



การแทนที่ข้อมูลสูญหายด้วยวิธีการเชิงพันธุกรรม และการถดถอยเชิงเส้น
พหุคูณ เพื่อปรับปรุงความแม่นยำของแบบจำลองทำนายข้อมูล

โดย

สุรวัช อำพัน

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)
สาขาวิชาวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
ปีการศึกษา 2565

การแทนที่ข้อมูลสูญหายด้วยวิธีการเชิงพันธุกรรม และการถดถอยเชิงเส้น
พหุคูณ เพื่อปรับปรุงความแม่นยำของแบบจำลองทำนายข้อมูล

โดย

สุรวัช อำพัน



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)
สาขาวิชาวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
ปีการศึกษา 2565

IMPUTATION WITH GENETIC ALGORITHM AND MULTIPLE LINEAR
REGRESSION FOR IMPROVING PREDICTION MODEL

BY

SURAWACH AMPHAN



A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE (COMPUTER SCIENCE)
DEPARTMENT OF COMPUTER SCIENCE
FACULTY OF SCIENCE AND TECHNOLOGY
THAMMASAT UNIVERSITY
ACADEMIC YEAR 2022

มหาวิทยาลัยธรรมศาสตร์
คณะวิทยาศาสตร์และเทคโนโลยี

วิทยานิพนธ์

ของ

สุรวิช อำพัน


เรื่อง

การแทนที่ข้อมูลสูญหายด้วยวิธีการเชิงพันธุกรรม และการถดถอยเชิงเส้นพหุคูณ เพื่อปรับปรุงความ
แม่นยำของแบบจำลองทำนายข้อมูล

ได้รับการตรวจสอบและอนุมัติ ให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต วิทยาการคอมพิวเตอร์

เมื่อ วันที่ 21 ธันวาคม พ.ศ. 2565

ประธานกรรมการสอบวิทยานิพนธ์


(ผู้ช่วยศาสตราจารย์ ดร.ปกรณ์ ลีสุทธิพรชัย)

กรรมการและอาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก


(ผู้ช่วยศาสตราจารย์ ดร.ปอง ส่องเมือง)

กรรมการสอบวิทยานิพนธ์


(ผู้ช่วยศาสตราจารย์ ดร.ธนาธร ทะนานทอง)

กรรมการสอบวิทยานิพนธ์


(ผู้ช่วยศาสตราจารย์ ดร.เกรียงศักดิ์ เตมีย์)

คณบดี


(รองศาสตราจารย์ ดร.สุเพชร จิระขจรกุล)

หัวข้อวิทยานิพนธ์	การแทนที่ข้อมูลสูญหายด้วยวิธีการเชิงพันธุกรรม และการถดถอยเชิงเส้นพหุคูณ เพื่อปรับปรุงความแม่นยำของแบบจำลองทำนายข้อมูล
ชื่อผู้เขียน	สุรัช อำพัน
ชื่อปริญญา	วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)
สาขาวิชา/คณะ/มหาวิทยาลัย	สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์	ผู้ช่วยศาสตราจารย์ ดร.ปกป้อง ส่องเมือง
ปีการศึกษา	2565

บทคัดย่อ

การทำเหมืองข้อมูลเป็นเป็นเทคนิคที่ใช้สำหรับพยากรณ์หรือสกัดข้อมูลบางอย่างจากข้อมูลขนาดใหญ่ แต่ปัญหาหนึ่งที่มีกพบในการทำเหมืองข้อมูลคือ การสูญหายของข้อมูล (Missing Values) และวิธีที่นิยมใช้ในการจัดการกับปัญหาดังกล่าวมีอยู่ 2 วิธีคือ การลบข้อมูลที่สูญหาย (Ignoring) และการแทนที่ข้อมูลที่สูญหาย (Imputation) อย่างไรก็ตามวิธีการตัดข้อมูลที่สูญหายออกอาจลดประสิทธิภาพในการทำนายของการทำเหมืองข้อมูล เนื่องจากข้อมูลที่ถูกลบออกเป็นข้อมูลสำคัญ เราจึงนำเสนอวิธีการแทนที่ข้อมูลในแก้ไขปัญหาดังกล่าวด้วยการปรับปรุง อัลกอริทึมการค้นหาเพื่อนบ้านใกล้สุด k ตัว (K-Nearest Neighbors Algorithm) ด้วยการวิเคราะห์การถดถอยเชิงเส้นตรง (Linear Regression) และการเลือกเสียงส่วนใหญ่ (Majority) โดยจะทำการเปรียบเทียบกับอัลกอริทึมอื่นได้แก่ ขั้นตอนวิธีการค้นหาเพื่อนบ้านใกล้สุด k ตัวแบบดั้งเดิม และ ขั้นตอนวิธีเชิงพันธุกรรม (Genetic Algorithm)

คำสำคัญ: ข้อมูลสูญหาย, การแทนที่ข้อมูลสูญหาย, ขั้นตอนเชิงพันธุกรรม, การถดถอยเชิงเส้นพหุคูณ

Thesis Title	IMPUTATION WITH GENETIC ALGORITHM AND MULTIPLE LINEAR REGRESSION FOR IMPROVING PREDICTION MODEL
Author	Surawach Amphan
Degree	Master of Science (Computer Science)
Department/Faculty/University	Department of Computer Science Faculty of Science and Technology Thammasat University
Thesis Advisor	Associate Professor Pokpong Songmuang, Ph.D.
Academic Year	2022

ABSTRACT

Prediction model is used to forecast or predict value from dataset. But one of the most common problems in training prediction model is there are missing values in datasets. Problem is usually managed by two methods for solving this problem. First is ignoring, but it reduces the predictive model's performance because of the data that was cut off may be important. Another method is replacing the missing values or data imputation. Benefit of imputation is it still keep all of data. It means an important data will not loss. Therefore, most researchers offer an imputation method for solving this problem. In the past most researches are proposed algorithm that trying to recover the original data, but main object of using prediction model is accuracy of prediction. Algorithm is based on Genetics Algorithm and Multiple Linear Regression is create for improving performance of prediction model

Keywords: MISSING VALUES, IMPUTATION, GENETIC ALGORITHM, MULTIPLE LINEAR REGRESSION

กิตติกรรมประกาศ

การทำวิจัยสำหรับวิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงอย่างเรียบร้อย เนื่องด้วยความกรุณาอย่างยิ่งจาก ผู้ช่วยศาสตราจารย์ ดร.ปกป้อง ส่องเมือง อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งกรุณาให้คำปรึกษาและคำแนะนำที่เป็นประโยชน์ในการดำเนินการวิจัยจนสำเร็จ ผู้วิจัยจะระลึกถึงพระคุณความทุ่มเทเสียสละที่ทุ่มเทให้ผู้วิจัย จึงขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ ที่นี้

ขอขอบคุณคณะกรรมการการสอบวิทยานิพนธ์ซึ่งประกอบด้วย ผู้ช่วยศาสตราจารย์ ดร.ปกรณ์ ลีสุทธิพรชัย ผู้ช่วยศาสตราจารย์ ดร.เกรียงศักดิ์ เตมีย์ และ ผู้ช่วยศาสตราจารย์ ดร.ธนาธร ทะทานทอง สำหรับข้อเสนอแนะเพิ่มเติมให้งานวิจัยสมบูรณ์ยิ่งขึ้นและการสละเวลาในการตรวจสอบวิทยานิพนธ์

ขอขอบพระคุณบุคลากรคณะวิทยาศาสตร์หลายท่านที่ช่วยดำเนินการเอกสารต่างๆให้สำเร็จลุล่วงไปได้เป็นอย่างดี

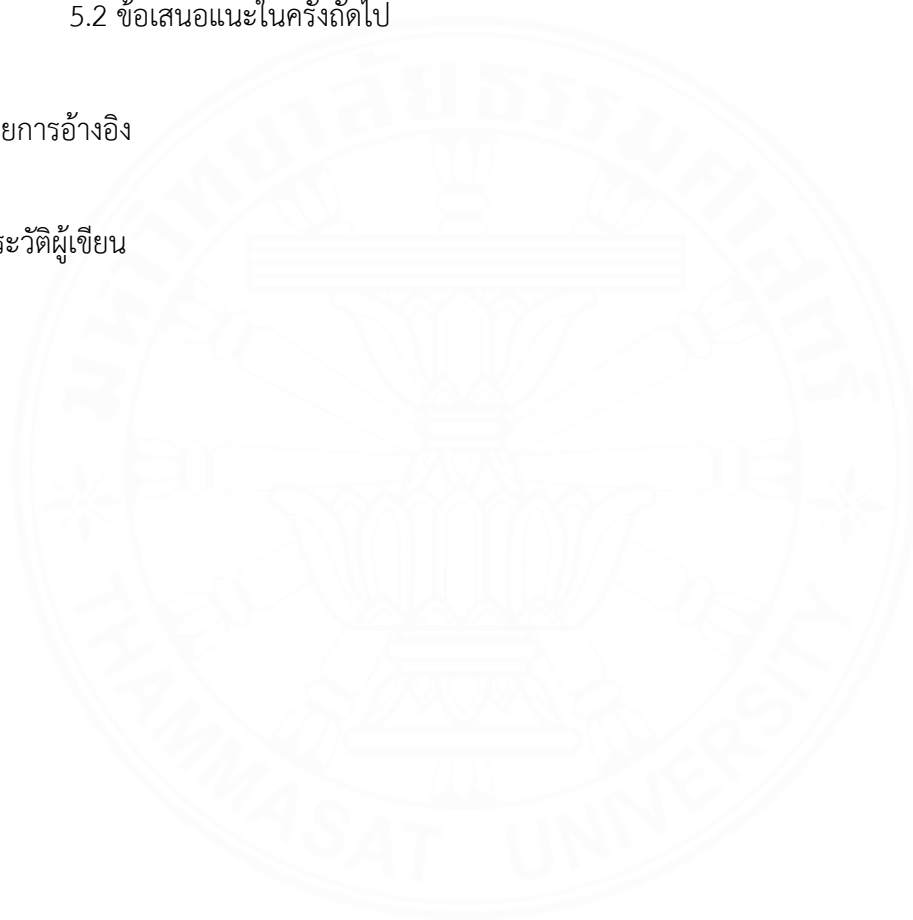
สุรวัช อำพัน

สารบัญ

	หน้า
บทคัดย่อภาษาไทย	(1)
บทคัดย่อภาษาอังกฤษ	(2)
กิตติกรรมประกาศ	(3)
สารบัญตาราง	(6)
สารบัญรูปภาพ	(7)
รายการสัญลักษณ์และคำย่อ	(8)
บทที่ 1 บทนำ	1
1.1 ความเป็นมา	1
1.1.1 การลบข้อมูลทั้งหมดที่มีสูญหายออก (Ignoring)	1
1.1.2 การแทนที่ข้อมูล (Missing Value Imputation)	1
1.1.2.1 วิธีการทางสถิติ (Statical Approach)	1
1.1.2.2 วิธีการการเรียนรู้ของเครื่อง (Machine Learning Approach)	2
1.2 วัตถุประสงค์งานวิจัย	2
บทที่ 2 วรรณกรรมและงานวิจัยที่เกี่ยวข้อง	3
2.1 K-nearest neighbors algorithm (KNN)	3
2.2 Predictive Mean Matching (PMM)	5
2.3 Linear Regression	6

2.4 Genetic Algorithm	7
2.4.1 การคัดเลือกผู้อยู่รอด (Selection)	7
2.4.2 สร้างลูกผสม (Crossover)	7
2.4.2.1 Roulette Wheel	7
2.4.2.2 Tournament Selection	8
2.4.3 กลายพันธุ์ (Mutation)	8
2.5 Multiple Linear Regression	9
2.6 Clusterwise Linear Regression	11
 บทที่ 3 วิธีการวิจัย	 12
3.1 ภาพรวมระบบ	12
3.2 ขั้นตอนการเตรียมการทดลอง	12
3.3 Combination Of Genetic Algorithm and Multiple Linear Regression	13
 บทที่ 4 ผลวิจัยและอภิปรายผล	
4.1 การทดลอง	19
4.1.1 การทดลองบนชุดข้อมูลสุญหาย Wine	20
4.1.2 การทดลองบนชุดข้อมูลสุญหาย Glass	21
4.1.3 การทดลองบนชุดข้อมูลสุญหาย Indian Patient Liver	22
4.1.4 การทดลองบนชุดข้อมูลสุญหาย Seed	23
4.2 การทดลองหาความผิดพลาดของการแทนที่ข้อมูลสุญหายด้วย RMSE	24
4.2.1 การทดลองบนชุดข้อมูลสุญหาย Wine	24
4.2.2 การทดลองบนชุดข้อมูลสุญหาย Glass	25

