**FMR**	**CAA**
34.	IF { exercise = exercise && weight = 4 } THEN { PA_level = moderate }	1	0.188025
35.	IF { exercise = exercise && waist = 4 } THEN { PA_level = moderate }	1	0.177064
36.	IF{ exercise = exercise && waist = 5 } THEN { PA_level = moderate }	1	0.170662

4.2.2การตีความของกฎความสัมพันธ์ (Interpreting rules)

จากตารางที่ 4.13 มีกฎความสัมพันธ์ทั้งสิ้น 36 กฎ ซึ่งทั้ง 36 กฎความสัมพันธ์นี้สามารถตีความได้ดังนี้

กฎความสัมพันธ์ลำดับที่1

IF{ exercise=no_exercise&& weight=3 } **THEN** { PA_level=low }

FMR = 3 CAA = 0.98914

ตีความได้ว่าถ้า ผู้ป่วยไม่มีประวัติการออกกำลังกายและน้ำหนักอยู่ในช่วงที่ 3 (52.43-60.03) แล้วPA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 3 และมีค่าความเชื่อมั่นแม่นตรงที่ 0.98914

กฎความสัมพันธ์ลำดับที่ 2

IF{ exercise=no_exercise&& height=4 } **THEN** { PA_level=low }

FMR = 3 CAA = 0.98897

ตีความได้ว่า ถ้า ผู้ป่วยไม่มีประวัติการออกกำลังกายและความสูงอยู่ในช่วงที่ 4 (156.4-161.4) แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 3 และมีค่าความเชื่อมั่นแม่นตรงที่ 0.98897

กฎความสัมพันธ์ลำดับที่ 3

IF{ exercise=no_exercise&& waist=3 } **THEN** { PA_level=low }

FMR = 3 CAA = 0.98683

ตีความได้ว่า ถ้า ผู้ป่วยไม่มีประวัติการออกกำลังกายและรอบเอวอยู่ในช่วงที่ 3 (69.5-75.6) แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 3 และมีค่าความเชื่อมั่นแม่นตรงที่ 0.98683

กฎความสัมพันธ์ลำดับที่ 4

IF{ exercise=no_exercise&& height=3 } **THEN** { PA_level=low }

FMR = 2 CAA = 0.99746

ตีความได้ว่า ถ้า ผู้ป่วยไม่มีประวัติการออกกำลังกายและความสูงอยู่ในช่วงที่ 3 (151.3-156.3) แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 2 และมีค่าความเชื่อมั่นแม่นตรงที่ 0.99746

กฎความสัมพันธ์ลำดับที่ 5

IF{ exercise=no_exercise&& bodyfat=6 } **THEN** { PA_level=low }

FMR = 1 CAA = 0.99487

ตีความได้ว่า ถ้า ผู้ป่วยไม่มีประวัติการออกกำลังกายและดัชนีมวลกายอยู่ในช่วงที่ 6 (29.66-33.24) แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1 และมีค่าความเชื่อมั่นแม่นตรงที่ 0.99487

กฎความสัมพันธ์ลำดับที่ 6

IF{ exercise=no_exercise&& bodyfat=5 } **THEN** { PA_level=low }

FMR = 1 CAA = 0.99486

ตีความได้ว่า ถ้า ผู้ป่วยไม่มีประวัติการออกกำลังกายและดัชนีมวลกายอยู่ในช่วงที่ 5 (26.07-29.65) แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1

และมีค่าความเชื่อมั่นแม่นตรงที่ 0.99486

กฎความสัมพันธ์ลำดับที่ 7

IF{ exercise=no_exercise&& height=5 } **THEN** { PA_level=low }

FMR = 1 CAA = 0.99486

ตีความได้ว่า ถ้า ผู้ป่วยไม่มีประวัติการออกกำลังกายและความสูงอยู่ในช่วงที่ 5 (161.5-166.5) แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1 และมีค่าความเชื่อมั่นแม่นตรงที่ 0.99486

กฎความสัมพันธ์ลำดับที่ 8

IF{ exercise=no_exercise&& waist=5 } **THEN** { PA_level=low }

FMR = 1 CAA = 0.99486

ตีความได้ว่า ถ้า ผู้ป่วยไม่มีประวัติการออกกำลังกายและรอบเอวอยู่ในช่วงที่ 5 (81.9-88) แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1 และมีค่าความเชื่อมั่นแม่นตรงที่ 0.99486

กฎความสัมพันธ์ลำดับที่ 9

IF{ exercise=no_exercise&& weight=2 } **THEN** { PA_level=low }

FMR = 1 CAA = 0.99485

ตีความได้ว่า ถ้า ผู้ป่วยไม่มีประวัติการออกกำลังกายและน้ำหนักอยู่ในช่วงที่ 2 (44.82-52.42) แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1 และมีค่าความเชื่อมั่นแม่นตรงที่ 0.99485

กฎความสัมพันธ์ลำดับที่ 10

IF{ exercise=no_exercise&& waist=4 } **THEN** { PA_level=low }

FMR = 1 CAA = 0.99485

ตีความได้ว่า ถ้า ผู้ป่วยไม่มีประวัติการออกกำลังกายและรอบเอวอยู่ในช่วงที่ 4 (75.7-81.8) แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1 และมีค่าความเชื่อมั่นแม่นตรงที่ 0.99485

กฎความสัมพันธ์ลำดับที่ 11

IF{ exercise=no_exercise&&bodyfat=7 } **THEN** { PA_level=low }

FMR = 1 CAA = 0.99483

ตีความได้ว่า ถ้า ผู้ป่วยไม่มีประวัติการออกกำลังกายและดัชนีมวลกายอยู่ในช่วงที่ 7 (33.25-36.83)

แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1

และมีค่าความเชื่อมั่นแม่นตรงที่ 0.99483

กฎความสัมพันธ์ลำดับที่ 12

IF{ exercise=no_exercise&& weight=4 } **THEN** { PA_level=low }

FMR = 1 CAA = 0.99482

ตีความได้ว่า ถ้า ผู้ป่วยไม่มีประวัติการออกกำลังกายและน้ำหนักอยู่ในช่วงที่ 4 (60.04-67.64)

แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1

และมีค่าความเชื่อมั่นแม่นตรงที่ 0.99482

กฎความสัมพันธ์ลำดับที่ 13

IF{ exercise=no_exercise&& weight=5 } **THEN** { PA_level=low }

FMR = 1 CAA = 0.99481

ตีความได้ว่า ถ้า ผู้ป่วยไม่มีประวัติการออกกำลังกายและน้ำหนักอยู่ในช่วงที่ 5 (67.65-75.25)

แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1

และมีค่าความเชื่อมั่นแม่นตรงที่ 0.99481

กฎความสัมพันธ์ลำดับที่ 14

IF{ age = 4 &&exercise=no_exercise } **THEN** { PA_level=low }

FMR = 1 CAA = 0.99476

ตีความได้ว่า ถ้า ผู้ป่วยมีอายุอยู่ในช่วงที่ 4 (43.8-46.6) และไม่มีประวัติการออกกำลังกาย

แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1

และมีค่าความเชื่อมั่นแม่นตรงที่ 0.99476

กฎความสัมพันธ์ลำดับที่ 15

IF{exercise=no_exercise&& bodyfat=4} **THEN** { PA_level=low }

FMR = 1 CAA = 0.99473

ตีความได้ว่า ถ้า ผู้ป่วยไม่มีประวัติการออกกำลังกายและดัชนีมวลกายอยู่ในช่วงที่ 4 (22.48-26.06)

แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1

และมีค่าความเชื่อมั่นแม่นตรงที่ 0.99473

กฎความสัมพันธ์ลำดับที่ 16

IF{ age = 3 &&exercise=no_exercise } **THEN** { PA_level=low }

FMR = 1 CAA = 0.9947

ตีความได้ว่า ถ้า ผู้ป่วยมีอายุอยู่ในช่วงที่ 3 (40.9-43.6) และไม่มีประวัติการออกกำลังกาย

แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1

และมีค่าความเชื่อมั่นแม่นตรงที่ 0.9947

กฎความสัมพันธ์ลำดับที่ 17

IF{ age = 5 &&exercise=no_exercise } **THEN** { PA_level=low }

FMR = 1 CAA = 0.9947

ตีความได้ว่า ถ้า ผู้ป่วยมีอายุอยู่ในช่วงที่ 5 (46.7-49.5) และไม่มีประวัติการออกกำลังกาย

แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1

และมีค่าความเชื่อมั่นแม่นตรงที่ 0.9947

กฎความสัมพันธ์ลำดับที่ 18

IF{exercise=no_exercise&&waist=2} **THEN** { PA_level=low }

FMR = 1 CAA = 0.9947

ตีความได้ว่า ถ้า ผู้ป่วยไม่มีประวัติการออกกำลังกายและรอบเอวอยู่ในช่วงที่ 2(63.3-69.4)

แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1

และมีค่าความเชื่อมั่นแม่นตรงที่ 0.9947

กฎความสัมพันธ์ลำดับที่ 19

IF{ age = 6 &&exercise=no_exercise } **THEN** { PA_level=low }

FMR = 1 CAA = 0.99466

ตีความได้ว่า ถ้า ผู้ป่วยมีอายุอยู่ในช่วงที่ 6 (49.6-52.4) และไม่มีประวัติการออกกำลังกาย แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1 และมีค่าความเชื่อมั่นแม่นตรงที่ 0.99466

กฎความสัมพันธ์ลำดับที่ 20

IF{exercise=no_exercise&&height=2} **THEN** { PA_level=low }

FMR = 1 CAA = 0.99466

ตีความได้ว่า ถ้า ผู้ป่วยไม่มีประวัติการออกกำลังกายและความสูงอยู่ในช่วงที่ 2 (146.2-151.2) แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1 และมีค่าความเชื่อมั่นแม่นตรงที่ 0.99466

กฎความสัมพันธ์ลำดับที่ 21

IF{exercise=no_exercise&&waist=6} **THEN** { PA_level=low }

FMR = 1 CAA = 0.99462

ตีความได้ว่า ถ้า ผู้ป่วยไม่มีประวัติการออกกำลังกายและรอบเอวอยู่ในช่วงที่ 6 (88.1-94.2) แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1 และมีค่าความเชื่อมั่นแม่นตรงที่ 0.99462

กฎความสัมพันธ์ลำดับที่ 22

IF{exercise=no_exercise&&bodyfat=8} **THEN** { PA_level=low }

FMR = 1 CAA = 0.99462

ตีความได้ว่า ถ้า ผู้ป่วยไม่มีประวัติการออกกำลังกายและดัชนีมวลกายอยู่ในช่วงที่ 8 (36.84-40.42) แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1 และมีค่าความเชื่อมั่นแม่นตรงที่ 0.99462

กฎความสัมพันธ์ลำดับที่ 23

IF{ age = 7 &&exercise=no_exercise } **THEN** { PA_level=low }

FMR = 1 CAA = 0.99458

ตีความได้ว่า ถ้า ผู้ป่วยมีอายุอยู่ในช่วงที่ 7 (52.5-55.3) และไม่มีประวัติการออกกำลังกาย
แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1
และมีค่าความเชื่อมั่นแม่นตรงที่ 0.99458

กฎความสัมพันธ์ลำดับที่ 24

IF{ age = 8&&exercise=no_exercise } **THEN** { PA_level=low }

FMR = 1 CAA = 0.99454

ตีความได้ว่า ถ้า ผู้ป่วยมีอายุอยู่ในช่วงที่ 8(55.4-58.2) และไม่มีประวัติการออกกำลังกาย
แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1
และมีค่าความเชื่อมั่นแม่นตรงที่ 0.99454

กฎความสัมพันธ์ลำดับที่25

IF{ age = 1&&exercise=no_exercise } **THEN** { PA_level=low }

FMR = 1 CAA = 0.99451

ตีความได้ว่า ถ้า ผู้ป่วยมีอายุอยู่ในช่วงที่ 1 (0-37.9) และไม่มีประวัติการออกกำลังกาย
แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1
และมีค่าความเชื่อมั่นแม่นตรงที่ 0.99451

กฎความสัมพันธ์ลำดับที่26

IF{exercise=no_exercise&&weight=6} **THEN** { PA_level=low }

FMR = 1 CAA = 0.99441

ตีความได้ว่า ถ้า ผู้ป่วยไม่มีประวัติการออกกำลังกายและน้ำหนักอยู่ในช่วงที่ 6 (75.26-82.86)
แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1
และมีค่าความเชื่อมั่นแม่นตรงที่ 0.99441

กฎความสัมพันธ์ลำดับที่ 27

IF{ age = 2 &&exercise=no_exercise } **THEN** { PA_level=low }

FMR = 1 CAA = 0.99433

ตีความได้ว่า ถ้า ผู้ป่วยมีอายุอยู่ในช่วงที่ 2 (38-40.8) และไม่มีประวัติการออกกำลังกาย แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1 และมีค่าความเชื่อมั่นแม่นตรงที่ 0.99433

