ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)

ปริญญา

วิศวกรรมคอมพิวเตอร์	วิศวกรรมคอมพิวเตอร์
สาขา	**ภาควิชา**

เรื่อง แนวทางการปรับปรุงประสิทธิภาพของการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์บนฐานข้อมูลที่ไม่สมดุล

An Approach for Improving Associative Classification in Imbalanced Datasets

นามผู้วิจัย นายพูนเพิ่ม สุวรรณรัฐภูมิ

ได้พิจารณาเห็นชอบโดย

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก .....

(รองศาสตราจารย์กฤษณะ ไวยมัย, Ph.D.)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม .....

(อาจารย์สุภาพร เอื้อจงมานี, Ph.D.)

หัวหน้าภาควิชา .....

(ผู้ช่วยศาสตราจารย์ภุชงค์ อุทโยภาศ, Ph.D.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

.....

(รองศาสตราจารย์กัญจนา ธีระกุล, D.Agr.)

คณบดีบัณฑิตวิทยาลัย

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

วิทยานิพนธ์

เรื่อง

แนวทางการปรับปรุงประสิทธิภาพของการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์บนฐานข้อมูลที่ไม่สมดุล

An Approach for Improving Associative Classification in Imbalanced Datasets

โดย

นายพูนเพิ่ม สุวรรณรัฐภูมิ

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์
เพื่อความสมบูรณ์แห่งปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)
พ.ศ. 2555

พูนเพิ่ม สุวรรณรัฐภูมิ 2555: แนวทางการปรับปรุงประสิทธิภาพของการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์บนฐานข้อมูลที่ไม่สมดุล ปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์) สาขาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก: รองศาสตราจารย์กฤษณะ ไวยมัย, Ph.D. 85 หน้า

การจำแนกประเภทด้วยกฎความสัมพันธ์เป็นหนึ่งในตัวจำแนกประเภทกฎ ซึ่งนำไปประยุกต์ใช้ในงานจริงหลายๆงานด้วยกัน ประโยชน์อย่างหนึ่งของตัวจำแนกประเภทด้วยกฎความสัมพันธ์คือง่ายต่อการตีความจากกฎการจำแนก อย่างไรก็ตามยังคงต้องพัฒนาการจำแนกประเภทด้วยกฎความสัมพันธ์เมื่อนำไปใช้ในงานจำแนกประเภทข้อมูลที่ไม่สมดุล ในหลายๆ งาน เช่นการวินิจฉัยทางการแพทย์ คลาสรองคือคลาสแรกที่เราสนใจและมีต้นทุนของการจำแนกผิดพลาดสูงมากกว่าคลาสหลัก ในงานวิจัยนี้ได้นำเสนอ SSCR เป็นการปรับปรุงประสิทธิภาพของการจำแนกประเภทด้วยกฎความสัมพันธ์ในชุดข้อมูลที่ไม่สมดุล SSCR ได้รวมเอากฎความสัมพันธ์ที่มีนัยความสำคัญทางสถิติเข้าไว้ด้วยกันกับการเรียนรู้แบบมีต้นทุนเพื่อสร้างเป็นตัวจำแนกประเภทด้วยกฎความสัมพันธ์ โดยผลการทดลองจะแสดงให้เห็นว่า SSCR ให้ประสิทธิภาพที่ดีกับชุดข้อมูลจริงที่ไม่สมดุลเมื่อเปรียบเทียบกับ CBA และ C4.5

/ /

ลายมือชื่อนิสิต ลายมือชื่ออาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

Phoonperm Suwannarattaphoom 2012: An Approach for Improving Associative Classification in Imbalanced Datasets. Master of Engineering (Computer Engineering), Major Field: Computer Engineering, Department of Computer Engineering. Thesis Advisor: Associate Professor Kitsana Waiyamai, Ph.D. 85 pages.

Associative classification is one of rule-based classifiers that has been applied in many real-world applications. One advantage of the associative classifier is its easy interpretability in terms of classification rules. However, there is room for improvement when associative classification is applied in the imbalanced classification task. In many applications such as medical diagnosis, the minority class can be the class of primary interest and it has a much higher misclassification cost than the majority class. Existing associative classification algorithms can be limited in their performance on highly imbalanced datasets in which the class of interest is a minority class. In this paper we present an approach, SSCR for improving associative classification in imbalanced data sets. SSCR combines statistically significant association rules with cost-sensitive learning to build as associative classifier. Experimental results show that SSCR archives best performance on real-world imbalanced datasets, compared with CBA and C4.5.

_____ _____ ___ / ___ / ___

Student's signature Thesis Advisor's signature

กิตติกรรมประกาศ

ข้าพเจ้าขอกราบขอบพระคุณอาจารย์กฤษณะ ไวยมัย อาจารย์ที่ปรึกษาวิทยานิพนธ์หลักที่ได้ประสิทธิ์ประสาทวิชาความรู้และทฤษฎีต่าง ๆ ตลอดจนเป็นที่ปรึกษาในการวางแนวทาง วางแผนงานแก้ปัญหา และตรวจสอบข้อบกพร่องต่าง ๆ จนงานวิจัยนี้สำเร็จลุล่วง ขอกราบขอบพระคุณ อาจารย์จิตร์ทัศน์ ฝักเจริญผล และ อาจารย์สุภาพร เอื้อจงมานี อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ตลอดจนอาจารย์ในภาควิชาวิศวกรรมคอมพิวเตอร์ ที่กรุณาให้คำปรึกษา และข้อเสนอแนะต่าง ๆ จนส่งผลให้งานวิจัยนี้สมบูรณ์ยิ่งขึ้นในทุก ๆ ด้าน และประสบผลสำเร็จลุล่วงไปด้วยดี

ขอขอบคุณสมาชิกห้องปฏิบัติการ DAKDL คุณท่านที่คอยแนะนำให้คำปรึกษาและคำอธิบายที่เป็นประโยชน์ต่องานวิจัย ไม่ว่าจะเป็นทิศทางของงานวิจัย จุดแข็งและจุดด้อย ของงานวิจัย ตลอดทั้งความรู้และทฤษฎีต่าง ๆ ที่จำเป็นสำหรับงานวิจัยและให้กำลังใจที่ดีเสมอมา

ขอขอบคุณเจ้าหน้าที่โครงการบัณฑิตศึกษา และเจ้าหน้าที่ธุรการภาควิชาวิศวกรรมคอมพิวเตอร์มหาวิทยาลัยเกษตรศาสตร์ที่ช่วยเหลือในการประสานงาน และดำเนินงานด้านเอกสารต่างๆ ให้เป็นไปอย่างสะดวกลุล่วงไปด้วยดี

คุณงามความดีหรือประโยชน์อันใดเนื่องจากวิทยานิพนธ์เล่มนี้ ขออุทิศให้แก่ บิดา มารดา พี่น้อง ญาติสนิทมิตรสหาย ตลอดทั้งครูอาจารย์และผู้มีพระคุณทุกท่าน ที่ได้อบรมและให้กำลังใจเสมอมาในทุกๆ เรื่อง

พูนเพิ่ม สุวรรณรัฐภูมิ

เมษายน 2555

สารบัญ

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญภาพ

สารบัญภาพ (ต่อ)

คำอธิบายสัญลักษณ์และคำย่อ

Atts	=	Attributes
AUC	=	Area under the ROC curve
CARs	=	Class-Association Rules
CBA	=	Classification Based on Associations Algorithm
CCR	=	Class Correlation Ratio
CCS	=	Complement Class Support
Cls	=	Classes
conf	=	Confident
corr	=	Correlation
FET	=	Fisher Exact Test
FN	=	False Negative
FNR	=	False Negative Rate
FP	=	False Positive
FTP	=	False Positive Rate
H_0	=	Null Hypothesis
H_1	=	Alternative Hypothesis
IMAC	=	Imbalanced Associative Classification
Ins	=	Instances
IR	=	Imbalanced Ratio
Ncs	=	Negative Classes
Pcs	=	Positive Classes
p_{value}	=	Fisher Exact Test Probability
TN	=	True Negative
TNR	=	True Negative Rate
TP	=	True Positive
TPR	=	True Positive Rate
ROC	=	Receiver Operating Characteristic Curve
ROC Area	=	Area under the ROC curve
α	=	Significant Level

คำอธิบายสัญลักษณ์และคำย่อ (ต่อ)

SPARCCC	=	Significant, Positively Associated and Relatively Class Correlated Classification
SSCR	=	Statistically Significant Cost-sensitive Rules
supp	=	Support
σ	=	Support Count

แนวทางการปรับปรุงประสิทธิภาพของการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์บนฐานข้อมูลที่ไม่สมดุล

An Approach for Improving Associative Classification in Imbalanced Datasets

คำนำ

ตัวจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์นั้นเป็นหนึ่งในตัวจำแนกประเภทที่ได้รับความนิยมเนื่องจากรูปแบบของผลลัพธ์ที่ได้อยู่ในรูปของกฎความสัมพันธ์ของข้อมูลซึ่งทำให้ง่ายต่อการตีความหมายและเหมาะกับฐานข้อมูลที่มีขนาดใหญ่ โดยตัวจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์เกิดจากการรวมเทคนิคการสืบค้นกฎความสัมพันธ์และการจำแนกข้อมูลเข้าไว้ด้วยกัน ซึ่งสามารถจำแนกประเภทและให้ความแม่นยำได้เป็นอย่างดี แต่เมื่อถูกนำมาใช้จำแนกประเภทบนฐานข้อมูลที่ไม่สมดุล ซึ่งก็คือกลุ่มของข้อมูลมีสัดส่วนไม่เท่ากันโดยคลาสที่เราให้ความสนใจมีสัดส่วนข้อมูลน้อยมากๆ เมื่อเทียบกับคลาสอื่นๆ จึงส่งผลให้ตัวจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์นั้นทำนายคลาสของกลุ่มที่มีข้อมูลน้อยๆ นั้นเกิดความผิดพลาดสูง ซึ่งในปัจจุบันงานวิจัยที่เกี่ยวข้องกับปัญหาการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์บนข้อมูลที่ไม่สมดุลและคลาสที่มีข้อมูลน้อยมากๆ คือคลาสที่เราให้ความสนใจเป็นพิเศษนั้นยังมีอยู่น้อยมาก ยังคงเป็นปัญหาที่ต้องค้นคว้าและวิจัยเพื่อปรับปรุงประสิทธิภาพให้ดียิ่งขึ้นต่อไป

ปัญหาความไม่สมดุลของคลาส โดยทั่วไปแล้วเกิดจากชุดข้อมูลที่มีการกระจายตัวของข้อมูลลักษณะลาดเอียงสูง (Skewed Distribution) และมีคลาสรองเป็นสัดส่วนที่น้อยมากๆ เมื่อเทียบกับคลาสหลัก ปัจจัยหลักอาจจะมาจากการมีข้อมูลที่ไม่เพียงพอ (Absolute Rarity) ซึ่งพบว่าคลาสรองที่เราสนใจนั้นมีความถี่ในการเกิดขึ้นอยู่น้อยมาก จึงส่งผลทำให้การหาความสัมพันธ์ของกฎที่เกี่ยวข้องกับคลาสรองนั้นทำได้ยาก (Relative Rarity) หรืออาจจะหาไม่พบ อันเนื่องมาจากมาตรวัดต่างๆ ที่ถูกนำมาใช้เพื่อสืบหากฎความสัมพันธ์นั้นเอง ดังนั้นจึงแสดงให้เห็นว่าชุดข้อมูลที่มีความไม่สมดุลหากเราเลือกใช้มาตรวัดที่ไม่เหมาะสม (Improper Evaluation Metrics) อาจจะทำให้เราสูญเสียความสัมพันธ์ของกฎที่น่าสนใจไป (Weiss, 2004) ส่งผลให้ตัวจำแนกประเภทของเรามีประสิทธิภาพที่แย่ลง

ตัวจำแนกประเภทด้วยกฎความสัมพันธ์ (Associative Classifiers) ส่วนใหญ่นิยมใช้การสืบค้นกฎความสัมพันธ์จากค่าสนับสนุน (Support) และ ค่าความเชื่อมั่น (Confidence) ในการรวบรวมกฎและจัดลำดับความสำคัญของกฎ ถึงแม้วิธีดังกล่าวจะให้ความแม่นยำที่ดี แต่พบว่าเมื่อชุดข้อมูลมีลักษณะที่ไม่สมดุลสูงและคลาสรองคือคลาสที่เราสนใจ พบว่าตัวจำแนกด้วยกฎความสัมพันธ์หลายๆ ตัว ให้ผลลัพธ์ที่ดีกับคลาสหลักแต่ให้ผลลัพธ์ที่แย่ลงกับคลาสรอง นำมาซึ่งการจำแนกประเภทข้อมูลที่เราสนใจผิดพลาด ซึ่งจะเห็นว่ามีต้นทุนในการจำแนกผิดพลาดที่สูงกว่า สาเหตุใหญ่เกิดจากชุดข้อมูลที่ประกอบด้วยคลาสหลักซึ่งมีสัดส่วนมากกว่าคลาสรองอยู่มาก และความสัมพันธ์ของกฎที่อยู่ในคลาสรองก็มีอยู่น้อยมากเช่นกัน ข้อเสียของการนำค่าสนับสนุนมาใช้เพียงอย่างเดียวก็คือ กฎที่มีความน่าสนใจที่มีค่าสนับสนุนต่ำอาจจะถูกกำจัดออกไปจากการแบ่งเกณฑ์ด้วยค่าสนับสนุนขั้นต่ำ (Tan *et al*., 2005) ดังนั้นการใช้ค่าสนับสนุนและค่าความเชื่อมั่นสำหรับค้นหาและตัดกฎทิ้งจะส่งผลให้การค้นหากฎความสัมพันธ์นั้นมีความโน้มเอียงเข้าหาคลาสหลัก ทำให้ความสัมพันธ์ของกฎในคลาสรองถูกกำจัดออกไป นั้นหมายถึงการสูญเสียกฎที่มีนัยสำคัญต่อคลาสรอง

จากการค้นคว้างานวิจัยส่วนใหญ่พบว่า ในการค้นหากฎที่มีนัยสำคัญบนชุดข้อมูลที่ไม่สมดุล ตัวจำแนกประเภทหลายๆ งานวิจัย ได้อาศัยสหสัมพันธ์เชิงบวก (Positively Correlation) มาช่วยในการพิจารณาเช่น Complement Class Support: CCS (Arunasalam and Chawla, 2006), Class Correlation Ratio: CCR (Verhein and Chawla, 2007) ซึ่งแสดงให้เห็นว่าส่งผลที่ดีต่อชุดข้อมูลที่มีความไม่สมดุล แต่ในการพิจารณาความสัมพันธ์ของกฎที่เป็นสหสัมพันธ์เชิงบวกเพียงอย่างเดียวนั้นพบว่ามีจำนวนกฎอยู่มากมายที่ถูกสร้างขึ้นมา ถึงแม้ว่าเราจะกำจัดกฎที่เป็นสหสัมพันธ์เชิงลบ (Negatively Correlation) ออกไปแล้วก็ตาม ซึ่งจำนวนกฎที่ถูกสร้างขึ้นมาอย่างมากมายนี้ประกอบไปด้วยกฎที่เป็นประโยชน์ (Productive Rules) และกฎที่ไม่มีประโยชน์ (Unproductive Rules) ในงานวิจัยที่ผ่านมาจึงได้มีการนำเสนอเทคนิคทางสถิติเพื่อค้นหากฎที่มีนัยสำคัญที่น่าสนใจ ซึ่ง Webb (2006) ได้ทำการทดสอบสมมุติฐานทางสถิติโดยใช้ Fisher Exact Test (FET) (Fisher, 1922) เพื่อกำจัดกฎที่ไม่เป็นประโยชน์ออกไป จากงานวิจัยของ Webb ได้มีการนำไปใช้ประโยชน์ในการสร้างตัวจำแนกประเภทบนชุดข้อมูลที่ไม่สมดุลเช่น SPARCCC (Verhein and Chawla, 2007) ซึ่งเห็นได้ชัดเจนว่ากฎที่มีนัยสำคัญทางสถิติเป็นกฎที่มีศักยภาพน่าสนใจจริงๆ โดย SPARCCC ได้อาศัยกฎที่มีนัยสำคัญทางสถิติร่วมกับมาตรวัด CCR ซึ่งเป็นมาตรวัดเพื่อใช้หาอัตราส่วนของสหสัมพันธ์เชิงบวกกับคลาสที่กฎนั้นทำนายเมื่อเทียบกับคลาสอื่นๆ ซึ่งข้อด้อยของ CCR ก็คือไม่รับรองเรื่องค่าความผิดพลาดประเภทลบ (False Negative Error) อันเนื่องมาจาก กฎที่มีค่า CCR สูงอาจจะมีค่าความผิดพลาดประเภทลบสูงกว่ากฎที่มีค่า CCR ที่น้อยกว่า

ในวิทยานิพนธ์เล่มนี้ ผู้วิจัยมุ่งเน้นการเพิ่มประสิทธิภาพการจำแนกประเภทคลาสรองให้มีความแม่นยำมากขึ้น โดยได้นำเสนอแนวทางการปรับปรุงการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์บนฐานข้อมูลที่ไม่สมดุล โดยอาศัยการเลือกกฎที่เป็นสหสัมพันธ์เชิงบวกเท่านั้นมาทำการทดสอบสมมุติฐานทางสถิติด้วย FET เพื่อให้ได้กฎที่เป็นประโยชน์ (Productive Rules) เท่านั้น จากนั้นนำกฎที่ได้มาทำการคิดต้นทุนด้วยค่าตารางเมตริกต้นทุน (Cost Matrix) เพื่อหาค่าต้นทุนรวมของกฎที่ทำการจำแนกผิดพลาดนั้นหมายถึงความเสี่ยงของกฎเมื่อถูกนำไปใช้ทำนายข้อมูลใหม่ (Cost-Sensitive Rule) โดยกำหนดให้ต้นทุนของความผิดพลาดที่จำแนกคลาสบวก (Positive Class) ผิดเป็นคลาสลบ (Negative Class) มีค่ามากกว่าต้นทุนของความผิดพลาดที่จำแนกคลาสลบผิดเป็นคลาสบวก ทั้งนี้เนื่องจากต้นทุนของการจำแนกคลาสที่เราสนใจผิดพลาด มีต้นทุนที่มากกว่าคลาสอื่นๆ

วัตถุประสงค์และขั้นตอนการวิจัย

วัตถุประสงค์ของการวิจัย

1. ศึกษาเทคนิคการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ ปัญหาผลกระทบของชุดข้อมูลที่ไม่สมดุลในการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์ และเทคนิคอื่นๆที่เกี่ยวข้อง เพื่อที่จะพัฒนาเทคนิคการจำแนกประเภทข้อมูลบนฐานข้อมูลที่ไม่สมดุลให้มีประสิทธิภาพมากขึ้น

2. พัฒนาเทคนิคการจำแนกประเภทข้อมูลบนฐานข้อมูลที่ไม่สมดุล ในส่วนของ

2.1. วิธีการคัดเลือกกฎความสัมพันธ์เพื่อใช้เป็นข้อมูลให้กับโมเดลการจำแนกประเภท

2.2. อัลกอริทึมสำหรับการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์

ขั้นตอนการวิจัย

1. ศึกษาทฤษฎีต่างๆ ของเทคนิคการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ เพื่อสร้างตัวจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์ให้มีประสิทธิภาพ

2. ศึกษาผลกระทบของการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์บนฐานข้อมูลที่ไม่สมดุล และคลาสรองคือคลาสแรกที่เราสนใจ เพื่อที่จะปรับปรุงความแม่นยำให้ดีที่สุด

3. ศึกษางานวิจัยก่อนหน้านี้ เพื่อวิเคราะห์ปัญหาและรวบรวมข้อดีและข้อด้อยต่างๆ เพื่อนำมาเป็นข้อมูลในการพัฒนาเทคนิคและอัลกอริทึมในการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์

4. รวบรวมฐานข้อมูลมาตรฐานที่มีสัดส่วนความไม่สมดุลของข้อมูลที่หลากหลาย เพื่อนำมาใช้ในการทดสอบและศึกษาผลลัพธ์จากโมเดลการทำนายรวมทั้งหาสาเหตุของปัญหาที่เกิดขึ้น

5. พัฒนาเทคนิคการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์บนฐานข้อมูลที่ไม่สมดุล

6. ทดสอบและวัดผลของการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์บนฐานข้อมูลที่ไม่สมดุล

7. สรุปผลการวิจัยและประโยชน์ที่ได้รับ

การตรวจเอกสาร

ความรู้พื้นฐานของดาต้าไมนิ่ง

ดาต้าไมนิ่ง (Data Mining) คือเทคโนโลยี่ที่ใช้สำหรับการค้นหาความรู้ในฐานข้อมูล ซึ่งเป็นเทคนิคหนึ่งที่ได้รับความนิยมในการค้นหารูปแบบหรือความสัมพันธ์จากข้อมูลจำนวนมหาศาลซึ่งมีขนาดใหญ่ เพื่อนำไปใช้ประโยชน์ในการวิเคราะห์ข้อมูลต่อไป เทคนิคของดาต้าไมนิ่งก่อนหน้านี้มีการนำเสนอด้วยกันหลากหลายวิธีการ โดยเทคนิคหนึ่งที่ได้รับความนิยมคือการค้นหากฎความสัมพันธ์ เนื่องจากง่ายต่อการแปลความหมายผลลัพธ์จากรูปแบบของกฎความสัมพันธ์

1. การสืบค้นกฎความสัมพันธ์ (Association Rule Discovery)

การสืบค้นกฎความสัมพันธ์ได้ถูกนำมาใช้เพื่อค้นหาความสัมพันธ์ระหว่างข้อมูลรายการหนึ่ง (Record) กับรายการอื่นๆ ในฐานข้อมูลที่เป็นแบบทรานแซกชั่น (Transaction) ซึ่งความสัมพันธ์ที่เกิดขึ้นไม่ได้เป็นความสัมพันธ์ที่ถาวรตามคุณสมบัติของฐานข้อมูล แต่เป็นความสัมพันธ์ที่เกิดขึ้นร่วมกันระหว่างข้อมูลรายการนั้นกับรายการอื่นๆ (Agrawal *et al.*, 1993) ด้วยคุณสมบัติลักษณะเช่นนี้ จึงได้นำไปประยุกต์ใช้ในการวิเคราะห์ข้อมูลในหลายๆ แอพพลิเคชั่น เช่น การวิเคราะห์พฤติกรรมการซื้อสินค้าของลูกค้าในห้างสรรพสินค้า การแนะนำรายการหนังสือที่น่าสนใจจากประวัติการเรียกดูรายการหนังสือของลูกค้า การวินิจฉัยโรคทางการแพทย์ เป็นต้น โดยตารางที่ 1 แสดงถึงตัวอย่างฐานข้อมูลที่เป็นแบบทรานแซกชั่น ซึ่งเป็นรูปแบบที่นิยมใช้อย่างแพร่หลายในการจัดเก็บข้อมูล

ตารางที่ 1 ตัวอย่างบัญชีรายการสินค้าในฐานข้อมูลแบบทรานแซกชั่น

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

ตารางที่ 1 (ต่อ)

ที่มา: Tan *et al*. (2006)

1.1. ไอเท็มเซต (Itemset) ถ้ากำหนดให้ $I = \{i_1, i_2, ...i_d\}$ เป็นเซตของไอเท็มทั้งหมด และกำหนดให้ $T = \{t_1, t_2, ...t_N\}$ เป็นทรานแซกชั่นทั้งหมด โดยแต่ละทรานแซกชั่น t_j จะประกอบด้วยสับเซตของไอเท็มที่ถูกเลือกมาจากไอเท็มเซต I ซึ่งในการวิเคราะห์ความสัมพันธ์ ไอเท็มอาจจะประกอบด้วยจำนวนตั้งแต่ 0 ไอเท็ม หรือประกอบด้วยหลายๆ ไอเท็มรวมกันอยู่ในรูปของไอเท็มเซต ถ้าไอเท็มเซตประกอบด้วยไอเท็มจำนวน k ไอเท็ม จะถูกเรียกว่า k-itemset ตัวอย่างเช่น {Beer, Diapers, Milk} ถือเป็น 3-itemset

1.2. การนับค่าสนับสนุน (Support Count) ถ้าทรานแซกชั่น t_j ประกอบด้วยไอเท็มเซต X ดังนั้น X จะเป็นสับเซทของ t_j จากตารางที่ 1 ทรานแซกชั่นที่ 2 ประกอบด้วยไอเท็มเซต {Bread, Diapers} แต่ไม่มีไอเท็มเซต {Bread, Milk} ดังนั้นคุณสมบัติที่สำคัญของไอเท็มเซตก็คือการนับค่าสนับสนุน (Agrawal *et al.*, 1993) เขียนอยู่ในรูปสมการที่ (1) ซึ่งจะเป็นจำนวนของทรานแซกชั่นที่มีไอเท็มเซตปรากฏอยู่ โดยจะแทนด้วยสัญลักษณ์ $\sigma(X)$

$$\sigma(X) = |\{t_j \mid X \subseteq t_j, t_j \in T\}| \tag{1}$$

1.3. กฎความสัมพันธ์ (Association Rule) จะแสดงอยู่ในรูป $X \rightarrow Y$ โดย X และ Y คือไอเท็มเซตที่ไม่ปรากฏร่วมกัน (Disjoint Itemsets) และโดยทั่วไปการวัดความแข็งแกร่งของกฎความสัมพันธ์จะอยู่ในรูปของค่าสนับสนุน (Support) สมการที่ (2) และ ค่าความเชื่อมั่น (Confidence) สมการที่ (3) ซึ่งค่าสนับสนุนจะเป็นการหาว่ากฎความสัมพันธ์นี้เกิดขึ้นในชุดข้อมูล (Data Set) เป็นความถี่เท่าไหร่ ขณะที่ค่าความเชื่อมั่นจะเป็นการหาว่าในทรานแซกชั่นที่ประกอบด้วยไอเท็มเซต X ทั้งหมดนั้น มีไอเท็มเซต Y เกิดขึ้นร่วมกันบนทรานแซกชั่นใดๆ เป็นความถี่เท่าไหร่

$$Support, s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N} \tag{2}$$

$$Confidence, c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \tag{3}$$

การวิเคราะห์ความสัมพันธ์จะใช้การสำรวจความสัมพันธ์ที่น่าสนใจซึ่งถูกซ่อนอยู่ในชุดข้อมูลขนาดใหญ่ โดยความสัมพันธ์สามารถนำมาแสดงอยู่ในรูปของกฎความสัมพันธ์ หรือไอเท็มเซตที่พบบ่อย ๆ (Frequent Itemset) จากตารางที่ 1 เราจะได้กฎความสัมพันธ์เช่น

$\{\text{Diapers}\} \rightarrow \{\text{Beer}\}$

จากตัวอย่างถ้าหากนับค่าสนับสนุนของกฎความสัมพันธ์ $\{\text{Diapers}\} \rightarrow \{\text{Beer}\}$ ถือว่าเป็นกฎที่มีความสัมพันธ์แข็งแกร่ง (Strong Relationship) ระหว่างการขาย ผ้าอ้อม {Diapers} และเบียร์ {Beer} เพราะว่ามีลูกค้าจำนวนมากซื้อผ้าอ้อมและซื้อเบียร์ด้วยพร้อมกัน ซึ่งในการวิเคราะห์ความสัมพันธ์จำเป็นต้องพิจารณาถึงการสำรวจรูปแบบ (Pattern) ที่มีการเกิดขึ้นซ้ำๆ ของทรานแซกชั่นในฐานข้อมูลขนาดใหญ่ โดยอาศัยการสร้างไอเท็มเซตทุกความเป็นไปได้ที่เกิดขึ้นบ่อยๆ (Frequent Itemset Generation) ซึ่งก็คือการสร้างไอเท็มเซตทุกๆ ความเป็นไปได้ของแต่ละไอเท็มที่เกิดขึ้นร่วมกัน แล้วทำการนับค่าสนับสนุนของแต่ละไอเท็มเซตที่มี และเลือกเอาเฉพาะไอเท็มเซตที่มีค่าสนับสนุนมากกว่าค่าสนับสนุนขั้นต่ำ (Minimum Support) ตามที่กำหนดไว้ ก็จะได้รูปแบบที่เกิดขึ้นซ้ำๆ หรือไอเท็มเซตที่เกิดขึ้นบ่อยมาใช้เป็นกฎความสัมพันธ์ ในการพิจารณาว่ากฎความสัมพันธ์นั้นมีศักยภาพเพียงพอหรือไม่ ก็จะอาศัยค่าความเชื่อมั่นมาเป็นเกณฑ์การพิจารณา ถ้าค่าความเชื่อมั่นมีค่ามากกว่าค่าความเชื่อมั่นขั้นต่ำ (Minimum Confidence) ก็จะพิจารณาว่าเป็นกฎที่มีความสัมพันธ์แข็งแกร่ง อัลกอริทึมหนึ่งที่นิยมใช้ในการสืบค้นกฎความสัมพันธ์ นั้นก็คือ อัลกอริทึม Apriori (Agrawal and Srikant, 1994)

1.4. อัลกอริทึม Apriori มีรายละเอียดดังนี้ ในการหารูปแบบของไอเท็มเซตที่เป็นไปได้ทั้งหมดจะเห็นว่ามีความเป็นไปได้จำนวนมากมายที่ไอเท็มเกิดขึ้นร่วมกัน เพื่อลดจำนวนของการหาทุกความเป็นไปได้ อัลกอริทึม Apriori จึงนำเสนอขั้นตอนการหาไอเท็มเซตที่เกิดขึ้นบ่อย โดยใช้ค่าสนับสนุนและค่าความเชื่อมั่นขั้นต่ำมาพิจารณาในการสร้างความเป็นไปได้ของไอเท็มเซตต่างๆ ซึ่งผลลัพธ์ที่ได้จากอัลกอริทึม Apriori ก็คือกฎที่มีความสัมพันธ์แข็งแกร่งเท่านั้น ส่วนกฎที่ไม่มีศักยภาพเพียงพอก็จะถูกกำจัดทิ้งไป (Pruned) โดยขั้นตอนการทำงานของอัลกอริทึม Apriori จะทำการสร้างไอเท็มเซตครั้งละ 1 ขนาด (k-itemset) เพิ่มขึ้นไปเรื่อยๆ ที่ละรอบการทำงาน ซึ่งในแต่ละรอบจะทำการนับจำนวนค่าสนับสนุนและค่าความเชื่อมั่น เพื่อพิจารณาว่าไอเท็มเซตตัวไหนมีค่าสนับสนุนและค่าความเชื่อมั่นมากกว่าค่าสนับสนุนและค่าความเชื่อมั่นขั้นต่ำ แล้วจึงนำไปสร้างเป็นไอเท็มเซตที่มีขนาดเพิ่มขึ้นในรอบถัดๆ ไป โดยตัวอย่างขั้นตอนการสร้างกฎความสัมพันธ์อ้างอิงข้อมูลทรานแซกชั่นจากตารางที่ 1 และกำหนดให้ค่าสนับสนุนขั้นต่ำคือ 3 ซึ่งสามารถดู

ขั้นตอนต่างๆ ได้จากตารางที่ 2 ถึง ตารางที่ 4 และกฎความสัมพันธ์ทั้งหมด สามารถดูได้จากตารางที่ 5

ตารางที่ 2 อัลกอริทึม Apriori สร้างไอเทมเซ็ตขนาด 1-itemsets

ไอเท็ม	ค่าสนับสนุน	ค่าสนับสนุน (%)	เลือก/ตัดทิ้ง
{Beer}	3	60	ถูกเลือก
{Bread}	4	80	ถูกเลือก
{Cola}	2	40	ตัดทิ้ง
{Diapers}	4	80	ถูกเลือก
{Milk}	4	80	ถูกเลือก
{Eggs}	1	20	ตัดทิ้ง

ตารางที่ 3 อัลกอริทึม Apriori สร้างไอเทมเซ็ตขนาด 2-itemsets

ไอเท็มเซต	ค่าสนับสนุน	ค่าสนับสนุน (%)	เลือก/ตัดทิ้ง
{Beer, Bread}	2	40	ตัดทิ้ง
{Beer, Diapers}	3	60	ถูกเลือก
{Bear, Milk}	2	40	ตัดทิ้ง
{Bread, Diapers}	3	60	ถูกเลือก
{Bread, Milk}	3	60	ถูกเลือก
{ Diapers, Milk}	3	60	ถูกเลือก

ตารางที่ 4 อัลกอริทึม Apriori สร้างไอเทมเซ็ตขนาด 3-itemsets

ไอเท็มเซต	ค่าสนับสนุน	ค่าสนับสนุน (%)	เลือก/ตัดทิ้ง
{Beer, Diapers, Milk}	3	60	ถูกเลือก

ตารางที่ 5 กฎความสัมพันธ์ทั้งหมดที่ถูกสร้างจากอัลกอริทึม Apriori

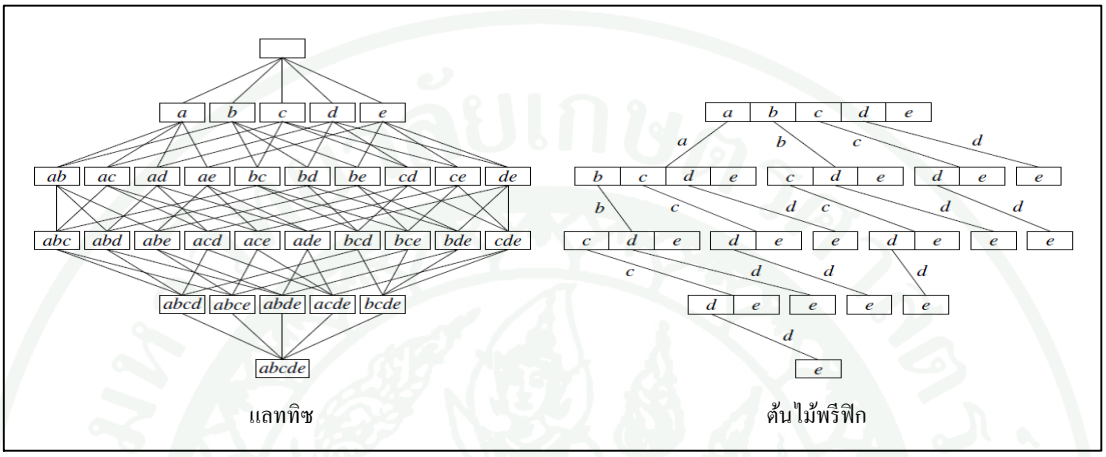
กฎความสัมพันธ์	ค่าสนับสนุน (%)	ค่าความเชื่อมั่น(%)
$\{Beer\} \rightarrow \{Diapers\}$	60	100
$\{Bread\} \rightarrow \{Diapers\}$	60	75
$\{Bread\} \rightarrow \{Milk\}$	60	75
$\{Diapers\} \rightarrow \{Milk\}$	60	75
$\{Beer, Diapers\} \rightarrow \{Milk\}$	60	100

