SDU
Research Journal

การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ รุ่นที่ 3 ที่ใช้เทคนิคการแบ่งข้อมูลที่แตกต่างกัน

Comparative Efficiency of Classification Data by ID3 with Different Discretization Techniques

สุระสิทธิ์ ทรงม้า*
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสวนดุสิต

Surasit Songma*
Faculty of Science and Technology, Suan Dusit Rajabhat University

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์ 1) เพื่อเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ รุ่นที่ 3 ที่ใช้เทคนิคการแบ่งข้อมูลที่แตกต่างกัน และ 2) เพื่อทดสอบประสิทธิภาพของเทคนิคการแบ่งข้อมูลชนิดแบบไม่มีผู้สอน ชุดข้อมูลที่ใช้ในการทดลอง คือ ชุดข้อมูล KDD Cup 1999 ประสิทธิภาพของการจำแนกข้อมูลวัดจากอัตราการจำแนกข้อมูลถูกต้อง และอัตราการจำแนกข้อมูลผิดพลาด โดยใช้โปรแกรมเวก้าและแมทแลปในการประมวลผล

จากการเฉลี่ยผลการทดลองจำนวน 10 รอบ โดยใช้เทคนิคในการแบ่งข้อมูลของชุดข้อมูลฝึกสอนให้อยู่ในรูปแบบไม่ต่อเนื่องและจำแนกข้อมูลด้วยต้นไม้ตัดสินใจ รุ่นที่ 3 พบว่าการแบ่งข้อมูลด้วยขนาดความถี่เท่ากัน จำนวน 10 ชั้น มีอัตราการจำแนกข้อมูลถูกต้องสูงที่สุด เท่ากับ 99.79% รองลงมาคือการแบ่งข้อมูลด้วยการจัดกลุ่มเคมีนส์ จำนวน 40 กลุ่ม เท่ากับ 99.75% และน้อยที่สุดคือการแบ่งข้อมูลด้วยขนาดความกว้างเท่ากัน จำนวน 20 ชั้น เท่ากับ 99.57% เมื่อเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลด้วยต้นไม้ตัดสินใจ รุ่นที่ 3 โดยนำกฎที่สร้างขึ้นจากชุดข้อมูลฝึกสอนไปใช้กับชุดข้อมูลทดสอบ พบว่าประสิทธิภาพของการใช้เทคนิคแบ่งข้อมูลด้วยขนาดความกว้างเท่ากันและจำแนกข้อมูลด้วยต้นไม้ตัดสินใจ รุ่นที่ 3 มีประสิทธิภาพสูงที่สุด โดยมีอัตราการจำแนกข้อมูลถูกต้อง เท่ากับ 92.30% มีอัตราการจำแนกข้อมูลผิดพลาด เท่ากับ 4.89% และสามารถจำแนกข้อมูลที่ไม่สามารถจัดเข้ากลุ่มได้ เท่ากับ 2.81% ผลการทดลองยังแสดงให้เห็นว่าเทคนิคการแบ่งข้อมูลชนิดแบบไม่มีผู้สอนแต่ละชนิดส่งผลต่อประสิทธิภาพ

* ผู้ประสานงานหลัก (Corresponding Author)
e-mail: Surasit.songma@gmail.com

การจำแนกข้อมูลของต้นไม้ตัดสินใจ รุ่นที่ 3 และพบว่าการใช้เทคนิคการแบ่งข้อมูลร่วมกับเทคนิคต้นไม้ตัดสินใจ รุ่นที่ 3 สามารถช่วยกรองข้อมูลที่ไม่อยู่ในกฎที่สร้างขึ้นได้ ซึ่งเป็นสิ่งที่แตกต่างกับเทคนิคต้นไม้ตัดสินใจ รุ่นที่ C4.5 ฉะนั้นวิธีที่นำเสนอหากเลือกนำไปใช้ได้อย่างเหมาะสมจะเกิดประโยชน์อย่างมาก

คำสำคัญ: ข้อมูลชนิดแบบต่อเนื่อง ข้อมูลชนิดแบบไม่ต่อเนื่อง การจำแนกข้อมูล เทคนิคการแบ่งข้อมูลชนิดแบบไม่มีผู้สอน และต้นไม้ตัดสินใจ

Abstract

The objectives of this research were; 1) to compare efficiency of classification data by Iterative Dichotomiser3 (ID3) with difference discretization techniques and 2) to test efficiency of unsupervised discretization technique. We use the so-called KDD Cup 1999 data set in our experiment. Efficiency of data classification was evaluated from detection rate and false positive rate processing by applying the WEKA and MATLAB programs.

According to average result of 10 rounds, which was undertaken by discretization technique with train data set into discrete data sets and classifying the data by the ID3 algorithm, revealed that 10 bin equal width discretization technique achieved the highest correct rate 99.79%, then 40 sets of *k*-means discretization technique 99.75% and 10 bin equal frequency discretization technique 99.57% respectively. To compare efficiency of data classification using the ID3 algorithm by applying rules contributed from supervised data on trial data, the result revealed that the most efficiency technique was the equal width discretization technique and the ID3 algorithm with 92.32% of the detection rate and 4.89% of the false positive rate. The proposed technique could classify 2.81% of non-classified data. Moreover, research result presented that each technique of unsupervised discretization effected efficiency of Classification continuous data set by ID3 algorithm and found that discretization technique combine ID3 algorithm can use filtering data which not in the rules. The new technique is different from decision tree C4.5 therefore, the proposed method, if applied properly, will benefit greatly.

Keywords: Continuous Data, Discrete Data, Classification, Unsupervised Discretization and ID3 Algorithm.

บทนำ

ปัจจุบันเทคโนโลยีสารสนเทศและการสื่อสารได้เข้ามามีบทบาท และมีความสำคัญอย่างมากในการดำรงชีวิตประจำวัน การศึกษา และการติดต่อสื่อสาร การประกอบธุรกิจที่มีการเปลี่ยนแปลงไปอย่างรวดเร็วตามเทคโนโลยีที่เปลี่ยนแปลงไป (Kamonwan, 2014) ส่งผลให้มีข้อมูลจำนวนมากที่ต้องถูกจัดเก็บในรูปแบบเอกสารหรือฐานข้อมูลขนาดใหญ่ เพื่อรอการประมวลผลสำหรับนำไปใช้ประโยชน์ (Sawit, Sunthorn, & Wacharakorn, 2013) โดยกระบวนการเปลี่ยนแปลงข้อมูลดังกล่าวให้เป็นสารสนเทศนั้นจะมีการประยุกต์ใช้เทคโนโลยีเข้ามาช่วยสนับสนุนให้บริการแก่ผู้ใช้ในรูปแบบต่างๆ (Pairachnop & Doungkamol, 2014) ทั้งนี้การทำเหมืองข้อมูล (Data Mining) เป็นเครื่องมือที่นิยมใช้งานอย่างมากในกระบวนการกลั่นกรองข้อมูลที่อยู่ในฐานข้อมูลขนาดใหญ่ (Knowledge Discovery in Database : KDD) เพื่อให้ได้สารสนเทศที่ต้องการสำหรับการทำนายแนวโน้มและพฤติกรรมต่างๆ (Patchaya, 2010)

