การวิเคราะห์การตลาดทางตรงของธนาคารบนข้อมูลแบบอสมมาตร โดยใช้การเรียนรู้ของเครื่อง

BANK DIRECT MARKETING ANALYSIS OF ASYMMETRIC DATA USING MACHINE LEARNING

ภูมิธารา เรืองทอง
PUMITARA RUANGTHONG

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต
ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2558
KMITL-2015-SC-M-002-067

การวิเคราะห์การตลาดทางตรงของธนาคารบนข้อมูลแบบอสมมาตร โดยใช้การเรียนรู้ของเครื่อง

BANK DIRECT MARKETING ANALYSIS OF ASYMMETRIC DATA USING MACHINE LEARNING

ภูมิธารา เรืองทอง

PUMITARA RUANGTHONG

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต

ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2558

KMITL-2015-SC-M-002-067

BANK DIRECT MARKETING ANALYSIS OF ASYMMETRIC DATA USING MACHINE LEARNING

PUMITARA RUANGTHONG

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-002-067

คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์ "การวิเคราะห์การตลาดทางตรงของธนาคารบนข้อมูลแบบอสมมาตร โดยใช้การเรียนรู้ของเครื่อง"
"BANK DIRECT MARKETING ANALYSIS OF ASYMMETRIC DATA USING MACHINE LEARNING"

ชื่อนักศึกษา นางสาวภูมิธารา เรืองทอง
รหัสประจำตัว 56605030
ปริญญา วิทยาศาสตรมหาบัณฑิต (สาขาวิทยาการคอมพิวเตอร์)
ภาควิชา วิทยาการคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์ ดร.สายชล ใจเย็น
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม (ถ้ามี) -----

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

วัน/ เดือน/ ปี ที่สอบ 1 ธันวาคม พ.ศ. 2558 เวลา 15.00 - 17.00 น.
สถานที่สอบ ณ ห้อง 306 ตึกปฏิบัติการหลังใหม่

คณะวิทยาศาสตร์รับรองแล้ว

(รองศาสตราจารย์ ดร.ดุษณี ธนะบริพัฒน์)
คณบดีคณะวิทยาศาสตร์
วันที่ 21 เดือน ธันวาคม พ.ศ. 58

หัวข้อวิทยานิพนธ์	การวิเคราะห์การตลาดทางตรงของธนาคารบนข้อมูลแบบอสมมาตรโดยใช้การเรียนรู้ของเครื่อง
ชื่อนักศึกษา	ภูมิธารา เรืองทอง
รหัสประจำตัว	56605030
ปริญญา	วิทยาศาสตรมหาบัณฑิต สาขาวิทยาการคอมพิวเตอร์
ภาควิชา	วิทยาการคอมพิวเตอร์
พ.ศ.	2558
อาจารย์ที่ปรึกษาวิทยานิพนธ์	อาจารย์ ดร.สายชล ใจเย็น

บทคัดย่อ

ปัญหาคลาสที่ไม่สมดุลเป็นปัญหาที่ทำให้การจำแนกประเภทได้ผลลัพธ์ไม่เป็นที่น่าพอใจ ไม่ว่าจะใช้การจำแนกประเภทแบบใดก็ไม่สามารถทำให้ผลลัพธ์สูงขึ้นได้ ดังนั้นในงานวิจัยฉบับนี้ จึงได้นำเสนอแบบจำลองการเรียนรู้แบบรวมกลุ่มผสมแบบใหม่ที่ใช้พื้นฐานการเรียนรู้ของเอดาบูสต์เอ็มทู และใช้อัลกอริทึมเอสเอ็มโอทีอี แก้ปัญหาความไม่สมดุลของคลาส เพื่อคาดการณ์ความเป็นไปได้ในการฝากเงินระยะยาวจากลูกค้าธนาคาร แบบจำลองผสมดังกล่าวประกอบไปด้วยอัลกอริทึมสำหรับจำแนกประเภทที่หลากหลาย ได้แก่โครงข่ายแบบเบย์ ต้นไม้ตัดสินใจแบบเอดีทรี ต้นไม้ตัดสินใจแบบเจโฟร์ทีเอต และต้นไม้ตัดสินใจแบบอาร์อีพีทรี จากผลการทดลองแบบจำลองที่งานวิจัยนี้นำเสนอ สามารถเพิ่มประสิทธิภาพการจำแนกได้ทุกด้านของการวัดผลเมื่อเทียบกับแบบจำลองการเรียนรู้แบบรวมกลุ่มปกติและแบบจำลองการเรียนรู้แบบรวมกลุ่มที่ใช้วิธีปรับลดคลาสที่มีจำนวนมากกว่า

คำสำคัญ : การตลาดทางตรง การเรียนรู้แบบต้นไม้ตัดสินใจ การเรียนรู้แบบกลุ่ม โครงข่ายแบบเบย์

Thesis Title	Bank Direct Marketing Analysis of Asymmetric Information Based Using Machine Learning
Student Name	Pumitara Ruangthong
Student ID	56605030
Degree	Master of Science Computer
Department	Computer Science
Year	2015
Thesis Advisor	Dr.Saichon Jaiyen

Abstract

Class imbalance problem is the main issue causing unsatisfactory outcome in classification. Any type of classification used still cannot improve the result. Therefore, in this research we propose a new hybrid ensemble model based on Adaboost.M2 and adopt SMOTE algorithm to solve the class imbalance problem in order to predict the probability of term deposit from bank customers. The proposed hybrid ensemble model consist of diverse based classifiers which are Bayesian network, Alternating decision tree, Tree-J48 and REPTree. From the experimental results, the proposed model can improve the prediction accuracy in all aspects when compared to a normal ensemble model and an ensemble model that uses majority class reduction.

Keywords: Bayesian Network, Decision Tree, Direct Marketing, Ensemble Learning

กิตติกรรมประกาศ

ขอขอบพระคุณอาจารย์ที่ปรึกษาอาจารย์ ดร.สายชล ใจเย็น ที่ได้สละเวลาตรวจแก้งานวิจัยเปิดสอนพิเศษการเรียนรู้ของเครื่อง (Machine Learning) โดยไม่คิดค่าใช้จ่าย หากมิได้รับคำแนะนำจากอาจารย์ งานวิจัยฉบับนี้คงไม่ได้สำเร็จลุล่วงไปด้วยดี จึงใคร่ขอขอบพระคุณอาจารย์เป็นอย่างสูง

ขอขอบพระคุณอาจารย์ ผศ.ดร.อนันตพร หรรษคุณาฒัย อาจารย์ ผศ.ดร.ศุภกานต์ พิมลธเรศ และอาจารย์ ผศ.ดร.ศรัณย์ อินทโกสุม กรรมการสอบวิทยานิพนธ์ ที่ได้ให้คำแนะนำ เพื่อปรับปรุงเล่มวิทยานิพนธ์ เสนอแนะแนวทางแก้ปัญหา เพื่อให้วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปด้วยดี ขอขอบพระคุณอาจารย์ทุกท่านมากค่ะ

ขอขอบคุณนายยศภัทร เรืองไพศาล นายนิพัทธ์ คล้ายโพธิ์ นายณัฐวุฒิ ชัยรัตนทรงพร นางสาวปาริฉัตร นาคสิทธิ์ ว่าที่ร.ต.เสฏฐนันท์ ทองสุวรรณ์ อาจารย์อนุสรณ์ เจริญนาน และเพื่อนๆ พี่ๆ ทุกท่านที่ได้คอยช่วยเหลือ ให้คำปรึกษาแก่ข้าพเจ้า

ขอขอบพระคุณครอบครัว คุณน้า คุณยาย ที่ได้ให้การสนับสนุนค่าใช้จ่ายตลอดการศึกษาเล่าเรียน และขอขอบพระคุณมารดา บิดา คุณพี่จอย ซึ่งเป็นกำลังใจให้แก่ข้าพเจ้าเสมอ

นางสาวภูมิธารา เรืองทอง

สารบัญ

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญรูป

รูปที่ **หน้า**

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การตลาดทางตรงคือรูปแบบการทำตลาดรูปแบบหนึ่ง ซึ่งเป็นการสื่อสารที่เจาะเข้าถึงผู้บริโภคเป็นรายบุคคลและสามารถจัดหาสินค้าและบริการที่เหมาะสมเพื่อให้สอดคล้องกับผู้บริโภคแต่ละราย เช่นการติดต่อผู้บริโภคผ่านทางโทรศัพท์ การส่งจดหมายนำเสนอสินค้าพร้อมแบบตอบรับทางไปรษณีย์ การขายสินค้าผ่านทางอินเตอร์เน็ต เป็นต้น ในงานวิจัยนี้ได้ศึกษาการตลาดทางตรงของธนาคาร ซึ่งใช้การติดต่อผู้บริโภคแต่ละรายด้วยวิธีการสื่อสารผ่านทางโทรศัพท์เพื่อนำเสนอผลิตภัณฑ์เงินฝากระยะยาว แต่เนื่องจากลูกค้าของธนาคารมีจำนวนมหาศาล การติดต่อลูกค้าทุกรายเป็นไปได้ยากรวมถึงต้องใช้ต้นทุนทั้งทางด้านการสื่อสาร ต้นทุนแรงงานที่ใช้ในการติดต่อลูกค้า และระยะเวลาในการติดต่อที่มากตามไปด้วย ดังนั้นธนาคารจึงจำเป็นต้องกำหนดกลุ่มเป้าหมายของลูกค้าที่มีแนวโน้มจะสนใจในผลิตภัณฑ์เงินฝากระยะยาว ซึ่งจะทำให้ปริมาณต้นทุนทรัพยากรที่ธนาคารจะใช้ในการลงทุนเพื่อติดต่อสื่อสารกับลูกค้าลดลง เช่นผู้ติดต่อเลือกลูกค้ากลุ่มเป้าหมายจำนวน 40,000 คน จากลูกค้าจำนวน 1,000,000 คน ซึ่งใน 40,000 คนนี้เป็นกลุ่มลูกค้าที่มีแนวโน้มสนใจในผลิตภัณฑ์เงินฝากระยะยาว ก็จะทำให้ต้นทุนในการสื่อสารกับลูกค้าน้อยลง แต่จะได้รับการตอบรับฝากเงินระยะยาวจากลูกค้าเป็นจำนวนที่มากขึ้นหรือน้อยลง ก็ขึ้นอยู่กับว่าธนาคารสามารถกำหนดกลุ่มเป้าหมายได้อย่างแม่นยำหรือไม่ ซึ่งในการทำธุรกรรมผ่านทางธนาคารนั้น ธนาคารจะมีข้อมูลของลูกค้าเก็บรวบรวมไว้แล้ว สามารถนำข้อมูลเหล่านั้นมาทำนายพฤติกรรมของผู้บริโภคและความต้องการในอนาคตได้ หากการจำแนกประเภทกลุ่มเป้าหมายมีความแม่นยำ ก็จะได้รับการตอบรับจากลูกค้าที่ดีตามไปด้วย แต่เนื่องจากปัญหาของข้อมูลที่มีความหลากหลายและซับซ้อน ทั้งส่วนที่นำมาวิเคราะห์ได้และส่วนที่ไม่สามารถนำมาวิเคราะห์ได้ ข้อมูลที่นำมาใช้จึงมีความไม่สมดุลกันของข้อมูลอยู่สูง ทำให้ความแม่นยำในการทำนายผลต่ำ

ดังนั้นวิทยานิพนธ์ฉบับนี้จึงเสนอวิธีเอสเอ็มโอทีอี (SMOTE) ซึ่งช่วยในการปรับเพิ่มข้อมูลให้มีความสมดุล และใช้การแยกประเภทด้วยวิธีการแบบต้นไม้ร่วมกันตัดสินใจ (Decision Tree Ensemble) โดยใช้การทำงานร่วมกันของต้นไม้ตัดสินใจแบบเจโฟร์ทีเอต (J48) ต้นไม้ตัดสินใจแบบอาร์อีพีทรี (REPTree) และต้นไม้ตัดสินใจแบบเอดีทรี (ADTree) ร่วมกับทฤษฎีโครงข่ายแบบเบย์ (Bayesian Network) ในการจำแนกประเภทข้อมูล เนื่องจากเป็นการจำแนกประเภทบนข้อมูลที่ไม่สมดุลจึงใช้วิธีวัดประสิทธิภาพในการจำแนกประเภทของแต่ละคลาสทั้งหมด 6 แบบ คือค่าความแม่นยำของคลาสบวก (Sensitivity) ค่าความแม่นยำของคลาสลบ (Specificity) ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าเอฟเมเซอร์ (F-measure)

1.2 วัตถุประสงค์ของงานวิจัย

1) เพื่อนำเสนอวิธีการปรับข้อมูลที่ไม่สมดุลให้มีความสมดุลสำหรับการวิเคราะห์การตลาดทางตรงของธนาคารโดยการใช้อัลกอริทึมเอสเอ็มโอทีอี
2) เพื่อศึกษาและเปรียบเทียบประสิทธิภาพการจำแนกประเภทของข้อมูลที่ยังไม่มีการปรับความสมดุลของข้อมูล กับข้อมูลที่มีการปรับความสมดุลของข้อมูลแล้ว
3) เพื่อนำเสนอวิธีการจำแนกประเภทข้อมูลด้วยวิธีแบบรวมกลุ่มผสมระหว่างต้นไม้ตัดสินใจแบบเจโฟร์ทีเอท ต้นไม้ตัดสินใจแบบอาร์อีพีทรี ต้นไม้ตัดสินใจแบบเอดีทรี และทฤษฎีโครงข่ายแบบเบย์

1.3 สมมติฐานของงานวิจัย

ปัญหาข้อมูลที่มีความไม่สมดุลกันอย่างสูง ทำให้ไม่ว่าจะใช้การจำแนกประเภทข้อมูลแบบใดก็ไม่สามารถเพิ่มประสิทธิภาพการจำแนกให้สูงขึ้นได้ ดังนั้นจึงต้องใช้วิธีที่สามารถช่วยปรับให้ข้อมูลมีความสมดุลกันขึ้นก่อน ซึ่งวิธีเอสเอ็มโอทีอี ช่วยปรับสมดุลข้อมูลของคลาส (Class) ที่มีน้อยกว่า จึงทำให้เพิ่มประสิทธิภาพในการจำแนก และในการจำแนกประเภทข้อมูลด้วยวิธีการแบบต้นไม้ร่วมกันตัดสินใจ โดยใช้การทำงานร่วมกันของต้นไม้ตัดสินใจแบบเจโฟร์ทีเอท ต้นไม้ตัดสินใจแบบอาร์อีพีทรี และต้นไม้ตัดสินใจแบบเอดีทรี ร่วมกับทฤษฎีโครงข่ายแบบเบย์ จะทำให้การจำแนกประเภทมีประสิทธิภาพสูงขึ้น

1.4 ขอบเขตการวิจัย

1) ข้อมูลที่ใช้ในงานวิจัยเกี่ยวกับการจัดการตลาดทางตรงของธนาคารโปรตุเกส โดยติดต่อลูกค้าผ่านทางโทรศัพท์เพื่อสอบถามความต้องการฝากเงินระยะยาวกับธนาคาร ซึ่งมีทั้งกลุ่มลูกค้าที่มีความต้องการฝากเงินระยะยาวกับกลุ่มลูกค้าที่ไม่มีความต้องการฝากเงินระยะยาวกับธนาคาร
2) วิทยานิพนธ์นี้ใช้วิธีการเรียนรู้ของเครื่อง (Machine Learning) ในการจำแนกประเภทแบบต้นไม้ร่วมกันตัดสินใจ โดยใช้การทำงานร่วมกันของต้นไม้ตัดสินใจแบบเจโฟร์ทีเอท ต้นไม้ตัดสินใจแบบอาร์อีพีทรี และต้นไม้ตัดสินใจแบบเอดีทรี ร่วมกับทฤษฎีโครงข่ายแบบเบย์ เปรียบเทียบประสิทธิภาพการทำนายผลของข้อมูลที่ไม่สมดุลกับข้อมูลที่มีการปรับสมดุลแล้ว

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1) สามารถใช้ในการวิเคราะห์ความต้องการฝากเงินของลูกค้าธนาคาร เพื่อปรับปรุงพัฒนาผลิตภัณฑ์ของธนาคารได้
2) สามารถนำไปประยุกต์ใช้กับการจำแนกประเภทข้อมูลที่มีความไม่สมดุลประเภทอื่น และนำแบบแผนงานวิจัยไปใช้ร่วมกับการทำนายผลข้อมูลในแบบอื่นได้
3) สามารถใช้เพื่อประโยชน์ทางด้านธุรกิจ เพื่อลดต้นทุน ค่าใช้จ่ายในด้านการตลาดได้

บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

ทฤษฎีที่เกี่ยวข้องในบทนี้ได้แก่ ความรู้เบื้องต้นเกี่ยวกับการตลาดทางตรง (Direct Marketing) เทคนิคการทำเหมืองข้อมูล เทคนิคการปรับข้อมูลให้มีความสมดุลโดยใช้อัลกอริทึมเอสเอ็มโอทีอี (SMOTE Algorithm) โครงข่ายแบบเบย์ (Bayesian Network) กระบวนการเรียนรู้แบบต้นไม้ตัดสินใจ (Decision Tree) กระบวนการเรียนรู้แบบรวมกลุ่ม (Ensemble Learning) และงานวิจัยที่เกี่ยวข้อง

2.1 ความรู้เบื้องต้นเกี่ยวกับการตลาดทางตรง

การตลาดทางตรง คือรูปแบบของการทำการตลาด โดยใช้เทคนิคการสื่อสารไปยังผู้บริโภคแบบการโฆษณาในหลากหลายช่องทาง เช่นการส่งข้อความไปยังโทรศัพท์มือถือ ส่งอีเมล โฆษณาออนไลน์ ผ่านทางหน้าเว็บไซต์ ใบปลิว เป็นต้น การตลาดทางตรงมุ่งเน้นการส่งข้อมูลไปที่ผู้บริโภคโดยตรง นอกเหนือจากการโฆษณาแล้วยังมีการจัดกิจกรรมทางการตลาด (Campaign) เพื่อดึงดูดผู้บริโภค

2.1.1 ประวัติการตลาดทางตรง

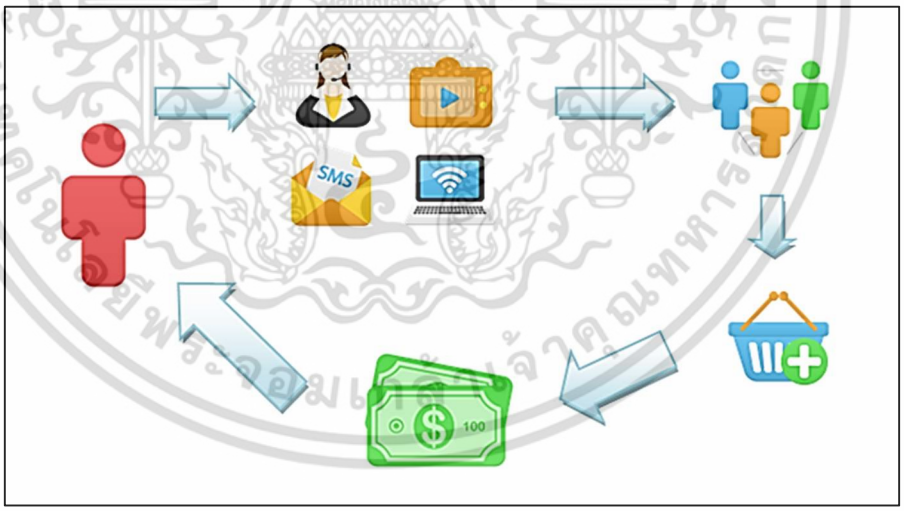
การสั่งซื้อสินค้าทางไปรษณีย์บุกเบิกโดย Aaron Montgomery Ward [1] ซึ่งเชื่อว่าเทคนิคการขายสินค้าโดยตรงให้ผู้บริโภคในราคาที่น่าสนใจสามารถทำได้อย่างมีประสิทธิภาพ จึงนำมาสู่การปฏิวัติอุตสาหกรรมการตลาด สร้างผลิตภัณฑ์การตลาดและสร้างความน่าเชื่อถือ ซึ่งเป็นที่มาของคำว่า “การตลาดทางตรง (Direct Marketing)”

ในปี 1872 Ward [2][3] ได้ผลิตแคตตาล็อก (Catalog) สำหรับการสั่งซื้อสินค้าทางไปรษณีย์ขึ้นเป็นครั้งแรกสำหรับธุรกิจของเขา โดยการเป็นพ่อค้าคนกลางรับซื้อสินค้ามาจำหน่าย และในปี 1967 Lester Wunderman บิดาแห่งการตลาดทางตรง ผู้อยู่เบื้องหลังเลขหมายโทรฟรี 1-800 ได้จัดโปรแกรมการตลาด (loyalty program) ขึ้นเพื่อใช้ในการสร้างความภักดีระหว่างผู้บริโภคกับแบรนด์ (Brand) สินค้า เช่นการออกบัตรสมาชิกโคลัมเบียคลับ (Columbia Record Club) บัตรสมาชิกนิตยสาร และแจกรางวัลให้ลูกค้าที่ใช้จ่ายผ่านบัตรอเมริกันเอ็กซ์เพรส (American Express) เป็นต้น

