ISSN 2229-1547

วารสารวิทยาศาสตร์และเทคโนโลยี มทร.ธัญบุรี
Science and Technology RMUTT Journal

ปีที่ 10 ฉบับที่ 1 (มกราคม - มิถุนายน 2563)
Vol.10 No.1 (January - June 2020)

คณะวิทยาศาสตร์และเทคโนโลยี
มหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรี
Faculty of Science and Technology
Rajamangala University of Technology Thanyaburi

Science and Technology
RMUTT Journal
Vol.10 No.1 (2020) : 215-225
www.sci.rmutt.ac.th/stj
Online ISSN 2229-1547

แบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุล โดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง

Fraud Detection Model in Imbalanced Data Using Dimension Reduction And Machine Learning Algorithms

นิเวศ จิระวิชิตชัย

คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยศรีปทุม เขตจตุจัตร กรุงเทพมหานคร 10900

E-mail: nivet.ch@spu.ac.th

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง เพื่อการจำแนกธุรกรรมที่มีความผิดปกติ (Fraud Detection) และหาความสัมพันธ์ของกลุ่มธุรกรรมผิดปกติ เพื่อป้องกันความเสียหายที่จะเกิดขึ้นจากธุรกรรมที่ทุจริตในระบบพาณิชย์อีเล็กทรอนิกส์ ผลการทดลองเมื่อวัดประสิทธิภาพแบบจำลองด้วยค่าความถูกต้อง (Accuracy) สรุปได้ว่าแบบจำลองที่ใช้อัลกอริทึม เอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) ให้ค่าความถูกต้องสูงที่สุดคือ 98.15 % ในขณะที่ใช้เวลาในการประมวลน้อยที่สุด จากการทดลองพบว่า แบบจำลองที่พัฒนาขึ้นนั้น ส่งผลให้อัลกอริทึมมีขีดความสามารถจำแนกธุรกรรมที่มีความผิดปกติได้อย่างมีประสิทธิภาพมากขึ้นอย่างชัดเจน

คำสำคัญ: การตรวจสอบการทุจริต เอ็กซ์ทรีมกาเดียนบูทติ้ง การเรียนรู้ของเครื่อง

Received: May 21, 2020
Revised: June 3, 20XX
Accepted: June 08, 20XX

Abstract

The objective of this research is to develop a method for fraud detection model in imbalanced data using dimension reduction combined with machine learning algorithms for fraud detection and finding the Relationship of irregular transaction groups. In order to prevent damage from fraudulent transactions in electronic commerce systems. The results of the experiment when testing the model performance with accuracy found the model that uses the Extreme Gradient Boosting algorithm gives the highest accuracy of 98.15%, while using the shortest processing time. From the experiment, it was found that the developed model resulted in the algorithm having the ability to more effectively.

Keywords: Fraud Detection, Extreme Gradient Boosting, Machine Learning

1. บทนำ

จากการขยายตัวด้านการใช้งานระบบพาณิชย์อีเล็กทรอนิกส์ ตลอดช่วงระยะเวลาที่ผ่านมา มีแนวโน้มในการใช้งานในด้านธุรกรรมการเงินเพิ่มมากขึ้นอย่างรวดเร็วและแพร่หลาย ส่งผลให้ผู้ใช้งานสามารถที่จะเข้าถึงข้อมูลการเงินส่วนตัวและส่งเสริมในสังคมเกิดการทำธุรกรรมผ่านช่องทางสื่อพาณิชย์อีเล็กทรอนิกส์ ระบบอินเตอร์แบงค์กิ้ง ระบบโมบายแบงกิ้ง ได้อย่างสะดวกสบายมากขึ้นและทำให้ได้รับความนิยมอย่างแพร่หลายในวงกว้าง และหนึ่งกรรมวิธีในการทำธุรกรรมออนไลน์ก็คือ การใช้บัตรเครดิตในการชำระค่าสินค้า

ปัจจุบันนอกจากบัตรเครดิตจะเป็นที่นิยมในการซื้อสินค้าตามราคาทั่วไปแล้ว ยังนิยมมาใช้ในการซื้อขายผ่านอินเทอร์เน็ตอีกด้วย เมื่อมีการซื้อขายสินค้าผ่านบัตรเครดิต ผู้ใช้บัตรเครดิตจะต้องแสดงความสมยอมว่าการซื้อขายนั้นได้เกิดขึ้นจริงด้วยการเซ็นชื่อในใบเสร็จ หากเป็นการซื้อขายทางอินเทอร์เน็ต ผู้ใช้อาจจะกรอกหมายเลขบัตรเครดิตและรหัสลับหลังบัตร เพื่อเป็นการแสดงความจำนงในการซื้อขาย ทำให้การใช้งานบัตรเครดิตซื้อสินค้าออนไลน์เติบโตอย่างรวดเร็ว แต่ปัญหาอย่างหนึ่งของธนาคารผู้ออกบัตรก็คือ การตรวจสอบการทำธุรกรรมออนไลน์ว่าเป็นการกระทำที่ผิดปกติหรือไม่ (Fraud detection) ซึ่งธุรกรรมทุจริตเหล่านี้ได้สร้างความเสียหายให้กับธุรกิจธนาคารเป็นอย่างมาก โดยเฉพาะอย่างยิ่งธุรกรรมทุจริตที่เกิดขึ้นจากการซื้อสินค้าผ่านช่องทางพาณิชย์อิเลคทรอนิคส์ (E-commerce) [1-3] ก็จะยิ่งเพิ่มความเสี่ยงจะที่เกิดธุรกรรมทุจริตมากกว่าธุรกรรมที่ต้องแสดงบัตรเครดิต ทั้งยังยากต่อการยืนยันว่าเป็นธุรกรรมทุจริตหรือไม่ ดังนั้นธนาคารผู้ออกเครดิตจึงได้รับความเสียหาย ทำให้เกิดต้นทุนกับธนาคารผู้ออกบัตรเพราะความเสียหายที่เกิดขึ้นธนาคารจะต้องเป็นผู้รับภาระเองทั้งหมด