จากตารางที่ 5 แสดงถึงกฎความสัมพันธ์ ค่าสนับสนุนของกฎและค่าความเชื่อมั่นของกฎทั้งหมดที่สร้างมาจากอัลกอริทึม Apriori ด้วยค่าสนับสนุนขั้นต่ำเท่ากับ 3 ซึ่งเห็นได้ว่า กฎความสัมพันธ์ทั้งหมดนั้นเป็นกฎที่แข็งแกร่ง เนื่องจากมีค่าความเชื่อมั่นที่สูง ตัวอย่างเช่นกฎความสัมพันธ์ $\{Beer, Diapers\} \rightarrow \{Milk\}$มีค่าสนับสนุน 60% ซึ่งแสดงให้เห็นว่าในฐานข้อมูลทั้งหมดมีการซื้อเบียร์, ผ้าอ้อม และนม พร้อมๆ กันมากถึง 60% จากข้อมูลการขายทั้งหมด ด้วยค่าความเชื่อมั่น 100% นั้นหมายถึงในทุกๆ ทรานแซกชั่นที่มีการซื้อเบียร์และผ้าอ้อมแล้วจะซื้อนมไปด้วยทุกครั้ง หรือคิดเป็น 100%

1.5. อัลกอริทึม Eclat ถูกพัฒนาและนำเสนอโดย Zaki *et al*. (1997) เพื่อปรับปรุงประสิทธิด้านความเร็วในการสืบค้นกฎความสัมพันธ์โดยอาศัยเทคนิคการอ่านข้อมูลเพียงครั้งเดียวสร้างเป็นครัซเตอร์ของไอเท็ม (Item Clustering) และการท่องไปในแลททิซ (Lattice) เพื่อสร้างไอเท็มเซตที่เกิดขึ้นบ่อย (Frequent Itemsets) ในแต่ละครัซเตอร์ของไอเท็ม โดย Eclat ได้ใช้โครงสร้างของต้นไม้พรีฟิก (Prefix Trees) แทนโครงสร้างแลททิซ แสดงไว้ในภาพที่ 1 โดยโครงสร้างของแลททิซใช้แสดงถึงรายการความเป็นไปได้ทั้งหมดของทุกไอเท็มเซต ซึ่ง Eclat อาศัยโครงสร้างต้นไม้พรีฟิก ในการสร้างไอเท็มเซตทุกความเป็นไปได้ โดยใช้อัลกอริทึมการแวะผ่านต้นไม้แบบลึกก่อน (Depth-First Traversal) ในการเข้าถึงโครงสร้างของต้นไม้พรีฟิก

1.5.1. การสร้างต้นไม้พรีฟิก ในแต่ละโหนด (Node) ของต้นไม้พรีฟิกจะมีโครงสร้างข้อมูลที่เรียกว่าบิทเวคเตอร์ (Bits Vector) ซึ่งแทนตำแหน่งของข้อมูลที่อยู่ในแต่ละทรานแซกชั่น เช่น ถ้ามีไอเท็ม *a* ที่ทรานแซกชั่นที่ 1 บิทที่ 1 ก็จะมีค่าเป็น 1 ถ้าไม่ปรากฏไอเท็ม *a* ที่ทรานแซกชั่นที่ 2 บิทที่ 2 ก็จะมีค่าเป็น 0 ดังนั้นแต่บิทจะแทนตำแหน่งของข้อมูลในทรานแซกชั่นทั้งหมดซึ่งทำให้มีขนาดที่เล็กกว่าฐานข้อมูลเดิม และ Eclat จะใช้บิทเวคเตอร์นี้แทนการเข้าถึงข้อมูล

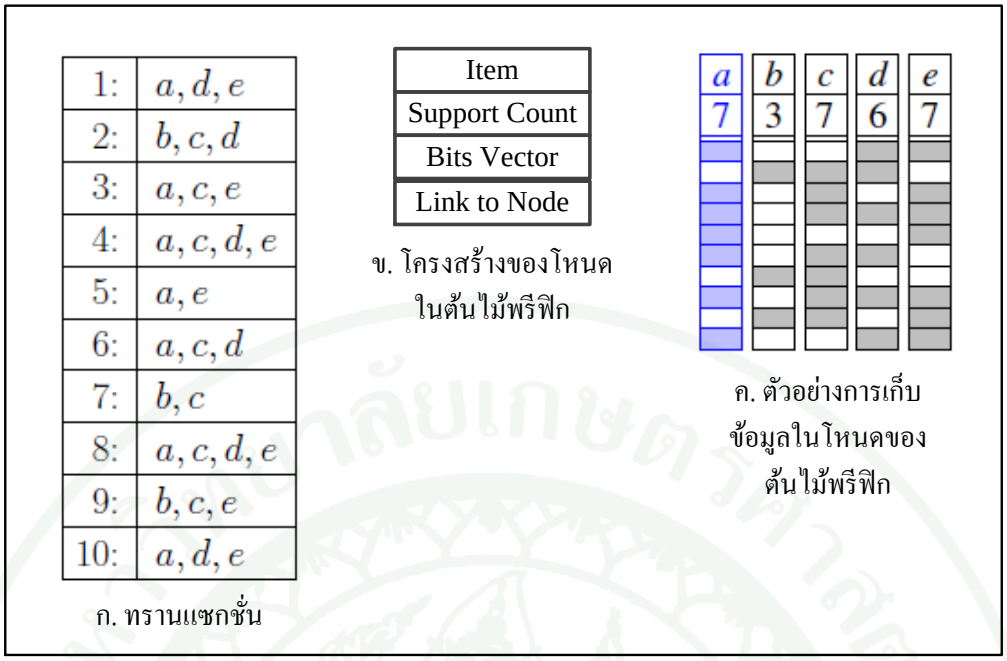
โดยตรง และในแต่ละโหนดของต้นไม้พรีฟิกจะมีการนับจำนวนค่าสนับสนุน (Support Count) ซึ่งเกิดจากการนับจำนวนบิทที่มีค่าเป็น 1 เท่านั้นในโหนดของมันเองเก็บเอาไว้ พิจารณาข้อมูลทรานแซกชั่นจากภาพที่ 2 ก. และโครงสร้างการเก็บข้อมูลในโหนดต่างๆ ของต้นไม้พรีฟิกในภาพที่ 2 ข. และ ค.



ภาพที่ 1 โครงสร้างเปรียบเทียบระหว่างไอเท็มเซตแลททิซและต้นไม้พรีฟิก

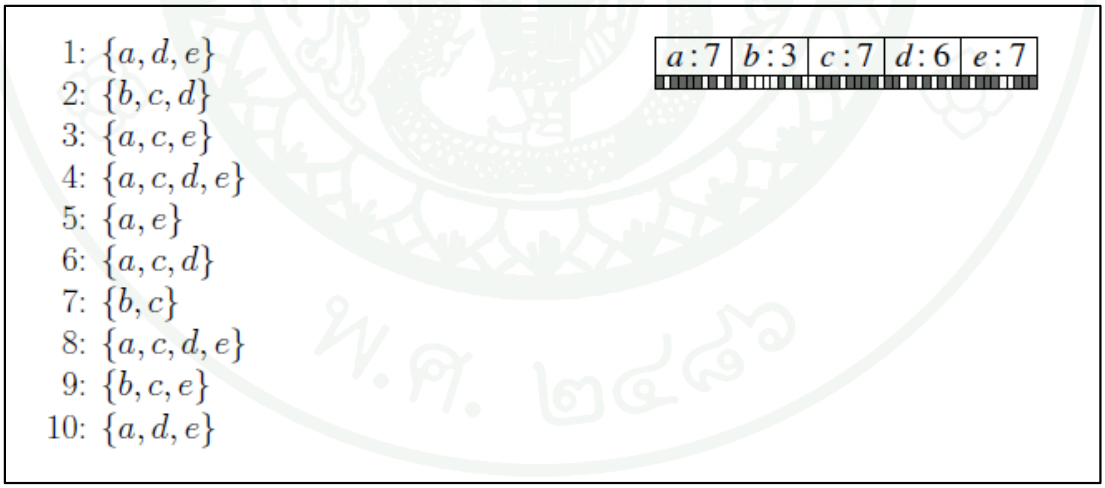
ที่มา: Borgelt (2010)

ในการอ่านข้อมูลรอบแรกจะได้ต้นไม้พรีฟิกของไอเท็มเซตที่มีขนาดเท่ากับ 1-itemset ซึ่งจะได้ข้อมูลของทุกทรานแซกชั่นเก็บเอาไว้ในบิทเวคเตอร์แต่ละโหนดของทุกแอตทริบิวต์ ซึ่งในรอบถัดๆ ไปไม่จำเป็นต้องอ่านข้อมูลจากฐานข้อมูลอีกแล้ว ใช้ข้อมูลจากบิทเวคเตอร์แทน พิจารณาจากภาพที่ 3 ด้านซ้ายคือข้อมูลทรานแซกชั่น ด้านขวาคือต้นไม้พรีฟิกโดยบรรทัดแรกแสดงถึง ชื่อของไอเท็ม:ค่าสนับสนุน บรรทัดที่สองเป็นช่องสี่เหลี่ยมเล็กๆ แสดงถึงบิทเวคเตอร์ โดยสีเทาแสดงถึงข้อมูลบิทที่เป็น 1 และสีขาว แสดงถึงบิทที่เป็น 0 ตามลำดับทรานแซกชั่น



ภาพที่ 2 โครงสร้างการจัดเก็บข้อมูลของแต่ละโหนดในต้นไม้พรีฟิก

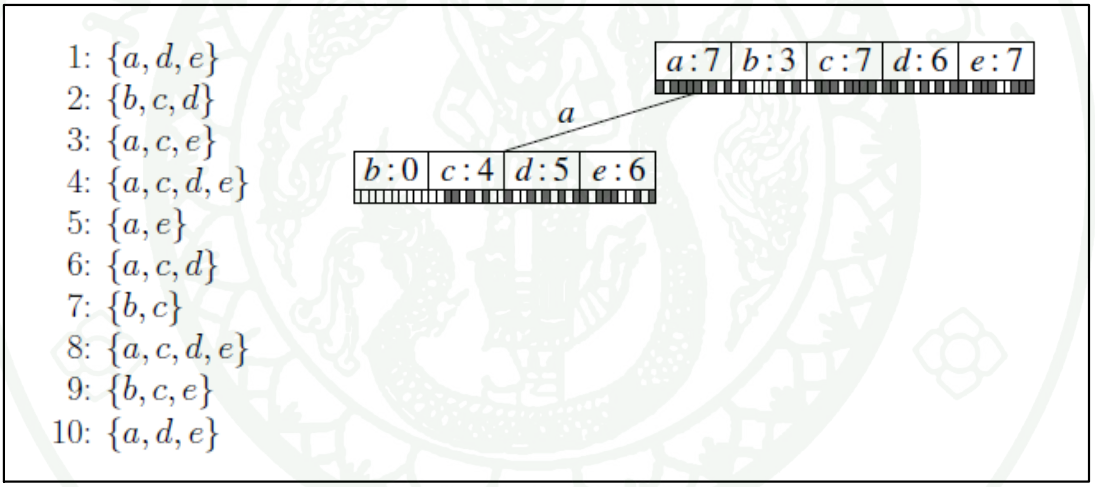
ที่มา: Borgelt (2010)



ภาพที่ 3 ขั้นตอนการสร้างต้นไม้พรีฟิกจากการอ่านข้อมูลทรานแซกชั่นรอบแรก

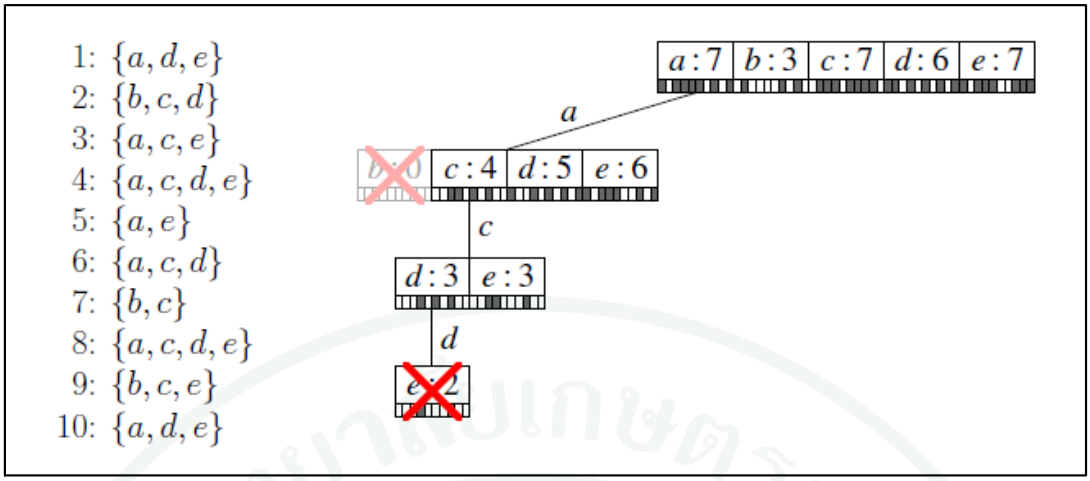
ที่มา: Borgelt (2010)

การสร้างต้นไม้พรีฟิกระดับที่ 2 ทำโดยการนำบิทเวคเตอร์ของต้นไม้พรีฟิกระดับที่ 1 มาอินเตอร์เซกกัน (Intersect) ก็คือการทำแอนด์บิทโอเปอร์เรเตอร์ (AND Bits) ทีละโหนดพร้อมทั้งนับค่าสนันสนุนจากบิทเวคเตอร์ ไล่จากโหนด $a \cap b$, $a \cap c$, $a \cap d$ และ $a \cap e$ แล้วนำไปสร้างเป็นต้นไม้พรีฟิกระดับที่ 2 ดังภาพที่ 4 เมื่อสร้างต้นไม้พรีฟิกระดับที่ 2 ของกลุ่มแรกเสร็จก็จะพิจารณาค่าสนับสนุนว่าผ่านเกณฑ์ค่าสนับสนุนขั้นต่ำหรือไม่ หากไม่ผ่านก็จะทำการตัดโหนดนั้นทิ้งไป ดังภาพที่ 4 แสดงให้เห็นว่าไอเท็มเซต $\{a, b\}$ มีค่าสนับสนุนเป็น 0 ซึ่งไม่ผ่านเกณฑ์ขั้นต่ำ (จากตัวอย่างกำหนดค่าสนับสนุนขั้นต่ำไว้ 20%) เมื่อสร้างต้นไม้พรีฟิกระดับที่ 2 ของกลุ่มแรกสมบูรณ์แล้วก็จะสร้างต้นไม้พรีฟิกระดับที่ 3, 4 ไปเรื่อยๆ ด้วยอัลกอริทึมการค้นหาแบบลึกก่อน (Depth-First Search) จนไม่สามารถสร้างต่อไปได้ดังภาพที่ 5 จากนั้นก็จะย้อนกลับ (Backtracks) ไปสร้างต้นไม้พรีฟิกในโหนดถัดไป ซึ่งทำเช่นนี้ไปจนกว่าครบทุกโหนดก็จะได้ต้นไม้พรีฟิกที่สมบูรณ์ดังภาพที่ 6



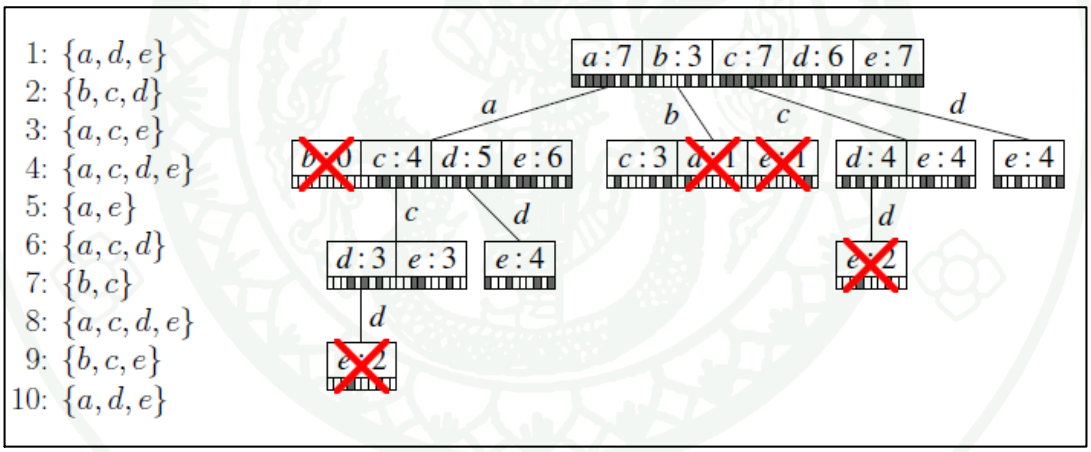
ภาพที่ 4 ขั้นตอนการสร้างต้นไม้พรีฟิกระดับที่ 2 ของกลุ่มแรก

ที่มา: Borgelt (2010)



ภาพที่ 5 ขั้นตอนการตัดโหนดที่ไม่ผ่านค่าสนับสนุนขั้นต่ำทิ้งไป

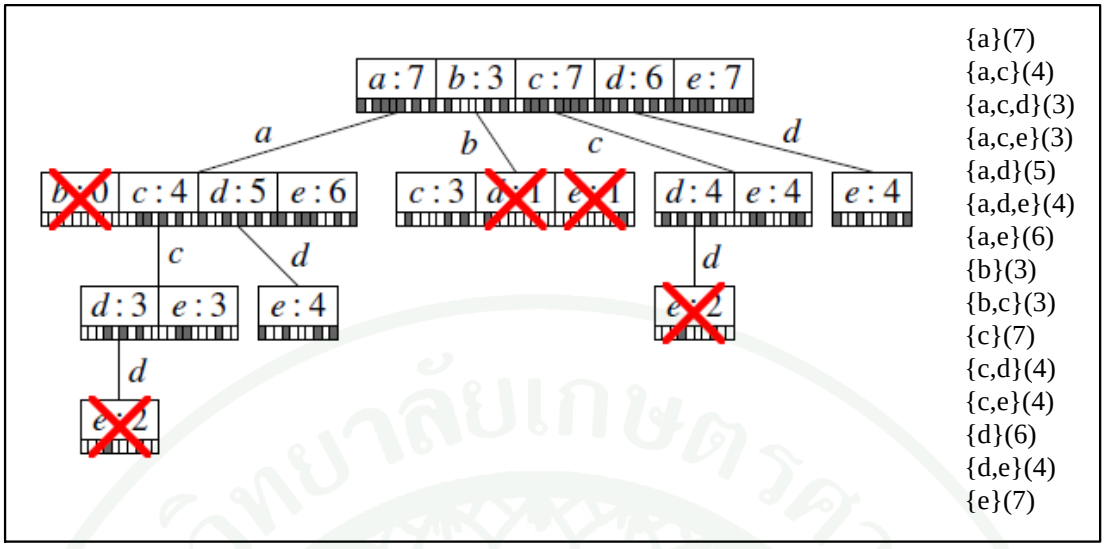
ที่มา: Borgelt (2010)



ภาพที่ 6 ต้นไม้พรีฟิกที่สมบูรณ์

ที่มา: Borgelt (2010)

1.5.2. การสร้างไอเท็มเซ็ตที่พบบ่อย (Generate Frequent Itemsets) ทำได้โดยการแวะผ่านไปในต้นไม้พรีฟิกแบบลึกก่อน (Depth-First Traversal) ขณะที่แวะผ่าน (Visit) โหนดไหนก็จะได้เซตของไอเท็มซึ่งนับรวมจากโหนดแรกของต้นไม้พรีฟิกระดับแรก ประกอบกันเป็นไอเท็มเซต และค่าสนับสนุนก็จะได้จากค่าสนับสนุนประจำโหนดนั้นๆ ภาพที่ 7 แสดงถึงไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำซึ่งถูกสร้างขึ้นจากการแวะผ่านไปในต้นไม้พรีฟิก



ภาพที่ 7 รายการไอเท็มเซตที่ปรากฏบ่อยซึ่งถูกสร้างจากต้นไม้พรีฟิก

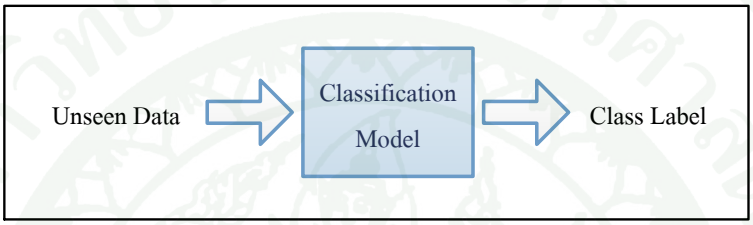
ที่มา: Borgelt (2010)

2. การจำแนกประเภทข้อมูล (Classification)

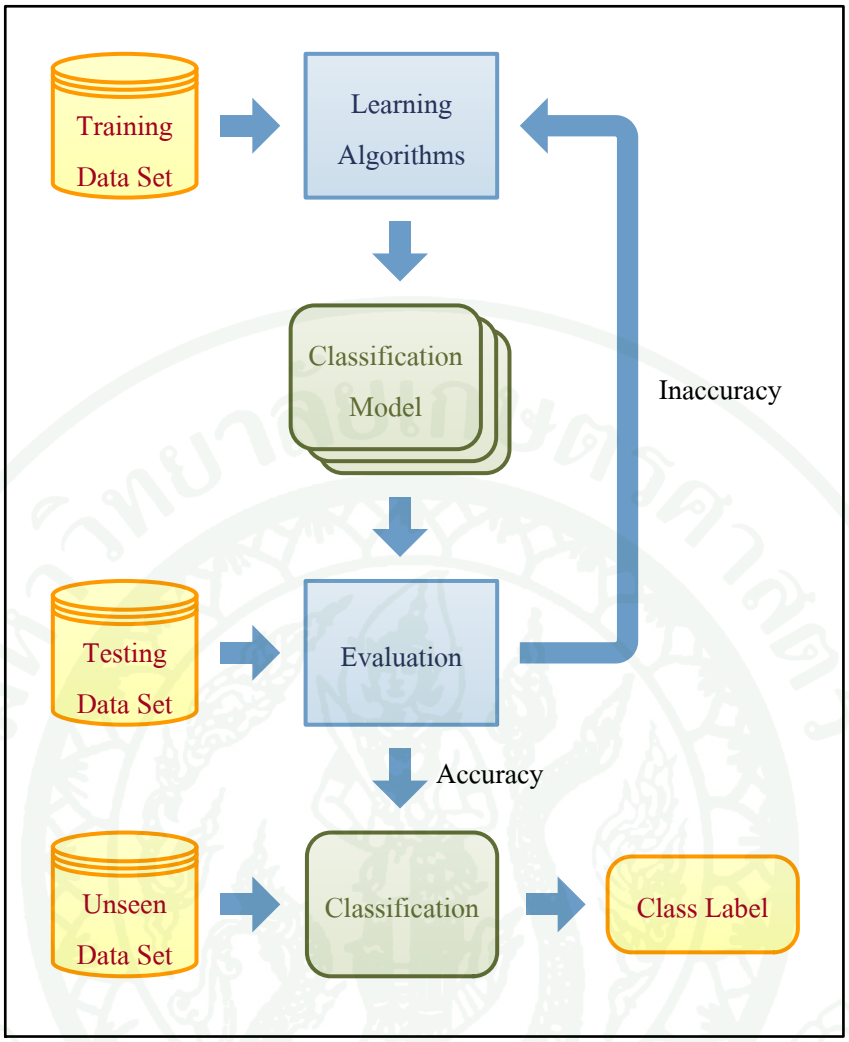
การจำแนกประเภทข้อมูลก็คือการระบุประเภทของข้อมูลจากหมวดหมู่ (Categories) ที่ถูกกำหนดไว้ ซึ่งนิยมใช้จำแนกประเภทบนข้อมูลที่เกิดขึ้นหรือเข้ามาใหม่ (Unseen Data) เทคนิคหนึ่งที่นิยมนำมาใช้จำแนกประเภทข้อมูลคือโมเดลการทำนาย (Predictive Modeling) ซึ่งถูกใช้ในการทำนายเพื่อระบุถึงประเภทหรือคลาส (Class Label) ของข้อมูลที่ยังไม่ทราบประเภท ภาพที่ 8 แสดงถึงลักษณะของโมเดลการทำนาย

เทคนิคการสร้างโมเดลการทำนายนั้นจะอาศัยเซตข้อมูล (Data Set) มาเป็นข้อมูลเบื้องต้นสำหรับใช้สร้างตัวจำแนกประเภทข้อมูล โดยข้อมูลที่นำมาใช้สร้างนี้จะเรียกว่าเซตข้อมูลสำหรับการเรียนรู้ (Training Data Set) เพื่อที่จะให้อัลกอริทึมสำหรับการเรียนรู้ (Learning Algorithms) ในการสร้างโมเดลจำแนกประเภทข้อมูลได้เรียนรู้ลักษณะรูปแบบของข้อมูล ดังนั้นอัลกอริทึมสำหรับการเรียนรู้จะสร้างเป็นโมเดลการทำนายที่มีความเหมาะสม (Fit) กับเซตข้อมูล และมีความถูกต้องสูงในการนำไปใช้จำแนกประเภทข้อมูลใหม่ ดังนั้นเพื่อที่จะตรวจสอบได้ว่าโมเดลการทำนายที่ได้มาจากการเรียนรู้นั้นมีความถูกต้องและมีคุณสมบัติที่ดีหรือแสดงถึงลักษณะโดยทั่วไปของข้อมูลได้ดี (Generalization Capability) จึงจำเป็นต้องมีเซตข้อมูลสำหรับนำมาทดสอบความถูกต้อง (Testing Data Set) โดยนำไปทดสอบโมเดลการทำนาย ซึ่งจะต้องสามารถระบุประเภทข้อมูลหรือ

คลาสของเซตข้อมูลสำหรับทดสอบได้ถูกต้องอย่างมีประสิทธิภาพ ซึ่งโดยทั่วไปแล้วจะแบ่งกลุ่มข้อมูลสำหรับการเรียนรู้และการทดสอบจากเซตข้อมูลออกเป็นสัดส่วน 70:30 โดยให้ข้อมูลสำหรับการเรียนรู้เป็น 70% และข้อมูลสำหรับการทดสอบเป็น 30% ถ้าหากโมเดลการทำนายที่สร้างขึ้นมีความถูกต้องในระดับที่ดีหรือยอมรับได้ ก็จะนำไปใช้ในการจำแนกประเภทข้อมูลใหม่ (Unseen Data) แต่ถ้าการทดสอบนั้นให้ความถูกต้องต่ำกว่าเกณฑ์ที่ยอมรับได้ ก็จะต้องปรับปรุงอัลกอริทึมสำหรับการเรียนรู้ให้มีประสิทธิภาพที่ดีขึ้นต่อไปจนกว่าจะได้ผลลัพธ์อยู่ในเกณฑ์ที่ยอมรับได้ โดยขั้นตอนการสร้างโมเดลการทำนายแสดงในภาพที่ 9



ภาพที่ 8 การจำแนกประเภทข้อมูลด้วยโมเดลการทำนาย



ภาพที่ 9 ขั้นตอนการสร้างโมเดลการทำนายสำหรับจำแนกประเภทข้อมูล

3. การจำแนกประเภทด้วยกฎความสัมพันธ์ (Associative Classification)

การจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์ (Liu *et al*., 1998) คือการรวมเทคนิคการสืบค้นกฎความสัมพันธ์เข้ากับการจำแนกประเภทข้อมูล เพื่อค้นหาเซตของกฎความสัมพันธ์ให้มีขนาดเล็กที่สุดจากฐานข้อมูลและนำมาสร้างเป็นโมเดลการจำแนกประเภทที่มีความถูกต้องแม่นยำมากที่สุด โดยกฎความสัมพันธ์จะแสดงถึงคุณลักษณะหรือรูปแบบความสัมพันธ์โดยทั่วไปของฐานข้อมูล ในการจำแนกประเภทด้วยกฎความสัมพันธ์นั้นได้อาศัยมาตรวัด 2 ชนิดด้วยกันมาเป็นเครื่องมือเพื่อเลือกกฎความสัมพันธ์ที่มีคุณลักษณะสมบัติที่ดีต่อฐานข้อมูลที่ใช้สำหรับการเรียนรู้ นั้นก็คือ มาตรวัดค่าสนับสนุน และมาตรวัดค่าความเชื่อมั่น โดยกฎความสัมพันธ์ทุกกฎที่ถูกเลือกจะต้องมีค่าสนับสนุนและค่าความเชื่อมั่นของกฎมากกว่าค่าสนับสนุนและค่าความเชื่อมั่นขั้นต่ำที่กำหนดไว้เป็นเกณฑ์การพิจารณา ซึ่งในการสร้างตัวจำแนกประเภทด้วยกฎความสัมพันธ์นั้นสามารถแบ่งออกเป็น 2 ส่วนหลักๆ ได้แก่ ส่วนที่ใช้ในการสร้างกฎความสัมพันธ์ (Rule Generator Phase) และส่วนที่นำกฎความสัมพันธ์ไปสร้างเป็นโมเดลการทำนาย (Classifier Builder Phase)

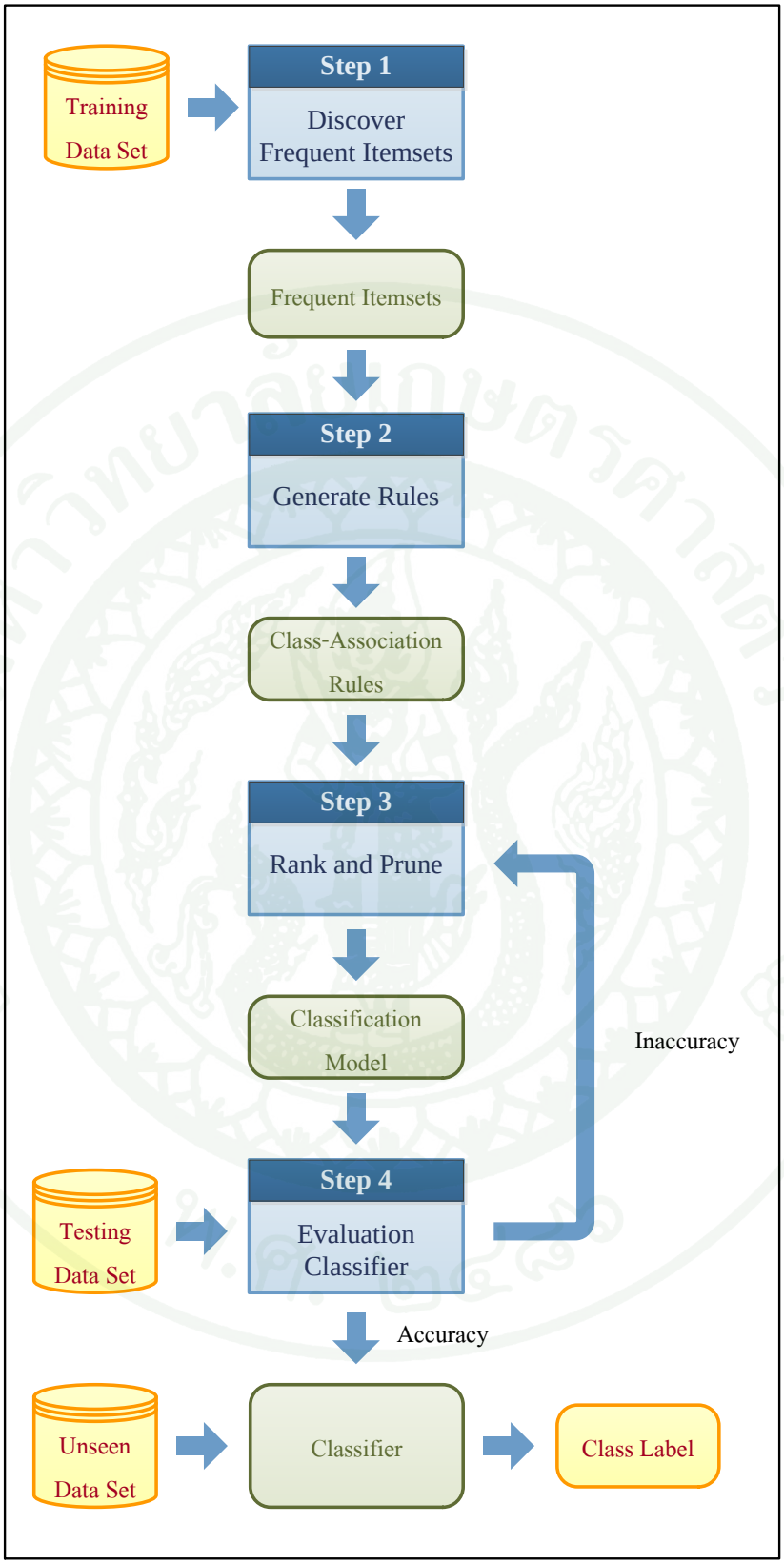
ขั้นตอนโดยทั่วไปของการสร้างโมเดลการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์จะประกอบไปด้วย 4 ขั้นตอนดังภาพที่ 10 ซึ่งมีขั้นตอนดังต่อไปนี้