2.1.2 ประโยชน์ของการตลาดทางตรง

การตลาดทางตรงเป็นที่สนใจของนักการตลาดจำนวนมาก เพราะสามารถวัดผลเชิงบวกได้โดยตรง ตัวอย่างเช่นถ้านักการตลาดส่งคำเชิญชวนทางไปรษณีย์ให้ผู้บริโภค 1,000 คน และได้รับการตอบรับกลับมาเป็นจำนวน 100 คน ถือว่าการจัดกิจกรรมการตลาดประสบความสำเร็จถึง 10 เปอร์เซ็นต์ ซึ่งเป็นตัวชี้วัดอัตราการตอบสนองเชิงปริมาณอย่างชัดเจน ในขณะที่การตลาดโดยทั่วไปใช้การวัดทางอ้อม เช่นการให้ผู้บริโภคมีส่วนรับรู้ถึงการโฆษณาสินค้า แต่ไม่มีการตอบรับเข้าร่วมกับสินค้าและบริการโดยตรงกลับมายังผู้ผลิต

การวัดผลเป็นองค์ประกอบพื้นฐานในการตลาดทางตรงที่ประสบความสำเร็จ [4] ในปัจจุบันอินเตอร์เน็ต (Internet) เข้าถึงผู้บริโภคและเชื่อมโยงไปยังการรับชมหน้าเว็บไซต์โดยตรง ดังนั้นสามารถจัดกิจกรรมส่งเสริมการขายผ่านทางหน้าเว็บไซต์ โดยกระจายการส่งข้อความ โดยมีวิธีการวัดผลคือเปรียบเทียบการคาดการณ์ยอดจำหน่ายที่ต้องการกับยอดจำหน่ายจริงที่เกิดขึ้น เพื่อนำไปสู่การจัดกิจกรรมส่งเสริมการขายของการตลาดทางตรง ความพยายามในการกำหนดเป้าหมายกลุ่มผู้บริโภค เป็นสิ่งที่นักการตลาดหลายคนตระหนักถึง หากกลุ่มผู้บริโภคไม่ตรงตามเป้าหมายอาจทำให้เสียเงินในการโฆษณาสื่อสารกับผู้บริโภคที่ไม่ได้สนใจในผลิตภัณฑ์ แผนภาพการตลาดทางตรงแสดงดังรูปที่ 2.1



รูปที่ 2.1 แผนภาพการตลาดทางตรง

2.2 เทคนิคการทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล คือเครื่องมือหนึ่งในด้านการวิเคราะห์ข้อมูลที่ออกแบบมาเพื่อการสำรวจข้อมูลที่มีจำนวนมาก โดยการค้นหารูปแบบและความสัมพันธ์ที่สอดคล้องกัน ซึ่งช่วยให้ผู้ใช้

สามารถวิเคราะห์ข้อมูลจากขนาดของข้อมูลที่มีความแตกต่างกันและสรุปความสัมพันธ์กันของข้อมูลได้ เช่นตัวอย่างข้อมูลที่ได้จากการวิเคราะห์สามารถลดต้นทุนหรือกำหนดเป้าหมายเพื่อความต้องการในด้านธุรกิจ การตลาด สังคม เศรษฐกิจและการศึกษา เป็นต้น

2.3 ขั้นตอนการทำเหมืองข้อมูล

1) การสำรวจข้อมูล (Exploration) [5] เริ่มต้นด้วยขั้นตอนการเตรียมข้อมูล การแปลงข้อมูลที่เลือกนำมาใช้ในการวิเคราะห์ โดยการเลือกช่วงข้อมูลและตัวแปรที่สามารถจัดการได้ โดยใช้วิธีการทางสถิติช่วยในการจัดเตรียมข้อมูล จากนั้นจะขึ้นอยู่กับคุณลักษณะ (Attribute) ของปัญหาที่วิเคราะห์ โดยการวิเคราะห์อาจเป็นการวิเคราะห์ในรูปแบบทางตรงหรือรูปแบบการถดถอย โดยใช้กระบวนการทางสถิติที่หลากหลายเพื่อระบุตัวแปรที่เกี่ยวข้องมากที่สุดและกำหนดรูปแบบความซ้ำซ้อนของข้อมูลเพื่อใช้ในการพิจารณาขั้นต่อไป
2) การสร้างแบบจำลองและการตรวจสอบ (Model building and validation) [6] ขั้นตอนนี้จะเกี่ยวข้องกับการพิจารณารูปแบบของแบบจำลองและเลือกรูปแบบจำลองที่ดีที่สุด ซึ่งขึ้นอยู่กับประสิทธิภาพการทำนาย
3) การปรับใช้ (Deployment) [7][8] คือการคัดเลือกรูปแบบจำลองที่ดีที่สุดและนำมาปรับใช้กับข้อมูล เพื่อทำนายหรือประมาณการผลลัพธ์ที่คาดหวัง

2.3.1 ประเภทข้อมูลที่นำมาใช้ในการบวนการทำเหมืองข้อมูล

1) ฐานข้อมูลเชิงสัมพันธ์ (Relational Database) เป็นข้อมูลที่มีความสัมพันธ์กันในรูปแบบการจัดเก็บแบบตาราง ซึ่งประกอบด้วยแถวและคอลัมน์ (Column) แสดงความสัมพันธ์ในรูปแบบ Entity Relationship หรือเรียกแบบย่อว่า “อีอาร์ (ER)”
2) รูปแบบของข้อมูลที่มีการเก็บรวบรวมจากหลายแหล่งที่มา ซึ่งเก็บไว้ในรูปแบบเดียวกันและรวบรวมไว้ในที่เดียวกัน (Data Warehouses)
3) ฐานข้อมูลแบบทรานแซกชัน (Transaction Database) จะแทนด้วยเหตุการณ์ในขณะใดขณะหนึ่ง เช่น การกดเงินจากตู้เอทีเอ็ม ข้อมูลที่ถูกเก็บอยู่ขณะนั้น คือ จำนวนเงินที่กด เวลา เลขที่บัญชี เป็นต้น
4) ฐานข้อมูลขั้นสูง (Advanced Database) [9] เป็นข้อมูลที่ถูกจัดเก็บลงฐานข้อมูลในรูปแบบพิเศษ เช่น ข้อมูลที่จัดเก็บในรูปแบบข้อความ (Text File) ข้อมูลมัลติมีเดีย เช่น รูปภาพ เสียง วีดีโอ ข้อมูลในรูปของเว็บไซต์ เป็นต้น

2.3.2 ลักษณะของข้อมูลที่สามารถนำมาทำเหมืองข้อมูล

1) ข้อมูลขนาดใหญ่ ที่ไม่สามารถหาความสัมพันธ์ได้ด้วยการพิจารณาแบบง่าย และไม่สามารถใช้ระบบการจัดการฐานข้อมูล (Database Management System) ในการหาความสัมพันธ์ได้
2) ข้อมูลที่ถูกเก็บรวบรวมมาจากหลากหลายแหล่ง และหลายระบบปฏิบัติการ เช่น Oracl, DB2, MS SQL, MS Access เป็นต้น
3) ข้อมูลที่ไม่มีการเปลี่ยนแปลงตลอดเวลา หากต้องการใช้ข้อมูลนี้ต้องแก้ปัญหาการเปลี่ยนแปลงของข้อมูลก่อน การที่ข้อมูลที่มีการเปลี่ยนแปลงตลอดเวลาไม่เหมาะสมในการทำเหมืองข้อมูลเนื่องจากผลลัพธ์ที่ได้อาจไม่เหมาะสมหรือทำให้การคาดการณ์ไม่ถูกต้องได้
4) ข้อมูลที่มีความซับซ้อน [9] เช่น เช่นข้อมูลเสียง รูปภาพ วีดีโอ แต่การทำเหมืองข้อมูลขั้นสูง มีการเตรียมข้อมูลที่ดี สามารถนำข้อมูลเหล่านี้มาทำเหมืองข้อมูลได้

2.3.3 การจำแนกประเภทข้อมูล (Classification)

การจำแนกประเภทข้อมูลเป็นหนึ่งในเทคนิคของกระบวนการทำเหมืองข้อมูล ใช้เพื่อคาดการณ์หรือทำนายรูปแบบของการจำแนกให้อยู่ในหมวดหมู่หรือกลุ่มที่กำหนด เช่น การแบ่งประเภทของความสนใจในผลิตภัณฑ์สินค้า โดยพิจารณาจากจำนวนการซื้อความถี่ของการซื้อสินค้า หรือการจัดหมวดหมู่ของสัตว์ โดยพิจารณาจากจำนวนขา และปีก เป็นต้น ดังแสดงใน
ตารางที่ 2.1

ตารางที่ 2.1 ตารางชนิดของสัตว์

สัตว์	จำนวนขา	มีปีก
วัว	4	0
สุนัข	4	0
ยุง	6	1
แมลงวัน	6	1

จากตารางจะเห็นว่าสามารถจัดหมวดหมู่สัตว์ออกเป็น 2 ชนิด คือแมลงกับสัตว์เลี้ยงลูกด้วยนม ด้วยลักษณะของขาที่ต่างกัน และแมลงจะมีปีก ดังแสดงในตารางที่ 2.2

ตารางที่ 2.2 การจัดหมวดหมู่ของสัตว์

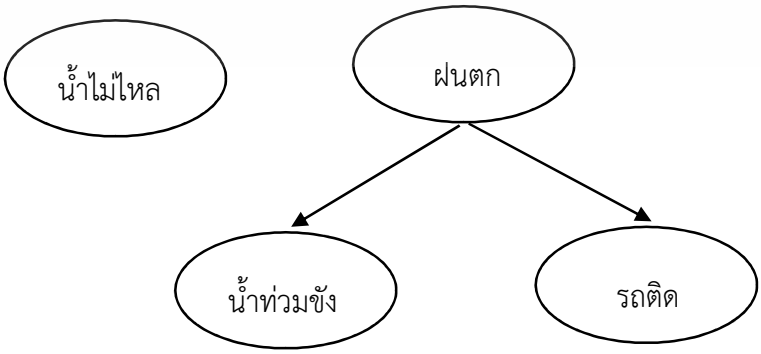
สัตว์	จำนวนขา	มีปีก	ประเภท
วัว	4	0	สัตว์เลี้ยงลูกด้วยนม
สุนัข	4	0	สัตว์เลี้ยงลูกด้วยนม
ยุง	6	1	แมลง
แมลงวัน	6	1	แมลง

2.4 ทฤษฎีโครงข่ายแบบเบย์ (Bayesian Network)

โครงข่ายแบบเบย์ [13] คือการเรียนรู้ที่ใช้หลักการของความน่าจะเป็น ซึ่งมีพื้นฐานมาจากทฤษฎีของเบย์ (Bayes Theorem) ซึ่งมีเงื่อนไขการทำงานดังนี้

1) เป็นแบบจำลองเครือข่ายแบบไม่วนกลับไปหาโหนด (Node) เดิม โดยแต่ละโหนดจะมีความสัมพันธ์ตามทิศทางที่แบบจำลองนำเสนอ
2) คุณสมบัติของแต่ละโหนดในโครงข่ายแบบเบย์ เป็นอิสระจากกันแบบมีเงื่อนไข ทำให้โหนดมีความสัมพันธ์กันแบบไม่ขึ้นต่อกันอย่างมีเงื่อนไข
3) โหนดทั้งหมดในแต่ละโหนดของโครงข่ายแบบเบย์ จะแทนด้วยตัวแปรที่เกี่ยวข้องกับเหตุการณ์หรือข้อมูลที่สนใจ
4) การเชื่อมต่อระหว่างโหนดด้วยสัญลักษณ์ลูกศร ถ้าลูกศรจากโหนด X ชี้ไปหาโหนด Y แสดงว่าโหนด X เป็นโหนดพ่อแม่ (Parents) ของโหนด Y
5) แต่ละโหนด X_i มีความน่าจะเป็นแบบมีเงื่อนไขเป็น $P(X_i|parents(X_i))$ ซึ่งส่งผลกับโหนดพ่อแม่ (Parents Node) ของแต่ละโหนด

ความไม่ขึ้นต่อกันอย่างมีเงื่อนไขของโครงข่ายแบบเบย์ ซึ่งในแต่ละโหนดนั้นมีความสัมพันธ์กัน แต่ก็อาจมีบางโหนดที่ไม่มีความสัมพันธ์ใดๆกับโหนดที่มีเลยดังรูปที่ 2.2



รูปที่ 2.2 เหตุการณ์ที่เกิดขึ้นเมื่อฝนตก

จากเหตุการณ์ฝนตกรูปที่ 2.2 แสดงให้เห็นว่าเมื่อฝนตกแล้วจะมีเหตุการณ์ที่ตามมาคือ รถติดและน้ำท่วมขัง ซึ่งในเหตุการณ์นี้โหนดพ่อแม่ คือ เหตุการณ์ฝนตก ซึ่งมีความสัมพันธ์กับอีกสองโหนดที่ตามมาคือเหตุการณ์น้ำท่วมขังและเหตุการณ์รถติดเสมอ แต่จะมีเหตุการณ์หนึ่ง ซึ่งไม่มีความสัมพันธ์กับเหตุการณ์ฝนตกเลย คือ เหตุการณ์น้ำไม่ไหล ซึ่งเรียกว่าความไม่ขึ้นต่อกันแบบมีเงื่อนไข อธิบายได้ดังดังสมการที่ (2.1)

$$P(X\,|Y,Z) = P(X|Z) \tag{2.1}$$

จากสมการแสดงถึงการไม่ขึ้นต่อกันอย่างมีเงื่อนไข คือ ค่าของ X ไม่จำเป็นต้องขึ้นตรงกับค่าของ Y เมื่อรู้ค่าของ Z แล้ว ดังนั้นในสมการจึงตัดค่าของ Y ออก เหลือเป็น P(X|Z) ความไม่ขึ้นต่อกันอย่างมีเงื่อนไข ทำให้หาความน่าจะเป็นของตัวแปรที่ต้องการได้ง่ายขึ้น เนื่องจากไม่จำเป็นต้องสนใจตัวแปรอื่นในโครงข่ายแบบเบย์ ตัวแปรแต่ละตัวจะมีความน่าจะเป็นเฉพาะ ที่อาจจะเป็นความน่าจะเป็นของโหนดเริ่มต้น ความน่าจะเป็นที่ได้จากความสัมพันธ์ที่มากกว่าหนึ่งโหนด หรือความน่าจะเป็นที่มาจากตัวแปรที่มากกว่าตัวแปรหนึ่งตัวเรียกว่า ความน่าจะเป็นร่วม (Joint Probability) สมการของความน่าจะเป็นร่วม แสดงได้ดังสมการที่ (2.2)

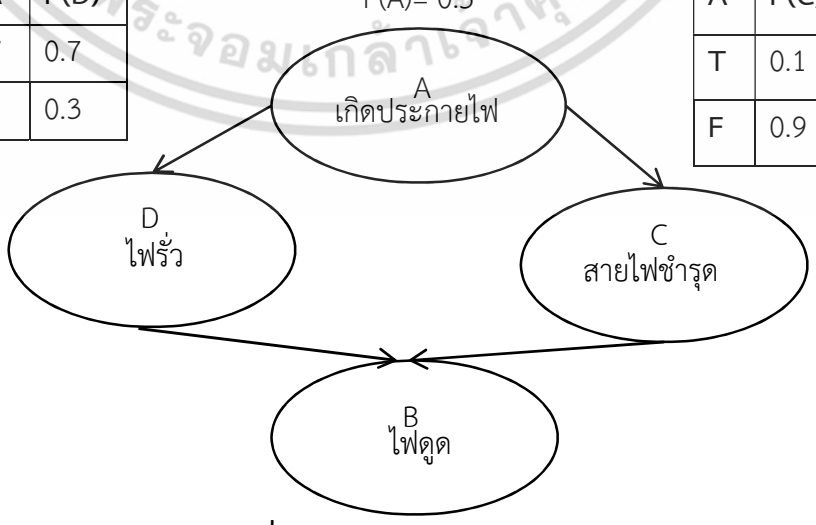
$$P(X_1, \ldots X_n) = \prod_{i=1}^{n} P(X_i|Parents(X_i)) \tag{2.2}$$

โดยที่ $Parents(X_i)$ หมายถึงโหนดพ่อแม่ของโหนดที่ X_i

รูปที่ 2.3 แสดงความสัมพันธ์ของเหตุการณ์ไฟดูด ในรูปแบบของโครงข่ายแบบเบย์

A	P(D)
T	0.7
F	0.3

D	C	P(B)
T	T	0.99
T	F	0.90
F	T	0.95
F	F	0.01



A	P(C)
T	0.1
F	0.9

รูปที่ 2.3 ความสัมพันธ์ของเหตุการณ์ไฟดูด

จากรูปที่ 2.3 แสดงโครงข่ายแบบเบย์ ซึ่งประมาณค่าความน่าจะเป็นของเหตุการณ์ ไฟดูด (B) จากโครงข่ายข้างต้นแล้วนำมาคำนวณความน่าจะเป็นในแต่ละเงื่อนไขเป็นตัวเลข หรือที่เรียกว่า “ตารางความน่าจะเป็นแบบมีเงื่อนไข” (Conditional Probability Table) ซึ่งใช้แสดงความสัมพันธ์ที่ขึ้นต่อกันในแต่ละตัวแปรที่เกี่ยวข้องกัน เช่น โอกาสที่จะเกิดเหตุการณ์ไฟดูด (B) หากเกิดไฟรั่ว (D=T) และสายไฟชำรุด (C=T) มีค่า 0.99 เป็นต้น

จากโครงข่ายแบบเบย์ดังรูปที่ 2.3 ที่แสดงเหตุการณ์ไฟดูด เราสามารถหาความน่าจะเป็นของการเกิดเหตุการณ์ไฟดูดเพราะไฟรั่ว ซึ่งสายไฟไม่ได้ชำรุด และไม่ได้เกิดประกายไฟขึ้นได้ดังนี้

$$P (B , \sim C , D , \sim A) = P (B \mid \sim C , D) P(\sim C|\sim A) P(D|\sim A) P(\sim A)$$

$$= 0.90 * 0.9 * 0.3 * 0.5$$

$$= 0.1215$$

ดังนั้นความน่าจะเป็นที่จะเกิดเหตุการณ์ไฟดูดเพราะไฟรั่ว โดยที่สายไฟไม่ได้ชำรุดและไม่ได้เกิดประกายไฟขึ้น คือ 0.1215 หรือ 12.15 เปอร์เซ็นต์

2.5 เทคนิคการปรับข้อมูลให้มีความสมดุล (SMOTE)

อัลกอริทึมเอสเอ็มโอทีอี (SMOTE Algorithm) [10] เป็นเทคนิคที่ใช้สำหรับการแก้ปัญหาความไม่สมดุลของข้อมูล โดยการปรับเพิ่มข้อมูลของคลาสที่มีข้อมูลน้อยกว่า ซึ่งเรียกว่า “ไมนอริทีคลาส” (Minority Class) ซึ่งคือการปรับความสมดุลข้อมูลของคลาสที่มีจำนวนข้อมูลน้อยกว่า ให้ข้อมูลมีความสมดุลกันกับคลาสที่มีจำนวนข้อมูลที่มากกว่า ซึ่งเรียกว่า “มาจอริทีคลาส” (Majority Class) เพื่อให้การทำนายผลมีประสิทธิภาพสูงขึ้น ในอัลกอริทึมเอสเอ็มโอทีอีนั้น การปรับเพิ่มไมนอริทีคลาส จะปรับเพิ่มได้ครั้งละ 100 เปอร์เซ็นต์ ดังอัลกอริทึมที่แสดงในรูปที่ 2.4

Algorithm : SMOTE (T, N, K)

Input: Number of minority class samples T; Amount of SMOTE $N\%$;

Number of nearest neighbors k

Output: $\left(\frac{N}{100}\right) * T$ synthetic minority class samples

1. (* If N is less than 100%, randomize the minority class samples as only a random percent of them will be SMOTEd. *)
2. if $N < 100$ then Randomize the T minority class samples
3. then Randomize the T minority class samples
4. $T = \left(\frac{N}{100}\right) * T$
5. $N = 100$
6. endif
7. N = (int)(N/100)(* The amount of SMOTE is assumed to be in integral multiples of 100.)
8. k = Number of nearest neighbors
9. numattrs = Number of attributes
10. Sample[ ][ ]: array for original minority class samples
11. Synthetic[ ][ ]: array for synthetic samples (* Compute k nearest neighbors for each minority class sample only. *)
12. for $i = 1$ **to** T
13. Compute k nearest neighbors for i and save the indices in the $nnarray$
14. Populate $(N, i, nnarray)$
15. endfor

Populate(N, i, nnarray) (* Function to generate the synthetic samples. *)

16. while $N \neq 0$
17. Choose a random number between 1 and k, call it nn. This step chooses one of the k nearest neighbors of i.
18. for $attr = 1$ **to** $numattrs$
19. Compute: $dif = Sample[nnarray[nn]][attr] - Sample[i][attr]$
20. Compute: $gap =$ random number between 0 and 1
21. $Synthetic[newindex][attr] = Sample[i][attr] + \text{gap} * \text{dif}$
22. endfor
23. $newindex++$
24. N = N – 1
25. endwhile
26. Return (* End of Populate *)

รูปที่ 2.4 รหัสเทียมการทำงานของอัลกอริทึมเอสเอ็มโอทีอี [10]

