การปรับปรุงการจำแนกแบบกึ่งมีผู้สอนด้วยการวิเคราะห์กลุ่มข้อมูล

นางสาวนรีพร พิรุฬห์ทรัพย์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2558

IMPROVED SEMI-SUPERVISED CLASSIFICATION WITH CLUSTER ANALYSIS

Ms.Nareeporn Piroonsup

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy Program in Computer Engineering
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2015

หัวข้อวิทยานิพนธ์	การปรับปรุงการจำแนกแบบกึ่งมีผู้สอนด้วยการวิเคราะห์กลุ่มข้อมูล
โดย	นางสาวนรีพร พิรุฬห์ทรัพย์
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาดุษฎีบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร.บัณฑิต เอื้ออาภรณ์)

คณะกรรมการสอบวิทยานิพนธ์

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

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

..... กรรมการ
(ศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.นัทที นิภานันท์)

..... กรรมการภายนอกมหาวิทยาลัย
(ผู้ช่วยศาสตราจารย์ ดร.ชลวิช นัทธี)

นรีพร พิรุฬห์ทรัพย์: การปรับปรุงการจำแนกแบบกึ่งมีผู้สอนด้วยการวิเคราะห์กลุ่มข้อมูล. (IMPROVED SEMI-SUPERVISED CLASSIFICATION WITH CLUSTER ANALYSIS) อ.ที่ปรึกษาวิทยานิพนธ์หลัก : ผศ. ดร.สุกรี สินธุภิญโญ, 117 หน้า.

ดุษฎีนิพนธ์นี้เสนอวิธีปรับปรุงการจำแนกแบบกึ่งมีผู้สอนซึ่งใช้ตัวอย่างมีป้ายกำกับและไม่มีป้ายกำกับเพื่อสร้างตัวจำแนก โดยปรับปรุงการติดป้ายกำกับให้แก่ตัวอย่างไม่มีป้ายกำกับซึ่งมักเกิดปัญหาเนื่องจากตัวอย่างจำนวนหนึ่งอาจถูกติดป้ายกำกับไม่ถูกต้อง และเมื่อนำตัวอย่างที่ติดป้ายกำกับผิดนั้นไปใช้สร้างตัวจำแนกสุดท้ายย่อมส่งผลเสียต่อประสิทธิภาพของตัวจำแนกกึ่งมีผู้สอนอย่างหลีกเลี่ยงไม่ได้

งานวิจัยส่วนแรกเสนอวิธีแบ่งกลุ่มย่อยตามค่าคุณลักษณะที่ถูกเลือกเพื่อปรับปรุงการเรียนรู้แบบกึ่งมีผู้สอนวิธีจัดกลุ่มและติดป้ายเพื่อปรับปรุงการติดป้ายกำกับตัวอย่างในกลุ่มคลาสปะปน ผลการทดลองบนชุดข้อมูลการลดสิ่งรบกวนในเอกสารภาษาไทยแสดงให้เห็นว่า ความถูกต้องของการติดป้ายกำกับและความถูกต้องของตัวจำแนกสุดท้ายของวิธีที่นำเสนอดีกว่าวิธีติดป้ายกำกับตามคลาสส่วนใหญ่อย่างเห็นได้ชัด นอกจากนั้นวิธีที่นำเสนอยังสามารถลดสิ่งรบกวนในเอกสารภาษาไทยได้ดีกว่าวิธีการลดสิ่งรบกวนวิธีอื่น ๆ ที่เปรียบเทียบ

งานวิจัยส่วนที่สองเสนอวิธีปรับปรุงการจำแนกกึ่งมีผู้สอนวิธีเรียนรู้ด้วยตนเองด้วยการวิเคราะห์ตัวอย่างที่ใช้ในการสอนด้วยการจัดกลุ่มข้อมูล ผลการทดลองแสดงให้เห็นว่าการใช้ตัวอย่างมีป้ายกำกับที่ไม่ครอบคลุมการกระจายตัวของตัวอย่างที่ใช้ในการสอนเพื่อสร้างตัวจำแนกกึ่งมีผู้สอน จะส่งผลต่อความถูกต้องของการจำแนกของตัวจำแนกสุดท้ายอย่างมีนัยสำคัญ งานวิจัยนี้จึงเสนอวิธีปรับปรุงตัวอย่างมีป้ายกำกับในบริเวณที่ไม่ครอบคลุมสองวิธี ได้แก่ การเพิ่มตัวอย่างมีป้ายกำกับโดยผู้ใช้และการเพิ่มตัวอย่างมีป้ายกำกับด้วยตัวจำแนกอื่น ผลการทดลองแสดงให้เห็นว่าวิธีที่นำเสนอสามารถเพิ่มค่าความถูกต้องในการทำนายของตัวจำแนกสุดท้ายอย่างมีนัยสำคัญ

ภาควิชา ...วิศวกรรมคอมพิวเตอร์... ลายมือชื่อนิสิต .....
สาขาวิชา...วิศวกรรมคอมพิวเตอร์... ลายมือชื่อ.ที่ปรึกษาหลัก .....
ปีการศึกษา .....2558.....

5371808121: MAJOR COMPUTER ENGINEERING
KEYWORDS: SEMI-SUPERVISED / CLUSTER / LEARNING / SELF-TRAINING / CLUSTER-AND-LABEL / CLASSIFICATION

NAREEPORN PIROONSUP : IMPROVED SEMI-SUPERVISED CLASSIFICATION WITH CLUSTER ANALYSIS. ADVISOR : ASST. PROF. SUKREE SINTHUPINYO, Ph.D., 117 pp.

We proposed a method to improve semi-supervised classification that is a classification with both labeled and unlabeled data. However, using unlabeled data can seriously degrade the classifier performance because the unlabeled data may incorrectly label. We aim to improve accuracy of unlabeled data labeling in two approaches, i.e., cluster-and-label and self-training. In cluster-and-label approach, we propose an improved labeling method for labeling data in mixed-class clusters, namely, feature selected sub-cluster labeling. The results on noise reduction in Thai document image dataset show that the accuracy of labeling and classification of the proposed method are obviously greater than the majority vote labeling. The proposed method can also significantly better on reducing noise than the comparative noise reduction approaches. In self-training approach, we found that performance of self-training classifier will be ineffective, if distribution of labeled data does not consistent with all training data. We then propose a training data analysis with clustering and suggest to enhance the labeled data distribution by labeling data in unknown clusters. The extensive experiments on UCI and real-world datasets show that our proposed method considerably improves the accuracy of the semi-supervised classifier with statistical significance. We also suggest that this data preprocessing is a necessary step for semi-supervised self-training.

Department : ... Computer Engineering ... Student's Signature .....
Field of Study : .. Computer Engineering .. Advisor's Signature .....
Academic Year : ..... 2015 .....

กิตติกรรมประกาศ

ดุษฎีนิพนธ์ฉบับนี้สำเร็จสมบูรณ์ได้ด้วยความกรุณาของ ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ อาจารย์ที่ปรึกษาดุษฎีนิพนธ์นี้ ผู้ให้ข้อแนะนำ คำปรึกษาและกำลังใจแก่ข้าพเจ้าตลอดระยะเวลาการศึกษา

ข้าพเจ้าขอขอบพระคุณคณะกรรมการสอบวิทยานิพนธ์ ศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา ศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล ผู้ช่วยศาสตราจารย์ ดร.ชลวิช นัทธี และดร.นัทที นิภานันท์ สำหรับข้อเสนอแนะเพื่อปรับปรุงดุษฎีนิพนธ์นี้ให้ถูกต้องครบถ้วน และข้อแนะนำในการวิจัยรวมถึงแนวคิดในการดำเนินชีวิตแก่ข้าพเจ้า

ข้าพเจ้าขอขอบคุณภาควิชาวิศวกรรมคอมพิวเตอร์และคณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย สำหรับทุนอุดหนุนค่าเล่าเรียนประเภท 50/50

ข้าพเจ้าขอขอบคุณกลุ่มเพื่อนวิจัยในแลป Machine learning and knowledge discovery (MIND) และกลุ่มเพื่อนที่ศึกษาร่วมกันในระดับดุษฎีบัณฑิตสำหรับกำลังใจและข้อเสนอแนะในการวิจัย

สุดท้ายนี้ข้าพเจ้าขอขอบคุณนายธีรชัย นางอรวรรณ พิรุฬห์ทรัพย์และนายจักษุเทพ ตี้กุล ครอบครัวของข้าพเจ้าผู้สนับสนุนด้านการศึกษาของข้าพเจ้า ผู้คอยกำลังใจและคำปรึกษาในการวิจัย และช่วยตรวจทานดุษฎีนิพนธ์ฉบับนี้

สารบัญ

บทที่ หน้า

บทที่ หน้า

สารบัญตาราง

ตารางที่ หน้า

ตารางที่ หน้า

สารบัญภาพ

ภาพที่ หน้า

บทที่ 1

บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ข้อมูลดิจิทัลในปัจจุบันมีอยู่จำนวนมากมายเนื่องจากผู้ใช้งานสามารถเป็นผู้ผลิตข้อมูลดิจิทัลได้อย่างง่ายดายด้วยอุปกรณ์อิเล็กทรอนิกส์ใกล้ตัว เช่น สมาร์ตโฟน (Smart phone) กล้องดิจิทัล หรืออุปกรณ์ดิจิทัลที่สามารถสวมใส่ได้ (Wearable device) เป็นต้น การเก็บรวบรวมข้อมูลดิจิทัลเหล่านี้สามารถทำได้สะดวกซึ่งเป็นผลมาจากการพัฒนาโครงข่ายอินเทอร์เน็ตความเร็วสูงและอินเทอร์เน็ตไร้สาย ขณะเดียวกันมนุษย์ต้องการให้ระบบดิจิทัลเหล่านี้สามารถทำงานที่ซับซ้อนมากยิ่งขึ้นทั้งเพื่ออำนวยความสะดวก เพื่อทำงานเสี่ยงภัย หรือเพื่อการศึกษาวิจัย ซึ่งข้อมูลดิจิทัลที่มีอยู่มากมายนี้ สามารถนำมาใช้เพื่อเพิ่มความสามารถของระบบ เพื่อให้ระบบทำงานที่ซับซ้อนคล้ายมนุษย์ได้ ด้วยการนำข้อมูลเหล่านั้นมาใช้ในกระบวนการเรียนรู้ของเครื่อง (machine learning) อย่างไรก็ดีข้อมูลดิจิทัลที่เข้าถึงได้โดยมากมักขาดข้อมูลป้ายกำกับคลาส (class label) ซึ่งเป็นส่วนสำคัญที่จำเป็นต้องใช้ในการเรียนรู้ของระบบ (training) ตัวอย่างเช่น เรามีข้อมูลข้อความจากผู้ใช้จำนวนมากที่ถูกโพสต์ (post) ขึ้นสู่เครือข่ายสังคมออนไลน์ (social network) ข้อความเหล่านี้สามารถนำมาใช้สร้างระบบจำแนกอารมณ์ในข้อความ เป็นอารมณ์ประเภทต่าง ๆ เช่น ข้อความตัดพ้อแสดงความรู้สึกเสียใจ ข้อความแสดงความฉุนเฉียว หรือข้อความแสดงอารมณ์มีความสุข เป็นต้น หากเราจำแนกข้อความตามอารมณ์ได้ จะสามารถนำไปใช้ประโยชน์ได้หลายประการ เช่น ใช้ช่วยในการติดตามอาการผู้ป่วยโรคซึมเศร้า หรือนำไปใช้พัฒนาระบบการอ่านออกเสียงของคอมพิวเตอร์ให้มีอรรถรสมากยิ่งขึ้น

อย่างไรก็ดีข้อความที่โพสต์ส่วนใหญ่นั้นไม่มีป้ายกำกับอารมณ์ต่าง ๆ กำกับมาด้วย เมื่อมีข้อมูลที่มีป้ายกำกับจำนวนไม่เพียงพอจึงไม่สามารถนำไปเรียนรู้ระบบที่มีประสิทธิภาพได้ หรือในปัญหาการจำแนกคลื่นไฟฟ้าหัวใจหรือคลื่นไฟฟ้าสมอง ซึ่งเราสามารถเก็บข้อมูลคลื่นไฟฟ้าของคนไข้ได้เป็นจำนวนเพียงพอภายในระยะเวลาไม่นาน แต่หากต้องการสร้างระบบเพื่อจำแนกคลื่นไฟฟ้าที่ปกติและผิดปกติเพื่อวินิจฉัยอาการของคนไข้เบื้องต้นเพื่อลดภาระของแพทย์เฉพาะทาง เราจำเป็นต้องสอนระบบให้รู้จักคลื่นไฟฟ้าที่ปกติและผิดปกติเสียก่อน ซึ่งจำเป็นต้องอาศัยแพทย์เฉพาะทางเป็นผู้ติดป้ายกำกับคลื่นไฟฟ้าเหล่านั้น แต่เนื่องจากภาระงานจำนวนมากของแพทย์เฉพาะทางทำให้แพทย์ไม่สามารถติดป้ายกำกับข้อมูลเหล่านั้นให้เพียงพอได้ การเรียนรู้แบบกึ่งมีผู้สอนเพื่อจำแนกคลื่นไฟฟ้าหัวใจ (Hughes et al., 2004) (Sun et al., 2012) (Zhang

et al., 2013) (Oster et al., 2015) และคลื่นไฟฟ้าสมอง (Li et al., 2008) (Guan G., 2013) จึงเป็นอีกปัญหาที่ได้รับความสนใจอย่างมาก ปัญหาการเรียนรู้ของเครื่องด้วยข้อมูลมีป้ายกำกับคลาสจำนวนไม่เพียงพอร่วมกับข้อมูลไม่มีป้ายกำกับคลาสซึ่งมีอยู่เป็นจำนวนมาก จึงเป็นปัญหาหนึ่งที่ได้รับความสนใจมากในปัจจุบัน โดยหนึ่งในแนวทางแก้ปัญหานั้นคือการเรียนรู้แบบกึ่งมีผู้สอน (semi-supervised learning)

การเรียนรู้แบบกึ่งมีผู้สอนใช้ตัวอย่างสองประเภทเพื่อเรียนรู้ระบบ ได้แก่ ตัวอย่างที่มีป้ายกำกับคลาสหรือตัวอย่างมีป้ายกำกับ (labeled data) และตัวอย่างไม่มีป้ายกำกับ (unlabeled data) ตัวอย่างมีป้ายกำกับประกอบด้วยข้อมูลสองส่วน ได้แก่ ค่าของคุณลักษณะ (attribute value) และป้ายกำกับคลาส (class label) ส่วนตัวอย่างไม่มีป้ายกำกับมีค่าของคุณลักษณะเพียงอย่างเดียว (Zhu, 2008) สำหรับตัวอย่างปัญหาการเรียนรู้เพื่อจำแนกข้อความตามอารมณ์และการจำแนกสัญญาณคลื่นไฟฟ้าที่ปกติและผิดปกติที่ยกตัวอย่างข้างต้นนั้นคือ ปัญหาการจำแนก (classification) ปัญหาการจำแนกเป็นรูปแบบปัญหาสำคัญในการเรียนรู้ของเครื่อง การแก้ปัญหาการจำแนกโดยทั่วไปใช้วิธีการจำแนกแบบมีผู้สอน (supervised classification) เพื่อสร้างแบบจำลองเพื่อจำแนกข้อมูลเป็นประเภทต่าง ๆ เช่น การสร้างแบบจำลองด้วยนิวรอลเน็ตเวิร์กเพื่อรู้จำตัวอักษรภาษาไทย หรือการหาซัพพอร์ตเวกเตอร์แมชชีนเพื่อจำแนกประเภทของเอกสาร เป็นต้น การจำแนกแบบมีผู้สอนดังกล่าวสร้างตัวจำแนก (classifier) จากการสอน (training) ระบบด้วยตัวอย่างมีป้ายกำกับ เนื่องจากข้อมูลมีป้ายกำกับนั้นมีจำนวนไม่เพียงพอในขณะที่ข้อมูลไม่มีป้ายกำกับคลาสนั้นมีอยู่จำนวนมากจึงมีงานวิจัยที่พยายามใช้วิธีการเรียนรู้กึ่งมีผู้สอนเพื่อแก้ปัญหาการจำแนกนี้หรือเรียกว่าการจำแนกแบบกึ่งมีผู้สอน (semi-supervised classification) การเรียนรู้แบบกึ่งมีผู้สอนที่นิยมใช้กับปัญหาการจำแนกนั้นมีอยู่หลายวิธีซึ่งแต่ละวิธีมีสมมติฐานที่แตกต่างกันและเหมาะสมกับปัญหาหรือชุดข้อมูลที่มีลักษณะแตกต่างกัน รายละเอียดวิธีการจำแนกกึ่งมีผู้สอนแบบต่าง ๆ จะนำเสนอในบทที่ โดยในงานวิจัยนี้สนใจศึกษาเพื่อปรับปรุงวิธีจัดกลุ่มและติดป้าย (cluster-and-label) และวิธีเรียนรู้ด้วยตนเอง (self-training) ซึ่งเป็นวิธีที่ได้การยอมรับว่ามีประสิทธิภาพและถูกนำไปใช้ในหลายงานวิจัย

วิธีจัดกลุ่มและติดป้ายเป็นเทคนิคหนึ่งของการเรียนรู้กึ่งมีผู้สอนด้วยวิธีเจเนเรทีฟโมเดล (Generative model) (Nigam et al., 2000) (Demiriz et al., 1999) (Dara et al., 2002) การเรียนรู้กึ่งมีผู้สอนวิธีนี้ถูกนำไปในหลายปัญหาเช่น การจำแนกประเภทเอกสาร (Nigam et al., 2000) (Zeng et al., 2003) (Zhang et al., 2015) การค้นหาสารชีวโมเลกุลประเภทโปรตีน (Sugiyama et al., 2012) การจำแนกปริมาณการใช้งานโพรโทคอลในเครือข่าย (Grimaudo et al., 2014) เป็นต้น ขั้นตอนวิธีจัดกลุ่มและติดป้ายใช้การจัดกลุ่มข้อมูลเพื่อจัดกลุ่มตัวอย่างทั้งหมดทั้งที่มีป้ายกำกับและไม่มีป้ายกำกับ หลังจากนั้นจึงติดป้ายกำกับตัวอย่างไม่มีป้ายกำกับในแต่ละกลุ่มตามคลาสของตัวอย่างมีป้ายกำกับส่วนใหญ่ในกลุ่มนั้น วิธีจัดกลุ่มและติดป้ายนี้มี

ประสิทธิภาพดีมาก เมื่อนำไปใช้แก้ปัญหาการจำแนกกึ่งมีผู้สอนเพื่อลดสิ่งรบกวนในเอกสารภาษาไทย ซึ่งเป็นงานที่ผู้วิจัยได้นำเสนอไว้ก่อนหน้า (Piroonsup and Sinthupinyo, 2010a) (Piroonsup and Sinthupinyo, 2010b)

วิธีเรียนรู้ด้วยตนเองเป็นวิธีหนึ่งที่มีประสิทธิภาพและง่ายต่อการนำไปใช้จึงถูกนำไปใช้ในหลายปัญหาและในหลายแขนงข้อมูล เช่น การตรวจจับวัตถุ (Rosenberg et al., 2005) การรู้จำใบหน้า (Roli and Marcialis, 2006) การแจกแจงรูปประโยค (parsing) (McClosky et al., 2006) การจำแนกคลื่นไฟฟ้าสมอง (Li et al., 2008) (Guan G., 2013) และปัญหาอนุกรมเวลา (time series) (Wei and Keogh, 2006) เป็นต้น วิธีเรียนรู้ด้วยตนเองไม่มีสมมติฐานที่กำหนดคุณลักษณะของชุดข้อมูลจึงสามารถนำไปใช้กับชุดข้อมูลได้หลากหลายและยังสามารถนำไปใช้เป็นวิธีห่อหุ้ม (wrapper method) วิธีการอื่น ๆ ได้อีกด้วย เช่น นำไปใช้ร่วมกับปัญหาการเรียนรู้ข้อมูลที่พิจารณาอิทธิพลของมูลค่าคลาส (cost-sensitive learning) (Liu et al., 2009) เป็นต้น ขั้นตอนวิธีของการจำแนกกึ่งมีผู้สอนวิธีเรียนรู้ด้วยตนเองเริ่มจากการสร้างตัวจำแนกตั้งต้น (initial classifier) จากตัวอย่างมีป้ายกำกับ ตัวจำแนกตั้งต้นนี้ใช้เพื่อติดป้ายกำกับให้แก่ตัวอย่างไม่มีป้ายกำกับที่ตัวจำแนกมั่นใจที่สุด (certain data) เพื่อเพิ่มจำนวนตัวอย่างมีป้ายกำกับ แล้วจึงใช้ตัวอย่างมีป้ายกำกับที่เพิ่มขึ้นมานั้นไปสร้างตัวจำแนกในขั้นตอนสุดท้าย (final classifier) สมมติฐานที่สำคัญในการนำวิธีเรียนรู้ด้วยตนเองไปใช้คือตัวอย่างมีป้ายกำกับตั้งต้นจะต้องเพียงพอที่จะสร้างตัวจำแนกตั้งต้น (initial classifier) ที่มีคุณภาพได้

แม้ว่าการใช้ตัวอย่างไม่มีป้ายกำกับร่วมกับตัวอย่างมีป้ายกำกับนี้สามารถปรับปรุงคุณภาพตัวจำแนกได้ดี แต่การใช้ตัวอย่างไม่มีป้ายกำกับก็มีโอกาสที่จะส่งผลเสียต่อตัวจำแนกได้เช่นกัน เนื่องจากตัวอย่างไม่มีป้ายกำกับนั้นถูกติดป้ายกำกับด้วยกระบวนการภายใต้สมมติฐานที่กำหนดจึงอาจมีบางตัวอย่างที่ถูกติดป้ายกำกับผิดไปจากคลาสที่แท้จริง อาทิเช่นในวิธีการเรียนรู้ด้วยตนเอง ซึ่งใช้ตัวจำแนกที่สร้างจากตัวอย่างมีป้ายกำกับตั้งต้นที่มีจำนวนไม่มากเพื่อติดป้ายกำกับให้แก่ตัวอย่างไม่มีป้ายกำกับ แต่ตัวอย่างมีป้ายกำกับตั้งต้นมีจำนวนน้อยมาก จึงอาจส่งผลให้ตัวจำแนกตั้งต้นติดป้ายกำกับผิดพลาด เมื่อตัวอย่างที่ติดป้ายกำกับผิดนั้นจะถูกนำไปใช้สร้างตัวจำแนกในขั้นตอนสุดท้ายย่อมส่งผลเสียต่อประสิทธิภาพของตัวจำแนก โดยทั่วไปสัดส่วนระหว่างตัวอย่างที่ติดป้ายกำกับใหม่นั้นมีจำนวนมากกว่าตัวอย่างมีป้ายกำกับตั้งต้นมาก จึงอาจมีอิทธิพลต่อการสร้างตัวจำแนกมากกว่าตัวอย่างที่ติดป้ายกำกับตั้งต้นได้ ในกรณีที่แย่ที่สุดอาจส่งผลให้ประสิทธิภาพของตัวจำแนกแบบกึ่งมีผู้สอนนั้นแย่กว่าการตัวจำแนกที่สร้างจากตัวอย่างมีป้ายกำกับตั้งต้นเพียงอย่างเดียว ความเสี่ยงในการใช้ตัวอย่างไม่มีป้ายกำกับในการเรียนรู้นั้นเป็นหนึ่งในปัญหาของการใช้ตัวอย่างไม่มีป้ายกำกับที่เริ่มได้รับความสนใจมากขึ้น (Singh et al., 2009) (Li and Zhou, 2015)

งานวิจัยนี้จึงเสนอวิธีปรับปรุงการเรียนรู้แบบกึ่งมีผู้สอนโดยแบ่งเป็นสองส่วน ได้แก่ ส่วนที่หนึ่งเสนอวิธีการปรับปรุงการเรียนรู้แบบกึ่งมีผู้สอนวิธีจัดกลุ่มและติดป้าย โดยเสนอวิธีติดป้ายกำกับในกลุ่มคลาสปะปนด้วยวิธีติดป้ายกำกับโดยแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือก (Feature selected sub-cluster labeling: FSL) และส่วนที่สองเสนอการวิธีปรับปรุงการเรียนรู้แบบกึ่งมีผู้สอนวิธีเรียนรู้ด้วยตนเองด้วยการวิเคราะห์ตัวอย่างที่ใช้ในการสอนด้วยการจัดกลุ่มข้อมูล (Training data Analysis with Clustering to Improve Semi-supervised Self-training: TACISS)

งานวิจัยในส่วนที่หนึ่งนั้นเสนอวิธีปรับปรุงการจำแนกกึ่งมีผู้สอนแบบจัดกลุ่มและติดป้ายโดยปรับปรุงการติดป้ายกำกับในกลุ่มที่ประกอบไปด้วยตัวอย่างมีป้ายกำกับมากกว่าหนึ่งคลาสหรือกลุ่มคลาสปะปน (mixed-class cluster) เนื่องจากการติดป้ายกำกับในแต่ละกลุ่มเลือกป้ายกำกับโดยพิจารณาคลาสของตัวอย่างมีป้ายกำกับในกลุ่มนั้น แต่หากกลุ่มข้อมูลที่ได้เป็นกลุ่มคลาสปะปนแนวทางการแก้ปัญหานั้นอาจเลือกไม่ติดป้ายกำกับในกลุ่มนั้น Dara et al. (2002) ซึ่งทำให้เราใช้ประโยชน์จากตัวอย่างไม่มีป้ายกำกับได้น้อยลงเนื่องจากตัวอย่างส่วนใหญ่อยู่ในกลุ่มคลาสปะปน หรืออีกแนวทางหนึ่งคือเลือกติดป้ายกำกับด้วยวิธีโหวตคลาสส่วนใหญ่ (majority vote) Demiriz et al. (1999) ซึ่งอาจส่งผลให้บางตัวอย่างติดป้ายกำกับผิดเนื่องจากไม่พิจารณาตัวอย่างในคลาสส่วนน้อยซึ่งมีโอกาสสูงที่จะมีตัวอย่างในคลาสส่วนน้อยในกลุ่มนั้นและตัวอย่างที่ติดป้ายกำกับผิดนั้น ย่อมส่งผลเสียต่อประสิทธิภาพของตัวจำแนกสุดท้ายอย่างหลีกเลี่ยงไม่ได้

งานวิจัยนี้จึงเสนอวิธีติดป้ายกำกับสำหรับตัวอย่างในกลุ่มคลาสปะปนโดยแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือกหรือวิธี FSL โดยตัวอย่างในกลุ่มคลาสปะปนจะถูกแบ่งเป็นกลุ่มย่อยแต่เนื่องจากค่าคุณลักษณะเดิมไม่สามารถแยกตัวอย่างต่างคลาสออกจากกันได้ ในงานนี้จึงเสนอให้ใช้คุณลักษณะชุดใหม่ที่ได้จากการเลือกคุณลักษณะ (feature selection) โดยเลือกคุณลักษณะที่เหมาะสมในการจำแนกตัวอย่างต่างคลาสออกจากกัน หลังจากได้ค่าคุณลักษณะนั้นแล้วจึงแบ่งตัวอย่างเป็นกลุ่มย่อย แล้วติดป้ายกำกับตัวอย่างไม่มีป้ายกำกับในกลุ่มย่อยนั้น จากการทดลองบนชุดข้อมูลการลดสิ่งรบกวนในเอกสารภาษาไทยซึ่งเป็นชุดข้อมูลจริง พบว่าความถูกต้องของการติดป้ายกำกับและความถูกต้องของตัวจำแนกสุดท้ายของวิธีที่นำเสนอดีกว่าวิธีโหวตคลาสส่วนใหญ่ งานวิจัยนี้ยังได้ทดลองวัดประสิทธิภาพเปรียบเทียบวิธีการลดสิ่งรบกวนที่นำเสนอกับวิธีลดสิ่งรบกวนลักษณะคล้ายตัวอักษร (Stroke-liked pattern noise removal: SPN) และการใช้ซอฟต์แวร์ลดสิ่งรบกวนเชิงพาณิชย์ ScanFix Xpress 6.0 พบว่าวิธีที่นำเสนอสามารถจำแนกสิ่งรบกวนและตัวอักษรออกจากกันได้ดีกว่าวิธีการที่เปรียบเทียบอย่างมีนัยสำคัญ รายละเอียดของวิธีการและผลการทดลองนำเสนออยู่ในบทที่ 5

งานวิจัยในส่วนที่สองเสนอการวิเคราะห์ตัวอย่างที่ใช้ในการเรียนรู้ (training data) ด้วยการจัดกลุ่มข้อมูลเพื่อปรับปรุงการจำแนกกึ่งมีผู้สอนวิธีเรียนรู้ด้วยตนเองหรือวิธี TACISS ซึ่งพิจารณาความครอบคลุมของตัวอย่างมีป้ายกำกับกับการกระจายตัวของตัวอย่างที่ใช้ในการเรียนรู้ทั้งหมด งานวิจัยนี้ใช้การจัดกลุ่มข้อมูลแบบกึ่งมีผู้สอนวิธีกำหนดตัวแทนกลุ่มตั้งต้น (semi-supervised clustering by seeding) (Basu et al., 2002) เพื่อจัดกลุ่มตัวอย่าง โดยสามารถแบ่งกลุ่มตัวอย่างได้เป็นสองประเภท คือ กลุ่มที่ประกอบไปด้วยตัวอย่างมีป้ายกำกับในกลุ่มนั้นหรือกลุ่มมีคลาส (labeled cluster) และกลุ่มที่ประกอบไปด้วยตัวอย่างไม่มีป้ายกำกับเท่านั้นหรือกลุ่มไม่ทราบคลาส (unknown cluster) จากการทดลองพบว่าชุดตัวอย่างมีป้ายกำกับที่ไม่ครอบคลุมการกระจายตัวของตัวอย่าง ส่งผลเสียต่อความถูกต้องของการจำแนกอย่างมีนัยสำคัญ งานวิจัยนี้จึงเสนอให้เพิ่มข้อมูลในกลุ่มไม่ทราบคลาส ด้วยการติดป้ายกำกับให้แก่ตัวอย่างในกลุ่มไม่ทราบคลาส

อย่างไรก็ดีการติดป้ายกำกับให้แก่ตัวอย่าง จำเป็นต้องอาศัยแรงงานมนุษย์ในการพิจารณาป้ายกำกับ ส่งผลให้ระบบไม่เป็นอัตโนมัติ (automatic) ในงานวิจัยนี้จึงศึกษาการใช้ตัวจำแนกอื่น ๆ เพื่อติดป้ายกำกับในกลุ่มไม่ทราบคลาสและพบว่าจำแนกป่าไม้แบบสุ่ม (random forest classifier) สามารถติดป้ายกำกับตัวอย่างที่เป็นตัวแทนของกลุ่มที่ไม่ทราบคลาสได้ถูกต้องมากที่สุด และเมื่อนำตัวอย่างที่ติดป้ายกำกับด้วยวิธีที่นำเสนอไปสร้างตัวจำแนกกึ่งมีผู้สอน พบว่าชุดตัวอย่างที่ปรับปรุงป้ายกำกับคลาสด้วยวิธีที่นำเสนอสามารถ สร้างตัวจำแนกกึ่งมีผู้สอนที่มีความถูกต้องในการจำแนกดีกว่าตัวจำแนกกึ่งมีผู้สอนที่สร้างจากตัวอย่างมีป้ายกำกับคลาสเดิมอย่างมีนัยสำคัญ นอกจากนี้ยังพบว่าวิธีปรับปรุงตัวอย่างมีป้ายกำกับที่นำเสนอ ช่วยปรับปรุงความถูกต้องของตัวจำแนกกึ่งมีผู้สอนที่เดิมให้ผลไม่แตกต่างหรือแย่กว่าการเรียนรู้โดยไม่ใช้ตัวอย่างมีป้ายกำกับ ให้ดีขึ้นอย่างมีนัยสำคัญในหลายชุดข้อมูล การทดลองนี้ใช้ชุดข้อมูลจาก UCI Machine Learning repository (Lichman, 2013) ซึ่งเป็นชุดข้อมูลมาตรฐานในการเรียนรู้ของเครื่อง จำนวน 16 ชุดข้อมูล และใช้ชุดข้อมูลจริง (real-world dataset) จากปัญหาการจำแนกสิ่งรบกวนออกจากตัวอักษรในภาพเอกสารภาษาไทย จำนวน 67 ภาพเอกสาร (Piroonsup and Sinthupinyo, 2010a) (Piroonsup and Sinthupinyo, 2010b)

งานวิจัยเกี่ยวกับการวิเคราะห์ความเสี่ยงของตัวอย่างไม่มีป้ายกำกับในการเรียนรู้แบบกึ่งมีผู้สอนเริ่มได้รับความสนใจมากขึ้น (Nigam et al., 2000) (Chapelle et al., 2006) (Cozman and Cohen, 2002) (Singh et al., 2009) (Ben-David et al., 2008) โดยงานวิจัยส่วนใหญ่สนใจศึกษาความเสี่ยงบนการเรียนรู้แบบกึ่งมีผู้สอนวิธีเจเนเรทีฟโมเดล สำหรับวิธีเรียนรู้ด้วยตนเองนั้นไม่มีงานวิจัยที่เสนอแนวทางลดความเสี่ยงนี้โดยตรง ส่วนมากเป็นงานวิจัยที่เสนอวิธีปรับปรุงคุณภาพตัวจำแนกจากวิธีเรียนรู้ด้วยตนเองด้วยการใช้เทคนิคต่าง ๆ เพื่อติดป้ายกำกับด้วยตนเอง งานวิจัยนี้เป็นงานวิจัยแรกที่เสนอการวิเคราะห์ตัวอย่างที่ใช้ในการเรียนรู้ก่อนที่จะ

นำไปใช้ในการเรียนรู้แบบกึ่งมีผู้สอน (pre-processing for semi-supervised learning) ส่วนงานวิจัยที่เกี่ยวข้องกับการใช้การจัดกลุ่มข้อมูลร่วมกับตัวอย่างไม่มีป้ายกำกับ ส่วนใหญ่ใช้การจัดกลุ่มข้อมูลเพื่อติดป้ายกำกับในการเรียนรู้กึ่งมีผู้สอน หรือใช้การจัดกลุ่มข้อมูลเพื่อเลือกตัวอย่างในการเรียนรู้แบบกัมมันต์ (active learning) ซึ่งวัตถุประสงค์และวิธีการของวิธีเหล่านี้แตกต่างจากการศึกษาในงานวิจัยนี้ รายละเอียดงานวิจัยที่เกี่ยวข้องและความแตกต่างกับงานวิจัยนี้จะนำเสนอในบทที่

1.2 วัตถุประสงค์ของการวิจัย

- ศึกษาแนวทางปรับปรุงความถูกต้องในการติดป้ายกำกับสำหรับตัวอย่างในกลุ่มคลาสปะปนในการจำแนกแบบกึ่งมีผู้สอนวิธีจัดกลุ่มและติดป้าย เพื่อเพิ่มความถูกต้องของตัวจำแนกกึ่งมีผู้สอนที่สร้างจากการตัวอย่างที่ติดป้ายกำกับใหม่นั้น
- ศึกษาวิธีวิเคราะห์คุณภาพของชุดตัวอย่างมีป้ายกำกับสำหรับการจำแนกแบบกึ่งมีผู้สอนวิธีเรียนรู้ด้วยตนเอง
- ศึกษาวิธีลดความเสี่ยงของตัวอย่างไม่มีป้ายกำกับที่ใช้ในการจำแนกแบบกึ่งมีผู้สอนวิธีเรียนรู้ด้วยตนเอง

1.3 ประโยชน์ที่คาดว่าจะได้รับ

- ได้วิธีการติดป้ายกำกับตัวอย่างไม่มีป้ายกำกับที่ช่วยปรับปรุงความถูกต้องในการติดป้ายกำกับ สำหรับตัวอย่างในกลุ่มคลาสปะปนในการจำแนกแบบกึ่งมีผู้สอนวิธีจัดกลุ่มและติดป้าย
- ได้วิธีการวิเคราะห์คุณภาพของชุดตัวอย่างมีป้ายกำกับสำหรับการจำแนกกึ่งมีผู้สอนวิธีเรียนรู้ด้วยตนเอง
- ได้วิธีปรับปรุงชุดตัวอย่างมีป้ายกำกับ ที่ช่วยปรับปรุงความถูกต้องในการติดป้ายกำกับ และปรับปรุงความถูกต้องในการจำแนกของตัวจำแนกกึ่งมีผู้สอน สำหรับการจำแนกกึ่งมีผู้สอนวิธีเรียนรู้ด้วยตนเอง

1.4 ขอบเขตของการวิจัย

- การทดลองใช้ชุดข้อมูลปัญหาสองคลาสที่ไม่ขาดค่าคุณลักษณะ (missing value) และไม่มีตัวอย่างที่ติดป้ายกำกับผิด
- กำหนดให้การติดป้ายกำกับจากผู้ใช้ถูกต้องเสมอ

1.5 ลำดับขั้นตอนในการเสนอผลการวิจัย

บทที่ 2 อธิบายแนวคิดและงานวิจัยที่เกี่ยวข้องกับการใช้ประโยชน์จากตัวอย่างไม่มีป้ายกำกับในการเรียนรู้ของเครื่อง บทที่ 3 อธิบายแนวคิดและงานวิจัยที่เกี่ยวข้องกับการใช้การจัดกลุ่มกับตัวอย่างไม่มีป้ายกำกับ โดยแบ่งเป็น การใช้การจัดกลุ่มกับการเรียนรู้แบบกึ่งมีผู้สอนและการใช้การจัดกลุ่มกับการเรียนรู้แบบกัมมันต์ บทที่ 4 เสนอวิธีปรับปรุงการจัดกลุ่มและติดป้ายกำกับในกลุ่มคลาสปะปนโดยใช้การติดป้ายกำกับกลุ่มย่อยตามคุณลักษณะที่ถูกเลือก และผลการทดลองเปรียบเทียบผลลัพธ์วิธีที่นำเสนอกับวิธีที่ใช้โดยทั่วไป บทที่ 5 เสนอการวิเคราะห์ตัวอย่างมีป้ายกำกับเพื่อปรับปรุงการจำแนกกึ่งมีผู้สอนวิธีเรียนรู้ด้วยตนเอง ได้แก่ วิธีหาค่าคะแนนความเชื่อมั่นสำหรับแต่ละตัวอย่างมีป้ายกำกับ และวิธีวิเคราะห์ชุดตัวอย่างมีป้ายกำกับด้วยการวิเคราะห์กลุ่มข้อมูล และผลการทดลองแสดงประสิทธิภาพของตัวจำแนกกึ่งมีผู้สอนด้วยวิธีปรับปรุงตัวอย่างมีป้ายกำกับที่นำเสนอ บทที่ 6 สรุปผลการทดลอง วิเคราะห์ผลการทดลอง และเสนอแนวทางและผลการทดลองเบื้องต้นสำหรับงานวิจัยในอนาคต

บทที่ 2

การเรียนรู้ด้วยตัวอย่างมีป้ายกำกับและไม่มีป้ายกำกับ

การเรียนรู้โดยใช้ตัวอย่างมีป้ายกำกับร่วมกับตัวอย่างไม่มีป้ายกำกับเป็นหัวข้อวิจัยที่ได้รับความสนใจอย่างมากในปัจจุบัน เนื่องจากข้อมูลดิจิทัลในปัจจุบันมีอยู่เป็นจำนวนมากและสามารถเข้าถึงได้ง่าย ข้อมูลดิจิทัลเหล่านี้ประกอบไปด้วยสารสนเทศ (information) ที่น่าสนใจและสามารถนำไปใช้ประโยชน์ได้หลายประการ อย่างไรก็ดีข้อมูลดิจิทัลที่เข้าถึงได้ส่วนมากขาดข้อมูลสำคัญ ได้แก่ ป้ายกำกับคลาสสำหรับแต่ละตัวอย่าง ซึ่งจำเป็นต้องใช้ในการเรียนรู้ของระบบเพื่อค้นหาสารสนเทศสำคัญจากข้อมูลเหล่านั้น สาเหตุที่ตัวอย่างจำนวนมากไม่มีป้ายกำกับเนื่องจากการติดป้ายกำกับให้แก่ตัวอย่างนั้นต้องใช้ทรัพยากรจำนวนมากทั้งเวลา แรงงาน และค่าใช้จ่าย ยกตัวอย่างเช่น

- ปัญหาที่ต้องใช้เวลานานในการติดป้ายกำกับ เช่น การจำแนกประเภทของเอกสาร (Nigam et al., 2000) ที่ผู้ใช้ต้องอ่านทำความเข้าใจเอกสารแต่ละฉบับเพื่อติดป้ายกำกับแก่เอกสารตามประเภทที่กำหนด
- ปัญหาที่ต้องใช้แรงงานมากในการติดป้ายกำกับ เช่น ปัญหาการจำแนกสิ่งรบกวนออกจากตัวอักษรในเอกสารภาษาไทย (Piroonsup and Sinthupinyo, 2010a) (Piroonsup and Sinthupinyo, 2010b) การสร้างตัวอย่างเริ่มจากการแปลงเอกสารกระดาษเข้าสู่ระบบดิจิทัลด้วยการสแกนหรือถ่ายภาพเอกสาร กระบวนการแปลงข้อมูลใช้เวลาไม่นานก็สามารถนำภาพเอกสารเข้าสู่ระบบได้จำนวนมาก แต่ในการเรียนรู้ของเครื่องเพื่อจำแนกสิ่งรบกวนออกจากตัวอักษรในเอกสารภาษาไทย ภาพเอกสารประกอบไปด้วยตัวอักษรและสิ่งรบกวนหลากหลายรูปแบบ ระบบจำเป็นต้องใช้ตัวอย่างแต่ละประเภทจำนวนมากเพียงพอในกระบวนการเรียนรู้ ซึ่งต้องอาศัยผู้ใช้ช่วยติดป้ายกำกับองค์ประกอบที่เชื่อมกัน (connected component) เพื่อสร้างตัวอย่างสำหรับแต่ละภาพเอกสาร การติดป้ายกำกับจนเพียงพอแก่การเรียนรู้ต้องใช้เวลาและแรงงานจำนวนมาก แต่จำนวนองค์ประกอบที่ไม่มีป้ายกำกับในภาพเอกสารหนึ่งภาพอาจมีจำนวนหลายหมื่นตัวอย่าง
- ปัญหาที่ต้องอาศัยผู้เชี่ยวชาญเฉพาะทางในการติดป้ายกำกับตัวอย่าง เช่น ปัญหาการติดป้ายกำกับคลื่นไฟฟ้าสมอง (Hughes et al., 2004) (Sun et al., 2012) (Zhang et al., 2013) (Oster et al., 2015) หรือคลื่นไฟฟ้าหัวใจ (Li et al., 2008) (Guan G., 2013) ที่การเก็บข้อมูลคลื่นไฟฟ้าสมองหรือคลื่นไฟฟ้าหัวใจนั้นสามารถเก็บได้เป็นจำนวนมากในระยะเวลาไม่นาน และยังมีเครื่องมือที่สามารถเก็บข้อมูลเหล่านี้ได้อย่างอัตโนมัติ แต่การติดป้ายกำกับแต่ละส่วนของคลื่นไฟฟ้านั้นต้องอาศัยแพทย์เฉพาะทางเป็นผู้ติดป้ายกำกับ

- ปัญหาที่ต้องใช้มูลค่าสูงในเพื่อติดป้ายกำกับแก่ข้อมูล เช่น การวินิจฉัยผู้ป่วยโรคมะเร็งชนิดต่าง ๆ (Bair and Tibshirani, 2004) (Koestler, 2010)
- ปัญหาที่ไม่สามารถหาป้ายกำกับให้แก่ข้อมูลบางส่วนได้ เช่น การทำนายระยะเวลารอดชีพ (survival times) ของคนไข้จากข้อมูลยีน โดยการทำนายระยะเวลารอดชีพจำเป็นต้องทราบชนิดของมะเร็ง แต่โรคมะเร็งหลายชนิดยังไม่สามารถวินิจฉัยได้ งานวิจัยนี้จึงสร้างแบบจำลองเพื่อทำนายชนิดของมะเร็งที่สัมพันธ์กับระยะเวลารอดชีพ โดยใช้ข้อมูลยีนและระยะเวลารอดชีพของคนไข้มะเร็งที่ไม่สามารถระบุชนิดมะเร็งในอดีตมาทำนายชนิดของมะเร็งตามระยะเวลารอดชีพ และนำแบบจำลองนั้นมาทำนายระยะเวลารอดชีพของคนไข้ในปัจจุบัน (Bair and Tibshirani, 2004) (Koestler, 2010)

งานวิจัยที่ศึกษาการเรียนรู้ด้วยตัวอย่างมีป้ายกำกับร่วมกับตัวอย่างไม่มีป้ายกำกับสามารถแบ่งตามวัตถุประสงค์ในการใช้ตัวอย่างไม่มีป้ายกำกับได้เป็นสองแนวทาง ได้แก่ การเรียนรู้แบบกึ่งมีผู้สอน (Semi-supervised learning) และการเรียนรู้แบบแอ็กทิฟ (Active learning) การเรียนรู้ทั้งสองวิธีนำมาใช้ในสถานการณ์ที่ตัวอย่างมีป้ายกำกับมีจำนวนน้อยหรือไม่เพียงพอที่จะเรียนรู้ระบบที่มีประสิทธิภาพ แต่ตัวอย่างไม่มีป้ายกำกับนั้นมีอยู่เป็นจำนวนมาก ข้อแตกต่างระหว่างสองวิธีนี้ คือ ขั้นตอนวิธีเลือกตัวอย่างไม่มีป้ายกำกับมาใช้ และขั้นตอนวิธีติดป้ายกำกับให้แก่ตัวอย่างไม่มีป้ายกำกับ

ขั้นตอนแรกอาจใช้ประโยชน์จากตัวอย่างมีป้ายกำกับด้วยการสร้างตัวจำแนกตั้งต้นที่สร้างจากตัวอย่างมีป้ายกำกับเท่านั้น แล้วจึงปรับปรุงตัวจำแนกตั้งต้นนั้นให้มีประสิทธิภาพมากขึ้นโดยการเรียนรู้แบบกึ่งมีผู้สอนเลือกตัวอย่างไม่มีป้ายกำกับที่ตัวจำแนกตั้งต้นมั่นใจที่จะติดป้ายกำกับที่สุด (certain data) มาติดป้ายกำกับ แล้วนำตัวอย่างที่ติดป้ายกำกับใหม่นั้นมาสร้างตัวจำแนกใหม่ในรอบต่อไป ส่วนการเรียนรู้แบบแอ็กทิฟเลือกตัวอย่างไม่มีป้ายกำกับที่จะช่วยปรับปรุงตัวจำแนกตั้งต้นได้มากที่สุด หรือเลือกตัวอย่างที่ตัวจำแนกตั้งต้นไม่มั่นใจมากที่สุด (uncertain data) แล้วนำตัวอย่างนั้นมาติดป้ายกำกับโดยผู้ใช้หรือผู้ที่ทราบป้ายกำกับ (oracle)

ในหัวข้อ 2.1 อธิบายหลักการการเรียนรู้แบบกึ่งมีผู้สอน วิธีการเรียนรู้แบบกึ่งมีผู้สอนวิธีต่าง ๆ รวมทั้งสมมติฐานที่จำเป็นในการเรียนรู้แบบกึ่งมีผู้สอนแต่ละวิธี โดยเฉพาะการเรียนรู้แบบกึ่งมีผู้สอนวิธีเรียนรู้ด้วยตนเอง (Self-training) และการเรียนรู้แบบกึ่งมีผู้สอนวิธีจัดกลุ่มและติดป้าย (Cluster-and-label) ซึ่งเป็นสองวิธีการที่งานวิจัยนี้มุ่งเน้นศึกษา หัวข้อนี้อธิบายรวมถึงการจัดกลุ่มแบบกึ่งมีผู้สอน (Semi-supervised clustering) ซึ่งเป็นเทคนิคที่นำมาประยุกต์ใช้ในการวิเคราะห์ตัวอย่างในงานวิจัยนี้ด้วย และหัวข้อย่อยสุดท้ายอธิบายถึงความเสี่ยงของการใช้ตัวอย่างไม่มีป้ายกำกับในการเรียนรู้แบบกึ่งมีผู้สอน รวมถึงงานวิจัยที่ศึกษาวิธีลดความผิดพลาดจากการใช้ตัวอย่างไม่มีป้ายกำกับในการเรียนรู้กึ่งมีผู้สอนวิธีต่าง ๆ

ในหัวข้อ 2.2 อธิบายหลักการการเรียนรู้แบบแอ็กทิฟและงานวิจัยที่นำเสนอวิธีการเรียนรู้แบบแอ็กทิฟที่เป็นที่นิยม ทั้งการเรียนรู้แบบกึ่งมีผู้สอนและการเรียนรู้แบบแอ็กทิฟใช้ประโยชน์จากตัวอย่างไม่มีป้ายกำกับเหมือนกัน แต่ใช้ประโยชน์ในต่างด้าน ทั้งสองวิธีสามารถนำมาประยุกต์ใช้ร่วมกันเพื่อใช้ประโยชน์จากตัวอย่างไม่มีป้ายกำกับอย่างสูงสุดได้ (Settles, 2010) (Zhu, 2008) โดยตัวอย่างงานวิจัยที่ประยุกต์ทั้งสองวิธีมาใช้ร่วมกันนำเสนอในหัวข้อ 2.3

2.1 การเรียนรู้แบบกึ่งมีผู้สอน

การเรียนรู้แบบกึ่งมีผู้สอนนี้อยู่ระหว่างการเรียนรู้แบบมีผู้สอน (Supervised learning) และการเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning) การเรียนรู้แบบมีผู้สอนซึ่งเป็นวิธีที่ใช้ในปัญหาการเรียนรู้ของเครื่องส่วนใหญ่ ใช้ตัวอย่างมีป้ายกำกับคลาสเท่านั้นในการเรียนรู้ ส่วนการเรียนรู้แบบไม่มีผู้สอนใช้ตัวอย่างไม่มีป้ายกำกับคลาสซึ่งมีเพียงค่าคุณลักษณะเท่านั้น เพื่อหาสารสนเทศแฝงในข้อมูลไม่มีป้ายกำกับ เช่น การจัดกลุ่มข้อมูลด้วยเค-มีนส์ (K-means clustering) การเรียนรู้แบบกึ่งมีผู้สอนสามารถนำไปใช้ในสถานการณ์ที่ต้องการเรียนรู้ระบบเช่นเดียวกับการเรียนรู้แบบมีผู้สอน แต่จำนวนตัวอย่างมีป้ายกำกับมีไม่เพียงพอ ในขณะที่จำนวนตัวอย่างไม่มีป้ายกำกับนั้นมีจำนวนมาก หรืออาจใช้ในการเรียนรู้แบบไม่มีผู้สอน โดยใช้ตัวอย่างมีป้ายกำกับเพื่อกำหนดเงื่อนไขในการจัดกลุ่มข้อมูล

