การรวมกลุ่มผสมระหว่างการเรียนรู้ของเครื่องและการเรียนรู้เชิงสถิติโดยใช้การบูสต์ด้วยค่าความเชื่อมั่น

A HYBRID ENSEMBLE OF MACHINE AND STATISTICAL LEARNING USING CONFIDENCE-BASED BOOSTING

ณัฐวุฒิ ชัยรัตนทรงพร
NATTAWUT CHAIRATANASONGPORN

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต
ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2558

KMITL-2015-SC-M-002-065

การรวมกลุ่มผสมระหว่างการเรียนรู้ของเครื่องและการเรียนรู้เชิงสถิติโดยใช้การบูสต์ด้วยค่าความเชื่อมั่น

A HYBRID ENSEMBLE OF MACHINE AND STATISTICAL LEARNING USING CONFIDENCE-BASED BOOSTING

ณัฐวุฒิ ชัยรัตนทรงพร

NATTAWUT CHAIRATANASONGPORN

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตร์มหาบัณฑิต

ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2558

KMITL-2015-SC-M-002-065

A HYBRID ENSEMBLE OF MACHINE AND STATISTICAL LEARNING USING CONFIDENCE-BASED BOOSTING

NATTAWUT CHAIRATANASONGPORN

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SCIENCE

FACULTY OF SCIENCE

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2015

KMITL-2015-SC-M-002-065

คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์ "การรวมกลุ่มผสมระหว่างการเรียนรู้ของเครื่องและการเรียนรู้เชิงสถิติโดยใช้การบูสต์ด้วยค่าความเชื่อมั่น"
"A HYBRID ENSEMBLE OF MACHINE AND STATISTICAL LEARNING USING CONFIDENCE-BASED BOOSTING"

ชื่อนักศึกษา นายณัฐวุฒิ ชัยรัตนทรงพร
รหัสประจำตัว 53650851
ปริญญา วิทยาศาสตรมหาบัณฑิต (สาขาวิทยาการคอมพิวเตอร์)
ภาควิชา วิทยาการคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์ ดร.สายชล ใจเย็น
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม (ถ้ามี) -----

คณะกรรมการสอบวิทยานิพนธ์	ลายมือชื่อ
ผศ.ดร.ศรัณย์ อินทโกสุม ประธานกรรมการ	
ผศ.ดร.อนันตพร หรรษคุณาตัย อาจารย์บัณฑิตประจำ (ในสาขาวิชาที่เกี่ยวข้อง)	
ผศ.ดร.ศุภกานต์ พิมลธเรศ ผู้ทรงคุณวุฒิจากภายนอกสถาบันฯ	
ดร.สายชล ใจเย็น อาจารย์ที่ปรึกษาวิทยานิพนธ์	

วัน/ เดือน/ ปี ที่สอบ 1 ธันวาคม พ.ศ. 2558 เวลา 13.00 - 15.00 น.
สถานที่สอบ ณ ห้อง 306 ตึกปฏิบัติการหลังใหม่

คณะวิทยาศาสตร์รับรองแล้ว

(รองศาสตราจารย์ ดร.ดุษณี ธนะบริพัฒน์)
คณบดีคณะวิทยาศาสตร์
วันที่ ๗ เดือน ธันวาคม พ.ศ. ๕๘

หัวข้อวิทยานิพนธ์	การรวมกลุ่มผสมระหว่างการเรียนรู้ของเครื่องและการเรียนรู้เชิงสถิติโดยใช้การบูสต์ด้วยค่าความเชื่อมั่น
ชื่อนักศึกษา	ณัฐวุฒิ ชัยรัตนทรงพร
รหัสประจำตัว	53650851
ปริญญา	วิทยาศาสตรมหาบัณฑิต สาขาวิทยาการคอมพิวเตอร์
ภาควิชา	วิทยาการคอมพิวเตอร์
พ.ศ.	2558
อาจารย์ที่ปรึกษาวิทยานิพนธ์	อาจารย์ ดร.สายชล ใจเย็น

บทคัดย่อ

ในปัจจุบัน ปัญหาการจำแนกประเภทข้อมูลนับเป็นปัญหาที่มีความท้าทายที่เพิ่มมากขึ้น เนื่องจากข้อมูลในปัจจุบันมีอยู่ด้วยกันมากมายหลากหลายรูปแบบ โดยที่ชุดข้อมูลเหล่านี้บางชุดข้อมูลอาจเหมาะกับการจำแนกประเภทด้วยการใช้เทคนิคการเรียนรู้ของเครื่อง ในขณะที่บางชุดข้อมูลอาจเหมาะกับการจำแนกประเภทด้วยเทคนิคเชิงสถิติ จึงเป็นที่มาของงานวิจัยชิ้นนี้ที่นำเสนอวิธีการรวมกลุ่มผสมระหว่างการเรียนรู้ของเครื่องและการเรียนรู้เชิงสถิติโดยใช้การบูสต์ด้วยค่าความเชื่อมั่น โดยวิธีการดังกล่าวจะใช้เทคนิคการเรียนรู้พื้นฐานที่แตกต่างกันจำนวนหนึ่ง ซึ่งสามารถให้ผลลัพธ์ในการจำแนกประเภทข้อมูลจากชุดข้อมูลต่าง ๆ ได้ดี นอกจากนี้การรวมเอาเทคนิคการคำนวณค่าความเชื่อมั่นมาช่วยเสริมกับเทคนิคการบูสต์ด้วยแล้วก็ยิ่งช่วยเพิ่มประสิทธิภาพในการจำแนกประเภทให้ได้ผลลัพธ์ที่ดียิ่งขึ้นอีกด้วย สำหรับการวัดประสิทธิภาพของงานวิจัยนี้จะทำการวัดประสิทธิภาพเปรียบเทียบกับวิธีการเรียนรู้แบบรวมกลุ่มเอดาบูสต์เอ็มวันที่ใช้เทคนิคการเรียนรู้ต้นไม้ตัดสินใจและเปรียบเทียบกับการเรียนรู้แบบรวมกลุ่มเอ็มอาร์เอ็น โดยจะใช้ชุดข้อมูลจากเว็บไซต์ยูซีไอในการทดสอบ จากผลการทดลองจะเห็นว่าวิธีการที่นำเสนอนี้สามารถเพิ่มประสิทธิภาพความถูกต้องในการจำแนกประเภทข้อมูลที่เป็นทั้งแบบสองคลาสและแบบหลายคลาสได้

คำสำคัญ : การเรียนรู้ของเครื่อง การเรียนรู้แบบรวมกลุ่ม การวิเคราะห์การถดถอยแบบโลจิสติก การวิเคราะห์เชิงเส้นแบบดิสคริมิแนนท์ โครงข่ายประสาทเทียมเพอร์เซ็ปตรอนหลายชั้น ต้นไม้ตัดสินใจ เอดาบูสต์

Thesis Title	A Hybrid Ensemble of Machine and Statistical Learning Using Confidence-Based Boosting
Student Name	Nattawut Chairatanasongporn
Student ID	53650851
Degree	Master of Science Computer
Department	Computer Science
Year	2015
Thesis Advisor	Dr. Saichon Jaiyen

Abstract

Nowadays, the classification problems have become more challenging due to the various types of data set. Some data are appropriated for machine learning techniques and some data are appropriated for statistical leaning techniques. This work proposes a new hybrid ensemble of machine and statistical learning models using confidence-based boosting. The proposed method which uses variants of based classifiers can solve classification problems in variant data set. Moreover, combining the confidence value to the current boosting method can improve the performance of classification. The performance of proposed method is compared to the ensemble of decision trees and MRN created by Adaboost.M1 on data sets from UCI. The experimental results show that the proposed method can improve the accuracy in both binary and multiclass classification problems.

Keywords : Machine Learning, Ensemble Learning, Logistic Regression, Linear Discriminant Analysis, Multi-layer Perceptron, Decision Tree, Adaboost

กิตติกรรมประกาศ

การทำวิทยานิพนธ์นี้จะไม่สามารถสำเร็จลุล่วงไปได้หากปราศจากความอนุเคราะห์และคำแนะนำจากอาจารย์ที่ปรึกษาปริญญานิพนธ์ ดร.สายชล ใจเย็น ซึ่งได้สละเวลาให้คำปรึกษา แนะนำ แก้ไข รวมทั้งให้การสนับสนุนทางด้านคู่มือและตำราในการศึกษาเพื่อใช้ในการทำวิทยานิพนธ์เล่มนี้ รวมถึงข้อแนะนำและแก้ไขจากคณะกรรมการสอบหัวข้อและโครงร่างวิทยานิพนธ์ ผศ.ดร.ศรัณย์ อินทโกสุม ผศ.ดร.อนันตพร หรรษคุณาฒัย และผศ.ดร.ศุภกานต์ พิมลธเรศ จนทำให้วิทยานิพนธ์เล่มนี้สามารถสำเร็จลุล่วงไปได้ ซึ่งในระหว่างการทำวิจัยและวิทยานิพนธ์เล่มนี้นั้น ผู้จัดทำได้พบกับอุปสรรคหลายประการ ไม่ว่าจะเป็นแนวความคิดที่จะใช้ในการทำวิจัย เทคนิคในการเขียนโปรแกรม การหาและแก้ไขจุดบกพร่องของการทดลอง และปัญหาอื่น ๆ อีกมากมาย ซึ่งข้าพเจ้าจะไม่สามารถผ่านอุปสรรคเหล่านั้นไปได้หากปราศจากผู้มีพระคุณตามที่ได้กล่าวมา ดังนั้นข้าพเจ้ารู้สึกเป็นเกียรติและซาบซึ้งในความกรุณาของทุกท่านเป็นอย่างยิ่ง จึงใคร่ขอกราบขอบพระคุณมาไว้ ณ ที่นี้

ขอขอบคุณนายยศภัทร เรืองไพศาล นางสาวภูมิธารา เรืองทอง นายนิพัทธ์ คล้ายโพธิ์ นายอนุสรณ์ เจริญนาน นายทักษ์ดนัย สุวรรณ นางสาวบุญหทัย เครือแก้ว นายมงคล ทองไกรแก้ว นายจนัสธา ศรีสรวล นายวริษฐ์ เพ็ชรประสิทธิ์ และเหล่านักศึกษาระดับบัณฑิตศึกษาที่ภาควิชาวิทยาการคอมพิวเตอร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบังทุกคนที่คอยให้ความช่วยเหลือตลอดจนอำนวยความสะดวกในการศึกษาเล่าเรียนและการทำวิทยานิพนธ์เล่มนี้เป็นอย่างดี

สุดท้ายนี้ขอขอบพระคุณบิดามารดาและสมาชิกในครอบครัวทุกคนที่เป็นกำลังใจให้แก่ข้าพเจ้าด้วยดีเสมอมา และขอขอบคุณบุคคลอื่น ๆ ที่มิได้กล่าวถึงมา ณ ที่นี้ด้วย ที่ซึ่งได้ให้กำลังใจและการสนับสนุนการจัดทำตลอดจนให้ข้อเสนอแนะที่เป็นประโยชน์ต่อการดำเนินงานในครั้งนี้จนประสบผลสำเร็จไปได้ด้วยดี

ณัฐวุฒิ ชัยรัตนทรงพร

สารบัญ

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญรูป

รูปที่ **หน้า**

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

โลกในยุคของข้อมูลข่าวสารและการติดต่อสื่อสารที่มีข้อมูลอยู่เป็นจำนวนมากปรากฏบนเว็บไซต์ จดหมายอิเล็กทรอนิกส์ (E-mail) สื่อสังคมออนไลน์ (Social media) หรือสื่ออื่น ๆ ซึ่งเป็นได้ทั้ง บทความต่าง ๆ กระทู้ถาม-ตอบ ข้อมูลเชิงสถิติต่าง ๆ โดยข้อมูลเหล่านี้จะอยู่ในรูปของข้อมูลแบบต่าง ๆ เช่น ตัวอักษร ข้อความ เสียง ภาพนิ่ง วีดิทัศน์ เป็นต้น อีกทั้งยังมีข้อมูลที่เกิดจากการทำงานของอุปกรณ์ที่มีเซ็นเซอร์ต่าง ๆ เช่น โทรศัพท์มือถือ ชุดอุปกรณ์ตรวจสมรรถภาพร่างกาย หรือแม้กระทั่งอุปกรณ์เครื่องใช้ไฟฟ้าภายในบ้าน เป็นต้น ทำให้ข้อมูลในปัจจุบันมีความหลากหลาย ซับซ้อน และมีขนาดที่มากขึ้นเรื่อย ๆ จนเกิดเป็นปัญหาที่เรียกว่า บิ๊กเดต้า (Big Data) ในขณะเดียวกันมนุษย์ก็ต้องการที่จะนำข้อมูลที่มีความซับซ้อนและมีจำนวนมากเหล่านี้มาใช้งานให้เกิดประโยชน์ โดยอาจเป็นการจำแนกประเภทข้อมูลออกเป็นกลุ่มต่าง ๆ เช่น การจำแนกประเภทผู้บริโภคออกเป็นกลุ่มต่าง ๆ เพื่อใช้สำหรับประกอบการตัดสินใจดำเนินกลยุทธ์ทางการตลาดให้ตรงกับกลุ่มลูกค้ากลุ่มนั้น ๆ มากที่สุด หรือการจำแนกประเภทระดับความคิดเห็นเชิงบวกหรือเชิงลบที่มีต่อกระทู้ต่าง ๆ โดยการอ่านความคิดเห็นที่ตอบกับกระทู้นั้น ๆ หรือกระทั่งการจำแนกประเภทอิริยาบถของมนุษย์โดยการอ่านค่าจากเซ็นเซอร์ที่มีอยู่ในโทรศัพท์มือถือ เป็นต้น ซึ่งวิธีการจำแนกประเภทข้อมูลเหล่านี้มนุษย์อาจให้เครื่องคอมพิวเตอร์ช่วยในการจำแนกประเภทข้อมูลโดยการป้อนชุดข้อมูลให้คอมพิวเตอร์เรียนรู้และค้นหารูปแบบของข้อมูลเพื่อใช้ในการจำแนกประเภทข้อมูลออกมา แม้ปัจจุบันจะมีวิธีการจำแนกประเภทข้อมูลอยู่หลายวิธีด้วยกัน แต่โดยส่วนมากแล้ววิธีการแต่ละวิธีนั้นจะสามารถใช้จำแนกประเภทข้อมูลได้อย่างมีประสิทธิภาพที่แม่นยำได้มักขึ้นอยู่กับลักษณะ ของข้อมูลที่ถูกนำมาใช้สร้างแบบจำลองการจำแนกประเภทข้อมูลด้วย และด้วยเหตุที่ข้อมูลมีความหลากหลายและมีความซับซ้อนที่เพิ่มขึ้นเรื่อย ๆ การจะใช้วิธีการจำแนกประเภทข้อมูลวิธีการใดเพียงวิธีเดียวนั้นอาจไม่เพียงพอที่จะให้ประสิทธิภาพในการจำแนกประเภทข้อมูลให้ผลเป็นที่น่าพอใจได้ ทั้งในเรื่องของเวลาที่ใช้ในการสร้างแบบจำลองเพื่อจำแนกประเภทข้อมูลและความถูกต้องเที่ยงตรงของผลลัพธ์

ด้วยเหตุนี้วิทยานิพนธ์นี้จึงนำเสนอวิธีการจำแนกประเภทข้อมูลแบบกลุ่มผสมระหว่างกลุ่มเทคนิคการเรียนรู้ของเครื่องและกลุ่มเทคนิคการเรียนรู้ด้วยวิธีเชิงสถิติ ซึ่งจะประกอบไปด้วยเทคนิคหลายเทคนิคเพื่อช่วยให้สามารถรองรับข้อมูลที่มีความหลายหลายและมีความซับซ้อนได้มากขึ้น โดยการเรียนรู้ของเครื่องจะใช้เทคนิคโครงข่ายประสาทเทียมแบบเพอร์เซ็ปตรอนหลายชั้น (Multilayer Perceptron: MLP) และเทคนิคการเรียนรู้ต้นไม้ตัดสินใจ (Decision Tree) ส่วนการเรียนรู้ด้วยวิธีเชิง

สถิติจะใช้เทคนิคการวิเคราะห์เชิงเส้นแบบดิสคริมิแนนท์ (Linear Discriminant Analysis: LDA) และเทคนิคการถดถอยแบบลอจิสติก (Logistic Regression: LR) นอกจากนี้จะมีการคำนวณค่าความเชื่อมั่นเพื่อใช้ช่วยส่งเสริมความแม่นยำในการจำแนกประเภทข้อมูลของแต่ละเทคนิคด้วย

1.2 วัตถุประสงค์

1) เพื่อนำเสนอและพัฒนาวิธีการจำแนกประเภทข้อมูลแบบรวมกลุ่มให้มีความสามารถในการรองรับการจำแนกประเภทข้อมูลได้หลากหลายมากขึ้นด้วยการนำเอาเทคนิคการเรียนรู้ของเครื่องและเทคนิคการจำแนกประเภทเชิงสถิติมาใช้ร่วมกัน
2) เพื่อศึกษาและเปรียบเทียบประสิทธิภาพวิธีการจำแนกประเภทข้อมูลแบบรวมกลุ่ม

1.3 สมมติฐานของงานวิจัย

แม้วิธีการในการจำแนกประเภทข้อมูลจะมีอยู่หลากหลายวิธี แต่เนื่องจากข้อมูลในปัจจุบันนั้นมีความหลากหลายและซับซ้อน การใช้วิธีการจำแนกประเภทข้อมูลเพียงวิธีการเดียวอาจไม่เหมาะสมกับชุดข้อมูลใดข้อมูลหนึ่งและทำให้ได้ผลลัพธ์ที่ไม่น่าพอใจ หากสามารถพัฒนาหรือหาวิธีในการเพิ่มขอบเขตและประสิทธิภาพในการรองรับข้อมูลได้กว้างมากขึ้น โดยการใช้วิธีการการจำแนกประเภทด้วยเทคนิคการเรียนรู้ของเครื่องผสมผสานกับเทคนิคการเรียนรู้เชิงสถิติ อาจทำให้ได้ผลลัพธ์ที่ดีขึ้น

1.4 ขอบเขตการวิจัย

1) ข้อมูลที่ใช้ทดสอบในงานวิจัยนี้จะเลือกชุดข้อมูลที่มีความหลากหลาย ได้แก่ ข้อมูลจากเซ็นเซอร์ในโทรศัพท์ ข้อมูลจากการเก็บสถิติ ข้อมูลที่สกัดมาจากข้อมูลภาพ เป็นต้น โดยข้อมูลจะมีคลาสคำตอบทั้งแบบสองมิติ (Binary Class) และหลายมิติ (Multi Class)
2) งานวิจัยนี้มุ่งเน้นที่การประยุกต์การเรียนรู้แบบรวมกลุ่ม (Ensemble Learning) ซึ่งเป็นการผสมผสานกันระหว่างกลุ่มเทคนิคการเรียนรู้ของเครื่องและกลุ่มเทคนิคการเรียนรู้เชิงสถิติ โดยเทคนิคในกลุ่มการเรียนรู้ของเครื่องจะใช้เทคนิคโครงข่ายประสาทเทียมแบบเพอร์เซ็ปตรอนหลายชั้นและเทคนิคการเรียนรู้ต้นไม้ตัดสินใจ ส่วนการเรียนรู้ด้วยวิธีเชิงสถิติจะใช้เทคนิคการวิเคราะห์เชิงเส้นแบบดิสคริมิแนนท์และเทคนิคการวิเคราะห์การถดถอยแบบลอจิสติก
3) การศึกษาเปรียบเทียบประสิทธิภาพของวิธีการจำแนกประเภทข้อมูลแบบรวมกลุ่มจะเป็นการศึกษาเปรียบเทียบด้วยวิธีการจำแนกประเภทแบบรวมกลุ่มผสมระหว่างการเรียนรู้ของเครื่องและการเรียนรู้เชิงสถิติโดยใช้การบูสต์ด้วยค่าความเชื่อมั่นกับ

วิธีการจำแนกประเภทแบบรวมกลุ่มเอดาบูสต์เอ็มวัน (Adaboost.M1) และวิธีการจำแนกประเภทแบบรวมกลุ่มเอ็มอาร์เอ็น (MRN)

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1) สามารถประยุกต์วิธีการเรียนรู้แบบรวมกลุ่มเพื่อใช้แก้ปัญหาการจำแนกประเภทข้อมูลให้มีความสามารถรองรับข้อมูลซึ่งมีที่มาของชุดข้อมูลได้กว้างหรือหลากหลายมากขึ้น
2) สามารถนำไปประยุกต์ใช้ได้ในหลายอุตสาหกรรม เช่น ด้านธุรกิจ ด้านการแพทย์ ด้านความบันเทิง เป็นต้น

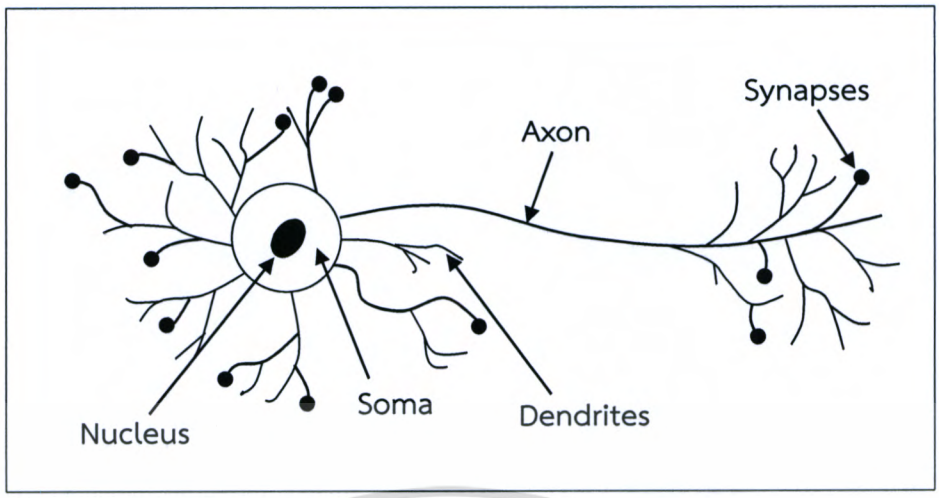
บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีและงานวิจัยต่าง ๆ ที่เกี่ยวข้องกับการวิจัยประกอบไปด้วย โครงข่ายประสาทเทียม (Artificial Neural Network) การเรียนรู้ต้นไม้ตัดสินใจ (Decision Tree) การวิเคราะห์เชิงเส้นแบบดิสคริมิแนนท์ (Linear Discriminant Analysis) การวิเคราะห์การถดถอยโลจิสติก (Logistic Regression Analysis) การเรียนรู้แบบรวมกลุ่ม (Ensemble Learning) และงานวิจัยที่เกี่ยวข้อง

2.1 โครงข่ายประสาทเทียม (Artificial Neural Network)

โครงข่ายประสาทเทียม (Artificial Neural Network) เป็นศาสตร์แขนงหนึ่งด้านปัญญาประดิษฐ์ (Artificial Intelligence : AI) ที่สามารถนำไปประยุกต์ใช้กับงานหลายด้านได้อย่างมีประสิทธิภาพ เช่น การจำแนกประเภท การจดจำรูปแบบ เป็นต้น โดยใช้โมเดลการคำนวณทางคณิตศาสตร์จำลองการทำงานให้คล้ายกับการทำงานของระบบประสาทในสมองมนุษย์ เพื่อสร้างเครื่องมือซึ่งมีความสามารถในการเรียนรู้การจดจำรูปแบบ (Pattern Recognition) และการสกัดความรู้ใหม่ (Knowledge Extraction) เช่นเดียวกับความสามารถที่มีในสมองมนุษย์

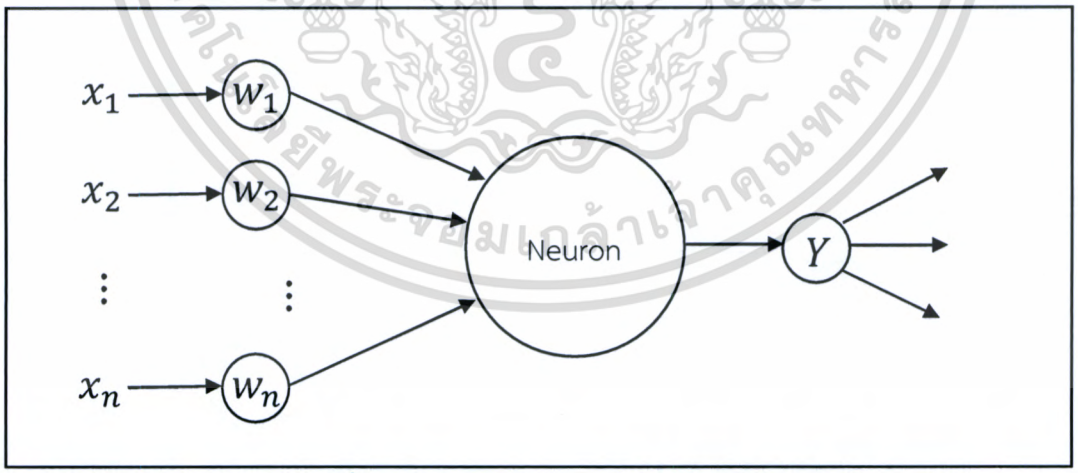
แบบจำลองเชิงคำนวณของเทคนิคนี้มีแนวคิดเริ่มต้นมาจากการศึกษาโครงข่ายไฟฟ้าชีวภาพ (Bioelectric Network) ในสมองมนุษย์ ซึ่งในสมองประกอบไปด้วยเซลล์ประสาท หรือ "นิวรอน" (Neuron) และ "จุดประสานประสาท" (Synapse) เป็นจำนวนมาก โดยแต่ละเซลล์จะประกอบด้วยปลายในการรับกระแสประสาท เรียกว่า "เดนไดรท์" (Dendrite) ซึ่งเป็นส่วนข้อมูลนำเข้า (Input) และปลายในการส่งกระแสประสาทเรียกว่า "แอคซอน" (Axon) ซึ่งเป็นเหมือนส่วนส่งออก (Output) ของเซลล์ประสาท โดยที่เซลล์เหล่านี้จะทำงานด้วยการถ่ายเทสารประกอบทางเคมี เมื่อมีการกระตุ้นด้วยสิ่งเร้าจากภายนอกหรือจากการถูกกระตุ้นด้วยเซลล์ด้วยกัน กระแสประสาทจะวิ่งผ่านเดนไดรท์เข้าสู่นิวเคลียส (Nucleus) ซึ่งจะเป็นตัวตัดสินว่าต้องกระตุ้นเซลล์อื่น ๆ ต่อหรือไม่ ถ้ากระแสประสาทแรงพอ นิวเคลียสก็จะกระตุ้นเซลล์อื่น ๆ ต่อไปผ่านทางแอคซอนของเซลล์นั้น ๆ ดังรูปที่ 2.1



รูปที่ 2.1 ภาพร่างเซลล์ประสาทของมนุษย์

เนื่องจากในการทำงานของระบบประสาทของมนุษย์นั้นเกิดขึ้นจากปฏิกิริยาทางเคมี แต่ในการจำลองเพื่อให้สามารถใช้งานในเครื่องคอมพิวเตอร์ได้นั้นจะเป็นการใช้หลักการทางคณิตศาสตร์มาช่วยสร้างแบบจำลองการทำงานของเซลล์ประสาท

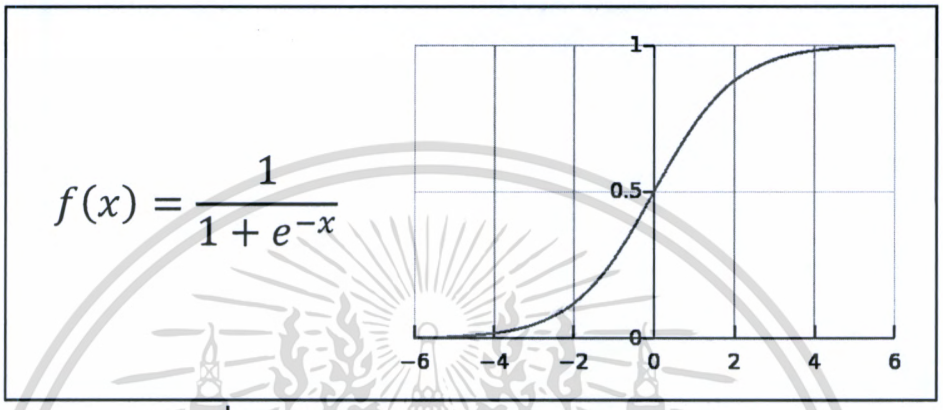
หลักการทำงานของโครงข่ายประสาทเทียมซึ่งจำลองมาจากการทำงานของระบบประสาทของมนุษย์จะมีส่วนประกอบต่าง ๆ ที่เลียนแบบมา ซึ่งประกอบไปด้วย ส่วนข้อมูลนำเข้า (Input) ที่เปรียบเสมือนเดรนไดรท์ในระบบประสาทของมนุษย์ และยังมีส่วนข้อมูลส่งออก (Output) ที่เปรียบได้กับแอคซอนดังแสดงในรูปที่ 2.2



รูปที่ 2.2 รูปแบบประสาทเทียมในคอมพิวเตอร์

จากรูปที่ 2.2 หากมีข้อมูลนำเข้า x ซึ่งมีจำนวนเท่ากับจำนวนคุณลักษณะ (Attribute) ที่มีอยู่ n จำนวน โดยข้อมูลนำเข้าแต่ละข้อมูลจะมีค่าน้ำหนักของตัวเอง คือ w ซึ่งเปรียบได้กับไซแนปส์ (Synapse) โดยเมื่อมีข้อมูลนำเข้าและค่าน้ำหนักแล้ว ค่าทั้งสองค่าจะถูกนำมาคูณกันและส่งไปที่

