Research Article

Received: August 8, 2018; Accepted: October 1, 2018

การเปรียบเทียบประสิทธิภาพในการทำนายผลการจำแนกเมล็ดพันธุ์ข้าวเปลือกขาวดอกมะลิ 105 ด้วยเทคนิคการทำเหมืองข้อมูล

An Efficiency Comparison in Prediction of Khao Dok Mali 105 Paddy Rice Classification with Data Mining Techniques

สายชล สินสมบูรณ์ทอง*

ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ถนนฉลองกรุง เขตลาดกระบัง กรุงเทพมหานคร 10520

Saichon Sinsomboonthong*

Department of Statistics, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang,

Chalongkrung Road, Ladkrabang, Bangkok 10520

บทคัดย่อ

การศึกษานี้เป็นการเปรียบเทียบประสิทธิภาพในการทำนายผลการจำแนกเมล็ดพันธุ์ข้าวเปลือกขาวดอกมะลิ 105 ด้วยเทคนิคการทำเหมืองข้อมูล วิธีการจำแนกที่นำมาเปรียบเทียบมี 7 วิธี ได้แก่ วิธีเพื่อนบ้านใกล้สุด k ตัว โดยใช้อัลกอริทึมชนิด IBk วิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึมชนิด J48 วิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึมชนิดเพอร์เซปตรอนแบบหลายชั้น วิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO ชนิดโพลิโนเมียลเคอร์เนล วิธีฐานกฎโดยใช้อัลกอริทึม decision table วิธีการถดถอยลอจิสติกทวิภาค และวิธีนาอีฟเบส์ การเปรียบเทียบประสิทธิภาพของวิธีการจำแนกจะพิจารณาจากค่าความถูกต้อง ค่าความระลึก ค่าความถ่วงดุล และค่าคลาดเคลื่อนกำลังสองเฉลี่ย ผลการศึกษาพบว่าวิธีเพื่อนบ้านใกล้สุด k ตัว ที่ random seed = 10, 20 และ 30 มีค่าความถูกต้อง ค่าความระลึก ค่าความถ่วงดุล และค่าคลาดเคลื่อนกำลังสองเฉลี่ยดีที่สุด คือ ร้อยละ 100, 1.000, 1.000, 0.00002 ตามลำดับ ส่วนวิธีวิธีซัพพอร์ตเวกเตอร์แมชชีนและวิธีฐานกฎที่ random seed = 10 และ 20 มีค่าความระลึกดีที่สุด คือ 1.000 เนื่องจากวิธีเพื่อนบ้านใกล้สุด k ตัว มีประสิทธิภาพในการทำนายผลดีที่สุดถึง 4 ค่า ดังนั้นวิธีเพื่อนบ้านใกล้สุด k ตัว เป็นวิธีดีที่สุด

คำสำคัญ : วิธีเพื่อนบ้านใกล้สุด k ตัว; วิธีต้นไม้ตัดสินใจ; วิธีโครงข่ายประสาทเทียม; วิธีซัพพอร์ตเวกเตอร์แมชชีน; วิธีฐานกฎ; วิธีการถดถอยลอจิสติกทวิภาค; วิธีนาอีฟเบส์

Abstract

In this study, an efficiency comparison in prediction of Khao Dok Mali 105 paddy rice classification with data mining techniques was compared. The seven classification methods were

***ผู้รับผิดชอบบทความ :** saichon.ss49@gmail.com **doi:** 10.14456/tstj.2020.32

the followings: (1) k-nearest neighbor method using IBk algorithm; (2) decision tree method using J48 algorithm; (3) artificial neural network method using multilayer perceptron algorithm; (4) support vector machine method using polynomial kernel; (5) rule-based method using decision table algorithm; (6) binary logistic regression method; and (7) naïve Bayes method. The following efficiency comparisons of classification were employed: accuracy, recall, F-measure, and mean square error (MSE). The important results are as follows. The k-nearest neighbor method using random seed = 10, 20 and 30 showed the best accuracy, recall, F-measure, and MSE at 100 %, 1.000, 1.000 and 0.00002 respectively. The support vector machine method and rule-based using random seed = 10, and 20 exhibited the best recall at 1.000. Since the k-nearest neighbor method offered the best efficiencies for all the 4 values, it was considered the best prediction method.

Keywords: k-nearest neighbor; decision tree; artificial neural network; support vector machine; rule-based; binary logistic regression; naïve-Bayes

