SPU
SRIPATUM
UNIVERSITY

รายงานการวิจัย

เรื่อง

แบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุล โดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง

FRAUD DETECTION MODEL IN IMBALANCED DATA USING DIMENSION REDUCTION AND MACHINE LEARNING ALGORITHMS

นิเวศ จิระวิชิตชัย

งานวิจัยนี้ ได้รับทุนอุดหนุนการวิจัยจากมหาวิทยาลัยศรีปทุม

ปีการศึกษา 2560

คำนำ

รายงานการวิจัยฉบับนี้ มีวัตถุประสงค์เพื่อให้นำเสนอการสร้างและทดสอบประสิทธิภาพของแบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง เพื่อการป้องกันธุรกรรมการทุจริต (Fraud Detection) เพื่อเป็นแนวทางในการเพิ่มขีดความสามารถขององค์กร ในการลดความเสียหายที่จะเกิดขึ้นจากธุรกรรมที่ทุจริต โดยมุ่งเน้นพัฒนาแบบจำลองการจำแนกหาความสัมพันธ์ของกลุ่มธุรกรรมที่ผิดปกติ โดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่องที่ประสิทธิภาพและความแม่นยำ ในการจำแนกธุรกรรมทุจริตที่กระทำผ่านช่องทางพาณิชย์อิเล็กทรอนิกส์ ผู้วิจัยขอขอบคุณ คณะเทคโนโลยีสารสนเทศ และศูนย์ส่งเสริมและพัฒนางานวิจัย มหาวิทยาลัยศรีปทุม ที่ให้ทุนอุดหนุนสำหรับการทำโครงการวิจัยนี้ และขอขอบคุณ ผู้ช่วยศาสตราจารย์ ดร.นิพัทธ์ จงสวัสดิ์ ที่ปรึกษาโครงการวิจัยที่เสียสละเวลาให้คำปรึกษาและคำแนะนำในการทำโครงการวิจัยในครั้งนี้เป็นอย่างดี

นิเวศ จิระวิชิตชัย

ผู้วิจัย

พฤษภาคม 2562

กิตติกรรมประกาศ

ขอขอบคุณ คณะเทคโนโลยีสารสนเทศ และศูนย์ส่งเสริมและพัฒนางานวิจัย มหาวิทยาลัยศรีปทุม ที่ให้ทุนอุดหนุนสำหรับการทำโครงการวิจัยนี้ และขอขอบคุณ ผู้ช่วยศาสตราจารย์ ดร.นิพัทธ์ จงสวัสดิ์ ที่ปรึกษาโครงการวิจัย ที่เสียสละเวลาให้คำปรึกษาและคำแนะนำในการทำโครงการวิจัยในครั้งนี้เป็นอย่างดี

นิเวศ จิระวิชิตชัย

ผู้วิจัย

พฤษภาคม 2562

หัวข้อวิจัย : แบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง
ผู้วิจัย : นาย นิเวศ จิระวิชิตชัย
หน่วยงาน : คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยศรีปทุม
ปีที่พิมพ์ : พ.ศ. 2562

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง เพื่อการจำแนกธุรกรรมที่มีความผิดปกติ (Fraud Detection) และหาความสัมพันธ์ของกลุ่มธุรกรรมผิดปกติ เพื่อป้องกันความเสียหายที่จะเกิดขึ้นจากธุรกรรมที่ทุจริตในระบบพาณิชย์อิเล็กทรอนิกส์ ผลการทดลองเมื่อวัดประสิทธิภาพแบบจำลองด้วยค่าความถูกต้อง (Accuracy) สรุปได้ว่า แบบจำลองที่ใช้อัลกอริทึม เอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) ให้ค่าความถูกต้องสูงที่สุดคือ 98.15 % รองลงมาเป็นแบบจำลองที่ใช้อัลกอริทึมการเรียนรู้เชิงลึก (Deep Learning) ให้ค่าความถูกต้อง 92.42% จากการทดลองพบว่า แบบจำลองที่พัฒนาขึ้นนั้น ส่งผลให้อัลกอริทึมมีขีดความสามารถจำแนกธุรกรรมที่มีความผิดปกติได้อย่างมีประสิทธิภาพขึ้นอย่างชัดเจน

คำสำคัญ: การตรวจสอบการทุจริต การเรียนรู้ของเครื่อง ข้อมูลที่ไม่สมดุล

Research Title : Fraud Detection Model in Imbalanced Data Using Dimension Reduction and Machine Learning Algorithms
Name of Researcher : Mr.Nivet Chirawichitchai
Name of Institution : Faculty of Information Technology, Sripatum University
Year of Publication : B.E. 2562

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%, followed by a model that uses Deep Learning gives the accuracy 92.42% respectively. From the experiment, it was found that the developed model resulted in the algorithm having the ability to more effectively.

Keywords: fraud detection, machine learning, imbalanced data

สารบัญ

บทที่ **หน้า**

สารบัญ (ต่อ)

บทที่ **หน้า**

สารบัญภาพประกอบ

ภาพประกอบ **หน้า**

บทที่ 1

บทนำ

ความเป็นมาและความสำคัญของปัญหา

จากการขยายตัวด้านการใช้งานระบบพาณิชย์อิเล็กทรอนิกส์ตลอดช่วงระยะเวลาที่ผ่านมา มีแนวโน้มในการใช้งานในด้านธุรกรรมการเงินเพิ่มมากขึ้นอย่างรวดเร็วและแพร่หลาย ส่งผลให้ผู้ใช้งานสามารถที่จะเข้าถึงข้อมูลการเงินส่วนตัว และส่งเสริมในสังคมเกิดการทำธุรกรรมผ่านช่องทางสื่อพาณิชย์อิเล็กทรอนิกส์ระบบอินเตอร์แบงค์กิ้ง ระบบโมบายแบงกิ้ง ได้อย่างสะดวกสบายมากขึ้นและทำให้ได้รับความนิยมอย่างแพร่หลายในวงกว้าง และหนึ่งกรรมวิธีในการทำธุรกรรมออนไลน์ก็คือ การใช้บัตรเครดิตในการชำระค่าสินค้า บัตรเครดิตแต่เดิมนั้นจัดเป็นบริการที่สถาบันทางการเงินต่าง ๆ ออกให้แก่ลูกค้า เพื่อใช้จ่ายแทนเงินสด บัตรเครดิตที่รู้จักกันตัวอย่างเช่น วีซ่า มาสเตอร์การ์ด เจซีบี ยูเนี่ยนเพย์ อเมริกันเอกซ์เพรส ดิสคัฟเวอร์ และ ไดเนอร์สคลับ สามารถใช้ได้ตามจำนวนวงเงินบัตรที่อนุมัติหักออกด้วยค่าสินค้าและบริการที่ใช้จ่ายผ่านบัตร ค่าธรรมเนียม ดอกเบี้ย และหนี้สินคงค้างที่ยังไม่ได้ชำระ ใช้สามารถนำบัตรมาซื้อ สินค้าและบริการได้ตามวงเงินที่ธนาคารอนุมัติ หลังจากผู้รับบริการได้ บัตรเครดิตแล้ว ผู้ขายหรือผู้ให้บริการจะต้องเช็คยอดที่จ่ายกับทางธนาคารก่อนและจะได้รับรหัสอนุมัติจากธนาคาร ในสมัยก่อนจะเป็นเครื่องรูดบัตร ร้านค้าต้องโทรศัพท์ไปที่ธนาคาร แต่ปัจจุบันนี้มีเครื่องรูดบัตรที่จะออนไลน์กับธนาคารเพื่อให้ได้รหัสอนุมัติได้ในทันที จากผู้ขายหรือผู้ให้บริการก็จะนำสลิปไปให้เจ้าของบัตรเซ็นชื่อ เพื่อตรวจสอบว่าเป็นเจ้าของบัตรจริงหรือไม่โดยเทียบกับลายเซนต์ที่เซนต์ไว้ด้านหลังของบัตรเครดิต และเก็บสำเนาไว้เพื่อส่งให้ธนาคารตรวจสอบได้ในภายหลัง

แต่ในปัจจุบันนอกจากบัตรเครดิตจะเป็นที่นิยมในการซื้อสินค้าตามราคาทั่วไปแล้ว ยังนิยมมาใช้ในการซื้อขายผ่านอินเทอร์เน็ตอีกด้วย เมื่อมีการซื้อขายสินค้าผ่านบัตรเครดิต ผู้ใช้บัตรเครดิตจะต้องแสดงความสมยอมว่าการซื้อขายนั้นได้เกิดขึ้นจริง ด้วยการเซ็นชื่อในใบเสร็จ หากเป็นการซื้อขายทางอินเทอร์เน็ต ผู้ใช้อาจจะกรอกหมายเลขบัตรเครดิต และรหัสลับหลังบัตร เพื่อเป็นการแสดงความจำนงในการซื้อขาย ปัจจุบันบัตรเครดิตถือเป็นสินเชื่อประเภทหนึ่งที่ได้รับความนิยมและใช้กันอย่างแพร่หลาย เนื่องจากความสะดวกในการใช้ชำระสินค้า และมีสิทธิประโยชน์ต่างๆที่สอดคล้องการพฤติกรรมของลูกค้า เช่น แลกคะแนนบัตรเครดิตกับของสมนาคุณ การซื้อสินค้าแบบผ่อนชำระทำให้ธุรกิจบัตรเครดิตเติบโตอย่างรวดเร็ว แต่ปัญหาอย่างหนึ่งของธนาคารผู้ออกบัตรคือ

การกระทำทุจริตทางบัตรเครดิต ซึ่งธุรกรรมทุจริตนี้ได้สร้างความเสียหายให้กับธุรกิจธนาคาร โดยเฉพาะอย่างยิ่งธุรกรรมทุจริตที่เกิดการซื้อสินค้าผ่านช่องทางพาณิชย์อิเล็กทรอนิกส์ (E-commerce) ก็จะยิ่งเพิ่มความเสี่ยงจะที่เกิดธุรกรรมทุจริตมากกว่าธุรกรรมที่ต้องแสดงบัตรเครดิต ทั้งยังยากต่อการยืนยันว่าเป็นธุรกรรมทุจริตหรือไม่ ดังนั้นธนาคารผู้ออกเครดิตจึงได้รับความเสียหายเป็นอย่างมาก ทำให้เกิดต้นทุนกับธนาคารผู้ออกบัตร เพราะความเสียหายที่เกิดขึ้นธนาคารผู้ออกบัตรจะต้องเป็นผู้รับภาระเองทั้งหมด (B. Baesens et al., 2015; E.Caldeira. et al, 2014)

ดังนั้นแนวทางอย่างหนึ่งที่ธนาคารผู้ออกบัตรเลือกใช้คือจัดตั้งหน่วยงานเพื่อตรวจสอบธุรกรรมทุจริตโดยเฉพาะ เพื่อดูแลธุรกรรมทุจริตโดยเฉพาะ ซึ่งถ้าลดธุรกรรมทุจริตลงได้จะทำให้ธนาคารลดค่าใช้จ่ายในการจัดการด้านนี้ลดลง และ เพื่อลดความเสียหายที่จะเกิดกับผู้ถือบัตรทั้งนี้ธนาคารพาณิชย์ต่างๆมีข้อมูลรายการใช้จ่ายผ่านบัตรเครดิตของลูกค้า จำนวนมากซึ่งข้อมูลเหล่านั้นเป็นสิ่งที่มีค่ามาก เพราะสามารถบ่งบอกถึงพฤติกรรมของลูกค้าในแต่ละรายแต่ธนาคารพาณิชย์ส่วนใหญ่ยังไม่สามารถนำข้อมูลในส่วนนี้มาใช้ประโยชน์ได้อย่างเต็มประสิทธิภาพ ขณะที่ข้อมูลธุรกรรมบัตรเครดิตมีจำนวนมากขึ้นเรื่อยๆ วิธีการที่ธนาคารพาณิชย์ใช้ในการคัดกรองธุรกรรมทุจริตออกจากธุรกรรมที่ใช้จริงคือ การนำหลักการวิเคราะห์ปัจจัยต่างที่มีผลต่อการทุจริตบัตรเครดิต เช่น สถานที่ซื้อสินค้า จำนวนเงินที่ใช้ซื้อสินค้าแต่ละครั้ง เป็นต้น มาช่วยวิเคราะห์เพื่อตรวจสอบธุรกรรมทุจริตของธนาคารผู้ออกบัตร ซึ่งใช้การเก็บข้อมูลจากธุรกรรมทั้งหมดแล้วนำมาใช้ในการคัดกรองธุรกรรมที่คาดว่าจะเป็นธุรกรรมทุจริต ทำให้ยากต่อการชี้ชัดได้ว่าธุรกรรมใดเป็นธุรกรรมทุจริตหรือธุรกรรมใดเป็นการใช้จริงเพราะธุรกรรมที่เข้าข่ายอาจจะมีจำนวนมาก เนื่องจากผู้ถือบัตรแต่ละรายมีพฤติกรรมการใช้ที่แตกต่างกัน จึงทำให้ธนาคารพาณิชย์มีต้นทุนที่เพิ่มขึ้น เนื่องจากต้องเสียทรัพยากรบุคคลในการตรวจสอบธุรกรรม ดังนั้นบริษัทผู้ให้บริการต่างออกมาตรการต่างๆ เพื่อการป้องกันธุรกรรมทุจริตผ่านช่องทางพาณิชย์อิเล็กทรอนิกส์ (Fraud Detection) (ปิยะ, 2553)

จากความสำคัญดังกล่าว ผู้วิจัยจึงเห็นความสำคัญในการศึกษาแบบจำลองการป้องกันธุรกรรมทุจริต (Fraud Detection) เพื่อเป็นแนวทางในการเพิ่มขีดความสามารถขององค์กรและลดความเสียหายที่จะเกิดขึ้นจากธุรกรรมทุจริต โดยมุ่งเน้นพัฒนาแบบจำลองการจำแนกหาความสัมพันธ์ของกลุ่มธุรกรรมที่ผิดปกติ โดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่องที่ประสิทธิภาพและความแม่นยำในการจำแนกธุรกรรมทุจริตผ่านช่องทางพาณิชย์อิเล็กทรอนิกส์เป็นปัจจัยหลัก ในขณะที่การใช้ทรัพยากรและเวลาเป็นปัจจัยรอง โดยองค์ความรู้ที่ได้จากงานวิจัยนี้ สามารถนำไปประยุกต์ใช้กับการพัฒนาระบบงานการวิเคราะห์พฤติกรรมของลูกค้าเพื่อป้องกันธุรกรรมทุจริต (Fraud Detection) (Sunder Gee, 2014) ตลอดจนสามารถนำ

แบบจำลองที่นำเสนอในงานวิจัยนี้ ตีพิมพ์ลงในวารสารวิชาการระดับชาติหรือนานาชาติ เพื่อตอบตัวชี้วัดด้านการพัฒนางานวิจัยของมหาวิทยาลัย ซึ่งกำหนดโดยสำนักงานคณะกรรมการการอุดมศึกษา (สกอ.) และ สำนักงานรับรองมาตรฐานและประเมินคุณภาพการศึกษา (สมศ.) อีกทางหนึ่งด้วย

วัตถุประสงค์ของการวิจัย

1. เพื่อสร้างแบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง

2. เพื่อทดสอบประสิทธิภาพของแบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง

สมมุติฐานการวิจัย

1. แบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่องที่พัฒนาขึ้นมีประสิทธิภาพอยู่ในระดับดี โดยวัดจากค่าความถูกต้อง (Accuracy) มากกว่า 80 %

2. แบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่องที่พัฒนาขึ้นใช้เวลาทรัพยากรที่เหมาะสมในการประมวลผล

นิยามศัพท์

การเรียนรู้ของเครื่อง (Machine Learning) คือ การทำให้เครื่องเรียนรู้ได้จากข้อมูลตัวอย่างหรือจากสภาพแวดล้อม จุดมุ่งหมายคือการพัฒนาหรือปรับปรุงประสิทธิภาพการทำงานของระบบให้ดีขึ้น เมื่อเรียนรู้แล้วความรู้ที่เรียนได้จะเก็บไว้ในฐานความรู้ด้วยรูปแบบการแทนความรู้อย่างใดอย่างหนึ่ง เช่น กฎ ฟังก์ชัน

ปัญญาประดิษฐ์ (Artificial Intelligence) หมายถึง ความฉลาดเทียมที่สร้างขึ้นให้กับสิ่งที่ไม่มีชีวิต ปัญญาประดิษฐ์เป็นสาขาหนึ่งในด้านวิทยาการคอมพิวเตอร์และวิศวกรรม ซึ่งสาขาปัญญาประดิษฐ์เป็นการเรียนรู้เกี่ยวกับกระบวนการการคิด การกระทำ การให้เหตุผล การปรับตัว หรือการอนุมาน และการทำงานของสมอง

การทำเหมืองข้อมูล (Data Mining) หรือการค้นหาความรู้ในฐานข้อมูล (Knowledge Discovery In Databases - KDD) หมายถึง เทคนิคเพื่อค้นหารูปแบบ (Pattern) จากข้อมูลจำนวน

มหาศาลโดยอัตโนมัติ โดยใช้ขั้นตอนวิธีจากวิชาสถิติ การเรียนรู้ของเครื่อง และการรู้จำแบบ หรือในอีกนิยามหนึ่ง การทำเหมืองข้อมูล คือ กระบวนการที่กระทำกับข้อมูลจำนวนมาก เพื่อค้นหารูปแบบ แนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น โดยอาศัยหลักสถิติ การรู้จำ การเรียนรู้ของเครื่อง และหลักคณิตศาสตร์

การเรียนรู้แบบมีผู้สอน (Supervised Learning) หมายถึง เทคนิคหนึ่งของการเรียนรู้ของเครื่องซึ่งสร้างฟังก์ชันจากข้อมูลสอน (Training Data) ข้อมูลสอนประกอบด้วยวัตถุเข้าและผลที่ต้องการ ผลจากการเรียนรู้จะเป็นฟังก์ชันที่อาจจะให้ค่าต่อเนื่อง (Regression) หรือ ใช้ทำนายประเภทของวัตถุ (Classification) ภารกิจของเครื่องเรียนรู้แบบมีผู้สอนคือการทำนายค่าของฟังก์ชันจากวัตถุเข้าที่ถูกต้องโดยใช้ตัวอย่างสอนจำนวนน้อย โดยเครื่องเรียนรู้จะต้องวางนัยทั่วไป (generalize) จากข้อมูลที่มีอยู่ไปยังกรณีที่ไม่เคยพบอย่างมีเหตุผล

การจำแนกประเภทข้อมูล (Classification) หมายถึง การปัญหาพื้นฐานของการเรียนรู้แบบมีผู้สอน โดยปัญหาคือการทำนายประเภทของวัตถุจากคุณสมบัติต่าง ๆ ของวัตถุ ซึ่งการเรียนรู้แบบมีผู้สอนจะสร้างฟังก์ชันเชื่อมโยงระหว่างคุณสมบัติของวัตถุกับประเภทของวัตถุจากตัวอย่างสอนแล้วจึงใช้ฟังก์ชันนี้ทำนายประเภทของวัตถุที่ไม่เคยพบ เครื่องมือหรือขั้นตอนวิธีที่ใช้สำหรับการแบ่งประเภทข้อมูลเช่น โครงข่ายประสาทเทียม ต้นไม้ตัดสินใจ ซัพพอร์ทเวกเตอร์แมชชีน เนอีฟเบย์ เป็นต้น

การแบ่งกลุ่มข้อมูล (Data Clustering) หมายถึง วิธีการวิเคราะห์ข้อมูล ซึ่งใช้ในการเรียนรู้ของเครื่อง การทำเหมืองข้อมูล โดยจะแบ่งชุดข้อมูลออกเป็นกลุ่ม (Cluster) นำข้อมูลที่มีคุณลักษณะเหมือนกันหรือคล้ายกันจัดไว้ในกลุ่มเดียวกัน ขั้นตอนวิธีที่ใช้ในการแบ่งกลุ่มจะอาศัยความเหมือน (Similarity) หรือ ความใกล้ชิด (Proximity) โดยคำนวณจากการวัดระยะระหว่างเวกเตอร์ของข้อมูลเข้า โดยใช้การวัดระยะแบบต่าง ๆ เช่น การวัดระยะแบบยูคลิด (Euclidean Distance)

การสกัดคุณลักษณะ (Feature Extraction) หมายถึง การดึงคุณลักษณะ (Feature) ของเอกสารออกมา ซึ่งการดึงคุณลักษณะออกมานั้นพบว่า ส่วนใหญ่จะใช้คำ เป็นตัวแทนคุณลักษณะของเอกสาร ในภาษาไทยนอกจากการใช้คำเดี่ยวแล้ว ยังสามารถใช้ วลี ประโยค เป็นตัวแทนคุณลักษณะของเอกสารได้เช่นกัน

การลดมิติข้อมูล (Dimension Reduction) หมายถึง กระบวนการลดขนาดของคุณลักษณะที่จะเรียนรู้ โดยมีขั้นตอนวิธีการเลือกคุณลักษณะที่ดีและสัมพันธ์กับกลุ่มเป้าหมาย และขจัดคุณลักษณะที่ไม่ดีออกไป

ขอบเขตการวิจัย

1. ขอบเขตด้านข้อมูลทดสอบ แบบจำลองการตรวจสอบการทุจริต ที่นำมาใช้ในงานวิจัยนี้ ใช้ฐานข้อมูลการทำธุรกรรมผ่านช่องทางพาณิชย์อิเล็กทรอนิกส์ขนาดใหญ่ ซึ่งเป็นฐานข้อมูลมาตรฐานของแล็บวิจัยมหาวิทยาลัยแคลิฟอร์เนีย ซานดิเอโก (University of California San Diego :UCSD) Transaction Fraud Detection Dataset ที่ใช้ในการแข่งขันจาก UCSD Data Mining Contest

2. ขอบเขตด้านเครื่องมือทดลอง เครื่องคอมพิวเตอร์ หน่วยประมวลผลความเร็วสูง CPU Quad Core Processor 3.0 GHz ขึ้นไป หน่วยความจำ 16 GB ระบบปฏิบัติการวินโดวส์ Windows 64-bit โปรแกรมจาวา Java Development Kit (JDK) 11 โปรแกรม RStudio 64-bit และ โปรแกรม RapidMiner 64-bit

3. ขอบเขตด้านการทดสอบระบบ งานวิจัยนี้มุ่งเน้นแบบจำลองการตรวจสอบการทุจริต โดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่องที่พัฒนาขึ้น มุ่งเน้นประสิทธิภาพในการจำแนกด้านความถูกต้อง ความแม่นยำ โดยใช้วิธีการประเมินความสามารถของแบบจำลอง โดยวัดที่ประสิทธิผลของการจำแนกหมวดหมู่ตามแนวคิดทางด้านการค้นคืนสารสนเทศ ซึ่งก็คือการวัดค่าความถูกต้อง (Accuracy) ในการจำแนกประเภทการทุจริต

ประโยชน์ของการวิจัย

1. ได้แบบจำลองการตรวจสอบการทุจริตโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่องที่มีประสิทธิภาพ ทั้งในด้านความถูกต้องในการจำแนก (Accuracy) ที่ใช้ทรัพยากรของระบบและหน่วยความจำอย่างเหมาะสม ซึ่งสามารถนำแบบจำลองนี้ไปประยุกต์ใช้กับงานพัฒนาระบบสารสนเทศด้าน Big Data Fraud Detection Analytics การตรวจสอบการทุจริตการใช้บัตรเครดิต และการวิเคราะห์พฤติกรรมการซื้อขาย (Transaction) ผู้บริโภคแบบออนไลน์ได้

2. ได้รับตีพิมพ์ลงในวารสารวิชาการระดับชาติหรือนานาชาติ เพื่อตอบตัวชี้วัดด้านการพัฒนางานวิจัยของมหาวิทยาลัย ซึ่งกำหนดโดยสำนักงานคณะกรรมการการอุดมศึกษา (สกอ.) และสำนักงานรับรองมาตรฐานและประเมินคุณภาพการศึกษา(สมศ.)

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้มีวัตถุประสงค์ที่จะศึกษาออกแบบขั้นตอนวิธีการเพื่อสร้างแบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง ผู้วิจัยได้ศึกษาค้นคว้าข้อมูลที่เกี่ยวข้องเพื่อให้ได้ข้อมูลสำหรับนำมาพัฒนางานวิจัย โดยรายละเอียดหัวข้อด้านต่าง ๆ ดังนี้

เหมืองข้อมูล

การทำเหมืองข้อมูล (Data Mining) หรือที่เรียกว่าการค้นหาองค์ความรู้ในฐานข้อมูล (Knowledge Discovery In Databases) เป็นเทคนิคเพื่อค้นหารูปแบบจากชุดข้อมูลขนาดใหญ่โดยอัตโนมัติ จัดเป็นขั้นตอนการของการดึงเอาองค์ความรู้ออกมาจากข้อมูลขนาดใหญ่ โดยใช้หลักวิชาวิธีทางคณิตศาสตร์ หลักการทางสถิติ และ การจดจำรูปแบบ หรือในอีกนิยามหนึ่งสามารถกล่าวได้ว่า การทำเหมืองข้อมูล คือ กระบวนการจัดการข้อมูลขนาดใหญ่ เพื่อค้นหา รูปแบบ แนวทาง และกฎความสัมพันธ์ที่ซ่อนอยู่ในฐานข้อมูลขนาดใหญ่นั้น โดยอาศัยหลักการทางสถิติ การรู้จำ การเรียนรู้ของเครื่อง การทำเหมืองข้อมูลเปรียบเสมือนวิวัฒนาการหนึ่งในการจัดเก็บและตีความหมายข้อมูลจากเดิม ที่มีการจัดเก็บข้อมูลแบบเดิมมาสู่การจัดเก็บในรูปฐานข้อมูลที่สามารถดึงเอาข้อมูลสารสนเทศออกมาใช้ประโยชน์ได้อย่างมีประสิทธิภาพ ตลอดจนถึงการขุดเหมืองข้อมูลที่สามารถค้นพบความรู้ที่ซ่อนอยู่ในชุดข้อมูลขนาดใหญ่ มีผู้ให้นิยามของการทำเหมืองข้อมูลไว้หลากหลาย เช่น การทำเหมืองข้อมูล คือ กระบวนการคัดเลือกและสำรวจข้อมูล ตลอดจนเป็นการสร้างแบบจำลองของข้อมูลเพื่อค้นหารูปแบบและค้นหาความสัมพันธ์จากข้อมูลจำนวนมากเพื่อให้ได้ผลลัพธ์ที่เป็นประโยชน์ อีกนัยหนึ่ง การทำเหมืองข้อข้อมูล คือ การวิเคราะห์ข้อมูลเพื่อหาความสัมพันธ์และสรุปผลจากข้อมูลจำนวนมาก เพื่อทำความเข้าใจและให้เกิดประโยชน์ต่อเจ้าของธุรกิจ สามารถสรุปได้ว่าการทำเหมืองข้อมูล เป็นการนำข้อมูล มาวิเคราะห์ เพื่อให้ได้ความรู้ใหม่ออกมาเพื่อนำมาช่วยในการตัดสินใจ (Witten, et al., 2005; Han and Kamber, 2006) เราสามารถสรุปขั้นตอนของการค้นหาความรู้ใหม่จากกระบวนการทำเหมืองข้อมูลได้ดังนี้

1. เรียนรู้และศึกษาเกี่ยวกับฐานข้อมูลและโปรแกรมที่จะใช้ในการทำเหมืองข้อมูล (Data Preprocessing)

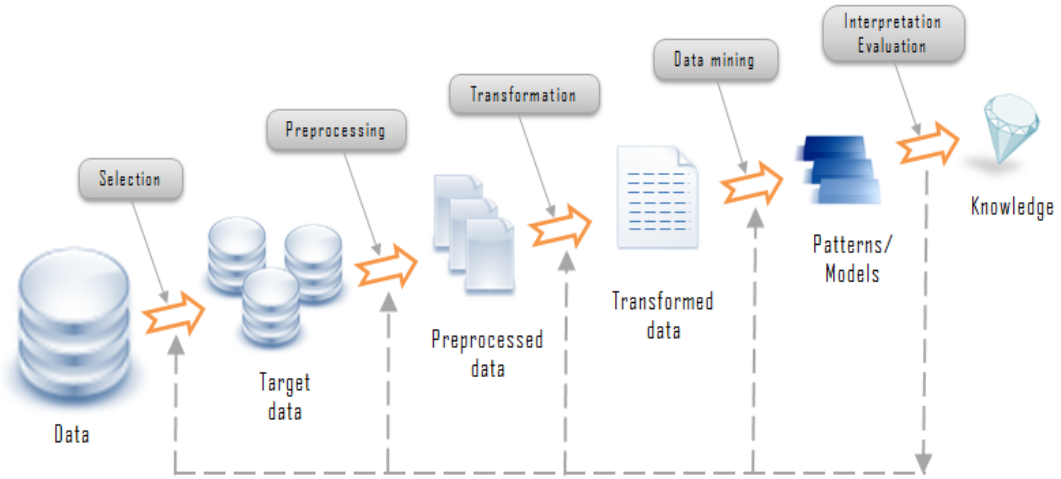
2. การกรองข้อมูลและประมวลผล (Data Cleaning and Integration) เป็นขั้นตอนสำหรับการคัดข้อมูลที่ไม่เกี่ยวข้องออกไป เพราะข้อมูลที่เก็บรวมรวมมามีเป็นจำนวนมาก จึงต้องนำมากรองเพื่อเลือกข้อมูลที่ตรงประเด็นเพราะบางข้อมูลอาจจะไม่เป็นประโยชน์กับเรา ตลอดจนเป็นขั้นตอนการรวมข้อมูลที่มีหลายแหล่งให้เป็นข้อมูลชุดเดียวกัน ซึ่งขั้นตอนนี้เป็นขั้นตอนที่เราจะได้มาซึ่งคุณภาพของข้อมูล ที่พร้อมจะนำไปวิเคราะห์