จากความสำคัญดังกล่าว ผู้วิจัยจึงเห็นความสำคัญในการศึกษาแบบจำลองการป้องกันธุรกรรมทุจริต (Fraud Detection) เพื่อเป็นแนวทางในการเพิ่มขีดความสามารถขององค์กรและลด

ความเสียหายที่จะเกิดขึ้นจากธุรกรรมทุจริต โดยมุ่งเน้นพัฒนาแบบจำลองการจำแนกหาความสัมพันธ์ของกลุ่มธุรกรรมที่ผิดปกติ โดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่องที่ประสิทธิภาพและความแม่นยำในการจำแนกธุรกรรมทุจริตผ่านช่องทางพาณิชย์อิเลคทรอนิคส์เป็นปัจจัยหลัก ในขณะที่การใช้ทรัพยากรและเวลาเป็นปัจจัยรอง โดยองค์ความรู้ที่ได้จากงานวิจัยนี้สามารถนำไปประยุกต์ใช้กับการพัฒนาระบบงานการวิเคราะห์พฤติกรรมของลูกค้าเพื่อป้องกันธุรกรรมทุจริต (Fraud Detection) ได้มีประสิทธิภาพมากยิ่งขึ้น

2. วัตถุประสงค์ของการวิจัย

2.1 เพื่อสร้างแบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง

2.2 เพื่อทดสอบประสิทธิภาพของแบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง

3. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

การเรียนรู้ของเครื่อง (Machine Learning) [4-5] จัดเป็นสาขาหนึ่งของปัญญาประดิษฐ์ ที่พัฒนามาจากการศึกษาการรู้จำแบบ เกี่ยวข้องกับการศึกษาและการสร้างอัลกอริทึมที่สามารถเรียนรู้ข้อมูลและทำนายข้อมูลได้ อัลกอริทึมนั้นจะทำงานโดยอาศัยโมเดลที่สร้างมาจากชุดข้อมูลตัวอย่างขาเข้าเพื่อการทำนายหรือตัดสินใจในภายหลัง การเรียนรู้แบบมีผู้สอน (Supervised Learning) กล่าวคือมีข้อมูลตัวอย่างและผลลัพธ์ที่ผู้สอนต้องการถูกป้อนเข้าสู่คอมพิวเตอร์ เป้าหมายคือการสร้างกฎทั่วไปที่สามารถเชื่อมโยงข้อมูลขาเข้ากับขาออกได้ เป็นเทคนิคการเรียนรู้ของเครื่องซึ่งสร้างฟังก์ชันจากข้อมูลสอน (Training Data) ข้อมูลสอนประกอบด้วยวัตถุนำเข้า และผลที่ต้องการ ผลจากการเรียนรู้จะเป็นฟังก์ชันที่อาจจะให้ค่าต่อเนื่องหรืออาจจะเรียกวิธีการว่า การถดถอย (Regression) หรือใช้ทำนายประเภทของวัตถุ อีกอย่างเรียกว่า การแบ่งประเภท (Classification) ภารกิจของเครื่องเรียนรู้แบบมีผู้สอนคือการทำนายค่าของฟังก์ชันจากวัตถุนำเข้าที่ถูกต้อง โดยใช้ตัวอย่างในการสอนโดยใช้ข้อมูลนำเข้าจำนวนมาก (Training Set) และผลที่เป็นเป้าหมาย โดยการเรียนรู้ของเครื่อง จะต้องวางนัยทั่วไป (Generalize) จากข้อมูลที่มีอยู่ ไปยังกรณีที่ไม่เคยพบอย่างมีเหตุมีผล โดยการเรียนรู้แบบมีผู้สอนนั้น มีขั้นตอนต่าง ๆ ที่ต้องพิจารณามากมาย ได้แก่ กำหนดชนิดของตัวอย่างสอน ก่อนจะเริ่มทำอย่างอื่น จะต้องตัดสินว่าข้อมูลชนิดใดที่จะใช้เป็นตัวอย่าง เก็บตัวอย่าง ชุดตัวอย่างสอนจะต้องมีลักษณะเป็นตามที่ใช้จริง ดังนั้นชุดข้อมูลตัวอย่างและผลที่สอดคล้องจะต้องถูกจัดเก็บจากผู้เชี่ยวชาญหรือจากการวัด กำหนดวิธีการแทนลักษณะ (Feature) ของข้อมูลเข้า ความถูกต้องของฟังก์ชันจะขึ้นอยู่กับการแทนข้อมูลอย่างมาก โดยทั่วไปวัตถุเข้าจะถูกแปลงเป็นเวกเตอร์ของลักษณะ ใช้อธิบายวัตถุที่ต้องการแบ่งประเภท จำนวนลักษณะจะต้องไม่มากจนเกินไป เพราะจะทำให้เกิดปัญหา Curse of dimensionality เนื่องจากมิติที่กว้างเกินไปจนทำให้มีพื้นที่ว่างมากจนเครื่องเรียนรู้ไม่สามารถวางนัยทั่วไปได้ แต่จำนวนลักษณะก็จะต้องมากพอที่จะทำให้สามารถทำนายผลได้แม่นยำ

