การจำแนกประเภทข้อมูลด้วยตัวจำแนกประเภทขั้นตอนวิธีเชิงพันธุกรรม ขนาดความยาวโครโมโซมเค

The data classification using Genetic Algorithm with k-length chromosome

ศิรพงศ์ โชคธนหิรัญย์[1,2,*] และสุพจน์ เฮงพระพรหม[1,2]

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาและพัฒนาตัวจำแนกประเภทข้อมูลฐานจีเอด้วยขั้นตอนวิธีเชิงพันธุกรรมขนาดความยาวโครโมโซมเค สำหรับการจำแนกประเภทข้อมูล 2 กลุ่ม และเปรียบเทียบประสิทธิภาพกับเทคนิคการจำแนกประเภทข้อมูลพื้นฐานทั่วไป โดยกระบวนการจะเริ่มด้วยการทำข้อมูลให้เป็นปกติด้วยวิธีคะแนนซี จากนั้นจะใช้ขั้นตอนวิธีเชิงพันธุกรรมเลือกคุณลักษณะที่สำคัญตามขนาดความยาวโครโมโซมที่กำหนด (ค่าเค) แล้วนำผลรวมของข้อมูลจากคุณลักษณะที่เลือกได้มาเปรียบเทียบกับ 0 ถ้ามีค่ามากกว่าจะจำแนกเป็นกลุ่มที่ 1 แต่ถ้าไม่ใช่จะจำแนกให้เป็นกลุ่ม 2 การทดสอบประสิทธิภาพของวิธีการที่นำเสนอได้ทำการทดสอบกับชุดข้อมูลเกณฑ์มาตรฐานจำนวน 9 ชุดข้อมูล แบ่งเป็นชุดข้อมูลที่มีจำนวนคุณลักษณะน้อย (ไม่เกิน 100 คุณลักษณะ) 3 ชุดข้อมูล จำนวนคุณลักษณะปานกลาง (100 - 1,000 คุณลักษณะ) 3 ชุดข้อมูล และจำนวนคุณลักษณะมาก (มากกว่า 1,000 คุณลักษณะ) 3 ชุดข้อมูล โดยใช้วิธีตรวจสอบแบบไขว้ 10 กลุ่ม และเปรียบเทียบประสิทธิภาพกับวิธีการจำแนกประเภทข้อมูลพื้นฐานทั่วไป จำนวน 3 วิธีการ ได้แก่ วิธีเพื่อนบ้านใกล้ที่สุดเค วิธีต้นไม้ตัดสินใจ และวิธีเบย์อย่างง่าย จากผลการทดลองพบว่า วิธีการฐานจีเอที่นำเสนอให้ค่าความถูกต้องในการจำแนกประเภทข้อมูลเทียบเท่ากับวิธีเพื่อนบ้านใกล้ที่สุดเค วิธีต้นไม้ตัดสินใจ และวิธีเบย์อย่างง่าย ซึ่งยืนยันวัตถุประสงค์ของการวิจัยในครั้งนี้

คำสำคัญ: การจำแนกประเภทข้อมูล, ตัวจำแนกประเภท, ขั้นตอนวิธีเชิงพันธุกรรม

Abstract

This research aims to study and develop the GA-based classifier using the Genetic Algorithm (GA) with k-length chromosome for 2-class data classification; and compare the accuracy with well-known basic methods. First of all, the data is normalized using the z-score technique. Then, the GA is used to select k informative features. Finally, the summation of data for the k features which have been selected is computed and compared with 0. If the summation is greater than 0, it is classified as class 1; otherwise, it is classified as class 2. Nine benchmark datasets are used to test the performance of the proposed method. The datasets are grouped as follows: three datasets with a small number of features (less than 100), three datasets with a medium number of features (100 – 1,000), and three datasets with a large number of features (more than 1,000). The 10-Fold Cross Validation technique is applied to compare the performance of the proposed GA-based method with the K-Nearest Neighbor (KNN), Decision Tree and Naïve Bayes. The experimental results show that the proposed GA-based method gives comparable performance to the KNN, Decision Tree and Naïve Bayes approaches in terms of accuracy and, therefore, confirms the validity of the purpose of this research.

