การปรับปรุงอัลกอริทึม ID3 โดยใช้วิธีการรวมแอตทริบิวต์ที่มีความสำคัญใกล้เคียงกัน

IMPROVING ID3 ALGORITHM BY COMBINING NEARLY IMPORTANT ATTRIBUTES

สุรัชนันท์ ไกรเดช

SURATCHANAN KRAIDECH

วิทยานิพนธ์นี้สำหรับการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ.2562

KMITL-2019-EN-M-070-047

การปรับปรุงอัลกอริทึม ID3 โดยใช้วิธีการรวมแอตทริบิวต์ที่มีความสำคัญใกล้เคียงกัน

IMPROVING ID3 ALGORITHM BY COMBINING NEARLY IMPORTANT ATTRIBUTES

สุรัชนันท์ ไกรเดช

SURATCHANAN KRAIDECH

วิทยานิพนธ์นี้สำหรับการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ.2562

KMITL-2019-EN-M-070-047

IMPROVING ID3 ALGORITHM BY COMBINING NEARLY IMPORTANT ATTRIBUTES

SURATCHANAN KRAIDECH

A THESIS SUBMITTED IN FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF ENGINEERING IN COMPUTER ENGINEERING

FACULTY OF ENGINEERING

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2019

KMITL-2019-EN-M-070-047

คณะวิศวกรรมศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์ การปรับปรุงอัลกอริทึม ID3 โดยใช้วิธีการรวมแอตทริบิวต์ที่มีความสำคัญใกล้เคียงกัน
Thesis Title Improving ID3 Algorithm by Combining Nearly Important Attributes
นักศึกษา นายสุรัชนันท์ ไกรเดช
รหัสประจำตัว 59601335
ปริญญา วิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชา วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์ รศ.ดร.เกียรติกูล เจียรนัยธนะกิจ
หมายเลขวิทยานิพนธ์ KMITL-2019-EN-M-070-047

คณะกรรมการสอบวิทยานิพนธ์		ลายมือชื่อ
รศ.ดร.สุรพงค์	เอื้อวัฒนามงคล	
รศ.ดร.บุญธีร์	เครือตราชู	
ผศ.ดร.ชุติเมษฎ์	ศรีนิลทา	
ผศ.ดร.รัฐชัย	ชาวอุทัย	
รศ.ดร.เกียรติกูล	เจียรนัยธนะกิจ	

วัน / เดือน/ ปี ที่สอบ วันอังคารที่ 18 มิถุนายน พ.ศ. 2562 เวลา 09.30-11.30 น.
สถานที่สอบ ณ ห้องประชุม 3 ชั้น 5 อาคาร A

คณะวิศวกรรมศาสตร์ รับรองแล้ว

(รองศาสตราจารย์ ดร. คมสัน มาลีสี)
คณบดี คณะวิศวกรรมศาสตร์
วันที่ 18 มิถุนายน พ.ศ. 2562

หัวข้อวิทยานิพนธ์	การปรับปรุงอัลกอริทึม ID3 โดยใช้วิธีการรวมแอตทริบิวต์ที่มีความสำคัญใกล้เคียงกัน
นักศึกษา	นายสุรัชนันท์ ไกรเดช
รหัสประจำตัว	59601335
ปริญญา	วิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
พ.ศ.	2562
อาจารย์ที่ปรึกษาวิทยานิพนธ์	รศ. ดร. เกียรติกูล เจียรนัยธนะกิจ

บทคัดย่อ

อัลกอริทึม ID3 เป็นหนึ่งในอัลกอริทึมสำหรับสร้างต้นไม้ตัดสินใจที่นิยมมากในหมู่งานด้านการจำแนกประเภท มีงานวิจัยหลายงานที่ทำการปรับปรุงอัลกอริทึม ID3 ด้วยวิธีการต่างๆ วิทยานิพนธ์นี้ได้เสนองานวิจัยซึ่งทำการปรับปรุงอัลกอริทึม ID3 โดยใช้วิธีการรวมแอตทริบิวต์ที่มีความสำคัญใกล้เคียงกัน วิธีการดังกล่าวจะทำการปรับเปลี่ยนวิธีการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 ให้สามารถเลือกแอตทริบิวต์ที่มีความสำคัญใกล้เคียงกันร่วมกับแอตทริบิวต์ที่มีความสำคัญที่สุดได้ งานวิจัยนี้ได้ทำการทดลองบนชุดข้อมูล 6 ชุดข้อมูล ซึ่งเป็นชุดข้อมูลมาตรฐานจาก UCI Machine Learning Repository ผลการทดลองแสดงให้เห็นถึงประสิทธิภาพในการลดความลึกสูงสุดและความลึกในการจำแนกของต้นไม้ตัดสินใจอย่างเห็นได้ชัด นอกจากนี้ เวลาในการทดสอบก็ลดลงในขณะที่ความแม่นยำในการจำแนกยังคงระดับเดิม

Thesis	Improving ID3 Algorithm by Combining Nearly Important Attributes
Student	Mr.Suratchanan Kraidech
Student ID.	59601335
Degree	Master of Engineering
Program	Computer Engineering
Year	2019
Thesis Advisor	Assoc.Prof.Dr. Kietikul Jearanaitanakij

ABSTRACT

The ID3 algorithm is one of the most popular decision tree algorithms which is mainly used in the classification task. There are many pieces of research about improving the ID3 algorithm by using various strategies. We improved the ID3 algorithm by combining nearly important attributes. This strategy changes the attribute selection to allow the neighboring second-place important attributes to be combined with the most important attributes. The proposed algorithm is tested on six standard benchmarks from the UCI repository. The experimental results indicate the significant reduction in the maximum depth and the classification depth of the decision tree. In addition, the testing time is also reduced while the classification accuracy is satisfying stable.

กิตติกรรมประกาศ

วิทยานิพนธ์เล่มนี้สำเร็จได้ด้วยความกรุณาจากอาจารย์ที่ปรึกษา รศ.ดร.เกียรติกูล เจียรนัยธนะกิจ ที่ให้ความช่วยเหลือ ให้คำชี้แนะช่วยแก้ปัญหา ตลอดจนให้ความรู้และประสบการณ์ที่ดีแก่ข้าพเจ้า

ขอขอบพระคุณกรรมการสอบหัวข้อและโครงร่างวิทยานิพนธ์ทุกท่านที่ได้กรุณาให้คำแนะนำตลอดจนข้อชี้แนะ จนในที่สุดทำให้วิทยานิพนธ์นี้สำเร็จลงได้

สุดท้ายต้องขอขอบคุณนางสาวณิชา แก้วรอด ที่คอยให้คำแนะนำช่วยแก้ปัญหาและเป็นกำลังใจให้กับข้าพเจ้าในการทำเล่มวิทยานิพนธ์ จนกระทั่งทำให้วิทยานิพนธ์นี้เสร็จสมบูรณ์ไปด้วยดี

สำหรับคุณงามความดีอันใดที่เกิดจากวิทยานิพนธ์นี้ ข้าพเจ้าขอมอบให้กับบิดามารดา ซึ่งเป็นที่รักและเคารพยิ่ง ตลอดจนครูอาจารย์ที่เคารพทุกท่านที่ได้ประสิทประสาทวิชาความรู้และถ่ายทอดประสบการณ์ที่ดีให้แก่ข้าพเจ้า

สุรัชนันท์ ไกรเดช

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญรูป

รูปที่ **หน้า**

สารบัญรูป (ต่อ)

รูปที่ **หน้า**

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การเรียนรู้ต้นไม้ตัดสินใจ (Decision tree learning) เป็นวิธีการเรียนรู้สำหรับการจำแนกประเภท (Classification) ของข้อมูลที่สนใจ โดยถูกนำไปใช้ในงานหลายด้าน เช่น การวิเคราะห์การตัดสินใจ, การวินิจฉัยทางการแพทย์ และการรู้จำรูปแบบ (Pattern recognition) เป็นต้น ในบรรดาอัลกอริทึมต้นไม้ตัดสินใจที่มีอยู่หลายชนิดนั้น อัลกอริทึมไอดีทรี (Iterative Dichotomiser 3, ID3) เป็นอัลกอริทึมที่นิยมใช้กันอย่างแพร่หลาย คิดค้นโดย Quinlan ในปี 1986 [1]

มีงานวิจัยหลายงานที่พยายามปรับปรุงและพัฒนาอัลกอริทึม ID3 เช่น Liu และ Li [2] ใช้หลักการระดับนัยสำคัญของแอตทริบิวต์ (Attribute significance) และฟังก์ชันนูน (Convex function) ในการปรับปรุงอัลกอริทึม ID3 งานวิจัยนี้ระบุว่าวิธีการที่เสนอสามารถเพิ่มความแม่นยำในการจำแนก (Classification accuracy) ลดจำนวนโหนด (Node) ที่สร้างในต้นไม้ตัดสินใจ รวมไปถึงลดเวลาในการสร้างต้นไม้ตัดสินใจอีกด้วย (Training time)

Luo, Chen และ Zhang [3] ใช้ทฤษฎีเทย์เลอร์ (Taylor's theorem) และน้ำหนักความสำคัญของแอตทริบิวต์ (Attribute importance-weighted) ในการปรับปรุงอัลกอริทึม ID3 ผลการทดลองในงานวิจัยนี้แสดงให้เห็นถึงการลดลงของเวลาในการรัน (Running time) และความแม่นยำในการจำแนกที่เพิ่มขึ้น

Liu, Hu และ Yan [4] ใช้ระยะทางแบบยุคลิดโดยเฉลี่ย (Average Euclidean distance) ในการปรับปรุงอัลกอริทึม ID3 งานวิจัยนี้สามารถเพิ่มความเที่ยงตรง (Precision) ในการจำแนก ลดเวลาในการสร้างต้นไม้ตัดสินใจ และลดเวลาในการทดสอบ (Testing time) ของอัลกอริทึม ID3 ได้

งานวิจัย 3 งานข้างต้นที่ทำการปรับปรุงและพัฒนาอัลกอริทึม ID3 นั้น ทั้งหมดล้วนจัดการกับปัญหาของอัลกอริทึม ID3 2 ปัญหาคือ ปัญหาการลำเอียง (bias) ของการเลือกแอตทริบิวต์ในอัลกอริทึม ID3 โดยใช้เกนความรู้ (Information gain) และปัญหาความซับซ้อนของการคำนวณค่าเกนความรู้ในอัลกอริทึม ID3 อย่างไรก็ตาม มีอีกปัญหาหนึ่งของอัลกอริทึม ID3 ที่ยังไม่มีงานวิจัยใดหรือผู้ใดเคยแก้ปัญหามาก่อน ปัญหาดังกล่าวคือ ปัญหาแอตทริบิวต์ที่มีความสำคัญเท่ากัน (Equally important attributes problem) เป็นปัญหาการเลือกแอตทริบิวต์ในอัลกอริทึม ID3 เมื่อมีแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันอย่างน้อย 2 แอตทริบิวต์ กล่าวคือแอตทริบิวต์เหล่านี้เป็นแอตทริบิวต์ที่มีความสำคัญเท่ากัน โดยปกติแล้วอัลกอริทึม ID3 เมื่อเจอสถานการณ์นี้ อัลกอริทึม ID3 จะสุ่มเลือกแอตทริบิวต์ใดแอตทริบิวต์หนึ่งจากกลุ่มของแอตทริบิวต์ที่มีความสำคัญเท่ากัน เพื่อนำไปไว้ในโหนดตัดสินใจที่กำลังพิจารณา วิธีการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 จากปัญหานี้อาจทำให้

ต้นไม้ตัดสินใจที่ได้มีความลึกที่มากเกินความจำเป็น และมีโอกาสทำให้การจำแนกตัวอย่างที่ไม่เคยฝึกฝนมาก่อนมีความช้าลง วิทยานิพนธ์นี้ได้ตีพิมพ์งานวิจัยงานแรก [5] ซึ่งทำการปรับปรุงอัลกอริทึม ID3 จากปัญหาที่ได้อธิบายก่อนหน้านี้ โดยใช้วิธีการรวมแอตทริบิวต์ที่มีความสำคัญเท่ากัน ผลการทดลองในงานวิจัยงานแรกนี้ แสดงให้เห็นถึงประสิทธิภาพของอัลกอริทึม ID3 ที่ปรับปรุงซึ่งสามารถลดความลึกสูงสุดของต้นไม้ตัดสินใจได้ ในขณะที่ความแม่นยำในการจำแนกยังคงระดับเดิม

ถึงแม้ว่าวิธีการรวมแอตทริบิวต์ที่มีความสำคัญเท่ากัน จะทำให้ต้นไม้ตัดสินใจที่ได้มีความลึกสูงสุดที่น้อยลง แต่เงื่อนไขในการเลือกแอตทริบิวต์ของวิธีการนี้นั้นมีความตึงเกินไป เพราะแอตทริบิวต์ที่จะสามารถเลือกมารวมด้วยกันได้นั้น ต้องมีค่าเกนความรู้สูงที่สุดและเท่ากันจริงๆ ด้วยเหตุนี้เอง วิธีการที่เสนอในงานวิจัยแรกของวิทยานิพนธ์นี้ จึงมีปัญหาความตึงของวิธีการเลือกแอตทริบิวต์ ปัญหานี้ส่งผลต่อความสามารถในการลดความลึกสูงสุดของต้นไม้ตัดสินใจที่สร้าง โดยเฉพาะอย่างยิ่งเมื่อใช้อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกกับชุดข้อมูลที่พบแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ต่อการพิจารณาเลือกแอตทริบิวต์โดยเฉลี่ยอยู่ในจำนวนที่น้อย ซึ่งจะมีผลทำให้ความสามารถในการลดความลึกสูงสุดของต้นไม้ตัดสินใจที่สร้างจากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกมีน้อย

จากปัญหาความตึงของวิธีการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกที่ได้อธิบายไปนั้น วิทยานิพนธ์นี้จึงมีแนวคิดที่จะทำการปรับปรุงอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกให้สามารถลดความลึกสูงสุดของต้นไม้ตัดสินใจได้มากขึ้นกว่าเดิม โดยใช้วิธีการรวมแอตทริบิวต์ที่มีความสำคัญใกล้เคียงกัน ซึ่งจะทำการลดความตึงของเงื่อนไขในการเลือกแอตทริบิวต์ลง ยอมให้แอตทริบิวต์ที่มีความสำคัญใกล้เคียงกับแอตทริบิวต์ที่มีความสำคัญที่สุด สามารถที่จะถูกเลือกมาอยู่ในโหนดตัดสินใจเดียวกันร่วมกับแอตทริบิวต์ที่มีความสำคัญที่สุดได้

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

เสนอวิธีการปรับปรุงอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกเพื่อจัดการกับปัญหาความตึงของวิธีการเลือกแอตทริบิวต์ และเพื่อลดความลึกสูงสุดของต้นไม้ตัดสินใจที่ได้ให้น้อยลงยิ่งขึ้น ในขณะที่ความแม่นยำในการจำแนกยังคงระดับเดิม

1.3 สมมติฐานของการศึกษา

วิธีการที่เสนอสามารถที่จะทำให้ต้นไม้ตัดสินใจที่ได้มีความลึกที่น้อยกว่าเดิมอย่างเห็นได้ชัดเมื่อเทียบกับต้นไม้ตัดสินใจที่ได้จากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกและอัลกอริทึม ID3 แบบดั้งเดิม ในขณะที่ความแม่นยำในการจำแนกยังคงเท่าเดิม

1.4 ขอบเขตของการวิจัย

1) วิธีการปรับปรุงอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกโดยใช้วิธีการรวมแอตทริบิวต์ที่มีความสำคัญใกล้เคียงกัน จะพิจารณาแอตทริบิวต์ที่มีความสำคัญอยู่อันดับที่ 1 และ 2 เป็นหลัก
2) การทดสอบต้นไม้ตัดสินใจที่ได้จากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ จะทำการทดสอบบนชุดข้อมูลมาตรฐาน (Benchmark dataset) จาก UCI Machine Learning Repository [6] จำนวน 6 ชุดข้อมูล

1.5 ขั้นตอนของการศึกษา

1) ศึกษาการทำงานของอัลกอริทึม ID3 แบบปกติ
2) ศึกษางานวิจัยที่มีการปรับปรุงอัลกอริทึม ID3 ในมุมมองที่แตกต่างกัน
3) ทำการคิดค้นวิธีการรวมแอตทริบิวต์ที่มีความสำคัญเท่ากัน ทดลองและตีพิมพ์เป็นงานวิจัยแรก
4) ทำการศึกษาหาปัญหาและข้อเสียของงานวิจัยแรกที่ได้ทำไป
5) ทำการศึกษาทดลอง โดยดัดแปลงวิธีการที่เสนอในงานวิจัยแรก ให้สามารถเลือกรวมแอตทริบิวต์ที่มีความสำคัญใกล้เคียงกันได้ และดูผลลัพธ์ที่ได้ว่าเป็นอย่างไร
6) ทำการคิดค้นวิธีการรวมแอตทริบิวต์ที่มีความสำคัญใกล้เคียงกันอย่างเป็นขั้นเป็นตอน
7) ทำการทดลองเบื้องต้นหาพารามิเตอร์ที่เหมาะสมที่สุด เพื่อใช้เป็นพารามิเตอร์ในการทดลองหลักของวิทยานิพนธ์
8) ทำการทดลองหลักโดยใช้พารามิเตอร์ที่หาได้จากขั้นตอนที่ 7 และบันทึกสรุปผล
9) เรียบเรียงสิ่งที่ได้ศึกษาวิจัย แล้วจัดทำเล่มวิทยานิพนธ์

1.6 เครื่องมือและอุปกรณ์ที่ใช้ในงานวิจัย

1) โน้ตบุ๊กส่วนบุคคล หน่วยประมวลผลกลาง Intel(R) Core(TM) i7-7700HQ ความเร็ว 2.8 GHz หน่วยความจำหลักขนาด 8 GB
2) ระบบปฏิบัติการ Windows 10 64-บิต
3) โปรแกรม Visual Studio 2017

1.7 โครงสร้างของวิทยานิพนธ์

วิทยานิพนธ์นี้ประกอบไปด้วย 7 บท ซึ่งแต่ละบทมีหัวข้อดังนี้

บทที่ 1 เป็นบทนำ

บทที่ 2 กล่าวถึงทฤษฎีที่เกี่ยวข้อง

บทที่ 3 นำเสนอเกี่ยวกับงานวิจัยที่เกี่ยวข้อง

บทที่ 4 อธิบายเกี่ยวกับอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก

บทที่ 5 อธิบายเกี่ยวกับอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ

บทที่ 6 นำเสนอเกี่ยวกับผลการทดลอง

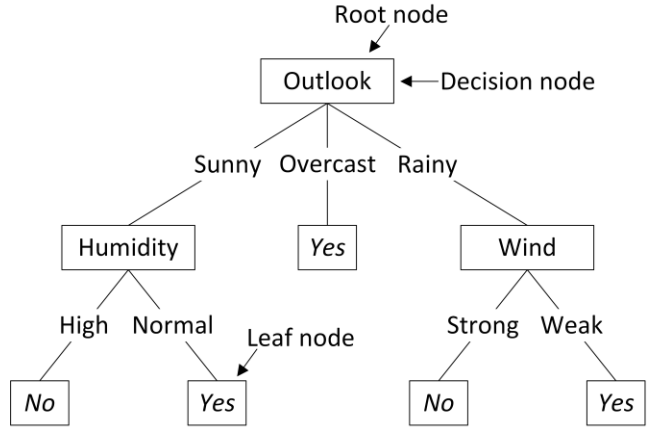
บทที่ 7 เป็นบทสรุปและข้อเสนอแนะ

บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

2.1 ต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจ (Decision tree) เป็นโครงสร้างข้อมูลชนิดหนึ่ง ที่มีพื้นฐานมาจากโครงสร้างข้อมูลที่เป็นต้นไม้ (Tree) ทำหน้าที่ในการตัดสินใจให้กับข้อมูล ซึ่งการตัดสินใจนั้นจะเป็นการหาคำตอบตรงๆ หรือเป็นการพยากรณ์คำตอบก็ได้ ต้นไม้ตัดสินใจมีส่วนประกอบอยู่ 3 ส่วนคือ

1) โหนดราก (Root node) เป็นโหนดเริ่มต้นของการนำข้อมูลมาตัดสินใจ
2) โหนดตัดสินใจ (Decision node) คือโหนดที่ทำหน้าที่ในการตัดสินใจ ว่าถ้าข้อมูลตรงกับค่านี้ ค่านั้น จะต้องไปหาโหนดไหนต่อ โหนดชนิดนี้จะเป็นโหนดที่มีลูกเสมอ
3) โหนดคำตอบ คือโหนดที่ทำหน้าที่ในการให้คำตอบกับข้อมูลที่ต้องการการตัดสินใจ โหนดชนิดนี้เป็นโหนดที่ไม่มีลูก (Leaf node) เสมอ



รูปที่ 2.1 โครงสร้างของต้นไม้ตัดสินใจ

2.2 อัลกอริทึม ID3

อัลกอริทึม ID3 (Iterative Dichotomiser 3, ID3) เป็นอัลกอริทึมการเรียนรู้ (Learning algorithm) สำหรับสร้างต้นไม้ตัดสินใจ คิดค้นโดย Quinlan ในปี 1986 [1] แนวคิดหลักของอัลกอริทึม ID3 คือ ใช้ค่าที่เรียกว่า เกนความรู้ (Information gain) เป็นเกณฑ์ในการเลือกแอตทริบิวต์ที่มีความสำคัญที่สุดเพื่อนำไปเป็นแอตทริบิวต์บนโหนดตัดสินใจ (Decision node) อัลกอริทึมนี้นิยมใช้กับชุดข้อมูลที่มีความไม่ต่อเนื่อง

2.2.1 เอนโทรปี (Entropy)

เอนโทรปี ในที่นี้จะกล่าวถึงเอนโทรปีทางทฤษฎีสารสนเทศ (Information theory) เอนโทรปี คือหน่วยวัดปริมาณสารสนเทศ (Information) โดยเฉลี่ยที่ถูกสร้างโดยแหล่งข้อมูล (Data) ที่มีความไม่แน่นอน เอนโทรปีมีหน่วยเป็นบิต (Bit) เอนโทรปีจะมีค่ามากหรือน้อยขึ้นอยู่กับความรู้ (knowledge) ที่ผู้รับสารสนเทศมีต่อข้อมูลนั้นมาก่อน โดยจะมีค่าน้อยถ้าผู้รับสารสนเทศมีความรู้อยู่แล้ว และจะมีค่ามากถ้าผู้รับสารสนเทศมีความรู้น้อย

สมมติว่ามีกล่อง 3 กล่อง โดยแต่ละกล่องจะมีตัวอักษรภาษาอังกฤษใส่ไว้อยู่ 4 ตัว

- กล่องที่ 1 มีตัว A จำนวน 4 ตัว
- กล่องที่ 2 มีตัว A จำนวน 3 ตัว และตัว B จำนวน 1 ตัว
- กล่องที่ 3 มีตัว A จำนวน 2 ตัว และตัว B จำนวน 2 ตัว

จากรายละเอียดกล่องทั้ง 3 กล่อง เรามีความรู้เกี่ยวกับการหยิบตัวอักษรแบบสุ่มจากทั้ง 3 กล่องดังนี้

1) กล่องที่ 1 เรารู้ว่ากล่องนี้หยิบได้ตัว A แน่นอน
2) กล่องที่ 2 เรารู้ว่า มีโอกาส 75% ที่จะหยิบได้ตัว A และมีโอกาส 25% ที่จะหยิบได้ตัว B
3) กล่องที่ 3 เรารู้ว่า มีโอกาส 50% ที่จะหยิบได้ตัว A และมีโอกาส 50% ที่จะหยิบได้ตัว B

จากความรู้ในการหยิบตัวอักษรแบบสุ่มของทั้ง 3 กล่องนี้ จะพบว่า เรามีความรู้ในกล่องใบแรกมากที่สุด ตามมาด้วยกล่องที่ 2 ซึ่งเรามีความรู้ระดับกลาง และกล่องที่ 3 เรามีความรู้น้อยที่สุด อย่างไรก็ตามเอนโทรปีของกล่องทั้ง 3 ใบนี้ กล่องใบแรกมีเอนโทรปีน้อยที่สุด กล่องที่ 2 มีเอนโทรปีระดับกลาง และกล่องที่ 3 มีเอนโทรปีมากที่สุด เพื่อพิสูจน์ค่าเอนโทรปีของแต่ละกล่องว่าเป็นไปตามที่ได้อธิบายมาจริงหรือไม่ เราจะใช้สมการเอนโทรปีในการหาค่าเอนโทรปีของแต่ละกล่อง

กำหนดให้ X เป็นตัวแปรสุ่ม (Random variable) ที่มีค่าที่เป็นไปได้ k ค่า,

$\{x_1, x_2, ..., x_k\}$ เป็นเซตของค่าที่เป็นไปได้ของตัวแปรสุ่ม X,

และ $p(x_i)$ แทนความน่าจะเป็นที่ค่าของตัวแปรสุ่มจะเป็น x_i โดยที่ i มีค่าตั้งแต่ 1 ถึง k สมการรูปทั่วไปของเอนโทรปีของตัวแปรสุ่ม X คือ

$$Entropy(X) = -\sum_{i=1}^{k} p(x_i) \log_2 p(x_i) \tag{2.1}$$

จากสมการรูปทั่วไปของเอนโทรปีที่แสดงไว้ด้านบน ทำการหาค่าเอนโทรปีของกล่องแต่ละกล่อง โดยที่กล่องแต่ละกล่องถือว่าเป็นตัวแปรสุ่ม ซึ่งมีค่าที่เป็นไปได้ก็คือ ชนิดของตัวอักษรที่อยู่ในกล่องแต่ละกล่อง แต่ละกล่องมีความน่าจะเป็นที่จะหยิบได้ตัวอักษรแต่ละชนิด

กล่องที่ 1 มีตัวอักษร A เพียงชนิดเดียว ความน่าจะเป็นในการหยิบได้ตัว A เป็น 100% หรือ 1 ทำการคำนวณหาเอนโทรปีของกล่องที่ 1 จะได้

$$Entropy(Box1) = -1\log_2 1 = 0 \tag{2.2}$$

กล่องที่ 2 มีตัวอักษรชนิด A และ B ความน่าจะเป็นในการหยิบได้ตัว A เป็น 75% หรือ 0.75 ในขณะที่ความน่าจะเป็นในการหยิบได้ตัว B เป็น 25% หรือ 0.25 ทำการคำนวณหาเอนโทรปีของกล่องที่ 2 จะได้

$$Entropy(Box2) = -(0.75\log_2 0.75 + 0.25\log_2 0.25) \approx 0.811 \tag{2.3}$$

กล่องที่ 3 มีตัวอักษรชนิด A และ B ความน่าจะเป็นในการหยิบได้ตัว A เป็น 50% หรือ 0.5 ในขณะที่ความน่าจะเป็นในการหยิบได้ตัว B เป็น 50% หรือ 0.5 ทำการคำนวณหาเอนโทรปีของกล่องที่ 3 จะได้

$$Entropy(Box3) = -(0.5\log_2 0.5 + 0.5\log_2 0.5) = 1 \tag{2.4}$$

จากค่าเอนโทรปีของแต่ละกล่องที่ได้คำนวณไป เราจะพบว่าค่าเอนโทรปีของแต่ละกล่องมีความสอดคล้องตรงตามที่ได้อธิบายไว้ก่อนหน้า กล่องที่ 1 นั้นมีเอนโทรปี 0 บิต สาเหตุหลักเป็นเพราะว่า กล่องที่ 1 เรา (ผู้รับสารสนเทศ) รู้อยู่แล้วว่ากล่องนี้หยิบตัวอักษรได้ตัว A แน่นอน ดังนั้นกล่องใบนี้จึงไม่มีสารสนเทศเลย สำหรับกล่องที่ 2 มีเอนโทรปี 0.811 บิต เนื่องจากกล่องที่ 2 เราค่อนข้างรู้ในระดับหนึ่งว่า เรามีโอกาสหยิบได้ตัว A เป็นส่วนใหญ่ มีเพียงส่วนน้อยที่จะมีโอกาสหยิบได้ตัว B (โอกาสในการหยิบตัว B คือ 25% หรือก็คือ หยิบตัวอักษรจากกล่องที่ 2 100 ครั้ง มีโอกาส 25 ครั้งที่จะหยิบได้ตัว B) ดังนั้นกล่องใบนี้จึงมีสารสนเทศในระดับหนึ่ง ส่วนกล่องที่ 3 ซึ่งมีเอนโทรปี 1 บิตนั้น เนื่องจากกล่องที่ 3 เรารู้ว่าโอกาสในการหยิบได้ตัวอักษร A และ B มีค่าเท่ากัน หรือมองอีกมุมหนึ่งก็คือ เราไม่รู้เลยว่ากล่องนี้หยิบครั้งต่อไปจะได้ตัวอะไร ความรู้ที่เกี่ยวกับตัวอักษรที่จะหยิบได้จากกล่องที่ 3 นั้นไม่มีเลย ดังนั้นกล่องที่ 3 จึงมีสารสนเทศมากที่สุด

สมมติว่ามีกล่องอีกใบหนึ่ง กล่องใบที่ 4 ข้างในนั้นมีตัว A จำนวน 2 ตัว, ตัว B จำนวน 2 ตัว, ตัว C จำนวน 2 ตัว และ ตัว D จำนวน 2 ตัว โอกาสในการหยิบได้ตัวอักษรแต่ละชนิดที่อยู่ในกล่องที่ 4 มีค่าเท่ากัน ซึ่งคล้ายกับกล่องที่ 3 ต่างกันตรงกล่องที่ 4 มีตัวอักษร 4 ชนิด ในขณะที่กล่องที่ 3 มีตัวอักษรเพียง 2 ชนิด จากค่าเอนโทรปีของกล่องที่ 3 ซึ่งมีค่า 1 บิต จินตนาการว่า ถ้าสารสนเทศ 1 บิตในกล่องที่ 3 ถูกตั้งค่าให้ ‘0’ แทนการหยิบได้ตัว A และ ‘1’ แทนการหยิบได้ตัว B เราจะพบว่าสารสนเทศ 1 บิตในกล่องที่ 3 สามารถแทนทางเลือกในการหยิบตัวอักษรได้ครบทั้ง 2 ชนิด กลับมาที่

กล่องที่ 4 มีตัวอักษร 4 ชนิดอยู่ในกล่อง โอกาสในการหยิบได้ตัวอักษรแต่ละชนิดมีค่าเท่ากัน คำถามก็คือ เอนโทรปีจะมีค่าเป็นเท่าไร โดยหาค่าแบบไม่ใช้สูตรเอนโทรปีช่วย คำตอบก็คือ เราสามารถหาเอนโทรปีของกล่องที่ 4 ได้ โดยการสังเกตค่าเอนโทรปีของกล่องที่ 3 และวิธีการแทนค่าบิตสำหรับแต่ละทางเลือกในการหยิบตัวอักษรของกล่องที่ 3 เนื่องจากกล่องที่ 4 มีตัวอักษรอยู่ในกล่อง 4 ชนิด และมีในสัดส่วนที่เท่าเทียมกัน เราจะต้องใช้ 2 บิต เพื่อรองรับทางเลือกในการหยิบตัวอักษรทั้ง 4 ชนิด เราอาจจะให้ ‘00’ แทนการหยิบได้ตัว A, ‘01’ แทนการหยิบได้ตัว B, ‘10’ แทนการหยิบได้ตัว C และ ‘11’ แทนการหยิบได้ตัว D ได้ ดังนั้นกล่องที่ 4 จึงมีเอนโทรปี 2 บิต เพื่อความน่าเชื่อถือที่มากขึ้น รูปถัดไปจะแสดงการคำนวณหาค่าเอนโทรปีของกล่องที่ 4

$$Entropy(Box4) = -(0.25\log_2 0.25 + 0.25\log_2 0.25 + 0.25\log_2 0.25 + 0.25\log_2 0.25) = 0.5 + 0.5 + 0.5 + 0.5 = 2 \quad (2.5)$$

จากรูปด้านบนแสดงให้เห็นว่า ค่าเอนโทรปีของกล่องที่ 4 ตรงตามที่ได้อธิบายไว้ก่อนหน้านี้จริงๆ

ถ้าหากมีกล่องที่ 5 ข้างในกล่องนี้ มีตัว A จำนวน 4 ตัว, ตัว B จำนวน 2 ตัว, ตัว C จำนวน 1 ตัว และ ตัว D จำนวน 1 ตัว เอนโทรปีของกล่องที่ 5 จะมีค่าน้อยกว่า 2 บิตเพราะในกล่องนี้ สัดส่วนของตัวอักษรทั้ง 4 ชนิดมีไม่เท่ากัน ตัว A มีมากที่สุด ตามมาด้วยตัว B, C และ D ตามลำดับ หากจะให้ประมาณค่าว่าเอนโทรปีของกล่องที่ 5 มีค่าที่แท้จริงเป็นเท่าไรโดยใช้วิธีการแทนค่าบิตเหมือนกับที่ทำในกล่องที่ 3 และ 4 ก็ย่อมลำบาก ครั้นจะแทนค่าบิตของการหยิบได้ตัว A ถึง D ให้มี 2 บิตเท่ากันเหมือนในกล่องที่ 4 ก็ทำได้เช่นกัน แต่ไม่ใช่วิธีที่ดีที่สุด เพราะตัว A มีโอกาสในการถูกหยิบมากกว่าตัว B, C และ D ตัวอย่างวิธีหนึ่งที่จะแทนค่าบิตสำหรับทางเลือกในการหยิบได้ตัวอักษรแต่ละตัวในกล่องที่ 5 คือ ‘0’ แทนการหยิบได้ตัว A, ‘10’ แทนการหยิบได้ตัว B, ‘110’ แทนการหยิบได้ตัว C และ ‘111’ แทนการหยิบได้ตัว D เมื่อพิจารณาวิธีการแทนค่าจากตัวอย่างนี้ จะพบว่าจำนวนบิตโดยเฉลี่ยที่ใช้ในการสร้างสารสนเทศต่อทางเลือกในการหยิบได้ตัวอักษรแต่ละชนิดในกล่องที่ 5 นั้น มีค่าที่น้อยกว่า 2 บิต จริงๆ สังเกตจากการที่ตัว A และ B ถูกแทนค่าด้วยจำนวนบิตที่น้อยกว่าตัว C และ D เมื่อนำมาเฉลี่ยในระยะยาว (อิงจากการหยิบตัวอักษรในกล่องที่ 5 หลายๆ รอบ) ดังนั้นจำนวนบิตที่ใช้ดังกล่าวก็จะมีค่าประมาณ 1 บิตกว่าๆ เท่านั้น รูปถัดไปจะแสดงการคำนวณหาค่าเอนโทรปีของกล่องที่ 5

$$Entropy(Box5) = -(0.5\log_2 0.5 + 0.25\log_2 0.25 + 0.125\log_2 0.125 + 0.125\log_2 0.125) = 0.5 + 0.5 + 0.375 + 0.375 = 1.75 \quad (2.6)$$

เอนโทรปีหรือปริมาณสารสนเทศ ตามที่ได้นำเสนอไปนั้น ถึงแม้ว่าจะเป็นปริมาณที่นิยามขึ้นมาเป็นนามธรรม แต่ก็สามารถใช้ประโยชน์ได้หลากหลาย โดยเฉพาะในด้านการสื่อสาร การส่งข้อมูลระหว่างผู้รับและผู้ส่งนั้น การหาค่าเอนโทรปีสามารถใช้ประมาณจำนวนบิตโดยเฉลี่ยอย่างน้อยที่สุดที่ใช้ในการเข้ารหัสข้อมูลและส่งต่อไปหาผู้รับได้ กล่าวคือ การส่งข้อมูลระหว่างผู้ส่งและผู้รับนั้น จะต้องใช้จำนวนบิตโดยเฉลี่ยในการส่ง ซึ่งมีค่าไม่ต่ำกว่าเอนโทรปีของแหล่งข้อมูลที่จะส่งไป อย่างไรก็ดี การเข้ารหัสเพื่อส่งข้อมูลไม่จำเป็นที่จะต้องทำให้จำนวนบิตโดยเฉลี่ยที่ส่งมีค่าเท่ากับเอนโทรปีก็ได้ เพราะในทางปฏิบัติ อัลกอริทึมในการเข้ารหัส-จัดสรร-แทนค่าบิตให้กับแหล่งข้อมูลที่จะส่งอาจจะไม่ให้ผลลัพธ์จำนวนบิตโดยเฉลี่ยที่ใช้ในการเข้ารหัสที่ดีเท่ากับเอนโทรปีของแหล่งข้อมูลที่จะส่งจริงๆ

2.2.2 เกนความรู้

ในหัวข้อที่แล้วได้อธิบายเกี่ยวกับ เอนโทรปี ซึ่งอธิบายในเชิงหลักการและแนวคิดแบบภาพรวม สำหรับในเรื่องราวของอัลกอริทึม ID3 นั้น ได้มีการใช้แนวคิดที่อยู่บนพื้นฐานของเอนโทรปีเข้ามาเกี่ยวข้องในการที่จะหาแอตทริบิวต์ที่มีความสำคัญที่สุด แนวคิดดังกล่าวคือค่าเกนความรู้ (Information gain)

เกนความรู้ คือส่วนต่างระหว่างเอนโทรปีของเซตของตัวอย่างฝึกฝน (Training examples) กับเอนโทรปีของเซตของตัวอย่างฝึกฝนที่เหลือหลังถูกทดสอบด้วยแอตทริบิวต์ที่ต้องการหาค่าความสำคัญ ส่วนต่างดังกล่าวเป็นค่าที่บ่งบอกว่า เราจะได้รับสารสนเทศ (ได้รับคำตอบ) มากน้อยเพียงใดเมื่อทำการถามหาคำตอบจากแอตทริบิวต์ที่สนใจ คำตอบในที่นี้ก็คือ ความสามารถในการแบ่งเซตของตัวอย่างฝึกฝนออกเป็นเซตย่อย ตามค่าที่เป็นไปได้ของแอตทริบิวต์ที่สนใจ แล้วทำให้เซตย่อยแต่ละเซตที่ได้จากการแบ่งอยู่ในคลาส (Class) เดียวกันให้ได้ทั้งหมด เพราะอัลกอริทึม ID3 จะได้สร้างโหนดคำตอบ ซึ่งตอบคลาสที่ตัวอย่างฝึกฝนอยู่ด้วยกันทั้งหมดในเซตย่อยได้ หากเจอคำตอบที่มีลักษณะนี้สูงในระหว่างการสร้างต้นไม้ตัดสินใจด้วยอัลกอริทึม ID3 ต้นไม้ตัดสินใจที่ได้จะมีความลึกโดยเฉลี่ยไม่มากในการหาคลาสของตัวอย่างที่ไม่เคยฝึกฝนมาก่อน ซึ่งเป็นตัวอย่างที่อยู่ในชุดข้อมูลทดสอบ (Test set) จากความหมายของเกนความรู้ สามารถอธิบายให้อยู่ในรูปสมการเกนความรู้ได้

กำหนดให้ T คือเซตของตัวอย่างฝึกฝน และ a คือแอตทริบิวต์ สมการเกนความรู้ของแอตทริบิวต์ a บนเซตของตัวอย่างฝึกฝน T (หรือเรียกสั้นๆ ว่า เกนความรู้ของแอตทริบิวต์ a) คือ

$$IG(T, a) = IG(a) = Entropy(T) - Remainder(T|a) \tag{2.7}$$

โดยที่ IG(T, a) หรือ IG(a) คือเกนความรู้ของแอตทริบิวต์ a,

Entropy(T) คือ เอนโทรปีของเซตของตัวอย่างฝึกฝน T จากสมการรูปทั่วไปของเอนโทรปีที่ได้อธิบายในหัวข้อก่อนหน้านี้ เอนโทรปีของเซตของตัวอย่างฝึกฝน T จะมีการเปลี่ยนแปลง

ตัวอักษรของตัวแปรในสมการ รวมไปถึงเปลี่ยนแปลงความหมายของตัวแปรที่เกี่ยวข้องเล็กน้อย เพื่อให้สอดคล้องกับเอนโทรปีในบริบทของค่าเกนความรู้จริงๆ ฉะนั้นสมการเอนโทรปีของเซตของตัวอย่างฝึกฝน T จะมีลักษณะดังสมการถัดไปที่จะได้เห็นต่อไปนี้

$$Entropy(T) = -\sum_{c \in C} p(c) \log_2 p(c) \tag{2.8}$$

โดยที่ C คือเซตของคลาสของตัวอย่างฝึกฝน,

c คือคลาสที่อยู่ในเซต C,

p(c) คือความน่าจะเป็นที่ตัวอย่างจากคลาส c ซึ่งอยู่ในเซตของตัวอย่างฝึกฝนจะปรากฏ

สำหรับ Remainder(T | a) นั้นเป็นเอนโทรปีของเซตของตัวอย่างฝึกฝน T ที่เหลือหลังถูกทดสอบด้วยแอตทริบิวต์ a ซึ่งรูปสมการจะมีความแตกต่างไปจากสมการเอนโทรปีของเซตของตัวอย่างฝึกฝน T กล่าวคือ ค่า Remainder(T | a) จะมีการทดสอบเซตของตัวอย่างฝึกฝนด้วยแอตทริบิวต์ a แบบล่วงหน้า เพื่อดูว่า ถ้าแบ่งเซตของตัวอย่างฝึกฝนตามค่าที่เป็นไปได้ของแอตทริบิวต์ a ออกเป็นเซตย่อย เซตย่อยแต่ละเซตที่แบ่งได้ จะมีค่าเอนโทรปีเป็นอย่างไร เซตย่อยใดที่ตัวอย่างภายในเซตอยู่คลาสเดียวกันทั้งหมด ค่าเอนโทรปีก็จะไม่มีเลย ในขณะที่เซตย่อยใดที่ตัวอย่างภายในเซตไม่ได้อยู่ในคลาสเดียวกัน ค่าเอนโทรปีก็จะมีอยู่ แต่จะมีค่ามากหรือน้อยขึ้นอยู่กับสัดส่วนของคลาสที่ปรากฏในตัวอย่างที่อยู่ในเซตย่อยว่าจะเป็นอย่างไร ผลรวมของเอนโทรปีของเซตย่อยแต่ละเซตที่สังกัดในแต่ละค่าของแอตทริบิวต์ a ก็คือเอนโทรปีของเซตของตัวอย่างฝึกฝน T ที่เหลือหลังถูกทดสอบด้วยแอตทริบิวต์ a นั่นเอง เอนโทรปีที่เหลือดังกล่าวเป็นค่าที่สามารถบอกได้ว่า หากอัลกอริทึม ID3 เลือกที่จะเอาแอตทริบิวต์ a เป็นแอตทริบิวต์ของโหนดตัดสินใจที่กำลังพิจารณาอยู่ เราจะต้องใช้สารสนเทศอีกเท่าไร เราถึงจะได้รับคำตอบที่ต้องการโดยสมบูรณ์ (คือความสามารถในการแบ่งเซตของตัวอย่างฝึกฝนตามค่าที่เป็นไปได้ของแอตทริบิวต์ ให้อยู่ในคลาสเดียวกันให้ได้ทั้งหมดทุกเซตย่อย ดังที่ได้อธิบายก่อนหน้า) สมการ Remainder(T | a) มีลักษณะดังสมการด้านล่างที่จะแสดงนี้

$$Remainder(T|a) = \sum_{v \in V} \frac{|T_v|}{|T|} \cdot Entropy(T_v) \tag{2.9}$$

โดยที่ V คือเซตของค่าที่เป็นไปได้ของแอตทริบิวต์ a,

v คือ ค่าที่อยู่ในเซต V,

T_v คือเซตของตัวอย่างฝึกฝนที่มีค่าของแอตทริบิวต์ a เป็น v

เนื่องจาก Entropy(T_v) ซึ่งเป็นเอนโทรปีของเซตย่อย T_v ที่สังกัดในค่า v ของแอตทริบิวต์ a ทำการหาค่าเอนโทรปีโดยใช้เซตย่อย T_v ซึ่งถูกแบ่งมาจากเซตของตัวอย่างฝึกฝนหลัก T หากนำไปหาผลรวมกับค่าเอนโทรปีของเซตย่อยอื่นๆ ที่ถูกแบ่งด้วยกันเลย ก็อาจจะทำให้ผลรวมของค่า

เอนโทรปีที่ได้มีค่ามากกว่าเอนโทรปีของเซตตัวอย่างฝึกฝนหลัก ซึ่งยังไม่ใช่ค่าของ Remainder(T | a) ที่แท้จริง เพราะค่า Remainder(T | a) จะคิดโดยใช้หลักการแบ่งเซตของตัวอย่างฝึกฝน T ออกเป็นเซตย่อยตามค่าที่เป็นไปได้ของแอตทริบิวต์ a และทำการหาค่าเอนโทรปีของเซตย่อยเหล่านั้น แต่ทว่าค่าเอนโทรปีที่หาได้ จะยังไม่สามารถนำไปหาผลรวมได้ทันที เพราะค่าเอนโทรปีของเซตย่อยแต่ละเซตนั้น เกี่ยวข้องกับสัดส่วนของจำนวนตัวอย่างภายในเซตของตัวอย่างฝึกฝนหลักที่มีค่าของแอตทริบิวต์ a เป็นค่าที่เซตย่อยแต่ละเซตสังกัดอีกด้วย ดังนั้นค่าเอนโทรปีของแต่ละเซตย่อยที่หาได้ จึงต้องนำไปคูณกับสัดส่วนดังกล่าว เพื่อให้มีความแตกต่างชัดเจนจากเอนโทรปีของเซตของตัวอย่างฝึกฝนหลัก และเพื่อแสดงให้เห็นถึงค่าเอนโทรปีของเซตย่อยแต่ละเซต ว่าเป็นการหามาจากเซตย่อยที่ถูกแบ่งออกมาจากเซตของตัวอย่างฝึกฝนหลักจริงๆ ค่าเอนโทรปีของเซตย่อยแต่ละเซตที่ผ่านการคูณกับสัดส่วนที่ได้กล่าวถึงไป สามารถที่จะนำไปหาผลรวมต่อได้ทันที นั่นคือเหตุผลว่าทำไมสมการ Remainder(T | a) จึงมีรูปสมการเป็นดังที่ได้แสดงไป

เพื่อความเข้าใจที่มากขึ้นเกี่ยวกับค่าเกนความรู้ วิทยานิพนธ์นี้จะยกตัวอย่างการคำนวณหาค่าเกนความรู้ เพื่อหาแอตทริบิวต์ที่มีค่าเกนความรู้มากที่สุด ซึ่งในมุมมองอัลกอริทึม ID3 ก็คือ แอตทริบิวต์ที่มีความสำคัญที่สุด โดยใช้ชุดข้อมูลเล่นเทนนิส (Play tennis) เป็นเซตของตัวอย่างฝึกฝนที่จะใช้ในตัวอย่างการคำนวณนี้

ชุดข้อมูลเล่นเทนนิส เป็นชุดข้อมูลที่นำเสนอรายละเอียดเกี่ยวกับ การที่จะเล่นหรือไม่เล่นเทนนิส (Play) โดยขึ้นอยู่กับเงื่อนไขของสภาพอากาศ ซึ่งประกอบไปด้วย ชนิดสภาพอากาศ (Outlook), อุณหภูมิ (Temperature), ความชื้น (Humidity) และลม (Wind) ตารางถัดไปแสดงรายละเอียดตัวอย่าง (Example) ทั้ง 14 ตัวอย่างที่มีในชุดข้อมูลเล่นเทนนิส

ตารางที่ 2.1 ชุดข้อมูลเล่นเทนนิส

Outlook	Temperature	Humidity	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rainy	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

จากตารางข้างบน ชุดข้อมูลเล่นเทนนิส มีแอตทริบิวต์ 4 แอตทริบิวต์ คือ Outlook, Temperature, Humidity และ Wind ส่วน Play เป็นตัวบอกคลาสของตัวอย่างแต่ละตัวอย่างว่าจะเล่นเทนนิสหรือไม่

ต่อไปจะเริ่มทำการคำนวณหาค่าเกนความรู้ของทุกแอตทริบิวต์ โดยจะแสดงวิธีทำแบบเต็มกับแอตทริบิวต์ Outlook เท่านั้น สำหรับแอตทริบิวต์อื่นๆ จะแสดงคำตอบที่หาได้เลย

กำหนดให้ T คือเซตของตัวอย่างฝึกฝน อันประกอบไปด้วยตัวอย่างที่อยู่ในชุดข้อมูลเล่นเทนนิสทั้งหมด

จากสมการ (2.5) สมการเกนความรู้ของแอตทริบิวต์ a เพื่อที่จะหาค่าเกนความรู้ของแอตทริบิวต์ Outlook ก่อน ทำการแทนตัวแปร a ด้วยแอตทริบิวต์ Outlook จะได้

$$IG(T, Outlook) = Entropy(T) - Remainder(T|Outlook) \quad (2.10)$$

ทำการหาค่าของแต่ละพจน์ในสมการเกนความรู้ของแอตทริบิวต์ Outlook เริ่มจาก Entropy(T) ซึ่งพจน์นี้จะใช้ร่วมกันในการหาค่าเกนความรู้ของทุกแอตทริบิวต์ (ในขณะที่พจน์ Remainder เป็นพจน์ที่หาค่าเจาะจงของแต่ละแอตทริบิวต์จริงๆ)

เซตของตัวอย่างฝึกฝน T เซตนี้ มีตัวอย่างทั้งหมด 14 ตัวอย่าง จากตัวอย่างทั้งหมดภายในเซต มีตัวอย่างที่คลาสเป็น Yes (จะเล่นเทนนิส) อยู่ 9 ตัวอย่าง และมีตัวอย่างที่คลาสเป็น No

(จะไม่เล่นเทนนิส) อยู่ 5 ตัวอย่าง ทำการหาความน่าจะเป็นที่ตัวอย่างจากแต่ละคลาส ซึ่งอยู่ในเซต T จะปรากฏ จะได้

ความน่าจะเป็นที่ตัวอย่างจากคลาส Yes ซึ่งอยู่ในเซต T จะปรากฏ p(Yes) = 9/14

ความน่าจะเป็นที่ตัวอย่างจากคลาส No ซึ่งอยู่ในเซต T จะปรากฏ p(No) = 5/14

ทำการแตกสมการ Entropy(T) ให้อยู่ในรูปผลรวมของการบวกที่เกี่ยวข้องกับคลาส Yes และ No จะได้

$$Entropy(T) = -(p(Yes)\log_2 p(Yes) + p(No)\log_2 p(No)) \tag{2.11}$$

แทนค่า p(Yes) และ p(No) ที่คำนวณไว้ลงในสมการ Entropy(T) ที่แตกออกมา จะได้

$$Entropy(T) = -\left(\frac{9}{14}\log_2\frac{9}{14} + \frac{5}{14}\log_2\frac{5}{14}\right) \approx 0.94 \tag{2.12}$$

ขั้นต่อไป ทำการคำนวณหาค่าของพจน์ Remainder(T | Outlook)

แอตทริบิวต์ Outlook มีค่าที่เป็นไปได้ 3 ค่า คือ Sunny, Overcast และ Rainy ในบรรดาตัวอย่างทั้ง 14 ตัวอย่างที่อยู่ในเซต T มีตัวอย่างที่ค่าของแอตทริบิวต์ Outlook เป็น Sunny อยู่ 5 ตัวอย่าง มีตัวอย่างที่ค่าของแอตทริบิวต์ Outlook เป็น Overcast อยู่ 4 ตัวอย่าง และมีตัวอย่างที่ค่าของแอตทริบิวต์ Outlook เป็น Rainy อยู่ 5 ตัวอย่าง ทำการแบ่งเซต T ออกเป็น 3 เซตย่อย คือ T_{Sunny}, $T_{Overcast}$ และ T_{Rainy} แต่ละเซตย่อยจะมีตัวอย่างที่มีค่าของแอตทริบิวต์ Outlook ตรงตามที่เซตย่อยแต่ละเซตกำหนด ดังนั้น จากการนับจำนวนตัวอย่างในเซต T ที่อยู่ในแต่ละค่าของแอตทริบิวต์ Outlook เราสามารถบอกได้ว่า เซต T_{Sunny} มีตัวอย่างทั้งหมด 5 ตัวอย่าง เซต $T_{Overcast}$ มีตัวอย่างทั้งหมด 4 ตัวอย่าง และเซต T_{Rainy} มีตัวอย่างทั้งหมด 5 ตัวอย่าง ทำการแตกสมการ Remainder(T | Outlook) ให้อยู่ในรูปผลรวมของการบวกที่เกี่ยวข้องกับค่าที่เป็นไปได้ของแอตทริบิวต์ Outlook จะได้

$$Remainder(T|Outlook) = \left(\frac{|T_{Sunny}|}{|T|}\cdot Entropy(T_{Sunny})\right) + \left(\frac{|T_{Overcast}|}{|T|}\cdot Entropy(T_{Overcast})\right) + \left(\frac{|T_{Rainy}|}{|T|}\cdot Entropy(T_{Rainy})\right) \tag{2.13}$$

จากสมการ Remainder(T | Outlook) ที่แตกออกมา จะพบว่าฝั่งขวาของสมการประกอบไปด้วยผลรวมของการบวก 3 พจน์ ฉะนั้น เราจะทำการคำนวณหาค่าของทั้ง 3 พจน์ โดยไล่

หาค่าตั้งแต่พจน์ของ T_{Sunny} ก่อน ไปจนถึงพจน์ของ T_{Rainy} จากนั้นเราจึงจะทำการนำค่าที่หาได้จากทั้ง 3 พจน์มาบวกด้วยกันเพื่อหาผลรวมต่อไป