3.1 เนอีฟเบย์ (Naïve Bayes) การเรียนรู้แบบเบย์ [5-6] เป็นวิธีการเรียนรู้ที่ใช้หลักการของความน่าจะเป็น ซึ่งมีพื้นฐานมาจากทฤษฎีของเบย์ (Bayes Theorem) เข้ามาช่วยในการเรียนรู้ จุดมุ่งหมาย ก็เพื่อต้องการสร้างโมเดลที่อยู่ในรูปของความน่าจะเป็น จัดเป็นขั้นตอนวิธีในการจำแนกข้อมูล โดยการเรียนรู้ปัญหาที่เกิดขึ้น เพื่อนามาสร้างเงื่อนไขการจำแนกข้อมูลใหม่ หลักการของนาอีฟเบย์ ใช้การคำนวณหาความน่าจะเป็นในการทานายผล เป็นเทคนิคในการแก้ปัญหาแบบจำแนกประเภทที่สามารถคาดการณ์ผลลัพธ์ได้ จะทาการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสาหรับแต่ละความสัมพันธ์ เหมาะกับกรณีของเซตตัวอย่างที่มีจำนวนมากและคุณสมบัติ (Attribute) ของตัวอย่างไม่ขึ้นต่อกัน โดยกาหนดให้ความน่าจะเป็นของข้อมูลเท่ากับสมการ

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \qquad (1)$$

3.2 การวิเคราะห์การถดถอยโลจิสติค [7-8] เป็นการวิเคราะห์ที่มีเป้าหมายเพื่อทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจ โดยอาศัยสมการโลจิสติคที่สร้างขึ้นจากชุดตัวแปรทำนาย ที่เป็นตัวแปรที่มีข้อมูลอยู่ ในระดับช่วงเป็นอย่างน้อย โดยที่ระหว่างตัวแปรทำนายจะต้องมีความสัมพันธ์กันต่ำ การวิเคราะห์การถดถอยโลจิสติค (Logistic Regression Analysis) เป็นเทคนิคการวิเคราะห์สถิติเชิงคุณภาพ (Qualitative Statistical Techniques) ที่แตกต่างไปจากเทคนิคการวิเคราะห์เชิงปริมาณ (Quantitative Techniques) อย่างน้อย ก็เรื่องของข้อมูลที่ตัวแปรตามเป็นตัวแปรเชิงคุณภาพ ซึ่งก็คือ เป็นตัวแปรเชิงกลุ่มนั่นเอง การวิเคราะห์การถดถอยโลจิสิคแบ่งเป็น 2 ประเภท คือ การวิเคราะห์การถดถอยโลจิสติคทวิ (Binary Logistic Regression Analysis) และการวิเคราะห์การถดถอยโลจิสติคพหุกลุ่ม (Multinomial Logistic Regression Analysis) การวิเคราะห์การถดถอยโลจิสติคทั้ง 2 ประเภท แตกต่างกันในด้านตัวแปรตาม โดยที่การวิเคราะห์การถดถอยโลจิสติคทวิใช้กับตัวแปรตามที่แบ่งออกเป็น 2 กลุ่มย่อย (Dichotomous Variable) มี 2 ค่า คือมีค่าเป็น 0 กับ 1 เช่น กลุ่มที่มีเหตุการณ์กับกลุ่มที่ไม่มีเหตุการณ์ ส่วนการวิเคราะห์โลจิสติคแบบพหุกลุ่มใช้กับตัวแปรตามที่มีหลายค่ามากกว่า 2 กลุ่ม (Polytomous Variable) การวิเคราะห์โลจิสติคมีเป้าหมายก็คือ เพื่อทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจ ซึ่งก็คือตัวแปรเกณฑ์ โดยอาศัยสมการโลจิสติคที่สร้างขึ้นจากชุดตัวแปรทำนาย ที่มีข้อมูลเป็นตัวแปรที่มีข้อมูลอยู่ในระดับช่วง (interval scale) เป็นอย่างน้อย หากเป็นข้อมูลเชิงกลุ่มจะต้องแปลงเป็นตัวแปรทวิที่มีค่า 0 กับ 1 ก่อน โดยที่ระหว่างตัวแปรทำนายจะต้องมีความสัมพันธ์กันต่ำ รูปแบบสมการ การวิเคราะห์การถดถอยโลจิสติกสำหรับการวิเคราะห์การถดถอย สมการพยากรณ์ที่ได้จากตัวแบบการวิเคราะห์จะเป็นสมการแสดงความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ (Probability of Event)

