ตัวปรับแบบยึดเกาะในต้นไม้ตัดสินใจสำหรับเซตข้อมูลไม่สมดุล

นางสาวอุไรรัตน์ กฤษดาวาณิชย์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2553

AN ADHESIVE MODIFIER IN DECISION TREES FOR IMBALANCED DATA SETS

Miss Urairat Kritsadawanit

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computer Science

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2010

หัวข้อวิทยานิพนธ์	ตัวปรับแบบยึดเกาะในต้นไม้ตัดสินใจสำหรับเซตข้อมูลไม่สมดุล
โดย	นางสาวอุไรรัตน์ กฤษดาวาณิชย์
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.บุญสม เลิศหิรัญวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ)

..... กรรมการ
(รองศาสตราจารย์ ดร.ญาใจ ลิ่มปิยะกรณ์)

..... กรรมการภายนอกมหาวิทยาลัย
(ผู้ช่วยศาสตราจารย์ ดร.รัชฎา คงคะจันทร์)

อุไรรัตน์ กฤษดาวาณิชย์ : ตัวปรับแบบยึดเกาะในต้นไม้ตัดสินใจสำหรับเซตข้อมูลไม่สมดุล. (AN ADHESIVE MODIFIER IN DECISION TREES FOR IMBALANCED DATA SETS) อ. ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ.ดร.สุกรี สินธุภิญโญ, 52 หน้า.

ต้นไม้ตัดสินใจเป็นเทคนิคการจำแนกข้อมูลที่ใช้กันอย่างแพร่หลายทางด้านการทำเหมืองข้อมูล การสร้างต้นไม้ตัดสินใจสามารถสร้างได้หลายรูปแบบขึ้นอยู่กับการเลือกตัววัดความสามารถในการแบ่งแยกข้อมูล วิธีหนึ่งที่นิยมนำมาใช้ คือ ID3 ซึ่งเลือกคุณลักษณะบนพื้นฐานของทฤษฎีสารสนเทศ และ C4.5 ได้พัฒนาต่อมาจาก ID3 ใช้ทฤษฎีสารสนเทศเช่นเดียวกับ ID3 และได้แก้ปัญหาไบแอสด้วยค่าสารสนเทศการแบ่งแยก ทั้งสองวิธีนี้รวดเร็วและเข้าใจง่ายเมื่อเทียบกับวิธีอื่น ๆ และเหมาะกับข้อมูลที่มีการกระจายแบบสมดุล แต่เมื่อนำมาจำแนกข้อมูลไม่สมดุล การเลือกคุณลักษณะจะให้ความสำคัญกับกลุ่มที่มีจำนวนตัวอย่างมาก ไม่สนใจกลุ่มที่มีจำนวนตัวอย่างน้อย ทำให้ได้ผลการทำนายสูงในกลุ่มที่มีมาก แต่ให้ผลการทำนายต่ำในกลุ่มที่มีน้อย

วิทยานิพนธ์ฉบับนี้จึงนำเสนอเอนโทรปีแบบใหม่สำหรับต้นไม้ตัดสินใจ โดยใช้วิธี C4.5 เป็นพื้นฐาน สำหรับการเรียนรู้ข้อมูลแบบสองกลุ่ม จุดประสงค์คือ เพื่อจำแนกตัวอย่างน้อยให้ดีขึ้น ในการทดลองนั้นใช้การทดสอบแบบไขว้ข้าม 5 กลุ่มกับ 16 ชุดข้อมูลไม่สมดุล และเปรียบเทียบผลการทดลองกับอัลกอริทึม C4.5, เอนโทรปีแบบอสมมาตร และเอนโทรปีแบบออกจากศูนย์กลาง ทดสอบประสิทธิภาพด้วยค่าความระลึก ค่าความเที่ยง และค่าเอฟ ซึ่งคำนวณได้จากตารางคอนฟิวชันเมตริกซ์ จากผลการทดลองพบว่าวิธีการที่นำเสนอสามารถสร้างกฎของกลุ่มที่มีน้อยได้ดีกว่าวิธีอื่นจึงทำให้จำแนกตัวอย่างในกลุ่มที่มีน้อยได้ดี

ภาควิชาวิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต อุไรรัตน์ กฤษดาวาณิชย์
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์ สุกรี สินธุภิญโญ
ปีการศึกษา 2553

5170536021 : MAJOR COMPUTER SCIENCE

KEYWORDS : DECISION TREES / ENTROPY / IMBALANCED DATA

URAIRAT KRITSADAWANIT : AN ADHESIVE MODIFIER IN DECISION TREES FOR IMBALANCED DATA SETS. THESIS ADVISOR: ASST. PROF. SUKREE SINTHUPINYO, Ph.D., 52 pp.

In data mining research, decision tree is a famous method for classification. It can build different forms of decision trees based on selected splitting attribute. One of the most famous algorithms is ID3, in which choice of splitting attributes is based on information theory. C4.5 is an improvement of ID3 which, in the same way as ID3, constructs a decision trees using information theory but reducing the bias of ID3 by splitting information. Both are relatively fast and easily understood. However they are suitable only for the balanced class distribution, we cannot achieve good results on imbalanced data set.

In this paper, we present a new entropy measure based on C4.5 method for decision trees learning on two-class data sets. We need a prediction model, which can improve the accuracy of the minority class. In our experiments, we tested our algorithm on 16 datasets using five-fold cross-validation method. We compared the results to C4.5, Asymmetric Entropy and Off-Center Entropy. Recall, precision, and f-measure were computed. The results show that the proposed method can construct the better rules which finally improve the accuracy of the minority class data.

Department : Computer Engineering Student's Signature อัเรรัตน์ กฤษดาวาณิชย์

Field of Study : Computer Science Advisor's Signature

Academic Year : 2010

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงด้วยความช่วยเหลือของ ผศ.ดร.สุกรี สินธุภิญโญ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ช่วยแนะนำแนวทางการทำวิจัย ให้คำปรึกษา และข้อคิดเห็นที่เป็นประโยชน์ต่องานวิจัยมาด้วยดีตลอด

ขอขอบพระคุณคณะกรรมการสอบวิทยานิพนธ์ ศ.ดร.บุญเสริม กิจสิริกุล รศ.ดร. ญาใจ ลิ่มปิยะกรณ์ และ ผศ.ดร. รัชฎา คงคะจันทร์ ที่สละเวลามาตรวจ ให้ข้อเสนอแนะ และให้ข้อคิดเห็นที่เป็นประโยชน์

ขอขอบคุณคณาจารย์จุฬาลงกรณ์มหาวิทยาลัยที่ได้ให้ความรู้ในด้านวิชาการ ขอขอบคุณเพื่อนๆ พี่ๆ น้องๆ สมาชิกห้องปฏิบัติการอัจฉริยะภาพเครื่องกล และการค้นพบความรู้ (MIND LAB) ที่คอยช่วยเหลือ ให้คำปรึกษา ให้ความรู้ต่างๆ และช่วยวิจารณ์ผลงาน

สุดท้ายนี้ ขอขอบคุณสมาชิกทุกคนในครอบครัว ที่สนับสนุน และให้กำลังใจเสมอมา จนสามารถทำวิทยานิพนธ์นี้ได้เสร็จสมบูรณ์

สารบัญ

หน้า

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1
บทนำ

1.1. ที่มาและความสำคัญของปัญหา

ในปัจจุบันฐานข้อมูลได้เพิ่มจำนวนมากขึ้น การจะนำข้อมูลมาทำให้เกิดประโยชน์โดยการวิเคราะห์ข้อมูลจำเป็นต้องใช้แรงงานคนจำนวนมาก และไม่สามารถใช้แรงงานคนในการวิเคราะห์ข้อมูลได้ทันเวลา อีกทั้งยังมีข้อมูลใหม่ ๆ ทำให้ต้องเสียเวลาในการวิเคราะห์ใหม่ และอาจเกิดความผิดพลาดจากผู้วิเคราะห์เอง ดังนั้นการเรียนรู้ของเครื่องจึงกลายเป็นส่วนหนึ่งในการทำงานแทนมนุษย์ ซึ่งเป็นการทำงานโดยใช้คอมพิวเตอร์มาช่วยในการวิเคราะห์ข้อมูล เพื่อความสะดวกสบายและความรวดเร็ว โดยการให้คำสั่งกับคอมพิวเตอร์ และให้คอมพิวเตอร์วิเคราะห์ข้อมูลได้นำไปประยุกต์ใช้หลาย ๆ ด้าน เช่น

ด้านการเงิน:	ใช้ในการตัดสินใจอนุมัติบัตรเครดิตให้ลูกค้า การพยากรณ์หุ้น
ด้านภูมิศาสตร์:	ใช้ในการพยากรณ์อากาศ
ด้านการแพทย์:	ใช้ในการวินิจฉัยโรค
ด้านการเกษตร	ใช้ในการจำแนกโรคพืช
ด้านการพิสูจน์ตัวตน:	การรู้จำตัวอักษร การรู้จำเสียง

ต้นไม้ตัดสินใจเป็นการเรียนรู้ของเครื่องประเภทหนึ่งที่นิยมนำมาใช้ในการจำแนกข้อมูล เพราะเป็นวิธีที่มนุษย์สามารถเข้าใจได้ง่าย การเรียนรู้มีความรวดเร็วเมื่อเทียบกับอัลกอริทึมประเภทอื่น ๆ อีกทั้งยังมีความทนทานต่อสิ่งรบกวนได้เป็นอย่างดี และสามารถนำแบบจำลองที่ได้มาเขียนให้อยู่ในรูปของกฎได้ การเรียนรู้เป็นแบบมีผู้สอนโดยใช้คุณลักษณะในการแบ่งแยกข้อมูล โดยให้ความสำคัญแต่ละคุณลักษณะต่างกัน อัลกอริทึมพื้นฐานที่นิยมนำมาใช้เพื่อเลือกคุณลักษณะเช่น ID3 [1] นำเสนอค่าเกนมาตรฐานในการเลือกคุณลักษณะมาเป็นโหนด ซึ่งค่าเกนมาตรฐานนี้บ่งบอกว่าคุณลักษณะนั้นจำแนกข้อมูลได้ดีเพียงใด โดยทดลองทุก ๆ คุณลักษณะ ถ้าคุณลักษณะใดให้ค่าเกนมาตรฐานสูงสุดแสดงว่าคุณลักษณะนั้นจำแนกข้อมูลได้ดีสุด ค่าเกนมาตรฐานนี้สามารถคำนวณได้จากทฤษฎีสารสนเทศ (information theory) [2] และได้ถูกพัฒนาต่อมาเป็น C4.5 [3] ซึ่งเสนอค่าอัตราส่วนเกน และได้เพิ่มค่าสารสนเทศการแบ่งแยก

ข้อมูลไม่สมดุล (Imbalanced data) [4] เกิดขึ้นเมื่อข้อมูลมีจำนวนตัวอย่างของกลุ่มหนึ่งน้อยกว่ากลุ่มอื่นมาก ๆ ซึ่งมักจะเกิดขึ้นบ่อยกับข้อมูลที่หายาก ตัวอย่างเช่น ข้อมูลการวินิจฉัยโรคทางการแพทย์ ซึ่งมีจำนวนผู้ป่วยน้อยกว่าผู้มีสุขภาพดี ข้อมูลการใช้งานบัตรเครดิตซึ่งมี

คนโกงน้อยกว่าคนใช้งานปกติ ข้อมูลการบุกรุกเครือข่ายซึ่งมีจำนวนผู้บุรุกน้อยกว่าผู้ใช้งานปกติ เมื่อสร้างต้นไม้ตัดสินใจด้วยวิธีการพื้นฐานที่ได้กล่าวไว้ข้างต้น เพื่อจำแนกข้อมูลไม่สมดุล พบว่าการจำแนกโน้มเอียงไปยังกลุ่มที่มีมาก จึงได้ผลการทำนายสูงในกลุ่มที่มีมาก และได้ผลการทำนายต่ำในกลุ่มที่มีน้อย

งานวิจัยที่เสนอแนวคิดในการแก้ปัญหาการจำแนกข้อมูลไม่สมดุลนั้น อาจแบ่งเป็น 4 ประเภทใหญ่ ๆ คือ 1) พัฒนาเครื่องมือวัดประสิทธิภาพการเรียนรู้ของอัลกอริทึม 2) เปลี่ยนการกระจายของข้อมูลการเรียนรู้ก่อนที่จะเรียนรู้ด้วยอัลกอริทึมพื้นฐาน 3) เปลี่ยนการวัดค่าเอนโทรปี เพื่อเลือกคุณลักษณะมาสร้างต้นไม้ตัดสินใจ 4) การผสานหลาย ๆ อัลกอริทึม

วิทยานิพนธ์ฉบับนี้ใช้วิธีการเปลี่ยนการหาค่าเอนโทรปี แต่ยังคงใช้วิธีพื้นฐานของอัลกอริทึมต้นไม้ตัดสินใจที่มีอยู่แล้วทางด้านการทำเหมืองข้อมูล เพื่อจำแนกตัวอย่างในกลุ่มที่มีน้อยให้ดีขึ้นสำหรับข้อมูลไม่สมดุลที่มีสองกลุ่ม โดยการใช้ตัวปรับที่สามารถปรับตัวเองให้สัมพันธ์กับความน่าจะเป็นของทั้งสองกลุ่มในการสร้างต้นไม้ตัดสินใจในแต่ละระดับ เพื่อให้ความสำคัญกับกลุ่มที่มีน้อย

1.2. วัตถุประสงค์ของการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อเพิ่มประสิทธิภาพของต้นไม้ตัดสินใจสำหรับจำแนกข้อมูลที่มีความไม่สมดุล โดยการปรับเปลี่ยนวิธีการหาค่าเอนโทรปีแบบใหม่บนพื้นฐานอัลกอริทึมที่มีอยู่แล้วทางด้านการทำเหมืองข้อมูล ที่สามารถจำแนกได้ดีกับตัวอย่างในกลุ่มที่มีน้อย เพื่อให้สามารถนำไปใช้กับข้อมูลด้านต่าง ๆ อีกทั้งยังเป็นพื้นฐานของงานวิจัยอื่น ๆ ต่อไป

1.3. ขอบเขตของการวิจัย

1) งานวิจัยนี้เสนอวิธีการจำแนกข้อมูลไม่สมดุลที่มีสองกลุ่ม

2) ข้อมูลที่ใช้ในการวิจัยได้คัดเลือกชุดข้อมูลมาจาก UCI Machine Learning Repository [23] 16 ชุดข้อมูล สำหรับข้อมูลที่มีมากกว่าสองกลุ่ม ได้เลือกกลุ่มที่มีน้อยเป็นกลุ่มหนึ่ง และให้กลุ่มที่เหลือรวมเป็นอีกกลุ่มหนึ่ง

3) ผลการทดลองที่แสดงเป็นค่าเฉลี่ยจากการทดสอบแบบไขว้ข้าม 5 ครั้ง (five-fold cross validation)

4) ใช้ค่าความแม่นยำและค่าเอฟเป็นเกณฑ์ในการวัดประสิทธิภาพของตัวจำแนก

1.4. ขั้นตอนและวิธีดำเนินการวิจัย

1) ศึกษาเกี่ยวกับปัญหาข้อมูลไม่สมดุล
2) ศึกษางานวิจัยที่เกี่ยวข้องกับปัญหาข้อมูลไม่สมดุล
3) กำหนดหัวข้อปัญหาของงานวิจัย
4) ศึกษาทฤษฎีการต้นไม้ตัดสินใจ
5) หาแนวทางการพัฒนาอัลกอริทึมพื้นฐาน
6) ศึกษาเครื่องมือ และซอฟต์แวร์ที่ใช้สำหรับงานวิจัย
7) พัฒนาเครื่องมือทดสอบ
8) ทำการทดลองแนวคิดและสมมติฐาน
9) วิเคราะห์และสรุปผลผลการทดลอง
10) เรียบเรียงวิทยานิพนธ์

1.5. คุณค่าทางวิชาการ

1) นำเสนอวิธีการจำแนกข้อมูลไม่สมดุล เพื่อจำแนกข้อมูลในกลุ่มที่มีน้อย
2) สามารถนำเทคนิคการจำแนกข้อมูลไปประยุกต์ใช้กับปัญหาจริงต่าง ๆ ได้
3) เป็นพื้นฐานของงานวิจัยอื่น ๆ ในอนาคต

1.6. ผลงานตีพิมพ์จากวิทยานิพนธ์

1) หัวเรื่อง “An Adhesive Modifier in Decision Trees for Imbalanced data sets” โดย อุไรรัตน์ กฤษดาวาณิชย์ และ สุกรี สินธุภิญโญ ในบันทึกการประชุม “International Joint Conference on Computer Science and Software Engineering (JCSSE 2011)” ซึ่งจัดขึ้น ณ จังหวัดนครปฐม ประเทศไทย ระหว่างวันที่ 11-13 พฤษภาคม 2554

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1. ทฤษฎีที่เกี่ยวข้อง

2.1.1. ต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจ (decision tree) คือ แผนผังต้นไม้ที่ช่วยในการตัดสินใจโดยใช้ทฤษฎีทางด้านคณิตศาสตร์ เป็นการเรียนรู้แบบมีผู้สอน (supervised learning) นิยมนำมาจำแนกข้อมูล เนื่องจากมีโครงสร้างที่เข้าใจง่าย การเรียนรู้เร็ว และยังสามารถแทนความรู้ในรูปของกฎ “IF THEN”

ส่วนประกอบของต้นไม้ตัดสินใจ

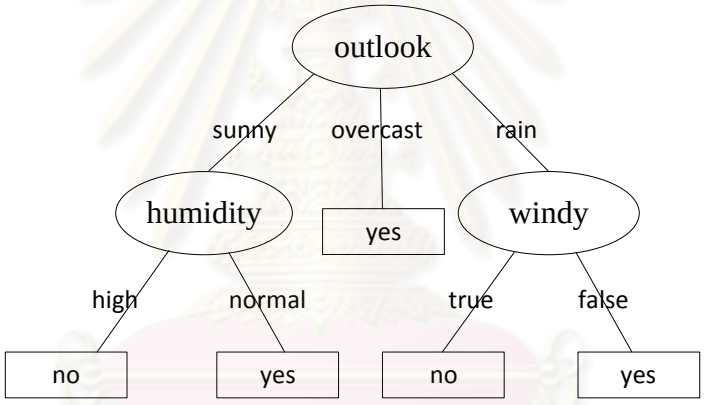
ต้นไม้ตัดสินใจมีลักษณะเหมือนโครงสร้างต้นไม้ ซึ่งภายในต้นไม้ประกอบไปด้วยโหนด (node) แต่ละโหนดแทนคุณลักษณะที่ใช้ในกระบวนการเรียนรู้ข้อมูล โดยโหนดบนสุดของต้นไม้เรียกว่าโหนดราก (root node) ส่วนกิ่ง (link) ของต้นไม้แทนค่าที่เป็นไปได้ของคุณลักษณะ และใบเป็นส่วนที่อยู่ล่างสุดของต้นไม้ซึ่งแสดงถึงกลุ่มหรือผลลัพธ์ที่ได้จากการทำนายข้อมูลนั้น ตัวอย่างสภาพอากาศในการตัดสินใจเล่นกอล์ฟ Quinlan [1] ตารางที่ 2-1 เป็นตัวอย่างชุดข้อมูลเรียนรู้ในการตัดสินใจออกไปตีกอล์ฟ โดยพิจารณาจากคุณลักษณะ outlook, temperature, humidity และ windy และมีคลาสคำตอบคือ yes และ no

ตารางที่ 2-1 ชุดข้อมูลเรียนรู้ในการตัดสินใจออกไปตีกอล์ฟ

Attributes				Class
Outlook	Temperature	Humidity	Windy	
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no

sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

เซตของข้อมูลเรียนรู้มีข้อมูลทั้งหมด 14 ระเบียนประกอบด้วย 4 คุณลักษณะ คือ Outlook, Temperature, Humidity, และ Windy คุณลักษณะ Outlook มีได้ 3 ค่าคือ {sunny, overcast, rain},Temperature มีได้ 3 ค่า คือ {cool, mild, hot}, Humidity มีได้ 3 ค่า คือ {high, normal} และ Windy มีได้ 2 ค่า คือ {true, false } สามารถสร้างแผนผังต้นไม้ตัดสินใจแบบง่าย ๆ ดังภาพที่ 2-1



ภาพที่ 2-1 แผนภาพต้นไม้ตัดสินใจ

การสร้างต้นไม้ตัดสินใจ

การสร้างต้นไม้ตัดสินใจสร้างทีละโหนดจากบนลงล่างแบบตะกราม (top-down greedy search) คือจะเริ่มจากโหนดรากต้นไม้แล้วแตกกิ่งไปเรื่อย ๆ จนกระทั่งถึงโหนดใบ แสดงความรู้ในรูปกฎ IF-THEN ได้ การสร้างต้นไม้ตัดสินใจ เพื่อใช้ในการจำแนกกลุ่มข้อมูลนั้น สามารถสร้างได้หลายรูปแบบขึ้นอยู่กับการเลือกคุณลักษณะในการแบ่งแยก ซึ่งมีหลายวิธีในการเลือกคุณลักษณะ อัลกอริทึมพื้นฐานที่นิยมนำมาใช้ เช่น ID3 และ C4.5 โดยขั้นตอนการสร้างต้นไม้ตัดสินใจมีดังนี้ Han, J. and Kamber, M. [5]