กำหนดให้แทนตัวเรียนรู้ด้วย L แทนตัวอย่างด้วย x แทนป้ายกำกับของตัวอย่างด้วย y ตัวอย่างทั้งหมดมีจำนวน n โดยแบ่งเป็นตัวอย่างที่ใช้ในการเรียนรู้ (training data) จำนวน t และตัวอย่างที่ใช้ในการทดสอบ (test data) T โดย $n = t + T$ ตัวอย่างที่ใช้ในการเรียนรู้แบบกึ่งมีผู้สอนมีสองประเภท ได้แก่ ตัวอย่างมีป้ายกำกับมีจำนวน l ตัวอย่างไม่มีป้ายกำกับมีจำนวน u โดยที่ $l << u$ และ $t = l + u$ ตัวอย่างมีป้ายกำกับ $(X_l, Y_l) = (x_{1:l}, y_{1:l})$ และตัวอย่างไม่มีป้ายกำกับ $X_u = x_{l+1:n}$ ตัวอย่างที่ใช้ในการทดสอบ $(X_T, Y_T) = (x_{1:T}, y_{1:T})$ การเรียนรู้แบบกึ่งมีผู้สอนใช้ตัวอย่าง (X_l, Y_l) และ X_u เพื่อสร้างตัวเรียนรู้หรือตัวจำแนก L

การเรียนรู้แบบกึ่งมีผู้สอนในงานวิจัยนี้เป็นการเรียนรู้แบบอุปนัย (inductive learning) โดยเรียนรู้เพื่อสร้างกฎ (general rule) สำหรับทำนายตัวอย่างที่ไม่เคยพบ (unseen data) การวัดประสิทธิภาพของตัวเรียนรู้อุปนัยจึงวัดบนตัวอย่างทดสอบ (X_T, Y_T) ที่แยกไว้ ตัวอย่างทดสอบนี้ไม่ได้นำมาใช้ในขั้นตอนการเรียนรู้ วิธีการเรียนรู้อีกวิธีหนึ่งคือการเรียนรู้แบบถ่ายโอน (transductive learning) (Vapnik, 1998) วิธีนี้ไม่ได้สร้างกฎสำหรับตัวอย่างที่ไม่เคยพบอย่างวิธีเรียนรู้แบบอุปนัย แต่วิธีนี้พยายามทำนายตัวอย่างทดสอบที่สนใจโดยนำข้อมูลค่าคุณลักษณะของตัวอย่างทดสอบมาพิจารณาร่วมในตอนการเรียนรู้ด้วย (Zhu, 2008) วิธีเรียนรู้แบบถ่ายโอนไม่แยกตัวอย่างเป็นตัวอย่างทดสอบแต่ตัวอย่างทั้งหมดถูกนำไปใช้ในขั้นตอนการเรียนรู้ ดังนั้นการวัดประสิทธิภาพของวิธีเรียนรู้แบบถ่ายโอนนี้จะพิจารณาความถูกต้องบนตัวอย่างไม่มีป้าย

กำกับ X_u นั้น

วิธีการเรียนรู้แบบกึ่งมีผู้สอนที่นิยมนำมาใช้ในปัญหาการจำแนกนั้นมีอยู่หลายวิธี เช่น วิธีเจเนเรทีฟโมเดล (Generative model) (Nigam et al., 2000) วิธีเรียนรู้ด้วยตนเอง (self-training) วิธีเรียนรู้โดยร่วมมือกันสองวิธี (co-training) (Blum and Mitchell, 1998) วิธีใช้พื้นฐานจากกราฟ (graph-based method) (Chapelle and Zien, 2005) (Belkin et al., 2006) ซัพพอร์ตเวกเตอร์แบบกึ่งมีผู้สอน (semi-supervised support vector machine: S3VM หรือ transductive support vector machine: TSVM) (Joachims, 1999) เป็นต้น

วิธีเรียนรู้แบบกึ่งมีผู้สอนแต่ละวิธีมีสมมติฐานเพื่อใช้ประโยชน์จากตัวอย่างไม่มีป้ายกำกับที่แตกต่างกันซึ่งเหมาะสมกับปัญหาและข้อมูลที่แตกต่างกัน ตัวอย่างเช่นวิธีเรียนรู้โดยร่วมมือกันสองวิธีใช้สมมติฐานเกี่ยวกับค่าคุณลักษณะ (feature หรือ attribute) ของตัวอย่าง โดยกำหนดว่าชุดข้อมูลมีค่าคุณลักษณะที่สามารถนำมาใช้ได้หลายมุมมอง (multi-view) โดยค่าคุณลักษณะนั้นมีความซ้ำซ้อนของมุมมองจึงสามารถแบ่งคุณลักษณะออกเป็นสองชุด แต่ละชุดต้องเพียงพอที่จะสร้างตัวเรียนรู้เพื่อเรียนรู้สมมติฐานบางส่วนที่สอดคล้องกับคลาสของตัวอย่างได้ ตัวอย่างเช่นในการจำแนกรายการโทรทัศน์ที่แตกต่างกันของมนุษย์ เราอาจจำแนกบางประเภทของรายการออกจากกันได้อย่างมั่นใจโดยดูจากภาพเท่านั้นหรือฟังจากเสียงเท่านั้น วิธีเรียนรู้โดยร่วมมือกันสองวิธีจึงแบ่งคุณลักษณะออกเป็นสองชุดซึ่งแต่ละชุดเพียงพอที่จะสร้างตัวเรียนรู้สำหรับสมมติฐานที่แตกต่างกันได้ โดยการสร้างตัวเรียนรู้ทั้งสองตัวนั้นใช้ตัวอย่างมีป้ายกำกับชุดเดียวกัน แต่อาจใช้ขั้นตอนวิธีในการเรียนรู้ที่แตกต่างกัน ใช้พารามิเตอร์ที่แตกต่างกัน หรือใช้ค่าคุณลักษ-ณะที่แตกต่างกัน และตัวเรียนรู้ทั้งสองตัวนี้จะถูกนำมาช่วยพิจารณาป้ายกำกับแก่ตัวอย่างไม่มีป้ายกำกับเพื่อติดป้ายกำกับตัวอย่างให้ได้ถูกต้องมากที่สุด ตัวอย่างที่ถูกติดป้ายกำกับด้วยตัวเรียนรู้หนึ่งจะถูกนำไปใช้เพื่อสอนตัวเรียนรู้อีกตัวหนึ่ง (Zhou and Li, 2010) วิธีเรียนรู้แบบร่วมมือกันสองวิธีนี้เป็นวิธีที่ยอมรับว่ามีประสิทธิภาพมากวิธีหนึ่ง แต่สมมติฐานในการนำวิธีนี้ไปใช้ซึ่งกำหนดว่าชุดข้อมูลมีค่าคุณลักษณะซ้ำซ้อนกันบางส่วน และสามารถแบ่งคุณลักษณะออกเป็นสองชุดที่สอดคล้องกับคลาสและคุณลักษณะทั้งสองชุดเป็นอิสระต่อกันแบบมีเงื่อนไข (conditional independence) นั้นพบได้ยากในชุดข้อมูลทั่วไป

วิธีการเรียนรู้แบบกึ่งมีผู้สอนที่ได้รับความสนใจมากอีกวิธีหนึ่งคือวิธีใช้พื้นฐานจากกราฟของตัวอย่าง โดยสร้างกราฟเพื่อเปลี่ยนการแทนข้อมูล (data representation) เพื่อหาขอบการตัดสินใจ (decision boundary) ของตัวอย่างแต่ละคลาส ซึ่งควรจะอยู่ในบริเวณที่ตัวอย่างมีความหนาแน่นน้อย (Chapelle and Zien, 2005) สมมติฐานสำคัญในวิธีนี้คือต้องสามารถสร้างกราฟที่เหมาะสมกับลักษณะของข้อมูลและสอดคล้องกับการกระจายตัวของคลาสได้ อีกวิธีหนึ่งที่ได้รับความสนใจมากเช่นกัน คือ วิธีซัพพอร์ตเวกเตอร์กึ่งมีผู้สอน โดยการใช้ตัวอย่างไม่มีป้าย

กำกับช่วยเลือกซัพพอร์ตเวกเตอร์ที่ดีที่สุด อย่างไรก็ดีด้วยคุณลักษณะของซัพพอร์ตเวกเตอร์-แมชชีนนั้นเหมาะสมกับการนำไปใช้แก้ปัญหาสองคลาสมากกว่าปัญหาหลายคลาส นอกจากนั้นวิธีใช้พื้นฐานจากกราฟและวิธีซัพพอร์ตเวกเตอร์กึ่งมีผู้สอนนั้นถูกออกแบบมาให้เหมาะกับการเรียนรู้แบบถ่ายโอน หรือวิธีการเรียนรู้ที่นำตัวอย่างที่ต้องการทราบคลาสมาร่วมพิจารณาในขั้นตอนการเรียนรู้มากกว่าการเรียนรู้แบบอุปนัย

งานวิจัยนี้เลือกศึกษาการเรียนรู้แบบกึ่งมีผู้สอนสองวิธี คือ วิธีเจเนเรทีฟโมเดลด้วยเทคนิคการจัดกลุ่มและติดป้ายและวิธีเรียนรู้ด้วยตนเอง ซึ่งทั้งสองวิธีเป็นวิธีที่ได้รับการยอมรับว่ามีประสิทธิภาพและมีสมมติฐานไม่ซับซ้อนจึงสามารถนำไปประยุกต์ใช้กับชุดข้อมูลทั่วไปได้ โดยรายละเอียดวิธีการและสมมติฐานของวิธีการจัดกลุ่มและติดป้ายนำเสนอในหัวข้อ 2.1.2 และวิธีเรียนรู้ด้วยตนเองนำเสนอในหัวข้อ 2.1.3

นอกจากการใช้ตัวอย่างไม่มีป้ายกำกับเพื่อช่วยสร้างตัวจำแนกแล้ว การเรียนรู้แบบกึ่งมีผู้สอนนั้นยังสามารถนำไปประยุกต์ใช้ร่วมกับปัญหาการจัดกลุ่มข้อมูลซึ่งเป็นปัญหาในการเรียนรู้แบบไม่มีผู้สอนได้ด้วย โดยในหัวข้อ 2.1.4 นำเสนอวิธีการจัดกลุ่มกึ่งมีผู้สอนวิธีต่าง ๆ และวิธีการจัดกลุ่มกึ่งมีผู้สอนที่นำมาประยุกต์ใช้ในงานวิจัยนี้

ทั้งนี้งานวิจัยเกี่ยวกับการเรียนรู้แบบกึ่งมีผู้สอนโดยละเอียดสามารถศึกษาเพิ่มเติมได้จากหนังสือของ Chapelle (Chapelle et al., 2006) งานทบทวนงานวิจัยการเรียนรู้แบบกึ่งมีผู้สอน (Zhu, 2008) และเอกสารสอนพื้นฐานการเรียนรู้แบบกึ่งมีผู้สอน (Zhu, 2007) ของ Zhu

2.1.1 สมมติฐานในการเรียนรู้แบบกึ่งมีผู้สอน

ตัวอย่างไม่มีป้ายกำกับนั้นจะมีประโยชน์ต่อการเรียนรู้ก็ต่อเมื่อมีการกำหนดสมมติฐานเพื่อใช้ประโยชน์จากตัวอย่างไม่มีป้ายกำกับนั้น สมมติฐานที่ใช้ในการเรียนรู้กึ่งมีผู้สอน เช่น

- สมมติฐานความราบรื่น (smoothness assumption) คือ คลาสของตัวอย่างที่อยู่ใกล้กันจะมีความคล้ายคลึงกัน
- สมมติฐานกลุ่มข้อมูล (cluster-assumption) คือ ตัวอย่างที่อยู่ถูกจัดกลุ่มอยู่ในกลุ่มเดียวกันจะมีคลาสเดียวกัน สมมติฐานนี้คล้ายกับสมมติฐานความราบรื่น โดยสมมติฐานนี้นำไปใช้ในการเรียนรู้แบบกึ่งมีผู้สอนหลายวิธี โดยเฉพาะวิธีจัดกลุ่มและติดป้าย
- สมมติฐานในบริเวณความหนาแน่นต่ำ (low-density assumption) คือ ขอบการตัดสินใจอยู่ในบริเวณที่ตัวอย่างมีความหนาแน่นต่ำ ๆ สมมติฐานนี้เป็นสมมติฐานสำคัญในการเรียนรู้

แบบกึ่งมีผู้สอนโดยการสร้างกราฟของตัวอย่างและการเรียนรู้กึ่งสอนด้วยซัพพอร์ตเวกเตอร์แมชชีน

- สมมติฐานมานิโฟลด์ (manifold assumption) คือ ตัวอย่างที่มีมิติมาก (high-dimensional data) นั้นอยู่บนมานิโฟลด์ในมิติที่ต่ำกว่า (low-dimension manifold) โดยฟังก์ชันการตัดสินใจนั้นอยู่บนมานิโฟลด์ในมิติต่ำ หากสามารถหาโครงสร้างของมานิโฟลด์ได้จะช่วยประมาณค่าคลาสของตัวอย่างไม่มีป้ายกำกับที่มีมิติมากนั้นได้ สมมติฐานนี้ใช้ในการเรียนรู้แบบกึ่งมีผู้สอนโดยการสร้างกราฟของตัวอย่าง การสร้างกราฟเสมือนการหามานิโฟลด์ในมิติที่ต่ำกว่าของตัวอย่าง โดยแสดงด้วยค่าความสัมพันธ์ของตัวอย่างบนกราฟ หากสามารถหากราฟความสัมพันธ์ของตัวอย่างและคลาสของตัวอย่างได้ จะสามารถกระจายป้ายกำกับคลาสให้แก่ตัวอย่างไม่มีป้ายกำกับตามค่าความสัมพันธ์บนกราฟได้ สมมติฐานนี้มักใช้กับการเรียนรู้แบบถ่ายโอน (transductive learning) หรือเมื่อตัวอย่างที่สนใจนั้นถูกนำมาพิจารณาร่วมด้วยในขั้นตอนการเรียนรู้

2.1.2 การเรียนรู้แบบกึ่งมีผู้สอนวิธีจัดกลุ่มและติดป้าย

วิธีจัดกลุ่มและติดป้ายเป็นเทคนิคหนึ่งของการเรียนรู้แบบกึ่งมีผู้สอนด้วยวิธีเจเนเรทีฟโมเดล (Nigam et al., 2000) Nigam นำเสนอวิธีใช้ตัวอย่างไม่มีป้ายกำกับร่วมกับตัวอย่างมีป้ายกำกับเพื่อจำแนกประเภทเอกสาร ซึ่งเป็นงานวิจัยเริ่มต้นสำคัญงานหนึ่งที่ทำให้การเรียนรู้แบบกึ่งมีผู้สอนได้รับความสนใจอย่างมาก

วิธีจัดกลุ่มและติดป้าย (Demiriz et al., 1999) (Dara et al., 2002) เป็นเทคนิคที่อยู่ในกลุ่มเจเนเรทีฟโมเดลเช่นเดียวกัน ขั้นตอนวิธีการจัดกลุ่มและติดป้ายเริ่มจากการจัดกลุ่มตัวอย่างทั้งหมดทั้งที่มีป้ายกำกับและไม่มีป้ายกำกับ หลังจากนั้นจึงติดป้ายกำกับตัวอย่างไม่มีป้ายกำกับในแต่ละกลุ่มตามคลาสของตัวอย่างมีป้ายกำกับส่วนใหญ่ในกลุ่มนั้น ขั้นตอนวิธีการจัดกลุ่มและติดป้ายแสดงอยู่ในขั้นตอนวิธีที่ 1

โดย (X_{tr}, Y_{tr}) คือตัวอย่างที่ใช้ในการเรียนรู้ โดยที่ $(X_{tr}, Y_{tr}) \leftarrow (X_l, Y_l) \cup (X_u, Y_u)$ C คือกลุ่มข้อมูล k คือจำนวนกลุ่มข้อมูล c_i คือตัวอย่างในกลุ่มที่ i โดยที่ $(c_1, c_2, ..., c_k) \in C$ (X_{new}, Y_{new}) คือชุดตัวอย่างมีป้ายกำกับใหม่

วิธีจัดกลุ่มและติดป้ายนี้ที่ได้รับความนิยมใช้ในหลายปัญหา เช่น การจำแนกประเภทเอกสาร (Nigam et al., 2000) (Zeng et al., 2003) (Zhang et al., 2015) การค้นหาโปรตีนโมเลกุน (Sugiyama et al., 2012) การจำแนกปริมาณการใช้งานโพรโทคอลในเครือข่าย (Grimaudo et al., 2014) เป็นต้น วิธีจัดกลุ่มและติดป้ายนี้ผู้วิจัยได้นำไปประยุกต์ใช้เพื่อแก้ปัญหาการ

Algorithm 1 Cluster-and-label

Intialize:
Cluster data (X_{tr}, Y_{tr}) into cluster C
$(X_{new}, Y_{new}) = (X_l, Y_l)$
for each c_i in C **do**
    Find majority class y_{major} of c_i
    for each (x_j, y_j) in c_i **do**
        if y_j is empty **then**
            y_j = y_{major}
            $(X_{new}, Y_{new}) = (X_{new}, Y_{new}) \cup (x_j, y_j)$
        end if
    end for
end for
Train final classifier L with (X_{new}, Y_{new})

จำแนกกึ่งมีผู้สอนเพื่อลดสิ่งรบกวนในภาพเอกสารภาษาไทยซึ่งสามารถลดสิ่งรบกวนในเอกสารภาษาไทยได้อย่างมีประสิทธิภาพ (Piroonsup and Sinthupinyo, 2010a) (Piroonsup and Sinthupinyo, 2010b) ข้อจำกัดหนึ่งของวิธีจัดกลุ่มและติดป้าย คือพารามิเตอร์ที่จำเป็นต้องกำหนด เพื่อให้สามารถจัดกลุ่มได้สอดคล้องกับการกระจายตัวของข้อมูล เช่น วิธีการจัดกลุ่ม วิธีการวัดระยะทาง และจำนวนกลุ่มข้อมูล เป็นต้น

2.1.3 การเรียนรู้แบบกึ่งมีผู้สอนวิธีเรียนรู้ด้วยตนเอง

การจำแนกกึ่งมีผู้สอนวิธีเรียนรู้ด้วยตนเองเป็นวิธีที่ถูกนำไปใช้อย่างแพร่หลายมากที่สุดวิธีหนึ่ง (Zhu, 2008) และน่าจะเป็นวิธีการเรียนรู้แบบกึ่งมีผู้สอนที่เก่าแก่ที่สุด แนวคิดการเรียนรู้ด้วยตนเองนี้ถูกนำเสนอในหลายงานวิจัย โดยปรากฏครั้งแรกในงานของ Scudder (Scudder, 1965) (Chapelle et al., 2006)

การจำแนกกึ่งมีผู้สอนวิธีเรียนรู้ด้วยตนเองใช้วิธีสร้างตัวจำแนกตั้งต้น (initial classifier) จากตัวอย่างมีป้ายกำกับ ตัวจำแนกตั้งต้นนั้นจะถูกนำมาใช้เพื่อติดป้ายกำกับให้แก่ตัวอย่างไม่มีป้ายกำกับที่ตัวจำแนกมั่นใจที่สุด จุดประสงค์เพื่อเพิ่มจำนวนตัวอย่างมีป้ายกำกับให้เพียงพอแล้วนำตัวอย่างมีป้ายกำกับชุดใหม่นั้นไปสร้างตัวจำแนกในขั้นตอนสุดท้าย (final classifier) ขั้นตอนวิธีการเรียนรู้ด้วยตนเองแสดงอยู่ในขั้นตอนวิธีที่ 2

วิธีเรียนรู้ด้วยตนเองนี้เป็นวิธีหนึ่งที่มีประสิทธิภาพและง่ายในการนำไปใช้ เนื่องจากไม่มีสมมติฐานที่กำหนด คุณลักษณะของชุดข้อมูลจึงสามารถนำไปใช้กับชุดข้อมูลได้หลากหลาย

Algorithm 2 Self-training

Initialize:
Labeled data (X_l, Y_l)
Unlabeled data (X_u)
Given $(X_{tr}, Y_{tr}) = (X_l, Y_l)$
while stopping criteria not met **do**
    Train classifier C_{int} from (X_{tr}, Y_{tr})
    Use C_{int} to predict class label Y_u of X_u
    Select confidence sample $(X_{conf}, Y_{conf}); (X_{conf}, Y_{conf}) \in (X_u, Y_u)$
    Remove selected unlabeled data $X_u \leftarrow X_u - X_{conf}$
    Combine newly labeled data $(X_{tr}, Y_{tr}) \leftarrow (X_l, Y_l) \cup (X_{conf}, Y_{conf})$
end while

เช่น การตรวจจับวัตถุ (Rosenberg et al., 2005) การรู้จำใบหน้า (Roli and Marcialis, 2006) การแจกแจงรูปประโยค (parsing) (McClosky et al., 2006) และการจำแนกคลื่นไฟฟ้าสมอง (Li et al., 2008) (Guan G., 2013) ปัญหาอนุกรมเวลา (time series) (Wei and Keogh, 2006) เป็นต้น นอกจากนี้วิธีเรียนรู้ด้วยตนเองยังสามารถนำไปใช้ร่วมกับวิธีการอื่นได้อีกด้วย เช่น การนำไปใช้ร่วมกับปัญหาการเรียนรู้ข้อมูลที่อ่อนไหวต่อมูลค่าของการทำนายแต่ละคลาส (cost-sensitive learning) (Liu et al., 2009) ใช้กับการรวมตัวเรียนรู้ตัวอื่น ๆ (Kumar Mallapragada et al., 2009) (Maulik and Chakraborty, 2011) หรือใช้กับระบบที่ใช้ตัวจำแนกหลายตัว (Multiple classifier systems) (Didaci and Roli, 2006) ได้อีกด้วย สมมติฐานที่สำคัญในการนำวิธีเรียนรู้ด้วยตนเองไปใช้ คือ ตัวจำแนกตั้งต้นที่ได้สามารถเลือกตัวอย่างไม่มีป้ายกำกับที่มั่นใจมาติดป้ายกำกับได้อย่างมีประสิทธิภาพ

2.1.4 การจัดกลุ่มแบบกึ่งมีผู้สอน

การจัดกลุ่มโดยทั่วไปเป็นการเรียนรู้แบบไม่มีผู้สอน ซึ่งพิจารณาค่าคุณลักษณะของตัวอย่างเพื่อวัดระยะ (distance measure) ระหว่างตัวอย่าง แล้วจึงจัดตัวอย่างที่อยู่ใกล้กันให้อยู่ในกลุ่มเดียวกัน การจัดกลุ่มข้อมูลแบบกึ่งมีผู้สอนนั้นพิจารณาทั้งค่าคุณลักษณะของตัวอย่างและเงื่อนไขเพิ่มเติมอื่น ๆ เช่น ป้ายกำกับคลาส เพื่อประกอบการจัดกลุ่มข้อมูล วิธีนี้จัดกลุ่มแบบกึ่งมีผู้สอนจึงได้ผลลัพธ์เป็นกลุ่มข้อมูลที่สอดคล้องกับทั้งการกระจายตัวของตัวอย่างและการกระจายตัวของคลาสของตัวอย่าง (class distribution)

ตัวอย่างการจัดกลุ่มแบบกึ่งมีผู้สอน เช่น การจัดกลุ่มข้อมูลแบบกึ่งมีผู้สอนโดยกำหนดตัวอย่างตั้งต้น (semi-supervised cluster by seeding) (Basu et al., 2002) เช่น การใช้ตัวอย่างมีป้ายกำกับคลาสช่วยในการกำหนดจุดศูนย์กลางของกลุ่มหรือเซนทรอยด์ (centroid)

ตั้งต้นของแต่ละกลุ่มข้อมูลในวิธีจัดกลุ่มเค-มีนส์ และการจัดกลุ่มแบบกึ่งมีผู้สอนโดยกำหนดเงื่อนไข (constraint) ระหว่างคู่ตัวอย่าง (Wagstaff et al., 2001) ซึ่งกำหนดเงื่อนไขให้บางคู่ตัวอย่างเป็นคู่ตัวอย่างที่ต้องอยู่ในกลุ่มเดียวกัน (must-link) หรือห้ามอยู่ในกลุ่มเดียวกัน (can not-link) เป็นต้น

2.1.5 ความเสี่ยงในการใช้ตัวอย่างไม่มีป้ายกำกับในการเรียนรู้แบบกึ่งมีผู้สอน

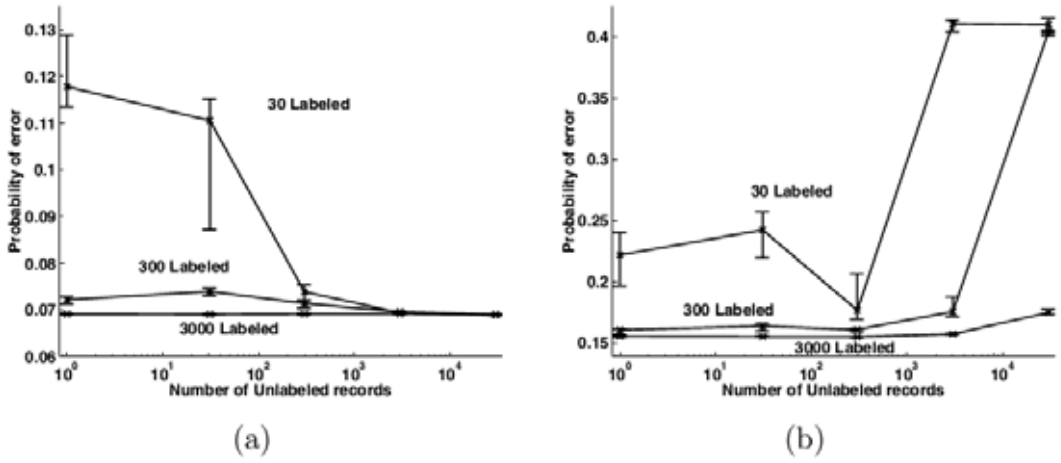
ถึงแม้ผลลัพธ์จากหลายงานวิจัยแสดงให้เห็นว่าการใช้ตัวอย่างไม่มีป้ายกำกับในการเรียนรู้แบบกึ่งมีผู้สอนสามารถเพิ่มประสิทธิภาพการเรียนรู้ได้ แต่การใช้ตัวอย่างไม่มีป้ายกำกับนั้นก็มีความเสี่ยง โดยการใช้ตัวอย่างไม่มีป้ายกำกับเข้ามาช่วยเรียนรู้อาจทำให้ได้ตัวเรียนรู้ที่มีความสามารถในการจำแนกแย่กว่าการตัวจำแนกที่สร้างจากตัวอย่างมีป้ายกำกับเท่านั้น

อาทิเช่นงานวิจัยที่พิสูจน์ทางทฤษฎีเกี่ยวกับประโยชน์ของตัวอย่างไม่มีป้ายกำกับ เช่น Singh (Singh et al., 2009) ซึ่งพิสูจน์ว่าการเรียนรู้กึ่งมีผู้สอนด้วยสมมติฐานกลุ่มข้อมูล (cluster assumption) นั้น ตัวอย่างไม่มีป้ายกำกับจะช่วยปรับปรุงการเรียนรู้ก็ต่อเมื่อความหนาแน่นของตัวอย่างจากแต่ละคลาสสามารถจำแนกได้โดยใช้ตัวอย่างไม่มีป้ายกำกับ m จำนวนแต่ไม่สามารถจำแนกได้เมื่อใช้แต่ตัวอย่างมีป้ายกำกับ n จำนวน เมื่อ $n < m$ แต่หากความหนาแน่นของตัวอย่างมีป้ายกำกับเพียงอย่างเดียวสามาถจำแนกข้อมูลได้แล้ว การใช้ตัวอย่างไม่มีป้ายกำกับจะไม่สามารถช่วยปรับปรุงการเรียนรู้นั้นให้ดีขึ้นได้ เช่นเดียวกับการศึกษาของ David (Ben-David et al., 2008) ที่พิสูจน์การเรียนรู้กึ่งมีผู้สอนด้วยสมมติฐานกลุ่มข้อมูลเช่นเดียวกัน ได้กล่าวว่าตัวอย่างไม่มีป้ายกำกับจะช่วยปรับปรุงการเรียนรู้ก็ต่อเมื่อมีสมมติฐานที่ชัดเจนเกี่ยวกับการกระจายตัวของคลาสของตัวอย่างเท่านั้น

งานวิจัยอื่น ๆ ที่ศึกษาความเสี่ยงในการเรียนรู้แบบกึ่งมีผู้สอน โดยแบ่งตามวิธีการเรียนรู้แบบกึ่งมีผู้สอน ได้แก่ วิธีเจเนเรทีฟโมเดล วิธีซัพพอร์ตเวกเตอร์กึ่งมีผู้สอน และวิธีเรียนรู้ด้วยตนเอง แสดงในหัวข้อต่อไปนี้

2.1.5.1 ความเสี่ยงในการเรียนรู้แบบกึ่งมีผู้สอนวิธีเจเนเรทีฟโมเดล

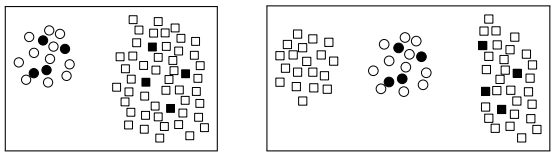
ในงานของ Nigam (Nigam et al., 2000) ซึ่งเสนอการใช้ตัวอย่างไม่มีป้ายกำกับด้วยอีเอ็มอัลกอริธึมซึ่งอยู่ในกลุ่มการเรียนรู้แบบกึ่งมีผู้สอนวิธีเจเนเรทีฟโมเดล ได้กล่าวถึงความน่าจะเป็นที่การใช้ตัวอย่างไม่มีป้ายกำกับส่งผลเสียต่อการสร้างตัวจำแนก ซึ่งอาจทำให้ได้ตัวเรียนรู้ที่มีคุณภาพแย่กว่าการเรียนรู้ด้วยตัวอย่างมีป้ายกำกับเท่านั้น หากสมมติฐานที่ใช้ไม่สอดคล้องกับการกระจายตัวของตัวอย่าง โดยผลการทดลองของ Cozman (Cozman and Cohen, 2002) (Chapelle et al., 2006) สนับสนุนในคำกล่าวนี้ Cozman ทดลองวัดประสิทธิภาพของการ



ภาพที่ 2.1: กราฟแสดงร้อยละความผิดพลาดของตัวเรียนรู้ที่สร้างจากการเรียนรู้แบบกึ่งมีผู้สอนด้วยนาอีฟเบย์บน (a) ชุดข้อมูลที่สมมติฐานสอดคล้องกับนาอีฟเบย์ และ (b) ชุดข้อมูลที่สมมติฐานไม่สอดคล้องกับนาอีฟเบย์ (Cozman and Cohen, 2002)

เรียนรู้ เมื่อสมมติฐานที่ใช้ในการสร้างแบบจำลองสอดคล้องและไม่สอดคล้องกับการกระจายตัวของตัวอย่าง โดยใช้ตัวจำแนกนาอีฟเบย์ในการเรียนรู้กึ่งมีผู้สอนบนชุดข้อมูลสังเคราะห์สองชุด ชุดข้อมูลแรกสร้างตัวอย่างด้วยสมมติฐานของนาอีฟเบย์ ซึ่งกำหนดให้คุณลักษณะทั้งหมด X_v เป็นอิสระต่อกันเมื่อคลาสของตัวอย่างคือ Y_v หรือ $p(X_v, Y_v) = p(Y_v)\Pi p(X_v i)$ ส่วนชุดข้อมูลที่สองสร้างด้วยสมมติฐานของแทน (TAN assumption) ซึ่งกำหนดให้ค่าของคุณลักษณะใด ๆ นั้นขึ้นต่อกันผลการทดลองในตัวอย่างชุดแรกที่สร้างโดยใช้สมมติฐานที่สอดคล้องกับแบบจำลองที่ใช้ โดยเมื่อเพิ่มจำนวนตัวอย่างไม่มีป้ายกำกับในการเรียนรู้ทำให้ได้ตัวเรียนรู้ที่มีประสิทธิภาพในการเรียนรู้ที่ดีขึ้น ต่างจากในตัวอย่างชุดที่สองที่พบว่าเมื่อเพิ่มจำนวนตัวอย่างไม่มีป้ายกำกับกลับส่งผลให้ประสิทธิภาพของตัวเรียนรู้แย่ลง ผลการทดลองเปรียบเทียบประสิทธิภาพตัวเรียนรู้บนสองชุดข้อมูลนั้นแสดงอยู่ในภาพที่ 2.1

การศึกษาของ Tian (Tian et al., 2004) พบว่าการกระจายตัวของตัวอย่างมีป้ายกำกับและตัวอย่างไม่มีป้ายกำกับส่งผลต่อประสิทธิภาพของการเรียนรู้แบบกึ่งมีผู้สอน หากการกระจายตัวของตัวอย่างสองกลุ่มนี้ไม่สอดคล้องกันจะส่งผลให้การใช้ตัวอย่างไม่มีป้ายกำกับนั้นส่งผลเสียต่อประสิทธิภาพของตัวเรียนรู้ โดยทดสอบประสิทธิภาพการเรียนรู้บนตัวอย่างสองชุดที่มีการกระจายตัวของตัวอย่างมีป้ายกำกับและตัวอย่างไม่มีป้ายกำกับดังภาพที่ 2.2 ผลการทดลองแสดงให้เห็นว่าการใช้ตัวอย่างไม่มีป้ายกำกับจะทำให้ได้ตัวจำแนกที่มีประสิทธิภาพดีขึ้น ก็ต่อเมื่อการกระจายตัวของตัวอย่างมีป้ายกำกับและตัวอย่างไม่มีป้ายกำกับนั้นสอดคล้องกัน แต่หากการกระ-



(a) (b)

ภาพที่ 2.2: การกระจายตัวของตัวอย่างคลาสวงกลมและสี่เหลี่ยม ตัวอย่างมีป้ายกำกับแสดงด้วยรูปวงกลมและสี่เหลี่ยมทึบ ตัวอย่างไม่มีป้ายกำกับแสดงด้วยรูปวงกลมและสี่เหลี่ยมโปร่ง (a) กรณีที่การกระจายตัวของตัวอย่างมีป้ายกำกับและตัวอย่างไม่มีป้ายกำกับสอดคล้องกัน (b) กรณีที่ไม่สอดคล้องกัน (Cozman and Cohen, 2002)

จายของตัวอย่างมีป้ายกำกับและตัวอย่างไม่มีป้ายกำกับแตกต่างกัน การเพิ่มตัวอย่างไม่มีป้ายกำกับในการเรียนรู้อาจจะส่งผลให้ตัวจำแนกมีประสิทธิภาพแย่ลง ซึ่งสอดคล้องกับสมมติฐานที่ใช้ในงานวิจัยนี้ที่เสนอการปรับปรุงชุดตัวอย่างมีป้ายกำกับสำหรับการเรียนรู้แบบกึ่งมีผู้สอน โดยวิเคราะห์การกระจายตัวของตัวอย่างมีป้ายกำกับและตัวอย่างไม่มีป้ายกำกับ

2.1.5.2 ความเสี่ยงในการเรียนรู้แบบกึ่งมีผู้สอนวิธีซัพพอร์ตเวกเตอร์แมชชีนกึ่งมีผู้สอน

การศึกษาของ Li (Li and Zhou, 2015) พบว่าการใช้ตัวอย่างไม่มีป้ายกำกับเพื่อสร้างซัพพอร์ตเวกเตอร์แมชชีน มีความน่าจะเป็นที่จะทำให้ประสิทธิภาพการจำแนกแย่กว่าซัพพอร์ตเวกเตอร์แมชชีนที่สร้างจากตัวอย่างมีป้ายกำกับเท่านั้น และได้เสนอวิธีการเรียนรู้แบบกึ่งมีผู้สอนด้วยซัพพอร์ตเวกเตอร์แมชชีนที่ปลอดภัย (safe semi-supervised support vector machine: S4VMs) วัตถุประสงค์เพื่อสร้างซัพพอร์ตเวกเตอร์แมชชีนแบบกึ่งมีผู้สอนที่มีประสิทธิภาพดีมากกว่าหรือเท่ากับซัพพอร์ตเวกเตอร์แมชชีนที่สร้างจากตัวอย่างมีป้ายกำกับเท่านั้น โดยวิธีนี้หาตัวจำแนกทั้งหมดที่เป็นไปได้หรือหาระนาบการตัดสินใจขอบกว้างที่อยู่บริเวณความหนาแน่นน้อย (large-margin low-density separators) ทั้งหมดที่เป็นไปได้ แล้วเลือกติดป้ายกำกับตัวอย่างที่ทำให้ตัวจำแนกนี้มีประสิทธิภาพดีขึ้น

2.1.5.3 ความเสี่ยงในการเรียนรู้แบบกึ่งมีผู้สอนด้วยการติดป้ายกำกับด้วยตนเอง (self-labeling)

การติดป้ายกำกับด้วยตนเอง (self-labeling) เป็นขั้นตอนหนึ่งในการเรียนรู้แบบกึ่งมีผู้สอนหลายวิธี เช่น วิธีเรียนรู้ด้วยตนเอง วิธีเรียนรู้แบบร่วมมือกันสองวิธี (co-training) และวิธีเรียนรู้แบบร่วมมือกันสามวิธี (tri-training) กระบวนการติดป้ายกำกับด้วยตนเองนั้นสร้างตัวจำแนกตั้งต้นเพื่อนำมาใช้ติดป้ายกำกับตัวอย่างไม่มีป้ายกำกับที่เหลือ อย่างไรก็ดีการสร้างตัวจำแนกตั้ง-

ต้นนั้นใช้ตัวอย่างมีป้ายกำกับที่มีอยู่จำนวนน้อยและอาจไม่เพียงพอที่จะสร้างตัวจำแนกตั้งต้นที่มีประสิทธิภาพ การใช้ตัวจำแนกตั้งต้นนี้ในการติดป้ายกำกับมีโอกาสที่จะติดป้ายกำกับผิดพลาดสูง

งานวิจัยส่วนใหญ่จึงเสนอวิธีปรับปรุงการติดป้ายกำกับด้วยตนเองโดยพยายามลดตัวอย่างที่ติดป้ายกำกับไม่ถูกต้องด้วยเทคนิคต่าง ๆ อาทิเช่น

- วิธีเซตเรด (self-training with editing: SETRED) (Li and Zhou, 2005) ที่อนุญาตให้แก้ไขตัวอย่างที่น่าจะถูกติดป้ายกำกับผิด
- วิธีเอสเอ็นเอ็นอาร์ซีอี (self-training nearest neighbor rule using cut edges: SNNRCE) (Wang et al., 2010) ซึ่งใช้กราฟของตัวอย่างเพื่อนบ้านในการตัดสินใจติดป้ายกำกับในช่วงต้นของกระบวนการที่มีโอกาสติดป้ายกำกับผิด
- วิธีใช้ตัวกรองสิ่งรบกวน (noise filters) (Triguero et al., 2014) โดยศึกษาการใช้ตัวกรองสิ่งรบกวนกรองตัวอย่างที่น่าจะติดป้ายกำกับผิดออกไป
- วิธีร่วมมือกันสองวิธีแบบเป็นประชาธิปไตย (Democratic co-learning) ซึ่งใช้ตัวเรียนรู้จากหลายอัลกอริธึมช่วยติดป้ายกำกับโดยเลือกติดป้ายกำกับตามตัวเรียนรู้ส่วนใหญ่ เป็นต้น

จากผลการศึกษาของ Triguero (Triguero and Salvador García, 2015) ซึ่งทดลองเปรียบเทียบประสิทธิภาพของวิธีติดป้ายกำกับด้วยตนเองที่กล่าวมาข้างต้นและวิธีอื่น ๆ ที่มีผู้นำเสนอไว้รวม 14 วิธีการ บนชุดข้อมูลจากฐานข้อมูลยูซีไอ (UCI) และคีล (KEEL) รวมทั้งสิ้น 55 ชุดข้อมูล ด้วยตัวเรียนรู้เพื่อนบ้านใกล้ที่สุดเคตัว ต้นไม้ตัดสินใจ นาอีฟเบย์และซัพพอร์ตเวกเตอร์แมชชีน ผลการศึกษาพบว่าไม่มีวิธีใดที่ดีที่สุดและวิธีเรียนรู้ด้วยตนเองแบบมาตรฐานเป็นวิธีที่ดีอย่างโดดเด่นในหลายชุดข้อมูล

การศึกษาของ Guo (Guo et al., 2010) ทดลองเปรียบเทียบประสิทธิภาพของการเรียนรู้บนตัวอย่างมีป้ายกำกับเท่านั้นกับการเรียนรู้แบบกึ่งมีผู้สอน 3 วิธี ได้แก่ วิธีเรียนรู้ด้วยตนเอง วิธีเรียนรู้แบบร่วมมือกันสองวิธี และวิธีโคอีเอ็ม โดยใช้ตัวจำแนกนาอีฟเบย์และตัวจำแนกเบย์ (Bayes) 6 ชนิด ผลการทดลองพบว่าในหลายชุดข้อมูลการใช้ตัวอย่างไม่มีป้ายกำกับในการเรียนรู้แบบกึ่งมีผู้สอนกลับทำให้ได้ตัวจำแนกที่มีประสิทธิภาพในการจำแนกแย่กว่าตัวจำแนกที่สร้างจากตัวอย่างมีป้ายกำกับเท่านั้น นอกจากนี้ยังพบว่าในบางชุดข้อมูลการเพิ่มตัวอย่างที่ติดป้ายกำกับอย่างถูกต้องแก่ตัวจำแนกอาจไม่ได้ส่งผลให้ตัวจำแนกมีประสิทธิภาพในการจำแนกดีขึ้นเสมอไป ในงานต่อมาของ Guo (Guo et al., 2011) จึงเสนอวิธีปรับปรุงการติดป้ายกำกับด้วยตนเองโดยนอกจากจะเลือกติดป้ายกำกับแก่ตัวอย่างที่ตัวจำแนกตั้งต้นมั่นใจที่จะติดป้ายกำกับที่สุดแล้ว

ตัวอย่างที่ติดป้ายกำกับใหม่นั้นต้องต้องไม่ทำให้ประสิทธิภาพในการจำแนกแย่ลง เมื่อประเมินประสิทธิภาพการจำแนกบนตัวอย่างมีป้ายกำกับตั้งต้นนั้น วิธีนี้ทำให้ได้ตัวจำแนกที่ประสิทธิภาพในการจำแนกไม่แย่ไปกว่าตัวจำแนกที่ใช้ตัวอย่างมีป้ายกำกับเท่านั้น อย่างไรก็ดีการทดลองในงานนี้ใช้ตัวจำแนกเบย์เท่านั้น และผลการทดลองพบว่ามีเพียงในสามของชุดข้อมูลทั้งหมดที่ประสิทธิภาพในการจำแนกด้วยวิธีของ Guo นี้ดีกว่าการใช้ตัวอย่างมีป้ายกำกับเท่านั้น

2.2 การเรียนรู้แบบแอ็กทิฟ

การเรียนรู้แบบแอ็กทิฟ เป็นวิธีการเรียนรู้ที่อนุญาตให้ตัวเรียนรู้ถามป้ายกำกับคลาสของตัวอย่างที่ตัวเรียนรู้ไม่มั่นใจได้ การเรียนรู้แบบแอ็กทิฟนี้มักจะสร้างตัวเรียนรู้ตั้งต้นจากตัวอย่างมีป้ายกำกับ แล้วเลือกตัวอย่างเพิ่มเติมมาปรับปรุงปรับปรุงประสิทธิภาพตัวเรียนรู้นี้ให้ได้มากที่สุด วัตถุประสงค์ของการเรียนรู้แบบแอ็กทิฟเพื่อสร้างตัวเรียนรู้ที่มีประสิทธิภาพ โดยสอบถาม (query) ป้ายกำกับจากผู้ทราบคำตอบ (oracle) หรือผู้ใช้ในจำนวนที่เหมาะสม (Settles, 2010) โดยสถานการณ์ (scenario) ในการสอบถามอาจใช้ตัวอย่างสังเคราะห์ด้วยวิธีสังเคราะห์ตัวอย่างที่เป็นสมาชิกของชุดข้อมูล (membership query synthesis) หรือสอบถามโดยเลือกจากตัวอย่างไม่มีป้ายกำกับด้วยวิธีสอบถามทีละหนึ่งตัวอย่าง (stream-based selective sampling) หรือวิธีสอบถามจากกลุ่มตัวอย่าง (pool-based sampling) โดยประเมินตัวอย่างทั้งกลุ่มแล้วจึงเลือกตัวอย่างที่ดีที่สุดในกลุ่มนั้น

ในปัญหาการจำแนกมีวัตถุประสงค์เพื่อสร้างตัวจำแนกที่มีความผิดพลาดโดยทั่วไป (generalized error) น้อยที่สุด ดังนั้นการเลือกตัวอย่างมาปรับปรุงตัวจำแนกควรเลือกตัวอย่างช่วยลดความเสี่ยงที่จะทำนายผิดพลาดของตัวจำแนกได้มากที่สุด หรือเลือกตัวอย่างที่ทำให้ค่าความผิดพลาดของตัวจำแนกในอนาคตน้อยที่สุด แม้หลักการในการเลือกตัวอย่างดังกล่าวจะชัดเจนตรงไปตรงมาแต่การคำนวณเพื่อประมาณค่าความผิดพลาดในอนาคตจำเป็นต้องใช้การคำนวณที่สูงมาก ทั้งเพื่อประมาณค่าความผิดพลาดในอนาคตของตัวอย่างไม่มีป้ายกำกับทั้งหมดที่มีอยู่จำนวนมาก อาจต้องสอนแบบจำลองใหม่ (re-train) สำหรับทุก ๆ ตัวอย่าง (Settles, 2010) จึงมีผู้นำเสนอกลยุทธ์ (strategy) สำหรับสอบถามเพื่อเลือกตัวอย่างที่ดีที่สุดหลายวิธี เช่น วิธีเลือกตัวอย่างที่ไม่มั่นใจมากที่สุด (uncertainty sampling) และวิธีสอบถามโดยคณะกรรมการ (query by committee)

วิธีเลือกตัวอย่างที่ไม่มั่นใจมากที่สุดเลือกตัวอย่างที่ใกล้กับขอบการตัดสินใจ (decision boundary) มากที่สุด เช่น การเรียนรู้แบบแอ็กทิฟด้วยซัพพอร์ตเวกเตอร์แมชชีนจะเลือกตัวอย่างที่ใกล้กับระนาบตัดสินใจที่สุด (Tong and Koller, 2002) (Tong and Chang, 2001) การจำแนกตัวอย่างสองคลาสด้วยขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดเลือกตัวอย่างที่ระยะห่างจากตัวอย่างไปยังเพื่อนบ้านที่ใกล้ที่สุดทั้งสองคลาสนั้นใกล้เคียงกัน หรือพิจารณาจากตัวเรียนรู้ปัจจุบันโดย

เลือกตัวอย่างที่ตัวเรียนรู้ปัจจุบันมั่นใจที่จะติดป้ายกำกับน้อยที่สุด (Tur et al., 2005) เป็นต้น

วิธีสอบถามโดยคณะกรรมการสร้างคณะกรรมการหรือกลุ่มของตัวเรียนรู้จากตัวอย่างมีป้ายกำกับชุดเดียวกัน โดยแต่ละตัวเรียนรู้อาจใช้คุณลักษณะที่แตกต่างกัน (Muslea et al., 2006) หรือใช้ขั้นตอนวิธีที่แตกต่างกัน กลุ่มของตัวเรียนรู้นี้จะถูกนำมาใช้เพื่อทำนายป้ายกำกับแก่ตัวอย่างไม่มีป้ายกำกับ ตัวอย่างที่คณะกรรมการตอบป้ายกำกับคลาสไม่ตรงกันหรือมีระดับความขัดแย้ง (level of disagreement) สูงจะถูกเลือกเพื่อนำไปถามป้ายกำกับคลาสจากผู้ใช้ (Zhou et al., 2006) (Muslea et al., 2006)

ทั้งวิธีการเลือกตัวอย่างที่ไม่มั่นใจที่สุดและการเลือกโดยใช้คณะกรรมการพิจารณาประสิทธิภาพแต่ละตัวอย่างเท่านั้น ทั้งสองวิธีไม่ได้พิจารณาการกระจายตัวของตัวอย่าง ทำให้วิธีดังกล่าวอาจเลือกตัวอย่างที่ผิดปกติ (outliers) ซึ่งไม่ใช่กลุ่มที่เป็นตัวแทนข้อมูลมาสอบถาม จึงมีผู้วิจัยนำเสนอวิธีเลือกตัวอย่างที่เป็นตัวแทนของข้อมูล (density-based sampling หรือ representative sampling) ซึ่งเลือกตัวอย่างโดยพิจารณาการกระจายตัวของตัวอย่างประกอบด้วย โดยเลือกตัวอย่างที่เป็นตัวแทนของกลุ่มข้อมูลเพื่อหลีกเลี่ยงตัวอย่างที่เป็นตัวอย่างผิดปกติในชุดข้อมูลนั้น วิธีการเลือกตัวอย่างที่เป็นตัวแทนของข้อมูลนี้สามารถนำไปใช้ร่วมกับวิธีอื่น ๆ ได้ด้วย เช่น การเลือกตัวอย่างที่ไม่มั่นใจด้วยการให้น้ำหนักจากความหนาแน่น (density-weighted uncertainty sampling: DWUS) (Zhu et al., 2008) ซึ่งพิจารณาตัวอย่างที่ตัวเรียนรู้มั่นใจน้อยที่สุดและค่าเป็นตัวแทนของข้อมูลด้วย (Nguyen and Smeulders, 2004) และ (Donmez et al., 2007)

รายละเอียดการเรียนรู้แบบแอ็กทิฟสามารถศึกษาเพิ่มเติมได้จากเอกสารทบทวนงานวิจัยของ Settles (Settles, 2010)

2.3 งานวิจัยที่ประยุกต์ใช้การเรียนรู้แบบกึ่งมีผู้สอนร่วมกับการเรียนรู้แบบแอ็กทิฟ

ดังที่กล่าวไปข้างต้นว่าการเรียนรู้แบบกึ่งมีผู้สอนและการเรียนรู้แบบแอ็กทิฟใช้ในสถานการณ์เดียวกัน คือ เมื่อชุดข้อมูลมีตัวอย่างมีป้ายกำกับจำนวนน้อยมาก ๆ แต่มีตัวอย่างไม่มีป้ายกำกับอยู่เป็นจำนวนมาก ข้อแตกต่างระหว่างสองวิธีนี้ คือ การเรียนรู้แบบกึ่งมีผู้สอนเลือกตัวอย่างที่ตัวเรียนรู้มั่นใจที่สุดมาติดป้ายกำกับ ส่วนการเรียนรู้แบบแอ็กทิฟพยายามหาข้อมูลที่ตัวเรียนรู้ปัจจุบันยังมีไม่เพียงพอจากตัวอย่างไม่มีป้ายกำกับ ทั้งสองวิธีสามารถนำมาใช้ร่วมกันเพื่อปรับปรุงประสิทธิภาพของตัวเรียนรู้โดยใช้ประโยชน์จากตัวอย่างไม่มีป้ายกำกับได้อย่างสูงสุด (Settles, 2010)

งานวิจัยที่ใช้การเรียนรู้แบบกึ่งมีผู้สอนและการเรียนรู้แบบแอ็กทิฟร่วมกันจึงเป็นแนวทาง

ที่ได้รับความสนใจมาก ตัวอย่างเช่นในงานของ Muslea นำเสนอวิธีโคอีเอ็มที (Co-EMT) (Muslea et al., 2002) ซึ่งรวมการเรียนรู้แบบกึ่งมีผู้สอนวิธีโคอีเอ็ม (Co-EM) กับการเรียนรู้แบบแอ็กทิฟวิธีโคเทสติง (Co-Testing) วิธีโคอีเอ็มพัฒนามาจากการเรียนรู้แบบกึ่งมีผู้สอนโดยร่วมมือกันสองวิธี (Co-Training) ซึ่งสร้างตัวเรียนรู้สองตัวจากค่าคุณลักษณะต่างชุดกันด้วยขั้นตอนวิธีอีเอ็ม ตัวอย่างที่มั่นใจและถูกติดป้ายกำกับด้วยตัวเรียนรู้หนึ่งจะถูกนำไปใช้สอนตัวเรียนรู้อีกตัวหนึ่ง ส่วนวิธีโคเทสติงเลือกตัวอย่างที่ตัวเรียนรู้ทั้งสองตัวให้คำตอบแตกต่างกันมาถามป้ายกำกับจากผู้ใช้ ผลการทดลองพบว่าวิธีโคอีเอ็มทีที่รวมระหว่างสองวิธีข้างต้นนี้สามารถจำแนกตัวอย่างโดยมีค่าความถูกต้องสูงกว่าการเรียนรู้แบบมีผู้สอนด้วยวิธีอีเอ็ม การเรียนรู้แบบกึ่งมีผู้สอนวิธีโคอีเอ็ม และการเรียนรู้แบบแอ็กทิฟวิธีโคเทสติง (Muslea et al., 2002)

ขั้นตอนวิธีของวิธีโคอีเอ็มทีนำเสนอในขั้นตอนวิธีที่ 3 โดยกำหนดให้ $V1$ และ $V2$ แทนเซตย่อยของคุณลักษณะสองชุดที่ซ้ำซ้อนกัน T คือขั้นตอนวิธีในการเรียนรู้ $h1$ และ $h2$ คือสมมติฐานที่ได้จากการเรียนรู้คุณลักษณะแต่ละส่วน L แทนตัวอย่างมีป้ายกำกับ U แทนตัวอย่างไม่มีป้ายกำกับ N คือจำนวนครั้งในการสอบถามป้ายกำกับ และ s คือจำนวนรอบในการเรียนด้วยวิธีโคอีเอ็ม

Algorithm 3 Co-EMT

1: **while** number of queries $\leq N$ **do**
2: Run Co-EM(T, $V1$, $V2$, L, U, s) to learn $h1$ and $h2$
3: $query = \{x \in U, h1(x) \neq h2(x)\}$.
4: Label $query$
5: Move newly labeled sample from U to T.
6: **end while**
7: Create final classifier that combines the prediction from $h1$ and $h2$.