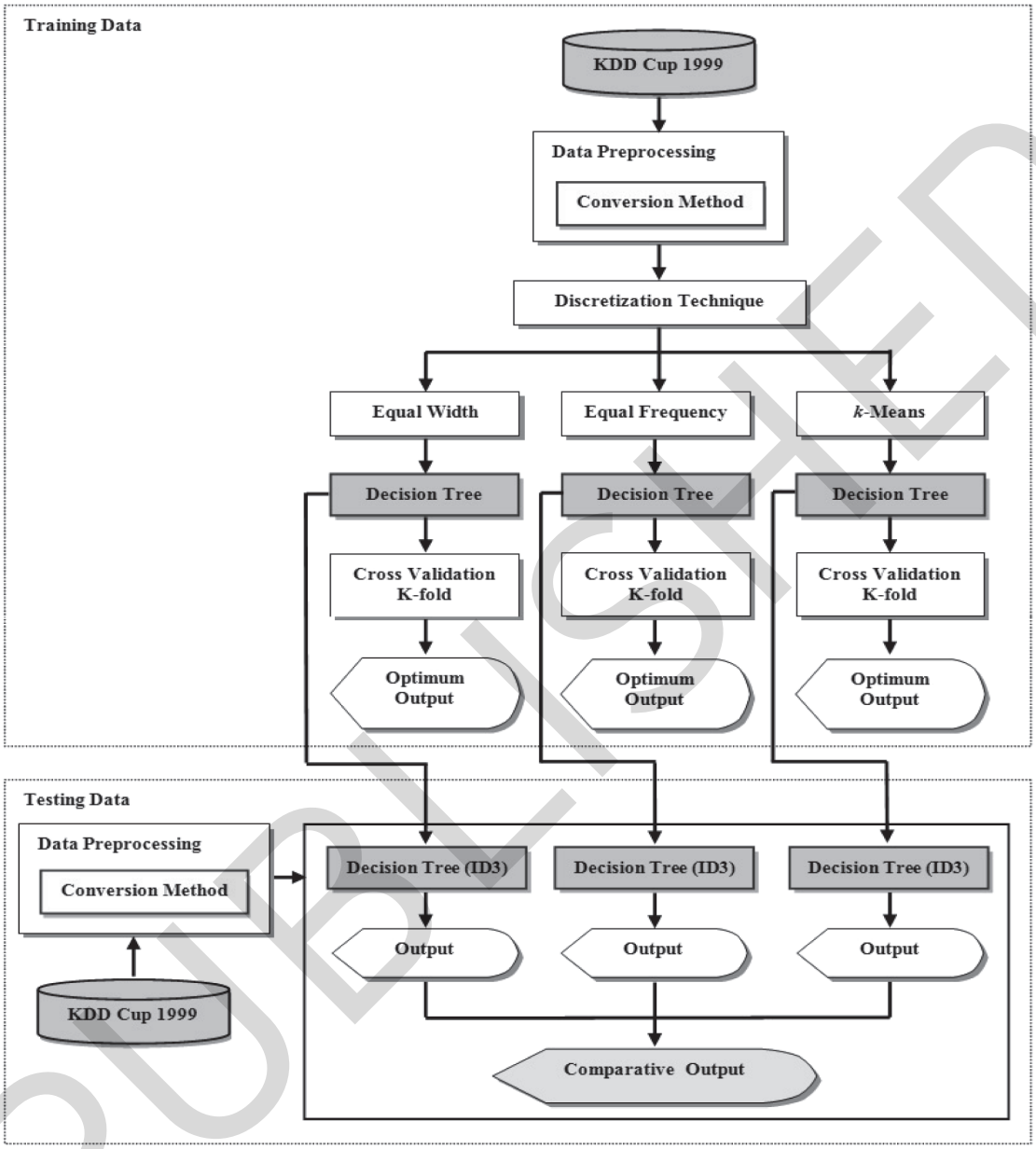
การจำแนกข้อมูล (Classification) เป็นเทคนิคหนึ่งของการทำเหมืองข้อมูลที่จำเป็นต้องมีการเรียนรู้ข้อมูลในอดีต (Supervised Learning) เพื่อสร้างแบบจำลองสำหรับใช้ทดสอบกับข้อมูลใหม่โดย (Rajashree, Rajib & Rasmita, 2011) กล่าวว่าการเรียนรู้ของเครื่อง (Machine Learning) และการทำเหมืองข้อมูลส่วนใหญ่จะรองรับข้อมูลที่มีคุณลักษณะ (Feature) ที่เป็นลักษณะไม่ต่อเนื่อง (Discrete Data) เท่านั้น แต่ในโลกของความเป็นจริงแล้วข้อมูลที่จะนำมาใช้จะมีคุณลักษณะเป็นข้อมูลแบบต่อเนื่อง (Continuous Data) ฉะนั้นจึงจำเป็นต้องมีวิธีการแบ่งข้อมูลแบบต่อเนื่องให้เป็นช่วงๆ เพื่อทำให้ข้อมูลอยู่ในลักษณะข้อมูลไม่ต่อเนื่อง เพื่อจะได้นำไปใช้กับอัลกอริทึมที่รองรับข้อมูลเฉพาะชนิดไม่ต่อเนื่องได้ สำหรับการทำเหมืองข้อมูลนั้นมีอยู่หลายวิธี โดยงานวิจัยนี้เราได้เลือกใช้เทคนิคต้นไม้ตัดสินใจ (Decision Tree) รุ่นที่ 3 ซึ่งสอดคล้องกับ (Prashant, Sanket, Kunal, Hemant & Abhilasha, 2015) ที่กล่าวว่าเป็นเทคนิคที่นิยมใช้งานกันอย่างแพร่หลาย เนื่องจากผู้ใช้สามารถทำความเข้าใจผลลัพธ์ได้ง่ายโดยเป็นเทคนิคที่ได้รับการพัฒนาขึ้นโดย J. R. Quinlanในปี ค.ศ. 1986 มีลักษณะโครงสร้างข้อมูลเป็นลำดับชั้น (Hierarchy) ในปัจจุบันได้มีการพัฒนาต้นไม้ตัดสินใจออกมาหลายรุ่น และรุ่นที่นิยมใช้กันคือ Iterative Dichotomiser 3 หรือนิยมเรียกว่า ไอดีสาม (ID3) (Chinnapat, 2010) โดยคำว่าต้นไม้ตัดสินใจในบทความนี้ผู้วิจัยขอใช้ในความหมายของต้นไม้ตัดสินใจ รุ่นที่ 3

การสร้างต้นไม้ตัดสินใจใช้หลักการของทฤษฎีข่าวสาร โดยเป็นการสร้างต้นไม้จากบนลงล่าง โดยจะเลือกจากคุณลักษณะที่มีค่าอินฟอร์เมชันเกน (Information gain) สูงสุดมาเป็นโหนดบนสุดหรือโหนดเริ่มต้นและข้อมูลถัดไปเป็นค่าลดหลั่นกันตามลำดับ โดยข้อจำกัดของเทคนิคต้นไม้ตัดสินใจ รุ่นที่ 3 นั้นคือจะไม่สามารถจำแนกข้อมูลแบบต่อเนื่อง จึงได้มีการพัฒนาต้นไม้ตัดสินใจใน รุ่นที่ 4.5 ขึ้นมาเพื่อแก้ไขปัญหาดังกล่าว คือจะสามารถรองรับข้อมูลได้ทั้งที่เป็นแบบต่อเนื่องและแบบไม่ต่อเนื่อง แต่อย่างไรก็ตามก็ยังมีงานบางชนิดยังจำเป็นต้องใช้ต้นไม้ตัดสินใจ รุ่นที่ 3 ในการแก้ไขปัญหา เช่น การสร้างต้นไม้ตัดสินใจเพื่อจำแนกข้อมูลที่มีอยู่ในกฎเท่านั้นและหากตรวจพบว่าไม่มีอยู่ในกฎก็จะไม่สามารถจำแนกได้ โดยจะแสดงเป็นกลุ่มไม่รู้จัก สำหรับต้นไม้ตัดสินใจรุ่นที่ 4.5 นั้นจะไม่สามารถทำได้ ผู้วิจัยจึงมีแนวคิดในการปรับปรุงให้เทคนิคต้นไม้ตัดสินใจ รุ่นที่ 3 ให้สามารถจำแนกข้อมูลชนิดแบบต่อเนื่องได้ และศึกษาประสิทธิภาพของเทคนิคการแบ่งข้อมูลชนิดการเรียนรู้แบบไม่มีผู้สอนที่ทำงานร่วมกับเทคนิคต้นไม้ตัดสินใจว่าเทคนิคใดมีประสิทธิภาพดีที่สุด