กฎความสัมพันธ์ลำดับที่ 28

IF{exercise=no_exercise&&height=6} **THEN** { PA_level=low }

FMR = 1 CAA = 0.99429

ตีความได้ว่า ถ้า ผู้ป่วยไม่มีประวัติการออกกำลังกายและความสูงอยู่ในช่วงที่ 6 (166.6-171.6) แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1 และมีค่าความเชื่อมั่นแม่นตรงที่ 0.99429

กฎความสัมพันธ์ลำดับที่ 29

IF{exercise=no_exercise&&bodyfat=3} **THEN** { PA_level=low }

FMR = 1 CAA = 0.99411

ตีความได้ว่า ถ้า ผู้ป่วยไม่มีประวัติการออกกำลังกายและดัชนีมวลกายอยู่ในช่วงที่ 3 (18.89-22.47) แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1 และมีค่าความเชื่อมั่นแม่นตรงที่ 0.99411

กฎความสัมพันธ์ลำดับที่ 30

IF{exercise=no_exercise&&waist=7} **THEN** { PA_level=low }

FMR = 1 CAA = 0.99364

ตีความได้ว่า ถ้า ผู้ป่วยไม่มีประวัติการออกกำลังกายและรอบเอวอยู่ในช่วงที่ 7(94.3-100.4) แล้ว PA_level อยู่ในคลาส lowมีค่าความถี่ของการรวมกฎเป็น 1 และมีค่าความเชื่อมั่นแม่นตรงที่ 0.99364

กฎความสัมพันธ์ลำดับที่ 31

IF{ exercise=exercise&&bodyfat=6} **THEN** { PA_level=moderate }

FMR = 1 CAA = 0.22925

ตีความได้ว่า ถ้า ผู้ป่วยมีประวัติการออกกำลังกายและดัชนีมวลกายอยู่ในช่วงที่ 6 (29.66-33.24)

แล้ว PA_level อยู่ในคลาส moderateมีค่าความถี่ของการรวมกฎเป็น 1

และมีค่าความเชื่อมั่นแม่นตรงที่ 0.229250

กฎความสัมพันธ์ลำดับที่ 32

IF{ exercise=exercise&&weight=3} **THEN** { PA_level=moderate }

FMR = 1 CAA = 0.22172

ตีความได้ว่า ถ้า ผู้ป่วยมีประวัติการออกกำลังกายและน้ำหนักอยู่ในช่วงที่ 3(52.43-60.03)

แล้ว PA_level อยู่ในคลาส moderateมีค่าความถี่ของการรวมกฎเป็น 1

และมีค่าความเชื่อมั่นแม่นตรงที่ 0.22172

กฎความสัมพันธ์ลำดับที่ 33

IF{ exercise=exercise&&height=4} **THEN** { PA_level=moderate }

FMR = 1 CAA = 0.22155

ตีความได้ว่า ถ้า ผู้ป่วยมีประวัติการออกกำลังกายและความสูงอยู่ในช่วงที่ 4(156.4-161.4)

แล้ว PA_level อยู่ในคลาส moderateมีค่าความถี่ของการรวมกฎเป็น 1

และมีค่าความเชื่อมั่นแม่นตรงที่ 0.22155

กฎความสัมพันธ์ลำดับที่ 34

IF{ exercise=exercise&&weight=4} **THEN** { PA_level=moderate }

FMR = 1 CAA = 0.18803

ตีความได้ว่า ถ้า ผู้ป่วยมีประวัติการออกกำลังกายและน้ำหนักอยู่ในช่วงที่ 4(60.04-67.64)

แล้ว PA_level อยู่ในคลาส moderateมีค่าความถี่ของการรวมกฎเป็น 1

และมีค่าความเชื่อมั่นแม่นตรงที่ 0.18803

กฎความสัมพันธ์ลำดับที่ 35

IF{ exercise=exercise&&waist=4} **THEN** { PA_level=moderate }

FMR = 1 CAA = 0.17706

ตีความได้ว่า ถ้า ผู้ป่วยมีประวัติการออกกำลังกายและรอบเอวอยู่ในช่วงที่ 4 (75.7-81.8) แล้ว PA_level อยู่ในคลาส moderateมีค่าความถี่ของการรวมกฎเป็น 1 และมีค่าความเชื่อมั่นแม่นตรงที่ 0.17706

กฎความสัมพันธ์ลำดับที่ 36

IF{ exercise=exercise&&waist=5} **THEN** { PA_level=moderate }

FMR = 1 CAA = 0.17066

ตีความได้ว่า ถ้า ผู้ป่วยมีประวัติการออกกำลังกายและรอบเอวอยู่ในช่วงที่ 5 (81.9-88) แล้ว PA_level อยู่ในคลาส moderateมีค่าความถี่ของการรวมกฎเป็น 1 และมีค่าความเชื่อมั่นแม่นตรงที่ 0.17066

4.3 การทดสอบเปรียบเทียบกับงานวิจัยอื่นโดยใช้ข้อมูลโรคหัวใจ

งานวิจัยของ Jesminและคณะได้ใช้อัลกอริทึมการหากฎความสัมพันธ์พื้นฐานอันได้แก่ วิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรีวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทีพเอไพรออรีและวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมเทอเชียส ซึ่งจะใช้ในการวิเคราะห์ข้อมูลผู้ป่วยโรคหัวใจ ที่ได้จากฐานข้อมูลมาตรฐานUC Irvineดังนั้นในส่วนนี้จะทำการเปรียบเทียบกฎความสัมพันธ์ที่หาได้โดยวิธีเดียวกันกับข้อมูลโรคหัวใจซึ่งเปรียบเทียบกับค่าความเชื่อมั่นและค่าความแม่นตรงที่ได้จากงานวิจัยของ Jesminและคณะดังตารางที่ 4.14และรูปที่ 4.1

ตารางที่ 4.14 กฎความสัมพันธ์ที่ได้จากวิธีที่วิทยานิพนธ์นี้เสนอ

Rank	Rules:	FMR	CAA
1.	IF { MaxHr='[162.7-175.8]'&& Oldpeak='[0-0.62]' && Thal=normal } THEN {Class=absence }	**2**	**0.981435**
2.	IF { MaxHr='[162.7-175.8]'&& Slope=up&& Thal=normal } THEN { Class=absence }	**2**	**0.981435**
3.	IF { MaxHr='[162.7-175.8]' && Ca='[0-0.3]' && Thal=normal } THEN { Class=absence }	**2**	**0.981335**
4.	IF { Sex=female&& Cpt=asympt&& Ca='[0.9-1.2]' } THEN { Class=presence }	**2**	**0.98093**
5.	IF { Restecg=norm && MaxHr='[162.7-175.8]' && Exang=no && Thal=normal } THEN { Class=absence }	**2**	**0.96652**
6.	IF {Fbs=F&& MaxHr='[162.7-175.8]'&& Exang=no&& Thal=normal } THEN { Class=absence }	**1**	**0.97**
7.	IF { Fbs=F&& MaxHr='[162.7-175.8]'&&Ca='[-inf-0.3]' &&Thal=normal } THEN { Class=absence }	**1**	**0.97**
8.	IF { Rbp='[125.8-136.4]'&& Exang=no&& Ca='[-inf-0.3]' } THEN { Class=absence }	**1**	**0.96**
9.	IF { Sex=female&& Cpt=asympt&& Fbs=F&& Ca='[0.9-1.2]' } THEN { Class=presence }	**1**	**0.96**
10.	IF { MaxHr='[162.7-175.8]'&& Exang=no&& Oldpeak='[-inf-0.62]'&& Thal=normal } THEN { Class=absence }	**1**	**0.96**

ตารางที่ 4.14 กฎความสัมพันธ์ที่ได้จากวิธีที่วิทยานิพนธ์นี้เสนอ (ต่อ)

Rank	Rules:	FMR	CAA
11.	IF { Sex=male&& Restecg=norm &&Exang=no&& Ca='[-inf-0.3]'} THEN { Class=absence }	**1**	**0.99346**
12.	IF { Sex=male&& Cpt=notang&& Ca='[-inf-0.3]' THEN { Class=absence }	**1**	**0.99334**
13.	IF { Restecg=norm && MaxHr='[162.7-175.8]' && Exang=no &&Thal=normal } THEN { Class=absence }	**1**	**0.99304**
14.	IF { Ca='[1.8-2.1]' && Thal=reversible_defect } THEN { Class=presence }	**1**	**0.99217**
15.	IF { Sex=female && Cpt=asympt && Restecg=hyp && Ca='[0.9-1.2]' } THEN { Class=presence }	**1**	**0.99186**
16.	IF { Exang=yes && Ca='[1.8-2.1]' } THEN { Class=presence }	**1**	**0.99107**
17.	IF { Ca='[-inf-0.3]'&& Thal = normal } THEN { Class=absence }	**1**	**0.540702**
18.	IF {Cpt = asympt } THEN { Class=presence}	**1**	**0.499227**

จากตารางที่ 4.14 กฎความสัมพันธ์ลำดับที่ 1 ถึงลำดับที่ 5 พบว่ามาจากการรวมกันของวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรีและวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทีพเอไพรออรี สามารถระบุคลาสได้ทั้งสองคลาส ประกอบด้วยคลาส absence คลาส presence ดังนั้นจะทำการเปรียบเทียบด้วยมาตรวัด CAA โดยเทียบระหว่างกฎความสัมพันธ์ของวิทยานิพนธ์ฉบับนี้กับกฎความสัมพันธ์ที่ได้จากงานวิจัยของ Jesmin ที่แสดงได้ดังตารางที่ 4.15

ตารางที่ 4.15กฎความสัมพันธ์ที่ได้จากงานวิจัยของ Jesmin

Rank	**Rules:**	**FMR**	**CAA**
1.	IF { Sex = female && Exang = no && Ca = 0 } THEN { Class=absence}	**2**	**0.9901**
2.	IF { Sex = female && Exang = no && Ca = 0 && Thal = normal } THEN { Class=absence}	**1**	**0.98**
3.	IF { Sex = female && Fbs = F && Ca = 0 } THEN { Class=absence}	**1**	**0.98**
4.	IF { Sex = female && Fbs = f && Exang = no && Thal = normal } THEN { Class=absence}	**1**	**0.95**
5.	IF { Rbp='[115.2-136.4]&& Exang = no && Ca = 0 && Thal = normal } THEN { Class=absence}	**1**	**0.94**
6.	IF { Cpt=asympt && Slope = flat && Thal = rev} THEN { Class=presence }	**1**	**0.96**
7.	IF { Cpt=asympt && Exang = yes && Thal = rev} THEN { Class=presence }	**1**	**0.94**
8.	IF { Sex = female && Fbs = f && Restecg && Exang = no && Thal = normal } THEN { Class=absence }	**1**	**0.9938**
9.	IF { Sex = female && Cpt=notang && Thal = normal }	**1**	**0.9935**

	THEN { Class=absence }		

ตารางที่ 4.15กฎความสัมพันธ์ที่ได้จากงานวิจัยของ Jesmin (ต่อ)

Rank	**Rules:**	**FMR**	**CAA**
10.	IF { Age='[48.2-57.8]'&& MaxHr='[149.6-175.8]'Exang = no && Ca = 0 } THEN {Class=absence }	**1**	**0.99314**
11.	IF {Age= '[36.6-48.2]'&&Rbp='[115.2-136.4]&&Thal = normal} THEN {Class=absence }	**1**	**0.9918**
12.	IF { Age=' [48.2-57.8]' && Slope = flat && Ca = 1 } THEN { Class=presence }	**1**	**0.9902**
13.	IF { MaxHr='[123.4-149.6]'&&Exang = Yes&& Thal=rev } THEN { Class=presence }	**1**	**0.9931**
14.	IF { Sex=male && Cpt=asympt && Ca=2} THEN { Class=presence }	**1**	**0.9915**
15.	IF {Age=' [57.8-67.4]' && Sex=male && Ca=2} THEN { Class=presence }	**1**	**0.94**

ตารางที่ 4.16ผลสรุปการเปรียบเทียบอัลกอริทึมของวิทยานิพนธ์ฉบับนี้กับงานวิจัยของ NaharJesminและคณะ

	กฎความสัมพันธ์ที่ดีที่สุด		กฎความสัมพันธ์ที่ดีที่สุด	
	Class = absence		Class = presence	
อัลกอริทึมของวิทยานิพนธ์ฉบับนี้	IF { MaxHr='[162.7-175.8]'&& Oldpeak='[0-0.62]' && Thal=normal } THEN {Class=absence }		IF { Sex=female&& Cpt=asympt&& Ca='[0.9-1.2]' } THEN { Class=presence }	
	FMR = 2	**CAA = 0.981435**	**FMR = 2**	**CAA = 0.98093**
งานวิจัยของ Nahar Jesminและคณะ	IF { Sex = female && Exang = No && Ca = 0 } THEN { Class=absence}		IF { Cpt=asympt && Slope = flat && Thal = rev} THEN { Class=presence }	
	FMR = 2	**CAA = 0.98505**	**FMR = 1**	**CAA = 0.96**

