การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลการใช้บริการเครือข่ายอินเทอร์เน็ตของสถาบันอุดมศึกษา

A Comparison of the Efficiency of the Classification of Internet Service Usage of Higher Education Institutions

พรเทพ ด่านน้อย[1*] และสุวิมล มรรควิบูลย์ชัย[2]

[1]สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม
[2]สาขาวิชาวิทยาการข้อมูล คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม
[*]jack.haper007@gmail.com

บทคัดย่อ

งานวิจัยนี้นำเสนอผลการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลการใช้บริการเครือข่ายอินเทอร์เน็ตของสถาบันอุดมศึกษา โดยเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลของ 3 เทคนิค คือ เทคนิค Decision Tree เทคนิค K-NN และเทคนิค Deep Learning โดยการนำข้อมูลเกี่ยวกับข้อมูลการใช้บริการเครือข่ายอินเทอร์เน็ตของสถาบันอุดมศึกษา มีข้อมูลจำนวนที่สามารถใช้ในงานวิจัยได้ 8,530 ชุด ซึ่งทำการแบ่งข้อมูลด้วยวิธี Cross-validation Test โดยการสุ่มข้อมูลเพื่อแบ่งข้อมูลออกเป็น 5 ส่วน ในแต่ละส่วนประกอบด้วยข้อมูล จำนวน 1,706 ชุด โดยสร้างโมเดลจากการเลือกข้อมูล 4 ส่วน และทดสอบประสิทธิภาพของระบบด้วยข้อมูล 1 ส่วน ผลการวิจัยพบว่า เทคนิคที่ใช้ในการจำแนกข้อมูลการใช้บริการเครือข่ายอินเทอร์เน็ตของสถาบันอุดมศึกษา ที่มีประสิทธิภาพสูงสุด คือ Deep Learning โดยมีค่าความแม่นตรง 88.79% ค่าความระลึก 88.55% ค่าความแม่นยำ 89.10% และค่าถ่วงดุล 0.289 ซึ่งเป็นระดับการประเมินที่สามารถยอมรับได้

คำสำคัญ: การจำแนกข้อมูล ต้นไม้ตัดสินใจ เคเนียเรสเนเบอร์ การเรียนรู้เชิงลึก

Abstract

This research presents the results of a comparison of the efficiency of the classification of internet service usage of higher education institutions. The efficiency of 3 classification techniques were compared namely, Decision Tree technique, K-NN technique and Deep Learning technique by taking internet service usage data of higher education institutions. There are 8,530 sets of data that can be used in this research, which were randomly divided by cross-validation test into 5 parts, consisting of 1,706 sets in each part. The model was created from 4 selected parts, then evaluated by the rest part. The results showed that the most efficiency technique in the classification of internet service usage of higher educational institutions was Deep Learning with 88.79 percent of precision, 88.55 percent of recall, 89.10 percent of accuracy and 0.289 of F-measure counterbalance, which were acceptable.

Keywords: data classification, decision tree, k-NN, deep learning

1. บทนำ

การเติบโตของการใช้อินเทอร์เน็ตเพิ่มขึ้นอย่างรวดเร็ว โดยเฉพาะอย่างยิ่งจากการใช้โทรศัพท์มือถือและอุปกรณ์แท็บเล็ตที่มีขนาดพกพา ทำให้สามารถเข้าถึงการใช้อินเทอร์เน็ตได้ทุกที่ทุกเวลา โดยเฉพาะอย่างยิ่ง ในยุคที่สื่อสังคมออนไลน์เข้ามามีบทบาทในชีวิตประจำวัน ทำให้อินเทอร์เน็ตเป็นสิ่งแรกที่ใครหลายคนนึกถึงยามตื่นนอน ยามกินอาหาร ยามเดินทาง ยามว่าง ยามออกกำลังกาย หรือแม้แต่ใครหลายคนอาจจะหลับไปพร้อมกับการใช้อินเทอร์เน็ตจากโทรศัพท์มือถือเลยก็ได้ นอกจากนี้ความพยายามที่จะทำให้อุปกรณ์ต่าง ๆ เชื่อมต่อกับอินเทอร์เน็ตมากขึ้น ทำให้สามารถเชื่อมต่อใช้งานอินเทอร์เน็ตได้หลากหลายรูปแบบและหลากหลายช่องทางไม่ว่าจะเป็นการศึกษา การวิจัย การทำงาน ความบันเทิง หรือแม้แต่สุขภาพ ทำให้มีการกล่าวถึง Internet of Things คือ อินเทอร์เน็ตทุกสรรพสิ่ง หรือทุกสิ่งทุกอย่างล้วนแล้วแต่ใช้อินเทอร์เน็ต ไม่ว่าจะเป็นเรื่องราวข่าวสาร การสร้างธุรกิจ การศึกษาข้อมูล การลงทุน ประวัติสุขภาพ หรือแม้แต่ชีวิตประจำวันอินเทอร์เน็ตก็เข้ามามีส่วนแทบทุกกระบวนการ