วัตถุประสงค์

1. เพื่อเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ รุ่นที่ 3 ที่ใช้เทคนิคการแบ่งข้อมูลที่แตกต่างกัน
2. เพื่อทดสอบประสิทธิภาพของเทคนิคการแบ่งข้อมูลชนิดแบบไม่มีผู้สอน

กรอบแนวคิด



ภาพที่ 1 กรอบแนวคิดในการวิจัย

วิธีการวิจัย

การวิจัยครั้งนี้ผู้วิจัยใช้วิธีการวิจัยเชิงทดลอง (Experimental Research) เครื่องมือที่ใช้ในการทดลองประกอบด้วยเครื่องคอมพิวเตอร์ที่มีความเร็วของหน่วยประมวลผลกลาง 2.10 GHz. รุ่น Intel® Core™ i7-3612QM หน่วยความจำ (DDR RAM) ขนาด 16 GB ใช้ระบบปฏิบัติการวินโดว์เซเว่น (Windows 7) และมีการใช้โปรแกรมเวก้า (WEKA) และโปรแกรมแมทแลป (MATLAB) ในการประมวลผล ซึ่งได้แบ่งขั้นตอนการทำงานออกเป็น 4 ขั้นตอนดังนี้

1. การเตรียมข้อมูล

ผู้วิจัยใช้ชุดข้อมูล KDD Cup 1999 ซึ่งเป็นข้อมูลมาตรฐานสำหรับการทดสอบระบบการรักษาความปลอดภัย สามารถแบ่งข้อมูลออกเป็น 2 ส่วน คือ ชุดข้อมูลฝึกสอน (Train Data) และชุดข้อมูลทดสอบ (Test Data) ภายในระเบียนจะเป็นการติดต่อสื่อสารข้อมูลที่อยู่ในลักษณะทีซีพีแพ็คเก็ตจากต้นทางไปยังปลายทาง ซึ่งมีการกำหนดโปรโตคอลที่ใช้ในการสื่อสาร หมายเลขไอพีแอดเดรสต้นทาง และหมายเลขแอดเดรสปลายทางไว้อย่างชัดเจน แต่ละระเบียนจะมีขนาดประมาณ 100 ไบต์ และพบว่าในแต่ละระเบียนประกอบด้วยคุณลักษณะที่เป็นคำตอบว่าระเบียนดังกล่าวเป็นปกติ (Normal) หรือการบุกรุก (Attack) จำนวน 1 ตัว และคุณลักษณะอื่นๆอีกจำนวน 41 ตัว ซึ่งแต่ละตัวจะมีความหมายแตกต่างกันออกไป เช่น ระยะเวลาในการเชื่อมต่อ (Duration) ชนิดของโปรโตคอล (Protocol Type) รวมไปถึงชนิดของการให้บริการ (Service) คุณลักษณะเหล่านี้มี 2 รูปแบบ คือ แบบต่อเนื่อง (Continuous) จำนวน 38 ตัว และแบบไม่ต่อเนื่อง (Discrete) จำนวน 3 ตัว ผลจากการวิเคราะห์ข้อมูลดังกล่าวโดยละเอียด พบว่าระเบียนมีความซ้ำซ้อนกันอยู่จำนวนมากจึงได้ทำการลดความซ้ำซ้อนของระเบียนตามหลักความน่าจะเป็นในการสุ่มโดยสุ่มตัวอย่างแบบแบ่งชั้นภูมิ (Stratified Sampling) เพื่อให้สะดวกต่อการดำเนินงานในขั้นตอนต่อไป โดยผลการลดความซ้ำซ้อนของระเบียนในชุดข้อมูลแสดงได้ดังตารางที่ 1-2 ซึ่งชุดข้อมูลดังกล่าวสามารถดาว์นโหลดได้จากเว็บไซต์ http://kdd.ics.uci.edu/databases/kddcup99/ โดยข้อมูลภายในระเบียนมีลักษณะเป็นตัวเลขผสมกับตัวอักษรและขั้นกลางด้วยเครื่องหมายมหัพภาค

ตารางที่ 1 อัตราส่วนของระเบียนข้อมูลฝึกสอน

ที่	ประเภทระเบียน	ระเบียนเดิม		ระเบียนที่ลดความซ้ำซ้อนแล้ว	
		จำนวน	อัตราส่วน(%)	จำนวน	อัตราส่วน (%)
1	ปกติ	97,278	19.69	87,832	60.33
2	การบุกรุก	396,743	80.31	57,754	39.67
	รวมทั้งสิ้น	494,021	100	145,586	100

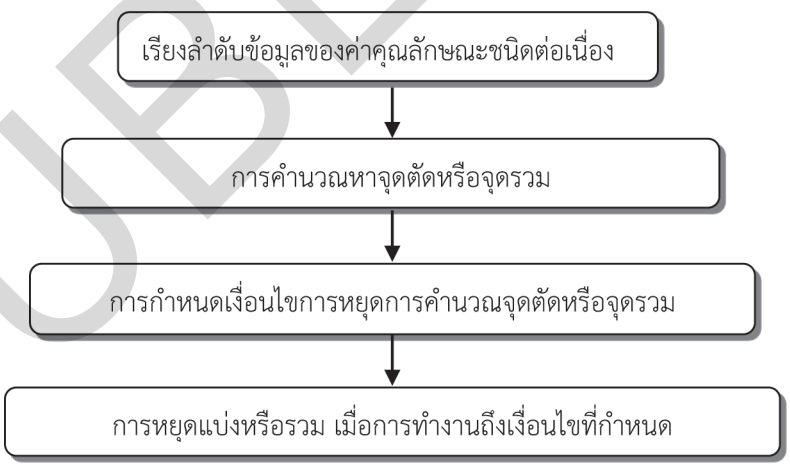
ตารางที่ 2 อัตราส่วนของระเบียนข้อมูลทดสอบ

ที่	ประเภทระเบียน	ระเบียนเดิม		ระเบียนที่ลดความซ้ำซ้อนแล้ว	
		จำนวน	อัตราส่วน(%)	จำนวน	อัตราส่วน (%)
1	ปกติ	60,593	19.48	47,913	61.99
2	การบุกรุก	250,436	80.51	29,378	38.01
	รวมทั้งสิ้น	311,029	100	77,291	100

ทั้งนี้มีการปรับค่าคุณลักษณะ (Conversion) ของชุดข้อมูลที่เป็นตัวอักษรให้อยู่ในรูปของตัวเลข ซึ่งมีอยู่จำนวน 3 คุณลักษณะ โดยคุณลักษณะเหล่านี้จะถูกแทนที่เป็นตัวเลขตั้งแต่ 1 ถึง N เมื่อ N คือจำนวนของตัวอักษรทั้งหมดที่มีในคุณลักษณะนั้น (Kandeeban, Selvakani & Rajesh, 2009)