Keywords: data classification, classifier, Genetic Algorithm

[1]สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม

[2]หน่วยวิจัยอัจฉริยภาพแห่งเครื่องจักร สถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏนครปฐม

[*]Corresponding author, E-mail: Peterpluger@gmail.com

บทนำ

ในปัจจุบันนักวิจัยจำนวนมากได้ให้ความสำคัญกับการแก้ปัญหาที่ต้องการวิธีการจำแนกประเภทข้อมูลให้มีความถูกต้องเพิ่มมากขึ้น โดยเทคนิคการจำแนกประเภทข้อมูลนั้นมีหลายวิธี นอกจากนี้ยังได้มีการนำเทคนิคการจำแนกประเภทข้อมูลแบบต่าง ๆ มาประยุกต์ใช้ร่วมกันกับเทคนิคอื่น ๆ ในศาสตร์ด้านการทำเหมืองข้อมูล สถิติ รวมถึงการเรียนรู้ของเครื่อง เพื่อให้ได้ตัวแบบในการจำแนกประเภทข้อมูลแบบใหม่ที่มีประสิทธิภาพเพิ่มมากขึ้นโดยขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดเค [1] เป็นเทคนิคหนึ่งที่ได้รับความนิยมในการใช้งานมากที่สุดสำหรับการนำมาประยุกต์ใช้ในการจำแนกประเภทข้อมูลสำหรับงานวิจัยต่าง ๆ

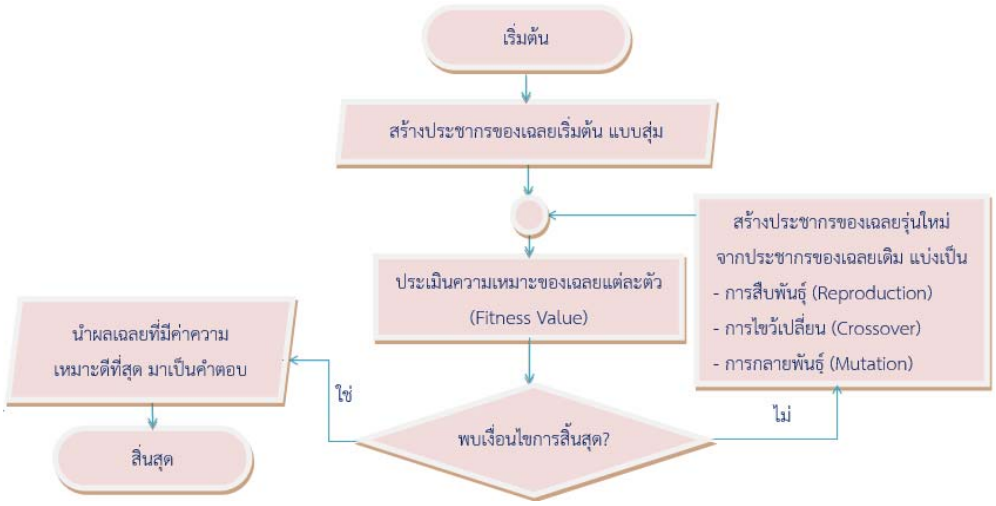
ขั้นตอนวิธีเชิงพันธุกรรม [2] เป็นอีกเทคนิคหนึ่งที่นิยมนำมาประยุกต์ใช้ร่วมกับเทคนิคการจำแนกประเภทพื้นฐานต่าง ๆ เพื่อทำให้ได้ตัวแบบการจำแนกประเภทที่มีประสิทธิภาพมากขึ้น เมื่อเทียบกับการใช้เทคนิคการจำแนกประเภทเพียงเทคนิคเดียว โดยมีงานวิจัยต่าง ๆ ได้นำขั้นตอนวิธีเชิงพันธุกรรมมาประยุกต์ใช้ เช่น Kelly & Davis [3] ได้นำเสนอการนำขั้นตอนวิธีเชิงพันธุกรรมมาใช้ในการค้นหาค่าน้ำหนักความสำคัญของคุณลักษณะ (attribute) ที่เหมาะสมเพื่อนำมาใช้จำแนกประเภทข้อมูลร่วมกับขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดเค ด้วยวิธีดังกล่าวทำให้การจำแนกประเภทข้อมูลมีประสิทธิภาพเพิ่มมากขึ้น และเมื่อไม่นานมานี้ Hengpraprohm *et al.* [4] ได้นำเสนอวิธีการสร้างตัวแบบจำแนกประเภทฐานจีเอ (GA - based classifier) สำหรับการจำแนกประเภทข้อมูลไมโครอาร์เรย์ (microarray data) โดยนำขั้นตอนวิธีเชิงพันธุกรรมมาใช้ในการเลือกคุณลักษณะที่สำคัญและทำการเปรียบเทียบค่าของข้อมูลจากคุณลักษณะที่เลือกได้ เพื่อใช้ในการจำแนกประเภทข้อมูลด้วยวิธีดังกล่าวทำให้การจำแนกประเภทข้อมูลมีประสิทธิภาพที่ดีเมื่อเทียบกับการจำแนกประเภทข้อมูลไมโครอาร์เรย์ด้วยเทคนิคการจำแนกประเภทพื้นฐานทั่วไป