จากตารางที่ 4.15 สามารถสรุปได้ว่า อัลกอริทึมของวิทยานิพนธ์ฉบับนี้ ให้ประสิทธิภาพของค่า CAA ค่อนข้างสูงมาก เมื่อเปรียบเทียบกับงานวิจัยของ Nahar Jesminและคณะ

4.4 อภิปรายผล

จากแนวคิดเบื้องต้นของอัลกอริทึมที่วิทยานิพนธ์นี้ได้เสนอซึ่งใช้ข้อมูลโรคหอบหืดที่เป็นข้อมูลตัวอย่างเบื้องต้นของแนวคิด และนำแนวคิดที่ได้มาทดสอบประสิทธิภาพกับข้อมูลทางการแพทย์ที่เป็นข้อมูลของผู้ป่วยโรคหัวใจที่ได้จากฐานข้อมูลมาตรฐาน เพื่อเปรียบเทียบกับงานวิจัยของ Jesminและคณะ ซึ่งพบว่ากฎความสัมพันธ์ที่สามารถนำมารวมได้นั้นมีค่า FMR สูงสุดที่ 2 เกิดจากวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรีรวมกับวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทีพเอไพรออรีและสามารถแสดงกฎของการจำแนกประเภทได้เป็น 2 คลาส ดังนั้นจึงทำการเปรียบเทียบได้ 2กรณีดังนี้

กรณีที่ 1 พิจารณากฎความสัมพันธ์จากกฎความสัมพันธ์ที่ดีที่สุดของการจำแนกประเภทคลาส ผู้ป่วยที่ไม่ได้เป็นโรคหัวใจ เปรียบเทียบกับค่า CAAและ FMRที่ได้จากอัลกอริทึมของวิทยานิพนธ์นี้ พบว่าค่า FMR จากอัลกอริทึมของวิทยานิพนธ์นี้มีค่าเท่ากันกับFMR ของงานวิจัย Jesminแต่อัลกอริทึมของวิทยานิพนธ์นี้ให้ค่า CAA ที่น้อยกว่าของงานวิจัย Jesmin

กรณีที่ 2พิจารณากฎความสัมพันธ์จากกฎความสัมพันธ์ที่ดีที่สุดของการจำแนกประเภทคลาส ผู้ป่วยที่เป็นโรคหัวใจ เปรียบเทียบกับค่า CAAและ FMRที่ได้จากอัลกอริทึมของวิทยานิพนธ์นี้ พบว่าค่า FMR และ CAA จากอัลกอริทึมของวิทยานิพนธ์นี้มีค่าสูงกว่า FMR และ CAA ของงานวิจัย Jesmin

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

ในปัจจุบันนั้นเทคนิคในด้านการค้นหากฎความสัมพันธ์มีอยู่มากมายหลากหลายเทคนิค แต่การเลือกเทคนิคหรือวิธีการค้นหากฎความสัมพันธ์ให้เหมาะสมกับข้อมูลที่สนใจขณะนั้น เป็นสิ่งสำคัญที่จะต้องคำนึงถึงเป็นอันดับแรก สำหรับงานทางด้านการค้นหากฎความสัมพันธ์ปัจจุบันได้ถูกนำมาใช้ได้หลากหลายวงการวิชาชีพ เช่น ด้านการตลาด ใช้สำหรับการหาความสัมพันธ์ของสินค้าที่ผู้บริโภคเลือกซื้อ ด้านการแพทย์ ใช้สำหรับการหาความสัมพันธ์ของความเสี่ยงที่จะนำไปสู่การเกิดโรคต่าง ๆ แต่การหากฎความสัมพันธ์แบบปกตินั้นพบว่ามีปัญหาในด้านของชนิดข้อมูลที่นำมาใช้ไม่สามารถกระทำกับชนิดของข้อมูลที่มีลักษณะเป็นตัวเลขดังนั้นงานวิจัยต่าง ๆ ที่ใช้ข้อมูลที่มีลักษณะตัวเลขมาทำการค้นหากฎความสัมพันธ์ มีความจำเป็นที่จะต้องใช้เทคนิคการแบ่งช่วงข้อมูลที่มีลักษณะเป็นตัวเลข ซึ่งเป็นกระบวนการจัดการข้อมูลก่อนการค้นหากฎความสัมพันธ์

ในงานวิจัยนี้มุ่งเน้นในกระบวนการวิเคราะห์ข้อมูลก่อนการค้นหากฎความสัมพันธ์และกระบวนการวิเคราะห์กฎความสัมพันธ์หลังการค้นหากฎความสัมพันธ์ สามารถแบ่งได้ดังนี้

- กระบวนการวิเคราะห์ข้อมูลก่อนการค้นหากฎความสัมพันธ์
  - กระบวนการลบข้อมูลที่ไม่สามารถระบุข้อมูลได้ชัดเจน
  - กระบวนการแบ่งช่วงข้อมูลที่มีลักษณะเป็นตัวเลข
- กระบวนการวิเคราะห์กฎความสัมพันธ์หลังการค้นหากฎความสัมพันธ์
  - กระบวนการรวมกฎความสัมพันธ์
  - กระบวนการจัดอันดับกฎความสัมพันธ์
  - กระบวนการขจัดกฎความสัมพันธ์ที่ซ้ำซ้อน

การทดสอบประสิทธิภาพจะใช้ชุดข้อมูลโรคหอบหืดซึ่งเป็นชุดข้อมูลจริง และชุดข้อมูลโรคหัวใจซึ่งเป็นชุดข้อมูลจากฐานข้อมูลมาตรฐานโดยใช้มาตรวัด CAA และ FMR เป็นเกณฑ์ในการพิจารณา

5.1 สรุปผลการวิจัย

ในการทดสอบประสิทธิภาพที่ใช้เกณฑ์การพิจารณาค่าความเชื่อมั่นแม่นตรงที่ได้จากการรวมกฎความสัมพันธ์ โดยใช้ข้อมูลโรคหอบหืดซึ่งเป็นข้อมูลจริงเป็นฐานรองรับแนวความคิดของมาตรวัดประสิทธิภาพใหม่ที่วิทยานิพนธ์ฉบับนี้เสนออันประกอบด้วย CAA ที่มาจากสมมุติฐานของความเหมือนกันระหว่างกฎความสัมพันธ์ที่มาจากอัลกอริทึมการค้นหากฎความสัมพันธ์ที่แตกต่างกัน ซึ่งจะนำกฎความสัมพันธ์ดังกล่าวมาใช้รวมกฎความสัมพันธ์จะได้กฎความสัมพันธ์เดิม ที่มีการเพิ่มมาตรวัดประสิทธิภาพใหม่ที่ถูกเรียกว่า CAA และเกณฑ์ FMR ที่มาจากแนวคิดสำหรับการรองรับมาตรวัด CAA ให้มีความน่าเชื่อถือ โดยจะทำการนับจำนวนกฎที่ถูกใช้ในการรวมกฎความสัมพันธ์ในแง่มุมของการเปรียบเทียบจะทำการเปรียบเทียบกับงานวิจัยของ Jesminและคณะ โดยใช้มาตรวัด CAA และทำการใช้ข้อมูลโรคหัวใจเป็นข้อมูลที่ใช้เปรียบเทียบซึ่งเป็นข้อมูลจากแหล่งข้อมูลมาตรฐาน UC Irvine

จากผลการทดสอบประสิทธิภาพเชิงเปรียบเทียบพบว่ามาตรวัดประสิทธิภาพใหม่ที่ได้จากวิธีที่วิทยานิพนธ์นี้เสนอ สำหรับกรณีจำแนกประเภทผู้ป่วยที่เป็นโรคหัวใจ ให้ค่าของมาตรวัดที่สูงกว่างานวิจัยของ Naharและคณะ แต่สำหรับกรณีจำแนกประเภทผู้ป่วยที่ไม่ได้เป็นโรคหัวใจ ให้ค่าของมาตรวัดที่ต่ำกว่าเล็กน้อย

5.2 ปัญหาและข้อเสนอแนะ

งานวิจัยนี้เกิดขึ้นเนื่องจากปัญหาการหากฎความสัมพันธ์แบบปกตินั้นไม่สามารถกระทำได้กับข้อมูลที่ผสมกันระหว่างข้อมูลที่เป็นข้อความและข้อมูลที่เป็นตัวเลข งานวิจัยนี้จึงใช้การแบ่งช่วงข้อมูลที่เป็นตัวเลขเข้ามาช่วยในการแปลงข้อมูลที่เป็นลักษณะตัวเลขให้เป็นข้อความ แต่กระนั้นข้อมูลทางการแพทย์ส่วนใหญ่นั้นจะพบข้อมูลบางส่วนที่ไม่สามารถระบุได้ชัดเจน เมื่อนำข้อมูลที่ไม่สามารถระบุได้ชัดเจนมาทำการหากฎความสัมพันธ์ จะได้กฎความสัมพันธ์ที่ผิดเพี้ยน งานวิจัยนี้จึงใช้การลบข้อมูลที่ไม่สามารถบ่งชี้ได้ชัดเจน จากนั้นจะทำการหากฎความสัมพันธ์ด้วยวิธีการหากฎความสัมพันธ์ทั้ง 3 แบบ อันได้แก่ วิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออรีวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมพรีดิคทีพเอไพรออรี และวิธีการหากฎความสัมพันธ์ด้วยอัลกอริทึมเทอเชียส ปัญหาที่พบคือต้องใช้ความละเอียดรอบคอบอย่างมากในการค้นหากฎความสัมพันธ์ที่มีลักษณะเหมือนกันที่จะสามารถนำมารวมกฎความสัมพันธ์กันได้ จึงมีความซับซ้อนในกระบวนการ และปัญหาทางด้านการประมวลผลที่จะช้ากว่าวิธีการค้นหากฎความสัมพันธ์กว่าปกติเล็กน้อย ดังนั้นแนวทางในการทำงานวิจัยต่อไปผู้วิจัยจะพัฒนาในส่วนอัลกอริทึมให้สามารถทำงานในการรวมกฎความสัมพันธ์แบบอัตโนมัติกับข้อมูลทางการแพทย์และ

ข้อเสนอแนะเพิ่มเติมสำหรับงานวิจัยต่าง ๆ ที่เกี่ยวข้องกับข้อมูลทางการแพทย์ในการค้นหากฎความสัมพันธ์ รวมถึงงานของวิทยานิพนธ์ฉบับนี้มีความจำเป็นที่จะต้องมีผู้เชี่ยวญชาญทางด้านการแพทย์เพื่อทำการตรวจสอบกฎความสัมพันธ์ที่ได้อีกครั้งก่อนการนำไปใช้งานจริงทางการแพทย์

รายการอ้างอิง

Agrawal, R., and Srikant, R. (1994). Fast algorithms for mining association rules. **In Proceedings of the 20th international conference on very large data bases**, Santiago, Chile. (487–499).

Cook, W.D., and Kress, M. (1990). A data envelopment model for aggregating preference rankings. **Management Science** 36: 1302–1310.

David, L.,Siau-Cheng, K., and Limsoon, W. (2009). Non-redundant sequential rules - Theory and algorithm. **Information Systems** 34:438-453.

Divya, B., and Lekha, B. (2013). Execution of APRIORI Algorithm of Data Mining Directed Towards Tumultuous Crimes Concerning Women. **International Journal of Advanced Research in Computer Science and Software Engineering** 3: 256-262.

Huawen, L., Lei, L., and Huijie, Z. (2011). A fast pruning redundant rule method using Galois connection. **Applied Soft Computing** 11: 130-137.

Jesmin, N., Tasadduq, I., Tickle, K., and Yi-Ping, C. (2013). Association rule mining to detect factors which contribute to heart disease in males and females. **Expert System with Applications** 40: 1086-1093.

Lobo, L.M.R.J., and Sunita, A. (2012). A Comparative Study of Association Rule Algorithms for Course Recommender System in E-learning. **International Journal of ComputerApplications**. 39: 231-237.

Meghana, N.,Praful, S., and Shivaji, M. (2013). Association Rule Mining Algorithms for Brain Tumour Detection. **International Journal of Advanced Research in Computer and Communication Engineering**2: 405-412.

Mehdi, T., Babak, S., and Soroosh, N. (2009). A new method for ranking discovered rules from data mining by DEA. **Expert System with Applications** 36: 8503-8508.

Mohammed, Z., and Wagner, M., (2014). **Data Mining and Analysis: Fundamental Conceptsand Algorithms.**Cambridge University Press, May 2014.

รายการอ้างอิง(ต่อ)

Mu, C. (2007) Ranking discovered rules from data mining with multiple criteria by data envelopment analysis. **Expert Systems with Applications** 33: 1110–1116.