3.1. ขั้นตอนการสืบค้นไอเท็มเซตที่พบบ่อย คือการหารูปแบบการเกิดขึ้นของไอเท็มเซตที่ถูกพบบ่อยๆ ซึ่งปกติแล้วอาศัยการสร้างทุกความเป็นไปได้ของไอเท็มแอตทริบิวต์ (Attributes) ที่มีอยู่ผ่านเทคนิคดังกล่าวข้างต้น เช่น Apriori และทำการนับค่าสนับสนุนของแต่ละไอเท็มเซตเพื่อใช้ในการพิจารณาว่าเกิดขึ้นบ่อยครั้งเพียงใด ปกตินิยมใช้มาตรวัดค่าสนับสนุนมาเป็นตัวบ่งชี้ว่าไอเท็มเซตต่างๆ เกิดขึ้นเป็นจำนวนเท่าไหร่ โดยกำหนดค่าสนับสนุนขั้นต่ำไว้เป็นเกณฑ์ หากไอเท็มเซตใดมีค่าสนับสนุนผ่านเกณฑ์ก็จะถูกพิจารณาเลือกเป็นไอเท็มเซตที่ถูกพบบ่อย (Frequent Itemsets)

3.2. ขั้นตอนสร้างกฎความสัมพันธ์ จะทำการเลือกเฉพาะกฎความสัมพันธ์ที่มีคุณสมบัติมีความสัมพันธ์กับคลาส หรือเรียกว่า กฎความสัมพันธ์แบบมีคลาสหรือ CARs (Class-Association Ruels) นั้นคือ กฎความสัมพันธ์ $X \rightarrow Y$ จะประกอบไปด้วย ไอเท็มเซต X เป็นแอตทริบิวต์อื่นๆ ซึ่งไม่มีแอตทริบิวต์ประเภทคลาสรวมอยู่ด้วย และมีความสัมพันธ์กับไอเท็มเซต Y ซึ่งเป็นแอตทริบิวต์ประเภทคลาสเท่านั้น

3.3. ขั้นตอนลำดับและกำจัดกฎความสัมพันธ์คือการพิจารณาเพื่อเลือกกฎความสัมพันธ์ที่มีคุณภาพและลำดับความสำคัญของกฎว่ากฎใดมีคุณภาพที่ดีกว่า ซึ่งในการพิจารณาคุณภาพของกฎความสัมพันธ์นิยมใช้มาตรวัดค่าความเชื่อมั่นมาเป็นตัวพิจารณาคุณภาพ โดยกำหนดเป็นค่าความเชื่อมั่นขั้นต่ำ หากกฎความสัมพันธ์ใดมีค่าความเชื่อมั่นผ่านเกณฑ์ความเชื่อมั่นขั้นต่ำนี้ก็ถือเป็นกฎที่มีคุณภาพซึ่งจะเก็บเอาไปใช้ในการสร้างโมเดลการทำนาย ส่วนกฎความสัมพันธ์ใดที่ไม่ผ่านเกณฑ์ความเชื่อมั่นขั้นต่ำก็จะถูกกำจัดทิ้งไป หลังจากที่ได้กฎความสัมพันธ์ที่มีคุณภาพมาแล้วก็จะทำการลำดับความสำคัญของกฎ โดยทั่วไปแล้วจะใช้ค่าความเชื่อมั่นของกฎความสัมพันธ์มาเรียงลำดับจากมากไปหาน้อย แต่ถ้าค่าความเชื่อมั่นของกฎความสัมพันธ์มีค่าเท่ากันก็จะใช้ค่าสนับสนุนของกฎความสัมพันธ์มาพิจารณาร่วมด้วยในการจัดลำดับ โดยเรียงลำดับจากค่าสนับสนุนมากไปหาน้อย สำหรับกฎที่มีค่าความเชื่อมั่นที่เท่ากัน แต่ถ้าค่าความเชื่อมั่นและค่าสนับสนุนของกฎความสัมพันธ์มีค่าเท่ากัน ก็จะพิจารณาจากลำดับของกฎที่ถูกสร้างขึ้น กฎความสัมพันธ์ตัวไหนถูกสร้างก่อน ก็จะเรียงเป็นลำดับแรก และกฎที่ถูกสร้างขึ้นครั้งหลังจะถูกเรียงเป็นลำดับถัดไป ตามลำดับที่เกิดขึ้นของกฎ

3.4. ขั้นตอนการทำนายจะอาศัยกฎความสัมพันธ์ที่ได้จากขั้นตอนก่อนหน้ามาสร้างเป็นโมเดลในการทำนายข้อมูล โดยในการทำนายข้อมูลนั้นคือการเลือกกฎที่เหมาะสมกับข้อมูลใหม่มาใช้ระบุคลาสของข้อมูลใหม่ ซึ่งสามารถแบ่งการพิจารณาเพื่อเลือกกฎในการทำนายได้เป็น 2 วิธี โดยวิธีที่ 1 คือการพิจารณาความสัมพันธ์ทีละกฎ (Single Rule) วิธีการพิจารณาแบบนี้จำเป็นต้องเรียงลำดับกฎความสัมพันธ์ก่อนจากขั้นตอนก่อนหน้านี้ หลังจากที่ได้ลำดับความสำคัญของกฎความสัมพันธ์เรียบร้อยแล้วก็จะได้เป็นโมเดลของกฎความสัมพันธ์ที่พร้อมนำไปทำนายข้อมูล โดยการทำนายข้อมูลนั้นจะทำนายตามคลาสของกฎที่มีศักย์สูงสุด (Precedence) วิธีที่ 2 คือการพิจารณากฎความสัมพันธ์หลายกฎร่วมกัน (Multiple Rules) ซึ่งจะอาศัยกฎความสัมพันธ์ที่อยู่ในกลุ่มของคลาสเดียวกัน นำมาคำนวณผ่านสูตรหรือวิธีการอื่นๆ ที่กำหนดไว้ว่า คำตอบของคลาสไหนให้ค่าคำตอบสูงสุด ก็จะพิจารณาตอบเป็นคลาสนั้น



ภาพที่ 10 ขั้นตอนการสร้างตัวจำแนกประเภทด้วยกฎความสัมพันธ์

ปัญหาความไม่สมดุลของคลาส

เซตข้อมูลโดยปกติจะประกอบไปด้วยคลาสต่างๆ มากมายซึ่งทำให้กลุ่มข้อมูลบางคลาสอาจจะเกิดความไม่สมดุลได้ อันเนื่องมาจากกลุ่มข้อมูลในแต่ละคลาสมีสัดส่วนที่ไม่เท่ากันหรือมีการแพร่กระจายตัวแบบไม่สมดุล หรือโน้มเอียงไปทางคลาสใดคลาสหนึ่งมากกว่า ซึ่งสามารถพบได้ในหลายๆ แอพพลิเคชั่น เช่น การจำแนกข้อมูลตัวอย่างของผู้ป่วยที่เป็นโรคมะเร็งออกจากข้อมูลตัวอย่างของคนสุขภาพปกติ จะเห็นว่าข้อมูลตัวอย่างของผู้ป่วยโรคมะเร็งมีปริมาณจำนวนน้อยมาก เพื่อให้ตัวจำแนกประเภทใช้ในการเรียนรู้เพื่อให้ได้คุณลักษณะที่เหมาะสมกับข้อมูลกลุ่มนี้เมื่อเทียบกับข้อมูลตัวอย่างของคนที่มีสุขภาพปกติ ทำให้ยากต่อการสร้างโมเดลการจำแนก ดังนั้นปัญหาของความไม่สมดุลของคลาสจะเกิดขึ้นเมื่อคลาสที่เราให้ความสนใจเป็นคลาสของกลุ่มข้อมูลที่มีอยู่ค่อนข้างน้อย ซึ่งปัจจุบันยังคงเป็นปัญหาหนึ่งที่เป็นเรื่องท้าทายในการทำดาต้าไมนิ่ง ปัญหาที่เกิดขึ้นจะเห็นว่าข้อมูลตัวอย่างผู้ป่วยโรคมะเร็งมีความถี่ในการเกิดขึ้นไม่บ่อยนัก ดังนั้นหากการจำแนกประเภททำการจำแนกคลาสของข้อมูลกลุ่มน้อยหรือคลาสรอง (Minority Class) ผิดพลาดนั้นมีต้นทุนการจำแนกผิดพลาดสูงกว่า (Cost of Misclassification) เมื่อเทียบกับการจำแนกคลาสข้อมูลกลุ่มใหญ่หรือคลาสหลัก (Majority Class) ผิดพลาด ตัวอย่างเช่น ถ้าโมเดลการจำแนกทำการระบุข้อมูลตัวอย่างของผู้ป่วยโรคมะเร็งว่าไม่ได้เป็นโรคมะเร็งเป็นคนสุขภาพปกติดีจะมีผลเสียมากกว่าการระบุว่าข้อมูลตัวอย่างคนสุขภาพปกติเป็นโรคมะเร็งเพราะว่าถ้าผู้ป่วยเป็นโรคมะเร็งแล้วระบุว่าเป็นคนสุขภาพปกติ อาจจะทำให้เขาไม่ได้รับการรักษาและนำไปสู่การเสียชีวิตจากโรคมะเร็งได้ แต่ในทางตรงกันข้ามถ้าคนสุขภาพปกติถูกระบุว่าเป็นโรคมะเร็งแล้ว ถึงแม้ว่าจะมีการรักษาเกิดขึ้นแต่ก็ไม่ส่งผลถึงชีวิตหรือการรักษานั้นไม่มีผลอะไรเกิดขึ้น ซึ่งกรณีตัวอย่างนี้จะเห็นว่ากลุ่มข้อมูลตัวอย่างของผู้ป่วยโรคมะเร็งและกลุ่มข้อมูลตัวอย่างของคนสุขภาพปกติมีลักษณะการแพร่กระจายตัวเป็นแบบไม่สมดุล (Class Imbalanced) และกลุ่มข้อมูลที่เราสนใจคือกลุ่มผู้ป่วยโรคมะเร็งหรือคลาสรอง ซึ่งปัจจุบันการแพร่กระจายของกลุ่มข้อมูลแบบไม่สมดุลก็ยังคงเป็นปัญหาใหญ่ในอัลกอริทึมของโมเดลการจำแนกประเภท

1. ปัญหาจากความเบาบาง (Rarity)

ในการศึกษาเรื่องความไม่สมดุลของกลุ่มข้อมูลในงานดาต้าไมนิ่ง สิ่งหนึ่งที่มีความสำคัญมากก็คือเรื่องความเบาบางของข้อมูล (Rarity) ซึ่งจะทำให้การระบุคลาสของข้อมูลทำได้ยาก เพราะว่าการระบุคลาสของข้อมูลที่มีอยู่น้อย (Rare Objects) นั้นทำได้ยากกว่าการระบุคลาสของข้อมูลที่เป็นข้อมูลกลุ่มหลัก (Common Objects) โดยก่อนหน้านี้ Weiss (2004) ได้ทำการศึกษาวิจัย

เพื่อสำรวจปัญหาความเบาบางของข้อมูลในการทำดาต้าไมนิ่งบนข้อมูลที่ไม่สมดุล ได้ทำการแบ่งประเภทของความเบาบางของข้อมูลออกเป็นสองประเภทดังต่อไปนี้

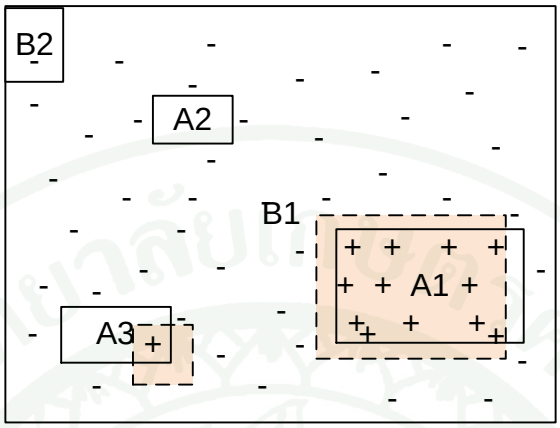
1.1. ความเบาบางของคลาส (Rare Classes) เป็นปัญหาหลักของการเกิดความไม่สมดุลของคลาส (Class Imbalance) ซึ่งความเบาบางประเภทนี้เกิดจากความต้องการข้อมูลตัวอย่างที่ระบุคลาสแล้ว (Labeled Examples) และมีความสัมพันธ์สอดคล้องกับปัญหาที่ต้องการจำแนกประเภทมาช่วยในการเรียนรู้ แต่พบว่ามีข้อมูลที่ไม่เพียงพอ

1.2. ความเบาบางของความสัมพันธ์ (Rare Cases) ถึงแม้ว่าเราจะทราบคลาสของกลุ่มข้อมูลจากการระบุชื่อคลาส (Labeled Data) แล้วก็ตาม แต่ก็ยังคงมีความสัมพันธ์ของข้อมูลที่เกิดขึ้นเป็นสับเซตเล็กๆบนข้อมูลแต่ละตัวซึ่งมีขอบเขตข้อมูลค่อนข้างเล็กจากข้อมูลที่มีอยู่ทั้งหมด (Instance Space) โดยกรณีความเบาบางของความสัมพันธ์ที่เกิดขึ้นนี้ ขึ้นอยู่กับการแพร่กระจายตัวของข้อมูลทั้งบนข้อมูลที่ระบุคลาส และข้อมูลที่ไม่ทราบคลาส (Unlabeled Data) ซึ่งในข้อมูลตัวอย่างแต่ละตัวอาจจะประกอบไปด้วยกลุ่มข้อมูลย่อยๆ (Subconcept) หรือมีคลาสย่อย (Subclass) รวมกัน ดังนั้นหากเรามีกลุ่มตัวอย่างน้อยแล้ว ก็จะทำให้การหาความสัมพันธ์ของกลุ่มข้อมูลย่อยๆ เหล่านี้เป็นไปได้ยาก ตัวอย่างเช่น ในการวินิจฉัยโรคทางการแพทย์เพื่อตรวจสอบว่าเป็นโรคมะเร็งหรือไม่ อาจจะมีการใช้หลากหลายวิธีร่วมกันในการวินิจฉัยโรค ซึ่งแต่ละวิธีเหล่านี้ก็คือกลุ่มข้อมูลย่อย (Subconcept) ที่ปรากฏอยู่บนข้อมูล ตัวอย่างเช่นการใช้ผลการตรวจเลือดทางห้องปฏิบัติการมาเป็นข้อมูลหนึ่งเพื่อบอกว่าเป็นโรคมะเร็งหรือไม่ ซึ่งจะเห็นว่ากรณีเช่นนี้ความสัมพันธ์มักจะเกิดขึ้นเบาบางตามข้อมูลที่มีอยู่น้อยไปด้วย จึงเป็นเรื่องที่ยากที่จะหาความสัมพันธ์ที่แท้จริงซึ่งถูกซ่อนอยู่

2. ปัญหาจากกลุ่มข้อมูลที่แยกจากกัน (Small Disjunct)

กลุ่มของข้อมูลที่แยกจากกันนั้นเกิดขึ้นจากโมเดลการจำแนกประเภทมีข้อมูลที่คลอบคลุมตัวอย่างสำหรับเรียนรู้น้อยเกินไปทำให้โมเดลการเรียนรู้แบ่งกลุ่มข้อมูลออกเป็นกลุ่มย่อยๆ พิจารณาจากภาพที่ 11 ซึ่งประกอบด้วยข้อมูล 2 ประเภทคือคลาส A และคลาส B โดยกำหนดให้คลาส A คือประเภทคลาสที่เบาบาง (Rare class หรือ Minority class) และคลาส B คือประเภทคลาสส่วนใหญ่หรือคลาสหลัก (Common class หรือ Majority class) และกำหนดให้ประเภทคลาสที่เบาบางเป็นคลาสบวก (Positive class) และประเภทคลาสส่วนใหญ่เป็นคลาสลบ (Negative class) โดยมีขอบเขตการตัดสินใจ (Decision Boundaries) ที่เป็นขอบเขตจริงซึ่งถูกแสดงด้วยเส้นทึบ ขณะที่ขอบเขตการตัดสินใจที่ได้จากโมเดลการเรียนรู้ (Learned Boundaries) จะถูกแสดงด้วยเส้นประ โดย

การระบุชื่อของข้อมูลตัวอย่างจะถูกแสดงด้วยเครื่องหมาย + ซึ่งแสดงถึงตัวอย่างข้อมูลที่เป็นคลาสบวก และเครื่องหมาย – แสดงถึงตัวอย่างข้อมูลที่เป็นคลาสลบ



ภาพที่ 11 ข้อมูลตัวอย่างมีทั้งกรณี Rare Classes และ Rare Cases

จากภาพที่ 11 คลาสรองจะประกอบด้วย 3 กลุ่มข้อมูลย่อย (Subconcepts) ซึ่งมีความสัมพันธ์กับคลาส A (Rare Classes) และถูกระบุชื่อเป็น A1-A3 โดย กลุ่มข้อมูลย่อย A2 และ A3 ก็คือ Rare Case โดยที่ A1 คือกรณีที่เกิดขึ้นส่วนใหญ่ของคลาส A และคลาสหลัก (Majority classes) มีอยู่ด้วยกัน 2 กลุ่มข้อมูลย่อย (Subconcepts) ซึ่งมีความสัมพันธ์กับคลาส B โดยถูกระบุชื่อเป็น B1 และ B2 โดยกลุ่มข้อมูลย่อย B1 เป็นกรณีทั่วไปส่วนใหญ่ที่ครอบคลุมอยู่ในข้อมูลที่มีทั้งหมด ส่วนกลุ่มข้อมูลย่อย B2 คือ Rare Case ซึ่งจะเห็นว่า ในกรณีที่เป็นกลุ่มข้อมูลคลาสหลัก ก็อาจจะมี Rare Case ได้เช่นกัน จากภาพที่ 11 จะเห็นว่า A3 คือกลุ่มข้อมูลที่ถูกแยกออกจากกันขนาดเล็กๆ (Small Disjunct) อันเนื่องมาจากการเรียนรู้ของโมเดลการจำแนก ส่วนกรณี A2 โมเดลการจำแนกไม่สามารถหาได้จากขั้นตอนการเรียนรู้ ดังนั้นจะเห็นได้ว่า Rare Cases จะมีความคล้ายกับ Rare Classes ซึ่งเป็นผลที่เกิดมาจากข้อมูลที่ไม่สมดุล

3. ปัญหาจากการไม่นิ่งข้อมูลที่เบาบาง

จากปัญหาความเบาบางของข้อมูล Weiss (2004) ได้รวบรวมปัญหาที่เกิดจากการไม่นิ่งข้อมูลที่เกิดจากกรณีความเบาบางของคลาสและความเบาบางของความสัมพันธ์ของข้อมูล ซึ่งได้สรุปปัญหาออกเป็นหมวดหมู่ดังนี้

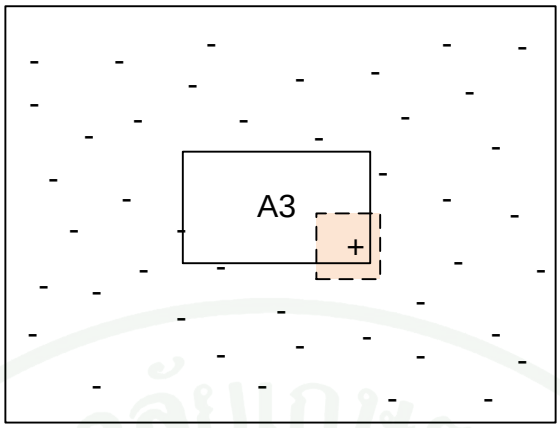
3.1. การเลือกมาตรวัดที่ไม่เหมาะสม (Improper Evaluation Metrics) โดยปกติมาตรวัดได้ถูกนำมาใช้เป็นแนวทางเพื่อหาผลลัพธ์จากอัลกอริทึมสำหรับไมนิ่งข้อมูล ดังนั้นถ้ามาตรวัดไม่มีความเหมาะสมเพียงพอที่จะให้ค่าผลลัพธ์ที่ดี ในการไมนิ่งข้อมูลที่เกิดขึ้นอย่างเบาบาง เมื่อทำการไมนิ่งข้อมูลก็เหมือนกับไม่ได้จัดการกับปัญหาของประเภทคลาสที่เบาบางและความสัมพันธ์ของข้อมูลที่เบาบาง ดังนั้นความแม่นยำของการจำแนกประเภทจะถูกคำนวณมาจากข้อมูลตัวอย่างที่สามารถจำแนกได้ถูกต้อง ซึ่งการใช้มาตรวัดในการหาค่าความแม่นยำนั้น รู้กันดีว่าในกลุ่มประเภทข้อมูลที่เบาบางจะมีผลกระทบต่อความแม่นยำค่อนข้างน้อย เมื่อเทียบกลับกลุ่มข้อมูลส่วนใหญ่

การไมนิ่งข้อมูลโดยอาศัยกฎความสัมพันธ์ (Association Rule Mining Systems) โดยทั่วไปจะอาศัยค่าสนับสนุนและค่าความเชื่อมั่นเป็นมาตรวัดหลัก และนำไปใช้ในการสืบค้นกฎความสัมพันธ์ โดยปกติแล้วระบบการไมนิ่งข้อมูลโดยอาศัยกฎความสัมพันธ์จะเลือกกฎที่มีค่าสนับสนุนมากกว่าค่าสนับสนุนขั้นต่ำ ที่กำหนดไว้ เพื่อประสิทธิภาพที่ดี จึงทำให้ค่าสนับสนุนขั้นต่ำนั้นไม่สามารถกำหนดได้ต่ำมากเพียงพอต่อในการสืบหาหาความสัมพันธ์ที่หาเกิดขึ้นอย่างเบาบาง (Rare Associations) ได้ จึงจำเป็นต้องเลือกมาตรวัดที่มีความเหมาะสม และสามารถให้ค่าคำตอบที่ดีต่อข้อมูลที่มีความเบาบาง

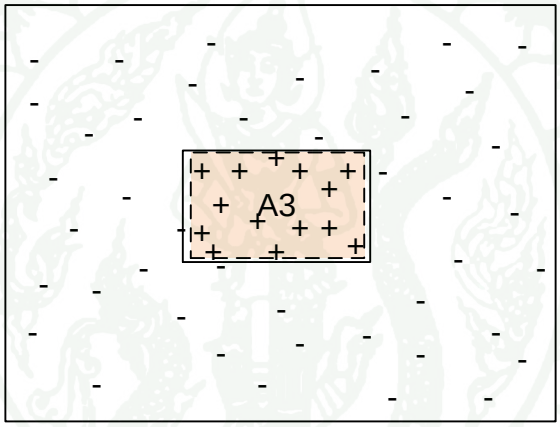
ในงานวิจัยที่ผ่านมามีการแก้ไขปัญหาความไม่สมดุลของข้อมูล โดยเลือกมาตรวัดที่เหมาะสมกับปัญหาข้อมูลที่ไม่สมดุล รวมทั้งนำเสนอมาตรวัดใหม่ๆ เพื่อเพิ่มประสิทธิภาพความแม่นยำให้กับการจำแนกประเภท

3.2. ความขาดแคลนข้อมูล (Lack of Data) เป็นปัญหาหลักที่ทำให้เกิดความเบาบางของคลาสและความเบาบางของความสัมพันธ์โดยตรง นั้นคือข้อมูลตัวอย่างมีความสัมพันธ์กับประเภทคลาสที่เบาบาง และความสัมพันธ์ของข้อมูลที่เบาบางนั้นมีขนาดเล็กหรือมีจำนวนน้อยเกินไป ในสถานการณ์เช่นนี้จะทำให้เกิดความยากลำบากในการหาขอบเขตการตัดสินใจของประเภทคลาสที่เบาบางและความสัมพันธ์ของข้อมูลที่เบาบาง

พิจารณาภาพที่ 12 จะเห็นว่าเมื่อตัวอย่างข้อมูลของประเภทคลาสที่เบาบางและความสัมพันธ์ของข้อมูลที่เบาบางของคลาส A3 มีจำนวนน้อยเกินไป ก็จะทำให้การไมนิ่งข้อมูลเพื่อหาขอบเขตการตัดสินใจ ทำได้ไม่ถูกต้องหรือผิดเพี้ยนไปจากความเป็นจริง



ภาพที่ 12 แสดงผลกระทบของการขาดแคลนข้อมูล



ภาพที่ 13 แสดงผลกระทบที่เกิดจากข้อมูลที่เพิ่มขึ้น

จากภาพที่ 13 จะเห็นว่าเมื่อเพิ่มข้อมูลตัวอย่างของประเภทคลาสที่เบาบางและความสัมพันธ์ของข้อมูลที่เบาบางของคลาส A3 เข้าไปจนมากเพียงพอ ก็จะทำให้การไมนิ่งข้อมูลเพื่อหาขอบเขตการตัดสินใจ มีความแม่นยำมากขึ้นและมีความใกล้เคียงกับค่าความเป็นจริง

ดังนั้นจึงพบว่าจำนวนข้อมูลตัวอย่างที่ป้อนให้กับโมเดลการเรียนรู้ เพื่อทำการจำแนกประเภทจะมีความสัมพันธ์กับขอบเขตการตัดสินใจโดยตรง และขอบเขตการตัดสินใจที่ได้มาจากโมเดลการเรียนรู้จะมีค่าประมาณใกล้เคียงกับขอบเขตการตัดสินใจจริงๆ ก็ต่อเมื่อมีข้อมูลตัวอย่างมากเพียงพอ

ในหลายงานวิจัยที่ผ่านมาได้มีการแก้ไขปัญหานี้ โดยใช้เทคนิคการเพิ่มข้อมูลตัวอย่างเข้าไปในประเภทคลาสที่เบาบางและความสัมพันธ์ของข้อมูลที่เบาบาง (Over-Sampling) เพื่อเพิ่มประสิทธิภาพให้กับโมเดลการเรียนรู้ ในการหาขอบเขตการตัดสินใจให้เข้าใกล้ขอบเขตการตัดสินใจจริงๆ

3.3. ความขาดแคลนความสัมพันธ์ของข้อมูล (Relative Lack of Data) เป็นปัญหาหนึ่งของความเบาบางเช่นกัน ซึ่งการสืบค้นความสัมพันธ์กับข้อมูลอื่นๆ เป็นเรื่องที่ทำได้ยากบนฐานข้อมูลที่ไม่สมดุล ยกตัวอย่างเช่น ข้อมูลการซื้อขายสินค้าในห้างสรรพสินค้าพบว่า จำนวนการซื้อเครื่องบดอาหารหรือกระทะปรุงอาหารเกิดขึ้นไม่บ่อยนัก ดังนั้นในการสืบค้นความสัมพันธ์ของเครื่องบดอาหารและกระทะปรุงอาหารที่ถูกซื้อร่วมกัน ยิ่งทำได้ยากมากหรืออาจจะไม่สามารถสืบค้นความสัมพันธ์ได้เลย เพราะทั้งสองเป็นสินค้าที่มีการซื้อน้อยในซุปเปอร์มาร์เก็ตและโอกาสที่สินค้าทั้งคู่จะถูกซื้อร่วมกันเมื่อมีการซื้อสินค้าอย่างใดอย่างหนึ่งนั้นมีน้อยมาก

ในหลายงานวิจัยที่ผ่านมาได้แก้ไขปัญหานี้ด้วยหลากหลายวิธี เช่น การเรียนรู้เฉพาะประเภทข้อมูลที่เบาบางเพียงอย่างเดียว (Learn only the rare class), การแบ่งข้อมูลออกเป็นส่วน ๆ (Segmenting the data), การเพิ่มและลดข้อมูลตัวอย่าง (Sampling over and under), หลีกเลี่ยงเทคนิคการค้นหาแบบกรีดี้ (Non greedy search techniques), การเรียนรู้แบบมีต้นทุน (Cost-sensitive learning), การเพิ่มกำลัง (Boosting) ซึ่งข้อมูลที่เบาบางนี้อาจจะอยู่บนความเชื่อมโยงระหว่างหลากหลายเงื่อนไข ดังนั้นการพิจารณาข้อมูลเพียงทีละเงื่อนไขอาจจะมีข้อมูลที่ไม่เพียงพอ

3.4. ข้อมูลที่กระจัดกระจาย (Data Fragmentation) หลายๆ อัลกอริทึมในการไมนิ่งข้อมูลอาศัยเทคนิคการแบ่งและเอาชนะ (Divide-and-conquer) ซึ่งก็คือเป็นการแบ่งย่อยปัญหาให้เล็กลงๆ และด้วยกระบวนการวิธีนี้อาจนำไปสู่ปัญหาการกระจัดกระจายของข้อมูล ส่งผลให้เกิดปัญหาของการขาดแคลนข้อมูลตามมา

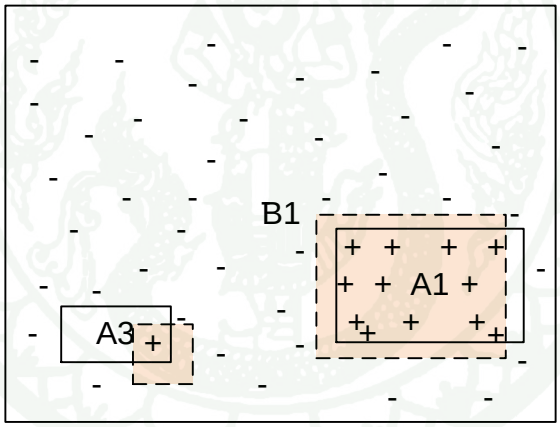
ในหลายงานวิจัยที่ผ่านมาได้แก้ไขปัญหานี้หลากหลายวิธี เช่น การเรียนรู้เฉพาะกลุ่มข้อมูลที่เบาบางเพียงอย่างเดียว (Learn only the rare class), การเพิ่มและลดข้อมูลตัวอย่าง (Sampling over and under), หลีกเลี่ยงเทคนิคค้นหาแบบกรีดี้ (Non greedy search techniques) เป็นต้น

3.5. การปรับลดที่ไม่เหมาะสม (Inappropriate Inductive Bias) มีผลกระทบต่อประสิทธิภาพโดยตรง หลายๆ อัลกอริทึมจะปรับลดเพื่อหลีกเลี่ยงการรู้จำคำตอบมากเกินไป (Over

fitting) ซึ่งการปรับลดจะมีผลกระทบกับความสามารถในการเรียนรู้ประเภทข้อมูลที่หายากและกรณีหายาก

ในหลายงานวิจัยที่ผ่านมาได้แก้ไขปัญหานี้หลากหลายวิธี เช่น การเลือกค่าปรับลดอื่นๆ ที่มีความเหมาะสมแทนรูปแบบเดิม, การเลือกมาตรวัดที่มีความเหมาะสม, การเรียนรู้แบบมีต้นทุน (Cost-sensitive learning)

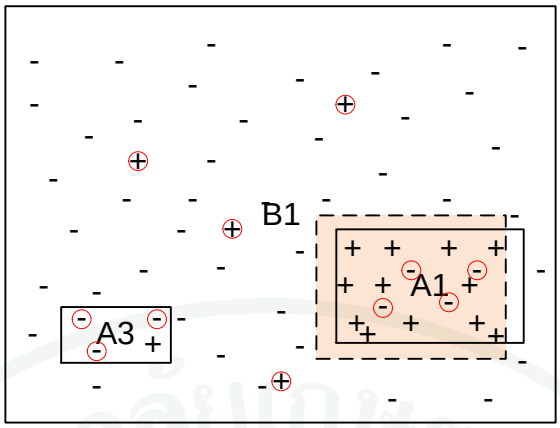
3.6. ข้อมูลรบกวน (Noise) มีผลกระทบสูงมากต่อความสัมพันธ์ของข้อมูลที่เบาบางเมื่อเทียบกับความสัมพันธ์ของข้อมูลหลัก ถ้าหากมีปริมาณข้อมูลรบกวนมากๆ ปรากฏอยู่บนความสัมพันธ์ของข้อมูลที่เบาบาง จะทำให้โมเดลการเรียนรู้ไม่สามารถเรียนรู้ขอบเขตการตัดสินใจได้เลยหรือผิดไปจากที่ควรจะเป็น แต่จะไม่ค่อยมีผลกระทบกับความสัมพันธ์ของข้อมูลหลัก เนื่องจากขอบเขตการตัดสินใจครอบคลุมบริเวณพื้นที่ของข้อมูลกลุ่มใหญ่อยู่แล้ว



ภาพที่ 14 แสดงถึงข้อมูลที่ไม่มีข้อมูลรบกวน

จากภาพที่ 15 จะพบว่าเมื่อมีข้อมูลรบกวนเกิดขึ้นซึ่งแสดงด้วยวงกลมล้อมรอบปรากฏอยู่มากมายบนความสัมพันธ์ของข้อมูลที่เบาบาง อาจจะส่งผลกระทบทำให้การเรียนรู้เพื่อหาขอบเขตการตัดสินใจเกิดความผิดพลาด เช่นในกรณี A3 เมื่อมีข้อมูลรบกวนจะทำให้โมเดลการเรียนรู้ไม่สามารถจำแนกได้ เพราะข้อมูลส่วนใหญ่บริเวณนั้นเป็นคลาสลบ

ในหลายงานวิจัยที่ผ่านมาได้แก้ไขปัญหานี้หลากหลายวิธี เช่น การเลือกค่าปรับลดอื่นๆ ที่มีความเหมาะสมแทนรูปแบบเดิม, การเพิ่มข้อมูลตัวอย่างแบบก้าวหน้า (Advanced sampling)



ภาพที่ 15 แสดงถึงข้อมูลที่มีข้อมูลรบกวนอยู่ใน Rare Cases

จากปัญหาดังกล่าวของฐานข้อมูลที่มีการกระจายตัวแบบไม่สมดุล Weiss ได้สรุปปัญหาและแนวทางการแก้ไขปัญหาจากงานวิจัยต่างๆ แสดงไว้ในตารางที่ 6

