รหัสโครงการ SUT7-705-60-12-29

# รายงานการวิจัย

# การสังเคราะห์แบบรูปความสัมพันธ์ครอบคลุมจากข้อมูลหลายแหล่งเพื่อสนับสนุนสนเทศศาสตร์สุขภาพแบบกระจาย

# (Synthesizing Global Association Patterns from Multiple Data Sources to Support Distributed Health Informatics)



ได้รับทุนอุดหนุนการวิจัยจาก
มหาวิทยาลัยเทคโนโลยีสุรนารี

ผลงานวิจัยเป็นความรับผิดชอบของหัวหน้าโครงการวิจัยแต่เพียงผู้เดียว

รหัสโครงการ SUT7-705-60-12-29

**รายงานการวิจัย**

# การสังเคราะห์แบบรูปความสัมพันธ์ครอบคลุมจากข้อมูลหลายแหล่งเพื่อสนับสนุนสนเทศศาสตร์สุขภาพแบบกระจาย

# (Synthesizing Global Association Patterns from Multiple Data Sources to Support Distributed Health Informatics)

**ผู้วิจัย**

| | |
|---|---|
| **รองศาสตราจารย์ ดร.นิตยา เกิดประสพ** | **หัวหน้าโครงการ** |
| **รองศาสตราจารย์ ดร.กิตติศักดิ์ เกิดประสพ** | **ผู้วิจัยร่วม** |
| **อาจารย์ ดร.นันทวุฒิ คะอังกุ** | **ผู้วิจัยร่วม** |

สาขาวิชาวิศวกรรมคอมพิวเตอร์

สำนักวิชาวิศวกรรมศาสตร์

**ได้รับทุนอุดหนุนการวิจัยจากมหาวิทยาลัยเทคโนโลยีสุรนารี ปีงบประมาณ พ.ศ. 2560**

**ผลงานวิจัยเป็นความรับผิดชอบของหัวหน้าโครงการวิจัยแต่เพียงผู้เดียว**

มกราคม 2562

# กิตติกรรมประกาศ

คณะผู้วิจัยขอขอบคุณมหาวิทยาลัยเทคโนโลยีสุรนารี และสำนักงานคณะกรรมการวิจัยแห่งชาติ ที่สนับสนุนโครงการวิจัยนี้ด้วยการจัดสรรงบประมาณให้ในปีงบประมาณ พ.ศ.2560 รวมถึงขอขอบคุณผู้ทรงคุณวุฒิทั้งภายนอกและภายในมหาวิทยาลัย ที่ได้เสียสละเวลาทำหน้าที่ตรวจข้อเสนอโครงการวิจัยและตรวจร่างรายงานการวิจัยฉบับสมบูรณ์ ข้อเสนอแนะจากผู้ทรงคุณวุฒิทุกท่านเป็นประโยชน์อย่างมากต่อคณะผู้วิจัยในการปรับปรุงการออกแบบ และขั้นตอนการดำเนินงานของโครงการวิจัย งานวิจัยนี้สำเร็จได้อย่างดีด้วยการมีส่วนร่วมจากนักศึกษาทั้งในระดับปริญญามหาบัณฑิตและปริญญาดุษฎีบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ ที่ได้ทำหน้าที่เป็นผู้ช่วยวิจัยในโครงการวิจัยนี้

# บทคัดย่อภาษาไทย

ในระบบสนเทศศาสตร์สุขภาพ ข้อมูลทุกด้านที่เกี่ยวข้องกับสุขภาพของประชาชนจะถูกบันทึกลงฐานข้อมูลเฉพาะท้องถิ่นและเก็บรักษาไว้ที่หน่วยงานสาธารณสุขแต่ละแห่ง ฐานข้อมูลแบบกระจายเหล่านี้มีความรู้ที่เป็นประโยชน์ซ่อนอยู่ ความรู้ในลักษณะของการค้นหาความสัมพันธ์ระหว่างแอททริบิวต์หรือลักษณะประจำภายในฐานข้อมูล สามารถเรียนรู้ได้ด้วยการทำเหมืองความสัมพันธ์ แต่การเรียนรู้แบบนี้มักจะมีข้อกำหนดเบื้องต้นว่าข้อมูลจะต้องถูกรวบรวมไว้เป็นไฟล์เดียว ซึ่งในระบบสนเทศศาสตร์สุขภาพที่ใช้งานอยู่จริง ฐานข้อมูลแต่ละแห่งจะมีข้อมูลในปริมาณมาก เมื่อรวมข้อมูลในทุกฐานข้อมูลเป็นไฟล์ขนาดใหญ่ไฟล์เดียวจะไม่สามารถประมวลผลได้เนื่องจากไฟล์จะมีขนาดใหญ่เกินกว่าความสามารถของหน่วยประมวลผลโดยทั่วไปจะทำงานได้

งานวิจัยนี้ได้เสนอวิธีแก้ปัญหาด้วยการหากฎความสัมพันธ์แบบกระจาย แต่การหากฎความสัมพันธ์เพื่อให้ได้แหล่งความรู้เพียงแหล่งเดียวในลักษณะนี้ทำได้ยาก เนื่องจากขั้นตอนการรวมกฎความสัมพันธ์นั้นอาจทำให้ได้กฎความสัมพันธ์ที่ขัดแย้งกันเอง หรือได้จำนวนกฎความสัมพันธ์ที่มากจนเกินไป หรือเกิดการขาดไปของกฎความสัมพันธ์ที่สำคัญ ดังนั้น งานวิจัยนี้ได้เสนอแนวทางแก้ไขปัญหาการหากฎความสัมพันธ์แบบกระจาย โดยในขั้นตอนการรวมกฎความสัมพันธ์จะนำมาเฉพาะกฎความสัมพันธ์ที่ปรากฏขึ้นบ่อยในทุก ๆ แหล่งความรู้ที่เรียนรู้ได้จากฐานข้อมูลย่อย จากนั้นนำกฎความสัมพันธ์ที่เรียนรู้ได้จากแต่ละฐานข้อมูลไปตรวจสอบความขัดแย้งและในขั้นตอนนี้สามารถสร้างกฎความสัมพันธ์ใหม่จากกฎความสัมพันธ์เดิมที่มีอยู่ด้วยวิธีการอนุมานเชิงตรรกศาสตร์ ซึ่งสามารถเติมเต็มในส่วนของกฎความสัมพันธ์ที่ขาดหายไปได้ สุดท้ายจะได้กฎความสัมพันธ์ที่มีประสิทธิภาพเพียงพอสำหรับการนำไปทำนายผลข้อมูลในอนาคตและไม่เกิดความขัดแย้งกันเอง

# บทคัดย่อภาษาอังกฤษ

In health informatics, all data related to people's health are recorded in the local database and stored at each healthcare organization. These distributed databases have valuable hidden knowledge. A specific kind of knowledge revealing relations among data attributes can be extracted by means of association mining. But the normal assumption of such knowledge mining is that all data to be learned have to be collected into a single file. However, in reality each database in health informatics stores a lot of information. Therefore, gathering all tremendous data as a single source will result in a too big file size to be practically processed by a typical processing unit.

This research proposes a distributed learning method to solve mining association knowledge from distributed databases. The difficulties of such distributed learning method are that the process of combining association rules may lead to inconsistent rules, there may be too many number of association rules, and some significant association rules are possibly missing. We thus propose the distributed association rule mining method. In the combining process, association rules that appear frequently in all knowledge bases are combined and then checked for inconsistency of rules. This process can generate new association rules from original association rule set with the inference feature of first-order logic. Finally, the efficient and consistent association rules are obtained.

# สารบัญ

หน้า

กิตติกรรมประกาศ ..... ก
บทคัดย่อภาษาไทย ..... ข
บทคัดย่อภาษาอังกฤษ ..... ค
สารบัญ ..... ง
สารบัญตาราง ..... ฉ
สารบัญภาพ ..... ช
บทที่ 1 บทนำ
1.1 ความสำคัญและที่มาของปัญหาการวิจัย ..... 1
1.2 วัตถุประสงค์ของโครงการวิจัย ..... 2
1.3 ขอบเขตของการวิจัย ..... 3
1.4 ประโยชน์ที่ได้รับ ..... 3
บทที่ 2 เอกสารและงานวิจัยที่เกี่ยวข้อง
2.1 การเรียนรู้กฎความสัมพันธ์จากข้อมูลหลายแหล่ง ..... 4
2.2 งานวิจัยที่เกี่ยวข้อง ..... 7
บทที่ 3 การออกแบบและพัฒนาวิธีการสังเคราะห์แบบรูปความสัมพันธ์จากข้อมูลหลายแหล่ง
3.1 กรอบแนวคิดของงานวิจัย ..... 10
3.2 ขั้นตอนการดำเนินงานวิจัย ..... 11
3.3 การออกแบบขั้นตอนวิธีการ ..... 14
3.3.1 วิธีการรวมกฎความสัมพันธ์ ..... 14
3.3.2 วิธีการแปลงรูปแบบกฎความสัมพันธ์ให้เป็นภาษาธรรมชาติ ..... 15
3.3.3 วิธีการตรวจสอบความขัดแย้งของกฎความสัมพันธ์ ..... 17
บทที่ 4 การทดสอบกลไกการค้นหาและรวมกฎความสัมพันธ์จากหลายแหล่ง
4.1 ข้อมูลที่ใช้ในการทดสอบ ..... 20
4.2 วิธีการทดสอบประสิทธิภาพกลไกการค้นหาและรวมกฎความสัมพันธ์ ..... 23
4.3 ผลการทดสอบกลไกการค้นหากฎความสัมพันธ์ใหม่ด้วยวิธีการเชิงตรรกะ ..... 25
4.4 ผลการทดสอบกลไกการค้นหาและรวมกฎความสัมพันธ์ด้วยค่าสนับสนุน ที่แตกต่างกัน ..... 29
4.5 ผลการทดสอบความถูกต้องของการค้นหากฎความสัมพันธ์จากหลายแหล่ง ..... 32
4.6 อภิปรายผล ..... 34

# สารบัญ(ต่อ)

หน้า

บทที่ 5 บทสรุป

5.1 สรุปผลการวิจัย ..... 36

5.2 ข้อจำกัดและข้อเสนอแนะ ..... 38

บรรณานุกรม ..... 39

ภาคผนวก ผลผลิตของงานวิจัย ..... 42

ภาคผนวก ก บทความวิจัยตีพิมพ์ในวารสารวิชาการ ..... 43

1. N. Kaoungku, K. Kerdprasop, N. Kerdprasop (2014). A technique to association rule mining on multiple datasets. *Journal of Advances in Information Technology*, vol.5, no.2, May, pp.53-57. (indexing: INSPEC, ISSN: 1798-2340)

2. N. Kaoungku, K. Kerdprasop, N. Kerdprasop (2017). A method to clustering the feature ranking on data classification using an ensemble feature selection. International *Journal of Future Computer and Communication*, vol. 6, no. 3, September 2017, pp. 81-85. (indexing: INSPEC, ISSN: 2010-3751)

3. N. Kaoungku, K. Suksut, R. Chanklan, K. Kerdprasop, N. Kerdprasop (2018). The silhouette width criterion for clustering and association mining to select image features. *International Journal of Machine Learning and Computing*, vol. 8, no. 1, pp. 69-73. (indexing: Scopus, ISSN: 2010-3700)

ภาคผนวก ข ลิขสิทธิ์โปรแกรม ..... 58

โปรแกรมรวมฎความสัมพันธ์ด้วยกลไกการให้เหตุผลเชิงตรรกะ
(Program to integrate association rules with reasoning mechanism)

ประวัติผู้วิจัย ..... 61

# สารบัญตาราง

หน้า

ตารางที่ 3.1 ข้อมูลตัวอย่างผู้ป่วยโรคมะเร็ง ..... 14
ตารางที่ 3.2 รูปแบบกฎความสัมพันธ์ที่ถูกแปลงเป็นภาษาธรรมชาติ ..... 16
ตารางที่ 3.3 ตัวอย่างของความรู้ใหม่ที่ได้จากการให้เหตุผลเชิงตรรกะจากภาษาธรรมชาติ ..... 18
ตารางที่ 4.1 รายละเอียดแอททริบิวต์ของข้อมูลผู้ป่วยโรคมะเร็งเต้านม ..... 21
ตารางที่ 4.2 รายละเอียดแอททริบิวต์ของข้อมูลผู้ป่วยโรคหัวใจ ..... 22
ตารางที่ 4.3 กฎความสัมพันธ์ที่ค้นพบได้ใหม่จากวิธีการเชิงตรรกะและใช้ค่าสนับสนุน
ขั้นต่ำ 0.1 ..... 26
ตารางที่ 4.4 กฎความสัมพันธ์ที่ค้นพบได้ใหม่จากวิธีการเชิงตรรกะและใช้ค่าสนับสนุน
ขั้นต่ำ 0.2 ..... 27
ตารางที่ 4.5 กฎความสัมพันธ์ที่ค้นพบได้ใหม่จากวิธีการเชิงตรรกะและใช้ค่าสนับสนุน
ขั้นต่ำ 0.3-0.5 ..... 28
ตารางที่ 4.6 กฎความสัมพันธ์ที่ค้นพบได้ใหม่จากวิธีการเชิงตรรกะและใช้ค่าสนับสนุน
ขั้นต่ำ 0.6 ..... 28
ตารางที่ 4.7 เปรียบเทียบจำนวนกฎความสัมพันธ์ระหว่างการหากฎความสัมพันธ์แบบดั้งเดิม
และการหากฎความสัมพันธ์จากข้อมูลหลายแหล่งในข้อมูลผู้ป่วยโรคมะเร็งเต้านม 30
ตารางที่ 4.8 เปรียบเทียบจำนวนกฎความสัมพันธ์ระหว่างการหากฎความสัมพันธ์แบบดั้งเดิม
และการหากฎความสัมพันธ์จากข้อมูลหลายแหล่งในข้อมูลผู้ป่วยโรคหัวใจ ..... 31
ตารางที่ 4.9 เปรียบเทียบความถูกต้องระหว่างกฎความสัมพันธ์แบบดั้งเดิมและ
กฎความสัมพันธ์จากข้อมูลหลายแหล่งด้วยค่าสนับสนุน 0.4 ในข้อมูลผู้ป่วย
โรคมะเร็งเต้านม ..... 32
ตารางที่ 4.10 เปรียบเทียบความถูกต้องระหว่างกฎความสัมพันธ์แบบดั้งเดิมและ
กฎความสัมพันธ์จากข้อมูลหลายแหล่งด้วยค่าสนับสนุน 0.4 ในข้อมูลผู้ป่วย
โรคมะเร็งหัวใจ ..... 34

# สารบัญภาพ

หน้า

รูปที่ 2.1 กฎความสัมพันธ์ที่ได้จากข้อมูล CRYPT ..... 5
รูปที่ 2.2 กฎความสัมพันธ์ที่ได้จากข้อมูล EUROCRYPT ..... 5
รูปที่ 2.3 กฎความสัมพันธ์ที่ได้จากข้อมูล COMPGEOM ..... 5
รูปที่ 2.4 กฎความสัมพันธ์ที่ได้จากเรียนรู้แบบกระจายจากสองกลุ่มข้อมูล CRYPT- EUROCRYPT และข้อมูล COMPGEOM ..... 6
รูปที่ 3.1 กรอบการวิจัยการสังเคราะห์แบบรูปความสัมพันธ์ครอบคลุมจากข้อมูลหลายแหล่งเพื่อสนับสนุนสนเทศศาสตร์สุขภาพแบบกระจาย ..... 11
รูปที่ 3.2 ขั้นตอนการพัฒนากลไกการค้นหาและรวมกฎความสัมพันธ์ที่เป็นแบบรูปเฉพาะที่ .... 13
รูปที่ 3.3 ผลลัพธ์ของการรวมกฎความสัมพันธ์ด้วยการทำอินเตอร์เซกชัน ..... 15
รูปที่ 3.4 ขั้นตอนการแปลงรูปแบบกฎความสัมพันธ์ทั่วไปให้อยู่ในรูปแบบภาษาธรรมชาติ ..... 17
รูปที่ 3.5 ตัวอย่างออนโทโลยีที่สร้างจากกฎความสัมพันธ์ ..... 18
รูปที่ 3.6 ขั้นตอนการตรวจสอบความขัดแย้งของกฎความสัมพันธ์ที่รวบรวมจากหลายแหล่ง .... 19
รูปที่ 4.1 วิธีการทดสอบประสิทธิภาพกลไกการค้นหาและรวมกฎความสัมพันธ์จากหลายแหล่ง 24
รูปที่ 4.2 กราฟปรียบเทียบจำนวนกฎความสัมพันธ์ระหว่างการหากฎความสัมพันธ์แบบดั้งเดิมและการหากฎความสัมพันธ์จากข้อมูลหลายแหล่งในข้อมูลผู้ป่วยโรคมะเร็งเต้านม .... 30
รูปที่ 4.3 กราฟปรียบเทียบจำนวนกฎความสัมพันธ์ระหว่างการหากฎความสัมพันธ์แบบดั้งเดิมและการหากฎความสัมพันธ์จากข้อมูลหลายแหล่งในข้อมูลผู้ป่วยโรคหัวใจ ..... 31
รูปที่ 5.1 แนวคิดการทำเหมืองข้อมูลแบบรวมฐานเปรียบเทียบกับแบบกระจาย ..... 36

# บทที่ 1
# บทนำ

## 1.1 ความสำคัญและที่มาของปัญหาการวิจัย

ในปัจจุบันข้อมูลสุขภาพของประชาชนที่มารับบริการจากหน่วยงานสาธารณสุขทั้งของรัฐและเอกชนมักจะถูกบันทึกไว้ในฐานข้อมูล ลักษณะของข้อมูลอิเล็กทรอนิกส์เหล่านี้ประกอบด้วยไฟล์หลายประเภท ได้แก่ ไฟล์ข้อความที่เป็นข้อมูลพื้นฐานของคนไข้เกี่ยวกับ ชื่อ ที่อยู่ อายุ เพศ อาชีพ และอื่น ๆ ไฟล์ภาพที่เป็นภาพถ่ายเอ็กซเรย์หรือภาพถ่ายประเภทอื่น ๆ เช่น ภาพอุลตราซาวด์ ภาพเอ็มอาร์ไอ ไฟล์ข้อมูลที่เป็นตัวเลขและสัญญาณ เช่น ค่าทางเคมีที่ได้จากการตรวจทางห้องปฏิบัติการ ค่าที่อ่านจากเซ็นเซอร์และเครื่องมือวัด เช่น คลื่นไฟฟ้าหัวใจ (electrocardiogram, ECG)

ไฟล์ข้อมูลต่างลักษณะเหล่านี้ประกอบกันขึ้นเป็นระเบียนสุขภาพอิเล็กทรอนิกส์ (electronic health record, EHR) แนวคิดเกี่ยวกับระเบียนสุขภาพอิเล็กทรอนิกส์นี้ ได้รับการพัฒนาต่อเนื่องมาจากเวชระเบียนอิเล็กทรอนิกส์ (electronic medical record, EMR) โดยเพิ่มขอบเขตให้บันทึกข้อมูลด้านอื่น ๆ ของผู้รับการรักษาหลายด้านมากขึ้น นอกเหนือจากด้านการตรวจรักษาของแพทย์เท่านั้น การพัฒนาระบบระเบียนสุขภาพอิเล็กทรอนิกส์ มีวัตถุประสงค์หลักเพื่อช่วยเพิ่มประสิทธิภาพการให้บริการ และการส่งต่อการรักษาผู้ป่วย รวมถึงลดโอกาสการเกิดข้อผิดพลาดในการดูแลรักษาผู้ป่วย (Theera-Ampornpunt, 2010)

ระบบระเบียนสุขภาพอิเล็กทรอนิกส์ที่จะสามารถบรรลุวัตถุประสงค์ดังกล่าวได้ จะต้องมีคุณสมบัติพื้นฐานของฐานข้อมูล คือ การบันทึกข้อมูลที่ต่างแบบกัน (heterogeneous data) ทำได้อย่างถูกต้อง การเรียกดูข้อมูลทำได้รวดเร็วทันต่อการใช้งาน และการเก็บข้อมูลเป็นปัจจุบัน นอกจากนี้ระบบยังจะต้องมีคุณสมบัติพิเศษอื่นเพิ่มเติมขึ้น เพื่อช่วยสนับสนุนขีดความสามารถด้านการลดโอกาสเกิดข้อผิดพลาด และช่วยการตัดสินใจของบุคลากรทั้งทางคลินิกและทางการดูแลสุขภาพในเชิงป้องกัน

การเพิ่มคุณสมบัติพิเศษดังกล่าวของระบบระเบียนสุขภาพอิเล็กทรอนิกส์ จะต้องอาศัยเทคโนโลยีปัญญาประดิษฐ์ด้านการเรียนรู้ของเครื่อง และการวิเคราะห์ข้อมูลอัตโนมัติในลักษณะการทำเหมืองข้อมูล เทคนิคการทำเหมืองข้อมูลในปัจจุบัน มักจะถูกออกแบบมาให้วิเคราะห์ข้อมูลที่ถูกรวบรวมไว้ในแหล่งเดียว การประยุกต์ใช้เทคนิคเหมืองข้อมูลกับข้อมูลสุขภาพให้ใช้งานได้จริง จะต้องมีการพัฒนารูปแบบการวิเคราะห์ให้รองรับข้อมูลแบบกระจาย เนื่องจากในสถานการณ์จริงข้อมูลสุขภาพของผู้ป่วยรวมถึงข้อมูลการตรวจเช็คสุขภาพของประชาชนทั่วไป กระจายอยู่ในฐานข้อมูลของ

โรงพยาบาลและสถานบริการสาธารณสุขทั่วประเทศ การรวบรวมข้อมูลจากแหล่งข้อมูลที่กระจายและใช้รูปแบบการบันทึกข้อมูลที่ยังไม่เป็นมาตรฐานเดียวกันเป็นสิ่งที่ทำได้ยาก และใช้เวลานานมากในการรวบรวมข้อมูลรวมและแปลงรูปแบบข้อมูลให้ตรงกัน เวลาที่ต้องใช้ในการรวบรวมและแปลงข้อมูล อาจทำให้ไม่ได้ผลการวิเคราะห์ข้อมูลทันต่อการใช้งาน

โครงการวิจัยนี้จึงเสนอแนวทางใหม่สำหรับแก้ปัญหาของการวิเคราะห์ข้อมูลอัตโนมัติ ในลักษณะการทำเหมืองข้อมูลจากแหล่งข้อมูลที่อยู่กระจายกัน โดยไม่ต้องใช้ขั้นตอนการรวมข้อมูล แต่จะใช้วิธีการวิเคราะห์ข้อมูล ณ แหล่งกำเนิดของข้อมูลนั้น (on-site analysis) ผลลัพธ์ที่ได้จะเป็นแบบรูปของข้อมูลเฉพาะแหล่งนั้น เรียกว่า แบบรูปเฉพาะที่ (local patterns) แบบรูปเฉพาะที่ที่วิเคราะห์ได้จากแต่ละฐานข้อมูล จะถูกส่งมาที่คอมพิวเตอร์ส่วนกลางเพื่อรวบรวมผลวิเคราะห์ให้เป็นแบบรูปกลุ่มเดียว เรียกว่า แบบรูปครอบคลุม (global patterns)

ในงานวิจัยนี้เน้นการวิเคราะห์แบบรูปที่แสดงความสัมพันธ์เชื่อมโยงอย่างเด่นชัดของลักษณะประจำ หรือแอททริบิวต์ที่เป็นฟิลด์ข้อมูลในระเบียนข้อมูลสุขภาพ แบบรูปประเภทนี้เรียกว่า แบบรูปความสัมพันธ์ (association patterns) ในการวิเคราะห์แบบรูปความสัมพันธ์ครอบคลุมจะใช้วิธีการอนุมาน (inference) จากแบบรูปเฉพาะที่หลายกลุ่มที่ถูกส่งมาจากฐานข้อมูลแบบกระจาย โดยจะแปลงแบบรูปเฉพาะที่เหล่านั้นให้อยู่ในลักษณะของภาษาธรรมชาติ (natural language) เพื่อให้สามารถใช้กลไกอนุมานช่วยในกระบวนการสังเคราะห์เป็นแบบรูปครอบคลุมได้ นอกจากนั้นการตรวจสอบความหมายของแบบรูปเฉพาะที่จากแต่ละฐานข้อมูล เพื่อเปรียบเทียบกับแบบรูปครอบคลุมที่ได้เป็นผลลัพธ์สุดท้ายจะใช้วิธีสร้างออนโทโลยี (ontology) เพื่อตรวจสอบการคงตัวของความหมายของแบบรูปครอบคลุมที่สังเคราะห์ขึ้น

## 1.2 วัตถุประสงค์ของโครงการวิจัย

โครงการวิจัยนี้เน้นการสร้างองค์ความรู้ใหม่ โดยมีวัตถุประสงค์หลักเพื่อพัฒนากลไกการรวมแบบรูปความสัมพันธ์เฉพาะที่ (local association patterns) ที่วิเคราะห์ได้จากฐานข้อมูลสุขภาพแบบกระจาย การรวมแบบรูปหรือแพทเทิร์นเป็นการรวมแบบรูปความสัมพันธ์เฉพาะที่หลายกลุ่มให้เป็นแบบรูปความสัมพันธ์ครอบคลุม (global association patterns) เพียงกลุ่มเดียว

นอกจากการสร้างกลไกในการรวมแบบรูปเฉพาะที่ งานวิจัยนี้ยังมีการพัฒนากลไกในการปรับปรุงแบบรูปความสัมพันธ์ที่สังเคราะห์และรวบรวมจากข้อมูลหลายแหล่ง โดยในขั้นตอนการสังเคราะห์แบบรูปความสัมพันธ์ครอบคลุมจะใช้กลไกอนุมาน (inference engine) และกลไกการให้เหตุผล (reasoning mechanism) ช่วยในการสังเคราะห์แบบรูปที่ปราศจากการซ้ำซ้อนหรือการขัดแย้งของความหมาย

### 1.3 ขอบเขตของการวิจัย

ข้อมูลที่ใช้ทดสอบกลไกการสังเคราะห์แบบรูปความสัมพันธ์ครอบคลุมที่พัฒนาขึ้นใหม่ในโครงการวิจัยนี้ จะใช้ข้อมูลที่เป็นข้อมูลระเบียนสุขภาพอิเล็กทรอนิกส์ที่เป็นข้อมูลสังเคราะห์ (www.emrbots.org) และข้อมูลจากฐานข้อมูลมาตรฐาน UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets.html)

ในส่วนเทคนิคการทำเหมืองข้อมูลเพื่อเรียนรู้ความสัมพันธ์ (association mining) งานวิจัยนี้ใช้อัลกอริทึม Apriori (Agrawal & Srikant, 1994) เป็นพื้นฐานในการทำงาน

### 1.4 ประโยชน์ที่ได้รับ

งานวิจัยนี้ได้รับประโยชน์จากการดำเนินงานโครงการ ในหลายด้านได้แก่

(1) แบบรูปความสัมพันธ์ (association patterns) ที่เรียนรู้จากระเบียนข้อมูลสุขภาพอิเล็กทรอนิกส์ จะช่วยเปิดเผยความเชื่อมโยงของลักษณะประจำ (attributes, features) ที่มักจะปรากฏร่วมกัน การเรียนรู้ความสัมพันธ์นี้จะเป็นประโยชน์ต่อการวิเคราะห์แบบรูปที่นำไปสู่ภาวะที่ก่อให้เกิดโรค ทำให้การวางแผนการรักษามีโอกาสประสบความสำเร็จสูงขึ้นและการเรียนรู้แบบรูปที่อาจเป็นจุดเริ่มต้นของการนำไปสู่การเกิดโรคหรือพยาธิสภาพ จะเป็นประโยชน์ต่องานสาธารณสุขเชิงป้องกัน

(2) งานวิจัยนี้เป็นการศึกษา ค้นคว้า ออกแบบ และพัฒนาขั้นตอนวิธีการ ทั้งในเชิงทฤษฎีและเชิงทดลองเพื่อให้ได้องค์ความรู้ใหม่ในด้านการรวมและวิเคราะห์กฎจากการทำเหมืองข้อมูลความสัมพันธ์ จากฐานข้อมูลแบบกระจาย ประโยชน์ที่ได้รับโดยตรงคือเทคนิคและอัลกอริทึมใหม่ที่สามารถตีพิมพ์ผลงานวิจัยในวารสารวิชาการระดับนานาชาติได้จำนวน 3 บทความ และยังช่วยพัฒนานักวิจัยรุ่นใหม่ที่เป็นนักศึกษาในระดับปริญญาโทและเอก ให้สามารถทำงานวิจัยในระดับสูงได้