ในงานวิจัยนี้ผู้วิจัยจึงต้องการศึกษาและพัฒนาวิธีการในการสร้างตัวแบบการจำแนกประเภทด้วยขั้นตอนวิธีเชิงพันธุกรรมให้สามารถใช้ได้กับปัญหาการจำแนกประเภทข้อมูลสำหรับชุดข้อมูลทั่วไปเพื่อให้สามารถใช้ได้อย่างกว้างขวางและเป็นอีกทางเลือกหนึ่งในการแก้ปัญหาการจำแนกประเภทข้อมูลให้มีประสิทธิภาพมากขึ้น

ทฤษฎีที่เกี่ยวข้อง

1. ขั้นตอนวิธีเชิงพันธุกรรม (Genetic Algorithm)

ขั้นตอนวิธีเชิงพันธุกรรม [2] เป็นวิธีการค้นหาคำตอบแบบสุ่ม (stochastic search) แก้ปัญหาด้วยการเลียนแบบกระบวนการวิวัฒนาการของสิ่งมีชีวิตตามธรรมชาติ โดยการถ่ายทอดลักษณะทางพันธุกรรมจากรุ่นหนึ่งสู่อีกรุ่นหนึ่งขั้นตอนการค้นหาคำตอบจะแบ่งออกเป็นขั้นตอนหลัก ๆ ได้แก่ 1) การสร้างประชากรของผลเฉลยเริ่มต้น 2) การประเมินค่าความเหมาะ (fitness Value) ของผลเฉลย 3) การสร้างประชากรของผลเฉลยรุ่นใหม่ และ 4) การแสดงคำตอบ ซึ่งกระบวนการของขั้นตอนวิธีเชิงพันธุกรรมแสดงในผังงาน (flowchart) ดังรูปภาพที่ 1



รูปภาพที่ 1 ขั้นตอนการทำงานของขั้นตอนวิธีเชิงพันธุกรรม

2. ขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดเค (K-Nearest Neighbor: KNN)

ขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดเค [1] เป็นวิธีการจำแนกประเภทข้อมูลที่ใช้วิธีการหาระยะห่างระหว่างคุณลักษณะของแต่ละข้อมูล ซึ่งวิธีนี้จะเหมาะสำหรับข้อมูลแบบตัวเลข โดยขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุดเค มีขั้นตอนโดยสรุปดังนี้

1) กำหนดจำนวนเพื่อนบ้าน k (นิยมกำหนดให้เป็นเลขคี่)

2) คำนวณระยะห่าง (distance) ของข้อมูลที่ต้องการพิจารณากับชุดข้อมูลสอน โดยสามารถคำนวณได้จากสมการระยะทางยูคริเดียน (Euclidean distance) ดังสมการ (1)

$$dist(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (1)$$

โดยที่ *dist(p,q)* หมายถึง ระยะห่างระหว่างข้อมูล p กับ q, p_i หมายถึง ค่าของข้อมูลคุณสมบัติที่ *i* ของข้อมูลที่ p และ q_i หมายถึง ค่าของข้อมูลคุณสมบัติที่ *i* ของข้อมูลที่ q

3) จัดลำดับของระยะห่างจากน้อยไปมากและเลือกชุดข้อมูลที่น้อยที่สุดตามจำนวน k ที่กำหนดไว้

4) กำหนดให้คำตอบของข้อมูลที่ต้องการทำนาย คือ กลุ่มที่มีจำนวนมากที่สุดในกลุ่มของชุดข้อมูล k ตัวแรก

3. การตรวจสอบแบบไขว้เคกลุ่ม (K-Folds cross validation)

การตรวจสอบแบบไขว้เคกลุ่ม [5] เป็นวิธีการวัดประสิทธิภาพของตัวแบบที่นิยมใช้กันอย่างแพร่หลาย การตรวจสอบแบบไขว้เคกลุ่มนั้นจะแบ่งข้อมูลออกเป็นกลุ่ม ตามจำนวนเคที่กำหนดกลุ่มละเท่า ๆ กันและนำมาแบ่งออกเป็น 2 ชุดได้แก่ ข้อมูลสอน (training data) และข้อมูลทดสอบ (testing data) โดยใช้ข้อมูล 1 กลุ่มเป็นข้อมูลทดสอบและข้อมูลส่วนที่เหลือเป็นข้อมูลสอน โดยจะวนสลับให้ข้อมูลทุกกลุ่มเป็นข้อมูลทดสอบครบทั้งเคกลุ่ม ซึ่งจะทำให้ได้การทดลองทั้งสิ้นเคการทดลอง