1) ต้นไม้จะเริ่มต้นด้วยโหนดเพียงโหนดเดียว ซึ่งเรียกว่า โหนดราก (root node)
2) ถ้าข้อมูลทั้งหมดในโหนดอยู่ในกลุ่มเดียวกัน ให้โหนดนั้นเป็นโหนดใบ

3) ถ้าข้อมูลทั้งหมดในโหนดยังไม่เป็นระเบียบหรือมีข้อมูลหลายกลุ่มปะปนกันอยู่ ต้องมีการคัดเลือกคุณลักษณะที่ดีที่สุดสำหรับการแบ่งแยกข้อมูลออกเป็นกลุ่ม ๆ เช่น ใช้อัลกอริทึมพื้นฐาน ID3 โดยใช้ค่าเกนมาตรฐาน (gain) ของแต่ละคุณลักษณะโดยคุณลักษณะที่มีค่าเกนมาตรฐานมากที่สุดจะถูกเลือกให้เป็นคุณลักษณะในการตัดสินใจ
4) เมื่อได้คุณลักษณะที่ดีที่สุดแล้ว คุณลักษณะนั้นจะถูกนำมาสร้างเป็นโหนดราก
5) ทำการวนซ้ำเพื่อหาคุณลักษณะที่มีค่าเกนมาตรฐานมากที่สุด เพื่อนำคุณลักษณะนี้มาสร้างเป็นโหนดในระดับถัดไป และแตกกิ่งตามค่าต่าง ๆ ของโหนดเรียนรู้นั้น จะสิ้นสุดการทำก็ต่อเมื่อ ข้อมูลทุกตัวในโหนดนั้นอยู่กลุ่มเดียวกัน หรือใช้คุณลักษณะในการเรียนรู้ครบทุกตัวแล้ว

การแทนความรู้ในรูปของกฎต้นไม้ตัดสินใจ

เมื่อสร้างต้นไม้ตัดสินใจแล้ว สามารถสร้างกฎของต้นไม้ตัดสินใจในรูป “IF THEN”จะเริ่มจากโหนดรากไปยังโหนดใบ หลัง IF ให้ใส่โหนดทดสอบ และค่าของโหนดนั้น ถ้ามีหลายเงื่อนไขให้เชื่อมด้วย AND เมื่อเจอใบให้ใส่ใบหลัง THEN จากภาพที่ 2-1 สามารถเขียนเป็นกฎได้ 5 กฎ ดังนี้

1) IF outlook=sunny AND humidity=high THEN class=no
2) IF outlook=sunny AND humidity=normal THEN class=yes
3) IF outlook=overcast THEN class=yes
4) IF outlook=rain AND windy= true THEN class=no
5) IF outlook=rain AND windy= false THEN class=yes

2.1.2. **อัลกอริทึม** ID3

เนื่องจากการสร้างต้นไม้ตัดสินใจในข้อมูลชุดเดียวกัน สามารถสร้างต้นไม้ได้หลายแบบขึ้นอยู่กับการเลือกคุณลักษณะมาเป็นตัวแบ่งแยก ดังนั้นจึงต้องมีเกณฑ์ในการวัดค่าความสามารถในการแบ่งแยกของแต่ละคุณลักษณะเช่น ID3 [3] เป็นอัลกอริทึมพื้นฐานในการสร้างต้นไม้ตัดสินใจที่นิยมนำมาใช้ในการจำแนกข้อมูล ในอัลกอริทึม ID3 นั้นได้มีการใช้ค่าเกนมาตรฐาน (gain) ในการคัดเลือก โดยจะหาค่าเกนมาตรฐานของทุก ๆ คุณลักษณะ ถ้าคุณลักษณะใดมีค่าเกนมาตรฐานมากที่สุดจะถูกเลือกให้เป็นโหนดในการตัดสินใจซึ่งค่าเกนมาตรฐานนี้สามารถคำนวณได้จากทฤษฎีสารสนเทศ (information theory) [2] ที่กล่าวไว้ว่าค่าสารสนเทศของข้อมูลขึ้นอยู่กับความน่าจะเป็นของข้อมูลที่อยู่ในรูปของบิต ค่าเอนโทรปีแสดงถึงปริมาณข้อมูลที่ต้องการเพื่อจำแนกกลุ่มข้อมูล ซึ่งคำนวณโดยใช้ทฤษฎีสารสนเทศ กำหนดตัวแปรต่าง ๆ ได้ดังนี้

m แทนจำนวนกลุ่มข้อมูลที่ต่างกัน

S แทนเซตข้อมูลทั้งหมด $S = \{s_1, s_2, s_3, \dots, s_m\}$

C_i แทนกลุ่มในลำดับที่ i โดยที่ $1 < i < m$

S_i แทนจำนวนที่เป็นสมาชิกของ S และอยู่ในกลุ่ม C_i

S_{ij} แทนจำนวนที่เป็นสมาชิกของ S และอยู่ในกลุ่ม C_i จากการแบ่งข้อมูลด้วยค่าที่เป็นไปได้ j ของคุณลักษณะ A โดยที่ $1 \leq j \leq v$

$P(S_i)$ แทนความน่าจะเป็นในการเกิดค่า S_i

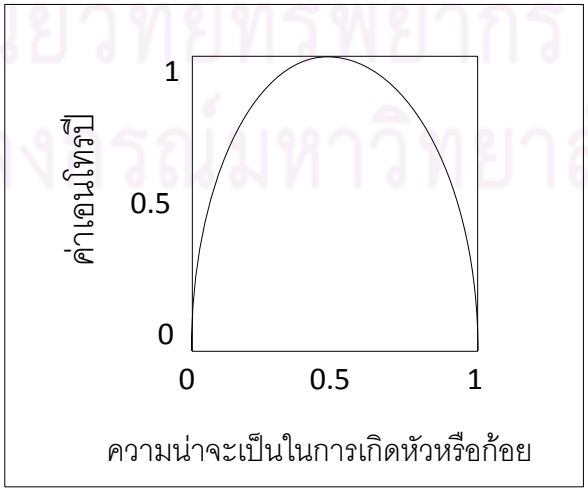
ค่าสารสนเทศที่ใช้ในการจำแนกประเภทข้อมูล สามารถคำนวณได้ดังสมการที่ (1)

$$I(S_1, S_2, S_3, \dots, S_m) = -\sum_{i=1}^{m} \frac{S_i}{S} log_2 \frac{S_i}{S} \quad (1)$$

ตัวอย่างในการโยนหัวโยนก้อย บุญเสริม กิจศิริกุล [6] ชุดข้อมูล S มีค่าที่เป็นไปได้คือ (หัว, ก้อย) ให้ความน่าจะเป็นในการเกิดหัวแทนด้วย $P(\text{หัว})$ และความน่าจะเป็นในการเกิดก้อยแทนด้วย $P(\text{ก้อย})$ ดังนั้นค่าเอนโทรปีในการโยนหัวโยนก้อย คำนวณได้ดังสมการที่ (2)

$$I(\text{การโยนหัวก้อย}) = -P(\text{หัว}) log_2 P(\text{หัว}) - P(\text{ก้อย}) log_2 P(\text{ก้อย}) \quad (2)$$

เมื่อคำนวณตามสูตรนี้แล้วจะได้ดังภาพที่ 2-2 จะสังเกตเห็นว่าถ้าออกหัวหมดหรือออกก้อยหมด ค่าเอนโทรปีจะเป็น 0 และค่าเอนโทรปีสูงสุดเมื่อความน่าจะเป็นของการเกิดหัวเท่ากับความน่าจะเป็นของการเกิดก้อย จึงสรุปได้ว่าถ้าข้อมูลแตกต่างกันน้อยจะให้ค่าเอนโทรปีต่ำ แต่ถ้าข้อมูลมีความแตกต่างกันมากจะให้ค่าเอนโทรปีสูง



ภาพที่2-2 ค่าเอนโทรปีของการโยนหัวโยนก้อย

ค่าเอนโทรปีที่ใช้จำแนกข้อมูลโดยใช้คุณลักษณะ A คำนวณได้ดังสมการที่ (3)

$$E(A) = -\sum_{j=1}^{v} \frac{S_{1j}+\ldots+S_{mj}}{S} I(S_{1j}, \ldots, S_{mj}) \quad (3)$$

ค่าเกนมาตรฐาน (Gain) คำนวณได้จากค่าเอนโทรปีก่อนการแบ่งแยกลบด้วยค่าเอนโทรปีที่ได้หลังจากการแบ่งด้วยคุณลักษณะที่ถูกเลือกเขียนได้ดังสมการที่ (4)

$$Gain(A) = I(S_1, S_2, S_3, \ldots, S_m) - E(A) \quad (4)$$

จากตารางที่ 2-1 ตัวอย่างชุดข้อมูลเรียนรู้การตัดสินใจออกไปตีกอล์ฟ มีจำนวนข้อมูล 14 ระเบียน มีข้อมูล 2 กลุ่ม คือ ข้อมูลตัดสินใจออกไปเล่นกอล์ฟ 9 ระเบียน และข้อมูลตัดสินใจไม่ออกไปเล่นกอล์ฟ 5 ระเบียนค่าสารสนเทศที่ใช้ในการจำแนกประเภทข้อมูลโดยใช้สมการที่ (1) คำนวณได้ดังนี้

$I(S) = -(9/14) \times \log_2 (9/14) - (5/14) \times \log_2 (5/14)$

$= 0.940$ บิต

ในการตัดสินใจนั้นจะต้องใช้คุณลักษณะต่าง ๆ ประกอบการตัดสินใจ ถ้าแบ่งข้อมูลชุดนี้ด้วยคุณลักษณะ outlook ตามค่าที่เป็นไปได้ 3 ค่า คือ sunny, overcast และ rainy สามารถคำนวณค่าเอนโทรปีตามสมการที่ (3) ได้ดังนี้

$E(\text{outlook}) = (5/14)\times(-(2/5)\times\log_2(2/5) - (3/5)\times\log_2(3/5))$

$+ (4/14)\times(-(4/4)\times\log_2(4/4) - (0/4)\times\log_2(0/4))$

$+ (5/14)\times(-(3/5)\times\log_2(3/5) - (2/5)\times\log_2(2/5))$

$= 0.693$ บิต

ดังนั้นสามารถคำนวณค่าเกนมาตรฐานของคุณลักษณะ outlook ได้ดังนี้

$\text{Gain(outlook)} = I(S) - E(\text{outlook})$

$= 0.940 - 0.693$

$= 0.247$ บิต

และหาค่าเกนมาตรฐานของคุณลักษณะที่เหลือ เพราะคุณลักษณะที่เหลือสามารถถูกเลือกมาสร้างเป็นโหนดในการแบ่งแยกได้เช่นกัน หาค่าเกนมาตรฐานคุณลักษณะ temperature, humidity และ windy ได้ดังนี้

Gain (temperature) = I (S) – E (temperature)

= 0.940 – 0.911

= 0.029 บิต

Gain (humidity) = I (S) – E (humidity)

= 0.940 – 0.788

= 0.152 บิต

Gain (windy) = I (S) – E (windy)

= 0.940 – 0.892

= 0.048 บิต

คุณลักษณะ outlook มีค่าเกนมาตรฐานสูงสุด คือ 0.247 ดังนั้นจึงเลือกคุณลักษณะ outlook เป็นโหนดรากของต้นไม้ตัดสินใจและสร้างโหนดในระดับที่สองต่อไป โดยพิจารณาค่าเกนมาตรฐานของคุณลักษณะที่เหลือ

2.1.3. **อัลกอริทึม** C4.5

เนื่องจากค่าเกนมาตรฐานในอัลกอริทึม ID3 มีความโน้มเอียงในกรณีที่แต่ละคุณลักษณะมีค่าที่เป็นไปได้จำนวนมาก ๆ เช่น มีคุณลักษณะหนึ่งมีค่าไม่ซ้ำกันเลย ซึ่งจะส่งผลให้ได้ค่าเอนโทรปีเท่ากับ 0 เนื่องจาก $\log_2 1 = 0$ ทำให้ค่าเกนมาตรฐานที่ได้มีค่าสูงที่สุด จึงส่งผลให้การทำนายผิดพลาดด้วย ดังนั้นจึงต้องมีการปรับค่าเกนมาตรฐานใหม่ Quinlan [3] เสนออัลกอริทึม C4.5 ซึ่งเป็นอัลกอริทึมที่ใช้กันอย่างแพร่หลาย โดยนำอัลกอริทึม ID3 มาพัฒนาต่อ โดยได้เพิ่มการใช้ค่าสารสนเทศการแบ่งแยก (split information) ในการตัดสินใจเลือกคุณลักษณะที่จะนำมาใช้เป็นโหนด กำหนดให้ T แทนชุดข้อมลเรียนรู้ แบ่งข้อมูลเป็น v ชุด โดยใช้คุณลักษณะ A ได้ชุดข้อมูลย่อยในแต่ละกิ่งเป็น $\{t_1, t_2, \ldots t_v\}$ สามารถคำนวณค่าสารสนเทศการแบ่งแยกได้ดังสมการที่ (5)

$$SplitInformation(A) = -\sum_{i=1}^{v} \frac{|t_i|}{|T|} log_2 \left(\frac{|t_i|}{|T|} \right) \quad (5)$$

เนื่องจากคุณลักษณะมีค่าที่เป็นไปได้จำนวนมาก ๆ จะให้ค่าเอนโทรปีสูงค่า $|t_i|$ จะมีค่าสูงสุดเมื่อ $|t_i|$ ทุกกิ่งมีค่าเป็น 1 และจะลดลงเมื่อ $|t_i|$ เพิ่มขึ้น จึงนำค่านี้ไปหารค่าเกนมาตรฐานเพื่อแก้ไขความลำเอียงโดยการทำให้ค่าเกนมาตรฐานคุณลักษณะที่มีค่าเป็นไปได้จำนวนมากถูกปรับลดลง สามารถคำนวณค่ามาตรฐานอัตราส่วนเกน (Gain Ratio) ได้ดังสมการที่ (6)

$$Gain\ Ratio(A) = \frac{Gain(A)}{SplitInformation(A)} \quad (6)$$

จากตารางที่ 2-1 ตัวอย่างชุดข้อมูลเรียนรู้การตัดสินใจออกไปตีกอล์ฟ ค่าสารสนเทศการแบ่งแยกคำนวณตามสมการที่ (5) และค่ามาตรฐานอัตราส่วนเกนของคุณลักษณะ outlook คำนวณได้จากสมการที่ (6) ได้ดังนี้

Split Information (outlook) = – (5/14) x $\log_2$ (5/14) – (4/14)x $\log_2$ (4/14)
– (5/14) xlog2 (5/14)
= 1.577 บิต

Gain ratio (outlook) = 0.247 / 1.577
= 0.156 บิต

คุณลักษณะที่เหลือสามารถถูกเลือกมาสร้างเป็นโหนดในการแบ่งแยกได้เช่นกัน ค่ามาตรฐานอัตราส่วนเกนคุณลักษณะ temperature, humidity และ windy ได้ดังนี้

Gain ratio (temperature) = 0.029 / 1.362
= 0.021 บิต

Gain ratio (humidity) = 0.152 / 1.000
= 0.152 บิต

Gain ratio (windy) = 0.048 / 0.985
= 0.049 บิต

เมื่อเปรียบเทียบทั้ง 4 คุณลักษณะ พบว่าคุณลักษณะ outlook มีค่ามาตรฐานอัตราส่วนเกนสูงสุด ดังนั้นจึงเลือกคุณลักษณะ outlook เป็นโหนดรากของต้นไม้ตัดสินใจและสร้างโหนดในระดับต่อไปโดยใช้คุณลักษณะที่เหลือ

2.2. งานวิจัยที่เกี่ยวข้อง

เริ่มมีนักวิจัยหลายคนสนใจการจำแนกข้อมูลไม่สมดุลมากขึ้นตั้งแต่ปี ค.ศ. 2000 ในงานประชุมวิชาการ AAAI มีสองประเด็นที่ได้รับความสนใจมาก ประเด็นแรก คือ จะประเมินประสิทธิภาพการเรียนรู้ของอัลกอริทึมในกรณีข้อมูลไม่สมดุลได้อย่างไร ซึ่งมีนักวิจัยได้เสนอเส้นโค้งอาร์โอซี (Receiver Operating Characteristic: Roc Curve) และเส้นโค้งค่าคอสต์ (Cost

Curve) ส่วนประเด็นที่สอง คือ ความสัมพันธ์ระหว่างคลาสไม่สมดุล และการเรียนรู้แบบไวต่อค่าคอสต์ (Cost-Sensitive Learning)

อีกประเด็นที่นักวิจัยสนใจคือ การแก้ปัญหานี้ด้วยเทคนิคการสุ่มซ้ำ เพื่อเปลี่ยนการกระจายของตัวอย่างในแต่ละกลุ่ม คือ การสุ่มให้มากขึ้น (Over-sampling) และ การสุ่มให้ลดลง (Under-sampling) และต้นไม้ตัดสินใจยังคงเป็นตัวจำแนกที่นิยมนำมาใช้หลังจากสุ่มซ้ำ

เมื่อมีหลายอัลกอริทึมเกิดขึ้น จึงมีนักวิจัยใช้หลาย ๆ อัลกอริทึมและเทคนิคต่างๆ รวมกัน (Ensemble method) เพื่อเรียนรู้หลาย ๆ อัลกอริทึมให้ได้หลายแบบจำลอง แล้วเอาแบบจำลองที่ได้มารวมกัน เพื่อโหวตหรือหาค่าเฉลี่ยให้เป็นคำตอบสุดท้าย หลายงานวิจัยมักจะรวมอัลกอริทึมพื้นฐานกับเทคนิคแบบต่าง ๆ เช่น วิธีการแบกกิ้ง (Bagging) และวิธีการบูสต์ (Boosting)

มีบางงานวิจัยที่ได้หาวิธีแบบใหม่ในการแยกข้อมูลเพื่อเรียนรู้ต้นไม้ตัดสินใจ มีทั้งใช้การปรับเปลี่ยนอัลกอริทึมพื้นฐานที่มีอยู่แล้วทางด้านการทำเหมืองข้อมูล และการใช้การวัดระยะห่างระหว่างตัวอย่างสองกลุ่ม

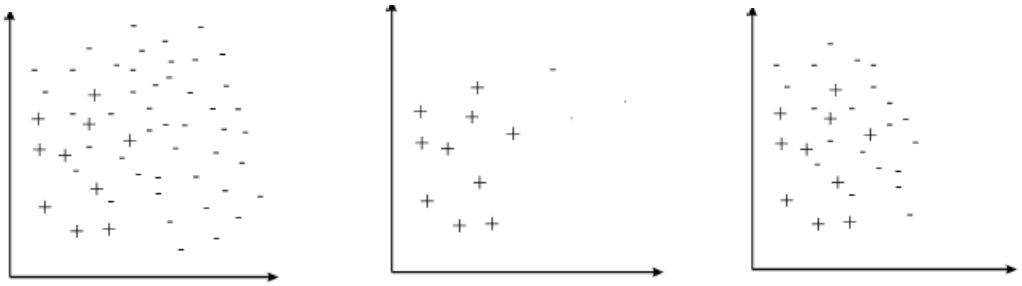
ต่อไปนี้จะยกตัวอย่างงานวิจัยโดยแบ่งออกเป็นประเภทต่าง ๆ ดังนี้

2.2.1. งานวิจัยที่ใช้เทคนิคการสุ่มในแต่ละกลุ่ม (re-sampling)

งานวิจัยประเภทนี้จะเปลี่ยนการกระจายตัวอย่างในแต่ละกลุ่มในชุดข้อมูลเรียนรู้ แบ่งเป็นสองวิธี คือ การเอาตัวอย่างออก (Under-sampling) เพื่อเอาตัวอย่างในกลุ่มที่มีมากออกไป และการเพิ่มตัวอย่าง (Over-sampling) เพื่อสร้างตัวอย่างเข้าไปเพิ่มในกลุ่มที่มีน้อย

ตัวอย่างงานวิจัยที่ใช้เทคนิคการเอาตัวอย่างออกและการเพิ่มตัวอย่างมีดังต่อไปนี้

Hart, P. E. [7] เสนอ กฎเพื่อนบ้านใกล้สุดแบบลดแน่น (Condensed Nearest Neighbor Rule: CNN) ในปี 1968 ใช้การหาสับเซตของตัวอย่าง เพื่อกำจัดสิ่งรบกวนในกลุ่มที่มีมากออก กล่าวคือ กำหนดให้ E คือ ข้อมูลเรียนรู้ทั้งหมด, ให้ E' คือ ตัวอย่างกลุ่มที่มีน้อยทั้งหมดรวมกับตัวอย่างกลุ่มที่มีมากหนึ่งตัวที่ได้จากการสุ่ม หลังจากนั้นทำการจำแนก E ด้วยกฎ 1-NN โดยใช้ข้อมูลใน E' และสร้างข้อมูลเรียนรู้ชุดใหม่ที่เอาตัวอย่างที่คาดว่าจำทำนายผิดออกไป แสดงการใช้กฎเพื่อนบ้านใกล้สุดแบบลดแน่นดังภาพที่ 2-3



