การจำแนกแบบหลายฉลากโดยใช้การเรียนรู้เชิงรุกบนชุดข้อมูลขนาดใหญ่และไม่สมดุล

นายไพโรจน์ ตันติวชิรฐากูร

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2559

Multi-Label Classification Using Active learning on Large Scale and Imbalanced Data Sets

Mr. Phairod Tuntiwachiratrakun

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 2016

หัวข้อวิทยานิพนธ์	การจำแนกแบบหลายฉลากโดยใช้การเรียนรู้เชิงรุกบนชุดข้อมูลขนาดใหญ่และไม่สมดุล
โดย	นายไพโรจน์ ตันติวชิรฐากูร
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	อาจารย์ ดร.พีรพล เวทีกูล

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

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

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(อาจารย์ ดร.พีรพล เวทีกูล)

.....กรรมการ
(รองศาสตราจารย์ ดร.โชติรัตน์ รัตนามหัทธนะ)

.....กรรมการภายนอกมหาวิทยาลัย
(รองศาสตราจารย์ ดร.กฤษณะ ไวยมัย)

ไพโรจน์ ตันติวชิรฐากูร : การจำแนกแบบหลายฉลากโดยใช้การเรียนรู้เชิงรุกบนชุดข้อมูลขนาดใหญ่และไม่สมดุล (Multi-Label Classification Using Active learning on Large Scale and Imbalanced Data Sets) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: อ. ดร.พีรพล เวทีกูล, 56 หน้า.

ปัจจุบันข้อมูลมีความซับซ้อนมากขึ้น ข้อมูลหนึ่งตัวอย่างสามารถเป็นได้มากกว่าหนึ่งคลาส ซึ่งข้อมูลมักจะมีขนาดใหญ่และเพิ่มขึ้นมหาศาล ดังนั้นในงานวิจัยชิ้นนี้จึงนำเสนอการเรียนรู้แบบเชิงรุก เป็นการเรียนรู้ที่ค่อยๆเรียนรู้จากข้อมูลขนาดใหญ่ โดยเริ่มต้นเรียนรู้จากกลุ่มตัวแทนข้อมูลที่ถูกเลือก แล้วเรียนรู้เพิ่มเติมจากข้อมูลที่จำแนกผิดพลาดในอดีตโดยแบบจำลอง จึงทำให้สามารถแก้ไขปัญหาการจำแนกข้อมูลขนาดใหญ่ได้ นอกจากนั้นข้อมูลแบบหลายฉลากมักมีปัญหาเรื่องความไม่สมดุลของข้อมูล จึงได้นำเสนอการแก้ไขปัญหาไม่สมดุลของข้อมูลด้วยเทคนิคที่เหมาะสมของแต่ละตัวจำแนก ซึ่งในการทดลองจะใช้ตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีน และนิวรอลเน็ตเวิร์กสำหรับการสร้างแบบจำลองเริ่มต้น และการเรียนรู้เพิ่มเติม

จากผลการทดลองเปรียบเทียบทั้งสองเทคนิค การเรียนรู้เชิงรุกประสิทธิภาพการจำแนกจะสูงกว่าการเรียนรู้เชิงรับเมื่อสร้างแบบจำลองด้วยตัวจำแนกนิวรอลเน็ตเวิร์กบนข้อมูลสองคลาสและข้อมูลแบบหลายฉลาก อย่างไรก็ตามเมื่อเทียบประสิทธิภาพการจำแนกกับการเรียนรู้เชิงรับที่สร้างแบบจำลองด้วยตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีนจะมีค่าที่ต่ำกว่ามากบนชุดข้อมูลหลายฉลาก จึงเป็นที่มาการนำเสนอการเรียนรู้เชิงรุกด้วยตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีน

จากผลการทดลองเมื่อเปรียบเทียบการเรียนรู้เชิงรุกและการเรียนรู้เชิงรับเมื่อสร้างแบบจำลองด้วยซัพพอร์ตเวกเตอร์แมชชีน บนข้อมูลแบบหลายฉลากขนาดใหญ่ แสดงให้เห็นว่าทั้งสองวิธีการมีค่าประสิทธิภาพจากการวัดค่าเฉลี่ยไมโคร ค่าเฉลี่ยแมโคร ที่ใกล้เคียงกันและการเรียนรู้เชิงรุกจะใช้ขนาดข้อมูลที่น้อยกว่าการเรียนรู้เชิงรับในการเรียนรู้ ซึ่งจะเลือกใช้กลยุทธ์การเลือกข้อมูลที่เหมาะสมกับการเรียนรู้เพิ่มเติม วิทยานิพนธ์ฉบับนี้จึงนำเสนอกลยุทธ์การเรียนรู้เชิงรุก ได้แก่ การเรียนรู้เชิงรุกไม่เอนเอียง (UAL) และการเรียนรู้เชิงรุกด้วยเอสวีเอ็มเลือกข้อมูลไม่มั่นใจบนซัพพอร์ตเวกเตอร์แบบใช้ข้อมูลซ้ำ (AL-SVM-SV-R)

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

ปีการศึกษา 2559

5770949121 : MAJOR COMPUTER SCIENCE
KEYWORDS: ACTIVE LEARNING / COST SENSITIVE / IMBALANCED ISSUE / LARGE SCALE DATA / MULTI-LABEL / NEURAL NETWORK / SUPPORT VECTOR MACHINE

PHAIROD TUNTIWACHIRATRAKUN: Multi-Label Classification Using Active learning on Large Scale and Imbalanced Data Sets. ADVISOR: DR.PEERAPON VATEEKUL, 56 pp.

Nowadays, data are getting more complicated, where an instance in a dataset can represent multiple classes. The volume of data is usually large and getting larger. This thesis proposes Active learning, and this method gradually learns from the whole data by initially learning from the selected sample data, and it incrementally learns from misclassified examples. Hence, it can solve the large data volume issue. Moreover, multi-label data usually has the imbalanced data issue, therefore, this thesis uses appropriate technique for it. The experiments described herein used the Support Vector Machine (SVM) and Neural Network classifiers for the construction of an initial and incremental model.

The experiments compared the results of both techniques in binary and multi-label data. The performance of Active learning was better than Passive learning, when both techniques were constructed by Neural Network. However, its performance was lower than the performance of Passive learning constructed by SVM classifier in multi-label data. Accordingly, Active learning with SVM was proposed instead.

The experiments show the comparison of both techniques with SVM classifier on the multi-label large data. Both techniques have obtained similar result measured by micro-average and macro-average. Moreover, Active learning uses sizing of data less than Passive learning for learning and selecting strategies which are suitable with the incremental learning. In this thesis two strategies are proposed: UAL and AL-SVM-SV-R.

Department:	Computer Engineering	Student's Signature	.....
Field of Study:	Computer Science	Advisor's Signature	.....
Academic Year:	2016		

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้อย่างสมบูรณ์ด้วยดีนั้น ต้องขอขอบคุณบุคคลเหล่านี้ที่ให้ความช่วยเหลือ เป็นที่ปรึกษา คอยให้กำลังใจ ตลอดชี้แนะให้แก่ผู้วิจัยจนเพียรพยายามจนงานวิจัยสำเร็จตามที่ตั้งใจไว้

ขอขอบคุณอาจารย์ที่ปรึกษา ดร. พีรพล เวทีกูล ผู้ที่ให้ความช่วยเหลืออย่างดีเยี่ยม ในการให้คำปรึกษา ทุ่มเทแรงกายและแรงใจให้กับตัวผู้วิจัยและเพื่อนๆในห้องปฏิบัติการณ์ อย่างไม่รู้จักเหน็ดเหนื่อย จนเป็นแรงบันดาลใจให้ตัวผู้วิจัยในการทำงานวิจัยจนสำเร็จลุล่วง

ขอขอบคุณ ศ. ดร. บุญเสริม กิจศิริกุล ผู้ให้คำแนะนำชี้แนะในการทำงานวิจัย รวมถึงการเป็นประธานในการสอบวิทยานิพนธ์

ขอขอบคุณกรรมการการสอบวิทยานิพนธ์ รศ. ดร. โชติรัตน์ รัตนามหัทธนะ และ รศ. ดร. กฤษณะ ไวยมัย ที่ให้คำแนะนำที่เป็นประโยชน์ปรับปรุงในการทำงาน

ขอขอบคุณอาจารย์ทุกท่านในหลักสูตร ที่ให้ความรู้พื้นฐาน และแนวทางในการนำความรู้ที่ได้ไปใช้ให้เกิดประโยชน์ในงานวิจัย จนทำให้ผู้วิจัยสามารถนำความรู้ที่ได้ไปประยุกต์ใช้ได้ในงานวิจัยอย่างดีเยี่ยม

ขอขอบคุณเพื่อนๆ ในห้องปฏิบัติการที่คอยให้ความช่วยเหลือในด้านต่างๆ และเป็นกำลังใจด้วยดีเสมอมาตลอดระยะเวลาที่ทำงานด้วยกัน

สุดท้ายขอขอบคุณคุณพ่อ คุณแม่ และครอบครัวที่ให้การสนับสนุนด้วยดีในทุกๆด้านตลอดระยะเวลาที่ผ่านมาจนถึงปัจจุบัน

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ง

บทคัดย่อภาษาอังกฤษ ..... จ

กิตติกรรมประกาศ ..... ฉ

สารบัญ ..... ช

สารบัญภาพ ..... ฎ

สารบัญตาราง ..... ฐ

บทที่ 1 บทนำ 1

1.1 ที่มาและความสำคัญของปัญหา 1

1.2 วัตถุประสงค์ 2

1.3 ขอบเขตการดำเนินงาน 2

1.4 ประโยชน์ที่คาดว่าจะได้รับ 3

1.5 วิธีดำเนินการวิจัย 3

1.6 ผลงานตีพิมพ์จากงานวิจัย 3

บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง 5

2.1 การเรียนรู้เชิงรุก (Active Learning) 5

2.1.1 การเลือกตัวอย่าง Stream-based (Stream-based Selective Sampling) 6

2.1.2 การเลือกตัวอย่าง Pool-based (Pool-based Selective Sampling) 7

2.1.3 การเลือกตัวอย่างข้อมูลที่มีความไม่มั่นใจ (Uncertainty Sampling) 8

2.2 การเรียนรู้เชิงรับ (Passive Learning) 8

2.3 การจำแนกสองคลาส (Binary Classification) 8

2.4 การจำแนกหลายคลาสและแบบหลายฉลาก (Multiclass and Multi-Label) 8

2.4.1 One-Versus-One 9

หน้า

2.4.2 One-Versus-All 9
2.5 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Manchine: SVM) 9
2.6 สโตแคสติกเกรเดียนเดสเซนท์ (Stochastic Gradient Descent: SGD) 11
2.7 นิวรอลเน็ตเวิร์ก (Neural Network) 12
2.8 การวัดความคล้ายคลึงเชิงมุมโคไซน์ (Cosine Similarity) 14
2.9 กลยุทธ์สำหรับข้อมูลไม่สมดุล 14
2.9.1 สุ่มเลือกตัวอย่างลด (Random Undersampling) 16
2.10 การวัดประสิทธิภาพการทำงาน (Performance Evaluation) 16
2.10.1 ตัววัดประสิทธิภาพการจำแนกข้อมูลสองคลาส (Binary Classification Performance Measurement) 16
2.10.2 ตัววัดประสิทธิภาพการจำแนกข้อมูลสองคลาสสำหรับข้อมูลไม่สมดุล (Binary Classification Performance Measurement with Imbalanced data) 17
2.10.3 ตัววัดประสิทธิภาพการจำแนกประเภทแบบหลายฉลาก (Multi-Label Classification Performance Measurement 17
2.11 งานวิจัยที่เกี่ยวข้อง (Related Work) 18
บทที่3 การใช้การเรียนรู้เชิงรุกกับชุดข้อมูลขนาดใหญ่และไม่สมดุล 20
3.1 รูปแบบที่ 1 เสนอวิธีการจำแนกข้อมูลขนาดใหญ่และไม่สมดุลด้วยนิวรอลเน็ตเวิร์ก 21
3.1.1 การจำแนกข้อมูลขนาดใหญ่และไม่สมดุลสำหรับข้อมูลสองคลาสด้วยนิวรอลเน็ตเวิร์ก 21
3.1.2 การจำแนกข้อมูลขนาดใหญ่และไม่สมดุลสำหรับข้อมูลแบบหลายฉลากด้วยนิวรอลเน็ตเวิร์ก 22
3.2 รูปแบบที่ 2 เสนอวิธีการจำแนกข้อมูลขนาดใหญ่และไม่สมดุลแบบหลายฉลากด้วยซัพพอร์ตเวกเตอร์แมชชีน 25
3.2.1 การเรียนรู้เชิงรุกเอสวีเอ็มเลือกข้อมูลไม่มั่นใจบนซัพพอร์ตเวกเตอร์แบบใช้ข้อมูลซ้ำ (AL-SVM-SV-R) 25

หน้า

3.2.2 การเรียนรู้เชิงรุกเอสวีเอ็มเลือกข้อมูลไม่มั่นใจบนซัพพอร์ตเวกเตอร์แบบไม่ใช้ข้อมูลซ้ำ (AL-SVM-SV-N) 27
บทที่ 4 การทดลองและผลการทดลอง 30
4.1 ระบบที่ใช้ในการทดลอง 30
4.2 ข้อมูลที่ใช้ในการทดลอง 30
4.2.1 ชุดข้อมูลแบบสองคลาส (Binary Class Data Sets) 30
4.2.2 ชุดข้อมูลแบบหลายฉลาก (Multi-Label Data Sets) 31
4.3 การทดลองเปรียบเทียบการจำแนกของการเรียนรู้เชิงรุกกับการเรียนรู้เชิงรับด้วยตัวจำแนกนิวรอลเน็ตเวิร์ก 31
4.3.1 การเปรียบเทียบเปอร์เซ็นต์ของจำนวนข้อมูลแบบสองคลาสในการสร้างแบบจำลองเริ่มต้น 32
4.3.2 การเปรียบเทียบประสิทธิภาพความแม่นยำบนข้อมูลแบบสองคลาสระหว่างการเรียนรู้เชิงรุกและการเรียนรู้เชิงรับ 33
4.3.3 สรุปผลการทดลองการเปรียบเทียบการเรียนรู้เชิงรุกกับการเรียนรู้เชิงรับบนข้อมูลสองคลาสขนาดใหญ่และไม่สมดุล 34
4.3.4 การทดลองการจำแนกข้อมูลขนาดใหญ่และไม่สมดุลหลายฉลากด้วยนิวรอลเน็ตเวิร์ก 35
4.4 การทดลองเปรียบเทียบการจำแนกของการเรียนรู้เชิงรุกกับการเรียนรู้เชิงรับด้วยตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีน 36
4.4.1 การทดลองเปรียบเทียบประสิทธิภาพการเรียนรู้เชิงรุกกับการเรียนรู้เชิงรับด้วยตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีนเบื้องต้นบนชุดข้อมูล NUS-WIDE 37
4.4.2 การเปรียบเทียบประสิทธิภาพของการเรียนรู้เชิงรุกทั้ง 2 แบบ AL-SVM-SV-R และ AL-SVM-SV-N 40
4.4.3 การวัดประสิทธิภาพของการเรียนรู้เชิงรุกด้วย 3 Fold Cross-Validation บนชุดข้อมูล NUS-WIDE 43

หน้า

4.4.4 การวัดประสิทธิภาพการเรียนรู้เชิงรุกด้วย 3 Fold Cross-Validation บนชุดข้อมูล RCV1V2 44
บทที่ 5 สรุปการวิจัยและแนวทางการวิจัยในขั้นถัดไป 45
5.1 สรุปการวิจัย 45
5.2 แนวทางการวิจัยในขั้นถัดไป 46
รายการอ้างอิง 47
ภาคผนวก ก รายละเอียดของชุดข้อมูลแบบหลายฉลาก NUS-WIDE 50
ภาคผนวก ข รายละเอียดของชุดข้อมูลแบบหลายฉลาก RCV1V2 53
ประวัติผู้เขียนวิทยานิพนธ์ 56

สารบัญภาพ

หน้า

รูปที่ 1.1 รายละเอียดขั้นตอนการดำเนินงานเรื่องการทำวิจัย....4
รูปที่ 2.1 การเปรียบเทียบขั้นตอนการทำงานระหว่างการเรียนรู้เชิงรุกและการเรียนรู้เชิงรับ....6
รูปที่ 2.2 การเลือกตัวอย่าง Stream-based....7
รูปที่ 2.3 แสดงการเลือกตัวอย่าง Pool-based....7
รูปที่ 2.4 แสดงตัวอย่างการทำ One-Versus-All ของคลาส C1, C2 และ C3....9
รูปที่ 2.5 การหา Decision Boundary ที่เหมาะสมของ ซัพพอร์ตเวกเตอร์แมชชีน....10
รูปที่ 2.6 แสดงแบบจำลองเชิงเส้นของตัวจำแนก SVM ระหว่าง 2 คลาส (-1 และ +1)....10
รูปที่ 2.7 โครงสร้างทั่วไปของนิวรอลเน็ตเวิร์กประกอบไปด้วย 3 ชั้น คือ ชั้นนำเข้า ชั้นซ่อน และชั้นนำออก....13
รูปที่ 2.8 แสดงการเพิ่มข้อมูลให้กับคลาสฝั่งข้างน้อยโดยใช้เทคนิคการเลือกตัวอย่างเพิ่ม (Oversampling) ให้เท่ากับข้อมูลคลาสฝั่งข้างมาก....15
รูปที่ 2.9 แสดงการลดจำนวนข้อมูลของคลาสฝ่ายข้างมากโดยใช้เทคนิคการเลือกตัวอย่างลด (Undersampling) ให้เท่ากับข้อมูลคลาสฝั่งข้างน้อย....15
รูปที่ 2.10 แสดงการเพิ่มน้ำหนักให้กับคลาสฝั่งข้างน้อยโดยใช้เทคนิค Cost Sensitive....16
รูปที่ 3.1 แสดงขั้นตอนวิธีการการเรียนรู้เชิงรุกสำหรับข้อมูลขนาดใหญ่และไม่สมดุล....21
รูปที่ 3.2 รหัสเทียมของกลยุทธ์การเรียนรู้เชิงรุกไม่เอนเอียงแบบมีสัดส่วน....23
รูปที่ 3.3 ขั้นตอนการจำแนกข้อมูลขนาดใหญ่และไม่สมดุลหลายฉลากด้วยนิวรอลเน็ตเวิร์ก....24
รูปที่ 3.4 ตัวอย่างการเรียนรู้เชิงรุกเอสวีเอ็มเลือกข้อมูลไม่มั่นใจบนซัพพอร์ตเวกเตอร์แบบใช้ข้อมูลซ้ำ....26
รูปที่ 3.5 การเลือกข้อมูลที่ไม่มั่นใจบนเส้นซัพพอร์ตเวกเตอร์ข้างลบถึงเส้นซัพพอร์ตเวกเตอร์ข้างบวก....26
รูปที่ 3.6 ตัวอย่างการเรียนรู้เชิงรุกเอสวีเอ็มเลือกข้อมูลไม่มั่นใจบนซัพพอร์ตเวกเตอร์แบบไม่ใช้ข้อมูลซ้ำ....28
รูปที่ 3.7 การเรียนรู้เชิงรุกเอสวีเอ็มเลือกข้อมูลไม่มั่นใจบนซัพพอร์ตเวกเตอร์แบบไม่ใช้ข้อมูลซ้ำ (AL-SVM-SV-N) โดยที่มีจุดเชื่อมไปยังรูปที่ 3.8....28
รูปที่ 3.8 การเรียนรู้เชิงรุกเอสวีเอ็มเลือกข้อมูลไม่มั่นใจบนซัพพอร์ตเวกเตอร์แบบใช้ข้อมูลซ้ำ (AL-SVM-SV-R)....29

รูปที่ 4.1 การทำคอสท์เซนซิทีฟ (Cost Sensitive) ของ PL-SVM เพื่อให้แบบจำลองสามารถทำนายคลาสที่มีข้อมูลปริมาณน้อยได้หรือแก้ไขปัญหาไม่สมดุลของคลาส (Label0-Label40).....39
รูปที่ 4.2 การทำคอสท์เซนซิทีฟ (Cost Sensitive) ของ PL-SVM เพื่อให้แบบจำลองสามารถทำนายคลาสที่มีข้อมูลปริมาณน้อยได้หรือแก้ไขปัญหาไม่สมดุลของคลาส (Label41-Label80).....39
รูปที่ 4.3 แสดงการเปรียบเทียบค่าประสิทธิภาพ F1 แต่ละ Classifier ของการทำ One-Versus-All ของการเรียนรู้เชิงรุก 2 รูปแบบ (Label0-Label80).....41
รูปที่ 4.4 แสดงการเปรียบเทียบจำนวนรอบของการเรียนรู้เพิ่มเติมของการเรียนรู้เชิงรุกทั้ง 2 รูปแบบ (Label0-Label60).....42
รูปที่ 4.5 แสดงการเปรียบเทียบจำนวนรอบของการเรียนรู้เพิ่มเติมของการเรียนรู้เชิงรุกทั้ง 2 รูปแบบ (Label61-Label80).....43
รูปที่ ก.1 แสดงจำนวนข้อมูลของแต่ละคลาสของชุดข้อมูล NUS-WIDE.....49
รูปที่ ข.1 แสดงจำนวนข้อมูลของแต่ละคลาสของชุดข้อมูล RCV1V2.....52

สารบัญตาราง

หน้า

ตารางที่ 2.1 แสดงความสัมพันธ์ Predicted คลาส และ True คลาส.....16
ตารางที่ 4.1 ลักษณะเฉพาะของชุดข้อมูลแบบสองคลาสที่ใช้ในการทดลอง.....30
ตารางที่ 4.2 ลักษณะเฉพาะของชุดข้อมูลแบบหลายฉลากที่ใช้ในการทดลอง.....31
ตารางที่ 4.3 แสดงการเปรียบเทียบเปอร์เซ็นต์ของจำนวนข้อมูลสำหรับการสร้างแบบจำลองเริ่มต้น.....32
ตารางที่ 4.4 แสดงการเปรียบเทียบประสิทธิภาพความแม่นยำระหว่างการเรียนรู้เชิงรุกและการเรียนรู้เชิงรับ โดยการใช้ตัวจำแนก ANN และการวัด G-Mean แบบ 5 Fold Cross-Validation.....33
ตารางที่ 4.5 แสดงการเปรียบเทียบประสิทธิภาพความแม่นยำระหว่างการเรียนรู้เชิงรุกด้วยตัวจำแนก ANN และการเรียนรู้เชิงรับด้วยตัวจำแนก NB kNN SVM และ DT และการวัด G-Mean แบบ 5 Fold Cross-Validation.....34
ตารางที่ 4.6 แสดงค่าพารามิเตอร์ของนิวรอลเน็ตเวิร์กสำหรับการเรียนรู้เชิงรับและการเรียนรู้เชิงรุกบนชุดข้อมูล NUS-WIDE.....35
ตารางที่ 4.7 แสดงผลการทดลองเปรียบเทียบประสิทธิภาพการจำแนกชุดข้อมูล NUS-WIDE ระหว่างการเรียนรู้เชิงรับด้วยนิวรอลเน็ตเวิร์ก (PL-ANN) การเรียนรู้เชิงรุกด้วยนิวรอลเน็ตเวิร์ก (AL-ANN) และการเรียนรู้เชิงรับด้วยซัพพอร์ตเวกเตอร์แมชชีน และมีการทำ 3 Fold Cross-Validation วัดด้วยค่าเฉลี่ยไมโครและค่าเฉลี่ยแมโคร.....35
ตารางที่ 4.8 การแบ่งข้อมูลเพื่อวัดประสิทธิภาพแต่ละเทคนิคในเบื้องต้นบนชุดข้อมูล NUS-WIDE และ RCV1V2.....36
ตารางที่ 4.9 แสดงค่าพารามิเตอร์ที่สำคัญของการใช้สโตแคสติกเกรเดียนเดสเซนท์แบบจำลองเส้นตรงในการเรียนรู้เริ่มต้นและเรียนรู้เพิ่มเติมของการเรียนรู้เชิงรุกบนชุดข้อมูล NUS-WIDE และ RCV1V2.....37
ตารางที่ 4.10 แสดงผลการทดลองเปรียบเทียบประสิทธิภาพเบื้องต้นการจำแนกชุดข้อมูล NUS-WIDE ระหว่างการเรียนรู้เชิงรับและการเรียนรู้เชิงรุกด้วยซัพพอร์ตเวกแมชชีน และเบื้องต้นแบ่งข้อมูลเป็นชุดฝึกสอน ชุดตรวจสอบและชุดทดสอบ โดยที่วัดด้วยค่าเฉลี่ยไมโครและค่าเฉลี่ยแมโคร.....38

ตารางที่ 4.11 แสดงผลการทดลองเปรียบเทียบประสิทธิภาพการจำแนกชุดข้อมูล NUS-WIDE โดยเปรียบเทียบการเรียนรู้ของแบบจำลองบนชุดตรวจสอบของการเรียนรู้ทั้งสองแบบ.....40
ตารางที่ 4.12 แสดงผลการทดลองเปรียบเทียบประสิทธิภาพการจำแนกชุดข้อมูล NUS-WIDE โดยเปรียบเทียบค่า F1 บนชุดทดสอบของทุกคลาสของการเรียนรู้เชิงรุกทั้ง 2 แบบ.....41
ตารางที่ 4.13 แสดงผลการทดลองเปรียบเทียบประสิทธิภาพการจำแนกชุดข้อมูล NUS-WIDE ระหว่างการเรียนรู้เชิงรุกและการเรียนรู้เชิงรับด้วยซัพพอร์ตเวกเตอร์แมชชีน และมีการทำ 3 Fold Cross-Validation วัดด้วยค่าเฉลี่ยไมโครและค่าเฉลี่ยแมโคร.....44
ตารางที่ 4.14 แสดงผลการทดลองเปรียบเทียบประสิทธิภาพการจำแนกชุดข้อมูล RCV1V2 ระหว่างการเรียนรู้เชิงรุกและการเรียนรู้เชิงรับด้วยซัพพอร์ตเวกเตอร์แมชชีน และมีการทำ 3 Fold Cross-Validation วัดด้วยค่าเฉลี่ยไมโครและค่าเฉลี่ยแมโคร.....44
ตารางที่ ก.1 แสดงค่าทางสถิติของจำนวนข้อมูลแต่ละคลาสของข้อมูล NUS-WIDE.....50
ตารางที่ ก.2 แสดงรายละเอียดจำนวนข้อมูลแต่ละคลาสของชุดข้อมูล NUS-WIDE.....50
ตารางที่ ข.1 แสดงค่าทางสถิติของจำนวนข้อมูลแต่ละคลาสของข้อมูล RCV1V2.....53
ตารางที่ ข.2 แสดงจำนวนข้อมูลแต่ละคลาสของชุดข้อมูล RCV1V2.....53