แนวคิดหลักในการทำงานของอัลกอริทึมเอสเอ็มโอทีอี คือสร้างข้อมูลชุดใหม่ระหว่างข้อมูลที่อยู่ในไมนอริทีคลาสกับเพื่อนบ้านใกล้เคียง กำหนดให้ x เป็นข้อมูลในไมนอริทีคลาส และ x_1 เป็นเพื่อนบ้านใกล้เคียงตัวหนึ่งในเพื่อนบ้านใกล้เคียง k ตัว ข้อมูลใหม่สามารถคำนวณได้จากสมการที่ (2.3)

$$y = x + rand(0,1) \times (x_1 - x) \tag{2.3}$$

โดยที่ $rand\ (0,1)$ คือ ค่าสุ่มที่อยู่ระหว่าง 0 กับ 1

y คือ ข้อมูลตัวใหม่

ตัวอย่างเช่น กำหนดให้ข้อมูลในไมนอริทีคลาส คือ $(8,6)$ และ $(5,4)$ เป็นเพื่อนบ้านใกล้เคียงตัวหนึ่งในเพื่อนบ้านใกล้เคียง k ตัว สมมุติให้ (x, y) เป็นข้อมูลใหม่ที่ต้องการเพิ่ม เราสามารถคำนวณข้อมูลใหม่ได้ดังนี้

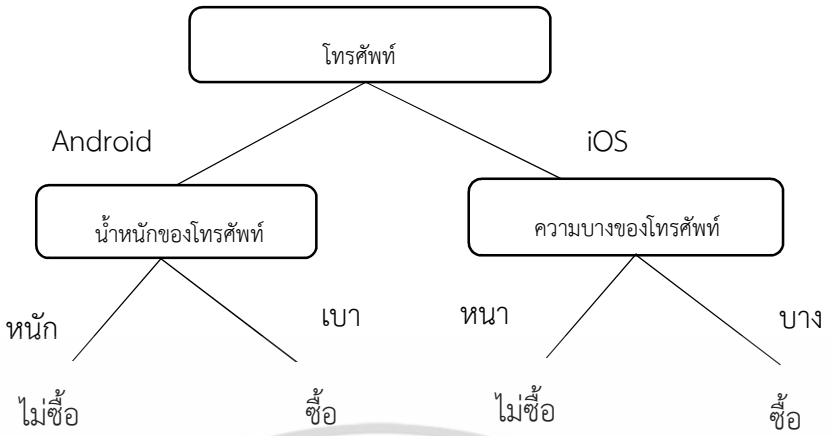
$$\begin{aligned}(x,y) &= (8,6) + rand(0,1) \times \big((5,4) - (8,6)\big) \\ &= (8,6) + rand(0,1) \times (-3,-2)\end{aligned}$$

สมมุติว่าสุ่มได้ค่า $rand(0,1) = 0.2$

ดังนั้น
$$\begin{aligned}(x,y) &= (8,6) + (0.2) \times (-3,-2) \\ &= (8,6) + (-0.6,-0.4) \\ &= (7.4,5.6)\end{aligned}$$

2.6 การเรียนรู้แบบต้นไม้ตัดสินใจ (Decision Tree)

การเรียนรู้แบบต้นไม้ตัดสินใจ [13] เป็นเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) แบบมีผู้สอน (Supervised Learning) โครงสร้างของต้นไม้ตัดสินใจประกอบด้วย ราก (Root) ใบ (Leaf) และกิ่งก้าน (Branch) แตกแขนงออกไปตามเงื่อนไข ซึ่งหลักการทำงานไม่มีสิ่งใดซับซ้อนมากนัก เป็นการแตกแขนงจากรากไปสู่ใบทีละกิ่งก้านตามค่าของคุณลักษณะของโหนดที่ใช้ในต้นไม้ตัดสินใจ รูปที่ 2.5 แสดงตัวอย่างเหตุการณ์การตัดสินใจเลือกซื้อโทรศัพท์



รูปที่ 2.5 เหตุการณ์การตัดสินใจเลือกซื้อโทรศัพท์ในรูปแบบต้นไม้ตัดสินใจ

ใบ คือข้อมูลที่สนใจ หรือสิ่งที่คาดการณ์ไว้ว่ามีโอกาสที่จะเกิดขึ้นตามเหตุการณ์แวดล้อม โดยแต่ละใบ จะถูกเชื่อมด้วยกิ่งก้าน ซึ่งเป็นข้อมูลที่แตกแขนงออกมาจากโหนดต่างๆ ดังตัวอย่างเช่น เหตุการณ์ของการตัดสินใจซื้อโทรศัพท์ เป็นต้น

จากตัวอย่างรูปที่ 2.5 แสดงเหตุการณ์ของการตัดสินใจซื้อโทรศัพท์โดยมีให้เลือก 2 ระบบปฏิบัติการระหว่างระบบปฏิบัติการแอนดรอยด์ (Android) และไอโอเอส (iOS) ถ้าเลือกระบบปฏิบัติการแอนดรอยด์ จะตัดสินใจซื้อโทรศัพท์ โดยดูจากโหนดน้ำหนักของโทรศัพท์ หากโทรศัพท์มีน้ำหนักที่หนักก็จะตัดสินใจไม่ซื้อ แต่หากโทรศัพท์มีน้ำหนักเบาก็จะเกิดการตัดสินใจซื้อเกิดขึ้น เช่นเดียวกันหากลูกค้าเลือกซื้อโทรศัพท์ระบบปฏิบัติการไอโอเอส โดยพิจารณาจากความบางของโทรศัพท์ หากโทรศัพท์มีความหนาก็จะตัดสินใจไม่ซื้อ แต่หากโทรศัพท์มีความบางก็จะตัดสินใจซื้อ เป็นต้น

2.6.1 ต้นไม้ตัดสินใจแบบไอดีทรี

ต้นไม้ตัดสินใจแบบไอดีทรี (ID3) [13] ใช้วิธีการวัดความเหมาะสม ของแต่ละคุณลักษณะ โดยใช้ค่าเกนสารสนเทศ (Information Gain) ซึ่งต้องเริ่มพิจารณาจากโหนดราก (Root Node) เป็นอันดับแรก ก่อนจะดำเนินการพิจารณาใบและกิ่งก้านที่แตกแขนงออกไป ซึ่งต้องคำนวณหาค่าความเหมาะสม โดยจำเป็นต้องหาค่าเอนโทรปี (Entropy) ซึ่งเป็นส่วนประกอบของค่าเกนสารสนเทศให้ได้ก่อน ดังสมการที่ (2.4)

$$Info(D) = -\sum_{i=1}^{m} p_i log_2(p_i) \quad (2.4)$$

โดยที่ p_i คือ ค่าความน่าจะเป็นของข้อมูล D ที่อยู่ในกลุ่ม i

m คือ จำนวนกลุ่มทั้งหมด

ซึ่งค่า $Info(D)$ คือ ค่าเอนโทรปีของข้อมูล D หลังจากที่คำนวณค่า $Info(D)$ แล้ว จะคำนวณค่า $Info_A(D)$ ของคุณลักษณะ A ดังสมการที่ (2.5)

$$Info_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \quad (2.5)$$

โดยที่ v คือ จำนวนค่าที่เป็นไปได้ทั้งหมดของคุณลักษณะ A

$|D_j|$ คือ จำนวนของข้อมูล D ที่มีค่าที่ j ของคุณลักษณะ A

j คือ ค่าที่ j ของคุณลักษณะ A

หลังจากที่ได้ค่า ค่า $Info(D)$ และ $Info_A(D)$ แล้ว จะหาค่าเกนสารสนเทศ เพื่อใช้พิจารณาเลือกคุณลักษณะที่เหมาะสมที่สุดมาเป็นโหนดของต้นไม้ซึ่งคำนวณได้จากผลต่างของ $Info(D)$ และ $Info_A(D)$เขียนได้ดังสมการที่ (2.6)

$$Gain(A) = Info(D) - Info_A(D) \quad (2.6)$$

ในการสร้างต้นไม้ตัดสินใจจะต้องพิจารณาเลือกคุณลักษณะที่เหมาะสมเพื่อนำมาเป็นโหนดของต้นไม้ตัดสินใจ โดยการคำนวณหาค่าเกนสารสนเทศของแต่ละคุณลักษณะ และเลือกคุณลักษณะที่มีค่าเกนสารสนเทศมากสุดมาเป็นโหนดของต้นไม้ตัดสินใจ

2.6.2 ต้นไม้ตัดสินใจแบบเจโฟร์ทีเอต

ต้นไม้ตัดสินใจแบบเจโฟร์ทีเอต (J48) หรือซีโฟร์พอยต์ไฟว์ (C4.5) คือขั้นตอนการสร้างต้นไม้ตัดสินใจที่พัฒนามาจากไอดีทรี (ID3) ผู้พัฒนาคือ Ross Quinlan ใช้วิธีการหาค่าเกนสารสนเทศ (Information Gain) เหมือนกับวิธีไอดีทรี แต่มีส่วนที่เพิ่มเติมเพื่อแก้ปัญหาความบกพร่องของไอดีทรีเพิ่มเข้ามา ข้อดีของต้นไม้ตัดสินใจแบบเจโฟร์ทีเอต คือสามารถใช้แก้ปัญหาค่าข้อมูลที่ขาดหาย (Missing Value Data) โดยการแทนค่าด้วยเครื่องหมายคำถาม ซึ่งมีสัญลักษณ์คือ “?” เพื่อที่จะไม่นำค่าที่ขาดหายนี้มาคำนวณในกระบวนการตัดสินใจ สามารถใช้กับชุดข้อมูลทดสอบ (Testing Set) ที่มีค่าผิดปกติได้ สามารถลดทอนจำนวนกิ่งของต้นไม้ โดยการตัดกิ่งขณะสร้างต้นไม้ตัดสินใจ โดยไม่ทำให้ค่าความถูกต้องลดลง

ในต้นไม้ตัดสินใจแบบเจโฟร์ทีเอตนั้นใช้ค่าอัตราส่วนเกน (Gain ratio criterion) ซึ่งเป็นค่าที่ใช้ในการตัดสินใจเลือกโหนด เนื่องจากค่าเกนสารสนเทศจะมีค่าความเอนเอียง (Bias) เมื่อข้อมูลนั้นมีค่าที่เป็นไปได้จำนวนมาก ซึ่งการแก้ไขปัญหาความเอนเอียงของค่าเกนสารสนเทศนั้นต้องใช้ค่าสารสนเทศของการแบ่งแยก ซึ่งคำนวณได้จากสมการที่ (2.7)

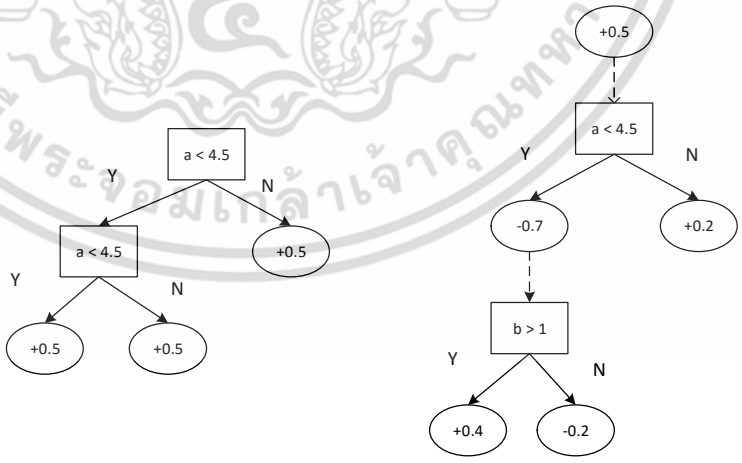
$$Splitinfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log\left(\frac{|\mathrm{D_j}|}{|\mathrm{D}|}\right) \quad (2.7)$$

ค่าสารสนเทศของการแบ่งแยกที่ได้นี้ ถูกนำไปหารค่าเกนสารสนเทศเพื่อหาค่าอัตราส่วนเกน ดังสมการที่ (2.8)

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)} \quad (2.8)$$

2.6.3 ต้นไม้ตัดสินใจแบบเอดีทรี

ต้นไม้ตัดสินใจแบบเอดีทรี (ADTree) [14][15][21] คือต้นไม้ตัดสินใจที่มีการทำงานแบบสองเงื่อนไข ประกอบด้วยโหนดราก โหนดตัดสินใจ (Decision node) และโหนดทำนาย (Prediction node) แสดงตัวอย่างการทำงานของต้นไม้ตัดสินใจเอดีทรีแสดงดังรูปที่ 2.6



รูปที่ 2.6 ตัวอย่างการทำงานของต้นไม้ตัดสินใจเอดีทรี [21]

จากรูปที่ 2.6 เป็นต้นไม้ตัดสินใจแบบสองเงื่อนไข ซึ่งในโหนดทำนายจะมีเฉพาะตัวเลขที่เป็นจำนวนจริงเท่านั้น การคำนวณผลลัพธ์ได้จากการหาผลรวมของค่าในโหนดทำนายบนเส้นทางจาก

โหนดรากถึงโหนดใบแล้วนำไปผ่านฟังก์ชันซายน์ (Sign function) ตัวอย่างการจำแนกคลาสของข้อมูลชุดหนึ่ง เมื่อกำหนดให้ a=0.5 และ b=0.5 คือ sign(0.5-0.7-0.2) = sign(-0.4) = -1 โดยถ้าค่าผลรวมมากกว่าหรือเท่ากับศูนย์ ฟังก์ชันซายน์จะให้ผลลัพธ์เป็นคลาส +1 แต่ถ้าผลรวมน้อยกว่า 0 จะให้ผลลัพธ์เป็นคลาส -1

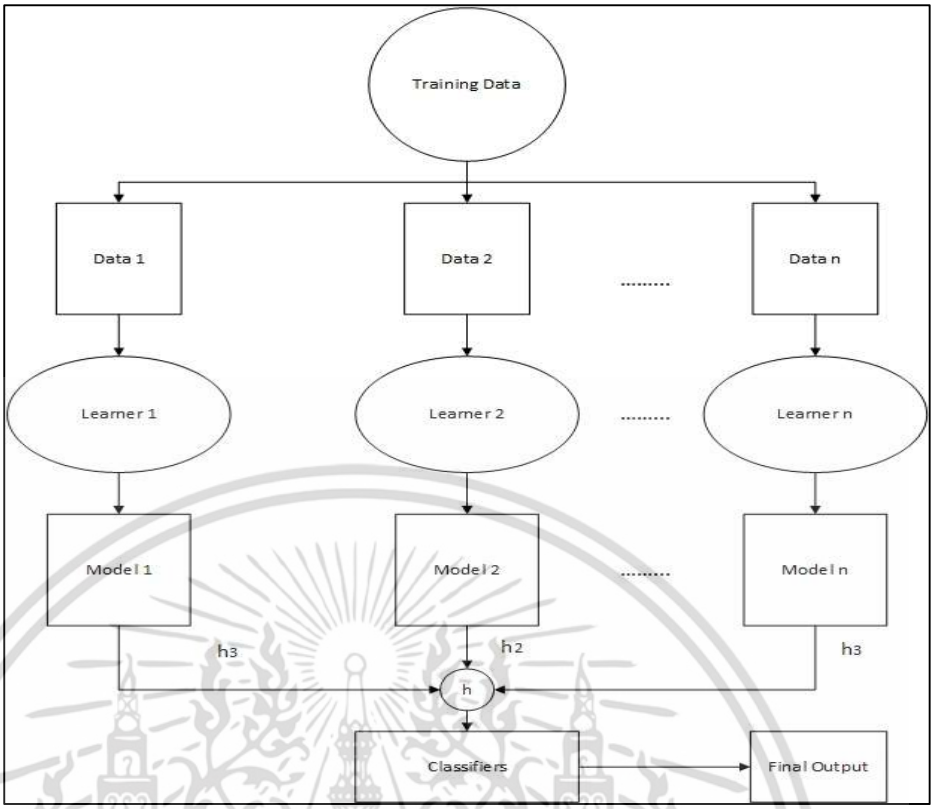
2.6.4 ต้นไม้ตัดสินใจแบบอาร์อีพีทรี

ต้นไม้ตัดสินใจอาร์อีพีทรี (REPTree) [16] คือการเรียนรู้แบบต้นไม้ตัดสินใจที่มีการทำงานที่รวดเร็ว โดยใช้วิธีการลดข้อผิดพลาดด้วยการตัดแต่งกิ่งของต้นไม้ (Reduced-Error Pruning) มีการเรียงลำดับค่าเฉพาะ และจัดการค่าที่ขาดหายไปด้วยการแยกกิ่งของต้นไม้ หากมีการแยกกิ่งของต้นไม้ออกมาซ้ำกันจำนวนมาก จะใช้วิธีรวมกิ่งและวัดเป็นค่าเฉลี่ยแทน ในการเรียนรู้แบบอาร์อีพีที จะใช้ชุดข้อมูลการตัดกิ่ง (Prunning Set) แยกออกจากชุดข้อมูลฝึกสอน เพื่อจะวัดความถูกต้องของส่วนโหนดและใบในการสร้างเป็นต้นไม้ตัดสินใจต่อไป

2.7 กระบวนการเรียนรู้แบบรวมกลุ่ม (Ensemble Method)

การเรียนรู้แบบรวมกลุ่ม (Ensemble Learning) [13] คือ เทคนิคการเรียนรู้แบบเครื่องรูปแบบหนึ่ง ซึ่งใช้วิธีการรวมเอาแบบจำลองการเรียนรู้หลายๆแบบจำลองเข้าไว้ด้วยกัน โดยแต่ละแบบจำลองจะเป็นตัวจำแนกพื้นฐานให้กับการเรียนรู้แบบรวมกลุ่ม ซึ่งการเรียนรู้แบบรวมกลุ่มเหมาะสมที่จะนำไปใช้กับชุดข้อมูลที่มีความซับซ้อน เพื่อเพิ่มประสิทธิภาพของผลลัพธ์ในการจำแนกที่ดี

การเรียนรู้แบบรวมกลุ่มแบ่งออกเป็น 2 ประเภท คือแบบแบ็กกิง (Bagging) และแบบบูสต์ติง (Boosting) การเรียนรู้แบบรวมกลุ่มแบบแบ็กกิง คือ การสร้างแบบจำลองด้วยชุดข้อมูลเรียนรู้ที่แตกต่างกัน ซึ่งช่วยในการปรับปรุงประสิทธิภาพในการจำแนกและการประมาณค่าได้ คำตอบของวิธีแบ็กกิงใช้วิธีการนับคะแนนผลโหวตที่ได้จากแบบจำลองการจำแนกพื้นฐานทั้งหมด และกำหนดกลุ่มจากค่าคะแนนสูงสุดที่ได้จากการโหวตให้กับชุดข้อมูลใหม่ การเรียนรู้แบบรวมกลุ่มแบบบูสต์ติง คือ การใช้แบบจำลองหลาย ๆ แบบจำลองในการตัดสินใจ โดยใช้การกำหนดน้ำหนักให้แต่ละแบบจำลอง ซึ่งค่าน้ำหนักจะได้จากความแม่นยำในการเรียนรู้ คำตอบสุดท้ายของวิธีแบบบูสต์ติง จะใช้การโหวตด้วยค่าน้ำหนักของตัวจำแนกพื้นฐาน รูปที่ 2.7 แสดงแบบจำลองการเรียนรู้แบบรวมกลุ่ม



รูปที่ 2.7 แผนภาพการเรียนรู้แบบรวมกลุ่ม

2.7.1 AdaBoost.M1 Algorithm

อัลกอริทึมเอดาบูสต์เอ็มวัน (AdaBoost.M1 Algorithm) [20] คือเทคนิคการเรียนรู้แบบรวมกลุ่มแบบบูสต์ติง ซึ่งใช้การกำหนดน้ำหนักให้กับแบบจำลองพื้นฐานสำหรับการโหวตหาค่าผลลัพธ์สุดท้าย กำหนดตัวแปรของอัลกอริทึมเอดาบูสต์เอ็มวันมีดังต่อไปนี้

1) นำเข้าข้อมูลชุดฝึกสอน m ชุด ชุดตัวอย่างฝึกสอนคือ
$S = \langle (x_1, y_1), \ldots, (x_m, y_m) \rangle$
2) กำหนดให้ T คือ จำนวนรอบของการสร้างตัวจำแนกพื้นฐาน
3) กำหนดให้ i คือ ลำดับของข้อมูล
4) กำหนดให้ t คือ ลำดับของการจำแนกข้อมูล
5) กำหนดให้ W คือ แบบจำลองการเรียนรู้พื้นฐาน (WeakLearn)

ขั้นตอนการทำงานของอัลกอริทึมเอดาบูสต์เอ็มวัน สามารถอธิบายได้ดังอัลกอริทึมในรูปที่ 2.8

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 2.8 อัลกอริทึมแสดงการทำงานของเอดาบูสต์เอ็มวัน [20]

กำหนดข้อมูลนำเข้า คือข้อมูลที่จะเป็นชุดฝึกสอน $S = \langle (x_1, y_1), \ldots, (x_m, y_m) \rangle$ เมื่อ m คือจำนวนข้อมูลในชุดฝึกสอน จำนวนการวนรอบการเรียนรู้ทั้งหมดจำนวน T ครั้ง ในแต่ละรอบการเรียนรู้ กำหนดด้วยตัวแปร t อัลกอริทึมจะทำการกำหนดค่าแจกแจง D_i ให้กับข้อมูลแต่ละตัว โดยในครั้งแรกของการวนรอบมีค่าเท่า $\frac{1}{m}$ โดยอัลกอริทึมจะส่งข้อมูลชุดฝึกสอน S และค่าการแจกแจง D_t เข้าสู่แบบจำลองการเรียนรู้ W ซึ่งจะได้ผลลัพธ์เป็นสมมติฐาน (Hypotesis) h_t ของแต่ละรอบการเรียนรู้ จากนั้นนำมาคำนวณหาค่าความผิดพลาด (Error rate) ε_t ซึ่งคำนวณได้จากการนำค่าการแจกแจง D_t คูณกับฟังก์ชันอินดิเคเตอร์ (Indicator function) คือเมื่อผลลัพธ์ h_t มีค่าที่ไม่ถูกต้องจะได้ค่า 1 หากถูกต้องจะมีค่าเป็น 0 หรือกล่าวให้เข้าใจอย่างได้ง่าย ๆ คือการหาผลรวมของค่า D_t ตัวที่ i แต่ละตัวมีสมมติฐานในการจำแนกประเภทที่ไม่ถูกต้อง ตรวจสอบค่าความผิดพลาดว่ามีค่าเกินกว่า 0.5 หรือไม่ หากค่าความผิดพลาดเกิน 0.5 จะยกเลิกการเรียนรู้ นำค่าความผิดพลาดที่ได้มาคำนวณหาค่าปัจจัยน้ำหนัก β_t โดยมีค่าเท่ากับ $\varepsilon_t / 1 - \varepsilon_t$ และค่าปัจจัยน้ำหนักนี้จะใช้สำหรับคำนวณหาค่าน้ำหนักของตัวจำแนกประเภทพื้นฐานและปรับปรุงค่าการแจกแจงของข้อมูลใหม่ในแต่ละรอบการ