เซต T_{Sunny} มีตัวอย่างทั้งหมด 5 ตัวอย่าง มีตัวอย่างที่คลาสเป็น Yes อยู่ 2 ตัวอย่าง และมีตัวอย่างที่คลาสเป็น No อยู่ 3 ตัวอย่าง ทำการหาความน่าจะเป็นที่ตัวอย่างจากแต่ละคลาส ซึ่งอยู่ในเซต T_{Sunny} จะปรากฏ จะได้

ความน่าจะเป็นที่ตัวอย่างจากคลาส Yes ซึ่งอยู่ในเซต T_{Sunny} จะปรากฏ $p(Yes_{Sunny}) = 2/5$

ความน่าจะเป็นที่ตัวอย่างจากคลาส No ซึ่งอยู่ในเซต T_{Sunny} จะปรากฏ $p(No_{Sunny}) = 3/5$

ทำการแตก Entropy(T_{Sunny}) ให้อยู่ในรูปผลรวมของการบวกที่เกี่ยวข้องกับคลาส Yes และ No จะได้

$$Entropy(T_{Sunny}) = -\left(p(Yes_{Sunny}) \log_2 p(Yes_{Sunny}) + p(No_{Sunny}) \log_2 p(No_{Sunny})\right) \quad (2.14)$$

แทนค่า $p(Yes_{Sunny})$ และ $p(No_{Sunny})$ ที่คำนวณไว้ลงในสมการ Entropy(T_{Sunny}) ที่แตกออกมา จะได้

$$Entropy(T_{Sunny}) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) \approx 0.971 \quad (2.15)$$

ทำการหาสัดส่วนของจำนวนตัวอย่างในเซต T_{Sunny} เทียบกับจำนวนตัวอย่างในเซต T จะได้

$$\frac{|T_{Sunny}|}{|T|} = \frac{5}{14} \quad (2.16)$$

แทนค่า Entropy(T_{Sunny}) และสัดส่วนของจำนวนตัวอย่างในเซต T_{Sunny} เทียบกับจำนวนตัวอย่างในเซต T ที่คำนวณได้ กลับลงไปในสมการ Remainder(T | Outlook) จะได้

$$Remainder(T|Outlook) = \left(\frac{5}{14} \cdot 0.971\right) + \left(\frac{|T_{Overcast}|}{|T|} \cdot Entropy(T_{Overcast})\right) + \left(\frac{|T_{Rainy}|}{|T|} \cdot Entropy(T_{Rainy})\right) \quad (2.17)$$

ต่อไปจะทำการคำนวณหาค่าในพจน์ของ $T_{Overcast}$

เซต $T_{Overcast}$ มีตัวอย่างทั้งหมด 4 ตัวอย่าง มีตัวอย่างที่คลาสเป็น Yes อยู่ 4 ตัวอย่าง และมีตัวอย่างที่คลาสเป็น No อยู่ 0 ตัวอย่าง ทำการหาความน่าจะเป็นที่ตัวอย่างจากแต่ละคลาส ซึ่งอยู่ในเซต $T_{Overcast}$ จะปรากฏ จะได้

ความน่าจะเป็นที่ตัวอย่างจากคลาส Yes ซึ่งอยู่ในเซต $T_{Overcast}$ จะปรากฏ $p(Yes_{Overcast}) = 4/4 = 1$

ความน่าจะเป็นที่ตัวอย่างจากคลาส No ซึ่งอยู่ในเซต $T_{Overcast}$ จะปรากฏ $p(No_{Overcast}) = 0/4 = 0$

ทำการแตก Entropy($T_{Overcast}$) ให้อยู่ในรูปผลรวมของการบวกที่เกี่ยวข้องกับคลาส Yes และ No จะได้

$$Entropy(T_{Overcast}) = -\begin{pmatrix} p(Yes_{Overcast}) \log_2 p(Yes_{Overcast}) + \\ p(No_{Overcast}) \log_2 p(No_{Overcast}) \end{pmatrix} \quad (2.18)$$

แทนค่า $p(Yes_{Overcast})$ และ $p(No_{Overcast})$ ที่ได้คำนวณไว้ลงในสมการ Entropy($T_{Overcast}$) ที่แตกออกมา จะได้

$$Entropy(T_{Overcast}) = -(1 \log_2 1 + 0 \log_2 0) = 0 \quad (2.19)$$

ในการคำนวณ Entropy($T_{Overcast}$) สังเกตว่ามีพจน์หลังซึ่งเป็นลอการิทึมฐาน 2 ของ 0 ด้วย พจน์นี้ไม่สามารถหาค่าได้ เนื่องจากไม่มีจำนวนจริงใดที่สามารถนำมาเป็นเลขชี้กำลัง แล้วดำเนินการนำฐานมายกกำลังกับเลขชี้กำลังเพื่อให้ค่า 0 ได้ ดังนั้น พจน์นี้จึงละเว้นไป การคำนวณ Entropy($T_{Overcast}$) จึงเหลือแต่ในส่วนของพจน์หน้าเท่านั้น

ทำการหาสัดส่วนของจำนวนตัวอย่างในเซต $T_{Overcast}$ เทียบกับจำนวนตัวอย่างในเซต T จะได้

$$\frac{|T_{Overcast}|}{|T|} = \frac{4}{14} \quad (2.20)$$

แทนค่า Entropy($T_{Overcast}$) และสัดส่วนของจำนวนตัวอย่างในเซต $T_{Overcast}$ เทียบกับจำนวนตัวอย่างในเซต T ที่คำนวณได้ กลับลงไปในสมการ Remainder(T | Outlook) จะได้

$$Remainder(T|Outlook) = \left(\frac{5}{14} \cdot 0.971\right) + \left(\frac{4}{14} \cdot 0\right) + \left(\frac{|T_{Rainy}|}{|T|} \cdot Entropy(T_{Rainy})\right) \quad (2.21)$$

ต่อไปจะทำการคำนวณหาค่าในพจน์ของ T_{Rainy}

เซต T_{Rainy} มีตัวอย่างทั้งหมด 5 ตัวอย่าง มีตัวอย่างที่คลาสเป็น Yes อยู่ 3 ตัวอย่าง และมีตัวอย่างที่คลาสเป็น No อยู่ 2 ตัวอย่าง ทำการหาความน่าจะเป็นที่ตัวอย่างจากแต่ละคลาส ซึ่งอยู่ในเซต T_{Rainy} จะปรากฏ จะได้

ความน่าจะเป็นที่ตัวอย่างจากคลาส Yes ซึ่งอยู่ในเซต T_{Rainy} จะปรากฏ $p(Yes_{Rainy}) = 3/5$

ความน่าจะเป็นที่ตัวอย่างจากคลาส No ซึ่งอยู่ในเซต T_{Rainy} จะปรากฏ $p(No_{Rainy}) = 2/5$

ทำการแตก Entropy(T_{Rainy}) ให้อยู่ในรูปผลรวมของการบวกที่เกี่ยวข้องกับคลาส Yes และ No จะได้

$$Entropy(T_{Rainy}) = -\left(p(Yes_{Rainy})\log_2 p(Yes_{Rainy}) + p(No_{Rainy})\log_2 p(No_{Rainy})\right) \quad (2.22)$$

แทนค่า $p(Yes_{Rainy})$ และ $p(No_{Rainy})$ ที่คำนวณไว้ลงในสมการ Entropy(T_{Rainy}) ที่แตกออกมา จะได้

$$Entropy(T_{Rainy}) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) \approx 0.971 \quad (2.23)$$

ทำการหาสัดส่วนของจำนวนตัวอย่างในเซต T_{Rainy} เทียบกับจำนวนตัวอย่างในเซต T จะได้

$$\frac{|T_{Rainy}|}{|T|} = \frac{5}{14} \quad (2.24)$$

แทนค่า Entropy(T_{Rainy}) และสัดส่วนของจำนวนตัวอย่างในเซต T_{Rainy} เทียบกับจำนวนตัวอย่างในเซต T ที่คำนวณได้ กลับลงไปในสมการ Remainder(T | Outlook) และหาผลรวมของการบวกทั้ง 3 พจน์ จะได้

$$Remainder(T|Outlook) = \left(\frac{5}{14}\cdot 0.971\right) + \left(\frac{4}{14}\cdot 0\right) + \left(\frac{5}{14}\cdot 0.971\right) \approx 0.694 \quad (2.25)$$

ตอนนี้เราคำนวณค่าของ Entropy(T) และ Remainder(T | Outlook) เสร็จเรียบร้อยแล้ว ขั้นต่อไปทำการคำนวณหาเกนความรู้ของแอตทริบิวต์ Outlook จะได้

$$IG(T, Outlook) = 0.94 - 0.694 = 0.246 \quad (2.26)$$

ทำการคำนวณหาเกนความรู้ของแอตทริบิวต์ทั้ง 3 ที่เหลือ จะได้

IG(T, Temperature) = 0.029

IG(T, Humidity) = 0.151

IG(T, Wind) = 0.048

จากค่าเกนความรู้ของแอตทริบิวต์ทั้ง 4 ของชุดข้อมูลเล่นเทนนิสที่หาได้ เราจะพบว่าแอตทริบิวต์ Outlook เป็นแอตทริบิวต์ที่มีค่าเกนความรู้มากที่สุด หรือเป็นแอตทริบิวต์ที่มีความสำคัญที่สุดตามมุมมองของอัลกอริทึม ID3 ดังนั้น ตัวอัลกอริทึม ID3 ก็จะทำการเลือกแอตทริบิวต์ Outlook ไปเป็นแอตทริบิวต์ของโหนดตัดสินใจที่กำลังพิจารณา

2.2.3 ขั้นตอนการทำงานของอัลกอริทึม ID3

อัลกอริทึม ID3 ต้องการอินพุต 2 อย่างคือ เซตของตัวอย่างฝึกฝน และเซตของแอตทริบิวต์ของตัวอย่างฝึกฝน ผลลัพธ์ที่ได้จากการทำงานของอัลกอริทึม ID3 ก็คือ ได้โหนดตัดสินใจ โดยภายในบริบทหลักของอัลกอริทึม ID3 ที่ทำงาน จะมีลักษณะการทำงานในรูปแบบที่มีการเรียกใช้ตัวเอง (Recursion) หรือมีการเรียกใช้อัลกอริทึม ID3 ในบริบทรอง เพื่อที่จะทำการสร้างโหนดที่เป็นลูกหลานของโหนดตัดสินใจที่สร้างจากอัลกอริทึม ID3 ในบริบทหลัก และประกอบรวมกันจนเป็นต้นไม้ตัดสินใจที่พร้อมจะนำไปใช้ทำนายคลาสของตัวอย่างที่ไม่เคยฝึกฝนมาก่อน ซึ่งเป็นตัวอย่างที่อยู่ในชุดข้อมูลทดสอบ ลักษณะการเรียกใช้ตัวเองของอัลกอริทึม ID3 นี้ สามารถเกิดขึ้นกับบริบทรองและบริบทรองอื่นๆ ที่อยู่ในระดับล่างลงไปได้เรื่อยๆ เช่นกัน ดังนั้นอินพุตของอัลกอริทึม ID3 ทั้ง 2 สามารถที่จะอยู่ในรูปของเซตย่อยได้ ขั้นตอนการทำงานของอัลกอริทึม ID3 มีดังนี้ กำหนดให้ T คือเซตของตัวอย่างฝึกฝนที่รับค่าเข้ามา, A คือเซตของแอตทริบิวต์ของตัวอย่างฝึกฝนที่รับค่าเข้ามา, Tc คือเซตที่เก็บสำเนาของตัวอย่างทั้งหมดของเซต T และ selAttr คือ แอตทริบิวต์ที่อัลกอริทึม ID3 เลือก

1) ทำการคำนวณค่าเกนความรู้ของทุกแอตทริบิวต์ที่อยู่ในเซต A โดยใช้สมการ (2.5) – (2.7) ร่วมกับเซต T
2) ทำการเลือกแอตทริบิวต์ในเซต A ที่มีความสำคัญที่สุด ซึ่งเป็นแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดเพียงแอตทริบิวต์เดียว ในกรณีที่มีแอตทริบิวต์ที่มีความสำคัญที่สุดอย่างน้อย 2 แอตทริบิวต์ ซึ่งต้องมีค่าเกนความรู้สูงที่สุดและเท่ากัน ให้ทำการสุ่มเลือก 1 แอตทริบิวต์จากแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันทั้งหมด นำแอตทริบิวต์ที่เลือกได้จากขั้นตอนนี้ไปเก็บไว้ในตัวแปร selAttr
3) ทำการสร้างโหนดตัดสินใจ แล้วใช้แอตทริบิวต์ selAttr เป็นแอตทริบิวต์บนโหนดตัดสินใจที่สร้าง

4) นำแอตทริบิวต์ที่ถูกเลือก selAttr ออกจากเซต A ; A = A – selAttr เนื่องจากจะไม่มีการนำแอตทริบิวต์ที่เคยเลือกไปแล้วมาพิจารณาซ้ำในการสร้างโหนดตัดสินใจอื่นๆ ที่อยู่ในระดับลูกหลานของโหนดตัดสินใจปัจจุบันที่เพิ่งสร้างไป
5) ทำการคัดลอกตัวอย่างภายในเซต T ทั้งหมดไปใส่ในเซต Tc
6) แบ่งเซต T ตามค่าที่เป็นไปได้ของแอตทริบิวต์ selAttr จะได้เซตย่อย k เซต ; $t_1 – t_k$ โดยที่ k คือจำนวนค่าที่เป็นไปได้ของแอตทริบิวต์ selAttr, $v_1 – v_k$ คือค่าที่เป็นไปได้ของแอตทริบิวต์ selAttr และ $t_1 – t_k$ คือเซตย่อยที่ตัวอย่างฝึกฝนทุกตัวอย่างภายในเซต มีค่าของแอตทริบิวต์ selAttr เป็น $v_1 – v_k$ ตามลำดับ
7) ทำการแตกกิ่ง (Branch) ออกมาจากโหนดตัดสินใจปัจจุบันที่สร้าง จำนวน k กิ่ง แต่ละกิ่งระบุว่าเป็นกิ่งของค่าในแอตทริบิวต์ selAttr ค่าใด ไล่ตั้งแต่ กิ่งแรกระบุเป็นกิ่งของค่า v_1 ไปจนถึงกิ่งสุดท้ายระบุเป็นกิ่งของค่า v_k
8) ทำการตรวจสอบเซตย่อยแต่ละเซตที่แบ่งได้ กำหนดให้ t_i คือเซตย่อยที่กำลังพิจารณา และ i มีค่าไล่ตามลำดับตั้งแต่ 1 ไปถึง k
    8.1) ถ้าเซตย่อย t_i ไม่มีตัวอย่างใดๆ อยู่ในเซตเลย ให้ทำการพิจารณาตัวอย่างภายในเซต Tc ดูว่าภายในเซต Tc ตัวอย่างส่วนใหญ่มีสัดส่วนอยู่ในคลาสใดมากที่สุด ให้ทำการสร้างโหนดคำตอบที่ตอบคลาสดังกล่าว แล้วนำโหนดคำตอบที่สร้างเชื่อมกับโหนดตัดสินใจปัจจุบันที่ได้สร้างไป ผ่านกิ่งของค่า v_i หากเงื่อนไขในขั้นตอนนี้ไม่เป็นจริง ให้ทำการตรวจสอบเซตย่อย t_i ตามเงื่อนไขในขั้นตอนถัดไป
    8.2) ถ้าตัวอย่างภายในเซตย่อย t_i อยู่ในคลาสเดียวกันทั้งหมด ให้ทำการสร้างโหนดคำตอบที่ตอบคลาสดังกล่าว แล้วนำโหนดคำตอบที่สร้างเชื่อมกับโหนดตัดสินใจปัจจุบันที่ได้สร้างไป ผ่านกิ่งของค่า v_i หากเงื่อนไขในขั้นตอนนี้ไม่เป็นจริง ให้ทำการตรวจสอบเซตย่อย t_i ตามเงื่อนไขในขั้นตอนถัดไป
    8.3) ถ้าเซต A ไม่มีแอตทริบิวต์ใดๆ เหลืออยู่ในเซตเลย ส่งผลให้อัลกอริทึม ID3 ไม่สามารถดำเนินการเลือกแอตทริบิวต์ ตามเงื่อนไขของตัวอัลกอริทึมที่กำหนดต่อไปได้อีก ทำให้ไม่สามารถสร้างโหนดตัดสินใจได้ และเมื่อสร้างโหนดตัดสินใจไม่ได้ ตัวอัลกอริทึมเองก็ย่อมไม่สามารถที่จะแบ่งเซตย่อย t_i ต่อไปได้อีกเช่นกัน หากเงื่อนไขในขั้นตอนนี้เป็นจริง ให้ทำการพิจารณาตัวอย่างภายในเซตย่อย t_i ดูว่าภายในเซตย่อย t_i ตัวอย่างส่วนใหญ่มีสัดส่วนอยู่ในคลาสใดมากที่สุด ให้ทำการสร้าง

โหนดคำตอบที่ตอบคลาสดังกล่าว แล้วนำโหนดคำตอบที่สร้างเชื่อมกับโหนดตัดสินใจปัจจุบันที่ได้สร้างไป ผ่านกิ่งของค่า v_i หากเงื่อนไขในขั้นตอนนี้ไม่เป็นจริง ให้ทำตามขั้นตอนสุดท้ายของการตรวจสอบเซตย่อย t_i ได้เลย

8.4) เนื่องจากตัวอย่างภายในเซตย่อย t_i ไม่ได้อยู่ในคลาสเดียวกัน และยังคงมีแอตทริบิวต์ในเซต A เหลืออยู่

8.4.1) ทำการคัดลอกเซต A จะได้เซตสำเนาของเซต A

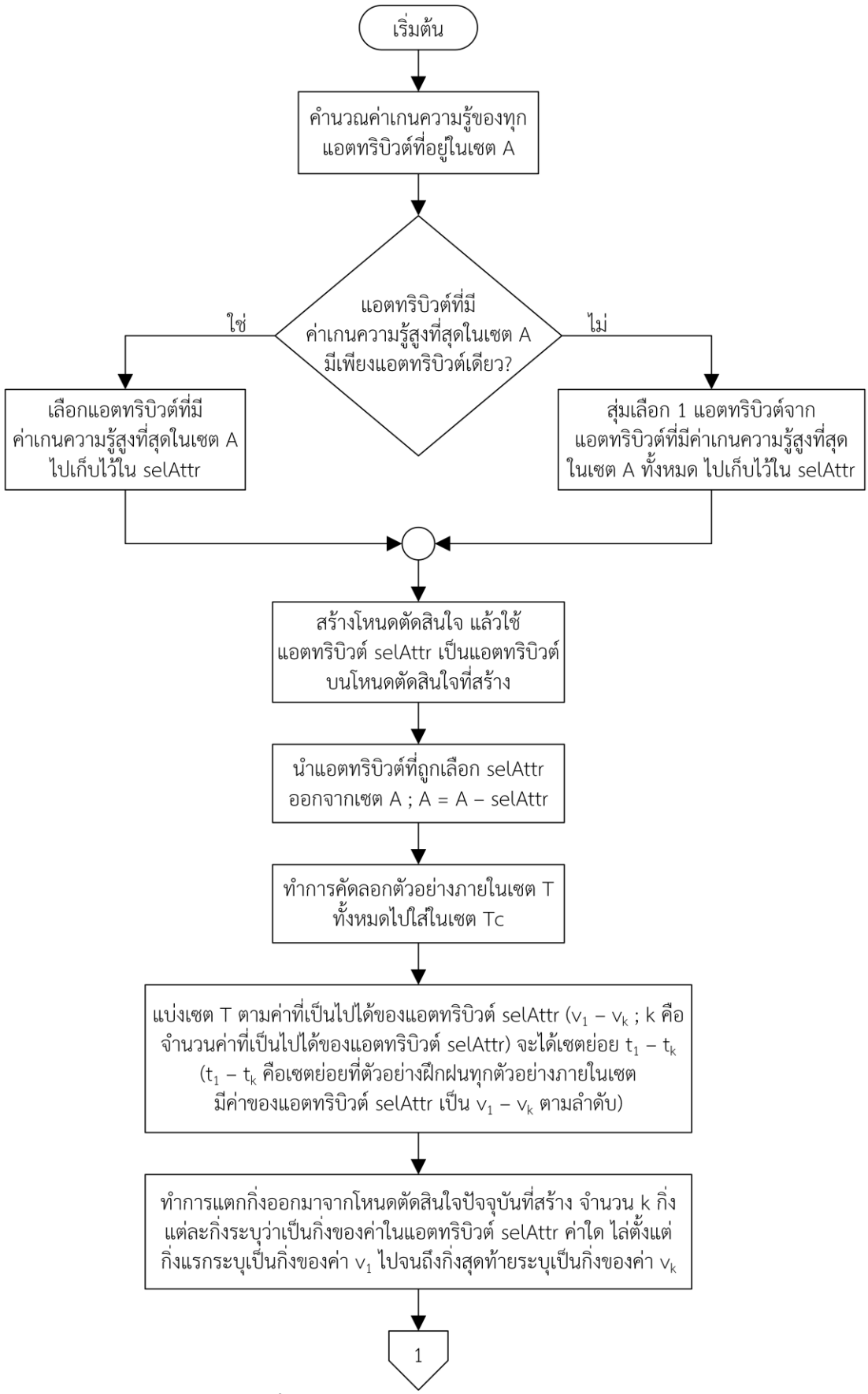
8.4.2) ทำการเรียกใช้งานอัลกอริทึม ID3 ในบริบทรองที่แยกต่างหากเลย (ไม่ใช่การวนกลับไปทำขั้นตอนที่ 1 ในบริบทของอัลกอริทึม ID3 ปัจจุบัน แต่เป็นการเรียกใช้งานอัลกอริทึม ID3 ในบริบทรอง ซึ่งกระทำกับอินพุตที่ต่างกัน บริบทรองนี้จะไม่ทับกับบริบทปัจจุบัน ดังนั้นการทำงานของอัลกอริทึม ID3 ในบริบทปัจจุบันยังคงสามารถกลับมาดำเนินการต่อได้ ถ้าอัลกอริทึม ID3 ในบริบทรองทำงานเสร็จสิ้นและมีการคืนค่าผลลัพธ์โหนดตัดสินใจที่สร้างกลับมา) โดยการเรียกใช้งานอัลกอริทึม ID3 ในบริบทรองนี้ ให้ส่งค่าเซตย่อย t_i ไปเป็นอินพุตเซตของตัวอย่างฝึกฝน และส่งค่าเซตสำเนาของเซต A ที่ได้จากการคัดลอกในขั้นตอนที่แล้วไปเป็นอินพุตเซตของแอตทริบิวต์ของตัวอย่างฝึกฝน

8.4.3) รออัลกอริทึม ID3 ในบริบทรองทำงานเสร็จสิ้นและมีการคืนค่าผลลัพธ์โหนดตัดสินใจกลับมา

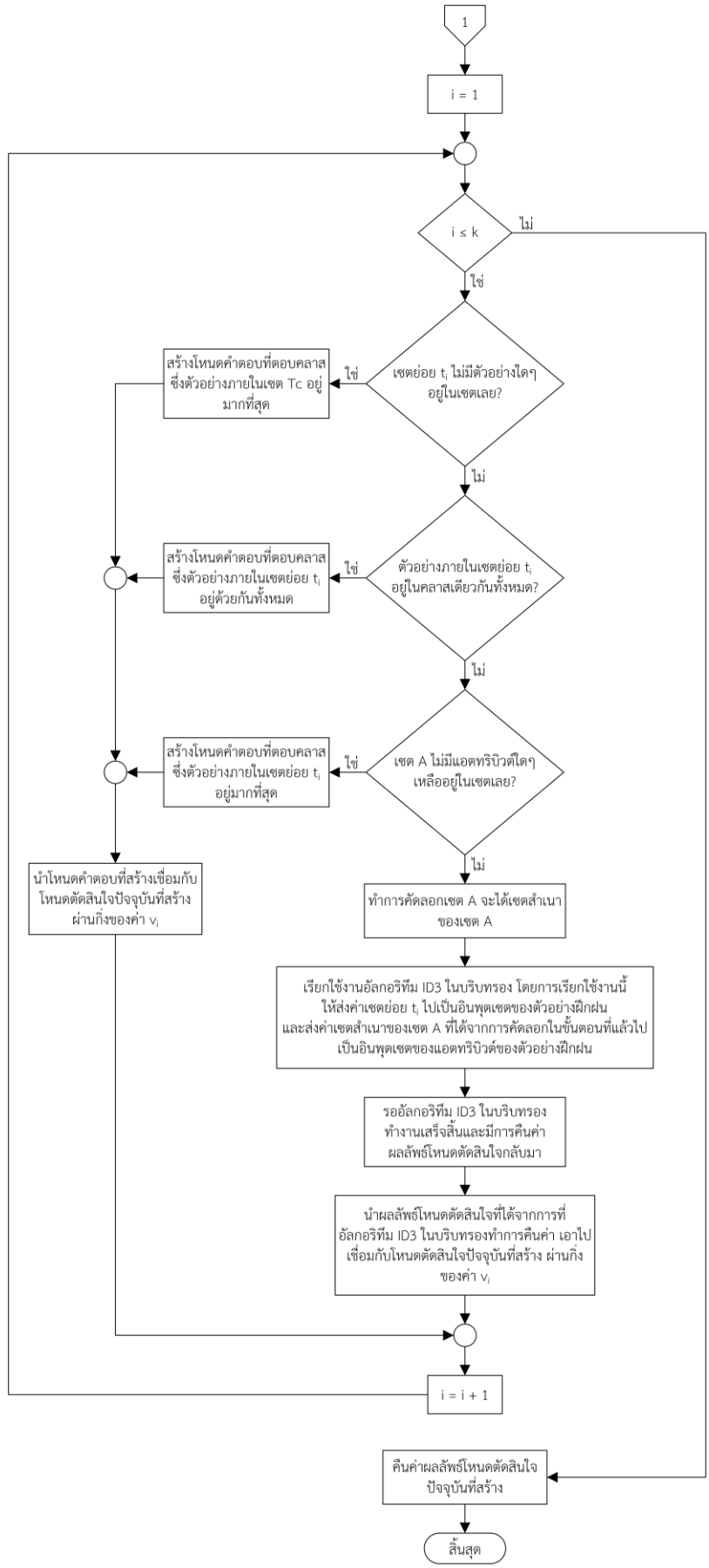
8.4.4) นำผลลัพธ์โหนดตัดสินใจที่ได้จากการที่อัลกอริทึม ID3 ในบริบทรองทำการคืนค่า (Return) เอาไปเชื่อมกับโหนดตัดสินใจปัจจุบันที่สร้าง ผ่านกิ่งของค่า v_i

9) ทำการคืนค่าผลลัพธ์โหนดตัดสินใจปัจจุบันที่สร้าง และจบการทำงานของอัลกอริทึม ID3 ในบริบทปัจจุบัน

ผังงานของอัลกอริทึม ID3 แสดงดัง 2 รูปถัดไปแบบต่อเนื่องกัน



รูปที่ 2.2 ผังงานของอัลกอริทึม ID3 (1/2)



รูปที่ 2.3 ผังงานของอัลกอริทึม ID3 (2/2)

บทที่ 3
งานวิจัยที่เกี่ยวข้อง

3.1 ข้อเสียของอัลกอริทึม ID3 ที่งานวิจัยที่เกี่ยวข้องนิยมให้ความสนใจ

3.1.1 ปัญหาการลำเอียง (bias) ของการเลือกแอตทริบิวต์ในอัลกอริทึม ID3 โดยใช้เกนความรู้

ปัญหาการลำเอียง (bias) ของการเลือกแอตทริบิวต์ในอัลกอริทึม ID3 โดยใช้เกนความรู้ เป็นปัญหาที่เกิดขึ้นเนื่องจากอัลกอริทึม ID3 มีแนวโน้มที่จะทำการเลือกแอตทริบิวต์ที่มีค่าที่เป็นไปได้หลายค่าไปเป็นแอตทริบิวต์บนโหนดตัดสินใจ ปัญหานี้อาจทำให้ต้นไม้ตัดสินใจที่ได้นั้นมีประสิทธิภาพที่ไม่ดี โดยอาจส่งผลกระทบต่อการสร้างโหนดลูกหลานที่มากขึ้น เนื่องจากมีการแตกกิ่งของโหนดตัดสินใจออกมามากขึ้นตามจำนวนของค่าที่เป็นไปได้ของแอตทริบิวต์ที่ถูกเลือกจากปัญหาการลำเอียงนี้ และอาจส่งผลต่อความแม่นยำในการจำแนกที่น้อยลงอีกด้วย เนื่องจากแอตทริบิวต์ที่มีค่าที่เป็นไปได้หลายค่า บางครั้ง อาจไม่ให้ความหมายของการตัดสินใจในเชิงโลกแห่งความเป็นจริงที่ดีนัก ตัวอย่างเช่น ถ้าจะสร้างต้นไม้ตัดสินใจที่ใช้ตัดสินใจว่า “จะเล่นเทนนิสหรือไม่” สมมติว่าถ้าแอตทริบิวต์สีเสื้อที่ใส่ ซึ่งมีค่าที่เป็นไปได้ถึง 10 ค่า ถูกเลือกเป็นแอตทริบิวต์บนโหนดตัดสินใจที่เป็นโหนดราก ทั้งที่ความเป็นจริงแล้ว แอตทริบิวต์ ชนิดสภาพอากาศ สมมติว่ามีค่าที่เป็นไปได้ 3 ค่า (Sunny, Overcast และ Rainy) เป็นแอตทริบิวต์ที่ดีกว่า สมเหตุสมผลในการตีความหมายตัดสินใจ และสมควรที่จะถูกเลือกแทนแอตทริบิวต์สีเสื้อ เพราะว่าชนิดสภาพอากาศ น่าจะมีผลต่อการตัดสินใจว่าจะเล่นเทนนิสหรือไม่ มากกว่าการดูที่สีเสื้อจริงๆ ดังนั้น หากนำแอตทริบิวต์สีเสื้อไปเป็นแอตทริบิวต์บนโหนดตัดสินใจ อาจจะส่งผลต่อความแม่นยำของต้นไม้ตัดสินใจที่น้อยลง ในขณะที่แอตทริบิวต์ชนิดสภาพอากาศมีแนวโน้มที่ทำให้ความแม่นยำของต้นไม้ตัดสินใจที่ได้ดีกว่าความแม่นยำที่ได้จากการเลือกแอตทริบิวต์สีเสื้อ ดังนั้นโดยสรุปแล้วการเลือกแอตทริบิวต์ที่มีค่าที่เป็นไปได้หลายค่าอาจไม่ทำให้ได้ผลที่ดีเสมอไป การเลือกแอตทริบิวต์ที่มีค่าที่เป็นไปได้น้อยค่า อาจให้ผลที่ดีกว่าก็เป็นได้

3.1.2 ปัญหาความซับซ้อนของการคำนวณค่าเกนความรู้ในอัลกอริทึม ID3

ปัญหาความซับซ้อนของการคำนวณค่าเกนความรู้ในอัลกอริทึม ID3 เป็นปัญหาที่เกิดขึ้นเนื่องจากสมการเกนความรู้ของอัลกอริทึม ID3 (สมการ 2.5) มีการคำนวณที่ใช้ลอการิทึม ซึ่งการคำนวณที่ใช้ลอการิทึมของสมการเกนความรู้ของอัลกอริทึม ID3 นั้น พบทั้งในพจน์ Entropy(T) (สมการ 2.6) และพจน์ Remainder(T | a) (สมการ 2.7) ในแง่ความซับซ้อนในการคำนวณของลอการิทึม ลอการิทึมมีความซับซ้อนในการคำนวณที่มากกว่าการบวก ลบ คูณ และหาร เป็นอย่างมาก เมื่อทำการพิจารณาพจน์ทั้ง 2 ภายในสมการเกนความรู้ จะพบว่าพจน์ Remainder(T | a) มีการคำนวณที่ใช้ลอการิทึมเยอะกว่าพจน์ Entropy(T) เพราะว่าพจน์ Remainder(T | a) จะทำการคำนวณ

ค่าเอนโทรปีของเซตย่อยที่ถูกแบ่งออกมาจากเซตของตัวอย่างฝึกฝนหลัก ซึ่งเซตของตัวอย่างฝึกฝนหลักนี้ถูกแบ่งออกมาเป็นเซตย่อยตามค่าที่เป็นไปได้ของแอตทริบิวต์ที่กำลังพิจารณา ดังนั้น ยิ่งค่าที่เป็นไปได้ของแอตทริบิวต์ที่กำลังพิจารณามีหลายค่ามากเท่าไร จำนวนครั้งที่ใช้ลอการิทึมในการคำนวณของพจน์ Remainder(T | a) ยิ่งมากขึ้นเช่นกัน ปัญหานี้ส่งผลกระทบต่อประสิทธิภาพในการสร้างต้นไม้ตัดสินใจของอัลกอริทึม ID3 ที่ช้าลง เพราะการทำงานของอัลกอริทึม ID3 ส่วนใหญ่เทหนักไปที่การคำนวณลอการิทึมเพื่อหาค่าเกนความรู้ ปัญหานี้จะหนักมากขึ้น ถ้าชุดข้อมูลที่ใช้ในการสร้างต้นไม้ตัดสินใจด้วยอัลกอริทึม ID3 มีขนาดใหญ่และมีจำนวนแอตทริบิวต์ที่มาก รวมทั้งแอตทริบิวต์ส่วนใหญ่ภายในชุดข้อมูลมีค่าที่เป็นไปได้หลายค่า

3.2 อัลกอริทึม ID3 ที่ปรับปรุงโดยการประยุกต์ใช้ทฤษฎีเทย์เลอร์และน้ำหนักความสำคัญของแอตทริบิวต์

อัลกอริทึม ID3 ที่ปรับปรุงโดยการประยุกต์ใช้ทฤษฎีเทย์เลอร์และน้ำหนักความสำคัญของแอตทริบิวต์ [3] เป็นงานวิจัยที่ทำการปรับปรุงอัลกอริทึม ID3 เพื่อจัดการกับปัญหาการลำเอียงของการเลือกแอตทริบิวต์ในอัลกอริทึม ID3 โดยใช้เกนความรู้ และจัดการกับปัญหาความซับซ้อนของการคำนวณค่าเกนความรู้ในอัลกอริทึม ID3

แนวคิดหลักของงานวิจัยนี้คือ ทำการปรับปรุงอัลกอริทึม ID3 โดยมุ่งไปที่การปรับปรุงสมการเกนความรู้ของอัลกอริทึม ID3 ซึ่งมี 2 ขั้นตอนหลักใหญ่คือ ใช้ทฤษฎีเทย์เลอร์ (Taylor's theorem) ลดความซับซ้อนของการคำนวณค่าเกนความรู้ในอัลกอริทึม ID3 โดยการเปลี่ยนส่วนที่มีการคำนวณด้วยลอการิทึมภายในสมการเกนความรู้ ให้อยู่ในรูปของการคำนวณด้วยการบวก ลบ คูณ และหาร และขั้นตอนถัดไปของการปรับปรุงสมการเกนความรู้ของอัลกอริทึม ID3 คือ ใช้ทฤษฎีความคล้ายของแอตทริบิวต์ (Attribute similarity theorem) ในการที่จะกำหนดน้ำหนักความสำคัญของแต่ละแอตทริบิวต์ โดยน้ำหนักดังกล่าวจะเป็นตัวปรับค่าในสมการเกนความรู้ที่ปรับปรุงจากขั้นตอนก่อนหน้า เพื่อจัดการและป้องกันแอตทริบิวต์ที่มีค่าที่เป็นไปได้หลายค่าถูกเลือกโดยพฤติกรรมการเลือกที่ลำเอียงของอัลกอริทึม ID3 สมการเกนความรู้ที่ปรับปรุงเสร็จแล้วนี้จะถูกนำไปใช้เป็นเกณฑ์ในการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 แทนการใช้สมการเกนความรู้แบบเดิม ซึ่งการเลือกแอตทริบิวต์ไปเป็นแอตทริบิวต์บนโหนดตัดสินใจนั้น ก็จะทำการเลือกจากแอตทริบิวต์ที่มีค่าเกนความรู้ที่ปรับปรุงมากที่สุด โดยรวมแล้ว กระบวนการทำงานของอัลกอริทึม ID3 ที่ปรับปรุงในงานวิจัยนี้ยังคงเดิมทุกอย่าง เปลี่ยนแค่สมการที่ใช้เป็นเกณฑ์ในการเลือกแอตทริบิวต์

3.2.1 การลดความซับซ้อนของการคำนวณค่าเกนความรู้ในอัลกอริทึม ID3 ด้วยทฤษฎีเทย์เลอร์

ขั้นตอนแรกของการปรับปรุงอัลกอริทึม ID3 ในงานวิจัยนี้ คือ การลดความซับซ้อนของการคำนวณค่าเกนความรู้ในอัลกอริทึม ID3 ด้วยการใช้ทฤษฎีเทย์เลอร์ เป็นการนำทฤษฎีดังกล่าวมาทำการปรับเปลี่ยนและจัดรูปสมการเกนความรู้แบบเดิม ซึ่งมีการคำนวณที่เกี่ยวกับลอการิทึมอยู่ ให้อยู่ในรูปของการคำนวณที่มีแต่การบวก ลบ คูณและหาร โดยจุดที่มีการปรับเปลี่ยนและจัดรูปจริงๆ ก็คือในส่วนของสมการ Entropy(T) (สมการ 2.6) ซึ่งสมการนี้ ถูกใช้เป็นตัวตั้งของสมการเกนความรู้แบบเดิมและใช้ในสมการ Remainder (สมการ 2.7) ภายใต้ผลรวมตามค่าที่เป็นไปได้ของแอตทริบิวต์ที่กำลังพิจารณา ดังนั้น จากการใช้ทฤษฎีเทย์เลอร์กับสมการ Entropy(T) จะได้สมการ Entropy(T) ที่ปรับปรุง $Entropy^{R1}(T)$ ดังสมการถัดไปที่แสดง

$$Entropy^{R1}(T) = \frac{1}{2|T|^3 \ln 2} \sum_{c \in C} |T_c|(|T| - |T_c|)(3|T| - |T_c|) \tag{3.1}$$

โดยที่ T คือเซตของตัวอย่างฝึกฝน,

C คือเซตของคลาสของตัวอย่างฝึกฝน,

c คือคลาสที่อยู่ในเซต C,

T_c คือเซตของตัวอย่างฝึกฝนที่มีคลาสเป็น c

จากสมการ $Entropy^{R1}(T)$ ที่ได้ ทำการแก้ไขสมการเกนความรู้แบบเดิม จะได้

$$IG^{R1}(T, a) = IG^{R1}(a) = Entropy^{R1}(T) - Remainder^{R1}(T|a) \tag{3.2}$$

โดยที่ T คือเซตของตัวอย่างฝึกฝน, a คือแอตทริบิวต์ และ $Remainder^{R1}(T \mid a)$ มีรูปสมการดังสมการถัดไป

$$Remainder^{R1}(T|a) = \sum_{v \in V} \frac{|T_v|}{|T|} \cdot Entropy^{R1}(T_v) \tag{3.3}$$

โดยที่ V คือเซตของค่าที่เป็นไปได้ของแอตทริบิวต์ a,

v คือ ค่าที่อยู่ในเซต V,

T_v คือเซตของตัวอย่างฝึกฝนที่มีค่าของแอตทริบิวต์ a เป็น v

สมการเกนความรู้ที่ปรับปรุงตามขั้นตอนแรกของการปรับปรุงอัลกอริทึม ID3 ในงานวิจัยนี้ $IG^{R1}(T, a)$ ยังไม่ใช่สมการที่จะนำไปใช้เป็นเกณฑ์ในการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 แทนการใช้สมการเกนความรู้แบบเดิมจริงๆ เนื่องจากสมการเกนความรู้ที่ปรับปรุงได้จากขั้นตอนนี้ยังคงมีปัญหาความลำเอียงของการเลือกแอตทริบิวต์ในอัลกอริทึม ID3 อยู่ เพราะขั้นตอนนี้ได้ปรับปรุง

สมการเกนความรู้แบบเดิมโดยทำการปรับเปลี่ยนและจัดรูปสมการเกนความรู้แบบเดิมผ่านการใช้ทฤษฎีเทย์เลอร์ ซึ่งการปรับเปลี่ยนและจัดรูปสมการดังกล่าวตามวิธีที่ใช้ในขั้นตอนนี้ ทำให้ความหมายและค่าของสมการเกนความรู้ที่ปรับปรุงได้จากขั้นตอนนี้ ต่างจากของสมการเกนความรู้แบบเดิมน้อยมาก ดังนั้น ปัญหาความลำเอียงของการเลือกแอตทริบิวต์ในอัลกอริทึม ID3 ยังคงอยู่ ส่งผลให้สมการเกนความรู้ที่ปรับปรุงได้จากขั้นตอนนี้ ต้องการการปรับปรุงเพิ่ม ดังที่จะอธิบายในขั้นตอนถัดไปของการปรับปรุงอัลกอริทึม ID3 ในงานวิจัยนี้ ซึ่งจะใช้ทฤษฎีความคล้ายของแอตทริบิวต์ มาช่วยจัดการกับปัญหาที่ยังคงเหลืออยู่

3.2.2 การจัดการปัญหาการลำเอียงของการเลือกแอตทริบิวต์ในอัลกอริทึม ID3 ด้วยการใช้ทฤษฎีความคล้ายของแอตทริบิวต์

จากขั้นตอนที่แล้ว สมการเกนความรู้ที่ปรับปรุงได้นั้น ยังคงมีปัญหาการลำเอียงของการเลือกแอตทริบิวต์ในอัลกอริทึม ID3 อยู่ งานวิจัยนี้จึงมีการใช้ทฤษฎีความคล้ายของแอตทริบิวต์เข้ามาช่วยจัดการกับปัญหาดังกล่าว โดยจะนำสมการเกนความรู้ที่ปรับปรุงได้จากขั้นตอนที่แล้ว มาทำการปรับปรุงเพิ่มเติมโดยใช้ทฤษฎีดังกล่าว เพื่อที่จะเพิ่มตัวปรับค่าซึ่งเป็นน้ำหนักความสำคัญของแต่ละแอตทริบิวต์ เข้าไปในตัวสมการ ตัวปรับค่านี้จะทำหน้าที่ในการเพิ่มค่าความสำคัญให้กับแอตทริบิวต์ที่มีความสำคัญจริงๆ เช่น แอตทริบิวต์ที่มีค่าที่เป็นไปได้น้อยค่าและมีแนวโน้มที่มีความสำคัญมากกว่าแอตทริบิวต์ที่อัลกอริทึม ID3 เลือกจากพฤติกรรมการลำเอียงซึ่งเป็นแอตทริบิวต์ที่มีค่าที่เป็นไปได้หลายค่า และตัวปรับค่านี้จะทำการลดค่าความสำคัญกับแอตทริบิวต์ที่มีความสำคัญสูงแต่กลับไม่ได้มีความสำคัญสูงอย่างแท้จริง เช่น แอตทริบิวต์ที่ถูกเลือกโดยพฤติกรรมการลำเอียงของอัลกอริทึม ID3 แอตทริบิวต์ในลักษณะนี้มีโอกาสที่จะถูกปรับค่าความสำคัญลง ถ้าหากว่าแอตทริบิวต์ในลักษณะนี้ไม่ได้มีความสำคัญสูงอย่างแท้จริง

ทฤษฎีความคล้ายของแอตทริบิวต์ที่งานวิจัยนี้ใช้ มีแนวคิดคือ กำหนดให้มีเซตของตัวอย่างฝึกฝน T, แอตทริบิวต์ a และเซตของคลาสของตัวอย่างฝึกฝน C ความคล้ายของแอตทริบิวต์ a ที่มีต่อเซตของคลาสของตัวอย่างฝึกฝน C เป็นการบอกว่า หากเอาแอตทริบิวต์ a มาทดสอบเพื่อแบ่งตัวอย่างภายในเซต T ออกไปตามค่าที่เป็นไปได้ของ a เซตย่อยแต่ละเซตจะมีคลาสของตัวอย่างภายในเซตสอดคล้องไปในทางเดียวกันมากน้อยแค่ไหน ถ้าโดยภาพรวมแล้ว เซตย่อยทุกเซตมีคลาสของตัวอย่างภายในเซตอยู่ในคลาสเดียวกันทั้งหมดทุกเซต นั่นก็หมายความว่า แอตทริบิวต์ a มีความคล้ายต่อ C มากที่สุด แต่ถ้าเกิดว่าไม่เป็นไปตามเงื่อนไขที่อธิบายก่อนหน้า นั่นก็หมายความว่า แอตทริบิวต์ a มีความคล้ายต่อ C อยู่ระหว่างระดับมีความคล้ายบางส่วน ไปจนถึงไม่มีความคล้ายต่อกันเลย สมการความคล้ายของแอตทริบิวต์ a ที่มีต่อเซตของคลาสของตัวอย่างฝึกฝน C แสดงดังสมการถัดไป

$$sim(a, C) = \frac{\langle a \cup C \rangle}{\sqrt{\langle a \rangle} \cdot \sqrt{\langle C \rangle}} \quad (3.4)$$

โดยที่ตัวดำเนินการ < > คือผลรวมของกำลังสองของจำนวนตัวอย่างจากเซต T ที่จำแนกไปอยู่ในแต่ละเซตย่อยของแต่ละค่าที่เป็นไปได้ของตัวแปรที่อยู่ภายใน < > ตัวอย่างเช่น <a> ตัวแปร a คือแอตทริบิวต์ a สมมติว่า a มีค่าที่เป็นไปได้ n ค่า ทำการจำแนกตัวอย่างจากเซต T ตามค่าที่เป็นไปได้ของ a จะได้เซตย่อย n เซต ให้เซตย่อยเหล่านี้มีชื่อเซตเป็น a_1, a_2, ..., a_n ค่าของ <a> สามารถหาได้โดยคำนวณผลรวมของกำลังสองของจำนวนตัวอย่างในเซต a_1, a_2, ..., a_n ดังสมการถัดไปที่แสดง

$$\langle a \rangle = |a_1|^2 + |a_2|^2 + \cdots + |a_n|^2 \tag{3.5}$$

จากสมการ sim(a, C) ทำการสร้างสมการตัวปรับค่าซึ่งเป็นน้ำหนักความสำคัญของแต่ละแอตทริบิวต์ จะได้

$$\omega_a = \frac{sim(a,C)}{\sum_{eachAttr \in A} sim(eachAttr,C)} \tag{3.6}$$

โดยที่ a คือแอตทริบิวต์ที่ต้องการหาน้ำหนักความสำคัญของแอตทริบิวต์,

A คือเซตของแอตทริบิวต์ของตัวอย่างฝึกฝน,

eachAttr คือแอตทริบิวต์ที่อยู่ในเซต A,

C คือ เซตของคลาสของตัวอย่างฝึกฝน

นำสมการตัวปรับค่าที่ได้ เพิ่มเข้าไปในสมการเกนความรู้ที่ปรับปรุงจากขั้นตอนที่แล้ว $IG^{R1}(T, a)$ จะได้สมการเกนความรู้ในรูปแบบที่ปรับปรุงเสร็จสิ้นของงานวิจัยนี้ ดังที่แสดงในสมการถัดไป

$$IG^{R1'}(T,a) = IG^{R1'}(a) = Entropy^{R1}(T) - Remainder^{R1}(T|a) \cdot \omega_a \tag{3.7}$$

ตัวแปร T ซึ่งเป็นเซตของตัวอย่างฝึกฝนที่อยู่ในสมการ $IG^{R1'}(T, a)$ และ A ซึ่งเป็นเซตของแอตทริบิวต์ที่ใช้ในการทำงานภายในบริบทปัจจุบันของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยนี้จะถูกนำไปใช้ในการคำนวณเพื่อหาค่าของน้ำหนักความสำคัญของแอตทริบิวต์ที่อยู่ตรงส่วนท้ายของสมการ $IG^{R1'}(T, a)$

สมการเกนความรู้ในรูปแบบที่ปรับปรุงเสร็จสิ้นของงานวิจัยนี้ จะถูกนำไปใช้เป็นเกณฑ์ในการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 แทนการใช้สมการเกนความรู้แบบเดิม และได้เป็นอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยนี้ ซึ่งกระบวนการทำงานโดยรวมของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยนี้ยังคงเหมือนอัลกอริทึม ID3 แบบดั้งเดิม เปลี่ยนแค่สมการที่ใช้เป็นเกณฑ์ในการเลือกแอตทริบิวต์

3.2.3 การทดลองประสิทธิภาพของอัลกอริทึม ID3 ที่ปรับปรุงในงานวิจัยนี้

งานวิจัยนี้ได้ทำการทดลองอัลกอริทึม ID3 ที่ปรับปรุง เพื่อทำการเปรียบเทียบประสิทธิภาพกับอัลกอริทึม ID3 แบบดั้งเดิม โดยใช้ชุดข้อมูล (Dataset) 2 ชุด ซึ่งนำมาจากตารางบันทึกการทำงานของอุปกรณ์ในฐานข้อมูลระบบการจัดการอุปกรณ์บ่อน้ำมัน Shengli แล้วทำการสร้าง-ทดสอบต้นไม้ตัดสินใจจากอัลกอริทึม ID3 ทั้ง 2 แบบ โดยใช้ชุดข้อมูลทั้ง 2 ชุดที่เตรียมไว้ ซึ่งชุดข้อมูลแต่ละชุด จะถูกแบ่งออกเป็น 2 ส่วน คือ ส่วนของชุดข้อมูลฝึกฝน (Training set) 70% และส่วนของชุดข้อมูลทดสอบ (Test set) 30% ผลการทดลองของอัลกอริทึม ID3 ทั้ง 2 แบบ แสดงดังตารางถัดไป

ตารางที่ 3.1 ผลการทดลองเปรียบเทียบระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัย [3]

	อัลกอริทึม ID3 แบบดั้งเดิม		**อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัย [3]**	
	เวลาในการรัน (ns)	**ความแม่นยำ**	**เวลาในการรัน (ns)**	**ความแม่นยำ**
ชุดข้อมูล 1	2235	0.78	1936	0.81
ชุดข้อมูล 2	2224	0.90	1879	0.95

ผลการทดลองในตารางการเปรียบเทียบข้างบนนี้ แสดงให้เห็นถึงประสิทธิภาพของเวลาในการรันของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยนี้ ซึ่งใช้เวลาน้อยลง เนื่องจากมีการลดความซับซ้อนของการคำนวณค่าเกนความรู้ในอัลกอริทึม ID3 แบบดั้งเดิมด้วยทฤษฎีเทย์เลอร์ และยังแสดงให้เห็นถึงประสิทธิภาพของความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยนี้ ซึ่งมีค่าเพิ่มขึ้น เนื่องจากมีการใช้ตัวปรับค่า ที่เป็นน้ำหนักความสำคัญของแอตทริบิวต์ อยู่ภายในสมการเกนความรู้ที่ปรับปรุงเสร็จสิ้นของงานวิจัยนี้

โดยสรุปแล้ว งานวิจัยนี้ สามารถลดความซับซ้อนของการคำนวณค่าเกนความรู้ในอัลกอริทึม ID3 แบบดั้งเดิมได้ รวมทั้งสามารถจัดการกับปัญหาการลำเอียงของการเลือกแอตทริบิวต์ในอัลกอริทึม ID3 แบบดั้งเดิมได้ด้วยเช่นกัน แต่งานวิจัยนี้มีข้อด้อยคือ งานวิจัยนี้ไม่ได้ระบุขนาดของชุดข้อมูลที่ใช้ในการทดลองว่าใหญ่แค่ไหน มีแอตทริบิวต์ภายในแต่ละชุดข้อมูลมากขนาดไหน รวมไปถึงจำนวนคลาสของแต่ละชุดข้อมูลด้วย

3.3 อัลกอริทึม ID3 ที่ปรับปรุงโดยใช้ระยะทางแบบยุคลิดโดยเฉลี่ย

อัลกอริทึม ID3 ที่ปรับปรุงโดยใช้ระยะทางแบบยุคลิดโดยเฉลี่ย [4] เป็นงานวิจัยที่ทำการปรับปรุงอัลกอริทึม ID3 เพื่อจัดการกับปัญหาการลำเอียงของการเลือกแอตทริบิวต์ในอัลกอริทึม ID3 โดยใช้เกนความรู้ และจัดการกับปัญหาความซับซ้อนของการคำนวณค่าเกนความรู้ในอัลกอริทึม ID3