นิวรอนซึ่งจะเป็นส่วนที่รวบรวมผลคูณของข้อมูลนำเข้ากับค่าน้ำหนักทั้งหมดเพื่อทำการจำแนกประเภท จากนั้นจะนำผลลัพธ์เข้าสู่ฟังก์ชันการกระตุ้น (Activation function) ซึ่งเป็นฟังก์ชันที่ทำการบีบอัดค่าเพื่อไม่ให้ได้ค่าข้อมูลที่กว้างเกินไป เช่น การนำผลลัพธ์เข้าฟังก์ชันซิกมอยด์ (Sigmoid function) ซึ่งเป็นฟังก์ชันที่ทำการปรับค่าในช่วงที่กว้างตั้งแต่ค่าที่ติดลบอนันต์จนถึงค่าบวกอนันต์ให้อยู่ในช่วงของค่าที่ต้องการดังรูปที่ 2.3



รูปที่ 2.3 ฟังก์ชันซิกมอยด์ (Sigmoid function)

การเปรียบเทียบระหว่างส่วนประกอบต่าง ๆ ระหว่างโครงข่ายระบบประสาทของสิ่งมีชีวิตกับโครงข่ายประสาทเทียมสามารถแสดงได้ดังตารางที่ 2.1

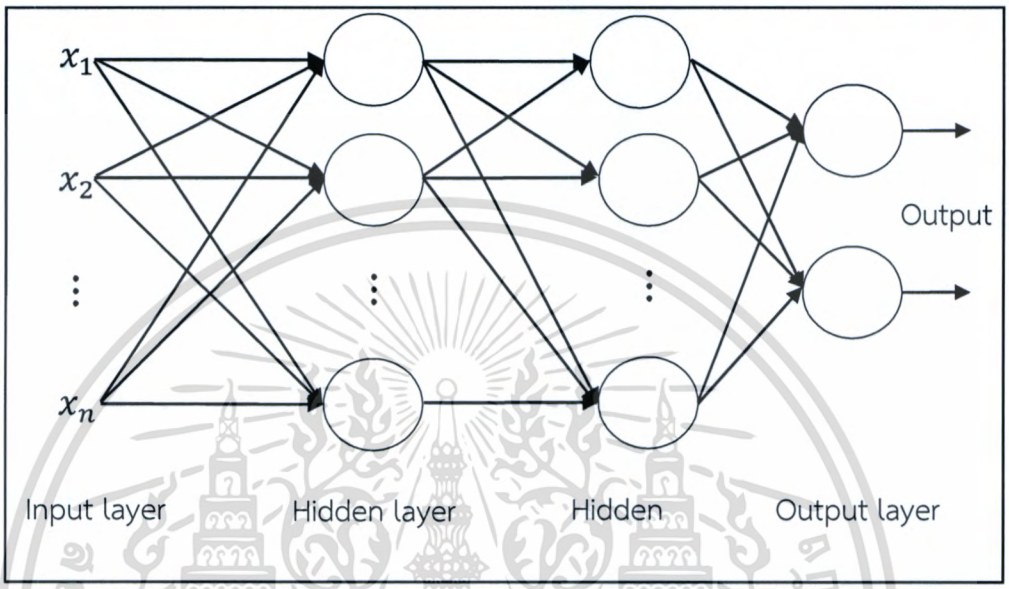
ตารางที่ 2.1 การเปรียบเทียบส่วนประกอบระหว่างโครงข่ายประสาทของสิ่งมีชีวิตและโครงข่ายประสาทเทียม

Biological neural network	Artificial neural network
Soma	Neuron
Dendrite	Input
Axon	Output
Synapse	Weight

2.1.1 โครงข่ายประสาทเทียมเพอร์เซ็ปตรอนหลายชั้น (Multi-layer Perceptron)

การสร้างโครงข่ายประสาทเทียมอย่างง่ายที่สุด คือ การสร้างโครงข่ายประสาทเทียมที่มีส่วนประกอบเพียงชั้นข้อมูลนำเข้า (Input layer) และชั้นข้อมูลส่งออก (Output layer) ซึ่งมีโครงสร้างที่ไม่ซับซ้อนหรือกล่าวได้ว่าเป็นเพอร์เซ็ปตรอนชั้นเดียว (Single-layer perceptron) แต่การสร้างเพอร์เซ็ปตรอนแบบนี้ไม่เหมาะสำหรับใช้ในการจำแนกประเภทข้อมูลที่มีความซับซ้อนที่มาก

ได้ ดังนั้นจึงมีการสร้างสถาปัตยกรรมเพอร์เซ็ปตรอนแบบหลายชั้น (Multi-layer Perceptron) ขึ้นมา โดยให้มีชั้นเพิ่มขึ้นมาอีกเรียกว่า ชั้นซ่อน (Hidden layer) โดยที่ชั้นซ่อนนี้อาจมีเพียงชั้นเดียวหรือหลายชั้นก็ได้ ซึ่งการส่งผ่านข้อมูลจะกระทำด้วยการส่งข้อมูลผ่านชั้นข้อมูลนำเข้าไปยังชั้นซ่อนโดยที่ในชั้นซ่อนนั้นก็ทำการส่งข้อมูลต่อไปยังชั้นซ่อนอื่น ๆ จนกระทั่งถึงชั้นข้อมูลส่งออกดังรูปที่ 2.4



รูปที่ 2.4 เพอร์เซ็ปตรอนแบบหลายชั้น

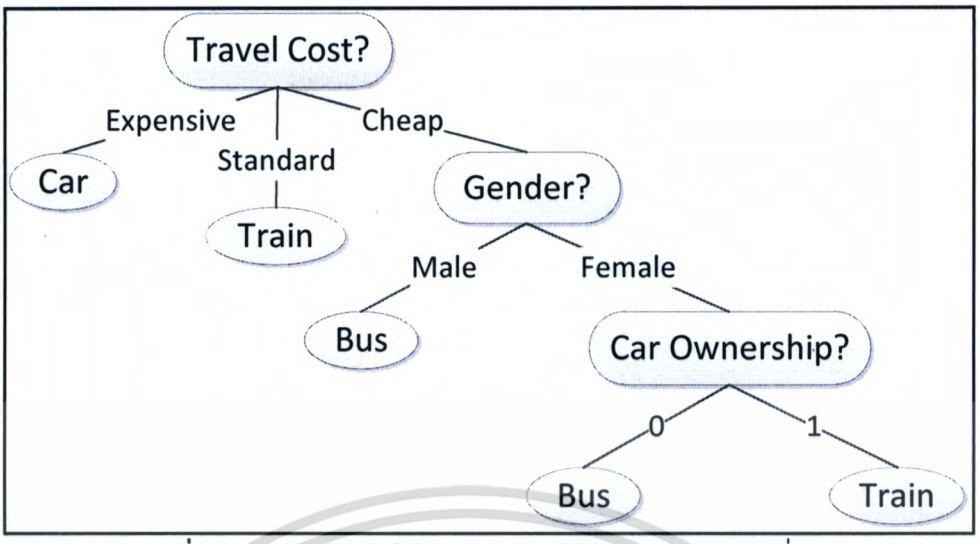
นอกจากนี้เพอร์เซ็ปตรอนแบบหลายชั้นยังใช้ขั้นตอนการเรียนรู้แบบแพร่กระจายย้อนกลับ (Back-propagation) โดยแบ่งเป็น 2 ส่วนย่อย ได้แก่ การส่งผ่านไปข้างหน้า (Forward pass) และการส่งผ่านย้อนกลับ (Backward pass) ซึ่งเป็นการทำให้โครงข่ายประสาทเทียมสามารถทำการปรับตัวเองเพื่อให้เข้ากับข้อมูลต่าง ๆ ได้ โดยในขั้นตอนการส่งผ่านไปข้างหน้าโครงข่ายจะทำการสุ่มค่าน้ำหนักและค่าไบแอส (Bias) ขึ้นมา เมื่อนิวรอนเริ่มรับข้อมูลสำหรับการสอน (Training set) ก็จะทำการคำนวณจนถึงชั้นข้อมูลส่งออก จากนั้นจะนำค่าผลลัพธ์ที่ได้มาเปรียบเทียบกับผลลัพธ์ที่คาดหวัง แล้วทำการคำนวณหาค่าความผิดพลาด ซึ่งค่าความผิดพลาดที่ได้นี้จะถูกส่งย้อนกลับเข้าสู่ชั้นต่าง ๆ ซึ่งเป็นขั้นตอนการส่งผ่านย้อนกลับเพื่อปรับปรุงค่าน้ำหนักต่อไป

2.2 การเรียนรู้ต้นไม้ตัดสินใจ (Decision Tree)

การเรียนรู้ต้นไม้ตัดสินใจ (Decision tree) คือ โมเดลที่มีโครงสร้างเป็นลำดับชั้นในลักษณะคล้ายต้นไม้กลับหัวโดยจะมีราก (Root) อยู่บนสุดและแตกแขนงออกเป็นกิ่งและใบ (leaf) ลงมาด้านล่าง เทคนิคนี้จัดเป็นเทคนิคการเรียนรู้แบบมีผู้สอน (Supervised Learning) ที่สามารถใช้จำแนกประเภทข้อมูลได้โดยการกำหนดคำถามขึ้นมาเพื่อตัดสินว่าข้อมูลที่ได้มาจะจัดอยู่ในประเภทข้อมูล (Class) อะไร ซึ่งข้อมูลที่สามารถนำมาใช้งานกับเทคนิคนี้สามารถมี คุณลักษณะ (Attribute) ของข้อมูลได้หลายชนิดตั้งแต่ ข้อมูลแบบไบนารี (Binary) ข้อมูลแบบนามบัญญัติ (Norminal) ข้อมูลแบบอันดับ (Ordinal) จนถึงข้อมูลเชิงปริมาณ (Quantitative) ในขณะที่ประเภทของข้อมูลนั้นต้องเป็นข้อมูลเชิงคุณภาพ (Qualitative) เท่านั้น รูปที่ 2.5 แสดงตัวอย่างโครงสร้างต้นไม้ตัดสินใจซึ่งใช้ข้อมูลจากตารางที่ 2.2

ตารางที่ 2.2 ตัวอย่างข้อมูลที่ใช้สร้างต้นไม้ตัดสินใจ

Attributes				Classes
Gender	Car ownership	Travel cost	Income level	Transportation mode
Male	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	1	Cheap	Medium	Train
Female	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Male	0	Standard	Medium	Train
Female	1	Standard	Medium	Train
Female	1	Expensive	High	Car
Male	2	Expensive	Medium	Car
Female	2	Expensive	High	Car



รูปที่ 2.5 โครงสร้างต้นไม้ตัดสินใจโดยใช้ข้อมูลจากตารางที่ 2.2

หลักการสร้างต้นไม้ตัดสินใจคือการเลือกคุณลักษณะที่เหมาะสมที่สุดมาสร้างเป็นโหนด (Node) เพื่อถามคำถามสำหรับตัดสินใจในการจำแนกประเภทข้อมูล โดยจะใช้วิธีการคำนวณหาค่าความเหมาะสมของคุณลักษณะแต่ละตัว ซึ่งมีอยู่ด้วยกันหลายวิธีขึ้นอยู่กับต้นไม้ตัดสินใจแต่ละแบบ เช่น ต้นไม้ตัดสินใจแบบไอดีทรี (ID3) จะหาค่าความเหมาะสมของคุณลักษณะได้จากค่าเกนสารสนเทศ (Information Gain) ดังสมการที่ (2.1)

$$Gain(A) = Info(D) - Info_A(D) \tag{2.1}$$

จากสมการที่ (2.1) เราสามารถหาค่าเกนสารสนเทศได้จากการหาผลต่างระหว่างค่าเอนโทรปี (Entropy) กับค่าความรู้ที่คาดหวังจากการจำแนกข้อมูล D ด้วยคุณลักษณะ A ซึ่งหาได้จากสมการที่ (2.2) และสมการที่ (2.4) ตามลำดับ

$$Info(D) = \sum_{i=1}^{m} -p_i \log_2 p_i \tag{2.2}$$

ซึ่งสามารถคำนวณค่า p_i ได้จากสมการ

$$p_i = \frac{|C_{i,D}|}{|D|} \tag{2.3}$$

โดยกำหนดให้

p_i คือ ค่าความน่าจะเป็นของข้อมูล D ที่จะอยู่ในกลุ่ม C_i

m คือ จำนวนกลุ่มทั้งหมด

C_i คือ กลุ่มของข้อมูลในลำดับที่ i

$|C_{i,D}|$ คือ จำนวนข้อมูล D ที่อยู่ในกลุ่ม C_i

$|D|$ คือ จำนวนข้อมูลของข้อมูล D

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \tag{2.4}$$

โดยที่ $|D_j|$ คือ จำนวนของข้อมูล D ที่ค่าคุณลักษณะตัวที่ j ของคุณลักษณะ A

j คือ ค่าของคุณลักษณะตัวที่ j ของคุณลักษณะ A

v คือ จำนวนของค่าที่เป็นไปได้ทั้งหมดของคุณลักษณะ A

สมการที่ (2.2) เป็นสมการหาค่าเอนโทรปี (Entropy) หรือค่าชี้วัดความไม่แน่นอน (Measure of uncertainty) ซึ่งเป็นค่าที่ใช้บอกความสามารถในการจำแนกประเภทข้อมูลเมื่อพิจารณาจากคุณลักษณะแต่ละคุณลักษณะ ยิ่งค่าเอนโทรปีมีค่าสูงมากเท่าใด ความไม่แน่นอนในการแยกประเภทก็ยิ่งมีมากขึ้นเท่านั้น

หรือวิธีการสร้างต้นไม้ตัดสินใจแบบซีโฟร์พอยต์ไฟว์ (C4.5) ที่ถูกออกแบบมาให้ลดความเอนเอียงในการจำแนกประเภทข้อมูลในการสร้างต้นไม้ตัดสินใจแบบไอดีทรี และสามารถใช้ได้กับข้อมูลที่มีค่าที่มีความผิดปกติหรือเสียหายได้ โดยการเพิ่มค่าอัตราส่วนเกน (Gain ratio) ในการตัดสินใจเลือกคุณลักษณะ และเนื่องจากค่าเกนสารสนเทศจะมีความเอนเอียงค่อนข้างสูงกับข้อมูลที่ประกอบด้วยคุณลักษณะที่มีค่าที่เป็นไปได้จำนวนมาก ซึ่งการแก้ไขความเอนเอียงสามารถทำได้ด้วยการปรับค่าเกนสารสนเทศโดยการใช้ค่าสารสนเทศของการแบ่งแยก (Split information) ซึ่งคำนวณได้จากสมการที่ (2.5)

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right) \tag{2.5}$$

ค่าสารสนเทศของการแบ่งแยกที่ได้นี้ เมื่อนำค่าดังกล่าวไปหารกับค่าเกนสารสนเทศก็จะได้อัตราส่วนเกนดังสมการที่ (2.6) และทำการเลือกคุณลักษณะที่มีค่าอัตราส่วนเกนมากที่สุดเป็นโหนดตัดสินใจ

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)} \tag{2.6}$$

2.3 การวิเคราะห์เชิงเส้นแบบดิสคริมิแนนท์ (Linear Discriminant Analysis)

การวิเคราะห์เชิงเส้นแบบดิสคริมิแนนท์ เป็นเทคนิคทางสถิติวิธีหนึ่งที่มักนิยมใช้ในการลดจำนวนมิติ (Dimensionality reduction) หรือใช้ในการจำแนกประเภทของข้อมูล (Classification) โดยการนำข้อมูลที่ประกอบไปด้วยข้อมูลที่เป็นตัวแปรต้นและตัวแปรตาม (คุณลักษณะและประเภทกลุ่มข้อมูล) มาหาสมการแสดงความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตามในรูปสมการเชิงเส้นดังสมการที่ (2.7)

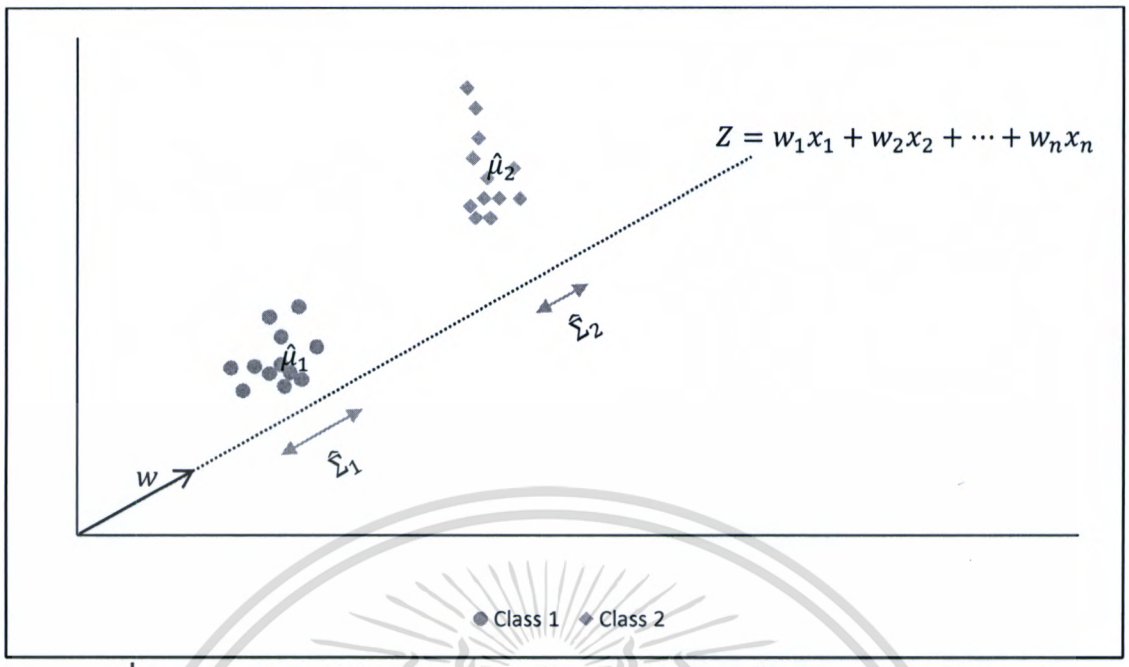
$$Z = w_1x_1 + w_2x_2 + \cdots + w_nx_n \tag{2.7}$$

โดยให้ Z คือ ตัวแปรตาม

x_i คือ ตัวแปรต้นตัวที่ i

w_i คือ สัมประสิทธิ์พจน์ที่ i

สมการที่ (2.7) เรียกว่า ฟังก์ชันดิสคริมิแนนท์ฟิชเชอร์ (Fisher discriminant function) คิดค้นโดย R. A. Fisher ซึ่งมีหลักการในการจำแนกประเภทข้อมูล คือ หากมีข้อมูลกลุ่มตัวอย่างอยู่ 2 กลุ่ม โดยข้อมูลแต่กลุ่มจะมีค่าเฉลี่ย (Mean) และค่าความแปรปรวน (Variance) ของตัวเองอยู่ การทำการจำแนกประเภทโดยการหาเส้นตรงใด ๆ มาแบ่งกลุ่มข้อมูลเพื่อให้ได้ผลลัพธ์ที่ดีที่สุดสามารถทำได้ด้วยการนำข้อมูลดังกล่าวมาฉาย (Project) ลงบนเวกเตอร์ใด ๆ ที่สามารถทำให้ค่าเฉลี่ยของแต่ละกลุ่มมีระยะห่างกันมากที่สุดและมีค่าความแปรปรวนของแต่ละกลุ่มน้อยที่สุด ซึ่งหลักการดังกล่าว หากพิจารณาการฉายข้อมูลลงบนเวกเตอร์ใหม่จะสามารถทำได้ด้วยการคูณกันของเวกเตอร์ตัวแปรต้นกับเวกเตอร์ที่ต้องการฉาย โดยหากให้ x เป็นเวกเตอร์ตัวแปรต้น เวกเตอร์ที่ต้องการฉายก็จะเป็น w ดังในฟังก์ชันดิสคริมิแนนท์ฟิชเชอร์นั่นเอง รูปที่ 2.6 จำลองหลักการฉายข้อมูลลงบนเวกเตอร์ใด ๆ ที่ทำให้ระยะห่างของค่าเฉลี่ยของแต่ละกลุ่มห่างกันมากที่สุดและมีค่าความความแปรปรวนของแต่ละกลุ่มน้อยที่สุด



รูปที่ 2.6 กราฟจำลองวิธีการจำแนกกลุ่มด้วยวิธีการวิเคราะห์เชิงเส้นแบบดิสคริมิแนนท์

สำหรับการจำแนกกลุ่มสามารถทำได้จากสมการที่ (2.8)

$$w^T\left(x - \left(\frac{\mu_1 + \mu_2}{2}\right)\right) > \log\frac{P(C_1)}{P(C_2)} \tag{2.8}$$

โดยที่ w คือ ค่าสัมประสิทธิ์ หรือเวกเตอร์ที่ต้องฉายข้อมูล

x คือ ข้อมูลที่ต้องการจำแนกกลุ่ม

μ คือ ค่าเฉลี่ยของข้อมูลตัวอย่างของกลุ่ม C

$P(C_i)$ คือ ความน่าจะเป็นของกลุ่ม i จากกลุ่มที่มีทั้งหมด

ตัวแปร w สามารถหาได้จากสมการที่ (2.9)

$$w = \Sigma^{-1}(\mu_1 - \mu_2) \tag{2.9}$$

ถ้าสมมติให้ M เป็นเมตริกซ์ขนาด $N \times K$ ค่า Σ จะสามารถหาได้จากสมการที่ (2.10)

$$\hat{\Sigma} = \frac{\sum_{n=1}^{N}\sum_{k=1}^{K} M_{nk}(x_n - \hat{\mu}_k)\,(x_n - \hat{\mu}_k)^T}{N - K} \tag{2.10}$$

โดยกำหนดให้ $M_{nk} = 1$ ถ้า n อยู่ในกลุ่ม k และให้เป็นค่า 0 หากเป็นอย่างอื่น

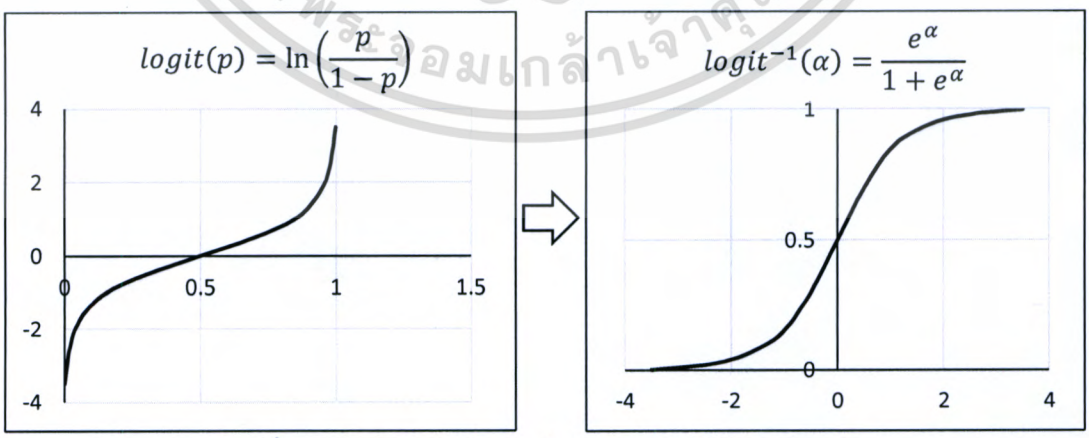
2.4 การวิเคราะห์การถดถอยแบบโลจิสติก (Logistic Regression Analysis)

การวิเคราะห์การถดถอยแบบโลจิสติกเป็นเทคนิคหนึ่งที่สามารถใช้ในงานจำแนกประเภทข้อมูลด้วยวิธีการคำนวณทางสถิติ ซึ่งสามารถใช้แก้ไขปัญหาการจำแนกประเภทข้อมูลที่การใช้วิธีการวิเคราะห์เชิงเส้นแบบดิสคริมิแนนท์ไม่สามารถให้ผลลัพธ์ที่ดีได้ ซึ่งเทคนิคนี้จะใช้วิธีการประมาณค่าความน่าจะเป็นสำหรับผลรวมเชิงเส้น (Linear combination) ของตัวแปรอิสระ (Independent variable) มาใช้ในการจำแนกประเภทข้อมูล โดยการใช้ฟังก์ชันลอการิทึมธรรมชาติ (Natural logarithm function) มาช่วยในการเชื่อมโยงกันระหว่างฟังก์ชันผลรวมเชิงเส้นกับช่วงของค่าความน่าจะเป็นที่ข้อมูลจะเป็นประเภทที่ 1 หรือประเภทที่ 2 ให้อยู่ในรูปของสมการการแจกแจงแบบเบอร์นูลี (Bernoulli probability distribution)

ฟังก์ชันลอกาลิธึมดังกล่าว คือ ฟังก์ชันลอกาลิธึมธรรมชาติของอัตราส่วนออด (Odds ratio) หรือสามารถเรียกได้อีกชื่อว่า ฟังก์ชันโลจิต (Logit Function) สามารถเขียนได้ดังนี้

$$logit(p) = \ln(Odd) = \ln\left(\frac{p}{1-p}\right) \tag{2.11}$$

จากสมการที่ (2.11) ออด (Odd) เป็นแนวคิดที่เกิดจากการเสี่ยงทาย โดยใช้หลักความน่าจะเป็นมาคิดอัตราส่วนของการเกิดเหตุการณ์หนึ่งต่อเหตุการณ์อื่นที่สามารถเกิดขึ้นได้ ซึ่งสมการดังกล่าวเป็นฟังก์ชันที่อยู่ในรูปของฟังก์ชันซิกมอยด์ซึ่งมีความสัมพันธ์กับค่าในแกนแนวนอน แต่เนื่องจากวิธีการจำแนกประเภทแบบโลจิสติกต้องการหาค่าความน่าจะเป็นในแกนแนวตั้ง จึงต้องใช้วิธีการหาสมการผกผันของฟังก์ชันโลจิตจึงจะทำให้สามารถหาค่าความน่าจะเป็นในแกนแนวนอนได้ดังแสดงได้ดังรูปที่ 2.7



รูปที่ 2.7 กราฟแสดงฟังก์ชันโลจิตและฟังก์ชันโลจิตผกผัน

การหาค่าสัมประสิทธิ์การถดถอย (Regression coefficient) สำหรับการถดถอยโลจิสติกสามารถทำได้ด้วยการคำนวณโดยใช้การประมาณค่ามากสุดของความน่าจะเป็นแบบมีเงื่อนไขของคุณลักษณะเมื่อกำหนดเป้าหมาย (Likelihood) เมื่อพิจารณาจากสมการที่ (2.11) จะเห็นว่าสมการลอกาลิธึมธรรมชาติมีความสมมูลกับสมการผลรวมเชิงเส้นของตัวแปรอิสระ ดังนั้นหากต้องการแก้สมการเพื่อหาค่าความน่าจะเป็น p จะสามารถทำได้ด้วยการใช้เทคนิคแอนติลอกาลิธึม (Antilogarithm)

จากสมการที่ (2.12)

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 \tag{2.12}$$

ใช้เทคนิคแอนติลอกาลิธึมเพื่อแก้สมการหาค่า p จะได้

$$antilog\big(logit(p)\big) = \frac{p}{1-p} = e^{\beta_0+\beta_1 x_1} \tag{2.13}$$

$$p = e^{\beta_0+\beta_1 x_1}(1-p)$$

$$p = e^{\beta_0+\beta_1 x_1} - e^{\beta_0+\beta_1 x_1} \cdot p$$

$$p + e^{\beta_0+\beta_1 x_1} \cdot p = e^{\beta_0+\beta_1 x_1}$$

$$p\big(1 + e^{\beta_0+\beta_1 x_1}\big) = e^{\beta_0+\beta_1 x_1}$$

ดังนั้น

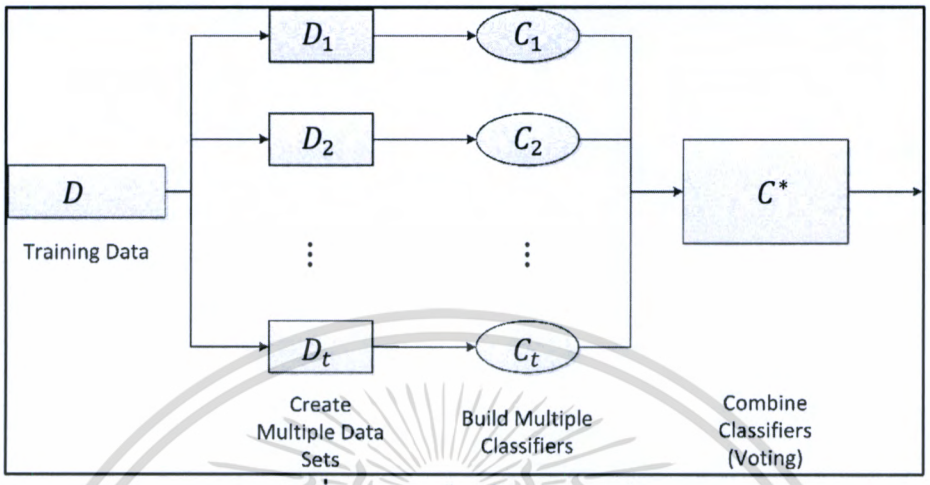
$$p = \frac{e^{\beta_0+\beta_1 x_1}}{1 + e^{\beta_0+\beta_1 x_1}} \tag{2.14}$$

จากสมการที่ (2.14) ซึ่งเป็นการประมาณค่าการถดถอย การหาค่าสัมประสิทธิ์ β สามารถทำได้ด้วยวิธีการหาจากค่าสัมประสิทธิ์การถดถอยเชิงเส้น (Linear regression) หากแทนค่าทั้งหมดแล้วจะได้ค่าประมาณของความเป็นไปได้ที่ข้อมูล x_i จะถูกจัดอยู่ในประเภทที่ 1 จากทั้งหมด 2 ประเภทโดยประเภทแรก คือ $y = 1$ และ $y = 2$ คือประเภทที่ 2

2.5 การเรียนรู้แบบรวมกลุ่ม (Ensemble Learning)