3. คัดเลือกข้อมูล (Data Selection) เป็นการระบุถึงแหล่งข้อมูลที่จะนำมาทำเหมืองข้อมูล รวมถึงการนำข้อมูลที่ต้องการออกจากฐานข้อมูล เพื่อสร้างกลุ่มข้อมูลสำหรับพิจารณาในเบื้องต้น และทำการแปลงภาพแบบข้อมูล (Data Transformation) ลดภาพและจัดข้อมูลให้อยู่ในภาพแบบเดียวกันมีรูปแบบ (Format) ที่เป็นมาตรฐานและเหมาะสมที่จะนำไปใช้กับอัลกอริทึมและแบบจำลองที่ใช้ทำเหมืองข้อมูล

4. เลือกรูปแบบของการทำเหมืองข้อมูล (Data Mining) เช่น Summarization, Classification, Regression, Association และ Clustering เป็นต้น และเลือกอัลกอริทึมที่เหมาะสมกับลักษณะของงาน จัดเป็นขั้นตอนการค้นหารูปแบบที่เป็นประโยชน์จากข้อมูลที่มีอยู่

5. ขั้นตอนการประเมินรูปแบบที่ได้จากการทำเหมืองข้อมูล (Pattern Evaluation) และ เป็นขั้นตอนการนำเสนอความรู้ที่ค้นพบ โดยใช้เทคนิคในการนำเสนอเพื่อให้เข้าใจ (Knowledge Representation) ในขั้นตอนนี้จะเป็นการวิเคราะห์ผลลัพธ์ที่ได้และแปลความหมาย และประเมินผลว่าผลลัพธ์นั้นเหมาะสมหรือตรงวัตถุประสงค์หรือไม่

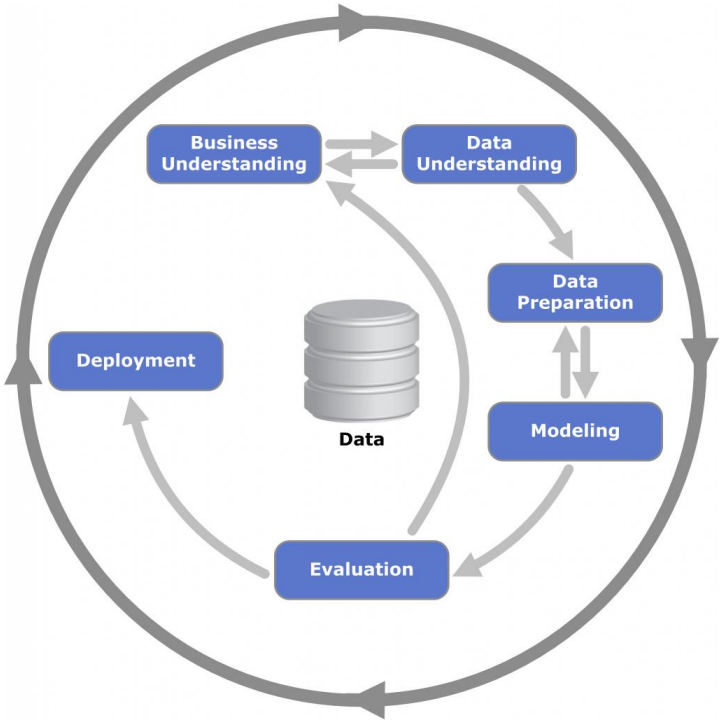
6. ใช้องค์ความรู้ที่ค้นพบขึ้น (Use of Discovered Knowledge) ซึ่งสามารถแสดงขั้นตอนในการทำเหมืองข้อมูล ระบุรายละเอียดขั้นตอนทั้งหมดเป็นแผนภาพได้ดังภาพ



ภาพประกอบ 1 ขั้นตอนกระบวนการทำเหมืองข้อมูล

กระบวนการการทำงานแบบ CRISP-DM

ในการวิเคราะห์ข้อมูล ด้วยเทคนิคการทำเหมืองข้อมูล มีกระบวนการมาตรฐานที่เรียกว่า “Cross-Industry Standard Process for Data Mining” หรือชื่อย่อ “CRISP-DM” ได้ถูกจัดคิดค้นเมื่อปลายปี 1996 แบบจำลองกระบวนการที่ใช้ในการเตรียมข้อมูล ทาง Data Mining จะมองตลอด อายุของโครงการซึ่งประกอบด้วยขั้นตอนต่างๆ ของโครงการ ตามลำดับงาน และความสัมพันธ์ระหว่างงาน ในการอธิบายในแต่ละระดับจะเป็นการยากที่จะบอกได้ถึงทุกความสัมพันธ์ ระหว่างงานแต่ละงานทั้งการทำเหมืองข้อมูลขึ้นกับวัตถุประสงค์ที่ตั้งไว้ด้วยกระบวนการในการทำเหมืองข้อมูล (Data Mining) ที่เป็นมาตรฐาน (Shearer, 2000) CRISP-DM มีขั้นตอนดังนี้



ภาพประกอบ 2 กระบวนการ CRISP-DM

1. ขั้นตอนเข้าธุรกิจ (Business Understanding) เป็นขั้นตอนแรกในกระบวนการ CRISP-DM ซึ่งเน้นไปที่การเข้าใจปัญหาและแปลงปัญหาที่ได้ให้อยู่ในรูปโจทย์ของการวิเคราะห์ข้อมูลทางการทำเหมืองข้อมูล พร้อมทั้งวางแผนในการดำเนินการคร่าวๆ ตัวอย่างการนำเทคนิคการทำเหมืองข้อมูล ไปใช้ในการวิเคราะห์ด้านต่างๆ

2. ขั้นตอนทำความเข้าใจข้อมูล (Data Understanding) ขั้นตอนนี้เริ่มจากการเก็บรวบรวมข้อมูล หลังจากนั้นจะเป็นการตรวจสอบข้อมูลที่ได้ทำการรวบรวมมาได้เพื่อดูความถูกต้องของข้อมูล และพิจารณาว่าจะใช้ข้อมูลทั้งหมดหรือจำเป็นต้องเลือกข้อมูลบางส่วนมาใช้ในการวิเคราะห์

3. ขั้นตอนเตรียมข้อมูล (Data Preparation) ขั้นตอนนี้เป็นขั้นตอนที่ทำการแปลงข้อมูลที่ได้ทำการเก็บรวบรวมมา (Raw Data) ให้กลายเป็นข้อมูลที่สามารถนำไปวิเคราะห์ในขั้นถัดไปได้ โดยการแปลงข้อมูลนี้อาจจะต้องมีการทำข้อมูลให้ถูกต้อง (Data Cleaning) เช่น การแปลงข้อมูลให้อยู่ในช่วง (Scale) เดียวกัน หรือการเติมข้อมูลที่ขาดหายไป เป็นต้น โดยขั้นตอนนี้จะเป็นขั้นตอนที่ใช้เวลามากที่สุดของกระบวนการ CRISP-DM

4. ขั้นตอนจัดทำแบบจำลอง (Modeling) ขั้นตอนนี้จะเป็นขั้นตอนการวิเคราะห์ข้อมูลด้วยเทคนิคทางการทำเหมืองข้อมูลที่ได้แนะนำไปแล้ว เช่น การจำแนกประเภทข้อมูล หรือ การแบ่งกลุ่มข้อมูล ซึ่งในขั้นตอนนี้หลายเทคนิคจะถูกนำมาใช้เพื่อให้ได้คำตอบที่ดีที่สุด ดังนั้นในบางครั้งอาจจะต้องมีการย้อนกลับไปที่ขั้นตอนที่ 3 Data Preparation เพื่อแปลงข้อมูลบางส่วนให้เหมาะสมกับแต่ละเทคนิคด้วย ตัวอย่างเทคนิคในการวิเคราะห์ข้อมูลต่างๆ เช่น การจำแนกประเภทข้อมูล (Classification) การแบ่งกลุ่มข้อมูล (Clustering) หรือ การหากฎความสัมพันธ์ (Association Rules)

5. ขั้นตอนประเมินผล (Evaluation) ในขั้นตอนนี้เราจะได้ผลการวิเคราะห์ข้อมูลด้วยเทคนิคการทำเหมืองข้อมูลแล้วแต่ก่อนที่จะนำผลลัพธ์ที่ได้ไปใช้งานต่อไปก็จะต้องมีการวัดประสิทธิภาพของผลลัพธ์ที่ได้ว่าตรงกับวัตถุประสงค์ที่ได้ตั้งไว้ในขั้นตอนแรก หรือ มีความน่าเชื่อถือมากน้อยเพียงใด ซึ่งอาจจะย้อนกลับไปยังขั้นตอนก่อนหน้าเพื่อเปลี่ยนแปลงแก้ไขเพื่อให้ได้ผลลัพธ์ตามที่ต้องการได้ สำหรับการสร้างโมเดลด้วยเทคนิค Classification มีการทดสอบประสิทธิภาพของโมเดลอยู่ 3 แบบใหญ่ คือ วิธี Self-Consitency Test วิธี Split Test หรือวิธี Cross-Validation Test ในขั้นตอนนี้จะทำการประเมินแต่ละโมเดลด้วยว่ามีส่วนดีส่วนด้อยอย่างไร และควรเลือกใช้โมเดลใด การทำงานในส่วนนี้ต้องอาศัยทักษะในการวิเคราะห์ข้อมูล และธุรกิจ เพื่อช่วยให้การวิเคราะห์ทำได้สะดวกและรวดเร็วขึ้น

6. ขั้นตอนการนำไปใช้งาน (Deployment) ในกระบวนการทำงานของ CRISP-DM นั้นไม่ได้หยุดเพียงแค่ผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูลด้วยเทคนิคทางการทำเหมืองข้อมูลเท่านั้น แม้ว่าผลลัพธ์ที่ได้จะแสดงถึงองค์ความรู้ที่มีประโยชน์ แต่จะต้องนำองค์ความรู้ที่ได้เหล่านี้ไปใช้ได้

จริงในองค์กรหรือบริษัท ตัวอย่างเช่น การสร้างรายงานเพื่อให้ผู้บริหารหรือนักการตลาดเข้าใจได้ง่ายและสามารถนำไปออกโปรโมชั่นได้ เป็นต้น

การเรียนรู้ของเครื่อง (Machine Learning)

จัดเป็นสาขาหนึ่งของปัญญาประดิษฐ์ ที่พัฒนามาจากการศึกษาการรู้จำรูปแบบ เกี่ยวข้องกับทำนายข้อมูล อัลกอริทึมจะทำงานโดยอาศัยโมเดลที่ศึกษาเรียนรู้และสร้างมาจากชุดข้อมูลตัวอย่างที่มีอยู่จำนวนมาก เพื่อการสร้างโมเดลการทำนายหรือตัดสินใจในภายหลัง ซึ่งมันจะแทนที่จะทำงานตามลำดับที่คล้ายกับชุดคำสั่งทางโปรแกรมคอมพิวเตอร์ การเรียนรู้ของเครื่อง มีส่วนเกี่ยวข้องอย่างมากกับ วิชาคณิตศาสตร์ หลักการทางสถิติ และความน่าจะเป็น เนื่องจากทั้งสองศาตร์สามารถทำการวิเคราะห์ข้อมูลที่มีอยู่จำนวนมากเพื่อการทำนายเช่นเดียวกัน นอกจากนี้การเรียนรู้ของเครื่องยังมีความสัมพันธ์กับการหาค่าเหมาะที่สุด (Optimization Techniques)ในทางคณิตศาสตร์ ที่ในแง่ของหลักวิธีการ และการประยุกต์ใช้ การเรียนรู้ของเครื่องสามารถนำไปประยุกต์ใช้งานได้หลากหลายมากมาย ไม่ว่าจะเป็นการจำแนกภาพ (Image Classification) และการจดจำวัตถุ (Object Recognition) การใช้ค้นหามะเร็งผ่านเครื่อง MRIs การกรองอีเมล์ขยะ การรู้จำตัวอักษร เครื่องมือค้นหารูปภาพ และคอมพิวเตอร์วิทัศน์ เป็นต้น โดยศาสตร์การเรียนรู้ของเครื่องสามารถแบ่งโดยกว้างๆได้เป็น 3 ประเภท ตามประเภทของข้อมูลฝึกสอนได้ดังนี้ (Witten, et al., 2005; Han and Kamber, 2006) คือ

1. การเรียนรู้แบบมีผู้สอน (Supervised Learning) กล่าวคือมีข้อมูลตัวอย่างและผลลัพธ์ที่ผู้สอนต้องการถูกป้อนเข้าสู่คอมพิวเตอร์ เป้าหมายคือการสร้างกฎทั่วไปที่สามารถเชื่อมโยงข้อมูลขาเข้ากับขาออกได้ เป็นเทคนิคการเรียนรู้ของเครื่องซึ่งสร้างฟังก์ชันจากข้อมูลสอน (Training Data) ข้อมูลสอนประกอบด้วย วัตถุนำเข้า และผลที่ต้องการ ผลจากการเรียนรู้จะเป็นฟังก์ชันที่อาจจะให้ค่าต่อเนื่องหรืออาจจะเรียกวิธีการว่า การถดถอย (Regression) หรือ ใช้ทำนายประเภทของวัตถุ อีกอย่างเรียกว่า การแบ่งประเภท (Classification) ภารกิจของเครื่องเรียนรู้แบบมีผู้สอนคือการทำนายค่าของฟังก์ชันจากวัตถุนำเข้าที่ถูกต้อง โดยใช้ตัวอย่างในการสอนโดยใช้ข้อมูลนำเข้าจำนวนมาก (Training Set) และผลที่เป็นเป้าหมาย โดยการเรียนรู้ของเครื่อง จะต้องวางนัยทั่วไป (Generalize) จากข้อมูลที่มีอยู่ ไปยังกรณีที่ไม่เคยพบอย่างมีเหตุมีผล โดยการเรียนรู้แบบมีผู้สอนนั้น มีขั้นตอนต่าง ๆ ที่ต้องพิจารณามากมาย ได้แก่ กำหนดชนิดของตัวอย่างสอน ก่อนจะเริ่มทำอย่างอื่น จะต้องตัดสินว่าข้อมูลชนิดใดที่จะใช้เป็นตัวอย่าง เช่นในกรณีการรู้จักลายมือ ตัวอย่างอาจจะเป็นตัวอักษรตัวเดียว คำ หรือบรรทัด เก็บตัวอย่าง ชุดตัวอย่างสอนจะต้องมีลักษณะเป็นตามที่ใช้จริง ดังนั้นชุด

ข้อมูลตัวอย่างและผลที่สอดคล้องจะต้องถูกจัดเก็บจากผู้เชี่ยวชาญหรือจากการวัดหรือแปลงค่า กำหนดวิธีการแทนลักษณะ (Feature) ของข้อมูลนำเข้า ความถูกต้องของฟังก์ชันจะขึ้นอยู่กับการแทนข้อมูลเป็นอย่างมาก โดยทั่วไปข้อมูลนำเข้าจะถูกแปลงเป็นเวกเตอร์ที่แทนที่คุณลักษณะ และใช้อธิบายวัตถุที่ต้องการจำแนกประเภท จำนวนคุณลักษณะจะต้องไม่มากจนเกินไป เพราะหากมีมากจนเกินไปจะทำให้เกิดปัญหาปัญหาของมิติข้อมูล (Curse of Dimensionality) เนื่องจากมิติข้อมูลที่กว้างเกินไป จะทำให้มีพื้นที่ว่างจำนวนมากจน การเรียนรู้ของเครื่องไม่สามารถวางนัยทั่วไปของสิ่งที่เรียนรู้ได้ ในทางเดียวกันจำนวนคุณลักษณะก็จะต้องมากพอที่จะทำให้สามารถเรียนรู้เพื่อที่จะทำนายผลได้อย่างถูกต้อง

2. การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) เป็นเทคนิคหนึ่งของการเรียนรู้ของเครื่อง โดยการสร้างโมเดลที่เหมาะสมกับข้อมูล การเรียนรู้แบบนี้แตกต่างจากการเรียนรู้แบบมีผู้สอน จัดเป็นเทคนิคการเรียนรู้อีกรูปแบบหนึ่งของการเรียนรู้ของเครื่อง โดยการใช้ชุดข้อมูลที่มีทั้งหมดในการให้เครื่องทำการเรียนรู้ โดยไม่จำเป็นต้องระบุผลลัพธ์ที่ต้องการ หรือ ชนิดประเภทข้อมูลไว้ก่อนที่จะให้เครื่องทำการเรียนรู้ การเรียนรู้แบบนี้จะทำหน้าที่แยกผลที่ต้องการ หรือประเภทให้เองจากชุดข้อมูลที่เราใช้ในการเรียนรู้ การเรียนรู้แบบนี้จะพิจารณาวัตถุเป็นเซตของตัวแปรแล้วจึงสร้างโมเดลความหนาแน่นร่วมของชุดข้อมูลเพื่อพิจารณา การเรียนรู้แบบไม่มีผู้สอนนรนั้นสามารถนำไปใช้ร่วมกับทฤษฎีความน่าจะเป็นแบบเบย์ เพื่อหาความน่าจะเป็นแบบมีเงื่อนไขของตัวแปรสุ่มโดยกำหนดตัวแปรที่เกี่ยวข้องให้เรียนรู้ การเรียนรู้แบบไม่มีผู้สอนในอีกรูปแบบหนึ่งก็คือการแบ่งกลุ่มข้อมูล (Clustering) นั่นเอง เทคนิคการเรียนรู้แบบไม่มีผู้สอนนี้ในเมื่อไม่มีผลเฉลยที่บอกว่าข้อมูลชนิดนั้นคืออะไรการเรียนรู้แบบไม่มีผู้สอนจะจัดข้อมูลนำเข้า (Input) จัดเป็นกลุ่ม (Cluster) พิจารณาจากพื้นฐานของความเหมือนกัน (Similarities) และความแตกต่าง (Differences)กัน ระหว่างรูปแบบของข้อมูลนำเข้า (Input Patterns) ตัวอย่างเช่น การหารูปแบบโครงสร้างที่ซ่อนอยู่ในชุดข้อมูลเป็นต้น

3. การเรียนรู้แบบเสริมกำลัง (Reinforcement Learning) จัดเป็นอีกเทคนิคหนึ่งของการเรียนรู้ก็คือจะเรียนรู้จากสิ่งแวดล้อมรอบตัวเอง นั่นก็คือเรียนรู้จากธรรมชาติรอบตัวที่มีอยู่ในชีวิตจริงนำมาดัดแปลงให้คอมพิวเตอร์ ตัวอย่างเช่น การเล่นหมากรุก หมากล้อม เราจะต้องมีการทำนายล่วงหน้าว่าจะสามารถเกิดอะไรขึ้นได้บ้าง ซึ่งการเดินแต่ละครั้งอาจจะไม่เป็นผลดีต่อครั้งนั้นแต่อาจมีผลดีในครั้งต่อจากนั้นก็ได้ หรือกล่าวอีกนัยหนึ่งคือ การเรียนรู้แบบเสริมกำลังจะพิจารณาว่า เอเยนต์ ควรจะมีการกระทำใดในสิ่งแวดล้อมใดเพื่อที่จะได้รางวัลสูงสุด อัลกอริทึมของการเรียนรู้แบบเสริมกำลังนี้ พยายามจะหานโยบาย ที่เชื่อมโยง สถานะของโลกเข้ากับการกระทำที่เอเยนต์ควร

จะทำในสถานะนั้นๆ การเรียนรู้แบบเสริมกำลังนี้มีความแตกต่างไปจากการเรียนรู้แบบมีผู้สอน ตรงที่ว่า คอมพิวเตอร์จะไม่รู้เลยว่าอะไรถูกอะไรผิด กล่าวคือ ไม่มีการบอกอย่างชัดเจนว่าการกระทำใดยังไม่ดี แต่อาศัยการเรียนรู้สะสมองคืความรู้เสริมทบต้นไปเรื่อยๆ

ศาสตร์ด้านการเรียนรู้ของเครื่องมีการเติบโตไปพร้อมๆกับศาสตร์ด้านปัญญาประดิษฐ์ ในความจริงนั้น การเรียนรู้ของเครื่องมีมาตั้งแต่ยุคแรกๆของปัญญาประดิษฐ์ นักวิทยาศาสตร์หลายคนสนใจการสร้างเครื่องจักรที่สามารถเรียนรู้จากข้อมูลได้ จึงเริ่มทดลองวิธีการหลายๆอย่าง ที่เด่นชัดสุดคือ โครงข่ายประสาทเทียม และในเวลาต่อมา ได้มีการคิดค้นโมเดลเชิงเส้นทั่วไปจากหลักการทางสถิติศาสตร์ ไปจนถึงการพัฒนาวิธีการให้เหตุผลตามหลักความน่าจะเป็น โดยเฉพาะในการประยุกต์ด้านการวินิจฉัยโรคอัตโนมัติ ส่วนการเรียนรู้ของเครื่อง กับการทำเหมืองข้อมูลมักจะใช้วิธีการเหมือนๆกันและ มีส่วนคาบเกี่ยวกันอย่างเห็นได้ชัด สิ่งที่แตกต่างระหว่างสองศาสตร์นี้คือ การเรียนรู้ของเครื่อง เน้นเรื่องการพยากรณ์ข้อมูลจากคุณสมบัติที่รู้แล้วที่ได้เรียนรู้มาจากข้อมูลชุดสอน ส่วนการทำเหมืองข้อมูล เน้นเรื่องการค้นหาคุณสมบัติที่ไม่รู้จากข้อมูลที่ได้มา กล่าวได้ว่าเป็นขั้นตอนการวิเคราะห์เพื่อค้นหาความรู้ในฐานข้อมูลนั่นเอง

แต่อย่างไรก็ตามสองศาสตร์นี้มีส่วนคาบเกี่ยวกันไม่น้อย คือ การทำเหมืองข้อมูลใช้วิธีการทางการเรียนรู้ของเครื่อง แต่มักจะมีเป้าหมายในใจที่แตกต่างออกไปเล็กน้อย ส่วนการเรียนรู้ของเครื่องก็ใช้วิธีการของการทำเหมืองข้อมูลบางอย่าง เช่น การเรียนรู้แบบไม่มีผู้สอน หรือขั้นตอนการเตรียมข้อมูลเพื่อปรับปรุงความถูกต้องของการเรียนรู้ บ่อยครั้งที่นักวิทยาศาสตร์ผสมสองสาขานี้เข้าด้วยกันด้วยเหตุผลที่ว่า ประสิทธิภาพของการเรียนรู้ของเครื่องมักจะดีขึ้นหากมีความสามารถในการรู้ความรู้บางอย่าง ในขณะที่การค้นหาความรู้และการทำเหมืองข้อมูลนั้น กุญแจสำคัญคือการค้นหาความรู้ที่ไม่รู้มาก่อน หากมีการวัดประสิทธิภาพจากสิ่งที่ไม่รู้มาก่อน วิธีการเรียนรู้แบบมีผู้สอนของการเรียนรู้ของเครื่อง ก็มักจะให้ผลได้ดีกว่าการใช้วิธีการเรียนรู้แบบไม่มีผู้สอนอย่างเดียว นั่นคือ นอกเหนือจากนั้นการเรียนรู้ของเครื่อง ยังมีความคล้ายคลึงกับการหาค่าเหมาะที่สุด (Optimization) นั่นคือ การเรียนรู้หลายอย่างมักจะถูกจัดให้อยู่ในรูปแบบของการหาค่าที่น้อยที่สุดของฟังก์ชันการสูญเสียบางอย่างจากข้อมูลชุดสอน ฟังก์ชันการสูญเสียหมายถึงความแตกต่างระหว่างสิ่งที่พยากรณ์ไว้กับสิ่งที่เป็นจริง

การจัดประเภทการเรียนรู้ของเครื่อง (Machine Learning)

เนื่องจากการเรียนรู้ของเครื่องเป็นเทคนิคในการค้นความรู้จากข้อมูลขนาดใหญ่ การทำเหมืองข้อมูลจึงเป็นการรวมเอาศาสตร์ต่าง ๆ หลายแขนงมารวมไว้ด้วยกันโดยไม่จำกัดวิธีการที่จะ

ใช้ ตัวอย่างศาสตร์ที่ใช้ เช่น เทคโนโลยีฐานข้อมูล (Database Technology) วิทยาศาสตร์สารสนเทศ (Information Science) สถิติ (Statistics) และระบบการเรียนรู้ (Machine Learning) เป็นต้น ซึ่งศาสตร์ต่างๆ เหล่านี้จะทำให้เกิดกระบวนการค้นความรู้ในแบบต่าง ๆ โดยภาพแบบการค้นความรู้หลักมีดังนี้ (Witten, et al., 2005; Han and Kamber, 2006)

การจำแนกประเภทและการทำนาย (Classification & Prediction) จัดเป็นกระบวนการที่ใช้ในการหาภาพแบบของชุดข้อมูลที่มีความใกล้เคียงกัน หรือเหมือนกันมากที่สุด เพื่อใช้ในการทำนายชุดข้อมูลว่าอยู่ในประเภทใดของชุดข้อมูลที่ได้ทำการแบ่งไว้แล้ว ซึ่งชุดข้อมูลที่แบ่งไว้เกิดจากการเรียนรู้จากชุดข้อมูลที่มีอยู่แล้ว (Training Data) แบบจำลองที่เกิดจากการเรียนรู้สามารถแสดงได้หลายภาพแบบ เช่น กฎการแบ่ง (Classification Rules, IF-THEN) การคำนวณแบบต้นไม้วิเคราะห์ (Decision Tree) การใช้สูตรทางคณิตศาสตร์ (Mathematical Formula) หรือโครงข่ายใยประสาทเทียม เป็นต้น ในส่วนของการทำต้นไม้วิเคราะห์ จะแสดงออกมาในลักษณะของแผนภูมิโครงสร้างต้นไม้ ซึ่งก้านของต้นไม้จะแสดงถึงความรู้ที่ได้ และใบไม้จะแสดงถึงประเภทชุดข้อมูลที่ถูกแบ่งออกมา แผนภูมิต้นไม้สามารถแปลงเป็นกฎการแบ่งได้ง่ายเพราะลักษณะของแผนภูมิสามารถเข้าใจได้ง่าย ในส่วนของโครงข่ายใยประสาทเทียมนั้น จะแสดงในลักษณะของการเชื่อมต่อระหว่างหน่วยที่เกิดขึ้น การทำการแบ่งประเภทนั้นมักจะใช้ประโยชน์ร่วมกับการทำนาย โดยเฉพาะข้อมูลที่เป็นตัวเลข เราจึงอาจมองได้ว่าการทำนายเป็นการบอกถึงค่าตัวเลขและการบ่งบอกประเภทของข้อมูลนั้นในลักษณะของการดูแนวโน้ม (Trends) ที่จะเกิดขึ้น ตัวอย่างเทคนิคของการแบ่งประเภทและการทำนายได้แก่ การคำนวณแบบพันธุกรรม (Genetic Algorithm) การคำนวณแบบต้นไม้วิเคราะห์ และโครงข่ายใยประสาทเทียม (Neural Network) เป็นต้น

การจัดกลุ่ม (Clustering Analysis) จัดเป็นการวิเคราะห์เพื่อจัดกลุ่ม ซึ่งจะแตกต่างกับการทำการแบ่งประเภทและการทำนายซึ่งวิเคราะห์กลุ่มข้อมูลที่มีความคล้ายกันมากที่สุด ซึ่งจะเป็นการจัดกลุ่มที่แบ่งประเภทโดยไม่มีการระบุชื่อกลุ่มในช่วงของการสอน แบบจำลองโดยทั่วไปแล้ววิธีแบบนี้จะใช้กับการจัดการแบ่งข้อมูลที่ไม่รู้ว่าจะจัดประเภทไว้ด้วยกันอย่างไรดี และการทำการวิเคราะห์นี้จะสามารถทำการบ่งบอกถึงชื่อของกลุ่มที่แบ่งขึ้นได้ด้วย ในการทำการวิเคราะห์เพื่อจัดกลุ่มนั้นจะอาศัยพื้นฐานของความเหมือนกันมากที่สุดและความเหมือนกันน้อยที่สุดของกลุ่ม คือ ข้อมูลที่ถูกจัดไว้ในกลุ่มเดียวกันจะมีความคล้ายกันสูงมาก แต่จะแตกต่างกันกับข้อมูลที่ถูกจัดไว้คนละกลุ่ม และตัวอย่างของการวิเคราะห์เพื่อจัดกลุ่มได้แก่ การหาค่าเฉลี่ย K (K-Mean Algorithm) การรวมและการแบ่งกลุ่มโดยจัดลำดับชั้น (Agglomerative and Divisive Hierarchical Clustering) และการ

ลำดับตำแหน่งเพื่อแสดงโครงสร้างการจัดกลุ่ม (Ordering Points To Identify The Clustering Structure) เป็นต้น

