# ใบรับรองวิทยานิพนธ์

**บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์**

วิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)

**ปริญญา**

| วิศวกรรมคอมพิวเตอร์ | วิศวกรรมคอมพิวเตอร์ |
|---|---|
| **สาขา** | **ภาควิชา** |

**เรื่อง** การใช้เกณฑ์ความต่างลำดับในการปรับปรุงกฎความสัมพันธ์จำแนกประเภทข้อมูล

Order Difference Criterion for Class Association Rule Update

**นามผู้วิจัย** นายกฤษฎากร ก๋งอุบล

**ได้พิจารณาเห็นชอบโดย**

**อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก** .....

( รองศาสตราจารย์กฤษณะ ไวยมัย, Ph.D. )

**อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม** .....

( ผู้ช่วยศาสตราจารย์พีรวัฒน์ วัฒนพงศ์, Ph.D. )

**อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม** .....

( ผู้ช่วยศาสตราจารย์จิตร์ทัศน์ ฝักเจริญผล, Ph.D. )

**หัวหน้าภาควิชา** .....

( ผู้ช่วยศาสตราจารย์เขมะฑัต วิภาตะวนิช, Ph.D. )

**บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว**

.....

( รองศาสตราจารย์กัญจนา ธีระกุล, D.Agr. )

**คณบดีบัณฑิตวิทยาลัย**

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

วิทยานิพนธ์

เรื่อง

# การใช้เกณฑ์ความต่างลำดับในการปรับปรุงกฎความสัมพันธ์จำแนกประเภทข้อมูล

Order Difference Criterion for Class Association Rule Update


โดย

นายกฤษฎากร ก๋ำอุบล


เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์
เพื่อความสมบูรณ์แห่งปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)
พ.ศ. 2552

กฤษฎากร ก๋งอุบล 2552: การใช้เกณฑ์ความต่างลำดับในการปรับปรุงกฎความสัมพันธ์จำแนกประเภทข้อมูล ปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์) สาขาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก: รองศาสตราจารย์กฤษณะ ไวยมัย, Ph.D. 67 หน้า

เทคนิคการสร้างโมเดลจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์ (Associative Classification) เป็นเทคนิคที่รู้จักกันดี สำหรับการจำแนกประเภทข้อมูล แต่เมื่อมีการเพิ่มของข้อมูล โมเดลจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์จะเกิดปัญหาความแม่นยำของโมเดลเปลี่ยน การแก้ปัญหาความแม่นยำของโมเดลเปลี่ยน โดยการสร้างโมเดลจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์ขึ้นมาใหม่ยังใช้เวลามาก แต่ความแม่นยำเพิ่มขึ้นน้อย ผู้วิจัยจึงเสนอแนวทางใหม่เรียกว่า ICMF (Incremental Classifier Model Framework) สำหรับการปรับปรุงความแม่นยำของโมเดลจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์ เมื่อมีการเพิ่มของข้อมูล โดยแนวทาง ICMF ใช้ค่าลำดับต่างกัน OD (Order Difference) เป็นเกณฑ์ในการตัดสินใจว่า ควรจะปรับปรุงโมเดลหรือไม่ ค่าลำดับต่างกัน คือ ค่าความแตกต่างของลำดับของกฎในโมเดล ระหว่างก่อนและหลังการเพิ่มของข้อมูล จากผลการทดลองแสดงให้เห็นว่า เมื่อมีการเพิ่มของข้อมูล แนวทาง ICMF ที่นำเสนอ ให้ความแม่นยำของโมเดลสูงกว่า แนวทาง ที่ปรับปรุงโมเดลทุกครั้ง และแนวทาง ที่ไม่มีการปรับปรุงโมเดลเลย

_____ _____ ___ / ___ / ___

ลายมือชื่อนิสิต ลายมือชื่ออาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

Kritsadakorn Kongubol 2009: Order Difference Criterion for Class Association Rule Update. Master of Engineering (Computer Engineering), Major Field: Computer Engineering, Department of Computer Engineering. Thesis Advisor: Associate Professor Kitsana Waiyamai, Ph.D. 67 pages.

Associative classification is one of on well-known data classification techniques. With the increasing number of data, the problem of classification accuracy maintenance is of great importance. Updating the class association rules requires large amount of time, and the prediction accuracy is not increased that much. In this research work, a novel framework for Incremental Associative Classification (ICMF) is proposed. ICMF uses Order Difference (OD) criterion to decide whether to update the class association rules. Order Difference is the difference between rule order before and after updating model. Rules are updated only if new data modify to a certain extent the order of the rules on the basis of some interestingness measures which can be support or confidence. Experiment results show that ICMF yields better accuracy compared to the framework with class association rule update and without class association rule update.

_____ _____ ___ / ___ / ___

Student's signature Thesis Advisor's signature

# กิตติกรรมประกาศ

ข้าพเจ้าขอกราบขอบพระคุณอาจารย์กฤษณะ ไวยมัย อาจารย์ที่ปรึกษาวิทยานิพนธ์หลักที่ได้ประสิทธิ์ประสาทวิชาความรู้และทฤษฎีต่าง ๆ ตลอดจนเป็นที่ปรึกษาในการวางแนวทาง วางแผนงาน แก้ปัญหา และตรวจสอบข้อบกพร่องต่าง ๆ จนงานวิจัยนี้สำเร็จลุล่วง ขอกราบขอบพระคุณ อาจารย์พีรวัฒน์ วัฒนพงศ์ และ อาจารย์จิตร์ทัศน์ ฝักเจริญผล อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ตลอดจนอาจารย์ในภาควิชาวิศวกรรมคอมพิวเตอร์ ที่กรุณาให้คำปรึกษา และข้อเสนอแนะต่าง ๆ จนส่งผลให้งานวิจัยนี้สมบูรณ์ยิ่งขึ้นในทุก ๆ ด้าน และประสบผลสำเร็จลุล่วงไปด้วยดี

ขอขอบคุณอาจารย์ธนาวินท์ รักธรรมานนท์ ที่ให้คำปรึกษาที่ดีเสมอมา ไม่ว่าทิศทางของงานวิจัย จุดแข็งและจุดด้อย ของงานวิจัย ตลอดทั้ง ความรู้และทฤษฎีต่าง ๆ ที่จำเป็นสำหรับงานวิจัย ขอขอบคุณพี่ธนภัทร ฆังคะจิตร พี่สาวิณี แสงสุริยันต์ ที่ให้คำปรึกษาและคำอธิบายที่เป็นประโยชน์ต่องานวิจัย และสุดท้ายขอขอบคุณสมาชิกห้องปฏิบัติการ DAKDL คุณท่านที่คอยให้คำแนะนำและกำลังใจที่ดีเสมอมา

ขอขอบคุณเจ้าหน้าที่โครงการบัณฑิตศึกษา และเจ้าหน้าที่ธุรการภาควิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเกษตรศาสตร์ที่ช่วยเหลือในการประสานงาน และดำเนินงานด้านเอกสารต่างๆ ให้เป็นไปอย่างสะดวกลุล่วงไปด้วยดี

คุณงามความดีหรือประโยชน์อันใดเนื่องจากวิทยานิพนธ์เล่มนี้ ขออุทิศให้แก่ บิดา มารดา พี่ น้อง ญาติสนิทมิตรสหาย ตลอดทั้งครูอาจารย์และผู้มีพระคุณทุกท่าน ที่ได้อบรมและให้กำลังใจเสมอมาในทุก ๆ เรื่อง

กฤษฎากร ก๋ิงอุบล

เมษายน 2552

# สารบัญ

# สารบัญตาราง

# สารบัญตาราง (ต่อ)

# สารบัญภาพ

# สารบัญภาพ (ต่อ)

# คำอธิบายสัญลักษณ์และคำย่อ

| | | |
|---|---|---|
| Avg | = | Average |
| CARs | = | Class-Association Rules |
| CBA | = | Classification Based on Associations Algorithm |
| CBA-RG | = | Classification Based on Associations (Rule generator phase) |
| CBA-CB | = | Classification Based on Associations (Classifier builder phase) |
| CMAR | = | Classification based on Multiple Class-Association Rules Algorithm |
| Comp | = | Computational |
| CPAR | = | Classification based on Predictive Association Rules Algorithm |
| FP-Growth | = | Frequent Pattern Growth Algorithm |
| FP-tree | = | Frequent pattern tree |
| FUFI | = | Fast Update of Frequent Itemsets Algorithm |
| FUP | = | Fast Update Algorithm |
| FUP**2** | **=** | Fast Update **2** Algorithm |
| FW | = | Framework |
| ICMF | = | Incremental Classifier Model Framework |
| LUCS-KDD | = | Liverpool University Computer Science - Knowledge Discovery in Datas |
| LUCS-KDD DN | = | Liverpool University Computer Science - Knowledge Discovery in Datas Discretization Normalisation Software |
| MOD | = | Maximum Order Difference |
| NUWEP | = | Negative border Update With Early Pruning Algorithm |
| OD | = | Order Difference |

# การใช้เกณฑ์ความต่างลำดับในการปรับปรุงกฎความสัมพันธ์จำแนกประเภทข้อมูล

## Order Difference Criterion for Class Association Rule Update

## คำนำ

เทคนิคการสร้างโมเดลจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์ ที่มีอยู่ในปัจจุบัน เช่น CBA (Liu *et al.*, 1998) , CMAR (Li *et al.*, 2001), CPAR (Yin and Han., 2003) จะแบ่งการทำงานเป็นสองขั้นตอนคือ ขั้นตอนแรก สร้างชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูล Class-Association Rules (CARs) ของโมเดลจากข้อมูลเรียนรู้ (Training data) ขั้นตอนที่สอง นำชุดกฎความสัมพันธ์จำแนกประเภทข้อมูลมาสร้างโมเดลการทำนายจากข้อมูลทดสอบ (Testing data) แต่อย่างไรก็ตามเมื่อข้อมูลเรียนรู้มีการเพิ่มขึ้น ชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูล ซึ่งพิจารณาค่าสนับสนุนเป็นเกณฑ์หลัก จะได้รับผลกระทบ จาก ค่าสนับสนุน (Support) ที่เปลี่ยนแปลง และเป็นสาเหตุทำให้ความแม่นยำของโมเดลเปลี่ยนแปลง ฉะนั้นเพื่อให้โมเดลยังคงรักษาระดับความแม่นยำในการทำนาย จึงจำเป็นต้องปรับปรุงและสร้างชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูล สำหรับโมเดลขึ้นมาใหม่ แต่จากการศึกษางานวิจัยต่างๆ พบว่า งานวิจัยส่วนใหญ่ FUP (Cheung, 1996), FUP2 (Cheung *et al.*, 1997), NUWEP (Imberman *et al.*, 2004), FUFI (กฤษฎากร และคณะ , 2550), (Thabtah, 2006, 2007) มุ่งเน้นไปขั้นตอนแรก คือ การปรับปรุงและสร้างชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดลขึ้นมาใหม่ เมื่อมีการเพิ่มของข้อมูล ซึ่งยังมีข้อเสียคือใช้เวลามาก และอาจจะไม่ทำให้ความแม่นยำของโมเดลเปลี่ยนแปลงหรือดีขึ้น จึงจำเป็นต้องมีแนวทางเพื่อพิจารณาว่า เมื่อมีการเพิ่มของข้อมูล ควรจะปรับปรุงชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลโมเดลทุกครั้งหรือไม่ เพื่อลดเวลาของการปรับปรุงโมเดล และยังคงรักษาระดับความแม่นยำให้สูงใกล้เคียงกับโมเดลเดิม กรณีมีการเพิ่มของข้อมูลแต่ละครั้ง

แนวทางของการปรับปรุงชุดกฎความสัมพันธ์จำแนกประเภทข้อมูลเมื่อมีการเพิ่มของข้อมูลจะมีสองแนวทางคือ แนวทางแรก จะใช้ชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลจากโมเดลเดิมตลอด ไม่สนใจการปรับปรุงและสร้างชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลสำหรับโมเดลขึ้นมาใหม่ ข้อดีก็คือมีความเร็วสูงในการทำงาน เพราะไม่เสียเวลาปรับปรุงและสร้างชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลใหม่ให้กับโมเดล แต่มีข้อเสียคือ ความแม่นยำจะลดลง และไม่สนับสนุนแนวคิดการปรับปรุงกฎของโมเดล เมื่อมีการเพิ่มของข้อมูล (Incremental

Associative Classification) สำหรับแนวทางที่สอง เมื่อมีการเพิ่มของข้อมูล จะสร้างหรือปรับปรุงชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดลทั้งหมดใหม่ จากข้อมูลเดิมและข้อมูลที่เพิ่มเข้ามา ซึ่งมีข้อดีก็คือชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดล ยังคงรักษาระดับความแม่นยำสูงในการทำนายได้ และมีความแม่นยำสูงกว่าแนวทางแรก แต่เทคนิคเหล่านี้ก็ยังมีข้อเสียคือ ใช้เวลามาก ในการสร้างกฎหรือปรับปรุงชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลใหม่ทุกครั้งเมื่อมีการเพิ่มของข้อมูล

ในวิทยานิพนธ์เล่มนี้ ผู้วิจัยได้นำเสนอแนวทางในการลดจำนวนครั้งของการปรับปรุงโมเดลเมื่อมีข้อมูลเพิ่มเข้ามา โดยใช้ค่าเกณฑ์ ลำดับต่างกัน OD (Order Difference) เพื่อเป็นเกณฑ์ในการพิจารณาว่า ข้อมูลที่เพิ่มเข้ามา มีผลต่อความแม่นยำของโมเดลและจำเป็นต้องปรับปรุงกฎของโมเดลหรือไม่ ซึ่งจะมีผลดีคือลดเวลาของการปรับปรุงโมเดล และยังคงรักษาระดับความแม่นยำให้สูงใกล้เคียงกับโมเดลเดิม กรณีมีการเพิ่มของข้อมูล

# วัตถุประสงค์และขั้นตอนการวิจัย

## วัตถุประสงค์ของการวิจัย

1. ศึกษาเทคนิคการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ และเทคนิคอื่นๆที่เกี่ยวข้อง เพื่อที่จะพัฒนาเทค นิคการจำแนกประเภทข้อมูลที่สนับสนุนการเพิ่มของข้อมูล ได้อย่างเหมาะสมและมีประสิทธิภาพ

2. พัฒนาเทคนิคการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ ที่สนับสนุนการเพิ่มของข้อมูล ในส่วนของ

2.1 เกณฑ์ตัดสินใจ เพื่อพิจารณาการสร้างกฎความสัมพันธ์ของโมเดลขึ้นมาใหม่หรือใช้ของเดิม

2.2 ความแม่นยำโดยเฉลี่ยของโมเดลการทำนาย

2.3 เวลาที่ใช้ในการทำงานของระบบ

## ขั้นตอนการวิจัย

1. ศึกษาทฤษฎีต่างๆ ของเทคนิคการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ รวมถึงศึกษาทฤษฎีที่เกี่ยวข้อง เพื่อที่จะนำความรู้ที่ได้มาใช้ในการวิจัย

2. รวบรวมฐานข้อมูลมาตรฐานเพื่อที่จะนำมาใช้ในการทดสอบเพื่อศึกษาผลจากโมเดลการทำนายและหาสาเหตุของปัญหาที่เกิดขึ้น

3. ศึกษาผลที่ได้จากการทดลอง เพื่อวิเคราะห์ปัญหาและรวบรวมข้อดีและข้อด้อยต่างๆ ของงานก่อนหน้านี้ เพื่อนำมาเป็น ข้อมูลในการพัฒนาเทคนิคและอัลกอริทึมในการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์

4. พัฒนาเทคนิคการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ ที่สนับสนุนการเพิ่มของข้อมูล

5. ทดสอบและวัดผลของการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ ที่สนับสนุนการเพิ่มของข้อมูล

6. สรุปผลการวิจัยและประโยชน์ที่ได้รับ

# การตรวจเอกสาร

## ความรู้พื้นฐานของเทคนิคดาต้าไมนิ่ง

### 1. การสืบค้นกฎความสัมพันธ์ (Association Rule Discovery)

การสืบค้นกฎความสัมพันธ์ (Association Rule Discovery) เป็นหนึ่งในเทคนิคของ Data mining ที่มีความสำคัญเทคนิคหนึ่ง โดยวิธีการสืบค้นกฎความสัมพันธ์นี้ เปรียบเสมือนกับการค้นหาทองจากเหมืองข้อมูลขนาดใหญ่ ซึ่งทองที่ได้กล่าวถึงนั้นก็คือ กฎ ที่มีความน่าสนใจ ที่บ่งบอกถึงลักษณะเฉพาะหรือคุณสมบัติเด่นของเหมืองข้อมูลหรือฐานข้อมูลนั้นๆ โดยที่เราไม่สามารถที่จะค้นหาได้มาก่อน โดยหลักการทำงานของเทคนิคนี้คือ การค้นหากฎความสัมพันธ์ของข้อมูลจากข้อมูลขนาดใหญ่ที่มีอยู่ เพื่ อนำกฎที่ได้เหล่านั้นไปวิเคราะห์เพื่อใช้ช่วยในการตัดสินใจ โดยส่วนใหญ่แล้วจะนำไปใช้ทางด้านธุรกิจ (Business decision making) เช่น การนำเทคนิคนี้ไปวิเคราะห์พฤติกรรมของลูกค้าที่ซื้อสินค้าในซุปเปอร์มาเก็ต (Market basket analysis) โดยดูว่าลูกค้ามักจะซื้อสินค้าอะไรด้วยกัน เพื่อที่จะนำข้อมูลการซื้อสินค้าของลูกค้าเหล่านั้นมาช่วยในการวางแผนทางการตลาด เช่น การจัดวางสินค้าที่มักจะถูกซื้อด้วยกันไว้ใกล้ๆ กันหรือการจัดโปรโมชั่นให้กับสินค้าเป็นต้น

หนึ่งในอัลกอริทึมในการสืบค้นกฎความสัมพันธ์ที่รู้จักกันดี นั่นคือ อัลกอริทึม Apriori (Rakesh and Srikant, 1994; Rakesh *et al.*, 1993) ซึ่งหลักการคือ จะทำการคำนวณหาความสัมพันธ์ของ Itemsets ที่มักจะเกิดขึ้นพร้อมๆกันในฐานข้อมูล โดยความสัมพันธ์ของ Itemsets นั้นเรียกว่ากฎความสัมพันธ์ (Association rule) ซึ่งจะอยู่ในรูปแบบดังต่อไปนี้

{Item1, Item2} -> {Item3}; ค่าสนับสนุน (Support), ค่าความเชื่อมั่น (Confidence)

อัลกอริทึม Apriori จะต้องมีการกำหนดค่าสนับสนุนขั้นต่ำ (Minimum support) และค่าความเชื่อมั่นขั้นต่ำ (Minimum confidence) ด้วยซึ่งในการกำหนดค่าขั้นต่ำทั้งสองค่านี้ จะขึ้นอยู่กับผู้ใช้ระบบเป็นผู้กำหนดเอง หรือจะใช้ผู้เชี่ยวชาญ (Expert user) เป็นผู้กำหนดให้ก็ได้ โดยกฎความสัมพันธ์ที่ได้นั้นจะต้องมีค่าสนับสนุน (Support) และค่าความเชื่อมั่น(Confidence) ไม่น้อยกว่าค่าขั้นต่ำที่ได้กำหนดเอาไว้ข้างต้น ค่าสนับสนุน (Support) คือ เปอร์เซ็นต์ของจำนวน Itemsets ทั้งหมดที่เกิดขึ้นในฐานข้อมูล ค่าความเชื่อมั่น (Confidence) คือ เปอร์เซ็นต์ของจำนวน Itemsets ทั้งหมดที่เกิดขึ้นในฐานข้อมูล ต่อ จำนวน Itemsets ที่เกิดขึ้นทางด้านซ้ายมือของกฎ

ตัวอย่างการใช้เทคนิค Association Rule Discovery (วีระพล, 2549) ในการค้นหากฎความสัมพันธ์ของข้อมูลโดยในตารางที่ 1 คือ ตัวอย่างชุดข้อมูลการซื้อสินค้า ซึ่งคอลัมน์ TID เปรียบเสมือนตะกร้าที่ใส่สินค้าที่ซื้อในครั้งหนึ่งๆ และคอลัมน์ Items คือรายการสินค้าที่ซื้อพร้อมกันใน TID ใดๆ และตัวอักษร A, B, C, D, และ E แทนชื่อสินค้าแต่ละชนิด โดยที่กำหนดค่าสนับสนุนขั้นต่ำ (Minimum support) เท่ากับ 50% และค่าความมั่นใจขั้นต่ำ (Minimum confidence) เท่ากับ 70%

**ตารางที่ 1** ตัวอย่างข้อมูลรายการซื้อสินค้า

| TID | Items |
|---|---|
| 1 | A C D |
| 2 | B C E |
| 3 | A B C E |
| 4 | B E |
| 5 | A B C E |

ข้อมูลในตารางที่ 1 ถูกนำเข้าสู่กระบวนการสร้างกฎความสัมพันธ์ โดยขั้นตอนวิธีการสร้างกฎความสัมพันธ์ สามารถดูได้จาก ตารางที่ 2 ถึง ตารางที่ 4 และกฎความสัมพันธ์ที่ได้ทั้งหมดสามารถดูได้จากตารางที่ 5

**ตารางที่ 2** อัลกอริทึม Apriori ทำการค้นหา Frequent itemsets รอบแรก

| 1-itemsets $C_1$ | 1-itemsets $C_1$ | Count | % | Large 1-itemsets $L_1$ | Count | % |
|---|---|---|---|---|---|---|
| {A} | {A} | 3 | 60 | {A} | 3 | 60 |
| {B} | {B} | 4 | 80 | {B} | 4 | 80 |
| {C} | {C} | 4 | 80 | {C} | 4 | 80 |
| {D} | {D} | 1 | 20 | | | |
| {E} | {E} | 4 | 80 | {E} | 4 | 80 |
| a) Generate phase | b1) Count phase | | | b2) Select phase | | |