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ปัจจุบันการจำแนกข้อมูลขนาดใหญ่ (Large Scale data Classification) ยังคงเป็นปัญหาในเรื่องของการบรรจุข้อมูลขนาดใหญ่เข้าหน่วยความจำในครั้งเดียวเพื่อประมวลผล และเรื่องของเวลาที่ใช้ดำเนินการ ตัวอย่างข้อมูลขนาดใหญ่ เช่น ข้อมูลสื่อสังคมออนไลน์ ข้อมูลข่าว ข้อมูลภาพยนตร์ ข้อมูลเพลง การเรียนรู้เชิงรับ (Passive Learning) เป็นเทคนิคที่นิยมใช้และถูกใช้ในการสร้างแบบจำลองเพื่อการจำแนกข้อมูลขนาดใหญ่ การเรียนรู้ของแบบจำลองจะเรียนรู้เพียงครั้งเดียวต่อหนึ่งชุดฝึกสอน (Training Set) กรณีถ้ามีตัวอย่างข้อมูลใหม่เพิ่มเข้ามา การเรียนรู้ไม่สามารถที่จะเรียนรู้เพิ่มเติมได้ต้องทำการเรียนรู้ข้อมูลทั้งหมดซ้ำ ซึ่งตัวจำแนกข้อมูล (Classifier) ที่นิยมใช้ในงานวิจัย เช่น นิวรอลเน็ตเวิร์ก (Neural Network, NN) ซัพเพอร์ตเวกเตอร์แมชชีน (Support Vector Machine, SVM) การเรียนรู้เบย์แบบง่าย (Naive Bayes, NB)

การเรียนรู้เชิงรุก (Active Learning) เป็นการเรียนรู้ที่ใช้วิธีการเลือกตัวแทนของข้อมูลของแต่ละคลาสจากขนาดข้อมูลทั้งหมดเพื่อนำมาใช้สร้างแบบจำลองเริ่มต้น (Initial Model) ซึ่งขนาดข้อมูลที่ใช้ในการเรียนรู้จะมีขนาดเล็กลง นอกจากนั้นแบบจำลองที่สร้างด้วยวิธีการนี้สามารถเรียนรู้ตัวอย่างข้อมูลใหม่จากชุดฝึกสอนและแบบจำลองสามารถเรียนรู้เพิ่มเติม (Incremental Learning) เฉพาะข้อมูลใหม่โดยที่ไม่ต้องเรียนรู้ข้อมูลทั้งหมดใหม่ซ้ำเหมือนกับวิธีการเรียนรู้เชิงรับ ดังนั้นเทคนิคในการเลือกข้อมูลของการเรียนรู้นี้มีความสำคัญเนื่องจากข้อมูลที่ได้จากเลือกจะถูกนำไปใช้ในกระบวนการเรียนรู้เพิ่มเติมต่อไป และส่งผลต่อประสิทธิภาพความแม่นยำการจำแนกข้อมูลของแบบจำลองอีกด้วย ตัวอย่างเทคนิคการเลือกข้อมูล เช่น การเลือกตัวอย่างข้อมูลที่มีความไม่มั่นใจ (Uncertainty Sampling), Query-By-Committee, การเลือกตัวอย่าง Stream-based โดยข้อดีของการเรียนรู้เชิงรุก คือ ขนาดจำนวนข้อมูลที่ใช้จะมีขนาดลดลงและมีประสิทธิภาพความแม่นยำในการจำแนกข้อมูลที่ดี ส่วนจุดด้อยของการเรียนรู้นี้ คือ การใช้วิธีการเลือกข้อมูลให้เหมาะกับคุณลักษณะของชุดข้อมูลตัวอย่างที่จะดำเนินการ เพราะจะมีผลต่อประสิทธิภาพความแม่นยำของการจำแนกข้อมูลของแบบจำลอง และวิธีการดำเนินการของการเรียนรู้ที่ซับซ้อนกว่าการเรียนรู้เชิงรับมาก

ปัญหาความไม่สมดุลของข้อมูล (Imbalanced data) เป็นปัญหาที่เกี่ยวกับจำนวนของข้อมูลระหว่างคลาสที่มีความแตกต่างกันมาก โดยคลาสที่มีปริมาณข้อมูลมากกว่าจะเรียกว่า คลาสฝ่ายข้างมาก (majority class) ส่วนคลาสที่มีปริมาณข้อมูลน้อยกว่ามากจะเรียกว่า คลาสฝ่ายข้างน้อย (minority class) ตัวอย่างข้อมูลไม่สมดุล เช่น ข้อมูลเกี่ยวกับการทุจริต (fraud data) ข้อมูลเกี่ยวกับผู้ป่วยโรคมะเร็ง โดยการแก้ไขปัญหาการจำแนกข้อมูลไม่สมดุลโดยทั่วไปจะเป็นการทำให้ข้อมูลทุก

คลาสมีความสมดุลก่อนบรรจุข้อมูลเข้าหน่วยความจำเพื่อสร้างแบบจำลอง ซึ่งกลยุทธ์ที่นิยมในการทำให้ข้อมูลมีความสมดุลระหว่างคลาสมีกลยุทธ์ดังนี้ [1] Oversampling เป็นการเพิ่มขนาดจำนวนข้อมูลในคลาสฝ่ายข้างน้อย เพื่อให้มีขนาดสมดุลกับจำนวนของคลาสฝ่ายข้างมาก ดังนั้นขนาดข้อมูลหลังจากที่ผ่านกลยุทธ์นี้จะมีขนาดเพิ่มขึ้นจากขนาดเดิม ดังนั้นสำหรับข้อมูลขนาดใหญ่การใช้กลยุทธ์นี้จะทำให้การจำแนกข้อมูลเพื่อสร้างแบบจำลองจะใช้หน่วยความจำและระยะเวลาในการดำเนินการมากกว่าเดิม Undersampling เป็นการลดขนาดจำนวนข้อมูลในคลาสฝ่ายข้างมากเพื่อให้สมดุลกับจำนวนของคลาสฝ่ายข้างน้อย ดังนั้นปริมาณข้อมูลขนาดใหญ่หลังจากที่ผ่านกลยุทธ์นี้จะมีขนาดลดลง ทำให้การจำแนกข้อมูลเพื่อสร้างแบบจำลองจะใช้หน่วยความจำและระยะเวลาลดลงจากเดิม Cost Sensitive เป็นเทคนิคการเพิ่มความสำคัญหรือน้ำหนัก (Weight) ให้กับคลาสฝ่ายข้างน้อยให้มีความสำคัญมากขึ้นเพื่อให้สมดุลกับคลาสฝ่ายข้างมาก โดยที่ขนาดข้อมูลที่ใช้ในการเรียนรู้ยังคงมีปริมาณเท่าเดิม ทำให้แบบจำลองสามารถที่จะจำแนกคลาสฝ่ายข้างน้อยได้ดีขึ้น

งานวิจัยนี้จะมุ่งเน้นการจำแนกข้อมูลสองคลาส (Binary Classification) แบบหลายฉลาก (Multi-Label Classification) สำหรับข้อมูลขนาดใหญ่และไม่สมดุลด้วยการเรียนรู้เชิงรุก ที่มีประสิทธิภาพการจำแนกข้อมูลที่ใกล้เคียงหรือสูงกว่าการเรียนรู้เชิงรับ บนชุดข้อมูลที่มีขนาดข้อมูลตัวอย่างมากกว่า 100,000 ตัวอย่างข้อมูล

1.2 วัตถุประสงค์

นำเสนอการเรียนรู้เชิงรุก (Active Learning) สำหรับการจำแนกข้อมูลที่มีประสิทธิภาพการจำแนกใกล้เคียงหรือสูงกว่าการเรียนรู้เชิงรับ (Passive Learning) บนข้อมูลแบบสองคลาสและแบบหลายฉลาก ซึ่งจะใช้เทคนิค One-Versus-All ในการจำแนกข้อมูลแบบหลายฉลากขนาดใหญ่และไม่สมดุล โดยที่จะมีการทดสอบบนข้อมูลขนาดใหญ่ที่มีขนาดข้อมูลตัวอย่างมากกว่า 100,000 ตัวอย่างข้อมูลต่อหนึ่งชุดข้อมูล

1.3 ขอบเขตการดำเนินงาน

ในงานวิจัยนี้จะประกอบด้วย 2 ส่วน คือ ส่วนแรกจะเป็นการเพิ่มประสิทธิภาพของแบบจำลองในการจำแนกข้อมูลขนาดใหญ่ และไม่สมดุลสำหรับข้อมูลสองคลาส โดยข้อมูลที่ใช้จะมาจากเว็ปไซต์ KDD และ UCI โดยวิธีการที่ใช้จะใช้ข้อมูลเพียงบางส่วนในการสร้างแบบจำลองเพื่อการจำแนกข้อมูลและได้ประสิทธิภาพในการทำนายที่เหนือกว่าการเรียนรู้เชิงรับ

ในส่วนที่สอง จะเป็นการเพิ่มประสิทธิภาพของแบบจำลองในการจำแนกข้อมูลขนาดใหญ่และไม่สมดุลสำหรับข้อมูลหลายฉลาก (Multi-Label Data Sets) โดยข้อมูลที่ใช้จะมาจากเว็ปไซต์ Mulan โดยวิธีการจะประยุกต์มาจากส่วนแรกนำมาใช้กับข้อมูลแบบหลายฉลาก โดยประสิทธิภาพใน

การจำแนกใกล้เคียงหรือสูงกว่าการเรียนรู้เชิงรับ ซึ่งการวัดประสิทธิภาพการจำแนกข้อมูลทั้งสองส่วนจะใช้ตัววัดที่แตกต่างกัน อ้างอิงจากหัวข้อ 2.10.2 และหัวข้อ 2.10.3

1.4 ประโยชน์ที่คาดว่าจะได้รับ

ได้วิธีการจำแนกข้อมูลขนาดใหญ่และไม่สมดุลสำหรับข้อมูลแบบสองคลาสและข้อมูลแบบหลายฉลากที่มีประสิทธิภาพการจำแนกใกล้เคียงหรือสูงกว่าแบบการเรียนรู้เชิงรับในด้านการทำนายความแม่นยำของการจำแนกประเภทข้อมูลดังกล่าว

1.5 วิธีดำเนินการวิจัย

วิธีการดำเนินการวิจัยสามารถแบ่งออกได้เป็นขั้นตอนดังนี้ และสามารถดูรายละเอียดการทำงานเพิ่มเติมได้รูปที่ 1.1

1. ศึกษางานวิจัยที่เกี่ยวข้อง
2. ศึกษาข้อมูลตัวอย่างที่จะใช้ทดสอบ
3. กำหนดตัววัดประสิทธิภาพการทำงาน
4. กำหนดแนวทางและวิธีการ
5. ทดลองเพื่อให้เกิดวิธีการใหม่
6. วิเคราะห์ผลการทดลองและปรับปรุง
7. สรุปผลการทดลองที่ได้ทั้งหมด
8. สอบโครงร่างวิทยานิพนธ์
9. ตีพิมพ์ผลงานทางวิชาการ
10. สรุปผลและจัดทำวิทยานิพนธ์
11. สอบวิทยานิพนธ์

1.6 ผลงานตีพิมพ์จากงานวิจัย

ส่วนหนึ่งของการศึกษาเบื้องต้นของงานวิจัยชิ้นนี้ ได้รับการตีพิมพ์ดังรายละเอียดต่อไปนี้

- Applying active learning to strategy to classify large scale data with imbalanced classes โดย ไพโรจน์ ตันติวชิรฐากูร และ พีรพล เวทีกูล ในงานประชุมวิชาการ “The 5th International Conference on Control, Automation and Information Sciences (ICCAIS 2016)” ซึ่งจัดขึ้น ณ เมืองอันซัน ประเทศเกาหลีใต้ ระหว่างวันที่ 27 ถึง 29 ตุลาคม 2559

เดือน / ปี	2557					2558												2559												2560						
	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7
1. ศึกษางานวิจัยที่เกี่ยวข้อง																																				
2. ศึกษาข้อมูลตัวอย่างที่จะใช้ทดสอบ																																				
3. กำหนดตัวชี้วัดประสิทธิภาพการทำงาน																																				
4. กำหนดแนวทางและวิธีการ																																				
5. ทดลองเพื่อให้เกิดวิธีการใหม่																																				
6. วิเคราะห์ผลการทดลองและปรับปรุง																																				
7. สรุปผลการทดลองที่ได้ทั้งหมด																																				
8. สอบโครงร่างวิทยานิพนธ์																																				
9. ตีพิมพ์ผลงานทางวิชาการ																																				
10. สรุปผลและจัดทำวิทยานิพนธ์																																				
11. สอบวิทยานิพนธ์																																				

รูปที่ 1.1 รายละเอียดขั้นตอนการดำเนินงานเรื่องการทำวิจัย

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 การเรียนรู้เชิงรุก (Active Learning)

การเรียนรู้เชิงรุก [2, 3] คือ การเรียนรู้ที่สามารถเลือกขนาดข้อมูลเพียงบางส่วนจากขนาดข้อมูลทั้งหมดเพื่อจะนำมาเรียนรู้และใช้สร้างแบบจำลอง ทำให้ข้อมูลที่ใช้ในการสร้างแบบจำลองมีขนาดลดลง เทคนิคนี้จะลดขนาดจำนวนข้อมูลตัวอย่างที่ติดฉลาก (Label) เนื่องจากข้อมูลติดฉลากจะมีค่าใช้จ่ายสูงสำหรับการติดฉลากจากผู้เชี่ยวชาญของชนิดข้อมูลกลุ่มนั้น การเลือกข้อมูลจะถูกใช้เพื่อคัดเลือกตัวแทนของข้อมูลตัวอย่างบางส่วนจากขนาดข้อมูลทั้งหมดด้วยวิธีการที่เรียกว่า ขั้นตอนวิธีสอบถาม (query algorithm) [4, 5] ตัวอย่างการใช้วิธีการคัดเลือกข้อมูลด้วยวิธีการสอบถามจากผู้เรียนหรือแบบจำลอง เช่น การเลือกตัวอย่างข้อมูลที่มีความไม่มั่นใจโดยข้อมูลที่ได้กลับมาจากการสอบถามจะเป็นข้อมูลที่ผู้เรียนตอบผิดหรือทำนายข้อมูลผลเฉลยของตัวอย่างข้อมูลนั้นผิดไป โดยการเรียนรู้เชิงรุกจะมีการเรียนรู้ที่สำคัญประกอบไปด้วย คือ การเรียนรู้เบื้องต้นหรือการสร้างแบบจำลองเริ่มต้น (Initial Model) และการเรียนรู้ข้อมูลเพิ่มเติมหรือการเรียนรู้เพิ่มเติม (Incremental Learning) [6] เพราะฉะนั้นการเรียนรู้จึงสามารถที่จะเรียนรู้จากขนาดข้อมูลที่ถูกเลือกบางส่วนและค่อยเรียนรู้เพิ่มเติมจากชุดข้อมูลฝึกสอนทำให้การจำแนกข้อมูลของแบบจำลองที่ใช้เทคนิคมีประสิทธิภาพโดยที่ใช้ขนาดข้อมูลที่ลดลง ตัวจำแนกที่นิยมใช้ในการเรียนรู้เชิงรุก เช่น นิวรอลเน็ตเวิร์ก (Neural Network) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) นาอีฟเบย์ (Naïve Bayes) สามารถดูการเปรียบเทียบการทำงานระหว่างการเรียนรู้เชิงรุกและการเรียนรู้เชิงรับ ดังรูปที่ 2.1

การสร้างแบบจำลองสำหรับการเรียนรู้เบื้องต้น เป็นการเลือกจำนวนข้อมูลตัวอย่างเพื่อที่จะนำมาใช้สร้างแบบจำลองเริ่มต้น เช่น การเลือกตัวแทนข้อมูลตัวอย่างจากการสุ่มจำนวนข้อมูลตัวอย่างทั้งหมด ซึ่งวิธีการเลือกข้อมูลตัวอย่างและชนิดตัวจำแนกข้อมูลอาจจะมีผลต่อประสิทธิภาพความแม่นยำในการจำแนกข้อมูล ควรพิจารณาคุณลักษณะของชุดข้อมูลตัวอย่างประกอบกับการเลือกใช้วิธีการเหล่านั้นให้เหมาะสม

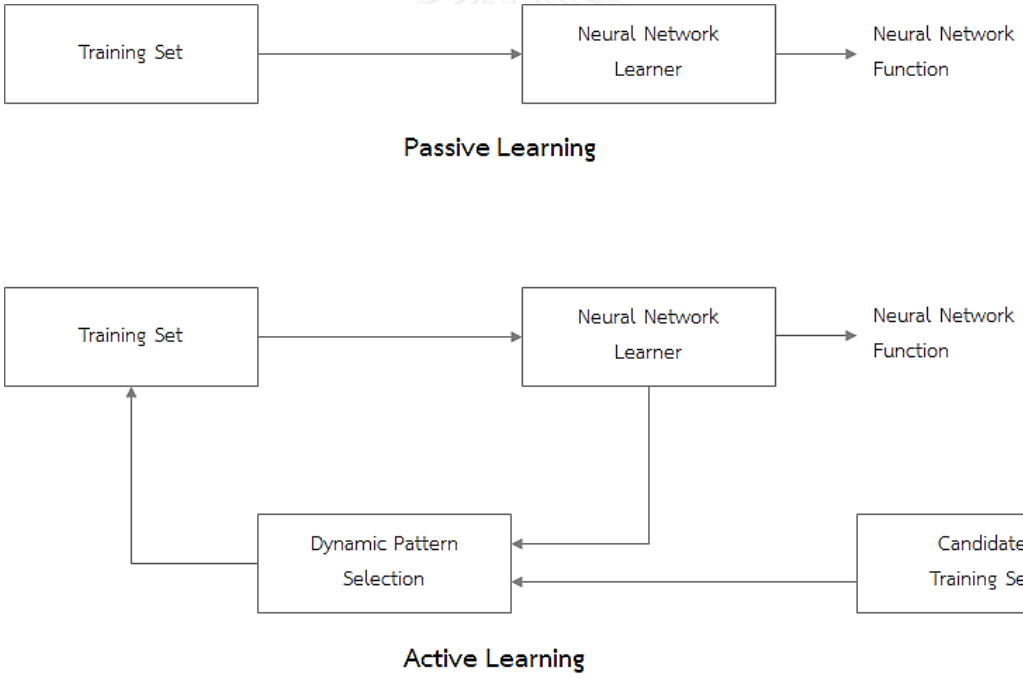
การเรียนรู้ข้อมูลเพิ่มเติม หลังจากที่ได้ทำการสร้างแบบจำลองเริ่มต้น จะมีการทำซ้ำของขั้นตอนวิธีการสอบถามเพื่อเลือกข้อมูลสำหรับการเรียนรู้เพิ่มเติม และขั้นตอนวิธีการฝึกอบรม เพื่อเลือกชุดข้อมูลตัวอย่างที่ยังไม่ได้ถูกเลือก และใช้วิธีการเลือกข้อมูลเพียงบางส่วนจากข้อมูลที่เหลือทั้งหมด ซึ่งข้อมูลตัวอย่างที่ถูกเลือกจะถูกนำมาอัพเดทความรู้เพิ่มเติมโดยตัวจำแนกข้อมูลลงในแบบจำลอง ซึ่งการเรียนรู้มีพารามิเตอร์ที่เกี่ยวข้องดังนี้

1) งบประมาณของการสอบถาม (budget of queries) คือ จำนวนข้อมูลที่ใช้ในการเรียนรู้เพิ่มเติม

2) จำนวนรอบในการเรียนรู้เพิ่มเติม คือ จำนวนรอบที่แบบจำลองใช้ข้อมูลที่ถูกเลือกแต่ละรอบมาอัพเดทความรู้เพิ่มเติม

3) จำนวนข้อมูลที่ใช้ในการเรียนรู้เพิ่มเติม คือ จำนวนข้อมูลที่ถูกเลือกในแต่ละรอบเพื่อนำมาใช้เรียนรู้เพิ่มเติม

โดยที่การเรียนรู้เพิ่มเติมแต่ละรอบจะถูกวัดประสิทธิภาพกับชุดข้อมูลการตรวจสอบ และมีการกำหนดพารามิเตอร์ เช่น งบประมาณการเรียนรู้ของแบบจำลอง จำนวนรอบในการเรียนรู้เพิ่มเติม เพื่อให้แบบจำลองหยุดทำงานลง ส่วนการเรียนรู้เชิงรับจะเป็นฝึกสอนการจำแนกข้อมูลที่ผู้เรียนเป็นผู้รับฝ่ายเดียวโดยที่ไม่มีการโต้ตอบระหว่างผู้เรียนและผู้สอน และใช้ข้อมูลทั้งหมดในการฝึกสอน งานวิจัยนี้จะมุ่งเน้นที่ทฤษฎีที่เกี่ยวข้องกับการเรียนรู้เชิงรุก ซึ่งจะประกอบด้วย 3 หัวข้อ คือ การเลือกตัวอย่าง Stream-based, การเลือกตัวอย่าง Pool-based และ การเลือกตัวอย่างข้อมูลที่มีความไม่มั่นใจ ดังจะได้กล่าวต่อไป

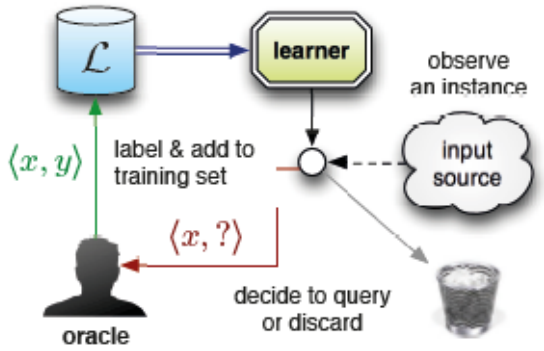


รูปที่ 2.1 การเปรียบเทียบขั้นตอนการทำงานระหว่างการเรียนรู้เชิงรุกและการเรียนรู้เชิงรับ (อ้างอิงจาก Fig. 1 ใน [7])

2.1.1 การเลือกตัวอย่าง Stream-based (Stream-based Selective Sampling)

การเลือกตัวอย่าง Stream-based [5] หรือเรียกอีกชื่อหนึ่งว่า การเรียนรู้เชิงรุกเชิงลำดับ ซึ่งการเลือกตัวอย่างข้อมูลที่ไม่ติดฉลาก (Unlabeled data) จะถูกแสดงเข้ามาแบบหนึ่งต่อหนึ่งจาก

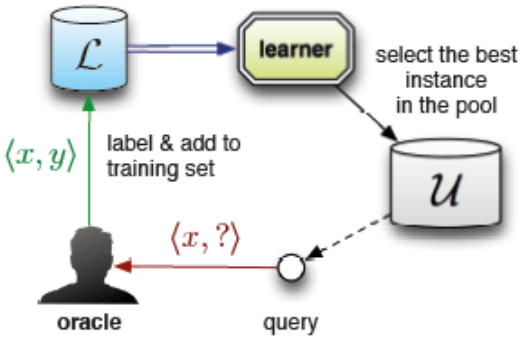
แหล่งข้อมูล ซึ่งผู้เรียนจะต้องพิจารณาข้อมูลที่ถูกแสดงนี้และตัดสินใจที่จะเลือกหลังจากที่ข้อมูลไม่ติดฉลากที่ได้ถูกเลือกจะถูกส่งต่อไปให้ผู้ชำนาญข้อมูลในกลุ่มนั้นเพื่อติดฉลาก (Labeled data) ส่วนข้อมูลไม่ติดฉลากและไม่ถูกเลือกจะถูกละทิ้งไป การเลือกตัวอย่างข้อมูลนี้จะมีการกำหนดขีดแบ่งขั้นต่ำ (Minimum thereshold) จะมีความแตกต่างกันออกไปตามแต่รูปแบบของข้อมูลและปัญหาแสดงได้ดังรูปที่ 2.2



รูปที่ 2.2 การเลือกตัวอย่าง Stream-based (อ้างอิงจาก Fig. 1.5a ใน [4])

2.1.2 การเลือกตัวอย่าง Pool-based (Pool-based Selective Sampling)

การเลือกตัวอย่าง Pool-based [4, 5] จะถูกใช้มากในโปรแกรมประยุกต์ในปัจจุบัน โดยการเลือกข้อมูลตัวอย่างจากชุดข้อมูลที่มีขนาดใหญ่จำนวนมากที่ไม่ติดฉลาก ความแตกต่างของ Pool-based เมื่อเทียบกับ Stream-based ที่เห็นได้ชัดคือ การประเมินตัวอย่างข้อมูลไม่ติดฉลากที่เข้ามาเป็นจำนวนมาก ส่วน Stream-based ข้อมูลไม่ติดฉลากจะเข้ามาแบบลำดับหนึ่งต่อหนึ่ง ดังรูปที่ 2.3



รูปที่ 2.3 แสดงการเลือกตัวอย่าง Pool-based (อ้างอิงจาก Fig. 1.5b ใน [4])

2.1.3 การเลือกตัวอย่างข้อมูลที่มีความไม่มั่นใจ (Uncertainty Sampling)

การเลือกตัวอย่าง ข้อมูลที่มีความไม่มั่นใจเป็นกลยุทธ์หนึ่งในการเรียนรู้เชิงรุก ซึ่งข้อมูลติดฉลากจะมีความไม่มั่นใจหรือมีความมั่นใจน้อย (least confident) [4] โดยที่กลยุทธ์การเลือกตัวอย่างข้อมูลจะมีหลักการในการเลือกดังนี้ เริ่มแรกจะใช้ข้อมูลติดฉลากจำนวนหนึ่งเพื่อฝึกสอนการจำแนกข้อมูล ซึ่งข้อมูลติดฉลากที่ใช้ในการฝึกสอนและทำนายผิดจะถูกนำมาฝึกสอนซ้ำใหม่อีกครั้ง หลังจากนั้นกระบวนการนี้จะถูกทำซ้ำจนประสิทธิภาพของแบบจำลองที่ถูกฝึกสอนโดยข้อมูลติดฉลากเหล่านั้นมีความแม่นยำในการทำนายสูงขึ้นจนเป็นที่พึงพอใจ โดยที่ความไม่มั่นใจของข้อมูลแบบสองคลาสจะอยู่ใกล้ค่า 0.5

2.2 การเรียนรู้เชิงรับ (Passive Learning)