ในกรณีที่มี ตัวแปรอิสระเพียง 1 ตัว ตัวแบบการวิเคราะห์ถดถอยโลจิสติกสามารถเขียนได้ดังสมการ ความน่าจะเป็นของการเกิดเหตุการณ์

$$\frac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}} \qquad (2)$$

ความน่าจะเป็นของการไม่เกิดเหตุการณ์

$$\frac{1}{1+e^{b_0+b_1x}} \qquad (3)$$

ถ้ากำหนดให้ $Z = b_0 + b_1 X$ จากสมการข้างต้นจะสามารถเขียนสมการใหม่ได้เป็นความน่าจะเป็นของการเกิดเหตุการณ์

$$\frac{e^z}{1+e^z} \quad (4)$$

ความน่าจะเป็นของการไม่เกิดเหตุการณ์

$$\frac{1}{1+e^z} \quad (5)$$

และในกรณีที่มีตัวแปรอิสระ p ตัว จะได้ว่า

$$Z = b_0 + b_1X_1 + b_2X_2 + \cdots + b_pX_p \quad (6)$$

3.3 ต้นไม้ตัดสินใจ (Decision Tree) เป็นการนำข้อมูลมาสร้างแบบจำลอง [9] มีลักษณะเป็นผังงาน (Flowchart) เหมือนโครงสร้างต้นไม้ เป็นการเรียนรู้แบบมีผู้สอน (Supervised Learning) วิธีการสร้างต้นไม้ตัดสินใจ อันดับแรกจะหาคุณลักษณะที่สำคัญที่สุดมาแบ่งข้อมูลโดยคุณลักษณะนี้จะถูกตั้งให้เป็นโหนดราก จากโหนดรากจะสร้างเส้นทางเชื่อมหรือกิ่งไปยังโหนดลูก โดยจำนวนเส้นทางเชื่อมจะเท่ากับจำนวนค่าที่เป็นไปได้ของคุณลักษณะของโหนดราก ถ้าโหนดลูกเป็นกลุ่มของข้อมูลที่อยู่ในกลุ่มเดียวกันทั้งหมดให้หยุดการสร้างต้นไม้ แต่ถ้าโหนดลูกมีข้อมูลของหลายกลุ่มปะปนกัน ต้องสร้างโหนดลูกเพื่อจำแนกข้อมูลต่อไป โดยวนกลับไปทำขั้นตอนแรก ซ้ำเพื่อเลือกคุณลักษณะที่สำคัญที่สุดมาเป็นตัวแบ่งข้อมูลต่อไป การแบ่งข้อมูลโดยวิธีการกำหนดโครงสร้างต้นไม้ตัดสินใจจะเป็นการเลือกข้อมูลตามลาดับของตัวชี้วัดหรือค่าเกน (Gain) สูงที่สุดเป็นข้อมูลเริ่มต้นและข้อมูลถัดไปที่มีค่าลดหลั่นกันตามลาดับ ตัวอย่างเช่น การพิจารณาจากกลุ่มข้อมูล 2 คลาสคือ P และ N โดยจานวนตัวอย่างในคลาส P คือ p ตัว และจานวนตัวอย่างในคลาส N คือ n ตัว ส่วนค่าของกลุ่มข้อมูลคือค่าคาดคะเนที่กลุ่มตัวอย่างต้องใช้จานวนบิตในการแยกคลาส P และ N โดยนิยามตามสมการ

$$I(p,n) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right) \quad (7)$$

ค่าคาดคะเนของข้อมูล (Entropy) เป็นค่าที่แยกโดยการใช้ลักษณะประจา A ซึ่งกาหนด A คือ ลักษณะประจาที่แบ่ง S ออกเป็น $\{S_1, S_2, , S_v\}$ โดยให้ S1 มีตัวอย่างจากคลาส P จานวน P_1 และตัวอย่างจากคลาส N จำนวน n_1 ดังสมการ

$$E(A) = \sum_{i=1}^{v} \frac{p_1+n_1}{p+n} I(p_1,n_1) \quad (8)$$

ดังนั้นค่าเกนข้อมูล (Data Gain) ที่ได้จากการแยกข้อมูลด้วยลักษณะประจำ A จะได้ดังสมการ

$$Gain(A) = I(p,n) - E(A) \quad (9)$$

3.4 ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) หลักการของวิธีการนี้ [10-11] ใช้เพื่อหาระนาบการตัดสินใจในการแบ่งข้อมูลออกเป็นสองส่วน โดยใช้สมการเส้นตรงเพื่อแบ่งเขตข้อมูล 2 กลุ่มออกจากกัน ซัพพอร์ตเวกเตอร์แมชชีนจัดเป็นการเรียนรู้ของเครื่องประเภทที่ต้องมีตัวอย่างในการเรียนรู้ (Supervised Learning) ประเภทหนึ่งซึ่งมีความสามารถในการคัดแยก (Classification) และการทำนาย (Regression) โดยเอสวีเอ็มมีพื้นฐานจากการคำนวณแบบ Linear Classifier ซึ่งจัดอยู่ในประเภทมุ่งหาผลลัพธ์ที่ดีที่สุดของการเรียนรู้ (Discriminative Training) บนการเรียนรู้จากสถิติของข้อมูล ซึ่งทำงานโดยการหาค่าระยะขอบที่มากที่สุด (Maximum Margin) ของระนาบตัดสินใจ