ตัวอย่างงานวิจัยที่รวมระหว่างการเรียนรู้แบบกึ่งมีผู้สอนและวิธีเรียนรู้แบบแอ็กทิฟในปัญหาต่าง ๆ เช่น การรู้จำเสียงพูด (Tur et al., 2005) การค้นคืนรูปภาพ (Zhou et al., 2006) (Zhou and Li, 2010) ซึ่งผลการทดลองพบว่าการรวมตัวเรียนรู้โดยพิจารณาความมั่นใจของตัวเรียนรู้นี้ สามารถปรับปรุงประสิทธิภาพตัวเรียนรู้ได้ดีกว่าการใช้การเรียนรู้แบบแอ็กทิฟหรือการเรียนรู้แบบกึ่งมีผู้สอนเพียงวิธีเดียว

บทที่ 3

งานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้เสนอการใช้การจัดกลุ่มข้อมูลเพื่อปรับปรุงการติดป้ายกำกับตัวอย่างไม่มีป้ายกำกับในการเรียนรู้แบบกึ่งมีผู้สอน โดยแบ่งเป็นสองส่วน งานวิจัยส่วนแรกเสนอวิธีปรับปรุงการติดป้ายกำกับสำหรับการเรียนรู้กึ่งมีผู้สอนวิธีจัดกลุ่มและติดป้าย งานวิจัยที่เกี่ยวข้องจึงเป็นงานที่ใช้การจัดกลุ่มข้อมูลเพื่อช่วยติดป้ายกำกับตัวอย่างในการเรียนรู้กึ่งมีผู้สอน นำเสนอในหัวข้อที่ 3.1

งานวิจัยในส่วนที่สองเสนอการวิเคราะห์กลุ่มข้อมูลเพื่อปรับปรุงตัวอย่างมีป้ายกำกับสำหรับการเรียนรู้กึ่งมีผู้สอน อย่างไรก็ดีงานวิจัยที่เกี่ยวกับการวิเคราะห์คุณภาพของชุดตัวอย่างมีป้ายกำกับสำหรับการเรียนรู้กึ่งสอนนั้นยังไม่ได้รับความสนใจมากนัก เท่าที่ผู้วิจัยสืบค้นไม่พบงานวิจัยที่นำเสนอการวิเคราะห์คุณลักษณะของตัวอย่างมีป้ายกำกับก่อนที่จะนำตัวอย่างชุดนั้นมาใช้ในการเรียนรู้กึ่งมีผู้สอน งานวิจัยนี้จึงเป็นงานแรกที่ใช้การจัดกลุ่มข้อมูลเพื่อวิเคราะห์คุณภาพของชุดตัวอย่างมีป้ายกำกับ งานวิจัยที่ใกล้เคียงสำหรับงานวิจัยส่วนที่สองจึงเกี่ยวกับการใช้การจัดกลุ่มข้อมูลในการเรียนรู้แบบแอ็กทิฟ ซึ่งนำเสนอในหัวข้อที่ 3.2

3.1 การใช้การจัดกลุ่มข้อมูลเพื่อช่วยติดป้ายกำกับตัวอย่าง

งานวิจัยเกี่ยวกับการใช้การจัดกลุ่มข้อมูลกับการเรียนรู้แบบกึ่งสอนมีวัตถุประสงค์ในการใช้การจัดกลุ่มข้อมูลเพื่อติดป้ายกำกับให้แก่ตัวอย่างที่มั่นใจที่สุด (certain data) อาทิเช่น การเรียนรู้แบบกึ่งสอนด้วยการจัดกลุ่มและติดป้าย (cluster-and-label) (Dara et al., 2002) ซึ่งรายละเอียดของวิธีการนี้อธิบายอยู่ในหัวข้อที่ 2.1.2 อย่างไรก็ดีผลลัพธ์จากขั้นตอนการจัดกลุ่มอาจได้กลุ่มที่ประกอบไปด้วยตัวอย่างมีป้ายกำกับมากกว่าหนึ่งคลาส หรือกลุ่มคลาสปะปน (mixed-class cluster) แนวทางการติดป้ายกำกับตัวอย่างในกลุ่มคลาสปะปนนี้มีสองแนวทาง แนวทางแรกคือการไม่ติดป้ายกำกับตัวอย่างในกลุ่มคลาสปะปน Dara et al. (2002) ซึ่งจากการทดลองบนชุดข้อมูลการลดสิ่งรบกวนในเอกสารภาษาไทยพบว่ามีตัวอย่างถึงร้อยละ 60 อยู่ในกลุ่มคลาสปะปน หากไม่ใช้ตัวอย่างในกลุ่มนี้จะลดจำนวนตัวอย่างที่น่าจะมีประโยชน์ต่อการเรียนรู้ของตัวจำแนกลงไปเป็นจำนวนมาก

อีกแนวทางหนึ่งสำหรับการติดป้ายกำกับตัวอย่างในกลุ่มคลาสปะปนคือ การเลือกติดป้ายกำกับตามคลาสของตัวอย่างมีป้ายกำกับส่วนใหญ่ Demiriz et al. (1999) ถึงแม้วิธีนี้จะใช้ประโยชน์จากตัวอย่างไม่มีป้ายกำกับได้สูงสุด แต่การติดป้ายกำกับโดยเลือกตามคลาสส่วนใหญ่เท่า

นั้นโดยไม่พิจารณาตัวอย่างในคลาสส่วนน้อย อาจส่งผลให้ตัวอย่างบางตัวถูกติดป้ายกำกับไม่ถูกต้อง เมื่อนำตัวอย่างที่มีป้ายกำกับที่ไม่ถูกต้องนั้นไปใช้สร้างตัวจำแนกสุดท้าย ย่อมส่งผลเสียต่อประสิทธิภาพของตัวจำแนกสุดท้ายอย่างหลีกเลี่ยงไม่ได้

ส่วนแรกของงานวิจัยนี้จึงเสนอวิธีการปรับปรุงการเรียนรู้แบบกึ่งสอนวิธีจัดกลุ่มและติดป้ายเพื่อปรับปรุงการติดป้ายกำกับในกลุ่มคลาสปะปน งานวิจัยที่เสนอวิธีปรับปรุงการเรียนรู้กึ่งมีผู้สอนแบบจัดกลุ่มและติดป้ายซึ่งใกล้เคียงกับงานที่นำเสนอ ได้แก่ PRC-Tree (Su et al., 2010) ซึ่งปรับปรุงขั้นตอนการจัดกลุ่มโดยใช้การจัดกลุ่มแบบลำดับขั้น (Hierarchical clustering) แบบบนลงล่าง (Top-down) วิธีนี้กำหนดให้ข้อมูลทั้งหมดอยู่ในกลุ่มเดียวกันในตอนเริ่มต้น จากนั้นพิจารณาความปะปนกันของคลาสของตัวอย่างมีป้ายกำกับ หากในกลุ่มตัวอย่างประกอบไปด้วยตัวอย่างมีป้ายกำกับต่างคลาสเกินกว่าค่าขีดแบ่งที่กำหนดให้จัดกลุ่มเป็นกลุ่มย่อยด้วยวิธีการเค-มีนส์ ซึ่งกำหนดศูนย์กลางของกลุ่มย่อยด้วยศูนย์กลางของตัวอย่างมีป้ายกำกับแต่ละคลาส และและกำหนดจำนวนกลุ่มย่อยเท่ากับจำนวนคลาสที่มีการปนกัน แล้วจึงติดป้ายกำกับตัวอย่างไม่มีป้ายกำกับตามคลาสของตัวอย่างมีป้ายกำกับในกลุ่มนั้น หรือหากกลุ่มย่อยนั้นไม่มีตัวอย่างมีป้ายกำกับจะติดป้ายกำกับตามกลุ่มบรรพบุรุษ (parent)

อีกงานวิจัยที่คล้ายคลึงกันคือ TESC (Zhang et al., 2015) มีแนวคิดใกล้เคียงกับงานข้างต้น (Su et al., 2010) ต่างกันที่งานของ Zhang นี้ใช้การจัดกลุ่มแบบล่างขึ้นบน (Bottom-up) วิธีนี้กำหนดให้ตัวอย่างแต่ละตัวอยู่แยกกันในแต่ละกลุ่มในตอนเริ่มต้น จากนั้นรวมกลุ่มที่ใกล้ที่สุดหากทั้งสองกลุ่มมีคลาสเดียวกัน (หรือเป็นกลุ่มไม่มีคลาสเช่นเดียวกัน) โดยจะไม่รวมกลุ่มหากกลุ่มที่ใกล้ที่สุดนั้นประกอบไปด้วยตัวอย่างที่มีป้ายกำกับต่างคลาสกัน สุดท้ายจึงติดป้ายกำกับตัวอย่างไม่มีป้ายกำกับตามกลุ่มที่ประกอบไปด้วยตัวอย่างมีป้ายกำกับที่ใกล้ที่สุด

ทั้งสองงานนี้ใช้การจัดกลุ่มแบบกึ่งสอนเพื่อกำหนดลักษณะของกลุ่มข้อมูลโดยทั้งสองงานนี้ใช้ป้ายกำกับคลาสเป็นเงื่อนไขในการจัดกลุ่มกึ่งมีผู้สอนสำหรับตัวอย่างทั้งหมด แตกต่างจากในงานวิจัยนี้ที่ใช้การจัดกลุ่มปกติแต่ปรับปรุงเฉพาะกลุ่มคลาสปะปนเท่านั้น นอกจากนี้ทั้งสองวิธีไม่ใช้ตัวอย่างที่ติดป้ายกำกับใหม่เพื่อสร้างตัวจำแนกในขั้นตอนสุดท้ายสำหรับทำนายตัวอย่างทดสอบ แต่ทั้งสองวิธีติดป้ายกำกับตัวอย่างทดสอบตามคลาสของกลุ่มข้อมูลที่ตัวอย่างทดสอบใกล้กับศูนย์กลางของกลุ่มมากที่สุด

งานวิจัยที่นำการจัดกลุ่มมาช่วยปรับปรุงการเรียนรู้กึ่งมีผู้สอนวิธีอื่น ๆ ได้แก่ วิธีซีบีซี (Cluster based classification: CBC) (Zeng et al., 2003) ซึ่งปรับปรุงการจัดกลุ่มโดยใช้การจัดกลุ่มเค-มีนส์ด้วยเงื่อนไขแบบอ่อน (soft-constraint k-means) เริ่มต้นกำหนดให้ศูนย์กลางของตัวอย่างที่มีป้ายกำกับแต่ละคลาสคือศูนย์กลางตั้งต้นของแต่ละกลุ่ม จากนั้นจัดกลุ่มตัวอย่างทั้งหมดตามศูนย์กลางตั้งต้นนั้นและปรับค่าศูนย์กลางของแต่ละกลุ่ม จนกว่าตัวอย่างไม่เปลี่ยนไป

อยู่กลุ่มอื่น หรือหยุดการรวมกลุ่มก่อนที่การรวมกลุ่มนั้นจะส่งผลให้ศูนย์กลางตั้งต้นต่างคลาสถูกรวมเข้าด้วยกัน หลังจากนั้นจึงติดป้ายกำกับให้แก่สมาชิกทั้งกลุ่มตามค่าคลาสของศูนย์กลางตั้งต้นของกลุ่มนั้น สุดท้ายจึงนำตัวอย่างมีป้ายกำกับใหม่และตัวอย่างมีป้ายกำกับตั้งต้นมาสร้างตัวจำแนกด้วยซัพพอร์ตเวกเตอร์แมชชีน

อีกวิธีหนึ่งคือ วิธีเอสเอสซีซีเอ็ม (Semi-supervised classification based on class membership: SSCCM) (Wang et al., 2012) ซึ่งพยายามหาฟังก์ชันตัดสินใจ (decision function) จากค่าความเป็นสมาชิกของป้ายกำกับ (label membership) ของตัวอย่าง โดยตัวอย่างแต่ละตัวมีค่าความเป็นสมาชิกได้มากกว่าหนึ่งคลาส วิธีนี้ใกล้เคียงกับแนวคิดของวิธีการจัดกลุ่มกึ่งมีผู้สอนด้วยวิธี Fuzzy c-means (semi-supervised fuzzy c-means) (Gan et al., 2013) โดยวิธีนี้ใช้การจัดกลุ่มข้อมูลช่วยปรับปรุงการเรียนรู้กึ่งมีผู้สอนวิธีเรียนรู้ด้วยตนเอง โดยจัดกลุ่มตัวอย่างทั้งหมดตามจำนวนคลาส ตัวอย่างแต่ละตัวสามารถเป็นสมาชิกมากกว่าหนึ่งกลุ่มได้โดยแต่ละตัวอย่างจะมีค่าระดับความเป็นสมาชิก (membership degree) ของแต่ละกลุ่ม ตัวอย่างที่มีค่าความเป็นสมาชิกสูงจะถูกเลือกไปใช้ติดป้ายกำกับด้วยซัพพอร์ตเวกเตอร์แมชชีนที่สร้างจากตัวอย่างมีป้ายกำกับ และตัวอย่างที่ซัพพอร์ตเวกเตอร์แมชชีนมั่นใจที่จะติดป้ายกำกับมากที่สุดจะถูกติดป้ายกำกับ และนำไปรวมกับตัวอย่างมีป้ายกำกับตั้งต้นเพื่อปรับปรุงในขั้นตอนการจัดกลุ่มและปรับปรุงซัพพอร์ตเวกเตอร์แมชชีนที่จะใช้ในรอบต่อไป

ทั้งวิธีซีบีซี วิธีเอสเอสซีซีเอ็มและการจัดกลุ่มกึ่งมีผู้สอนด้วยวิธี Fuzzy c-means กำหนดให้จำนวนกลุ่มข้อมูลเท่ากับจำนวนคลาส เพื่อใช้คุณลักษณะของกลุ่มข้อมูลเพื่อเลือกตัวอย่างที่มั่นใจที่สุดในแต่ละคลาส ซึ่งแตกต่างจากการจัดกลุ่มและติดป้ายที่ใช้ในงานวิจัยนี้ที่ไม่ได้กำหนดให้หนึ่งคลาสต้องมีเพียงหนึ่งกลุ่ม แต่ตัวอย่างในคลาสใด ๆ อาจประกอบไปด้วยกลุ่มข้อมูลมากกว่าหนึ่งกลุ่มได้ ซึ่งสอดคล้องกันลักษณะข้อมูลที่แท้จริงในปัญหาส่วนใหญ่ (Piroonsup and Sinthupinyo, 2010a) (Wang et al., 2012)

3.2 การใช้การจัดกลุ่มข้อมูลในการเรียนรู้แบบแอ็กทิฟ

งานวิจัยที่ใช้การจัดกลุ่มข้อมูลในการเรียนรู้แบบแอ็กทิฟใช้การจัดกลุ่มข้อมูลเพื่อเลือกตัวอย่างมาติดป้ายกำกับเพื่อปรับปรุงคุณภาพตัวเรียนรู้ เช่น การเลือกตัวอย่างที่เป็นตัวแทนของข้อมูล (representative sampling) ในการเรียนรู้แบบแอ็กทิฟด้วยซัพพอร์ตเวกเตอร์แมชชีน (Xu et al., 2003) และการเรียนรู้แบบแอ็กทิฟด้วยการจัดกลุ่มข้อมูลก่อน (pre-clustering) (Nguyen and Smeulders, 2004) ทั้งสองวิธีใช้การจัดกลุ่มแบบไม่มีผู้สอนโดยกำหนดจำนวนกลุ่มข้อมูลตามที่ผู้ใช้กำหนดหรือตามเงื่อนไขที่กำหนด แล้วเสนอให้เลือกตัวอย่างโดยพิจารณาทั้งความไม่มั่นใจของตัวจำแนกปัจจุบันและตำแหน่งของตัวอย่างกับระยะจากศูนย์กลางของกลุ่มข้อมูล โดยเสนอให้เลือกตัวอย่างที่ตัวจำแนกไม่มั่นใจและอยู่ที่บริเวณศูนย์กลางของกลุ่มข้อมูล

มาถามป้ายกำกับจากผู้ใช้

งานวิจัยที่ใช้การจัดกลุ่มข้อมูลกับการเรียนรู้แบบแอ็กทิฟนี้ สนับสนุนแนวคิดในการเลือกตัวอย่างที่ศูนย์กลางของกลุ่มข้อมูล เพื่อเป็นตัวแทนของตัวอย่าง และหลีกเลี่ยงการติดป้ายกำกับตัวอย่างในกลุ่มเดียวกัน ซึ่งสอดคล้องกับแนวคิดการพิจารณาตัวอย่างในกลุ่มไม่ทราบคลาสในงานวิจัยนี้ อย่างไรก็ดีงานวิจัยที่ใช้การจัดกลุ่มข้อมูลในการเรียนรู้แบบแอ็กทิฟนั้นใช้การจัดกลุ่มแบบไม่มีผู้สอน แต่งานวิจัยนี้ใช้การจัดกลุ่มแบบกึ่งสอน โดยใช้ตัวอย่างมีป้ายกำกับในการกำหนดศูนย์กลางตั้งต้นและจำนวนกลุ่มข้อมูล เพื่อวิเคราะห์การกระจายตัวของตัวอย่างและการกระจายตัวของคลาส นอกจากนั้นการเรียนรู้แบบแอ็กทิฟอาศัยผู้ใช้ช่วยติดป้ายกำกับให้แก่ตัวอย่างที่ถูกเลือก ซึ่งทำให้ระบบไม่เป็นอัตโนมัติ งานวิจัยนี้จึงทดลองหาวิธีติดป้ายกำกับให้ตัวอย่างที่เลือกมาจากกลุ่มที่ไม่มีตัวอย่างมีป้ายกำกับด้วยตัวจำแนกวิธีต่าง ๆ ผลการทดลองพบว่าตัวจำแนกป่าไม้แบบสุ่มช่วยติดป้ายกำกับก่อนนำตัวอย่างไปใช้ในการเรียนรู้แบบกึ่งสอนได้ดีที่สุด รายละเอียดของวิธีการวิเคราะห์ตัวอย่างมีป้ายกำกับเพื่อปรับปรุงการจำแนกกึ่งมีผู้สอนวิธีเรียนรู้ด้วยตนเองจะนำเสนอในบทที่ 5

บทที่ 4

วิธีติดป้ายกำกับโดยแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือกเพื่อปรับปรุงวิธีจัดกลุ่มและติดป้าย

วิธีติดป้ายกำกับโดยแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือกนี้นำเสนอเพื่อแก้ปัญหาในการติดป้ายกำกับในกลุ่มที่มีตัวอย่างต่างคลาสปะปนกัน เนื่องจากในขั้นตอนการติดป้ายกำกับของการเรียนรู้กึ่งมีผู้สอนวิธีจัดกลุ่มและติดป้ายโดยทั่วไป หากกลุ่มตัวอย่างนั้นประกอบไปด้วยตัวอย่างมากกว่าหนึ่งคลาสจะเลือกแก้ปัญหาสองแนวทาง แนวทางแรกคือไม่นำตัวอย่างไม่มีป้ายกำกับที่อยู่ในกลุ่มคลาสปะปนนั้นมาใช้งาน Dara et al. (2002) ซึ่งอาจลดตัวอย่างที่มีประโยชน์ต่อการปรับปรุงตัวจำแนกลงไปด้วย อีกแนวทางหนึ่งคือติดป้ายกำกับตัวอย่างทั้งกลุ่มด้วยวิธี-โหวตคลาสส่วนใหญ่ (majority vote) ซึ่งติดป้ายกำกับตัวอย่างตามคลาสส่วนใหญ่ของตัวอย่างมีป้ายกำกับในกลุ่มนั้น Demiriz et al. (1999) วิธีนี้อาจส่งผลให้ติดป้ายกำกับผิดแก่บางตัวอย่างได้และตัวอย่างที่ติดป้ายกำกับผิดนั้นอาจส่งผลเสียต่อการสร้างตัวจำแนกสุดท้ายด้วย

ขั้นตอนวิธีติดป้ายกำกับโดยแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือกแบ่งกลุ่มที่มีตัวอย่างต่างคลาสหรือกลุ่มคลาสปะปนเป็นกลุ่มย่อย ๆ แต่เนื่องจากคุณลักษณะชุดที่ใช้ในการจัดกลุ่มข้อมูลไม่สามารถจำแนกตัวอย่างในกลุ่มนี้ได้ จึงเสนอใช้เลือกคุณลักษณะใหม่ ที่สามารถจำแนกตัวอย่างต่างคลาสในกลุ่มคลาสปะปนได้ แล้วจึงจัดกลุ่มตัวอย่างที่อยู่ในคลาสปะปนนั้นใหม่ด้วยค่าคุณลักษณะที่ถูกเลือก โดยจะได้กลุ่มข้อมูลที่ละเอียดขึ้นและลดความน่าจะเป็นที่จะติดป้ายกำกับตัวอย่างผิดได้

การทดลองนี้ใช้ชุดข้อมูลปัญหาจริงจากปัญหาการลดสิ่งรบกวนในเอกสารภาษาไทยโดยรายละเอียดของชุดข้อมูลอธิบายอยู่ในหัวข้อที่ 4.1 ขั้นตอนวิธีโหวตคลาสส่วนใหญ่ซึ่งเป็นวิธีที่ใช้โดยทั่วไปสำหรับการเรียนรู้กึ่งมีผู้สอนวิธีจัดกลุ่มและติดป้ายอธิบายอยู่ในหัวข้อที่ 4.2 ขั้นตอนวิธีของวิธีติดป้ายกำกับโดยแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือกที่นำเสนออธิบายอยู่ในหัวข้อที่ 4.3 และเสนอผลการทดลองนำเสนอในหัวข้อที่ 4.4 ซึ่งเปรียบเทียบทั้งความถูกต้องในการติดป้ายกำกับและความถูกต้องในการจำแนกของตัวจำแนก ที่สร้างจากตัวอย่างที่ติดป้ายกำกับด้วยวิธีเดิมเปรียบเทียบกับผลลัพธ์จากวิธีที่นำเสนอ ในหัวข้อนี้ยังได้เปรียบเทียบประสิทธิภาพในการลดสิ่งรบกวนระหว่างวิธีที่นำเสนอกับอีกสองวิธีการ ได้แก่ การลดสิ่งรบกวนวิธีเอสพีเอ็นสองขั้นตอน (Two-phased SPN) ซึ่งเป็นการลดสิ่งรบกวนที่คล้ายคลึงกับวิธีที่นำเสนอ และซอฟต์แวร์ที่ใช้ในการลดสิ่งรบกวนในเอกสาร ได้แก่ ScanFix Xpress 6.0 ด้วย

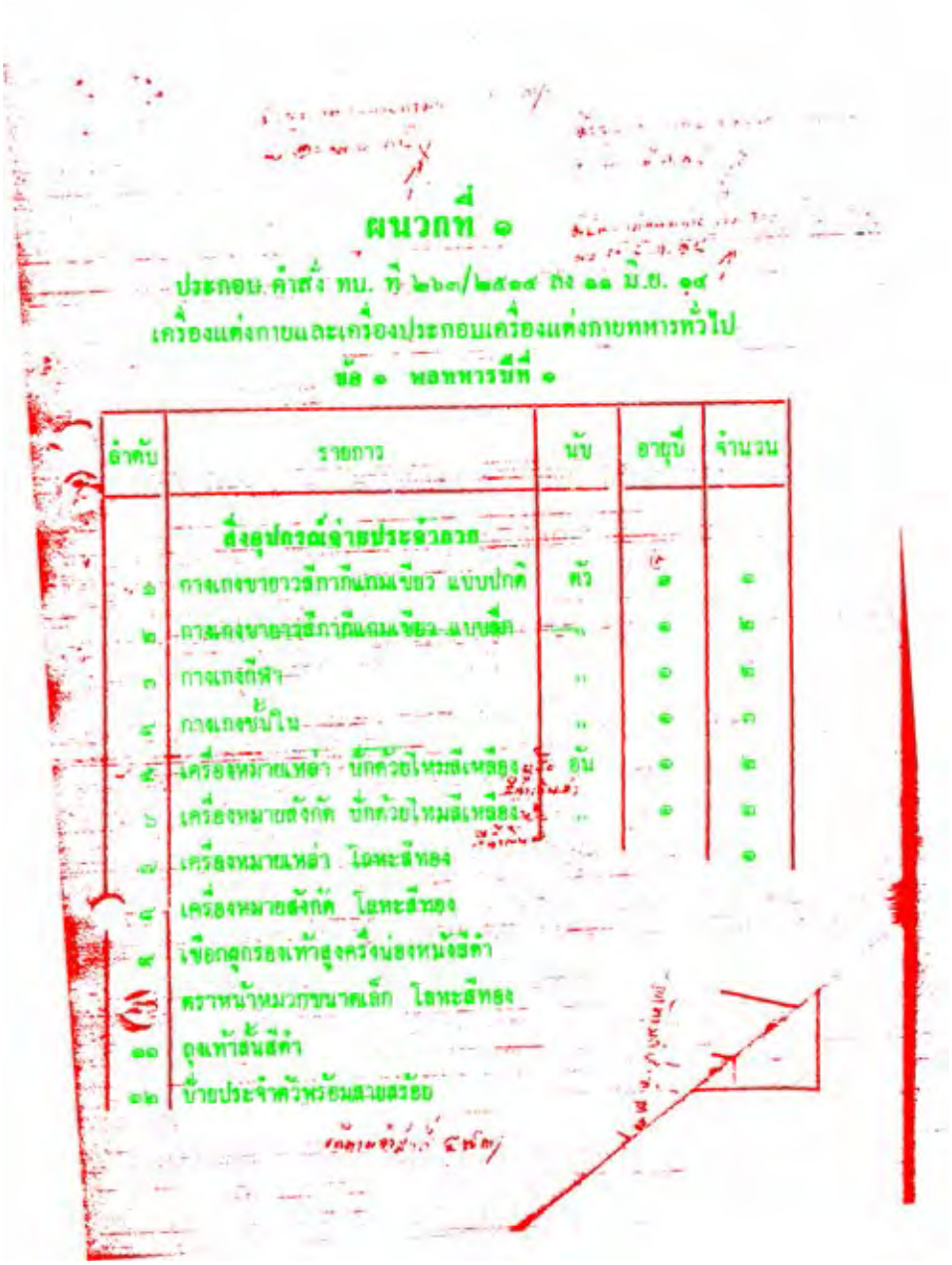
4.1 ชุดข้อมูล

ชุดข้อมูลที่ใช้ในการทดลองเป็นชุดข้อมูลจริง สร้างจากภาพเอกสารภาษาไทยซึ่งเป็นเอกสารเก่าซึ่งถูกสแกนเก็บไว้ประกอบไปด้วย 182 ภาพเอกสาร งานวิจัยนี้กำหนดให้องค์ประกอบที่อยู่ติดกัน (connected component) ในละองค์ประกอบคือหนึ่งตัวอย่าง และแบ่งตัวอย่างเป็นตัวอย่างประเภทตัวอักษรภาษาไทยและประเภทสิ่งรบกวน แต่ละตัวอย่างประกอบไปด้วยค่าคุณลักษณะที่พิจารณาจากโครงสร้างขององค์ประกอบ 9 คุณลักษณะ ได้แก่ ความกว้าง ความสูง อัตราส่วนระหว่างความกว้างต่อความสูง ความหนาแน่นของจุดดำ ความหนาของเส้น จำนวนขาบน จำนวนขาล่าง จำนวนจุดตัดและจำนวนห่วง รายละเอียดของตัวอักษรภาษาไทยรวมถึงกระบวนการสกัดตัวอย่างและค่าคุณลักษณะสามารถศึกษาเพิ่มเติมได้จาก (Piroonsup, 2009)

ชุดข้อมูลการจำแนกสิ่งรบกวนกับตัวอักษรในเอกสารภาษาไทยนี้ประกอบไปด้วยภาพเอกสารทั้งหมดประกอบไปด้วย 999,520 องค์ประกอบ แบ่งเป็นประเภทตัวอักษรจำนวน 137,444 องค์ประกอบ และประเภทสิ่งรบกวนจำนวน 862,076 องค์ประกอบ โดยการทดลองแบ่งตัวอย่าง 7 ภาพเอกสารเป็นชุดเรียนรู้และ 175 ภาพเอกสารเป็นชุดทดสอบ ชุดเรียนรู้ประกอบไปด้วยตัวอย่างจำนวน 40,842 องค์ประกอบ กำหนดให้ผู้ใช้ติดป้ายกำกับประมาณร้อยละ 4 ขององค์ประกอบที่ใช้ในการเรียนรู้ ตัวอย่างที่เหลือร้อยละ 96 เป็นตัวอย่างไม่มีป้ายกำกับ ชุดทดสอบประกอบไปด้วย 966,911 องค์ประกอบซึ่งถูกติดป้ายกำกับทั้งหมด ตัวอย่างภาพเอกสารในชุดทดสอบที่ถูกติดป้ายกำกับองค์ประกอบทั้งหมดแสดงอยู่ในภาพที่ 4.1 โดยองค์ประกอบสีเขียวคือประเภทตัวอักษรและสีแดงคือประเภทสิ่งรบกวน

4.2 ขั้นตอนวิธีโหวตคลาสส่วนใหญ่

การติดป้ายกำกับด้วยวิธีโหวตคลาสส่วนใหญ่เป็นวิธีที่ใช้โดยทั่วไปในการติดป้ายกำกับสำหรับตัวอย่างในกลุ่มคลาสปะปน Demiriz et al. (1999) คลาสส่วนใหญ่ คือ คลาสของตัวอย่างมีป้ายกำกับส่วนใหญ่ในกลุ่มนั้น ซึ่งมีอัตราส่วนของตัวอย่างของคลาสนั้นเกินกว่าค่าขีดแบ่ง (threshold) ที่กำหนด ขั้นตอนวิธีพิจารณาคลาสส่วนใหญ่เริ่มจากการนับจำนวนตัวอย่างในแต่ละคลาสในกลุ่มพิจารณาตามสมการที่ 4.1 โดยที่ c_j แทนป้ายกำกับคลาส N_{c_j} คือจำนวนตัวอย่างซึ่งมีป้ายกำกับเป็นคลาส c_j n คือจำนวนตัวอย่างในกลุ่มนั้น และ y_i แทนป้ายกำกับคลาสของตัวอย่าง x_i



ผนวก ๑

ประกอบคำสั่ง ทบ. ที่ ๒๖๓/๒๕๑๙ ลง ๑๑ มิ.ย. ๑๙

เครื่องแต่งกายและเครื่องประกอบเครื่องแต่งกายทหารทั่วไป

หมวด ๑ พลทหารยศที่ ๑

ลำดับ	รายการ	นับ	อายุปี	จำนวน
	สิ่งอุปกรณ์จ่ายประจำตัวบุคคล			
๑	กางเกงขายาวสีกากีแถมเขียว แบบปกติ	ตัว	๑	๑
๒	กางเกงขายาวสีกากีแถมเขียว แบบพิเศษ	,,	๑	๒
๓	กางเกงยีนส์	,,	๑	๒
๔	กางเกงชั้นใน	,,	๑	๓
๕	เครื่องหมายเหล่า ปักด้วยไหมสีเหลือง	อัน	๑	๒
๖	เครื่องหมายสังกัด ปักด้วยไหมสีเหลือง	,,	๑	๒
๗	เครื่องหมายเหล่า โลหะสีทอง			๑
๘	เครื่องหมายสังกัด โลหะสีทอง			
๙	เชือกสูตรองเท้าสูงครึ่งน่องหนังสีดำ			
๑๐	ตราหน้าหมวกขนาดเล็ก โลหะสีทอง			
๑๑	ถุงเท้าสั้นสีดำ			
๑๒	ป้ายประจำตัวพร้อมสายสวย			

ภาพที่ 4.1: ตัวอย่างภาพเอกสารในชุดทดสอบที่นำมาใช้วัดผล องค์ประกอบสีเขียวคือประเภทตัวอักษร และสีแดงคือประเภทสิ่งรบกวน

$$N_{c_j} = \sum_{i=1}^{n} f_{c_j}(x_i) \tag{4.1}$$

$$f_{c_j}(x_i) = \begin{cases} 1 & \text{if } y_i = c_j \\ 0 & \text{if } y_i \neq c_j \end{cases} \tag{4.2}$$

หลังจากนั้นจึงหาอัตราส่วนคลาสปะปน (mixed-class ratio) ในแต่ละกลุ่มตามสมการที่ 4.3 โดยที่ C คือจำนวนคลาส N_l คือจำนวนตัวอย่างมีป้ายกำกับในกลุ่มนั้น

$$\text{mixed-class ratio} = \frac{\max_{j=1}^{C}(N_{c_j})}{N_l} \tag{4.3}$$

อัตราส่วนคลาสปะปน คือ อัตราส่วนระหว่างจำนวนตัวอย่างในคลาสที่ปรากฏมากที่สุดในกลุ่มนั้นกับจำนวนตัวอย่างมีป้ายกำกับทั้งหมดในกลุ่มนั้น ตัวอย่างไม่มีป้ายกำกับจะถูกติดป้ายกำกับตามคลาสส่วนใหญ่ในกลุ่มนั้นถ้าอัตราส่วนคลาสปะปนมีค่ามากกว่าค่าขีดแบ่งที่กำหนด แต่หากอัตราส่วนคลาสปะปนมีค่าน้อยกว่าค่าขีดแบ่งตัวอย่างไม่มีป้ายกำกับในกลุ่มนั้นจะไม่ถูกติดป้ายกำกับ โดยค่าขีดแบ่งที่ใช้ในการกำหนดการติดป้ายกำกับคลาสนี้มีค่าอยู่ระหว่าง 0.5 ถึง 1 หากค่าขีดแบ่งเท่ากับ 1 ตัวอย่างในกลุ่มคลาสปะปนทั้งหมดจะไม่ถูกติดป้ายกำกับ หากค่าขีดแบ่งเท่ากับ 0.5 ตัวอย่างในกลุ่มคลาสปะปนทุกกลุ่มที่มีจำนวนตัวอย่างในคลาสใดคลาสหนึ่งมากกว่าคลาสอื่น ๆ จะถูกติดป้ายกำกับตามคลาสส่วนใหญ่ในกลุ่มนั้น

แม้ว่าการติดป้ายกำกับวิธีโหวตคลาสส่วนใหญ่จะสามารถติดป้ายกำกับได้ค่อนข้างมีประสิทธิภาพ แต่วิธีนี้อาจติดป้ายกำกับไม่ถูกต้องโดยเฉพาะตัวอย่างไม่มีป้ายกำกับที่เป็นสมาชิกในกลุ่มคลาสปะปน เนื่องจากกลุ่มคลาสปะปนนี้มีความน่าจะเป็นที่จะประกอบไปด้วยตัวอย่างมากกว่าหนึ่งคลาส แต่การติดป้ายกำกับวิธีโหวตคลาสส่วนใหญ่นั้นเลือกเพียงคลาสเดียวในการติดป้ายกำกับให้แก่ตัวอย่างทั้งหมดในกลุ่มนั้น โดยไม่พิจารณาตัวอย่างในคลาสส่วนน้อย โดยตัวอย่างที่ถูกติดป้ายกำกับผิดนั้นจะถูกนำไปใช้สร้างตัวจำแนกในขั้นตอนสุดท้าย ซึ่งย่อมส่งผลเสียต่อประสิทธิภาพของจำแนกอย่างหลีกเลี่ยงไม่ได้

4.3 ขั้นตอนวิธีติดป้ายกำกับโดยแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือก

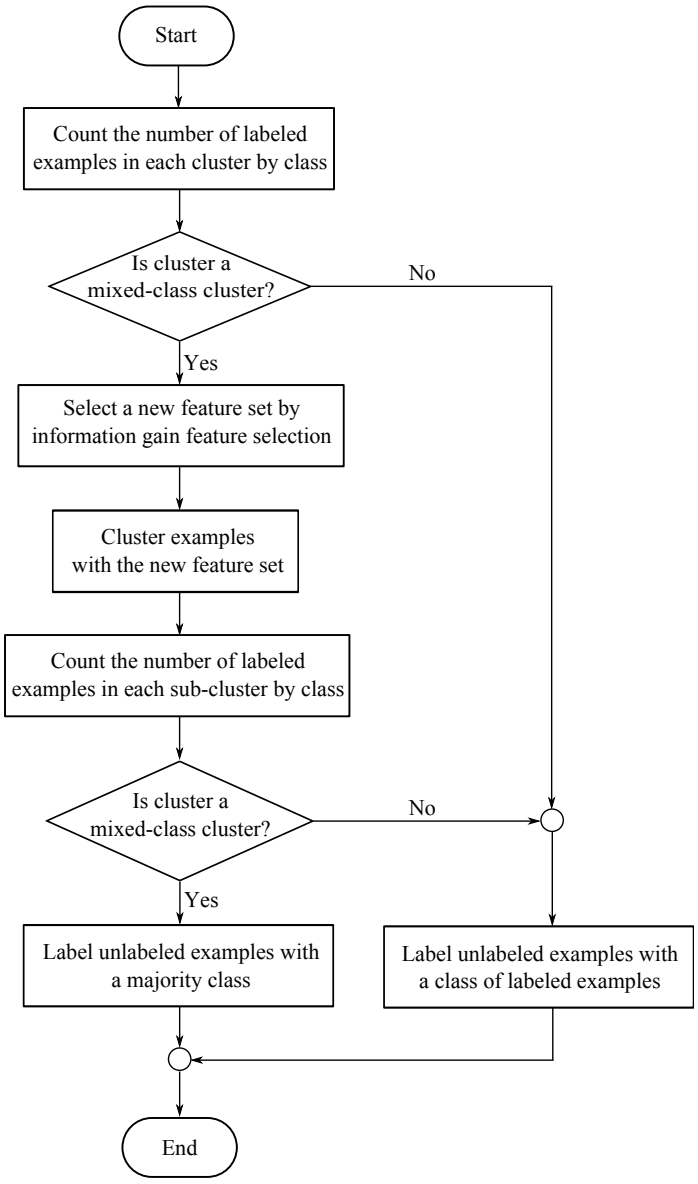
งานวิจัยนี้เสนอวิธีติดป้ายกำกับโดยแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือก เพื่อปรับปรุงการติดป้ายกำกับในกลุ่มคลาสปะปน วิธีนี้แบ่งตัวอย่างในกลุ่มคลาสปะปนออกเป็นกลุ่มย่อย แต่เนื่องจากคุณลักษณะชุดที่ใช้ในการจัดกลุ่มข้อมูลไม่สามารถจำแนกตัวอย่างในกลุ่มนี้ได้ จึงเสนอ

ให้เลือกคุณลักษณะที่สามารถจำแนกตัวอย่างต่างคลาสในกลุ่มคลาสปะปนได้ โดยในงานวิจัยนี้เลือกคุณลักษณะชุดใหม่ด้วยค่าอินฟอร์เมชันเกน (information gain) Quinlan (1986) ผลลัพธ์ที่ได้ คือกลุ่มข้อมูลที่ละเอียดขึ้นซึ่งลดความน่าจะเป็นที่จะติดป้ายกำกับตัวอย่างผิดลงได้ แผนภาพแสดงภาพรวมการทำงานของกระบวนการแสดงดังรูปที่ 4.2 และขั้นตอนวิธีของการติดป้ายกำกับด้วยการติดป้ายกำกับกลุ่มย่อยตามคุณลักษณะที่ถูกเลือกแสดงอยู่ในขั้นตอนวิธีที่ 4 โดย y_l คือ ตัวอย่างมีป้ายกำกับ และ y_u คือตัวอย่างไม่มีป้ายกำกับ y_{major} คือคลาสของตัวอย่างมีป้ายกำกับส่วนใหญ่ N_{major} คือจำนวนตัวอย่างในคลาสส่วนใหญ่ N_l คือจำนวนตัวอย่างมีป้ายกำกับ

Algorithm 4 ขั้นตอนวิธีติดป้ายกำกับโดยแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือก

```
1: procedure ClusterLabeling(C)
2:     for each cluster c_i in C do
3:         Find majority class of labeled examples y_major
4:         Count the number of labeled examples of majority class N_major
5:         if N_major = N_l then
6:             y_u = y_major
7:         else
8:             y_u = SplitCluster(x_i,y_l)
9:         end if
10:    end for
11: end procedure
12: procedure SplitCluster(x_i,y_l)
13:    f_new = FeatureSelection(x_l,y_l)
14:    C_split = SubClustering(x_i,f_new)
15:    for each cluster c_j in C_split do
16:        Find majority class of labeled examples y_major
17:        y_split = y_major
18:    end for
19: return y_split
20: end procedure
```

วิธีนี้พิจารณาคลาสของตัวอย่างมีป้ายกำกับโดยตัวอย่างที่เป็นสมาชิกกลุ่มที่มีตัวอย่างมีป้ายกำกับเพียงคลาสเดียวจะถูกติดป้ายกำกับด้วยคลาสของตัวอย่างนั้น ส่วนตัวอย่างที่เป็นสมาชิกกลุ่มคลาสปะปนจะถูกติดป้ายกำกับด้วยฟังก์ชัน SplitCluster ซึ่งเริ่มจากการหาชุดคุณลักษณะใหม่ด้วยการเลือกคุณลักษณะด้วยค่าอินฟอร์เมชันเกน โดยเลือกคุณลักษณะที่สามารถจำแนกตัวอย่างต่างคลาสออกจากกันได้ดีที่สุด คุณลักษณะชุดใหม่ที่เลือกมานั้นจะถูกนำไปใช้จัดกลุ่มตัวอย่างในกลุ่มคลาสปะปนด้วยขั้นตอนวิธีการจัดกลุ่มแบบเกาะกลุ่ม (agglomerative cluster-



ภาพที่ 4.2: แผนภาพแสดงกระบวนการติดป้ายกำกับโดยแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือก

ing) Vesanto and Alhoniemi (2000) สุดท้ายจึงติดป้ายกำกับให้แก่ตัวอย่างไม่มีป้ายกำกับในกลุ่มย่อยนั้น

วิธีที่นำเสนอนี้ช่วยประมาณขอบการตัดสินใจในการจำแนกคลาสของตัวอย่างในกลุ่มตัวอย่างที่มีการปะปนกัน โดยการเลือกค่าคุณลักษณะที่สามารถจำแนกตัวอย่างออกจากกันได้ดีที่สุด ขอบการตัดสินใจที่ได้นี้ช่วยให้การติดป้ายกำกันในกลุ่มคลาสปะปนดีขึ้นกว่าวิธีโหวตคลาสส่วนใหญ่ซึ่งไม่ได้พิจารณาตัวอย่างในคลาสส่วนน้อย

4.4 ผลการทดลอง

หัวข้อผลการทดลองนี้ประกอบไปด้วย 4 หัวข้อย่อย ได้แก่ วิธีการวัดผล ผลลัพธ์การจัดกลุ่ม ผลลัพธ์เปรียบเทียบประสิทธิภาพการติดป้ายกำกับ ผลลัพธ์เปรียบเทียบประสิทธิภาพการลดสิ่งรบกวน ซึ่งจะนำเสนอในหัวข้อย่อยต่อไปนี้ตามลำดับ

4.4.1 วิธีการวัดผล

งานวิจัยนี้วัดประสิทธิภาพการลดสิ่งรบกวนด้วยค่าความถูกต้องของการทำนายตัวอย่างทั้งสองคลาสตามสมการที่ 5.3 และค่าเอฟของแต่ละคลาสซึ่งแสดงถึงความสามารถในการจำแนกตัวอย่างทั้งสองคลาสดังสมการที่ 4.5 - 4.9 ด้านล่าง ค่าเอฟของตัวอักษรที่สูงแสดงถึงความแม่นยำและความครอบคลุมในการคงตัวอักษรไว้ในภาพเอกสาร ค่าเอฟของสิ่งรบกวนที่สูงแสดงถึงความแม่นยำและความครอบคลุมในการลดสิ่งรบกวนจากภาพเอกสารนั้น

$$\text{Accuracy} = \frac{\text{char predict as char} + \text{noise predict as noise}}{\text{number of all examples}} \tag{4.4}$$

$$\text{Precision of char} = \frac{\text{char predict as char}}{\text{char predict as char} + \text{noise predict as char}} \tag{4.5}$$

$$\text{Recall of char} = \frac{\text{char predict as char}}{\text{char predict as char} + \text{char predict as noise}} \tag{4.6}$$

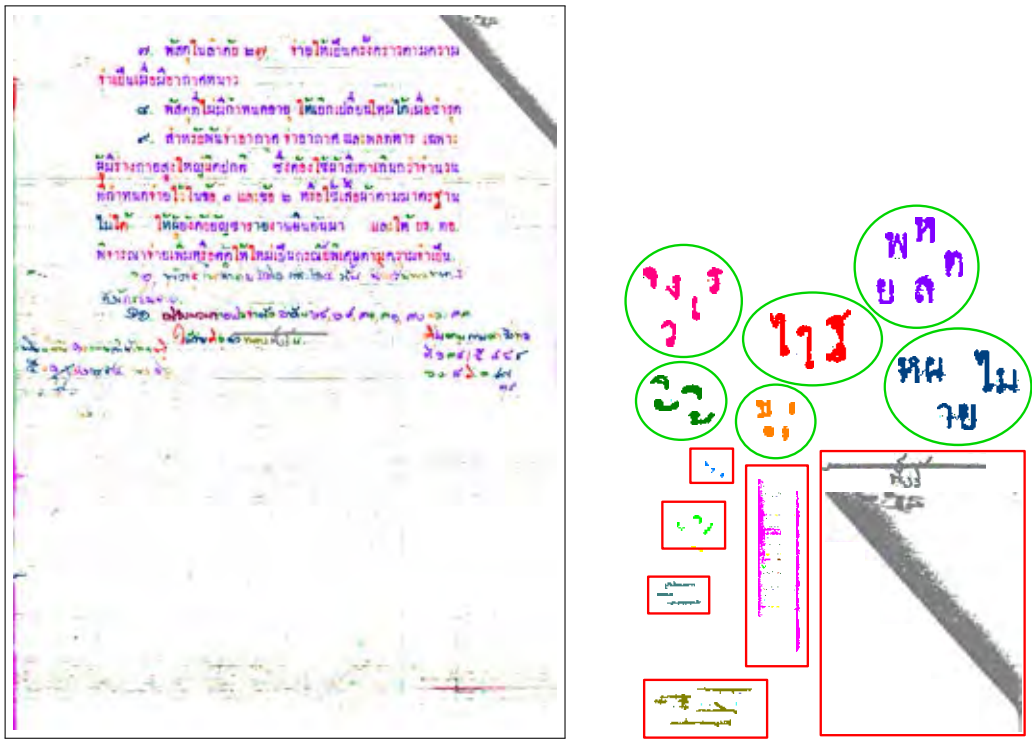
$$\text{Precision of noise} = \frac{\text{noise predict as noise}}{\text{noise predict as noise} + \text{char predict as noise}} \tag{4.7}$$

$$\text{Recall of noise} = \frac{\text{noise predict as noise}}{\text{noise predict as noise} + \text{noise predict as char}} \tag{4.8}$$

$$\text{F-measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.9}$$

4.4.2 ผลลัพธ์การจัดกลุ่ม

ภาพที่ 4.3 แสดงตัวอย่างผลลัพธ์ที่ได้จากการจัดกลุ่มตัวอย่าง ในภาพดังกล่าวประกอบไปด้วยตัวอย่างที่ถูกกำหนดสีต่าง ๆ โดยตัวอย่างที่มีสีเดียวกันคือตัวอย่างในกลุ่มเดียวกัน ซึ่งจะเห็นได้ว่าการจัดกลุ่มข้อมูลสามารถจัดกลุ่มตัวอย่างประเภทตัวอักษรที่มีความคล้ายคลึงกัน