1. บทนำ

การทำเหมืองข้อมูลนับเป็นสาขาวิจัยทางการเกษตรสาขาหนึ่งที่เกิดขึ้นเมื่อไม่นานมานี้ ข้าวเป็นธัญพืชด้านอาหารที่มีการเจริญเติบโตอย่างรวดเร็วและมีความสำคัญต่อประชากรโลก ข้าวเป็นอาหารหลักของประชากรโลกมากกว่า 60 % การจำแนกเมล็ดข้าวจึงมีความสำคัญในการหาคุณค่าทางการตลาดของพันธุ์ข้าว การจำแนกกลุ่มพันธุ์ข้าวเป็นสิ่งจำเป็นสำหรับนักผสมพันธุ์พืชเพื่อทำนายผลผลิตและคุณภาพ [1] ข้าวเป็นธัญพืชที่คนไทยรับประทานมากที่สุดและเป็นสินค้าส่งออกที่สำคัญของประเทศไทย คุณภาพเมล็ดพันธุ์ข้าวเป็นปัจจัยหลักที่ส่งผลต่อคุณภาพของข้าวสาร เมล็ดพันธุ์ข้าวมักมีการปนพันธุ์มาจากกระบวนการผลิต ซึ่งทำให้ผลผลิตข้าวสารต่ำลง [2] ในปี พ.ศ. 2558 ข้าวขาวดอกมะลิ 105 เป็นข้าวที่มีปริมาณการส่งออก 1,963.356 พันเมตริกตัน มีมูลค่าการส่งออกรวม 52,391.05 ล้านบาท ประเทศที่ไทยส่งออกไปมากที่สุด คือ ประเทศจีน มีมูลค่าส่งออก 26,429 ล้านบาท คิดเป็น 50 % ทำให้มีการปลอมปนข้าวพันธุ์อื่นลงไปในข้าวขาวดอกมะลิ 105 เช่น พันธุ์ปทุมธานี 1 พันธุ์ชัยนาท 1 เนื่องจากมีขนาด สี และรูปร่างใกล้เคียงกับข้าวขาวดอกมะลิ 105 [3]

การตรวจสอบการปลอมปนมีหลายวิธี ได้แก่ การตรวจสอบรูปร่างลักษณะภายนอก ได้แก่ น้ำหนักเมล็ด ความยาวเมล็ด ความกว้างเมล็ด ความหนาเมล็ด พื้นที่เมล็ด และเส้นรอบวงเมล็ด สามารถบอกว่าเมล็ดข้าวนั้นเป็นข้าวสาลีหรือไม่ โดยดูจากระยะแกนเมล็ด ระยะแนวขวาง ระยะรอบเมล็ด แล้วมาหาค่า aspect ratio กับ form factor การวิเคราะห์ข้อมูลโดยนำข้อมูลของเมล็ดข้าวสาลีและใช้โครงข่ายประสาทเทียมมาสร้างตัวจำแนก ซึ่งการทดลองพบว่าวิธีนี้มีความแม่นยำ [4] ส่วนลักษณะสัณฐานและสีเมล็ดเป็นลักษณะเด่นเพื่อจำแนกหาความแตกต่างของแต่ละพันธุ์ด้วยการวิเคราะห์องค์ประกอบหลัก ในการวิเคราะห์ข้อมูลที่ใช้ในการทดสอบมีจำนวนน้อย เมล็ดข้าวที่ใช้มีขนาดที่แตกต่างกันมาก บางพันธุ์มีลักษณะเมล็ดอ้วนป้อม บางพันธุ์มีลักษณะเรียวและไม่สามารถแยกข้าวที่มีจำนวนมาก [5] นอกจากนี้มุมหัวข้าวและ

มุมหางข้าวเป็นลักษณะเด่นที่มีประสิทธิผลต่อตัวจำแนกแบบนาอีฟเบส์และซัพพอร์ตเวกเตอร์แมชชีน โดยมีความแม่นยำเฉลี่ย 83.01 และ 99.49 % ตามลำดับ ซึ่งมุมหัวข้าว มุมหาง และซัพพอร์ตเวกเตอร์แมชชีนเป็นองค์ประกอบในการจำแนกข้าวได้ดีมาก [2] สำหรับการตรวจสอบทางเคมีมีหลายวิธี วิธีที่ 1 คือ การละลายด้วยโพแทสเซียมไฮดรอกไซด์ที่ความเข้มข้น 1.7 % เป็นเวลา 23 ชั่วโมง พบว่าข้าวขาวดอกมะลิ 105 จะละลายทั้งหมด วิธีที่ 2 คือ การย้อมสี พบว่าข้าวขาวดอกมะลิ 105 จะติดสีน้อยมากเมื่อเปรียบเทียบกับข้าวขาวพันธุ์อื่น ๆ วิธีที่ 3 คือ การต้มแล้วบด โดยนำข้าวสารไปต้มด้วยเวลาที่ต่างกัน แล้วนำไปบดบนแผ่นกระจก พบว่าข้าวขาวดอกมะลิ 105 จะสุกหมด ไม่มีเนื้อแป้งหลงเหลืออยู่ ส่วนข้าวพันธุ์อื่น ๆ จะมีเนื้อแป้งหลงเหลืออยู่ [6] ส่วนวิธีที่ 4 คือ การใช้เทคโนโลยีดีเอ็นเอ ซึ่งมีความแม่นยำสูงและใช้ในปัจจุบัน [7] ในที่นี้จะเลือกใช้วิธีที่ 2 และ 3 คือ การย้อมสี และการต้มแล้วบด เนื่องจากทำได้โดยสะดวกและใช้เวลาอันรวดเร็ว