ภาพที่ 2-3 กฎเพื่อนบ้านใกล้สุดแบบลดแน่น

Tomek, I. [8] เสนอรอยเชื่อมโทเม็ก (Tomek links) ในปี 1976 ซึ่งเป็นเทคนิคการหาขอบเขตอันตราย ที่เรียกว่า Tomek links เพื่อกำจัดสิ่งรบกวนในกลุ่มที่มีมากออก โดยรอยเชื่อมโทเม็กนั้นหาได้จาก กำหนดให้ E_i, E_j คือ กลุ่มที่แตกต่างกัน, $d(E_i,E_j)$ คือระยะห่างระหว่างสองกลุ่ม ถ้าไม่มีตัวอย่าง E_1 ที่ $d(E_i,E_1) < d(E_i,E_j)$, $d(E_j,E_1) < d(E_i,E_j)$ เรียกว่า คู่รอยเชื่อมโทเม็ก แสดงการใช้รอยเชื่อมโทเม็กดังภาพที่ 2-4

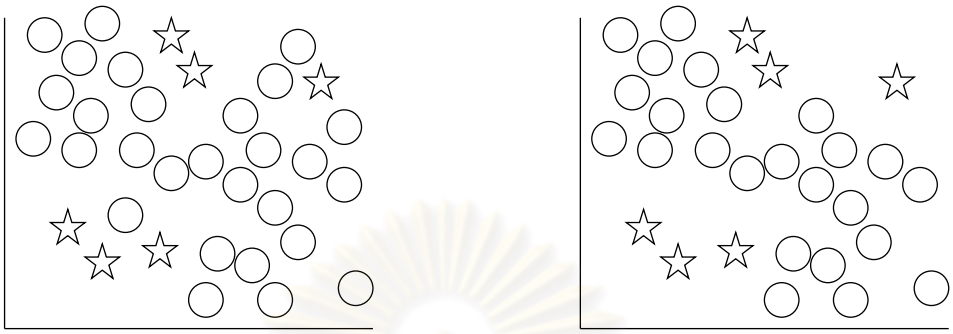
ภาพที่ 2-4 รอยเชื่อมโทเม็ก

Kubat, M. และ Matwin, S. [9] เสนอ วิธีการเลือกข้างเดียว (One-sided selection: OSS) ในปี 1997 ใช้ทั้งวิธีรอยเชื่อมโทเม็ก และเพื่อนบ้านใกล้สุดแบบลดแน่นโดยใช้วิธีรอยเชื่อมโทเม็กเอาตัวอย่างที่อันตรายออกไปก่อน ดังนั้นจะเหลือแต่ตัวอย่างที่ปลอดภัยแล้วจึงใช้กฎเพื่อนบ้านใกล้สุดแบบลดแน่นเอาตัวอย่างออกไปอีก

Batista, G.E.A.P.A., Prati, R.C. และ Monard, M.C. [10] เสนอ CNN+Tomek links ในปี 2004 ซึ่งจะคล้ายกับ OSS แต่วิธีนี้จะใช้กฎเพื่อนบ้านใกล้สุดแบบลดแน่นก่อนรอยเชื่อมโทเม็ก

Laurikkala, J. [11] เสนอ กฎการทำความสะอาดบริเวณใกล้เคียง (Neighborhood Cleaning Rule: NCL) ในปี 2001 เพื่อกำจัดสิ่งรบกวนให้ได้มากขึ้น วิธีนี้จะเอาตัวอย่างออกมากกว่าวิธีที่กล่าวมาก่อนหน้านี้ กำหนดถ้าให้ E_i เป็นตัวอย่างอยู่ในกลุ่มที่มีมาก และตัวอย่างที่อยู่ใกล้ E_i สามตัว (3-NN) เป็นตัวอย่างในกลุ่มที่มีน้อย ดังนั้นให้เอา E_i ออก หรือถ้า E_i

เป็นตัวอย่างในกลุ่มที่มีน้อยและตัวอย่างที่อยู่ใกล้ E_i สามตัว (3-NN) เป็นตัวอย่างในกลุ่มที่มีมาก ดังนั้นให้เอาตัวอย่างที่อยู่ใกล้สามตัวออก แสดงการใช้กฎการทำความสะอาดบริเวณใกล้เคียงดังภาพที่ 2-5



ก่อน หลัง

ภาพที่ 2-5 ก่อนและหลังการใช้กฎการทำความสะอาดบริเวณใกล้เคียง

ส่วนเทคนิคการเพิ่มข้อมูลนั้น Chawla, N. V., Bowyer, K. W. L., Hall, O., และ Kegelmeyer, W. P. [12] เสนอ SMOTE ในปี 2002 ใช้เทคนิคหาเพื่อนบ้านใกล้สุด k ตัว เพื่อสร้างตัวอย่างเสมือนจริงเพิ่มเข้าไปในกลุ่มที่มีจำนวนตัวอย่างน้อยอาจจะ over-sampling 100%, 200%, 300% หรือ 400% ของข้อมูลต้นฉบับ

ขั้นตอนการสร้างตัวเสมือนจริงมีดังนี้

1) กำหนดให้ $\forall S_{min} \in S$

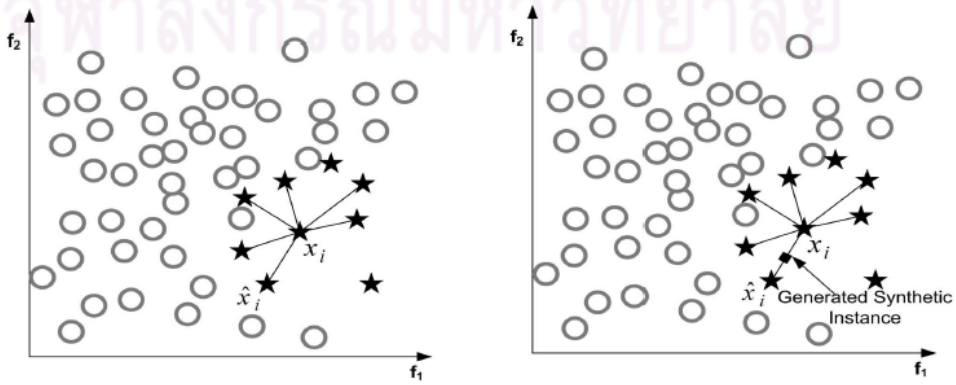
2) สุ่ม x_i โดยที่ $\forall \mathrm{x}_i \in S_{min}$

3) หาเพื่อนบ้านใกล้สุด k ตัว$(\hat{x})$, $\forall \check{x} \in S_{min}$

4) สุ่ม δ โดยที่ $\delta \in [0,1]$

5) สร้างตัวอย่างเสมือนจริงได้ดังสมการที่ (7)

$$X_{\mathrm{new}} = \mathrm{x}_i + (\check{x} - \mathrm{x}_i) \times \delta \tag{7}$$



a) b)

ภาพที่ 2-6 การใช้ SMOTE

จากภาพที่ 2-6 รูปวงกลมแทนกลุ่มที่มีมาก และรูปดาวแทนกลุ่มที่มีน้อย a) สุ่มตัวอย่างในกลุ่มที่มีน้อย (x_i) หนึ่งตัว และหาตัวอย่างในกลุ่มที่มีน้อยที่ใกล้สุด $(\hat{x})$ 6 ตัว b) หาตัวเสมือนจริงเพิ่มเข้าไปในข้อมูลเรียนรู้

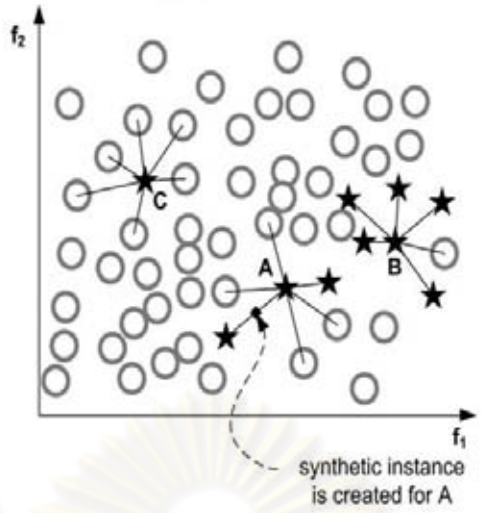
Han, H., Wang W., และ Mao, B. [13] เสนอ การหาขอบเขต SMOTE (Borderline-SMOTE) ในปี 2005 เนื่องวิธี SMOTE อาจจะสร้างตัวอย่างเพิ่มไปในเขตที่ไม่ควรสร้าง เพราะอาจทำให้การจำแนกผิดพลาดได้จึงต้องมีการกำหนดขอบเขต 3 ขอบเขต คือ เขตสิ่งรบกวน เขตอันตราย เขตปลอดภัย เพื่อเป็นเงื่อนไขในการสร้างตัวเสมือนจริง

ขั้นตอนการสร้างตัวเสมือนจริงมีดังนี้

1) สำหรับทุก ๆ ตัวอย่างในกลุ่มที่มีน้อย $P_i\{i = 1,2, \dots p_{num}\}$ หา m เพื่อนบ้านใกล้สุดจากข้อมูลเรียนรู้ทั้งหมด จำนวนตัวอย่างในกลุ่มที่มีมากท่ามกลาง mเพื่อนบ้านใกล้สุด ให้แทนด้วย $\acute{m}$ ซึ่ง $0 \le \acute{m} \le m$
2) ถ้า $m = \acute{m}$ นั่นคือ m เพื่อนบ้านใกล้สุดทั้งหมดเป็นตัวอย่างในกลุ่มที่มีมากกำหนดให้เป็นเขตสิ่งรบกวนและไม่เลือกมาใช้
3) ถ้า $m/2 \le \acute{m} < m$ ให้เป็นเขตอันตราย หรือ $DANGER$ เนื่องจากมีเพื่อนบ้านใกล้สุดที่อยู่ตัวอย่างที่อยู่ในกลุ่มที่มีมากเท่ากับหรือเกินกว่าครึ่งหนึ่งของตัวอย่างในกลุ่มที่มีน้อย ถ้าสร้างตัวอย่างเพิ่มไปในบริเวณนี้ จะเกิดโอกาสจำแนกผิดได้มาก
4) ถ้า $0 \le \acute{m} < m/2$ ให้เป็นเขตปลอดภัยเนื่องจากมีเพื่อนบ้านใกล้สุดที่อยู่ตัวอย่างที่อยู่ในกลุ่มที่มีมากน้อยกว่าครึ่งหนึ่งของตัวอย่างในกลุ่มที่มีน้อย
5) ตัวอย่างที่อยู่ในเขตอันตราย ต้องมีการกำหนดขอบเขตข้อมูลใหม่สามารถเขียนได้ดังนี้ $DANGER = \{\acute{P}_1, \acute{P}_2, \acute{P}_3, \dots \acute{P}_{dnum}\},\ 0 \le d_{num} \le p_{num}$
6) สร้างตัวอย่างเสมือนในเขตอันตราย สุ่มเลือก P_i มาและหาเพื่อนบ้านใกล้สุด m ตัว ใน P หาระยะห่าง $dif_j(j = 1,2, \dots, m)$ ระหว่าง $\acute{P}_i$ และ mเพื่อนบ้านใกล้สุด หลังจากนั้นสุ่ม r_j ระหว่าง 0-1 สร้างตัวอย่างเสมือนจริงแบบใหม่ตามสมการที่ (8)

$$Synthetic_j = \acute{P}_i + r_j \times dif_j, j = 1,2, \dots, s \qquad (8)$$

7) ทำการสร้างตัวเสมือนจริงซ้ำไปเรื่อย ๆ จนครบ m ตัว



ภาพที่ 2-7 Borderline-SMOTE1

จากภาพที่ 2-7 แสดง Borderline-SMOTE1 A คือ ตำแหน่งอันตราย B คือ ตำแหน่งปลอดภัย C คือตำแหน่งสิ่งรบกวน และจุดสี่เหลี่ยม คือ ตำแหน่งที่สร้างตัวเสมือนจริงเข้าไปเพิ่ม ส่วน Borderline-SMOTE 2 นั้นทำเช่นเดียวกับ Borderline-SMOTE1 แต่จะสุ่ม r_j ระหว่าง 0-0.5 ดังนั้นตัวเสมือนจริงที่สร้างขึ้นมาใหม่นั้นจะเข้าใกล้ตัวอย่างอื่น ๆ มากขึ้นกว่าวิธี Borderline-SMOTE1

Batista, G.E.A.P.A. Prati, R.C., และ Monard, M.C. [10] เสนอ SMOTE รวมกับรอยเชื่อมโทเม็กและ SMOTE รวมกับวิธีแก้ไขเพื่อนบ้านใกล้เคียงของวิลสัน (Wilson's Edited Nearest Neighbor Rule) [14] ซึ่งทั้งสองวิธีใช้ทั้งเทคนิคการเพิ่มตัวอย่างและการเลือกตัวอย่างออก แต่วิธีแก้ไขเพื่อนบ้านใกล้เคียงของวิลสันเอาตัวอย่างในกลุ่มที่มีมากออกมากกว่าวิธีรอยเชื่อมโทเม็ก ผลที่ได้คือ ทั้งสองวิธีนี้ให้ผลดีเมื่อจำแนกข้อมูลที่ไม่สมดุลมาก ๆ และได้เปรียบเทียบการใช้เทคนิคการเอาตัวอย่างออกและเพิ่มตัวอย่างกับข้อมูลไม่สมดุล ผลปรากฏว่าเทคนิคการเพิ่มตัวอย่างมีประสิทธิภาพที่ดีกว่าการเอาตัวอย่างออก

2.2.2. **งานวิจัยที่ใช้อัลกอริทึมหลายตัวรวมกัน** (ensemble method)

งานวิจัยประเภทนี้จะใช้อัลกอริทึมหลาย ๆ ตัวรวมกันหรือนำแต่ละวิธีมาผสานกัน ส่วนมากในการแก้ปัญหาข้อมูลไม่สมดุลนั้นผู้วิจัยมักจะใช้อัลกอริทึมพื้นฐานร่วมกับเทคนิคการสุ่มแบบต่าง ๆ ซึ่งมักจะให้ผลลัพธ์ที่ดีกว่าการใช้เพียงอัลกอริทึมเดียวแต่เนื่องจากต้องสร้างหลายแบบจำลอง ทำให้เสียเวลาในการจำแนกนานกว่าการใช้อัลกอริทึมเดียว

ตัวอย่างงานวิจัยที่ใช้การผสานหลายอัลกอริทึม มีดังต่อไปนี้

Chawla, N. V., Lazarevic, A., Hall, L. O. และ Bowyer, K. W. [15] ได้เสนออัลกอริทึม SMOTEBoost ในปี 2003 ซึ่งได้พัฒนาจาก SMOTE โดยการใช้กระบวนการ boosting ร่วมด้วย โดยจะสร้างตัวจำแนกขึ้นมาจะพบว่าตัวจำแนกนี้จำแนกข้อมูลใดผิดบ้าง และจะสร้างตัวจำแนกใหม่โดยเพิ่มน้ำหนักกับข้อมูลที่ตัวจำแนกผิด และสร้างตัวจำแนกใหม่พร้อมกับปรับน้ำหนักใหม่ไปเรื่อย ๆ เพื่อให้การทำนายคำตอบดีขึ้น

Chen C., Liaw, A., และ Breiman, L. [16] เสนอป่าแบบสุ่มแบบสมดุล (Balanced Random Forest) และถ่วงน้ำหนักป่าแบบสุ่ม (Weighted Random Forest) ในปี 2004 ทั้งสองวิธีนี้พัฒนามาจากป่าแบบสุ่ม (Random Forest) [17] ซึ่งเป็นอัลกอริทึมที่สร้างต้นไม้หลาย ๆ ต้น และโหวตเลือกต้นที่ดีที่สุด ป่าแบบสุ่มแบบสมดุลนั้นจะสุ่มลดตัวอย่างในกลุ่มที่มีจำนวนตัวอย่างมากให้เท่ากับกลุ่มที่มีจำนวนตัวอย่างน้อย ส่วนการถ่วงน้ำหนักป่าแบบสุ่มจะถ่วงน้ำหนักในสัมประสิทธิ์จีนี (Gini) ของอัลกอริทึมการจำแนกและการถดถอย (CART) [18] โดยให้น้ำหนักมากกับกลุ่มที่มีจำนวนตัวอย่างน้อย หรือให้น้ำหนักน้อยกับกลุ่มที่มีจำนวนตัวอย่างมากแล้วจึงสร้างต้นไม้หลาย ๆ ต้น ผลที่ได้คือ ป่าแบบสุ่มแบบสมดุลและการถ่วงน้ำหนักป่าแบบสุ่มให้ประสิทธิภาพที่ดีกว่าการใช้ต้นไม้ต้นเดียวและทั้งสองวิธีนี้ให้ประสิทธิภาพใกล้เคียงกัน

2.2.3. งานวิจัยที่ใช้เอนโทรปีแบบใหม่

เนื่องจากเอนโทรปีที่ใช้กันอยู่ในปัจจุบันไบแอสไปทางกลุ่มที่มีมาก ดังนั้นจึงมีหลายงานวิจัยสนใจเปลี่ยนการหาค่าเอนโทรปีแบบใหม่ เพื่อให้เหมาะสมกับข้อมูลไม่สมดุล

ตัวอย่างงานวิจัยที่ใช้การเปลี่ยนเอนโทรปี มีดังต่อไปนี้

Marcellin, S., Zighed, D. A., และ Ritschard, G. [19] ได้เสนอ เอนโทรปีแบบอสมมาตร (Asymmetric Entropy: AE) ในปี 2006 ซึ่งเป็นการวัดค่าเอนโทรปีแบบไม่สมมาตร ค่าเอนโทรปีไม่จำเป็นที่จะมีค่าสูงสุดที่ค่าความน่าจะเป็นเท่ากับ 0.5 ตามวิธี C4.5 แต่จะมีค่าสูงสุด เมื่อ p=w ซึ่ง w กำหนดโดยผู้ใช้งาน

แสดงคุณสมบัติของเอนโทรปีแบบอสมมาตรได้ดังนี้

1) $\frac{^2h}{p^2}$ 0 ตามกฎความเว้า เมื่ออนุพันธ์อันดับที่สองน้อยกว่าศูนย์ จึงได้เส้นโค้งคว่ำ
2) เอนโทรปีมีค่าต่ำสุด เมื่อ $h(p=1)=0$ และ $h(p=0)=0$
3) เอนโทรปีมีค่าสูงสุด $\frac{h}{p}=0;$ เมื่ออนุพันธ์อันดับหนึ่งเท่ากับศูนย์จะได้จุดสูงสุดของเส้นโค้ง

4) จากกฎข้อที่ 1, 2 และ 3 จึงได้ฟังก์ชันความสัมพันธ์ดังสมการที่ (9)

$$h(p) = \frac{ap^2 + bp + c}{dp + e} \tag{9}$$

ซึ่งกำหนดให้ $a = -1,\ b = 1, c = 0, d = 0, e = -2w + 1$และ $f = w^2$

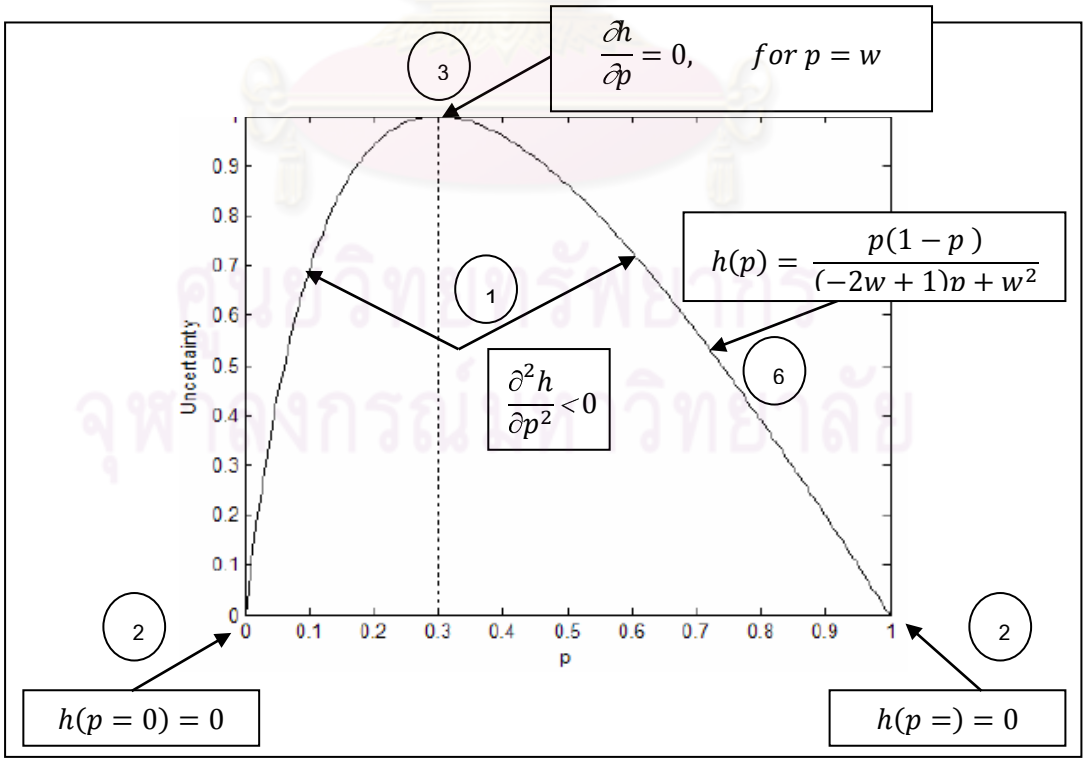
5) $h\ (p = w)\ =\ 1$

6) จากกฎข้อ 2, 5 และฟังก์ชันในข้อ 4 จึงได้เอนโทรปีดังสมการที่ (10)

$$h(\mathrm{p}_1, \mathrm{p}_2, \mathrm{p}_3 \ldots, \mathrm{p}_\mathrm{k}) = -\sum_{i=1}^{k} h(\mathrm{p}_\mathrm{i}) \tag{10}$$

$$h(p) = \frac{-p^2 + p}{(-2w + 1)p + w^2} = \frac{p(1 - p\ )}{(-2w + 1)p + w^2} \tag{11}$$