(a) (b)

ภาพที่ 4.3: ผลการจัดกลุ่มตัวอย่าง (a) ภาพเอกสารที่ให้สีตัวอย่างตามกลุ่ม (b) องค์ประกอบที่ถูกจัดกลุ่มตามประเภท โดยองค์ประกอบประเภทสิ่งรบกวนอยู่ในกลุ่มสี่เหลี่ยมสีแดง และองค์ประกอบประเภทตัวอักษรอยู่ในกลุ่มวงกลมสีเขียว

เป็นกลุ่มเดียวกัน เช่น ตัวอักษรขนาดเล็ก ตัวอักษรลักษณะผอม ตัวอักษรที่อยู่ติดกับ เป็นต้น และสามารถจัดกลุ่มสิ่งรบกวนที่มีลักษณะคล้ายคลึงกันเป็นกลุ่มเดียวกัน เช่น สิ่งรบกวนที่เป็นเส้นแนวนอน สิ่งรบกวนเป็นเส้นแนวตั้ง สิ่งรบกวนที่เป็นกลุ่มขนาดต่าง ๆ กัน เป็นต้น

4.4.3 ผลลัพธ์เปรียบเทียบประสิทธิภาพในการติดป้ายกำกับ

หัวข้อนี้แสดงผลการทดลองเปรียบเทียบประสิทธิภาพในการติดป้ายกำกับตัวอย่างระหว่างวิธีโหวตคลาสส่วนใหญ่กับวิธีติดป้ายกำกับโดยแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือก เนื่องจากการวิธีโหวตคลาสส่วนใหญ่จำเป็นต้องระบุค่าขีดแบ่งของอัตราส่วนคลาสปะปน การทดลองนี้จึงทดลองบนค่าขีดแบ่งตั้งแต่ค่าที่น้อยที่สุดที่เป็นไปได้คือ 0.5 ถึงค่าขีดแบ่งที่มากที่สุดคือ 1 โดยเพิ่มขึ้นทีละ 0.1 และการทดลองนี้ใช้ไลบรารี (library) ของ Weka (Hall et al., 2009) เพื่อเลือกค่าคุณลักษณะด้วยอินฟอร์เมชันเกน และสร้างตัวจำแนกสุดท้ายด้วยขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดเคจำนวน (k-nearest neighbors) ซึ่งเป็นตัวจำแนกขึ้นกับตัวอย่าง (instance based classifier) โดยกำหนดค่าเคเท่ากับ 3

วิธีการ	ความถูกต้อง	ค่าเอฟของ คลาสตัวอักษร	ค่าเอฟของ คลาสสิ่งรบกวน
วิธีโหวตคลาสส่วนใหญ่ที่			
ค่าขีดแบ่งเท่ากับ 0.5	93.98 $\pm$ 0.3	81.93 $\pm$ 0.48	96.39 $\pm$ 0.23
ค่าขีดแบ่งเท่ากับ 0.6	94.37 $\pm$ 0.29	82.86 $\pm$ 0.47	96.63 $\pm$ 0.23
ค่าขีดแบ่งเท่ากับ 0.7	93.95 $\pm$ 0.3	82.49 $\pm$ 0.48	96.35 $\pm$ 0.23
ค่าขีดแบ่งเท่ากับ 0.8	93.95 $\pm$ 0.3	82.49 $\pm$ 0.48	96.35 $\pm$ 0.23
ค่าขีดแบ่งเท่ากับ 0.9	94.47 $\pm$ 0.29	83.6 $\pm$ 0.46	96.67 $\pm$ 0.22
ค่าขีดแบ่งเท่ากับ 1	94.69 $\pm$ 0.28	83.94 $\pm$ 0.46	96.82 $\pm$ 0.22
วิธีแบ่งกลุ่มย่อยตาม คุณลักษณะที่ถูกเลือก	**95.3 $\pm$ 0.26**	**84.99 $\pm$ 0.45**	**97.22 $\pm$ 0.2**

ตารางที่ 4.1: เปรียบเทียบประสิทธิภาพของตัวจำแนกบนตัวอย่างทดสอบที่อยู่ในกลุ่มคลาสปะปน ระหว่างตัวจำแนกที่สร้างจากตัวอย่างที่ติดป้ายกำกับด้วยวิธีโหวตคลาสส่วนใหญ่ และตัวจำแนกที่สร้างจากตัวอย่างที่ติดป้ายกำกับด้วยวิธีแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือก

วิธีที่นำเสนอมุ่งเน้นปรับปรุงการติดป้ายกำกับตัวอย่างในกลุ่มที่ยากที่จะติดป้ายกำกับหรือกลุ่มคลาสปะปนซึ่งเป็นกลุ่มของตัวอย่างส่วนใหญ่ โดยจำนวนตัวอย่างในชุดเรียนรู้ที่เป็นสมาชิกกลุ่มคลาสปะปนมีทั้งสิ้น 24,539 ตัวอย่างจากทั้งหมด 40,842 ตัวอย่าง หรือคิดเป็นร้อยละ 60.08 เช่นเดียวกับตัวอย่างในชุดทดสอบซึ่งจำนวนตัวอย่างทดสอบที่เป็นสมาชิกกลุ่มคลาสปะปนมีจำนวนทั้งสิ้น 550,243 จากทั้งหมด 966,911 หรือคิดเป็นร้อยละ 56.91

ผลการทดลองเปรียบเทียบประสิทธิภาพในการจำแนกตัวอย่างทดสอบในกลุ่มคลาสปะปนด้วยวิธีที่นำเสนอเทียบกับวิธีโหวตคลาสส่วนใหญ่ที่ระดับความเชื่อมั่นร้อยละ 95 แสดงดังตารางที่ 4.1 โดยผลลัพธ์แสดงให้เห็นว่าตัวจำแนกที่สร้างจากตัวอย่างที่ติดป้ายกำกับด้วยวิธีที่นำเสนอสามารถจำแนกตัวอย่างทดสอบที่เป็นสมาชิกกลุ่มคลาสปะปนได้ดีกว่าวิธีโหวตคลาสส่วนใหญ่ในทุก ๆ ค่าขีดแบ่ง

และผลการทดลองเปรียบเทียบประสิทธิภาพในการจำแนกระหว่างทั้งสองวิธีบนตัวอย่างชุดทดสอบทั้งหมดที่ระดับความเชื่อมั่นร้อยละ 95 แสดงดังตารางที่ 4.2 ซึ่งให้ผลลัพธ์ไปในแนวทางเดียวกัน คือ วิธีติดป้ายกำกับโดยแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือกทำให้ได้ตัวจำแนกที่มีประสิทธิภาพดีกว่าวิธีโหวตคลาสส่วนใหญ่

ผลการทดลองวัดความถูกต้องในการติดป้ายกำกับที่ระดับความเชื่อมั่นร้อยละ 95 แสดงในตารางที่ 4.3 โดยเปรียบเทียบความถูกต้องของป้ายกำกับของตัวอย่างไม่มีป้ายกำกับ ระหว่าง

วิธีการ	ความถูกต้อง	ค่าเอฟของคลาสตัวอักษร	ค่าเอฟของคลาสสิ่งรบกวน
วิธีโหวตคลาสส่วนใหญ่ที่			
ค่าขีดแบ่งเท่ากับ 0.5	95.44 $\pm$ 0.2	84.09 $\pm$ 0.35	97.34 $\pm$ 0.16
ค่าขีดแบ่งเท่ากับ 0.6	95.71 $\pm$ 0.2	84.89 $\pm$ 0.35	97.5 $\pm$ 0.15
ค่าขีดแบ่งเท่ากับ 0.7	95.47 $\pm$ 0.2	84.55 $\pm$ 0.35	97.35 $\pm$ 0.16
ค่าขีดแบ่งเท่ากับ 0.8	95.47 $\pm$ 0.2	84.55 $\pm$ 0.35	97.35 $\pm$ 0.16
ค่าขีดแบ่งเท่ากับ 0.9	95.79 $\pm$ 0.19	85.41 $\pm$ 0.34	97.54 $\pm$ 0.15
ค่าขีดแบ่งเท่ากับ 1	95.93 $\pm$ 0.19	85.68 $\pm$ 0.34	97.62 $\pm$ 0.15
วิธีแบ่งกลุ่มย่อยตาม			
คุณลักษณะที่ถูกเลือก	**96.33 $\pm$ 0.18**	**86.58 $\pm$ 0.33**	**97.87 $\pm$ 0.14**
วิธีจำแนกแบบมีผู้สอน	95.89 $\pm$ 0.19	85.67 $\pm$ 0.34	97.6 $\pm$ 0.15

ตารางที่ 4.2: เปรียบเทียบประสิทธิภาพของตัวจำแนกบนตัวอย่างทดสอบทั้งหมดระหว่างตัวจำแนกที่สร้างจากตัวอย่างที่ติดป้ายกำกับด้วยวิธีโหวตคลาสส่วนใหญ่ ตัวจำแนกที่สร้างจากตัวอย่างที่ติดป้ายกำกับด้วยวิธีแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือก และตัวจำแนกแบบมีผู้สอนที่สร้างจากตัวอย่างมีป้ายกำกับเท่านั้น

วิธีติดป้ายกำกับ	ความถูกต้อง
วิธีโหวตคลาสส่วนใหญ่	41.31 $\pm$ 0.48
วิธีแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือก	**94.65 $\pm$ 0.22**

ตารางที่ 4.3: ความถูกต้องของการติดป้ายกำกับตัวอย่างไม่มีป้ายกำกับเปรียบเทียบระหว่างวิธีโหวตคลาสส่วนใหญ่กับวิธีติดป้ายกำกับโดยแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือก

ตัวอย่างที่ติดป้ายกำกับด้วยวิธีที่นำเสนอกับวิธีโหวตคลาสส่วนใหญ่ที่สร้างตัวจำแนกที่มีประสิทธิภาพดีที่สุด หรือที่ค่าขีดแบ่งเท่ากับ 1 ผลลัพธ์แสดงให้เห็นว่าความถูกต้องในการติดป้ายกำกับด้วยวิธีที่นำเสนอนั้นดีกว่าวิธีโหวตคลาสส่วนใหญ่ ส่วนหนึ่งเนื่องจากการติดป้ายกำกับตามคลาสส่วนใหญ่ด้วยค่าขีดแบ่งเท่ากับ 1 นั้นไม่ติดป้ายให้แก่ตัวอย่างในกลุ่มคลาสปะปนเลย ซึ่งประสิทธิภาพในการติดป้ายกำกับนี้เป็นส่วนหนึ่งที่ส่งผลให้ตัวจำแนกที่ได้จากวิธีติดป้ายกำกับโดยแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือกนี้มีประสิทธิภาพ

4.4.4 ผลลัพธ์เปรียบเทียบประสิทธิภาพการลดสิ่งรบกวนในเอกสารภาษาไทย

หัวข้อนี้แสดงผลการทดลองเปรียบเทียบประสิทธิภาพในการลดสิ่งรบกวนในเอกสารภาษาไทย ระหว่างวิธีลดสิ่งรบกวนด้วยการเรียนรู้แบบกึ่งมีผู้สอนวิธีที่ติดป้ายกำกับโดยแบ่งกลุ่ม

ย่อยตามคุณลักษณะที่ถูกเลือกกับวิธีที่ใกล้เคียงกัน ได้แก่ วิธีเอสพีเอ็นแบบสองเฟส (two-phased stroke-like pattern noise removal หรือ two-phased SPN) (Agrawal and Doermann, 2011) และโปรแกรมลดสิ่งรบกวนในเอกสาร ScanFix Xpress 6.0 sca (2010)

วิธีเอสพีเอ็นแบบสองเฟสออกแบบมาเพื่อลดสิ่งรบกวนในเอกสารอารบิกซึ่งประกอบไปด้วยองค์ประกอบขนาดเล็กที่คล้ายคลึงกับสิ่งรบกวนเป็นจำนวนมาก ซึ่งมีลักษณะคล้ายภาษาไทยที่ประกอบไปด้วยตัวอักษรขนาดเล็กที่คล้ายคลึงกับสิ่งรบกวนเช่นกัน เช่น วรรณยุกต์ วิธีเอสพีเอ็นแบบสองเฟสนี้แบ่งองค์ประกอบเป็นสองประเภท ได้แก่ องค์ประกอบที่เชื่อมั่นว่าเป็นตัวอักษร (prominent text component: PTC) และองค์ประกอบที่อาจไม่ใช่ตัวอักษร (non-prominent text component: non-PTC) องค์ประกอบประเภท PTC คือองค์ประกอบที่ระบบมั่นใจว่าเป็นตัวอักษรเนื่องจากมีคุณลักษณะแตกต่างจากองค์ประกอบประเภทสิ่งรบกวนอย่างชัดเจน ส่วนองค์ประกอบประเภท non-PTC ประกอบไปด้วยตัวอักษรที่คล้ายกับสิ่งรบกวนและสิ่งรบกวนในเฟสแรกของเอสพีเอ็นแบบสองเฟสนี้ใช้ตัวอย่างมีป้ายกำกับเพื่อสร้างตัวจำแนกสำหรับจำแนกองค์ประกอบประเภท PTC และ non-PTC ออกจากกัน ตัวจำแนกนี้ใช้ค่าคุณลักษณะขององค์-ประกอบ เช่น พื้นที่ เส้นรอบวง ทิศทางของวงรีที่ล้อมพอดีองค์ประกอบ ความเยื้องศูนย์กลาง (eccentricity) ความยาวในแกนเอก (major axis) และโท (minor axis) ของวงรีนั้น เป็นต้น สำหรับการเรียนรู้ของตัวจำแนก จากนั้นองค์ประกอบที่ถูกจำแนกเป็นประเภท PTC จะถูกนำมาใช้เพื่อคำนวณค่าเฉลี่ยความกว้างของเส้นของตัวอักษร และค่าเฉลี่ยระยะห่างระหว่างตัวอักษรเพื่อใช้ในการจำแนกสิ่งรบกวนและตัวอักษรที่คล้ายสิ่งรบกวนในเฟสที่สอง ในเฟสที่สองพิจารณาทั้งสองค่านี้เพื่อจัดกลุ่มองค์ประกอบประเภท non-PTC ด้วยขั้นตอนวิธีเค-มีนส์ ซึ่งองค์ประ-กอบจะถูกจัดกลุ่มเป็นเป็นองค์ประกอบประเภทสิ่งรบกวนหรือตัวอักษร

โปรแกรม ScanFix Xpress 6.0 เป็นโปรแกรมพาณิชย์สำหรับปรับปรุงคุณภาพภาพเอก-สาร โดยการทดลองนี้เลือกใช้วิธีลดสิ่งรบกวนทั้งหมดในโปรแกรม ได้แก่ ปรับองค์ประกอบให้เรียบ (smooth objects) ลดสิ่งรบกวนขนาดเล็ก (despeckle) ลบเส้น (line removal) ลบสิ่งรบกวนที่ขาดช่วง (comb removal) ลดสิ่งรบกวนบริเวณขอบ (border removal) และลดสิ่งรบกวนขนาดใหญ่ (blob removal) และการทดลองนี้ปรับค่าพารามิเตอร์ของโปรแกรมให้สามารถลดสิ่งรบกวนในเอกสารชุดเรียนรู้ได้ดีที่สุด โดยรายละเอียดค่าพารามิเตอร์ของโปรแกรม ScanFix Xpress 6.0 แสดงอยู่ในบทเสริม ก

ผลการทดลองเปรียบเทียบประสิทธิภาพในการลดสิ่งรบกวนระหว่างวิธีที่นำเสนอ เปรียบ-เทียบกับวิธีเอสพีเอ็นแบบสองเฟสและโปรแกรม ScanFix Xpress 6.0 ที่ระดับความเชื่อมั่นร้อยละ 95 แสดงในตารางที่ 4.4 ผลการทดลองแสดงให้เห็นว่าวิธีที่นำเสนอมีค่าความถูกต้องและค่าเอฟทั้งสองคลาสดีกว่าวิธีที่เปรียบเทียบ ซึ่งแสดงให้เห็นว่าวิธีที่นำเสนอสามารถลดสิ่งรบกวน

วิธีลดสิ่งรบกวน	ความถูกต้อง	ค่าเอฟของ คลาสตัวอักษร	ค่าเอฟของ คลาสสิ่งรบกวน
วิธีเอสพีเอ็นแบบสองเฟส	91.7 $\pm$ 0.27	70.80 $\pm$ 0.44	95.16 $\pm$ 0.2
โปรแกรม ScanFix Xpress 6.0	93.16 $\pm$ 0.25	77.94 $\pm$ 0. 4	95.95 $\pm$ 0.19
วิธีแบ่งกลุ่มย่อยตาม คุณลักษณะที่ถูกเลือก	**96.33 $\pm$ 0.18**	**86.58 $\pm$ 0.33**	**97.87 $\pm$ 0.14**

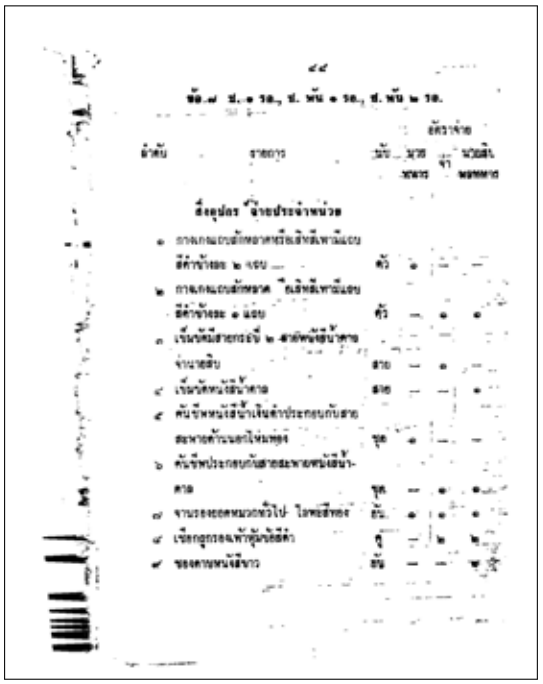
ตารางที่ 4.4: เปรียบเทียบประสิทธิภาพการลดสิ่งรบกวนระหว่าง วิธีเอสพีเอ็นแบบสองเฟส โปรแกรม ScanFix Xpress 6.0 และการเรียนรู้กึ่งมีผู้สอนวิธีจัดกลุ่มและติดป้ายซึ่งติดป้ายกำกับด้วยวิธีแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือก

ออกจากภาพเอกสาร และคงรักษาตัวอักษรไว้ในภาพเอกสารได้มากกว่าวิธีที่เปรียบเทียบทั้งสองวิธี ตัวอย่างภาพเอกสารที่ลดสิ่งรบกวนด้วยวิธีเอสพีเอ็นแบบสองเฟส โปรแกรม ScanFix X-press 6.0 และการเรียนรู้กึ่งมีผู้สอนวิธีจัดกลุ่มและติดป้ายซึ่งติดป้ายกำกับด้วยวิธีแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือกแสดงอยู่ในภาพที่ 4.4 และภาพที่ 4.5 ภาพเอกสารต้นฉบับและภาพเอกสารผลลัพธ์ซึ่งผ่านกระบวนการลดสิ่งรบกวนด้วยวิธีที่นำเสนอและวิธีอื่น ๆ ที่เปรียบเทียบสามารถเข้าถึงได้จากลิงก์ข้อมูลต่อไปนี้
https://www.dropbox.com/s/xjrzh04qdxxb49c/NoiseReductionResult.zip?dl=0

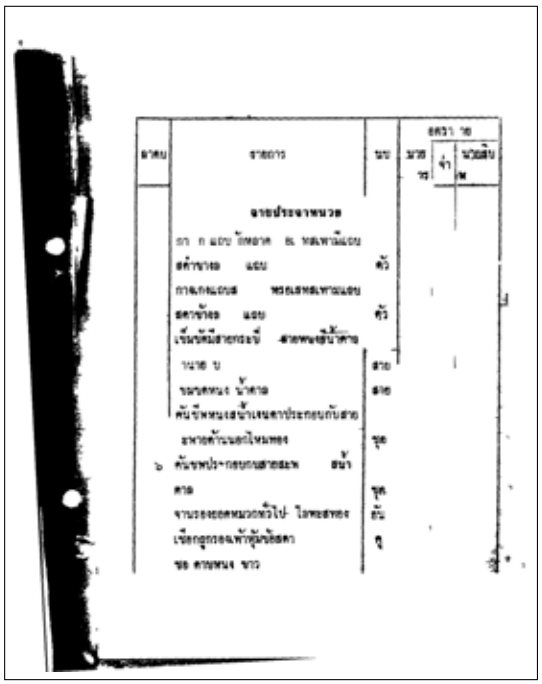
แนวคิดของวิธีเอสพีเอ็นแบบสองเฟสนั้นใกล้เคียงกับวิธีการลดสิ่งรบกวนด้วยการเรียนรู้แบบกึ่งมีผู้สอนที่นำเสนอ โดยทั้งสองวิธีใช้เทคนิคการเรียนรู้ของเครื่องเพื่อลดสิ่งรบกวนบนภาพเอกสารขาวดำเช่นกัน อย่างไรก็ดีวิธีเอสพีเอ็นแบบสองเฟสและวิธีที่นำเสนอมีรายละเอียดที่แตกต่างกันในหลายประเด็น ประเด็นแรกคือความแตกต่างของคุณลักษณะที่ใช้ คุณลักษณะของวิธีที่นำเสนอคือโครงสร้างของตัวอักษรภาษาไทยวัตถุประสงค์เพื่อลดสิ่งรบกวนในเอกสารภาษาไทยเท่านั้น แต่คุณลักษณะของวิธีเอสพีเอ็นแบบสองเฟสคือคุณลักษณะขององค์ประกอบโดยทั่วไปวัตถุประสงค์เพื่อลดสิ่งรบกวนในเอกสารภาษาอารบิกและภาษาอื่น ๆ ดังนั้นการใช้ค่าคุณลักษณะของวิธีที่นำเสนอจึงเหมาะสมกับชุดข้อมูลเอกสารภาษาไทยในการทดลองนี้มากกว่ามากกว่า เพื่อความเป็นธรรมจึงทดลองเพิ่มเติมโดยเปลี่ยนค่าคุณลักษณะที่ใช้ในวิธีเอสพีเอ็นแบบสองเฟสเป็นค่าคุณลักษณะของตัวอักษรภาษาไทยที่ใช้ในงานวิจัยนี้ ผลการทดลองเปรียบเทียบประสิทธิภาพของวิธีเอสพีเอ็นแบบสองเฟสเมื่อใช้ค่าคุณลักษณะเดิมเปรียบเทียบกับค่าคุณลักษณะโครงสร้างภาษาไทยและวิธีที่นำเสนอ ที่ระดับความเชื่อมั่นร้อยละ 95 แสดงดังตารางที่ 4.5 ซึ่งแสดงให้เห็นว่าประสิทธิภาพในการจำแนกของวิธีเอสพีเอ็นแบบสองเฟสเมื่อเปลี่ยนมาใช้ค่าคุณลักษณะที่นำเสนอให้ผลดีขึ้นกว่าเมื่อใช้ค่าคุณลักษณะเดิม แต่ก็ยังไม่ดีไปกว่าวิธีที่นำเสนอ

ทั้งวิธีเอสพีเอ็นแบบสองเฟสและวิธีลดสิ่งรบกวนด้วยการเรียนรู้กึ่งมีผู้สอนวิธีจัดกลุ่มและ

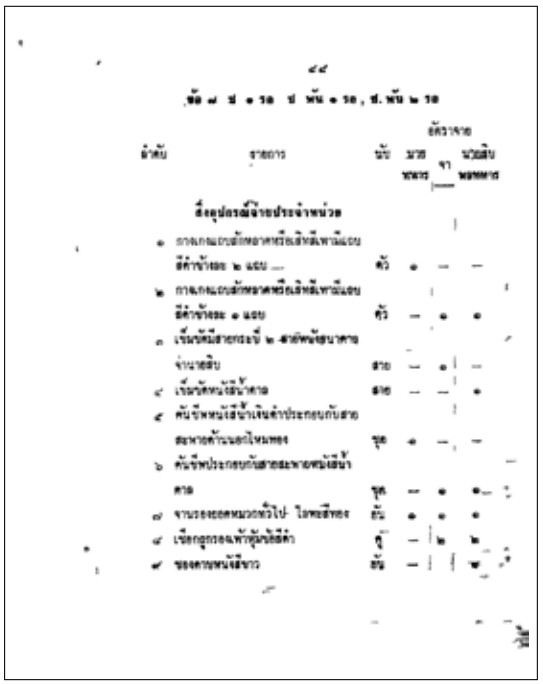
(a)



(b)



(c)



(d)

ภาพที่ 4.4: เปรียบเทียบภาพเอกสารผลลัพธ์จากกระบวนการลดสิ่งรบกวนด้วยวิธีต่าง ๆ (a) ภาพต้นฉบับ (b) ผลลัพธ์จากโปรแกรม ScanFix Xpress 6.0, (c) ผลลัพธ์จากวิธีเอสพีเอ็นแบบสองเฟส และ (d) ผลลัพธ์จากการเรียนรู้กึ่งมีผู้สอนวิธีจัดกลุ่มและติดป้ายซึ่งติดป้ายกำกับด้วยวิธีแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือก

(a)

(b)

(c)

(d)

ภาพที่ 4.5: เปรียบเทียบภาพเอกสารผลลัพธ์จากกระบวนการลดสิ่งรบกวนด้วยวิธีต่าง ๆ (a) ภาพต้นฉบับ (b) ผลลัพธ์จากโปรแกรม ScanFix Xpress 6.0, (c) ผลลัพธ์จากวิธีเอสพีเอ็นแบบสองเฟส และ (d) ผลลัพธ์จากวิธีลดสิ่งรบกวนด้วยการเรียนรู้กึ่งมีผู้สอนวิธีจัดกลุ่มและติดป้ายซึ่งติดป้ายกำกับด้วยวิธีแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือก

วิธีลดสิ่งรบกวน	ค่าความถูกต้อง	ค่าเอฟของคลาสตัวอักษร	ค่าเอฟของคลาสสิ่งรบกวน
วิธีเอสพีเอ็นแบบสองเฟส ด้วยค่าคุณลักษณะเดิม	91.7 $\pm$ 0.27	70.80 $\pm$ 0.44	95.16 $\pm$ 0.2
วิธีเอสพีเอ็นแบบสองเฟส ด้วยค่าคุณลักษณะโครงสร้างภาษาไทย	95.81 $\pm$ 0.19	82.80 $\pm$ 0.37	97.62 $\pm$ 0.15
วิธีแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือก	**96.33 $\pm$ 0.18**	**86.58 $\pm$ 0.33**	**97.87 $\pm$ 0.14**

ตารางที่ 4.5: เปรียบเทียบประสิทธิภาพของวิธีเอสพีเอ็นแบบสองเฟสระหว่างการใช้ค่าคุณลักษณะเดิม กับการใช้ค่าคุณลักษณะโครงสร้างภาษาไทย และเทียบกับวิธีลดสิ่งรบกวนด้วยการเรียนรู้กึ่งมีผู้สอนวิธีจัดกลุ่มและติดป้ายซึ่งติดป้ายกำกับด้วยวิธีแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือก

ติดป้ายซึ่งติดป้ายกำกับด้วยวิธีแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือกใช้การเรียนรู้แบบมีผู้สอนร่วมกับการเรียนรู้แบบไม่มีผู้สอน ข้อแตกต่างคือลำดับการใช้วิธีดังกล่าว โดยวิธีที่นำเสนอใช้ใช้การเรียนรู้แบบไม่มีผู้
สอนเพื่อช่วยติดป้ายกำกับตัวอย่าง หลังจากนั้นจึงใช้การเรียนรู้แบบมีผู้สอนเพื่อสร้างตัวจำแนกสุดท้าย ในทางกลับกันวิธีเอสพีเอ็นแบบสองเฟสใช้การเรียนรู้แบบมีผู้สอนเพื่อจำแนกองค์ประ-กอบประเภท PTC และ non-PTC ในขั้นตอนแรก แล้วจึงใช้ค่าคุณลักษณะขององค์ประกอบประเภท PTC ในการเลือกตัวอักษรในกลุ่ม non-PTC ด้วยการเรียนรู้แบบไม่มีผู้สอน เนื่องจากวิธีเอสพีเอ็นแบบสองเฟสใช้การเรียนรู้แบบมีผู้สอนในขั้นตอนแรก จึงจำเป็นต้องใช้ตัวอย่างมีป้ายกำกับจำนวนมากเพื่อให้เพียงพอต่อการนำไปใช้สร้างตัวจำแนก จึงจำเป็นต้องใช้แรงงานในการติดป้ายกำกับจำนวนมาก ต่างจากวิธีที่นำเสนอซึ่งใช้การจัดกลุ่มเพื่อช่วยเพิ่มตัวอย่างมีป้ายกำกับเพื่อใช้ในการสร้างตัวจำแนกในขั้นตอนสุดท้าย แรงงานที่ใช้เพื่อติดป้ายกำกับของวิธีที่นำเสนอจึงน้อยกว่าวิธีเอสพีเอ็นสองเฟส ดังแสดงในตารางที่ 4.6 ตัวอย่างที่ใช้ในการเรียนรู้ทั้งหมด 40,842 องค์ประกอบนั้นไม่มีป้ายกำกับตั้งแต่ต้น วิธีที่นำเสนอใช้แรงงานคลิกติดป้ายกำกับทั้งสิ้น 1,485 ครั้งแต่วิธีเอสพีเอ็นสองเฟสใช้แรงงานทั้งสิ้น 31,482 ครั้ง (โดยนับจำนวนครั้งในการติดป้ายกำกับใหม่เมื่อผู้ใช้ติดป้ายกำกับผิดแก่บางตัวอย่างด้วย)

จำนวนตัวอย่างที่ถูกติดป้ายกำกับนั้นแสดงในตารางที่ 4.6 เช่นกัน สำหรับวิธีเอสพีเอ็นแบบสองเฟสตัวอย่างชุดเรียนรู้ควรจะถูกติดป้ายกำกับทั้งหมด แต่เนื่องจากตัวอย่างบางตัวมี-ขนาดเล็กมากผู้ใช้จึงอาจไม่เห็นจึงมีบางตัวอย่างที่ไม่ถูกติดป้ายกำกับ ผลลัพธ์จำนวนตัวอย่างที่ติดป้ายกำกับเปรียบเทียบระหว่างวิธีติดป้ายกำกับทั้งสองวิธี พบว่าวิธีการเรียนรู้กึ่งมีผู้สอนวิธีจัดกลุ่มและติดป้ายซึ่งติดป้ายกำกับด้วยวิธีแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือกติดป้ายกำกับได้ทั้งสิ้น 39,216 ตัวอย่างหรือคิดเป็นร้อยละ 96.02 ของตัวอย่างชุดเรียนรู้ ซึ่งมากกว่าการติด

วิธีลดสิ่งรบกวน	จำนวนครั้งของการกดเมาส์	จำนวนตัวอย่าง
วิธีเอสพีเอ็นแบบสองเฟส	32,360	31,482
วิธีแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือก	1,485	39,216

ตารางที่ 4.6: จำนวนแรงงานเพื่อติดป้ายกำกับและจำนวนตัวอย่างที่ถูกติดป้ายกำกับเปรียบเทียบระหว่างวิธีเอสพีเอ็นแบบสองเฟส กับการเรียนรู้กึ่งมีผู้สอนวิธีจัดกลุ่มและติดป้ายซึ่งติดป้ายกำกับด้วยวิธีแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือก

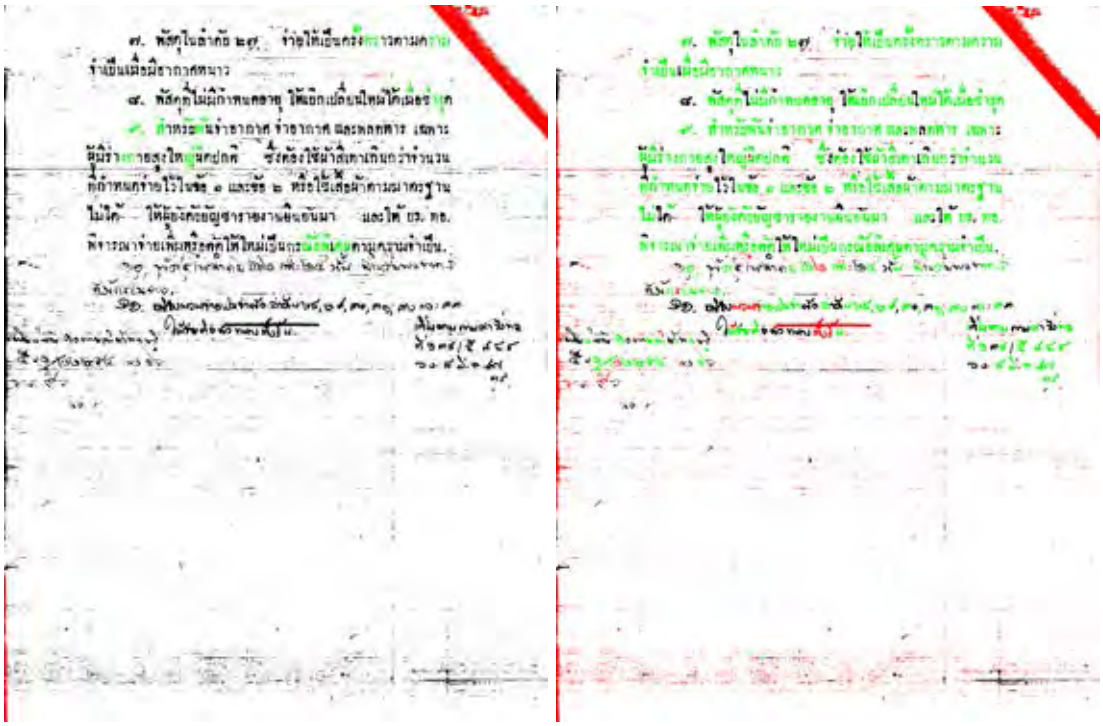
ป้ายกำกับโดยผู้ใช้ด้วยวิธีเอสพีเอ็นแบบสองเฟสซึ่งติดป้ายกำกับทั้งสิ้น 31,482 ตัวอย่างหรือคิดเป็นร้อยละ 77.08 ของตัวอย่างชุดเรียนรู้ และการเรียนรู้กึ่งมีผู้สอนสามารถเพิ่มจำนวนตัวอย่างมีป้ายกำกับได้ถึง 26 เท่าจากจำนวนตัวอย่างที่ติดป้ายกำกับโดยผู้ใช้ โดยภาพที่ 4.6 แสดงภาพเอกสารที่ประกอบไปด้วยตัวอย่างที่ติดป้ายกำกับโดยผู้ใช้ เปรียบเทียบกับตัวอย่างที่ถูกติดป้ายกำกับเพิ่มด้วยการจัดกลุ่มและติดป้ายวิธีที่นำเสนอ

4.4.5 การทดลองบนชุดข้อมูลอื่น ๆ

ผลการทดลองบนชุดข้อมูลการลดสิ่งรบกวนในเอกสารภาษาไทยแสดงให้เห็นว่าวิธีแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือกสามารถปรับปรุงความถูกต้องในการติดป้ายกำกับ และช่วยเพิ่มความถูกต้องในการทำนายของตัวจำแนกกึ่งมีผู้สอนได้ดีกว่าวิธีโหวตคลาสส่วนใหญ่ อย่างไรก็ดีประสิทธิภาพของวิธีที่นำเสนอบนชุดข้อมูลอื่น ๆ ซึ่งเลือกมาจากฐานข้อมูล UCI กลับให้ผลไม่สอดคล้องกัน โดยรายละเอียดของชุดข้อมูล UCI จะนำเสนอในหัวข้อ 5.1.2 ในบทต่อไป

ตารางที่ 4.7 แสดงผลการทดลองเปรียบเทียบประสิทธิภาพของตัวจำแนกกึ่งมีผู้สอนด้วยวิธีที่นำเสนอ เปรียบเทียบกับวิธีโหวตคลาสส่วนใหญ่ที่ค่าขีดแบ่งต่าง ๆ บนชุดข้อมูลอื่น ๆ โดยค่าความถูกต้องนี้เป็นค่าเฉลี่ยจากการทดสอบไขว้ข้ามสิบพับ และเครื่องหมายดอกจัน (*) แสดงอยู่ท้ายค่าความถูกต้องที่สูงที่สุด ผลการทดลองแสดงให้เห็นว่ามีเพียงชุดข้อมูล splice เพียงชุดเดียวที่วิธีที่นำเสนอให้ค่าความถูกต้องของตัวจำแนกกึ่งมีผู้สอนดีกว่าวิธีโหวตคลาสส่วนใหญ่ ในขณะที่ในชุดข้อมูลอื่น ๆ วิธีที่นำเสนอให้ค่าความถูกต้องไม่ดีไปกว่าวิธีโหวตคลาสส่วนใหญ่ โดยอาจเกิดได้จากหลายสาเหตุ

ประการแรก วิธีเรียนรู้โดยแบ่งกลุ่มและติดป้ายนั้นมีสมมติฐานว่าข้อมูลนั้นสามารถจัดกลุ่มโดยสอดคล้องกับคลาสได้ และเมื่อนำวิธีจัดกลุ่มและติดป้ายไปใช้แก้ปัญหาการลดสิ่งรบกวนในเอกสารภาษาไทยนั้นได้ผลดี เนื่องจากลักษณะข้อมูลในชุดข้อมูลเอกสารภาษาไทยมีกลุ่มข้อมูลแฝงในแต่ละคลาส เช่น กลุ่มของตัวอักษรที่คล้ายคลึงกัน กลุ่มของสิ่งรบกวนขนาดเล็ก เป็นต้น แต่ชุดข้อมูลอื่น ๆ ที่นำมาใช้ในการทดลองนี้อาจไม่มีลักษณะเป็นกลุ่มข้อมูล เมื่อสมมติฐานไม่



(a) (b)

ภาพที่ 4.6: ภาพเอกสารที่ประกอบไปด้วยตัวอย่างที่ติดป้ายกำกับ (a) ตัวอย่างถูกติดป้ายกำกับโดยผู้ใช้ และ (b) ตัวอย่างถูกติดป้ายกำกับเพิ่มจากการจัดกลุ่มและติดป้าย โดยตัวอย่างสีดำแสดงถึงตัวอย่างไม่มีป้ายกำกับ ตัวอย่างสีเขียวแสดงถึงตัวอย่างที่ติดป้ายกำกับเป็นคลาสตัวอักษร และตัวอย่างสีแดงแสดงตัวอย่างที่ติดป้ายกำกับเป็นคลาสสิ่งรบกวน

ชุดข้อมูล	ค่าความถูกต้องของวิธี			
	วิธีแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือก	วิธีโหวตตามคลาสส่วนใหญ่ที่กำหนดค่าขีดแบ่งเท่ากับ		
		0.5	0.75	1
banknote	63.06	54.60	73.43	**74.60***
eeg	55.59	55.51	**71.94***	**71.94***
mnist 1vs7	96.03	96.03	96.03	**98.33***
mnist 3vs8	79.55	79.55	79.62	**91.84***
mnist 4vs9	58.68	51.68	**85.41***	**85.41***
mnist 7vs9	89.31	60.24	**90.95***	**90.95***
mushroom	88.76	81.26	93.84	**96.9***
splice	**62.63***	61.84	54.37	60.22

ตารางที่ 4.7: ค่าความถูกต้องของตัวจำแนกกึ่งมีผู้สอนที่ได้จากวิธีแบ่งกลุ่มและติดป้ายเปรียบเทียบระหว่างวิธีติดป้ายกำกับตัวอย่างในกลุ่มคลาสปะปนด้วยวิธีที่นำเสนอ และวิธีโหวตคลาสส่วนใหญ่ที่ค่าขีดแบ่งต่าง ๆ

สอดคล้องกับลักษณะข้อมูลย่อมส่งผลเสียต่อความถูกต้องตัวจำแนก

ประการที่สอง วิธีจัดกลุ่มและติดป้ายจำเป็นต้องกำหนดค่าตัวแปรต่าง ๆ เพื่อให้สามารถจัดกลุ่มได้สอดคล้องกับการกระจายตัวของข้อมูล เช่น วิธีการจัดกลุ่ม วิธีการวัดระยะทาง และจำนวนกลุ่มข้อมูล เป็นต้น ซึ่งในการทดลองนี้ใช้ค่าตัวแปรเพียงค่าเดียวบนทุกชุดข้อมูล ซึ่งอาจไม่เหมาะสมกับหลายชุดข้อมูล และการทดลองเพื่อหาค่าตัวแปรที่เหมาะสมสำหรับแต่ละชุดข้อมูลนั้นต้องใช้ทรัพยากรจำนวนมาก

สาเหตุประการสุดท้าย อาจเพราะการพยายามติดป้ายกำกับในกลุ่มคลาสปะปนทำให้มีบางตัวอย่างที่ติดป้ายกำกับผิดแม้จะมีจำนวนไม่มาก แต่บางชุดข้อมูลอาจอ่อนไหวต่อตัวอย่างที่ติดป้ายกำกับผิด โดยตัวอย่างที่ติดป้ายกำกับผิดบางส่วนนั้นอาจส่งผลเสียต่อตัวจำแนกอย่างมาก เนื่องจากเมื่อวิเคราะห์ผลการทดลองข้างต้นพบว่าวิธีที่มีค่าความถูกต้องของตัวจำแนกสูงที่สุดในชุดข้อมูลส่วนใหญ่ คือ วิธีโหวตคลาสส่วนใหญ่ที่กำหนดค่าขีดแบ่งเท่ากับ 1 ซึ่งวิธีนี้จะติดป้ายกำกับให้แก่ตัวอย่างที่เป็นสมาชิกกลุ่มที่มีตัวอย่างมีป้ายกำกับที่มีเพียงคลาสเดียวเท่านั้นและจะไม่ติดป้ายกำกับตัวอย่างในกลุ่มคลาสปะปนเลย

นอกจากนี้เมื่อเปรียบเทียบวิธีการที่นำเสนอกับการเรียนรู้แบบกึ่งมีผู้สอนวิธีเรียนรู้ด้วยตนเอง ดังแสดงในตารางที่ 4.8 พบว่าวิธีเรียนรู้ด้วยตนเองให้ค่าความถูกต้องที่สูงกว่าวิธีที่นำเสนอในชุดข้อมูลส่วนใหญ่ นอกจากนี้วิธีเรียนรู้ด้วยตนเองนี้ไม่จำเป็นต้องกำหนดค่าตัวแปรดังวิธีจัด

ชุดข้อมูล	ค่าความถูกต้องของวิธี	
	วิธีแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือก	วิธีเรียนรู้ด้วยตนเอง
banknote	63.06	**76.06***
eeg	55.59	**72.16***
mnist 1vs7	96.03	**97.6***
mnist 3vs8	79.55	**92.03***
mnist 4vs9	58.68	**84.94***
mnist 7vs9	89.31	**91.31***
mushroom	88.76	**96.97***
splice	**62.63***	57.34

ตารางที่ 4.8: เปรียบเทียบค่าความถูกต้องของตัวจำแนกกึ่งมีผู้สอนที่ได้จากวิธีแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือกที่นำเสนอและวิธีเรียนรู้ด้วยตนเอง

กลุ่มและติดป้ายอีกด้วย จึงสามารถนำไปใช้กับชุดข้อมูลอื่น ๆ รวมทั้งชุดข้อมูลจริงได้อีกด้วย

ด้วยเหตุนี้ผู้วิจัยจึงศึกษาวิธีเรียนรู้ด้วยตนเองเพิ่มเติม และพบว่าแม้วิธีเรียนรู้ด้วยตนเองนี้จะมีประสิทธิภาพดีมาก แต่ยังมีจุดอ่อนที่การติดป้ายกำกับตัวอย่างไม่มีป้ายกำกับของการเรียนรู้ด้วยตนเอง ซึ่งใช้ตัวจำแนกตั้งต้นที่สร้างจากตัวอย่างมีป้ายกำกับเท่านั้นในการติดป้ายกำกับ ซึ่งอาจติดป้ายกำกับผิดแก่บางตัวอย่างได้เช่นกัน งานวิจัยในส่วนที่สองผู้วิจัยจึงศึกษาวิธีการเพื่อปรับปรุงวิธีเรียนรู้ด้วยตนเอง โดยเสนอการวิเคราะห์ตัวอย่างที่ใช้ในการเรียนรู้เพื่อลดความเสี่ยงในการติดป้ายกำกับผิด วัตถุประสงค์หลักคือเพื่อให้ได้ตัวจำแนกกึ่งมีผู้สอนที่มีประสิทธิภาพดีขึ้น และรายละเอียดของวิธีดังกล่าวจะนำเสนอในบทต่อไป

บทที่ 5

การวิเคราะห์ตัวอย่างที่ใช้ในการเรียนรู้ด้วยการจัดกลุ่มข้อมูลเพื่อปรับปรุงวิธีเรียนรู้ด้วยตนเอง

วิธีเรียนรู้ด้วยตนเองเป็นวิธีที่มีประสิทธิภาพและถูกใช้อย่างแพร่หลายในหลายปัญหา อย่างไรก็ดีขั้นตอนการติดป้ายกำกับตัวอย่างไม่มีป้ายกำกับของการเรียนรู้ด้วยตนเองนี้ใช้ตัวจำแนกตั้งต้นที่สร้างจากตัวอย่างมีป้ายกำกับเท่านั้นในการติดป้ายกำกับ ซึ่งอาจติดป้ายกำกับผิดแก่บางตัวอย่าง และตัวอย่างที่ติดป้ายกำกับผิดนั้นจะถูกนำไปใช้สร้างตัวจำแนกสุดท้าย ซึ่งอาจส่งผลเสียต่อความถูกต้องของการจำแนกได้

งานวิจัยนี้ใช้การวิเคราะห์ตัวอย่างที่ใช้ในการเรียนรู้ โดยประมาณการกระจายตัวของตัวอย่างมีป้ายกำกับและการกระจายตัวของตัวอย่างไม่มีป้ายกำกับด้วยการจัดกลุ่มข้อมูล และเสนอวิธีปรับปรุงชุดตัวอย่างมีป้ายกำกับที่อาจทำให้ความถูกต้องในการติดป้ายกำกับลดลง งานวิจัยนี้ใช้การจัดกลุ่มข้อมูลกึ่งมีผู้สอนเพื่อพิจารณาว่าชุดตัวอย่างมีป้ายกำกับนี้เพียงพอสำหรับการนำไปใช้สร้างตัวจำแนกกึ่งมีผู้สอนหรือไม่ เท่าที่ผู้วิจัยสืบค้นมาไม่พบงานวิจัยอื่นที่เสนอแนวทางในการวิเคราะห์ลักษณะของชุดตัวอย่างมีป้ายกำกับก่อนการนำไปใช้เรียนรู้กึ่งมีผู้สอน อาจกล่าวได้ว่างานวิจัยนี้เป็นงานวิจัยแรกที่เสนอว่าการเรียนรู้กึ่งมีผู้สอนควรมีกระบวนการก่อน (pre-processing) เพื่อวิเคราะห์ตัวอย่างที่จะนำมาใช้ในการเรียนรู้

ลำดับการนำเสนอในบทนี้เป็นดังนี้ หัวข้อ 5.1 อธิบายรายละเอียดชุดข้อมูลและเงื่อนไขที่ใช้ในการเลือกชุดข้อมูล หัวข้อ 5.2 อธิบายวิธีการวัดผล หัวข้อ 5.3 อธิบายรายละเอียดการเลือกตัวอย่างที่มั่นใจมาติดป้ายกำกับในการเรียนรู้ด้วยตนเอง ส่วนสุดท้ายของบทนี้นำเสนอวิธีวิเคราะห์ตัวอย่างมีป้ายกำกับและผลการวิเคราะห์สองวิธี หัวข้อ 5.4 นำเสนอวิธีการและผลการทดลองสำหรับวิธีแรก คือการวิเคราะห์ตัวอย่างมีป้ายกำกับโดยกำหนดค่าความเชื่อมั่นแก่ตัวอย่างมีป้ายกำกับด้วยการทดสอบแบบไขว้ข้ามดึงออกหนึ่งตัว ส่วนหัวข้อ 5.5 นำเสนอวิธีที่สอง คือการวิเคราะห์ชุดตัวอย่างมีป้ายกำกับด้วยการวิเคราะห์กลุ่มข้อมูล และในหัวข้อสุดท้ายนำเสนอวิธีปรับปรุงชุดตัวอย่างมีป้ายกำกับ และผลการปรับปรุงชุดตัวอย่างมีป้ายกำกับด้วยวิธีเพิ่มป้ายกำกับโดยผู้ใช้และวิธีเพิ่มป้ายกำกับด้วยตัวจำแนกต่าง ๆ

5.1 ชุดข้อมูล

หัวข้อนี้ประกอบไปด้วยสามหัวข้อย่อย ได้แก่ การเลือกชุดข้อมูล รายละเอียดชุดข้อมูลจากฐานข้อมูล UCI Machine Learning repository และรายละเอียดชุดข้อมูลจริงจากปัญหา