2. กระบวนการแบ่งข้อมูลชนิดแบบไม่มีผู้สอน

(Hemada & Vijaya, 2013) อธิบายถึงเทคนิคการแบ่งข้อมูลว่าเป็นกระบวนที่ใช้ในการแปลงข้อมูลจากคุณลักษณะที่เป็นค่าต่อเนื่องให้อยู่ในลักษณะค่าไม่ต่อเนื่อง โดยมีจุดมุ่งหมายหลักเพื่อให้สามารถทำเข้าใจข้อมูลได้ง่ายขึ้น ลดเวลา และรายจ่ายอื่นๆในการนำไปใช้งาน อีกทั้งยังช่วยเพิ่มความแม่นยำและประสิทธิภาพในการจำแนกประเภท กระบวนการทำงานหลักของการแบ่งข้อมูลจะมี 4 ส่วน คือ เรียงลำดับข้อมูลของค่าคุณลักษณะชนิดต่อเนื่อง การคำนวณหาจุดตัดหรือจุดรวม การกำหนดเงื่อนไขการหยุดการคำนวณจุดตัดหรือจุดรวม และการหยุดแบ่งหรือรวมเมื่อการทำงานถึงเงื่อนไขที่กำหนด ดังภาพที่ 2



ภาพที่ 2 ขั้นตอนการทำงานของเทคนิคการแบ่งข้อมูล

ผู้วิจัยเลือกใช้เทคนิคการแบ่งข้อมูลแบบไม่มีผู้สอนที่นิยมใช้งาน ซึ่งสอดคล้องกับผลการวิจัยที่ผ่านมา (Palaniappan & Hong, 2009) และ (Rajashree, Rajib & Rasmita, 2011) ประกอบไปด้วย การแบ่งข้อมูลด้วยขนาดความกว้างเท่ากัน (Equal Width) การแบ่งข้อมูลด้วยขนาดความถี่เท่ากัน (Equal Frequency) และการแบ่งข้อมูลด้วยการจัดกลุ่มเคมีนส์ (*k*-means Cluster) โดยมีรายละเอียดการทำงานดังนี้

2.1 การแบ่งข้อมูลด้วยขนาดความกว้างเท่ากัน

เป็นกระบวนการแบ่งข้อมูลโดยกำหนดขนาดความกว้างข้อมูลให้เท่ากันทุกชั้นของคุณลักษณะทั้งหมด เรียกจำนวนชั้นนี้ว่าค่า *k* โดยค่า *k* จะถูกกำหนดจากผู้ใช้งานและต้องเป็นค่าที่เหมาะสมจึงจะทำให้ประสิทธิภาพการจำแนกข้อมูลด้วยต้นไม้ตัดสินใจได้ผลดี มีขั้นตอนการทำงานมีดังนี้

1) เรียงลำดับข้อมูลของค่าคุณลักษณะชนิดต่อเนื่อง *(V)*
2) คำนวณหาค่าต่ำสุดของแต่ละคุณลักษณะ (V_{min})
3) คำนวณหาค่าสูงสุดของแต่ละคุณลักษณะ (V_{max})
4) ค่าพิสัยหรือจำนวนชั้นหาได้จากคำนวณได้จากสมการ

$$Interval = \frac{(V_{max} - V_{min})}{k} \quad (1)$$

$$Boundaries = V_{min} + (i + Interval) \quad (2)$$

โดยที่ *Boundaries* สามารถมีได้ตั้งแต่ $i = 1$ ถึง $k - 1$

2.2 การแบ่งข้อมูลด้วยขนาดความถี่เท่ากัน

เป็นกระบวนการที่มีลักษณะคล้ายกับการแบ่งข้อมูลด้วยขนาดความกว้างเท่ากัน ยกเว้นในส่วนของของค่าของคุณลักษณะจะนับค่าเดียวหากมีการซ้ำกันของข้อมูล (Unique Values) โดยค่าของจำนวนชั้นที่แบ่งจะถูกกำหนดโดยผู้ใช้งาน เรียกว่าค่า *n* สมาชิกในแต่ละชั้นจะคำนวณได้จากสมการ

$$Interval\ Frequency = \frac{nb\_unique\_values}{n} \quad (3)$$

โดยที่ *nb_unique_values* คือ จำนวนข้อมูลที่ไม่นับค่าซ้ำของคุณลักษณะแบบต่อเนื่อง

2.3 การแบ่งข้อมูลด้วยการจัดกลุ่มเคมีนส์

ผู้วิจัยเลือกใช้เทคนิคการจัดกลุ่มแบบเคมีนส์ (*k*-means) เนื่องจากเป็นเทคนิคการจัดกลุ่มที่ได้รับความนิยมตั้งแต่อดีตจนถึงปัจจุบันและมีการใช้งานกันอย่างแพร่หลาย ซึ่งสอดคล้องกับงานผลงานวิจัยของ (Surasit, Witcha, Kiattsak & Parinya, 2012a) โดย (Sirapat, 2009) ได้อธิบายว่าเป็นเทคนิคที่

จัดอยู่ในกลุ่มการเรียนรู้แบบไม่มีผู้สอน หรือวิธีการที่ไม่นำคุณลักษณะที่เป็นคำตอบของระเบียนมาใช้ในการพิจารณา โดยจะเป็นการพยายามจัดกลุ่มระเบียนที่มีความคล้ายคลึงกันหรืออยู่ใกล้กันมากที่สุด (Intra-Cluster Distance) ไว้กลุ่มหนึ่ง และกลุ่มระเบียนที่มีความแตกต่างกันมากที่สุด (Inter-Cluster Distance) ไว้อีกกลุ่มหนึ่งการจัดข้อมูลเข้ากลุ่มพิจารณาจากการนำข้อมูลที่มีคุณสมบัติหรือลักษณะเหมือนกันหรือมีความคล้ายกันมาจัดไว้ในกลุ่มเดียวกัน ซึ่งมีการคำนวณความเหมือนของกลุ่มข้อมูลหรือความใกล้กันของข้อมูล โดยใช้มาตรวัดความคล้าย (Similarity Measure) หรือมาตรวัดระยะห่าง (Distance Measure) ผู้วิจัยเลือกใช้มาตรวัดที่นิยมใช้กันมากที่สุด คือ มาตรวัดยูคลิค ซึ่งสอดคล้องกับงานวิจัยของ (Witcha, 2008) ซึ่งนำมาใช้วัดระยะห่าง (Distance) เพื่อจัดกลุ่มข้อมูล และ (Watthananon & Mingkhwan, 2012) ได้สนับสนุนคำกล่าวข้างต้นว่าการจัดกลุ่มแบบเคมีนส์มีขั้นตอนวิธีในการปฏิบัติที่ง่ายและได้ผลดี (Supawee, 2011) ได้สรุปขั้นตอนการจัดกลุ่มแบบเคมีนส์ไว้ดังนี้

1) สุ่มข้อมูลออกเป็นกลุ่มย่อย จำนวน *k* กลุ่ม โดยที่ค่า *k* ต้องเป็นเลขจำนวนเต็มและมีค่าน้อยกว่าจำนวนข้อมูลทั้งหมด

2) เมื่อได้กลุ่มข้อมูลแต่ละกลุ่มแล้ว ทำการคำนวณหาจุดเซนทรอยด์ โดยใช้ค่าเฉลี่ยเลขคณิต (Arithmetic Mean)

3) ตรวจดูระยะห่างระหว่างข้อมูลกับจุดเซนทรอยด์ โดยข้อมูลใดที่มีระยะห่างสั้นที่สุดก็นำข้อมูลไปสังกัดกลุ่มข้อมูลดังกล่าว (ทำให้เกิดการย้ายกลุ่มข้อมูล)

4) ทำการคำนวณจุดเซนทรอยด์ และกำหนดข้อมูลให้กับกลุ่มข้อมูลใหม่ไปเรื่อยๆ จนกว่าข้อมูลในแต่ละกลุ่มไม่สามารถเปลี่ยนกลุ่มได้ดีกว่าเดิม