Peter, F., and Nicolas, L. (2001). Confirmation-guided discovery of first-order rules with tertius.Springer. **Machine Learning**, 42: 61–95.

Peter, F., Valentina, M., and Fabrizio, R. (2006). Algorithms for efficiently and effectively using background knowledge in Tertius. **Department of computer science**, University of Bristol, Udine, Italy. (522–535).

Yan, P.,Haixia, L.,Hongrui, Q., and Yong, T. (2011). Transductive learning to rank using association rules. **Expert System with Applications** 38: 12839-1284

Rahman, G., andZahidul, I. (2013). Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques. **Knowledge-Based Systems** 53: 51-65.

Ramaraj, E., and Remeshkumar, K. (2009). Ranking mined Association Rule: A new measure.**Journal of Technology and Engineering Sciences** 1:569-574.

Scheffer, T. (2001). Finding association rules that trade support optimally against confidence.**Principles of Data Mining and Knowledge DiscoveryLecture Notes in ComputerScience** 2168: 424-435.

UCI. (2010). **Cleveland Heart disease data details**.<http://archive.ics.uci.edu/ml/machine learning-databases/heart-disease/heart-disease.names> Accessed 8.02.10.

ภาคผนวก ก

บทความวิจัยที่ได้รับการตีพิมพ์เผยแพร่ในระหว่างศึกษา

รายชื่อบทความวิจัยที่ได้รับการตีพิมพ์เผยแพร่ในระหว่างศึกษา

NuntawutKaoungku, TippayaThinsungnoen, PongsakornDurongdumrongchai,KittisakKerdprasop, NittayaKerdprasop. 2015. **Discretization Based on Chi2 Algorithm and Visualize Technique for Association Rule Mining.** Proceedings of the 3rd International Conference on Industrial Application Engineering 28 - 31 March 2015.

TippayaThinsungnoen, NuntawutKaoungku, PongsakornDurongdumronchai, KittisakKerdprasop, NittayaKerdprasop. 2015. **The Clustering Validity with Silhouette and Sum of Squared Errors.** Proceedings of the 3rd International Conference on Industrial Application Engineering 28 - 31 March 2015.

PongsakornDurongdumrongchai, NuntawutKaoungku,KittisakKerdprasop, NittayaKerdprasop. 2015.**IMPROVING MEDICAL DIAGNOSTIC MODEL WITH FILTRATION AND DISCRETIZATION TECHNIQUES.** 9th SOUTH EAST ASIAN TECHNICAL UNIVERSITY CONSORTIUM (SEATUC) SYMPOSIUM 29 July 2015.

Proceedings of the 3rd International Conference on Industrial Application Engineering 2015

Discretization Based on Chi2 Algorithm and Visualize Technique for Association Rule Mining

Nuntawut Kaoungku*, Tippaya Thinsungnoen, Pongsakorn Durongdumrongchai,
Kittisak Kerdprasop, Nittaya Kerdprasop
School of Computer Engineering, Institute of Engineering, Suranaree University of Technology, Thailand.

*Corresponding Author: b5111299@gmail.com

Abstract

This research aims at studying the discretization based on Chi2 algorithm and visualize technique for association rule mining. Numeric attributes with large distinct values normally do not appear in the association rules. We thus study the discretization method for numeric attributes with the integrated Chi2 algorithm and visualize technique to handle numeric attributes prior to the association analysis phase. We comparatively experiment with our proposed method against existing techniques. The comparative metrics are accuracy and number of rules.

Keywords: Discretization, Association Rule Mining, Chi2, Visualize.

1. Introduction

Currently, many organizations and merchants do not use paper to record data, but they use computers for recording. When the data are in the digital form, we can use computer to analyze these data in order to obtain the model for future use such as to predict patient's disease, to understand customer purchase behavior, or to assess learning behavior of student. The automatic induction of model from electronic data is known as data mining [1].

Association rule mining is one well-known technique in data mining. It is the induction of relationships of events or objects and generate these relationship as association rules to understand the current data or to predict the occurrence of an event or object in the future. There are many researches about the increase in efficiency in association rule mining, such as increase the speed or the accuracy of the association rule mining.

The data currently available are in a variety of types, such as numeric, text or character. But, the result of relationship induction from the numerical data in association rule mining was not good enough because the numerical data have a wide range of values. Thus, the solution of numerical data handling for association rule mining is discretization technique. There exist have many techniques for discretizing numerical data[2], such as Chi2 algorithm and Extend-Chi2 algorithm[3].

This research aims at proposing the efficient discretization technique based on Chi2 algorithm and visualized cut-point analysis for association rule mining to handle numerical data. The problem caused by the numerical data in association rule mining is that they make association rules disappear due to the sparseness of each numeric value. We, therefore, propose an efficient algorithm to solve this problem.

2. Related Work

Related researches can be divided into several types: research for proposing the new idea, research to improve the original algorithm, research to discretize numerical data for classification, and research to discretize numerical data for association rule mining. Thus, we study related work along these research themes with the details as follows:

Gyenesei[4] has proposed an algorithm to discretize numerical data by using fuzzy sets technique to reduce runtime and to increase number of effective rules in the association rule mining. The experimental result has been compared between discretization by non-normal distributed data and normal distributed data. The result of the discretization by normal distributed data gives more effective association rules than non-normal distributed data,

DOI: 10.12792/iciae2015.047

which is measured by the minimum support, minimum confidence and runtime in association rule mining.

Tong et al.[5] has proposed a method for association rule mining with numerical data using k-means clustering algorithm and Euclidean-based distance calculation. They used synthetic data to test the performance of the algorithm. Their algorithm results in less number of association rules, but higher in the values of support and confidence compared to association rule mining without any handling method for numerical data.

Ke et al.[6] has proposed a method of the association rule mining with numerical data using mutual information and clique (MIC). This technique has divided the work into 3 parts. First, applying discretization to numerical data. Second, using the data obtained from the first step to create the MI graph. And finally, the result of the second step was used in the finding of frequent itemsets. The experimental result used 6 datasets and compared the runtime, number of rules and other measures.

Wei[7] has proposed discretization technique to handle numerical data for association rule mining with clustering and genetic algorithm. His proposed technique is multivariate discretization based on density-based clustering and genetic algorithm (MVD-CG), which is an algorithm that improves the multivariate discretization algorithm (MVD). The experiment used real data and compared between MVD-CG algorithm and MVD algorithm. Measurement metrics are number of rules and effective of rules. The result is that MVD-CG algorithm has higher confident than MVD algorithm.

Sug[8] has proposed multi-dimensional association rule mining, which is different from original association rule mining in term of the difference of column. This method can reduce the data size and runtime in association rule mining. The experiment used real data from UCI. The result is that the algorithm can create small multi-dimensional table and reduce the number of rules.

From the related work, it can be seen that there exist many techniques for discretization. However, most researches do not take into consideration the distribution of the data in each rang of the divided value. We notice that it may be possible that some of the data that was in the range with very small amount might be unnecessary to be used in the association rule mining. We thus propose the visualize technique to detect ranges of values with limited amount of data in order to consider a cut point during discretization process.

3. Background

This research aims at studying the discretization and proposing a new method based on Chi2 algorithm and visualize technique for association rule mining. The related theories are divided into 4 parts, that is, association rule mining, discretization algorithms, the cut points in Chi2 algorithm, and visualization by cut points.

3.1 Association Rule Mining

Association rule mining is a popular analysis technique to automatically find the relationships between the data. There are many methods for association rule mining. In this papers we use Apriori[9] algorithm for association rule mining. In the table 1 is a list of customer purchases that will be used as an example to explain the association rule mining process. The data are counted to find the frequent customer purchases on each itemset, and then take the frequent itemsets to generate the association rules, which is a rule in the form of "If condition Then result". The measurement metrics used in the selection of frequent itemsets and rules are the following:

- Support is the frequency of the occurring event. Give the items A and B, the computation for support of A and B to be purchased is as follows:

$$Support(A \rightarrow B) = P(A \wedge B) \quad (1)$$

As an example, support (Coca cola → Bread) = 2/5 = 0.4 or 40%.

- Confidence is the frequency of the incident with other events occurring together. The computation for confident is as follows:

$$Confidence(A \rightarrow B) = \frac{Support(A \rightarrow B)}{Support(A)} \quad (2)$$

For the same example as above, confidence (Coca cola → Bread) = 0.4/0.4 = 1.0 or 100%.

Table 1. Purchase transactions of customers.

Order	Coca cola	Bread	Candy	Milk
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

Table 2. Contingency table for cut point computing demonstration.

	Class 1	Class 2	Sum
Interval 1	A_{11}	A_{12}	R_1
Interval 2	A_{21}	A_{22}	R_2
Sum	C_1	C_2	N

Sample	K=1	K=2	
2	0	1	1
3	1	0	1
total	1	1	2

$E_{11} = (1/2)*1 = .05$
$E_{12} = (1/2)*1 = .05$
$E_{21} = (1/2)*1 = .05$
$E_{22} = (1/2)*1 = .05$

$X^2 = (0-.5)^2/.5 + (1-.5)^2/.5 + (1-.5)^2/.5 + (0-.5)^2/.5 = \mathbf{2}$

Sample	K=1	K=2	
3	1	0	1
4	1	0	1
total	2	0	2

$E_{11} = (1/2)*2 = 1$
$E_{12} = (0/2)*2 = 0$
$E_{21} = (1/2)*2 = 1$
$E_{22} = (0/2)*2 = 0$

$X^2 = (1-1)^2/1 + (0-0)^2/0 + (1-1)^2/1 + (0-0)^2/0 = \mathbf{0}$

Fig. 1. Example of calculation the Chi2 cut point between intervals.

3.2 Discretization algorithms

(a) Chi2 algorithm

Chi2 algorithm[10] that is based on the X^2 statistics was used to perform discretization over the numerical data. The computation for x^2 is as follows:

$$X^2 = \sum_{i=1}^{x}\sum_{j=1}^{k}\frac{(A_{ij} - E_{ij})^2}{E_{ij}} \quad (3)$$

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 jth class $= \sum_{i=1}^{2} A_{ij}$,

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

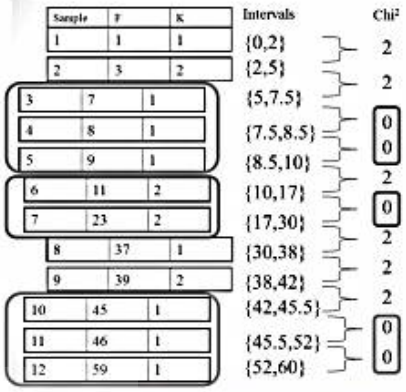


Fig. 2. Example of interval integration by Chi2 considering value.

(b) CAIM algorithm

Class-Attribute Interdependence Maximization (CAIM)[11] is a discretization algorithm by supervised learning. Main idea of the CAIM algorithm is to use class-attributed interdependence of the class and the numeric column to create minimal interval. The computation for CIAM is as follows:

$$CAIM(C, D \mid F) = \frac{\sum_{r=1}^{n}\frac{\max_r^2}{M_{+r}}}{n} \quad (4)$$

where:

C = class,

D = discretization,

F = columns,

n = number of intervals,

$\max_r$ = maximum of the q_{ir},

M_{+r} = number of all continuous columns

(c) CACC algorithm

Class-Attribute Contingency Coefficient (CACC)[12] is a discretization algorithm developed from CAIM algorithm for solving the overfitting problem. The computation for CACC is as follows:

$$CACC = \sqrt{\frac{y'}{y'+M'}} \quad (5)$$

where:

M' = number of sampling data

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

when

M = number of sampling data,

n = number of data,

q_{ir} = number of class i by sampling data (i=1,2,..,S and r=1,2,..,n) in the interval $(d_{r-1},d_r]$,

M_{i*} = number of class i by sampling data

(d) AMEVA algorithm

AMEVA algorithm[13] is discretization algorithm by supervised learning. This algorithm increase performance in terms of correlations between parameter and decrease number of intervals of the Chi2 algorithm. The computation for AMEVA is as follows:

$$Ameva(k) = \frac{x^2(k)}{k(l-1)} \quad (6)$$

where:

k = number of intervals,

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

3.3 Cut point in Chi2 Algorithm

Discretization by Chi2 algorithm is to find the cut points to divide the numerical data to interval data. The algorithm is based on the bottom-up division of numerical data in each row into intervals, and then gradually merging each interval based on independence, which can be computed by Chi2 value. We can be demonstrate the cut point calculation from contingency table in table 2 and equation (3). Figure 1 shows an example of calculating the Chi2 in each interval, such as intervals 2, 3 and 4 has Chi2 0. Figure 2 shows an example of integration the intervals by considering minimal Chi2 value, because the minimal Chi2 value means less independent between intervals. If the intervals are less independent, they should be in the same interval. For instance, the intervals 3, 4 and 5 have the Chi2 values 0. They should be in the same interval. Repeat the interval merging until Chi2 values of all intervals are greater than threshold that the user has defined.