4.2.3 การทดลองบนชุดข้อมูลสุญหาย Indian Patient Liver	25
4.2.4 การทดลองบนชุดข้อมูลสุญหาย Seed	26
บทที่ 5 สรุปผลวิจัยและข้อเสนอแนะ	27
5.1 อภิปราย	27
5.2 ข้อเสนอแนะในครั้งถัดไป	28
รายการอ้างอิง	29
ประวัติผู้เขียน	31



สารบัญตาราง

ตารางที่	หน้า
2.1 ตารางแสดงชุดข้อมูลที่นำมาทดสอบ Cumulative Linear Regression	6
3.1 ตารางแสดงชุดข้อมูล UCI ที่มาใช้ในการทดสอบ	13
4.1 ตารางแสดงผลการเปรียบเทียบการทดลองบนชุดข้อมูล Wine	20
4.2 ตารางแสดงผลการเปรียบเทียบการทดลองบนชุดข้อมูล Glass	21
4.3 ตารางแสดงผลการเปรียบเทียบการทดลองบนชุดข้อมูล Indian Patien Liver	22
4.4 ตารางแสดงผลการเปรียบเทียบการทดลองบนชุดข้อมูล Seed	23
4.5 ตารางเปรียบเทียบค่า RMSE บนชุดข้อมูล Wine	24
4.6 ตารางเปรียบเทียบค่า RMSE บนชุดข้อมูล Glass	24
4.7 ตารางเปรียบเทียบค่า RMSE บนชุดข้อมูล Indian Liver Patient	25
4.8 ตารางเปรียบเทียบค่า RMSE บนชุดข้อมูล Seed	26

สารบัญรูปภาพ

ภาพที่	หน้า
2.1 ภาพแสดงการเลือกพ่อแม่ด้วยวิธี Roulette Wheel	7
2.2 ภาพแสดงการเลือกพ่อแม่ด้วยวิธี Tournament Selection	8
2.3 ภาพแสดงการกลายพันธุ์ หรือ Mutation	8
2.4 ภาพแสดงการกลายพันธุ์ หรือ Mutation	9
2.5 ภาพแสดงค่าสัมประสิทธิ์ β ของสมการ Multiple Linear Regression	10
3.1 ภาพแสดง Genetic Algorithm and Multiple Linear Regression Algorithm	12
3.2 ภาพที่ 3.2 ภาพแสดงการแทนที่ข้อมูลสูญหายจากค่า min-max	14
3.3 ภาพที่ 3.3 ภาพแสดงการสร้าง chromosome จากชุดข้อมูลสูญหาย	14
3.4 ภาพที่ 3.4 ภาพแสดงตารางการเก็บ chromosomes พร้อม fitness	15
3.5 ภาพที่ 3.5 ภาพแสดงการ crossover ของรุ่นพ่อแม่เพื่อให้กำเนิดรุ่นลูก	16
3.6 ภาพที่ 3.6 ภาพแสดงการสร้างชุดข้อมูลที่สมบูรณ์เพื่อใช้ในสมการ Multiple Linear Regression	17
3.7 แผนภาพแสดงการทำของ Genetic Algorithm and Multiple Linear Regression Algorithm	18

รายการสัญลักษณ์และคำย่อ

สัญลักษณ์/คำย่อ

คำเต็ม/คำจำกัดความ

KNN

K-nearest Neighbors Algorithm

GA

Genetic Algorithm

LR

Linear Regression

MLR

Multiple Linear Regression

MAR

Missing at Random

MCAR

Missing Completely at Random

MNAR

Missing Not at Random



บทที่ 1

บทนำ

1.1 ความเป็นมา

การทำเหมืองข้อมูล (Data Mining) คือวิธีการ ค้นหา หรือ สกัด รูปแบบหรือความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูล ปัจจุบันการทำเหมืองข้อมูลถูกนำไปใช้ในหลายด้าน อาทิเช่น ด้านการเงิน (Finance) ธนาคารหรือองค์กรด้านการเงินสามารถเสนอสินเชื่อให้กับกลุ่มลูกค้าใหม่ที่กำลังจะมีความต้องการในการกู้เงิน หรือสามารถประมาณวงเงินกู้ที่เหมาะสมให้กับลูกค้าที่มาขอสินเชื่อได้ ซึ่งคุณภาพของผลลัพธ์ที่ได้จากการทำเหมืองข้อมูล จะมีผลมาจากข้อมูลที่ทำมาใช้ โดยยังมีจำนวนและความสมบูรณ์ของข้อมูลที่น่าสนใจมากเท่าใด ยิ่งจะทำให้ผลลัพธ์ออกมาเชื่อถือในทางสถิติเท่านั้น

อย่างไรก็ตามปัญหาหนึ่งที่นักวิจัยมักพบคือ การสูญหายของข้อมูล (Missing Values) และหากนำข้อมูลไปใช้ทันทีอาจส่งผลให้ความแม่นยำในการทำเหมืองข้อมูลลดลง จึงต้องมีขั้นตอนการเตรียมข้อมูล (Data Preparation) เพื่อทำให้ข้อมูลมีความเหมาะสมที่จะนำไปทำเหมืองข้อมูล โดยสามารถแบ่งประเภทของการสูญหายของข้อมูลได้เป็น 3 ประเภท คือ (1) Missing completely at random (MCAR) เป็นลักษณะของข้อมูลสูญหายที่เกิดขึ้นอย่างสุ่มจากค่าสังเกตทั้งหมด (2) Missing at random (MAR) เป็นลักษณะของข้อมูลสูญหายที่ไม่ได้เกิดขึ้นอย่างสุ่มจากค่าสังเกตทั้งหมด แต่เกิดขึ้นอย่างสุ่มภายในบางส่วนหรือบางกลุ่มของค่าสังเกต และสุดท้ายคือ (3) Missing Not at Random (MNAR) เป็นลักษณะของข้อมูลสูญหายซึ่งไม่ได้เกิดขึ้นอย่างสุ่ม

ในปี 2012 Han กับ Kamber¹ ได้แบ่งวิธีการจัดการกับข้อมูลที่มีการสูญหายไว้ด้วยกัน 2 วิธีดังนี้

1.1.1. การลบข้อมูลทั้งหมดที่มีสูญหายออก (Ignoring) วิธีนี้จะเป็นการลบข้อมูลออกทั้งแถว หากข้อมูลในแถวนั้นมีการสูญหายเกิดขึ้น ซึ่งจำทำให้อาจมีการตัดข้อมูลสำคัญหรือส่งผลอย่างมากต่อการทำเหมืองข้อมูลออก ทำให้ความน่าเชื่อถือทางสถิติลดลง

1.1.2. การแทนที่ข้อมูล (Missing Value Imputation) เป็นการเติมข้อมูลที่สูญหายจากการหาความสัมพันธ์ของข้อมูลที่มีอยู่ และพยากรณ์ว่าข้อมูลที่หายไปนั้นควรมีค่าเป็นเท่าใด โดยแบ่งออกได้เป็น วิธีการทางสถิติและวิธีการการเรียนรู้ของเครื่องดังนี้

1.1.2.1 วิธีการทางสถิติ (Statistical Approach) เป็นการประยุกต์ใช้หลักการทางด้านสถิติเพื่อแทนที่ข้อมูลสูญหาย ซึ่งมีอยู่ด้วยกันหลายวิธีการ อาทิเช่น วิธีการแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ย (Mean Imputation) วิธีการแทนที่ข้อมูลสูญหายด้วยฐานนิยม (Mode Imputation) และการแทนที่ข้อมูลด้วยวิธีการแบบถดถอย (Regression Imputation)

1.1.2.2 วิธีการการเรียนรู้ของเครื่อง (Machine Learning Approach) เป็นการใช้ประโยชน์ของเทคนิคการทำเหมืองข้อมูลเรื่องการหาความสัมพันธ์บางส่วน เพื่อมาแก้ไขปัญหาเรื่องข้อมูลสูญหาย เช่น การจำแนกประเภท (Classification) การแบ่งกลุ่ม (Clustering) โดยวิธีที่อยู่บนพื้นฐานของการจำแนกประเภทประกอบไปด้วย วิธีการแทนที่ข้อมูลด้วยการค้นหาเพื่อนบ้านใกล้สุด k ตัว (K-Nearest Neighbors Imputation) วิธีการแทนที่ข้อมูลด้วยการวิเคราะห์การถดถอยโลจิสติก (Logistic Regression Imputation) วิเคราะห์จำแนกประเภทข้อมูลด้วยเครื่องเวกเตอร์ค้ำยัน (Support Imputation) ในส่วนของการแทนที่ข้อมูลด้วยการแบ่งกลุ่มจะหลายวิธีการด้วยกัน เช่น วิธีการแบ่งกลุ่มข้อมูลแบบเคมีน (KMI) และ การวิเคราะห์กลุ่มแบบลำดับขั้น (Hierarchical Clustering) เป็นต้น

อีกหนึ่งงานวิจัยที่ผู้วิจัยให้ความสนใจคือ งานวิจัยของ K. Chingnong ในปี 2022¹⁰ ซึ่งเป็นการปรับปรุงประสิทธิภาพความแม่นยำการจำแนกข้อมูล (Classification) โดยใช้อัลกอริทึมอาณานิคมผึ้งเทียมเป็นพื้นฐาน เพื่อใช้ในการจำแนกข้อมูลกับชุดข้อมูลประเภทหมวดหมู่ (category data) ซึ่งแตกต่างเป้าหมายของงานวิจัยนี้ที่สนใจชุดข้อมูลประเภทตัวเลข (numerical data)

ในงานวิจัยนี้ จะมุ่งสนใจในการปรับปรุงประสิทธิภาพความแม่นยำของแบบจำลองการทำนายข้อมูล โดยวิธีการแทนที่ข้อมูลสูญหายประเภทตัวเลข โดยได้เลือกชุดข้อมูลทดสอบจากคลังข้อมูล UCI ได้แก่ Wine , Glass identification, Indian Patient Liver และ Seeds ซึ่งเป็นชุดข้อมูลประเภทตัวเลข ที่มีการสูญหายของข้อมูลในลักษณะ MAR มาทดสอบ

1.2 วัตถุประสงค์งานวิจัย

1.2.1. เพื่อศึกษาวิธีการแทนที่ข้อมูลสูญหายภายในชุดข้อมูลที่เป็นตัวเลข (Numeric Attribute) ให้ใกล้เคียงกับข้อมูลต้นฉบับ

1.2.2. เพื่อศึกษาและประยุกต์ใช้วิธีเชิงพันธุกรรม (Genetic Algorithm, GA) เพื่อแทนที่ข้อมูลสูญหาย

1.2.3. เพื่อศึกษาแล้วปรับปรุงด้วยวิธีวิธีเชิงพันธุกรรม ร่วมกับการถดถอยเชิงเส้นพหุคูณ (Multiple Linear Regression)

บทที่ 2

วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

บทนี้จะกล่าวถึงองค์ประกอบที่นำมาใช้ในการปรับปรุงอัลกอริทึมใหม่ที่เรานำเสนอ ซึ่งประกอบด้วย วิธีการค้นหาเพื่อนบ้านใกล้ที่สุด k ตัว (K-nearest neighbors algorithm) ร่วมกับการวิเคราะห์การถดถอยเชิงเส้นตรง (Linear regression) และ วิธีเชิงพันธุกรรม โดยจะมีรายละเอียดดังต่อไปนี้

2.1 K-nearest neighbors algorithm (KNN)

เป็นส่วนหนึ่งในอัลกอริทึมการจัดกลุ่ม (Classification Algorithm) เพื่อหาค่ามาสมมติมาแทนค่าที่สูญหายไป ซึ่งวิธีนี้เหมาะกับข้อมูลประเภทตัวเลข หลักการทำงานของอัลกอริทึมนี้คือ ใช้ความคล้ายกันของข้อมูลระหว่างข้อมูลในการแบ่งกลุ่ม (ความคล้ายกันของข้อมูลในที่นี้คือ ระยะห่างแบบเวกเตอร์ระหว่างข้อมูล) และจัดสมาชิกจากความคล้ายกันของข้อมูลที่มีความห่างน้อยที่สุดจำนวน k ตัว² โดยมีขั้นตอนการทำงานดังนี้

2.1.1. กำหนด instance เป้าหมาย ซึ่งแทนจุดนี้ด้วย d_i กับ ค่า instance ที่สมบูรณ์อื่นๆ ซึ่งแทนจุดนี้ด้วย d_j จนครบทุกค่า

2.1.2 คำนวณหาความคล้ายคลึงกันระหว่าง d_i กับ d_j เพื่อให้ได้ค่าความคล้ายคลึงกัน d_{ij} โดยใช้สมการคำนวณความคล้ายกันดังนี้

$$d_{ij} = \text{dist}(d_i, d_j) = \sqrt{\sum_{k=1}^n (d_{ik} + d_{jk})^2} \quad (2.1)$$

2.1.3. จัดกลุ่มความสัมพันธ์ใกล้เคียงกัน D_i มากที่สุดจำนวน k ค่า ก็จะได้เพื่อนบ้านจำนวน k ตัวตามอัลกอริทึม

