การปรับปรุงอัลกอริทึม ID3 โดยการละเลยกรณีตัวอย่างส่วนน้อย

IMPROVING ID3 ALGORITHM BY IGNORING MINOR INSTANCES

ณิชา แก้วรอด

NICHA KAEWROD

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ.2562

KMITL-2019-EN-M-070-025

การปรับปรุงอัลกอริทึม ID3 โดยการละเลยกรณีตัวอย่างส่วนน้อย

IMPROVING ID3 ALGORITHM BY IGNORING MINOR INSTANCES

ณิชา แก้วรอด

NICHA KAEWROD

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ.2562

KMITL-2019-EN-M-070-025

IMPROVING ID3 ALGORITHM BY IGNORING MINOR INSTANCES

NICHA KAEWROD

A THESIS SUBMITTED IN PARTIAL 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-070-025

หัวข้อวิทยานิพนธ์	การปรับปรุงอัลกอริทึม ID3 โดยการละเลยกรณีตัวอย่างส่วนน้อย
นักศึกษา	นางสาวณิชา แก้วรอด
รหัสประจำตัว	59601087
ปริญญา	วิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
พ.ศ.	2562
อาจารย์ที่ปรึกษาวิทยานิพนธ์	รศ.ดร.เกียรติกูล เจียรนัยธนะกิจ

บทคัดย่อ

อัลกอริทึมไอดีทรี (ID3 algorithms) เป็นหนึ่งในอัลกอริทึมที่ใช้สร้างต้นไม้ตัดสินใจ (Decision tree) เพื่อการจำแนกข้อมูล (Classification) และเป็นที่นิยมใช้กันอย่างกว้างขวาง อย่างไรก็ตาม อัลกอริทึม ID3 ประสบปัญหาความเข้มงวดในการสร้างกฎการตัดสินใจ (Decision rule) ส่งผลให้ต้นไม้ตัดสินใจที่ได้อาจมีกฎการตัดสินใจที่มากเกินจำเป็น บางส่วนของกฎการตัดสินใจเหล่านี้อาจมีจำนวนกรณีตัวอย่าง (Instance) ที่น้อยเกินไปซึ่งอาจส่งผลกระทบต่อความแม่นยำในการจำแนก (Classification accuracy) เพียงเล็กน้อย ดังนั้นเป้าหมายของงานวิจัยนี้คือเพื่อนำเสนอวิธีลดความเข้มงวดของอัลกอริทึม ID3 โดยการละเลยกรณีตัวอย่างส่วนน้อย อัลกอริทึมที่เสนอถูกทดสอบประสิทธิภาพโดยใช้ชุดข้อมูล (Dataset) 17 ชุดจากฐานข้อมูลยูซีไอ (UCI repository) ผลการทดลองแสดงให้เห็นว่าอัลกอริทึมที่นำเสนอไม่เพียงแต่ลดจำนวนกฎการตัดสินใจและความลึกสูงสุดของต้นไม้ตัดสินใจได้อย่างชัดเจน แต่ยังคงรักษาความแม่นยำในการจำแนกให้ลดลงเพียงเล็กน้อยในชุดข้อมูลส่วนใหญ่ ยิ่งไปกว่านั้นเวลาในการฝึกฝน (Training time) และเวลาในการทดสอบ (Testing time) ยังน้อยกว่าอัลกอริทึม ID3 แบบดั้งเดิมอย่างชัดเจน

Thesis	Improving ID3 Algorithm by Ignoring Minor Instances
Student	Miss Nicha Kaewrod
Student ID.	59601087
Degree	Master of Engineering
Program	Computer Engineering
Year	2019
Thesis Advisor	Assoc. Prof. Dr. Kietikul Jearanaitanakij

ABSTRACT

Among various classification algorithms, ID3 is one of the most widely used and well-known tools that generates an efficient decision tree. Nevertheless, ID3 is too rigorous in generating the decision rules. As a result, the final decision tree may carry too many decision rules. Some of these decision rules may have very low number of instances which do not make significant change to the classification accuracy. The aim of this paper is to propose an approach to relax the rigorousness of the conventional ID3 algorithm by ignoring minor instances so that the resulting decision tree will have the lower number of depths yet produce promising accuracy. The proposed algorithm is examined on seventeen datasets from UCI repository. The experimental results indicate that the proposed algorithm not only significantly reduces the maximum number of depths of the decision tree, but also retains the classification accuracy in the satisfying level of most datasets. Moreover, the training time, the classification time, and the number of decision rules of the proposed algorithm are lower than those of the conventional ID3.