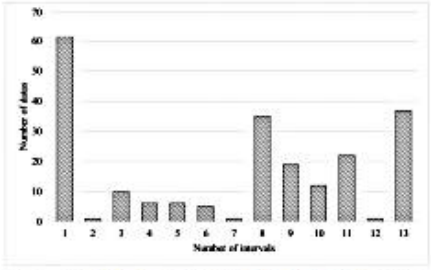


Fig. 3. Example of comparing the number of rules in each interval.

3.4 Visualization Cut Point Consideration Through

Discretization numerical data by various algorithms has to consider the cut points, which can be used for grouping the numerical data to intervals. But in some situation the data in the discretized interval has too small amount of data. This can lead to an inefficient association rule mining. Thus, we propose to use visualization technique[14] in the post-processing of the discretization steps to see the distribution of data in each interval, and then adjust the cut points to fit the data distribution in each interval. Figure 3 shows an example chart comparing the number of data in each interval. It can be seen that the data in the intervals 1, 7 and 12 are minimal comparing to other intervals. The cut points should be adjusted to re-distribute data in the discretized intervals. Figure 4 example of interval integration by visualize technique.

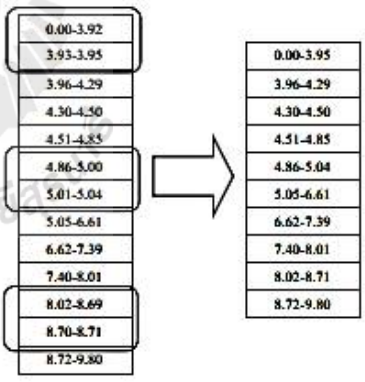


Fig. 4. Example of interval integration by visualize technique

Table 3. Comparative results of accuracy and number of rules by five discretization algorithms.

Datasets / Algorithms	Simulated Data		ECOLI		APD		CLE	
	#Rules	Acc.	#Rules	Acc.	#Rules	Acc.	#Rules	Acc.
AMEVA	81	88.19	107	97.57	987	95.29	111	45.31
CAIM	79	88.39	52	97.47	971	94.87	43	83.13
CACC	81	88.19	107	97.57	987	95.21	75	65.99
Chi2	41	88.17	47	97.67	40	81.85	239	81.37
Chi2+Visualize	37	95.41	47	97.68	36	87.08	220	81.58

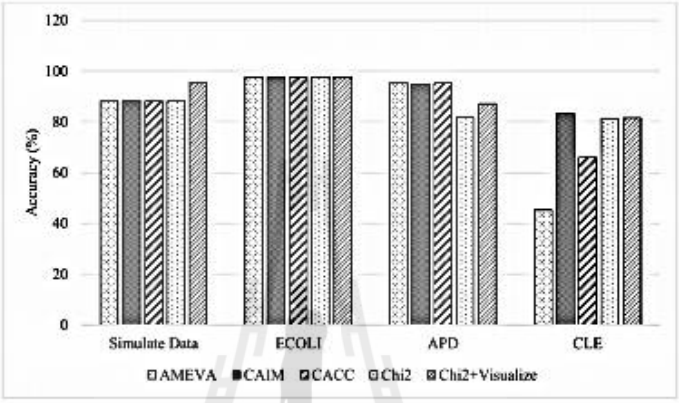


Fig. 5. The accuracy comparison of the five discretization algorithms.

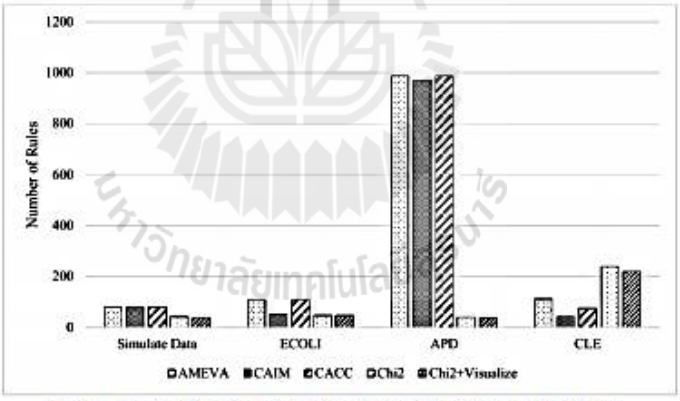


Fig. 6. The number of rules comparison of the five discretization algorithms.

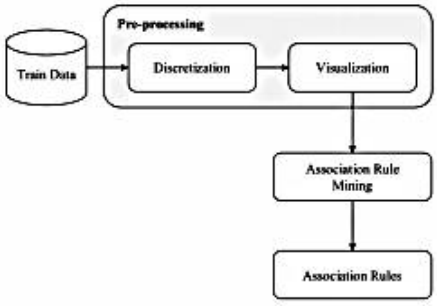


Fig. 7. Conceptual framework of the research.

4. Proposed Discretization Method

This research has proposed a methodology to perform discretization based on the Chi2 algorithm and the visualize technique for association rule mining. Figure 7 sketches the proposed method to discretize numerical attributed for association rule mining. The pre-processing module can be divided into two parts: discretization step and visualization by cut point consideration step. Discretization in our method is based on the Chi2 algorithm because it is easy to understand. After the discretization step to visualization has been applied to see distribution of data in each interval. If amount of data in any interval is too small, that interval will be merged with the previous interval or the next interval, which is considered by the distance to the nearest interval. Finally, the result is data discretized by new cut points. These data will be further used in the association rule mining

5. Experimental Setting and Results

The proposed discretization method has been experimented with both synthetic data and real data from the UCI Machine Learning Repository. The UCI data are Appendicitis data (APD) with 106 records and 7 attributes. Ecoli data with 336 records and 8 attributes. Cleveland data (CLE) with 303 records and 13 attributes. Each of these datasets hasbeen divided into training dataset (70%) and test dataset (30%). The discretization algorithm is encoded with the R language[15]. The algorithm proposed in this research is called Chi2+Visualize discretization algorithm. Its performance has been compared with the AMEVA, CAIM, CACC, and Chi2 algorithms. The performance metrics are number of rules and accuracy of the algorithms.

Table 3 shows the number of rules and accuracy of algorithms with different datasets. It can be seen that the Chi2+Visualize algorithm with the synthetic data and Ecoli data shows higher accuracy with less number of rules when compared to other algorithms. But the performance of our algorithm on the CLE and APD data shows lower accuracy than some algorithms.

Figure 5 shows a chart comparing the accuracy of algorithms with different datasets. It can be seen that the Chi2+Visualize with synthetic data and Ecoli data show higher accuracy when compared to other algorithms. Figure 6 shows a chart comparing the number of rules with different datasets. It can be seen that the Chi2+Visualize with synthetic, Ecoli and APD data has less number of rules when compared with other algorithms.

6. Conclusions

This research aims at studying the discretization method based on Chi2 algorithm and visualize technique for association rule mining. The problem of association rule mining with numerical data is that there will be large number of rules and the obtained association rules are not effective enough to predict the future data. Thus, we propose to use the cut point from discretization by Chi2 algorithm to see the distributed data in each interval, and then adjust the cut points to fit the distribution in each interval. The experimental results reveal that the proposed algorithm can reduce the number of rules and increase accuracy in predicting the future data. However, the application of our method over some data show low accuracy, but can be traded-off by small number of rules. But in some data our method shows low accuracy and large number of rules. We hypothesize that this dataset may be non-normal distribution and this kind of distribution has strong effect to our method. However, this hypothesis needs theoretical and experimental proofs further.

References

(1) Berry, Michael J., and Gordon Linoff. : "Data mining techniques: for marketing, sales, and customer support.", John Wiley & Sons, Inc., 1997.

(2) Liu, Huan, et al. : "Discretization: An enabling technique.", Data mining and knowledge discovery, vol. 6, No. 4, pp. 393-423, 2002

(3) Su, Chao-Ton, and Jyh-Hwa Hsu. : "An extended chi2 algorithm for discretization of real value attributes.", Knowledge and Data Engineering, IEEE Transactions on, vol. 17, No. 3, pp. 437-441, 2005

(4) Gyenesei, Attila : "A Fuzzy Approach for Mining Quantitative Association Rules.", Acta Cybern, Vol. 15, No. 2, pp. 305-320, 2001

(5) Tong, Qiang, et al. : "A method for mining quantitative association rules.", Jisuanji Gongcheng/ Computer Engineering, vol. 33, No. 10, pp. 34-35, 2007

(6) Ke, Yiping, James Cheng, and Wilfred Ng. : "MIC framework: an information-theoretic approach to quantitative association rule mining.", Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on. IEEE, pp. 112-112, 2006.

(7) Wei, Hantian. : "A novel multivariate discretization method for mining association rules.", Information Processing, 2009. APCIP 2009. Asia-Pacific Conference on, Vol. 1, pp. 378-381, 2009

(8) Sug, Hyontai. : "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, 2011

(9) Agrawal, Rakesh, Tomasz Imieliński, and Arun Swami : "Mining association rules between sets of items in large databases.", ACM SIGMOD Record, Vol. 22, No. 2, ACM, 1993.

(10) Liu, Huan, and Rudy Setiono. : "Chi2: Feature selection and discretization of numeric attributes.", 2012 IEEE 24th International Conference on Tools with Artificial Intelligence. IEEE Computer Society, pp. 388-388, 1995

(11) Kurgan, Lukasz A., and Krzysztof J. Cios. : "CAIM discretization algorithm.", Knowledge and Data Engineering, IEEE Transactions on, vol. 16, No. 2, pp. 145-153, 2004

(12) Tsai, Cheng-Jung, Chien-I. Lee, and Wei-Pang Yang. : "A discretization algorithm based on class-attribute contingency coefficient.", Information Sciences, vol. 178, No. 3, pp. 714-731, 2008

(13) Gonzalez-Abril, L., et al. : "Ameva: An autonomous discretization algorithm.", Expert Systems with Applications, Vol. 36, No. 3, pp. 5327-5332, 2009

(14) Keim, Daniel A. : "Information visualization and visual data mining.", Visualization and Computer Graphics, IEEE Transactions on, vol. 8, No. 1, pp. 1-8, 2002

(15) Ihaka, Ross, and Robert Gentleman : "R: a language for data analysis and graphics.", Journal of computational and graphical statistics, Vol. 5, No. 3, pp. 299-314, 1996

Proceedings of the 3rd International Conference on Industrial Application Engineering 2015

The Clustering Validity with Silhouette and Sum of Squared Errors

Tippaya Thinsungnoen[a*], Nuntawut Kaoungku[b], Pongsakorn Durongdumronchai[b],
Kittisak Kerdprasop[b], Nittaya Kerdprasop[b]
[a]Informatics Program, Faculty of Science and Technology, Nakhon Ratchasima Rajabhat University, Thailand
[b]Data Engineering Research Unit, School of Computer Engineering, Institute of Engineering,
Suranaree University of Technology, Thailand
*Corresponding Author: tippayasot@hotmail.com

Abstract

The data clustering with automatic program such as k-means has been a popular technique widely used in many general applications. Two interesting sub-activity of clustering process are studied in this paper, selection the number of clusters and analysis the result of data clustering. This research aims at studying the clustering validation to find appropriate number of clusters for k-means method. The characteristics of experimental data have 3 shapes and each shape have 4 datasets (100 items), which diffusion is achieved by applying a Gaussian distributed (normal distribution). This research used two techniques for clustering validation: Silhouette and Sum of Squared Errors (SSE). The research shows comparative results on data clustering configuration k from 2 to 10. The results of both Silhouette and SSE are consistent in the sense that Silhouette and SSE present appropriate number of clusters at the same k-value (Silhouette value: maximum average, SSE-value: knee point).

Keywords: Clustering Validity, Silhouette Measure, Sum of Squared Errors, k-means Algorithm.

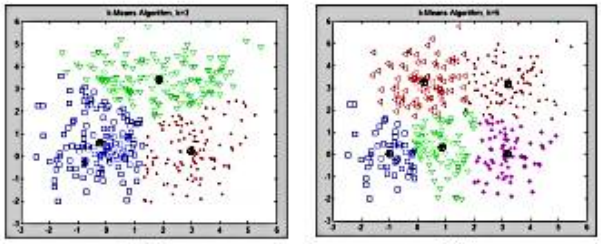
1. Introduction

A clustering is to group data. Although the clustering is similar to the data classification in terms of data input, the clustering is learning without target class. The clustering algorithm forms groups based on object similarities[1]. The clustering was applied to many fields such as bioinformatics, genetics, image processing, speech recognition, market research, document classification, and weather classification[2]. In addition, the clustering was applied to document data analysis that was one of big data learning[3-7].