แนวคิดหลักของงานวิจัยนี้คือ ทำการปรับปรุงอัลกอริทึม ID3 โดยมุ่งไปที่การเปลี่ยนเกณฑ์ในการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 จากเดิมที่ใช้เกนความรู้ งานวิจัยนี้จะใช้สมการที่ขึ้นกับระยะทางแบบยุคลิดโดยเฉลี่ย (Average Euclidean distance) แทน ซึ่งจะช่วยจัดการกับปัญหาทั้ง 2 ได้ โดยปัญหาการลำเอียงนั้น ตัวสมการที่ขึ้นกับหลักการดังกล่าวจะทำการตรวจสอบว่าแอตทริบิวต์ใดที่สามารถจำแนกคลาสของตัวอย่างภายในเซตของตัวอย่างฝึกฝนได้สอดคล้องมากที่สุด ในขณะที่ปัญหาความซับซ้อนของการคำนวณค่าเกนความรู้ในอัลกอริทึม ID3 นั้น ตัวสมการที่ขึ้นกับหลักการดังกล่าวจะไม่มีการคำนวณที่เกี่ยวข้องกับลอการิทึมเลย สมการที่งานวิจัยนี้เสนอเป็นสมการใหม่ ไม่ได้มีพื้นฐานหรือแก้ไขมาจากสมการเกนความรู้แต่อย่างใด ทำให้สามารถจัดการกับปัญหาความซับซ้อนของการคำนวณค่าเกนความรู้ในอัลกอริทึม ID3 ได้ สมการใหม่ที่งานวิจัยนี้เสนอ จะถูกนำไปใช้เป็นเกณฑ์ในการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 แทนการใช้สมการเกนความรู้แบบเดิม ซึ่งการเลือกแอตทริบิวต์ไปเป็นแอตทริบิวต์บนโหนดตัดสินใจนั้น ก็จะทำการเลือกจากแอตทริบิวต์ที่มีค่าของสมการใหม่ของงานวิจัยนี้มากที่สุด โดยรวมแล้ว กระบวนการทำงานของอัลกอริทึม ID3 ที่ปรับปรุงในงานวิจัยนี้ยังคงเดิมทุกอย่าง เปลี่ยนแค่สมการที่ใช้เป็นเกณฑ์ในการเลือกแอตทริบิวต์

3.3.1 เกณฑ์แบบใหม่ในการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 ที่ขึ้นกับระยะทางแบบยุคลิดโดยเฉลี่ย

เกณฑ์แบบใหม่ในการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 ที่ขึ้นกับระยะทางแบบยุคลิดโดยเฉลี่ยนี้ มีที่มาจากระยะทางแบบยุคลิด ซึ่งเป็นระยะทางระหว่างจุด 2 จุดบนปริภูมิ n มิติ (n-dimensional space) สมการระยะทางแบบยุคลิดนั้นมีรูปแบบการคำนวณที่มาจากการคำนวณตามทฤษฎีบทพีทาโกรัส คือ หาระยะทางที่สั้นที่สุด ที่สามารถตรงดิ่งจากจุดเริ่มต้นไปถึงที่หมายได้เลยโดยไม่มีการอ้อมหรือเลี้ยวโค้ง สมการระยะทางแบบยุคลิดแสดงดังสมการถัดไป กำหนดให้ X เป็นปริภูมิ n มิติ x และ y เป็นเวกเตอร์ n มิติที่อยู่ใน X

$$d(x, y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \tag{3.8}$$

โดยที่ d(x, y) คือระยะทางแบบยุคลิดระหว่างเวกเตอร์ x และ y บนปริภูมิ X ที่มี n มิติ,

x_k และ y_k คือค่าของเวกเตอร์ x และ y ในมิติที่ k

งานวิจัยนี้ได้ระบุว่าสมการที่ขึ้นกับระยะทางแบบยุคลิด ยังไม่สามารถนำไปใช้เป็นเกณฑ์ในการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 แทนการใช้สมการเกนความรู้แบบเดิมได้ เนื่องจากสมการที่ขึ้นกับระยะทางแบบยุคลิดไม่สามารถจัดการกับปัญหาการลำเอียงของการเลือกแอตทริบิวต์ในอัลกอริทึม ID3 โดยใช้เกนความรู้ ดังนั้นงานวิจัยนี้จึงเลือกใช้ระยะทางแบบยุคลิดโดยเฉลี่ยเป็น

พื้นฐานในการสร้างสมการใหม่เพื่อใช้เป็นเกณฑ์ในการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 ที่ปรับปรุงในงานวิจัยนี้จริงๆ

กำหนดให้ T เป็นเซตของตัวอย่างฝึกฝน ที่มีคลาสที่เป็นไปได้ 2 คลาส, a เป็นแอตทริบิวต์ ที่มีค่าที่เป็นไปได้ n ค่า สมการระยะทางแบบยุคลิดโดยเฉลี่ยของแอตทริบิวต์ a หรือ AED(a) แสดงดังสมการถัดไป

$$AED(a) = \frac{\sqrt{\sum_{i=1}^{n}(x_{i1}-x_{i2})^2}}{n} \quad (3.9)$$

โดยที่ x_{i1} กับ x_{i2} คือจำนวนของตัวอย่างภายในเซต T ที่มีค่าของแอตทริบิวต์ a เป็นค่าของแอตทริบิวต์ a ในลำดับที่ i และมีคลาสเป็นคลาสในลำดับที่ 1 กับ 2 ตามลำดับ

สมการระยะทางแบบยุคลิดโดยเฉลี่ยของแอตทริบิวต์ a จะทำการตรวจสอบแอตทริบิวต์ a โดยนำแอตทริบิวต์ a ไปทดสอบแบ่งตัวอย่างภายในเซต T ออกไปตามค่าที่เป็นไปได้ของ a แล้วดูว่าเซตย่อยแต่ละเซตจะมีคลาสของตัวอย่างภายในเซตสอดคล้องไปในทางเดียวกันมากน้อยแค่ไหน ถ้าโดยภาพรวมแล้ว เซตย่อยทุกเซตมีคลาสของตัวอย่างภายในเซตอยู่ในคลาสเดียวกันทั้งหมดทุกเซต นั่นก็หมายความว่า ระยะทางแบบยุคลิดโดยเฉลี่ยของแอตทริบิวต์ a มีค่ามากที่สุด แต่ถ้าเกิดว่าไม่เป็นไปตามเงื่อนไขที่อธิบายก่อนหน้า นั่นก็หมายความว่า ระยะทางแบบยุคลิดโดยเฉลี่ยของแอตทริบิวต์ a มีค่าอยู่ระหว่างระดับปานกลาง ไปจนถึงมีค่าเป็น 0

สมการระยะทางแบบยุคลิดโดยเฉลี่ยของแอตทริบิวต์ a จะถูกนำไปใช้เป็นเกณฑ์ในการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 แทนการใช้สมการเกนความรู้แบบเดิม และได้เป็นอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยนี้ ซึ่งกระบวนการทำงานโดยรวมของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยนี้ยังคงเหมือนอัลกอริทึม ID3 แบบดั้งเดิม เปลี่ยนแค่สมการที่ใช้เป็นเกณฑ์ในการเลือกแอตทริบิวต์

3.3.2 การทดลองประสิทธิภาพของอัลกอริทึม ID3 ที่ปรับปรุงในงานวิจัยนี้

งานวิจัยนี้ได้ทำการทดลองอัลกอริทึม ID3 ที่ปรับปรุง เพื่อทำการเปรียบเทียบประสิทธิภาพกับอัลกอริทึม ID3 แบบดั้งเดิม โดยใช้ชุดข้อมูล 5 ชุด ซึ่งนำมาจาก UCI Machine Learning Repository [6] แล้วทำการสร้าง-ทดสอบต้นไม้ตัดสินใจจากอัลกอริทึม ID3 ทั้ง 2 แบบ โดยใช้ชุดข้อมูลทั้ง 5 ชุดที่เตรียมไว้ ซึ่งชุดข้อมูลแต่ละชุด จะมีการจัดการกับปัญหาของตัวอย่างที่ค่าในบางแอตทริบิวต์ไม่ทราบค่า และสำหรับอัลกอริทึม ID3 ที่ปรับปรุงในงานวิจัยนี้ ต้องการชุดข้อมูลที่มีคลาสที่เป็นไปได้ 2 คลาสเท่านั้น จึงจำเป็นที่จะต้องปรับชุดข้อมูลที่เตรียมมา จากมีหลายคลาสที่เป็นไปได้ให้เหลือแค่ 2 คลาสเท่านั้น ในด้านของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนั้น ใช้จากชุดข้อมูลที่มีการแบ่งไว้ให้ก่อนแล้วภายในชุดข้อมูลที่งานวิจัยนี้เตรียมมา ถ้าชุดข้อมูลที่งานวิจัยนี้เตรียมมาชุดใดไม่มีการแบ่งชุดข้อมูลไว้ให้ก่อน ชุดข้อมูลเหล่านั้นก็จะถูกดำเนินการแบ่งออกเป็น 2 ส่วน คือ ส่วนของชุด

ข้อมูลฝึกฝน 70% และส่วนของชุดข้อมูลทดสอบ 30% ผลการทดลองของอัลกอริทึม ID3 ทั้ง 2 แบบ แสดงดังตาราง 2 ตารางข้างล่างนี้

ตารางที่ 3.2 ผลการทดลองเปรียบเทียบระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัย [4] ด้านความเที่ยงตรง (Precision)

ชุดข้อมูล	อัลกอริทึม ID3 แบบดั้งเดิม		อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัย [4]	
	ความเที่ยงตรงของการฝึกฝน	ความเที่ยงตรงของการทดสอบ	ความเที่ยงตรงของการฝึกฝน	ความเที่ยงตรงของการทดสอบ
SPECT Heart	0.937	0.711	0.787	0.791
Monks-3	1.0	0.944	1.0	0.975
Monks-1	0.983	0.866	1.0	0.884
Monks-2	0.988	0.646	1.0	0.812
Car	1.0	0.888	0.984	0.982

ตารางที่ 3.3 ผลการทดลองเปรียบเทียบระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัย [4] ด้านเวลา

ชุดข้อมูล	อัลกอริทึม ID3 แบบดั้งเดิม		อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัย [4]	
	เวลาในการฝึกฝน	เวลาในการทดสอบ	เวลาในการฝึกฝน	เวลาในการทดสอบ
SPECT Heart	0.697	0.597	0.663	0.567
Monks-3	0.669	0.545	0.594	0.539
Monks-1	0.670	0.546	0.652	0.540
Monks-2	0.703	0.647	0.668	0.597
Car	1.058	0.766	0.860	0.657

ผลการทดลองในตารางการเปรียบเทียบข้างบนทั้ง 2 ตารางนี้ แสดงให้เห็นถึงประสิทธิภาพของเวลาในการฝึกฝนและเวลาในการทดสอบของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยนี้ ซึ่งใช้เวลาน้อยลง เวลาในการฝึกฝนนั้นมีผลโดยตรงมาจากตัวสมการระยะทางแบบยุคลิดโดยเฉลี่ยของแอตทริบิวต์ ซึ่งไม่มีการคำนวณที่เกี่ยวข้องกับลอการิทึมเลย ทำให้ความซับซ้อนในการคำนวณของกระบวนการเลือกแอตทริบิวต์ในอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยนี้ลดลง ส่งผลให้ความเร็วในการสร้างต้นไม้ตัดสินใจของอัลกอริทึม ID3 ที่ปรับปรุงดังกล่าวเพิ่มขึ้น สำหรับเวลาในการทดสอบนั้นมีผลมาจากสมการระยะทางแบบยุคลิดโดยเฉลี่ยของแอตทริบิวต์เช่นกัน ซึ่งทำหน้าที่ในการตรวจสอบว่าแอตทริบิวต์ใด ที่สามารถจำแนกคลาสของตัวอย่างภายในเซตของตัวอย่างฝึกฝนได้สอดคล้องมากที่สุด เมื่อพิจารณาจากการแบ่งตัวอย่างภายในเซตของตัวอย่างฝึกฝน ตามค่าที่เป็นไปได้

ของแอตทริบิวต์ที่ตรวจสอบ นอกเหนือจากประสิทธิภาพด้านเวลาทั้ง 2 ที่น้อยลงแล้ว ความเที่ยงตรงของการทดสอบของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยนี้ ก็มีค่าที่มากกว่าของอัลกอริทึม ID3 แบบดั้งเดิมเช่นกัน

โดยสรุปแล้ว งานวิจัยนี้ สามารถลดความซับซ้อนของการคำนวณค่าเกนความรู้ในอัลกอริทึม ID3 แบบดั้งเดิมได้ รวมทั้งสามารถจัดการกับปัญหาการลำเอียงของการเลือกแอตทริบิวต์ในอัลกอริทึม ID3 แบบดั้งเดิมได้ด้วยเช่นกัน แต่งานวิจัยนี้มีข้อด้อยคือ สมการระยะทางแบบยุคลิดโดยเฉลี่ยของแอตทริบิวต์ที่งานวิจัยนี้เสนอ รองรับการคำนวณกับชุดข้อมูลที่มีคลาสที่เป็นไปได้ไม่เกิน 2 คลาสเท่านั้น หากเกินกว่านี้ต้องใช้วิธีการปรับคลาสภายในชุดข้อมูลให้เหลือเพียง 2 คลาส ซึ่งอาจส่งผลต่อความเที่ยงตรง / ความแม่นยำได้ อีกข้อด้อยหนึ่งก็คือ งานวิจัยนี้ไม่ได้ทำการทดลองกับชุดข้อมูลที่มีขนาดใหญ่มาก จึงไม่สามารถทราบได้ว่า ประสิทธิภาพโดยรวมของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยนี้ ยังสามารถทำงานได้ดีเหมือนตอนที่ทดลองกับชุดข้อมูลขนาดเล็กหรือไม่

3.4 อัลกอริทึม ID3 ที่ปรับปรุงโดยใช้ระดับนัยสำคัญของแอตทริบิวต์และฟังก์ชันนูน

อัลกอริทึม ID3 ที่ปรับปรุงโดยใช้ระดับนัยสำคัญของแอตทริบิวต์และฟังก์ชันนูน [2] เป็นงานวิจัยที่ทำการปรับปรุงอัลกอริทึม ID3 เพื่อจัดการกับปัญหาการลำเอียงของการเลือกแอตทริบิวต์ในอัลกอริทึม ID3 โดยใช้เกนความรู้ และจัดการกับปัญหาความซับซ้อนของการคำนวณค่าเกนความรู้ในอัลกอริทึม ID3

แนวคิดหลักของงานวิจัยนี้คือ ทำการปรับปรุงอัลกอริทึม ID3 โดยมุ่งไปที่การปรับปรุงสมการเกนความรู้ของอัลกอริทึม ID3 ซึ่งมี 2 ขั้นตอนหลักใหญ่คือ ใช้ระดับนัยสำคัญของแอตทริบิวต์ (Attribute significance) เป็นตัวปรับค่าในสมการเกนความรู้แบบเดิมของอัลกอริทึม ID3 เพื่อจัดการกับปัญหาการลำเอียงของการเลือกแอตทริบิวต์ในอัลกอริทึม ID3 และขั้นตอนถัดไปของการปรับปรุงสมการเกนความรู้ของอัลกอริทึม ID3 คือ ใช้ฟังก์ชันนูน (Convex function) ในการลดความซับซ้อนของการคำนวณค่าเกนความรู้ในอัลกอริทึม ID3 โดยการเปลี่ยนส่วนที่มีการคำนวณด้วยลอการิทึมภายในสมการเกนความรู้ที่ปรับปรุงจากขั้นตอนก่อนหน้า ให้อยู่ในรูปของการคำนวณด้วยการบวก ลบ คูณ และหาร สมการเกนความรู้ที่ปรับปรุงเสร็จแล้วนี้จะถูกนำไปใช้เป็นเกณฑ์ในการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 แทนการใช้สมการเกนความรู้แบบเดิม ซึ่งการเลือกแอตทริบิวต์ไปเป็นแอตทริบิวต์บนโหนดตัดสินใจนั้น ก็จะทำการเลือกจากแอตทริบิวต์ที่มีค่าเกนความรู้ที่ปรับปรุงมากที่สุด โดยรวมแล้ว กระบวนการทำงานของอัลกอริทึม ID3 ที่ปรับปรุงในงานวิจัยนี้ยังคงเดิมทุกอย่าง เปลี่ยนแค่สมการที่ใช้เป็นเกณฑ์ในการเลือกแอตทริบิวต์

3.4.1 การจัดการปัญหาการลำเอียงของการเลือกแอตทริบิวต์ในอัลกอริทึม ID3 ด้วยการใช้ระดับนัยสำคัญของแอตทริบิวต์

ขั้นตอนแรกของการปรับปรุงอัลกอริทึม ID3 ในงานวิจัยนี้ คือ การจัดการปัญหาการลำเอียงของการเลือกแอตทริบิวต์ในอัลกอริทึม ID3 ด้วยการใช้ระดับนัยสำคัญของแอตทริบิวต์ เป็นตัวปรับค่าในสมการเกนความรู้ ตัวปรับค่านี้จะถูกใช้อยู่ในสมการ Remainder ซึ่งเป็นพจน์ตัวลบของสมการเกนความรู้ แนวคิดการทำงานของตัวปรับค่านี้คล้ายกับงานวิจัยที่ได้อธิบายไป [3] ซึ่งใช้น้ำหนักความสำคัญของแอตทริบิวต์ แต่ที่มาของตัวปรับค่านี้ต่างกัน ดังที่จะอธิบายในย่อหน้าถัดไป

ตัวปรับค่าที่เป็นระดับนัยสำคัญของแอตทริบิวต์นี้ มีพื้นฐานมาจากทฤษฎีเซตหยาบ (Rough set theory) กำหนดให้ T เป็นเซตของตัวอย่างฝึกฝน, C เป็นเซตของคลาสของตัวอย่างฝึกฝน , A เป็นเซตของแอตทริบิวต์ของตัวอย่างฝึกฝน และ a เป็นแอตทริบิวต์ที่ต้องการหาระดับนัยสำคัญของแอตทริบิวต์ สมการระดับนัยสำคัญของแอตทริบิวต์ a หรือ SGF(a) แสดงดังสมการถัดไป

$$\boldsymbol{SGF(a) = \gamma(T, A, C) - \gamma(T, A - \{a\}, C)} \tag{3.10}$$

โดยฟังก์ชันที่มีอักษรกรีกแกมมานำหน้า แสดงดังสมการถัดไป

$$\boldsymbol{\gamma(U, Conds, Dcns) = \frac{|POS_{Conds}(Dcns)|}{|U|}} \tag{3.11}$$

โดยที่ U เป็นเซตของตัวอย่างที่ใช้ในการพิจารณา,
Conds เป็นเซตของแอตทริบิวต์เงื่อนไข,
Dcns เป็นเซตของแอตทริบิวต์ตัดสินใจ (แอตทริบิวต์ที่ระบุคลาสของตัวอย่าง),
$POS_{Conds}(Dcns)$ เป็นฟังก์ชันในทฤษฎีเซตหยาบ
นำสมการ SGF(a) ไปเพิ่มในสมการ Remainder แบบเดิมจะได้

$$\boldsymbol{Remainder^{R2}(T|a) = \sum_{v \in V} \frac{|T_v|}{|T|} \cdot SGF(a) \cdot Entropy(T_v)} \tag{3.12}$$

ตัวแปร T กับ a ซึ่งเป็นเซตของตัวอย่างฝึกฝนกับแอตทริบิวต์ที่อยู่ในสมการ $Remainder^{R2}(T \mid a)$ พร้อมด้วย A กับ C ซึ่งเป็นเซตของแอตทริบิวต์ของตัวอย่างฝึกฝนกับเซตของคลาสของตัวอย่างฝึกฝน ที่ใช้ในการทำงานภายในบริบทปัจจุบันของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยนี้ จะถูกนำไปใช้ในการคำนวณเพื่อหาค่า SGF(a) ที่อยู่ภายในสมการ $Remainder^{R2}(T \mid a)$

จากสมการ $Remainder^{R2}(T \mid a)$ จะได้สมการเกนความรู้ที่ปรับปรุงในขั้นตอนแรกของการปรับปรุงอัลกอริทึม ID3 ในงานวิจัยนี้ แสดงดังสมการถัดไป

$$IG^{R2}(T,a) = IG^{R2}(a) = Entropy(T) - Remainder^{R2}(T|a) \quad (3.13)$$

สมการเกนความรู้ที่ปรับปรุงในขั้นตอนแรกของการปรับปรุงอัลกอริทึม ID3 ในงานวิจัยนี้ $IG^{R2}(T, a)$ ยังไม่ใช่สมการที่จะนำไปใช้เป็นเกณฑ์ในการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 แทนการใช้สมการเกนความรู้แบบเดิมจริงๆ เนื่องจากสมการเกนความรู้ที่ปรับปรุงได้จากขั้นตอนนี้ยังคงมีความซับซ้อนในการคำนวณที่สูงอยู่ เพราะตัวสมการเกนความรู้ที่ปรับปรุงได้จากขั้นตอนนี้ไม่ได้มีความแตกต่างจากสมการเกนความรู้แบบเดิมมากนัก ขั้นตอนแรกของงานวิจัยนี้ทำการเพิ่ม SGF(a) เข้าไปในพจน์ของสมการ Remainder แบบเดิมเท่านั้น โดยที่ส่วนอื่นของสมการเกนความรู้ที่ปรับปรุงได้จากขั้นตอนนี้ไม่มีการเปลี่ยนแปลง ด้วยเหตุผลนี้ จึงส่งผลให้สมการเกนความรู้ที่ปรับปรุงได้จากขั้นตอนนี้ ต้องการการปรับปรุงเพิ่ม ดังที่จะอธิบายในขั้นตอนถัดไปของการปรับปรุงอัลกอริทึม ID3 ในงานวิจัยนี้ ซึ่งจะใช้ฟังก์ชันนูน มาช่วยจัดการกับปัญหาที่ยังคงเหลืออยู่

3.4.2 การลดความซับซ้อนของการคำนวณค่าเกนความรู้ในอัลกอริทึม ID3 ด้วยฟังก์ชันนูน

จากขั้นตอนที่แล้ว ปัญหาความซับซ้อนของการคำนวณค่าเกนความรู้ในอัลกอริทึม ID3 ยังไม่ได้ถูกจัดการ สมการเกนความรู้ที่ปรับปรุงได้ในขั้นตอนที่แล้ว ยังคงมีความซับซ้อนในการคำนวณที่สูงอยู่ ดังนั้นงานวิจัยนี้จึงใช้ฟังก์ชันนูน เข้ามาช่วยจัดการกับปัญหาดังกล่าว โดยจะนำสมการเกนความรู้ที่ปรับปรุงได้จากขั้นตอนที่แล้ว มาทำการลดความซับซ้อนของการคำนวณโดยทำการปรับเปลี่ยนและจัดรูปสมการเกนความรู้ที่ปรับปรุงได้จากขั้นตอนที่แล้ว ซึ่งมีการคำนวณที่เกี่ยวกับลอการิทึมอยู่ ให้อยู่ในรูปของการคำนวณที่มีแต่การบวก ลบ คูณและหาร จากการใช้ฟังก์ชันนูนกับสมการดังกล่าว จะได้สมการเกนความรู้ในรูปแบบที่ปรับปรุงเสร็จสิ้นของงานวิจัยนี้ ดังที่แสดงในรูปถัดไป

$$IG^{R2'}(T,a) = IG^{R2'}(a) = \sum_{v \in V} \frac{|T_v|}{|T|} \cdot SGF(a) \cdot \left(\sum_{c1 \in C} \sum_{c2 \in C \wedge c2 \neq c1} p(T_{vc1}) \times p(T_{vc2}) \right) \quad (3.14)$$

โดยที่ T เป็นเซตของตัวอย่างฝึกฝน,

a เป็นแอตทริบิวต์,

A กับ C เป็นเซตของแอตทริบิวต์ของตัวอย่างฝึกฝนกับเซตของคลาสของตัวอย่างฝึกฝน ที่ใช้ในการทำงานภายในบริบทปัจจุบันของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยนี้,

V เป็นเซตของค่าที่เป็นไปได้ของแอตทริบิวต์ a,

v เป็นค่าที่อยู่ในเซต V,

T_v เป็นเซตของตัวอย่างฝึกฝนที่มีค่าของแอตทริบิวต์ a เป็น v,

c1 และ c2 เป็นคลาสที่อยู่ในเซต C,

$p(T_{vc1})$ และ $p(T_{vc2})$ คือความน่าจะเป็นที่ตัวอย่างจากคลาส c1 และ c2 ซึ่งอยู่ในเซต T_v จะปรากฏ

สมการเกนความรู้ในรูปแบบที่ปรับปรุงเสร็จสิ้นของงานวิจัยนี้ จะถูกนำไปใช้เป็นเกณฑ์ในการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 แทนการใช้สมการเกนความรู้แบบเดิม และได้เป็นอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยนี้ ซึ่งกระบวนการทำงานโดยรวมของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยนี้ยังคงเหมือนอัลกอริทึม ID3 แบบดั้งเดิม เปลี่ยนแค่สมการที่ใช้เป็นเกณฑ์ในการเลือกแอตทริบิวต์

3.4.3 การทดลองประสิทธิภาพของอัลกอริทึม ID3 ที่ปรับปรุงในงานวิจัยนี้

งานวิจัยนี้ได้ทำการทดลองอัลกอริทึม ID3 ที่ปรับปรุง เพื่อทำการเปรียบเทียบประสิทธิภาพกับอัลกอริทึม ID3 แบบดั้งเดิม โดยใช้ชุดข้อมูล 6 ชุด ซึ่งนำมาจาก UCI Machine Learning Repository ชุดข้อมูลทั้ง 6 ชุดนี้ผ่านการปรับทำให้เป็นชุดข้อมูลที่มีความไม่ต่อเนื่อง

งานวิจัยนี้มีการทดลอง 3 แบบที่จะทดลองประสิทธิภาพของอัลกอริทึม ID3 ทั้ง 2 ชนิด ทุกการทดลองใช้ชุดข้อมูลทั้ง 6 ชุดที่เตรียมไว้ โดยรายละเอียดของแต่ละการทดลองมีดังนี้

การทดลองที่ 1 หาความแม่นยำในการจำแนกโดยเฉลี่ยของต้นไม้ตัดสินใจที่สร้างได้จากอัลกอริทึม ID3 ทั้ง 2 ชนิด โดยใช้วิธีการแบ่งครึ่งช่วง 10 ชั้น (10 layer bracketing method)

การทดลองที่ 2 หาจำนวนโหนดคำตอบ (Leaf node) โดยเฉลี่ยของต้นไม้ตัดสินใจที่สร้างได้จากอัลกอริทึม ID3 ทั้ง 2 ชนิด โดยการทดลองนี้จะทำการแบ่งชุดข้อมูลที่เตรียมมาแต่ละชุดออกเป็นชุดข้อมูลย่อย 10 ส่วน แต่ละส่วนจะมีตัวอย่างที่มาจากการสุ่มเลือกทั้งหมด แล้วใช้ชุดข้อมูลย่อยทั้ง 10 ส่วนที่แบ่งได้ของแต่ละชุดข้อมูล ไปทำการสร้างต้นไม้ตัดสินใจโดยใช้อัลกอริทึม ID3 ทั้ง 2 ชนิด เป็นจำนวน 10 ครั้ง/อัลกอริทึม แต่ละครั้งใช้ชุดข้อมูลย่อยที่ไม่ซ้ำกับชุดข้อมูลย่อยที่เคยใช้แล้ว จากนั้นทำการหาค่าเฉลี่ยของจำนวนโหนดคำตอบของต้นไม้ตัดสินใจที่สร้างได้จากอัลกอริทึม ID3 ทั้ง 2 ชนิด

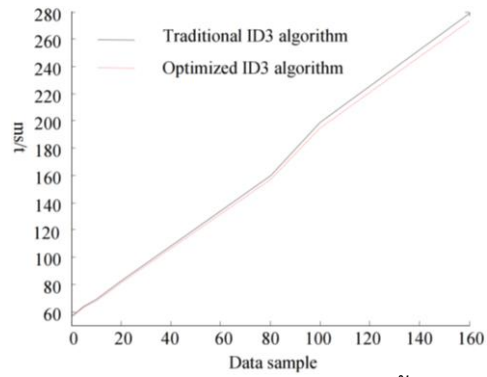
การทดลองที่ 3 หาเวลาในการสร้างต้นไม้ตัดสินใจโดยเฉลี่ย (เวลาในการฝึกฝนโดยเฉลี่ย) ของอัลกอริทึม ID3 ทั้ง 2 ชนิด โดยการทดลองนี้จะทำการแบ่งชุดข้อมูลที่เตรียมมาแต่ละชุดออกเป็นชุดข้อมูลย่อย 20 ส่วน แต่ละส่วนจะมีตัวอย่างที่มาจากการสุ่มเลือกทั้งหมด แล้วใช้ชุดข้อมูลย่อยทั้ง 20 ส่วนที่แบ่งได้ของแต่ละชุดข้อมูล ไปทำการสร้างต้นไม้ตัดสินใจโดยใช้อัลกอริทึม ID3 ทั้ง 2 ชนิด เป็นจำนวน 20 ครั้ง/อัลกอริทึม แต่ละครั้งใช้ชุดข้อมูลย่อยที่ไม่ซ้ำกับชุดข้อมูลย่อยที่เคยใช้แล้ว จากนั้นทำการหาค่าเฉลี่ยของเวลาในการสร้างต้นไม้ตัดสินใจของอัลกอริทึม ID3 ทั้ง 2 ชนิด

ผลการทดลองที่ 1-2 ของอัลกอริทึม ID3 ทั้ง 2 แบบ แสดงดังตารางถัดไป

ตารางที่ 3.4 ผลการทดลองที่ 1-2 ของอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัย [2]

ชุดข้อมูล	อัลกอริทึม ID3 แบบดั้งเดิม		อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัย [2]	
	ความแม่นยำโดยเฉลี่ย	จำนวนโหนดคำตอบโดยเฉลี่ย	ความแม่นยำโดยเฉลี่ย	จำนวนโหนดคำตอบโดยเฉลี่ย
Breast	85.8%	8.0	90.5%	6.2
Diabetes	71.5%	16.5	79.2%	10.8
Iris	73.1%	7.0	78.7%	5.0
Lymph	73.2%	8.5	77.8%	7.2
Bupa	82.8%	10.0	88.9%	7.8
Segmentation	73.2%	10.8	84.1%	8.1

ผลการทดลองในตารางข้างบนนี้ แสดงให้เห็นถึงประสิทธิภาพความแม่นยำโดยเฉลี่ยและจำนวนโหนดคำตอบโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยนี้ ซึ่งมีประสิทธิภาพที่ดีกว่าประสิทธิภาพทั้ง 2 ของอัลกอริทึม ID3 แบบดั้งเดิม เนื่องจากมีการใช้ตัวปรับค่า ที่เป็นระดับนัยสำคัญของแอตทริบิวต์ อยู่ภายในสมการเกนความรู้ที่ปรับปรุงเสร็จสิ้นของงานวิจัยนี้ กราฟผลการทดลองที่ 3 ของอัลกอริทึม ID3 ทั้ง 2 แบบ แสดงดังรูปถัดไป



รูปที่ 3.1 กราฟผลการทดลองที่ 3 ของอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัย [2]

กราฟข้างบนนี้ แสดงให้เห็นถึงประสิทธิภาพเวลาในการสร้างต้นไม้ตัดสินใจโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยนี้ ซึ่งใช้เวลาน้อยกว่าเวลาของอัลกอริทึม ID3 แบบดั้งเดิม เนื่องจากมีการลดความซับซ้อนของการคำนวณค่าเกนความรู้ในอัลกอริทึม ID3 แบบดั้งเดิมด้วยการใช้ฟังก์ชันนูน

โดยสรุปแล้ว งานวิจัยนี้ สามารถลดความซับซ้อนของการคำนวณค่าเกนความรู้ในอัลกอริทึม ID3 แบบดั้งเดิมได้ รวมทั้งสามารถจัดการกับปัญหาการลำเอียงของการเลือกแอตทริบิวต์

ในอัลกอริทึม ID3 แบบดั้งเดิมได้ด้วยเช่นกัน แต่งานวิจัยนี้มีข้อด้อยคือ งานวิจัยนี้ไม่ได้ทำการทดลองกับชุดข้อมูลที่มีขนาดใหญ่มาก จึงไม่สามารถทราบได้ว่า ประสิทธิภาพโดยรวมของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยนี้ ยังสามารถทำงานได้ดีเหมือนตอนที่ทดลองกับชุดข้อมูลขนาดเล็กหรือไม่ และอีกข้อด้อยก็คือ งานวิจัยนี้ไม่ได้แสดงผลการทดลองด้านเวลาในการทดสอบ ผลการทดลองในงานวิจัยนี้แสดงให้เห็นถึงเวลาในการสร้างต้นไม้ตัดสินใจโดยเฉลี่ยและจำนวนโหนดคำตอบที่ได้ แต่ไม่ได้สะท้อนและแสดงให้เห็นถึงเวลาในการทดสอบหรือความเร็วในการจำแนกตัวอย่างที่ไม่เคยฝึกฝนมาก่อน ว่าเป็นอย่างไร

บทที่ 4
อัลกอริทึม ID3 กับแอตทริบิวต์ที่มีความสำคัญเท่ากัน

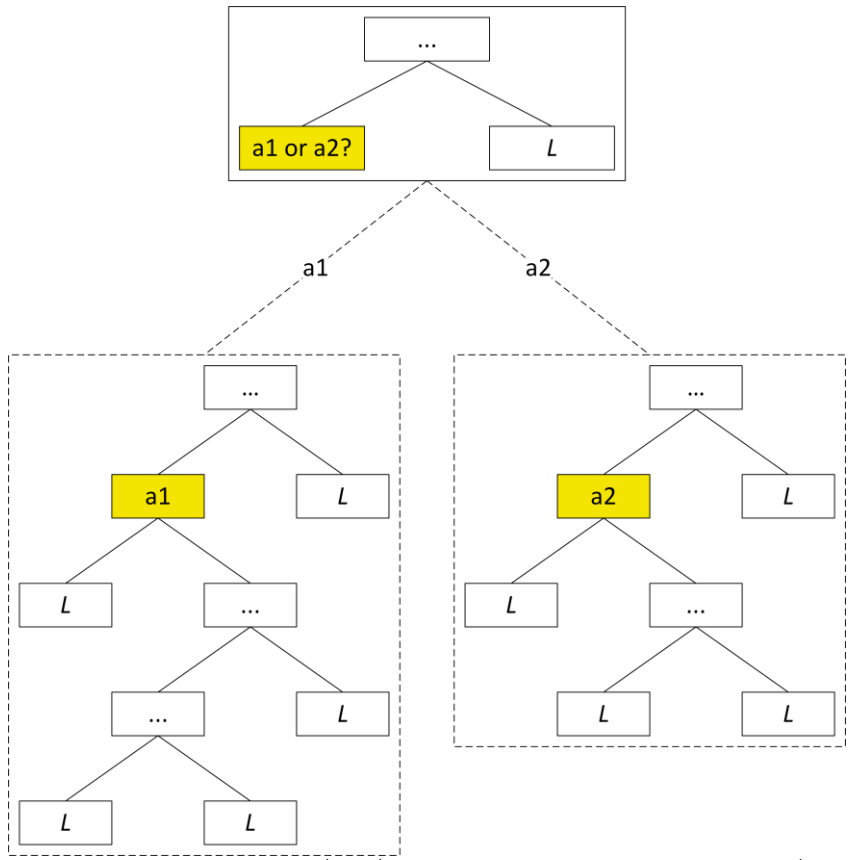
4.1 ปัญหาแอตทริบิวต์ที่มีความสำคัญเท่ากัน

ปัญหาแอตทริบิวต์ที่มีความสำคัญเท่ากัน (Equally important attributes problem) เป็นปัญหาที่เกิดขึ้นในขั้นตอนการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 เมื่อจะทำการเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดเพียงแอตทริบิวต์เดียว เพื่อนำมาวางบนโหนดตัดสินใจที่จะถูกสร้างภายในบริบทปัจจุบันของอัลกอริทึม ID3 แต่กลับพบว่ามีแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันอย่างน้อย 2 แอตทริบิวต์

ปกติแล้วในขั้นตอนการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 ตัวอัลกอริทึม ID3 จะทำการเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดเพียงแอตทริบิวต์เดียว ซึ่งในมุมมองอัลกอริทึม ID3 ถือว่าเป็นแอตทริบิวต์ที่มีความสำคัญที่สุด แอตทริบิวต์ที่ถูกเลือกตามขั้นตอนนี้ จะนำไปเป็นแอตทริบิวต์บนโหนดตัดสินใจที่ถูกสร้างภายในบริบทปัจจุบันของอัลกอริทึม ID3 ถ้าแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุด มีเพียงแอตทริบิวต์เดียว อัลกอริทึม ID3 สามารถเลือกแอตทริบิวต์ดังกล่าวได้ทันที แต่ในกรณีที่มีแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันอย่างน้อย 2 แอตทริบิวต์ จะเกิดปัญหาแอตทริบิวต์ที่มีความสำคัญเท่ากันขึ้น

การเลือกแอตทริบิวต์ของอัลกอริทึม ID3 ในกรณีที่มีแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันอย่างน้อย 2 แอตทริบิวต์นั้น ไม่ว่าจะเลือกแอตทริบิวต์ใดก็ตามจากแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันทั้งหมด ต้นไม้ตัดสินใจที่ได้ย่อมมีความแม่นยำที่ใกล้เคียงกันเสมอ เมื่อทำการทดสอบด้วยตัวอย่างทดสอบในสัดส่วนที่เพียงพอ ซึ่งมีค่าของแต่ละแอตทริบิวต์และคลาสภายในตัวอย่างทดสอบทุกตัวอย่างคละอย่างเท่าเทียมกัน อย่างไรก็ตาม ต้นไม้ตัดสินใจที่ได้ อาจจะมีความลึกที่ไม่เท่ากัน กล่าวคือ แอตทริบิวต์ที่ถูกเลือกในกรณีนี้ ส่งผลต่อความลึกของต้นไม้ตัดสินใจที่ได้ ต้นไม้ตัดสินใจที่มีความลึกน้อย จะมีโอกาสช่วยให้การจำแนกตัวอย่างที่ไม่เคยฝึกฝนมาก่อนมีความรวดเร็วมากขึ้น แต่อัลกอริทึม ID3 ไม่มีการรู้ล่วงหน้าว่าแอตทริบิวต์ใดจากแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันทั้งหมด เมื่อเลือกแล้วจะทำให้ได้ต้นไม้ตัดสินใจที่มีความลึกที่น้อยกว่า ดังนั้นอัลกอริทึม ID3 จึงจัดการกับกรณีที่มีแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันอย่างน้อย 2 แอตทริบิวต์ โดยการสุ่มเลือก 1 แอตทริบิวต์จากแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันทั้งหมด ซึ่งการสุ่มเลือก 1 แอตทริบิวต์ของอัลกอริทึม ID3 เพื่อจัดการกับกรณีดังกล่าว อาจทำให้ต้นไม้ตัดสินใจที่ได้มีความลึกที่มากเกินความจำเป็น และมีโอกาสทำให้การจำแนกตัวอย่างที่ไม่เคยฝึกฝนมาก่อนมีความช้าลง อันเนื่องมาจากต้นไม้ตัดสินใจที่มีความลึกมาก

สิ่งที่เกิดขึ้นทั้งหมดจากการที่อัลกอริทึม ID3 จัดการกับกรณีที่มีแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันอย่างน้อย 2 แอตทริบิวต์ ด้วยวิธีการสุ่มเลือก 1 แอตทริบิวต์จากแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันทั้งหมดนั้น กลายเป็นผลกระทบของปัญหาแอตทริบิวต์ที่มีความสำคัญเท่ากัน รูปถัดไปแสดงทางเลือกของต้นไม้ตัดสินใจที่ได้เมื่ออัลกอริทึม ID3 เลือกแอตทริบิวต์ที่ต่างกันจากแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันทั้งหมด กำหนดให้แอตทริบิวต์ a1 และ a2 เป็นแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากัน โหนดที่มี ... คือ โหนดตัดสินใจอื่นๆ ที่อยู่ภายในต้นไม้ตัดสินใจที่สร้างได้ โหนดที่มีตัว L คือ โหนดคำตอบ และโหนดที่ระบายสีเหลือง คือโหนดตัดสินใจที่แอตทริบิวต์ a1 และ a2 มีค่าเกนความรู้สูงที่สุดและเท่ากัน



รูปที่ 4.1 ทางเลือกของต้นไม้ตัดสินใจที่ได้เมื่ออัลกอริทึม ID3 เลือกแอตทริบิวต์ที่ต่างกัน จากแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันทั้งหมด

จากรูปข้างบนนี้ ต้นไม้ตัดสินใจที่ได้จากการเลือกแอตทริบิวต์ a2 ในโหนดสีเหลือง มีความลึกน้อยกว่าความลึกของต้นไม้ตัดสินใจที่ได้จากการเลือกแอตทริบิวต์ a1 ในโหนดสีเหลือง ดังนั้นแอตทริบิวต์ a2 จึงเป็นแอตทริบิวต์ที่ดีกว่า อัลกอริทึม ID3 ก็ควรที่จะเลือกแอตทริบิวต์ a2 เป็นแอตทริบิวต์บนโหนดสีเหลือง แต่เนื่องจากอัลกอริทึม ID3 ทำการเลือก 1 แอตทริบิวต์แบบสุ่มจากแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันทั้งหมด เพื่อจัดการกับกรณีที่มีแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุด

และเท่ากัน เพราะฉะนั้นเราจึงไม่สามารถบังคับให้อัลกอริทึม ID3 เลือกแอตทริบิวต์ a2 ตามที่ต้องการได้โดยตรง อัลกอริทึม ID3 จะเลือกแอตทริบิวต์ใดจากกรณีดังกล่าว ขึ้นอยู่กับการสุ่มเพียงอย่างเดียว ทำให้ต้นไม้ตัดสินใจที่ได้มีความลึกที่ขึ้นกับแอตทริบิวต์ที่ถูกเลือก หากโชคไม่ดีก็จะได้ต้นไม้ตัดสินใจที่มีความลึกมาก ยิ่งถ้าอัลกอริทึม ID3 พบแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันในระดับความลึกต้นๆ ของกระบวนการสร้างต้นไม้ตัดสินใจ ก็จะยิ่งมีโอกาสส่งผลต่อความลึกของต้นไม้ตัดสินใจที่สร้างเสร็จได้ โอกาสส่งผลดังกล่าว คือ ส่งผลทำให้ได้ต้นไม้ตัดสินใจที่มีความลึกมาก หรือ ส่งผลทำให้ได้ต้นไม้ตัดสินใจที่มีความลึกน้อย

4.2 อัลกอริทึม ID3 ที่ปรับปรุงโดยใช้วิธีการรวมแอตทริบิวต์ที่มีความสำคัญเท่ากัน

วิทยานิพนธ์นี้ได้ศึกษาและคิดค้นงานวิจัยแรก [5] ซึ่งเป็นงานวิจัยที่ทำการปรับปรุงอัลกอริทึม ID3 โดยใช้วิธีการรวมแอตทริบิวต์ที่มีความสำคัญเท่ากัน เพื่อจัดการกับปัญหาแอตทริบิวต์ที่มีความสำคัญเท่ากัน แนวคิดหลักของงานวิจัยแรกคือ ทำการปรับปรุงอัลกอริทึม ID3 โดยเปลี่ยนวิธีการเลือกแอตทริบิวต์ จากเดิมเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดเพียงแอตทริบิวต์เดียว โดยไม่สนว่าจะมีแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันกี่แอตทริบิวต์ก็ตาม เปลี่ยนเป็นเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันมาพร้อมกัน แอตทริบิวต์ที่ถูกเลือกพร้อมกันจะนำไปเป็นแอตทริบิวต์บนโหนดตัดสินใจเดียวกัน และทำการพิจารณาแตกกิ่ง (Branch) จากโหนดตัดสินใจปัจจุบันที่ถูกสร้าง ตามค่าที่เป็นไปได้ของแอตทริบิวต์ที่ถูกเลือกพร้อมกันแบบร่วมกัน ซึ่งจะอธิบายรายละเอียดเชิงลึกในหัวข้อย่อยต่อไป

4.2.1 การปรับปรุงอัลกอริทึม ID3 โดยใช้วิธีการรวมแอตทริบิวต์ที่มีความสำคัญเท่ากัน

วิธีการที่งานวิจัยแรก [5] ใช้ในการปรับปรุงอัลกอริทึม ID3 คือ วิธีการรวมแอตทริบิวต์ที่มีความสำคัญเท่ากัน เป็นวิธีการที่คิดค้นขึ้นเอง แนวคิดของวิธีการนี้อยู่บนพื้นฐานของปัญหาแอตทริบิวต์ที่มีความสำคัญเท่ากัน เมื่ออัลกอริทึม ID3 จะทำการเลือกแอตทริบิวต์เพื่อนำมาวางในโหนดตัดสินใจ และพบกรณีที่มีแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันอย่างน้อย 2 แอตทริบิวต์ จะเกิดปัญหาแอตทริบิวต์ที่มีความสำคัญเท่ากันขึ้น ปัญหานี้อาจทำให้ต้นไม้ตัดสินใจที่ได้มีความลึกที่มากเกินความจำเป็น ในขณะที่ความแม่นยำของต้นไม้ตัดสินใจที่ได้ ยังคงอยู่ในระดับที่ใกล้เคียงกันเหมือนเดิม ไม่ว่าแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดที่ถูกเลือกจะเป็นแอตทริบิวต์ใดก็ตาม

จากปัญหาแอตทริบิวต์ที่มีความสำคัญเท่ากัน งานวิจัยแรกจึงได้ทำการปรับปรุงอัลกอริทึม ID3 เพื่อจัดการกับปัญหาดังกล่าว โดยใช้วิธีการรวมแอตทริบิวต์ที่มีความสำคัญเท่ากัน วิธีการนี้จะทำการเปลี่ยนวิธีการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 เพื่อรองรับการเลือกแอตทริบิวต์ในกรณีที่มีแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากัน ซึ่งปกติแล้วเมื่ออัลกอริทึม ID3 พบกรณีดังกล่าวในระหว่างการเลือกแอตทริบิวต์ อัลกอริทึม ID3 จะทำการสุ่มเลือก 1 แอตทริบิวต์จากแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันทั้งหมด วิธีการรวมแอตทริบิวต์ที่มีความสำคัญเท่ากัน ได้ทำ

การเปลี่ยนวิธีการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 ในกรณีดังกล่าวเป็น เลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันมาพร้อมกัน แอตทริบิวต์ที่ถูกเลือกพร้อมกันจะนำไปอยู่บนโหนดตัดสินใจเดียวกัน ทำให้โหนดตัดสินใจของอัลกอริทึม ID3 ที่ปรับปรุงในงานวิจัยแรกนี้ สามารถมีแอตทริบิวต์ที่เกี่ยวข้องกำกับอยู่บนโหนดมากกว่า 1 แอตทริบิวต์ได้ วิธีการรวมแอตทริบิวต์ที่มีความสำคัญเท่ากัน จะไม่ไปจัดการหรือยุ่งกับกรณีที่อัลกอริทึม ID3 จะทำการเลือกแอตทริบิวต์แล้วพบแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดเพียงแอตทริบิวต์เดียว เนื่องจากกรณีนี้ อัลกอริทึม ID3 สามารถเลือกแอตทริบิวต์ดังกล่าวมาเป็นแอตทริบิวต์บนโหนดตัดสินใจได้ทันที

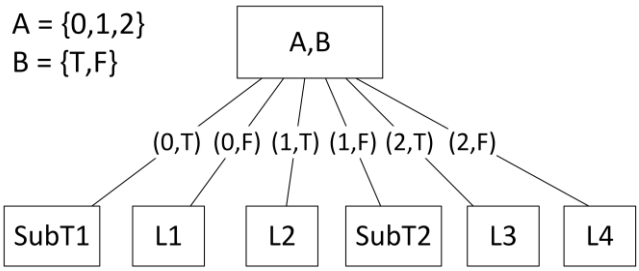
การแตกกิ่งออกมาจากโหนดตัดสินใจของอัลกอริทึม ID3 ที่ปรับปรุงโดยใช้วิธีการรวมแอตทริบิวต์ที่มีความสำคัญเท่ากัน จะมีลักษณะการแตกกิ่งออกมาจากโหนดตัดสินใจที่ไม่เหมือนกัน ขึ้นอยู่กับจำนวนแอตทริบิวต์ที่อยู่ในโหนดตัดสินใจ การแตกกิ่งออกมาจากโหนดตัดสินใจที่มีแอตทริบิวต์อยู่ในโหนดเพียงแอตทริบิวต์เดียว สามารถที่จะแตกตามค่าที่เป็นไปได้ของแอตทริบิวต์ที่อยู่ในโหนดนั้นได้ทันที แต่การแตกกิ่งออกมาจากโหนดตัดสินใจที่มีแอตทริบิวต์อยู่ในโหนดมากกว่า 1 แอตทริบิวต์ ไม่สามารถที่จะเลือกแตกตามค่าที่เป็นไปได้ของแอตทริบิวต์ใดแอตทริบิวต์หนึ่งจากแอตทริบิวต์ที่อยู่ในโหนดตัดสินใจเดียวกันทั้งหมดได้ เนื่องจากแอตทริบิวต์ทุกแอตทริบิวต์ที่อยู่ในโหนดตัดสินใจเดียวกันนั้น จะต้องถูกพิจารณาร่วมกันทั้งในการแตกกิ่งออกมาจากโหนดตัดสินใจและการทดสอบต้นไม้ตัดสินใจที่ได้ ดังนั้นการแตกกิ่งออกมาจากโหนดตัดสินใจที่มีแอตทริบิวต์อยู่ในโหนดมากกว่า 1 แอตทริบิวต์ จึงจำเป็นที่จะต้องนำค่าที่เป็นไปได้ของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในโหนด มาพิจารณาร่วมกัน โดยทำการนำเซตของค่าที่เป็นไปได้ของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในโหนดตัดสินใจ มาหาหลายสิ่งอันดับ (Tuple) ด้วยการใช้ผลคูณคาร์ทีเซียน (Cartesian product) ซึ่งจะได้เซตของหลายสิ่งอันดับ เซตของหลายสิ่งอันดับแสดงให้เห็นถึงวิธีการจัดหมู่ (Combination) ที่เป็นไปได้ทั้งหมดในการรวมค่าที่เป็นไปได้ของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในโหนดตัดสินใจ แต่ละหลายสิ่งอันดับที่อยู่ในเซตที่ได้ จะประกอบไปด้วยค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในโหนดตัดสินใจ ซึ่งถูกจัดหมู่ให้อยู่ด้วยกันเป็นกลุ่ม หลายสิ่งอันดับ 1 หลายสิ่งอันดับเป็น 1 กลุ่มของค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในโหนดตัดสินใจ จากนั้นจึงทำการแตกกิ่งออกมาจากโหนดตัดสินใจตามจำนวนของหลายสิ่งอันดับที่อยู่ในเซตที่ได้ ซึ่งแต่ละกิ่งที่แตกออกมาจากโหนดตัดสินใจจะมีการระบุกำกับด้วยกลุ่มของค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในโหนดตัดสินใจ (หลายสิ่งอันดับ) เพื่อเป็นการบอกว่ากิ่งต่างๆ ที่แตกออกมา เป็นกิ่งของกลุ่มของค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในโหนดตัดสินใจกลุ่มใด

สมมติว่ามีแอตทริบิวต์ A และ B เป็นแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดและเท่ากัน แอตทริบิวต์ทั้ง 2 ถูกเลือกไปเป็นแอตทริบิวต์บนโหนดตัดสินใจโหนดเดียวกัน แอตทริบิวต์ A มีค่าที่เป็นไปได้ 3 ค่า เซตของค่าที่เป็นไปได้ของแอตทริบิวต์ A คือ $V_A = \{0,1,2\}$ แอตทริบิวต์ B มีค่าที่เป็นไป

ได้ 2 ค่า เซตของค่าที่เป็นไปได้ของแอตทริบิวต์ B คือ V_B = {T,F} โดยการใช้ผลคูณคาร์ทีเซียนระหว่างเซต V_A และ V_B จะได้เซตของหลายสิ่งอันดับดังสมการถัดไปที่แสดง

$$V_A \times V_B = \{(0,\mathrm{T}),(0,\mathrm{F}),(1,\mathrm{T}),(1,\mathrm{F}),(2,\mathrm{T}),(2,\mathrm{F})\} \quad (4.1)$$

จากสมการผลคูณคาร์ทีเซียนระหว่างเซต V_A และ V_B ที่แสดงข้างบนนี้ จะเห็นว่าผลลัพธ์ที่ได้เป็นเซตของหลายสิ่งอันดับซึ่งมีสมาชิก 6 ตัว สมาชิกแต่ละตัวเป็นหลายสิ่งอันดับ ซึ่งภายในหลายสิ่งอันดับแต่ละหลายสิ่งอันดับ จะประกอบไปด้วย 2 ค่าคือ ค่าของแอตทริบิวต์ A และค่าของแอตทริบิวต์ B ตัวอย่างเช่น (0,T) เป็นหลายสิ่งอันดับซึ่งประกอบไปด้วยค่าของแอตทริบิวต์ A เป็น 0 และค่าของแอตทริบิวต์ B เป็น T ส่วน (1,F) เป็นหลายสิ่งอันดับซึ่งประกอบไปด้วยค่าของแอตทริบิวต์ A เป็น 1 และค่าของแอตทริบิวต์ B เป็น F อย่างนี้เป็นต้น เซตของหลายสิ่งอันดับที่ได้ในสมการข้างบนนี้ จะแทนวิธีการจัดหมู่ (Combination) ที่เป็นไปได้ทั้งหมดในการรวมค่าที่เป็นไปได้ของแอตทริบิวต์ A และ B ไล่ตั้งแต่กลุ่มแรกคือ ค่าของแอตทริบิวต์ A เป็น 0 และค่าของแอตทริบิวต์ B เป็น T ไปจนถึงกลุ่มสุดท้ายคือ ค่าของแอตทริบิวต์ A เป็น 2 และค่าของแอตทริบิวต์ B เป็น F เมื่อได้เซตของหลายสิ่งอันดับซึ่งมีรายละเอียดวิธีการจัดหมู่ในการรวมค่าที่เป็นไปได้ของแอตทริบิวต์ A และ B แล้ว ขั้นตอนต่อไปก็จะทำการแตกกิ่งออกมาจากโหนดตัดสินใจ ตามจำนวนของสมาชิกในเซตของหลายสิ่งอันดับที่ได้จากการคำนวณในสมการข้างบน ซึ่งก็คือ 6 กิ่ง แต่ละกิ่งที่แตกออกมาจะมีการระบุกำกับด้วยกลุ่มของค่าของแอตทริบิวต์ A และ B ตามที่ปรากฏในเซตของหลายสิ่งอันดับที่ได้ดังกล่าว รูปถัดไปแสดงการแตกกิ่งของโหนดตัดสินใจที่มีแอตทริบิวต์ A และ B อยู่ด้วยกัน โดย SubT1 และ SubT2 เป็นต้นไม้ตัดสินใจย่อยที่อยู่ในกิ่งที่แตกออกมาจากโหนดตัดสินใจหลัก ส่วน L1 – L4 เป็นโหนดคำตอบที่อยู่ในกิ่งที่แตกออกมาจากโหนดตัดสินใจหลัก