4. การทำข้อมูลให้เป็นปกติ (data normalization)

การทำข้อมูลให้เป็นปกติ [6] เป็นวิธีการหนึ่งของการแปลงข้อมูล (data transformation) ในกระบวนการเตรียมข้อมูล (data preparation) ของการทำเหมืองข้อมูล วิธีการพื้นฐานที่นิยมใช้ มีดังนี้

1) วิธีน้อย-มาก (min-max) เป็นการปรับค่าคุณลักษณะของข้อมูลให้อยู่ในช่วงค่าน้อยสุด และค่ามากสุดตามที่กำหนด ซึ่งนิยมใช้ค่าน้อยสุดเป็น 0 และ ค่ามากสุดเป็น 1 บางครั้งนิยมเรียกวิธีการนี้ว่า การทำข้อมูลให้เป็นปกติแบบ 0-1 (0-1 normalization) วิธีการแปลงค่าคำนวณได้จากสมการที่ (2)

$$v'_i = \frac{v_i - MIN_v}{MAX_v - MIN_v} \quad (2)$$

โดยที่ v'_i หมายถึงค่าใหม่ของคุณลักษณะตัวที่ *i* ของข้อมูล *v*, v_i หมายถึง ค่าของคุณลักษณะตัวที่ *i* ของข้อมูล *v* เดิม, MAX_v หมายถึง ค่าที่มากที่สุดของคุณลักษณะนั้น และ MIN_v หมายถึง ค่าที่น้อยที่สุดของคุณลักษณะนั้น

2) วิธีคะแนนซี (z-score) เป็นการปรับค่าของคุณลักษณะโดยการแปลงให้เป็นคะแนนมาตรฐานซี เพื่อปรับให้ค่าเฉลี่ยของข้อมูลมีค่าเป็น 0 และ ค่าส่วนเบี่ยงเบนมาตรฐานมีค่าเป็น 1 คำนวณได้ตามสมการที่ (3)

$$v'_i = \frac{v_i - Mean_v}{SD_v} \quad (3)$$

โดยที่ v'_i หมายถึง ค่าใหม่ของคุณลักษณะตัวที่ *i* ของข้อมูล *v*, v_i หมายถึง ค่าของคุณลักษณะตัวที่ *i* ของข้อมูล *v* เดิม $Mean_v$ หมายถึง ค่าเฉลี่ยของคุณลักษณะนั้น และ SD_v หมายถึง ค่าส่วนเบี่ยงเบนมาตรฐานของคุณลักษณะนั้น

วิธีดำเนินการวิจัย

1. ข้อมูลที่ใช้ในงานวิจัย

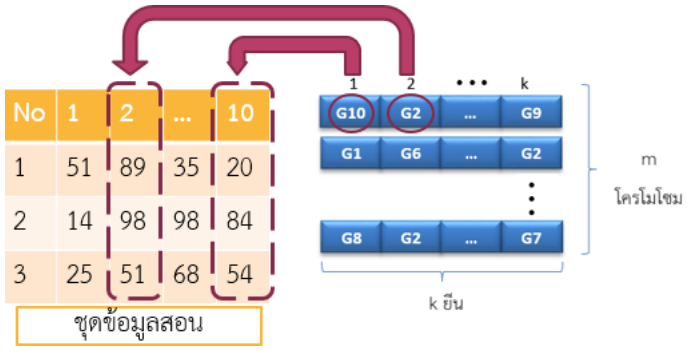
ข้อมูลที่ใช้ในงานวิจัยเป็นชุดข้อมูลจากเว็บไซต์ UCI [7] ซึ่งเป็นแหล่งรวบรวมข้อมูลเกณฑ์มาตรฐาน (benchmark data) โดยได้ทำการเลือกข้อมูลสำหรับปัญหาการจำแนกประเภทข้อมูลแบบ 2 กลุ่มจำนวน 9 ชุดข้อมูล โดยแบ่งออกเป็น 3 กลุ่ม ดังนี้