There are various algorithms for the data clustering. But the most popular one is k-means algorithm. The k-means algorithm is very simple in operation and suitable for unraveling compact clusters and a fast iterative algorithm[8]. The principle of k-means algorithm has divide n objects from dataset for k clusters that used center-based clustering methods[2]. In addition, each cluster has represented by the means of objects[9]. Although k-means is a popular technique, k-means is not known the correct number of clusters a priori. Consequently, the main challenge for these clustering methods is in determining the number of clusters[2]. In general, the number of clusters has been set by users or archives from knowledge of research[1-9-11].

Fig. 1 shows the distribution of each cluster when k=3, and k=4. The researcher found that the determination of suitable k value is not clear as shown in fig. 1a and fig. 1b. As mentioned above about problem of clustering, there are various research for selecting an appropriate number of clusters[10-13]. Each of the proposed technique is suitable for each of data distribution such as Gaussianity and non-Gaussianity[14]. Therefore, finding the correct k-value for clustering is still a fundamental problem of clustering methods[15-16].

In this research, we study the clustering validity techniques to quantify the appropriate number of clusters for k-means algorithm. These techniques are Silhouette and Sum of Squared Errors. The rest of this paper is organized as follows. Section 2 discusses related research. Section 3 contains a description of methodology. Section 4 presents the results of experiments. The last section contains conclusions.



(a) k=3 (b) k=5

Fig. 1. Data clustering when k=3 and k=5

2. Related Research

Rousseeuw[12] have proposed the concept of the monitoring cluster. In the research proposed for Silhouette technique, which is based on the comparison of objects tightness and separation. The silhouette can reflect the data is grouped that objects are organized into groups that match it. This is a tool to assess the validity of the clustering to be used for selecting the optimal k in the cluster.

Kwedlo[17] have proposed the concept of problem solving in order to know the number of cluster using the Sum of Squared Errors (SSE). The research for developed a new method, called DE-KM (Differential Evolution Algorithm: DE) technique is the combination of algorithms, k-means clustering by tuning in DE and sort data to see the evolution. Experimental results show that the highest k values appropriate to the clustering, and DE-KM SSE values than the other methods they tested.

Shahbaba and Beheshti[2] have proposed the concept of dealing with the problem of determining the correct number of cluster to be grouped. The method used to estimate the probability of error point average (ACE), which is the difference between the actual point and estimate point. The idea is to explore how to use the k-means clustering with ACE k-means low (MACE-Means) is used with the UCI data and the other synthetic. They found that a correct number of cluster that can be spent on items that are sturdy little overlap and time less.

Jun et al.[9] have proposed the concept of clustering the documents based on the concept of reducing the dimension of the data, combined with the clustering k-means based on clustering with support vector and silhouette measure. They have experimented with the patent, documents from UCI to analyze separately each group of documents to clustering for technology forecasting.

3. Proposed Methodology

3.1 A Framework of Data Clustering and Validation Approach

The clustering is a data mining at an unsupervised learning technique[2, 17-20]. The principle of data clustering that objects in the same cluster will have to look very similar, while objects in other similar less[17]. There are various algorithms of clustering technique, for example, Basic Sequential Algorithms Scheme (BSAS), Partitioning Around Medoids Algorithm (PAM), Fuzzy c-Means Algorithm (FCM), k-means Algorithm[8] and so on.

The main steps in the work of the clustering has 5 steps[21]. There are (a) set a number of cluster for clustering (k) and cluster feature, (b) set a function for objects similarity measurement, (c) run clustering algorithm, (d) set Visualization to display cluster and (e) clustering validity analysis, as fig. 2.

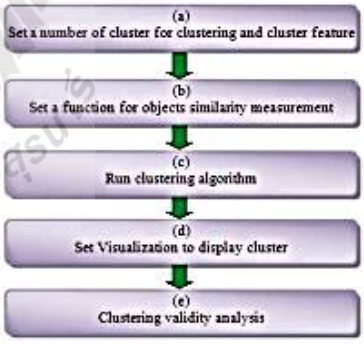


Fig. 2. Show 5 steps in clustering process.

3.2 k-Means Clustering

The k-means clustering is a technique that relies on the center of cluster. This is often represented by the average (Means) of cluster. The clustering measure the similarity of the group by iterating the measurement distance between each object and the center of each cluster[2] using Euclidean distance measuring.

The k-means algorithm is an iterative algorithm which can be described by the following steps.

Algorithm: k-means Clustering[17].

(a) Choose initial centroids $\{m_1,...,m_k\}$ of the clusters $\{C_1,...,C_k\}$.

(b) Calculate new cluster membership. A feature vector x_j is assigned to the cluster C_i if and only if

$$i = \arg\min_{k=1,...,k} ||x_j - m_k||^2 \quad (1)$$

(c) Recalculate centroids for the cluster according.

$$m_i = \frac{1}{|C_i|}\sum_{x_j \in C_i} x_j. \quad (2)$$

(d) If none of the cluster centroids have changed, finish the algorithm. Otherwise go to Step (b).

3.3 Clustering Validity Methods

3.3.1 Silhouette Measure

The concept of Rousseeuw[12] is described as follows: the Silhouette is a tool used to assess the validity of clustering. The silhouette constructed to select the optimal number of cluster with a ratio scale data (as in the case of Euclidean distances) that suitable for clearly separated cluster. The clustering are considered average proximities as the two are dissimilarities and similarities, which work best in a situation with roughly spherical clusters.

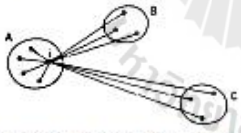


Fig. 3. Show computation *s(i)* for each object, where object *i* belong to cluster A[12].

Case #1 considered dissimilarities[12].

From fig. 3 described for take the object *i* in the data set, and assigned to cluster *A*, then define as follows:

s(i) = in case of dissimilarities.

i = object *i* belong to cluster *A*.

a(i) = average dissimilarity of *i* to all other objects of *A*.

d(i, C) = average dissimilarity of i to all objects of *C*.

b(i) = minimum *d(i, C)*, where $C \neq A$.

B = the cluster *B* for which minimum is attained the neighbor of object *i*

The cluster *B* is like the second-best choice for object *i*: if it could not be accommodated into cluster *A*, which cluster B would be the closest competitor In Fig. 3. The number *s(i)* write this in formula:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}. \quad (3)$$

The number *s(i)* is obtained by combining *a(i)* and *b(i)* as follows:

$$s(i) = \begin{cases} 1 - a(i)/b(i) & if\ a(i) < b(i), \\ 0 & if\ a(i) = b(i), \\ b(i)/a(i) - 1 & if\ a(i) > b(i), \end{cases}$$

s(i) can will be $-1 \leq s(i) \leq 1$

Case #2 considered similarities[12].

In this case consideration similarities and define *a'(i)*, *d'(i, C)*, and put *b'(i)* = maximum *d'(i, C)*, where $C \neq A$. The numbers *s(i)* is obtained by

$$s(i) = \begin{cases} 1 - b'(i)/a'(i) & if\ a'(i) > b'(i), \\ 0 & if\ a'(i) = b'(i), \\ a'(i)/b'(i) - 1 & if\ a'(i) < b'(i), \end{cases}$$

For Example, fig. 4 shows the results silhouette of clustering, when fig. 4 (a) present clustering on k = 2 and fig. 4 (b) clustering on k = 3. The Figure shows the comparison of result: density and separation, Neighbors, the average Silhouette of each cluster. Which silhouette is used to support the evaluation clustering with the maximum of silhouette.

3.3.2 Sum of Squared Errors

The k-means clustering techniques defines the target object (x_i) to each group (C_i), which relies on the Euclidean distance measurement (m_i) is the reference point to check the quality of clustering. The Sum of Squared Errors: SSE is another technique for clustering validity. SSE is defined as follows[17].

$$SSE(X, \Pi) = \sum_{i=1}^{K}\sum_{x_j \in C_i} ||x_j - m_i||^2 \quad (4)$$

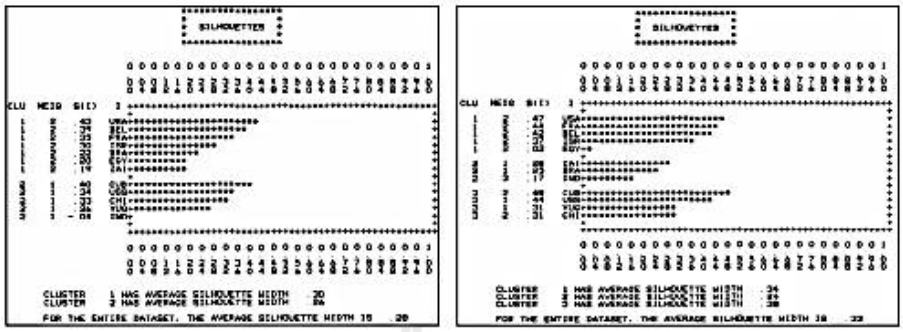
where

N = Feature vectors

$X = \{x_1, ..., x_i, ..., x_N\}$, $x_i \in \Re^M$

$\Pi = \{C_1, C_2, ... C_K\}$, $\forall_{i \neq j} C_i \cap C_j = \emptyset, \cup_{i=1}^{K} C_i = X, \forall_i C_i \neq \emptyset$.

$||\cdot||$ = Euclidean distance and m_i is centroid of cluster C_i which can computed as Eq.(2).



(a) k=2. (b) k=3.

Fig. 4. Shown Silhouette was present clustering when k=2 and k=3[12].

Conditions of applied the SSE for clustering, is to determine $k \geq 2$[22]. When the SSE is applied in graph that generated from the relationship between the SSE and k value at knee point (Significant "knee"), which is positioned to indicate the appropriate number of cluster in the k-means clustering [8] as shown in fig. 5.

3.4 Selection an Appropriate Number of Cluster

The principle of the monitoring tool for clustering, can support the selection of correct k values for the k-means clustering, consider the following.

Fig. 4, Silhouette is used to assist in cluster monitoring. This analysis is compared between Fig. 4 (a) and (b) it is found that the average silhouette of clustering when k = 3, the value 33 will be greater than k = 2, the value 28.

Fig. 5 the SSE is used in the inspection cluster. This analysis was shows the appropriate number at the knee clearly was 5(a) k = 3, 5(b) k = 4 and 5(c) k = 5, which the appropriate number of cluster.

4. Experimentation and Results

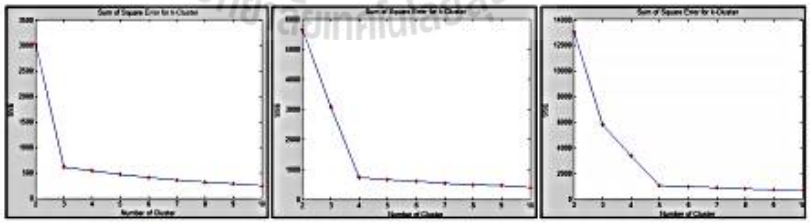
4.1 Experimental Data

The research uses data synthesized with 3 shapes and each shape have 4 datasets (100 items), which is applying a Gaussian distribution (normal distribution). Fig. 6 is the distribution of a spherical around the center of dataset, fig. 7 the distribution is non-spherical lying on the x-axis, and fig. 8 the distribution is spherical, but each group will have some overlap.

4.2 Results of k-Means Clustering Method

In experiments, the researchers repeated the k-means clustering algorithm with datasets by changing the value of k, set k = 2 to k = 10, which illustrate the specific clustering when k = 2, 4, and 6 shown in fig. 6, fig. 7 and fig. 8.

The next step is to investigate the cluster. This relies on the analysis of both Silhouette and SSE of above mentioned, are as follows.



(a) k=3 (b) k=4 (c) k=5

Fig. 5. Number of cluster consideration from the relationship between SSE and the k value.