การเรียนรู้แบบรวมกลุ่ม (Ensemble learning) เป็นเทคนิคการเรียนรู้ของเครื่องรูปแบบหนึ่งที่ใช้วิธีการรวมเอาโมเดลการเรียนรู้หลาย ๆ โมเดลเข้าไว้ด้วยกัน โดยแต่ละโมเดลจะเป็นโมเดลการจำแนกประเภทพื้นฐาน (Based classifier) ให้กับการเรียนรู้แบบกลุ่มเพื่อให้ได้ผลลัพธ์ที่ดีขึ้น ซึ่งโดยปกติแล้วการเรียนรู้แบบรวมกลุ่มจะต้องการการคำนวณที่มากกว่าการเรียนรู้แบบใช้เพียงโมเดลเดี่ยว ๆ ดังนั้นการเรียนรู้แบบกลุ่มจึงเหมาะที่จะนำไปใช้ในการเรียนรู้กับชุดข้อมูลที่มีความซับซ้อน

หรือชุดข้อมูลที่ยากต่อการเรียนรู้โดยการเพิ่มการคำนวณการเรียนรู้ให้มากขึ้น โดยสามารถแสดงให้เห็นแผนผังขั้นตอนการทำงานได้ดังรูปที่ 2.8



รูปที่ 2.8 การเรียนรู้แบบรวมกลุ่ม

การเรียนรู้แบบรวมกลุ่มโดยทั่วไปจะแบ่งออกเป็น 2 ประเภท คือ แบบแบ็กกิ้ง (Bagging) และแบบบูสติ้ง (Boosting)

การเรียนรู้แบบรวมกลุ่มแบบแบ็กกิ้งเป็นการเรียนรู้แบบรวมกลุ่มที่สามารถลดความผิดพลาดของผลลัพธ์ได้โดยการจำลองข้อมูลเพิ่มขึ้นมาจากชุดข้อมูลต้นฉบับให้เป็นชุดข้อมูลหลายชุดด้วยการผสมกันของข้อมูลที่ทำซ้ำขึ้นมา และใช้ลักษณะของการนับคะแนนเสียงจากการคำนวณของโมเดลการเรียนรู้พื้นฐาน ซึ่งได้ผลลัพธ์การจำแนกที่แตกต่างกันออกไปตามแต่ละโมเดล โดยจะใช้วิธีการเลือกผลลัพธ์ที่ได้รับคะแนนจากการนับมากที่สุดเป็นผลลัพธ์สุดท้าย

สำหรับการเรียนรู้แบบรวมกลุ่มแบบบูสติ้งจะคล้ายกับการเรียนรู้แบบแบ็กกิ้งแต่ต่างกันที่การเรียนรู้แบบบูสติ้งจะมีการคำนวณค่าน้ำหนักสำหรับทุก ๆ โมเดลพื้นฐานซึ่งมีอิทธิพลต่อการเลือกผลลัพธ์สุดท้าย ยิ่งค่าน้ำหนักมีค่ามากเพียงใด ก็ยิ่งส่งผลให้โมเดลนั้นมีอิทธิพลต่อการตัดสินใจในการเลือกผลลัพธ์มากขึ้นตามไปด้วย โดยในแต่ละรอบของการเรียนรู้ เวกเตอร์ค่าน้ำหนักจะถูกปรับค่าเพื่อสะท้อนถึงประสิทธิภาพในการจำแนกประเภทของแต่ละโมเดลพื้นฐาน

การเรียนรู้แบบรวมกลุ่มเอดาบูสต์เอ็มวัน (Adaboost.M1) เป็นหนึ่งในเทคนิคในการเรียนรู้แบบรวมกลุ่มที่เป็นแบบบูสติ้ง ซึ่งมีขั้นตอนการทำงานที่สามารถแสดงได้ดังรูปที่ 2.9

Input: Training set $S = \{x_i, y_i\}, i = 1, \ldots, N$ and $y_i \in \mathbb{C}, \mathbb{C} = \{c_1, \ldots, c_m\}$ T: Nummber of iterations; W: WeakLearn

Output: Boosted classifier:

$H(x) = \arg\max_{y \in \mathbb{C}} \sum_{t=1}^{T} \ln\left(\frac{1}{\beta_t}\right) I\,[h_t(x) = y]$ where h_t, β_t are the induced classifiers (with $h_t(x) \in \mathbb{C}$) and their assigned weighs, respectively

Method:

1. $D_1(i) \leftarrow \frac{1}{N}$, $i = 1, \ldots, N$
2. **for** t = 1 to T **do**
3. $h_t \leftarrow W(S, D_t)$
4. $\varepsilon_t \leftarrow \sum_{i=1}^{N} D_t(i) I[h_t(x_i) \neq y_i]$
5. **if** $\varepsilon_t > 0.5$ **then**
6. $T \leftarrow t - 1$
7. **return**
8. **end if**
9. $\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t}$
10. $D_{t+1}(i) = D_t(i) \cdot \beta_t^{1-I[h_t(x_i) \neq y_i]}$ for $i = 1, \ldots, N$
11. Normalize D_{t+1} to be a proper distribution
12. **end for**

รูปที่ 2.9 ขั้นตอนวิธีของการเรียนรู้แบบกลุ่มเอดาบูสท์เอ็มวัน

จากรูปที่ 2.9 กำหนดให้ข้อมูลนำเข้า $S = \{x_i, y_i\}$ โดยที่ i มีค่าตั้ง 1 ถึง N และ y_i เป็นสมาชิกของ $\mathbb{C}$ ซึ่งมีสมาชิกเป็นประเภทข้อมูลจำนวน m ตัว และให้ T คือจำนวนรอบของการเรียนรู้ ส่วน W คือ โมเดลการเรียนรู้พื้นฐาน (WeakLearn)

สำหรับขั้นตอนการเรียนรู้ด้วยวิธีการบูสต์จะเป็นขั้นตอนของการวนซ้ำเพื่อทำการเรียนรู้หลาย ๆ รอบ ซึ่งเรียกฟังก์ชันการเรียนรู้ในแต่ละรอบว่า โมเดลการเรียนรู้พื้นฐาน (WeakLearn) ตามที่ได้กล่าวไปแล้ว โดยมีจำนวนการวนรอบทั้งหมด T ครั้ง ซึ่งในแต่ละครั้งของการวนรอบ (กำหนดด้วยตัวแปร t) ตัวบูสต์ (Booster) จะทำการกำหนดค่าการแจกแจง (Distribution) $D_t(i)$ ให้กับข้อมูลแต่ละตัว โดยในครั้งแรกของการวนรอบจะกำหนดให้ทุกตัวมีค่าเท่ากับ $1/N$ จากนั้นตัวบูสต์จะส่งข้อมูลสำหรับการเรียนรู้ S และค่าการแจกแจง D_t เข้าสู่โมเดลการเรียนรู้ W ซึ่งจะได้ผลลัพธ์เป็นสมมติฐาน (Hypotesis) h_t ของแต่ละรอบออกมา กล่าวคือ หน้าที่ของโมเดลการเรียนรู้ในแต่ละรอบคือการหาสมมติฐานที่มีความผิดพลาด ε_t ในการจำแนกประเภทให้น้อยที่สุด ซึ่งในขั้นตอนสุดท้าย

ของแต่ละรอบการเรียนรู้จะใช้ค่าความผิดพลาดดังกล่าวเพื่อทำการคำนวณหาค่าน้ำหนักของแต่ละสมมติฐานต่อไป

การคำนวณค่าความผิดพลาด (Error rate) ε_t คำนวณได้จากการเอาค่าการแจกแจง $D_t(i)$ คูณกับฟังก์ชันอินดิเคเตอร์ (Indicator function) (เมื่อนิพจน์ในฟังก์ชันมีค่าเป็นจริงจะได้ค่า 1 แต่หากเป็นเท็จจะมีค่าเป็น 0) หรือกล่าวให้เข้าใจอย่างได้ง่าย ๆ คือการหาผลรวมของค่า D_t ที่ i แต่ละตัวมีสมมติฐานในการจำแนกประเภทที่ไม่ถูกต้อง นอกจากนี้ให้ทำการตรวจสอบค่าความผิดพลาดที่มีค่าเกินกว่า 0.5 โดยจะให้ทำการยกเลิกการเรียนรู้รอบนั้น ๆ หากมีค่าความผิดพลาดเกินกว่าที่กำหนด ซึ่งค่าความผิดพลาดที่ได้มาจะใช้คำนวณหาค่าปัจจัยน้ำหนัก β_t โดยมีค่าเท่ากับ $\varepsilon_t/(1-\varepsilon_t)$ และค่าปัจจัยน้ำหนักนี้จะใช้สำหรับคำนวณหาค่าน้ำหนักต่อไป

ในขั้นตอนสุดท้ายของแต่ละรอบการเรียนรู้ให้ทำการปรับค่าการแจกแจงใหม่ซึ่งคำนวณได้จากเงื่อนไขดังนี้ หากสมมติฐานของข้อมูลที่ i มีความถูกต้องให้ทำการคูณ $D_t(i)$ กับค่า β_t และหากสมมติฐานของข้อมูลที่ i นั้น ๆ ไม่ถูกต้อง ก็ไม่ต้องทำการปรับค่า $D_t(i)$ ดังกล่าว เมื่อทำการปรับค่าการกระจายแล้ว ให้ทำการลดทอนค่าการแจกแจงหรือการนอมอลไลซ์ (Normalize) ให้มีค่าอยู่ระหว่างค่า 0 ถึง 1

สำหรับผลลัพธ์สุดท้ายที่ได้ (Output) จะได้จากการคำนวณจากสมการที่ (2.15)

$$H(x) = arg\max_{y \in \mathbb{C}} \sum_{t=1}^{T} ln\left(\frac{1}{\beta_t}\right) I\left[h_t(x) = y\right] \qquad (2.15)$$

โดยที่ h_t คือ ผลลัพธ์ของแต่ละรอบการเรียนรู้ และ β_t คือ ค่าปัจจัยน้ำหนักของรอบนั้น

2.6 งานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้มุ่งเน้นไปที่การปรับปรุงการเรียนรู้แบบรวมกลุ่มแบบบูสติ้งเป็นหลัก โดยมีงานวิจัยที่เกี่ยวข้องดังต่อไปนี้

2.6.1 A novel confidence-based multiclass boosting algorithm for mobile physical activity monitoring

งานวิจัยนี้เป็นงานวิจัยที่ทำการดัดแปลงวิธีการเรียนรู้แบบรวมกลุ่มเอดาบูสต์เอ็มวัน โดยใช้ค่าความเชื่อมั่น (Confidence) มาคิดคำนวณและเรียกวิธีการนี้ว่า การเรียนรู้แบบรวมกลุ่มเอดาบูสต์เอ็มวันด้วยค่าความเชื่อมั่น (ConfAdaboost.M1) [1] ซึ่งค่าความเชื่อมั่นจะถูกนำมาใช้คำนวณในขั้นตอนการหาค่าความผิดพลาด (Error rate) โดยนำไปคูณกับค่าการแจกแจง (Distribution) ดังสมการที่ (2.16)

$$e_t = \sum_{i:y_i \neq f_t(x_i)} p_{ti} w_i \tag{2.16}$$

โดยที่ i มีค่าตั้งแต่ 1 ถึง N ซึ่งเป็นจำนวนของข้อมูลที่ใช้ในการเรียนรู้

w_i คือ ค่าการแจกแจงของข้อมูลตัวที่ i

$f_t()$ คือ ฟังก์ชันการเรียนรู้รอบที่ t

x_i คือ ข้อมูลตัวที่ i

p_{ti} คือ ค่าความเชื่อมั่นของข้อมูลตัวที่ i ในรอบการเรียนรู้ t สามารถคำนวณได้จากสมการที่ (2.17)

$$p_{ti} = \frac{\sum_{j \in S_c} w_j}{\sum_{j \in S} w_j} \tag{2.17}$$

โดยที่ j มีค่าตั้งแต่ 1 ถึง N ซึ่งเป็นจำนวนของข้อมูลที่ใช้ในการเรียนรู้

w_j คือ ค่าการแจกแจงของข้อมูลตัวที่ j

S คือ ชุดข้อมูลที่ใช้ในการเรียนรู้

S_c คือ ชุดข้อมูลที่ใช้ในการเรียนรู้ ซึ่งมีประเภทข้อมูลเป็นประเภท c

นอกจากการปรับปรุงการคำนวณค่าความผิดพลาดแล้ว ค่าความเชื่อมั่นยังถูกใช้ในการปรับค่าการแจกแจงทุกครั้งที่มีการสอนอีกด้วยดังสมการที่ (2.18)

$$w_i \leftarrow w_i e^{\left(\frac{1}{2} - I\left(y_i = f_t(\underline{x_i})\right)\right) p_{ti} \alpha_t} \tag{2.18}$$

โดยที่ α_t คือ ค่าปัจจัยน้ำหนักที่เป็นเลขจำนวนจริงตั้งแต่ 0 ถึง 1

และเมื่อต้องการจำแนกประเภทกลุ่มของข้อมูล x_n ก็สามารถทำได้ด้วยการเลือกประเภทที่มีคะแนนมากที่สุดโดยคิดตามค่าน้ำหนักและค่าความเชื่อมั่นของแต่ละสมมติฐาน ซึ่งสามารถคำนวณได้จากสมการที่ (2.19)

$$\mu_c \leftarrow \mu_c + p_t(\underline{x_n})\alpha_t \tag{2.19}$$

โดยที่ c คือ ประเภทของข้อมูล มีค่าตั้งแต่ 1 ถึง C ซึ่งเป็นจำนวนของประเภทข้อมูลที่สามารถแบ่งได้ทั้งหมด

μ_c คือ ค่าน้ำหนักของแต่ละประเภทข้อมูล

$\underline{x_n}$ คือ ข้อมูลที่ต้องการทราบประเภท

$p_t(\underline{x_n})$ คือ ค่าความเชื่อมั่นในการจำแนกประเภทข้อมูลด้วยสมมติฐาน t

งานวิจัยที่กล่าวถึงนี้ได้ทำการทดลองวัดประสิทธิภาพจากการสำรวจค่าความผิดพลาดที่เกิดจากการจำแนก 10 ครั้ง และหาค่าเฉลี่ยของค่าความผิดพลาด โดยการใช้ชุดข้อมูลจากเว็บไซต์ของมหาวิทยาลัยแคลิฟอร์เนีย ไอร์วีน (UCI: University of California, Irvine) มาทำการทดลองและเปรียบเทียบผลกับวิธีการจำแนกประเภทแบบรวมกลุ่มหลายแบบรวมถึงเอดาบูสต์เอ็มวัน โดยงานงานวิจัยแสดงให้เห็นถึงผลลัพธ์ที่ได้จากการจำแนกประเภทด้วยวิธีการใช้ค่าความเชื่อมั่นเป็นวิธีที่มีค่าความผิดพลาดเฉลี่ยน้อยที่สุด

2.6.2 A New Ensemble Model based on Linear Mapping, Nonlinear Mapping, and Probability Theory for Classification Problems

งานวิจัยนี้นำเสนอการประยุกต์ใช้โมเดลการจำแนกประเภทข้อมูลแบบรวมกลุ่มเอดาบูสต์เอ็มวันโดยผู้วิจัยเรียกงานวิจัยของตนเองว่า การเรียนรู้แบบรวมกลุ่มเอ็มอาร์เอ็น (MRN ensemble) [2] ด้วยสมมติฐานของการใช้การจำแนกประเภทข้อมูลด้วยวิธีการคำนวณเชิงเส้น (Linear) การคำนวณไม่เชิงเส้น (non-linear) และการคำนวณโดยใช้หลักความน่าจะเป็น (Probability) โดยใช้การจำแนกประเภทด้วยวิธีการคำนวณแบบโครงข่ายประสาทเทียมเพอร์เซ็ปตรอนหลายชั้น แบบเรเดียลเบสิสฟังก์ชัน (Radial basis function) และนาอีฟเบย์ (Naïve Bayes) ตามลำดับ กล่าวคือ เป็นการนำเอาขั้นตอนวิธีการจำแนกประเภทแบบรวมกลุ่มเอดาบูสต์มาประยุกต์ใช้ และกำหนดรอบของการเรียนรู้ไว้ 3 รอบ โดยในแต่ละรอบจะใช้วิธีการดังที่กล่าวมาเป็นโมเดลการเรียนรู้พื้นฐาน

บทที่ 3
โมเดลการรวมกลุ่มผสมระหว่างการเรียนรู้ของเครื่องและการเรียนรู้เชิงสถิติโดยใช้การบูสต์ด้วยค่าความเชื่อมั่น

ในบทนี้จะอธิบายถึงแนวคิดและขั้นตอนการรวมกลุ่มผสมระหว่างการเรียนรู้ของเครื่องและการเรียนรู้เชิงสถิติโดยใช้การบูสต์ด้วยค่าความเชื่อมั่น

3.1 ขั้นตอนการเตรียมข้อมูล

เนื่องจากในงานวิจัยชิ้นนี้มุ่งเน้นไปที่การพัฒนาปรับปรุงขั้นตอนวิธีการเรียนรู้แบบรวมกลุ่ม ซึ่งจะใช้ชุดข้อมูลจำนวนหนึ่งในการนำมาทดสอบเพื่อวัดประสิทธิภาพในการเรียนรู้การจำแนกประเภทของงานวิจัยนี้ โดยผู้วิจัยได้เลือกใช้ชุดข้อมูลที่มีอยู่บนเว็บไซต์ UCI [3] ซึ่งเป็นชุดข้อมูลที่มีลักษณะที่แตกต่างกันออกไปโดยสามารถอธิบายได้ดังนี้

ตารางที่ 3.1 ชุดข้อมูลที่นำมาใช้วัดประสิทธิภาพในงานวิจัย

ชื่อชุดข้อมูล (Data set)	**จำนวนประเภท** (Class)	**จำนวนคุณลักษณะ** (Attribute)	**จำนวนข้อมูล** (instance)
Banknote Authentication	2	5	1372
Connectionist Bench	2	60	208
Iris	3	4	150
PAMAP	18	39	21401
Ecoli	8	7	336
Glass Identification	7	9	214
Car Evaluation	2	6	1728

3.1.1 ชุดข้อมูล Banknote Authentication

จากตารางที่ 3.1 ชุดข้อมูลที่ใช้ในการวัดประสิทธิภาพชุดแรกคือชุดข้อมูลสำหรับทดสอบการตรวจจับธนบัตรปลอม [4] โดยเป็นชุดข้อมูลที่ผ่านการสกัด (Extract) ก่อนจะมีการนำมาเผยแพร่ด้วยการสกัดข้อมูลจากภาพถ่ายของธนบัตรตัวอย่างทั้งธนบัตรที่ถูกต้องและธนบัตรที่ถูกทำปลอมขึ้น โดย

ได้มีการนำกล้องที่ใช้สำหรับตรวจสอบงานพิมพ์ในระดับอุตสาหกรรมมาใช้ในการถ่ายภาพ ซึ่งภาพสุดท้ายที่ได้มามีขนาดความกว้าง 400 ความยาว 400 พิกเซล (pixel) และมีภาพเป็นภาพขาว-ดำที่ความละเอียด 660 จุดต่อตารางนิ้ว (dpi) อีกทั้งยังผ่านขั้นตอนการแปลงคลื่นสัญญาณด้วยวิธีเวฟเลต (Wavelet transform) เพื่อให้ข้อมูลอยู่ในรูปของตัวเลข

ประเภทของข้อมูลชุดนี้มีทั้งสิ้น 2 ประเภท คือ ธนบัตรจริง และธนบัตรปลอม โดยใช้ค่า 1 และ 0 ตามลำดับเพื่อแทนค่าสำหรับการสอน และมีคุณลักษณะของข้อมูลอีก 4 คุณลักษณะ โดย 1 ใน 4 ของคุณลักษณะคือค่าเอนโทรปีของภาพ และส่วนที่เหลือคือค่าของการแปลงคลื่นสัญญาณด้วยวิธีเวฟเลต ซึ่งคุณลักษณะทั้งหมดยกเว้นประเภทของข้อมูลเป็นข้อมูลเชิงปริมาณ

3.1.2 ชุดข้อมูล Connectionist Bench

Connectionist Bench [5] เป็นชุดข้อมูลที่เก็บคลื่นสัญญาณการสะท้อนกลับของโซนาร์ (Sonar) ที่ตกกระทบกับแร่เหล็กและก้อนหินธรรมดา โดยข้อมูลที่ได้มาจะถูกแบ่งออกเป็น 2 ส่วนตามประเภทของคลื่นที่ตกกระทบ คือ คลื่นที่ตกกระทบกับแร่เหล็กจำนวน 111 ตัวอย่าง และคลื่นที่ตกกระทบกับก้อนหินจำนวน 97 ตัวอย่าง โดยในการเก็บข้อมูลของแต่ละประเภทจะมีการเก็บคลื่นสัญญาณที่สะท้อนกลับมาจากวัตถุในหลายทิศทาง ในช่วงมุมไม่เกิน 90 องศาสำหรับแร่เหล็กและช่วงมุมไม่เกิน 180 องศาสำหรับก้อนหิน

สำหรับในแต่ละเรคคอร์ด (Record) ตัวอย่างจะมีคุณลักษณะซึ่งเป็นตัวเลขที่มีค่าระหว่าง 0.0 ถึง 1.0 จำนวน 60 ตัว โดยที่เลขแต่ละตัวคือค่าพลังงานที่วัดได้ในแต่ละคลื่นความถี่และผ่านการรวมค่าเมื่อได้รับสัญญาณเป็นระยะเวลาหนึ่ง นอกจากนี้ในแต่ละเรคคอร์ดจะมีคุณลักษณะกำกับประเภทของสัญญาณด้วยตัวอักษรภาษาอังกฤษ M สำหรับแร่เหล็กและ R สำหรับก้อนหิน

ข้อมูลชุดนี้นำมาใช้ในการทดลองเพื่อวัดประสิทธิภาพในการจำแนกประเภทที่มีประเภทของข้อมูล 2 แบบ หรือเรียกว่าการจำแนกประเภทแบบไบนารี (Binary classification)

3.1.3 ชุดข้อมูล Iris

ชุดข้อมูล Iris [6] ถูกจัดเก็บและสร้างขึ้นโดย Ronald A. Fisher นักสถิติศาสตร์ผู้คิดค้นสมการวิเคราะห์การถดถอยเชิงเส้น (Linear Discriminant Analysis) และทฤษฎีทางสถิติศาสตร์ที่สำคัญอื่น ๆ โดยชุดข้อมูลชุดนี้เป็นชุดข้อมูลที่มีความนิยมอย่างมากในการนำมาใช้ในการทดสอบประสิทธิภาพการจำแนกประเภทในหลายงานวิจัย

ชุดข้อมูลนี้จัดเก็บข้อมูลขนาดของกลีบดอกและกลีบเลี้ยงของดอกไอริส ซึ่งมีคุณลักษณะของข้อมูลอยู่ 4 คุณลักษณะด้วยกัน แบ่งเป็นข้อมูลความกว้างของกลีบดอก ความยาวของกลีบดอก ความกว้างของกลีบเลี้ยง ความยาวของกลีบเลี้ยง โดยทั้งหมดอยู่ในหน่วยเซนติเมตร นอกจากนี้ข้อมูลของแต่ละดอกจะถูกแยกประเภทออกตามแต่ละสายพันธุ์ซึ่งมีอยู่ในชุดข้อมูลทั้งสิ้น 3 สายพันธุ์ ได้แก่

Setosa Versicolour และVirginica โดยแต่ละสายพันธุ์จะมีข้อมูลตัวอย่างอยู่สายพันธุ์ละ 50 ตัวอย่าง เหตุผลหลักที่เลือกข้อมูลชุดนี้มาใช้เพราะเป็นชุดข้อมูลที่ผู้วิจัยต้องการนำมาทดสอบประสิทธิภาพในการจำแนกประเภทแบบหลายประเภท (Multi-class classification)

3.1.4 ชุดข้อมูล PAMAP2

PAMAP2 หรือเรียกชื่อเต็มทางภาษาอังกฤษว่า *"Physical Activity Monitoring Data Set"* [7] เป็นข้อมูลการทำกิจกรรมต่าง ๆ ของมนุษย์โดยการจัดเก็บการเคลื่อนไหวส่วนต่าง ๆ ของร่างกาย ซึ่งอาสาสมัครจำนวน 9 รายได้ถูกทำการติดเซ็นเซอร์วัดอัตราการเต้นของหัวใจและเซ็นเซอร์ IMU เอาไว้ โดยที่เซ็นเซอร์ IMU เหล่านี้จะถูกนำไปติดไว้ที่ตามส่วนต่าง ๆ ของร่ายกาย เช่น มือ หน้าอก และข้อเท้า ทั้งนี้เซ็นเซอร์ IMU นั้นประกอบไปด้วยเซ็นเซอร์ย่อยอีก 6 เซ็นเซอร์ ได้แก่ เซ็นเซอร์ตรวจวัดอุณหภูมิ เซ็นเซอร์ตรวจจับความเร่ง 3 มิติ (3D-acceleration) ขนาด 13 บิต ที่อัตราส่วน 16 g เซ็นเซอร์ตรวจจับความเร่ง 3 มิติ ขนาด 13 บิต ที่อัตราส่วน 6 g เซ็นเซอร์ไจโรสโคป 3 มิติ (3D-Gyroscope) เซ็นเซอร์ตรวจจับสนามแม่เหล็ก (3D-magnetometer) และเซ็นเซอร์ตรวจจับการตั้ง-การนอน (Orientation) เป็นต้น และสำหรับกิจกรรมที่ให้อาสาสมัครทำสามารถแบ่งได้ทั้งสิ้น 19 กิจกรรมดังนี้

1) นอนราบ
2) นั่ง
3) ยืน
4) เดิน
5) วิ่ง
6) ปั่นจักรยาน
7) เดินแบบนอร์ดิก (Nordic walking - เป็นการเดินโดยมีไม้ค้ำเป็นอุปกรณ์ร่วม)
8) ชมรายการโทรทัศน์
9) ทำงานที่เครื่องคอมพิวเตอร์
10) ขับรถยนต์
11) เดินขึ้นบันได
12) เดินลงบันได
13) ดูดฝุ่น
14) รีดผ้า
15) พับผ้า
16) ทำความสะอาดบ้าน
17) เล่นฟุตบอล
18) กระโดดเชือก

19) อื่น ๆ

จากข้อมูลเซ็นเซอร์ดังกล่าวสรุปแล้วชุดข้อมูลนี้มีคุณลักษณะทั้งหมด 52 คุณลักษณะ มีประเภทของข้อมูลทั้งหมด 19 ประเภท และจำนวนเรคคอร์ดข้อมูลกว่า 3.8 ล้านเรคคอร์ด ซึ่งถือว่าเป็นข้อมูลที่มีจำนวนค่อนข้างมาก อีกทั้งข้อมูลถูกเก็บจากหลายอิริยาบถ หลายกิจกรรมของมนุษย์ ซึ่งอาจทำให้เกิดข้อมูลรบกวน (Noise) ได้บ้างจึงอาจเป็นการยากในการเรียนรู้การจำแนกประเภทกิจกรรมดังกล่าว

อย่างไรก็ตาม เนื่องจากในชุดข้อมูลชุดนี้ยังมีข้อมูลบางส่วนที่ขาดหายไป ซึ่งเกิดจากปัญหาการส่งข้อมูลจากเซ็นเซอร์ผ่านสัญญาณไร้สาย เช่น ข้อมูลอุณหภูมิที่เซ็นเซอร์ IMU ข้อมูลจากเซ็นเซอร์ตรวจจับแนวตั้ง-แนวนอน และกิจกรรมบางกิจกรรมที่มีข้อมูลที่น้อยเกินไปสำหรับการเรียนรู้ (บางกิจกรรมมีตัวอย่างจากอาสาสมัครเพียง 1 หรือ 2 คนเท่านั้น) ดังนั้นในงานวิจัยชิ้นนี้จะมีการลดทอนข้อมูลของชุดข้อมูลนี้ลงไปบ้าง โดยมีรายละเอียดการลดทอนข้อมูลดังนี้

1) ทำการตัดคุณลักษณะที่เป็นส่วนเก็บข้อมูลจากเซ็นเซอร์ IMU บางชนิดออก ได้แก่ ค่าอุณหภูมิและค่าจากเซ็นเซอร์ตรวจจับการตั้ง-การนอน โดยตัดคุณลักษณะที่ 1 17 18 19 20 34 35 36 37 51 52 53 54 ออก
2) ทำการเลือกข้อมูลมา 200 เรคคอร์ดแรกจากทุกประเภทกิจกรรมของอาสาสมัครแต่ละคน โดยหากมีข้อมูลไม่ถึง 200 เรคคอร์ดในกิจกรรมประเภทนั้น ก็ให้นำข้อมูลในการทำกิจกรรมนั้นมาเท่าที่ข้อมูลจะสามารถเอื้อได้
3) เนื่องจากข้อมูลจาก UCI มีการจัดเก็บในลักษณะ 1 ไฟล์ต่ออาสาสมัคร 1 คน ซึ่งในแต่ละไฟล์จะมีข้อมูลการทำกิจกรรมต่าง ๆ ของอาสาสมัครคนนั้นอยู่ อย่างไรก็ดีการเก็บข้อมูลของการทำกิจกรรมจากอาสาสมัครทุกคนจะจัดเก็บเพียงแค่ 12 กิจกรรมเท่านั้น ได้แก่ นอนราบ นั่ง ยืน รีดผ้า ดูดฝุ่น เดินขึ้นบันได เดินลงบันได เดินปกติ เดินแบบนอร์ดิก ปั่นจักรยาน วิ่ง กระโดดเชือก เป็นต้น และกิจกรรมที่เหลือทางผู้สร้างชุดข้อมูลได้เตรียมแยกไว้ต่างหาก แต่ผู้วิจัยได้ทำการรวมข้อมูลการทำกิจกรรมที่นอกเหนือจากกิจกรรมข้างต้นเหล่านั้นเข้ามาด้วย