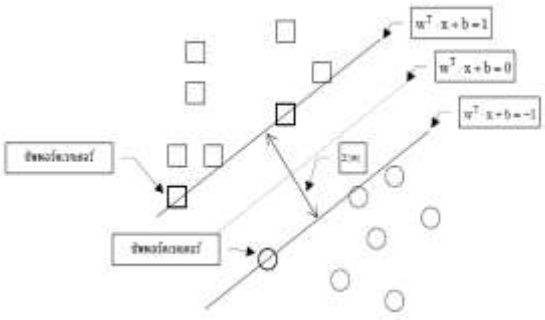
(Decision Hyperplane) ในการแบ่งแยกกลุ่มข้อมูลที่ใช้ฝึกฝนออกจากกัน โดยจะใช้ฟังก์ชั่นแม็ปข้อมูลจาก Input Space ไปยัง Feature Space และสร้างฟังก์ชั่นวัดความคล้ายที่เรียกว่าเคอร์เนลฟังก์ชั่น (Kernel Function) บน Feature Space โดยมีวัตถุประสงค์ที่จะพยายามที่จะทำการลดความผิดพลาดจากการทำนาย (Minimize Error) พร้อมกับเพิ่มระยะแยกแยะให้มากที่สุด (Maximized Margin) ดังนั้นเมื่อข้อมูลที่ใช้ฝึกฝนเป็นข้อมูลที่สามารถแบ่งกลุ่มได้ด้วยเส้นตรงใด ๆ ระยะที่กว้างที่สุดของ Hyper plane ทั้งสองที่ขยายออกไปจนกว่าจะพบจุดข้อมูลของทั้งสองกลุ่มคือ 2/|w| โดยค่า |w| มีค่าน้อยที่สุด ค่าข้อมูลแต่ละ x_i จัดจำแนกอยู่ในกลุ่มใดสามารถพิจารณาได้จากเงื่อนไขดังนี้ ถ้า $W^T \cdot X_i + \mathrm{b} \geq 1$ แสดงว่า x_iเป็นกลุ่มที่ 1และ ถ้า $W^T \cdot X_i + \mathrm{b} \leq -1$ แสดงว่า x_i เป็นกลุ่มที่ 2 ดังนั้นในการคัดแยกจุดข้อมูลใด ๆ ที่นำมาฝึกฝนเพื่อกรองว่าเป็นกลุ่ม 1 สามารถตรวจสอบได้จากสมการ

$$c_i(W \cdot X_i - b) \geq 1 \text{ เมื่อ } 1 \leq i \leq n \quad (10)$$

ดังนั้นสามารถเขียนเป็นรูปสั้นเพื่อหาระยะขอบที่น้อยที่สุด โดยที่สามารถคำนวณการแบ่งกลุ่มได้ดังสมการ

$$Minimize_{w,b}|W| \text{ โดยตรวจสอบ}$$

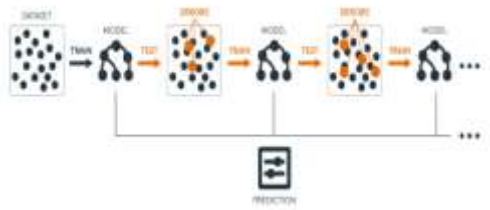
$$c_i(W \cdot X_i - b) \geq 1 \text{ เมื่อ } 1 \leq i \leq n \quad (11)$$



รูปที่ 1 ระยะขอบที่กว้างที่สุดซัพพอร์ตเวกเตอร์

3.5 เอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) บูทติ้ง (Boosting) [12-13] จัดเป็นเทคนิคเพื่อลดความแปรปรวนและเพิ่มความแม่นยำในการทำนายของตัวจำแนกประเภท โดยใช้วิธีลดความอคติ (Bias) และมีแนวคิดที่ให้ตัวเรียนรู้ที่อ่อนแอ (Weak Learner) ชุดหนึ่ง ทำงานร่วมกันจนสามารถพัฒนามาเป็นตัวเรียนรู้ที่เข้มแข็ง (Strong Learner) ได้ การสร้างตัวเรียนรู้ที่อ่อนแอ (Weak Learner) แต่ละตัวสามารถทำได้ โดยการปรับเพิ่มน้ำหนักของการทำนายที่ผิดพลาดให้มากขึ้นในแต่ละรอบ แล้วทำการเรียนรู้ใหม่ ซึ่งจะทำให้โมเดลของตัวจำแนก (Classifier) เปลี่ยนไปโดยให้ความสำคัญกับความผิดพลาดในรอบที่แล้วมากขึ้น เมื่อได้จำนวนตัวเรียนรู้ที่อ่อนแอมากพอแล้ว จึงนำมารวมกันสร้างเป็นตัวเรียนรู้ที่เข้มแข็งต่อไป (Strong Learner) โดยผลรวมของตัวจำแนกประเภท (Aggregating) จะเกิดเป็นตัวจำแนกประเภทใหม่ขึ้นมา เราจะทำแบบนี้ไปเรื่อย ๆ (Recursive) จนได้

โมเดลที่ดีที่สุดจากผลรวมของการจำแนก หลักการสร้างต้นไม้แต่ละต้น จะเป็นแบบเรียงลำดับ (Sequence) โดยข้อมูลนำเข้า (Input) แต่ละของต้นไม้แต่ละต้น จะเป็นเอาท์พุต (Output) จากต้นไม้ก่อนหน้า โดยหลักการคือเอ็กซ์ทรีมกาเดียนบูทติ้ง จะทำการสร้างต้นไม้แต่ละต้น เพื่อลดค่าความผิดพลาด (Error) ที่เกิดจาก ต้นไม้ก่อนหน้า โดยใช้วิธี Gradient Descend แล้วนำผลลัพธ์ที่ได้มารวมกัน ก็จะทำให้ได้ค่าใกล้เคียงกับค่าที่จะทำนาย Y(actual) ซึ่งข้อดีของ เอ็กซ์ทรีมกาเดียนบูทติ้ง คือ ความอคติ (Bias) และความแปรปรวน (Variance) ลดลง เนื่องจากความผิดพลาดก่อนหน้า (Error) ถูกแก้ไขแค่ ความลึกของชั้นต้นไม้ (Tree Depth) แค่หนึ่งชั้นก็เพียงพอที่จะได้ค่าประสิทธิภาพ (Performance) ที่ดีขึ้นมาก เมื่อเทียบกับ (Bagging Tree) และแรนดอมฟอเรส (Random Forest) ที่ต้องเพิ่มชั้นความลึกมากขึ้นมากขึ้น เพื่อให้ประสิทธิภาพที่ใกล้เคียง