กิตติกรรมประกาศ

วิทยานิพนธ์เล่มนี้สำเร็จได้ด้วยความกรุณาจากอาจารย์ที่ปรึกษา รศ.ดร.เกียรติกูล เจียรนัยธนะกิจ ที่ให้ความช่วยเหลือ ให้คำชี้แนะช่วยแก้ปัญหาตลอดจนให้ความรู้และประสบการณ์ที่ดีแก่ข้าพเจ้า

ขอขอบพระคุณ รศ.ดร.สุรพงค์ เอื้อวัฒนามงคล ผศ.ดร.ชุติเมษฎ์ ศรีนิลทา รศ.ดร.บุญธีร์ เครือตราชู และ ดร.รัฐชัย ชาวอุทัย กรรมการสอบหัวข้อและโครงร่างวิทยานิพนธ์ที่ได้กรุณาให้คำแนะนำตลอดจนข้อชี้แนะ จนในที่สุดทำให้วิทยานิพนธ์ฉบับนี้สำเร็จลงได้

สุดท้ายต้องขอขอบคุณนายสุรัชนันท์ ไกรเดช ที่คอยให้คำแนะนำช่วยแก้ปัญหาและช่วยสอนการเขียนโปรแกรมในระดับที่ยากขึ้น ส่งผลให้ข้าพเจ้ามีความชำนาญยิ่งขึ้น

สำหรับคุณงามความดีอันใดที่เกิดจากวิทยานิพนธ์ฉบับนี้ ข้าพเจ้าขอมอบให้กับบิดามารดา ซึ่งเป็นที่รักและเคารพยิ่ง ตลอดจนครูอาจารย์ที่เคารพทุกท่านที่ได้ประสิทประสาทวิชาความรู้และถ่ายทอดประสบการณ์ที่ดีให้แก่ข้าพเจ้า

ณิชา แก้วรอด

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญรูป

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

อัลกอริทึมไอดีทรี (ID3 หรือ Iterative Dichotomiser 3) [1] ถูกพัฒนาขึ้นโดยควินแลน (Quinlan) ในปี 1986 เพื่อใช้ในการจำแนกประเภทของข้อมูลในรูปแบบของต้นไม้ตัดสินใจ (Decision tree) ด้วยความที่ ID3 เป็นอัลกอริทึมที่ค่อนข้างง่าย ไม่ซับซ้อน และให้ผลการจำแนกข้อมูลที่มีความแม่นยำค่อนข้างสูง จึงเป็นที่นิยมใช้กันอย่างแพร่หลาย อัลกอริทึม ID3 ใช้การคำนวณเกนความรู้ (Information gain) เพื่อเลือกแอตทริบิวต์ (Attribute) ที่มีความสำคัญสูงสุดมาสร้างโหนด (Node) และทำการแบ่งกรณีตัวอย่าง (Instance) ตามค่าที่ไม่ซ้ำกัน (Value) ของแอตทริบิวต์นั้น ๆ ไปเรื่อย ๆ จนกว่ากรณีตัวอย่างทั้งหมดจะถูกจำแนกประเภทจนเสร็จสิ้น แต่อย่างไรก็ตามแอตทริบิวต์ที่มีจำนวนค่าที่ไม่ซ้ำกันมาก ๆ มักถูกเลือกก่อนเสมอ ทั้ง ๆ ที่แอตทริบิวต์ที่มีจำนวนค่าที่ไม่ซ้ำกันที่น้อยกว่าอาจให้ผลลัพธ์ที่ดีกว่า นี้คือปัญหาการลำเอียงในการเลือกแอตทริบิวต์ ซึ่งเป็นข้อเสียหลักของการคัดเลือกแอตทริบิวต์โดยใช้การคำนวณเกนความรู้ ทำให้มีงานวิจัยที่พยายามคิดค้นเกณฑ์ในการเลือกแอตทริบิวต์สำหรับอัลกอริทึม ID3 ที่ไร้ซึ่งปัญหาข้างต้น ยกตัวอย่างเช่นการประยุกต์ทฤษฎีเซตอย่างหยาบ (Rough set theory) [2] เพื่อคำนวณค่าความมั่นคงของแอตทริบิวต์ (Consistency of attribute) และเลือกแอตทริบิวต์ที่มีค่า Consistency สูงสุดเพื่อสร้างโหนด ผลการทดลองแสดงให้เห็นว่าวิธีการนี้สามารถเพิ่มความแม่นยำในการจำแนก (Accuracy) ได้ชัดเจน แต่ใช้ได้เฉพาะชุดข้อมูลที่มีการจำแนกแค่ 2 คลาส (Class) เท่านั้น ตัวอย่างงานสุดท้ายคือการปรับปรุงค่าเกนความรู้โดยการถ่วงน้ำหนัก (Weighted Modified Information gain) [3] เพื่อเพิ่มความแม่นยำในการจำแนก อย่างไรก็ตามวิธีการนี้ทำให้ต้นไม้ตัดสินใจที่ได้มีความซับซ้อนมากกว่าต้นไม้ตัดสินใจที่ได้จากอัลกอริทึม ID3 แบบดั้งเดิม อีกหนึ่งข้อเสียของอัลกอริทึม ID3 คือปัญหาแอตทริบิวต์ที่มีความสำคัญเท่ากัน [4,5] ในระหว่างการสร้างต้นไม้ตัดสินใจอาจเกิดเหตุการณ์ที่มีแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเท่ากันมากกว่าหนึ่งแอตทริบิวต์ เมื่อเกิดเหตุการณ์นี้ขึ้นอัลกอริทึม ID3 แบบดั้งเดิมจะสุ่มเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดเหล่านั้นมา 1 แอตทริบิวต์และนำไปสร้างโหนดของต้นไม้ตัดสินใจ พฤติกรรมนี้นำไปสู่ความไม่มั่นคงของความลึกสูงสุด (Maximum depth) ในต้นไม้ตัดสินใจผลลัพธ์ ตัวอย่างงานวิจัยที่พยายามปรับปรุงอัลกอริทึม ID3 แบบดั้งเดิมเพื่อลดจำนวนความลึกสูงสุดเช่น ตัวอย่างแรกคือการปรับปรุงอัลกอริทึม ID3 โดยการประยุกต์ใช้การค้นหาแบบเอสตาร์ (A* Search) [4] แม้ว่างานวิจัยนี้สามารถรักษาความแม่นยำในการจำแนกให้อยู่ในระดับที่น่าพอใจในชุดข้อมูลทั้งหมด แต่ความลึกสูงสุดลดลงอย่างชัดเจนในชุดข้อมูลที่ซับซ้อนและมีขนาดใหญ่เพียงชุดข้อมูลเดียวเท่านั้น อีกทั้งยังใช้เวลาในการสร้างต้นไม้ตัดสินใจที่นาน ตัวอย่างสุดท้ายคือการปรับปรุง