การวิเคราะห์ความสัมพันธ์ (Association Analysis) (Daniel, et al., 2009) จัดเป็นภาพแบบการค้นความรู้โดยการหาสิ่งที่เรียกว่า กฎความของสัมพันธ์ (Association Rules) ซึ่งจะแสดงความสัมพันธ์ของค่าที่มีความสัมพันธ์และมีเงื่อนไขที่ตรงกับข้อกำหนดและลักษณะของข้อมูลที่มีการเรียนรู้ในภาพของตะกร้าจ่ายตลาด (Market Basket) หรือการซื้อขาย (Transaction) ในการทำการวิเคราะห์ความสัมพันธ์กฎที่เกิดขึ้นนี้จำเป็นที่จะต้องกำหนดค่าสนับสนุน (Support) และค่าความมั่นใจ (Confidence) ซึ่งจะเป็นตัวกำหนดว่ากฎที่เกิดขึ้นนั้นมีความสัมพันธ์กันในระดับใดและยังเป็นการช่วยยับยั้งการเกิดกฎที่ไม่จำเป็นหรือกฎที่มีความเกี่ยวข้องกันน้อยมาก ตัวอย่างเทคนิคของการวิเคราะห์ความสัมพันธ์ได้แก่ การวิเคราะห์แบบตะกร้าสินค้า (Market Basket Analysis) การคำนวณแบบแอพพริออรี (Apriori Algorithm) และกฎความสัมพันธ์แบบหลายระดับ (Multilevel Association Rules) เป็นต้น

การลดมิติของข้อมูล (Dimension Reduction)

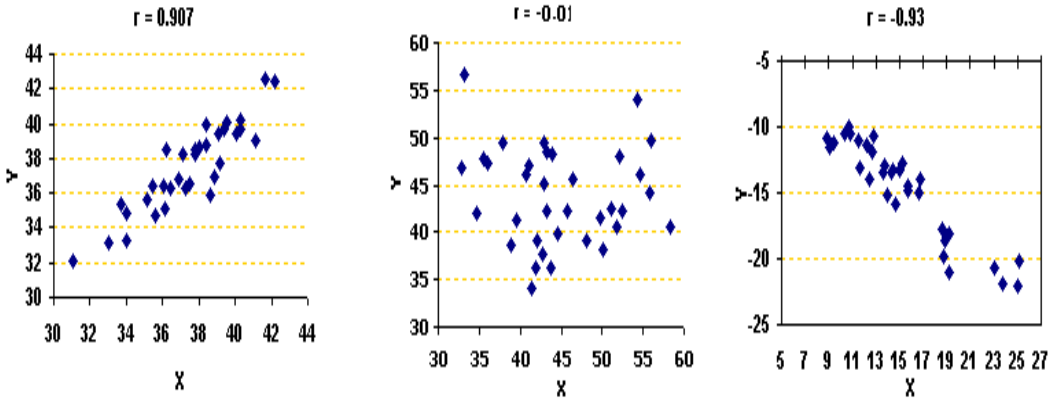
การลดมิติของข้อมูลถือเป็นงานสำคัญในงานการเรียนรู้ของเครื่องและการรู้จำ เนื่องจากเทคโนโลยีในปัจจุบันนี้ มีความสามารถในการเก็บข้อมูลทุกๆ คุณลักษณะที่เราสามารถจะเก็บได้ ซึ่งนำไปสู่การเก็บข้อมูลขนาดใหญ่ที่มีคุณลักษณะมากเกินกว่าที่จะประมวลผลได้ เทคนิคลดมิติของข้อมูลช่วยลดความซ้ำซ้อน ระหว่างคุณลักษณะและช่วยขจัดความห่างมากในชุดข้อมูล การเลือกฐานหลักถือเป็นปัญหาเปิด ว่าเราจะสามารถลดมิติของข้อมูลเพื่อให้ได้เซตย่อยของข้อมูลที่เหมาะที่สุดอย่างอัตโนมัติได้อย่างไร ในงานการเรียนรู้ของเครื่อง จากการศึกษาพบว่าวิธีลดมิติของข้อมูล (Data Reduction) (Nicholls, 2010) จัดเป็นกระบวนการหนึ่งในขั้นตอนการเตรียมข้อมูล นั่นคือการทำให้ข้อมูลตั้งต้นมีขนาดลดลง โดยสูญเสียลักษณะสำคัญของข้อมูลน้อยที่สุดและสูญเสียความถูกต้องของผลลัพธ์น้อยที่สุด เนื่องจากคุณลักษณะของฐานข้อมูลแต่ละตัว จะมีความสำคัญต่อการจำแนกไม่เท่ากัน ดังนั้นด้วยเทคนิคการเลือกข้อมูลที่ดี จะทำให้สามารถเลือกข้อมูลที่มีความสำคัญและสามารถใช้เป็นตัวแทนของข้อมูลส่วนใหญ่ได้ และในความเป็นจริงมักจะเกิดเหตุการณ์ที่ เรียกกันว่าปัญหาของมิติข้อมูล (Curse of Dimensionality) ขึ้นเสมอ นั้นหมายความว่าจำเป็นต้องลดขนาดมิติของข้อมูลลง (Dimensionality Reduction) เพื่อให้ตัวจำแนกประเภทสามารถทำงานได้ถูกต้องมากขึ้น ซึ่งงานวิจัยนี้ใช้ ค่าความสัมพันธ์ (Correlation) เพื่อลดมิติของข้อมูล เพราะเป็นการคัดเลือกมิติข้อมูล โดยใช้การคำนวณหาค่าน้ำหนัก ซึ่งอาจจะเป็นค่า

ความสัมพันธ์ระหว่างแต่ละตัวแปรการอธิบายความสัมพันธ์ของตัวแปร 2 ตัว โดยใช้ค่าเชิงปริมาณ เพราะบางครั้งการดูแต่ การวากกราฟเพียงอย่างเดียว ก็อธิบายระดับความแตกต่างได้ไม่ละเอียดพอ ค่าความสัมพันธ์(Correlation) คือค่าที่ใช้บอกระดับความสัมพันธ์ และยังบอกด้วยว่าความสัมพันธ์ดังกล่าวเป็นชนิดใด Coefficient of correlation เป็นค่าที่ใช้บ่งบอกระดับความสัมพันธ์เชิงเส้นดังกล่าว โดยจะมีค่าอยู่ระหว่าง -1.0 ถึง +1.0 โดยที่ค่าที่อยู่ใกล้ -1.0 หรือ +1.0 ถือว่ามีความสัมพันธ์กันมากที่สุด ส่วน 0 หมายความว่า ตัวแปรทั้งสองไม่มีความสัมพันธ์กันแม้แต่น้อย ส่วนเครื่องหมาย + หรือ - บงบอกว่าความสัมพันธ์นั้น เป็นตามกันหรือตรงกันข้าม เช่น ตัวแปรหนึ่งเพิ่มค่าขึ้นอีกตัวแปรหนึ่งก็จะเพิ่มตาม แต่ถ้าลดก็จะลดตาม ลักษณะเช่นนี้ ค่า r จะเป็นบวก แต่ในกรณีที่ตัวแปรหนึ่งเพิ่มค่า แต่อีกตัวแปรจะลดค่าลง แต่ตัวแปรหนึ่งลดลงอีกตัวแปรจะเพิ่มขึ้น ลักษณะเช่นนี้ค่า r จะมีเครื่องหมาย - โดยการเลือกมิติข้อมูลนี้จะพิจารณาโดยเรียงลำดับตามค่าน้ำหนักความสัมพันธ์ที่คำนวณได้ แล้วเลือกมิติข้อมูลที่มีค่าน้ำหนักมากกว่าที่ต้องการมาใช้งานต่อไป

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

เมื่อ x_i, y_i คือค่าใดๆของแต่ละตัวแปร ที่เป็นคู่กัน

$\bar{x}, \bar{y}$ คือค่า เฉลี่ยของแต่ละตัวแปร



ภาพประกอบ 3 ความสัมพันธ์ระหว่าง 2 ตัวแปร ในรูปแบบ 3 ลักษณะ

การเรียนรู้ของเครื่องการจำแนกประเภท (Machine Learning Classifier)

อัลกอริทึมในการจำแนกประเภทการเรียนรู้แบบมีผลเฉลย (Supervised Learning) สามารถแบ่งขั้นตอนวิธีการจำแนกได้เป็น 2 ขั้นตอนใหญ่ๆ คือ การเรียนรู้เพื่อสร้างแบบจำลองต้นแบบและแยกแยะประเภทของสิ่งที่สนใจ โดยการตรวจสอบหาความคล้ายกับแบบจำลองต้นแบบ

1. เนอีฟเบย์ (Naïve Bayes)

การเรียนรู้แบบเบย์ (Panigrahi, et al., 2009; Vijayshree, 2016) เป็นวิธีการเรียนรู้ที่ใช้หลักการของความน่าจะเป็น ซึ่งมีพื้นฐานมาจากทฤษฎีของเบย์ (Bayes Theorem) เข้ามาช่วยในการเรียนรู้ จุดมุ่งหมายก็เพื่อต้องการสร้างโมเดลที่อยู่ในรูปของความน่าจะเป็น ซึ่งเป็นค่าที่บันทึกได้จากการสังเกต จากนั้นนำโมเดลมาหาว่าสมมติฐานใดถูกต้องที่สุดโดยใช้ความน่าจะเป็นเข้ามาช่วย ความรู้ก่อนหน้า หมายถึง ความรู้ที่เรามีเกี่ยวกับสมมติฐานแต่ละตัวก่อนที่เราจะเก็บข้อมูล เมื่อใช้งานเราจะนำความน่าจะเป็นของข้อมูลที่เก็บได้มาปรับสมมติฐานซ้ำอีกครั้ง ข้อดีของวิธีการเรียนรู้แบบนี้ คือสามารถใช้ข้อมูลและความรู้ก่อนหน้า (Prior Knowledge)

วิธีการเรียนรู้เบย์อย่างง่าย (Naïve Bayesian Learning) การเรียนรู้เบย์อย่างง่าย เป็นวิธีจำแนกประเภทข้อมูลที่มีประสิทธิภาพวิธีหนึ่ง โดยที่ใช้งานได้ดี เหมาะกับกรณีของเซตตัวอย่างมีจำนวนมากและคุณสมบัติ (Attribute) ของตัวอย่างไม่ขึ้นต่อกัน มีการนำจำแนกประเภทเบย์อย่างง่ายไปประยุกต์ใช้งานในด้านการจำแนกประเภทข้อความ (Text Classification) การวินิจฉัย (Diagnosis) และพบว่าใช้งานได้ดีไม่ต่างจากการจำแนกประเภทวิธีการอื่น เนื่องจากเป็นวิธีการจำแนกข้อมูลได้อย่างมีประสิทธิภาพและอัลกอริทึมในการทำงานที่ไม่ซับซ้อน

หลักการของวิธีการนี้ ใช้การคำนวณความน่าจะเป็นซึ่งถูกใช้ในการทำนายผล จัดเป็นเทคนิคในการแก้ปัญหาแบบที่สามารถคาดการณ์ผลลัพธ์ได้และสามารถอธิบายได้ด้วย จะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปร เพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ เหมาะกับกรณีของเซตตัวอย่างมีจำนวนมากและคุณสมบัติ (Attribute) ของตัวอย่างไม่ขึ้นต่อกัน โดยกำหนดให้ความน่าจะเป็นของข้อมูลที่จะเป็น ดังสมการ กำหนดให้ความน่าจะเป็นของข้อมูลที่จะเป็นกลุ่ม v_j สำหรับข้อมูลที่มีคุณสมบัติ n ตัว

$X = \{a_1, a_2, a_3, a_n\}$ หรือ ใช้สัญลักษณ์ว่า $P(a_1, a_2..., a_3 | v_j)$ คือ

$$P(a_1, a_2..., a_3 | v_j) = \prod_{i=1}^{n} P(a_1 | v_j)$$

$P(a_i | v_j)$

$P(v_j)$

V_{NB}

โดยที่ $\prod$ หมายถึง ผลคูณของค่า $P(a_i | v_j)$ ทั้งหมด i = 1,2,3,…,n และ j = 1,2,3,…,n ทำการหาค่าความน่าจะเป็นของคำที่พบในแต่ละกลุ่มโดยนำค่า $P(a_1, a_2..., a_3 | v_j)$ จากสมการ มาคูณกับค่าความน่าจะเป็นของกลุ่มนั้น ๆ คือ $P(v_j)$ ได้เท่ากับ V_{NB} นำค่าที่ได้ มาเปรียบเทียบกัน กลุ่มที่มีค่าความน่าจะเป็นสูงสุด คือ คำตอบ ดังนั้นจะได้ว่า วิธีการจำแนกประเภทแบบเบย์อย่างง่าย ดังสมการ

$$V_{NB} = \arg\max_{v_j \in V} P(v_j) \times \prod_{i=1}^{n} P(a_1 | v_j)$$

จากข้อมูลการเล่นเทนนิส สามารถคำนวณการเล่นเทนนิสได้ดังนี้ สมมติต้องการทราบว่า ณ สภาพอากาศ Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong แล้วจะเล่นเทนนิสหรือไม่? คำตอบก็คือ “no” เพราะว่าค่าความน่าจะเป็นในการเกิด no มากกว่า yes ดังตัวอย่าง

$$v_{NB} = \arg\max_{v_k \in [yes, no]} P(v_k) P(Outlook = sunny | v_k) P(Temp = cool | v_k)$$
$$P(Humidity = high | v_k) P(Wind = strong | v_k)$$
$$P(PlayTennis = yes) = 9/14 = 0.64$$
$$P(PlayTennis = no) = 5/14 = 0.36$$
$$P(Wind = strong | PlayTennis = yes) = 3/9 = 0.33$$
$$P(Wind = strong | PlayTennis = no) = 3/5 = 0.60$$
$$...$$
$$P(yes) P(sunny | yes) P(cool | yes) P(high | yes) P(strong | yes) = 0.0053$$
$$P(no) P(sunny | no) P(cool | no) P(high | no) P(strong | no) = \mathbf{0.0206}$$
$$\Rightarrow answer : PlayTennis(x) = no$$

2. ตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model)

การวิเคราะห์ตัวแบบเชิงเส้นนัยทั่วไป (Jon, 2007; Maria, et al., 2016) จัดเป็นวิธีการทางสถิติที่มีแนวคิดเพื่อหาความสัมพันธ์ระหว่างตัวแปรต้นกับตัวแปรตาม ที่ไม่มีข้อจำกัดว่าการกระจายตัวของค่าความคลาดเคลื่อนในตัวแปรตามจะต้องเป็นแบบปกติเท่านั้น และข้อมูลที่นำมาวิเคราะห์ความสัมพันธ์เป็นได้ทั้งข้อมูลตัวเลขและไม่เป็นตัวเลข ตัวแบบเชิงเส้นแบบนี้เป็นที่นิยมใช้กันอย่างแพร่หลายเนื่องจากเป็นตัวแบบที่ง่ายและไม่ซับซ้อน จนเกินไป มีเครื่องมือให้เลือกใช้หลากหลายและสามารถให้ผลข้อมูลได้รวดเร็ว โครงสร้างของตัวแบบเชิงเส้นนัยทั่วไปประกอบด้วย 3 องค์ประกอบคือ องค์ประกอบแบบสุ่ม องค์ประกอบเชิงระบบ และฟังก์ชันเชื่อมโยง องค์ประกอบแบบสุ่ม (Random Component) เป็นองค์ประกอบที่เกี่ยวข้องกับคุณลักษณะของการแจกแจงของตัวแปรตาม โดยตัวแปรตามจะมีค่าเป็นอิสระต่อกันและมีการแจกแจงแบบปกติ หรือไม่ปกติก็ได้แต่ต้องเป็นการแจกแจงที่อยู่ในชนิดของตระกูลเอกโพเนนเชียล (Exponential Family) สมมุติฐานของตัวแบบเชิงเส้นวางนัยทั่วไปมีดังนี้

1) องค์ประกอบแบบสุ่ม (Random Component) ค่าของ Y เป็นอิสระต่อกัน และมีการแจกแจงแบบใดแบบหนึ่งในวงศ์ชี้กาลัง (Exponential Family)

2) องค์ประกอบแบบเป็นระบบ (Systematic Component) สามารถเขียนตัวแปรต้นให้อยู่ในรูปตัวประมาณเชิงเส้น η ได้ โดยที่ $\eta = \boldsymbol{X}\beta$

3) ฟังก์ชันเชื่อมโยง (Link Function) ความสัมพันธ์ระหว่างองค์ประกอบแบบสุ่มและองค์ประกอบแบบเป็นระบบถูกกาหนดด้วยฟังก์ชันเชื่อมโยง (g) โดยที่ฟังก์ชันเชื่อมโยงนั้นสามารถหาอนุพันธ์ได้ (Differentiable) และเป็นฟังก์ชันทางเดียว (Monotonic) โดยที่

$$E[\underline{Y}] \equiv \underline{\mu} = g^{-1}(\underline{\eta})$$

การแจกแจงที่อยู่ในวงศ์ชี้กาลัง (Exponential Family) มีคุณสมบัติ 2 ประการ ดังนี้ 1. การแจกแจงสามารถเขียนได้ในรูปแบบของค่าเฉลี่ย และความแปรปรวน 2. ความแปรปรวนเป็นฟังก์ชันของค่าเฉลี่ย จากคุณสมบัติข้อที่ 2 เราสามารถเขียนให้อยู่ในรูปของสมการได้ดังสมการ

$$Var(Y_i) = \frac{\phi V(\mu_i)}{\omega_i}$$

โดยที่ ϕ เป็นพารามิเตอร์ที่กำหนดขนาดของความแปรปรวน (Scaled Parameter) และ ω_i เป็นค่าคงที่ที่ กำหนดน้ำหนักให้กับค่าสังเกต (Prior Weight) แต่ละตัว โดยการแจกแจงที่อยู่ในวงศ์ชี้กาลังและความแปรปรวนของการแจกแจงต่างๆ สามารถสรุปได้โดยย่อตามตาราง

การแจกแจงที่อยู่ในวงศ์ชี้กาลัง (Exponential Family)

การแจกแจง	ความแปรปรวน
Normal	1
Poisson	x
Gamma	x^2
Binomial	$x(1-x)$ เมื่อจำนวนการทดลองเท่ากับ 1 ครั้ง
Inverse Gaussian	x^3

3. การวิเคราะห์การถดถอยโลจิสติก (Logistic Regression**)**

การวิเคราะห์การถดถอยโลจิสติก (Jon, 2007; Francisca, 2011) เป็นการวิเคราะห์ที่มีเป้าหมายเพื่อทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจ โดยอาศัยสมการโลจิสติคที่สร้างขึ้นจากชุดตัวแปรทำนาย ที่เป็นตัวแปรที่มีข้อมูลอยู่ ในระดับช่วงเป็นอย่างน้อย โดยที่ระหว่างตัวแปรทำนายจะต้องมีความสัมพันธ์กันต่ำ การวิเคราะห์การถดถอยโลจิสติก (Logistic Regression Analysis) เป็นเทคนิคการวิเคราะห์สถิติเชิงคุณภาพ (Qualitative Statistical Techniques) ที่แตกต่างไปจากเทคนิคการวิเคราะห์เชิงปริมาณ (Quantitative Techniques) อย่างน้อย ก็เรื่องของข้อมูลที่ตัวแปรตามเป็นตัวแปรเชิงคุณภาพ ซึ่งก็คือ เป็นตัวแปรเชิงกลุ่มนั่นเอง การวิเคราะห์การถดถอยโลจิสิคแบ่งเป็น 2 ประเภท คือ การวิเคราะห์การถดถอยโลจิสติกทวิ (Binary Logistic Regression Analysis) และการวิเคราะห์การถดถอยโลจิสติกพหุกลุ่ม (Multinomial Logistic Regression Analysis) การวิเคราะห์การถดถอยโลจิสติกทั้ง 2 ประเภท แตกต่างกันในด้านตัวแปรตาม โดยที่การวิเคราะห์การถดถอยโลจิสติกทวิใช้กับตัวแปรตามที่แบ่งออกเป็น 2 กลุ่มย่อย (Dichotomous Variable) มี 2 ค่า คือมีค่าเป็น 0 กับ 1 เช่น กลุ่มที่มีเหตุการณ์กับกลุ่มที่ไม่มีเหตุการณ์ ส่วนการวิเคราะห์โลจิสติคแบบพหุกลุ่มใช้กับตัวแปรตามที่มีหลายค่ามากกว่า 2 กลุ่ม (Polytomous Variable) การวิเคราะห์โลจิสติคมีเป้าหมายก็คือ เพื่อทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจ ซึ่งก็คือ ตัวแปรเกณฑ์ โดยอาศัยสมการโลจิสติคที่สร้างขึ้น จากชุดตัวแปรทำนาย (x's) ที่มีข้อมูลเป็น ตัวแปรที่มีข้อมูลอยู่ในระดับช่วง (interval scale) เป็นอย่างน้อย หากเป็นข้อมูลเชิงกลุ่มจะต้องแปลงเป็นตัวแปรทวิ ที่มีค่า 0 กับ 1 ก่อน โดยที่ระหว่างตัวแปรทำนายจะต้องมีความสัมพันธ์กันต่ำ รูปแบบสมการ การวิเคราะห์การถดถอยโลจิสติกสำหรับการวิเคราะห์การถดถอย สมการพยากรณ์ที่ได้จากตัวแบบการวิเคราะห์จะเป็นสมการแสดงความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ (Probability of Event)

$$\hat{Y}_i = \frac{e^u}{1+e^u}$$

เมื่อ $\hat{Y}$ เป็นค่าความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ

$$\ln\left[\frac{\hat{Y}}{1-\hat{Y}}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

และสามารถทำให้อยู่ในรูปเชิงเส้น (Linear model) ได้เป็น

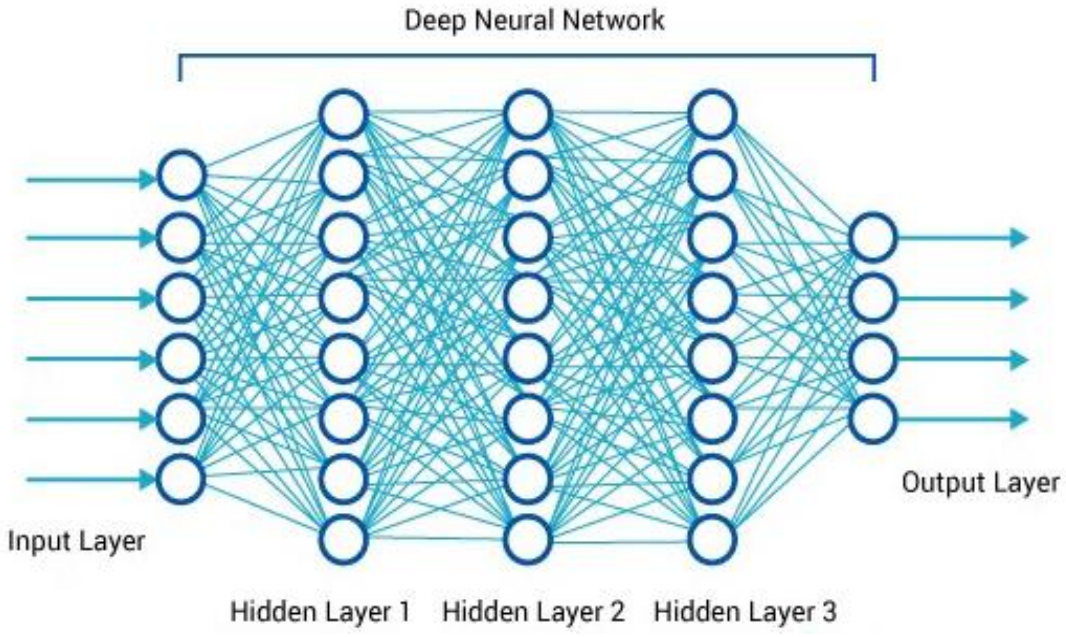
$$u = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

4. การเรียนรู้เชิงลึก (Deep Learning)

การเรียนรู้เชิงลึก (Beysolow, 2017; Francois, 2017) จัดเป็นส่วนหนึ่งของปัญญาประดิษฐ์ (AI) ที่เรียนแบบการทำงานของสมองมนุษย์ในกระบวนการประมวลผลข้อมูลและเป็นการสร้างรูปแบบสำหรับใช้ในการตัดสินใจ พื้นฐานของการเรียนรู้เชิงลึกนั้น จัดเป็นเป็นสาขาของการเรียนรู้ของเครื่องกล่าวคือ อัลกอริทึมจะพยายามจะสร้างแบบจำลองเพื่อแทนความหมายของข้อมูลในระดับสูงโดยการสร้างสถาปัตยกรรมข้อมูลขึ้นมาที่ประกอบด้วยโครงข่ายงานประสาทเทียม โดยจะพยายามหาโครงสร้างของโครงข่ายงานประสาทเทียมที่มีลักษณะเป็นกราฟที่มีจุดยอด (Node) และเส้นเชื่อม (Edge) ที่เหมาะสมและอัลกอริธึมที่จะทำให้การหาน้ำหนักที่เหมาะสม โดยเน้นความรู้ที่เก็บอยู่ในรูปแบบนามธรรม (Tacit Knowledge) การเรียนรู้เชิงลึกถือว่าเป็นวิธีการที่มีศักยภาพสูงในการจัดการกับฟีเจอร์สำหรับการเรียนรู้แบบไม่มีผู้สอนหรือการเรียนรู้แบบกึ่งมีผู้สอน โดยงานวิจัยในสาขานี้พยายามจะหาวิธีการที่ดีขึ้นในการแทนข้อมูลแล้วสร้างแบบจำลองเพื่อเรียนรู้จากตัวแทนของข้อมูลเหล่านี้ในระดับใหญ่ มีรากฐานมาจากโครงข่ายงานประสาทเทียมโดยเฉพาะเรื่องกระบวนการตีความหมายในกระบวนการประมวลผลข้อมูล ตัวอย่างของ กระบวนการที่การเรียนรู้เชิงลึกนำไปประยุกต์ใช้ได้แก่ โครงข่ายงานประสาทเทียม อันเป็นกระบวนการหาความสัมพันธ์ระหว่างตัวกระตุ้นกับการตอบสนองของเซลล์ประสาทในสมอง งานวิจัยด้าน การเรียนรู้เชิงลึก มีการออกแบบสถาปัตยกรรมการเรียนรู้หลายแบบ ได้แก่ โครงข่ายประสาทเทียมแบบลึก (Deep Artificial Neural Networks) โครงข่ายประสาทเทียมแบบสังวัตนาการ (Convolutional Neural Networks) โครงข่ายความเชื่อแบบลึก (Deep Belief Networks) และโครงข่ายประสาทเทียมแบบวนซ้ำ (Recurrent Neural Network) ซึ่งมีการนำมาใช้งานอย่างแพร่หลายในทางคอมพิวเตอร์วิทัศน์ การ

รู้จำเสียงพูด การประมวลผลภาษาธรรมชาติ การรู้จำเสียง การตรวจจับการทุจริต และชีวสารสนเทศศาสตร์ เป็นต้น

หลักแนวคิดพื้นฐานของการเรียนรู้เชิงลึกคือการมีหน่วยประมวลผลหลายๆชั้น ข้อมูลขาเข้าในแต่ละชั้นได้มาจากปฏิสัมพันธ์กับชั้นอื่นๆ ทั้งนี้ การเรียนรู้เชิงลึกพยายามหาความสัมพันธ์ที่ล้ำลึกมากขึ้น นั่นคือ เมื่อมีจำนวนของชั้นและหน่วยประมวลผลที่อยู่ในชั้นมากขึ้น ข้อมูลในชั้นสูงๆก็จะยิ่งล้ำลึกซับซ้อนมากขึ้น การเรียนรู้เชิงลึก ประกอบไปด้วยชั้นของหน่วยประมวลผลแบบไม่เป็นเชิงเส้นหลายๆชั้น ข้อมูลขาออกของแต่ละชั้นก่อนหน้า จะเป็นข้อมูลขาเข้าของชั้นต่อไป พื้นฐานมาจากการเรียนรู้ฟีเจอร์หลายๆชั้นหรือการแทนข้อมูลแบบหลายๆชั้น (แบบไม่มีผู้สอน) กล่าวคือ ฟีเจอร์ในชั้นสูงๆจะได้มาจากฟีเจอร์ในชั้นที่ต่ำกว่า เพื่อสร้างมาเป็นการแทนข้อมูลแบบหลายๆชั้นเป็นส่วนหนึ่งของสาขาการเรียนรู้ของเครื่องในการเรียนรู้การแทนข้อมูล กล่าวคือ การเรียนรู้เชิงลึก ก็คือโครงข่ายประสาทเทียม (Artificial Neuron Networks) นั่นเอง โดยการเรียนรู้เชิงลึกและ โครงข่ายใยประสาทเป็นอัลกอริทึมที่ถูกสร้างขึ้นมาเพื่อการเรียนรู้ของเครื่อง แต่ความแตกต่างระหว่าง การเรียนรู้เชิงลึก กับ โครงข่ายใยประสาท ก็คือระดับ Hidden Layer ที่ในการเรียนรู้เชิงลึกมี ชั้นซ่อน (Hidden Layer) มากกว่าในโครงข่ายประสาทเทียม จะว่าไปแล้ว การเรียนรู้เชิงลึกก็ไม่ใช่เรื่องใหม่อะไร โดยโครงข่ายประสาทเทียม เสมือนนั้นอาศัยแนวคิดและเทคนิคจากการทำงานของระบบโครงข่ายประสาทเทียม ในระบบประสาทของมนุษย์ โดยจำลองการทำงานเหมือนกับกลุ่มเซลล์ประสาทที่เชื่อมโยงกันเป็นระบบประสาทที่สามารถรับรู้หลายๆ สิ่งในเวลาเดียวกัน ด้วยการประมวลผลแบบขนาน ทำให้ระบบสามารถตัดสินใจได้ใกล้เคียงกับมนุษย์ ในการที่เครื่องจะสามารถเข้าใจสิ่งต่างๆ ได้ก็จำเป็นที่จะต้องมีองค์ความรู้ (Knowledge) เสียก่อน [Input Layer] จากนั้นก็จะประเมินชุดข้อมูลชั้นซ่อนและนำเสนอหรือแทนองค์ความรู้นั้น [Output Layer] หรืออีกนัยหนึ่งก็คือ การเรียนรู้จากตัวอย่าง โดยใช้โครงข่ายประสาทเทียม ของ Deep-Learning แบบ Multi Layered ซึ่งเลียนแบบเครือข่ายเซลล์ประสาทในสมองมนุษย์ สิ่งนี้ทำให้เทคโนโลยีมีความสามารถใหม่ที่น่าทึ่งในการจดจำรูปภาพ แยกแยะเทรนด์ และสามารถคาดการณ์และ ตัดสินใจได้อย่างชาญฉลาด เริ่มจากตรรกะหลักที่พัฒนาขึ้นในระหว่างการฝึกอบรมระบบในเบื้องต้น และหลังจากนั้นโครงข่ายประสาทเทียมของเทคโนโลยีการเรียนรู้เชิงลึกสามารถปรับปรุงประสิทธิภาพของตัวเองได้อย่างต่อเนื่องเมื่อมีการป้อนข้อมูลใหม่ๆ เข้าไป ดังภาพ