จากการเพิ่มปริมาณการใช้ข้อมูลบนเครือข่ายอินเทอร์เน็ตอย่างต่อเนื่อง ทำให้เกิดการกระทำความผิดเกี่ยวกับคอมพิวเตอร์เพิ่มขึ้น ข้อมูลจราจรทางคอมพิวเตอร์จึงมีความสำคัญต่อหน่วยงานทั้งภาครัฐและเอกชนและผู้ให้บริการเป็นอย่างมาก เนื่องจากต้องมีการจัดเก็บข้อมูลจราจรทางคอมพิวเตอร์อย่างเป็นระบบ ถูกต้อง ทันสมัยและสามารถนำมาใช้ได้จริง สอดคล้องกับหลักเกณฑ์การเก็บข้อมูลจราจรทางคอมพิวเตอร์ตามประกาศกระทรวงเทคโนโลยีสารสนเทศและการสื่อสาร (กระทรวงเทคโนโลยีสารสนเทศและการสื่อสาร, 2550) การใช้อินเทอร์เน็ตภายในองค์กรที่มีขนาดใหญ่ เช่น สถาบันอุดมศึกษา ที่ต้องให้บริการผู้ใช้จำนวนมากทั้งการเข้าถึงเว็บไซต์ที่อยู่ภายในมหาวิทยาลัย ภายนอกมหาวิทยาลัย ทั้งภายในและต่างประเทศ ทำให้แฟ้มข้อมูลที่เก็บร่องรอยการใช้งานเครือข่ายอินเทอร์เน็ตของผู้ใช้ภายในมหาวิทยาลัย (internet usage log file) มีขนาดใหญ่มาก เนื่องจากแต่ละวินาทีมีผู้ใช้พร้อม ๆ กันจำนวนมาก

การวิเคราะห์การใช้งานของอินเทอร์เน็ต จึงทำได้ยากเนื่องจากมีรายการข้อมูลเกิดขึ้นปริมาณมากตลอดเวลา และมากขึ้นทุก ๆ วัน บางหน่วยงานมีปริมาณการเก็บข้อมูลถึงวันละ 4 GB โดยไม่มีการนำไปใช้ประโยชน์ ผู้วิจัยจึงได้นำวิทยาการในการจัดการกับข้อมูลขนาดใหญ่มาใช้ เพื่อนำข้อมูลเหล่านี้มาวิเคราะห์แนวทางการจำแนกข้อมูลเพื่อให้เกิดประโยชน์ต่อการบริหารจัดการข้อมูลการใช้งานเครือข่ายอินเทอร์เน็ตภายในมหาวิทยาลัย เพื่อใช้ประกอบการวางแผนเกี่ยวกับการให้บริการข้อมูลตลอดจนการจัดระเบียบข้อมูลจราจรทางคอมพิวเตอร์ต่อไป

2. งานวิจัยและทฤษฎีที่เกี่ยวข้อง

2.1 ข้อมูลจราจรทางคอมพิวเตอร์และส่วนประกอบของระบบเก็บข้อมูลล็อก

ข้อมูลจราจรทางคอมพิวเตอร์ หมายถึง ข้อมูลเกี่ยวกับการติดต่อสื่อสารของระบบคอมพิวเตอร์ ที่แสดงถึงต้นทาง ปลายทาง เส้นทาง เวลา วันที่ ปริมาณ ระยะเวลา ชนิดของบริการ หรือข้อมูลอื่นที่เกี่ยวข้องกับการติดต่อสื่อสารของระบบคอมพิวเตอร์ โดยต้องเก็บรักษาข้อมูลด้วยวิธีการที่มั่นคงปลอดภัย เช่น เก็บในสื่อที่มีระบบการเก็บรักษาความลับของข้อมูลที่สามารถระบุรายละเอียดของผู้ใช้บริการรายบุคคลได้ เป็นต้น (กระทรวงเทคโนโลยีสารสนเทศและการสื่อสาร, 2550)

ส่วนประกอบของระบบเก็บข้อมูลล็อก (วัชรินทร์ จิรโสภณ, 2555 อ้างถึง กระทรวงเทคโนโลยีสารสนเทศและการสื่อสาร, 2550)

1) Log Generation หรือ Log Source เป็นเซิร์ฟเวอร์หรืออุปกรณ์บนเครือข่ายที่เป็นแหล่งกำเนิดข้อมูลล็อกจากระบบปฏิบัติการ และ แอปพลิเคชัน การจัดเก็บข้อมูลล็อกบนเครื่องเซิร์ฟเวอร์หรืออุปกรณ์ในตัวเองเรียกว่า Primary Logging ในกรณีที่มีการจัดส่งข้อมูลล็อกไปยังล็อกเซิร์ฟเวอร์ (Log server) จะเรียกลักษณะการส่งข้อมูลล็อกนี้ว่า Secondary Logging

2) Log Storage and Correlation เป็นล็อกเซิร์ฟเวอร์สำหรับรับข้อมูลล็อกจากแหล่งกำเนิดข้อมูลล็อก (Log Generation) เพื่อจัดเก็บตามรูปแบบที่กำหนดไว้ รวมทั้ง การแปลงข้อมูลล็อกให้อยู่ในรูปแบบที่สามารถจัดเก็บและพร้อมนำไปวิเคราะห์ต่อได้ ในกรณีที่เซิร์ฟเวอร์รับข้อมูลล็อกจากแหล่งกำเนิดข้อมูลจำนวนมากจะเรียกว่า Collectors หรือ Aggregators

3) Log Analysis and Monitoring เป็นหน้าต่างสำหรับผู้ดูแลระบบ หรือผู้ที่มีหน้าที่รับผิดชอบในการวิเคราะห์ข้อมูลล็อก และติดตามตรวจสอบความถูกต้องของข้อมูลล็อกระบบจัดเก็บข้อมูลล็อกบางระบบสนับสนุนการสร้างรายงานการวิเคราะห์ข้อมูลล็อก ทั้งนี้เพื่อให้ข้อมูลเร็วและตรงกับความเป็นจริงในปัจจุบันที่สุด

2.2 ระบบจัดเก็บข้อมูลล็อกและลักษณะข้อมูลล็อก

จากการค้นคว้าพบว่า ระบบการจัดเก็บข้อมูลล็อกมีความสำคัญกับหน่วยงานเป็นอย่างมาก เมื่อพิจารณาความสามารถของระบบจัดเก็บและลักษณะข้อมูลล็อกสามารถอธิบายได้ ดังนี้