อัลกอริทึม ID3 โดยการรวมค่าระหว่างแอตทริบิวต์ที่มีความสำคัญเท่ากัน [5] แม้ว่าวิธีการนี้จะสามารถลดจำนวนความลึกสูงสุดของต้นไม้ตัดสินใจและยังสามารถรักษาอัตราความแม่นยำในการจำแนกไว้ได้ แต่จำนวนของกฎการตัดสินใจ (Decision rule) เพิ่มขึ้นอย่างเห็นได้ชัด ซึ่งตามปกติอัลกอริทึม ID3 แบบดั้งเดิมก็มีปัญหาความเข้มงวดในการสร้างกฎการตัดสินใจอยู่แล้ว ซึ่งกฎการตัดสินใจที่มีมากเกินไปต้องใช้พื้นที่เพื่อเก็บต้นไม้ตัดสินใจในหน่วยความจำมากขึ้นอีกด้วย เนื่องจากถ้ามีจำนวนกฎการตัดสินใจมากต้นไม้ตัดสินใจผลลัพธ์จะมีความซับซ้อนและใหญ่ตามไปด้วย และบางส่วนของกฎการตัดสินใจเหล่านี้อาจจะมีบางโหนดที่มีจำนวนกรณีตัวอย่างที่น้อยมาก ๆ ซึ่งอาจส่งผลกระทบต่อความแม่นยำในการจำแนกของต้นไม้ตัดสินใจผลลัพธ์เพียงเล็กน้อยเท่านั้น ดังนั้นผู้เขียนจึงเกิดแนวคิดที่จะลดความเข้มงวดของกฎการตัดสินใจในอัลกอริทึม ID3 หรืออีกความหมายคือต้องการลดจำนวนกฎการตัดสินใจในต้นไม้ตัดสินใจที่สร้างจากอัลกอริทึม ID3 แบบดั้งเดิมโดยการละเลยกรณีตัวอย่างที่มีจำนวนน้อย ๆ