ผู้วิจัยมีความสนใจในการเปรียบเทียบประสิทธิภาพในการทำนายผลการจำแนกเมล็ดพันธุ์ข้าวเปลือกขาวดอกมะลิ 105 ด้วยเทคนิคการทำเหมืองข้อมูล ซึ่งการเปรียบเทียบประสิทธิภาพในการทำนายผลการจำแนกเมล็ดพันธุ์ข้าวเปลือกขาวดอกมะลิ 105 ขึ้นอยู่กับปัจจัยหลายอย่าง ได้แก่ น้ำหนักเมล็ด ความยาวเมล็ด ความกว้างเมล็ด พื้นที่เมล็ด เส้นรอบวงเมล็ด รูปร่างเมล็ด สีเปลือกเมล็ด สีปลายยอดเมล็ด ขนบนเปลือกเมล็ด การมีหาง ชื่อพันธุ์ข้าว เป็นต้น โดยผู้วิจัยสนใจศึกษาการเปรียบเทียบประสิทธิภาพในการทำนายผลการจำแนกเมล็ดพันธุ์ข้าวเปลือกขาวดอกมะลิ 105 ด้วยเทคนิคการทำเหมืองข้อมูล 7 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว (k-nearest neighbor) วิธีต้นไม้ตัดสินใจ (decision tree) วิธีโครงข่ายประสาทเทียม (artificial neural network) วิธีซัพพอร์ตเวกเตอร์แมชชีน (support vector machine) วิธีฐานกฎ (rules-based) วิธีการถดถอยลอจิสติกทวิภาค (binary logistic regression) และวิธีนาอีฟเบส์ (naïve Bayes) เพื่อหาวิธีที่มีประสิทธิภาพในการทำนายผลการจำแนกเมล็ดพันธุ์ข้าวเปลือกขาวดอกมะลิ 105 ที่เหมาะสมต่อไป

2. วิธีการวิจัย

2.1 เครื่องมือที่ใช้ในการวิจัย

การวิจัยนี้ใช้โปรแกรม WEKA (Waikato Environment for Knowledge Analysis) เวอร์ชั่น 3.9 เป็นโปรแกรมที่สามารถดาวน์โหลดจากเว็บไซต์ ซึ่งอยู่ภายใต้การควบคุมของ GPL License โดยโปรแกรม WEKA นี้พัฒนามาจากภาษาจาวาทั้งหมด นิยมใช้งานในด้านการทำเหมืองข้อมูล

2.2 การเก็บรวบรวมข้อมูล การบันทึกข้อมูล และการแบ่งข้อมูล

2.2.1 การเก็บรวบรวมข้อมูล โดยเก็บรวบรวมข้อมูลจากการบันทึกน้ำหนักเมล็ด ความยาวเมล็ด ความกว้างเมล็ด พื้นที่เมล็ด เส้นรอบวงเมล็ด รูปร่างเมล็ด สีเปลือกเมล็ด สีปลายยอดเมล็ด ขนบนเปลือกเมล็ด การมีหาง และชื่อพันธุ์ข้าว ตั้งแต่เดือนมกราคมถึงเดือนมีนาคม พ.ศ. 2561 จำนวน 1,000 เมล็ด โดยข้อมูลประกอบด้วยคุณลักษณะ (attribute) ต่าง ๆ ดังนี้ (1) น้ำหนักเมล็ด (หน่วย : มิลลิกรัม) (2) ความยาวเมล็ด (หน่วย : มิลลิเมตร) (3) ความกว้างเมล็ด (หน่วย : มิลลิเมตร) (4) พื้นที่เมล็ด (area = πr^2 หน่วย : ตารางมิลลิเมตร) (5) เส้นรอบวงเมล็ด (circumference = $2\pi r$ หน่วย : มิลลิเมตร) (6) รูปร่างเมล็ด (แบนวงรี / ป้อมวงรี) (7) สีเปลือกเมล็ด (น้ำตาลอ่อน / น้ำตาลเข้ม) (8) สีปลายยอดเมล็ด (น้ำตาลอ่อน / น้ำตาลเข้ม) (9) ขนบนเปลือกเมล็ด (ไม่

มีขน / ขนสั้น / ขนยาว) (10) การมีหาง (ไม่มีหาง / หางสั้น / หางยาว) (11) ชื่อพันธุ์ข้าว (ข้าวเปลือกขาวดอกมะลิ 105 / ข้าวเปลือกขาวพันธุ์อื่น ๆ เช่น พันธุ์ปทุมธานี 1 พันธุ์ชัยนาท 1)

2.2.2 การบันทึกข้อมูล โดยบันทึกข้อมูลผลการตรวจสอบการจำแนกเมล็ดพันธุ์ข้าวเปลือกขาวดอกมะลิจำนวน 1,000 เมล็ด ลงในโปรแกรม Microsoft Excel ให้แนวคอลัมน์เป็นน้ำหนักเมล็ด ความยาวเมล็ด ความกว้างเมล็ด พื้นที่เมล็ด เส้นรอบวงเมล็ด รูปร่างเมล็ด สีเปลือกเมล็ด สีปลายยอดเมล็ด ขนบนเปลือกเมล็ด การมีหาง และชื่อพันธุ์ข้าว ส่วนแนวแถวเป็นลำดับที่ของเมล็ดข้าวที่ทำการบันทึกผลลำดับที่ 1-1,000