การลดสิ่งรบกวนในภาพเอกสารภาษาไทย

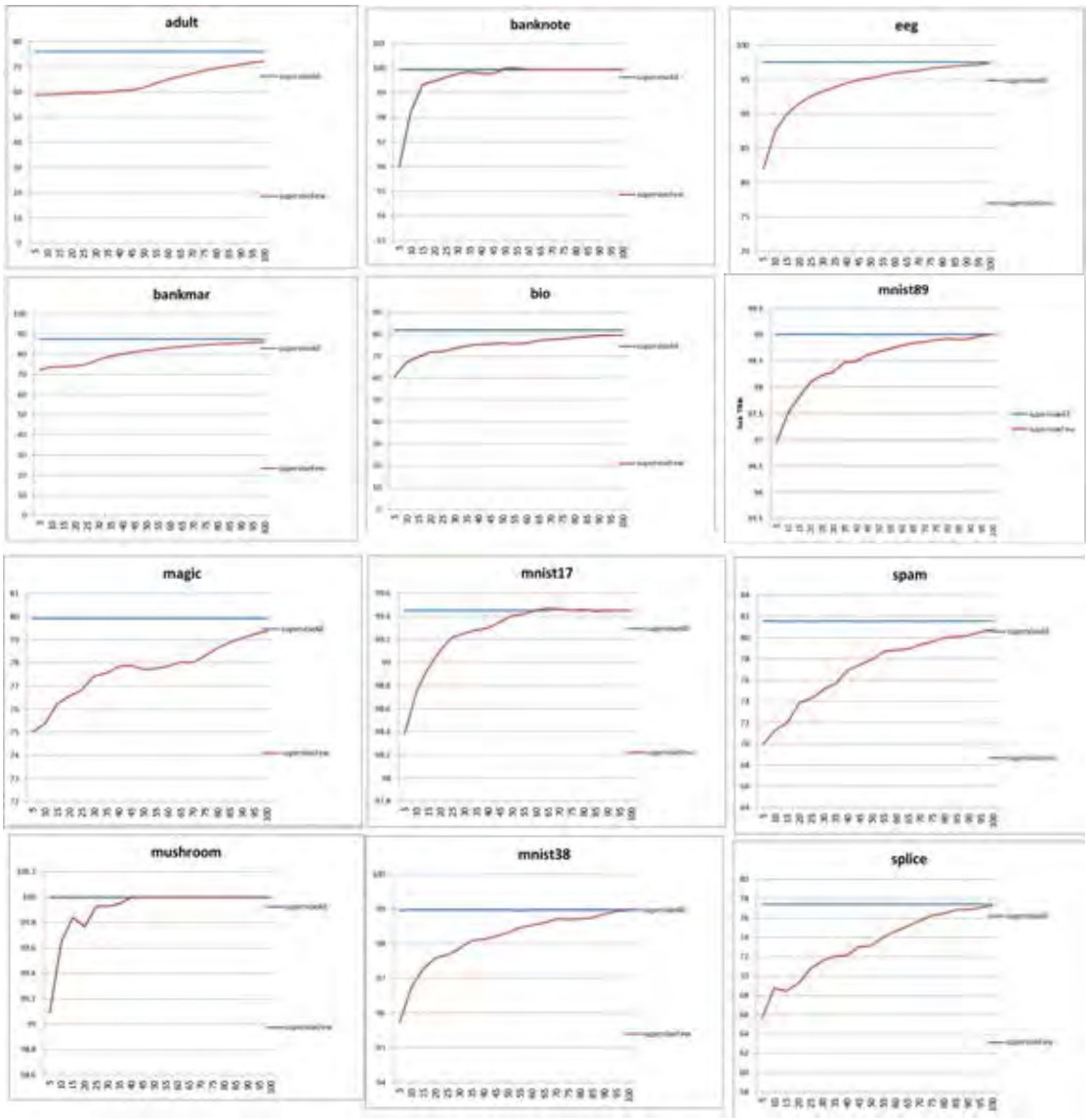
5.1.1 การเลือกชุดข้อมูล

งานวิจัยที่สร้างแบบจำลองในการเรียนรู้ของเครื่องส่วนใหญ่เชื่อว่าการเพิ่มจำนวนตัวอย่างเพื่อใช้สอน (training) ตัวเรียนรู้หรือตัวจำแนก จะทำให้ได้ตัวจำแนกที่มีประสิทธิภาพในการจำแนกที่ดีขึ้น การเรียนรู้กึ่งมีผู้สอนโดยการเรียนรู้ด้วยตนเองจึงพยายามเพิ่มตัวอย่างสำหรับใช้ในการสร้างตัวจำแนกนั้น แต่สมมติฐานนี้อาจไม่เฉพาะเจาะจงเกินกว่าจะนำไปใช้งานจริงได้ เช่นงานวิจัยของ Guo (Guo et al., 2010) กลับพบว่าการเพิ่มตัวอย่างในการเรียนรู้อาจทำให้ได้ตัวจำแนกที่มีประสิทธิภาพแย่ลงได้ในบางชุดการทดลอง ซึ่งสอดคล้องกับผลการทดลองที่พบในงานวิจัยนี้

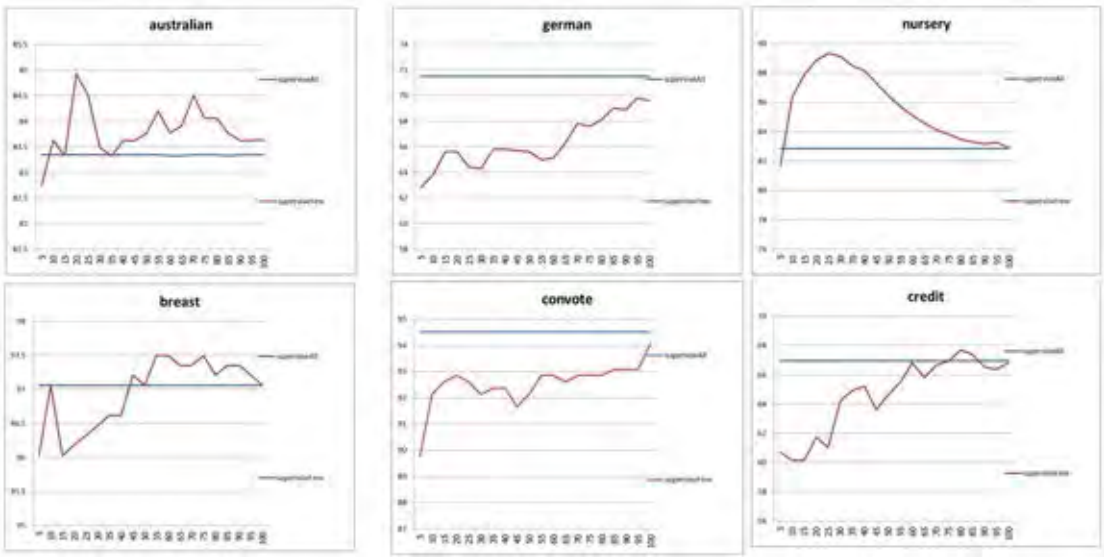
เมื่อทดลองเพิ่มจำนวนตัวอย่างเพื่อใช้สอนตัวจำแนกเพื่อนบ้านใกล้ที่สุดสามตัว โดยเพิ่มตัวอย่างรอบละ 5% ของตัวอย่างทั้งหมด ผลการทดลองแสดงให้เห็นว่าในบางชุดข้อมูลเมื่อเพิ่มตัวอย่างมีป้ายกำกับที่ถูกต้องทำให้ได้ตัวจำแนกที่มีประสิทธิภาพดีขึ้นจริง ตัวอย่างชุดข้อมูลดังกล่าวแสดงในภาพที่ 5.1 แต่ในบางชุดข้อมูลประสิทธิภาพของตัวจำแนกกลับแย่ลงหรือผันผวนเมื่อเพิ่มจำนวนตัวอย่างในการเรียนรู้มากขึ้น ตัวอย่างชุดข้อมูลดังกล่าวแสดงในภาพที่ 5.2

สาเหตุที่บางชุดข้อมูลการเพิ่มตัวอย่างในการเรียนรู้กลับทำให้ได้ตัวจำแนกที่ดีขึ้นหรือแย่ลงขึ้นกับหลายปัจจัย เช่น ความสอดคล้องระหว่างชุดข้อมูลกับขั้นตอนวิธี คุณลักษณะที่นำมาใช้ ความซับซ้อนของปัญหา สมมติฐานที่ใช้ และการกระจายตัวของข้อมูล เป็นต้น ในขณะที่งานของ Guo พบว่าชุดข้อมูล mushroom ซึ่งเป็นชุดข้อมูลจากฐานข้อมูล UCI Machine Learning repository (Lichman, 2013) นั้นให้ผลแย่ลงเมื่อเพิ่มตัวอย่างสำหรับการเรียนรู้ แต่ในงานวิจัยนี้กลับพบว่าผลเป็นในทางตรงข้าม ข้อแตกต่างระหว่างงานของ Guo และงานวิจัยนี้คือตัวจำแนกที่ใช้ โดยงานของ Guo ใช้ตัวจำแนกเบย์แต่ในงานวิจัยนี้ใช้ตัวจำแนกเพื่อนบ้านใกล้ที่สุด ตัวจำแนกที่แตกต่างกันนี้อาจให้ผลการจำแนกที่แตกต่างกันได้ หรือกล่าวได้ว่าชุดข้อมูล mushroom นี้เหมาะสมกับการใช้กับตัวจำแนกเพื่อนบ้านใกล้ที่สุดมากกว่าตัวจำแนกเบย์

งานวิจัยนี้จึงเลือกชุดข้อมูลโดยทดสอบความสอดคล้องระหว่างชุดข้อมูล กับตัวจำแนกเพื่อนบ้านใกล้ที่สุด ซึ่งเป็นตัวจำแนกที่ใช้ในการเรียนรู้กึ่งมีผู้สอนในงานวิจัยนี้ โดยชุดข้อมูลที่ใช้ในการทดลองประกอบไปด้วย 2 ชุด ได้แก่ ชุดข้อมูลจากฐานข้อมูล UCI และ ชุดข้อมูลจริงจากปัญหาการลดสิ่งรบกวนในภาพเอกสารภาษาไทย



ภาพที่ 5.1: กราฟแสดงประสิทธิภาพในการจำแนกที่เพิ่มขึ้นตามจำนวนตัวอย่าง



ภาพที่ 5.2: กราฟแสดงประสิทธิภาพในการจำแนกที่ลดลงเมื่อเพิ่มจำนวนตัวอย่าง

5.1.2 ชุดข้อมูลจากฐานข้อมูล UCI Machine Learning repository

การทดลองนี้เลือกชุดข้อมูลปัญหาสองคลาสที่ไม่มีตัวอย่างที่ไม่มีค่าคุณลักษณะ (missing value) สำหรับชุดข้อมูลการรู้จำตัวเลขจากลายมือเขียน (MNIST) ซึ่งเป็นปัญหาการจำแนกภาพตัวเลข 10 รูปแบบเป็นปัญหาหลายคลาส การทดลองนี้เลือกตัวอย่างเพื่อจำแนกคู่ของตัวเลขที่ยากต่อการจำแนกตามงานวิจัยของ Valizadegan (Valizadegan and Jin, 2007) และ Zhang (Zhang et al., 2009) สัดส่วนของจำนวนตัวอย่างแต่ละคลาสในชุดข้อมูลมีทั้งสมดุลและไม่สมดุลในอัตราส่วนต่าง ๆ รายละเอียดชุดข้อมูลทั้ง 16 ชุดแสดงในตารางที่ 5.1

5.1.3 ชุดข้อมูลจริงจากปัญหาการลดสิ่งรบกวนในภาพเอกสารภาษาไทย

กำหนดให้หนึ่งชุดข้อมูลคือหนึ่งภาพเอกสาร การทดลองนี้เลือกภาพเอกสาร 67 ภาพเอกสารที่มีตัวอย่างในคลาสส่วนน้อยไม่น้อยกว่าร้อยละ 20 ของตัวอย่างทั้งหมดในภาพเอกสารนั้น ตัวอย่างในชุดข้อมูลนี้ประกอบไปด้วยสองคลาส คือ คลาสตัวอักษร และคลาสสิ่งรบกวน และมีจำนวนคุณลักษณะ 9 คุณลักษณะ ดังรายละเอียดในหัวข้อที่ 4.1 ตัวอย่างในแต่ละชุดข้อมูลมีจำนวนตั้งแต่ 510 ถึง 7930 ตัวอย่าง ดังรายละเอียดในตารางที่ 5.2

5.2 การวัดผล

การทดลองในแต่ละชุดการทดลองและแต่ละวิธีใช้การทดสอบไขว้ข้าม (cross validation) วัดประสิทธิภาพในสองส่วน ส่วนแรกคือประสิทธิภาพของการติดป้ายกำกับ โดยวัดจาก

ชื่อชุดข้อมูล	จำนวนคุณลักษณะ	จำนวนตัวอย่าง	อัตราส่วนตัวอย่างแต่ละคลาส
ชุดข้อมูลขนาดเล็ก			
banknote	4	1370	56:44
bioNRB	41	1050	66:34
splice	60	2160	52:48
wilt	5	4830	95:5
ชุดข้อมูลขนาดกลาง			
adult	104	32550	76:24
bankmarket	43	45210	88:12
eeg	14	14980	55:45
magicgamma	10	19020	65:35
mushroom	121	8120	50:50
spam	57	4590	60:40
ชุดข้อมูลขนาดใหญ่			
mnist 1vs7	784	13000	52:48
mnist 2vs7	784	12220	51:49
mnist 3vs8	784	11980	51:49
mnist 4vs9	784	11790	50:50
mnist 7vs9	784	12210	51:49
mnist 8vs9	784	11800	50:50

ตารางที่ 5.1: รายละเอียดชุดข้อมูลจากฐานข้อมูล UCI

ชื่อชุดข้อมูล	จำนวนตัวอย่างประเภท	
	ตัวอักษร	สิ่งรบกวน
noise001	731	329
noise004	517	1313
noise005	1036	304
noise006	746	294
noise007	937	3443
noise008	1240	2170
noise009	1196	2004
noise013	560	270
noise016	671	309
noise020	821	519
noise022	582	848
noise023	760	480
noise024	648	232
noise026	241	549
noise027	814	2126
noise028	174	686
noise033	889	721
noise036	566	584
noise038	790	2940
noise039	898	2642
noise040	995	2195
noise048	1970	5860
noise070	685	2315
noise071	612	1468
noise072	569	701
noise073	540	1010
noise074	442	798
noise077	632	2448
noise100	797	2763
noise112	1106	1534
noise113	609	1541
noise117	1798	2162
noise118	830	750
noise120	2022	1568

ชื่อชุดข้อมูล	จำนวนตัวอย่างประเภท	
	ตัวอักษร	สิ่งรบกวน
noise121	936	2124
noise122	1042	2638
noise125	2029	5901
noise126	944	1026
noise131	1295	1695
noise133	1454	1386
noise134	1880	2330
noise135	1032	998
noise147	1545	5045
noise148	1228	3642
noise150	936	2854
noise152	960	3490
noise154	1447	2403
noise155	1773	3767
noise156	1021	3039
noise158	423	1437
noise160	987	903
noise161	1306	424
noise162	683	247
noise163	436	224
noise164	1528	592
noise165	371	809
noise166	575	385
noise170	901	379
noise171	798	442
noise172	1053	507
noise173	989	311
noise174	516	194
noise177	356	734
noise180	347	163
noise184	917	283
noise185	841	219
noise188	549	141

ตารางที่ 5.2: รายละเอียดชุดข้อมูลการจำแนกสิ่งรบกวนในภาพเอกสารภาษาไทย

ค่าความถูกต้องในการติดป้ายกำกับ (labeling accuracy) ซึ่งคำนวณจากสมการ 5.1 และความแม่นยำในการติดป้ายกำกับ (labeling precision) ซึ่งคำนวณจากสมการ 5.2

$$\text{Labeling accuracy} = \frac{\text{number of correctly label data}}{\text{number of unlabeled data}} \tag{5.1}$$

$$\text{Labeling precision} = \frac{\text{number of correctly label data}}{\text{number of newly label data}} \tag{5.2}$$

ส่วนที่สองวัดประสิทธิภาพของตัวจำแนกกึ่งมีผู้สอนที่สร้างจากตัวอย่างติดป้ายกำกับที่แตกต่างกัน โดยวัดค่าความถูกต้องของการทำนายคลาสตัวอย่างตามสมการที่ 5.3

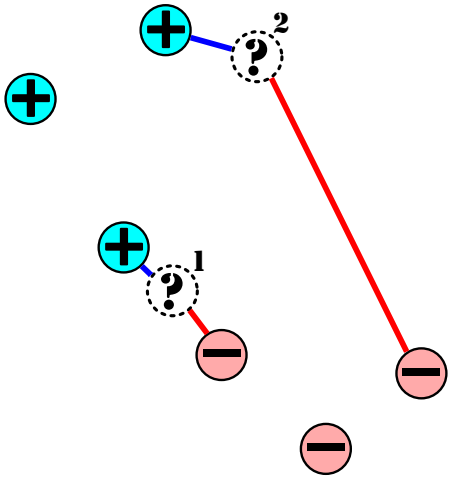
$$\text{Accuracy} = \frac{\text{number of correctly predict data}}{\text{number of data}} \tag{5.3}$$

5.3 วิธีเลือกตัวอย่างเพื่อติดป้ายกำกับในการเรียนรู้ด้วยตนเอง

การเรียนรู้ด้วยตนเองนั้นใช้ตัวอย่างมีป้ายกำกับตั้งต้นในการเลือกติดป้ายกำกับตัวอย่างไม่มีป้ายกำกับที่มั่นใจที่สุด ซึ่งการคำนวณค่าความมั่นใจ (confident value) สำหรับตัวจำแนกแต่ละขั้นตอนวิธีนั้นก็แตกต่างกัน ในหัวข้อนี้อธิบายวิธีคำนวณค่าความมั่นใจสำหรับตัวจำแนกเนื่องจากงานวิจัยนี้มุ่งเน้นศึกษาอิทธิพลของตัวอย่างมีป้ายกำกับต่อการจำแนกกึ่งมีผู้สอน จึงเลือกใช้ตัวจำแนกขึ้นกับตัวอย่างเช่นเดียวกับในหัวข้อก่อนหน้า โดยใช้ตัวจำแนกเพื่อนบ้านใกล้ที่สุดเคจำนวน โดยทดลองที่ค่าเคเท่ากับ 1 และ 3

การคำนวณค่าความมั่นใจสำหรับตัวจำแนกเพื่อนบ้านใกล้ที่สุดนั้นคำนวณจากระยะห่างของตัวอย่างไปยังตัวอย่างมีป้ายกำกับแต่ละคลาสที่ใกล้ที่สุด (Li and Zhou, 2005) ตัวอย่างที่มีระยะห่างระหว่างสองคลาสที่ใกล้กันที่สุดแตกต่างกันมากที่สุด หรือเป็นตัวอย่างที่ใกล้คลาสหนึ่งที่สุดและอยู่ห่างจากอีกคลาสที่สุด ดังภาพที่ 5.3

กำหนดให้คลาสที่เป็นไปได้คือ $c_1, c_2, ...c_i \in C$ ตัวอย่างมีป้ายกำกับของคลาส c_1 คือ $L_{c_1} = x_{c_1}, y_{c_1}$ ตัวอย่างไม่มีป้ายกำกับคือ $U = x_u$ ตัวอย่างไม่มีป้ายกำกับที่มั่นใจที่จะติดป้ายกำกับที่สุดคือ x_{conf} ค่าความมั่นใจและการเลือกตัวอย่างที่มั่นใจที่สุดสำหรับเพื่อนบ้านใกล้ที่สุดคำนวณดังสมการที่ 5.4 ถึงสมการที่ 5.8 สำหรับเพื่อนบ้านใกล้ที่สุดเคตัวนั้นในขั้นตอนการหาระยะห่างไปยังแต่ะคลาสคำนวณจากระยะห่างเฉลี่ยไปยังเพื่อนบ้านแต่ละคลาสเคตัวอย่าง



ภาพที่ 5.3: ตัวอย่างค่าความมั่นใจของตัวอย่างไม่มีป้ายกำกับหมายเลข 1 และหมายเลข 2 ในกรณีนี้ แม้ว่าตัวอย่างหมายเลข 1 จะอยู่ใกล้ตัวอย่างมีป้ายกำกับ
คลาสบวกมากกว่าแต่ก็ไม่ห่างจากตัวอย่างในคลาสลบ ตัวอย่างหมายเลข 2 ที่อยู่ใกล้คลาสบวกและไกลคลาสลบจึงมีค่าความมั่นใจมากกว่าและจะถูกติดป้ายกำกับในรอบนี้

$$D_{i-c1} = min(dist(x_i, x_j)); \forall x_j \in L_{c_1} \tag{5.4}$$

$$D_{1stNearest} = min(D_p); \forall p \in C \tag{5.5}$$

$$D_{2ndNearest} = min(D_q); \forall q \in C - C_{1stNearest} \tag{5.6}$$

$$Conf_i = |D_{1stNearest} - D_{2ndNearest}| \tag{5.7}$$

$$x_{conf} = x_i; \text{when } Conf_{x_i} = max(Conf_i); \forall x_i \in U \tag{5.8}$$

อย่างไรก็ดีการเลือกตัวอย่างติดป้ายกำกับโดยเลือกตัวอย่างที่ตัวจำแนกมั่นใจที่สุดนั้นจำเป็นต้องใช้ทรัพยากรเพื่อคำนวณค่าความมั่นใจสูงมาก อีกทั้งจำนวนตัวอย่างที่เลือกในแต่ละรอบที่เหมาะสมและการหยุดการเลือกตัวอย่างนั้นยังเป็นเป็นคำถามที่ไม่ทราบคำตอบ นอกจากนั้นผลการทดลองของ Guo แสดงให้เห็นว่าการเลือกตัวอย่างที่มั่นใจที่สุดนั้นให้ผลลัพธ์ไม่แตกต่างจากการเลือกตัวอย่างด้วยการสุ่ม (Guo et al., 2010) วิธีเลือกตัวอย่างที่มั่นใจที่สุดนี้จึงอาจไม่ใช่วิธีที่เหมาะสมที่สุด

งานวิจัยนี้จึงทดลองเพื่อเปรียบเทียบประสิทธิภาพระหว่าง วิธีที่หนึ่งเลือกตัวอย่างที่มั่นใจที่สุดมาติดป้ายกำกับแล้วสร้างตัวจำแนกใหม่เป็นรอบ ๆ กับวิธีที่สองติดป้ายกำกับตัวอย่างทั้งหมด โดยทดสอบไขว้ข้ามสิบพับด้วยการเรียนรู้กึ่งมีผู้สอนวิธีเรียนรู้ด้วยตนเองด้วยตัวจำแนกเพื่อนบ้านใกล้ที่สุดสามตัว กำหนดให้จำนวนตัวอย่างมีป้ายกำกับเท่ากับร้อยละหนึ่งของจำนวน

ชุดข้อมูล	เลือกตัวอย่างทั้งหมด	เลือกตัวอย่างที่มั่นใจที่สุด
banknote	**85.33*** **	59.71
bioNRB	**47.81**	45.05
mushroom	**94.50*** **	91.90
spam	**66.17**	65.43
splice	**56.77*** **	49.97
wilt	62.73	**70.89**

ตารางที่ 5.3: เปรียบเทียบความถูกต้องของตัวจำแนกกึ่งมีผู้สอนระหว่างวิธีเลือกตัวอย่างทั้งหมดกับวิธีเลือกตัวอย่างที่มั่นใจที่สุดในแต่ละรอบ

ตัวอย่างทั้งหมด ผลการทดลองแสดงในตารางที่ 5.3 พบว่าการเลือกตัวอย่างที่มั่นใจที่สุดทีละรอบให้ผลไม่ดีไปกว่าวิธีเลือกตัวอย่างทั้งหมด โดยการเลือกตัวอย่างทั้งหมดดีกว่าการเลือกตัวอย่างที่มั่นใจอย่างมีนัยสำคัญในสองชุดข้อมูล และสองวิธีนี้ให้ผลไม่แตกต่างกันในชุดข้อมูลที่เหลือ

สาเหตุส่วนหนึ่งเนื่องจากการเลือกตัวอย่างที่มั่นใจที่สุดมาติดป้ายกำกับนั้นไม่สามารถรับรองได้ว่าตัวอย่างที่เลือกมานั้นจะถูกติดป้ายกำกับอย่างถูกต้อง ตัวอย่างที่ถูกติดป้ายกำกับผิดในขั้นตอนแรกจะส่งผลต่อการติดป้ายกำกับในขั้นตอนต่อไป และตัวอย่างที่ติดป้ายกำกับผิดสะสมทั้งหมดจะส่งผลต่อการสร้างตัวจำแนกสุดท้าย การเลือกตัวอย่างที่มั่นใจที่สุดจึงเหมาะสมสำหรับในบางชุดข้อมูลเท่านั้น ดังนั้นในการทดลองต่อไปในงานวิจัยนี้ใช้วิธีเลือกติดป้ายกำกับแก่ตัวอย่างไม่มีป้ายกำกับทั้งหมด

5.4 การวิเคราะห์ตัวอย่างมีป้ายกำกับโดยกำหนดค่าน้ำหนักแก่ตัวอย่างมีป้ายกำกับด้วยการทดสอบแบบไขว้ข้ามดึงออกหนึ่งตัว

ในหัวข้อนี้ประกอบไปด้วยสามหัวข้อย่อย ได้แก่ แรงจูงใจ การให้ค่าน้ำหนักตัวอย่างมีป้ายกำกับ และการวิเคราะห์ค่าน้ำหนักตัวอย่างมีป้ายกำกับ

5.4.1 แรงจูงใจ

การทดลองเบื้องต้นบนชุดข้อมูล banknote ซึ่งมีตัวอย่างมีป้ายกำกับคลาสละ 6 ตัวอย่าง รวมเป็นจำนวนตัวอย่างมีป้ายกำกับทั้งหมด 12 ตัวอย่างหรือคิดเป็นร้อยละหนึ่งของตัวอย่างชุดเรียนรู้ โดยทดสอบไขว้ข้ามสิบพับด้วยการเรียนรู้กึ่งมีผู้สอนวิธีเรียนรู้ด้วยตนเองด้วยตัวจำแนกเพื่อนบ้านใกล้ที่สุดสามตัว เพื่อเปรียบเทียบประสิทธิภาพตัวจำแนกกึ่งมีผู้สอนด้วยวิธีเรียนรู้ด้วยตนเอง (Semi-supervised classifier from self-training: SemiSelf) กับตัวจำแนกพื้นฐานสอง

พับที่	ค่าความถูกต้องของ		
	ตัวจำแนกมีผู้สอนโดยใช้		ตัวจำแนกกึ่งมีผู้สอน
	ตัวอย่างทั้งหมด	ตัวอย่างมีป้ายกำกับ	วิธีเรียนรู้ด้วยตนเอง
1	100	82.48	84.67
2	100	91.24	91.24
3	100	81.02	81.75
4	100	93.43	94.16
5	100	91.24	89.05
6	100	83.94	83.94
7	99.27	81.02	81.02
8	100	90.51	91.24
9	100	86.86	87.59
10	100	56.2	56.93

ตารางที่ 5.4: ค่าความถูกต้องของตัวจำแนกบนสิบพับ

ตัวจำแนก ได้แก่ ตัวจำแนกที่สร้างจากการเรียนรู้แบบมีผู้สอนด้วยตัวอย่างในชุดเรียนรู้ทั้งหมด (Supervised classifier from all training data: SupervisedAll) ซึ่งน่าจะมีประสิทธิภาพสูงที่สุด และตัวจำแนกที่สร้างจากการเรียนรู้แบบมีผู้สอนด้วยตัวอย่างที่ติดป้ายกำกับเท่านั้น (Supervised classifier from labeled data: SupervisedLabel) ซึ่งน่าจะมีประสิทธิภาพต่ำที่สุด

ตารางที่ 5.4 แสดงผลการทดลองเปรียบเทียบตัวจำแนกสามตัว พบว่าตัวจำแนกที่สร้างจากวิธี SupervisedLabel และวิธี SemiSelf ในพับที่ 10 มีประสิทธิภาพแตกต่างจากพับอื่นมาก โดยผลการทำนายในพับที่ 10 นั้นมีค่าความถูกต้องของการทำนายประมาณร้อยละ 56 ในขณะที่พับอื่น ๆ ค่าความถูกต้องของการทำนายมากกว่าร้อยละ 80 และบางพับค่าความถูกต้องของการทำนายสูงถึงร้อยละ 90

จากการทดลองพบว่าค่าความถูกต้องในการทำนายบนตัวอย่างทดสอบด้วยวิธี SupervisedLabel ส่งผลต่อค่าความถูกต้องของวิธี SemiSelf เมื่อวิธี SupervisedLabel มีค่าความถูกต้องต่ำเช่นในพับที่ 10 จะส่งผลให้วิธี SemiSelf มีค่าความถูกต้องต่ำไปด้วย เนื่องจากตัวจำแนกตั้งต้นของวิธี SemiSelf นั้นสร้างจากตัวอย่างมีป้ายกำกับเท่านั้นซึ่งคือตัวจำแนกจากวิธี SupervisedLabel นั่นเอง ตัวจำแนกตั้งต้นนี้จะถูกใช้เพื่อติดป้ายกำกับให้แก่ตัวอย่างไม่มีป้ายกำกับ ตัวจำแนกตั้งต้นที่มีประสิทธิภาพไม่ดีจึงอาจส่งผลให้การติดป้ายกำกับตัวอย่างนั้นไม่ถูกต้อง ทั้งนี้ค่าความถูกต้องในการติดป้ายกำกับในพับที่ 10 เท่ากับร้อยละ 63.47 ซึ่งค่อนข้างต่ำประสิทธิภาพของตัวจำแนกที่ไม่ดีนี้น่าจะได้รับอิทธิพลจากตัวอย่างมีป้ายกำกับบางกลุ่มที่เมื่อนำ

ตัวอย่างกลุ่มนั้นไปพิจารณาจะส่งผลให้ติดป้ายกำกับผิด ด้วยสมมติฐานดังกล่าวงานวิจัยนี้จึงออกแบบการทดลองเพื่อให้ค่าน้ำหนักแก่ตัวอย่างมีป้ายกำกับ เพื่อลดค่าน้ำหนักแก่ตัวอย่างมีป้ายกำกับบางตัวที่มีแนวโน้มจะทำให้การติดป้ายกำกับนั้นไม่ถูกต้อง เนื่องจากการทดลองนี้มีสมมติฐานว่าตัวอย่างมีป้ายกำกับบางตัวส่งผลเสียต่อการติดป้ายกำกับด้วยตนเอง จึงออกแบบการทดลองเพื่อวิเคราะห์คุณภาพตัวอย่างมีป้ายกำกับแต่ละตัวอย่าง และให้ค่าน้ำหนักตัวอย่างมีป้ายกำกับโดยลดค่าน้ำหนักตัวอย่างที่มีแนวโน้มว่าจะส่งผลเสียและเพิ่มค่าน้ำหนักตัวอย่างที่มีแนวโน้มจะส่งผลดีต่อการติดป้ายกำกับตัวอย่างอื่น ๆ

5.4.2 การให้ค่าน้ำหนักตัวอย่างมีป้ายกำกับ

การให้ค่าน้ำหนักตัวอย่างมีป้ายกำกับนั้นใช้วิธีทดสอบไขว้ข้ามแบบดึงออกหนึ่งตัว (leave-one-out cross validation) บนตัวอย่างมีป้ายกำกับทั้งหมด เพื่อทดสอบคุณภาพของตัวอย่างมีป้ายกำกับแต่ละตัวและวิเคราะห์อิทธิพลของตัวอย่างมีป้ายกำกับต่อการทำนายตัวอย่างมีป้ายกำกับอื่น ๆ วิธีนี้สอดคล้องกับงานของ Guo (Guo et al., 2011) ซึ่งวัดประสิทธิภาพของตัวอย่างที่ติดป้ายกำกับใหม่ก่อนการนำตัวอย่างนั้นไปใช้ โดยวัดประสิทธิภาพบนการทำนายตัวอย่างมีป้ายกำกับตั้งต้น และเลือกตัวอย่างมีป้ายกำกับใหม่ที่ทำให้ประสิทธิภาพของตัวจำแนกกึ่งมีผู้สอนนั้นดีขึ้นหรือไม่แย่ลง ข้อแตกต่างคือในงานของ Guo ใช้เพื่อเลือกตัวอย่างมีป้ายกำกับใหม่ แต่ในงานนี้ใช้เพื่อให้ค่าน้ำหนักแก่ตัวอย่างมีป้ายกำกับตั้งต้น วัตถุประสงค์ในงานของ Guo เพื่อเลือกตัวอย่างมีป้ายกำกับใหม่ที่ดีที่สุด หากแต่ในงานนี้วัตถุประสงค์เพื่อวิเคราะห์คุณภาพตัวอย่างมีป้ายกำกับตั้งต้นที่อาจส่งผลเสียเมื่อนำไปใช้สร้างตัวจำแนกกึ่งมีผู้สอน โดยค่าความเชื่อมั่นของตัวอย่างมีป้ายกำกับแต่ละตัวใช้วิธีให้คะแนนดังต่อไปนี้

- ตัวอย่างมีป้ายกำกับแต่ละตัวจะได้รับค่าคะแนนเพิ่มขึ้นเมื่อมีอิทธิพลเชิงบวก (positive influence) ต่อตัวอย่างมีป้ายกำกับอื่น ๆ หรือตัวอย่างมีป้ายกำกับนั้นส่งผลให้การทำนายตัวอย่างมีป้ายกำกับตัวอื่น ๆ ถูกต้อง
- ตัวอย่างมีป้ายกำกับแต่ละตัวจะถูกลดค่าคะแนนมีค่าอิทธิพลเชิงลบ (negative influence) ต่อตัวอย่างมีป้ายกำกับอื่น ๆ หรือตัวอย่างมีป้ายกำกับนั้นส่งผลให้การทำนายคลาสตัวอย่างมีป้ายกำกับตัวอื่น ๆ ผิด

ค่าคะแนนของตัวอย่างมีป้ายกำกับแต่ละตัวอย่างจะถูกนำมาใช้ในการวิเคราะห์คุณภาพตัวอย่างมีป้ายกำกับ รวมทั้งใช้เป็นค่าน้ำหนักในการติดป้ายกำกับแก่ตัวอย่างไม่มีป้ายกำกับด้วย การทดลองเบื้องต้นบนชุดข้อมูล banknote ในพับที่ค่าความถูกต้องต่ำกว่าพับอื่นมากหรือในพับที่ 10 ค่าน้ำหนักของตัวอย่างมีป้ายกำกับแต่ละตัวอย่างที่ได้จากวิธีข้างต้นแสดงในตารางที่ 5.5 กำ-

หมายเลขตัวอย่างมีป้ายกำกับ	ค่าน้ำหนัก	คลาส
3	-0.5	banknote_zero
99	1	banknote_zero
251	-0.5	banknote_zero
444	-0.5	banknote_zero
570	-1	banknote_zero
678	-1	banknote_zero
686	1	banknote_one
782	0.67	banknote_one
934	0.5	banknote_one
1125	1	banknote_one
1127	0.67	banknote_one
1172	0.67	banknote_one

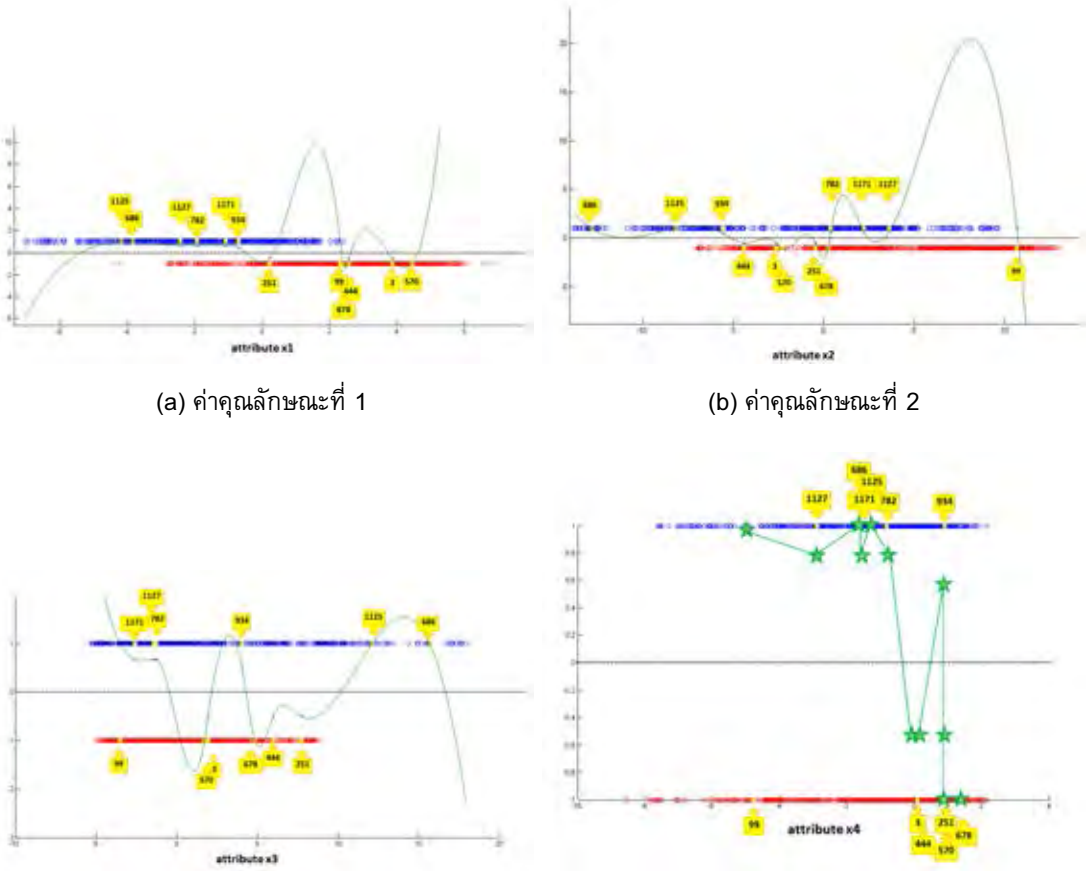
ตารางที่ 5.5: ค่าน้ำหนักตัวอย่างมีป้ายกำกับ กำหนดให้ค่าน้ำหนักสูงสุดของคลาส banknote_zero เท่ากับ -1 และคลาส banknote_one เท่ากับ 1

หนดให้ค่าน้ำหนักสูงสุดของคลาส banknote_zero เท่ากับ -1 และคลาส banknote_one เท่ากับ 1 โดยพบว่าตัวอย่างที่น่าจะทำให้การทำนายผิดพลาดคือตัวอย่างหมายเลข 99 ซึ่งเป็นตัวอย่างคลาส banknote_zero แต่มีค่าน้ำหนักเป็นคลาสตรงข้าม

5.4.3 การวิเคราะห์ค่าน้ำหนักตัวอย่างมีป้ายกำกับ

การวิเคราะห์ตัวอย่างมีป้ายกำกับโดยติดป้ายกำกับตัวอย่างไม่มีป้ายกำกับด้วยค่าคะแนนจากการประมาณค่าในช่วงในหนึ่งมิติ (1-dimension interpolation) ด้วยการประมาณค่าในช่วงแบบ spline (spline interpolation) โดยสร้างกราฟแสดงความสัมพันธ์ระหว่างค่าคุณลักษณะของตัวอย่างกับค่าคะแนนดังแสดงในภาพที่ 5.4 สำหรับค่าคุณลักษณะที่ 4 ไม่สามารถประมาณค่าในช่วงด้วย spline interpolation ได้จึงใช้วิธีประมาณค่าในช่วงแบบ linear interpolation แทน

อย่างไรก็ดีการวิเคราะห์คุณภาพตัวอย่างมีป้ายกำกับด้วยการพิจารณาทีละหนึ่งค่าคุณลักษณะนั้นทำได้ยาก การทดลองนี้จึงแปลงค่าคุณลักษณะ (Feature transformation) ด้วยเทคนิคการวิเคราะห์องค์ประกอบหลัก (Principle component analysis: PCA) เพื่อลดจำนวนมิติของค่าคุณลักษณะ โดยลดเหลือองค์ประกอบหลักสองมิติ จากนั้นจึงสร้างระนาบความมั่นใจ (confident plane) ของตัวอย่างมีป้ายกำกับบนองค์ประกอบหลักนั้น โดยมีสมมติฐานว่าตัวอย่างมีป้ายกำกับที่อยู่บริเวณที่ค่าความมั่นใจสูงสุดในแต่ละคลาสหรืออยู่บริเวณยอดของระนาบทั้งด้าน



(a) ค่าคุณลักษณะที่ 1

(b) ค่าคุณลักษณะที่ 2

(c) ค่าคุณลักษณะที่ 3

(d) ค่าคุณลักษณะที่ 4

ภาพที่ 5.4: กราฟแสดงการประมาณค่าน้ำหนักตัวอย่างบนค่าคุณลักษณะต่าง ๆ

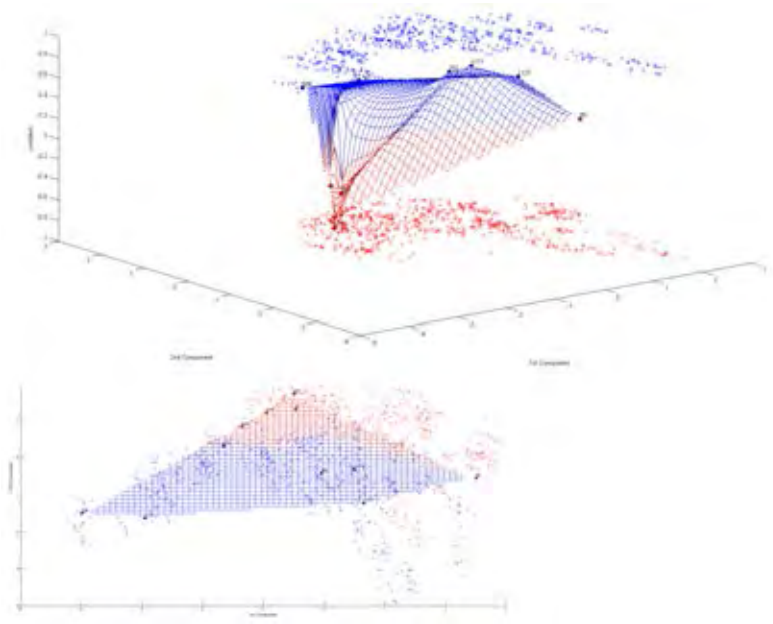
บนและด้านล่าง (ยอดเขาและท้องหุบเหว) นั้นเป็นตัวอย่างที่มีความน่าจะเป็นที่จะถูกติดป้ายกำกับได้อย่างถูกต้อง แต่ตัวอย่างที่ค่าความมั่นใจอยู่ระหว่างทั้งสองคลาส (บนพื้นราบ) อาจถูกติดป้ายกำกับผิดได้ ระนาบความมั่นใจที่บริเวณที่ค่าความมั่นใจระหว่างสองคลาสใกล้เคียงกันคือบริเวณที่ระนาบมีความชันต่ำใกล้จุดศูนย์ ตัวอย่างมีป้ายกำกับชุดที่ทำให้ได้ระนาบความมั่นใจที่มีบริเวณที่ระนาบมีความชันต่ำนั้นเป็นบริเวณกว้าง (แบนและกว้าง) มีความน่าจะเป็นที่จะส่งผลให้การเรียนรู้กึ่งมีผู้สอนนั้นมีประสิทธิภาพไม่ดี

อย่างไรก็ดีเมื่อเปรียบเทียบระนาบแสดงค่าความมั่นใจของแต่ละพับที่มีประสิทธิภาพการจำแนกแตกต่างกัน ดังภาพที่ 5.5 และ 5.6 พบว่าไม่สามารถหาความสัมพันธ์ของลักษณะระนาบกับประสิทธิภาพของตัวจำแนกได้ เมื่อเปรียบเทียบลักษณะของระนาบความมั่นใจในภาพที่ 5.7 พบว่าลักษณะของระนาบความมั่นใจที่ดีตามสมมติฐานที่ตั้งไว้ไม่สอดคล้องกับผลการจำแนกที่ได้ นอกจากนั้นเมื่อลองหาตำแหน่งของตัวอย่างที่ติดป้ายกำกับถูกและผิด ยังพบว่าตัวอย่างที่ติดป้ายกำกับผิดไม่ได้อยู่ใกล้กับตัวอย่างที่มีค่าน้ำหนักต่ำหรือตัวอย่างในวงกลมสีแดงในภาพที่ 5.8 และตัวอย่างที่ติดป้ายกำกับผิดเหล่านั้นไม่ได้อยู่ในระนาบบริเวณที่มีความมั่นใจต่ำ ๆ หรืออยู่ใกล้กับตัวอย่างมีป้ายกำกับที่มีค่าน้ำหนักต่ำที่สุดแต่กระจายอยู่ทั้งบนระนาบที่มีค่าความมั่นใจสูงมาก เช่น บริเวณใกล้ค่า -1 บริเวณความมั่นใจต่ำ เช่น ตัวอย่างในบริเวณใกล้ค่า 0 ดังภาพที่ 5.9

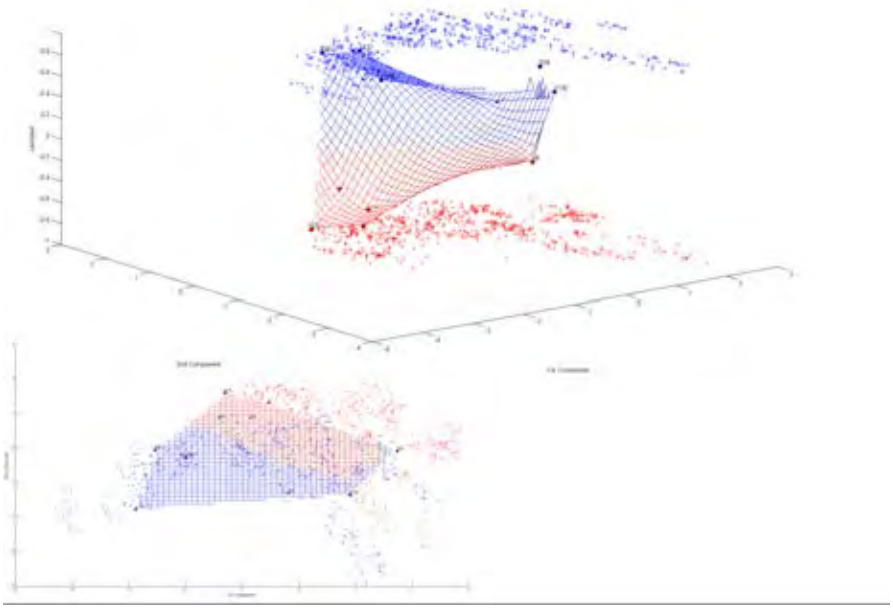
สรุปจากการทดลองนี้พบว่าลักษณะของระนาบไม่สามารถบอกความแตกต่างของคุณภาพตัวอย่างมีป้ายกำกับที่ทำให้ประสิทธิภาพในการจำแนกกึ่งสอนแตกต่างกันได้ ระนาบที่มีความชันต่ำเป็นบริเวณกว้างอาจจะทำให้ได้ตัวจำแนกที่มีประสิทธิภาพสูง และตำแหน่งตัวอย่างที่ถูกติดป้ายกำกับผิดไม่ได้สัมพันธ์กับตัวอย่างมีป้ายกำกับที่ค่าน้ำหนักอยู่ในช่วงที่ไม่มั่นใจ

5.5 การวิเคราะห์ชุดตัวอย่างมีป้ายกำกับด้วยการวิเคราะห์กลุ่มข้อมูล

จากการทดลองข้างต้นพบว่าวิธีวิเคราะห์ตัวอย่างมีป้ายกำกับโดยให้น้ำหนักทีละตัวอย่างไม่สามารถวิเคราะห์คุณภาพของตัวอย่างมีป้ายกำกับที่มีผลต่อตัวจำแนกกึ่งมีผู้สอนได้ เนื่องจากการวิเคราะห์โดยพิจารณาแต่ละตัวอย่างมีป้ายกำกับอาจเป็นการวิเคราะห์ที่หน่วยย่อยเกินไป จึงไม่เพียงพอที่บอกคุณภาพของตัวอย่างมีป้ายกำกับได้ นอกจากนี้ประสิทธิภาพของการติดป้ายกำกับผิดน่าจะได้รับอิทธิพลจากตัวอย่างมีป้ายกำกับทั้งชุดมากกว่าตัวอย่างใดตัวอย่างหนึ่ง การกระจายตัวของตัวอย่างมีป้ายกำกับกับการกระจายตัวของตัวอย่างไม่มีป้ายกำกับน่าจะเป็นประเด็นหนึ่งที่ส่งผลต่อความถูกต้องของการติดป้ายกำกับ ซึ่งการวิเคราะห์ด้วยวิธีแรกนั้นไม่ได้นำการกระจายตัวของตัวอย่างมาพิจารณาประกอบด้วย ทั้งนี้การกระจายตัวของคลาสเป็นสมมติฐานสำคัญประการหนึ่งที่อาจส่งผลต่อประสิทธิภาพของการเรียนรู้กึ่งมีผู้สอน ดังที่กล่าวไว้ในงานของ Tian (Tian et al., 2004) และในหัวข้อที่ 2.1.5

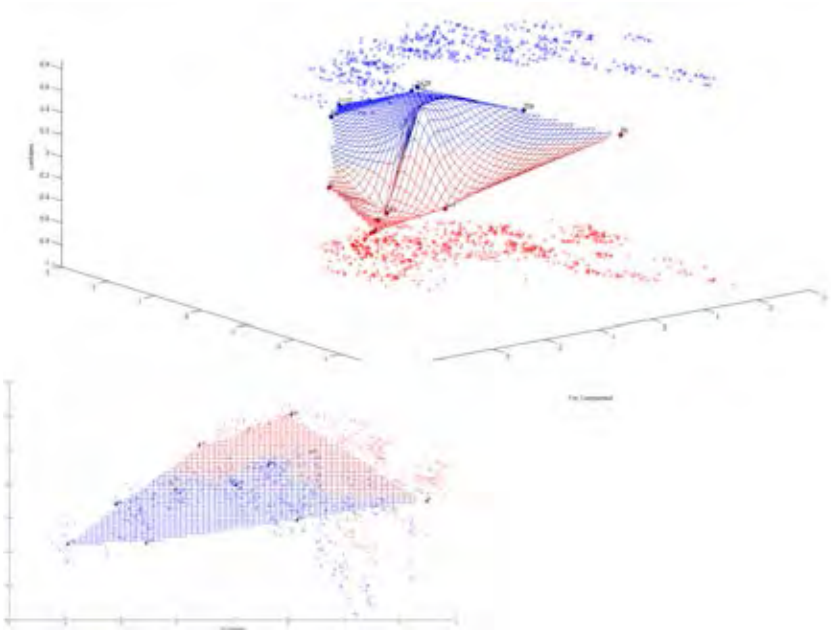


(a) ระนาบความมั่นใจของชุดตัวอย่างมีป้ายกำกับบนองค์ประกอบหลักสองมิติ ในพับที่ค่าความถูกต้องของตัวจำแนกต่ำผิดปกติ