2.2.1 การจัดเก็บข้อมูลล็อก

1) Log rotation เป็นการจัดเก็บล็อกไฟล์โดยการหมุนเวียนข้อมูลล็อก โดยจะดำเนินการตามระยะเวลาที่เหมาะสม เช่น ทุกวัน ทุกสัปดาห์ หรือ เมื่อขนาดของไฟล์ข้อมูลล็อกมีขนาดถึงที่กำหนดไว้ เพื่อไม่ให้ไฟล์มีขนาดใหญ่เกินไป

2) Log archival คือการสำรองข้อมูลล็อกเพื่อให้สามารถรักษาระยะเวลาในการจัดเก็บข้อมูลล็อกตามความต้องการ แบ่งเป็นสองแบบ คือ Log retention เป็นการบันทึกข้อมูลล็อกของเหตุการณ์จากระบบสม่ำเสมอ และ Log preservation เป็นกระบวนการรักษาข้อมูลล็อก โดยการบันทึกข้อมูลล็อกบนสื่อบันทึกข้อมูลภายนอก หรือการบันทึกข้อมูลบน SAN (Storage Area Network) หรือการบันทึกบนเซิร์ฟเวอร์ เป็นต้น

3) Log compression คือ การบีบอัดข้อมูลล็อก ทำงานต่อจาก Log rotation หรือ Log archival เพื่อเพิ่มพื้นที่ในการจัดเก็บข้อมูลล็อกและง่ายต่อการสำรองข้อมูลล็อก หรือ การย้ายข้อมูลล็อกไปเก็บไว้บนสื่อบันทึกข้อมูลอื่น

4) Log reduction เป็นการตัด ลบ หรือ ลดข้อมูลล็อกบางส่วนที่ไม่เกี่ยวข้อง ทำงานร่วมกับกระบวนการ Log archival เพื่อลดข้อมูลล็อกที่ไม่เกี่ยวข้องก่อนจะบันทึกข้อมูลล็อกในสื่อบันทึกข้อมูล เช่น การลบตัวอักษรหรืออักขระที่ไม่จำเป็นต่อการเก็บบันทึกข้อมูลล็อก

5) Log conversion เป็นการแปลงรูปแบบการจัดเก็บข้อมูลล็อก เช่น แปลงข้อมูลล็อกจากรูปแบบของไฟล์ Text เป็นรูปแบบข้อมูลล็อกแบบ XML เป็นต้น โดย Log conversion มักจะทำกระบวนการ Event filtering และ Event aggregation จนถึง Log normalization

6) Log normalization เป็นการปรับรูปแบบของข้อมูลล็อกให้อยู่ในรูปแบบเดียวกัน ให้อยู่ในรูปแบบที่สามารถนำไปจัดเก็บ สืบค้น และวิเคราะห์ข้อมูลล็อกโดยผู้เชี่ยวชาญต่อไป เช่น การปรับรูปแบบของวันที่ที่แตกต่างกัน หรือ ความแตกต่างของชื่อตำแหน่งของข้อมูลล็อก มีความสำคัญกับการใช้ล็อกเซิร์ฟเวอร์แบบศูนย์กลาง

7) Log file integrity checking เป็นกระบวนการตรวจสอบความถูกต้องของล็อกไฟล์โดยการทำ Data Hashing กับล็อกไฟล์ที่ไม่มีการเขียนข้อมูลแล้ว เช่น การนำล็อกไฟล์มาบีบอัดและคำนาณด้วยวิธี Message Digest หรือการคำนาณด้วยอัลกอริทึม MD5 ขนาด128 บิต หรือ อัลกอริทึม SHA-1 ขนาด 128 บิต เป็นต้น ผลลัพธ์ที่ได้จะมีขนาดความยาวขนาด 128 บิต เพื่อใช้เป็นตัวแทนของล็อกไฟล์ และจัดเก็บไว้ในสื่อบันทึกข้อมูลที่ปลอดภัย เช่น สื่อบันทึกที่เขียนได้อย่างเดียว

2.3 เทคนิคการทำเหมืองข้อมูล

เนื่องจากการทำเหมืองข้อมูลเป็นเทคนิคในการค้นความรู้จากข้อมูลขนาดใหญ่ การทำเหมืองข้อมูลจึงเป็นการรวมเอาศาสตร์ต่าง ๆ หลายแขนงมารวมไว้ด้วยกันโดยไม่จำกัดวิธีการที่จะใช้ ตัวอย่างศาสตร์ที่ใช้ เช่น เทคโนโลยีฐานข้อมูล (Database technology) วิทยาศาสตร์สารสนเทศ (Information science) สถิติ (Statistics) และระบบการเรียนรู้

(Machine learning) เป็นต้น ซึ่งศาสตร์ต่าง ๆ เหล่านี้จะทำให้เกิดกระบวนการค้นความรู้ในแบบต่าง ๆ (เดช ธรรมศิริ และพยุง มีสัจ. 2556)

1) การแบ่งประเภทและการทำนาย (Classification and Prediction)