(3) การพัฒนากลไกการสังเคราะห์แบบรูปความสัมพันธ์ครอบคลุมจากข้อมูลที่กระจายอยู่ในหลายแหล่ง ในลักษณะของโปรแกรมต้นแบบหรือ prototype ทำให้ได้โปรแกรมต้นแบบที่สามารถจดลิขสิทธิ์ได้จำนวน 1 โปรแกรม ได้แก่ โปรแกรมรวมกฎความสัมพันธ์ด้วยกลไกการให้เหตุผลเชิงตรรกะ (program to integrate association rules with reasoning mechanism) ลิขสิทธิ์เลขที่ ว1. 5360 ออกให้ ณ วันที่ 8 เมษายน 2558

# บทที่ 2
# เอกสารและงานวิจัยที่เกี่ยวข้อง

## 2.1 การเรียนรู้กฎความสัมพันธ์จากข้อมูลหลายแหล่ง

เทคนิคการทำเหมืองข้อมูลเพื่อเรียนรู้กฎความสัมพันธ์ (association rule mining) โดยทั่วไปสามารถเรียนรู้ได้จากข้อมูลในแหล่งเดียว แต่ในการทำงานจริงข้อมูลบางประเภทไม่ได้มีการจัดเก็บข้อมูลไว้ในแหล่งข้อมูลเพียงแหล่งเดียว ทำให้การนำข้อมูลเหล่านี้มาเรียนรู้เพื่อหากฎความสัมพันธ์จำเป็นต้องรวมข้อมูลที่ถูกกระจายกันอยู่ให้เป็นข้อมูลเพียงชุดเดียว เพื่อสามารถนำไปหากฎความสัมพันธ์ได้ แต่เนื่องจากข้อมูลที่กระจายกันอยู่นั้นมีการจัดเก็บข้อมูลที่มีลักษณะแตกต่างกันออกไป ซึ่งอาจทำให้การนำข้อมูลเหล่านี้มารวมกันแล้วนำไปหากฎความสัมพันธ์นั้นได้ประสิทธิภาพที่ไม่ดีพอสำหรับการนำไปใช้งาน ดังนั้นนักวิจัย (Li et al., 2003) ได้เสนอแนวคิดของการประเมินความคล้ายคลึงของข้อมูลก่อนที่จะรวมข้อมูลที่กระจายกันอยู่ตามแหล่งต่าง ๆ โดยสามารถวัดความคล้ายคลึงกันของข้อมูลได้จากมาตรวัดความคล้ายคลึง (Similarity Measure) ซึ่งคำนวณได้จากสมการที่ 2-1 (Li et al., 2003) ซึ่งเป็นการวัดความคล้ายคลึงของข้อมูลสองชุดคือ ข้อมูล A และ B โดยค่าความคล้ายคลึงจะมีค่าอยู่ระหว่าง 0 ถึง 1 ค่าที่เข้าใกล้ 1 แสดงว่าข้อมูล A และ B มีความคล้ายคลึงกันมาก แต่ถ้าค่าเข้าใกล้ 0 แสดงว่าข้อมูล A และ B มีความคล้ายคลึงกันน้อย

$$Sim(A,B) = \frac{2I_3}{I_1 + I_2} \tag{2-1}$$

โดยกำหนดให้

$$I_1 = \sum_{i,j} \frac{|A_i \cap A_j|}{|A_i \cup A_j|} \log\left(1 + \frac{|A_i \cap A_j|}{|A_i \cup A_j|}\right) \min\{C_{Ai}, C_{Aj}\},$$

$$I_2 = \sum_{i,j} \frac{|B_i \cap B_j|}{|B_i \cup B_j|} \log\left(1 + \frac{|B_i \cap B_j|}{|B_i \cup B_j|}\right) \min\{C_{Bi}, C_{Bj}\},$$

$$I_3 = \sum_{i,j} \frac{|A_i \cap B_j|}{|A_i \cup B_j|} \log\left(1 + \frac{|A_i \cap B_j|}{|A_i \cup B_j|}\right) \min\{C_{Ai}, C_{Bj}\},$$

เมื่อ $|A_i \cap A_j|$ แทนจำนวนข้อมูลใน $A_i \cap A_j$ โดย $A_i$ หมายถึง เซตของแอททริบิวต์ที่ปรากฏบ่อยในข้อมูลกลุ่มย่อยที่ i ของชุดข้อมูล A

$2/(I_1 + I_2)$ ทำหน้าที่เป็น normalization factor

$|A_i \cap B_j| / |A_i \cup B_j|$ ใช้ตรวจสอบความคล้ายของเซตย่อย $A_i$ และ $B_j$

$\min\{C_{Ai}, C_{Bj}\}$ คือค่าสนับสนุนขั้นต่ำของเซตย่อย $A_i$ และ $B_j$ และค่านี้ทำหน้าที่เป็น weight factor

ตัวอย่างการนำมาตรวัดความคล้ายคลึงกันไปใช้สำหรับการเรียนรู้กฎความสัมพันธ์ โดยข้อมูลที่นำมาใช้ได้แก่ ฐานข้อมูลที่รวมบทความจากการประชุมวิชาการ 3 แหล่ง คือ Computational Geometry (COMPGEOM), Conference on Cryptography (CRYPT) และ European Conference on Cryptography (EUROCRYPT) ทั้งสามฐานข้อมูลนี้มีโครงสร้างตารางในฐานข้อมูลเป็นแบบเดียวกัน

การค้นหากฎความสัมพันธ์จากแต่ละฐานข้อมูลด้วยค่า Minimum support = 1% และ Minimum confidence = 80% ได้ผลลัพธ์เป็นชุดของกฎความสัมพันธ์ โดยแต่ละกฎเขียนอยู่ในรูปแบบ a,b --> c(x%, y%) เมื่อ a, b, c คือ item หรือแอททริบิวต์ที่ปรากฏในฐานข้อมูล x คือ ค่า support ของกฎ และ y คือค่า confidence ของกฎ รูปที่ 2.1 2.2 และ 2.3 แสดงกฎความสัมพันธ์ที่ได้จากข้อมูล CRYPT ข้อมูล EUROCRYPT และข้อมูล COMPGEOM ตามลำดับ

1: schem,secret ⟶ shar (1.9%,85.7%)
2: schem,shar ⟶ secret (2.0%,80%)
3: key,cryptosystem ⟶ public (1.6%,83.3%)
4: public,cryptosystem ⟶ key (1.6%,83.3%)

รูปที่ 2.1 กฎความสัมพันธ์ที่ได้จากข้อมูล CRYPT (Li et al., 2003)

1. adapt ⟶ secur (1.2%,87.5%)
2. boolean ⟶ func (1.5%,90.0%)
3. digit ⟶ signatur (1.5%,80.0%)
4. public ⟶ key (3.0%,80.0%)
5. shar ⟶ secret (3.4%,82.6%)
6. logarithm ⟶ discrete (2.1%,100.0%)
7. low ⟶ bound (1.2%,87.5%)
8: schem,secret ⟶ shar (1.8%,100.0%)
9: schem,shar ⟶ secret (2.1%,85.7%)
10: public,cryptosystem ⟶ key (1.3%,100.0%)

รูปที่ 2.2 กฎความสัมพันธ์ที่ได้จากข้อมูล EUROCRYPT (Li et al., 2003)

1. hull ⟶ convex (2.4%,94.1%)
2. short ⟶ path (3.9%,89.3%)
3. voronoi ⟶ diagram (3.9%,89.3%)
4. diagram ⟶ voronoi (3.6%,96.2%)
5. algorithm, convex ⟶ hull (1.1%,87.5%)
6. simpl,visibl ⟶ polygon (1.1%,87.5%)
7. minim,tree ⟶ span (1.2%,88.9%)
8: minim,span ⟶ tree (1.1%,100.0%)

รูปที่ 2.3 กฎความสัมพันธ์ที่ได้จากข้อมูล COMPGEOM (Li et al., 2003)

รูปที่ 2.1-2.3 แสดงกฎความสัมพันธ์ที่ได้จากการเรียนรู้ในแต่ละฐานข้อมูล ในกรณีที่เป็นการเรียนรู้ความสัมพันธ์ในแบบกระจายโดยเรียนรู้จากแต่ละแหล่งข้อมูล เทคนิคการเรียนรู้แบบกระจายที่ใช้วิธีการวัดความคล้ายคลึงของความสัมพันธ์ในฐานข้อมูล (Li et al., 2003) จะวัดความคล้ายของฐานข้อมูลแต่ละคู่ คือ คู่ของข้อมูล CRYPT-EUROCRYPT คู่ของข้อมูล CRYPT-COMPGEOM และคู่ของข้อมูล EUROCRYPT-COMPGEOM จะสังเกตได้ว่าบางกฎความสัมพันธ์ที่เรียนรู้ได้จากข้อมูล CRYPT และข้อมูล EUROCRYPT มีลักษณะเหมือนกัน และเมื่อคำนวณค่าความคล้ายคลึง ผลลัพธ์ที่ได้คือคู่ของข้อมูล CRYPT-EUROCRYPT ให้ค่ามากที่สุด ในขณะที่คู่ของข้อมูล CRYPT-COMPGEOM และ EUROCRYPT-COMPGEOM มีค่าความคล้ายคลึงต่ำ ดังนั้นการหากฎความสัมพันธ์แบบกระจายจากสามฐานข้อมูลจึงสามารถแบ่งออกได้เป็น 2 กลุ่ม คือในกลุ่มแรกเป็นการหากฎความสัมพันธ์ระหว่างข้อมูล CRYPT และ EUROCRYPT ในกลุ่มที่สองเป็นการหากฎความสัมพันธ์จากข้อมูล COMPGEOM ผลลัพธ์ที่ได้แสดงดังรูปที่ 2.4

1. logarithm ⟶ discrete (2.2%,87.1%)
2: schem,secret ⟶ shar (1.8%,92.3%)
3: schem,shar ⟶ secret (2.0%,82.8%)
4: public,cryptosystem ⟶ key (1.5%,90.5%)

กลุ่ม CRYPT-EUROCRYPT

1. hull ⟶ convex (2.4%,94.1%)
2. short ⟶ path (3.9%,89.3%)
3. voronoi ⟶ diagram (3.9%,89.3%)
4. diagram ⟶ voronoi (3.6%,96.2%)
5. algorithm, convex ⟶ hull (1.1%,87.5%)
6. simpl,visibl ⟶ polygon (1.1%,87.5%)
7. minim,tree ⟶ span (1.2%,88.9%)
8: minim,span ⟶ tree (1.1%,100.0%)

กลุ่ม COMPGEOM

รูปที่ 2.4 กฎความสัมพันธ์ที่ได้จากเรียนรู้แบบกระจายจากสองกลุ่มข้อมูล CRYPT- EUROCRYPT และข้อมูล COMPGEOM (Li et al., 2003)



จากตัวอย่างข้างต้นของการเรียนรู้กฎความสัมพันธ์แบบกระจาย จะเห็นได้ว่าการเรียนรู้กฎความสัมพันธ์จะต้องมีการจัดลำดับความคล้ายของเนื้อหาในฐานข้อมูลก่อนที่จะสร้างกฎความสัมพันธ์จากฐานข้อมูลแบบกระจาย โครงการวิจัยนี้เสนอแนวทางที่แตกต่างจากรูปแบบตามตัวอย่างข่างต้น โดยในงานวิจัยนี้คณะผู้วิจัยเสนอแนวคิดของการเรียนรู้กฎความสัมพันธ์จากแต่ละแหล่งฐานข้อมูลโดยอิสระโดยไม่ต้องคำนวณค่าความคล้าย การลดทอนกฎที่ซ้ำซ้อนจะเกิดขึ้นในขั้นตอนหลังจากที่รวมผลลัพธ์ที่เรียนรู้ได้จากแต่ละแหล่งโดยอาศัยกลไกการอนุมานเชิงตรรกะซึ่งให้ประสิทธิภาพที่ดี

## 2.2 งานวิจัยที่เกี่ยวข้อง

ระเบียนสุขภาพอิเล็กทรอนิกส์ หรือ EHR ถูกออกแบบมาให้อำนวยประโยชน์ต่อคนไข้มากที่สุด ในด้านการได้รับการวางแผนรักษา การดูแลสุขภาพ และการส่งต่อการดูแลรักษาอย่างรวดเร็วและมีประสิทธิภาพ โดยมีพื้นฐานของแนวคิดคือ ในการดูแลรักษาผู้ป่วยจะต้องมีหลายหน่วยงานร่วมมือกัน ดังนั้นการกำหนดรูปแบบมาตรฐานกลางของระเบียนข้อมูลอิเล็กทรอนิกส์ที่จะต้องถูกส่งต่อจากหน่วยงานหนึ่งไปยังอีกหน่วยงานหนึ่ง จะมีประโยชน์และก่อให้เกิดประสิทธิผลดีที่สุดต่อการรักษาผู้ป่วย มาตรฐานกลางที่กำหนดโดยหน่วยงานไอเอสโอ (ISO/TR 20514, 2005) เป็นเกณฑ์พื้นฐานที่จะช่วยให้ข้อมูลอิเล็กทรอนิกส์มีฟังก์ชันและความหมายที่แลกเปลี่ยนกันได้ระหว่างหน่วยงาน รายละเอียดอื่น ๆ ที่เพิ่มเติมจากมาตรฐานไอเอสโอ จะขึ้นอยู่กับหน่วยงานกลางของแต่ละประเทศหรือแต่ละภูมิภาค (Hayrinen et al., 2008)

การมีระบบ EHR ที่ช่วยให้การเข้าถึงข้อมูลผู้ป่วยทำได้รวดเร็วและเป็นระบบ สร้างความตื่นตัวอย่างมากให้กับนักวิจัยที่ทำงานเกี่ยวข้องกับข้อมูลสุขภาพ ให้หันมาพิจารณาใช้ประโยชน์จากข้อมูลอิเล็กทรอนิกส์เหล่านี้ รวมถึงให้ความสนใจในการประเมินความคุ้มค่าของระบบ (Peterson et al., 2011; Coorevits et al., 2013; Ravel at al., 2015)

แต่ถึงแม้จะมีระบบ EHR ที่ช่วยให้การเก็บข้อมูลทั้งที่มีโครงสร้างและไม่มีโครงสร้าง ทำได้อย่างสมบูรณ์ แต่การใช้ประโยชน์จากข้อมูลเหล่านั้นยังเป็นเรื่องที่ต้องได้รับการพัฒนาอีกมาก แนวทางการพัฒนาความเร็วและความสะดวกในการวิเคราะห์ข้อมูลจากฐานข้อมูลสุขภาพ ด้วยการใช้เทคนิคเหมืองข้อมูลเป็นแนวทางหนึ่งที่ได้รับความสนใจอย่างกว้างขวางมาเป็นระยะเวลานาน ดังปรากฏในงานวิจัยของดีเลนและคณะ (Delen at al., 2005) ที่ประยุกต์ใช้เทคนิคโครงข่ายประสาทเทียม ต้นไม้ตัดสินใจ และการวิเคราะห์การถดถอยโลจิสติก สร้างโมเดลเพื่อทำนายการอยู่รอดของคนไข้จากโรคมะเร็งเต้านมโดยใช้การเรียนรู้โมเดลจากข้อมูลคนไข้ 200,000 ราย

ทีมงานของฟิลลิป-เรน (Phillips-Wren et al., 2008) วิเคราะห์การใช้ประโยชน์จากทรัพยากรสาธารณสุขเพื่อการดูแลรักษาคนไข้โรคมะเร็งปอด ด้วยการใช้เทคนิคต้นไม้ตัดสินใจร่วมกับโครงข่ายประสาทเทียม ทั้งสองงานวิจัยนี้เป็นการทำเหมืองข้อมูลในลักษณะการจำแนกประเภท (classification)

งานวิจัยของชิโนะซะวะและคณะ (Shinozawa et al., 2009) เสนอการใช้เทคนิคเหมืองข้อมูลประเภทการค้นหาความสัมพันธ์ (association) เพื่อแสดงความเชื่อมโยงของข้อมูลที่ปรากฏในผลตรวจเลือดและปัสสาวะจากการตรวจสุขภาพประจำปี และในช่วงเวลาสี่ปีหลังจากนั้น (Sakata et al., 2013) ทีมงานนี้ได้พัฒนาแนวคิดเพิ่มขึ้นให้นำไปสู่การออกแบบระบบผู้เชี่ยวชาญเพื่อสนับสนุนการตัดสินใจทางคลินิก

ทีมงานวิจัยจากตุรกี (Yuregir et al., 2010) ได้เสนอแนวคิดของระบบสนับสนุนการตัดสินใจทางการแพทย์เช่นเดียวกัน โดยได้เสนอการรวมโมดูล self organizing map (SOM) ไว้ในระบบเพื่อเชื่อมโยงแบบรูปสำหรับแจ้งเตือนการระบาดของโรคที่เกิดจากเชื้อแบคทีเรีย *Ligionella pneumophila*

การทำเหมืองข้อมูลกับระบบระเบียนสุขภาพอิเล็กทรอนิกส์ ซึ่งมีขนาดใหญ่กว่าฐานข้อมูลสุขภาพเฉพาะด้าน เริ่มปรากฏในงานวิจัยที่เป็นความร่วมมือระหว่างนักวิจัยจากบริษัทไอบีเอ็มและทีมวิจัยจากมหาวิทยาลัยโคลัมเบีย ประเทศสหรัฐอเมริกา (Lee et al., 2011) ทีมวิจัยขนาดใหญ่นี้วิเคราะห์แบบรูปที่ซ่อนอยู่ เพื่อค้นหาความเชื่อมโยงของเหตุการณ์ที่เกี่ยวข้องกับเวลาจากฐานข้อมูล EHR ขนาดใหญ่

ในปี 2013 ทีมวิจัยจากประเทศสวีเดน (Banaee et al., 2013) ได้ศึกษาการใช้เทคนิคเหมืองข้อมูลสำหรับเฝ้าติดตามกระบวนการทำงานของอวัยวะต่าง ๆ ผ่านอุปกรณ์ส่งสัญญาณซึ่งเป็นตัวเซ็นเซอร์ที่ใช้ติดที่ส่วนต่าง ๆ ของร่างกาย (wearable sensors)

ถึงแม้การประยุกต์เทคนิคเหมืองข้อมูลกับฐานข้อมูลสุขภาพ จะประสบความสำเร็จในเบื้องต้นดังที่ปรากฏในงานวิจัยจำนวนมาก โดยสามารถค้นพบโมเดลหรือแบบรูปที่เป็นประโยชน์ต่อการทำนายการเกิดโรค ค้นพบความเชื่อมโยงของเหตุแห่งการนำไปสู่ภาวะเสี่ยงของการเกิดโรคซึ่งจะช่วยให้การป้องกันทำได้ทันเวลา แต่ข้อมูลสุขภาพในปัจจุบันที่ส่วนใหญ่ได้รับการบันทึกอยู่ในรูปแบบอิเล็กทรอนิกส์มีปริมาณเพิ่มขึ้นอย่างก้าวกระโดด ทำให้การวิเคราะห์ข้อมูลกับข้อมูลขนาดใหญ่ต้องใช้เทคนิคและวิธีการที่มีประสิทธิภาพมากขึ้น (Herland et al., 2014)

นอกจากนี้อีกสาเหตุหนึ่งที่ทำให้การวิเคราะห์ข้อมูลจากฐานข้อมูล EHR ทำได้ยากคือลักษณะของแหล่งข้อมูลที่เป็นฐานข้อมูลแบบกระจาย (Atilgan & Dogan, 2008) การวิเคราะห์ข้อมูลจึงมักกระทำที่แหล่งข้อมูล (on-site analysis) เช่น งานวิจัยที่เสนอโดยทีมงานนานาชาติที่เป็นผู้เชี่ยวชาญด้านออร์โทพีดิกส์ (Banerjee et al., 2014) ได้ใช้วิธีการวิเคราะห์อัตราการรอดชีวิตของคนไข้ที่ฐานข้อมูลแต่ละแหล่ง จากนั้นรวบรวมเฉพาะผลการวิเคราะห์นำมาสร้างเป็นการวิเคราะห์รวบยอด (meta-analysis) ที่คอมพิวเตอร์ส่วนกลาง บางทีมวิจัยใช้การพัฒนาด้วยการใช้เอเจนต์เพื่อทำเหมืองข้อมูลแบบกระจาย (Zaidi et al., 2002) รวมไปถึงการนำเสนอสถาปัตยกรรมระบบสนับสนุนการตัดสินใจทางคลินิกแบบกระจาย (El-Sappgh & El-Masri, 2014)

การพัฒนาเทคนิคการทำเหมืองข้อมูลให้ทำงานได้ดีกับฐานข้อมูลสุขภาพ ที่มีลักษณะการจัดเก็บข้อมูลแบบกระจาย นอกจากการออกแบบเฟรมเวิร์คและสถาปัตยกรรมแล้ว จำเป็นต้องมีการพัฒนาในระดับขั้นตอนวิธีของกระบวนการทำเหมืองข้อมูล การจัดการกับข้อมูลแบบกระจายด้วยวิธีการค้นหาแบบรูปเฉพาะที่เพื่อรวมเป็นแบบรูปครอบคลุม เริ่มได้รับความสนใจจากนักวิจัยในสาขาการเรียนรู้ของเครื่องและการทำเหมืองข้อมูล (Yin et al., 2010; Lin et al., 2013) แต่การรวมแบบ

รูปเฉพาะที่หลายกลุ่มให้เป็นแบบรูปครอบคลุมเพียงกลุ่มเดียวยังมีแนวทางที่ค่อนข้างจำกัด (Sulzmann & Furnkranz, 2008)

โครงการวิจัยนี้จึงมีแนวคิดในการการรวมแบบรูปเฉพาะที่ ซึ่งเรียนรู้ได้จากแต่ละฐานข้อมูล ผ่านกระบวนการอนุมานและกลไกการรวมกฎ เพื่อให้ได้แบบรูปครอบคลุมที่ปรากฏบ่อยอย่างครบถ้วน และไม่มีความขัดแย้งในความหมายของแบบรูปที่เรียนรู้ได้

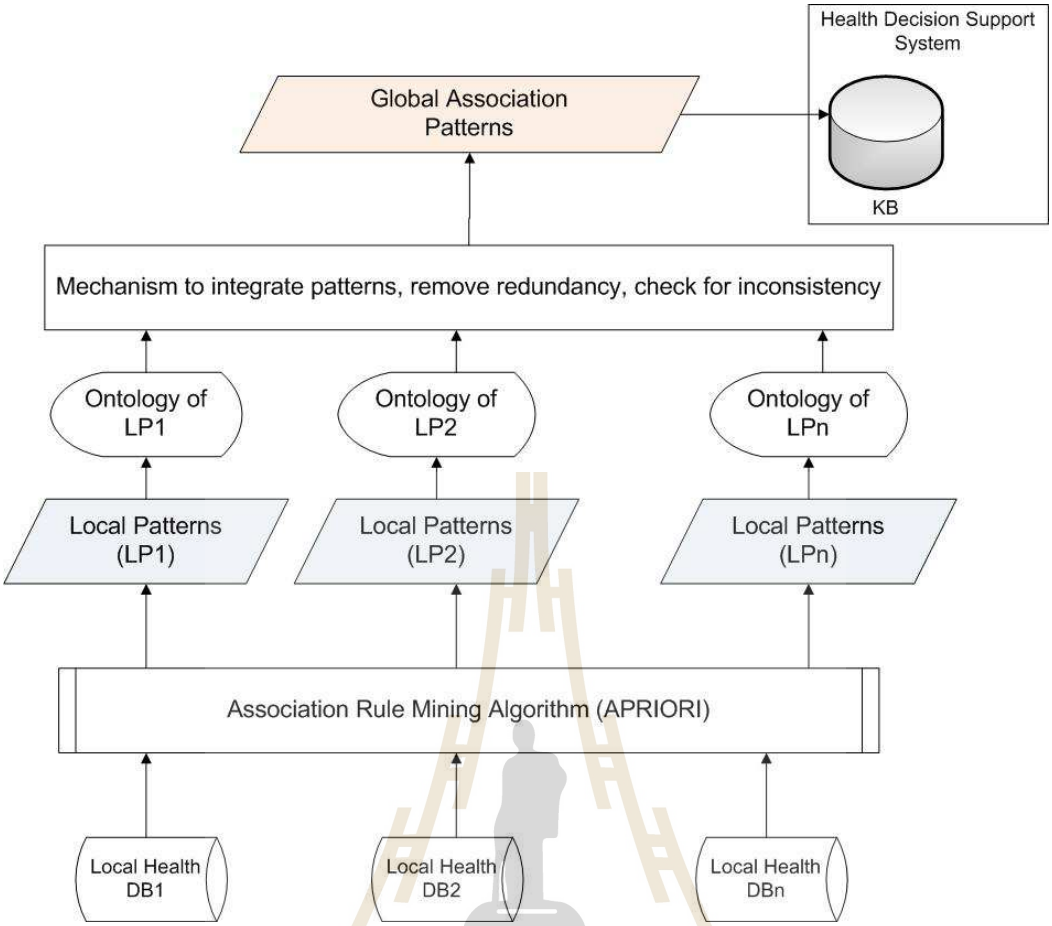
# บทที่ 3
# การออกแบบและพัฒนาวิธีการสังเคราะห์แบบรูปความสัมพันธ์จากข้อมูลหลายแหล่ง

## 3.1 กรอบแนวคิดของงานวิจัย

โครงการวิจัยนี้ต้องการหาแนวทางเรียนรู้จากข้อมูลแบบกระจาย เพื่อให้ได้ผลลัพธ์เป็นแบบรูปความสัมพันธ์ที่แสดงในลักษณะของกฎความสัมพันธ์ “ถ้าเกิดลักษณะ X แล้วจะเกิดลักษณะ Y ร่วมด้วย” หรือเขียนในรูปแบบย่อได้เป็น X -> Y การพัฒนางานวิจัยจะไม่ได้เน้นที่เทคนิคการทำเหมืองข้อมูลแบบหาความสัมพันธ์ (association mining) เนื่องจากเทคนิคที่ใช้อยู่ทั่วไปในปัจจุบัน มีประสิทธิภาพสูงเพียงพอสำหรับการค้นหาความสัมพันธ์จากฐานข้อมูลแหล่งเดียว แต่ในกรณีที่เป็นฐานข้อมูลแบบกระจาย จำเป็นต้องพัฒนาเทคนิคใหม่มาช่วยในกระบวนการค้นหาและรวมกฎความสัมพันธ์ ซึ่งงานวิจัยในลักษณะนี้ยังปรากฏไม่มากนักและยังต้องการการค้นคว้าแนวทางต่าง ๆ ที่จะช่วยเพิ่มประสิทธิภาพการรวมความรู้ที่เรียนรู้ได้จากแต่ละฐานข้อมูล

จุดเน้นของโครงการวิจัยนี้ จึงมุ่งที่การออกแบบขั้นตอนต่อจากการค้นหาความสัมพันธ์ที่ได้จากแต่ละแหล่งข้อมูลในลักษณะของแบบรูปเฉพาะที่ (local patterns) ซึ่งหมายถึงกฎความสัมพันธ์ที่เรียนรู้ได้จากฐานข้อมูลย่อย เมื่อต้องการได้กฎความสัมพันธ์รวมจากทุกฐานข้อมูลย่อย จึงต้องมีการรวมแบบรูปเฉพาะที่แต่ละกลุ่มเข้าด้วยกันให้เป็นแบบรูปครอบคลุม (global patterns) กลุ่มเดียว ในขั้นตอนการรวมกฎความสัมพันธ์ที่เป็นแบบรูปเฉพาะที่ จำเป็นต้องมีเทคนิคในการวิเคราะห์และจัดการกับแต่ละกลุ่มของแบบรูปเฉพาะที่ ในโครงการวิจัยนี้เสนอแนวทางเชิงตรรกะด้วยการใช้กลไกการอนุมานเพื่อวิเคราะห์ความซ้ำซ้อนหรือการขาดหายไปของความรู้ที่ได้จากแต่ละฐานข้อมูล รวมถึงออกแบบแนวทางวิเคราะห์และจัดการกับความขัดแย้งที่อาจปรากฏในแบบรูปเฉพาะที่ที่ได้จากฐานข้อมูลต่างแหล่ง ผลลัพธ์สุดท้ายจะเป็นแบบรูปครอบคลุมเพียงกลุ่มเดียวที่อยู่ในลักษณะกฎความสัมพันธ์ ที่จะสามารถส่งต่อไปเป็นฐานความรู้ของระบบสนับสนุนการตัดสินใจด้านสุขภาพได้