**ตารางที่ 3** อัลกอริทึม Apriori ทำการค้นหา Frequent itemsets รอบที่สอง

| 2-itemsets $C_2$ | 2-itemsets $C_2$ | Count | % | Large 2-itemsets $L_2$ | Count | % |
|---|---|---|---|---|---|---|
| {A, B} | {A, B} | 2 | 40 | | | |
| {A, C} | {A, C} | 3 | 60 | {A, C} | 3 | 60 |
| {A, E} | {A, E} | 2 | 40 | | | |
| {B, C} | {B, C} | 3 | 60 | {B, C} | 3 | 60 |
| {B, E} | {B, E} | 4 | 80 | {B, E} | 4 | 80 |
| {C, E} | {C, E} | 3 | 60 | {C, E} | 3 | 60 |
| a) Generate phase | b1) Count phase | | | b2) Select phase | | |

**ตารางที่ 4** อัลกอริทึม Apriori ทำการค้นหา Frequent itemsets รอบที่สาม

| 3-itemsets $C_3$ | 3-itemsets $C_3$ | Count | % | Large 3-itemsets $L_3$ | Count | % |
|---|---|---|---|---|---|---|
| {B, C, E} | {B, C, E} | 3 | 60 | {B, C, E} | 3 | 60 |
| a) Generate phase | b1) Count phase | | | b2) Select phase | | |

**ตารางที่ 5** กฎความสัมพันธ์ทั้งหมดที่ถูกสร้างโดย อัลกอริทึม Apriori

| กฎความสัมพันธ์ | ค่าสนับสนุน (%) | ค่าความเชื่อมั่น(%) |
|---|---|---|
| {BC} => {E} | 60 | 3/3 = 100 |
| {CE} => {B} | 60 | 3/3 = 100 |
| {BE} => {C} | 60 | 3/4 = 75 |
| {B} => {CE} | 60 | 3/4 = 75 |
| {C} => {BE} | 60 | 3/4 = 75 |
| {E} => {BC} | 60 | 3/4 = 75 |
| {A} => {C} | 60 | 3/3 = 100 |
| {B} => {C} | 60 | 3/4 = 75 |
| {B} => {E} | 80 | 4/4 = 100 |
| {C} => {E} | 60 | 3/4 = 75 |

ตารางที่ 5 แสดงกฎความสัมพันธ์ทั้งหมดที่ถูกสร้างโดย อัลกอริทึม Apriori ซึ่งข้อมูลในตารางจะประกอบไปด้วย กฎความสัมพันธ์ ค่าสนับสนุนของกฎความสัมพันธ์ (Support) และค่าความมั่นใจของกฎความสัมพันธ์ (Confidence)

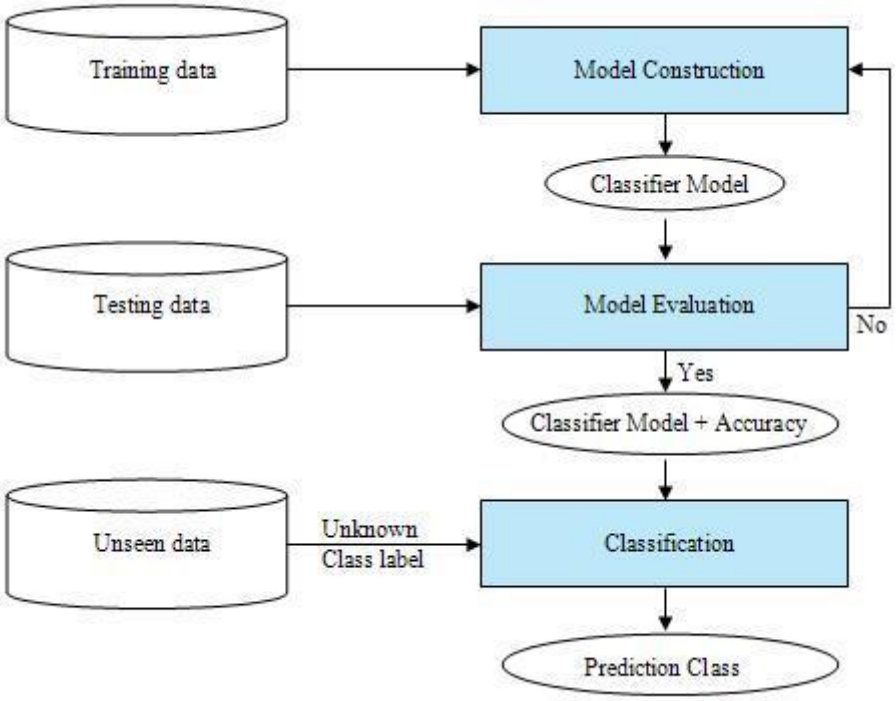
ตัวอย่างกฎความสัมพันธ์ในตารางที่ 5 กฎความสัมพันธ์ {BC} => {E} มีค่าสนับสนุนเท่ากับ 60% และค่าความเชื่อมั่นเท่ากับ 100% หมายความว่า จำนวนครั้งการซื้อสินค้าที่มีการซื้อ B, C, และ E พร้อมกันมีจำนวน 3 ครั้ง จากจำนวนรายการทั้งหมด และความน่าจะเป็นที่เมื่อมีการซื้อสินค้า B และ C พร้อมกันแล้ว จะซื้อสินค้า E ด้วยเสมอ คิดเป็น 100%

**2. การจำแนกประเภทข้อมูล (Data Classification)**

การจำแนกประเภทข้อมูล(Data Classification) เป็นอีกหนึ่งเทคนิคใน Data Mining ซึ่งทำหน้าที่สืบค้นความรู้เพื่อสรุปหาแบบจำลองหรือโมเดลของฐานข้อมูลนั้นๆ (Quinlan, 1993; Wang *et al.*, 2000) เพื่อใช้ในการทำนายข้อมูลใหม่ (unseen data) โดยเทคนิคนี้จะหาความสัมพันธ์ของข้อมูลจากฐานข้อมูลขนาดใหญ่ เพื่อนำมาสร้างโมเดลเพื่อใช้ในการจำแนกประเภทข้อมูล ซึ่งจะสามารถนำไปจำแนกประเภทข้อมูลใหม่ๆ ที่ยังไม่ทราบประเภทได้ (Unknown class label)

เนื่องจากการจำแนกประเภทข้อมูลเป็นเทคนิคแบบ Supervise learning นั่นคือ การจะสร้างโมเดลออกมาได้นั้นจะต้องทำการสอนระบบเสียก่อน ดังนั้นจำเป็นจะต้องทราบจำนวนคลาสปลายทาง (Class label) และจำนวนแอตทริบิวต์ (Attribute) ที่แน่นอน และส่วนของข้อมูลจะต้องแบ่งออกเป็นสองส่วน ส่วนหนึ่งใช้สอนระบบ (Training data) อีกส่วนหนึ่งใช้ทดสอบความแม่นยำของโมเดลที่ถูกสร้างออกมา (Testing data) โดยปรกติสัดส่วนระหว่าง Training กับ Testing จะอยู่ที่ประมาณ 70 ต่อ 30 โดยในการที่จะสร้างโมเดลออกมาเพื่อใช้สำหรับทำนายข้อมูลได้นั้น จะต้องผ่านขั้นตอนดังต่อไปนี้ เริ่มจากการนำข้อมูลสอนระบบ (Training data) เข้ามาสู่กระบวนการสร้างโมเดลจำแนกประเภทข้อมูล (Model construction) เพื่อให้ได้โมเดลจำแนกประเภทข้อมูล (Classifier model) ออกมา และหลังจากได้โมเดลจำแนกประเภทข้อมูลแล้ว วิธีการทดสอบว่าโมเดลที่ถูกสร้างขึ้นมามีความแม่นยำมากเพียงพอที่จะนำไปใช้ได้หรือไม่นั้น จะใช้ข้อมูลทดสอบระบบ หรือ Testing data เพื่อทดสอบความแม่นยำของโมเดลที่ถูกสร้างขึ้นมา (Model evaluation) ถ้าโมเดลที่สร้างขึ้นมามีความแม่นยำไม่ผ่านเกณฑ์ที่กำหนด จะต้องกลับไปปรับปรุงในส่วนของกระบวนการสร้างโมเดลจำแนกประเภทข้อมูลเสียก่อน แต่ถ้าโมเดลที่สร้างขึ้นมามีความแม่นยำผ่านเกณฑ์ที่ต้องการแล้ว ก็สามารถที่จะนำโมเดลที่สร้างมานั้นไปประยุกต์ใช้เพื่อทำนายประเภทข้อมูลใหม่ (Unseen data) ที่ไม่ทราบประเภทของข้อมูล (Unknown class label) ต่อไปได้

โดยภาพรวมขั้นตอนทั้งหมดของเทคนิคการจำแนกประเภทข้อมูลที่อธิบายไว้ข้างต้น สามารถดูได้จากภาพที่ 1




**ภาพที่ 1** ขั้นตอนของเทคนิคการจำแนกประเภทข้อมูล

**ที่มา:** วีระพล (2549)

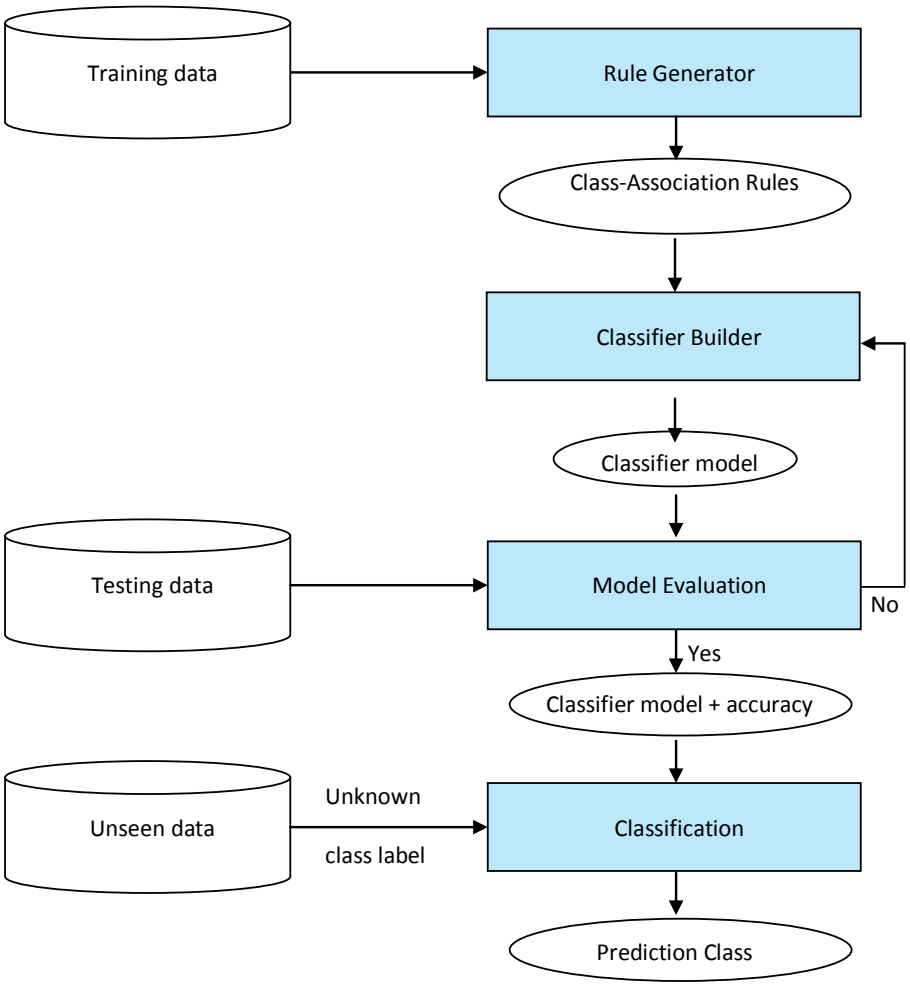
**3. การจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ (Associative Classification)**

การจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ (Associative classification) เป็นเทคนิคที่เกิดจากการรวมกันระหว่าง 2 เทคนิค (Liu *et al.*, 1998) ที่ได้กล่าวในหัวข้อข้างต้น คือ การจำแนกประเภทข้อมูล (Data classification) และ การสืบค้นกฎความสัมพันธ์ (Association rule discovery) โดยวัตถุประสงค์ของเทคนิคการจำแนกประเภทข้อมูลคือ ค้นหาโมเดลหรือเซตของกฎความสัมพันธ์ที่เล็กที่สุดในฐานข้อมูล เพื่อสร้างโมเดลจำแนกประเภทข้อมูลที่มีความถูกต้องแม่นยำมากที่สุด และ วัตถุประสงค์ของเทคนิคการสืบค้นกฎความสัมพันธ์คือ ค้นหากฎความสัมพันธ์ทั้งหมดที่มีความสำคัญและบ่งบอกถึงคุณลักษณะของฐานข้อมูล โดยที่กฎเหล่านั้นจะต้องผ่านค่าสนับสนุนและค่าความเชื่อมั่นขั้นต่ำด้วย เทคนิคการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ (Associative classification) แบ่งออกเป็น 2 ส่วนหลักๆ คือ ส่วนที่ใช้ในการสร้างกฎความสัมพันธ์ (Rule generator phase) และส่วนที่นำกฎความสัมพันธ์ไปสร้างโมเดลเพื่อใช้ทำนายข้อมูล (Classifier builder phase)

ส่วนของการสร้างกฎความสัมพันธ์ (Rule generator phase) ใช้หลักการหรือวิธีการเดียวกันกับเทคนิค Association rule discovery เกือบทั้งหมด ยกเว้นกฎที่ถูกสร้างจากกระบวนการสร้างกฎความสัมพันธ์นั้นจะต้องเป็นกฎเฉพาะที่เรียกว่า กฎความสัมพันธ์จำแนกประเภทข้อมูล หรือ CARs (Class-Association Rules) นั่นคือกฎความสัมพันธ์ที่สับเซตของกฎทางด้านขวามือจะต้องเป็นคลาสแอททริบิวต์ (Class Attribute) เท่านั้น เช่น กฎ {A, B, C → **Class**} ด้านขวาของกฎเป็นคลาส อัลกอริทึมการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ที่รู้จักกันดี (Liu *et al.*, 1998; Li *et al.*, 2001) จะค้นหา CARs ทั้งหมดที่ผ่านค่าสนับสนุนขั้นต่ำ (Minimum support) และค่าความมั่นใจขั้นต่ำ (Minimum confidence)

ส่วนของการสร้างโมเดลในการทำนาย (Classifier builder phase) จะนำกฎความสัมพันธ์ที่ได้จากส่วนการสร้างกฎมาใช้เพื่อสร้างโมเดลในการทำนายข้อมูล โดยในการทำนายข้อมูลนั้น จะมีการพิจารณาแบ่งออกเป็น 2 วิธี วิธีที่ 1 จะทำการพิจารณากฎความสัมพันธ์ทีละกฎ (Single rule) โดยวิธีการพิจารณาแบบนี้ จะต้องทำการเรียงลำดับกฎความสัมพันธ์ ที่มีศักย์ (Precedence) สูงก่อน การเรียงลำดับของกฎจะเรียงลำดับเริ่มจากกฎความสัมพันธ์ที่มีค่าความเชื่อมั่น(Confidence) สูงก่อน แต่ถ้าค่าความเชื่อมั่นของกฎความสัมพันธ์เท่ากัน ก็จะเรียงลำดับของกฎความสัมพันธ์ตามค่าสนับสนุน (Support) แต่ถ้าทั้งค่าความเชื่อมั่นและ ค่าสนับสนุนของกฎเกิดเท่ากันอีก ก็จะเรียงลำดับ

กฎโดยดูจาก กฎไหนถูกส ร้างมาก่อน ก็จะเรียงกฎนั้นก่อนตามลำดับ หลังจากเรียงลำดับกฎความสัมพันธ์เป็นที่เรียบร้อยแล้วก็พร้อมที่จะทำนายข้อมูล โดยการทำนายข้อมูลนั้นจะทำนายตาม class ของกฎที่มีศักย์ (Precedence) สูงที่สุด ส่วนวิธีที่ 2 จะทำการพิจารณากฎความสัมพันธ์ทีละหลายๆกฎพร้อมกัน (Multiple rules) โดยการทำนายข้อมูลนั้นจะนำกลุ่มของกฎที่มีอยู่ในคลาสเดียวกัน มาคำนวณผ่านสูตรที่ได้กำหนดเอาไว้แล้วดูว่าคลาสไหนที่ให้ค่ามากที่สุดคลาสนั้นก็จะเป็นคำตอบ โดยที่ภาพรวมขั้นตอนทั้งหมดของเทคนิคการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ สามารถดูได้จากภาพที่ 2




**ภาพที่ 2** ขั้นตอนของเทคนิคการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์

**ที่มา:** วีระพล (2549)

**4. ค่าวัดผลที่น่าสนใจ (Interesting measures)**

อัลกอริทึมต่างๆ เช่น Apriori, FP-Growth (Han *et al.*, 2004) ในเทคนิคดาต้าไมน์นิ่ง ได้เสนอวิธีการค้นหากฎความสัมพันธ์จากฐานข้อมูล แต่อย่างไรก็ตามกฎความสัมพันธ์ที่ได้จากอัลกอริทึมต่างๆ เหล่านี้มีจำนวนมากมาย จนไม่สามารถวิเคราะห์กฎทั้งหมดได้ จึงมีการนำเสนอค่าวัดผลต่างๆ (Interesting measures) (Sheikh *et al.*, 2004; Geng and Hamilton., 2006; Lenca *et al.*, 2008) สำหรับการเลือกกฎที่มีความสำคัญ ถูกต้องน่าเชื่อถือ หรือน่าสนใจจริงๆ ไปใช้ ในที่นี้จะสนใจค่าวัดผลดังต่อไปนี้

1. ค่าสนับสนุน (Support)
2. ค่าความเชื่อมั่น (Confidence)
3. ค่าคอนวิคชั่น (Conviction)
4. ค่าลิฟท์ (Lift)

**ตารางที่ 6** ตัวอย่างฐานข้อมูลทรานเซ็กชั่น

| Transaction ID | Items |
|---|---|
| 1 | A B |
| 2 | B Y |
| 3 | C |
| 4 | A B Y |
| 5 | B |

ตารางที่ 6 กฎความสัมพันธ์แบบมีคลาส (CARs) จะถูกกำหนดโดยรูปแบบกฎดังตัวอย่างข้างล่าง

A -> Y

เมื่อ

A= {A, B, C}

Y= {Y}

เมื่อคำนวณค่าวัดผลสำหรับกฎความสัมพันธ์แบบมีคลาส ค่าวัดผลแต่ละแบบจะคำนวณค่าอยู่บนความน่าจะเป็น ตามสมการข้างล่าง

$$P(A) = count(A)/|D|$$

$$P(Y) = count(Y)/|D|$$

$$P(AY) = count(AY)/|D|$$

โดยที่

1. count(A) คือจำนวนทรานเซ็กชั่น (transaction) หรือเร็คคอร์ดในฐานข้อมูลที่ประกอบไปด้วยไอเท็ม A
2. count(Y) คือจำนวนทรานเซ็กชั่น (transaction) หรือเร็คคอร์ดในฐานข้อมูลที่ประกอบไปด้วยไอเท็ม Y
3. count(XY) คือจำนวนทรานเซ็กชั่น (transaction) หรือเร็คคอร์ดในฐานข้อมูลที่ประกอบไปด้วยไอเท็ม A และ Y
4. |D| คือจำนวนทรานเซ็กชั่น (transaction) หรือเร็คคอร์ดในฐานข้อมูล

4.1 ค่าสนับสนุน (Support)

ค่าสนับสนุน (Rakesh *et al.*, 1993) คือ เปอ ร์เซ็นต์ของจำนวน Itemsets ทั้งหมดที่เกิดขึ้นในฐานข้อมูล เขียนอยู่ในรูปสมการที่ (1)

$$Support(A) = P(A) \quad \text{.....(1)}$$

จากตารางที่ 6 เราสามารถคำนวณค่าสนับสนุนได้เช่น

P(A) = 2/5 = 0.4

P(B) = 4/5 = 0.8

P(C) = 1/5 = 0.2

P(Y) = 2/5 = 0.4

4.2 ค่าความเชื่อมั่น (Confidence)

ค่าความเชื่อมั่น (Rakesh *et al.*, 1993) คือ เปอร์เซ็นต์ของจำนวน Itemsets ของกฎที่เกิดขึ้นในฐานข้อมูล ต่อ จำนวน Itemsets ที่เกิดขึ้นทางด้านซ้ายมือของกฎ เขียนอยู่ในรูปสมการที่ (2)

$$Confidence(A -> Y) = P(A\ and\ Y)/P(A) \quad \text{..... (2)}$$

จากตารางที่ 6 เราสามารถคำนวณค่าความเชื่อมั่นได้เช่น

P(A->Y) = (1/5)/(2/5) = 0.5

P(B->Y) = (2/5)/(2/5) = 1

P(AB->Y) = (1/5)/(2/5) = 0.5

4.3 ค่าคอนวิคชั่น (Conviction)

ค่าคอนวิคชั่น (Brin *et al.*, 2004) คือ ผลคูณระหว่างเปอร์เซ็นต์ของจำนวน Itemsets ทางด้านซ้ายมือของกฎและ เปอร์เซ็นต์ของจำนวน Itemsets ที่เป็นคลาสแอททริบิวต์ ยกเว้น Itemsets ที่เป็นคลาสแอททริบิวต์ทางด้านขวามือของกฎ ต่อ เปอร์เซ็นต์ของ Itemsets ทางด้านซ้ายมือของกฎ ที่ไม่มีคลาส แอททริบิวต์ทางด้านขวามือของกฎประกอบอยู่ด้วย เขียนอยู่ในรูปสมการที่ (3)

$$\text{Conviction}(A \rightarrow Y) = P(A)P(\bar{Y})/P(A \text{ and } \bar{Y}) \quad \text{..... (3)}$$