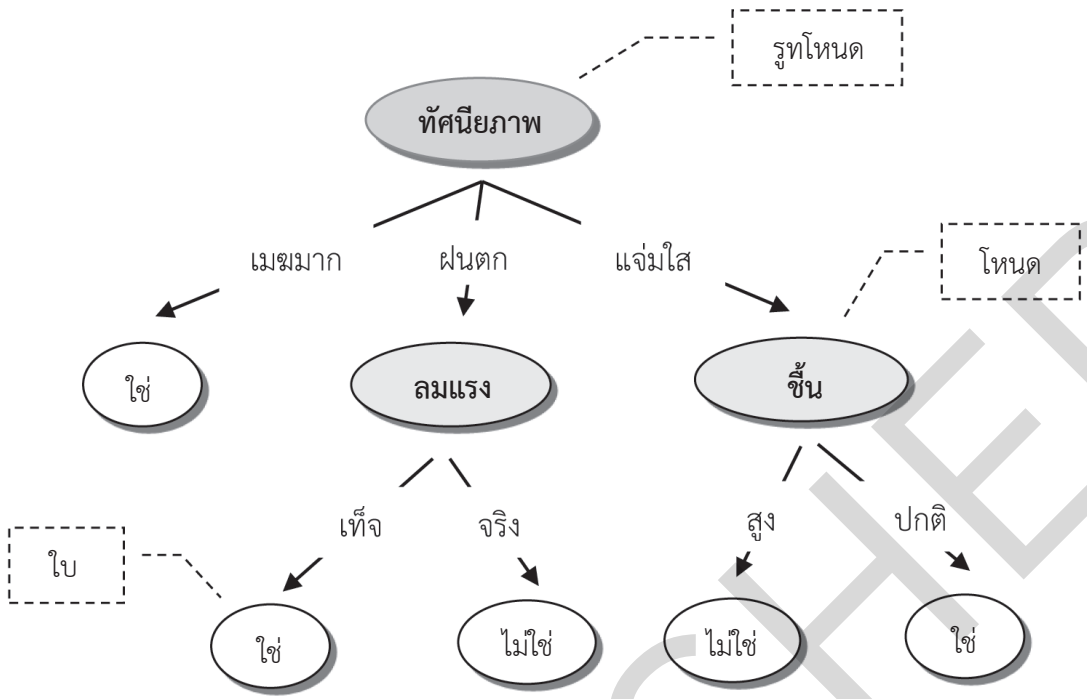
3. การจำแนกข้อมูลโดยใช้เทคนิคต้นไม้ตัดสินใจ

เทคนิคต้นไม้ตัดสินใจจัดอยู่ในกลุ่มการเรียนรู้แบบมีผู้สอน หรือวิธีการที่มีการนำคุณลักษณะที่เป็นคำตอบของระเบียนมาใช้ในการพิจารณาด้วย มีลักษณะโครงสร้างข้อมูลเป็นลำดับชั้น ซึ่งเป็นที่นิยมใช้กันอย่างแพร่หลาย และอาศัยหลักการสร้างกฎในรูปแบบ IF-THEN ต้นไม้ตัดสินใจประกอบด้วยโหนด (Node) กิ่ง (Branch) และใบ (Leaf) แสดงได้ดังภาพที่ 3 โดยขอยกตัวอย่างการจำแนกข้อมูลโดยใช้ต้นไม้ตัดสินใจว่าจะสามารถเล่นเทนนิสได้ ใช่หรือไม่ใช่ โดยพิจารณาจากองค์ประกอบด้านทัศนียภาพของอากาศ ด้านความแรงของลม และด้านความชื้นของอากาศ เช่น กฎที่สร้างขึ้นจากเทคนิคต้นไม้ตัดสินใจ กรณีที่สามารถเล่นเทนนิสได้

IF ทัศนียภาพ = “เมฆมาก” THEN “ใช่”

IF ทัศนียภาพ = “ฝนตก” and ลมแรง = “เท็จ” THEN “ใช่”

IF ทัศนียภาพ = “แจ่มใส” and ความชื้น = “ปกติ” THEN “ใช่”



ภาพที่ 3 ส่วนประกอบของต้นไม้ตัดสินใจ

หลักการพื้นฐานของการสร้างต้นไม้ตัดสินใจ รุ่น 3 คือ การสร้างรูทโหนดลงมาก่อน แล้วตามด้วยโหนดใบจนถึงกิ่ง และสร้างจากบนลงล่าง (Han & Kamber, 2006) ได้อธิบายขั้นตอนการสร้างต้นไม้ตัดสินใจดังนี้

1) เริ่มต้นสร้างโหนดขึ้นมาหนึ่งโหนดจากชุดข้อมูล

2) ถ้าข้อมูลทั้งหมดอยู่ในกลุ่มเดียวกันแล้ว ให้โหนดที่สร้างขึ้นนั้นเป็นโหนดใบและตั้งชื่อแยกตามกลุ่มของข้อมูลนั้น

3) ถ้าข้อมูลไม่มีคุณลักษณะใดที่เหมาะสมในการแบ่งกลุ่ม ให้โหนดที่สร้างขึ้นนั้นเป็นโหนดใบและตั้งชื่อตามกลุ่มที่มีข้อมูลสนับสนุนมากที่สุด

4) ถ้าข้อมูลภายในโหนดมีหลากหลายกลุ่มปะปนกันให้ทำการเลือกคุณลักษณะที่มีความเหมาะสมที่สุด โดยพิจารณาจากอินฟอร์เมชันเกนซึ่งเป็นตัวชี้วัดความสามารถในการจำแนกกลุ่ม

5) เมื่อได้ตัวทดสอบการตัดสินใจ ให้สร้างกิ่งของต้นไม้ด้วยค่าต่างๆ ที่เป็นไปได้ของตัวทดสอบและแบ่งข้อมูลออกตามกิ่งต่างๆ ที่สร้างขึ้น

6) พิจารณาข้อมูลในแต่ละกิ่ง ถ้าพบว่าข้อมูลทั้งหมดอยู่ในกลุ่มเดียวกัน ให้ต่อกิ่งด้วยโหนดใบและกำหนดค่าด้วยกลุ่มของข้อมูลนั้น แต่ถ้าพบว่าข้อมูลมีหลากหลายกลุ่มปะปนกัน ให้ทำการวนซ้ำการหาตัวทดสอบการตัดสินใจที่เหมาะสมต่อไป

7) ทำการวนซ้ำเพื่อแบ่งข้อมูลและแตกกิ่งของต้นไม้ไปเรื่อยๆ โดยการวนซ้ำจะสิ้นสุดก็ต่อเมื่อเงื่อนไขข้อใดข้อหนึ่งต่อไปนี้เป็นจริง

7.1) ข้อมูลทั้งหมดในโหนดอยู่ในกลุ่มเดียวกัน ให้โหนดที่สร้างขึ้นนั้นเป็นโหนดใบและตั้งชื่อกลุ่มของข้อมูลนั้น

7.2) ไม่มีคุณลักษณะใดที่เหมาะสมในการแบ่งกลุ่ม ให้โหนดที่สร้างขึ้นนั้นเป็นโหนดใบและกำหนดค่าด้วยกลุ่มที่มีข้อมูลสนับสนุนมากที่สุด