ตารางที่ 6 ปัญหาของการไมนิ่งข้อมูลที่ไม่สมดุลและวิธีแก้ไขปัญหา

ปัญหาของการไมน์ข้อมูล	วิธีการแก้ไขปัญหา
1. การเลือกมาตรวัดที่ไม่เหมาะสม (Improper Evaluation Metrics)	- เลือกมาตรวัดอื่น (More appropriate evaluation metrics) - การเรียนรู้แบบมีต้นทุน (Cost-Sensitive Learning)
2. ความขาดแคลนข้อมูล (Lack of Data)	- การเพิ่มข้อมูลตัวอย่าง (Over Sampling)
3. ความขาดแคลนความสัมพันธ์ของข้อมูล (Relative Lack of Data)	- การเรียนรู้เฉพาะประเภทข้อมูลที่หายาก เพียงอย่างเดียว (Learn only the rare class) - การแบ่งข้อมูลออกเป็นส่วน ๆ (Segmenting the data) - การเพิ่มลดข้อมูลตัวอย่าง (Sampling over-and-under) - หลีกเลี่ยงเทคนิคการค้นหาแบบกรีดี้ (Non greedy search techniques) - การใช้กฎเกณฑ์แบบ 2 เฟส (Two-phase rule induction) - การบันทึกรายการที่หายาก (Accounting for rare items) - การเรียนรู้แบบมีต้นทุน (Cost-sensitive learning) - การป้อนความรู้ร่วมกับผู้ใช้ (Knowledge/human interaction)

ตารางที่ 6 (ต่อ)

ปัญหาของการไมน์ข้อมูล	วิธีการแก้ไขปัญหา
	- การแยกคลาสสำหรับกรณีที่หายาก (Rare cases into separate classes) - เลือกมาตรวัดอื่นๆ (More appropriate evaluation metrics) - การปรับลดอื่นๆ ที่เหมาะสม (More appropriate inductive bias) - การเพิ่มกำลัง (Boosting)
4. ข้อมูลที่กระจัดกระจาย (Data Fragmentation)	- หลีกเลี่ยงเทคนิคการค้นหาแบบกรีดี้ (Non greedy search techniques) - การใช้กฎเกณฑ์แบบ 2 เฟส (Two-phase rule induction) - การเรียนรู้เฉพาะประเภทข้อมูลที่หายาก เพียงอย่างเดียว (Learn only the rare class) - การแยกคลาสสำหรับกรณีที่หายาก (Rare cases into separate classes) - การเพิ่มลดข้อมูลตัวอย่าง (Sampling)
5. การปรับลดที่ไม่เหมาะสม (Inappropriate Inductive Bias)	- การปรับลดอื่นๆ ที่เหมาะสม (More appropriate inductive bias) - เลือกมาตรวัดอื่น (More appropriate evaluation metrics) - การเรียนรู้แบบมีต้นทุน (Cost-Sensitive Learning)
6. ข้อมูลรบกวน (Noise)	- การเพิ่มข้อมูลตัวอย่างแบบก้าวหน้า (Advanced Sampling) - การปรับลดอื่นๆ ที่เหมาะสม (More appropriate inductive bias)

มาตรวัดผล (Measures)

มาตรวัดที่ให้ค่าความแม่นยำของทุกๆ คลาสมีค่าความสำคัญเท่าๆ กันอาจจะไม่เหมาะสมในการวิเคราะห์ปัญหาข้อมูลที่ไม่สมดุล ซึ่งประเภทข้อมูลคลาสที่เบาบางจะต้องมีการพิจารณาให้ความสำคัญมากกว่าประเภทข้อมูลคลาสหลัก สำหรับการจำแนกประเภทข้อมูลแบบไบนารี่ (Binary Classification) ในปัญหาของความไม่สมดุลของข้อมูลและคลาสแรกที่เราให้ความสนใจคือประเภทข้อมูลคลาสที่เบาบางหรือคลาสรอง ดังนั้นการเลือกมาตรวัดที่เหมาะสมจะช่วยเพิ่มประสิทธิภาพในการไมนิ่งข้อมูลให้ดีขึ้นได้จากกระบวนการสืบค้นและคัดเลือกกฎความสัมพันธ์ที่มีคุณภาพ

ในขั้นตอนของการสร้างกฎความสัมพันธ์จะเห็นได้ว่ามีกฎจำนวนมากมายที่ถูกสร้างขึ้นจนไม่สามารถนำกฎความสัมพันธ์ทั้งหมดเหล่านั้นไปใช้ในการวิเคราะห์ได้ทั้งหมด ดังนั้นจึงจำเป็นต้องคัดเลือกเฉพาะกฎความสัมพันธ์ $X \rightarrow Y$ ที่มีคุณภาพที่ดีมาใช้วิเคราะห์ โดยในงานวิจัยที่ผ่านมาก็ได้มีการนำเสนอหลากหลายมาตรวัดเพื่อคัดเลือกกฎความสัมพันธ์ให้มีคุณภาพที่ดีหลายตัวด้วยกัน ซึ่งในที่นี้ผู้วิจัยขอกล่าวถึงเฉพาะมาตรวัดที่เกี่ยวข้องกับวิทยานิพนธ์เล่มนี้

1. มาตรวัดคุณภาพของข้อมูล

1.1. มาตรวัดค่าสนับสนุน (Support) (Agrawal *et al.*, 1993) เป็นมาตรวัดเพื่อพิจารณาว่ากฎความสัมพันธ์ $X \rightarrow Y$ ที่เกิดขึ้นในเซตข้อมูลมีความถี่เท่าไหร่ โดยคำนวณได้จากสมการที่ (4)

$$supp(X \rightarrow Y) = P(X,Y) = \frac{\sigma(X \cup Y)}{N} \tag{4}$$

1.2. มาตรวัดค่าความเชื่อมั่น (Confident) (Agrawal *et al.*, 1993) เป็นมาตรวัดเพื่อพิจารณากฎความสัมพันธ์ $X \rightarrow Y$ ว่าในทรานแซกชั่นที่ประกอบด้วยไอเท็มเซต X ทั้งหมดนั้น มีไอเท็มเซต Y เกิดขึ้นร่วมกันบนทรานแซกชั่นใดๆ มีความถี่เท่าไหร่ โดยคำนวณได้จากสมการที่ (5)

$$conf(X \rightarrow Y) = P(Y \mid X) = \frac{\sigma(X \cup Y)}{\sigma(X)} \tag{5}$$

1.3. มาตรวัดความน่าสนใจ (Interest Factor) (Silverstein *et al*.,1998) เป็นมาตรวัดเพื่อพิจารณาความสัมพันธ์และความเป็นอิสระของไอเท็มเซต X และไอเท็มเซต Y ที่ปรากฏในกฎความสัมพันธ์ $X \rightarrow Y$ ซึ่งถูกนำไปใช้ในการวิเคราะห์สหสัมพันธ์ของกฎ (Correlation Rules) โดยคำนวณได้จากสมการที่ (6)

$$corr(X \rightarrow Y) = \frac{P(X,Y)}{P(X)P(Y)} \tag{6}$$

จากสมการที่ 6 สามารถวิเคราะห์ผลลัพธ์จากสมการได้ดังนี้

1.3.1. สหสัมพันธ์เชิงลบ (Negatively Correlated) หรือ $corr(X \rightarrow Y) < 1$ แสดงถึงการเกิดขึ้นของไอเท็มเซต X มีความสัมพันธ์ตรงข้ามกับการเกิดขึ้นของไอเท็มเซต Y ตัวอย่างเช่น คนที่มีส่วนสูงเกิน 170 เซนติเมตร (ไอเท็มเซต X) จะมีน้ำหนักตัวน้อยกว่า 45 กิโลกรัม (ไอเท็มเซต Y) นั้นคือ $corr(X \rightarrow Y) < 1$

1.3.2. สหสัมพันธ์เชิงบวก (Positively Correlated) หรือ $corr(X \rightarrow Y) > 1$ แสดงถึงการเกิดขึ้นของไอเท็มเซต X มีความสัมพันธ์ร่วมกันกับการเกิดขึ้นของไอเท็มเซต Y ตัวอย่างเช่น คนที่มีส่วนสูงเกิน 170 เซนติเมตร (ไอเท็มเซต X) จะมีน้ำหนักตัวมากกว่า 45 กิโลกรัม (ไอเท็มเซต Y) นั้นคือ $corr(X \rightarrow Y) > 1$

1.3.3. สหสัมพันธ์อิสระ (Independent) หรือ $corr(X \rightarrow Y) = 1$ แสดงถึงการเกิดขึ้นของไอเท็มเซต X ไม่มีความสัมพันธ์กับการเกิดขึ้นของไอเท็มเซต Y ตัวอย่างเช่น เพศชาย (ไอเท็มเซต X) มีการตั้งครรภ์ (ไอเท็มเซต Y) นั้นคือ $corr(X \rightarrow Y) = 1$

1.4. มาตรวัดตามวัตถุประสงค์ของความน่าสนใจ (Objective Measures of Interestingness) ใช้สำหรับการหาคุณภาพของรูปแบบความสัมพันธ์ (Quality of association patterns) ซึ่งเป็นอิสระจากโดแมนของข้อมูล (Domain Independent) โดยอาศัยการคำนวณความถี่ที่เกิดขึ้นผ่านตารางการณ์จร (Contingency Table) ที่แสดงไว้ในตารางที่ 7 ซึ่งแสดงถึงตัวอย่างของตารางการณ์จรระหว่างตัวแปร 2 ตัวคือ A และ B โดยสัญลักษณ์ $\overline{A}$ แสดงถึง A ที่ไม่ปรากฏอยู่ในทรานแซกชั่น และ $\overline{B}$ แสดงถึง B ที่ไม่ปรากฏอยู่ในทรานแซกชั่น แต่ละค่าของ f_{ij} ที่อยู่ในตารางขนาด 2×2 แสดงถึงค่าความถี่ที่เกิดขึ้น โดย f_{11} คือค่าความถี่หรือจำนวนครั้งของ A และ B ที่ปรากฏร่วมกันอยู่ในทรานแซกชั่นเดียวกัน f_{01} คือค่าความถี่จำนวนครั้งของ B แต่ไม่ปรากฏว่ามี

A ร่วมอยู่ในทรานแซกชั่นเดียวกัน f_{10} คือค่าความถี่จำนวนครั้งของ A แต่ไม่ปรากฏว่ามี B ร่วมอยู่ในทรานแซกชั่นเดียวกัน f_{00} คือค่าความถี่จำนวนครั้งที่ไม่ปรากฏว่ามี A และ B ร่วมกันอยู่ในทรานแซกชั่นเดียวกัน ผลรวมของแถว f_{1+} แสดงถึงค่าสนับสนุนของ A ขณะที่ ผลรวมของแถว f_{+1} แสดงถึงค่าสนับสนุนของ B และผลรวมของแถว f_{0+} แสดงถึงค่าสนับสนุนที่ไม่มี A ขณะที่ ผลรวมของแถว f_{+0} แสดงถึงค่าสนับสนุนที่ไม่มี B และ N คือผลรวมของทั้งหมด และกำหนดให้สัญลักษณ์ [f_{11}, f_{10}; f_{01}, f_{00}] แทนตารางการณ์จรของตัวแปร A และ B

ตารางที่ 7 ตารางการณ์จรขนาด 2 กรณีสำหรับตัวแปร A และ B

ตัวแปร	**ตัวแปร**		**ผลรวม**
	B	$\overline{B}$	
A	f_{11}	f_{10}	f_{1+}
$\overline{A}$	f_{01}	f_{00}	f_{0+}
ผลรวม	f_{+1}	f_{+0}	N

2. มาตรวัดประสิทธิภาพโมเดลการจำแนกประเภท

การประเมินประสิทธิภาพของโมเดลการจำแนกประเภทนิยมประเมินจากผลการทำนายข้อมูลสำหรับทดสอบว่าสามารถจำแนกประเภทได้ถูกต้องหรือผิดพลาดเกิดขึ้นเป็นจำนวนเท่าไหร่ ผ่านทางตารางเมตริกความสับสน (Confusion Matrix) โดยแสดงไว้ในตารางที่ 8

ตารางที่ 8 ตารางเมตริกความสับสนสำหรับปัญหา 2 คลาส

คลาสจริง	**คลาสที่ทำนาย**	
	+	–
+	f_{++}	f_{+-}
–	f_{-+}	f_{--}

ข้อมูลในแต่ละช่องของเมตริกความสับสนจะถูกแทนด้วยค่าจำนวนนับ ซึ่งสามารถอ้างอิงได้ดังนี้

2.1. ค่าความถูกต้องประเภทบวก (True Positive : TP) หรือ f_{++} คือค่าของจำนวนข้อมูลตัวอย่างคลาสบวกที่โมเดลจำแนกประเภทสามารถทำนายเป็นคลาสบวกได้ถูกต้อง

2.2. ค่าความผิดพลาดประเภทลบ (False Negative : FN) หรือ f_{+-} คือค่าของจำนวนข้อมูลตัวอย่างคลาสบวกที่โมเดลจำแนกประเภททำนายผิดเป็นคลาสลบ

2.3. ค่าความผิดพลาดประเภทบวก (False Positive : FP) หรือ f_{-+} คือค่าของจำนวนข้อมูลตัวอย่างคลาสลบที่โมเดลจำแนกประเภททำนายผิดเป็นคลาสบวก

2.4. ค่าความถูกต้องประเภทลบ (True Negative : TN) หรือ f_{--} คือค่าของจำนวนข้อมูลตัวอย่างคลาสลบที่โมเดลจำแนกประเภทสามารถทำนายเป็นคลาสลบได้ถูกต้อง

การนับค่าในเมตริกความสับสนสามารถแสดงอยู่ในรูปแบบของเปอร์เซ็นต์ความถูกต้องและเปอร์เซนต์ความผิดพลาดได้ดังนี้

2.5. อัตราความถูกต้องประเภทบวก (True Positive Rate : TPR) หรือมาตรวัดความไว (Sensitivity) ซึ่งจะแสดงถึงประสิทธิภาพความแม่นยำในการทำนายข้อมูลตัวอย่างคลาสบวกได้ถูกต้องด้วยโมเดลการจำแนกประเภทและสามารถเขียนเป็นสมการได้ดังสมการที่ (7)

$$TPR = \frac{TP}{(TP + FN)} \tag{7}$$

2.6. อัตราความถูกต้องประเภทลบ (True Negative Rate : TNR) หรือมาตรวัดความจำเพาะ (Specificity) ซึ่งจะแสดงถึงประสิทธิภาพความแม่นยำในการทำนายข้อมูลตัวอย่างคลาสลบได้ถูกต้องด้วยโมเดลการจำแนกประเภทและสามารถเขียนเป็นสมการได้ดังสมการที่ (8)

$$TNR = \frac{TN}{(TN + FP)} \tag{8}$$

2.7. อัตราความผิดพลาดประเภทบวก (False Positive Rate : FPR) ซึ่งจะแสดงถึงความความผิดพลาดในการทำนายข้อมูลตัวอย่างคลาสลบผิดเป็นคลาสบวกของโมเดลการจำแนกประเภทและสามารถเขียนเป็นสมการได้ดังสมการที่ (9)

$$FPR = \frac{FP}{(TN + FP)} \tag{9}$$

2.8. อัตราความผิดพลาดประเภทลบ (False Negative Rate : FNR) ซึ่งจะแสดงถึงความความผิดพลาดในการทำนายข้อมูลตัวอย่างคลาสบวกผิดเป็นคลาสลบของโมเดลการจำแนกประเภทและสามารถเขียนเป็นสมการได้ดังสมการที่ (10)

$$FNR = \frac{FN}{(TP + FN)} \tag{10}$$

2.9. มาตรวัดความเที่ยงตรง (Precision) ใช้สำหรับการพิจารณาความเที่ยงตรงในกลุ่มข้อมูลตัวอย่างที่ถูกโมเดลการจำแนกประเภททำนายเป็นคลาสบวกทั้งหมด ค่าความเที่ยงตรงยิ่งมีค่ามากๆ จะแสดงถึงโมเดลการจำแนกประเภทมีความผิดพลาดประเภทบวกน้อย สามารถเขียนเป็นสมการได้ดังสมการที่ (11)

$$Precision,\ p = \frac{TP}{(TP + FP)} \tag{11}$$

2.10. มาตรวัดค่าจดจำ (Recall) ใช้สำหรับการพิจารณาความถูกต้องของการทำนายข้อมูลตัวอย่างคลาสบวกด้วยโมเดลการจำแนกประเภท ค่าจดจำยิ่งมีค่ามากๆ จะแสดงถึงข้อมูลตัวอย่างคลาสบวกถูกทำนายผิดพลาดเป็นคลาสลบน้อยมาก ซึ่งเหมือนกับมาตรวัดอัตราความถูกต้องประเภทบวก (TPR) สามารถเขียนเป็นสมการได้ดังสมการที่ (12)

$$Recall\ ,r = \frac{TP}{(TP + FN)} \tag{12}$$

2.11. มาตรวัดเอฟ (F-Measure) เป็นมาตรวัดที่สอดผสานทั้งความเที่ยงตรงและค่าจดจำเข้าด้วยกัน ซึ่งถ้าโมเดลการจำแนกประเภทระบุทุกๆ รายการเป็นคลาสบวกก็จะมีค่าความจดจำที่ดีเยี่ยมแต่มีค่าความเที่ยงตรงที่แย่มาก ในทางตรงกันข้ามถ้าโมเดลการจำแนกประเภทระบุข้อมูลสำหรับทดสอบเป็นคลาสบวกทุกๆ รายการก็จะมีค่าความเที่ยงตรงสูงมากแต่มีค่าความจดจำที่แย่มากเช่นกัน ดังนั้นในการสร้างโมเดลการจำแนกประเภทข้อมูลให้มีความเที่ยงตรงและค่าจดจำที่ให้ค่าสูงที่สุดทั้งคู่ยังคงเป็นเรื่องที่ท้าทายในการสร้างอัลกอริทึมสำหรับการจำแนกประเภท ดังนั้นมาตร

วัดเอฟจึงถูกนำมาแสดงถึงความกลมกลืนระหว่างความเที่ยงตรงและค่าจดจำของโมเดลการจำแนกประเภท และสามารถเขียนเป็นสมการได้ดังสมการที่ (13)

$$F_1 = \frac{2}{\frac{1}{r}+\frac{1}{p}} = \frac{2rp}{r+p} = \frac{2\times TP}{2\times TP + FP + FN} \tag{13}$$

2.12. มาตรวัดความแม่นยำ (Accuracy) ใช้สำหรับวัดความแม่นยำในการทำนายได้ถูกต้องของโมเดลการจำแนกประเภทจากทุกคลาสเทียบกับข้อมูลที่ใช้ทดสอบทั้งหมด และสามารถเขียนเป็นสมการได้ดังสมการที่ (14)

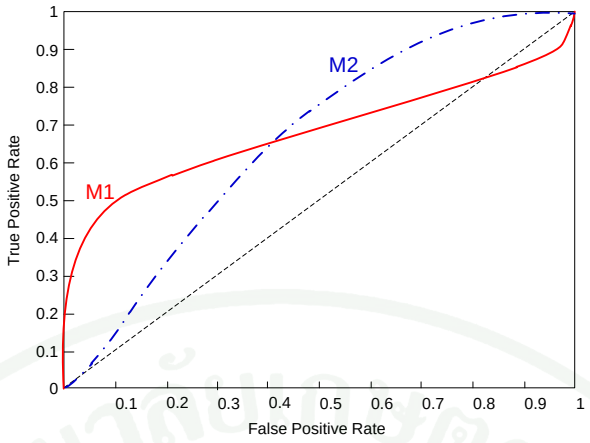
$$Accuracy = \frac{TP + TN}{(TP + FN + FP + TN)} \tag{14}$$

2.13. เส้นโค้งรับคุณลักษณะการทำงาน (Receiver Operating Characteristic Curve : ROC) มีลักษณะการแสดงผลเป็นกราฟิกซึ่งแสดงภาวะถ่วงดุล (Tradeoff) ระหว่างอัตราความถูกต้องประเภทบวก (TPR) และอัตราความผิดพลาดประเภทบวก (FPR) ของโมเดลการจำแนกประเภท โดยอัตราความถูกต้องประเภทบวกจะถูกแสดงผลบนแกน Y และอัตราความผิดพลาดประเภทบวกจะถูกแสดงผลบนแกน X ซึ่งแต่ละจุดที่วาดลงบนเส้นโค้งได้มาจากโมเดลการจำแนกประเภทแต่ละตัว ตัวอย่างเช่น M1 และ M2 ดังภาพที่ 16 ซึ่งเส้นโค้ง ROC นั้นสามารถตีความได้ดังนี้

(TPR=0, FPR=0) โมเดลทำนายทุกตัวอย่างข้อมูลเป็นคลาสลบ

(TPR=1, FPR=1) โมเดลทำนายทุกตัวอย่างข้อมูลเป็นคลาสบวก

(TPR=1, FPR=0) โมเดลในอุดมคติ



ภาพที่ 16 เส้นโค้ง ROC ของโมเดลจำแนกประเภท 2 ตัวที่ต่างกัน

ดังนั้นโมเดลการจำแนกประเภทควรจะให้ค่าใกล้เคียงมุมบนซ้ายให้มากที่สุด จากภาพที่ 8 สามารถเปรียบเทียบประสิทธิภาพของโมเดลการจำแนกประเภทได้ว่าโมเดล M1 มีความสามารถจำแนกได้ดีกว่าโมเดล M2 เมื่อ FPR มีค่าน้อยกว่า 0.4 ขณะที่ M2 สามารถจำแนกได้ดีกว่าเมื่อ FPR มีค่ามากกว่า 0.4

ในการหาค่าเฉลี่ยว่าโมเดลการจำแนกประเภทตัวไหนดีกว่า จะใช้การหาพื้นที่ใต้เส้นโค้ง ROC (Area under the ROC curve: AUC) มาเป็นมาตรวัด ถ้าโมเดลที่ดีสมบูรณ์จะให้ค่า AUC เท่ากับ 1 และโมเดลที่ให้ค่าลักษณะเดาสุ่มจะมีค่าประมาณ 0.5 ดังนั้นโมเดลที่ดีกว่าจะต้องมีค่า AUC ที่มากกว่า

การเรียนรู้แบบมีต้นทุน

อัลกอริทึมส่วนใหญ่มักอยู่บนสมมุติฐานที่ว่าชุดข้อมูลที่ทำการเรียนรู้และทดสอบนั้นมีความสมดุลหรือคลาสที่ให้ความสนใจคือคลาสหลักซึ่งจะพิจารณาว่าต้นทุนการทำนายผิดพลาดของทุกคลาสนั้นมีต้นทุนที่เท่ากัน แต่ในขณะที่ชุดข้อมูลที่ไม่สมดุลและคลาสแรกที่เราให้ความสนใจนั้นคือคลาสรอง ดังนั้นต้นทุนการทำนายคลาสรองผิดพลาดนั้นย่อมมีต้นทุนสูงกว่าการทำนายคลาสหลักผิดพลาด การเรียนรู้แบบมีต้นทุนจึงเป็นอีกเทคนิคหนึ่งของการไมนิ่งข้อมูลที่ถูกนำมาจัดการกับข้อมูลที่มีต้นทุนความผิดพลาดที่ไม่เท่ากัน โดยการเรียนรู้แบบมีต้นทุนนั้นจะอาศัยเมตริกต้นทุน (Cost Matrix) มาใช้เป็นตัวกำหนดค่าปรับและรางวัลหากโมเดลการจำแนกประเภททำการจำแนกข้อมูลจากคลาสหนึ่งไปเป็นคลาสอื่นๆ ถ้ากำหนดให้ $C(i, j)$ แสดงถึงต้นทุนการทำนายข้อมูลจากคลาส i เป็นคลาส j ดังนั้น $C(+,-)$ ก็คือต้นทุนของความผิดพลาดประเภทลบ และ $C(-,+)$ ก็คือต้นทุนของความผิดพลาดประเภทบวก ส่วนค่าต้นทุน $C(+,+)$ และ $C(-,-)$ ในเมตริกต้นทุนก็คือรางวัลของตัวจำแนกประเภทเมื่อทำนายได้ถูกต้อง ดังนั้นในการพิจารณาต้นทุนรวมของโมเดลการจำแนกประเภท M สามารถหาได้จากสมการที่ (15) และตัวอย่างเมตริกต้นทุนแสดงในตารางที่ 9

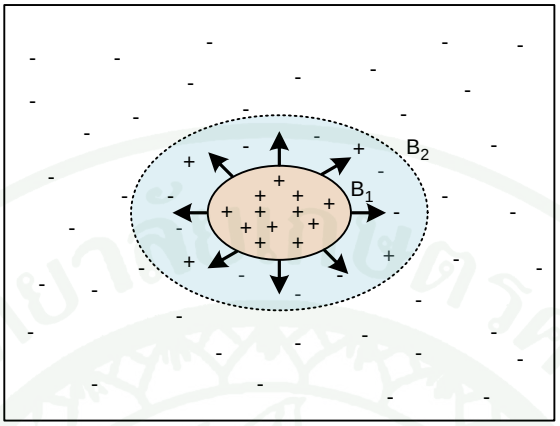
$$C_t(M) = TP \times C(+,+) + FP \times C(-,+) + FN \times C(+,-) + TN \times C(-,-) \quad (15)$$

ตารางที่ 9 ตัวอย่างเมตริกต้นทุน [-1,100;1,0]

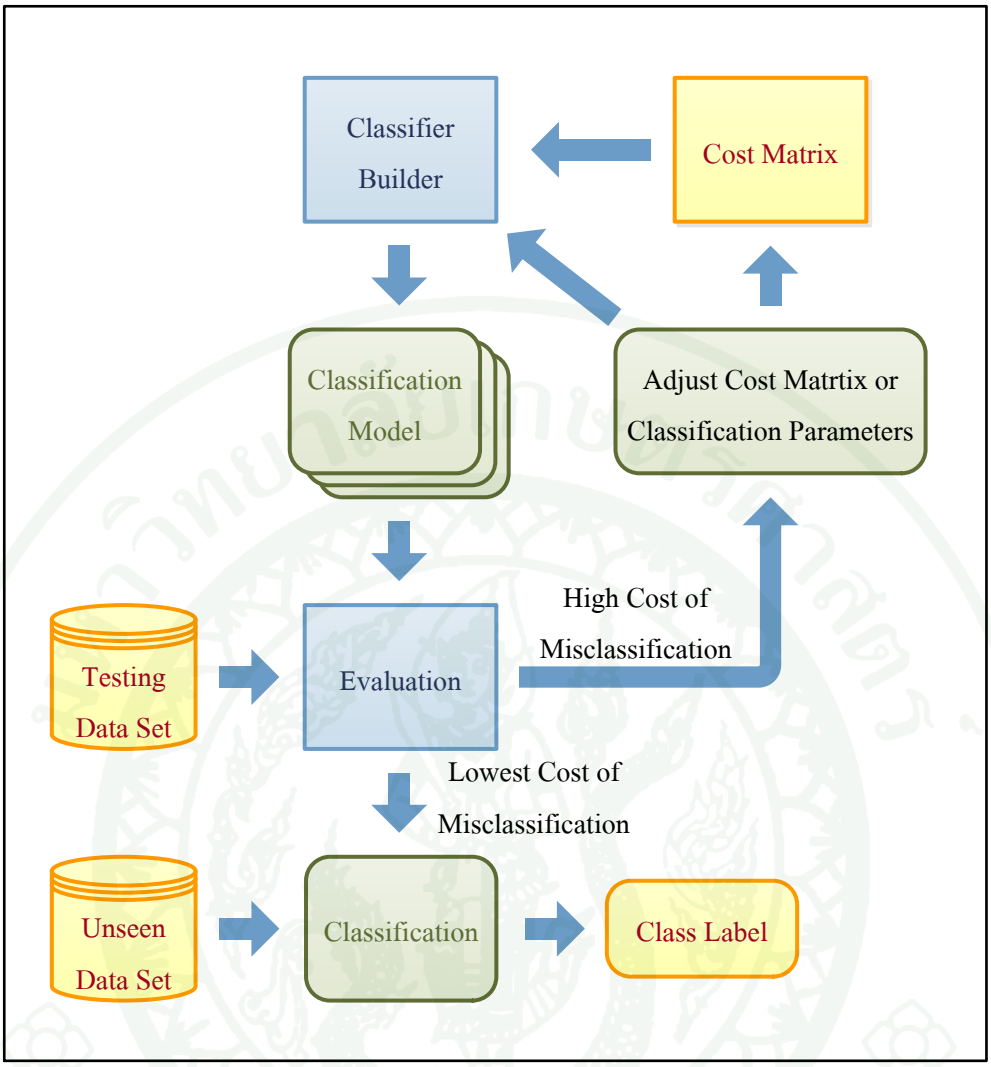
คลาสจริง	**คลาสที่ทำนาย**	
	+	–
+	-1	100
–	1	0

เทคนิคการจำแนกประเภทแบบมีต้นทุน (Cost-Sensitive Classification) คือการอาศัยเมตริกต้นทุนมาพิจารณาขณะที่สร้างโมเดลเพื่อให้มีต้นทุนความผิดพลาดที่ต่ำที่สุด เช่นถ้ากำหนดให้ความผิดพลาดประเภทลบมีต้นทุนที่สูงมาก ดังนั้นอัลกอริทึมสำหรับเรียนรู้จะพยายามลดความผิดพลาดประเภทลบด้วยการขยายขอบเขตการตัดสินใจออกไปยังบริเวรที่มีคลาสลบโดยการสร้างเป็นโมเดลการจำแนกตัวใหม่ เพื่อให้สามารถคลอบคลุมตัวอย่างข้อมูลคลาสบวกให้มากขึ้น ด้วยเหตุผลที่ว่าค่าความผิดพลาดประเภทบวกนั้นมีต้นทุนน้อยกว่าค่าความผิดพลาดประเภทลบ

ซึ่งสามารถพิจารณาได้ดังภาพที่ 17 และขั้นตอนการสร้างตัวจำแนกประเภทแบบมีต้นทุนในภาพที่ 18



ภาพที่ 17 การแก้ไขขอบเขตการตัดสินใจ (จาก B1 เป็น B2) เพื่อลดความผิดพลาดในข้อมูลประเภทลบของโมเดลการจำแนกประเภท



ภาพที่ 18 ขั้นตอนการสร้างตัวจำแนกประเภทแบบมีต้นทุน

ความน่าจะเป็นและสถิติ

ในการวิเคราะห์ถึงคุณภาพของกฎความสัมพันธ์นั้นหลายงานวิจัยได้อาศัยหลักความน่าจะเป็นและหลักทางสถิติมาช่วยในการวิเคราะห์ความสัมพันธ์ที่เกิดขึ้น ซึ่งในวิทยานิพนธ์เล่มนี้จะกล่าวถึงเฉพาะในส่วนที่เกี่ยวข้องกับงานวิจัย

1. ความน่าจะเป็นแบบมีเงื่อนไข (Conditional Probability)

ถ้ากำหนดให้ตัวแปร X และ Y เป็นตัวแปรสุ่ม ดังนั้น $P(X,Y)$ แสดงถึงความน่าจะเป็นของทั้งสองตัวแปรที่เกิดขึ้นร่วมกัน ขณะที่ $P(Y \mid X)$ คือความน่าจะเป็นของตัวแปร Y เมื่อกำหนดตัวแปร X มาให้ ซึ่ง $P(Y \mid X)$ ก็คือความน่าจะเป็นแบบมีเงื่อนไข การคำนวณความน่าจะเป็นแบบมีเงื่อนไขสามารถคำนวณได้จากสมการที่ (16)

$$P(Y \mid X) = \frac{P(X,Y)}{P(X)} \tag{16}$$

ถ้ากำหนดให้ตัวแปร X และ Y เป็นอิสระจากกันแล้ว $P(Y \mid X) = P(Y)$ ความน่าจะเป็นแบบมีเงื่อนไขสามารถแสดงอยู่ในรูปอื่นได้ดังสมการที่ (17) หรือที่รู้จักกันในทฤษฎีของเบย์ (Bayes Theorem) ซึ่งก็คือการหาความความน่าจะเป็นของเหตุการณ์ที่มีอยู่เมื่อทราบข้อมูลหรือหลักฐานเพิ่มเติม (Evidence) หรือเป็นการรวมข้อมูลใหม่เข้ากับความรู้ที่มีอยู่เดิม ดังนั้นค่าความน่าจะเป็นแบบมีเงื่อนไขมากหรือน้อย จะแสดงถึงความน่าจะเป็นของเหตุการณ์ที่เกิดขึ้น บนหลักฐานที่กำหนดให้ และสามารถเขียนอยู่ในรูปของความหมายได้ดังสมการที่ (18)

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)} \tag{17}$$

$$posterior = \frac{likelihood \times prior}{evidence} \tag{18}$$

$P(Y \mid X)$ หรือ *Posterior Probability* คือความน่าจะเป็นของเหตุการณ์ Y ที่เกิดขึ้นภายหลังเกิดเหตุการณ์ X แล้ว เมื่อพิจารณาในรูปของกฎความสัมพันธ์ $X \rightarrow Y$ จะได้ว่าถ้ามีไอเท็มเซต X แล้วมีความน่าจะเป็นที่จะเกิดไอเท็มเซต Y เป็นเท่าไหร่ ซึ่งเทียบได้กับค่าความเชื่อมั่นนั้นเอง $conf(X \rightarrow Y)$

$P(X|Y)$ หรือ *Posterior Probability* คือความน่าจะเป็นของเหตุการณ์ X ที่เกิดขึ้นภายหลังเกิดเหตุการณ์ Y แล้ว หรือเรียกอีกอย่างว่า *Likelihood* เมื่อพิจารณาในรูปของกฎความสัมพันธ์ $X \rightarrow Y$ จะได้ว่าภายใต้ไอเท็มเซต Y มีความน่าจะเป็นที่จะมีไอเท็มเซต X ปรากฏอยู่เป็นเท่าไหร่

$P(Y)$ หรือ *Prior Probability* คือความน่าจะเป็นของตัวแปรหรือเหตุการณ์ Y เมื่อพิจารณาในรูปของกฎความสัมพันธ์ $X \rightarrow Y$ จะได้ว่าความน่าจะเป็นในการแพร่กระจายของคลาสหรือไอเท็มเซต Y เกิดขึ้นเท่าไหร่

$P(X)$ หรือ *Marginal Probability* คือความน่าจะเป็นของตัวแปรหรือเหตุการณ์ X หรือเรียกอีกอย่างว่าหลักฐาน (Evident) เมื่อพิจารณาในรูปของกฎความสัมพันธ์ $X \rightarrow Y$ จะได้ว่ามีความน่าจะเป็นที่จะพบไอเท็มเซต X เป็นเท่าไหร่