งานวิจัยของ Anil Jadhav, Dhanya Pramod & Krishnan Ramanathan ใน ปี 2019⁴ ได้มีการนำอัลกอริทึมการแทนที่ข้อมูลสูญหายชนิดตัวเลขหลายๆประเภทมาเปรียบเทียบประสิทธิภาพ โดยได้ทำการทดลองงานทำวิจัยไว้ดังนี้

ขั้นตอนที่ 1 เตรียมชุดข้อมูลจาก UCI Machine Learning Repository โดยชุดข้อมูลที่นำมาใช้คือชุดข้อมูล ไวน์(Wine Dataset) การแบ่งประเภทของหญ้า(Glass Identification) ความแข็งแรงทนทานของคอนกรีต (Concrete Comprehensive Strength) ขนาดของตับในผู้ป่วยโรคตับชาวอินเดีย (Indian Liver Patient) และ ลักษณะของเมล็ดพืช (Seeds Dataset)

ขั้นตอนที่ 2 ทำให้เกิดข้อมูลสูญหายในแต่ละชุดข้อมูลเป็นจำนวน 10% 20% 30% 40% และ 50% ของทั้งชุดข้อมูล

ขั้นตอนที่ 3 หลังจากนั้นจะนำชุดข้อมูลที่สร้างขึ้นมาไปทดสอบการแทนที่ข้อมูลด้วยวิธีการ ค่าเฉลี่ย(Mean), ค่ามัธยฐาน(Median) เพื่อนบ้านใกล้ที่สุด k ตัว(kNN) การจับคู่ค่าเฉลี่ยจากการทำนาย(Predictive Mean Matching) การวิเคราะห์ถดถอยเชิงเส้นของเบย์ (Bayesian Linear Regression) การวิเคราะห์ถดถอยเชิงเส้นโดยไม่เกี่ยวข้องกับเบย์(Linear Regression–Non-Bayesian) และ วิธีการแทนข้อมูลด้วยข้อมูลตัวอย่าง(Sample Imputation Method) โดยการแทนที่ข้อมูลแต่ละวิธีใช้อัลกอริทึมดังนี้

i. Mean และ Median ใช้การคำนวณค่าเฉลี่ยและค่ามัธยฐานของทั้งข้อมูลและนำไปเติมในข้อมูลสูญหาย

ii. kNN ใช้ VIM package ในภาษา R⁵

iii. Predictive Mean Matching, Bayesian Linear Regression, Linear Regression–Non-Bayesian และ Sample Imputation Method ใช้ mic package ในภาษา R ซึ่งถูกจัดทำโดยนักวิจัยเหล่านี้ Van Buuren and Groothuis-Oudshoorn ในปี 2011⁶, Sterne et al. ในปี 2009 Patric และ White ในปี 2011 White, Royston และ Wood ในปี 2011

ขั้นตอนที่ 4 วิเคราะห์ประสิทธิภาพของแต่ละวิธีวิธี มีด้วยกันหลากหลายวิธี อาทิเช่น ความถูกต้อง(accuracy) ความถูกต้องสัมพัทธ์(relative accuracy) ค่าเฉลี่ยความผิดพลาดกำลังสอง(mean absolute error) และ รากที่สองของค่าเฉลี่ยความผิดพลาดกำลังสอง(root mean square error)

ซึ่งในงานวิจัยได้เลือกใช้การวัดประสิทธิภาพของอัลกอริทึมด้วย การหาค่าเฉลี่ยของค่ารากที่สองของค่าเฉลี่ยความผิดพลาดกำลังสองแบบบรรทัดฐาน หรือ Mean of Normalize RMSE ซึ่งวิธีที่ได้รับความนิยมมากที่สุดในการเปรียบเทียบประสิทธิภาพการอัลกอริทึมในการแทนที่ข้อมูล⁷ โดยแสดงได้ดังสมการนี้

$$RMSE = \sqrt{\frac{(a_1 - y_1)^2 + \dots + (a_n - y_n)^2}{n}} \quad (2.2)$$

เมื่อกำหนดให้	a	คือ	ค่าของข้อมูลที่ถูกต้อง
	y	คือ	ค่าที่ได้จากการพยากรณ์
	n	คือ	จำนวนข้อมูลทั้งหมดที่มีการแทนที่ข้อมูล

อีกเหตุผลที่ทำให้ผู้วิจัยเลือกการวัดประสิทธิภาพด้วยวิธีนี้คือ ข้อมูลมีขนาดข้อมูลที่แตกต่างกัน ทำให้ต้องหาค่าเฉลี่ย เพื่อประเมินประสิทธิภาพของอัลกอริทึมออกมาได้

2.2 Predictive Mean Matching (PMM)

Predictive Mean matching (PMM) เป็นวิธีการแทนที่ข้อมูลที่เหมาะสมกับการประมาณค่าสูญหายจากชุดข้อมูลที่ไม่ได้มีการกระจายตัวแบบปกติ ข้อดีของ PMM คือ ค่าที่ได้จากการพยากรณ์จะมีความใกล้เคียงกับค่าจริง เนื่องจากผลลัพธ์ที่ได้มาจากค่าจริงที่มีที่ไม่สูญหายอื่นๆ ตัวอย่างเช่น หากชุดข้อมูลมีลักษณะข้อมูลเบ้ขวา ผลที่ได้จากการทำ PMM จะออกมาเป็นข้อมูลเบ้ขวาเหมือนกัน หรือหากชุดข้อมูลอยู่ในช่วง 0-100 ผลที่ได้จากค่าประมาณก็จะอยู่ในช่วง 0-100 เช่นเดียวกัน โดยหลักการการทำงานของ PMM มีดังนี้

Gerko Vink ⁹ ได้ทำงานวิจัยหัวข้อ Predictive mean matching imputation of semicontinuous variables ซึ่งงานวิจัยนี้เป็นการศึกษาประสิทธิภาพของ PMM เมื่อเปรียบเทียบกับอัลกอริทึมอื่นๆ โดยผู้วิจัยได้ทำการทดลองเพื่อหาประสิทธิภาพของอัลกอริทึมนี้ว่าสามารถทำการแทนที่ข้อมูลสูญหายได้ดีเท่าใด เมื่อเทียบกับอัลกอริทึมอื่นๆ โดยทำการเปรียบเทียบกับ PMM กับ TEMPS, 2-Part, MI, IRMI, BGLon, และชุดข้อมูลที่เป็นกึ่งต่อเนื่อง โดยชุดข้อมูล 2 ชุด ประกอบด้วย 1. ข้อมูลจากในโลกสังคมออนไลน์ (social statistics [The Hague Twitter Scene (HTS) data]) และ 2 จาก official statistics (Dutch Wholesalers Statistics 2008) โดยทั้งสองชุดข้อมูลเป็น complete dataset ทางผู้ทดลองจึงได้จำลองข้อมูลสูญหายให้กับชุดข้อมูลชนิด MAR จำนวน 50% ให้กับชุดข้อมูล จากผลการทดลองในการแทนที่ข้อมูลของอัลกอริทึม PMM ให้ผลการทดลองที่มีความเอนเอียงน้อยกว่าทุกอัลกอริทึมที่ใช้ในการทดสอบของงานวิจัยนี้ และประโยชน์อีกหนึ่งสิ่งของ PMM คือ สามารถคงรูปแบบและความสัมพันธ์ของข้อมูลสูญหายประเภท MCAR และ MAR

2.3 Linear Regression

งานวิจัยของ Samih M. Mostafa ได้ตีพิมพ์งานวิจัยในชื่อ Imputing Missing Values Using Cumulative Linear Regression¹³ โดยในงานวิจัยนี้ได้นำเสนอ Cumulative Linear Regression เพื่อใช้ในการแทนที่ข้อมูลสูญหาย และได้ทำการทดสอบการแทนที่ข้อมูลสูญหาย โดยเปรียบเทียบประสิทธิภาพกับอัลกอริทึมดังต่อไปนี้ ampute, mice, ForImp, missForest, impute_lm, regressionImp, IterativeImputer, KNN และ SoftImpute และในการทดลองนี้ได้ใช้ชุดข้อมูล 5 ชุดข้อมูล diabetes, graduate admissions, profit estimation of companies, red & white wine dataset, California, และ diamonds ดังตารางที่ 2.1 ซึ่งในแต่ละข้อมูลจะมีการสร้างลักษณะของข้อมูลสูญหายที่ต่างกันไปและปริมาณข้อมูลสูญหายที่ต่างกันไปดังต่อไปนี้ โดยในแต่ละชุดข้อมูลได้สร้างข้อมูลสูญหายแบ่งออกเป็น 3 ประเภท ได้แก่ MCR MCAR และ MNAR โดยในแต่ละประเภทของ missing value จะมีอัตราส่วนการสูญหายคือ 5 10 15 20 และ 25 เปอร์เซ็นต์

ในการทดลองนี้ได้เปรียบเทียบประสิทธิภาพของ Cumulative Linear Regression กับอัลกอริทึมข้างต้น โดยใช้วิธีการ เวลาในการแทนที่ข้อมูลสูญหาย(imputation time) RMSE MAE และ coefficient of determination [inline-formula] จากการสำรวจพบว่าประสิทธิภาพของอัลกอริทึมขึ้นกับขนาดของชุดข้อมูลและปริมาณการสูญหายของข้อมูล และประเภทข้อมูลสูญหาย จากการทดลองพบว่าประสิทธิภาพของ c ให้ผลที่ดีกว่าเล็กน้อย

ตารางที่ 2.1

ตารางแสดงชุดข้อมูลที่น่ามาทดสอบ Cumulative Linear Regression

ชื่อชุดข้อมูล	ข้อมูลตัวอย่าง	ข้อมูลลักษณะ
diabetes	424	11
graduate admissions	500	8
profit estimation of companies	1,000	6
red & white wine dataset	4,898	12
California	20,640	9
diamonds	53,940	10

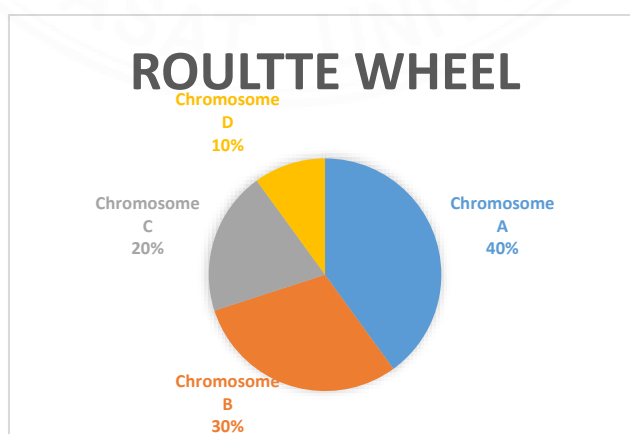
2.4 Genetic Algorithm

วิธีเชิงพันธุกรรม (Genetics Algorithm หรือ GA) เป็นวิธีหาคำตอบด้วยการคัดเลือกทางธรรมชาติ โดยจำลองมาจากทฤษฎีวิวัฒนาการโดยการคัดเลือกตามธรรมชาติของ ชาร์ลส์ ดาร์วิน กระบวนการทำงานของอัลกอริทึมนี้ มีด้วย 5 ขั้นตอนคือ การสร้างต้นแบบ (Initial population) ขยายจำนวนประชากร (Scale population) สกวนไว้เฉพาะผู้ที่เหมาะสม (Selection) สร้างลูกผสม (Crossover) และกลายพันธุ์ (Mutation) โดยจุดที่น่าสนใจของอัลกอริทึมมีรายมี 3 ขั้นตอนที่ทำให้มีความแตกต่างจากการสุ่มใส่ค่าแบบทั่วไปคือ สกวนไว้เฉพาะผู้ที่เหมาะสม (Selection) สร้างลูกผสม (Crossover) และกลายพันธุ์ (Mutation) ซึ่งมีรายละเอียดดังนี้

2.4.1 การคัดเลือกผู้อยู่รอด (Selection) จะเป็นการรักษาประชากรไว้เฉพาะประชากรที่มีค่าความเหมาะสมในการอยู่รอด (fitness) ไว้ ซึ่งจะเก็บเฉพาะกลุ่มของผู้ที่มีค่าสูงสุดไว้ และกลุ่มที่มีค่าความเหมาะสมน้อยจะถูกตัดออก เปรียบเหมือนการสุญพันธุ์ไปและเก็บไว้เฉพาะผู้อยู่รอด