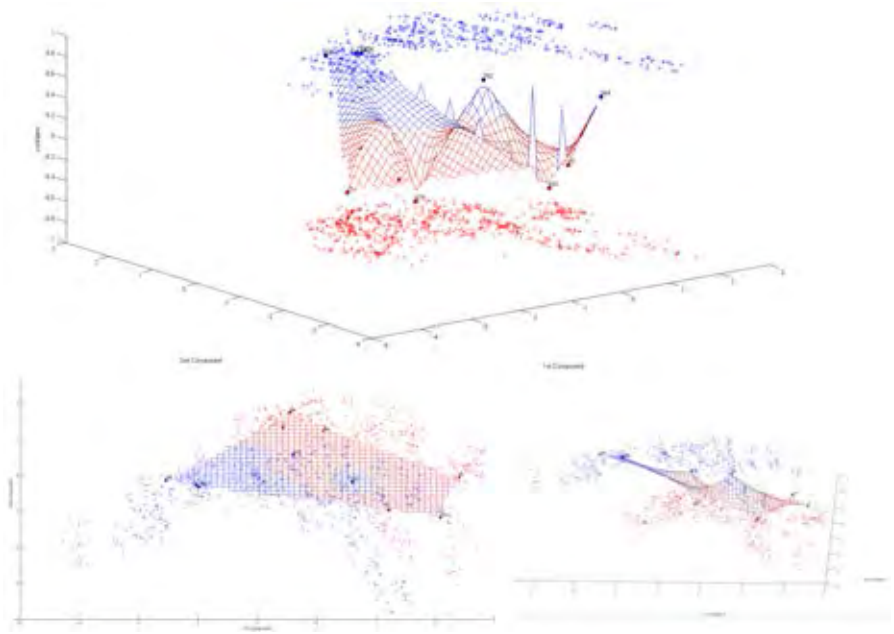


(b) ระนาบความมั่นใจของชุดตัวอย่างมีป้ายกำกับบนองค์ประกอบหลักสองมิติ ในพับที่ค่าความถูกต้องของตัวจำแนกสูงที่สุด

ภาพที่ 5.5: เปรียบเทียบระนาบความมั่นใจที่สร้างจากการให้ค่าน้ำหนักตัวอย่างมีป้ายกำกับบนพับที่ประสิทธิภาพการจำแนกแตกต่างกัน ตัวอย่างสีแดงและน้ำเงินแสดงถึงตัวอย่างในคลาส banknote_zero และ banknote_one ตามลำดับ

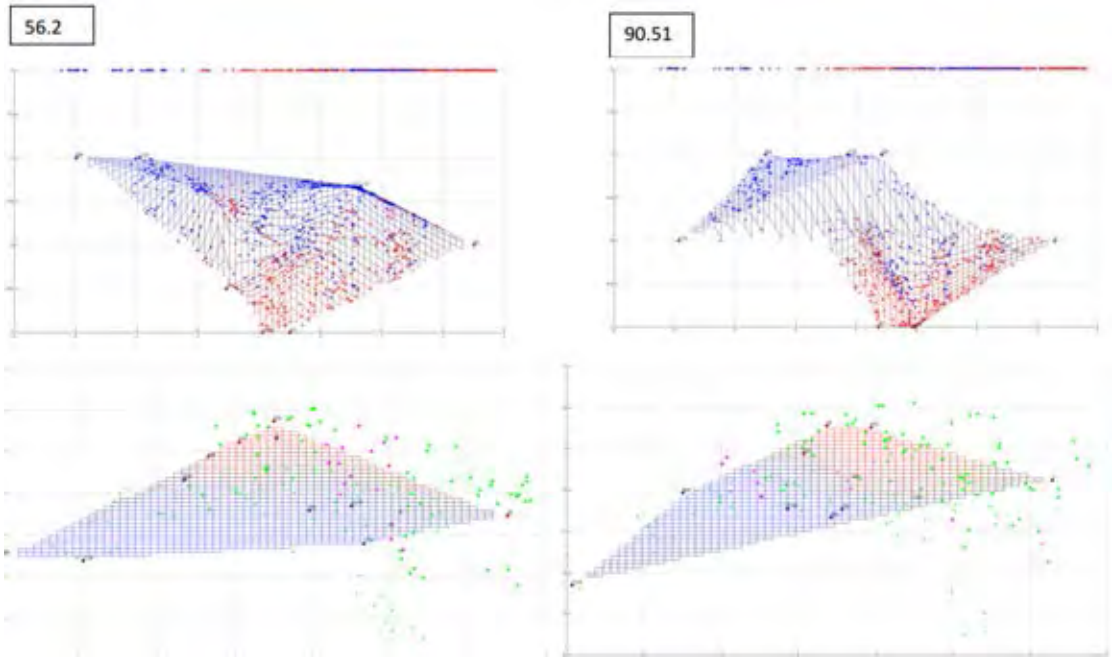


(a) ระนาบความมั่นใจของชุดตัวอย่างมีป้ายกำกับบนพับที่ความถูกต้องในการทำนายเท่ากับ 91.24 ลักษณะระนาบมีความชันต่ำเป็นบริเวณกว้าง

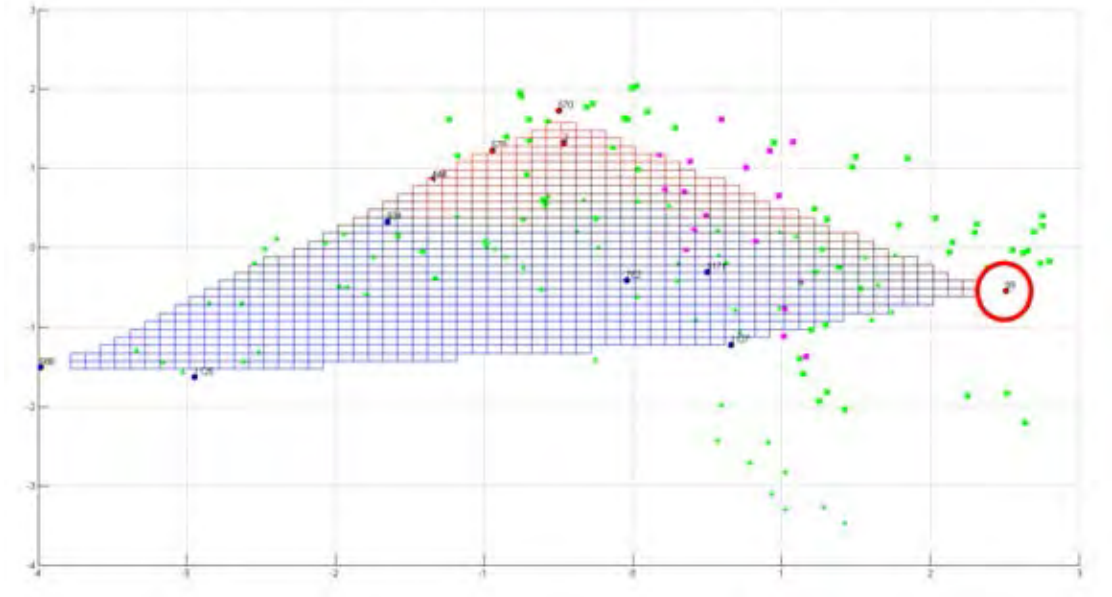


(b) ระนาบความมั่นใจของชุดตัวอย่างมีป้ายกำกับบนพับที่ความถูกต้องในการทำนายเท่ากับ 91.24 ลักษณะระนาบมีความชันสูงเป็นช่วง ๆ

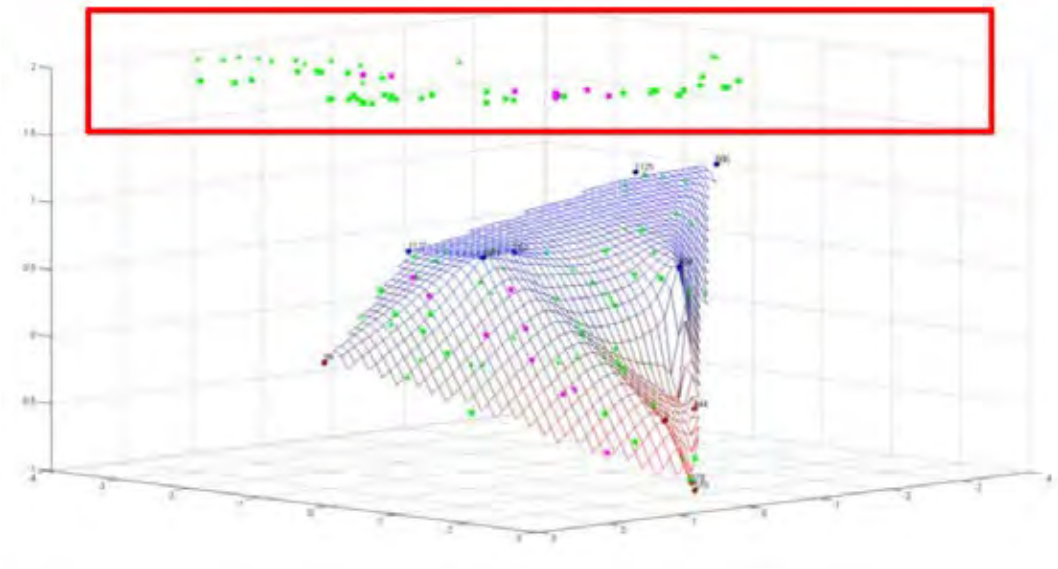
ภาพที่ 5.6: เปรียบเทียบระนาบความมั่นใจของชุดตัวอย่างมีป้ายกำกับบนพับที่ความถูกต้องในการทำนายสูงเท่ากัน แต่ลักษณะของระนาบแตกต่างกัน



ภาพที่ 5.7: เปรียบเทียบระนาบความมั่นใจบนชุดตัวอย่างมีป้ายกำกับบนพับที่ความถูกต้องในการทำนายต่ำ (ภาพบนและล่างซ้าย) และบนพับที่ความถูกต้องในการทำนายสูง (ภาพบบนและล่างขวา) ซึ่งมีลักษณะเป็นระนาบที่มีความชันต่ำเป็นบริเวณกว้างเช่นเดียวกัน



ภาพที่ 5.8: ตำแหน่งและความถูกต้องของการติดป้ายกำกับตัวอย่างเทียบกับระนาบการตัดสินใจ ตัวอย่างสีเขียวคือตัวอย่างที่ถูกติดป้ายกำกับถูกต้อง ตัวอย่างสีชมพูคือตัวอย่างที่ถูกติดป้ายกำกับผิด ตัวอย่างในวงกลมสีแดงคือตัวอย่างที่ถูกให้ค่าน้ำหนักต่ำที่สุด



ภาพที่ 5.9: ภาพแสดงความถูกต้องของการติดป้ายกำกับตัวอย่างไม่มีป้ายกำกับ โดยตัวอย่างในสี่เหลี่ยมสีแดงคือตัวอย่างไม่มีป้ายกำกับที่ไม่สามารถประมาณค่าน้ำหนักจากระนาบนี้ความมั่นใจได้ จากภาพนี้ตัวอย่างไม่มีป้ายกำกับที่ถูกติดป้ายกำกับผิดไม่ได้อยู่ในช่วงที่มีความมั่นใจต่ำ

งานวิจัยนี้จึงเสนอการวิเคราะห์ชุดตัวอย่างมีป้ายกำกับด้วยการวิเคราะห์กลุ่มข้อมูล โดยมีสมมติฐานว่า การจัดกลุ่มสามารถประมาณการกระจายตัวของตัวอย่างที่ใช้ในการเรียนรู้ได้ และการวิเคราะห์ลักษณะของกลุ่มข้อมูลที่ได้สามารถนำมาใช้เพื่อปรับปรุงชุดตัวอย่างมีป้ายกำกับ เพื่อปรับปรุงประสิทธิภาพของตัวจำแนกกึ่งมีผู้สอนได้ โดยการวิเคราะห์ข้อมูลนี้ประกอบไปด้วยสามขั้นตอน ได้แก่ การจัดกลุ่มแบบกึ่งมีผู้สอน การวิเคราะห์ประสิทธิภาพการทำนายของตัวจำแนกกึ่งมีผู้สอนด้วยลักษณะของกลุ่มข้อมูลของตัวอย่างที่ใช้ในการเรียนรู้ และการปรับปรุงชุดตัวอย่างมีป้ายกำกับจากผลการวิเคราะห์กลุ่มข้อมูล ซึ่งจะนำเสนอในหัวข้อย่อยต่อไปตามลำดับ

การทดลองใช้การทดสอบแบบไขว้ข้ามหลายพับและในแต่ละพับสุ่มตัวอย่างมีป้ายกำกับหลายครั้ง เพื่อวิเคราะห์อิทธิพลของตัวอย่างมีป้ายกำกับและหลีกเลี่ยงอิทธิพลของการกระจายตัวของตัวอย่างในแต่ละพับ ในชุดข้อมูลจริงใช้การทดสอบไขว้ข้ามสิบพับและสุ่มตัวอย่างมีป้ายกำกับสิบครั้งในแต่ละพับ ในชุดข้อมูลจากฐานข้อมูล UCI กำหนดจำนวนพับและจำนวนครั้งในการสุ่มตัวอย่างมีป้ายกำกับตามขนาดชุดข้อมูลดังแสดงในตารางที่ 5.6

5.5.1 การจัดกลุ่มแบบกึ่งมีผู้สอน

งานวิจัยนี้ประมาณการกระจายตัวของตัวอย่างที่ใช้ในการเรียนรู้ด้วยการจัดกลุ่มข้อมูลแบบกึ่งมีผู้สอน การจัดกลุ่มข้อมูลแบบกึ่งมีผู้สอนนี้พิจารณาทั้งค่าคุณลักษณะของตัวอย่าง และ

ชื่อชุดข้อมูล	จำนวนพับ	จำนวนครั้งในการสุ่ม
ชุดข้อมูลขนาดเล็ก		
banknote	10	10
bioNRB	10	10
splice	10	10
wilt	10	10
ชุดข้อมูลขนาดกลาง		
adult	10	1
bankmarket	10	1
eeg	10	1
magicgamma	10	1
mushroom	10	1
spam	10	1
ชุดข้อมูลขนาดใหญ่		
mnist 1vs7	5	1
mnist 2vs7	5	1
mnist 3vs8	5	1
mnist 4vs9	5	1
mnist 7vs9	5	1
mnist 8vs9	5	1

ตารางที่ 5.6: จำนวนพับและจำนวนครั้งในการสุ่มตัวอย่างมีป้ายกำกับสำหรับชุดข้อมูลจากฐานข้อมูล U-CI

คลาสของตัวอย่างในการจัดกลุ่มข้อมูล โดยงานวิจัยนี้เลือกใช้การจัดกลุ่มข้อมูลแบบกึ่งมีผู้สอนวิธีกำหนดตัวอย่างตั้งต้น (semi-supervised clustering by seeding) ร่วมกับการจัดกลุ่มด้วยขั้นตอนวิธีเค-มีนส์

ขั้นตอนวิธีจัดกลุ่มข้อมูลแบบกึ่งมีผู้สอนวิธีกำหนดตัวอย่างตั้งต้น เริ่มจากกำหนดศูนย์กลางของกลุ่มข้อมูล โดยในงานวิจัยนี้กำหนดให้ตัวอย่างมีป้ายกำกับแต่ละตัวเป็นตัวแทนศูนย์กลางของหนึ่งกลุ่ม ดังนั้นจำนวนกลุ่มตั้งต้นจึงเท่ากับจำนวนตัวอย่างมีป้ายกำกับทั้งหมด จากนั้นจัดกลุ่มข้อมูลด้วยขั้นตอนวิธีเคมีนส์โดยปกติ ขณะจัดกลุ่มอาจเกิดการเปลี่ยนกลุ่มของตัวอย่างมีป้ายกำกับหรือเกิดการรวมกันระหว่างกลุ่มได้ หากกลุ่มใดไม่เหลือสมาชิกแล้วกำหนดให้ยุบกลุ่มนั้นทิ้ง โดยจะหยุดกระบวนการจัดกลุ่มเมื่อตัวอย่างไม่มีการเปลี่ยนแปลงกลุ่มแล้ว ผลลัพธ์ที่ได้คือกลุ่มข้อมูลที่สัมพันธ์กับการกระจายตัวของข้อมูลและและสอดคล้องกับการกระจายตัวของตัวอย่างมีป้ายกำกับ แต่ละกลุ่มอาจประกอบไปด้วยตัวอย่างมีป้ายกำกับและตัวอย่างไม่มีป้ายกำกับ หรือประกอบไปด้วยตัวอย่างประเภทใดประเภทหนึ่ง งานวิจัยนี้แบ่งกลุ่มข้อมูลออกเป็นสองประเภท ได้แก่ กลุ่มที่มีตัวอย่างมีป้ายกำกับในกลุ่มนั้นหรือกลุ่มมีคลาส (labeled cluster) และกลุ่มที่ไม่มีตัวอย่างมีป้ายกำกับในกลุ่มนั้นหรือกลุ่มไม่ทราบคลาส (unknown cluster)

5.5.2 การวิเคราะห์ประสิทธิภาพการทำนายของตัวจำแนกกึ่งมีผู้สอนด้วยลักษณะของกลุ่มข้อมูลของตัวอย่างที่ใช้ในการเรียนรู้

งานวิจัยนี้มีสมมติฐานว่าตัวอย่างที่เป็นสมาชิกกลุ่มไม่ทราบคลาส จะถูกทำนายผิดมากกว่าตัวอย่างที่เป็นสมาชิกกลุ่มมีคลาส การทดลองแรกจึงทดสอบสมมติฐานนี้ โดยวัดประสิทธิภาพในการจำแนกตัวอย่างทดสอบในแต่ละกลุ่มข้อมูล โดยแบ่งตัวอย่างชุดทดสอบตามกลุ่มที่ได้จากการจัดกลุ่มกึ่งมีผู้สอนในขั้นตอนก่อนหน้านี้ ด้วยการวัดระยะห่างระหว่างตัวอย่างทดสอบกับศูนย์กลางของแต่ละกลุ่ม ตัวอย่างที่อยู่ใกล้กลุ่มใดที่สุดถือว่าเป็นสมาชิกในกลุ่มนั้น ผลการทดลองพบว่าค่าความถูกต้องในการจำแนกตัวอย่างทดสอบที่เป็นสมาชิกกลุ่มไม่ทราบคลาสต่ำกว่าตัวอย่างทดสอบในกลุ่มมีคลาสอย่างมีนัยสำคัญ โดยผลการทดลองบนชุดข้อมูล UCI แสดงในตารางที่ 5.7 และผลการทดลองค่าความถูกต้องเฉลี่ยบนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย แสดงในตารางที่ 5.8 ส่วนผลการทดลองโดยละเอียดบนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทยแสดงในตารางที่ ข.1 ในภาคผนวก ข และจำนวนของเครื่องหมายดอกจัน (*) แสดงถึงค่านัยสำคัญทางสถิติ เครื่องหมายจำนวน 3 ตัว 2 ตัว และ 1 ตัว แสดงถึงค่าความเชื่อมั่นร้อยละ 99 ร้อยละ 95 และร้อยละ 90 ตามลำดับ

การทดลองนี้สามารถบอกได้ว่าตัวจำแนกกึ่งมีผู้สอนที่สร้างขึ้นไม่สามารถทำนายตัวอย่างที่อยู่ในกลุ่มไม่ทราบคลาสได้อย่างมีประสิทธิภาพ จึงควรปรับปรุงชุดตัวอย่างมีป้ายกำกับ ให้

ครอบคลุมตัวอย่างกลุ่มไม่ทราบคลาสให้ได้มากขึ้น

	ความถูกต้องของ 1-nn บนตัวอย่างทดสอบที่อยู่ใน		ความถูกต้องของ 3-nn บนตัวอย่างทดสอบที่อยู่ใน	
ชุดข้อมูล	กลุ่มมีคลาส	กลุ่มไม่ทราบคลาส	กลุ่มมีคลาส	กลุ่มไม่ทราบคลาส
adult	**56.23**	55.52	**57.90**	56.39
bankmarket	**70.63**	68.66	**72.14**	70.8
banknote	**93.28*** **	78.91	**87.17*** **	71.92
bioNRB	**58.53** **	53.33	**57.28** **	52.22
eeg	**75.57*** **	65.31	**72.60*** **	64.15
magicgamma	**68.37**	68.07	**71.45**	69.69
mnist17	**98.41**	97.13	**98.17**	97.36
mnist27	**96.25***	94.16	**96.27** **	93.86
mnist38	**93.39*** **	88.55	**94.08** **	90.58
mnist49	**87.99***	82.29	**89.44**	85.76
mnist79	**92.99** **	87.93	**95.59***	92.21
mnist89	**95.66***	93.97	**95.97**	95.52
mushroom	**99.23*** **	65.55	**97.18*** **	64.53
spam	**68.06***	62.49	**66.86**	63.77
splice	55.86	**56.10**	**54.29**	53.35
wilt	66.53	**68.21**	60.91	**62.72**

ตารางที่ 5.7: เปรียบเทียบความถูกต้องของตัวจำแนกกึ่งมีผู้สอนเมื่อทำนายบนตัวอย่างทดสอบที่เป็นสมาชิกกลุ่มมีคลาสและกลุ่มไม่ทราบคลาสบนชุดข้อมูลจากฐานข้อมูล UCI

ตารางที่ 5.7 แสดงให้เห็นว่าความถูกต้องของการทำนายตัวอย่างที่เป็นสมาชิกกลุ่มไม่ทราบคลาสต่ำกว่าตัวอย่างที่เป็นสมาชิกกลุ่มมีคลาส เมื่อใช้ขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุด พบว่าชุดข้อมูลส่วนใหญ่ค่าความถูกต้องของการทำนายตัวอย่างที่เป็นสมาชิกกลุ่มไม่ทราบคลาสต่ำกว่าตัวอย่างที่เป็นสมาชิกกลุ่มมีคลาส โดยมี 10 ชุดข้อมูลที่แตกต่างอย่างมีนัยสำคัญทางสถิติ มีสองชุดข้อมูล ได้แก่ splice และ wilt ค่าความถูกต้องของการทำนายตัวอย่างที่เป็นสมาชิกกลุ่มไม่ทราบคลาสสูงกว่าตัวอย่างที่เป็นสมาชิกกลุ่มมีคลาส เมื่อใช้ขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดสามตัว พบว่าชุดข้อมูลส่วนใหญ่ค่าความถูกต้องของการทำนายตัวอย่างที่เป็นสมาชิกกลุ่มไม่ทราบคลาสต่ำกว่าตัวอย่างที่เป็นสมาชิกกลุ่มมีคลาส โดยมี 7 ชุดข้อมูลที่แตกต่างอย่างมีนัยสำคัญทางสถิติ และชุดข้อมูล wilt ค่าความถูกต้องของการทำนายตัวอย่างที่เป็นสมาชิกกลุ่มไม่ทราบคลาสสูงกว่าตัวอย่างที่เป็นสมาชิกกลุ่มมีคลาส

ชุดข้อมูล	ความถูกต้องของ 1-nn บนตัวอย่างทดสอบที่อยู่ใน		ความถูกต้องของ 3-nn บนตัวอย่างทดสอบที่อยู่ใน	
	กลุ่มมีคลาส	กลุ่มไม่ทราบคลาส	กลุ่มมีคลาส	กลุ่มไม่ทราบคลาส
ค่าเฉลี่ยบนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย	**91.59**	72.96	**92.37**	71.82

ตารางที่ 5.8: เปรียบเทียบความถูกต้องของตัวจำแนกกึ่งมีผู้สอนเมื่อทำนายบนตัวอย่างทดสอบที่เป็นสมาชิกกลุ่มมีคลาสและกลุ่มไม่ทราบคลาสบนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย

ตารางที่ 5.8 ทดลองบนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย แสดงให้เห็นว่าความถูกต้องเฉลี่ยของการทำนายตัวอย่างที่เป็นสมาชิกกลุ่มไม่ทราบคลาสโดยเฉลี่ยต่ำกว่าตัวอย่างที่เป็นสมาชิกกลุ่มมีคลาส ทั้งเมื่อใช้ขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดและเพื่อนบ้านใกล้ที่สุดสามตัว

5.5.3 การปรับปรุงชุดตัวอย่างมีป้ายกำกับจากผลการวิเคราะห์กลุ่มข้อมูล

การทดลองในหัวข้อก่อนหน้าแสดงให้เห็นว่า ตัวจำแนกกึ่งมีผู้สอนไม่สามารถจำแนกตัวอย่างในกลุ่มไม่ทราบคลาสได้อย่างมีประสิทธิภาพ เนื่องจากการสร้างตัวจำแนกตั้งต้นนั้นสร้างจากตัวอย่างมีป้ายกำกับซึ่งไม่ครอบคลุมตัวอย่างในกลุ่มไม่ทราบคลาส ตัวจำแนกตั้งต้นจึงอาจติดป้ายกำกับตัวอย่างในกลุ่มนี้ผิดและส่งผลต่อประสิทธิภาพของตัวจำแนกสุดท้ายที่สร้างจากตัวอย่างที่ติดป้ายกำกับใหม่ด้วย

สมมติฐานต่อมาของงานวิจัยนี้ คือหากสามารถเพิ่มจำนวนตัวอย่างมีป้ายกำกับให้แก่กลุ่มไม่ทราบคลาส จะส่งผลให้ตัวอย่างประสิทธิภาพในการจำแนกแบบกึ่งสอนดีขึ้น หัวข้อนี้จึงเสนอวิธีเพิ่มข้อมูลตัวอย่างมีป้ายกำกับเพื่อปรับปรุงประสิทธิภาพการจำแนกกึ่งมีผู้สอน โดยติดป้ายกำกับเพิ่มให้แก่กลุ่มไม่ทราบคลาสหนึ่งตัวอย่างต่อหนึ่งกลุ่มไม่ทราบคลาส ในหัวข้อที่ 5.5.3.1 แสดงผลการทดลองที่ให้ผู้ใช้ช่วยติดป้ายกำกับ และหัวข้อที่ 5.5.3.2 แสดงผลการทดลองที่ใช้ตัวจำแนกอื่น ๆ ช่วยติดป้ายกำกับ

5.5.3.1 การเพิ่มตัวอย่างมีป้ายกำกับโดยผู้ใช้ (active labeling)

การปรับปรุงตัวอย่างมีป้ายกำกับวิธีแรกอาศัยผู้ใช้ช่วยติดป้ายกำกับในกลุ่มไม่ทราบคลาส ซึ่งผลการทดลองพบว่าการเพิ่มตัวอย่างในกลุ่มไม่ทราบคลาสนี้ สามารถปรับปรุงประสิทธิภาพการจำ-แนกจากที่ใช้ตัวอย่างป้ายกำกับเดิมได้อย่างมีนัยสำคัญ ผลการทดลองบนชุดข้อมูลจาก

ชุดข้อมูล	ความถูกต้องของ 1-nn		ความถูกต้องของ 3-nn	
	ตัวอย่าง มีป้ายกำกับเดิม	เพิ่มตัวอย่างในกลุ่ม ไม่ทราบคลาส	ตัวอย่าง มีป้ายกำกับเดิม	เพิ่มตัวอย่างในกลุ่ม ไม่ทราบคลาส
adult	55.24	**59.49*** **	56.90	**61.86*** **
bankmarket	68.79	**74.32*** **	70.78	**77.28*** **
banknote	88.31	**93.73*** **	81.66	**89.70*** **
bioNRB	56.05	**60.41*** **	55.42	**58.67*** **
eeg	71.63	**78.08*** **	69.32	**74.48*** **
magicgamma	67.14	**69.57*** **	70.03	**73.12*** **
mnist17	98.08	**98.71*** **	97.96	**98.45*** **
mnist27	95.88	**96.86***	95.95	**97.00**
mnist38	92.24	**94.27*** **	93.19	**95.24*** **
mnist49	86.41	**89.95*** **	88.45	**92.17*** **
mnist79	91.21	**93.19*** **	94.47	**95.83** *
mnist89	95.36	**96.05** *	95.83	**96.59** *
mushroom	96.33	**98.90*** **	94.42	**95.32**
spam	64.07	**64.27**	64.42	**64.71**
splice	56.45	**57.31**	54.49	**55.19**
wilt	67.23	**77.70*** **	60.48	**72.43*** **

ตารางที่ 5.9: เปรียบเทียบความถูกต้องของตัวจำแนกกึ่งมีผู้สอนระหว่างการเรียนรู้ด้วยป้ายกำกับเดิมกับเมื่อผู้ใช้เพิ่มป้ายกำกับในกลุ่มไม่ทราบคลาสบนชุดข้อมูล UCI

ฐานข้อมูล UCI แสดงในตารางที่ 5.9 และผลการทดลองเฉลี่ยบนชุดข้อมูลการลดสิ่งรบกวนในเอกสารภาษาไทยแสดงในตารางที่ 5.10 ส่วนผลการทดลองโดยสมบูรณ์ของชุดข้อมูลการลดสิ่งรบกวนในเอกสารภาษาไทยด้วยเพื่อนบ้านใกล้ที่สุดหนึ่งและสามตัว แสดงในตารางที่ ข.2 ในภาคผนวก ข

ตารางที่ 5.9 แสดงผลการทดลองเปรียบเทียบความถูกต้องของตัวจำแนกกึ่งมีผู้สอนด้วยตัวอย่างมีป้ายกำกับเดิม กับตัวจำแนกที่สร้างจากตัวอย่างที่เพิ่มตัวอย่างในกลุ่มไม่ทราบคลาสบนชุดข้อมูล UCI ผลการทดลองแสดงให้เห็นว่าการปรับปรุงตัวอย่างมีป้ายกำกับทำให้ความถูกต้องของตัวจำแนกกึ่งมีผู้สอนสูงขึ้นในทุกชุดข้อมูล เมื่อใช้ขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุด พบว่ามี 14 ชุดข้อมูลที่ค่าความถูกต้องของการเพิ่มตัวอย่างสูงขึ้นอย่างมีนัยสำคัญทางสถิติ เมื่อใช้ขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดสามตัว พบว่ามี 12 ชุดข้อมูลที่ค่าความถูกต้องของการเพิ่มตัวอย่างสูงขึ้นอย่างมีนัยสำคัญทางสถิติ

ชุดข้อมูล	ความถูกต้องของ 1-nn		ความถูกต้องของ 3-nn	
	ตัวอย่าง มีป้ายกำกับเดิม	เพิ่มตัวอย่างในกลุ่ม ไม่ทราบคลาส	ตัวอย่าง มีป้ายกำกับเดิม	เพิ่มตัวอย่างในกลุ่ม ไม่ทราบคลาส
ค่าเฉลี่ยบนชุดข้อมูล การลดสิ่งรบกวนใน ภาพเอกสารภาษาไทย	89.27	**90.72**	89.95	**91.15**

ตารางที่ 5.10: เปรียบเทียบความถูกต้องของตัวจำแนกกึ่งมีผู้สอนระหว่างการเรียนรู้ด้วยป้ายกำกับเดิมกับเมื่อผู้ใช้เพิ่มป้ายกำกับในกลุ่มไม่ทราบคลาสบนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย

ตารางที่ 5.10 แสดงผลการทดลองเปรียบเทียบความถูกต้องของตัวจำแนกกึ่งมีผู้สอนด้วยตัวอย่างมีป้ายกำกับเดิม กับตัวจำแนกที่สร้างจากตัวอย่างที่เพิ่มตัวอย่างในกลุ่มไม่ทราบคลาสบนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย ผลการทดลองแสดงให้เห็นว่าการปรับปรุงตัวอย่างมีป้ายกำกับทำให้ความถูกต้องโดยเฉลี่ยของตัวจำแนกกึ่งมีผู้สอนสูงขึ้นเช่นเดียวกัน

ทั้งนี้เมื่อเปรียบเทียบการเลือกตัวอย่างมีป้ายกำกับระหว่างการเพิ่มตัวอย่างมีป้ายกำกับในกลุ่มมีคลาสกับการเลือกเพิ่มตัวอย่างในกลุ่มไม่ทราบคลาส พบว่าการเพิ่มตัวอย่างในกลุ่มไม่ทราบคลาสเพิ่มประสิทธิภาพได้ดีกว่ากลุ่มที่มีป้ายกำกับอยู่แล้วอย่างมีนัยสำคัญ ผลการทดลองบนทั้งสองชุดข้อมูลแสดงในตารางที่ 5.11 ตารางที่ 5.12 และตารางที่ ข.3 ในภาคผนวก ข

ตารางที่ 5.11 แสดงผลการทดลองเปรียบเทียบความถูกต้องของตัวจำแนกกึ่งมีผู้สอนเมื่อเพิ่มตัวอย่างมีป้ายกำกับในกลุ่มมีคลาสกับกลุ่มไม่ทราบคลาสบนชุดข้อมูลบนชุดข้อมูล U-CI ผลการทดลองแสดงให้เห็นว่าการเพิ่มตัวอย่างมีป้ายกำกับในกลุ่มไม่ทราบคลาสทำให้ได้ตัวจำแนกที่มีความถูกต้องสูงกว่าการเพิ่มตัวอย่างในกลุ่มมีคลาส เมื่อใช้ขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุด พบว่าค่าความถูกต้องของการเพิ่มตัวอย่างในกลุ่มไม่ทราบคลาสสูงกว่าการเพิ่มตัวอย่างในกลุ่มมีคลาสในชุดข้อมูลส่วนใหญ่ โดยมี 9 ชุดข้อมูลที่การเพิ่มตัวอย่างในกลุ่มไม่ทราบคลาสมีค่าความถูกต้องสูงกว่าอย่างมีนัยสำคัญทางสถิติ และมีสองชุดข้อมูล ได้แก่ spam และ wilt ที่การเพิ่มตัวอย่างในกลุ่มมีคลาสให้ค่าความถูกต้องของตัวจำแนกสูงกว่า เมื่อใช้ขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดสามตัว พบว่าค่าความถูกต้องของการเพิ่มตัวอย่างในกลุ่มไม่ทราบคลาสสูงกว่าการเพิ่มตัวอย่างในกลุ่มมีคลาสในชุดข้อมูลส่วนใหญ่ โดยมี 6 ชุดข้อมูลที่การเพิ่มตัวอย่างในกลุ่มไม่ทราบคลาสมีค่าความถูกต้องสูงกว่าอย่างมีนัยสำคัญทางสถิติ มีสามชุดข้อมูล ได้แก่ bioNRB spam และ wilt ที่การเพิ่มตัวอย่างในกลุ่มมีคลาสให้ค่าความถูกต้องของตัวจำแนกสูงกว่า

ตารางที่ 5.12 แสดงผลการทดลองเปรียบเทียบความถูกต้องเฉลี่ยของตัวจำแนกกึ่งมีผู้

ชุดข้อมูล	ความถูกต้องของ 1-nn เมื่อเพิ่มตัวอย่างมีป้ายกำกับในกลุ่ม		ความถูกต้องของ 3-nn เมื่อเพิ่มตัวอย่างมีป้ายกำกับในกลุ่ม	
	มีคลาส	ไม่ทราบคลาส	มีคลาส	ไม่ทราบคลาส
adult	58.93	**59.49**	61.63	**61.86**
bankmarket	73.68	**74.32**	77.07	**77.28**
banknote	89.4	**93.73*** **	85.43	**89.70*** **
bioNRB	58.64	**60.41** **	**58.88**	58.67
eeg	73.8	**78.08*** **	72.09	**74.48*** **
magicgamma	69.38	**69.57**	72.39	**73.12**
mnist17	98.17	**98.71*** **	98.06	**98.45** **
mnist27	96.01	**96.86***	96.07	**97.00**
mnist38	92.77	**94.27*** **	93.88	**95.24*** **
mnist49	87.55	**89.95***	89.73	**92.17***
mnist79	91.89	**93.19** **	94.83	**95.83***
mnist89	95.76	**96.05**	96.55	**96.59**
mushroom	96.36	**98.90*** **	94.74	**95.32**
spam	**65.60**	64.27	**67.45*** **	64.71
splice	56.74	**57.31**	54.52	**55.19**
wilt	**78.98** **	77.7	**76.65*** **	72.43

ตารางที่ 5.11: เปรียบเทียบความถูกต้องของตัวจำแนกกึ่งมีผู้สอนเมื่อเพิ่มตัวอย่างมีป้ายกำกับในกลุ่มมีคลาสกับกลุ่มไม่ทราบคลาสบนชุดข้อมูล UCI

ชุดข้อมูล	ความถูกต้องของ 1-nn เมื่อเพิ่มตัวอย่างมีป้ายกำกับในกลุ่ม		ความถูกต้องของ 3-nn เมื่อเพิ่มตัวอย่างมีป้ายกำกับในกลุ่ม	
	มีคลาส	ไม่ทราบคลาส	มีคลาส	ไม่ทราบคลาส
ค่าเฉลี่ยบนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย	89.81	**90.72**	90.43	**91.15**

ตารางที่ 5.12: เปรียบเทียบความถูกต้องเฉลี่ยของตัวจำแนกกึ่งมีผู้สอนเมื่อเพิ่มตัวอย่างมีป้ายกำกับในกลุ่มมีคลาสกับกลุ่มไม่ทราบคลาสบนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย

สอนเมื่อเพิ่มตัวอย่างมีป้ายกำกับในกลุ่มมีคลาสกับกลุ่มไม่ทราบคลาส บนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย ผลการทดลองแสดงให้เห็นว่าการเพิ่มตัวอย่างมีป้ายกำกับในกลุ่มไม่ทราบคลาสทำให้ได้ตัวจำแนกที่มีความถูกต้องโดยเฉลี่ยสูงกว่าการเพิ่มตัวอย่างในกลุ่มมีคลาส ทั้งเมื่อใช้ขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดและเพื่อนบ้านใกล้ที่สุดสามตัว

งานวิจัยนี้เสนอให้เพิ่มตัวอย่างมีป้ายกำกับโดยติดป้ายกำกับตัวอย่างที่บริเวณศูนย์กลางของกลุ่มไม่ทราบคลาส เหตุผลที่เลือกติดป้ายกำกับที่ศูนย์กลางของกลุ่มเนื่องจากบริเวณนั้นตัวอย่างมีความหนาแน่นสูงและมีระยะห่างจากทุกตัวอย่างน้อยที่สุดจึงน่าจะเป็นตัวแทนของกลุ่มได้ดีที่สุด การทดลองเปรียบเทียบระหว่างการเลือกเพิ่มตัวอย่างที่ศูนย์กลางของกลุ่มไม่ทราบคลาสกับการเลือกที่บริเวณสุ่มในกลุ่มไม่ทราบคลาสนั้น พบว่าการเลือกตัวอย่างที่ศูนย์กลางของกลุ่มสามารถปรับปรุงประสิทธิภาพของตัวจำแนกกึ่งมีผู้สอนได้ดีกว่าอย่างมีนัยสำคัญ โดยผลการทดลองบนทั้งสองชุดข้อมูลแสดงในตารางที่ 5.13 ตารางที่ 5.14 และตารางที่ ข.4 ในภาคผนวก ข

ตารางที่ 5.13 แสดงผลการทดลองเปรียบเทียบความถูกต้องของตัวจำแนกกึ่งมีผู้สอนเมื่อเพิ่มป้ายกำกับในบริเวณศูนย์กลางกับบริเวณอื่น ๆ ของกลุ่มไม่ทราบคลาสบนชุดข้อมูล U-CI ผลการทดลองแสดงให้เห็นว่าการเพิ่มตัวอย่างมีป้ายกำกับที่ศูนย์กลางกลุ่มไม่ทราบคลาสทำให้ได้ตัวจำแนกที่มีความถูกต้องสูงกว่าการเพิ่มตัวอย่างที่ตำแหน่งอื่น ๆ ในกลุ่ม เมื่อใช้ขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุด พบว่าทุกชุดข้อมูลค่าความถูกต้องของการเพิ่มตัวอย่างที่ศูนย์กลางกลุ่มสูงกว่าการเพิ่มตัวอย่างที่ตำแหน่งอื่น ๆ ในกลุ่ม โดยมี 11 ชุดข้อมูลที่แตกต่างอย่างมีนัยสำคัญทางสถิติ เมื่อใช้ขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดสามตัว พบว่าทุกชุดข้อมูลค่าความถูกต้องของการเพิ่มตัวอย่างที่ศูนย์กลางกลุ่มสูงกว่าการเพิ่มตัวอย่างที่ตำแหน่งอื่น ๆ ในกลุ่ม โดยมี 8 ชุดข้อมูลที่แตกต่างอย่างมีนัยสำคัญทางสถิติ และมีสามชุดข้อมูล ได้แก่ bioNRB mnist89 และ mushroom ที่ค่าความถูกต้องของการเพิ่มตัวอย่างที่ตำแหน่งอื่น ๆ สูงกว่าการเพิ่มตัวอย่างที่ศูนย์กลางกลุ่ม

ตารางที่ 5.14 แสดงผลการทดลองเปรียบเทียบความถูกต้องเฉลี่ยของตัวจำแนกกึ่งมีผู้สอนเมื่อเพิ่มป้ายกำกับในบริเวณศูนย์กลางกับบริเวณอื่น ๆ ของกลุ่มไม่ทราบคลาส บนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย ผลการทดลองแสดงให้เห็นว่าการเพิ่มตัวอย่างมีป้ายกำกับที่ศูนย์กลางกลุ่มไม่ทราบคลาสทำให้ได้ตัวจำแนกที่มีความถูกต้องโดยเฉลี่ยสูงกว่าการเพิ่มตัวอย่างที่ตำแหน่งอื่น ๆ ทั้งเมื่อใช้ขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดและเพื่อนบ้านใกล้ที่สุดสามตัว

ชุดข้อมูล	ความถูกต้องของ 1-nn เมื่อเพิ่มตัวอย่างที่		ความถูกต้องของ 3-nn เมื่อเพิ่มตัวอย่างที่	
	สุ่มตำแหน่ง	ศูนย์กลางกลุ่ม	สุ่มตำแหน่ง	ศูนย์กลางกลุ่ม
adult	59	**59.49**	61.47	**61.86**
bankmarket	73.34	**74.32*** **	76.27	**77.28*** **
banknote	93.23	**93.73** **	89.01	**89.70***
bioNRB	59.18	**60.41** **	**59.81**	58.67
eeg	75.66	**78.08*** **	73.06	**74.48** **
magicgamma	68.82	**69.57** **	72.2	**73.12***
mnist17	98.52	**98.71***	98.26	**98.45*** **
mnist27	96.57	**96.86***	96.68	**97.00**
mnist38	92.99	**94.27*** **	94.25	**95.24** **
mnist49	88.29	**89.95** **	90.81	**92.17** **
mnist79	92.44	**93.19** **	95.54	**95.83**
mnist89	95.88	**96.05**	**96.63**	96.59
mushroom	98.82	**98.90**	**95.42**	95.32
spam	64.05	**64.27**	64.23	**64.71**
splice	56.17	**57.31**	54.84	**55.19**
wilt	75.65	**77.70*** **	70.73	**72.43*** **

ตารางที่ 5.13: เปรียบเทียบความถูกต้องของตัวจำแนกกึ่งมีผู้สอนเมื่อเพิ่มป้ายกำกับในบริเวณศูนย์กลางกับบริเวณอื่น ๆ ของกลุ่มไม่ทราบคลาสบนชุดข้อมูล UCI

ชุดข้อมูล	ความถูกต้องของ 1-nn เมื่อเพิ่มตัวอย่างที่		ความถูกต้องของ 3-nn เมื่อเพิ่มตัวอย่างที่	
	สุ่มตำแหน่ง	ศูนย์กลางกลุ่ม	สุ่มตำแหน่ง	ศูนย์กลางกลุ่ม
ค่าเฉลี่ยบนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย	90.42	**90.72**	91.00	**91.15**

ตารางที่ 5.14: เปรียบเทียบความถูกต้องเฉลี่ยของตัวจำแนกกึ่งมีผู้สอนเมื่อเพิ่มป้ายกำกับในบริเวณศูนย์กลางกับบริเวณอื่น ๆ ของกลุ่มไม่ทราบคลาส บนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย

5.5.3.2 การเพิ่มตัวอย่างมีป้ายกำกับโดยตัวจำแนกอื่น ๆ (co-labeling)

อย่างไรก็ดีการให้ผู้ใช้ช่วยติดป้ายกำกับตัวอย่างที่เป็นตัวแทนของกลุ่มไม่ทราบคลาสทำให้ระบบไม่เป็นอัตโนมัติ งานวิจัยนี้จึงทดลองใช้ตัวจำแนกที่สร้างจากขั้นตอนวิธีอื่น ๆ เพื่อช่วยติดป้ายกำกับตัวอย่างตัวแทนของกลุ่มไม่ทราบคลาส ผลการทดลองเปรียบเทียบตัวจำแนกที่สร้างจากขั้นตอนวิธีต่าง ๆ 6 วิธี ได้แก่ เพื่อนบ้านใกล้ที่สุดหนึ่งตัว เพื่อนบ้านใกล้ที่สุดสามตัว นิวรอลเน็ตเวิร์ก ซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ และป่าไม้แบบสุ่ม พบว่าการเพิ่มป้ายกำกับให้แก่กลุ่มไม่ทราบคลาสด้วยป่าไม้แบบสุ่มสามารถปรับปรุงประสิทธิภาพของตัวจำแนกกึ่งมีผู้สอนได้ดีที่สุด

ผลการทดลองบนชุดข้อมูล UCI ด้วยขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดแสดงในตารางที่ 5.15 แสดงให้เห็นว่าการเพิ่มป้ายกำกับให้แก่กลุ่มไม่ทราบคลาสด้วยป่าไม้แบบสุ่มสามารถปรับปรุงประสิทธิภาพของตัวจำแนกกึ่งมีผู้สอนให้ดีขึ้นอย่างมีนัยสำคัญที่ระดับความเชื่อมั่นร้อยละ 99 ถึง 7 ชุดข้อมูล และแย่กว่าการติดป้ายกำกับด้วยวิธีเดิม 1 ชุดข้อมูล โดยต้นไม้ตัดสินใจเป็นตัวจำแนกที่ดีรองลงมา ตัวจำแนกอื่น ๆ ได้แก่ ซัพพอร์ตเวกเตอร์แมชชีน เพื่อนบ้านใกล้ที่สุดสามและหนึ่งตัวอย่าง และนิวรอลเน็ตเวิร์ก อาจไม่เหมาะสมในการติดป้ายกำกับนัก เนื่องจากการติดป้ายกำกับด้วยตัวจำแนกดังกล่าวให้ผลแย่ลงในหลายชุดข้อมูล

ผลการทดลองบนชุดข้อมูล UCI ด้วยขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดสามตัวแสดงในตารางที่ 5.16 แสดงให้เห็นว่าการเพิ่มป้ายกำกับให้แก่กลุ่มไม่ทราบคลาสด้วยป่าไม้แบบสุ่มสามารถปรับปรุงประสิทธิภาพของตัวจำแนกกึ่งมีผู้สอนให้ดีขึ้นอย่างมีนัยสำคัญถึง 8 ชุดข้อมูล และแย่กว่าการติดป้ายกำกับด้วยวิธีเดิม 1 ชุดข้อมูล โดยเพื่อนบ้านใกล้ที่สุดหนึ่งตัวเป็นตัวจำแนกที่ดีรองลงมา

ผลการทดลองเฉลี่ยบนชุดข้อมูลการลดสิ่งรบกวนในเอกสารภาษาไทย ด้วยขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดแสดงในตารางที่ 5.17 แสดงให้เห็นว่าการเพิ่มป้ายกำกับให้แก่กลุ่มไม่ทราบคลาสด้วยป่าไม้แบบสุ่มสามารถปรับปรุงประสิทธิภาพของตัวจำแนกกึ่งมีผู้สอนให้ดีขึ้นอย่างมีนัยสำคัญถึง 26 ชุดข้อมูล และแย่กว่าการติดป้ายกำกับด้วยวิธีเดิม 7 ชุดข้อมูล โดยเพื่อนบ้านใกล้ที่สุดสามตัวเป็นตัวจำแนกที่ดีรองลงมา

ผลการทดลองเฉลี่ยบนชุดข้อมูลการลดสิ่งรบกวนในเอกสารภาษาไทย ด้วยขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดสามตัวแสดงในตารางที่ 5.18 แสดงให้เห็นว่าการเพิ่มป้ายกำกับให้แก่กลุ่มไม่ทราบคลาสด้วยป่าไม้แบบสุ่มสามารถปรับปรุงประสิทธิภาพของตัวจำแนกกึ่งมีผู้สอนให้ดีขึ้นอย่างมีนัยสำคัญถึง 27 ชุดข้อมูล และแย่กว่าการติดป้ายกำกับด้วยวิธีเดิม 7 ชุดข้อมูล โดยต้นไม้ตัดสินใจเป็นตัวจำแนกที่ดีรองลงมา

ชุดข้อมูล	ความถูกต้องของตัวจำแนก 1-nn กึ่งมีผู้สอนที่สร้างจากตัวอย่างมีป้ายกำกับ							
	ชุดตั้งต้น	ชุดปรับปรุงด้วยผู้ใช้	ชุดปรับปรุงด้วยตัวจำแนก					
			1-nn	3-nn	Neural Network	SVM	Tree	Random Forest
adult	55.24	**59.49+++**	**55.30**	55.16	**56.85++**	**59.39+++**	**57.73+++**	**57.19+++**
bankmarket	68.79	**74.32+++**	**69.14**	**69.43++**	**70.92+**	58.01–	**70.59+++**	**72.11+++**
banknote	88.31	**93.73+++**	88.29	86.47–	86.76-	86.99–	**88.57**	88.23
bioNRB	56.05	**60.41+++**	55.67	54.83–	55.13	**57.72+**	**56.60**	**57.65++**
eeg	71.63	**78.08+++**	70.35–	70.35–	69.15–	71.05	68.99–	71.07
magicgamma	67.14	**69.57+++**	66.44	**67.49**	**68.29**	**69.01+++**	**68.89+++**	**69.65+++**
mnist17	98.08	**98.71+++**	**98.24**	**98.12**	**98.30**	85.16–	97.88	**98.55+++**
mnist27	95.88	**96.86+**	95.07–	94.84–	95.74	82.76–	95.64	95.81
mnist38	92.24	**94.27+++**	91.77	91.58	89.88	83.97–	**92.66**	**92.56**
mnist49	86.41	**89.95+++**	**87.03**	85.95	**87.41**	76.03–	**88.43++**	**88.99+++**
mnist79	91.21	**93.19+++**	**91.79**	**92.02**	90.37	77.63–	90.19	90.87
mnist89	95.36	**96.05++**	95.64	**95.85+**	95.82	81.03–	93.46	95.06
mushroom	96.33	**98.90+++**	**96.53**	**96.97**	**96.40**	96.03	**96.55**	**96.65**
spam	64.07	**64.27**	63.9	**64.47**	**64.25**	61.77-	**64.18**	**64.66**
splice	56.45	**57.31**	53.64-	53.77–	54.08	53.33–	**57.09**	53.29–
wilt	67.23	**77.70+++**	67.12	65.57–	65.16–	**69.59+++**	**73.14+++**	**71.89+++**
ดีกว่าใช้ป้ายกำกับเดิมจำนวน		**14**	0	1	2	4	5	**7**
แย่กว่าใช้ป้ายกำกับเดิมจำนวน		0	3	6	3	10	1	1