จัดเป็นกระบวนการที่ใช้ในการหาภาพแบบของชุดข้อมูลที่มีความใกล้เคียงกัน หรือเหมือนกันมากที่สุด เพื่อใช้ในการทำนายชุดข้อมูลว่าอยู่ในประเภทใดของชุดข้อมูลที่ได้ทำการแบ่งไว้แล้ว ซึ่งชุดข้อมูลที่แบ่งไว้เกิดจากการเรียนรู้จากชุดข้อมูลที่มีอยู่แล้ว (Training data) แบบจำลองที่เกิดจากการเรียนรู้ สามารถแสดงได้หลายภาพแบบ เช่น กฎการแบ่ง (Classification rules, IF-THEN) การคำนวณแบบต้นไม้วิเคราะห์ (Decision Tree) การใช้สูตรทางคณิตศาสตร์ (mathematical formula) หรือโครงข่ายใยประสาทเทียม เป็นต้น ในส่วนของการทำต้นไม้วิเคราะห์จะแสดงออกมาในลักษณะของแผนภูมิโครงสร้างต้นไม้ ซึ่งก้านของต้นไม้จะแสดงถึงความรู้ที่ได้และใบไม้จะแสดงถึงประเภทชุดข้อมูลที่ถูกแบ่งออกมา แผนภูมิต้นไม้สามารถแปลงเป็นกฎการแบ่งได้ง่ายเพราะลักษณะของแผนภูมิสามารถเข้าใจได้ง่าย ในส่วนของโครงข่ายใยประสาทเทียมนั้นจะแสดงในลักษณะของการเชื่อมต่อระหว่างหน่วยที่เกิดขึ้นการทำการแบ่งประเภทนั้นมักจะใช้ประโยชน์ร่วมกับการทำนายโดยเฉพาะข้อมูลที่เป็นตัวเลข จึงอาจมองได้ว่าการทำนายเป็นการบอกถึงค่าตัวเลขและการบ่งบอกประเภทของข้อมูลนั้นในลักษณะของการดูแนวโน้ม (trends) ที่จะเกิดขึ้นตัวอย่างเทคนิคของการแบ่งประเภทและการทำนายได้แก่ การคำนวณแบบพันธุกรรม (Genetic Algorithm) การคำนวณแบบต้นไม้วิเคราะห์ และโครงข่ายใยประสาทเทียม เป็นต้น

2) อัลกอริทึมการแบ่งประเภท (Classifier Algorithms)

2.1) ต้นไม้ตัดสินใจ (Decision Tree) ต้นไม้ตัดสินใจประกอบด้วยโหนดจำนวนมากที่แทนค่าด้วยสิ่งที่ใช้ในการพิจารณาความน่าจะเป็น เช่น คุณลักษณะ และหมวดหมู่ การสร้างกิ่งสาขาจะพิจารณาจากค่าความจริงของคุณลักษณะ โดยใช้ค่าจะมาจากการคำนวณจากค่า Information Gain การสร้างต้นไม้ตัดสินใจ C4.5 ใช้ค่ามาตรฐานอัตราส่วนเกน (Gain Ratio) เพื่อเลือกคุณลักษณะที่จะใช้เป็นรากหรือโหนด โดยถ้าให้ชุดข้อมูล M ประกอบด้วยค่าที่เป็นไปได้ n ค่า คือ {m1, m2, ..., mn} และให้ความน่าจะเป็นที่จะเกิดค่า m1 มีค่าเท่ากับ P(mi) จะได้ค่าเกนสารสนเทศ (Information Gain) ของ M และถ้าให้ T คือ ข้อมูลสอนและ x คือ คุณลักษณะที่เป็นโหนด โดยมีค่าทั้งหมดที่เป็นไปได้ n ค่า โหนดปัจจุบันจะแบ่งตัวอย่าง T ออกตามกิ่งเป็น {t1, t2, ..., tn} ตามค่าที่เป็นไปได้ของ x ดังนั้นจึงสามารถคำนวณค่าเกนสารสนเทศหลังจากแบ่ง เขียนแทนด้วยสมการ (พรพล ธรรมรงค์รัตน์ และคณะ, 2551 และ Quinlan, J. R., 1993)

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

$$I_x(T) = \sum_{i=1}^{n} \frac{|t_i|}{|\mathrm{T}|} I(t_i) \quad (2)$$

$$Gain(x) = \mathrm{I(T)} = \mathrm{I_x(T)} \quad (3)$$

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

2.2) เคเนียเรสเนเบอร์ (K-Nearest Neighbor) หลักการของวิธีการนี้จะจำแนกประเภทข้อมูลโดยข้อมูลที่มีคุณสมบัติใกล้เคียงที่สุด K ตัวจากข้อมูลบนชุดข้อมูลตัวอย่างทำงานโดยขึ้นกับระยะทางน้อยสุดจากสมาชิกใหม่ หรือข้อมูลที่ป้อนถาม (input query instance) กับข้อมูลตัวอย่างฝึกฝน จะคำนวณหาเพื่อนบ้านที่ใกล้ที่สุด K ตัว หลังจากนั้นจะรวบรวมสมาชิกที่ใกล้เคียงที่สุด K ตัวแล้วเลือกคลาสที่สมาชิกส่วนใหญ่ที่ในกลุ่ม K ดังกล่าวสังกัดอยู่มากที่สุดให้กับสมาชิกใหม่ ข้อมูลการจำแนกโดยใช้ข้อมูลข้างเคียง K ตัว ประกอบด้วยแอททริบิวต์หลายตัวแปร Xi ซึ่งจะนำมาใช้ในการแบ่งกลุ่ม Yi โดยระบุค่า

ตัวเลขจำนวนเต็มบวกให้กับ K ซึ่งค่านี้จะเป็นตัวบอกจำนวนของกรณี (case) ที่จะต้องค้นหาในการทำนายกรณีใหม่ อัลกอริทึมแบบ KNN ได้แก่ 1-NN , 2-NN , 3-NN , ... K-NN ตัวอย่าง 2-KNN หมายถึง อัลกอริทึมนี้จะค้นหา 2 กรณีที่มีลักษณะใกล้เคียงกับกรณีใหม่ (2 Nearest Cases) การนำระยะทางที่หาได้จากสมาชิกในข้อมูลตัวอย่างฝึกฝน มาเรียงลำดับจากน้อยไปหามากแล้วเลือกสมาชิกที่มีระยะทาง (Distance) ใกล้เคียงที่สุดออกมา K ตัวโดยใช้การวัดระยะทางแบบ Euclidean distance มีหลักการคือ การวัดระยะทางระหว่างสองวัตถุ ถ้าวัตถุห่างกันมากแสดงว่าวัตถุนั้นมีความคล้ายกันน้อย ถ้ามีค่าน้อยก็แสดงว่ามีความคล้ายคลึงกันมาก (บุญเสริม กิจศิริกุล, 2546)