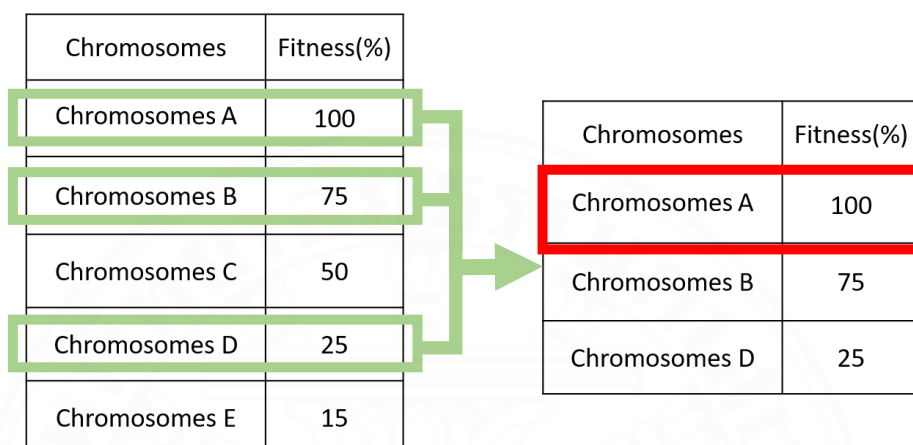
2.4.2 สร้างลูกผสม (Crossover) เป็นการสร้างรุ่นต่อไปจากรุ่นปัจจุบัน หรือ หรือรุ่นลูก ซึ่งวิธีการเลือกพ่อ-แม่ที่นิยมกันมีด้วย 2 วิธีคือ Roulette Wheel และ Tournament Selection

2.4.2.1 Roulette Wheel จะมีลักษณะคล้ายกับวงล้อ Roulette ที่ใช้ในคาสิโน โดยจะกำหนดความกว้างของแต่ละช่อง Roulette ด้วยค่าความเหมาะสมที่จะอยู่รอด (fitness) จึงทำให้ chromosome ที่มีค่า fitness สูงจะมีโอกาสการถูกเลือกให้อยู่รอดสูงตามไปด้วย ซึ่งแสดงให้เห็นตัวอย่างในภาพที่ 2.1



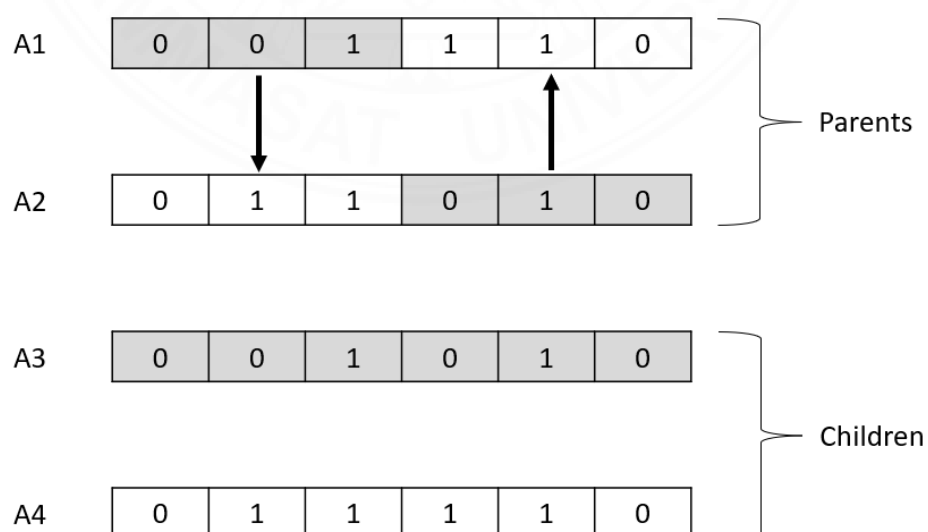
ภาพที่ 2.2 ภาพแสดงการเลือกพ่อแม่ด้วยวิธี Roulette Wheel

2.4.2.2 Tournament Selection วิธีนี้จะเป็นการคัดสรรประชากรให้อยู่รอดโดยทำการสุ่มเลือก chromosomes ที่มีอยู่มา เช่น หากสุ่มเลือกมา 3 chromosomes เลือก chromosome ที่มีค่า fitness สูงที่สุดมาเป็นพ่อ-แม่สำหรับรุ่นต่อไป ซึ่งได้ทำการแสดงตัวอย่างไว้ในภาพที่ 2.2



ภาพที่ 2.2 ภาพแสดงการเลือกพ่อแม่ด้วยวิธี Tournament Selection

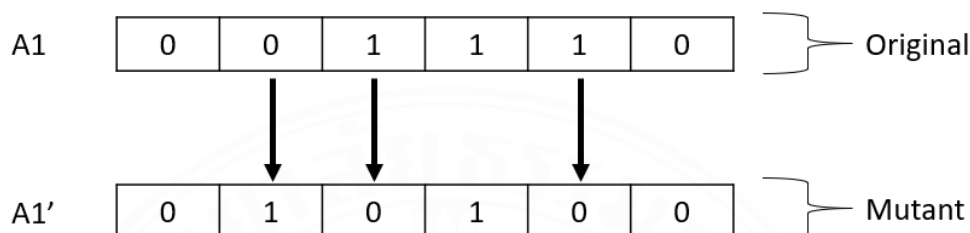
กลุ่มผู้อยู่รอดจากขั้นตอน Selection โดยอาจจะใช้การสุ่มเลือกประชากรจำนวน 2 หน่วย เพื่อนำมาเป็นพ่อ-แม่ของรุ่นถัดไป ซึ่งจะมีขั้นตอนคือ การนำ chromosomes บางเส้นของพ่อ-แม่มาสลับกัน ดังรูปด้านล่างนี้



ภาพที่ 2.3 ภาพแสดงการสร้างลูกผสม หรือ Crossover

จากภาพจะเห็นได้ว่า A1 และ A2 คือพ่อแม่ ซึ่งมีการแบ่งสาย chromosome ไว้เป็น 2 ส่วน หลังจากนั้นจะทำการผสม 2 ส่วนของ chromosomes เพื่อสร้างรุ่นลูกออกมาเป็น A3 และ A4

2.4.3 กลายพันธุ์ (Mutation) จะเกิดขึ้นหลังจากขั้นตอน Crossover แล้ว หรือคือกลายพันธุ์หลังการผสมให้เกิดรุ่นลูก ซึ่งจะเกิดขึ้นกับยีนส์บางตัวของรุ่นลูกดังภาพต่อไปนี้



ภาพที่ 2.4 ภาพแสดงการกลายพันธุ์ หรือ Mutation

จากขั้นตอนของวิธีเชิงพันธุกรรมที่กล่าวมานั้น จะเห็นได้ว่าจุดสำคัญที่ทำให้อัลกอริทึมนี้แตกต่างจากการสุ่มทั่วไปคือ จะมีการเก็บค่าของมูลที่มีค่า fitness สูงไว้เพื่อนำมาเป็นข้อมูลในการสุ่มครั้งถัดไป แต่จะยังมีการประชากรที่มีค่าสูงไว้กลุ่มหนึ่งไว้เป็นพ่อแม่ของรุ่นถัดไป เพื่อว่าในรุ่นถัดไปจะยังมีการสุ่มเลือกพ่อแม่ ไม่ใช่การเลือกจากประชากรที่มีค่าสูงสุด 2 ค่าเท่านั้น เพื่อป้องกันการเกิด local optimal solution เพราะ ประชากรที่มีค่าสูงสุดในรุ่นๆหนึ่ง อาจจะไม่นำไปสู่คำตอบที่ดีที่สุดก็เป็นได้

2.5 Multiple Linear Regression

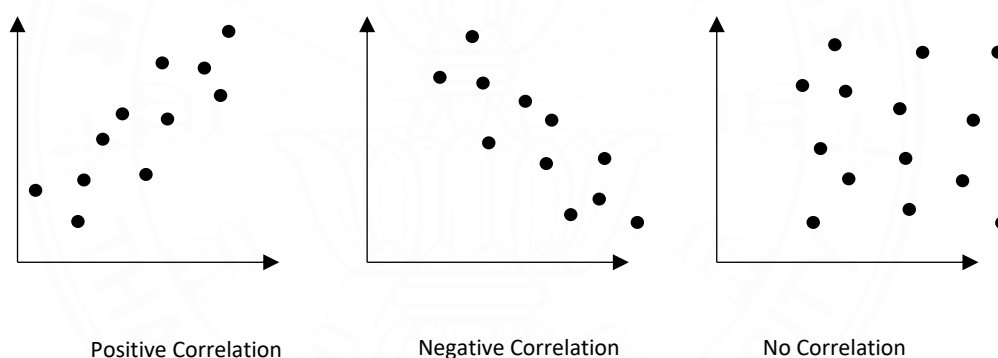
สมการถดถอยเชิงเส้นพหุคูณ หรือ Multiple Linear Regression (MLR) เป็นวิธีหาคำตอบด้วยสมการซึ่งมีพื้นฐานมาจากการวิเคราะห์การถดถอยเชิงเส้นแบบดั้งเดิม (Linear Regression) ที่ใช้การหาค่าจากความสัมพันธ์ระหว่าง 1 ตัวแปรต้นและ 1 ตัวแปรตาม แต่สำหรับ MLR นั้นจะเป็นการหาความสัมพันธ์จากหลายตัวแปรต้น เพื่อหาค่าของ 1 ตัวแปรตาม

ในปี 2013 ได้มีการศึกษาการวิเคราะห์การถดถอย หรือ Regression Analysis¹¹ เป็นวิธีการทางสถิติเพื่อประมาณค่า โดยการใช้ความสัมพันธ์ระหว่างหลายตัวแปรต้น (X) และ 1 ตัวแปรตาม (y) ซึ่งสามารถเขียนสมการออกมาได้ดังนี้

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon \quad (2.3)$$

กำหนดให้	i	แทน แถวที่ต้องการหาค่า
	y	แทน ค่าตัวแปรตามที่ได้จากสมการในแถวที่ i
	p	แทน จำนวนตัวแปรต้นในแถวที่ i
	x_{ip}	แทน ตัวแปรต้นที่ p ในแถว i
	β_0	แทน จุดกราฟตัดแกน y (หากค่า β อื่นปกติ β_0 จะเท่ากับ 0)
	β_p	แทน ค่าความชันระหว่าง y_i กับ x_{ip}
	ϵ	แทน ค่าความต่างระหว่างค่า y จริง กับ ค่า y ที่ได้จากสมการ

จากสมการข้างต้นทำให้เห็นข้อดีของ multiple linear regression ว่าสามารถหาค่า y ได้โดยมีการใช้ความสัมพันธ์ของของแต่ละตัวแปรผ่านค่า β ซึ่งค่า β มีค่าอยู่ในช่วง -1 ถึง 1 ดังภาพด้านล่างนี้



ภาพที่ 2.5 ภาพแสดงค่าสัมประสิทธิ์ β ของสมการ Multiple Linear Regression

จากภาพใน จะเป็นการแสดงค่าความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตาม (β) หากมีการกระจายตัวปกติปกติ ค่า Positive Correlation หรือความสัมพันธ์เชิงบวกจะมีค่าอยู่ในช่วงมากกว่า 0 ถึง 1 และ ค่า Positive Correlation หรือความสัมพันธ์เชิงลบจะมีค่าอยู่ในช่วงน้อยกว่า 0 ถึง -1 แต่หากการกระจายตัวไม่ปกติ ทำให้ไม่สามารถหาความสัมพันธ์ของกราฟได้ ทำให้ค่า β เป็น 0 และตัวแปร x ก็จะไม่ส่งผลต่อค่าของ y หรือหายไปนั่นเอง

2.6 Clusterwise Linear Regression

โดยในปี 2022 นักวิจัยที่ชื่อ Napsu Karmita¹² ได้นำเสนอ Clusterwise Linear Regression มีขั้นตอนการทำงานที่ผสมระหว่าง การแบ่งกลุ่ม (Clustering) และ สมการถดถอยเชิงเส้น (Linear Regression) ซึ่งมีการนำ CLR (Clusterwise Linear Regression) มาใช้ในการแทนที่ข้อมูลสูญหาย โดยตั้งชื่ออัลกอริทึมนี้ว่า IviaCLR (Imputation via Clusterwise Linear Regression) ซึ่งมีกระบวนการทำงานอยู่ 3 ขั้นตอนคือ Initial Imputation, CLR Method และ Prediction

Initial Imputation เป็นการเติมข้อมูลสูญหายด้วยวิธีต่างๆก่อนที่จะนำไปสู่ขั้นตอนถัด ซึ่งในการทดลองจะมี 3 วิธีที่นำมาใช้ในการเติมข้อมูลสูญหาย คือ Mean Imputation, Linear Regression Imputation และ Recursive Regression Imputation เพื่อทำให้เกิดชุดข้อมูลที่สมบูรณ์ก่อนที่จะนำไปใช้กับ CLR แบบดั้งเดิม

CLR Method เป็นวิธีการแทนที่ข้อมูลสูญหายโดยการหาข้อมูลที่มีความใกล้เคียงกับแถวของข้อมูลที่มีความสนใจด้วยวิธี kNN เพื่อหาว่าข้อมูลใดมีความใกล้เคียงกับข้อมูลที่สนใจโดยจะค้นหาจากชุดข้อมูลที่ได้จากขั้นตอน Initial Imputation จากนั้นจะใช้ Linear Regression จากข้อมูลที่ได้จาก kNN เพื่อหาค่าเพื่อใช้ในการแทนที่ข้อมูลสูญหาย