สำหรับปัญหาสองกลุ่ม ให้ $p_2 = 1 - p_1$ และ $w_2 = 1 - w_1$ ซึ่ง w_1 กำหนดโดยผู้ใช้งาน การเลือกคุณลักษณะนั้นจะเลือกจากคุณลักษณะที่มีค่าเอนโทรปีน้อยที่สุดเป็นโหนดในการสร้างต้นไม้เส้นโค้งแสดงคุณสมบัติต่าง ๆ ของเอนโทรปีแบบอสมมาตร สำหรับ $w_1 = 0.3$ ได้ดังภาพที่ 2-8



ภาพที่ 2-8 คุณสมบัติของเอนโทรปีแบบอสมมาตร

Lallich, S., Lenca, P., และ Vaillant, B. [20] เสนอ เอนโทรปีแบบออกจากศูนย์กลาง (Off-Centered Entropy: OCE) ในปี 2007 ซึ่งพัฒนามาจากเอนโทรปีแบบอสมมาตร กำหนดให้มีการหาค่าเอนโทรปีได้ 2 ฟังก์ชัน 2 กรณี โดยใช้แชนนอนเอนโทรปี [2] เป็นพื้นฐาน ดังนั้นจากเอนโทรปีของแชนนอน $h(p) = -plog_2 p - (1-p)log_2(1-p)$จะถูกแปลงเป็นค่าเอนโทรปีแบบใหม่ดังสมการที่ (12)

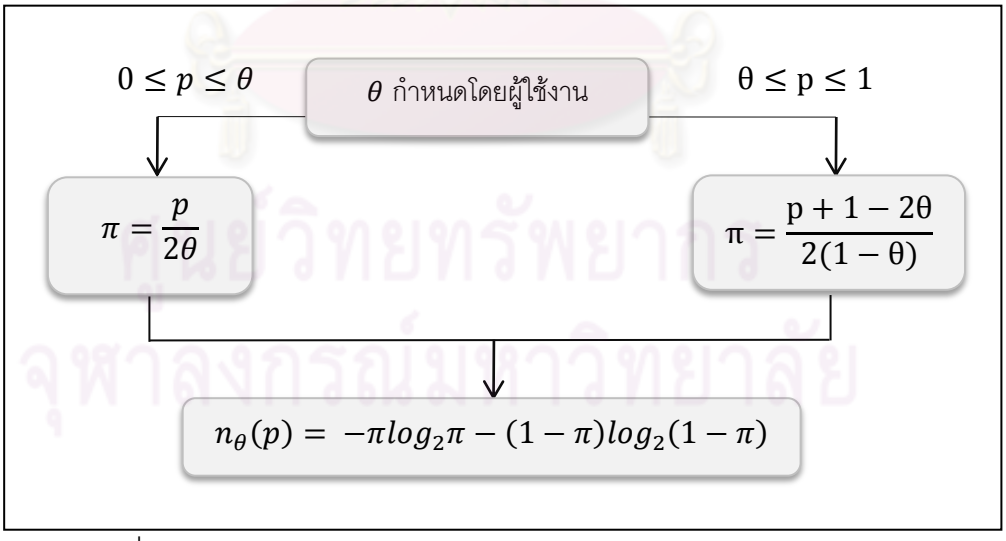
$$n_\theta(p) = -\pi log_2 \pi - (1-\pi)log_2(1-\pi) \tag{12}$$

เอนโทรปีแบบออกจากศูนย์กลางนี้มีค่าสูงสุดเมื่อ โดยที่กำหนดโดยผู้ใช้งาน กรณีปัญหาสองกลุ่มที่ $\mathrm{k} = 2$ แบ่งเป็น 2 กรณี ดังสมการที่ (13) และ (14)

$$\text{ถ้า } 0 \le p \le \theta \text{ แล้ว } \pi = \frac{p}{2\theta} \tag{13}$$

$$\text{ถ้า } \theta \le \mathrm{p} \le 1 \text{แล้ว} \pi = \frac{\mathrm{p}+1-2\theta}{2(1-\theta)} \tag{14}$$

ถ้า p เพิ่มขึ้นจาก 0 ถึง θ ดังนั้น π จะเพิ่มขึ้นจาก 0 ถึง 1/2
ถ้า p เพิ่มขึ้นจาก θ ถึง 1 ดังนั้น π จะเพิ่มขึ้นจาก 1/2 ถึง 1

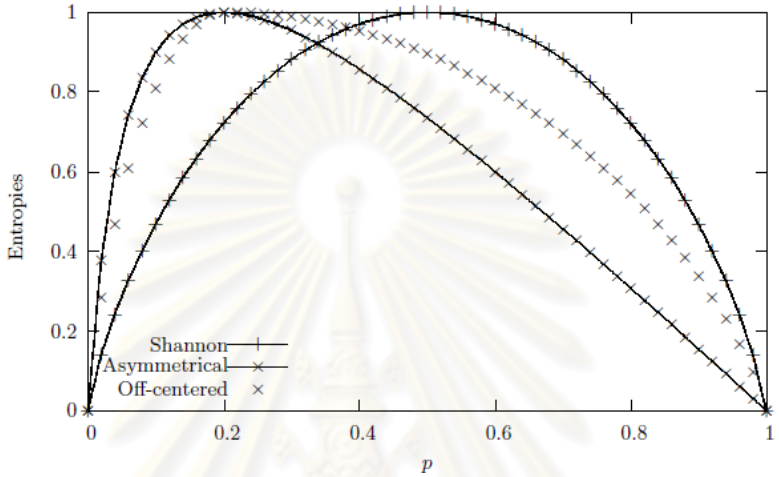


ภาพที่ 2-9 การหาค่าเอนโทรปีแบบออกจากศูนย์กลางแบบปัญหาสองกลุ่ม

ส่วนในกรณี $k > 2$ แบ่งเป็น 2 กรณี ดังสมการที่ (15) และ (16)

$$\text{ถ้า } 0 \leq p \leq \theta \text{ แล้ว } \pi = \frac{p}{k\theta} \tag{15}$$

$$\text{ถ้า } \theta \leq p \leq 1 \text{ แล้ว} \pi = \frac{k(p-\theta)+1-\theta}{k(1-\theta)} \tag{16}$$



ภาพที่ 2-10 เอนโทรปีแบบออกจากศูนย์กลาง,เอนโทรปีแบบอสมมาตร และแชนนอนเอนโทรปี

Cielak, D. A. และ Chawla, N. V. [21] ได้นำทฤษฎีระยะฮิลลิงเกอร์ (Hellinger Distance) ในปี 2008 ซึ่งเป็นทฤษฎีในการหาความแตกต่างของการกระจายทั้งสองกลุ่มมาประยุกต์ใช้กับต้นไม้ตัดสินใจ โดยใช้ระยะฮิลลิงเกอร์เป็นเกณฑ์แบ่งแยกต้นไม้ตัดสินใจระหว่างสองกลุ่ม เรียกว่า ต้นไม้ตัดสินใจระยะฮิลลิงเกอร์ (Hellinger Distance Decision Tree: HDDT)

กำหนดให้

X_+ แทนจำนวนของตัวอย่างในกลุ่มบวก

X_- แทนจำนวนของตัวอย่างในกลุ่มลบ

j แทนคุณสมบัติของคุณลักษณะ A ซึ่ง $1 \leq j \leq k$

แสดงคุณสมบัติของระยะฮิลลิงเกอร์ได้ดังนี้

1) $d_H(X_+, X_-)$ คือ ระยะห่างระหว่าง X_+ และ X_- อยู่ในช่วง $[0,\sqrt{2}]$
2) ถ้า $X_+ = X_-$ ดังนั้น $d_H(X_+, X_-) = 0$
3) ถ้า $X_+ \neq X_-$ ดังนั้น $d_H(X_+, X_-) = \sqrt{2}$
4) d_H มีคุณสมบัติสมมาตร ดังนั้น $d_H(X_+, X_-) = d_H(X_-, +)$
5) ระยะฮิลลิงเกอร์ระหว่าง X_+ และ X_- คำนวณได้ดังสมการที่ (17)

$$d_H(X_+, X_-) = \sqrt{\sum_{j=1}^{p}\left(\sqrt{\frac{|X_{+j}|}{|\mathrm{X}_+|}} - \sqrt{\frac{|X_{-j}|}{|\mathrm{X}_-|}}\right)^2} \quad (17)$$

ตารางที่ 2-2 ข้อมูลการตัดสินใจ

Att1	Att2	Att3	Class
Sunny	Yes	True	+
Rain	Yes	False	-
Sunny	no	True	+
Sunny	Yes	False	-
Rain	Yes	False	+

การสร้างต้นไม้ตัดสินใจโดยใช้ระยะฮิลลิงเกอร์เป็นเกนการแบ่งแยกทดสอบหาความแตกต่างทุก ๆ คุณลักษณะ คุณลักษณะใดมีค่าความแตกต่างมากที่สุด คุณลักษณะนั้นแบ่งแยกได้ดีที่สุด แต่ถ้าคุณลักษณะใดมีค่าความแตกต่างน้อยที่สุดแสดงว่าคุณลักษณะนั้นมีข้อมูลสองกลุ่มปะปนกันมากที่สุด ตัวอย่าง เช่น ข้อมูลมีสองกลุ่มคือ + และ – มี 3 คุณลักษณะ จากตารางที่ 2-2 สามารถคำนวณค่าความแตกต่างของแต่ละคุณลักษณะได้ดังนี้

$$d_H(X_+, X_-)(Att1) = \sqrt{\left(\sqrt{\frac{|X_{+sunny}|}{|\mathrm{X}_+|}} - \sqrt{\frac{|X_{-sunny}|}{|\mathrm{X}_-|}}\right)^2 + \left(\sqrt{\frac{|X_{+rain}|}{|\mathrm{X}_+|}} - \sqrt{\frac{|X_{-rain}|}{|\mathrm{X}_-|}}\right)^2}$$

$$d_H(X_+, X_-)(Att2) = \sqrt{\left(\sqrt{\frac{|X_{+yes}|}{|\mathrm{X}_+|}} - \sqrt{\frac{|X_{-yes}|}{|\mathrm{X}_-|}}\right)^2 + \left(\sqrt{\frac{|X_{+no}|}{|\mathrm{X}_+|}} - \sqrt{\frac{|X_{-no}|}{|\mathrm{X}_-|}}\right)^2}$$

$$d_H(X_+, X_-)(Att3) = \sqrt{\left(\sqrt{\frac{|X_{+True}|}{|\mathrm{X}_+|}} - \sqrt{\frac{|X_{-\mathrm{True}}|}{|\mathrm{X}_-|}}\right)^2 + \left(\sqrt{\frac{|X_{+\mathrm{False}}|}{|\mathrm{X}_+|}} - \sqrt{\frac{|X_{-\mathrm{False}}|}{|\mathrm{X}_-|}}\right)^2}$$

จากนั้นทำการเปรียบเทียบทั้ง 3 คุณลักษณะคุณลักษณะไหนมีค่าความแตกต่างมากที่สุดเลือกเป็นโหนดราก และวนทำซ้ำไปจนกว่าข้อมูลทุกตัวในโหนดนั้นอยู่กลุ่มเดียวกัน หรือใช้คุณลักษณะในการเรียนรู้ครบทุกตัวแล้ว

บทที่ 3
การออกแบบการสร้างต้นไม้ตัดสินใจและเอนโทรปี

วิธี C4.5 เป็นอัลกอริทึมพื้นฐานในการสร้างต้นไม้ตัดสินใจที่เป็นที่รู้จักกันอย่างแพร่หลายในการจำแนกข้อมูล ให้ผลความแม่นยำในการจำแนกสูง แต่เมื่อจำแนกข้อมูลไม่สมดุล ทำให้ผลการทำนายโอนเอียงไปในกลุ่มที่มีมาก และให้ผลการทำนายไม่ดีในกลุ่มที่มีน้อย เนื่องจากอัลกอริทึมนี้ถูกออกแบบมาสำหรับข้อมูลสมดุล

จากเหตุผลดังกล่าว จึงได้วิเคราะห์ปัญหาของวิธีการ C4.5 และออกแบบเอนโทรปีใหม่ เพื่อใช้เป็นตัววัดค่าคุณลักษณะ โดยได้เลือกใช้แชนนอนเอนโทรปีเป็นพื้นฐาน และนำมาปรับปรุงเพื่อให้ความสำคัญกับกลุ่มที่มีน้อย และทำให้ได้ผลความแม่นยำที่ดีขึ้นในกลุ่มที่มีน้อย เมื่อใช้กับข้อมูลไม่สมดุล

ในการออกแบบเอนโทรปีนั้น ได้เสนอตัวปรับที่สามารถปรับตัวมันเองให้สัมพันธ์กับความน่าจะเป็นทั้งสองกลุ่มในแต่ละระดับของการสร้างต้นไม้ตัดสินใจ มีข้อสันนิษฐานว่าเมื่อสร้างต้นไม้ตัดสินใจแล้วจะทำให้แยกตัวอย่างในกลุ่มที่มีน้อยให้เกาะติดกันเป็นกลุ่มก้อน

ข้อจำกัดอย่างหนึ่งของการใช้ต้นไม้ตัดสินใจ คือ ข้อมูลที่ใช้ในการทดสอบต้องไม่มีข้อมูลที่ไม่ทราบค่า และข้อมูลชนิดตัวเลข ถ้ามีต้องกำจัดข้อมูลไม่ทราบค่าออก ทั้งนี้ผู้วิจัยได้แทนด้วยค่ามัธยฐานสำหรับข้อมูลชนิดค่าโนมินัล (nominal) และแทนด้วยค่าเฉลี่ยสำหรับข้อมูลชนิดตัวเลข (numeric) ส่วนข้อมูลชนิดตัวเลขได้แปลงเป็นข้อมูลชนิดโนมินัล และข้อจำกัดอีกประการหนึ่งคือ เอนโทรปีที่ออกแบบใหม่นี้ไม่สามารถใช้กับข้อมูลแบบหลายกลุ่มได้

3.1. การสร้างต้นไม้ตัดสินใจ

3.1.1. โครงสร้างต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจประกอบด้วยแต่ละโหนด กิ่ง และใบ เริ่มด้วยการสร้างโหนดหนึ่งโหนด เพื่อเป็นรากของต้นไม้ โดยเลือกโหนดจากคุณลักษณะที่มีค่าอัตราส่วนตัวปรับแบบยึดเกาะ (AMIERatio) สูงสุด และสร้างกิ่งตามค่าของคุณลักษณะที่เลือกมา สร้างโหนดถัดไปเรื่อย ๆ โดยจะหยุด และเปลี่ยนปลายโหนดนี้เป็นใบเมื่อ

1) ตัวอย่างทั้งหมดในแต่ละกิ่งอยู่ในกลุ่มเดียวกัน
2) ไม่มีตัวอย่างใดตกในค่าตามกิ่งนั้น ๆ หลังจากการแบ่ง ให้เลือกค่าของกลุ่มที่มีความถี่สูงสุดของโหนดก่อนหน้าเป็นกลุ่มของใบนี้

3) กิ่งนั้นมีจำนวนของตัวอย่างกลุ่มใดกลุ่มหนึ่งตั้งแต่ 90% ของจำนวนตัวอย่างทั้งหมดในโหนดนั้น ๆ ให้เลือกค่าของกลุ่มที่มีความถี่สูงสุดของโหนดก่อนหน้าเป็นกลุ่มของใบนี้

4) ไม่เหลือคุณลักษณะที่จะมาสร้างเป็นโหนดแล้ว

3.2 **เอนโทรปีตัวปรับแบบยึดเกาะ** (Adhesively Modlfied Entropy : AMIE)

ผู้วิจัยเสนอ เอนโทรปีตัวปรับแบบยึดเกาะเป็นตัววัดค่าความสามารถในการแบ่งแยก โดยเลือกใช้แซนนอนเอนโทรปีเป็นพื้นฐานในสร้างต้นไม้ตัดสินใจ และใช้ตัวปรับที่สามารถปรับตัวเองให้สัมพันธ์กับความน่าจะเป็นของทั้งสองกลุ่มได้ในทุก ๆ ระดับ เพื่อหาค่าเอนโทรปีที่เหมาะสม ซึ่งจะพยายามเลือกคุณลักษณะที่ทำให้ตัวอย่างในกลุ่มที่มีน้อยอยู่เกาะกลุ่มกันให้ได้มาก

กำหนดค่าต่างๆในการคำนวณดังนี้

S_1 แทน จำนวนตัวอย่างในกลุ่มที่มีตัวอย่างมาก

S_2 แทนจำนวนตัวอย่างในกลุ่มที่มีตัวอย่างน้อย

C_i แทน กลุ่มของตัวอย่าง i กลุ่ม

S_{1j} แทนจำนวนตัวอย่างที่เป็นสมาชิกในกลุ่ม S_1 จากการแบ่งข้อมูลด้วยค่าที่เป็นไปได้ j ของคุณลักษณะ A

S_{2j} แทนจำนวนตัวอย่างที่เป็นสมาชิกในกลุ่ม S_2 จากการแบ่งข้อมูลด้วยค่าที่เป็นไปได้ j ของคุณลักษณะ A

สำหรับตัวปรับ $(\propto)$ นั้นสามารถแบ่งได้สองกรณี เพื่อให้เหมาะสมกับความน่าจะเป็นของทั้งสองกลุ่ม นั่นคือ กรณีที่ S_2 มากกว่า S_1 กำหนดให้ตัวปรับ $\propto = 1$ ส่วนกรณีที่ S_2 น้อยกว่า S_1 กำหนดให้ตัวปรับ $(\propto)$ เท่ากับสัดส่วนของจำนวนตัวอย่างในกลุ่มที่มีน้อยต่อจำนวนตัวอย่างในกลุ่มที่มีมาก นั่นคือ $\propto = S_2/S_1$

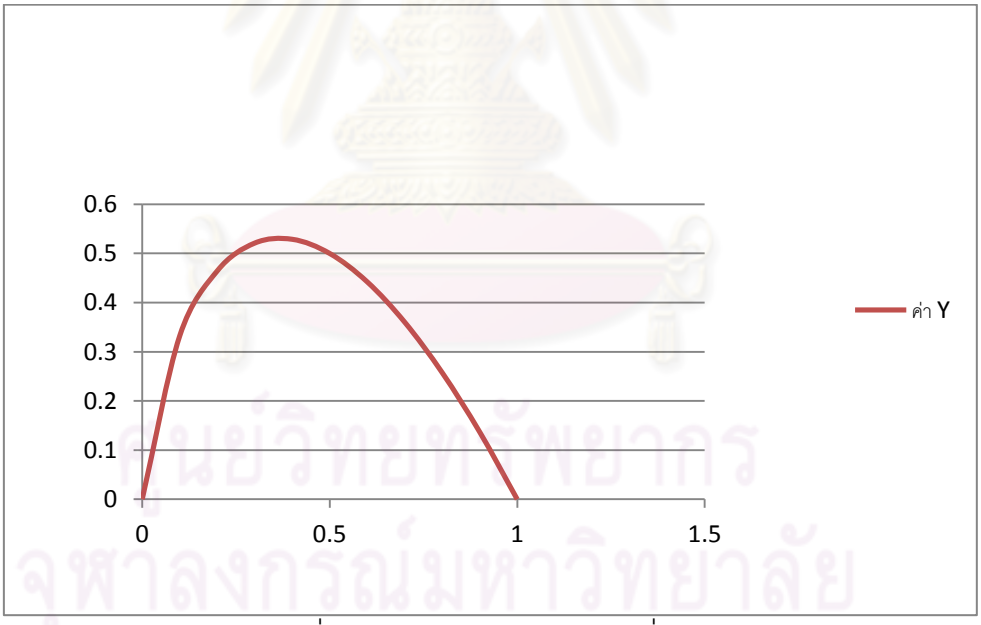
ค่าเอนโทรปีก่อนการแบ่งแยก $AMIE(C_i)$ และค่าเอนโทรปีหลังจากถูกแบ่งแยก $AMIE(C_i|A)$ แสดงได้ดังสมการที่ (1) และ (2) ได้ดังนี้

$$AMIE(C_i) = (\propto) \times \left[-\left(\frac{S_1}{S} log_2 \frac{S_1}{S}\right) - \left(\frac{S_2}{S} log_2 \frac{S_2}{S}\right) \right] \quad (1)$$

$$AMIE(C_i|A) = \left[(\propto) \times \left(-\sum_{j=1}^{V} P(a_j) \times P(C_i|a_j) log_2 P(C_i|a_j) \right) \right]$$

$$+ \left[(1-\propto) \times \left(-\sum_{j=1}^{V} \frac{S_{2j}}{S} \times \left(\frac{S_{2j}}{S_2} log_2 \frac{S_{2j}}{S_2} \right) \right) \right] \quad (2)$$