3) Deep Learning คือ ศาสตร์แขนงหนึ่งของ Machine Learning ที่เลียนแบบระบบเซลล์ประสาทในสมองของมนุษย์ (Neural Network) โดยเพิ่มความสามารถของโครงข่ายประสาทเทียม (Artificial Neural Networks – ANN) คือ การสร้างคอมพิวเตอร์ที่จำลองเอาวิธีการทำงานของสมองมนุษย์ หรือทำให้คอมพิวเตอร์รู้จักคิดและจดจำในแนวเดียวกับโครงข่ายประสาทของมนุษย์ เพื่อช่วยให้คอมพิวเตอร์ฟังภาษามนุษย์ได้เข้าใจ อ่านออก และรู้จำได้ ซึ่งอาจเรียกได้ว่าเป็น “สมองกล” เป็นหนึ่งในเทคนิคของการทำเหมืองข้อมูล (Data Mining) คือ โมเดลทางคณิตศาสตร์ สำหรับประมวลผลสารสนเทศด้วยการคำนวณแบบคอนเนคชันนิสต์ (Connectionist) เพื่อจำลองการทำงานของเครือข่ายประสาทในสมองมนุษย์ ด้วยวัตถุประสงค์ที่จะสร้างเครื่องมือซึ่งมีความสามารถในการเรียนรู้การจดจำรูปแบบ (Pattern Recognition) และการสร้างความรู้ใหม่ (Knowledge Extraction) เช่นเดียวกับความสามารถที่มีในสมองมนุษย์ แนวคิดเริ่มต้นของเทคนิคนี้ได้มาจากการศึกษาโครงข่ายไฟฟ้าชีวภาพ (Bioelectric Network) ในสมอง ซึ่งประกอบด้วย เซลล์ประสาท หรือ "นิวรอน" (Neurons) และ "จุดประสานประสาท" (Synapses) แต่ละเซลล์ประสาทประกอบด้วยปลายในการรับกระแสประสาทเรียกว่า "เดนไดรท์" (Dendrite) ซึ่งเป็น input และปลายในการส่งกระแสประสาทเรียกว่า "แอคซอน" (Axon) ซึ่งเป็นเหมือน output ของเซลล์ เซลล์เหล่านี้ทำงานด้วยปฏิกิริยาไฟฟ้าเคมี เมื่อมีการกระตุ้นด้วยสิ่งเร้าภายนอกหรือกระตุ้นด้วยเซลล์ด้วยกัน กระแสประสาทจะวิ่งผ่านเดนไดรท์เข้าสู่นิวเคลียสซึ่งจะเป็นตัวตัดสินว่าต้องกระตุ้นเซลล์อื่น ๆ ต่อหรือไม่ ถ้ากระแสประสาทแรงพอ นิวเคลียสก็จะกระตุ้นเซลล์อื่น ๆ ต่อไปผ่านทางแอคซอนของมัน เซลล์ประสาทแต่ละตัว จะเชื่อมต่อกับเส้นประสาทหลาย ๆ เส้น โดยมี เส้นประสาทและเซลล์ประสาท (เรียกว่า Node) ออกมาเป็น Artificial Neural Network (ANN) ประกอบด้วย 3 ส่วน คือ ส่วนนำเข้าข้อมูล คือ input layer ส่วนประมวลผล คือ hidden layer และส่วนผลลัพธ์หรือส่วนหลังประมวลผล คือ output layer (Dayong W., et al., 2016 และ Breiman, L., 2001)

3. วิธีดำเนินการศึกษา/การวิจัย

3.1 วัตถุประสงค์การวิจัย

การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลล็อกการใช้บริการเครือข่ายอินเทอร์เน็ตของสถาบันอุดมศึกษาระหว่างเทคนิค Decision Tree เทคนิค K-NN และเทคนิค Deep Learning โดยการทดสอบประสิทธิภาพด้วยวิธี Cross-validation Test ด้วยการแยกข้อมูลสำหรับสร้าง Training data และ Testing data

3.2 เครื่องมือในการวิจัย

เครื่องมือในการวิจัย คือ โปรแกรม Rapid Miner Studio และ Microsoft Excel โดยใช้ข้อมูลในการวิจัย เป็นข้อมูลล็อกที่ได้มีการจัดเก็บระหว่างการใช้งานเครือข่ายอินเทอร์เน็ตของสถาบันอุดมศึกษา

3.3 วิธีดำเนินการวิจัย

ผู้วิจัยดำเนินการวิจัยตามกระบวนการทำงาน Cross-Industry Standard Process for Data Mining หรือ CRISP-DM ซึ่งเป็นกระบวนการในการวิเคราะห์ข้อมูล โดยแบ่งขั้นตอนการวิจัยออกเป็น 6 ขั้นตอน (Shearer C., 2000) ดังนี้

1) Business Understanding เป็นขั้นตอนการเข้าใจปัญหา โดยเริ่มจากการค้นคว้าและศึกษาชุดข้อมูลล็อกที่ได้จากเครื่องแม่ข่าย โดยเป็นข้อมูลการใช้บริการเครือข่ายอินเทอร์เน็ตของสถาบันอุดมศึกษาในช่วงระยะเวลา 1 เดือน