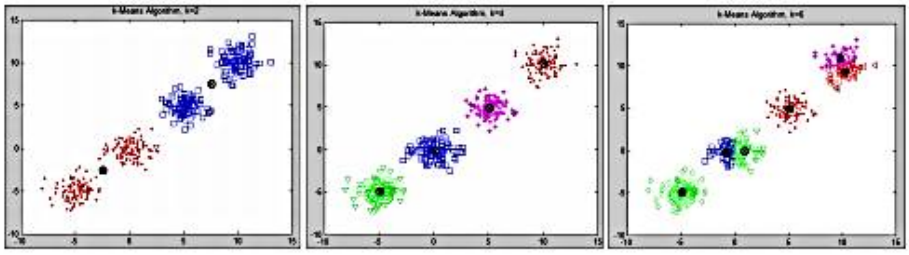


Fig. 6. Clustering with spherical data shape where k=2 (left), k=4 (middle), and k=6 (right)

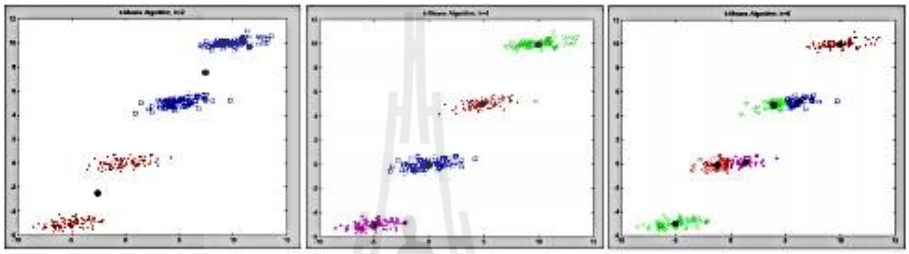


Fig. 7. Clustering with non-spherical data shape where k=2 (left), k=4 (middle), and k=6 (right)

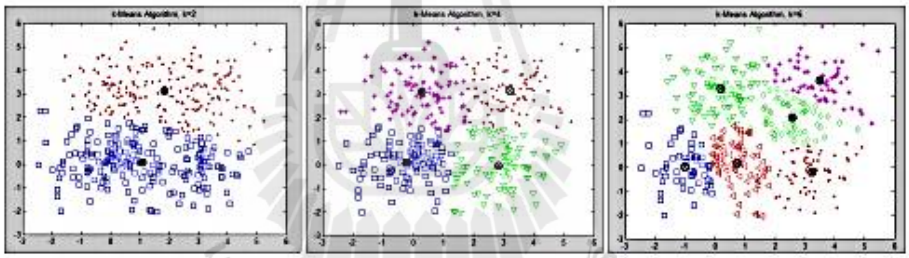


Fig. 8. Clustering with overlap data where k=2 (left), k=4 (middle), and k=6 (right)

4.3 Clustering Validity with Silhouette Measure

Consider Fig. 9 is an illustration Silhouette of a clustering technique to the k-means repeating the grouping by changing the value of k from 2 to 10, which shows a comparison of the density and separation of each cluster. Which found that the density of the k values of k = 2 and k = 4 show the density and separation is optimal.

Using the silhouette to assess the quality of clustering not silhouette diagrams only in addition need to consider the average of silhouette. It was found that the average of all silhouette values when k=4 the highest shown in table 1.

4.4 Clustering Validity with SSE

The Result of SSE for inspection the cluster is shown in Table 2. The table 2 shows the SSE value and rate of change of the SSE when k = 2 to 10, found that when k = 4 SSE is the maximum rate of change. The rate of change (%Change) defined as follows.

$$\%Change = \frac{(SSEofK_{i-1} - SSEofK_i) * 100}{SSEofK_i} \qquad (5)$$

i can will be $i \geq 2$

where

$SSEofK_{k-1}$ = SSE values of k_{k-1}

$SSEofK_k$ = SSE values of k_t

Table 1. Show comparison of the average of the Silhouette of a k-means clustering when k = 2 to 10.

Number of Cluster	Average of Silhouette		
	Spherical	Non-Spherical	Spherical & Overlap
K=2	0.8305	0.8318	0.5262
K=3	0.7715	0.7553	0.5603
K=4	**0.9117**	**0.9018**	**0.6150**
K=5	0.8182	0.8558	0.5720
K=6	0.7109	0.7982	0.5407
K=7	0.5639	0.7466	0.5177
K=8	0.6198	0.7333	0.5104
K=9	0.5072	0.6945	0.5217
K=10	0.5162	0.6850	0.5177

As the Silhouette to assess the quality of clustering not the data in table 2 that should set correct k value only. In order to investigate the effect is therefore necessary to consider a graph showing the relationship between k and the SSE values at the knee point are shown in fig. 10.

Table 2. Show SSE values and %change from k-means algorithm when k=2 to 10.

Number of Cluster	Sum of Squared Errors					
	Spherical		Non-Spherical		Spherical & Overlap	
	SSE	%Change	SSE	%Change	SSE	%Change
K=2	5,972.97	-	6,042.24	-	1505.40	-
K=3	3,179.66	87.85	3,265.32	85.04	958.94	56.99
K=4	**771.76**	**312.00**	**834.01**	**291.52**	**608.08**	**57.70**
K=5	682.02	13.16	679.50	22.74	518.62	17.25
K=6	612.04	11.43	552.51	22.98	441.48	17.47
K=7	544.41	12.42	430.75	28.27	387.16	14.03
K=8	512.90	6.14	403.10	6.86	337.27	14.79
K=9	436.02	17.63	364.19	10.68	302.21	11.60
K=10	403.84	7.97	245.55	48.32	276.06	9.47

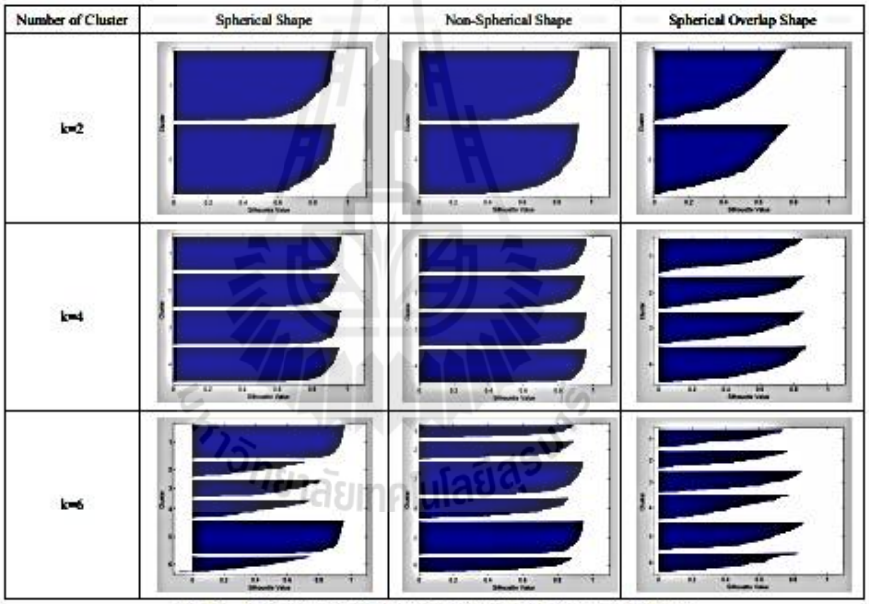
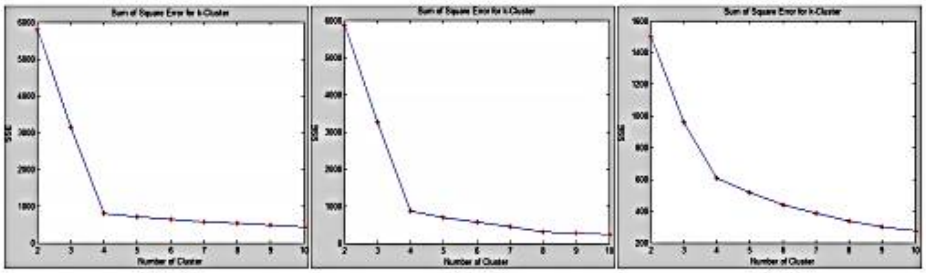


Fig. 9. Silhouette of a clustering technique when k=2, k=4, and k=6



(a) Spherical (b) Non-Spherical (c) Spherical & Overlap

Fig. 10. The graph showing the relationship between k and the SSE values of different data shapes

5. Conclusions

The results of research above when examining the Silhouette clustering analysis is to determine the k = 4 was the highest average Silhouette with all the data sets. When examining the clustering of the graph that shows the relationship between the SSE and k value with k = 4 the result was the knee point. That means the examination of both Silhouette and SSE are result inconsistent. Is that the number of cluster as the same number that k = 4.

However, a comparison of SSE and Silhouette have to attention is if the data does not overlap. Assessment the number of cluster is appropriate both SSE and Silhouette. However, when the data begin to overlap SSE will provide an assessment that is more close to the true value.

Acknowledgment

The first author has been supported by grant from the Informatics Program, Faculty of Science and Technology, Rajabaht Nakhon Ratchasrima University (NRRU). Data Engineering Research Unit has been funded by Suranaree University of Technology.

References

(1) Han, J., and Kamber, M. (2006). Data mining concepts and techniques (2nd ed.). United States of America: Morgan Kaufman Publishers.

(2) Shahbaba, M., Beheshti, S. (2014). MACE-means clustering. Singnal Processing. Vol.105, pp.216-225.

(3) Aliguliyev, R. M. (2009). Clustering of document collection–A weighting approach. Expert Systems with Applications, Vol.36, pp. 7904-7916.

(4) Isa, D., Kallimani, V. P., and Lee, L. H. (2009). Using the Self Organizing map for Clustering of Text Documents. Expert Systems with Applications, Vol.36, pp.9584–9591.

(5) Maziere, P. A. D., and Hulle, M. M. V. (2011). A clustering study of a 7000 EU document inventory using MDS and SOM. Expert Systems with Applications, Vol.38, pp. 8835–8849.

(6) Saracoglu, R., Tutuncu, K., and Allahverdi, N. (2007). A fuzzy clustering approach for finding similar documents using a novel similarity measure. Expert Systems with Applications, Vol.33, pp.600–605.

(7) Tseng, Y. H. (2010). Generic title labeling for clustered documents. Expert Systems with Applications, Vol.37, pp.2247–2254.

(8) Theodoridis, S., Pikrakis, A., Koutroumbas, K., Cavouras, D. (2010). An Introduction to Pattern Recognition : A MATLAB Approach. Academic Press, USA.

(9) Jun, S., Park, S., and Jang, D. (2014). Document clustering method using dimension reduction and support vector clustering to overcome sparseness. Expert Systems with Applications. Vol.41, pp.3204–3212.

(10) Everitt, B. S., Landau, S., and Leese, M. (2001). Cluster analysis (4th ed.). Arnold.

(11) Jun, S., and Uhm, D. (2010). Patent and statistics, What's the connection? Communications of the Korean Statistical Society, Vol.17(2), pp.205–222.

(12) Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. Vol.20, pp.53-65.

(13) Wang, L., Leckie, C., Ramamohanarao, K., and Bezdek, J. (2009). Automatically determining the number of clusters in unlabeled data sets. IEEE Transactions on Knowledge and Data Engineering, Vol.21(3), pp.335–350.

(14) McNicholas, P. D., Subedi, S. (2012). Clustering gene expression time course data using mixtures of multivariate t-distributions, J. Stat. Plan. Inference 142 (May (5)). p.1114–1127.

(15) Jain, A. K. (2010). Data clustering: 50 years beyond k-means, Pattern Recogn. Lett. Vol.(8) 31, pp.651–666, http://dx.doi.org//10.1016j.patrec..2009.09.011

(16) Aggarwal, C. C., Reddy, C. K. (2013). Data Clustering: Algorithms and Applications, Vol.31, CRC Press, Hoboken, New Jersey, p.648. (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, ISBN: 1466558210).

(17) Kwedlo, W. (2011). A clustering method combining differential evolution with the k-means algorithm. Pattern Recognition Letters. Vol.32, pp.1613–1621.

(18) Roiger, R. J., Geatz, M. W. (2003). Data Mining A Tutorial – Based Primer. Pearson Education, Inc. Addison Wesley. pp. 11-12.

(19) Jain, A., Murty, M. N., Flynn, P. J. (1999). Data clustering: a review. ACM Comput. Surv. Vol.31(3), pp.264–323.

(20) Kaufman, L., Rousseeuw, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley.

(21) Kerdprasop, K. (2006). Density Biased Sampling for Incremental Data Clustering. Research Final Report. School of Computer Engineering, Suranaree University of Technology.

(22) Aloise, D., Deshpande, A., Hansen, P., Popat, P. (2009). NP-hardness of Euclidean sum-of-squares clustering. Mach. Learn Vol.75 (2), pp.245–248.

IMPROVING MEDICAL DIAGNOSTIC MODEL WITH FILTRATION AND DISCRETIZATION TECHNIQUES

Pongsakorn Durongdumrongchai, Nuntawut Kaoungku, Kittisak Kerdprasop, Nittaya Kerdprasop
School of Computer Engineering, Suranaree University of Technology, Thailand

ABSTRACT

To apply data mining classification technique to the medical data, it is important to have a classification model that can perform effective analysis with high accuracy. There are significant amount of researches trying to find ways to improve classification model accuracy. One of the methods successfully applied is the filter technique. It is the process of preparing data before modeling. This paper presents the filter technique together with the discretization method for the important of medical diagnostic model. The proposed data filtration and discretization techniques has been tested with the six classification algorithm: support vector machine, naïve Bayes, multilayer perceptron, k-nearest neighbor, decision stump, and decision trees. The six medical data and one generated data were used for accuracy evaluation. Five medical data showed the improvement on classification accuracy

Keywords: Filters, Discretization, Classifications, Accuracy, Medical data

1. INTRODUCTION

Data mining model for classification of medical data has proved to be a success in the prediction of new data with the relatively high accuracy. Guyon & Elisseeff (2003) stated that data filtration by feature selection can help the discovery of relationships hidden in data

Feature selection is in the preprocessing of machine learning as a way to reduce the dimension by reducing the redundant and irrelevant information. This will increase the accuracy. Feature selection generally fall into two categories: filter models and wrapper model. Filter model will select some features without involving algorithm to learn any model. Wrapper model is based on the one that is preset on the selected feature. The choice of filter models studied by Guz (2011) often provide higher accuracy.