เมื่อพิจารณาจากสมการที่ (2) ในส่วนของ $-\frac{S_{2j}}{S_2} log_2 \frac{S_{2j}}{S_2}$ และภาพที่ 3-1 ถ้าความน่าจะเป็นในกลุ่มที่มีน้อยในคุณลักษณะนั้นตามค่าที่เป็นไปได้ j เมื่อพิจารณาจากจำนวนตัวอย่างในกลุ่มที่มีน้อย คือ $\frac{S_{2j}}{S_2}$ มีค่าเท่ากับ 0.9 ดังนั้นค่าของ $-\frac{S_{2j}}{S_2} log_2 \frac{S_{2j}}{S_2}$ จะมีค่าเท่ากับ -0.9log$_2$0.9 จะให้ค่าเอนโทรปี 0.1368 ซึ่งน้อยกว่า กรณีที่ $\frac{S_{2j}}{S_2}$ มีค่าเท่ากับ 0.1 ที่ค่าของ $-\frac{S_{2j}}{S_2} log_2 \frac{S_{2j}}{S_2}$ จะมีค่าเท่ากับ -0.1log$_2$0.1 จะให้ค่าเอนโทรปี 0.3322 หมายความว่ายิ่งมีตัวอย่างในกลุ่มที่มีน้อยเกาะกลุ่มกันมากยิ่งดีให้ค่าเอนโทรปีน้อย



ภาพที่ 3-1 ความน่าจะเป็นในกลุ่มที่มีน้อย

เนื่องจากคุณลักษณะอาจมีจำนวนค่าที่เป็นไปได้จำนวนมาก ๆ ซึ่งจะส่งผลให้เกิดความลำเอียง เพราะคุณลักษณะที่มีค่าเป็นไปได้จำนวนมาก จะให้ค่าเอนโทรปีสูง จึงทำให้ค่าเอนโทรปีถูกปรับลดลงด้วยค่าสารสนเทศการแบ่งแยก (Split Information) โดยคำนวณเช่นเดียวกับวิธี C4.5 ดังนั้นค่าอัตราส่วนตัวปรับแบบยึดเกาะสามารถเขียนได้ดังสมการที่ (3)

$$AMIERatio(A) = \frac{AMIE(C_i) - AMIE(C_i|A)}{SplitInformation(A)} \quad (3)$$

และเลือกคุณลักษณะที่มีค่าอัตราส่วนตัวปรับแบบยึดเกาะ (AMIERatio) มากที่สุดมาเป็นโหนดในการสร้างต้นไม้ตัดสินใจจากตารางที่ 2-1 ตัวอย่างชุดข้อมูลเรียนรู้การตัดสินใจออกไปตีกอล์ฟ มีจำนวนข้อมูล 14 ระเบียน มีข้อมูล 2 กลุ่ม คือ ข้อมูลตัดสินใจออกไปเล่นกอล์ฟ 9 ระเบียน และข้อมูลตัดสินใจไม่ออกไปเล่นกอล์ฟ 5 ระเบียนค่าสารสนเทศที่ใช้ในการจำแนกประเภทข้อมูลใช้สมการที่ (1)

AMIE(C) = (5/9)x(– (9/14)x$\log_2$ (9/14) – (5/14)x$\log_2$ (5/14))

= 0.5223 บิต

ในการตัดสินใจนั้น จะต้องใช้คุณลักษณะต่าง ๆ ประกอบการตัดสินใจถ้าแบ่งข้อมูลชุดนี้ด้วยคุณลักษณะ outlook ตามค่าที่เป็นไปได้ 3 ค่า คือ sunny, overcast, rainy สามารถคำนวณค่าเอนโทรปีตามสมการที่ (2) ได้ดังนี้

AMIE (outlook) = (5/9)x(5/14)x(– (2/5)x$\log_2$ (2/5) – (3/5)x$\log_2$ (3/5))

+(4/9)x(3/14)x(-(3/5)x$\log_2$ (3/5))

+ (5/9)x(4/14)x(– (4/4)x$\log_2$ (4/4) – (0/4)x$\log_2$ (0/4))

+(4/9)x(0/14)x(-(0/5)x$\log_2$ (0/5))

+ (5/14)x(– (3/5)x$\log_2$ (3/5) – (2/5)x$\log_2$ (2/5))

+(4/9)x(2/14)x(-(2/5)x$\log_2$ (2/5))

= 0.4609บิต

SplitInformation (outlook) = 1.5774

ดังนั้นสามารถคำนวณค่าอัตราส่วนตัวปรับแบบยึดเกาะของคุณลักษณะ outlook ได้ดังนี้

AMIERatio (outlook) = (0.5223 – 0.4609)/ 1.5774

= 0.0389 บิต

และหาค่าอัตราส่วนตัวปรับแบบยึดเกาะของคุณลักษณะที่เหลือ เพราะคุณลักษณะที่เหลือสามารถถูกเลือกมาสร้างเป็นโหนดในการแบ่งแยกได้เช่นกัน หาค่าอัตราส่วนตัวปรับแบบยึดเกาะของคุณลักษณะ temperature, humidity และ windy ได้ดังนี้

AMIERatio (temperature) = (0.5223 – 0.5880)/ 1.5566

= -0.0421 บิต

AMIERatio (humidity) = (0.5223 – 0.4854)/1

= 0.0369 บิต

AMIERatio(windy) = (0.5223 – 0.5713)/ 0.9852

= -0.0496 บิต

คุณลักษณะ outlook มีค่า AMIERatio สูงสุด คือ 0.0389 ดังนั้นจึงเลือกคุณลักษณะ outlook เป็นโหนดรากของต้นไม้ตัดสินใจและสร้างโหนดในระดับที่สองต่อไป โดยพิจารณาค่า AMIERatio คุณลักษณะที่เหลือ

บทที่ 4

ผลการทดลองและวิเคราะห์ผล

ในบทนี้จะแสดงขั้นตอนการทดลองการจำแนกข้อมูลเมื่อใช้วิธีการที่ได้นำเสนอ โดยเริ่มจาก 1. ชุดข้อมูลที่นำมาใช้ในการทดลอง และรายละเอียดของข้อมูล 2. การทำความสะอาดข้อมูล 3. การทดสอบแบบไขว้ข้าม 5 กลุ่ม 4. การกำหนดตัววัดผล 5. ผลการทดลองที่ได้จากการใช้วิธีที่ได้นำเสนอจำแนกข้อมูลไม่สมดุล โดยเปรียบเทียบผลการทดลองกับวิธีอื่นที่มีมาก่อนหน้านี้และวิเคราะห์ผลการทดลองด้วยค่าความแม่นยำทั้งหมด ค่าความแม่นยำในกลุ่มที่มีมาก ความแม่นยำในกลุ่มที่มีน้อย และเนื่องจากข้อมูลที่นำมาวิเคราะห์นี้เป็นข้อมูลไม่สมดุล จึงแสดงผลค่าความระลึก ค่าความเที่ยง และค่าเอฟในกลุ่มที่มีน้อย เพื่อแสดงให้เห็นประสิทธิภาพระหว่างวิธีที่นำเสนอกับวิธีอื่น ๆ แสดงขั้นตอนการทดลองได้ดังภาพที่ 4-1

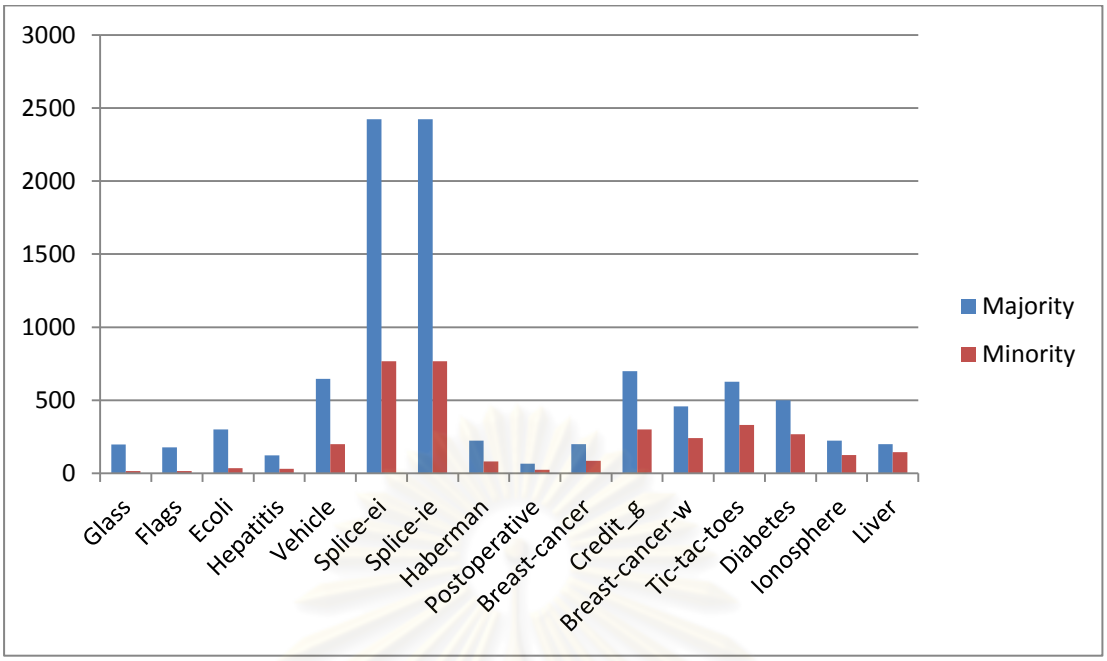
ภาพที่ 4-1 ภาพขั้นตอนการทดลอง

4.1. ข้อมูลที่ใช้ในการทดลอง

ข้อมูลที่ใช้ในการวิจัยได้คัดเลือกชุดข้อมูลที่ไม่สมดุลมาจาก UCI Machine Learning Repository [22] ทั้งหมด 16 ชุดข้อมูล สำหรับข้อมูลชุดที่ 1,2,3,5,6,7 และ 9 เป็นชุดข้อมูลมีมากกว่าสองกลุ่ม จึงได้เลือกกลุ่มที่มีน้อยเป็นกลุ่มหนึ่ง และให้กลุ่มที่เหลือรวมเป็นอีกกลุ่มหนึ่ง และข้อมูล Splice มีกลุ่มที่มีน้อยสองกลุ่มใกล้เคียงกัน จึงสร้างเป็นสองชุดข้อมูล มีรายละเอียดดังแสดงในตารางที่ 4-1 (คอลัมน์แสดงลำดับ ชื่อชุดข้อมูล จำนวนคุณลักษณะ ชื่อกลุ่มที่มีจำนวนตัวอย่างน้อย จำนวนตัวอย่างในกลุ่มที่มีมาก จำนวนตัวอย่างในกลุ่มที่มีน้อย จำนวนตัวอย่างทั้งหมด และเปอร์เซ็นของกลุ่มที่มีจำนวนตัวอย่างน้อย ตามลำดับ)

ตารางที่ 4-1 รายละเอียดข้อมูลไม่สมดุลที่ใช้ในการวิจัย

ลำดับ	ชื่อชุดข้อมูล	คุณลักษณะ	ชื่อกลุ่มน้อย	จำนวนกลุ่มมาก	จำนวนกลุ่มน้อย	จำนวนทั้งหมด	เปอร์เซ็นต์กลุ่มน้อย
1	Glasse	9	Vehic wind float	197	17	214	7.9439%
2	Flags	28	White	177	17	194	8.7629%
3	Ecoli	7	Imu	301	35	336	10.4167%
4	Hepatitis	19	Die	123	32	155	20.645%
5	Vehicle	18	Van	647	199	846	23.5225%
6	Splice-ei	60	EI	2423	767	3190	24.0439%
7	Splice-ie	60	IE	2422	768	3190	24.0752%
8	Haberman	3	2	225	81	306	24.470%
9	Post-operative	8	S	66	24	90	26.6667%
10	Breast-cancer	9	Recurrence-events	201	85	286	29.720%
11	Credit_g	20	Bad	700	300	1000	30%
12	Breast-cancer-w	9	Malignant	458	241	699	34.478%
13	Tic-tac-toes	9	Negative	626	332	958	34.655%
14	Diabetes	8	Tested-positive	500	268	768	34.895%
15	Ionosphere	34	B	225	126	351	35.8974%
16	Liver	6	1	200	145	345	42.0289%



ภาพที่ 4-2 แผนภูมิจำนวนตัวอย่างในกลุ่มที่มีมากและกลุ่มที่มีน้อยของ16 ชุดตัวอย่าง

ข้อมูลแต่ละชุดมีรายระเอียดการจำแนกกลุ่มของข้อมูลดังนี้

1) Glass เป็นข้อมูลจำแนกประเภทกระจก

มีจำนวนข้อมูล 214 ระเบียน จำแนกออกเป็น 2 กลุ่ม คือ

Vehicwind float (จำนวน 17 ระเบียน)

Remainder (จำนวน 197 ระเบียน)

2) Flags เป็นข้อมูลการจำแนกธง

มีจำนวนข้อมูล 194 ระเบียน จำแนกออกเป็น 2 กลุ่ม คือ

White (จำนวน 17 ระเบียน)

Remainder (จำนวน 177 ระเบียน)

3) Ecoli เป็นข้อมูลเกี่ยวกับโปรตีนที่อยู่ในร่างกาย

มีจำนวนข้อมูล 336 ระเบียนจำแนกออกเป็น 2 กลุ่ม คือ

Imu (จำนวน 35 ระเบียน)

Remainder (จำนวน 301 ระเบียน)

4) Hepatitis เป็นข้อมูลบันทึกว่าคนไข้โรคตับอักเสบยังมีชีวิตอยู่หรือไม่

มีจำนวนข้อมูล 155 ระเบียน จำแนกออกเป็น 2 กลุ่ม คือ

Die (จำนวน 32 ระเบียน)

Live (จำนวน 123 ระเบียน)

5) Vehicle เป็นข้อมูลจำแนกประเภทยานพาหนะ

มีจำนวนข้อมูล 846 ระเบียน จำแนกออกเป็น 2 กลุ่ม คือ

Van	(จำนวน 199 ระเบียน)
Remainder	(จำนวน 647 ระเบียน)

6) Splice-ei เป็นข้อมูลการจำแนกโครงสร้างทางพันธุกรรมของมนุษย์

มีจำนวนข้อมูล 3,190 ระเบียน จำแนกออกเป็น 2 กลุ่ม คือ

EI	(จำนวน 767 ระเบียน)
Remainder	(จำนวน 2423 ระเบียน)

7) Splice-ie เป็นข้อมูลการจำแนกโครงสร้างทางพันธุกรรมของมนุษย์

มีจำนวนข้อมูล 3,190 ระเบียน จำแนกออกเป็น 2 กลุ่ม คือ

IE	(จำนวน 768 ระเบียน)
Remainder	(จำนวน 2422 ระเบียน)

8) Haberman เป็นข้อมูลผู้ป่วยหลังการผ่าตัดมะเร็งเต้านมในโรงพยาบาลในชิคาโก

มีจำนวนข้อมูล 306 ระเบียน จำแนกออกเป็น 2 กลุ่ม คือ

2	(จำนวน 81 ระเบียน)
1	(จำนวน 255 ระเบียน)

9) Postoperative เป็นข้อมูลผู้ป่วยหลังการผ่าตัด

มีจำนวนข้อมูล 90 ระเบียน จำแนกออกเป็น 2 กลุ่ม คือ

S	(จำนวน 24 ระเบียน)
Remainder	(จำนวน 66 ระเบียน)

10) Breast-cancer เป็นข้อมูลการวินิจฉัยการเกิดมะเร็งเต้านมว่าจะเกิดขึ้นใหม่ได้หรือไม่

มีจำนวนข้อมูล 286 ระเบียน จำแนกออกเป็น 2 กลุ่ม คือ

no recurrence events	(จำนวน 201 ระเบียน)
recurrence events	(จำนวน 85 ระเบียน)

11) Credit_g เป็นข้อมูลการใช้บัตรเครดิต

มีจำนวนข้อมูล 1,000 ระเบียน จำแนกออกเป็น 2 กลุ่ม คือ

Bad	(จำนวน 300 ระเบียน)
Good	(จำนวน 700 ระเบียน)

12) Breast-cancer-w เป็นข้อมูลการวินิจฉัยมะเร็งเต้านมว่าเป็นชนิดร้ายแรงหรือไม่

มีจำนวนข้อมูล 699 ระเบียน จำแนกออกเป็น 2 กลุ่ม คือ

malignant	(จำนวน 241 ระเบียน)
benign	(จำนวน 458 ระเบียน)

13) Tic-tac-toes เป็นข้อมูลการเล่นเกมส์โอเอ๊กซ์

มีจำนวนข้อมูล 958 ระเบียน จำแนกออกเป็น 2 กลุ่ม คือ

Negative	(จำนวน 332 ระเบียน)
Positive	(จำนวน 626 ระเบียน)

14) Diabetes เป็นข้อมูลทดสอบว่าคนไข้มีอาการโรคเบาหวานหรือไม่ โดยใช้มาตรฐานองค์การอนามัยโลก ทดสอบกับคนไข้เพศหญิงที่เป็นชนเผ่าพื้นเมืองอินเดียแดง รัฐอริโซน่า

มีจำนวนข้อมูล 768 ระเบียน จำแนกออกเป็น 2 กลุ่ม คือ

Tested negative	(จำนวน 500 ระเบียน)
Tested positive	(จำนวน 268 ระเบียน)

15) Ionosphere เป็นข้อมูลการวัดสัญญาณ จำนวน 34 จุด เพื่อตรวจสอบว่าสัญญาณ good (ดี) หรือ bad (ไม่ดี)

มีจำนวนข้อมูล 351 ระเบียน จำแนกออกเป็น 2 กลุ่ม คือ

B	(จำนวน 126 ระเบียน)
Remainder	(จำนวน 225 ระเบียน)

16) Liver เป็นข้อมูลทดสอบความผิดปกติของตับของชายโสดที่นิยมดื่มแอลกอฮอร์

มีจำนวนข้อมูล 345 ระเบียน จำแนกออกเป็น 2 กลุ่ม คือ

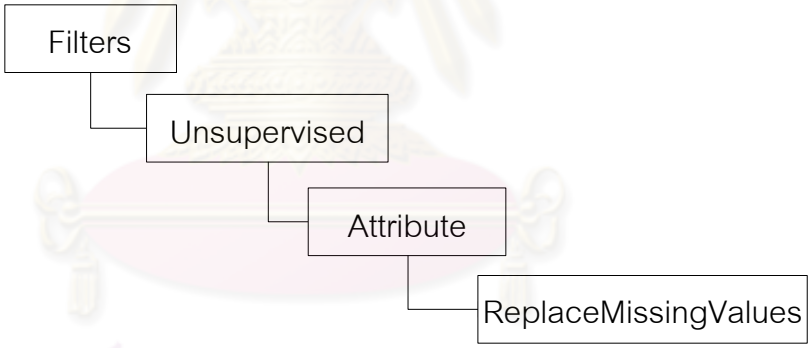
Class 1	(จำนวน 145 ระเบียน)
Class 2	(จำนวน 200 ระเบียน)

4.2. การทำความสะอาดข้อมูล

การทำเหมืองข้อมูลให้มีประสิทธิภาพ จะต้องเริ่มจากการเตรียมข้อมูล (Data Preprocessing) ให้มีคุณภาพ ซึ่งการเตรียมข้อมูลเป็นขั้นตอนก่อนที่จะนำข้อมูลนั้นเข้าสู่กระบวนการจำแนกข้อมูลแต่เนื่องจากข้อมูลที่ได้คัดเลือกมานั้นอาจมีความไม่สมบูรณ์ เช่น อาจจะมีข้อมูลสูญหายไป (Missing Value) หรือ มีข้อมูลชนิดตัวเลข ดังนั้นจึงใช้ตัวกรองในโปรแกรม WEKAเวอร์ชัน 3.6.0 [23]