นอกจากนี้ในกระบวนการวิเคราะห์แบบรูปเฉพาะที่ที่ได้จากแต่ละฐานข้อมูล จะใช้ออนโทโลยีช่วยในการพิจารณาความหมายที่เหมือนหรือต่างกันในความรู้ที่เรียนรู้ได้จากแต่ละฐานข้อมูล ขั้นตอนหลักของงานวิจัยนี้แสดงได้ดังแผนภาพในรูปที่ 3.1




**รูปที่ 3.1** กรอบการวิจัยการสังเคราะห์แบบรูปความสัมพันธ์ครอบคลุมจากข้อมูลหลายแหล่งเพื่อสนับสนุนสนเทศศาสตร์สุขภาพแบบกระจาย

### 3.2 ขั้นตอนการดำเนินงานวิจัย

การดำเนินงานหลักของงานวิจัยนี้คือ การพัฒนากลไกการค้นหาและรวมแบบรูปเฉพาะที่ ซึ่งได้แก่ขั้นตอน Mechanism to integrate patterns, remove redundancy, check for inconsistency ในรูปที่ 2.1 ทั้งนี้ในการอธิบายรายละเอียดของขั้นตอนการพัฒนากลไกการค้นหาและรวมแบบรูปเฉพาะที่ จะเรียกแบบรูปเฉพาะที่ว่ากฎความสัมพันธ์เพื่อความเข้าใจที่ต่อเนื่องจากขั้นตอนการค้นหากฎความสัมพันธ์ด้วยอัลกอริทึม APRIORI จากฐานข้อมูลที่กระจายอยู่ในหลายแหล่ง ขั้นตอนการดำเนินงานแบ่งออกเป็น 5 ขั้นตอน ดังนี้

ขั้นตอนที่ 1 การหากฎความสัมพันธ์แบบกระจายจากฐานข้อมูลในหลายแหล่ง

การทำงานในขั้นตอนนี้มีจุดมุ่งหมายเพื่อหากฎความสัมพันธ์จากข้อมูลที่กระจายกันอยู่ โดยในขั้นตอนนี้กระบวนการหากฎความสัมพันธ์ในแต่ละชุดข้อมูลนั้นจะไม่ขึ้นต่อกัน ผลลัพธ์ที่ได้จากขั้นตอนนี้ คือฐานความรู้ที่เก็บกฎความสัมพันธ์จากข้อมูลที่

กระจายกันอยู่ ซึ่งจำนวนของฐานความรู้ที่จะได้นั้นจะขึ้นอยู่กับจำนวนของชุดข้อมูลที่กระจายกันอยู่

ขั้นตอนที่ 2 การรวมกฎความสัมพันธ์

การทำงานในขั้นตอนนี้ต่อเนื่องจากขั้นตอนที่ 1 โดยจะดึงกฎความสัมพันธ์ที่มีลักษณะเหมือนกันออกมาจากฐานความรู้ที่กระจายกันอยู่เพื่อให้ได้ฐานความรู้เพียงชุดเดียว

ขั้นตอนที่ 3 การแปลงกฎความสัมพันธ์ให้เป็นภาษาธรรมชาติ

เนื่องจากกฎความสัมพันธ์ที่ได้จากขั้นตอนที่ 2 เป็นกฎความสัมพันธ์แบบกระจายที่ได้จากฐานข้อมูลหลายแหล่ง ปัญหาที่อาจจะเกิดขึ้นคืออาจเกิดความขัดแย้งกันเองของกฎความสัมพันธ์ และอาจเกิดการขาดหายไปของกฎความสัมพันธ์เมื่อเทียบกับการหากฎความสัมพันธ์แบบฐานข้อมูลรวม ดังนั้นขั้นตอนนี้จะเป็นการแปลงรูปแบบของกฎความสัมพันธ์จากรูปแบบทั่วไปให้อยู่ในรูปแบบของภาษาธรรมชาติเพื่อนำไปใช้ในขั้นตอนต่อไป

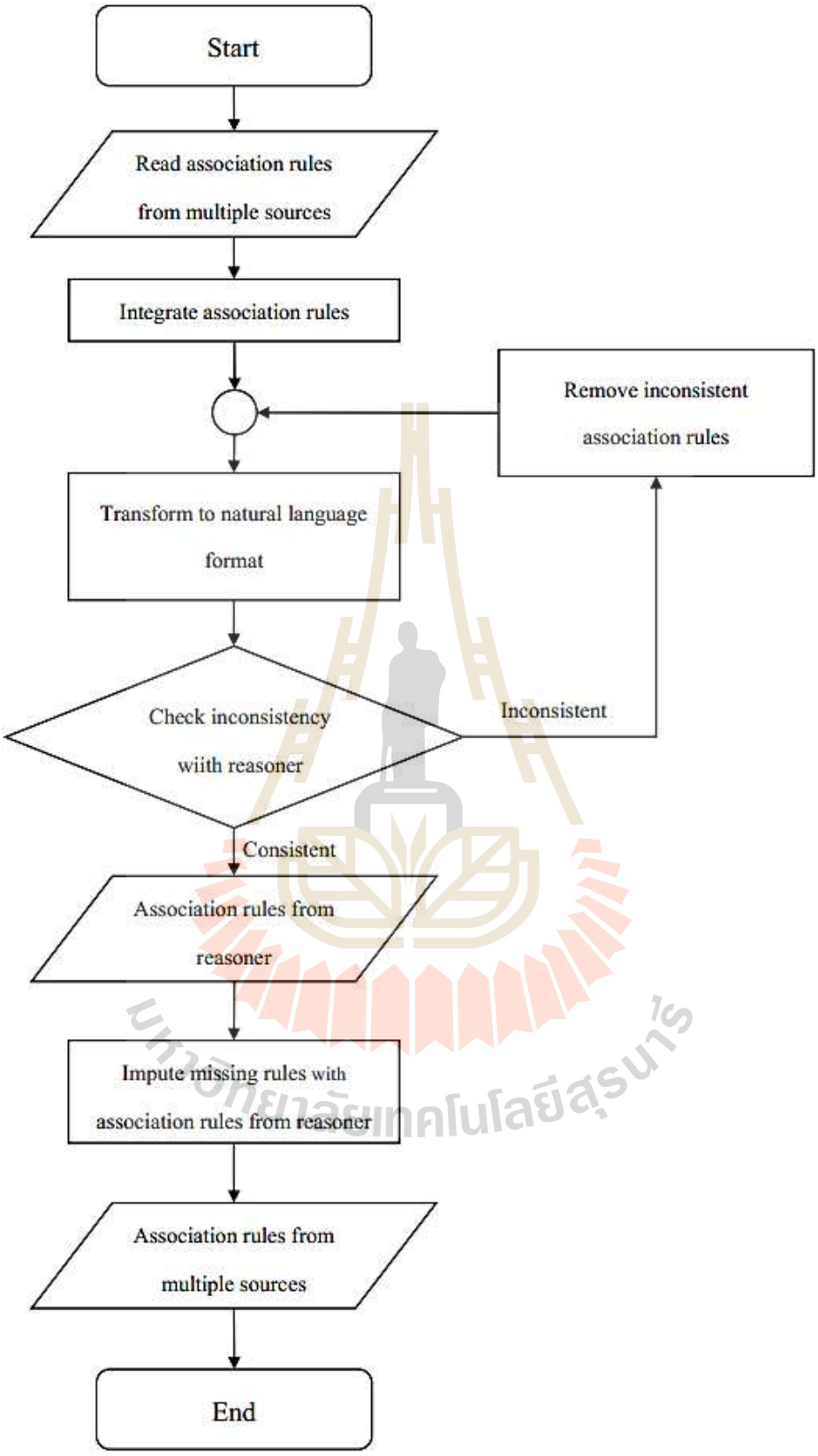
ขั้นตอนที่ 4 การตรวจสอบความขัดแย้งของกฎความสัมพันธ์

กฎความสัมพันธ์ที่อยู่ในรูปแบบของภาษาธรรมชาติที่ได้จากขั้นตอนที่ 3 จะถูกนำไปตรวจสอบความขัดแย้ง โดยผลลัพธ์ที่ได้จากขั้นตอนนี้คือผลการตรวจสอบที่สามารถบอกได้ว่ากฎความสัมพันธ์ที่ได้จากกระบวนการก่อนหน้านี้มีความขัดแย้งกันหรือไม่ ถ้าเกิดความขัดแย้งขึ้นจริงจะทำการลบกฎความสัมพันธ์นั้นทิ้งไปและทำการตรวจสอบความขัดแย้งใหม่ ในขั้นตอนนี้ยังได้ความรู้ใหม่จากกฎความสัมพันธ์ที่มีอยู่เดิม

ขั้นตอนที่ 5 การสร้างกฎความสัมพันธ์ที่ขาดหายไป

จากความรู้ใหม่ที่ได้จากกระบวนการของขั้นตอนที่ 4 สามารถนำไปสร้างเป็นกฎความสัมพันธ์แล้วนำไปเพิ่มเติมจากกฎความสัมพันธ์ของเดิมที่มีอยู่แล้วได้ ซึ่งในส่วนนี้สามารถชดเชยกฎความสัมพันธ์ที่ขาดหายไปจากการเรียนรู้ในฐานข้อมูลแบบกระจายได้ สุดท้ายจะได้ฐานความรู้ของกฎความสัมพันธ์เพียงชุดเดียวที่มาจากการหากฎความสัมพันธ์แบบกระจายที่มีประสิทธิภาพใกล้เคียงกับการหากฎความสัมพันธ์แบบที่ใช้ฐานข้อมูลรวมเพียงแหล่งเดียว

ขั้นตอนทั้ง 5 ขั้นตอนนี้แสดงเป็นผังงานได้ดังรูปที่ 3.2




**รูปที่ 3.2** ขั้นตอนการพัฒนากลไกการค้นหาและรวมกฎความสัมพันธ์ที่เป็นแบบรูปเฉพาะที่

## 3.3 การออกแบบขั้นตอนวิธีการ

### *3.3.1 วิธีการรวมกฎความสัมพันธ์*

การรวมกฎความสัมพันธ์ที่ได้จากฐานข้อมูลแบบกระจายนั้น จะรวมโดยนำมาเฉพาะกฎความสัมพันธ์ที่ปรากฏขึ้นในทุกแหล่ง เนื่องจากกฎความสัมพันธ์ที่ปรากฏในทุกแหล่งนั้นหมายถึงกฎความสัมพันธ์นั้น ๆ มีความสำคัญและมีประสิทธิภาพในการนำไปใช้งานในอนาคต แต่ในทางตรงข้ามการนำทุกกฎความสัมพันธ์มารวมกัน อาจได้กฎความสัมพันธ์ที่มากเกินไปและบางกฎความสัมพันธ์ที่ได้อาจไม่มีประสิทธิภาพเพียงพอสำหรับการนำไปใช้ประโยชน์เพื่อการทำนายข้อมูลในอนาคต ดังนั้นการรวมกฎความสัมพันธ์จะทำการดึงมาเฉพาะกฎความสัมพันธ์ที่ปรากฏในทุก ๆ แหล่งฐานข้อมูล

ข้อมูลตัวอย่างในตารางที่ 3.1 เป็นข้อมูลผู้ป่วยโรคมะเร็งเต้านม การจำลองข้อมูลนี้ให้เป็นลักษณะฐานข้อมูลแบบกระจาย จะใช้การแบ่งข้อมูลออกเป็น 3 ชุด ข้อมูลย่อยแต่ละชุดจะถูกนำไปหากฎความสัมพันธ์ ได้กฎความสัมพันธ์จำนวน 3 เซตคือ $R_1$, $R_2$, $R_3$ จากนั้นนำกฎความสัมพันธ์ที่ได้ไปสู่ขั้นตอนของการรวมกฎความสัมพันธ์ด้วยวิธีการอินเตอร์เซกชัน (intersection) จะได้ผลลัพธ์ดังรูปที่ 3.3

**ตารางที่ 3.1** ข้อมูลตัวอย่างผู้ป่วยโรคมะเร็ง

| age | menopause | tumor-size | inv-nodes | node-caps | Class |
|---|---|---|---|---|---|
| 40-49 | premeno | 15-19 | 0-2 | yes | recurrence-events |
| 50-59 | ge40 | 15-19 | 0-2 | no | no-recurrence-events |
| 50-59 | ge40 | 35-39 | 0-2 | no | recurrence-events |
| 50-59 | ge40 | 15-19 | 0-2 | no | no-recurrence-events |
| 40-49 | premeno | 15-19 | 0-2 | yes | recurrence-events |

| $R_1$ | $R_2$ | $R_3$ |
|---|---|---|
| {Class=no-recurrence-events}=>{inv-nodes=0-2} | - | - |
| {inv-nodes=0-2}=>{irradiat=no} | {inv-nodes=0-2}=>{irradiat=no} | {inv-nodes=0-2}=>{irradiat=no} |
| {irradiat=no}=>{node-caps=no} | {irradiat=no}=>{node-caps=no} | {irradiat=no}=>{node-caps=no} |

| R |
|---|
| {inv-nodes=0-2}=>{irradiat=no} |
| {irradiat=no}=>{node-caps=no} |

**รูปที่ 3.3** ผลลัพธ์ของการรวมกฎความสัมพันธ์ด้วยการทำอินเตอร์เซกชัน

*3.3.2 วิธีการแปลงรูปแบบกฎความสัมพันธ์ให้เป็นภาษาธรรมชาติ*

การหากฎความสัมพันธ์จากฐานข้อมูลแบบกระจายนั้น ไม่สามารถหากฎความสัมพันธ์ได้แบบตรงไปตรงมา เนื่องจากกฎความสัมพันธ์แบบกระจายที่หามาได้จากฐานข้อมูลย่อยแต่ละชุดโดยอิสระ อาจทำให้ได้กฎความสัมพันธ์ที่เกิดความขัดแย้งกันเอง ตัวอย่างเช่น

Rule 1: If X is a man, then X is a human.

Rule 2: If X is John, then X is a man.

Rule 3: If X is John, then X is not a human.

จากตัวอย่างกฎความสัมพันธ์ที่ได้ จะเห็นว่าเกิดความขัดแย้งของกฎความสัมพันธ์ที่ 2 และ 3 เนื่องจากทั้งสามกฎนี้เรียนรู้มาจากต่างฐานข้อมูล ดังนั้นจึงจำเป็นต้องมีเครื่องมือสำหรับการตรวจสอบความขัดแย้งซึ่งในงานวิจัยนี้ใช้ Attempto Controlled English หรือ ACE (Kuhn, 2014) เป็นเครื่องมือช่วยในการตรวจสอบ แต่ก่อนที่จะตรวจสอบความขัดแย้งได้ จะต้องแปลงกฎความสัมพันธ์ให้อยู่ในรูปแบบของภาษาธรรมชาติ โดยใช้เทคนิคการค้นหาและแทนที่ข้อความเพื่อทำให้กฎความสัมพันธ์ในรูปแบบทั่วไป เช่น ‘{A=B} => {B=C}’ อยู่ในรูปแบบภาษาธรรมชาติ เช่น ‘if X is a n:A_equal_B then X is a n:B_equal_C’ เป็นต้น ซึ่งข้อความหรือตัวอักษรที่จะต้องถูกแปลงมีดังต่อไปนี้

‘{’ แทนที่ด้วย ‘if X is a n:’

‘=>’ แทนที่ด้วย ‘then X is a’

‘=’ แทนที่ด้วย ‘_equal_’

‘,’ แทนที่ด้วย ‘and X is a n:’
‘{’ แทนที่ด้วย ‘n:’
‘}’ แทนที่ด้วย empty string

โดยภาษาธรรมชาติในที่นี้คือประโยคภาษาอังกฤษทั่วไป แต่จะเขียนโดยมีข้อกำหนดของภาษา ACE ซึ่งเป็นภาษาธรรวมชาติชนิดหนึ่งที่สามารถแทนความรู้แล้วนำไปสร้างเป็นออนโทโลยีได้ ดังตารางที่ 3.2 แสดงตัวอย่างการแปลงรูปแบบกฎความสัมพันธ์ทั่วไปให้อยู่ในรูปแบบภาษาธรรมชาติ จากรูปที่ 3.4 แสดงคำสั่งเทียมขั้นตอนการแปลงรูปแบบกฎความสัมพันธ์ทั่วไปให้อยู่ในรูปแบบภาษาธรรมชาติ โดยจะรับข้อมูลเป็นกฎความสัมพันธ์ที่ได้จากการรวมกฎความสัมพันธ์จากแหล่งความรู้ต่าง ๆ ซึ่งการแปลงรูปแบบกฎความสัมพันธ์นี้จะใช้เทคนิคของการค้นหาและแทนที่ในการค้นหารูปแบบของประโยคหรือสัญลักณ์ที่ต้องการแทนที่ด้วยประโยคหรือสัญลักณ์ที่ต้องแทนที่ลงไป ผลลัพธ์ที่ได้คือกฎความสัมพันธ์ที่อยู่ในรูปแบบของภาษาธรรมชาติที่สามารถนำไปแทนความรู้เพื่อนำไปสร้างเป็นออนโทโลยีได้

**ตารางที่ 3.2** รูปแบบกฎความสัมพันธ์ที่ถูกแปลงเป็นภาษาธรรมชาติ

| Original association rules | Association rules in ACE format |
|---|---|
| {Class=no-recurrence-events}=>{inv-nodes=0-2} | If X is a n: Class_equal_no-recurrence-events then X is a n:inv-nodes_equal_0-2. |
| {inv-nodes=0-2}=>{irradiat=no} | If X is a n:inv-nodes_equal_0-2 then X is a n:irradiat_equal_no. |
| {irradiat=no}=>{node-caps=no} | If X is a n:irradiat_equal_no then X is a n: node-caps_equal_no. |

**Algorithm Transform_to_Natural_Language**

//Input: C, an association rule set.

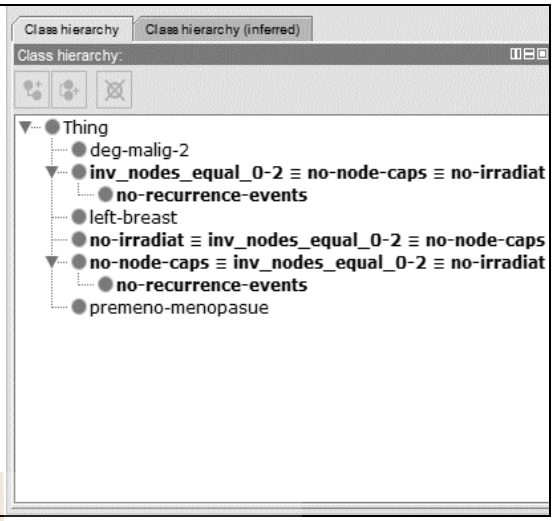
//Output: CACE, an association rule set in the form of natural language.

```
1.  Create CACE as empty list;
2.  Create D as dictionary = { '\A{' :'if X is a n:'
3.                             '=>' : 'then X is a',
4.                             '=' : '_equal_',
5.                             ',' : ' and X is a n:',
6.                             '{' : 'n:',
7.                             '}' : '',
8.  }
9.  For i =1 ← to length(C) do
10.         RN = multiple_replace(D, Ci);
11.         add RN to CACE;
12. End for
13. Return CACE
```

**รูปที่ 3.4** ขั้นตอนการแปลงรูปแบบกฎความสัมพันธ์ทั่วไปให้อยู่ในรูปแบบภาษาธรรมชาติ

*3.3.3 วิธีการตรวจสอบความขัดแย้งของกฎความสัมพันธ์*

ขั้นตอนวิธีในรูปที่ 3.4 เป็นขั้นตอนการแปลงรูปแบบกฎความสัมพันธ์ทั่วไปให้อยู่ในรูปแบบภาษาธรรมชาติ เพื่อใช้สำหรับการนำกฎความสัมพันธ์ไปสร้างเป็นออนโทโลยีหรือฐานข้อมูลที่มีการระบุความสัมพันธ์ของข้อมูล ดังแสดงในรูปที่ 3.5 ขั้นตอนต่อจากแปลงรูปแบบกฎความสัมพันธ์ให้เป็นรูปแบบภาษาธรรมชาติคือขั้นตอนการตรวจสอบความขัดแย้งของกฎ และนอกเหนือจากผลลัพธ์ที่ได้ว่ากฎความสัมพันธ์เกิดความขัดแย้งกันหรือไม่ ยังสามารถสร้างความรู้ใหม่หรือกฎความสัมพันธ์ใหม่จากกฎความสัมพันธ์ที่มีอยู่เดิมได้ ดังตารางที่ 3.3 แสดงตัวอย่างความรู้ใหม่ที่ได้จากการตรวจสอบความขัดแย้งของกฎความสัมพันธ์ที่ได้จากแหล่งต่าง ๆ

<table>
<tr>
<td>If X is a n: Class_equal_no-recurrence-events<br>then X is a n:inv-nodes_equal_0-2.<br>If X is a n:inv-nodes_equal_0-2<br>then X is a n:irradiat_equal_no.<br>If X is a n:irradiat_equal_no<br>then X is a n: node-caps_equal_no.</td>
<td> 
</td>
</tr>
</table>

**รูปที่ 3.5** ตัวอย่างออนโทโลยีที่สร้างจากกฎความสัมพันธ์

**ตารางที่ 3.3** ตัวอย่างของความรู้ใหม่ที่ได้จากการให้เหตุผลเชิงตรรกะจากภาษาธรรมชาติ

| Entailment | New rules in ACE format |
|---|---|
| Every inv_nodes_equal_0-2 is a no-irradiat that is a no-node-caps. | If X is a n:inv_nodes_equal_0-2 and X is a n:no-irradiat<br>then X is a n:no-node-caps . |
| Every no-irradiat is an inv_nodes_equal_0-2 that is a no-node-caps. | If X is a n:no-irradiat and X is a n:inv_nodes_equal_0-2<br>then X is a n:no-node-caps . |
| Every no-node-caps is an inv_nodes_equal_0-2 that is a no-irradiat. | If X is a n:no-node-caps and X is a n:inv_nodes_equal_0-2<br>then X is a n:no-irradiat . |

ขั้นตอนการตรวจสอบความขัดแย้งของกฎความสัมพันธ์ที่ได้จากแหล่งความรู้ต่าง ๆ แสดงได้เป็นคำสั่งเทียมได้ดังรูปที่ 3.6 โดยจะรับข้อมูลเข้าเป็นกฎความสัมพันธ์ที่อยู่ในรูปแบบของภาษาธรรมชาติ แล้วนำแต่ละกฎความสัมพันธ์ไปสร้างเป็นออนโทโลยีสำหรับการนำไปใช้กับเครื่องมือตรวจสอบความขัดแย้ง หลังจากได้กฎความสัมพันธ์ที่อยู่ในรูปแบบของออนโทโลยีแล้วจะนำไปตรวจสอบความขัดแย้งด้วย FaCT++ Reasoner (Tsarkov and Ian, 2006) ซึ่งถ้าเกิดความขัดแย้งจะทำการลบกฎความสัมพันธ์นั้น ๆ โดยผลลัพธ์ที่ได้คือสามารถบอกได้ว่ากฎความสัมพันธ์ที่ได้จากการรวมกฎความสัมพันธ์จากแหล่งความรู้ต่าง ๆ นั้นขัดแย้งกันหรือไม่ และกฎความสัมพันธ์ใหม่ที่ได้

จากกฎความสัมพันธ์เดิมเพื่อนำไปเพิ่มเติมในส่วนของกฎความสัมพันธ์ที่ขาดหายไป สุดท้ายจะได้ฐานความรู้ของกฎความสัมพันธ์เพียงชุดเดียวที่มาจากการหากฎความสัมพันธ์แบบกระจายที่มีประสิทธิภาพใกล้เคียงกับการหากฎความแบบดั้งเดิม

**Algorithm Check_Inconsistency_with_Reasoner**

//Input: CACE, an association rule set in the form of natural language.

//Output: I, Inconsistency of Association Rules as true or false.