ภาพประกอบ 4 การเรียนรู้เชิงลึก

ตัวอย่างของ การเรียนรู้เชิงลึก (Deep Learning) เช่น การสอนให้หุ่นยนต์และเครื่องจักรเลียนแบบพฤติกรรมของมนุษย์ การแยกแยะใบหน้าแต่ละคน ตัวอย่างเช่นในการติดแท็กรูปภาพเพื่อนใน Facebook หรือการแยกวัตถุที่ไม่ใช่คน หรือใช้เป็นส่วนหนึ่งในระบบรถยนต์ไร้คนขับ ปัจจุบันการเรียนรู้เชิงลึกถูกใช้งานต่างๆได้อย่างประสิทธิภาพ โดยเฉพาะการประมวลผลภาพเพื่อระบุวัตถุมีอยู่ในภาพ การประมวลผลสัญญาณเพื่อจำแนกเหตุการณ์ที่สำคัญ การรู้จำเสียงพูด การรู้จำตัวอักษรเขียน การรู้จะป้ายจราจร การบังคับรถอัตโนมัติ เป็นต้น

5. **ต้นไม้ตัดสินใจ (Decision Tree)**

เป็นการนำข้อมูลมาสร้างแบบจำลอง (Quinlan, 1993; Sahin, 2011) มีลักษณะเป็นผังงาน (Flowchart) เหมือนโครงสร้างต้นไม้ เป็นการเรียนรู้แบบมีผู้สอน (Supervised Learning) คือ สร้างแบบจำลองขึ้นมาจากข้อมูลที่นำมาใช้เรียนรู้ โดยต้นไม้ตัดสินใจสามารถนำมาใช้ในการทำนายค่าต่าง ๆ ได้ โดยผลการทำนายจะขึ้นอยู่กับตัวแปรต้น รูปแบบของต้นไม้ ประกอบด้วยโหนดแรกสุดที่เรียกว่า โหนดราก (Root Node) จากโหนดรากก็จะแตกออกเป็นโหนดลูกซึ่งมีกิ่งในการเชื่อมระหว่างโหนด และโหนดลูกก็จะมีลูกของตัวเองซึ่งที่โหนดระดับสุดท้ายจะเรียกว่าโหนดใบ (Leaf Node) แต่ละโหนดของโหนดรากและโหนดลูกจะแสดงค่าคุณลักษณะ (Attribute) ที่ใช้ทดสอบข้อมูล ส่วนโหนดใบจะแสดงกลุ่ม (Class) ที่กำหนดไว้ เมื่อมีข้อมูลที่ต้องการทำนาย การทำงานจะ

เริ่มต้นจากโหนดราก ซึ่งจะนำค่าคุณลักษณะต่าง ๆ ของข้อมูลนั้นไปเปรียบเทียบกับคุณลักษณะของโหนด และทำการตัดสินใจว่าจะเดินทางไปตามกิ่งใด หลังจากนั้นจะเดินทางผ่านโหนดลูก และทำการเปรียบเทียบคุณลักษณะไปเรื่อย ๆ จนกระทั่งสุดท้ายไปถึงโหนดใบ ก็จะได้กลุ่มที่ถูกกำหนด

วิธีการสร้างต้นไม้ตัดสินใจ อันดับแรกจะหาคุณลักษณะที่สำคัญที่สุดมาแบ่งข้อมูลโดยคุณลักษณะนี้จะถูกตั้งให้เป็นโหนดราก จากโหนดรากจะสร้างเส้นทางเชื่อมหรือกิ่งไปยังโหนดลูก โดยจำนวนเส้นทางเชื่อมจะเท่ากับจำนวนค่าที่เป็นไปได้ของคุณลักษณะของโหนดราก ถ้าโหนดลูกเป็นกลุ่มของข้อมูลที่อยู่ในกลุ่มเดียวกันทั้งหมดให้หยุดการสร้างต้นไม้ แต่ถ้าโหนดลูกมีข้อมูลของหลายกลุ่มปะปนกัน ต้องสร้างโหนดลูกเพื่อจำแนกข้อมูลต่อไป โดยวนกลับไปทำขั้นตอนแรก ซ้ำเพื่อเลือกคุณลักษณะที่สำคัญที่สุดมาเป็นตัวแบ่งข้อมูลต่อไป สำหรับความสำคัญของคุณลักษณะสามารถหาได้จากการคำนวณค่าการเพิ่มสารสนเทศ (Information Gain)

ต้นไม้ตัดสินใจ สร้างกิ่งสาขาจะพิจารณาจากค่าความจริงของคุณลักษณะ โดยค่าที่ใช้จะมาจากการคำนวณจากค่าการเพิ่มสารสนเทศ การสร้างต้นไม้ตัดสินใจ C4.5 ใช้ค่ามาตรฐานอัตราส่วนเกน (Gain Ratio) เพื่อเลือกคุณลักษณะที่จะใช้เป็นรากหรือโหนด ถ้าให้ชุดข้อมูล M ประกอบด้วยค่าที่เป็นไปได้ คือ {m1, m2, …, mn} และให้ความน่าจะเป็นที่จะเกิดค่า m1 มีค่าเท่ากับ P(mi) จะได้ว่าค่าเกนสารสนเทศของ M เขียนแทนด้วย I(M) คำนวณได้ดังสมการ

$$I(M) = \sum_{i=1}^{n} -P(m_i)\log_2 P(m_i)$$

ถ้าให้ข้อมูลสอน คือ T และคุณลักษณะที่เป็นโหนด คือ x และมีค่าทั้งหมดที่เป็นไปได้ n ค่า โหนดปัจจุบันจะแบ่งตัวอย่าง T ออกตามกิ่งเป็น $\{t_1, t_2, \ldots, t_n\}$ ตามค่าที่เป็นไปได้ของ x ดังนั้นจึงสามารถคำนวณค่าเกนสารสนเทศหลังจากแบ่งตามคุณลักษณะและค่ามาตรฐานเกน(Gain) ของคุณลักษณะ x ได้ดังสมการ

$$I_x(T) = \sum_{i=1}^{n} \frac{|t_i|}{|\mathrm{T}|} I(t_i)$$

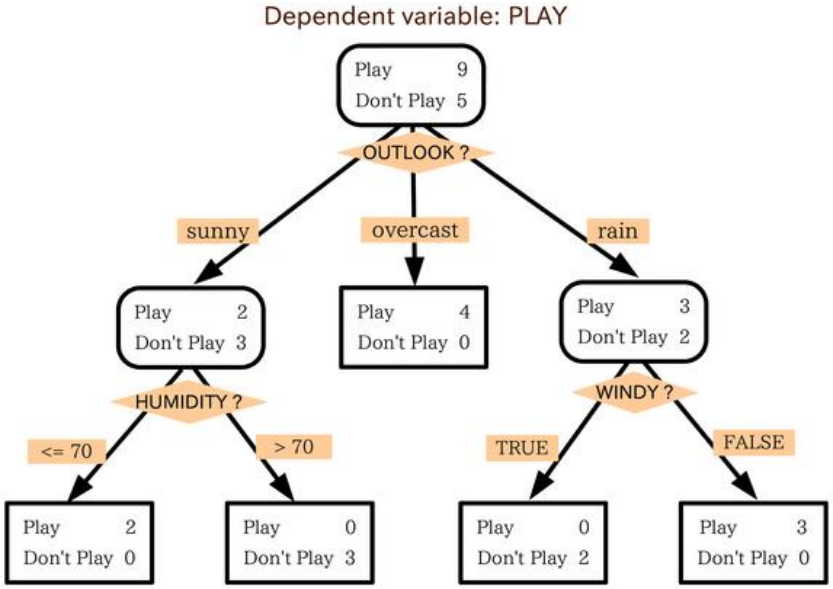
$$Gain(x) = \mathrm{I(T)} = \mathrm{I_x(T)}$$

จากนั้นคำนวณค่าสารสนเทศของการแบ่งแยก (Split Information) ของคุณลักษณะแต่ละตัว ถ้าให้ T คือ ชุดของตัวอย่าง เมื่อแบ่งตัวอย่างนี้ตามคุณลักษณะ x จะได้ชุดของตัวอย่างย่อยในแต่

ละกิ่ง คือ $\{t_1, t_2, \ldots, t_n\}$ จำนวน n ชุด ตามค่าที่เป็นไปได้ในคุณสมบัติ x เมื่อคำนวณค่าสารสนเทศของการแบ่งแยกได้ดังสมการ

$$\text{Split Information} = \sum_{i=1}^{n} \frac{|t_i|}{|\mathbf{T}|} \log_2 \frac{|t_i|}{|\mathbf{T}|}$$

คำนวนค่ามาตรฐานอัตราส่วนเกน (Gain ratio) ได้จาก Gain Ratio = Gain – Split Information ท้ายสุดจึงเลือกค่า Gain ratio สูงสุดเป็นคุณลักษณะเริ่มต้น และเลือกคุณสมบัติถัดไปตามค่า Gain ratio น้อยลงตามลำดับ



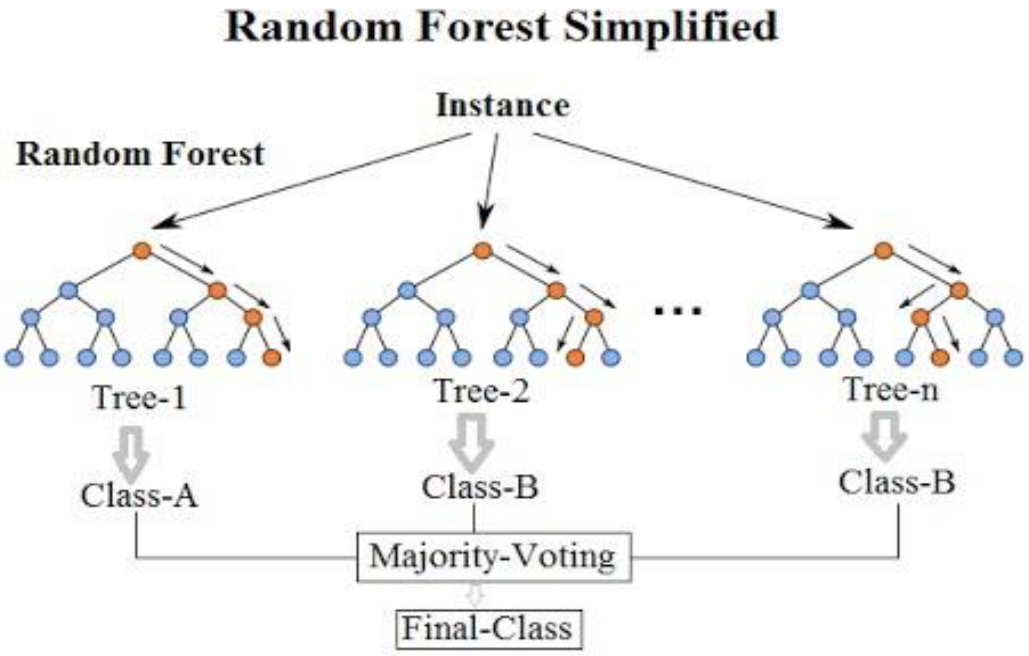
ภาพประกอบ 5 ต้นไม้ตัดสินใจ

ลักษณะของปัญหาที่เหมาะกับการใช้ต้นไม้ตัดสินใจถึงแม้ว่าวิธีการเรียนรู้ของต้นไม้ตัดสินใจจะถูกพัฒนาขึ้นมาในหลายรูปแบบเพื่อให้เหมาะสมกับรูปแบบของปัญหาต่างๆ แต่โดยทั่วไปแล้วลักษณะของปัญหาที่เหมาะกับการใช้ต้นไม้ตัดสินใจ มีขั้นตอนดังต่อไปนี้ นำข้อมูลแสดงอยู่ในรูปของชุดของคุณลักษณะ (เช่น อุณหภูมิ) และค่าของคุณลักษณะ (เช่น ร้อน) สำหรับรูปแบบที่ง่ายที่สุดสำหรับการเรียนรู้ของต้นไม้ตัดสินใจคือ เมื่อคุณลักษณะมีกลุ่มอยู่เพียงไม่กี่ตัว (เช่น ร้อน เย็น ปกติ) อย่างไรก็ตามมีอัลกอริทึมที่พัฒนาขึ้นมาเพื่อรองรับค่าของคุณลักษณะที่เป็นจำนวนจริง แล้วหาฟังก์ชันเป้าหมาย (Target Function) มีเอาต์พุตเป็นแบบไม่ต่อเนื่อง (Discrete Output Value) แต่ต้นไม้ตัดสินใจก็มีอัลกอริทึมเสริมซึ่งอนุญาตให้เอาต์พุตของฟังก์ชันเป้าหมายสามารถเป็นจำนวนจริงได้ โดยอนุญาตให้มีข้อมูลที่ผิดพลาด (Error) อยู่ในข้อมูลที่ใช้ในการสอน

ได้ เนื่องจากวิธีการเรียนรู้ของต้นไม้ตัดสินใจนั้นทนทาน (Robust) ต่อความผิดพลาด โดยยอมให้มีความผิดพลาดได้ทั้งค่าคุณลักษณะ และการแบ่งกลุ่ม และ ข้อมูลที่ใช้ในการเรียนรู้ไม่จำเป็นต้องมีค่าของคุณลักษณะอย่างครบถ้วน นั่นคือวิธีการของต้นไม้ตัดสินใจสามารถใช้ได้ ถึงแม้ข้อมูลนั้นจะมีค่าบางค่าของคุณลักษณะบางตัวหายไปหรือไม่ทราบค่าก็ตาม ดังภาพ

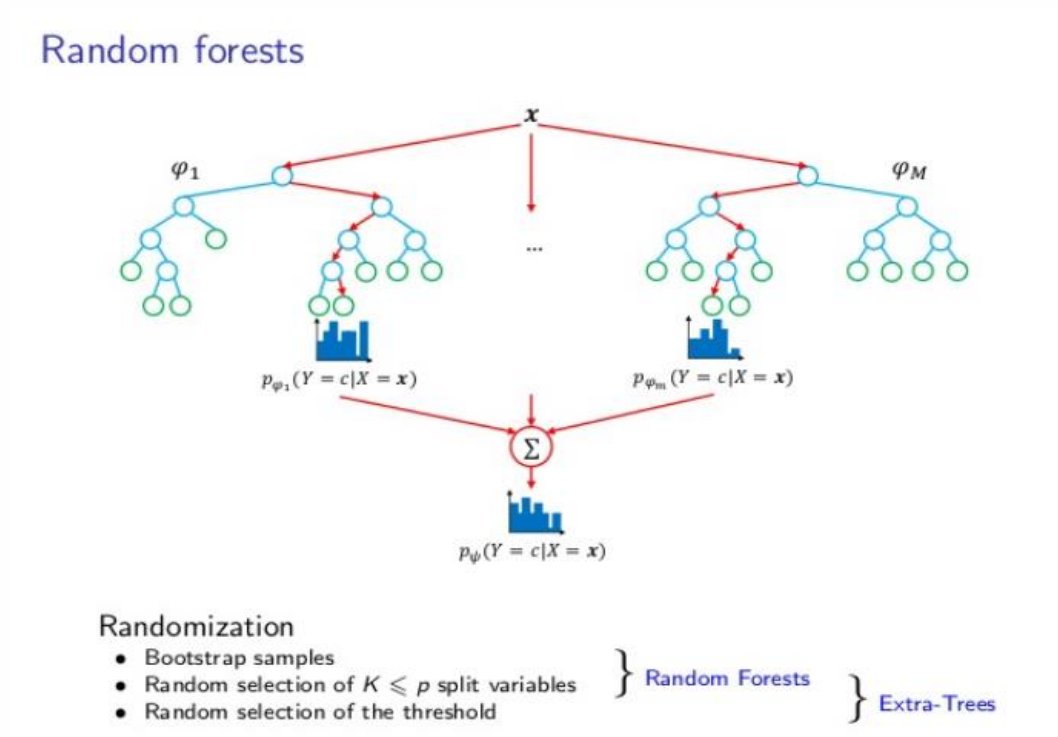
6. **แรนดอมฟอเรส (Random Forest)**

เกิดจากการรวมกลุ่มกันของโครงสร้างต้นไม้ (Leo, 2001; Scott, 2016)ซึ่งค่าความคลาดเคลื่อนโดยรวมของป่าไม้จะถูกเปลี่ยนให้เป็นค่าลิมิต ทำให้จำนวนของต้นไม้ในป่าเพิ่มขึ้น ค่าความคลาดเคลื่อนโดยรวมจะขึ้นกบความมั่นคง (Strength) ของต้นไม้แต่ละต้น รวมถึงความสัมพันธ์กันระหว่างต้นไม้เหล่านั้น โดยจะใช้วิธีการสุ่มเลือกคุณสมบัติเพื่อการแบ่งแยกโหนดทำให้ค่าความผิดพลาดลดลง ขั้นตอนวิธีนี้ จะมีประสิทธิภาพมากเมื่อนำไปใช้วิเคราะห์เกี่ยวกบเสียงและการประมาณการขนาดใหญ่ เราสามารถสร้างแบบจำลองที่ใช้ต้นไม้หลาย ๆ ต้นในการตัดสินใจเพื่อนำมาประมวลผล ซึ่ งมีความแม่นยำสูง สามารถจัดการข้อมูลได้มากและเหมาะสำหรับข้อมูลที่มีความสำคัญ ดังภาพ



ภาพประกอบ 6 การรวมกลุ่มกันของโครงสร้างต้นไม้ตัดสินใจ

ในขั้นตอนการทำงานของแรนดอมฟอเรส จะทำการจำแนกต้นไม้หลาย ๆ ต้น ซึ่งในต้นไม้ แต่ละต้นมีการแบ่งเป็น นคลาส โดยที่ต้นไม้แต่ละต้นจะถูกสร้างขึ้นจากกลุ่มตัวอย่างที่แตกต่างกัน จากกระบวนการของต้นไม้ตัดสินใจ (Tree Classification Algorithm) ถูกสร้างขึ้นจนกลายเป็นป่า (forest) จนกระทั่งวิเคราะห์การตัดสินใจจากต้นไม้แต่ละต้นที่อยู่ในป่า ดังนั้นสามารถสรุป ได้ว่าอัลกอริทึมแรนดอมฟอเรส เป็นอัลกอริทึมประเภทหนึ่งของอัลกอริทึมต้นไม้ตัดสินใจที่มีลักษณะแบบไม่ตัดแต่งกิ่ง (Unpruned) หรือต้นไม้ถดถอย (Regression Trees) ซึ่งถูกสร้างจากการนำข้อมูลฝึกสอนไปสุ่มเลือกตัวอย่างข้อมูล และคุณลักษณะข้อมูลแล้วนำมาสร้างเป็นต้นไม้ตัดสินใจซึ่งมีตัวอย่างส่วนหนึ่งที่ไม่ถูกเลือกจะถูกนำมาใช้ในการทดสอบต้นไม้ตัดสินใจ เรียกข้อมูลส่วนนี้ว่า Out-of-Bag (OOB) ซึ่งวิธีการนี้เรียกว่า Bagging ผลลัพธ์ที่ได้อย่างอิสระจากต้นไม้ตัดสินใจแต่ละต้นถูกนำมาคิดเป็นผลการโหวตที่มากที่สุด อัลกอริทึมแรนดอมฟอเรส ไม่จำเป็นต้องมีข้อมูลทดสอบเพื่อประมาณความผิดพลาดเพราะข้อมูล Out-of-Bag นั้นถูกนำมาใช้ทดสอบต้นไม้ตัดสินใจแล้ว ดังภาพ



ภาพประกอบ 7 อัลกอริทึมแรนดอมฟอเรส

7. **ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine)**

หลักการของวิธีการนี้ (Farquad, et al., 2012; Farquad, et al., 2014) 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) ซึ่งต่างจากเทคนิคโดยทั่วไปเช่น โครงข่ายประสาทเทียม (Artificial Neural Network: ANN) ที่มุ่งเพียงทำให้ความผิดพลาดจากการทำนายให้ต่ำที่สุดเพียงอย่างเดียว เหมาะใช้สำหรับข้อมูลที่มีลักษณะมิติของข้อมูลที่มีปริมาณมาก หลักการทำงานของเอสวีเอ็ม ข้อมูลจะถูกเขียนในรูปสมาชิกคู่อันดับดังนี้

$$\{(x_1, c_1), (x_2, c_2), (x_3, c_3), (x_n, c_n)\}$$

เมื่อ c_i มีค่าเป็น 1 หรือ -1 ซึ่งกำหนดให้เป็นข้อมูลแบ่งกลุ่มของข้อมูล x_i ที่มีค่า c_i เป็น 1 และ x_i ที่มีค่า c_i เป็น -1 โดยที่แต่ละ x_i เป็นค่าข้อมูลมิติของเวกเตอร์จริง เมื่อกำหนดให้ข้อมูลนี้เป็นข้อมูลสำหรับฝึกฝน ซึ่งหมายความว่าการแบ่งกลุ่มของข้อมูลมีความถูกต้อง ดังนั้นเส้นแบ่งกลุ่มข้อมูลที่ถูกสร้างขึ้นซึ่งเป็นสมการเส้นตรงทั่วไปคือ y = mx+b โดยในที่นี้จะแทน m ด้วย w^T เพื่อกำหนดเป็นสมการ (2-2) แสดงการแบ่งข้อมูลดังภาพ

$$W^T \cdot X + b = 0$$

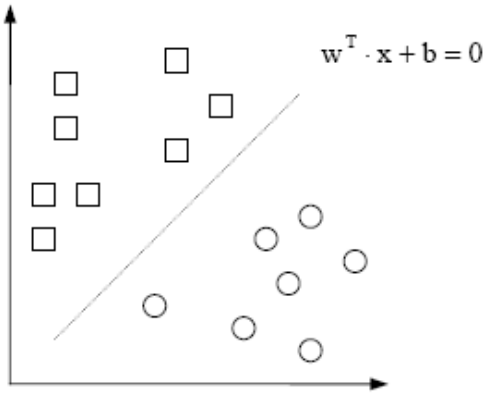
เมื่อ W^T เป็นเวกเตอร์ตั้งฉากของค่าความชัน m ของเส้นแบ่ง

b เป็นค่าคงที่ที่ได้จากค่าของแกน y ของแต่ละข้อมูล x

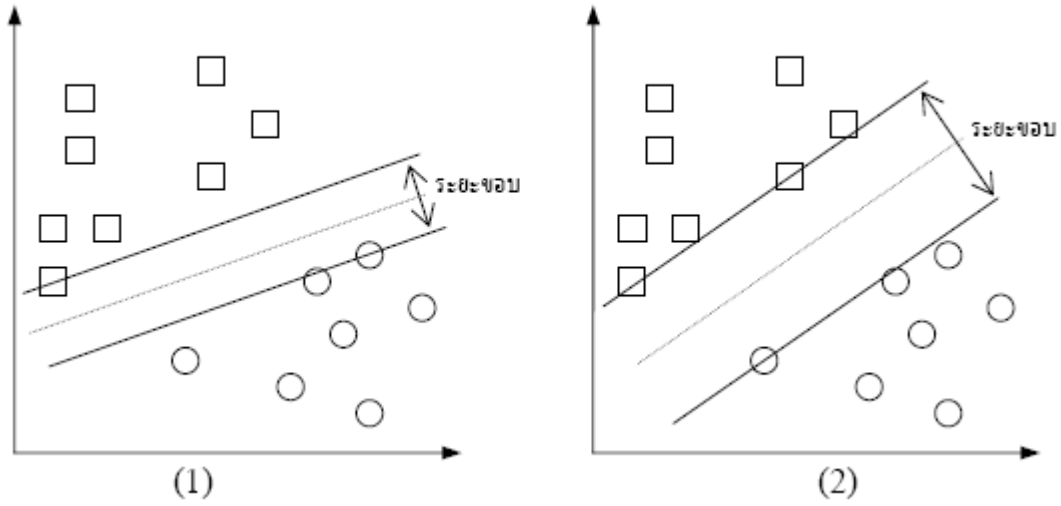
ดังนั้นในทางอุดมคติ เส้นแบ่งกลุ่มที่ดีที่สุดคือ เส้นแบ่งกลุ่มที่ทำให้มีระยะขอบ (Margin)

มากที่สุดของเส้นคู่ขนานที่ขยายออกจากเส้นแบ่งไปสัมผัสจุดข้อมูลของทั้งสองกลุ่มที่ใกล้ที่สุด โดยเรียกเส้นขนานนั้นว่า Parallel Hyper plane หรือกล่าวได้ว่าเป็นเส้นแบ่งที่มีระยะขอบกว้างที่สุด ดังภาพ เรียกจุดข้อมูลอย่างน้อยหนึ่งจุดจากทั้งสองกลุ่มที่สัมผัสกับเส้นขนานที่ขยายออกได้มากที่สุดว่าซัพพอร์ตเวกเตอร์ โดยค่าของเส้นขอบทั้งสองที่ใช้แบ่งกลุ่มข้อมูลเป็นสองกลุ่มคำนวณได้จากสมการ

$$W^T \cdot X + b = 1$$
$$W^T \cdot X + b = -1$$



ภาพประกอบ 8 ข้อมูลสองกลุ่มที่ถูกแบ่งด้วยเส้นตรง



ภาพประกอบ 9 การปรับความชันของเส้นแบ่งแล้วทำให้ได้ระยะขอบที่มากที่สุด

ดังนั้นเมื่อข้อมูลที่ใช้ฝึกฝนเป็นข้อมูลที่สามารถแบ่งกลุ่มได้ด้วยเส้นตรงใด ๆ ระยะที่กว้างที่สุดของ Hyper plane ทั้งสองที่ขยายออกไปจนกว่าจะพบจุดข้อมูลของทั้งสองกลุ่มคือ 2/|w| โดยค่า |w| มีค่าน้อยที่สุด ค่าข้อมูลแต่ละ x_i จัดจำแนกอยู่ในกลุ่มใดสามารถพิจารณาได้จากเงื่อนไขดังนี้

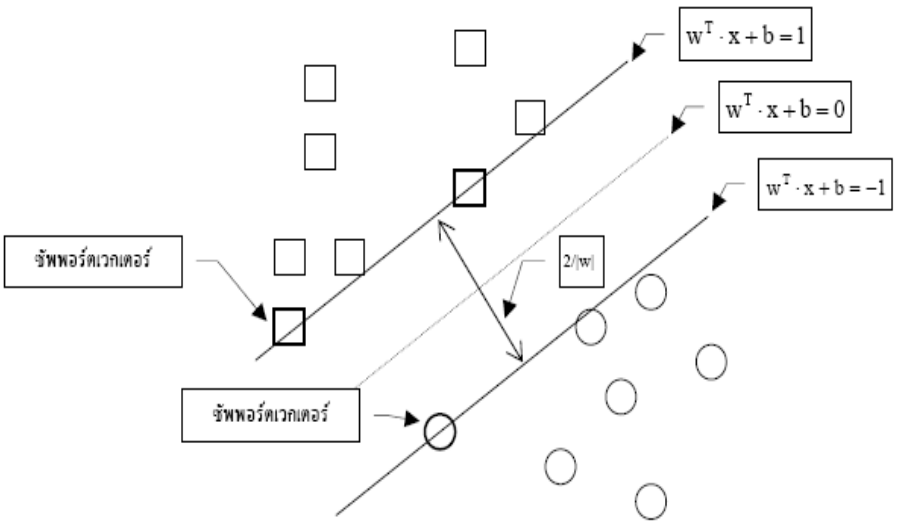
ถ้า $W^T \cdot X_i + b \geq 1$ แสดงว่า x_i เป็นกลุ่มที่ 1
และ ถ้า $W^T \cdot X_i + b \leq -1$ แสดงว่า x_i เป็นกลุ่มที่ 2