ตัวอย่างการคำนวณกฎความสัมพันธ์ $\{Beer\} \rightarrow \{Diaper\}$ ของข้อมูลทรานแซกชั่นตารางที่ 1

$$P(X) = P\{Beer\} = \frac{3}{5} = 0.6$$
$$P(Y) = P\{Diaper\} = \frac{4}{5} = 0.8$$
$$P(X|Y) = P\{Beer \mid Diaper\} = \frac{3}{4} = 0.75$$
$$P(Y|X) = P\{Diaper \mid Beer\} = \frac{0.75 \times 0.8}{0.6} = 1$$

2. การทดสอบความถูกต้องของฟิชเชอร์ (Fisher Exact Test)

การทดสอบความถูกต้องของฟิชเชอร์ก็คือการทดสอบนัยสำคัญทางสถิติโดยใช้การวิเคราะห์ผ่านตารางการณ์จร เพื่อทดสอบว่าข้อมูลทั้งสองกลุ่มนั้นแตกต่างกันอย่างมีนัยสำคัญหรือไม่ โดยที่ข้อมูลทั้งสองกลุ่มนั้นเป็นอิสระจากกัน และสามารถแบ่งได้เป็น 2 ประเภท พิจารณาได้จากตารางที่ 10

ตารางที่ 10 การทดสอบความถูกต้องของฟิชเชอร์ผ่านตารางการณ์จร

ประเภทข้อมูล	กลุ่มข้อมูล		ผลรวม
	I	II	
1	A	B	A+B
2	C	D	C+D
ผลรวม	A+C	B+D	N

การทดสอบความถูกต้องของฟิชเชอร์มีขั้นตอนการทดสอบตวามแตกต่างของกลุ่มตัวอย่างข้อมูล 2 กลุ่มที่เป็นอิสระจากกันดังนี้

2.1. กำหนดสมมุติฐาน

H_0 คือสมมุติฐานหลักที่ว่า โอกาสที่กลุ่มข้อมูลทั้ง 2 กลุ่มจะถูกจำแนกประเภทออกเป็น ประเภทที่ 1 และ 2 นั้นไม่ได้แตกต่างกัน $H_0 : P_1 = P_2$

H_1 คือสมมุติฐานรองที่ว่า โอกาสที่กลุ่มข้อมูลทั้ง 2 กลุ่มจะถูกจำแนกประเภทออกเป็น ประเภทที่ 1 และ 2 นั้นแตกต่างกัน $H_1 : P_1 \neq P_2$

2.2. กำหนดค่าวิกฤตหรือระดับนัยสำคัญ α เช่น 0.01 หรือ 0.05 โดยค่ามาตรฐานคือ 0.05

2.3. คำนวณหาค่าความน่าจะเป็น p_{value} จากค่าต่างๆ บนตารางการณ์จรด้วยสมการที่ (19) หรือ (20)

$$p_{value} = \frac{\binom{A+C}{A}\binom{B+D}{B}}{\binom{N}{A+B}} \tag{19}$$

$$p_{value} = \frac{(A+B)!(C+D)!(A+C)!(B+D)!}{N!A!B!C!D!} \tag{20}$$

2.4. ทดสอบนัยสำคัญโดยการเปรียบเทียบค่า p_{value} ที่คำนวณได้จากขั้นตอนก่อนหน้าเทียบกับค่าวิกฤตที่กำหนดไว้ โดยสามารถแปลผลได้ดังนี้

ถ้าค่า $p_{value} > \alpha$ คือการยอมรับในสมมุติฐานหลัก H_0 นั้นคือโอกาสของข้อมูลกลุ่ม I และ II สามารถนำมาจำแนกเป็นประเภทที่ 1 และ 2 ได้ไม่แตกต่างกัน

ถ้าค่า $p_{value} \leq \alpha$ คือการปฏิเสธในสมมุติฐานหลัก H_0 หรือยอมรับในสมมุติฐานรอง H_1 นั้นคือโอกาสของข้อมูลกลุ่ม I และ II สามารถนำมาจำแนกเป็นประเภทที่ 1 และ 2 ได้แตกต่างกัน

ตัวอย่างการคำนวณการทดสอบความถูกต้องของฟิชเชอร์เช่น การทดสอบความแตกต่างของโอกาสที่นิสิตชายและหญิงจะมีความสนใจในวิชาดาต้าไมนิ่งมากและน้อยที่ระดับนัยสำคัญ 0.01 โดยสามารถแทนค่าลงในตารางการณ์จรได้ดังตารางที่ 11

ตารางที่ 11 ตัวอย่างข้อมูลนิสิตที่มีความสนใจในวิชาดาต้าไมนิ่ง

ความสนใจวิชาดาต้าไมนิ่ง	**นิสิต**		**ผลรวม**
	ชาย	หญิง	
มาก	4	11	15
น้อย	7	3	10
ผลรวม	11	14	N=25

H_0 คือสมมุติฐานหลักที่ว่า นิสิตชายหญิงมีความสนใจในวิชาดาต้าไมนิ่งมากและน้อยนั้นไม่แตกต่างกัน

H_1 คือสมมุติฐานรองที่ว่า นิสิตชายหญิงมีความสนใจในวิชาดาต้าไมนิ่งมากและน้อยนั้นแตกต่างกัน

คำนวณค่าความน่าจะเป็น $p_{value} = \dfrac{(4+11)!(7+3)!(4+7)!(11+3)!}{25!4!11!7!3!} = 0.0367$

ดังนั้น $p_{value} > \alpha$ $(0.0367 > 0.01)$ คือการยอมรับในสมมุติฐานหลัก H_0 ซึ่งหมายความว่านิสิตชายและหญิงมีความสนใจวิชาดาต้าไมนิ่งไม่แตกต่างกัน ที่ระดับนัยสำคัญ 0.01

งานวิจัยที่เกี่ยวข้อง

วิทยานิพนธ์เล่มนี้ผู้วิจัยนำเสนองานวิจัยแนวทางการปรับปรุงประสิทธิภาพของการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์บนฐานข้อมูลที่ไม่สมดุลซึ่งมีการปรับปรุงประสิทธิภาพด้านความเร็วในการทำงานของ Imbalanced Associative Classification : IMAC ที่นำเสนอโดย พูนเพิ่มและกฤษณะ (2555, 2554) ซึ่งจากที่ผ่านมาการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์นั้นไม่ได้มุ่งแก้ปัญหาของความไม่สมดุล ทำให้มีงานวิจัยที่เกี่ยวข้องกับปัญหาเรื่องนี้มีไม่มากนัก ซึ่งปัญหาความไม่สมดุลของข้อมูลนั้นเกิดขึ้นจากสาเหตุหลักคือเรื่องของการมีข้อมูลไม่เพียงพอของคลาสที่เบาบาง ทำให้สืบค้นหาความสัมพันธ์ของข้อมูลเป็นไปได้ยากส่งผลให้ตัวจำแนกประเภทด้วยกฎความสัมพันธ์ไม่มีประสิทธิภาพบนข้อมูลคลาสรอง ดังนั้นในขั้นตอนการสร้างกฎความสัมพันธ์จึงจำเป็นที่ต้องคัดเลือกกฎที่มีคุณภาพมาให้ได้จากจำนวนกฎความสัมพันธ์มากมายที่เกิดขึ้น โดยผู้วิจัยขออธิบายเป็นหัวข้อตามลักษณะของปัญหาในการทำไมนิ่งข้อมูลที่ไม่สมดุล

1. ปัญหาจากมาตรวัด

ปัญหาการจำแนกประเภทด้วยกฎความสัมพันธ์ CBA (Liu *et al*., 1998) บนฐานข้อมูลที่ไม่สมดุลส่วนหนึ่งนั้นเกิดจากข้อจำกัดของมาตรวัดค่าสนับสนุนและมาตรวัดค่าความเชื่อมั่น โดยมาตรวัดทั้งสองจะส่งผลให้กฎความสัมพันธ์ที่มีศักยภาพเพียงพอที่จะเป็นกฎความสัมพันธ์ที่มีนัยน่าสนใจนั้นถูกกำจัดทิ้งไป เพราะมีค่าสนับสนุนของกฎที่ต่ำ ซึ่งไม่สามารถผ่านเกณฑ์ค่าสนับสนุนขั้นต่ำที่กำหนดไว้ได้ เมื่อพิจารณาถึงมาตรวัดความความเชื่อมั่น จะขออธิบายผ่านตัวอย่างที่อยู่ในตารางที่ 12

ตารางที่ 12 ตัวอย่างข้อมูลของนักดื่มจำนวน 1000 ราย

ตัวแปร	**ตัวแปร**		**ผลรวม**
	coffee	$\overline{coffee}$	
tea	150	50	200
$\overline{tea}$	650	150	800
ผลรวม	800	200	1000

จากข้อมูลที่กำหนดให้ในตารางที่ 12 ถ้าต้องการวิเคราะห์กฎความสัมพันธ์ $\{tea\} \rightarrow \{coffee\}$ จะเห็นว่าคนที่ดื่มชาแล้วจะดื่มกาแฟด้วยนั้นมีค่าสนับสนุน 15% และค่าความเชื่อมั่น 75% ซึ่งถือว่ามีเหตุผลเพียงพอที่จะพิจารณาว่าเป็นกฎที่น่าสนใจ หากพิจารณาลึกลงไปจะเห็นว่าจำนวนคนที่ดื่มกาแฟโดยไม่สนใจว่าจะดื่มชาด้วยหรือไม่นั้นมีถึง 80% $conf(\{\} \rightarrow \{coffee\}) = 80\%$ แต่ขณะที่นักดื่มชาและดื่มกาแฟด้วยนั้นมีถึง 75% $conf(\{tea\} \rightarrow \{coffee\}) = 75\%$ หากพิจารณาจากข้อมูลดังกล่าวนี้อาจจะเข้าใจได้ว่านักดื่มชานั้นพิจารณามาจากความน่าจะเป็นของนักดื่มกาแฟทั้งหมดจาก 80% ลดลงเป็น 75% นั้นคืออาจจะทำให้เกิดความเข้าใจผิดพลาดได้จากการที่มีค่าความเชื่อมั่นที่สูง ซึ่งเป็นหลุมพลางที่อาจจะทำให้ละเลยการพิจารณาค่าสนับสนุนร่วมด้วยจากลำดับของกฎความสัมพันธ์ของไอเท็มเซตทางด้านขวา หากพิจารณาค่าสนับสนุนของนักดื่มกาแฟร่วมด้วยแล้วจะไม่ประหลาดใจเลยว่านักดื่มชานั้นจะดื่มกาแฟด้วย จากตัวอย่างที่ผ่านมาจะเห็นได้ว่า การเลือกกฎความสัมพันธ์ตามค่าความเชื่อมั่นนั้นก็อาจจะทำให้ได้กฎความสัมพันธ์ที่ไม่น่าสนใจมาด้วยเช่นกัน

เพื่อแก้ไขปัญหาข้อจำกัดของค่าสนับสนุนและค่าความเชื่อมั่น Silverstein *et al*. (1998) ได้นำเสนอมาตรวัดความน่าสนใจ ซึ่งเป็นการพิจารณาความถี่ที่เกิดขึ้นของกฎเทียบกับความถี่เส้นฐาน (Baseline) ที่ถูกคำนวณภายใต้สมมุติฐานความเป็นอิสระทางสถิติ โดยความถี่เส้นฐานถูกคำนวณมาจากความถี่ของตัวแปรอิสระ 2 ตัวคู่กัน ดังสมการที่ (21) โดย f_{11} คือความถี่รวมทั้งหมด f_{1+} คือความถี่ของตัวแปรที่ 1 และ f_{+1} คือความถี่ของตัวแปรที่ 2

$$\frac{f_{11}}{N} = \frac{f_{1+}}{N} \times \frac{f_{+1}}{N} \tag{21}$$

$\frac{f_{11}}{N}$ คือความน่าจะเป็นของตัวแปร A และ B ที่เกิดขึ้นร่วมกัน $P(A,B)$ ขณะที่ $\frac{f_{1+}}{N}$ คือความน่าจะเป็นของตัวแปร A และ $\frac{f_{+1}}{N}$ คือความน่าจะเป็นของตัวแปร B ดังนั้นในทางสถิติแล้ว ถ้า A และ B เป็นอิสระต่อกันจะได้ว่า $P(A,B) = P(A) \times P(B)$

จากตารางที่ 12 หากพิจารณาถึงกฎความสัมพันธ์ $\{tea\} \rightarrow \{coffee\}$ แล้วจะได้ว่า

$$corr(\{tea\} \rightarrow \{coffee\}) = \frac{P(tea, coffee)}{P(tea) \times P(coffee)}$$

$$corr(\{tea\} \rightarrow \{coffee\}) = \frac{0.15}{0.2 \times 0.8} = 0.9375$$

ซึ่งพบว่ามีสหสัมพันธ์เชิงลบเล็กน้อยระหว่างนักดื่มชาและดื่มกาแฟด้วย ทำให้กฎความสัมพันธ์ $\{tea\} \rightarrow \{coffee\}$ อาจเป็นกฎที่ไม่น่าสนใจ ถึงแม้ว่ามาตรวัดความน่าสนใจจะทำให้สามารถพิจารณาถึงสหสัมพันธ์เชิงบวกหรือลบรวมทั้งความเป็นอสิระของกฎได้ แต่ในบางสถานะการณ์มาตรวัดปัจจัยความน่าสนใจก็มีข้อด้อยในตัวมันเองได้เช่นกัน พิจารณาตารางการณ์จร $T_1 = [880,50;50,20]$ และ $T_2 = [20,50;50,880]$ เมื่อคำนวณด้วยมาตรวัดความน่าสนใจจะได้ว่า $corr(T_1) = \frac{0.88}{0.93 \times 0.93} = 1.02$ และ $corr(T_2) = \frac{0.02}{0.07 \times 0.07} = 4.08$ ซึ่งจะเห็นว่ากรณีนี้ T_2 มีสหสัมพันธ์เชิงบวกแข็งแกร่งมากกว่า T_1 มาก ถ้าพิจารณาจากมาตรวัดความน่าสนใจก็จะบอกได้ว่า T_2 มีความน่าสนใจกว่า T_1 แต่หากพิจารณาความน่าจะเป็นของกฎความสัมพันธ์บนตาราง T_1 พบว่ากฎความสัมพันธ์นี้มีความน่าจะเป็นสูงถึง 88% แต่ในขณะที่กฎความสัมพันธ์จากตาราง T_2 นั้นมีความน่าจะเป็นเพียง 2% และเมื่อพิจารณาค่าความเชื่อมั่นจะเห็นว่ากฎความสัมพันธ์บนตาราง T_1 นั้นมีค่าความเชื่อมั่นสูงถึง 94.6% ส่วนกฎความสัมพันธ์บนตาราง T_2 มีค่าความเชื่อมั่น 28.6% หากเป็นสถานการณ์ของการพิจารณาคลาสหลักแล้ว มาตรวัดความเชื่อมั่นให้ผลที่มีนัยสำคัญได้ดีกว่ามาตรวัดความน่าสนใจ ดังนั้นการพิจารณาเลือกใช้มาตรวัดความน่าสนใจนั้นจะต้องดูตามสถานการณ์ว่าเหมาะสมกับการกระจายตัวของเซตข้อมูลด้วยหรือไม่ ข้อด้อยของมาตรวัดความน่าสนใจอีกกรณีก็คือการเพิ่มลดขนาดของฐานข้อมูลมีผลโดยตรงกับสหสัมพันธ์ของกฎ พิจารณาจากตัวอย่างเช่น ให้ตารางการณ์จร $T_1 = [100,20;20,10]$ จะได้ว่ากฎความสัมพันธ์ $X \rightarrow Y$ มีสหสัมพันธ์ $corr(T_1) = 1.04$ ซึ่งเกือบจะเป็นอิสระจากกัน ถ้าทำการเพิ่มขนาดข้อมูลใหม่เป็น $T_2 = [100,20;20,200]$ จะได้ว่ากฎความสัมพันธ์ $X \rightarrow Y$ มีสหสัมพันธ์ $corr(T_2) = 2.36$ เห็นได้ว่าข้อมูลที่เพิ่มเข้าไปเป็น T_2 นั้นได้เพิ่มข้อมูลในส่วนของ $\neg X$ และ $\neg Y$ ซึ่งไม่มีความเกี่ยวข้องโดยตรงกับข้อมูลของกฎความสัมพันธ์ $X \rightarrow Y$ แต่ค่าของสหสัมพันธ์เปลี่ยนแปลงไปตามขนาดของข้อมูล (Verhein and Sanjay, 2007)

การจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์ ไม่ได้ถูกออกแบบมาให้รองรับปัญหาความไม่สมดุลจนกระทั้งได้มีการนำเสนอตัวจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์บนฐานข้อมูลที่ไม่สมดุลเป็นครั้งแรก มีชื่อว่า CCCS (Arunasalam and Chawla, 2006) ซึ่งได้นำเสนอมาตรวัดค่าสนับสนุนของคลาสตรงข้าม (Complement Class Support: CCS) ร่วมกับอัลกอริทึมการจำแนกประเภทข้อมูล เพื่อรับมือกับปัญหาความไม่สมดุล โดยมาตรวัดค่าสนับสนุนของคลาสตรงข้ามเป็นการวัดความแข็งแกร่งของกฎความสัมพันธ์กับคลาสตรงข้ามว่ามีความสัมพันธ์แข็งแกร่งเพียงใด ซึ่งกฎความสัมพันธ์ที่ดีที่สุดจะมีค่าสนับสนุนของคลาสตรงข้ามน้อยที่สุด โดยแสดงอยู่ในสมการที่ (22)

$$CCS(X \rightarrow Y) = \frac{\sigma(X \bigcup \neg Y)}{\sigma(\neg Y)} \quad (22)$$

มาตรวัดค่าสนับสนุนของคลาสตรงข้ามนั้นสามารถบ่งบอกถึงความสัมพันธ์ที่มีกับคลาสตรงข้ามได้ดี แต่ในทางกลับกันกลับมีข้อด้อยเรื่องการตอบสนองต่อคลาสของกฎความสัมพันธ์ของตัวมันเอง ตัวอย่างเช่น ให้ตารางการณ์จร $T_1 = [10,1000;10,100]$ จะได้ว่ากฎความสัมพันธ์ $X \rightarrow Y$ มีค่าสนับสนุนของคลาสตรงข้าม $CCS(T_1) = 0.09$ และถ้ากำหนดให้ตารางการณ์จร $T_2 = [1000,10;10,100]$ จะได้ว่ามีค่าสนับสนุนของคลาสตรงข้าม $CCS(T_2) = 0.09$ ซึ่งพบว่าค่าสนับสนุนของคลาสตรงข้ามของทั้งสองตารางมีค่าเท่ากัน เมื่อพิจารณาที่ T_2 จะเห็นว่ากฎความสัมพันธ์ $X \rightarrow Y$ นั้นมีความสัมพันธ์แข็งแกร่งมากคือเกิดขึ้นร่วมกัน 1000 ทรานแซกชั่น เมื่อเทียบกับ T_1 แล้วกฎความสัมพันธ์ $X \rightarrow Y$ นั้นเกิดขึ้นร่วมกันเพียง 10 ทรานแซกชั่นเท่านั้น ซึ่งจะเห็นว่ามาตรวัดค่าสนับสนุนของคลาสตรงข้ามนั้นไม่ตอบสนองต่อความสัมพันธ์ของคลาสของตัวมันเอง

เพื่อแก้ไขปัญหาของมาตรวัดความน่าสนใจและมาตรวัดค่าสนับสนุนของคลาสตรงข้าม Verhein and Chawla (2007) ได้นำเสนอมาตรวัดอัตราส่วนสหสัมพันธ์ของคลาส (Class Correlation Ratio : CCR) ร่วมกับอัลกอริทึมการจำแนกประเภทบนฐานข้อมูลที่ไม่สมดุล โดยมีชื่อว่า SPARCCC ซึ่ง CCR เป็นมาตรวัดที่พิจารณาสหสัมพันธ์ของกฎความสัมพันธ์ $X \rightarrow Y$ เทียบกับสหสัมพันธ์ของกฎความสัมพันธ์ $X \rightarrow \neg Y$ โดยเขียนอยู่ในรูปของสมการได้ดั่งที่แสดงในสมการที่ (23)

$$CCR(X \rightarrow Y) = \frac{corr(X \rightarrow Y)}{corr(X \rightarrow \neg Y)} \quad (23)$$

มาตรวัดอัตราส่วนสหสัมพันธ์ของคลาส นั้นคือการวัดสหสัมพันธ์เชิงบวกของไอเท็มเซตด้านซ้ายว่ามีสหสัมพันธ์หรือเกี่ยวข้องกับคลาสที่มันทำนายหรือไม่ พิจารณาจากตัวอย่างบนตารางการณ์จรเดิม $T_1 = [100,20;20,10]$ เมื่อคำนวณด้วยมาตรวัดอัตราส่วนสหสัมพันธ์ของคลาสจะได้ $CCR(T_1) = 1.25$ และตารางการณ์จร $T_2 = [100,20;20,200]$ มีค่า $CCR(T_2) = 9.17$ ดังนั้นจะบอกได้ว่ากฎความสัมพันธ์ $X \rightarrow \neg Y$ บนตาราง T_2 นั้นดีกว่า T_1 เพราะว่ามีสหสัมพันธ์กับคลาสที่มันทำนายมากกว่าเมื่อเทียบกับคลาสอื่นๆ แล้ว ซึ่งในการพิจารณากฎความสัมพันธ์ด้วยมาตรวัดอัตราส่วนสหสัมพันธ์ของคลาสจะเลือกพิจารณาเฉพาะกฎที่มีค่า $CCR(X \rightarrow Y) > 1$ เท่านั้น

แม้ว่ามาตรวัดอัตราส่วนสหสัมพันธ์ของคลาสนั้นจะมีความน่าสนใจแต่ค่าของ *CCR* ที่สูงกว่ากฎอื่นๆ ก็ไม่ได้รับประกันว่าจะเกิดความผิดพลาดประเภทลบ (FN) น้อยกว่ากฎความสัมพันธ์อื่นๆ ที่ให้ค่าต่ำกว่า ตัวอย่างเช่นตารางการณ์จร $T_1 = [10,10;100,1000]$ แทนความถี่ของกฎความสัมพันธ์ $R_1 = X_1 \to Y$ ที่นำไปใช้จำแนกประเภทข้อมูล เมื่อคำนวณค่าตามมาตรวัดอัตราส่วนสหสัมพันธ์ของคลาสจะได้ว่า $CCR(T_1) = 5.5$ และตารางการณ์จร $T_2 = [10,30;80,1000]$ แทนความถี่ของกฎความสัมพันธ์ $R_2 = X_2 \to Y$ ที่นำไปใช้จำแนกประเภทข้อมูล มีค่า $CCR(T_2) = 3.375$ ซึ่งจากกฎความสัมพันธ์ R_1 และ R_2 จะเห็นว่ากฎความสัมพันธ์ R_1 นั้นมีโอกาสทำนายเป็นคลาส $Y = \frac{10}{1120} \times 100 = 0.89\%$ เท่ากับกฎความสัมพันธ์ R_2 แต่กฎความสัมพันธ์ R_1 นั้นมีโอกาสทำนายเป็นคลาสอื่นๆ $\neg Y = \frac{100}{1120} \times 100 = 8.92\%$ และกฎความสัมพันธ์ R_2 มีโอกาสทำนายเป็นคลาสอื่นๆ $\neg Y = \frac{80}{1120} \times 100 = 7.14\%$ ซึ่งเห็นได้ว่ากฎความสัมพันธ์ R_2 มีโอกาสเกิดความผิดพลาดประเภทลบน้อยกว่ากฎความสัมพันธ์ R_1 แต่มีค่า *CCR* ที่น้อยกว่าเมื่อถูกนำไปใช้สำหรับการจำแนกประเภท โดยแนวทางแก้ไขจะนำเสนอในวิทยานิพนธ์เล่มนี้ในลำดับถัดไป

2. ปัญหาการพิจารณาคุณภาพของกฎความสัมพันธ์

ปัญหาอย่างหนึ่งของการสืบค้นกฎความสัมพันธ์ก็คือมีจำนวนกฎมากมายถูกสร้างขึ้นโดยถูกเลือกจากมาตรวัดต่างๆ ที่ได้นำเสนอมาข้างต้น แม้ว่ากฎที่ถูกเลือกมาแล้วนั้นล้วนมีคุณภาพตามมาตรวัดที่เลือกใช้ แต่ก็ยังพบว่าจำนวนกฎเหล่านั้นประกอบไปด้วยกฎที่เป็นประโยชน์ (Productive Rules) และกฎที่ไม่เป็นประโยชน์ (Unproductive Rules) ซึ่งจะส่งผลกับขั้นตอนการจำแนกประเภท เช่น กฎ $\{pregnant,female\} \to \{oedema\}$ และกฎ $\{pregnant,dataminer\} \to \{oedema\}$ จะเห็นว่ากฎ $\{pregnant,female\} \to \{oedema\}$ เป็นกฎที่เป็นประโยชน์เพราะว่าการเกิดโรคอาการบวมน้ำนั้น ไม่ได้เกี่ยวข้องกับการประกอบอาชีพเป็นนักดาต้าไมนิ่งดังที่ปรากฏอยู่ในกฎความสัมพันธ์ $\{pregnant,dataminer\} \to \{oedema\}$ ดังนั้นจึงถือว่ากฎความสัมพันธ์นี้เป็นกฎที่ไม่เป็นประโยชน์สมควรที่ถูกคัดเลือกออก

จากปัญหาดังกล่าวข้างต้น Webb (2006) ได้นำเสนอการทดสอบสมมุติฐานทางสถิติด้วยการทดสอบความถูกต้องของฟิชเชอร์ (FET) มาทำการคัดเลือกเพื่อให้ได้กฎความสัมพันธ์ที่เป็นประโยชน์โดยทดสอบสมมุติฐานดังนี้ H_0 เป็นสมมุติฐานหลักที่แสดงว่ากฎความสัมพันธ์ $X \to Y$ และกฎความสัมพันธ์ $X \setminus \{Z\} \to Y$ เป็นอิสระจากกัน เมื่อ $Z \in X$ ซึ่งเป็นการพิจารณา

ว่ากฎความสัมพันธ์ $X \rightarrow Y$ ขัดแย้งกับลักษณะโดยทั่วไป (Immediate Generalization) ของกฎ $X \setminus \{Z\} \rightarrow Y$ หรือไม่ โดยพิจารณาจากทุกๆ สมาชิกของไอเท็มเซต X ด้วยวิธีการทดสอบความถูกต้องของฟิชเชอร์โดยกำหนดให้กฎความสัมพันธ์ $X \rightarrow Y$ แบ่งกลุ่มข้อมูลได้เป็นกลุ่มที่ 1 และกฎความสัมพันธ์ $X \setminus \{Z\} \rightarrow Y$ แบ่งกลุ่มข้อมูลได้เป็นกลุ่มที่ 2 ซึ่งทั้งสองกลุ่มนั้นสามารถแยกประเภทข้อมูลได้ 2 ประเภทคือคลาส Y และ $\neg Y$ จากนั้นก็จะคำนวณค่าความน่าจะเป็น p_{value} ของทุกๆ กฎความสัมพันธ์ $X \setminus \{Z\} \rightarrow Y$ เมื่อ $Z \in X$ โดยหยิบไอเท็มเซต Z ออกครั้งละ 1 ไอเท็มสลับไปเรื่อยๆ จนครบทุกสมาชิกในไอเท็มเซต X แล้วจึงพิจารณาดูว่าการทดสอบด้วย FET แต่ละกฎนั้น มีกฎใดกฎหนึ่งปฏิเสธสมมุติฐานหลัก H_0 เกิดขึ้นหรือไม่ ($p_{value} \leq \alpha$) ถ้ามีแสดงว่ากฎความสัมพันธ์ $X \rightarrow Y$ เป็นกฎที่เป็นประโยชน์ไม่สามารถตัดทิ้งได้ เพราะว่าไม่ได้เป็นอิสระจากกฎความสัมพันธ์ $X \setminus \{Z\} \rightarrow Y$ ในทางกลับกันหากพบว่าการทดสอบนั้นยอมรับในสมมุติฐานหลักทั้งหมดแสดงว่า $X \rightarrow Y$ เป็นอิสระจากกฎ $X \setminus \{Z\} \rightarrow Y$ ในทุกๆ สมาชิกของ X ซึ่งหมายถึงว่า กฎ $X \rightarrow Y$ เป็นกฎที่ไม่เป็นประโยชน์ สามารถใช้กฎความสัมพันธ์ $X \setminus \{Z\} \rightarrow Y$ ทำนายแทนได้ตัวอย่างเช่น กฎ $\{pregnant,dataminer\} \rightarrow \{oedema\}$ และกฎ $\{pregnant\} \rightarrow \{oedema\}$ สามารถจำแนกข้อมูลได้เหมือนกัน ดังนั้นกฎความสัมพันธ์ $\{pregnant,dataminer\} \rightarrow \{oedema\}$ จึงตัดทิ้งได้ส่วนกฎ $\{pregnant\} \rightarrow \{oedema\}$ ถือเป็นกฎที่มีนัยสำคัญทางสถิติ

การทดสอบนัยสำคัญทางสถิติของกฎความสัมพันธ์ซึ่งถูกนำเสนอโดย Webb (2006) สามารถคำนวณได้ด้วยสมการที่ (24) ซึ่งกำหนดให้

$$a = \sigma(X \rightarrow Y)$$
$$b = \sigma(X \setminus \{Z\} \rightarrow Y) - \sigma(X \rightarrow Y)$$
$$c = \sigma(X \rightarrow \neg Y)$$
$$d = \sigma(X \setminus \{Z\} \rightarrow \neg Y) - \sigma(X \rightarrow \neg Y)$$

$$p = \sum_{i=0}^{\min(b,c)} \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{(a+b+c+d)!(a+i)!(b-i)!(c-i)!(d+i)!} \qquad (24)$$

เมื่อ $n!$ แสดงถึงแฟคทอเรียลของ n

วิธีการพิจารณานัยสำคัญทางสถิติเพื่อคัดเลือกกฎความสัมพันธ์ของ Webb (2006) นั้นถือได้ว่ามีประโยชน์อย่างมากในการพิจารณากฎความสัมพันธ์บนฐานข้อมูลที่ไม่สมดุล เพราะกฎ

ความสัมพันธ์ที่ได้นั้นล้วนแล้วแต่เป็นกฎที่มีคุณภาพและจำนวนกฎที่ถูกคัดเลือกมาแล้วนั้นมีจำนวนไม่มากเกินไปที่จะนำไปสร้างตัวจำแนกประเภทข้อมูล จากผลการวิจัยของ Weeb (2006) ได้ถูกนำไปใช้สร้างตัวจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์เพื่อแก้ไขปัญหาความไม่สมดุลของข้อมูล ชื่อว่า SPARCCC นำเสนอโดย Verhein and Chawla (2007) โดยนำเสนอพร้อมมาตรวัด *CCR* เพื่อพิจารณาปัญหาความไม่สมดุลของคลาสโดยเฉพาะ และในวิทยานิพนธ์เล่มนี้ได้อาศัยการพิจารณาคุณภาพของกฎความสัมพันธ์ด้วยวิธีของ Webb เช่นเดียวกัน

3. สรุปข้อดีข้อเสียของงานวิจัยก่อนหน้าในปัญหาของความไม่สมดุล

3.1. มาตรวัดค่าสนับสนุนและค่าความเชื่อมั่น

3.1.1. ข้อดีของมาตรวัดค่าสนับสนุนและค่าความเชื่อมั่น ง่ายต่อความเข้าใจ มีการคำนวณไม่ซับซ้อน

3.1.2. ข้อเสียของมาตรวัดค่าสนับสนุนและค่าความเชื่อมั่น ทำให้กฎที่มีศักยภาพน่าสนใจแต่มีค่าสนับสนุนต่ำ ถูกกำจัดทิ้งไป

3.1.3. การเลือกกฎความสัมพันธ์ของ CBA ตั้งอยู่บนค่าความเชื่อมั่นและค่าสนับสนุน ทำให้กฎที่มีอยู่ส่วนใหญ่แล้วเป็นกฎที่สอดคล้องกับคลาสหลักทำให้การทำนายแอนเอียงไปทางคลาสหลักมากกว่า

3.2. มาตรวัดความน่าสนใจ

3.2.1. ข้อดีของมาตรวัดความน่าสนใจ คือการพิจารณาความเกี่ยวข้องหรือสหสัมพันธ์ของข้อมูลเป็นหลัก ทำให้เชื่อได้ว่าความสัมพันธ์ของกฎหากมีสหสัมพันธ์เชิงบวกแล้วเป็นข้อมูลที่สอดคล้องกันจริง

3.2.2. ปัญหาของคลาสที่เบาบางทำให้มีค่าสนับสนุนและค่าความเชื่อมั่นต่ำแต่มาตรวัดความน่าสนใจยังคงให้ผลลัพธ์ที่ดีได้ ซึ่งเหมาะแก่ความสัมพันธ์ของข้อมูลที่เบาบาง

3.2.3. ข้อเสียคือการเพิ่มขึ้นของข้อมูลที่ไม่เกี่ยวข้องกับความสัมพันธ์ของข้อมูลมีผลให้ค่าสหสัมพันธ์เปลี่ยนแปลงไป

3.3. มาตรวัดค่าสนับสนุนของคลาสตรงข้าม

3.3.1. ข้อดีของมาตรวัดค่าสนับสนุนของคลาสตรงข้ามคือทำให้ทราบถึงความสัมพันธ์ของกฎความสัมพันธ์ที่มีต่อคลาสตรงข้ามว่ามีความแข็งแกร่งเพียงใด

3.3.2. ข้อเสียคือไม่ตอบสนองต่อคลาสของกฎความสัมพันธ์ของตัวมันเอง

3.4. มาตรวัดอัตราส่วนสหสัมพันธ์ของคลาส

3.4.1. ข้อดีของมาตรวัดความน่าสนใจ พิจารณาความเกี่ยวข้องหรือสหสัมพันธ์ของข้อมูลเป็นหลัก ทำให้เชื่อได้ว่าความสัมพันธ์ของกฎหากมีสหสัมพันธ์เชิงบวกแล้วเป็นข้อมูลที่สอดคล้องกันจริง