เรียนรู้ ซึ่งคำนวณได้จากเงื่อนไขสมมติฐานของชุดข้อมูลที่ i มีความถูกต้องให้ทำการคูณ D_t ตัวที่ i กับค่า β_t และหากสมมติฐานของข้อมูลที่ i ไม่ถูกต้อง จะไม่ทำการปรับค่า D_t ตัวที่ i ดังกล่าว เมื่อทำการปรับค่าการกระจายแล้ว ให้ทำการลดทอนค่าการแจกแจงหรือการนอร์มัลไลซ์ (Normalize) ของข้อมูล ให้มีค่าอยู่ระหว่างค่า 0 ถึง 1

2.7.2 AdaBoost.M2 Algorithm

อัลกอริทึมเอดาบูสต์เอ็มทู (AdaBoost.M2 Algorithm) [20] คือเทคนิคการเรียนรู้แบบกลุ่มแบบบูสต์ติง ซึ่งคล้ายกับการเรียนรู้ของเอดาบูสต์เอ็มวัน แต่จะมีส่วนที่แตกต่างจากการเรียนรู้ของเอดาบูสต์เอ็มวัน อยู่ตรงที่เอดาบูสต์เอ็มวันนั้นจะตรวจสอบความผิดพลาด ซึ่งหากมีค่าความผิดพลาดเกิน 0.5 จะยกเลิกการเรียนรู้ทันที แต่ในเอดาบูสต์เอ็มทูนั้น ไม่มีการกำหนดว่าหากค่าความผิดพลาดเกิน 0.5 จะหยุดการเรียนรู้ แต่จะใช้การทดแทนค่าความผิดพลาดด้วยการคำนวณซูโดลอส (Pseudo-loss) เพื่อปรับค่าน้ำหนักใหม่ให้ได้สมมติฐานที่ดีที่สุดซึ่งซูโดลอสสามารถคำนวณได้จากสมการที่ (2.9)

$$h_t: \epsilon_t = \frac{1}{2} \sum_{(i,y)\epsilon B} D_t(i,y)\left(1 - h_t(x_i, y_i) + h_t(x_i, y)\right) \qquad (2.9)$$

โดยที่ D_t คือ ค่าการแจกแจงของข้อมูลตัวที่ t

h_t คือ สมมติฐานของการเรียนรู้ในรอบที่ t

ขั้นตอนการเรียนรู้ของเอดาบูสต์เอ็มทู สามารถเขียนเป็นอัลกอริทึมได้ดังรูปที่ 2.9

Algorithm : AdaBoost.M2

Input: sequence of m examples $\langle (x_1, y_1), \dots, (x_m, y_m) \rangle$ with labels $y_i \epsilon\, Y = \{1, \dots, k\}$, weak leaning algorithm WeakLearn , T is Number of iteration

Let $B = \{(i, y) : i \epsilon \{1, \dots, m\}, y \neq y_i\}$

Initialize $D_1(i, y) = \frac{1}{|B|} for\ (i, y) \epsilon B.$

Do for $t = 1,2, \dots T$

1. Call Weaklearn, providing it with mislabel distribution D_t
2. Get back a hypothesis $h_t : X \times Y \rightarrow [0,1].$
3. Calculate the pseudo-loss of

$$h_t : \epsilon_t = \frac{1}{2} \sum_{(i,y) \epsilon B} D_t(i, y)\big(1 - h_t(x_i, y_i) + h_t(x_i, y)\big)$$

4. Set $\beta_t = \epsilon_t / (1 - \epsilon_t)$
5. Update $D_t : D_{t+1}(i, y) = \frac{D_t(i,y)}{Z_t} . \beta_t^{(1/2)(1 + h_t(x_i, y_i) - h_t(x_i, y))}$ where Z_t is a normalization constant (chosen so that D_{t+1} will be distribution)
6. Output the hypothesis:

$$h_{fin}{}^{x} = \underset{y \epsilon Y}{\arg max} \sum_{t=1}^{T} \left(log \frac{1}{\beta_t} \right) h_t(x, y)$$

รูปที่ 2.9 อัลกอริทึมแสดงการทำงานของเอดาบูสต์เอ็มทู [20]

2.8 ตัววัดประสิทธิภาพของแบบจำลอง

ในงานวิจัยนี้เป็นการวัดประสิทธิภาพของข้อมูลที่มีความแตกต่างกันระหว่าง 2 คลาส คือ คลาสบวก และคลาสลบ ซึ่งใช้การวัดประสิทธิภาพการจำแนกประเภททั้งหมด 6 แบบ คือค่าความแม่นยำของคลาสบวก (Sensitivity) ค่าความแม่นยำของคลาสลบ (Specificity) ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าเอฟเมเซอร์ (F-measure)

กำหนดให้ TP คือจำนวนข้อมูลที่จำแนกถูกต้องของคลาสบวก

TN คือจำนวนข้อมูลที่จำแนกถูกต้องของคลาสลบ

FN คือจำนวนข้อมูลที่จำแนกไม่ถูกต้องของคลาสลบ

FP คือจำนวนข้อมูลที่จำแนกไม่ถูกต้องของคลาสบวก

P คือจำนวนข้อมูลที่เป็นคลาสบวก

N คือจำนวนข้อมูลที่เป็นคลาสลบ

1) ค่าความแม่นยำของคลาสบวก คือค่าความถูกต้องของคลาสที่เป็นบวกดังสมการที่ (2.10)

$$\mathrm{TPR} = \mathrm{TP} / (\mathrm{TP} + \mathrm{FN}) \tag{2.10}$$

2) ค่าความแม่นยำของคลาสลบ คือค่าความถูกต้องของคลาสที่เป็นลบดังสมการที่ (2.11)

$$\mathrm{TNR} = \mathrm{TN} / (\mathrm{FP} + \mathrm{TN}) \tag{2.11}$$

3) ค่าความถูกต้อง คือค่าที่วัดประสิทธิภาพทั้งคลาสบวก และคลาสลบ ออกมาในรูปแบบการวัดผลรวมดังสมการที่ (2.12)

$$\mathrm{ACC} = (\mathrm{TP} + \mathrm{TN}) / (\mathrm{P} + \mathrm{N}) \tag{2.12}$$

ตัวอย่าง กำหนดข้อมูลการจำแนกประเภท โดยให้จำนวนข้อมูลนำเข้าของคลาสลบ มีจำนวน 7,860 จำแนกถูกต้องจำนวน 7,310 จำแนกไม่ถูกต้องจำนวน 550 ข้อมูลนำเข้าของคลาสบวก มีจำนวน 5,018 จำแนกไม่ถูกต้องจำนวน 0 จำแนกถูกต้องจำนวน 5,018 ดังแสดงในตารางที่ 2.3

ตารางที่ 2.3 จำนวนข้อมูลการจำแนกประเภททั้งคลาสลบและคลาสบวกที่แสดงการจำแนกที่ถูกต้องและการจำแนกที่ไม่ถูกต้อง

คลาส	**จำนวนข้อมูล**	
	จำแนกถูกต้อง	**จำแนกไม่ถูกต้อง**
คลาสลบ	7310	550
คลาสบวก	5018	0

ตัวอย่าง การคำนวณค่าความแม่นยำของคลาสบวก

Sensitivity = TP / (TP+FN)

= 5018 / (5018+550)

= 5018/5568

= 0.9012 คิดเป็น 90.1 เปอร์เซ็นต์

ตัวอย่าง การคำนวณค่าความแม่นยำของคลาสลบ

Specificity = TN / (FP+TN)

= 7310 / (0+7310)

= 7310/7310

= 1.000 คิดเป็น 100 เปอร์เซ็นต์

ตัวอย่าง การคำนวณค่าความถูกต้อง

Accuracy = (TP+TN) / (P+N)

= (5018+7310) / (5018+7860)

= 12328 / 12878

= 0.9572 คิดเป็น 95.7 เปอร์เซ็นต์

สำหรับตัววัดประสิทธิภาพของแบบจำลองด้วยค่าความแม่นยำ ค่าความระลึก และค่าเอฟเมเชอร์ กำหนดตัวแปรดังต่อไปนี้

กำหนดให้ tp คือจำนวนข้อมูลที่จำแนกถูกต้องของคลาสบวก

tn คือจำนวนข้อมูลที่จำแนกถูกต้องของคลาสลบ

fn คือจำนวนข้อมูลที่จำแนกไม่ถูกต้องของคลาสลบ

fp คือจำนวนข้อมูลที่จำแนกไม่ถูกต้องของคลาสบวก

1) Precision คือค่าความแม่นยำ ที่บ่งบอกความสามารถในการทำนายคลาสบวก ซึ่งคำนวณได้จากสมการ (2.13)

$$\frac{tp}{tp + fp} \tag{2.13}$$

2) Recall คือค่าความระลึก ที่บ่งบอกความสามารถในการจดจำคลาสบวก คำนวณได้จากสมการ (2.14)

$$\frac{tp}{tp + fn} \tag{2.14}$$

3) F-measure คือค่าการวัดประสิทธิภาพในการจำแนกข้อมูล ซึ่งเกิดจากการนำค่าความแม่นยำและค่าความระลึกมาคำนวณ คำนวณได้จากสมการ (2.15)

$$F - measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{2.15}$$

จากตัวอย่างตารางที่ 2.3 สามารถนำมาคำนวณทั้ง 3 แบบของการวัดประสิทธิภาพการจำแนกประเภท ได้ดังตัวอย่างต่อไปนี้

<u>ตัวอย่าง</u> การคำนวณค่าความแม่นยำ

P $= \frac{tp}{tp+fp}$

$= \frac{5018}{5018+}$

= 1 คิดเป็น 100 เปอร์เซ็นต์

<u>ตัวอย่าง</u> การคำนวณค่าความระลึก

Recall $= \frac{tp}{tp+fn}$

$= \frac{5018}{5018+550}$

= 0.9012 คิดเป็น 90.12 เปอร์เซ็นต์

<u>ตัวอย่าง</u> การคำนวณค่าเอฟเมเซอร์

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

$$= \frac{2 \times 1 \times 0.9012}{1 + 0.9012}$$

= 0.9480 คิดเป็น 94.80 เปอร์เซ็นต์

2.9 งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้องมุ่งเน้นไปที่การศึกษาต้นแบบกระบวนการทำงานและอัลกอริทึมของงานวิจัยนั้น เพื่อนำมาประยุกต์ใช้กับงานวิจัยของตนเอง และเปรียบเทียบประสิทธิภาพการจำแนกกับงานวิจัยต้นแบบ

2.9.1 RUSBoost: A Hybrid Approach to Alleviating Class Imbalance

ในปี 2009 C. Seiffert, FL. Boca Raton, T.M. Khoshgoftaar J. Van Hulse and A. Napolitano [18] ได้พัฒนาอัลกอริทึมที่ช่วยแก้ปัญหาความไม่สมดุลของข้อมูลด้วยการลดทอนจำนวนของมาจอริทีคลาส อัลกอริทึมมีชื่อเรียกว่า “อาร์ยูเอสบูสต์” (RUSBoost) โดยใช้ข้อมูลทดสอบจากเว็บไซต์ยูซีไอ (UCI Website) ทั้งหมด 15 ชุดข้อมูล ดังตารางที่ 2.4

ตารางที่ 2.4 ตัวอย่าง 15 ชุดข้อมูลทดสอบของงานวิจัยอาร์ยูเอสบูสต์

ชุดข้อมูล	จำนวนข้อมูล	จำนวนคอลัมน์
SP3	3541	43
MAMMOGRAPHY	11183	7
SOLARFLAREF	1389	13
CAR3	1728	7
CCCS12	282	9
SP1	3649	43
PC1	1107	16
GLASS3	214	10
CM1	505	16
PENDIGITS5	10992	17
SATIMAGE4	6435	37
ECOLI4	336	8
SEGMENT5	2310	20
CONTRA2	1473	10
VEHICLE1	846	19

การเรียนรู้ของอาร์ยูเอสบูสต์ ใช้วิธีการลดทอนจำนวนของคลาสที่มีมากกว่าหรือลดทอนคลาสที่มีข้อมูลเยอะกว่า เช่นตัวอย่างคลาสลบมี 90 คลาสบวกมี 10 รวมทั้งหมดจำนวน 100 ข้อมูล กำหนดให้ปรับลดทอนด้วยอาร์ยูเอสบูสต์จำนวน 5 เท่า แสดงวิธีคิดคำนวณแบบง่าย กำหนดให้คลาสที่มีมากกว่า แทนด้วยตัวแปร p กำหนดให้คลาสที่มีน้อยกว่าแทนด้วยตัวแปร n กำหนดให้ตัวแปร i เป็นจำนวนของการปรับลดทอนคลาสที่มีมากกว่า คือ $i \leq p/i$

A = (p/i)+(n)

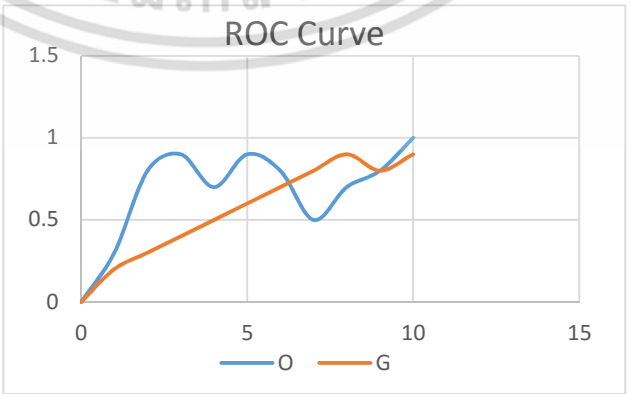
= (90/5)+10

= 28

จากการคำนวณจะได้ค่าจำนวนข้อมูลของคลาสที่มีมากกว่า เปลี่ยนจากจำนวนข้อมูลเดิมที่มี 100 เหลือเพียง 28 ข้อมูล เมื่อเข้าสู่การปรับลดทอนข้อมูลแล้ว ซึ่งในการปรับลดทอนจำนวนข้อมูลในคลาสใดคลาสหนึ่งนั้น การปรับลดทอนจะทำได้ต่อเมื่อคลาสนั้นๆ มีข้อมูลที่เยอะกว่าอีกคลาส ซึ่งจะใช้ไม่ได้สำหรับข้อมูลที่มีจำนวนคลาสเท่ากัน

โดยเปรียบเทียบการวัดประสิทธิภาพทั้งหมด 4 แบบ คือแบบพื้นที่โค้งแบบอาร์โอซี (ROC Curves) แบบพื้นที่โค้งแบบอาร์พีซี (PRC Curves) แบบการวัดด้วยค่าทางสถิติแบบเคเอส (K-S Statistic) และวัดด้วยค่าเอฟเมเซอร์ (F-measure)

1) พื้นที่โค้งแบบอาร์โอซี มีวิธีการวัดประสิทธิภาพด้วยการสร้างกราฟความสัมพันธ์ระหว่างค่าความแม่นยำของคลาสบวก กับค่าความไม่ถูกต้องของคลาสบวก โดยการแปรค่าจุดตัด (Cut point) เพื่อดูค่าความเอนเอียงของกราฟ ตัวอย่างรูปที่ 2.10 กำหนดให้ O และ G คือชุดข้อมูล ในการวัดค่าประสิทธิภาพของข้อมูลที่มีความไม่สมดุล ควรมีค่าความแม่นยำของคลาสบวก และค่าความแม่นยำของคลาสลบที่สูง ซึ่งจะทำให้ค่ากราฟเส้นเอนชิดมุมซ้ายมากที่สุดดังรูปที่ 2.10



รูปที่ 2.10 กราฟพื้นที่โค้งแบบอาร์โอซีเพื่อดูค่าความเอนเอียง

2) พื้นที่โค้งแบบพีอาร์ซี คือกราฟการวัดประสิทธิภาพ โดยดูจากค่าความเอนเอียง มีลักษณะคล้ายกับพื้นที่โค้งแบบอาร์โอซี แต่ใช้ค่าความระลึก (Recall) เป็นตัวชี้วัด
3) ค่าสถิติแบบเคเอส (K-S Statistic) คือการใช้ค่าทางสถิติเป็นตัววัดประสิทธิภาพของแต่ละคลาส โดยลำดับแรกคำนวณหาค่าสัดส่วนระหว่างคลาส เขียนเป็นสมการได้ดังสมการที่ (2.16)

$$F_{c_i}(t) = P(p(x)) \leq t|c_i \tag{2.16}$$

โดยที่ F_{c_i} คือ ค่าสัดส่วนของ Class c_i

t คือ จำนวนข้อมูล

เมื่อได้ค่าสัดส่วนระหว่างคลาสแล้ว ลำดับต่อไปคือการหาค่าทางสถิติเคเอส ดังสมการที่ (2.17)

$$K - S = \max_{t \in [0,1]} |F_{c_1}(t) - F_{c_2}(t)| \tag{2.17}$$

โดยที่ $[0,1]$ คือ จำนวนคลาสของข้อมูล

4) ค่าเอฟเมเชอร์ (F-Measure) คือค่าการวัดประสิทธิภาพในการจำแนกข้อมูล ซึ่งเกิดจากการนำค่าความแม่นยำและค่าความระลึกมาคำนวณโดยพิจารณาจากข้อมูลของผลลัพธ์เดียวกัน ดังสมการที่ (2.18)

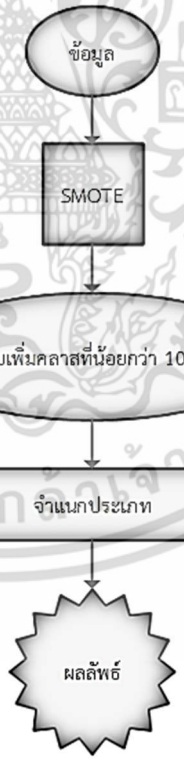
$$F - measure = \frac{2 \times precision \times recall}{precision + recall} \tag{2.18}$$

ในงานวิจัยนี้ทดลองจำแนกประเภททั้งหมด 15 ชุดข้อมูล โดยเปรียบเทียบระหว่างอาร์ยูเอสบูสต์ (RUSBoost) กับเอสเอ็มโอทีอีบูสต์ (SMOTEBoost) เอดาบูสต์ (AdaBoost) อาร์ยูเอส (RUS) เอสเอ็มโอทีอี (SMOTE) และแบบที่ไม่ใช้กระบวนการแปลงข้อมูลใดๆเลย (None) จากการทดลองอาร์ยูเอสบูสต์ ให้ผลลัพธ์ดีที่สุดที่ค่าเฉลี่ยของการวัดประสิทธิภาพทั้ง 15 ชุดข้อมูลคือ 0.8725 รองลงมาคือเอสเอ็มโอทีอีบูสต์ได้ 0.8683 เอดาบูสต์ได้ 0.8492 อาร์ยูเอสได้ 0.8098 เอสเอ็มโอทีอีได้ 0.7780 และแบบที่ไม่ใช้กระบวนการแปลงข้อมูลได้ 0.7049 ค่าที่ได้ทั้งหมดเปรียบเทียบโดยใช้ตัววัดประสิทธิภาพแบบพื้นที่โค้งอาร์โอซี

2.9.2 Bank direct marketing analysis of asymmetric information based on machine learning

ในปี 2015 P. Ruangthong และ S. Jaiyen ได้แก้ปัญหาความไม่สมดุลของข้อมูล (Imbalanced Data) ไว้ด้วยการใช้อัลกอริทึมเอสเอ็มโอทีอี ร่วมกับการจำแนกประเภทโดยใช้การเรียนรู้แบบรวมกลุ่มด้วยอัลกอริทึมชนิดเดียว ซึ่งโรเทชันฟอเรสท์ เจโฟร์ทีเอต (Rotation Forest-J48) ให้ผลลัพธ์ในคลาสบวกสูงที่สุด [19]

ในงานวิจัยนี้ใช้ตัวอย่างชุดข้อมูลทดสอบจากเว็บไซต์ยูซีไอเป็นชุดข้อมูลทางด้านการตลาดทางตรงของธนาคาร ที่มีความไม่สมดุลของข้อมูลอยู่สูง มีจำนวนข้อมูลทั้งหมด 45,211 และมี 17 คุณลักษณะ 2 คลาส ซึ่งคลาสบวก มี 5,289 ข้อมูล คลาสลบ มี 39,922 ข้อมูล ความต่างระหว่าง 2 คลาสอยู่ที่ 34,633 จำนวนข้อมูล ซึ่งหากนำมาใช้ในด้านการจำแนกประเภทเลยนั้น ทำให้ประสิทธิภาพของความแม่นยำต่ำมาก ดังนั้นในงานวิจัยจึงเสนอการปรับเพิ่มข้อมูลด้วยการใช้เอสเอ็มโอทีอี อัลกอริทึม และเปรียบเทียบประสิทธิภาพการจำแนกทั้งหมด 14 ตัวจำแนกพื้นฐาน เพื่อหาประสิทธิภาพของการจำแนกที่ให้ผลลัพธ์ที่ดีที่สุด แสดงกระบวนการทำงานของงานวิจัยดังรูปที่ 2.11