E, Association rules entailed from reasoner.

```
1.  I = true
2.  Ontology = ACE_views (CACE);
2.  While I = true{
3.        (I, E) = Reasoner (Ontology)
4.        If I = true
5.              Ontology = Remove_rules_inconsistent (CACE)
6.   }
7.  Return (I, E)
```

**รูปที่ 3.6** ขั้นตอนการตรวจสอบความขัดแย้งของกฎความสัมพันธ์ที่รวบรวมจากหลายแหล่ง

# บทที่ 4
# การทดสอบกลไกการค้นหาและรวมกฎความสัมพันธ์จากหลายแหล่ง

โครงการวิจัยนี้นำเสนอวิธีการทำเหมืองความสัมพันธ์แบบกระจาย ด้วยการออกแบบและพัฒนากลไกเชิงตรรกะเพื่อรวมกฎความสัมพันธ์ที่กระจายอยู่ในหลายฐานข้อมูล รวมถึงตรวจสอบความขัดแย้งของกฎและอนุมานกฎที่อาจจะขาดหายไป การทดสอบประสิทธิภาพของวิธีการที่เสนอขึ้นใหม่นั้น จะทดสอบประสิทธิภาพกลไกการค้นหาและรวมกฎความสัมพันธ์จากหลายแหล่งเปรียบเทียบกับการหากฎความสัมพันธ์แบบดั้งเดิมที่ข้อมูลทั้งหมดต้องถูกรวบรวมไว้ในแหล่งเดียว โดยจะพิจารณาจากจำนวนกฎความสัมพันธ์ที่ได้เป็นเกณฑ์ในการเปรียบเทียบ

## 4.1 ข้อมูลที่ใช้ในการทดสอบ

การทดสอบกลไกการค้นหาและรวมกฎความสัมพันธ์จากหลายแหล่ง จะใช้ข้อมูลมาตรฐานจาก UCI Machine Learning Repository ซึ่งเป็นข้อมูลผู้ป่วยโรคมะเร็งเต้านม (Breast Cancer) และข้อมูลผู้ป่วยโรคหัวใจ (Heart Disease) ค่าในทุกแอททริบิวต์ของข้อมูลทั้งสองชุดนี้ถูกแปลงให้เป็นประเภทข้อมูลเชิงกลุ่ม (categorical data type) เนื่องจากการทำเหมืองข้อมูลประเภทการค้นหาความสัมพันธ์ ใช้หลักการนับความถี่ของแต่ละค่าข้อมูลในทุกแอททริบิวต์ ประเภทของข้อมูลจึงต้องเป็นชนิดที่สามารถแจงนับได้

ข้อมูลผู้ป่วยโรคมะเร็งเต้านม เป็นข้อมูลการวินิจฉัยการเกิดซ้ำของมะเร็ง มีข้อมูลทั้งหมด 286 เรคคอร์ด แต่ละเรคคอร์ดประกอบด้วย 9 แอททริบิวต์ สามารถดาวน์โหลดได้จาก https://archive.ics.uci.edu/ml/datasets/Breast+Cancer รายละเอียดและความหมายของแต่ละแอททริบิวต์ในข้อมูลผู้ป่วยโรคมะเร็งเต้านมแสดงได้ดังตารางที่ 4.1

ข้อมูลผู้ป่วยโรคหัวใจ เป็นข้อมูลเกี่ยวกับคนไข้ที่มีความเสี่ยงเป็นโรคหัวใจ มีข้อมูลทั้งหมด 303 เรคคอร์ด แต่ละเรคคอร์ดมี 14 แอททริบิวต์ สามารถดาวน์โหลดข้อมูลได้จาก https://archive.ics.uci.edu/ml/datasets/heart+Disease รายละเอียดและความหมายของแต่ละแอททริบิวต์ในข้อมูลผู้ป่วยโรคหัวใจแสดงได้ดังตารางที่ 4.2

**ตารางที่ 4.1** รายละเอียดแอททริบิวต์ของข้อมูลผู้ป่วยโรคมะเร็งเต้านม

| ชื่อแอททริบิวต์ | ความหมายของแอททริบิวต์ | ค่าที่เป็นไปได้ |
|---|---|---|
| *age* | Age of patient (in the range 29-77) | {10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99} |
| menopause | Age of a woman at menopause | {lt40, ge40, premeno} |
| tumor-size | Size of the tumor in millimeters | {0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59} |
| inv-nodes | The number (range 0 - 39) of axillary lymph nodes that contain metastatic breast cancer visible on histological examination | {0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39} |
| node-caps | Does the cancer metastasize to a lymph node? | {yes, no} |
| deg-malig | Degree of malignant tumor | {1, 2, 3} |
| breast | Position of breast that cancer occurs | {left, right} |
| breast-quad | Quadrant on the cancerous breast | {left-up, left-low, right-up, right-low, central} |
| irradiat | Does the patient underwent the radiation treatment? | {yes, no} |
| Class | Does this cancer a recurrent one? | {no-recurrence-events, recurrence-events} |

**ตารางที่ 4.2** รายละเอียดแอททริบิวต์ของข้อมูลผู้ป่วยโรคหัวใจ

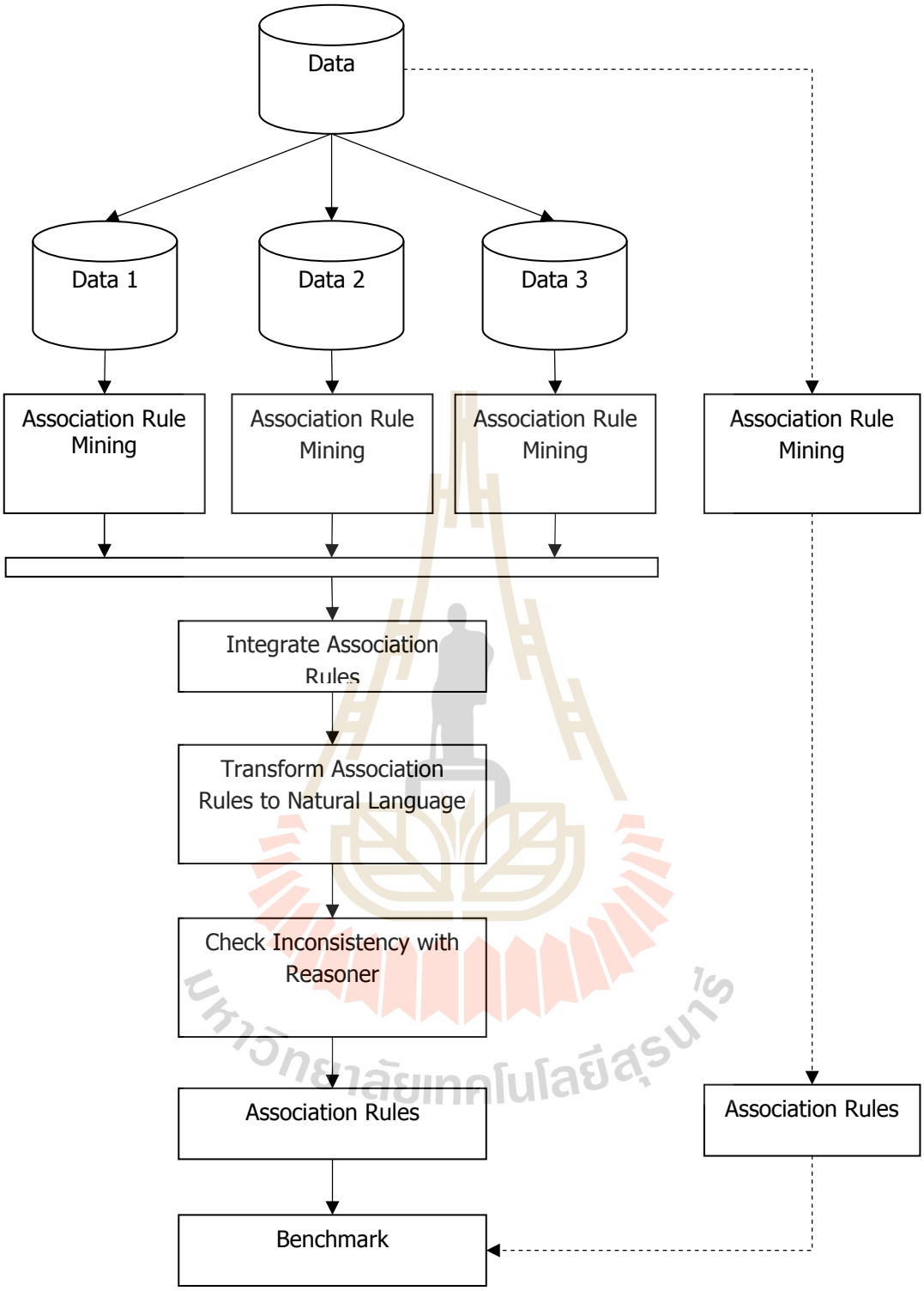
| ชื่อแอททริบิวต์ | ความหมายของแอททริบิวต์ | ค่าที่เป็นไปได้ |
|---|---|---|
| *age* | Age of patient (in the range 29-77) | {29-45, 46-60, 61-77} |
| *sex* | Sex of patient | {male, female} |
| *cp* | Chest pain type (4 types):<br>1 = typical angina, 2 = atypical angina,<br>3 = non-anginal pain , 4 =asymptom | {typ_angina,<br>atyp_angina,<br>non_anginal, asympt} |
| *trestbps* | Blood pressure when resting (in mm Hg),<br>values are in the range 94-200 | {94-129, 130-164,<br>165-200} |
| *chol* | Serum cholesterol in mg/dl, values are in the range 126-564 | {126-272, 273-418,<br>419-564} |
| *fbs* | Blood sugar when fasting (> 120 mg/dl) | {t, f} |
| *restecg* | Electrocardiographic results when resting:<br>0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria | {left_vent_hyper,<br>normal, abnormal} |
| *thalach* | Maximum heart rate achieved (in the range 71-202) | {71-114, 115-158,<br>159-202} |
| *exang* | Exercise induced angina | {yes, no} |
| *oldpeak* | ST depression induced by exercise relative to rest (values are in the range 0-62) | {1, 2, 3} |
| *slope* | The slope of the peak exercise for the ST depression segment | {down, flat, up} |
| *ca* | Number of major vessels (0-3) colored by fluoroscopy | {0, 1, 2, 3} |
| *thal* | Thalassemia (3 = normal; 6 = fixed defect; 7 = reversible defect) | {normal, fixed_defect,<br>reversible_defect } |
| *num* | Risk factor of heart disease:<br><50 = no disease,<br>>50_1 = presence of heart disease | {<50, >50_1} |

### 4.2 วิธีการทดสอบประสิทธิภาพกลไกการค้นหาและรวมกฎความสัมพันธ์

การทดสอบประสิทธิภาพกลไกการค้นหาและรวมกฎความสัมพันธ์จากหลายแหล่งจะเป็นการจำลองการหากฎความสัมพันธ์แบบกระจาย โดยแบ่งข้อมูลออกเป็นชุดย่อย ๆ หลายชุดเพื่อนำไปหากฎความสัมพันธ์ โดยที่กระบวนการของการทำงานบนข้อมูลแต่ละชุดย่อยจะไม่ขึ้นต่อกัน เกณฑ์การประเมินประสิทธิภาพ จะใช้จำนวนกฎความสัมพันธ์ที่ให้ผลลัพธ์ตรงกันกับการหากฎความสัมพันธ์แบบดั้งเดิมที่ข้อมูลทั้งหมดรวมอยู่ในแหล่งเดียวเป็นตัววัดประสิทธิภาพ

รูปที่ 4.1 แสดงแผนภาพที่เป็นแนวคิดของวิธีการทดสอบประสิทธิภาพกลไกการค้นหาและรวมกฎความสัมพันธ์จากหลายแหล่ง โดยในภาพจะแบ่งข้อมูลออกเป็น 3 ชุด คือ Data1 Data2 และ Data 3 แล้วนำข้อมูลทั้ง 3 ชุดไปหากฎความสัมพันธ์ ซึ่งในส่วนนี้การหากฎความสัมพันธ์ของข้อมูลแต่ละชุดจะไม่ขึ้นต่อกัน ผลลัพธ์ที่ได้คือชุดกฎความสัมพันธ์ของข้อมูลแต่ละชุด หลังจากนั้นจะดึงกฎความสัมพันธ์ที่ได้จากข้อมูลแต่ละชุดมารวมให้เป็นกฎความสัมพันธ์เพียงชุดเดียวด้วยกลไกการค้นหาและรวมกฎความสัมพันธ์จากหลายแหล่ง ผลลัพธ์สุดท้ายที่ได้คือชุดกฎความสัมพันธ์เพียงชุดเดียวที่ได้จากการรวมกฎความสัมพันธ์จากหลายแหล่ง ในการวัดประสิทธิภาพของกฎความสัมพันธ์ที่ได้นั้นจะเปรียบเทียบจากกฎความสัมพันธ์ที่ได้จากการหากฎความสัมพันธ์ด้วยข้อมูลดั้งเดิมที่ไม่ได้ถูกแบ่งออกเป็นชุด ๆ ซึ่งจะพิจารณาจากจำนวนกฎความสัมพันธ์ที่เหมือนกันระหว่างกฎความสัมพันธ์ที่ได้จากกลไกการค้นหาและรวมกฎความสัมพันธ์จากหลายแหล่ง เปรียบเทียบกับการหากฎความสัมพันธ์แบบดั้งเดิม

การทดสอบประสิทธิภาพกลไกการค้นหาและรวมกฎความสัมพันธ์จากหลายแหล่ง นอกจากจะใช้จำนวนกฎความสัมพันธ์ที่เหมือนกันกับการหากฎความสัมพันธ์แบบดั้งเดิมเป็นตัววัดประสิทธิภาพแล้ว ในขั้นตอนของการหากฎความสัมพันธ์นั้นจะเลือกใช้ค่าสนับสนุนที่แตกต่างกันออกไป เพื่อต้องการทดสอบว่าเมื่อใช้ค่าสนับสนุนที่แตกต่างกัน จะได้จำนวนกฎความสัมพันธ์ที่เหมือนกันกับการหากฎความสัมพันธ์แบบดั้งเดิมหรือไม่ โดยค่าสนับสนุนที่เลือกใช้ในการทดสอบได้แก่ ค่าสนับสนุนที่ 0.1 0.2 0.3 0.4 0.5 และ 0.6 ในกรณีที่ค่าสนับสนุนมีค่าตั้งแต่ 0.7 ขึ้นไปจะไม่ปรากฏกฎความสัมพันธ์




**รูปที่ 4.1** วิธีการทดสอบประสิทธิภาพกลไกการค้นหาและรวมกฎความสัมพันธ์จากหลายแหล่ง

### 4.3 ผลการทดสอบกลไกการค้นหากฎความสัมพันธ์ใหม่ด้วยวิธีการเชิงตรรกะ

ในการกำหนดค่าสนับสนุนขั้นต่ำเพื่อค้นหากฎความสัมพันธ์เป็น 0.1 กลไกไกการค้นหากฎความสัมพันธ์ใหม่ด้วยการให้เหตุผลเชิงตรรกะ สามารถค้นพบกฎความสัมพันธ์ใหม่จากข้อมูลผู้ป่วยโรคมะเร็งเต้านมจำนวน 8 กฎ และค้นพบกฎความสัมพันธ์ใหม่จากข้อมูลผู้ป่วยโรคหัวใจจำนวน 2 กฎ แสดงได้ดังตารางที่ 4.3

เมื่อเพิ่มเกณฑ์การกำหนดค่าสนับสนุนขั้นต่ำเพื่อค้นหากฎความสัมพันธ์เป็น 0.2 กลไกไกการค้นหากฎความสัมพันธ์ใหม่ด้วยการให้เหตุผลเชิงตรรกะ ค้นพบกฎความสัมพันธ์ใหม่ได้ลดลง โดยในข้อมูลผู้ป่วยโรคมะเร็งเต้านมได้กฎใหม่จำนวน 7 กฎ และค้นพบกฎใหม่จากข้อมูลผู้ป่วยโรคหัวใจได้จำนวน 1 กฎ แสดงได้ดังตารางที่ 4.4

เมื่อกำหนดค่าสนับสนุนขั้นต่ำเพื่อค้นหากฎความสัมพันธ์เป็น 0.3 กลไกไกการค้นหากฎความสัมพันธ์ใหม่ด้วยการให้เหตุผลเชิงตรรกะ ค้นพบกฎความสัมพันธ์ใหม่ได้ลดลง โดยในข้อมูลผู้ป่วยโรคมะเร็งเต้านมได้กฎใหม่จำนวน 5 กฎ และไม่สามารถค้นพบกฎใหม่จากข้อมูลผู้ป่วยโรคหัวใจได้ ซึ่งผลลัพธ์นี้จะเป็นเช่นเดียวกับกรณีกำหนดค่าสนับสนุนขั้นต่ำเป็น 0.4 และ 0.5 แสดงรายละเอียดของกฎใหม่ที่ค้นพบได้ดังตารางที่ 4.5

ในกรณีที่กำหนดค่าสนับสนุนขั้นต่ำเป็น 0.6 กลไกไกการค้นหากฎความสัมพันธ์ใหม่ด้วยการให้เหตุผลเชิงตรรกะค้นพบกฎความสัมพันธ์ใหม่ได้ลดลง โดยในข้อมูลผู้ป่วยโรคมะเร็งเต้านมได้กฎใหม่จำนวน 3 กฎ และไม่สามารถค้นพบกฎใหม่จากข้อมูลผู้ป่วยโรคหัวใจได้ แสดงรายละเอียดของกฎใหม่ที่ค้นพบได้ดังตารางที่ 4.6

**ตารางที่ 4.3** กฎความสัมพันธ์ที่ค้นพบได้ใหม่จากวิธีการเชิงตรรกะและใช้ค่าสนับสนุนขั้นต่ำ 0.1

| Result / Dataset | Entailment | If-Then Rules |
|---|---|---|
| **Breast Cancer** | Every age_equal_50-59 is an inv-nodes_equal_0-2 . | If age=50-59<br>Then inv-nodes=0-2 |
| | Every inv-nodes_equal_0-2 is an irradiat_equal_no that is a node-caps_equal_no . | If inv-nodes=0-2 and irradiat=no<br>Then node-caps=no |
| | Every irradiat_equal_no is an inv-nodes_equal_0-2 that is a node-caps_equal_no . | If irradiat=no and inv-nodes=0-2<br>Then node-caps=no |
| | Every node-caps_equal_no is an inv-nodes_equal_0-2 that is an irradiat_equal_no . | If node-caps=no and inv-nodes=0-2<br>Then irradiat=no |
| | Every deg-malig_equal_1 is a Class_equal_no-recurrence-events . | If deg-malig=1<br>Then Class=no-recurrence-events |
| | Every Class_equal_no-recurrence-events is an inv-nodes_equal_0-2 . | If Class=no-recurrence-events<br>Then inv-nodes=0-2 |
| | Every age_equal_50-59 is a node-caps_equal_no . | If age=50-59<br>Then node-caps=no |
| | Every Class_equal_no-recurrence-events is an irradiat_equal_no . | If Class=no-recurrence-events<br>Then irradiat=no |
| **Heart Disease** | Every sex_equal_female is a thal_equal_normal. | If sex=female<br>Then thal=normal |
| | Every cp_equal_atyp_angina is a fbs_equal_f . | If cp=atyp_angina<br>Then fbs=f |

**ตารางที่ 4.4** กฎความสัมพันธ์ที่ค้นพบได้ใหม่จากวิธีการเชิงตรรกะและใช้ค่าสนับสนุนขั้นต่ำ 0.2

| Result / Dataset | Entailment | If-Then Rules |
|---|---|---|
| **Breast Cancer** | Every age_equal_50-59 is an inv-nodes_equal_0-2 . | If age=50-59<br>Then inv-nodes=0-2 |
| | Every inv-nodes_equal_0-2 is an irradiat_equal_no that is a node-caps_equal_no . | If inv-nodes=0-2 and irradiat=no<br>Then node-caps=no |
| | Every irradiat_equal_no is an inv-nodes_equal_0-2 that is a node-caps_equal_no . | If irradiat=no and inv-nodes=0-2<br>Then node-caps=no |
| | Every node-caps_equal_no is an inv-nodes_equal_0-2 that is an irradiat_equal_no . | If node-caps=no and inv-nodes=0-2<br>Then irradiat=no |
| | Every Class_equal_no-recurrence-events is an inv-nodes_equal_0-2 . | If Class=no-recurrence-events<br>Then inv-nodes=0-2 |
| | Every age_equal_50-59 is a node-caps_equal_no . | If age=50-59<br>Then node-caps=no |
| | Every Class_equal_no-recurrence-events is an irradiat_equal_no . | If Class=no-recurrence-events<br>Then irradiat=no |
| **Heart Disease** | Every num_equal_less_than_50 is a thal_equal_normal . | If num=< 50<br>Then thal=normal |

**ตารางที่ 4.5** กฎความสัมพันธ์ที่ค้นพบได้ใหม่จากวิธีการเชิงตรรกะและใช้ค่าสนับสนุนขั้นต่ำ 0.3-0.5

| Result / Dataset | Entailment | If-Then Rules |
|---|---|---|
| **Breast Cancer** | Every inv-nodes_equal_0-2 is an irradiat_equal_no that is a node-caps_equal_no . | If inv-nodes=0-2 and irradiat=no Then node-caps=no |
| | Every irradiat_equal_no is an inv-nodes_equal_0-2 that is a node-caps_equal_no . | If irradiat=no and inv-nodes=0-2 Then node-caps=no |
| | Every node-caps_equal_no is an inv-nodes_equal_0-2 that is an irradiat_equal_no . | If node-caps=no and inv-nodes=0-2 Then irradiat=no |
| | Every Class_equal_no-recurrence-events is an inv-nodes_equal_0-2 . | If Class=no-recurrence-events Then inv-nodes=0-2 |
| | Every Class_equal_no-recurrence-events is an irradiat_equal_no . | If Class=no-recurrence-events Then irradiat=no |
| **Heart Disease** | - | - |

**ตารางที่ 4.6** กฎความสัมพันธ์ที่ค้นพบได้ใหม่จากวิธีการเชิงตรรกะและใช้ค่าสนับสนุนขั้นต่ำ 0.6

| Result / Dataset | Entailment | If-Then Rules |
|---|---|---|
| **Breast Cancer** | Every inv-nodes_equal_0-2 is an irradiat_equal_no that is a node-caps_equal_no . | If inv-nodes=0-2 and irradiat=no Then node-caps=no |
| | Every irradiat_equal_no is an inv-nodes_equal_0-2 that is a node-caps_equal_no . | If irradiat=no and inv-nodes=0-2 Then node-caps=no |
| | Every node-caps_equal_no is an inv-nodes_equal_0-2 that is an irradiat_equal_no . | If node-caps=no and inv-nodes=0-2 Then irradiat=no |
| **Heart Disease** | - | - |

### 4.4 ผลการทดสอบกลไกการค้นหาและรวมกฎความสัมพันธ์ด้วยค่าสนับสนุนที่แตกต่างกัน

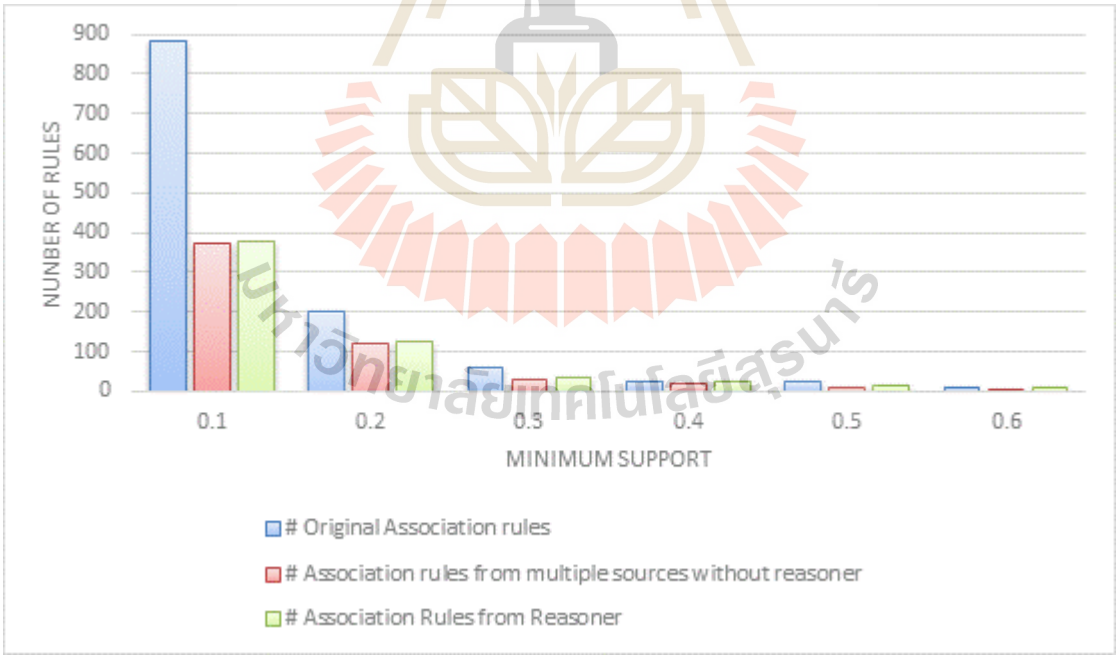
การทดสอบการหากฎความสัมพันธ์แบบกระจายนั้นจะเลือกใช้ค่าสนับสนุนที่แตกต่างกัน เพื่อต้องการทดสอบจำนวนกฎความสัมพันธ์ที่ได้ว่าจะตรงกันหรือใกล้เคียงกับการหากฎความสัมพันธ์แบบดั้งเดิมที่ข้อมูลรวมอยู่ในแหล่งเดียวหรือไม่

ผลการทดสอบกับข้อมูลผู้ป่วยโรคมะเร็งเต้านมสรุปได้ดังตารางที่ 4.7 ซึ่งจากผลการทดสอบจะสังเกตได้ว่าเปอร์เซ็นต์จำนวนกฎความสัมพันธ์ที่ถูกปรับปรุงด้วยการอนุมานความรู้ใหม่นั้นสามารถช่วยเพิ่มประสิทธิภาพให้กับการหากฎความสัมพันธ์จากข้อมูลหลายแหล่งได้อย่างชัดเจนเมื่อมีค่าสนับสนุนที่มาก (ดังรูปที่ 4.2).

ในผลการทดสอบการหากฎความสัมพันธ์จากข้อมูลผู้ป่วยโรคหัวใจ จะสังเกตได้ว่าเปอร์เซ็นต์จำนวนกฎความสัมพันธ์ที่ถูกปรับปรุงด้วยการอนุมานความรู้ใหม่ที่ค่าสนับสนุนที่น้อยนั้นสามารถช่วยเพิ่มประสิทธิภาพได้ในระดับหนึ่ง แต่เมื่อเพิ่มค่าสนับสนุนจำนวนของกฎความสัมพันธ์ที่ได้มีจำนวนใกล้เคียงกับการหากฎความสัมพันธ์แบบดั้งเดิม ทำให้เปอร์เซ็นต์จำนวนกฎความสัมพันธ์ที่ถูกปรับปรุงด้วยการค้นหากฎความสัมพันธ์ใหม่มีค่าเป็น 0.00 % ดังปรากฏในตารางที่ 4.8 และแสดงภาพกราฟเปรียบเทียบได้ดังรูปที่ 4.3

**ตารางที่ 4.7** เปรียบเทียบจำนวนกฎความสัมพันธ์ระหว่างการหากฎความสัมพันธ์แบบดั้งเดิมและการหากฎความสัมพันธ์จากข้อมูลหลายแหล่งในข้อมูลผู้ป่วยโรคมะเร็งเต้านม

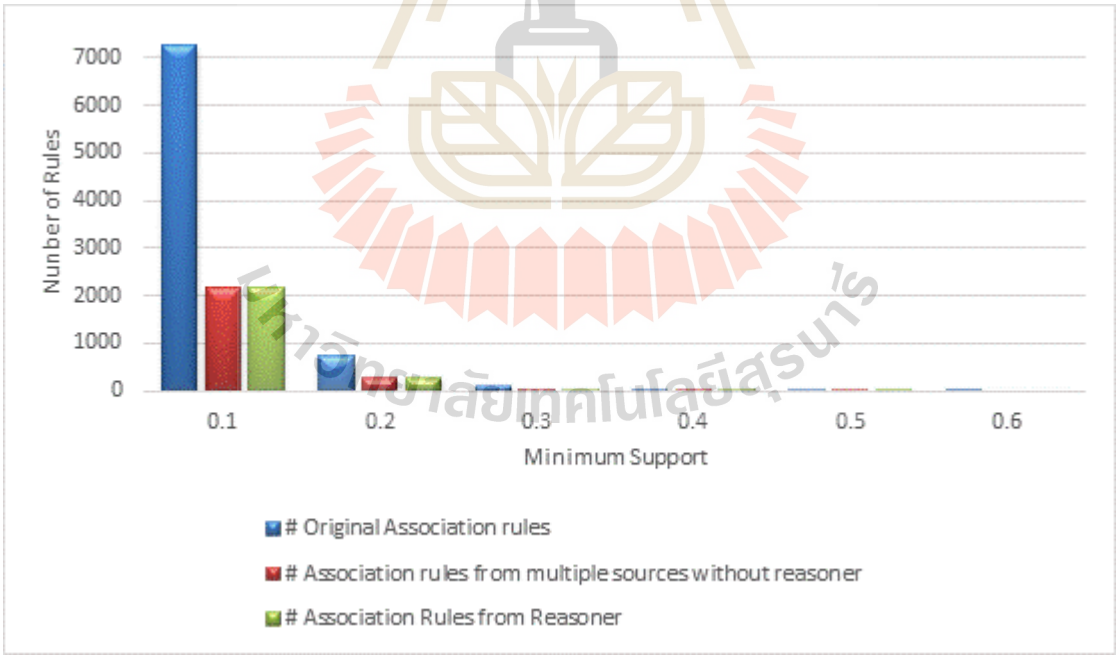
| Support value | # Original Association rules | # Association rules from multiple sources without reasoner | # Association Rules from Reasoner | % Improvement by Reasoner |
|---|---|---|---|---|
| **0.1** | 883 | 372 | 380 | 2.15% |
| **0.2** | 203 | 119 | 126 | 5.88% |
| **0.3** | 62 | 30 | 35 | 16.67% |
| **0.4** | 24 | 19 | 24 | 26.32% |
| **0.5** | 24 | 10 | 15 | 50.00% |
| **0.6** | 9 | 6 | 9 | 50.00% |




**รูปที่ 4.2** กราฟปรียบเทียบจำนวนกฎความสัมพันธ์ระหว่างการหากฎความสัมพันธ์แบบดั้งเดิมและการหากฎความสัมพันธ์จากข้อมูลหลายแหล่งในข้อมูลผู้ป่วยโรคมะเร็งเต้านม

**ตารางที่ 4.8** เปรียบเทียบจำนวนกฎความสัมพันธ์ระหว่างการหากฎความสัมพันธ์แบบดั้งเดิมและการหากฎความสัมพันธ์จากข้อมูลหลายแหล่งในข้อมูลผู้ป่วยโรคหัวใจ

| Support value | # Original Association rules | # Association rules from multiple sources without reasoner | # Association Rules from Reasoner | % Improvement by Reasoner |
|---|---|---|---|---|
| **0.1** | 7245 | 2158 | 2160 | 0.09% |
| **0.2** | 726 | 282 | 283 | 0.35% |
| **0.3** | 118 | 42 | 42 | 0.00% |
| **0.4** | 18 | 10 | 10 | 0.00% |
| **0.5** | 4 | 4 | 4 | 0.00% |
| **0.6** | 1 | 0 | 0 | 0.00% |




**รูปที่ 4.3** กราฟปรียบเทียบจำนวนกฎความสัมพันธ์ระหว่างการหากฎความสัมพันธ์แบบดั้งเดิมและการหากฎความสัมพันธ์จากข้อมูลหลายแหล่งในข้อมูลผู้ป่วยโรคหัวใจ