กล่าวโดยสรุปคืองานวิจัยนี้จะใช้ข้อมูลจากชุดข้อมูล PAMAP2 โดยรวมทั้งข้อมูลหลักและข้อมูลที่เป็นข้อมูลส่วนข้อมูลตัวเลือกมาได้จำนวนประเภทกิจกรรมทั้งสิ้น 18 กิจกรรม และลดทอนคุณลักษณะลงเหลือเป็นจำนวน 39 คุณลักษณะและจำนวนข้อมูล 21,401 เรคคอร์ด ซึ่งจากจำนวนของข้อมูลดังกล่าว ข้อมูลชุดนี้จึงจัดว่าเป็นข้อมูลอีกชุดหนึ่งที่มีความน่าสนใจในการนำมาวัดประสิทธิภาพการเรียนรู้การจำแนกประเภทได้เป็นอย่างดี ตารางที่ 3.2 แสดงจำนวนข้อมูลของกิจกรรมต่าง ๆ ที่ได้หลังการลดทอนข้อมูล โดยมีกิจกรรมบางอย่างที่ได้ข้อมูลมาน้อย เช่น การเล่นฟุตบอล และการกระโดดเชือก เป็นต้น สาเหตุที่ทำให้ข้อมูลกิจกรรมดังกล่าวมีน้อยนั่นก็คือ ข้อมูลที่ผู้สร้างชุดข้อมูลนี้จัดทำขึ้นมาสามารถจัดเก็บข้อมูลในการทำกิจกรรมบางกิจกรรมได้จากอาสาสมัคร

เพียงคนเดียวหรือ 2 คน จึงทำให้หลังจากทำการรวมข้อมูลแล้วจะมีข้อมูลจากอาสาสมัครเพียงคนเดียวจึงทำให้ข้อมูลมีจำนวนน้อยตามไปด้วย

ตารางที่ 3.2 จำนวนข้อมูลของกิจกรรมต่าง ๆ หลังทำการลดทอนจำนวนชุดข้อมูล

กิจกรรม	จำนวนที่สกัดออกมา
นอนราบ	1927
นั่ง	1846
ยืน	1890
เดิน	2297
วิ่ง	957
ปั่นจักรยาน	1631
เดินแบบนอร์ดิก	1845
ชมรายการโทรทัศน์	836
ทำงานด้วยเครื่องคอมพิวเตอร์	3099
ขับรถยนต์	545
เดินขึ้นบันได	1172
เดินลงบันได	1047
ดูดฝุ่น	1752
รีดผ้า	2378
พับผ้า	995
ทำความสะอาดบ้าน	1865
เล่นฟุตบอล	459
กระโดดเชือก	477

3.1.5 ชุดข้อมูล Ecoli

ชุดข้อมูล Ecoli [8] จัดเก็บข้อมูลสัญญาณเฉพาะของยูแคริโอต (Eukaryote) จำนวน 336 ตัวอย่าง ซึ่งมีจำนวนคุณลักษณะอยู่ 7 คุณลักษณะ โดยที่คุณลักษณะแต่ละตัว คือ ค่าของคะแนน (score) สำหรับระบุประเภทการจัดเรียงของโปรตีน (protein sequence) โดยจะเป็นค่าทศนิยมตั้งแต่ 0 ถึง 1 โดยหากมีค่าคะแนนมากเท่าไหร่ ก็ยิ่งมีความน่าจะเป็นที่จะถูกจัดอยู่ในรูปแบบการจัดเรียงของโปรตีนประเภทนั้น ๆ มากขึ้นตามไปด้วย ซึ่งคุณลักษณะเหล่านี้ ได้แก่

1) mcg วิธีการจัดเรียงแบบแม็คโกช (McGeoch's method for signal sequence recognition.)
2) gvh วิธีการจัดเรียงแบบวอน ไฮจีน (Von Heijne's method for signal sequence recognition.)
3) lip คะแนนความสอดคล้องของการจัดเรียงกันแบบวอน ไฮจีน ทู (Von Heijne's Signal Peptidase II consensus sequence score.)
4) chg การปรากฏของปลายเอ็นของไลโปโปรตีนที่คาดหวัง (Presence of charge on N-terminus of predicted lipoproteins.)
5) aac คะแนนการวิเคราะห์การถดถอยแบบดิสคริมิแนนท์ด้วยการวิเคราะห์กรดอะมิโนจากเยื่อหุ้มเซลล์ชั้นนอกและช่องว่างระหว่างเยื่อหุ้มเซลล์ (Score of discriminant analysis of the amino acid content of outer membrane and periplasmic proteins.)
6) alm1 คะแนนที่ได้จากการคาดเดาด้วยโปรแกรม ALOM (Score of the ALOM membrane spanning region prediction program.)
7) alm2 คะแนนที่ได้จากการคาดเดาด้วยโปรแกรม ALOM หลังจากทำการตัดค่า (Score of ALOM program after excluding putative cleavable signal regions from the sequence.)

นอกจากคุณลักษณะที่กล่าวมาแล้วในชุดข้อมูลดังกล่าวยังมีคุณลักษณะอีก 1 คุณลักษณะ ซึ่งผู้วิจัยได้ทำการตัดออกไปก่อนนำข้อมูลมาใช้ทำการทดลอง นั่นก็คือคุณลักษณะในลำดับที่ 1 ซึ่งเป็นคุณลักษณะที่เก็บข้อมูลชื่อการจัดเรียงของโปรตีน (Sequence name) ในรูปแบบรหัสตัวเลขตัวอักษร ส่วนประเภทกลุ่มข้อมูลของชุดข้อมูลนี้จะมีอยู่ด้วยกัน 8 ประเภท ได้แก่

1) cp (Cytoplasm) จะแทนด้วยค่า 1
2) im (inner membrane without signal sequence) จะแทนด้วยค่า 2
3) imS (inner membrane, cleavable signal sequence) จะแทนด้วยค่า 3
4) imL (inner membrane lipoprotein) จะแทนด้วยค่า 4
5) imU (inner membrane, uncleavable signal sequence) จะแทนด้วยค่า 5
6) om (outer membrane) จะแทนด้วยค่า 6
7) omL (outer membrane lipoprotein) จะแทนด้วยค่า 7
8) pp (perisplasm) จะแทนด้วยค่า 8

3.1.6 ชุดข้อมูล Glass Identification

Glass Identification [9] เป็นชุดข้อมูลที่ถูกสร้างขึ้นจากแรงบันดาลใจในการสืบสวนคดีความต่าง ๆ ด้วยความพยายามที่จะทำการระบุถึงชิ้นส่วนของเศษกระจกกับกระจกจากที่ต่าง ๆ ที่อาจเป็นกระจกบานเดียวกันชิ้นส่วนเศษกระจกที่พบอยู่ในสถานที่เกิดเหตุ ซึ่งอาจเป็นประโยชน์ในการใช้เป็นหลักฐานสืบสวนในคดีความต่าง ๆ ได้ ชุดข้อมูลนี้มีจำนวนข้อมูลตัวอย่าง 214 ตัวอย่าง และมีคุณลักษณะอยู่ทั้งสิ้น 11 คุณลักษณะ โดยในงานวิจัยนี้จะทำการตัดคุณลักษณะที่ 1 ซึ่งเป็นรหัสของชิ้นส่วนเศษกระจกออก ส่วนคุณลักษณะที่เหลือที่ใช้ในการทดลองส่วนใหญ่จะเป็นข้อมูลสัดส่วนของแร่ธาตุที่เป็นส่วนประกอบของชิ้นส่วนกระจก ซึ่งมีดังต่อไปนี้

1) RI ดรรชนีความหักเหของแสง
2) Na สัดส่วนธาตุโซเดียม
3) Mg สัดส่วนธาตุแมกนีเซียม
4) Al สัดส่วนธาตุอลูมิเนียม
5) Si สัดส่วนแร่ธาตุซิลิคอน
6) K สัดส่วนแร่ธาตุโพแทสเซียม
7) Ca สัดส่วนแร่ธาตุแคลเซียม
8) Ba สัดส่วนแร่ธาตุแบเรียม
9) Fe สัดส่วนแร่ธาตุเหล็ก

และสำหรับจำนวนกลุ่มประเภทข้อมูลมีทั้งสิ้น 7 ประเภท ได้แก่

1) building_windows_float_processed (จำนวนข้อมูล 70)
2) building_windows_non_float_processed (จำนวนข้อมูล 76)
3) vehicle_windows_float_processed (จำนวนข้อมูล 17)
4) vehicle_windows_non_float_processed (ไม่มีข้อมูลประเภทนี้)
5) containers (จำนวนข้อมูล 13)
6) tableware (จำนวนข้อมูล 9)
7) headlamps (จำนวนข้อมูล 29)

ซึ่งกลุ่มประเภทข้อมูลทั้ง 7 กลุ่มนี้เป็นกระจกที่มาจากหน้าต่างอาคาร หน้าต่างยานพาหนะ ภาชนะกระจก โต๊ะกระจก โคมไฟ โดยกระจกที่เป็น Float process คือ บานกระจกที่ผ่านการกระบวนการผลิตพิเศษที่ทำให้กระจกที่มีลักษณะกว้างแต่บางและสามารถคงอยู่ในรูปที่เป็นลักษณะดังกล่าวได้หรือจะแตกได้ยากกว่ากระจกที่ไม่ผ่านกระบวนการดังกล่าว อย่างไรก็ตามข้อมูลชุดนี้จะไม่มีข้อมูลที่เป็นประเภท vehicle_windows_float_processed

3.1.7 ชุดข้อมูล Car Evaluation

ชุดข้อมูล Car Evaluation [10] เป็นชุดข้อมูลที่ได้มาจากการสร้างแบบจำลองการตัดสินใจซื้อรถยนต์อย่างง่าย ซึ่งเดิมถูกสร้างขึ้นเพื่อใช้ในการสาธิตเทคนิคการตัดสินใจแบบเชี่ยวชาญ (DEX: Decision Expert) [11] โดยมีแนวคิดในการวางโครงสร้างการตัดสินใจจากข้อมูลด้านต่าง ๆ 2 ด้าน คือ

1) ด้านราคา (PRICE: overall price) ประกอบด้วยข้อมูลราคาขาย (buying: buying price) และข้อมูลราคาการซ่อมบำรุง (maint: price of the maintenance)
2) ด้านคุณลักษณะทางเทคนิค (TECH: technical characteristics) แบ่งเป็น
    ก) ด้านความสะดวกสบาย (COMFORT) ประกอบด้วย จำนวนประตู (doors: number of doors) ความจุผู้โดยสาร (persons: capacity in terms of persons to carry) ความจุช่องใส่สัมภาระ (lug_boot: the size of luggage boot)
    ข) ด้านความปลอดภัย (safety: estimated safety of the car)

แต่ชุดข้อมูลที่นำมาใช้ในการทดลองซึ่งนำมาจากเว็บไซต์ UCI จะเป็นชุดข้อมูลที่ถูกทำการลดทอนจำนวนข้อมูลออกไปจำนวนหนึ่งแล้วจากผู้ที่บริจาคข้อมูล โดยได้มีการตัดเอาข้อมูลออกจนเหลือคุณลักษณะอยู่ 6 คุณลักษณะ ได้แก่ ข้อมูลราคาขาย ข้อมูลราคาซ่อมบำรุง ข้อมูลจำนวนประตู ข้อมูลความจุผู้โดยสาร ข้อมูลความจุสัมภาระ และข้อมูลความปลอดภัย และเนื่องจากข้อมูลที่ได้มาจะอยู่ในรูปของข้อมูลแบบนามบัญญัติที่เป็นข้อความ ดังนั้นผู้วิจัยจึงได้มีการแปลงข้อมูลดังกล่าวให้เป็นข้อมูลที่เป็นตัวเลขเพื่อให้สามารถทำการคำนวณได้ โดยข้อมูลจะทำการแปลงตามลำดับที่แสดงดังในตารางที่ 3.3 เช่น คุณลักษณะ persons ที่มีข้อมูลเดิมเป็น “2” “4” และ “more” จะถูกเปลี่ยนให้เป็นค่า 1 2 และ 3 ตามลำดับ

ตารางที่ 3.3 การเปรียบเทียบข้อมูลของแต่ละคุณลักษณะก่อนและหลังการแปลงข้อมูล

ชื่อคุณลักษณะ	ข้อมูลเดิม	ข้อมูลหลังการแปลง
buying	v-high, high, med, low	1, 2, 3, 4
maint	v-high, high, med, low	1, 2, 3, 4
doors	2, 3, 4, 5-more	1, 2, 3, 4
persons	2, 4, more	1, 2, 3
lug_boot	small, med, big	1, 2, 3
safety	low, med, high	1, 2, 3

นอกจากการแปลงข้อมูลคุณลักษณะแล้วยังได้มีการแปลงประเภทข้อมูลให้อยู่ในรูปของตัวเลขอีกด้วย ซึ่งประเภทข้อมูลของชุดข้อมูลนี้มีทั้งสิ้น 4 ประเภท ได้แก่ unacc acc good และ vgood โดยจะแปลงประเภทข้อมูลดังตารางที่ 3.4

ตารางที่ 3.4 เปรียบเทียบประเภทข้อมูลก่อนและหลังการแปลงข้อมูล

ข้อมูลเดิม	ข้อมูลหลังการแปลง	จำนวนข้อมูล	สัดส่วนร้อยละของจำนวน
unacc	1	1210	70.023
acc	2	384	22.222
good	3	69	3.993
vgood	4	65	3.762

3.2 ขั้นตอนวิธีการรวมกลุ่มผสมระหว่างการเรียนรู้ของเครื่องและการเรียนรู้เชิงสถิติโดยใช้การบูสต์ด้วยค่าความเชื่อมั่น

เทคนิคที่นำเสนอนี้เป็นเทคนิคการเรียนรู้แบบรวมกลุ่มที่มุ่งเน้นสำหรับใช้ในการแก้ปัญหาการจำแนกประเภทข้อมูลด้วยการนำเอาเทคนิคการเรียนรู้ของเครื่อง (Machine learning) มาประยุกต์ใช้ผสมกับเทคนิคการจำแนกประเภทด้วยวิธีการทางสถิติ (Statistical learning) โดยใช้แนวคิดการเรียนรู้แบบรวมกลุ่มเอ็มอาร์เอ็น (ดูข้อ 2.6.2) ที่ใช้เทคนิคการแยกประเภทต่าง ๆ มาใช้ร่วมกันในเทคนิคการเรียนรู้แบบรวมกลุ่มกลุ่มเดียวกัน อีกทั้งยังนำแนวคิดการใช้ค่าความเชื่อมั่น (ดูข้อ 2.6.1) มาประยุกต์ใช้ด้วย เพื่อเพิ่มประสิทธิภาพในการให้ค่าน้ำหนักสำหรับทำการจำแนกประเภทให้ดียิ่งขึ้น

การเรียนรู้แบบรวมกลุ่มที่นำเสนอในงานวิจัยนี้จะใช้หลักการทำงานที่ประยุกต์มาจากการทำงานแบบรวมกลุ่มแบบบูสติ้ง โดยเลือกใช้เทคนิคการเรียนรู้แบบรวมกลุ่มเอดาบูสต์เอ็มวันมาเป็นเทคนิคต้นแบบ แต่ดัดแปลงในขั้นตอนการคำนวณหาค่าน้ำหนักโดยการเพิ่มค่าความเชื่อมั่นเข้าไปด้วย ซึ่งขั้นตอนการทำงานของการรวมกลุ่มผสมระหว่างการเรียนรู้ของเครื่องและการเรียนรู้เชิงสถิติโดยใช้การบูสต์ด้วยค่าความเชื่อมั่นสามารถแสดงได้ดังนี้

ขั้นตอนการทำงานของการรวมกลุ่มผสมระหว่างการเรียนรู้ของเครื่องและการเรียนรู้เชิงสถิติโดยใช้การบูสต์ด้วยค่าความเชื่อมั่น

ข้อมูลนำเข้า: กำหนดให้ชุดข้อมูลเรียนรู้ $S = \{x_i, y_i\}, i = 1, \dots, N$ โดยมีประเภทของข้อมูล คือ $y_i \in \Omega, \Omega = \{y_1, \dots, y_c\}$

ขั้นตอนวิธี:

1. กำหนดค่าการแจกแจง D_t เริ่มต้นให้กับ x แต่ละตัว โดย $D_1(i)$ มีค่าเท่ากับ $\frac{1}{n}$ โดยที่ i เป็นเลขลำดับของสมาชิกที่มีค่าตั้งแต่ 1 ถึง n
2. กำหนดค่า T ให้มีค่าเท่ากับจำนวนของ WeakLearn และให้ t มีค่าเริ่มต้นเป็นศูนย์
3. ปรับปรุงค่า t โดยให้ $t = \mathrm{t} + 1$
4. ทำการคำนวณหาผลสมมติฐานของ WeakLearn เขียนเป็นสมการได้เป็น

$$h_t : S \to \Omega \tag{3.1}$$

5. คำนวณค่าความเชื่อมั่นจากสมการ

$$p_t = \frac{\sum_{j:h(x_j)=y_j} D_j}{\sum_j D_j} \tag{3.2}$$

โดยที่ p_t คือ ค่าความเชื่อมั่น

$\sum_{j:h(x_j)=y_j} D_j$ คือ ผลรวมของค่าการแจกแจงที่ได้สมมติฐานที่ถูก

$\sum_j D_j$ คือ ผลรวมของค่าการแจกแจงทั้งหมด

6. คำนวณหาค่าความผิดพลาด (Error rate) ของ h_t จากสมการ

$$\varepsilon_t = \sum_{i=1}^{n} I[h_t(x_i) \neq y_i] \cdot D_i \cdot p_t \tag{3.3}$$

7. เปรียบเทียบค่าความผิดพลาด หากค่าความผิดพลาดมีค่ามากกว่า 0.5 ให้ย้อนกลับไปทำการหาสมมติฐานใหม่ที่ขั้นตอนที่ 4
8. คำนวณหาค่าน้ำหนัก β_t จากสมการ

$$\beta_t = \begin{cases} \beta_t = 1 \; if \; \varepsilon_t = 0 \\ \frac{1}{2} log \frac{(1-\varepsilon_t)}{\varepsilon_t} \; otherwise \end{cases} \tag{3.4}$$

9. ปรับปรุงค่าการแจกแจง D_t จากสมการ

$$D_{t+1}(i) = D_t(i) \cdot e^{\left(\frac{1}{2} - I[h_t(x_i)=y_i]\right)\beta_t p_t} \tag{3.5}$$

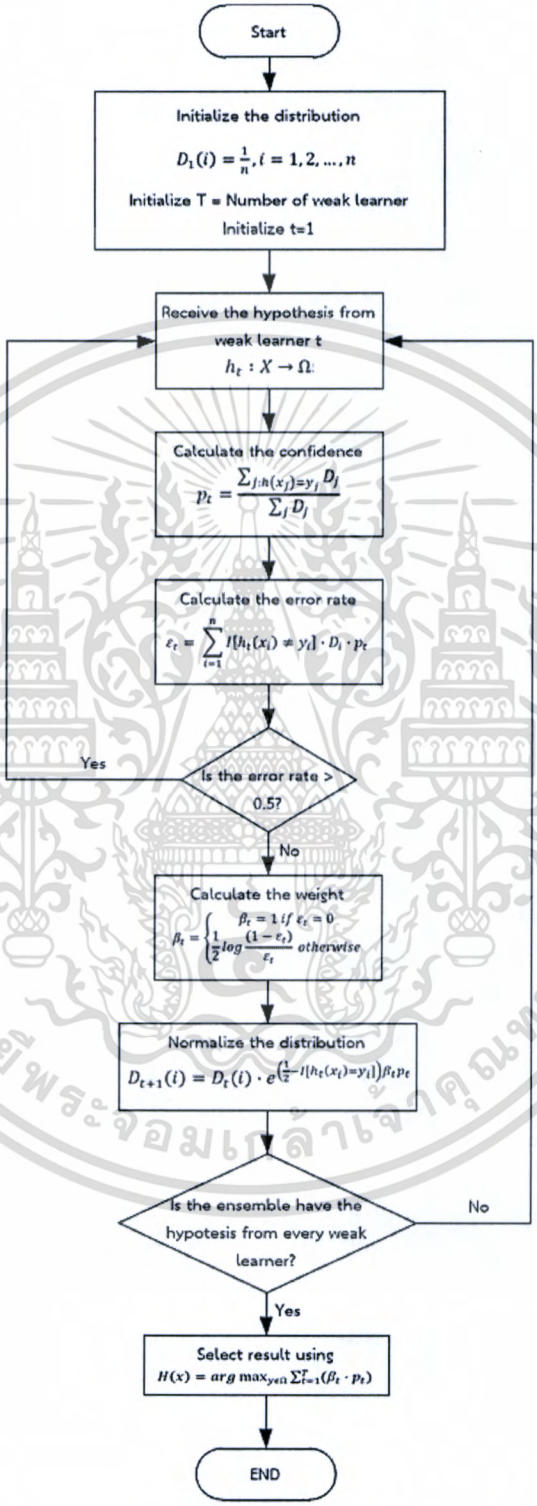
10. เปรียบเทียบค่า t กับ T โดยให้วนรอบทำซ้ำตั้งแต่ขั้นตอนที่ 3 จนกว่า t จะมีค่ามากกว่า T

ข้อมูลนำออก: ให้ทำการเลือกสมมติฐานที่มีผลคูณระหว่างค่าน้ำหนักกับค่าความเชื่อมั่นสูงที่สุด มาเป็นแบบจำลองในการจำแนกประเภทข้อมูล โดยแสดงได้ในสมการที่ (3.6)

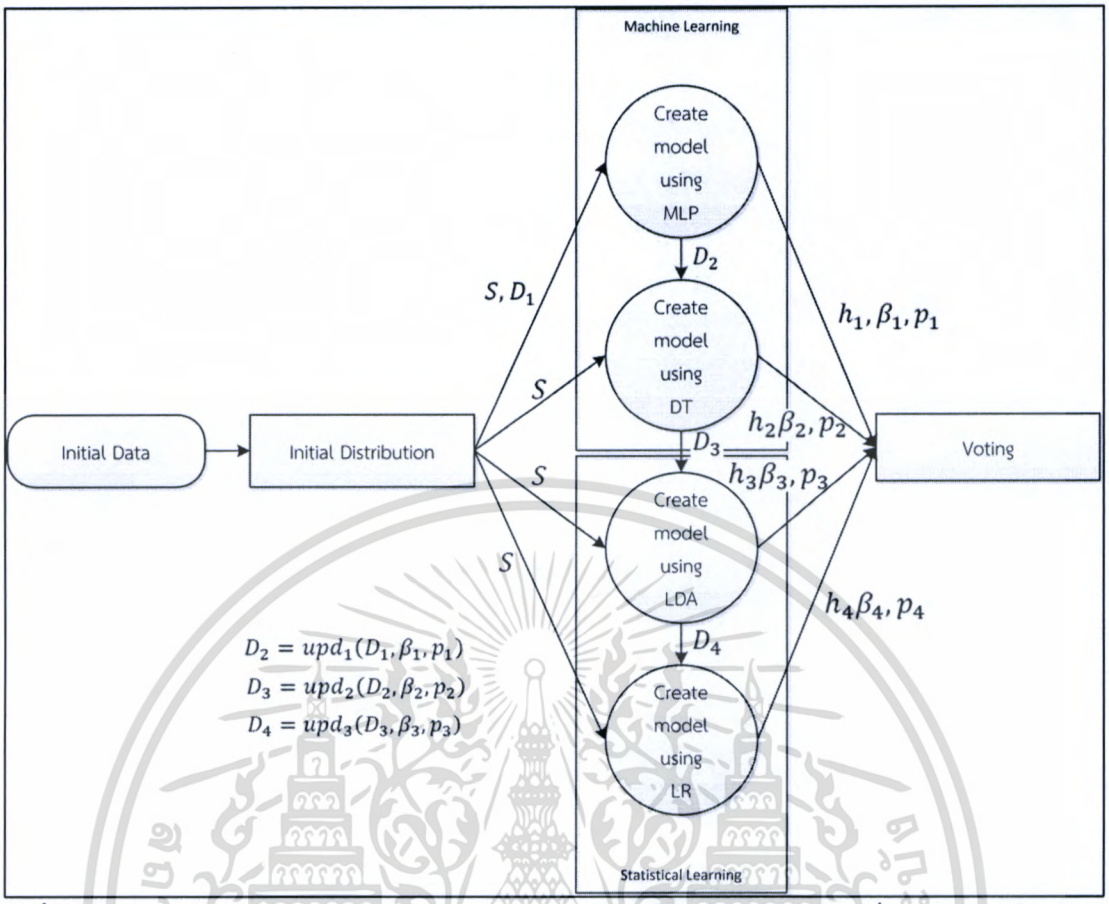
$$H(x) = arg \max_{y \epsilon \Omega} \sum_{t=1}^{T} (\beta_t \cdot p_t) \tag{3.6}$$

ในขั้นตอนการนำข้อมูลเข้ากำหนดให้ S เป็นชุดข้อมูลสำหรับสอนตัวจำแนกประเภทโดยมีสมาชิกเป็น x_i ที่เป็นเวกเตอร์ข้อมูลนำเข้าและ y_i ซึ่งเป็นเวกเตอร์กลุ่มของข้อมูล สำหรับทุก ๆ ตัวจำแนกประเภทพื้นฐาน (Based classifier) ตั้งแต่ตัวที่ 1 จนถึงตัวที่ T ให้ทำการกำหนดค่าเริ่มต้นให้กับค่าการแจกแจง D โดยให้มีค่าเท่ากับ $1/N$ กับสมาชิกทุกตัว (ตั้งแต่ตัวที่ 1 จนถึงตัวที่ N) จากนั้นให้เรียกการทำงานของ WeakLearn โดยให้ข้อมูลสำหรับการสอน S และค่าการแจกแจง D_t เข้าไป ซึ่งจะได้ผลลัพธ์ออกมาเป็นสมมติฐาน (Hypothesis) ของประเภทข้อมูลที่ได้ส่งเข้าไปใน WeakLearn ในขั้นตอนต่อมาให้คำนวณหาค่าความความเชื่อมั่นโดยคำนวณได้จากการหาผลรวมของค่าการแจกแจงของทุก ๆ i ที่มีประเภทตรงกันกับสมมติฐานดังสมการที่ (3.2) จากนั้นใช้ค่าความเชื่อมั่นที่คำนวณได้มาคำนวณหาค่าความผิดพลาดต่อดังในสมการที่ (3.3) ซึ่งเป็นผลรวมของผลคูณระหว่างค่าความเชื่อมั่นกับค่าการแจกแจงของข้อมูลที่มีประเภทไม่ตรงกับสมมติฐานที่ได้ ต่อมาในขั้นตอนที่ 7 เป็นการตรวจค่าความผิดพลาดว่าตัวจำแนกประเภทพื้นฐานตัวนั้น (ตัวที่ t ใด ๆ) มีความผิดพลาดมากกว่าการความผิดพลาดจากการสุ่มโดยทั่วไปหรือไม่ ซึ่งกำหนดไว้ที่ร้อยละ 50 หากมีค่าความผิดพลาดสูงกว่าก็ให้ย้อนกลับไปเรียนรู้ใหม่จนกว่าค่าความผิดพลาดจะมีค่าน้อยกว่าร้อยละ 50 ในขั้นตอนถัดมาเป็นขั้นตอนการหาค่าน้ำหนักซึ่งจะเป็นค่าที่ส่งผลในตอนนับคะแนนเสียงในการเลือกสมมติฐานการจำแนกประเภท โดยหากมีค่าน้ำหนักที่มาก นั่นหมายความว่า WeakLearn ตัว

นั้นยิ่งมีความน่าเชื่อถือสูงตามไปด้วย สุดท้ายเป็นการปรับปรุงและลดทอนค่าน้ำหนักเพื่อไม่ให้มีช่วงที่กว้างจนเกินไป สำหรับผังงาน (Flowchart) ขั้นตอนการทำงานสามารถแสดงได้ดังรูปที่ 3.1



รูปที่ 3.1 แผนผังการทำงานการรวมกลุ่มผสมระหว่างการเรียนรู้ของเครื่องและการเรียนรู้เชิงสถิติโดยใช้การบูสต์ด้วยค่าความเชื่อมั่น



รูปที่ 3.2 แผนภาพจำลองการทำงานการรวมกลุ่มผสมระหว่างการเรียนรู้ของเครื่องและการเรียนรู้เชิงสถิติโดยใช้การบูสต์ด้วยค่าความเชื่อมั่น

เนื่องจากงานวิจัยนี้ได้ใช้เทคนิคการจำแนกประเภทด้วยวิธีการใช้โครงข่ายประสาทเทียมเพอร์เซ็ปตรอนหลายชั้น ต้นไม้ตัดสินใจ การวิเคราะห์เชิงเส้นแบบดิสคริมิแนนท์ และการวิเคราะห์การถดถอยแบบโลจิสติก ซึ่งเป็นเทคนิคที่นำมาจากเทคนิคการเรียนรู้ของเครื่องและการคำนวณเชิงสถิติมาใช้เป็นตัวจำแนกประเภทพื้นฐานดังรูปที่ 3.2 โดยจะใช้เทคนิคการการจำแนกประเภทด้วยโครงข่ายประสาทเทียมเพอร์เซ็ปตรอนเป็นรอบการคำนวณครั้งแรก เมื่อคำนวณเสร็จก็จะได้ผลลัพธ์เป็นสมมติฐาน h_1 ค่าน้ำหนัก β_1 และค่าความเชื่อมั่น p_t จากนั้นจะทำการคำนวณเพื่อทำการปรับค่าการแจกแจงของ D_2 ของทุก i จากสมการที่ (3.5) จากนั้นในการเรียนรู้ครั้งที่ 2 ก็จะมีขั้นตอนเช่นเดียวกันกับในครั้งที่ 1 เพียงแต่จะทำการเรียนรู้การจำแนกประเภทด้วยเทคนิคต้นไม้ตัดสินใจเท่านั้น ส่วนในครั้งที่ 3 และครั้งที่ 4 จะใช้เทคนิคการวิเคราะห์เชิงเส้นแบบดิสคริมิแนนท์และการวิเคราะห์การถดถอยแบบโลจิสติกมาเป็นวิธีการเรียนรู้การจำแนกตามลำดับ สุดท้ายเมื่อทำงานจนครบทุกเทคนิคแล้ว จึงเอาสมมติฐานที่ได้ทั้งหมดเข้าสู่ขั้นตอนการโหวต โดยในขั้นตอนนี้จะทำการเลือกสมมติฐานที่ได้จากตัวจำแนกประเภทพื้นฐานที่มีค่าน้ำหนักมากที่สุด