ตารางที่ 5.15: เปรียบเทียบความถูกต้องของตัวจำแนกกึ่งมีผู้สอนด้วยเพื่อนบ้านใกล้ที่สุด (1-nn) เมื่อเพิ่มป้ายกำกับในกลุ่มไม่ทราบคลาสด้วยตัวจำแนกที่สร้างจากขั้นตอนวิธีต่าง ๆ บนชุดข้อมูล UCI

ชุดข้อมูล	ความถูกต้องของตัวจำแนก 3-nn กึ่งมีผู้สอนที่สร้างจากตัวอย่างมีป้ายกำกับ							
	ชุดตั้งต้น	ชุดปรับปรุงด้วยผู้ใช้	ชุดปรับปรุงด้วยตัวจำแนก					
			1-nn	3-nn	Neural Network	SVM	Tree	Random Forest
adult	56.90	**61.86+++**	**59.89+++**	**62.48+++**	56.58	56.88	**59.13++**	**59.66+++**
bankmarket	70.78	**77.28+++**	**72.86+++**	62.42–	70.56	**71.08**	**73.78+++**	**74.16+++**
banknote	81.66	**89.70+++**	**86.11+++**	**84.12+++**	**84.80+++**	**82.26**	**84.68+++**	**85.62+++**
bioNRB	55.42	**58.67+++**	**56.01**	**57.60++**	54.66	54.30–	54.52	**56.61**
eeg	69.32	**74.48+++**	67.93–	**69.57**	68.77	68.63	68.22-	**69.46**
magicgamma	70.03	**73.12+++**	**72.12+++**	**72.05++**	69.28-	**70.38**	**71.44**	**72.84+++**
mnist17	97.96	**98.45+++**	**98.07**	93.45–	**98.29**	**98.17**	**98.34+++**	**98.36++**
mnist27	95.95	**97.00**	95.75	95.5	**96.44**	87.03–	**96.42**	**96.46**
mnist38	93.19	**95.24+++**	**94.17**	90.00-	**93.37**	**93.47**	92.19	**93.67**
mnist49	88.45	**92.17+++**	**91.20+++**	79.64–	**88.96**	87.69	**90.81+**	**91.34+++**
mnist79	94.47	**95.83++**	94.45	84.32–	**94.73**	**94.94**	94.18	**94.57**
mnist89	95.83	**96.59++**	**96.24**	**96.28**	**96.41+**	87.69–	**96.20**	**96.41**
mushroom	94.42	**95.32**	**94.57**	**94.87**	94.39	**94.90**	**94.43**	**94.70++**
spam	64.42	**64.71**	**65.05**	62.98	63.83	63.58	**64.47**	**65.51**
splice	54.49	**55.19**	**56.17**	52.41	51.84	52.28–	53.32	50.79-
wilt	60.48	**72.43+++**	**68.41+++**	**64.97+++**	**63.14+++**	**61.26++**	**61.18**	**67.26+++**
ดีกว่าใช้ป้ายกำกับเดิมจำนวน		**12**	6	5	3	1	5	**8**
แย่กว่าใช้ป้ายกำกับเดิมจำนวน		0	1	5	1	4	1	1

ตารางที่ 5.16: เปรียบเทียบความถูกต้องของตัวจำแนกกึ่งมีผู้สอนด้วยเพื่อนบ้านใกล้ที่สุดสามตัว (3-nn) เมื่อเพิ่มป้ายกำกับในกลุ่มไม่ทราบคลาสด้วยตัวจำแนกที่สร้างจากขั้นตอนวิธีต่าง ๆ บนชุดข้อมูล UCI

ชุดข้อมูล	ความถูกต้องของตัวจำแนก 1-nn กึ่งมีผู้สอนที่สร้างจากตัวอย่างมีป้ายกำกับ							
	ชุดตั้งต้น	ชุดปรับปรุงด้วยผู้ใช้	ชุดปรับปรุงด้วยตัวจำแนก					
			1-nn	3-nn	Neural Network	SVM	Tree	Random Forest
ค่าเฉลี่ยบนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย	89.27	**90.72**	89.14	89.25	88.28	88.21	88.92	**89.43**
ดีกว่าใช้ป้ายกำกับเดิมจำนวน		**62**	3	15	1	4	8	**26**
แย่กว่าใช้ป้ายกำกับเดิมจำนวน		2	19	14	41	49	26	7

ตารางที่ 5.17: เปรียบเทียบความถูกต้องเฉลี่ยของตัวจำแนกกึ่งมีผู้สอนด้วยเพื่อนบ้านใกล้ที่สุด (1-nn) เมื่อเพิ่มป้ายกำกับในกลุ่มไม่ทราบคลาสด้วยตัวจำแนกที่สร้างจากขั้นตอนวิธีต่าง ๆ บนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย

ชุดข้อมูล	ความถูกต้องของตัวจำแนก 3-nn กึ่งมีผู้สอนที่สร้างจากตัวอย่างมีป้ายกำกับ							
	ชุดตั้งต้น	ชุดปรับปรุงด้วยผู้ใช้	ชุดปรับปรุงด้วยตัวจำแนก					
			1-nn	3-nn	Neural Network	SVM	Tree	Random Forest
ค่าเฉลี่ยบนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย	89.95	91.15	89.82	89.80	89.82	89.92	90.11	**90.22**
ดีกว่าใช้ป้ายกำกับเดิมจำนวน		**59**	14	6	11	14	18	**27**
แย่กว่าใช้ป้ายกำกับเดิมจำนวน		0	25	13	19	23	14	9

ตารางที่ 5.18: เปรียบเทียบความถูกต้องเฉลี่ยของตัวจำแนกกึ่งมีผู้สอนด้วยเพื่อนบ้านใกล้ที่สุดสามตัว (3-nn) เมื่อเพิ่มป้ายกำกับในกลุ่มไม่ทราบคลาสด้วยตัวจำแนกที่สร้างจากขั้นตอนวิธีต่าง ๆ บนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย

ชุดข้อมูล	ตัวจำแนกมีผู้สอน บนตัวอย่างมีป้ายกำกับเดิม	ตัวจำแนกกึ่งมีผู้สอน	
		บนตัวอย่าง มีป้ายกำกับเดิม	เพิ่มตัวอย่างกลุ่มไม่ทราบคลาส ด้วยป่าไม้แบบสุ่ม
adult	55.24	55.24	**57.19*** **
bankmarket	68.78	**68.79**	**72.11*** **
bioNRB	56.71	56.05 - - -	**57.65**
magicgamma	67.41	67.14	**69.65*** **
mnist49	85.67	**86.41**	**88.99** **
mushroom	96.43	96.33	**96.65**
spam	64.53	64.07 - -	**64.66**
wilt	67.36	67.23	**71.89*** **

ตารางที่ 5.19: เปรียบเทียบความถูกต้องของตัวจำแนกกึ่งมีผู้สอนเพื่อนบ้านใกล้ที่สุด (1-nn) ระหว่างการเรียนรู้แบบมีผู้สอน การเรียนรู้แบบกึ่งมีผู้สอนด้วยชุดป้ายกำกับเดิม และการเรียนรู้แบบกึ่งมีผู้สอนที่เพิ่มตัวอย่างในกลุ่มไม่ทราบคลาสด้วยป่าไม้แบบสุ่ม บนชุดข้อมูล UCI

ผลการทดลองเปรียบเทียบความถูกต้องของตัวจำแนกกึ่งมีผู้สอนเมื่อเพิ่มป้ายกำกับในกลุ่มไม่ทราบคลาสด้วยตัวจำแนกที่สร้างจากขั้นตอนวิธีต่าง ๆ บนชุดข้อมูลการลดสิ่งรบกวนในเอกสารภาษาไทยโดยสมบูรณ์ ด้วยขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดแสดงในตารางที่ ข.5 และเพื่อนบ้านใกล้ที่สุดสามตัวแสดงในตารางที่ ข.6 ในภาคผนวกที่ ข

นอกจากนี้การเพิ่มตัวอย่างด้วยป่าไม้แบบสุ่มยังช่วยปรับปรุงความถูกต้องของตัวจำแนกกึ่งมีผู้สอนจากเดิมที่ไม่แตกต่างหรือแย่กว่าการเรียนรู้แบบมีผู้สอนโดยไม่ใช้ตัวอย่างไม่มีป้ายกำกับ ให้ดีขึ้นอย่างมีนัยสำคัญในหลายชุดข้อมูล โดยผลการทดลองแสดงในตารางที่ 5.19 ถึงตารางที่ 5.22

ตารางที่ 5.19 แสดงผลการทดลองเปรียบเทียบความถูกต้องของตัวจำแนกกึ่งมีผู้สอนเพื่อนบ้านใกล้ที่สุด ระหว่างการเรียนรู้แบบมีผู้สอน การเรียนรู้แบบกึ่งมีผู้สอนด้วยชุดป้ายกำกับเดิม และการเรียนรู้แบบกึ่งมีผู้สอนที่เพิ่มตัวอย่างในกลุ่มไม่ทราบคลาสด้วยป่าไม้แบบสุ่ม บนชุดข้อมูล UCI ผลการทดลองแสดงให้เห็นว่าในชุดข้อมูล bioNRB และ spam การเรียนรู้แบบกึ่งมีผู้สอนด้วยตัวอย่างมีป้ายกำกับเดิมให้ความถูกต้องต่ำกว่าการเรียนรู้แบบมีผู้สอนด้วยตัวอย่างมีป้ายกำกับเท่านั้นที่ระดับความเชื่อมั่นร้อยละ 99 และร้อยละ 95 ตามลำดับ เมื่อปรับปรุงตัวอย่างมีป้ายกำกับทำให้ค่าความถูกต้องดีขึ้นกว่าการเรียนรู้แบบมีผู้สอน ในชุดข้อมูล adult bankmarket magicgamma mnist49 และ wilt การเรียนรู้แบบกึ่งมีผู้สอนด้วยตัวอย่างมีป้ายกำกับเดิมให้ความถูกต้องไม่แตกต่างกับการเรียนรู้แบบมีผู้สอนด้วยตัวอย่างมีป้ายกำกับเท่านั้น

ชุดข้อมูล	ตัวจำแนกมีผู้สอน บนตัวอย่างมีป้ายกำกับเดิม	ตัวจำแนกกึ่งมีผู้สอน	
		บนตัวอย่าง มีป้ายกำกับเดิม	เพิ่มตัวอย่างกลุ่มไม่ทราบคลาส ด้วยป่าไม้แบบสุ่ม
adult	56.87	**56.90**	**59.66*** **
bankmarket	70.81	70.78	**74.16*** **
bioNRB	55.62	55.42	**56.61**
magicgamma	69.70	**70.03***	**72.84*** **
mnist49	86.26	**88.45*** **	**91.34*** **
mushroom	94.26	**94.42**	**94.70*** **
spam	64.47	64.42	**65.51**
wilt	60.37	**60.48**	**67.26*** **

ตารางที่ 5.20: เปรียบเทียบความถูกต้องของตัวจำแนกกึ่งมีผู้สอนเพื่อนบ้านใกล้ที่สุดสามตัว (3-nn) ระหว่างการเรียนรู้แบบมีผู้สอน การเรียนรู้แบบกึ่งมีผู้สอนด้วยชุดป้ายกำกับเดิม และการเรียนรู้แบบกึ่งมีผู้สอนที่เพิ่มตัวอย่างในกลุ่มไม่ทราบคลาสด้วยป่าไม้แบบสุ่ม บนชุดข้อมูล UCI

อย่างมีนัยสำคัญ เมื่อปรับปรุงตัวอย่างมีป้ายกำกับทำให้ความถูกต้องของตัวจำแนกกึ่งมีผู้สอนสูงกว่าการเรียนรู้แบบมีผู้สอนอย่างมีนัยสำคัญที่ระดับความเชื่อมั่นร้อยละ 95 สำหรับชุดข้อมูล mnist49 และที่ระดับความเชื่อมั่นร้อยละ 99 สำหรับชุดข้อมูล adult bankmarket magicgamma และ wilt และชุดข้อมูล mushroom แม้ผลลัพธ์จะไม่แตกต่างอย่างมีนัยสำคัญทางสถิติแต่ตัวจำแนกกึ่งมีผู้สอนที่ได้จากการปรับปรุงชุดตัวอย่างมีป้ายกำกับก็ให้ค่าความถูกต้องสูงขึ้นกว่าการใช้ตัวอย่างมีป้ายกำกับเดิม

ตารางที่ 5.20 แสดงผลการทดลองเปรียบเทียบความถูกต้องของตัวจำแนกกึ่งมีผู้สอนเพื่อนบ้านใกล้ที่สุดสามตัว ระหว่างการเรียนรู้แบบมีผู้สอน การเรียนรู้แบบกึ่งมีผู้สอนด้วยชุดป้ายกำกับเดิม และการเรียนรู้แบบกึ่งมีผู้สอนที่เพิ่มตัวอย่างในกลุ่มไม่ทราบคลาสด้วยป่าไม้แบบสุ่ม บนชุดข้อมูล UCI ผลการทดลองแสดงให้เห็นว่าในชุดข้อมูล adult bankmarket mushroom และ wilt การเรียนรู้แบบกึ่งมีผู้สอนด้วยตัวอย่างมีป้ายกำกับเดิมให้ความถูกต้องไม่แตกต่างกับการเรียนรู้แบบมีผู้สอนด้วยตัวอย่างมีป้ายกำกับเท่านั้นอย่างมีนัยสำคัญ เมื่อปรับปรุงตัวอย่างมีป้ายกำกับทำให้ความถูกต้องของตัวจำแนกกึ่งมีผู้สอนสูงกว่าการเรียนรู้แบบมีผู้สอนอย่างมีนัยสำคัญที่ระดับความเชื่อมั่นร้อยละ 99 ในชุดข้อมูล magicgamma การเพิ่มตัวอย่างมีป้ายกำกับด้วยป่าไม้แบบสุ่มทำให้ความถูกต้องเพิ่มขึ้น และระดับความเชื่อมั่นเพิ่มขึ้นจากร้อยละ 90 เป็นร้อยละ 99 และในชุดข้อมูล bioNRB และ spam แม้ผลลัพธ์จะไม่แตกต่างอย่างมีนัยสำคัญทางสถิติแต่ตัวจำแนกกึ่งมีผู้สอนที่ได้จากการปรับปรุงชุดตัวอย่างมีป้ายกำกับก็ให้ค่าความถูกต้องสูงขึ้นกว่าการใช้ตัวอย่างมีป้ายกำกับเดิม

ชุดข้อมูล	ตัวจำแนกมีผู้สอน บนตัวอย่างมีป้ายกำกับเดิม	ตัวจำแนกกึ่งมีผู้สอน	
		บนตัวอย่าง มีป้ายกำกับเดิม	เพิ่มตัวอย่างกลุ่มไม่ทราบคลาส ด้วยป่าไม้แบบสุ่ม
noise020	77.44	77.32	**78.57*** **
noise040	93.31	93.32	**93.52** **
noise077	93.76	93.80	**94.21*** **
noise120	96.67	96.64	**96.84**
noise122	95.31	95.3	**95.79*** **
noise125	94.67	94.65	**94.85*** **
noise134	92.01	92.08	**92.74*** **
noise135	91.46	91.48	**91.99** **
noise160	97.03	97.01	**97.73*** **
noise161	90.11	90.11	**90.93*** **
noise162	84.36	84.46	**85.21***
noise163	96.98	96.97	**97.57** **
noise164	94.99	94.88 - -	**95.66** **
noise170	82.44	82.38	**82.59**
noise177	83.59	83.54	**85.02*** **
noise184	84.10	84.03	**84.18**
noise188	84.61	84.51	**85.81*** **

ตารางที่ 5.21: เปรียบเทียบความถูกต้องเฉลี่ยของตัวจำแนกกึ่งมีผู้สอนเพื่อนบ้านใกล้ที่สุด (1-nn) ระหว่างการเรียนรู้แบบมีผู้สอน การเรียนรู้แบบกึ่งมีผู้สอนด้วยชุดป้ายกำกับเดิม และการเรียนรู้แบบกึ่งมีผู้สอนที่เพิ่มตัวอย่างในกลุ่มไม่ทราบคลาสด้วยป่าไม้แบบสุ่ม บนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย

ตารางที่ 5.21 แสดงผลการทดลองเปรียบเทียบความถูกต้องของตัวจำแนกกึ่งมีผู้สอนเพื่อนบ้านใกล้ที่สุด ระหว่างการเรียนรู้แบบมีผู้สอน การเรียนรู้แบบกึ่งมีผู้สอนด้วยชุดป้ายกำกับเดิม และการเรียนรู้แบบกึ่งมีผู้สอนที่เพิ่มตัวอย่างในกลุ่มไม่ทราบคลาสด้วยป่าไม้แบบสุ่ม บนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย ผลการทดลองแสดงให้เห็นว่าในชุดข้อมูล noise164 การเรียนรู้แบบกึ่งมีผู้สอนด้วยตัวอย่างมีป้ายกำกับเดิมให้ความถูกต้องต่ำกว่าวิธีเรียนรู้แบบมีผู้สอนอย่างมีนัยสำคัญ เมื่อปรับปรุงตัวอย่างมีป้ายกำกับทำให้ค่าความถูกต้องดีขึ้นกว่าการเรียนรู้แบบมีผู้สอนอย่างมีนัยสำคัญที่ระดับความเชื่อมั่นร้อยละ 95 การปรับปรุงตัวอย่างมีป้ายกำกับในกลุ่มไม่ทราบคลาสทำให้ความถูกต้องของตัวจำแนกกึ่งมีผู้สอนจากเดิมไม่แตกต่างจากการเรียนรู้แบบมีผู้สอนเปลี่ยนเป็นสูงกว่าการเรียนรู้แบบมีผู้สอนอย่างมีนัยสำคัญที่ระดับความเชื่อมั่นร้อยละ 99 ใน 9 ชุดข้อมูล ที่ระดับความเชื่อมั่นร้อยละ 95 ใน 3 ชุดข้อมูล ที่ระดับความเชื่อมั่นร้อยละ 90 ในชุดข้อมูล noise162 สำหรับชุดข้อมูล noise120 noise170 และ noise184 แม้ผลลัพธ์จะไม่แตกต่างอย่างมีนัยสำคัญทางสถิติแต่ตัวจำแนกกึ่งมีผู้สอนที่ได้จากการปรับปรุงชุดตัวอย่างมีป้ายกำกับก็ให้ค่าความถูกต้องสูงขึ้นกว่าการใช้ตัวอย่างมีป้ายกำกับเดิม

ตารางที่ 5.22 แสดงผลการทดลองเปรียบเทียบความถูกต้องของตัวจำแนกกึ่งมีผู้สอนเพื่อนบ้านใกล้ที่สุด ระหว่างการเรียนรู้แบบมีผู้สอน การเรียนรู้แบบกึ่งมีผู้สอนด้วยชุดป้ายกำกับเดิม และการเรียนรู้แบบกึ่งมีผู้สอนที่เพิ่มตัวอย่างในกลุ่มไม่ทราบคลาสด้วยป่าไม้แบบสุ่ม บนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย ผลการทดลองแสดงให้เห็นว่าในชุดข้อมูล noise125 การเรียนรู้แบบกึ่งมีผู้สอนด้วยตัวอย่างมีป้ายกำกับเดิมให้ความถูกต้องต่ำกว่าการเรียนรู้แบบมีผู้สอนอย่างมีนัยสำคัญ เมื่อปรับปรุงตัวอย่างมีป้ายกำกับทำให้ค่าความถูกต้องดีขึ้นกว่าการเรียนรู้แบบมีผู้สอน การปรับปรุงตัวอย่างมีป้ายกำกับในกลุ่มไม่ทราบคลาสทำให้ความถูกต้องของตัวจำแนกกึ่งมีผู้สอนจากเดิมไม่แตกต่างจากการเรียนรู้แบบมีผู้สอนเปลี่ยนเป็นสูงกว่าการเรียนรู้แบบมีผู้สอนอย่างมีนัยสำคัญที่ระดับความเชื่อมั่นร้อยละ 99 ใน 11 ชุดข้อมูล ที่ระดับความเชื่อมั่นร้อยละ 95 ใน 3 ชุดข้อมูล ที่ระดับความเชื่อมั่นร้อยละ 90 ในชุดข้อมูล noise161 สำหรับชุดข้อมูล noise020 แม้ผลลัพธ์จะไม่แตกต่างอย่างมีนัยสำคัญทางสถิติแต่ตัวจำแนกกึ่งมีผู้สอนที่ได้จากการปรับปรุงชุดตัวอย่างมีป้ายกำกับก็ให้ค่าความถูกต้องสูงขึ้นกว่าการใช้ตัวอย่างมีป้ายกำกับเดิม

ชุดข้อมูล	ตัวจำแนกมีผู้สอน บนตัวอย่างมีป้ายกำกับเดิม	ตัวจำแนกกึ่งมีผู้สอน	
		บนตัวอย่าง มีป้ายกำกับเดิม	เพิ่มตัวอย่างกลุ่มไม่ทราบคลาส ด้วยป่าไม้แบบสุ่ม
noise020	79.61	79.52	**79.66**
noise040	93.72	93.76	**94.07*** **
noise077	95.00	95.04	**95.18** **
noise120	95.81	95.78	**96.88*** **
noise122	95.79	95.80	**96.27*** **
noise125	94.80	94.76 - -	**94.80**
noise134	93.15	93.19	**93.42** **
noise135	91.77	91.83	**92.31*** **
noise160	96.90	96.95	**97.51*** **
noise161	91.48	91.53	**92.08***
noise162	86.08	86.02	**87.56*** **
noise163	96.98	96.92	**97.77*** **
noise164	94.76	94.80	**95.54** **
noise170	81.06	81.16	**82.90*** **
noise177	90.24	90.28	**91.10*** **
noise184	80.97	81.03	**82.40*** **
noise188	80.25	80.45	**83.03*** **

ตารางที่ 5.22: เปรียบเทียบความถูกต้องเฉลี่ยของตัวจำแนกกึ่งมีผู้สอนเพื่อนบ้านใกล้ที่สุดสามตัว (3-nn) ระหว่างการเรียนรู้แบบมีผู้สอน การเรียนรู้แบบกึ่งมีผู้สอนด้วยชุดป้ายกำกับเดิม และการเรียนรู้แบบกึ่งมีผู้สอนที่เพิ่มตัวอย่างในกลุ่มไม่ทราบคลาสด้วยป่าไม้แบบสุ่ม บนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย

บทที่ 6

สรุปผลการทดลอง

การจำแนกแบบกึ่งมีผู้สอนเป็นวิธีที่ได้รับการยอมรับว่ามีประสิทธิภาพเมื่อนำไปใช้แก้ปัญหาในหลายแขนงปัญหา อย่างไรก็ดีการใช้ตัวอย่างไม่มีป้ายกำกับในการสร้างตัวจำแนกนั้นอาจส่งผลเสียต่อประสิทธิภาพของตัวจำแนกได้ โดยเฉพาะในวิธีการเรียนรู้แบบกึ่งมีผู้สอนที่ต้องติดป้ายกำกับตัวอย่างไม่มีป้ายกำกับก่อนการนำไปใช้ เช่น วิธีจัดกลุ่มและติดป้าย และวิธีเรียนรู้ด้วยตนเอง งานวิจัยนี้จึงนำเสนอวิธีปรับปรุงการติดป้ายกำกับตัวอย่างโดยมุ่งเน้นปรับปรุงการติดป้ายกำกับในวิธีการเรียนรู้แบบกึ่งมีผู้สอนทั้งสองวิธีนี้ จุดประสงค์เพื่อปรับปรุงความถูกต้องของตัวจำแนกกึ่งมีผู้สอนที่สร้างจากตัวอย่างที่ติดป้ายกำกับใหม่นั้น

ขั้นตอนวิธีติดป้ายกำกับในวิธีจัดกลุ่มและติดป้ายเริ่มจากการจัดกลุ่มตัวอย่าง จากนั้นจึงติดป้ายกำกับตัวอย่างไม่มีป้ายกำกับในแต่ละกลุ่มในขั้นตอนต่อมา การติดป้ายกำกับตัวอย่างในแต่ละกลุ่มพิจารณาจากคลาสของตัวอย่างมีป้ายกำกับในกลุ่มนั้น ในกรณีที่สมาชิกในกลุ่มประกอบไปด้วยตัวอย่างมีป้ายกำกับมากกว่าหนึ่งคลาสหรือเป็นกลุ่มคลาสปะปน วิธีจัดกลุ่มและติดป้ายโดยทั่วไปใช้วิธีโหวตคลาสส่วนใหญ่เพื่อเลือกคลาสตัวแทนสำหรับติดป้ายกำกับตัวอย่างไม่มีป้ายกำกับในกลุ่มนั้น อย่างไรก็ดีวิธีโหวตคลาสส่วนใหญ่นี้ไม่พิจารณาตัวอย่างในคลาสส่วนน้อยจึงส่งผลให้หลายตัวอย่างถูกติดป้ายกำกับไม่ถูกต้อง งานวิจัยนี้จึงเสนอวิธีติดป้ายกำกับตัวอย่างในกลุ่มคลาสปะปน ด้วยวิธีแบ่งกลุ่มย่อยตามคุณลักษณะที่ถูกเลือก วิธีที่นำเสนอแบ่งตัวอย่างในกลุ่มคลาสปะปนเป็นกลุ่มย่อยด้วยค่าคุณลักษณะใหม่ที่ถูกเลือกเพื่อให้สามารถจำแนกตัวอย่างมีป้ายกำกับต่างคลาสได้ จากนั้นจึงติดป้ายกำกับตัวอย่างในแต่ละกลุ่มย่อย ผลการทดลองแสดงให้เห็นว่าวิธีที่นำเสนอมีค่าความถูกต้องของการติดป้ายกำกับและความถูกต้องของตัวจำแนกสูงกว่าวิธีโหวตคลาสส่วนใหญ่ และเมื่อเปรียบเทียบวิธีการลดสิ่งรบกวนด้วยวิธีที่นำเสนอกับวิธีลดสิ่งรบกวนที่ใกล้เคียงกัน ได้แก่ วิธีเอสพีเอ็นแบบสองเฟสและซอฟต์แวร์ลดสิ่งรบกวนในเอกสาร ScanFix Xpress 6.0 พบว่าวิธีที่นำเสนอสามารถจำแนกสิ่งรบกวนและตัวอักษรออกจากกันได้ดีกว่าวิธีการที่เปรียบเทียบ

วิธีเรียนรู้ด้วยตนเองเป็นวิธีเรียนรู้แบบกึ่งมีผู้สอนอีกวิธีหนึ่งที่ได้รับความนิยมมาก กระบวนการติดป้ายกำกับในวิธีเรียนรู้ด้วยตนเองใช้ตัวจำแนกตั้งต้นที่สร้างจากตัวอย่างมีป้ายกำกับเท่านั้นเพื่อติดป้ายกำกับตัวอย่างไม่มีป้ายกำกับที่เหลือ แต่เนื่องจากตัวอย่างมีป้ายกำกับนั้นมีจำนวนไม่มาก จึงอาจไม่เพียงพอที่จะสร้างตัวจำแนกตั้งต้นที่มีประสิทธิภาพสำหรับการนำไปใช้ติดป้ายกำกับตัวอย่างไม่มีป้ายกำกับที่เหลือ งานวิจัยนี้จึงเสนอการวิเคราะห์ตัวอย่างที่ใช้ในการเรียนรู้ด้วยการจัดกลุ่มข้อมูลด้วยการจัดกลุ่มข้อมูลแบบกึ่งมีผู้สอน ซึ่งเป็นการประมวลผลก่อน

เพื่อวิเคราะห์การกระจายตัวของตัวอย่างมีป้ายกำกับ เท่าที่ผู้วิจัยสืบค้นมาไม่พบงานวิจัยอื่นที่เสนอแนวทางในการวิเคราะห์ลักษณะของชุดตัวอย่างมีป้ายกำกับก่อนการนำไปใช้เรียนรู้กึ่งมีผู้สอน งานวิจัยเพื่อปรับปรุงวิธีเรียนรู้ด้วยตนเองมุ่งเน้นศึกษาวิธีแก้ไขป้ายกำกับของตัวอย่างที่ถูกติดป้ายกำกับผิด หรือเลือกใช้ตัวอย่างที่ติดป้ายกำกับใหม่โดยประเมินความมั่นใจด้วยเทคนิคต่าง ๆ อาจกล่าวได้ว่างานวิจัยนี้เป็นงานวิจัยแรกที่เสนอว่าการเรียนรู้กึ่งมีผู้สอนควรมีกระบวนการก่อนเพื่อวิเคราะห์ตัวอย่างมีป้ายกำกับที่จะนำมาใช้ในการเรียนรู้กึ่งมีผู้สอน โดยผลการวิเคราะห์กลุ่มข้อมูลพบว่าตัวอย่างที่เป็นสมาชิกกลุ่มที่ไม่มีตัวอย่างมีป้ายกำกับจะถูกทำนายผิดมากกว่าตัวอย่างในกลุ่มมีป้ายกำกับอย่างมีนัยสำคัญ งานวิจัยนี้จึงเสนอให้ปรับปรุงชุดตัวอย่างมีป้ายกำกับด้วยการเพิ่มตัวอย่างมีป้ายกำกับในกลุ่มไม่ทราบคลาส จากการทดลองพบกว่าการเพิ่มตัวอย่างมีป้ายกำกับในกลุ่มไม่ทราบคลาสส่งผลให้ประสิทธิภาพตัวจำแนกนั้นดีขึ้นกว่าการเพิ่มตัวอย่างในกลุ่มมีคลาสอย่างมีนัยสำคัญ โดยการเพิ่มตัวอย่างมีป้ายกำกับเลือกเพิ่มหนึ่งตัวอย่างที่บริเวณศูนย์กลางของกลุ่มซึ่งส่งผลให้ประสิทธิภาพตัวจำแนกนั้นดีขึ้นกว่าการเพิ่มตัวอย่างในบริเวณอื่น ๆ ของกลุ่มข้อมูลอย่างมีนัยสำคัญ

แต่เนื่องจากการเพิ่มข้อมูลตัวอย่างมีป้ายกำกับจำเป็นต้องใช้แรงงานมนุษย์เพื่อพิจารณาป้ายกำกับ จึงส่งผลให้ระบบไม่เป็นอัตโนมัติ ในงานวิจัยนี้จึงศึกษาการใช้ตัวจำแนกต่าง ๆ เพื่อติดป้ายกำกับในกลุ่มไม่ทราบคลาส ได้แก่ เพื่อนบ้านใกล้ที่สุดหนึ่งตัว เพื่อนบ้านใกล้ที่สุดสามตัว นิวรอลเน็ตเวิร์ก ซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ และป่าไม้แบบสุ่ม ผลการทดลองแสดงให้เห็นว่าตัวจำแนกป่าไม้แบบสุ่มสามารถติดป้ายกำกับตัวอย่างที่เป็นตัวแทนของกลุ่มที่ไม่ทราบคลาสได้ถูกต้องมากที่สุด และเมื่อนำตัวอย่างที่ติดป้ายกำกับด้วยวิธีที่นำเสนอไปสร้างตัวจำแนก พบว่าชุดตัวอย่างที่ปรับปรุงป้ายกำกับคลาสด้วยวิธีที่นำเสนอ สามารถสร้างตัวจำแนกกึ่งมีผู้สอนที่มีความถูกต้องในการจำแนกสูงกว่าตัวจำแนกที่สร้างจากตัวอย่างมีป้ายกำกับคลาสเดิมอย่างมีนัยสำคัญ นอกจากนี้การเพิ่มตัวอย่างด้วยป่าไม้แบบสุ่มยังช่วยปรับปรุงความถูกต้องของตัวจำแนกกึ่งมีผู้สอน จากเดิมซึ่งข้อมูลตัวอย่างมีป้ายกำกับเดิมทำให้ได้ตัวจำแนกกึ่งมีผู้สอนที่มีค่าความถูกต้องไม่แตกต่างหรือต่ำกว่าการเรียนรู้แบบมีผู้สอน เมื่อปรับปรุงตัวอย่างด้วยป่าไม้แบบสุ่มทำให้ได้ตัวจำแนกกึ่งมีผู้สอนที่มีค่าความถูกต้องสูงกว่าการเรียนรู้แบบมีผู้สอนอย่างมีนัยสำคัญในหลายชุดข้อมูล โดยการทดลองบนชุดข้อมูลจาก UCI ซึ่งเป็นชุดข้อมูลมาตรฐานในการเรียนรู้ของเครื่อง ซึ่งประกอบไปด้วยชุดข้อมูลจากหลากหลายแขนงปัญหาโดยเลือกใช้ทั้งสิ้นจำนวน 15 ชุดข้อมูล และบนชุดข้อมูลจริงจากปัญหาการจำแนกสิ่งรบกวนออกจากตัวอักษรในภาพเอกสารภาษาไทยจำนวน 67 ภาพเอกสาร

6.1 แนวทางการพัฒนางานวิจัยต่อในอนาคต

งานวิจัยนี้พิจารณาความครอบคลุมของตัวอย่างมีป้ายกำกับด้วยการจัดกลุ่มข้อมูล โดยพิจารณาการปรากฏของตัวอย่างมีป้ายกำกับในแต่ละกลุ่ม แบ่งเป็นกลุ่มมีคลาสและกลุ่มไม่ทราบคลาส ทั้งนี้นอกจากการพิจารณาความครอบคลุมของตัวอย่างมีป้ายกำกับแล้ว เราอาจใช้กลุ่มข้อมูลที่ได้มาพิจารณาความซ้ำซ้อนของตัวอย่างมีป้ายกำกับ เพื่อปรับสมดุลของตัวอย่างมีป้ายกำกับสำหรับตัวอย่างแต่ละกลุ่ม เพื่อให้ได้ตัวอย่างมีป้ายกำกับที่เหมาะสมที่จะสร้างตัวจำแนกที่ติดป้ายกำกับผิดพลาดน้อยที่สุด

นอกจากประเภทของกลุ่มข้อมูลตามการปรากฏของตัวอย่างมีป้ายกำกับแล้ว ข้อมูลอื่นที่ได้จากการจัดกลุ่มข้อมูลน่าจะสามารถนำมาประมาณคุณภาพตัวจำแนกในเบื้องต้นได้ งานวิจัยนี้ได้ทดลองสกัดค่าคุณลักษณะอื่น ๆ ที่น่าสนใจดังแสดงอยู่ในตารางที่ 6.1

โดยการทดลองเบื้องต้นเพื่อพิจารณาความสัมพันธ์ของคุณลักษณะของกลุ่มข้อมูลกับประสิทธิภาพการจำแนกตัวอย่างในกลุ่มนั้นด้วยค่าคุณลักษณะข้างต้น การทดลองนี้แบ่งกลุ่มข้อมูลตามความถูกต้องในการจำแนกตัวอย่างในกลุ่มนั้น โดยแบ่งเป็น กลุ่มมีประสิทธิภาพหรือกลุ่มที่ค่าความถูกต้องสูงกว่าร้อยละ 80 และกลุ่มต้องปรับปรุงหรือกลุ่มที่ค่าความถูกต้องต่ำกว่าร้อยละ 80

เมื่อวิเคราะห์ด้วยค่าทางสถิติเพื่อหาความสัมพันธ์ของค่าคุณลักษณะกับประเภทของกลุ่มตัวอย่าง พบว่าค่าคุณลักษณะส่วนใหญ่มีอิทธิพลต่อความถูกต้องของการทำนายตัวอย่าง ยกเว้นค่า DBI และค่า L2C2 ที่ไม่มีผลอย่างมีนัยสำคัญทางสถิติ และเมื่อเรียงลำดับคุณลักษณะตามค่าอินฟอร์เมชันเกนพบว่าค่าคุณลักษณะที่มีค่าอินฟอร์เมชันเกนสูงที่สุดในหลายชุดข้อมูล ได้แก่ ชนิดของกลุ่มข้อมูล ค่าเฉลี่ยระยะห่างระหว่างตัวอย่าง และค่าเบี่ยงเบนมาตรฐานระยะห่างระหว่างตัวอย่าง ดังแสดงในตารางที่ 6.2

ชื่อคุณลักษณะ	รายละเอียด	ค่าคุณลักษณะ
ชนิดของกลุ่มข้อมูล (type)		{คลาสเดียว, คลาสปะปน, ไม่ทราบคลาส}
อัตราส่วนของจำนวนตัวอย่างเรียนรู้ (numTr)	อัตราส่วนระหว่างจำนวนตัวอย่างเรียนรู้ในกลุ่ม กับจำนวนตัวอย่างเรียนรู้ทั้งหมด	[0,1]
อัตราส่วนของจำนวนตัวอย่างทดสอบ (numTst)	อัตราส่วนระหว่างจำนวนตัวอย่างทดสอบในกลุ่ม กับจำนวนตัวอย่างทดสอบทั้งหมด	[0,1]
ค่าเฉลี่ยระยะห่างระหว่างตัวอย่าง (avgDist)	ค่าเฉลี่ยของระยะห่างของตัวอย่างกับศูนย์กลาง	$\mathbb{R}$
ค่าเบี่ยงเบนมาตรฐานระยะห่างระหว่างตัวอย่าง (stdDist)	ค่าเบี่ยงเบนมาตรฐานของระยะห่างของตัวอย่างกับศูนย์กลาง	$\mathbb{R}$
รัศมีของกลุ่มข้อมูล (radius)	ระยะห่างระหว่างศูนย์กลางกับตัวอย่างที่ไกลที่สุด	$\mathbb{R}$
ค่าประมาณความหนาแน่น (density)	อัตราส่วนระหว่างจำนวนตัวอย่างเรียนรู้ กับรัศมีของกลุ่มข้อมูล	$\mathbb{R}$
Davies-Bouldin index (DBI)	ค่าประมาณคุณภาพกลุ่มข้อมูล	$\mathbb{R}$
ตำแหน่งตัวอย่างมีป้ายกำกับในกลุ่ม (L1C1)	อัตราส่วนระหว่าง ระยะห่างระหว่างตัวอย่างมีป้ายกำกับกับศูนย์กลางกลุ่มข้อมูล กับรัศมีของกลุ่มข้อมูล	[0,1]
ตำแหน่งตัวอย่างมีป้ายกำกับในกลุ่มที่ใกล้ที่สุด (L2C2)	อัตราส่วนระหว่าง ระยะห่างระหว่างตัวอย่างมีป้ายกำกับกับศูนย์กลางของกลุ่มเพื่อนบ้าน กับรัศมีของกลุ่มเพื่อนบ้าน	[0,1]
ระยะห่างระหว่างตัวอย่างมีป้ายกำกับ		$\mathbb{R}$
กับกลุ่มข้อมูลเพื่อนบ้านที่ใกล้ที่สุด (L1C2)		$\mathbb{R}$
ระยะห่างระหว่างตัวอย่างมีป้ายกำกับ		$\mathbb{R}$
ของกลุ่มที่ใกล้ที่สุดกับศูนย์กลางกลุ่มข้อมูล (L2C1)		$\mathbb{R}$

ตารางที่ 6.1: ค่าคุณลักษณะของกลุ่มข้อมูลสำปรับการวิเคราะห์กลุ่มข้อมูล

ชุดข้อมูล	คุณลักษณะเมื่อเรียงตามค่าอินฟอร์เมชันเกน		
	ลำดับที่ 1	ลำดับที่ 2	ลำดับที่ 3
adult	numTst	numTr	**avgDist**
bankmar	numTst	density	**stdDist**
banknote	L1C1	**avgDist**	**type**
bio	numTrPer	numTstPer	radius
eeg	density	**stdDist**	L1C2
madelon	**type**	**stdDist**	**avgDist**
magicgamma	**avgDist**	**stdDist**	radius
mnist1	**avgDist**	**stdDist**	numTrPer
mnist3	radius	**avgDist**	**stdDist**
mnist4	**type**	**stdDist**	**avgDist**
mnist7	**type**	**stdDist**	**avgDist**
mushroom	L1C1	**type**	L1C2
spam	**avgDist**	L2C2	radius
splice	**type**	**stdDist**	**avgDist**
wilt	radius	**avgDist**	**stdDist**

ตารางที่ 6.2: คุณลักษณะที่มีค่าอินฟอร์เมชันเกนสูงที่สุดสามอันดับแรก

รายการอ้างอิง

ScanFix Xpress [Computer software]. Available from: http://www.accusoft.com/scanfix.htm [2010].

Agrawal, M. and Doermann, D. 2011. Stroke-like pattern noise removal in binary document images. In Proceedings of the 2011 International Conference on Document Analysis and Recognition.

Bair, E. and Tibshirani, R. 2004. Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol 2.4 (2004): e108.

Basu, S., Banerjee, A., and Mooney, R. J. 2002. Semi-supervised clustering by seeding. In Proceedings of the Nineteenth International Conference on Machine Learning, ICML '02, pp. 27–34. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Belkin, M., Niyogi, P., and Sindhwani, V. 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. J. Mach. Learn. Res. 7 (Dec 2006): 2399–2434.

Ben-David, S., Lu, T., and Pál, D. 2008. Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In Proceedings of the Twenty-first Annual conference on Learning Theory, COLT, pp. 33–44.

Blum, A. and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. In Proceedings of the Eleventh Annual Conference on Computational Learning Theory, COLT' 98, pp. 92–100. New York, NY, USA: ACM.

Chapelle, O. and Zien, A. 2005. Semi-supervised classification by low density separation. In Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, pp. 57–64.

Chapelle, O., Schölkopf, B., and Zien, A. 2006. Semi-Supervised Learning. The MIT Press.

Cozman, F. G. and Cohen, I. 2002. Unlabeled data can degrade classification performance of generative classifiers. In Proceedings of the Fifteenth International Florida Artificial Intelligence Society Conference, pp. 327–331.

Dara, R., Kremer, C., and Stacey, D. A. 2002. Clustering unlabeled data with soms improves classification of labeled real-world data. In Proceedings of the 2002 International Joint Conference on Neural Networks.

Demiriz, A., Bennett, K. P., and Embrechts, M. J. 1999. Semi-supervised clustering using genetic algorithms. In Proceedings of Artificial neural networks in engineering, pp. 809–814.

Didaci, L. and Roli, F. 2006. Using co-training and self-training in semi-supervised multiple classifier systems. In Structural, Syntactic, and Statistical Pattern Recognition, volume 4109 of Lecture Notes in Computer Science, pp. 522–530. : Springer Berlin Heidelberg.

Donmez, P., Carbonell, J. G., and Bennett, P. N. 2007. Dual strategy active learning. In Proceedings of the 18th European Conference on Machine Learning, ECML '07, pp. 116–127. Berlin, Heidelberg: Springer-Verlag.

Gan, H., Sang, N., Huang, R., Tong, X., and Dan, Z. 2013. Using clustering analysis to improve semi-supervised classification. Neurocomputing 101 (2013): 290–298.

Grimaudo, L., Mellia, M., Baralis, E., and Keralapura, R. 2014. SeLeCT: Self-learning classifier for internet traffic. IEEE Transactions on Network and Service Management 11.2 (June 2014): 144–157.

Guan G., e. 2013. Joint rayleigh coefficient maximization and graph based semi-supervised for the classification of motor imagery eeg. In Proceedings of the 2013 IEEE International Conference on Information and Automation, pp. 379–383.

Guo, Y., Niu, X., and Zhang, H. 2010. An extensive empirical study on semi-supervised learning. In Proceedings of the Tenth IEEE International Conference on Data Mining (ICDM), pp. 186–195.

Guo, Y., Zhang, H., and Liu, X. 2011. Instance selection in semi-supervised learning. In Advances in Artificial Intelligence, volume 6657 of Lecture Notes in Computer Science, pp. 158–169. : Springer Berlin Heidelberg.

Hall, M. et al. 2009. The weka data mining software: An update. SIGKDD Explorations 11 (2009):

Hughes, N., Roberts, S., and Tarassenko, L. 2004. Semi-supervised learning of probabilistic models for ecg segmentation. In Proceedings on the Twentieth-sixth Annual International Conference of the IEEE on Engineering in Medicine and Biology Society, volume 1, pp. 434–437.

Joachims, T. 1999. Transductive inference for text classification using support vector machines. In ICML, volume 99, pp. 200–209.

Koestler, e. a., D.. 2010. Semi-supervised recursively partitioned mixture models for identifying cancer subtypes. Bioinformatics 26.20 (2010): 2578–2585.

Kumar Mallapragada, P., Jin, R., Jain, A., and Liu, Y. 2009. Semiboost: Boosting for semi-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 31.11 (Nov 2009): 2000–2014.

Li, M. and Zhou, Z.-H. 2005. Setred: Self-training with editing. In Advances in Knowledge Discovery and Data Mining, volume 3518 of Lecture Notes in Computer Science, pp. 611–621. : Springer Berlin Heidelberg.

Li, Y.-F. and Zhou, Z.-H. 2015. Towards making unlabeled data never hurt. IEEE Transactions on Pattern Analysis and Machine Intelligence 37.1 (Jan 2015): 175–188.

Li, Y., Guan, C., Li, H., and Chin, Z. 2008. A self-training semi-supervised {SVM} algorithm and its application in an eeg-based brain computer interface speller system. Pattern Recognition Letters 29.9 (2008): 1285–1294.

Lichman, M. 2013. UCI machine learning repository [Online]. Available from: http://archive.ics.uci.edu/ml [2013].

Liu, A., Jun, G., and Ghosh, J. 2009. A self-training approach to cost sensitive uncertainty sampling. Machine Learning 76.2-3 (2009): 257–270.

Maulik, U. and Chakraborty, D. 2011. A self-trained ensemble with semisupervised svm: An application to pixel classification of remote sensing imagery. Pattern Recognition 44.3 (2011): 615 – 623.

McClosky, D., Charniak, E., and Johnson, M. 2006. Effective self-training for parsing. In Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, HLT-NAACL '06, pp. 152–159. Stroudsburg, PA, USA: Association for Computational Linguistics.

Muslea, I., Minton, S., and Knoblock, C. A. 2002. Active + semi-supervised learning = robust multi-view learning. In Proceedings of ICML-2002, pp. 435–442.

Muslea, I., Minton, S., and Knoblock, C. A. 2006. Active learning with multiple views. Journal of artificial intelligence research 27 (2006): 203–233.

Nguyen, H. T. and Smeulders, A. 2004. Active learning using pre-clustering. In Proceedings of the Twenty-first International Conference on Machine Learning, ICML '04, pp. 79–86. New York, NY, USA: ACM.

Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. 2000. Text classification from labeled and unlabeled documents using EM. Mach. Learn. 39.2-3 (may 2000): 103–134.

Oster, J., Behar, J., Sayadi, O., Nemati, S., Johnson, A., and Clifford, G. 2015. Semi-supervised ecg ventricular beat classification with novelty detection based on switching kalman filters. IEEE Transactions on Biomedical Engineering PP.99 (2015): 1–1.

Piroonsup, N. and Sinthupinyo, S. 2010a. Applying a semi-supervised learning approach to reduce noise in thai-ocr. In Proceedings of The Second International Conference on Computer Engineering and Technology, pp. 444–449.

Piroonsup, N. and Sinthupinyo, S. 2010b. A combination of som and decision tree learning to reduce noise in thai-ocr. In Proceedings of The Seventh International Joint Conference on Computer Science and Software Engineering.

Piroonsup, N. 2009. Noise reduction in thai-ocr using semi-supervised learning. Master's thesis, Chulalongkorn University.

Quinlan, J. R. 1986. Induction of decision trees. Machine learning 1.1 (1986): 81–106.

Roli, F. and Marcialis, G. 2006. Semi-supervised pca-based face recognition using self-training. In Structural, Syntactic, and Statistical Pattern Recognition, volume 4109 of Lecture Notes in Computer Science, pp. 560–568. : Springer Berlin Heidelberg.

Rosenberg, C., Hebert, M., and Schneiderman, H. 2005. Semi-supervised self-training of object detection models. In Proceedings of the Seventh IEEE Workshops on Application of Computer Vision, volume 1, pp. 29–36.

Scudder, H. J. 1965. Probability of error of some adaptive pattern-recognition machines. IEEE Transactions on Information Theory (1965): 363–371.

Settles, B. 2010. Active learning literature survey. Technical report, University of Wisconsin-Madison.

Singh, A., Nowak, R., and Zhu, X. 2009. Unlabeled data: Now it helps, now it doesn't. In Advances in Neural Information Processing Systems 21, pp. 1513–1520. : Curran Associates, Inc.

Su, H., Chen, L., Ye, Y., Sun, Z., and Wu, Q. 2010. A refinement approach to handling model misfit in semi-supervised learning. In Advanced Data Mining and Applications, volume 6441 of Lecture Notes in Computer Science, pp. 75–86. : Springer Berlin Heidelberg.

Sugiyama, M., Imajo, K., otaki, K., and yamamoto, A. 2012. Semi-supervised ligand finding using formal concept analysis. IPSJ Transactions on Mathematical Modeling and Its Applications 5 (2012): 39–48.

Sun, L., Lu, Y., Yang, K., and Li, S. 2012. Ecg analysis using multiple instance learning for myocardial infarction detection. IEEE Transactions on Biomedical Engineering 59.12 (Dec 2012): 3348–3356.

Tian, Q., Yu, J., Xue, Q., and Sebe, N. 2004. A new analysis of the value of unlabeled data in semi-supervised learning for image retrieval. In Proceedings of the

2004 IEEE International Conference on Multimedia and Expo, volume 2, pp. 1019–1022 Vol.2.

Tong, S. and Chang, E. 2001. Support vector machine active learning for image retrieval. In Proceedings of the ninth ACM international conference on Multimedia, pp. 107–118.