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

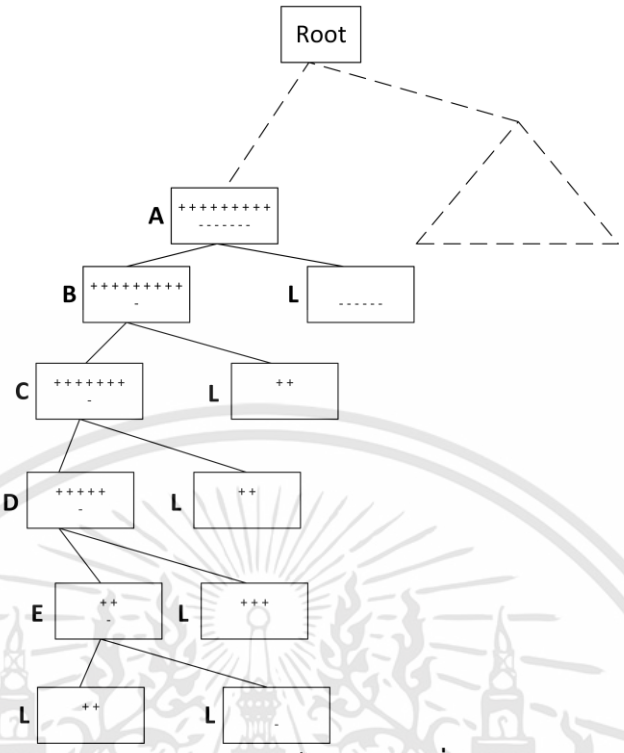
เสนออัลกอริทึมแบบใหม่เพื่อลดความเข้มงวดในการสร้างกฎการตัดสินใจและความลึกสูงสุดของอัลกอริทึม ID3 แบบดั้งเดิม ในขณะที่ความแม่นยำลดลงเพียงเล็กน้อย

1.3 สมมติฐานของการศึกษา

อัลกอริทึมที่เสนอสามารถลดจำนวนกฎการตัดสินใจและจำนวนความลึกสูงสุดของต้นไม้ตัดสินใจได้อย่างชัดเจน ทั้งยังสามารถรักษาแม่นยำในการจำแนกให้ลดลงเพียงเล็กน้อย

1.4 ทฤษฎีหรือแนวความคิดที่ใช้ในการวิจัย

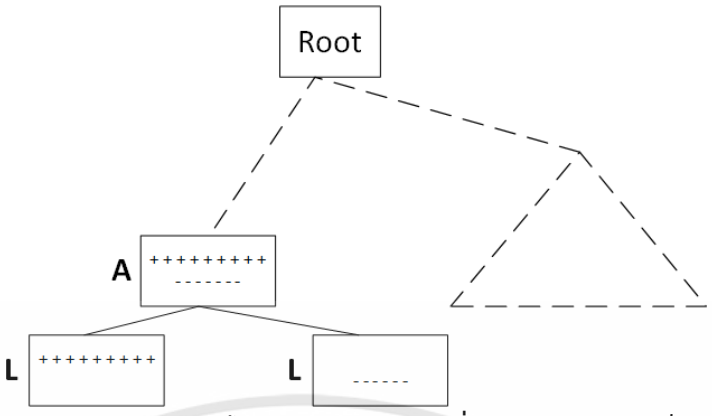
อัลกอริทึม ID3 มีปัญหาความเข้มงวดในการสร้างกฎการตัดสินใจ โดยปกติแล้วถ้าชุดข้อมูลสำหรับฝึกฝน (Training set) มีจำนวนแอตทริบิวต์หรือกรณีตัวอย่างที่มากเกินไป ต้นไม้ตัดสินใจที่ได้อาจจะมีจำนวนกฎการตัดสินใจที่มากเกินจำเป็น บางส่วนของกฎการตัดสินใจเหล่านี้อาจจะมีบางโหนดที่มีจำนวนกรณีตัวอย่างที่น้อยมาก ๆ ถ้าเราตัดกรณีตัวอย่างส่วนน้อยนี้ออกไปความแม่นยำในการจำแนกประเภทจะลดลงเพียงเล็กน้อย เพื่อให้เข้าใจแนวคิดนี้มากขึ้นให้พิจารณารูปที่ 1.1



รูปที่ 1.1 แสดงตัวอย่างของบางส่วนของต้นไม้ตัดสินใจที่มีกฎการตัดสินใจที่มากเกินจำเป็น

จากรูปที่ 1.1 สมมติว่าส่วนที่เป็นเส้นประคือส่วนบนของต้นไม้ตัดสินใจที่มีแอตทริบิวต์บางส่วนถูกใช้ในการสร้างโหนดส่วนบน โหนด L คือโหนดคำตอบ (Leaf node) จากรูปเส้นทางจากโหนดราก (Root node) ลงมาถึงโหนดคำตอบ 1 โหนด นับเป็น 1 กฎการตัดสินใจ ดั้งนั้นจำนวนกฎการตัดสินใจของต้นไม้ตัดสินใจจึงเท่ากับจำนวนโหนดคำตอบทั้งหมด ส่วนโหนด A, B, C และ D คือโหนดการตัดสินใจ (Decision node) ในงานวิจัยนี้เรานิยามโหนดการตัดสินใจคือโหนดที่ยังไม่สามารถจำแนกประเภทของกรณีตัวอย่างได้ทั้งหมด หรือโหนดที่มีโหนดลูกบางส่วนไม่เป็นโหนดคำตอบ และโหนด E คือโหนดการจำแนกประเภท (Classified node) ก็คือโหนดที่สามารถจำแนกประเภทของกรณีตัวอย่างได้เสร็จสิ้นโดยที่ไม่ต้องผ่านไปที่โหนดตัวอื่น หรือกล่าวง่าย ๆ ว่าเป็นโหนดที่มีโหนดลูกเป็นโหนดคำตอบทั้งหมด ส่วนเครื่องหมาย + และ - ในรูปหมายถึงกรณีตัวอย่างที่อยู่ในคลาสบวกและลบตามลำดับ