4.2.1. การแทนข้อมูลไม่ทราบค่า

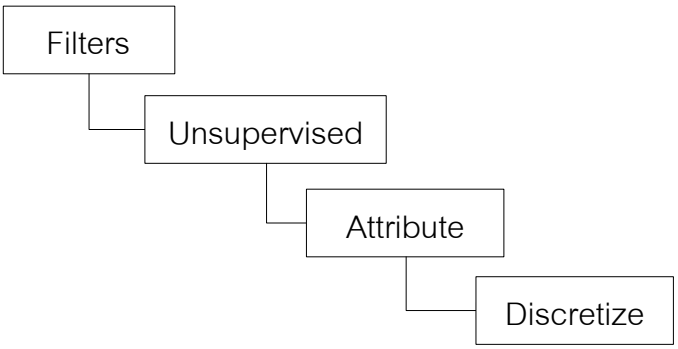
เนื่องจากข้อมูลมีคุณลักษณะบางอย่างของตัวอย่างประกอบด้วยข้อมูลที่ไม่ทราบค่าซึ่งมี 2 ทางเลือกในการจัดการกับข้อมูลประเภทนี้วิธีที่หนึ่งเอาตัวอย่างที่มีข้อมูลไม่ทราบค่าออก แต่เนื่องจากข้อมูลเรียนรู้น้อยลง จึงทำให้ข้อมูลที่มีประโยชน์อาจสูญหายไปได้ ดังนั้นผู้วิจัยจึงเลือกแนวทางที่สอง โดยการกำจัดข้อมูลสูญหายโดยการแทนด้วยค่าฐานนิยมสำหรับข้อมูลชนิดค่าคุณลักษณะที่เป็นค่าโนมินัล (nominal) และค่าเฉลี่ยสำหรับข้อมูลชนิดค่าคุณลักษณะที่เป็นตัวเลข (numeric) ด้วยฟังก์ชัน weka.filters.unsupervised.attribute.ReplaceMissingValues



ภาพที่ 4-3 การแทนค่าข้อมูลไม่ทราบค่า

4.2.2. การแปลงข้อมูลชนิดตัวเลขให้เป็นข้อมูลชนิดไม่ต่อเนื่อง

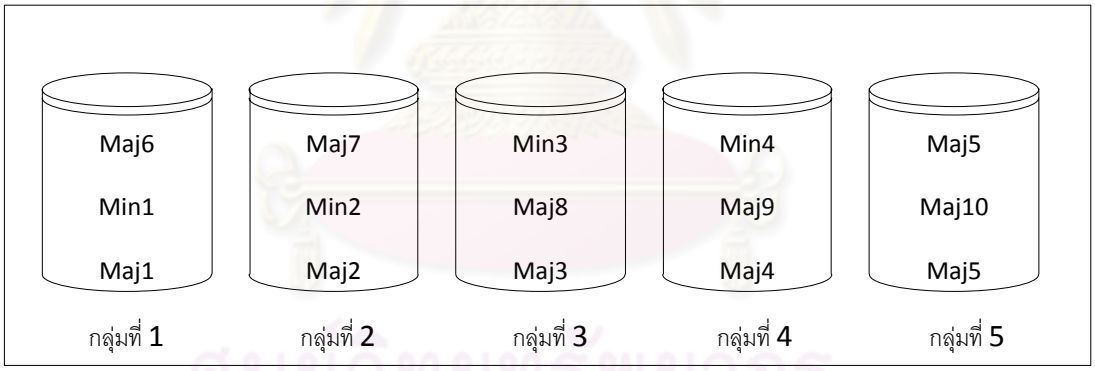
กรณีข้อมูลชนิดตัวเลขแปลงข้อมูลชนิดตัวเลขให้เป็นข้อมูลชนิดค่าไม่ต่อเนื่องด้วยการแบ่งข้อมูลชนิดตัวเลขในคุณลักษณะนั้นเป็นช่วงๆ 10 ช่วง ด้วยฟังก์ชัน weka.filters.unsupervised.attribute.discretize



ภาพที่ 4-4 การแปลงข้อมูลชนิดตัวเลขให้เป็นข้อมูลชนิดไม่ต่อเนื่อง

4.3. การทดสอบแบบไขว้ข้าม 5 กลุ่ม

วิทยานิพนธ์ฉบับนี้ใช้การทดสอบแบบไขว้ข้าม 5 กลุ่ม (five-fold cross-validation) ซึ่งเป็นวิธีการตรวจสอบการทำนายค่าความถูกต้องของแบบจำลองโดยการทำงานของการไขว้ข้าม 5 กลุ่มนั้นจะเริ่มจากแบ่งข้อมูลเป็น 5 กลุ่มเท่า ๆ กัน ในการทดลองนี้ได้แบ่งในแต่ละกลุ่มให้มีจำนวนตัวอย่างในกลุ่มที่มีมากประมาณ 20% และจำนวนตัวอย่างในกลุ่มที่มีน้อยประมาณ 20% และใช้ 4 กลุ่มเป็นชุดข้อมูลเรียนรู้ และ 1 กลุ่มเป็นข้อมูลทดสอบ สลับกันเป็นข้อมูลเรียนรู้และข้อมูลทดสอบ 5 ครั้ง เพื่อให้ทุกส่วนถูกใช้เป็นตัวทดสอบดังภาพที่ 4-2



ภาพที่ 4-5 การแบ่งข้อมูล 5 กลุ่ม

4.4. ตัววัดผล

ในส่วนนี้กล่าวถึง ตัววัดผลการทดลองในวิทยานิพนธ์ฉบับนี้ ได้แก่ ค่าความแม่นยำ (Accuracy) ค่าความแม่นยำในกลุ่มที่มีมาก (Accuracy of Majority) และค่าความแม่นยำในกลุ่มที่มีน้อย (Accuracy of Minority) แต่หากต้องอ้างทั้งสองกลุ่มจะทำให้กลุ่มตัวอย่างที่มีน้อยเป็นตัวอย่างที่เราสนใจถูกบดบัง ดังนั้นเพื่อให้เปรียบเทียบผลการทดลองได้ชัดเจนขึ้น ในการทดลองนี้เราจึงใช้ค่าความระลึก (Recall) ค่าความเที่ยง (Precision) และค่าเอฟ (F-measure) เฉพาะกลุ่มที่มีจำนวนตัวอย่างน้อย โดยการวัดค่าทั้งหมดนี้อาศัยการคำนวณจากตาราง

คอนฟิวชันเมทริกซ์แสดงดังตารางที่ 4-2 แถวของตารางแทนค่าจริงของกลุ่ม และคอลัมน์แทนค่าการทำนายของกลุ่ม

ตารางที่ 4-2 คอนฟิวชันเมทริกซ์

	ผลการทำนายตอบกลุ่มบวก	**ผลการทำนายตอบกลุ่มลบ**
ค่าจริงกลุ่มบวก	True Positive :TP	False Negative: FN
ค่าจริงกลุ่มลบ	False Positive :FP	True Negative : TN

ค่าต่างๆในตารางคอนฟิวชันเมทริกซ์อธิบายได้ดังนี้

1) ค่า True Positive (TP) คือ ค่าความถูกต้องในการจำแนกข้อมูล ซึ่งมีค่าที่แท้จริงอยู่ในกลุ่มบวก และผลการทำนายว่าอยู่ในกลุ่มบวก
2) ค่า False Positive (FP) คือ ค่าความผิดพลาดในการจำแนกข้อมูล ซึ่งมีค่าที่แท้จริงอยู่ในกลุ่มบวก แต่ผลการทำนายว่าอยู่ในกลุ่มลบ
3) ค่า True Negative (TN) คือ ค่าความถูกต้องในการจำแนกข้อมูล ซึ่งมีค่าที่แท้จริงอยู่ในกลุ่มลบ และผลการทำนายว่าอยู่ในกลุ่มลบ
4) ค่า False Negative (FN) คือ ค่าความผิดพลาดในการจำแนกข้อมูล ซึ่งมีค่าที่แท้จริงอยู่ในกลุ่มลบ แต่ผลการทำนายว่าอยู่ในกลุ่มบวก

4.4.1. ค่าความแม่นยำ

ค่าความแม่นยำ $(Accuracy)$: เป็นค่าความแม่นยำรวมทั้งหมด โดยพิจารณาจากข้อมูลทั้งหมดโดยคำนวณจากสมการ (1) ดังนี้

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (1)$$

ค่าความแม่นยำในกลุ่มที่มีน้อย $(Accuracy\ of\ Minority)$: เป็นค่าความแม่นยำในกลุ่มที่มีน้อย โดยพิจารณาจากข้อมูลในกลุ่มที่มีน้อย คำนวณจากสมการ (2) ดังนี้

$$Accuracy\ of\ Minority = \frac{TP}{TP + FN} \quad (2)$$

ค่าความแม่นยำในกลุ่มที่มีมาก $(Accuracy\ of\ Majority)$: เป็นค่าความแม่นยำในกลุ่มที่มีมาก โดยพิจารณาจากข้อมูลในกลุ่มที่มีมาก โดยคำนวณจากสมการ (3) ดังนี้

$$Accuracy\ of\ Majority = \frac{TN}{TN + FP} \tag{3}$$

4.4.2 ค่าความระลึก ค่าความเที่ยง ค่าเอฟ ที่ใช้ทดสอบเฉพาะในกลุ่มที่มีน้อย

1) ค่าความระลึก $(Recall)$: เป็นค่าที่ใช้ทดสอบประสิทธิภาพความแม่นยำหาได้จากค่าของข้อมูลที่ทำนายถูกต้องในกลุ่มที่มีน้อย โดยพิจารณาจากข้อมูลในกลุ่มที่มีน้อย โดยคำนวณจากสมการ (4) ดังนี้

$$Recall = \frac{TP}{TP\ + FN} \tag{4}$$

2) ค่าความเที่ยง $(Precision)$: คำนวณจากค่าของข้อมูลที่ทำนายถูกต้องในกลุ่มที่มีน้อย โดยพิจารณาจากจำนวนข้อมูลที่มีการทำนายในกลุ่มที่มีน้อย โดยคำนวณจากสมการ (5) ดังนี้

$$Precision\ = \frac{TP}{TP + FP} \tag{5}$$

3) ค่าเอฟ ($F - measure$): เป็นค่าที่วัดประสิทธิภาพโดยรวมเนื่องจากในการจำแนกกรณีที่มีค่าความระลึกสูง แต่อาจมีค่าความเที่ยงต่ำได้ หรือในกรณีมีค่าความระลึกต่ำแต่อาจมีค่าความเที่ยงสูงได้ ค่านี้คำนวณได้จากการเฉลี่ยค่าความระลึก และค่าความเที่ยง โดยคำนวณจากสมการ (6) ดังนี้

$$F - measure\ = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6}$$

4.2.3 ทดสอบสมมติฐานทางสถิติ

ทดสอบสมมติฐานทางสถิติเพื่อวิเคราะห์เปรียบเทียบความแตกต่างอย่างมีนัยสำคัญทางสถิติ (Level of Significant) ของค่าทีแบบทางเดียว (One Tail Paired t-test) โดยกำหนดระดับความมีนัยสำคัญทางสถิติที่ 0.05 และ 0.1

4.5. ผลการทดลอง และวิเคราะห์ผล

4.5.1. ผลการทดลองวัดค่าความแม่นยำ

ในส่วนนี้จะแสดงผลค่าความแม่นยำในกลุ่มที่มีมาก (Accuracy of Majority) ค่าความแม่นยำในกลุ่มที่มีน้อย (Accuracy of Minority) ค่าความแม่นยำทั้งหมด (Accuracy) ของการทดสอบแบบไขว้ข้าม 5 จากการทดลองจำแนกด้วยวิธีที่แตกต่างกัน คือ การจำแนกด้วยวิธี C4.5, AE, OCE และ AMIE ผู้วิจัยได้กำหนดค่าพารามิเตอร์ w ด้วยความน่าจะเป็นของกลุ่มที่มีจำนวนตัวอย่างน้อยในแต่ละชุดข้อมูลในตารางที่ 4-3

ตารางที่ 4-3 สรุปผลค่าความแม่นยำในแต่ละกลุ่ม และค่าความแม่นยำทั้งหมด

ชุดข้อมูล	อัลกอริทึม	ความแม่นยำกลุ่มมาก	ความแม่นยำกลุ่มน้อย	ความแม่นยำทั้งหมด
Glass	*C4.5	91.37	5.88	84.58%
	AE w=0.08	91.37	**11.76**	85.05%
	OCE w=0.08	92.39	**11.76**	85.98%
	AMIE	**94.92**	**11.76**	**88.32%**
Flag	C4.5	**98.87**	0	90.21%
	**AE w=0.08	90.4	**47.06**	86.60%
	**OCE w=0.08	90.4	**47.06**	86.60%
	AMIE	95.48	41.18	**90.72%**
Ecoli	C4.5	**93.02**	**48.57**	**88.39%**
	AE w=0.1	91.36	42.86	86.31%
	OCE w=0.1	91.36	42.86	86.31%
	AMIE	89.37	0.60	86.31%
Hepatitis	C4.5	**90.24**	43.75	80.65%
	AE w=0.2	89.43	53.12	81.94%
	OCE w=0.2	**90.24**	**62.5**	**84.52%**
	AMIE	86.18	59.38	80.65%
Vehicle	**C4.5	96.29	85.93	93.85%
	**AE w=0.25	93.51	82.41	90.90%
	**OCE w=0.25	93.97	83.42	91.49%
	AMIE	**96.91**	**90.45**	**95.39%**
Splice-ei	C4.5	**97.03**	94.39	**96.40%**
	**AE w=0.25	96.41	90.74	95.05%
	**OCE w=0.25	95.91	91.13	94.76%
	AMIE	96.66	**94.92**	96.24%
Splice-ie	C4.5	96.78	88.93	94.89%
	AE w=0.25	96.08	88.8	94.33%
	OCE w=0.25	**96.99**	89.58	**95.20%**
	AMIE	96.49	**90.23**	94.98%

Haberman	C4.5	**80.44**	32.10	67.65%
	AE w=0.25	78.22	**48.15**	**70.26%**
	OCE w=0.25	78.22	**48.15**	**70.26%**
	AMIE	78.67	33.33	66.67%
Post-operative	C4.5	**75.76**	8.33	57.78%
	AE w=0.25	74.24	20.83	60%
	OCE w=0.25	71.21	16.67	56.67%
	AMIE	71.21	**37.50**	**62.22%**
Breast-cancer	C4.5	**72.64**	**72.64**	**64.34%**
	AE w=0.3	63.18	63.53	63.29%
	OCE w=0.3	61.69	63.53	62.24%
	AMIE	70.15	49.41	63.99%
Credit_g	C4.5	**78.14**	43	**67.60%**
	AE w=0.3	71.57	47.67	64.40%
	*OCE w=0.3	70.14	**50.33**	64.20%
	AMIE	74.86	49	67.10%
Breast-cancer-w	C4.5	**96.29**	87.55	**93.28%**
	AE w=0.65	93.89	90.04	92.56%
	OCE w=0.65	94.1	**90.87**	92.99%
	AMIE	94.54	90.46	93.13%
Tic-tac-toes	C4.5	**90.58**	77.41	**86.01%**
	AE w=0.65	87.54	81.33	85.39%
	OCE w=0.65	87.54	**87.54**	85.39%
	AMIE	87.54	80.12	84.97%
Diabetes	C4.5	**81**	44.78	68.36%
	AE w=0.65	73.6	**62.69**	**69.79%**
	**OCE w=0.65	67.6	60.82	65.23%
	AMIE	74	59.70	69.01%
Ionosphere	*C4.5	88.44	79.37	85.19%
	AE w=0.65	87.11	**85.71**	86.61%
	OCE w=0.65	87.11	**85.71**	86.61%
	AMIE	**90.22**	**85.71**	**88.60%**
Liver	C4.5	**57**	60	58.26%
	*AE w=0.65	49	64.83	55.65%
	*OCE w=0.65	48.5	64.83	55.36%
	AMIE	54.5	**65.52**	**59.13%**

เครื่องหมาย * หมายถึงนัยสำคัญทางสถิติของค่าความแม่นยำทั้งหมดที่ระดับ 0.1

เครื่องหมาย** หมายถึงนัยสำคัญทางสถิติของค่าความแม่นยำทั้งหมดที่ระดับ 0.05

การทดลองได้เปรียบเทียบประสิทธิภาพการทำนายของอัลกอริทึม C4.5 AE OCE และ AMIE โดยทดสอบข้อมูลที่ไม่สมดุลทั้งหมด 16 ชุดตัวอย่าง เมื่อเปรียบเทียบค่าความแม่นยำวิธี AMIE กับวิธี

อื่น ๆ นั้น มากกว่า C4.5 8 ชุดข้อมูล เท่ากัน 1 และน้อยกว่า 7 ชุดข้อมูลมากกว่า AE 11 ชุดข้อมูล เท่ากัน 1 และน้อยกว่า 4 ชุดข้อมูล มากกว่า OCE 11 ชุดข้อมูล เท่ากัน 1 และน้อยกว่า 4 ชุดข้อมูล

ในส่วนนี้จะแสดงรายละเอียดจากตารางที่ 4-3 ผลค่าความแม่นยำในกลุ่มที่มีมาก (Accuracy of Majority) ค่าความแม่นยำในกลุ่มที่มีน้อย (Accuracy of Minority) ค่าความแม่นยำทั้งหมด (Accuracy) ของการทดสอบแบบไขว้ข้าม 5 กลุ่มแยกทีละครั้ง ใน 16 ชุดข้อมูล จากการทดลองจำแนกด้วยวิธีที่แตกต่างกัน คือ การจำแนกด้วยวิธี C4.5, AE, OCE และ AMIE ในตารางที่ 4-4 ถึง 4-19

ตารางที่ 4-4 ค่าความแม่นยำในแต่ละกลุ่ม และค่าความแม่นยำทั้งหมดของ glass

กลุ่ม	ความแม่นยำกลุ่มมาก				ความแม่นยำกลุ่มน้อย				ความแม่นยำทั้งหมด			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	100	100	100	100	0	0	0	0	90.91	90.91	90.91	90.91
2	100	100	100	100	0	0	0	0	90.91	90.91	90.91	90.91
3	82.05	89.74	89.74	87.18	0	0	0	0	76.19	83.33	83.33	80.95
4	89.74	82.05	87.18	87.18	0	33.33	33.33	66.67	83.33	78.57	83.33	85.71
5	84.62	84.62	84.62	100	33.33	33.33	33.33	0	80.95	80.95	80.95	92.86
ค่าเฉลี่ย	91.37	91.37	92.39	94.92	5.88	11.76	11.76	11.76	84.58	85.05	85.98	88.32

ตารางที่ 4-5 ค่าความแม่นยำในแต่ละกลุ่ม และค่าความแม่นยำทั้งหมดของ Flags

กลุ่ม	ความแม่นยำกลุ่มมาก				ความแม่นยำกลุ่มน้อย				ความแม่นยำทั้งหมด			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	97.22	94.44	94.44	97.22	0	50	50	25	87.50	90.00	90.00	90.0
2	100	94.44	94.44	97.22	0	50	50	50	90.00	90.00	90.00	92.50
3	97.14	82.86	82.86	88.57	0	33.33	33.33	66.67	89.47	78.95	78.95	86.84
4	100	91.43	91.43	97.14	0	33.33	33.33	33.33	92.11	86.84	86.84	92.11
5	100	88.57	88.57	97.14	0	66.67	66.67	33.33	92.11	86.84	86.84	92.11
ค่าเฉลี่ย	98.87	90.40	90.40	95.48	0	47.06	47.06	41.18	90.21	86.60	86.60	90.72

ตารางที่ 4-6 ค่าความแม่นยำในแต่ละกลุ่ม และค่าความแม่นยำทั้งหมดของ ecoli

กลุ่ม	ความแม่นยำกลุ่มมาก				ความแม่นยำกลุ่มน้อย				ความแม่นยำทั้งหมด			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	95.08	95.08	95.08	90.16	42.86	57.14	57.14	57.14	89.71	91.18	91.18	86.76
2	91.67	88.33	88.33	85	42.86	0	0	85.71	86.58	79.10	79.10	85.07
3	91.67	93.33	93.33	91.67	85.71	57.14	57.14	71.43	91.04	89.55	89.55	89.55
4	95	86.67	86.67	88.33	42.86	42.86	42.86	28.57	89.55	82.09	82.09	82.09
5	91.67	93.33	93.33	91.67	2857	57.14	57.14	57.14	85.07	89.55	89.55	88.06
ค่าเฉลี่ย	93.02	91.36	91.36	89.37	48.57	42.86	42.86	0.60	88.39	86.31	86.31	86.31

ตารางที่ 4-7 ค่าความแม่นยำในแต่ละกลุ่ม และค่าความแม่นยำทั้งหมดของ Hepatitis

กลุ่ม	ความแม่นยำกลุ่มมาก				ความแม่นยำกลุ่มน้อย				ความแม่นยำทั้งหมด			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	88	88	88	80	42.86	71.43	85.71	85.71	78.16	84.375	87.50	81.25
2	84	76	84	80	28.57	42.86	57.14	42.86	71.88	68.75	78.13	71.88
3	88	96	96	92	50	33.33	33.33	50	80.65	83.87	83.87	83.87
4	91.67	95.83	95.83	87.50	50	50	50	66.67	83.33	86.67	86.67	83.33
5	1	91.67	87.50	91.67	50	66.67	83.33	50	90.00	86.67	86.67	83.33
ค่าเฉลี่ย	90.24	89.43	90.24	86.18	43.75	53.12	62.50	59.38	80.65	81.94	84.57	80.65