จากตารางที่ 6 เราสามารถคำนวณค่าคอนวิคชั่นได้เช่น

P(A->Y) = (2/5)*(3/5)/(1/5) = 1.2

P(B->Y) = (3/5)*(3/5)/(1/5) = 1.8

P(AB->Y) = (2/5)*(3/5)/(1/5) = 1.2

4.4 ค่าลิฟท์ (Lift)

ค่าลิฟท์ (Brin *et al.*, 2004) คือเปอร์เซ็นต์ของจำนวน Itemsets ของกฎที่เกิดขึ้นในฐานข้อมูล ต่อ จำนวน Itemsets ที่เกิดขึ้นทางด้านซ้ายมือของกฎ และ จำนวน Itemsets ที่เกิดขึ้นทางด้านขวามือของกฎ เขียนอยู่ในรูปสมการที่ (4)

$$Lift(X -> Y) = \frac{P(X \text{ and } Y)}{P(X)P(Y)} \quad \text{..... (4)}$$

จากตารางที่ 6 เราสามารถคำนวณค่าความลิฟท์ได้เช่น

P(A->Y) = (1/5)/((2/5)*(2/5)) = 0.5

P(B->Y) = (2/5)/((4/5)*(2/5)) = 1.25

P(AB->Y) = (1/5)/((2/5)*(2/5)) = 0.2

# งานวิจัยที่เกี่ยวข้อง

ในวิทยานิพนธ์เล่มนี้ผู้วิจัยนำเสนองานวิจัยสำหรับการสร้างโมเดลจำแนกประเภทข้อมูล โดยใช้กฎความสัมพันธ์ (Association Classification) CBA (Liu *et al.*, 1998) และแนวทางสำหรับการปรับปรุงชุดกฎความสัมพันธ์จำแนกประเภทข้อมูลเมื่อมีการเพิ่มของข้อมูล ซึ่งแนวทางทั่วไปจะมีสองแนวทางคือ แนวทางแรก (Framework 1) จะใช้ชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลจากโมเดลเดิมตลอด และแนวทางที่สอง (Framework 2) เมื่อมีการเพิ่มของข้อมูล จะสร้างหรือปรับปรุงชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดลทั้งหมดใหม่ จากข้อมูลเดิมและข้อมูลที่เพิ่มเข้ามา โดยในการนำเสนอ จะอธิบายรายละเอียดพร้อมตัวอย่างของแต่ละอัลกอริทึมหรือแนวทางด้วย

## 1. CBA อัลกอริทึม

Classification Based on Associations (CBA) เป็นงานแรกในการเสนอวิธีการรวมอัลกอริทึมการสืบค้นกฎความสัมพันธ์เข้ากับเทคนิคการจำแนกประเภทข้อมูล โดยวิธีการแบบนี้ถูกเรียกว่า การจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ (Associative Classification) ซึ่งอัลกอริทึม CBA ประกอบด้วย 2 ส่วนคือ ส่วนที่หนึ่ง การสร้างกฎความสัมพันธ์ (Rule generator) และส่วนที่สอง การสร้างโมเดลในการทำนายข้อมูล (Classifier builder)

### 1.1 การสร้างกฎความสัมพันธ์ (CBA-Rule Generator)

ขั้นตอนแรก การสร้างกฎความสัมพันธ์ของอัลกอริทึม CBAจะสร้างเหมือนกับเทคนิค Association Rule Discovery, Apriori เกือบทั้งหมด ยกเว้นกฎที่ถูกสร้างจากกระบวนการสร้างกฎความสัมพันธ์นั้นจะต้องเป็นกฎที่ประกอบด้วยคลาส (Class label) ที่เรียกว่าชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูล (Class-Association Rules หรือ CARs) โดยการสร้างกฎความสัมพันธ์ (Rules) จะเริ่มจากการนับความถี่ของชุดข้อมูลที่เกิดในฐานข้อมูลเรียนรู้ เรียกว่าไอเท็มเซต (itemsets) แล้วนำไอเท็มเซตเหล่านี้ มาสร้างกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดล โดยพิจารณาเฉพาะกฎความสัมพันธ์จำแนกประเภทข้อมูล ที่มีความถี่ผ่านค่าสนับสนุนขั้นต่ำ (Minimum support หรือ Minsup) และค่าความเชื่อมั่นขั้นต่ำ (Minimum confidence หรือ Minconf) แต่อย่างไรก็ตามการสร้างกฎความสัมพันธ์จากอัลกอริทึมค้นหากฎความสัมพันธ์

อย่างเช่น อัลกอริทึม Apriori หรือ FP-Growth ไม่รองรับการเพิ่มของข้อมูล จึงมีการนำ อัลกอริทึมการสร้างกฎความสัมพันธ์ ที่สนับสนุนการเพิ่มของข้อมูล FUP , FUP2 , NUWEP , FUFI มาใช้ โดยอัลกอริทึมเหล่านี้ จะพยายามเก็บและเลือกไอเท็มเซตที่น่าสนใจในการหากฎความสัมพันธ์ครั้งแรก เพื่อนำไปปรับปรุงกฎความสัมพันธ์ในครั้งต่อไป เพื่อลดเวลาในการปรับปรุงกฎความสัมพันธ์ทั้งหมด แต่อย่างไรก็ตามอัลกอริทึมเหล่านี้ไม่ได้ลดจำนวนรอบของการปรับปรุงโมเดลการทำนาย เมื่อมีการเพิ่มของข้อมูล ทำให้เวลาโดยรวมยังมาก เมื่อเทียบการลดจำนวนรอบของการปรั บปรุงโมเดลการทำนาย

### 1.2 การสร้างโมเดลในการทำนายข้อมูล (CBA-Classifier Builder)

ขั้นตอนที่สอง การสร้างโมเดลในการทำนายข้อมูล จะนำชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลมาพิจารณาว่ากฎไหน สมควรที่จะนำมาสร้างโมเดลในการทำนายข้อมูล โดยเทคนิค CBA จะเรียงลำดับชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูล ด้วยเกณฑ์ของค่าความเชื่อมั่น (Confidence) ค่าสนับสนุน (Support) แล้วนำชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลที่ได้จากการเรียงลำดับมาสร้างเป็นโมเดลการทำนาย โดยแนวคิดการเรียงลำดับชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลสามารถดูได้จากหัวข้อ 1.2.1 และเมื่อสร้างโมเดลการทำนายเสร็จ จะทดสอบความแม่นยำของโมเดลกับข้อมูลทดสอบ (Testing data) เพื่อหาค่าความแม่นยำของโมเดล

#### 1.2.1 แนวคิดการเรียงลำดับชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูล

กำหนดให้มี 2 กฎ นั่นคือกฎ $r_i$ และ $r_j$ เราสามารถกำหนดว่า $r_i \succ r_j$ ($r_i$ มีศักย์ลำดับ (Precedence)สูงกว่า $r_j$)ได้ก็ต่อเมื่อ

1. ค่าความเชื่อมั่น (Confidence) ของ $r_i$ มากกว่า $r_j$ หรือ
2. ค่าความเชื่อมั่น (Confidence) เท่ากัน แต่ค่าสนับสนุน (Support) ของ $r_i$ มากกว่า $r_j$ หรือ
3. ทั้งค่าความเชื่อมั่น (Confidence) และค่าสนับสนุน (Support) เท่ากัน แต่ $r_i$ ถูกสร้างมาก่อน $r_j$

1.3 ตัวอย่างการทำงานของอัลกอริทึม CBA

ตัวอย่างการทำงานของอัลกอริทึม CBA ตั้งแต่การค้นหากฎความสัมพันธ์ (CBA-RG) และการสร้างโมเดลในการทำนายข้อมูล (CBA-CB) ตารางที่ 7 คือ ตัวอย่าง ข้อมูลจากฐานข้อมูลทรานเซ็กชั่น ประกอบด้วย 6 คอลัมน์ มีคลาสปลายทาง คือ X และ Y

กำหนดค่าสนับสนุนขั้นต่ำ (Minimum support) เท่ากับ 2/8=25% และค่าความเชื่อมั่นขั้นต่ำ (Minimum confidence) เท่ากับ 50%

**ตารางที่ 7** ฐานข้อมูลทรานเซ็กชั่น

| Transaction ID | Itemsets |
|---|---|
| 1 | A B C X |
| 2 | A X |
| 3 | B X |
| 4 | A C D Y |
| 5 | C Y |
| 6 | D Y |
| 7 | B C X |
| 8 | A D Y |

การสร้างกฎความสัมพันธ์ (CBA-RG) เริ่มต้นจากตัวอย่างข้อมูลทรานเซ็กชั่นในตารางที่ 7 โดยขั้นตอนวิธีการสร้างกฎความสัมพันธ์ สามารถดูได้จาก ตารางที่ 8 ถึง ตารางที่ 10 และกฎความสัมพันธ์ที่ได้ทั้งหมดที่ผ่านค่าสนับสนุนขั้นต่ำสามารถดูได้จากตารางที่ 11 ส่วนตารางที่ 12 จะแสดงกฎความสัมพันธ์ที่ผ่านค่าสนับสนุนขั้นต่ำและค่าความมั่นใจขั้นต่ำ หลังจากนั้นอัลกอริทึม CBA-RG จะทำการจัดเรียงกฎความสัมพันธ์จำแนกประเภทข้อมูล (CARs) ตามค่าความเชื่อมั่น โดยดูจากตารางที่ 13

**ตารางที่ 8** อัลกอริทึม CBA-RG ทำการค้นหา Frequent itemsets รอบแรก

| 1-itemsets $C_1$ | 1-itemsets $C_1$ | Count | Support % | Large 1-itemsets $L_1$ | Count | Support % |
|---|---|---|---|---|---|---|
| {A} | {A} | 4 | 50 | {A} | 4 | 50 |
| {B} | {B} | 3 | 37.5 | {B} | 3 | 37.5 |
| {C} | {C} | 4 | 50 | {C} | 4 | 50 |
| {D} | {D} | 3 | 37.5 | {D} | 3 | 37.5 |
| {X} | {X} | 4 | 37.5 | {X} | 4 | 37.5 |
| {Y} | {Y} | 4 | 37.5 | {Y} | 4 | 37.5 |
| a) Generated phase | b1) Counted phase | | | b2) Selected phase | | |

ตารางที่ 8 แสดงขั้นตอนของอัลกอริทึม CBA-RG ทำการค้นหาไอเท็มเซต ที่มีค่าสนับสนุนมากกว่า ค่าสนับสนุนขั้นต่ำ ซึ่งได้กำหนด ไว้คือ 25 เปอร์เซ็นต์ โดยไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำ จะถูกนำไปใช้ในการหาไอเท็มเซตในระดับที่สอง ถัดไป

**ตารางที่ 9** อัลกอริทึม CBA-RG ทำการค้นหา Frequent itemsets รอบที่สอง

| 2-itemsets $C_2$ | 2-itemsets $C_2$ | Count | Support % | Large 2-itemsets $L_2$ | Count | Support % |
|---|---|---|---|---|---|---|
| {A, B} | {A, B} | 1 | 12.5 | | | |
| {A, C} | {A, C} | 2 | 25 | {A, C} | 2 | 25 |
| {A, D} | {A, D} | 2 | 25 | {A, D} | 2 | 25 |
| {A, X} | {A, X} | 2 | 25 | {A, X} | 2 | 25 |
| {A, Y} | {A, Y} | 2 | 25 | {A, Y} | 2 | 25 |
| {B, C} | {B, C} | 2 | 25 | {B, C} | 2 | 25 |
| {B, D} | {B, D} | 0 | 0 | | | |
| {B, X} | {B, X} | 3 | 37.5 | {B, X} | 3 | 37.5 |
| {B, Y} | {B, Y} | 0 | 0 | | | |
| {C, D} | {C, D} | 1 | 12.5 | | | |
| {C, X} | {C, X} | 2 | 25 | {C, X} | 2 | 25 |
| {C, Y} | {C, Y} | 2 | 25 | {C, Y} | 2 | 25 |
| {D, X} | {D, X} | 0 | 0 | | | |
| {D, Y} | {D, Y} | 3 | 37.5 | {D, Y} | 3 | 37.5 |
| a) Generated phase | b1) Counted phase | | | b2) Selected phase | | |

ตารางที่ 9 แสดงขั้นตอนของอัลกอริทึม CBA-RG ทำการค้นหาไอเท็มเซตในระดับที่สอง ที่มีค่าสนับสนุนมากกว่า ค่าสนับสนุนขั้นต่ำ โดยไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำ จะถูกนำไปใช้ในการหาไอเท็มเซตในระดับที่สาม ถัดไป

**ตารางที่ 10** อัลกอริทึม CBA-RG ทำการค้นหา Frequent itemsets รอบที่สาม

| 3-itemsets $C_c$ | 3-itemsets $C_3$ | Count | Support % | Large 3-itemsets $L_3$ | Count | Support % |
|---|---|---|---|---|---|---|
| {A, C, D} | {A, C, D} | 1 | 12.5 | | | |
| {A, C, X} | {A, C, X} | 1 | 12.5 | | | |
| {A, C, Y} | {A, C, Y} | 1 | 12.5 | | | |
| {A, D, X} | {A, D, X} | 0 | 0 | | | |
| {A, D, Y} | {A, D, Y} | 2 | 25 | {A, D, Y} | 2 | 25 |
| {A, X, Y} | {A, X, Y} | 0 | 0 | | | |
| {B, C, X} | {B, C, X} | 2 | 25 | {B, C, X} | 2 | 25 |
| {C, X, Y} | {C, X, Y} | 0 | 0 | | | |
| a) Generated phase | b1) Counted phase | | | b2) Selected phase | | |

ตารางที่ 10 แสดงขั้นตอนในการหาไอเท็มเซตในระดับที่สาม เมื่อเราพิจารณาในคอลัมน์ Large 3-itemsets พบว่าเหลือไอเท็มเซตที่ผ่านค่าสนับสนุนในระดับที่สามเพียงสองตัวคือ {A,D,Y} และ {B,C,X} ฉะนั้นการสร้างไอเท็มเซตในระดับที่สี่ไม่สามารถทำได้เนื่องจากการสร้างไอเท็มเซตในระดับที่สี่ต้องมีไอเท็มเซ็ตอย่างน้อยกว่าสามตัวขึ้นไป ระดับที่สามนี้จึงเป็นระดับสุดท้ายของการสร้างไอเท็มเซ็ต

**ตารางที่ 11** CBA-RG สร้าง Class-Association Rules (CARs) ที่ผ่านค่าสนับสนุนขั้นต่ำ 25%

| CARs | Support (%) | Confidence (%) |
|---|---|---|
| {A, X} | 25 | 2/4=50 |
| {A, Y} | 25 | 2/4=50 |
| {B, X} | 37.5 | 3/3=100 |
| {C, X} | 25 | 2/4=50 |
| {C, Y} | 25 | 2/4=50 |
| {D, Y} | 37.5 | 3/3=100 |
| {A, D, Y} | 25 | 2/4=50 |
| {B, C, X} | 25 | 2/4=50 |

ตารางที่ 11 แสดงกฎความสัมพันธ์จำแนกประเภทข้อมูล (CARs) ที่ผ่านค่าสนับสนุนขั้นต่ำ 25 % ไอเท็มเซ็ตในคอลัมน์ CARs อย่างเช่น {A, X} สามารถเขียนในรูปกฎความสัมพันธ์จำแนกประเภทข้อมูลคือ A->X

**ตารางที่ 12** CBA-RG สร้าง Class-Association Rules (CARs) ที่ผ่านค่าความเชื่อมั่นขั้นต่ำ 50%

| CARs | Support (%) | Confidence (%) |
|---|---|---|
| {A, X} | 25 | 2/4=50 |
| {A, Y} | 25 | 2/4=50 |
| {B, X} | 37.5 | 3/3=100 |
| {C, X} | 25 | 2/4=50 |
| {C, Y} | 25 | 2/4=50 |
| {D, Y} | 37.5 | 3/3=100 |
| {A, D, Y} | 25 | 2/4=50 |
| {B, C, X} | 25 | 2/4=50 |

ตารางที่ 12 แสดงกฎความสัมพันธ์จำแนกประเภทข้อมูล (CARs) ในทุกๆ ระดับของไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำ และค่าเชื่อมั่นขั้นต่ำ หลังจากตารางที่ 12 จะเป็นการจบในส่วน

ของการสร้างกฎความสัมพันธ์ (CBA-RG) ซึ่งในขั้นตอนถัดไปนั้นจะเป็นขั้นตอนในส่วนการสร้างโมเดลในการทำนาย (CBA-CB)

**ตารางที่ 13** เรียงกฎตามค่าความเชื่อมั่น

| RID | CARs | Support (%) | Confidence (%) |
|---|---|---|---|
| 1 | {B, X} | 37.5 | 3/3=100 |
| 2 | {D, Y} | 37.5 | 3/3=100 |
| 3 | {A, X} | 25 | 2/4=50 |
| 4 | {A, Y} | 25 | 2/4=50 |
| 5 | {C, X} | 25 | 2/4=50 |
| 6 | {C, Y} | 25 | 2/4=50 |
| 7 | {A, D, Y} | 25 | 2/4=50 |
| 8 | {B, C, X} | 25 | 2/4=50 |

ตารางที่ 13 แสดงกฎความสัมพันธ์แบบมีคลาส (CARs) ที่ถูกจัดเรียงใหม่ตามค่าความมั่นใจเพื่อที่จะนำกฎเหล่านี้ไปใช้ในการสร้างโมเดลในการทำนายต่อไป

**ตารางที่ 14** CBA-CB สร้างโมเดลในการทำนาย

| รอบการทำงาน | โมเดลในการทำนาย | ความถูกต้อง (%) |
|---|---|---|
| 1 | R1 = {r1, Default class = Y} | 90% |
| 2 | R1 = {r1, r2, Default class = X} | 90% |
| 3 | R1 = {r1, r2, r3, Default class = Y} | 100% |

จากตารางที่ 14 เราจะเห็นว่าอัลกอริทึม CBA-CB หรือส่วนในการสร้างโมเดลในการทำนาย จะมีการสร้างโมเดลออกมาจำนวนหลายชุด ดังนั้นจะต้องทำการเลือกชุดโมเดลที่ดีที่สุดโดยดูจากค่าเปอร์เซ็นต์ความถูกต้องของโมเดลชุดนั้นๆ แต่ถ้ามีหลายชุดโมเดลที่มีความถูกต้อง

เท่ากัน ก็จะพิจารณาตามหลักเกณฑ์ที่ว่า จะเลือกโมเดลที่มีเซตของกฎ ความสัมพันธ์แบบมีคลาส (CARs) ที่สั้นที่สุด ดังนั้นโมเดลที่สั้นที่สุดและมีความถูกต้องมากที่สุด นั่นคือโมเดลชุดที่ R3

## 2. แนวทางการสร้างโมเดลเมื่อมีการเพิ่มของข้อมูล (Frameworks of Associative Classification Model for Incremental databases)

แนวทางทั่วไปในการสร้างหรือปรับปรุงชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดล เมื่อมีการเพิ่มของข้อมูล สามารถแบ่งได้เป็น 2 แนวทางคือ

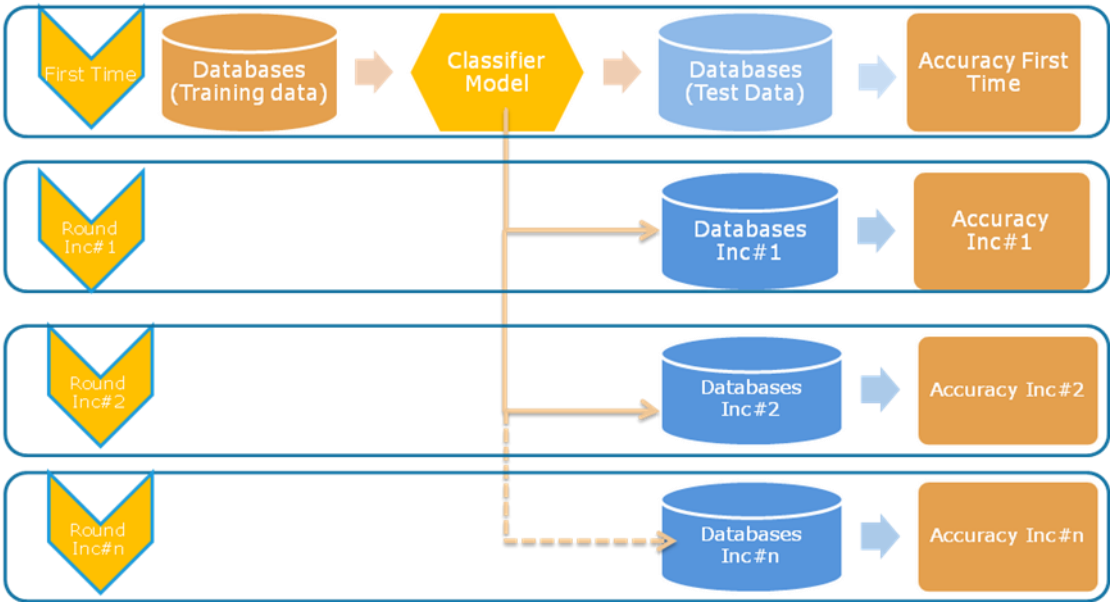
1. แนวทางที่ 1 (Framework 1) ไม่ปรับปรุงชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดลใหม่
2. แนวทางที่ 2 (Framework 2) ปรับปรุงชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดลใหม่