ตารางที่ 1 รายละเอียดชุดข้อมูลที่ใช้ในงานวิจัย

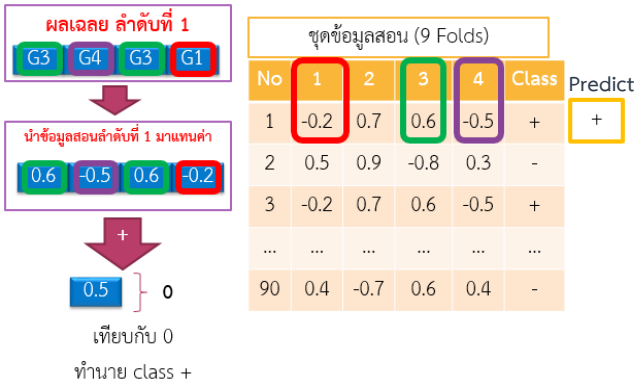
กลุ่มข้อมูล	ชุดข้อมูลที่ใช้ในงานวิจัย	จำนวนคุณลักษณะ	ข้อมูลตัวอย่าง
จำนวนคุณลักษณะน้อย (น้อยกว่า 100)	1. Indian Liver Patient	10	583
	2. Breast Cancer	30	569
	3. Ionosphere	34	351
จำนวนคุณลักษณะปานกลาง (100-1,000)	4. LSVT	309	126
	5. Madelon	500	4,000
	6. SECOM	591	1,567
จำนวนคุณลักษณะมาก (มากกว่า 1,000)	7. Ad	1,558	3,279
	8. DLBCL	4,026	47
	9. Lung Cancer	12,533	181

2. วิธีการที่นำเสนอ

ในการพัฒนาตัวแบบที่นำเสนอนั้น จะนำชุดข้อมูลเกณฑ์มาตรฐานไปแปลงข้อมูลด้วยวิธีคะแนนซีและได้นำการออกแบบโครโมโซมสำหรับขั้นตอนวิธีเชิงพันธุกรรมของ Hengpraprohm *et al.* [4] มาประยุกต์ใช้ในงานวิจัยนี้ โดยกำหนดจำนวนยีนเท่ากับ k ซึ่งหมายถึงขนาดความยาวโครโมโซม และ m คือจำนวนโครโมโซม (จำนวนประชากรของผลเฉลย ซึ่งในงานวิจัยนี้กำหนดให้ m=400) และค่าตัวเลขในแต่ละยีน (G) คือค่าตำแหน่งคุณลักษณะของข้อมูลแสดงดังรูปภาพที่ 3 การหาค่าความเหมาะจะคำนวณได้จากความถูกต้องของการจำแนกประเภทบนชุดข้อมูลสอนโดยจะนำโครโมโซมแต่ละตัวมาแทนค่าด้วยข้อมูลในตำแหน่งคุณลักษณะที่ยีนในโครโมโซมนั้นสุ่มได้ จากนั้นหาค่าผลรวมของข้อมูลจากยีนที่เลือกได้ด้วยขั้นตอนวิธีเชิงพันธุกรรม และนำไปเปรียบเทียบกับ 0 ถ้ามีค่ามากกว่าจะจำแนกเป็นประเภทที่ 1 แต่ถ้าไม่ใช่จะจำแนกให้เป็นประเภทที่ 2 และนำผลการจำแนกที่ได้มาคำนวณหาความถูกต้อง ตัวอย่างของการคำนวณแสดงดังรูปภาพที่ 4



รูปภาพที่ 3 โครงสร้างโครโมโซมที่ใช้ในการทดลอง



รูปภาพที่ 4 ตัวอย่างการประเมินค่าความเหมาะจากข้อมูลสอน

ในแต่ละการทดลองจะกำหนดค่า k=1 ไปถึง k=10 และใช้การตรวจสอบความถูกต้องด้วยวิธีการตรวจสอบแบบไขว้ 10 กลุ่ม โดยมีการกำหนดค่าพารามิเตอร์ที่ใช้สำหรับขั้นตอนวิธีเชิงพันธุกรรม แสดงดังตารางที่ 2

ตารางที่ 2 รายละเอียดค่าพารามิเตอร์ที่ใช้สำหรับขั้นตอนวิธีเชิงพันธุกรรม

พารามิเตอร์	กำหนด	หมายถึง
m	400	จำนวนโครโมโซม
k	1-10	จำนวนยีน
gen	800	จำนวนรุ่น
Reproduction rate	5%	อัตราการสืบพันธุ์
Crossoverrate	93%	อัตราการไขว้เปลี่ยน
Mutationrate	2%	อัตราการกลายพันธุ์