(Eakasit, 2014) อธิบายถึงการสร้างกฎด้วยต้นไม้ตัดสินใจว่าจะทำการคัดเลือกคุณลักษณะที่มีความสัมพันธ์กับคลาสมากที่สุดเป็นรูทโหนด การหาความสัมพันธ์ของคุณลักษณะนี้จะใช้ตัววัดที่เรียกว่าอินฟอร์เมชันเกน ซึ่งถ้าคุณลักษณะใดมีค่าอินฟอร์เมชันเกนสูง แสดงว่าคุณลักษณะนั้นสามารถจำแนกกลุ่มได้ดี ช่วยลดจำนวนครั้งของการทดสอบในการแยกแยะข้อมูลและรับประกันว่าต้นไม้ตัดสินใจที่ได้จะไม่มีความซับซ้อนมากจนเกินไป โดยค่าอินฟอร์เมชันเกนคำนวณได้จากสมการ

$$IG\ (\text{parent, child}) = Entropy\ (\text{parent}) - \sum_{i} [p(c_i) \times Entropy\ (c_i)] \qquad (4)$$

โดยที่ $Entropy\ (c_i) : = - p\ (ci)\ \log_2 p(c_i)$ คือ เอนโทรปีของ c_i

$p(c_i)$ คือ ค่าความน่าจะเป็นของ c_i

เอนโทรปีนี้จะใช้ในการวัดความแตกต่างของข้อมูล นั่นคือ ถ้าข้อมูลมีความแตกต่างกันน้อยเอนโทรปีจะมีค่าต่ำ แต่ถ้าข้อมูลมีความแตกต่างกันมาก เอนโทรปีจะมีค่าสูง ดังนั้นถ้าโหนดลูก (Child) สามารถแบ่งแยกข้อมูลได้ดีจะมีค่าเอนโทรปีต่ำ และจะทำให้ค่าอินฟอร์เมชันเกนสูงเมื่อเทียบกับโหนดแม่ (Parent) ในขั้นตอนการสร้างแบบจำลองของต้นไม้ตัดสินใจ จะคำนวณอินฟอร์เมชันเกนของแต่ละคุณลักษณะเทียบกับคลาสเพื่อหาคุณลักษณะที่มีค่าอินฟอร์เมชันเกนมากที่สุดและให้เป็นรูทโหนดของต้นไม้ตัดสินใจ

4. การวัดประสิทธิภาพ

ผู้วิจัยเลือกใช้วิธีการวัดประสิทธิภาพที่นิยมใช้ ซึ่งสอดคล้องกับ (Kohavi, 1995) โดยมีรายละเอียดยังนี้

4.1 วิธีการแยกทดสอบ (Split Test) เพื่อใช้วัดประสิทธิภาพการทำจำแนกข้อมูลสำหรับชุดข้อมูลทดสอบ โดยจะมีการแบ่งข้อมูลด้วยวิธีการสุ่มออกมาเป็น 2 ส่วน เช่น 70% ต่อ 30% โดย 70% จะเป็นชุดข้อมูลที่ใช้ในการฝึกสอน และ 30% ใช้ในการทดสอบ ซึ่งในงานวิจัยนี้ใช้ชุดข้อมูลมาตรฐาน ซึ่งมีการแบ่งข้อมูลสำหรับฝึกสอนและทดสอบไว้ดังรายละเอียดในตารางที่ 1-2

4.2 วิธีการตรวจสอบไขว้ (Cross-Validation) เพื่อใช้สร้างแบบจำลองเพื่อวัดประสิทธิภาพของชุดข้อมูลฝึกสอนโดยจะมีการแบ่งข้อมูลออกเป็นหลายส่วน ในวิจัยนี้เรียกว่า k -Fold Cross Validation ซึ่งค่า k คือจำนวนที่ชุดข้อมูลที่แบ่งโดยงานวิจัยนี้กำหนดให้ k = 10 หมายถึง แบ่งข้อมูลออกเป็น 10 ชุดเท่าๆ กันและในแต่ละการวนรอบจะมีการใช้ชุดข้อมูล จำนวน 1 ชุด สำหรับใช้ทดสอบ และจำนวน 9 ชุดสำหรับใช้ฝึกสอน โดยจะมีการนำผลลัพธ์ของแต่ละรอบการคำนวณมาหาค่าเฉลี่ยร่วมกันเพื่อใช้วัดประสิทธิภาพแบบจำลอง ภาพที่ 4 เป็นการแสดงตัวอย่างวิธีการตรวจสอบไขว้ แบบ 5-Fold Cross Validation

	ชุดข้อมูลฝึกสอน				ชุดข้อมูลทดสอบ
รอบที่ 1	2	3	4	5	1
รอบที่ 2	1	3	4	5	2
รอบที่ 3	1	2	4	5	3
รอบที่ 4	1	2	3	5	4
รอบที่ 5	1	2	3	4	5

ภาพที่ 4 วิธีการตรวจสอบไขว้แบบ 5-fold Cross Validation

4.3 การวัดประสิทธิภาพของแบบจำลองวัดจากความอัตราการจำแนกข้อมูลที่ถูกต้อง (DTR) และอัตราการจำแนกข้อมูลที่ผิดพลาด (FPR) ซึ่งสอดคล้องกับผลงานวิจัยของ (Surasit, Witcha, Kiattsak & Parinya, 2012b) มีสมการสำหรับการคำนวณดังนี้

$$DTR = [TP/(TP + FN)] \times 100\% \quad (5)$$

$$FPR = [FP/(TN + FP)] \times 100\% \quad (6)$$

โดยที่

TP คือ จำนวนระเบียนที่เป็นการบุกรุกและจำแนกได้ถูกต้องว่าเป็นระเบียนการบุกรุก

TM คือ จำนวนระเบียนทีเป็นปกติและจำแนกได้ถูกต้องว่าเป็นระเบียนปกติ

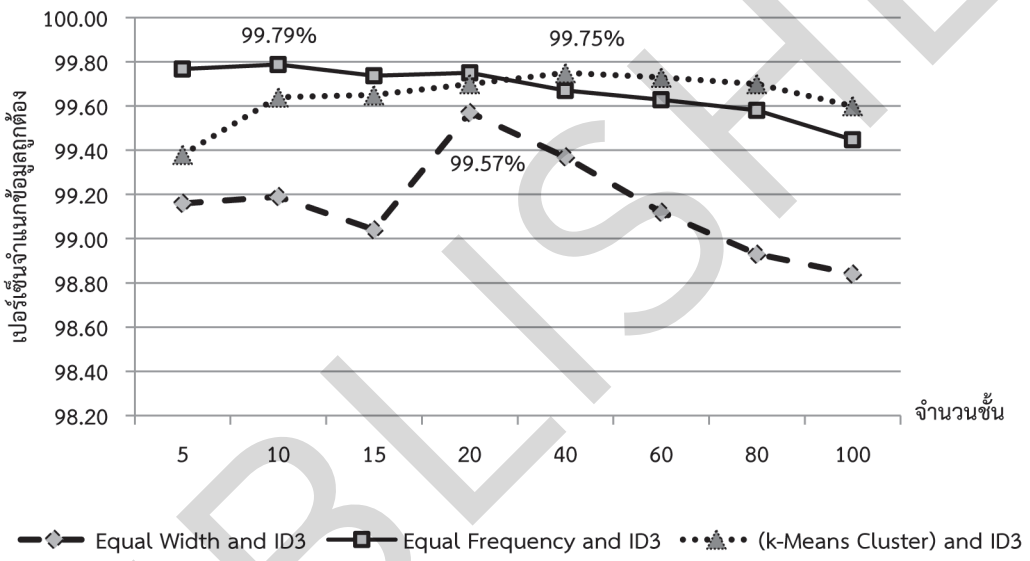
FP คือ จำนวนระเบียนที่เป็นปกติและจำแนกผิดว่าเป็นระเบียนการบุกรุก