พิจารณาที่โหนด B พบว่าโหนด B ประกอบไปด้วยกรณีตัวอย่างที่เป็นคลาสบวก 9 กรณี และกรณีตัวอย่างที่เป็นคลาสลบ 1 กรณี ซึ่งกรณีตัวอย่างที่เป็นคลาสลบมีจำนวนน้อยมาก ๆ เมื่อเปรียบเทียบกับกรณีตัวอย่างที่เป็นคลาสบวกในโหนด B เราเรียกกรณีตัวอย่างที่มีจำนวนน้อย ๆ นี้ว่ากรณีตัวอย่างส่วนน้อย (Minor instance) จะเห็นได้ว่ามันคุ้มค่าที่จะละเลยกรณีตัวอย่างส่วนน้อยเหล่านี้ เพื่อป้องกันมิให้ต้นไม้ตัดสินใจที่ได้มีกฎการตัดสินใจที่มากเกินจำเป็น ดังนั้นเมื่อเราละเลยกรณีตัวอย่างส่วนน้อยในโหนด B ส่งผลให้โหนด B จะเหลือแค่กรณีตัวอย่างใน 1 คลาส ซึ่งก็คือกรณีตัวอย่างที่เป็นคลาสบวก โหนด B จึงกลายเป็นโหนดคำตอบดังในรูปที่ 1.2



รูปที่ 1.2 แสดงบางส่วนของต้นไม้ตัดสินใจสุดท้ายเมื่อโหนด B กลายเป็นโหนดคำตอบ

จากรูปที่ 1.2 จะเห็นว่าเมื่อเราละเลยกรณีตัวอย่างส่วนน้อยในโหนด B จำนวนกฎการตัดสินใจและระดับความลึกสูงสุดลดลงจาก 6 กฎการตัดสินใจและระดับความลึกสูงสุด 5 เมื่อเริ่มนับความลึกสูงสุดที่โหนด A ในรูปที่ 1.1 เหลือเพียง 2 กฎการตัดสินใจและระดับความลึกสูงสุด 1 ในรูปที่ 1.2 แต่อย่างไรก็ตามมันมีความเสี่ยงสูง คือเราไม่ควรละเลยกรณีตัวอย่างส่วนน้อยของโหนดที่อยู่บริเวณต้น ๆ ของต้นไม้ตัดสินใจ เพราะว่าโหนดด้านบนนั้นมีความสำคัญอย่างมาก ถ้าเราทำการแปลงโหนดด้านบนของต้นไม้ตัดสินใจให้กลายเป็นโหนดคำตอบ หรือก็คือการละเลยกรณีตัวอย่างส่วนน้อยบริเวณตอนต้นของการสร้างต้นไม้ตัดสินใจ ความแม่นยำในการจำแนกจะลดลงอย่างเห็นได้ชัด นี้คือแนวคิดที่นำไปสู่การพัฒนาอัลกอริทึมที่เสนอ

1.5 ขอบเขตการวิจัย

1. ศึกษาลักษณะการทำงานและปัญหาของอัลกอริทึม ID3 แบบดั้งเดิม

2. อัลกอริทึมที่เสนอมุ่งเน้นเพื่อแก้ปัญหาความเข้มงวดในการสร้างกฎการตัดสินใจของอัลกอริทึม ID3 แบบดั้งเดิม

3. ทดสอบประสิทธิภาพของอัลกอริทึมที่เสนอโดยใช้ชุดข้อมูล 17 ชุดจากฐานข้อมูลยูซีไอ [6]

1.6 ขั้นตอนของการศึกษา

1. ศึกษาทฤษฎีพื้นฐานที่เกี่ยวข้อง
2. ศึกษางานวิจัยที่เกี่ยวข้องและปัญหาที่มุ่งเน้น
3. คิดค้นและพัฒนาอัลกอริทึมเพื่อแก้ปัญหาที่มุ่งเน้น
4. ทำการทดสอบประสิทธิภาพของอัลกอริทึมที่เสนอและปรับปรุงเพื่อผลการทดลองที่ดีขึ้น
5. วิเคราะห์และสรุปผลการทดลอง
6. จัดทำเอกสารประกอบวิทยานิพนธ์