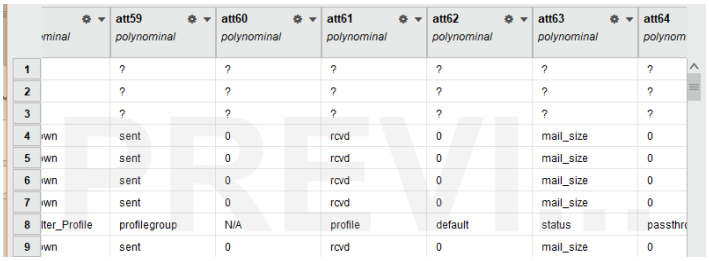
2) Data Understanding เป็นขั้นตอนทำความเข้าใจข้อมูล โดยการรวบรวมข้อมูลเพื่อหาความสัมพันธ์ของข้อมูล และตัดข้อมูลที่ไม่จำเป็นออก จากการตรวจสอบและวิเคราะห์ความสมบูรณ์ของข้อมูล พบว่า ข้อมูลที่จัดเก็บเป็นข้อมูลล็อกจะอยู่ในรูปแบบที่เข้าใจยาก โดยมีการแบ่งข้อมูลออกเป็นส่วนย่อยหรือแอททริบิวต์ที่ใช้ชื่อย่อเป็นจำนวนมาก เช่น att01, att02, ..., att60 เป็นต้น เมื่อวิเคราะห์และทำความเข้าใจข้อมูลล็อก พบว่า ข้อมูลที่สามารถใช้ในการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลการใช้บริการเครือข่ายอินเทอร์เน็ตของสถาบันอุดมศึกษาซึ่งได้จากเครื่องเซิร์ฟเวอร์ ประกอบด้วยข้อมูลที่มีความสัมพันธ์กัน จำนวน 18 แอททริบิวต์ แสดงดังตารางที่ 1

ตารางที่ 1 แอททริบิวต์ข้อมูลล็อกที่ใช้ในการวิจัย

attribute	meaning	attribute	meaning	attribute	meaning
auth	Authentication	mark	Event time	local	Local message
authpriv	Private authentication	crit	conditions to be aware	emerg	Emergency event
cron	Cron daemon	security	auth security	alert	Warning alert
deamon	System deamon	syslog	Internal log data	err	Error
kern	Kernel	user	User process	warning	Warning
mail	Mail Software / mail service	debug	detecting errors	serv	service

จากตารางที่ 1 พบว่า แอททริบิวต์ที่นำมาใช้ในการวิจัย เกี่ยวข้องกับการยืนยันตัวบุคคล การอำนวยความสะดวก การลำดับความสำคัญ และการแจ้งเตือนเหตุการณ์บางอย่าง ซึ่งมีความสัมพันธ์กับบริการที่มีการใช้งานบนเครือข่ายอินเทอร์เน็ตที่สามารถนำมาใช้ในการจำแนกข้อมูลเบื้องต้นเพื่อใช้ประกอบการพิจารณาเกี่ยวกับการให้บริการบนระบบเครือข่ายและการวางแผนการจราจรทางคอมพิวเตอร์ของหน่วยงานต่าง ๆ ได้

3) Data Preparation เป็นขั้นตอนการแปลงข้อมูล เพื่อให้เหมาะสมกับเทคนิคการจำแนกข้อมูลที่มีความหลากหลาย ทำให้ข้อมูลที่เบื้องต้นไม่สามารถใช้งานร่วมกับบางเทคนิคได้ จึงต้องทำการแปลงข้อมูล โดยเริ่มจากการทำข้อมูลในแต่ละระเบียนให้มีความถูกต้อง (data cleaning) เช่น การแปลงข้อมูลให้อยู่ในช่วงเดียวกัน หรือการเติมข้อมูลที่ขาดหายไป หรือลบข้อมูลที่ไม่มีค่าออกไป เป็นต้น ตัวอย่างข้อมูลเบื้องต้นที่จำเป็นต้องแปลงข้อมูล ดังภาพที่ 1

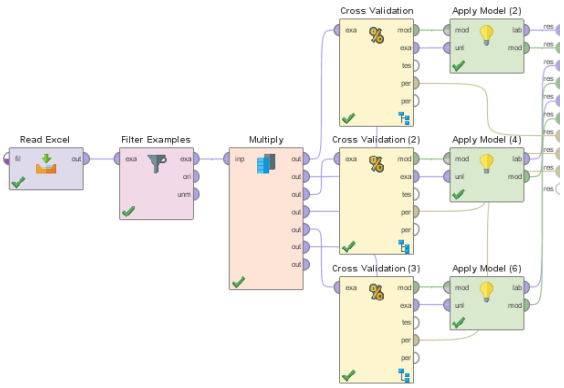


	minal	att59 polynominal	att60 polynominal	att61 polynominal	att62 polynominal	att63 polynominal	att64 polynom
1		?	?	?	?	?	?
2		?	?	?	?	?	?
3		?	?	?	?	?	?
4	wn	sent	0	rcvd	0	mail_size	0
5	wn	sent	0	rcvd	0	mail_size	0
6	wn	sent	0	rcvd	0	mail_size	0
7	wn	sent	0	rcvd	0	mail_size	0
8	lter_Profile	profilegroup	N/A	profile	default	status	passthr
9	wn	sent	0	rcvd	0	mail_size	0