### 4.5 ผลการทดสอบความถูกต้องของการค้นหากฎความสัมพันธ์จากหลายแหล่ง

การทดสอบวิธีการค้นหากฎความสัมพันธ์จากฐานข้อมูลแบบกระจาย นอกจากจะพิจารณาที่จำนวนของกฎความสัมพันธ์ที่ค้นพบ จำเป็นต้องพิจารณาในด้านความถูกต้องของกฎความสัมพันธ์ โดยใช้เกณฑ์เปรียบเทียบกับวิธีการค้นหากฎความสัมพันธ์แบบดั้งเดิมที่ข้อมูลรวมอยู่ในฐานข้อมูลเดียว ผลการทดสอบกับข้อมูลข้อมูลผู้ป่วยโรคมะเร็งเต้านมที่ค่าสนับสนุนขั้นต่ำ 0.4 แสดงได้ดังตารางที่ 4.9 และผลการทดสอบกับข้อมูลผู้ป่วยโรคหัวใจด้วยค่าสนับสนุนที่ 0.4 แสดงได้ดังตารางที่ 4.10 จากผลการทดสอบจะเห็นว่าในข้อมูลผู้ป่วยโรคมะเร็งเต้านมให้ความถูกต้องของกฎความสัมพันธ์คิดเป็น 87.50% และข้อมูลผู้ป่วยโรคหัวใจให้ความถูกต้องของกฎความสัมพันธ์คิดเป็น 100%

**ตารางที่ 4.9** เปรียบเทียบความถูกต้องระหว่างกฎความสัมพันธ์แบบดั้งเดิมและกฎความสัมพันธ์จากข้อมูลหลายแหล่งด้วยค่าสนับสนุน 0.4 ในข้อมูลผู้ป่วยโรคมะเร็งเต้านม

| กฎความสัมพันธ์แบบดั้งเดิม | กฎความสัมพันธ์จากหลายแหล่ง |
|---|---|
| {Class=no-recurrence-events}=>{inv-nodes=0-2} | {Class=no-recurrence-events}=>{inv-nodes=0-2} |
| {Class=no-recurrence-events}=>{irradiat=no} | {Class=no-recurrence-events}=>{irradiat=no} |
| {Class=no-recurrence-events}=>{node-caps=no} | {Class=no-recurrence-events}=>{node-caps=no} |
| {inv-nodes=0-2}=>{irradiat=no} | {inv-nodes=0-2}=>{irradiat=no} |
| {irradiat=no}=>{inv-nodes=0-2} | {irradiat=no}=>{inv-nodes=0-2} |
| {inv-nodes=0-2}=>{node-caps=no} | {inv-nodes=0-2}=>{node-caps=no} |
| {node-caps=no}=>{inv-nodes=0-2} | {node-caps=no}=>{inv-nodes=0-2} |
| {irradiat=no}=>{node-caps=no} | {irradiat=no}=>{node-caps=no} |
| {node-caps=no}=>{irradiat=no} | {node-caps=no}=>{irradiat=no} |
| {inv-nodes=0-2,Class=no-recurrence-events}=>{irradiat=no} | {inv-nodes=0-2,Class=no-recurrence-events}=>{irradiat=no} |
| {irradiat=no,Class=no-recurrence-events}=>{inv-nodes=0-2} | {irradiat=no,Class=no-recurrence-events}=>{inv-nodes=0-2} |
| {inv-nodes=0-2,irradiat=no}=>{Class=no-recurrence-events} | - |

**ตารางที่ 4.9** เปรียบเทียบความถูกต้องระหว่างกฎความสัมพันธ์แบบดั้งเดิมและกฎความสัมพันธ์จากข้อมูลหลายแหล่งด้วยค่าสนับสนุน 0.4 ในข้อมูลผู้ป่วยโรคหัวใจ (ต่อ)

| กฎความสัมพันธ์แบบดั้งเดิม | กฎความสัมพันธ์จากหลายแหล่ง |
|---|---|
| {inv-nodes=0-2,Class=no-recurrence-events}=>{node-caps=no} | {inv-nodes=0-2,Class=no-recurrence-events}=>{node-caps=no} |
| {node-caps=no,Class=no-recurrence-events}=>{inv-nodes=0-2} | {node-caps=no,Class=no-recurrence-events}=>{inv-nodes=0-2} |
| {irradiat=no,Class=no-recurrence-events}=>{node-caps=no} | {irradiat=no,Class=no-recurrence-events}=>{node-caps=no} |
| {node-caps=no,Class=no-recurrence-events}=>{irradiat=no} | {node-caps=no,Class=no-recurrence-events}=>{irradiat=no} |
| {node-caps=no,irradiat=no}=>{Class=no-recurrence-events} | - |
| {inv-nodes=0-2,irradiat=no}=>{node-caps=no} | {inv-nodes=0-2,irradiat=no}=>{node-caps=no} |
| {inv-nodes=0-2,node-caps=no}=>{irradiat=no} | {inv-nodes=0-2,node-caps=no}=>{irradiat=no} |
| {node-caps=no,irradiat=no}=>{inv-nodes=0-2} | {node-caps=no,irradiat=no}=>{inv-nodes=0-2} |
| {inv-nodes=0-2,irradiat=no,Class=no-recurrence-events}=>{node-caps=no} | {inv-nodes=0-2,irradiat=no,Class=no-recurrence-events}=>{node-caps=no} |
| {inv-nodes=0-2,node-caps=no,Class=no-recurrence-events}=>{irradiat=no} | {inv-nodes=0-2,node-caps=no,Class=no-recurrence-events}=>{irradiat=no} |
| {node-caps=no,irradiat=no,Class=no-recurrence-events}=>{inv-nodes=0-2} | {node-caps=no,irradiat=no,Class=no-recurrence-events}=>{inv-nodes=0-2} |
| {inv-nodes=0-2,node-caps=no,irradiat=no}=>{Class=no-recurrence-events} | - |

**ตารางที่ 4.10** เปรียบเทียบความถูกต้องระหว่างกฎความสัมพันธ์แบบดั้งเดิมและกฎความสัมพันธ์จากข้อมูลหลายแหล่งด้วยค่าสนับสนุน 0.4 ในข้อมูลผู้ป่วยโรคหัวใจ

| กฎความสัมพันธ์แบบดั้งเดิม | กฎความสัมพันธ์จากหลายแหล่ง |
|---|---|
| {slope=up}=>{fbs=f} | - |
| {cp=asympt}=>{fbs=f} | - |
| {restecg=normal}=>{fbs=f} | - |
| {thalach=115-158}=>{fbs=f} | - |
| {trestbps=129-164}=>{fbs=f} | {trestbps=129-164}=>{fbs=f} |
| {num=<50}=>{exang=no} | {num=<50}=>{exang=no} |
| {num=<50}=>{fbs=f} | {num=<50}=>{fbs=f} |
| {oldpeak=0}=>{fbs=f} | {oldpeak=0}=>{fbs=f} |
| {thal=normal}=>{exang=no} | - |
| {thal=normal}=>{fbs=f} | {thal=normal}=>{fbs=f} |
| {age=46-60}=>{fbs=f} | - |
| {ca=0.0}=>{fbs=f} | {ca=0.0}=>{fbs=f} |
| {exang=no}=>{fbs=f} | {exang=no}=>{fbs=f} |
| {sex=male}=>{fbs=f} | {sex=male}=>{fbs=f} |
| {chol=126-272}=>{fbs=f} | {chol=126-272}=>{fbs=f} |
| {chol=126-272,ca=0.0}=>{fbs=f} | - |
| {chol=126-272,exang=no}=>{fbs=f} | {chol=126-272,exang=no}=>{fbs=f} |
| {sex=male,chol=126-272}=>{fbs=f} | - |



### 4.6 อภิปรายผล

จากผลการทดสอบประสิทธิภาพการหากฎความสัมพันธ์จากข้อมูลผู้ป่วยโรคมะเร็งเต้านม และข้อมูลผู้ป่วยโรคหัวใจด้วยกลไกการค้นหาและรวมกฎความสัมพันธ์จากหลายแหล่งด้วยค่าสนับสนุนที่แตกต่างกัน ได้แก่ ค่าสนับสนุนที่ 0.1 0.2 0.3 0.4 0.5 และ 0.6 สามารถสรุปผลการทดสอบได้ดังนี้

1. การเปรียบเทียบจำนวนกฎความสัมพันธ์จากค่าสนับสนุนที่แตกต่างกัน จะเห็นได้ว่าเมื่อหากฎความสัมพันธ์ด้วยเกณฑ์ค่าสนับสนุนที่ต่ำ จะให้กฎความสัมพันธ์จากกลไกการค้นหาและรวมกฎความสัมพันธ์จากหลายแหล่งมีจำนวนที่น้อยเมื่อเทียบกับการหากฎความสัมพันธ์แบบดั้งเดิมที่

ข้อมูลอยู่ในแหล่งเดียว แต่เมื่อหากฎความสัมพันธ์ด้วยค่าสนับสนุนที่มากขึ้น จะให้จำนวนกฎความสัมพันธ์ที่ใกล้เคียงกับการหากฎความสัมพันธ์แบบดั้งเดิม ดังรูปที่ 4.2 และ 4.3

2. กฎความสัมพันธ์ใหม่ที่ได้จากกฎความสัมพันธ์เดิมเมื่อใช้ค่าสนับสนุนที่แตกต่างกัน จะเห็นได้ว่าในข้อมูลผู้ป่วยโรคมะเร็งเต้านมและข้อมูลผู้ป่วยโรคหัวใจ สามารถอนุมานกฎความสัมพันธ์ใหม่จากกฎความสัมพันธ์เดิมนั้นได้ ซึ่งเมื่อพิจารณาจากค่าสนับสนุนที่แตกต่างกันพบว่าจำนวนกฎความสัมพันธ์ใหม่ที่ได้ลดลงเมื่อใช้ค่าสนับสนุนที่มากขึ้น

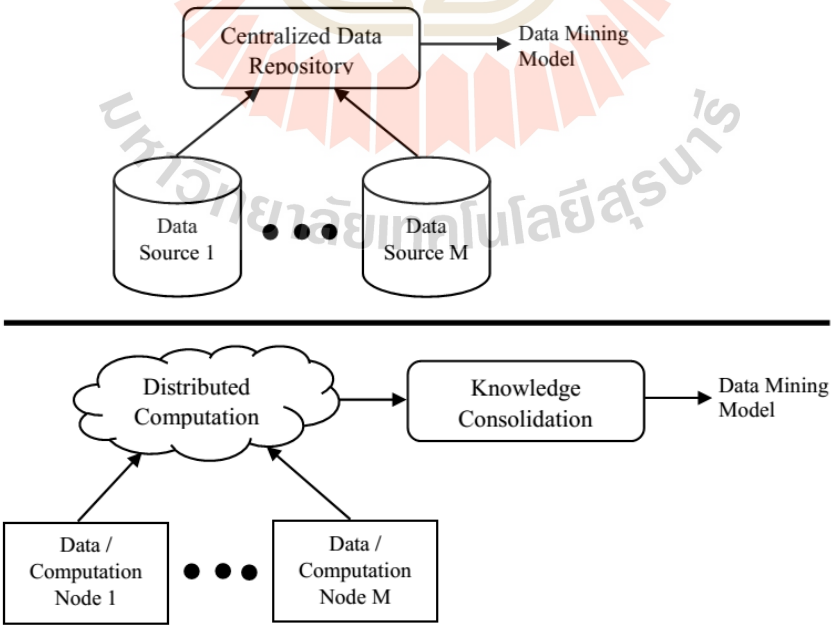
3. จากผลการทดสอบวิธีการค้นหากฎความสัมพันธ์ใหม่ที่ใช้กลไกการอนุมานได้จากกฎความสัมพันธ์เดิม สามารถนำกฎความสัมพันธ์ใหม่ที่ได้นั้นไปเพิ่มเติมในส่วนของกฎความสัมพันธ์ที่ขาดหายไปได้ ดังนั้นความสามารถในส่วนนี้สามารถช่วยเพิ่มประสิทธิภาพให้กับการหากฎความสัมพันธ์ด้วยกลไกการค้นหาและรวมกฎความสัมพันธ์จากหลายแหล่งให้มีจำนวนที่เหมือนหรือใกล้เคียงกับการการหากฎความสัมพันธ์แบบดั้งเดิมมากที่สุด

4. การทดสอบความถูกต้องของการค้นหากฎความสัมพันธ์จากข้อมูลหลายแหล่ง เมื่อเปรียบเทียบกับวิธีการค้นหากฎความสัมพันธ์แบบดั้งเดิมที่ข้อมูลอยู่ในแหล่งเดียว จะเห็นได้ว่ากฎความสัมพันธ์จากข้อมูลหลายแหล่งให้ความถูกต้องของกฎความสัมพันธ์ที่สูง แต่ความถูกต้องยังไม่ถึง 100% เนื่องจากเป็นกฎความสัมพันธ์ที่ได้จากการอนุมานความรู้ใหม่ ซึ่งมีบางกฎที่ไม่เหมือนกับกฎความสัมพันธ์แบบดั้งเดิม

# บทที่ 5
# บทสรุป

## 5.1 สรุปผลการวิจัย

ปัจจุบันการจัดเก็บข้อมูลในรูปแบบอิเล็กทรอนิกส์มีความสำคัญต่อองค์กรและหน่วยงานในยุคใหม่ เนื่องจากสามารถที่จะสกัดความรู้จากข้อมูลเหล่านั้นเพื่อนำความรู้ที่ได้ไปวางแผนพัฒนาองค์กรได้อย่างมีประสิทธิภาพ การสกัดความรู้ด้วยเทคนิคการทำเหมืองข้อมูลถูกนำมาใช้อย่างแพร่หลายโดยเฉพาะในองค์กรภาคธุรกิจ โดยกระบวนการสกัดความรู้มักจะกระทำกับฐานข้อมูลที่มีขนาดใหญ่เพียงแหล่งเดียวเพื่อค้นหาความรู้ใหม่ในลักษณะต่าง ๆ เช่น รูปแบบโดยรวมของข้อมูล แนวโน้มของข้อมูล หรือความสัมพันธ์ภายในกลุ่มข้อมูล เป็นต้น ความรู้ที่ได้จะถูกนำไปประยุกต์ใช้เพื่อวิเคราะห์ข้อมูลหลากหลายด้าน เช่น ใช้วิเคราะห์ข้อมูลการตลาดเพื่อคาดการณ์แนวโน้มการเติบโตของยอดขายสินค้า เป็นต้น แต่การประยุกต์เทคโนโลยีเหมืองข้อมูลกับหน่วยงานสาธารณสุขเป็นสิ่งที่ทำได้ยาก เนื่องจากการจะรวมข้อมูลจากสถานพยาบาลที่กระจายอยู่ทั่วประเทศเพื่อให้ได้ฐานข้อมูลขนาดใหญ่เพียงแหล่งเดียวไม่สามารถปฏิบัติได้จริง ทำให้การทำเหมืองข้อมูลสำหรับงานด้านสาธารณสุขที่เป็นลักษณะฐานข้อมูลแบบกระจาย จำเป็นต้องมีการเปลี่ยนรูปแบบการวิเคราะห์ข้อมูลจากเดิมที่เป็นลักษณะรวมฐานให้เป็นลักษณะกระจาย แสดงภาพเปรียบเทียบได้ดังรูปที่ 5.1 ที่ภาพบนเป็นลักษณะการรวมฐานข้อมูลและภาพล่างเป็นการทำเหมืองข้อมูลแบบกระจาย




**รูปที่ 5.1** แนวคิดการทำเหมืองข้อมูลแบบรวมฐานเปรียบเทียบกับแบบกระจาย

งานวิจัยนี้เน้นการทำเหมืองข้อมูลประเภทการค้นหาความสัมพันธ์ (association mining) โดยปรับปรุงกระบวนการค้นหาความสัมพันธ์ให้เป็นแบบกระจายด้วยการออกแบบอัลกอริทึมและพัฒนาโปรแกรมเพื่อค้นหาและรวมกฎความสัมพันธ์จากหลายแหล่ง ซึ่งจะทำให้ได้กฎความสัมพันธ์ที่มีประสิทธิภาพใกล้เคียงกับการหากฎความสัมพันธ์แบบดั้งเดิมที่ข้อมูลรวมฐานอยู่ในแหล่งเดียว การประเมินประสิทธิภาพพิจารณาจากจำนวนกฎความสัมพันธ์ที่เหมือนกันกับกฎความสัมพันธ์ที่ได้จากการหากฎความสัมพันธ์แบบดั้งเดิม ดังนั้นงานวิจัยนี้ได้พัฒนาและออกแบบอัลกอริทึมใหม่ในส่วนต่าง ๆ ดังนี้

(1) ส่วนของการรวมกฎความสัมพันธ์ โดยจะค้นหากฎความสัมพันธ์ที่ปรากฏในทุก ๆ แหล่งความรู้เพื่อนำมาเก็บไว้ในแหล่งความรู้เพียงแหล่งเดียว

(2) ส่วนของการแปลงกฎความสัมพันธ์ที่อยู่ในรูปแบบทั่วไป ให้อยู่ในรูปแบบของภาษาธรรมชาติ เพื่อนำกฎความสัมพันธ์ไปแทนความรู้สำหรับการนำไปสร้างเป็นออนโทโลยี

(3) ส่วนของการสร้างกฎใหม่และการตรวจสอบความขัดแย้งของกฎ ด้วยการนำกฎความสัมพันธ์ที่ถูกแทนความรู้แล้วนำไปสร้างเป็นออนโทโลยีสำหรับนำไปตรวจสอบความขัดแย้ง และอนุมานกฎความสัมพันธ์ใหม่จากกฎความสัมพันธ์เดิม โดยความรู้ใหม่ที่ได้สามารถนำไปเพิ่มเติมในส่วนของกฎความสัมพันธ์ที่ขาดหายไปได้

ผลการทดสอบประสิทธิภาพกลไกการค้นหาและรวมกฎความสัมพันธ์จากหลายแหล่งโดยใช้ข้อมูลผู้ป่วยโรคมะเร็งเต้านมและข้อมูลผู้ป่วยโรคหัวใจด้วยค่าสนับสนุนที่แตกต่างกัน พบว่าเมื่อใช้ค่าสนับสนุนที่มีค่าต่ำ ข้อมูลทั้งสองชุดจะให้จำนวนกฎความสัมพันธ์ที่แตกต่างจากการหากฎความสัมพันธ์แบบดั้งเดิม แต่เมื่อเพิ่มเกณฑ์ค่าสนับสนุนให้สูงขึ้นพบว่าข้อมูลทั้งสองชุดจะให้จำนวนกฎความสัมพันธ์ที่ใกล้เคียงหรือเทียบเท่ากับการหากฎความสัมพันธ์แบบดั้งเดิม

ในขั้นตอนของการตรวจสอบความขัดแย้งและอนุมานความรู้ใหม่ที่ได้จากความรู้เดิมนั้น การหากฎความสัมพันธ์ด้วยกลไกการค้นหาและรวมกฎความสัมพันธ์จากหลายแหล่งด้วยข้อมูลผู้ป่วยโรคมะเร็งเต้านมและข้อมูลผู้ป่วยโรคหัวใจ พบว่าสามารถนำกฎความสัมพันธ์ใหม่ที่ได้ไปเพิ่มเติมในส่วนของกฎความสัมพันธ์ที่ขาดหายไปได้

ดังนั้นจึงสามารถสรุปได้ว่างานวิจัยที่ได้เสนอกลไกการค้นหาและรวมกฎความสัมพันธ์จากหลายแหล่ง เมื่อกำหนดเกณฑ์สนับสนุนให้มีค่าค่อนข้างสูงสามารถช่วยค้นหาและรวมกฎความสัมพันธ์ที่ได้จากข้อมูลที่กระจายตัวกันอยู่หรือข้อมูลที่มีขนาดใหญ่ได้ ผลลัพธ์ที่ได้เป็นเซตของกฎความสัมพันธ์ที่มีจำนวนและรายละเอียดในแต่ละกฎใกล้เคียงกับการหากฎความสัมพันธ์แบบดั้งเดิม

## 5.2 ข้อจำกัดและข้อเสนอแนะ

กลไกการค้นหาและรวมกฎความสัมพันธ์จากหลายแหล่งนั้นเป็นการทำงานแบบกึ่งอัตโนมัติ ซึ่งในขั้นตอนของการแทนความรู้ด้วยภาษาธรรมชาติสำหรับใช้ในขั้นตอนการตรวจสอบความขัดแย้งของกฎความสัมพันธ์นั้นผู้ใช้จะต้องเป็นผู้ดำเนินการเอง โปรแกรมที่พัฒนาขึ้นยังไม่สามารถทำงานในส่วนนี้ได้ และเมื่อใช้ค่าสนับสนุนระดับต่ำในการหากฎความสัมพันธ์ กลไกการค้นหาและรวมกฎความสัมพันธ์จากหลายแหล่งที่พัฒนาขึ้นนี้ยังให้จำนวนของกฎความสัมพันธ์ที่น้อยเมื่อเทียบกับการหากฎความสัมพันธ์แบบดั้งเดิม

อย่างไรก็ตามกลไกการค้นหาและรวมกฎความสัมพันธ์จากหลายแหล่ง ให้ประสิทธิภาพที่ใกล้เคียงกับการหากฎความสัมพันธ์จากข้อมูลเพียงแหล่งเดียว จึงสามารถใช้เพื่อรวมกฎความสัมพันธ์ที่ได้จากการหากฎความสัมพันธ์ในแต่ละแหล่งข้อมูล โดยไม่จำเป็นต้องรวมข้อมูลจากแหล่งต่าง ๆ ให้อยู่ในแหล่งข้อมูลเพียงแหล่งเดียว ซึ่งในกรณีของการรวมแหล่งข้อมูลจำเป็นต้องใช้คอมพิวเตอร์ที่มีประสิทธิภาพสูงสำหรับการนำมาประมวลผล แต่ด้วยกลไกที่พัฒนาขึ้นนี้สามารถนำไปประยุกต์ใช้กับงานหากฎความสัมพันธ์จากข้อมูลทางการแพทย์ที่กระจายอยู่ตามคลินิก หรือโรงพยาบาลต่าง ๆ เป็นต้น

ข้อเสนอแนะสำหรับการพัฒนาต่อยอดกระบวนการค้นหาและรวมกฎความสัมพันธ์จากหลายแหล่ง ควรมีการปรับปรุงอัลกอริทึมให้สามารถทำงานแบบอัตโนมัติเพื่อลดความผิดพลาดของผู้ใช้และช่วยลดเวลาการทำงาน ในด้านการเพิ่มความถูกต้องของจำนวนกฎที่รวบรวมจากหลายแหล่งอาจปรับปรุงด้วยการใช้เครื่องมือหรืออัลกอริทึมอื่นที่สามารถอนุมานความรู้ได้ดีกว่าเครื่องมือที่ใช้ในงานวิจัยนี้

# บรรณานุกรม


Agrawal R., Srikant R. (1994), Fast algorithms for mining association rules in large databases, Proceedings of the 20th International Conference on Very Large Data Bases, pp.487-499.

Atligan Y., Dogan F. (2008). Data mining on distributed medical databases: Recent trends and future directions. Proceedings of the 1st ICST Conference on IT Revolution, pp. 216-224.

Banaee H., Ahmed M.U., Loutfi A. (2013). Data mining for wearable sensors in health monitoring systems: A review of recent trends and challenges. Sensors, 13: 17472-17500.

Banerjee S., Cafri G., Isaacs A.J., Graves S., Paxton E., Marinac-Dabic D., Sedrakyan A. (2014). A distributed health data network analysis of survival outcomes: The international consortium of orthopaedic registries perspective. The Journal of Bone & Joint Surgery, 96(Supplement 1): 7-11.

Coorevits P., Sundgren M., Klein G.O., Bahr A., Claerhout B., Daniel C., Dugas M., Dupont D., Schmidt A., Singleton P., De Moor G., Kalra D. (2013). Electronic health records: new opportunities for clinical research. Journal of Internal Medicine, 274(6): 547-560.

Delen D., Walker G., Kadam A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. Artificial Intelligence in Medicine, 34: 113-127.

El-Sappagh S.H., El-Masri S. (2014). A distributed clinical decision support system architecture. Journal of King Saud University – Computer and Information Sciences, 26: 69-78.

Hayrinen K., Saranto K., Nykanen P. (2008). Definition, structure, content, use and impacts of electronic health records: A review of the research literature. International Journal of Medical Informatics, 77: 291-304.

Herland M., Khoshgoftaar T.M., Wald R. (2014). A review of data mining using big data in health informatics. Journal of Big Data, 1(2): 1-35.

ISO/TR 20514. (2005). Health Informatics – Electronic health record – Definition, scope and context.

Kuhn T. (2014). A survey and classification of controlled natural languages. Journal of Computational Linguistics, 40(1): 121-170.

Lee N., Laine A.F., Hu J., Wang F., Sun J., Ebadollahi S. (2011). Mining electronic medical records to explore the linkage between healthcare resource utilization and disease severity in diabetic patients. Proceedings of the 1st International Conference on Healthcare Informatics, Imaging and Systems Biology, pp. 250-257.

Li T., Zhu S., Ogihara M. (2003). A new distributed data mining model based on similarity. Proceedings of the 2003 ACM Symposium on Applied Computing, pp. 432-436.

Lin Y., Hu X., Li X., Wu X. (2013). Mining stable patterns in multiple correlated databases. Decision Support Systems, 56: 202-210.

Peterson L.T., Ford E.W., Eberhardt J., Huerta T.R., Menachemi N. (2011). Assessing differences between physicians' realized and anticipated gains from electronic health record adoption. Journal of Medical Systems, 35: 151-161.

Phillips-Wren G., Sharkey P., Dy S.M. (2008). Mining lung cancer patient data to assess healthcare resource utilization. Expert Systems with Applications, 35: 1611-1619.

Raval M.V., Rust L., Thakkar R.K., Kurtovic K.J., Nwomeh B.C., Besner G.E., Kenney B.D. (2015). Development and implementation of an electronic health record generated surgical handoff and rounding tool. Journal of Medical Systems, 39: 8pp.

Sakata M., Yucel Z., Shinozawa K., Hagita N., Imai M., Furutani M., Matsuoka R. (2013). An inference engine for estimating outside states of clinical test items. ACM Transactions on Management Information Systems, 4(3), Article 13, 13:1-13:21.

Shinozawa K., Hagita N., Furutani M., Matsuoka R. (2009). A data mining method for finding hidden relationship in blood and urine examination items for health check. Proceedings of the 9th Industrial Conference on Data Mining, pp. 44-50.

Sulzmann J., Furnkranz J. (2008). A comparison of techniques for selecting and combining class association rules. Proceedings of ECML PKDD 2008 Workshop: From Local Patterns to Global Models, pp. 87-93.

Theera-Ampornpunt N. (2010). Electronic health records: what does the HITECH act teach Thailand? Proceedings of Thai Medical Informatics Association's 19th Annual Conference – Health Informatics: From Standards to Practice, 9 pp.

Tsarkov D., Horrocks I. (2006). FaCT++ description logic reasoner: system description. Proceedings of the 3rd International Joint Conference on Automated Reasoning, pp. 292-297.

Yin K., Hsieh Y., Yang D. (2010). GLFMiner: Global and local frequent pattern mining with temporal intervals. Proceedings of the 5th IEEE Conference on Industrial Electronics and Applications, pp. 2248-2253.

Yuregir O.H., Oral M., Kalan O. (2010). A decision support system for preventing Ligionella disease. Journal of Medical Systems, 34: 875-881.

Zaidi S.Z.H., Abidi S.S.R., Manickam S. (2002). Distributed data mining from heterogeneous healthcare data repositories: Towards an intelligent agent-based framework. Proceedings of the 15th IEEE Symposium on Computer-Based Medical Systems, pp. 339-342.