รูปที่ 2.11 กระบวนการทำงานของงานวิจัย

สำหรับการวัดประสิทธิภาพใช้ทั้งหมด 3 แบบ คือการวัดค่าความแม่นยำของคลาสบวก วัดค่าความแม่นยำของคลาสลบ และวัดค่าความถูกต้อง ผลลัพธ์ของการจำแนกเมื่อเปรียบเทียบระหว่างข้อมูลเดิมที่ยังไม่มีการแปลงด้วยเอสเอ็มโอทีอี กับข้อมูลที่มีการแปลงด้วยเอสเอ็มโอทีอีแล้ว ค่าความแม่นยำของคลาสบวกของข้อมูลเดิมนั้นโรเทชันฟอเรสท์ เอดีทรี ให้ค่าสูงสุดที่ 67.73 แต่เมื่อมีการปรับเพิ่มข้อมูลด้วยเอสเอ็มโอทีอีแล้วโครงข่ายแบบเบย์ ให้ค่าความแม่นยำของคลาสบวก สูงสุดที่ 94.48 ส่วนค่าความแม่นยำของคลาสลบ ทรีเจโฟร์ทีเอต ให้ค่าการจำแนกสำหรับข้อมูลเดิมที่ 93.19 และเมื่อมีการปรับด้วยเอสเอ็มโอทีอีแล้วโรเทชันฟอเรสท์ เจโฟร์ทีเอต ให้ค่าสูงสุดที่ 93.17 ค่าความถูกต้องสูงสุดสำหรับข้อมูลเดิม คือโรเทชันฟอเรสท์ เจโฟร์ทีเอต ให้ค่า 90.05 เมื่อปรับด้วยเอสเอ็มโอทีอี ค่าเพิ่มมาที่ 90.81 ดังแสดงในตารางที่ 2.5

ตารางที่ 2.5 ผลการทดลองของแต่ละตัวจำแนกพื้นฐานที่ให้ค่าการจำแนกสูงที่สุด

ตัวจำแนกพื้นฐาน	Sensitivity		Specificity		Accuracy	
	Original Data	SMOTE	Original Data	SMOTE	Original Data	SMOTE
Rotation Forest (PCA) ADTree	67.73	80.12	89.78	87.81	89.12	86.75
BayesNet	50.61	94.48	92.50	89.63	88.18	90.28
J48	57.11	76.63	93.19	92.75	89.51	89.43
Rotation Forest (PCA) J48	63.35	81.23	92.21	93.17	90.05	90.81

บทที่ 3
วิธีดำเนินงานวิจัย

งานวิจัยนี้นำเสนอการแก้ปัญหาความไม่สมดุลของข้อมูลและพัฒนาแบบจำลองของการจำแนกประเภทเพื่อให้ประสิทธิภาพในการจำแนกประเภทของข้อมูลมีผลลัพธ์ที่สูงขึ้น โดยข้อมูลที่ใช้จากเว็บไซต์ยูซีไอเป็นข้อมูลที่ถูกเก็บรวมรวมจากธนาคารโปรตุเกส (Portuguese Banking Institution) ซึ่งเกี่ยวข้องกับการตลาด โดยการทำการตลาดนั้นใช้การโทรศัพท์ติดต่อลูกค้าของธนาคาร โดยมีเงื่อนไขเป็นความต้องการฝากเงินระยะยาวกับธนาคารหรือไม่

3.1 ชุดข้อมูลที่ใช้สำหรับการทดลอง

สำหรับข้อมูลที่ใช้ในการทดลอง ใช้ข้อมูลทั้งหมด 41,188 ข้อมูล ซึ่งถูกเก็บรวบรวมตั้งแต่เดือนเมษายน ปี ค.ศ. 2008 ถึงเดือนพฤศจิกายน ปี ค.ศ. 2010 ข้อมูลประกอบด้วย 20 คุณลักษณะ แบ่งเป็นข้อมูลที่ถูกเก็บรวบรวมจากลูกค้าของธนาคาร ข้อมูลที่เก็บรวบรวมมาจากการจัดรายการส่งเสริมการขายและข้อมูลที่ถูกเก็บรวบรวมจากแหล่งข้อมูลอื่นในช่วงที่มีการจัดกิจกรรม ตารางที่ 3.1 แสดงรายละเอียดคุณลักษณะของข้อมูลดังกล่าว

ตารางที่ 3.1 แสดงรายละเอียดของคุณลักษณะข้อมูลที่เก็บรวบรวมจากลูกค้าของเครือข่ายธนาคาร

ลำดับ	ชื่อของคุณลักษณะ	ชนิดข้อมูล	คำอธิบาย
1	age	numeric	อายุ
2	job	categorical	ชนิดของงาน
3	marital	categorical	สถานะการแต่งงาน
4	education	categorical	การศึกษา
5	default	categorical	มีเครดิตหรือไม่
6	housing	categorical	มีสินเชื่อบ้านหรือไม่
7	loan	categorical	มีสินเชื่อส่วนบุคคลหรือไม่

ข้อมูลของลูกค้าที่ถูกเก็บรวบรวมจากการจัดกิจกรรมส่งเสริมการขายของธนาคารเอง ซึ่งรายละเอียดโดยทั่วไปจะเกี่ยวข้องกับวันที่ติดต่อลูกค้า เดือน และช่วงเวลา ซึ่งรายละเอียดของคุณลักษณะข้อมูลดังกล่าวแสดงอยู่ในตารางที่ 3.2

ตารางที่ 3.2 แสดงรายละเอียดของคุณลักษณะข้อมูลของลูกค้าที่ถูกเก็บรวบรวมจากการเริ่มต้นจัดกิจกรรมส่งเสริมการขายและมีการติดต่อไปหาลูกค้า

ลำดับ	ชื่อของคุณลักษณะ	ชนิดข้อมูล	คำอธิบาย
1	contact	categorical	ชนิดของการติดต่อ เช่น โทรศัพท์บ้าน โทรศัพท์มือถือ
2	month	categorical	เดือนที่ติดต่อล่าสุด
3	day_of_week	categorical	สัปดาห์ที่ติดต่อล่าสุด
4	duration	numeric	ช่วงเวลาที่ติดต่อล่าสุด

ข้อมูลที่เก็บรวบรวมจากส่วนการติดต่อของลูกค้าแต่ละราย เป็นข้อมูลในรูปแบบของจำนวนครั้งในการติดต่อ จำนวนวันที่ขาดการติดต่อกับลูกค้า ซึ่งรายละเอียดของคุณลักษณะข้อมูลดังกล่าวแสดงอยู่ในตารางที่ 3.3

ตารางที่ 3.3 แสดงคุณลักษณะรายละเอียดของข้อมูลที่เก็บรวบรวมจากส่วนการติดต่อของลูกค้าแต่ละราย

ลำดับ	ชื่อของคุณลักษณะ	ชนิดข้อมูล	อธิบาย
1	campaign	numeric	จำนวนครั้งในการติดต่อลูกค้า
2	pdays	numeric	จำนวนวันที่ขาดการติดต่อกับลูกค้า
3	previous	numeric	จำนวนครั้งที่ติดต่อก่อนที่จะมีการจัดกิจกรรม
4	poutcome	categorical	ผลการตอบรับของลูกค้า จากการจัดกิจกรรมในครั้งที่แล้ว

ข้อมูลที่ถูกเก็บรวบรวมจากแหล่งข้อมูลอื่น ที่เกี่ยวข้องทางด้านเศรษฐกิจและสังคมของลูกค้าแต่ละราย โดยมีอัตราเงินเดือน ดัชนีราคาผู้บริโภค ซึ่งเป็นดัชนีราคาที่ใช้วัดการเปลี่ยนแปลงของ

ราคาขายปลีกของสินค้าและบริการที่ผู้บริโภคจ่ายซื้อ เพื่อการบริโภค ณ ตลาดและร้านค้าปลีกในปีใดปีหนึ่ง เปรียบเทียบกับราคาสินค้าชนิดเดียวกันและจำนวนเดียวกันในปีฐาน ดัชนีความเชื่อมั่นต่อผู้บริโภค ซึ่งเป็นการประมาณหรือคาดการณ์ความรู้สึกของผู้บริโภคที่มีต่อเศรษฐกิจโดยรวมและสภาพเศรษฐกิจของตนเอง ซึ่งหากผู้บริโภคมีความมั่นใจในฐานะและรายได้ของตนเอง ย่อมส่งผลถึงการตัดสินใจใช้จ่าย ตารางที่ 3.4 แสดงรายละเอียดของคุณลักษณะข้อมูลดังกล่าว

ตารางที่ 3.4 แสดงรายละเอียดของข้อมูลที่ถูกเก็บรวบรวมจากแหล่งข้อมูลอื่นทางด้านเศรษฐกิจและสังคมในช่วงเวลาของการจัดกิจกรรม

ลำดับ	ชื่อของคุณลักษณะ	ชนิดข้อมูล	อธิบาย
1	emp.var.rate	numeric	เรทอัตราเงินเดือน ณ ช่วงเวลาของการจัดกิจกรรม
2	cons.price.idx	numeric	ดัชนีราคาผู้บริโภค
3	cons.conf.idx	numeric	ดัชนีความเชื่อมั่นต่อผู้บริโภค
4	euribor3m	numeric	เรทอัตราการกู้ยืมระหว่างธนาคาร (euribor) ใน 3 เดือน
5	nr.employed	numeric	จำนวนพนักงาน

3.2 ขั้นตอนการเตรียมข้อมูลสำหรับการทดลอง

เนื่องจากข้อมูลในบางคุณลักษณะเป็นข้อมูลแบบนามบัญญัติ ซึ่งอาจจะยังไม่มีความชัดเจนมากนัก ทำให้การจำแนกประเภทได้ผลไม่เป็นที่น่าพอใจ ซึ่งการแปลงข้อมูลแบบนามบัญญัติเป็นแบบตัวเลขแบบปกตินั้น ยังไม่สามารถให้ประสิทธิภาพการจำแนกเป็นที่น่าพอใจได้ จึงได้ทำการแปลงข้อมูลแบบนามบัญญัติเป็นข้อมูลแบบไบนารีแทน ซึ่งทำให้เพิ่มจำนวนคุณลักษณะของข้อมูลและทำให้ข้อมูลมีความชัดเจนมากขึ้น ส่งผลให้การจำแนกประเภทมีประสิทธิภาพมากขึ้น ดังตารางที่ 3.5 แสดงตัวอย่างสถานะข้อมูลแบบนามบัญญัติที่มีการแปลงเป็นแบบตัวเลข

ตารางที่ 3.5 ข้อมูลแบบตัวเลขของคุณลักษณะของสถานะการสมรส

ข้อมูลแบบนามบัญญัติ	ข้อมูลแบบตัวเลข
สมรส	1
โสด	2
หย่าร้าง	3
ไม่ทราบข้อมูล	4

จากตารางที่ 3.5 ข้อมูลสถานะการสมรส จะเห็นได้ว่าข้อมูลแบบตัวเลขในแต่ละสถานะสมรสนั้นมี 4 สถานะ คือ สมรสแล้ว โสด หย่าร้าง และไม่ทราบข้อมูล แต่ในข้อมูลชุดทดลองจริงมีถึง 20 คุณลักษณะ และในบางคุณลักษณะมีสถานะที่หลากหลายและมีจำนวนที่มาก ดังนั้นการแปลงข้อมูลให้เหลือเพียงแค่ศูนย์กับหนึ่งเป็นแบบไบนารีนั้น จะทำให้ข้อมูลมีแบบแผนและชัดเจน เหมาะสมที่จะนำข้อมูลไปใช้ต่อไป

ตารางที่ 3.6 ตัวอย่างการแปลงข้อมูลแบบนามบัญญัติเป็นข้อมูลแบบไบนารี

ลูกค้า	สมรส	โสด	หย่าร้าง	ไม่ทราบข้อมูล
A	1	0	0	0
B	0	1	0	0
C	0	0	1	0
D	1	0	0	0
E	0	0	0	1

จากตารางที่ 3.6 แสดงตัวอย่างในส่วนของคุณลักษณะที่เพิ่มขึ้นมา ซึ่งข้อมูลเดิมเป็นข้อมูลแบบนามบัญญัติ หากแปลงข้อมูลเป็นแบบไบนารีแล้ว จะเห็นได้ว่าจำนวนของคุณลักษณะมีการเพิ่มขึ้นจากเดิมมีเพียงหนึ่งคุณลักษณะและมีค่าตั้งแต่หนึ่งถึงสี่ แต่พอแปลงข้อมูลเป็นแบบไบนารีนั้นทำให้เพิ่มจำนวนคุณลักษณะเป็นสี่และมีจำนวนข้อมูลเพียงแค่ค่าศูนย์กับหนึ่ง ทำให้ข้อมูลมีความชัดเจนเหมาะสมสำหรับการนำไปใช้ในการจำแนกประเภทต่อไป

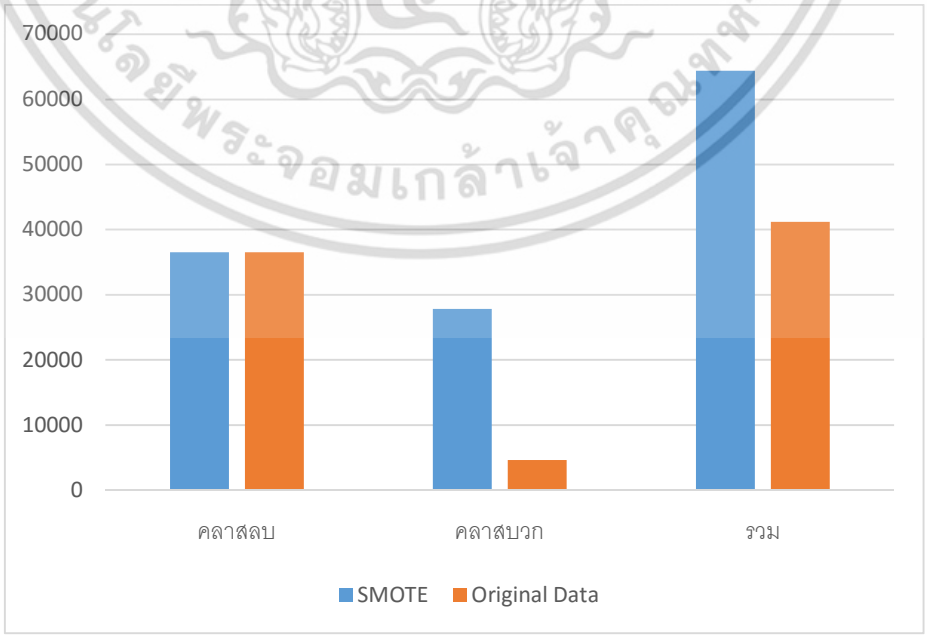
3.3 การปรับเพิ่มข้อมูลเพื่อให้มีความสมดุลโดยการใช้อัลกอริทึมเอสเอ็มโอทีอี (SMOTE Algorithm)

จากข้อมูลที่ได้นำมาทำการทดลองเป็นข้อมูลในลักษณะที่ไม่สมดุล คือเป็นข้อมูลที่มีความแตกต่างระหว่างสองคลาสอย่างชัดเจน ซึ่งในชุดข้อมูลทดสอบนี้มีคลาสลบจำนวน 36,548 ข้อมูล แต่คลาสบวกมีข้อมูลอยู่เพียง 4,640 ข้อมูล รวมทั้งหมดเป็น 41,188 ข้อมูล และใช้อัลกอริทึมเอสเอ็มโอทีอี ในการปรับสมดุลข้อมูล โดยการปรับมาจอริทีคลาสเพิ่มเป็น 500% จากเดิมจำนวนข้อมูลมี 4,640 ข้อมูล เพิ่มเป็น 27,840 ข้อมูล ดังนั้นข้อมูลที่จากเดิมมี 41,188 ข้อมูล ถูกปรับเพิ่มเป็น 64,388 ข้อมูล ตารางที่ 3.7 แสดงการเปรียบเทียบระหว่างความแตกต่างของข้อมูลที่มีการปรับเพิ่มขึ้นแล้ว กับข้อมูลที่ยังไม่มีการปรับเพิ่ม

ตารางที่ 3.7 แสดงการเปรียบเทียบระหว่างข้อมูลที่มีการปรับแล้วกับข้อมูลที่ยังไม่มีการปรับเพิ่มขึ้น

ข้อมูล	คลาสลบ	คลาสบวก	รวม
SMOTE	36,548	27,840	64,388
Original Data	36,548	4,640	41,188

ตารางที่ 3.1 แสดงการเปรียบเทียบระหว่างข้อมูลเดิมกับข้อมูลที่มีการแปลงด้วยอัลกอริทึม เอสเอ็มโอทีอี



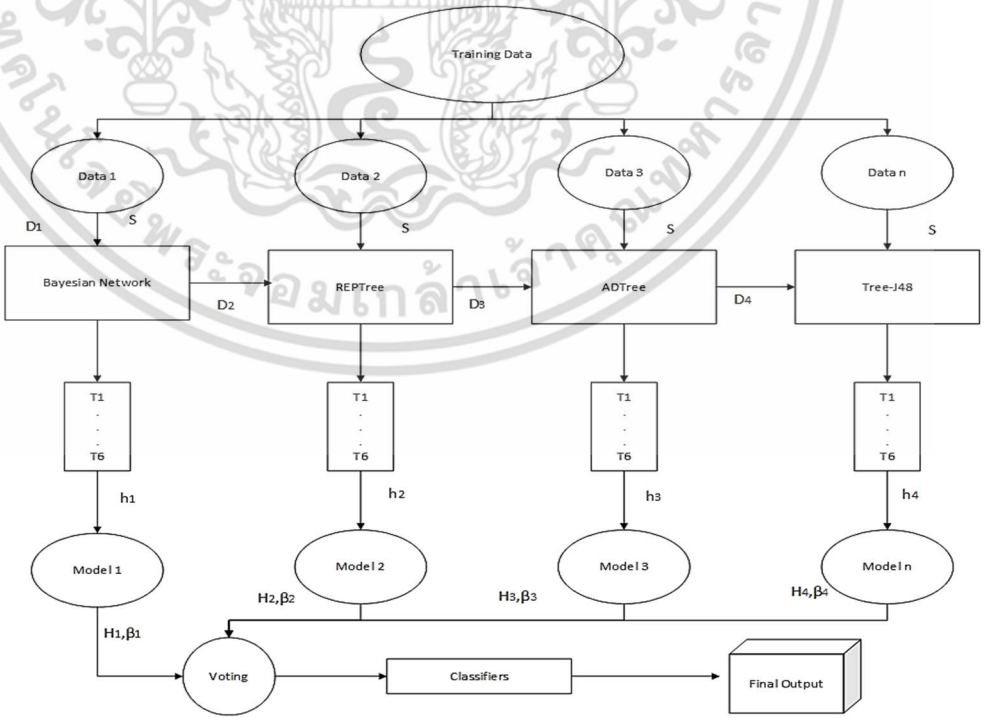
รูปที่ 3.1 แสดงการเปรียบเทียบระหว่างข้อมูลเดิมกับข้อมูลที่ผ่านการแปลงด้วยเอสเอ็มโอทีอี

3.4 ขั้นตอนการวัดประสิทธิภาพ

ในงานวิจัยนี้ได้ทำการใช้ชุดข้อมูลทดสอบจากเว็บไซต์ยูซีไอ มาใช้สำหรับการทดสอบวัดประสิทธิภาพการจำแนกประเภท โดยการปรับความสมดุลของข้อมูลด้วยอัลกอริทึมเอสเอ็มโอทีอี และในส่วนของการจำแนก ได้ทำการแบ่งชุดข้อมูลสำหรับฝึกสอนเป็น 80 เปอร์เซ็นต์ และแบ่งชุดข้อมูลสำหรับใช้ทดสอบเป็น 20 เปอร์เซ็นต์ ทำงานบนโปรแกรม MATLAB เวอร์ชัน R2014a และ WEKA เวอร์ชัน 3.6.13 โดยใช้ MATLAB สร้างแบบจำลองการเรียนรู้แบบรวมกลุ่มร่วมกับเอดาบูสต์เอ็มทู และใช้อัลกอริทึมที่อยู่ใน WEKA เป็นตัวจำแนกพื้นฐาน

ขั้นตอนการวัดประสิทธิภาพได้แบ่งการทดสอบออกเป็นการเรียนรู้แบบสลับตำแหน่งทั้งหมด 24 ลำดับ โดยสลับตำแหน่งระหว่างโครงข่ายแบบเบย์ ต้นไม้ตัดสินใจแบบเอดีทรี ต้นไม้ตัดสินใจแบบอาร์อีพีทรี ต้นไม้ตัดสินใจแบบเจโฟร์ทีเอต เพื่อหาลำดับการเรียนรู้ที่ดีที่สุดในการสร้างแบบจำลอง และทดสอบการวนรอบการเรียนรู้ตั้งแต่สองถึงหกครั้ง เพื่อหาจำนวนครั้งที่ดีที่สุด สำหรับนำไปใช้ทดสอบการวัดประสิทธิภาพทั้ง 6 แบบ