การเรียนรู้เชิงรับ เป็นการเรียนรู้ข้อมูลทั้งหมดภายในครั้งเดียวโดยการโหลดข้อมูลทั้งหมดเข้าหน่วยความจำ ซึ่งระยะเวลาที่ใช้ในฝึกสอนกับแบบจำลอง (Model) ขึ้นกับรูปแบบของตัวจำแนกที่ใช้ โดยรูปแบบตัวจำแนกที่นิยม ได้แก่ ซัพพอร์ตเวกเตอร์แมชชีน (SVM) นาอีพเบย์ (Naïve Bayes) นิวรอลเน็ตเวิร์ก (Neural Network) ต้นไม้ตัดสินใจ (decision tree) เพื่อนบ้านใกล้สุด K (k-Nearest Neighbors)

2.3 การจำแนกสองคลาส (Binary Classification)

ข้อมูลที่ถูกใช้ในการจำแนกจะมีเพียง 2 คลาสเท่านั้น ตัวอย่างข้อมูลที่ใช้ในการจำแนก เช่น คำตอบของข้อสอบที่ใช้ในการวัดผลถูกหรือผิด การทดสอบสำหรับรายบุคคลว่าเป็นโรคมะเร็งหรือไม่ หรือการทดสอบการตั้งครรภ์ โดยการจำแนกในรูปแบบนี้เป็นแบบที่ง่ายกว่าการจำแนกหลายประเภทและแบบหลายฉลาก ซึ่งรูปแบบข้อมูลไม่มีความซับซ้อนทำให้สามารถนำมาใช้ทดสอบสมมติฐานและวิธีการของงานวิจัยในเบื้องต้น ก่อนที่จะนำไปประยุกต์ใช้ทดลองการจำแนกกับข้อมูลที่มีความซับซ้อนเหล่านั้น

2.4 การจำแนกหลายคลาสและแบบหลายฉลาก (Multiclass and Multi-Label)

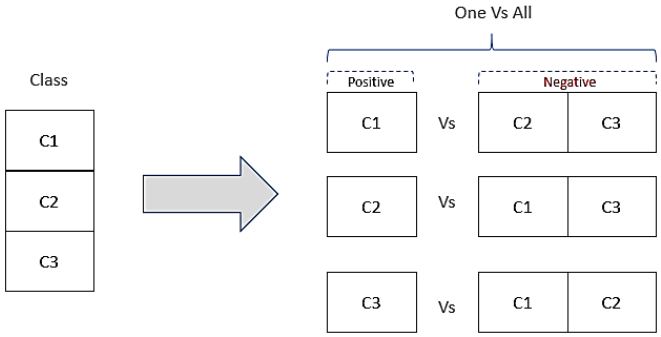
การจำแนกหลายคลาส ข้อมูลที่ถูกใช้ในการจำแนกจะมีมากกว่าสองคลาสขึ้นไป เช่น การจำแนกยี่ห้อของรถยนต์ โตโยต้า ฮอนด้า เบนซ์ ซึ่งข้อมูลที่ถูกจำแนกจะเป็นยี่ห้อใดยี่ห้อหนึ่งเท่านั้น ส่วนการจำแนกแบบหลายฉลาก ข้อมูลที่ถูกใช้ในการจำแนกสามารถมีเป็นได้มากกว่าหนึ่งคลาส เช่น ภาพยนตร์หนึ่งเรื่องที่สามารถเป็นได้มากกว่าหนึ่งคลาส Action และ Adventure หรือ การจำแนกเพลงซึ่งข้อมูลที่ถูกจำแนกได้มากกว่าหนึ่งคลาส

2.4.1 One-Versus-One

วิธีการนี้เป็นเทคนิคในการจำแนกข้อมูลระหว่างคลาสแบบหนึ่งต่อหนึ่งเพื่อฝึกสอนให้กับแบบจำลอง ซึ่งการจับคู่แต่ละคลาสเป็นจำนวนครั้ง $\frac{K(K-1)}{2}$ โดยที่ค่า K คือจำนวนคลาสของข้อมูล (Class) โดยที่เทคนิคนี้จะมีความเร็วกว่า One-Versus-All และนิยมใช้ในการจำแนกแบบหลายคลาส (Multiclass Classification) ตัวอย่าง ให้ C1, C2 และ C3 เป็นตัวแทนในแต่ละคลาส โดยมีทั้งหมด 3 คลาส การทำ One-Versus-One ได้แก่ C1 vs C2, C1 vs C3 และ C2 vs C3 เป็นต้น

2.4.2 One-Versus-All

วิธีการนี้เป็นเทคนิคการจำแนกข้อมูลระหว่างคลาสแบบหนึ่งต่อ $(K-1)$ โดยที่ค่า K เป็นจำนวนคลาสทั้งหมด (class) ของข้อมูลชุดนั้น ซึ่งเทคนิคนี้เป็นที่นิยมในการจำแนกคลาสแบบหลายฉลาก (Multi-Label Classification) [8] โดยที่เทคนิค One-Versus-All ความเร็วในการฝึกสอนให้กับผู้เรียนจะช้าและใช้เวลานานกว่าเทคนิค One-Versus-One มาก ตัวอย่าง ให้ C1, C2 และ C3 เป็นตัวแทนในแต่ละคลาส โดยมีทั้งหมด 3 คลาส การทำ One-Versus-All ได้แก่ C1 vs C2+C3, C2 vs C1+C3 และ C3 vs C1+C2 เป็นต้น ดังรูปที่ 2.4

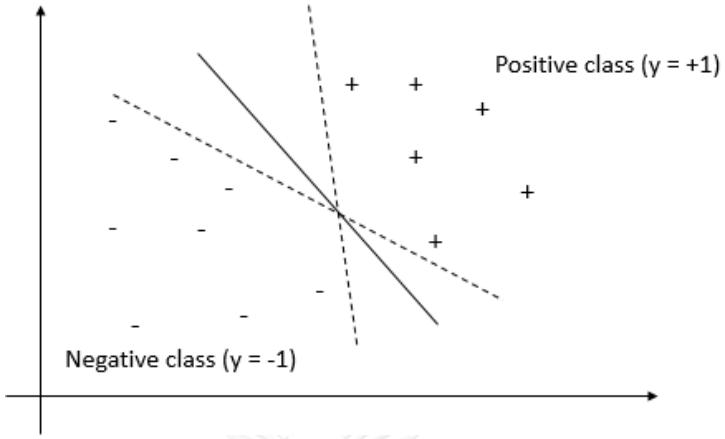


รูปที่ 2.4 แสดงตัวอย่างการทำ One-Versus-All ของคลาส C1, C2 และ C3

2.5 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Manchine: SVM)

ซัพพอร์ตเวกเตอร์แมชชีน (SVM) เป็นแบบจำลองที่ได้รับความนิยม เนื่องจากมีความแม่นยำสูงในการจำแนกประเภทของข้อมูล จากการเรียนรู้คุณลักษณะต่างๆ ของข้อมูลที่ถูกนำมาฝึกสอน ซึ่งข้อมูลที่นำฝึกสอน จะเป็นข้อมูลรูปแบบเชิงเส้น (linear) หรือข้อมูลรูปแบบไม่เชิงเส้น (nonlinear) โดยใช้ เคอร์เนลฟังก์ชัน (Kernel Function) ปรับเปลี่ยนมิติของข้อมูลจากพื้นที่มิติต่ำ (lower dimensional space) ไปยัง พื้นที่มิติสูง (higher dimensional space) ซัพพอร์ตเวกเตอร์แมชชีน สามารถถูกจัดกลุ่มให้อยู่ใน Supervised learning algorithm หรือการเรียนรู้จากข้อมูลที่มีผลเฉลย และถูกใช้แก้ไขปัญหาการจำแนกข้อมูลแบบสองคลาส โดยหลักการของซัพพอร์ตเวกเตอร์แมชชีน จะ

หา Decision Boundary ที่เหมาะสมโดยให้คลาสของข้อมูลมีระยะห่างจากเส้น Hyperplane มากที่สุด ดังรูปที่ 2.5

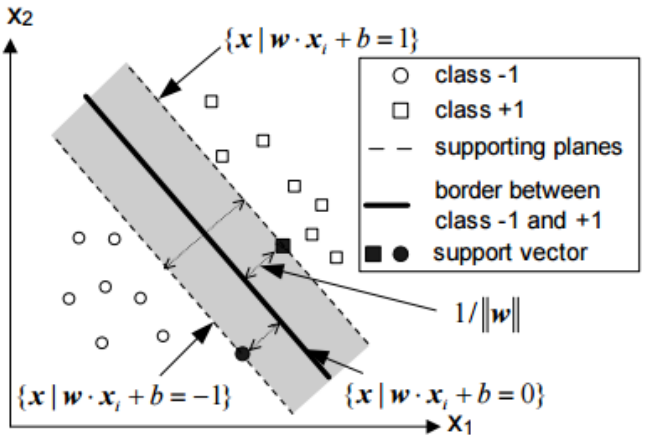


รูปที่ 2.5 การหา Decision Boundary ที่เหมาะสมของ ซัพพอร์ตเวกเตอร์แมชชีน

แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน เลือกเส้นตรง (linear) ที่มีระยะห่าง 2 คลาส มากที่สุด การเลือกเส้นตรงที่ใกล้กับข้อมูลแต่ละคลาสมากไป กรณีถ้ามีข้อมูลที่ห่างจากเส้นออกไปเล็กน้อยแบบจำลองก็จะทำนายข้อมูลนั้นผิดไป สามารถกำหนดได้ดังสมการที่ (1)

$$y(x) = sgn((w.x) + b) \quad (1)$$

โดยกำหนดให้ w คือ เวกเตอร์ของน้ำหนัก (weight) b คือ ค่าคงที่เกี่ยวกับความสัมพันธ์ระนาบ (plane) ที่เลื่อนออกไปจากจุดตั้งต้น จากรูปที่ 2.5 จะแสดงแบบจำลองเชิงเส้นของตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีน ระหว่าง 2 คลาส คือ คลาสบวก (positive class) และคลาสลบ (negative class)



รูปที่ 2.6 แสดงแบบจำลองเชิงเส้นของตัวจำแนก SVM ระหว่าง 2 คลาส (-1 และ +1)
(อ้างอิงจาก Fig. 1 ใน [9])

แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน โดยทั่วไปสามารถจำแนกข้อมูลแบบเชิงเส้น และข้อมูลแบบไม่เชิงเส้น ตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีนสามารถแก้ไขปัญหาการเรียนรู้และการจำแนกข้อมูลแบบไม่เชิงเส้น โดยการใช้เคอร์เนลฟังก์ชัน (Kernel Function) ดังสมการที่ (2) เปลี่ยนแปลงมิติของข้อมูล ซึ่งเคอร์เนลฟังก์ชันที่นิยมใช้มีอยู่ 4 ประเภท ดังสมการ ที่ (3) (4) (5) และ (6) ดังนี้

1) เคอร์เนลฟังก์ชัน(Kernel Function) :

$$K(x_i, x_j) = \emptyset(x_i)\emptyset(x_j) \quad (2)$$

2) เชิงเส้น (Linear):

$$K(x_i, x_j) = x_i . x_j \quad (3)$$

3) โพลิโนเมียล (Polynomial):

$$K(x_i, x_j) = (x_i . x_j + 1)^h \quad (4)$$

4) เกาเชียนเรเดียลเบสิสฟังก์ชัน (Gaussian Radial Basis Function-RBF) :

$$K(x_i, x_j) = e^{-\|x_i - x_j\|^2 / 2\sigma} \quad (5)$$

5) ซิกมอยด์ (Sigmoid):

$$K(x_i, x_j) = \tanh(\kappa x_i . x_j - \delta) \quad (6)$$

2.6 สโตแคสติกเกรเดียนเดสเซนท์ (Stochastic Gradient Descent: SGD)

การเรียนรู้ข้อมูลฝึกสอนด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน สำหรับข้อมูลขนาดใหญ่มีข้อจำกัดในการเรียนรู้เพิ่มเติม (Incremental Learning) กรณีที่มีข้อมูลใหม่เข้ามา ไม่สามารถที่จะเรียนรู้เพิ่มเติมได้ [10] ได้นำเสนอ สโตแคสติกเกรเดียนเดสเซนท์ หรือ SGD มาช่วยให้แบบจำลองเชิงเส้น (linear model) สามารถเรียนรู้เพิ่มเติมได้ โดยที่ไม่ต้องเรียนรู้ข้อมูลฝึกสอนทั้งหมดใหม่ (Passive learning) กำหนดให้ w คือ น้ำหนัก (weight) ที่ต้องการจะปรับค่า α คือ อัตราการเรียนรู้ $\frac{\partial J_t}{\partial w}$ คือเกรเดียน (gradient) ของฟังก์ชันต้นทุน ดังสมการที่ (7)

$$w_t = w_{t-1} - \alpha \frac{\partial J_t}{\partial w} \quad (7)$$

ฟังก์ชันการสูญเสีย (Loss Function) [10, 11] ที่สำคัญที่เกี่ยวเนื่องกับ สโตแคสติกเกรเดียนเดสเซนท์ มีดังนี้ ลอจิสติก (Logistic) การสูญเสียฮินจ์ (Hinge loss) เอสวีเอ็มปรับเรียบยกกำลังสอง (Quadratically smoothed SVM) การสูญเสียฮูเบอร์ (Huber loss) และ ฮูเบอร์ดัดแปลง (Modified Huber) สามารถแสดงดังสมการที่ (8) (9) (10) (11) (12) และ (13)

1) ฟังก์ชันการสูญเสีย (Loss Function):

$$\emptyset(p, y) \tag{8}$$

2) ลอจิสติก (Logistic):

$$\emptyset(p, y) = \ln(1 + \exp(-py)) \tag{9}$$

3) การสูญเสียฮินจ์ (Hinge loss):

$$\emptyset(p, y) = \max(0, 1 - py) \tag{10}$$

4) เอสวีเอ็มปรับเรียบยกกำลังสอง (Quadratically smoothed SVM):

$$\emptyset(p, y) = \frac{1}{2\gamma}\max(0, 1 - py)^2 \tag{11}$$

5) การสูญเสียฮูเบอร์ (Huber loss):

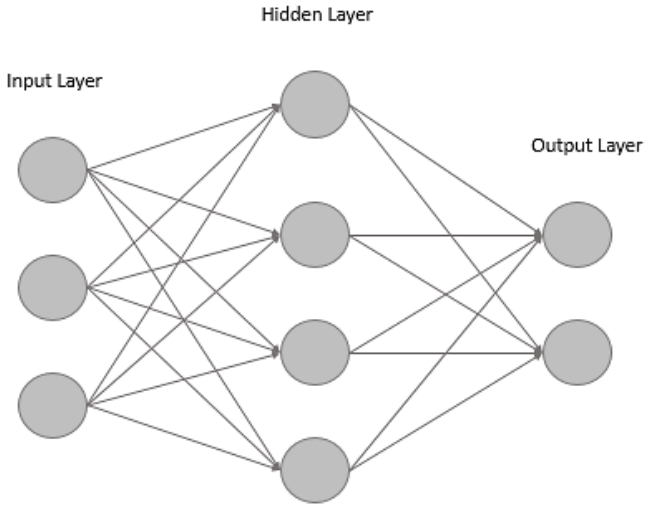
$$\emptyset(p, y) = (p - y)^2 \tag{12}$$

6) ฮูเบอร์ดัดแปลง (Modified Huber):

$$\emptyset(p, y) = max(0,1 - py)^2 \tag{13}$$

2.7 นิวรอลเน็ตเวิร์ก (Neural Network)

นิวรอลเน็ตเวิร์ก เป็นแบบจำลองที่ได้รับความนิยมอย่างสูง ในปัจจุบันนำมาใช้ในงานต่างๆ เช่น งานจดจำรูปภาพ การพยากรณ์อากาศ เป็นต้น นิวรอลเน็ตเวิร์กจัดเป็นการเรียนรู้แบบมีผู้สอน (Supervised Learning) มีหน่วยที่เล็กที่สุดที่เรียกว่าเพอร์เซ็ปตรอน (Perceptron) โครงสร้างทั่วไปของนิวรอลเน็ตเวิร์กประกอบไปด้วย 3 ชั้น คือ ชั้นนำเข้า (Input Layer) ชั้นซ่อน (Hidden Layer) และชั้นนำออก (Output Layer) ซึ่งในแต่ละชั้นจะมีเส้นกับเพอร์เซ็ปตรอนตัวอื่นที่อยู่ในชั้นติดกันทั้งหมด และในชั้นเดียวกันเพอร์เซ็ปตรอนจะไม่มีเส้นเชื่อมถึงกัน โดยที่ข้อมูลจะถูกป้อนเข้าชั้นนำเข้า ข้อมูลจะส่งผ่านจากชั้นหนึ่งไปสู่อีกชั้นหนึ่งและผลลัพธ์จะออกจากชั้นนำออก ดังรูปที่ 2.6



รูปที่ 2.7 โครงสร้างทั่วไปของนิวรอลเน็ตเวิร์กประกอบไปด้วย 3 ชั้น คือ ชั้นนำเข้า ชั้นซ่อน และชั้นนำออก

กำหนดให้ $f(x)$ แทนฟังก์ชันของเพอร์เซ็ปตรอน x คือข้อมูลป้อนเข้าชั้นนำเข้า n เป็นจำนวนข้อมูล w คือค่าเวกเตอร์น้ำหนัก (weight vector) b คือค่าไบแอส (bias) ดังสมการที่ (14)

$$f(x) = \begin{cases} 1 & if \ \sum_{i=1}^{n} w_i x_i + b > 0 \\ -1 & otherwise \end{cases} \quad (14)$$

ค่าเวกเตอร์น้ำหนักจะเปลี่ยนไปแต่ละรอบ (Iteration) การเรียนรู้ซึ่งเป็นการทำซ้ำจนกระทั่งเพอร์เซ็ปตรอนสามารถจำแนกข้อมูลได้ถูกต้อง โดยที่ช่วงเริ่มต้นจะมีการสุ่มค่าน้ำหนักให้ กำหนด ให้ t คือ ผลเฉลย o คือ ผลการทำนาย ถ้า $(t-o)$ เป็นศูนย์จะไม่มีการอัพเดทเวกเตอร์น้ำหนัก η คือ อัตราการเรียนรู้ (learning rate) ดังสมการที่ (15) และ (16)

$$w_i \leftarrow w_i + \Delta w_i \quad (15)$$

$$\Delta w_i = \eta(t-o)x_i \quad (16)$$

ฟังก์ชันกระตุ้น (Activation Function) จะถูกใช้ในชั้นส่งออกที่แต่ละเพอร์เซ็ปตรอน โดยฟังก์ชันกระตุ้นที่นิยมได้แก่ ฟังก์ชันซิกมอยด์ (Sigmoid Function) ฟังก์ชันซอฟต์แมกซ์ (Softmax Function) ฟังก์ชันเรคติไฟต์เชิงเส้น (Rectified Linear Unit Function) ฟังก์ชันแทนเจนต์ไฮเพอร์โบลิก (Hyperbolic Tangent Function) และฟังกชันขีดแบ่ง (Threshold Function)

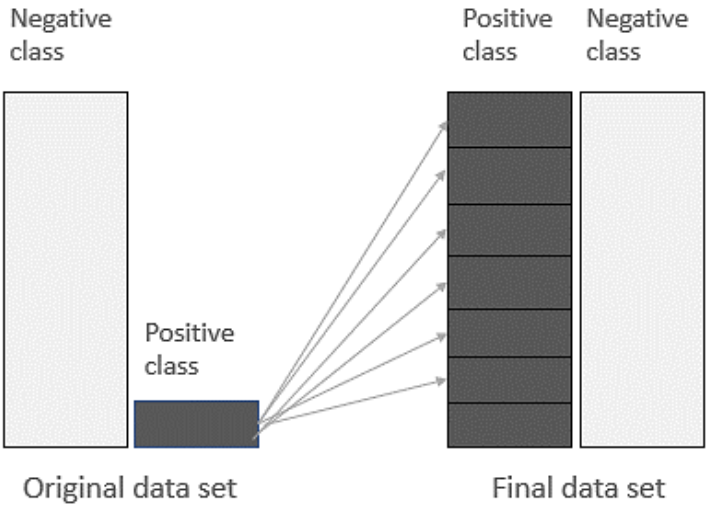
2.8 การวัดความคล้ายคลึงเชิงมุมโคไซน์ (Cosine Similarity)

การวัดความคล้ายคลึงเชิงมุมโคไซน์เป็นการหาความคล้ายกันของข้อมูล ซึ่งใช้วิธีการกำหนดตัวแทนกลุ่มของข้อมูล เพื่อคำนวณหาค่าเชิงมุมโคไซน์กับตัวอย่างข้อมูล โดยที่ถ้าค่าผลลัพธ์ได้ค่าเท่ากับ 1 ข้อมูลทั้งสองจะมีความคล้ายกันมากและมีทิศทาง (vector) ไปทางเดียวกัน กำหนดให้ x คือ ตัวแทนของกลุ่มข้อมูล y คือ ตัวอย่างข้อมูลที่ต้องการเปรียบเทียบ ดังสมการที่ (17)

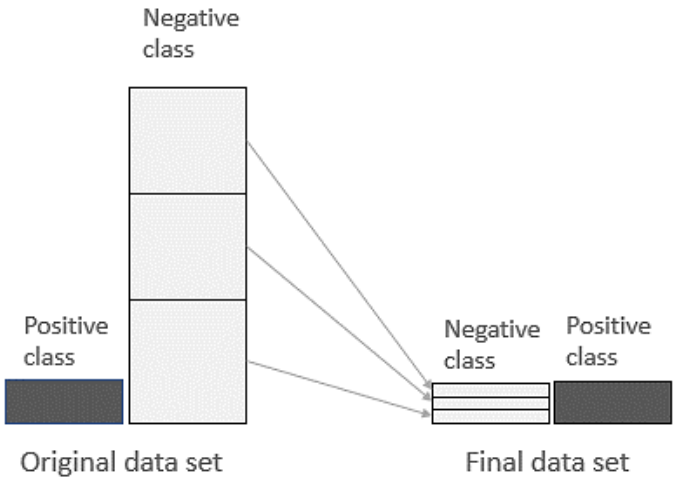
$$k(x,y) = \frac{xy}{\|x\|\|y\|} \tag{17}$$

2.9 กลยุทธ์สำหรับข้อมูลไม่สมดุล

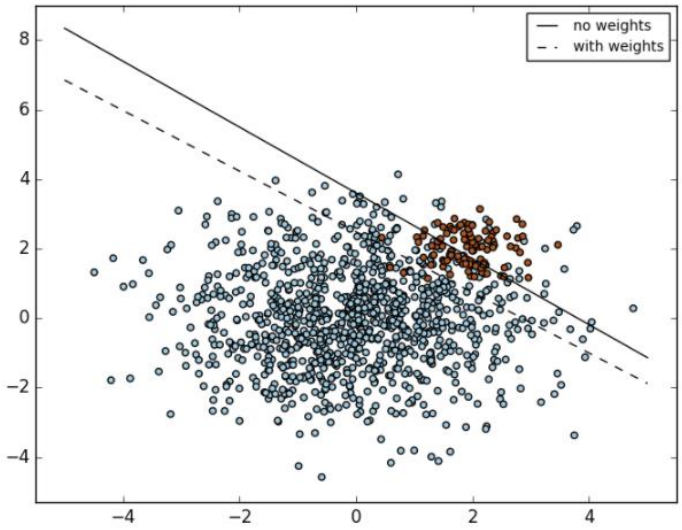
ในส่วนนี้จะกล่าวถึงวิธีการที่ทำให้ความแตกต่างระหว่างสัดส่วนอย่างเห็นได้ชัดของคลาสฝ่ายข้างมากและ คลาสฝ่ายข้างน้อย ซึ่งเป็นแนวทางโดยทั่วไปคือ การทำให้สมดุลกันระหว่างข้อมูล 2 ประเภท ดังจะกล่าวต่อไปสำหรับ การเลือกตัวอย่างเพิ่ม (Oversampling) การเลือกตัวอย่างลด (Undersampling) และ คอสท์เซนซิทีฟ (Cost Sensitive) [1] ซึ่งเทคนิคการเลือกตัวอย่างเพิ่มจะดำเนินการโดยการเพิ่มข้อมูลของคลาสฝ่ายข้างน้อย ให้มีจำนวนสัดส่วนของข้อมูลที่เพิ่มขึ้นเพื่อสมดุลกับคลาสฝ่ายข้างมาก สำหรับการเรียนรู้และสร้างแบบจำลองเพื่อให้มีประสิทธิภาพในการทำนายสำหรับการจำแนกข้อมูลมีความแม่นยำขึ้น อย่างไรก็ตามปัญหาของการใช้วิธีการนี้ที่ตามมาแบบจำลองจดจำรูปแบบข้อมูลฝึกสอนมากเกินไป (Overfitting) การเรียนรู้จากข้อมูลไม่สมดุลกับวิธีการดังที่กล่าวมาข้างต้นอาจ จะมีผลกับข้อมูลชุดเดียวเท่านั้น แต่เมื่อนำไปใช้กับข้อมูลชุดอื่นประสิทธิภาพของการจำแนกข้อมูลไม่ดีเท่ากับการใช้ชุดข้อมูลก่อนหน้านี้ ในทางตรงข้ามการเลือกข้อมูลตัวอย่างลดจะเป็นการลดปริมาณตัวอย่างข้อมูลในคลาสฝ่ายข้างมากให้มีจำนวนสัดส่วนข้อมูลสมดุลกับคลาสฝ่ายข้างน้อย เพื่อให้ประสิทธิภาพของแบบจำลองที่ได้มีการจำแนกข้อมูลได้ดีขึ้น แต่จุดด้อยของวิธีการนี้ได้แก่ ข้อมูลที่ถูกขจัดออกไปจากคลาสฝ่ายข้างมาก อาจจะเป็นข้อมูลที่สำคัญต่อการจำแนกข้อมูล ที่มีผลต่อประสิทธิภาพของแบบจำลอง นอกเหนือจากนั้นคอสท์เซนซิทีฟ คือ การให้น้ำหนักเพิ่มกับคลาสฝ่ายข้างน้อยเพื่อให้แบบจำลองสามารถจำแนกคลาสฝ่ายข้างน้อยได้แม่นยำขึ้น สามารถแสดงวิธีการเลือกตัวอย่างเพิ่ม การเลือกตัวอย่างลด และคอสท์เซนซิทีฟ ได้ดังรูปที่ 2.8 2.9 และ 2.10 ตามลำดับ



รูปที่ 2.8 แสดงการเพิ่มข้อมูลให้กับคลาสฝั่งข้างน้อยโดยใช้เทคนิคการเลือกตัวอย่างเพิ่ม (Oversampling) ให้เท่ากับข้อมูลคลาสฝั่งข้างมาก



รูปที่ 2.9 แสดงการลดจำนวนข้อมูลของคลาสฝ่ายข้างมากโดยใช้เทคนิคการเลือกตัวอย่างลด (Undersampling) ให้เท่ากับข้อมูลคลาสฝั่งข้างน้อย