2.2.3 การแบ่งข้อมูล โดยสุ่มด้วยโปรแกรม WEKA จำนวน 3 รอบ ซึ่งกำหนดตัวเลขสุ่มเทียม (random seed) เป็น 10, 20 และ 30 แล้วนำข้อมูลทั้งหมดมาแบ่งเป็น 2 ส่วน การตรวจสอบการจำแนกเมล็ดพันธุ์ข้าวเปลือกขาวดอกมะลิ 105 จำนวน 1,000 เมล็ด เพื่อใช้สำหรับการวิเคราะห์ข้อมูล โดยแบ่งข้อมูลออกเป็นสัดส่วนดังนี้ [8] (1) ส่วนที่ 1 ข้อมูลชุดฝึกหัด (training data set) เพื่อนำไปสร้างตัวแบบ มีข้อมูลร้อยละ 80 ของข้อมูลทั้งหมด ซึ่งจะได้ข้อมูลในส่วนที่ 1 จำนวน 800 เมล็ด และส่วนที่ 2 ข้อมูลชุดทดสอบ (testing data set) เพื่อนำไปทำนายตัวแบบ มีข้อมูลร้อยละ 20 ของข้อมูลทั้งหมด ซึ่งจะได้ข้อมูลในส่วนที่ 2 จำนวน 200 เมล็ด

2.3 การวิเคราะห์ข้อมูล

การจัยครั้งนี้ได้สุ่มแบ่งข้อมูลเป็น 2 ส่วน โดยส่วนที่ 1 ข้อมูลชุดฝึกหัด ใช้ข้อมูลร้อยละ 80 ของข้อมูลทั้งหมดในการสร้างตัวแบบ ส่วนที่ 2 ข้อมูลชุดทดสอบ ใช้ข้อมูลร้อยละ 20 ของข้อมูลทั้งหมดในการทำนายตัวแบบ แล้วแปลงไฟล์ข้อมูลให้เป็นนามสกุล *.csv เพื่อใช้วิเคราะห์ประสิทธิภาพการจำแนกข้อมูลในโปรแกรม WEKA ซึ่งเป็นโปรแกรมที่สามารถนำมาทดสอบอัลกอริทึมของวิธีการจำแนกได้เนื่องจากมีอัลกอริทึมที่ได้ระบุไว้ให้เลือกใช้ในโปรแกรมครบตามที่กำหนด ผู้วิจัยได้กำหนดวิธีการจำแนกเพื่อนำมาทดสอบดังนี้

2.3.1 วิธีเพื่อนบ้านใกล้สุด k ตัว ใช้อัลกอริทึมชนิด IBk เนื่องจากเป็นฟังก์ชันหลักที่สนใจ ซึ่งเป็นพื้นฐานของอัลกอริทึม 8.1 อัลกอริทึม IBk ยังสามารถกำหนดน้ำหนักระยะห่างและทางเลือก (option) เพื่อกำหนดค่า k โดยใช้ cross-validation [9]

2.3.2 วิธีต้นไม้ตัดสินใจ ใช้อัลกอริทึมชนิด J48 ซึ่งพัฒนาจาก ID3 สามารถใช้กับข้อมูลแบบไม่ต่อเนื่องและแบบต่อเนื่อง ต่างจาก ID3 ที่ใช้ได้เพียงข้อมูลแบบไม่ต่อเนื่องเท่านั้น [10]

2.3.3 วิธีโครงข่ายประสาทเทียม ใช้อัลกอริทึมชนิดเพอร์เซปตรอนหลายชั้น (multilayer perceptron) โดยกำหนดค่าอัตราการเรียนรู้ (learning rate) เป็น 0.1 ค่าโมเมนตัม (momentum) เป็น 0.9 จำนวนรอบการสอน (training time) 20,000 รอบ การวิจัยครั้งนี้ใช้อัลกอริทึมของวิธีโครงข่ายประสาทเทียมชนิดเพอร์เซปตรอนหลายชั้นที่มีชั้นซ่อน (hidden layer) 1 ชั้น แม้ว่าโครงสร้างโครงข่ายประสาทเทียมที่ซับซ้อนสามารถมีชั้นซ่อนมากกว่า 1 ชั้น แต่ทางปฏิบัติการกำหนดชั้นซ่อน 1 ชั้น ก็เพียงพอต่อการวิเคราะห์ข้อมูล [11]

2.3.4 วิธีซัพพอร์ตเวกเตอร์แมชชีน ใช้อัลกอริทึม SMO ชนิดโพลิโนเมียลเคอร์เนล (polynomial kernel) อ้างอิงจากงานวิจัยของ วาทินี และคณะ [12] ที่พบว่าวิธีซัพพอร์ตเวกเตอร์แมชชีนที่ใช้อัลกอริทึมชนิดโพลิโนเมียลเคอร์เนลดีที่สุด