# ภาคผนวก

## ผลผลิตของงานวิจัย

# ภาคผนวก ก

บทความวิจัยตีพิมพ์ในวารสารวิชาการ

1. N. Kaoungku, K. Kerdprasop, N. Kerdprasop (2014). A technique to association rule mining on multiple datasets. *Journal of Advances in Information Technology*, vol.5, no.2, May, pp.53-57. (indexing: INSPEC, ISSN: 1798-2340)

2. N. Kaoungku, K. Kerdprasop, N. Kerdprasop (2017). A method to clustering the feature ranking on data classification using an ensemble feature selection. International *Journal of Future Computer and Communication*, vol. 6, no. 3, September 2017, pp. 81-85. (indexing: INSPEC, ISSN: 2010-3751)

3. N. Kaoungku, K. Suksut, R. Chanklan, K. Kerdprasop, N. Kerdprasop (2018). The silhouette width criterion for clustering and association mining to select image features. *International Journal of Machine Learning and Computing*, vol. 8, no. 1, pp. 69-73. (indexing: Scopus, ISSN: 2010-3700)

# A Technique to Association Rule Mining on Multiple Datasets

Nuntawut Kaoungku
School of Computer Engineering, Institute of Engineering Suranaree University of Technology, Thailand.
Email: b5111299@gmail.com

Kittisak Kerdprasop and Nittaya Kerdprasop
School of Computer Engineering, Institute of Engineering Suranaree University of Technology, Thailand.
Email: KittisakThailand@gmail.com and nittaya@sut.ac.th

***Abstract*— This research aims at studying the method for association rule mining on multiple datasets. Current with technology and information systems enabling agencies or organization has a data-storage system, but the problem is that those with a larger data set, which is difficult in the association rule mining, because it requires a computer with a high-performance to process, which was followed by a cost increase. How it can help solve this problem is to distribute data process according to multiple computers, then combined rules of each machine using Fact + + Reasoner for check conflicts of rules, and will therefore have powerful association rules similar to the method for association rule mining on one dataset. We thus propose a technique for association rule mining on multiple datasets.**

***Index Terms*—Association rule mining, Controlled language, Attempto Controlled English**

## I. Introduction

Current, with the rapid development of technology allows agencies and organizations have adopted various technologies applied to the agency or the organization even more. These technologies make it possible to easily and systematically, but what follows is data that is stored large, which is difficult in association rule mining. As it required a computer with high-performance to processing and high cost, which the large organizations to have the financial resources to association rule mining from large data set. There is a technique to help fix this problem is to distribute the data set to be processed by multiple computers, by the computer, it does not require a high-performance to processing in association rule mining. However, it may have a conflict with the association rules in the process of combining association rules from each machine, and association rules from multiple datasets may be inefficient compared to the association rules from only data set. So in the process of combining association rules requires a technique to help fix the problems mentioned above.

Manuscript received November 30, 2013; revised December 15, 2013; accepted December 31, 2013.
Journal of Computers (JCP, ISSN 1796-203X), corresponding author.

Step in the combine association rules from distributed data is essential as well, as association rules from multiple datasets must be close to the most powerful association rules from one datasets and association rules must be inconsistency. Examination of conflict in association rules is used Fact + + Reasoner [7] and need to write rules in the form of Attempto Controlled English (ACE) [2], which is a Controlled Natural Language (CNL) on the Protégé.

Researches related to association rule mining on multiple datasets have to appear very little. Probably, due to the association rule mining on multiple dataset that is difficult process of combined association rules from distributed data, association rules with efficient close to that of association rule mining from one datasets. The researchers appeared, there was an inefficient comparison clearly. [1, 3, 8]

From the above it can be seen that association rule mining relations from large dataset it is difficult, there is a need to distribute data processing according to multiple computers. Combining association rule from each of computers may be a problem in the conflict of association rules and the efficiency of association rules. We thus propose a technique to association rule mining on multiple datasets.

## II. Background

### A. Association Rule Mining

Association Rule Mining is a process that has been popular in the relationship between the data that is how most association rule mining in a variety of ways. In this paper, the algorithm Apriori [6] of the association rule mining.

TABLE 1
PURCHASE TRANSACTIONS OF ALL CUSTOMERS

| Order | Milk | Water | Candy | Sausage |
|---|---|---|---|---|
| 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 |


doi:10.4304/jait.5.2.53-57

Table 1 is show a purchase transaction of all customers and then fined the frequency pattern of purchases of customers in each piece, to find the relationship of each product which is shown in Table 2, after which it will be used with high frequency items set to generate association rules, which is in the form of IF condition Then result by the criteria used in the present are the following:

- Support is the frequency of the event occurring
- Confidence is the frequency of the incident with other events occurring together.

TABLE 2
THE FREQUENCY OF CUSTOMER PURCHASES.

| | Milk | Water | Candy | Sausage |
|---|---|---|---|---|
| **Milk** | 2* | 2 | 1 | 0 |
| **Water** | 2 | 4* | 2 | 0 |
| **Candy** | 1 | 1 | 2* | 0 |
| **Sausage** | 0 | 0 | 0 | 1* |

### *B. Attempto Controlled English*

ACE is a controlled natural language that is based on first-order logic language, which combines the advantages of natural languages ˚ı ¸ formal language to want to make the writing language in the form of human and machine can understand, can be written in the form of simple English sentences [2], as shown by Figure 1 is an example of comparison between FOL, DL, OWL, UML, and ACE. ACE is a plugin of Protégé editor, in this research association rules are written in the form of ACE because will lead association rules to check conflicts with Fact++ Reasoner in Protégé editor. Table 3 shows an example of converting association rules in the form of ACE.

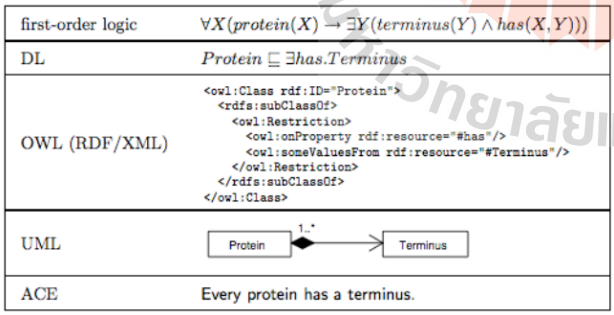



Figure 1 Example of comparison between FOL, DL, OWL, UML, and ACE

TABLE 3
EXAMPLE OF CONVERTING ASSOCIATION RULES IN THE FORM OF ACE

| Original association rules | Association rules in ACE |
|---|---|
| {CLASS=crew} => {SEX=male} | If X is a **crew** then X is a **male**. |
| {CLASS=crew, AGE=adult} => {SEX=male} | If X is a **crew** and X is an **adult** then X is a **male**. |
| {CLASS=crew, AGE=adult, SURVIVED=no} => {SEX=male} | If X is a **crew** and X is an **adult** and X is a n: **not-survivor** then X is a **male**. |

### *C. FaCT++ Reasoner*

FaCT + + is reasoner was developed from FaCT algorithm using C + + language development, which is based on Description Logics (DL), to be used for checking the inconsistency of Ontology [7], for example following:

> Every man is a human.
> John is a man.
> John is not a human.

For example, it can be seen that the conflict in the sentence “John is not a human”, because two sentences have previously said that “John is a human” and could not use sentences in an example to created ontology. In this research, association rule mining from distribute data, may be association rules is a conflict, so it need FaCT++ Reasoner to checking the conflict of the association rules from multiple datasets

## III. METHODOLOGY

This research proposed a technique for association rule mining on multiple datasets, the data is divided according to multiple computers to help in the association rule mining, replace association rule mining from large dataset, which require a computer with high-performance to process. But in the process of combined association rules from multiple datasets is difficult, because to the association rules with performance close to the association rule mining from large dataset and association rules that may conflict.

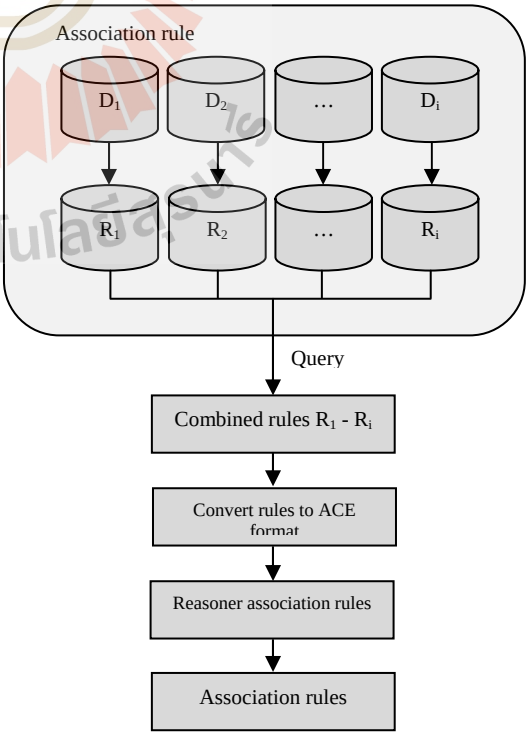



Figure 2 Conceptual framework of the research

| $R_1$ | $R_2$ | $R_3$ |
|---|---|---|
| {CLASS=third} => {AGE=adult} | | |
| {CLASS=crew} => {SEX=male} | {CLASS=crew} => {SEX=male} | {CLASS=crew} => {SEX=male} |
| {CLASS=crew} => {AGE=adult} | {CLASS=crew} => {AGE=adult} | {CLASS=crew} => {AGE=adult} |

| R |
|---|
| {CLASS=crew} => {SEX=male} |
| {CLASS=crew} => {AGE=adult} |

Figure 3 Example combined association rules

Figure 1 shows conceptual framework of the research. First, association rule mining from multiple datasets by $D_1, D_2, \ldots, D_t$ with $t = 1, 2, \ldots, n$. Second, combined association rules from first step by $R = R_1 \cap R_2 \cap \ldots \cap R_t$ with $t = 1, 2, \ldots, n$ which is shown in figure 3. Third, converting association rules in the form of ACE. Forth, checking the conflict of the association rules with FaCT++ Reasoner. Finally, the association rules from multiple datasets with similar efficient to the association rules from one dataset, this can be checked from the ontology created from Protégé editor.

## IV. Experiment Result

This research experimented to compare the results from the association rule mining on multiple datasets and the association rule mining on one dataset used Breast-cancer dataset from the UCI Machine Learning Repository. Breast-cancer dataset has 10 attributes and 286 data instances, figure 4 is an example of Breast-cancer dataset are 5 instants.

The experiment will divided breast-cancer dataset for association rule mining to three datasets, which use minimum support for association rule mining at 0.3 and 0.5. Table 4 show comparative results of association rule mining on multiple dataset and association rule mining on one dataset with minimum support 0.3, it can be seen that the association rules from multiple dataset missing a lot, and when considering ontology Figure 5 shows that association rules from multiple dataset and association rules from one dataset are clearly different. Table 6 show comparative results of association rule mining on multiple dataset and association rule mining on one dataset with minimum support 0.5, it can be seen that there are some association rules from multiple dataset still missing, and when considering Ontology of Figure 6 shows that association rules from multiple dataset effectively close association rules from one dataset.

Association rule mining with minimum support 0.3 and 0.5, table 8 show association rule mining from multiple dataset with minimum support 0.5 it provides the number of rules closely to association rule mining from one dataset more than minimum support 0.3. Association rule mining from multiple dataset there is a missing, can be certain association rules of table 5 and table 7 to fill in some missing association rules.

| age | menopause | tumor-size | inv-nodes | node-caps | deg-malig | breast | breast-quad | irradiat | Class |
|---|---|---|---|---|---|---|---|---|---|
| 40-49 | premeno | 15-19 | 0-2 | yes | 3 | right | left_up | no | recurrence-events |
| 50-59 | ge40 | 15-19 | 0-2 | no | 1 | right | central | no | no-recurrence-events |
| 50-59 | ge40 | 35-39 | 0-2 | no | 2 | left | left_low | no | recurrence-events |
| 40-49 | premeno | 35-39 | 0-2 | yes | 3 | right | left_low | yes | no-recurrence-events |
| 40-49 | premeno | 30-34 | 3-5 | yes | 2 | left | right_up | no | recurrence-events |

Figure 4 Example of Breast-cancer dataset

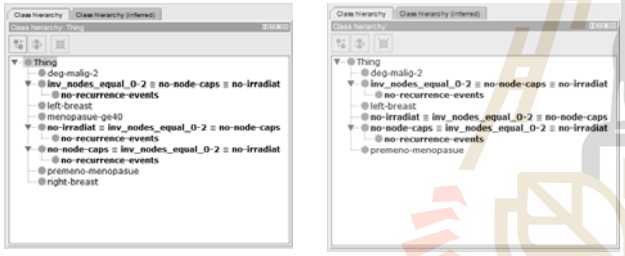
TABLE 4
COMPARATIVE RESULTS OF ASSOCIATION RULE MINING ON MULTIPLE DATASET AND ASSOCIATION RULE MINING ON ONE DATASET WITH MINIMUM SUPPORT 0.3

| Original association rules | Combined association rules |
|---|---|
| if X is a n:inv_nodes_equal_0-2 and X is a n:left-breast then X is a n:no-node-caps. | if X is a n:inv_nodes_equal_0-2 and X is a n:left-breast then X is a n:no-node-caps. |
| if X is a n:no-node-caps and X is a n:left-breast then X is a n:inv_nodes_equal_0-2. | if X is a n:no-node-caps and X is a n:left-breast then X is a n:inv_nodes_equal_0-2. |
| if X is a n:premeno-menopasue and X is a n:inv_nodes_equal_0-2 then X is a n:no-node-caps. | if X is a n:premeno-menopasue and X is a n:inv_nodes_equal_0-2 then X is a n:no-node-caps. |
| if X is a n:premeno-menopasue and X is a n:no-node-caps then X is a n:inv_nodes_equal_0-2. | if X is a n:premeno-menopasue and X is a n:no-node-caps then X is a n:inv_nodes_equal_0-2. |
| if X is a n:left-breast and X is a n:no-irradiat then X is a n:no-node-caps. | if X is a n:left-breast and X is a n:no-irradiat then X is a n:no-node-caps. |
| if X is a n:no-node-caps and X is a n:left-breast then X is a n:no-irradiat. | if X is a n:no-node-caps and X is a n:left-breast then X is a n:no-irradiat. |
| if X is a n:inv_nodes_equal_0-2 and X is a n:left-breast then X is a n:no-irradiat. | if X is a n:inv_nodes_equal_0-2 and X is a n:left-breast then X is a n:no-irradiat. |
| if X is a n:left-breast and X is a n:no-irradiat then X is a n:inv_nodes_equal_0-2. | if X is a n:left-breast and X is a n:no-irradiat then X is a n:inv_nodes_equal_0-2. |
| if X is a n:inv_nodes_equal_0-2 and X is a n:left-breast and X is a n:no-irradiat then X is a n:no-node-caps. | if X is a n:inv_nodes_equal_0-2 and X is a n:left-breast and X is a n:no-irradiat then X is a n:no-node-caps. |
| if X is a n:inv_nodes_equal_0-2 and X is a n:no-node-caps and X is a n:left-breast then X is a n:no-irradiat. | if X is a n:inv_nodes_equal_0-2 and X is a n:no-node-caps and X is a n:left-breast then X is a n:no-irradiat. |
| if X is a n:no-node-caps and X is a n:left-breast and X is a n:no-irradiat then X is a n:inv_nodes_equal_0-2. | if X is a n:no-node-caps and X is a n:left-breast and X is a n:no-irradiat then X is a n:inv_nodes_equal_0-2. |
| if X is a n:premeno-menopasue and X is a n:no-irradiat then X is a n:no-node-caps. | if X is a n:premeno-menopasue and X is a n:no-irradiat then X is a n:no-node-caps. |
| if X is a n:premeno-menopasue and X is a n:no-node-caps then X is a n:no-irradiat. | if X is a n:premeno-menopasue and X is a n:no-node-caps then X is a n:no-irradiat. |
| if X is a n:premeno-menopasue and X is a n:inv_nodes_equal_0-2 then X is a n:no-irradiat. | |
| if X is a n:premeno-menopasue and X is a n:no-irradiat then X is a n:inv_nodes_equal_0-2. | |
| if X is a n:inv_nodes_equal_0-2 and X is a n:deg-malig-2 then X is a n:no-node-caps. | if X is a n:inv_nodes_equal_0-2 and X is a n:deg-malig-2 then X is a n:no-node-caps. |
| if X is a n:no-node-caps and X is a n:deg-malig-2 then X is a n:inv_nodes_equal_0-2. | if X is a n:no-node-caps and X is a n:deg-malig-2 then X is a n:inv_nodes_equal_0-2. |
| if X is a n:left-breast and X is a n:no-recurrence-events then X is a n:inv_nodes_equal_0-2. | if X is a n:left-breast and X is a n:no-recurrence-events then X is a n:inv_nodes_equal_0-2. |
| if X is a n:inv_nodes_equal_0-2 and X is a n:left-breast then X is a n:no-recurrence-events. | |
| if X is a n:menopasue-ge40 and X is a n:inv_nodes_equal_0-2 then X is a n:no-node-caps. | |
| if X is a n:menopasue-ge40 and X is a n:no-node-caps then X is a n:inv_nodes_equal_0-2. | |
| if X is a n:inv_nodes_equal_0-2 and X is a n:right-breast then X is a n:no-node-caps. | |
| if X is a n:no-node-caps and X is a n:right-breast then X is a n:inv_nodes_equal_0-2. | |
| if X is a n:left-breast and X is a n:no-recurrence-events then X is a n:no-node-caps. | |
| if X is a n:premeno-menopasue and X is a n:inv_nodes_equal_0-2 and X is a n:no-irradiat then X is a n:no-node-caps. | |
| if X is a n:premeno-menopasue and X is a n:inv_nodes_equal_0-2 and X is a n:no-node-caps then X is a n:no-irradiat. | |
| if X is a n:premeno-menopasue and X is a n:no-node-caps and X is a n:no-irradiat then X is a n:inv_nodes_equal_0-2. | |
| if X is a n:left-breast and X is a n:no-recurrence-events then X is a n:no-irradiat. | |
| if X is a n:inv_nodes_equal_0-2 and X is a n:left-breast and X is a n:no-recurrence-events then X is a n:no-node-caps. | |
| if X is a n:no-node-caps and X is a n:left-breast and X is a n:no-recurrence-events then X is a n:inv_nodes_equal_0-2. | |

| | |
|---|---|
| if X is a n:inv_nodes_equal_0-2 and X is a n:no-node-caps and X is a n:left-breast then X is a n:no-recurrence-events. | |
| if X is a n:menopasue-ge40 and X is a n:no-irradiat then X is a n:no-node-caps. | |
| if X is a n:menopasue-ge40 and X is a n:no-node-caps then X is a n:no-irradiat. | |
| if X is a n:premeno-menopasue and X is a n:no-recurrence-events then X is a n:no-node-caps. | |
| if X is a n:deg-malig-2 and X is a n:no-irradiat then X is a n:no-node-caps. | if X is a n:deg-malig-2 and X is a n:no-irradiat then X is a n:no-node-caps. |
| if X is a n:no-node-caps and X is a n:deg-malig-2 then X is a n:no-irradiat. | if X is a n:no-node-caps and X is a n:deg-malig-2 then X is a n:no-irradiat. |
| if X is a n:right-breast and X is a n:no-irradiat then X is a n:no-node-caps. | |
| if X is a n:no-node-caps and X is a n:right-breast then X is a n:no-irradiat. | |

TABLE 5
ENTAILMENT FROM REASONER ASSOCIATION RULES WITH MINIMUM SUPPORT 0.3

| Entailment | ACE If-then |
|---|---|
| Every inv_nodes_equal_0-2 is a no-irradiat that is a no-node-caps . | If X is a n:inv_nodes_equal_0-2 and X is a n:no-irradiat then X is a n:no-node-caps . |
| Every no-irradiat is an inv_nodes_equal_0-2 that is a no-node-caps . | If X is a n:no-irradiat and X is a n:inv_nodes_equal_0-2 then X is a n:no-node-caps . |
| Every no-node-caps is an inv_nodes_equal_0-2 that is a no-irradiat . | If X is a n:no-node-caps and X is a n:inv_nodes_equal_0-2 then X is a n:no-irradiat . |



(a) (b)

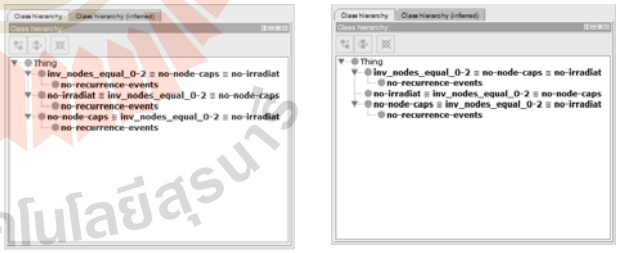
Figure 5 Ontology from association rule mining on one dataset (a) and association rule mining on multiple dataset (b) with minimum support 0.3

TABLE 6
COMPARATIVE RESULTS OF ASSOCIATION RULE MINING ON MULTIPLE DATASET AND ASSOCIATION RULE MINING ON ONE DATASET WITH MINIMUM SUPPORT 0.5

| Original association rules | Combined association rules |
|---|---|
| if X is a n:inv_nodes_equal_0-2 then X is a n:no-node-caps. | if X is a n:inv_nodes_equal_0-2 then X is a n:no-node-caps. |
| if X is a n:no-node-caps then X is a n:inv_nodes_equal_0-2. | if X is a n:no-node-caps then X is a n:inv_nodes_equal_0-2. |
| if X is a n:no-irradiat then X is a n:no-node-caps. | if X is a n:no-irradiat then X is a n:no-node-caps. |
| if X is a n:no-node-caps then X is a n:no-irradiat. | if X is a n:no-node-caps then X is a n:no-irradiat. |
| if X is a n:inv_nodes_equal_0-2 then X is a n:no-irradiat. | |
| if X is a n:no-irradiat then X is a n:inv_nodes_equal_0-2. | if X is a n:no-irradiat then X is a n:inv_nodes_equal_0-2. |
| if X is a n:inv_nodes_equal_0-2 and X is a n:no-irradiat then X is a n:no-node-caps. | if X is a n:inv_nodes_equal_0-2 and X is a n:no-irradiat then X is a n:no-node-caps. |
| if X is a n:inv_nodes_equal_0-2 and X is a n:no-node-caps then X is a n:no-irradiat. | if X is a n:inv_nodes_equal_0-2 and X is a n:no-node-caps then X is a n:no-irradiat. |
| if X is a n:no-node-caps and X is a n:no-irradiat then X is a n:inv_nodes_equal_0-2. | if X is a n:no-node-caps and X is a n:no-irradiat then X is a n:inv_nodes_equal_0-2. |
| if X is a n:no-recurrence-events then X is a n:no-node-caps. | if X is a n:no-recurrence-events then X is a n:no-node-caps. |
| if X is a n:no-recurrence-events then X is a n:inv_nodes_equal_0-2. | if X is a n:no-recurrence-events then X is a n:inv_nodes_equal_0-2. |
| if X is a n:no-recurrence-events then X is a n:no-irradiat. | |
| if X is a n:inv_nodes_equal_0-2 and X is a n:no-recurrence-events then X is a n:no-node-caps. | if X is a n:inv_nodes_equal_0-2 and X is a n:no-recurrence-events then X is a n:no-node-caps. |
| if X is a n:no-node-caps and X is a n:no-recurrence-events then X is a n:inv_nodes_equal_0-2. | if X is a n:no-node-caps and X is a n:no-recurrence-events then X is a n:inv_nodes_equal_0-2. |
| if X is a n:no-irradiat and X is a n:no-recurrence-events then X is a n:no-node-caps. | if X is a n:no-irradiat and X is a n:no-recurrence-events then X is a n:no-node-caps. |
| if X is a n:no-node-caps and X is a n:no-recurrence-events then X is a n:no-irradiat. | if X is a n:no-node-caps and X is a n:no-recurrence-events then X is a n:no-irradiat. |
| if X is a n:no-node-caps and X is a n:no-irradiat then X is a n:no-recurrence-events. | |
| if X is a n:inv_nodes_equal_0-2 and X is a n:no-recurrence-events then X is a n:no-irradiat. | if X is a n:inv_nodes_equal_0-2 and X is a n:no-recurrence-events then X is a n:no-irradiat. |
| if X is a n:no-irradiat and X is a n:no-recurrence-events then X is a n:inv_nodes_equal_0-2. | if X is a n:no-irradiat and X is a n:no-recurrence-events then X is a n:inv_nodes_equal_0-2. |
| if X is a n:inv_nodes_equal_0-2 and X is a n:no-irradiat then X is a n:no-recurrence-events. | |
| if X is a n:inv_nodes_equal_0-2 and X is a n:no-irradiat and X is a n:no-recurrence-events then X is a n:no-node-caps. | if X is a n:inv_nodes_equal_0-2 and X is a n:no-irradiat and X is a n:no-recurrence-events then X is a n:no-node-caps. |
| if X is a n:inv_nodes_equal_0-2 and X is a n:no-node-caps and X is a n:no-recurrence-events then X is a n:no-irradiat. | if X is a n:inv_nodes_equal_0-2 and X is a n:no-node-caps and X is a n:no-recurrence-events then X is a n:no-irradiat. |
| if X is a n:no-node-caps and X is a n:no-irradiat and X is a n:no-recurrence-events then X is a n:inv_nodes_equal_0-2. | if X is a n:no-node-caps and X is a n:no-irradiat and X is a n:no-recurrence-events then X is a n:inv_nodes_equal_0-2. |
| if X is a n:inv_nodes_equal_0-2 and X is a n:no-node-caps and X is a n:no-irradiat then X is a n:no-recurrence-events. | |

TABLE 7
ENTAILMENT FROM REASONER ASSOCIATION RULES WITH MINIMUM SUPPORT 0.5

| Entailment | ACE If-then |
|---|---|
| Every no-recurrence-events is a no-irradiat . | If X is a n:no-recurrence-events then X is a n:no-irradiat . |
| Every inv_nodes_equal_0-2 is a no-irradiat that is a no-node-caps . | If X is a n: inv_nodes_equal_0-2 and X is a n: no-irradiat then X is a n:no-node-caps . |
| Every no-irradiat is an inv_nodes_equal_0-2 that is a no-node-caps . | If X is a n:no-irradiat and X is a n:inv_nodes_equal_0-2 then X is a n:no-node-caps . |
| Every no-node-caps is an inv_nodes_equal_0-2 that is a no-irradiat . | If X is a n:no-node-caps and X is a n:inv_nodes_equal_0-2 then X is a n:no-irradiat . |



(a) (b)

Figure 6 Ontology from association rule mining on one dataset (a) and association rule mining on multiple dataset (b) with minimum support 0.5

TABLE 8
COMPERATIVE RESULTS OF NUMBER OF RULES FROM ONE DATASET AND NUMBER OF RULES FROM MULTIPLE DATASET

| Minimum support | Number of rules from one dataset | Number of rules from multiple dataset |
|---|---|---|
| 0.3 | 65 | 37 |
| 0.5 | 24 | 19 |

## V. Conclusion

Association rule mining from large dataset, need a computer with high-performance to process and high cost. There is a technique to help fix this problem is to distribute datasets to be processed by multiple computers. The process combined association rules from distribute datasets take the same association rules and checking the conflict of the association rules.

From such experiments can be seen that association rule mining from multiple dataset with minimum support that many have a closely efficient the association rule mining from one dataset. This consider from ontology and inconsistency association rules, but association rule mining from multiple dataset there is a missing, can be result of reasoned process to fill missing association rules.

## References

[1] Domingos, Pedro. Prospects and challenges for multi-relational data mining. ACM SIGKDD explorations newsletter, vol.5, no.1, pp.80-83.
[2] Fuchs, Norbert E., Kaarel Kaljurand, and Tobias Kuhn (2008). Attempto controlled english for knowledge representation. Reasoning Web. Springer Berlin Heidelberg, pp.104-124.
[3] Han, Jiawei, and Yongjian Fu (1999). Mining multiple-level association rules in large databases. Knowledge and Data Engineering, IEEE Transactions, vol.11, no.5, pp.798-805.
[4] Kaljurand, Kaarel (2008). ACE View-An Ontology and Rule Editor based on Controlled English. International Semantic Web Conference (Posters & Demos). vol. 401.
[5] Norbert E. Fuchs and Kaarel Kaljurand (2006). Attempto Controlled English: Language, Tools and Applications [Online]. Available URL: http://attempto.ifi.uzh.ch/site/courses/files/ACE.Course.UniZH.1206.Getting_Started.pdf
[6] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami (1993). Mining Association Rules between Sets of Items in Large Database. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207-216
[7] Tsarkov, Dmitry, and Ian Horrocks (2006). FaCT++ description logic reasoner: System description. Automated reasoning. Springer Berlin Heidelberg, pp.292-297.
[8] Zhao Yanchang, et al (2007). Mining for combined association rules on multiple datasets. Proceedings of the 2007 international workshop on Domain driven data mining, pp.18-23.

**Nuntawut Kaoungku** is currently a doctoral student with the School of Computer Engineering, Suranaree University of Technology, Thailand. He received his bachelor degree in Computer Engineering from Suranaree University of Technology, Thailand, in 2012, and master degree in Computer Engineering from Suranaree University of Technology, Thailand, in 2013. He current research includes semantic web and association.