ดังนั้นในการคัดแยกจุดข้อมูลใด ๆ ที่นำมาฝึกฝนเพื่อกรองว่าเป็นกลุ่ม 1 สามารถตรวจสอบได้จากสมการ

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

ดังนั้นสามารถเขียนเป็นรูปสั้นเพื่อหาระยะขอบที่น้อยที่สุด โดยที่สามารถคำนวณการแบ่งกลุ่มได้ดังสมการ

$$Minimize_{w,b} |W| \text{ โดยตรวจสอบ } c_i(W \cdot X_i - b) \geq 1 \quad \text{เมื่อ } 1 \leq i \leq n$$

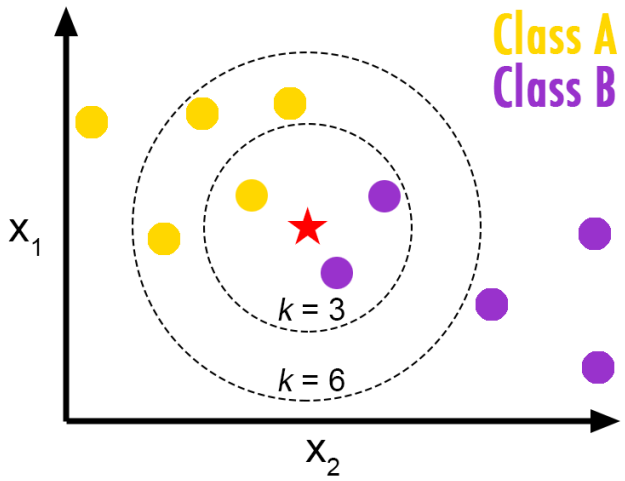


ภาพประกอบ 10 ระยะขอบที่กว้างที่สุดเมื่อสัมผัสกับซัพพอร์ตเวกเตอร์

8. **เคเนียเรสเนเบอร์ (K-Nearest Neighbor)**

เคเนียเรสเนเบอร์ (K-Nearest Neighbor) (Witten, et al., 2005; Han and Kamber, 2006) หลักการของวิธีการนี้ จะจำแนกประเภทข้อมูลโดยขึ้นกับข้อมูลที่มีคุณสมบัติใกล้เคียงที่สุด K ตัว จากข้อมูลบนชุดข้อมูลตัวอย่าง เป็นอัลกอริทึ่มที่เรียบง่าย การทำงานโดยขึ้นกับระยะทางน้อยสุดจากสมาชิกใหม่ หรือข้อมูลที่ป้อนถาม (Input Query Instance) กับข้อมูลตัวอย่างฝึกฝน จะคำนวณหาเพื่อนบ้านที่ใกล้ที่สุด K ตัว หลังจากนั้นเราจะรวบรวมสมาชิกที่ใกล้เคียงที่สุด K ตัวแล้วเลือกคลาสที่สมาชิกส่วนใหญ่ที่ในกลุ่ม K ดังกล่าวสังกัดอยู่มากที่สุดให้กับสมาชิกใหม่ ข้อมูลการจำแนกโดยใช้ข้อมูลข้างเคียง K ตัว ประกอบด้วยแอททริบิวต์หลายตัวแปร X_i ซึ่งจะนำมาใช้ในการแบ่งกลุ่ม Y_i โดยระบุค่าตัวเลขจำนวนเต็มบวกให้กับ K ซึ่งค่านี้จะเป็นตัวบอกจำนวนของกรณี (Case) ที่จะต้องค้นหาในการทำนายกรณีใหม่ โดยงานวิจัยนี้กำหนด 1-KNN หมายถึงจะค้นหา 1 กรณีที่มีลักษณะใกล้เคียงกับกรณีใหม่ (1- Nearest Cases) การนำระยะทางที่หาได้จากสมาชิกในข้อมูลตัวอย่างฝึกฝน มาเรียงลำดับจากน้อยไปหามากแล้วเลือกสมาชิกที่มีระยะทาง (Distance) ใกล้เคียงที่สุดออกมา K ตัวโดยใช้การวัดระยะทางแบบ Euclidean Distance มีหลักการคือ การวัดระยะทางระหว่างสองวัตถุ ถ้าวัตถุห่างกันมากแสดงว่าวัตถุนั้นมีความคล้ายกันน้อย ถ้ามีค่าน้อยก็แสดงว่ามีความคล้ายคลึงกันมาก โดยที่ ค่า P_i แทนคุณสมบัติจากฐานข้อมูล Q_i แทนคุณสมบัติที่ผู้ใช้ระบุ ดังแสดงในสมการ

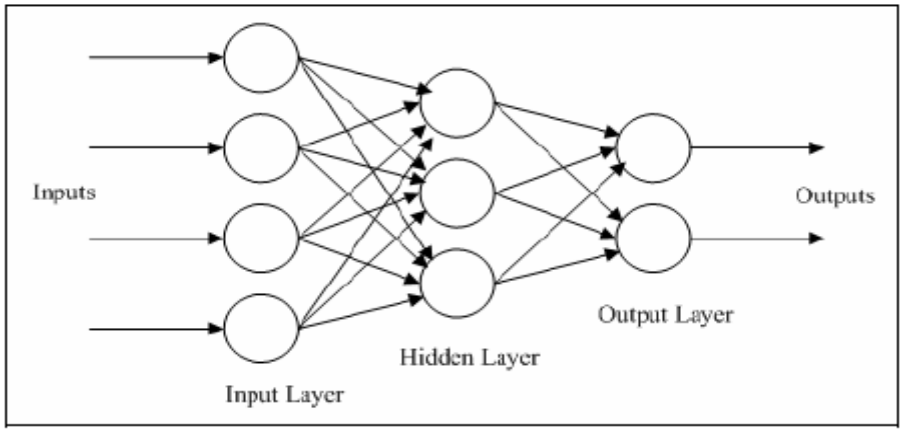
$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}.$$



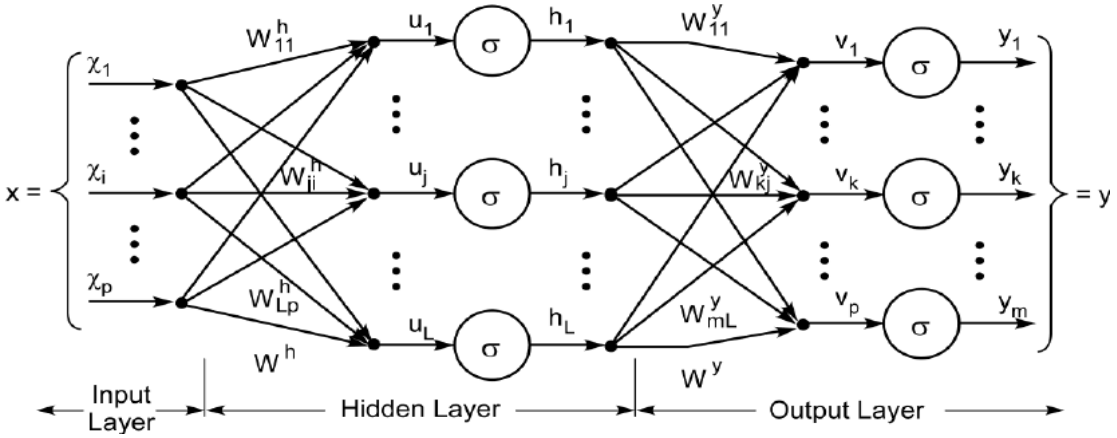
ภาพประกอบ 11 เคเนียเรสเนเบอร์

9. **โครงข่ายประสาทเทียมพอร์เซฟตรอนแบบหลายชั้น (Multi-Layer Perceptron Neural Networks)**

สถาปัตยกรรมโครงข่ายเพอร์เซฟตรอนแบบหลายชั้น เป็นโครงข่ายประเภทการคาดเดา (Networks for Prediction) (Witten, et al., 2005; Han and Kamber, 2006) ที่มักใช้ทำงานกับงานคาดเดาสามารถช่วยกำหนดลำดับความสำคัญได้ สถาปัตยกรรมโครงข่ายเพอร์เซฟตรอนแบบหลายชั้น หรือ โครงข่ายแบบป้อนไปข้างหน้าโดยมีการเรียนรู้แบบย้อนกลับ นั้นถูกพัฒนาในต้นปี 1970 โดยจากหลายๆ แหล่งและเป็นการทำงานพัฒนาร่วมกันอย่างอิสระ โดยในปัจจุบันสถาปัตยกรรมแบ็คพรอพาเกชั่นหรือแบบแพร่ย้อนกลับนี้เป็นที่นิยมสูงสุดและยังมีประสิทธิภาพมาก รวมถึงยังมีความง่ายสำหรับการเป็นต้นแบบสำหรับโครงข่ายประสาทเทียมที่มีความซับซ้อนมากขึ้นที่เป็นแบบหลายเลเยอร์ โดยเทคนิคการแพร่ย้อนของสถาปัตยกรรมแบบนี้นั้นได้ถูกใช้ในหลายแอพพลิเคชั่นด้วยกัน และยังมีผลต่อชนิดของโครงข่ายขนาดใหญ่ในด้านของรูปร่างและวิธีการฝึกที่แตกต่างกันไป เพราะมีจุดแข็งที่สำคัญของเทคนิค คือ วิธีการทำงานแบบไม่เชิงเส้น (Non-Linear) ที่มีความเหมาะสมต่อการแก้ปัญหาที่มีความไม่ชัดเจน โครงข่ายประสาทเทียมเพอร์เซฟตรอนแบบหลายชั้นนั้น มีลักษณะต้นแบบ คือ จะมีจำนวนหนึ่งชั้นอินพุต (Input Layer) หนึ่งชั้นเอาท์พุต (Output Layer) และอย่างน้อยหนึ่งชั้น ซ่อน (Hidden Layer) ซึ่งลักษณะของโครงข่ายประสาทเทียมนั้น ไม่มีข้อจำกัดทางทฤษฎีต่อจำนวนของชั้นซ่อน แต่ตามแบบต้นฉบับจะมีเพียงหนึ่งชั้นหรือสองชั้นเท่านั้น โดยบางการทำงานที่แก้ปัญหาที่ซับซ้อนจะต้องมีอย่างน้อยที่สุดสี่ชั้น (สามชั้นซ่อน กับหนึ่งชั้นเอาท์พุต) แต่ละชั้นเชื่อมต่อกับชั้นที่ตามมา ดังแสดงในภาพประกอบ



ภาพประกอบ 12 โครงข่ายเพอร์เซฟตรอนแบบหลายชั้น 1



ภาพประกอบ 13 โครงข่ายเพอร์เซฟตรอนแบบหลายชั้น 2

เซลล์ประสาททำหน้าที่รวบรวมข้อมูลนำเข้าจากหน่วยอื่น หรือเรียกว่า อินพุต (Input) เข้าสู่ระบบ ซึ่งอินพุตจะถูกปรับระดับไปตามค่าน้ำหนัก (Weighted Connections) จากนั้น ก่อนที่ค่าผลรวมของอินพุตแต่ละตัวที่ถูกปรับระดับค่าสัญญาณตามค่าน้ำหนัก จะถูกส่งออกไปยังภายนอกผ่านแกนประสาทนำออก หรือเรียกว่า เอาต์พุต (Output) ค่าผลรวมอินพุตจะถูกกระตุ้นให้มีการเปลี่ยนแปลงค่าอีกครั้ง โดยฟังก์ชันกระตุ้น หรือฟังก์ชันการเปลี่ยนแปลง (Activation Function or Transfer Function) ซึ่งเพอร์เซปตรอนสามารถแทนด้วยแบบจำลองทางคณิตศาสตร์ดังนี้

$$o(x_1, x_2, \cdots, x_n) = \begin{cases} 1 \; if \; w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n > 0 \\ -1 \; if \; w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n > 0 \end{cases}$$

โดยที่ $x_1,x_2,\ldots,x_n$ คืออินพุตเข้าสู่ระบบ และ $w_1,w_2,\ldots,w_n$ คือค่าน้ำหนักของอินพุตแต่ละตัว เอาต์พุต $o(x_1,x_2,\ldots,x_n)$ เป็นฟังก์ชันของอินพุตในรูปของผลรวมเชิงเส้นแบบถ่วงน้ำหนักน้ำหนักจะเป็นตัวกำหนดว่าในจำนวนอินพุตนั้น อินพุตตัวใด (x_i) มีความสำคัญต่อการกำหนดค่าเอาต์พุต ตัวที่มีความสำคัญมากจะมีค่าสัมบูรณ์ของน้ำหนักมาก ตรงกันข้ามตัวที่มีความสำคัญน้อยจะมีค่าใกล้ศูนย์ หากค่าผลรวมเท่ากับศูนย์ จะให้ค่าเอาต์พุตเป็นหนึ่ง หรือลบหนึ่ง (-1) ก็ได้ เมื่อกำหนดให้

$g(\bar{x}) = \sum_{i=0}^{n} wx_i = \bar{w} \cdot \bar{x}$ โดยที่ $\bar{x}$ คือเวกเตอร์อินพุต จะได้ฟังก์ชันของเอาต์พุตได้ดังนี้

$$o(x_1, x_2, \cdots, x_n) = \begin{cases} 1 \; if \; g(\bar{x}_i) > 0 \\ -1 \; if \; g(\bar{x}_i) < 0 \end{cases}$$

ฟังก์ชันกระตุ้นที่นิยมใช้ในเพอร์เซปตรอน คือ ฟังก์ชันสองขั้ว (Bipolar Function) ซึ่งแสดงผลของเอาต์พุตเป็น 1 กับ -1 หรือ ฟังก์ชันไบนารี (Binary Function) ซึ่งแสดงผลของเอาต์พุตเป็น 1 กับ 0 เพอร์เซปตรอนสามารถมีอินพุตได้หลายตัว หากมีอินพุต 2 ตัว เพอร์เซปตรอนจะเป็นเส้นตรง ในกรณีที่อินพุตมากกว่าสอง เพอร์เซปตรอนจะเป็นระนาบตัดสินใจหลายมิติ (Hyperplane Decision Surface) การเรียนรู้ของเพอร์เซปตรอนจะเกี่ยวข้องกับการหาค่าเวกเตอร์น้ำหนัก (w) ที่เหมาะสมในการจำแนกประเภทของข้อมูลสอน (Training Data) เพื่อให้เพอร์เซปตรอนแสดงเอาต์พุตได้ตรงกับค่าที่สอน โดยอาศัยกฎการเรียนรู้เพอร์เซปตรอน (Perceptron Learning Rule)

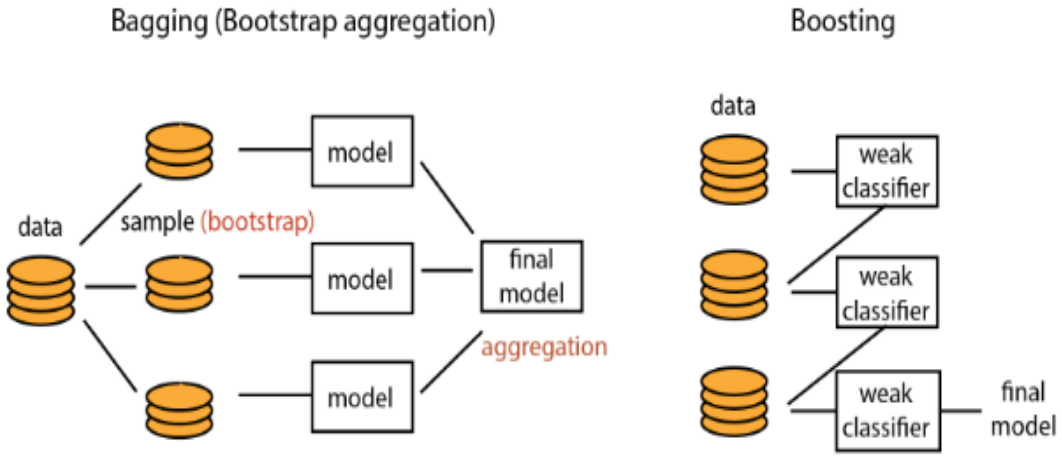
$$w_i \leftarrow w_i + \Delta w_i$$
$$\Delta w_i \leftarrow \alpha(t - o)x_i$$

โดย α คืออัตราการเรียนรู้ เป็นค่าคงที่ตัวเลขบวก t เป็นเอาต์พุตเป้าหมายของเพอร์เซปตรอนและ O เป็นเอาต์พุตของเพอรเซปตรอน เพอร์เซปตรอนไม่สามารถเรียนรู้บางฟังก์ชันได้ ฟังก์ชันเหล่านี้เรียกว่า ฟังก์ชันแยกเชิงเส้นไม่ได้ (Linearly Non-separable Function) อาทิ ฟังก์ชัน XOR (Exclusive-Or) ซึ่งเป็นข้อจำกัดของเพอร์เซปตรอน ส่วนฟังก์ชันที่แยกได้เรียกว่า ฟังก์ชันแยกเชิงเส้นได้ (Linearly Separable Function) ดังนั้นข่ายงานประสาทเทียมแบบหลายชั้น (Multilayer Perceptron or Multilayer Neural Network) ถูกออกแบบขึ้นมาเพื่อแก้ปัญหาดังกล่าว ข่ายงานประสาทเทียมแบบหลายชั้นประกอบด้วยเพอร์เซปตรอนหลายเพอร์เซปตรอน มาเชื่อมต่อกันในหลายรูปแบบ ลักษณะการเชื่อมต่อจะเป็นแบบป้อนไปข้างหน้า (Feedforward) และอยู่ในรูปของชั้น (Layer) ของเพอร์เซปตรอน โดยชั้นที่รับข้อมูลเข้าเรียกว่า ชั้นอินพุต (Input Layer) ชั้นที่ทำการประมวลผลภายในเรียกว่าชั้นซ่อน (Hidden Layer) ซึ่งอาจมีหลายชั้นได้ ชั้นสุดท้ายคือชั้นที่ให้ผลลัพธ์กับข่ายงาน เรียกว่าชั้นเอาต์พุต (Output Layer) ข่ายงานประสาทเทียมแบบหลายชั้นที่นิยมใช้ในงานวิจัย คือข่ายงานประสาทเทียมที่ใช้ อัลกอริทึมแพร่กระจายย้อนกลับ (Backpropagation) และเรียกข่ายงานประสาทเทียมแบบนี้ว่า ข่ายงานประสาทเทียมแบบแพร่กระจายย้อนกลับหลายชั้น (Multilayer Backpropagation Neural Network) ความสามารถพิเศษที่เพิ่มขึ้นมาก็คือ สามารถสร้างพื้นผิวการตัดสินใจแบบไม่เป็นเชิงเส้นได้ (Nonlinear Decision Surface) ที่แบ่งแยกตัวอย่างได้ดีกว่าพื้นผิวการตัดสินใจแบบเชิงเส้น (Linearly Decision Surface) อัลกอริทึมแพร่กระจายย้อนกลับใช้กฎการเรียนรู้แบบใหม่คือ กฎเดลต้า (Delta Rule) ซึ่งมีข้อดีตรงที่การเรียนรู้จะลู่เข้าสู่ระนาบหลายมิติที่ให้ค่าผิดพลาดน้อยที่สุด โดยใช้หลักการเคลื่อนลงตามความชัน (Gradient Descent) กฎเดลต้าจะหาเวกเตอร์น้ำหนักที่ให้ค่าผิดพลาดของตัวอย่างสอนน้อยที่สุดโดยการหาอนุพันธ์ทาง

คณิตศาสตร์ ดังนั้นจึงต้องใช้ฟังก์ชันกระตุ้นที่สามารถหาอนุพันธ์ได้ อาทิ ฟังก์ชันเชิงเส้น (Linear Function) หรือฟังก์ชันซิกมอยด์ (Sigmoid Function)

10. เอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting)

บูทติ้ง (Boosting) (Robert, et al., 2014; Nonita, 2017) จัดเป็นเทคนิคเพื่อลดความแปรปรวนและเพิ่มความแม่นยำในการทำนายของตัวจำแนกประเภท โดยใช้วิธีลดความอคติ (Bias) และมีแนวคิดที่ให้ตัวเรียนรู้ที่อ่อนแอ (Weak Learner) ชุดหนึ่ง ทำงานร่วมกันจนสามารถพัฒนามาเป็นตัวเรียนรู้ที่เข้มแข็ง (Strong Learner) ได้ การสร้างตัวเรียนรู้ที่อ่อนแอ (Weak Learner) แต่ละตัวสามารถทำได้ โดยการปรับเพิ่มน้ำหนักของการทำนายที่ผิดพลาดให้มากขึ้นในแต่ละรอบ แล้วทำการเรียนรู้ใหม่ ซึ่งจะทำให้โมเดลของตัวจำแนก (Classifier) เปลี่ยนไปโดยให้ความสำคัญกับความผิดพลาดในรอบที่แล้วมากขึ้น เมื่อได้จำนวนตัวเรียนรู้ที่อ่อนแอมากพอแล้ว จึงนำมารวมกันสร้างเป็นตัวเรียนรู้ที่เข้มแข็งต่อไป (Strong Learner) ในขั้นตอนการทำงานบูทติ้งนั้น มีวิธีคิดจะไม่เหมือนกับแบ็กกิ้ง (Bagging) เพราะบูทติ้ง คือการนำตัวจำแนกที่อ่อนแอที่มีความแม่นยำต่ำมาทำนายข้อมูลที่เรามี จากนั้นเราจะให้ (Weak Classifier) ตัวใหม่มาแก้ไข (Error) ที่เรามี โดยผลรวมของตัวจำแนกประเภท (Aggregating) จะเกิดเป็นตัวจำแนกประเภทใหม่ขึ้นมา เราจะทำแบบนี้ไปเรื่อยๆ (Recursive) จนได้โมเดลที่ดีที่สุดจากผลรวมของการจำแนก ถ้าให้มองภาพรวมการทำงานของบูทติ้ง (Boosting) ก็เหมือนการทำงานเป็นทีมนั่นเอง โดยการเอา ตัวจำแนกที่ไม่ได้ดีมากมารวมกันจนทำนายข้อมูลที่ซับซ้อนมากๆได้ แต่ข้อเสียของการใช้บูทติ้ง (Boosting) ก็คือต้องรันหลายครั้งและเป็นลำดับ กว่าจะได้โมเดลที่ต้องการ ต่างจากเหมือนกับแบ็กกิ้ง (Bagging) ที่สามาถสุ่มข้อมูลได้แล้วเทรนโมเดลได้พร้อมกันเลย แต่ความฉลาดของเอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) นั้น จะสามารถเลือกชนิดของ ตัวเรียนรู้ที่อ่อนแอ (Weak Learner) หรือ ตัวจำแนกที่อ่อนแอ (Weak Classifier) ได้ด้วย ไม่ว่าจะเป็นในรูปแบบต้นไม้ (tree) หรือแบบเส้นตรง (linear) และในหลายๆครั้งนั้นเทคนิคบูทติ้ง สามารถทำนายข้อมูลที่มีความซับซ้อนมากๆได้มากกว่าการใช้แบ็กกิ้ง เอ็กซ์ทรีมกาเดียนบูทติ้งมีพื้นฐานอยู่บนหลักการของ Gradient Boosting ซึ่งจัดเป็นใช้โมเดลตัวจำแนกประเภทหลายๆ โมเดล มาช่วยในการหาคำตอบ (Ensemble Classifier) ในกลุ่มบูทติ้งประเภทหนึ่ง แต่ได้รับการปรับปรุงให้สามารถทำงานได้อย่างรวดเร็วและมีประสิทธิภาพมากขึ้นกว่าเดิม สามารถใช้ประโยชน์จาก Multithread ได้อย่างเต็มที่ และเพิ่มตัวแปร Regularization เข้าไปเพื่อลดการเกิด การที่โมเดลจดจำรูปแบบของข้อมูล (Overfitting) ลงได้



ภาพประกอบ 14 ความแตกต่างระหว่างบูทติ้ง (Boosting) และแบ็กกิ้ง (Bagging)

เอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) ใช้หลักการสร้างต้นไม้แต่ละต้นจะเป็นแบบเรียงลำดับ (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) ถูกแก้ไข แค่ ความลึกของชั้นต้นไม้ (Tree Depth) แค่หนึ่งชั้น ก็เพียงพอที่จะได้ค่าประสิทธิภาพ (Performance) ที่ดีขึ้นมาก เมื่อเทียบกับ (Bagging Tree) และแรนดอมฟอเรส (Random Forest) ที่ต้องเพิ่มชั้นความลึกมากขึ้นมากขึ้น เพื่อให้ประสิทธิภาพที่ใกล้เคียง ซึ่งเราสามารถเขียนลำดับขั้นตอนการทำงานของเอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) แบบซูโดโค้ด (Pseudo Code) ที่ใช้เป็นตัวแทนของอัลกอริทึมได้ดังนี้

Algorithm 2 Gradient boosting.

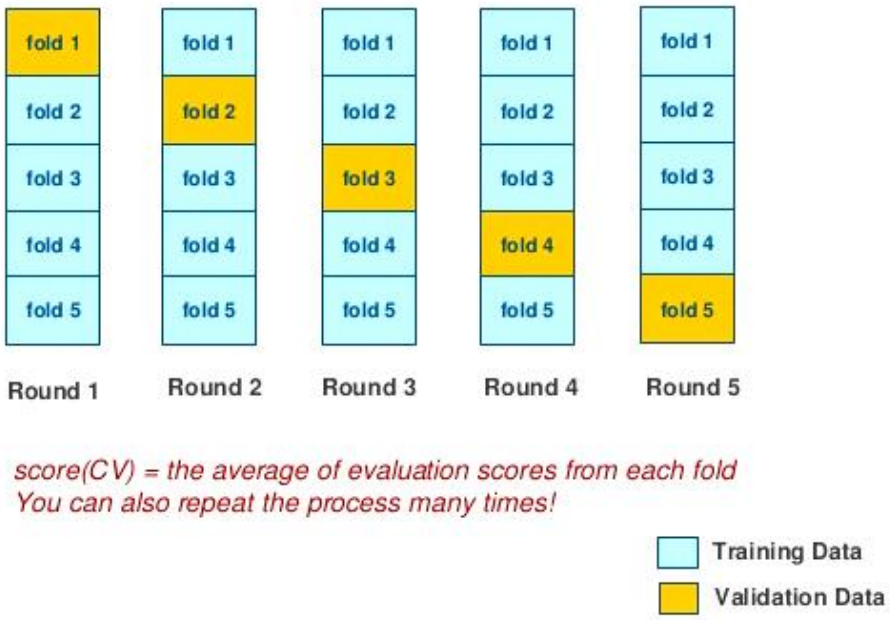
let F_0 be a "dummy" constant model
for $m = 1, \ldots, M$
  for each pair (x_i, y_i) in the training set
    compute the *pseudo-residual* $R(y_i, F_{m-1}(x_i))$ = negative gradient of the loss
  train a regression sub-model h_m on the pseudo-residuals
  add h_m to the ensemble: $F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x)$
return the ensemble F_M

การประเมินผลโมเดล

ครอสวาลิเดชั่น (Cross Validation) (Witten, et al., 2005; Han and Kamber, 2006) คือ วิธีการในการคาดการค่าความผิดพลาดของโมเดลหรือ วิธีการที่เรานำเสนอโดยพื้นฐานของวิธีการครอสวาลิเดชั่นคือ การสุ่มตัวอย่าง โดยเริ่มจากแบ่งชุดข้อมูลออกเป็นส่วน ๆ และนำบางส่วนจากชุดข้อมูลนั้นมาตรวจสอบ ผลลัพธ์จากการทำครอสวาลิเดชั่นมักถูกใช้เป็นตัวเลือกในการกำหนดโมเดล เช่น สถาปัตยกรรมเครือข่ายการสื่อสาร (Network Architecture) โมเดลในการคัดแยกประเภท (Classification Model) นั้นจะต้องมีการแบ่งข้อมูลออกเป็นชุดสอนและชุดทดสอบ แต่ในบางครั้งอาจเกิดปัญหาจากการเลือกข้อมูลที่ดีและง่ายมาเป็นข้อมูลชุดทดสอบทำให้ผลการ Classify นั้นดีเกินจริง ดังนั้นจะมีการคิดวิธี K-Fold Cross Validation ขึ้นมาแก้ปัญหา การตรวจสอบความถูกต้องของการเรียนรู้ สามารถสังเกตได้จากค่าความแม่นยำหรือค่าความผิดพลาดที่ได้จากการทำครอสวาลิเดชั่น ระหว่างการแบ่งชุดทดสอบในการเรียนรู้ ในที่นี้ผู้วิจัยเลือกพิจารณาความแม่นยำ ซึ่งเป็นสัดส่วนร้อยละของการคัดแยกได้ถูกต้องต่อจำนวนตัวอย่างทั้งหมด โดยใช้การครอสวาลิเดชั่นแบบ K-fold ซึ่งพื้นฐานของวิธีการครอสวาลิเดชั่นคือ การสุ่มตัวอย่างข้อมูลไปเป็นชุดฝึกฝนและชุดทดสอบแบบสลับกัน โดยเริ่มจากแบ่งชุดข้อมูลออกเป็นส่วนๆ และนำบางส่วนจากชุดข้อมูลนั้นมาเป็นชุดทดสอบเพื่อจำลองผลลัพธ์จากการทำ ครอสวาลิเดชั่น การแบ่งชุดทดสอบแบบนี้มักถูกใช้เป็นตัวเลือกในการกำหนดโมเดล โดยสังเกตจากผลของการทดสอบในแต่ละครั้งของการปรับค่าพารามิเตอร์แต่ละโมเดล การทำครอสวาลิเดชั่น