**ตารางที่ 15** ตารางค่าพารามิเตอร์สำหรับอธิบายสัญลักษณ์ในภาพที่ 3 และ 4

| ตัวแปร | คำอธิบาย |
|---|---|
| DB | ฐานข้อมูลเรียนรู้ |
| Inc#n | ข้อมูลใหม่ที่เพิ่มเข้ามาในรอบที่ n อย่างเช่น รอบที่ 1, Inc#1 รอบที่ n, Inc#n |
| db | ข้อมูลใหม่ที่เพิ่มเข้ามา |
| DB+db | ฐานข้อมูลรวม(ข้อมูลเรียนรู้ + ข้อมูลที่เพิ่มเข้ามา) |
| $L_2$ | กฎจำแนกประเภทข้อมูลที่มีคลาสและความยาวสอง |

2.1 แนวทางที่ 1 (Framework 1) : ไม่ปรับปรุงชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดลใหม่

แนวทางที่ 1 จะใช้ชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลจากโมเดลเดิม ในการทำนายทุกรอบของการเพิ่มของข้อมูล ไม่มีการปรับปรุงและสร้างชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดลขึ้นมาใหม่ ดังแสดงในภาพที่ 3 และดูค่าพารามิเตอร์จากตารางที่ 15 ประกอบ

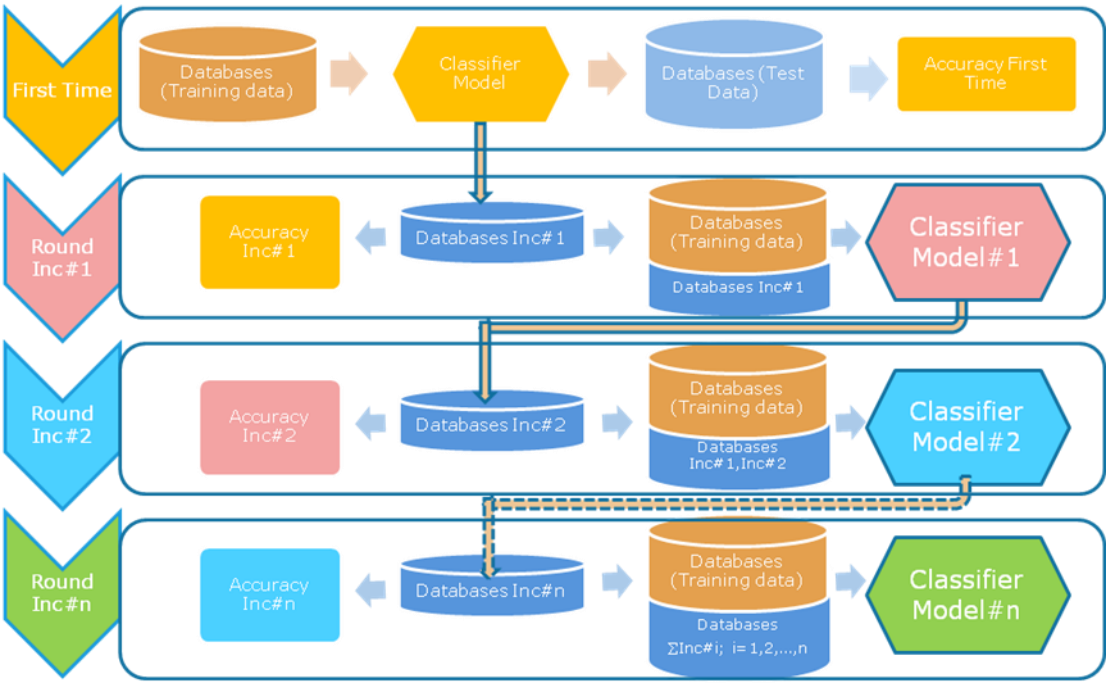



**ภาพที่ 3** แนวทางที่ 1 ไม่มีการปรับปรุงชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดลใหม่ เมื่อมีการเพิ่มของข้อมูล

ภาพที่ 3 แสดงโมเดลที่ถูกสร้างขึ้นมาในรอบแรก (First Time) จะนำไปใช้ทดสอบกับข้อมูลที่เพิ่มเข้ามาแต่ละรอบ และได้ความแม่นยำของโมเดลคืนกลับมา โดยที่ทุกรอบของการเพิ่มของข้อมูล จะใช้โมเดลเดิมทำนายตลอด ไม่มีการปรับปรุงโมเดล

2.2 แนวทางที่ 2 (Framework 2) ปรับปรุงชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดลใหม่

แนวทางที่สอง จะสร้างหรือปรับปรุงชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดลทั้งหมดใหม่จากข้อมูลรวม (ข้อมูลเรียนรู้และข้อมูลที่เพิ่มเข้ามา) ทุกรอบของการเพิ่มของข้อมูล ดังแสดงในภาพที่ 4 และดูค่าพารามิเตอร์จากตารางที่ 15 ประกอบ




**ภาพที่ 4** แนวทางที่ 2 ปรับปรุงชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูล ของโมเดลใหม่ทุกครั้งที่มีการเพิ่มของข้อมูล

ภาพที่ 4 ในรอบแรก(First Time) โมเดลจะถูกสร้างขึ้นมา และนำโมเดล ที่ได้ไปทดสอบ กับข้อมูลที่เพิ่มเข้ามาในรอบที่ 1 (Inc#1) หลังจากการทดสอบก็จะได้ความแม่นยำของโมเดลคืนกลับมา (Accuracy Inc#1) แล้วนำข้อมูลเรียนรู้และข้อมูลที่เพิ่มเข้ามา (Databases&Inc#1) มาสร้างเป็นโมเดลใหม่ (Classifier Model#1) เมื่อข้อมูลรอบถัดไปเพิ่มเข้า (Inc#2) ก็นำโมเดลใหม่ (Classifier Model#1) ไปทดสอบก็จะได้ความแม่นยำของโมเดลคืนกลับมา (Accuracy Inc#2) แล้วนำข้อมูลเรียนรู้และข้อมูลที่เพิ่มเข้ามา (Databases&Inc#1&Inc#2) มาสร้างเป็นโมเดลใหม่ (Classifier Model#2) สำหรับทดสอบรอบถัดไป (Inc#n) ทำแบบนี้ไปเรื่อยๆ ทุกรอบของการเพิ่มข้อมูล

**3. สรุปข้อดีและข้อเสียของอัลกอริทึมที่เกี่ยวข้องกับงานวิจัย**

3.1 อัลกอริทึม CBA

อัลกอริทึม CBA เป็นงานวิจัยแรก ที่เสนอ อัลกอริทึมการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ โดยมีส่วนการทำงานที่สำคัญอยู่ 2 ส่วน คือ ส่วนของการสร้างกฎความสัมพั นธ์ (Rule generator phase) และส่วนของการสร้างโมเดลในการทำนาย (Classifier builder phase โดยอัลกอริทึมดังกล่าวมีข้อดีและข้อเสีย ดังต่อไปนี้

3.1.1 ข้อดีของอัลกอริทึม CBA

1. เป็นอัลกอริทึมที่สามารถศึกษาทำความเข้าใจได้ง่าย
2. แบ่งการทำงานเป็นสองขั้นตอน คือ ขั้นตอนที่หนึ่ง การสร้างกฎความสัมพันธ์(Rule generator) และขั้นตอนที่สอง การสร้างโมเดลในการทำนายข้อมูล (Classifier builder) สำหรับขั้นตอนที่หนึ่งสามารถประยุกต์ใช้กับอัลกอริทึมการสร้างกฎความสัมพันธ์แบบเพิ่มเติม (Incremental Association Rules) ที่สนับสนุนการเพิ่มของข้อมูล (Incremental data) เช่น อัลกอริทึม FUP FUP2 NUWEP และ FUFI
3. อัลกอริทึม CBA สามารถนำไปใช้กับแนวทางที่ 1 ไม่ปรับปรุงชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดลใหม่ ทุกรอบของการเพิ่มของข้อมูล หรือแนวทางที่ 2 ปรับปรุงชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดลใหม่ทุกรอบของการเพิ่มของข้อมูล

3.1.2 ข้อเสียของอัลกอริทึม CBA

1. อัลกอริทึม CBA จะสร้างโมเดลจากกฎความสัมพันธ์จำแนกประเภทข้อมูล(CARs) แต่เมื่อมีการเพิ่มขึ้นของข้อมูล กฎความสัมพันธ์จำแนกประเภทข้อมูลจะเปลี่ยนไป จึงจำเป็นต้องปรับปรุงกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดลใหม่ทุกครั้ง แต่การ

ปรับปรุงกฎความสัมพันธ์ของโมเดลให้สอดคล้องกับข้อมูลที่เพิ่มเข้ามา อาจไม่คุ้มค่าเพราะใช้เวลามาก และความแม่นยำของโมเดลอาจเพิ่มขึ้นเล็กน้อย และถึงแม้ว่าจะนำเทคนิคการสร้างกฎความสัมพันธ์แบบเพิ่มเติมมาใช้เพื่อลดเวลาการปรับปรุงกฎความสัมพันธ์ เวลาโดยรวมก็ยังถือว่ามากอยู่

2. การนำอัลกอริทึม CBA ไปใช้กับแนวทางที่ 1 ไม่สามารถแก้ปัญหาความแม่นยำของโมเดลที่เปลี่ยนแปลงในแต่ละรอบของการเพิ่มของข้อมูลได้

3. การนำอัลกอริทึม CBA ไปใช้กับแนวทางที่ 2 ไม่สามารถแก้ปัญหาเวลาในการปรับปรุงกฎความสัมพันธ์จำแนกประเภทข้อมูลในแต่ละรอบของการเพิ่มของข้อมูลให้ลดลงได้

## 3.2 แนวทางที่ 1 (Framework 1)

แนวทางที่ 1 (Framework 1) ไม่ปรับปรุงชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดลใหม่ ทุกรอบของการเพิ่มของข้อมูล แนวทางที่ 1 จะใช้ชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลจากโมเดลเดิม ในการทำนายข้อมูลทุกรอบของการเพิ่มของข้อมูล

### 3.2.1 ข้อดีของแนวทางที่ 1

1. แนวทางที่ 1 ให้เวลาในการประมวลผลน้อย เพราะไม่ต้องเสียเวลาปรับปรุงและสร้างชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลใหม่ให้กับโมเดล

### 3.2.2 ข้อเสียของแนวทางที่ 1

1. เมื่อนำแนวทางที่ 1 ไปทำนายข้อมูลที่เพิ่มเข้ามา ความแม่นยำของโมเดลจะลดลง และไม่สนับสนุนแนวคิดการปรับปรุงกฎของโมเดล เมื่อมีการเพิ่มของข้อมูล (Incremental Associative Classification)

3.3 แนวทางที่ 2 (Framework 2)

แนวทางที่ 2 (Framework 2) ปรับปรุงชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดลใหม่ ทุกรอบของการเพิ่มของข้อมูล การสร้างหรือปรับปรุงชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดลทั้งหมดใหม่จะใช้ข้อมูลจากข้อมูลรวม (ข้อมูลเรียนรู้และข้อมูลที่เพิ่มเข้ามา)

3.3.1 ข้อดีของแนวทางที่ 2

1. แนวทางที่ 2 สร้างชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดลขึ้นมาใหม่ ทุกรอบของการเพิ่มขอข้อมูล ทำให้ผลระดับความแม่นยำในการทำนายสูงขึ้นและมีความแม่นยำสูงกว่าแนวทางที่ 1

3.3.2 ข้อเสียของแนวทางที่ 2

1. แนวทางที่ 2 ต้องสร้างกฎหรือปรับปรุงชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลใหม่ทุกรอบของการเพิ่มของข้อมูล ทำให้ใช้เวลาในการประมวลผลมาก

# อุปกรณ์และวิธีการ

## อุปกรณ์

### 1. ฮาร์ดแวร์

1.1 เครื่องคอมพิวเตอร์โน๊ตบุค 1 เครื่อง ประกอบด้วยอุปกรณ์ดังต่อไปนี้

1. ซีพียู (CPU) อินเทล คอร์ทูดูโอ ความเร็ว 2.1 GHz
2. หน่วยความจำหลัก 4 GB
3. ฮาร์ดดิสก์ขนาด 300 GB

### 2. ซอฟต์แวร์

2.1 ระบบปฏิบัติการ Windows Vista Ultimate SP1

2.2 Eclipse

## วิธีการ

### 1. ภาพรวมของระบบ

จากข้อดีและข้อเสียของของการนำอัลกอริทึม CBA มาใช้กับแนวทางที่ 1 และแนวที่ทาง 2 มาใช้สร้างหรือปรับปรุงกฎความสัมพันธ์เมื่อมีการเพิ่มของข้อมูล ที่ได้กล่าวถึงในส่วนของงานวิจัยที่เกี่ยวข้องในหัวข้อที่ผ่านมา จะเห็นว่า อัลกอริทึม CBA ยังสามารถพัฒนาให้ใช้แนวทางใหม่ (New Framework) ที่สามารถแก้ข้อเสียของแนวทางที่ 1 ความแม่นยำของโมเดลเปลี่ยนหรือลดลงเมื่อมีการเพิ่มของข้อมูล และข้อเสียของแนวทางที่ 2 เวลาประมวลผลมาก ทุกรอบของการเพิ่มของข้อมูลเนื่องจากการปรับปรุงกฎความสัมพันธ์ของโมเดล

แต่แรงจูงใจในการเสนอแนวทางใหม่ เพื่อแก้ปัญหาของแนวทางที่ 1 และแนวทางที่ 2 พบว่าอัลกอริทึม CBA จะมีการแบ่งข้อมูลทั้งหมดสำหรับใช้ในโมเดล ออกเป็น 2 ชุดคือ ข้อมูลเรียนรู้ (Training data) และข้อมูลทดสอบ (Testing data) เมื่อมีการเพิ่มของข้อมูลสำหรับใช้ในโมเดล โดยเฉพาะการเพิ่มของข้อมูลเรียนรู้ (Training data) ทำให้ชุดกฎความสัมพันธ์จำแนกประเภทข้อมูล Class-Association Rules (CARs) มีค่าสนับสนุน (Support) เปลี่ยนแปลงไป เมื่อค่าสนับสนุนของชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลเปลี่ยนแปลง ชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลที่ใช้ในโมเดลและความแม่นยำของโมเดลจะเปลี่ยนแปลง เพื่อให้โมเดลยังคงรักษาระดับความแม่นยำในการทำนาย จึงจำเป็นต้องปรับปรุงและสร้างชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลสำหรับโมเดลขึ้นมาใหม่ ซึ่งล้วนแต่ใช้เวลามากทุกคร้ัง แต่อย่างไรก็ตาม หลังจากสร้างหรือปรับปรุงชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดลใหม่แล้ว พบว่าความแม่นยำของโมเดลยังใกล้เคียงกับค่าเดิม ทำให้สามารถสรุปได้ว่า การสร้างหรือปรับปรุงชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดลใหม่ทุกครั้ง ไม่จำเป็นหรือคุ้มค่าที่ต้องทำใหม่ทุกครั้งเมื่อมีการเพิ่มของข้อมูล

แต่ปัญหาสำคัญคือการหาเกณฑ์มาทำนายว่า การเพิ่มของข้อมูลเข้ามา จำเป็นหรือไม่ที่ต้องทำการสร้างหรือปรับปรุงชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดลใหม่ทุกครั้ง และยังคงรักษาระดับความแม่นยำของโมเดลให้สูงกว่าหรือใกล้เคียงค่าเดิมไว้ได้ ผู้วิจัยจึงได้นำเสนอแนวทาง ICMF (Incremental Classifier Model Framework: ICMF) (กฤษฎากร และคณะ, 2550) เพื่อหาเกณฑ์ที่ใช้ทำนายว่า การเพิ่มของข้อมูล ควรหรือไม่ควรสร้างปรับปรุงชุดกฎของโมเดลใหม่

สำหรับแนวคิดเรื่องกฎความสัมพันธ์ของโมเดลไม่จำเป็นต้องสร้างหรือปรับปรุงกฎของโมเดลทุกครั้งของการเพิ่มของข้อมูล ดูได้จากตัวอย่างข้างล่าง

1.1 ตัวอย่างกฎของโมเดลก่อนและหลังการเพิ่มของข้อมูล

กำหนดให้ข้อมูลเรียนรู้ (Training data) มีค่าสนับสนุนขั้นต่ำ (Minimum support, minsup=2), ค่าความเชื่อมั่นขั้นต่ำ (Minimum confidence, minconf=50%) โดยแสดงผลการคำนวณดังตัวอย่างข้างล่าง

Databases before insertions

| TID | Items | Class |
|---|---|---|
| 1 | A,B,C | X |
| 2 | A,C,D | Y |
| 3 | A,B,D | X |
| 4 | B,C,D | Y |
| 5 | A,D | X |

Databases after insertions

| TID | Items | Class |
|---|---|---|
| 1 | A,B,C | X |
| 2 | A,C,D | Y |
| 3 | A,B,D | X |
| 4 | B,C,D | Y |
| 5 | A,D | X |
| 6 | A,B | X |
| 7 | C,D | Y |
| 8 | D | X |

**ภาพที่ 5** ข้อมูลในฐานข้อมูลก่อนและหลังของการเพิ่มของข้อมูล

Rules before insertions

| RID | Rule | Support | Confidence |
|---|---|---|---|
| 1 | A,B->X | 2 | 2/2 = 100% |
| 2 | C,D->Y | 2 | 2/2 = 100% |
| 3 | A->X | 3 | 3/4 = 75% |
| 4 | B->X | 2 | 2/3 = 67% |
| 5 | C->Y | 2 | 2/3 = 67% |
| 6 | D->X | 2 | 2/4 = 50% |
| 7 | D->Y | 2 | 2/4 = 50% |

Rules after insertions

| RID | Rule | Support | Confidence |
|---|---|---|---|
| 1 | A,B->X | 3 | 3/3 = 100% |
| 2 | C,D->Y | 3 | 3/3 = 100% |
| 3 | A->X | 4 | 4/5 = 80% |
| 4 | B->X | 3 | 3/4 = 75% |
| 5 | C->Y | 3 | 3/4 = 75% |
| 6 | D->X | 3 | 3/6 = 50% |
| 7 | D->Y | 3 | 3/6 = 50% |

**ภาพที่ 6** กฎความสัมพันธ์ก่อนและหลังของการเพิ่มของข้อมูล

จากภาพที่ 5 เมื่อพิจารณาหลังจากเมื่อมีการเพิ่มของข้อมูล (Database after insertions) ค่าสนับสนุนและค่าความเชื่อมั่นของกฎใหม่ (Rules after insertions) ในภาพที่ 6 มีการเปลี่ยนแปลงไปจากค่าเดิม แต่อย่างไรก็ตาม กฎที่ได้หลังจากการเพิ่มของข้อมูลยังเหมือนเดิม เพียงแต่ค่าสนับสนุนและค่าความเชื่อมั่นเปลี่ยนแปลง จากตัวอย่างจะเห็นว่า ถ้าการเพิ่มของข้อมูลมีการกระจายตัว ในทิศทางส่งเสริมชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดลเดิมหรือไอเท็มเซ็ตที่เข้ามาลักษณะเหมือนเดิม การปรับปรุงค่าสนับสนุนก็ไม่ได้ทำชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดลเปลี่ยนแปลงไปมากนัก ชุดของกฎของโมเดลจึงไม่ได้รับ

ผลกระทบจากการเพิ่มของข้อมูล ดังนั้นการปรับปรุงชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดลทุกครั้งของการเพิ่มของข้อมูลจึงไม่จำเป็นเสมอไป

จากตัวอย่างนี้ เราสามารถอุปมานได้ว่า ถ้าการกระจายตัวของข้อมูลเรียนรู้และข้อมูลที่เพิ่มเข้ามาคล้ายกันหรือไอเท็มเซ็ตที่เข้ามาลักษณะเหมือนเดิม การเปลี่ยนแปลงค่าสนับสนุนของกฎยังคงเพิ่มเหมือนกันๆ ในทุกๆกฎ ค่าความแม่นยำของโมเดลก็ยังใกล้เคียงค่าเดิม และถ้าค่าความแม่นยำของโมเดลใกล้เคียงค่าเดิม ก็ไม่จำเป็นต้องปรับปรุงโมเดลใหม่ทุกครั้งเมื่อมีการเพิ่มของข้อมูล

**2. Incremental Classifier Model Framework: ICMF**

ผู้วิจัยเสนอแนวทาง การลดจำนวนครั้งของการปรับปรุงโมเดล เมื่อมีข้อมูลเพิ่มเข้ามา โดยใช้ค่าลำดับต่างกัน (Order Difference, OD) เป็นเกณฑ์ในการพิจารณาว่า ข้อมูลที่เพิ่มเข้ามา มีผลต่อความแม่นยำของโมเดล และจำเป็นต้องปรับปรุงชุดกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดลหรือไม่ โดยถ้าข้อมูลที่เพิ่มเข้ามา มีการกระจายตัวของข้อมูล หรือไอเท็มเซ็ตที่เข้ามาคล้ายๆกับข้อมูลเรียนรู้เริ่มต้น ค่าผลต่างของลำดับกฎก่อนและหลังการเพิ่มของข้อมูลจะน้อย ความแม่นยำของโมเดลจะใกล้เคียงค่าเดิม และไม่จำเป็นต้องปรับปรับปรุงชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดล โดยแนวทางของผู้วิจัยสามารถแบ่งได้เป็นหัวข้อดังต่อไปนี้

หัวข้อที่ 2.1 กล่าวถึงการเลือกชุดกฎที่จะใช้คำนวณค่าลำดับต่างกัน ผู้วิจัยได้เสนอการใช้ชุดกฎที่มีความยาวสอง และอธิบายเหตุผลที่ใช้ชุดกฎที่มีความยาวสอง เพื่อคำนวณค่า ลำดับต่างกัน (OD) เพื่อพิจารณาว่าข้อมูลที่เพิ่มเข้ามา มีการกระจายตัวคล้ายข้อมูลเดิมหรือไม่