ตารางที่ 4-8 ค่าความแม่นยำในแต่ละกลุ่ม และค่าความแม่นยำทั้งหมดของ Vehicle

กลุ่ม	ความแม่นยำกลุ่มมาก				ความแม่นยำกลุ่มน้อย				ความแม่นยำทั้งหมด			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	94.62	92.31	93.08	97.69	80	85	80	85	91.18	90.59	90.00	94.71
2	90	93.08	93.85	97.69	96.15	80	82.50	92.50	94.71	90.00	91.18	96.47
3	95.35	93.80	94.57	96.12	92.50	87.50	90	95	94.67	92.31	93.49	95.86
4	96.9	93.02	93.02	95.35	92.50	82.50	87.5	95	95.86	90.53	91.72	95.27
5	98.45	95.35	95.35	97.67	74.36	76.92	76.92	84.62	92.86	91.07	91.07	94.64
ค่าเฉลี่ย	96.29	93.51	93.97	96.91	85.93	82.41	83.42	90.45	93.85	90.90	91.49	95.39

ตารางที่ 4-9 ค่าความแม่นยำในแต่ละกลุ่ม และค่าความแม่นยำทั้งหมดของ Splice-ei

กลุ่ม	ความแม่นยำกลุ่มมาก				ความแม่นยำกลุ่มน้อย				ความแม่นยำทั้งหมด			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	96.29	95.46	94.85	96.29	91.56	90.26	92.86	92.86	95.15	94.21	94.37	95.46
2	96.70	96.08	95.67	95.88	94.81	92.21	88.31	96.10	96.24	95.15	93.90	95.93
3	96.70	95.88	94.23	96.70	92.16	87.58	90.85	92.81	95.61	93.89	93.42	95.77
4	97.93	97.11	97.31	96.90	98.04	92.81	92.16	98.04	97.96	96.08	96.08	97.17
5	97.52	97.52	97.52	97.52	95.42	90.85	91.5	94.77	97.02	95.92	96.08	96.86
ค่าเฉลี่ย	97.03	96.41	95.91	96.66	94.39	90.74	91.13	94.92	96.40	95.05	94.76	96.24

ตารางที่ 4-10 ค่าความแม่นยำในแต่ละกลุ่ม และค่าความแม่นยำทั้งหมดของ Splice-ie

กลุ่ม	ความแม่นยำกลุ่มมาก				ความแม่นยำกลุ่มน้อย				ความแม่นยำทั้งหมด			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	96.29	97.11	97.94	96.29	85.71	85.06	84.42	83.12	93.74	94.21	94.68	93.11
2	95.88	95.88	96.91	96.29	94.81	88.31	90.91	94.16	95.62	94.05	95.46	95.77
3	97.31	95.45	95.66	95.66	79.22	89.61	89.61	88.31	92.95	94.04	94.20	93.89
4	97.52	96.07	97.31	97.31	88.89	90.85	89.54	89.54	95.45	94.82	95.45	95.45
5	96.90	95.87	97.11	96.90	96.08	90.20	93.46	96.08	96.70	94.51	96.23	96.70
ค่าเฉลี่ย	96.78	96.08	96.99	96.49	88.93	88.8	89.58	90.23	94.89	94.33	95.20	94.98

ตารางที่ 4-11 ค่าความแม่นยำในแต่ละกลุ่ม และค่าความแม่นยำทั้งหมดของ Haberman

กลุ่ม	ความแม่นยำกลุ่มมาก				ความแม่นยำกลุ่มน้อย				ความแม่นยำทั้งหมด			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	75.56	77.78	77.78	73.33	35.29	41.18	41.18	29.41	64.52	67.74	67.74	61.29
2	84.44	84.44	84.44	84.44	43.75	50	50	37.50	73.77	75.41	75.41	72.13
3	86.67	84.44	84.44	86.67	37.50	43.75	43.75	31.25	73.77	73.77	73.77	72.13
4	82.22	73.33	73.33	71.11	12.5	50	50	43.75	63.93	67.21	67.21	63.93
5	73.33	71.11	71.11	77.78	31.25	56.25	56.25	25	62.30	67.21	67.21	63.93
ค่าเฉลี่ย	80.44	78.22	78.22	78.67	32.10	48.15	48.15	33.33	67.65	70.26	70.26	66.67

ตารางที่ 4-12 ค่าความแม่นยำในแต่ละกลุ่ม และค่าความแม่นยำทั้งหมดของ Post-operative

กลุ่ม	ความแม่นยำกลุ่มมาก				ความแม่นยำกลุ่มน้อย				ความแม่นยำทั้งหมด			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	78.57	71.43	71.43	78.57	20	20	20	60	63.16	57.89	57.89	73.68
2	61.54	84.62	84.62	53.85	20	40	40	60	50.0	72.22	72.22	55.56
3	61.54	61.54	61.54	76.92	0	0	0	0	44.44	44.44	44.44	55.56
4	100	92.31	92.31	76.92	0	20	20	20	72.22	72.22	72.22	61.11
5	76.92	61.54	46.15	69.23	0	25	0	50	58.82	52.94	35.29	64.71
ค่าเฉลี่ย	75.76	74.24	71.21	71.21	8.33	20.83	16.67	37.5	57.78	60	56.67	62.22

ตารางที่ 4-13 ค่าความแม่นยำในแต่ละกลุ่ม และค่าความแม่นยำทั้งหมดของ Breast-cancer

กลุ่ม	ความแม่นยำกลุ่มมาก				ความแม่นยำกลุ่มน้อย				ความแม่นยำทั้งหมด			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	68.29	68.29	68.29	63.41	64.71	88.24	82.35	58.82	67.24	74.14	72.41	62.06
2	70	57.50	47.50	77.50	23.53	58.82	58.82	47.06	56.14	57.89	50.88	68.42
3	82.50	60	60	75	41.18	58.82	58.82	35.29	70.18	59.65	59.65	63.16
4	72.50	67.50	70	72.50	41.18	58.82	58.82	47.06	63.16	64.91	66.67	64.91
5	70	62.50	62.5	62.50	52.94	52.94	58.82	58.82	64.91	59.65	61.40	61.40
ค่าเฉลี่ย	72.64	63.18	61.69	70.15	72.64	63.53	63.53	49.41	64.34	63.29	62.24	63.99

ตารางที่ 4-14 ค่าความแม่นยำในแต่ละกลุ่ม และค่าความแม่นยำทั้งหมดของ Credit-g

กลุ่ม	ความแม่นยำกลุ่มมาก				ความแม่นยำกลุ่มน้อย				ความแม่นยำทั้งหมด			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	74.29	69.29	70	73.57	48.33	40	45	60	66.50	60.5	62.50	69.50
2	81.43	74.29	70	70.71	33.33	43.33	53.33	45	67.00	65.00	65.00	63.00
3	87.14	73.57	75	77.86	40	53.33	51.67	45	73.00	67.50	68.00	68.0
4	75.71	67.86	71.43	79.29	55	51.67	51.67	50	69.50	63.00	65.50	70.50
5	72.14	72.86	64.29	72.86	38.33	0.50	50	45	62.00	66.00	60.00	0.50
ค่าเฉลี่ย	78.14	71.57	70.14	74.86	43	47.67	50.33	49	67.60	64.40	64.20	67.10

ตารางที่ 4-15 ค่าความแม่นยำในแต่ละกลุ่ม และค่าความแม่นยำทั้งหมดของ Breast-cancer-w

กลุ่ม	ความแม่นยำกลุ่มมาก				ความแม่นยำกลุ่มน้อย				ความแม่นยำทั้งหมด			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	97.83	94.57	96.74	92.39	83.67	89.80	93.88	85.71	92.91	92.91	95.74	90.07
2	95.65	94.57	94.57	93.48	81.25	87.50	87.50	87.50	90.71	92.14	92.14	91.43
3	95.65	91.30	91.3	93.48	87.50	89.58	89.58	89.58	92.86	90.71	90.71	92.14
4	96.70	93.41	93.41	96.70	89.58	91.67	91.67	93.75	94.24	92.81	92.86	95.68
5	95.60	95.60	94.51	96.70	95.83	91.67	91.67	95.83	95.68	94.24	93.53	96.40
ค่าเฉลี่ย	96.29	93.89	94.1	94.54	87.55	90.04	90.87	90.46	93.28	92.56	92.99	93.13

ตารางที่ 4-16 ค่าความแม่นยำในแต่ละกลุ่ม และค่าความแม่นยำทั้งหมดของ Tic-tac-toes

กลุ่ม	ความแม่นยำกลุ่มมาก				ความแม่นยำกลุ่มน้อย				ความแม่นยำทั้งหมด			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	92.86	87.30	89.68	88.89	70.15	86.57	85.07	79.10	84.97	87.05	88.08	85.49
2	88.80	84.80	83.20	86.40	85.07	85.07	85.07	85.07	87.50	84.90	83.85	85.94
3	88.80	88.80	85.60	81.60	74.24	72.73	71.21	75.76	83.77	83.25	80.63	79.58
4	92.80	91.20	91.20	94.40	77.27	77.27	75.76	77.27	87.43	86.39	85.86	88.48
5	89.60	85.60	88	86.40	80.3	84.85	89.39	83.33	86.39	85.34	88.48	85.34
ค่าเฉลี่ย	90.58	87.54	87.54	87.54	77.41	81.33	87.54	80.12	86.01	85.39	85.39	84.97

ตารางที่ 4-17 ค่าความแม่นยำในแต่ละกลุ่ม และค่าความแม่นยำทั้งหมดของ Diabetes

กลุ่ม	ความแม่นยำกลุ่มมาก				ความแม่นยำกลุ่มน้อย				ความแม่นยำทั้งหมด			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	77	73	68	75	53.7	55.56	59.26	48.15	68.83	66.88	64.94	65.58
2	80	69	68	76	50	61.11	62.96	68.52	69.48	66.23	66.23	73.38
3	83	65	65	71	33.33	64.81	65	55.56	65.58	64.94	64.94	0.69
4	83	79	73	75	49.06	66.04	73	66.04	71.24	74.51	66.01	71.90
5	82	82	64	73	37.74	66.04	64.15	60.38	66.67	76.47	64.05	68.63
ค่าเฉลี่ย	81	73.60	67.60	74	44.78	62.69	60.82	59.70	68.36	69.79	65.23	69.01

ตารางที่ 4-18 ค่าความแม่นยำในแต่ละกลุ่ม และค่าความแม่นยำทั้งหมดของ Ionosphere

กลุ่ม	ความแม่นยำกลุ่มมาก				ความแม่นยำกลุ่มน้อย				ความแม่นยำทั้งหมด			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	91.11	82.22	82.22	91.11	73.08	100	100	92.31	84.51	88.73	88.73	91.55
2	91.11	93.33	93.33	91.11	72	76	76	72	84.29	87.14	87.14	84.29
3	95.56	91.11	91.11	95.56	76	72	72	80	88.57	84.29	84.29	90.00
4	86.67	88.89	88.89	86.67	92	100	100	92	88.57	92.86	92.86	88.57
5	77.78	80	80	86.67	84	80	80	92	80.0	80.00	80.00	88.57
ค่าเฉลี่ย	88.44	87.11	87.11	90.22	79.37	85.71	85.71	85.71	85.19	86.61	86.61	88.60

ตารางที่ 4-19 ค่าความแม่นยำในแต่ละกลุ่ม และค่าความแม่นยำทั้งหมดของ Liver

กลุ่ม	ความแม่นยำกลุ่มมาก				ความแม่นยำกลุ่มน้อย				ความแม่นยำทั้งหมด			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	40	55	55	50	72.41	58.62	58.62	68.97	53.62	56.52	56.52	57.97
2	72.50	42.50	42.50	57.50	51.72	72.41	72.41	68.97	63.77	55.07	55.07	62.32
3	47.50	47.50	50	50	62.07	65.52	65.52	62.07	53.62	55.07	56.52	55.07
4	62.50	52.50	47.50	60	55.17	75.86	75.86	62.07	59.42	62.32	59.42	60.87
5	62.50	47.50	47.50	55	58.62	51.72	51.72	65.52	60.87	49.27	49.28	59.42
ค่าเฉลี่ย	57	49	48.50	54.50	60	64.83	64.83	65.52	58.26	55.65	55.36	59.13

4.5.2. ผลการทดลองเฉพาะในกลุ่มที่มีน้อย

เนื่องจากงานวิจัยนี้สนใจปัญหาการเรียนรู้ข้อมูลไม่สมดุล เพื่อต้องการเพิ่มประสิทธิภาพในกลุ่มที่มีน้อยให้ดีขึ้น ดังนั้นจึงได้ทำการทดลองเปรียบเทียบเฉพาะค่าความระลึก ค่าความเที่ยง และค่าเอฟในกลุ่มที่มีน้อยด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามทั้ง 5 ครั้ง ในแต่ละชุดข้อมูลอัลกอริทึมที่ให้ค่าความระลึก ค่าความเที่ยง และค่าเอฟสูงสุดจะถูกแสดงเป็นตัวหนา สรุปได้ดังตารางที่ 4-20

ตารางที่ 4-20 สรุปผลค่าความระลึก ค่าความเที่ยง และค่าเอฟ ในกลุ่มที่มีน้อย

ชุดข้อมูล	อัลกอริทึม	ความระลึก	ความเที่ยง	เอฟ
Glass	C4.5	0.06	0.06	0.06
	AE w=0.08	**0.12**	0.11	0.11
	OCE w=0.08	**0.12**	0.17	0.12
	AMIE	**0.12**	**0.17**	**0.14**
Flag	C4.5	0	0	0
	AE w=0.08	**0.47**	0.32	0.38
	OCE w=0.08	**0.47**	0.32	0.38
	AMIE	0.41	**0.47**	**0.44**
Ecoli	C4.5	0.49	**0.45**	0.47
	AE w=0.1	0.43	0.37	0.40
	OCE w=0.1	0.43	0.37	0.39
	AMIE	**0.60**	0.40	**0.48**
Hepatitis	C4.5	0.44	0.54	0.48
	AE w=0.2	0.53	0.57	0.55
	OCE w=0.2	**0.63**	**0.63**	**0.63**
	AMIE	0.59	0.53	0.56
Vehicle	C4.5	0.86	0.88	0.87
	AE w=0.25	0.82	0.80	0.81
	OCE w=0.25	0.83	0.81	0.82
	AMIE	**0.90**	**0.90**	**0.90**

Splice-ei	C4.5	0.94	**0.90**	**0.93**
	AE w=0.25	0.91	0.89	0.90
	OCE w=0.25	0.91	0.88	0.89
	AMIE	**0.95**	0.90	0.92
Splice-ie	C4.5	0.89	0.90	0.89
	AE w=0.25	0.89	0.88	0.88
	OCE w=0.25	0.90	**0.90**	**0.90**
	AMIE	**0.90**	0.89	0.90
Haberman	C4.5	0.32	0.37	0.34
	AE w=0.75	**0.48**	**0.44**	**0.46**
	OCE w=0.75	**0.48**	**0.44**	**0.46**
	AMIE	0.33	0.36	0.35
Post-operative	C4.5	0.08	0.11	0.09
	AE w=0.25	0.21	0.23	0.22
	OCE w=0.25	0.17	0.17	0.17
	AMIE	**0.38**	**0.32**	**0.35**
Breast-cancer	C4.5	0.45	0.41	0.43
	AE w=0.3	**0.64**	**0.42**	**0.51**
	OCE w=0.3	**0.64**	0.41	0.50
	AMIE	0.49	0.41	0.45
Credit_g	C4.5	0.43	**0.46**	0.44
	AE w=0.3	0.48	0.42	0.45
	OCE w=0.3	**0.50**	0.42	0.46
	AMIE	0.49	0.45	**0.47**
Breast-cancer-w	C4.5	0.88	**0.93**	0.89
	AE w=0.35	0.90	0.89	0.89
	OCE w=0.35	**0.91**	0.89	0.89
	AMIE	0.90	0.89	**0.90**
Tic-tac-toes	C4.5	0.77	**0.81**	0.79
	AE w=0.35	**0.81**	0.77	**0.79**
	OCE w=0.35	**0.81**	0.77	**0.79**
	AMIE	0.80	0.77	0.78
Diabetes	C4.5	0.44	0.55	0.49
	AE w=0.35	**0.62**	**0.56**	**0.59**
	OCE w=0.35	0.60	0.50	0.55
	AMIE	0.59	0.55	0.57
Ionosphere	C4.5	0.79	0.79	0.79
	AE w=0.35	**0.86**	0.79	0.82
	OCE w=0.35	**0.86**	0.79	0.82
	AMIE	**0.86**	**0.83**	**0.84**
Liver	C4.5	0.60	0.50	0.54
	AE w=0.4	0.64	0.47	0.55
	OCE w=0.4	0.64	0.47	0.55
	AMIE	**0.65**	**0.51**	**0.57**

ในส่วนนี้จะแสดงรายละเอียดจากตารางที่ 4-4 ผลค่าความระลึก ค่าความเที่ยง และค่าเอฟของการทดสอบแบบไขว้ข้าม 5 กลุ่มแยกทีละครั้ง ใน 16 ชุดข้อมูล จากการทดลองจำแนกด้วยวิธีที่แตกต่างกัน คือ การจำแนกด้วยวิธี C4.5, AE, OCE และ AMIE โดยพิจารณาเฉพาะในกลุ่มที่มีจำนวนตัวอย่างน้อย และใช้ค่าเฉลี่ยของการทดสอบทั้ง 5 ครั้ง ดังในตารางที่ 4-21 ถึง 4-36

ตารางที่ 4-21 ค่าความระลึก ค่าความเที่ยง และค่าเอฟของข้อมูล Glasses

กลุ่ม	ความระลึก				ความเที่ยง				เอฟ			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0.33	0.33	0.66	0	0.12	0.16	0.29	0	0.18	0.22	0.4
5	0.33	0.33	0.33	0	0.14	0.14	0.14	0	0.2	0.2	0.2	0
ค่าเฉลี่ย	0.06	0.12	0.12	0.12	0.06	0.11	0.17	0.17	0.06	0.11	0.12	0.14

ตารางที่ 4-22 ค่าความระลึก ค่าความเที่ยง และค่าเอฟของเซตข้อมูล Flags

กลุ่ม	ความระลึก				ความเที่ยง				เอฟ			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	0	0.5	0.5	0.25	0	0.5	0.5	0.5	0	0.5	0.5	0.33
2	0	0.5	0.5	0.5	0	0.5	0.5	0.66	0	0.5	0.5	0.57
3	0	0.33	0.33	0.66	0	0.14	0.14	0.33	0	0.2	0.2	0.44
4	0	0.33	0.33	0.33	0	0.25	0.25	0.5	0	0.28	0.28	0.4
5	0	0.66	0.66	0.33	0	0.33	0.33	0.5	0	0.44	0.44	0.4
ค่าเฉลี่ย	0	0.47	0.47	0.41	0	0.32	0.32	0.47	0	0.38	0.38	0.44

ตารางที่ 4-23 ค่าความระลึก ค่าความเที่ยง และค่าเอฟของเซตข้อมูล Ecoli

กลุ่ม	ความระลึก				Precision				เอฟ			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	0.42	0.57	0.57	0.57	0.5	0.57	0.57	0.4	0.46	0.57	0.57	0.47
2	0.42	0	0	0.85	0.37	0	0	0.4	0.4	0	0	0.54
3	0.85	0.57	0.57	0.71	0.54	0.5	0.5	0.5	0.66	0.53	0.53	0.58
4	0.42	0.42	0.42	0.28	0.5	0.27	0.27	0.22	0.46	0.33	0.33	0.25
5	0.28	0.57	0.57	0.57	0.28	0.5	0.5	0.44	0.28	0.53	0.53	0.5
ค่าเฉลี่ย	0.49	0.43	0.43	0.60	0.45	0.37	0.37	0.40	0.47	0.40	0.39	0.48

ตารางที่ 4-24 ค่าความระลึก ค่าความเที่ยง และค่าเอฟของเซตข้อมูล Hepatitis

กลุ่ม	ความระลึก				ความเที่ยง				เอฟ			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	0.42	0.71	0.85	0.85	0.5	0.62	0.66	0.54	0.46	0.66	0.75	0.66
2	0.28	0.42	0.57	0.42	0.33	0.33	0.5	0.37	0.30	0.375	0.53	0.4
3	0.5	0.33	0.33	0.5	0.5	0.66	0.66	0.6	0.5	0.44	0.44	0.54
4	0.5	0.5	0.5	0.66	0.6	0.75	0.75	0.57	0.54	0.6	0.6	0.61
5	0.5	0.66	0.83	0.5	1	0.66	0.62	0.6	0.66	0.66	0.71	0.54
ค่าเฉลี่ย	0.44	0.53	0.62	0.59	0.54	0.57	0.63	0.53	0.48	0.55	0.63	0.56