เนื่องจากขั้นตอนวิธีเชิงพันธุกรรมเป็นวิธีการเชิงสุ่มทำให้ผลลัพธ์ที่ได้ในแต่ละรอบไม่เหมือนเดิมเพื่อให้เกิดความเชื่อมั่นตามหลักสถิติจึงได้ทำการทดลองซ้ำ 10 รอบ (10 รอบ x การตรวจสอบแบบไขว้ 10 กลุ่ม = 100 การทดลอง) และรายงานผลเป็นค่าเฉลี่ยของความถูกต้องและส่วนเบี่ยงเบนมาตรฐานโดยได้ดำเนินการทดลองปรับค่า k=1 ไปถึง k=10 และนำค่าเฉลี่ยของความถูกต้องและส่วนเบี่ยงเบนมาตรฐานมาเปรียบเทียบกับวิธีเพื่อนบ้านใกล้ที่สุดเคซึ่งเป็นค่าเฉลี่ยของการปรับค่า k=1 ไปถึง k=10 โดยใช้การตรวจสอบความถูกต้องด้วยวิธีการตรวจสอบแบบไขว้ 10 กลุ่มเช่นกัน นอกจากนี้ยังได้นำผลการทดลองไปเปรียบเทียบประสิทธิภาพของการจำแนกประเภทข้อมูลด้วยวิธีต้นไม้ตัดสินใจและวิธีเบย์อย่างง่าย เพื่อเป็นการยืนยันประสิทธิภาพของวิธีการที่นำเสนอด้วย

ผลการทดลอง

การทดลองเปรียบเทียบประสิทธิภาพของวิธีการที่นำเสนอเทียบกับขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดเคได้ทำการทดสอบกับชุดข้อมูลที่มีคุณลักษณะแตกต่างกัน 3 กลุ่ม กลุ่มละ 3 ชุดข้อมูลรวมทั้งหมด 9 ชุดข้อมูลตามที่ได้อธิบายไว้ในหัวข้อที่ผ่านมา และทำการวัดประสิทธิภาพด้วยวิธีการตรวจสอบแบบไขว้ 10 กลุ่ม โดยได้ทำการทดลองซ้ำ 10 รอบ (รวมเป็น 100 การทดลอง) และรายงานผลเป็นค่าเฉลี่ยของความถูกต้องและส่วนเบี่ยงเบนมาตรฐานโดยนำมาเปรียบเทียบกับวิธีเพื่อนบ้านใกล้ที่สุดเคซึ่งได้ทำการทดลองกับทั้งชุดข้อมูลเดิม และทดลองกับชุดข้อมูลด้วยการทำข้อมูลให้เป็นปกติอีก 2 แบบ ได้แก่ วิธีน้อย-มากและวิธีคะแนนซี ผลการทดลองแสดงดังตารางที่ 3

ตารางที่ 3 ผลการเปรียบเทียบค่าเฉลี่ยของความถูกต้องของวิธีการที่นำเสนอกับขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดเค

ชุดข้อมูล	K-GA	KNN (ข้อมูลเดิม)	KNN (min-max)	KNN (z-score)
1. Indian Liver Patient	62.61%±1.78	62.47%±1.70 p-value : 0.85	**58.99%±2.21** **p-value : 0.01***	**60.92%±2.27** **p-value : 0.04***
2. Breast Cancer	95.96%±1.28	**92.08%±0.42** **p-value : 7.96E-07***	96.19%±0.68 p-value : 0.46	95.95%±0.60 p-value : 0.96
3. Ionosphere	85.23%±2.68	83.93%±1.27 p-value : 0.17	**83.28%±0.88** **p-value : 0.04***	84.44%±1.32 p-value : 0.30
4. LSVT	68.94%±2.54	**55.76%±3.84** **p-value : 2.38E-06***	78.18%±4.10 p-value : 0.01*	75.96%±6.11 p-value : 0.01*
5. Madelon	55.73%±1.99	67.95%±0.68 p-value : 5.16E-09*	58.26%±0.80 p-value : 0.01*	54.70%±0.71 p-value : 0.19
6. SECOM	81.96%±11.10	91.71%±2.79 p-value : 0.04*	90.13%±3.91 p-value : 0.10	89.51%±4.74 p-value : 0.12

ตารางที่ 3 (ต่อ) ผลการเปรียบเทียบค่าเฉลี่ยของความถูกต้องของวิธีการที่นำเสนอกับขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดเค

ชุดข้อมูล	K-GA	KNN (ข้อมูลเดิม)	KNN (min-max)	KNN (z-score)
7. Ad	92.65%±2.61	92.63%±0.32 p-value : 0.96	93.35%±0.30 p-value : 0.38	92.55%±0.34 p-value : 0.89
8. DLBCL	77.11%±6.78	76.40%±4.17 p-value : 0.83	74.75%±3.90 p-value : 0.45	74.20%±3.86 p-value : 0.36
9. Lung Cancer	83.66%±8.86	98.28%±1.55 p-value : 0.01*	91.76%±2.37 p-value : 0.01*	90.23%±2.57 p-value : 0.01*