รูปที่ 4.2 การแตกกิ่งของโหนดตัดสินใจที่มีแอตทริบิวต์ A และ B อยู่ด้วยกัน

จากรูป โหนดตัดสินใจที่มีแอตทริบิวต์ A และ B อยู่ด้วยกัน เมื่อทำการแตกกิ่งแล้วจะพบว่ามีกิ่งทั้งหมด 6 กิ่ง โดยแต่ละกิ่งจะมีกลุ่มของค่าของแอตทริบิวต์ A และ B กำกับ ไล่จากซ้ายสุดกิ่งแรก มี (0,T) กำกับ หมายถึงกิ่งนี้เป็นกิ่งที่เกี่ยวข้องกับค่าของแอตทริบิวต์ A เป็น 0 และค่าของแอตทริบิวต์ B เป็น T ไปจนถึงขวาสุดกิ่งสุดท้าย มี (2,F) กำกับ หมายถึงกิ่งนี้เป็นกิ่งที่เกี่ยวข้องกับค่า

ของแอตทริบิวต์ A เป็น 2 และค่าของแอตทริบิวต์ B เป็น F จะเห็นได้อย่างชัดเจนว่า โหนดตัดสินใจที่มีแอตทริบิวต์อยู่ในโหนดมากกว่า 1 แอตทริบิวต์ มีลักษณะของกิ่งซึ่งมาจากการพิจารณาค่าที่เป็นไปได้ของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในโหนดตัดสินใจแบบร่วมกันจริงๆ ลักษณะกิ่งที่ได้จากโหนดตัดสินใจที่มีแอตทริบิวต์อยู่ในโหนดมากกว่า 1 แอตทริบิวต์นั้น ทำให้การสร้างโหนดลูกตามกิ่งต่างๆ ที่แตกออกมาจากโหนดตัดสินใจที่มีแอตทริบิวต์อยู่ในโหนดมากกว่า 1 แอตทริบิวต์ จะต้องใช้เซตย่อยทั้งหมดของเซตของตัวอย่างฝึกฝนหลัก ที่มาจากการแบ่งเซตของตัวอย่างฝึกฝนหลักตามกลุ่มของค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในโหนดตัดสินใจ โดยเซตย่อยแต่ละเซตนี้จะนำไปใช้สร้างกับกิ่งของโหนดตัดสินใจซึ่งมีกลุ่มของค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในโหนดตัดสินใจของกิ่งสอดคล้องกับกลุ่มของค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในโหนดตัดสินใจของเซตย่อยที่ใช้สร้าง ในตอนทดสอบต้นไม้ตัดสินใจที่สร้างด้วยการใช้ตัวอย่างทดสอบ หากตัวอย่างทดสอบดำเนินการหาคำตอบจากต้นไม้ตัดสินใจที่สร้าง แล้วพบโหนดตัดสินใจที่มีแอตทริบิวต์อยู่ในโหนดมากกว่า 1 แอตทริบิวต์ ตัวอย่างทดสอบก็จะทำการเลือกกิ่งของโหนดตัดสินใจที่จะไปต่อ โดยเลือกกิ่งที่มีกลุ่มของค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในโหนดตัดสินใจของกิ่ง สอดคล้องกับกลุ่มของค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในโหนดตัดสินใจของตัวอย่างทดสอบ

การใช้ผลคูณคาร์ทีเซียน ในการช่วยหากลุ่มของค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในโหนดตัดสินใจทุกกลุ่มที่เป็นไปได้ทั้งหมด ผ่านผลลัพธ์ที่เป็นเซตของหลายสิ่งอันดับ ทำให้ทราบข้อเท็จจริงสองอย่างคือ หนึ่ง ยิ่งแอตทริบิวต์ที่ถูกเลือกพร้อมกันมีค่าที่เป็นไปได้หลายค่า จำนวนกิ่งที่แตกออกมาจากโหนดตัดสินใจก็จะมีมากขึ้น ทำให้เกิดการสร้างโหนดในระดับลูกหลานมากขึ้น และสอง ยิ่งเลือกแอตทริบิวต์มาพร้อมกันมากเท่าไร ก็จะเกิดผลเหมือนข้อแรกแต่มากขึ้นยิ่งกว่าเดิม ผลของการแตกกิ่งออกมาจากโหนดตัดสินใจที่มากขึ้นนี้ ส่งผลต่อกระบวนการสร้างต้นไม้ตัดสินใจของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกอย่างชัดเจน ทั้งเวลาที่ใช้ในการสร้างซึ่งใช้เวลามากขึ้น และใช้พื้นที่มากขึ้นในการเก็บรายละเอียดของต้นไม้ตัดสินใจที่สร้างได้ จากข้อเท็จจริงสองอย่างที่กล่าวมา อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกนี้ จึงกำหนดให้มีพารามิเตอร์ 1 ตัว คือ N ซึ่งเป็นพารามิเตอร์ที่กำหนดจำนวนแอตทริบิวต์สูงสุดที่สามารถเลือกมาไว้ในโหนดตัดสินใจเดียวกัน ทำให้สามารถควบคุมขนาดของต้นไม้ตัดสินใจที่สร้างผ่านพารามิเตอร์ N ได้

4.2.2 ขั้นตอนการทำงานของอัลกอริทึม ID3 ที่ปรับปรุงโดยใช้วิธีการรวมแอตทริบิวต์ที่มีความสำคัญเท่ากัน

อัลกอริทึม ID3 ที่ปรับปรุงโดยใช้วิธีการรวมแอตทริบิวต์ที่มีความสำคัญเท่ากัน (อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก [5]) ต้องการอินพุต 3 อย่างคือ เซตของตัวอย่างฝึกฝน เซตของแอตทริบิวต์ของตัวอย่างฝึกฝน และพารามิเตอร์ N (จำนวนแอตทริบิวต์สูงสุดที่สามารถเลือกมาไว้ในโหนดตัดสินใจเดียวกัน) ลักษณะการทำงานของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก คล้ายกับของอัลกอริทึม ID3 แบบดั้งเดิมทุกประการ ต่างกันตรงวิธีการเลือกแอตทริบิวต์มาไว้ในโหนด

ตัดสินใจซึ่งเพิ่มวิธีการเลือกแอตทริบิวต์สำหรับกรณีที่มีแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากัน วิธีการแตกกิ่งออกมาจากโหนดตัดสินใจ และวิธีการแบ่งตัวอย่างฝึกฝนที่อยู่ในเซตของตัวอย่างฝึกฝนเพื่อนำไปใช้สร้างโหนดลูกตามกิ่งต่างๆ ที่แตกออกมาจากโหนดตัดสินใจ ขั้นตอนการทำงานของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกมีดังนี้ กำหนดให้ T คือเซตของตัวอย่างฝึกฝนที่รับค่าเข้ามา , A คือเซตของแอตทริบิวต์ของตัวอย่างฝึกฝนที่รับค่าเข้ามา, Tc คือเซตที่เก็บสำเนาของตัวอย่างทั้งหมดของเซต T, selAttrs คือ เซตของแอตทริบิวต์ที่อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกเลือก และ combVals คือ เซตของกลุ่มของค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในเซต selAttrs

1) ทำการคำนวณค่าเกนความรู้ของทุกแอตทริบิวต์ที่อยู่ในเซต A โดยใช้สมการ (2.5) – (2.7) ร่วมกับเซต T
2) ทำการเลือกแอตทริบิวต์ในเซต A มาเก็บไว้ในเซต selAttrs ตามเงื่อนไขที่สอดคล้อง
    2.1) ถ้าแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดมีเพียงแอตทริบิวต์เดียว ให้ทำการเลือกแอตทริบิวต์ดังกล่าวได้ทันที
    2.2) ถ้ามีแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากัน
        2.2.1) ถ้าจำนวนของแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากัน มีน้อยกว่าหรือเท่ากับ N แอตทริบิวต์ ให้ทำการเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันทั้งหมดได้ทันที
        2.2.2) ถ้าจำนวนของแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากัน มีมากกว่า N แอตทริบิวต์ ให้ทำการสุ่มเลือก N แอตทริบิวต์จากแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันทั้งหมด
3) ทำการสร้างโหนดตัดสินใจ แล้วใช้แอตทริบิวต์ที่อยู่ในเซต selAttrs ทั้งหมดเป็นแอตทริบิวต์บนโหนดตัดสินใจที่สร้าง
4) นำแอตทริบิวต์ที่ถูกเลือก ซึ่งเป็นแอตทริบิวต์ที่อยู่ในเซต selAttrs ทั้งหมดออกจากเซต A ; A = A – selAttrs เนื่องจากจะไม่มีการนำแอตทริบิวต์ที่เคยเลือกไปแล้วมาพิจารณาซ้ำในการสร้างโหนดตัดสินใจอื่นๆ ที่อยู่ในระดับลูกหลานของโหนดตัดสินใจปัจจุบันที่เพิ่งสร้างไป
5) ทำการคัดลอกตัวอย่างภายในเซต T ทั้งหมดไปใส่ในเซต Tc
6) ทำการพิจารณาหากลุ่มของค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในเซต selAttrs ทุกกลุ่มที่เป็นไปได้ แล้วนำกลุ่มของค่าดังกล่าวที่หาได้ทั้งหมดไปเก็บ

ไว้ในเซต combVals โดยพิจารณาหากลุ่มของค่าดังกล่าวตามเงื่อนไขของจำนวนแอตทริบิวต์ที่อยู่ในเซต selAttrs

6.1) ถ้าแอตทริบิวต์ที่อยู่ในเซต selAttrs มีเพียงแอตทริบิวต์เดียว ให้ทำการหากลุ่มของค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในเซต selAttrs ผ่านการใช้ค่าที่เป็นไปได้ทุกค่าของแอตทริบิวต์ที่อยู่ในเซต selAttrs เป็นกลุ่มของค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในเซต selAttrs ได้โดยตรง

6.2) ถ้าแอตทริบิวต์ที่อยู่ในเซต selAttrs มีมากกว่า 1 แอตทริบิวต์ ให้ทำการหากลุ่มของค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในเซต selAttrs ผ่านการใช้ผลคูณคาร์ทีเซียนกับเซตของค่าที่เป็นไปได้ของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในเซต selAttrs

7) ใช้เซต combVals ในการแบ่งเซต T ตามกลุ่มของค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในเซต selAttrs ที่เป็นไปได้ จะได้เซตย่อย k เซต ; $t_1 – t_k$ โดยที่ k คือจำนวนกลุ่มของค่าทั้งหมดในเซต combVals, $cv_1 – cv_k$ คือกลุ่มของค่าที่อยู่ในเซต combVals ที่เป็นไปได้ทั้งหมด และ $t_1 – t_k$ คือเซตย่อยที่ตัวอย่างฝึกฝนทุกตัวอย่างภายในเซต มีค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในเซต selAttrs เป็น $cv_1 – cv_k$ ตามลำดับ

8) ทำการแตกกิ่ง (Branch) ออกมาจากโหนดตัดสินใจปัจจุบันที่สร้าง จำนวน k กิ่ง แต่ละกิ่งระบุว่าเป็นกิ่งของกลุ่มของค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในเซต selAttrs กลุ่มใด ไล่ตั้งแต่ กิ่งแรกระบุเป็นกิ่งของกลุ่มของค่า cv_1 ไปจนถึงกิ่งสุดท้ายระบุเป็นกิ่งของกลุ่มของค่า cv_k

9) ทำการตรวจสอบเซตย่อยแต่ละเซตที่แบ่งได้ กำหนดให้ t_i คือเซตย่อยที่กำลังพิจารณา และ i มีค่าไล่ตามลำดับตั้งแต่ 1 ไปถึง k

9.1) ถ้าเซตย่อย t_i ไม่มีตัวอย่างใดๆ อยู่ในเซตเลย ให้ทำการพิจารณาตัวอย่างภายในเซต Tc ดูว่าภายในเซต Tc ตัวอย่างส่วนใหญ่มีสัดส่วนอยู่ในคลาสใดมากที่สุด ให้ทำการสร้างโหนดคำตอบที่ตอบคลาสดังกล่าว แล้วนำโหนดคำตอบที่สร้างเชื่อมกับโหนดตัดสินใจปัจจุบันที่ได้สร้างไป ผ่านกิ่งของกลุ่มของค่า cv_i หากเงื่อนไขในขั้นตอนนี้ไม่เป็นจริง ให้ทำการตรวจสอบเซตย่อย t_i ตามเงื่อนไขในขั้นตอนถัดไป

9.2) ถ้าตัวอย่างภายในเซตย่อย t_i อยู่ในคลาสเดียวกันทั้งหมด ให้ทำการสร้างโหนดคำตอบที่ตอบคลาสดังกล่าว แล้วนำโหนดคำตอบที่สร้างเชื่อมกับโหนดตัดสินใจปัจจุบันที่ได้สร้างไป ผ่านกิ่งของกลุ่มของค่า cv_i

หากเงื่อนไขในขั้นตอนนี้ไม่เป็นจริง ให้ทำการตรวจสอบเซตย่อย t_i ตามเงื่อนไขในขั้นตอนถัดไป

9.3) ถ้าเซต A ไม่มีแอตทริบิวต์ใดๆ เหลืออยู่ในเซตเลย ส่งผลให้อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกไม่สามารถดำเนินการเลือกแอตทริบิวต์ ตามเงื่อนไขของตัวอัลกอริทึมที่กำหนดต่อไปได้อีก ทำให้ไม่สามารถสร้างโหนดตัดสินใจได้ และเมื่อสร้างโหนดตัดสินใจไม่ได้ ตัวอัลกอริทึมเองก็ย่อมไม่สามารถที่จะแบ่งเซตย่อย t_i ต่อไปได้อีกเช่นกัน หากเงื่อนไขในขั้นตอนนี้เป็นจริง ให้ทำการพิจารณาตัวอย่างภายในเซตย่อย t_i ดูว่าภายในเซตย่อย t_i ตัวอย่างส่วนใหญ่มีสัดส่วนอยู่ในคลาสใดมากที่สุด ให้ทำการสร้างโหนดคำตอบที่ตอบคลาสดังกล่าว แล้วนำโหนดคำตอบที่สร้างเชื่อมกับโหนดตัดสินใจปัจจุบันที่ได้สร้างไปผ่านกิ่งของกลุ่มของค่า cv_i หากเงื่อนไขในขั้นตอนนี้ไม่เป็นจริง ให้ทำตามขั้นตอนสุดท้ายของการตรวจสอบเซตย่อย t_i ได้เลย

9.4) เนื่องจากตัวอย่างภายในเซตย่อย t_i ไม่ได้อยู่ในคลาสเดียวกัน และยังคงมีแอตทริบิวต์ในเซต A เหลืออยู่

9.4.1) ทำการคัดลอกเซต A จะได้เซตสำเนาของเซต A

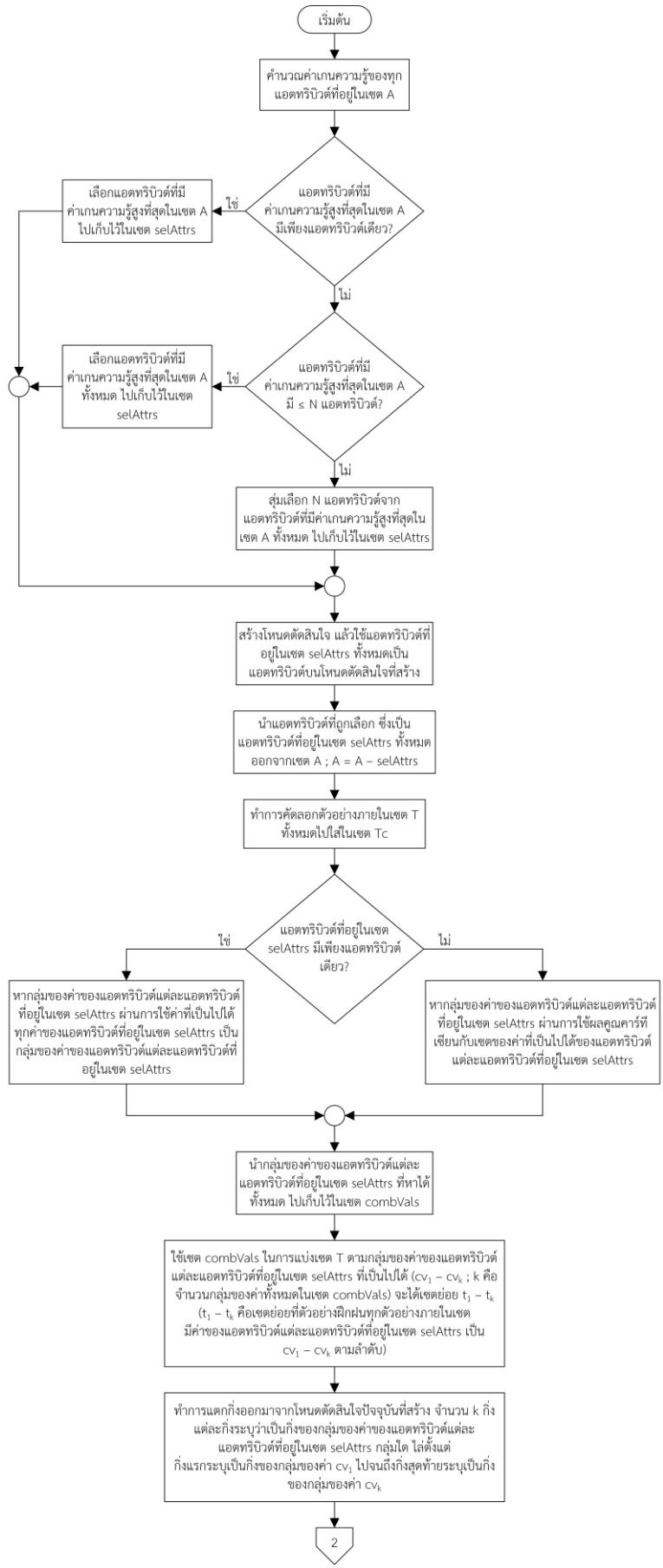
9.4.2) ทำการเรียกใช้งานอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในบริบทรองที่แยกต่างหากเลย (ไม่ใช่การวนกลับไปทำขั้นตอนที่ 1 ของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในบริบทปัจจุบัน แต่เป็นการเรียกใช้งานอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในบริบทรอง ซึ่งกระทำกับอินพุตที่ต่างกัน บริบทรองนี้จะไม่ทับกับบริบทปัจจุบัน ดังนั้นการทำงานของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในบริบทปัจจุบันยังคงสามารถกลับมาดำเนินการต่อได้ ถ้าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในบริบทรองทำงานเสร็จสิ้นและมีการคืนค่าผลลัพธ์โหนดตัดสินใจที่สร้างกลับมา) โดยการเรียกใช้งานอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในบริบทรองนี้ ให้ส่งค่าเซตย่อย t_i ไปเป็นอินพุตเซตของตัวอย่างฝึกฝน, ส่งค่าเซตสำเนาของเซต A ที่ได้จากการคัดลอกในขั้นตอนที่แล้วไปเป็นอินพุตเซตของแอตทริบิวต์ของตัวอย่างฝึกฝน และส่งค่า N ไปเป็นอินพุตพารามิเตอร์ N

9.4.3) รออัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในบริบทรองทำงานเสร็จสิ้นและมีการคืนค่าผลลัพธ์โหนดตัดสินใจกลับมา

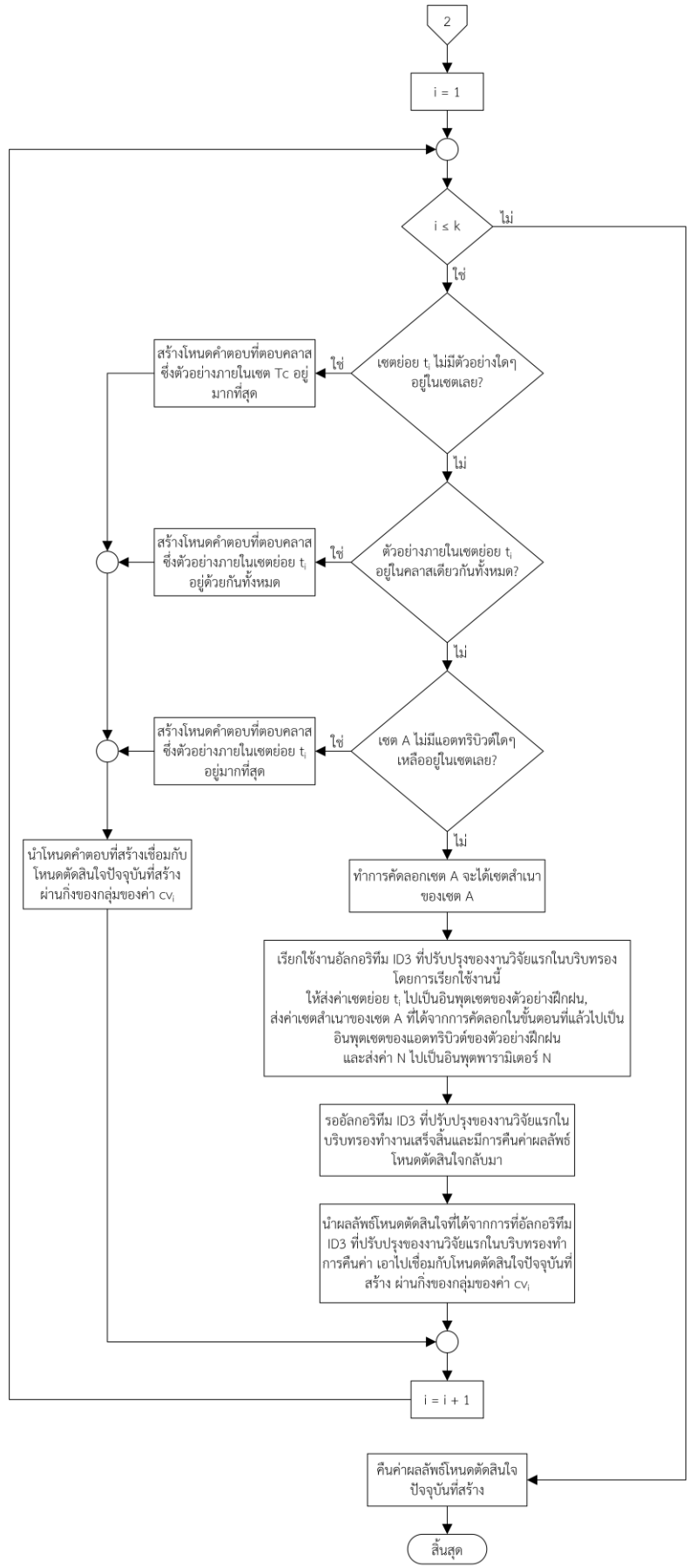
9.4.4) นำผลลัพธ์โหนดตัดสินใจที่ได้จากการที่อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในบริบทรองทำการคืนค่า (Return) เอาไปเชื่อมกับโหนดตัดสินใจปัจจุบันที่สร้าง ผ่านกิ่งของกลุ่มของค่า cv_i

10) ทำการคืนค่าผลลัพธ์โหนดตัดสินใจปัจจุบันที่สร้าง และจบการทำงานของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในบริบทปัจจุบัน

ผังงานของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก แสดงดัง 2 รูปถัดไปแบบต่อเนื่องกัน



รูปที่ 4.3 ผังงานของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก (1/2)



รูปที่ 4.4 ผังงานของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก (2/2)

4.2.3 ประสิทธิภาพและข้อจำกัด

อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก ซึ่งใช้วิธีการรวมแอตทริบิวต์ที่มีความสำคัญเท่ากัน สามารถลดความลึกสูงสุดของต้นไม้ตัดสินใจได้ ในขณะที่ความแม่นยำอยู่ในระดับที่ใกล้เคียงกับความแม่นยำของต้นไม้ตัดสินใจที่สร้างได้จากอัลกอริทึม ID3 แบบดั้งเดิม อย่างไรก็ตาม อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก ก็มีข้อจำกัดที่เห็นได้ชัดเจนที่สุดคือ ชุดข้อมูลที่ใช้ในการสร้างต้นไม้ตัดสินใจจากอัลกอริทึมนี้ ต้องมีขนาดใหญ่และมีการกระจายของค่าที่เป็นไปได้ของแอตทริบิวต์ทุกแอตทริบิวต์อย่างมากพอ เพื่อที่จะทำให้มีตัวอย่างฝึกฝนที่สามารถใช้สร้างโหนดลูกตามกิ่งทุกกิ่งที่แตกออกมาจากโหนดตัดสินใจทุกโหนดได้ ทั้งนี้เนื่องจากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกมีการแตกกิ่งออกมาจากโหนดตัดสินใจมากขึ้น ซึ่งการแตกกิ่งออกมาจากโหนดตัดสินใจมากขึ้นนั้น มาจากการที่อัลกอริทึมเองสามารถเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดและเท่ากันหลายแอตทริบิวต์เข้าไปอยู่ในโหนดตัดสินใจเดียวกันได้และแอตทริบิวต์ที่ถูกเลือกพร้อมกันอาจมีค่าที่เป็นไปได้หลายค่า เพราะฉะนั้นอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกจึงมีการแตกกิ่งออกมาจากโหนดตัดสินใจมากขึ้น และการสร้างโหนดลูกของกิ่งที่แตกออกมาหลายกิ่ง แต่ละกิ่งจะต้องใช้ตัวอย่างฝึกฝนที่มีกลุ่มของค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในโหนดตัดสินใจ สอดคล้องกับกลุ่มของค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในโหนดตัดสินใจของกิ่งที่จะทำการสร้างโหนดลูกด้วย ถ้าชุดข้อมูลที่ใช้ในการสร้างต้นไม้ตัดสินใจจากอัลกอริทึมนี้ มีขนาดที่ไม่ใหญ่หรือไม่มีการกระจายของค่าที่เป็นไปได้ของแอตทริบิวต์ทุกแอตทริบิวต์อย่างมากพอ อาจทำให้ตอนสร้างต้นไม้ตัดสินใจมีโอกาสที่กิ่งของโหนดตัดสินใจบางโหนดที่แตกออกมาบางกิ่ง จะไม่มีตัวอย่างฝึกฝนที่สามารถใช้ในการสร้างโหนดลูกต่อไปได้ ซึ่งท้ายที่สุดอาจจะส่งผลทำให้ความแม่นยำในการจำแนกของต้นไม้ตัดสินใจที่ได้จากอัลกอริทึมนี้มีค่าที่น้อยลง

จากการที่อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก มีการแตกกิ่งออกมาจากโหนดตัดสินใจมากขึ้นเนื่องจากอัลกอริทึมนี้สามารถเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดและเท่ากันหลายแอตทริบิวต์เข้าไปอยู่ในโหนดตัดสินใจเดียวกันได้และแอตทริบิวต์ที่ถูกเลือกพร้อมกันอาจมีค่าที่เป็นไปได้หลายค่า ดังนั้นอัลกอริทึมนี้จึงมีข้อจำกัดถัดมาคือ ความต้องการด้านพื้นที่ (Space requirement) ที่ใช้ในการเก็บรายละเอียดของต้นไม้ตัดสินใจที่สร้าง ซึ่งความต้องการด้านพื้นที่ของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก มากกว่าความต้องการด้านพื้นที่ของอัลกอริทึม ID3 แบบดั้งเดิม ข้อจำกัดนี้ไม่สามารถปฏิเสธหรือหลีกเลี่ยงได้ เพราะการแตกกิ่งออกมาจากโหนดตัดสินใจมากขึ้น ย่อมทำให้มีการสร้างโหนดในระดับลูกหลานที่มากขึ้นและนำไปสู่ความต้องการด้านพื้นที่ที่มากขึ้นในที่สุด

นอกจากข้อจำกัดชุดข้อมูลที่ใช้ในการสร้างต้นไม้ตัดสินใจและความต้องการด้านพื้นที่ของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกแล้ว มีข้อจำกัดอีกข้อหนึ่งที่เป็นข้อจำกัดที่สำคัญซึ่งจะนำไปสู่การปรับปรุงต่อยอดจากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก คือ ความตึงของ

วิธีการเลือกแอตทริบิวต์ เป็นข้อจำกัดที่ส่งผลต่อความสามารถในการลดความลึกสูงสุดของต้นไม้ตัดสินใจที่สร้างจากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก ซึ่งจะถูกอธิบายอย่างละเอียดในบทต่อไปของวิทยานิพนธ์นี้

บทที่ 5
อัลกอริทึม ID3 กับแอตทริบิวต์ที่มีความสำคัญใกล้เคียงกัน

5.1 ความตึงของวิธีการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก

ตามที่ได้อธิบายปิดท้ายในบทที่แล้ว อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก [5] มีข้อจำกัดที่สำคัญซึ่งจะอธิบายในบทนี้ ข้อจำกัดนั้นคือ ความตึงของวิธีการเลือกแอตทริบิวต์

ความตึงของวิธีการเลือกแอตทริบิวต์ เป็นข้อจำกัดของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก ซึ่งเกิดจากวิธีการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกเอง ปกติแล้ววิธีการเลือกแอตทริบิวต์เพื่อนำมาอยู่ในโหนดตัดสินใจของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกนั้นมี 2 เงื่อนไข คือ เงื่อนไขแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดเพียงแอตทริบิวต์เดียว และ เงื่อนไขแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากัน ซึ่งเงื่อนไขหลังนี้ อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกทำการจัดการโดยเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันมาพร้อมกัน แล้วนำแอตทริบิวต์ที่ถูกเลือกพร้อมกันไปอยู่บนโหนดตัดสินใจเดียวกัน เงื่อนไขดังกล่าวได้มีการระบุไว้ชัดเจนว่าเป็นเงื่อนไขแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากัน คือต้องการแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันจริงๆ จึงจะสามารถดำเนินการเลือกแอตทริบิวต์เหล่านี้ไปเป็นแอตทริบิวต์บนโหนดตัดสินใจเดียวกันได้ เพราะฉะนั้นเงื่อนไขหลังนี้เป็นสาเหตุที่ทำให้วิธีการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกมีความตึง

ความตึงของวิธีการเลือกแอตทริบิวต์ ส่งผลต่อความสามารถในการลดความลึกสูงสุดของต้นไม้ตัดสินใจที่สร้างจากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก โดยผลกระทบจากความตึงของวิธีการเลือกแอตทริบิวต์ จะมีมากน้อยไม่เหมือนกัน ขึ้นอยู่กับชุดข้อมูลที่ใช้ในการสร้างต้นไม้ตัดสินใจของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก

ชุดข้อมูลที่ใช้ในการสร้างต้นไม้ตัดสินใจนั้น เป็นปัจจัยหลักที่ส่งผลต่อความมากน้อยของผลกระทบจากความตึงของวิธีการเลือกแอตทริบิวต์ โดยถ้าชุดข้อมูลที่ใช้ในการสร้างต้นไม้ตัดสินใจชุดใด เมื่อนำไปใช้ในอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก แล้วโดยเฉลี่ยพบแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันต่อการพิจารณาเลือกแอตทริบิวต์อยู่ในจำนวนที่น้อย ก็จะทำให้อัลกอริทึมเองสามารถเลือกแอตทริบิวต์มาอยู่ในโหนดตัดสินใจเดียวกันต่อโหนดโดยเฉลี่ยได้ในจำนวนที่น้อย และทำให้ผลกระทบจากความตึงของวิธีการเลือกแอตทริบิวต์มีมาก กล่าวคือ ความสามารถในการลดความลึกสูงสุดของต้นไม้ตัดสินใจจะมีน้อย ทำให้ความลึกของต้นไม้ตัดสินใจที่ได้จากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก มีแนวโน้มที่ลดลงในระดับที่น้อย เมื่อวัดจากความลึกของต้นไม้ตัดสินใจที่ได้จากอัลกอริทึม ID3 แบบดั้งเดิม ถ้าชุดข้อมูลที่ใช้ในการสร้างต้นไม้ตัดสินใจชุดใด เมื่อนำไปใช้ใน

อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก แล้วโดยเฉลี่ยพบแอตทริบิวต์ที่มีค่าเกนความรู้สูงที่สุดและเท่ากันต่อการพิจารณาเลือกแอตทริบิวต์อยู่ในจำนวนที่มาก ก็จะทำให้ตัวอัลกอริทึมสามารถเลือกแอตทริบิวต์มาอยู่ในโหนดตัดสินใจเดียวกันต่อโหนดโดยเฉลี่ยได้ในจำนวนที่มาก และทำให้ผลกระทบจากความตึงของวิธีการเลือกแอตทริบิวต์มีน้อย (แต่ไม่ใช่ไม่มี) กล่าวคือ ความสามารถในการลดความลึกสูงสุดของต้นไม้ตัดสินใจจะมีมาก ทำให้ความลึกของต้นไม้ตัดสินใจที่ได้จากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก มีแนวโน้มที่ลดลงในระดับที่มาก เมื่อวัดจากความลึกของต้นไม้ตัดสินใจที่ได้จากอัลกอริทึม ID3 แบบดั้งเดิม

5.2 อัลกอริทึม ID3 ที่ปรับปรุงโดยใช้วิธีการรวมแอตทริบิวต์ที่มีความสำคัญใกล้เคียงกัน

จากข้อจำกัดด้านความตึงของวิธีการเลือกแอตทริบิวต์ที่พบในอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก [5] วิทยานิพนธ์นี้จึงเสนอวิธีการปรับปรุงอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก โดยใช้วิธีการรวมแอตทริบิวต์ที่มีความสำคัญใกล้เคียงกัน เพื่อลดข้อจำกัดด้านความตึงของวิธีการเลือกแอตทริบิวต์ที่พบในอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกให้น้อยลง วิธีการปรับปรุงดังกล่าวเป็นงานวิจัยหลักที่เสนอของวิทยานิพนธ์นี้

แนวคิดของงานวิจัยหลักที่เสนอคือ ทำการปรับปรุงอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกโดยใช้วิธีการรวมแอตทริบิวต์ที่มีความสำคัญใกล้เคียงกัน ซึ่งจะทำการแก้ไขวิธีการเลือกแอตทริบิวต์ให้สามารถเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 ร่วมกับแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ได้ อย่างไรก็ตามแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 จะมีสิทธิ์ในการถูกเลือกก่อน ส่วนแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 จะถูกพิจารณาให้สามารถเลือกร่วมกับแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ได้หลังจากที่ทำการเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ไปอยู่บนโหนดตัดสินใจเดียวกันทุกแอตทริบิวต์แล้ว แต่จำนวนแอตทริบิวต์ที่ถูกเลือกไปอยู่บนโหนดตัดสินใจเดียวกันยังไม่ครบตามจำนวนที่กำหนด โดยรวมแล้วอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอนี้คล้ายกับอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก ต่างกันตรงวิธีการเลือกแอตทริบิวต์เท่านั้น

5.2.1 การปรับปรุงอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกโดยใช้วิธีการรวมแอตทริบิวต์ที่มีความสำคัญใกล้เคียงกัน

วิธีการที่งานวิจัยหลักที่เสนอ ใช้ในการปรับปรุงอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกคือ วิธีการรวมแอตทริบิวต์ที่มีความสำคัญใกล้เคียงกัน แนวคิดของวิธีการนี้อยู่บนพื้นฐานของข้อจำกัดด้านความตึงของวิธีการเลือกแอตทริบิวต์ที่พบในอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก ซึ่งเป็นข้อจำกัดที่เกิดจากวิธีการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกเอง แอตทริบิวต์ที่จะถูกเลือกในอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกนั้นต้องมีค่าเกนความรู้

สูงสุดเป็นอันดับที่ 1 จริงๆ แอตทริบิวต์อื่นที่มีค่าเกนความรู้ใกล้เคียงกับค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ไม่ว่าจะมีค่าเกนความรู้ใกล้เคียงมากขนาดไหนแต่ก็ไม่สามารถถูกเลือกร่วมด้วยได้ ข้อจำกัดด้านความตึงของวิธีการเลือกแอตทริบิวต์นี้ ส่งผลต่อความสามารถในการลดความลึกสูงสุดของต้นไม้ตัดสินใจที่สร้างจากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก ผลกระทบจากความตึงของวิธีการเลือกแอตทริบิวต์นี้จะเห็นได้ชัดเจนที่สุด เมื่อทำการสร้างต้นไม้ตัดสินใจจากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก โดยใช้ชุดข้อมูลที่พบแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ต่อการพิจารณาเลือกแอตทริบิวต์โดยเฉลี่ยอยู่ในจำนวนที่น้อย ซึ่งจะทำให้ผลกระทบดังกล่าวมีมาก กล่าวคือ ทำให้ความสามารถในการลดความลึกสูงสุดของต้นไม้ตัดสินใจที่สร้างจากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกมีน้อย

งานวิจัยหลักที่เสนอทำการปรับปรุงอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก โดยใช้วิธีการรวมแอตทริบิวต์ที่มีความสำคัญใกล้เคียงกัน เพื่อลดข้อจำกัดที่ได้กล่าวถึงไปก่อนหน้าให้น้อยลง ซึ่งหลักการปรับปรุงอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกโดยใช้วิธีการนี้คือ ทำการแก้ไขวิธีการเลือกแอตทริบิวต์ โดยเป็นการเพิ่มความสามารถในการเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 ร่วมกับแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ได้ ซึ่งจะเป็นการเปิดทางให้ตัวอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอนี้ สามารถที่จะพบและเลือกแอตทริบิวต์เข้ามาอยู่ในโหนดตัดสินใจเดียวกันได้มากขึ้น เมื่อเทียบกับวิธีการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก ตัวอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ ยังคงมีการจำกัดจำนวนแอตทริบิวต์สูงสุดที่สามารถเลือกมาอยู่ในโหนดตัดสินใจเดียวกันเหมือนในอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก โดยมีการนำพารามิเตอร์ N จากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกมาใช้ในอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ ซึ่งพารามิเตอร์ N นี้จะทำหน้าที่เดิมคือกำหนดจำนวนแอตทริบิวต์สูงสุดที่สามารถเลือกมาไว้ในโหนดตัดสินใจเดียวกัน

ถึงแม้ว่าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีความสามารถที่จะเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 ร่วมกับแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ได้ แต่แอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 ไม่สามารถที่จะถูกเลือกได้โดยตรง การที่จะเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 ร่วมกับแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ได้นั้นจะต้องผ่านเงื่อนไข 2 อย่างก่อน

เงื่อนไขแรกในการพิจารณาว่าจะสามารถเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 ร่วมกับแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ได้หรือไม่นั้นคือ การเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ก่อนแล้วตรวจสอบจำนวนแอตทริบิวต์ที่ถูกเลือกไปอยู่ในโหนดตัดสินใจเดียวกันว่ามีไม่ครบตามจำนวนที่กำหนดหรือไม่ เนื่องจากแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 นั้น ในมุมมองของอัลกอริทึม ID3 โดยทั่วไปแล้วถือว่าเป็นแอตทริบิวต์ที่มีความสำคัญที่สุด แต่ในอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอนี้ มีการเปิดทางให้แอตทริ

บิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 สามารถที่จะถูกเลือกได้ ซึ่งแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 มีความสำคัญน้อยกว่าแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ดังนั้นแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 จะไม่สามารถถูกเลือกได้ทันที ตัวอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอนี้ จะยึดลำดับความสำคัญของแอตทริบิวต์เป็นหลัก โดยทำการเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ไปอยู่ในโหนดตัดสินใจก่อนเสมอ ซึ่งหลักการเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 นี้ จะเหมือนกับในอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกคือ สามารถเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 พร้อมกันมากกว่า 1 แอตทริบิวต์ได้ อย่างไรก็ตาม จากการที่อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีการจำกัดจำนวนแอตทริบิวต์สูงสุดที่สามารถเลือกมาอยู่ในโหนดตัดสินใจเดียวกัน ภายใต้พารามิเตอร์ N ที่กำหนด ทำให้อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ สามารถที่จะเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 พร้อมกันได้อย่างมากที่สุดไม่เกิน N แอตทริบิวต์ หลังจากที่ตัวอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอได้ทำการเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ไปอยู่ในโหนดตัดสินใจเดียวกันเรียบร้อยแล้ว ตัวอัลกอริทึมนี้ก็จะทำการตรวจสอบว่าแอตทริบิวต์ที่ถูกเลือกไปอยู่ในโหนดตัดสินใจเดียวกันมีไม่ครบตามจำนวนที่กำหนดหรือไม่ โดยจำนวนที่กำหนดในที่นี้จะอิงจากค่าของพารามิเตอร์ N ที่ได้ตั้งไว้ เพราะฉะนั้นเงื่อนไขแรกในการพิจารณาว่าจะสามารถเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 ร่วมกับแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ได้หรือไม่ก็คือ การตรวจสอบว่าหลังจากเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ไปอยู่ในโหนดตัดสินใจเดียวกันเรียบร้อยแล้ว จำนวนแอตทริบิวต์ที่ถูกเลือกไปอยู่ในโหนดตัดสินใจเดียวกันมีไม่ครบ N แอตทริบิวต์หรือไม่

ถ้าจำนวนแอตทริบิวต์ที่ถูกเลือกไปอยู่ในโหนดตัดสินใจเดียวกันมีครบ N แอตทริบิวต์แล้ว อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอก็จะหยุดกระบวนการเลือกแอตทริบิวต์ทันทีโดยที่ไม่มีการตรวจสอบเงื่อนไขถัดไปของการพิจารณาว่าจะเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 ร่วมกับแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ได้หรือไม่ เนื่องจากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอสามารถเลือกแอตทริบิวต์เข้ามาอยู่ในโหนดตัดสินใจเดียวกันได้เพียง N แอตทริบิวต์เท่านั้น ไม่สามารถเลือกเกินกว่านี้ได้ แต่ถ้าจำนวนแอตทริบิวต์ที่ถูกเลือกไปอยู่ในโหนดตัดสินใจเดียวกันยังมีไม่ครบ N แอตทริบิวต์ อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอก็จะทำการตรวจสอบเงื่อนไขที่สอง เพื่อที่จะตัดสินว่าจะสามารถเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 ร่วมกับแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ได้หรือไม่

เงื่อนไขที่สองในการพิจารณาว่าจะสามารถเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 ร่วมกับแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ได้หรือไม่นั้นคือ การตรวจสอบค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 เมื่อเทียบกับค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ว่ามีค่าต่างกันไม่เกินร้อยละที่กำหนดของ

ค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 หรือไม่ เหตุผลที่ต้องทำการตรวจสอบเงื่อนไขนี้คือ แอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 ควรมีค่าเกนความรู้ที่ต่างจากค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 อยู่ในระดับที่ไม่มากเกินไป จึงจะเหมาะแก่การเลือกมาเป็นแอตทริบิวต์บนโหนดตัดสินใจร่วมกับแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 แอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 ยิ่งมีค่าเกนความรู้ที่ต่างจากค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 น้อยลงเท่าไร หรือมีความใกล้เคียงกับค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 มากเท่าไร ต้นไม้ตัดสินใจที่ได้จากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอเมื่อดำเนินการเลือกแอตทริบิวต์ที่มีลักษณะดังกล่าว จะมีความแม่นยำในการจำแนกอยู่ในระดับคงเดิมหรือใกล้เคียงกับความแม่นยำในการจำแนกของต้นไม้ตัดสินใจที่ได้จากอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกมากขึ้น ในทางกลับกันแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 ยิ่งมีค่าเกนความรู้ที่ต่างจากค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 มากขึ้นเท่าไร หรือมีความใกล้เคียงกับค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 น้อยเท่าไร ต้นไม้ตัดสินใจที่ได้จากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอเมื่อดำเนินการเลือกแอตทริบิวต์ที่มีลักษณะดังกล่าว อาจจะมีความแม่นยำในการจำแนกที่น้อยลง เมื่อเทียบกับความแม่นยำในการจำแนกของต้นไม้ตัดสินใจที่ได้จากอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก เพราะฉะนั้นการเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 เลยโดยที่ไม่ตรวจสอบความต่างของค่าเกนความรู้เมื่อเทียบกับแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 อาจส่งผลต่อความแม่นยำในการจำแนกของต้นไม้ตัดสินใจที่ได้จากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอซึ่งมีค่าน้อยลงเมื่อเทียบกับความแม่นยำในการจำแนกของต้นไม้ตัดสินใจที่ได้จากอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก ด้วยสาเหตุดังกล่าวจึงเป็นที่มาของการตรวจสอบในเงื่อนไขที่สองของการพิจารณาว่าจะสามารถเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 ร่วมกับแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ได้หรือไม่

เงื่อนไขที่สองในการพิจารณาว่าจะสามารถเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 ร่วมกับแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ได้หรือไม่ จะทำการตรวจสอบความใกล้เคียงของค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 เมื่อเทียบกับค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ว่ามีความใกล้เคียงกันหรือไม่ ซึ่งความใกล้เคียงนี้จะวัดจากผลต่างระหว่าง ค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 และค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ว่ามีค่าต่างกันไม่เกินร้อยละที่กำหนดของค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 หรือไม่ ร้อยละที่กำหนดดังกล่าวนั้น จะอิงจากค่าของพารามิเตอร์ที่ 2 ของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ พารามิเตอร์นี้คือ D ทำหน้าที่ในการกำหนดร้อยละของค่าเกนความรู้ของแอ

ตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 เพื่อใช้วัดความต่างระหว่างค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 และค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 (รายละเอียดของพารามิเตอร์ที่มีในอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอจะกล่าวถึงอีกครั้งในส่วนที่อธิบายเกี่ยวกับพารามิเตอร์) ดังนั้นการตรวจสอบในเงื่อนไขที่สองนี้ก็จะวัดความใกล้เคียงหรือความต่างจากพารามิเตอร์ D เป็นหลัก ถ้าค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 เมื่อเทียบกับค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 แล้วพบว่ามีค่าต่างกันเกินร้อยละ D ของค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอจะไม่ทำการเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 เข้าไปอยู่ในโหนดตัดสินใจเดียวกันร่วมกับแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 แต่ถ้าค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 เมื่อเทียบกับค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 แล้วพบว่ามีค่าต่างกันไม่เกินร้อยละ D ของค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอจะทำการเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 เข้าไปอยู่ในโหนดตัดสินใจเดียวกันร่วมกับแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ซึ่งหลักการเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 นี้ จะเหมือนกับหลักการเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอคือสามารถเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 พร้อมกันมากกว่า 1 แอตทริบิวต์ได้ อย่างไรก็ตาม จากการที่อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีการจำกัดจำนวนแอตทริบิวต์สูงสุดที่สามารถเลือกมาอยู่ในโหนดตัดสินใจเดียวกัน ภายใต้พารามิเตอร์ N ที่กำหนด และมีการเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 เข้าไปอยู่ในโหนดตัดสินใจเดียวกันก่อนแล้ว ทำให้อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ สามารถที่จะเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 พร้อมกันได้อย่างมากที่สุดไม่เกิน N แอตทริบิวต์ ลบด้วยจำนวนแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ที่ถูกเลือกไปอยู่ในโหนดตัดสินใจเดียวกันก่อนแล้ว อนึ่ง ถ้าไม่มีแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 อยู่ในเซตของแอตทริบิวต์ที่กำลังพิจารณา อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอจะหยุดกระบวนการเลือกแอตทริบิวต์ทันที เนื่องจากไม่สามารถทำการตรวจสอบเงื่อนไขที่ 2 ของการพิจารณาว่าจะสามารถเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 ร่วมกับแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ได้หรือไม่ต่อไปได้

หลังจากทำการตรวจสอบเงื่อนไขทั้งสองของการพิจารณาว่าจะสามารถเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 ร่วมกับแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ได้หรือไม่ และได้ทำการเลือกหรือไม่เลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 ร่วมกับแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ตามเงื่อนไขที่สอดคล้องแล้ว อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอก็จะดำเนินการขั้นตอนในส่วนของการแตกกิ่งออกมาจากโหนดตัดสินใจและ

แบ่งตัวอย่างฝึกฝนทั้งหมดที่อยู่ในเซตของตัวอย่างฝึกฝนเพื่อไปอยู่ในกิ่งที่แตกออกมาต่อไป วิธีการดำเนินการขั้นตอนในส่วนนี้จะเหมือนกับในอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกทุกประการ

5.2.2 พารามิเตอร์

อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีพารามิเตอร์ 2 พารามิเตอร์คือ N และ D โดยพารามิเตอร์แรก N เป็นพารามิเตอร์ที่มีอยู่ในอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก และได้นำมาใช้ต่อในอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ พารามิเตอร์ N นี้ยังคงทำหน้าที่เดิมคือ กำหนดจำนวนแอตทริบิวต์สูงสุดที่สามารถเลือกมาไว้ในโหนดตัดสินใจเดียวกัน ทำให้อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอสามารถจำกัดจำนวนแอตทริบิวต์สูงสุดที่สามารถเลือกมาอยู่ในโหนดตัดสินใจเดียวกันได้ สำหรับในส่วนของพารามิเตอร์ถัดไปคือ D เป็นพารามิเตอร์ใหม่ที่กำหนดขึ้นในอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ พารามิเตอร์ D ทำหน้าที่กำหนดร้อยละของค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 เพื่อใช้วัดความต่างระหว่างค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 และค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ทำให้อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอสามารถกรองแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 ที่มีสิทธิถูกเลือก ตามระดับความใกล้เคียงของค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 เมื่อเทียบกับค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ได้

5.2.3 ขั้นตอนการทำงานของอัลกอริทึม ID3 ที่ปรับปรุงโดยใช้วิธีการรวมแอตทริบิวต์ที่มีความสำคัญใกล้เคียงกัน

อัลกอริทึม ID3 ที่ปรับปรุงโดยใช้วิธีการรวมแอตทริบิวต์ที่มีความสำคัญใกล้เคียงกัน (อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ) ต้องการอินพุต 4 อย่างคือ เซตของตัวอย่างฝึกฝน , เซตของแอตทริบิวต์ของตัวอย่างฝึกฝน, พารามิเตอร์ N (จำนวนแอตทริบิวต์สูงสุดที่สามารถเลือกมาไว้ในโหนดตัดสินใจเดียวกัน) และพารามิเตอร์ D (ร้อยละของค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 สำหรับใช้วัดความต่างระหว่างค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 และค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1) ลักษณะการทำงานของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอคล้ายกับของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก ต่างกันตรงวิธีการเลือกแอตทริบิวต์ ซึ่งมีการเพิ่มความสามารถให้แอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 มีสิทธิถูกเลือกไปอยู่ในโหนดตัดสินใจเดียวกันร่วมกับแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ขั้นตอนการทำงานของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมีดังนี้ กำหนดให้ T คือเซตของตัวอย่างฝึกฝนที่รับค่าเข้ามา, A คือเซตของแอตทริบิวต์ของตัวอย่างฝึกฝนที่รับค่าเข้ามา, Tc คือเซตที่เก็บสำเนาของตัวอย่างทั้งหมดของเซต T, selAttrs

คือ เซตของแอตทริบิวต์ที่อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอเลือกและ combVals คือ เซตของกลุ่มของค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในเซต selAttrs

1) ทำการคำนวณค่าเกนความรู้ของทุกแอตทริบิวต์ที่อยู่ในเซต A โดยใช้สมการ (2.5) – (2.7) ร่วมกับเซต T
2) ทำการเลือกแอตทริบิวต์ในเซต A ซึ่งเป็นแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 มาเก็บไว้ในเซต selAttrs ตามเงื่อนไขที่สอดคล้อง
    2.1) ถ้าแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 มีเพียงแอตทริบิวต์เดียว ให้ทำการเลือกแอตทริบิวต์ดังกล่าวได้ทันที
    2.2) ถ้าแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 มีมากกว่า 1 แอตทริบิวต์
        2.2.1) ถ้าจำนวนของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 มีน้อยกว่าหรือเท่ากับ N แอตทริบิวต์ ให้ทำการเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ทั้งหมดได้ทันที
        2.2.2) ถ้าจำนวนของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 มีมากกว่า N แอตทริบิวต์ ให้ทำการสุ่มเลือก N แอตทริบิวต์จากแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ทั้งหมด
3) ทำการตรวจสอบจำนวนแอตทริบิวต์ที่อยู่ในเซต selAttrs
    3.1) ถ้าจำนวนแอตทริบิวต์ที่อยู่ในเซต selAttrs มีครบ N แอตทริบิวต์แล้ว ให้ข้ามไปทำขั้นตอนที่ 7 ทันที
    3.2) ถ้าจำนวนแอตทริบิวต์ที่อยู่ในเซต selAttrs มีไม่ครบ N แอตทริบิวต์ ให้ทำขั้นตอนถัดไป
4) ทำการตรวจสอบการมีอยู่ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 ในเซต A
    4.1) ถ้าไม่มีแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 อยู่ในเซต A ให้ข้ามไปทำขั้นตอนที่ 7 ทันที
    4.2) ถ้ามีแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 อยู่ในเซต A ให้ทำขั้นตอนถัดไป