3.3 ขั้นตอนการวัดประสิทธิภาพ

ในงานวิจัยนี้ได้นำชุดข้อมูลจากเว็บไซต์ UCI มาใช้ในการวัดประสิทธิภาพการเรียนรู้การจำแนกประเภทโดยได้เลือกใช้โปรแกรม MATLAB เวอร์ชัน R2015a ในสร้างแบบจำลองสำหรับการทดลอง ทั้งยังใช้ในการลดทอนจำนวนชุดข้อมูล PAMAP2 ด้วย ซึ่งเมื่อได้ทำการจัดเตรียมข้อมูลสำหรับทำการทดลองเรียบร้อยแล้ว ผู้วิจัยจึงได้นำเอาชุดข้อมูลต่าง ๆ ที่เตรียมไว้มาวัดประสิทธิภาพด้วยการให้เครื่องสร้างแบบจำลองการเรียนรู้และนำแบบจำลองการเรียนรู้ที่ได้มาทดสอบวัดผลความถูกต้อง (Accuracy)

ขั้นตอนการวัดความถูกต้องจะนำชุดข้อมูลแต่ละชุดมาทำการทดลองทั้งหมด 5 ครั้ง โดยในแต่ละครั้งชุดข้อมูลจะถูกแบ่งใช้สำหรับทำการเรียนรู้และใช้ทดสอบความถูกต้องโดยการใช้เทคนิค k-fold cross validation ซึ่งเป็นเทคนิคที่ทำการสุ่มข้อมูลจากชุดข้อมูลขึ้นมาแบ่งเป็นจำนวน k ส่วน โดยแบ่งเป็นส่วนชุดข้อมูลที่ใช้ในการเรียนรู้จำนวน k - 1 ส่วน และส่วนข้อมูลที่ใช้ในการทดสอบจำนวน 1 ส่วน โดยจะมีการวนรอบการทดสอบเพื่อเปลี่ยนข้อมูลส่วนที่ใช้ในการทดสอบไปจนครบข้อมูลทุกส่วน ซึ่งในงานวิจัยนี้จะใช้ค่า k เท่ากับ 3 ซึ่งก็คือการแบ่งข้อมูลออกเป็น 3 ส่วน โดยที่ 1 ส่วนจะถูกนำไปใช้ในการทดสอบ และอีก 2 ส่วนที่เหลือจะถูกในการใช้สำหรับการเรียนรู้ อีกทั้งจะมีการวนจำนวนรอบการทดสอบเพื่อเปลี่ยนส่วนข้อมูลที่ใช้สำหรับทดสอบ จนข้อมูลทุกส่วนได้ถูกนำไปใช้ในการทดสอบ จากนั้นนำผลความถูกต้องในการจำแนกประเภทที่ทดสอบแต่ละรอบมาหาค่าเฉลี่ยความถูกต้องรวมจึงถือว่าจบการทดลอง 1 ครั้ง ซึ่งในงานวิจัยนี้จะทำการทดลองเช่นนี้จนครบ 5 ครั้ง แล้วจึงหาค่าเฉลี่ยจากการทดลองทั้ง 5 ครั้งมาแสดงผล

เนื่องจากงานวิจัยนี้เป็นงานวิจัยที่ใช้เทคนิคการเรียนรู้ของเครื่องและเทคนิคการคำนวณเชิงสถิติมาผสมรวมกันด้วยเทคนิคการเรียนรู้โครงข่ายประสาทเทียมเพอร์เซ็ปตรอนหลายชั้น การเรียนรู้ต้นไม้ตัดสินใจ การวิเคราะห์เชิงเส้นแบบดิสคริมิแนนท์ การวิเคราะห์การถดถอยแบบโลจิสติก อีกทั้งยังเป็นงานวิจัยที่มุ่งเน้นการปรับปรุงพัฒนาขั้นตอนวิธีการเรียนรู้แบบรวมกลุ่ม ดังนั้นในขั้นตอนการวัดประสิทธิภาพของงานวิจัยนี้จึงถูกแบ่งออกเป็น 2 ส่วนใหญ่ ได้แก่ การทดลองเพื่อทดสอบประสิทธิภาพการเรียงสับเปลี่ยนของตัวเรียนรู้พื้นฐานและการทดสอบประสิทธิภาพเพื่อเปรียบเทียบกับวิธีการเรียนรู้แบบรวมกลุ่มวิธีอื่น

การทดลองเพื่อทดสอบประสิทธิภาพการเรียงสับเปลี่ยนของตัวเรียนรู้พื้นฐานจะเป็นการนำชุดข้อมูลที่ได้เตรียมไว้มาใช้ทดลอง เพื่อหาการเรียงสับเปลี่ยนระหว่างตัวเรียนรู้พื้นฐานต่าง ๆ โดยจะเป็นการเรียงกันด้วยตัวเรียนรู้พื้นฐานที่แตกต่างกัน 4 แบบ ดังที่ได้กล่าวมาในย่อหน้าที่แล้ว ซึ่งคิดเป็นจำนวนการเรียงกันทั้งสิ้น 24 แบบ เพื่อหาคำตอบว่าการเรียงสับเปลี่ยนในรูปแบบต่าง ๆ นั้นมีผลต่อความถูกต้องในการจำแนกประเภทที่ต่างกันหรือไม่ อย่างไร และวิธีการเรียงกันแบบใดที่สามารถให้ผลลัพธ์ในการจำแนกประเภทได้ดีที่สุด และเนื่องจากงานวิจัยชิ้นนี้มุ่งเน้นที่การพัฒนาขั้นตอนวิธีการ

เรียนรู้แบบรวมกลุ่ม ดังนั้นในการทดสอบวัดประสิทธิภาพส่วนที่ 2 ผู้วิจัยจึงได้นำวิธีการเรียนรู้แบบรวมกลุ่มเอดาบูสต์เอ็มวันและการเรียนรู้แบบรวมกลุ่มเอ็มอาร์เอ็นมาใช้เปรียบเทียบประสิทธิภาพความถูกต้อง โดยผู้วิจัยจะใช้ต้นไม้ตัดสินใจซีโฟร์พอยต์ไฟว์มาเป็น WeakLearn ให้กับการเรียนรู้แบบรวมกลุ่มเอดาบูสต์เอ็มวัน และใช้เทคนิคการเรียนรู้โครงข่ายประสาทเทียมเพอร์เซ็ปตรอนหลายชั้น เทคนิคเรเดียลเบสิสฟังก์ชัน (Radial basis function: RBF) เทคนิคการเรียนรู้ด้วยตัวจำแนกแบบเบย์อย่างง่าย (Naïve Bayes classifier) มาใช้ในการเรียนรู้แบบรวมกลุ่มเอ็มอาร์เอ็น

สำหรับการคำนวณหาค่าความถูกต้องจะคำนวณจากอัตราส่วนร้อยละของจำนวนของข้อมูลที่ได้ผลลัพธ์การจำแนกประเภทที่ถูกต้องต่อจำนวนของชุดข้อมูลทดสอบทั้งหมดตามสมการที่ (3.7)

$$\text{ค่าความถูกต้อง} = \frac{\text{จำนวนของข้อมูลที่ได้ผลลัพธ์การจำแนกประเภทที่ถูกต้องจากชุดข้อมูลทดสอบ}}{\text{จำนวนข้อมูลของชุดข้อมูลทดสอบทั้งหมด}} \qquad (3.7)$$

บทที่ 4

การทดลองและผลการทดลอง

ในบทนี้จะกล่าวถึงการทดลองและผลการทดลองต่าง ๆ ที่ได้ทั้งจากการวัดประสิทธิภาพจากการวัดประสิทธิภาพการรวมกลุ่มผสมระหว่างการเรียนรู้ของเครื่องและการเรียนรู้เชิงสถิติโดยใช้การบูสต์ด้วยค่าความเชื่อมั่นกับวิธีการเรียนรู้แบบรวมกลุ่มเอดาบูสต์เอ็มวันและการเรียนรู้แบบรวมกลุ่มเอ็มอาร์เอ็น โดยจะมีการใช้ชุดข้อมูลที่มีการจำแนกประเภททั้งแบบสองประเภทและแบบหลายประเภทดังที่ได้กล่าวมาในบทที่แล้ว

4.1 การทดลอง

4.1.1 โปรแกรมที่ใช้และการกำหนดค่าพารามิเตอร์

การทดลองในงานวิจัยนี้ได้ใช้โปรแกรม MATLAB ในการทำการทดลองวัดประสิทธิภาพการเรียนรู้แบบรวมกลุ่มโดยได้นำเอาเทคนิคการจำแนกประเภทโครงข่ายประสาทเทียมเพอร์เซ็ปตรอนหลายชั้น ต้นไม้ตัดสินใจ การวิเคราะห์เชิงเส้นแบบดิสคริมิแนนท์ และการวิเคราะห์การถดถอยแบบโลจิสติกมาใช้ ซึ่งแต่ละเทคนิคมีการใช้ค่าพารามิเตอร์ในการเรียกใช้คำสั่งใน MATLAB ดังนี้

1) โครงข่ายประสาทเทียมเพอร์เซ็ปตรอนหลายชั้น ใช้ฟังก์ชัน trainscg เป็นฟังก์ชันการคำนวณย้อนกลับโดยกำหนด epochs ไว้ที่ 1000 goal เท่ากับ 0.01 และตั้งค่าการทำ feed forward ด้วยการกำหนดชั้นซ่อน 10 ชั้น และใช้ค่าปริยายของโปรแกรมเป็นค่าพารามิเตอร์ที่เหลือ
2) ต้นไม้ตัดสินใจ ใช้ฟังก์ชัน fitctree ของโปรแกรม MATLAB ในการสร้างต้นไม้ตัดสินใจซึ่งใช้ค่าปริยายเป็นค่าพารามิเตอร์
3) การวิเคราะห์เชิงเส้นแบบดิสคริมิแนนท์ ใช้ฟังก์ชัน fitcdiscr ของโปรแกรม MATLAB ในการคำนวณหาสมการเชิงเส้นแบบดิสคริมิแนนท์ และกำหนดให้ประเภทการถดถอย (Discriminant Type) เป็น pseudoLinear
4) การวิเคราะห์การถดถอยแบบโลจิสติก ใช้ชุดฟังก์ชันการคำนวณจากโปรแกรม WEKA ด้วยการเรียกผ่านโปรแกรม MATLAB โดยกำหนดค่าปริยายเป็นค่าพารามิเตอร์

4.2 ผลการทดลอง

การทดลองในงานวิจัยนี้จะแบ่งการทดลองออกเป็น 2 ส่วนที่สำคัญ ได้แก่ ส่วนการทดลองเพื่อทดสอบประสิทธิภาพการเรียงสับเปลี่ยนตัวเรียนรู้พื้นฐานและส่วนการทดสอบประสิทธิภาพเพื่อ

เปรียบเทียบกับวิธีการเรียนรู้แบบรวมกลุ่มอื่น ๆ โดยผลการทดลองดังกล่าวสามารถอธิบายได้ดังต่อไปนี้

4.2.1 การทดลองเพื่อทดสอบประสิทธิภาพการเรียงสับเปลี่ยนของตัวเรียนรู้พื้นฐาน

การทดลองในส่วนนี้จะเป็นการทดลองเพื่อค้นหาว่าการสลับสับเปลี่ยนการจัดเรียงกันของตัวเรียนรู้พื้นฐานมีผลต่อความถูกต้องในการจำแนกประเภทหรือไม่ และหากลำดับการเรียงตัวของตัวเรียนรู้พื้นฐานมีผลต่อความถูกต้องแล้วจะมีลำดับหรือวิธีการเรียงกันแบบใดที่ให้ผลการจำแนกได้ถูกต้องที่สุด

เนื่องจากในการทดลองนี้จะทำการทดสอบการเรียงกันของตัวเรียนรู้พื้นฐานจำนวน 4 ตัวด้วยกัน ได้แก่ ตัวเรียนรู้พื้นฐานที่ใช้เทคนิคโครงข่ายประสาทเทียมเพอร์เซ็ปตรอนหลายชั้น ตัวเรียนรู้พื้นฐานที่ใช้เทคนิคการเรียนรู้ต้นไม้ตัดสินใจ ตัวเรียนรู้พื้นฐานที่ใช้เทคนิคการวิเคราะห์เชิงเส้นแบบดิสคริมิแนนท์ และตัวเรียนรู้พื้นฐานที่ใช้เทคนิคการถดถอยแบบโลจิสติก ซึ่งจะเขียนแทนด้วยอักษรภาษาอังกฤษ M T D และ L ตามลำดับ เพื่อให้ตารางมีขนาดที่ไม่ใหญ่โตมาก โดยการเรียงจะเป็นการเรียงกันโดยไม่ซ้ำเทคนิคกันซึ่งมีความเป็นไปได้ทั้งหมดเท่ากับ 4! หรือ 24 แบบ และจะทำการทดลองกับชุดข้อมูลจำนวน 6 ชุดด้วยกัน นอกจากนี้การแสดงผลการทดลองในตารางต่าง ๆ นั้น จะทำการจัดเรียงตามอันดับที่สามารถให้ค่าเฉลี่ยความถูกต้องจากการคำนวณทั้งหมด 5 รอบ จากค่ามากที่สุดไปหาค่าเฉลี่ยความถูกต้องน้อยที่สุด ซึ่งสามารถแสดงผลการทดลองได้ดังต่อไปนี้

4.2.1.1 การทดลองด้วยชุดข้อมูล Banknote Authentication

การทดลองกับชุดข้อมูล Banknote Authentication เป็นการทดลองกับชุดข้อมูลที่มีการจำแนกประเภทแบบไบนารีหรือมีประเภทข้อมูลทั้งหมด 2 ประเภท โดยผลการทดลองที่ได้แสดงให้เห็นว่าการจัดเรียงตัวเรียนรู้พื้นฐานแบบต่าง ๆ นั้น ให้ผลที่แตกต่างกันน้อยมาก โดยการเรียงกันแบบแบบ DTML จะให้ผลลัพธ์ความถูกต้องเฉลี่ยดีที่สุดที่ร้อยละ 99.606 ส่วนสองอันดับรองลงมาจะเป็นการเรียงกันแบบ LMDT และแบบ MTDL ที่ความถูกต้องเฉลี่ยร้อยละ 99.548 เท่ากัน ส่วนแบบที่ได้ที่ความถูกต้องเฉลี่ยน้อยที่สุดคือการเรียงกันแบบ MDTL ที่ความถูกต้องเฉลี่ยร้อยละ 99.228 ดังในตารางที่ 4.1

นอกจากนี้ผลการเปรียบเทียบประสิทธิภาพความถูกต้องของการเรียงกันด้วยตัวเรียนรู้พื้นฐานแบบเดียวกันจำนวน 4 ตัว หรือใช้เทคนิคการเรียนรู้เพียงเทคนิคเดียวเป็นตัวเรียนรู้พื้นฐานและให้ทำการวนรอบการเรียนรู้ 4 รอบ ซึ่งผลค่าความถูกต้องเฉลี่ยที่ได้จากการใช้เทคนิคการเรียนรู้ต้นไม้ตัดสินใจอย่างเดียว (TTTT) จะให้ค่าความถูกต้องเฉลี่ยน้อยที่สุดที่ร้อยละ 97.52 ในขณะที่การใช้เทคนิคการวิเคราะห์เชิงเส้นแบบดิสคริมิแนนท์เพียงอย่างเดียว (DDDD) และการใช้เทคนิคการวิเคราะห์การถดถอยแบบโลจิสติก (LLLL) ให้ค่าความถูกต้องเฉลี่ยที่ร้อยละ 97.524 และร้อยละ

98.73 ตามลำดับ ซึ่งทั้ง 3 แบบนี้มีค่าความถูกต้องเฉลี่ยน้อยกว่าการเรียงแบบผสมทั้งสิ้น อย่างไรก็ตามการใช้เทคนิคโครงข่ายประสาทเทียมเพียงอย่างเดียว (MMMM) มีความถูกต้องเฉลี่ยที่ร้อยละ 99.358 ซึ่งมากกว่าการใช้เทคนิคเดียวเรียงกันทั้ง 3 แบบดังที่ได้กล่าวมาแล้วก่อนหน้านี้ และมีค่าความถูกต้องเฉลี่ยมากกว่าการเรียนรู้แบบผสมที่มีการจัดเรียงแบบ MDTL MTLD LDMT TLDM MDLT เพียง 5 แบบเท่านั้น จึงน่าจะกล่าวได้ว่าการเรียนรู้แบบผสมสามารถให้ผลความถูกต้องในการจำแนกประเภทได้มากกว่าการเรียนรู้ที่ใช้เทคนิคเดียวเป็นตัวเรียนรู้พื้นฐาน

ตารางที่ 4.1 เปรียบเทียบประสิทธิภาพการจำแนกประเภทด้วยการสับเปลี่ยนลำดับการเรียงของตัวเรียนรู้พื้นฐานโดยใช้ชุดข้อมูล Banknote Authentication

อันดับที่	การเรียงตัวจำแนก	รอบการคำนวณ					ค่าเฉลี่ยความถูกต้อง	ค่าเบี่ยงเบนมาตรฐาน
		1	2	3	4	5		
1	DTML	99.27	99.85	99.71	99.49	99.71	99.606	0.204
2	LMDT	99.49	99.49	99.64	99.27	99.85	99.548	0.192
3	MTDL	99.49	99.64	99.78	99.27	99.56	99.548	0.169
4	TMLD	99.56	99.78	99.49	99.2	99.56	99.518	0.187
5	DMLT	99.49	99.34	99.42	99.71	99.49	99.49	0.123
6	LTDM	99.13	99.78	99.71	99.42	99.34	99.476	0.240
7	MLTD	99.27	99.85	99.05	99.71	99.49	99.474	0.289
8	DTLM	99.56	99.42	99.64	99.34	99.34	99.46	0.121
9	TMDL	99.71	99.34	99.71	99.2	99.34	99.46	0.210
10	MLDT	99.2	99.34	99.78	99.71	99.2	99.446	0.250
11	DMTL	99.34	99.27	99.34	99.56	99.64	99.43	0.143
12	TLMD	99.49	99.71	99.2	99.13	99.56	99.418	0.220
13	LTMD	99.34	99.2	99.56	99.56	99.42	99.416	0.137
14	DLTM	99.27	99.27	99.56	99.42	99.49	99.402	0.117
15	TDLM	98.98	99.49	99.27	99.56	99.71	99.402	0.254
16	LMTD	99.27	99.71	99.34	99.49	99.2	99.402	0.181
17	TDML	99.34	99.27	99.49	99.56	99.27	99.386	0.118
18	DLMT	99.49	99.49	99.71	98.98	99.2	99.374	0.255
19	LDTM	99.42	99.49	99.2	99.56	99.2	99.374	0.149
20	MMMM	99.05	99.34	99.42	99.56	99.42	99.358	0.170
21	MDLT	99.27	99.2	99.78	99.2	99.05	99.3	0.251
22	TLDM	99.27	99.27	99.27	99.05	99.42	99.256	0.118
23	MTLD	99.13	99.2	99.78	99.56	98.54	99.242	0.424
24	LDMT	99.27	99.34	99.42	98.98	99.2	99.242	0.150
25	MDTL	99.49	99.27	99.49	99.27	98.62	99.228	0.320
26	LLLL	98.83	98.54	98.76	98.69	98.83	98.73	0.108
27	DDDD	97.52	97.38	97.67	97.38	97.67	97.524	0.130
28	TTTT	98.54	96.72	98.03	96.79	97.52	97.52	0.703

4.2.1.2 การทดลองด้วยชุดข้อมูล Connectionist Bench

การทดลองด้วยชุดข้อมูลนี้จะเป็นการทดลองเพื่อวัดประสิทธิภาพการจำแนกประเภทข้อมูลแบบไบนารีอีกหนึ่งชุดข้อมูล ซึ่งผลค่าความถูกต้องเฉลี่ยที่ได้ก็มีความแตกต่างที่น้อยมากเช่นเดียวกัน โดยการจัดเรียงแบบ TDLM จะมีค่าความถูกต้องเฉลี่ยสูงที่สุดที่ร้อยละ 78.456 รองลงมาคือ LTMD และ MDTL ที่ค่าความถูกต้องเฉลี่ยร้อยละ 77.878 และร้อยละ 77.776 ตามลำดับ ส่วนการเรียนรู้แบบผสมที่มีการเรียงกันของตัวเรียนรู้พื้นฐานที่มีค่าความถูกต้องน้อยที่สุดที่ร้อยละ 73.166 คือ DMTL

ในขณะที่การใช้เทคนิคการจำแนกประเภทที่ใช้เพียงเทคนิคเดียวเป็นตัวเรียนรู้พื้นฐาน และมีการกำหนดรอบการเรียนรู้ที่ 4 รอบ อาทิ การใช้เทคนิคต้นไม้ตัดสินใจเพียงอย่างเดียว การใช้เทคนิคการวิเคราะห์เชิงเส้นแบบดิสคริมิแนนท์เพียงอย่างเดียว และการใช้การวิเคราะห์การถดถอยเพียงอย่างเดียวยังคงให้ผลลัพธ์ค่าความถูกต้องเฉลี่ยที่น้อยกว่าการเรียนรู้แบบผสมที่ร้อยละ 68.748 ร้อยละ 72.304 และร้อยละ 72.97 ตามลำดับ โดยเรียงจากน้อยสุดไปมาก แต่การใช้เทคนิคโครงข่ายประสาทเทียมกลับสามารถให้ค่าความถูกต้องเฉลี่ยในการจำแนกประเภทที่ร้อยละ 77.612 ซึ่งมีค่ามากรองจากการใช้วิธีการจัดเรียงแบบผสม MDTL ที่มีค่าความถูกต้องเฉลี่ยสูงเป็นอันดับที่ 3 โดยผลการการทดลองทั้งหมดสามารถแสดงได้ดังตารางที่ 4.2

ตารางที่ 4.2 เปรียบเทียบประสิทธิภาพการจำแนกประเภทด้วยการสับเปลี่ยนลำดับการเรียงของตัวเรียนรู้พื้นฐานโดยใช้ชุดข้อมูล Connectionist Bench

อันดับที่	การเรียงตัวจำแนก	รอบการคำนวณ					ค่าเฉลี่ยความถูกต้อง	ค่าเบี่ยงเบนมาตรฐาน
		1	2	3	4	5		
1	TLDM	81.24	76.44	77.87	80.3	76.43	78.456	1.983
2	LTMD	80.3	78.34	74.02	82.71	74.02	77.878	3.441
3	MDTL	77.42	76.41	76.91	78.82	79.32	77.776	1.115
4	MMMM	78.86	77.41	76.47	76.47	78.85	77.612	1.071
5	DLMT	79.33	79.3	78.85	72.15	76.45	77.216	2.747
6	MTLD	75.96	79.32	73.57	78.37	77.94	77.032	2.049
7	DTML	76.43	76.44	76.95	75.48	77.88	76.636	0.783
8	LMDT	73.54	74.48	75.48	77.88	81.73	76.622	2.935
9	MLTD	77.9	75.03	75	77.43	77.43	76.558	1.272
10	DLTM	74.56	76.92	76.89	80.29	72.6	76.252	2.583
11	LTDM	75.49	75.97	79.8	70.19	79.8	76.25	3.538
12	TDLM	79.78	79.82	73.55	73.57	74.51	76.246	2.923
13	MTDL	75.49	76.5	74.48	76.92	77.43	76.164	1.056
14	TMDL	74.51	75.94	78.38	74.56	77.41	76.16	1.538
15	DMLT	77.41	74.98	74.02	77.41	75.99	75.962	1.336
16	TLMD	74.03	77.38	76	78.36	74.03	75.96	1.745
17	LMTD	75.93	79.8	73.55	69.21	79.83	75.664	4.016
18	TDML	76.91	75.46	76.88	79.32	69.7	75.654	3.225
19	MDLT	70.68	75.49	78.38	72.14	79.3	75.198	3.370
20	TMLD	77.92	72.12	75.51	73.6	75.94	75.018	1.996
21	MLDT	77.43	77.87	73.55	77.93	67.31	74.818	4.094
22	LDTM	72.64	79.84	72.13	71.13	76.48	74.444	3.251
23	LDMT	73.06	72.64	75.98	75.47	73.57	74.144	1.334
24	DTLM	75.49	74.56	71.16	73.06	74.04	73.662	1.477
25	DMTL	73.08	69.68	75.02	71.62	76.43	73.166	2.393
26	LLLL	72.12	75.49	72.56	73.08	71.6	72.97	1.351
27	DDDD	71.64	68.74	74.5	75	71.64	72.304	2.266
28	TTTT	68.78	74.04	65.89	64.83	70.2	68.748	3.275

4.2.1.3 การทดลองด้วยชุดข้อมูล Iris

ชุดข้อมูล Iris มีจำนวนประเภทข้อมูลอยู่ 3 ประเภท จึงเหมาะในการนำใช้ทดสอบการจำแนกประเภทที่มีประเภทมากกว่า 2 หรือการจำแนกแบบหลายคลาสได้ ซึ่งจากผลการทดลองวัดประสิทธิภาพการสับเปลี่ยนลำดับของตัวเรียนรู้พื้นฐานด้วยการใช้ชุดข้อมูล Iris ดังแสดงในตารางที่ 4.3 การจัดเรียงตัวเรียนรู้พื้นฐานที่ให้ผลการจำแนกประเภทได้ดีที่สุดคือการจัดเรียงโดยใช้เทคนิคการวิเคราะห์เชิงเส้นแบบดิสคริมิแนนท์เพียงเทคนิคเดียวจำนวน 4 ตัว ซึ่งได้ค่าความถูกต้องเฉลี่ยร้อยละ 98 โดยมีการจัดเรียงแบบผสม LMDT และ DMTL เป็นแบบการจัดเรียงที่ได้ค่าความถูกต้องเฉลี่ยที่รองลงมาที่ร้อยละ 97.72 และร้อยละ 97.62 ตามลำดับ และถัดมาไม่ห่างมากในอันดับที่ 6 ที่ใช้เทคนิคการวิเคราะห์การถดถอยแบบโลจิสติกเพียงเทคนิคเดียวเป็นตัวเรียนรู้พื้นฐานได้ค่าความถูกต้องที่ 97.22 แต่ในขณะเดียวกันแบบการจัดเรียงที่ใช้เทคนิคต้นไม้ตัดสินใจเพียงเทคนิคเดียวเป็นตัวเรียนรู้พื้นฐานกลับให้ค่าความถูกต้องเฉลี่ยที่ต่ำที่สุดที่ร้อยละ 94.39 ส่วนการจัดเรียงที่ใช้เทคนิคโครงข่ายประสาทเทียมเพอร์เซ็ปตรอนหลายชั้นเพียงอย่างเดียวนั้นก็ให้ผลที่อยู่ในลำดับท้าย ๆ เช่นกัน โดยเป็นรูปแบบที่ทำอันดับความถูกต้องเฉลี่ยได้ลำดับที่ 24 จากรูปแบบการจัดเรียงทั้งหมด 28 รูปแบบที่ร้อยละ 96.136 และเมื่อสังเกตดูการเรียงลำดับที่เป็นการเรียงโดยใช้เทคนิคผสมอื่น ๆ ประกอบแล้วจะเห็นได้ว่า วิธีการจัดเรียงลำดับที่นำเอาเทคนิคการคำนวณเชิงสถิติมาเป็นตัวเรียนรู้พื้นฐานในลำดับแรก ๆ นั้น มีแนวโน้มที่จะให้ค่าความถูกต้องที่สูงกว่าการเรียงโดยใช้เทคนิคการเรียนรู้ของเครื่องมาเป็นตัวเรียนรู้พื้นฐานตัวแรก ๆ ซึ่งสามารถอนุมานได้ว่าการใช้เทคนิคการวิเคราะห์เชิงเส้นแบบดิสคริมิแนนท์และการใช้เทคนิคการวิเคราะห์การถดถอยแบบโลจิสติกซึ่งทั้งสองจัดเป็นเทคนิคการจำแนกประเภทเชิงสถิตินั้น มีความเหมาะสมในการจำแนกประเภทข้อมูลชุดนี้มากกว่าการใช้เทคนิคโครงข่ายประสาทเทียมเพอร์เซ็ปตรอนหลายชั้นและการเรียนรู้ต้นไม้ตัดสินใจที่จัดเป็นเทคนิคการเรียนรู้ของเครื่อง

ตารางที่ 4.3 เปรียบเทียบประสิทธิภาพการจำแนกด้วยการสับเปลี่ยนลำดับการเรียงของตัวเรียนรู้พื้นฐานโดยใช้ชุดข้อมูล Iris