FN คือ จำนวนระเบียนที่เป็นการบุกรุกและจำแนกผิดว่าเป็นระเบียนปกติ

ผลการวิจัย

1. ผลการจำแนกข้อมูลของเทคนิคต้นไม้ตัดสินใจกับชุดข้อมูลฝึกสอน

ข้อมูลที่ใช้ในการวิจัยเป็นข้อมูลชนิดต่อเนื่อง ซึ่งไม่สามารถนำไปใช้กับต้นไม้ตัดสินใจ รุ่นที่ 3 ได้ ทำให้จำเป็นต้องมีการแบ่งข้อมูลให้อยู่ในรูปแบบไม่ต่อเนื่อง โดยผู้วิจัยใช้เทคนิคการแบ่งข้อมูลแบบไม่มีผู้สอน ทั้งนี้จะต้องกำหนดจำนวนชั้นหรือกลุ่มของข้อมูลให้เหมาะสมด้วยตนเอง ซึ่งหากกำหนดไม่เหมาะสมจะทำให้ประสิทธิภาพในการจำแนกข้อมูลของต้นไม้ตัดสินใจไม่ดี ผู้วิจัยจึงได้ทำการทดสอบเพื่อหาค่าจำนวนชั้นที่ดีที่สุด โดยมีการเปลี่ยนค่าจำนวนชั้น ตั้งแต่ 5, 10, 15, 20, 40, 60, 80 และ 100 โดยในการทำแต่ละรอบจะใช้วิธีการตรวจสอบไขว้ (10-Fold Cross Validation) ผลการจำแนกข้อมูลที่ถูกต้องด้วยต้นไม้ตัดสินใจ แสดงได้ภาพที่ 5



ภาพที่ 5 ผลการจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจของชุดข้อมูลฝึกสอน

จากภาพที่ 5 แสดงให้เห็นว่าการใช้เทคนิคในการแบ่งข้อมูลเพื่อแบ่งข้อมูลแบบต่อเนื่องของชุดข้อมูลฝึกสอนให้อยู่ในรูปแบบไม่ต่อเนื่องและจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ เฉลี่ยจำนวน 10 รอบ พบว่าเทคนิคการแบ่งข้อมูลด้วยขนาดความถี่เท่ากัน จำนวน 10 ชั้น อัตราการจำแนกข้อมูลถูกต้องสูงที่สุดเท่ากับ 99.79% รองลงมาคือเทคนิคการแบ่งข้อมูลด้วยการจัดกลุ่มเคมีนส์ จำนวน 40 กลุ่ม เท่ากับ 99.75% และน้อยที่สุดคือเทคนิคการแบ่งข้อมูลด้วยขนาดความกว้างเท่ากัน จำนวน 20 ชั้น เท่ากับ 99.57%

2. ผลการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลของเทคนิคต้นไม้ตัดสินใจกับชุดข้อมูลทดสอบ

ผลการจำแนกข้อมูลของเทคนิคต้นไม้ตัดสินใจกับชุดข้อมูลฝึกสอน ทำให้ทราบว่าเทคนิคการแบ่งข้อมูลแบบไม่มีผู้สอนแต่ละแบบ ควรเลือกจำนวนชั้นเท่าใดจึงทำให้ได้ประสิทธิภาพสูงสุด ผู้วิจัยจึงนำเอากฎที่สร้างขึ้นนั้นไปใช้กับชุดข้อมูลทดสอบ เพื่อเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลด้วยต้นไม้ตัดสินใจ โดยวัดประสิทธิภาพอัตราการจำแนกที่ถูกต้อง อัตราการจำแนกที่ผิดพลาด และข้อมูลที่ไม่สามารถจำแนกได้ ซึ่งผลการจำแนกข้อมูลแสดงได้ดังตารางที่ 3

ตารางที่ 3 ผลการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลของเทคนิคต้นไม้ตัดสินใจกับชุดข้อมูลทดสอบ

เทคนิค	ชั้น/กลุ่ม	เปอร์เซ็น		
		อัตราการจำแนกข้อมูลที่ถูกต้อง	อัตราการจำแนกข้อมูลที่ผิดพลาด	ข้อมูลที่ไม่สามารถจัดเข้ากลุ่มได้
Equal Width and ID3	20	92.30	4.89	2.81
Equal Frequency and ID3	10	91.23	7.85	0.92
k-means and ID3	40	90.92	5.61	3.48

จากตารางที่ 3 พบว่าประสิทธิภาพของการใช้เทคนิคการแบ่งข้อมูลแบบขนาดความกว้างเท่ากันและการจำแนกข้อมูลด้วยต้นไม้ตัดสินใจ มีประสิทธิภาพสูงที่สุด โดยมีอัตราการจำแนกถูกต้องเท่ากับ 92.30% มีอัตราการจำแนกผิดพลาด เท่ากับ 4.89% และข้อมูลที่ไม่สามารถจัดเข้ากลุ่มได้ เท่ากับ 2.81%

สรุปและอภิปรายผลการวิจัย

จากผลการทดลองแสดงให้เห็นว่าสามารถใช้ต้นไม้ตัดสินใจ รุ่นที่ 3 ในการจำแนกข้อมูลชนิดต่อเนื่องได้ โดยมีการใช้เทคนิคการแบ่งข้อมูลให้อยู่ในรูปแบบไม่ต่อเนื่องเข้ามาช่วย ซึ่งในการทดลองครั้งนี้ใช้เทคนิคการแบ่งข้อมูลแบบขนาดความกว้างเท่ากัน เทคนิคการแบ่งข้อมูลที่ขนาดความถี่เท่ากัน และเทคนิคการจัดกลุ่มข้อมูลด้วยเทคนิคเคมีนส์ เมื่อนำชุดข้อมูลฝึกสอนผ่านกระบวนการแบ่งข้อมูลและจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ โดยมีการเปลี่ยนค่าจำนวนชั้นหรือกลุ่ม เพื่อหาค่าที่เหมาะสมในการใช้งานและใช้วิธีการตรวจสอบไขว้แบบ 10-fold Cross Validation ผลการทดลองที่ได้จากภาพที่ 5 พบว่าเทคนิคการแบ่งข้อมูลด้วยขนาดความกว้างเท่ากันแบ่งได้ จำนวน 20 ชั้น เทคนิคการแบ่งข้อมูลด้วยขนาดความถี่เท่ากัน

แบ่งได้ จำนวน 10 ชั้น และเทคนิคการแบ่งข้อมูลด้วยการจัดกลุ่มเคมีนส์ จำนวน 40 กลุ่ม มีความเหมาะสมมากที่สุดโดยพิจารณาจากประสิทธิภาพการจำแนกข้อมูลได้ถูกต้องสูงที่สุด