รูปที่ 2.10 แสดงการเพิ่มน้ำหนักให้กับคลาสฝั่งข้างน้อยโดยใช้เทคนิค Cost Sensitive [12]

2.9.1 สุ่มเลือกตัวอย่างลด (Random Undersampling)

เป็นเทคนิคที่ง่ายต่อการแก้ไขปัญหาข้อมูลไม่สมดุลของการเรียนรู้เชิงรับซึ่งการแก้ไขปัญหาเบื้องต้นโดยการทำให้มีความสมดุลระหว่างคลาสฝ่ายข้างน้อย และคลาสฝ่ายข้างมาก วิธีการจะทำการสุ่มเลือกข้อมูลจากคลาสฝ่ายข้างมากในจำนวนที่เท่ากับคลาสฝ่ายข้างน้อย และนำจำนวนข้อมูลที่ได้จากทั้งสองคลาสไปสร้างแบบจำลอง วิธีการนี้จะเป็นการลดจำนวนข้อมูลในคลาสฝ่ายข้างมากทำให้ประสิทธิภาพความแม่นยำของแบบจำลองมีค่าที่สูงขึ้นกว่าวิธีการเรียนรู้เชิงรับแบบทั่วไป

2.10 การวัดประสิทธิภาพการทำงาน (Performance Evaluation)

2.10.1 ตัววัดประสิทธิภาพการจำแนกข้อมูลสองคลาส (Binary Classification Performance Measurement)

ในหัวข้อนี้จะกล่าวถึงตัววัดประสิทธิภาพที่นิยมใช้ในการวัดประสิทธิภาพการจำแนกข้อมูลสองคลาส [1] ซึ่งประกอบด้วย TP (True Positive), TN (True Negative), FP (False Positive) และ FN (False Negative) ดังตารางที่ 1 โดยที่จะนำค่าเหล่านี้มาใช้ในการหาค่า Precision, Recall, Accuracy และค่า F1 ต่อไปดังสมการ (18) (19) (20) และ (21)

ตารางที่ **2.1** แสดงความสัมพันธ์ Predicted คลาส และ **True** คลาส

	Predicted negative	**Predicted positive**
True negative	TN	FP
True positive	FN	TP

$$Precision = \frac{TP}{TP + FP} \quad (18)$$

$$Recall = \frac{TP}{TP + FN} \quad (19)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (20)$$

$$F\text{–}Measure = \frac{(1 + \beta^2) \times Recall \times Precision}{\beta^2 \times Recall + Precision} \quad (21)$$

กำหนดให้ค่า β = 1 จะได้ดังสมการ (22)

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (22)$$

2.10.2 ตัววัดประสิทธิภาพการจำแนกข้อมูลสองคลาสสำหรับข้อมูลไม่สมดุล (Binary Classification Performance Measurement with Imbalanced data)

ในหัวข้อนี้จะแสดงการใช้ตัววัดผล Geometric Mean (G-mean) [1] เพื่อวัดประสิทธิภาพของแบบจำลอง ซึ่งเป็นการเปรียบเทียบของผลลัพธ์ที่ได้จากการจำแนกกับผลลัพธ์ที่ถูกต้องสำหรับข้อมูลสองคลาสและไม่สมดุล ดังตารางที่ 1 โดยที่สมการที่ (25) จะเป็นตัวประสิทธิภาพที่ใช้สำหรับข้อมูลสองคลาสที่ไม่สมดุล

$$a^+ = \frac{TP}{TP + FN} \quad (23)$$

$$a^- = \frac{TN}{TN + FP} \quad (24)$$

$$G = \sqrt{a^+ \times a^-} \quad (25)$$

2.10.3 ตัววัดประสิทธิภาพการจำแนกประเภทแบบหลายฉลาก (Multi-Label Classification Performance Measurement

การวัดประสิทธิภาพการจำแนกประเภทแบบหลายฉลากจะใช้การเฉลี่ยจากค่า Precision, Recall และ F1 [1] ซึ่งการเฉลี่ยจะมีสองวิธีที่นิยมคือ ค่าเฉลี่ยแมโคร (macro-average) และค่าเฉลี่ยไมโคร (micro-average) [1] โดยวิธีของค่าเฉลี่ยแมโคร จะทำการคำนวณค่า Presicsion, Recall และ F1 ก่อนหลังจากนั้นจะนำมาค่าเหล่านั้นมาเฉลี่ย ดังสมการที่ (26) ส่วนวิธีการของค่าเฉลี่ยไมโคร จะรวมค่า TP, FP, TN และ FN แต่ละประเภทก่อน ดังสมการ (27)

โดยที่ k เป็นจำนวนคลาส และ $B(TP_k, FP_k, TN_k, FN_k)$ ทำหน้าที่เป็นตัวแทนของ “Binary Classification Metric”

$$B_{Marcro}(h) = \frac{1}{q} \sum_{k=1}^{q} B(TP_k, FP_k, TN_k, FN_k) \quad (26)$$

$$B_{Micro}(h) = B(\sum_{k=1}^{q} TP_k, \sum_{k=1}^{q} FP_k, \sum_{k=1}^{q} TN_k, \sum_{k=1}^{q} FN_k) \quad (27)$$

2.11 งานวิจัยที่เกี่ยวข้อง (Related Work)

งานวิจัย [13] นำเสนอ “Mean Score on Unlabeled set (MSU)” ซึ่งถูกใช้วัดประสิทธิภาพของแบบจำลอง การแก้ปัญหาชุดข้อมูลไม่สมดุล โดยกลยุทธ์การเลือกข้อมูลของการเรียนรู้เชิงรุก ในรูปแบบที่แตกต่างกันไป ซึ่งสามารถทำได้ดีสำหรับชุดข้อมูลไม่สมดุล 4 ชุดทดสอบดังต่อไปนี้ Healthcare Insurance claim (CLAIMS), Network Instrustion detection (KDD), HIV Active compound (HIVA) และ Embryology (ZEBRA) ซึ่งในงานวิจัย [13] จะใช้ตัวจำแนกข้อมูลเป็น “Linear SVM” และมีการทำ “5-fold cross-validation” เพื่อการเรียนรู้และการวัดประสิทธิภาพของแบบจำลอง โดยที่การเริ่มสร้างแบบจำลองจะใช้ข้อมูลที่ได้จากสุ่มข้อมูลตัวอย่าง หลังจากนั้นจะเปรียบเทียบใช้กลยุทธ์การเลือกข้อมูลแบบต่างๆ ได้แก่ “Uncertainty Sampling” “Density Sampling” “Hybrid Sampling” “Certainty Sampling” และ “Sparsity Sampling” ซึ่งผลลัพธ์ที่ได้กลยุทธ์การเลือกข้อมูล “Uncertainty Sampling” ได้ผลลัพธ์ที่ดีที่สุดและผลลัพธ์มีส่วนสำคัญการปรับปรุงประสิทธิภาพของแบบจำลอง โดยเหนือกว่ากลยุทธ์ “Pool-based instance selection strategies (ISS)” และมีประสิทธิภาพของแบบจำลองอยู่ที่ 80%-100% โดยการใช้วิธีการวัดของงานวิจัย [13] อย่างไรก็ตามยังมีข้อจำกัดในเรื่องของตัวจำแนกที่เป็น “Linear SVM” และการใช้วิธีการเลือกข้อมูลของงานวิจัย [13] กับชุดข้อมูลไม่ติดฉลาก

งานวิจัย [14] นำเสนอวิธีการใหม่สำหรับการเลือกข้อมูลโดยใช้ความขัดแย้ง (disagreement) เพื่อตัดสินว่าเป็นข้อมูลที่สำคัญ โดยวิธีการนี้จัดหามาจาก “Estmation-Exploration Algorithm (EEA)” การเลือกข้อมูลใช้ในการเรียนรู้เชิงรุก ซึ่งมีประสิทธิภาพเกี่ยวกับความเร็วและลดจำนวนตัวอย่างข้อมูลเพื่อใช้ในการเรียนรู้ของแบบจำลอง ตัวจำแนกที่ใช้ในงานวิจัย [14] จะเป็น Artificial Neural Network (ANN) เป็นมาตรฐานทั่วโปคือมี 3 ชั้น โดยที่ชั้นรับเข้า (input layer) จะมี 16 โหนด ชั้นซ่อน (hidden layer) จะมี 8 โหนด และชั้นนำออก (output layer) จะมี 1 โหนด (binary node) และชุดข้อมูลที่ใช้เป็นข้อมูลที่เกี่ยวกับทางด้านการแพทย์ เรียกว่า “National Truma Data Bank (NTDB)” โดยที่ชุดข้อมูลมีขนาดใหญ่และไม่สมดุลมาก (Highly Imbalanced data set) ซึ่งผลลัพธ์ของการทดลอง [14] คือการเปรียบเทียบระหว่าง “Informative Sampling” ซึ่งเป็นวิธีการใหม่ในงานวิจัย [14] กับ “Random Sampling” และ “Balanced Sampling” โดยที่ผลลัพธ์ประสิทธิภาพของแบบจำลองจะเรียงจากมากไปน้อย ได้แก่ “Informative Sampling” “Random Sampling” และ “Balanced Sampling” และผลลัพธ์ค่าเฉลี่ยจำนวนข้อมูลการทำนายผิดของวิธีการที่กล่าวมาคือ 233.6 407.5 และ 447.2 ตามลำดับ จากจำนวนตัวอย่างในการทดสอบทั้งสิ้น 2000 ตัวอย่าง และใช้จำนวนรอบในการทำ 600 รอบ และทำเป็น

จำนวน 30 ครั้งในแต่ละวิธี อย่างไรก็ตาม ข้อจำกัดของงานวิจัย [14] คือจำนวนรอบในการทำซ้ำที่มากและทำซ้ำเป็นจำนวนหลายครั้งในแต่ละวิธีการในการทดลอง

งานวิจัย [15] นำเสนอวิธีการจำแนกแบบหลายฉลากสำหรับข้อมูลประเภทข้อความ โดยที่ต้องการลดจำนวนข้อมูลที่ใช้ในการเรียนรู้เพื่อสร้างแบบจำลอง เนื่องจากข้อมูลที่ติดฉลากมีราคาแพง จึงต้องใช้เวลาและผู้ชำนาญในข้อมูลที่เกี่ยวข้อง ข้อมูลแบบหลายฉลากจะต้องใช้ความพยายามในการจำแนกข้อมูลมากกว่าข้อมูลฉลากเดียว (Single-Label) มาก งานวิจัย [15] ต้องการความคาดหมายการลดการสูญเสียที่เหมาะที่สุด (Optimize the expected loss reduction) ซึ่งการสูญเสียของแบบจำลองคือการประมาณโดยแบบปริภูมิ (Version space) โดยการใช้การเรียนรู้เชิงรุกแบบ pool-based (The pool-based Active Learning) ซึ่งจะใช้ตัวจำแนก SVM การเลือกข้อมูลตัวอย่างมาใช้ในการเรียนรู้เชิงรุกและใช้วิธีการความไม่แน่นอนแบบ Query By Committee สำหรับข้อมูลที่มีความขัดแย้งกัน (disagreement) เทคนิคที่นิยมใช้สำหรับข้อมูลแบบหลายฉลากและใช้ในงานวิจัย [15] คือ One-Versus-All ซึ่งจะถูกใช้จำแนกความเป็นไปได้ของคลาสทั้งหมด ในการทดลองของงานวิจัย [15] มีการนำเสนอวิธีการที่เรียกว่า MMC เปรียบเทียบกับวิธีการ Random BinMin และ MML ชุดข้อมูลที่ใช้ในงานวิจัย [15] จะมีทั้งสิ้น 7 ชุด ได้แก่ ชุดข้อมูล RCV1-R2 และ ชุดข้อมูล yahoo 6 ชุด เริ่มแรกจะสุ่มข้อมูลตัวอย่างจำนวน 500 ตัวอย่าง ในการเริ่มสร้างแบบจำลอง โดยที่การทำซ้ำมีจำนวน 50 รอบ แต่ละรอบจะเลือกข้อมูลไม่ติดฉลากจำนวน 20 ตัวอย่าง มาใช้ในการเรียนรู้เชิงรุก ซึ่งเมื่อวัดผลจากข้อมูลตัวอย่างจำนวน 1000 ตัวอย่างในชุดข้อมูล RCV1-R2 ผลปรากฏว่าวิธีการ MMC อยู่ที่ 82.88% ซึ่งสูงกว่าวิธีการ BinMin MML Random อยู่ที่ 77.11 75.77 และ 75.12 ตามลำดับ ในขณะเดียวสำหรับชุดข้อมูลตัวอย่าง yahoo 6 ชุด (Art&Humanities, Business&Economy, Computer&Internet, Education, Entetainment, Health) วิธีการ MMC ยังคงเหนือกว่าวิธีการ BinMin MML Random ทุกชุดข้อมูล อยู่ที่ 66.50 78.97 74.40 68.99 73.40 และ 79.78 ตามลำดับของชุดข้อมูล อย่างไรก็ตามจากงานวิจัย [15] ยังมี ข้อจำกัดของตัวจำแนกที่ใช้คือ SVM และผลการทดลองที่ยังมีค่าที่ไม่สูงมาก

ในงานวิจัยนี้จะนำเสนอกลยุทธ์ใหม่สำหรับข้อมูลขนาดใหญ่และไม่สมดุล โดยจะประยุกต์แนวคิดและวิธีการไปยังชุดข้อมูลแบบสองคลาส (Binary Data sets) และข้อมูลแบบหลายฉลาก (Multi-Label Data sets) ซึ่งแนวคิดวิธีการและชุดข้อมูลที่ใช้จะมีความแตกต่างจากงานวิจัย [13-15] โดยจุดประสงค์ในงานวิจัยต้องการปรับปรุงความแม่นยำของแบบจำลอง โดยเปรียบเทียบกับการเรียนรู้เชิงรับเท่านั้น เนื่องจากข้อจำกัดของงานวิจัย [13-15] ที่กล่าวมาข้างต้น

บทที่3
การใช้การเรียนรู้เชิงรุกกับชุดข้อมูลขนาดใหญ่และไม่สมดุล

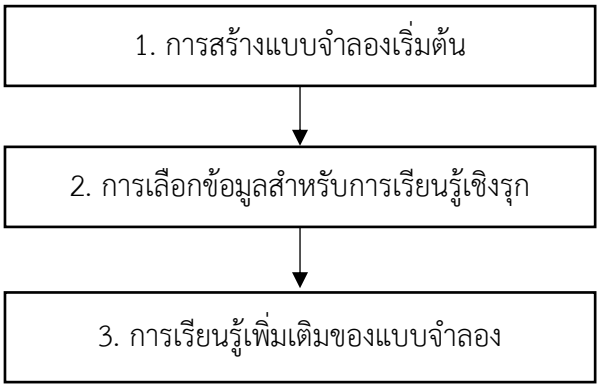
ในบทนี้จะนำเสนอการจำแนกข้อมูลขนาดใหญ่และไม่สมดุลจำนวน 2 รูปแบบ รูปแบบที่ 1 เสนอวิธีการจำแนกข้อมูลขนาดใหญ่และไม่สมดุลด้วยนิวรอลเน็ตเวิร์ก และรูปแบบที่ 2 เสนอวิธีการจำแนกข้อมูลขนาดใหญ่และไม่สมดุลแบบหลายฉลากด้วยซัพพอร์ตเวกเตอร์แมชชีน

รูปแบบที่ 1 เป็นการจำแนกข้อมูลสองคลาสและข้อมูลแบบหลายฉลากด้วยตัวจำแนกนิวรอลเน็ตเวิร์ก โดยเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้เชิงรุกและการเรียนรู้เชิงรับ ด้วยการวัดจีมีน (G-mean) สำหรับข้อมูลสองคลาส ส่วนข้อมูลหลายฉลากจะวัดด้วย ค่าเฉลี่ยไมโคร (Micro-average) และค่าเฉลี่ยแมโคร (Macro-average) แต่เนื่องด้วยผลการทดลองในบทที่ 4 เมื่อเปรียบเทียบผลการวัดประสิทธิภาพที่ได้ระหว่างการเรียนรู้เชิงรุกและการเรียนรู้เชิงรับด้วยนิวรอลเน็ตเวิร์ก กับการเรียนรู้เชิงรับด้วยซัพพอร์ตเวกเตอร์แมชชีนบนข้อมูลหลายฉลาก ผลการวัดของแบบจำลองที่สร้างด้วยนิวรอลเน็ตเวิร์คจะมีค่าน้อยกว่าแบบจำลองที่สร้างด้วยซัพพอร์ตเวกเตอร์แมชชีนค่อนข้างมาก ดังนั้นจึงเป็นที่มาของการนำเสนอรูปแบบที่ 2

รูปแบบที่ 2 เป็นการจำแนกข้อมูลหลายฉลากด้วยตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีน และมีการเปรียบเทียบประสิทธิภาพของการเรียนรู้เชิงรุกและการเรียนรู้เชิงรับ โดยที่งานวิจัยชิ้นนี้ได้นำเสนอกลยุทธ์สำหรับการเรียนรู้เชิงรุก ได้แก่ การเรียนรู้เชิงรุกเอสวีเอ็มเลือกข้อมูลไม่มั่นใจบนซัพพอร์ตเวกเตอร์แบบใช้ข้อมูลซ้ำ (Active Learning SVM selective uncertain data with support vector repeat: AL-SVM-SV-R) และ การเรียนรู้เชิงรุกเอสวีเอ็มเลือกข้อมูลไม่มั่นใจบนซัพพอร์ตเวกเตอร์แบบไม่ใช้ข้อมูลซ้ำ (Active Learning SVM selective uncertain data with support vector no repeat: AL-SVM-SV-N)

การเรียนรู้เชิงรุกเป็นการเรียนรู้ที่ค่อยๆเรียนรู้จากข้อมูลขนาดใหญ่ โดยเริ่มต้นเรียนรู้จากกลุ่มตัวแทนที่ถูกเลือก และเรียนรู้เพิ่มเติมจากข้อมูลที่ทำนายผิดพลาดในอดีต ทำให้สามารถแก้ไขปัญหาการจำแนกข้อมูลขนาดใหญ่ ดังนั้นในงานวิจัยชิ้นนี้ จึงนำเสนอการเรียนรู้เชิงรุกเพื่อการจำแนกข้อมูลขนาดใหญ่ 3 ขั้นตอน ประกอบด้วย การสร้างแบบจำลองเริ่มต้น (Initial Model Construction), การเลือกข้อมูลสำหรับการเรียนรู้เชิงรุก (Active Data Selection) และการเรียนรู้เพิ่มเติมของแบบจำลอง (Active Model Learning) โดยสามารถดูรายละเอียดขั้นตอนได้ดังรูปที่ 3.1

1) การสร้างแบบจำลองเริ่มต้น (Initial Model Construction) เป็นการนำกลุ่มตัวแทนข้อมูลจากตัวอย่างข้อมูลทั้งหมดที่ถูกเลือกมานำใช้ในการสร้างแบบจำลองเริ่มต้น



รูปที่ 3.1 ขั้นตอนวิธีการการเรียนรู้เชิงรุกสำหรับข้อมูลขนาดใหญ่และไม่สมดุล

2) การเลือกข้อมูลสำหรับการเรียนรู้เชิงรุก (Active Data Selection) เป็นการเลือกตัวอย่างข้อมูลเพื่อใช้สำหรับการเรียนรู้เพิ่มเติมของแบบจำลอง โดยการเลือกข้อมูลของขั้นตอนนี้จะถูกกำหนดโดยค่างบประมาณของการสอบถาม (budget of queries) และจำนวนรอบของการเรียนรู้ โดยกำหนดค่า β คือ งบประมาณของการสอบถาม และ n คือ จำนวนรอบของการเรียนรู้

3) การเรียนรู้เพิ่มเติมของแบบจำลอง (Active Model Learning) เป็นนำข้อมูลที่ได้จากการเลือกข้อมูลสำหรับการเรียนรู้เชิงรุกมาใช้ในการเรียนรู้เพิ่มเติม เพื่อเพิ่มประสิทธิภาพของแบบจำลองเริ่มต้น โดยที่การเรียนรู้เพิ่มเติมจะหยุดทำก็ต่อเมื่อ β หรือ n หรือขนาดจำนวนข้อมูลที่ถูกเลือกจะมีค่าเท่ากับศูนย์ ซึ่งแต่ละรอบของการเรียนรู้จะถูกวัดด้วยชุดข้อมูลตรวจสอบ (validation set) โดยที่การเรียนรู้เชิงรุกจะเลือกแบบจำลองที่ดีที่สุดบนชุดตรวจสอบและทดสอบด้วยชุดทดสอบ (testing set) อีกครั้งเพื่อวัดประสิทธิภาพของแบบจำลอง

3.1 รูปแบบที่ 1 เสนอวิธีการจำแนกข้อมูลขนาดใหญ่และไม่สมดุลด้วยนิวรอลเน็ตเวิร์ก

ในหัวข้อนี้จะอธิบายขั้นตอนการทำงานของการเรียนรู้เชิงรุกที่สร้างแบบจำลองด้วยนิวรอลเน็ตเวิร์กบนชุดข้อมูลสองคลาส และบนชุดข้อมูลแบบหลายฉลาก โดยใช้ขั้นตอนที่ได้กล่าวมาข้างต้น 3 ขั้นตอน ประกอบด้วย การสร้างแบบจำลองเริ่มต้น การเลือกข้อมูลสำหรับการเรียนรู้เชิงรุก และการเรียนรู้เพิ่มเติมของแบบจำลอง

3.1.1 การจำแนกข้อมูลขนาดใหญ่และไม่สมดุลสำหรับข้อมูลสองคลาสด้วยนิวรอลเน็ตเวิร์ก

สำหรับหัวข้อนี้ได้มีกำหนดกลยุทธ์สำหรับการจำแนกข้อมูลสองคลาสด้วยนิวรอลเน็ตเวิร์ก ได้แก่ การเรียนรู้เชิงรุกไม่เอนเอียง (Unbiased Active Learning: UAL) และการเรียนรู้เชิงรุกไม่เอนเอียงแบบมีสัดส่วน (Unbiased Active Learning Proportion: UAL-P) ซึ่งทั้งสองกลยุทธ์ถูกนำเสนอเพื่อแก้ไขปัญหาไม่สมดุลของข้อมูลสองคลาส สำหรับการเรียนรู้เชิงรุก

1) การเรียนรู้เชิงรุกไม่เอนเอียง (Unbiased Active Learning: UAL) เริ่มต้นการเลือกข้อมูลตัวแทนสำหรับการสร้างแบบจำลองเริ่มต้นของกลยุทธ์นี้จะเลือกตัวแทนข้อมูลของคลาสฝ่ายข้างน้อย (Minority class) มาจำนวนทั้งหมดจากชุดข้อมูล โดยที่ข้อมูลของคลาสฝ่ายข้างมาก (Majority class) จะถูกเลือกจากค่าความน่าจะเป็นที่สูง (ความน่าเป็นมากกว่า 0.6 โดยการใช้แบบจำลองนาอีฟเบย์) ในขนาดจำนวนที่เท่ากับคลาสฝ่ายข้างน้อย เพื่อสำหรับการสร้างแบบจำลองเริ่มต้น ขั้นตอนที่สองการเลือกข้อมูลสำหรับการเรียนรู้เชิงรุก แบบจำลองที่ถูกสร้างจะทำการเลือกข้อมูลที่ทำนายผิด โดยเลือกขนาดข้อมูลคลาสฝ่ายข้างมากที่ทำนายผิดให้มีขนาดเท่ากับขนาดจำนวนของคลาสฝ่ายน้อยที่ทำนายผิด (balancing classes) โดยเลือกข้อมูลที่มีค่าความน่าจะเป็นอยู่ระหว่าง 0.45-0.55 ซึ่งค่าได้มาจากฟังก์ชันการสูญเสียคลอสเอ็นโทรปี ซึ่งข้อมูลที่ถูกเลือกในขั้นตอนนี้สามารถนำมาใช้ซ้ำได้ ขั้นตอนที่สาม เป็นการเรียนรู้เพิ่มเติมของแบบจำลอง ซึ่งจะนำข้อมูลที่ได้จากการเลือกจากขั้นตอนที่สอง มาใช้ในการเรียนรู้เพิ่มเติม โดยมีพารามิเตอร์ที่เกี่ยวข้องคือ β, n ซึ่งประสิทธิภาพของแบบจำลองแต่ละรอบของการเรียนรู้จะถูกวัดด้วยชุดตรวจสอบ เมื่อการเรียนรู้เพิ่มเติมหยุดแบบจำลองที่ดีที่สุดบนชุดตรวจสอบจะถูกเลือก และทดสอบด้วยชุดทดสอบอีกครั้ง

2) การเรียนรู้เชิงรุกไม่เอนเอียงแบบมีสัดส่วน (Unbiased Active Learning Proportion: UAL-P) โดยกลยุทธ์นี้มีความคล้ายคลึงกับกลยุทธ์ UAL ต่างกันที่ UAL-P มีการกำหนดสัดส่วนของข้อมูลคลาสฝ่ายข้างน้อยเป็น 15%, 20%, 30%, 35% และ 40% ตามลำดับ และกำหนดให้ UAL-P (15%), UAL-P (20%), UAL-P (25%), UAL-P (30%), UAL-P (35%) และ UAL-P (40%) แทนสัดส่วนข้างต้นตามลำดับ ซึ่งการทำงานขั้นตอนที่หนึ่งถึงขั้นตอนที่สามจะเหมือนกับกลยุทธ์ UAL โดยสามารถดูรหัสเทียมได้ดังรูปที่ 3.2

3.1.2 การจำแนกข้อมูลขนาดใหญ่และไม่สมดุลสำหรับข้อมูลแบบหลายฉลากด้วยนิวรอลเน็ตเวิร์ก

ชุดข้อมูลแบบหลายฉลากเป็นข้อมูลที่มีความซับซ้อนและยากต่อการจำแนกมากกว่าชุดข้อมูลสองคลาส ในหนึ่งตัวอย่างข้อมูลสามารถเป็นได้มากกว่าหนึ่งคลาส (การเลือกตัวแทนโดยใช้นาอีฟเบย์เพื่อหาตัวแทนข้อมูลสองคลาสจากหัวข้อ 3.1.1 จึงไม่เหมาะสม) โดยที่ขั้นตอนการเรียนรู้ อ้างอิงจากรูปที่ 3.1 ขั้นตอนที่หนึ่ง การสร้างแบบจำลองเริ่มต้น จะหากลุ่มตัวแทนของข้อมูลในแต่ละคลาสโดยใช้การวัดความคล้ายคลึงเชิงมุมโคไซน์ของข้อมูลมาใช้ในการสร้างแบบจำลองเริ่มต้น ซึ่งการแก้ไขปัญหาไม่สมดุลของข้อมูลจะใช้คอสท์เซนซิทีฟ (Cost Sensitive) ขั้นตอนที่สอง การเลือกข้อมูลสำหรับการเรียนรู้เชิงรุกแบบจำลองจะทำการเลือกข้อมูลสำหรับเรียนรู้เพิ่มเติมโดยการนำชุดข้อมูลฝึกสอน (Training Set) มาทำนายแล้วเลือกข้อมูลที่ทำนายผิดที่มีค่าความน่าจะเป็นอยู่ระหว่าง 0.45-0.55 ซึ่งค่าได้มาจากฟังก์ชันการสูญเสียคลอสเอ็นโทรปี โดยข้อมูลที่ถูกเลือกสามารถนำมาใช้ซ้ำได้