การวัดประสิทธิภาพด้วยแบบจำลอง BRAC โดยข้อมูลนำเข้าคือชุดข้อมูลทดสอบ โดยเริ่มที่โครงข่ายแบบเบย์เป็นลำดับแรก ลำดับที่สองคือต้นไม้ตัดสินใจแบบอาร์อีพีทรี ต้นไม้ตัดสินใจแบบเอดีทรี และลำดับสุดท้ายต้นไม้ตัดสินใจแบบเจโฟร์ทีเอต ทดสอบการวนรอบในส่วนการเรียนรู้ของเอดาบูสต์เอ็มทูจำนวน 6 ครั้ง เพื่อคัดเลือกสมมติฐานที่ดีที่สุด สำหรับนำไปใช้ในการจำแนกประเภท ดังแสดงในรูปที่ 3.2



รูปที่ 3.2 การเรียนรู้แบบรวมกลุ่มของแบบจำลอง BRAC

บทที่ 4
ผลการวิจัยและการอภิปรายผล

ในบทนี้กล่าวถึงผลการทดลองเปรียบเทียบประสิทธิภาพความแม่นยำของแบบจำลองที่พัฒนาขึ้น โดยการเปรียบเทียบการสลับตำแหน่งทั้ง 24 แบบที่ไม่ซ้ำกัน โดยเปรียบเทียบประสิทธิภาพการจำแนกด้วยข้อมูลที่ผ่านการปรับสมดุลแล้วกับข้อมูลเดิม ซึ่งใช้การวัดประสิทธิภาพด้วยค่าความแม่นยำของคลาสบวก (Sensitivity) ค่าความแม่นยำของคลาสลบ (Specificity) ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าเอฟเมเชอร์ (F-measure) ในการเปรียบเทียบ พร้อมทั้งเปรียบเทียบประสิทธิภาพการจำแนกโดยการกำหนดจำนวนรอบการเรียนรู้ตั้งแต่สองรอบถึงหกรอบดังรายละเอียดต่อไปนี้

4.1 ข้อมูลที่นำมาใช้สำหรับการทดลอง

ข้อมูลที่นำมาใช้สำหรับการทดลองได้จากการเก็บรวบรวมจากธนาคารโปรตุเกส ถูกแบ่งเป็นสองคลาส คือ คลาสบวกและคลาสลบ ซึ่งมีจำนวนข้อมูลทั้งหมด 41,188 เรคอร์ด (Record) มี 20 คุณลักษณะ โดยข้อมูลถูกเก็บรวบรวมตั้งแต่เดือนเมษายน ปี ค.ศ. 2008 ถึงเดือนพฤศจิกายน ปี ค.ศ. 2010 เผยแพร่ในเว็บไซต์ยูซีไอซึ่งมีรายละเอียดของข้อมูลดังแสดงในตารางที่ 4.1

ตารางที่ 4.1 รายละเอียดข้อมูลที่นำมาใช้สำหรับทำการทดลอง

รายการ	คุณลักษณะ
ลักษณะของข้อมูล	หลายตัวแปร
ลักษณะของคุณลักษณะข้อมูล	ข้อมูลจำนวนจริง
ลักษณะงานที่นำมาใช้	การจำแนกประเภท
จำนวนข้อมูล	41,188
จำนวนคุณลักษณะ	20
ค่าที่ขาดหาย (Missing Value)	ไม่มี
ขอบเขตของข้อมูล	ข้อมูลทางด้านธุรกิจ
วันที่เก็บรวบรวมข้อมูล	2008/04-2010/11
จำนวนคลาส	2

4.2 เครื่องมือที่ใช้สำหรับการทดลอง

เครื่องมือที่ใช้สำหรับทำการทดสอบการเรียนรู้ของเครื่องและทดสอบประสิทธิภาพของการจำแนกประเภท สำหรับการดำเนินการตรวจสอบสมมติฐานและทำการทดลองของงานวิจัย มีดังต่อไปนี้

1) MATLAB Version R2014a (64-bit)
2) Waikato Environment for Knowledge Analysis (WEKA) Version 3.6.13
3) ทำงานบนระบบปฏิบัติการณ์ Window 10 (2015) Pro Processor: Intel(R) Core(TM) i7-5500U CPU: 2.40 GHz 64-bit RAM: 8.00 GB

4.3 ขั้นตอนการกำหนดค่าสำหรับทำการทดลอง

สำหรับงานวิจัยฉบับนี้ได้ปรับปรุงแบบจำลองการเรียนรู้แบบรวมกลุ่ม ซึ่งประกอบด้วยโครงข่ายแบบเบย์ การเรียนรู้แบบต้นไม้ตัดสินใจทั้งหมด 3 แบบ คือ เอดีทรี อาร์อีพีทรี และทรีเจโฟร์ทีเอต การกำหนดค่าสำหรับการทดลองมีดังนี้

1) กำหนดจำนวนการวนรอบทั้งหมด 4 รอบ ยกเว้นในตารางผลการทดลองที่แสดงผลการทดลองสำหรับจำนวนการวนรอบโดยเฉพาะ ใช้ทั้งหมด 6 รอบ
2) สำหรับโครงข่ายแบบเบย์ ใช้อัลกอริทึม K2 ในการเรียนรู้
3) กำหนดให้จำนวนเพื่อนบ้านใกล้เคียงของเอสเอ็มโอทีอี เท่ากับ 5 ตัว จำนวนเปอร์เซ็นต์การปรับเพิ่มขึ้นเท่ากับ 500 เปอร์เซ็นต์ กำหนดการสุ่มตัวอย่างข้อมูลจำนวน 1 ชุดข้อมูลตัวอย่าง
4) สำหรับการเรียนรู้ต้นไม้ตัดสินใจอาร์อีพีทรี กำหนดให้ค่าความลึกสูงสุดของต้นไม้ (The maximum of tree depth) เท่ากับ -1 เสมอ กำหนดค่าน้ำหนักรวมต่ำสุดของการเกิดใบใหม่เท่ากับ 2.0 กำหนดค่าสัดส่วนความแปรปรวนไม่เกิน 0.001 กำหนดจำนวนของข้อมูลที่ใช้สำหรับการตัดแต่งกิ่งต้นไม้เท่ากับ 3
5) สำหรับต้นไม้ตัดสินใจทรีเจโฟร์ทีเอต กำหนดให้ค่าปัจจัยความเชื่อมั่นในการตัดแต่งกิ่งเท่ากับ 0.25 กำหนดจำนวนขั้นต่ำในการเกิดใบเท่ากับ 2.0 กำหนดจำนวนข้อมูลที่ใช้สำหรับการตัดแต่งกิ่งที่ผิดพลาด เพื่อต้นไม้จะได้เจริญเติบโตต่อไปเท่ากับ 3
6) สำหรับต้นไม้ตัดสินใจเอดีทรี กำหนดจำนวนการวนรอบเท่ากับ 4 รอบ ค่าการขยายโหนดเท่ากับ -3

4.4 ผลการทดลอง

ผลการทดลองแบบสลับตำแหน่งในแต่ละแบบจำลองการเรียนรู้ทั้งหมด 24 แบบ โดยกำหนดให้แต่ละตัวจำแนกพื้นฐานได้เริ่มทำงานเป็นลำดับแรกทุกครั้ง และแสดงผลการทดลองทุกตำแหน่งที่มีการสลับกัน ซึ่งการสลับตำแหน่งของตัวจำแนกพื้นฐานในการเรียนรู้แบบรวมกลุ่มนั้น เพื่อทดสอบประสิทธิภาพการวัดผลที่ครอบคลุมการเรียนรู้ที่สามารถเป็นไปได้ทั้งหมด ในการสร้างแบบจำลองที่มีประสิทธิภาพการจำแนกที่สูงที่สุด ในงานวิจัยนี้ใช้ตัวจำแนกพื้นฐานทั้งหมด 4 ตัว จำแนก คือโครงข่ายแบบเบย์ ต้นไม้ตัดสินใจแบบอาร์อีพีทรี ต้นไม้ตัดสินใจแบบเอดีทรี และต้นไม้ตัดสินใจแบบเจโฟร์ทีเอต ตารางที่ 4.2 แสดงการสลับลำดับของแบบจำลองทั้ง 24 แบบ

โดยกำหนดให้ B คือ “Bayesian Network”

R คือ “REPTree”

A คือ “ADTree”

C คือ “Tree-J48”

สำหรับการเปรียบเทียบการสลับตำแหน่งทั้ง 24 ตำแหน่งพบว่าประสิทธิภาพของการจำแนกไม่แตกต่างกันมากนักทั้งจากการวัดประสิทธิภาพทั้งหมด 6 แบบคือค่าความแม่นยำของคลาสบวก ค่าความแม่นยำของคลาสลบ ค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าเอฟเมเซอร์ ซึ่งในการสลับตำแหน่งที่แบบจำลอง BRAC โดยมีโครงข่ายแบบเบย์เป็นลำดับที่ 1 อาร์อีพีทรีเป็นลำดับที่ 2 เอดีทรีเป็นลำดับที่ 3 และทรีเจโฟร์ทีเอต เป็นลำดับสุดท้ายนั้นให้ประสิทธิภาพการจำแนกประเภทสูงสุด โดยที่ค่าความแม่นยำของคลาสบวกได้ 0.9149 ค่าความแม่นยำของคลาสลบได้ 1.000 ค่าความถูกต้องได้ 0.9632 ค่าความแม่นยำได้ 1.000 ค่าความระลึกได้ 0.9149 และค่าเอฟเมเซอร์ได้ 0.9555 หากเปรียบเทียบการวัดประสิทธิภาพด้วยค่าความแม่นยำของคลาสบวก ซึ่งเป็นค่าที่ใช้วัดประสิทธิภาพการจำแนกของคลาสบวกนั้น แบบจำลอง BRAC ยังคงให้ประสิทธิภาพสูงสุดที่ 0.9149 รองลงมาคือแบบจำลอง RBAC 0.9109 และแบบจำลอง CBAR ได้ 0.9086 จากผลการทดลองสลับตำแหน่งทั้ง 24 ตำแหน่ง พบว่าค่าความแม่นยำของคลาสลบและค่าความแม่นยำ ซึ่งเป็นตัววัดประสิทธิภาพการจำแนกของคลาสลบสำหรับการทดลองในส่วนของข้อมูลที่มีความไม่สมดุลอยู่สูงนั้นจะทำให้การเรียนรู้นั้นแม่นยำเสมอสำหรับคลาสที่มีจำนวนมากกว่า ซึ่งหากใช้แบบจำลองการเรียนรู้แบบรวมกลุ่มซึ่งได้พัฒนาขึ้นมาในการจำแนกประเภท การทำนายประสิทธิภาพจะให้ผลแม่นยำสูงสุดที่ 1.000 หรือ 100 เปอร์เซ็นต์ ตารางที่ 4.3 แสดงผลการเปรียบเทียบประสิทธิภาพของแบบจำลองทั้ง 24 แบบ

ตารางที่ 4.2 การสลับลำดับของแบบจำลองทั้ง 24 แบบ

แบบจำลอง	ตำแหน่งของการเรียงลำดับ			
	1	2	3	4
BRAC	Bayesian Network	REPTree	ADTree	TREE-J48
BARC	Bayesian Network	ADTree	REPTree	TREE-J48
BCRA	Bayesian Network	TREE-J48	REPTree	ADTree
BRCA	Bayesian Network	REPTree	TREE-J48	ADTree
BCAR	Bayesian Network	TREE-J48	ADTree	REPTree
BACR	Bayesian Network	ADTree	TREE-J48	REPTree
RACB	REPTree	ADTree	TREE-J48	Bayesian Network
RCBA	REPTree	TREE-J48	Bayesian Network	ADTree
RABC	REPTree	ADTree	Bayesian Network	TREE-J48
RCAB	REPTree	TREE-J48	ADTree	Bayesian Network
RBCA	REPTree	Bayesian Network	TREE-J48	ADTree
RBAC	REPTree	Bayesian Network	ADTree	TREE-J48
ACBR	ADTree	TREE-J48	Bayesian Network	REPTree
ARCB	ADTree	REPTree	TREE-J48	Bayesian Network
ABCR	ADTree	Bayesian Network	TREE-J48	REPTree
ABRC	ADTree	Bayesian Network	REPTree	TREE-J48
ARBC	ADTree	REPTree	Bayesian Network	TREE-J48
ACRB	ADTree	TREE-J48	REPTree	Bayesian Network
CBRA	TREE-J48	Bayesian Network	REPTree	ADTree
CBAR	TREE-J48	Bayesian Network	ADTree	REPTree
CRAB	TREE-J48	REPTree	ADTree	Bayesian Network
CABR	TREE-J48	ADTree	Bayesian Network	REPTree
CARB	TREE-J48	ADTree	REPTree	Bayesian Network
CRBA	TREE-J48	REPTree	Bayesian Network	ADTree

ตารางที่ 4.3 ผลการวัดประสิทธิภาพการจำแนกของแบบจำลองทั้ง 24 แบบ

ลำดับ	แบบจำลอง	Sensitivity	Specificity	Accuracy	Precision	Recall	F-measure
1	**BRAC**	**0.9149**	**1.0000**	**0.9632**	**1.0000**	**0.9149**	**0.9555**
2	BARC	0.8820	1.0000	0.9490	1.0000	0.8820	0.9373
3	BCRA	0.8899	1.0000	0.9524	1.0000	0.8899	0.9417
4	BRCA	0.8940	1.0000	0.9542	1.0000	0.8940	0.9441
5	BCAR	0.8960	1.0000	0.9550	1.0000	0.8960	0.9452
6	BACR	0.8890	1.0000	0.9520	1.0000	0.8890	0.9412
7	RACB	0.8874	1.0000	0.9513	1.0000	0.8874	0.9403
8	RCBA	0.8883	1.0000	0.9517	1.0000	0.8883	0.9408
9	RABC	0.8885	1.0000	0.9518	1.0000	0.8885	0.9409
10	RCAB	0.8851	1.0000	0.9503	1.0000	0.8851	0.9390
11	RBCA	0.8903	1.0000	0.9526	1.0000	0.8903	0.9419
12	RBAC	0.9109	1.0000	0.9615	1.0000	0.9109	0.9534
13	ACBR	0.8924	1.0000	0.9535	1.0000	0.8924	0.9432
14	ARCB	0.8957	1.0000	0.9549	1.0000	0.8957	0.9450
15	ABCR	0.9007	1.0000	0.9571	1.0000	0.9007	0.9477
16	ABRC	0.8935	1.0000	0.9540	1.0000	0.8935	0.9438
17	ARBC	0.8967	1.0000	0.9554	1.0000	0.8967	0.9456
18	ACRB	0.8885	1.0000	0.9518	1.0000	0.8885	0.9409
19	CBRA	0.8865	1.0000	0.9509	1.0000	0.8865	0.9398
20	CBAR	0.9086	1.0000	0.9605	1.0000	0.9086	0.9521
21	CRAB	0.8962	1.0000	0.9551	1.0000	0.8962	0.9453
22	CABR	0.8906	1.0000	0.9527	1.0000	0.8906	0.9421
23	CARB	0.9036	1.0000	0.9583	1.0000	0.9036	0.9493
24	CRBA	0.8955	1.0000	0.9548	1.0000	0.8955	0.9449

จากผลการวัดประสิทธิภาพการจำแนกแบบสลับกันทั้ง 24 ตำแหน่งนั้น แบบจำลอง BRAC ให้ผลการวัดประสิทธิภาพสูงสุดในทุกค่าการเปรียบเทียบ จึงนำแบบจำลอง BRAC มาใช้สำหรับทดสอบประสิทธิภาพการจำแนกด้วยการวนรอบการเรียนรู้จำนวน 6 รอบ สำหรับแบบจำลองการเรียนรู้แบบรวมกลุ่มที่ใช้เอดาบูสต์เอ็มทูนั้น ในแต่ครั้งของการวนรอบจะมีการปรับปรุงสมมติฐานและทำการคัดเลือกตัวจำแนกที่ดีที่สุด ดังนั้นการวนรอบในแต่ละครั้งจะให้ค่าการวัดประสิทธิภาพแตกต่างกันดังแสดงในตารางที่ 4.4

ตารางที่ 4.4 ผลการทดลองวัดประสิทธิภาพในแต่ละรอบของการเรียนรู้

จำนวนรอบ	Sensitivity	Specificity	Accuracy	Precision	Recall	F-measure
2	0.8994	1.0000	0.9565	1.0000	0.8994	0.9470
3	0.8996	1.0000	0.9566	1.0000	0.8996	0.9471
4	**0.9149**	**1.0000**	**0.9632**	**1.0000**	**0.9149**	**0.9555**
5	0.8967	1.0000	0.9554	1.0000	0.8967	0.9456
6	0.8901	1.0000	0.9525	1.0000	0.8901	0.9418

จากตารางที่ 4.4 การวนรอบครั้งที่ 4 ให้ประสิทธิภาพการจำแนกสูงสุด โดยค่าความแม่นยำของคลาสบวกได้ 0.9149 ค่าความแม่นยำของคลาสลบได้ 1.000 ค่าความถูกต้องได้ 0.9632 ค่าความแม่นยำได้ 1.000 ค่าความระลึกได้ 0.9149 ค่าเอฟเมเชอร์ได้ 0.9555 ซึ่งในการวนรอบ 5 ครั้งและ 6 ครั้ง นั้นให้ประสิทธิภาพการจำแนกใกล้เคียงกับ 2 ครั้ง และ 3 ครั้ง แสดงให้เห็นว่าในการสุ่มข้อมูลชุดตัวอย่างขึ้นมาเพื่อนำมาคัดเลือกในแต่ละครั้งนั้น ข้อมูลชุดตัวอย่างที่ได้จะไม่เหมือนเดิมในแต่ละการวนรอบ ดังนั้นค่าที่ได้จึงมีความแตกต่างกัน แต่ไม่ได้แตกต่างกันมากนักในแต่ละวิธีวัดประสิทธิภาพการจำแนก

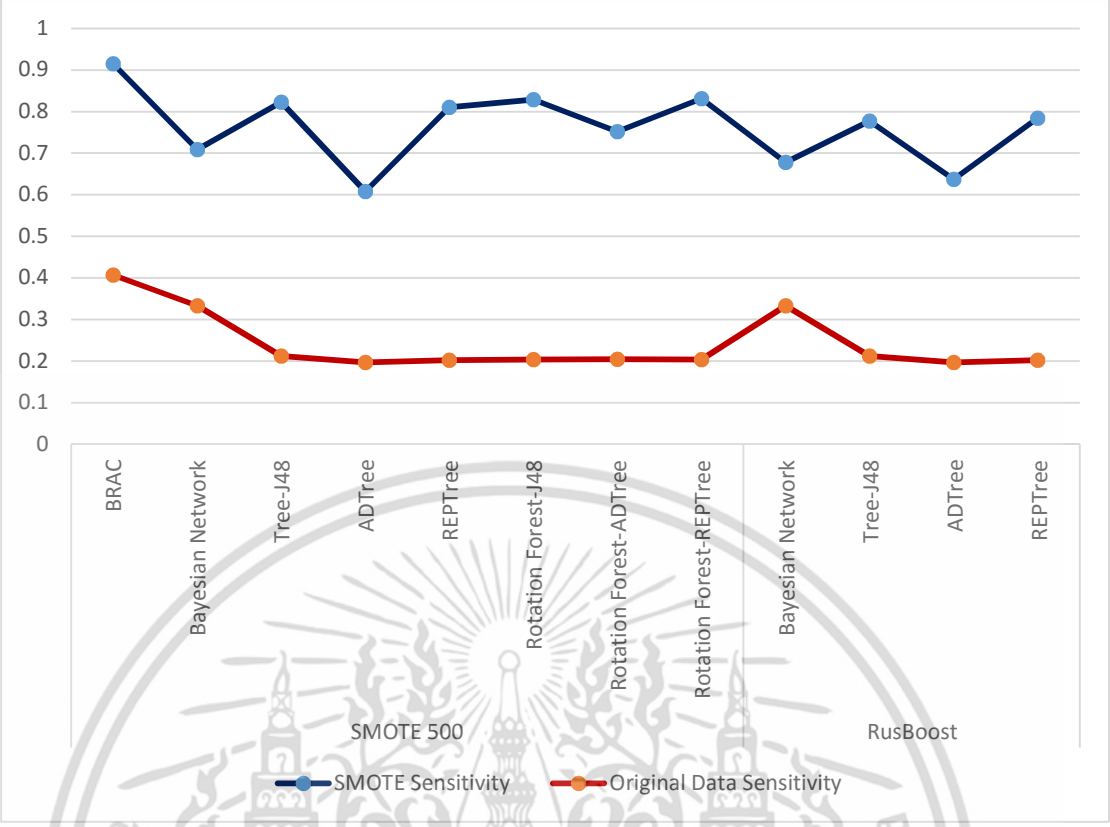
จากผลการวัดประสิทธิภาพการจำแนกด้วยการเปรียบเทียบระหว่างแบบจำลองการเรียนรู้แบบ BRAC กับแบบจำลองการเรียนรู้แบบอื่น ที่มีการปรับสมดุลข้อมูลแล้ว กับชุดข้อมูลเดิมที่ยังไม่มีการปรับสมดุลข้อมูลโดยใช้การวัดประสิทธิภาพด้วยค่าความแม่นยำของคลาสบวก ค่าความแม่นยำของคลาสลบ ค่าความถูกต้อง พบว่าการเรียนรู้กับชุดข้อมูลที่มีการปรับสมดุลแล้วให้ค่าการวัดประสิทธิภาพการจำแนกสำหรับคลาสบวกหรือสำหรับชุดข้อมูลที่ใช้ทดลองคือคลาส Yes นั้นให้ค่าความแม่นยำของคลาสบวกนั้นสูงสุดในทุกการเปรียบเทียบ โดยที่แบบจำลอง BRAC นั้นให้ค่าสูงสุดที่ 0.9149 เมื่อเปรียบเทียบกับการใช้ชุดข้อมูลเดิมในการทดสอบ แบบจำลอง BRAC ให้ประสิทธิภาพการจำแนกด้วยค่าความแม่นยำของคลาสบวกที่ 0.4073 ซึ่งสูงกว่าในทุกแบบจำลองสำหรับการ