หัวข้อที่ 2.2 กล่าวถึง การตั้งค่าลำดับต่างกัน สูงสุด (Maximum Order Difference, MOD) และการเปรียบเทียบค่าระหว่าง ค่าลำดับต่างกันสูงสุด (MOD) กับค่า ลำดับต่างกัน (OD) ถ้าค่าลำดับต่างกัน (OD) มากกว่าค่าลำดับลำดับต่างกันสูงสุด (MOD) แสดงว่าการกระจายตัวของข้อมูลที่เพิ่มเข้ามา มีการกระจายตัว หรือไอเท็มเซ็ตที่เข้ามาเปลี่ยนมีลักษณะเปลี่ยนไปจากข้อมูลเดิมมาก กฎความสัมพันธ์ของโมเดลจะเปลี่ยนไป และ ความแม่นยำของโมเดลก็จะเปลี่ยนไปมากด้วย จำเป็นต้องปรับปรุงหรือสร้างชุดของกฎความสัมพันธ์จำแนกประเภทข้อมูลของโมเดลใหม่

2.1 การประเมินโมเดลโดยค่าลำดับต่างกัน (Order Difference, OD)

ผู้วิจัยเสนอเกณฑ์ประเมินโมเดลด้วยค่าลำดับต่างกัน ว่าโมเดลสมควรปรับปรุงหรือไม่ ในกรณีการเพิ่มของข้อมูล แต่ในการประเมินโมเดลจะต้องพิจารณากฎความสัมพันธ์ของโมเดลซึ่ง หากประเมินทุกความยาวของกฎความสัมพันธ์ของโมเดลก็จะใช้เวลามาก ไม่เหมาะกับเทคนิคการ เพิ่มของข้อมูลที่ต้องการลดเวลาการทำงานลง และกรณีที่มีกฎความสัมพันธ์ของโมเดลเกิดขึ้นมาใหม่หลังจากเพิ่มของข้อมูล การประเมินโมเดลด้วยค่าลำดับต่างกัน ไม่สามารถคำนวณได้

ฉะนั้น การเลือกกฎความสัมพันธ์ที่สามารถเป็นตัวแทนของกฎทั้งหมดของโมเดล และรองรับการคำนวณค่าลำดับต่างกันของกฎใหม่ที่เกิดจากการเพิ่มของข้อมูล จะช่วยลดเวลาคำนวณค่าลำดับต่างกันลงได้ ผู้วิจัยจึงนำเสนอ กฎความสัมพันธ์ที่มีคลาสและความยาวสอง (R2) มาใช้เป็นตัวแทนของโมเดล เนื่องจาก เป็นกฎที่สั้นและเป็นกฎตั้งต้นของกฎทุกกฎในโมเดล ทำให้อนุมานได้ว่ากฎความสัมพันธ์ที่มีคลาสและความยาวสอง มีคุณสมบัติในการประเมินโมเดลเทียบเท่ากับกฎทั้งหมด ผู้วิจัยเสนอว่า ถ้านำกฎความสัมพันธ์ที่มีคลาสและความยาวสอง มาเรียงตามค่า วัดผล (Interesting measures) จากมากไปน้อย จะได้ลำดับของกฎทุกตัว เมื่อมีการเพิ่มของข้อมูล ค่าสนับสนุนของกฎจะเปลี่ยนไป ทำให้ ลำดับของกฎเปลี่ยนไปด้วย การที่ลำดับกฎของเปลี่ยนไปมาก แสดงว่าการกระจายตัวของข้อมูลหรือรูปแบบของไอเท็มเซ็ตที่เข้ามาเปลี่ยนไปมากด้วย ความแม่นยำโมเดลการทำนายจะเปลี่ยนตามไปด้วย

การคำนวณหาลำดับต่างกัน (OD) จะพิจารณาจากความแตกต่างลำดับของกฎความสัมพันธ์ที่มีความยาวสองจากฐานข้อมูลเรียนรู้เริ่มต้น กับลำดับของกฎความสัมพันธ์ ที่มีความยาวสองจากฐานข้อมูลรวม (ฐานข้อมูลเรียนรู้เริ่มต้นและข้อมูลที่เพิ่มขึ้น) โดยจับคู่กฎที่เหมือนกันเปรียบเทียบกัน

กำหนดให้ฟังก์ชัน f(r), f'(r') หมายถึง

f(r) คือลำดับของชุดกฎที่เกิดฐานข้อมูลเริ่มต้น (Original database (D))

f'(r') คือลำดับของชุดกฎที่เกิดฐานข้อมูลเริ่มต้นรวมกับข้อมูลที่เพิ่มเข้ามา (Incremented databases (Dd))

โดย

$f: r \rightarrow Z_+$ ; $D_{f=r \in R} R_f \subset \{1,2,3,\ldots\}$ และ

$f': r' \rightarrow Z_+$ ; $D_{f'=r' \in R} R_{f'} \subset \{1,2,3,\ldots\}$

สำหรับสูตรใช้คำนวณค่าลำดับต่างกัน มี 2 สมการคือ

1. สมการที่ (5) เป็นสมการที่หาค่าลำดับต่างกันในรูปแบบของระยะทางแมนอัตตัน (Manhattan Distance) เหมาะสำหรับกฎที่เรียงลำดับด้วยค่าวัดผลดังต่อไปนี้คือ ค่าความเชื่อมั่น (Confidence), ค่าลิฟท์ (Lift) และค่าคอนวิคชั่น (Conviction) เนื่องจากเมื่อมีข้อมูลเข้ามาเล็กน้อย ลำดับของกฎจะเปลี่ยนไปมาก ผลรวมของค่าลำดับต่างกันของกฎจะมีมาก การกำหนดค่า MOD จะเป็นไปได้ยาก จึงต้องใช้ค่าลำดับต่างกันในรูปแบบแมนอัตตัน (Manhattan Distance) เพราะจะทำให้ค่าลำดับของกฎจะค่อยๆ เปลี่ยนอย่างค่อยเป็นค่อยไปสอดคล้องกับลักษณะข้อมูลที่เปลี่ยนไป และง่ายต่อการกำหนดค่า MOD

$$OD = \sum_{r,r' \in R} |(f(r) - f'(r)) * log|(\frac{1}{f(r)-f(r')})|| \quad \text{.....(5)}$$

$$where\ log\left|\frac{1}{f(r)-f(r')}\right| = 0\ if\ f(r) - f'(r) = 0$$

2. สมการที่ (6) เป็นสมการที่หาค่าลำดับต่างกันในรูปแบบของระยะทางยูคลิดเดียน (Euclidian Distance) เหมาะสำหรับกฎที่เรียงลำดับด้วยค่าวัดผลคือ ค่าสนับสนุน (Support) เนื่องจากเมื่อมีข้อมูลเข้ามามาก ลำดับของกฎจะเปลี่ยนไปน้อยมาก ผลรวมของค่าลำดับต่างกันของกฎจะน้อยการกำหนดค่า MOD จะเป็นไปได้ยาก จึงต้องใช้ค่าลำดับต่างกันในรูปแบบแมนอัตตัน (Manhattan Distance) เพราะจะทำให้ค่าลำดับของกฎจะค่อยๆ เปลี่ยนเพิ่มขึ้นสอดคล้องกับลักษณะข้อมูลที่เปลี่ยนไป และง่ายต่อการกำหนดค่า MOD สมการที่ (6) อยู่ในหน้าถัดไป

$$OD = \sum_{r,r' \in R} \left|\left(f(r) - f'(r')\right)^2 * log\left(\frac{1}{(f(r)-f(r'))^2}\right)\right| \quad \text{.....(6)}$$

$$where\ log\left|\frac{1}{(f(r)-f(r'))^2}\right| = 0\ if\ (f(r) - f(r'))^2 = 0$$

2.2 ค่าลำดับต่างกันสูงสุด (Maximum Order Difference: MOD)

ค่าลำดับต่างกันสูงสุด (Maximum Order Difference, MOD) เป็นค่าเกณฑ์ที่ใช้ตัดสินใจว่าควรจะปรับปรุงโมเดลหรือไม่ โดยแนวคิดในการกำหนดค่าลำดับต่างกันสูงสุดได้จากสมมติฐานดังตัวอย่างภาพที่ 7

| RID | R2 | Original Databases | | Incremental Databases | |
|---|---|---|---|---|---|
| | | Support | Rule Order | Support | Rule Order |
| 1 | A->X | 10 | 1 | 11 | 4 |
| 2 | B->X | 8 | 2 | 13 | 3 |
| 3 | C->Y | 7 | 3 | 14 | 2 |
| 4 | A->Y | 6 | 4 | 8 | 5 |
| 5 | B->Y | 4 | 5 | 5 | 6 |
| 6 | C->X | 3 | 6 | 15 | 1 |

**ภาพที่ 7** แสดงลำดับ (Rule Order) ของกฎเก่าก่อนการเพิ่มของข้อมูล(Original Databases)และลำดับของกฎใหม่หลังจากการเพิ่มของข้อมูล(Incremental Databases)

เมื่อเราพิจารณาภาพที่ 7 กฎความสัมพันธ์ที่มีคลาสและมีความยาวสอง R2, RID=6 ลำดับของกฎ (Original Databases, Rule Order คอลัมน์) มีค่าเท่ากับ 6 และหลังจากการเพิ่มของข้อมูล ลำดับของกฎ (Incremental Databases, Rule Order คอลัมน์) มีค่าเท่ากับ 1 และกฎอื่นๆ RID=1, 2, 3, 4 และ 6 มีลำดับก่อนและหลังการเพิ่มของข้อมูลเปลี่ยนไปมาก ฉะนั้นจะเห็นได้ กฎที่เปลี่ยนแปลงจากลำดับสุดท้ายมายังลำดับที่หนึ่งจะมีผลกระทบต่อลำดับของกฎอื่นๆของโมเดลมาก ถ้าเราสนใจลำดับของกฎที่เป็นอันดับสุดท้ายในตอนแรกแล้วกลายเป็นอันดับหนึ่งเมื่อมีการเพิ่มข้อมูล เราสามารถอนุมานได้ว่าลำดับของกฎในโมเดลมีการเปลี่ยนแปลงไปมาก โมเดลจำเป็นต้องมีการปรับปรุง เมื่อเราคำนวณลำดับต่างกันของกฎที่ความยาวสอง RID=6 จะได้ค่าลำดับต่างกัน

เท่ากับ 6-1=5 หรือเท่ากับจำนวนกฎที่ความยาวสองทั้งหมด – 1 และเมื่อคำนวณค่าลำดับลำดับต่างกันของกฎที่เหลืออื่นๆ จะได้ผลรวมของค่าลำดับต่างกันของกฎทั้งหมดที่มากกว่าจำนวนกฎที่มีความยาวสองทั้งหมด ถ้าเรากำหนดค่าลำดับต่างกันสูงสุด MOD เท่ากับจำนวนกฎที่มีความยาวสองทั้งหมด เราสามารถอนุมานได้ว่าถ้าผลรวมของค่าลำดับต่างกันของกฎทั้งหมดมากกว่าหรือเท่ากับค่าลำดับต่างกันสูงสุด ลำดับของกฎในโมเดลมีการเปลี่ยนแปลงไปมาก โมเดลจำเป็นต้องปรับปรุง

ฉะนั้นการตั้งค่าลำดับต่างกันสูงสุด MOD เท่ากับจำนวนกฎความสัมพันธ์ที่มีคลาสและมีความยาวสอง R2 สามารถคำนวณได้จากสมการที่ (7)

$$MOD = C \times L \times Y \quad \text{..... (7)}$$

โดย

C = ค่าสัมประสิทธิ์ที่ผู้ใช้กำหนด โดยปกติ กำหนดให้เท่ากับ 1

L = จำนวนไอเท็มเซตที่ไม่ใช่คลาส

Y = จำนวนคลาส

ในบางกรณีเช่น จำนวนทรานเซ็คชั่น (Transactions) ของฐานข้อมูลที่เพิ่มเข้ามามีจำนวนน้อยมากเมื่อเทียบกับจำนวนทรานเซ็คชั่นของฐานข้อมูลเดิม กฎความสัมพันธ์ของโมเด ลก็ได้รับผลกระทบไม่มาก ไม่จำเป็นต้องปรับปรุงโมเดลบ่อย แต่ถ้าจำนวนกฎความสัมพันธ์ที่มีคลาสและมีความยาวสอง R2 มีจำนวนมาก แนวโน้มของผลรวมค่าลำดับต่างกันของกฎจะมีค่ามากกว่า MOD มากขึ้นทำให้ต้องปรับปรุงโมเดลบ่อยขึ้น จึงควรกำหนดค่าสัมประสิทธิ์ที่ใช้กำหนดค่า MOD ให้มีค่ามากขึ้น เช่น 1.5, 2 หรือ จำนวนเท่าของกฎความสัมพันธ์ที่มีคลาสและมีความยาวสอง R2 ในทางกลับกัน ถ้าจำนวนทรานเซ็คชั่น (Transactions) ของฐานข้อมูลที่เพิ่มเข้ามามีจำนวนมากเมื่อเทียบกับจำนวนทรานเซ็คชั่นของฐานข้อมูลเดิม กฎความสัมพันธ์ของโมเดลจะได้รับผลกระทบมาก จำเป็นต้องปรับปรุงโมเดลบ่อยๆ แต่จำนวนกฎความสัมพันธ์ที่มีคลาสและมีความยาวสอง R2 มีจำนวนน้อยมาก แนวโน้มของผลรวมค่าลำดับต่างกันของกฎจะมีมากกว่า MOD น้อย การจะทำให้โมเดลต้องปรับปรุงโมเดลบ่อยขึ้น จึงควรกำหนดค่าสัมประสิทธิ์ที่ใช้กำหนดค่า MOD ให้มีค่าน้อยขึ้น เช่น 0.5, 0.75 หรือ จำนวนเท่าของกฎความสัมพันธ์ที่มีคลาสและมีความยาวสอง R2

การตัดสินใจว่าควรจะปรับปรุงโมเดลหรือ พิจารณาจากค่าลำดับต่างกัน (OD) เปรียบเทียบกับ ค่าลำดับต่างกันสูงสุด (MOD) มีกรณีที่เป็นไปได้ดังนี้

1. กรณี ที่ 1: OD > MOD

   ค่าลำดับต่างกัน มากกว่าค่าลำดับต่างกันสูงสุด แสดงว่าควรจะปรับปรุงโมเดลใหม่

2. กรณี ที่ 2: OD < = MOD

   ค่าลำดับต่างกัน น้อยกว่าหรือเท่ากับค่าลำดับต่างกันสูงสุด ไม่จำเป็นต้องปรับปรุงโมเดลให้ใช้โมเดลเดิมต่อไป

### 3. อัลกอริทึม ICMF (Incremental Classifier Model Framework)

อัลกอริทึม ICMF มี 2 เฟส คือ เฟสเริ่มต้น (Initial Phase) และ เฟสการเพิ่มของข้อมูล (Incremental Phase) เฟสเริ่มต้นจะหาลำดับของชุดกฎความสัมพันธ์ที่มีคลาสและความยาว 2 (R2) จากฐานข้อมูลเริ่มต้น (Original Database) และเฟสการเพิ่มของข้อมูลจะคำนวณหาค่าลำดับต่างกันระหว่างลำดับของกฎในชุดกฎที่หาจากเฟสเริ่มต้นและลำดับของกฎในชุดกฎที่หาจากข้อมูลที่รวมระหว่างข้อมูลเริ่มต้นกับข้อมูลที่เพิ่มเข้า สัญลักษณ์ที่ใช้ในอัลกอริทึม ICMF แสดงในตารางที่ 16 และรหัสเทียมของอัลกอริทึม ICMF แสดงในรูปที่ 8

**ตารางที่ 16** ตารางสัญลักษณ์ที่ใช้ในอัลกอริทึม ICMF

| Parameter | Description |
|---|---|
| $D$ | Original databases |
| $d$ | Incremental databases |
| $i$ | $i^{th}$ round; i=1,2,3,…,n |
| $d_i$ | Incremental databases in $i^{th}$ round |
| $Dd$ | Incremented databases use data from both original database and cumulative incremental databases to compute Rules $R$ (without merging both databases). |
| $R$ | Size-2 rules |
| $r$ , $R^D$ | CAR Rules of size 2 from the original database |
| $r'$ ,$R^{Dd}$ | CAR Rules of size 2 from the incremented databases |
| $\alpha$ | Phase number, =0 for initial phase; phase number, =1 for incremental phase |

```
Algorithm ICMF
Input: Original databases D, Incremental databases d_i, MOD
Output: CAR Rules of original databases R^D, CAR Rules of incremented
databases R^Dd
Procedure ICMF
1)     α ← 0
2)     i  ← 0
3)     While (D ≠ ∅)
4)         If α = 0  Then //Initial phase
5)             BuildClassifier (D);
6)             f(r) = ComputeOrder (R^D);
7)             α ← 1;
8)         Else   //Incremental phase
9)         d ← d_i++;
10)        Dd ← D ∪ d;
11)            f'(r')= ComputeOrder (R^Dd);
12)            //compute OD
13)            For each  r ∈ R^D = r' ∈ R^Dd
14)            // If R order by confidence, lift and conviction then
15)            // calculate f(r) Δ f'(r') from equation 1
16)            // If R order by support then
17)            // calculate f(r) Δ f'(r') from equation 2
18)              OD'_r =f(r) Δ f'(r');
19)            End
20)            OD = Σ_{r∈R^D} OD'_r;
21)            If OD > MOD  Then
22)              D ← D ∪ d;
23)              α ← 0;
24)              i← 0;
25)            Else
26)              // current classifier model
27)              // predict incremental databases
28)              ModelPredict (d);
29)              α ← 1;
30)            End
31)       End
32)    End
```

**ภาพที่ 8** รหัสเทียมของอัลกอริทึม ICMF

อัลกอริทึม ICMF เริ่มต้นที่เฟสเริ่มต้น (initial phase) (บรรทัดที่ 4-7) อัลกอริทึมนี้สร้างโมเดลการทำนายจากฐานข้อมูลเริ่มต้น (original databases D) หลังจากนั้นก็คำนวณลำดับของกฎความสัมพันธ์ที่มีคลาสและมีความยาวสอง ($R^D$) จากฐานข้อมูลเริ่มต้น แล้วก็เริ่มทำต่อไปในเฟสการเพิ่มของข้อมูล (incremental phase) เฟสการเพิ่มของข้อมูล (บรรทัดที่ 9-10) กำหนดค่าเริ่มต้นของฐานข้อมูลที่เพิ่มเข้ามา (incremental databases, d) และค่าเริ่มต้นของฐานข้อมูลรวมระหว่างข้อมูลเริ่มต้นและข้อมูลที่เพิ่มเข้ามา (incremented databases, Dd) (บรรทัดที่ 11) คำนวณลำดับของกฎความสัมพันธ์ที่มีคลาสและมีความยาวสอง ($R^{Dd}$) จากฐานข้อมูลรวมระหว่างข้อมูลเริ่มต้นและ

ข้อมูลที่เพิ่มเข้ามา (บรรทัดที่ 13-20) คำนวณลำดับแต่งต่างกันระหว่างกฎความสัมพันธ์ที่มีคลาสและมีความยาวสองที่ได้จากฐานข้อมูลเริ่มต้น (original databases, D) และกฎความสัมพันธ์ที่มีคลาสและมีความยาวสองที่ได้จากฐานข้อมูลรวมระหว่างข้อมูลเริ่มต้นและข้อมูลที่เพิ่มเข้ามา (incremented databases, Dd) แล้วคำนวณผลรวมของค่าลำดับต่างกันของทุกกฎ(OD) (บรรทัดที่ 21-31) ตรวจสอบว่าค่าลำดับต่างกัน (OD) มากกว่าค่าลำดับต่างกันสูงสุด (MOD) หรือไม่ ถ้ามากกว่าก็ให้อัลกอริทึม ICMF รวมฐานข้อมูลเริ่มต้นและข้อมูลที่เพิ่มเข้ามาเป็นฐานข้อมูลเริ่มต้นใหม่ และกลับไปเริ่มต้นที่เฟสเริ่มต้น ถ้าไม่มากกว่าก็ให้ใช้โมเดลเดิมทำนายข้อมูลที่เพิ่มเข้ามา และกลับไปเริ่มที่เฟสการเพิ่มของข้อมูลอีกครั้ง

## 4. ตัวอย่างการทำงานของอัลกอริทึม ICMF

เพื่อให้ง่ายต่อการเข้าใจ เรามาดูตัวอย่างการทำงานของอัลกอริทึม ICMF เฉพาะการคำนวณค่าลำดับต่างกัน (OD) และเปรียบเทียบค่าลำดับต่างกันขั้นต่ำ (MOD) ดังแสดงในภาพที่ 9, 10 และ 11 ตัวอย่างการทำงานของอัลกอริทึม ICMF ใช้กฎความสัมพันธ์ที่มีคลาสและมีความยาวสอง R2 และลำดับกฎ R2 ด้วยค่าสนับสนุน (Support) และใช้สมการที่ (6) ในหน้า 37 คำนวณ กำหนดให้ค่า MOD = 6

| RID | R2 | Initial Databases | |
|---|---|---|---|
| | | Support | Rule Order |
| 1 | A->X | 10 | 1 |
| 2 | B->X | 8 | 2 |
| 3 | C->Y | 7 | 3 |
| 4 | A->Y | 6 | 4 |
| 5 | B->Y | 4 | 5 |
| 6 | C->X | 3 | 6 |