รูปที่ 2 เอ็กซ์ทรีมกาเดียนบูทติ้ง

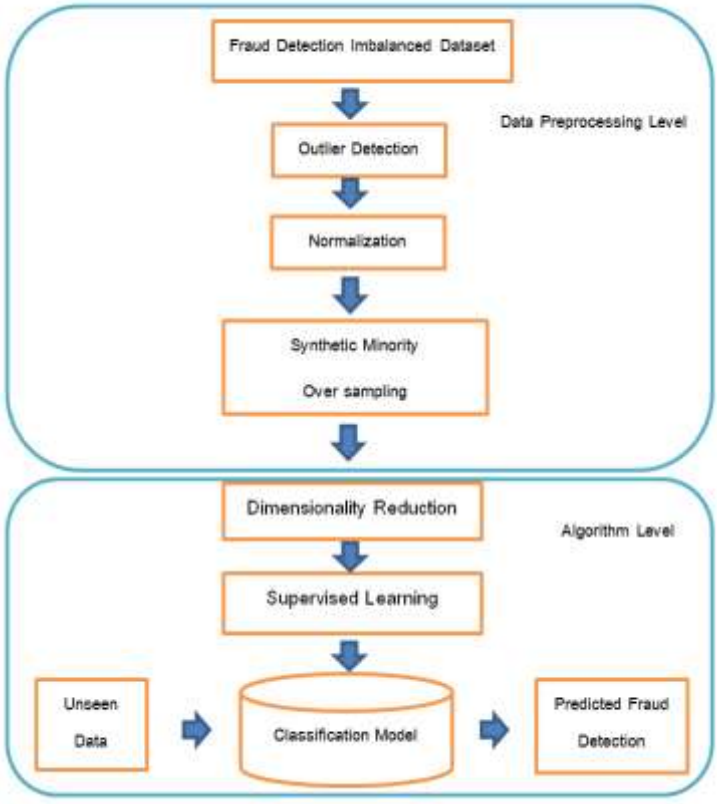
3.6 การลดมิติของข้อมูล (Dimension Reduction) [14] ถือเป็นงานสำคัญในงานการเรียนรู้ของเครื่องและการรู้จำ เนื่องจากเทคโนโลยีในปัจจุบันนี้ มีความสามารถในการเก็บข้อมูลทุก ๆ คุณลักษณะที่เราสามารถจะเก็บได้ ซึ่งนำไปสู่การเก็บข้อมูลขนาดใหญ่ที่มีคุณลักษณะมากเกินกว่าที่จะประมวลผลได้ เทคนิคลดมิติของข้อมูลช่วยลดความซ้ำซ้อน ระหว่างคุณลักษณะและช่วยขจัดความห่างมากในชุดข้อมูล การเลือกฐานหลักถือเป็นปัญหาเปิด ว่าเราจะสามารถลดมิติของข้อมูลเพื่อให้ได้เซตย่อยของข้อมูลที่เหมาะที่สุดอย่างอัตโนมัติได้อย่างไร ในงานการเรียนรู้ของเครื่องจากการศึกษาพบว่าวิธีลดมิติของข้อมูล (Data Reduction) จัดเป็นกระบวนการหนึ่งในขั้นตอนการเตรียมข้อมูล นั่นคือการทำให้ข้อมูลตั้งต้นมีขนาดลดลง ซึ่งงานวิจัยนี้ใช้ ค่าความสัมพันธ์ (Correlation) เพื่อลดมิติของข้อมูล เพราะเป็นการคัดเลือกมิติข้อมูล โดยใช้การคำนวณหาค่าน้ำหนักซึ่งอาจจะเป็นค่าความสัมพันธ์ระหว่างแต่ละตัวแปร การอธิบายความสัมพันธ์ของตัวแปร 2 ตัว โดยใช้ค่าเชิงปริมาณ เพราะบางครั้งการดูแต่ การวากกราฟเพียงอย่างเดียว ก็อธิบายระดับความแตกต่างได้ไม่ละเอียดพอ ค่าความ สัมพันธ์ (Correlation) คือค่าที่ใช้บอกระดับความสัมพันธ์ของแต่ละตัวแปร ที่เป็นคู่

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (12)$$

4. วิธีดำเนินงานวิจัย

โดยแบบจำลองที่นำเสนอในงานวิจัยนี้ ใช้การตรวจสอบข้อมูลก่อนการวิเคราะห์ เพื่อกำจัดข้อมูลค่าสุดโต่งแฝง (Outlier) ออกก่อน จากนั้นทำการแปลงข้อมูล (Data Transformation) ด้วยวิธีการนอร์มอลไลซ์ (Normalization) ทำการแปลงค่าข้อมูลให้เป็นช่วงข้อมูลใหม่ที่กำหนดขึ้นให้มีสเกลที่เท่ากัน แล้วจึงทำการปรับปรุงชุดข้อมูลกลุ่มตัวอย่าง [15] ของการตรวจสอบการทุจริตไม่สมดุลด้วยวิธีการสุ่มตัวอย่าง (Sampling Methods) โดยวิธีการ Over-sampling จะทำการสุ่มข้อมูลในกลุ่มรองเพื่อ