Prediction หรือการทำนายค่า จะเกิดขึ้นเมื่อได้ค่าใหม่จาก CLR ที่จะนำมาแทนที่ในข้อมูลที่เกิดข้อมูลสูญหาย แต่ก่อนที่จะแทนที่ข้อมูลสูญหายทันที IviaCLR จะทำการเปรียบเทียบก่อนว่าค่าที่ได้จาก CLR นั้นเมื่อเทียบกับข้อมูลที่มีอยู่เดิมจากขั้นตอน Initial Imputation มีค่า ϵ จากสมการ Regression เท่าใด หากน้อยกว่าจึงจะทำการแทนที่ข้อมูลเดิม หากมากกว่าจะไม่ทำการแทนที่ข้อมูล

บทที่ 3

วิธีการวิจัย

3.1 ภาพรวมระบบ

การสูญหายของข้อมูลในชุดข้อมูล (Missing Value) สำหรับการทำเหมืองข้อมูล (Data Mining) ทำให้ความแม่นยำของผลลัพธ์และความน่าเชื่อถือทางสถิติของการทำเหมืองของข้อมูลลดลง เราจึงแก้ไขปัญหาเหล่านี้ด้วยการแทนที่ข้อมูลสูญหาย (Imputation) โดยประยุกต์ใช้วิธีการ Combination of K-Nearest Neighbors and Linear Regression (KNNLR) เพื่อแทนที่ข้อมูลสูญหายให้ใกล้เคียงกับข้อมูลต้นฉบับมากที่สุด โดยภาพรวมของการแทนที่ข้อมูลสูญหายเป็นไปตามรูปที่ 3.1 ดังนี้



ภาพที่ 3.1 ภาพแสดง Genetic Algorithm and Multiple Linear Regression Algorithm

ขั้นตอนแรกนำเข้าข้อมูลที่เกิดการสูญหายของข้อมูล โดยจะเรียกข้อมูลต่อไปนี้ว่า Incomplete dataset จากนั้นข้อมูลข้างต้นไปผ่านกระบวนการแทนที่ข้อมูลสูญหายด้วยอัลกอริทึมที่เรานำเสนอในงานวิจัยนี้ โดยอัลกอริทึมดังกล่าวจะทำหน้าที่หาคำตอบสำหรับทุกข้อมูลสูญหายที่เกิดขึ้นภายใน Incomplete Dataset ซึ่งหลังจากได้ชุดคำตอบของข้อมูลสูญหายมาแล้ว จะนำคำตอบที่ได้ไปแทนที่ข้อมูลสูญหายใน Incomplete Dataset หลังจากจบกระบวนการข้างต้นแล้ว เราจะได้ชุดข้อมูลที่สมบูรณ์ (Complete Dataset) เพื่อนำไปทดสอบความแม่นยำของ prediction model และนำไปเปรียบเทียบกับผลด้านความแม่นยำของการแทนที่ข้อมูลสูญหายด้วยวิธีการอื่น

3.2 ตารางแสดงชุดข้อมูลที่นำมาทดสอบ Cumulative Linear Regression

ในขั้นตอนการเตรียมข้อมูล จะทำการนำข้อมูลที่ต้องการทดสอบมาแบ่งออกเป็น 2 ชุด คือ ชุดข้อมูลที่มีข้อมูลสูญหาย และ ชุดข้อมูลทดสอบ ซึ่งมีลักษณะข้อมูลดังตารางที่ 3.1

ตารางที่ 3.1

ตารางแสดงชุดข้อมูล UCI ที่มาใช้ในการทดสอบ

ชื่อฐานชุดข้อมูล	จำนวนข้อมูลในชุด (Instances)	จำนวนคุณลักษณะ (Attributes / features)
Wine	178	13
Glass Identification	214	10
Indian Liver Patient	583	10
Seeds	210	7


ชุดข้อมูลที่มีข้อมูลสูญหาย จะถูกแบ่งระดับการสูญหายของข้อมูลเป็น 5 ระดับ โดยเริ่มจาก 5 10 15 20 และ 25 เปอร์เซ็นต์

ชุดข้อมูลทดสอบ มีไว้สำหรับทดสอบประสิทธิภาพโมเดลพยากรณ์ เพื่อคำนวณค่าความเหมาะสม หรือ fitness ในเทียบประสิทธิภาพของอัลกอริทึมต่างๆ

3.3 Genetic Algorithm and Multiple Linear Regression (GAMLR)

ขั้นแรกของ GAMLR จะเริ่มต้นเหมือน Genetic Algorithm ซึ่งเป็นการสร้างประชากรชุดเริ่มต้น โดยการสุ่มค่าจาก min-max ของข้อมูลที่มีอยู่ในคอลัมน์ที่สนใจ และทำการเติมลงในชุดข้อมูลสูญหาย โดยทุกข้อมูลสูญหายจะทำการสุ่มใหม่ทุกครั้ง (ไม่ใช่การสุ่มครั้งเดียวและเติมลงทุกข้อมูลสูญหายในคอลัมน์) ตัวอย่างการสุ่มของคอลัมน์ที่ 1 (col#1) จะเป็นจะสุ่มในช่วง 9 ถึง 13 และคอลัมน์ที่ 2 (col#2) จะสุ่มในช่วง 1.9 ถึง 2.8 ซึ่งจะแสดงในภาพที่ 3.1

Class	Col#1	Col#2
1	11	1.9
1	?	2.0
1	13	?
1	?	2.4
2	?	2.8
2	9	2.7

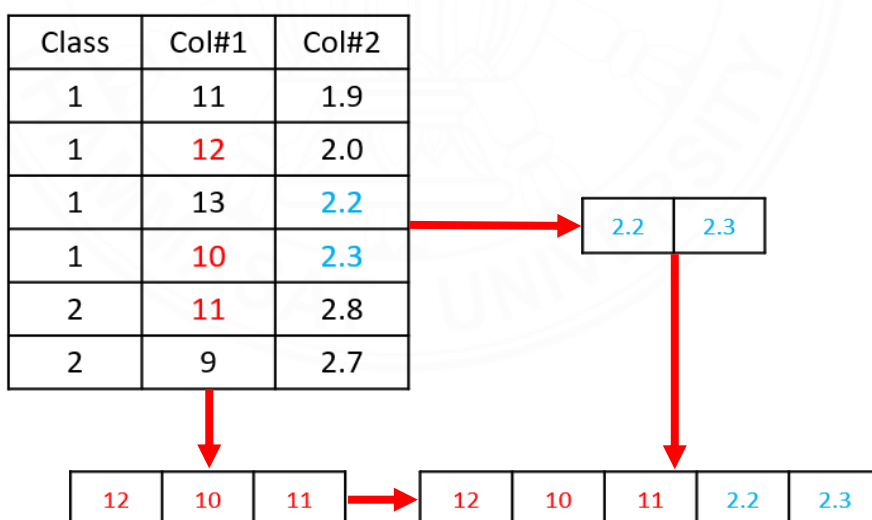


Class	Col#1	Col#2
1	11	1.9
1	12	2.0
1	13	2.2
1	10	2.3
2	11	2.8
2	9	2.7

ภาพที่ 3.2 ภาพแสดงการแทนที่ข้อมูลสูญหายจากค่า min-max

จะเห็นได้ว่า แต่ละค่าในคอลัมน์จะต่างกัน เนื่องจากการสุ่มใหม่ทุกครั้งที่เจอค่าของข้อมูลสูญหาย โดยคิดจากค่า min-max ในคอลัมน์นั้น

หลังจากที่ทำการเติมข้อมูลสูญหายจนครบแล้ว จะทำการเก็บข้อมูลสูญหายที่เติมลงไปในรูปแบบของ chromosome ซึ่งโครงสร้างของ chromosome ของประชากรนี้จะมีลักษณะเป็นการนำค่าที่ถูกแทนที่ในข้อมูลสูญหายมาเรียงต่อกัน โดยจะเรียงเป็นชุดของแต่ละคอลัมน์ และจากนั้นนำแต่ละคอลัมน์มาต่อกันอีกครั้ง ซึ่งจะมีลักษณะตามภาพ 3.2



ภาพที่ 3.3 ภาพแสดงการสร้าง chromosome จากชุดข้อมูลสูญหาย

จากภาพที่ 3.2 จะเห็นว่า chromosome ถูกสร้างจากการนำของข้อมูลสูญหายในแต่ละช่องมาเรียงต่อกันเป็น vector ก่อน หลังจากนั้นจึงนำ vector นั้นมาต่อกันเป็นสาย

chromosome โดยที่ช่องที่ตัวเลขเป็นสีแดง จะแทนค่าข้อมูลสุญหายในคอลัมน์ที่ 1 และช่องที่มีตัวเลขเป็นสีฟ้า จะแทนค่าของข้อมูลสุญหายในคอลัมน์ที่ 2

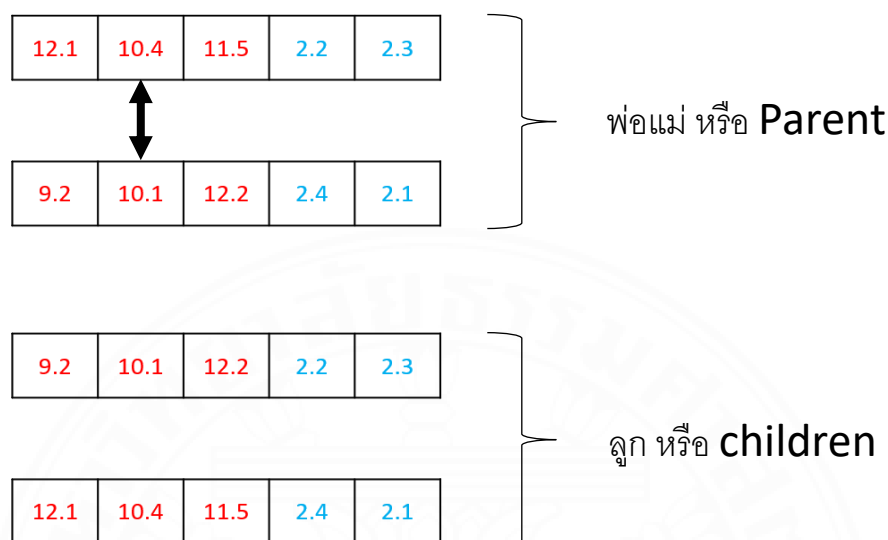
ขั้นตอนต่อมาเป็นการ คัดสรรประชากรที่อยู่รอด หรือ Selection ซึ่งจะเป็นการหาว่าประชากรคนใดที่เหมาะสมจะอยู่รอดบ้าง ซึ่งในขั้นตอนนี้จะใช้อัลกอริทึม Classification and Regression Trees หรือ CART ในการหาว่า chromosomes มีค่า fitness เท่าใด โดยคิดจากเปอร์เซ็นต์ความถูกต้องในการทำนายข้อมูลว่าได้คำตอบของ class ถูกต้องหรือไม่ และเก็บ chromosomes พร้อมกับค่า fitness ไว้ในตารางประชากร โดยเก็บแบบเรียงลำดับค่า fitness จากมากไปน้อย ดังภาพที่ 3.4 ด้านล่างนี้

Chromosomes					Fitness(%)
12.1	10.4	11.5	2.2	2.3	100
9.2	10.1	12.2	2.4	2.1	75
10.5	11.2	11.4	2.0	2.0	50
12.4	9.8	9.4	1.9	2.7	25

ภาพที่ 3.4 ภาพแสดงตารางการเก็บ chromosomes พร้อม fitness

ขั้นที่ตอน 3 การผสมข้ามสายพันธุ์ หรือ Crossover เป็นหนึ่งในกระบวนการของอัลกอริทึมเชิงพันธุกรรม ซึ่งในขั้นตอนนี้จะเป็นการผสม 2 chromosomes ที่เปรียบเป็นพ่อแม่ หรือ parents เข้าด้วยกัน เพื่อให้กำเนิดลูกรุ่นใหม่ ซึ่งการเลือกเลือกพ่อแม่นั้นจะใช้การสุ่มบนพื้นฐานของค่า fitness ซึ่งหากมีค่า fitness มากก็จะมีโอกาสถูกเลือกมาก โดยใช้วิธี Tournament Selection เมื่อเลือกมาได้แล้วจะทำการ crossover ซึ่งจะเลือกเฉพาะบางยีนส์ของทั้ง 2 chromosomes มา