ขั้นตอนที่สาม การเรียนรู้เพิ่มเติมของแบบจำลอง โดยที่แบบจำลองจะนำข้อมูลที่ถูกเลือกในขั้นตอนที่สอง มาใช้เรียนรู้เพิ่มเติม โดยที่มีพารามิเตอร์ที่สำคัญดังนี้ U_n คือ ข้อมูลไม่มั่นใจที่ถูกเลือกในแต่ละรอบ n โดยที่ขนาดของข้อมูลที่ถูกเลือกต้องมีมากกว่าศูนย์ n คือ จำนวนรอบที่ใช้ในการเรียนรู้เพิ่มเติม ซึ่งการเรียนรู้เพิ่มเติมจะต้องไม่เกินจากค่า n ที่กำหนดไว้ β คือ งบประมาณการสอบถามสำหรับการเรียนรู้เพิ่มเติมหรือขนาดข้อมูลสำหรับการเรียนรู้เพิ่มเติม แต่ละรอบในการเรียนรู้เพิ่มเติมแบบจำลองจะถูกวัดประสิทธิภาพค่าเฉลี่ยไมโครด้วยชุดตรวจสอบ โดยจะเลือกแบบจำลองที่มีค่าเฉลี่ยไมโครที่สูงที่สุดที่วัดด้วยชุดตรวจสอบ หลังจากนั้นจะทำการวัดประสิทธิภาพแบบจำลองด้วยชุดทดสอบ โดยสามารถแสดงขั้นตอนการจำแนกข้อมูลได้ดังรูปที่ 3.3

กลยุทธ์การเรียนรู้เชิงรุกไม่เอนเอียงแบบมีสัดส่วน (UAL-P)

L ← Set of Labeled data

L_i ← Set of Labeled data at iteration i ($0 \leq i \leq 100$)

β ← Budget of queries

P ← Percentage proportion of minority class

n ← Number of iteration ($n = 100$)

X_i = Set of selected data from L_i $(0 \leq i \leq 100)$, $i \leftarrow 0$

$Model_{opt} \leftarrow f(X_i, P, \beta, n)$

$L_i \leftarrow L$

$X_i \leftarrow L_i$ (Using high probability data and balancing classes with P parameter)

$Model_i \leftarrow \text{Train}\ (X_i)$

$L_i \leftarrow L - X_i$ (To eliminate majority class of X_i in L)

$i \leftarrow i + 1$

while $(X_i \leq \beta\ and\ i \leq n)$ do

  $Model_{i-1} \leftarrow \text{Train}\ (L_{i-1})$

  $X_i \leftarrow$ misclassify data of L_{i-1} (Balancing classes)

  $Model_i \leftarrow$ Incremental train (X_i) , Validated $Model_i$ by G-mean

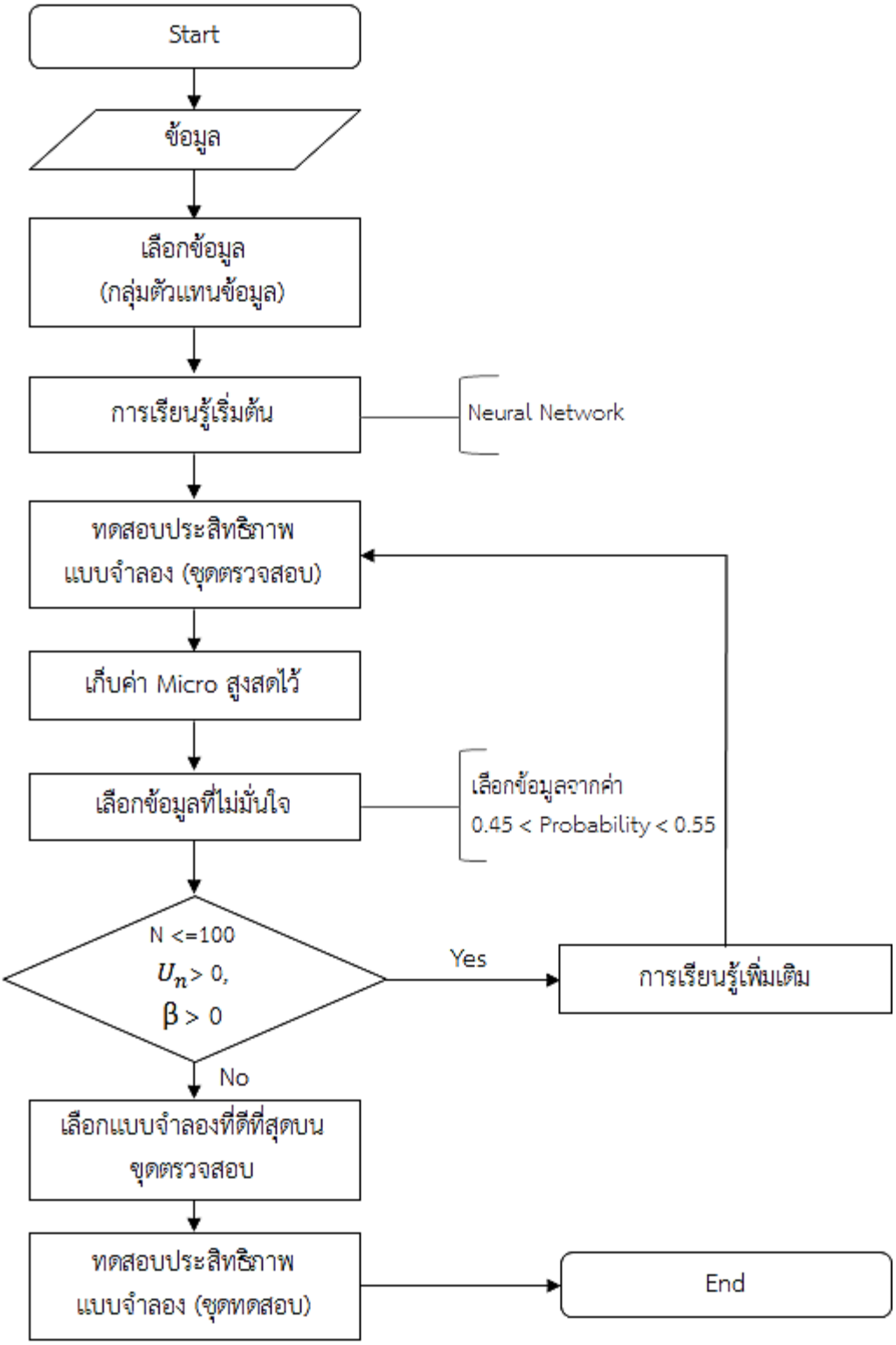
  $L_i \leftarrow L_{i-1} - X_i$ (Specific to eliminate majority class of X_i in L)

  $i \leftarrow i + 1$

end while

return $Model_{opt}$

รูปที่ 3.2 รหัสเทียมของกลยุทธ์การเรียนรู้เชิงรุกไม่เอนเอียงแบบมีสัดส่วน



รูปที่ 3.3 ขั้นตอนการจำแนกข้อมูลขนาดใหญ่และไม่สมดุลหลายฉลากด้วยนิวรอลเน็ตเวิร์ก

3.2 รูปแบบที่ 2 เสนอวิธีการจำแนกข้อมูลขนาดใหญ่และไม่สมดุลแบบหลายฉลากด้วยซัพพอร์ตเวกเตอร์แมชชีน

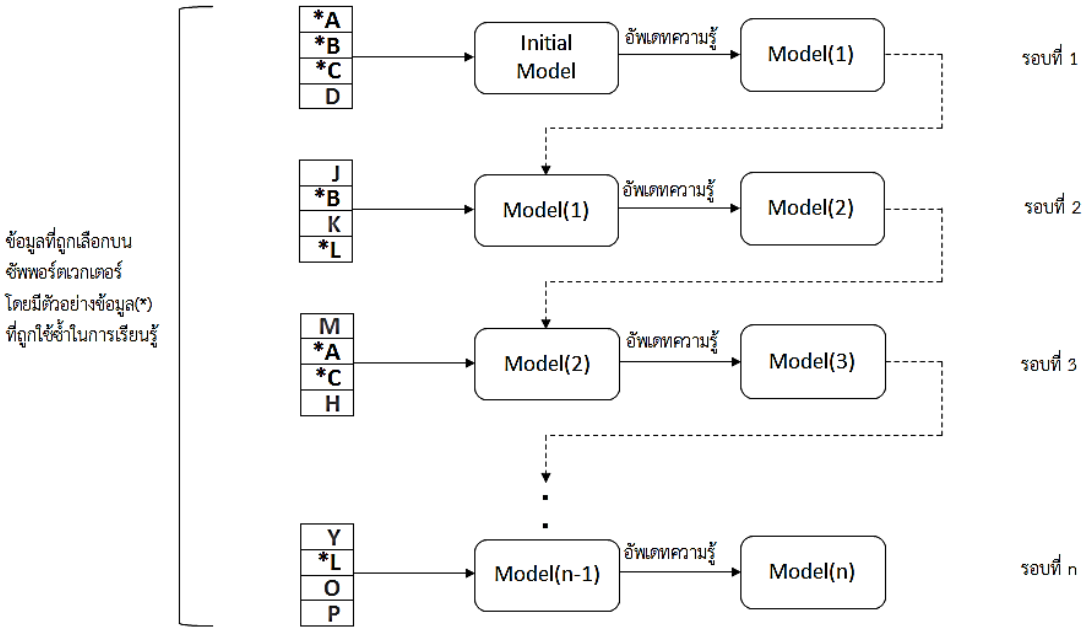
จากผลการวัดประสิทธิภาพของแบบจำลองการเรียนรู้เชิงรุกบนชุดข้อมูลแบบหลายฉลากที่สร้างด้วยนิวรอลเน็ตเวิร์คจะมีค่าน้อยกว่าแบบจำลองที่สร้างด้วยซัพพอร์ตเวกเตอร์แมชชีนค่อนข้างมาก ดังนั้นจึงเป็นที่มาของการนำเสนอรูปแบบที่ 2 หัวข้อนี้เป็นนำเสนอกลยุทธ์เพิ่มเติมในการจำแนกข้อมูลขนาดใหญ่และไม่สมดุลหลายฉลากด้วยซัพพอร์ตเวกเตอร์แมชชีนซึ่งเป็นที่นิยมและมีประสิทธิภาพในการจำแนกข้อมูล ซึ่งจะนำตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีนมาใช้ในการเรียนรู้เชิงรุก โดยใช้วิธี One-Versus-All เพื่อให้การจำแนกข้อมูลหลายฉลากมีประสิทธิภาพ สำหรับการใช้ซัพพอร์ตเวกเตอร์ที่เป็นแบบจำลองเส้นตรง (Linear Model) ทั่วไปไม่สามารถที่จะทำการเรียนรู้เพิ่มเติมได้ ดังนั้นในงานวิจัยชิ้นนี้ได้นำ สโตแคสติกเกรเดียนเดสเซนท์ (Stochastic Gradient Descent หรือ SGD) แบบจำลองเส้นตรง สร้างแบบจำลองการเรียนรู้เริ่มต้น และการเรียนรู้เพิ่มเติม สามารถแบ่งเป็นกลยุทธ์เพิ่มเติมสำหรับการเรียนรู้เชิงรุกได้ 2 แบบ คือ การเรียนรู้เชิงรุกเอสวีเอ็มเลือกข้อมูลไม่มั่นใจบนซัพพอร์ตเวกเตอร์แบบใช้ข้อมูลซ้ำ (AL-SVM-SV-R) และ การเรียนรู้เชิงรุกเอสวีเอ็มเลือกข้อมูลไม่มั่นใจบนซัพพอร์ตเวกเตอร์แบบไม่ใช้ข้อมูลซ้ำ(AL-SVM-SV-N)

3.2.1 การเรียนรู้เชิงรุกเอสวีเอ็มเลือกข้อมูลไม่มั่นใจบนซัพพอร์ตเวกเตอร์แบบใช้ข้อมูลซ้ำ (AL-SVM-SV-R)

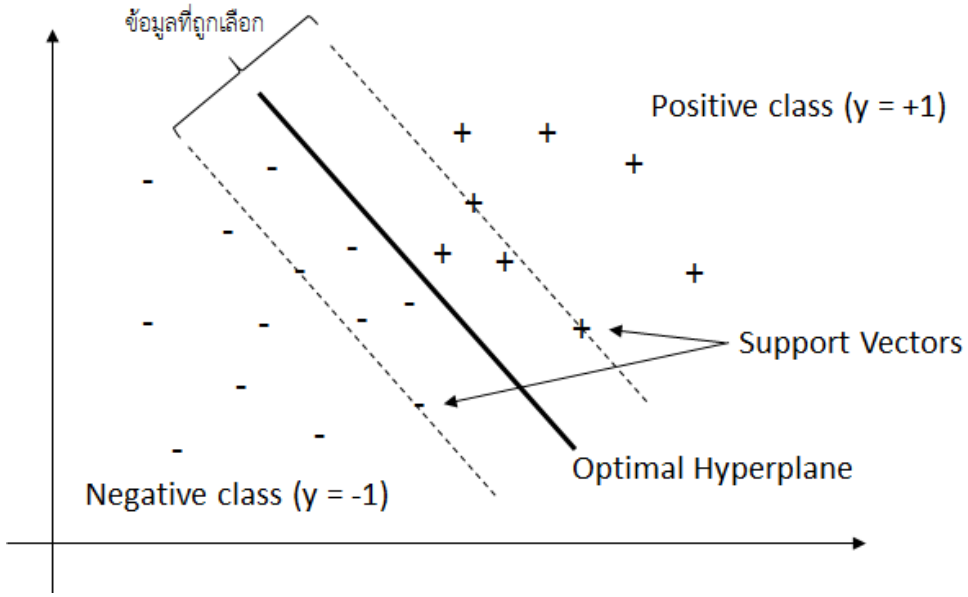
การเรียนรู้เชิงรุกเอสวีเอ็มเลือกข้อมูลไม่มั่นใจบนซัพพอร์ตเวกเตอร์แบบใช้ข้อมูลซ้ำ เป็นการเรียนรู้เชิงรุกที่สามารถเลือกข้อมูลที่เคยเรียนรู้ในอดีตแล้วมาใช้ในการเรียนรู้ซ้ำใหม่ได้ โดยข้อมูลการเรียนรู้ซ้ำจะถูกใช้ในการเรียนรู้เพิ่มเติมและถูกจำกัดให้อยู่บนซัพพอร์ตเวกเตอร์ คือ จากบนเส้นซัพพอร์ตเวกเตอร์ข้างลบถึงบนเส้นซัพพอร์ตเวกเตอร์ข้างบวก กำหนดให้ค่าซัพพอร์ตเวกเตอร์ข้างลบเท่ากับ -1 และค่าซัพพอร์ตเวกเตอร์ข้างบวกเท่ากับ +1 การเลือกข้อมูลดังกล่าวในระนาบจะเป็นบริเวณที่แบบจำลองที่สร้างด้วยซัพพอร์ตเวกเตอร์แมชชีนจะไม่สามารถจำแนกตัวอย่างข้อมูลได้อย่างถูกต้อง โดยการค้นหาข้อมูลที่อยู่บนซัพพอร์ตในแต่ละรอบของการเรียนรู้จะใช้แบบจำลองในการทำนายชุดข้อมูลฝึกสอน แล้วทำการเลือกข้อมูลที่อยู่บนซัพพอร์ตเวกเตอร์ทั้งหมด มาใช้ในการเรียนรู้เพิ่ม สามารถดูตัวอย่างการใช้ข้อมูลซ้ำในการเรียนรู้เพิ่มเติมของการเรียนรู้เชิงรุกได้ดังรูปที่ 3.4 และการเลือกข้อมูลไม่มั่นใจบนซัพพอร์ตเวกเตอร์ดังรูปที่ 3.5

อ้างอิงจาก รูปที่ 3.1 ขั้นตอนที่หนึ่ง การสร้างแบบจำลองเริ่มต้น การเรียนรู้เชิงรุกจะทำการหาตัวแทนของกลุ่มข้อมูลของแต่ละคลาสโดยใช้การวัดความคล้ายคลึงเชิงมุมโคไซน์ของข้อมูลมาใช้ในหาข้อมูลเพื่อการสร้างแบบจำลองเริ่มต้นด้วย SGD แบบจำลองเส้นตรงและวิธีการ One-Versus-All ซึ่งจะทำให้เกิดปัญหาไม่สมดุลด้วยวิธีการนี้ เนื่องจากต้องสร้างแบบจำลองตามจำนวนของคลาส แบบ

ตัวอย่างข้อมูลหนึ่งคลาสต่อตัวอย่างข้อมูลของจำนวนคลาสทั้งหมด คอสท์เซนซิทีฟจะถูกนำมาใช้แก้ไขปัญหาไม่สมดุลระหว่างคลาส



รูปที่ 3.4 ตัวอย่างการเรียนรู้เชิงรุกเอสวีเอ็มเลือกข้อมูลไม่มั่นใจบนซัพพอร์ตเวกเตอร์แบบใช้ข้อมูลซ้ำ



รูปที่ 3.5 การเลือกข้อมูลที่ไม่มั่นใจบนเส้นซัพพอร์ตเวกเตอร์ข้างลบถึงเส้นซัพพอร์ตเวกเตอร์ข้างบวก

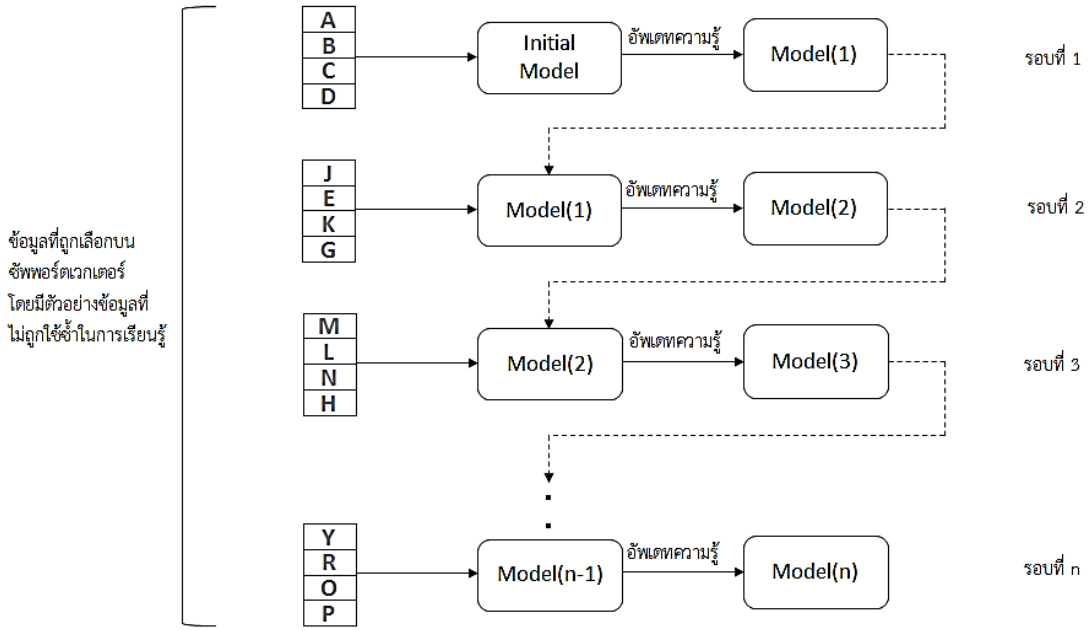
ขั้นตอนที่สอง การเลือกข้อมูลสำหรับการเรียนรู้เชิงรุก โดยขั้นตอนนี้จะเป็นการเลือกข้อมูลโดยใช้แบบจำลองมาทำนาย แล้วเลือกข้อมูลที่อยู่บนเส้นซัพพอร์ตเวกเตอร์จำนวนทั้งหมด ซึ่งข้อมูลที่ถูกเลือกสามารถเป็นข้อมูลที่เคยถูกเรียนรู้ สามารถนำมาใช้ซ้ำในการเรียนรู้เพิ่มเติมใหม่ได้

ขั้นตอนที่สาม การเรียนรู้เพิ่มเติมของแบบจำลอง โดยข้อมูลที่ถูกเลือกในขั้นตอนที่สองจะถูกนำมาใช้เรียนรู้เพิ่มเติม กำหนดให้ n คือจำนวนรอบของการเรียนรู้เพิ่มเติม U_n คือจำนวนข้อมูลที่ไม่มั่นใจแต่ละรอบ (n) ของการเรียนรู้เพิ่มเติมที่อยู่บนซัพพอร์ตเวกเตอร์ข้างลบถึงซัพพอร์ตเวกเตอร์ข้างบวก งบประมาณของการสอบถาม คือ ขนาดของข้อมูลที่ใช้ในการเรียนรู้เพิ่มเติม (ในหัวข้อนี้กำหนดให้ขนาดของการเรียนรู้เชิงรุกมีขนาดน้อยกว่าหรือเท่ากับ 40 เปอร์เซ็นต์ของขนาดข้อมูลฝึกสอน) การเรียนรู้เพิ่มเติมของแบบจำลองจะหยุดเมื่อจำนวนรอบการเรียนรู้เกินค่า n ที่กำหนดไว้ หรือจำนวนข้อมูลของแบบจำลองที่ใช้ในการเรียนรู้เริ่มต้นและเรียนรู้เพิ่มเติมเกินกว่าค่าที่กำหนดไว้ หรือ U_n เป็นศูนย์ และให้แต่ละรอบของการเรียนรู้ทดสอบด้วยชุดตรวจสอบเพื่อวัดประสิทธิภาพ F1 เมื่อแบบจำลองหยุดการเรียนรู้เพิ่มเติม แบบจำลองที่ดีที่สุดบนชุดตรวจสอบจะถูกเลือกและจะถูกทดสอบด้วยชุดทดสอบเพื่อวัดประสิทธิภาพแบบค่าเฉลี่ยไมโคร และค่าเฉลี่ยแมโคร และจบการทำงาน สามารถดูขั้นตอนการจำแนกข้อมูล ได้ดังรูปที่ 3.8

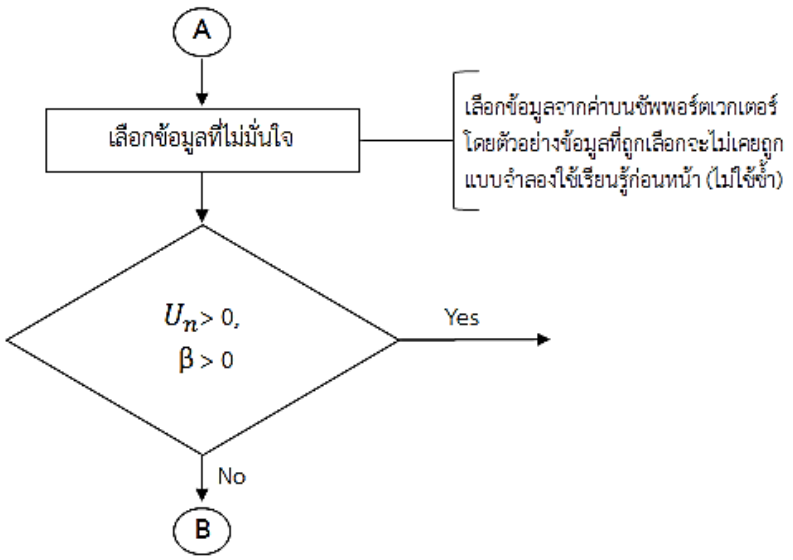
3.2.2 การเรียนรู้เชิงรุกเอสวีเอ็มเลือกข้อมูลไม่มั่นใจบนซัพพอร์ตเวกเตอร์แบบไม่ใช้ข้อมูลซ้ำ (AL-SVM-SV-N)

การเรียนรู้เชิงรุกเอสวีเอ็มเลือกข้อมูลไม่มั่นใจบนซัพพอร์ตเวกเตอร์แบบไม่ใช้ข้อมูลซ้ำ เป็นการเลือกข้อมูลที่ไม่เคยเรียนรู้หรือไม่เคยถูกใช้ซ้ำ มาใช้ในการสร้างแบบจำลองเริ่มต้นและการเรียนรู้เพิ่มเติม อ้างอิงจากรูปที่ 3.1 ขั้นตอนที่หนึ่ง การสร้างแบบจำลองเริ่มต้นจะทำการค้นหากลุ่มตัวแทนข้อมูลของแต่ละคลาสโดยใช้การวัดความคล้ายคลึงเชิงมุมโคไซน์ของข้อมูลมาใช้ในหาข้อมูลเพื่อการสร้างแบบจำลองเริ่มต้นด้วย SGD แบบจำลองเส้นตรงและวิธีการ One-Versus-All และทำการแก้ปัญหาไม่สมดุลด้วยเทคนิคคอสท์เซนซิทีฟ ขั้นตอนที่สองการเลือกข้อมูลสำหรับการเรียนรู้เชิงรุกซึ่งจะใช้แบบจำลองเลือกข้อมูลที่อยู่บนเส้นซัพพอร์ตเวกเตอร์ทั้งหมด สามารถดูได้ดังรูปที่ 3.5 โดยที่ข้อมูลที่เคยถูกใช้ในอดีตจะไม่ถูกเลือก ซึ่งเป็นจุดที่ต่างกับหัวข้อ 3.2.1 ขั้นตอนที่สามการเรียนรู้เพิ่มเติมของแบบจำลอง ข้อมูลที่ถูกเลือกแล้วจะถูกนำไปเรียนรู้เพิ่มเติมในแบบจำลอง ซึ่งในแต่ละรอบของการเรียนรู้เพิ่มเติมจะมีการวัดประสิทธิภาพ F1 ด้วยชุดข้อมูลตรวจสอบ กำหนดให้ n คือจำนวนรอบของการเรียนรู้เพิ่มเติม U_n คือจำนวนข้อมูลที่ไม่มั่นใจแต่ละรอบ การเรียนรู้เพิ่มเติมทำจนกระทั่งไม่พบข้อมูลบนซัพพอร์ตเวกเตอร์ หรือขนาดข้อมูลที่ใช้ในการเรียนรู้เริ่มต้นและการเรียนรู้เพิ่มเติมของแบบจำลองมีขนาดจำนวนน้อยกว่าหรือเท่ากับที่กำหนดไว้คือ 40 เปอร์เซ็นต์ของขนาดข้อมูลฝึกสอน