2.3.5 วิธีฐานกฎ ใช้ชุดลำดับของกฎมาสร้างรูปแบบการแยกประเภทข้อมูล โดยส่วนใหญ่แล้ว

จะใช้กฎที่เป็น If ... then ซึ่งเป็นกฎอย่างง่าย ใช้อัลกอริทึม decision table เป็นเครื่องมือที่ใช้แสดงเงื่อนไขการตัดสินใจและเลือกการทำงานหรือกระทำกิจกรรมภายใต้เหตุการณ์ของเงื่อนไขที่ระบุ วิธีการตัดสินใจแบบ decision table จะเป็นตาราง 2 มิติ [13]

2.3.6 วิธีการถดถอยลอจิสติกทวิภาค เป็นการวิเคราะห์การถดถอยแบบหนึ่งโดยที่ตัวแปรตามเป็นตัวแปรเชิงคุณภาพ ส่วนตัวแปรอิสระอาจเป็นตัวแปรเชิงปริมาณหรือเชิงคุณภาพ หรืออาจมีทั้งตัวแปรเชิงปริมาณและตัวแปรเชิงคุณภาพก็ได้ [14]

2.3.7 วิธีนาอีฟเบส์ คือ อัลกอริทึมที่ใช้หลักการของความน่าจะเป็นในการคัดกรองแต่ละคำตอบ (class) โดยมี 2 คำตอบ [15]

การนำผลการวิเคราะห์ข้อมูลมาประเมินผลเพื่อเปรียบเทียบประสิทธิภาพในการทำนายผลการจำแนกเมล็ดพันธุ์ข้าวเปลือกขาวดอกมะลิ 105 ด้วยวิธีการจำแนก 7 วิธี ว่าวิธีใดมีค่าความถูกต้อง (accuracy) ค่าความระลึก (recall) ค่าความถ่วงดุล (F-measure) สูงที่สุด และมีค่าคลาดเคลื่อนกำลังสองเฉลี่ย (mean square error, MSE) ต่ำที่สุด ซึ่งจะเป็นวิธีที่มีประสิทธิภาพในการทำนายผลดีที่สุด โดยที่ค่าคลาดเคลื่อนกำลังสองเฉลี่ยเป็นมาตรวัดการประเมินค่าได้ดี เนื่องจากค่าคลาดเคลื่อนกำลังสองเฉลี่ยประกอบด้วยทั้งความแปรปรวนและความเอนเอียง [16] ดังตารางที่ 1

Table 1 Confusion matrix

Class of Real Value	Class of Predicted Value	
	A	B
A	TP	FN
B	FP	TN

ตารางที่ 1 แสดงเมทริกซ์ความสับสน ซึ่งใช้เป็นค่าพื้นฐานในการวัดประสิทธิภาพในการทำนาย โดยที่ true positive (TP) คือ จำนวนข้อมูลที่จำแนกถูกว่าเป็นชั้น A; true negative (TN) คือ จำนวนข้อมูลที่จำแนกถูกว่าเป็นชั้น B; false positive (FP) คือ จำนวนข้อมูลที่จำแนกผิดว่าเป็นชั้น A แต่ชั้นที่แท้จริงเป็นชั้น B; false negative (FN) คือ จำนวนข้อมูลที่จำแนกผิดว่าเป็นชั้น B แต่ชั้นที่แท้จริงเป็นชั้น A

โดยมีสูตรในการคำนวณค่าต่าง ๆ คือ (1) ค่าความถูกต้อง = จำนวนข้อมูลที่จำแนกถูกว่าเป็นชั้น A และ B ÷ จำนวนข้อมูลทั้งหมด = (TP + TN) ÷ (TP + TN + FP + FN) (2) ค่าความระลึก = TP rate = TP ÷ (TP + FN) (3) ค่าความถ่วงดุล = [2 (ค่าความแม่นยำ) (ค่าความระลึก)] ÷ (ค่าความแม่นยำ + ค่าความระลึก) โดยที่ค่าความแม่นยำ (precision) = (จำนวนข้อมูลที่จำแนกถูกว่าเป็นชั้น A) ÷ (จำนวนข้อมูลที่ทำนายได้ในชั้น A) = TP ÷ (TP + FN) และ (4) ค่าคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) = $\sum_{i=1}^{n}\frac{(y_i-\hat{y}_i)^2}{n} = \sum_{i=1}^{n}\frac{e_i^2}{n}$ โดยที่ y_i แทนค่าที่แท้จริง และ $\hat{y}_i$ แทนค่าที่ทำนายได้

3. ผลการวิจัย

3.1 ผลการเปรียบเทียบประสิทธิภาพในการทำนายผลของวิธีการจำแนก

การเปรียบเทียบประสิทธิภาพในการทำนายผลการจำแนกเมล็ดพันธุ์ข้าวเปลือกขาวดอกมะลิ 105 ด้วยเทคนิคการทำเหมืองข้อมูล ระหว่างวิธีเพื่อนบ้านใกล้สุด k ตัว วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีฐานกฎ วิธีการถดถอยลอจิสติกทวิภาค และวิธีนาอีฟเบส์ โดยพิจารณาจากค่าความถูกต้อง ค่าความ