3.4.2. ข้อดีมีความสัมพันธ์สอดคล้องกับคลาสที่ทำนาย

3.4.3. ไม่แตกต่างจากมาตรวัดความน่าสนใจในเทอมของขนาดข้อมูลที่เพิ่มขึ้นหรือลดลง

3.4.4. ค่า *CCR* ที่มากกว่าอาจจะเกิดความผิดพลาดประเภทลบได้มากกว่ากฎที่มีค่า *CCR* น้อยกว่า

3.5. การทดสอบคุณภาพกฎความสัมพันธ์ด้วย FET

3.5.1. ข้อดีคือไม่มีกฎที่ไม่เป็นประโยชน์หรือซ้ำซ้อน ทำให้จำนวนกฎลดลง

3.5.2. กฎที่ได้มามีคุณภาพอย่างมีนัยสำคัญทางสถิติ

3.5.3. ข้อเสียคือการคำนวณมีความซับซ้อน

3.5.4. การทดสอบต้องใช้เวลานานเพราะต้องทดสอบทุกๆ สมาชิกในไอเท็มเซตด้านซ้าย

4. แนวทางการแก้ไขปัญหาข้อมูลที่ไม่สมดุล

จากปัญหาที่ผ่านมาจะเห็นได้ว่าการจำแนกข้อมูลที่ไม่สมดุลนั้นอาจจะต้องอาศัยหลากหลายวิธีการเพื่อรับมือกับความเบาบางของข้อมูลที่เราให้ความสนใจเป็นพิเศษ โดยผู้วิจัยมุ่งเน้นการลดข้อผิดพลาดประเภทลบ (FN) เพราะมีต้นทุนการทำนายผิดพลาดที่สูงกว่าความผิดพลาดประเภทบวก (FP) โดยในวิทยานิพนธ์เล่มนี้จะนำเสนอแนวทางที่ใช้ในการแก้ไขปัญหาความไม่สมดุลของคลาส ซึ่งแบ่งตามขั้นตอนการทำงานดังนี้

4.1. ขั้นตอนการสร้างกฎความสัมพันธ์

การสร้างกฎความสัมพันธ์นั้นเราจำเป็นที่จะต้องสร้างกฎความสัมพันธ์จากทุกความเป็นไปได้ เพื่อหลีกเลี่ยงการสูญเสียกฎความสัมพันธ์ที่มีศักย์ภาพอันเนื่องมาจากความสัมพันธ์ของกฎเกิดขึ้นอย่างเบาบาง ดังนั้นจึงจำเป็นที่จะต้องอาศัยอัลกอริทึมที่มีประสิทธิภาพมาช่วยในการสร้างกฎจำนวนมหาศาลให้ได้อย่างรวดเร็ว

4.2. ขั้นตอนการคัดเลือกและกำจัดกฎความสัมพันธ์ทิ้ง

จากกฎความสัมพันธ์ในขั้นตอนแรกที่ได้มาจะมีจำนวนมากมายเพื่อคัดเลือกกฎให้มีคุณภาพผู้วิจัยได้นำแนวทางการทดสอบสมมุติฐานทางสถิติด้วยการทดสอบความถูกต้องของฟิชเชอร์ (FET) ที่ Webb ได้เสนอมาก่อนหน้านี้ ซึ่งเป็นวิธีที่เหมาะกับการคัดเลือกกฎความสัมพันธ์ให้ได้กฎที่มีนัยสำคัญบนชุดข้อมูลที่ไม่สมดุล มาใช้เพื่อคัดเลือกเฉพาะกฎที่มีนัยสำคัญทางสถิติเท่านั้น

4.3. ขั้นตอนลำดับความสำคัญของกฎ

เพื่อหลีกเลี่ยงความผิดพลาดประเภทลบให้ได้มากที่สุด ผู้วิจัยจึงนำเสนอแนวคิดเรื่องต้นทุนการทำนายผิดพลาดของกฎความสัมพันธ์ (Cost-Sensitive Rules) ซึ่งเป็นการคำนวณต้นทุนความเสี่ยงในการทำนายผิดพลาดของกฎความสัมพันธ์ หากหยิบไปใช้ในการทำนายข้อมูลใหม่ โดยเรียงลำดับกฎจากความเสี่ยงน้อยไปหาความเสี่ยงมาก

4.4. ขั้นตอนการจำแนกประเภทข้อมูล

การทำนายข้อมูลใหม่อาศัยการพิจารณาเรื่องต้นทุนความเสี่ยงของกฎความสัมพันธ์มาใช้เลือกกฎความสัมพันธ์ที่เข้ากันได้กับข้อมูลใหม่ (Unseen Data Matching) ด้วยความเสี่ยงที่น้อยที่สุด และหากไม่มีกฎใดๆ เลยเข้ากับข้อมูลใหม่ได้ ก็จะใช้วิธีทำนายจากการอำพลางข้อมูล (Attribute Value Hiding) ซึ่งจะนำเสนอรายละเอียดทั้งหมดในหัวข้อถัดไป

อุปกรณ์และวิธีการ

อุปกรณ์

1. ฮาร์ดแวร์

1.1. เครื่องคอมพิวเตอร์โน๊ตบุค 1 เครื่อง ประกอบด้วยอุปกรณ์ดังต่อไปนี้

1.1.1. ซีพียู Intel Core 2 Duo P8400 (2.26GHz)
1.1.2. หน่วยความจำหลัก 4GB
1.1.3. ฮาร์ดดิสก์ขนาด 120GB

2. ซอฟต์แวร์

2.1. ระบบปฏิบัติการ Microsoft Windows 7 Ultimate Edition Service Pack 1

2.1.1. คอมไพเลอร์ภาษา Java
2.1.2. Weka Libraries
2.1.3. โปรแกรม Eclipse

วิธีการ

การสร้างตัวจำแนกประเภทข้อมูลบนฐานข้อมูลที่ไม่สมดุลในงานวิจัยนี้ใช้ชื่อว่า Statistically Significant Cost-sensitive Rules for Associative Classification in Imbalanced Datasets หรือ SSCR ซึ่งประกอบด้วยขั้นตอนสำคัญหลัก 3 ขั้นตอนด้วยกัน ประกอบด้วย

1. ขั้นตอนการสืบค้นและตัดกฎความสัมพันธ์ (Searching and Prunning Rules)

1.1. ขั้นตอนการสืบค้นกฎความสัมพันธ์นั้นเราจำเป็นที่จะต้องสร้างกฎความสัมพันธ์จากทุกความเป็นไปได้เพื่อหลีกเลี่ยงการสูญเสียกฎความสัมพันธ์ที่มีศักย์ภาพอันเนื่องมาจากความสัมพันธ์ของกฎเกิดขึ้นอย่างเบาบาง ดังนั้นในวิทยานิพนธ์เล่มนี้จึงได้เลือกอัลกอริทึม Eclat

(Zaki *et al*., 1997) เนื่องจากเป็นอัลกอริทึมที่มีประสิทธิภาพทั้งความเร็วและมีโครงสร้างข้อมูลที่เอื้อต่อการคำนวณค่าต่างๆ ที่ใช้ในวิทยานิพนธ์เล่มนี้ โดยผู้วิจัยได้ทำการปรับปรุงอัลกอริทึม Eclat ให้เหมาะสมกับงานวิจัยโดยมีรายละเอียดดังนี้

1.2. ขั้นตอนการสร้างต้นไม้พรีฟิกในขั้นตอนนี้ไม่ได้ปรับปรุงอะไรเพียงแต่กำหนดค่าสนับสนุนขั้นต่ำเป็น 1 ซึ่งก็คือ กฎความสัมพันธ์ที่จะสร้างทุกความเป็นไปได้จะต้องเกิดจากการมีข้อมูลอย่างน้อย 1 ตัว เนื่องจากในที่สุดแล้วการทดสอบด้วย FET ก็จะกำจัดกฎที่มีไอเท็มซึ่งไม่มีค่าสนับสนุนออกไป

1.3. ขั้นตอนการสร้างไอเท็มเซต ขณะที่แวะผ่านลงไปในแต่ละโหนดจะพิจารณาเฉพาะไอเท็มเซตที่สร้างเป็น CARs ได้เท่านั้น ส่วนไอเท็มเซตที่ไม่ใช่ CARs จะถูกตัดทิ้งไป เมื่อพิจารณาแล้วว่าไอเท็มเซตนี้เป็น CARs ก็จะทำการตรวจสอบคุณสมบัติพื้นฐานของกฎความสัมพันธ์ด้วยมาตรวัดความน่าสนใจ (สมการที่ 6)

1.3.1. พิจารณา $corr(X \rightarrow Y) > 1$ ซึ่งในขั้นตอนนี้จะเลือกเฉพาะกฎที่มีสหสัมพันธ์เชิงบวกเท่านั้น การเลือกมาตรวัดความน่าสนใจมาใช้ตรวจสอบตรวจสอบคุณสมบัติพื้นฐานของกฎเพราะว่าต้องการหลีกเลี่ยงการคำนวณ FET ที่ไม่จำเป็นของกฎที่มีสหสัมพันธ์เชิงลบและกฎที่เป็นอิสระ กฎความสัมพันธ์ที่ผ่านคุณสมบัติเบื้องต้นจะถูกทดสอบ FET ตามวิธีการของ Webb (2006)

1.3.2. ทดสอบกฎความสัมพันธ์ด้วย FET พิจารณาเฉพาะกฎที่มี $p_{value} \leq \alpha$ เพื่อคัดเลือกเฉพาะกฎที่มีนัยสำคัญทางสถิติเท่านั้น ซึ่งถูกกำหนดเกณฑ์การยอมรับผ่านค่าวิกฤษ α โดยการทดสอบความเป็นอิสระของกฎความสัมพันธ์ $X \rightarrow Y$ และ $X \setminus \{Z\} \rightarrow Y$ เมื่อ $Z \in X$ จะแบ่งออกเป็น 2 กรณี คือ เมื่อกฎความสัมพันธ์ด้านซ้ายมีขนาดไอเท็มเซตเท่ากับ 1-itemset จะคำนวณผ่านตารางการณ์จรที่แสดงไว้ในภาพที่ 19 หากขนาดของไอเท็มเซตมีค่ามากกว่า 1-itemset จะคำนวณผ่านตารางการณ์จรที่แสดงไว้ในภาพที่ 20 ซึ่งทั้งสองตารางจะถูกแทนค่าลงในสมการที่ (24) เพื่อคำนวณหาค่า p_{value} และจะพิจารณาค่า p_{value} ที่น้อยที่สุดว่า $p_{value} \leq \alpha$ หรือไม่ ถ้า $p_{value} \leq \alpha$ แสดงถึงสมมุติฐานหลักจะถูกปฏิเสธ ซึ่งหมายถึงกฎ $X \rightarrow Y$ และกฎ $X \setminus \{Z\} \rightarrow Y$ แบ่งกลุ่มข้อมูลออกเป็นคลาส Y และ $\neg Y$ ได้แตกต่างกัน ดังนั้นจะเห็นว่ากฎ $X \rightarrow Y$ เป็นกฎที่เป็นประโยชน์ไม่สามารถตัดทิ้งได้ แต่ถ้ากฎใดมีค่า $p_{value} > \alpha$ แสดงถึงสมมุติฐานหลักถูกยอมรับ

นั้นหมายถึงกฎ $X \rightarrow Y$ และกฎ $X \setminus \{Z\} \rightarrow Y$ แบ่งกลุ่มข้อมูลออกเป็นคลาส Y และ $\neg Y$ ได้เหมือนกัน ดังนั้นกฎ $X \rightarrow Y$ เป็นกฎที่ไม่เป็นประโยชน์สามารถตัดทิ้งได้

จากภาพที่ 19 และ 20 จะเห็นว่ามีการคำนวณค่าสนับสนุน σ ของตัวแปร a, b, c และ d ซึ่งเกิดจากการนับบิทที่เป็น 1 ในบิทเวคเตอร์ของแต่ละโหนดมาเป็นค่าสนับสนุนของกฎความสัมพันธ์ หลังจากที่พิจารณาว่าเป็นกฎที่มีนัยสำคัญแล้ว ก็จะคำนวณความเสี่ยงของกฎ (Cost-Sensitive Rule) เก็บเอาไว้

	$i: X \subset t_i$	$i: X\setminus\{Z\} \subset t_i \wedge Z \notin t_i$	$i: X\setminus\{Z\} \subset t_i$
$i: Y \in t_i$	$a = \sigma(X \rightarrow Y)$	$b = \sigma(Y) - \sigma(X \rightarrow Y)$	$a + b = \sigma(Y)$
$i: \neg Y \in t_i$	$c = \sigma(X \rightarrow \neg Y)$	$d = \sigma(\neg Y) - \sigma(X \rightarrow \neg Y)$	$c + d = \sigma(\neg Y)$
	$a + c = \sigma(X)$	$b + d = \sigma(\neg X)$	$a + b + c + d = \sigma(Y) + \sigma(\neg Y)$

ภาพที่ 19 ตารางการณ์จร [a,b;c,d] สำหรับทดสอบนัยสำคัญทางสถิติของกฎความสัมพันธ์ $X \rightarrow Y$ เมื่อ $|X| = 1$

	$i: X \subset t_i$	$i: X\setminus\{Z\} \subset t_i \wedge Z \notin t_i$	$i: X\setminus\{Z\} \subset t_i$
$i: Y \in t_i$	$a = \sigma(X \rightarrow Y)$	$b = \sigma(X\setminus\{Z\} \rightarrow Y) - \sigma(X \rightarrow Y)$	$a + b = \sigma(X\setminus\{Z\} \rightarrow Y)$
$i: \neg Y \in t_i$	$c = \sigma(X \rightarrow \neg Y)$	$d = \sigma(X\setminus\{Z\} \rightarrow \neg Y) - \sigma(X \rightarrow \neg Y)$	$c + d = \sigma(X\setminus\{Z\} \rightarrow \neg Y)$
	$a + c = \sigma(X)$	$b + d = \sigma(X\setminus\{Z\}) - \sigma(X)$	$a + b + c + d = \sigma(X\setminus\{Z\})$

ภาพที่ 20 ตารางการณ์จร [a,b;c,d] สำหรับทดสอบนัยสำคัญทางสถิติของกฎความสัมพันธ์ $X \rightarrow Y$ เมื่อ $|X| > 1$

1.3.3. การคำนวณความเสี่ยงของกฎ ก็เพื่อรับมือกับปัญหาความไม่สมดุลของคลาส การเรียนรู้แบบมีต้นทุนถูกนำมาใช้เพื่อช่วยเพิ่มประสิทธิภาพในการให้ความสำคัญกับคลาสที่เบาบางของคลาสรอง โดยอาศัยตารางเมตริกต้นทุนที่แสดงไว้ในตารางที่ 9 [-1,100;1,0] มาใช้เป็นค่าปรับเมื่อจำแนกข้อมูลผิดพลาด โดยผู้วิจัยได้นำการเรียนรู้แบบมีต้นทุนมาใช้เพื่อลำดับความสำคัญของกฎและประเมินค่าความเสี่ยงของกฎให้ออกมาในรูปแบบของต้นทุนในแต่ละกฎ (Rule Cost) เพื่อให้ทราบถึงความเสี่ยงของกฎเมื่อเราหยิบนำไปใช้เพื่อจำแนกประเภทตัวอย่างข้อมูลชุดทดสอบ ซึ่งได้มาจากการวิเคราะห์หาค่าความเสี่ยงของกฎ $X \rightarrow Y$ จากชุดข้อมูลสำหรับเรียนรู้นำมาคำนวณต้นทุนด้วยสมการที่ (25) โดยค่าความเสี่ยงของกฎ $C_R(X \rightarrow Y)$ ที่มีค่าน้อยที่สุดคือกฎที่มีลำดับความสำคัญสูงสุด

$$
\begin{aligned}
C_R(X \rightarrow Y) \quad &= TP_R \times C(+,+) \\
&+ FP_R \times C(-,+) \\
&+ FN_R \times C(+,-) \\
&+ TN_R \times C(-,-)
\end{aligned} \tag{25}
$$

จากโครงสร้างข้อมูลของต้นไม้พรีฟิกที่จัดเก็บบิทเวคเตอร์ ทำให้สามารถคำนวณค่าได้โดยง่าย โดยแต่ละค่าคำนวณได้ดังนี้ $TP_R = \sigma(X \cap Y)$, $FP_R = \sigma(\neg X \cap Y)$, $FN_R = \sigma(X \cap \neg Y)$ และ $TN_R = \sigma(\neg X \cap \neg Y)$ ภาพที่ 21 แสดงรหัสเทียมการคำนวณต้นทุนความเสี่ยงของกฎความสัมพันธ์

```
1.  Let r denote the rule
2.  Let N denote the node of prefix trees
3.  Let M denote the cost matrix
4.  calculateRuleCost(r, N, M)
5.  X = r.X.getBitVector(N)
6.  Y = r.Y.getBitVector(N)
7.  if (r.Y ∈ PositiveClass) then
8.     TP = σ(X ∩ Y), FP = σ(X ∩ ¬Y), FN = 0, TN = 0
9.  else
10.    TN = σ(X ∩ Y), FN = σ(X ∩ ¬Y), FP = 0, TP = 0
11. end if
12. r.cost=(TP × M[+,+]) + (FP × M[−,+]) + (FN × M[+,−]) + (TN × M[−,−])
```

ภาพที่ 21 รหัสเทียมการคำนวณต้นทุนความเสี่ยงของกฎความสัมพันธ์

ตัวอย่างการคิดต้นทุนของกฎความสัมพันธ์ ขอยกตัวอย่างเดิมของตารางการณ์จร $T_1 = [10,10;100,1000]$ เมื่อนำมาคำนวณต้นทุนความเสี่ยงของกฎ $X \rightarrow Y$ ที่ได้ผลลัพธ์เป็น T_1 ด้วยตารางเมตริกต้นทุนที่แสดงไว้ในตารางที่ 9 [-1,1;100,0] จะมีค่าความเสียงของกฎ $C_R(T_1) = (10 \times -1) + (100 \times 1) + (0 \times 100) + (0 \times 0) = 90$ สาเหตุที่เทอมของ FN_R และ TN_R เป็น 0 เพราะว่าในการเลือกกฎความสัมพันธ์ไปทำนายนั้น จะทำการตรวจสอบไอเท็มเซตด้านซ้ายหรือ X ว่าเข้ากันได้กับข้อมูลใหม่ (Unseen Data) หรือไม่ ซึ่งถ้าเข้ากันได้โอกาสตอบด้วยกฎความสัมพันธ์นี้มีแค่ตอบเป็นคลาส Y เท่านั้นซึ่งถ้าคลาสจริงคือ $\neg Y$ แล้วกฎนี้ตอบ Y ก็คือเป็นความผิดพลาดประเภทบวก FP_R ทั้งนี้ต้องพิจารณาจากคลาสคำตอบของ Y ด้วยเช่นกันว่าเป็นคลาสบวกหรือคลาสลบ ถ้าหากเป็นคลาสลบเทอมของ FP_R และ TP_R จะเป็น 0 แทน จากกรณีของตัวอย่างเพื่อให้ง่ายต่อการอธิบายให้เห็นภาพโดยจะสมมุติให้คลาสคำตอบคือคลาสบวกทั้งหมด

ส่วนกฎ $X \rightarrow Y$ ที่ได้ผลลัพธ์เป็น $T_2 = [10,30;80,1000]$ จะมีค่าความเสี่ยงของกฎ $C_R(T_2) = (10 \times -1) + (80 \times 1) + (0 \times 100) + (0 \times 0) = 70$ เมื่อเปรียบเทียบกับมาตรวัด CCR แล้วจะได้ว่า $C_R(T_1) = 90$ ขณะที่ $CCR(T_1) = 5.5$ และ $C_R(T_2) = 70$ มีค่า $CCR(T_2) = 3.375$ ซึ่งเห็นได้ว่า $C_R(T_2)$ มีความเสี่ยงต่อการเกิดความผิดพลาดน้อยกว่า $C_R(T_1)$ ในขณะที่มาตรวัด CCR นั้นให้ค่า $CCR(T_1)$ มีผลลัพธ์ที่ดีกว่า $CCR(T_2)$

จากคำอธิบายข้างต้นสามารถเขียนเป็นรหัสเทียมของอัลกอริทึมสำหรับสืบค้นและตัดกฎความสัมพันธ์ได้ดังภาพที่ 22

```
1.  Let D denote the dataset
2.  Let α denote the significant level
3.  Let M denote the cost matrix
4.  Let P denote the prefix trees
5.  Let R denote the set of rules
6.  generateRules(D, α , M)
7.  P=buildPrefixTrees(D)
8.  R = φ
9.  while not visit (P.node)        // Use depth-first search approach
10.   r=createRule(P.node)          // Create association rule from path of current node
11.   if r is CARs then             // Select class association rules only
12.     if corr(r) > 1 then         // Positively correlated & Independence
13.       r.pvalue = FET(r)         // Statically Significant Test
14.       if r.pvalue ≤ α then
15.         calculateRuleCost(r, P.node, M)  // Evaluate risks of rule
16.         R = R ∪ r
17.       end if
18.     end if
19.   end if
20. end while
21. return R
```

ภาพที่ 22 รหัสเทียมของอัลกอริทึมสำหรับสืบค้นและตัดกฎความสัมพันธ์

2. ขั้นตอนการจัดลำดับความสำคัญของกฎ (Rule Ranking)

จากกฎความสัมพันธ์ที่ได้มาจากขั้นตอนสืบค้นและกำจัดกฎทิ้ง เราจะได้กฎความสัมพันธ์ที่มีนัยสำคัญทางสถิติซึ่งถือเป็นกฎที่มีคุณภาพสูง รวมทั้งการคำนวณเพื่อหาต้นทุนการทำนายผิดพลาดจากกฎความสัมพันธ์จะช่วยลดโอกาสการเกิดข้อผิดพลาดประเภทลบ (FN) ดังนั้นด้วยแนวคิดของกฎที่มีต้นทุนความเสี่ยงนี้ ผู้วิจัยจึงลำดับความสำคัญของกฎด้วยขนาดของกฎ

ความสัมพันธ์ด้านซ้าย $|X|$ จากขนาดที่ยาวกว่าไปหาขนาดสั้น ซึ่ง X เป็นไอเท็มเซตที่จะถูกตรวจสอบความเข้ากันได้กับข้อมูลใหม่ในขั้นตอนการจำแนกประเภท ด้วยคุณสมบัติของกฎความสัมพันธ์ กฎที่มีไอเท็มเซทร่วมกันแต่มีขนาดยาวกว่าจะมีค่าสนับสนุนของกฎเท่ากับหรือน้อยกว่ากฎที่มีขนาดสั้นกว่า ดังนั้นในการจำแนกประเภทจึงควรพิจารณากฎที่มีขนาดยาวกว่าก่อน เพราะจะเข้ากับข้อมูลได้ดีกว่า (Fit) กฎความสัมพันธ์ใดๆ ที่มี $|X|$ เท่ากันก็จะถูกลำดับด้วยค่าต้นทุนความเสี่ยงของกฎจากน้อยไปหามาก โดยแสดงตัวอย่างการลำดับความสำคัญของกฎความสัมพันธ์ไว้ในภาพที่ 23

Rule	Class	Cost
$\{a, b, c\}$	$C1$	-10
{a, c, d}	$C2$	2
{b, c}	$C1$	-5
$\{a\}$	$C2$	0

ภาพที่ 23 ตัวอย่างการลำดับความสำคัญของกฎความสัมพันธ์

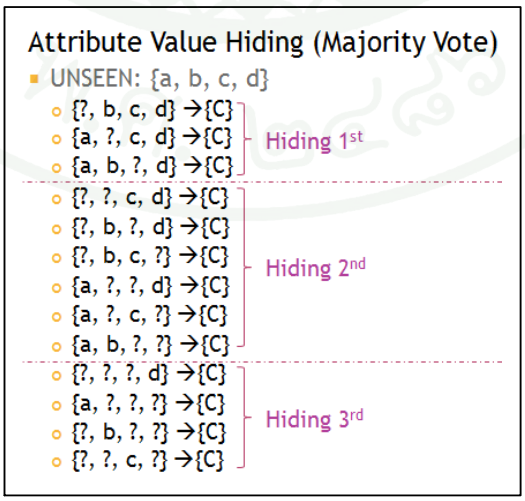
3. ขั้นตอนการจำแนกประเภท (Classification)

3.1. การจำแนกประเภทจากข้อมูลชุดทดสอบ อัลกอริทึมของตัวจำแนกประเภทอาศัยการตัดสินใจอยู่บนกฎที่เข้ากันได้กับข้อมูลชุดทดสอบและมีต้นทุนการทำนายผิดพลาดที่น้อยที่สุด ถ้ากฎที่เข้ากันได้กับข้อมูลชุดทดสอบซึ่งมีคะแนนต้นทุนต่ำที่สุดเท่ากัน และกฎเหล่านี้ทำนายคลาสเดียวกันทั้งหมดตัวจำแนกประเภทก็จะตอบด้วยคลาสของกฎเหล่านี้ ถ้ากฎที่เข้ากันได้กับข้อมูลทดสอบทำนายต่างคลาสกัน ผู้วิจัยจะใช้การโหวตจากคลาสหลักที่อยู่ในกฎเหล่านี้ แต่ถ้าไม่มีคลาสหลักในกฎเหล่านี้เลย หรือมีจำนวนความถี่ของคลาสต่างๆ เท่ากัน และมีคะแนนต้นทุนเท่ากัน ตัวจำแนกประเภทข้อมูลจะเลือกกฎที่เข้ากันได้กับข้อมูลทดสอบที่มีคะแนนต้นทุนต่ำที่สุดอันดับรองลงมาเพื่อใช้จำแนกประเภทข้อมูลตามวิธีที่กล่าวมาข้างต้น ในกรณีที่มีกฎที่เข้ากันได้กับข้อมูลทดสอบแต่ไม่สามารถเลือกคลาสจากวิธีดังกล่าวข้างต้นมาตอบได้ ตัวจำแนกประเภทจะเลือกคลาสหลักจากกฎทั้งหมดเหล่านี้มาตอบ แต่ในกรณีที่ไม่มีคลาสหลักของกฎที่เข้ากันได้กับข้อมูลทดสอบหรือไม่มีกฎใดที่เข้ากันได้กับข้อมูลทดสอบเลย ตัวจำแนกประเภทข้อมูลจะใช้วิธีทำนายจากการอำพลางข้อมูล

3.2. การทำนายจากการอำพลางข้อมูล (Attribute Value Hiding) สืบเนื่องจากปัญหาจากความเบาบางของคลาสและความสัมพันธ์ของข้อมูล อาจจะทำให้ไม่สามารถหากฎความสัมพันธ์ใดๆ ได้เลยเพื่อนำไปใช้ในการทำนายข้อมูลใหม่ ดังนั้นเพื่อจัดการกับปัญหานี้ผู้วิจัยจึงนำเสนอเทคนิคการทำนายจากการอำพลางข้อมูล โดยมีแนวคิดดังนี้ เริ่มจากทำการปิดข้อมูลใหม่ครั้งละ 1 ไอเท็ม จากนั้นใช้อัลกอริทึมเดิม (การจำแนกประเภทจากกฎที่เข้ากันได้กับข้อมูลชุดทดสอบ) ทำการทำนายข้อมูลใหม่ที่ถูกอำพลางไปแล้ว 1 ไอเท็ม แล้วเก็บค่าคำตอบไว้ และทำซ้ำบนไอเท็มอื่นๆ ของการอำพลางข้อมูลครั้งแรก (1 ไอเท็ม) จนครบทุกไอเท็ม จากนั้นนำผลลัพธ์ที่ได้จากการอำพลาง 1 ไอเท็มมาทำการโหวตคลาสเพื่อทำนาย ถ้าผลโหวตสามารถระบุคลาสคำตอบได้ ก็จะตอบคลาสที่ถูกโหวตทันที แต่ถ้าไม่สามารถหาผลโหวตได้ ก็จะทำการอำพลางข้อมูลใหม่เพิ่มขึ้นอีกครั้งละ 1 ไอเท็มและทำซ้ำกระบวนการเดิม จนกว่าจะได้คำตอบหรือจนกว่าอำพลางจนหมดทุกไอเท็มแล้วก็ไม่ได้คำตอบ ก็จะสิ้นสุดการทำงานของวิธีการนี้ แนวคิดของวิธีการอำพลางก็คือการใช้กฎความสัมพันธ์ที่เป็นกฎพ่อ (Parent Rules) และเข้ากันได้กับข้อมูลใหม่ทุกตัวมาทำการโหวตผลการทำนาย ตัวอย่างการอำพลางข้อมูลใหม่แสดงไว้ในภาพที่ 24 ข้อมูลที่ถูกอำพลางจะแสดงด้วยสัญลักษณ์ ?

3.3. การตอบคลาสพื้นฐาน (Default Class) ในกรณีที่ไม่สามารถทำนายได้จากทุกวิธีที่กล่าวมา ตัวจำแนกประเภทจะเลือกคลาสพื้นฐาน (Default Class) ซึ่งก็คือคลาสหลักมาใช้เป็นคำตอบ

อัลกอริทึมสำหรับจำแนกประเภทข้อมูลแสดงอยู่ในภาพของรหัสเทียมดังภาพที่ 25 และ 26



ภาพที่ 24 ตัวอย่างการอำพลางข้อมูลใหม่

1. Let R denote the set of rules
2. Let u denote the unseen data
3. Let h_p denote the attribute value hiding index
4. Let C_p denote the predict class
5. Let U_p denote the set of instances which generate from unseen data with h_p degree
6. Let C_{HP} denote the set of predicted class from attribute value hiding
7. **classification***(R, u)*
8. $h_p = 0$ // attribute value hiding degree
9. $C_p = classifier(R, u)$
10. **while** $(C_p = \phi) \vee (h_p < |u|)$
11. $h_p = h_p + 1$ // No of attribute value hiding degree
12. U_p = generate instances with attribute value hiding h_p degree
13. $C_{HP} = \phi$
14. **for each** $u \in U_p$ **do**
15. $C_{HP} = C_{HP} \cup$ classifier(R, u)
16. **end for**
17. C_p = majority vote from $\forall\{c \mid c \in C_{HP}\}$
18. **end while**
19. **if** $C_p = \phi$ **then**
20. C_p = Default class
21. **end if**
22. **return** C_p

ภาพที่ 25 รหัสเทียมของอัลกอริทึมสำหรับการจำแนกประเภท

1. Let R denote the set of rules
2. Let u denote the unseen data
3. Let C_P denote the predict class
4. Let R_M denote the set of matched rules
5. Let C_M denote the majority class from set of matched rules
6. Let R_{Low} denote the set of rules which equally rule size and lowest cost from R_M
7. **classifier**(R, u)
8. $C_P = \phi$
9. $R_M = \{r \mid r \in R, r.X \subset u\}$ // Find all matched rules
10. C_M = *majority vote from* $\forall\{r.Y \mid r \in R_M\}$
11. **repeat**
12. R_{Low} = *set of rules which equally rule size and lowest cost from* R_M
13. **case**
14. **when** $\forall\{r \mid r \in R_{Low}\}$ *are the same* $\mid r.X \mid$ *and r.cost and* $r.Y$ **then**
15. // prediction with its class
$C_P = r.Y$
16. **when** $\forall\{r \mid r \in R_{Low}\}$ *are the same* $\mid r.X \mid$ *and r.cost but difference* $r.Y$ **then**
17. // prediction with majority votes from lowest cost of rules set.
C_P = *majority vote from* $\forall\{r.Y \mid r \in R_{Low}\}$
18. **end case**
19. **if** $C_P = \phi$ **then**
20. $R_M = R_M - R_{Low}$ // Remove all the rules which used
21. **end if**
22. **until** $(C_P \neq \phi) \vee (R_M = \phi)$
23. **if** $C_P = \phi$ **then**
24. $C_p = C_M$ // prediction with majority votes from all matched rules
25. **end if**
26. **return** C_P

ภาพที่ 26 รหัสเทียมของอัลกอริทึมการจำแนกประเภทด้วยกฎความสัมพันธ์ที่เข้ากันได้กับข้อมูลใหม่

ผลและวิจารณ์

ผล

1. วิธีวัดผลการทดลอง

ผู้วิจัยได้ทำการทดลองเปรียบเทียบประสิทธิภาพการทำงานของตัวจำแนกประเภทด้วยกฎความสัมพันธ์ตามที่นำเสนอในวิทยานิพนธ์เล่มนี้ซึ่งมีชื่อว่า SSCR เทียบกับตัวจำแนกประเภทด้วยกฎความสัมพันธ์ CBA และตัวจำแนกประเภทต้นไม้การตัดสินใจ C4.5 (Quinlan, 1993) ในเวอร์ชั่นที่อยู่ในซอฟต์แวร์ Weka (Hall *et al*., 2009) ในการทดลองผู้วิจัยได้อาศัยชุดข้อมูล UCI (Frank and Asuncion, 2010) จำนวน 4 ชุด คือ breast-cancer, yeast3, yeast6 และ abalone19 ซึ่งมีสัดส่วนของความไม่สมดุล (Imbalanced Ratio: IR) ไม่เท่ากัน โดยมีสัดส่วนความไม่สมดุลน้อยจนไปถึงความไม่สมดุลสูง ในตารางที่ 13 คือคุณสมบัติของชุดข้อมูลที่นำมาทำการทดลองโดยในแต่ละชุดข้อมูลจะแสดงถึง จำนวนข้อมูลตัวอย่าง (#Ins), จำนวนคุณสมบัติ (#Atts), จำนวนข้อมูลตัวอย่างที่เป็นคลาสบวก (#Pcs), จำนวนข้อมูลตัวอย่างที่เป็นคลาสลบ (#Ncs) และจำนวนชื่อคลาสทั้งหมด (#Cls)

ตารางที่ 13 รายละเอียดของชุดข้อมูลที่ไม่สมดุลที่ใช้ในการทดลองผล

Datasets	#Ins	#Atts	#Pcs	#Ncs	#Cls
Breast-Cancer	286	10	85	201	2
Yeast3	1484	9	163	1321	2
Yeast6	1484	9	35	1449	2
Abalone19	4174	9	32	4142	2