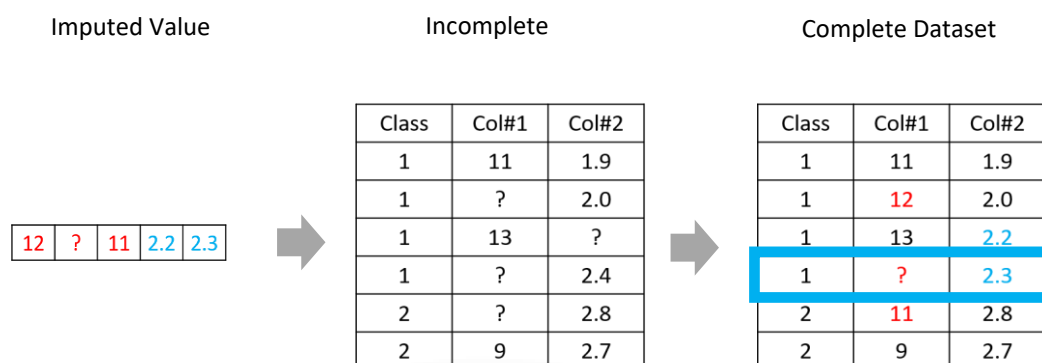
สลับกัน โดยในวิธีที่นำเสนอจะการใช้การสลับทั้ง vector ของ column แทนการสลับที่ละยีนส์ ซึ่งจะแสดงขั้นตอนดังภาพที่ 3.5



ภาพที่ 3.5 ภาพแสดงการ crossover ของรุ่นพ่อแม่เพื่อกำเนิดรุ่นลูก

จากภาพที่ 3.5 จะเป็นตัวอย่างการ crossover ที่ 50% โดยจะสุ่มเลือก ยีนส์ หรือในที่นี้คือ vector ของคอลัมน์ที่ 1 (สีแดง) และทำการสลับยีนส์ของพ่อแม่ เพื่อให้กำเนิดขึ้นมา 2 chromosomes ซึ่งจะถูกรับเรียกว่าลูกหรือรุ่นใหม่ (new generation) หลังจากได้รุ่นลูกมาแล้ว จะทำการนำไปเติมในชุดข้อมูลสุญหาย เพื่อให้ได้ชุดข้อมูลที่สมบูรณ์และนำไปสร้างโมเดลพยากรณ์ด้วย CART และทดสอบโมเดลเพื่อหาความแม่นยำของโมเดลพยากรณ์ที่ถูกสร้างด้วยข้อมูลสุญหายนี้ หรือค่า fitness นั้นเอง จากนั้นจะทำการเก็บ chromosomes รุ่นใหม่และค่า fitness ลงในตารางประชากร

ขั้นตอนที่ 4 การกลายพันธุ์หรือ Mutation กระบวนการนี้จะเกิดกับรุ่นลูกที่เกิดจากขั้นตอนการ crossover ซึ่งตามหลักการเชิงพันธุศาสตร์จะกลายพันธุ์เกิดขึ้นเมื่อทำการผสมของโครโมโซมจากรุ่นพ่อแม่สู่รุ่นลูก ในกระบวนการนี้จะเลียนแบบวิธิต่างพันธุกรรมในธรรมชาติ โดยจะสุ่มให้ยีนส์บางยีนส์เกิดการกลายพันธุ์ แต่การเลือกยีนส์ของ mutation จะแตกต่างจากขั้นตอน crossover ตรงที่จะมอง 1 ช่องในสายโครโมโซมเป็น 1 ยีนส์เลย และในการกลายพันธุ์นี้จะใช้ Multiple Linear Regression (MLR ที่กล่าวไปในบทบทวนวรรณกรรมข้อ 2.5) ในหาค่าเพื่อมาเป็นค่าการกลายพันธุ์และใส่ลงไปแทนค่าข้อมูลเดิม ดังตัวอย่างในภาพที่ 3.6 นี้



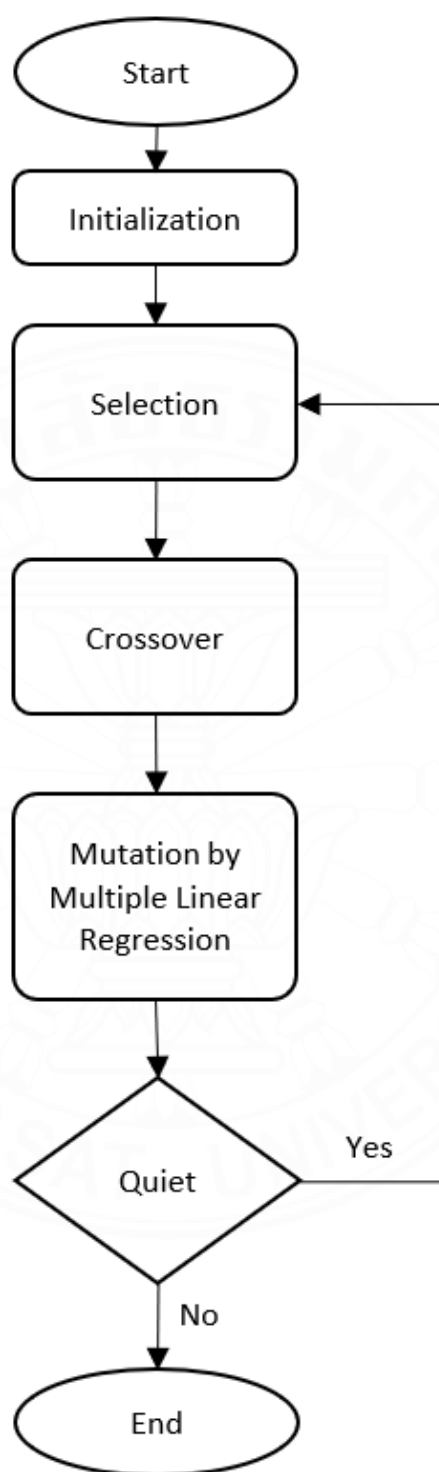
ภาพที่ 3.6 ภาพแสดงการสร้างชุดข้อมูลที่สมบูรณ์เพื่อใช้ในสมการ Multiple Linear Regression

จากภาพที่ 3.6 จะเป็นตัวอย่างการกลายพันธุ์ที่ช่องที่มีเครื่องหมาย ? โดยจะนำทั้งโครโมโซมไปเติมในชุดข้อมูลที่มีข้อมูลสูญหาย (Incomplete Dataset) เพื่อให้เกิดเป็นชุดข้อมูลที่สมบูรณ์ (Complete Dataset) จากนั้นจะใช้ชุดข้อมูลสมบูรณ์ทั้งหมด ยกเว้น แถวที่มีข้อมูลที่ต้องการให้เกิดกลายพันธุ์ (แถวที่มีเครื่องหมาย ?) เพื่อใช้ในการสร้างค่าสัมประสิทธิ์ของสมการ MLR และเมื่อได้ค่าสัมประสิทธิ์แล้ว จะนำข้อมูลในแถวที่ต้องการหาค่าไปแทนลงในสมการ MLR เพื่อหาค่าคำตอบของสมการ และนำมาแทนที่ข้อมูลที่เกิดการกลายพันธุ์ เมื่อการเกิดกลายพันธุ์ครบทุกค่าที่ต้องการแล้ว จะนำไปหาค่า fitness และเก็บข้อมูลสูญหายให้อยู่ในรูปแบบของโครโมโซม เพื่อนำไปเก็บไว้ในตารางประชากรต่อไป

ขั้นตอนที่ 5 จะเป็นการเช็คเงื่อนไขว่าจะต้องการออกจากการกระบวนการเชิงพันธุกรรมหรือไม่ โดยเงื่อนไขนี้จะถูกตั้งโดยใช้รอบในการเกิดกระบวนการเชิงพันธุกรรม หากยังไม่ครบ จะวนกลับไปที่กระบวนการ Selection อีกครั้ง หากครบแล้ว จะออกจากกระบวนการนี้

ขั้นตอนที่ 6 ขั้นตอนนี้จะเป็นการเลือกข้อมูลสูญหายที่มีประสิทธิภาพสูงสุด เพื่อนำไปใช้ในการเปรียบเทียบกับอัลกอริทึมแทนที่ข้อมูลสูญหายอื่น โดยวิธีการเลือกนั้นจะใช้ chromosomes ที่มีค่า fitness สูงสุดจากตารางประชากร เพื่อเป็นผลการแทนที่ข้อมูลสูญหายจากอัลกอริทึม GAML

จากขั้นตอนทั้งหมดที่ได้กล่าวมานั้น วิธีใหม่ที่ผู้วิจัยนำเสนอสามารถแสดงขั้นตอนการทำงานทั้งหมดออกมาเป็นแผนภาพได้ดังรูปที่ 3.3



ภาพที่ 3.7 แผนภาพแสดงการทำของ Genetic Algorithm and Multiple Linear Regression Algorithm

บทที่ 4

ผลการวิจัยและอภิปรายผล

4.1 การทดลอง

ในการทดลองนี้ได้แบ่งการทดลองเปรียบเทียบเชิงคุณภาพความแม่นยำในการทำผลของของโมเดลพยากรณ์ด้วยชุดข้อมูลที่ถูกแทนที่ข้อมูลสูญหายด้วยวิธีการแทนที่ข้อมูลศูนย์หายที่มีอยู่แล้ว เช่น การแทนที่ข้อมูลศูนย์หายด้วยค่าเฉลี่ย (Mean) การแทนที่ข้อมูลศูนย์หายด้วยค่ามัธยฐาน (Median) การแทนที่ข้อมูลศูนย์หายด้วยการถดถอยเชิงพหุ (Predictive Mean Matching) การแทนที่ข้อมูลศูนย์หายด้วยเพื่อนบ้านใกล้ที่สุด k ตัว (K-Nearest Neighbor Algorithm) การแทนที่ข้อมูลสูญหายด้วยการจัดกลุ่มและสมการถดถอยเชิงเส้น (IviaCLR) โดยมีการใช้ความแม่นยำของโมเดลผลลัพธ์ที่เกิดจากชุดข้อมูลที่ถูกแทนที่สมบูรณ์แล้วเป็นเกณฑ์ในการประเมินประสิทธิภาพของอัลกอริทึมในการแทนที่ข้อมูลศูนย์หาย โดยโมเดลพยากรณ์ที่ถูกนำมาใช้นั้นคือ ต้นไม้ตัดสินใจด้วยการถดถอยแบบพหุ (Decision Tree Multi Regression) หลังจากได้ผลจากการทำนายของโมเดลพยากรณ์แล้ว จะนำผลที่ทำนายได้ไปทำการเปรียบเทียบความถูกต้องการโมเดลพยากรณ์ด้วยชุดข้อมูลสมบูรณ์ที่ถูกแทนข้อมูลสูญหาย ด้วยอัลกอริทึมที่กล่าวมาข้างต้นพ เทียบกับผลความถูกต้องที่ได้จากโมเดลพยากรณ์ที่ใช้ข้อมูลสมบูรณ์ที่ได้จากวิธีที่นำเสนอใหม่นี้

การทดลองบนชุดข้อมูลสุญหาย Wine

ตารางที่ 4.1

ตารางเปรียบเทียบค่าความถูกต้องของโมเดลที่ได้จากข้อมูล Wine

Imputation Algorithm	5%	10%	15%	20%	25%
Mean	43.16	42.74	42.25	46.78	35.75
Median	42.41	41.50	43.91	46.58	35.50
PMM	41.58	42.08	42.66	46.25	34.58
kNN(k=3)	41.16	43.41	43.00	47.50	35.75
IviaCLR	43.01	44.01	43.91	48.23	36.17
GAMLR	65.50	61.83	61.91	60.91	55.75

จากตารางที่ 4.1 แสดงการเปรียบเทียบค่าความถูกต้องของโมเดลที่ได้จากข้อมูล Wine ที่มีชนิดข้อมูลการสุญหายลักษณะ Missing at Random (MAR) ในปริมาณ 5% 10% 15% 20% และ 25% ระหว่าง 5 อัลกอริทึม ดังนี้ การแทนที่ข้อมูลสุญหายด้วยค่าเฉลี่ย (Mean) การแทนที่ข้อมูลสุญหายด้วยค่ามัธยฐาน (Median) การแทนที่ข้อมูลสุญหายด้วยการถดถอยเชิงพหุ (Predictive Mean Matching) การแทนที่ข้อมูลสุญหายด้วยเพื่อนบ้านใกล้เคียงที่สุด k ตัว (K-Nearest Neighbor Algorithm) การแทนที่ข้อมูลสุญหายด้วยการจัดกลุ่มและสมการถดถอยเชิงเส้น (IviaCLR) และวิธีการที่นำเสนอใหม่นี้ จากภาพข้อมูลในแนวนอน X แสดงถึงค่าเปอร์เซ็นต์ข้อมูลสุญหายตั้งแต่ 5% 10% 15% 20% และ 25% ข้อมูลในแนวนอน Y แสดงถึงค่าความถูกต้องในการทำนายผลของโมเดลทำนายข้อมูลที่ใช้ข้อมูลจากการแทนที่ข้อมูลด้วยวิธีการแทนที่ข้อมูลสุญหายดังนี้ค่าเฉลี่ย (Mean) ค่ามัธยฐาน (Median) การถดถอยเชิงพหุ (Predictive Mean Matching) เพื่อนบ้านใกล้เคียงที่สุด k ตัว (K-Nearest Neighbor Algorithm) และวิธีการที่นำเสนอใหม่ จากภาพที่ 4.1 นี้แสดงให้เห็นว่าวิธีการที่นำเสนอใหม่นี้มีค่าความแม่นยำโดยเฉลี่ยสูงกว่าทุกอัลกอริทึมที่นำมาเปรียบเทียบ