ภาพที่ 1 ตัวอย่างข้อมูลเบื้องต้น

จากภาพที่ 1 เมื่อตรวจสอบประเภทของข้อมูล พบว่า มีข้อมูลบางแอททริบิวต์ไม่เหมาะสมในการนำไปใช้งาน ตัวอย่างการแปลงข้อมูลที่อยู่ในรูปแบบ polynomial ให้อยู่ในรูปแบบ Integer เช่น ข้อมูลของแอททริบิวต์ art59 ถูกเปลี่ยนจาก sent เป็น 1 และ profilegroup เป็น 2 และทำการลบระเบียนที่มีค่าเป็น ? ออก

4) Modeling เป็นขั้นตอนการวิเคราะห์ข้อมูลด้วยเทคนิคเหมืองข้อมูล โดยเลือกใช้การจำแนกประเภทข้อมูล (Classification) ในการทดลอง ประกอบด้วย เทคนิค Decision Tree ในการสร้างโมเดลต้นไม้ เทคนิค Deep Learning เป็นการเรียนรู้เชิงลึกด้วยการประมวลผลหลายชั้น และเทคนิค Naive Bayes เป็นการทำ Classification ด้วยการหาความน่าจะเป็นจากชุดข้อมูล โดยทดสอบประสิทธิภาพด้วยวิธี Cross-validation Test โดยมีรูปแบบการทดลอง ดังภาพที่ 2



ภาพที่ 2 รูปแบบการทดลองสร้างโมเดลในการวิจัย

จากภาพที่ 2 ผู้วิจัยนำเข้าข้อมูลในรูปแบบไฟล์ .xlsx คัดกรองข้อมูลเบื้องต้นด้วยเครื่องมือ Filter Examples จากนั้นจำลองชุดข้อมูลเพื่อนำใช้ในการสร้างและทดสอบโมเดลต่าง ๆ ด้วยเครื่องมือ Multiply โดยข้อมูลจะถูกส่งไปสร้างชุดข้อมูลด้วยเครื่องมือ Cross-validation Test โดยเลือกใช้โมเดลในการจำแนกข้อมูล ประกอบด้วย เทคนิค Decision Tree เทคนิค Deep Learning และเทคนิค K-NN และเรียกดูผลของการสร้างโมเดลด้วยเครื่องมือ Apply Model ซึ่งสามารถตรวจสอบประสิทธิภาพด้วยเครื่องมือ Performance

5) Evaluation เป็นขั้นตอนการทดสอบประสิทธิภาพโมเดล โดยมีข้อมูลที่สามารถใช้ในการวิจัยได้ จำนวน 8,530 ชุด ทำการแบ่งข้อมูลที่ใช้ในการวิจัยเพื่อนำไปใช้ในการทดสอบประสิทธิภาพโมเดลด้วยวิธี Cross-validation Test โดยผู้วิจัยกำหนดให้มีการแบ่งข้อมูลออกเป็น 5 ส่วน แต่ละส่วนประกอบด้วยข้อมูล จำนวน 1,706 ชุด โดยทำการเลือกข้อมูล 4 ส่วน สำหรับใช้ในการสร้างโมเดล และข้อมูล 1 ส่วน สำหรับใช้ในการทดสอบประสิทธิภาพของระบบ

6) Deployment เป็นขั้นตอนการปรับใช้ เมื่อทดสอบจนได้โมเดลที่ให้ค่าความแม่นตรง ค่าความระลึก ค่าความแม่นยำ และค่าความถูกต้องที่สามารถยอมรับได้แล้ว สามารถนำรูปแบบการจำแนกข้อมูลที่ได้จากงานวิจัยต่อไปได้

4. ผลการศึกษา/การวิจัย

ผลการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลการใช้บริการเครือข่ายอินเทอร์เน็ตของสถาบันอุดมศึกษาระหว่างเทคนิค Decision Tree เทคนิค K-NN และเทคนิค Deep Learning โดยการทดสอบประสิทธิภาพด้วยวิธี Cross-validation Test ด้วยการแยกข้อมูลสำหรับสร้าง Training data และ Testing data ใช้เกณฑ์การวัดประสิทธิภาพของตัวแบบรู้จำด้วยวิธี Predictive Modeling ซึ่งประกอบด้วย ค่าความแม่นยำ (Accuracy) ค่าความแม่นตรง (Precision) ค่าความระลึก (Recall) และค่าถ่วงดุล (F-Measure) แสดงดังตารางที่ 2

ตารางที่ 2 ผลการวัดประสิทธิภาพการจำแนกข้อมูล

Technique	Accuracy	Precision	Recall	F-Measure
Decision Tree	78.28%	77.35%	75.21%	76.26
K-NN	64.88%	63.51%	61.32%	62.39
Deep Learning	88.79%	87.55%	85.10%	86.30

จากตารางที่ 2 ผลการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลการใช้บริการเครือข่ายอินเทอร์เน็ตของสถาบันอุดมศึกษา พบว่า เทคนิค Deep Learning เป็นเทคนิคที่มีประสิทธิภาพที่สุด โดยให้ค่าความแม่นยำ 88.79% ค่าความแม่นตรง 87.55% ค่าความระลึก 85.10% และค่าถ่วงดุล 0.289