รูปแบบการทดลองผู้วิจัยได้เลือกใช้การทดสอบแบบ 10-Fold Cross Validation ซึ่งเป็นแนวทางการทดสอบที่ได้รับการยอมรับน่าเชื่อถือ โดยวิธีการก็คือการแบ่งข้อมูลออกเป็น 10 ส่วนเท่าๆ กันแล้วเก็บไว้ 1 ส่วนสำหรับเป็นชุดข้อมูลทดสอบ ส่วนที่เหลือนำมารวมกันใช้เป็นข้อมูลสำหรับการเรียนรู้ และทดสอบวัดประสิทธิภาพ ทำซ้ำเช่นนี้จำนวน 10 รอบ ซึ่งแต่ละรอบก็จะเปลี่ยนชุดข้อมูลสำหรับทดสอบและการเรียนรู้ตามที่แบ่งไว้ 10 ส่วนไปเรื่อยๆ จนครบ 10 รอบ โดยข้อมูลที่แบ่งไว้ 10 ส่วน แต่ละส่วนจะถูกทดสอบเพียง 1 ครั้งเท่านั้น ผลการวัดประสิทธิจะถูกเฉลี่ยออกมาเป็นประสิทธิภาพของตัวจำแนกประเภท

มาตรวัดผลที่ใช้วัดประสิทธภาพของตัวจำแนกประเภทข้อมูลในการทดลองครั้งนี้ได้ใช้มาตรวัดผล 5 ชนิดด้วยกัน ประกอบไปด้วย มาตรวัดความเที่ยงตรง (Precision), มาตรวัดค่าจดจำ (Recall), มาตรวัดเอฟ (F-Measure), มาตรวัดพื้นที่ใต้เส้นโค้งรับคุณลักษณะการทำงาน (ROC Area) และ มาตรวัดความแม่นยำ (Accuracy)

การกำหนดค่าสำหรับทดสอบ ตัวจำแนกประเภท SSCR ใช้ค่าวิกฤต α และตารางเมตริกต้นทุน (Cost Matrix) ในการเรียนรู้เพื่อสร้างเป็นโมเดลจำแนกประเภทข้อมูล ตัวจำแนกประเภท CBA ใช้ค่าสนับสนุน (Support) และค่าความเชื่อมั่น (Confident) ในการเรียนรู้ ส่วนตัวจำแนกประเภท C4.5 ใช้ค่าปัจจัยความเชื่อมั่น (Confident Factor) ในการเรียนรู้ สำหรับการทดลองตัวจำแนกประเภท SSCR ได้แบ่งออกเป็น 2 รูปแบบ ดังนี้

1.1. การทดสอบประสิทธิภาพ สำหรับการทดสอบรูปแบบนี้มีวัตถุประสงค์เพื่อต้องการทราบถึงประสิทธิภาพของตัวจำแนกประเภท SSCR ในการจำแนกคลาสรองบนชุดข้อมูลที่ไม่สมดุล ซึ่งคลาสรองเป็นคลาสแรกที่เราให้ความสนใจในการทำนาย ว่ามีความผิดพลาดประเภทลบเกิดขึ้นมากน้อยเพียงไร และมีความแม่นยำเพียงใด เมื่อเทียบกับตัวจำแนกประเภท CBA และ C4.5 การตั้งค่าสำหรับ SSCR ผู้วิจัยกำหนดให้ค่าวิกฤต $\alpha = 0.01$ และเมตริกต้นทุน $C = [-1{,}100;1{,}0]$ จากตารางที่ 9 มาเป็นค่าทดสอบ โดยค่า $\alpha = 0.01$ เพื่อต้องให้กฎความสัมพันธ์มีนัยทางสถิติอย่างมีนัยสำคัญ (Statistically Significant) ส่วนตารางเมตริกต้นทุนกำหนดให้ C(+,+)=-1 เป็นค่ารางวัลนั้นคือถ้ากฎทำนายคลาสรองถูกต้องอยู่แล้ว ก็ควรจะมีต้นทุนต่ำว่าตัวอื่นๆ C(+,-)=100 เพื่อเป็นค่าปรับในการทำนายคลาสรองผิด ซึ่งหากทำนายผิดควรจะต้องมีต้นทุนที่สูงมากเนื่องจากคลาสรองคือคลาสที่เราให้ความสำคัญเป็นอันดับแรก C(-,+)=1 เป็นค่าปรับเมื่อทำนายคลาสหลักผิดแต่เราให้ความสำคัญน้อยจึงกำหนดค่าปรับด้วยค่าเล็กน้อย C(-,-)=0 ไม่มีค่าปรับและไม่มีรางวัลเนื่องจากทำนายคลาสหลักถูกอยู่แล้ว ส่วนตัวจำแนกประเภท CBA กำหนดให้ค่าสนับสนุนขั้นต่ำเป็น 1% ($minSup = 1\%$) และค่าสนับสนุนเป็น 1% ($minConf = 1\%$) เพื่อให้มีกฎความสัมพันธ์มากเพียงพอต่อการจำแนกประเภทข้อมูลที่ไม่สมดุล สำหรับ C4.5 ซึ่งใช้ค่าปัจจัยความเชื่อมั่นในการสร้างต้นไม้ เพื่อให้ได้ต้นไม้การตัดสินใจที่มีจำนวนโหนดมากที่สุด จึงกำหนดค่าปัจจัยความเชื่อมั่นเป็น 100% ($conf = 100\%$)

1.2. การทดสอบผลกระทบของตารางเมตริกต้นทุน เพื่อศึกษาและวิเคราะห์ถึงผลกระทบของเมตริกต้นทุนที่เปลี่ยนแปลงหลากหลายค่า ว่ามีผลกับประสิทธิภาพของตัวจำแนกประเภท

SSCR อย่างไรบ้างผู้วิจัยได้กำหนดค่าตารางเมตริกต้นทุนหลากหลายค่าโดยแสดงรายละเอียดไว้ในตารางที่ 14 ซึ่งทุกการทดลองกำหนดค่าวิกฤต $\alpha = 0.01$

ตารางที่ 14 เมตริกต้นทุนที่ใช้วิเคราะห์ผลกระทบของตัวจำแนกประเภท SSCR

	C(+,+)	**C(+,-)**	**C(-,+)**	**C(-,-)**
C0	0	1	1	0
CDP	0	#Ncs/#Pcs	1	0
CP1	-1	50	1	0
CP2	-1	100	1	0
CP3	-1	250	1	0
CP4	-1	500	1	0
CP5	-1	1000	1	0
CPN1	-1	100	50	0
CDN	0	1	#Ncs/#Pcs	0
CN1	-1	1	50	0
CN2	-1	1	100	0

2. ผลการทดสอบประสิทธภาพ

2.1. อัตราความถูกต้องประเภทบวก (TPR) จากผลการทดสอบประสิทธิภาพความถูกต้องของคลาสรองซึ่งแสดงในตารางที่ 15 แสดงให้เห็นได้ว่า SSCR ให้อัตราความถูกต้องของคลาสบวกสูง ทำให้มีความผิดพลาดในข้อมูลประเภทลบ (FNR) ต่ำกว่า CBA และ C4.5 ในขณะที่ข้อมูลมีความไม่สมดุลสูงมากเช่น abalone19 จะเห็นได้ชัดเจนว่าทั้ง CBA และ C4.5 ต่างทำนายเข้าหาคลาสหลักทั้งหมดจนทำให้เกิดความผิดพลาดในข้อมูลประเภทลบ (FNR) 100% ในขณะที่ SSCR ยังคงสามารถให้ผลการทำนายในคลาสบวกได้ดี ซึ่งเป็นผลมาจากกฎความสัมพันธ์ที่มีคุณภาพและมีนัยสำคัญ

ตารางที่ 15 เปรียบเทียบอัลกอริทึมด้วยมาตรวัด TPR บนคลาสบวก

Algorithm	Measure	Breast Cancer	Yeast3	Yeast6	Abalone19
CBA	TPR	36.5%	23.3%	0.0%	0.0%
C4.5	TPR	37.6%	69.9%	42.9%	0.0%
SSCR	TPR	**60.00%**	**91.41%**	**82.86%**	**68.75%**

2.2. อัตราความผิดพลาดประเภทบวก (FPR) ความผิดพลาดประเภทบวกเกิดจากการทำนายข้อมูลคลาสลบผิดเป็นคลาสบวก ซึ่งในผลการทดสอบประสิทธิภาพความผิดพลาดประเภทบวกที่ได้แสดงในตารางที่ 16 พบว่า SSCR มีความผิดพลาดประเภทบวกสูงกว่าอัลกอริทึมอื่นๆ ปัญหาเกิดมากจาก อัลกอริทึมของ SSCR เน้นการจัดการคลาสรองด้วยค่าความเสี่ยงของกฎความสัมพันธ์ อีกทั้งการกำหนดค่าวิกฤตที่ต่ำมากๆ จะทำให้สูญเสียกฎความสัมพันธ์บางส่วนไป ส่งผลให้เกิดความผิดพลาดประเภทบวกมากกว่าอัลกอริทึมอื่นๆ จากการทดสอบพบว่าอัลกอริทึมที่เกิดความผิดพลาดประเภทบวกน้อยที่สุดคือ CBA สืบเนื่องจากมีจำนวนกฎความสัมพันธ์จำนวนมากและทำนายเข้าสู่คลาสหลักมากกว่าคลาสรอง

ตารางที่ 16 เปรียบเทียบอัลกอริทึมด้วยมาตรวัด FPR บนคลาสบวก

Algorithm	Measure	Breast Cancer	Yeast3	Yeast6	Abalone19
CBA	FPR	**16.9%**	**0.2%**	**0.0%**	**0.0%**
C4.5	FPR	19.9%	4.2%	0.8%	0.1%
SSCR	FPR	35.32%	16.12%	22.84%	23.47%

2.3. ประสิทธิภาพความเที่ยงตรง (Precision) จะเห็นได้ชัดเจนว่าตัวจำแนกประเภท CBA และ C4.5 ให้ผลลัพธ์ที่ดีกว่าซึ่งสาเหตุมาจากตัวจำแนกทั้งคู่สามารถจัดการคลาสหลักได้ดีกว่าคลาสรอง และจำนวนตัวอย่างของคลาสหลักมีสัดส่วนมากกว่าคลาสรองมาก จึงส่งผลให้เกิด TN และ FN มากกว่า ในขณะที่ SSCR แม้จะสามารถจัดการกับคลาสรองได้ดี แต่ก็เกิด FPR มากกว่า อันเนื่องจากการใช้ต้นทุนในการเรียนรู้ซึ่งยอมให้เกิด FP มากกว่า FN เพราะว่าการทำนายคลาสบวกผิดเป็นคลาสลบมีต้นทุนการทำนายผิดพลาดสูงกว่า ดังนั้นเมื่อพิจารณาจากมาตรวัดความเที่ยงตรงแล้วจึงส่งผลให้ทั้ง CBA และ C4.5 มีความเที่ยงตรงมากกว่า

ตารางที่ 17 เปรียบเทียบอัลกอริทึมด้วยมาตรวัด Precision บนคลาสบวก

Algorithm	Measure	Breast Cancer	Yeast3	Yeast6	Abalone19
CBA	Precision	**47.7%**	**92.7%**	0.0%	0.0%
C4.5	Precision	44.4%	67.5%	**55.6%**	0.0%
SSCR	Precision	41.80%	41.16%	8.06%	**2.21%**

2.4. ประสิทธิภาพความจดจำ (Recall) อัลกอริทึม SSCR ให้ค่า TPR ที่สูงซึ่งทำให้เกิด FN ที่ต่ำกว่าอัลกอริทึมอื่นๆ รวมทั้ง SSCR ถูกออกแบบมาเพื่อจัดการปัญหาความไม่สมดุลที่มีคลาสรองเป็นคลาสที่ให้ความสนใจเป็นพิเศษ จึงส่งผลให้ SSCR ยังคงทำงานได้ดีแม้จะเกิดความไม่สมดุลของคลาสสูงมากก็ตามเช่น abalone19 จะเห็นได้ว่าทั้ง CBA และ C4.5 ทำนายคลาสรองผิดพลาดหมดเลย เกิด FNR 100% นั้นก็คือเป็นการทำนายเข้าหาคลาสหลักทั้งหมด ตารางที่ 18 จะแสดงให้เห็นถึงประสิทธิภาพความจดจำของอัลกอริทึม SSCR

ตารางที่ 18 เปรียบเทียบอัลกอริทึมด้วยมาตรวัด Recall บนคลาสบวก

Algorithm	Measure	Breast Cancer	Yeast3	Yeast6	Abalone19
CBA	Recall	36.5%	23.3%	0.0%	0.0%
C4.5	Recall	37.6%	69.9%	42.9%	0.0%
SSCR	Recall	**60.00%**	**91.41%**	**82.86%**	**68.75%**

2.5. ประสิทธิภาพจากมาตรวัดเอฟ (F-Measure) เมื่อพิจารณาทั้งความเที่ยงตรงและความจดจำด้วยมาตรวัดเอฟเป็นการยากที่จะออกแบบอัลกอริทึมที่ให้ผลลัพธ์ที่ดีทั้งความเที่ยงตรงและความจดจำ ซึ่ง SSCR เองก็ให้ค่าความจดจำที่ดีแต่มีค่าความเที่ยงตรงที่ต่ำจึงส่งผลให้ประสิทธิภาพจากมาตรวัดเอฟไม่สูงมาก เมื่อเทียบกับ C4.5 แต่ SSCR ก็ยังคงให้ผลลัพธ์ที่ดีกว่าเมื่อข้อมูลเกิดความไม่สมดุลสูงมาก ผลการทดสอบถูกแสดงไว้ในตารางที่ 19

ตารางที่ 19 เปรียบเทียบอัลกอริทึมด้วยมาตรวัด F-Measure บนคลาสบวก

Algorithm	Measure	Breast Cancer	Yeast3	Yeast6	Abalone19
CBA	F-Measure	41.3%	37.3%	0.0%	0.0%
C4.5	F-Measure	40.8%	**68.7%**	**48.4%**	0.0%
SSCR	F-Measure	**49.28%**	56.76%	14.68%	**4.29%**

2.6. ประสิทธิภาพจากมาตรวัดพื้นที่ใต้เส้นโค้งรับคุณลักษณะการทำงาน (ROC Area) เมื่อพิจารณาการถ่วงดุลระหว่างความถูกต้องประเภทบวก TPR และความผิดพลาดประเภทบวก FPR แล้วจะเห็นว่าอัลกอริทึม SSCR นั้นให้ประสิทธิภาพโดยรวมที่ดีกว่า CBA และ C4.5 นั้นหมายถึง SSCR สามารถจำแนกประเภทคลาสรองได้ดีกว่าอัลกอริทึมอื่นๆ ซึ่งแสดงผลการทดสอบไว้ในตารางที่ 20

ตารางที่ 20 เปรียบเทียบอัลกอริทึมด้วยมาตรวัด ROC Area บนคลาสบวก

Algorithm	Measure	Breast Cancer	Yeast3	Yeast6	Abalone19
CBA	ROC Area	59.8%	61.5%	50.0%	50.0%
C4.5	ROC Area	58.1%	**88.9%**	72.8%	69.0%
SSCR	ROC Area	**62.30%**	87.60%	**79.90%**	**72.20%**

2.7. ประสิทธิภาพความแม่นยำ เพื่อพิจารณาประสิทธิภาพความแม่นยำโดยรวมของอัลกอริทึม SSCR แล้ว จึงได้ทำการทดสอบผ่านมาตรวัดความแม่นยำซึ่งจากผลการทดสอบก็ได้แสดงให้เห็นชัดเจนว่า อัลกอริทึม SSCR ให้ความแม่นยำโดยรวมได้ดีกว่า CBA และ C4.5 เมื่อข้อมูลมีลักษณะไม่สมดุลสูงมากๆ จะเห็นได้ว่าความแม่นยำของ CBA และ C4.5 มีค่าเท่ากับการเดาคือมีความแม่นยำเพียง 50% แต่ SSCR ยังคงให้ความแม่นยำที่ดีกว่าในทุกๆ อัตราสัดส่วนของความไม่สมดุลมากและน้อย ซึ่งแสดงผลการทดสอบไว้ในตารางที่ 21

ตารางที่ 21 เปรียบเทียบอัลกอริทึมด้วยมาตรวัด Accuracy

Algorithm	Measure	Breast Cancer	Yeast3	Yeast6	Abalone19
CBA	Accuracy	59.8%	61.6%	50.0%	50.0%
C4.5	Accuracy	58.9%	82.9%	71.1%	50.0%
SSCR	Accuracy	**63.29%**	**84.70%**	**77.29%**	**76.47%**

3. ผลการทดสอบผลกระทบของเมตริกต้นทุน

ในการวิเคราะห์ถึงผลกระทบของเมตริกต้นทุนผู้วิจัยได้พิจารณาเริ่มจากการจำแนกประเภทด้วยต้นทุนปริยาย (Cost Default) คือใช้เมตริกต้นทุน C0 [0,1;1,0] นั้นคือ หากทำนายผิดพลาดทั้งคลาสหลักและคลาสรองจะมีค่าปรับเท่ากับ 1 เท่ากัน กรณีเมตริกต้นทุน CDP [0, #Ncs/#Pcs;1,0] คือการปรับค่าต้นทุนตามสัดส่วนการจะจายตัวของคลาส (Class Distribution) โดยค่าปรับของ FN เกิดจากจำนวนตัวอย่างข้อมูลของคลาสหลักหารด้วยจำนวนตัวอย่างข้อมูลของคลาสรอง กรณีเมตริกต้นทุน CP1 [-1,50;1,0], CP2 [-1,100;1,0], CP3 [-1,250;1,0], CP4 [-1,500;1,0] และ CP5 [-1,1000;1,0] คือการพยายามลดความผิดพลาดของการเกิด FN ซึ่งกำหนดให้มีค่าปรับที่สูงกว่าความผิดพลาดอื่นๆ โดยมีค่าปรับเริ่มจาก 50, 100, 250, 500 และ 1000 ตามลำดับ กรณีเมตริกต้นทุน CPN1 [-1,100;50,0] คือการพยายามลดความผิดพลาดทั้ง FN และ FP โดยยังคงให้ความผิดพลาดของ FN เกิดได้น้อยกว่า FP ซึ่งทั้ง C0, CP1-5 และ CPN1 จะพิจารณาดูผลลัพธ์ที่คลาสรอง ส่วนกรณีเมตริกต้นทุน CDN [0,1; #Ncs/#Pcs,0], CN1 [-1,1;50,0] และ CN2 [-1,1;100,0] คือการพยายามลดความผิดพลาดของการเกิด FP ซึ่งจะพิจารณาดูผลลัพธ์ที่คลาสหลัก

จากภาพที่ 27 แสดงให้เห็นว่ากรณีที่ปรับลดความผิดพลาดของ FN เพียงค่าเดียว ส่งผลดีต่อ TPR บนชุดข้อมูลต่างๆ ด้วยค่าต้นทุนที่ต่างกัน แต่เมื่อปรับค่าจนถึงจุดหนึ่งแล้วพบว่าค่า TPR จะคงที่ ซึ่งเห็นได้ชัดเจนว่าค่าต้นทุนมีผลกระทบต่อ TPR ที่สูงขึ้นและลดลง สาเหตุเกิดจากกฎที่ทำนายคลาสรองมีค่อนข้างน้อยทำให้การปรับค่าต้นทุนของ FN ส่งผลโดยตรงต่อต้นทุนความเสี่ยงของกฎที่ทำนายคลาสรองในการเลือกไปทำนาย เมื่อพิจารณาเปรียบเทียบกับการปรับตารางต้นทุนด้วยสัดส่วนการกระจายตัวของคลาสเช่นกรณี CDP จะเห็นว่าส่งผลให้มี TPR ที่ดีขึ้นเมื่อเทียบกับค่าต้นทุนปริยาย แต่ค่า CDP นั้นยังไม่สามารถให้ประสิทธิภาพที่ดีที่สุดของ TPR เมื่อเปรียบเทียบกับ CP1 สำหรับ Brest-Cancer, Yeast3, Yeast6 และ CP3 สำหรับ Abalone19 ดังนั้นการเลือกค่าตารางเมตริกต้นทุนที่เหมาะสมจะช่วยทำให้ SSCR มีประสิทธิภาพสูงสุด และหากต้องการค่าเริ่มต้นใน

การปรับตารางต้นทุน วิธีการใช้สัดส่วนการกระจายตัวของคลาสเช่น CDP อาจจะเป็นค่าเริ่มต้นที่เหมาะสม

การปรับลดความผิดพลาดของ FN และ FP ในเวลาเดียวกัน ซึ่งแสดงผลลัพธ์อยู่ในภาพที่ 28 แสดงให้เห็นว่าค่าตารางเมตริกต้นทุนส่งผลให้ TPR ต่ำลง และทำให้ FPR ลดลงด้วยเช่นกัน ซึ่งเป็นผลจากกฎความสัมพันธ์ทั้งหมดที่มีนั้นมีโอกาสที่กฎจะตอบคลาสหลักมากกว่าคลาสรองเพิ่มขึ้นจากค่าต้นทุนความเสี่ยงของกฎที่เปลี่ยนไป ดังนั้นการลดความผิดพลาดของ FP ก็เท่ากับเพิ่มโอกาสให้กฎที่เป็นคลาสหลักมีโอกาสถูกเลือกไปทำนายมากขึ้น และทำให้กฎที่ทำนายคลาสรองมีโอกาสน้อยลง ถึงแม้จะมีค่าต้นทุนที่สูงกว่าก็ตาม

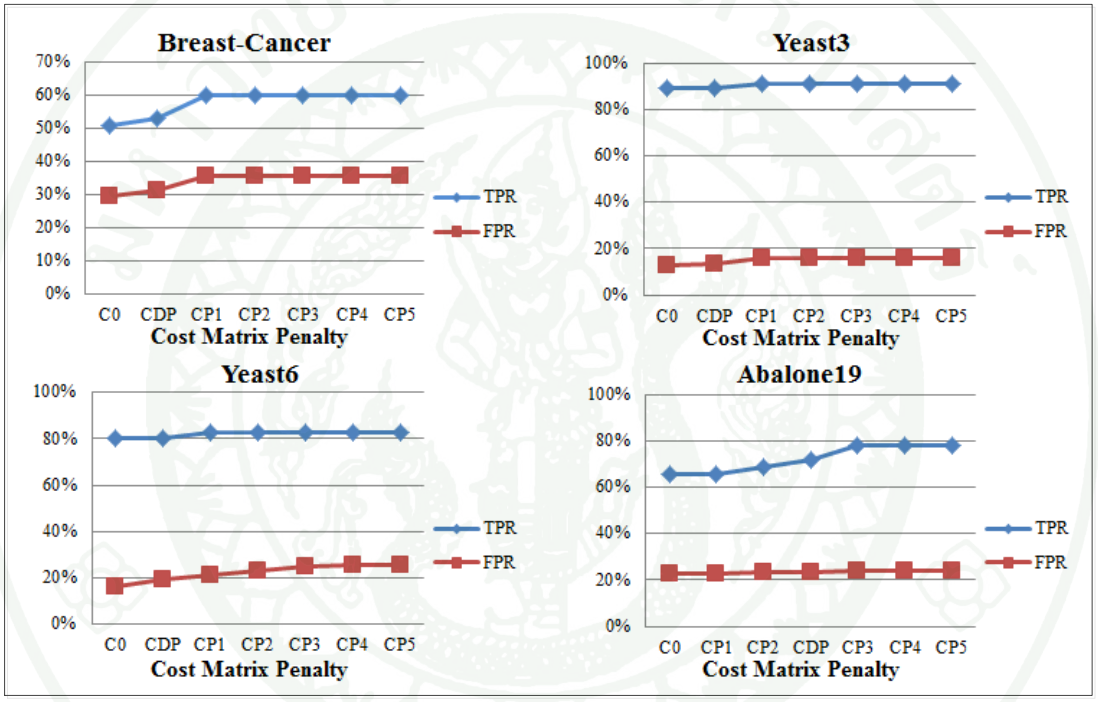
ในทางกลับกันเมื่อปรับลดความผิดพลาดของ FP และดูผลกระทบบนคลาสหลัก ซึ่งแสดงผลลัพธ์อยู่ในภาพที่ 29 พบว่าค่าเมตริกต้นทุนส่งผลให้ TNR และ FNR สูงขึ้นบนฐานข้อมูล Brest-Cancer ซึ่งเกิดจากกฎความสัมพันธ์ที่ได้มานั้นมีสัดส่วนของคลาสหลักและคลาสรองไม่ต่างกันมากนัก ดังนั้นในการปรับค่า FP จึงส่งผลดีกับชุดข้อมูล Brest-Cancer แต่เมื่อพิจารณาที่ชุดข้อมูลอื่นๆ ที่มีความไม่สมดุลมากกว่าพบว่าอัตรา TNR นั้นคงที่ และ FNR มีอัตราที่ลดลงเล็กน้อย ซึ่งเกิดจากกฎความสัมพันธ์ที่มีอยู่ส่วนมากทำนายคลาสหลักอยู่แล้ว ดังนั้นการลดความผิดพลาดของ FP จึงเป็นการเพิ่มความน่าจะเป็นให้กฎเหล่านี้ถูกพิจารณาในการทำนายเพิ่มมากขึ้นซึ่งไม่ได้ส่งผลในเรื่องประสิทธิภาพในการจำแนกของคลาสหลัก ดังนั้นจึงสรุปได้ว่าหากชุดข้อมูลมีความไม่สมดุลสูงและคลาสที่เราสนใจคือคลาสหลัก การปรับ FP ให้มีค่ามากๆ อาจจะทำให้ FNR ลดลงเล็กน้อยบางชุดข้อมูล แต่ไม่ส่งผลกับค่า TNR

เมื่อพิจารณาจาก F-Measure ซึ่งแสดงไว้ในภาพที่ 30 จะเห็นว่ากรณีปรับลดความผิดพลาดของ FN ส่งผลให้ F-Measure บนคลาสรองของชุดข้อมูลที่ไม่สมดุลสูงมากๆ เช่น Abalone19 มีผลลัพธ์ที่ดีขึ้นกว่าการที่ไม่คิดต้นทุนใดๆ เลย แต่ในขณะที่ Yeast3 และ Yeast6 กลับมี F-Measure ที่ลดลง แต่หากพิจารณาแล้วพบว่า ทั้ง Yeast3, Yeast6 และ Abalone19 นั้นมีค่า TPR ที่สูงขึ้นและ FPR ที่สูงขึ้นเช่นเดียวกัน ซึ่งเมื่อเปรียบเทียบบนสัดส่วนของจำนวนข้อมูลแล้ว Abalone19 มีจำนวนข้อมูลที่มากกว่า ดังนั้นจึงทำให้ F-Measure ของ Abalone19 มีประสิทธิภาพที่ดีขึ้น ขณะที่ Yeast3 และ Yeast6 มีค่า F-Measure ที่ลดลง

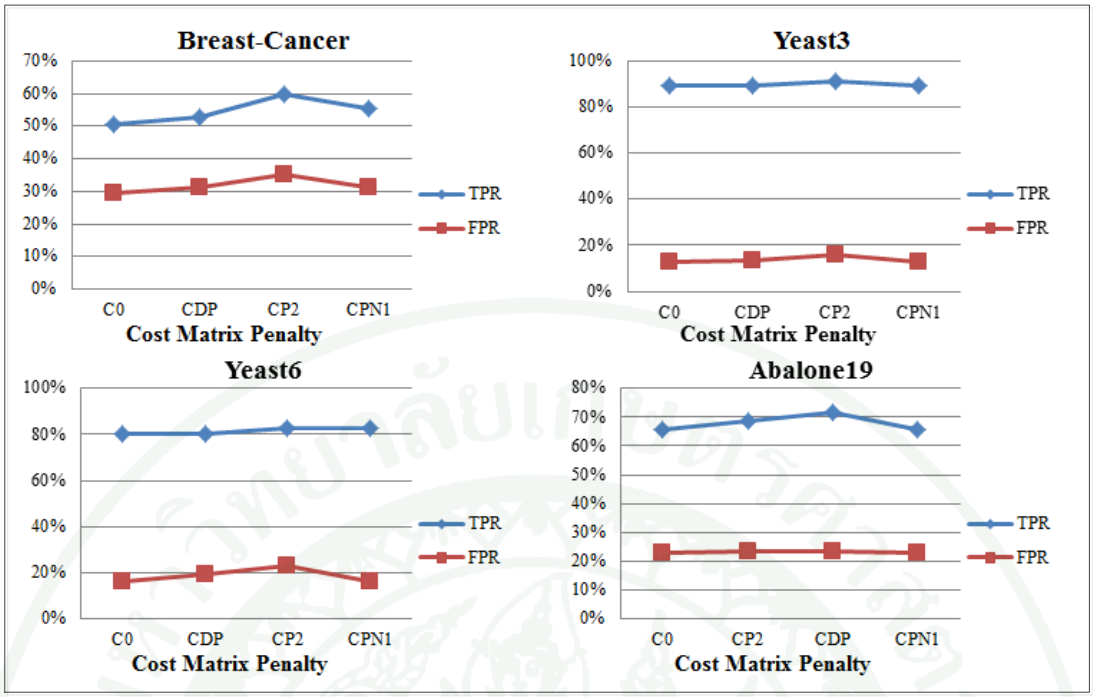
ในกรณีที่ปรับลดความผิดพลาดทั้ง FN และ FP ซึ่งแสดงผลลัพธ์ไว้ในภาพที่ 31 พบว่าการเกิด TPR และ FPR นั้นลดลงทำให้เห็นผล F-Measure เปลี่ยนแปลงลดลงได้อย่างชัดเจนใน Breast-

Cancer และ Abalone19 แต่ในขณะที่ Yeast3 และ Yeast6 นั้นมีการเปลี่ยนแปลง TPR ลดลงเล็กน้อยแต่มีการเปลี่ยนแปลง FPR ลดลงมากกว่า จึงทำให้เห็นค่า F-Measure ส่งผลดีขึ้น

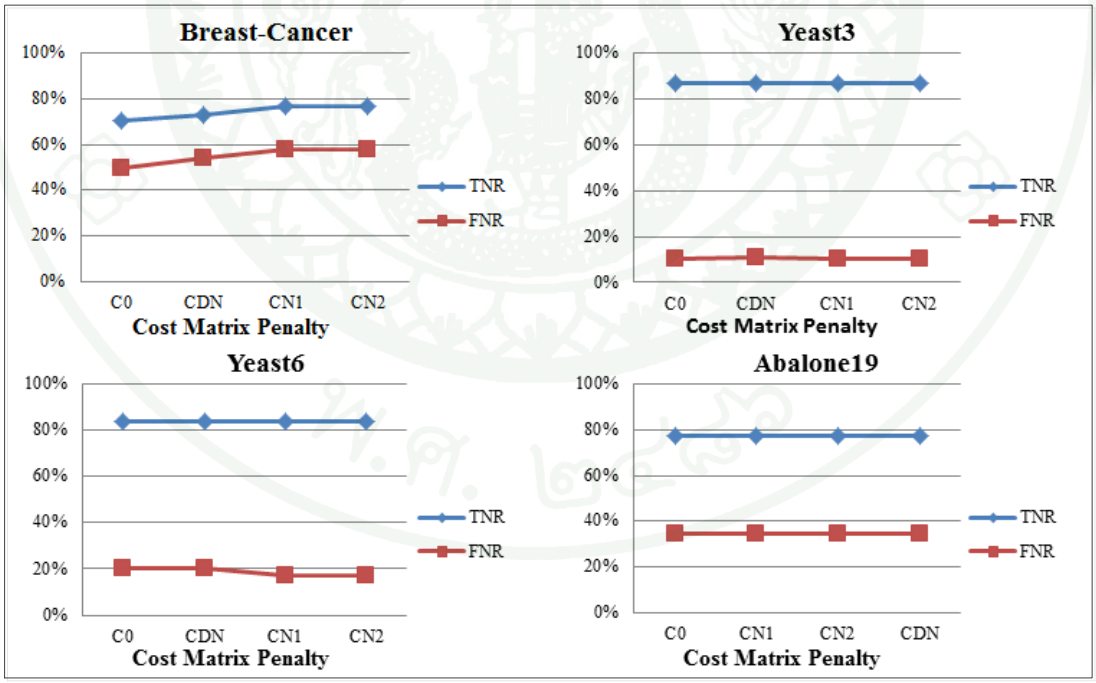
ในกรณีของการปรับลดความผิดพลาดของ FP เพื่อพิจารณาผลกระทบบนคลาสหลัก ซึ่งแสดงไว้ในภาพที่ 32 จะเห็นว่าค่าเมตริกต้นทุนก็ยังส่งผลให้ F-Measure ของคลาสหลักมีค่าที่ดีขึ้นบนชุดข้อมูล Breast-Cancer แต่ไม่ส่งผลกับชุดข้อมูลอื่นๆ ที่มีความไม่สมดุลสูง อันเนื่องจากการมีกฎของคลาสหลักที่มากกว่านั้นเอง