ตารางที่ 4-25 ค่าความระลึก ค่าความเที่ยง และค่าเอฟของเซตข้อมูล Vehicle

กลุ่ม	ความระลึก				ความเที่ยง				เอฟ			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	0.8	0.85	0.8	0.85	0.82	0.77	0.78	0.91	0.81	0.80	0.79	0.88
2	0.9	0.8	0.82	0.92	0.87	0.78	0.80	0.92	0.88	0.79	0.81	0.92
3	0.92	0.87	0.9	0.95	0.86	0.81	0.83	0.88	0.89	0.84	0.86	0.91
4	0.92	0.82	0.87	0.95	0.90	0.78	0.79	0.86	0.91	0.80	0.83	0.90
5	0.74	0.76	0.76	0.84	0.93	0.83	0.83	0.91	0.82	0.8	0.8	0.88
ค่าเฉลี่ย	0.86	0.82	0.83	0.90	0.88	0.80	0.81	0.90	0.87	0.81	0.82	0.90

ตารางที่ 4-26 ค่าความระลึก ค่าความเที่ยง และค่าเอฟของเซตข้อมูล Splice-ei

กลุ่ม	ความระลึก				ความเที่ยง				เอฟ			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	0.91	0.90	0.92	0.92	0.88	0.86	0.85	0.88	0.90	0.88	0.88	0.90
2	0.94	0.92	0.88	0.96	0.90	0.88	0.86	0.881	0.92	0.90	0.87	0.91
3	0.92	0.87	0.90	0.92	0.89	0.87	0.83	0.89	0.90	0.87	0.86	0.91
4	0.98	0.92	0.92	0.98	0.93	0.91	0.91	0.90	0.95	0.91	0.91	0.94
5	0.95	0.90	0.91	0.94	0.92	0.92	0.92	0.92	0.93	0.91	0.91	0.93
ค่าเฉลี่ย	0.94	0.91	0.91	0.95	0.90	0.89	0.88	0.90	0.93	0.90	0.89	0.92

ตารางที่ 4-27 ค่าความระลึก ค่าความเที่ยง และค่าเอฟของเซตข้อมูล Splice-ie

กลุ่ม	ความระลึก				ความเที่ยง				เอฟ			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	0.85	0.85	0.84	0.83	0.88	0.90	0.92	0.87	0.868	0.87	0.88	0.85
2	0.94	0.88	0.90	0.94	0.87	0.87	0.90	0.88	0.91	0.87	0.90	0.91
3	0.79	0.89	0.89	0.88	0.90	0.86	0.86	0.86	0.84	0.8	0.88	0.87
4	0.88	0.90	0.89	0.89	0.91	0.87	0.91	0.91	0.90	0.89	0.90	0.90
5	0.96	0.90	0.93	0.96	0.90	0.87	0.91	0.90	0.93	0.88	0.92	0.93
ค่าเฉลี่ย	0.89	0.89	0.90	0.90	0.90	0.88	0.90	0.89	0.89	0.88	0.89	0.90

ตารางที่ 4-28 ค่าความระลึก ค่าความเที่ยง และค่าเอฟของเซตข้อมูล Haberman

กลุ่ม	ความระลึก				ความเที่ยง				เอฟ			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	0.35	0.41	0.41	0.29	0.35	0.41	0.41	0.29	0.35	0.41	0.41	0.29
2	0.43	0.5	0.5	0.3	0.5	0.53	0.53	0.46	0.46	0.51	0.51	0.41
3	0.37	0.43	0.43	0.31	0.5	0.5	0.5	0.45	0.42	0.46	0.46	0.37
4	0.12	0.5	0.5	0.43	0.2	0.4	0.4	0.35	0.15	0.44	0.44	0.38
5	0.31	0.56	0.56	0.25	0.29	0.40	0.40	0.28	0.30	0.47	0.47	0.26
ค่าเฉลี่ย	0.32	0.48	0.48	0.33	0.37	0.44	0.44	0.36	0.34	0.46	0.46	0.35

ตารางที่ 4-29 ค่าความระลึก ค่าความเที่ยง และค่าเอฟของเซตข้อมูล Postoperative

กลุ่ม	ความระลึก				ความเที่ยง				เอฟ			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	0.2	0.2	0.2	0.6	0.25	0.2	0.2	0.5	0.22	0.2	0.2	0.54
2	0.2	0.4	0.4	0.6	0.16	0.5	0.5	0.33	0.18	0.44	0.44	0.42
3	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0.2	0.2	0.2	0	0.5	0.5	0.25	0	0.28	0.28	0.22
5	0	0.25	0	0.5	0	0.16	0	0.33	0	0.2	0	0.4
ค่าเฉลี่ย	0.08	0.20	0.16	0.37	0.11	0.22	0.17	0.32	0.09	0.21	0.17	0.34

ตารางที่ 4-30 ค่าความระลึก ค่าความเที่ยง และค่าเอฟของเซตข้อมูล Breast-cancer

กลุ่ม	ความระลึก				ความเที่ยง				เอฟ			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	0.64	0.88	0.82	0.58	0.45	0.53	0.51	0.4	0.53	0.66	0.63	0.47
2	0.23	0.58	0.58	0.47	0.25	0.37	0.32	0.47	0.24	0.45	0.41	0.47
3	0.41	0.58	0.58	0.35	0.5	0.38	0.38	0.37	0.45	0.46	0.46	0.36
4	0.41	0.58	0.58	0.47	0.38	0.43	0.45	0.42	0.4	0.5	0.51	0.44
5	0.52	0.52	0.58	0.58	0.42	0.37	0.4	0.4	0.47	0.43	0.47	0.47
ค่าเฉลี่ย	0.45	0.64	0.64	0.49	0.41	0.42	0.41	0.41	0.43	0.51	0.50	0.45

ตารางที่ 4-31 ค่าความระลึก ค่าความเที่ยง และค่าเอฟของเซตข้อมูล Credit_g

กลุ่ม	ความระลึก				ความเที่ยง				เอฟ			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	0.48	0.4	0.45	0.6	0.44	0.35	0.39	0.49	0.46	0.37	0.4186	0.5414
2	0.33	0.43	0.53	0.45	0.43	0.41	0.43	0.39	0.37	0.42	0.47	0.42
3	0.4	0.53	0.51	0.45	0.57	0.46	0.46	0.46	0.47	0.49	0.49	0.45
4	0.55	0.51	0.51	0.5	0.49	0.40	0.43	0.50	0.51	0.45	0.47	0.50
5	0.38	0.5	0.5	0.45	0.37	0.44	0.37	0.41	0.37	0.46	0.42	0.43
ค่าเฉลี่ย	0.43	0.48	0.50	0.49	0.46	0.42	0.42	0.45	0.44	0.45	0.46	0.47

ตารางที่ 4-32 ค่าความระลึก ค่าความเที่ยง และค่าเอฟของเซตข้อมูล Breast-cancer-w

กลุ่ม	ความระลึก				ความเที่ยง				เอฟ			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	0.84	0.90	0.94	0.86	0.95	0.90	0.93	0.86	0.89	0.90	0.93	0.86
2	0.81	0.87	0.88	0.88	0.90	0.89	0.90	0.88	0.86	0.88	0.88	0.87
3	0.88	0.90	0.90	0.90	0.91	0.84	0.84	0.88	0.89	0.87	0.86	0.89
4	0.90	0.92	0.92	0.94	0.93	0.88	0.88	0.94	0.91	0.90	0.90	0.94
5	0.96	0.92	0.92	0.96	0.92	0.92	0.90	0.94	0.94	0.92	0.91	0.95
ค่าเฉลี่ย	0.88	0.90	0.91	0.90	0.93	0.89	0.89	0.89	0.89	0.89	0.89	0.90

ตารางที่ 4-33 ค่าความระลึก ค่าความเที่ยง และค่าเอฟของเซตข้อมูล Tic-tac-toes

กลุ่ม	ความระลึก				ความเที่ยง				เอฟ			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	0.70	0.86	0.85	0.79	0.83	0.78	0.81	0.79	0.76	0.82	0.83	0.7
2	0.85	0.85	0.85	0.85	0.80	0.75	0.73	0.77	0.82	0.79	0.78	0.80
3	0.74	0.72	0.71	0.75	0.77	0.77	0.72	0.68	0.75	0.75	0.71	0.71
4	0.77	0.77	0.75	0.77	0.85	0.82	0.81	0.87	0.80	0.79	0.78	0.82
5	0.80	0.84	0.89	0.83	0.80	0.75	0.79	0.76	0.8	0.8	0.84	0.79
ค่าเฉลี่ย	0.77	0.81	0.81	0.80	0.81	0.77	0.77	0.77	0.79	0.79	0.79	0.78

ตารางที่ 4-34 ค่าความระลึก ค่าความเที่ยง และค่าเอฟของเซตข้อมูล Diabetes

กลุ่ม	ความระลึก				ความเที่ยง				เอฟ			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	0.53	0.55	0.59	0.48	0.55	0.52	0.5	0.50	0.54	0.54	0.54	0.49
2	0.5	0.61	0.62	0.68	0.57	0.51	0.51	0.60	0.53	0.55	0.56	0.64
3	0.33	0.64	0.64	0.55	0.51	0.5	0.5	0.50	0.40	0.56	0.56	0.53
4	0.49	0.66	0.52	0.66	0.60	0.62	0.50	0.58	0.54	0.64	0.51	0.61
5	0.37	0.66	0.64	0.60	0.52	0.66	0.48	0.54	0.43	0.66	0.55	0.57
ค่าเฉลี่ย	0.44	0.62	0.60	0.59	0.55	0.56	0.50	0.55	0.49	0.59	0.55	0.57

ตารางที่ 4-35 ค่าความระลึก ค่าความเที่ยง และค่าเอฟของเซตข้อมูล Ionosphere

กลุ่ม	ความระลึก				ความเที่ยง				เอฟ			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	0.73	1	1	0.92	0.82	0.76	0.76	0.85	0.77	0.86	0.86	0.88
2	0.72	0.76	0.76	0.72	0.81	0.86	0.86	0.81	0.766	0.80	0.80	0.76
3	0.76	0.72	0.72	0.8	0.90	0.81	0.81	0.90	0.82	0.76	0.76	0.85
4	0.92	1	1	0.92	0.79	0.83	0.83	0.79	0.85	0.90	0.90	0.85
5	0.84	0.8	0.8	0.92	0.67	0.68	0.68	0.79	0.75	0.74	0.74	0.85
ค่าเฉลี่ย	0.79	0.86	0.86	0.86	0.79	0.79	0.79	0.83	0.79	0.82	0.82	0.84

ตารางที่ 4-36 ค่าความระลึก ค่าความเที่ยง และค่าเอฟของเซตข้อมูล Liver

กลุ่ม	ความระลึก				ความเที่ยง				เอฟ			
	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE	C4.5	AE	OCE	AMIE
1	0.72	0.58	0.58	0.68	0.46	0.48	0.48	0.5	0.57	0.53	0.53	0.58
2	0.51	0.72	0.72	0.68	0.57	0.47	0.47	0.54	0.55	0.57	0.57	0.60
3	0.62	0.65	0.65	0.62	0.46	0.4	0.48	0.47	0.52	0.55	0.55	0.53
4	0.55	0.75	0.75	0.62	0.51	0.53	0.51	0.52	0.53	0.62	0.61	0.57
5	0.58	0.51	0.51	0.65	0.53	0.41	0.41	0.51	0.55	0.46	0.46	0.57
ค่าเฉลี่ย	0.6	0.64	0.64	0.65	0.50	0.47	0.47	0.51	0.54	0.55	0.55	0.57

และได้เปรียบเทียบค่าเอฟ และค่านัยสำคัญทางสถิติ (Significant) ที่ระดับ 0.1 ของทั้ง 4 อัลกอริทึมแสดงดังตารางที่ 4-37 ซึ่งได้เปรียบเทียบวิธีที่เสนอกับวิธีอื่นๆแบบทางเดียว คอลัมน์ 2, 3 แสดงจำนวนชุดข้อมูลเมื่อเปรียบเทียบค่าเอฟ และค่านัยสำคัญทางสถิติ

ตารางที่ 4-37 เปรียบเทียบค่าเอฟ และค่านัยสำคัญทางสถิติ

	เอฟ	**นัยสำคัญทางสถิติ**
AMIE vs. C4.5		
AMIE wins	14	7
C4.5 wins	2	0
AMIE vs. AE		
AMIE wins	12	2
AE wins	4	1
AMIE vs. OCE		
AMIE wins	11	2
OCE wins	5	1

ในส่วนการทดลองได้เปรียบเทียบประสิทธิภาพการทำนายของอัลกอริทึม C4.5, AE, OCE และ AMIE โดยทดสอบข้อมูลที่ไม่สมดุลทั้งหมด 16 ชุดตัวอย่าง เมื่อเปรียบเทียบค่าเอฟ วิธี AMIE กับวิธีอื่นๆนั้น มากกว่า C4.5 14 ชุดข้อมูล และน้อยกว่า 2 ชุดข้อมูล มากกว่า AE 12 ชุดข้อมูล และน้อยกว่า 4 ชุดข้อมูล มากกว่า OCE 11 ชุดข้อมูล และน้อยกว่า 5 ชุดชุดข้อมูล และเมื่อพิจารณาค่าเอฟร่วมกับค่านัยสำคัญทางสถิติที่ระดับ 0.1 พบว่า AMIE ดีกว่าวิธีอื่นๆ

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1. สรุปผลการวิจัย

ในงานวิจัยนี้ได้ศึกษาตัวจำแนกพื้นฐานที่นิยมใช้กันในการทำเหมืองข้อมูล และวิธีนี้มีจุดอ่อนที่ไม่สามารถทำงานได้ดีเมื่อข้อมูลมีความไม่สมดุล เพราะวิธีนี้ออกแบบมาสำหรับการจำแนกสำหรับข้อมูลที่มีความสมดุล

ในวิทยานิพนธ์ฉบับนี้ เสนอตัวปรับแบบยึดเกาะ (Adhesive Modifier) เป็นเอนโทรปีแบบใหม่ โดยใช้แซนนอนเอนโทรปี [2] เป็นพื้นฐานในสร้างต้นไม้ตัดสินใจ จุดประสงค์เพื่อเพิ่มประสิทธิภาพการจำแนกข้อมูลไม่สมดุลในกลุ่มที่มีน้อยให้ดีขึ้น โดยใช้ตัวปรับที่สามารถปรับตัวมันเองให้สัมพันธ์กับความน่าจะเป็นทั้งสองกลุ่มในแต่ละระดับของการสร้างต้นไม้ตัดสินใจ ได้แก้ไขความลำเอียงในวิธี C4.5 และดีกว่าวิธีเอนโทรปีแบบอสมมาตร และ เอนโทรปีออกจากศูนย์กลาง เพราะทั้งสองวิธีนี้กำหนดพารามิเตอร์เพียงครั้งเดียวที่อาจจะเกิดความไม่แน่นอนของการกระจายความน่าจะเป็นในการเลือกโหนดระดับถัดไปได้ โดยพิจารณาจากค่าเอฟ และเปรียบเทียบระดับความมีนัยสำคัญทางสถิติที่ระดับ 0.1 จากการทดลองข้อมูลทั้ง 16 ชุด พบว่าวิธีการที่เสนอนี้สามารถสร้างกฎตัวอย่างในกลุ่มที่มีน้อยได้ดี และให้ค่าเอฟค่อนข้างสูงกว่าวิธี C4.5, AE และ OCE

5.2. ข้อจำกัด

แม้ว่าวิทยานิพนธ์ฉบับนี้จะแสดงให้เห็นถึงความสามารถในการจำแนกข้อมูลไม่สมดุลได้ แต่วิธีที่เสนอในวิทยานิพนธ์ฉบับนี้ยังมีข้อจำกัดดังต่อไปนี้

ประการแรกคือ หากมีจำนวนคุณลักษณะของข้อมูลน้อยเกินไปประสิทธิภาพการทำงานจะลดลงและหากมีจำนวนมากต้นไม้การตัดสินใจที่ได้จะมีขนาดใหญ่และมีจำนวนกฎจำนวนมาก

ประการที่สอง คือ ไม่สามารถจำแนกข้อมูลที่มีมากกว่าสองกลุ่ม เนื่องจากวิธีที่เสนอถูกออกแบบมาสำหรับข้อมูลแบบสองกลุ่มเท่านั้น

5.3. ข้อเสนอแนะ

1. การจำแนกข้อมูลไม่สมดุล สามารถนำไปประยุกต์ใช้กับการจำแนกข้อมูลในชีวิตประจำวันได้ ซึ่งส่วนใหญ่แล้วจะเป็นข้อมูลทางด้านการแพทย์ เพื่อการวินิจฉัยโรค
2. งานวิจัยนี้เสนอวิธีการจำแนกข้อมูลแบบสองกลุ่ม ดังนั้นหากมีการเพิ่มการจำแนกข้อมูลแบบหลายกลุ่มได้ จะทำให้นำไปใช้ประโยชน์กับข้อมูลจริงได้มากขึ้น
3. หากเพิ่มวิธีการตัดเล็มต้นไม้ตัดสินใจเข้าไปในระหว่างการเรียนรู้ อาจทำให้โครงสร้างต้นไม้ตัดสินใจมีความแม่นยำมากขึ้น

รายการอ้างอิง

[1] Quinlan, J. R. Induction of decision trees, Machine Learning. (1986): 81–106.

[2] Shannon, C.E. A Mathematical Theory of Communication, Bell System Technical Journal, 27, July & October, (1948).

[3] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann. San Mateo. (1993).

[4] Chawla, N. V., Japkowicz, N., และ Kolcz, A. Special Issue on Class Imbalances. SIGKDD Explorations. (2004).

[5] Han, J., และ Kamber, M. Data Mining: Concepts and Techniques. San Francisco, CA, (2001)

[6] บุญเสริม กิจศิริกุล. ปัญญาประดิษฐ์. (2548): 169-185.

[7] Hart, P. E. The Condensed Nearest Neighbor Rule. IEEE Transactions on Information Theory. (1968): 515–516.

[8] Tomek, I. Two Modifications of CNN. IEEE Transactions on Systems Man and Communications. (1976): 769–772.

[9] Kubat, M., และ Matwin, S. Addressing the Course of Imbalanced Training Sets: One-sided Selection. ICML. (1997): 179–186.

[10] Batista, G.E.A.P.A., Prati, R.C., และ Monard, M.C. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. ACM SIGKDD Explorations Newsletter, vol. 6, no. 1. (2004): 20-29.

[11] Laurikkala, J. Improving Identification of Difficult Small Classes by Balancing Class Distribution. Tech. University of Tampere. (2001): A-2001-2.

[12] Chawla, N. V., Bowyer, K. W., Hall, L. O., และ Kegelmeyer, W. P.. SMOTE: Synthetic Minority Over-sampling Technique. JAIR. (2002).

[13] Han, H., Wang, W., และ Mao, B. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Huang, D.-S., Zhang, X.-P., Huang, G.-B. (eds.) ICIC. (2005). LNCS, vol. 3644, pp. 878–887, Springer Heidelberg. (2005).

[14] Wilson, D. L. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. IEEE Transactions on Systems, Man, and Communications 2, 3 (1972): 408–421.

[15] Chawla, N. V., Lazarevic, A., Hall L. O., และ Bowyer, K. W. SMOTEBoost: Improving Prediction of the Minority Class in Boosting. PKDD. (2003).

[16] Chen, C., Liaw, A., และ Breiman, L. Using random forest to learn imbalanced data Technical report. Department of Statistics, University of California, Berkeley, (2004).

[17] Breiman, L. Random forests. Machine Learning. (2001): 5–32.

[18] Breiman, L. Friedman, J. H., Olshen, R. A., and Stone, C. J. Classification and Regression Trees. Wadsworth International. (1984).

[19] Marcellin, S. D., Zighed, A. และ Ritschard, G. An asymmetric entropy measure for decision trees. IPMU. (2006):1292–1299.

[20] Lallich, S., Lenca, P. และ Vaillant, B. Construction of an off-centered entropy for supervised learning. ASMDA. (2007): 8.

[21] Cielak, D. A. และ Chawla. N. V. Learning decision trees for unbalanced data. ECML/PKDD. (2008).

[22] C. L. Blake and C. J. Merz, UCI repository of machine learning databases, Department of Information and Computer Science, [Online], 1998, Available from: http://www.ics.uci.edu/~mlearn/MLRepository.html

[23] Weka 3, Data Minig with Open Source Machine Learning Software, [Online], 1999. Available from: http://www.cs.waikato.ac.nz/~ml/.

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวอุไรรัตน์ กฤษดาวาณิชย์ เกิดเมื่อวันที่ 22 ธันวาคม พ.ศ. 2528 ที่จังหวัดสุโขทัย สำเร็จการศึกษาปริญญาวิทยาศาสตรบัณฑิต สาขาวิชาคณิตศาสตร์ ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร ในปีการศึกษา 2550 และเข้าศึกษาในหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2551