Tong, S. and Koller, D. 2002. Support vector machine active learning with applications to text classification. The Journal of Machine Learning Research 2 (2002): 45–66.

Triguero, I. and Salvador García, F. H. 2015. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. Knowledge and Information Systems 42.2 (2015): 245–284.

Triguero, I., Sáez, J. A., Luengo, J., García, S., and Herrera, F. 2014. On the characterization of noise filters for self-training semi-supervised in nearest neighbor classification. Neurocomputing 132.0 (2014): 30–41.

Tur, G., Tur, D., and Schapire, R. 2005. Combining active and semi-supervised learning for spoken language understanding. Speech communication 45 (2005): 171–186.

Valizadegan, H. and Jin, R. 2007. Generalized maximum margin clustering and unsupervised kernel learning. Advances in Neural Infromation Processing Systems 19 (2007): 1417–1424.

Vapnik, V. N. 1998. Statistical Learning Theory. Wiley.

Vesanto, J. and Alhoniemi, E. 2000. Clustering of the self-organizing map. IEEE Transactions on Neural Networks 11.3 (May 2000): 586–600.

Wagstaff, K., Cardie, C., Rogers, S., and Schrödl, S. 2001. Constrained k-means clustering with background knowledge. In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, pp. 577–584. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Wang, Y., Xu, X., Zhao, H., and Hua, Z. 2010. Semi-supervised learning based on nearest neighbor rule and cut edges. Knowledge-Based Systems 23.6 (aug 2010): 547–554.

Wang, Y., Chen, S., and Zhou, Z.-H. 2012. New semi-supervised classification method based on modified cluster assumption. IEEE Transactions on Neural Networks and Learning Systems 23 (2012): 689–702.

Wei, L. and Keogh, E. 2006. Semi-supervised time series classification. In Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '06, pp. 748–753. New York, NY, USA: ACM.

Xu, Z., Yu, K., Tresp, V., Xu, X., and Wang, J. 2003. Representative sampling for text classification using support vector machines. In Advances in Information Retrieval, volume 2633 of Lecture Notes in Computer Science, pp. 393–407. : Springer Berlin Heidelberg.

Zeng, H. J., Wang, X. H., Chen, Z., Lu, H., and Ma, W. Y. 2003. CBC: clustering based text classification requiring minimal labeled data. In Proceedings of the Third IEEE International Conference on Data Mining, pp. 443–450.

Zhang, H., Huang, K., Li, D., and Zhang, L. 2013. A 12-lead clinical ecg classification method based on semi-supervised discriminant analysis. In Proceedings of the Sixth International Conference on Biomedical Engineering and Informatics, pp. 177–181.

Zhang, K., Tsang, I., and Kwok, J. 2009. Maximum margin clustering made practical. IEEE Transactions on Neural Networks 20.4 (April 2009): 583–596.

Zhang, W., Tang, X., and Yoshida, T. 2015. TESC: An approach to text classification using semi-supervised clustering. Knowledge-Based Systems 75.0 (2015): 152 – 160.

Zhou, Z. H., Chen, K. J., and Dai, H. 2006. Enhancing relevance feedback in image retrieval using unlabeled data. ACM Transactions on information systems 24.2 (April 2006): 219–244.

Zhou, Z.-H. and Li, M. 2010. Semi-supervised learning by disagreement. Knowledge and Information Systems 24.3 (2010): 415–439.

Zhu, J., Wang, H., Yao, T., and Tsou, B. K. 2008. Active learning with sampling by uncertainty and density for word sense disambiguation and text classifica-

tion. In Proceedings of the 22nd international conference on computational linguistics, volume 1, pp. 1137–1144.

Zhu, X. 2008. Semi-supervised learning literal survey. Technical Report TR-1530, University of Wisconsic Madison.

Zhu, X. 2007. Semi-supervised learning tutorial. ICML 2007 Tutorial (2007):

ภาคผนวก

ภาคผนวก ก

ค่าพารามิเตอร์ของโปรแกรม SCANFIX XPRESS 6.0

รายละเอียดค่าพารามิเตอร์ของโปรแกรม ScanFix Xpress 6.0 แสดงดังตารางที่ ก.1

วิธีลดสิ่งรบกวน	ชื่อพารามิเตอร์	ค่าพารามิเตอร์
Smooth objects	Amount	1
Despeckle	Speck width	5
	Speck height	5
Line removal	Maximum character repair size	15
	Maximum gap	5
	Maximum thickness	20
	Maximum aspect ratio	10
	Minimum length	50
Comb removal	Left	0
	Top	0
	Width	1000
	Height	1000
	Comb height	20
	Comb spacing	25
	Horizontal line thickness	4
	Vertical line thickness	4
	Minimum comb length	50
	Minimum confidence	50
Border removal	Amount to expand page	0
	Border speck size	100
	Crop border	no
	Deskew border	no
	Maximum page height	214 748
	Maximum page width	214 748
	Minimum page height	0

วิธีลดสิ่งรบกวน	ชื่อพารามิเตอร์	ค่าพารามิเตอร์
	Minimum page width	0
	Pad color	black
	Page speck size	30
	Quality	80
	Replace border	yes
Blob removal	Left	0
	Top	0
	Width	1000
	Height	1000
	Minimum density	60
	Minimum pixel count	300
	Maximum pixel count	100 000

ตารางที่ ก.1: ค่าพารามิเตอร์ของโปรแกรม ScanFix Xpress 6.0 ที่ใช้ลดสิ่งรบกวนในเอกสาร

ภาคผนวก ข

ผลการทดลองบนชุดข้อมูลการลดสิ่งรบกวนในภาษาไทย

ชุดข้อมูล	ความถูกต้องของ 1-nn บนตัวอย่างทดสอบที่อยู่ใน		ความถูกต้องของ 3-nn บนตัวอย่างทดสอบที่อยู่ใน	
	กลุ่มไม่ทราบคลาส	กลุ่มมีคลาส	กลุ่มไม่ทราบคลาส	กลุ่มมีคลาส
noise001	**98.05*** **	71.42	**97.39*** **	73.99
noise004	**96.01*** **	70.66	**96.25*** **	72.53
noise005	**95.91*** **	76.92	**95.32*** **	77.70
noise006	**94.06*** **	64.05	**94.92*** **	66.47
noise007	**93.28*** **	66.24	**95.02*** **	64.47
noise008	**96.40*** **	76.76	**96.46*** **	75.78
noise009	**94.24*** **	68.57	**94.20*** **	66.16
noise013	**91.00*** **	48.61	**91.10*** **	46.31
noise016	**93.27*** **	62.58	**94.27*** **	60.67
noise020	**80.13*** **	55.81	**82.31*** **	53.79
noise022	**94.65*** **	57.89	**95.08*** **	56.85
noise023	**74.76*** **	50.73	**76.31*** **	42.14
noise024	**98.22*** **	66.84	**98.61*** **	59.20
noise026	**96.70*** **	65.37	**97.04*** **	64.10
noise027	**95.72*** **	75.72	**95.81*** **	74.02
noise028	**96.63*** **	57.40	**96.50*** **	54.06
noise033	**98.42*** **	88.99	**98.32*** **	89.19
noise036	**97.53*** **	84.32	**97.81*** **	85.47
noise038	**82.48*** **	51.07	**86.25*** **	52.97
noise039	**93.19*** **	73.12	**94.17*** **	75.19
noise040	**94.48*** **	84.00	**94.94*** **	84.02
noise048	**78.34*** **	64.16	**79.56*** **	63.46
noise070	**86.55*** **	56.76	**88.93*** **	58.70
noise071	**87.83*** **	66.19	**89.25*** **	67.01
noise072	**82.44*** **	72.20	**88.83*** **	71.89

ชุดข้อมูล	ความถูกต้องของ 1-nn บนตัวอย่างทดสอบที่อยู่ใน		ความถูกต้องของ 3-nn บนตัวอย่างทดสอบที่อยู่ใน	
	กลุ่มไม่ทราบคลาส	กลุ่มมีคลาส	กลุ่มไม่ทราบคลาส	กลุ่มมีคลาส
noise073	83.07***	64.41	89.81***	67.09
noise074	96.31***	74.23	96.70***	73.88
noise077	95.29***	80.24	96.44***	82.05
noise100	97.99***	85.80	98.44***	86.16
noise112	98.80***	85.56	98.67***	85.58
noise113	98.55***	82.39	98.39***	84.19
noise117	97.45***	81.20	97.38***	80.42
noise118	92.82***	73.45	95.21***	78.64
noise120	98.83***	87.79	98.34***	84.98
noise121	95.95***	72.20	95.66***	69.03
noise122	96.94***	84.34	97.55***	83.71
noise125	96.38***	86.83	96.51***	86.79
noise126	95.59***	82.15	97.20***	85.35
noise131	81.86***	73.25	87.69***	75.01
noise133	74.14***	65.97	75.57***	68.31
noise134	94.21***	82.66	95.14***	84.42
noise135	94.50***	79.07	94.77***	79.83
noise147	97.46***	85.44	98.13***	86.58
noise148	97.62***	86.07	98.14***	86.47
noise150	92.71***	80.70	97.01***	83.06
noise152	98.20***	78.01	97.98***	77.64
noise154	70.28	70.1	73.49	71.62
noise155	98.04***	85.71	98.01***	86.63
noise156	97.38***	80.08	97.46***	82.07
noise158	92.49***	76.17	94.46***	79.96
noise160	98.38***	90.29	98.04***	91.59
noise161	93.81***	76.98	94.76***	79.35
noise162	86.94***	73.61	88.14***	77.11
noise163	99.17***	80.96	99.16***	81.19
noise164	97.58***	84.18	97.17***	85.33

ชุดข้อมูล	ความถูกต้องของ 1-nn บนตัวอย่างทดสอบที่อยู่ใน		ความถูกต้องของ 3-nn บนตัวอย่างทดสอบที่อยู่ใน	
	กลุ่มไม่ทราบคลาส	กลุ่มมีคลาส	กลุ่มไม่ทราบคลาส	กลุ่มมีคลาส
noise165	**97.02*** **	73.76	**96.05*** **	73.40
noise166	**84.29*** **	68.94	**87.17*** **	74.23
noise170	**83.59*** **	74.55	**83.21*** **	65.63
noise171	**92.04*** **	76.64	**89.04*** **	66.80
noise172	**88.59*** **	72.01	**88.15*** **	70.36
noise173	**93.03*** **	67.30	**92.59*** **	66.09
noise174	**75.25** **	61.55	**76.48*** **	47.57
noise177	**86.34*** **	65.51	**93.89*** **	68.51
noise180	**72.90*** **	52.84	**68.39*** **	28.95
noise184	**86.94*** **	69.69	**84.89*** **	57.77
noise185	**90.92*** **	78.85	**88.64*** **	80.05
noise188	**86.56*** **	60.32	**83.97*** **	32.44

ตารางที่ ข.1: เปรียบเทียบความถูกต้องของตัวจำแนกกึ่งมีผู้สอนเมื่อทำนายบนตัวอย่างทดสอบที่อยู่ในกลุ่มมีคลาสและกลุ่มไม่ทราบคลาสบนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย

ชุดข้อมูล	ความถูกต้องของ 1-nn		ความถูกต้องของ 3-nn	
	ตัวอย่างมีป้ายกำกับเดิม	เพิ่มตัวอย่างในกลุ่มไม่ทราบคลาส	ตัวอย่างมีป้ายกำกับเดิม	เพิ่มตัวอย่างในกลุ่มไม่ทราบคลาส
noise001	93.82	**97.00*** **	93.55	**96.01*** **
noise004	93.17	**93.99*** **	93.68	**93.90***
noise005	92.9	**95.18*** **	92.93	**94.17*** **
noise006	89.59	**92.22*** **	90.88	**91.98*** **
noise007	90.66	**91.93*** **	92.07	**93.39*** **
noise008	94.18	**95.07*** **	94.14	**94.94*** **
noise009	90.18	**91.65*** **	89.82	**91.52*** **
noise013	87.1	**90.11*** **	87.21	**88.04*** **
noise016	89.43	**91.84*** **	90.25	**92.15*** **
noise020	77.32	**79.77*** **	79.52	**80.20*** **

ชุดข้อมูล	ความถูกต้องของ 1-nn		ความถูกต้องของ 3-nn	
	ตัวอย่างมีป้ายกำกับเดิม	เพิ่มตัวอย่างในกลุ่มไม่ทราบคลาส	ตัวอย่างมีป้ายกำกับเดิม	เพิ่มตัวอย่างในกลุ่มไม่ทราบคลาส
noise022	91.46	**92.22*** **	91.8	**92.68*** **
noise023	72.52	**74.56*** **	73.3	**74.59** **
noise024	94.43	**97.76*** **	94.34	**96.68*** **
noise026	**93.43*** **	92.80	93.75	**93.94**
noise027	93.61	**94.78*** **	93.53	**94.55*** **
noise028	94.83	**95.64*** **	94.79	**95.28*** **
noise033	96.8	**96.88**	96.85	**97.20** **
noise036	**95.69** *	95.28	96.12	**96.24**
noise038	78.89	**81.00*** **	82.45	**84.00*** **
noise039	90.99	**91.99*** **	92.1	**92.95*** **
noise040	93.32	**93.81*** **	93.76	**94.32*** **
noise048	77.01	**78.35*** **	78.07	**79.51*** **
noise070	82.8	**85.30*** **	85.09	**86.66*** **
noise071	85.67	**86.83*** **	87.1	**88.12*** **
noise072	80.83	**82.96*** **	86.53	**87.65*** **
noise073	80.31	**82.24*** **	86.54	**87.69*** **
noise074	93.71	**94.65*** **	94.13	**94.18**
noise077	93.8	**94.39*** **	95.04	**95.26*** **
noise100	96.7	**97.51*** **	97.15	**98.04*** **
noise112	97.1	**97.62*** **	97.03	**97.67*** **
noise113	96.57	**97.50*** **	96.74	**97.17*** **
noise117	94.71	**96.42*** **	94.51	**96.12*** **
noise118	89.51	**91.87*** **	92.39	**93.51*** **
noise120	96.64	**98.12*** **	95.78	**97.93*** **
noise121	93.07	**94.37*** **	92.5	**93.96*** **
noise122	95.3	**96.13*** **	95.8	**96.46*** **
noise125	94.65	**95.49*** **	94.76	**95.27*** **
noise126	93.37	**94.63*** **	95.21	**95.74*** **
noise131	80.78	**81.76*** **	85.94	**86.60*** **
noise133	72.98	**73.97*** **	74.55	**75.15*** **

ชุดข้อมูล	ความถูกต้องของ 1-nn		ความถูกต้องของ 3-nn	
	ตัวอย่างมีป้ายกำกับเดิม	เพิ่มตัวอย่างในกลุ่มไม่ทราบคลาส	ตัวอย่างมีป้ายกำกับเดิม	เพิ่มตัวอย่างในกลุ่มไม่ทราบคลาส
noise134	92.08	**93.11*** **	93.19	**93.85*** **
noise135	91.48	**92.62*** **	91.83	**92.63*** **
noise147	95.92	**96.82*** **	96.68	**97.11*** **
noise148	96.18	**97.29*** **	96.71	**97.47*** **
noise150	91.55	**92.04*** **	95.56	**95.69**
noise152	96.03	**97.45*** **	95.82	**96.80*** **
noise154	70.05	**70.15**	73.09	**73.22**
noise155	96.27	**96.92*** **	96.37	**96.73*** **
noise156	95.51	**96.45*** **	95.8	**95.87**
noise158	89.66	**91.47*** **	92.05	**93.09*** **
noise160	97.01	**98.10*** **	96.95	**97.82*** **
noise161	90.11	**92.56*** **	91.53	**93.81*** **
noise162	84.46	**84.85**	86.02	**86.46**
noise163	96.97	**98.23*** **	96.92	**97.79*** **
noise164	94.88	**97.03*** **	94.8	**97.23*** **
noise165	94.95	**95.72*** **	94.09	**94.73*** **
noise166	81.03	**82.48*** **	84.37	**84.69**
noise170	82.38	**83.69*** **	81.16	**82.70*** **
noise171	89.44	**90.26** *	85.82	**87.28*** **
noise172	85.52	**87.55*** **	85.14	**87.25*** **
noise173	89.43	**91.45*** **	89.19	**90.67*** **
noise174	76.65	**79.06*** **	75.93	**82.04*** **
noise177	83.54	**87.07*** **	90.28	**92.95*** **
noise180	72.08	**74.84*** **	66.61	**70.02*** **
noise184	84.03	**86.97*** **	81.03	**85.12*** **
noise185	89.45	**91.56*** **	87.59	**88.96*** **
noise188	84.51	**86.74*** **	80.45	**83.58*** **

ตารางที่ ข.2: เปรียบเทียบประสิทธิภาพของตัวจำแนกกึ่งมีผู้สอนระหว่างใช้ป้ายกำกับเดิมกับเมื่อผู้ใช้เพิ่มป้ายกำกับในกลุ่มไม่ทราบคลาสบนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย

ชุดข้อมูล	ความถูกต้องของ 1-nn เมื่อเพิ่มตัวอย่างมีป้ายกำกับในกลุ่ม		ความถูกต้องของ 3-nn เมื่อเพิ่มตัวอย่างมีป้ายกำกับในกลุ่ม	
	ไม่ทราบคลาส	มีคลาส	ไม่ทราบคลาส	มีคลาส
noise001	**97*** **	93.95	**96.01*** **	94.13
noise004	**93.99*** **	93.10	**93.9***	93.67
noise005	**95.18*** **	93.60	**94.17**	93.92
noise006	**92.22*** **	90.84	**91.98**	91.8
noise007	**91.93*** **	90.95	**93.39*** **	92.15
noise008	**95.07*** **	94.35	**94.94*** **	94.28
noise009	**91.65*** **	90.54	**91.52*** **	90.11
noise013	**90.11*** **	86.74	**88.04*** **	87.34
noise016	**91.84*** **	89.73	**92.15*** **	90.43
noise020	**79.77**	79.59	80.2	**80.79**
noise022	**92.22*** **	91.06	**92.68*** **	91.76
noise023	**74.56**	74.16	**74.59**	74.06
noise024	**97.76*** **	94.70	**96.68**	94.86
noise026	92.8	**93.71*** **	**93.94** *	93.56
noise027	**94.78*** **	93.85	**94.55*** **	93.74
noise028	**95.64*** **	94.83	**95.28*** **	94.57
noise033	**96.88**	96.86	**97.2**	97.02
noise036	95.28	**95.69***	**96.24*** **	95.86
noise038	**81*** **	79.81	**84*** **	83.18
noise039	**91.99***	91.51	**92.95***	92.57
noise040	**93.81** *	93.37	**94.32** *	93.97
noise048	**78.35*** **	77.74	**79.51** *	79.04
noise070	**85.3*** **	83.84	**86.66**	86.41
noise071	**86.83** *	86.23	**88.12*** **	87.38
noise072	**82.96*** **	81.67	**87.65*** **	86.50
noise073	**82.24*** **	80.39	**87.69** *	86.39
noise074	**94.65*** **	93.73	**94.18**	94.08
noise077	**94.39** *	94.07	**95.26**	95.15
noise100	**97.51*** **	96.96	**98.04*** **	97.31
noise112	**97.62*** **	97.19	**97.67*** **	97.06

ชุดข้อมูล	ความถูกต้องของ 1-nn เมื่อเพิ่มตัวอย่างมีป้ายกำกับในกลุ่ม		ความถูกต้องของ 3-nn เมื่อเพิ่มตัวอย่างมีป้ายกำกับในกลุ่ม	
	ไม่ทราบคลาส	มีคลาส	ไม่ทราบคลาส	มีคลาส
noise113	**97.5*** **	96.72	**97.17*** **	96.72
noise117	**96.42*** **	94.87	**96.12*** **	94.85
noise118	**91.87*** **	90.38	**93.51*** **	92.54
noise120	**98.12*** **	96.79	**97.93*** **	96.26
noise121	**94.37*** **	93.34	**93.96*** **	92.70
noise122	**96.13*** **	95.50	**96.46*** **	95.86
noise125	**95.49*** **	94.83	**95.27*** **	94.84
noise126	**94.63*** **	93.75	**95.74*** **	95.21
noise131	**81.76**	81.29	**86.6** *	85.90
noise133	**73.97*** **	73.17	**75.15**	74.92
noise134	**93.11*** **	92.31	**93.85** *	93.52
noise135	**92.62*** **	91.84	**92.63*** **	91.94
noise147	**96.82*** **	96.28	**97.11*** **	96.89
noise148	**97.29*** **	96.49	**97.47*** **	96.91
noise150	92.04	**92.18**	95.69	**95.73**
noise152	**97.45*** **	96.24	**96.8*** **	96.02
noise154	**70.15**	69.85	73.22	**73.64**
noise155	**96.92*** **	96.34	**96.73*** **	96.41
noise156	**96.45*** **	95.68	**95.87**	95.8
noise158	**91.47**	91.44	**93.09**	92.74
noise160	**98.1*** **	97.08	**97.82*** **	96.86
noise161	**92.56*** **	90.91	**93.81*** **	92.32
noise162	84.85	**85.98**	86.46	**87.20**
noise163	**98.23*** **	97.00	**97.79** *	97.27
noise164	**97.03*** **	95.47	**97.23*** **	95.85
noise165	**95.72*** **	94.90	**94.73**	94.57
noise166	**82.48** *	81.39	**84.69**	84.57
noise170	**83.69** *	82.99	**82.7*** **	81.40
noise171	**90.26**	89.96	87.28	**87.41**
noise172	**87.55*** **	86.40	**87.25*** **	85.76

ชุดข้อมูล	ความถูกต้องของ 1-nn เมื่อเพิ่มตัวอย่างมีป้ายกำกับในกลุ่ม		ความถูกต้องของ 3-nn เมื่อเพิ่มตัวอย่างมีป้ายกำกับในกลุ่ม	
	ไม่ทราบคลาส	มีคลาส	ไม่ทราบคลาส	มีคลาส
noise173	**91.45*** **	90.45	**90.67*** **	89.90
noise174	**79.06**	78.38	82.04	**82.21**
noise177	**87.07**	86.11	**92.95*** **	90.98
noise180	**74.84**	73.53	**70.02**	68.45
noise184	**86.97*** **	85.23	**85.12*** **	81.98
noise185	**91.56***	90.88	**88.96**	88.43
noise188	**86.74**	86.65	**83.58*** **	81.48

ตารางที่ ข.3: เปรียบเทียบประสิทธิภาพตัวจำแนกกึ่งมีผู้สอนเมื่อเพิ่มตัวอย่างมีป้ายกำกับในกลุ่มมีป้ายกำกับกับกลุ่มไม่มีป้ายกำกับบนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย

ชุดข้อมูล	ความถูกต้องของ 1-nn เมื่อเพิ่มตัวอย่างที่		ความถูกต้องของ 3-nn เมื่อเพิ่มตัวอย่างที่	
	ศูนย์กลางกลุ่ม	สุ่มตำแหน่ง	ศูนย์กลางกลุ่ม	สุ่มตำแหน่ง
noise001	**97*** **	96.40	**96.01** *	95.55
noise004	**93.99**	93.86	93.9	**93.94**
noise005	**95.18** *	94.33	**94.17**	94.1
noise006	**92.22**	92.16	**91.98**	91.53
noise007	**91.93**	91.9	**93.39**	93.32
noise008	**95.07**	95	94.94	**94.95**
noise009	91.65	**91.77**	**91.52**	91.49
noise013	**90.11*** **	88.94	**88.04** *	87.65
noise016	**91.84*** **	90.94	**92.15*** **	91.53
noise020	**79.77*** **	78.35	**80.2***	79.43
noise022	92.22	**92.34**	92.68	**92.71**
noise023	**74.56**	74.31	**74.59**	74.52
noise024	**97.76*** **	97.09	**96.68**	96.47
noise026	92.8	**93.44*** **	93.94	**93.96**
noise027	**94.78**	94.63	**94.55** *	94.32
noise028	**95.64***	95.35	**95.28**	95.27

ชุดข้อมูล	ความถูกต้องของ 1-nn เมื่อเพิ่มตัวอย่างที่		ความถูกต้องของ 3-nn เมื่อเพิ่มตัวอย่างที่	
	ศูนย์กลางกลุ่ม	สุ่มตำแหน่ง	ศูนย์กลางกลุ่ม	สุ่มตำแหน่ง
noise033	96.88	**96.94**	**97.2**	97.17
noise036	95.28	**95.49**	**96.24**	96.14
noise038	81	**81.15**	84	**84.30**
noise039	**91.99**	91.8	**92.95**	92.88
noise040	**93.81**	93.71	**94.32**	94.23
noise048	**78.35*** **	78.02	**79.51**	79.45
noise070	**85.3*** **	84.70	**86.66**	86.5
noise071	**86.83**	86.7	**88.12***	87.91
noise072	**82.96*** **	82.02	**87.65***	87.33
noise073	**82.24** **	81.30	**87.69**	87.23
noise074	**94.65*** **	94.04	**94.18**	94.1
noise077	94.39	**94.42**	95.26	**95.28**
noise100	97.51	**97.55**	**98.04** **	97.89
noise112	**97.62***	97.45	**97.67**	97.63
noise113	**97.5**	97.4	97.17	**97.18**
noise117	**96.42*** **	95.92	**96.12*** **	95.80
noise118	**91.87*** **	91.06	**93.51*** **	92.77
noise120	**98.12** **	97.88	**97.93*** **	97.68
noise121	94.37	94.33	**93.96**	93.88
noise122	96.13	96.06	**96.46**	96.35
noise125	**95.49***	95.37	95.27	**95.30**
noise126	**94.63***	94.37	**95.74**	95.6
noise131	**81.76**	81.41	**86.6**	86.44
noise133	**73.97***	73.56	**75.15**	74.86
noise134	**93.11*** **	92.53	**93.85** **	93.52
noise135	**92.62*** **	91.68	**92.63** **	92.06
noise147	**96.82*** **	96.62	**97.11**	97.06
noise148	**97.29*** **	96.93	**97.47*** **	97.27
noise150	92.04	**92.22**	95.69	**95.78**
noise152	**97.45**	97.32	**96.8**	96.75

ชุดข้อมูล	ความถูกต้องของ 1-nn เมื่อเพิ่มตัวอย่างที่		ความถูกต้องของ 3-nn เมื่อเพิ่มตัวอย่างที่	
	ศูนย์กลางกลุ่ม	สุ่มตำแหน่ง	ศูนย์กลางกลุ่ม	สุ่มตำแหน่ง
noise154	**70.15**	70.03	73.22	**73.40**
noise155	**96.92**	96.79	**96.73**	96.68
noise156	**96.45*****	95.99	**95.87*****	95.62
noise158	**91.47***	90.95	**93.09****	92.61
noise160	**98.1*****	97.54	**97.82*****	97.49
noise161	**92.56****	92.08	**93.81****	93.33
noise162	84.85	**85.05**	86.46	**88.19*****
noise163	**98.23****	97.86	**97.79****	97.66
noise164	**97.03**	97.01	**97.23*****	96.66
noise165	**95.72**	95.7	**94.73**	94.51
noise166	**82.48*****	80.90	**84.69**	84.29
noise170	83.69	**83.76**	82.7	**83.21***
noise171	**90.26**	90.23	87.28	**87.76***
noise172	**87.55**	87.19	87.25	**87.43**
noise173	91.45	**91.62**	**90.67**	90.59
noise174	**79.06***	78.77	**82.04**	81.39
noise177	**87.07****	86.53	**92.95*****	92.49
noise180	74.84	**75.14**	70.02	**70.22**
noise184	**86.97*****	86.22	**85.12*****	84.30
noise185	**91.56*****	91.02	**88.96**	88.87
noise188	**86.74**	86.73	**83.58**	83.51

ตารางที่ ข.4: เปรียบเทียบประสิทธิภาพตัวจำแนกกึ่งมีผู้สอนเมื่อเพิ่มป้ายกำกับในบริเวณศูนย์กลางและบริเวณอื่น ๆ ของกลุ่มไม่มีป้ายกำกับบนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย

ชุดข้อมูล	ความถูกต้องของตัวจำแนก 1-nn กึ่งมีผู้สอนที่สร้างจากตัวอย่างมีป้ายกำกับ							
	ชุดตั้งต้น	ชุดปรับปรุงด้วยผู้ใช้	ชุดปรับปรุงด้วยตัวจำแนก					
			1-nn	3-nn	Neural Network	SVM	Tree	Random Forest
noise001	93.82	**97.00+++**	93.74	93.5	92.91-	93.49	**94.85++**	**95.02+++**
noise004	93.17	**93.99+++**	**93.18**	**93.58+++**	91.36–	90.83–	92.32–	92.94
noise005	92.90	**95.18+++**	92.52-	**93.22**	90.98–	92.6	90.68–	**93.04**
noise006	89.59	**92.22+++**	89.39	**90.18**	88.73	**90.23**	89.22	89.47
noise007	90.66	**91.93+++**	**90.68**	**90.69**	89.78–	89.24–	90.52	90.52
noise008	94.18	**95.07+++**	94.12	94.12	93.35–	92.75–	93.41–	94.15
noise009	90.18	**91.65+++**	90.18	89.88–	89.32–	88.32–	89.93	89.93
noise013	87.10	**90.11+++**	87.05	**87.43**	85.00–	87.05	**88.19+++**	86.52
noise016	89.43	**91.84+++**	89.10	**89.43**	88.75	89.08	**89.86**	**89.72**
noise020	77.32	**79.77+++**	77.23	**78.21+++**	**77.87**	76.28-	**78.29++**	**78.57+++**
noise022	91.46	**92.22+++**	91.10–	91.13–	90.18–	89.85–	90.58–	90.62–
noise023	72.52	**74.56+++**	**72.86+**	**72.52**	72.46	**72.59**	**72.84**	72.48
noise024	94.43	**97.76+++**	94.33	93.85–	94.11	**95.25+**	**96.41+++**	94.39
noise026	93.43	92.80–	92.76–	92.70–	91.79–	91.11–	90.67–	92.54–
noise027	93.61	**94.78+++**	**93.68**	93.6	92.97–	92.43–	92.62–	93.43
noise028	94.83	**95.64+++**	**94.93**	**94.84**	93.87–	93.43–	92.50–	**94.84**
noise033	96.80	**96.88**	96.52–	96.73	94.09–	95.24–	96.03–	96.72
noise036	95.69	95.28-	95.16–	95.57	94.40–	94.70–	94.62–	94.82–

ชุดข้อมูล	ความถูกต้องของตัวจำแนก 1-nn กึ่งมีผู้สอนที่สร้างจากตัวอย่างมีป้ายกำกับ							
	ชุดตั้งต้น	ชุดปรับปรุงด้วยผู้ใช้	ชุดปรับปรุงด้วยตัวจำแนก					
			1-nn	3-nn	Neural Network	SVM	Tree	Random Forest
noise038	78.89	**81.00+++**	**78.92**	**79.08**	**79.17**	77.11–	**79.12**	78.75
noise039	90.99	**91.99+++**	90.96	**91.09**	**91.12**	89.96–	90.96	**91.22**
noise040	93.32	**93.81+++**	93.32	93.27	93.04	92.87–	**93.46**	**93.52++**
noise048	77.01	**78.35+++**	76.90	77	76.94	76.29–	**77.14**	**77.14**
noise070	82.80	**85.30+++**	82.65	**83.00**	82.34	80.88–	82.54	**82.94**
noise071	85.67	**86.83+++**	85.41–	85.63	**86.05+**	84.90–	85.67	85.61
noise072	80.83	**82.96+++**	80.24–	**81.00**	79.93-	80.04	80.65	**81.41++**
noise073	80.31	**82.24+++**	80.00–	**80.84+**	79.61	**80.41**	80.24	**80.67**
noise074	93.71	**94.65+++**	93.27–	93.28–	92.26–	92.28–	92.44–	**93.78**
noise077	93.80	**94.39+++**	**93.90**	**94.02++**	93.65	92.13–	93.37–	**94.21+++**
noise100	96.70	**97.51+++**	96.63	96.61	96.67	94.65–	96.07–	**97.30+++**
noise112	97.10	**97.62+++**	**97.13**	**97.10**	95.98–	95.61–	96.44–	97.07
noise113	96.57	**97.50+++**	**96.67**	**96.99+++**	95.82–	94.13–	95.14–	**97.07+++**
noise117	94.71	**96.42+++**	94.55–	94.43–	94.61	93.33–	94.49	**95.46+++**
noise118	89.51	**91.87+++**	89.34	**90.30+++**	88.39–	88.58–	89.2	**90.39+++**
noise120	96.64	**98.12+++**	96.48	96.06–	96.12-	96.14–	96.16–	**96.84**
noise121	93.07	**94.37+++**	**93.08**	92.83–	91.92–	91.27–	92.41–	92.82–
noise122	95.30	**96.13+++**	**95.39**	**95.39**	94.85–	93.83–	94.80–	**95.79+++**

ชุดข้อมูล	ความถูกต้องของตัวจำแนก 1-nn กึ่งมีผู้สอนที่สร้างจากตัวอย่างมีป้ายกำกับ							
	ชุดตั้งต้น	ชุดปรับปรุงด้วยผู้ใช้	ชุดปรับปรุงด้วยตัวจำแนก					
			1-nn	3-nn	Neural Network	SVM	Tree	Random Forest
noise125	94.65	95.49+++	94.63	94.73	94.28	94.28–	94.75	94.85+++
noise126	93.37	94.63+++	93.13-	93.49	91.97–	92.23–	92.63–	92.98
noise131	80.78	81.76+++	80.49–	80.76	80.10–	79.86–	80.90	80.88
noise133	72.98	73.97+++	72.82	73.36+	72.27–	73.68+++	72.7	73.35+
noise134	92.08	93.11+++	91.78–	92.38+	91.55-	91.15–	92.12	92.74+++
noise135	91.48	92.62+++	91.21–	91.52	90.36–	90.31–	91.39	91.99++
noise147	95.92	96.82+++	95.88	96.20+++	95.77	94.90–	96.10	96.32+++
noise148	96.18	97.29+++	96.08	96.25	96.26	95.26–	96.18	96.34
noise150	91.55	92.04+++	91.57	91.93+++	90.85–	90.61–	91.25-	91.55
noise152	96.03	97.45+++	95.97	95.94	96.28	94.47–	96.17	96.82+++
noise154	70.05	70.15	69.96	70.15	69.50–	69.82	70.12	70.27+
noise155	96.27	96.92+++	96.29	96.37+	96.07	95.22–	96.28	96.65+++
noise156	95.51	96.45+++	95.55	95.80+++	95.61	94.99–	95.97+++	96.07+++
noise158	89.66	91.47+++	89.45	90.10	86.99–	86.01–	85.89–	87.54–
noise160	97.01	98.10+++	97.31+++	97.33+	96.04–	96.35–	97.55+++	97.73+++
noise161	90.11	92.56+++	89.73–	90.64	88.18–	91.01+++	89.17–	90.93+++
noise162	84.46	84.85	83.71–	84.88	80.76–	85.00	86.42+++	85.21
noise163	96.97	98.23+++	97.60+++	96.57	94.01–	94.60–	95.42–	97.57++

ชุดข้อมูล	ความถูกต้องของตัวจำแนก 1-nn กึ่งมีผู้สอนที่สร้างจากตัวอย่างมีป้ายกำกับ							
	ชุดตั้งต้น	ชุดปรับปรุงด้วยผู้ใช้	ชุดปรับปรุงด้วยตัวจำแนก					
			1-nn	3-nn	Neural Network	SVM	Tree	Random Forest
noise164	94.88	**97.03+++**	94.69	**95.06**	92.30–	**95.84+++**	94.65	**95.66+++**
noise165	94.95	**95.72+++**	**95.00**	**95.09**	94.21–	92.75–	94.88	94.79
noise166	81.03	**82.48+++**	**81.23**	**82.17+++**	80.38	80.36	80.40	**81.43**
noise170	82.38	**83.69+++**	81.70–	81.00–	80.52–	81.18–	82.29	**82.59**
noise171	89.44	**90.26++**	88.60–	88.12–	87.71–	87.61–	89.04	89.11
noise172	85.52	**87.55+++**	85.22-	85.36	84.13–	84.95-	**85.83**	**85.84**
noise173	89.43	**91.45+++**	**89.45**	**89.67**	87.12–	**89.79**	**89.82**	**89.98+**
noise174	76.65	**79.06+++**	76.35	74.82-	71.46–	72.03–	**77.93++**	74.25–
noise177	83.54	**87.07+++**	**83.78**	**84.39+++**	82.98	81.43–	82.86–	**85.02+++**
noise180	72.08	**74.84+++**	72.04	70.94–	71.18	71.61	69.69–	**72.43**
noise184	84.03	**86.97+++**	**84.12**	83.03–	83.58	83.01–	**84.08**	**84.18**
noise185	89.45	**91.56+++**	89.15	89.21	87.49–	86.26–	88.67–	88.88-
noise188	84.51	**86.74+++**	**84.78**	83.30–	84.17	84.48	82.93–	**85.81+++**
ดีกว่าชุดตั้งต้น		**62**	3	15	1	4	8	**26**
แย่กว่าชุดตั้งต้น		2	19	14	41	49	26	7

ตารางที่ ข.5: เปรียบเทียบประสิทธิภาพของตัวจำแนกกึ่งมีผู้สอนด้วยเพื่อนบ้านใกล้ที่สุด ระหว่างการใช้ตัวอย่างมีป้ายกำกับเดิม กับเมื่อเพิ่มป้ายกำกับในกลุ่มไม่มีป้ายกำกับด้วยตัวจำแนกที่สร้างจากขั้นตอนวิธีต่าง ๆ บนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย

ชุดข้อมูล	ความถูกต้องของตัวจำแนก 3-nn กึ่งมีผู้สอนที่สร้างจากตัวอย่างมีป้ายกำกับ							
	ชุดตั้งต้น	ชุดปรับปรุงด้วยผู้ใช้	ชุดปรับปรุงด้วยตัวจำแนก					
			1-nn	3-nn	Neural Network	SVM	Tree	Random Forest
noise001	93.55	**96.01+++**	92.67–	92.60–	**93.75**	**94.56++**	**94.17**	**94.32+**
noise004	93.68	**93.90+**	93.32–	93.62	93.00–	92.78–	93.29–	93.33–
noise005	92.93	**94.17+++**	90.58–	91.04–	91.72–	**93.87+++**	91.62–	92.27
noise006	90.88	**91.98+++**	88.96–	89.05–	89.38–	90.57	89.22–	89.53–
noise007	92.07	**93.39+++**	**92.18**	**92.17**	91.73–	91.40–	**92.20**	91.98
noise008	94.14	**94.94+++**	**94.23**	94.09	94.07	93.95	94.12	**94.35++**
noise009	89.82	**91.52+++**	**90.29+++**	**89.82**	**89.99**	89.56	**90.37+++**	**89.98**
noise013	87.21	**88.04+++**	**87.56+**	**87.50+**	87.03	**87.69+++**	**87.82+++**	**87.25**
noise016	90.25	**92.15+++**	89.36–	89.00–	89.74	**90.54**	89.83	90.22
noise020	79.52	**80.20+++**	79.06	79.35	**79.69**	79.09	**79.84**	**79.66**
noise022	91.80	**92.68+++**	91.55–	91.43–	91.43–	91.35–	91.43–	91.41–
noise023	73.30	**74.59++**	**73.63**	73.19	**73.76**	**73.31**	**73.65**	73.23
noise024	94.34	**96.68+++**	**94.81++**	**94.39**	**94.89+**	**95.34+++**	**95.95+++**	**94.73**
noise026	93.75	**93.94**	93.68	93.67	93.37	93.61	93.72	93.67
noise027	93.53	**94.55+++**	**93.73+**	**93.55**	93.5	93.53	**93.55**	**93.77**
noise028	94.79	**95.28+++**	**94.83**	94.76	**94.88**	**94.80**	**94.84**	**94.81**
noise033	96.85	**97.20++**	96.84	96.82	96.42-	**97.07**	**96.94**	**97.05**
noise036	96.12	**96.24**	95.93	96.01	95.85	96.01	95.71–	95.87-

ชุดข้อมูล	ความถูกต้องของตัวจำแนก 3-nn กึ่งมีผู้สอนที่สร้างจากตัวอย่างมีป้ายกำกับ							
	ชุดตั้งต้น	ชุดปรับปรุงด้วยผู้ใช้	ชุดปรับปรุงด้วยตัวจำแนก					
			1-nn	3-nn	Neural Network	SVM	Tree	Random Forest
noise038	82.45	**84.00+++**	82.07–	82.34	**82.48**	81.01–	82.23	81.99–
noise039	92.10	**92.95+++**	91.88	91.97	**92.37++**	91.55–	**92.16**	**92.20**
noise040	93.76	**94.32+++**	**93.86**	**93.79**	**93.88**	**93.79**	**94.10+++**	**94.07+++**
noise048	78.07	**79.51+++**	78	**78.14**	**78.22**	77.94	**78.28++**	**78.32++**
noise070	85.09	**86.66+++**	84.74	**85.13**	84.74	83.77–	84.97	84.89
noise071	87.10	**88.12+++**	87.04	87.05	**87.52+++**	86.65–	**87.18**	**87.11**
noise072	86.53	**87.65+++**	85.93–	86.36	86.03	85.87	86.31	**86.83**
noise073	86.54	**87.69+++**	85.65–	86.44	85.95-	**86.66**	86.43	86.25
noise074	94.13	**94.18**	93.62–	93.59–	93.62–	93.69–	93.71–	93.72–
noise077	95.04	**95.26+++**	94.91	94.98	95	94.46–	94.92	**95.18+**
noise100	97.15	**98.04+++**	**97.25**	**97.25+**	**97.57+++**	96.56–	**97.22**	**97.74+++**
noise112	97.03	**97.67+++**	**97.29+++**	**97.30+++**	**97.05**	97.01	**97.05**	**97.28+++**
noise113	96.74	**97.17+++**	96.6	96.73	96.66	96.04–	96.24–	**96.86**
noise117	94.51	**96.12+++**	**94.64**	94.27–	**94.95+++**	94.10–	**94.99+++**	**95.12+++**
noise118	92.39	**93.51+++**	90.91–	92.01	91.70-	92.33	91.85–	92.35
noise120	95.78	**97.93+++**	**96.57+++**	**95.94**	**96.52+++**	**96.74+++**	**96.42+++**	**96.88+++**
noise121	92.50	**93.96+++**	**92.83++**	92.49	92.45	91.99–	**92.60**	92.41
noise122	95.80	**96.46+++**	**96.04++**	**95.95++**	**95.85**	95.78	95.78	**96.27+++**

ชุดข้อมูล	ความถูกต้องของตัวจำแนก 3-nn กึ่งมีผู้สอนที่สร้างจากตัวอย่างมีป้ายกำกับ							
	ชุดตั้งต้น	ชุดปรับปรุงด้วยผู้ใช้	ชุดปรับปรุงด้วยตัวจำแนก					
			1-nn	3-nn	Neural Network	SVM	Tree	Random Forest
noise125	94.76	**95.27+++**	**94.76**	**94.77**	94.75	94.64–	**94.91+++**	**94.80**
noise126	95.21	**95.74+++**	94.60–	94.84–	94.29–	94.70–	94.81–	94.98
noise131	85.94	**86.60+++**	85.64–	85.71–	85.68	85.22–	**85.97**	85.84
noise133	74.55	**75.15+++**	73.82–	**74.62**	73.63–	**74.92+**	74.11-	**74.69**
noise134	93.19	**93.85+++**	92.66–	93.01-	92.86–	**93.31**	93.07	**93.42++**
noise135	91.83	**92.63+++**	91.64	91.76	91.51	91.69	**92.00**	**92.31+++**
noise147	96.68	**97.11+++**	96.40-	**96.70**	96.63	96.42-	**96.72**	**96.83++**
noise148	96.71	**97.47+++**	96.6	**96.74**	**96.75**	96.50	**96.73**	**96.84**
noise150	95.56	**95.69**	95.22–	95.43-	94.81–	94.68–	95.17–	95.22–
noise152	95.82	**96.80+++**	**95.91**	95.82	**96.25+++**	95.36–	**96.11+++**	**96.37+++**
noise154	73.09	**73.22**	72.64–	72.89–	72.70–	72.9	72.95	73.05
noise155	96.37	**96.73+++**	96.29	96.33	**96.40**	96.05–	**96.39**	**96.54+++**
noise156	95.80	**95.87**	95.63-	**95.90**	95.77	95.71	**95.89**	**95.98++**
noise158	92.05	**93.09+++**	91.01–	91.63	90.25–	88.89–	89.59–	90.48–
noise160	96.95	**97.82+++**	**97.19**	**97.26++**	**96.98**	**97.21**	**97.46+++**	**97.51+++**
noise161	91.53	**93.81+++**	90.59-	91.32	90.92	**92.25+**	91.25	**92.08**
noise162	86.02	**86.46**	84.41-	**86.75**	84.29-	**86.63**	**89.01+++**	**87.56+++**
noise163	96.92	**97.79+++**	**97.68+++**	**97.15**	96.63	**97.50++**	**97.71+++**	**97.77+++**

ชุดข้อมูล	ความถูกต้องของตัวจำแนก 3-nn กึ่งมีผู้สอนที่สร้างจากตัวอย่างมีป้ายกำกับ							
	ชุดตั้งต้น	ชุดปรับปรุงด้วยผู้ใช้	ชุดปรับปรุงด้วยตัวจำแนก					
			1-nn	3-nn	Neural Network	SVM	Tree	Random Forest
noise164	94.80	**97.23+++**	94.24-	94.12–	93.47–	**96.06+++**	**95.05**	**95.54++**
noise165	94.09	**94.73+++**	94.01	94.04	**94.35**	**94.58**	**94.37**	**94.12**
noise166	84.37	**84.69**	83.51–	84.09	83.46–	83.66-	83.33–	83.86-
noise170	81.16	**82.70+++**	**81.87++**	80.96	**82.20+++**	**81.97++**	**82.57+++**	**82.90+++**
noise171	85.82	**87.28+++**	**86.12**	85.8	**86.77++**	**86.96+++**	**87.24+++**	**87.56+++**
noise172	85.14	**87.25+++**	84.92	84.85	84.65	**85.28**	**85.99+++**	**85.64+**
noise173	89.19	**90.67+++**	88.86	88.99	88.10–	**89.42**	**89.60**	**89.55**
noise174	75.93	**82.04+++**	**78.34+**	75.62	**77.15**	**77.61**	**81.63+++**	**77.07**
noise177	90.28	**92.95+++**	89.82-	90.2	**90.32**	89.02–	89.87	**91.10+++**
noise180	66.61	**70.02+++**	**68.12+++**	**66.63**	**67.35**	**67.94+**	65.96–	**68.86+++**
noise184	81.03	**85.12+++**	**82.29+++**	80.78	**82.60+++**	**82.38+++**	**82.72+++**	**82.40+++**
noise185	87.59	**88.96+++**	**87.77**	**87.87++**	87.5	87.39	**88.12+**	**87.89**
noise188	80.45	**83.58+++**	**82.26+++**	**80.51**	**82.17+++**	**82.51+++**	80.42	**83.03+++**
ดีกว่าชุดตั้งต้น		**59**	14	6	11	14	18	**27**
แย่กว่าชุดตั้งต้น		0	25	13	19	23	14	9

ตารางที่ ข.6: เปรียบเทียบประสิทธิภาพของตัวจำแนกกึ่งมีผู้สอนด้วยเพื่อนบ้านใกล้ที่สุดสามตัว ระหว่างการใช้ตัวอย่างมีป้ายกำกับเดิม กับเมื่อเพิ่มป้ายกำกับในกลุ่มไม่มีป้ายกำกับด้วยตัวจำแนกที่สร้างจากขั้นตอนวิธีต่าง ๆ บนชุดข้อมูลการลดสิ่งรบกวนในภาพเอกสารภาษาไทย

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวนรีพร พิรุฬห์ทรัพย์ เกิดเมื่อวันที่ 11 กันยายน พ.ศ.2527 ที่กรุงเทพมหานคร สำเร็จการศึกษาระดับบัณฑิตศึกษาเมื่อปี พ.ศ. 2549 ในสาขาวิชาศาสตร์คอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยธรรมศาสตร์ ได้รับวิทยาศาสตรบัณฑิต เกียรตินิยมอันดับสอง ด้วยคะแนนสูงสุดในสาขาวิชา

หัวข้อโครงงานระดับบัณฑิตศึกษาชื่อหัวข้อ "หุ่นยนต์เพื่อนเล่นเด็ก" ได้รับรางวัลขวัญใจมหาชน จากสำนักงานส่งเสริมอุตสาหกรรมซอฟต์แวร์แห่งชาติ ในงาน Thailand Animation and Multimedia (TAM) 2006 และผ่านเข้ารอบสุดท้ายในโครงการแข่งขันพัฒนาโปรแกรมคอมพิวเตอร์แห่งประเทศไทย (NCSEC) ครั้งที่ 8

สำเร็จการศึกษาระดับมหาบัณฑิตเมื่อปี พ.ศ. 2553 ในสาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ชื่อหัวข้อวิทยานิพนธ์ "การลดสิ่งรบกวนในไทยโอซีอาร์โดยการเรียนรู้แบบกึ่งสอน" โดยเผยแพร่ผลงานงานวิจัยในงานประชุมวิชาการที่ The 2nd international conference on computer engineering and technology เมืองเฉิงตู ประเทศจีน และ The 7th international joint conference on computer science and software engineering จังหวัดกรุงเทพมหานคร ประเทศไทย

และเข้าศึกษาระดับดุษฎีบัณฑิตสาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย โดยงานวิจัยมุ่งเน้นศึกษาในหัวข้อ ปัญญาประดิษฐ์ (Artificial intelligence) การเรียนรู้ของเครื่อง (Machine learning) การเรียนรู้แบบกึ่งมีผู้สอน (Semi-supervised learning) ปัญหาการจำแนก (Classification problems) และการวิเคราะห์เอกสารภาษาไทย

ประสบการณ์การทำงาน ได้แก่ ตำแหน่งวิศวกรซอฟต์แวร์ที่บริษัทรอยเตอร์ ซอฟต์แวร์ไทยแลนด์ ตั้งแต่ พ.ศ. 2549 ถึง 2551 ตำแหน่งนักพัฒนาอาวุโสที่บริษัท ซอฟสเคปเอเชีย พ.ศ. 2552 และตำแหน่งอาจารย์ประจำ สาขาวิชาคอมพิวเตอร์เพื่อการสื่อสาร วิทยาลัยนวัตกรรมสื่อสารสังคม มหาวิทยาลัยศรีนครินทรวิโรฒ ตั้งแต่ พ.ศ. 2553 ถึง พ.ศ. 2554