1.7 เครื่องมือและอุปกรณ์ที่ใช้ในงานวิจัย

1. เครื่องคอมพิวเตอร์ซึ่งมีหน่วยประมวลผลกลาง Intel(R) Core(TM) i7-7700HQ 2.80 GHz 64 บิต และหน่วยความจำหลักขนาด 8 GB ลงระบบปฏิบัติการ Windows 10 Pro 64 บิต

2. โปรแกรม Microsoft Visual Studio 2017 สำหรับเขียนโปรแกรมอัลกอริทึมที่เสนอด้วยภาษา C#

1.8 โครงสร้างของวิทยานิพนธ์

บทที่ 1 อธิบายถึงความเป็นมาและความสำคัญของปัญหา ความมุ่งหมาย วัตถุประสงค์ สมมติฐานของการศึกษา ทฤษฎี แนวคิดที่ใช้ในการวิจัย ขอบเขตการวิจัย ขั้นตอนของการศึกษา เครื่องมือและอุปกรณ์ที่ใช้ในงานวิจัย

บทที่ 2 อธิบายถึงนิยามรวมถึงทฤษฎีพื้นฐานที่เกี่ยวข้อง

บทที่ 3 อธิบายถึงงานวิจัยที่เกี่ยวข้อง

บทที่ 4 อธิบายถึงวิธีการที่พัฒนาขึ้นคือการปรับปรุงอัลกอริทึม ID3 โดยการละเลยกรณีตัวอย่างส่วนน้อย

บทที่ 5 อธิบายถึงเงื่อนไขในการทดลองและผลการทดลอง

บทที่ 6 กล่าวสรุปผลการวิจัยและข้อเสนอแนะ

บทที่ 2
นิยามและทฤษฎีพื้นฐานที่เกี่ยวข้อง

2.1 เอนโทรปี (Entropy)

เอนโทรปีในทางทฤษฎีของข้อมูลหมายถึงค่าความไม่เป็นระเบียบของข้อมูลมีหน่วยเป็นบิต (Bit) ยิ่งข้อมูลมีการกระจายมาก ค่าเอนโทรปีจะยิ่งมากยกตัวอย่างเช่น การโยนเหรียญเมื่อเราโยนเหรียญ 1 เหรียญ ผลลัพธ์ที่เป็นไปได้มีอยู่ 2 ผลลัพธ์คือเหรียญออกหัวและเหรียญออกก้อย สมมติว่าเราโยนเหรียญ 10 ครั้ง มีครั้งที่เหรียญออกหัวและก้อยเป็น 5 ครั้งเท่ากัน นั้นหมายความว่าค่าความน่าจะเป็นที่เหรียญออกหัวและก้อยมีค่าเท่ากันคือ 0.5 การโยนเหรียญ 10 ครั้ง ก็เปรียบเหมือนกับเรามีชุดข้อมูลที่มีข้อมูลอยู่ 10 ค่า ผลลัพธ์ของการโยนเหรียญออกหัวหรือก้อยก็คือจำนวนกลุ่มของข้อมูล ดังนั้นจากตัวอย่างการโยนเหรียญนี้ กลุ่มที่ออกหัวมี 5 ค่าเท่ากับกลุ่มที่ออกก้อย จะเห็นว่าข้อมูลชุดนี้มีความไม่เป็นระเบียบสูงมาก เพราะข้อมูลมีการกระจายแยกออกจากกันเป็น 2 กลุ่ม กลุ่มละเท่า ๆ กัน ดังนั้นค่าเอนโทรปีของการโยนเหรียญจะมีค่ามาก เมื่อเราจะโยนเหรียญครั้งต่อไป เราจึงไม่สามารถยืนยันได้ว่าเหรียญจะออกหัวหรือก้อย นั้นก็เพราะค่าเอนโทรปีหรือความไม่แน่นอนของข้อมูลมีค่ามากนั้นเอง ตามทฤษฎีแล้วค่าเอนโทรปีจะมีค่าสูงสุดเมื่อความน่าจะเป็นที่จะเกิดผลลัพธ์ในแต่ละผลลัพธ์มีค่าเท่ากันและค่าเอนโทรปีสูงสุดจะเท่ากับ $\log_2(N)$ เมื่อ N คือจำนวนของผลลัพธ์ที่เป็นไปได้ทั้งหมด ค่าเอนโทรปีมีสมการดังนี้

$$H(S) = -\sum_{n \in N} p(x_n) \log_2 p(x_n) \qquad (2.1)$$

H(S) คือค่าเอนโทรปี

N คือเซตของผลลัพธ์ที่เป็นไปได้ทั้งหมด

$p(x_n)$ คือความน่าจะเป็นที่จะเกิดผลลัพธ์ n หรือจำนวนข้อมูลที่จะเกิดผลลัพธ์ n หารด้วยจำนวนข้อมูลทั้งหมด