Discretization is one of the preprocessing of the machine learning. How is the nature of the data are numeric will provide a range of features that will help increase the accuracy of prediction. Many works have been action in the field of dimensionality reduction for medical diagnosis.

The following section presents the summary of those works, highlighting the strengths and weaknesses of each method. It may be observed that the Sequential Forward Feature Selection (SFFS) is impractical for feature subset selection from a large number of samples of high-dimensional features (Hall, 1998). Therefore, Hab & Yu (2010) propose the Filter-Dominating Hybrid Sequential Forward Feature Selection (FDHSFFS) algorithm for high dimensional feature subset selection. This method had been proved to be fast but wanted greatly computational complexity. A new approach called Multi Filtration Feature Selection (MFFS) had been proposed by Sasikala et al., (2014). The method used a holistic and universal method that achieved the best classification accuracy with fewest features possible.

This paper makes an attempt to design such a feature selection sequence and discretization. It is called "Multi Filtration with Feature Selection Discretization (MFFDS)".This paper is organized as follows: Section 3 describes the proposed method. Experimental results and discussions are presented in Section 4.

2. METHODOLOGY

Figure 1 shows the flow of our methodology. The feature selection technique has been applied to the data. Then the discretization technique is used. The accuracy is tested with the six classification algorithms

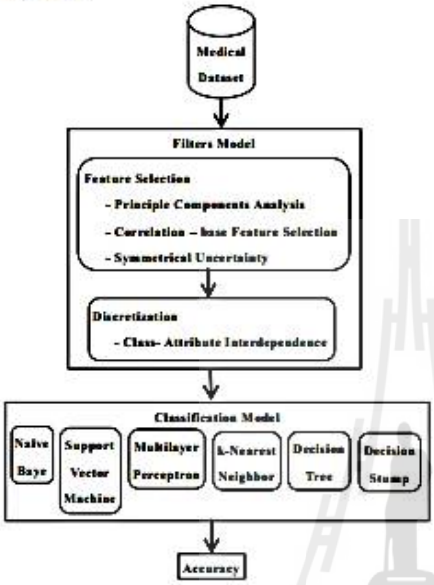


Fig. 1 System flow of the proposed MFFDS model

2.1 Principle Component Analysis (PCA)

A represented of unsupervised dimensionality reduction method is Principal Component Analysis (PCA) Moutselos et al., (1986) which aims to identifying a lower-dimensional space maximizing the variance among data Sasikala et al., (2014). PCA is a very effective approach of extracting features Vinh et al., (2011). The computation for PCA is as follows:

$$AA^{T}P = LP \qquad (1)$$

Let AA^{T} is matrix L is covariance matrix and P is eigenvector.

2.2 Correlation – base Feature Selection (CFS)

The correlation between each feature and the class and between two features can be measured and best-first search. This is realized in the Correlation-based Feature Selection (CFS) method Xie, & Wang (2011). The computation for CFS is as follows:

$$\text{Merit}_s = \frac{k\overline{\omega_{cf}}}{\sqrt{k+k(k-1)\overline{\omega_{ff}}}} \qquad (2)$$

$$\omega_{ij} = \frac{\sum(i-\bar{i})(j-\bar{j})}{\sqrt{[\sum(i-\bar{i})^2][\sum(j-\bar{j})^2]}} \qquad (3)$$

Let s is feature subset k is feature $\overline{\omega_{cf}}$ is the average value of feature-class and $\overline{\omega_{ff}}$ is average value of feature-feature.
Let i and j is correlation between entities.

2.3 Class-Attribute Interdependence maximization (CAIM)

The goal of the CAIM algorithm is to maximize the class-attribute interdependence and to generate a (possibly) minimal number of discrete intervals Yazdani et al., (2012). The computation for CAIM is as follows:

$$\text{CAIM}(C, D|F) = \frac{\sum_{r=1}^{n} \frac{\max_r^2}{M_{+r}}}{n} \qquad (4)$$

Let C is class variable and D is discretization variable for attribute F.
Let n is number of intervals r = 1, 2. . . n, $\max_r$ is maximum value among and M_{+r} is total number of continuous values.

3. EXPERIMENT

3.1 Experimental Apparatus

The tests are carried out in a system with Intel i7-4500U, CPU 1.80GHz, 8 GB RAM, 1TB hard drive on a Windows 8 Enterprise operating system.

The proposed algorithm is implemented using Weka (2013). The input to the system is given in the Attribute-Relation File Format (ARFF). Ten-fold cross validation is performed for all classifiers (Zahedi & Sorkhi 2013). Data sets used in the expertiment are listed in Table1.

Data set names, number of rows, number of features, and the number of classes.

Name	Instance	Attribute	Class
Breast Cancer Wisconsin	569	32	2
E.coli	336	8	8
Indian Liver Patient Dataset	583	11	2
Liver	345	7	2
SPECT Heart Dataset	80	45	2
Parkinson	195	23	2
Generate Dataset	570	3	2

3.2 Result

Figure 2–7 show the experimental results of the proposed method (MFFDS) compared to the MFFS methods and the classification without any data preprocessing technique. The X-axis represents accuracy and the Y-axis represents medical data.

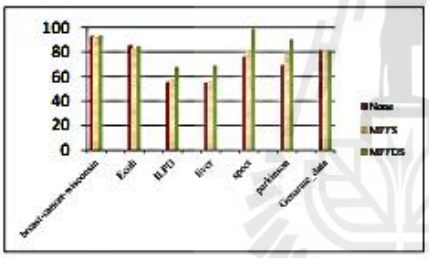


Fig. 2 Classification accuracy obtained for existing and the proposed MFFDS by naïve Bayes.

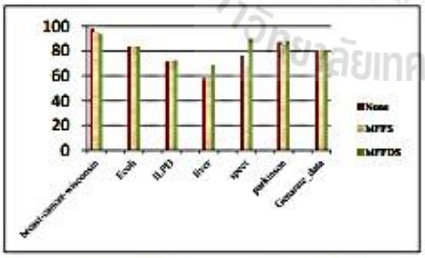


Fig. 3 Classification accuracy obtained for existing and the proposed MFFDS by support vector machine.

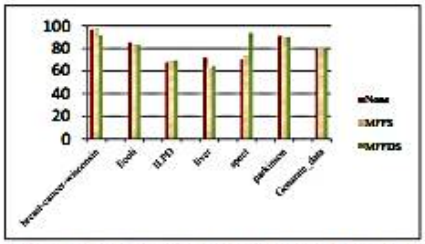


Fig. 4 Classification accuracy obtained for existing and the proposed MFFDS by multilayer perceptron.

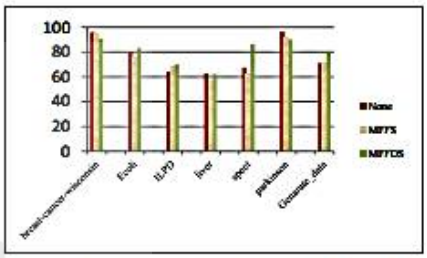


Fig. 5 Classification accuracy obtained for existing and the proposed MFFDS by k-nearest neighbor.

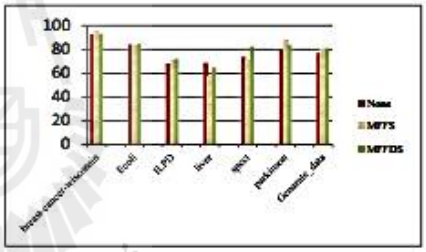


Fig. 6 Classification accuracy obtained for existing and the proposed MFFDS by decision tree.

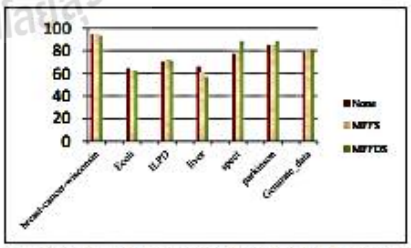


Fig. 7 Classification accuracy obtained for existing and the proposed MFFDS by decision stump.

CONCLUSION

MFFDS model performance testing with naïve Bayes, support vector machine, multilayer perceptron, k-nearest neighbor, decision tree and decision stump by using Breast Cancer Wisconsin, E.coli, Indian Liver, Patient Dataset, Liver, SPECT Heart Dataset, Parkinson and Generate data. The implication of the research is MFFDS Model for best results, performance testing is five data.

REFERENCE

Guyon, A.I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn Res. 3, 1157–1182.

Guz, H.U., 2011. A hybrid system based on information gain and principal component analysis for the classification of transcranial Doppler signals. Comput. Methods Programs Biomed. 107, 598–609.

Hall, M.A., 1998. Correlation based feature selection for machine learning (PhD Dissertation), Dept. of Comp. Science, Univ. of Waikato, Hamilton, New Zealand.

Han, Y., Yu, L., 2010. A variance reduction framework for stable feature selection. In: Webb, G.I., Liu, B., Zhang, C., Gunopulos, D., Wu, X. (Eds.), Data Mining. IEEE Computer Society, Sydney, Australia, pp. 206– 215.

Jolliffe, I.T., 1986. Principal Component Analysis. Springer-Verlag, New York, NY.

Moutselos, K., Maglogiannis, I., Chatziioannou, A., 2014. Integration of high-volume molecular and imaging data for composite biomarker discovery in the study of melanoma. Biomed. Res. Int. http://dx.doi.org/10.1155/2014/145243.

Sasikala, S. Appavu alias Balamurugan, Geetha ,S., 2014. Multi Filtration Feature Selection (MFFS) to improve discriminatory ability in clinical data set. Applied Computing and Informatics, In Press, Corrected Proof, Available online 5 April 2014

Vinh, L.T., Lee, S., Park, Y., Auriol, B.J., 2011. A novel feature selection method based on normalized mutual information. Int. J. Appl. Intell 37, 100–120

Xie, J., Wang, C., 2011. Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases. Expert Syst. Appl. 38 (5), 5809–5815.

Yazdani, S., Shanbehzadeh, J., Shalmani, Mohammad Taghi Manzuri, 2012. RPCA: a novel pre-processing method for PCA. Adv. Artif. Intell. 1, 1–7.

Zahedi, M., Sorkhi, A.G., 2013. Improving text classification performance using PCA and recall-precision. Arab J Sci Eng (2013) 38:2095–2102.

Pongsakorn Durongdumrongchai
He is currently a master student with the School of Computer Engineering, Suranaree University of Technology, Thailand. His current research of interest includes association and Classification.

Nuntawut Kaoungku
He is currently a doctoral student with the School of Computer Engineering, Suranaree University of Technology, Thailand. His current research includes semantic web and association.

Kittisak Kerdprasop
He is an associate professor and chair of the School of Computer Engineering, Suranaree University of Technology, Thailand. His current research of interest includes Data mining, Artificial Intelligence, Functional and Logic Programming Languages and Computational Statistics.

Nittaya Kerdprasop
She is an associate professor at the School of Computer Engineering, Suranaree University of Technology, Thailand. She is a member of ACM and IEEE Computer Society. Her research of interest includes Knowledge Discovery in Databases, Artificial Intelligence and Logic Programming.

ประวัติผู้เขียน

นายพงศกร ดุรงค์ดำรงชัยเกิดเมื่อวันที่ 18 มิถุนายนพ.ศ. 2535 ที่อำเภอสีคิ้ว จังหวัดนครราชสีมา เริ่มเข้าศึกษาระดับชั้นอนุบาล 1 ถึงชั้นประถมศึกษาปีที่ 6 ที่โรงเรียนบ้านมงคลกุลวิทยา อำเภอสีคิ้วจังหวัดนครราชสีมาจากนั้นได้เข้าศึกษาต่อในระดับมัธยมศึกษาตอนต้นและตอนปลายที่โรงเรียนราชสีมาวิทยาลัยอำเภอเมืองจังหวัดนครราชสีมาปีการศึกษา 2553 ได้เข้าศึกษาต่อระดับปริญญาตรีในสาขาวิชาวิศวกรรมคอมพิวเตอร์สำนักวิชาวิศวกรรมศาสตร์มหาวิทยาลัยเทคโนโลยีสุรานารีและสำเร็จการศึกษาเมื่อปีพ.ศ. 2556 ภายหลังสำเร็การศึกษาในระดับปริญญาตรีได้เข้าศึกษาในระดับปริญญาโทสาขาวิชาวิศวกรรมคอมพิวเตอร์สำนักวิชาวิศวกรรมศาสตร์มหาวิทยาลัยเทคโนโลยีสุรานารีในปี 2556ในระหว่างการศึกษาได้รับความอนุเคราะห์อย่างดียิ่งจากอาจารย์ประจำวิชา Database System ได้รับความไว้วางใจให้เป็นผู้ช่วยสอนปฏิบัติการ