อันดับที่	การเรียงตัวจำแนก	รอบการคำนวณ					ค่าเฉลี่ยความถูกต้อง	ค่าเบี่ยงเบนมาตรฐาน
		1	2	3	4	5		
1	DDDD	98	98	98.01	98.01	98	98.004	0.005
2	LMDT	97.33	97.31	98	97.99	98	97.726	0.332
3	DMTL	98.04	98	97.36	98.03	96.67	97.62	0.540
4	DLTM	98.04	98	97.35	96.68	98.01	97.616	0.535
5	DTLM	98	97.35	96.67	97.35	97.33	97.34	0.421
6	LLLL	97.36	98.01	98.04	96.67	96.04	97.224	0.776
7	MTDL	97.35	97.31	97.36	97.33	96.71	97.212	0.252
8	LTMD	96.67	97.99	97.33	97.33	96.65	97.194	0.498
9	MDLT	97.33	97.33	97.29	97.33	96.67	97.19	0.260
10	DTML	97.33	97.35	96	96.68	98.03	97.078	0.688
11	LDTM	98.03	98	98.01	95.33	96	97.074	1.170
12	LMTD	96.67	96.68	96.71	97.33	97.37	96.952	0.325
13	DMLT	96.64	95.41	97.99	98	96.67	96.942	0.973
14	DLMT	95.33	96.67	97.97	96.67	98.03	96.934	0.999
15	LDMT	98.01	98	97.32	96.69	94.61	96.926	1.257
16	TDML	97.33	96.67	95.33	98.04	96.69	96.812	0.895
17	LTDM	95.29	97.97	97.36	98.04	94.67	96.666	1.410
18	MTLD	96.67	97.33	97.33	95.97	95.96	96.652	0.611
19	MDTL	97.33	96	96.67	95.37	97.33	96.54	0.765
20	TLMD	96.67	94.66	97.33	98	95.99	96.53	1.150
21	TDLM	97.33	95.32	94.67	97.29	97.33	96.388	1.156
22	TMDL	97.32	94.71	96.03	96	97.35	96.282	0.983
23	MLTD	97.99	95.33	94.72	97.33	96	96.274	1.220
24	MMMM	94.67	96.69	95.35	96.64	97.33	96.136	0.975
25	MLDT	97.32	96.03	95.99	95.99	95.33	96.132	0.649
26	TLDM	94.67	94.68	97.31	97.33	95.31	95.86	1.214
27	TMLD	96	96.68	96.01	96.64	93.3	95.726	1.248
28	TTTT	93.33	95.97	94.68	94	93.99	94.394	0.896

4.2.1.4 การทดลองด้วยชุดข้อมูล Ecoli

อีกหนึ่งชุดข้อมูลที่สามารถใช้ทดสอบประสิทธิภาพการเรียงสับเปลี่ยนลำดับตัวเรียนรู้พื้นฐานที่ต้องการจำแนกประเภทมากกว่า 2 ประเภท โดยชุดข้อมูลนี้จะมีจำนวนประเภทอยู่ 8 ประเภท ซึ่งผลการทดลองที่ได้จะแสดงดังในตารางที่ 4.4

ผลการทดลองจากตารางดังกล่าวแสดงให้เห็นว่าการจัดเรียงแบบ DLTM ได้ค่าความถูกต้องเฉลี่ยสูงที่สุดที่ร้อยละ 87.14 และการจัดเรียงที่ได้อันดับรองลงมาในอันดับที่ 2 และ 3 ได้แก่ DTML ที่ร้อยละ 87.02 และ LDMT ที่ค่าความถูกต้องเฉลี่ยร้อยละ 86.9 ตามลำดับ ส่วนในอันดับท้ายสุดซึ่งได้ค่าความถูกต้องเฉลี่ยน้อยที่สุดคือการจัดเรียงแบบ TTTT ที่ค่าความถูกต้องเฉลี่ยร้อยละ 79.59 และมีการจัดเรียงแบบ TMDL ที่ได้ค่าความถูกต้องเฉลี่ยที่ร้อยละ 83.69 ซึ่งเป็นอันดับน้อยที่สุดรองจากอันดับสุดท้าย โดยความห่างของค่าความถูกต้องเฉลี่ยที่ใช้การจัดเรียงแบบ DLTM กับ TTTT นั้นมีค่าประมาณที่ต่างกันร้อยละ 7.55

นอกจากนี้หากพิจารณาค่าความถูกต้องเฉลี่ยที่ใช้การจัดเรียงที่ขึ้นต้นด้วยการใช้เทคนิคการวิเคราะห์เชิงเส้นแบบดิสคริมิแนนท์ เช่น DLMT DTLM DMLT DMTL ที่ได้ค่าความถูกต้องเฉลี่ยที่ร้อยละ 86.61 86.6 86.42 และ 86.31 ตามลำดับ รวมถึงการจัดเรียงที่ใช้เพียงเทคนิคการวิเคราะห์เชิงเส้นแบบดิสคริมิแนนท์เพียงเทคนิคเดียวเป็นตัวเรียนรู้พื้นฐานที่ได้ค่าความถูกต้องเฉลี่ย 86.48 ด้วยแล้ว จะสามารถอนุมานได้ว่าสำหรับชุดข้อมูลนี้การใช้วิธีการจำแนกประเภทด้วยเทคนิคการวิเคราะห์เชิงเส้นแบบดิสคริมิแนนท์นั้นมีความเหมาะสมที่สุด แต่อย่างไรก็ตามหากทำการพิจารณาโดยรวมแล้ว วิธีการจัดเรียงแบบผสมก็ยังให้ค่าความถูกต้องเฉลี่ยที่มากกว่าการใช้เทคนิคใดเพียงเทคนิคเดียวอยู่ดี ซึ่งก็อนุมานได้อีกเช่นกันว่าการใช้การจัดเรียงแบบผสมสามารถให้ผลการจำแนกที่ดีกว่าการจัดเรียงที่ใช้เพียงเทคนิคเดียวในการจัดเรียงตัวเรียนรู้พื้นฐาน

ตารางที่ 4.4 เปรียบเทียบประสิทธิภาพการจำแนกด้วยการสับเปลี่ยนลำดับการเรียงของตัวเรียนรู้พื้นฐานโดยใช้ชุดข้อมูล Ecoli

อันดับที่	การเรียงตัวจำแนก	รอบการคำนวณ					ค่าเฉลี่ยความถูกต้อง	ค่าเบี่ยงเบนมาตรฐาน
		1	2	3	4	5		
1	DLTM	87.2	86.61	87.5	87.19	87.2	87.14	0.290
2	DTML	87.51	87.8	84.82	88.39	86.61	87.026	1.244
3	LDMT	87.52	87.5	86.89	84.83	87.8	86.908	1.081
4	LTMD	88.39	85.12	86.01	88.09	86.3	86.782	1.256
5	MDTL	85.98	88.39	87.48	86.32	85.43	86.72	1.071
6	DLMT	86.9	86.89	85.45	86.61	87.2	86.61	0.609
7	DTLM	86.6	85.72	86.61	86.9	87.19	86.604	0.492
8	DDDD	86.61	85.12	86.29	87.82	86.6	86.488	0.862
9	LDTM	85.43	86.9	87.2	86.63	86	86.432	0.639
10	MLDT	87.2	85.71	86.9	86.01	86.31	86.426	0.552
11	DMLT	85.4	85.71	87.2	86.31	87.5	86.424	0.816
12	DMTL	87.21	84.83	83.63	88.39	87.5	86.312	1.785
13	LMDT	86.91	86.01	86.31	86.02	86.29	86.308	0.327
14	LMTD	86.01	88.99	85.42	85.68	84.84	86.188	1.452
15	MLTD	85.41	87.2	86.31	84.84	86.92	86.136	0.893
16	MDLT	86.33	85.13	86.01	86.01	86	85.896	0.403
17	MTDL	86.01	86.61	87.2	86.31	83.03	85.832	1.455
18	LTDM	85.12	85.4	86.31	86.61	85.41	85.77	0.581
19	LLLL	84.82	86.61	85.12	85.72	85.13	85.48	0.636
20	TMLD	86.01	85.4	86.31	84.82	84.21	85.35	0.767
21	TDML	86.01	85.42	84.52	84.53	85.12	85.12	0.564
22	MTLD	85.42	84.21	85.11	84.23	86.61	85.116	0.887
23	TDLM	86.9	84.22	85.71	85.12	83.35	85.06	1.220
24	MMMM	83.32	83.06	84.24	86.9	86.9	84.884	1.692
25	TLMD	83.94	85.71	86.9	82.42	85.13	84.82	1.533
26	TLDM	85.42	82.44	86.9	83.63	85.42	84.762	1.556
27	TMDL	82.73	83.66	84.84	83.93	83.32	83.696	0.698
28	TTTT	78.29	79.19	81.54	80.95	77.99	79.592	1.419

4.2.1.5 การทดลองด้วยชุดข้อมูล Glass Identification

ชุดข้อมูล Glass Identification เป็นชุดข้อมูลที่มีจำนวนประเภทข้อมูลมาก โดยมีประเภทข้อมูลด้วยกัน 7 ประเภทในชุดข้อมูลที่นำมาใช้ในการทดลอง ทั้งยังมีจำนวนของข้อมูลในแต่ละประเภทที่ไม่เท่ากันหรือมีข้อมูลในลักษณะที่เป็นอสมมาตร (Imbalanced dataset) รายละเอียดข้อมูลสามารถดูได้ที่ข้อ 3.1.6

ตารางที่ 4.5 แสดงผลการทดลองโดยใช้ชุดข้อมูลดังกล่าว ซึ่งการจัดเรียงแบบ MLDT จะให้ค่าความถูกต้องเฉลี่ยมากที่สุดที่ร้อยละ 67.94 โดยมีการจัดเรียงแบบ DMLT และ MTDL ที่ได้ค่าความถูกต้องเฉลี่ยมากเป็นอันดับรองลงมาที่ร้อยละ 67.56 และ 67.48 ตามลำดับ ส่วนการเรียงอันดับที่ใช้การจัดเรียงแบบผสมอันดับสุดท้ายจะเป็นการเรียงแบบ TMLD ที่ร้อยละ 64.6 และมีการจัดเรียงแบบ DLTM และ MLTD ที่ให้ค่าน้อยที่สุดรองจากการเรียงแบบ TMLD โดยมีค่าความถูกต้องเฉลี่ยที่ร้อยละ 64.84 และ 64.96 ตามลำดับ

อย่างไรก็ตาม ถึงแม้ว่าการจัดเรียงแบบใช้เทคนิคการเรียนรู้ต้นไม้ตัดสินใจเพียงเทคนิคเดียวเป็นตัวเรียนรู้พื้นฐาน (TTTT) ที่ให้ค่าความถูกต้องเฉลี่ยร้อยละ 65.69 ซึ่งมากกว่าการเรียนรู้ที่ใช้การจัดเรียงแบบผสมหลาย ๆ รูปแบบ อาทิ DLMT TDML LDTM LTMD MLTD DLTM และ TMLD แต่ก็ยังมีค่าความถูกต้องเฉลี่ยที่น้อยกว่าการใช้วิธีการจัดเรียงแบบผสมอีกหลายรูปแบบมาก อีกทั้งการใช้การเรียงด้วยวิธีที่ใช้เทคนิคอื่น ๆ เพียงแบบเดียวเป็นตัวเรียนรู้พื้นฐาน เช่น LLLL DDDD และ MMMM นั้น ก็ยังให้ค่าความถูกต้องเฉลี่ยที่ต่ำกว่าวิธีที่มีการจัดเรียงแบบผสมทั้งหมด โดยวิธีที่ได้ค่าความถูกต้องเฉลี่ยน้อยที่สุดคือวิธีการจัดเรียงที่ใช้เทคนิคโครงข่ายประสาทเทียมเทคนิคเดียวเป็นตัวเรียนรู้พื้นฐานโดยได้ค่าความถูกต้องเฉลี่ยที่ร้อยละ 58.42 ซึ่งเมื่อเทียบกับการจัดเรียงแบบ MLDT ที่ได้ค่าความถูกต้องสูงที่สุดแล้วจะมีค่าที่ต่างกันถึงร้อยละ 9.51

ตารางที่ 4.5 เปรียบเทียบประสิทธิภาพการจำแนกด้วยการสับเปลี่ยนลำดับการเรียงของตัวเรียนรู้พื้นฐานโดยใช้ชุดข้อมูล Glass Identification

อันดับที่	การเรียงตัวจำแนก	รอบการคำนวณ					ค่าเฉลี่ยความถูกต้อง	ค่าเบี่ยงเบนมาตรฐาน
		1	2	3	4	5		
1	MLDT	72.42	64.94	67.76	66.82	67.76	67.94	2.465
2	DMLT	64.95	67.27	72.89	67.27	65.45	67.566	2.823
3	MTDL	67.77	68.69	66.39	67.27	67.29	67.482	0.751
4	LTDM	64.98	67.3	68.23	70.57	65.43	67.302	2.022
5	DMTL	70.1	68.67	66.84	67.3	63.09	67.2	2.350
6	MDTL	71.03	65.9	66.86	64.51	66.8	67.02	2.178
7	DTLM	62.17	72.42	65.88	68.21	66.34	67.004	3.342
8	MTLD	65.41	71.04	68.73	64.98	64.5	66.932	2.536
9	TDLM	65.9	64.02	68.66	68.26	65.85	66.538	1.714
10	TLDM	64.95	66.35	63.06	68.69	69.63	66.536	2.402
11	TLMD	64.95	69.18	68.7	64.04	64.95	66.364	2.135
12	LDMT	66.39	66.33	66.36	68.24	64.47	66.358	1.192
13	LMTD	67.76	66.81	64.52	65.88	66.37	66.268	1.071
14	DTML	64.93	69.18	62.62	66.34	68.21	66.256	2.339
15	LMDT	66.79	66.85	67.76	65.42	63.05	65.974	1.642
16	TMDL	64.48	69.64	66.86	64.96	63.56	65.9	2.158
17	MDLT	63.99	70.57	64.96	63.55	65.88	65.79	2.522
18	TTTT	56.08	65.85	68.22	67.77	70.57	65.698	5.038
19	DLMT	61.71	64.95	67.75	66.36	67.31	65.616	2.176
20	TDML	66.79	63.05	67.31	64.49	64.48	65.224	1.589
21	LDTM	71.49	61.72	68.2	62.14	62.15	65.14	3.983
22	LTMD	64.47	65.9	62.14	64.02	64.98	64.98	1.249
23	MLTD	66.39	63.55	65.9	63.09	65.89	64.964	1.362
24	DLTM	64.93	63.07	65.39	64.95	65.88	64.844	0.953
25	TMLD	61.68	64.03	60.72	71.02	65.88	64.666	3.653
26	LLLL	61.65	64.93	61.24	62.12	64.51	62.89	1.526
27	DDDD	63.09	61.74	63.56	62.14	62.58	62.622	0.650
28	MMMM	61.72	60.75	57.02	57.45	55.2	58.428	2.433

4.2.1.6 การทดลองด้วยชุดข้อมูล Car Evaluation

แม้ชุดข้อมูล Car Evaluation จะมีประเภทข้อมูลอยู่ 2 ประเภทเท่านั้น แต่ก็เป็นชุดข้อมูลที่มีลักษณะแบบอสมมาตรที่มีจำนวนของข้อมูลในแต่ละประเภทที่ไม่เท่ากันค่อนข้างมาก ผู้วิจัยจึงนำชุดข้อมูลนี้มาใช้ในการทดลอง ซึ่งวิธีการจัดเรียงตัวเรียนรู้พื้นฐานที่ใช้เทคนิคการเรียนรู้ต้นไม้ตัดสินใจเป็นตัวแรกทั้งที่ใช้การจัดเรียงแบบผสมและแบบที่ใช้เทคนิคการเรียนรู้ต้นไม้ตัดสินใจเทคนิคเดียวเป็นตัวเรียนรู้พื้นฐาน สามารถให้ค่าความถูกต้องเฉลี่ยที่สูงกว่าแบบอื่น ๆ โดยแบบการจัดเรียงที่ได้ค่าความถูกต้องเฉลี่ยมากที่สุดคือ TLMD ที่ร้อยละ 94.78 โดยมีการจัดเรียงแบบ TTTT TDLM TDML TMLD TLDM TMDL เป็นอันดับที่รองลงมา โดยมีค่าความถูกต้องเฉลี่ยที่ร้อยละ 94.62 94.58 94.15 94.05 93.94 และ 93.89 ตามลำดับ ในขณะที่การใช้เทคนิคการเรียนรู้แบบอื่น ๆ เพียงเทคนิคเดียวเป็นตัวเรียนรู้พื้นฐานนั้น กลับให้ค่าความถูกต้องเฉลี่ยน้อยที่สุด เช่น การใช้เทคนิคการวิเคราะห์เชิงเส้นแบบดิสคริมิแนนท์เป็นตัวเรียนรู้พื้นฐานเทคนิคเดียวให้ค่าความถูกต้องเฉลี่ยที่ร้อยละ 81.48 ซึ่งน้อยกว่าแบบ TLMD อยู่ประมาณร้อยละ 13.3 และแม้ว่าการจัดเรียงที่ใช้เทคนิคการวิเคราะห์การถดถอยแบบโลจิสติกเทคนิคเดียวนั้นจะให้ค่าความถูกต้องเฉลี่ยที่แทบไม่ต่างจากแบบ DDDD มากนัก แต่แบบที่ใช้เทคนิคโครงข่ายประสาทเทียมเพอร์เซ็ปตรอนหลายชั้นเพียงเทคนิคเดียวนั้นให้ค่าความถูกต้องเฉลี่ยที่มากกว่าที่ร้อยละ 90.42 ซึ่งมีความต่างถึงร้อยละ 8.946 ดังนั้นอนุมานได้ว่าการใช้เทคนิคการวิเคราะห์เชิงเส้นแบบดิสคริมิแนนท์และการใช้เทคนิคการวิเคราะห์การถดถอยแบบโลจิสติกที่จัดเป็นเทคนิคการจำแนกเชิงสถิติมาทำการจำแนกประเภทข้อมูลชุดนี้นั้นอาจไม่เหมาะสม ซึ่งผลการทดลองกับชุดข้อมูล Car Evaluation สามารถแสดงดังในตารางที่ 4.6 โดยเรียงจากค่าเฉลี่ยความถูกต้องมากไปหาน้อย

ตารางที่ 4.6 เปรียบเทียบประสิทธิภาพการจำแนกด้วยการสับเปลี่ยนลำดับการเรียงของตัวเรียนรู้พื้นฐานโดยใช้ชุดข้อมูล Car Evaluation

อันดับที่	การเรียงตัวจำแนก	รอบการคำนวณ					ค่าเฉลี่ยความถูกต้อง	ค่าเบี่ยงเบนมาตรฐาน
		1	2	3	4	5		
1	TLMD	94.68	95.97	94.68	94.09	94.5	94.784	0.631
2	TTTT	94.79	94.04	94.56	94.5	95.25	94.628	0.395
3	TDLM	93.93	94.73	95.78	93.34	95.14	94.584	0.864
4	TDML	94.56	94.44	93.98	93.52	94.27	94.154	0.372
5	TMLD	94.79	93.52	94.16	93.69	94.1	94.052	0.441
6	TLDM	93.29	93.4	93.75	94.91	94.39	93.948	0.615
7	TMDL	93.98	94.16	94.56	94.22	92.53	93.89	0.705
8	MTDL	94.04	92.88	93.63	93.4	93.11	93.412	0.404
9	DMTL	92.36	93.75	92.82	94.27	93.06	93.252	0.679
10	LTMD	91.67	93.29	93.23	93.05	94.21	93.09	0.816
11	LTDM	92.42	93.06	92.94	93.11	93.17	92.94	0.271
12	DTLM	92.59	92.71	93.23	93	92.65	92.836	0.242
13	LDMT	93.46	92.82	92.42	92.19	93.06	92.79	0.452
14	MDLT	93.17	91.2	92.94	93.4	92.77	92.696	0.778
15	DMLT	93.11	92.53	92.94	92.19	92.71	92.696	0.321
16	MDTL	91.44	93.58	93.63	93.05	91.67	92.674	0.939
17	MTLD	94.15	93	92.48	91.49	91.38	92.5	1.024
18	LDTM	92.02	92.42	92.36	91.32	93.52	92.328	0.713
19	MLTD	91.78	90.97	93.06	92.76	92.59	92.232	0.760
20	DLMT	91.72	92.88	91.26	92.53	92.48	92.174	0.593
21	LMTD	90.4	92.77	92.47	93.11	91.67	92.084	0.967
22	DTML	92.65	91.44	91.96	92.42	91.72	92.038	0.444
23	DLTM	92.25	90.16	92.65	90.97	93.81	91.968	1.281
24	MLDT	92.77	92.02	90.74	91.21	92.19	91.786	0.723
25	LMDT	92.48	89.18	91.03	91.38	92.76	91.366	1.271
26	MMMM	91.26	90.74	91.55	89.47	89.12	90.428	0.967
27	LLLL	83.28	83.22	83.04	82.93	82.23	82.94	0.376
28	DDDD	81.65	81.37	81.31	81.66	81.42	81.482	0.146

4.2.1.7 การเปรียบเทียบประสิทธิภาพการจำแนกประเภทด้วยการสับเปลี่ยนตัวเรียนรู้พื้นฐานโดยภาพรวม

ในตารางที่ 4.7 แสดงการจัดอันดับค่าเฉลี่ยของผลการเรียงลำดับค่าความถูกต้องเฉลี่ยในแต่ละชุดข้อมูล โดยค่ายิ่งมากยิ่งหมายถึงค่าอันดับที่มากในการจัดอันดับค่าความถูกต้องเฉลี่ยที่มีการเรียงจากน้อยไปมาก (ค่ายิ่งมากยิ่งมีความถูกต้องเฉลี่ยที่มาก) ซึ่งการจัดเรียงตัวเรียนรู้พื้นฐานแบบ MTDL ที่ได้อันดับที่ 26 16 22 12 26 และ 21 ในชุดข้อมูล Banknote Authentication, Connectionist Bench, Iris, Ecoli, Glass Identification และ Car Evaluation ตามลำดับ มีค่าอันดับเฉลี่ยที่ 20.5 ซึ่งเป็นค่าเฉลี่ยที่มากที่สุด หมายความว่าจากผลการทดลองจากชุดข้อมูล 6 ชุด การใช้การจัดเรียงแบบ MTDL มักจะได้ค่าความถูกต้องเฉลี่ยในอันดับที่ดี และในทางกลับกันการจัดเรียงแบบผสมด้วยวิธี TMLD ก็ได้ค่าอันดับเฉลี่ยน้อยที่สุดสำหรับการวิธีการจัดเรียงแบบผสมที่ค่าเฉลี่ย 12.33 อย่างไรก็ตาม เมื่อดูจากภาพรวมจากตารางดังกล่าวแล้วดูเหมือนว่าการใช้เทคนิคการจัดเรียงที่ใช้เทคนิคการเรียนรู้เทคนิคเดียวมาเป็นตัวเรียนรู้พื้นฐานนั้น ทั้งหมดจะอยู่ในอันดับท้าย ๆ นั่นอาจหมายความว่าการใช้วิธีการจัดเรียงแบบผสมนั้นมีโอกาสที่จะให้ผลในการจำแนกประเภทข้อมูลได้ดีกว่าแบบที่ใช้เทคนิคเดียว

สำหรับในตารางที่ 4.8 จะแสดงการจัดอันดับจากค่าเฉลี่ยของค่าความถูกต้องเฉลี่ยจากทุกชุดข้อมูล โดยวิธีการจัดเรียงที่ได้ค่าความถูกต้องเฉลี่ยจากทุกชุดข้อมูลที่สูงที่สุดคือ MDTL ที่ค่าเฉลี่ยร้อยละ 86.66 และมีอันดับที่ 2 และ 3 ที่เป็นอันดับรองลงมา ได้แก่ MTDL ที่ร้อยละ 86.61 และ LTMD ที่ร้อยละ 86.56 ตามลำดับ ส่วนวิธีการเรียงที่ใช้เทคนิคแบบผสมที่ได้ค่าเฉลี่ยน้อยที่สุดคือแบบ TMLD ที่ได้ค่าเฉลี่ยที่ร้อยละ 85.72 ซึ่งมีค่าที่ห่างจากอันดับแรกอยู่ไม่มากเท่าใดนัก แต่สำหรับวิธีการที่ใช้เทคนิคการเรียนรู้เพียงเทคนิคเดียวเป็นตัวเรียนรู้พื้นฐาน เช่น MMMM TTTT LLLL และ DDDD นั้นจะให้ค่าเฉลี่ยที่ต่ำที่สุด โดยมีค่าเฉลี่ยร้อยละ 84.47 83.43 83.37 และ 83.07 ตามลำดับ ซึ่งเมื่อพิจารณาเบื้องต้นค่าเฉลี่ยที่ได้จากการใช้เทคนิคเดียวเป็นตัวเรียนรู้พื้นฐานเหล่านี้ คล้ายกับจะมีระยะห่างบ้างแม้จะมีเพียงเล็กน้อย โดยเฉพาะกับแบบที่ใช้เทคนิคการวิเคราะห์เชิงเส้นแบบดิสคริมิแนนท์เพียงเทคนิคเดียวนั้นที่ได้ค่าเฉลี่ยในการจำแนกประเภทน้อยที่สุด

ตารางที่ 4.7 เปรียบเทียบลำดับ (Ranking) ของการจัดอันดับค่าความถูกต้องเฉลี่ยในแต่ละชุดข้อมูล โดยเรียงจากค่ามากไปหาค่าน้อย

การจัดเรียงตัวเรียนรู้พื้นฐาน	Banknote Authenti-cation	Connectio-nist Bench	Iris	Ecoli	Glass	Car	ค่าเฉลี่ย (ร้อยละ)
MTDL	26	16	22	12	26	21	20.5
DTML	28	23	19	27	15	7	19.83
LTMD	16	29	21	25	7	19	19.5
DMLT	24	15	16	18	27	14	19
DTLM	21	6	24	22	22	17	18.67
DMTL	18	5	26	17	24	20	18.33
LMDT	27	22	27	16	14	4	18.33
LTDM	23	19	12	11	25	18	18
MDTL	4	28	10	24	23	13	17
DLTM	15	20	25	28	5	6	16.5
DLMT	10	24	15	23	10	9	15.17
TDLM	13	18	8	6	20	26	15.17
TLMD	17	14	9	4	18	28	15
LDMT	5	7	14	26	17	16	14.16667
TLDM	7	30	3	3	19	23	14.16667
MLDT	19	9	4	19	28	5	14
LMTD	13	13	17	15	16	8	13.67
MTLD	5	24	11	7	21	12	13.33
TMDL	20	15	7	2	13	22	13.17
MLTD	22	21	6	14	6	10	13.17
TDML	12	12	13	8	9	25	13.17
MDLT	8	11	20	13	12	14	13
LDTM	10	8	18	20	8	11	12.5
TMLD	25	10	2	9	4	24	12.33
DDDD	2	3	28	21	2	1	9.5
MMMM	9	27	5	5	1	3	8.33
LLLL	3	4	23	10	3	2	7.5
TTTT	1	2	1	1	11	27	7.167

ตารางที่ 4.8 เปรียบเทียบลำดับ (Ranking) ของค่าความถูกต้องเฉลี่ยของแต่ละชุดข้อมูล โดยเรียงจากค่ามากไปหาค่าน้อย

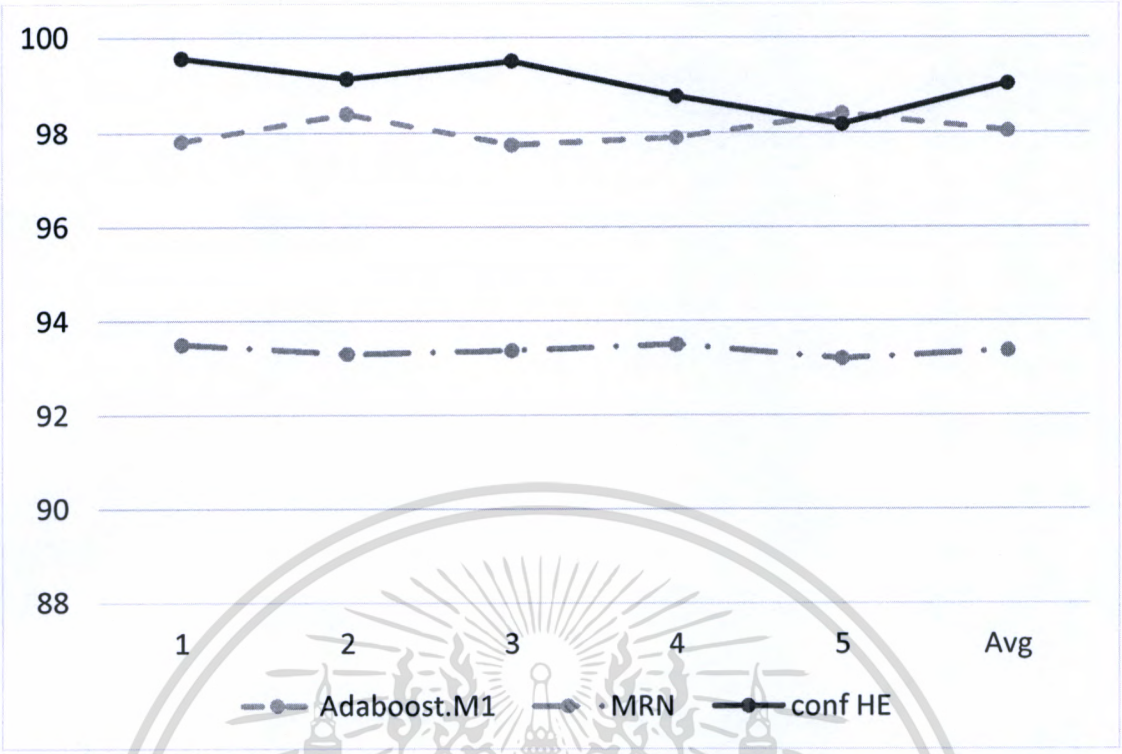
การจัดเรียงตัวเรียนรู้พื้นฐาน	Banknote Authenti-cation	Connectio-nist Bench	Iris	Ecoli	Glass	Car	ค่าเฉลี่ย (ร้อยละ)
MDTL	99.228	77.776	96.54	86.72	67.02	92.674	86.66
MTDL	99.548	76.164	97.212	85.832	67.482	93.412	86.61
LTMD	99.416	77.878	97.194	86.782	64.98	93.09	86.56
DMLT	99.49	75.962	96.942	86.424	67.566	92.696	86.51
TLDM	99.256	78.456	95.86	84.762	66.536	93.948	86.47
DTML	99.606	76.636	97.078	87.026	66.256	92.038	86.44
LTDM	99.476	76.25	96.666	85.77	67.302	92.94	86.4
TDLM	99.402	76.246	96.388	85.06	66.538	94.584	86.37
DLMT	99.374	77.216	96.934	86.61	65.616	92.174	86.32
TLMD	99.418	75.96	96.53	84.82	66.364	94.784	86.31
LMDT	99.548	76.622	97.726	86.308	65.974	91.366	86.26
MTLD	99.242	77.032	96.652	85.116	66.932	92.5	86.25
DLTM	99.402	76.252	97.616	87.14	64.844	91.968	86.2
DMTL	99.43	73.166	97.62	86.312	67.2	93.252	86.16
DTLM	99.46	73.662	97.34	86.604	67.004	92.836	86.15
LMTD	99.402	75.664	96.952	86.188	66.268	92.084	86.09
MLDT	99.446	74.818	96.132	86.426	67.94	91.786	86.09
LDMT	99.242	74.144	96.926	86.908	66.358	92.79	86.06
TDML	99.386	75.654	96.812	85.12	65.224	94.154	86.06
MDLT	99.3	75.198	97.19	85.896	65.79	92.696	86.01
MLTD	99.474	76.558	96.274	86.136	64.964	92.232	85.94
TMDL	99.46	76.16	96.282	83.696	65.9	93.89	85.9
LDTM	99.374	74.444	97.074	86.432	65.14	92.328	85.8
TMLD	99.518	75.018	95.726	85.35	64.666	94.052	85.72
MMMM	99.358	77.612	96.136	84.884	58.428	90.428	84.47
TTTT	97.52	68.748	94.394	79.592	65.698	94.628	83.43
LLLL	98.73	72.97	97.224	85.48	62.89	82.94	83.37
DDDD	97.524	72.304	98.004	86.488	62.622	81.482	83.07