ทดสอบกับชุดข้อมูลเดิม แต่ค่าการวัดประสิทธิภาพในคลาสลบ หรือในชุดข้อมูลนี้คลาสลบ คือคลาส No ชุดข้อมูลเดิมที่ยังไม่มีการปรับสมดุลข้อมูลให้ค่าความแม่นยำของคลาสลบสูงสุดในทุกแบบจำลองการเรียนรู้ ยกเว้นแบบจำลอง BRAC ที่ทดสอบกับชุดข้อมูลที่มีการปรับแล้ว ให้ค่าความแม่นยำของคลาสลบเท่ากับการทดลองกับชุดข้อมูลเดิมที่ 1.000 เท่ากัน สำหรับค่าความถูกต้องแบบจำลอง BRAC ได้ค่าสูงสุดทั้งจากการทดสอบด้วยข้อมูลที่มีการปรับสมดุลแล้วกับข้อมูลที่ยังไม่มีการปรับสมดุล ได้ค่า 0.9632 และสำหรับชุดข้อมูลเดิมได้ค่า 0.9332 ดังตารางที่ 4.5

ตารางที่ 4.5 เปรียบเทียบประสิทธิภาพการจำแนกกับแบบจำลองการเรียนรู้แบบอื่นโดยใช้การวัดค่าความแม่นยำของคลาสบวก ค่าความแม่นยำของคลาสลบ และค่าความถูกต้อง

แบบจำลอง		SMOTE			Original Data		
		Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
SMOTE 500	BRAC	**0.9149**	**1.0000**	**0.9632**	**0.4073**	**1.0000**	**0.9332**
	Bayesian Network	0.7087	0.7944	0.7558	0.3330	0.9524	0.8826
	Tree-J48	0.8226	0.8475	0.8371	0.2123	0.9799	0.8934
	ADTree	0.6077	0.8177	0.6943	0.1972	0.9862	0.8973
	REPTree	0.8101	0.8402	0.8276	0.2026	0.9814	0.8937
	Rotation Forest-J48	0.8285	0.8532	0.8428	0.2037	0.9807	0.8932
	Rotation Forest-ADTree	0.7520	0.7899	0.7743	0.2047	0.9848	0.8969
	Rotation Forest-REPTree	0.8312	0.8479	0.8410	0.2037	0.9832	0.8954
RusBoost	Bayesian Network	0.6781	0.8129	0.7455	0.3330	0.9524	0.8826
	Tree-J48	0.7773	0.8736	0.8283	0.2123	0.9799	0.8934
	ADTree	0.6371	0.8469	0.7259	0.1972	0.9862	0.8973
	REPTree	0.7844	0.8551	0.8233	0.2026	0.9814	0.8937

รูปที่ 4.1 แสดงกราฟเส้นเปรียบเทียบประสิทธิภาพการจำแนกประเภทด้วยค่าความแม่นยำของคลาสบวก โดยการเปรียบเทียบระหว่างข้อมูลเดิมกับข้อมูลที่มีการปรับสมดุลแล้ว กำหนดให้เส้นสีน้ำเงินแทนข้อมูลที่มีการปรับสมดุลแล้ว และเส้นสีแดงแทนข้อมูลเดิม จากกราฟเส้นจะเห็นได้ว่าในส่วนค่าของการวัดประสิทธิภาพนั้น แบบจำลอง BRAC ให้ค่าความแม่นยำของคลาสบวกทั้งจากการวัดผลกับข้อมูลเดิมและการวัดผลของข้อมูลที่มีการปรับสมดุลแล้วสูงที่สุด



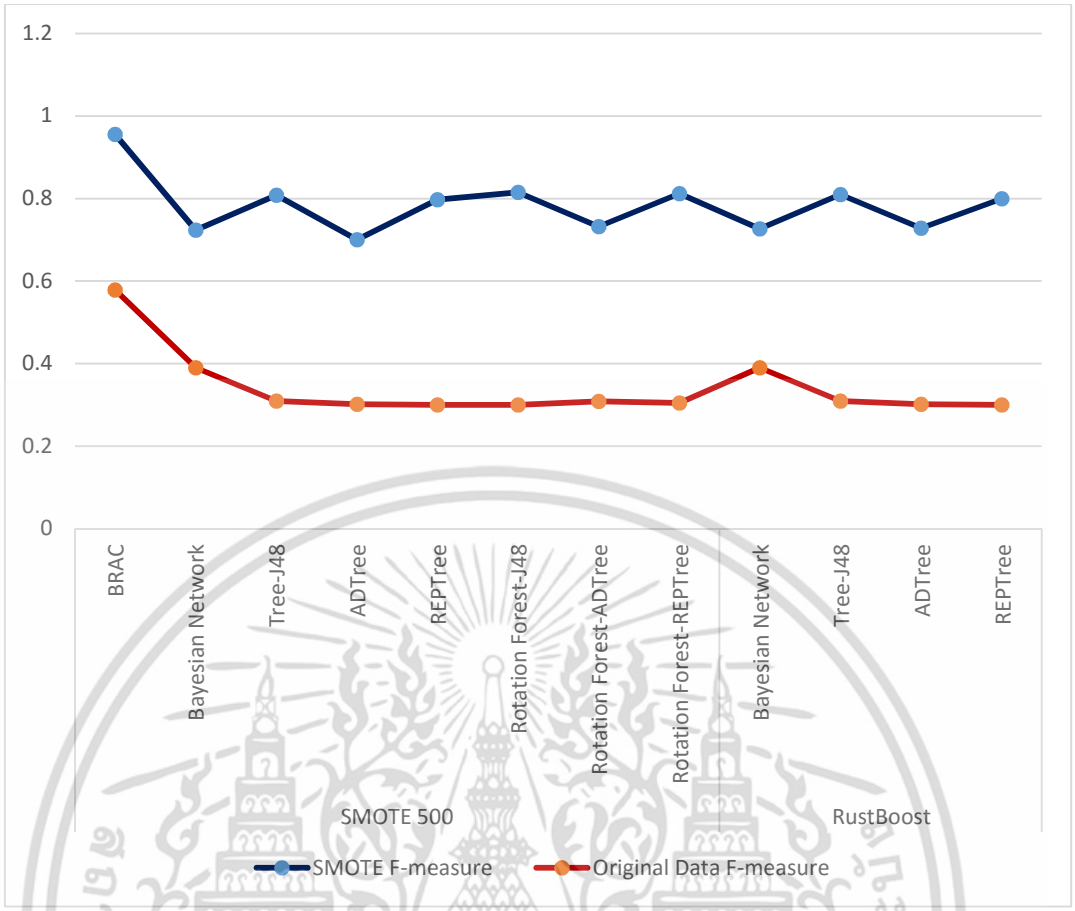
รูปที่ 4.1 กราฟเส้นแสดงการเปรียบเทียบประสิทธิภาพการจำแนกด้วยค่าความแม่นยำของคลาสบวก

จากผลการวัดประสิทธิภาพการจำแนกด้วยการเปรียบเทียบระหว่างแบบจำลองการเรียนรู้แบบ BRAC กับแบบจำลองการเรียนรู้แบบอื่น ที่มีการปรับสมดุลข้อมูลแล้ว กับชุดข้อมูลเดิมที่ยังไม่มีการปรับสมดุลข้อมูลโดยใช้การวัดประสิทธิภาพด้วยค่าความแม่นยำ ค่าความระลึก และค่าเอฟเมเชอร์ พบว่าแบบจำลองการเรียนรู้แบบรวมกลุ่มแบบ BRAC ให้ค่าการวัดประสิทธิภาพสูงสุดในทุกแบบจำลอง หากพิจารณาเปรียบเทียบในแต่ละค่าการวัดประสิทธิภาพ โดยเปรียบเทียบค่าความแม่นยำ ทั้งจากชุดข้อมูลเดิมและชุดข้อมูลที่มีการปรับสมดุลแล้ว แบบจำลอง BRAC ให้ค่า 1.0000 รองลงมาคืออาร์ยูเอสบูสต์ เอดีทรี ให้ค่า 0.8502 สำหรับชุดข้อมูลเดิมแบบจำลอง BRAC ให้ค่า 0.4073 รองลงมาคือโครงข่ายแบบเบย์ ให้ค่า 0.3330 ในการเปรียบเทียบค่าความระลึกกับชุดข้อมูลเดิมและชุดข้อมูลที่ยังไม่มีการปรับสมดุล แบบจำลอง BRAC ให้ค่า 0.9149 รองลงมาคือโรเทชันฟอเรสท์ อาร์อีพีทรี ให้ค่า 0.8312 และชุดข้อมูลเดิมแบบจำลอง BRAC ให้ค่า 1.000 รองลงมาคือเอดีทรี ให้ค่า 0.6444 เปรียบเทียบค่าเอฟเมเชอร์กับชุดข้อมูลเดิมและชุดข้อมูลที่ยังไม่มีการปรับสมดุล BRAC ให้ค่าสูงสุดที่ 0.9555 รองลงมาคือโรเทชันฟอเรสท์ อาร์อีพีทรี ให้ค่า 0.8119 และชุดข้อมูลเดิมแบบจำลอง BRAC ให้ค่า 0.5789 รองลงมาคือโครงข่ายแบบเบย์ ให้ค่า 0.3899 ดังแสดงในตารางที่ 4.6

ตารางที่ 4.6 เปรียบเทียบประสิทธิภาพการจำแนกกับแบบทำลองการเรียนรู้แบบอื่นโดยใช้การวัดค่าความแม่นยำของข้อมูลที่ดึงมาทั้งหมด ค่าความระลึก และค่าเอฟเมเซอร์

แบบจำลอง		SMOTE			Original Data		
		Precision	Recall	F-measure	Precision	Recall	F-measure
SMOTE 500	BRAC	**1.0000**	**0.9149**	**0.9555**	**0.4073**	**1.0000**	**0.5789**
	Bayesian Network	0.7389	0.7087	0.7235	0.3330	0.4703	0.3899
	Tree-J48	0.7945	0.8226	0.8083	0.2123	0.5727	0.3097
	ADTree	0.8261	0.6077	0.7003	0.1972	0.6444	0.3020
	REPTree	0.7854	0.8101	0.7976	0.2026	0.5802	0.3003
	Rotation Forest-J48	0.8026	0.8285	0.8154	0.2037	0.5727	0.3005
	Rotation Forest-ADTree	0.7134	0.7520	0.7322	0.2047	0.6312	0.3092
	Rotation Forest-REPTree	0.7935	0.8312	0.8119	0.2037	0.6058	0.3048
RustBoost	Bayesian Network	0.7834	0.6781	0.7269	0.3330	0.4703	0.3899
	Tree-J48	0.8450	0.7773	0.8097	0.2123	0.5727	0.3097
	ADTree	0.8502	0.6371	0.7284	0.1972	0.6444	0.3020
	REPTree	0.8156	0.7844	0.7997	0.2026	0.5802	0.3003

จากรูปที่ 4.2 แสดงกราฟเส้นเปรียบเทียบประสิทธิภาพการจำแนกประเภทด้วยค่าเอฟเมเซอร์ โดยการเปรียบเทียบระหว่างข้อมูลเดิมกับข้อมูลที่มีการปรับสมดุลแล้ว กำหนดให้เส้นสีน้ำเงินแทนข้อมูลที่มีการปรับสมดุลแล้ว และเส้นสีแดงแทนข้อมูลเดิม จากกราฟเส้นจะเห็นได้ว่าในส่วนค่าของการวัดประสิทธิภาพนั้น แบบจำลอง BRAC ให้ค่าความแม่นยำสูงที่สุด จากตำแหน่งเส้นกราฟของข้อมูลเดิม แสดงให้เห็นว่าในส่วนของค่าเอฟเมเซอร์มีค่าไม่แตกต่างกันมากนัก ดังจะเห็นได้จากเส้นสีแดงที่เกือบเป็นเส้นตรงต่อกัน เนื่องจากข้อมูลที่ยังไม่มีการปรับสมดุลจะทำให้ประสิทธิภาพการจำแนกต่ำมาก ไม่มีแบบจำลองใดให้ค่าสูงเกิน 50 เปอร์เซ็นต์ มีเพียงแบบจำลอง BRAC ที่ให้ค่าเกิน 50 เปอร์เซ็นต์ ซึ่งถ้าเปรียบเทียบการวัดประสิทธิภาพกับข้อมูลที่มีการปรับสมดุลแล้วนั้น ได้ค่าเกิน 50 เปอร์เซ็นต์ทุกแบบจำลอง



รูปที่ 4.2 กราฟเส้นแสดงการเปรียบเทียบประสิทธิภาพการจำแนกด้วยค่าเอฟเมเชอร์

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

งานวิจัยนี้ต้องการแก้ปัญหาความไม่สมดุลของข้อมูล ซึ่งถือว่าเป็นปัญหาหนึ่งที่ยากต่อการจำแนกประเภท ทั้งยังส่งผลในทางลบต่อความแม่นยำในการทำนายมาก ดังนั้นผู้วิจัยจึงได้ศึกษาและหาวิธีการแก้ปัญหาดังกล่าวโดยการตั้งสมมติฐานว่า หากเราสามารถลดจำนวนคลาสที่มีจำนวนที่มากกว่าหรือสามารถเพิ่มจำนวนของคลาสที่มีจำนวนน้อยกว่าได้ ก็จะสามารถทำให้ประสิทธิภาพในการจำแนกนั้นสูงขึ้นได้ ซึ่งเมื่อได้ทำการทดลองเปรียบเทียบประสิทธิภาพการจำแนกประเภทระหว่างข้อมูลเดิมที่ไม่ผ่านการปรับเพิ่มความสมดุลกับข้อมูลที่มีการปรับเพิ่มความสมดุลแล้วนั้น ผลการทดลองแสดงให้เห็นว่าประสิทธิภาพการจำแนกประเภทด้วยการใช้ข้อมูลที่มีการปรับเพิ่มความสมดุลนั้น สามารถให้ผลการจำแนกประเภทได้ดีกว่าข้อมูลเดิมที่ไม่ผ่านการปรับเพิ่มความสมดุล

อย่างไรก็ตาม ผู้วิจัยยังได้ตั้งข้อสมมติฐานเพิ่มเติมด้วยว่า การนำการจำแนกประเภทการเรียนรู้แบบรวมกลุ่มมาใช้ในการจำแนกประเภทนั้นน่าจะให้ประสิทธิภาพการจำแนกประเภทข้อมูลที่ไม่สมดุลได้ดีกว่าการใช้แบบจำลองการจำแนกประเภทแบบปกติ ดังนั้นผู้วิจัยจึงได้ทำการทดลองหาตัวจำแนกพื้นฐานแบบต่าง ๆ ที่สามารถให้ผลลัพธ์ในการจำแนกประเภทข้อมูลชุดนี้ได้ดี ซึ่งจากการทดลองผู้วิจัยค้นพบว่าต้นไม้ตัดสินใจแบบเอดีทรี ต้นไม้ตัดสินใจเจโฟร์ทีเอต ต้นไม้ตัดสินใจแบบอาร์อีพีทรี และโครงข่ายแบบเบย์ สามารถให้ประสิทธิภาพการจำแนกประเภทแบบเดี่ยวได้สูงที่สุด [19] ดังนั้นผู้วิจัยจึงได้นำตัวจำแนกพื้นฐานเหล่านี้มาพัฒนาแบบจำลองการเรียนรู้แบบรวมกลุ่ม โดยได้ทำการเลือกตัวจำแนกพื้นฐานที่สามารถให้ผลการจำแนกได้ดีที่สุดในการทดลองก่อนหน้า 4 ตัวมาใช้ ซึ่งผลการทดลองวัดประสิทธิภาพเปรียบเทียบระหว่างการใช้ตัวจำแนกพื้นฐานแบบเดี่ยวกับแบบรวมกลุ่มที่ใช้ตัวจำแนกพื้นฐานเป็นต้นไม้ตัดสินใจแบบเอดีทรี ต้นไม้ตัดสินใจเจโฟร์ทีเอต ต้นไม้ตัดสินใจแบบอาร์อีพีทรี และโครงข่ายแบบเบย์ นั้นแสดงให้เห็นว่า การเรียนรู้แบบรวมกลุ่มดังกล่าวสามารถให้ผลลัพธ์ความแม่นยำได้สูงที่สุด และแม่นยำมากกว่าตัวจำแนกประเภทเดี่ยวแบบต่าง ๆ ที่ได้นำมาทดลองทั้งหมด

นอกจากนี้ ในส่วนของการวัดประสิทธิภาพการจำแนกประเภทโดยใช้แบบจำลองที่พัฒนาขึ้นนั้น สามารถให้ประสิทธิภาพการจำแนกได้สูงที่สุดสำหรับทุกการวัดผลและทุกการเปรียบเทียบกับแบบจำลองอื่น แต่ในการเปรียบเทียบค่าความแม่นยำของคลาสลบระหว่างการปรับเพิ่มด้วยเอสเอ็มโอทีอีกับชุดข้อมูลเดิมที่ยังไม่มีการปรับเพิ่มนั้น แบบจำลองที่ได้พัฒนาขึ้นสามารถให้ผลลัพธ์การจำแนกได้เท่ากันกับชุดข้อมูลเดิมที่ 100 เปอร์เซ็นต์ ทั้งนี้มีสาเหตุมาจากอัลกอริทึมเอส

เอ็มโอทีอีจะทำการปรับเพิ่มเฉพาะคลาสที่มีจำนวนน้อยกว่า แต่ในคลาสที่มีจำนวนมากกว่านั้นข้อมูลจะยังคงเท่าเดิม

5.2 ข้อเสนอแนะ

ในงานวิจัยนี้ได้ใช้เทคนิคการเรียนรู้แบบรวมกลุ่มผสม คือการพัฒนาแบบจำลองขึ้นมาโดยใช้ตัวจำแนกพื้นฐาน 4 แบบ ซึ่งทำให้ผลการจำแนกมีประสิทธิภาพสูง แต่เวลาในการทำงานของแบบจำลองก็จะมากตามไปด้วย ยิ่งทดลองกับชุดข้อมูลที่มีปริมาณมากจะยิ่งทำให้ใช้เวลาในการคำนวณผลมากขึ้น การปรับเปลี่ยนอัลกอริทึมที่ใช้เป็นตัวจำแนกพื้นฐานที่นอกเหนือจากงานวิจัยนี้ ที่ใช้เวลาในการคำนวณน้อยลง อาจทำให้ผลลัพธ์ของประสิทธิภาพการจำแนกดีขึ้นและลดเวลาในการคำนวณผลลัพธ์ลงได้

เอกสารอ้างอิง

[1] "Direct Marketing Association emphasizes the importance of being 'Data Driven.'" 04-07-2015.

[2] Brandweek. 2009. “The Next Generation of Direct Marketing.” Academic Search Complete. 50(36): 6.

[3] "Postal Services OPERATIONS." 14-6-2015.

[4] "Marketing Legend Lester Wunderman Live on ‘The Alan Levy Show.” 14-6-2015

[5] Edelstein, H. A. 1999. “Introduction to data mining and knowledge discovery.” Two Crows Corp.

[6] Fayyad U. M., Piatetsky Shapiro, G. Smyth, P. and Uthurusamy R. 1996. “Advances in knowledge discovery & data mining.” Cambridge MA.

[7] Hastie, T. Tibshirani, R. and Friedman J. H. 2001. “The elements of statistical learning.” “Data mining inference and prediction.” New York: Springer.

[8] Han J., Kamber M. 2000. “Data mining: Concepts and Techniques.” New York: Morgan-Kaufman.

[9] กิติ ภักดีวัฒนะกุล. 2546. คัมภีร์ระบบสนับสนุนการตัดสินใจ และระบบผู้เชี่ยวชาญ. กรุงเทพฯ: เคทีพี คอมพ์ แอนด์ คอนซัลท์

[10] N. V. Chawla, W. K. Bowyer, O. L. Hall and W. P. Kegelmeyer. 2002 “SMOTE:Synthetic Minority Over-sampling Technique.” Journal of Artificial Intelligence Research. 321-357.

[11] S. Moro, P. Cortez and P. Rita. 2014. “A Data-Driven Approach to Predict the Success of Bank Telemarketing.” Decision Support Systems. Elsevier: 22-31.

[12] S. Moro, R. Laureano and P. Cortez. 2011. “Using Data Mining for Bank Direct Marketing.” An Application of the CRISP-DM Methodology. Proceedings of the European Simulation and Modelling Conference. Guimaraes. Portugal: EUROSIS. 117-121.

[13] ณัฐพงษ์ วารีประเสริฐ และณรงค์ ล่ำดี. 2552 ปัญญาประดิษฐ์ (Artificial Intelligence). กรุงเทพฯ: เคทีพี คอมพ์ แอนด์ คอนซัลท์.

[14] Yoav F. and Llew M. 1999. “The Alternating Decision Tree Algorithm.” Proceedings of the 16th International Conference on Machine Learning. 124-133

[15] Bernhard P., Geoffrey H. and Richard K. 2001. “Optimizing the Induction of Alternating Decision Trees.” Proceedings of the Fifth Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. 477-487

[16] Machine Learning Group. 2011. The University of Waikato. [Online} Available : http://weka.sourceforge.net/doc.stable/weka/classifiers/trees/REPTree.html

[17] Freund and Schapire. 1997. “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting.” University of Minnesota Department of Computer Science and Engineering.