4.2.1. การทดลองบนชุดข้อมูลสูญหาย Glass

ตารางที่ 4.2

ตารางเปรียบเทียบค่าความถูกต้องของโมเดลที่ได้จากข้อมูล Glass

Imputation Algorithm	5%	10%	15%	20%	25%
Mean	40.75	42.74	42.25	46.58	35.75
Median	44.66	41.50	43.91	46.58	35.50
PMM	42.24	42.08	42.66	46.25	34.58
kNN(k=3)	41.16	43.41	43.00	47.50	35.75
IviaCLR	41.10	46.75	42.41	46.91	38.75
GAMLR	60.08	61.83	61.91	60.91	55.75

จากตารางที่ 4.2 แสดงการเปรียบเทียบค่าความถูกต้องของโมเดลที่ได้จากข้อมูล Glass ที่มีชนิดข้อมูลการสูญหายลักษณะ Missing at Random (MAR) ในปริมาณ 5% 10% 15% 20% และ 25% ระหว่าง 5 อัลกอริทึม ดังนี้ การแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ย (Mean) การแทนที่ข้อมูลสูญหายด้วยค่ามัธยฐาน (Median) การแทนที่ข้อมูลสูญหายด้วยการถดถอยเชิงพหุ (Predictive Mean Matching) การแทนที่ข้อมูลสูญหายด้วยเพื่อนบ้านใกล้เคียงที่สุด k ตัว (K-Nearest Neighbor Algorithm) การแทนที่ข้อมูลสูญหายด้วยการจัดกลุ่มและสมการถดถอยเชิงเส้น (IviaCLR) และวิธีการที่นำเสนอใหม่นี้ จากภาพข้อมูลในแนวนอน X แสดงถึงค่าเปอร์เซ็นต์ข้อมูลสูญหายตั้งแต่ 5% 10% 15% 20% และ 25% ข้อมูลในแนวนอน Y แสดงถึงค่าความถูกต้องในการทำนายผลของโมเดลทำนายข้อมูลที่ใช้ข้อมูลจากการแทนที่ข้อมูลด้วยวิธีการแทนที่ข้อมูลสูญหายดังนี้ค่าเฉลี่ย (Mean) ค่ามัธยฐาน (Median) การถดถอยเชิงพหุ (Predictive Mean Matching) เพื่อนบ้านใกล้เคียงที่สุด k ตัว (K-Nearest Neighbor Algorithm) และวิธีการที่นำเสนอใหม่ จากภาพที่ 4.2 นี้แสดงให้เห็นว่าวิธีที่นำเสนอใหม่นี้มีค่าความแม่นยำโดยเฉลี่ยสูงกว่าทุกอัลกอริทึมที่นำมาเปรียบเทียบ

4.2.2. การทดลองบนชุดข้อมูลผู้ป่วย Indian Patient Liver

ตารางที่ 4.3

ตารางเปรียบเทียบค่าความถูกต้องของโมเดลที่ได้จากข้อมูล Indian Patient Liver

Imputation Algorithm	5%	10%	15%	20%	25%
Mean	58.03	56.35	58.35	56.99	58.39
Median	57.85	56.10	58.96	58.17	58.50
PMM	58.32	56.67	58.75	57.96	58.32
kNN(k=3)	57.46	56.21	59.57	57.49	57.74
IviaCLR	58.39	58.92	58.60	59.42	58.53
GAMLR	72.82	73.32	71.96	73.60	73.64

จากตารางที่ 4.3 แสดงการเปรียบเทียบค่าความถูกต้องของโมเดลที่ได้จากข้อมูล Indian Patient Liver ที่มีชนิดข้อมูลการสูญเสียหายลักษณะ Missing at Random (MAR) ในปริมาณ 5% 10% 15% 20% และ 25% ระหว่าง 5 อัลกอริทึม ดังนี้ การแทนที่ข้อมูลสูญเสียหายด้วยค่าเฉลี่ย (Mean) การแทนที่ข้อมูลสูญเสียหายด้วยค่ามัธยฐาน (Median) การแทนที่ข้อมูลสูญเสียหายด้วยการถดถอยเชิงพหุ (Predictive Mean Matching) การแทนที่ข้อมูลสูญเสียหายด้วยเพื่อนบ้านใกล้เคียงที่สุด k ตัว (K-Nearest Neighbor Algorithm) การแทนที่ข้อมูลสูญเสียหายด้วยการจัดกลุ่มและสมการถดถอยเชิงเส้น (IviaCLR) และวิธีการที่นำเสนอใหม่นี้ จากภาพข้อมูลในแนวนอน X แสดงถึงค่าเปอร์เซ็นต์ข้อมูลสูญเสียหายตั้งแต่ 5% 10% 15% 20% และ 25% ข้อมูลในแนวแกน Y แสดงถึงค่าความถูกต้องในการทำนายผลของโมเดลทำนายข้อมูลที่ใช้ข้อมูลจากการแทนที่ข้อมูลด้วยวิธีการแทนที่ข้อมูลสูญเสียหาย ดังนี้ค่าเฉลี่ย (Mean) ค่ามัธยฐาน (Median) การถดถอยเชิงพหุ (Predictive Mean Matching) เพื่อนบ้านใกล้เคียงที่สุด k ตัว (K-Nearest Neighbor Algorithm) และวิธีการที่นำเสนอใหม่ จากภาพที่ 4.3 นี้แสดงให้เห็นว่าวิธีที่นำเสนอใหม่นี้มีค่าความแม่นยำโดยเฉลี่ยสูงกว่าทุกอัลกอริทึม ที่นำมาเปรียบเทียบ

4.2.3. การทดลองบนชุดข้อมูลสูญหาย Seed

ตารางที่ 4.4

ตารางเปรียบเทียบค่าความถูกต้องของโมเดลที่ได้จากข้อมูล Seed

Imputation Algorithm	5%	10%	15%	20%	25%
Mean	87.52	88.38	86.38	86.76	91.04
Median	86.47	88.19	86.76	87.42	90.85
PMM	85.90	85.04	83.42	87.42	89.71
kNN(k=3)	86.66	89.71	86.19	86.57	88.38
IviaCLR	86.28	88.761	85.61	86.57	91.71
GAMLR	98.19	100.0	99.52	100.0	100.0

จากตารางที่ 4.4 แสดงการเปรียบเทียบค่าความถูกต้องของโมเดลที่ได้จากข้อมูล Seed ที่มีชนิดข้อมูลการสูญเสียลักษณะ Missing at Random (MAR) ในปริมาณ 5% 10% 15% 20% และ 25% ระหว่าง 5 อัลกอริทึม ดังนี้ การแทนที่ข้อมูลสูญเสียด้วยค่าเฉลี่ย (Mean) การแทนที่ข้อมูลสูญเสียด้วยค่ามัธยฐาน (Median) การแทนที่ข้อมูลสูญเสียด้วยการถดถอยเชิงพหุ (Predictive Mean Matching) การแทนที่ข้อมูลสูญเสียด้วยเพื่อนบ้านใกล้เคียงที่สุด k ตัว (K-Nearest Neighbor Algorithm) การแทนที่ข้อมูลสูญหายด้วยการจัดกลุ่มและสมการถดถอยเชิงเส้น (IviaCLR) และวิธีการที่นำเสนอใหม่นี้ จากภาพข้อมูลในแนวนอน X แสดงถึงค่าเปอร์เซ็นต์ข้อมูลสูญหายตั้งแต่ 5% 10% 15% 20% และ 25% ข้อมูลในแนวแกน Y แสดงถึงค่าความถูกต้องในการทำนายผลของโมเดลทำนายข้อมูลที่ใช้ข้อมูลจากการแทนที่ข้อมูลด้วยวิธีการแทนที่ข้อมูลสูญหายดังนี้ค่าเฉลี่ย (Mean) ค่ามัธยฐาน (Median) การถดถอยเชิงพหุ (Predictive Mean Matching) เพื่อนบ้านใกล้เคียงที่สุด k ตัว (K-Nearest Neighbor Algorithm) และวิธีการที่นำเสนอใหม่ จากภาพที่ 4.3 นี้แสดงให้เห็นว่าวิธีที่นำเสนอใหม่นี้มีค่าความแม่นยำโดยเฉลี่ยสูงกว่าทุกอัลกอริทึมที่นำมาเปรียบเทียบ

4.2 การทดสอบหาความผิดพลาดของการแทนที่ข้อมูลสูญหายด้วย RMSE

วัตถุประสงค์ของ GAMLR นั้นถูกสร้างขึ้นเพื่อเพิ่มประสิทธิภาพให้กับโมเดลพยากรณ์ แต่ในการทดลองย่อยนี้ทำขึ้นเพื่อให้เห็นผลอีกด้านหนึ่งของ GAMLR ว่ามีความสามารถในการนำค่าที่ใกล้เคียงค่าเดิมกลับคืนมา (data recovering)

โดยการเทียบประสิทธิภาพ GAMLR กับอัลกอริทึมแทนที่ข้อมูลสูญหายอื่น เพื่อหาความผิดพลาดของของค่าที่ได้จากอัลกอริทึมการแทนที่ข้อมูลสูญหายเทียบกับค่าที่แท้จริงของข้อมูล ก่อนถูกทำให้เกิดเป็นข้อมูลสูญหายชนิด MAR โดยวิธี RMSE (สมการที่ 2 ในหน้า 4)

4.2.1. การทดลองบนชุดข้อมูลสูญหาย Wine

ตารางที่ 4.5

ตารางเปรียบเทียบค่า RMSE บนชุดข้อมูล Wine

Imputation Algorithm	5%	10%	15%	20%	25%
Mean	0.058	0.073	0.102	0.290	0.530
Median	0.058	0.075	0.098	0.254	0.432
PMM	0.030	0.037	0.060	0.224	0.415
kNN(k=3)	0.054	0.067	0.088	0.207	0.341
IviaCLR	0.058	0.073	0.102	0.290	0.530
GAMLR	0.153	0.166	0.196	0.388	0.573

จากตารางที่ 4.5 จะพบว่าผลการทดลองเปรียบเทียบความคลาดเคลื่อนของค่าข้อมูลสูญหายที่ถูกแทนที่ด้วยอัลกอริทึมแทนที่ข้อมูลสูญหายเทียบกับค่าของข้อมูลจริงบนชุดข้อมูล Wine ด้วยวิธีการเปรียบเทียบ RMSE พบว่า GAMLR มีค่า RMSE สูงที่สุดในทุกอัตราการสูญหายของข้อมูลที่ 5% 10% 15% 20% และ 25% จึงแสดงให้เห็นว่า GAMLR ไม่เหมาะแก่การนำไปแทนที่ข้อมูลสูญหายเพื่อการนำค่าที่ใกล้เคียงค่าเดิมกลับคืนมา (data recovering)

4.2.2. การทดลองบนชุดข้อมูลสูญหาย Glass

ตารางที่ 4.6

ตารางเปรียบเทียบค่า RMSE บนชุดข้อมูล Glass

Imputation Algorithm	5%	10%	15%	20%	25%
Mean	0.038	0.074	0.128	0.259	0.402
Median	0.041	0.081	0.137	0.268	0.416
PMM	0.009	0.035	0.053	0.093	0.223
kNN(k=3)	0.018	0.045	0.080	0.184	0.270
IviaCLR	0.038	0.074	0.128	0.259	0.402
GAMLR	0.191	0.207	0.281	0.387	0.525

จากตารางที่ 4.6 จะพบว่าผลการทดลองเปรียบเทียบความคลาดเคลื่อนของค่าข้อมูลสูญหายที่ถูกแทนที่ด้วยอัลกอริทึมแทนที่ข้อมูลสูญหายเทียบกับค่าของข้อมูลจริงบนชุดข้อมูล Glass ด้วยวิธีการเปรียบเทียบ RMSE พบว่า GAMLR มีค่า RMSE สูงที่สุดในทุกอัตราการสูญหายของข้อมูลที่ 5% 10% 15% 20% และ 25% จึงแสดงให้เห็นว่า GAMLR ไม่เหมาะแก่การนำไปแทนที่ข้อมูลสูญหายเพื่อการนำค่าที่ใกล้เคียงค่าเดิมกลับคืนมา (data recovering)