4.2.2 การทดสอบประสิทธิภาพเพื่อเปรียบเทียบกับวิธีการเรียนรู้แบบรวมกลุ่มวิธีอื่น

ในส่วนของการทดลองเปรียบเทียบโดยใช้การเรียนรู้แบบรวมกลุ่มเอ็มอาร์เอ็นนั้น ผู้วิจัยได้รับความอนุเคราะห์ชุดคำสั่งโปรแกรม MATLAB จากนายอนุสรณ์ เจริญนาน ซึ่งเป็นผู้วิจัยวิธีการดังกล่าวมาใช้ร่วมในการเปรียบเทียบในงานวิจัยนี้ด้วย โดยในส่วนการตั้งค่าพารามิเตอร์โครงข่ายประสาทเทียมนั้นได้ใช้การปรับตั้งค่าที่เหมือนกัน แต่ในส่วนการตั้งค่าพารามิเตอร์ในการเรียกใช้งานเทคนิคเรเดียลเบสิสฟังก์ชันและเทคนิคการเรียนรู้ด้วยตัวจำแนกแบบเบย์อย่างง่ายนั้น ผู้วิจัยได้ใช้ค่าพารามิเตอร์ตามในงานวิจัยการเรียนรู้แบบรวมกลุ่มเอ็มอาร์เอ็น ซึ่งใช้ฟังก์ชัน newgrnn สำหรับการใช้เทคนิคเรเดียลเบสิสฟังก์ชันและกำหนดให้พารามิเตอร์ตัวที่ 3 ซึ่งเป็นค่าพารามิเตอร์ความกว้างของเบสิส (Spread of radial basis functions) มีค่าเป็น 0.5 และใช้ฟังก์ชัน fitNaiveBayes เป็นฟังก์ชันการหาผลลัพธ์ด้วยเทคนิคการเรียนรู้ด้วยตัวจำแนกแบบเบย์อย่างง่ายโดยใช้ค่าปริยายเป็นค่าพารามิเตอร์

การแสดงผลการทดลองในรูปแบบกราฟเส้นเปรียบเทียบและตารางเปรียบเทียบผู้เขียนขอใช้ชื่อ conf HE แทนวิธีการรวมกลุ่มผสมระหว่างการเรียนรู้ของเครื่องและการเรียนรู้เชิงสถิติโดยใช้การบูสต์ด้วยค่าความเชื่อมั่นเพื่อให้กราฟและตารางสามารถดูได้ง่ายและเป็นระเบียบเรียบร้อย

โดยผลการทดลองที่ได้จากการวัดประสิทธิภาพระหว่างการจำแนกประเภทแบบรวมกลุ่มของวิธีการที่นำเสนอกับวิธีการเรียนรู้แบบรวมกลุ่มเอดาบูสต์เอ็มวันและการเรียนรู้แบบรวมกลุ่มเอ็มอาร์เอ็นสามารถแสดงได้ดังนี้



รูปที่ 4.1 กราฟเส้นวัดประสิทธิภาพการจำแนกประเภทชุดข้อมูล Banknote Authentication

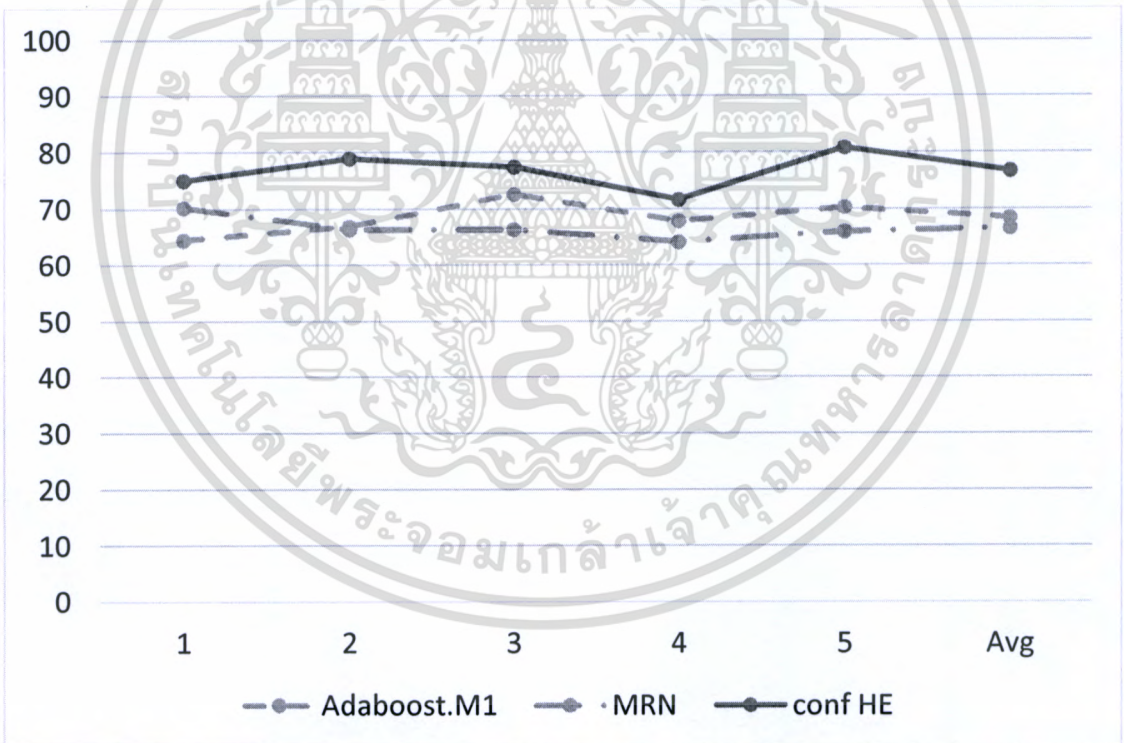
ตารางที่ 4.9 เปรียบเทียบประสิทธิภาพการจำแนกประเภทชุดข้อมูล Banknote Authentication

รอบการทดลอง	เอดาบูสต์เอ็มวัน	เอ็มอาร์เอ็น	conf HE
	ความถูกต้อง (ร้อยละ)	ความถูกต้อง (ร้อยละ)	ความถูกต้อง (ร้อยละ)
1	97.81	93.5	99.56
2	98.40	93.3	99.13
3	97.74	93.37	99.5
4	97.89	93.5	98.76
5	98.40	93.2	98.18
ค่าเฉลี่ย	98.04	93.37	99.02
ค่าเบี่ยงเบนมาตรฐาน	0.291	0.116	0.511

รูปที่ 4.1 และตารางที่ 4.9 เป็นผลการทดลองการจำแนกประเภทแบบไบนารีด้วยชุดข้อมูล Banknote Authentication ซึ่งเป็นชุดข้อมูลที่ได้มาจากการสกัดข้อมูลภาพถ่ายธนบัตรจริงและธนบัตรปลอม ด้วยวิธีการจำแนกประเภทแบบรวมกลุ่มวิธีต่าง ๆ ซึ่งจะเห็นได้ว่า ในขณะที่การเรียนรู้แบบรวมกลุ่มเอ็มอาร์เอ็นให้ผลลัพธ์เฉลี่ยที่ร้อยละ 93.37 แต่การเรียนรู้แบบรวมกลุ่มเอดาบูสต์เอ็มวันและการวิธีการเรียนรู้ที่นำเสนอนั้นให้ผลลัพธ์ที่ใกล้เคียงกันและมีความถูกต้องที่สูงกว่าการเรียนรู้แบบรวมกลุ่มเอ็มอาร์เอ็น โดยวิธีเอดาบูสต์เอ็มวันได้ค่าความถูกต้องเฉลี่ยอยู่ที่ร้อยละ 98.04 ส่วนวิธีการที่

นำเสนอจะมีค่าความถูกต้องเฉลี่ยที่ร้อยละ 99.02 ซึ่งเป็นค่าเฉลี่ยที่สูงที่สุด และยังได้ค่าความถูกต้องสูงสุดที่ร้อยละ 99.56 อีกด้วย

ชุดข้อมูลที่ใช้สำหรับการทดลองจำแนกประเภทแบบไบนารีชุดถัดมา คือ ชุดข้อมูล Connectionist Bench ซึ่งเป็นชุดข้อมูลที่ได้มาจากการเก็บข้อมูลของเครื่องโซนาร์ โดยเป็นข้อมูลสัญญาณการสะท้อนกลับของก้อนหินและแร่เหล็กในมุมต่าง ๆ ซึ่งวิธีการที่นำเสนอก็ยังสามารถให้ความถูกต้องในการจำแนกประเภทได้สูงที่สุดที่ร้อยละ 78.85 และได้ความถูกต้องเฉลี่ยที่ร้อยละ 76.73 ซึ่งเป็นความถูกต้องเฉลี่ยที่สูงที่สุดในสามวิธีเช่นกัน ในขณะที่การใช้วิธีการจำแนกประเภทเอดาบูสต์เอ็มวันและเอ็มอาร์เอ็นนั้นได้ค่าความถูกต้องสูงสุดที่ร้อยละ 72.57 และร้อยละ 70.2 ตามลำดับ โดยที่วิธีการจำแนกประเภทเอดาบูสต์เอ็มวันมีค่าความถูกต้องเฉลี่ยที่ร้อยละ 68.37 ส่วนวิธีการจำแนกประเภทเอ็มอาร์เอ็นนั้นได้ค่าความถูกต้องเฉลี่ยที่ร้อยละ 66.55 ซึ่งผลการทดลองชุดข้อมูล Connectionist Bench นั้นจะแสดงในตารางที่ 4.10 และสามารถแสดงการเปรียบเทียบได้ดังรูปที่ 4.2

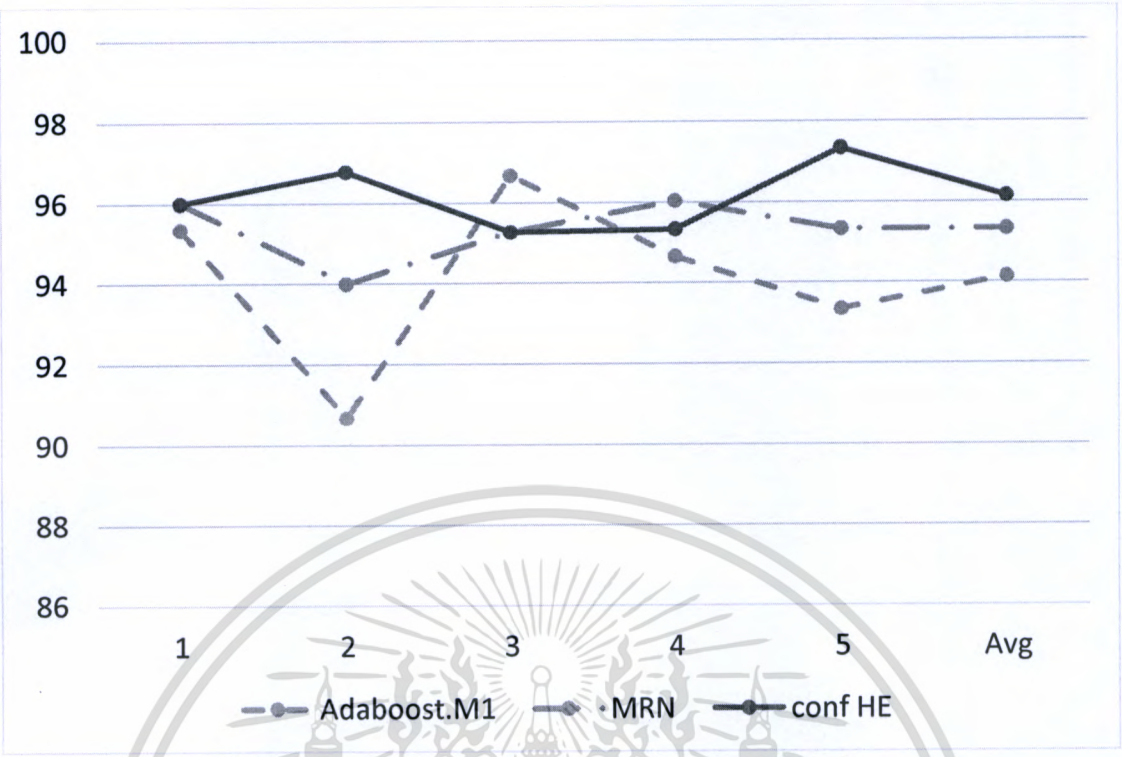


รูปที่ 4.2 กราฟเส้นวัดประสิทธิภาพการจำแนกประเภทชุดข้อมูล Connectionist Bench

ตารางที่ 4.10 เปรียบเทียบประสิทธิภาพการจำแนกประเภทชุดข้อมูล Connectionist Bench

รอบการทดลอง	เอดาบูสต์เอ็มวัน	เอ็มอาร์เอ็น	conf HE
	ความถูกต้อง (ร้อยละ)	ความถูกต้อง (ร้อยละ)	ความถูกต้อง (ร้อยละ)
1	64.44	70.2	74.99
2	66.83	66.3	78.85
3	72.57	66.3	77.4
4	67.81	64.1	71.61
5	70.21	65.86	80.80
ค่าเฉลี่ย	68.37	66.55	76.73
ค่าเบี่ยงเบนมาตรฐาน	2.798	1.996	3.186

ชุดข้อมูลถัดมา คือ ชุดข้อมูล Iris ซึ่งเป็นชุดข้อมูลที่มักถูกนำมาใช้อ้างอิงในงานการจำแนกประเภทข้อมูลในหลาย ๆ งาน ทั้งเป็นชุดข้อมูลที่มีจำนวนประเภทข้อมูลมากกว่า 2 ประเภท โดยเป็นชุดข้อมูลที่ได้มาจากการวัดขนาดของส่วนประกอบต่าง ๆ ของดอกไอริสสายพันธุ์ต่าง ๆ ซึ่งจากผลการทดลองที่ได้แสดงให้เห็นว่าในการจำแนกประเภทข้อมูลที่มีประเภทข้อมูลมากกว่า 2 ประเภทนั้น วิธีการที่นำเสนอก็สามารถให้ผลลัพธ์ที่ดีที่สุดได้แม้จะมีความถูกต้องที่ไม่ทิ้งห่างกันมากนักก็ตาม โดยวิธีการที่นำเสนอได้ค่าความถูกต้องเฉลี่ยที่ร้อยละ 96.14 และได้ค่าความถูกต้องสูงสุดที่ร้อยละ 97.33 ในขณะที่การใช้วิธีการจำแนกประเภทเอ็มอาร์เอ็นได้ค่าความถูกต้องเฉลี่ยรองลงมาที่ร้อยละ 95.33 และมีค่าความถูกต้องสูงสุดที่ร้อยละ 96.03 ส่วนวิธีการจำแนกประเภทเอดาบูสต์เอ็มวันได้ค่าความถูกต้องเฉลี่ยต่ำที่สุดที่ร้อยละ 94.14 แต่มีค่าความถูกต้องสูงสุดมากกว่าวิธีเอ็มอาร์เอ็นที่ร้อยละ 96.67 ซึ่งหากสังเกตกราฟเปรียบเทียบในรูปที่ 4.3 จะสามารถเห็นได้ว่าการใช้วิธีการจำแนกประเภทเอดาบูสต์เอ็มวันจะให้ผลความถูกต้องที่มีลักษณะของกราฟที่เหวี่ยงขึ้นเหวี่ยงลงมากกว่าอีกสองวิธี และรวมไปถึงค่าเบี่ยงเบนมาตรฐานที่ประมาณ 2.038 ซึ่งทิ้งห่างจากวิธีรวมกลุ่มแบบอื่น จึงเป็นที่น่าสังเกตว่าหากต้องการใช้วิธีการจำแนกประเภทด้วยวิธีการรวมกลุ่มเอดาบูสต์เอ็มวันนั้น ควรจะทำการกำหนดจำนวนรอบให้สูงกว่านี้เพื่อให้การปรับค่าการแจกแจงนั้นให้มีความเสถียรที่มากขึ้นเสียก่อน ซึ่งก็จะส่งผลให้การคำนวณในรอบที่สูง ๆ สามารถให้ค่าน้ำหนักที่ดีขึ้นและสามารถให้ผลลัพธ์ในการจำแนกที่แม่นยำมากขึ้นไปด้วย แต่อย่างไรก็ตามการกำหนดจำนวนรอบที่มากก็จะทำให้ต้องใช้เวลาในการคำนวณมากขึ้นตามจำนวนรอบไปด้วย สำหรับผลการทดลองของชุดข้อมูลนี้จะแสดงดังในตารางที่ 4.11



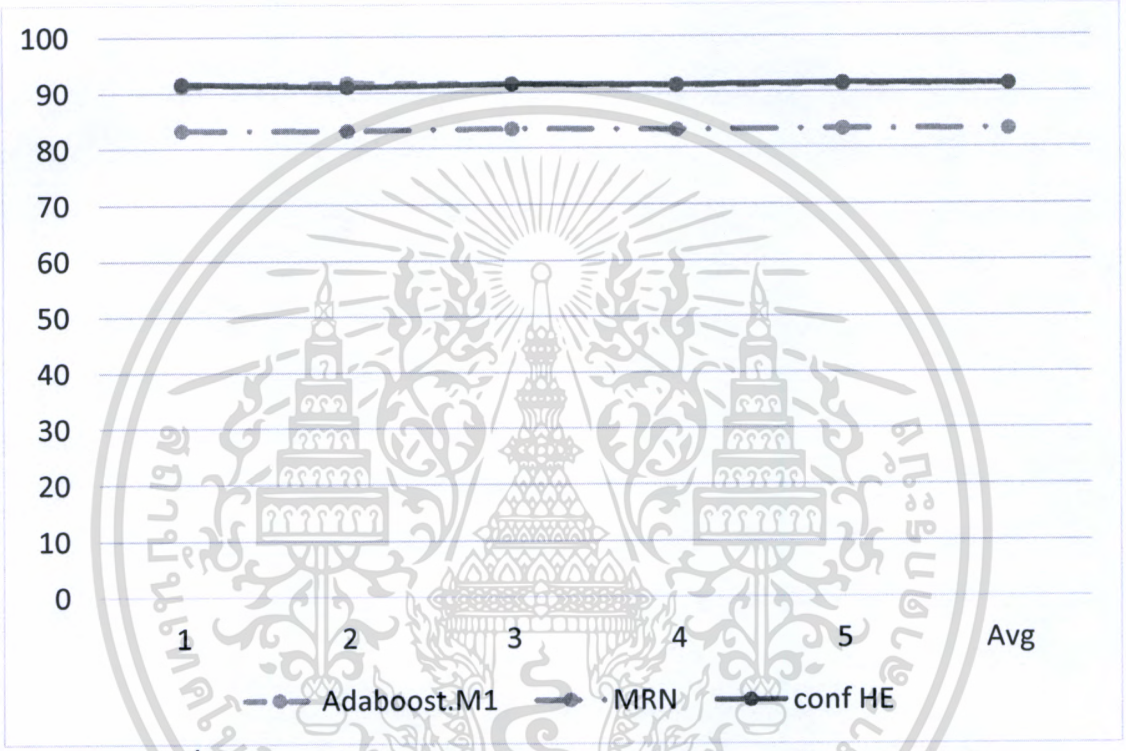
รูปที่ 4.3 กราฟเส้นวัดประสิทธิภาพการจำแนกประเภทชุดข้อมูล Iris

ตารางที่ 4.11 เปรียบเทียบประสิทธิภาพการจำแนกประเภทชุดข้อมูล Iris

รอบการทดลอง	เอดาบูสต์เอ็มวัน	เอ็มอาร์เอ็น	conf HE
	ความถูกต้อง (ร้อยละ)	ความถูกต้อง (ร้อยละ)	ความถูกต้อง (ร้อยละ)
1	95.34	96	95.99
2	90.67	94	96.77
3	96.67	95.27	95.27
4	94.65	96.03	95.33
5	93.35	95.33	97.33
ค่าเฉลี่ย	94.14	95.33	96.14
ค่าเบี่ยงเบนมาตรฐาน	2.038	0.736	0.806

ชุดข้อมูลถัดมาเป็นชุดข้อมูลที่ได้มาจากการเก็บข้อมูลด้วยอุปกรณ์เซ็นเซอร์ที่ติดตามร่างกายของอาสาสมัครจำนวน 9 คน โดยมีข้อมูลการทำกิจกรรมต่าง ๆ ของอาสาสมัครจำนวน 18 กิจกรรม ซึ่งในตารางที่ 4.12 จะแสดงผลการทดลองการวัดประสิทธิภาพการจำแนกประเภทข้อมูลของชุดข้อมูลนี้ โดยเอ็มอาร์เอ็นมีค่าความถูกต้องเฉลี่ยน้อยที่สุดที่ร้อยละ 83.34 และได้ค่าความถูกต้องสูงสุดที่ร้อยละ 83.5 ในขณะที่ค่าความถูกต้องเฉลี่ยของการเรียนรู้แบบรวมกลุ่มเอดาบูสต์และวิธีการรวมกลุ่มผสมด้วยค่าความเชื่อมั่นได้ค่าที่ใกล้เคียงกันมากที่ร้อยละ 91.4 และ 91.46 ตามลำดับ ทั้งนี้

ในส่วนของการเรียนรู้แบบรวมกลุ่มเอดาบูสต์เอ็มวันนั้นได้ค่าความถูกต้องสูงที่สุดที่ร้อยละ 91.77 และวิธีที่นำเสนอนั้นได้ต่ำกว่าเพียงเล็กน้อยที่ร้อยละ 91.67 ทั้งนี้ผู้วิจัยตั้งข้อสังเกตว่าสาเหตุที่ทำให้ได้ผลความถูกต้องในการจำแนกประเภทอย่างนี้เนื่องมาจากชุดข้อมูลชุดนี้อาจมีความเหมาะสมกับการใช้วิธีการจำแนกประเภทด้วยวิธีการใช้ต้นไม้ตัดสินใจ ซึ่งเมื่อประกอบการการบูสต์ด้วยต้นไม้ตัดสินใจในจำนวนรอบที่มากกว่าวิธีอื่นจึงอาจส่งผลให้ได้ผลที่ดีกว่าในบางครั้ง โดยผลการเปรียบเทียบในรูปของเส้นเปรียบเทียบสามารถแสดงได้ดังรูปที่ 4.4



รูปที่ 4.4 กราฟเส้นวัดประสิทธิภาพการจำแนกประเภทชุดข้อมูล PAMAP2

ตารางที่ 4.12 เปรียบเทียบประสิทธิภาพการจำแนกประเภทชุดข้อมูล PAMAP2

รอบการทดลอง	เอดาบูสต์เอ็มวัน	เอ็มอาร์เอ็น	conf HE
	ความถูกต้อง (ร้อยละ)	ความถูกต้อง (ร้อยละ)	ความถูกต้อง (ร้อยละ)
1	91.22	83.3	91.67
2	91.77	83.2	91.08
3	91.49	83.5	91.58
4	91.16	83.3	91.37
5	91.36	83.4	91.63
ค่าเฉลี่ย	91.40	83.34	91.46
ค่าเบี่ยงเบนมาตรฐาน	0.218	0.102	0.219

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

งานวิจัยนี้นำเสนอวิธีการจำแนกประเภทข้อมูลโดยการนำวิธีการเรียนรู้ของเครื่องมาผสมรวมกับการจำแนกประเภทด้วยวิธีการคำนวณทางสถิติ โดยใช้เอดาบูสต์เอ็มวัน นอกจากนี้ในขั้นตอนการพิจารณาเลือกสมมติฐานสุดท้ายยังได้นำวิธีการคำนวณค่าความเชื่อมั่นมาประยุกต์ใช้ร่วมกับการใช้ค่าน้ำหนักด้วย ซึ่งผลลัพธ์ที่ได้จากการทดลองแสดงให้เห็นว่า แม้จะมีการสลับสับเปลี่ยนการเรียงเทคนิคในรูปแบบต่าง ๆ แต่ผลลัพธ์ที่ได้ก็แทบจะไม่มีความแตกต่างกันมากนัก นอกจากนี้การทดลองยังแสดงให้เห็นว่า การใช้เทคนิคการจำแนกประเภทด้วยวิธีการเรียนรู้ของเครื่องนั้นอาจเหมาะกับการจำแนกประเภทในบางชุดข้อมูล และในขณะเดียวกันการใช้วิธีการจำแนกประเภทด้วยการคำนวณเชิงสถิติก็สามารถให้ผลการจำแนกที่ดีกว่ากับบางชุดข้อมูลเช่นกัน ทั้งยังแสดงให้เห็นด้วยว่าการใช้วิธีการจำแนกประเภทด้วยวิธีการผสมระหว่างเทคนิคการเรียนรู้ของเครื่องและการเรียนรู้การจำแนกประเภทเชิงสถิตินั้นสามารถให้ผลที่ดีกว่าการใช้เทคนิคใดเพียงเทคนิคเดียวได้

และเมื่อทำการวัดประสิทธิภาพเปรียบเทียบกับวิธีการเรียนรู้แบบรวมกลุ่มกับวิธีอื่นก็แสดงให้เห็นเช่นกันว่า วิธีการจำแนกประเภทแบบรวมกลุ่มด้วยการผสมระหว่างการเรียนรู้ของเครื่องและการเรียนรู้เชิงสถิติโดยใช้การบูสต์ด้วยค่าความเชื่อมั่นสามารถให้ประสิทธิภาพความถูกต้องที่ดีกว่าในการจำแนกประเภทข้อมูลที่ได้จากข้อมูลประเภทต่าง ๆ เช่น ข้อมูลลักษณะของดอกไม้ที่เป็นข้อมูลเชิงปริมาณ ข้อมูลที่ได้จากภาพ ข้อมูลที่ได้จากสัญญาณสะท้อนของคลื่นโซนาร์ หรือแม้กระทั่งข้อมูลที่ได้มาจากเซ็นเซอร์ตรวจจับการเคลื่อนไหว IMU เป็นต้น ในขณะเดียวกันข้อมูลที่ถูกนำมาใช้ในการทดลองที่มีทั้งการจำแนกประเภทแบบไบนารีและการจำแนกประเภทแบบหลายประเภทก็สามารถทำการจำแนกได้ดีเช่นกันทั้ง 2 แบบ โดยเมื่อทำการเปรียบเทียบความถูกต้องระหว่างวิธีการที่นำเสนอกับวิธีการเรียนรู้แบบรวมกลุ่มเอดาบูสต์เอ็มวันที่ใช้ต้นไม้ตัดสินใจเป็นตัวเรียนรู้พื้นฐานและกับวิธีการเรียนรู้แบบรวมกลุ่มเอ็มอาร์เอ็นแล้วนั้น วิธีการที่นำเสนอจะได้ความถูกต้องที่สูงกว่า

แต่อย่างไรก็ตามจากผลการทดลองในบางชุดข้อมูลดังเช่นชุดข้อมูล Iris หากทำการเพิ่มจำนวนรอบการเรียนรู้พื้นฐานของเอดาบูสต์เอ็มวันก็อาจได้ผลลัพธ์การจำแนกข้อมูลที่มีความถูกต้องที่มากกว่าวิธีการที่นำเสนอได้ เนื่องจากการบูสต์ในแต่ละรอบนั้นจะมีการปรับปรุงค่าน้ำหนักให้กับคุณลักษณะบางตัวที่มีผลต่อความถูกต้องในการจำแนกประเภทด้วย แต่การเพิ่มจำนวนรอบการเรียนรู้นั้นก็ต้องแลกกับเวลาและทรัพยากรที่มากขึ้นในการคำนวณด้วยเช่นกัน นอกจากนี้ผลการทดลองในบางชุดข้อมูลก็แสดงให้เห็นว่าชุดข้อมูลนั้น ๆ มีความเหมาะสมกับการใช้เทคนิคการเรียนรู้ด้วยบางเทคนิคเท่านั้น อย่างเช่นการทดลองกับชุดข้อมูล PAMAP2 ที่ผลการทดลองแสดงให้เห็นถึงค่า

ความถูกต้องที่สูงกว่าหากใช้เทคนิคการเรียนรู้ด้วยต้นไม้ตัดสินใจ แต่ถึงกระนั้นการที่วิธีการที่นำเสนอสามารถให้ความถูกต้องในการจำแนกประเภทเฉลี่ยที่มากกว่าได้นั้น ส่วนหนึ่งเป็นผลมาจากการใช้เทคนิคการเรียนรู้ด้วยโครงข่ายประสาทเทียมเพอร์เซ็ปตรอนแบบหลายชั้นมาช่วยปรับค่าน้ำหนักของคุณลักษณะของข้อมูลก่อน ซึ่งทำให้เมื่อถึงขั้นตอนของการจำแนกประเภทด้วยต้นไม้ตัดสินใจ ค่าน้ำหนักที่ผ่านการปรับค่าดังกล่าวจึงไปเน้นที่คุณลักษณะข้อมูลที่มีผลต่อความถูกต้องในการจำแนกประเภท ส่งผลให้การจำแนกประเภทในรอบดังกล่าวได้ผลที่ดีขึ้นไปด้วย

5.2 ข้อเสนอแนะ

ในงานวิจัยนี้ได้นำเอาเทคนิคการเรียนรู้ของเครื่องมาทำงานร่วมกับการใช้เทคนิคการคำนวณทางสถิติเพื่อใช้ในการจำแนกประเภทข้อมูล ซึ่งหากมองตามกระบวนการของการเรียนรู้แบบรวมกลุ่มด้วยวิธีการบูสต์ปกติทั่วไปที่จะมองตัวเรียนรู้พื้นฐาน 1 ตัว คือ 1 รอบการคำนวณ แต่หากเปลี่ยนมุมมองเสียใหม่ โดยให้ 1 รอบการทำงานคือกลุ่มของการเรียนรู้ที่ประกอบไปด้วยตัวเรียนรู้พื้นฐานที่ใช้เทคนิคการเรียนรู้ของเครื่องและการคำนวณทางสถิติ แล้วทำการกำหนดรอบการทำงานโดยให้มีการวนรอบในจำนวนที่มากขึ้น ซึ่งก็น่าจะทำให้การจำแนกข้อมูลมีค่าความถูกต้องที่มากขึ้นตามไปด้วย นอกจากนี้วิธีการที่ได้นำเสนอยังสามารถทำการปรับเปลี่ยนเทคนิคการเรียนรู้พื้นฐานทั้งเทคนิคการเรียนรู้ของเครื่องและเทคนิคการคำนวณเชิงสถิติ เพื่อให้มีความเหมาะสมสำหรับโจทย์ปัญหาที่แตกต่างกันไปหรือตามความซับซ้อนของข้อมูลเพื่อหาผลลัพธ์ที่ให้ผลในการจำแนกประเภทได้ดีที่สุด

เอกสารอ้างอิง

[1] A. Reiss, G. Hendeby and D. Stricker, "A novel confidence-based multiclass boosting algorithm for mobile physical activity monitoring," Spinger, Personal and Ubiquitous Computing, vol. 19, pp. 105 – 121, January 2015.