จากตัวอย่างการโยนเหรียญ ถ้าเราโยนเหรียญแบบปกติโดยไม่มีการถ่วงน้ำหนักที่เหรียญความน่าจะเป็นที่เหรียญออกหัวจะเท่ากับค่าความน่าจะเป็นที่เหรียญออกก้อยดังนั้นจะได้ค่าเอนโทรปีสูงสุดเท่ากับ 1 บิต ในกรณีที่เรามีการถ่วงน้ำหนักให้เหรียญออกหัวเท่านั้นไม่ว่าจะโยนกี่ครั้งเหรียญจะออกหัวเสมอ ค่าเอนโทรปีที่ได้จะเท่ากับ 0 บิต เพราะมีแค่ผลลัพธ์เดียวที่จะเกิดขึ้น เมื่อมีผลลัพธ์เดียวก็หมายความว่าข้อมูลมีกลุ่มเดียวและไม่มีการแบ่งแยกหรือกระจายออกไป ดังนั้นค่าเอนโทรปีหรือค่าความไม่เป็นระเบียบของข้อมูลจะเท่ากับ 0 ในกรณีต่อไปถ้าเราลองลดน้ำหนักของตัวถ่วงน้ำหนักลงเล็กน้อย ทำให้โอกาสที่เหรียญจะออกก้อยมีเพิ่มขึ้นเล็กน้อยจากที่ตอนแรกไม่มีโอกาสเลย เมื่อโยนเหรียญ 10 ครั้ง พบว่าเหรียญออกหัว 9 ครั้ง ออกก้อยแค่ 1 ครั้ง เมื่อคำนวณค่าเอนโทรปีของกรณีนี้

จะได้ประมาณ 0.47 บิต ค่าเอนโทรปีจะมากกว่ากรณีที่ผลลัพธ์ออกหัวอย่างเดียว (Entropy = 0) ขึ้นมาเล็กน้อยเพราะข้อมูลส่วนใหญ่ค่อนข้างจะเป็นระเบียบแล้ว มีแค่ 1 ค่าเท่านั้นที่แยกไปอยู่อีกกลุ่ม

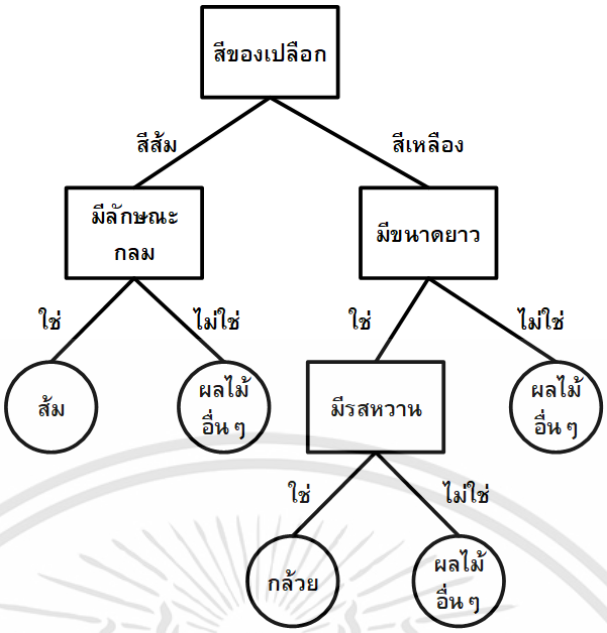
ตารางที่ 2.1 แสดงค่าเอนโทรปีของตัวอย่างผลลัพธ์ในการโยนเหรียญ 10 ครั้ง

จำนวนครั้งที่เหรียญออกหัว	จำนวนครั้งที่เหรียญออกก้อย	ค่าเอนโทรปี (บิต)
10	0	0
9	1	0.47
8	2	0.72
7	3	0.89
6	4	0.97
5	5	1
6	4	0.97
7	3	0.89
8	2	0.72
9	1	0.47
10	0	0