5) ทำการตรวจสอบความต่างระหว่างค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 และค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ในเซต A
   5.1) ถ้าค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 ต่างจากค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 เกินร้อยละ D ของค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ให้ข้ามไปทำขั้นตอนที่ 7 ทันที
   5.2) ถ้าค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 ต่างจากค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ไม่เกินร้อยละ D ของค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ให้ทำขั้นตอนถัดไป
6) ทำการเลือกแอตทริบิวต์ในเซต A ซึ่งเป็นแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 มาเก็บไว้ในเซต selAttrs ตามเงื่อนไขที่สอดคล้อง
   6.1) ถ้าแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 มีเพียงแอตทริบิวต์เดียว ให้ทำการเลือกแอตทริบิวต์ดังกล่าวได้ทันที
   6.2) ถ้าแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 มีมากกว่า 1 แอตทริบิวต์
      6.2.1) ถ้าจำนวนของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 มีน้อยกว่าหรือเท่ากับ N แอตทริบิวต์ ลบด้วยจำนวนแอตทริบิวต์ที่อยู่ในเซต selAttrs ให้ทำการเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 ทั้งหมดได้ทันที
      6.2.2) ถ้าจำนวนของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 มีมากกว่า N แอตทริบิวต์ ลบด้วยจำนวนแอตทริบิวต์ที่อยู่ในเซต selAttrs ให้ทำการสุ่มเลือก N แอตทริบิวต์ ลบด้วยจำนวนแอตทริบิวต์ที่อยู่ในเซต selAttrs จากแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 ทั้งหมด
7) ทำการสร้างโหนดตัดสินใจ แล้วใช้แอตทริบิวต์ที่อยู่ในเซต selAttrs ทั้งหมดเป็นแอตทริบิวต์บนโหนดตัดสินใจที่สร้าง
8) นำแอตทริบิวต์ที่ถูกเลือก ซึ่งเป็นแอตทริบิวต์ที่อยู่ในเซต selAttrs ทั้งหมดออกจากเซต A ; A = A – selAttrs เนื่องจากจะไม่มีการนำแอตทริบิวต์ที่เคยเลือกไปแล้วมาพิจารณาซ้ำในการสร้างโหนดตัดสินใจอื่นๆ ที่อยู่ในระดับลูกหลานของโหนดตัดสินใจปัจจุบันที่เพิ่งสร้างไป

9) ทำการคัดลอกตัวอย่างภายในเซต T ทั้งหมดไปใส่ในเซต Tc

10) ทำการพิจารณาหากลุ่มของค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในเซต selAttrs ทุกกลุ่มที่เป็นไปได้ แล้วนำกลุ่มของค่าดังกล่าวที่หาได้ทั้งหมดไปเก็บไว้ในเซต combVals โดยพิจารณาหากลุ่มของค่าดังกล่าวตามเงื่อนไขของจำนวนแอตทริบิวต์ที่อยู่ในเซต selAttrs

   10.1) ถ้าแอตทริบิวต์ที่อยู่ในเซต selAttrs มีเพียงแอตทริบิวต์เดียว ให้ทำการหากลุ่มของค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในเซต selAttrs ผ่านการใช้ค่าที่เป็นไปได้ทุกค่าของแอตทริบิวต์ที่อยู่ในเซต selAttrs เป็นกลุ่มของค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในเซต selAttrs ได้โดยตรง

   10.2) ถ้าแอตทริบิวต์ที่อยู่ในเซต selAttrs มีมากกว่า 1 แอตทริบิวต์ ให้ทำการหากลุ่มของค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในเซต selAttrs ผ่านการใช้ผลคูณคาร์ทีเซียนกับเซตของค่าที่เป็นไปได้ของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในเซต selAttrs

11) ใช้เซต combVals ในการแบ่งเซต T ตามกลุ่มของค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในเซต selAttrs ที่เป็นไปได้ จะได้เซตย่อย k เซต ; $t_1 - t_k$ โดยที่ k คือจำนวนกลุ่มของค่าทั้งหมดในเซต combVals, $cv_1 - cv_k$ คือกลุ่มของค่าที่อยู่ในเซต combVals ที่เป็นไปได้ทั้งหมด และ $t_1 - t_k$ คือเซตย่อยที่ตัวอย่างฝึกฝนทุกตัวอย่างภายในเซต มีค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในเซต selAttrs เป็น $cv_1 - cv_k$ ตามลำดับ

12) ทำการแตกกิ่ง (Branch) ออกมาจากโหนดตัดสินใจปัจจุบันที่สร้าง จำนวน k กิ่ง แต่ละกิ่งระบุว่าเป็นกิ่งของกลุ่มของค่าของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในเซต selAttrs กลุ่มใด ไล่ตั้งแต่ กิ่งแรกระบุเป็นกิ่งของกลุ่มของค่า cv_1 ไปจนถึงกิ่งสุดท้ายระบุเป็นกิ่งของกลุ่มของค่า cv_k

13) ทำการตรวจสอบเซตย่อยแต่ละเซตที่แบ่งได้ กำหนดให้ t_i คือเซตย่อยที่กำลังพิจารณา และ i มีค่าไล่ตามลำดับตั้งแต่ 1 ไปถึง k

   13.1) ถ้าเซตย่อย t_i ไม่มีตัวอย่างใดๆ อยู่ในเซตเลย ให้ทำการพิจารณาตัวอย่างภายในเซต Tc ดูว่าภายในเซต Tc ตัวอย่างส่วนใหญ่มีสัดส่วนอยู่ในคลาสใดมากที่สุด ให้ทำการสร้างโหนดคำตอบที่ตอบคลาสดังกล่าว แล้วนำโหนดคำตอบที่สร้างเชื่อมกับโหนดตัดสินใจปัจจุบันที่ได้สร้างไป ผ่านกิ่งของกลุ่มของค่า cv_i หากเงื่อนไขในขั้นตอนนี้ไม่เป็นจริง ให้ทำการตรวจสอบเซตย่อย t_i ตามเงื่อนไขในขั้นตอนถัดไป

13.2) ถ้าตัวอย่างภายในเซตย่อย t_i อยู่ในคลาสเดียวกันทั้งหมด ให้ทำการสร้างโหนดคำตอบที่ตอบคลาสดังกล่าว แล้วนำโหนดคำตอบที่สร้างเชื่อมกับโหนดตัดสินใจปัจจุบันที่ได้สร้างไป ผ่านกิ่งของกลุ่มของค่า cv_i หากเงื่อนไขในขั้นตอนนี้ไม่เป็นจริง ให้ทำการตรวจสอบเซตย่อย t_i ตามเงื่อนไขในขั้นตอนถัดไป

13.3) ถ้าเซต A ไม่มีแอตทริบิวต์ใดๆ เหลืออยู่ในเซตเลย ส่งผลให้อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอไม่สามารถดำเนินการเลือกแอตทริบิวต์ ตามเงื่อนไขของตัวอัลกอริทึมที่กำหนดต่อไปได้อีก ทำให้ไม่สามารถสร้างโหนดตัดสินใจได้ และเมื่อสร้างโหนดตัดสินใจไม่ได้ ตัวอัลกอริทึมเองก็ย่อมไม่สามารถที่จะแบ่งเซตย่อย t_i ต่อไปได้อีกเช่นกัน หากเงื่อนไขในขั้นตอนนี้เป็นจริง ให้ทำการพิจารณาตัวอย่างภายในเซตย่อย t_i ดูว่าภายในเซตย่อย t_i ตัวอย่างส่วนใหญ่มีสัดส่วนอยู่ในคลาสใดมากที่สุด ให้ทำการสร้างโหนดคำตอบที่ตอบคลาสดังกล่าว แล้วนำโหนดคำตอบที่สร้างเชื่อมกับโหนดตัดสินใจปัจจุบันที่ได้สร้างไป ผ่านกิ่งของกลุ่มของค่า cv_i หากเงื่อนไขในขั้นตอนนี้ไม่เป็นจริง ให้ทำตามขั้นตอนสุดท้ายของการตรวจสอบเซตย่อย t_i ได้เลย

13.4) เนื่องจากตัวอย่างภายในเซตย่อย t_i ไม่ได้อยู่ในคลาสเดียวกัน และยังคงมีแอตทริบิวต์ในเซต A เหลืออยู่

13.4.1) ทำการคัดลอกเซต A จะได้เซตสำเนาของเซต A

13.4.2) ทำการเรียกใช้งานอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในบริบทรองที่แยกต่างหากเลย (ไม่ใช่การวนกลับไปทำขั้นตอนที่ 1 ของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในบริบทปัจจุบัน แต่เป็นการเรียกใช้งานอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในบริบทรอง ซึ่งกระทำกับอินพุตที่ต่างกัน บริบทรองนี้จะไม่ทับกับบริบทปัจจุบัน ดังนั้นการทำงานของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในบริบทปัจจุบันยังคงสามารถกลับมาดำเนินการต่อได้ ถ้าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในบริบทรองทำงานเสร็จสิ้นและมีการคืนค่าผลลัพธ์โหนดตัดสินใจที่สร้างกลับมา) โดยการเรียกใช้งานอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในบริบทรองนี้ ให้ส่งค่าเซตย่อย t_i ไปเป็นอินพุตเซต

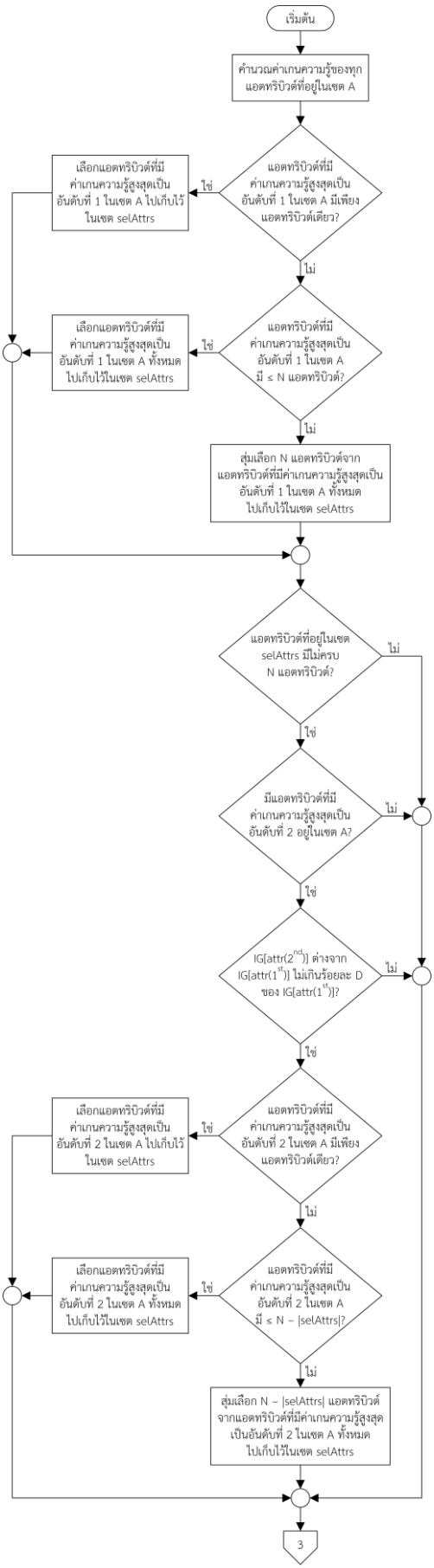
ของตัวอย่างฝึกฝน, ส่งค่าเซตสำเนาของเซต A ที่ได้จากการคัดลอกในขั้นตอนที่แล้วไปเป็นอินพุตเซตของแอตทริบิวต์ของตัวอย่างฝึกฝน, ส่งค่า N ไปเป็นอินพุตพารามิเตอร์ N และส่งค่า D ไปเป็นอินพุตพารามิเตอร์ D

13.4.3) รออัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในบริบทรองทำงานเสร็จสิ้นและมีการคืนค่าผลลัพธ์โหนดตัดสินใจกลับมา

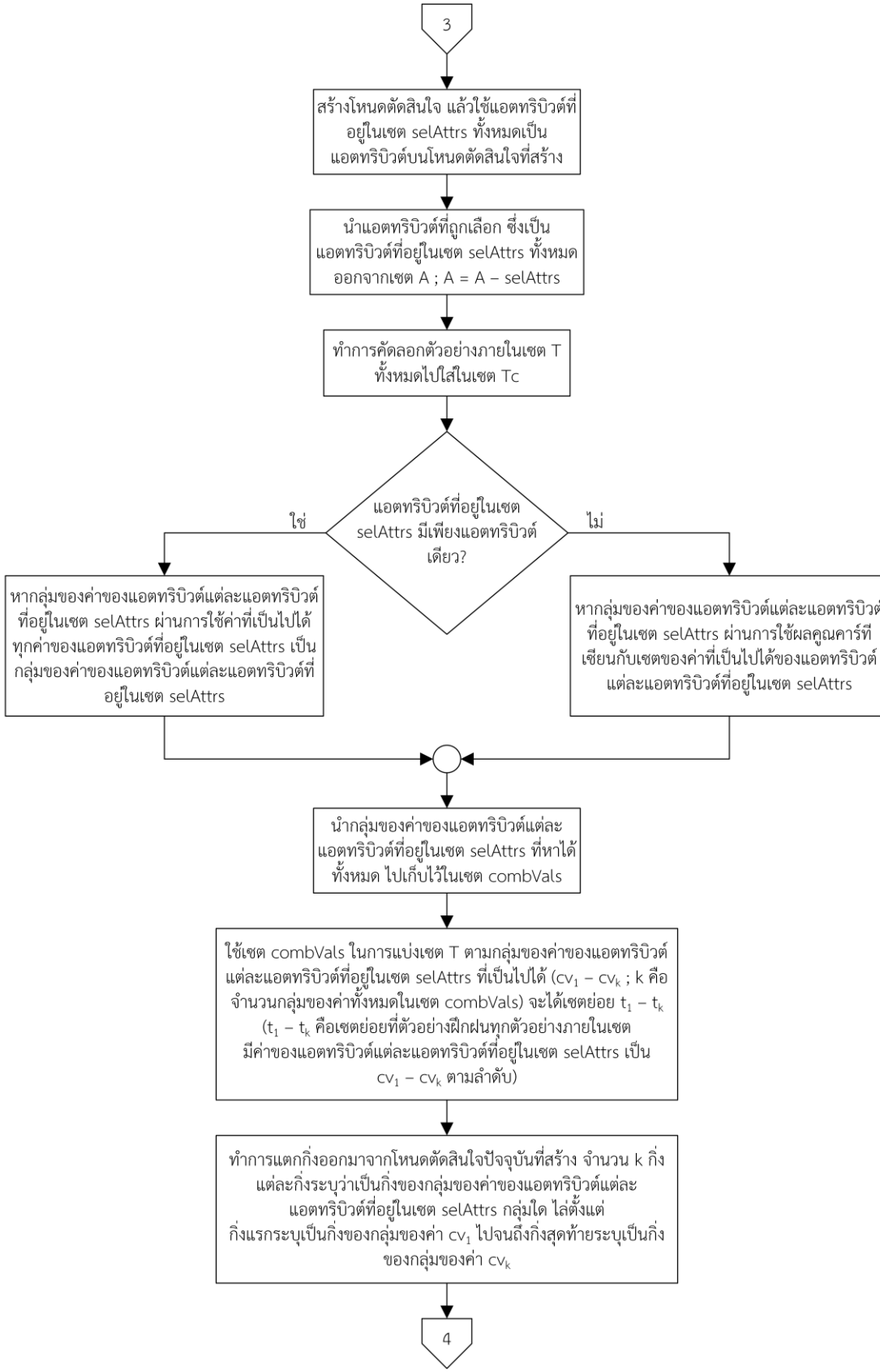
13.4.4) นำผลลัพธ์โหนดตัดสินใจที่ได้จากการที่อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในบริบทรองทำการคืนค่า (Return) เอาไปเชื่อมกับโหนดตัดสินใจปัจจุบันที่สร้าง ผ่านกิ่งของกลุ่มของค่า cv_i

14) ทำการคืนค่าผลลัพธ์โหนดตัดสินใจปัจจุบันที่สร้าง และจบการทำงานของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในบริบทปัจจุบัน

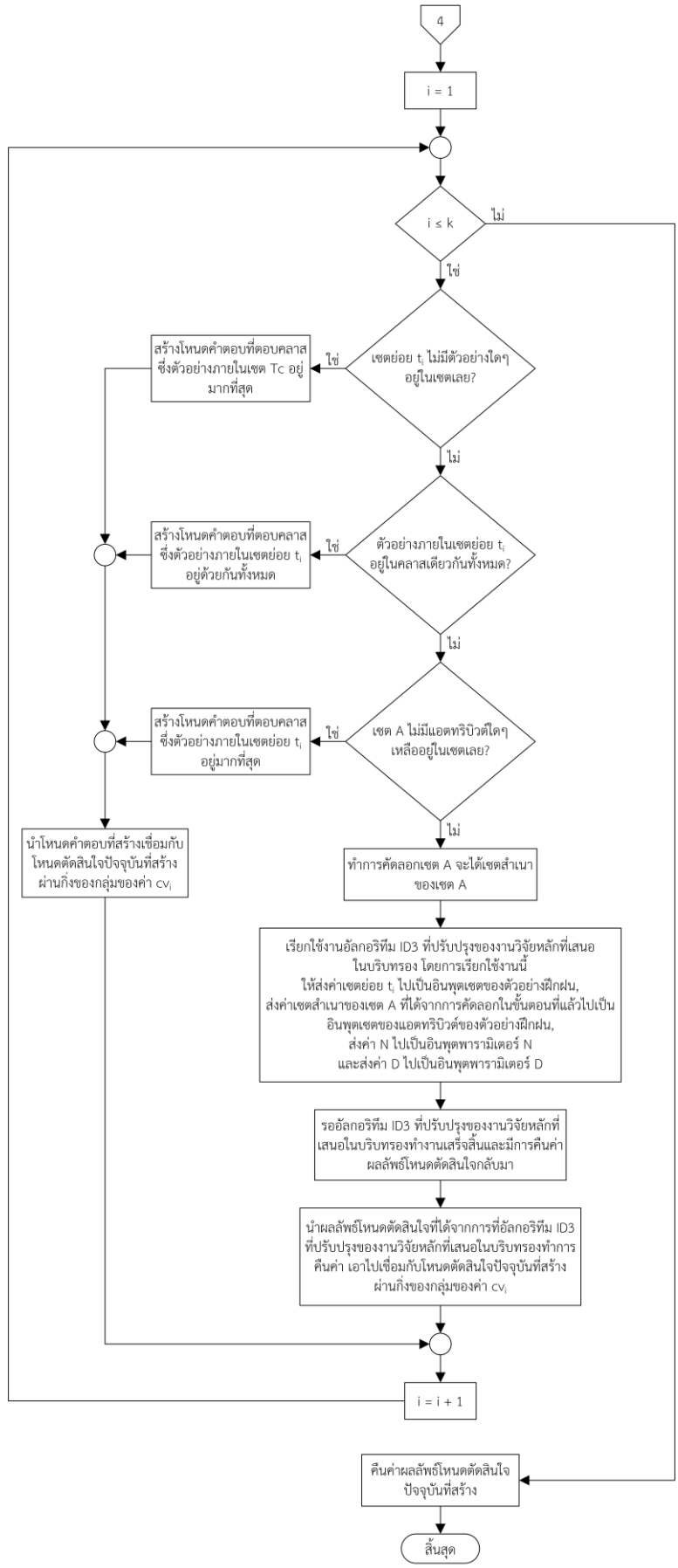
ผังงานของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ แสดงดัง 3 รูปถัดไปแบบต่อเนื่องกัน โดย IG[attr(1^{st})] คือ ค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1, IG[attr(2^{nd})] คือ ค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 2 และ |selAttrs| คือ จำนวนแอตทริบิวต์ที่อยู่ในเซต selAttrs



รูปที่ 5.1 ผังงานของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ (1/3)



รูปที่ 5.2 ผังงานของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ (2/3)



รูปที่ 5.3 ผังงานของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ (3/3)

บทที่ 6
ผลการทดลอง

6.1 ชุดข้อมูลที่ใช้ในการทดลอง

การทดลองอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ (อัลกอริทึม ID3 ที่ปรับปรุงโดยใช้วิธีการรวมแอตทริบิวต์ที่มีความสำคัญใกล้เคียงกัน) จะใช้ชุดข้อมูลจำนวน 6 ชุดข้อมูล ซึ่งเป็นชุดข้อมูลมาตรฐาน (Benchmark dataset) จาก UCI Machine Learning Repository [6] คุณสมบัติของชุดข้อมูลที่ใช้ในการทดลองทั้งหมดแสดงดังตารางถัดไป

ตารางที่ 6.1 คุณสมบัติของชุดข้อมูลที่ใช้ในการทดลอง

ชุดข้อมูล	จำนวนตัวอย่าง	จำนวนแอตทริบิวต์	จำนวนคลาส
Chess (King-Rook vs. King-Pawn)	3,196	36	2
Connect-4	67,557	42	3
Firm-Teacher	10,800	16	7
Phishing Websites	11,055	30	2
Insurance Company Benchmark	9,822	85	2
Bach Choral Harmony	5,665	14	102

6.2 เงื่อนไขในการทดลอง

6.2.1 การแบ่งชุดข้อมูลที่ใช้ในการทดลอง

ชุดข้อมูลที่ใช้ในการทดลองทั้งหมด จะถูกแบ่งออกเป็น 2 ส่วนแบบ 50% : 50% คือ ส่วนที่เป็นชุดข้อมูลฝึกฝนและส่วนที่เป็นชุดข้อมูลทดสอบ การแบ่งนั้นจะพยายามรักษาสมดุลของจำนวนตัวอย่างที่อยู่ในแต่ละคลาสของทั้งสองส่วน ให้มีจำนวนที่ใกล้เคียงกันมากที่สุด ตัวอย่างเช่น ชุดข้อมูล A มี 2 คลาส คือ Yes และ No คลาส Yes มี 10 ตัวอย่าง ในขณะที่คลาส No มี 9 ตัวอย่าง ใช้วิธีการแบ่งชุดข้อมูลตามวิธีข้างต้น จะได้ชุดข้อมูลฝึกฝนซึ่งประกอบไปด้วยตัวอย่างที่อยู่ในคลาส Yes 5 ตัวอย่างและตัวอย่างที่อยู่ในคลาส No 4 ตัวอย่างรวมทั้งหมด 9 ตัวอย่าง ส่วนตัวอย่างที่เหลือในชุดข้อมูล A ที่ไม่ได้แบ่งมาอยู่ในชุดข้อมูลฝึกฝนก็จะถูกแบ่งไปไว้ในชุดข้อมูลทดสอบ ดังนั้นชุดข้อมูลทดสอบจะประกอบไปด้วยตัวอย่างที่อยู่ในคลาส Yes 5 ตัวอย่างและตัวอย่างที่อยู่ในคลาส No 5 ตัวอย่างรวมทั้งหมด 10 ตัวอย่าง จะเห็นว่าชุดข้อมูลทดสอบมีตัวอย่างมากกว่าชุดข้อมูลฝึกฝน เพราะสัดส่วนของจำนวนตัวอย่างในชุดข้อมูล A ที่อยู่ในคลาส No เป็นเลขคี่ ทำให้ไม่สามารถแบ่งตัวอย่างในชุดข้อมูล A ที่อยู่ในคลาส No ไปอยู่ในชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบได้อย่างลงตัว ดังนั้นในการ

ทดลองของวิทยานิพนธ์นี้ จึงได้จัดการกับการแบ่งชุดข้อมูลที่มีลักษณะดังกล่าว ด้วยการแบ่งตัวอย่างส่วนเกินที่ไม่ลงตัวของแต่ละคลาสไปไว้ในชุดข้อมูลทดสอบทั้งหมด

6.2.2 รูปแบบการทดลอง

การทดลองอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ จะทำการทดลองกับแต่ละชุดข้อมูลที่เตรียมมา โดยหลังจากที่แบ่งชุดข้อมูลตามวิธีการที่ได้อธิบายไปในหัวข้อที่แล้ว ก็จะเริ่มกระบวนการทดลองโดยสร้างต้นไม้ตัดสินใจจากชุดข้อมูลฝึกฝน และทำการทดสอบต้นไม้ตัดสินใจที่สร้างด้วยชุดข้อมูลทดสอบ กระบวนการนี้จะทำทั้งหมด 1,000 ครั้งโดยใช้ชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบแบบเดียวกันทั้งหมด (ไม่มีการแบ่งชุดข้อมูลใหม่) ผลการทดลองที่ได้จากการสร้างและทดสอบต้นไม้ตัดสินใจของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ จะนำไปเปรียบเทียบกับอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก [5] ซึ่งใช้ชุดข้อมูลแบบเดียวกันและรูปแบบการทดลองที่เหมือนกัน

6.2.3 พารามิเตอร์

อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอต้องการพารามิเตอร์ 2 ตัวคือ N (จำนวนแอตทริบิวต์สูงสุดที่สามารถเลือกมาไว้ในโหนดตัดสินใจเดียวกัน) และ D (ร้อยละของค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1) การทดลองนี้จะกำหนดพารามิเตอร์ที่ใช้ในการทดลองของแต่ละชุดข้อมูล แล้วนำผลลัพธ์ที่ได้จากการทดลองด้วยพารามิเตอร์ที่กำหนดดังกล่าว ไปเปรียบเทียบกับอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก ซึ่งอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกจะใช้พารามิเตอร์ N เป็นค่าเดียวกันกับที่ใช้ในอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ สำหรับการกำหนดพารามิเตอร์ที่ใช้ในการทดลองอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอนั้น จะทำการกำหนดผ่านการหาพารามิเตอร์ N และ D ที่ดีที่สุดจากการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในชุดข้อมูลทั้งหมด การทดลองเบื้องต้นจะทำการสร้างและทดสอบต้นไม้ตัดสินใจภายใต้พารามิเตอร์ N และ D หลายแบบเพื่อเลือกพารามิเตอร์ N และ D ที่ดีที่สุดของแต่ละชุดข้อมูล โดยการทดลองเบื้องต้นนี้จะทำการทดลองทั้งหมด 1,000 ครั้ง โดยการเลือกพารามิเตอร์ N และ D ที่ดีที่สุดจะพิจารณาจากเกณฑ์ในการวัดเรียงตามลำดับต่อไปนี้

เกณฑ์ในการวัดที่ 1 ความแม่นยำโดยเฉลี่ย : เลือกพารามิเตอร์แบบที่ให้ความแม่นยำโดยเฉลี่ยมากที่สุด หากความแม่นยำโดยเฉลี่ยสูงสุดมีค่าใกล้เคียงกันกับความแม่นยำโดยเฉลี่ยของพารามิเตอร์แบบอื่นๆ ให้ทำการพิจารณาเกณฑ์ถัดไป

เกณฑ์ในการวัดที่ 2 ความลึกสูงสุดโดยเฉลี่ย : เลือกพารามิเตอร์แบบที่ให้ความลึกสูงสุดโดยเฉลี่ยน้อยที่สุด หากความลึกสูงสุดโดยเฉลี่ยมีค่าเท่ากันให้ทำการพิจารณาเกณฑ์สุดท้าย

เกณฑ์ในการวัดที่ 3 จำนวนโหนดโดยเฉลี่ย : เลือกพารามิเตอร์แบบที่ให้จำนวนโหนดโดยเฉลี่ยน้อยที่สุด

ในการทดลองเบื้องต้นเพื่อหาพารามิเตอร์ที่ดีที่สุด จะทำการทดลองโดยกำหนดค่าพารามิเตอร์ D เป็น 0.25 (หรือ 25%), 0.5 (หรือ 50%), 0.75 (หรือ 75%), 0.85 (หรือ 85%) และ 0.95 (หรือ 95%) ในแต่ละค่าของพารามิเตอร์ D ที่กำหนด จะนำมาทดลองกับการกำหนดค่าพารามิเตอร์ N ตั้งแต่ 2 ถึง 4 โดยจะทำการแบ่งกลุ่มการทดลองดังตารางถัดไป

ตารางที่ 6.2 การแบ่งกลุ่มการทดลองเบื้องต้นของค่าพารามิเตอร์ D และ N

กลุ่มที่ 1 (D = 0.25)	**กลุ่มที่ 2 (D = 0.5)**	**กลุ่มที่ 3 (D = 0.75)**	**กลุ่มที่ 4 (D = 0.85)**	**กลุ่มที่ 5 (D = 0.95)**
D = 0.25, N = 2	D = 0.5, N = 2	D = 0.75, N = 2	D = 0.85, N = 2	D = 0.95, N = 2
D = 0.25, N = 3	D = 0.5, N = 3	D = 0.75, N = 3	D = 0.85, N = 3	D = 0.95, N = 3
D = 0.25, N = 4	D = 0.5, N = 4	D = 0.75, N = 4	D = 0.85, N = 4	D = 0.95, N = 4

จากตารางข้างบนแสดงการแบ่งกลุ่มการทดลองออกเป็น 5 กลุ่ม ในแต่ละกลุ่มจะมีการทดลอง 3 การทดลอง โดยมีค่าพารามิเตอร์ D เป็นตัวหลัก พิจารณาที่กลุ่มที่ 1 จากตารางข้างบนจะเห็นว่ากลุ่มที่ 1 มีค่าพารามิเตอร์ D = 0.25 เป็นหลัก ภายในกลุ่มที่ 1 จะมีการทดลอง 3 การทดลอง การทดลองแรกคือการกำหนดพารามิเตอร์ D = 0.25 และ N = 2 การทดลองที่สองคือการกำหนดพารามิเตอร์ D = 0.25 และ N = 3 การทดลองสุดท้ายคือการกำหนดพารามิเตอร์ D = 0.25 และ N = 4 แต่ละกลุ่มการทดลองจะทำการหาค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดซึ่งจะทำการเปรียบเทียบการทดลองภายในกลุ่ม โดยใช้เกณฑ์ในการวัดเรียงตามลำดับดังที่ได้อธิบายไว้เพื่อเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละกลุ่มออกมา เมื่อได้ค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละกลุ่มแล้ว สุดท้ายจะทำการเลือกค่าพารามิเตอร์ N และ D ที่ดีที่สุดจากค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละกลุ่ม ซึ่งจะทำการเลือกโดยใช้เกณฑ์ในการวัดเรียงตามลำดับแบบเดียวกันกับที่เลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละกลุ่ม

6.2.3.1 การหาค่าพารามิเตอร์ N และ D ที่ดีที่สุดของชุดข้อมูล Connect-4

ตารางที่ 6.3 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 1 ที่มีค่าพารามิเตอร์ D = 0.25 บนชุดข้อมูล Connect-4

เกณฑ์ในการวัด	**D = 0.25**		
	N = 2	**N = 3**	**N = 4**
ความลึกสูงสุดโดยเฉลี่ย	15.27	13.68	**13.16**
ความแม่นยำโดยเฉลี่ย	72.15%	72.25%	**72.24%**
จำนวนโหนดโดยเฉลี่ย	38,707.49	49,974.29	**65,506.57**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 1 คือ D = 0.25 และ N = 4 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.25 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม

ตารางที่ 6.4 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 2 ที่มีค่าพารามิเตอร์ D = 0.5 บนชุดข้อมูล Connect-4

เกณฑ์ในการวัด	**D = 0.5**		
	N = 2	**N = 3**	**N = 4**
ความลึกสูงสุดโดยเฉลี่ย	13.18	11.68	**11.17**
ความแม่นยำโดยเฉลี่ย	72.43%	72.6%	**72.6%**
จำนวนโหนดโดยเฉลี่ย	40,693.55	55,801.35	**78,241.73**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 2 คือ D = 0.5 และ N = 4 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.5 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม

ตารางที่ 6.5 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 3 ที่มีค่าพารามิเตอร์ D = 0.75 บนชุดข้อมูล Connect-4

เกณฑ์ในการวัด	**D = 0.75**		
	N = 2	**N = 3**	**N = 4**
ความลึกสูงสุดโดยเฉลี่ย	12.9	11.01	**10.44**
ความแม่นยำโดยเฉลี่ย	72.61%	72.81%	**72.82%**
จำนวนโหนดโดยเฉลี่ย	45,787.98	76,391.34	**121,445.86**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 3 คือ D = 0.75 และ N = 4 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.75 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม

ตารางที่ 6.6 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 4 ที่มีค่าพารามิเตอร์ D = 0.85 บนชุดข้อมูล Connect-4

เกณฑ์ในการวัด	**D = 0.85**		
	N = 2	**N = 3**	**N = 4**
ความลึกสูงสุดโดยเฉลี่ย	12.1	10.73	**10.26**
ความแม่นยำโดยเฉลี่ย	72.65%	72.84%	**72.86%**
จำนวนโหนดโดยเฉลี่ย	46,434.56	79,032.2	**129,456.31**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 4 คือ D = 0.85 และ N = 4 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.85 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม

ตารางที่ 6.7 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 5 ที่มีค่าพารามิเตอร์ D = 0.95 บนชุดข้อมูล Connect-4

เกณฑ์ในการวัด	**D = 0.95**		
	N = 2	**N = 3**	**N = 4**
ความลึกสูงสุดโดยเฉลี่ย	11.53	10.42	**10.14**
ความแม่นยำโดยเฉลี่ย	72.66%	72.86%	**72.87%**
จำนวนโหนดโดยเฉลี่ย	46,461.33	79,174.26	**130,123.28**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 5 คือ D = 0.95 และ N = 4 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.95 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม เมื่อได้ค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละกลุ่มแล้ว ต่อมาให้ทำการเลือกค่าพารามิเตอร์ N และ D ที่ดีที่สุดโดยการเปรียบเทียบระหว่างค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละกลุ่ม

ตารางที่ 6.8 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอของค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละกลุ่ม บนชุดข้อมูล Connect-4

เกณฑ์ในการวัด	**กลุ่มที่ 1**	**กลุ่มที่ 2**	**กลุ่มที่ 3**	**กลุ่มที่ 4**	**กลุ่มที่ 5**
	D=0.25, N=4	**D=0.5, N=4**	**D=0.75, N=4**	**D=0.85, N=4**	**D=0.95, N=4**
ความลึกสูงสุดโดยเฉลี่ย	13.16	11.17	10.44	10.26	**10.14**
ความแม่นยำโดยเฉลี่ย	72.24%	72.6%	72.82%	72.86%	**72.87%**
จำนวนโหนดโดยเฉลี่ย	65,506.57	78,241.73	121,445.86	129,456.31	**130,123.28**

จากตารางข้างบนค่าพารามิเตอร์ N และ D ที่ดีที่สุดคือ D = 0.95 และ N = 4 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.95 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D ที่ดีที่สุดของชุดข้อมูล Connect-4

6.2.3.2 การหาค่าพารามิเตอร์ N และ D ที่ดีที่สุดของชุดข้อมูล Chess

ตารางที่ 6.9 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 1 ที่มีค่าพารามิเตอร์ D = 0.25 บนชุดข้อมูล Chess

เกณฑ์ในการวัด	D = 0.25		
	N = 2	N = 3	N = 4
ความลึกสูงสุดโดยเฉลี่ย	7.83	**7**	7
ความแม่นยำโดยเฉลี่ย	99.17%	**99.22%**	99.23%
จำนวนโหนดโดยเฉลี่ย	75.31	**92**	100

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 1 คือ D = 0.25 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.25 และ N = 3 และค่าพารามิเตอร์ D = 0.25 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.25 และ N = 3 มีจำนวนโหนดโดยเฉลี่ยน้อยกว่า ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม

ตารางที่ 6.10 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 2 ที่มีค่าพารามิเตอร์ D = 0.5 บนชุดข้อมูล Chess

เกณฑ์ในการวัด	D = 0.5		
	N = 2	N = 3	N = 4
ความลึกสูงสุดโดยเฉลี่ย	7.33	**7**	7
ความแม่นยำโดยเฉลี่ย	99.11%	**99.12%**	99.12%
จำนวนโหนดโดยเฉลี่ย	81.7	**96.51**	104

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 2 คือ D = 0.5 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.5 และ N = 3 และค่าพารามิเตอร์ D = 0.5 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.5 และ N = 3 มีจำนวนโหนดโดยเฉลี่ยน้อยกว่า ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม

ตารางที่ 6.11 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 3 ที่มีค่าพารามิเตอร์ D = 0.75 บนชุดข้อมูล Chess

เกณฑ์ในการวัด	**D = 0.75**		
	N = 2	**N = 3**	**N = 4**
ความลึกสูงสุดโดยเฉลี่ย	6.37	**6**	6
ความแม่นยำโดยเฉลี่ย	99.08%	**99.06%**	98.97%
จำนวนโหนดโดยเฉลี่ย	89.77	**111**	127

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 3 คือ D = 0.75 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.75 และ N = 3 และค่าพารามิเตอร์ D = 0.75 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.75 และ N = 3 มีจำนวนโหนดโดยเฉลี่ยน้อยกว่า ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม

ตารางที่ 6.12 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 4 ที่มีค่าพารามิเตอร์ D = 0.85 บนชุดข้อมูล Chess

เกณฑ์ในการวัด	**D = 0.85**		
	N = 2	**N = 3**	**N = 4**
ความลึกสูงสุดโดยเฉลี่ย	6.32	**6**	6
ความแม่นยำโดยเฉลี่ย	98.96%	**98.94%**	98.72%
จำนวนโหนดโดยเฉลี่ย	91.66	**117**	141

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 4 คือ D = 0.85 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.85 และ N = 3 และค่าพารามิเตอร์ D = 0.85 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.85 และ N = 3 มีจำนวนโหนดโดยเฉลี่ยน้อยกว่า ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม

ตารางที่ 6.13 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 5 ที่มีค่าพารามิเตอร์ D = 0.95 บนชุดข้อมูล Chess

เกณฑ์ในการวัด	D = 0.95		
	N = 2	N = 3	N = 4
ความลึกสูงสุดโดยเฉลี่ย	6.31	**6**	6
ความแม่นยำโดยเฉลี่ย	98.96%	**98.94%**	98.72%
จำนวนโหนดโดยเฉลี่ย	91.55	**117**	141

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 5 คือ D = 0.95 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.95 และ N = 3 และค่าพารามิเตอร์ D = 0.95 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.95 และ N = 3 มีจำนวนโหนดโดยเฉลี่ยน้อยกว่า ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม เมื่อได้ค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละกลุ่มแล้ว ต่อมาให้ทำการเลือกค่าพารามิเตอร์ N และ D ที่ดีที่สุดโดยการเปรียบเทียบระหว่างค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละกลุ่ม

ตารางที่ 6.14 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอของค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละกลุ่ม บนชุดข้อมูล Chess

เกณฑ์ในการวัด	กลุ่มที่ 1	กลุ่มที่ 2	กลุ่มที่ 3	กลุ่มที่ 4	กลุ่มที่ 5
	D=0.25, N=3	D=0.5, N=3	D=0.75, N=3	D=0.85, N=3	D=0.95, N=3
ความลึกสูงสุดโดยเฉลี่ย	7	7	**6**	6	6
ความแม่นยำโดยเฉลี่ย	99.22%	99.12%	**99.06%**	98.94%	98.94%
จำนวนโหนดโดยเฉลี่ย	92	96.51	**111**	117	117

จากตารางข้างบนค่าพารามิเตอร์ N และ D ที่ดีที่สุดคือ D = 0.75 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.75 และ N = 3, ค่าพารามิเตอร์ D = 0.85 และ N = 3 และค่าพารามิเตอร์ D = 0.95 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.75 และ N = 3 มีจำนวนโหนดโดยเฉลี่ยน้อยที่สุดจากค่าพารามิเตอร์ทั้ง 3 แบบที่มีความลึกสูงสุดโดยเฉลี่ยเท่ากัน ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D ที่ดีที่สุดของชุดข้อมูล Chess

6.2.3.3 การหาค่าพารามิเตอร์ N และ D ที่ดีที่สุดของชุดข้อมูล Phishing Websites

ตารางที่ 6.15 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 1 ที่มีค่าพารามิเตอร์ D = 0.25 บนชุดข้อมูล Phishing Websites

เกณฑ์ในการวัด	**D = 0.25**		
	N = 2	**N = 3**	**N = 4**
ความลึกสูงสุดโดยเฉลี่ย	20	18	**17**
ความแม่นยำโดยเฉลี่ย	95.51%	95.51%	**95.5%**
จำนวนโหนดโดยเฉลี่ย	2,255.12	2,937.03	**4,112.88**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 1 คือ D = 0.25 และ N = 4 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.25 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม

ตารางที่ 6.16 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 2 ที่มีค่าพารามิเตอร์ D = 0.5 บนชุดข้อมูล Phishing Websites

เกณฑ์ในการวัด	**D = 0.5**		
	N = 2	**N = 3**	**N = 4**
ความลึกสูงสุดโดยเฉลี่ย	18	16	**15**
ความแม่นยำโดยเฉลี่ย	95.5%	95.43%	**95.43%**
จำนวนโหนดโดยเฉลี่ย	2,278.95	3,004.08	**4,149.63**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 2 คือ D = 0.5 และ N = 4 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.5 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม

ตารางที่ 6.17 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 3 ที่มีค่าพารามิเตอร์ D = 0.75 บนชุดข้อมูล Phishing Websites

เกณฑ์ในการวัด	**D = 0.75**		
	N = 2	**N = 3**	**N = 4**
ความลึกสูงสุดโดยเฉลี่ย	16	14	**12**
ความแม่นยำโดยเฉลี่ย	95.18%	95.09%	**95.05%**
จำนวนโหนดโดยเฉลี่ย	2,320.31	3,113.77	**4,399.06**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 3 คือ D = 0.75 และ N = 4 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.75 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม

ตารางที่ 6.18 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 4 ที่มีค่าพารามิเตอร์ D = 0.85 บนชุดข้อมูล Phishing Websites

เกณฑ์ในการวัด	**D = 0.85**		
	N = 2	**N = 3**	**N = 4**
ความลึกสูงสุดโดยเฉลี่ย	16	14	**12**
ความแม่นยำโดยเฉลี่ย	95.18%	95.09%	**95.05%**
จำนวนโหนดโดยเฉลี่ย	2,341.9	3,145.43	**4,453.57**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 4 คือ D = 0.85 และ N = 4 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.85 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม

ตารางที่ 6.19 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 5 ที่มีค่าพารามิเตอร์ D = 0.95 บนชุดข้อมูล Phishing Websites

เกณฑ์ในการวัด	**D = 0.95**		
	N = 2	**N = 3**	**N = 4**
ความลึกสูงสุดโดยเฉลี่ย	16	14	**12**
ความแม่นยำโดยเฉลี่ย	95.23%	95.14%	**95.1%**
จำนวนโหนดโดยเฉลี่ย	2,343.88	3,148	**4,451.74**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 5 คือ D = 0.95 และ N = 4 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.95 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม เมื่อได้ค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละกลุ่มแล้ว ต่อมาให้ทำการเลือกค่าพารามิเตอร์ N และ D ที่ดีที่สุดโดยการเปรียบเทียบระหว่างค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละกลุ่ม

ตารางที่ 6.20 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอของค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละกลุ่ม บนชุดข้อมูล Phishing Websites

เกณฑ์ในการวัด	**กลุ่มที่ 1**	**กลุ่มที่ 2**	**กลุ่มที่ 3**	**กลุ่มที่ 4**	**กลุ่มที่ 5**
	D=0.25, N=4	**D=0.5, N=4**	**D=0.75, N=4**	**D=0.85, N=4**	**D=0.95, N=4**
ความลึกสูงสุดโดยเฉลี่ย	17	15	**12**	12	12
ความแม่นยำโดยเฉลี่ย	95.5%	95.43%	**95.05%**	95.05%	95.1%
จำนวนโหนดโดยเฉลี่ย	4,112.88	4,149.63	**4,399.06**	4,453.57	4,451.74

จากตารางข้างบนค่าพารามิเตอร์ N และ D ที่ดีที่สุดคือ D = 0.75 และ N = 4 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.75 และ N = 4, ค่าพารามิเตอร์ D = 0.85 และ N = 4 และค่าพารามิเตอร์ D = 0.95 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.75 และ N = 4 มีจำนวนโหนดโดยเฉลี่ยน้อยที่สุดจากค่าพารามิเตอร์ทั้ง 3 แบบที่มีความลึกสูงสุดโดยเฉลี่ยเท่ากัน ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D ที่ดีที่สุดของชุดข้อมูล Phishing Websites

6.2.3.4 การหาค่าพารามิเตอร์ N และ D ที่ดีที่สุดของชุดข้อมูล Insurance Company Benchmark

ตารางที่ 6.21 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 1 ที่มีค่าพารามิเตอร์ D = 0.25 บนชุดข้อมูล Insurance Company Benchmark

เกณฑ์ในการวัด	D = 0.25		
	N = 2	N = 3	N = 4
ความลึกสูงสุดโดยเฉลี่ย	43	30	**24**
ความแม่นยำโดยเฉลี่ย	91.75%	92.04%	**92.08%**
จำนวนโหนดโดยเฉลี่ย	83,745.83	409,504.92	**2,320,341.78**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 1 คือ D = 0.25 และ N = 4 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.25 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม

ตารางที่ 6.22 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 2 ที่มีค่าพารามิเตอร์ D = 0.5 บนชุดข้อมูล Insurance Company Benchmark

เกณฑ์ในการวัด	D = 0.5		
	N = 2	N = 3	N = 4
ความลึกสูงสุดโดยเฉลี่ย	43	30	**24**
ความแม่นยำโดยเฉลี่ย	91.84%	92.11%	**92.16%**
จำนวนโหนดโดยเฉลี่ย	84,200.42	416,785.09	**2,392,720.09**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 2 คือ D = 0.5 และ N = 4 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.5 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม

ตารางที่ 6.23 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 3 ที่มีค่าพารามิเตอร์ D = 0.75 บนชุดข้อมูล Insurance Company Benchmark

เกณฑ์ในการวัด	**D = 0.75**		
	N = 2	**N = 3**	**N = 4**
ความลึกสูงสุดโดยเฉลี่ย	43	30	**23**
ความแม่นยำโดยเฉลี่ย	91.92%	92.25%	**92.34%**
จำนวนโหนดโดยเฉลี่ย	84,588.51	423,023.62	**2,449,687.43**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 3 คือ D = 0.75 และ N = 4 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.75 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม

ตารางที่ 6.24 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 4 ที่มีค่าพารามิเตอร์ D = 0.85 บนชุดข้อมูล Insurance Company Benchmark

เกณฑ์ในการวัด	**D = 0.85**		
	N = 2	**N = 3**	**N = 4**
ความลึกสูงสุดโดยเฉลี่ย	43	30	**23**
ความแม่นยำโดยเฉลี่ย	91.92%	92.25%	**92.34%**
จำนวนโหนดโดยเฉลี่ย	84,598.68	422,763.84	**2,445,035.14**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 4 คือ D = 0.85 และ N = 4 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.85 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม

ตารางที่ 6.25 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 5 ที่มีค่าพารามิเตอร์ D = 0.95 บนชุดข้อมูล Insurance Company Benchmark

เกณฑ์ในการวัด	D = 0.95		
	N = 2	N = 3	N = 4
ความลึกสูงสุดโดยเฉลี่ย	43	30	**23**
ความแม่นยำโดยเฉลี่ย	91.92%	92.25%	**92.34%**
จำนวนโหนดโดยเฉลี่ย	84,613.76	422,367.37	**2,443,915.74**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 5 คือ D = 0.95 และ N = 4 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.95 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม เมื่อได้ค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละกลุ่มแล้ว ต่อมาให้ทำการเลือกค่าพารามิเตอร์ N และ D ที่ดีที่สุดโดยการเปรียบเทียบระหว่างค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละกลุ่ม

ตารางที่ 6.26 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอของค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละกลุ่ม บนชุดข้อมูล Insurance Company Benchmark

เกณฑ์ในการวัด	กลุ่มที่ 1	กลุ่มที่ 2	กลุ่มที่ 3	กลุ่มที่ 4	กลุ่มที่ 5
	D=0.25, N=4	D=0.5, N=4	D=0.75, N=4	D=0.85, N=4	D=0.95, N=4
ความลึกสูงสุดโดยเฉลี่ย	24	24	23	23	**23**
ความแม่นยำโดยเฉลี่ย	92.08%	92.16%	92.34%	92.34%	**92.34%**
จำนวนโหนดโดยเฉลี่ย	2,320,341.78	2,392,720.09	2,449,687.43	2,445,035.14	**2,443,915.74**

จากตารางข้างบนค่าพารามิเตอร์ N และ D ที่ดีที่สุดคือ D = 0.95 และ N = 4 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.75 และ N = 4, ค่าพารามิเตอร์ D = 0.85 และ N = 4 และค่าพารามิเตอร์ D = 0.95 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.95 และ N = 4 มีจำนวนโหนดโดยเฉลี่ยน้อยที่สุดจากค่าพารามิเตอร์ทั้ง 3 แบบที่มีความลึกสูงสุดโดยเฉลี่ยเท่ากัน ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D ที่ดีที่สุดของชุดข้อมูล Insurance Company Benchmark

6.2.3.5 การหาค่าพารามิเตอร์ N และ D ที่ดีที่สุดของชุดข้อมูล Firm-Teacher

ตารางที่ 6.27 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 1 ที่มีค่าพารามิเตอร์ D = 0.25 บนชุดข้อมูล Firm-Teacher

เกณฑ์ในการวัด	**D = 0.25**		
	N = 2	**N = 3**	**N = 4**
ความลึกสูงสุดโดยเฉลี่ย	**10**	10	10
ความแม่นยำโดยเฉลี่ย	**73.39%**	73.71%	73.52%
จำนวนโหนดโดยเฉลี่ย	**3,508.37**	4,216.93	4,859.91

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 1 คือ D = 0.25 และ N = 2 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าความลึกสูงสุดโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.25 และ N = 2 มีจำนวนโหนดโดยเฉลี่ยน้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม

ตารางที่ 6.28 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 2 ที่มีค่าพารามิเตอร์ D = 0.5 บนชุดข้อมูล Firm-Teacher

เกณฑ์ในการวัด	**D = 0.5**		
	N = 2	**N = 3**	**N = 4**
ความลึกสูงสุดโดยเฉลี่ย	9.61	**9**	9
ความแม่นยำโดยเฉลี่ย	73.36%	**73.73%**	73.89%
จำนวนโหนดโดยเฉลี่ย	3,387.7	**4,161.43**	5,045.52

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 2 คือ D = 0.5 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.5 และ N = 3 และค่าพารามิเตอร์ D = 0.5 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่า

ค่าพารามิเตอร์ D = 0.5 และ N = 3 มีจำนวนโหนดโดยเฉลี่ยน้อยกว่า ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม

ตารางที่ 6.29 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 3 ที่มีค่าพารามิเตอร์ D = 0.75 บนชุดข้อมูล Firm-Teacher

เกณฑ์ในการวัด	**D = 0.75**		
	N = 2	**N = 3**	**N = 4**
ความลึกสูงสุดโดยเฉลี่ย	8.56	8.23	**8**
ความแม่นยำโดยเฉลี่ย	73.56%	73.83%	**73.98%**
จำนวนโหนดโดยเฉลี่ย	3,581.59	4,718.72	**6,252.35**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 3 คือ D = 0.75 และ N = 4 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.75 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม

ตารางที่ 6.30 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 4 ที่มีค่าพารามิเตอร์ D = 0.85 บนชุดข้อมูล Firm-Teacher

เกณฑ์ในการวัด	**D = 0.85**		
	N = 2	**N = 3**	**N = 4**
ความลึกสูงสุดโดยเฉลี่ย	8.32	**8**	8
ความแม่นยำโดยเฉลี่ย	73.56%	**73.98%**	74.11%
จำนวนโหนดโดยเฉลี่ย	3,602.94	**4,801.48**	6,420.08

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 4 คือ D = 0.85 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.85 และ N = 3 และค่าพารามิเตอร์ D = 0.85 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่า

ค่าพารามิเตอร์ D = 0.85 และ N = 3 มีจำนวนโหนดโดยเฉลี่ยน้อยกว่า ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม

ตารางที่ 6.31 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 5 ที่มีค่าพารามิเตอร์ D = 0.95 บนชุดข้อมูล Firm-Teacher

เกณฑ์ในการวัด	**D = 0.95**		
	N = 2	**N = 3**	**N = 4**
ความลึกสูงสุดโดยเฉลี่ย	**8**	8	8
ความแม่นยำโดยเฉลี่ย	**73.54%**	73.98%	74.11%
จำนวนโหนดโดยเฉลี่ย	**3,611.24**	4,812.36	6,465.25