**ภาพที่ 9** แสดงตัวอย่างการทำงานเบื้องต้นของอัลกอริทึม ICMF ในรอบเริ่มต้น

รอบเริ่มต้น (Original databases round) อัลกอริทึม ICMF คำนวณค่าสนับสนุนของแต่ละกฎและเรียงลำดับกฎจากกฎที่มีค่าสนับสนุนมากไปน้อย ตัวอย่างเช่นกฎ A->X, A->X มีค่าสนับสนุนสูงสุดและมีลำดับเป็น 1

| RID | R2 | Initial Databases | | Incremental Databases#1 | | Order Difference (OD) |
|---|---|---|---|---|---|---|
| | | Support | Rule Order | Support | Rule Order | |
| 1 | A->X | 10 | 1 | 10 | 3 | $=\|(1-3)^2 x \log(1/(1-3)^2)\|=2.41$ |
| 2 | B->X | 8 | 2 | 11 | 2 | $=\|(2-2)^2 x \log(1/(2-2)^2)\|=0$ |
| 3 | C->Y | 7 | 3 | 13 | 1 | $=\|(3-1)^2 x \log(1/(3-1)^2)\|=2.41$ |
| 4 | A->Y | 6 | 4 | 9 | 4 | $=\|(4-4)^2 x \log(1/(4-4)^2)\|=0$ |
| 5 | B->Y | 4 | 5 | 5 | 5 | $=\|(5-5)^2 x \log(1/(5-5)^2)\|=0$ |
| 6 | C->X | 3 | 6 | 4 | 6 | $=\|(6-6)^2 x \log(1/(6-6)^2)\|=0$ |
| Total OD | | | | | | 4.82 |

**ภาพที่ 10** แสดงตัวอย่างการทำงานเบื้องต้นของอัลกอริทึม ICMF ในรอบรอบที่ 1 ของการเพิ่มของข้อมูล (Incremental database round #1)

รอบที่ 1 ของการเพิ่มของข้อมูล (Incremental database round #1) อัลกอริทึม ICMF คำนวณค่าสนับสนุนใหม่ของแต่ละกฎโดยรวมค่าสนับสนุนระหว่างฐานข้อมูลเริ่มต้นและข้อมูลที่เพิ่มเข้าของรอบที่ 1 หลังจากนั้น เรียงลำดับกฎจากกฎที่มีค่าสนับสนุนมากไปน้อย ตัวอย่างกฎ A->X, ค่าสนับสนุนของกฎ A->X ยังเท่ากับ 10 แต่ลำดับของกฎเปลี่ยนจาก 1 เป็น 3 ดังนั้นค่าลำดับต่างกัน (OD) ของกฎ A->X คือ $|(1-3)^2 * \log |1/(1-3)^2||=2.41$ หลังจากนั้นหารวมผลรวมของค่าลำดับต่างกันของแต่ละกฎได้เท่ากับ 4.82 ซึ่งน้อยกว่าค่าลำดับต่างกันสูงสุด (MOD) =6 ดังนั้นอัลกอริทึม ICMF ยังใช้โมเดลเดิมทำนายข้อมูลที่เพิ่มเข้ามา

| RID | R2 | Initial Databases | | Incremental Databases#2 | | Order Difference (OD) |
|---|---|---|---|---|---|---|
| | | Support | Rule Order | Support | Rule Order | |
| 1 | A->X | 10 | 1 | 10 | 4 | $=\|(1-4)^2 x \log(1/(1-4)^2)\|=8.59$ |
| 2 | B->X | 8 | 2 | 13 | 3 | $=\|(2-3)^2 x \log(1/(2-3)^2)\|=0$ |
| 3 | C->Y | 7 | 3 | 14 | 2 | $=\|(3-2)^2 x \log(1/(3-2)^2)\|=0$ |
| 4 | A->Y | 6 | 4 | 15 | 1 | $=\|(4-1)^2 x \log(1/(1-4)^2)\|=8.59$ |
| 5 | B->Y | 4 | 5 | 5 | 5 | $=\|(5-5)^2 x \log(1/(5-5)^2)\|=0$ |
| 6 | C->X | 3 | 6 | 4 | 6 | $=\|(6-6)^2 x \log(1/(6-6)^2)\|=0$ |
| Total OD | | | | | | 17.18 |

**ภาพที่ 11** แสดงตัวอย่างการทำงานเบื้องต้นของอัลกอริทึม ICMF ในรอบรอบที่ 2 ของการเพิ่มของข้อมูล (Incremental database round #2)

รอบที่ 2 ของการเพิ่มของข้อมูล Incremental database round #2 อัลกอริทึม ICMF คำนวณค่าสนับสนุนใหม่ของแต่ละกฎโดยรวมค่าสนับสนุนระหว่างฐานข้อมูลเริ่มต้นและข้อมูลที่เพิ่มเข้าของรอบที่ 1 และรอบที่ 2 หลังจากนั้น เรียงลำดับกฎจากกฎที่มีค่าสนับสนุนมากไปน้อย ตัวอย่างกฎ A->X ค่าสนับสนุนของกฎ A->X ยังเท่ากับ 10 แต่ค่าลำดับของกฎเปลี่ยนจาก 1 เป็น 4. ดังนั้นค่าลำดับต่างกัน (OD) ของกฎ A->X คือ $|(1-4)^2 * \log|1/(1-4)^2||=8.59$ หลังจากนั้นรวมผลรวมของค่าลำดับต่างกันของแต่ละกฎได้เท่ากับ 17.18 ซึ่งมากกว่าค่าลำดับต่างกันสูงสุด (MOD) =6 ดังนั้นอัลกอริทึม ICMF จะสร้างโมเดลขึ้นมาใหม่เพื่อทำนายข้อมูลที่เพิ่มเข้ามา

## 5. เปรียบเทียบคุณลักษณะของแต่ละอัลกอริทึม

จากที่ได้กล่าวมานั้น เราสังเกตเห็นว่า อัลกอริทึม ICMF ได้นำเสนอค่าลำดับต่างกันของกฎ (Order different) เพื่อพิจารณาว่าควรปรับปรุงโมเดลหรือใช้โมเดลเดิมเมื่อมีข้อมูลเพิ่มเข้ามา โดยแนวทางของอัลกอริทึม ICMF ถูกออกแบบมาเพื่อลดเวลารวมในการปรับปรุงโมเดลเมื่อมีข้อมูลเพิ่มเข้ามาและแก้ปัญหาเวลารวมที่มากของแนวทางที่ 2 (Framework 2) โดยยังรักษาระดับความแม่นยำของโมเดลให้ใกล้เคียงหรือดีกว่าแนวทางที่ 2 และเวลารวมยังคงใกล้เคียงกับแนวทางที่ 1 (Framework 1)

ความแตกต่างระหว่างอัลกอริทึม ICMF กับแนวทาง (Framework) อื่นๆ ที่เกี่ยวข้อง จะเป็นไปดังตารางที่ 17 ซึ่งเป็นตารางเปรียบเทียบคุณลักษณะต่างๆ ของแต่ละอัลกอริทึม

**ตารางที่ 17** ตารางเปรียบเทียบคุณลักษณะของแต่ละอัลกอริทึม

| Algorithms | Created New Classifier Model | | Total Computational Time | Average Accuracy |
|---|---|---|---|---|
| | Original Round | Incremental Round | | |
| CBA with Framework 1 | ✓ | ✗ | Low-Medium | Medium-High |
| CBA with Framework 2 | ✓ | ✓ | Very High | High |
| CBA with ICMF | ✓ | ✓✗ Depends on ICMF's criteria to create or not | Low-Medium | High |

# ผลและวิจารณ์

## ผล

### 1. วิธีวัดผลการทดลอง

ผู้วิจัยได้ทำการวิเคราะห์และเปรียบเทียบเวลาและความแม่นยำของอัลกอริทึม CBA ที่เปลี่ยนแปลงหลังจากการมีการเพิ่มของข้อมูลโดยเปรียบเทียบระหว่างแนวทางที่ 1 (Framework 1) ไม่มีการปรับปรุงโมเดล เมื่อมีการเพิ่มของข้อมูล ใช้โมเดลเดิม ไปทำนายข้อมูลที่เพิ่มขึ้นไปเรื่อยๆ แนวทางที่ 2 (Framework 2) โมเดลการทำนายข้อมูลมีการปรับปรุงให้โมเดลทันสมัย ทุกครั้งเมื่อมีการเพิ่มของข้อมูล และแนวทางของอัลกอริทึม ICMF ซึ่งพิจารณาว่าโมเดลสมควรปรับปรุงเมื่อมีการเพิ่มของข้อมูลหรือไม่ โดยอัลกอริทึม CBA สำหรับแนวทางที่ 1, แนวทางที่ 2 และอัลกอริทึม ICMF ถูกเขียนด้วยภาษาจาวา (Java) พัฒนาต่อยอดมาจาก LUCS-KDD implementation of CBA software (Coenen, 2004) และใช้ฐานข้อมูลสาธารณะจาก UCI dataset และแปลงข้อมูล (pre-processed) ให้อยู่ในรูปที่เหมาะสมสำหรับอัลกอริทึม CBA โดยใช้ LUCS-KDD DN software (Coenen, 2003)

**ตารางที่ 18** รายละเอียดของฐานข้อมูลแต่ละชุดจาก UCI Repository of Machine Learning Database

| Datasets | #Attribute | #Records | #Classes | #Records per class | | |
|---|---|---|---|---|---|---|
| Adult | 97 | 48842 | 2 | Class | Num. Rec. | % |
| | | | | 96 | 11687 | 23.93 |
| | | | | 97 | 37155 | 76.07 |
| Connect4 | 128 | 67557 | 3 | Class | Num. Rec. | % |
| | | | | 129 | 6449 | 9.55 |
| | | | | 128 | 16635 | 24.62 |
| | | | | 127 | 44473 | 65.83 |

**ตารางที่ 18** (ต่อ)

| Datasets | #Attribute | #Records | #Classes | #Records per class | | |
|---|---|---|---|---|---|---|
| Led7 | 24 | 3200 | 10 | Class | Num. Rec. | % |
| | | | | 21 | 301 | 9.41 |
| | | | | 18 | 307 | 9.59 |
| | | | | 19 | 312 | 9.75 |
| | | | | 17 | 313 | 9.78 |
| | | | | 20 | 313 | 9.78 |
| | | | | 22 | 314 | 9.81 |
| | | | | 23 | 327 | 10.22 |
| | | | | 15 | 329 | 10.28 |
| | | | | 24 | 334 | 10.44 |
| | | | | 16 | 350 | 10.94 |
| Nursery | 32 | 12960 | 5 | Class | Num. Rec. | % |
| | | | | 29 | 2 | 0.02 |
| | | | | 30 | 328 | 2.53 |
| | | | | 32 | 4044 | 31.20 |
| | | | | 31 | 4266 | 32.92 |
| | | | | 28 | 4320 | 33.33 |

**ตารางที่ 18** (ต่อ)

| Datasets | #Attribute | #Records | #Classes | #Records per class | | |
|---|---|---|---|---|---|---|
| PageBlocks | 46 | 5473 | 5 | Class | Num. Rec. | % |
| | | | | 44 | 28 | 0.51 |
| | | | | 45 | 88 | 1.61 |
| | | | | 46 | 115 | 2.10 |
| | | | | 43 | 329 | 6.01 |
| | | | | 42 | 4913 | 89.77 |
| PenDigits | 89 | 10992 | 10 | Class | Num. Rec. | % |
| | | | | 83 | 1055 | 9.60 |
| | | | | 85 | 1055 | 9.60 |
| | | | | 88 | 1055 | 9.60 |
| | | | | 89 | 1055 | 9.60 |
| | | | | 86 | 1056 | 9.61 |
| | | | | 87 | 1142 | 10.39 |
| | | | | 80 | 1143 | 10.40 |
| | | | | 81 | 1143 | 10.40 |
| | | | | 82 | 1144 | 10.41 |
| | | | | 84 | 1144 | 10.41 |

ตารางที่ 18 แสดงให้เห็นถึงรายละเอียดของทั้ง 6 ฐานข้อมูลที่นำมาใช้ โดยบอกรายละเอียดดังต่อไปนี้

#Attribute คือ จำนวนของข้อมูลในแต่ละแถว

#Records คือ จำนวนแถวของฐานข้อมูล

#Classes คือ จำนวนคลาสปลายทาง ที่ต้องการจะทำนาย

#Records per class, Num. Rec. คือ จำนวนแถวของฐานข้อมูลของแต่ละคลาส

เมื่อพิจารณาลักษณะของแต่ละดาต้าเซ็ต (Datasets) เราพบว่าสามารถแบ่งดาต้าเซ็ต ได้เป็นสองกลุ่ม คือ

1. Imbalanced Datasets คือ ดาต้าเซ็ตที่มีจำนวนของคลาสหลักมากกว่าคลาสอื่นเป็นจำนวนมาก ได้แก่ ดาต้าเซ็ต Adult Connect4 และ PageBlocks
2. Balanced Datasets คือ ดาต้าเซ็ตที่มีจำนวนของแต่ละคลาสใกล้เคียงกัน ได้แก่ ดาต้าเซ็ต Led7 Nursery และ PenDigits

รูปแบบการทดลองจะแบ่งข้อมูลแต่ละดาต้าเซต(Datasets) ออกเป็นสองชุด ชุดแรกเป็นข้อมูลสำหรับที่ใช้ในการสร้างโมเดลครั้งแรก ชุดที่สองสำข้อมูลที่เพิ่มเข้ามา ยกตัวอย่างเช่น ข้อมูล nursery ดาต้าเซต มี 12960 เร็คคอร์ด จะถูกแบ่งเป็นข้อมูลชุดแรกจำนวน 10960 เร็คคอร์ด สำหรับใช้ในการสร้างโมเดลครั้งแรก และชุดที่สอง สำหรับข้อมูลที่เพิ่มเข้าแต่ละผู้วิจัยกำหนดให้เข้ามาครั้งละ 200 เร็คคอร์ด จำนวน 10 ครั้ง แต่อย่างไรการกำหนดจำนวนเร็คคอร์ดของข้อมูลที่เพิ่มเข้า เราควรจะกำหนดเป็นจำนวนข้อมูลที่เพิ่มเข้ามาเป็น % ของจำนวนเร็คคอร์ดทั้งหมดของดาต้าเซ็ต อย่างเช่น เราแบ่ง ดาต้าเซ็ต เป็นสองส่วน ส่วนแรกกำหนดให้ข้อมูลเริ่มต้นเป็น 90% ส่วนที่สองกำหนดให้ข้อมูลที่เพิ่มเข้ามาเป็น 10% โดยกำหนดให้แต่ละครั้งของการเพิ่มข้อมูลเข้าครั้งละ 1 % จำนวน 10 ครั้ง หรือเราควรจะไปสำรวจว่าลักษณะงานจริงๆ มีลักษณะจำนวนการเพิ่มของข้อมูลเป็นเท่าไร เช่น ข้อมูลของการซื้อของในซุปเปอร์มาเก็ต มีจำนวนเร็คคอร์ดข้ อมูลการซื้อของลูกค้าเก็บไว้เป็นจำนวนเท่าไร แล้วจำนวนเร็คคอร์ดข้อมูลการซื้อของลูกค้าที่ซื้อในแต่ละวัน แต่ละอาทิตย์ แต่ละเดือน เพื่อจะได้กำหนดข้อมูลเริ่มต้นและข้อมูลที่เพิ่มเข้ามาได้เหมาะสมและสอดคล้องกับลักษณะงานจริงๆ

ข้อมูลชุดแรกจะถูกนำมาใช้ในการสร้างโมเดลครั้งแรกโดยแบ่งเป็น Training data 70% และ Testing data 30% หลังจากนั้นเมื่อมีการเพิ่มของข้อมูล ต้องสร้างโมเดลใหม่ ข้อมูลชุดที่สองจะถูกนำไปใช้ โดยจะแบ่งข้อมูลออกเป็น 70/30, โดยที่ 70% ของข้อมูลชุดที่สองจะถูกนำไปรวมกับ Training data ของข้อมูลชุดแรก และ 30% ของข้อมูลชุดที่สองจะถูกนำไปรวมกับ Testing data ของข้อมูลชุดแรก

**2. ผลการทดลอง**

ตารางที่ 19 และ 20 แสดงผลการทดทองของอัลกอริทึม ICMF เปรียบเทียบกับแนวทางที่ 1 (Framwork1) และ แนวทางที่ 2 (Framework2) ในส่วนของ ค่าแม่นยำเฉลี่ยของโมเดล (Average accuracy) และเวลาประมวลผลรวมเฉลี่ยของการสร้างและปรับปรุงโมเดล (Average computational time) โดยผลการทดลองของอัลกอริทึมของ ICMF ยังสามารถแบ่งได้ย่อยอีกเป็น 4 แบบคือ

1. อัลกอริทึม ICMF ใช้ค่าสนับสนุน (Support) ในการเรียงกฎ (ICMF Support)
2. อัลกอริทึม ICMF ใช้ค่าความเชื่อมั่น (Confidence) ในการเรียงกฎ (ICMF Confidence)
3. อัลกอริทึม ICMF ใช้ค่าลิฟท์ (Lift) ในการเรียงกฎ (ICMF Lift)
4. อัลกอริทึม ICMF ใช้ค่าคอนวิคชั่น (Conviction) ในการเรียงกฎ (ICMF Conviction)

**ตารางที่ 19** เปรียบเทียบความแม่นยำเฉลี่ยของโมเดล (Average accuracy) แต่ละอัลกอริทึม

| Datasets | FW1 Avg.Accuary | FW2 Avg.Accuary | ICMF (Support) Avg.Accuary | ICMF (Confidence) Avg.Accuary | ICMF (Lift) Avg.Accuary | ICMF (Conviction) Avg.Accuary |
|---|---|---|---|---|---|---|
| Adult | 76.10 | 76.10 | 76.10 | 76.10 | 76.10 | 76.10 |
| Connect4 | 65.89 | 65.89 | 65.89 | 65.89 | 65.89 | 65.89 |
| PageBlocks | 89.78 | 89.78 | 89.78 | 89.78 | 89.78 | 89.78 |
| Led7 | **73.24** | **73.29** | **73.29** | **73.29** | **73.29** | **73.29** |
| Nursery | **79.32** | **79.95** | **80.46** | **80.14** | **80.14** | **80.14** |
| PenDigits | **76.65** | **78.15** | **78.15** | **78.15** | **78.15** | **78.15** |

**ตารางที่ 20** เวลาประมวลผลรวมเฉลี่ยของการสร้างและปรับปรุงโมเดล (Average computational time) แต่ละอัลกอริทึม

| Datasets | FW1 Avg.Comp. Time(ms) | FW2 Avg.Comp. Time(ms) | ICMF (Support) Avg.Comp. Time(ms) /#rebuilt classifier | ICMF (Confidence) Avg.Comp. Time(ms) /#rebuilt classifier | ICMF (Lift) Avg.Comp. Time(ms) /#rebuilt classifier) | ICMF (Conviction) Avg.Comp. Time(ms) /#rebuilt classifier |
|---|---|---|---|---|---|---|
| Adult | 19119 | 229434 | 20942/0 | 20833/0 | 40137/1 | 19542/0 |
| Connect4 | 40082 | 458752 | 40672/0 | 40482/0 | 168503/3 | 83432/1 |
| PageBlocks | 163 | 913 | 698/7 | 402/3 | 432/3 | 347/2 |
| Led7 | 123 | 317 | 339/10 | 283/8 | 348/10 | 300/8 |
| Nursery | 265 | 808 | 437/4 | 295/4 | 250/4 | 249/4 |
| PenDigits | 238 | 1205 | 1244/10 | 1281/10 | 1349/10 | 1351/10 |

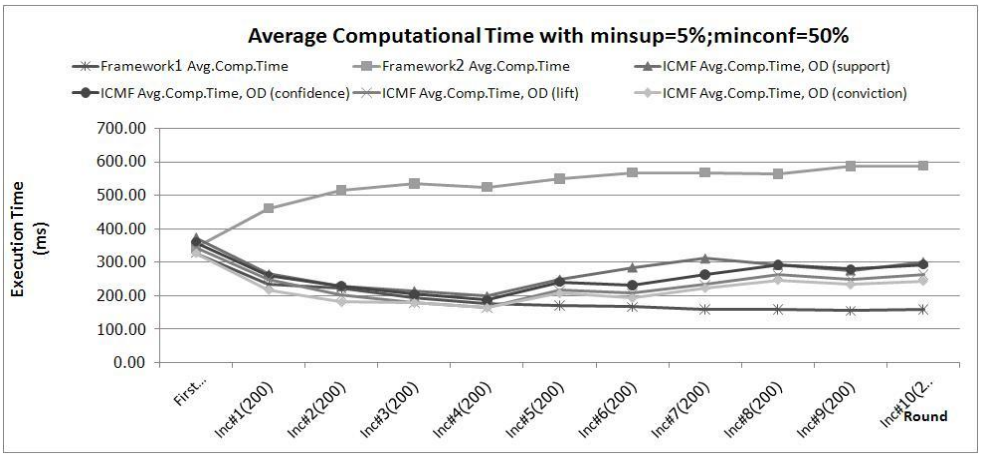
ตารางที่ 19 เราจะพบว่า Imbalance Datasets ได้แก่ดาต้าเซ็ต Adult, Connect4 และ PageBlocks มีความแม่นยำเฉลี่ยของโมเดลเท่ากันในแนวทางที่ 1 (ไม่มีการปรับปรุงโมเดล ใช้โมเดลเดิมทำนายข้อมูลที่เพิ่มขึ้นไปเรื่อยๆ) แนวทางที่ 2 (โมเดลการทำนายข้อมูลมีการปรับปรุงให้โมเดลทันสมัย ทุกครั้งเมื่อมีการเพิ่มของข้อมูล ) และแนวทางของอัลกอริทึม ICMF (ใช้เกณฑ์พิจารณาว่าควรจะปรับปรุงโมเดลหรือไม่) แสดงว่าการเพิ่มของข้อมูลมีลักษณะคล้ายคลึงกันหรือรูปแบบไอเท็มเซ็ตที่เข้ามาเหมือนหรือคล้ายเดิม ทำให้ค่าความแม่นยำของโมเดลเท่ากันทั้งหมด เมื่อพิจารณาลักษณะของแต่ละดาต้าเซ็ต พบว่าจำนวนแถวของฐานข้อมูลของแต่ละคลาสในแต่ละดาต้าเซ็ต คลาสอันดับที่หนึ่งที่มีจำนวนแถวของฐานข้อมูลมากที่สุด และคลา สอันดับที่สองที่มีจำแนวของฐานข้อมูลรองลงมามีจำนวนแถวแตกต่างกันมาก (อ้างถึงตารางที่ 18 รายละเอียดของฐานข้อมูลแต่ละชุด คอลัมน์ #Records per class) แม้ว่าจะมีการเพิ่มของข้อมูล ลักษณะกฎของโมเดลก็มักจะทำนาย ดีฟอลท์คลาส (Default class) เป็นคลาสอันดับที่หนึ่งก่อนเสมอ ทำให้ความแม่นยำของโมเดลยังใกล้เคียงหรือเท่ากับโมเดลเดิม ฉะนั้นในกรณีที่ความแม่นยำของโมเดลไม่เปลี่ยนหรือใกล้เคียงค่าเดิมเมื่อมีการเพิ่มของข้อมูล แนวทางที่ 1 จึงเป็นแนวทางที่เหมาะสมกว่า แต่แนวทางที่ 1 ไม่สามารถบอกได้ว่าการเพิ่มของข้อมูลมีลักษณะคล้ายคลึงกันหรือรูปแบบไอเท็มเซ็ตที่เข้ามาเหมือนหรือคล้ายเดิม จึงต้องเอาอัลกอริทึม ICMF มาพิจารณา ซึ่งจากตารางที่ 20 อัลกอริทึม ICMF ก็ให้เวลารวมเฉลี่ยใกล้เคียงแนวทางที่ 1 หรือในกรณีเวลารวมเฉลี่ยมากกว่าแนวทางที่ 1 เวลารวมเฉลี่ยก็ยังน้อยกว่าแนวทางที่ 2 มาก อย่างเช่น อัลกอริทึม ICMF ที่ใช้ค่าความเชื่อมั่นในการเรียง