ถ่วงดุล และค่าคลาดเคลื่อนกำลังสองเฉลี่ย ได้ผลดังตารางที่ 2, 3 และรูปที่ 1, 2, 3, 4

ตารางที่ 2 พบว่าวิธีเพื่อนบ้านใกล้สุด k ตัว มีค่าความถูกต้องสูงสุดร้อยละ 100, 99.875 และ 100 ที่ random seed = 10, 20 และ 30 มีค่าความระลึกสูงสุด = 1.000 ที่ random seed = 10, 20 และ 30 มีค่าความถ่วงดุลสูงสุด 1.000, 0.999 และ 1.000 ที่ random seed = 10, 20 และ 30 และมีค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด = 0.0000008, 0.00009 และ 0.0000008 ที่ random seed = 10, 20 และ 30

Table 2 Accuracy, recall, F-measure and mean square error in prediction of Khao Dok Mali 105 paddy rice classification for training data set

Classification	Accuracy (percentage)			Recall			F-measure			Mean square error		
	Random seed			Random seed			Random seed			Random seed		
	10	20	30	10	20	30	10	20	30	10	20	30
k-Nearest neighbor	***100***	***99.875***	***100***	***1.000***	***1.000***	***1.000***	***1.000***	***0.999***	***1.000***	***0.0000008***	***0.00009***	***0.0000008***
Decision tree	90.5	89.75	91.25	0.946	0.978	0.971	0.928	0.924	0.934	0.07426	0.08655	0.06991
Artificial neural network	87.125	86.125	84.75	0.963	0.982	0.937	0.906	0.901	0.887	0.11096	0.11965	0.12867
Support vector machine	84.375	84.25	83	0.938	0.949	0.957	0.886	0.885	0.878	0.15626	0.15753	0.16999
Rule based	85.25	85.125	83	0.975	0.973	0.957	0.895	0.893	0.878	0.11142	0.11676	0.13410
Binary logistic regression	85.75	85	84	0.965	0.969	0.965	0.898	0.892	0.885	0.10465	0.11236	0.11834
Naïve Bayes	81.75	81	79.5	0.900	0.885	0.886	0.865	0.857	0.847	0.13506	0.14341	0.14799

Table 3 Accuracy, recall, F-measure and mean square error in prediction of Khao Dok Mali 105 paddy rice classification for testing data set

Classification	Accuracy (percentage)			Recall			F-Measure			Mean square error		
	Random seed			Random seed			Random seed			Random seed		
	10	20	30	10	20	30	10	20	30	10	20	30
k-Nearest neighbor	***100***	***100***	***100***	***1.000***	***1.000***	***1.000***	***1.000***	***1.000***	***1.000***	***0.00002***	***0.00002***	***0.00002***
Decision tree	87.5	88	90	0.970	0.978	0.965	0.911	0.916	0.916	0.10349	0.09449	0.08020
Artificial neural network	90	89	82	0.985	0.978	0.991	0.929	0.923	0.862	0.08934	0.09784	0.13861
Support vector machine	85	84	79	***1.000***	***1.000***	0.929	0.898	0.893	0.833	0.15000	0.16000	0.21004
Rule based	85	84	79	***1.000***	***1.000***	0.929	0.898	0.893	0.833	0.12236	0.12931	0.15571
Binary logistic regression	87	86	81	0.962	0.970	0.920	0.907	0.903	0.846	0.09685	0.10278	0.13359
Naïve Bayes	87.5	81	77.5	0.962	0.873	0.876	0.910	0.860	0.815	0.11196	0.14608	0.15280

ตารางที่ 3 พบว่าวิธีเพื่อนบ้านใกล้สุด k ตัว มีค่าความถูกต้อง ค่าความระลึกและค่าความถ่วงดุลสูงสุดร้อยละ 100, 1.000 และ 1.000 ตามลำดับ ที่ random seed = 10, 20 และ 30 นอกจากนี้วิธีเพื่อนบ้านใกล้สุด k ตัว ยังมีค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด = 0.00002 ที่ random seed = 10, 20 และ 30 ส่วนวิธีซัพพอร์ตเวกเตอร์แมชชีนและวิธีฐานกฎมีค่าความระลึกสูงสุด = 1.000 ที่ random seed = 10 และ 20

การนำค่าในตารางที่ 3 ข้อมูลชุดทดสอบมา

เขียนกราฟเพื่อแสดงถึงการเปรียบเทียบที่ชัดเจนยิ่งขึ้น โดยแบ่งแยกตามค่าความถูกต้อง ค่าความระลึก ค่าความถ่วงดุล และค่าคลาดเคลื่อนกำลังสองเฉลี่ย ดังแสดงในรูปที่ 1, 2, 3, 4