**Nittaya Kerdprasop** is an associate professor at the School of Computer Engineering, Suranaree University of Technology, Thailand. She received her bachelor degree in Radiation Techniques from Mahidol University, Thailand, in 1985, master degree in Computer Science from the Prince of Songkla University, Thailand, in 1991 and doctoral degree in Computer Science from Nova Southeastern University, U.S.A, in 1999. She is a member of ACM and IEEE Computer Society. Her research of interest includes Knowledge Discovery in Databases, Artificial Intelligence, Logic Programming, and Intelligent Databases.

**Kittisak Kerdprasop** is an associate professor and chair of the School of Computer Engineering, Suranaree University of Technology, Thailand. He received his bachelor degree in Mathematics from Srinakarinwirot University, Thailand, in 1986, mas ter degree in Computer Science from the Prince of Songkla University, Thailand, in 1991 and doctoral degree in Computer Science from Nova Southeastern University, U.S.A., in 1999. His current research includes Data mining, Artificial Intelligence, Functional and Logic Programming Languages, Computational Statistics.

# A Method to Clustering the Feature Ranking on Data Classification Using an Ensemble Feature Selection

Nuntawut Kaoungku, Kittisak Kerdprasop, and Nittaya Kerdprasop

***Abstract*—The aim of this paper is to improve the predictive performance of the classification process by means of building multiple data classification models based on the output from feature selection methods that use ensemble strategy to find the optimal set of features. Currently, the data volume has grown at an extreme rate causing a variety of problems. The big data situation has made automatic analysis tasks such as data classification facing low performance and high computational time problems when dealing with big data that are huge in both volume and dimensions. In this research work, we tackle the big data problem in the high dimensionality aspect. We propose an ensemble method to reduce data dimension by means of feature clustering to rank the potential features and also return suitable subset of features for further classifying the training data. The two paradigms of feature selection based on ensemble strategy are proposed and evaluated. Experimental results confirm the efficacy of our proposed feature ensemble method.**

***Index Terms*—Feature selection, ensemble learning, clustering, classification.**

## I. Introduction

Traditional data classification seems to be an easy and straightforward task when applying a single classification model to predict future data. Currently, electronic equipments are ubiquitous and extensively used, thus, causing a variety of data forms such as numeric, categorical, time series, images, and so on. It is difficult to build a single model from these data to make a high performance classifier for accurately predicting future or unseen data. The basic solution idea is to build multiple models from the same dataset and then combine the predicted results from those multiple models to output a final prediction. This technique is called an ensemble learning.

Ensemble learning is basically a technique to use multiple models or multiple learning algorithms to predicted future data with the major purpose of better classification in terms of accuracy. Which combining results from multiple models built from various methods, the popular result combining method is a simple voting [1]. Typically, ensemble learning can be achieved from a wide range of methods, but the popular methods are bagging [2] and booting [3]. The two ensemble methods have long applied by many researchers, and they have been proven to provide better classification performance.

From the continuous and increasing advancement of software and hardware technologies, new structured and unstructured data have been generated every day. It is difficult to analyze and build a model from these mixed type data, even with the aid of ensemble learning method, because these data are high in dimensionality. The technique to solve this problem is the use of filter of find and extract only the optimal set of features for building classification model. The filtering techniques can be generally divided into 2 groups: feature selection and feature extraction. The research focusing on feature selection method uses some measures to calculate weight and then choosing a subset of features ordered by the weight [4], [5]. The feature selection methods can be further divided into 2 sub-groups: those that automatically return optimal set of features, and those that return weight of features. It is, however, difficult to choose the optimal weight of features for data classification.

Therefore, many researchers try to solve the optimal feature selection problem by proposing the ensemble feature selection method. Bolón-Canedo et al. [6] have shown that data classification using an ensemble of filters by using five groups of different feature selection methods for building instance-based learning (IB1) model [7] and support vector machine (SVM) model [8]. Seijo-Pardo et al. [9] propose technique to select optimal set of features by using several different threshold values, such as fisher discriminant ratio, $\log_2(n)$, and top percent of features, for ensemble feature selection. These research works [6]-[9] report a promising performance of ensemble feature selection strategy to increase classification accuracy.

This research, thus, aims at proposing a method to improve data classification accuracy by means of an ensemble feature selection. We propose a hybrid ensemble feature selection method by both automatically return optimal set of features and return weight of features. Our proposed method selects the optimal set of the feature from the return weight of features reported by the clustering method using k-Means algorithm [10], [11].

The contributions of this paper are as follows:

- With the proposed method, k-Means clustering can be applied as feature selection tool to select the optimal subset of the features.
- The proposed method can be applied to ensemble learning using a variety of learning algorithms and can

Manuscript received February 15, 2017; revised April 10, 2017. This work was supported in part by grants from Suranaree University of Technology through the funding of Knowledge and Data Engineering Research Units.

N. Kaoungku is with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand (e-mail: nuntawut@sut.ac.th).

K. Kerdprasop is with the School of Computer Engineering. He is also with Knowledge Engineering Research Unit, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand (e-mail: kerdpras@sut.ac.th).

N. Kerdprasop is with the School of Computer Engineering. She is also with Data Engineering Research Unit, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand (e-mail: nittaya@sut.ac.th).

doi: 10.18178/ijfcc.2017.6.3.494

increase the predictive accuracy.

## II. Materials and Methods

### A. Ensemble Learning

Ensemble learning is a technique to build multiple models from training data. The main purpose of this technique is to increase the model accuracy. Fig. 1 shows the main concept of ensemble learning. The ensemble process starts by taking the training data to build multiple models using either the same algorithm, or different algorithms. Then, combine the results from all the models to generate a single output. There are various strategies to combine results, but the most applicable one is a majority vote.

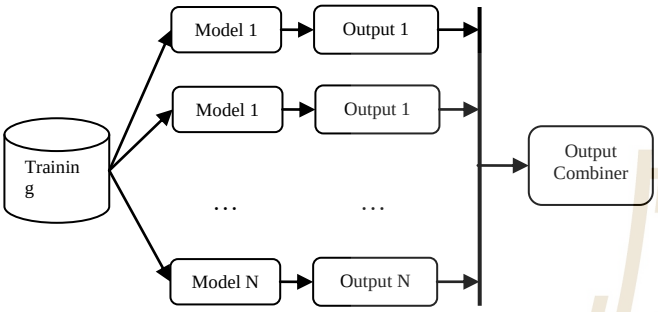



Fig. 1. The concept of ensemble learning.

Ensemble learning can be divided further into three classes of techniques [1]:

- Vote ensemble. It performs ensemble learning by building multiple models from one training dataset. To classify new data, it uses a majority vote to predict class of the new data.
- Bagging. It starts ensemble learning by dividing data, using random sampling technique, into several equal subsets. Each data subset is used to build the model. All built models are then used for classifying new data based on a majority vote.
- Random forest. It is ensemble learning method that is similar to bagging technique but it selects some of the features to each data subset.

### B. Feature Selection Method

Feature selection is a method to handle high dimensional data by reducing the data features based on some selection criteria. This method can reduce data dimensions and at the same time can increase the model accuracy. The examples of criteria for selecting feature subsets and returning the feature ranking score are as follows:

- Association rule mining-based feature selection (AFS) [12]. It is a method based on association analysis for analyzing features that are most influencing the class attribute. The calculation of frequent features from association rules is shown in equation (1). If the feature has the highest *FrequentFeature* score, that feature is the most influencing factor to the class attribute.

$$FrequentFeature(A) = \frac{AppearFrequency(A)}{\#Rules} \quad (1)$$

- Information Gain (IG) [13]. It selects features by measuring entropy, which is the measurement for purity of data with the same class. The computation of IG is shown in equations (2) and (3). The feature with high value of IG means the high potential of that feature on classifying data into class $c_1$ to $c_n$.

$$InfoGain = Entropy(initial) - [P(c_1) \times Entropy(c_1) + \ldots + P(c_n) \times Entropy(c_n)] \quad (2)$$

where

$$Entropy\ (c_1, c_2, \ldots, c_n) = -P(c_1)\log_2 P(c_1) - P(c_2)\log_2 P(c_2) - \ldots - P(c_n)\log_2 P(c_n) \quad (3)$$

### C. k-Means Clustering

k-Means clustering [10], [11] is an unsupervised learning algorithm for partitioning data into groups such that data subsets sharing similar attributes are assigned to be in the same group. This algorithm groups data into clusters by measuring the distance between data points. The most popular measure is Euclidean distance [14], as shown in equation (4).

$$dist(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \quad (4)$$

Fig. 2 presents the detail of the k-Means algorithm, which consists of five steps.

Step 1: at line 1, define the number of clusters (K) and the initial centroid, or central point, of each cluster.

Step 2: from lines 2 to 3, assign all data points to the closest centroid by measuring the distance between a data point to each centroid.

Step 3: at line 4, recompute the centroid of each cluster by calculating the average attribute value among all the points in each cluster.

Step 4: repeat steps 2 and 3 until the centroid does not change.

**Algorithm k-Means**

1. Select K point as the initial K centroids.
2. Repeat
3. Form K clusters by assigning all points to the closet centroid.
4. Recomputed the centroid of each cluster.
5. Until the centroid does not change

Fig. 2. k-Means algorithm.

## III. Proposed Work

In this section, we present the proposed process of clustering the feature ranking on data classification using an ensemble feature selection. The idea is that we use the k-Means algorithm to find the best cluster of the features from feature ranking scores and use these results to build the model for data classification. The objectives are to reduce the data dimensions and to increases the predictive accuracy.

The method of clustering the feature ranking on data classification using an ensemble feature selection is graphically shown in Fig. 3 and 4.

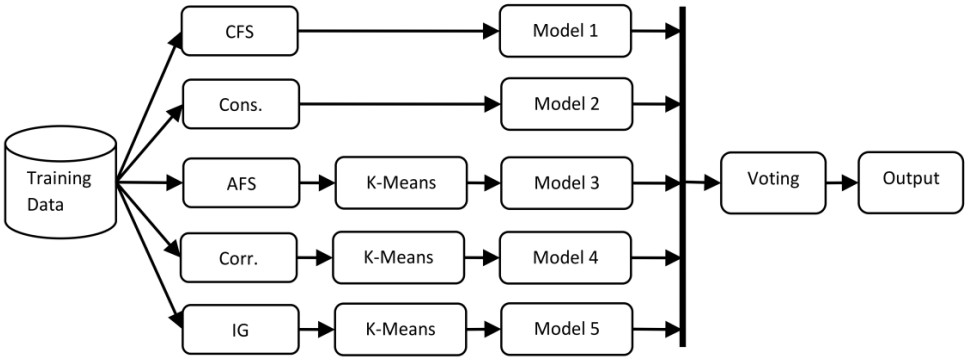



Fig. 3. The concept of ensemble 1.

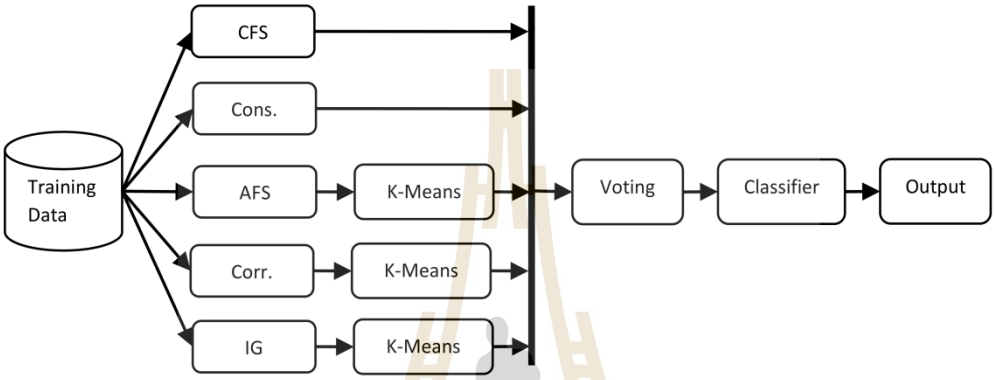



Fig. 4. The concept of ensemble 2.

**Feature Ranking Score**

| Features | Score |
|---|---|
| Attr.1 | 4 |
| Attr.2 | 4 |
| Attr.3 | 3 |
| Attr.4 | 3 |
| Attr.5 | 2 |
| Attr.6 | 2 |
| Attr.7 | 2 |
| Attr.8 | 1 |
| Attr.9 | 1 |
| Attr.10 | 1 |

**Cluster 0**

| Features | Score |
|---|---|
| Attr.1 | 4 |
| Attr.2 | 4 |
| Attr.3 | 3 |
| Attr.4 | 3 |

**Optimal set**

**Cluster 1**

| Features | Score |
|---|---|
| Attr.5 | 2 |
| Attr.6 | 2 |
| Attr.7 | 2 |
| Attr.8 | 1 |
| Attr.9 | 1 |
| Attr.10 | 1 |

Fig. 5. The concept of cluster the feature ranking score.

Our method consists of two parts: ensemble 1 and ensemble 2. Fig. 3 shows the steps in ensemble 1, which consists of three phases. The first phase of ensemble 1 feature selection method is reducing dimensions of the training data by using 5 feature selection methods including the correlation-based feature selection (CFS) [15], the consistency-based filter (Cons.) [16], the association rule mining-based feature selection (AFS), the correlation-based filter (Corr.), and the information gain (IG).

Ensemble 1 phase 2 is the clustering of feature ranking scores with the k-Means algorithm. The three ranking score methods used for clustering are the scores from the AFS, Corr, and IG methods. Fig. 5 shows running example for phase 2 of ensemble 1. The feature weight in Attr.1 to Attr.10 are clustered by k-Means (set k=2, user can increase k when the optimal set of feature is needed to be small size). The optimal set of the feature is cluster 0, which contains a set of features to be used for building a classification model.

Ensemble 1 phase 3, build the model with an optimal set of the feature from phase 2 with any data classification algorithm. The final step of ensemble 1 is the combining of the outputs from multiple models using a majority vote scheme to predicted class of data.

Fig. 4 shows concept of ensemble 2, which shares similar main idea to ensemble 1, but the ensemble 2 build a single classifier. At phase 3 of the ensemble 2 method, a majority vote of the feature from several feature selection methods generate a single set of optimal features. Then the optimal set of features is used for classification model building. The classification algorithm can be any one such as SVM and C4.5.

## IV. Experimental Results

The proposed ensemble feature selection methods have been experimented with data taken from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/). Table I shows details of the five data sets used in our experimentation. Each of these datasets has been divided into training dataset (70%) and test dataset (30%). We use the C4.5 and SVM algorithms for classification and use five feature selection methods, which are correlation-based feature selection (CFS), consistency-based (Cons.), association rule mining-based feature selection (AFS), correlation-based (Corr.), and information gain-based (IG).

Table II shows comparative results of classification accuracy and error. It can be seen that the ensemble 1 on C4.5

algorithm can improve the performance of accuracy on Spambase (92.72%) and Arrhythmia (66.91%) data sets. The proposed ensemble 1 and 2 on SVM algorithm can improve the performance of accuracy on Splice (96.68%) data set when compared to raw data set with no feature selection method and other feature selection algorithms.

Table III shows comparative results of average classification accuracy and error. It can be seen that the ensemble 1 performs well on the C4.5 algorithm (87.66%) when compared to ensemble 2 that is good on SVM algorithm (88.43%). When compared against other feature selection algorithms, it can be seen that our proposed ensemble feature selection algorithms using k-Means to cluster feature ranking can improve the performance of accuracy on C4.5 (IG = 86.63%) and SVM (Corr. = 88.39%). Table IV shows comparative results of the number of features selected by five feature selection algorithms. It can be seen the all five feature selection methods can reduce data dimensions.

TABLE I: DETAILS OF DATASETS

| Datasets | # Instances | # Features |
|---|---|---|
| Spambase | 4601 | 58 |
| Splice | 3190 | 62 |
| Opt digits | 5620 | 65 |
| Ozone | 2534 | 74 |
| Arrhythmia | 452 | 280 |

TABLE II: COMPARATIVE RESULTS OF CLASSIFICATION ACCURACY AND ERROR

| Methods | Spambase | | Ozone | | Splice | | Arrhythmia | | Opt digits | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Acc. | Err. | Acc. | Err. | Acc. | Err. | Acc. | Err. | Acc. | Err. |
| *Raw Data* | | | | | | | | | | |
| Raw + SVM | 92.65% | 7.35% | **94.00%** | 6.00% | 66.35% | 33.65% | 60.29% | 39.71% | **98.92%** | 1.08% |
| Raw + C4.5 | 92.13% | 7.87% | 92.09% | 7.91% | 93.57% | 6.43% | 62.50% | 37.50% | 89.70% | 10.30% |
| *SVM* | | | | | | | | | | |
| CFS + SVM | 88.75% | 11.25% | 92.36% | 7.64% | **96.57%** | 3.43% | 63.97% | 36.03% | 98.73% | 1.27% |
| Cons. + SVM | 89.71% | 10.29% | 93.86% | 6.14% | 93.03% | 6.97% | 60.29% | 39.71% | 86.33% | 13.67% |
| AFS + SVM | 89.63% | 10.37% | 93.45% | 6.55% | 94.75% | 5.25% | 63.77% | 36.23% | **98.98%** | 1.02% |
| Corr. + SVM | 90.59% | 9.41% | 94.00% | 6.00% | 96.14% | 3.86% | 62.50% | 37.50% | 98.73% | 1.27% |
| IG + SVM | 88.60% | 11.40% | 93.45% | 6.55% | 96.46% | 3.54% | 62.50% | 37.50% | 98.61% | 1.39% |
| *C4.5* | | | | | | | | | | |
| CFS + C4.5 | 91.91% | 8.09% | 91.27% | 8.73% | 93.57% | 6.43% | **66.18%** | 33.82% | 89.88% | 10.12% |
| Cons. + C4.5 | **92.06%** | 7.94% | 91.95% | 8.05% | 93.25% | 6.75% | 61.76% | 38.24% | 80.48% | 19.52% |
| AFS. + C4.5 | **92.06%** | 7.94% | 93.04% | 6.96% | 93.68% | 6.32% | 62.50% | 37.50% | 90.42% | 9.58% |
| Corr. + C4.5 | 91.47% | 8.53% | **94.13%** | 5.87% | 94.00% | 6.00% | 60.29% | 39.71% | 90.48% | 9.52% |
| IG + C4.5 | 91.91% | 8.09% | 93.18% | 6.82% | 94.00% | 6.00% | 65.44% | 34.56% | 88.61% | 11.39% |
| *Ensembles* | | | | | | | | | | |
| Emsemble1 + SVM | 90.07% | 9.93% | **93.45%** | 6.55% | 95.61% | 4.39% | 63.24% | 36.76% | **98.67%** | 1.33% |
| Emsemble1+ C4.5 | **92.72%** | 7.28% | 93.86% | 6.14% | 93.68% | 6.32% | **66.91%** | 33.09% | 91.14% | 8.86% |
| Emsemble2 + SVM | 89.78% | 10.22% | **93.04%** | 6.96% | **96.68%** | 3.32% | 63.97% | 36.03% | **98.67%** | 1.33% |
| Emsemble2 + C4.5 | 91.54% | 8.46% | 92.77% | 7.23% | 94.00% | 6.00% | 65.44% | 34.56% | 89.28% | 10.72% |

TABLE III: COMPARATIVE RESULTS OF NUMBER OF FEATURES BY FIVE FEATURE SELECTION ALGORITHMS

| Datasets | Raw | A | B | C | D | E |
|---|---|---|---|---|---|---|
| Spambase | 58 | 16 | 18 | 15 | 21 | **13** |
| Splice | 62 | 20 | 11 | **8** | 24 | 23 |
| Opt digits | 65 | 36 | **10** | 42 | 40 | 33 |
| Ozone | 74 | 15 | 23 | 26 | 23 | **13** |
| Arrhythmia | 280 | 32 | **19** | 50 | 46 | 25 |

TABLE IV: COMPARATIVE RESULTS OF AVERAGE ACCURACY AND ERROR

| Methods | SVM | | C4.5 | |
|---|---|---|---|---|
| | Accuracy | Error | Accuracy | Error |
| *Raw data* | 82.44 | 17.56 | 86.00 | 14.00 |
| *Features Selection* | | | | |
| CFS | 88.08 | 11.92 | 86.56 | 13.44 |
| Cons. | 84.64 | 15.36 | 83.90 | 16.10 |
| AFS | 88.12 | 11.88 | 86.34 | 13.66 |
| Corr. | **88.39** | **11.61** | 86.07 | 13.93 |
| IG | 87.92 | 12.08 | **86.63** | **13.37** |
| *Ensembles* | | | | |
| Ensemble 1 | 88.21 | 11.79 | **87.66** | **12.34** |
| Ensemble 2 | **88.43** | **11.57** | 86.61 | 13.39 |

## V. CONCLUSION

This research aims at studying a method to clustering the feature ranking on data classification using an ensemble feature selection. The problem of learning efficient model from data with high dimensionality can cause trouble to most algorithms. Thus, we propose to use the ensemble method at the feature selection step prior to the application of learning algorithm in order to increase accuracy and reduce learning problem due to dimensionality. We present clustering method using the k-Means algorithm to cluster the feature ranking scores for choosing an optimal set from feature ranking score.

From experimental results, it has been revealed that the proposed ensemble feature selection method can increase the accuracy of data classification, and can reduce high dimensional data problem by obtaining a small set of features. However, in some datasets our proposed ensemble method shows lower accuracy than the raw dataset with no feature selection applied. Even though the proposed method can

reduce data dimensions and hence expected to remedy the over-fitting problem, it still needs further improvement to perform well on every dataset. Such improvement is obviously planned as our future work.

## REFERENCES


[1] T. G. Dietterich, "Ensemble methods in machine learning," *International Workshop on Multiple Classifier Systems*, pp. 1-15, June 2000.
[2] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996.
[3] R. E. Schapire, "The strength of weak learnability," *Machine Learning*, vol. 5, no. 2, pp. 197-227, 1990.
[4] I. Guyon, "Practical feature selection: from correlation to causality," *NATO Science for Peace and Security*, vol. 19, pp. 27-43, 2008.
[5] H. G. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Proc. the Eleventh International Conference on Machine Learning*, pp. 121-129, 1994.
[6] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "Data classification using an ensemble of filters," *Neurocomputing*, vol. 135, pp. 13-20, 2014.
[7] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine Learning*, vol. 6, no. 1, pp. 37-66, 1991.
[8] N. Cristianini and J. Shawe-Taylor, "An introduction to support vector machines and other kernel-based learning methods," Cambridge University Press, 2000.
[9] B. Seijo-Pardo, I. Porto-Díaz, V. Bolón-Canedo, and A. Alonso-Betanzos, "Ensemble feature selection: Homogeneous and heterogeneous approaches," *Knowledge-Based Systems*, vol. 118, pp. 124-139. 2017.
[10] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967.
[11] M. R. Anderberg, "Cluster analysis for applications," *Academic Press*, 1973.
[12] N. Kaoungku, K. Kerdprasop, and N. Kerdprasop, "Data classification based on feature selection with association rule mining," *The International Multi Conference of Engineers and Computer Scientists 2017*, 2017.
[13] M. A. Hall, "Smith, Practical feature subset selection for machine learning," *Comput. Sci*, vol. 98, pp. 181–191, 1998.
[14] M. M. Deza and E. Deza, "Encyclopedia of distances," *Encyclopedia of Distances*, pp. 1-583, 2009.
[15] M. A. Hall, "Correlation-based feature selection for machine learning (Ph.D. thesis)," *Citeseer*, 1999.
[16] M. Dash and H. Liu, "Consistency-based search in feature selection," *Artif. Intell*, pp, 155–176, 2003.


**Nuntawut Kaoungku** is currently a lecturer at School of Computer Engineering, Suranaree University of Technology, Thailand. He received his doctoral degree in computer engineering from Suranaree University of Technology, Thailand, in 2014, the bachelor degree in computer engineering from Suranaree University of Technology, Thailand, in 2012, and the master degree in computer engineering from Suranaree University of Technology, Thailand, in 2013. His current research includes data mining and semantic web.

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

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

# The Silhouette Width Criterion for Clustering and Association Mining to Select Image Features

Nuntawut Kaoungku, Keerachart Suksut, Ratiporn Chanklan, Kittisak Kerdprasop, and Nittaya Kerdprasop

***Abstract*—Image data are normally unstructured and high dimensional due to the photography technology advancement such that an image can be taken at a wide range of resolution levels. To overcome such problem, data miners may consider selecting only a minimal set of features that are really important for classifying their images. Feature selection is a popular method for reducing dimensions in data. However, most feature selection algorithms return results in form of score for each feature. It is still difficult for data miners to choose features based on such scoring scheme because they may not know which score range is the best for their data classification at hand. Therefore, in this research, we aim to assist data miners and novice data analysts on solving dimensionality problem by finding for them the best optimal set of features, instead of just reporting the scores of all features and leaving the selection step to be the burden of miners. We select optimal set of features by firstly apply clustering technique to group similar features based on their scores. We thus propose the silhouette width criterion for selecting the optimal number of clusters during the cluster analysis step. After that we perform association mining to analyze relationships that may exist among different subsets of features toward the target attribute. Our method finally reports user the best subset of features to be potentially used further for data classification. We demonstrate performance of our proposed method on the satellite forest image data in Japan.**

***Index Terms*—Image data, feature selection, clustering, silhouette criterion, forrest type classification.**

## I. Introduction

With the rapid development of current electronic devices such as sensors and cameras, the outputs from these devices are of high quality and also high dimensions. Unfortunately, high dimensionality is still an unsolvable issue for many existing data mining and machine learning algorithms. Data with overwhelming attributes or dimensions can be a major cause of low computational performance. It can be even worse when such data may cause a creation of classifying model with low predictive accuracy due to the search for discriminative set of features is obscured by so many irrelevant features. Most classification algorithms are not designed to efficiently handle such high dimensionality problem.

Therefore, the numerous feature selection techniques have been proposed as a pre-classification step for solving the high dimensionality problem. Several research teams introduce many ways to reduce the number of features. The reduced set of features has been proven experimentally increasing the performance of learning process and also being able to build an accurate classification model. Generally, feature selection techniques can be divided into three classes [1]. The first class is called filter method, such as CfsSubsetEval [2], Information Gain, and Chi-Square [3]. The second class is the wrapper method [4], [5]. The filter method introduces some form of scoring computation without actually building a model, whereas the wrapper approach scoring the selected set of features by observing the error made by the classifying model. The last class is called the embedded method; it combines the advantages from both the filter and wrapper methods [5].

Xie *et al.* [6] proposed the association rule mining technique to calculate weight for find the optimal features that are closely correlative with the class attribute, but the proposed technique is quite complex and performance test with cross validation. Nuntawut *et al.* [7] proposed the filter method for feature selection based on association rule mining such that the specific set of association rules that the rules' consequence is the target class. But this feature selection algorithm does not work automatically because human is the one who select the features one by one based on the feature scores reported from the algorithm. Therefore, Nuntawut *et al.* [8] improved the algorithm by proposing clustering technique to cluster the feature scores to assist users on finding an appropriate groups of features. The clustering process is supposed to be automatic in the sense that the number of clusters should be judged by the process itself. However, the clustering algorithm is still semi-automatic in the sense that users must specify the suitable number of feature clusters.

This research, thus, aims at extending the previous work of Nuntawut *et al.* [7], [8] by proposing a silhouette width criterion for automatic setting of initial cluster numbers. We also add confidence criteria into feature selection based on association rule mining technique to increase performance. Experimental results confirm the efficacy of our proposed method that can extract only relevant set of image features from ASTER satellite resulting in better recognition for each forest type.

## II. Materials and Methods

### A. Feature Selection Based on Association Rule Mining

Association rule mining is finding the frequent patterns in

Manuscript received September 20, 2017; revised January 10, 2017. This work was supported by grant from Suranaree University of Technology through the funding of Knowledge Engineering Research Unit.