กฎ (ICMF confidence) ให้เวลารวมเฉลี่ยในดาต้าเซ็ต Adult และ Connect4 เทียบเท่าแนวทางที่ 1 และดาต้าเซ็ต PageBlock มากกว่าแนวทางที่ 1 ประมาณ 3 เท่า แต่น้อยกว่าแนวทางที่ 2 ประมาณ 2 เท่า

ตารางที่ 19 เราพิจารณา Imbalance Datasets ได้แก่ดาต้าเซ็ต Led7 และ PenDigits มีความแม่นยำเฉลี่ยของแนวทางที่ 2 และอัลกอริทึม ICMF มากกว่าแนวทางที่ 1 ถ้าพิจารณาความแม่นยำเฉลี่ยของแนวทางที่ 2 และแนวทางที่ 1 พบว่าแนวทางที่ 2 ให้ความแม่นยำเฉลี่ยสูงกว่าแนวทางที่ 1 แสดงว่าการการกระจายตัวของข้อมูลที่เพิ่มเข้ามาหรือ รูปแบบไอเท็มเซ็ตที่เพิ่มเข้ามา เปลี่ยนไป แนวทางที่ 2 จึงเป็นแนวทางที่เหมาะสมกว่า เมื่อพิจารณาตารางที่ 20 จะพบว่า อัลกอริทึม ICMF ทำงานเหมือนแนวทางที่ 2 เพราะว่ามีการสร้างโมเดลใหม่ทุกครั้งเมื่อมีการเพิ่มของข้อมูล (10 ครั้ง)

ตารางที่ 19 เราพิจารณา Imbalance Datasets ได้แก่ดาต้าเซ็ต Nursery ให้ความแม่นยำเฉลี่ยของอัลกอริทึม ICMF สูงกว่าทั้งแนวทางที่ 1 และแนวทาง 2 จึงสมควรวิเคราะห์ผลการทดลองของความแม่นยำเฉลี่ยและเวลารวมเฉลี่ยของแต่ละรอบการเพิ่มของข้อมูล

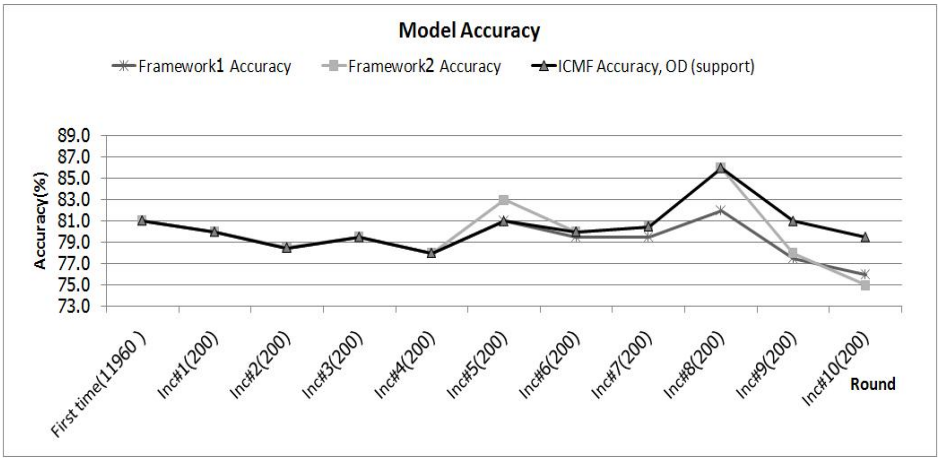
2.1 เวลาประมวลรวมเฉลี่ยของการสร้างและปรับปรุงโมเดลในแต่ละรอบของการเพิ่มของข้อมูล



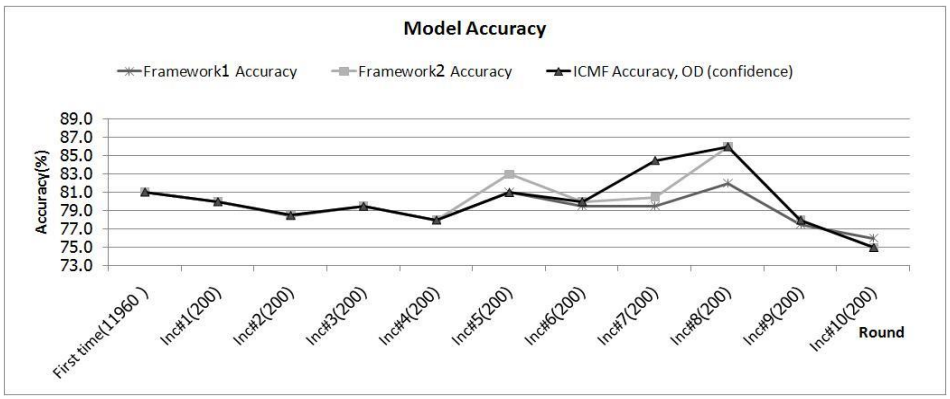

**ภาพที่ 12** แสดงเวลาประมวลรวมเฉลี่ยของการสร้างและปรับปรุงโมเดลระหว่างแนวทางที่ 1, แนวทางที่สอง และ อัลกอริทึม ICMF ในแต่ละรอบของการเพิ่มของข้อมูล

กราฟในภาพที่ 12 เราพบว่า เวลาประมวลผลเฉลี่ย (Avg. Comp. Time) ในแนวทางที่ 1 น้อยกว่าแนวทางที่สอง และ อัลกอริทึม ICMF ทุกรอบของการเพิ่มข้อมูล แต่ถึงแม้ว่าเวลาประมวลผลของแนวทางที่ 1 จะน้อยที่สุด แต่ยังมีข้อเสีย คือโมเดลที่ใช้แนวทางที่ 1 ไม่มีการปรับปรุงโมเดลใหม่เลย ทำให้เมื่อมี ข้อมูลใหม่เพิ่มเข้ามา เรื่อยๆ ความแม่นยำของโมเดล มีแนวโน้มลดลง แต่เมื่อพิจารณาเฉพาะแนวทางที่ 2 และอัลกอริทึม ICMF พบว่าเวลาประมวลผลเฉลี่ยของอัลกอริทึม ICMF จะน้อยกว่าแนวทางที่สองในเกือบทุกรอบของการเพิ่มข้อมูล ยกเว้นรอบแรกที่มากกว่าเนื่องต้องสร้างเกณฑ์ในการประเมินโมเดล และเมื่อพิจารณาเวลาประมวลผลเฉลี่ยรวมของ อัลกอริทึม ICMF พบว่าเวลาประมวลเฉลี่ยต่ำกว่าของ Framework2 มาก

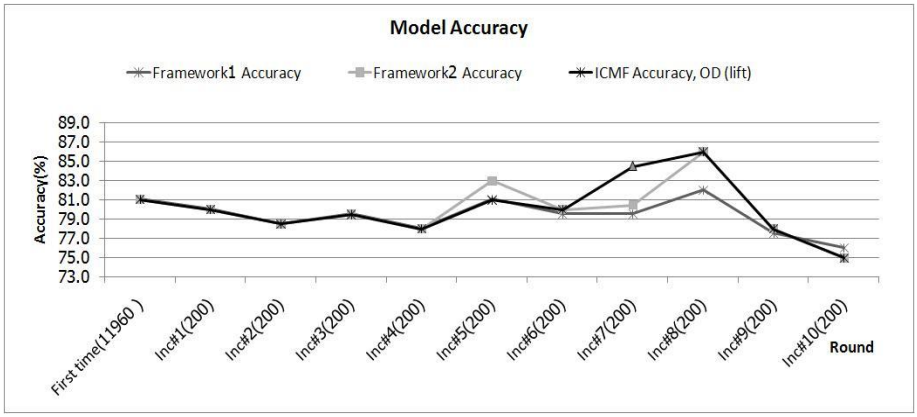
2.2 ความแม่นยำของโมเดลในแต่ละรอบของการเพิ่มของข้อมูล



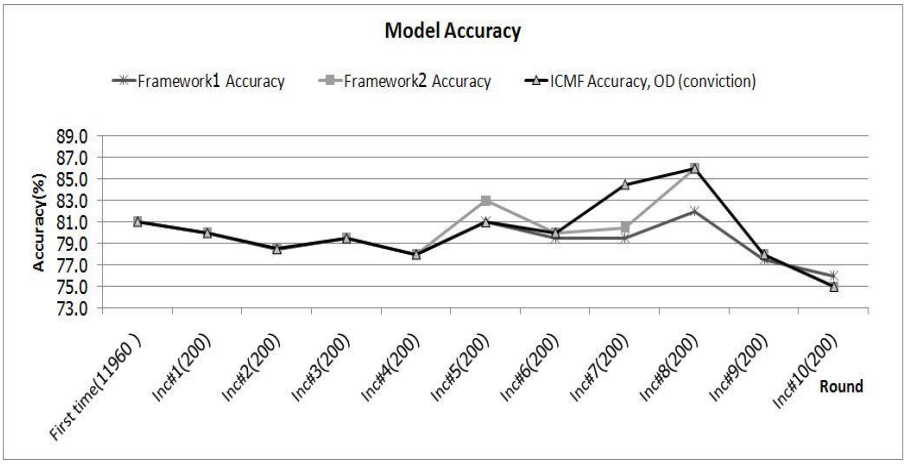

**ภาพที่ 13** แสดงความแม่นยำของโมเดลระหว่างแนวทางที่ 1, แนวทางที่สอง และ อัลกอริทึม ICMF, OD (Support) ในแต่ละรอบของการเพิ่มของข้อมูล




**ภาพที่ 14** แสดงความแม่นยำของโมเดลระหว่างแนวทางที่ 1, แนวทางที่สอง และ อัลกอริทึม ICMF, OD (Confidence) ในแต่ละรอบของการเพิ่มของข้อมูล




**ภาพที่ 15** แสดงความแม่นยำของโมเดลระหว่างแนวทางที่ 1, แนวทางที่สอง และ อัลกอริทึม ICMF, OD (Lift) ในแต่ละรอบของการเพิ่มของข้อมูล




**ภาพที่ 16** แสดงความแม่นยำของโมเดลระหว่างแนวทางที่ 1, แนวทางที่สอง และ อัลกอริทึม ICMF, OD (Conviction) ในแต่ละรอบของการเพิ่มของข้อมูล

กราฟในภาพที่ 13, 14, 15 และ 16 เราสามารถพิจารณาความแม่นยำของแนวทางที่ 1 เปรียบเทียบกับอัลกอริทึม ICMF เราพบว่าค่าความแม่นยำของของโมเดลจะน้อยกว่าหรือเท่ากับอัลกอริทึม ICMF เกือบทุกรอบของการเพิ่มข้อมูล ฉะนั้นจากผลการทดลองแสดงว่า เมื่อมีการเพิ่มของข้อมูล อัลกอริทึม ICMF มีความแม่นยำเฉลี่ยสูงกว่า แนวทางที่ 1

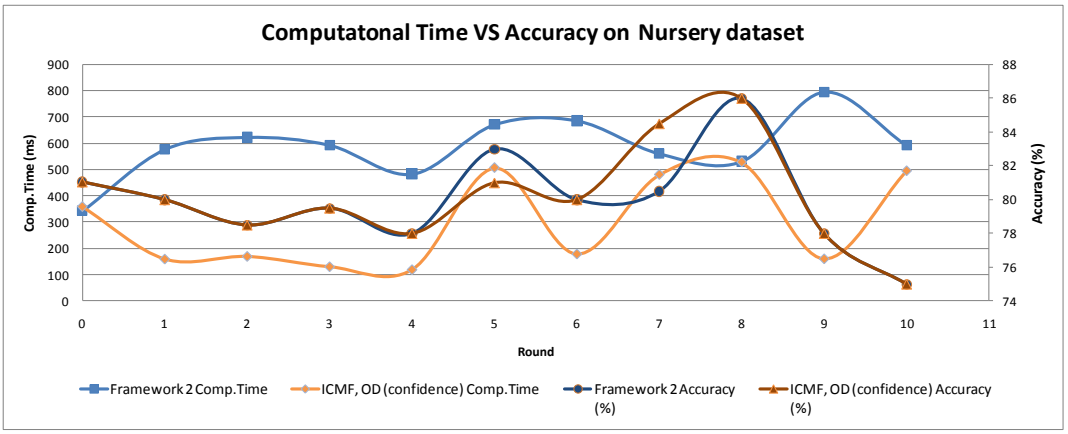
กราฟในภาพที่ 14, 15 และ 16 เราสามารถพิจาณาความแม่นยำของ แนวทางที่ 2 เปรียบเทียบกับ อัลกอริทึม ICMF เราพบว่าในบางรอบคือรอบที่ Inc#5 ความแม่นยำของแนวทางที่ 2 มีความแม่นยำมากกว่าอัลกอริทึม ICMF แต่หลังจากนั้นอัลกอริทึม ICMF ก็สามารถปรับตัวเองให้มีความแม่นยำใกล้เคียงหรือเท่ากับแนวทางที่ 2 แต่พบว่าบางรอบคือรอบ Inc#7 พบว่าอัลกอริทึม ICMF มีความแม่นยำมากกว่า แนวทางที่ 2 เนื่องจากในรอบ Inc#6, Inc#7 อัลกอริทึม ICMF ใช้โมเดลที่เกิดจากสร้างใหม่ในรอบ Inc#5 ขณะที่แนวทางที่ 2 ใช้โมเดลที่เกิดจากสร้างใหม่ในรอบ Inc#6 ซึ่งเป็นคนละโมเดลกันทำให้ความแม่นยำของโมเดลแตกต่างกัน หรือในภาพที่ 13 รอบ Inc#9 และ Inc#10 อัลกอริทึม ICMF มีความแม่นยำมากกว่า แนวทางที่ 2

จากผลการทดลองที่กล่าวมา เราสามารถสรุปได้ว่าบางรอบของการเพิ่มของข้อมูล การปรับปรุงโมเดลใหม่ทุกครั้งของ แนวทางที่ 2 ความแม่นยำอาจจะไม่เพิ่มขึ้น การทำใหม่ทุกครั้ งไม่จำเป็น แต่เมื่อมีข้อมูลเพิ่มเข้าเรื่อยๆ ก็จำเป็นต้องปรับปรุงโมเดล ซึ่ง อัลกอริทึม ICMF สามารถรองรับความต้องการดังกล่าว โดย อัลกอริทึม ICMF ทำนายโดยใช้โมเดลเดิมสำหรับข้อมูลที่เพิ่มเข้ามา แต่เมื่อข้อมูลใหม่ที่เพิ่มเข้ามา ที่ทำให้โมเดลเปลี่ยน อัลกอริทึม ICMF ก็สามารถที่จะบอกได้

ว่าควรจะปรับปรุงโมเดลใหม่ จากเกณฑ์ค่าลำดับต่างกันสูงสุด (MOD) ทำให้โมเดลมีความทันสมัยกับข้อมูลที่เพิ่มเข้ามาในรอบถัดไป

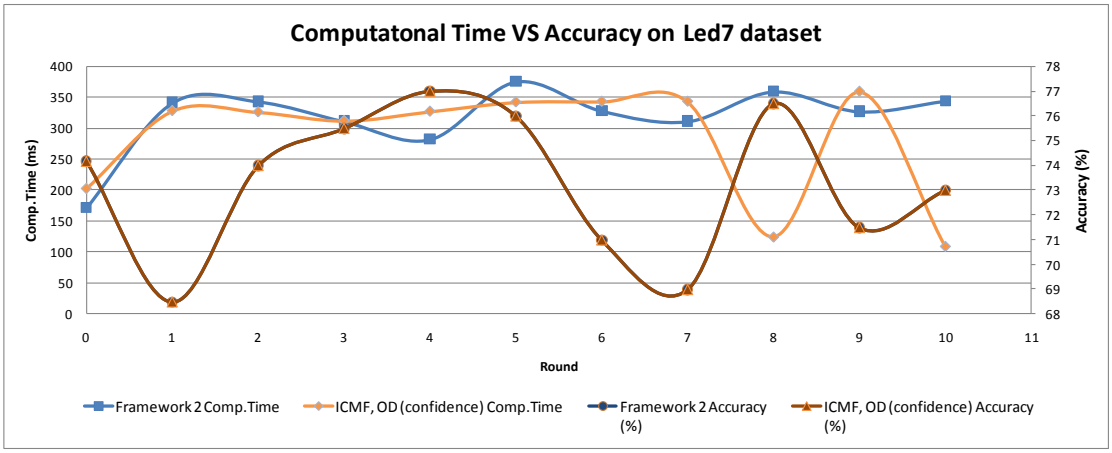
2.3 ความสัมพันธ์ระหว่างความแม่นยำและเวลาประมวล

เราสนใจความสัมพันธ์ระหว่างความแม่นยำและเวลาประมวลรวมในแต่ละรอบของการเพิ่มของข้อมูล โดยพิจารณ าเฉพาะ Balanced Datasets ได้แก่ดาต้าเซ็ต Led7, Nursery และ PenDigits เนื่องจากค่าความแม่นยำเฉลี่ยจากตารางที่ 19 และเวลาประมวลเฉลี่ยจากตารางที่ 20 ของแนวทางที่ 2 และอัลกอริทึม ICMF ที่เรียงกฎด้วยค่าความเชื่อมั่น (Confidence) มีค่าความแม่นยำเฉลี่ยสูงและเวลาประมวลเฉลี่ยผลใกล้เคียงกัน



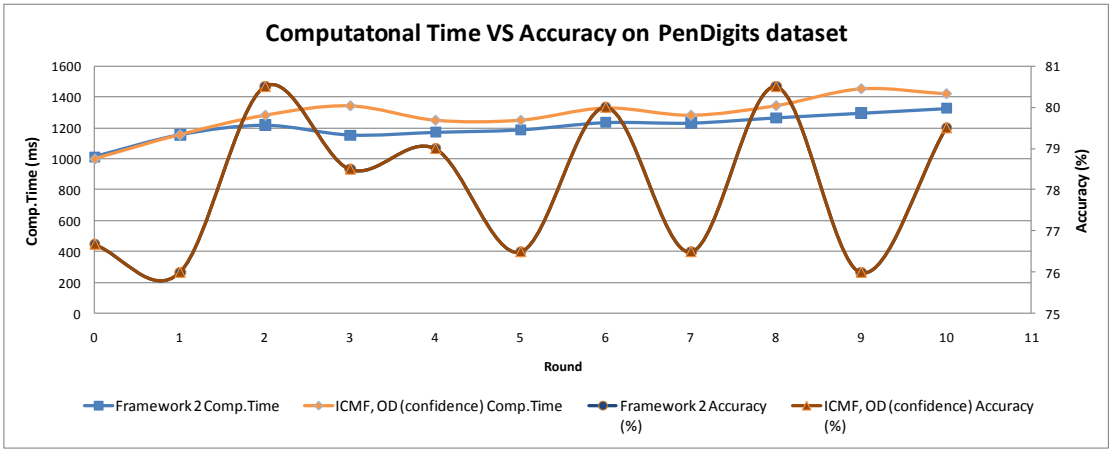

**ภาพที่ 17** แสดงเวลาประมวลผลและความแม่นยำของโมเดลแต่ละรอบของการเพิ่มของข้อมูลระหว่างแนวทางที่สอง และ อัลกอริทึม ICMF, OD (Confidence) ของ Nursery ดาต้าเซ็ต

กราฟในภาพที่ 17 เราสามารถพิจารณาได้ว่าเวลาประมวลผลของอัลกอริทึม ICMF, OD (Confidence) น้อยกว่าแนวทางที่ 2 (Framework 2) อยู่เกือบทุกรอบของการเพิ่มของข้อมูลและความแม่นยำของโมเดลมีค่าเท่ากับหรือสูงกว่าแนวทางที่ 2 ในเกือบทุกรอบของการเพิ่มของข้อมูล เราสามารถสรุปได้ว่าการประมวลผลเพื่อปรับปรุงโมเดลทุกรอบของการเพิ่มของข้อมูลในแนวทางที่ 2 ไม่ทำให้ความแม่นยำของโมเดลเพิ่มขึ้นทุกครั้ง การปรับปรุงโมเดลทุกรอบของการเพิ่มของข้อมูลจึงไม่จำเป็น




**ภาพที่ 18** แสดงเวลาประมวลผลและความแม่นยำของโมเดลแต่ละรอบของการเพิ่มของข้อมูลระหว่างแนวทางที่สอง และ อัลกอริทึม ICMF, OD (Confidence) ของ Led7 ดาต้าเซ็ต