4.2.3. การทดลองบนชุดข้อมูลสูญหาย Indian Liver Patient

ตารางที่ 4.7

ตารางเปรียบเทียบค่า RMSE บนชุดข้อมูล Indian Liver Patient

Imputation Algorithm	5%	10%	15%	20%	25%
Mean	0.055	0.045	0.068	0.140	0.590
Median	0.061	0.051	0.076	0.150	0.606
PMM	0.041	0.040	0.069	0.153	0.402
kNN(k=3)	0.054	0.045	0.068	0.131	0.509

IviaCLR	0.055	0.045	0.069	0.140	0.590
GAMLR	0.238	0.228	0.237	0.301	0.714

จากตารางที่ 4.7 จะพบว่าผลการทดลองเปรียบเทียบความคลาดเคลื่อนของค่าข้อมูลสูญหายที่ถูกแทนที่ด้วยอัลกอริทึมแทนที่ข้อมูลสูญหายเทียบกับค่าของข้อมูลจริงบนชุดข้อมูล Indian Liver Patient ด้วยวิธีการเปรียบเทียบ RMSE พบว่า GAMLR มีค่า RMSE สูงที่สุดในทุกอัตราการสูญหายของข้อมูลที่ 5% 10% 15% 20% และ 25% จึงแสดงให้เห็นว่า GAMLR ไม่เหมาะแก่การนำไปแทนที่ข้อมูลสูญหายเพื่อการนำค่าที่ใกล้เคียงค่าเดิมกลับคืนมา (data recovering)

4.2.4. การทดลองบนชุดข้อมูลสูญหาย Seed

ตารางที่ 4.8

ตารางเปรียบเทียบค่า RMSE บนชุดข้อมูล Seed

Imputation Algorithm	5%	10%	15%	20%	25%
Mean	0.064	0.071	0.075	0.083	0.109
Median	0.068	0.075	0.789	0.083	0.109
PMM	0.005	0.007	0.008	0.009	0.021
kNN(k=3)	0.011	0.014	0.016	0.017	0.026
IviaCLR	0.064	0.071	0.075	0.083	0.109
GAMLR	0.168	0.158	0.166	0.163	0.193

จากตารางที่ 4.8 จะพบว่าผลการทดลองเปรียบเทียบความคลาดเคลื่อนของค่าข้อมูลสูญหายที่ถูกแทนที่ด้วยอัลกอริทึมแทนที่ข้อมูลสูญหายเทียบกับค่าของข้อมูลจริงบนชุดข้อมูล Seed ด้วยวิธีการเปรียบเทียบ RMSE พบว่า GAMLR มีค่า RMSE สูงที่สุดในทุกอัตราการสูญหายของข้อมูลที่ 5% 10% 15% 20% และ 25% จึงแสดงให้เห็นว่า GAMLR ไม่เหมาะแก่การนำไปแทนที่ข้อมูลสูญหายเพื่อการนำค่าที่ใกล้เคียงค่าเดิมกลับคืนมา (data recovering)

บทที่ 5

สรุปผลวิจัยและข้อเสนอแนะ

5.1 อภิปราย

งานวิจัยนี้เป็นการทดลองแทนที่ข้อมูลสูญหายโดยประยุกต์ใช้ Genetic Algorithm เป็นพื้นฐานในการแทนที่ข้อมูลสูญหาย ซึ่งที่ผ่านมายังพบงานวิจัยลักษณะนี้ได้น้อย และงานวิจัยที่พบนั้นได้ทำการทดลองในลักษณะของชุดข้อมูลชนิดหมวดหมู่ (category) ซึ่งงานวิจัยในข้อมูลประเภทตัวเลขที่มีความต่อเนื่อง และทำการทดลองในการเทียบผลความถูกต้องของโมเดลพยากรณ์เมื่อใช้ข้อมูลที่สมบูรณ์ที่ถูกแทนที่ด้วยอัลกอริทึมการแทนที่ข้อมูลสูญหายแล้ว จากการทดลอง Genetic Algorithm เพื่อทดสอบผลการทดลอง ทำให้ทราบว่าสามารถแทนที่ข้อมูลได้ดีกว่า Mean, Median, PMM และ KNN แต่เนื่องจาก GA มีพื้นฐานจากการ Random ทำให้จำเป็นที่จะต้องใช้เวลาหรือรอบในการสร้างประชากรรุ่นใหม่ขึ้นมาเพื่อให้ได้ผลลัพธ์ที่ดีกว่าอัลกอริทึมที่ต้องการเปรียบเทียบ จึงทำให้เกิดการพัฒนาอัลกอริทึมที่มีการปรับปรุงจาก Genetic Algorithm มาประยุกต์ใช้กับ วิธีการถดถอยเชิงเส้นพหุ (Linear Regression) และ การค้นหาด้วยวิธีเพื่อนบ้านใกล้เคียง k ตัว ให้พัฒนา Genetic Algorithm and Multiple Regression ขึ้นมาเพื่อจำกัดขอบเขตของการ Random ใน Genetic Algorithm

Genetic Algorithm and Multiple Regression เป็นการพัฒนาต่อจาก Genetic Algorithm แบบดั้งเดิม โดยมีการประยุกต์ใช้ KNN และ Linear Regression มาเป็น Heuristic Function เพื่อกำหนดทิศทางในการเลือกค่าสำหรับการแทนที่ข้อมูลสูญหายเพื่อนำไปใช้หาผลจากการทำนายข้อมูลของโมเดลพยากรณ์ ซึ่งผลการทดลองแสดงให้เห็นว่าอัลกอริทึมนี้มีประสิทธิภาพสูงหรือไม่น้อยกว่าวิธีการแทนที่ข้อมูลแบบดั้งเดิมที่ปริมาณข้อมูลสูญหายที่ 5 10 15 20 และ 25 เปอร์เซ็นต์ ผลการทดลองแสดงให้เห็นว่าการเพิ่มประสิทธิภาพของอัลกอริทึมนี้ได้ผลดีกว่าอัลกอริทึมอื่นเฉลี่ยที่ 7.2% ที่ชุดข้อมูลทดสอบ Indian Patient Liver ซึ่งเป็นชุดข้อมูลที่มีขนาดใหญ่ที่สุดที่นำมาทดสอบและมีปริมาณข้อมูลมากกว่าชุดข้อมูลที่มีปริมาณรองลงมามากถึง 2.7 เท่า จึงแสดงให้เห็นว่า ยังมีขนาดข้อมูลสูญหายมากเท่าใด ยิ่งให้เห็นถึงประสิทธิภาพของอัลกอริทึมนี้มากขึ้นเท่านั้น

ในการทดลองนี้มีข้อจำกัดอยู่ในด้านของ Hardware เนื่องจาก Genetic Algorithm จำเป็นจะต้องเก็บรักษาข้อมูลของประชากร (population) และข้อมูลสูญหายที่อยู่ในรูปของเวกเตอร์ (chromosome) โดยยังมีชุดข้อมูลขนาดใหญ่ ยิ่งต้องการพื้นที่ในการเก็บข้อมูลขึ้นเท่านั้น อีกทั้งการหาผลความถูกต้องของโมเดลพยากรณ์นั้นต้องทำในทุกครั้งที่มีการเกิดของประชากรใน Genetic

Algorithm ใหม่ ทำให้ต้องใช้การคำนวณและใช้เวลาเป็นอย่างมาก ซึ่งปริมาณเวลาที่ใช้นั้นแปรผันตรงกับปริมาณข้อมูลในชุดข้อมูลที่นำมาใช้ หมายความว่า หากเป็นชุดข้อมูลขนาดใหญ่ จะทำให้ใช้เวลาในการหาคำตอบของผลการทดลองนานขึ้นไปด้วย

5.2 ข้อเสนอแนะในครั้งถัดไป

จากการทดลองครั้งนี้ ทำให้ทราบว่ายังมีสิ่งที่ควรปรับปรุงหรือเพิ่มเติม เพื่อให้งานวิจัยในในครั้งถัดไปมีประสิทธิภาพดีขึ้น โดยสามารถสรุปได้ดังนี้

5.2.1 สามารถใช้อัลกอริทึมสำหรับการเตรียมชุดคำตอบ เพื่อใช้ในการเตรียมชุดข้อมูลสำหรับทดลอง นอกจากวิธีการสุ่มออกแบบ MAR ซึ่งจะทำให้มีการเปรียบเทียบผลการทดลองที่หลากหลายรูปแบบมากยิ่งขึ้น

5.2.2 ประยุกต์ใช้ Feature Selection Algorithm ในการหาค่าความสัมพันธ์ของแต่คุณลักษณะในชุดข้อมูล ซึ่งจะส่งผลให้สามารถหาคำตอบของ KNN และ Linear Regression ได้ดียิ่งขึ้น

5.2.3 หา Error Rate ในการแทนที่ข้อมูลสูญหายว่ามีความคลาดเคลื่อนจากคำตอบจริงเท่าใด ซึ่งอาจนำมาปรับปรุง Heuristic Function ของ Genetic Algorithm ได้

รายการอ้างอิง

บทความวารสาร

1. Han J, Kamber M, Pei J. Data mining concepts and techniques, 3rd ed. Waltham, Mass.: Morgan Kaufmann; 2012.
2. Cormen TH, Leiserson CE, Rivest RL, Stein C. Introduction to Algorithms. 2nd ed. USA: The MIT Press; 2001.
3. Feng X, Wu S, Liu Y. Imputing Missing Values for Mixed Numeric and Categorical Attributes Based on Incomplete Data Hierarchical Clustering [Internet]. Knowledge Science, Engineering and Management. Springer Berlin Heidelberg; 2011. p. 414-24. Available: http://dx.doi.org/10.1007/978-3-642-25975-3_37
4. Jadhav A, Pramod D, Ramanathan K. Comparison of Performance of Data Imputation Methods for Numeric Dataset [Internet]. Vol. 33, Applied Artificial Intelligence. Informa UK Limited; 2019. p. 913-33. Available from: <http://dx.doi.org/10.1080/08839514.2019.1637138>
5. Kowarik A, Templ M. Imputation with the R Package VIM [Internet]. Vol. 74, Journal of Statistical Software. Foundation for Open Access Statistics; 2016. Available from: <http://dx.doi.org/10.18637/jss.v074.i07>
6. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R [Internet]. Vol. 45, Journal of Statistical Software. Foundation for Open Access Statistics; 2011. Available from: <http://dx.doi.org/10.18637/jss.v045.i03>
7. Mandel J SP. A Comparison of Six Methods for Missing Data Imputation [Internet]. Vol. 06, Journal of Biometrics & Biostatistics. OMICS Publishing Group; 2015. Available from: <http://dx.doi.org/10.4172/2155-6180.1000224>
8. Kleinke K. Multiple Imputation by Predictive Mean Matching When Sample Size Is Small [Internet]. Vol. 14, Methodology. Hogrefe Publishing Group; 2018. p. 3-15. Available from: <http://dx.doi.org/10.1027/1614-2241/a000141>
9. Vink G, Frank LE, Pannekoek J, van Buuren S. Predictive mean matching imputation of semicontinuous variables [Internet]. Vol. 68, Statistica Neerlandica. Wiley; 2014. p. 61–90. Available from: <http://dx.doi.org/10.1111/stan.12023>

10. Chungnoy K, Tanantong T, Songmuang P. Missing Value Imputation on Gene Expression Data using Bee-based algorithm to Improve Classification Performance. Research Square Platform LLC; 2022. Available from: <http://dx.doi.org/10.21203/rs.3.rs-2186533/v1>
11. Uyanık GK, Güler N. A Study on Multiple Linear Regression Analysis [Internet]. Vol. 106, Procedia - Social and Behavioral Sciences. Elsevier BV; 2013. p. 234-40. Available from: <http://dx.doi.org/10.1016/j.sbspro.2013.12.027>
12. Karmita N, Taheri S, Bagirov A, Makinen P. Missing Value Imputation via Clusterwise Linear Regression [Internet]. IEEE Transactions on Knowledge and Data Engineering. Institute of Electrical and Electronics Engineers (IEEE); 2020. p. 1-1. Available from: <http://dx.doi.org/10.1109/TKDE.2020.3001694>
13. Mostafa SM. Imputing missing values using cumulative linear regression [Internet]. Vol. 4, CAAI Transactions on Intelligence Technology. Institution of Engineering and Technology (IET); 2019. p. 182-200. Available from: <http://dx.doi.org/10.1049/trit.2019.0032>

ประวัติผู้เขียน

ชื่อ	สุรวัช อำพัน
วันเดือนปีเกิด	กรกฎาคม 2535
วุฒิการศึกษา	ปีการศึกษา 2558: วิทยาศาสตร์บัณฑิต มหาวิทยาลัยธรรมศาสตร์
ตำแหน่ง	พนักงานโปรแกรมคอมพิวเตอร์5 บริษัท โทรคมนาคม แห่งชาติ จำกัด (มหาชน)

ผลงานทางวิชาการ

Surawach Amphan and Pokpong Songmuang “Data Imputation with Genetic Algorithm and Multiple Linear Regression for Improving Performance of Prediction Model” 2023 The International Conference on Cybernetics and Innovations (ICCI), March 2023, Phetchaburi, Thailand