วิธี K-fold ครอสวาลิเดชั่นเป็นการแบ่งข้อมูลออกเป็น K ชุด เท่า ๆ กัน และทำการคำนวณค่าความผิดพลาด K รอบ โดยแต่ละรอบการคำนวณ ข้อมูลชุดหนึ่งจากข้อมูล K ชุดจะถูกเลือกออกมาเพื่อเป็นข้อมูลทดสอบ และข้อมูลอีก K-1 ชุดจะถูกใช้เป็นข้อมูลสำหรับการเรียนรู้ แล้วนำผลความถูกต้องหรือค่าความผิดพลาดของแต่ละรอบมารวมกันและหาค่าเฉลี่ยเพื่อเป็นค่าสะท้อนประสิทธิภาพของการฝึกฝน ดังตัวอย่างในภาพประกอบ 2-9 เป็นการกำหนดค่า K =5 หมายถึง การนำข้อมูลทั้งหมดมาแบ่งเป็น 5 ชุด และจะดำเนินการเรียนรู้และทดสอบจำนวน 5 รอบ โดยในแต่ละรอบจะเลือกข้อมูลมา 1 ชุดไม่ซ้ำกันมาเป็นชุดทดสอบ ด้วยการเรียนรู้ของชุดเรียนรู้ที่เหลือ เมื่อทำครบ 5 รอบ หาผลรวมของค่าความผิดพลาดหรือความแม่นยำจากชุดทดสอบในแต่ละรอบแล้วหารด้วยจำนวนรอบ ผลที่ได้เป็นค่าเฉลี่ยของความผิดพลาดในการเรียนรู้ของโมเดล ข้อดีของวิธีการนี้คือ ข้อมูลในแต่ละชุดที่ทำการแบ่งจะถูกทดสอบอย่างน้อย 1 ครั้ง และถูกเรียนรู้ทั้งหมด K-1 ครั้ง โดยในขั้นตอนเหล่านี้สามารถกำหนดได้ว่าต้องการขนาดข้อมูลขนาดใด และต้องการทำการ

คำนวณเป็นจำนวนรอบเท่าใด ซึ่งเหมาะสำหรับการประมวลผลทดสอบกับข้อมูลที่มีมิติขนาดจำนวนมาก



ภาพประกอบ 15 ตัวอย่างการแบ่งชุดข้อมูลของ k-fold

การประเมินความสามารถของระบบการจำแนกประเภทนั้น เน้นความสามารถในการตัดสินใจหรือการจำแนกหมวดหมู่ที่ถูกต้องของวิธีการเพื่อสร้างแบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง วิธีการทดสอบเพื่อเปรียบเทียบประสิทธิภาพของโมเดลการพัฒนาประสิทธิภาพการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่องนั้น สามารถพิจารณาได้จากค่าความถูกต้อง โดยวัดที่ประสิทธิผลของการจำแนกข้อมูลตามแนวคิดทางด้านการค้นคืนสารสนเทศ ซึ่งก็คือการวัดค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) ค่า F-Measure ซึ่งคำนวณได้ดังตารางการณ์ (Contingency Table) ได้ดังนี้

		Predicted class: P	Predicted class: N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

ภาพประกอบ 16 ตาราง Confusion matrix

TP = จำนวนตัวอย่างที่อยู่ในกลุ่ม C_j และตัวจำแนกทำนายว่าอยู่ในกลุ่ม C_j

FP = จำนวนตัวอย่างที่ไม่อยู่ในกลุ่ม C_j และตัวจำแนกทำนายว่าอยู่ในกลุ่ม C_j

FN = จำนวนตัวอย่างที่อยู่ในกลุ่ม C_j และตัวจำแนกทำนายว่าไม่อยู่ในกลุ่ม C_j

TN = จำนวนตัวอย่างที่ไม่อยู่ในกลุ่ม C_j และตัวจำแนกทำนายว่าไม่อยู่ในกลุ่ม C_j

C_j = กลุ่มประเภทของการตรวจสอบการทุจริตที่สนใจวัดประสิทธิภาพ

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

$$\text{Precision (p)} = \frac{TP}{TP + FP}$$

$$\text{Recall (r)} = \frac{TP}{TP + FN}$$

$$\text{F-measure (F)} = \frac{2rp}{r + p} = \frac{2TP}{2TP + FN + FP}$$

บทที่ 3

วิธีดำเนินงานวิจัย

งานวิจัยนี้มีวัตถุประสงค์ที่จะศึกษาออกแบบขั้นตอนวิธีการเพื่อสร้างแบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง โดยวัดประสิทธิภาพในด้านความถูกต้องในการจำแนก (Accuracy) ค่า F-Measure ค่าความแม่นยำ (Precision) ค่าเรียกกลับ (Recall) ค่า ROC curve (Receiver Operating Characteristic Curve) และ ปัจจัยด้านเวลาในการประมวลผล (Runtime) ตารางการวัดประสิทธิภาพการจำแนก (Confusion Matrix) ซึ่งผู้วิจัยได้แบ่งวิธีการดำเนินงานออกเป็น ขั้นตอนดังนี้

ข้อมูลผิดปกติ (Outlier)

การตรวจสอบข้อมูลก่อนการวิเคราะห์ออกแบบขั้นตอนวิธีการเพื่อสร้างแบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่องนั้น ถือเป็นเรื่องที่ดี เพราะข้อมูลที่ได้จากกลุ่มตัวอย่าง หรือกลุ่มเป้าหมายนั้น แสดงผลออกในรูปผลลัพธ์ของการวิจัย หากละเลยไม่มีการตรวจสอบก็อาจทำให้ผลการวิจัยคลาดเคลื่อน หรือมีความน่าเชื่อถือน้อยลง การเก็บข้อมูลการวิจัยแต่ละครั้งนักวิจัยอาจประสบปัญหา เช่น การได้ข้อมูลมาไม่ครบถ้วนอันจะส่งผลทำให้เกิดปรากฏการณ์ ความไม่ครบถ้วนหรือภาวะข้อมูลสูญหาย (Missing Data) และอีกปรากฏการณ์หนึ่งที่เกิดขึ้นได้หากไม่มีการตรวจสอบก่อนการวิเคราะห์ข้อมูลคือ การรายงานผลลัพธ์โดยมีค่าสุดโต่งแฝง (Outlier) ในการวิเคราะห์ข้อมูลนั้น โดยธรรมชาติของการวิจัยจะเป็นการวัดพฤติกรรมผ่านตัวแปรแฝง หรือพูดอีกอย่างคือการวัดพฤติกรรมทางอ้อม จะเห็นได้ว่าหากมีการบวนการเก็บข้อมูลที่ไม่ดี จะทำให้ข้อมูลที่ได้มาตอบคำถามไม่ตรงประเด็น เสียเวลา และเปล่าประโยชน์ เพราะการเก็บข้อมูลที่ผ่านกระบวนการสุ่มนั้น ค่าความคลาดเคลื่อนที่นักวิจัยยอมรับได้ คือ ค่าความคลาดเคลื่อนที่เกิดอย่างสุ่ม สามารถสรุปเหตุการณ์ที่จะพบโดยส่วนมากมีอยู่ 2 ประเด็นคือ ความไม่ครบถ้วนของข้อมูลและค่าสุดโต่ง (Missing Data and Outlier) แยกสรุปได้ ดังนี้

1. ความไม่ครบถ้วนของข้อมูล (Missing Data) ความไม่ครบถ้วนหรือการไม่สมบูรณ์ของข้อมูลนั้น เกิดได้หลายปัจจัย เกิดได้จากการละเลยหรือความบังเอิญ ข้อมูลที่ขาดหายไปจะส่งผลต่อการอ้างอิงกลับไปยังกลุ่มประชากร หากเกิดการขาดหายไปเยอะ นักวิจัยอาจแก้ปัญหาด้วยการเก็บ

ข้อมูลซ้ำ หรือหากขาดหลายไปบางช่วงบางตอน นักวิจัยอาจพิจารณาเป็นกรณีไป เช่น การตัดตัวแปรที่มีปัญหาออก การแทนที่ด้วยค่าเฉลี่ยของชุดข้อมูลนั้น ๆ การละเลยโดยให้โปรแกรมวิเคราะห์จัดการให้ เป็นต้น

2. ค่าสุดโต่งของข้อมูล (Outlier) จัดเป็นข้อมูลผิดปกติ จัดเป็นข้อมูลที่มีค่าแยกออกจากกลุ่มหรือผิดแผกแตกต่างไปจากข้อมูลค่าอื่น ๆ ซึ่งค่าผิดปกติมีโอกาสเกิดขึ้นได้บนพื้นฐานของเหตุผลคือ การจดบันทึกหรือเก็บข้อมูลมีความคลาดเคลื่อน หรือ กลุ่มตัวอย่างที่เก็บรวบรวมข้อมูลมา มีความแตกต่างไปจากกลุ่มจริง ซึ่งการเกิดค่าผิดปกติประการแรกนั้น สามารถเกิดขึ้นได้เสมอ จึงควรมีการตรวจสอบข้อมูลให้ถูกต้องก่อนวิเคราะห์ จะเห็นได้ว่าค่าสุดโต่งนี้ ก็มีผลต่อผลลัพธ์ทางการวิจัยอย่างมาก หากละเลยที่จะตรวจสอบค่าสุดโต่งจะทำให้ผลการวิจัยคลาดเคลื่อนและสรุปผลได้ไม่ถูกต้อง

งานวิจัยนี้มีวัตถุประสงค์ที่จะศึกษาออกแบบขั้นตอนวิธีการเพื่อสร้างแบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง โดยใช้เทคนิคการกรองข้อมูลผิดปกติเป็นการจำแนกความผิดปกติ โดยใช้ค่าพิสัยตรวจจับกลุ่มตัวอย่างที่ผิดปกติโดยการใช้ k-Nearest Neighbors (K-NN) ซึ่งข้อมูลจะเป็นข้อมูลในลักษณะกลุ่มส่วนการกรองข้อมูลด้วยเทคนิคต่างๆ ในเหมืองข้อมูลก็ได้มีการทำกันอย่างแพร่หลาย และปรากฏผลที่ดี

ข้อมูลที่ไม่สมดุล (Imbalanced Data)

ข้อมูลไม่สมดุลเป็นข้อมูลที่สามารถพบเจอได้จริงในชีวิตประจำวัน เช่น ข้อมูลการวินิจฉัยโรคบางประเภท การตรวจจับการทุจริตการใช้บัตรเครดิต เมื่อนำข้อมูลเหล่านี้มาใช้งานทางด้านการเรียนรู้ของเครื่องจักรและการทำเหมืองข้อมูลจะส่งผลกระทบต่อการเรียนรู้ของอัลกอริทึม เนื่องจากข้อมูลที่ใช้ในการเรียนรู้มีและเป็นกลุ่มที่ให้ความสนใจมีจำนวนข้อมูลที่น้อยมากเมื่อเทียบกับกลุ่มอื่นๆ ที่เหลือ อัลกอริทึมทางด้านการเรียนรู้ของเครื่องจักรนั้นสามารถทำงานได้ดีในกรณีที่ข้อมูลสมดุล สำหรับข้อมูลไม่สมดุลนั้นขอบเขตของการตัดสินใจของอัลกอริทึมการเรียนรู้ของเครื่องจักรนั้นจะมีความเอนเอียงไปทางกลุ่มข้อมูลส่วนมากส่งผลให้การจัดกลุ่มของข้อมูลส่วนน้อยมีแนวโน้มที่จะได้รับการจัดกลุ่มที่ผิดประเภท ลักษณะโดยทั่วไปของข้อมูลไม่สมดุล คือ ข้อมูลที่มีจำนวนข้อมูลของกลุ่มหนึ่งมากกว่าจำนวนข้อมูลของกลุ่มที่เหลือเป็นจำนวนมาก ซึ่งข้อมูลไม่สมดุลนี้จะส่งผลกระทบต่อการจำแนกประเภทข้อมูลทำให้ไม่สามารถจำแนกประเภทข้อมูลของกลุ่มที่มีจำนวนข้อมูลน้อยได้ถูกต้องแม่นยำ ในขณะเดียวกันจะสามารถจำแนกประเภทข้อมูลของกลุ่มที่มีจำนวนมากได้อย่างแม่นยำ โดยทั่วไปข้อมูลกลุ่มที่มีจำนวนมากจะถูก

เรียกว่า คลาสส่วนมาก (Majority Class หรือ Negative Class) และข้อมูลกลุ่มที่มีจำนวนน้อยจะถูกเรียกว่า คลาสส่วนน้อย (Minority Classหรือ Positive Class) (Farquad, 2012)

ซึ่งข้อมูลที่อยู่ในคลาสส่วนน้อยจะเป็นข้อมูลที่งานวิจัยนี้ให้ความสำคัญมากกว่าข้อมูลที่อยู่ในคลาสส่วนมาก ข้อมูลไม่สมดุลนั้นสามารถพบเห็นได้ทั่วไป ซึ่งสาเหตุของการเกิดความไม่สมดุลนั้นอาจจะมาจากหลายสาเหตุ เช่น ข้อมูลไม่สมดุลที่เกิดจากธรรมชาติของข้อมูลเองซึ่งสามารถพบเจอได้ในข้อมูลการวินิจฉัยทางด้านการแพทย์ ที่มีข้อมูลผู้ป่วยด้วยโรคร้ายแรงน้อยกว่าข้อมูลของผู้ที่มีสุขภาพดีเป็นจำนวนมาก ข้อมูลของบัตรเครดิตที่มีข้อมูลลูกค้าปกติมากกว่าลูกค้าที่ผิดปกติ ข้อมูลการตรวจจับผู้บุกรุกของเครือข่ายข้อมูล หรือข้อมูลไม่สมดุลอาจจะเกิดจากข้อจำกัดในการจัดเก็บเช่น ค่าใช้จ่ายที่สูงมาก อันตรายที่เกิดจากการรวบรวมข้อมูลเป็นต้น

ดังนั้นในงานวิจัยนี้ใช้ วิธีการสุ่มตัวอย่าง (Sampling Methods) โดยการเลือกสมาชิกจากกลุ่มตัวอย่างโดยพยายามทำให้สมาชิกที่เลือกมาเหล่านั้น เป็นตัวแทนที่ดีของกลุ่มตัวอย่างทั้งหมด การที่จะเลือกตัวอย่างให้เป็นตัวแทนที่ดีของกลุ่มตัวอย่างทั้งหมด ได้นั้น จะต้องทำการเลือกแบบสุ่ม (random) หรือเลือกอย่างไม่ลำเอียง (Unbias) คือ พยายามให้สมาชิกแต่ละตัวของกลุ่มตัวอย่างทั้งหมด มีโอกาสที่จะได้รับการเลือกเป็นตัวแทนเท่า ๆ กัน สำหรับขั้นตอนวิธีที่ใช้สำหรับการแบ่งกลุ่มข้อมูลที่ไม่สมดุลนี้ จะเป็นการประยุกต์เอาวิธีการสุ่มตัวอย่างซึ่งเป็นวิธีการทางสถิติ เพื่อสร้างข้อมูลสำหรับการสอน โดยมีจุดประสงค์เพื่อให้จำนวนสมาชิกในข้อมูลทั้งสองกลุ่มมีความสมดุลกัน โดยใช้วิธี Oversampling จะทำการสุ่มข้อมูลในกลุ่มรองเพื่อสร้างข้อมูลใหม่ของกลุ่มรองให้มีจำนวนเพิ่มมากขึ้น ให้ใกล้เคียงหรือเท่ากับจำนวนข้อมูลในกลุ่มหลัก โดยในงานวิจัยนี้ใช้เทคนิคการแก้ปัญหาโดยปรับความสมดุลของข้อมูลด้วยเทคนิคการสุ่มเพิ่มตัวอย่างเพิ่ม (Synthetic Minority Oversampling Technique) วิธี Synthetic Minority Oversampling Technique จัดเป็นเทคนิคที่ ใช้ในการแก้ปัญหาที่ต้องการจำแนกข้อมูลไม่สมดุล ซึ่งข้อมูลจำนวนตัวอย่างแตกต่างกันมากในแต่ละคลาส เมื่อทำการจำแนกประเภท จะทำให้ตัวโมเดลมีการเรียนรู้แต่ในข้อมูลกลุ่มที่มาก ทำให้เกิดการจดจำลักษณะรูปแบบของข้อมูลกลุ่มนี้มากไปจนเกิด Overfitting กับโมเดล ผลที่ได้ก็จะจำแนกไปในข้อมูลกลุ่มมากไปหลัก Synthetic Minority Oversampling Technique จัดเป็นวิธีการเพิ่มจำนวนข้อมูลประเภทที่มีข้อมูลน้อยให้เพิ่มปริมาณข้อมูลให้ใกล้เคียงกับประเภทที่มีมากที่สุด โดยสุ่มค่าขึ้นมาหนึ่งค่าและหาค่าระยะห่างระหว่างค่าที่ เลือกกับทุกๆค่า โดยเลือกค่าที่ใกล้เคียงที่สุด ดังสมการ

$$X_{new} = x_i + (\hat{x}_i - x_i) \times \delta$$

x_{new} คือ ข้อมูลใหม่

x_i คือ ข้อมูลที่สุ่มในตอนแรก

$\hat{x}_i$ คือ ข้อมูลที่สุ่มเพิ่มขึ้น

δ คือ ค่าที่สุ่มช่วงระหว่าง 0-1

การลดมิติข้อมูล (Dimensionality Reduction)

เนื่องจากกลุ่มตัวอย่างของการตรวจสอบการทุจริตมีแนวโน้มที่จะเพิ่มปริมาณสูงขึ้น ทุกวัน ทำให้ข้อมูลการตรวจสอบการทุจริตบางประเภทมีจำนวนคุณลักษณะมากขึ้นซึ่งจำนวนคุณลักษณะที่มากนี้ มีผลต่อประสิทธิภาพของการจำแนกประเภทของการตรวจสอบการทุจริต เนื่องจากแบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง โดยทั่วไปไม่สามารถรองรับการทำงานกับจำนวนคุณลักษณะของข้อมูลที่สูงมากได้ดี การลดขนาดข้อมูลจึงเป็นขั้นตอนหนึ่ง ที่จะต้องทำก่อนการเรียนรู้ด้วยการเรียนรู้ของเครื่อง (Machine learning) แต่การลดมิติข้อมูล ต้องพิจารณาด้วยความระมัดระวัง เนื่องจากมีความเสี่ยงในการที่จะกำจัดคุณลักษณะที่สำคัญต่อการจำแนกประเภทของการตรวจสอบการทุจริตออกไป จากการศึกษาพบว่าวิธีการลดคุณลักษะมีหลากหลายวิธี ขึ้นอยู่บนพื้นฐานของการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรกับกลุ่มเป้าหมาย (Class) รวมถึงการสร้างคุณลักษณะใหม่จากคุณลักษณะเดิม อาจจะนำคุณลักษณะพื้นฐานเหล่านี้มารวมกันเพื่อให้เป็นคุณลักษณะใหม่ ซึ่งงานวิจัยนี้ได้ใช้ค่าสถิติความสัมพันธ์ (Correlation Based Feature Selection) อัลกอริธึมนี้ มีหลักการกรองที่ง่าย ๆ โดยค่าสถิติความสัมพันธ์ จะจัดอันดับกลุ่มย่อยของมิติข้อมูล ตามความสัมพันธ์ที่อยู่บนพื้นฐานของฟังก์ชันการประมาณแบบการแก้ปัญหา (Heuristic) ซึ่งกลุ่มย่อยของมิติข้อมูล จะมีความสัมพันธ์กันสูงกับคลาส และไม่มีความสัมพันธ์กับคลาสอื่น ๆ สำหรับมิติข้อมูลที่ไม่เกี่ยวข้องอาจจะถูกลดทอนน้ำหนักลง เพราะมิติข้อมูลเหล่านี้อาจจะมีความสัมพันธ์ต่ำกับกลุ่มเป้าหมาย มิติข้อมูลที่ซ้ำซ้อนอาจจะถูกขจัดออกไป เหลือแต่กลุ่มมิติข้อมูลที่มีความสัมพันธ์สูง สามารถนิยามได้โดย

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}$$

โดยที่ M_i คือ ค่าที่ค้นหาได้ของมิติข้อมูลกลุ่มย่อย S ซึ่งประกอบด้วยมิติข้อมูล k

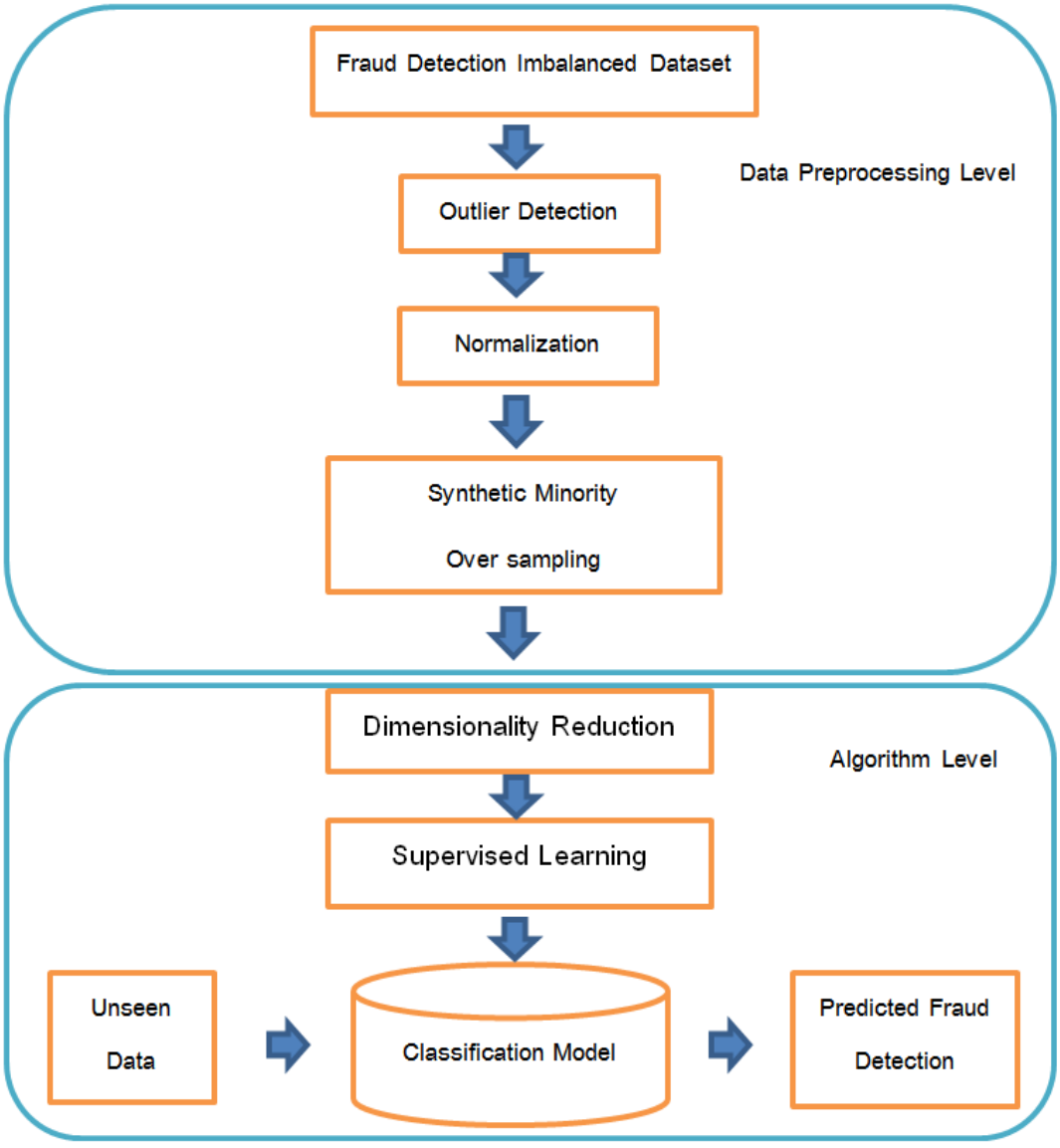
$\overline{r_{cf}}$ คือ ค่าเฉลี่ยความสัมพันธ์ของตัวแปรกับคลาส $(f \in S)$

$\overline{r_{ff}}$ คือ ค่าเฉลี่ยความสัมพันธ์ระหว่างมิติของข้อมูล

ขั้นตอนวิธีการสร้างการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุล

จากการศึกษาและวิเคราะห์ปัญหางานวิจัยที่เกี่ยวข้องกับสร้างแบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่องดังกล่าว ในส่วนนี้ผู้วิจัยขอนำเสนอขั้นตอนวิธีการสร้างแบบจำลอง เพื่อการจำแนกการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุล โดยมีวัตถุประสงค์ในการเพิ่มประสิทธิภาพในจำแนกการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุล โดยแบบจำลองที่นำเสนอในงานวิจัยนี้ ใช้การตรวจสอบข้อมูลก่อนการวิเคราะห์ เพื่อกำจัดข้อมูลค่าสุดโต่งแฝง(Outlier) ออกก่อน จากนั้นทำการแปลงข้อมูล (Data Transformation) ด้วยวิธีการนอร์มอลไลซ์ (Normalization) ทำการแปลงค่าข้อมูลให้อยู่ในช่วงสั้นๆ โดยแปลงข้อมูลเชิงเส้นจากช่วงที่เป็นไปได้เดิมของค่าอินพุต ให้เป็นช่วงข้อมูลใหม่ที่กำหนดขึ้นให้มีสเกลที่เท่ากัน โดยการปรับการกระจายของข้อมูล และค่าเบี่ยงเบนมาตรฐาน (Z-Score Normalization) แล้วจึงทำการปรับปรุงชุดข้อมูลกลุ่มตัวอย่างของการตรวจสอบการทุจริตไม่สมดุลด้วยวิธีการสุ่มตัวอย่าง (Sampling Methods) โดยวิธีการ Over-sampling จะทำการสุ่มข้อมูลในกลุ่มรองเพื่อสร้างข้อมูลใหม่ของกลุ่มรองให้มีจำนวนเพิ่มมากขึ้นให้ใกล้เคียงหรือเท่ากับจำนวนข้อมูลในกลุ่มหลัก จากนั้นทำการลดมิติข้อมูลของกลุ่มตัวอย่างลงโดยพิจารณาคุณลักษณะที่มีค่าสถิติความสัมพันธ์กับกลุ่มเป้าหมาย สูง จะให้ถ่วงน้ำหนักมาก ก่อนส่งเข้าเครื่องจักรการเรียนรู้แบบมีผลเฉลย (Supervised Learning) ซึ่งประกอบด้วย อัลกอริทึม 1. เนอีฟเบย์ (Naïve Bayes) 2. ตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) 3. การวิเคราะห์การถดถอยโลจิสติก (Logistic Regression) 4. การเรียนรู้เชิงลึก (Deep Learning) 5. ต้นไม้ตัดสินใจ (Decision Tree) 6. แรนดอมฟอเรส (Random Forest) 7. ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) 8. เคเนียเรสเนเบอร์ (K-Nearest Neighbor) 9. โครงข่ายประสาทเทียมพอร์เซฟตรอนแบบหลายชั้น (Multi-Layer Perceptron Neural Networks) 10. เอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) โดยอัลกอริทึมทั้งหมดใช้ค่าพารามิเตอร์มาตรฐานมาทำการเรียนรู้แล้วทำการทดสอบเปรียบเทียบประสิทธิภาพของแบบจำลอง การประเมินความสามารถของ

แบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่องนั้น เน้นความสามารถในการตัดสินใจหรือ การจำแนกการตรวจสอบการทุจริตที่ถูกต้อง วิธีการทดสอบเพื่อเปรียบเทียบประสิทธิภาพของโมเดล โดยพิจารณาด้านความถูกต้อง (Accuracy) ค่า F-Measure ค่าความแม่นยำ (Precision) ค่าเรียกกลับ (Recall) ค่า ROC curve (Receiver Operating Characteristic Curve) และปัจจัยด้านเวลาในการประมวลผล (Runtime) โดยทำการทดสอบแบบจำลองด้วยวิธี 10-ครอสวาลิเดชั่น (10-Cross Validation) ซึ่งเป็นวิธีที่เป็นมาตรฐานในการทดสอบแบบจำลองด้านการเรียนรู้แบบมีผลเฉลย