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 5 คือ D = 0.95 และ N = 2 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าความลึกสูงสุดโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.95 และ N = 2 มีจำนวนโหนดโดยเฉลี่ยน้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม เมื่อได้ค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละกลุ่มแล้ว ต่อมาให้ทำการเลือกค่าพารามิเตอร์ N และ D ที่ดีที่สุดโดยการเปรียบเทียบระหว่างค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละกลุ่ม

ตารางที่ 6.32 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอของค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละกลุ่ม บนชุดข้อมูล Firm-Teacher

เกณฑ์ในการวัด	**กลุ่มที่ 1**	**กลุ่มที่ 2**	**กลุ่มที่ 3**	**กลุ่มที่ 4**	**กลุ่มที่ 5**
	D=0.25, N=2	**D=0.5, N=3**	**D=0.75, N=4**	**D=0.85, N=3**	**D=0.95, N=2**
ความลึกสูงสุดโดยเฉลี่ย	10	9	8	8	**8**
ความแม่นยำโดยเฉลี่ย	73.39%	73.73%	73.98%	73.98%	**73.54%**
จำนวนโหนดโดยเฉลี่ย	3,508.37	4,161.43	6,252.35	4,801.48	**3,611.24**

จากตารางข้างบนค่าพารามิเตอร์ N และ D ที่ดีที่สุดคือ D = 0.95 และ N = 2 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.75 และ N = 4, ค่าพารามิเตอร์ D = 0.85 และ N = 3 และค่าพารามิเตอร์ D = 0.95 และ N = 2 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวน

โหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.95 และ N = 2 มีจำนวนโหนดโดยเฉลี่ยน้อยที่สุดจากค่าพารามิเตอร์ทั้ง 3 แบบที่มีความลึกสูงสุดโดยเฉลี่ยเท่ากัน ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D ที่ดีที่สุดของชุดข้อมูล Firm-Teacher

6.2.3.6 การหาค่าพารามิเตอร์ N และ D ที่ดีที่สุดของชุดข้อมูล Bach Choral Harmony

ตารางที่ 6.33 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 1 ที่มีค่าพารามิเตอร์ D = 0.25 บนชุดข้อมูล Bach Choral Harmony

เกณฑ์ในการวัด	**D = 0.25**		
	N = 2	**N = 3**	**N = 4**
ความลึกสูงสุดโดยเฉลี่ย	11	**10**	10
ความแม่นยำโดยเฉลี่ย	72.31%	**72.3%**	72.3%
จำนวนโหนดโดยเฉลี่ย	4,780.61	**5,814.07**	7,486.44

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 1 คือ D = 0.25 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.25 และ N = 3 และค่าพารามิเตอร์ D = 0.25 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.25 และ N = 3 มีจำนวนโหนดโดยเฉลี่ยน้อยกว่า ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม

ตารางที่ 6.34 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 2 ที่มีค่าพารามิเตอร์ D = 0.5 บนชุดข้อมูล Bach Choral Harmony

เกณฑ์ในการวัด	**D = 0.5**		
	N = 2	**N = 3**	**N = 4**
ความลึกสูงสุดโดยเฉลี่ย	9	9	**8**
ความแม่นยำโดยเฉลี่ย	72.29%	72.29%	**72.27%**
จำนวนโหนดโดยเฉลี่ย	4,722.52	5,737.08	**7,380.84**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 2 คือ D = 0.5 และ N = 4 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.5 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม

ตารางที่ 6.35 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 3 ที่มีค่าพารามิเตอร์ D = 0.75 บนชุดข้อมูล Bach Choral Harmony

เกณฑ์ในการวัด	**D = 0.75**		
	N = 2	**N = 3**	**N = 4**
ความลึกสูงสุดโดยเฉลี่ย	8	8	**7**
ความแม่นยำโดยเฉลี่ย	72.17%	72.1%	**72.18%**
จำนวนโหนดโดยเฉลี่ย	4,682.07	5,711.16	**7,333.89**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 3 คือ D = 0.75 และ N = 4 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.75 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม

ตารางที่ 6.36 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 4 ที่มีค่าพารามิเตอร์ D = 0.85 บนชุดข้อมูล Bach Choral Harmony

เกณฑ์ในการวัด	**D = 0.85**		
	N = 2	**N = 3**	**N = 4**
ความลึกสูงสุดโดยเฉลี่ย	8	8	**7**
ความแม่นยำโดยเฉลี่ย	72.17%	72.1%	**72.18%**
จำนวนโหนดโดยเฉลี่ย	4,692.79	5,713.71	**7,336.34**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 4 คือ D = 0.85 และ N = 4 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึก

สูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.85 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม

ตารางที่ 6.37 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N ของกลุ่มที่ 5 ที่มีค่าพารามิเตอร์ D = 0.95 บนชุดข้อมูล Bach Choral Harmony

เกณฑ์ในการวัด	**D = 0.95**		
	N = 2	**N = 3**	**N = 4**
ความลึกสูงสุดโดยเฉลี่ย	8	8	**7**
ความแม่นยำโดยเฉลี่ย	72.17%	72.1%	**72.18%**
จำนวนโหนดโดยเฉลี่ย	4,694.34	5,710.21	**7,335.17**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่มที่ 5 คือ D = 0.95 และ N = 4 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.95 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในกลุ่ม เมื่อได้ค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละกลุ่มแล้ว ต่อมาให้ทำการเลือกค่าพารามิเตอร์ N และ D ที่ดีที่สุดโดยการเปรียบเทียบระหว่างค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละกลุ่ม

ตารางที่ 6.38 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอของค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละกลุ่ม บนชุดข้อมูล Bach Choral Harmony

เกณฑ์ในการวัด	**กลุ่มที่ 1**	**กลุ่มที่ 2**	**กลุ่มที่ 3**	**กลุ่มที่ 4**	**กลุ่มที่ 5**
	D=0.25, N=3	**D=0.5, N=4**	**D=0.75, N=4**	**D=0.85, N=4**	**D=0.95, N=4**
ความลึกสูงสุดโดยเฉลี่ย	10	8	**7**	7	7
ความแม่นยำโดยเฉลี่ย	72.3%	72.27%	**72.18%**	72.18%	72.18%
จำนวนโหนดโดยเฉลี่ย	5,814.07	7,380.84	**7,333.89**	7,336.34	7,335.17

จากตารางข้างบนค่าพารามิเตอร์ N และ D ที่ดีที่สุดคือ D = 0.75 และ N = 4 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.75 และ N = 4, ค่าพารามิเตอร์ D = 0.85 และ N = 4 และค่าพารามิเตอร์ D = 0.95 และ N = 4 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวน

โหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.75 และ N = 4 มีจำนวนโหนดโดยเฉลี่ยน้อยที่สุดจากค่าพารามิเตอร์ทั้ง 3 แบบที่มีความลึกสูงสุดโดยเฉลี่ยเท่ากัน ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D ที่ดีที่สุดของชุดข้อมูล Bach Choral Harmony

6.2.3.7 สรุปค่าพารามิเตอร์ที่ดีที่สุดสำหรับการทดลองของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอของชุดข้อมูลทั้งหมด

ตารางที่ 6.39 ค่าพารามิเตอร์ที่ดีที่สุดสำหรับการทดลองของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอของชุดข้อมูลทั้งหมด

ชุดข้อมูล	พารามิเตอร์	
	N	D
Chess	3	0.75
Connect-4	4	0.95
Firm-Teacher	2	0.95
Phishing Websites	4	0.75
Insurance Company Benchmark	4	0.95
Bach Choral Harmony	4	0.75

จากตารางข้างบน ค่า N ในแต่ละชุดข้อมูลในตารางข้างบนนี้จะนำไปใช้เป็นค่าพารามิเตอร์ N ในการทดลองของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกด้วย

6.3 ผลการทดลองระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ

6.3.1 ความแม่นยำโดยเฉลี่ย

ตารางที่ 6.40 ความแม่นยำโดยเฉลี่ยระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ

ชุดข้อมูล	ความแม่นยำโดยเฉลี่ย	
	ID3 แบบดั้งเดิม	ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ
Chess	99.14%	99.06%
Connect-4	72.95%	72.87%
Firm-Teacher	72.26%	73.54%
Phishing Websites	95.27%	95.05%
Insurance Company Benchmark	90.34%	92.34%
Bach Choral Harmony	71.67%	72.18%

จากตารางข้างบน ความแม่นยำโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าน้อยกว่าความแม่นยำโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมเพียงเล็กน้อยในชุดข้อมูล Chess, Connect-4 และ Phishing Websites ในขณะที่ชุดข้อมูลที่เหลือคือ Firm-Teacher, Insurance Company Benchmark และ Bach Choral Harmony นั้นมีความแม่นยำโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมากกว่าความแม่นยำโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมเพียงเล็กน้อยเช่นกัน ซึ่งสามารถสรุปได้ว่า ความแม่นยำโดยเฉลี่ยของอัลกอริทึม ID3 ทั้งสองแบบนี้มีค่าที่ใกล้เคียงกัน

6.3.2 ความลึกสูงสุดโดยเฉลี่ย

ตารางที่ 6.41 ความลึกสูงสุดโดยเฉลี่ยระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ

ชุดข้อมูล	ความลึกสูงสุดโดยเฉลี่ย	
	ID3 แบบดั้งเดิม	ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ
Chess	12	6
Connect-4	24.12	10.14
Firm-Teacher	15	8
Phishing Websites	30	12
Insurance Company Benchmark	85	23
Bach Choral Harmony	14	7

จากตารางข้างบน ความลึกสูงสุดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าน้อยกว่าความลึกสูงสุดโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูลทั้งหมดอย่างเห็นได้ชัด ซึ่งความลึกสูงสุดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าที่ลดลงมากกว่า 45% จากความลึกสูงสุดโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูลทั้งหมด ซึ่งสามารถสรุปได้ว่าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอสามารถลดความลึกสูงสุดโดยเฉลี่ยจากความลึกสูงสุดโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูลทั้งหมดได้อย่างเห็นได้ชัด

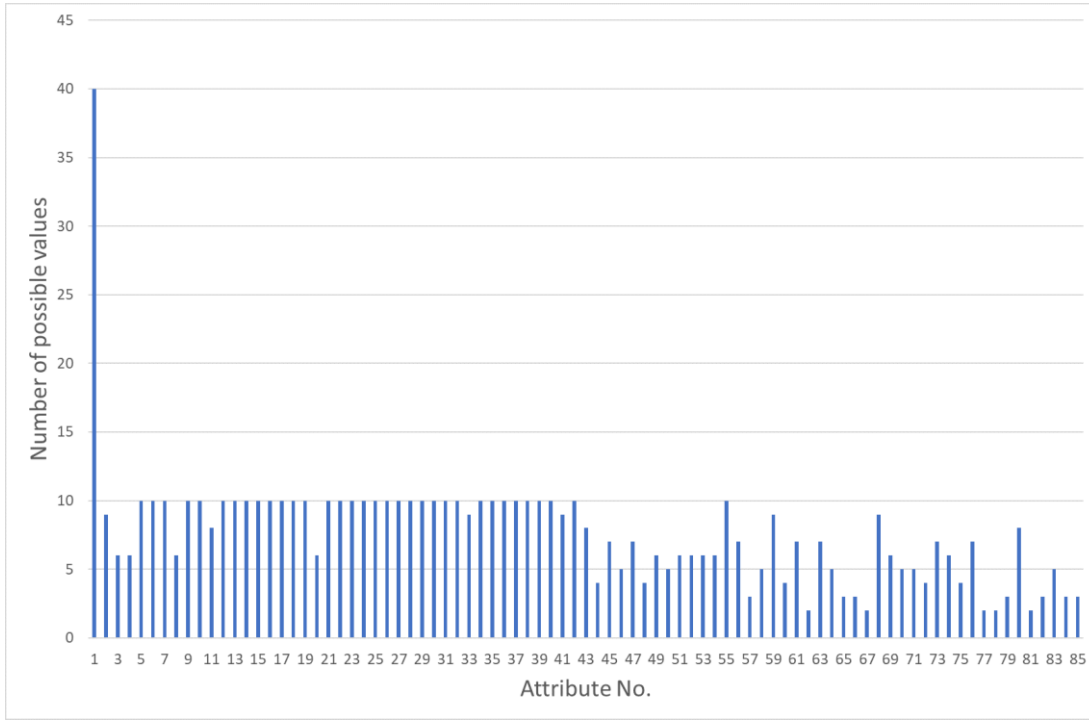
6.3.3 จำนวนโหนดโดยเฉลี่ย

ตารางที่ 6.42 จำนวนโหนดโดยเฉลี่ยระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ

ชุดข้อมูล	จำนวนโหนดโดยเฉลี่ย	
	ID3 แบบดั้งเดิม	ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ
Chess	69	111
Connect-4	22,560.1	130,123.28
Firm-Teacher	3,126	3,611.24
Phishing Websites	2,095.67	4,399.06
Insurance Company Benchmark	21,850.54	2,443,915.74
Bach Choral Harmony	4,911.6	7,333.89

จากตารางข้างบน จำนวนโหนดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่ามากกว่าจำนวนโหนดโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูลทั้งหมดอย่างเห็นได้ชัด ซึ่งจำนวนโหนดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าที่เพิ่มขึ้นมากกว่า 15% จากจำนวนโหนดโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูล Firm-Teacher และมีค่าที่เพิ่มขึ้นมากกว่า 45% จากจำนวนโหนดโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูลที่เหลือทั้งหมด

สำหรับจำนวนโหนดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในชุดข้อมูล Insurance Company Benchmark ซึ่งมีค่าที่เพิ่มขึ้นมากหลายเท่าตัวจากจำนวนโหนดโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูล Insurance Company Benchmark นั้น เพราะเนื่องจากว่าชุดข้อมูล Insurance Company Benchmark มีจำนวนค่าที่เป็นไปได้ของแอตทริบิวต์แต่ละแอตทริบิวต์เป็นไปตามที่แสดงในรูปถัดไป



รูปที่ 6.1 จำนวนค่าที่เป็นไปได้ของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในชุดข้อมูล Insurance Company Benchmark

จากรูปข้างบน จะเห็นว่าแอตทริบิวต์ที่ 1 มีค่าที่เป็นไปได้มากถึง 40 ค่า ในขณะที่แอตทริบิวต์อื่นๆ ส่วนมากมีค่าที่เป็นไปได้ถึง 10 ค่า ถ้าแอตทริบิวต์ที่ 1 และแอตทริบิวต์อื่นๆ ถูกอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอเลือกไปเป็นแอตทริบิวต์บนโหนดตัดสินใจเดียวกัน ก็จะทำให้มีการแตกกิ่งออกมาจากโหนดตัดสินใจเป็นจำนวนมาก และทำให้มีการสร้างโหนดในระดับลูกหลานเป็นจำนวนที่มากขึ้นหลายเท่าตัว ซึ่งในท้ายที่สุดทำให้จำนวนโหนดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในชุดข้อมูล Insurance Company Benchmark มีจำนวนที่มากขึ้นหลายเท่าตัว เมื่อเทียบกับจำนวนโหนดโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูล Insurance Company Benchmark

จากผลการทดลองจำนวนโหนดโดยเฉลี่ยระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ สามารถสรุปได้ว่า จำนวนโหนดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าที่เพิ่มขึ้นจากจำนวนโหนดโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมอย่างชัดเจนในชุดข้อมูลทั้งหมด

6.3.4 เวลาในการฝึกฝนโดยเฉลี่ย

ตารางที่ 6.43 เวลาในการฝึกฝนโดยเฉลี่ยระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ

ชุดข้อมูล	เวลาในการฝึกฝนโดยเฉลี่ย (มิลลิวินาที)	
	ID3 แบบดั้งเดิม	ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ
Chess	15.28	10.19
Connect-4	1,481.8	1,774.06
Firm-Teacher	97.26	63.27
Phishing Websites	63.49	174.55
Insurance Company Benchmark	395.99	127,870.45
Bach Choral Harmony	439.93	353.17

จากตารางข้างบน เวลาในการฝึกฝนโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่ามากกว่าเวลาในการฝึกฝนโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูล Connect-4, Phishing Websites และ Insurance Company Benchmark ในทางกลับกัน เวลาในการฝึกฝนโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าน้อยกว่าเวลาในการฝึกฝนโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูล Chess, Firm-Teacher และ Bach Choral Harmony สาเหตุที่ผลการทดลองด้านเวลาในการฝึกฝนโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอเมื่อเทียบกับเวลาในการฝึกฝนโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมเป็นตามที่ได้กล่าวมา เพราะว่าเวลาในการฝึกฝนโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ จะขึ้นอยู่กับขนาดและความซับซ้อนของชุดข้อมูลที่ใช้ในการทดลอง และยังขึ้นอยู่กับจำนวนค่าที่เป็นไปได้ของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในชุดข้อมูลที่ใช้ในการทดลอง

เมื่อพิจารณาเวลาในการฝึกฝนโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในชุดข้อมูล Connect-4, Phishing Websites และ Insurance Company Benchmark จะพบว่า เวลาในการฝึกฝนโดยเฉลี่ยเพิ่มขึ้นจากเวลาในการฝึกฝนโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมตามที่ได้กล่าวไว้ เนื่องจากชุดข้อมูลทั้ง 3 มีจำนวนตัวอย่างและจำนวนแอตทริบิวต์เป็นจำนวนมากและมีความซับซ้อนมากกว่าชุดข้อมูลอีก 3 ชุดข้อมูลที่เหลือ อีกทั้งยังมีจำนวนค่าที่เป็นไปได้ของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในชุดข้อมูล Connect-4, Phishing Websites และ Insurance Company Benchmark เป็นจำนวนที่มาก จึงส่งผลทำให้เวลาในการฝึกฝนโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในชุดข้อมูล Connect-4, Phishing Websites และ

Insurance Company Benchmark เพิ่มมากขึ้นจากเวลาในการฝึกฝนโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิม

ในทางตรงกันข้าม เมื่อพิจารณาเวลาในการฝึกฝนโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในชุดข้อมูล Chess, Firm-Teacher และ Bach Choral Harmony จะพบว่า เวลาในการฝึกฝนโดยเฉลี่ยลดลงจากเวลาในการฝึกฝนโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมตามที่ได้กล่าวไว้ เนื่องจากชุดข้อมูลทั้ง 3 นี้มีจำนวนตัวอย่างหรือจำนวนแอตทริบิวต์ที่น้อยหรือจำนวนทั้งสองมีน้อยทั้งคู่ อีกทั้งยังมีจำนวนค่าที่เป็นไปได้ของแอตทริบิวต์แต่ละแอตทริบิวต์ที่อยู่ในชุดข้อมูล Chess, Firm-Teacher และ Bach Choral Harmony เป็นจำนวนที่น้อย และจากผลการทดลองด้านความลึกสูงสุดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในชุดข้อมูลทั้ง 3 นี้ซึ่งมีความลึกที่ลดลงมากกว่า 45% จากความลึกสูงสุดโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิม เพราะฉะนั้นเวลาในการฝึกฝนโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในชุดข้อมูล Chess, Firm-Teacher และ Bach Choral Harmony จึงลดลงจากเวลาในการฝึกฝนโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิม

6.3.5 เวลาในการทดสอบโดยเฉลี่ย

ตารางที่ 6.44 เวลาในการทดสอบโดยเฉลี่ยระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ

ชุดข้อมูล	เวลาในการทดสอบโดยเฉลี่ย (มิลลิวินาที)	
	ID3 แบบดั้งเดิม	ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ
Chess	0.0075	0.0050
Connect-4	0.3395	0.1855
Firm-Teacher	0.0554	0.0298
Phishing Websites	0.0333	0.0180
Insurance Company Benchmark	0.0201	0.0113
Bach Choral Harmony	0.0279	0.0142

จากตารางข้างบน เวลาในการทดสอบโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าน้อยกว่าเวลาในการทดสอบโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูลทั้งหมดอย่างเห็นได้ชัด ซึ่งเวลาในการทดสอบโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าที่ลดลงมากกว่า 30% จากเวลาในการทดสอบโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูลทั้งหมด ซึ่งสามารถสรุปได้ว่าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่

เสนอสามารถลดเวลาในการทดสอบโดยเฉลี่ยจากเวลาในการทดสอบโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูลทั้งหมดได้อย่างเห็นได้ชัด

การที่เวลาในการทดสอบโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าที่น้อยกว่าเวลาในการทดสอบโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูลทั้งหมดอย่างเห็นได้ชัด สาเหตุนั้นไม่ได้มาจากความลึกสูงสุดโดยเฉลี่ยของต้นไม้ตัดสินใจที่สร้างจากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอโดยตรง แต่มาจากความลึกโดยเฉลี่ยในการจำแนกตัวอย่างในกระบวนการทดสอบต้นไม้ตัดสินใจที่สร้าง การที่ความลึกสูงสุดโดยเฉลี่ยของต้นไม้ตัดสินใจที่สร้างจากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมีค่าที่ลดลง ย่อมมีโอกาสทำให้ความลึกโดยเฉลี่ยในการจำแนกตัวอย่างในกระบวนการทดสอบต้นไม้ตัดสินใจที่สร้างมีค่าที่ลดลงตามไปด้วย ซึ่งตัวอย่างที่อยู่ในชุดข้อมูลทดสอบของการทดลองนี้นั้นส่วนใหญ่ถูกจำแนกในความลึกที่น้อยกว่าความลึกสูงสุดโดยเฉลี่ยของต้นไม้ตัดสินใจที่สร้างจากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ ดังนั้นจึงเป็นประเด็นที่น่าสนใจที่จะเปรียบเทียบร้อยละการลดลงของเวลาในการทดสอบโดยเฉลี่ยและความลึกโดยเฉลี่ยในการจำแนกของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอกับอัลกอริทึม ID3 แบบดั้งเดิม ตารางถัดไปแสดงความลึกโดยเฉลี่ยในการจำแนกระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ

ตารางที่ 6.45 ความลึกโดยเฉลี่ยในการจำแนกระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ

ชุดข้อมูล	ความลึกโดยเฉลี่ยในการจำแนก	
	ID3 แบบดั้งเดิม	ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ
Chess	4.33	2.7
Connect-4	9.43	5.17
Firm-Teacher	9.6	5.21
Phishing Websites	5.61	3.04
Insurance Company Benchmark	4.17	2.21
Bach Choral Harmony	9.28	4.7

จากตารางข้างบน จะเห็นว่าความลึกโดยเฉลี่ยในการจำแนกของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าน้อยกว่าความลึกโดยเฉลี่ยในการจำแนกของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูลทั้งหมดอย่างเห็นได้ชัด ตารางถัดไปแสดงร้อยละการลดลงของเวลาในการทดสอบโดยเฉลี่ยและความลึกโดยเฉลี่ยในการจำแนกของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เปรียบเทียบกับอัลกอริทึม ID3 แบบดั้งเดิม

ตารางที่ 6.46 ร้อยละการลดลงของเวลาในการทดสอบโดยเฉลี่ยและความลึกโดยเฉลี่ยในการจำแนกของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เปรียบเทียบกับอัลกอริทึม ID3 แบบดั้งเดิม

ชุดข้อมูล	ร้อยละการลดลงของเวลาในการทดสอบโดยเฉลี่ย	ร้อยละการลดลงของความลึกโดยเฉลี่ยในการจำแนก
Chess	33.59%	37.66%
Connect-4	45.29%	45.17%
Firm-Teacher	46.21%	45.73%
Phishing Websites	45.95%	45.81%
Insurance Company Benchmark	43.78%	47.00%
Bach Choral Harmony	49.10%	49.35%

จากตารางข้างบน จะเห็นว่าร้อยละการลดลงของเวลาในการทดสอบโดยเฉลี่ยมีค่าที่ใกล้เคียงกับร้อยละการลดลงของความลึกโดยเฉลี่ยในการจำแนก ดังนั้นจึงเห็นได้ชัดเจนว่า การลดลงของความลึกโดยเฉลี่ยในการจำแนกส่งผลต่อการลดลงของเวลาในการทดสอบโดยเฉลี่ย และนำไปสู่ผลลัพธ์ที่ได้ของเวลาในการทดสอบโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ ซึ่งมีค่าที่น้อยกว่าเวลาในการทดสอบโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูลทั้งหมด

6.4 ผลการทดลองระหว่างอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ

6.4.1 ความแม่นยำโดยเฉลี่ย

ตารางที่ 6.47 ความแม่นยำโดยเฉลี่ยระหว่างอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ

ชุดข้อมูล	ความแม่นยำโดยเฉลี่ย	
	ID3 ที่ปรับปรุงของงานวิจัยแรก	ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ
Chess	99.22%	99.06%
Connect-4	73.68%	72.87%
Firm-Teacher	72.55%	73.54%
Phishing Websites	95.35%	95.05%
Insurance Company Benchmark	91.33%	92.34%
Bach Choral Harmony	71.6%	72.18%

จากตารางข้างบน ความแม่นยำโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าน้อยกว่าความแม่นยำโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกเพียงเล็กน้อยในชุดข้อมูล Chess, Connect-4, Phishing Websites ในขณะที่ชุดข้อมูลที่เหลือคือ Firm-Teacher, Insurance Company Benchmark และ Bach Choral Harmony นั้นมีความแม่นยำโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมากกว่าความแม่นยำโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกเพียงเล็กน้อยเช่นกัน ซึ่งสามารถสรุปได้ว่า ความแม่นยำโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงทั้งสองแบบนี้มีค่าที่ใกล้เคียงกัน

6.4.2 ความลึกสูงสุดโดยเฉลี่ย

ตารางที่ 6.48 ความลึกสูงสุดโดยเฉลี่ยระหว่างอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ

ชุดข้อมูล	ความลึกสูงสุดโดยเฉลี่ย	
	ID3 ที่ปรับปรุงของงานวิจัยแรก	ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ
Chess	11	6
Connect-4	18	10.14
Firm-Teacher	14	8
Phishing Websites	21	12
Insurance Company Benchmark	26	23
Bach Choral Harmony	11	7

จากตารางข้างบน ความลึกสูงสุดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าน้อยกว่าความลึกสูงสุดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในชุดข้อมูลทั้งหมดอย่างเห็นได้ชัด ซึ่งความลึกสูงสุดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าที่ลดลงมากกว่า 35% จากความลึกสูงสุดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในชุดข้อมูลทั้งหมดยกเว้นชุดข้อมูล Insurance Company Benchmark เท่านั้นที่ความลึกสูงสุดโดยเฉลี่ยลดลง 11.54% แต่ก็เป็นการลดลงที่เห็นได้ชัดเหมือนเดิม ซึ่งสามารถสรุปได้ว่าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอสามารถลดความลึกสูงสุดโดยเฉลี่ยจากความลึกสูงสุดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในชุดข้อมูลทั้งหมดได้อย่างเห็นได้ชัด

การที่ความลึกสูงสุดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมีค่าที่น้อยกว่าความลึกสูงสุดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในชุดข้อมูลทั้งหมดอย่างเห็นได้ชัดนั้น สาเหตุมาจากจำนวนแอตทริบิวต์ที่ถูกเลือกไปอยู่ในโหนดตัดสินใจเดียวกันต่อโหนดโดยเฉลี่ย ซึ่งมีผลต่อความลึกสูงสุดโดยเฉลี่ยของต้นไม้ตัดสินใจที่สร้าง จำนวนแอตทริบิวต์ที่ถูกเลือกไปอยู่ในโหนดตัดสินใจเดียวกันต่อโหนดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงทั้งสองแบบแสดงดังตารางถัดไป

ตารางที่ 6.49 จำนวนแอตทริบิวต์ที่ถูกเลือกไปอยู่ในโหนดตัดสินใจเดียวกันต่อโหนดโดยเฉลี่ยระหว่างอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ

ชุดข้อมูล	จำนวนแอตทริบิวต์ที่ถูกเลือกไปอยู่ในโหนดตัดสินใจเดียวกันต่อโหนดโดยเฉลี่ย	
	ID3 ที่ปรับปรุงของงานวิจัยแรก	ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ
Chess	1.27	2.24
Connect-4	1.4	2.6
Firm-Teacher	1.33	1.95
Phishing Websites	2.35	3.05
Insurance Company Benchmark	3.45	3.77
Bach Choral Harmony	2.18	2.6

จากตารางข้างบนนี้จะเห็นว่า จำนวนแอตทริบิวต์ที่ถูกเลือกไปอยู่ในโหนดตัดสินใจเดียวกันต่อโหนดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่ามากกว่าจำนวนแอตทริบิวต์ที่ถูกเลือกไปอยู่ในโหนดตัดสินใจเดียวกันต่อโหนดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในชุดข้อมูลทั้งหมดอย่างเห็นได้ชัด ซึ่งเป็นสาเหตุที่ทำให้ความลึกสูงสุดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าที่น้อยกว่าความลึกสูงสุดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในชุดข้อมูลทั้งหมดอย่างเห็นได้ชัด จากตารางข้างบนนี้ยังสามารถสรุปได้อีกว่า อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอสามารถจัดการกับปัญหาความตึงของวิธีการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกได้ ดังจะเห็นได้จากการที่จำนวนแอตทริบิวต์ที่ถูกเลือกไปอยู่ในโหนดตัดสินใจเดียวกันต่อโหนดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ ซึ่งมีค่าที่มากกว่าจำนวนแอตทริบิวต์ที่ถูกเลือกไปอยู่ในโหนดตัดสินใจเดียวกันต่อโหนดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในชุดข้อมูลทั้งหมด

6.4.3 จำนวนโหนดโดยเฉลี่ย

ตารางที่ 6.50 จำนวนโหนดโดยเฉลี่ยระหว่างอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ

ชุดข้อมูล	จำนวนโหนดโดยเฉลี่ย	
	ID3 ที่ปรับปรุงของงานวิจัยแรก	ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ
Chess	85	111
Connect-4	55,127.28	130,123.28
Firm-Teacher	3,517.07	3,611.24
Phishing Websites	4,157.93	4,399.06
Insurance Company Benchmark	2,327,201.11	2,443,915.74
Bach Choral Harmony	7,956.71	7,333.89

จากตารางข้างบน จำนวนโหนดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่ามากกว่าจำนวนโหนดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในชุดข้อมูลส่วนมากอย่างเห็นได้ชัด ซึ่งจำนวนโหนดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าที่เพิ่มขึ้นมากกว่า 2% จากจำนวนโหนดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในชุดข้อมูล Chess, Connect-4, Firm-Teacher, Phishing Websites และ Insurance Company Benchmark โดยในชุดข้อมูล Chess และ Connect-4 มีจำนวนโหนดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอเพิ่มขึ้นมากกว่า 30% จากจำนวนโหนดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก ในชุดข้อมูล Bach Choral Harmony เป็นชุดข้อมูลเดียวที่มีจำนวนโหนดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอน้อยกว่าจำนวนโหนดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก 7.83% ซึ่งสามารถสรุปได้ว่า จำนวนโหนดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าที่เพิ่มขึ้นจากจำนวนโหนดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกอย่างชัดเจนภายในชุดข้อมูลส่วนใหญ่

6.4.4 เวลาในการฝึกฝนโดยเฉลี่ย

ตารางที่ 6.51 เวลาในการฝึกฝนโดยเฉลี่ยระหว่างอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ

ชุดข้อมูล	เวลาในการฝึกฝนโดยเฉลี่ย (มิลลิวินาที)	
	ID3 ที่ปรับปรุงของงานวิจัยแรก	ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ
Chess	15.46	10.19
Connect-4	2,272.29	1,774.06
Firm-Teacher	101.38	63.27
Phishing Websites	222.57	174.55
Insurance Company Benchmark	131,322.21	127,870.45
Bach Choral Harmony	464.66	353.17

จากตารางข้างบน เวลาในการฝึกฝนโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าน้อยกว่าเวลาในการฝึกฝนโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในชุดข้อมูลทั้งหมดอย่างเห็นได้ชัด ซึ่งสามารถสรุปได้ว่าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอสามารถลดเวลาในการฝึกฝนโดยเฉลี่ยจากเวลาในการฝึกฝนโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในชุดข้อมูลทั้งหมดได้อย่างเห็นได้ชัด

6.4.5 เวลาในการทดสอบโดยเฉลี่ย

ตารางที่ 6.52 เวลาในการทดสอบโดยเฉลี่ยระหว่างอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ

ชุดข้อมูล	เวลาในการทดสอบโดยเฉลี่ย (มิลลิวินาที)	
	ID3 ที่ปรับปรุงของงานวิจัยแรก	ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ
Chess	0.0074	0.0050
Connect-4	0.3385	0.1855
Firm-Teacher	0.0541	0.0298
Phishing Websites	0.0314	0.0180
Insurance Company Benchmark	0.0178	0.0113
Bach Choral Harmony	0.0219	0.0142

จากตารางข้างบน เวลาในการทดสอบโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าน้อยกว่าเวลาในการทดสอบโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในชุดข้อมูลทั้งหมดอย่างเห็นได้ชัด ซึ่งเวลาในการทดสอบโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าที่ลดลงมากกว่า 30% จากเวลาในการทดสอบโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในชุดข้อมูลทั้งหมด ซึ่งสามารถสรุปได้ว่าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอสามารถลดเวลาในการทดสอบโดยเฉลี่ยจากเวลาในการทดสอบโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในชุดข้อมูลทั้งหมดได้อย่างเห็นได้ชัด

ตามที่ได้กล่าวไว้ในผลการทดลองด้านเวลาในการทดสอบโดยเฉลี่ยระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เวลาในการทดสอบโดยเฉลี่ย จะขึ้นอยู่กับความลึกโดยเฉลี่ยในการจำแนกโดยตรง ตารางถัดไปแสดงความลึกโดยเฉลี่ยในการจำแนกระหว่างอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ

ตารางที่ 6.53 ความลึกโดยเฉลี่ยในการจำแนกระหว่างอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ

ชุดข้อมูล	ความลึกโดยเฉลี่ยในการจำแนก	
	ID3 ที่ปรับปรุงของงานวิจัยแรก	ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ
Chess	4.29	2.7
Connect-4	9.36	5.17
Firm-Teacher	9.36	5.21
Phishing Websites	5.25	3.04
Insurance Company Benchmark	3.5	2.21
Bach Choral Harmony	7.16	4.7

จากตารางข้างบน จะเห็นว่าความลึกโดยเฉลี่ยในการจำแนกของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าน้อยกว่าความลึกโดยเฉลี่ยในการจำแนกของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในชุดข้อมูลทั้งหมดอย่างเห็นได้ชัด ตารางถัดไปแสดงร้อยละการลดลงของเวลาในการทดสอบโดยเฉลี่ยและความลึกโดยเฉลี่ยในการจำแนกของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เปรียบเทียบกับอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก

ตารางที่ 6.54 ร้อยละการลดลงของเวลาในการทดสอบโดยเฉลี่ยและความลึกโดยเฉลี่ยในการจำแนกของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เปรียบเทียบกับอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก

ชุดข้อมูล	ร้อยละการลดลงของเวลาในการทดสอบโดยเฉลี่ย	ร้อยละการลดลงของความลึกโดยเฉลี่ยในการจำแนก
Chess	32.54%	37.07%
Connect-4	44.97%	44.76%
Firm-Teacher	44.92%	44.34%
Phishing Websites	42.68%	42.10%
Insurance Company Benchmark	36.52%	36.86%
Bach Choral Harmony	35.16%	34.36%

จากตารางข้างบน จะเห็นว่าร้อยละการลดลงของเวลาในการทดสอบโดยเฉลี่ยมีค่าที่ใกล้เคียงกับร้อยละการลดลงของความลึกโดยเฉลี่ยในการจำแนก ตามที่ได้กล่าวไว้ในผลการทดลองด้านเวลาในการทดสอบโดยเฉลี่ยระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ การลดลงของความลึกโดยเฉลี่ยในการจำแนกส่งผลต่อการลดลงของเวลาใน

การทดสอบโดยเฉลี่ย ดังนั้นเวลาในการทดสอบโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ จึงมีค่าน้อยกว่าเวลาในการทดสอบโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในชุดข้อมูลทั้งหมด

6.5 การทดลองระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝนที่น้อย

เนื่องจากผลการทดลองในหัวข้อด้านบนที่ได้ทดลองมา ใช้อัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 50% : 50% จะเห็นว่าอัตราส่วนของชุดข้อมูลฝึกฝนคือ 50% ของชุดข้อมูลทั้งหมด ซึ่งถือว่ามากโดยเฉพาะในชุดข้อมูลขนาดใหญ่ที่มีจำนวนตัวอย่างมาก การใช้อัตราส่วนของชุดข้อมูลฝึกฝนที่มากตามที่การทดลองในหัวข้อด้านบนกำหนดไว้ ผลที่ได้คือความแม่นยำโดยเฉลี่ยระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าใกล้เคียงกัน ดังนั้นการทดลองนี้ทำขึ้นเพื่อสังเกตและวิเคราะห์ลักษณะของความเป็นไปในความแม่นยำโดยเฉลี่ยระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝนที่น้อย ซึ่งการทดลองนี้จะใช้อัตราส่วนของชุดข้อมูลฝึกฝน 15 – 25%

6.5.1 รูปแบบการทดลอง

การทดลองอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอและอัลกอริทึม ID3 แบบดั้งเดิมนี้ จะทำการทดลองโดยใช้ชุดข้อมูลเดียวกันกับการทดลองในหัวข้อด้านบนทั้งหมด แต่ละชุดข้อมูลนั้นจะนำมาทดลองโดยทำการสร้างต้นไม้ตัดสินใจจากชุดข้อมูลฝึกฝน และทำการทดสอบต้นไม้ตัดสินใจที่สร้างด้วยชุดข้อมูลทดสอบ กระบวนการนี้จะทำทั้งหมด 6 ครั้ง แต่ละครั้งใช้ชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบในอัตราส่วนที่ต่างกัน (ชุดข้อมูลหลักที่เตรียมมาจะทำการแบ่งออกเป็น 2 ส่วน : ชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ ซึ่งแต่ละครั้งที่ทำการทดลองจะทำการแบ่งชุดข้อมูลโดยใช้อัตราส่วนที่ไม่เหมือนกัน การแบ่งชุดข้อมูลยังคงมีการรักษาสมดุลของจำนวนตัวอย่างที่อยู่ในแต่ละคลาสของทั้งสองส่วนตามเดิม) อัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบในแต่ละครั้งของการทดลองแสดงดังตารางถัดไป

ตารางที่ 6.55 อัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบในแต่ละครั้งของการทดลอง

ครั้งที่	อัตราส่วน	
	ชุดข้อมูลฝึกฝน	ชุดข้อมูลทดสอบ
1	15%	85%
2	17%	83%
3	19%	81%
4	21%	79%
5	23%	77%
6	25%	75%

6.5.2 พารามิเตอร์

เนื่องจากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอต้องการพารามิเตอร์ 2 ตัวคือ N (จำนวนแอตทริบิวต์สูงสุดที่สามารถเลือกมาไว้ในโหนดตัดสินใจเดียวกัน) และ D (ร้อยละของค่าเกนความรู้ของแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1) เพราะฉะนั้นจึงต้องทำการทดลองเบื้องต้นเพื่อหาค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ โดยการเลือกพารามิเตอร์ N และ D แบบที่ดีที่สุด จะพิจารณาจากเกณฑ์ในการวัดเรียงตามลำดับแบบเดียวกันกับที่ใช้ในการทดลองเบื้องต้นของหัวข้อ 6.2.3

สำหรับการทดลองเบื้องต้นในหัวข้อนี้ จะทำการทดลองโดยกำหนดค่าพารามิเตอร์ D เป็น 0.2 และ 0.4 ในแต่ละค่าของพารามิเตอร์ D ที่กำหนด จะนำมาทดลองกับการกำหนดค่าพารามิเตอร์ N เป็น 2 และ 3

6.5.2.1 การหาค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบของชุดข้อมูล Connect-4

ทำการพิจารณาเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของแต่ละอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ โดยเริ่มจากการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 15% : 85% ดังตารางถัดไป

ตารางที่ 6.56 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 15% : 85% บนชุดข้อมูล Connect-4

เกณฑ์ในการวัด	D = 0.2		D = 0.4	
	N = 2	N = 3	N = 2	N = 3
ความลึกสูงสุดโดยเฉลี่ย	11	11	**10**	10
ความแม่นยำโดยเฉลี่ย	71.06%	71.29%	**71.27%**	71.36%
จำนวนโหนดโดยเฉลี่ย	11,545	17,026	**11,962**	18,364

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 15% : 85% คือ D = 0.4 และ N = 2 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 2 และค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 2 มีจำนวนโหนดโดยเฉลี่ยน้อยกว่า ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 17% : 83% ดังตารางถัดไป

ตารางที่ 6.57 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 17% : 83% บนชุดข้อมูล Connect-4

เกณฑ์ในการวัด	D = 0.2		D = 0.4	
	N = 2	N = 3	N = 2	N = 3
ความลึกสูงสุดโดยเฉลี่ย	11	11	**9**	9
ความแม่นยำโดยเฉลี่ย	70.91%	70.93%	**71.07%**	71.40%
จำนวนโหนดโดยเฉลี่ย	13,132	19,606	**13,855**	21,205

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 17% : 83% คือ D = 0.4 และ N = 2 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่า

ค่าพารามิเตอร์ D = 0.4 และ N = 2 และค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 2 มีจำนวนโหนดโดยเฉลี่ยน้อยกว่า ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 19% : 81% ดังตารางถัดไป

ตารางที่ 6.58 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 19% : 81% บนชุดข้อมูล Connect-4

เกณฑ์ในการวัด	**D = 0.2**		**D = 0.4**	
	N = 2	**N = 3**	**N = 2**	**N = 3**
ความลึกสูงสุดโดยเฉลี่ย	11	11	**9**	9
ความแม่นยำโดยเฉลี่ย	71.39%	71.62%	**71.60%**	71.81%
จำนวนโหนดโดยเฉลี่ย	14,218	20,575	**14,923**	22,525

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 19% : 81% คือ D = 0.4 และ N = 2 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 2 และค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 2 มีจำนวนโหนดโดยเฉลี่ยน้อยกว่า ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 21% : 79% ดังตารางถัดไป

ตารางที่ 6.59 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 21% : 79% บนชุดข้อมูล Connect-4

เกณฑ์ในการวัด	D = 0.2		D = 0.4	
	N = 2	N = 3	N = 2	N = 3
ความลึกสูงสุดโดยเฉลี่ย	12	11	**10**	14
ความแม่นยำโดยเฉลี่ย	72.30%	72.53%	**72.22%**	72.30%
จำนวนโหนดโดยเฉลี่ย	14,959	20,515	**16,279**	22,942

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 21% : 79% คือ D = 0.4 และ N = 2 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 2 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 23% : 77% ดังตารางถัดไป

ตารางที่ 6.60 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 23% : 77% บนชุดข้อมูล Connect-4

เกณฑ์ในการวัด	D = 0.2		D = 0.4	
	N = 2	N = 3	N = 2	N = 3
ความลึกสูงสุดโดยเฉลี่ย	12	12	**10**	10
ความแม่นยำโดยเฉลี่ย	71.97%	72.01%	**72.02%**	72.24%
จำนวนโหนดโดยเฉลี่ย	16,747	23,383	**17,944**	25,525

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 23% : 77% คือ D = 0.4 และ N = 2 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 2 และค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4

และ N = 2 มีจำนวนโหนดโดยเฉลี่ยน้อยกว่า ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 25% : 75% ดังตารางถัดไป

ตารางที่ 6.61 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 25% : 75% บนชุดข้อมูล Connect-4

เกณฑ์ในการวัด	**D = 0.2**		**D = 0.4**	
	N = 2	**N = 3**	**N = 2**	**N = 3**
ความลึกสูงสุดโดยเฉลี่ย	12	13	**10**	11
ความแม่นยำโดยเฉลี่ย	69.96%	70.09%	**70.39%**	70.51%
จำนวนโหนดโดยเฉลี่ย	20,212	27,943	**20,884**	30,589

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 25% : 75% คือ D = 0.4 และ N = 2 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 2 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

6.5.2.2 การหาค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบของชุดข้อมูล Chess

ทำการพิจารณาเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของแต่ละอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ โดยเริ่มจากการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 15% : 85% ดังตารางถัดไป

ตารางที่ 6.62 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 15% : 85% บนชุดข้อมูล Chess

เกณฑ์ในการวัด	D = 0.2		D = 0.4	
	N = 2	N = 3	N = 2	N = 3
ความลึกสูงสุดโดยเฉลี่ย	9	8	**7**	7
ความแม่นยำโดยเฉลี่ย	95.40%	95.10%	**95.07%**	95.58%
จำนวนโหนดโดยเฉลี่ย	73	91	**71**	87

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 15% : 85% คือ D = 0.4 และ N = 2 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 2 และค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 2 มีจำนวนโหนดโดยเฉลี่ยน้อยกว่า ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 17% : 83% ดังตารางถัดไป

ตารางที่ 6.63 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 17% : 83% บนชุดข้อมูล Chess

เกณฑ์ในการวัด	D = 0.2		D = 0.4	
	N = 2	N = 3	N = 2	N = 3
ความลึกสูงสุดโดยเฉลี่ย	8	8	**8**	8
ความแม่นยำโดยเฉลี่ย	96.91%	96.91%	**96.83%**	96.72%
จำนวนโหนดโดยเฉลี่ย	74	96	**72**	96

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 17% : 83% คือ D = 0.4 และ N = 2 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าความลึก

สูงสุดโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 2 มีจำนวนโหนดโดยเฉลี่ยน้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 19% : 81% ดังตารางถัดไป

ตารางที่ 6.64 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 19% : 81% บนชุดข้อมูล Chess

เกณฑ์ในการวัด	**D = 0.2**		**D = 0.4**	
	N = 2	**N = 3**	**N = 2**	**N = 3**
ความลึกสูงสุดโดยเฉลี่ย	8	8	7	**6**
ความแม่นยำโดยเฉลี่ย	96.64%	96.79%	97.18%	**97.18%**
จำนวนโหนดโดยเฉลี่ย	62	62	60	**68**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 19% : 81% คือ D = 0.4 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 21% : 79% ดังตารางถัดไป

ตารางที่ 6.65 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 21% : 79% บนชุดข้อมูล Chess

เกณฑ์ในการวัด	D = 0.2		D = 0.4	
	N = 2	N = 3	N = 2	N = 3
ความลึกสูงสุดโดยเฉลี่ย	8	7	7	**6**
ความแม่นยำโดยเฉลี่ย	96.99%	97.23%	96.67%	**97.07%**
จำนวนโหนดโดยเฉลี่ย	85	95	87	**101**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 21% : 79% คือ D = 0.4 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 23% : 77% ดังตารางถัดไป

ตารางที่ 6.66 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 23% : 77% บนชุดข้อมูล Chess

เกณฑ์ในการวัด	D = 0.2		D = 0.4	
	N = 2	N = 3	N = 2	N = 3
ความลึกสูงสุดโดยเฉลี่ย	8	8	**7**	8
ความแม่นยำโดยเฉลี่ย	97.48%	97.36%	**96.79%**	96.79%
จำนวนโหนดโดยเฉลี่ย	57	63	**65**	83

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 23% : 77% คือ D = 0.4 และ N = 2 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 2 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 25% : 75% ดังตารางถัดไป

ตารางที่ 6.67 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 25% : 75% บนชุดข้อมูล Chess

เกณฑ์ในการวัด	**D = 0.2**		**D = 0.4**	
	N = 2	**N = 3**	**N = 2**	**N = 3**
ความลึกสูงสุดโดยเฉลี่ย	7	7	8	**6**
ความแม่นยำโดยเฉลี่ย	97.21%	97.75%	97.91%	**97.66%**
จำนวนโหนดโดยเฉลี่ย	73	93	77	**79**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 25% : 75% คือ D = 0.4 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

6.5.2.3 การหาค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบของชุดข้อมูล Phishing Websites

ทำการพิจารณาเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของแต่ละอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ โดยเริ่มจากการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 15% : 85% ดังตารางถัดไป

ตารางที่ 6.68 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 15% : 85% บนชุดข้อมูล Phishing Websites

เกณฑ์ในการวัด	D = 0.2		D = 0.4	
	N = 2	N = 3	N = 2	N = 3
ความลึกสูงสุดโดยเฉลี่ย	20	17	17	**14**
ความแม่นยำโดยเฉลี่ย	93.08%	93.07%	93.36%	**93.31%**
จำนวนโหนดโดยเฉลี่ย	737	957	787	**1,043**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 15% : 85% คือ D = 0.4 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 17% : 83% ดังตารางถัดไป

ตารางที่ 6.69 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 17% : 83% บนชุดข้อมูล Phishing Websites

เกณฑ์ในการวัด	D = 0.2		D = 0.4	
	N = 2	N = 3	N = 2	N = 3
ความลึกสูงสุดโดยเฉลี่ย	19	**15**	18	15
ความแม่นยำโดยเฉลี่ย	93.72%	**93.83%**	93.59%	93.66%
จำนวนโหนดโดยเฉลี่ย	886	**1,145**	872	1,166

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 17% : 83% คือ D = 0.2 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.2 และ N = 3 และค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.2

และ N = 3 มีจำนวนโหนดโดยเฉลี่ยน้อยกว่า ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 19% : 81% ดังตารางถัดไป

ตารางที่ 6.70 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 19% : 81% บนชุดข้อมูล Phishing Websites

เกณฑ์ในการวัด	**D = 0.2**		**D = 0.4**	
	N = 2	**N = 3**	**N = 2**	**N = 3**
ความลึกสูงสุดโดยเฉลี่ย	18	**15**	18	15
ความแม่นยำโดยเฉลี่ย	93.13%	**93.01%**	93.22%	93.03%
จำนวนโหนดโดยเฉลี่ย	787	**1,009**	806	1,045

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 19% : 81% คือ D = 0.2 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.2 และ N = 3 และค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.2 และ N = 3 มีจำนวนโหนดโดยเฉลี่ยน้อยกว่า ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 21% : 79% ดังตารางถัดไป

ตารางที่ 6.71 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 21% : 79% บนชุดข้อมูล Phishing Websites

เกณฑ์ในการวัด	D = 0.2		D = 0.4	
	N = 2	N = 3	N = 2	N = 3
ความลึกสูงสุดโดยเฉลี่ย	18	16	17	**14**
ความแม่นยำโดยเฉลี่ย	93.52%	93.50%	93.41%	**93.47%**
จำนวนโหนดโดยเฉลี่ย	909	1,214	886	**1,262**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 21% : 79% คือ D = 0.4 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 23% : 77% ดังตารางถัดไป

ตารางที่ 6.72 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 23% : 77% บนชุดข้อมูล Phishing Websites

เกณฑ์ในการวัด	D = 0.2		D = 0.4	
	N = 2	N = 3	N = 2	N = 3
ความลึกสูงสุดโดยเฉลี่ย	19	16	17	**15**
ความแม่นยำโดยเฉลี่ย	93.39%	93.27%	93.53%	**93.39%**
จำนวนโหนดโดยเฉลี่ย	1,068	1,356	1,084	**1,409**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 23% : 77% คือ D = 0.4 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 25% : 75% ดังตารางถัดไป

ตารางที่ 6.73 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 25% : 75% บนชุดข้อมูล Phishing Websites

เกณฑ์ในการวัด	**D = 0.2**		**D = 0.4**	
	N = 2	**N = 3**	**N = 2**	**N = 3**
ความลึกสูงสุดโดยเฉลี่ย	19	16	18	**15**
ความแม่นยำโดยเฉลี่ย	93.99%	94.05%	94.39%	**94.48%**
จำนวนโหนดโดยเฉลี่ย	1,537	1,991	1,507	**1,991**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 25% : 75% คือ D = 0.4 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

6.5.2.4 การหาค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบของชุดข้อมูล Insurance Company Benchmark

ทำการพิจารณาเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของแต่ละอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ โดยเริ่มจากการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 15% : 85% ดังตารางถัดไป

ตารางที่ 6.74 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 15% : 85% บนชุดข้อมูล Insurance Company Benchmark

เกณฑ์ในการวัด	D = 0.2		D = 0.4	
	N = 2	N = 3	N = 2	N = 3
ความลึกสูงสุดโดยเฉลี่ย	43	30	43	**30**
ความแม่นยำโดยเฉลี่ย	92.01%	92.38%	91.92%	**92.43%**
จำนวนโหนดโดยเฉลี่ย	14,078	79,564	13,734	**79,000**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 15% : 85% คือ D = 0.4 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.2 และ N = 3 และค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 3 มีจำนวนโหนดโดยเฉลี่ยน้อยกว่า ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 17% : 83% ดังตารางถัดไป

ตารางที่ 6.75 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 17% : 83% บนชุดข้อมูล Insurance Company Benchmark

เกณฑ์ในการวัด	D = 0.2		D = 0.4	
	N = 2	N = 3	N = 2	N = 3
ความลึกสูงสุดโดยเฉลี่ย	43	**30**	43	30
ความแม่นยำโดยเฉลี่ย	91.65%	**92.19%**	92.10%	92.43%
จำนวนโหนดโดยเฉลี่ย	15,258	**82,606**	15,253	87,159

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 17% : 83% คือ D = 0.2 และ N = 3 เนื่องจาก

เมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.2 และ N = 3 และค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.2 และ N = 3 มีจำนวนโหนดโดยเฉลี่ยน้อยกว่า ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 19% : 81% ดังตารางถัดไป

ตารางที่ 6.76 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 19% : 81% บนชุดข้อมูล Insurance Company Benchmark