The authors are with the School of Computer Engineering, Suranaree University of Technology (SUT), Nakhon Ratchasima 30000, Thailand (corresponding author: N. Kaoungku; Tel: +66872155059; e-mail: nuntawut@sut.ac.th, mikaiterng@gmail.com, arc_angle@hotmail.com, kittisakThailand@gmail.com, nittaya@sut.ac.th).

database and present them in the form of association rules [9]. Generally, there can be so many possible association rules from this technique. Therefore, some constraints are necessary for reducing such exponential growth. There are two popular criteria: support and confidence. Support is the frequency of the occurring event, as shown in (1). Confidence is the proportion of frequency of co-occurring events to the frequency of antecedent event, as shown in (2).

$$\text{Support}, \text{Supp}(X \rightarrow Y) = P(X \wedge Y) \quad (1)$$

$$\text{Confidence}, \text{Conf}(X \rightarrow Y) = \frac{\text{Supp}(X \rightarrow Y)}{\text{Supp}(X)} \quad (2)$$

This technique had been successfully applied to multiple disciplines such as marketing to increase sales. Nuntawut *et al*. [7] applied this technique to find optimal feature set from high dimensional dataset by finding association rules that the target class appears in the consequence of the rule. Then, consider the features or attributes that are most influencing the target class. The algorithm consists of 4 steps:

Step 1: define minimum frequency threshold, support, and confidence. Find frequent patterns and then generate association rules based on the Apriori algorithm [10].

Step 2: select only association rules that their consequence is target class.

Step 3: count features that appear on association rules.

Step 4: calculate frequent features in percentage, as in (3). Then, remove any feature having percentage of frequency appearance in the set of association rules lower than the specified minimum frequency threshold.

$$\text{FrequentFeature} = \frac{\text{AppearFrequency}}{\#\text{Rules}} \times 100 \quad (3)$$

### B. k-Means Clustering

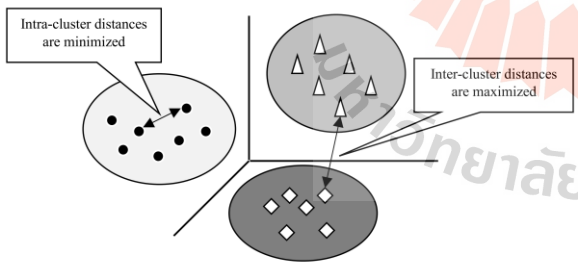



Fig. 1. Forming three clusters with minimized intra-cluster distance but maximized the inter-cluster distance.

k-Means algorithm is unsupervised learning method to form data into clusters based on data similarity regardless of the target class information. Fig. 1 depicts the idea of clustering such that distance between data in the same cluster (intra-cluster) is low, whereas the distance between data in one cluster to data in another cluster (inter-cluster) is high [11], [12]. The k-means algorithm can be explained by the following 4 steps:

Step 1: define the number of clusters (K) and randomly pick k instances as the initial cluster centroids.

Step 2: assign all data points to the closest centroid by measuring the distance such as the Euclidean distance.

Step 3: re-compute the centroid of each cluster by calculating mean value of all the data points in the cluster.

Step 4: repeat steps 2 and 3 until the centroid does not change.

### C. Silhouette Coefficient

The shortcoming of k-means clustering is the appropriate choice of *k*, which is the number of clusters. Silhouette coefficient is a popular measure for considering such parameter. The silhouette coefficient can be computed by using average distance between data points in the same cluster compared against average distances between data points in other clusters. Fig. 2 shows main concept of the silhouette coefficient to calculate the silhouette average of all cluster.

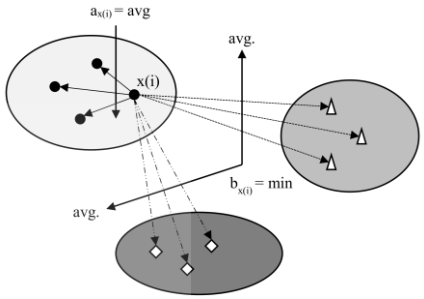



Fig. 2. Concept of the silhouette coefficient.

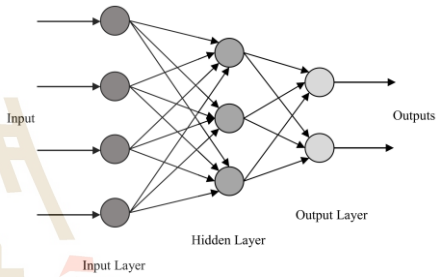



Fig. 3. The architecture of artificial neural networks.

Define *K* to be cluster composing of data $x(i)$ and $a_{x(i)}$ is average distance between $x_{(i)}$ to every data point in the cluster *K*. The notation $b_x$ is minimum average distance between $x(i)$ and every data point in other clusters that are not a member in *K*. The calculation [13] of the silhouette coefficient of $x(i)$, the silhouette average of each cluster, and the silhouette average of all cluster can be shown as in (4), (5), and (6), respectively.

$$S_{x(i)} = \frac{b_{x(i)} - a_{x(i)}}{\max(a_{x(i)}, b_{x(i)})} \quad (4)$$

where

$x(i)$ = data point in the cluster, $i$ = 1, 2, 3, …, $n$,

$a_{x(i)}$ = average distance between $x_i$ and every data point in the same cluster, and

$b_{x(i)}$ = minimum average distance between $x_i$ and every data point in other clusters.

$$S_k = \frac{1}{n}\sum_{i=1}^{n} S_{x(i)} \quad (5)$$

where $k$ = number of clusters, and

$n$ = number of data points in the same cluster.

$$S_{avg} = \frac{1}{m}\sum_{k=1}^{m} S_k \quad (6)$$

where $m$ is number of all clusters.

### D. Artificial Neural Networks

Artificial neural networks is a simulation of human brain with computer program that can self-adjusting from learning the input values. The remarkable feature of this technique is that it consists of many nodes in the hidden layer in which parallel connections are effective for data classification [14]. Fig. 3 shows general architecture of artificial neural networks consisting of nodes and edges between nodes. Form the figure, the network can be partitioned based on node layout into 3 layers. The first layer is input layer; the second is hidden layer (this layer can have more than 1 layer), and the final layer is output layer.

## III. Proposed Work

In this section, we present the proposed process of silhouette width criterion consideration for automatic clustering of feature sets with the main focus of finding optimal feature to be discovered by association rule mining. The idea is that we use the silhouette coefficient to find the appropriate number of clusters for clustering the feature scores from feature selection obtained from the association rule mining. The objectives are to increase the predictive accuracy and to reduce the data dimensions of forest type dataset.

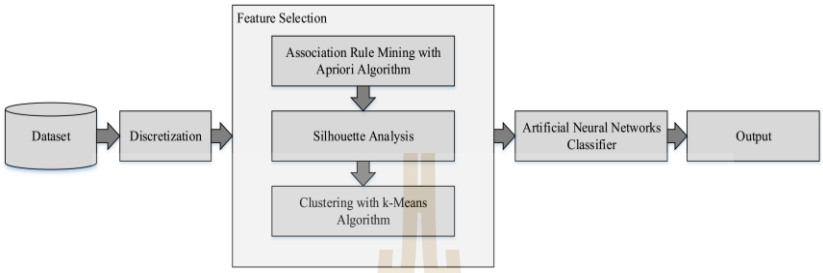



Fig. 4. The concept of feature selection based on association rule mining.

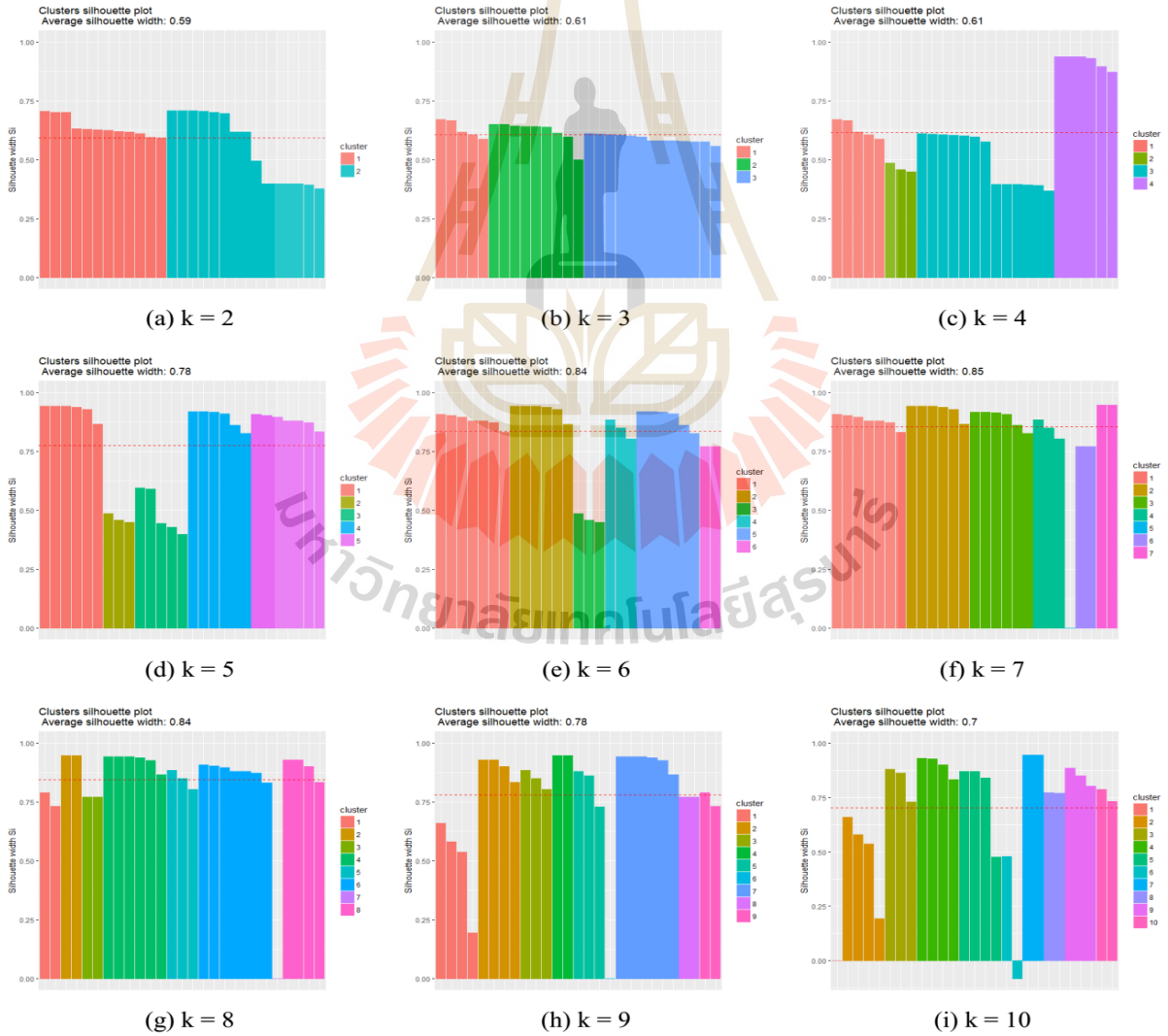



Fig. 5. Comparative graphs showing average silhouette widths of different cluster numbers.

Fig 4 shows the steps of the proposed process, which consists of three phases: phase 1, read the dataset from file or database and then perform discretization with chi-square algorithm if the data type is numeric. This discretization step is necessary for association rule mining that can handle only categorical values. Phase 2 is the feature selection method that

consists of three steps:

Step 1: find frequent patterns and generate from these patterns association rules in a format "*IF condition THEN consequence.*" This step is done through Apriori algorithm with initial 2 thresholds: support and confidence. We also constrain the algorithm to generate rules with target class appeared in the consequence part of the rule. The result from this step is a set of features with scores computed as feature frequency in association rules and average confidence of each feature.

Step 2: cluster the features based on their scores with different number of clusters. For each number of clusters, calculate the silhouette coefficient to find the best number of clusters. The higher silhouette coefficient means the better formation of clusters. The result in this step is the optimal parameter *k* to be used in the k-means clustering on step 3.

Step 3: perform k-means clustering with the initial number of cluster (k) according to the recommend value from step 2. We then select a set of features from a cluster showing mean confidence higher than other clusters. The result in this step is optimal feature set to be used for classification.

Finally, phase 3 is the building of classifier using artificial neural networks.

TABLE I: COMPARATIVE RESULTS OF CLASSIFICATION ACCURACY, NUMBER OF FEATURES, AND AVERAGE SILHOUETTE WIDTH

| Number of Clusters (k) | Accuracy | Number of Features | Average Silhouette Width |
|---|---|---|---|
| 2 | 80.31% | 20 | 0.59 |
| 3 | 81.54% | 14 | 0.61 |
| 4 | 82.77% | 11 | 0.61 |
| 5 | 82.77% | 11 | 0.78 |
| 6 | 82.77% | 11 | 0.84 |
| **7** | **84.62%** | 10 | **0.85** |
| 8 | 80.00% | 7 | 0.84 |
| 9 | 79.69% | 5 | 0.78 |
| 10 | 78.46% | **3** | 0.70 |

## IV. EXPERIMENTAL RESULTS

To test performance of the proposed method of feature selection based on the silhouette width criterion for clustering relevant featured discovered by association rule mining, we use the forest type with high-resolution imaging from ASTER satellite that has been publicly available at the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml). The data are divided into training dataset (198 instances) and test dataset (325 instances). We initialize Apriori algorithm to discover feature sets with support = 0.1 and confidence = 0.1. We experiment with number of clusters (k) between 2 to 10 clusters.

Table I shows comparative results of classification accuracy, number of features, and average silhouette. Fig. 5 shows comparative average silhouette widths of different clusters. It can be seen that when the number of cluster = 7, the average silhouette coefficient is maximized (0.85). At this maximum coefficient value, the predictive accuracy is as high as 84.62%. Moreover, the number of features can be reduced from 26 down to 10. Characteristic of number of features according to the changing number of clusters has been captured and shown in Fig. 6.

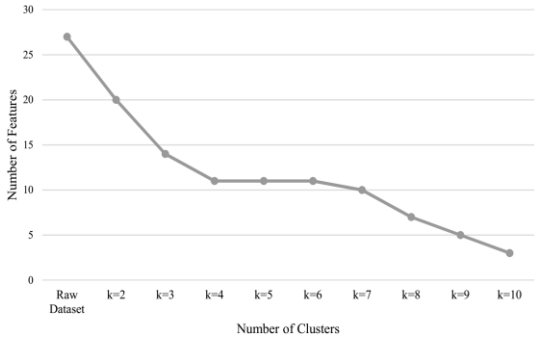



Fig. 6. The effect of cluster numbers to the number of features.

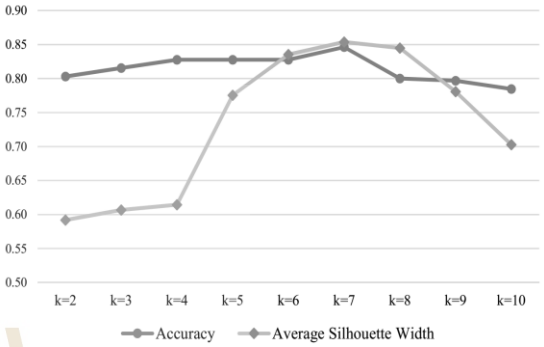



Fig. 7. The accuracy and average silhouette width characteristics.

Fig. 7 shows the comparisons of predictive accuracy and average silhouette width as the number of clusters has been varied from 2 to 10. It can be seen that the average silhouette has direct and positive impact to the classification accuracy. This is observed from the graph that when the silhouette width is low, the accuracy is also low. When the silhouette width is high, the accuracy is high as well.

## V. CONCLUSION

This research aims at studying a novel method to use silhouette width criterion for cluster analysis with the main focus of finding optimal feature set to be used for building classification model. Set of features are discovered with the association rule mining method. The proposed feature subset selection method is to be applied on classifying data with high dimensionality such as satellite image data. Our proposed method works with three main phases. Firstly, find and score relevant set of features based on association rule mining technique. Secondly, apply silhouette width criterion to find optimal parameter *k* for the next phase of feature clustering and add average confidence threshold of each cluster to feature score for increasing clustering performance. From the experimental results, we can conclude that the proposed method can select a discriminative set of features resulting in a highly accurate classification model.

## REFERENCES

[1] M. Hilario and A. Kalousis, "Approaches to dimensionality reduction in proteomic biomarker studies," *Briefings in Bioinformatics*, vol. 9, no. 2, pp. 102-118, 2008.

[2] Z. N. Hamilton, "Correlation-based feature subset selection for machine learning," Ph.D. Dissertation, Department of Computer Science, Waikato University, 1998.

[3] X. Jin, A. Xu, R. Bie, and P. Guo, "Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles," in *Proc. International Workshop on Data Mining for Biomedical Applications*, 2006, pp. 106-115.
[4] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of Machine Learning Research*, pp. 1205-1224, 2004.
[5] Y. Saeys, P. Rouzé, and Y. Van de Peer, "In search of the small ones: Improved prediction of short exons in vertebrates, plants, fungi and protists," *Bioinformatics*, vol. 23, no. 4, pp. 414-420, 2007.
[6] J. Xie, J. Wu, and Q. Qian, "Feature selection algorithm based on association rules mining method," *In Proc. the ACIS International Conference on Computer and Information Science*, pp. 357-3622009.
[7] N. Kaoungku, K. Suksut, R. Chanklan, K. Kerdprasop, and N. Kerdprasop, "Data classification based on feature selection with association rule mining," *The 25th Int. MultiConference of Engineers and Computer Scientists (IMECS2017)*, Hong Kong, China, 15-17 March 2017, pp. 321-326.
[8] N. Kaoungku, K. Kerdprasop, and N. Kerdprasop, "A method to clustering the feature ranking on data classification using an ensemble feature selection," *International Journal of Future Computer and Communication*, vol. 6, no. 3, September, pp. 81-85, 2017.
[9] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *ACM SIGMOD Record*, vol. 22, no. 2, pp. 207-216, 1993.
[10] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. The 20th Int. Conf. on Very Large Data Bases (VLDB)*, 1994, vol. 1215, pp. 487-499.
[11] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281-297.
[12] M. R. Anderberg, *Cluster Analysis for Applications*, Academic Press, 1973.
[13] S. Aranganayagi and K. Thangavel, "Clustering categorical data using silhouette coefficient as a relocating measure," in *Proc. IEEE Int. Conf. on Computational Intelligence and Multimedia Applications*, 2007, vol. 2, pp. 13-17.
[14] B. Yegnanarayana, *Artificial Neural Networks*, PHI Learning Pvt. Ltd., 2009.

**N. Kaoungku** is currently a lecturer at School of Computer Engineering, Suranaree University of Technology (SUT), Thailand. He received his doctoral degree, master degree, and bachelor degree in computer engineering from SUT, in 2015, 2013, and 2012, respectively. His current research includes data mining, knowledge engineering, and semantic web.

**K. Suksut** is currently a doctoral student with the School of Computer Engineering, SUT, Thailand. He received his bachelor degree in computer engineering from SUT in 2011, master degree in computer engineering from SUT in 2013. His current research of interest includes data mining, genetic algorithm, and imbalanced data classification.

**R. Chanklan** is currently a doctoral student with the School of Computer Engineering, SUT. She received her bachelor degree in computer engineering from SUT in 2013, master degree in Computer Engineering from SUT in 2014. Her current research of interest is data mining and artificial intelligence.

**K. Kerdprasop** is an associate professor and chair of Computer Engineering School, SUT. He received bachelor degree in mathematics from Srinakarinwirot University, Thailand, in 1986, MS in Computer Science from the Prince of Songkla University, in 1991, and PhD in computer science from Nova Southeastern University, U.S.A., in 1999.

**N. Kerdprasop** is an associate professor at the School of Computer Engineering, SUT. She received her bachelor degree in radiation techniques from Mahidol University, Thailand, in 1985, MS in Computer Science from the Prince of Songkla University in 1991, and PhD in computer science from Nova Southeastern University, U.S.A, in 1999.

# ภาคผนวก ข

ลิขสิทธิ์โปรแกรม

โปรแกรมรวมกฎความสัมพันธ์ด้วยกลไกการให้เหตุผลเชิงตรรกะ

(Program to integrate association rules with reasoning mechanism)

| | |
|---|---|
| ชื่อภาษาไทย | โปรแกรมรวมกฎความสัมพันธ์ด้วยกลไกการให้เหตุผลเชิงตรรกะ |
| ชื่อภาษาอังกฤษ | Program to integrate association rules with reasoning mechanism |
| ทะเบียนข้อมูลเลขที่ | ว1. 5360 |
| ให้ไว้ ณ วันที่ | 8 เมษายน พ.ศ. 2558 |
| คำอธิบายโปรแกรมโดยย่อ | โปรแกรมรวมกฎความสัมพันธ์ด้วยกลไกให้เหตุผลเชิงตรรกะหรือการอนุมาน พัฒนาด้วยภาษาไพธอน (Python) ทำงานในส่วนการอนุมานและรวมกฎที่กระจายอยู่ในหลายฐานข้อมูล การหากฎความสัมพันธ์แบบดั้งเดิมจะทำกับข้อมูลที่ถูกรวบรวมไว้ในแหล่งข้อมูลเพียงแหล่งเดียว ซึ่งสามารถหากฎความสัมพันธ์ได้แบบตรงไปตรงมา แต่ถ้าข้อมูลถูกกระจายตัวกันอยู่ตามแหล่งต่าง ๆ และด้วยข้อจำกัดของทรัพยากรทางด้านคอมพิวเตอร์ ไม่สามารถรวบรวมข้อมูลที่กระจายตัวกันอยู่ไว้ในแหล่งข้อมูลเพียงแหล่งเดียวเพื่อนำไปหากฎความสัมพันธ์ได้เนื่องจากขนาดของข้อมูลจะใหญ่กว่าขนาดของหน่วยความจำ จึงทำให้การหากฎความสัมพันธ์จะต้องใช้ลักษณะของการหากฎความสัมพันธ์แบบกระจาย แต่การหากฎความสัมพันธ์ในลักษณะนี้ทำได้ยาก เนื่องจากขั้นตอนการรวมกฎความสัมพันธ์นั้นอาจทำให้ได้กฎความสัมพันธ์ที่ขัดแย้งกันเอง หรือได้จำนวนกฎความสัมพันธ์ที่มากจนเกินไป หรือเกิดการขาดไปของกฎความสัมพันธ์ที่สำคัญ<br>ดังนั้นการพัฒนาโปรแกรมนี้ได้เสนอแนวทางใหม่ในการแก้ไขปัญหาการหากฎความสัมพันธ์แบบกระจาย โดยในขั้นตอนการรวมกฎความสัมพันธ์จะนำมาเฉพาะกฎความสัมพันธ์ที่ปรากฏขึ้นบ่อยในทุก ๆ แหล่งความรู้ แล้วนำเฉพาะกฎความสัมพันธ์ที่ได้ไปตรวจสอบความขัดแย้งและในขั้นตอนนี้สามารถสร้างกฎความสัมพันธ์ใหม่จากกฎความสัมพันธ์เดิมที่มีอยู่ด้วยวิธีการอนุมานซึ่งเป็นการให้เหตุผลเชิงตรรกะ ซึ่งสามารถเติมเต็มในส่วนของกฎความสัมพันธ์ที่ขาดหายไปได้ สุดท้ายจะได้กฎความสัมพันธ์ที่มีประสิทธิภาพเพียงพอสำหรับการนำไปทำนายผลข้อมูลในอนาคตและไม่เกิดความขัดแย้งกันเอง |

รลข.01 ทะเบียนข้อมูลเลขที่ ว1. 5360

# หนังสือรับรองการแจ้งข้อมูล
## ลิขสิทธิ์
## ออกให้เพื่อแสดงว่า
## มหาวิทยาลัยเทคโนโลยีสุรนารี

ได้แจ้งข้อมูลลิขสิทธิ์ ประเภทงาน วรรณกรรม

ลักษณะงาน โปรแกรมคอมพิวเตอร์

ชื่อผลงาน โปรแกรมรวมกฎความสัมพันธ์ด้วยกลไกการให้เหตุผลเชิงตรรกะ

ไว้ต่อกรมทรัพย์สินทางปัญญา ตามคำขอแจ้งข้อมูลลิขสิทธิ์ เลขที่ 322085

เมื่อวันที่ 1 เดือน เมษายน พ.ศ. 2558

ให้ไว้ ณ วันที่ 8 เดือน เมษายน พ.ศ. 2558

ลงชื่อ.....

นางสาวศิริวรรณ นพรัก

นักวิชาการพาณิชย์ปฏิบัติการ

ปฏิบัติราชการแทนผู้อำนวยการสำนักลิขสิทธิ์

**หมายเหตุ** 1. เอกสารนี้มิได้รับรองความเป็นเจ้าของลิขสิทธิ์

2. การเปลี่ยนแปลงรายการข้างต้น ให้ดูด้านหลัง

# ประวัติผู้วิจัย

รองศาสตราจารย์ ดร.นิตยา เกิดประสพ สำเร็จการศึกษาในระดับปริญญาเอกสาขา Computer Science จาก Nova Southeastern University เมือง Fort Lauderdale รัฐฟลอริดา สหรัฐอเมริกา เมื่อปีพุทธศักราช 2542 (ค.ศ. 1999) ด้วยทุนการศึกษาของกระทรวงวิทยาศาสตร์และเทคโนโลยี หลังสำเร็จการศึกษาได้ปฏิบัติราชการในตำแหน่งอาจารย์ ประจำสาขาคอมพิวเตอร์ ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ต่อมาในปีพุทธศักราช 2543 ได้มาปฏิบัติงานในตำแหน่งอาจารย์ประจำสาขาวิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี จนถึงปัจจุบัน งานวิจัยที่ทำในขณะนี้คือการประยุกต์เทคโนโลยีเหมืองข้อมูลกับงานด้านการแพทย์ การสาธารณสุขและสิ่งแวดล้อม รวมถึงการพัฒนาเทคนิคเพื่อเพิ่มความสามารถในการจัดการความรู้ของระบบเหมืองข้อมูล

รองศาสตราจารย์ ดร.กิตติศักดิ์ เกิดประสพ สำเร็จการศึกษาในระดับปริญญาเอกสาขา Computer Science จาก Nova Southeastern University เมือง Fort Lauderdale รัฐฟลอริดา สหรัฐอเมริกา เมื่อปีพุทธศักราช 2542 (ค.ศ. 1999) ด้วยทุนการศึกษาของทบวงมหาวิทยาลัย (หรือสำนักงานคณะกรรมการอุดมศึกษาในปัจจุบัน) หลังสำเร็จการศึกษาได้ปฏิบัติงานในตำแหน่งอาจารย์ประจำสาขาวิชาวิศวกรรมคอมพิวเตอร์ สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี ปัจจุบันดำรงตำแหน่งหัวหน้าหน่วยวิจัยวิศวกรรมความรู้ เน้นการวิจัยเกี่ยวกับการพัฒนาระบบเหมืองข้อมูลประสิทธิภาพสูง การประยุกต์เหมืองข้อมูลกับงานวิศวกรรม และการวิเคราะห์ข้อมูลเชิงสถิติ รวมถึงการวิจัยพื้นฐานเกี่ยวกับเทคนิคการวิเคราะห์ข้อมูลโดยวิธีอัตโนมัติ โดยมีผลงานวิจัยในด้านฐานข้อมูล การวิเคราะห์ข้อมูล การทำเหมืองข้อมูล และการค้นหาความรู้ ตีพิมพ์ในวารสารวิชาการและเอกสารการประชุมวิชาการทั้งระดับชาติและนานาชาติจำนวนมากกว่า 300 เรื่อง

อาจารย์ ดร.นันทวุฒิ คะอังกุ สำเร็จการศึกษาในระดับปริญญาเอกสาขาวิชาวิศวกรรมคอมพิวเตอร์ จากมหาวิทยาลัยเทคโนโลยีสุรนารี เมื่อปีพุทธศักราช 2558 ด้วยทุนการศึกษาสำหรับผู้มีศักยภาพเข้าศึกษาระดับบัณฑิตศึกษาของมหาวิทยาลัยเทคโนโลยีสุรนารี หลังสำเร็จการศึกษาได้ปฏิบัติงานในตำแหน่งอาจารย์ ประจำสาขาวิชาวิศวกรรมคอมพิวเตอร์ สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี เน้นการวิจัยเกี่ยวกับเทคนิคการเพิ่มประสิทธิภาพในการวิเคราะห์ข้อมูลโดยวิธีอัตโนมัติ และการพัฒนาเว็บเชิงความหมาย โดยมีผลงานวิจัยตีพิมพ์ในวารสารวิชาการและเอกสารการประชุมวิชาการจำนวนมากกว่า 20 เรื่อง

---