[18] C. Seiffert, T.M. Khoshgoftaar, J. Van Hulse and A. Napolitano, “RUSBoost: A Hybrid Approach to Alleviating Class Imbalance.” Systems Man and Cybernetics. Part A: Systems and Humans. 40. 185-197.

[19] P. Ruangthong and J. Saiyen. 2015. “Bank direct marketing analysis of asymmetric information based on machine learning.” Computer Science and Software Engineering (JCSSE). Songkhla: 93-96.

[20] F. Yoav and S. E. Robert. “Experiments with a New Boosting Algorithm.” AT&T Research. Murray Hill: NJ, 07974-0636.

[21] F. Yoav and M. Llew. “The alternating decision tree learning algorithm” AT&T Research. Florham Park: NJ, 07932.

ภาคผนวก ก

งานวิจัยที่ตีพิมพ์

Bank Direct Marketing Analysis of Asymmetric Information Based on Machine Learning

Pumitara Ruangthong and Saichon Jaiyen
Department of Computer Science, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Thailand
s6605030@kmitl.ac.th, kjsaicho@kmitl.ac.th

***Abstract*—The bank direct marketing campaign for offering products that meet the customers' needs is the challenge problems. The bank direct marketing data analysis is important work that helps the banks predict whether customers will sign a long term deposits with the banks. The method that can predict such customers' needs can be profitable to the banks for improving their marketing campaign strategies. Unfortunately, it is very hard to predict the customers' needs because the available information is asymmetric. In this paper, we propose the method to analyze asymmetric information using SMOTE algorithm and Rotation Forest (PCA)-J48. The SMOTE method is used to modify the data and improve the accuracy of the prediction. The performance of the proposed method is evaluated and compared to Decision Tree, Rotation Forest, Navie Bayes, BayesNet, Multilayer Perceptron Neural Network, RBF Neural Network. The experimental results show the predicting accuracies of all predictors. The experiments show that Rotation Forest (PCA)-J48 can achieve the highest value of accuracy and specificity. However, the sensitivity of Rotation Forest (PCA)-J48 is higher than all methods except BayesNet and Rotation Forest (PCA) RandomTree.**

Keywords-Direct Marketing; SMOTE; Decision Tree; Rotation Forest(PCA); MLP (MultilayerPerceptron); NavieBayes; BayesNet; RBFNetwork

I. INTRODUCTION

The bank direct marketing campaign is very useful for offering the new products to their customers. The bank direct marketing data analysis can be used to select the type of bank direct marketing. Viral marketing is a marketing method to reach customers directly via the internet and communication technology. They may be offered via email, post, and phone to the customers. The product offering can be possibly directly done through the customers and allows them to decide the products for themselves. However, if the market does not meet the needs of their customers, the customers are likely to reject the products. In 2014, F. Koto [2] proposed SMOTEOut, SMOTE-Cosine and Selected-SMOTE algorithm for solving cases that have not been done by SMOTE. After testing on eighteen data sets, the experimental results showed that the proposed SMOTE can improve B-ACC and F1-Score. For solving a problem of imbalance data, M. H. Nguyen, W. E. Cooper and K. Kamei, [3] suggested the Multiple Random Under-Sampling (MRUS) technique for imbalanced and streaming data, while G. C. Weng and J. Poon [4] presented re-sampling method for imbalanced data sets. A. Sarmanova and S. Albayrak [5] proposed a method called GASEN (Genetic Algorithm based on Selective Ensemble Network) to enhance a performance of classifiers on class-imbalanced data sets. In addition, Z. Lei, P. Shaoning, C. Gang and A. Sarrafzadeh [6] suggested an incremental LPSVM (Linear Proximal Support Vector Machines) called DCIL-IncLPSVM for a classification of class imbalanced data set. In this experiment, DCIL-IncLPSVM was compared with SVM and LPSVM classifiers. The result has shown that the algorithm is highly effective in class imbalanced data. Furthermore, L. Tian-yu [7] proposed an algorithm called MIEE (Mutual Information based feature selection for EasyEnsemble) for improving classification accuracy in imbalanced data sets, and the experimental result demonstrated that MIEE can obtain better performance than bagging and EasyEnsemble. M. Tsunoda, A. Monden, J. Shibata, K. Matsumoto [8] compared the results of five classifiers, including linear discriminant analysis, logistic regression, classification tree, Mahalanobis-Taguchi method, and collaborative filtering methods to evaluate the classification performance of classifiers on imbalanced data. The results disclosed that collaborative filtering method gained the highest accuracy on the imbalanced data sets compared with the others.

In 2013, A. H. Elsalamony and A. M. Elsayad [9] has applied bank direct marketing domain to data classification. The data was prepared by using correlation, statistical measures, and cluster analysis. For this experiment, MLP and C5.0 were adopted as classifiers, and the data was divided into training and testing sets at 70:30 by the results. After testing, MLP gained 90.32% of accuracy, 60.12% of sensitivity and 93.45% of specificity, while C5.0 algorithm obtained 90.09% of accuracy, 59.06% of sensitivity and 93.23% of specificity. This experiment illustrated that MLP has slightly higher performance than C5.0. Additionally, P. Youqin and T. Zaiyong (2014) [10] used an ensemble method to classify the bank direct marketing domain. Bagging and boosting models was utilized to separate classes. The results showed ROC curve [11, 12] because of the class imbalance. It can be concluded that bagging neuron network released a separate class at best.

Therefore, the forecast of the customer response obtained from bank marketing is the target for increasing the accepted opportunity of their customers. Many data mining techniques are adopted as classifiers for this experiment because a large number of information and customers joining the long term deposits are clearly differentiated. In addition, the prediction of the imbalanced data resulted in less accurate [13] so SMOTE algorithm was suggested to deal with this problem, and it can present a reliable classification performance by integrating

978-1-4799-1966-6/15/$31.00 ©2015 IEEE

SMOTE algorithm with Rotation Forest (PCA)-J48 [14]. Consequently, this paper has proposed a method based on SMOTE algorithm with Rotation Forest (PCA)-J48 to predict the exceedingly imbalanced data.

II. THE PROPOSED METHOD

A. SMOTE Algorithm

SMOTE (Synthetic Minority Over-sampling Technique) [13] is an oversampling technique for improving minority class recognition. The classification model can recognize patterns of data with large quantities as well. Since a minority class is much smaller than other majority classes, oversampling or up-sampling is used to increase the number of the minority class in order to improve minority class recognition. In this research, we are interested in the minority class which is the class that customers agree to deposit with banks in long term. However, the data set is unbalanced and the minority class is very small compared to the majority class. Since SMOTE algorithm is the oversampling technique for improving minority class, we adopt SMOTE algorithm to solve our problem. The pseudo code of SMOTE algorithm is demonstrated in Algorithm 1.

Algorithm 1: SMOTE
Input:
Number of minority class samples T; Amount of SMOTE N%;
Number of nearest neighbors k
Output: (N/100)* T synthetic minority class samples

1. **if** N < 100 **then** Randomize the T minority class samples
2. $T = (N/100) * T$
3. $N = 100$
4. **Endif**
5. $\boldsymbol{N} = \left\lfloor \frac{N}{100} \right\rfloor$
6. **for** $i = 1$ **to** T
7. Compute k nearest neighbors for i and save the indices in the $nnarray$
8. Populate $(N, i, nnarray)$
9. **endfor**
10. **while** $N \neq 0$
11. Choose a random number between 1 and k, call it nn.
12. **for** $attr = 1$ **to** $numattrs$
13. Compute: $dif = Sample[nnarray[nn]][attr] - Sample[i][attr]$
14. Compute: gap = random number between 0 and 1
15. $Synthetic[newindex][attr] = Sample[i][attr] + gap * dif$
16. **Endfor**
17. $newindex++$
18. $N = N - 1$
19. **endwhile**

In Algorithm 1, the parameter, k, is the number of nearest neighbors. The parameter, *numattrs*, is the number of attributes. Sample[ ][ ] is an two dimensional array for original minority class samples. The parameter, *newindex*, is used to keep a count of number of synthetic samples generated, which is initialized to 0. Synthetic[ ][ ] is an two dimensional array for synthetic samples. Populate(N, i, *nnarray*) is a function to generate the synthetic samples.

B. Rotation Forest(PCA)-J48

J48 Decision Tree [18] is an algorithm for generating a recursive pruned or un-pruned C4.5 decision tree by using the information entropy on the set of training data. The feature data is divided into sub-sections and get regular information by measuring the difference in entropy used to measure these subsidiaries to identify the best attribute as a node in the cut. The decision tree is created with the J48. In Rotation forest [14], the feature set is divided into subsets and then, Principal Component Analysis (PCA) is applied to each subset. The pseudo code of Rotation Forest is showed in Algorithm 2.

Algorithm 2: ROTATION FOREST
Given:
X: the objects in the training data set (an N×n matrix)
Y: the labels of the training set (an N×1 matrix)
L: the number of classifiers in the ensemble
K: the number of subsets
$\{\omega_1, \omega_2, \ldots, \omega_c\}$: the set of class labels
For i = 1…L
1. Split **F** (the feature set) into K subsets: $\mathbf{F}_{i,j}$ (for j=1…K)
2. For j=1…K
   - Let $X_{i,j}$ be the data set X for the features in $\mathbf{F}_{i,j}$
   - Eliminate from $X_{i,j}$ a random subsets of classes
   - Select a bootstrap sample from $X_{i,j}$ of size 75% of the number of objects in $X_{i,j}$. Denote the new set by $X'_{i,j}$
   - Apply PCA on $X'_{i,j}$ to obtain the coefficients in a matrix $C_{i,j}$
3. Arrange the $C_{i,j}$, for j=1…K in a rotation matrix R_i as

$$R_i = \begin{bmatrix} a_{i,1}^{(1)}, a_{i,1}^{(2)}, \ldots, a_{i,1}^{(M_1)} & [0] & \ldots & [0] \\ [0] & a_{i,2}^{(1)}, a_{i,2}^{(2)}, \ldots, a_{i,2}^{(M_2)} & \ldots & [0] \\ \vdots & \vdots & \ddots & \vdots \\ [0] & [0] & \ldots & a_{i,K}^{(1)}, a_{i,K}^{(2)}, \ldots, a_{i,K}^{(M_K)} \end{bmatrix}$$

4. Construct R_i^α by rearranging the column of R_i so as to match the order of features in **F**
5. Build classifier D_i using (XR_i^α ,Y) as the training set

Classification Phase
For a given **x**, let $d_{i,j}(XR_i^\alpha)$ be the probability assigned by the classifier D_i to the hypothesis that **x** comes from class ω_j. Calculate the confidence for each class, ω_j. Calculate the confidence for each class, ω_j, by the average combination method:

$$\mu_j(\mathbf{x}) = \frac{1}{L}\sum_{i=1}^{L} d_{i,j(xR_i^\alpha),j=1,\ldots c.}$$

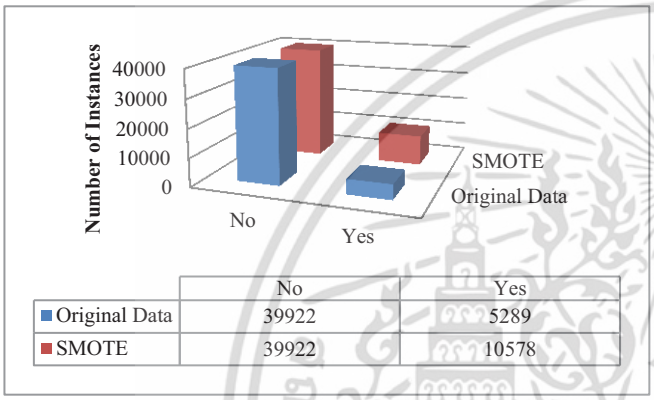
Assign **x** to the class with the largest confidence.

III. EXPERIMENT

A. Data Sets

In this experiment, data associated with direct marketing campaigns of the Portuguese banking institutions was collected from UCI Machine Learning Repository [15, 16, 17]. The

classifiers were implemented by WEKA data mining software in weka-3-6-12jre-x64 versions. The marketing campaign was offered via telephone. Deposit products were obtained from contacting customers via telephone, probably more than once, by banks. This data set contains 45211 instances with 17 attributes of 2 classes (yes/no). The classification goal is to predict whether customers will subscribe a term deposit. There are 39922 instances of no classes and 5289 instances of yes classes. Since the minority class (yes classes) contain 5289 instances which are very less than the majority class (no classes), the over-sampling technique which is SMOTE is applied to increase the number of instances of minority class. After applying SMOTE, the number of instances in minority class are increased to 10578 instances while the instances of majority class remain unchanged. In this examination, the number of yes classes is increased at 100%. The number of instances in original data and data improved by SMOTE are showed in figure 1.



	No	Yes
Original Data	39922	5289
SMOTE	39922	10578

Figure 1. The number of instances in original data and data improved by SMOTE algorithm.

In this experiment, the ratio of training and testing sets are 70:30. The parameters of SMOTE algorithm consists of S (the random number seed) which is set as 1, P (percentage of SMOTE instances to create.) which is set as 100, K (the number of nearest neighbors to use.) which is set as 5, C (the index of the nominal class value to SMOTE) which is set as 0. For Rotation Forest (PCA)-J48, maximum number of attributes to include in transformed attribute names is set as 5 and retain enough PC attributes to account for this proportion of variance in the original data is set as 1.0.

B. Performance Measurement

In this paper, the classification performance was measured by sensitivity, specificity and accuracy. For the following equations, TPR is the true value of positive rate; P is the positive class or yes class; N is the negative class or no class; T is the correct value; and F is the incorrect value.

1) Sensitivity is the true positive rate (TPR) which is defined as the fraction of positive instances predicted correctly by the model,

$$TPR = TP / (TP+FN). \quad (1)$$

2) Specificity is the true negative rate (TNR) which is defined as the fraction of negative instances predicted correctly by the model,

$$TNR = TN / (FP+TN). \quad (2)$$

3) Accuracy (ACC) is the overall success rate of the classifier defined as

$$ACC = (TP+TN) / (P+N) \quad (3)$$

C. Experimental Results

In this paper, the performance of the proposed model is evaluated and compared to Decision Tree, Rotation Forest, MLP neural network, RBF neural network, Bayesian network, and Naïve Bayes as shown in Table I. In addition, the experiments are conducted on two data sets which are the original data and the data improved by SMOTE algorithm. These two data sets are trained and tested on 14 machine learning algorithms. The comparative results are shown in Table I. From the experimental results, the algorithm that gives the highest sensitivity value on the original data is Rotation Forest (PCA)-ADTree that can achieve 67.73% of sensitivity on the original data set. The algorithm that produces the highest specificity value is Dicision Tree J48 that can achieve 93.19% of specificity on the original data set. However, after SMOTE algorithm improvement, the algorithm that gives the highest sensitivity value is BayesNet which can achieve 94.48% of sensitivity. In addition, the algorithm that gives the highest specificity and highest accuracy value are Rotation Forest (PCA)-J48 which can achieve 93.17% of specificity and 90.81% of accuracy, respectively. Since the number of instances between two classes is very different, techniques that can resolve imbalanced data should be applied. The performance metrics including sensitivity, specificity, and accuracy should be used to evaluate the performance of the classifiers on the imbalanced data set. The goal of this research is to predict whether customers will subscribe a term deposit based on asymmetric information. Therefore, we concentrate on both classes which are yes classes and no classes. So, the accuracy matric should be used to evaluate the classifiers. However, the sensitivity and specificity are also essential to evaluate the classifiers in order to analyze the performance of classifiers on each class. From the experimental results, it can be seen that the proposed method which adopts Rotation Forest (PCA)-J48 and SMOTE algorithm can achieve the highest accuracy value. This can signify that the proposed method that adopts Rotation Forest (PCA)-J48 and SMOTE algorithm is appropriated to classify the bank direct marketing data set.

IV. CONCLUSION

In this paper, the ensemble method which is Rotation Forest (PCA)-J48 with SMOTE has been proposed to solve the classification of imbalanced data set. The experimental results show that the proposed method that adopts Rotation Forest (PCA)-J48 and SMOTE algorithm is appropriated to classify the bank direct marketing data which is the imbalanced data. Since the imbalanced data set makes it difficult to classify,

using the algorithm to improve the data can make the imbalanced resolution better. From the experimental results, SMOTE algorithm shows that it can effectively solve the inequality of data. The obtained data is stable and can lead to reliable results. The imbalance problem is a challenging task. In the further work, another imbalanced data set is required for evaluating the classification performance of the proposed method.

TABLE I. THE COMPARATIVE RESULTS OF DECISION TREE, ROTATION FOREST, MLP, RBFNETWORK, BAYESNET, NAIVEBAYES.

Model	sensitivity(%)		specificity(%)		accuracy(%)	
	Original Data	SMOTE	Original Data	SMOTE	Original Data	SMOTE
J48	57.11	76.63	**93.19**	92.75	89.51	89.43
ADTree	60.03	80.10	90.98	88.77	89.21	87.47
LADTree	60.87	75.78	92.00	90.79	89.68	88.01
RandomTree	47.61	71.67	92.64	91.75	87.51	87.54
REPTree	60.05	79.26	92.78	92.15	89.85	89.68
Rotation Forest (PCA) J48	63.35	81.23	92.21	**93.17**	90.05	**90.81**
Rotation Forest (PCA) ADTree	**67.73**	80.12	89.78	87.81	89.12	86.75
Rotation Forest (PCA) LADTree	65.46	77.52	90.93	87.72	89.63	86.28
Rotation Forest (PCA) RandomTree	62.42	82.70	91.89	91.73	89.81	90.12
Rotation Forest (PCA) REPTree	63.48	81.09	91.90	92.08	89.93	90.03
MultilayerPerceptron	57.56	73.83	93.12	91.77	89.57	88.13
RBFNetwork	53.21	61.86	90.94	86.09	88.48	82.24
BayesNet	50.61	**94.48**	92.50	89.63	88.18	90.28
NaiveBayes	42.16	54.09	92.07	89.34	86.22	80.51

REFERENCES

[1] P. David Bianco, Direct Marketing Reference for Business URL: http://www.referenceforbusiness.com/encyclopedia/Dev-Eco/Direct-Marketing.html, 2015.

[2] F. Koto, "SMOTE-Out, SMOTE-Cosine, and Selected-SMOTE: An enhancement strategy to handle imbalance in data level", Advanced Computer Science and Information Systems (ICACSIS), 2014, pp. 280-284.

[3] M. H. Nguyen, W. E. Cooper and K. Kamei, "A comparative study on sampling techniques for handling class imbalance in streaming data", Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), 2012, pp. 1762-1767.

[4] G. C. Weng and J. Poon, "A Data Complexity Analysis on Imbalanced Datasets and an Alternative Imbalance Recovering Strategy", Web Intelligence, 2006, pp. 270-276.

[5] A. Sarmanova and S. Albayrak, "Alleviating class imbalance problem in data mining", Signal Processing and Communications Applications Conference (SIU), 2013, pp. 1-4.

[6] Z. Lei, P. Shaoning, C. Gang and A. Sarrafzadeh, "Class imbalance robust incremental LPSVM for data streams learning", Neural Networks (IJCNN), 2012, pp. 1-8.

[7] L. Tian-yu, "EasyEnsemble and Feature Selection for Imbalance Data Sets", Bioinformatics, Systems Biology and Intelligent Computing, 2009, pp. 517-520.

[8] M. Tsunoda, A. Monden, J. Shibata, K. Matsumoto, "Empirical Evaluation of Cost Overrun Prediction with Imbalance Data", Computer and Information Science (ICIS), 2011, pp. 415-420.

[9] A. H. Elsalamony and A. M. Elsayad, "Bank Direct Marketing Based on Neural Network", International Journal of Engineering and Advanced Technology (IJEAT), vol. 2, pp. 392-400.

[10] P. Youqin and T. Zaiyong, "Ensemble method in bank direc marketing", Service Systems and Service Management (ICSSSM), 2014, pp.1-5.

[11] W. M. David and Power, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation", Journal of Machine Learning Technologies, 2011, pp. 37–63.

[12] J. Furnkranz, "Ensemble Classifiers" (class notes)

[13] N. V. Chawla, W. K. Bowyer, O. L. Hall and W. P. Kegelmeyer, "SMOTE:Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research, 2002, pp. 321-357.

[14] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, "Rotation Forest: A New Classifier Ensemble Method," IEEE Transactions on *Pattern Analysis and Machine Intelligence*,2006, vol. 28, no. 10, pp. 1619-1630.

[15] S. Moro, P. Cortez and P. Rita. "A Data-Driven Approach to Predict the Success of Bank Telemarketing". Decision Support Systems, Elsevier, 2014, pp. 22-31.

[16] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS.

[17] M. Sergio, C. Paulo and R. Paulo, "A data-driven approach to predict the success of bank telemarketing", Decision Support System, 2014, pp. 22-31.

[18] J. R. Quinlan, "Improved use of continuous attributes in c4.5.", Journal of Artificial Intelligence Research, 1996, pp. 77-90.

ประวัติผู้เขียน

ชื่อ	นางสาวภูมิธารา เรืองทอง
ที่อยู่ปัจจุบัน	97/249 ม.การ์เด้นซูท ถ.ราษฎร์พัฒนา แขวง/เขต สะพานสูง กทม.
ประวัติการศึกษา	(2558) วิทยาศาสตรมหาบัณฑิต สาขา วิทยาการคอมพิวเตอร์ เกรดเฉลี่ย 3.50 สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ผลงานทางวิชาการ	1. Bank Direct Marketing Analysis of Asymmetric Information Based on Machine Learning (JCSSE 2015)