เกณฑ์ในการวัด	**D = 0.2**		**D = 0.4**	
	N = 2	**N = 3**	**N = 2**	**N = 3**
ความลึกสูงสุดโดยเฉลี่ย	43	30	43	**29**
ความแม่นยำโดยเฉลี่ย	91.60%	91.83%	91.78%	**92.12%**
จำนวนโหนดโดยเฉลี่ย	23,473	120,735	23,269	**124,099**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 19% : 81% คือ D = 0.4 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 21% : 79% ดังตารางถัดไป

ตารางที่ 6.77 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 21% : 79% บนชุดข้อมูล Insurance Company Benchmark

เกณฑ์ในการวัด	D = 0.2		D = 0.4	
	N = 2	N = 3	N = 2	N = 3
ความลึกสูงสุดโดยเฉลี่ย	45	31	43	**30**
ความแม่นยำโดยเฉลี่ย	91.15%	91.21%	91.42%	**91.62%**
จำนวนโหนดโดยเฉลี่ย	22,885	117,119	24,335	**129,508**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 21% : 79% คือ D = 0.4 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 23% : 77% ดังตารางถัดไป

ตารางที่ 6.78 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 23% : 77% บนชุดข้อมูล Insurance Company Benchmark

เกณฑ์ในการวัด	D = 0.2		D = 0.4	
	N = 2	N = 3	N = 2	N = 3
ความลึกสูงสุดโดยเฉลี่ย	44	31	43	**30**
ความแม่นยำโดยเฉลี่ย	91.71%	91.94%	91.78%	**92.04%**
จำนวนโหนดโดยเฉลี่ย	28,582	135,523	27,952	**137,877**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 23% : 77% คือ D = 0.4 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่า

ค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 25% : 75% ดังตารางถัดไป

ตารางที่ 6.79 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 25% : 75% บนชุดข้อมูล Insurance Company Benchmark

เกณฑ์ในการวัด	**D = 0.2**		**D = 0.4**	
	N = 2	**N = 3**	**N = 2**	**N = 3**
ความลึกสูงสุดโดยเฉลี่ย	43	**30**	43	30
ความแม่นยำโดยเฉลี่ย	91.35%	**91.52%**	91.29%	91.54%
จำนวนโหนดโดยเฉลี่ย	29,680	**148,656**	29,884	149,785

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 25% : 75% คือ D = 0.2 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.2 และ N = 3 และค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.2 และ N = 3 มีจำนวนโหนดโดยเฉลี่ยน้อยกว่า ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

6.5.2.5 การหาค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบของชุดข้อมูล Firm-Teacher

ทำการพิจารณาเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของแต่ละอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ โดยเริ่มจากการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 15% : 85% ดังตารางถัดไป

ตารางที่ 6.80 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 15% : 85% บนชุดข้อมูล Firm-Teacher

เกณฑ์ในการวัด	D = 0.2		D = 0.4	
	N = 2	N = 3	N = 2	N = 3
ความลึกสูงสุดโดยเฉลี่ย	9	9	**8**	8
ความแม่นยำโดยเฉลี่ย	67.46%	67.86%	**67.73%**	68.51%
จำนวนโหนดโดยเฉลี่ย	1,117	1,365	**1,147**	1,431

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 15% : 85% คือ D = 0.4 และ N = 2 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 2 และค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 2 มีจำนวนโหนดโดยเฉลี่ยน้อยกว่า ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 17% : 83% ดังตารางถัดไป

ตารางที่ 6.81 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 17% : 83% บนชุดข้อมูล Firm-Teacher

เกณฑ์ในการวัด	D = 0.2		D = 0.4	
	N = 2	N = 3	N = 2	N = 3
ความลึกสูงสุดโดยเฉลี่ย	9	9	**9**	9
ความแม่นยำโดยเฉลี่ย	67.66%	68.18%	**67.15%**	67.61%
จำนวนโหนดโดยเฉลี่ย	1,285	1,583	**1,255**	1,597

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 17% : 83% คือ D = 0.4 และ N = 2 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าความลึก

สูงสุดโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 2 มีจำนวนโหนดโดยเฉลี่ยน้อยที่สุดดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 19% : 81% ดังตารางถัดไป

ตารางที่ 6.82 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 19% : 81% บนชุดข้อมูล Firm-Teacher

เกณฑ์ในการวัด	**D = 0.2**		**D = 0.4**	
	N = 2	**N = 3**	**N = 2**	**N = 3**
ความลึกสูงสุดโดยเฉลี่ย	10	10	9	**8**
ความแม่นยำโดยเฉลี่ย	69.07%	69.14%	69.23%	**69.52%**
จำนวนโหนดโดยเฉลี่ย	1,403	1,723	1,415	**1,781**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 19% : 81% คือ D = 0.4 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 21% : 79% ดังตารางถัดไป

ตารางที่ 6.83 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 21% : 79% บนชุดข้อมูล Firm-Teacher

เกณฑ์ในการวัด	**D = 0.2**		**D = 0.4**	
	N = 2	**N = 3**	**N = 2**	**N = 3**
ความลึกสูงสุดโดยเฉลี่ย	11	11	**8**	8
ความแม่นยำโดยเฉลี่ย	68.56%	68.28%	**68.76%**	68.97%
จำนวนโหนดโดยเฉลี่ย	1,535	1,863	**1,503**	1,877

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 21% : 79% คือ D = 0.4 และ N = 2 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 2 และค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 2 มีจำนวนโหนดโดยเฉลี่ยน้อยกว่า ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 23% : 77% ดังตารางถัดไป

ตารางที่ 6.84 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 23% : 77% บนชุดข้อมูล Firm-Teacher

เกณฑ์ในการวัด	**D = 0.2**		**D = 0.4**	
	N = 2	**N = 3**	**N = 2**	**N = 3**
ความลึกสูงสุดโดยเฉลี่ย	10	10	9	**8**
ความแม่นยำโดยเฉลี่ย	68.15%	68.51%	68.83%	**68.86%**
จำนวนโหนดโดยเฉลี่ย	1,685	2,099	1,731	**2,137**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 23% : 77% คือ D = 0.4 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่า

ค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 25% : 75% ดังตารางถัดไป

ตารางที่ 6.85 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 25% : 75% บนชุดข้อมูล Firm-Teacher

เกณฑ์ในการวัด	**D = 0.2**		**D = 0.4**	
	N = 2	**N = 3**	**N = 2**	**N = 3**
ความลึกสูงสุดโดยเฉลี่ย	10	10	**9**	9
ความแม่นยำโดยเฉลี่ย	69.20%	69.37%	**70.48%**	70.68%
จำนวนโหนดโดยเฉลี่ย	1,783	2,129	**1,825**	2,267

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 25% : 75% คือ D = 0.4 และ N = 2 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 2 และค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 2 มีจำนวนโหนดโดยเฉลี่ยน้อยกว่า ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

6.5.2.6 การหาค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดในแต่ละอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบของชุดข้อมูล Bach Choral Harmony

ทำการพิจารณาเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของแต่ละอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ โดยเริ่มจากการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 15% : 85% ดังตารางถัดไป

ตารางที่ 6.86 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 15% : 85% บนชุดข้อมูล Bach Choral Harmony

เกณฑ์ในการวัด	D = 0.2		D = 0.4	
	N = 2	N = 3	N = 2	N = 3
ความลึกสูงสุดโดยเฉลี่ย	9	**8**	9	8
ความแม่นยำโดยเฉลี่ย	61.92%	**62.00%**	61.90%	61.94%
จำนวนโหนดโดยเฉลี่ย	1,419	**1,708**	1,423	1,729

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 15% : 85% คือ D = 0.2 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.2 และ N = 3 และค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.2 และ N = 3 มีจำนวนโหนดโดยเฉลี่ยน้อยกว่า ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 17% : 83% ดังตารางถัดไป

ตารางที่ 6.87 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 17% : 83% บนชุดข้อมูล Bach Choral Harmony

เกณฑ์ในการวัด	D = 0.2		D = 0.4	
	N = 2	N = 3	N = 2	N = 3
ความลึกสูงสุดโดยเฉลี่ย	10	9	9	**8**
ความแม่นยำโดยเฉลี่ย	62.79%	62.98%	63.17%	**63.30%**
จำนวนโหนดโดยเฉลี่ย	1,620	1,948	1,620	**1,965**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 17% : 83% คือ D = 0.4 และ N = 3 เนื่องจาก

เมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 19% : 81% ดังตารางถัดไป

ตารางที่ 6.88 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 19% : 81% บนชุดข้อมูล Bach Choral Harmony

เกณฑ์ในการวัด	**D = 0.2**		**D = 0.4**	
	N = 2	**N = 3**	**N = 2**	**N = 3**
ความลึกสูงสุดโดยเฉลี่ย	10	**9**	10	9
ความแม่นยำโดยเฉลี่ย	63.30%	**63.24%**	63.73%	63.58%
จำนวนโหนดโดยเฉลี่ย	1,787	**2,178**	1,777	2,207

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 19% : 81% คือ D = 0.2 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.2 และ N = 3 และค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.2 และ N = 3 มีจำนวนโหนดโดยเฉลี่ยน้อยกว่า ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 21% : 79% ดังตารางถัดไป

ตารางที่ 6.89 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 21% : 79% บนชุดข้อมูล Bach Choral Harmony

เกณฑ์ในการวัด	**D = 0.2**		**D = 0.4**	
	N = 2	**N = 3**	**N = 2**	**N = 3**
ความลึกสูงสุดโดยเฉลี่ย	10	9	10	**9**
ความแม่นยำโดยเฉลี่ย	64.01%	63.99%	63.41%	**64.07%**
จำนวนโหนดโดยเฉลี่ย	1,783	2,193	1,772	**2,186**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 21% : 79% คือ D = 0.4 และ N = 3 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.2 และ N = 3 และค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 3 มีจำนวนโหนดโดยเฉลี่ยน้อยกว่า ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 23% : 77% ดังตารางถัดไป

ตารางที่ 6.90 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 23% : 77% บนชุดข้อมูล Bach Choral Harmony

เกณฑ์ในการวัด	**D = 0.2**		**D = 0.4**	
	N = 2	**N = 3**	**N = 2**	**N = 3**
ความลึกสูงสุดโดยเฉลี่ย	10	9	9	**8**
ความแม่นยำโดยเฉลี่ย	63.92%	63.83%	63.69%	**63.58%**
จำนวนโหนดโดยเฉลี่ย	2,065	2,562	2,082	**2,558**

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 23% : 77% คือ D = 0.4 และ N = 3 เนื่องจาก

เมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ดังนั้นจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยที่น้อยที่สุด ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

ทำการเลือกค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 25% : 75% ดังตารางถัดไป

ตารางที่ 6.91 ผลการทดลองเบื้องต้นของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในการปรับค่าพารามิเตอร์ N = 2, 3 ภายใต้ค่าพารามิเตอร์ D = 0.2, 0.4 ของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 25% : 75% บนชุดข้อมูล Bach Choral Harmony

เกณฑ์ในการวัด	**D = 0.2**		**D = 0.4**	
	N = 2	**N = 3**	**N = 2**	**N = 3**
ความลึกสูงสุดโดยเฉลี่ย	10	10	**9**	9
ความแม่นยำโดยเฉลี่ย	66.15%	65.78%	**66.22%**	65.99%
จำนวนโหนดโดยเฉลี่ย	2,636	3,246	**2,651**	3,283

จากตารางข้างบนค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ 25% : 75% คือ D = 0.4 และ N = 2 เนื่องจากเมื่อพิจารณาเกณฑ์ในการวัดเรียงตามลำดับพบว่า ความแม่นยำโดยเฉลี่ยของการทดลองทั้งหมดในตารางข้างบน มีค่าใกล้เคียงกัน ต่อมาจึงต้องมาพิจารณาที่ความลึกสูงสุดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 2 และค่าพารามิเตอร์ D = 0.4 และ N = 3 มีความลึกสูงสุดโดยเฉลี่ยน้อยที่สุดและเท่ากัน สุดท้ายต้องพิจารณาที่จำนวนโหนดโดยเฉลี่ย พบว่าค่าพารามิเตอร์ D = 0.4 และ N = 2 มีจำนวนโหนดโดยเฉลี่ยน้อยกว่า ดังนั้นจึงถูกเลือกให้เป็นค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดของอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบนี้

6.5.2.7 สรุปค่าพารามิเตอร์แบบที่ดีที่สุดสำหรับการทดลองของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในแต่ละอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบของชุดข้อมูลทั้งหมด

ตารางที่ 6.92 ค่าพารามิเตอร์ N และ D แบบที่ดีที่สุดสำหรับการทดลองของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอในแต่ละอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบของชุดข้อมูลทั้งหมด

ชุดข้อมูล	อัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ					
	15% : 85%	17% : 83%	19% : 81%	21% : 79%	23% : 77%	25% : 75%
Chess	D=0.4, N=2	D=0.4, N=2	D=0.4, N=3	D=0.4, N=3	D=0.4, N=2	D=0.4, N=3
Connect-4	D=0.4, N=2	D=0.4, N=2	D=0.4, N=2	D=0.4, N=2	D=0.4, N=2	D=0.4, N=2
Firm-Teacher	D=0.4, N=2	D=0.4, N=2	D=0.4, N=3	D=0.4, N=2	D=0.4, N=3	D=0.4, N=2
Phishing Websites	D=0.4, N=3	D=0.2, N=3	D=0.2, N=3	D=0.4, N=3	D=0.4, N=3	D=0.4, N=3
Insurance Company Benchmark	D=0.4, N=3	D=0.2, N=3	D=0.4, N=3	D=0.4, N=3	D=0.4, N=3	D=0.2, N=3
Bach Choral Harmony	D=0.2, N=3	D=0.4, N=3	D=0.2, N=3	D=0.4, N=3	D=0.4, N=3	D=0.4, N=2

6.5.3 ผลการทดลองระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝนที่น้อย

6.5.3.1 ความแม่นยำโดยเฉลี่ย

ตารางที่ 6.93 ความแม่นยำโดยเฉลี่ยระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝนที่น้อย

ชุดข้อมูล	ความแม่นยำโดยเฉลี่ย	
	ID3 แบบดั้งเดิม	ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ
Chess	97.24%	96.77%
Connect-4	69.98%	71.43%
Firm-Teacher	68.22%	68.75%
Phishing Websites	93.19%	93.58%
Insurance Company Benchmark	89.92%	91.99%
Bach Choral Harmony	63.37%	63.74%

จากตารางข้างบน ความแม่นยำโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าน้อยกว่าความแม่นยำโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมเพียงเล็กน้อยในชุดข้อมูล Chess ส่วนในชุดข้อมูล Connect-4 และ Insurance Company Benchmark มีความแม่นยำโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมากกว่าความแม่นยำโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมอย่างเห็นได้ชัด สำหรับชุดข้อมูลที่เหลือคือ Firm-Teacher, Phishing Websites และ Bach Choral Harmony นั้นมีความแม่นยำโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมากกว่าความแม่นยำโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมเพียงเล็กน้อย ซึ่งสามารถสรุปได้ว่า เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝนที่น้อย ความแม่นยำโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าใกล้เคียงกับความแม่นยำโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูลส่วนใหญ่

เนื่องจากอัตราส่วนของชุดข้อมูลฝึกฝนที่ใช้ในการทดลองนี้คือ 15 – 25% ยังคงให้ผลด้านความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอใกล้เคียงกับความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูลส่วนใหญ่ ดังนั้นการทดลองในหัวข้อนี้จึงมีการทดลองด้านความแม่นยำเพิ่มเติมโดยใช้อัตราส่วนของชุดข้อมูลฝึกฝนที่น้อยยิ่งขึ้น เพื่อดูลักษณะของความแม่นยำที่ได้ในอัลกอริทึม ID3 ทั้ง 2 แบบต่อไปอีก ตารางถัดไปแสดงความแม่นยำระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 3%

ตารางที่ 6.94 ความแม่นยำระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 3%

ชุดข้อมูล	ความแม่นยำ	
	ID3 แบบดั้งเดิม	ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ
Chess	90.52%	91.13%
Connect-4	62.88%	64.89%
Firm-Teacher	58.90%	61.17%
Phishing Websites	88.34%	90.39%
Insurance Company Benchmark	89.63%	92.08%
Bach Choral Harmony	42.09%	44.08%

จากตารางข้างบนพบว่า การใช้อัตราส่วนของชุดข้อมูลฝึกฝน 3% ทำให้ความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่ามากกว่าความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมอย่างเห็นได้ชัดในชุดข้อมูลส่วนใหญ่ ตารางถัดไปแสดงความแม่นยำระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 1%

ตารางที่ 6.95 ความแม่นยำระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 1%

ชุดข้อมูล	ความแม่นยำ	
	ID3 แบบดั้งเดิม	ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ
Chess	90.33%	90.33%
Connect-4	59.94%	64.15%
Firm-Teacher	55.33%	55.20%
Phishing Websites	89.09%	90.05%
Insurance Company Benchmark	92.42%	93.14%
Bach Choral Harmony	31.92%	36.92%

จากตารางข้างบนพบว่า การใช้อัตราส่วนของชุดข้อมูลฝึกฝน 1% ทำให้ความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมีค่าใกล้เคียงกับความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูลส่วนใหญ่

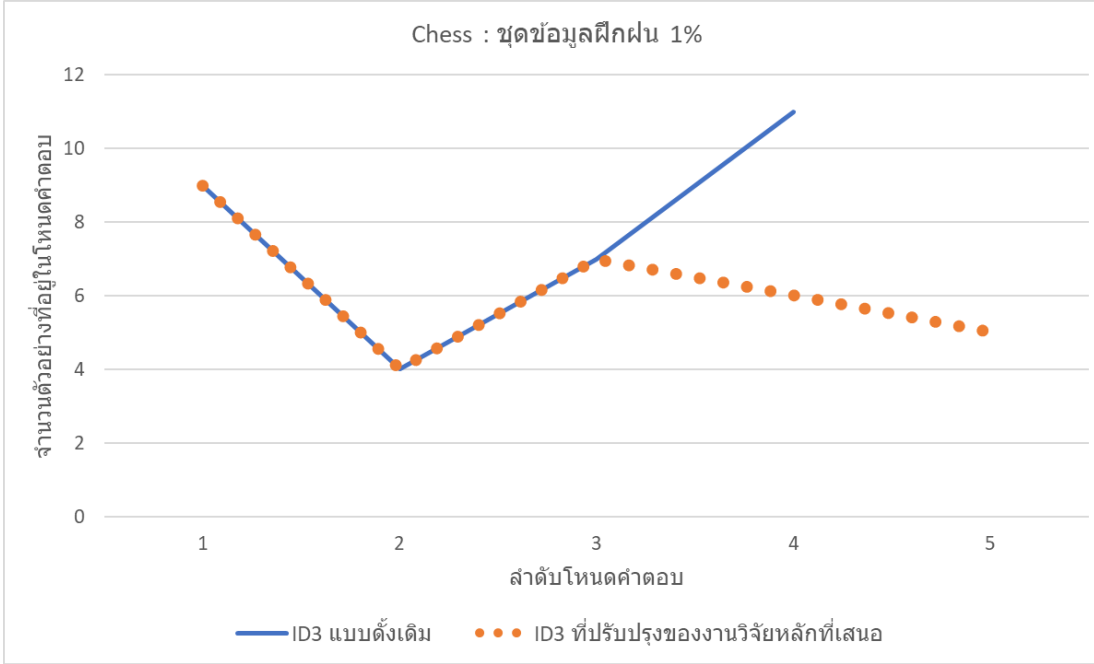
แม้ว่าอัตราส่วนของชุดข้อมูลฝึกฝนจะเหลือ 3% และ 1% แต่ความแม่นยำของทั้งสองอัลกอริทึมยังคงใกล้เคียงกันในชุดข้อมูลส่วนใหญ่ ตารางถัดไปแสดงความแม่นยำระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 0.3% โดยที่ 0.3% เป็นอัตราส่วนของชุดข้อมูลฝึกฝนที่เล็กที่สุดเท่าที่เป็นไปได้สำหรับชุดข้อมูลทั้งหมดที่ใช้ในการทดลอง

ตารางที่ 6.96 ความแม่นยำระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 0.3%

ชุดข้อมูล	**ความแม่นยำ**	
	ID3 แบบดั้งเดิม	**ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ**
Chess	60.43%	56.73%
Connect-4	57.72%	61.85%
Firm-Teacher	50.93%	44.50%
Phishing Websites	86.80%	64.23%
Insurance Company Benchmark	93.98%	93.99%
Bach Choral Harmony	24.17%	24.17%

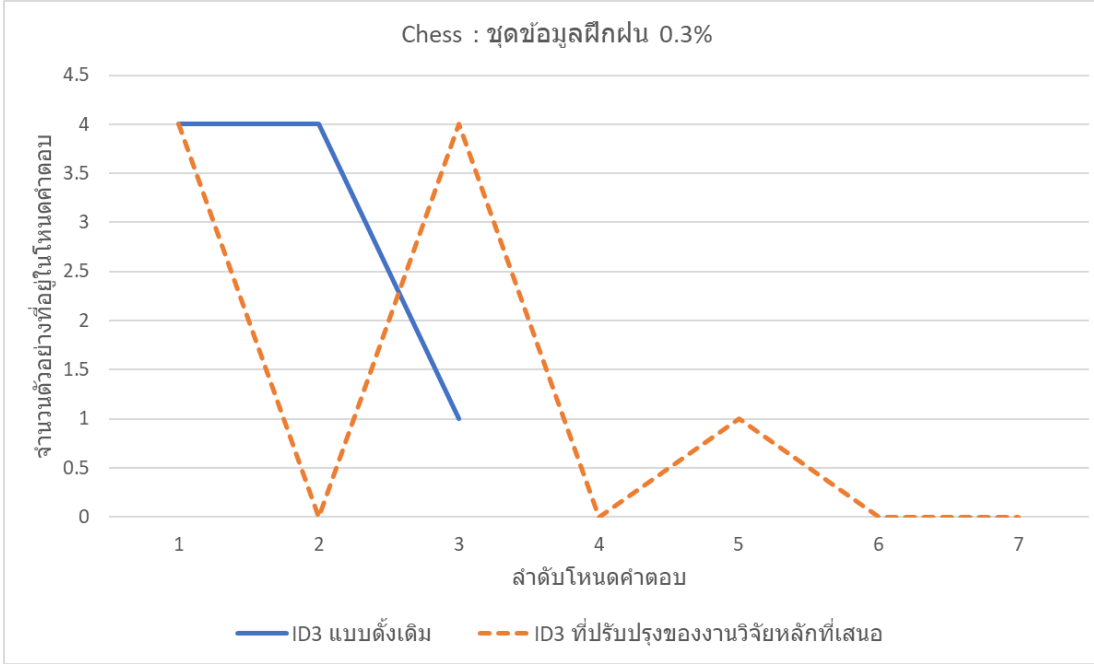
จากตารางข้างบนพบว่า การใช้อัตราส่วนของชุดข้อมูลฝึกฝน 0.3% ทำให้ความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าน้อยกว่าความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมอย่างเห็นได้ชัดในชุดข้อมูล Chess, Firm-Teacher และ Phishing Websites ในขณะที่ชุดข้อมูล Insurance Company Benchmark และ Bach Choral Harmony มีความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ ใกล้เคียงกับความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิม ส่วนชุดข้อมูล Connect-4 มีความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มากกว่าความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมอย่างเห็นได้ชัด

จะเห็นได้อย่างชัดเจนว่าเมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 0.3% ทำให้ชุดข้อมูลจำนวนครึ่งหนึ่งมีความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอน้อยกว่าความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมอย่างเห็นได้ชัด ซึ่งก่อนหน้านี้ตอนที่ใช้อัตราส่วนของชุดข้อมูลฝึกฝน 3% และ 1% ความแม่นยำของทั้งสองอัลกอริทึมยังคงใกล้เคียงกัน สาเหตุที่ผลออกมาเป็นแบบนี้เพราะจำนวนตัวอย่างที่อยู่ในโหนดคำตอบทั้งหมดของอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ



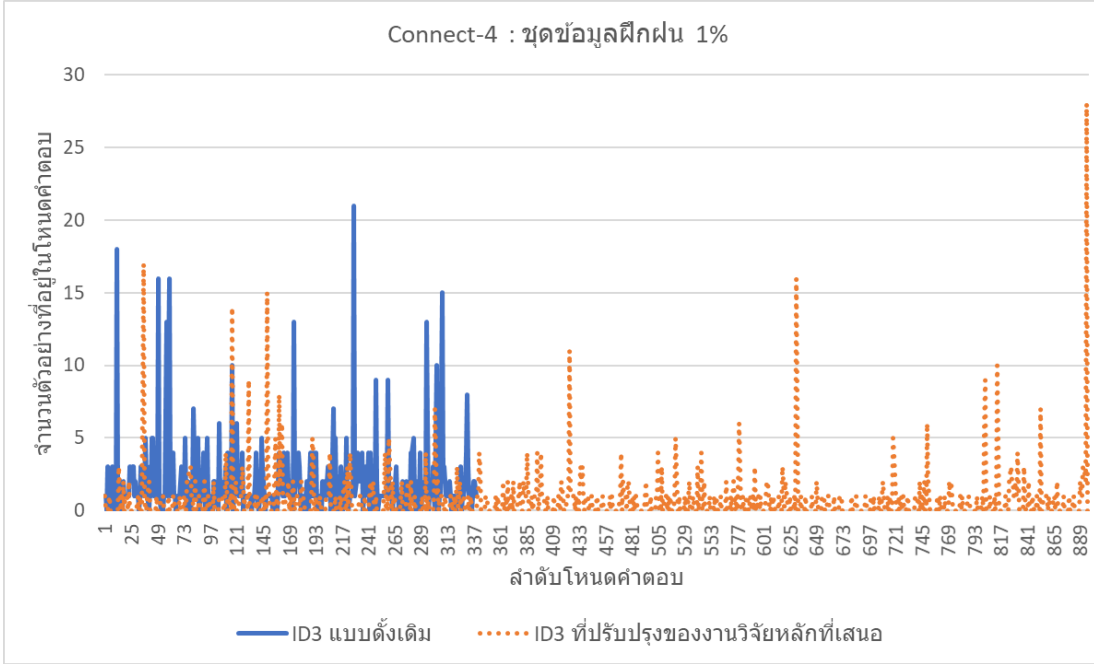
รูปที่ 6.2 จำนวนตัวอย่างที่อยู่ในโหนดคำตอบทั้งหมดของอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 1% บนชุดข้อมูล Chess

จากรูปด้านบนแสดงให้เห็นว่าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมีจำนวนตัวอย่างในโหนดคำตอบส่วนใหญ่เป็นไปในทางเดียวกันกับอัลกอริทึม ID3 แบบดั้งเดิม แม้ว่าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมีจำนวนตัวอย่างในโหนดคำตอบช่วงท้าย น้อยกว่าอัลกอริทึม ID3 แบบดั้งเดิม อย่างไรก็ตามจำนวนตัวอย่างในโหนดคำตอบช่วงท้ายของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ ไม่ได้น้อยลงจากจำนวนตัวอย่างในโหนดคำตอบช่วงท้ายของอัลกอริทึม ID3 แบบดั้งเดิมมากจนมีจำนวนตัวอย่างแค่ 1 หรือ 0 ตัวอย่าง ดังนั้นความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ จึงมีค่าใกล้เคียงกับความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิม เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 1% บนชุดข้อมูล Chess



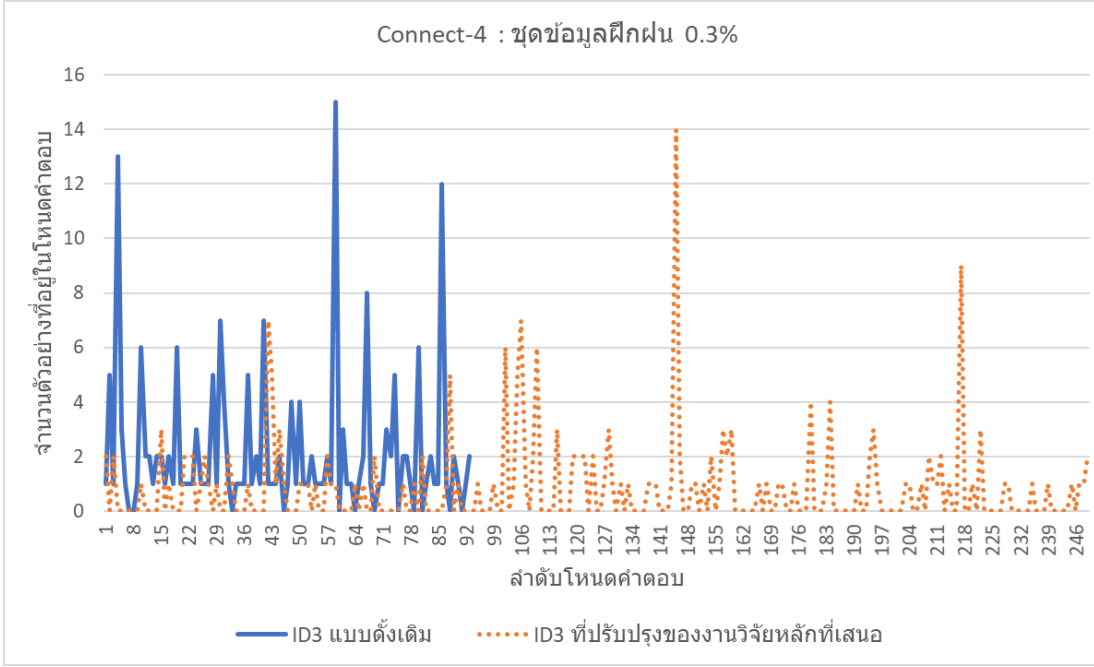
รูปที่ 6.3 จำนวนตัวอย่างที่อยู่ในโหนดคำตอบทั้งหมดของอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 0.3% บนชุดข้อมูล Chess

จากรูปด้านบนแสดงให้เห็นว่าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมีจำนวนตัวอย่างที่อยู่ในโหนดคำตอบเป็น 0 หรือ 1 เป็นจำนวนมากกว่าอัลกอริทึม ID3 แบบดั้งเดิม ดังนั้นความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ จึงมีค่าน้อยกว่าความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมอย่างเห็นได้ชัด เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 0.3% บนชุดข้อมูล Chess



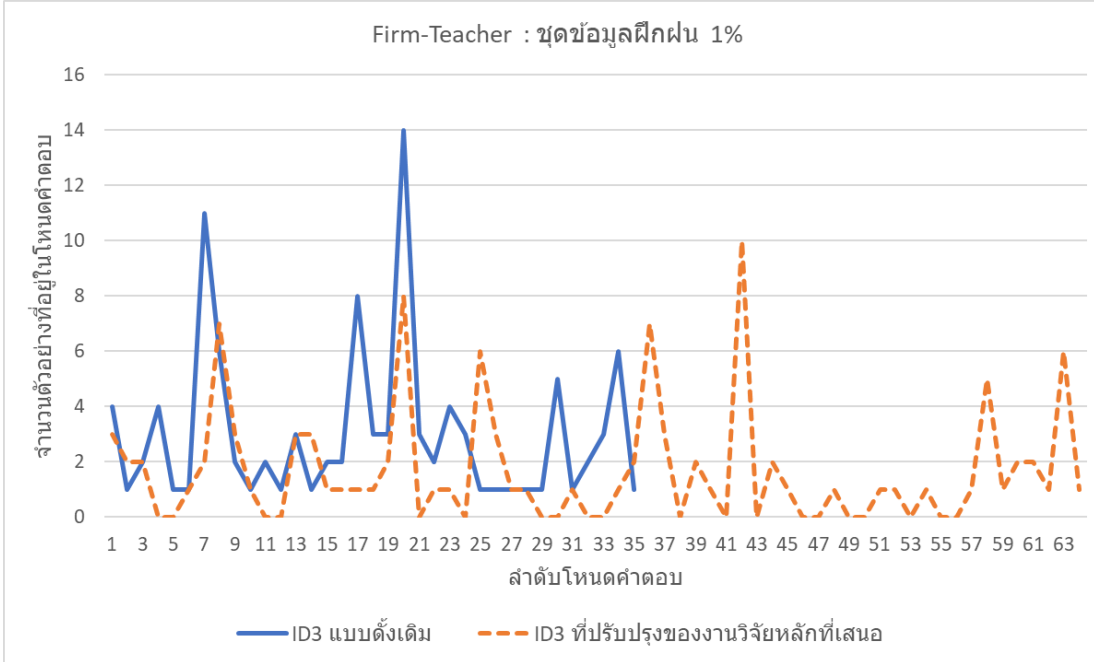
รูปที่ 6.4 จำนวนตัวอย่างที่อยู่ในโหนดคำตอบทั้งหมดของอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 1% บนชุดข้อมูล Connect-4

จากรูปด้านบนแสดงให้เห็นว่าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมีจำนวนตัวอย่างที่อยู่ในโหนดคำตอบเป็น 0 หรือ 1 เป็นจำนวนมากกว่าอัลกอริทึม ID3 แบบดั้งเดิม อย่างไรก็ตามความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ ยังคงมีค่ามากกว่าความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมอย่างเห็นได้ชัด เพราะถึงแม้ว่าการที่อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอจะมีจำนวนตัวอย่างที่อยู่ในโหนดคำตอบเป็น 0 หรือ 1 เป็นจำนวนมากกว่าอัลกอริทึม ID3 แบบดั้งเดิม แต่เวลาทำการทดสอบแล้ว ตัวอย่างในชุดข้อมูลทดสอบส่วนใหญ่อาจไม่ได้ถูกจำแนกในโหนดคำตอบที่มีจำนวนตัวอย่างเป็น 0 หรือ 1 ดังนั้นจึงมีโอกาสที่ความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ จะมีค่ามากกว่าความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมอย่างเห็นได้ชัด เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 1% บนชุดข้อมูล Connect-4



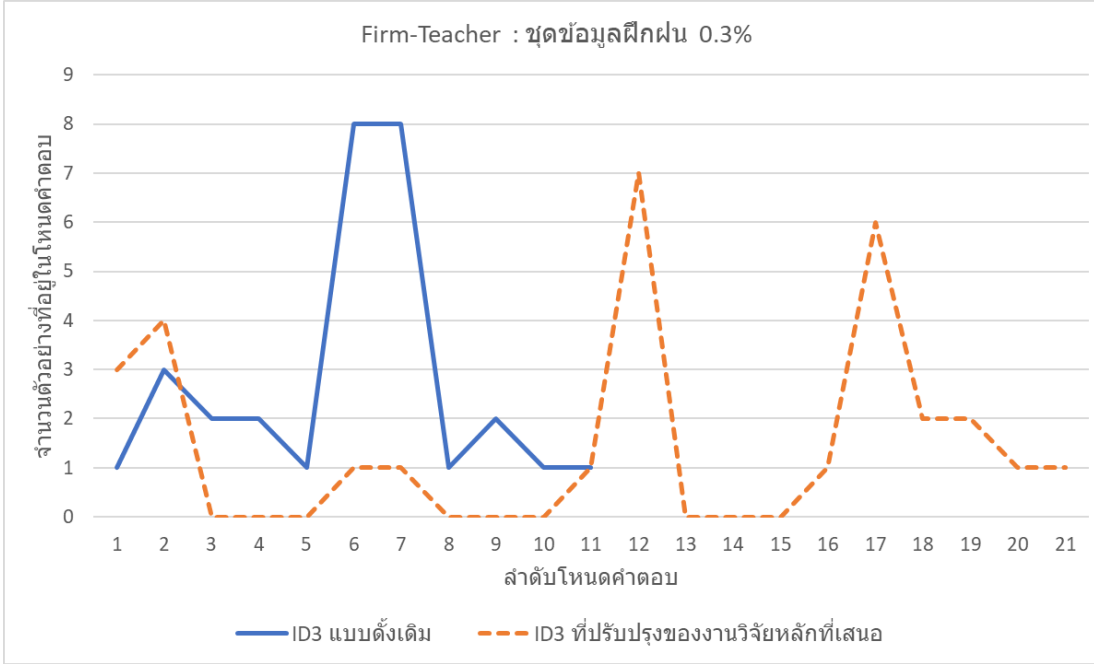
รูปที่ 6.5 จำนวนตัวอย่างที่อยู่ในโหนดคำตอบทั้งหมดของอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 0.3% บนชุดข้อมูล Connect-4

จากรูปด้านบนแสดงให้เห็นว่าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมีจำนวนตัวอย่างที่อยู่ในโหนดคำตอบเป็น 0 หรือ 1 เป็นจำนวนมากกว่าอัลกอริทึม ID3 แบบดั้งเดิม อย่างไรก็ตามความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ ยังคงมีค่ามากกว่าความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมอย่างเห็นได้ชัด เพราะถึงแม้ว่าการที่อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอจะมีจำนวนตัวอย่างที่อยู่ในโหนดคำตอบเป็น 0 หรือ 1 เป็นจำนวนมากกว่าอัลกอริทึม ID3 แบบดั้งเดิม แต่เวลาทำการทดสอบแล้ว ตัวอย่างในชุดข้อมูลทดสอบส่วนใหญ่อาจไม่ได้ถูกจำแนกในโหนดคำตอบที่มีจำนวนตัวอย่างเป็น 0 หรือ 1 ดังนั้นจึงมีโอกาสที่ความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ จะมีค่ามากกว่าความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมอย่างเห็นได้ชัด เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 0.3% บนชุดข้อมูล Connect-4



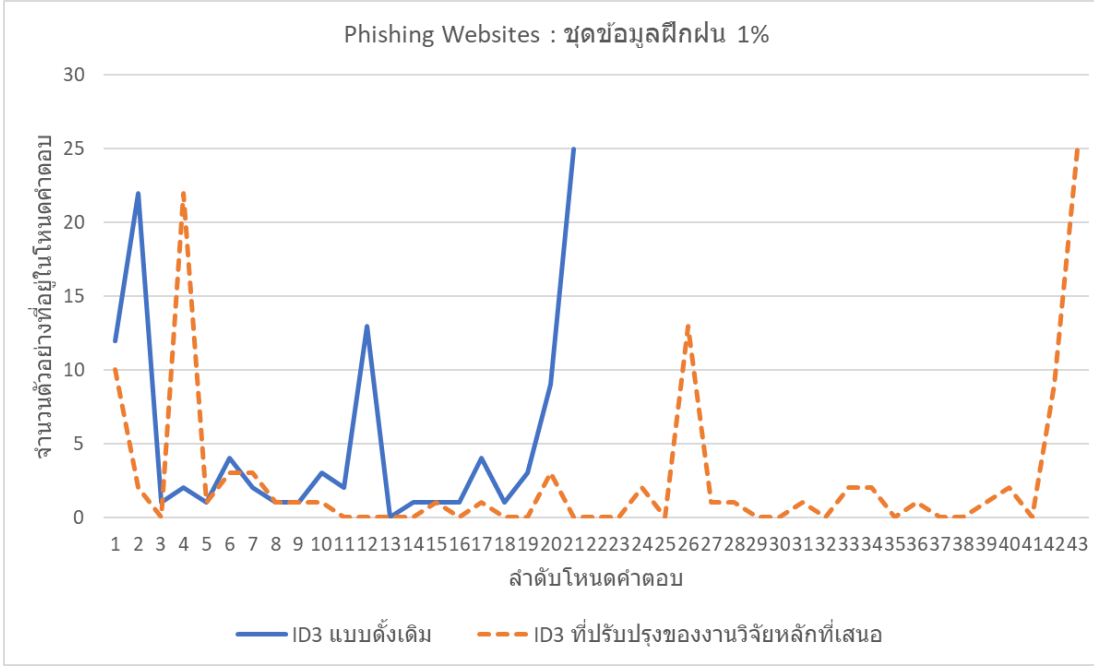
รูปที่ 6.6 จำนวนตัวอย่างที่อยู่ในโหนดคำตอบทั้งหมดของอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 1% บนชุดข้อมูล Firm-Teacher

จากรูปด้านบนแสดงให้เห็นว่าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมีจำนวนตัวอย่างที่อยู่ในโหนดคำตอบเป็น 0 หรือ 1 เป็นจำนวนมากกว่าอัลกอริทึม ID3 แบบดั้งเดิม อย่างไรก็ตามความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ ยังคงมีค่าใกล้เคียงกับความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิม เพราะถึงแม้ว่าการที่อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอจะมีจำนวนตัวอย่างที่อยู่ในโหนดคำตอบเป็น 0 หรือ 1 เป็นจำนวนมากกว่าอัลกอริทึม ID3 แบบดั้งเดิม แต่เวลาทำการทดสอบแล้ว ตัวอย่างในชุดข้อมูลทดสอบอาจไม่ได้ถูกจำแนกในโหนดคำตอบที่มีจำนวนตัวอย่างเป็น 0 หรือ 1 เป็นจำนวนที่มากพอที่จะส่งผลให้ความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอน้อยกว่าอัลกอริทึม ID3 แบบดั้งเดิมอย่างเห็นได้ชัด ดังนั้นความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ จึงมีค่าใกล้เคียงกันกับความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิม เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 1% บนชุดข้อมูล Firm-Teacher



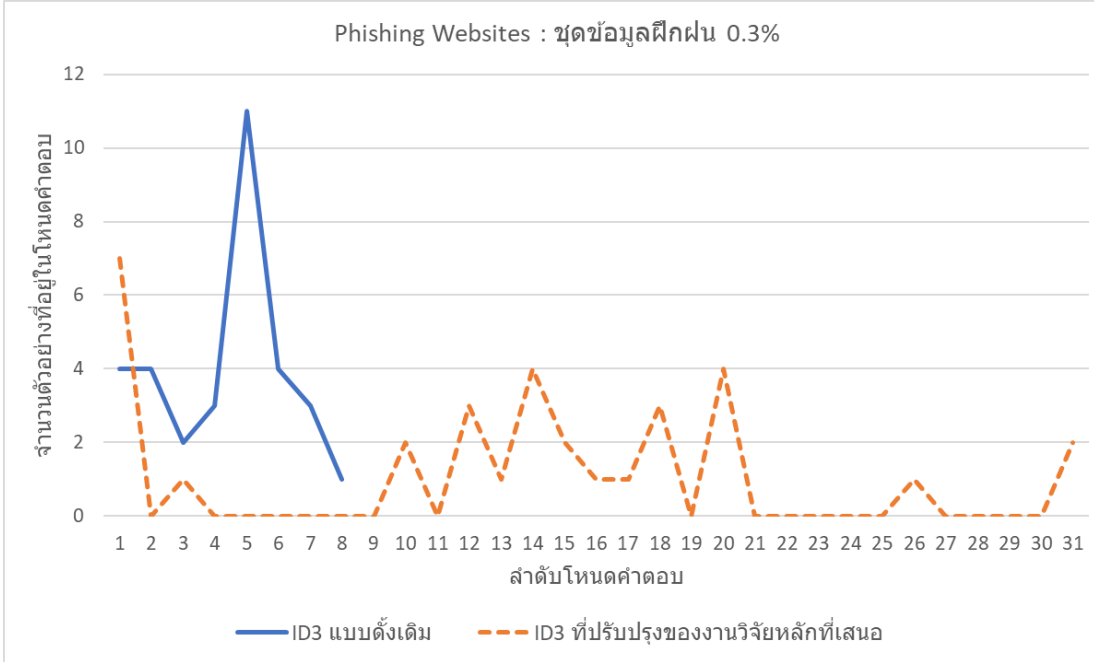
รูปที่ 6.7 จำนวนตัวอย่างที่อยู่ในโหนดคำตอบทั้งหมดของอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 0.3% บนชุดข้อมูล Firm-Teacher

จากรูปด้านบนแสดงให้เห็นว่าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมีจำนวนตัวอย่างที่อยู่ในโหนดคำตอบเป็น 0 หรือ 1 เป็นจำนวนมากกว่าอัลกอริทึม ID3 แบบดั้งเดิม ดังนั้นความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ จึงมีค่าน้อยกว่าความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมอย่างเห็นได้ชัด เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 0.3% บนชุดข้อมูล Firm-Teacher



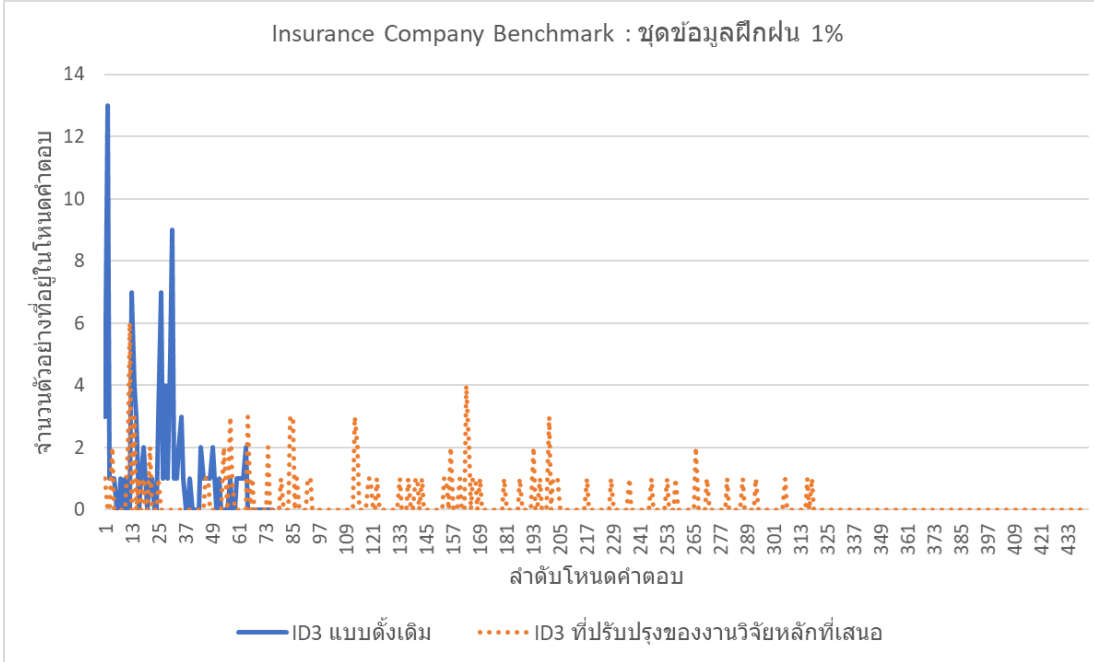
รูปที่ 6.8 จำนวนตัวอย่างที่อยู่ในโหนดคำตอบทั้งหมดของอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 1% บนชุดข้อมูล Phishing Websites

จากรูปด้านบนแสดงให้เห็นว่าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมีจำนวนตัวอย่างที่อยู่ในโหนดคำตอบเป็น 0 หรือ 1 เป็นจำนวนมากกว่าอัลกอริทึม ID3 แบบดั้งเดิม อย่างไรก็ตามความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ ยังคงมีค่าใกล้เคียงกับความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิม เพราะถึงแม้ว่าการที่อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอจะมีจำนวนตัวอย่างที่อยู่ในโหนดคำตอบเป็น 0 หรือ 1 เป็นจำนวนมากกว่าอัลกอริทึม ID3 แบบดั้งเดิม แต่เวลาทำการทดสอบแล้ว ตัวอย่างในชุดข้อมูลทดสอบอาจไม่ได้ถูกจำแนกในโหนดคำตอบที่มีจำนวนตัวอย่างเป็น 0 หรือ 1 เป็นจำนวนที่มากพอที่จะส่งผลให้ความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอน้อยกว่าอัลกอริทึม ID3 แบบดั้งเดิมอย่างเห็นได้ชัด ดังนั้นความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ จึงมีค่าใกล้เคียงกันกับความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิม เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 1% บนชุดข้อมูล Phishing Websites



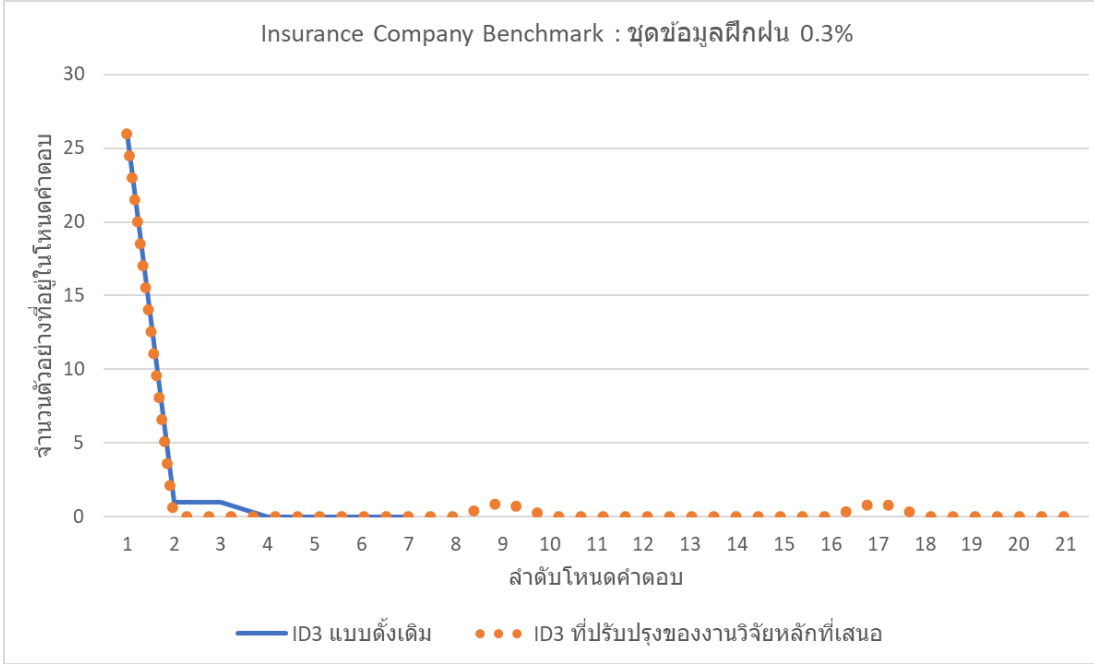
รูปที่ 6.9 จำนวนตัวอย่างที่อยู่ในโหนดคำตอบทั้งหมดของอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 0.3% บนชุดข้อมูล Phishing Websites

จากรูปด้านบนแสดงให้เห็นว่าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมีจำนวนตัวอย่างที่อยู่ในโหนดคำตอบเป็น 0 หรือ 1 เป็นจำนวนมากกว่าอัลกอริทึม ID3 แบบดั้งเดิม ดังนั้นความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ จึงมีค่าน้อยกว่าความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมอย่างเห็นได้ชัด เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 0.3% บนชุดข้อมูล Phishing Websites



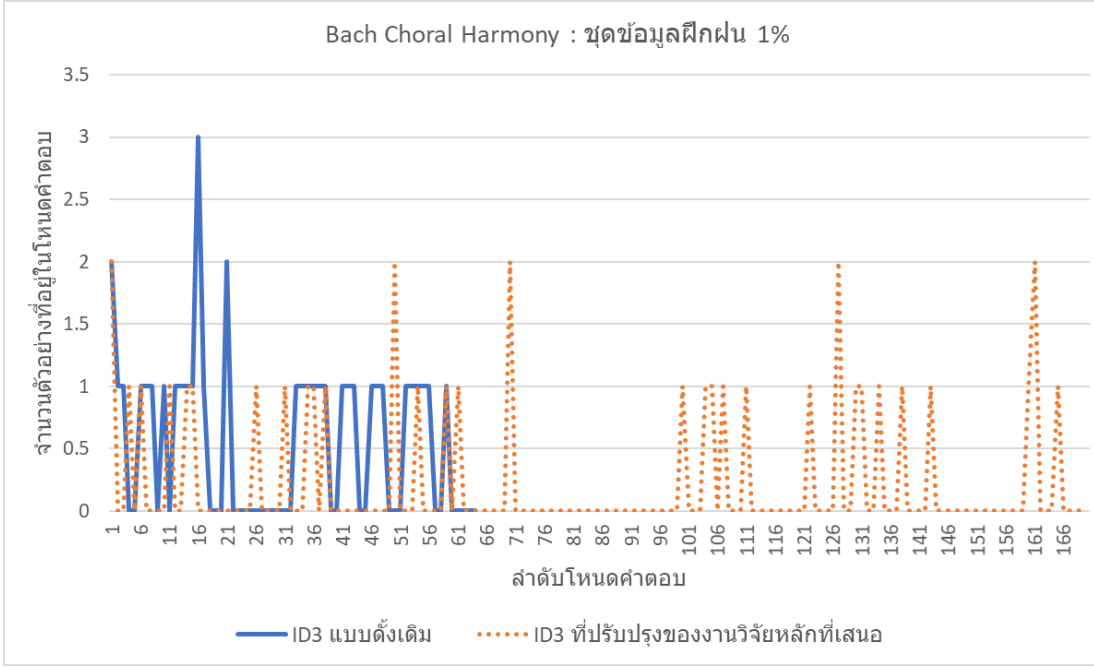
รูปที่ 6.10 จำนวนตัวอย่างที่อยู่ในโหนดคำตอบทั้งหมดของอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 1% บนชุดข้อมูล Insurance Company Benchmark