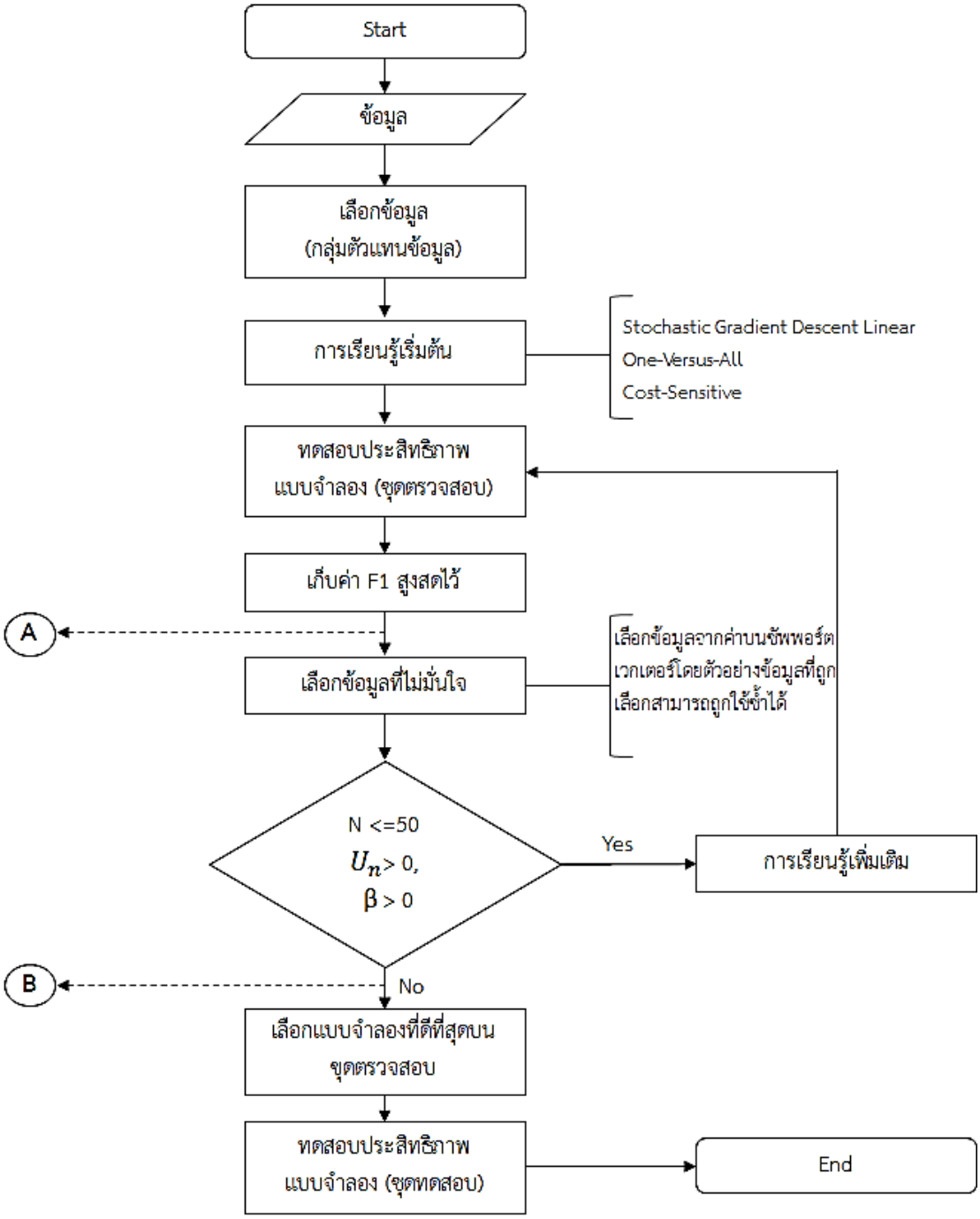
หลังจากนั้นแบบจำลองที่ดีที่สุดที่ทดสอบบนชุดตรวจสอบจะถูกเลือก และวัดประสิทธิภาพด้วยชุดทดสอบและจบการทำงาน สามารถดูตัวอย่างการเรียนรู้เชิงรุกเอสวีเอ็มเลือกข้อมูลไม่มั่นใจบนซัพพอร์ตเวกเตอร์แบบไม่ใช้ข้อมูลซ้ำได้ดังรูปที่ 3.6 และดูขั้นตอนการจำแนกข้อมูลได้ดังรูปที่ 3.7



รูปที่ 3.6 ตัวอย่างการเรียนรู้เชิงรุกเอสวีเอ็มเลือกข้อมูลไม่มั่นใจบนซัพพอร์ตเวกเตอร์แบบไม่ใช้ข้อมูลซ้ำ



รูปที่ 3.7 การเรียนรู้เชิงรุกเอสวีเอ็มเลือกข้อมูลไม่มั่นใจบนซัพพอร์ตเวกเตอร์แบบไม่ใช้ข้อมูลซ้ำ (AL-SVM-SV-N) โดยที่มีจุดเชื่อมไปยังรูปที่ 3.8



รูปที่ 3.8 การเรียนรู้เชิงรุกเอสวีเอ็มเลือกข้อมูลไม่มั่นใจบนซัพพอร์ตเวกเตอร์แบบใช้ข้อมูลซ้ำ (AL-SVM-SV-R)

บทที่ 4
การทดลองและผลการทดลอง

จากแนวคิดและวิธีการในบทที่ 3 ได้แบ่งการทดลองเป็นจำนวน 2 ส่วน เพื่อแสดงให้เห็นการใช้การเรียนรู้เชิงรุกเปรียบเทียบกับการเรียนรู้เชิงรับ บนข้อมูลขนาดใหญ่และไม่สมดุล สามารถแบ่งได้ดังนี้ 1) การทดลองเปรียบเทียบการจำแนกของการเรียนรู้เชิงรุกกับการเรียนรู้เชิงรับด้วยตัวจำแนกนิวรอลเน็ตเวิร์ค 2) การทดลองเปรียบเทียบการจำแนกของการเรียนรู้เชิงรุกกับการเรียนรู้เชิงรับด้วยตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีน

4.1 ระบบที่ใช้ในการทดลอง

คอมพิวเตอร์ที่ใช้ทำการทดลอง มีหน่วยประมวลผลกลาง Intel® Xeon CPU E3-1225 v3 ความเร็ว 3.2 GHz มีหน่วยความจำ 20 GB ระบบปฏิบัติการณ์ Windows 11 64 bits ภาษาโปรแกรมที่ใช้ Python 2.7 โดยใช้ไลบรารี Scikit Learn

4.2 ข้อมูลที่ใช้ในการทดลอง

ชุดข้อมูลที่ใช้ในการทดลองจะมี 2 ประเภท ได้แก่ ชุดข้อมูลแบบสองคลาส (Binary Class Data Sets) ชุดข้อมูลแบบหลายฉลาก (Multi-Label Data Sets)

4.2.1 ชุดข้อมูลแบบสองคลาส (Binary Class Data Sets)

งานวิจัยนี้มุ่งเน้นไปยังชุดข้อมูลขนาดใหญ่และไม่สมดุล ชุดข้อมูลแบบสองคลาสมาจาก “IJCNN 2007 Workshop on Agnostic Learning vs. Prior Knowledge” [16] และ “KDD CUP” [17] โดยที่ชุดข้อมูลแบบสองคลาสจะถูกแบ่งข้อมูลเป็น 3 ชุด ได้แก่ ชุดฝึกสอน ชุดตรวจสอบ และชุดทดสอบ |D| เป็นจำนวนตัวอย่างข้อมูล แสดงรายละเอียดได้ดังตารางที่ 4.1

ตารางที่ 4.1 ลักษณะเฉพาะของชุดข้อมูลแบบสองคลาสที่ใช้ในการทดลอง

Data Sets	|D|	Positive	Features	Classes	Imbalanced Ratio (Positive/Negative)
ADA	4,146	1,029	49	2	0.33
HIVA	3,844	135	1,617	2	0.03

Data Sets	|D|	Positive	Features	Classes	Imbalanced Ratio (Positive/Negative)
Protein Homology	145,751	1,296	78	2	0.01

4.2.2 ชุดข้อมูลแบบหลายฉลาก (Multi-Label Data Sets)

ชุดข้อมูลแบบหลายฉลากที่ใช้ในการทดลองจะนำมาจาก [18, 19] ซึ่งข้อมูลแบบหลายฉลากจะมีความซับซ้อนและยากต่อการจำแนกข้อมูลมากกว่าชุดข้อมูลแบบสองคลาส โดยที่ข้อมูลหนึ่งตัวอย่างสามารถเป็นได้มากกว่าหนึ่งคลาส กำหนดให้ชุดข้อมูลแบบหลายฉลากจะถูกแบ่งข้อมูลเป็น 3 ชุด ได้แก่ ชุดฝึกสอน ชุดตรวจสอบ และชุดทดสอบ |D| เป็นจำนวนตัวอย่างข้อมูล รายละเอียดดังตารางที่ 4.2 และสามารถดูรายละเอียดได้เพิ่มเติมที่ภาคผนวก ก และภาคผนวก ข

ตารางที่ 4.2 ลักษณะเฉพาะของชุดข้อมูลแบบหลายฉลากที่ใช้ในการทดลอง

Data Sets	|D|	Features	Classes
NUS-WIDE	269,648	128	81
RCV1V2	804,414	47,236	101

4.3 การทดลองเปรียบเทียบการจำแนกของการเรียนรู้เชิงรุกกับการเรียนรู้เชิงรับด้วยตัวจำแนกนิวรอลเน็ตเวิร์ก

สำหรับการทดลองนี้จะแบ่งการทดลองเป็นการเปรียบเทียบการจำแนกของการเรียนรู้เชิงรุกกับการเรียนรู้เชิงรับบนชุดของสองคลาสในหัวข้อ 4.3.1 4.3.2 และ 4.3.3 โดยใช้ข้อมูลสองคลาสจากตารางที่ 4.1 สำหรับในการทดลอง ส่วนหัวข้อ 4.3.4 จะเป็นการทดลองการจำแนกข้อมูลขนาดใหญ่และไม่สมดุลหลายฉลากด้วยนิวรอลเน็ตเวิร์ก ซึ่งจะใช้ชุดข้อมูลแบบหลายฉลากจากตารางที่ 4.2 สำหรับการทดลอง

4.3.1 การเปรียบเทียบเปอร์เซ็นต์ของจำนวนข้อมูลแบบสองคลาสในการสร้างแบบจำลองเริ่มต้น

จากผลการทดลองในตารางที่ 3 จะแสดงการเปรียบเทียบการใช้ข้อมูลสร้างแบบจำลองเริ่มต้นในสัดส่วนที่แตกต่างกันของคลาสฝ่ายข้างน้อย 15% 20% 25% 30% 35% และ 40% ซึ่งตัวจำแนกที่ใช้จะเป็น Artificial Neural Network (ANN) การวัดจะใช้ G-Mean และทำ 5 Fold Cross-Validation ผลการทดลองที่ได้ “UAL (100%)” จะได้ค่าสูงที่สุดบนชุดข้อมูล ADA HIVA และ Protein Homology อยู่ที่ 0.78 (+/- 0.03) 0.69 (+/- 0.08) และ 0.92 (+/- 0.01) ตามลำดับ ดังนั้นในหัวข้อที่ 4.3.2 การเปรียบเทียบประสิทธิภาพความแม่นยำระหว่างการเรียนรู้เชิงรุก (Active learning) และการเรียนรู้เชิงรับ (Passive learning) จะใช้ “UAL (100%)” เป็นตัวเปรียบเทียบ

ตารางที่ 4.3 แสดงการเปรียบเทียบเปอร์เซ็นต์ของจำนวนข้อมูลสำหรับการสร้างแบบจำลองเริ่มต้น

Active learning (ANN)	**Data Sets**		
	ADA	**HIVA**	**Protein Homology**
UAL (100%)	**0.78 (+/- 0.03) (100%)**	**0.69 (+/- 0.08) (100%)**	**0.92 (+/- 0.01) (40%)**
UAL-P (15%)	0.54 (+/- 0.08) (100%)	0.65 (+/- 0.08) (97%)	0.89 (+/- 0.03) (44%)
UAL-P (20%)	0.54 (+/- 0.08) (100%)	0.67 (+/- 0.10) (98%)	0.88 (+/- 0.02) (45%)
UAL-P (25%)	0.55 (+/- 0.06) (100%)	0.68 (+/- 0.07) (97%)	0.88 (+/- 0.01) (45%)
UAL-P (30%)	0.55 (+/- 0.11) (100%)	0.69 (+/- 0.10) (99%)	0.90 (+/- 0.02) (45%)
UAL-P (35%)	0.56 (+/- 0.08) (100%)	0.69 (+/- 0.09) (100%)	0.88 (+/- 0.03) (46%)
UAL-P (40%)	0.57 (+/- 0.13) (100%)	0.69 (+/- 0.03) (100%)	0.89 (+/- 0.02) (43%)

4.3.2 การเปรียบเทียบประสิทธิภาพความแม่นยำบนข้อมูลแบบสองคลาสระหว่างการเรียนรู้เชิงรุกและการเรียนรู้เชิงรับ

สำหรับการทดลองในหัวนี้จะเปรียบเทียบประสิทธิภาพของแบบจำลอง Artificial Neural Network หรือ ANN ระหว่างการเรียนรู้เชิงรุกและการเรียนรู้เชิงรับ ในตารางที่ 4.4 โดยใช้ค่าเริ่มต้นของแบบจำลองที่ดีที่สุดของการทดลองที่ 4.3.1 คือ “UAL (100%)” สำหรับ 3 ชุดข้อมูล จากตารางที่ 4.3 และใช้ Active learning (ANN) เปรียบเทียบกับ Passive learning ด้วยตัวจำแนกนาอีฟเบย์ (NB) เพื่อนบ้านใกล้ที่สุด K (kNN) ซัพพอร์ตเวกเตอร์แมชชีน (SVM) และต้นไม้ตัดสินใจ (DT) ในตารางที่ 4.5

จากผลการทดลองตารางที่ 4.4 จะแสดงผลการเปรียบเทียบประสิทธิภาพความแม่นยำระหว่างการเรียนรู้เชิงรุกและการเรียนรู้เชิงรับ ตัวจำแนกที่ใช้คือ ANN ใช้การวัด G-Mean และทำ 5 Fold Cross-Validation บนข้อมูลชุด ADA HIVA และ Protein Homology ซึ่งผลการทดลองที่ได้ Active learning จะมีประสิทธิภาพสูงกว่า Passive learning ประมาณ 8%-13% โดยจะนำวิธีการ Active learning – UAL (ANN) ในตาราง 4.4 ไปใช้อ้างอิงในผลการทดลองในตารางที่ 4.5

ตารางที่ 4.4 แสดงการเปรียบเทียบประสิทธิภาพความแม่นยำระหว่างการเรียนรู้เชิงรุกและการเรียนรู้เชิงรับ โดยการใช้ตัวจำแนก ANN และการวัด G-Mean แบบ 5 Fold Cross-Validation

Data Sets	**Passive learning (ANN)**	**Active learning - UAL (ANN)**
ADA	0.70 (+/- 0.05)	**0.78 (+/- 0.03) (100%)**
HIVA	0.56 (+/- 0.08)	**0.69 (+/- 0.08) (100%)**
Protein Homology	0.82 (+/- 0.05)	**0.92 (+/- 0.01) (40%)**

จากผลการทดลองในตารางที่ 4.5 จะแสดงผลการเปรียบเทียบประสิทธิภาพความแม่นยำระหว่างการเรียนรู้เชิงรุก (Active learning) ด้วยตัวจำแนก ANN และการเรียนรู้เชิงรับด้วยตัวจำแนก NB kNN SVM และ DT ซึ่งค่าการวัด G-Mean ของ Active Learn (ANN) จะมีค่าสูงกว่า Passive learning ด้วยตัวจำแนกทั้ง 4 แบบ บน 3 ขุดข้อมูลแบบสองคลาส

ตารางที่ 4.5 แสดงการเปรียบเทียบประสิทธิภาพความแม่นยำระหว่างการเรียนรู้เชิงรุกด้วยตัวจำแนก ANN และการเรียนรู้เชิงรับด้วยตัวจำแนก NB kNN SVM และ DT และการวัด G-Mean แบบ 5 Fold Cross-Validation

Data Sets	Active learning – UAL (ANN)	Passive learning			
		NB	kNN	SVM	DT
ADA	**0.78 (+/- 0.03) (100%)**	0.76 (+/- 0.02)	0.64 (+/- 0.04)	0.31 (+/- 0.05)	0.71 (+/- 0.04)
HIVA	**0.69 (+/- 0.08) (100%)**	0.63 (+/- 0.07)	0.50 (+/- 0.14)	0.56 (+/- 0.14)	0.54 (+/- 0.10)
Protein Homology	**0.92 (+/- 0.01) (40%)**	0.87 (+/- 0.02)	0.63 (+/- 0.05)	0.63 (+/- 0.05)	0.87 (+/- 0.03)

4.3.3 สรุปผลการทดลองการเปรียบเทียบการเรียนรู้เชิงรุกกับการเรียนรู้เชิงรับบนข้อมูลสองคลาสขนาดใหญ่และไม่สมดุล

จากผลการทดลองตารางที่ 4.3 4.4 และ 4.5 สามารถสรุปได้ดังนี้ สำหรับข้อมูลแบบสองคลาสจากตารางที่ 4.1 การเรียนรู้เชิงรุกด้วยตัวจำแนก ANN เมื่อวัดด้วย G-Mean แบบ 5-Fold Cross-Validation ค่าจะสูงกว่าการเรียนรู้เชิงรับด้วยตัวจำแนก ANN NB kNN SVM และ DT

Active learning – UAL (ANN) เริ่มจากการสร้างแบบจำลองเริ่มต้น (Initial Model) ด้วยวิธีการเลือกตัวอย่างลด (Undersampling) จากคลาสฝ่ายข้างมากสมดุลกับ ตัวอย่างข้อมูล 100% ของคลาส ฝ่ายข้างน้อย และใช้งบประมาณสำหรับการเรียนรู้เพิ่มเติมของแบบจำลอง และจำนวนในการทำซ้ำอยู่ที่ 100 รอบ สำหรับชุดข้อมูลแบบสองคลาส โดยที่ชุดข้อมูล Protein Homology ใช้ข้อมูลในการเรียนรู้ 40% ของปริมาณข้อมูลทั้งหมด ส่วนชุดข้อมูล ADA และ HIVA ใช้ข้อมูล 100% และทั้งสามชุดข้อมูลมีสิทธิภาพความแม่นยำที่สูงขึ้น โดยขั้นตอนวิธีจะเหมาะกับข้อมูลขนาดใหญ่ที่ไม่สมดุลมากกว่าข้อมูลขนาดเล็กที่ไม่สมดุล

4.3.4 การทดลองการจำแนกข้อมูลขนาดใหญ่และไม่สมดุลหลายฉลากด้วยนิวรอลเน็ตเวิร์ก

การทดลองนี้จะแสดงผลการจำแนกชุดข้อมูล NUS-WIDE จากตารางที่ 4.2 ซึ่งเป็นชุดข้อมูลแบบหลายฉลากด้วยการเรียนรู้เชิงรุกและการเรียนรู้เชิงรับที่สร้างด้วยตัวจำแนกนิวรอลเน็ตเวิร์กและการเรียนรู้เชิงรับที่สร้างด้วยตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีน เพื่อเปรียบเทียบประสิทธิภาพ ซึ่งข้อมูลหลายฉลากที่ใช้ในการทดลอง 4.3.4 จะถูกแบ่งเป็น 3 ส่วน ในปริมาณที่เท่ากันโดยการทำ Stratify Sampling เพื่อทำการทดสอบแบบ 3 Fold Cross-Validation และวัดประสิทธิภาพด้วยค่าเฉลี่ยไมโคร และค่าเฉลี่ยแมโคร โดยกำหนดพารามิเตอร์ต่างๆ ดังตารางที่ 4.6

ตารางที่ 4.6 แสดงค่าพารามิเตอร์ของนิวรอลเน็ตเวิร์กสำหรับการเรียนรู้เชิงรับและการเรียนรู้เชิงรุกบนชุดข้อมูล NUS-WIDE

ชื่อพารามิเตอร์	ค่าที่กำหนด
Activation Function	Sigmoid Function
Hidden Layer	2 ชั้น (450, 450)
Batch Size	1,000
Learning Rate	0.001
Iteration	3,000

ผลการทดลองการจำแนกชุดข้อมูล NUS-WIDE ด้วยนิวรอลเน็ตเวิร์ก โดยใช้พารามิเตอร์ดังตารางที่ 4.6 และเปรียบเทียบผลลัพธ์ค่าประสิทธิภาพระหว่างการเรียนรู้เชิงรับและการเรียนเชิงรุกซึ่งใช้วิธีการอ้างอิงจาก 3.1.2 สามารถแสดงผลลัพธ์การทดลองดังตารางที่ 4.7

ตารางที่ 4.7 แสดงผลการทดลองเปรียบเทียบประสิทธิภาพการจำแนกชุดข้อมูล NUS-WIDE ระหว่างการเรียนรู้เชิงรับด้วยนิวรอลเน็ตเวิร์ก (PL-ANN) การเรียนรู้เชิงรุกด้วยนิวรอลเน็ตเวิร์ก (AL-ANN) และการเรียนรู้เชิงรับด้วยซัพพอร์ตเวกเตอร์แมชชีน และมีการทำ 3 Fold Cross-Validation วัดด้วยค่าเฉลี่ยไมโครและค่าเฉลี่ยแมโคร

ชื่อเทคนิค	ค่าเฉลี่ยไมโคร	ค่าเฉลี่ยแมโคร	ขนาดข้อมูลที่ใช้	เวลาที่ใช้
PL-ANN	0.3265 (+/- 0.0042)	0.1294 (+/- 0.0047)	100.00%	14 ชั่วโมง 30 นาที
AL-ANN	0.3331 (+/- 0.0007)	0.1313 (+/- 0.0024)	39.27%	10 ชั่วโมง 26 นาที

ชื่อเทคนิค	ค่าเฉลี่ยไมโคร	ค่าเฉลี่ยแมโคร	ขนาดข้อมูลที่ใช้	เวลาที่ใช้
PL-SVM	**0.3698 (+/- 0.0007)**	**0.1866 (+/- 0.0033)**	100.00%	5 ชั่วโมง

จากผลการทดลองตารางที่ 4.7 แสดงการเปรียบเทียบสิทธิภาพการจำแนกชุดข้อมูล NUS-WIDE ด้วยนิวรอลเน็ตเวิร์กและซัพพอร์ตเวกเตอร์แมชชีน ประสิทธิภาพค่าเฉลี่ยไมโครและแมโครของ AL-ANN จะสูงกว่า PL-ANN โดยที่ขนาดข้อมูลและเวลาที่ใช้จะลดลงประมาณ 60 เปอร์เซ็นต์ และ 4 ชั่วโมง ตามลำดับ อย่างไรก็ตามการจำแนกด้วยนิวรอลเน็ตเวิร์กเมื่อเทียบกับซัพพอร์ตเวกเตอร์แมชชีน (PL-SVM) ค่าเฉลี่ยไมโคร และค่าเฉลี่ยไมโคร มีค่าน้อยกว่าประมาณ 0.03673 และ 0.05532 ตามลำดับ และเวลาจะใช้มากกว่า 5 ชั่วโมง ดังนั้นจึงทำการทดลองกลยุทธ์เพิ่มเติมในหัวข้อ 3.2 เพื่อวัดประสิทธิภาพการจำแนกข้อมูลขนาดใหญ่และไม่สมดุลหลายฉลากด้วยซัพพอร์ตเวกเตอร์แมชชีนสำหรับการเรียนรู้เชิงรุก

4.4 การทดลองเปรียบเทียบการจำแนกของการเรียนรู้เชิงรุกกับการเรียนรู้เชิงรับด้วยตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีน

ข้อมูลหลายฉลากที่ใช้ในการทดลอง 4.4 จะถูกแบ่งเป็น 3 ส่วน ในปริมาณที่เท่ากันโดยการทำ Stratify Sampling เพื่อการทำการทดสอบแบบ 3 Fold Cross-Validation สำหรับการเรียนรู้เชิงรับ การทดสอบแต่ละ Fold จะแบ่งเป็น ชุดฝึกสอน 70 เปอร์เซ็นต์ และชุดทดสอบ 30 เปอร์เซ็นต์ การเรียนรู้เชิงรุก แต่ละ Fold จะแบ่งเป็น ชุดฝึกสอน 60 เปอร์เซ็นต์ ชุดตรวจสอบ 10 เปอร์เซ็นต์ และชุดทดสอบ 30 เปอร์เซ็นต์ โดยสามารถสรุปได้ดังตารางที่ 4.8

ตารางที่ 4.8 การแบ่งข้อมูลเพื่อวัดประสิทธิภาพแต่ละเทคนิคในเบื้องต้น
บนชุดข้อมูล NUS-WIDE และ RCV1V2

ชื่อเทคนิค	ขนาดชุดฝึกฝนโดยประมาณ	ขนาดชุดตรวจสอบโดยประมาณ	ขนาดชุดทดสอบโดยประมาณ
Passive learning	70%	-	30%
Active learning	60%	10%	30%

ในหัวข้อนี้จะทำการทดลองเพิ่มเติมสำหรับตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีน โดยการเปรียบเทียบประสิทธิภาพระหว่างการเรียนรู้เชิงรุกเทียบกับการเรียนรู้เชิงรับ โดยจะแบ่งออกเป็น 4 การทดสอบ ดังนี้

1) การทดลองเปรียบเทียบประสิทธิภาพการเรียนรู้เชิงรุกกับการเรียนรู้เชิงรับด้วยตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีนเบื้องต้นบนชุดข้อมูล NUS-WIDE

2) การเปรียบเทียบประสิทธิภาพของการเรียนรู้เชิงรุกทั้ง 2 แบบ AL-SVM-SV-R และAL-SVM-SV-N

3) การวัดประสิทธิภาพการเรียนรู้เชิงรุกด้วย 3 Fold Cross-Validation บนชุดข้อมูล NUS-WIDE

4) การวัดประสิทธิภาพการเรียนรู้เชิงรุกด้วย 3 Fold Cross-Validation บนชุดข้อมูล RCV1V2