[2] Anusorn, C., and Saichon J. A New Ensemble Model based on Linear Mapping, Nonlinear Mapping, and Probability Theory for Classification Problems. 12th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2015.

[3] Lichman, M. (2013). **UCI Machine Learning Repository.** [Online]. Available : http://archive.ics.uci.edu/ml Irvine, CA: University of California, School of Information and Computer Science.

[4] Volker, L. (2012). **UCI Machine Learning Repository : Banknote Authentication.** [Online]. Available : https://archive.ics.uci.edu/ml/datasets/banknote+authentication.

[5] Terry, S. (1988). **UCI Machine Learning Repository : Connectionist Bench (Sonar, Mines vs. Rocks).** [Online]. Available : https://archive.ics.uci.edu/ml/datasets/Connectionist+Bench+%28Sonar%2C+Mines+vs.+Rocks%29.

[6] Fisher, R. A. (1988). **UCI Machine Learning Repository : Iris.** [Online]. Available : https://archive.ics.uci.edu/ml/datasets/Iris.

[7] Attila, R. (2012). **UCI Machine Learning Repository : PAMAP2 Physical Activity Monitoring.** [Online]. Available : https://archive.ics.uci.edu/ml/datasets/PAMAP2+Physical+Activity+Monitoring.

[8] Nakai, K. (1996). **UCI Machine Learning Repository : Ecoli.** [Online]. Available : https://archive.ics.uci.edu/ml/datasets/Ecoli.

[9] German, B. (1987). **UCI Machine Learning Repository : Glass Identification.** [Online]. Available : https://archive.ics.uci.edu/ml/datasets/Glass+Identification.

[10] Bohanec, M. (1997). **UCI Machine Learning Repository : Car Evaluation.** [Online]. Available : https://archive.ics.uci.edu/ml/datasets/Car+Evaluation.

[11] Bohanec, M., Rajkovic, V. 1990. "Expert System for Decision Making." **Sistemica 1.** เล่มที่ 1 : 145-157.

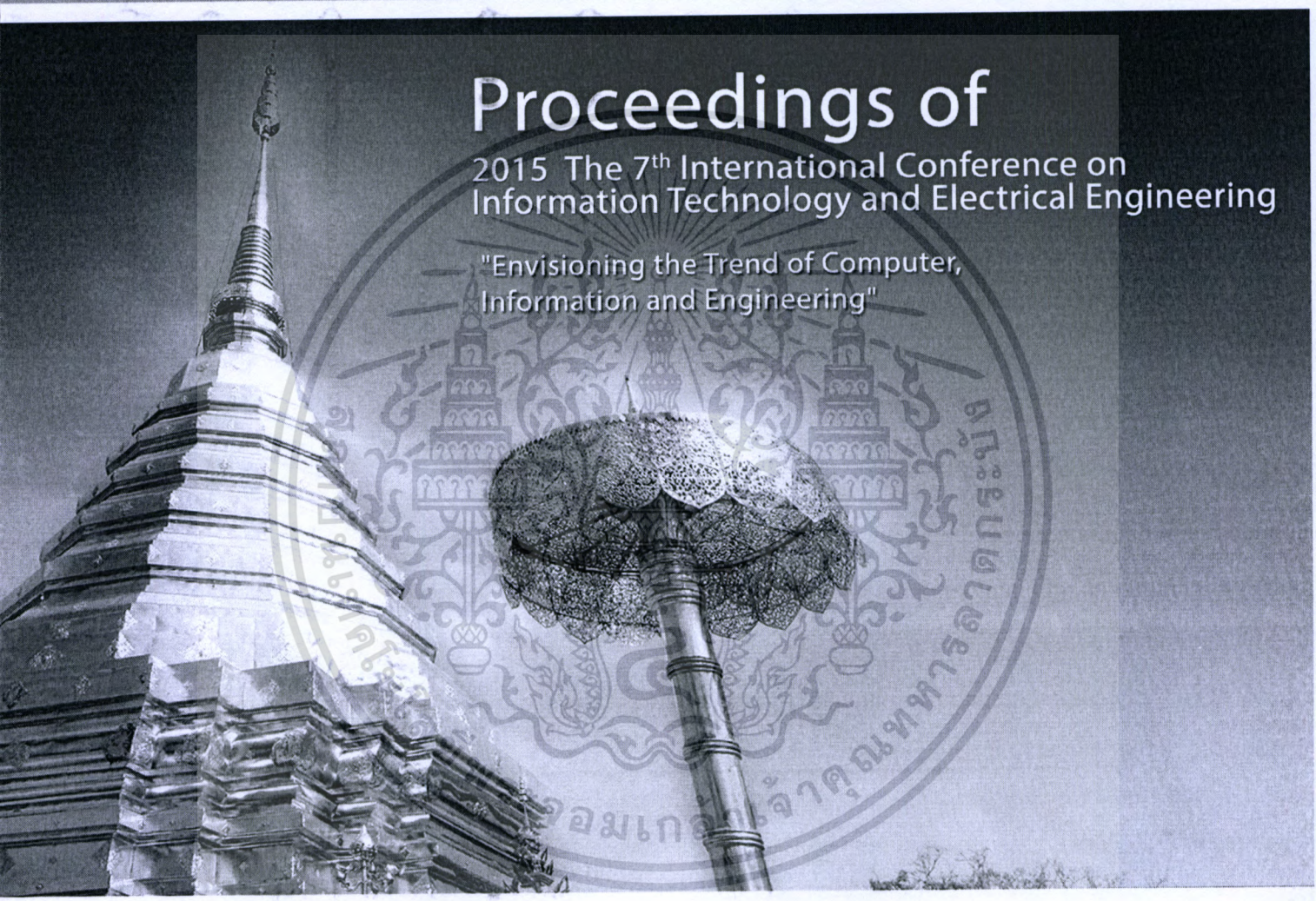
ภาคผนวก ก
งานวิจัยที่ตีพิมพ์

ISBN: 978-1-4673-7862-8

IEEE
Advancing Technology for Humanity

ICITEE2015

2015 The 7th International Conference on
Information Technology and Electrical Engineering



Proceedings of

2015 The 7th International Conference on Information Technology and Electrical Engineering

"Envisioning the Trend of Computer, Information and Engineering"

Le Méridien Chiang Mai Hotel, Thailand

29-30 October 2015

Organized by
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang, THAILAND

Co-organized by
Department of Electrical Engineering and Information Technology
Universitas Gadjah Mada, INDONESIA

A Hybrid Ensemble of Machine and Statistical Learning Using Confidence-Based Boosting

Nattawut Chairatanasongporn
Department of Computer Science, Faculty of Science
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
S3650851@kmitl.ac.th

Saichon Jaiyen
Department of Computer Science, Faculty of Science
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
kjsaicho@kmitl.ac.th

***Abstract*—Nowadays, the classification problems have become more challenging due to the various types of data set. Some data are appropriated for machine learning techniques and some data are appropriated for statistical leaning techniques. This work proposes a new hybrid ensemble of machine and statistical learning models using confidence-based boosting. The proposed method which uses variants of based classifiers can solve classification problems in variant data set. Moreover, combining the confidence value to the current boosting method can improve the performance of classification. The performance of proposed method is compared to the ensemble of decision trees and MRN created by Adaboost.M1 on data sets from UCI. The experimental results show that the proposed method can improve the accuracy in both binary and multiclass classification problems.**

Keywords—Machine learning; Adaboost; Multilayer Perceptron Neural Network; Decision Tree; Discriminant

I. INTRODUCTION

Ensemble learning combines classifiers for improving classification accuracy. In machine learning, the ensemble method is used for solving the classification and pattern recognition problems. AdaBoost algorithm is the popular method for creating ensemble [1] [2]. Reiss, Hendeby and Stricker [3] proposed ConfAdaBoost.M1 algorithm for activity recognition and an intensity estimation problem. The dataset used in the paper was physical activities in mobile systems called PAMAP2 dataset. The algorithm was a variant of the AdaBoost.M1 that incorporated well-established ideas for confidence-based boosting. From the experimental results, the new classification method outperformed the commonly used classifiers, such as decision trees or AdaBoost.M1. Rodriguez, Kuncheva and Alonso [4] suggested a new classifier ensemble method called Rotation Forest. This algorithm generated the ensemble of classifiers based on Principal Component Analysis (PCA). Rotation forest split the feature set into K subsets. From the experimental result, those Rotation Forest ensembles were more accurate than AdaBoost and Random Forest method. Yan, Yu, cooper and Ling [5] suggested a novel parallel BN learning algorithm called PENBays (Parallel Ensemble based Bayesian Networks Learning). This method consisted of three phases which were Data Preprocess (DP), Individual Ensemble Learning (IEL) and Central Ensemble. Learning (CNL) used Bayesian Network (BN) to analyze Big Data. Woo and Park [6] proposed ensemble learning method for semi supervised classification. This method combined label propagation and ensemble learning for predicted unlabeled data samples. Fu, Zhang, Zhao and Li [7] proposed AdaBoost.FT algorithm which each weak classifier of AdaBoost.FT algorithm used the floating threshold to obtain the outputs of classifiers by the distribution on the training samples. From these experimental results, AdaBoost.FT outperformed the real AdaBoost. An and Kim [8] proposed a novel method to inject diversity into the AdaBoost (DAdaBoost algorithm) process to improve the performance of the AdaBoost classifiers. They designed optimized ensemble classifiers with diversity. Thai Hoang Le and Len Tien Bui [9] proposed a hybrid model of combining AdaBoost and Artificial Neural Network for detecting human faces. From the experimental result, this method was able to efficiently detect more than 96%. B. Zhang and G. Han [10] proposed subcellular phenotype images classification by RS-MLP ensembles with the benchmarking 2D HeLa images to solve the challenging phenotype classification problem. From experimental result, random subspace ensemble was able to give satisfactory result of 95% classification accuracy. C. Anusorn [11] proposed the MRN algorithm using Adaboost.M1 as ensemble learning method. The ensemble was based on 3 different classification categories which were the linear mapping, non-linear mapping, and probability theory. The experiment used various data set and the results showed in good accuracy.

AdaBoost method has been popular to create ensemble classifiers for solving classification problems and pattern recognition problems. The ensemble models consist of many weak classifiers and used weighted majority voting in order to improve the accuracy of classifiers. Nowadays, statistical learning and machine learning techniques are popular and widely used to analyze big data. In addition, some data are appropriated for machine learning techniques and some data are suitable for statistical learning techniques. In this paper, we propose a new hybrid ensemble of machine learning and statistical learning techniques using confidence-based boosting to decide the class of input vectors.

978-1-4673-7863-5/15/$31.00 ©2015 IEEE

II. Related Work

A. Multilayer Perceptron Neural Network

It is likely to say that one of the most popular machine learning technique nowadays is artificial neural network. It plays a huge role in many classification problems for the last few decades. It is a machine learning model inspired by biological mechanism of human brain to solve a specific problem by learning from a given data. In addition, Multilayer Perceptron (MLP) Neural Network is widely used due to its reliability and efficiency in classification problems. MLP is a feedforward artificial neural network that utilizes a supervised learning technique called the back propagation to train the network which can be described in Algorithm 1.

Algorithm 1: Back propagation algorithm

1. Initialize all weights to small random numbers.
2. For each $(\mathbf{x_i}, d_i)$ in training examples
3. Input $\mathbf{x}_i$ to the network and compute y_u of every unit u
4. For each output unit k, calculate the local gradient δ_k
$$\delta_k = (d_k - y_k)y_k(1 - y_k)$$
5. For each hidden unit j, calculate the local gradient δ_j
$$\delta_j = y_i(1 - y_j)\sum_k \delta_k w_{jk}$$
6. Update each weight w_{ij}
$w_{ij} = w_{ij} + \Delta w_{ij}$ where $\Delta w_{ij} = \eta \times \delta_j \times y_i$
7. Repeat steps 2 to 6 until the stopping criteria is satisfied.

B. Decision Tree

Decision tree (DT) is a tree-like predictive model technique that can be used for classification problems. The most commonly decision tree model is like an up-side down tree-like. The root on the top of the model called root node can be expand down to many branches called branch node downward along the model. At the bottom of the model are leaf. Each leaf represents a class. Each internal node represents a test on an attribute. Each branch represents the outcome of the test.

C. Discriminant analysis

Discriminant analysis is a technique that used for predicting a class of data by attempting to determine several discriminant functions. R. Fisher [12] introduced the linear discriminant analysis (LDA) for classification problems. Unlike other generalized linear models, it assumes that the independent variables follow a multivariate normal distribution. The predict classifier model for discriminant analysis is defined as follows:

$$\hat{y} = arg\ \min_{y=1,\dots,K} \sum_{k=1}^{K} \hat{P}(k|x)C(y|k) \quad (1)$$

where $\hat{y}$ is the predicted class. K is the number of classes. $\hat{P}(k|x)$ is the posterior probability of class k for observation x. $C(y|k)$ is the conditional probability of y given k.

D. Ensemble Learning

The ensemble learning is machine learning technique based on supervised learning. It is used for generating a group of classifiers called weak learner to solve classification problems. Typically, the ensemble learning technique can be divided into two sub categories. The first one called bagging, proposed by Breiman [13], and the second one is boosting proposed by Freund and Schapire [14] [15]. Since ConfAdaboost.M1 and MRN both are Adaboost.M1 with modified version, it should be good to look into these methods before proposing a new model which is modified from these algorithms. The Adaboost.M1 algorithm can be shown in Algorithm 2.

Algorithm 2: Adaboost.M1

Input: Training set $S = \{x_i, y_i\}, i = 1, \dots, N$ and $y_i \in \mathbb{C}, \mathbb{C} = \{c_1, \dots, c_m\}$ T: Nummber of iterations; W: Weak learner

Output: Boosted classifier:
$H(x) = \arg\max_{y \in \mathbb{C}} \sum_{t=1}^{T} \ln\left(\frac{1}{\beta_t}\right) I\,[h_t(x) = y]$ where h_t, β_t are the induced classifiers (with $h_t(x) \in \mathbb{C}$) and their assigned weighs, respectively

Method:

1. $D_1(i) \leftarrow \frac{1}{N}$, $i = 1, \dots, N$
2. **for** t = 1 to T **do**
3. $h_t \leftarrow W(S, D_t)$
4. $\varepsilon_t \leftarrow \sum_{i=1}^{N} D_t(i) I[h_t(x_i) \neq y_i]$
5. **if** $\varepsilon_t > 0.5$ **then**
6. $T \leftarrow t - 1$
7. **return**
8. **end if**
9. $\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t}$
10. $D_{t+1}(i) = D_t(i) \cdot \beta_t^{1-I[h_t(x_i) \neq y_i]}$ for $i = 1, \dots, N$
11. Normalize D_{t+1} to be a proper distribution
12. **end for**

III. Proposed Method

The proposed method adopt machine learning and statistical learning to create the ensemble model by using the idea of confidence values to boost a weight for each weak learner. In this paper, MLP and Decision Tree which are machine learning techniques are used to create the ensemble model. In addition, Discriminant Analysis which is the statistical learning technique is also used to create the ensemble model. Thus, the base classifiers in our ensemble model

consists of MLP, Decision Tree, and Discriminant Analysis. The proposed method can be described in detail as follow.

Algorithm 3: Hybrid Ensemble with Confidence-Based Boosting (conf HE)

Input: Training data $S = \{x_1, \dots, x_n\}, x_i \in X$ with correct class $y_i \in \Omega, \Omega = \{y_1, \dots, y_c\}$

Output: $H(x) = arg\ \max_{y \in \Omega} \sum_{t=1}^{T} (\beta_t \cdot p_t)$

Method:

1. Initialize the distribution $D_1(i) = \frac{1}{n}, i = 1, 2, \dots, n$
2. Initialize T equal to the number of classifiers
3. **for** t = 1 to T **do**
4. **do**
5. Receive the hypothesis $h_t : X \to \Omega$,
$$h_t = \begin{cases} \text{MLP } if\ t = 1 \\ \text{DT } \ if\ t = 2 \\ \text{LDA } if\ t = 3 \end{cases}$$
6. Compute confidence value for each model
$$p_t = \frac{\sum_{j:H(x_j)=y_j} D_j}{\sum_j D_j}$$
7. Compute the error of h_t:
$$\varepsilon_t = \sum_{i=1}^{n} I[h_t(x_i) \neq y_i] \cdot D_i \cdot p_t$$
where ε_t is error of h_t and y_i is the desired class.
8. **while** $\varepsilon_t > 0.5$
9. Calculate the normalized error
$$\beta_t = \begin{cases} \beta_t = 1\ if\ \varepsilon_t = 0 \\ \frac{1}{2} log \frac{(1-\varepsilon_t)}{\varepsilon_t}\ otherwise \end{cases}$$
10. Update the distribution D_t:
$$D_{t+1}(i) = D_t(i) \cdot e^{\left(\frac{1}{2} - I[h_t(x_i)=y_i]\right)\beta_t p_t}$$
where z_t is a normalization constant.
11. **end for**

Let $S = \{x_1, x_2, \dots, x_n\}$ be the training data set where x_i is an input vector. First, initialize weight distribution D_i and the number of based classifiers T (which is 3 in this case). After initialization, the weak learners, which are the MLP, DT and LDA will be trained using training set respectively. Each time it is passed to weak learning process it will calculate confidence value p_t for each input vector in line 6 which later will be used to calculate the error rate ε_t in the next line. Unlike the other Adaboost algorithm such as the Adaboost.M1 or the confAdaboost.M1, the proposed model will not terminate model generation when error rate is more than 0.5. However, the model will continue to try to generate another weak learner that has error rate less than the criterion instead. After calculated β_t which are a voting weights for each weak learner (In this case MLP, Decision Tree and Discriminant respectively), it will use the confidence value to adjust the weight. The more confident the weak learner is to correct or miss in classification, the more that instance's error weight is reduced or increased respectively. After predicting with current model, the confidence value is used to boost the model weight. The more confidence of the model, the more boost for voting weight. That means if the model is likely to correct classification in training weak learner, when it comes to voting a class, it will have more voting weight than another weak learner that has not likely to correct classification.

IV. Experiments and Result

In this paper, the propose method is ensemble method used to solve classification problems. The experiment results of the propose method are compared with Decision Tree Ensemble using AdaboostM1 algorithm and the MRN algorithm. The experiments are conducted on Matlab software. The weak learners in the proposed model consist of three classifiers including MLP, Decision Tree, and Discriminant Analysis (DA) algorithm. Thus, there are three base classifiers in our model. The MRN algorithm includes MLP, radial basis function (RBF), and Naive Bayes classifier. In this experiment, we use three fold cross validation to train and test the ensemble models. The data set is divided into three subsets, the two subsets are used for training and another is used for testing. The accuracy of all methods in the experimental results are the averaged value of three subsets of each data set.

A. The data sets

In this experiment, the data sets are collected from UCI machine learning database that are used for training and testing the ensemble model. The details of the data sets are shown in Table I. The PAMAP2 [16] [17] data set from UCI contain 3,850,505 instances with 52 features. Since this data set is very large and could use a lot of time to run the experiment, so the data set is preprocessed to reduce the amount of instances and features. The instances are reduced into 21,401 instances and features are reducing into 52 features. In addition, the number of activities are reduced from 24 activities to 18 activities. The other data set was roughly preprocessed but the amount of instance wasn't reduced.

B. The Experimental Results

For this experiment, the proposed model is compared with Adaboost.M1 and MRN algorithms. The tests are run 5 times using 3 fold cross validation technique. The Adaboost.M1 use 3 iterations with decision tree algorithm. Table II shows the comparative results using IRIS data set. The proposed model (conf HE) can achieve the highest averaged accuracy of 96.14%. The Adaboost.M1 can achieve the averaged accuracy at approximately 94.14% while the MRN method can achieve the averaged accuracy of 95.33%.

TABLE I. THE DATA SET USED IN THE EXPERIMENTS

Data sets	No. of class	No. of attributes	No. of instances
Iris	3	4	150
Banknote Authentication	3	5	1372
PAMAP2	18	52	21401
Connectionist Bench	2	60	208

TABLE II. THE COMPARATIVE RESULT USING UCI IRIS DATA SET

Round	AdaboostM1	MRN	conf HE
	Accuracy (%)	Accuracy (%)	Accuracy (%)
1	95.34	96	95.99
2	90.67	94	96.77
3	96.67	95.27	95.27
4	94.65	96.03	95.33
5	93.35	95.33	97.33
Average	94.14	95.33	96.14

Table III demonstrates the comparative results using banknote authentication data set. From the experimental results, the average accuracy of the proposed model (conf HE) can produce the highest average accuracy while Adaboost.M1 and MRN ensemble can achieve the average accuracy at approximately 98.04% and 93.37%, respectively.

Table IV shows the comparative results using PAMAP2 data set. From the experimental results, it can be seen that the proposed model (conf HE) can get slightly higher average accuracy than the adaboostM1 which can achieve at 91.46% and 91.4%, respectively. Whereas the MRN can achieve the least accuracy at 83.34%.

Table V demonstrates the comparative results using connectionist bench (Sonar, Mines Vs. Rocks) data set. From the experimental results, the proposed model (conf HE) can obviously perform the best average accuracy at approximately 76.73%. The adaboost.M1 and MRN ensemble achieve the average accuracy at 68.37% and 66.55%, respectively.

TABLE III. THE COMPARATIVE RESULT USING UCI BANKNOTE AUTHENTICATION

Round	AdaboostM1	MRN	conf HE
	Accuracy (%)	Accuracy (%)	Accuracy (%)
1	97.81	93.5	99.56
2	98.40	93.3	99.13
3	97.74	93.37	99.5
4	97.89	93.5	98.76
5	98.40	93.2	98.18
Average	98.04	93.37	99.02

TABLE IV. THE COMPARATIVE RESULT USING UCI PAMAP2 DATA SET

Round	AdaboostM1	MRN	conf HE
	Accuracy (%)	Accuracy (%)	Accuracy (%)
1	91.22	83.3	91.67
2	91.77	83.2	91.08
3	91.49	83.5	91.58
4	91.16	83.3	91.37
5	91.36	83.4	91.63
Average	91.40	83.34	91.46

TABLE V. THE COMPARATIVE RESULT USING UCI CONNECTIONIST BENCH (SONAR, MINES VS. ROCKS) DATA SET

Round	AdaboostM1	MRN	conf HE
	Accuracy (%)	Accuracy (%)	Accuracy (%)
1	64.44	70.2	74.99
2	66.83	66.3	78.85
3	72.57	66.3	77.4
4	67.81	64.1	71.61
5	70.21	65.86	80.80
Average	68.37	66.55	76.73

V. CONCLUSION

This paper proposes the new hybrid ensemble of MLP, Decision Tree, and Discriminant Analysis using Adaboost.M1 with confident value. The proposed model consists of three base classifiers that come from machine learning techniques and statistical learning techniques for solving the diversity of data set. Using the confidence value to boost the weights of data and classifiers can improve the performance of classifiers. The performance of the proposed model is evaluated and compared to the ensemble of Decision Trees generated by Adaboost.M1 and MRN algorithms. The experimental results show that the proposed model can achieve the best average accuracy in all data sets.

REFERENCES

[1] M. Galar, A. Fernández, E. Barrenechea, H. Bustince and F. Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," IEEE Trans. Systems, Man, and Cybernetics, vol. 42, pp. 463 - 484, July 2012.

[2] R. Polikar, "Bootstrap - Inspired Techniques in Computation Intelligence," IEEE Tarns. Signal Processing Magazine, vol. 24, pp. 59 – 72, July 2007.

[3] A. Reiss, G. Hendeby and D. Stricker, "A novel confidence-based multiclass boosting algorithm for mobile physical activity monitoring," Spinger, Personal and Ubiquitous Computing, vol. 19, pp. 105 – 121, January 2015.

[4] J.J. Rodriguez, L.I. Kuncheva and C.J. Alonso, "Rotation Forest: A New Classifier Ensemble Method," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, pp. 1619 – 1630, Oct. 2006.

[5] T. Yan, W. Yu, K.M.l. Cooper and L. Li, "Towards Big Data Bayesian Network Learning - An Ensemble Learning Based Approach," IEEE Big Data (BigData Congress), pp. 355 - 357, , July 2 2014.

[6] H. Woo and C.H. Park, "Semi-supervised Ensemble Learning Using Label Propagation," IEEE. Computer and Information Technology, pp. 421 – 426, Oct. 2012.

[7] Z. Fu, D. Zhang, X. Zhao , X. Li, "Adaboost algorithm with floating threshold," IEEE. Automatic Control and Artificial Intelligence. pp. 349-354, March 2012.

[8] 13. T.An and M. Kim, "A New Diverse AdaBoost Classifier,". IEEE Artificial Intelligence and Computational Intelligence, pp. 359-363, Oct. 2010.

[9] T. H. Le and L. T. Bui, "A Hybrid Approach of AdaBoost and Artificial Neural Network for Detecting Human Faces," IEEE. Research, Innovation and Vision for the Future. pp. 79-85, July 2008.

[10] B.l. Zhang and G. Han, "Subcellular Phenotype Images Classification by MLP Ensembles with Random Linear Oracle," IEEE. Bioinformatics and Biomedical Engineering, pp. 1-4, May 2011.

[11] Anusorn, C., and Saichon J. A New Ensemble Model based on Linear Mapping, Nonlinear Mapping, and Probability Theory for Classification Problems. 12th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2015

[12] Guo, Y., T. Hastie, and R. Tibshirani. Regularized Discriminant Analysis and Its Application in Microarray. Biostatistics, Vol. 8, No. 1, pp. 86–100, 2007.

[13] Breiman, L. 1996. Bagging predictors. Machine Learning, forthcoming.

[14] Freund, Y., and Schapire, R. E. 1996a. A decision-theoretic generalization of on-lin learning and an application to boosting. Unpublished manuscript, available from the authors' home page ("http://www.research.att.com/org/ssr/people/{yoav,schapire}"). An extened abstact appears in Compuional Learning Theory: Second European Conference, EuroCOLT '95, 23-27, Springer-Verlag, 1995.

[15] Freund, Y., and Schapire, R, E. 1996b. Experiments with a new boosting algorithm. Unpublished manuscript.

[16] A. Reiss and D. Stricker. "Introducing a New Benchmarked Dataset for Activity Monitoring," The 16th IEEE International Symposium on Wearable Computers (ISWC), 2012.

[17] A. Reiss and D. Stricker. "Creating and Benchmarking a New Dataset for Physical Activity Monitoring," The 5th Workshop on Affect and Behaviour Related Assistance (ABRA), 2012.

ประวัติผู้เขียน

ชื่อ	นายณัฐวุฒิ ชัยรัตนทรงพร
วัน เดือน ปีเกิด	3 มีนาคม พ.ศ. 2531
ที่อยู่ปัจจุบัน	66/6 หมู่14 ตำบลตะพง อำเภอเมืองระยอง จังหวัดระยอง 21000
ประวัติการศึกษา	2553 วิทยาศาสตรบัณฑิต สาขาวิทยาการคอมพิวเตอร์ประยุกต์ เกรดเฉลี่ย 2.51 มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
	2558 วิทยาศาสตรมหาบัณฑิต สาขาวิทยาการคอมพิวเตอร์ เกรดเฉลี่ย 3.56
	สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ผลงานทางวิชาการ	1.“A Hybrid Ensemble of Machine and Statistical Learning Using Confidence-Based Boosting”, 2015 7th International Conference on Information Technology and Electrical Engineering: ICITEE