จากรูปด้านบนแสดงให้เห็นว่าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมีจำนวนตัวอย่างที่อยู่ในโหนดคำตอบเป็น 0 หรือ 1 เป็นจำนวนมากกว่าอัลกอริทึม ID3 แบบดั้งเดิม อย่างไรก็ตามความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ ยังคงมีค่าใกล้เคียงกับความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิม เพราะถึงแม้ว่าการที่อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอจะมีจำนวนตัวอย่างที่อยู่ในโหนดคำตอบเป็น 0 หรือ 1 เป็นจำนวนมากกว่าอัลกอริทึม ID3 แบบดั้งเดิม แต่เวลาทำการทดสอบแล้ว ตัวอย่างในชุดข้อมูลทดสอบอาจไม่ได้ถูกจำแนกในโหนดคำตอบที่มีจำนวนตัวอย่างเป็น 0 หรือ 1 เป็นจำนวนที่มากพอที่จะส่งผลให้ความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอน้อยกว่าอัลกอริทึม ID3 แบบดั้งเดิมอย่างเห็นได้ชัด ดังนั้นความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ จึงมีค่าใกล้เคียงกันกับความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิม เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 1% บนชุดข้อมูล Insurance Company Benchmark



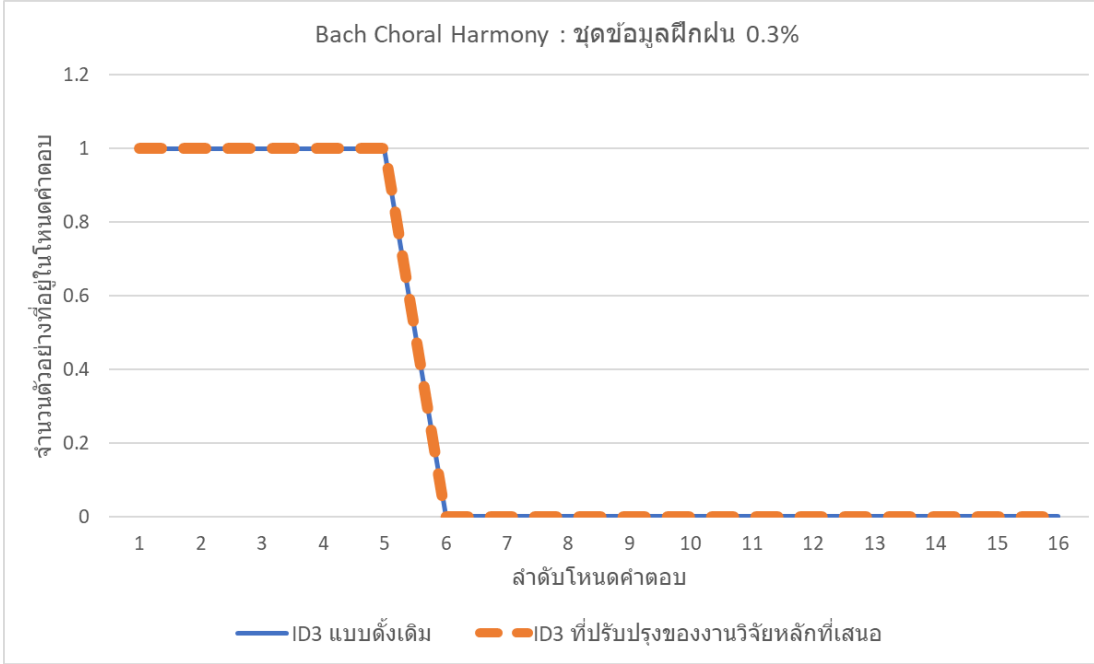
รูปที่ 6.11 จำนวนตัวอย่างที่อยู่ในโหนดคำตอบทั้งหมดของอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 0.3% บนชุดข้อมูล Insurance Company Benchmark

จากรูปด้านบนแสดงให้เห็นว่าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมีจำนวนตัวอย่างที่อยู่ในโหนดคำตอบเป็นไปในทางเดียวกันกับอัลกอริทึม ID3 แบบดั้งเดิม ดังนั้นความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ จึงมีค่าใกล้เคียงกับความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิม เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 0.3% บนชุดข้อมูล Insurance Company Benchmark



รูปที่ 6.12 จำนวนตัวอย่างที่อยู่ในโหนดคำตอบทั้งหมดของอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 1% บนชุดข้อมูล Bach Choral Harmony

จากรูปด้านบนแสดงให้เห็นว่าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมีจำนวนตัวอย่างที่อยู่ในโหนดคำตอบเป็น 0 หรือ 1 เป็นจำนวนมากกว่าอัลกอริทึม ID3 แบบดั้งเดิม อย่างไรก็ตามความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ ยังคงมีค่ามากกว่าความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมอย่างเห็นได้ชัด เพราะถึงแม้ว่าการที่อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอจะมีจำนวนตัวอย่างที่อยู่ในโหนดคำตอบเป็น 0 หรือ 1 เป็นจำนวนมากกว่าอัลกอริทึม ID3 แบบดั้งเดิม แต่เวลาทำการทดสอบแล้ว ตัวอย่างในชุดข้อมูลทดสอบส่วนใหญ่อาจไม่ได้ถูกจำแนกในโหนดคำตอบที่มีจำนวนตัวอย่างเป็น 0 หรือ 1 ดังนั้นจึงมีโอกาสที่ความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ จะมีค่ามากกว่าความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมอย่างเห็นได้ชัด เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 1% บนชุดข้อมูล Bach Choral Harmony



รูปที่ 6.13 จำนวนตัวอย่างที่อยู่ในโหนดคำตอบทั้งหมดของอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 0.3% บนชุดข้อมูล Bach Choral Harmony

จากรูปด้านบนแสดงให้เห็นว่าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมีจำนวนตัวอย่างที่อยู่ในโหนดคำตอบเป็นไปในทางเดียวกันกับอัลกอริทึม ID3 แบบดั้งเดิม ดังนั้นความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ จึงมีค่าใกล้เคียงกับความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิม เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝน 0.3% บนชุดข้อมูล Bach Choral Harmony

6.5.3.2 ค่าสูงสุดของความแม่นยำ

ตารางที่ 6.97 ค่าสูงสุดของความแม่นยำระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝนที่น้อย

ชุดข้อมูล	ค่าสูงสุดของความแม่นยำ	
	ID3 แบบดั้งเดิม	ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ
Chess	98.13%	97.66%
Connect-4	70.63%	72.22%
Firm-Teacher	69.62%	70.48%
Phishing Websites	93.83%	94.48%
Insurance Company Benchmark	90.50%	92.43%
Bach Choral Harmony	65.41%	66.22%

จากตารางข้างบน ค่าสูงสุดของความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าน้อยกว่าค่าสูงสุดของความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมเพียงเล็กน้อยในชุดข้อมูล Chess ส่วนในชุดข้อมูล Connect-4 และ Insurance Company Benchmark มีค่าสูงสุดของความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมากกว่าค่าสูงสุดของความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมอย่างเห็นได้ชัด สำหรับชุดข้อมูลที่เหลือคือ Firm-Teacher, Phishing Websites และ Bach Choral Harmony นั้นมีค่าสูงสุดของความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมากกว่าค่าสูงสุดของความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมเพียงเล็กน้อย ซึ่งสามารถสรุปได้ว่า เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝนที่น้อย ค่าสูงสุดของความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าใกล้เคียงกับค่าสูงสุดของความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูลส่วนใหญ่

6.5.3.3 ค่าต่ำสุดของความแม่นยำ

ตารางที่ 6.98 ค่าต่ำสุดของความแม่นยำระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝนที่น้อย

ชุดข้อมูล	ค่าต่ำสุดของความแม่นยำ	
	ID3 แบบดั้งเดิม	ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ
Chess	95.88%	95.07%
Connect-4	68.88%	70.39%
Firm-Teacher	67.15%	67.15%
Phishing Websites	92.54%	93.01%
Insurance Company Benchmark	89.13%	91.52%
Bach Choral Harmony	62.18%	62.00%

จากตารางข้างบน ค่าต่ำสุดของความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าน้อยกว่าค่าต่ำสุดของความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมเพียงเล็กน้อยในชุดข้อมูล Chess และ Bach Choral Harmony ส่วนในชุดข้อมูล Connect-4 และ Insurance Company Benchmark มีค่าต่ำสุดของความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมากกว่าค่าต่ำสุดของความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมอย่างเห็นได้ชัด ในด้านของชุดข้อมูล Phishing Websites นั้นมีค่าต่ำสุดของความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมากกว่าค่าต่ำสุดของความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมเพียงเล็กน้อย สำหรับชุดข้อมูลที่เหลือคือ Firm-Teacher นั้นมีค่าต่ำสุดของความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอเท่ากับค่าต่ำสุดของความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิม ดังนั้นจึงสามารถสรุปได้ว่า เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝนที่น้อย ค่าต่ำสุดของความแม่นยำของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าใกล้เคียงกับค่าต่ำสุดของความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูลส่วนใหญ่

6.5.3.4 ความลึกสูงสุดโดยเฉลี่ย

ตารางที่ 6.99 ความลึกสูงสุดโดยเฉลี่ยระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝนที่น้อย

ชุดข้อมูล	ความลึกสูงสุดโดยเฉลี่ย	
	ID3 แบบดั้งเดิม	ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ
Chess	10.67	6.67
Connect-4	16.67	9.67
Firm-Teacher	13.33	8.33
Phishing Websites	30	14.67
Insurance Company Benchmark	85	29.83
Bach Choral Harmony	14	8.5

จากตารางข้างบน ความลึกสูงสุดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าน้อยกว่าความลึกสูงสุดโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูลทั้งหมดอย่างเห็นได้ชัด ซึ่งสามารถสรุปได้ว่า เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝนที่น้อย อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอสามารถลดความลึกสูงสุดโดยเฉลี่ยจากความลึกสูงสุดโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูลทั้งหมดได้อย่างเห็นได้ชัด

6.5.3.5 ค่าสูงสุดของความลึกสูงสุด

ตารางที่ 6.100 ค่าสูงสุดของความลึกสูงสุดระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝนที่น้อย

ชุดข้อมูล	ค่าสูงสุดของความลึกสูงสุด	
	ID3 แบบดั้งเดิม	ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ
Chess	12	8
Connect-4	19	10
Firm-Teacher	14	9
Phishing Websites	30	15
Insurance Company Benchmark	85	30
Bach Choral Harmony	14	9

จากตารางข้างบน ค่าสูงสุดของความลึกสูงสุดของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าน้อยกว่าค่าสูงสุดของความลึกสูงสุดของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูลทั้งหมดอย่างเห็นได้ชัด

6.5.3.6 ค่าต่ำสุดของความลึกสูงสุด

ตารางที่ 6.101 ค่าต่ำสุดของความลึกสูงสุดระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เมื่อใช้อัตราส่วนของชุดข้อมูลฝึกฝนที่น้อย

ชุดข้อมูล	ค่าต่ำสุดของความลึกสูงสุด	
	ID3 แบบดั้งเดิม	ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ
Chess	10	6
Connect-4	16	9
Firm-Teacher	13	8
Phishing Websites	30	14
Insurance Company Benchmark	85	29
Bach Choral Harmony	14	8

จากตารางข้างบน ค่าต่ำสุดของความลึกสูงสุดของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าน้อยกว่าค่าต่ำสุดของความลึกสูงสุดของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูลทั้งหมดอย่างเห็นได้ชัด

บทที่ 7
บทสรุปและข้อเสนอแนะ

7.1 สรุป

อัลกอริทึม ID3 เป็นอัลกอริทึมการเรียนรู้ (Learning algorithm) สำหรับใช้ในการสร้างต้นไม้ตัดสินใจ ซึ่งเป็นอัลกอริทึมที่นิยมใช้กันมากในกลุ่มงานด้านการจำแนกประเภท (Classification) ของข้อมูลที่สนใจ ซึ่งงานวิจัยแรก [5] ของวิทยานิพนธ์นี้ได้ทำการปรับปรุงอัลกอริทึม ID3 เพื่อจัดการกับปัญหาแอตทริบิวต์ที่มีความสำคัญเท่ากัน ถึงแม้ว่าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกจะสามารถจัดการกับปัญหาแอตทริบิวต์ที่มีความสำคัญเท่ากันและลดความลึกสูงสุดของต้นไม้ตัดสินใจได้ ในขณะที่ความแม่นยำในการจำแนกยังคงระดับเดิม แต่ก็มีปัญหาความตึงของวิธีการเลือกแอตทริบิวต์ซึ่งส่งผลต่อความสามารถในการลดความลึกสูงสุดของต้นไม้ตัดสินใจที่สร้าง โดยเฉพาะอย่างยิ่งเมื่อใช้อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกกับชุดข้อมูลที่พบแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเป็นอันดับที่ 1 ต่อการพิจารณาเลือกแอตทริบิวต์โดยเฉลี่ยอยู่ในจำนวนที่น้อย ซึ่งจะมีผลทำให้ความสามารถในการลดความลึกสูงสุดของต้นไม้ตัดสินใจที่สร้างจากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกมีน้อย ดังนั้นวิทยานิพนธ์นี้จึงเสนอวิธีการปรับปรุงอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกเพื่อจัดการกับปัญหาความตึงของวิธีการเลือกแอตทริบิวต์ และเพื่อลดความลึกสูงสุดของต้นไม้ตัดสินใจที่ได้ให้น้อยลงยิ่งขึ้น ในขณะที่ความแม่นยำในการจำแนกยังคงระดับเดิม

ผลการทดลองแสดงให้เห็นว่า ความลึกสูงสุดโดยเฉลี่ย, เวลาในการทดสอบโดยเฉลี่ยและความลึกโดยเฉลี่ยในการจำแนกของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอน้อยกว่าความลึกสูงสุดโดยเฉลี่ย, เวลาในการทดสอบโดยเฉลี่ยและความลึกโดยเฉลี่ยในการจำแนกของอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกอย่างเห็นได้ชัดในชุดข้อมูลทั้งหมด ในขณะที่ความแม่นยำโดยเฉลี่ยมีค่าที่ใกล้เคียงกันทั้ง 3 อัลกอริทึม อย่างไรก็ตามอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมีจำนวนโหนดโดยเฉลี่ยที่มากกว่าจำนวนโหนดโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกอย่างเห็นได้ชัดในชุดข้อมูลส่วนใหญ่ ซึ่งจุดนี้ถือว่าเป็นข้อเสียที่มีอย่างแน่นอนในอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ เนื่องจากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอพัฒนาต่อยอดมาจากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก ซึ่งมีข้อเสียด้านความต้องการด้านพื้นที่ (Space requirement) ที่ใช้ในการเก็บรายละเอียดของต้นไม้ตัดสินใจที่สร้างอยู่แล้ว ข้อจำกัดนี้จึงหลีกเลี่ยงไม่ได้ในอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ ส่วนเรื่องเวลาในการฝึกฝนโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอนั้น อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอใช้เวลาในการฝึกฝนโดยเฉลี่ยน้อยกว่าเวลาในการฝึกฝนโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกในชุดข้อมูล

ทั้งหมด และเมื่อเปรียบเทียบเวลาในการฝึกฝนโดยเฉลี่ยระหว่างอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอและอัลกอริทึม ID3 แบบดั้งเดิม พบว่าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอใช้เวลาในการฝึกฝนโดยเฉลี่ยน้อยกว่าอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูล 3 ชุดข้อมูลจากทั้งหมด 6 ชุดข้อมูล ในขณะที่อัลกอริทึม ID3 แบบดั้งเดิมใช้เวลาในการฝึกฝนโดยเฉลี่ยมากกว่าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอเพียงเล็กน้อยใน 3 ชุดข้อมูล แต่เมื่อพิจารณาโดยรวมแล้ว ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอใช้เวลาในการฝึกฝนโดยเฉลี่ย มากกว่า ID3 แบบดั้งเดิม ดังนั้นจึงสามารถสรุปได้ว่าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอใช้เวลาในการฝึกฝนโดยเฉลี่ยมากกว่าเวลาในการฝึกฝนโดยเฉลี่ยของอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูลส่วนมาก

สำหรับปัญหาความตึงของวิธีการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกนั้น พบว่าอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมีจำนวนแอตทริบิวต์ที่ถูกเลือกไปอยู่ในโหนดตัดสินใจเดียวกันต่อโหนดโดยเฉลี่ยมากกว่าจำนวนแอตทริบิวต์ที่ถูกเลือกไปอยู่ในโหนดตัดสินใจเดียวกันต่อโหนดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกอย่างเห็นได้ชัด ดังนั้นอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอสามารถจัดการปัญหาความตึงของวิธีการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกได้อย่างมีประสิทธิภาพและสามารถลดความลึกสูงสุดโดยเฉลี่ยจากความลึกสูงสุดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกให้น้อยลงกว่าเดิมอย่างเห็นได้ชัด ในขณะที่ความแม่นยำโดยเฉลี่ยยังคงใกล้เคียงกันเหมือนเดิม อย่างไรก็ตามอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอยังคงมีข้อจำกัดที่คล้ายกับข้อจำกัดของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก ซึ่งก็คือชุดข้อมูลที่ใช้ในการสร้างต้นไม้ตัดสินใจและความต้องการด้านพื้นที่ โดยความต้องการด้านพื้นที่ของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอนั้น จะมากกว่าความต้องการด้านพื้นที่ของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก ทั้งนี้เนื่องจากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอสามารถพบและเลือกแอตทริบิวต์เข้ามาอยู่ในโหนดตัดสินใจเดียวกันได้มากขึ้นเมื่อเทียบกับอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก ซึ่งทำให้จำนวนโหนดที่สร้างโดยรวมของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอเพิ่มมากขึ้น ดังจะเห็นได้จากผลการทดลองด้านจำนวนโหนดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ มีค่าที่เพิ่มขึ้นจากจำนวนโหนดโดยเฉลี่ยของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรกอย่างชัดเจนภายในชุดข้อมูลส่วนใหญ่ เพราะฉะนั้นความต้องการด้านพื้นที่ของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอจึงมากกว่าความต้องการด้านพื้นที่ของอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยแรก

7.2 ข้อเสนอแนะ

อัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอสามารถนำไปปรับปรุงพัฒนาต่อได้ โดยสามารถทำการปรับปรุงในส่วนของจำนวนโหนดที่สร้าง ซึ่งอาจจะทำการยุบโหนดคำตอบที่ไม่จำเป็นที่อยู่ในระดับความลึกที่มากๆ เพราะตัวอย่างที่นำมาจำแนกในต้นไม้ตัดสินใจที่สร้างจากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ ส่วนใหญ่แล้วถูกจำแนกในระดับความลึกที่น้อยกว่าความลึกสูงสุดของต้นไม้ตัดสินใจ โหนดคำตอบใดที่สร้างมาแล้วแต่ไม่ได้ใช้งานเลย คือไม่มีตัวอย่างที่ถูกจำแนกภายในโหนดคำตอบ โหนดคำตอบเหล่านั้นก็ควรที่จะถูกยุบ ทั้งนี้การที่จะยุบโหนดคำตอบที่มีลักษณะดังกล่าวได้จะต้องอาศัยข้อมูลทางสถิติเกี่ยวกับความลึกโดยเฉลี่ยในการจำแนกของต้นไม้ตัดสินใจที่สร้างจากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอ ผลจากการยุบโหนดคำตอบที่มีลักษณะดังกล่าวจะทำให้ความต้องการด้านพื้นที่ ที่ใช้ในการเก็บรายละเอียดของต้นไม้ตัดสินใจที่สร้างจากอัลกอริทึม ID3 ที่ปรับปรุงของงานวิจัยหลักที่เสนอมีขนาดที่ลดลงได้

เอกสารอ้างอิง

[1] J.R. Quinlan. "Induction of decision trees" Mach. Leaning, vol. 1, 1986. pp. 81-106.

[2] J. Liu and N. Li. "Optimized ID3 algorithm based on attribute importance and convex function" IEEE Int. Symposium on IT in Medicine and Education, Cuangzhou, China, 2011. pp. 136-139.

[3] H. Luo, Y. Chen, and W. Zhang. "An Improved ID3 Algorithm Based on Attribute Importance-Weighted" 2nd Int. Workshop on Database Technology and Applications, Wuhan, China, 2010.

[4] Q. Liu, D. Hu, and Q. Yan. "Decision tree algorithm based on average Euclidean distance" 2nd Int. Conf. on Future Computer and Communication, Wuha, China, 2010. pp. 507-511.

[5] S. Kraidech and K. Jearanaitanakij. "Improving ID3 Algorithm by Combining Values from Equally Important Attributes" 21st Int. Computer Science and Engineering Conf., Bangkok, Thailand, 2017. pp. 102-105.

[6] D. Dua and E. K. Taniskidou. "UCI Machine Learning Repository." [online]. Available: http://archive.ics.uci.edu/ml/datasets.html. 2017.

ภาคผนวก ก.

งานวิจัยที่ได้รับการตีพิมพ์

ก.1 Improving ID3 Algorithm by Combining Values from Equally Important Attributes, ICSEC 2017

The 21st International Computer Science and Engineering Conference 2017 , November 15-18, 2017 Bangkok, Thailand.

Improving ID3 Algorithm by Combining Values from Equally Important Attributes

Suratchanan Kraidech, Kietikul Jearanaitanakij
Department of Computer Engineering, Faculty of Engineering
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
Email: 55011362@kmitl.ac.th, kietikul.je@kmitl.ac.th

***Abstract*—ID3 is a well-known algorithm which is used in the classification task of the decision tree learning. Although a lot of research provides improvements on the traditional ID3 algorithm with various strategies, no attempt was made on ID3 to address the problem when there are more than one attribute that can be placed at a particular node, i.e. those attributes are equally important. This paper proposes a new variation of ID3 to combine equally important attributes into a single node of the decision tree classification. The Connect-4 dataset from UCI is used in our experiment since the dataset contains many attributes and instances which can easily encounter the equally important attribute problem. The experimental results show that our proposed method significantly reduces the average depth of the decision tree generated by ID3 algorithm while the average accuracy rate is still preserved.**

Keywords—Decision tree; ID3 algorithm; Classification; Attribute combining

I. INTRODUCTION

Machine learning has an important role in data analysis of a wide variety of fields such as financial, medical, government, transportation and others. One of the popular models is the decision tree learning which is simple and easy to implement. Among many versions of decision tree algorithm, ID3 is the most well-known method which is proposed by Quinlan in 1986 [1].

There are many attempts to improve ID3 algorithm. Shichao et al. [2] found a new strategy for attributes selection in the ID3 algorithm with a trade-off method between attributes' information and cost-sensitive learning to increasing the classification's accuracy. This research is designed in important attributes selection with average gain which reduces disadvantage of attributes selection by considering only information gain. Baoshi et al. [3] proposed the attributes selection in the ID3 algorithm with rough set theory by considering information gain and relationship between condition attributes and decision attributes. However, the experiment was conducted on the low-complexity dataset (165 instances and 13 attributes.) It is questionable whether it can be applied to more complex datasets. Hongwu et al. [4] proposed another way to improve performance in ID3 algorithm by reduce the complexity of information gain calculation using Taylor's theorem and attribute similarity theorem. The attribute importance-weighted is added into the modified information gain calculation. The result shows that the improved algorithm can significantly improve accuracy and reduce running time. This improved algorithm also works well in complex datasets.

One issue that has been overlooked from above research is the situation when we choose the most important attribute for a particular node and there are at least two most important attributes, i.e., they are equally important. The traditional ID3 algorithm randomly chooses one from a set equally important attributes. However, this seems not to be a good idea since the classification tree may end up with a lengthy depth which may increase the running time. For example, suppose some dataset is training with the traditional ID3 algorithm, when the algorithm chooses the most important attribute for a particular node, it finds that attribute A and B are candidates of this node. If it chooses A, the decision tree will have depth 10. On the other hand, if it chooses B, the resulting tree will have depth 6. Therefore, the traditional ID3 should select B as the attribute for the current decision node since it produces a smaller depth value.

The purpose of this paper is to propose a novel approach to optimize an average depth by expanding the combination of values of equally important attributes in a single node, while the accuracy rate of the ID3 algorithm is still preserved. The experimental results on the Connect-4 dataset from UCI indicate the significant improvement of the proposed algorithm over the traditional ID3 algorithm. The rest of this paper is organized as the following sections. Section II describes a brief background about a traditional ID3 algorithm. Section III introduces a problem statement along with the proposed method. In section IV, dataset, experimental conditions, results and comparisons are provided. Lastly, discussion and future work are given in section V.

II. THE TRADITIONAL ID3 ALGORITHM

ID3 is the most popular algorithm for generating the decision tree from a given dataset. The conventional ID3 algorithm [6] is briefed below along with the flowchart in Fig. 1.

978-1-5386-0787-9/17/$31.00 ©2017 IEEE

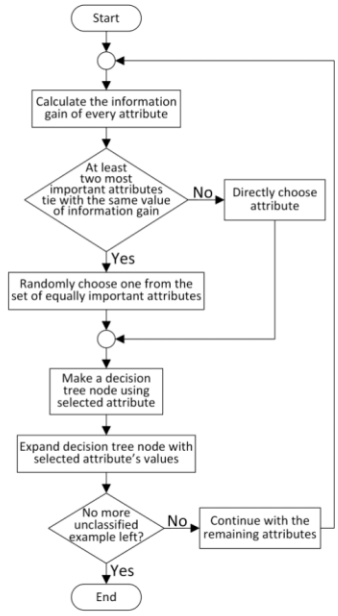


Fig. 1. Flowchart of the traditional ID3 algorithm

Step 1: Calculate the information gain of every attribute by using Eq. (1) and (2).

$$IG(A) = -\sum_{x \in X} p(x) \log_2 p(x) - Remainder(A) \qquad (1)$$

Where

IG(A) is the information gain of attribute A,

X is the set of classes in data,

p(x) is the proportion of the number of elements in class X to the number of elements in data.

The formula for the second term of Eq. (1) is listed in Eq. (2):

$$Remainder(A) = \sum_{v \in V} \left\{ - p(v) \cdot \sum_{x \in X} p(x_v) \log_2 p(x_v) \right\} \qquad (2)$$

Where

V is the set of values in A,

p(v) is the proportion of the number of elements of value in A to the number of elements of all values in A,

$p(x_v)$ is the proportion of the number of elements of classes of value in A to the number of elements of all classes of value in A.

Step 2: Select the attribute that have the maximum value of information gain. If there are at least two attributes tie with the same information gain at maximum value, randomly choose one from them.

Step 3: Make a decision tree node by using the attribute selected from step 2.

Step 4: Expand decision tree node with the selected attribute's values and repeat steps 1 to 4 using the remaining attributes until there is no unclassified example left.

III. THE PROPOSED ALGORITHM

We modify ID3 algorithm to handle the situation when the algorithm chooses the most important attribute for the current node and there are at least two attributes that can be chosen, i.e. those attributes are equally important. The traditional ID3 algorithm will randomly select one attribute from these candidates. The resulting classification tree may contain unnecessary levels of depth which increases the running time.

Conventionally, the number of branches of the decision node in ID3 algorithm is equal to the number of values of the attribute. We propose a new idea that the number of branches of the decision node is equal to the number of tuple which can be determined from the Cartesian product between candidate attribute's values. For example, suppose that attributes A and B are our candidates for the current node. An attribute A has three values and an attribute B has two values. By combining values from both attributes, the number of branches is equal to 6 (3x2). Obviously, if we have more candidate attributes, the number of branches will increase. The following Fig. 2 shows the decision node with the combined attributes {A, B}. Denote that SubT1 and SubT2 are two subtrees. L1 - L4 are leaf nodes.

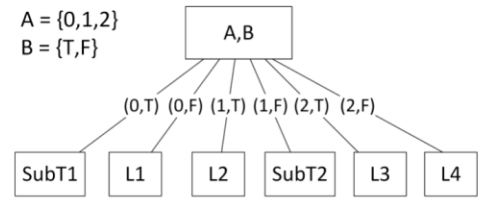


Fig. 2. Decision node with combined attributes {A,B}

The proposed algorithm is described below along with the flowchart in Fig. 3.

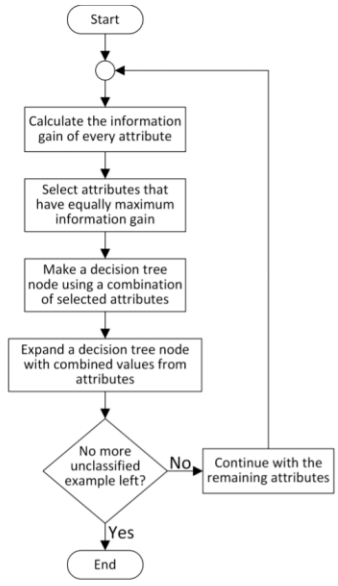


Fig. 3. Flowchart of the proposed ID3 algorithm

Step 1: Calculate the information gain of every attribute by using Eq. (1) and (2).

Step 2: Select most important attributes which tie with the value of information gain.

Step 3: Make a decision tree node by using a combination of the selected attributes from step 2.

Step 4: Expand a decision tree node by using the combination values of the combined attributes and repeat from steps 1 to 4 using the remaining attributes until there is no unclassified example left.

IV. Experiments

A. Dataset

The Connect-4 dataset consists of 42 attributes which represent the number of positions in the game. There are 67,557 instances and each instance is a game playing between two players in 8-ply which neither player has won yet. The next moving is not a compulsion. The number of classes is 3, i.e. {win, loss, draw}, which predicts from the game's theoretical value. The number of values for each attribute is 3, consisting of {x, o, b}, where x is a position which the player x has taken, o is a position which the player o has taken, b is a position which neither player x nor player o has taken. We choose the Connect-4 dataset for the experiment because it is the only one dataset in UCI repository that contains the equally important attributes when classifying with the ID3 algorithm.

B. Experimental conditions and Results

We conduct the experiment to compare the results between the conventional ID3 algorithm and the proposed ID3 improvement. Those experimental results consist of depth, accuracy, training time, testing time, total number of nodes, number of expanded nodes (i.e. non-leaf nodes) and number of expanded nodes that encounter equally important attributes problem. Before starting the experiment, the Connect-4 dataset is randomly separated into two parts, i.e. training set and testing set. The proportion of separation is 50 percent and the number of instances for both sets is balanced by the number of classifying classes. We repeat the experiment 1000 times on the same set of {train set/ test set} for both algorithms. For simplicity, the number of candidate attributes in the proposed algorithm is limited to 2. In addition, values of these candidate attributes must be discrete. In order to measure the testing time, we filter instances in the dataset to have only instances that the traditional ID3 classified them around the maximum depth. Afterward, we duplicate those instances to some large amounts so that the time difference between both algorithms is noticeable. The following Table I shows the comparison of the experiment between traditional ID3 and proposed ID3 improvement.

TABLE I. The experiment 1000 times between the traditional ID3 and the proposed ID3

Measurement	Algorithm	
	Traditional ID3	***Proposed ID3***
Min Depth	18	18
Max Depth	39	25
Avg. Depth	23.97	18.51
Avg. Accuracy	72.95	73.53
Avg. Number of all nodes	22562	33250
Avg. Number of non-leaf nodes	7520	6906
Avg. Number of non-leaf nodes that encounter equally important attributes problem	2459	2088
Avg. Training time	3.0903 sec.	3.1552 sec.
Avg. Testing time	0.0133 sec.	0.0079 sec.

The comparison in Table I indicates that both the average depth and the maximum depth of the proposed ID3 algorithm are significantly less than those of the traditional ID3 algorithm. The proposed ID3 can significantly reduce the average depth from the traditional ID3 by 22.78%. Moreover, the accuracy rates of both algorithms are very competitive. The average testing time is reduced from traditional ID3 by 40.61% while the average training time is slightly more than the traditional ID3. However, the number of all nodes in the proposed ID3 algorithm is significantly increased since it expands all cases in the combination of equally important attributes' values. The proposed algorithm creates nodes via combining equally important attributes 2088 nodes from all 6906 non-leaf nodes. This result indicates that this proposed ID3 algorithm can handle decision nodes that have the equally important attributes problem by 30.23%. This experiment implies that the proposed ID3 not only reduces the depth of the decision tree, but also preserves the accuracy of the classification.

V. Conclusion

This paper proposes a variation of ID3 algorithm to combine equally important attributes into a single node of the decision tree classification. The proposed algorithm can optimize an average depth of the decision tree, while the accuracy rate of the decision tree is still preserved. The decision tree which has less depth will usually spend less classification time. The space requirement of this algorithm is more than traditional ID3 since the proposed algorithm must expand all possible branches in the combination of equally important attributes' values. One limitation of the proposed algorithm is that it needs the dataset that is large enough and has various discrete values to exercise all combination branches in the tree. Otherwise, an unseen example may fall into the choice of combination which has not been exercised, i.e. there is no subtree prepared for the unseen patterns. As one of the future works, we plan to improve the proposed algorithm to have less depth than the current version by considering the attributes, whose the values of information gain are close to the information gain of the most important attribute. Those attributes can also be selected as the candidates of the most important attribute as well.

References

[1] R. Quinlan,"Induction of decision trees",Machine Learning, Vol. 1, No. 1, pp.81-106, 1986.

[2] Z. Shichao et al., "A Strategy for Attributes Selection in Cost-Sensitive Decision Trees Induction," in IEEE 8th Int. Conf. on Computer and Information Technology (CIT) Workshops, Sydney, Australia, 2008, pp.8-13.

[3] D. Baoshi et al., "A New Decision Tree Algorithm Based on Rough Set Theory," in Asia-Pacific Conf. on Information Processing (APCIP), Shenzhen, China, 2009, pp.326-329.

[4] L. Hongwu et al., "An Improved ID3 Algorithm Based on Attribute Importance-Weighted," in 2nd Int. Workshop on Database Technology and Applications (DBTA), Wuhan, China, 2010, pp.1-4.

[5] A. Arthur and N. David. (2007). Machine Learning Repository : Connect-4 Data Set [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Connect-4

[6] R. Stuart and N. Peter, *Artificial Intelligence A Modern Approach,* 3rd ed. Essex, England: Pearson, 2014, ch. 18, sec. 3, pp. 708-718.

ก.2 Reducing the Depth of ID3 Algorithm by Combining Values from Neighboring Important Attributes, ICSEC 2018

22nd International Computer Science and Engineering Conference (ICSEC 2018)

Reducing the Depth of ID3 Algorithm by Combining Values from Neighboring Important Attributes

Suratchanan Kraidech, Kietikul Jearanaitanakij
Department of Computer Engineering, Faculty of Engineering
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
Email: 55011362@kmitl.ac.th, kietikul.je@kmitl.ac.th

***Abstract*—The ID3 algorithm is one of the most popular decision tree algorithms which is mainly used in the classification task. There are many pieces of research about improving the ID3 algorithm by using various strategies. We improved the ID3 algorithm by addressing the equally important attributes problem in our previous work. In this paper, we extend our previous algorithm by changing the attribute selection to allow the neighboring second-place important attributes to be combined with the most important attributes. The proposed algorithm is tested on four standard benchmarks from the UCI repository. The experimental results indicate the significant reduction in the maximum depth and the classification depth of the decision tree. In addition, the testing time is also reduced while the classification accuracy is satisfying stable.**

Keywords—Decision tree; ID3 algorithm; Classification; Attribute combining; Neighboring important attributes

I. INTRODUCTION

Decision tree learning is a simple classification model which is used in various fields, e.g., decision analysis, medical diagnosis, pattern recognition. Among decision tree algorithms, ID3, invented by Quinlan in 1986 [1], is one of the most popular approach. There are many efforts to improve the ID3 algorithm.

Liu and Li [2] presented a new method to upgrade performance of ID3 algorithm by using attribute importance and convex function to reduce the complexity of information gain computation. The experimental results showed that their algorithm is significantly better than the traditional ID3 in both accuracy and number of leaf nodes. Moreover, it can save the training time to create the decision tree.

Luo, Chen and Zhang [3] found another approach to reduce complexity of calculation in the traditional ID3 by using Taylor's theorem and attribute similarity theorem. They evaluated their approach on two datasets which do not inform characteristics, i.e., the number of attributes and instances are unknown. The experimental results indicate that their algorithm not only improves the running time but also the classification accuracy.

Liu, Hu and Yan [4] proposed the strategy to improve the ID3 algorithm by using an average Euclidean distance (AED) to handle its defect, i.e., the traditional ID3 algorithm tends to select the attribute which has many values for serving as the decision node. In fact, using the attribute which has fewer values may lead to better decision tree with higher accuracy rate. They also demonstrated that the AED algorithm can improve the precision, running time, and training time of the traditional ID3.

Apart from the above problems, there is another issue that need to be considered. Kraidech and Jearanaitanakij pointed out the equally important attributes problem which it usually occurs in many datasets [5]. This is an incidence when at least two attributes have the same largest information gain. The traditional ID3 will randomly choose one of these equally important attributes to serve as the decision node. The result of this behavior may lead to the decision tree which contains unnecessary depths. Therefore, the classification time is expensively spent. Although their algorithm can effectively reduce the depth of the decision tree, the condition for combining attributes is too rigid as the information gains from important attributes have to be exactly equal.

In this paper, we propose the improvement on our previous work [5] by allowing the neighboring second-place important attributes to be selected and combined with the most important attributes. The experimental results indicate that the proposed algorithm can significantly reduce the number of depths of the final decision tree, while the classification accuracy is still preserved, comparing among the previous work and the traditional ID3. The rest of this paper is organized as follows. In section II, we briefly describe the fundamental of the traditional ID3 algorithm along with the previous work algorithm. Section III explains the process of the proposed algorithm. Section IV describes the benchmark datasets, the experimental conditions, and the experimental results. Finally, the conclusion is provided in section V.

II. RELATED BACKGROUNDS

A. The traditional ID3 algorithm

ID3, is an algorithm that generates a decision tree from the training set and verifies it by an unseen test set. The basic technique of the traditional ID3 algorithm is briefed below.

Step 1: Computes the information gain for every remaining attribute by using Eq. (1) - (3) and generates the current decision node by labeling it with the attribute which has the largest information gain. If there are at least two attributes having the same largest information gain, randomly select one from them.

$$IG(x) = Entropy(x) - Remainder(x) \quad (1)$$

978-1-5386-8164-0/18/$31.00 ©2018 IEEE

$$Entropy(x) = -\sum_{c \in C} p(c) \log_2 p(c) \quad (2)$$

$$Remainder(x) = \sum_{f \in F} \left\{ -p(f) \sum_{h \in C} p(h) \log_2 p(h) \right\} \quad (3)$$

Where

IG(x) is the information gain of an attribute x,

C is a set of classes in the dataset,

p(c) is the proportion of the number of examples in class c to the total number of remaining examples,

F is a set of values in an attribute x,

p(f) is the proportion of the number of examples having value f to the number of remaining examples,

p(h) is the proportion of the number of examples having class h and value f to the total number of examples having value f.

Step 2: Split the training examples of the current decision node depending on the possible values of the selected attribute.

Step 3: Repeat steps 1 to 2 by using the split training examples and the remaining attributes until there are no unclassified examples left, i.e., every example is classified. Fig. 1 shows the flowchart of the traditional ID3 algorithm.

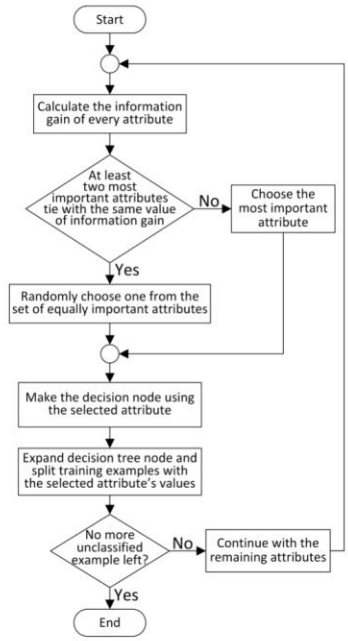


Fig. 1. Flowchart of the traditional ID3 algorithm

B. *Improving ID3 algorithm with the combination of equally important attributes*

Our previous work [5] was proposed to handle equally important attributes problem. The basic concept of this algorithm is to combine the most important attributes into a single decision node. The branches of the decision node are determined from the Cartesian product between possible values of the selected attributes. For example, suppose that attributes A and B are combined into a decision node. The attribute A has 3 possible values, while B has 2. By combining possible values from both attributes, the decision node will have 6 branches, i.e., 3 x 2. The Fig. 2 illustrates how attributes A and B are combined into a single decision node and how branches of decision node are determined.

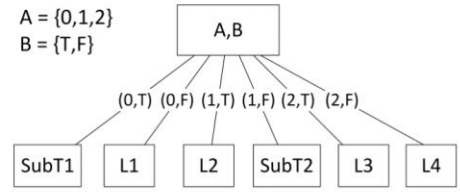


Fig. 2. Decision node with combined attributes {A, B}. Denote that SubT1 and SubT2 are subtrees, while L1-L4 are leaf nodes.

The experimental results in our previous work shows that the depth of the decision tree is significantly decreased while the classification accuracy is still preserved. However, the condition for combining attributes is too rigid for some datasets because the information gains of important attributes must be exactly equal. We propose a method to relax this condition for generating more shallow decision tree in the next section.

III. THE PROPOSED ALGORITHM

From our previous work [5], we found that if we increase the number of attributes per decision node, e.g., 3, 4, 5, the depth of the decision tree will remain stable. This experiment implies that the algorithm cannot reduce the depth further. It is perhaps because the dataset itself does not contain enough equally important attributes. Therefore, attributes cannot be combined into a single decision node. Consequently, we decide to modify our previous algorithm to further reduce the depth of the decision tree while the classification accuracy is still preserved.

The main concept of our proposed algorithm is similar to our previous work, except how important attributes are chosen. Instead of combining only the attributes which are most important, we allow the algorithm to combine the neighboring second-place important attributes together with the most important attributes. As a result, we can apply the proposed algorithm to the dataset whose attributes are competitively important but have insignificantly different values of information gain. It is worth to note that the second-place important attributes should not be combined to the most important attributes if their information gains are not competitive. Otherwise, the decision node will contain unimportant attributes which may deteriorate the classification accuracy.

The steps of the proposed algorithm are described along with a flowchart in Fig. 3.

Step 1: For each remaining attribute, calculate its information gain by using Eq. (1) - (3).

Step 2: Create an empty decision node, i.e., a node without any combination of attributes' values, and assign it as the current decision node.

Step 3: Select the most important attributes which tie with the same value of information gain and combine them into the current decision node. Repeat this step until the current decision node contains a combination of N attributes or the most important attributes are used up.

Step 4: If the attribute combination of the current decision node contains less than N attributes, choose the neighboring second-place important attribute by using the following criterion. If the information gain of the second-place important attribute is less than that of the most important attribute by the range of D percent, randomly select the second-place important attribute into the current decision node. Afterward, keep randomly selecting another neighboring second-place important attribute until we have a combination of N attributes in the current decision node or the second-place important attributes are used up.

Step 5: Expand a decision node by splitting the training examples based on the combination values of selected attributes.

Step 6: Repeat steps 1 – 5 on the remaining attributes until there is no unclassified example left.

From steps explained above, there are 2 parameters in our proposed algorithm. The first parameter is N which defines the maximum number of selected attributes per decision node. Another parameter is D which designates the percentage of the maximum difference between the neighboring second-place important attribute's information gain and the most important attribute's information gain. It prevents unimportant attributes from being chosen into the decision node. For example, the attribute A is the most important attribute which has the information gain 0.96. Suppose B is the second-place important attribute whose information gain is 0.5. If the parameter D was set at 5%, the attribute B will not be combined into a decision node with the attribute A since the difference between the information gain of A and B exceeds 5%.

IV. Experiments

A. Datasets

We conduct the experiment on four benchmarks from the UCI repository [6]. The characteristics of these datasets are described in Table I.

TABLE I. Characteristics of Datasets

Dataset	Instances	Attributes	Classes
Connect-4	67,557	42	3
Firm-Teacher	10,800	16	7
Phishing Websites	11,055	30	2
Insurance Company Benchmark	9,822	85	2

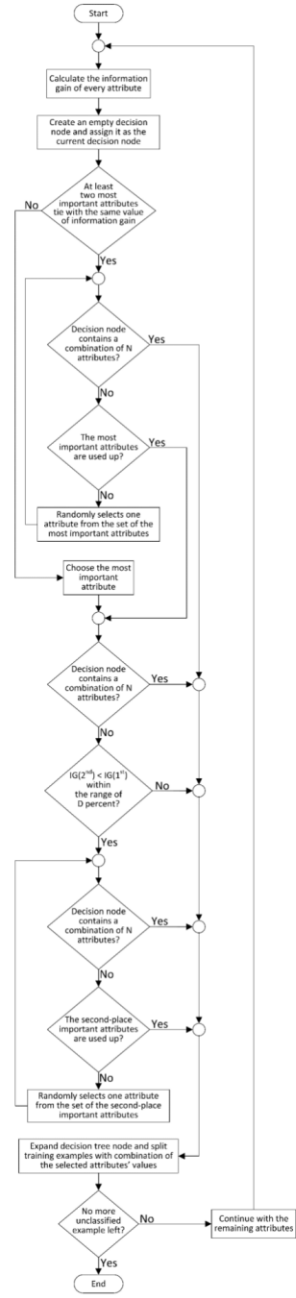


Fig. 3. Flowchart of the proposed ID3 algorithm (IG : information gain, 1st : the most important attribute, 2nd : the second-place important attribute)

B. Experimental conditions

We split every dataset into two parts; training set and test set. Both parts have equal number of instances. We also balance the number of instances based on classes in the dataset.

For significant and clear results, we find the best parameters via preliminary trials and use them in our main experiment. We found that if parameters N and D are large, the maximum depth of the decision tree will be very shallow. However, setting both parameters to very large value may not be useful. The decision tree building process may take very long time since there are many attributes that can be combined into each decision node. Moreover, each decision node contains many branches to expand in the next level. Consequently, we perform the preliminary run to find the best parameters by varying the parameters N and D by 2 values and executing 1000 preliminary runs. Table II shows an example of the preliminary run on Phishing Websites dataset.

TABLE II. THE PRELIMINARY RUN ON PHISHING WEBSITES DATASET

	N = 2		N = 3	
	D = 0.005	***D = 0.05***	***D = 0.005***	***D = 0.05***
Max depth	23	17.65	20	**15**
Classification depth	4.79	3.44	4.71	**3.35**
Accuracy	**95.57%**	95.43%	95.56%	95.38%
Training time (sec.)	0.07	**0.05**	0.1	0.08
Testing time (sec.)	2.84e-5	2.05e-5	2.8e-5	**1.99e-5**

From the results in Table II, we can see that the parameter setting {N = 3, D = 0.05} gives the best result in most measurements except the classification accuracy and the training time which are nearly stable. Therefore, we use this parameter setting in our main experiment.

In order to perform a fair comparison, we control the common experimental conditions, e.g., the number of runs = 1000, the maximum number of selected equally important attributes per decision node = 3, to be same for all applicable algorithms.

C. Experimental Results

We measure the maximum depth, the classification depth (the depth where an instance was classified), accuracy, training time and testing time of the resulting decision tree. The experimental results are compared among the traditional ID3 algorithm and our previous algorithm [5] as shown in Tables III - VI.

TABLE III. EXPERIMENTAL RESULTS OF THE CONNECT-4 DATASET

	Traditional	Previous work [5]	Proposed
Max depth	24.19	18.06	**12.03**
Classification depth	9.43	9.36	**5.57**
Accuracy	72.95%	**73.66%**	72.25%
Training time (sec.)	**1.49**	2.12	1.58
Testing time (sec.)	3.4e-4	3.38e-4	**2e-4**

TABLE IV. EXPERIMENTAL RESULTS OF THE FIRM-TEACHER CLAVE-DIRECTION CLASSIFICATION DATASET

	Traditional	Previous work [5]	Proposed
Max depth	15	14	**10**
Classification depth	9.6	9.33	**5.83**
Accuracy	72.26%	72.64%	**73.04%**
Training time (sec.)	0.1	0.12	**0.08**
Testing time (sec.)	5.54e-5	5.39e-5	**3.35e-5**

TABLE V. EXPERIMENTAL RESULTS OF THE PHISHING WEBSITES DATASET

	Traditional	Previous work [5]	Proposed
Max depth	30	22	**15**
Classification depth	5.61	5.28	**3.35**
Accuracy	95.26%	95.34%	**95.38%**
Training time (sec.)	**0.06**	0.11	0.08
Testing time (sec.)	3.33e-5	3.17e-5	**1.99e-5**

TABLE VI. EXPERIMENTAL RESULTS OF THE INSURANCE COMPANY BENCHMARK DATASET

	Traditional	Previous work [5]	Proposed
Max depth	85	32	**30**
Classification depth	4.17	3.57	**2.3**
Accuracy	90.35%	91.27%	**91.97%**
Training time (sec.)	**0.4**	6.79	7.14
Testing time (sec.)	2.01e-5	1.8e-5	**1.13e-5**

The experimental results from Tables III - VI show the significantly decreasing in the maximum depth and the classification depth of the proposed algorithm. In addition, the classification accuracy of the proposed algorithm is competitive to other algorithms except the classification accuracy of the Insurance Company Benchmark dataset which slightly increases from the traditional algorithm by 1.79% and from the previous work algorithm by 0.77%.

It is interesting to discuss about the training time of three algorithms on the Insurance Company Benchmark dataset since the traditional ID3 achieves the significantly low training time. The reason of this strange result can be explained by Fig. 4 which shows the number of possible values for each attribute in the Insurance Company Benchmark dataset.

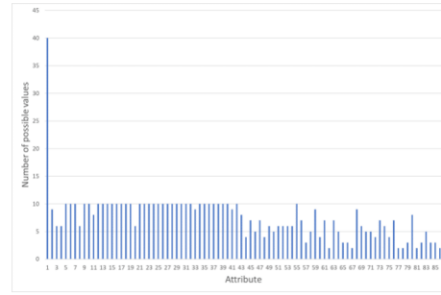


Fig. 4. Number of possible values in each attribute in the Insurance Company Benchmark dataset

We can see from Fig. 4 that the first attribute has very high number of possible values comparing among other attributes. If the first attribute was combined with other attributes, there will be a large number of value combinations (branches) in a decision node. As a result, both the proposed algorithm and the previous work take longer time to build the decision tree than the traditional ID3 for this dataset.

Optimizing the maximum depth of the decision tree may not help the classification reducing the testing time on unseen instances if most of them are classified at the shallow depths. Therefore, measuring the reduction of testing time would tell us how fast the proposed algorithm in classifying unseen data. Table VII shows the improvement on both the classification depth and the testing time of the proposed algorithm, compared to our previous algorithm [5].

TABLE VII. THE IMPROVEMENT ON THE CLASSIFICATION DEPTH AND THE TESTING TIME OF THE PROPOSED ALGORITHM, COMPARED TO OUR PREVIOUS ALGORITHM

	Classification depth reducing	**Testing time reducing**
Connect-4	40.56%	40.71%
Firm-Teacher	37.56%	37.80%
Phishing Websites	36.53%	37.07%
Insurance Company Benchmark	35.60%	36.88%

The percentages of the classification depth reduction are approximately equal to the percentages of the testing time reduction in all datasets. Therefore, the reduction of the classification depth of the proposed algorithm is concordant with the reduction of its testing time. We can conclude from the experimental results in this section that the proposed algorithm can improve our previous work by reducing the maximum depth of decision tree effectively while the classification accuracy is still preserved. Moreover, its testing time and the classification depth are also reduced congruently.

V. CONCLUSION

In this paper, we extend the previous work [5] to improve the ID3 algorithm by relaxing the attribute selection to allow the neighboring second-place important attributes to be combined with the most important attributes. The proposed algorithm can efficiently reduce the maximum depth and the classification depth of the decision tree. It can also preserve the classification accuracy. Furthermore, the testing time is significantly reduced and concordant with the reduction of classification depth. However, the proposed algorithm requires space more than the previous work since it combines more important attributes into each decision node which causes more branch expansions in the deeper level. Another limitation of the proposed algorithm is that it needs a sufficiently large number of instances to train every branch of the attribute combinations in each decision node. Otherwise, some branches which have never been trained may miss classifying unseen patterns.

REFERENCES

[1] J. R. Quinlan, "Induction of decision trees," *Mach. Learning*, vol. 1, pp.81-106, 1986.

[2] J. Liu and N. Li, "Optimized ID3 algorithm based on attribute importance and convex function," in *2011 IEEE Int. Symp. IT in Medicine and Education*, Cuangzhou, China, pp. 136-139.

[3] H. Luo, Y. Chen, and W. Zhang, "An improved ID3 algorithm based on attribute importance-weighted," in *2010 2nd Int. Workshop Database Technology and Applications*, Wuhan, China, pp. 1-4.

[4] Q. Liu, D. Hu, and Q. Yan, "Decision tree algorithm based on average Euclidean distance," in *2010 2nd Int. Conf. Future Computer and Communication*, Wuhan, China, pp. V1-507 - V1-511.

[5] S. Kraidech and K. Jearanaitanakij, "Improving ID3 algorithm by combining values from equally important attributes," in *2017 21st Int. Computer Science and Engineering Conf.*, Bangkok, Thailand, pp. 102-105.

[6] D. Dua and E. K. Taniskidou. *UCI Machine Learning Repository*. (2017) [Online]. Available: http://archive.ics.uci.edu/ml

ประวัติผู้เขียน

ชื่อ-นามสกุล	สุรัชนันท์ ไกรเดช
วัน เดือน ปีเกิด	23 กรกฎาคม 2536
ที่อยู่	บ้านเลขที่ 1 ซ.ชลบุรี-บ้านบึง 1 ถ.ชลบุรี-บ้านบึง ต.บ้านบึง อ.บ้านบึง จ.ชลบุรี 20170
ประวัติการศึกษา	วศ.บ. (เกียรตินิยมอันดับ 1) วิศวกรรมคอมพิวเตอร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง, 2558