โดยกำหนดค่าพารามิเตอร์ที่สำคัญของการใช้สโตแคสติกเกรเดียนเดสเซนท์แบบจำลองเส้นตรงสำหรับใช้ในการเรียนรู้เชิงรุกบนชุดข้อมูล NUS-Wide และ RCV1V2 ดังตารางที่ 4.9

ตารางที่ 4.9 แสดงค่าพารามิเตอร์ที่สำคัญของการใช้สโตแคสติกเกรเดียนเดสเซนท์แบบจำลองเส้นตรงในการเรียนรู้เริ่มต้นและเรียนรู้เพิ่มเติมของการเรียนรู้เชิงรุกบนชุดข้อมูล NUS-WIDE และ RCV1V2

ชื่อพารามิเตอร์	**ค่าที่กำหนดของ NUS-WIDE**	**ค่าที่กำหนดของ RCV1V2**
Loss Function	Logistic	Logistic
Learning Rate	0.001	0.001
Iteration	2,000	2200

4.4.1 การทดลองเปรียบเทียบประสิทธิภาพการเรียนรู้เชิงรุกกับการเรียนรู้เชิงรับด้วยตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีนเบื้องต้นบนชุดข้อมูล NUS-WIDE

การทดลองนี้ทำเพื่อวัตถุประสงค์ในการวัดประสิทธิภาพโดยรวมแต่ละวิธีการในเบื้องต้น การเรียนรู้เชิงรับข้อมูลจะมีชุดฝึกฝนและชุดทดสอบเท่านั้น ส่วนการเรียนรู้เชิงรุกจะแบ่งข้อมูลเป็น ชุดฝึกฝน ชุดตรวจสอบ และชุดทดสอบ ดังตารางที่ 4.8 สำหรับการทดลองนี้จะทำบนชุดข้อมูล NUS-WIDE

จากผลการทดลองที่ 4.10 แสดงการเปรียบเทียบการทดสอบการเรียนรู้เชิงรุกและการเรียนรู้เชิงรับด้วยตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีน โดยมีรายละเอียดเกี่ยวกับเทคนิคดังนี้

1) PL-SVM การเรียนรู้เชิงรับด้วยตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีน
2) PL-SVM Undersampling (90%) การเรียนรู้เชิงรับด้วยตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีน โดยการลดขนาดข้อมูลคลาสฝั่งข้างมากให้เหลือข้อมูล 90 เปอร์เซ็นต์ ซึ่งจะทำทุก Classifier ของการทำ One-Versus-All
3) PL-SVM Selective Representation data การเรียนรู้เชิงรับด้วยตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีน โดยการเลือกตัวแทนข้อมูลแล้วทำการสร้างแบบจำลองโดยใช้ Passive learning

4) AL-SVM-SV-R การเรียนรู้เชิงรุกด้วยเอสวีเอ็มเลือกข้อมูลไม่มั่นใจบนซัพพอร์ตเวกเตอร์แบบใช้ข้อมูลซ้ำ โดยอ้างอิงจากหัวข้อ 3.2.1
5) AL-SVM-SV-N การเรียนรู้เชิงรุกด้วยเอสวีเอ็มเลือกข้อมูลไม่มั่นใจบนซัพพอร์ตเวกเตอร์แบบไม่ใช้ข้อมูลซ้ำ โดยอ้างอิงจากหัวข้อ 3.2.2

การทดลองในตารางที่ 4.10 แสดงการเปรียบเทียบทั้ง 5 รูปแบบ โดยที่รูปแบบที่ 1 PL-SVM จะมีการทำ Cost Sensitive ดังรูปที่ 4.1 และ 4.2 ทำให้สามารถทำนายคลาสที่มีปริมาณน้อยได้ และมีค่าเฉลี่ยแมโครที่สูงที่สุด และใช้เวลามากที่สุดเมื่อเทียบ รูปแบบที่ 3 ที่ใช้เวลาน้อยที่สุด โดยมากกว่าประมาณ 1 เท่า ส่วนการเรียนรู้เชิงรุกในรูปแบบที่ 4 AL-SVM-SV-R และรูปแบบที่ 5 AL-SVM-SV-N ค่าเฉลี่ยไมโครจะสูงขึ้นเมื่อเทียบการเรียนรู้เชิงรับ

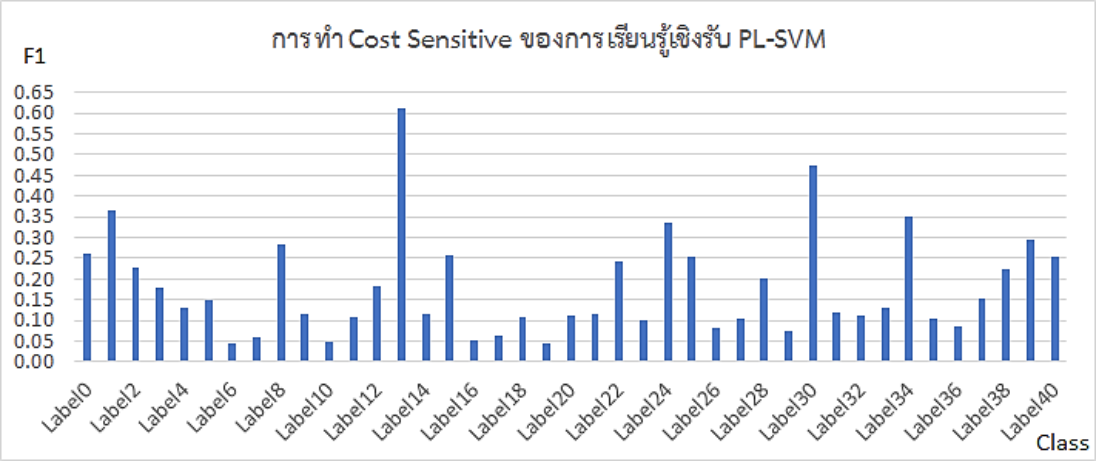
AL-SVM-SV-R ค่าเฉลี่ยไมโครจะสูงที่สุดเมื่อเปรียบเทียบกับการเรียนรู้เชิงรุกแบบที่เหลือและการเรียนรู้เชิงรับ โดยมากกว่าการเรียนรู้เชิงรับ แบบที่ 1 2 และ 3 คือ 0.02733 0.0311 และ 0.00929 ตามลำดับ ในขณะที่เวลาและขนาดข้อมูลที่ใช้ในการทดสอบลดลงไปประมาณ 20 นาที และ 65 เปอร์เซ็นต์ ตามลำดับ เมื่อเปรียบเทียบกับ PL-SVM

จากทดลองในหัวนี้แสดงว่า การเรียนรู้เชิงรุกทั้ง 2 แบบ คือ AL-SVM-SV-R และ AL-SVM-SV-N ได้เพิ่มประสิทธิภาพค่าเฉลี่ยไมโครให้สูงขึ้นเมื่อเทียบกับ PL-SVM อยู่ในช่วง 0.02208-0.02733 ซึ่งเวลาและขนาดข้อมูลที่ใช้ในการฝึกฝนแบบจำลองก็ลดลงเมื่อเปรียบเทียบกับการเรียนรู้เชิงรับ โดยที่ AL-SVM-SV-R มีค่าเฉลี่ยไมโครที่ดีที่สุด อย่างไรก็ตามค่าเฉลี่ยแมโครของการเรียนรู้เชิงรุกก็ลดลงไปจากเดิม

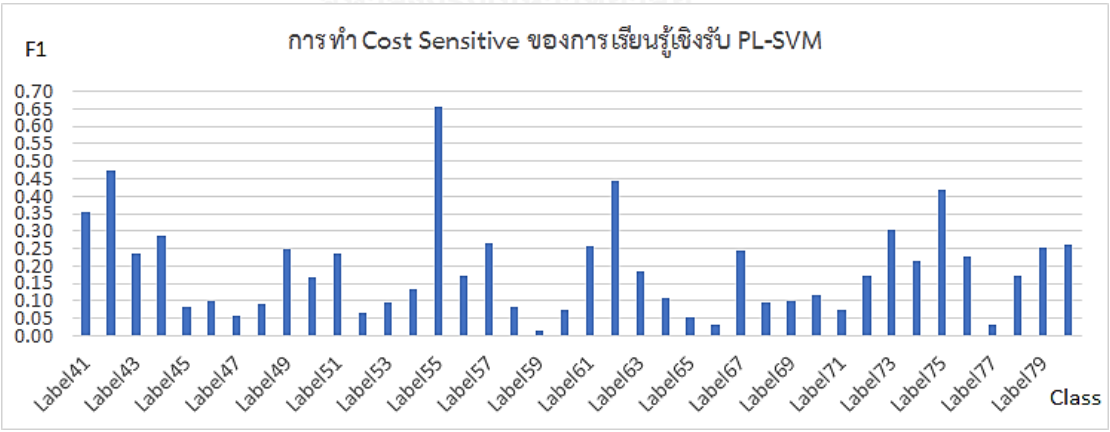
ตารางที่ 4.10 แสดงผลการทดลองเปรียบเทียบประสิทธิภาพเบื้องต้นการจำแนกชุดข้อมูล NUS-WIDE ระหว่างการเรียนรู้เชิงรับและการเรียนรู้เชิงรุกด้วยซัพพอร์ตเวกแมชชีน และเบื้องต้นแบ่งข้อมูลเป็นชุดฝึกสอน ชุดตรวจสอบและชุดทดสอบ โดยที่วัดด้วยค่าเฉลี่ยไมโครและค่าเฉลี่ยแมโคร

ชื่อเทคนิค	ค่าเฉลี่ยไมโคร	ค่าเฉลี่ยแมโคร	ขนาดข้อมูลที่ใช้	เวลาที่ใช้
PL-SVM	0.3695	**0.1852**	100.00%	1 ชั่วโมง 42 นาที
PL-SVM Undersampling (90%)	0.3657	0.1828	90.00%	1 ชั่วโมง 18 นาที
PL-SVM Selective Representation data	0.3875	0.1829	26.16%	52 นาที

ชื่อเทคนิค	ค่าเฉลี่ยไมโคร	ค่าเฉลี่ยแมโคร	ขนาดข้อมูลที่ใช้	เวลาที่ใช้
AL-SVM-SV-R	**0.3968**	0.1803	33.64%	1 ชั่วโมง 20 นาที
AL-SVM-SV-N	0.3915	0.1819	27.39%	59 นาที



รูปที่ 4.1 การทำคอสท์เซนซิทีฟ (Cost Sensitive) ของ PL-SVM เพื่อให้แบบจำลองสามารถทำนายคลาสที่มีข้อมูลปริมาณน้อยได้หรือแก้ไขปัญหาไม่สมดุลของคลาส (Label0-Label40)



รูปที่ 4.2 การทำคอสท์เซนซิทีฟ (Cost Sensitive) ของ PL-SVM เพื่อให้แบบจำลองสามารถทำนายคลาสที่มีข้อมูลปริมาณน้อยได้หรือแก้ไขปัญหาไม่สมดุลของคลาส (Label41-Label80)

4.4.2 การเปรียบเทียบประสิทธิภาพของการเรียนรู้เชิงรุกทั้ง 2 แบบ AL-SVM-SV-R และ AL-SVM-SV-N

จากผลการทดลองการเปรียบเทียบประสิทธิภาพของการเรียนรู้เชิงรุกทั้ง 2 แบบ อ้างอิงจากตารางที่ 4.10 โดยค่าเฉลี่ยไมโครของ AL-SVM-SV-R จะมีค่าที่สูงกว่า AL-SVM-SV-N โดยที่เมื่อเปรียบเทียบการเรียนรู้เชิงรุกแต่ละแบบเพิ่มเติมด้วยการวัดค่าประสิทธิภาพ F1 แยกแต่ละ Classifier ของการทำ One-Vs-All พบว่าภาพโดยรวมจากผลการทดลอง ดังตารางที่ 4.11 รูปที่ 4.3 4.4 และ 4.5 AL-SVM-SV-R จะทำได้ดีกว่า AL-SVM-SV-N เมื่อพิจารณาจากข้อมูลการเรียนรู้เริ่มต้นและการเรียนรู้เพิ่มเติมถึงจำนวนรอบทั้งหมดและรอบที่ดีที่สุดของการวัดค่าประสิทธิภาพ F1 บนชุดตรวจสอบของการเรียนรู้เชิงรุกทั้ง 2 รูปแบบ ดังรูปที่ 4.4 และ 4.5 แสดงการเปรียบเทียบจำนวนรอบของการเรียนรู้เพิ่มเติมของการเรียนรู้เชิงรุกทั้ง 2 รูปแบบ

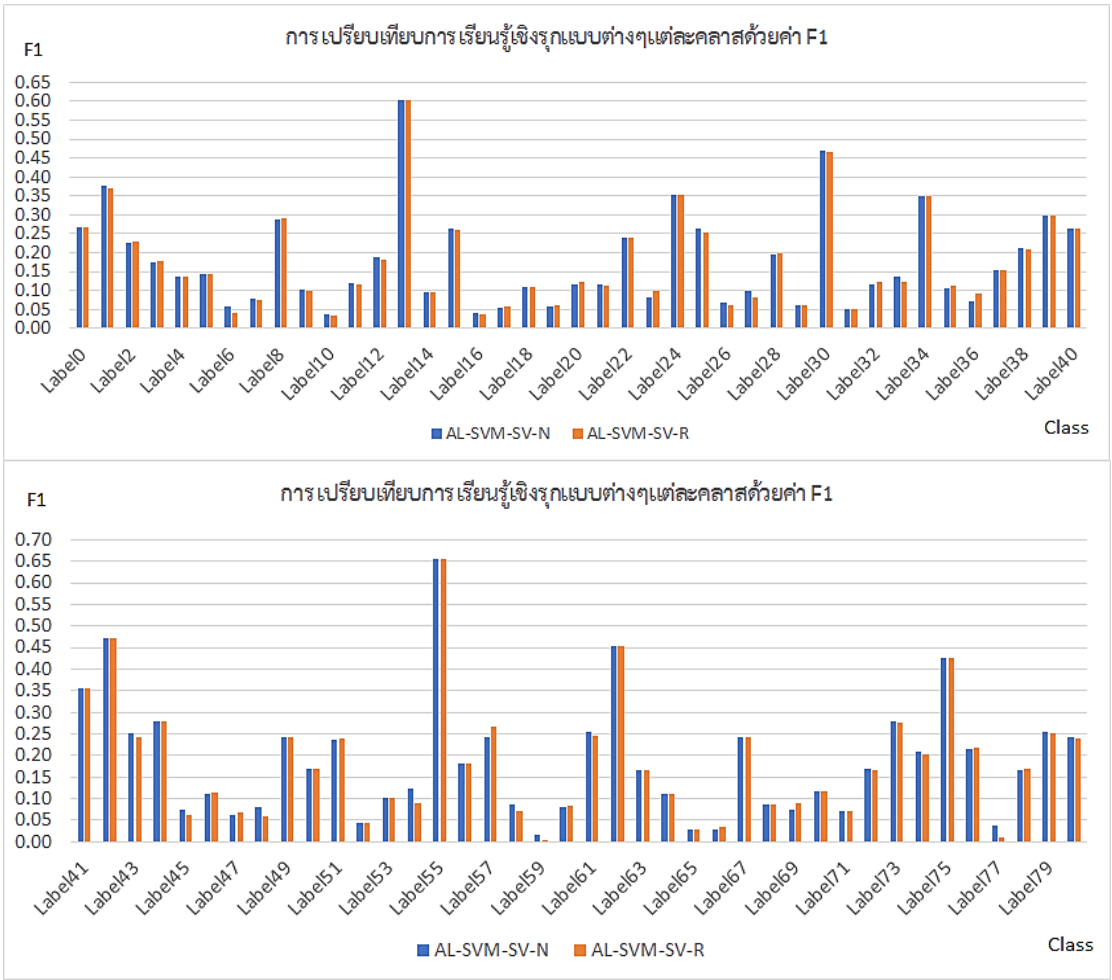
การเรียนรู้เชิงรุก AL-SVM-SV-R จะมีการเรียนรู้เพิ่มเติมที่มีจำนวนรอบการเรียนรู้ที่มากกว่า AL-SVM-SV-N ในแต่ละคลาส (ดูจากจำนวนรอบทั้งหมด และจำนวนรอบที่ดีที่สุดที่มีการเรียนรู้เพิ่มเติม) และรอบที่ดีที่สุดที่หยุดที่แบบจำลองเริ่มต้นของแต่ละคลาสจะมีจำนวนน้อยกว่า โดย AL-SVM-SV-R และ AL-SVM-SV-N จะมีรอบที่ดีที่สุดที่หยุดที่แบบจำลองเริ่มต้นจำนวน 30 คลาส และ 44 คลาส ตามลำดับ จากข้อมูลของผลการทดลองแสดงว่า AL-SVM-SV-R สามารถเรียนรู้เพิ่มเติมได้ดีกว่า AL-SVM-SV-N อย่างไรก็ตามเมื่อพิจารณาผลเปรียบเทียบดังตารางที่ 4.12 บนชุดข้อมูลทดสอบจากค่า F1 ของจำนวนคลาสชนะ AL-SVM-SV-N จะมีจำนวนมากกว่า AL-SVM-SV-R ถึง 11 คลาส

ตารางที่ 4.11 แสดงผลการทดลองเปรียบเทียบประสิทธิภาพการจำแนกชุดข้อมูล NUS-WIDE โดยเปรียบเทียบการเรียนรู้ของแบบจำลองบนชุดตรวจสอบของการเรียนรู้เชิงรุกทั้ง 2 แบบ

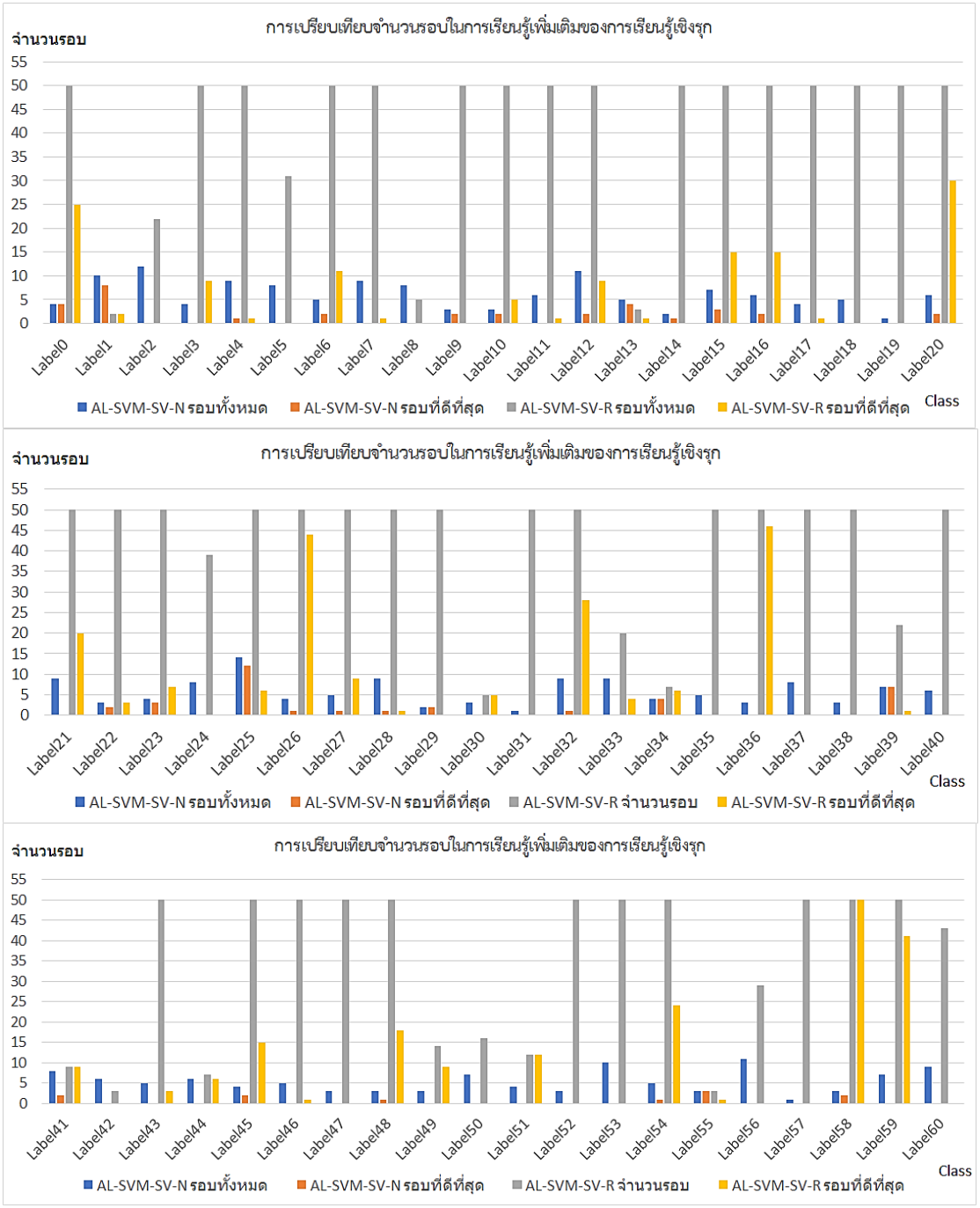
ชื่อเทคนิค	ค่า F1ของรอบที่ดีที่สุดเป็นแบบจำลองเริ่มต้น (จำนวนคลาส)	ค่า F1 ของรอบที่ดีที่สุดไม่ใช่แบบจำลองเริ่มต้น (จำนวนคลาส)
AL-SVM-SV-R	**30**	51
AL-SVM-SV-N	44	37

ตารางที่ 4.12 แสดงผลการทดลองเปรียบเทียบประสิทธิภาพการจำแนกชุดข้อมูล NUS-WIDE โดยเปรียบเทียบค่า F1 บนชุดทดสอบของทุกคลาส ของการเรียนรู้เชิงรุกทั้ง 2 แบบ

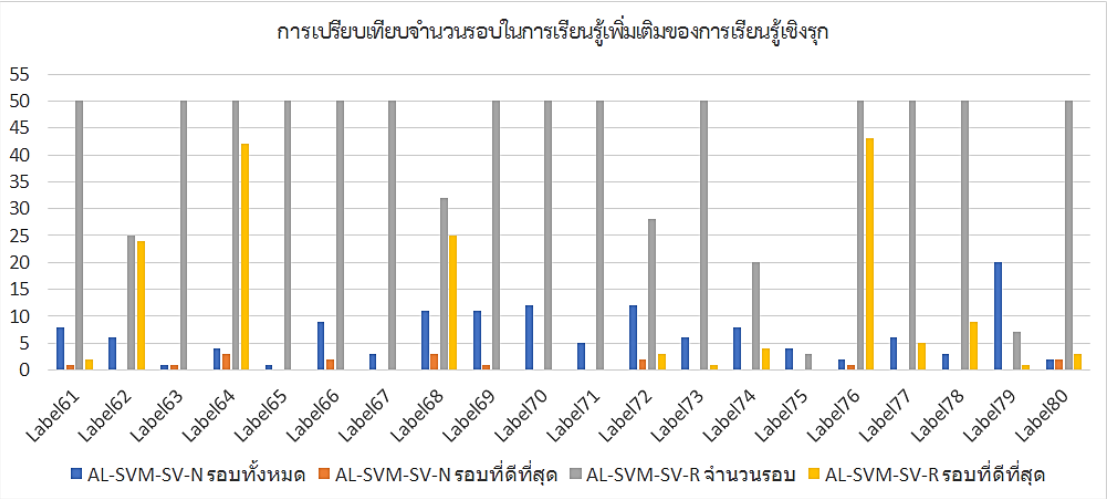
ชื่อเทคนิค	ค่าเฉลี่ย F1	ค่าต่ำสุด F1	ค่าสูงสุด F1	ค่าเบี่ยงเบนมาตรฐาน	จำนวนคลาสชนะ	จำนวนคลาสแพ้	จำนวนคลาสเสมอ
AL-SVM-SV-R	0.1803	0.0029	**0.6566**	**0.1327**	33	44	4
AL-SVM-SV-N	**0.1819**	0.0175	0.6558	0.1318	**44**	33	4



รูปที่ 4.3 แสดงการเปรียบเทียบค่าประสิทธิภาพ F1 แต่ละ Classifier ของการทำ One-Versus-All ของการเรียนรู้เชิงรุก 2 รูปแบบ (Label0-Label80)



รูปที่ 4.4 แสดงการเปรียบเทียบจำนวนรอบของการเรียนรู้เพิ่มเติมของการเรียนรู้เชิงรุกทั้ง 2 รูปแบบ (Label0-Label60)



รูปที่ 4.5 แสดงการเปรียบเทียบจำนวนรอบของการเรียนรู้เพิ่มเติมของการเรียนรู้เชิงรุกทั้ง 2 รูปแบบ (Label61-Label80)

4.4.3 การวัดประสิทธิภาพของการเรียนรู้เชิงรุกด้วย 3 Fold Cross-Validation บนชุดข้อมูล NUS-WIDE

การทดลองหัวข้อนี้จะนำเทคนิคที่วัดประสิทธิภาพได้ผลการทดลองจากหัวข้อ 4.4.1 อ้างอิงตารางที่ 4.10 ที่มีผลค่าเฉลี่ยไมโครและค่าเฉลี่ยแมโครดีที่สุด คือ AL-SVM-SV-R และ PL-SVM ตามลำดับ นอกเหนือจากนั้นได้เลือก AL-SVM-SV-N เพิ่มเติมเพื่อนำมาทำการทดลองในการวัดประสิทธิภาพการจำแนกข้อมูลแบบ 3 Fold Cross-Validation บนชุดข้อมูล NUS-WIDE และวัดด้วยค่าเฉลี่ยไมโครและค่าเฉลี่ยแมโคร ดังตารางที่ 4.11

จากผลการทดลองในหัวข้อนี้สามารถสรุปได้ว่า ประสิทธิภาพการวัดจากการทำ 3 Fold Cross-Validation บนชุดข้อมูล NUS-Wide ค่าเฉลี่ยไมโครของเทคนิค AL-SVM-SV-R มีค่าสูงสุดและมีค่ามากกว่า PL-SVM, AL-SVM-SV-N อยู่ประมาณ 0.02679 และ 0.00397 ตามลำดับ ค่าเฉลี่ยแมโครของเทคนิค PL-SVM มีค่าสูงสุดและมีค่ามากกว่า AL-SVM-SV-R และ AL-SVM-SV-N อยู่ประมาณ 0.00396 และ 0.0033 ตามลำดับ สำหรับประสิทธิภาพเรื่องจำนวนขนาดข้อมูลและเวลาที่ใช้สร้างแบบจำลองบนชุดข้อมูล NUS-WIDE การเรียนรู้เชิงรุกทั้ง 2 รูปแบบ ใช้ขนาดข้อมูลและเวลาน้อยกว่าการเรียนรู้เชิงรับ โดยขนาดข้อมูลใช้ลดลงประมาณ 65% และใช้เวลาลดลงประมาณ 1 ชั่วโมง