สร้างข้อมูลใหม่ของกลุ่มรองให้มีจำนวนเพิ่มมากขึ้น ให้ใกล้เคียงหรือเท่ากับจำนวนข้อมูลในกลุ่มหลัก จากนั้นทำการลดมิติข้อมูลของกลุ่มตัวอย่างลง โดยพิจารณาคุณลักษณะที่มีค่าสถิติความสัมพันธ์กับกลุ่มเป้าหมายสูง ก่อนส่งเข้าเครื่องจักรการเรียนรู้แบบมีผลเฉลย (Supervised Learning) ซึ่งประกอบด้วย อัลกอริทึม เนอีฟเบย์ (Naïve Bayes) การวิเคราะห์การถดถอยโลจิสติค (Logistic Regression) ต้นไม้ตัดสินใจ (Decision Tree) ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) เอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) โดยอัลกอริทึมทั้งหมดใช้ค่าพารามิเตอร์มาตรฐานมาทำการเรียนรู้ แล้วทำการทดสอบเปรียบเทียบประสิทธิภาพของแบบจำลอง การประเมินความสามารถของแบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่องนั้น เน้นความสามารถในการตัดสินใจหรือ การจำแนกการตรวจสอบการทุจริตที่ถูกต้อง วิธีการทดสอบเพื่อเปรียบเทียบประสิทธิภาพของโมเดล โดยพิจารณาด้านความถูกต้อง (Accuracy) และปัจจัยด้านเวลาในการประมวลผล (Runtime) โดยทำการทดสอบแบบจำลองด้วยวิธี 10-ครอสวาลิเดชั่น (10-Cross Validation) ซึ่งเป็นวิธีที่เป็นมาตรฐานในการทดสอบแบบจำลองด้านการเรียนรู้แบบมีผลเฉลย



รูปที่ 3 แบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุล
โดยใช้การลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง

รูปที่ 4 ผลการทดสอบประสิทธิภาพของโมเดลการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุล โดยใช้การลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง

4. สรุปผลการวิจัย

จากการทดลองทดสอบประสิทธิภาพของโมเดลการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง โดยทำการสุ่มตัวอย่าง (Sampling Methods) โดยการเลือกสมาชิกจากกลุ่มตัวอย่างโดยพยายามทำให้สมาชิกที่เลือกมาเหล่านั้น เป็นตัวแทนที่ดีของกลุ่มตัวอย่างทั้งหมด โดยทำการเลือกแบบสุ่ม (Random) หลังจากนั้นหาค่าสถิติความสัมพันธ์ (Correlation Based Feature Selection) ระหว่างตัวแปรเพื่อใช้ลดมิติของข้อมูล ก่อนส่งเข้าเครื่องจักรการเรียนรู้แบบมีผลเฉลย (Supervised Learning) สามารถสรุปในแต่ละประเด็นได้ว่า อัลกอริทึมที่ให้ประสิทธิภาพการจำแนกการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่องออกมาดีที่สุด เมื่อเปรียบเทียบกับวิธีการอื่น ๆ และเมื่อพิจารณาค่าพารามิเตอร์ที่ส่งผลให้ประสิทธิภาพของโมเดลที่ดีสุด คือ อัลกอริทึม เอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) ให้ค่าความถูกต้อง (Accuracy) สูงที่สุดคือ 98.15 % รองลงมาเป็นอัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) ให้ค่าความถูกต้อง 85.82% ซัพพอร์ทเวกเตอร์แมชชีน

(Support Vector Machine) ให้ค่าความถูกต้อง80.12% อัลกอริทึมการวิเคราะห์การถดถอยโลจิสติค (Logistic Regression) ให้ค่าความถูกต้องเท่ากันที่ 74.57 % และสุดท้ายอัลกอริทึมเนอีฟเบย์ (Naïve Bayes) ให้ค่าความถูกต้องน้อยที่สุด 72.96 % ตามลำดับ

เมื่อทำการเปรียบเทียบเวลาในการสร้างและทดสอบแบบจำลอง การจำแนกการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง ในด้านเวลาเป็นนาที หรือที่เรียกว่าค่า Runtime พบว่าอัลกอริทึมเอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) ใช้เวลา Runtime น้อยที่สุดคือ 15.23 นาที รองลงมาเป็นอัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) ใช้เวลา Runtime 42.27 นาที อัลกอริทึมการวิเคราะห์การถดถอยโลจิสติค (Logistic Regression) ใช้เวลา Runtime 44.19 นาที อัลกอริทึมเนอีฟเบย์ (Naïve Bayes) ใช้เวลา Runtime 44.39 นาที และสุดท้ายอัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) ใช้เวลา Runtime 69.47 นาที ตามลำดับ