หมายเหตุ: * หมายถึง ค่าที่ความถูกต้องแตกต่างกันกับวิธีที่นำเสนออย่างมีนัยสำคัญทางสถิติที่ระดับ .05

จากตารางที่ 3 ค่าที่ถูกแรเงา คือ ค่าที่ไม่แตกต่างกันทางสถิติเมื่อเปรียบเทียบกับวิธีการที่นำเสนอ และค่าที่เป็นตัวหนาคือ ค่าที่วิธีการที่นำเสนอให้ค่าความถูกต้องสูงกว่าอย่างมีนัยสำคัญทางสถิติ โดยพบว่า ตัวแบบที่นำเสนอสามารถให้ค่าเฉลี่ยของความถูกต้องที่ไม่แตกต่างทางสถิติจำนวน 14 ค่า จากทั้งหมด 27 ค่า คิดเป็นร้อยละ 51.85 และให้ค่าเฉลี่ยของความถูกต้องสูงกว่าวิธีเพื่อนบ้านใกล้ที่สุดเคอย่างมีนัยสำคัญทางสถิติจำนวน 5 ค่าซึ่งเมื่อรวมกับวิธีการที่ให้ค่าความถูกต้องที่ไม่แตกต่างกันทางสถิติจะมีค่าเป็น 19 ค่า จากทั้งหมด 27 ค่า คิดเป็นร้อยละ 70.37 และพบว่า วิธีการที่นำเสนอให้ประสิทธิภาพที่ดี เมื่อข้อมูลมีจำนวนคุณลักษณะน้อย (น้อยกว่า 100) คือ วิธีการที่นำเสนอให้ค่าความถูกต้องเทียบเท่าหรือดีกว่าอย่างมีนัยสำคัญทางสถิติที่ระดับ .05

นอกจากนี้ผู้วิจัยได้ทำการเปรียบเทียบกับวิธีการจำแนกประเภทข้อมูลแบบอื่น ๆ อีก 2 วิธี ได้แก่ วิธีต้นไม้ตัดสินใจและวิธีเบย์อย่างง่ายโดยนำชุดข้อมูลไปเตรียมข้อมูลด้วยวิธีการทำข้อมูลแบบไม่ต่อเนื่อง โดยการแบ่งช่วงข้อมูลออกเป็น 3 5 และ 7 ช่วง และใช้โปรแกรม Weka ในการหาค่าความถูกต้องและกำหนดการตรวจสอบแบบไขว้เป็น 10 กลุ่ม ผลการเปรียบเทียบแสดงดังตารางที่ 4

ตารางที่ 4 ผลการเปรียบเทียบค่าเฉลี่ยของความถูกต้องของวิธีการที่นำเสนอกับวิธีต้นไม้ตัดสินใจและวิธีเบย์อย่างง่าย

ชุดข้อมูล	K-GA	ช่วง 3		ช่วง 5		ช่วง 7	
		Decision Tree	Naïve Bayes	Decision Tree	Naïve Bayes	Decision Tree	Naïve Bayes
1. Indian Liver Patient	65.30%	**65.93%**	66.91%	65.68%	**68.87%**	64.32%	65.52%
2. Breast Cancer	**97.22%**	93.31%	94.19%	91.19%	94.19%	91.19%	94.36%
3. Ionosphere	87.46%	**89.14%**	77.14%	87.14%	88.57%	87.71%	**90.57%**
4. LSVT	70.79%	73.01%	**73.81%**	**76.98%**	67.46%	71.42%	69.04%
5. Madelon	**58.25%**	55.50%	57.16%	56.66%	53.16%	56.33%	57.00%
6. SECOM	**93.07%**	87.93%	89.72%	85.83%	88.32%	86.02%	88.96%
7. Ad	94.72%	**96.88%**	96.37%	96.43%	**96.52%**	96.49%	96.40%
8. DLBCL	**86.35%**	68.08%	95.74%	70.21%	**97.87%**	85.10%	95.74%
9. Lung Cancer	**94.65%**	91.16%	**95.02%**	93.92%	**95.02%**	93.92%	92.81%