ตารางที่ 4.13 แสดงผลการทดลองเปรียบเทียบประสิทธิภาพการจำแนกชุดข้อมูล NUS-WIDE ระหว่างการเรียนรู้เชิงรับและการเรียนรู้เชิงรับด้วยซัพพอร์ตเวกเตอร์แมชชีน และมีการทำ 3 Fold Cross-Validation วัดด้วยค่าเฉลี่ยไมโครและค่าเฉลี่ยแมโคร

ชื่อเทคนิค	ค่าเฉลี่ยไมโคร	ค่าเฉลี่ยแมโคร	ขนาดข้อมูลที่ใช้	เวลาที่ใช้
PL-SVM	0.3698 (+/- 0.0007)	**0.1866 (+/- 0.0033)**	100.00%	5 ชั่วโมง
AL-SVM-SV-R	**0.3966 (+/- 0.0014)**	0.1827 (+/- 0.0041)	33.79%	3 ชั่วโมง 52 นาที
AL-SVM-SV-N	0.3926 (+/- 0.0013)	0.1833 (+/- 0.0040)	27.38%	2 ชั่วโมง 50 นาที

4.4.4 การวัดประสิทธิภาพการเรียนรู้เชิงรุกด้วย 3 Fold Cross-Validation บนชุดข้อมูล RCV1V2

ในการทดลองนี้จะนำผลจากการทดลองที่ 4.4.2 การเรียนรู้เชิงรุกกลยุทธ์ AL-SVM-SV-R ที่มีประสิทธิภาพในการเรียนรู้เพิ่มเติมจากชุดข้อมูลตรวจสอบ มาทำการทดลองเพิ่มเติมในชุดข้อมูล RCV1V2 จากผลการทดลองตารางที่ 4.12 การวัดประสิทธิภาพการเรียนรู้เชิงรุกด้วย 3 Fold Cross-Validation เมื่อเปรียบเทียบกับ PL-SVM ได้ผลดังนี้คือ ค่าเฉลี่ยไมโครและแมโครของกลยุทธ์ที่นำเสนอเพื่อเปรียบเทียบการเรียนรู้เชิงรับจะน้อยกว่าประมาณ 0.01804 และ 0.04798 ตามลำดับ และใช้เวลามากกว่าประมาณ 14 ชั่วโมง สาเหตุที่ PL-SVM ใช้เวลาน้อยกว่ากลยุทธ์ที่นำเสนอไปมาก เนื่องจากข้อมูลถูกจัดเก็บอยู่ในรูปแบบ Sparse Matrix ส่วนกลยุทธ์การเรียนรู้เชิงรุกที่นำเสนอต้องทำกระบวนการเลือกข้อมูลเพิ่มเติมเพื่อเรียนรู้ผ่านการจัดเก็บข้อมูลแบบอะเรย์ (array) โดยข้อมูลที่ใช้ในการเรียนรู้เชิงรุกจะน้อยกว่าการเรียนรู้เชิงรับอยู่ประมาณ 60 เปอร์เซ็นต์

ตารางที่ 4.14 แสดงผลการทดลองเปรียบเทียบประสิทธิภาพการจำแนกชุดข้อมูล RCV1V2 ระหว่างการเรียนรู้เชิงรุกและการเรียนรู้เชิงรับด้วยซัพพอร์ตเวกเตอร์แมชชีน และมีการทำ 3 Fold Cross-Validation วัดด้วยค่าเฉลี่ยไมโครและค่าเฉลี่ยแมโคร

ชื่อเทคนิค	ค่าเฉลี่ยไมโคร	ค่าเฉลี่ยแมโคร	ขนาดข้อมูลที่ใช้	เวลาที่ใช้
PL-SVM	**0.8181 (+/- 0.0005)**	**0.6999 (+/- 0.0031)**	100.00%	1 ชั่วโมง 30 นาที
AL-SVM-SV-R	0.8001 (+/- 0.0008)	0.6520 (+/- 0.0033)	32.13%	14 ชั่วโมง 40 นาที

บทที่ 5
สรุปการวิจัยและแนวทางการวิจัยในขั้นถัดไป

5.1 สรุปการวิจัย

วิทยานิพนธ์ชิ้นนี้ได้นำเสนอการจำแนกโดยใช้การเรียนรู้เชิงรุกบนชุดข้อมูลขนาดใหญ่ โดยที่ในปัจจุบันข้อมูลมีความซับซ้อนและมีความหลากหลายมากขึ้น ซึ่งข้อมูลหนึ่งตัวอย่างสามารถเป็นได้มากกว่าหนึ่งคลาส และข้อมูลมักจะมีขนาดใหญ่และเพิ่มขึ้นมหาศาล นอกจากนั้นข้อมูลแบบหลายฉลากมักจะมีความไม่สมดุลระหว่างคลาส จึงได้นำเสนอเทคนิคคอสท์เซนซิทีฟเพิ่มเติมมาช่วยแก้ไขปัญหาความไม่สมดุลของข้อมูลประเภทนี้ด้วย วิทยานิพนธ์ฉบับนี้มุ่งเน้นที่ข้อมูลแบบหลายฉลาก จึงนำเสนอการจำแนกแบบหลายฉลากโดยใช้การเรียนรู้เชิงรุกบนชุดข้อมูลขนาดใหญ่และไม่สมดุล เพื่อเปรียบเทียบประสิทธิภาพความแม่นยำกับการเรียนรู้เชิงรับ โดยจะใช้ตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีน และนิวรอลเน็ตเวิร์ก

จากการทดลองการเปรียบเทียบประสิทธิภาพของการเรียนรู้เชิงรุกและการเรียนรู้เชิงรับเมื่อสร้างด้วยตัวจำแนกนิวรอลเน็ตเวิร์ก สำหรับบนข้อมูลสองคลาสที่ใช้ในงานวิจัยชิ้นนี้เมื่อวัดประสิทธิภาพด้วยตัววัดจีมีน ผลจากการทดลองค่าจีมีนของการเรียนรู้เชิงรุกจะสูงกว่าการเรียนรู้เชิงรับ และสำหรับบนชุดข้อมูลหลายฉลาก เมื่อวัดประสิทธิภาพด้วยค่าเฉลี่ยไมโครและค่าเฉลี่ยแมโคร ผลจากการทดลอง การเรียนรู้เชิงรุกจะมีค่าเฉลี่ยไมโครและค่าเฉลี่ยแมโครที่สูงกว่าการเรียนรู้เชิงรับ และใช้ขนาดข้อมูลในการเรียนรู้และเวลาที่น้อยกว่า อย่างไรก็ตามเมื่อเปรียบเทียบระหว่างแบบจำลองที่สร้างด้วยตัวจำแนกนิวรอลเน็ตเวิร์กและตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีนจะพบว่า ประสิทธิภาพการวัดด้วยค่าเฉลี่ยไมโครและแมโครและเวลาของแบบจำลองที่สร้างด้วยตัวจำแนกนิวรอลเน็ตเวิร์กมีประสิทธิภาพที่ต่ำกว่าแบบจำลองที่สร้างด้วยตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีนค่อนข้างมาก

จากผลการทดลองในข้างต้นจึงเป็นที่มาของการทดลองเพิ่มเติมสำหรับการเรียนรู้เชิงรุกที่สร้างแบบจำลองด้วยซัพพอร์ตเวกเตอร์แมชชีน โดยบนชุดข้อมูลหลายฉลาก โดยที่ได้มีนำเสนอกลยุทธ์เพิ่มเติมสองกลยุทธ์ ได้แก่ การเลือกข้อมูลไม่มั่นใจแบบไม่ใช้ซ้ำบนซัพพอร์ตเวกเตอร์ (AL-SVM-SV-N) และการเรียนรู้เชิงรุกด้วยเอสวีเอ็มเลือกข้อมูลไม่มั่นใจบนซัพพอร์ตเวกเตอร์แบบใช้ข้อมูลซ้ำ (AL-SVM-SV-R) เมื่อทำการทดลองเปรียบเทียบประสิทธิภาพระหว่างการเรียนรู้เชิงรุกทั้งสองกลยุทธ์และการเรียนรู้เชิงรับที่สร้างด้วยตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีน ประสิทธิภาพที่ได้ใกล้เคียงกับการเรียนรู้เชิงรับหรือด้อยกว่าเล็กน้อย แต่ขนาดข้อมูลที่ใช้ในการเรียนรู้จะน้อยกว่า 60% และกลยุทธ์ AL-SVM-SV-R จะมีเรียนรู้เพิ่มเติมได้ดีกว่ากลยุทธ์ AL-SVM-SV-N จากผลการทดลอง

5.2 แนวทางการวิจัยในขั้นถัดไป

แนวทางในการวิจัยในขั้นถัดไปของงาน มีดังนี้

1) ทดสอบกับข้อมูลขนาดใหญ่ที่มีมากขึ้นเพิ่มเติม เพื่อเปรียบเทียบดูประสิทธิภาพของวิธีการที่ได้นำเสนอไปเมื่อทำการทดลองกับข้อมูลที่มีขนาดใหญ่มากกว่า 1,000,000 ตัวอย่างข้อมูล
2) การนำการเรียนรู้เชิงรุกมาใช้ในการช่วยลดการติดฉลากจากผู้เชี่ยวชาญให้กับข้อมูลที่ไม่มีผลเฉลย ซึ่งเป็นอีกแนวทางหนึ่งที่จะปรับกลยุทธ์การเรียนรู้เชิงรุกที่นำเสนอไปทำการทดลองกับข้อมูลที่ไม่มีผลเฉลยเพิ่มเติม เพื่อให้เกิดคุณค่าของงานวิจัยที่นำเสนอไปเพิ่มขึ้นจากเดิม

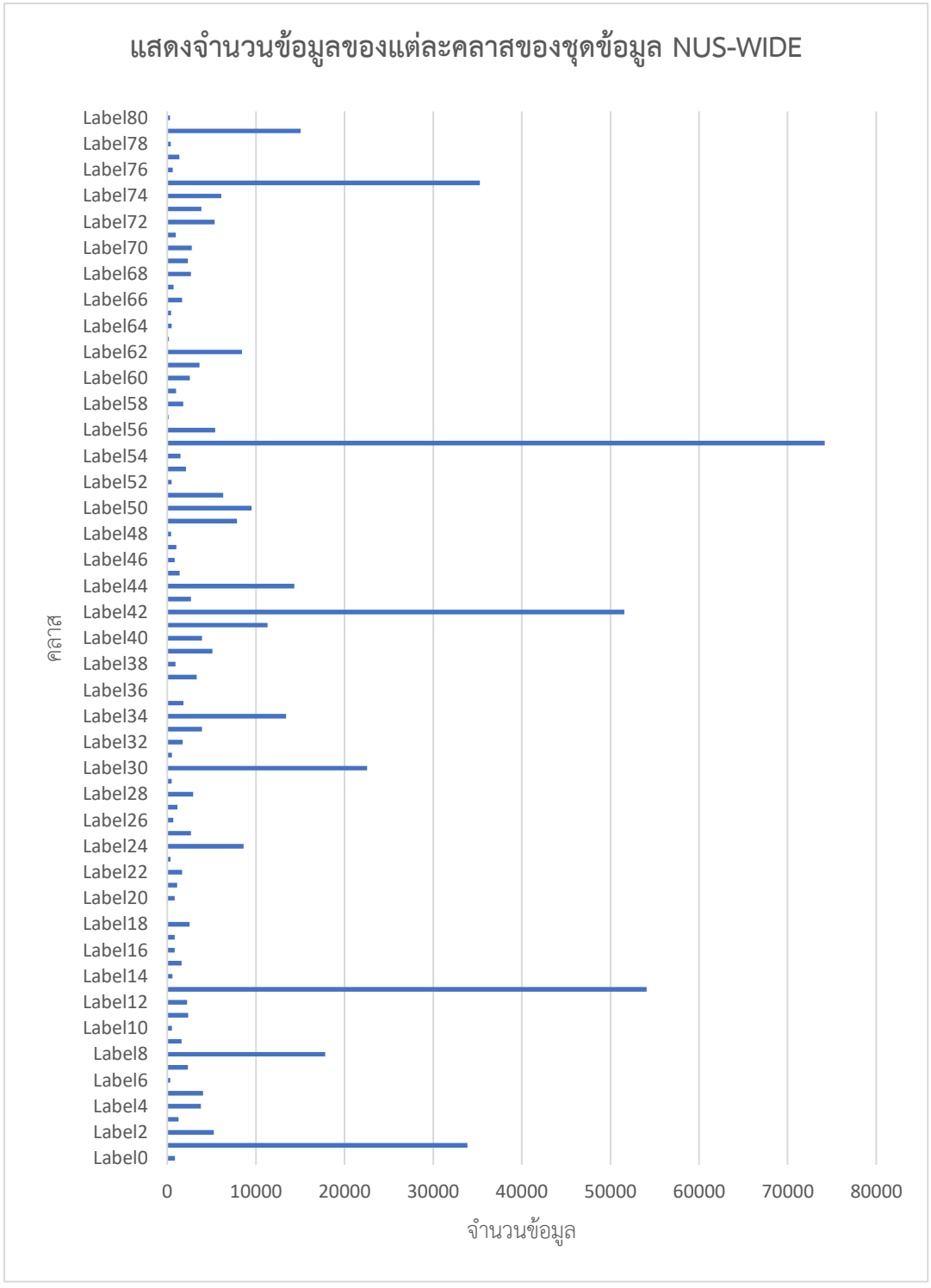
รายการอ้างอิง

1. He, H. and E.A. Garcia, *Learning from imbalanced data.* IEEE Transactions on Knowledge and Data Engineering, 2009. **21**(9): p. 1263-1284.
2. Fu, J.H. and S.L. Lee. *Certainty-enhanced active learning for improving imbalanced data classification*. in *Proceedings - IEEE International Conference on Data Mining, ICDM*. 2011.
3. Tong, S. and D. Koller, *Support vector machine active learning with applications to text classification.* Journal of machine learning research, 2001. **2**(Nov): p. 45-66.
4. Settles, B., *Active learning.* Synthesis Lectures on Artificial Intelligence and Machine Learning, 2012. **6**(1): p. 1-114.
5. Sun, L.-L. and X.-Z. Wang. *A survey on active learning strategy*. in *Machine Learning and Cybernetics (ICMLC), 2010 International Conference on*. 2010. IEEE.
6. Ross, D.A., et al., *Incremental learning for robust visual tracking.* International Journal of Computer Vision, 2008. **77**(1): p. 125-141.
7. Engelbrecht, A.P. and R. Brits, *Supervised training using an unsupervised approach to active learning.* Neural processing letters, 2002. **15**(3): p. 247-260.
8. Zhang, M.-L. and Z.-H. Zhou, *A review on multi-label learning algorithms.* IEEE transactions on knowledge and data engineering, 2014. **26**(8): p. 1819-1837.
9. Lessmann, S., et al. *An evaluation of discrete support vector machines for cost-sensitive learning*. in *IEEE International Conference on Neural Networks - Conference Proceedings*. 2006.
10. Zhang, T. *Solving large scale linear prediction problems using stochastic gradient descent algorithms*. in *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*. 2004.
11. Mining, W.I.D., *Data Mining: Concepts and Techniques.* Morgan Kaufinann, 2006.
12. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* Journal of Machine Learning Research, 2011. **12**: p. 2825-2830.

13. Ferdowsi, Z., R. Ghani, and R. Settimi. *Online active learning with imbalanced classes*. in *Proceedings - IEEE International Conference on Data Mining, ICDM*. 2013.
14. Lu, Z., et al. *Informative sampling for large unbalanced data sets*. in *GECCO'08: Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation 2008*. 2008.
15. Yang, B., et al. *Effective multi-label active learning for text classification*. in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2009.
16. Guyon, I., *Agnostic learning vs. prior knowledge challenge, 2007.* URL http://www. agnostic. Inf. ethz. ch/index. php. Last accessed August, 2010.
17. Caruana, R., T. Joachims, and L. Backstrom, *KDD-Cup 2004: results and analysis.* ACM SIGKDD Explorations Newsletter, 2004. **6**(2): p. 95-108.
18. *LIBSVM Data: Multi-label Classification*. Available from: https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multilabel.html.
19. Tsoumakas, G., et al., *MULAN: A Java library for multi-label learning.* Journal of Machine Learning Research, 2011. **12**: p. 2411-2414.

ภาคผนวก

ภาคผนวก ก
รายละเอียดของชุดข้อมูลแบบหลายฉลาก NUS-WIDE



รูปที่ ก.1 แสดงจำนวนข้อมูลของแต่ละคลาสของชุดข้อมูล NUS-WIDE

ตารางที่ ก.1 แสดงค่าทางสถิติของจำนวนข้อมูลแต่ละคลาสของข้อมูล NUS-WIDE

ค่าเฉลี่ย	ค่าเบี่ยงเบนมาตรฐาน	ค่าสูงสุด	ค่าต่ำสุด
6,220.34	12,559.51	74,190.00	60.00

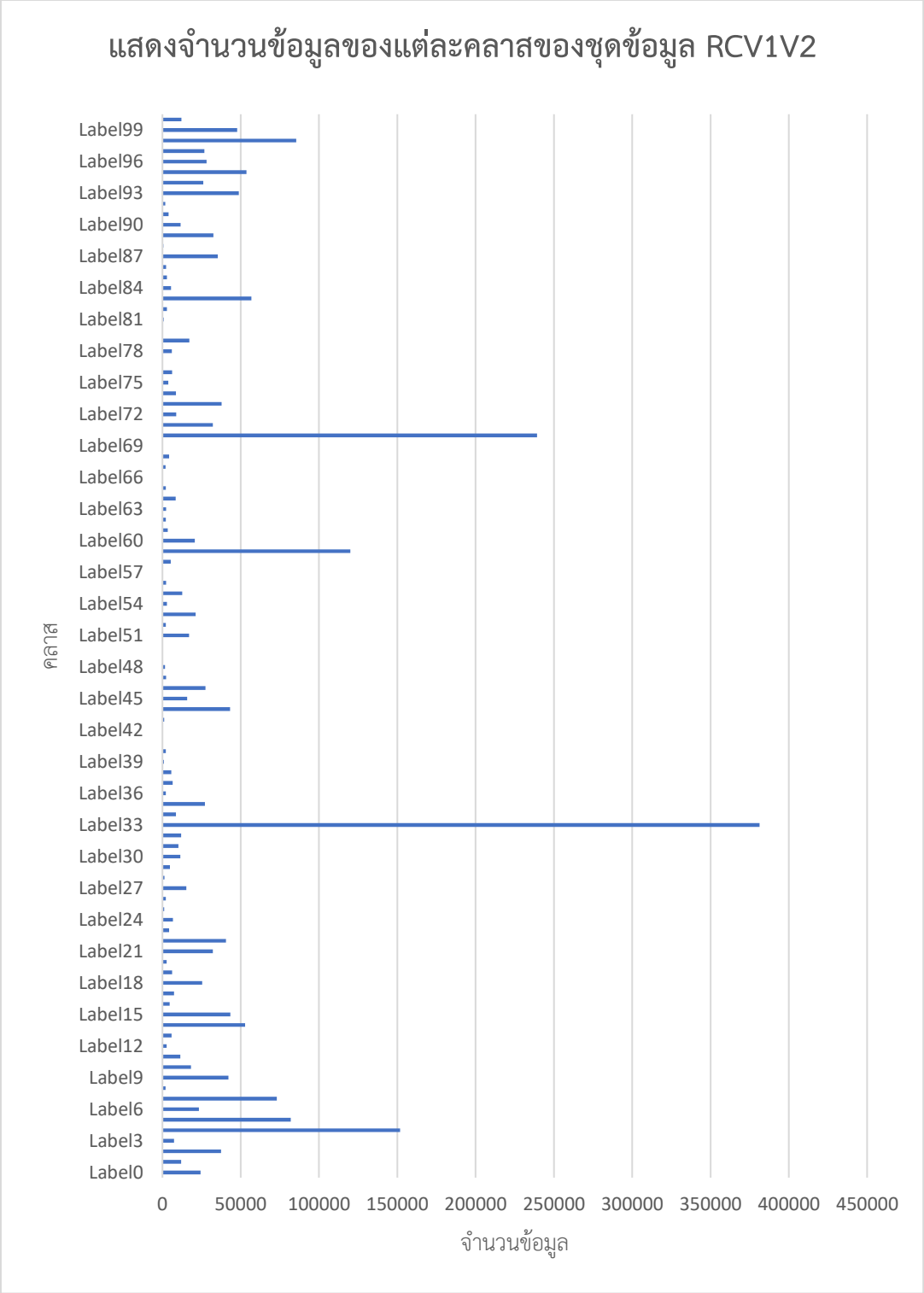
ตารางที่ ก.2 แสดงรายละเอียดจำนวนข้อมูลแต่ละคลาสของชุดข้อมูล NUS-WIDE

คลาส	Label0	Label1	Label2	Label3	Label4	Label5
จำนวนข้อมูล	864	33,887	5,239	1,271	3,780	4,038
คลาส	Label6	Label7	Label8	Label9	Label10	Label11
จำนวนข้อมูล	338	2,337	17,835	1,612	537	2,376
คลาส	Label12	Label13	Label14	Label15	Label16	Label17
จำนวนข้อมูล	2,236	54,087	591	1,601	830	841
คลาส	Label18	Label19	Label20	Label21	Label22	Label23
จำนวนข้อมูล	2,504	63	832	1,114	1,681	386
คลาส	Label24	Label25	Label26	Label27	Label28	Label29
จำนวนข้อมูล	8,605	2,685	668	1,149	2,929	504
คลาส	Label30	Label31	Label32	Label33	Label34	Label35
จำนวนข้อมูล	22,561	541	1,748	3,916	13,392	1,821
คลาส	Label36	Label37	Label38	Label39	Label40	Label41
จำนวนข้อมูล	60	3,340	941	5,099	3,925	11,307
คลาส	Label42	Label43	Label44	Label45	Label46	Label47
จำนวนข้อมูล	51,577	2,663	14,345	1,392	845	1,023
คลาส	Label48	Label49	Label50	Label51	Label52	Label53
จำนวนข้อมูล	420	7,875	9,524	6,327	484	2,114
คลาส	Label54	Label55	Label56	Label57	Label58	Label59
จำนวนข้อมูล	1,483	74,190	5,404	152	1,810	995

คลาส	Label60	Label61	Label62	Label63	Label64	Label65
จำนวนข้อมูล	2,556	3,645	8,418	193	495	426
คลาส	Label66	Label67	Label68	Label69	Label70	Label71
จำนวนข้อมูล	1,666	710	2,683	2,344	2,770	951
คลาส	Label72	Label73	Label74	Label75	Label76	Label77
จำนวนข้อมูล	5,352	3,843	6,099	35,264	627	1,353
คลาส	Label78	Label79	Label80			
จำนวนข้อมูล	399	15,051	309			

ภาคผนวก ข

รายละเอียดของชุดข้อมูลแบบหลายฉลาก RCV1V2



รูปที่ ข.1 แสดงจำนวนข้อมูลของแต่ละคลาสของชุดข้อมูล RCV1V2

ตารางที่ ข.1 แสดงค่าทางสถิติของจำนวนข้อมูลแต่ละคลาสของข้อมูล RCV1V2

ค่าเฉลี่ย	ค่าเบี่ยงเบนมาตราฐาน	ค่าสูงสุด	ค่าต่ำสุด
23,565.33	49,096.19	381,327.00	5.00

ตารางที่ ข.2 แสดงจำนวนข้อมูลแต่ละคลาสของชุดข้อมูล RCV1V2

คลาส	Label0	Label1	Label2	Label3	Label4	Label5
จำนวนข้อมูล	24,325	11,944	37,410	7,410	151,785	81,890
คลาส	Label6	Label7	Label8	Label9	Label10	Label11
จำนวนข้อมูล	23,211	73,092	1,920	42,155	18,313	11,487
คลาส	Label12	Label13	Label14	Label15	Label16	Label17
จำนวนข้อมูล	2,636	5,871	52,817	43,374	4,671	7,406
คลาส	Label18	Label19	Label20	Label21	Label22	Label23
จำนวนข้อมูล	25,403	6,119	2,625	32,153	40,509	4,299
คลาส	Label24	Label25	Label26	Label27	Label28	Label29
จำนวนข้อมูล	6,648	1,115	2,084	15,332	1,210	4,835
คลาส	Label30	Label31	Label32	Label33	Label34	Label35
จำนวนข้อมูล	11,355	10,272	11,878	381,327	8,568	27,100
คลาส	Label36	Label37	Label38	Label39	Label40	Label41
จำนวนข้อมูล	2,182	6,603	5,659	939	2,177	376
คลาส	Label42	Label43	Label44	Label45	Label46	Label47
จำนวนข้อมูล	200	1,206	43,130	15,768	27,405	2,415
คลาส	Label48	Label49	Label50	Label51	Label52	Label53
จำนวนข้อมูล	1,701	52	111	17,035	2,136	21,280
คลาส	Label54	Label55	Label56	Label57	Label58	Label59
จำนวนข้อมูล	2,933	12,634	2,290	391	5,268	119,920

คลาส	Label60	Label61	Label62	Label63	Label64	Label65
จำนวนข้อมูล	20,672	3,307	2,107	2,360	8,404	2,124
คลาส	Label66	Label67	Label68	Label69	Label70	Label71
จำนวนข้อมูล	260	2,036	4,300	40	239,267	32,219
คลาส	Label72	Label73	Label74	Label75	Label76	Label77
จำนวนข้อมูล	8,842	37,739	8,657	3,801	6,261	313
คลาส	Label78	Label79	Label80	Label81	Label82	Label83
จำนวนข้อมูล	6,030	17,241	5	844	2,802	56,878
คลาส	Label84	Label85	Label86	Label87	Label88	Label89
จำนวนข้อมูล	5,498	2,849	2,410	35,317	680	32,615
คลาส	Label90	Label91	Label92	Label93	Label94	Label95
จำนวนข้อมูล	11,532	3,878	1,869	48,696	26,036	53,634
คลาส	Label96	Label97	Label98	Label99	Label100	
จำนวนข้อมูล	28,185	26,752	85,440	47,708	12,130	

ประวัติผู้เขียนวิทยานิพนธ์

นายไพโรจน์ ตันติวชิรฐากูร เกิดเมื่อวันที่ 7 มกราคม พ.ศ. 2517 ที่จังหวัดสุราษฎร์ธานี สำเร็จการศึกษาระดับปริญญาตรีหลักสูตรวิศวกรรมศาตรบัณฑิต สาขาวิศกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยสงขลานครินทร์ ในปีการศึกษา 2539 และเข้าศึกษาในหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2557