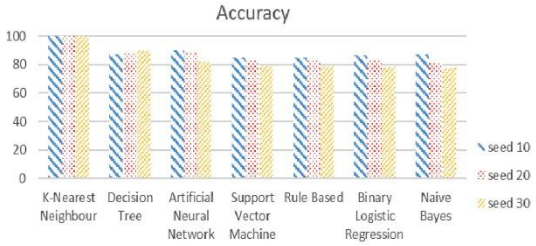

Figure 1 Comparison of efficiency in prediction of Khao Dok Mali 105 paddy rice classification with seven data mining techniques for testing data set by accuracy

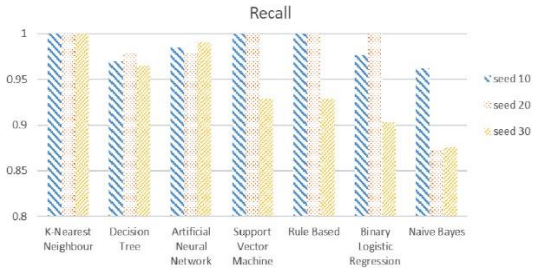

Figure 2 Comparison of efficiency in prediction of Khao Dok Mali 105 paddy rice classification with seven data mining techniques for testing data set by recall

รูปที่ 1 พบว่าวิธีเพื่อนบ้านใกล้สุด k ตัว มีค่าความถูกต้องสูงสุดร้อยละ 100 ที่ random seed = 10, 20 และ 30

รูปที่ 2 พบว่าวิธีเพื่อนบ้านใกล้สุด k ตัว มีค่าความระลึกสูงสุด = 1.000 ที่ random seed = 10, 20 และ 30 ส่วนวิธีซัพพอร์ตเวกเตอร์แมชชีนและวิธีฐานกฎมีค่าความระลึกสูงสุด = 1.000 ที่ random seed = 10 และ 20

รูปที่ 3 พบว่าวิธีเพื่อนบ้านใกล้สุด k ตัว มีค่าความถ่วงดุลสูงสุด = 1.0000 ที่ random seed = 10, 20 และ 30

รูปที่ 4 พบว่าวิธีเพื่อนบ้านใกล้สุด k ตัว มีค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด = 0.00002 ที่ random seed = 10, 20 และ 30

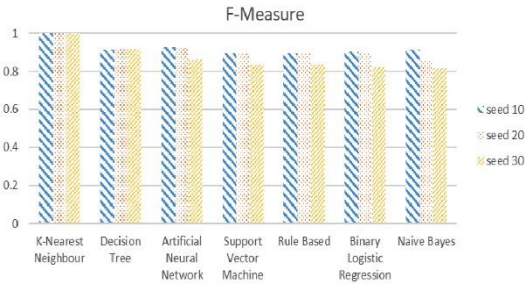

Figure 3 Comparison of efficiency in prediction of Khao Dok Mali 105 paddy rice classification with seven data mining techniques for testing data set by F-measure

Figure 4 Comparison of efficiency in prediction of Khao Dok Mali 105 paddy rice classification with seven data mining techniques for testing data set by mean square error

4. สรุปผลการวิจัยและข้อเสนอแนะ

4.1 สรุปผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพในการทำนายผลการจำแนกเมล็ดพันธุ์ข้าวเปลือกขาวดอกมะลิ 105 ด้วยเทคนิคการทำเหมืองข้อมูล โดยใช้วิธีการจำแนก 7 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีฐานกฎ วิธีการถดถอยลอจิสติกทวิภาค และวิธีนาอีฟเบส์ ประสิทธิภาพของวิธีการจำแนกพิจารณาจากค่าความถูกต้อง ค่าความระลึก ค่าความถ่วงดุลที่มีค่าที่สูงที่สุด และพิจารณาจากค่าคลาดเคลื่อนกำลังสองเฉลี่ยที่มีค่าที่ต่ำที่สุด โดยใช้หลักการของการทำเหมืองข้อมูลมาใช้ในการจำแนกข้อมูล ผลการวิจัยพบว่าวิธีเพื่อนบ้านใกล้สุด k ตัว มีค่าความถูกต้อง ค่าความระลึก ค่าความถ่วงดุล และค่าคลาดเคลื่อนกำลังสองเฉลี่ยดีที่สุด ส่วนวิธีวิธีซัพพอร์ตเวกเตอร์แมชชีนและวิธีฐานกฎมีค่าความระลึกดีที่สุด เนื่องจากวิธีเพื่อนบ้านใกล้สุด k ตัว มีประสิทธิภาพในการทำนายผลดีที่สุดถึง 4 ค่า คือ ค่าความถูกต้อง ค่าความระลึก ค่าความถ่วงดุล และค่าคลาดเคลื่อนกำลังสองเฉลี่ย ดังนั้นวิธีเพื่อนบ้านใกล้สุด k ตัว เป็นวิธีดีที่สุด