จากข้อสรุปผลการวิจัยพบว่า ปัจจัยหลักที่ส่งผลให้ประสิทธิภาพของแบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่องมีประสิทธิภาพสูงนั้น เกิดจากปัจจัยหลักด้านการผสมผสานกระบวนการใช้การตรวจสอบข้อมูลก่อนการวิเคราะห์ เพื่อกำจัดข้อมูลค่าสุดโต่งแฝง (Outlier) ออกก่อน จากนั้นทำการแปลงข้อมูล (Data Transformation) ด้วยวิธีการนอร์มอลไลซ์ (Normalization) แล้วจึงทำการปรับปรุงชุดข้อมูลกลุ่มตัวอย่างของการตรวจสอบการทุจริตไม่สมดุลด้วยวิธีการสุ่มตัวอย่าง (Sampling Methods) โดยวิธีการ Oversampling จะทำการสุ่มข้อมูลในกลุ่มรองเพื่อสร้างข้อมูลใหม่ของกลุ่มรองให้มีจำนวนเพิ่มมากขึ้น ให้ใกล้เคียงหรือเท่ากับจำนวนข้อมูลในกลุ่มหลัก จากนั้นทำการลดมิติข้อมูลของกลุ่มตัวอย่างลง โดยพิจารณาคุณลักษณะที่มีค่าสถิติความสัมพันธ์กับกลุ่มเป้าหมาย ก่อนส่งเข้าเครื่องจักรการเรียนรู้แบบมีผลเฉลย (Supervised Learning) แบบต่าง ๆ ประกอบกับอัลกอริทึมเอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) จัดเป็นเทคนิคเพื่อลดความแปรปรวนและเพิ่มความแม่นยำในการทำนายของตัวจำแนกประเภทโดยใช้วิธีลดความอคติ (Bias) และมีแนวคิดที่ให้ตัวเรียนรู้ที่อ่อนแอ (Weak Learner) ชุดหนึ่ง ทำงานร่วมกันจนสามารถพัฒนามาเป็นตัวเรียนรู้ที่เข้มแข็ง (Strong Learner) ได้ การสร้างตัวเรียนรู้ที่อ่อนแอแต่ละตัว สามารถทำได้โดยการปรับเพิ่มน้ำหนักของการทำนายที่ผิดพลาดให้มากขึ้นในแต่ละรอบแล้วทำการเรียนรู้ใหม่ ซึ่งจะทำให้โมเดลของตัวจำแนก (Classifier) เปลี่ยนไปโดยให้ความสำคัญกับความผิดพลาดในรอบที่แล้วมากขึ้น เมื่อได้จำนวนตัวเรียนรู้ที่อ่อนแอมากพอแล้ว จึงนำมารวมกันสร้างเป็นตัวเรียนรู้ที่เข้มแข็งต่อไป โดยผลรวมของตัวจำแนกประเภท (Aggregating) จะเกิดเป็นตัวจำแนกประเภทใหม่ขึ้นมา เราจะทำแบบนี้ไปเรื่อย ๆ (Recursive) จนได้โมเดลที่ดีที่สุดจากผลรวมของการจำแนก

5. เอกสารอ้างอิง

[1] B. Baesens, V. Van Vlasselaer, W. Verbeke. Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud

Detection. *John Wiley & Sons Publisher*. 2015.

[2] E. Caldeira. et al. , Fraud Analysis and Prevention in e- Commerce Transactions. *Proceedings 9th Latin American Web Congress.* 2014.

[3] Sunder Gee. Fraud and Fraud Detection: A Data Analytics Approach. *John Wiley & Sons Publisher.* 2014.

[4] Maria R. et al. Credit Card Fraud Detection with Unsupervised Algorithms. , *Journal of Advances in Information Technology*, Volume. 7, No. 1. 2016.

[5] Ian H. Witten, Eibe Frank and Mark A. Data mining : Practical machine learning tools and techniques (3rd). *Boston: Morgan Kaufman Publisher*. 2005.

[6] S Panigrahi, A Kundu, S Sural, AK Majumdar. Credit card fraud detection: A fusion Approach using Dempster– Shafer theory and Bayesian learning. *Elsevier Information Fusion. 2009 ; 10(4) : 354-363*

[7] Jon T. S. Quah, Sriganesh Srihari. Real Time Credit Card Fraud Detection using Computational Intelligence. *Proceedings of the International Joint Conference on Neural Networks*, Florida, USA. 2007.

[8] Francisca N. Ogwueleka. Data Mining Application In Credit Card Fraud Detection System. *Journal Of Engineering Science And Technology*, Volume. 6, No. 3. 2011.

[9] Quinlan, J. R. C4. 5 Programs for machine learning. San Francisco: *Morgan Kaufmann Publisher*. 1993.

[10] M. A. H. Farquad, Indranil Bose. Preprocessing unbalanced data using support vector machine. *Journal Decision Support Systems.* 2012 *; 53(1) : 226-233*

[11] M. A. H. Farquad, Vadlamani Ravi, S. Bapi Raju. Churn prediction using comprehensible support vector machine: An analytical CRM application. *Applied Soft Computing*. Volume 19. 2014.

[12] Robert E. Schapire and Yoav Freund. Boosting Foundations and Algorithms. *The MIT Press Publisher*. 2014.

[13] Sharma N. XGBoost. The Extreme Gradient Boosting for Mining *Applications. Technical Report*. 2017.

[14] Chirawichitichai, N. , Sanguansat, P. & Meesad, P. , A Comparative Study on Feature Weight in Thai Document Categorization Framework, 1*0th International Conference on Innovative Internet Community Services (I2CS)* , IICS, pp. 257-266, 2010.

[15] Fraud Detection Dataset UCSD: University of California, San Diego Data Mining Contest 2009.