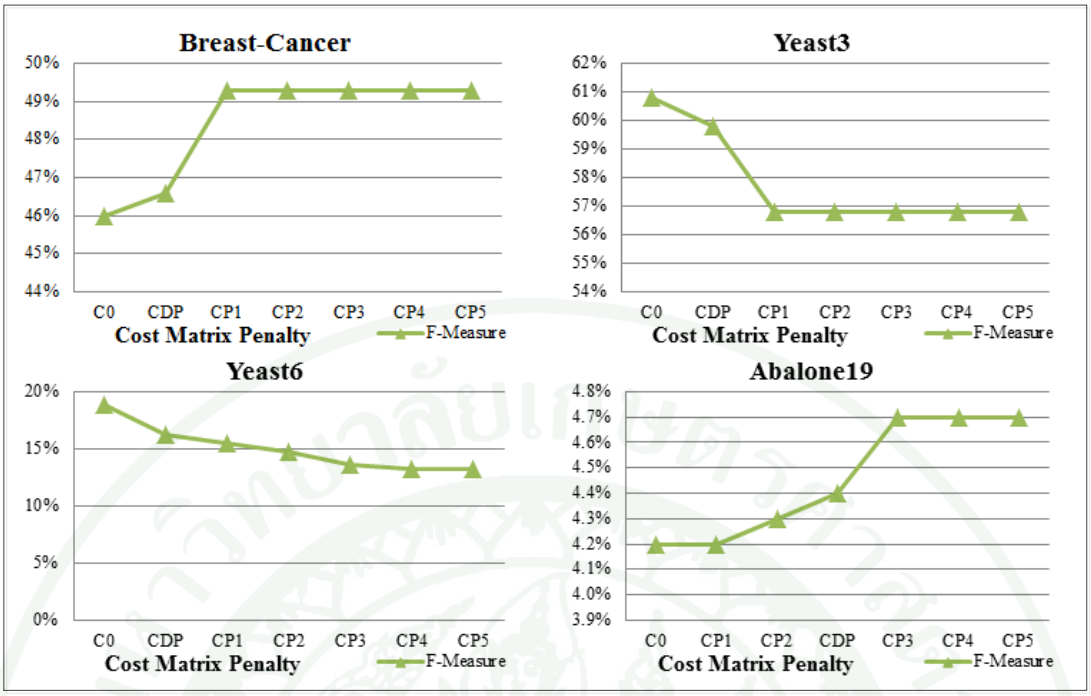
ภาพที่ 27 ผลกระทบ TPR, FPR บนคลาสรอง กรณีปรับลดความผิดพลาดของ FN



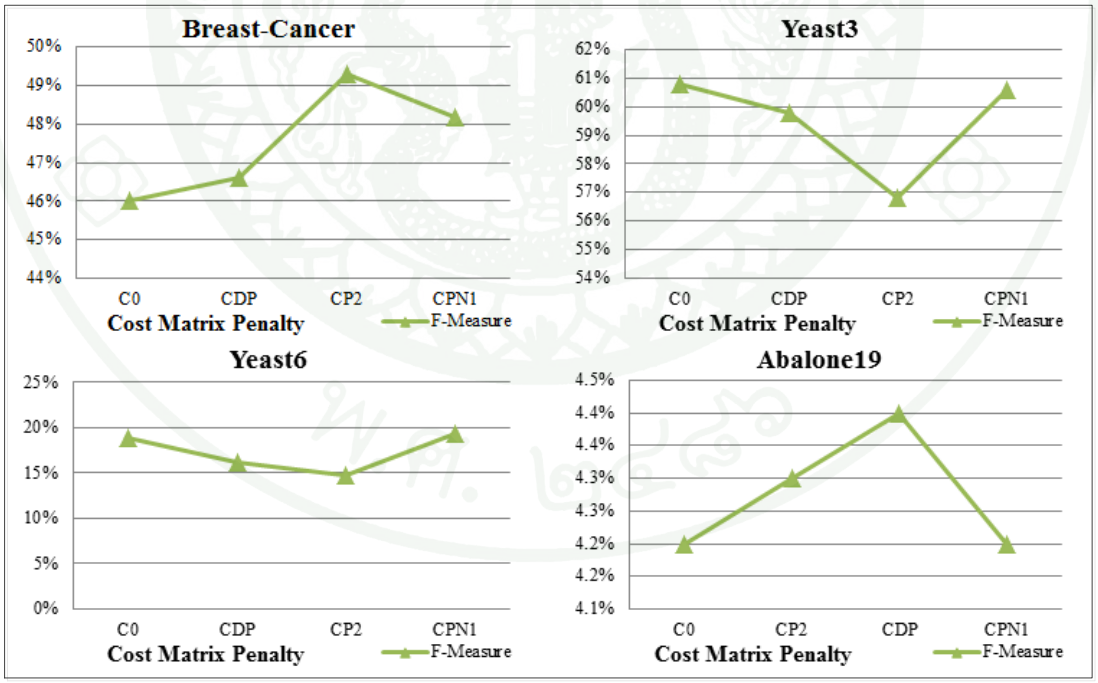
ภาพที่ 28 ผลกระทบ TPR, FPR บนคลาสรอง กรณีปรับลดความผิดพลาดของ FN และ FP



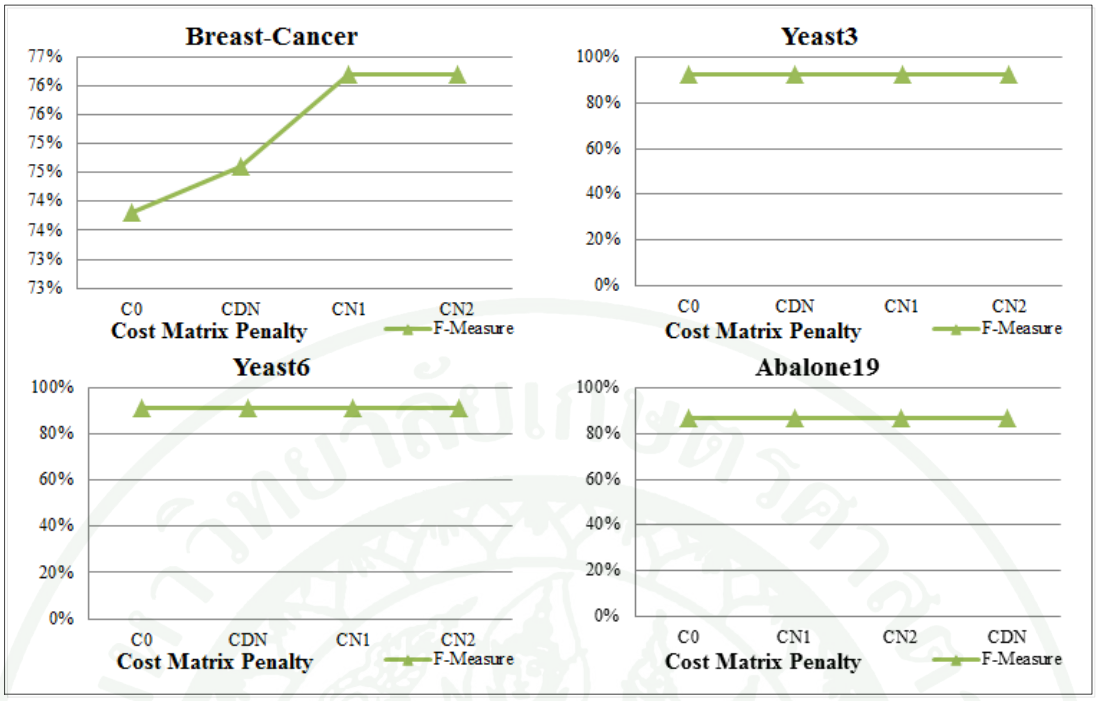
ภาพที่ 29 ผลกระทบ TNR, FNR บนคลาสหลัก กรณีปรับลดความผิดพลาดของ FP



ภาพที่ 30 ผลกระทบ F-Measure บนคลาสรอง กรณีปรับลดความผิดพลาดของ FN



ภาพที่ 31 ผลกระทบ F-Measure บนคลาสรอง กรณีปรับลดความผิดพลาดของ FNและ FP



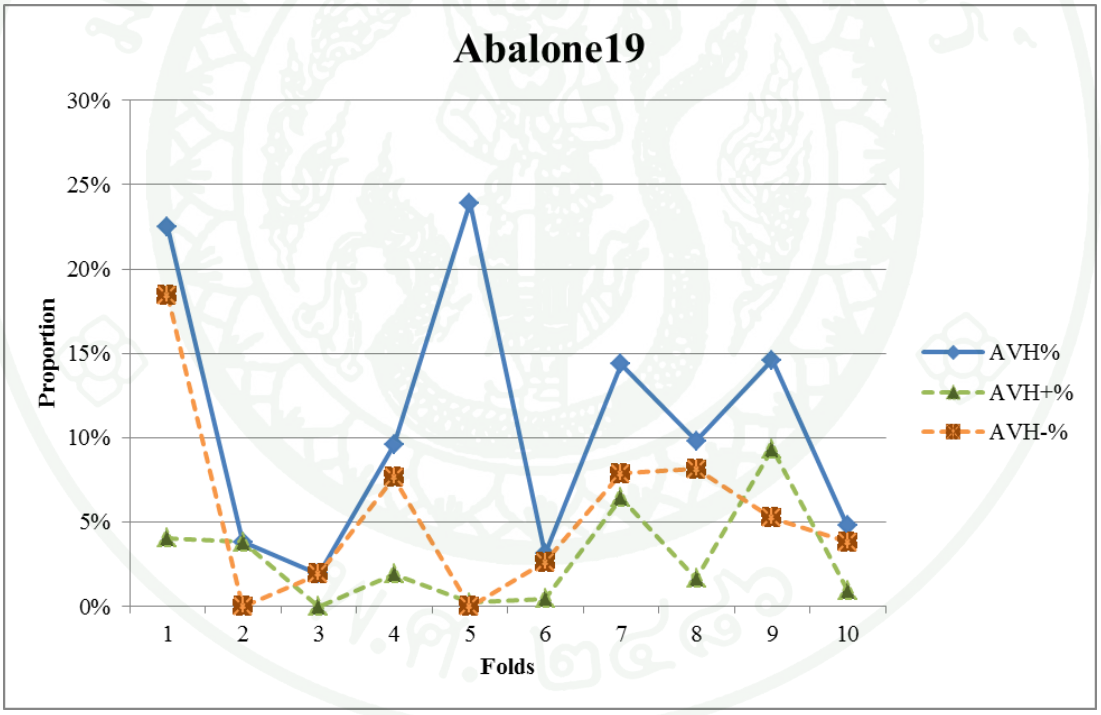
ภาพที่ 32 ผลกระทบ F-Measure บนคลาสหลัก กรณีปรับลดความผิดพลาดของ FP

4. ผลการทดสอบประสิทธิภาพของวิธีการอำพลางข้อมูล

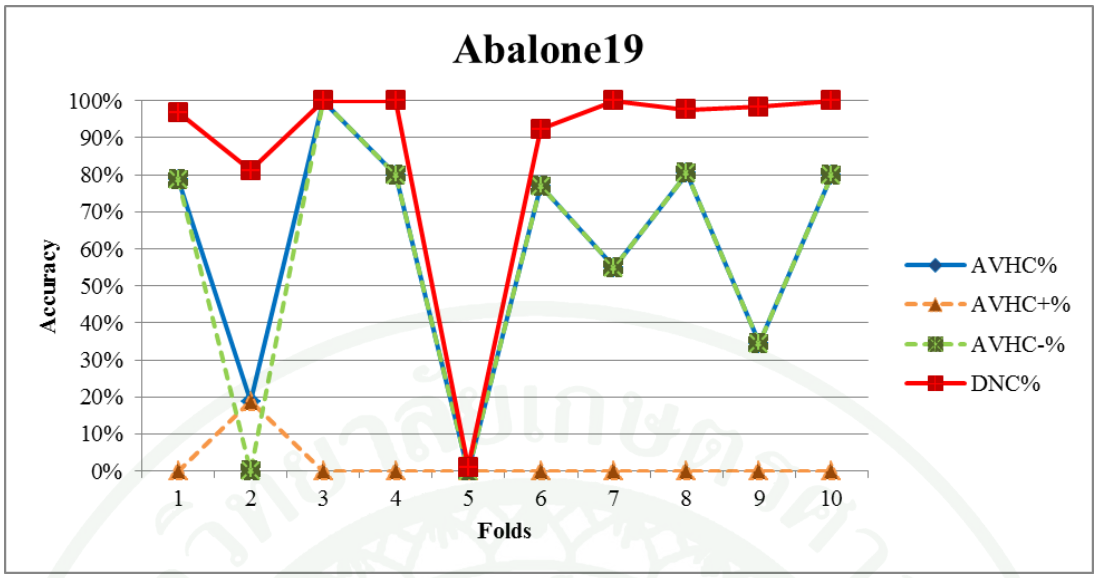
ในการทดสอบเพื่อวิเคราะห์ถึงประสิทธิภาพของเทคนิคการอำพลางข้อมูลผู้วิจัยได้ทำการเปรียบเทียบกับการตอบด้วยคลาสหลักเสมอ (Default Negative Class) บนชุดข้อมูล Abalone19 ซึ่งมีความไม่สมดุลสูงสุดและมีจำนวนกฎความสัมพันธ์ที่มีนัยสำคัญทางสถิติน้อยที่สุดมาทำการทดสอบและอ่านผลจากแต่ละ Fold ของ 10-Fold Cross Validation ซึ่งจากภาพที่ 33 แสดงให้เห็นถึง AVH% คือจำนวนเปอร์เซนต์ของข้อมูลชุดทดสอบที่ตอบด้วยวิธีการอำพลางข้อมูล โดยที่ AVH+% คือจำนวนเปอร์เซ็นต์ที่วิธีการอำพลางข้อมูลตอบด้วยคลาสรอง และ AVH-% คือจำนวนเปอร์เซ็นต์ที่วิธีการอำพลางข้อมูลตอบด้วยคลาสหลัก ซึ่งโดยเฉลี่ยจากทุก Fold จะตอบด้วยคลาสหลักมากกว่าคลาสรอง

ผลการเปรียบเทียบด้วยการตอบด้วยคลาสหลักเสมอและการตอบด้วยวิธีการอำพลางข้อมูลที่ถูกแสดงในภาพที่ 34 โดย AVHC% แสดงถึงจำนวนเปอร์เซ็นต์ที่วิธีการอำพลางข้อมูลทำนายได้ถูกต้อง AVHC+% แสดงถึงจำนวนเปอร์เซ็นต์ที่วิธีการอำพลางข้อมูลทำนายคลาสรองได้ถูกต้อง AVHC-% แสดงถึงจำนวนเปอร์เซ็นต์ที่วิธีการอำพลางข้อมูลทำนายคลาสหลักได้ถูกต้อง และ DNC% แสดงถึงจำนวนเปอร์เซ็นต์ที่วิธีตอบด้วยคลาสหลักเสมอทำนายได้ถูกต้อง ซึ่งจากผลการ

ทดสอบจะเห็นได้ว่าการตอบด้วยคลาสหลักเสมอนั้นให้ความถูกต้องได้มากกว่า อันเนื่องมาจากจำนวนกฎของคลาสหลักที่ถูกกำจัดทิ้งไปจำนวนมากนั้น อาจจะทำให้เราสูญเสียกฎที่มีไอเท็มเซต X ที่มีค่าของแอตทริบิวต์ตรงกับข้อมูลใหม่ (Unseen Data) ดังนั้นเมื่อเทียบกับสัดส่วนของข้อมูลบนคลาสรองแล้วจะเห็นว่าโอกาสที่ไม่มีกฎที่เหมาะสมกับข้อมูลใหม่นั้นเกิดขึ้นได้น้อยกว่าหรือมีความถี่ในการเกิดขึ้นต่ำกว่า ซึ่งแสดงให้เห็นว่าบนข้อมูลที่ไม่สมดุลมากๆ นั้น เช่น Abalone19 การเลือกตอบด้วยคลาสหลักเสมอมีโอกาสที่จะทำนายได้ถูกต้องมากกว่า แม้ว่าวิธีการอำพลางข้อมูลจะให้ประสิทธิภาพที่ด้อยกว่าการตอบด้วยคลาสหลักเสมอ แต่ในผลการทดสอบจะเห็นว่าวิธีการอำพลางข้อมูลก็มีประสิทธิภาพบนสถานการณ์ของคลาสรองอยู่บ้าง แต่เนื่องด้วยสัดส่วนของตัวอย่างข้อมูลที่คลาสรองมีนั้นน้อยกว่าข้อมูลของคลาสหลักอยู่อย่างมากแล้วนั้น โอกาสที่ข้อมูลใหม่จะเป็นคลาสรอง จึงมีน้อยกว่ามาก ดังนั้นในสถานการณ์เช่นนี้การเลือกตอบด้วยคลาสหลักเสมออาจจะได้ประสิทธิภาพที่ดีกว่า



ภาพที่ 33 จำนวนเปอร์เซ็นต์ของข้อมูลชุดทดสอบที่ตอบด้วยวิธีการอำพลางข้อมูล



ภาพที่ 34 เปรียบเทียบเปอร์เซ็นต์ความถูกต้องในการตอบด้วยคลาสหลักเสมอและการตอบด้วยวิธีการอำพลางข้อมูล

วิจารณ์

1. ประสิทธิภาพของตัวจำแนกประเภทข้อมูล SSCR

จากผลการทดลองเราสามารถสรุปได้ว่ากฎที่มีนัยสำคัญทางสถิติมีประสิทธิภาพอย่างมากกับชุดข้อมูลที่ไม่สมดุลสูง อีกทั้งการคิดต้นทุนความเสี่ยงการของกฎช่วยรับประกันความผิดพลาดประเภทลบ (FN) ได้ว่าจะเกิดขึ้นน้อยที่สุด ซึ่งเราได้เห็นประสิทธิภาพของความผิดพลาดประเภทลบแล้วในผลการทดสอบ จากมาตรวัดหลายๆ ตัวชี้ให้เห็นว่าอัลกอริทึม SSCR นั้นเหมาะสมกับข้อมูลที่มีความไม่สมดุลสูงและคลาสรองคือคลาสที่ให้ความสนใจเป็นพิเศษและต้องการความแม่นยำที่อยู่ในเกณฑ์ที่ดี

ปัญหาของความไม่สมดุลนั้นอาจจะมาจากหลากหลายปัจจัยแต่ในกรณีของปัญหาที่เกิดจากการมีข้อมูลตัวอย่างของคลาสรองน้อยมากๆ หรือมีตัวอย่างไม่เพียงพอ (Lack of Data) อัลกอริทึม SSCR ให้ประสิทธิภาพที่ดีในสถานการณ์ เช่นนี้ ซึ่งจะเห็นได้ชุดข้อมูล Abalone19 แต่ในกรณีที่ข้อมูลมีความไม่สมดุลสูงแต่มีตัวอย่างข้อมูลของคลาสรองเพียงพอเช่น คลาสรองมีตัวอย่างข้อมูล 100,000 ตัวอย่าง และมีตัวอย่างคลาสหลัก 10,000,000 ตัวอย่าง ซึ่งจะเห็นได้ว่ามีสัดส่วนความไม่สมดุลสูงมาก 1:100 แต่กรณีเช่นนี้ข้อมูลของทั้งสองคลาสต่างก็อธิบายคุณลักษณะการกระจายตัวของคลาสตัวเองได้อย่างสมบูรณ์ดีอยู่แล้ว ดังนั้นสถานการณ์เช่นนี้จึงอาจจะไม่เหมาะที่จะทำการจำแนกประเภทข้อมูลด้วยอัลกอริทึม SSCR เนื่องจากการคำนวณด้วย FET นั้นเป็นการคำนวณแบบแฟคทอเรียล ดังนั้นจำนวนตัวอย่างข้อมูลที่มีจำนวนมากจะส่งผลให้อัลกอริทึม SSCR อาจจะคำนวณข้อมูลแล้วเกิดความผิดพลาดได้อันเนื่องจากข้อจำกัดในประเภทของตัวแปรที่ใช้จัดเก็บข้อมูล (Data Type Overflow)

ถึงแม้ว่าอัลกอริทึม SSCR จะให้ประสิทธิภาพที่ดีกับคลาสรอง แต่ประสิทธิของคลาสหลักกลับยังทำได้ไม่ดีนัก ซึ่งสาเหตุส่วนหนึ่งเกิดจากการมีกฎความสัมพันธ์น้อยเกินไปยังไม่คลอบคลุมถึงลักษณะการกระจายตัวของข้อมูล อีกทั้งความไวของค่าความเสี่ยงอาจจะส่งผลให้กฎที่เป็นคลาสหลักมีโอกาสถูกเลือกไปทำนายลดลง ซึ่งประสิทธิภาพของคลาสหลักยังคงเป็นเรื่องที่ต้องปรับปรุง

2. การเลือกใช้สถิติสำหรับทดสอบนัยสำคัญในอัลกอริทึม SSCR

ในวิทยานิพนธ์เล่มนี้ได้พิจารณาเลือกใช้การทดสอบความถูกต้องของฟิชเชอร์ ซึ่งในทางสถิติแล้วยังมีการทดสอบชนิดอื่นๆ ที่สามารถนำมาทดสอบความเป็นอิสระของข้อมูลได้เช่นเดียวกันกับการทดสอบความถูกต้องของฟิชเชอร์ โดยในการพิจารณาเลือกสถิติทดสอบได้พิจารณาจากข้อดีข้อเสียดังต่อไปนี้

2.1. สถิติทดสอบความถูกต้องของฟิชเชอร์ (Fisher Exact Test) อาศัยการสรุปอ้างอิงจากการแจกแจงที่แท้จริง (Exact Distribution) ของกลุ่มตัวอย่างที่มีขนาดเล็ก มากกว่าการใช้วิธีประมาณการแจกแจงของกลุ่มตัวอย่างที่มีขนาดใหญ่ และมีวิธีการคำนวณค่านัยสำคัญทางสถิติด้วยการแจกแจงที่แท้จริงแบบไฮเปอร์ยีออเมตริก (Hypergeomatric Distribution) ซึ่งในการคำนวณความน่าจะเป็นจะต้องใช้หลักการจัดหมู่ (Combination) นั้นคืออาศัยวิธีการสุ่มแบบไม่ใส่กลับคืน (Sampling without Replacement) เรียกว่าการทดสอบ Exact Test และเรียกค่านัยสำคัญที่คำนวณได้ว่า Fisher's Exact Sig ซึ่งเป็นนัยสำคัญที่แท้จริงของสถิติทดสอบ และมีความถูกต้องมากกว่าการคำนวณค่านัยสำคัญจากวิธีการประมาณการแจกแจงของสถิติทดสอบให้เป็นแบบไคสแควร์

2.2. สถิติทดสอบเพียร์สันไคสแควร์ (Pearson Chi-Square) (Pearson, 1900) ซึ่งอาศัยการแจกแจงโดยประมาณแบบไคสแควร์สำหรับกลุ่มตัวอย่างที่มีขนาดใหญ่ โดยความถี่คาดหวังในแต่ละค่าของตารางแจกแจงความถี่จะต้องมีความถี่ที่มากกว่าหรือเท่ากับ 5 หรือถ้ามีจำนวนความถี่น้อยกว่า 5 แต่ต้องไม่เกิน 20% ของค่าในตารางแจกแจงความถี่ของทั้งหมด ซึ่งเป็นข้อกำหนดเบื้องต้นของสถิติทดสอบเพียร์สันไคสแควร์ ดังนั้นการที่มีข้อกำหนดให้กลุ่มตัวอย่างต้องมีขนาดใหญ่นั้นในปัญหาของความไม่สมดุลของคลาสบางครั้งอาจจะไม่สามารถทดสอบได้ถ้ากลุ่มตัวอย่างของกฎความสัมพันธ์นั้นมีตัวอย่างไม่มากพอหรือน้อยกว่า 5 และค่านัยสำคัญของสถิติทดสอบแบบไคสแควร์นั้นเป็นค่านัยสำคัญโดยประมาณ เรียกว่า Asymptotic Sig

2.3. สถิติทดสอบแม็กซิมั่มไลค์ลิฮูดเรโชไคสแควร์ (Likelihood Ratio Chi-Square) (Neyman and Pearson, 1933) เป็นสถิติทดสอบที่มีคุณสมบัติหลายอย่างร่วมกันกับเพียร์สันไคสแควร์ถึงแม้ว่าจะเป็นสถิติทดสอบคนละตัวกันก็ตาม ซึ่งให้ผลสรุปที่เหมือนกันเมื่อสมมุติฐานหลักเป็นจริง และจำนวนความถี่ในตารางข้อมูลที่มีจำนวนมาก ค่าสถิติทดสอบของแม็กซิมั่มไลค์ลิฮูดเรโชไคสแควร์จะมีการแจกแจงไคสแควร์เหมือนกัน และมีค่าสถิติที่ใกล้เคียงกัน

ดังนั้นจากสถิติทดสอบที่กล่าวมาจะเห็นได้ว่าการทดสอบความถูกต้องของฟิชเชอร์นั้นมีความเหมาะสมกับปัญหาความไม่สมดุลของคลาสมากกว่า เนื่องจากค่านัยสำคัญที่คำนวณได้เป็นค่านัยสำคัญที่แท้จริงไม่ใช่ค่านัยสำคัญประมาณการแบบไคสแควร์

3. ปัจจัยที่กระทบต่อประสิทธิภาพของอัลกอริทึม SSCR

3.1. จำนวนของไอเท็มและแอตทริบิวต์ อัลกอริทึม SSCR ใช้โครงสร้างข้อมูลที่เป็นต้นไม้พรีฟิก และบิทเวคเตอร์ เพื่อการทำงานที่รวดเร็ว ดังนั้นจำนวนไอเท็มที่มีขนาดใหญ่จะทำให้มีการใช้หน่วยความจำมากขึ้น และจำนวนแอตทริบิวต์ที่มีหลายแอตทริบิวต์จะทำให้เกิดกฎความสัมพันธ์จำนวนมากมายซึ่งอาจจะส่งผลต่อเวลาการคำนวณและหน่วยความจำที่ต้องใช้ ซึ่งจะส่งผลให้อัลกอริทึมมีประสิทธิภาพด้านเวลาที่แย่ลง

3.2. การกำหนดค่าวิกฤต α ในทางสถิติแล้วค่าที่เหมาะสมคือ 0.05 แต่ในชุดข้อมูลแต่ละชุดมีการกระจายตัวของข้อมูลที่ไม่เหมือนกัน ดังนั้นจึงต้องกำหนดค่าวิกฤตให้เหมาะสม การกำหนดค่าวิกฤตมากเกินไป จะส่งผลให้มีกฎความสัมพันธ์จำนวนมากมาย ซึ่งอาจจะแวดล้อมไปด้วยกฎที่ทำนายคลาสหลัก หากกำหนดค่าวิกฤตน้อยเกินไปก็จะทำให้มีจำนวนกฎความสัมพันธ์น้อยเกินไป จนไม่สอดคล้องกับการกระจายตัวของข้อมูล ซึ่งค่าวิกฤตนี้จะส่งผลโดยตรงกับประสิทธิภาพการทำนายของอัลกอริทึม

3.3. ตารางเมตริกต้นทุน เนื่องจากอัลกอริทึม SSCR อาศัยแนวคิดของการเรียนรู้แบบมีต้นทุนมาเพื่อวิเคราะห์ความเสี่ยงของกฎความสัมพันธ์ ดังนั้นการกำหนดค่าเมตริกต้นทุนจะส่งผลกับลำดับความสำคัญของกฎความสัมพันธ์ และก็เป็นเรื่องที่ยากที่จะกำหนดค่าตารางเมตริกต้นทุนให้เหมาะสมกับชุดข้อมูลแต่ละตัว จากการทดสอบผลกระทบของเมตริกต้นทุนจะเห็นว่า แต่ละชุดข้อมูลมีค่าเมตริกต้นทุนที่ให้ประสิทธิภาพสูงสุดแก่อัลกอริทึม SSCR ไม่เท่ากัน ซึ่งนี่ก็เป็นอีกปัจจัยหนึ่งที่กระทบต่อประสิทธิภาพโดยตรง

สรุปและข้อเสนอแนะ

สรุป

ผู้วิจัยได้นำเสนอแนวทางการปรับปรุงประสิทธิภาพของการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์บนฐานข้อมูลที่ไม่สมดุล โดยมุ่งแน้นการลดความผิดพลาดประเภทลบ FN ให้น้อยที่สุด โดยได้อาศัยการทดสอบทางสถิติด้วย FET มาคัดเลือกกฎความสัมพันธ์ให้มีคุณภาพที่ดีและมีนัยสำคัญทางสถิติ ร่วมกับการวิเคราะห์ค่าความเสี่ยงของกฎความสัมพันธ์เพื่อลด FN ซึ่งจากผลการทดลองจะเห็นว่า แนวทางที่นำเสนอมีประสิทธิภาพที่ดีบนฐานข้อมูลที่ไม่สมดุลอีกทั้งมีความแม่นยำสูงบนคลาสรองที่เราสนใจ ดังนั้น SSCR จึงเป็นอัลกอริทึมที่เหมาะกับข้อมูลที่มีความไม่สมดุลสูง แต่อัลกอริทึมเองยังมีข้อด้อยในเรื่องประสิทธิภาพของคลาสหลัก ซึ่งยังมีประสิทธิภาพไม่ดีนักเมื่อเทียบกับ CBA และ C4.5 อีกทั้งการเลือกค่าเมตริกต้นทุนที่เหมาะสมก็ยังคงเป็นเรื่องที่ต้องทำการปรับปรุงต่อไป

ข้อเสนอแนะ

อัลกอริทึม SSCR สามารถพัฒนาปรับปรุงให้มีประสิทธิภาพที่ดีขึ้นได้อีก ดังนี้

1. พัฒนาอัลกอริทึมให้มีความสามารถหาค่าตารางเมตริกต้นทุนที่เหมาะสมได้เอง เช่น การใช้ Genetic Algorithm (Goldberg, 1989) มาช่วยในการหาค่าที่เหมาะสมจากสแปซคำตอบทั้งหมด 1 เมตริกต้นทุนคือ 1 คำตอบที่เป็นไปได้ในสแปซคำตอบทั้งหมด (Solutions Space)

2. ปรับปรุงวิธีการคำนวณต้นทุนความเสี่ยงของกฎความสัมพันธ์ให้มีประสิทธิภาพมากยิ่งขึ้นทั้งคลาสรองและคลาสหลัก เช่นการใช้ Risk Fuction ตามกฎของเบย์มาคำนวณค่าความเสี่ยง การพิจารณาค่าความเสี่ยงของกฎร่วมกับค่า p_{value} ที่ได้จากการทดสอบด้วย FET

3. ปรับปรุงวิธีการควบคุมคุณภาพของกฎความสัมพันธ์เพื่อให้มีกฎความสัมพันธ์ที่มากเพียงพอทั้งคลาสหลักและคลาสรอง เช่นการพิจารณา FET ร่วมกับมาตรวัดอื่นๆ เพิ่มเติม

เอกสารและสิ่งอ้างอิง

พูนเพิ่ม สุวรรณรัฐภูมิ และ กฤษณะ ไวยมัย. 2555. แนวทางการปรับปรุงประสิทธิภาพของการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์บนฐานข้อมูลที่ไม่สมดุล. **วิศวกรรมสาร มก.** 25 (79): 36-49.

พูนเพิ่ม สุวรรณรัฐภูมิ และ กฤษณะ ไวยมัย. 2554. แนวทางการปรับปรุงประสิทธิภาพของการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์บนฐานข้อมูลที่ไม่สมดุล, น. 37-42. *ใน* **The 2011 International Computer Science and Engineering Conference (ICSEC 2011).**

Agrawal, R. and R. Srikant. 1994. Fast algorithm for mining association rules in large databases. pp. 487-499. *In* **Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94).** Santiago, Chile.

Agrawal R., T. Imielinski and A. 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.

Arunasalam, B. and S. Chawla. 2006 Cccs: a top-down associative classifier for imbalanced class distribution. pp. 517–522. *In* **Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining.** ACM Press. New York, NY, USA.

Borgelt, C. 2010. **Frequent Pattern Mining.** Available Source: http://www.borgelt.net/teach/fpm/, October 04, 2010.

Fisher, R. A. 1922. On the interpretation of X^2 from contingency tables, and the calculation of P. **Journal of the Royal Statistical Society.** 85(1): 87–94.

Frank, A. and A. Asuncion. 2010. **UCI Machine Learning Repository.** Irvine, CA: University of California, School of Information and Computer Science. Available Source: http://archive.ics.uci.edu/ml.

Goldberg, D., E. 1989. Genetic Algorithms in Search Optimization and Machine Learning. **Addison Wesley.**

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten. 2009. The WEKA Data Mining Software: An Update; **SIGKDD Explorations.** 11(1).

Liu, B., W. Hsu 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.

Neyman, J. And E. Pearson. 1933. On the Problem of the Most Efficient Tests of Statistical Hypotheses. **Philosophical Transactions of the Royal Society of London. Series A.** 231: 289-337.

Pearson, K. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to hove arisen from random sampling. **Philosophical Magazine Series 5.** 50(302): 157–175.

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

Silverstein, C., S. Brin and R. Motwani. 1998. Beyond market baskets: Generalizing association rules to dependence rules. **Data Mining and Knowledge Discovery.** 2 (1): 39-68.

Tan, P., M. Steinbach and V. Kumar. 2006. **Introduction to Data Mining**. Pearson International Edition, Addison-Wesley.

Verhein, F. and S. Chawla. 2007. Using Significant, Positively Associated and Relatively Class Correlated Rules for Associative Classification of Imbalanced Datasets. pp. 679-684. *In* **Seventh IEEE International Conference on 28-31 Oct.** IEEE Computer Society Press, Los Alamitos.

Webb, G. I. 2006. Discovering significant rules. pp. 434–443. *In* **KDD '06: Proceedings of the 12th ACMSIGKDD international conference on Knowledge discovery and data mining.** ACM Press. New York, NY, USA.

Weiss, G. M. 2004. Mining with rarity: a unifying framework. *In* **SIGKDD Explorations Special Issue on Learning from Imbalanced Datasets.** 6(1): 7-19.

Zaki, M. J, S. Parthasarathy, M. Ogihara and W. Li. 1997. New Algorithms for Fast Discovery of Association Rules. pp.283-286. *In* **Proceedings of KDD'1997.**

ประวัติการศึกษาและการทำงาน

ชื่อ	นายพูนเพิ่ม สุวรรณรัฐภูมิ
เกิดวันที่	5 กันยายน 2516
สถานที่เกิด	ภูเก็ต
ประวัติการศึกษา	วศ.บ. (วิศวกรรมคอมพิวเตอร์) คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้า เจ้าคุณทหาร ลาดกระบัง (2540)
ตำแหน่งปัจจุบัน	นักวิเคราะห์ระบบอาวุโส
สถานที่ทำงานปัจจุบัน	บริษัท ซอฟต์แสควร์อินเตอร์เนชั่นเนล จำกัด
ผลงานดีเด่นและ/หรือรางวัลทางวิชาการ	งานวิจัยเรื่องแนวทางการปรับปรุงประสิทธิภาพของการจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์บนฐานข้อมูลที่ไม่สมดุล
ทุนการศึกษาที่ได้รับ	-