4.2 ข้อเสนอแนะ

4.2.1 เพื่อให้ได้ข้อสรุปของผลการวิเคราะห์ข้อมูลที่มีความครอบคลุมมากขึ้น อาจใช้วิธีการจำแนกโดยใช้อัลกอริทึมประเภทอื่น ๆ เช่น วิธีเพื่อนบ้านใกล้สุด k ตัว อาจใช้อัลกอริทึม KStar และ LWL วิธีต้นไม้ตัดสินใจอาจใช้อัลกอริทึม decision stump, LMT, random forest, random tree และ REP tree วิธีโครงข่ายประสาทเทียมใช้อัลกอริทึมเพอร์เซปตรอนหลายชั้นอาจกำหนดอัตราการเรียนรู้เป็น 0.2 และค่าโมเมนตัมเป็น 0.8 วิธีซัพพอร์ตเวกเตอร์แมชชีนใช้อัลกอริทึม SMO และอาจใช้ฟังก์ชันเคอร์เนลแบบ normalized poly Kernel, Puk และ RBF Kernel วิธีฐานกฎอาจใช้อัลกอริทึม JRip, OneR, PART และ ZeroR วิธีนาอีพเบส์อาจใช้ Naïve Bayes updatable เป็นต้น

4.2.2 ตัวแปรที่นำมาใช้ในงานวิจัยนี้เป็นเพียงส่วนหนึ่งของการพิจารณาการจำแนกเมล็ดพันธุ์ข้าวเปลือกขาวดอกมะลิ 105 เท่านั้น เพื่อให้การทำนายมีประสิทธิภาพมากขึ้นควรเพิ่มตัวแปรที่เกี่ยวข้องอื่น ๆ อีก เช่น มุมหัวข้าว มุมหางข้าว การละลายด้วยโพแทสเซียมไฮดรอกไซด์ การย้อมสี การต้มแล้วบด และการใช้เทคโนโลยีดีเอ็นเอ

5. References

[1] Divya, K. and Sangeetha, Y., 2016, Paddy seeds categorizing based on morpho logical feature using data mining algorithms, Int. J. Res. Comp. Commun. Technol. 5(8): 25-33.

[2] Seekuka, J, 2014, Features for Classifying Rice Grains by Image Analysis, Master Thesis, Kasetsart University, Bangkok, 52 p. (in Thai)

[3] Ruangphayak, S., 2016, Checking the Purity and Quality of Rice Quickly by KASP SNPline Detection, Rice Science Center, Kasetsart University, Bangkok, 54 p. (in Thai)

[4] Aluru S., 2011, Morphology based feature extraction and recognition for enhanced wheat quality evaluation, Int. Conf. Contemp. Comp. 4: 41-50.

[5] Zhao-yan, C.F.L., Ying, Y. and Rao, X., 2005, Identification of rice seed varieties

using neural network, J. Zhejiang Univ. Sci. 6: 1095-1100.

[6] Kongseri, N., Bangvarg, J., Cheapun, K., Wongpiyachon, S., Sukvivat, V., Sawangjit, P. and Tangvisuttijit, S., 2004, Quality and Investigation of Thai Jasmine Rice, Academic Agriculture Department, Ministry of Agriculture and Cooperatives, 62 p. (in Thai)

[7] Vanavichit, A., Tragoonrung, S. and Toojinda, T., 2003, Biotechnology and Rice Varieties Improvement, Chap. in Science and Technology with Thai Rice, Thailand's National Science and Technology Development Agency, 85 p. (in Thai)

[8] Panichkul, P., 2005, Development Data Mining System by Decision Tree, Work System Development Project, Master Thesis, King Montkut's Institute of Technology Ladkrabang, Bangkok, 62 p. (in Thai)

[9] Wu, X. and Kumar, V., 2009, The Top Ten Algorithms in Data Mining, University of Minnesota Department of Computer Science and Engineering, CRC Press, Minneapolis, 215 p.

[10] Thammasombut, R., 2012, Decision Support System for Selection the Mobile Internet Package Using Decision Tree, Major of Business Computer, Faculty of Business Administration, Rajapruek College, Sakon Nakhon, 77 p. (in Thai)

[11] Berson, A. and Smith, S.J., 1997, Data Warehousing, Data Mining, and OLAP, McGraw-Hill, New York, 612 p.

[12] Nuipian, V., 2010, Comparison of efficiency and analysis of data classification using artificial neural network, support vector machine, Naïve Bayes and k-nearest neighbor, Natl. Conf. Comp. Inform. Technol. 5: 131-138. (in Thai)

[13] Murti, S. and Mahantappa, M., 2012, Using rule based classifiers for the predictive analysis of breast cancer recurrence, J. Inform. Eng. Appl. 2(2): 12-19.

[14] Vanichbuncha, K., 2009, Multivariate Analysis, Thammasan Co., Ltd., Bangkok, 589 p. (in Thai)

[15] Sinsomboonthong, S., 2017, Data Mining 1: Discovering Knowledge in Data, 2nd Ed., Chamchuree Products Co., Ltd., Bangkok, 512 p. (in Thai)

[16] Larose, D.T., 2005, Discovering Knowledge in Data: An Introduction to Data Mining, John Wiley and Sons, New Jersey, 222 p.