5. สรุปผลการดำเนินงาน

การหาประสิทธิภาพการจำแนกข้อมูลการใช้บริการเครือข่ายอินเทอร์เน็ตของสถาบันอุดมศึกษา ด้วยการประเมินและเปรียบเทียบประสิทธิภาพของตัวแบบระหว่าง เทคนิค Decision Tree เทคนิค K-NN และเทคนิค Deep Learning โดยการทดสอบประสิทธิภาพด้วยวิธี Cross-validation Test พบว่า เมื่อผ่านขั้นตอน Business Understanding ขั้นตอน Data Understanding และขั้นตอน Data Preparation จะมีข้อมูลที่สามารถใช้ในการวิจัยได้ จำนวน 8,530 ชุด และเมื่อทำการแบ่งข้อมูลด้วยวิธี Cross-validation Test ออกเป็น 5 ส่วน ประกอบด้วยข้อมูลส่วนละ 1,706 ชุด และทำการเลือกข้อมูล 4 ส่วน สำหรับใช้ในการสร้างโมเดล และเลือกข้อมูล 1 ส่วน สำหรับใช้ในการทดสอบประสิทธิภาพ ผลการวิจัย พบว่า เทคนิคที่ใช้ในการจำแนกข้อมูลที่มีประสิทธิภาพสูงสุด คือ เทคนิค Deep Learning เป็นเทคนิคที่มีประสิทธิภาพที่สุด โดยให้ค่าความแม่นยำ 88.79% ค่าความแม่นตรง 87.55% ค่าความระลึก 85.10% และค่าถ่วงดุล 0.289 ซึ่งเป็นระดับการประเมินที่สามารถยอมรับได้

6. อภิปรายผลการศึกษา

การหาประสิทธิภาพการจำแนกข้อมูลการใช้บริการเครือข่ายอินเทอร์เน็ตของสถาบันอุดมศึกษา พบว่า เทคนิคที่ใช้ในการจำแนกข้อมูลที่มีประสิทธิภาพสูงสุด คือ เทคนิค Deep Learning เป็นเทคนิคที่มีประสิทธิภาพมากที่สุด โดยให้ค่าความแม่นยำ 88.79% ค่าความแม่นตรง 87.55% ค่าความระลึก 85.10% และค่าถ่วงดุล 0.289 ซึ่งเป็นระดับการประเมินที่สามารถยอมรับได้ สอดคล้องกับงานวิจัยของ ไพศาล สิมาเลาเต่า และจรัญ แสนราช (Paisan Simalaotao and Charan Sanrach, 2019) ซึ่งทำการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลด้วยเทคนิค Random Forest เทคนิค Deep Learning และเทคนิค Naive Bayes ผลการวิจัยพบว่า เทคนิคที่ใช้ในการจำแนกข้อมูลปัจจัยสนับสนุนการเรียนรู้ด้วยสื่อการสอนอิเล็กทรอนิกส์ในระบบเปิดของผู้เรียนระดับอุดมศึกษา ที่มีประสิทธิภาพสูงสุด คือ Deep Learning ซึ่งได้ระดับการประเมินที่สามารถยอมรับได้

7. ข้อเสนอแนะ

7.1 ข้อเสนอแนะทางการใช้งาน สามารถนำผลการวิจัยไปใช้ในการวิเคราะห์เพื่อพัฒนาระบบสารสนเทศในการจัดเก็บข้อมูลล็อกของสถาบันอุดมศึกษา หรือนำผลการวิจัยไปใช้ในการวิเคราะห์เพื่อพัฒนาระบบสารสนเทศเฝ้าระวังและทำนายการเกิดข้อผิดพลาดที่เกิดจากการใช้งานระบบเครือข่ายอินเทอร์เน็ตของหน่วยงานต่อไป

7.2 ข้อเสนอทางการวิจัย สามารถแนวทางการวิจัยไปใช้ในการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลด้านอื่นโดยเลือกเทคนิคที่มีความเหมาะสมกับข้อมูลที่ใช้ในการวิจัยต่อไป

8. เอกสารอ้างอิง

กระทรวงเทคโนโลยีสารสนเทศและการสื่อสาร. **หลักเกณฑ์การเก็บรักษาข้อมูลจราจรทางคอมพิวเตอร์ของผู้ให้บริการ.** ราชกิจจานุเบกษา. ประกาศกระทรวงเทคโนโลยีสารสนเทศและการสื่อสาร. เล่มที่ 124 ตอน 27 ก หน้า 4 18 มิถุนายน 2550.

เดช ธรรมศิริ และ พยุง มีสัจ. (2556). Ensemble Data Classification Based on Decision Tree, Artificial Neuron Network and Support Vector Machine Optimized by Genetic Algorithm. **The Journal of KMUTNB.** Vol.21 No.2 : 239-303.

บุญเสริม กิจสิริกุล. (2546). ปัญญาประดิษฐ์, **เอกสารประกอบคำสอนวิชา 2110654**, ภาควิชาวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย.

พรพล ธรรมรงค์รัตน์ ลัดดา ปรีชาวีรกุล และวิภาดา เวทย์ประสิทธิ์. (2551). **การจำแนกประเภทเว็บเพจโดยใช้ค่าความถี่เอกสารและซัพพอร์ตเวกเตอร์แมชชีน**, The 12th National Computer Science and Engineering Conference.

วัชรินทร์ จิรโสภณ. (2555). **ระบบการจัดเก็บข้อมูลจราจรคอมพิวเตอร์.** สารนิพนธ์ สาขาวิชาเทคโนโลยีสารสนเทศ. คณะวิทยาการและเทคโนโลยีสารสนเทศ. หน้า 19

Breiman, L. (2001). Deep Learning. **Machine Learning**. Vol. 45 (October 2001): 5–32. Shearer C. (2000). The CRISP-DM model: The new blueprint for data mining. **Journal of Data Warehousing.** Vol. 5. No.4 : 13–22.

Dayong W., et al. (2016). **Deep Learning for Identifying Metastatic Breast Cancer.** CSAIL. Massachusetts Institute of Technology, p. 1-6.

Paisan Simalaotao and Charan Sanrach (2019). A Comparison of the Efficiency of Data Classification in Learning Factors through Open Educational System with Electronic Teaching Aids in Tertiary Level. **Technical Education Journ**al. Vol 10 No. 1.

Quinlan, J. R., (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.