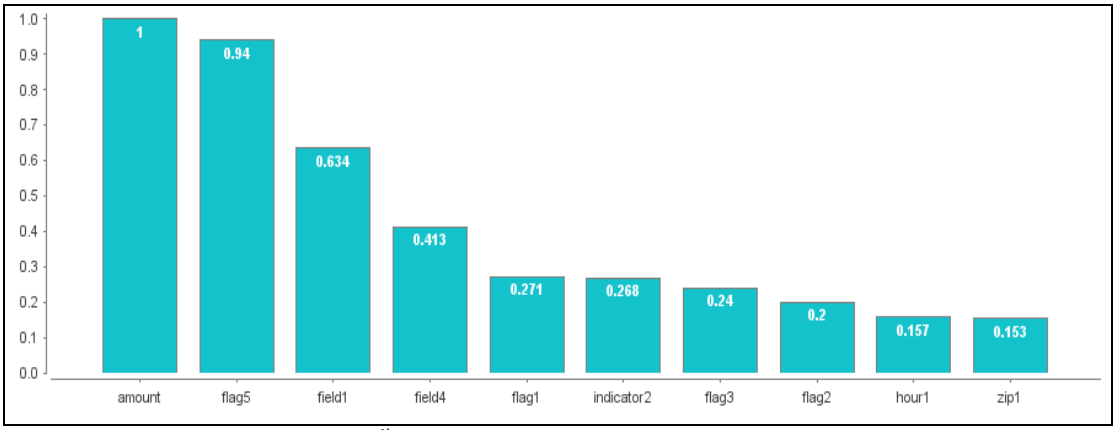


ภาพประกอบ 17 แบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง

บทที่ 4

ผลการทดลอง

งานวิจัยนี้มีวัตถุประสงค์เพื่อสร้างแบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง มุ่งเน้นความถูกต้องของการจำแนกการตรวจสอบการทุจริตเป็นหลัก โดยที่จำลองนี้ใช้ทรัพยากรของระบบและหน่วยความจำอย่างเหมาะสม ทำการทดสอบกับกลุ่มตัวอย่างฐานข้อมูลการทำธุรกรรมผ่านช่องทางพาณิชย์อิเลคทรอนิคส์ขนาดใหญ่ ซึ่งเป็นฐานข้อมูลมาตรฐานของแล็บวิจัยมหาวิทยาลัยแคลิฟอร์เนียซานดิเอโก (University of California San Diego :UCSD) Transaction Fraud Detection Dataset และใช้ในการแข่งขันจาก Data Mining Contest ซึ่งมีจำนวนคุณลักษณะที่ใช้ในการวิเคราะห์ทั้งหมด 16 ตัวแปร จำนวนรายการเท่ากับ 97,346 ระเบียน โดยทำการทดลองการลดคุณลักษณะโดยใช้ ค่าความสัมพันธ์ (Correlation) เพื่อลดมิติของข้อมูลแล้วส่งเข้าการเรียนรู้ของเครื่อง (Machine Learning) ประเภทต่างๆ ประกอบด้วย เนอีฟเบย์ (Naïve Bayes) ตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) การวิเคราะห์การถดถอยโลจิสติก (Logistic Regression) การเรียนรู้เชิงลึก (Deep Learning) ต้นไม้ตัดสินใจ (Decision Tree) แรนดอมฟอเรส (Random Forest) ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) เคเนียเรสเนเบอร์ (K-Nearest Neighbor) โครงข่ายประสาทเทียมพอร์เซฟตรอนแบบหลายชั้น (Multi-Layer Perceptron Neural Networks) เอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) และทำการทดสอบด้วยวิธี 10-fold cross validation สามารถสรุปผลการทดลอง ดังนี้



ภาพประกอบ 18 ค่าน้ำหนักความสัมพันธ์ (Correlation) ของแต่ละตัวแปร

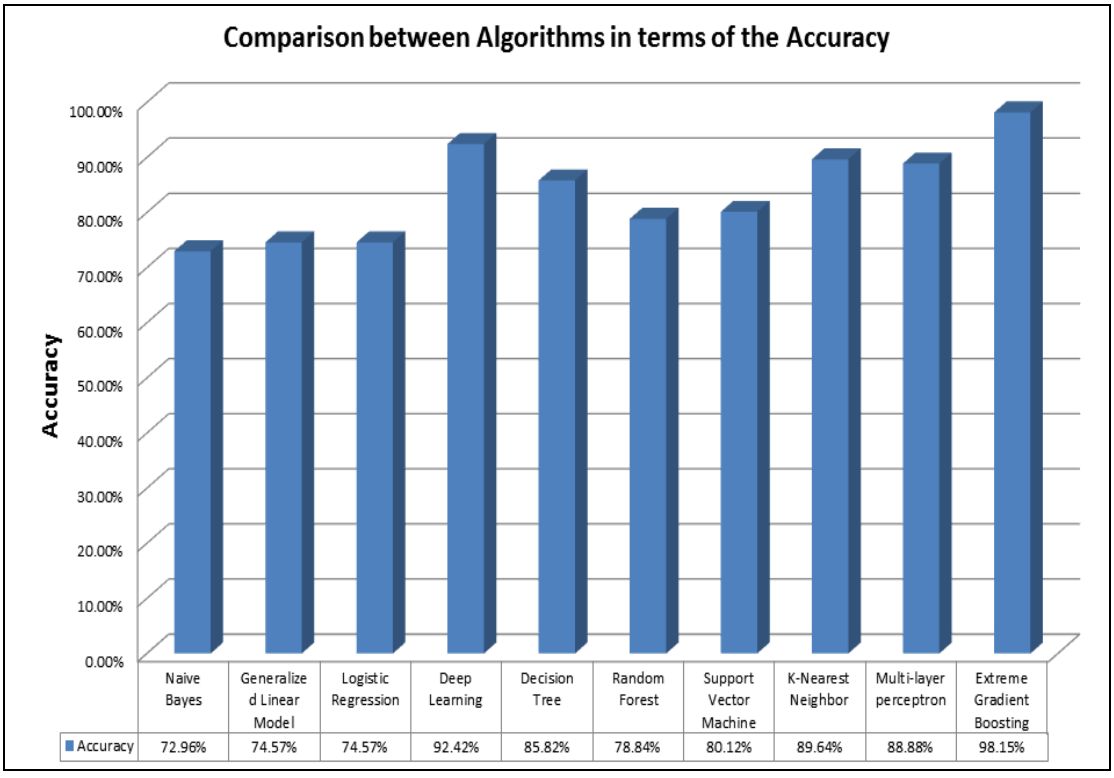
Attributes	amount	class = ...	field1	field2	field3	field4	flag1	flag2
amount	1	0.375	-0.295	-0.011	-0.051	0.114	-0.083	-0.018
class = 'not fraud'	0.375	1	-0.233	0.056	-0.021	0.154	-0.118	0.082
field1	-0.295	-0.233	1	0.077	0.069	-0.035	-0.099	0.130
field2	-0.011	0.056	0.077	1	0.034	0.059	-0.117	0.052
field3	-0.051	-0.021	0.069	0.034	1	0.008	0.075	-0.049
field4	0.114	0.154	-0.035	0.059	0.008	1	-0.396	0.020
flag1	-0.083	-0.118	-0.099	-0.117	0.075	-0.396	1	-0.112
flag2	-0.018	0.082	0.130	0.052	-0.049	0.020	-0.112	1
flag3	-0.138	-0.102	0.194	0.028	-0.040	-0.007	-0.084	0.504
flag4	0.006	0.003	-0.015	-0.016	-0.053	-0.086	0.105	-0.047
flag5	-0.396	-0.351	0.368	0.080	-0.027	-0.110	-0.158	0.257
hour1	0.058	0.062	-0.046	0.012	0.013	0.018	-0.018	-0.035
indicator1	-0.045	-0.052	0.006	0.009	0.150	-0.024	0.028	0.002
indicator2	-0.078	-0.102	0.087	-0.008	-0.034	-0.048	0.024	0.018
zip1	-0.048	-0.063	0.071	-0.034	-0.023	-0.025	0.026	-0.008

field3	field4	flag1	flag2	flag3	flag4	flag5	hour1	indicato...	indicator2
-0.051	0.114	-0.083	-0.018	-0.138	0.006	-0.396	0.058	-0.045	-0.078
-0.021	0.154	-0.118	0.082	-0.102	0.003	-0.351	0.062	-0.052	-0.102
0.069	-0.035	-0.099	0.130	0.194	-0.015	0.368	-0.046	0.006	0.087
0.034	0.059	-0.117	0.052	0.028	-0.016	0.080	0.012	0.009	-0.008
1	0.008	0.075	-0.049	-0.040	-0.053	-0.027	0.013	0.150	-0.034
0.008	1	-0.396	0.020	-0.007	-0.086	-0.110	0.018	-0.024	-0.048
0.075	-0.396	1	-0.112	-0.084	0.105	-0.158	-0.018	0.028	0.024
-0.049	0.020	-0.112	1	0.504	-0.047	0.257	-0.035	0.002	0.018
-0.040	-0.007	-0.084	0.504	1	-0.043	0.319	-0.046	-0.009	0.035
-0.053	-0.086	0.105	-0.047	-0.043	1	-0.033	-0.003	-0.012	0.007
-0.027	-0.110	-0.158	0.257	0.319	-0.033	1	-0.087	0.041	0.123
0.013	0.018	-0.018	-0.035	-0.046	-0.003	-0.087	1	0.008	-0.019
0.150	-0.024	0.028	0.002	-0.009	-0.012	0.041	0.008	1	-0.062
-0.034	-0.048	0.024	0.018	0.035	0.007	0.123	-0.019	-0.062	1
-0.023	-0.025	0.026	-0.008	0.006	-0.016	0.050	0.042	0.075	0.007

ภาพประกอบ 19 ตารางแสดงเมทริกซ์สหสัมพันธ์ (Correlation Matrix)

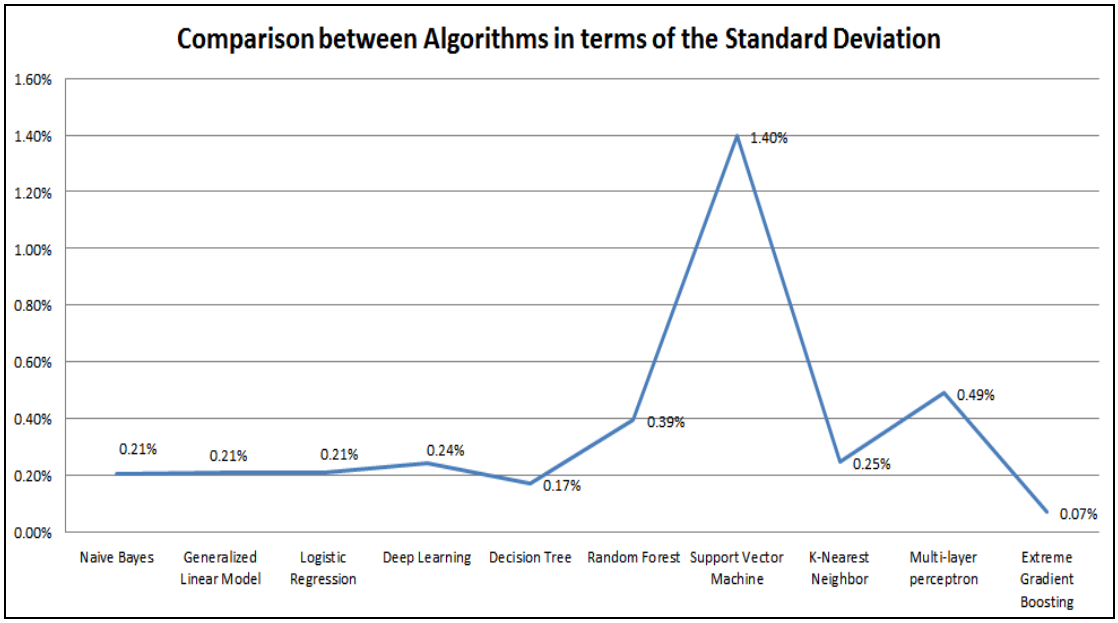
ผลการทดลอง

จากการทดลองทดสอบประสิทธิภาพของโมเดลการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง โดยทำการสุ่มตัวอย่าง (Sampling Methods) โดยการเลือกสมาชิกจากกลุ่มตัวอย่างโดยพยายามทำให้สมาชิกที่เลือกมาเหล่านั้น เป็นตัวแทนที่ดีของกลุ่มตัวอย่างทั้งหมด โดยทำการเลือกแบบสุ่ม (Random) หลังจากนั้นหาค่าสถิติความสัมพันธ์ (Correlation Based Feature Selection)) ระหว่างตัวแปรเพื่อใช้ลดมิติของข้อมูล ก่อนส่งเข้าเครื่องจักรการเรียนรู้แบบมีผลเฉลย (Supervised Learning) ประกอบด้วยอัลกอริทึม 1. เนอีฟเบย์ (Naïve Bayes) 2. ตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) 3. การวิเคราะห์การถดถอยโลจิสติก (Logistic Regression) 4. การเรียนรู้เชิงลึก (Deep Learning) 5. ต้นไม้ตัดสินใจ (Decision Tree) 6. แรนดอมฟอเรส (Random Forest) 7. ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) 8. เคเนียเรสเนเบอร์ (K-Nearest Neighbor) 9. โครงข่ายประสาทเทียมพอร์เซฟตรอนแบบหลายชั้น (Multi-Layer Perceptron Neural Networks) 10. เอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) โดยทำการวัดประสิทธิภาพจากค่าความถูกต้อง (Accuracy) สามารถสรุปในแต่ละประเด็นได้ว่า



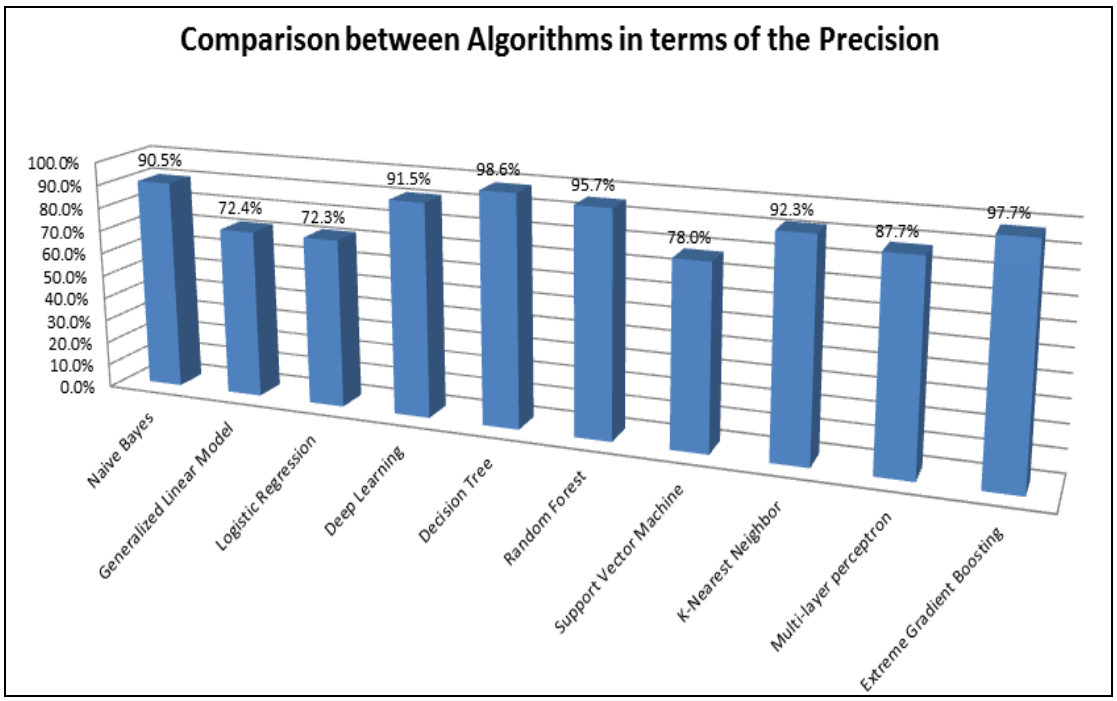
ภาพประกอบ 20 กราฟเปรียบเทียบประสิทธิภาพด้านความถูกต้องของแต่ละอัลกอริทึม

อัลกอริทึมที่ให้ประสิทธิภาพการจำแนกการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่องออกมาดีที่สุด เมื่อปเรียบเทียบกับวิธีการอื่นๆ และเมื่อพิจารณาค่าพารามิเตอร์ที่ส่งผลให้ประสิทธิภาพของโมเดลที่ดีสุด คือ อัลกอริทึมเอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) ให้ค่าความถูกต้อง (Accuracy) สูงที่สุดคือ 98.15 % รองลงมาเป็นอัลกอริทึมการเรียนรู้เชิงลึก (Deep Learning) ให้ค่าความถูกต้อง 92.42% อัลกอริทึมเคเนียเรสเนเบอร์ (K-Nearest Neighbor) ให้ค่าความถูกต้อง 89.64% อัลกอริทึมโครงข่ายประสาทเทียมพอร์เซฟตรอนแบบหลายชั้น (Multi-Layer Perceptron Neural Networks) ให้ค่าความถูกต้อง 88.88% อัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) ให้ค่าความถูกต้อง 85.82% ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) ให้ค่าความถูกต้อง 80.12% อัลกอริทึมแรนดอมฟอเรส (Random Forest) ให้ค่าความถูกต้อง 78.84% อัลกอริทึมตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) และอัลกอริทึมการวิเคราะห์การถดถอยโลจิสติก (Logistic Regression) ให้ค่าความถูกต้องเท่ากันที่ 74.57 % และสุดท้ายอัลกอริทึมเนอีฟเบย์ (Naïve Bayes) ให้ค่าความถูกต้องน้อยที่สุด 72.96 % ตามลำดับ



ภาพประกอบ 21 กราฟเปรียบเทียบค่าเบี่ยงเบนมาตรฐานของแต่ละอัลกอริทึม

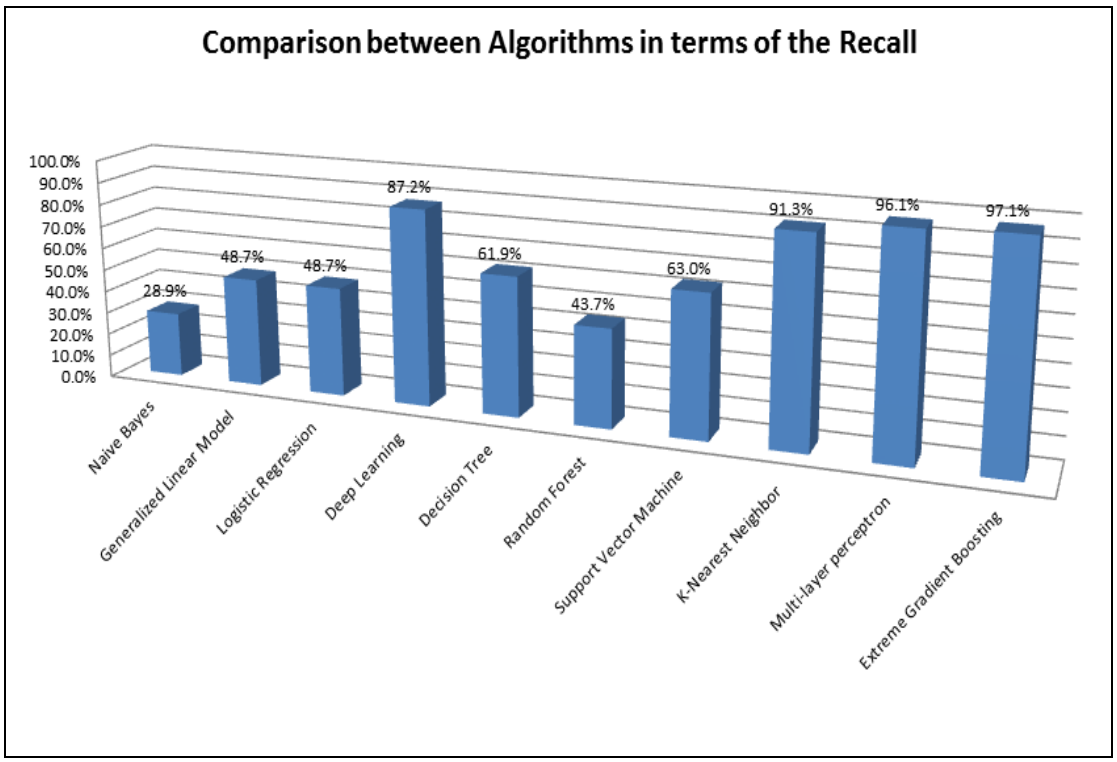
เมื่อพิจารณาถึงปัจจัยด้านค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) ของโมเดล พบว่า อัลกอริทึม เอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) ให้ค่าเบี่ยงเบนมาตรฐานน้อยที่สุด คือ 0.1% รองลงมาเป็นอัลกอริทึมเนอีฟเบย์ (Naïve Bayes) อัลกอริทึมตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) อัลกอริทึมการวิเคราะห์การถดถอยโลจิสติก (Logistic Regression) อัลกอริทึมการเรียนรู้เชิงลึก (Deep Learning) อัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) ให้ค่าเบี่ยงเบนมาตรฐานเท่ากันคือ 0.2% ถัดไปเป็นอัลกอริทึมเคเนียเรสเนเบอร์ (K-Nearest Neighbor) ให้ค่าเบี่ยงเบนมาตรฐาน 0.3% อัลกอริทึมแรนดอมฟอเรส (Random Forest) ให้ค่าเบี่ยงเบนมาตรฐาน 0.4% อัลกอริทึมโครงข่ายประสาทเทียมพอร์เซฟตรอนแบบหลายชั้น (Multi-Layer Perceptron Neural Networks) ให้ค่าเบี่ยงเบนมาตรฐาน 0.5% และสุดท้ายซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) ให้ค่าเบี่ยงเบนมาตรฐานสูงที่สุดคือ 1.4%



ภาพประกอบ 22 กราฟเปรียบเทียบประสิทธิภาพด้านความแม่นยำ (Precision)

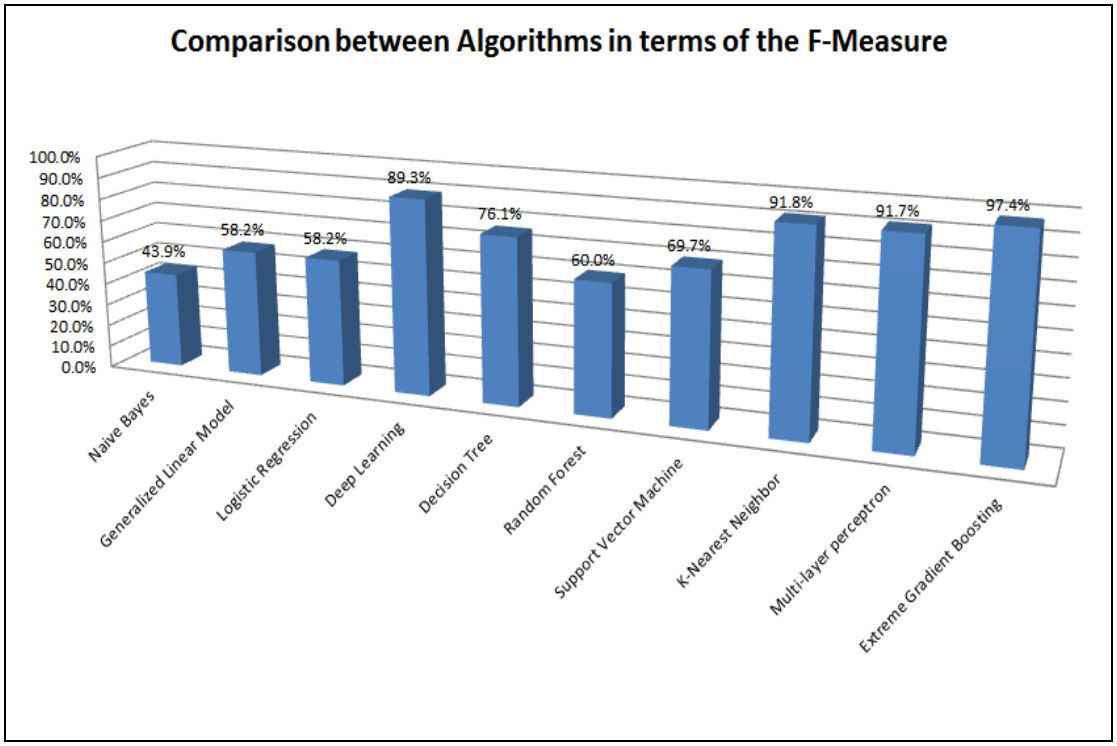
เมื่อทำการทดสอบประสิทธิภาพของโมเดล การจำแนกการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง ในด้านความแม่นยำ (Precision) กล่าวคือเป็นอัตราส่วนของการค้นพบข้อมูลที่ถูกต้องจากจำนวนข้อมูลทั้งหมดที่ทำการค้นคืนมาได้ พบว่า อัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) ให้ค่าความแม่นยำ สูงที่สุดคือ 98.6% รองลงมาเป็นอัลกอริทึมเอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) ให้ค่าความ

แม่นยำ 97.7 % อัลกอริทึมแรนดอมฟอเรส (Random Forest) ให้ค่าความแม่นยำ 95.7% อัลกอริทึมเคเนียเรสเนเบอร์ (K-Nearest Neighbor) ให้ค่าความแม่นยำ 92.3 % อัลกอริทึมการเรียนรู้เชิงลึก (Deep Learning) ให้ความแม่นยำ 91.5 % อัลกอริทึมเนอีฟเบย์ (Naïve Bayes) ให้ความแม่นยำ 90.5 % อัลกอริทึมโครงข่ายประสาทเทียมพอร์เซฟตรอนแบบหลายชั้น (Multi-Layer Perceptron Neural Networks) ให้ความแม่นยำ 87.7 % อัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) ให้ความแม่นยำ 78.0 % อัลกอริทึมตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) ให้ความแม่นยำ 72.4 % และสุดท้าย อัลกอริทึมการวิเคราะห์การถดถอยโลจิสติก (Logistic Regression) ให้ค่าความแม่นยำน้อยที่สุด 72.3 % ตามลำดับ



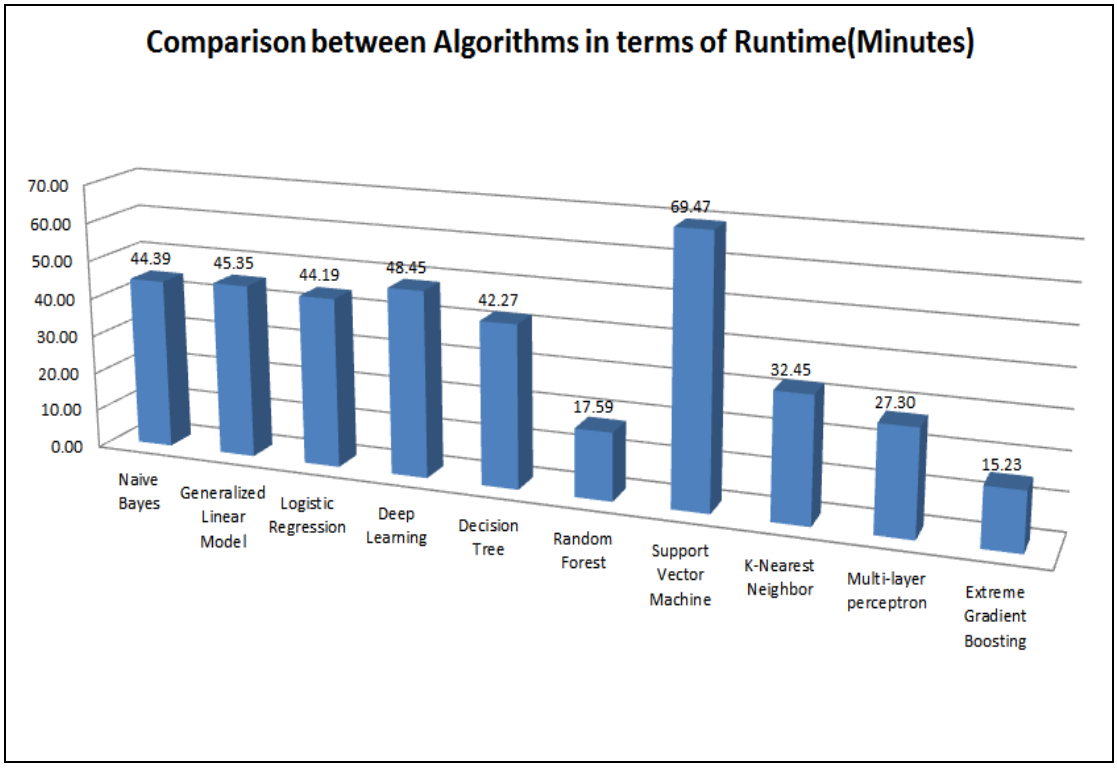
ภาพประกอบ 23 กราฟเปรียบเทียบประสิทธิภาพด้านค่าความระลึก (Recall)

เมื่อทำการทดสอบประสิทธิภาพของโมเดล การจำแนกการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง ในด้านค่าความระลึกหรือเรียกอย่างว่าค่าเรียกกลับ (Recall) เป็นอัตราส่วนของการค้นพบข้อมูลที่ถูกต้องจากจำนวนข้อมูลที่ถูกต้องทั้งหมด พบว่า อัลกอริทึมเอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) ให้ค่าความระลึก สูงที่สุดคือ 97.1 % รองลงมาเป็นอัลกอริทึมโครงข่ายประสาทเทียมพอร์เซฟตรอนแบบหลายชั้น (Multi-Layer Perceptron Neural Networks) ให้ค่าความระลึก 96.1 % อัลกอริทึมเคเนียเรสเนเบอร์ (K-Nearest Neighbor) ให้ค่าความระลึก 91.3% อัลกอริทึมการเรียนรู้เชิงลึก (Deep Learning) ให้ค่าความระลึก 87.2% อัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) ให้ค่าความระลึก 63.0% อัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) ให้ค่าความระลึก 61.9% อัลกอริทึมตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) และ อัลกอริทึมการวิเคราะห์การถดถอยโลจิสติก (Logistic Regression) ให้ค่าความระลึกเท่ากันคือ 48.7% อัลกอริทึมแรนดอมฟอเรส (Random Forest) ให้ค่าความระลึกเท่ากันคือ 43.7% และสุดท้ายอัลกอริทึมเนอีฟเบย์ (Naïve Bayes) ให้ค่าความระลึก 28.9% ตามลำดับ