อนึ่งเมื่อนำกฎที่สร้างขึ้นจากต้นไม้ตัดสินใจที่ใช้จำแนกข้อมูลกับชุดข้อมูลฝึกสอนและมีประสิทธิภาพในการจำแนกข้อมูลที่ถูกต้องสูงสุดไปใช้กับชุดข้อมูลทดสอบ พบว่าเทคนิคการแบ่งข้อมูลด้วยขนาดความกว้างเท่ากัน จำนวน 20 ชั้น มีประสิทธิภาพสูงสุด เท่ากัน 92.30% เมื่อพิจารณาจากอัตราการจำแนกข้อมูลที่ถูกต้อง รองลงมาคือเทคนิคการแบ่งข้อมูลด้วยขนาดความถี่เท่ากัน จำนวน 10 ชั้น เท่ากับ 91.23% และน้อยที่สุดคือการแบ่งข้อมูลด้วยการจัดกลุ่มเคมีนส์ จำนวน 40 กลุ่ม เท่ากับ 90.92% ทั้งนี้พิจารณาโดยละเอียดพบว่าการใช้เทคนิคการแบ่งข้อมูลด้วยขนาดความถี่เท่ากันจะมีประสิทธิภาพสูงสุดในชุดข้อมูลฝึกสอน แต่เมื่อนำมาใช้กับข้อมูลทดสอบแล้วมีประสิทธิภาพไม่ดี ทั้งนี้เนื่องจากเทคนิคการแบ่งข้อมูลแบบไม่มีผู้สอนแต่ละประเภท จะไม่มีการใช้งานส่วนคำตอบของแต่ละระเบียนมาพิจารณา และจะเหมาะสมกับข้อมูลที่มีขอบเขตจำกัดเท่านั้น จึงส่งผลทำให้เมื่อนำกฎที่สร้างขึ้นจากชุดข้อมูลฝึกสอนมาใช้กับชุดข้อมูลทดสอบแล้วได้ประสิทธิภาพไม่ดีดังเดิม ทั้งนี้เนื่องเทคนิคการแบ่งข้อมูลแบบไม่มีผู้สอนนั้นกระบวนการแบ่งข้อมูลจะใช้ข้อมูลภายในขอบเขตจำกัดที่มีเท่านั้น ดังนั้นการเลือกใช้เทคนิคการแบ่งข้อมูลแบบไม่มีผู้สอนควรจะมีการทดสอบกับทั้งชุดข้อมูลฝึกสอนและชุดข้อมูลทดสอบ เพื่อได้เทคนิคการแบ่งข้อมูลที่เหมาะสมที่สุด ซึ่งสอดคล้องกับผลงานวิจัยของ (Rajashree, Rajib & Rasmita, 2011) ที่กล่าวว่าการเลือกใช้เทคนิคการแบ่งข้อมูลนั้นแต่ละเทคนิคจะมีข้อดีและข้อด้อยแตกต่างกันในการใช้งานควรมีการทดสอบเพื่อเลือกใช้ได้อย่างถูกต้องเหมาะสม การใช้เทคนิคการแบ่งข้อมูลร่วมกับเทคนิคต้นไม้ตัดสินใจ รุ่นที่ 3 สามารถช่วยกรองข้อมูลที่ไม่อยู่ในกฎที่สร้างขึ้นจากชุดข้อมูลฝึกสอนได้ ซึ่งเป็นสิ่งที่แตกต่างกับเทคนิคต้นไม้ตัดสินใจ รุ่นที่ C4.5 ฉะนั้นวิธีที่นำเสนอหากเลือกนำไปใช้ได้อย่างเหมาะสมกับข้อมูลที่มีขอบเขตจำกัดจะเกิดประโยชน์อย่างมาก เช่น การกรองข้อมูลเพื่อหาข้อมูลที่ไม่อยู่ในกฎของข้อมูลแบบต่อเนื่อง

ข้อเสนอแนะ

งานวิจัยที่นำเสนอนี้เป็นการใช้เทคนิคการแบ่งข้อมูลแบบไม่มีผู้สอนและการจำแนกข้อมูลด้วยต้นไม้ตัดสินใจ รุ่นที่ 3 ทำในแบบออฟไลน์และเปรียบเทียบเฉพาะเทคนิคการแบ่งข้อมูลแบบไม่มีผู้สอนเท่านั้น แนวทางการวิจัยในอนาคต ได้แก่

1) ทำการทดลองหาประสิทธิภาพของการใช้เทคนิคการแบ่งข้อมูลแบบมีผู้สอน (Supervised Learning) และจำแนกข้อมูลด้วยเทคนิคต้นไม้ตันสินใจ รุ่นที่ 3

2) ทำการทดลองเปรียบเทียบประสิทธิภาพเทคนิคการแบ่งข้อมูลทั้งแบบไม่มีผู้สอนและแบบมีผู้สอน

References

Chinnapat, K. (2010). *Data Classification*. Retrieved from http://scriptslines.com/blog (in Thai)

Eakasit, P. (2014). Introduction to Data Mining Techniques. Bangkok : Asia Digital Printing. (in Thai)

Han & Kamber. (2006). *Data Mining: Concepts and Techniques*. MA : Morgan Kaufmann.

Hemada, B. & Vijaya, L. (2013). A Study On Discretization Technique. *International Journal of Engineering Research & Technology*, 2 (8), 1887-1892.

Kamonwan, K. (2014). Brand Building Strategies for e-Commerce Business Website. *SDU Research Journal Humanities and Social Sciences*, 10 (1), 77-96. (in Thai)

Kandeeban, Selvakani & Rajesh. (2009). Integrated Intrusion Detection Using SCT. *International Journal of Computer and Network Security*, 1, 128-133.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 2 (12), 1137–1143.

Pairachnop, V. & Doungkamol, P. (2014). Google Apps for Education an Educational Innovation in Digital Age. *SDU Research Journal Sciences and Technology*, 7 (3), 103-111. (in Thai)

Palaniappan, S. & Hong, Tan Kim., (2009). Discretization of Continuous Valued dimensions in OLAP Data Cubes. *International Journal of Computer Science and Network Security*, 8 (11), 116-126.

Patchaya, B. (2010). Decision Support System Using Tree Techniques. *Conference on Computer Science A Graduate of the year 2553 19 September 2010 (pp.9_1-9_5)*, Chiang Mai: Chiang Mai University. (in Thai)

Prashant, G. A., Sanket, K., Kunal, K., Hemant, K. & Abhilasha, B. (2015). Implementation of Improved ID3 Algorithm to Obtain more Optimal Decision Tree. *International Journal of Engineering Research and Development*, 11 (2), 44-47

Rajashree, D., Paramguru, Rajib, L. D. & Rasmita, D. (2011). Comparative Analysis of Supervised and Unsupervised Discretization Techniques. *International Journal of Advances in Science and Technology*, 2 (3), 29-37.

Sawit, C., Sunthorn, S. & Wacharakorn, N., (2013). The Development of the Automatic Scanner. *SUD Research Journal Science and Technology*, 6 (2), 39-47. (in Thai)

Sirapat, C. (2009). Unsupervised Learning. In Artificial Neural Networks, Department of Computer Science Faculty of Science , Khon Kaen University. (in Thai)

Supawee, M. (2011). *Invariant Range Image Multi - Pose Face Recognition Using Fuzzy Ant Algorithm Center of Gravity Search And Membership Matching Score*. Dissertation in Information Technology, Department of Faculty of Information Technology, Rangsit University, Thailand. (in Thai)

Surasit, S., Witcha, C., Kiattsak, M. & Parinya., (2012). Classification via k-Means Clustering and Distance-Based Outlier Detection. *Proceedings of the Tenth International Conference on ICT and Knowledge Engineering 21-23 November 2012 (pp.125-128)*, Bangkok: Siam University.

Surasit, S., Witcha, C., Kiattsak, M. & Parinya., (2012). Implementation of Fuzzy *c*-Means and Outlier Detection for Intrusion with KDD Cup 1999 Data Set. *International Journal of Engineering Research and Development*, 2 (2), 44-48.

Watthananon, J. & Mingkhwan, A. (2012). Comparative Efficiency of Correlation Plot Data Classification. *Journal of KMUTNB*, 22 (1), 77-89.

Witcha, C. (2008). *Hybrid Fuzzy Techniques of Unsupervised Intrusion Detection System*. Dissertation in Computer Science, Department of Faculty of Computer Science and Information Systems, University of Technology, Malaysia.

ผู้เขียน

อาจารย์สุระสิทธิ์ ทรงม้า

อาจารย์ประจำคณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสวนดุสิต

295 ถนนราชสีมา เขตดุสิต กรุงเทพมหานคร 10300

email : surasit.songma@gmail.com