กราฟในภาพที่ 18 เราสามารถพิจารณาได้การประมวลผลเพื่อปรับปรุงโมเดลทุกรอบของการเพิ่มของข้อมูลในแนวทางที่ 2 ไม่ทำให้ความแม่นยำของโมเดลเพิ่มขึ้นทุกครั้ง การปรับปรุงโมเดลทุกรอบของการเพิ่มของข้อมูลจึงไม่จำเป็น และการไม่ปรับปรุงโมเดล ในบางรอบการเพิ่มของข้อมูลของอัลกอริทึม ICMF, OD (Confidence) ก็ยังให้ความแม่นยำของโมเดลเทียบเท่าแนวทางที่ 2




**ภาพที่ 19** แสดงเวลาประมวลผลและความแม่นยำของโมเดลแต่ละรอบของการเพิ่มของข้อมูลระหว่างแนวทางที่สอง และ อัลกอริทึม ICMF, OD (Confidence) ของ PenDigits ดาต้าเซ็ต

กราฟในภาพที่ 19 เราสามารถพิจารณาได้ว่า การประมวลผลเพื่อปรับปรุงโมเดลทุกรอบของการเพิ่มของข้อมูลในแนวทางที่ 2 ไม่ทำให้ความแม่นยำของโมเดลเพิ่มขึ้นทุกครั้ง การปรับปรุงโมเดลทุกรอบของการเพิ่มของข้อมูลจึงไม่จำเป็น และเมื่อเราสังเกตความแม่นยำของโมเดลจะพบว่ามีค่าเปลี่ยนขึ้นลงในเกือบทุกรอบของการเพิ่มของข้อมูล ทำให้ อัลกอริทึม ICMF, OD (Confidence) ต้องปรับปรุงโมเดลทุกครั้งเนื่องจากความแม่นยำของโมเดลเปลี่ยนไป อย่างไรการปรับปรุงโมเดลทุกครั้งอัลกอริทึม ICMF, OD (Confidence) ใช้เวลามากกว่าแนวทางที่ 2 เล็กน้อย เนื่องจากต้องใช้เวลาส่วนหนึ่งเพื่อพิจารณาค่าลำดับต่างกันของกฎความสัมพันธ์ที่มีความยาวสองและมีคลาส R2

จากผลการทดลองที่กล่าวมาในรูปที่ 17-19 เราสามารถสรุปได้ว่า อัลกอริทึม ICMF, OD (Confidence) สามารถปรับปรุงโมเดลให้มีความแม่นยำของโมเดลสอดคล้องกับความแม่นยำของโมเดลใหม่ที่สร้างจากการเพิ่มข้อมูล แต่ในบางกรณีความแม่นยำของโมเดลใหม่อาจจะเพิ่มขึ้นหรือลดลงก็ได้ และ อัลกอริทึม ICMF, OD (Confidence) ใช้เวลาประมวลผลน้อยกว่าแนวทางที่ 2 มาก ในรอบที่ไม่มีปรับปรุงโมเดล และมากกว่าเล็กน้อยในรอบที่มีการปรับปรุงโมเดล

2.4 การตั้งค่าลำดับต่างกันสูงสุด (MOD)

สำหรับอัลกอริทึม ICMF นี้ การตั้ง ค่าลำดับต่างกันสูงสุด (MOD) มีความสำคัญต่อความแม่นยำเฉลี่ยของโมเดลและเวลาประมวลรวมเฉลี่ย เนื่องจากถ้าตั้งค่าลำดับต่างกันสูงสุดไว้มาก โมเดลจะไม่มีการปรับปรุงเลย ทำให้ประสิทธิภาพเทียบเท่าแนวทา งที่ 1 ซึ่งให้เวลาประมวลรวมเฉลี่ยน้อยแต่ความแม่นยำเฉลี่ยของโมเดลจะน้อยด้วย ในทางกลับกันถ้าตั้งค่าลำดับต่างกันสูงสุดไว้มาก โมเดลจะปรับปรุงทุกครั้งทุกรอบของการเพิ่มของข้อมูล ทำให้ประสิทธิภาพเทียบเท่าแนวทางที่ 2 ซึ่งให้เวลาประมวลรวมเฉลี่ยมากและความแม่นยำเฉลี่ยของโมเดลมากด้วย แต่ในบางกรณีแนวทางที่ 1 อาจจะทำให้เวลาประมวลรวมเฉลี่ยและความแม่นยำเฉลี่ยของโมเดลดีกว่าแนวทางที่ 2เนื่องจากข้อมูลที่เข้าผิดปกติ ทำให้แนวทางที่ 2 ที่นำข้อมูลที่เข้าไปใช้สร้างโมเดลใหม่สร้างโมเดลที่ความคลาดเคลื่อนจากความเป็นจริงด้วย ในที่นี้ จะทดลองตั้งค่าลำดับต่างกันสูงสุดของอัลกอริทึม (MOD) ที่พิจารณาค่าลำดับต่างกันของกฎที่เรียงด้วยค่าสนับสนุน (Support)

**ตารางที่ 21** เปรียบเทียบเวลาประมวลผลและความแม่นยำของโมเดล เมื่อตั้งค่า MOD ต่างกันและมีการเพิ่มของข้อมูล 10 ครั้ง ของ Nursery ดาต้าเซ็ต

| Compared | MOD=67.5 | | MOD=135 | | MOD=202.5 | |
|---|---|---|---|---|---|---|
| | Total | Average | Total | Average | Total | Average |
| Comp. Time(ms) | 3974 | 361.27 | 3383 | 307.55 | 2935 | 266.82 |
| Accuracy (%) | | 79.78 | | 80.46 | | 80.41 |
| # rebuilt classifier | 6 | | 4 | | 3 | |

ตารางที่ 21 เราจะเห็นว่า ถ้าเราเปรียบเทียบระหว่างค่า MOD=135 กับ MOD=202.5 ค่า MOD ที่ต่ำ (MOD=135) จะมีข้อดีคือความแม่นยำเฉลี่ยจะสูง แต่มีข้อเสียคือ เวลาในการประมวลผลจะสูงกว่าการค่า MOD ที่สูง (MOD=202.5) แต่มีกรณียกเว้น สำหรับค่า MOD=67.5 การตั้งค่า MOD ไว้ต่ำมาก โมเดลจะมีการปรับปรุงทุกรอบของการเพิ่มของข้อ มูล ซึ่งการปรับปรุงโมเดลใหม่ทุกรอบของการเพิ่มของข้อมูล อาจจะไม่ทำความแม่นยำเฉลี่ยเพิ่มขึ้น เนื่องจากโมเดลที่สร้างขึ้นมาใหม่แต่ละครั้ง อาจมีบางรอบการเพิ่มของข้อมูล มีข้อมูลที่เข้าผิดปกติ อย่างเช่น มีข้อมูลบางคลาสเพิ่มขึ้นมาผิดปกติ ทำให้ default class เปลี่ยนไป และมีการทำนายผิลพลาด จึงเป็นสาเหตุทำให้ความแม่นยำเฉลี่ยของโมเดลลดลง

## วิจารณ์

จากผลการทดลอง เราสามารถสรุปว่า กรณีมีการเพิ่มของข้อมูล อัลกอริทึม ICMF (กฤษฎา กร และคณะ , 2550) มีประสิทธิภาพในแง่ของการความแม่นยำเฉลี่ยของโมเดลดีกว่าแนวทางที่ 1 (ไม่ปรับปรุงโมเดล) และแนวทางที่ 2 (ปรับปรุงโมเดลทุกรอบของการเพิ่มของข้อมูลเข้ามา) เมื่อเราพิจารณากรณีข้อมูลที่เพิ่มเข้ามามีลักษณะการกระจายตัวหรือรูปแบบไอเท็มเซ็ตคล้ายกับก่อนและหลังการเพิ่มของข้อมูล อัลกอริทึม ICMF สามารถที่จะใช้ค่าเกณฑ์ค่าลำดับต่างกันสูงสุด (MOD) พิจารณาว่าไม่ควรปรับปรุงโมเดล ทำให้ยังรักษาความแม่นยำของโมเดลเฉลี่ยไม่ด้อยกว่าแนวทางที่ 1 และแนวทางที่ 2 และเวลาประมวลผลที่ลดลงน้อยกว่าแนวทางที่ 2 และเทียบเท่าแนวทางที่ 1 และในทางกลับกันกรณีก่อนและเพิ่มของข้อมูลมีลักษณะการกระจายตัวหรือรูปแบบไอเท็มเซ็ตแตกต่างกันมาก อัลกอริทึม ICMF สามารถที่จะใช้ค่าเกณฑ์ค่าลำดับต่างกันสูงสุด (MOD) พิจารณา

ว่าควรปรับปรุงโมเดล ทำให้มีความแม่นยำเฉลี่ยของโมเดลใกล้เคียงแนวทางที่ 2 และสูงกว่าแนวทางที่ 1

อัลกอริทึม ICMF ที่หาค่าลำดับต่างกันของกฎที่เรียงด้วยค่าสนับสนุน(support), ค่าความเชื่อมั่น (confidence), ค่าลิฟท์ (Lift), ค่าคอนวิคชั่น (Conviction) ต่างมีปัญหาที่ในการกำหนดค่าลำดับต่างกันสูงสุด (Maximum Order Difference) ที่เหมาะสม การกำหนดค่าลำดับต่างกันสูงสุดที่เหมาะสมทำได้ยาก กรณีที่ตั้งค่าลำดับต่างกันสูงสุดไว้มาก จะทำให้ประสิทธิผลข องอัลกอริทึม ICMF เทียบเท่าแนวทางที่ 1 แต่ถ้าตั้งค่าลำดับต่างกันสูงสุดไว้ต่ำมาก จะทำให้ประสิทธิผลของอัลกอริทึม ICMF เทียบเท่าแนวทางที่ 2

ปัจจัยที่กระทบต่อประสิทธิภาพของอัลกอริทึม ICMF มีดังต่อไปนี้คือ

1. จำนวนของไอเท็มและคลาส

1 อัลกอริทึม ICMF ไม่เหมาะกับลักษณะข้อมูลที่มีจำนวนของไอเท็มและคลาสมาก เนื่องจากค่าลำดับต่างกันของกฎความสัมพันธ์ความยาวสองและมีคลาส จะเปลี่ยนแปลงไปค่อนข้างมากแม้จะมีข้อมูลเพิ่มเข้ามาเล็กน้อย ทำให้ต้องปรับปรุงโมเดลบ่อย ประสิทธิภาพเทียบเท่าแนวทางที่ 2

2 อัลกอริทึม ICMF ไม่เหมาะกับกรณีที่จำนวนกฎความสัมพันธ์ที่มีความยาวสองและมีคลาส (R2) มีจำนวนน้อยกว่า 10 กฎ (R2 เท่ากับผลคูณระหว่างจำนวนไอเท็มและจำนวนคลาสไอเท็ม) เพราะถ้าจำนวนกฎน้อยจะการคำนวณค่าลำดับต่างกันจะคำนวณได้น้อย ทำให้แทบจะไม่ปรับปรุงโมเดลเลย ประสิทธิภาพเทียบเท่าแนวทางที่ 1

2. จำนวนเร็คคอร์ดหรือทรานเซ็กชั่นที่เพิ่มข้ามา

1. อัลกอริทึม ICMF ไม่เหมาะกับในกรณีที่ข้อมูลเร็คคอร์ดเพิ่มเข้ามาน้อยมาก เนื่องจากความแม่นยำของโมเดลอาจจะไม่เปลี่ยนเลย เวลาประมวลผลและความแม่นยำของโมเดลจะด้อยกว่าแนวทางที่ 1

3. จำนวนครั้งของการเพิ่มของข้อมูล

1. อัลกอริทึม ICMF ไม่เหมาะกับในกรณีที่จำนวนครั้งของการเพิ่มของข้อมูลมีจำนวนน้อย เช่น 2-3 ครั้ง เพราะ ความแม่นยำของโมเดลในแต่ละครั้งของการเพิ่มของข้อมูลอาจจะไม่เปลี่ยนเลย ประสิทธิภาพของโมเดลเทียบเท่าแนวทางที่ 1

# สรุปและข้อเสนอแนะ

## สรุป

ผู้วิจัยได้นำเสนอ แนวทางการปรับปรุงโมเดลจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์เมื่อมีการเพิ่มของข้อมูล โดยใช้ค่า ลำดับต่างกัน Order Difference (OD) ของอัลกอริทึม ICMF เพื่อพิจารณาว่าจำเป็นหรือไม่ที่ต้องปรับปรุงหรือสร้างโมเดลใหม่เมื่อมีการเพิ่มของข้อมูล จากผลการทดลองพบว่า แนวทางที่นำเสนอ ช่วยลดเวลาในการปรับปรุงโมเดลใหม่ให้มีเวลาเฉลี่ยน้อยกว่าและความแม่นยำเฉลี่ยเพิ่มสูงกว่า แนวทางที่ 2 สร้างโมเดลใหม่ทุกครั้งเมื่อมีการเพิ่มข้อมูล แต่ยังมีข้อด้อยที่เวลาเฉลี่ยมากกว่า แนวทางที่ 1, ไม่มีการปรับปรุงโมเดลเลย แต่ถ้าให้ความสำคัญความแม่นยำเป็นหลัก อัลกอริทึม ICMF จะดีกว่า แนวทางที่ 1 และแนวทางที่ 2 เพราะให้ความแม่นยำเฉลี่ยของโมเดลสูงกว่า

## ข้อเสนอแนะ

อัลกอริทึม ICMF สามารถปรับปรุง พัฒนา ต่อยอดได้ดังต่อไปนี้ คือ

1. เราสามารถนำค่าวัดผล (Interesting Measures) อื่นๆ เช่น ค่า All Confidence, Centered Confidence เป็นต้นมาใช้ในการเรียงลำดับกฎที่มีความยาวสองและมีคลาส

2. อัลกอริทึม ICMF คำนวณค่าลำดับต่างกันจากกฎความสัมพันธ์ที่เป็น Generic Associative Classification Rules เพื่อเป็นตัวแทนกฎความสัมพันธ์ทั้งหมด แทนที่จะใช้กฎที่มีความยาวสองและมีคลาส R2 เป็นตัวแทนกฎความสัมพันธ์ทั้งหมด

3. เราสามารถนำหลักการของสถิติเรื่องการทดสอบสมมติฐานมาตัดสินใจ มาใช้แทนค่าลำดับต่างกันสูงสุด MOD ว่าควรจะปรับปรุงโมเดลหรือสร้างโมเดลหรือไม่ โดยอาจทดสอบภาวะเท่ากันของค่าเฉลี่ยของ ค่าสนับสนุน , ค่าความเชื่อมั่น, ค่าลิฟท์ และค่าคอนวิคชั่น ของชุดกฎตัวอย่างก่อนและหลังการเพิ่มของข้อมูล ถ้าสร้างสมมติฐานว่าค่าเฉลี่ยก่อนและหลังการเพิ่มข้อมูลเท่ากัน ด้วยระดับนัยสำคัญ 95% แล้วคำนวณค่าตัวสถิติ เช่น แซด สกอร์ (Z-score) ตกอยู่ในบริเวณวิกฤต ก็ให้สมมติฐานเป็นเท็จ ควรจะปรับปรุงโมเดล

4. เราสามารถประยุกต์อัลกอริทึม ICMF ไปใช้กับอัลกอริทึม CMAR CPAR และ MMAC ซึ่งเป็นการสร้างโมเดลที่ละหลายๆ กฎ (Multiple rules) ที่ให้ความแม่นยำสูงกว่าอัลกอริทึม CBA

# เอกสารและสิ่งอ้างอิง


กฤษฎากร ก๋ิงอุบล, ธนาวินท์ รักธรรมานนท์ และ กฤษณะ ไวยมัย. 2550. เทคนิคการเก็บไอเท็มเซ็ตที่เกิดขึ้นบ่อย โดยพิจารณาค่าความเชื่อมั่นขั้นต่ำเพื่อรองรับการเพิ่มของข้อมูล, *ใน* **The 3rd National Conference on Computing and Information Technology Conference (NCCIT'07)**.

กฤษฎากร ก๋ิงอุบล, ธนาวินท์ รักธรรมานนท์ และ กฤษณะ ไวยมัย. 2550. แนวทางใหม่ สำหรับการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์ เมื่อมีการเพิ่มของข้อมูล, น. 596-605. *ใน* **The 11st National Computer Science and Engineering Conference (NCSEC'07).**

วีระพล หาญโชติช่วง. 2549. **ระบบการจัดกลุ่มและทำนายข้อมูลโดยใช้กฎความสัมพันธ์.** วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยเกษตรศาสตร์.

Blake, C., and C. Merz. 1998. **UCI repository of machine learning database.** Department of Information and Computer Science, University of California, Irvine, Irvine, CA. Available Source: http://www.ics.uci.edu/~mlearn/MLRepository.html, April 27, 2009.

Brin, S., R. Motwani, J. D. Ullman and S. Tsur. 1997. Dynamic Itemset Counting and Implication Rules for Market Basket Data, pp. 255-264. *In* Joan Peckham, ed. **SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data** . ACM Press, Tucson, Arizona, USA.

Cheung, D. W. 1996. Maintenance of Discovered Association Rules in Large Databases: An Incremental Updating Technique, pp. 106-114. *In* Stanley Y. W. Su, ed. **Proceedings of the Twelfth International Conference on Data Engineering** . IEEE Computer Society, Orleans, Louisiana.

Cheung, D. W., S. D. Lee and B. Kao. 1997. A General Incremental Technique for Maintaining Discovered Association Rules, pp. 185-194. *In* Rodney W. Topor and Katsumi Tanaka, eds. **Database Systems for Advanced Applications '97, Proceedings of the Fifth International Conference on Database Systems for Advanced Applications (DASFAA) 6. ed.** World Scientific, Melbourne, Australia.

Coenen, F. 2003. **LUCS-KDD CARM Discretisation/ Normalisation (DN) software.** Department of Computer Science, The University of Liverpool, UK. Available Source:

http://www.csc.liv.ac.uk/~frans/KDD/Software/LUCS-KDD-DN/lucs-kdd_DN.html, April 27, 2009.

Coenen, F. 2004. **LUCS KDD implementation of CBA (Classification Based on Associations),** Department of Computer Science, The University of Liverpool, UK. Available Source: http://www.csc.liv.ac.uk/~frans/KDD/Software/CBA/cba.html, April 27, 2009.

Geng, L. and H. J. Hamilton. 2006. Interestingness measures for data mining: A survey. **ACM Comput. Surv.** 2006 (38):

Han, J., J. Pei, Y. Yin and R. Mao. 2004. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. **Data Min. Knowl. Discov.** 2004 (8): 53-87.

Imberman, S. P., A. U. Tansel and E. Pacuit. 2004. An Efficient Method For Finding Emerging Large Itemsets, *In* **Proceeding of the Third Workshop on Mining Temporal and Sequential Data, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining** .

Lenca, P., P. Meyer, B. Vaillant and S. Lallich. 2008. On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. **European Journal of Operational Research.** 2008 (184): 610-626.

Li, W., J. Han and J. Pei. 2001. CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules., pp. 369-376. *In* Nick Cercone, Tsau Young Lin and Xindong Wu, eds. **Proceedings of the 2001 IEEE International Conference on Data Mining, 29 November - 2 December 2001, San Jose, California, USA** . IEEE Computer Society.

Liu, B., W. H., and Y. Ma. 1998. Integrating classification and association rule mining. *In* **Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD'98)** 4: 80-86.

Quinlan, J.R. 1993. C4.5: Programs for machine learning. **Morgan Kaufmann**

Rakesh, A., and R. Srikant. 1994. Fast algorithm for mining association rules in large databases. *In* **Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94)** Santiago, Chile.

Rakesh, A., T. Imielinski and A. N. Swami. 1993. Mining Association Rules between Sets of Items in Large Databases, pp. 207-216. *In* Peter Buneman and Sushil Jajodia, eds. **Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data** . ACM Press, Washington, D.C.

Sheikh, L. M., B. Tanveer and S. M. A. Hamdani. 2004. Interesting measures for mining association rules, *In* **Proceedings of IEEE International Multi Topic Conference (IEEE INMIC 2004)** . Piscataway, NJ.

Thabtah, F. A. 2007. A review of associative classification mining,. **Knowledge Eng. Review** 2007 (22): 37-65.

Thabtah, F. A. 2006. Challenges and Interesting Research Directions in Associative Classification, pp. 785-792. *In* **Workshops Proceedings of the 6th IEEE International Conference on Data Mining (ICDM 2006), 18-22 December 2006, Hong Kong, China** . IEEE Computer Society.

Wang, K., S. Zhou and Y. He. 2000. Growing decision tree on support-less association rules. *In* **Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining (KDD'00)** Boston, MA, Aug.

Yin, X. and J. Han. 2003. CPAR: Classification based on Predictive Association Rules, *In* Daniel Barbara and Chandrika Kamath, eds. **Proceedings of the Third SIAM International Conference on Data Mining, San Francisco, CA, USA, May 1-3, 2003**.

# ประวัติการศึกษา และการทำงาน

| | |
|---|---|
| ชื่อ –นามสกุล | นายกฤษฎากร ก๋ิงอุบล |
| วัน เดือน ปี ที่เกิด | วันที่ 14 กรกฎาคม 2524 |
| สถานที่เกิด | นครศรีธรรมราช |
| ประวัติการศึกษา | วศ.บ. (วิศวกรรมคอมพิวเตอร์) คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยี พระจอมเกล้า เจ้าคุณทหาร ลาดกระบัง (2547) |
| ตำแหน่งหน้าที่การงานปัจจุบัน | วิศวกรอาวุโส ทดสอบคุณภาพโปรแกรม |
| สถานที่ทำงานปัจจุบัน | บริษัท รอยเตอร์ ซอฟท์แวร์ ไทยแลนด์ จำกัด |
| ผลงานดีเด่นและรางวัลทางวิชาการ | - |
| ทุนการศึกษาที่ได้รับ | - |