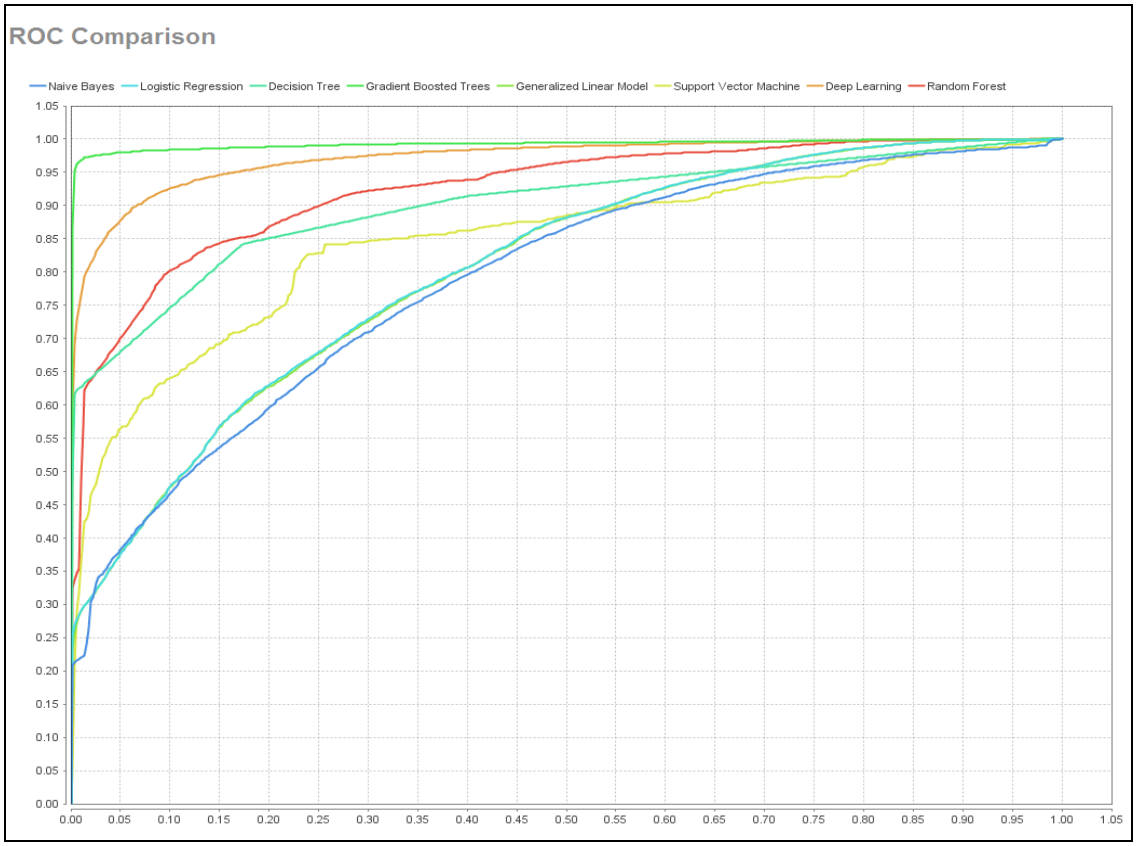
ภาพประกอบ 24 กราฟเปรียบเทียบประสิทธิภาพด้านค่า F-Measure

เมื่อทำการทดสอบประสิทธิภาพของโมเดล การจำแนกการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง ในด้านค่าความระลึกหรือที่เรียกว่าค่า F-Measure พบว่าอัลกอริทึมเอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) ให้ค่า F-Measure สูงที่สุดคือ 97.4 % รองลงมาเป็นอัลกอริทึมเคเนียเรสเนเบอร์ (K-Nearest Neighbor) ให้ค่า F-Measure 91.8 % อัลกอริทึมโครงข่ายประสาทเทียมพอร์เซฟตรอนแบบหลายชั้น (Multi-Layer Perceptron Neural Networks) ให้ค่า F-Measure 91.7 % อัลกอริทึมการเรียนรู้เชิงลึก (Deep Learning) ให้ค่า F-Measure 89.3% อัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) ให้ค่า F-Measure 76.1% อัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) ให้ค่า F-Measure 69.7% อัลกอริทึมแรนดอมฟอเรส (Random Forest) ให้ค่า F-Measure 60.0% อัลกอริทึมตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) และ อัลกอริทึมการวิเคราะห์การถดถอยโลจิสติก (Logistic Regression) ให้ค่า F-Measure เท่ากันคือ 58.2% และสุดท้ายอัลกอริทึมเนอีฟเบย์ (Naïve Bayes) ให้ค่า F-Measure 43.9% ตามลำดับ



ภาพประกอบ 25 กราฟเปรียบเทียบเวลาในการสร้างและทดสอบแบบจำลอง

เมื่อทำการเปรียบเทียบเวลาในการสร้างและทดสอบแบบจำลอง การจำแนกการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง ในด้านเวลาเป็นนาที หรือที่เรียกว่าค่า Runtime พบว่าอัลกอริทึมเอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) ใช้เวลา Runtime น้อยที่สุดคือ 15.23 นาที รองลงมาเป็นอัลกอริทึมแรนดอมฟอเรส (Random Forest) ใช้เวลา Runtime 17.59 นาที อัลกอริทึมโครงข่ายประสาทเทียมพอร์เซฟตรอนแบบหลายชั้น (Multi-Layer Perceptron Neural Networks) ใช้เวลา Runtime 27.30 นาที อัลกอริทึมเคเนียเรสเนเบอร์ (K-Nearest Neighbor) ใช้เวลา Runtime 32.45 นาที อัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) ใช้เวลา Runtime 42.27 นาที อัลกอริทึมการวิเคราะห์การถดถอยโลจิสติก (Logistic Regression) ใช้เวลา Runtime 44.19 นาที อัลกอริทึมเนอีฟเบย์ (Naïve Bayes) ใช้เวลา Runtime 44.39 นาที อัลกอริทึมตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) ใช้เวลา Runtime 45.35 นาที อัลกอริทึมการเรียนรู้เชิงลึก (Deep Learning) ใช้เวลา Runtime 48.45 นาที และสุดท้าย อัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) ใช้เวลา Runtime 69.47 นาที ตามลำดับ



ภาพประกอบ 26 กราฟเปรียบเทียบประสิทธิภาพ ROC curve แต่ละอัลกอริทึม

เมื่อทำพล็อตกราฟตรวจสอบการพยากรณ์ของตัวแบบโมเดลที่ถูกต้องด้วยเส้นโค้ง ROC curve (Receiver Operating Characteristic Curve) และเพื่อเปรียบเทียบประสิทธิภาพของแต่ละโมเดลในการจำแนกการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง โดยแสดงกราฟความสัมพันธ์ การทำนายถูกไปทางแกนตั้ง (Y) และถ้าทายผิดไปแนวแกนนอน (X) และถ้าค่า ROC Curve มีค่าเข้าใกล้ 1 จะแสดงว่าโมเดลนั้นมีประสิทธิภาพที่ดีกว่า จากการทดสอบประสิทธิภาพของโมเดลพบว่า โมเดลที่ประสิทธิภาพดีสุด คือ อัลกอริทึมเอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) รองลงมาเป็นอัลกอริทึมการเรียนรู้เชิงลึก (Deep Learning) อัลกอริทึมแรนดอมฟอเรส (Random Forest) อัลกอริทึมเคเนียเรสเนเบอร์ (K-Nearest Neighbor) อัลกอริทึมโครงข่ายประสาทเทียมพอร์เซฟตรอนแบบหลายชั้น (Multi-Layer Perceptron Neural Networks) อัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) อัลกอริทึมการวิเคราะห์การถดถอยโลจิสติก (Logistic Regression) และอัลกอริทึมตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) ประสิทธิภาพใกล้เคียงกัน และสุดท้ายอัลกอริทึมเนอีฟเบย์ (Naïve Bayes) ด้อยที่สุดตามลำดับ

บทที่ 5

สรุปผล อภิปรายผลและข้อเสนอแนะ

การนำเสนอในบทนี้ เป็นการรวบรวมสาระสำคัญจากการพัฒนาสร้างแบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง ที่มีประสิทธิภาพทั้งในด้านความถูกต้องในการจำแนกประเภท และความถูกต้องในภาพรวมด้านการค้นคืนสารสนเทศ ที่ใช้ทรัพยากรของระบบและหน่วยความจำอย่างเหมาะสม แต่สามารถจำแนกประเภทได้อย่างมีประสิทธิภาพและประสิทธิผล ซึ่งแบบจำลองดังกล่าวสามารถนำไปประยุกต์ใช้กับการพัฒนาระบบงานการตรวจสอบและการป้องกันธุรกรรมทุจริต (Fraud Detection) ตลอดจนเป็นแนวทางในการเพิ่มขีดความสามารถขององค์กรที่จะใช้การเรียนรู้ของเครื่อง (Machine Learning) มาลดความเสียหายที่จะเกิดขึ้นจากธุรกรรมทุจริต และสามารถนำแบบจำลองนี้ไปประยุกต์ใช้กับงานพัฒนาระบบสารสนเทศด้าน Big Data Fraud Detection Analytics การตรวจสอบการทุจริต การวิเคราะห์พฤติกรรมการซื้อขาย (Transaction) ที่ผิดปกติของผู้บริโภคแบบออนไลน์ได้ อีกทั้งรวมทั้งสรุป การอภิปรายผลและข้อเสนอแนะ ตลอดจนแนวทางและข้อเสนอแนะการพัฒนางานวิจัยในด้านนี้ต่อไป

สรุปผลการวิจัย

งานวิจัยนี้ได้นำเสนอวิธีการสร้างและทดสอบประสิทธิภาพของแบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง โดยงานวิจัยนี้มุ่งเน้นที่จะพัฒนาประสิทธิภาพของแบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการเรียนรู้ของเครื่อง โดยแบบจำลองที่นำเสนอในงานวิจัยนี้ ใช้การตรวจสอบข้อมูลก่อนการวิเคราะห์ เพื่อกำจัดข้อมูลค่าสุดโต่งแฝง(Outlier) ออกก่อน จากนั้นทำการแปลงข้อมูล (Data Transformation) ด้วยวิธีการนอร์มอลไลซ์ (Normalization) ทำการแปลงค่าข้อมูลให้อยู่ในช่วงสั้นๆ โดยแปลงข้อมูลเชิงเส้นจากช่วงที่เป็นไปได้เดิมของค่าอินพุต ให้เป็นช่วงข้อมูลใหม่ที่กำหนดขึ้นให้มีสเกลที่เท่ากัน โดยการปรับการกระจายของข้อมูล และค่าเบี่ยงเบนมาตรฐาน (Z-Score Normalization) แล้วจึงทำการปรับปรุงชุดข้อมูลกลุ่มตัวอย่างของการตรวจสอบการทุจริตไม่สมดุลด้วยวิธีการสุ่มตัวอย่าง (Sampling Methods) โดยวิธีการ Over-sampling จะทำการสุ่มข้อมูลในกลุ่มรองเพื่อสร้างข้อมูลใหม่ของกลุ่มรองให้มีจำนวนเพิ่มมากขึ้น

ให้ใกล้เคียงหรือเท่ากับจำนวนข้อมูลในกลุ่มหลัก จากนั้นทำการลดมิติข้อมูลของกลุ่มตัวอย่างลง โดยพิจารณาคุณลักษณะที่มีค่าสถิติความสัมพันธ์กับกลุ่มเป้าหมาย สูง จะให้ถ่วงน้ำหนักมาก ก่อนส่งเข้าเครื่องจักรการเรียนรู้แบบมีผลเฉลย (Supervised Learning) ซึ่งประกอบด้วยอัลกอริทึม 1. เนอีฟเบย์ (Naïve Bayes) 2. ตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) 3. การวิเคราะห์การถดถอยโลจิสติก (Logistic Regression) 4. การเรียนรู้เชิงลึก (Deep Learning) 5. ต้นไม้ตัดสินใจ (Decision Tree) 6. แรนดอมฟอเรส (Random Forest) 7. ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) 8. เคเนียเรสเนเบอร์ (K-Nearest Neighbor) 9.โครงข่ายประสาทเทียมพอร์เซฟตรอนแบบหลายชั้น (Multi-Layer Perceptron Neural Networks) 10.เอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) โดยมีข้อสรุปและข้อค้นพบที่ได้จากการวิจัยดังนี้

จากการทดลองทดสอบประสิทธิภาพของโมเดลการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง สามารถสรุปได้ว่า อัลกอริทึมที่ให้ประสิทธิภาพการจำแนกการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่องออกมาดีที่สุด เมื่อปเรียบเทียบกับวิธีการอื่นๆ และเมื่อพิจารณาค่าพารามิเตอร์ที่ส่งผลให้ประสิทธิภาพของโมเดลที่ดีสุด คืออัลกอริทึม เอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) ให้ค่าความถูกต้อง (Accuracy) สูงที่สุดคือ 98.15 % รองลงมาเป็นอัลกอริทึมการเรียนรู้เชิงลึก (Deep Learning) ให้ค่าความถูกต้อง 92.42% อัลกอริทึมเคเนียเรสเนเบอร์ (K-Nearest Neighbor) ให้ค่าความถูกต้อง 89.64% อัลกอริทึมโครงข่ายประสาทเทียมพอร์เซฟตรอนแบบหลายชั้น (Multi-Layer Perceptron Neural Networks) ให้ค่าความถูกต้อง 88.88% อัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) ให้ค่าความถูกต้อง 85.82% ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) ให้ค่าความถูกต้อง 80.12% อัลกอริทึมแรนดอมฟอเรส (Random Forest) ให้ค่าความถูกต้อง 78.84% อัลกอริทึมตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) และอัลกอริทึมการวิเคราะห์การถดถอยโลจิสติก (Logistic Regression) ให้ค่าความถูกต้องเท่ากันที่ 74.57 % และสุดท้ายอัลกอริทึมเนอีฟเบย์ (Naïve Bayes) ให้ค่าความถูกต้องน้อยที่สุด 72.96 % ตามลำดับ

เมื่อพิจารณาถึงปัจจัยด้านค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) ของโมเดล พบว่าอัลกอริทึม เอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) ให้ค่าเบี่ยงเบนมาตรฐานน้อยที่สุด คือ 0.1% รองลงมาเป็นอัลกอริทึมเนอีฟเบย์ (Naïve Bayes) อัลกอริทึมตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) อัลกอริทึมการวิเคราะห์การถดถอยโลจิสติก (Logistic Regression) อัลกอริทึมการเรียนรู้เชิงลึก (Deep Learning) อัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) ให้ค่าเบี่ยงเบนมาตรฐานเท่ากันคือ 0.2% ถัดไปเป็นอัลกอริทึมเคเนียเรสเนเบอร์ (K-Nearest

Neighbor) ให้ค่าเบี่ยงเบนมาตรฐาน 0.3% อัลกอริทึมแรนดอมฟอเรส (Random Forest) ให้ค่าเบี่ยงเบนมาตรฐาน 0.4% อัลกอริทึมโครงข่ายประสาทเทียมพอร์เซฟตรอนแบบหลายชั้น (Multi-Layer Perceptron Neural Networks) ให้ค่าเบี่ยงเบนมาตรฐาน 0.5% และสุดท้ายซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) ให้ค่าเบี่ยงเบนมาตรฐานสูงที่สุดคือ 1.4% ตามลำดับ

และเมื่อทำพล็อตกราฟตรวจสอบการพยากรณ์ของตัวแบบโมเดลที่ถูกต้องด้วยเส้นโค้ง ROC curve (Receiver Operating Characteristic Curve) และเพื่อเปรียบเทียบประสิทธิภาพของแต่ละโมเดลในการจำแนกการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง โดยแสดงกราฟความสัมพันธ์ การทำนายถูกไปทางแกนตั้ง (Y) และถ้าทายผิดไปแนวแกนนอน (X) และถ้าค่า ROC Curve มีค่าเข้าใกล้ 1 จะแสดงว่าโมเดลนั้นมีประสิทธิภาพที่ดีกว่า จากการทดสอบประสิทธิภาพของโมเดลพบว่า โมเดลที่ประสิทธิภาพดีสุด คือ อัลกอริทึมเอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) รองลงมาเป็นอัลกอริทึมการเรียนรู้เชิงลึก (Deep Learning) อัลกอริทึมแรนดอมฟอเรส (Random Forest) อัลกอริทึมเคเนียเรสเนเบอร์ (K-Nearest Neighbor) อัลกอริทึมโครงข่ายประสาทเทียมพอร์เซฟตรอนแบบหลายชั้น (Multi-Layer Perceptron Neural Networks) อัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) อัลกอริทึมการวิเคราะห์การถดถอยโลจิสติก (Logistic Regression) และอัลกอริทึมตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) ประสิทธิภาพใกล้เคียงกัน และสุดท้ายอัลกอริทึมเนอีฟเบย์ (Naïve Bayes) ด้อยที่สุดตามลำดับ

อภิปรายผลการวิจัย

จากข้อสรุปผลการวิจัยพบว่า ปัจจัยหลักที่ส่งผลให้ประสิทธิภาพของแบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง มีประสิทธิภาพสูงนั้น เกิดจากปัจจัยหลักด้านการผสมผสานกระบวนการใช้การตรวจสอบข้อมูลก่อนการวิเคราะห์ เพื่อกำจัดข้อมูลค่าสุดโต่งแฝง(Outlier) ออกก่อน จากนั้นทำการแปลงข้อมูล (Data Transformation) ด้วยวิธีการนอร์มอลไลซ์ (Normalization) ทำการแปลงค่าข้อมูลให้อยู่ในช่วงสั้นๆ โดยแปลงข้อมูลเชิงเส้นจากช่วงที่เป็นไปได้เดิมของค่าอินพุต ให้เป็นช่วงข้อมูลใหม่ที่กำหนดขึ้นให้มีสเกลที่เท่ากัน โดยการปรับการกระจายของข้อมูล และค่าเบี่ยงเบนมาตรฐาน (Z-Score Normalization) แล้วจึงทำการปรับปรุงชุดข้อมูลกลุ่มตัวอย่างของการตรวจสอบการทุจริตไม่สมดุลด้วยวิธีการสุ่มตัวอย่าง (Sampling Methods) โดยวิธีการ Oversampling จะทำการสุ่มข้อมูลในกลุ่มรองเพื่อสร้างข้อมูลใหม่ของกลุ่มรองให้มีจำนวนเพิ่มมากขึ้น ให้ใกล้เคียงหรือเท่ากับจำนวนข้อมูลในกลุ่มหลัก จากนั้นทำการลดมิติข้อมูลของกลุ่มตัวอย่างลง โดยพิจารณาคุณลักษณะที่มี

ค่าสถิติความสัมพันธ์กับกลุ่มเป้าหมาย สูง จะให้ถ่วงน้ำหนักมาก ก่อนส่งเข้าเครื่องจักรการเรียนรู้แบบมีผลเฉลย (Supervised Learning) แบบต่างๆ จากการศึกษาพบว่ากระบวนการแปลงสภาพข้อมูลนั้นมีสำคัญมาก ต่อประสิทธิภาพของพัฒนาแบบจำลอง ไม่ว่าจะเป็นขั้นตอนสำหรับการคัดข้อมูลที่ไม่เกี่ยวข้องออกไป (Data Cleaning) ขั้นตอนการรวมข้อมูลที่มีหลายแหล่งให้เป็นข้อมูลชุดเดียวกัน (Data Integration) และขั้นตอนการแปลงข้อมูลให้เหมาะสม (Data Transformation) ส่งเข้าเครื่องจักรการเรียนรู้แบบมีผลเฉลย (Supervised Learning) ประกอบกับอัลกอริทึมเอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) จัดเป็นเทคนิคเพื่อลดความแปรปรวนและเพิ่มความแม่นยำในการทำนายของตัวจำแนกประเภทโดยใช้วิธีลดความอคติ (Bias) และมีแนวคิดที่ให้ตัวเรียนรู้ที่อ่อนแอ (Weak Learner) ชุดหนึ่ง ทำงานร่วมกันจนสามารถพัฒนามาเป็นตัวเรียนรู้ที่เข้มแข็ง (Strong Learner) ได้ การสร้างตัวเรียนรู้ที่อ่อนแอแต่ละตัว สามารถทำได้โดยการปรับเพิ่มน้ำหนักของการทำนายที่ผิดพลาดให้มากขึ้นในแต่ละรอบ แล้วทำการเรียนรู้ใหม่ ซึ่งจะทำให้โมเดลของตัวจำแนก (Classifier) เปลี่ยนไปโดยให้ความสำคัญกับความผิดพลาดในรอบที่แล้วมากขึ้น เมื่อได้จำนวนตัวเรียนรู้ที่อ่อนแอมากพอแล้ว จึงนำมารวมกันสร้างเป็นตัวเรียนรู้ที่เข้มแข็งต่อไป โดยผลรวมของตัวจำแนกประเภท (Aggregating) จะเกิดเป็นตัวจำแนกประเภทใหม่ขึ้นมา เราจะทำแบบนี้ไปเรื่อยๆ (Recursive) จนได้โมเดลที่ดีที่สุดจากผลรวมของการจำแนก

ข้อเสนอแนะ

จากข้อค้นพบในการพัฒนาประสิทธิภาพของแบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง ที่ได้จากงานวิจัยนี้นั้น ผู้วิจัยของเสนอแนะเพื่อพัฒนาปรับปรุงประสิทธิภาพของงานวิจัยดังต่อไปนี้

งานวิจัยการพัฒนาประสิทธิภาพแบบจำลองแบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่องนี้ มุ่งเน้นเพื่อพัฒนาประสิทธิภาพในการการจำแนกประเภท โดยใช้อัลกอริทึมจำแนกประเภท (Classifier) ซึ่งประกอบด้วย 1. เนอีฟเบย์ (Naïve Bayes) 2. ตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) 3. การวิเคราะห์การถดถอยโลจิสติก (Logistic Regression) 4. การเรียนรู้เชิงลึก (Deep Learning) 5. ต้นไม้ตัดสินใจ (Decision Tree) 6. แรนดอมฟอเรส (Random Forest) 7. ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) 8. เคเนียเรสเนเบอร์ (K-Nearest Neighbor) 9.โครงข่ายประสาทเทียมพอร์เซฟตรอนแบบหลายชั้น (Multi-Layer Perceptron Neural Networks) 10.เอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) แต่มิได้ทดสอบกับอัลกอริทึมที่มีความซับซ้อนประเภทอื่น (Complexity Classifier) ตลอดจนการปรับแต่งแก้ไขพามิเตอร์การเพิ่มประสิทธิภาพ

(Optimization) ของแต่ละอัลกอริทึม ดังนั้นจึงควรมีการศึกษาวิธีการปรับแต่งอัลกอริทึมและพารามิเตอร์ต่างๆ(Parameter Tuning) เพื่อเพิ่มประสิทธิภาพ (Optimization) ในการจำแนกการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลให้มีประสิทธิภาพมากขึ้น งานวิจัยนี้มุ่งเน้นการวัดประสิทธิภาพด้านความถูกต้อง (Accuracy) ในการจำแนกการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลหลัก โดยคำนึงถึงระยะเวลาที่ใช้ในการประมวลผลเพื่อสร้างแบบจำลองเป็นปัจจัยรอง ดังนั้นจึงควรมีการศึกษาด้านการปรับปรุงประสิทธิภาพด้านความเร็ว และการบริหารหน่วยความจำที่ใช้ในการประมวลผล เพื่อสร้างแบบจำลองให้เกิดประสิทธิภาพทั้งในด้านความแม่นยำสูงสุด ในขณะเดียวกันก็ใช้เวลา (Runtime) ต่ำสุด ไม่ว่าจะเป็นการนำเทคนิคการประมวลผลแบบขนาน (Parallel Processing) การประมวลผลจำแนกประเภทแบบทันเวลา (Real Time Processing) การประมวลผลที่อยู่ในรูปแบบที่ไม่มีโครงสร้าง (Unstructured Data) การประมวลผลข้อมูลบนฐานขนาดใหญ่ (Big Data Analytics Platform) ตลอดจนประยุกต์ใช้สถาปัตยกรรมบนกลุ่มเมฆ (Cloud Architecture) และการประมวลผลบนกลุ่มเมฆ (Cloud Computing)

บรรณานุกรม

บรรณานุกรม

ปิยะ ปานทอง. 2553. **“**การตรวจจับธุรกรรมทุจริตทางบัตรเครดิตจากรายการชำระสินค้าพาณิชย์อิเล็กทรอนิกส์โดยการทำเหมืองข้อมูล กรณีศึกษา ธนาคารพาณิชย์แห่งหนึ่ง. ” วิทยานิพนธวิทยาศาสตรมหาบัณฑิต สาขาการบริหารเทคโนโลยี บัณฑิตวิทยาลัย มหาวิทยาลัยธรรมศาสตร์.

B. Baesens, V. Van Vlasselaer, W. Verbeke. 2015. **Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection.** New Jersey: Wiley Publisher.

Beysolow II, Taweh. 2017. **Introduction to Deep Learning Using R.** New York: Apress Publisher.

Daniel Sánchez et al. 2009. “Association Rules Applied to Credit Card Fraud Detection.” **Expert Systems with Applications** 36, 2: 3630-3640.

E.Caldeira. et al., 2014. “Fraud Analysis and Prevention in e-Commerce Transactions.” **Proceedings of the 9th Latin American Web Congress** 2014: 42-49.

Francois Chollet. 2017. **Deep Learning with Python**. New York: Manning Publisher.

Francisca N.Ogwueleka. 2011. “Data Mining Application In Credit Card Fraud Detection System.” **Journal of Engineering Science And Technology** 6, 3: 311–322.

Han, J., Kamber, M. 2006. **Data mining: Concepts and techniques**, 2nd ed. San Francisco: Morgan Kaufmann Publisher.

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

Jon T. S. Quah, Sriganesh Srihari. 2008. **“**Real Time Credit Card Fraud Detection using Computational Intelligence.**” Expert Systems with Applications** 35, 4: 1721-1732.

Leo Breiman. 2001. Random Forests . **Journal Machine Learning** 45,1: 5–32.

Maria R. et al. 2016. “Credit Card Fraud Detection with Unsupervised Algorithms.”, **Journal of Advances in Information Technology** 7,1: 34-38.

M.A.H. Farquad, Indranil Bose. 2012. “Preprocessing unbalanced data using support vector machine.” **Decision Support Systems** 53,1: 226-233.

บรรณานุกรม (ต่อ)

M. A. H. Farquad, Vadlamani Ravi, S. Bapi Raju. 2014. “Churn prediction using comprehensible support vector machine: An analytical CRM application.” **Applied Soft Computing**. 19,1: 31-40.

Nicholls, C., and Song, F. 2010. **“**Comparison of Feature Selection Methods for Sentiment Analysis.” **Advances in Artificial Intelligence,** Lecture Notes in Computer Science 2010: 286-289.

Nonita Sharma. 2017. **The Extreme Gradient Boosting for Mining Applications**. Latvia, European Union: LAMBERT Academic Publisher.

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

Robert E. Schapire and Yoav Freund. 2014. **Boosting Foundations and Algorithms**. Massachusetts: MIT Press Publisher.

Shearer C. 2000. “The CRISP-DM model: the new blueprint for data mining.” **Journal of Data Warehousing** 5: 13-22.

Sunder Gee. 2014. **Fraud and Fraud Detection: A Data Analytics Approach.** New Jersey: Wiley Publisher.

Scott Hartshorn. 2016. **Machine Learning With Random Forests And Decision Trees: A Visual Guide For Beginners.** Seattle: Amazon Digital Services Publisher.

S Panigrahi, A Kundu, S Sural, AK Majumdar. 2009. “Credit card fraud detection: A fusion Approach using Dempster–Shafer theory and Bayesian learning.” **Information Fusion** 10,4: 354-363.

Vijayshree B. Nipane et al. 2016. “Fraudulent Detection in Credit Card System Using SVM & Decision Tree.” **International Journal of Scientific Development and Research** 1,5: 590-594.

Y. Sahin , E. Duman. 2011. “Detecting Credit Card Fraud by Decision Trees and Support Vector Machines.” **Proceedings of the International Multi Conference of Engineers and Computer Scientists** 2011,1: 442-447.

ประวัติย่อผู้วิจัย

ชื่อ	นายนิเวศ จิระวิชิตชัย
วัน เดือน ปีเกิด	วันที่ 30 เมษายน 2517
สถานที่เกิด	กรุงเทพฯ
สถานที่อยู่ปัจจุบัน	บ้านเลขที่ 195 ถนนจารุเมือง อำเภอปทุมวัน กรุงเทพฯ 10330
ตำแหน่งหน้าที่การงานปัจจุบัน	อาจารย์ประจำ หลักสูตรวิทยาศาสตรมหาบัณฑิต
	สาขาวิชาเทคโนโลยีสารสนเทศ
สถานที่ทำงานปัจจุบัน	คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยศรีปทุม
ประวัติการศึกษา	พ.ศ. 2540 บธ.บ. สาขาการจัดการอุตสาหกรรม
	จาก มหาวิทยาลัยรามคำแหง
	พ.ศ. 2546 คอ.ม. สาขาเทคโนโลยีคอมพิวเตอร์
	จาก สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ
	พ.ศ. 2553 ปร.ด. สาขาเทคโนโลยีสารสนเทศ
	จาก สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