จากตารางที่ 4 ค่าที่ถูกแรเงา คือ ค่าความถูกต้องที่ต่ำกว่าวิธีการที่นำเสนอ และตัวเลขที่เป็นตัวเข้ม คือ ค่าความถูกต้องที่มีประสิทธิภาพดีที่สุด โดยพบว่า วิธีการที่นำเสนอให้ค่าความถูกต้องดีกว่าวิธีต้นไม้ตัดสินใจ 17 ค่า จากทั้งหมด 27 ค่า คิดเป็นร้อยละ 62.96 และให้ค่าความถูกต้องดีกว่าวิธีเบย์อย่างง่าย 13 ค่าจากทั้งหมด 27 ค่าคิดเป็นร้อยละ 48.15 ดังนั้น โดยภาพรวมวิธีการที่นำเสนอให้ค่าความถูกต้องที่ดีกว่าทั้ง 2 วิธี คิดเป็นร้อยละ 55.55 และเมื่อพิจารณาค่าความถูกต้องที่ดีที่สุดพบว่า วิธีการที่นำเสนอให้ค่าความถูกต้องดีที่สุด 5 ชุดข้อมูล จาก 9 ชุดข้อมูล

สรุปผลการวิจัย

งานวิจัยนี้ได้ทำการศึกษาและพัฒนาการจำแนกประเภทข้อมูลด้วยตัวจำแนกประเภทขั้นตอนวิธีเชิงพันธุกรรมขนาดความยาวโครโมโซมเคให้สามารถใช้ได้กับการแก้ไขปัญหาการจำแนกประเภทข้อมูลแบบ 2 กลุ่ม ซึ่งได้ทดสอบกับชุดข้อมูล 3 กลุ่ม กลุ่มละ 3 ชุดข้อมูลรวมเป็น 9 ชุดข้อมูล โดยนำขั้นตอนวิธีเชิงพันธุกรรมมาพัฒนาให้สามารถจำแนกประเภทข้อมูลได้อย่างมีประสิทธิภาพเทียบเท่าหรือดีกว่าเทคนิคการจำแนกประเภทข้อมูลพื้นฐานทั่วไปซึ่งเทคนิคที่ได้นำมาทดสอบเปรียบเทียบนั้นประกอบด้วย วิธีเพื่อนบ้านใกล้ที่สุดเค วิธีต้นไม้ตัดสินใจ และวิธีเบย์อย่างง่าย โดยวิธีเพื่อนบ้านใกล้ที่สุดเคนั้นได้ทำการเปรียบเทียบกับชุดข้อมูลแบบดั้งเดิมและข้อมูลที่ผ่านการทำข้อมูลให้เป็นปกติอีก 2 วิธี ส่วนวิธีต้นไม้ตัดสินใจและวิธีเบย์อย่างง่ายจะทำการเตรียมข้อมูลด้วยวิธีการทำข้อมูลแบบไม่ต่อเนื่อง โดยการแบ่งช่วงข้อมูลที่เท่ากัน 3 5 และ 7 ช่วง โดยวิธีทั้งหมดจะใช้การตัวสอบความถูกต้องด้วยวิธีการตรวจสอบประสิทธิภาพแบบไขว้ 10 กลุ่มเช่นกัน จากการทดลองพบว่า ตัวแบบที่นำเสนอนั้นสามารถให้ประสิทธิภาพในด้านความถูกต้องของการจำแนกประเภทข้อมูลเทียบเคียงกับวิธีการจำแนกประเภทข้อมูลพื้นฐานทั่วไป

เอกสารอ้างอิง

[1] Song, Y., Huang, J., Zhou, D., Zha, H., & Lee Giles, C. (2007). IKNN: Informative k-nearest neighbor pattern classification. **Knowledge Discovery in Databases: PKDD 2007**, 4702, 248-264.

[2] Holland, J.H. (1975). **Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence.** USA: MIT Press Cambridge.

[3] Kelly, J.D. & Davis, L. (1991). A hybrid genetic algorithm for classification. **In Proceedings of the 12th International joint Conference on Artificial Intelligence**, 2, 645-650.

[4] Hengpraprohm, S., Mukviboonchai, S., Thammasang, R., & Chongstitvatana, P. (2010). A GA-based classifier for microarray data classification. **Intelligent Computing and Cognitive Informatics (ICICCI): 2010 International Conference**, 199-202.

[5] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. **In Proceedings of the 14th international joint conference on Artificial intelligence**, 2, 1137-1143.

[6] Al Shalabi, L., Shaaban, Z., & Kasasbeh, B. (2006). Data mining: A preprocessing engine. **Journal of Computer Science**, 2 (9), 735-739.

[7] **The UC Irvine Machine Learning Repository**. From http://archive.ics.uci.edu/ml/index.html.