จากตารางที่ 2.1 แสดงให้เห็นว่ายิ่งข้อมูลมีความไม่เป็นระเบียบหรือมีการกระจายมาก ค่าเอนโทรปีจะยิ่งมากขึ้นเรื่อย ๆ และค่าเอนโทรปีจะมากที่สุดเมื่อข้อมูลกระจายออกในลักษณะที่เท่ากันในทุก ๆ กลุ่ม ดังเช่นในกรณีที่เราโยนเหรียญปกติโดยไม่มีการถ่วงน้ำหนัก 10 ครั้ง ผลลัพธ์ที่ได้คือเหรียญออกหัวและก้อยเท่ากันอย่างละ 5 ครั้ง ดังนั้นความน่าจะเป็นที่เหรียญออกหัวจะเท่ากับความน่าจะเป็นที่เหรียญออกก้อย จึงส่งผลให้ค่าเอนโทรปีมีค่าสูงสุด ในทางกลับกันถ้าลองถ่วงน้ำหนักเพื่อให้เหรียญออกก้อย แล้วเพิ่มน้ำหนักที่ถ่วงไปเรื่อย ๆ จนเหรียญออกก้อยในทุก ๆ ครั้ง (แถวสุดท้ายในตารางที่ 2.1) จากค่าเอนโทรปีสูงสุด (แถวที่ 7 ในตารางที่ 2.1) จะค่อย ๆ ลดลงเรื่อย ๆ จนถึงค่าต่ำสุดคือ 0

2.2 การเรียนรู้ต้นไม้ตัดสินใจ (Decision tree learning)

การเรียนรู้ต้นไม้ตัดสินใจเป็นวิธีการทั่วไปที่ใช้ในดาต้าไมนิง (Data mining) มีเป้าหมายเพื่อสร้างแบบจำลองในรูปแบบของต้นไม้ตัดสินใจ (Decision tree) ที่ใช้ในการทำนายผลลัพธ์ของตัวแปรเป้าหมาย (Target variable) จากตัวแปรที่รับเข้ามาหรือตัวแปรอินพุตหลาย ๆ ตัว (Input variable) ดังตัวอย่างในรูปที่ 2.1



รูปที่ 2.1 แสดงตัวอย่างต้นไม้ตัดสินใจสำหรับจำแนกประเภท ส้ม กล้วย และผลไม้อื่น ๆ

จากรูปที่ 2.1 โหนดที่เป็นรูปสี่เหลี่ยมหมายถึงตัวแปรอินพุตที่รับเข้ามาหลาย ๆ ค่าหรือในทางการจำแนกประเภทข้อมูลจะหมายถึงโหนดที่ถูกระบุด้วยแอตทริบิวต์ (Attribute node) ได้แก่ สีของเปลือก (มีค่าที่ไม่ซ้ำกันคือ “สีส้ม” และ “สีเหลือง”), มีลักษณะกลม (มีค่าที่ไม่ซ้ำกันคือ “ใช่” และ “ไม่ใช่”), มีขนาดยาว (มีค่าที่ไม่ซ้ำกันคือ “ใช่” และ “ไม่ใช่”) และมีรสหวาน (มีค่าที่ไม่ซ้ำกันคือ “ใช่” และ “ไม่ใช่”) โหนดที่อยู่ด้านบนสุด (สีของเปลือก) ถูกเรียกว่าโหนดราก (Root node) โหนดวงกลมหมายถึงตัวแปรเป้าหมายหรือในทางการจำแนกประเภทข้อมูลจะหมายถึงโหนดคำตอบ (Leaf node) ที่ถูกระบุด้วยค่าของแอตทริบิวต์เป้าหมาย (Target attribute) แต่ละค่าของแอตทริบิวต์เป้าหมายได้แก่ “ส้ม” “กล้วย” และ “ผลไม้อื่น ๆ” จะถูกเรียกว่าคลาส (Class) จากต้นไม้ตัดสินใจในรูปที่ 2.1 สามารถนำมาเขียนใหม่ในรูปแบบของกฎการตัดสินใจ (Decision rule) ดังนี้

กฎที่ 1 : ถ้า สีของเปลือก = สีส้ม และ มีลักษณะกลม = ใช่ ดังนั้นผลลัพธ์ = ส้ม

กฎที่ 2 : ถ้า สีของเปลือก = สีส้ม และ มีลักษณะกลม = ไม่ใช่ ดังนั้นผลลัพธ์ = ผลไม้อื่น ๆ

กฎที่ 3 : ถ้า สีของเปลือก = สีเหลือง และ มีขนาดยาว = ใช่ และ มีรสหวาน = ใช่ ดังนั้น ผลลัพธ์ = กล้วย

กฎที่ 4 : ถ้า สีของเปลือก = สีเหลือง และ มีขนาดยาว = ใช่ และ มีรสหวาน = ไม่ใช่ ดังนั้น ผลลัพธ์ = ผลไม้อื่น ๆ

กฎที่ 5 : ถ้า สีของเปลือก = สีเหลือง และ มีขนาดยาว = ไม่ใช่ ดังนั้นผลลัพธ์ = ผลไม้อื่น ๆ

จากตัวอย่างการแปลงต้นไม้ตัดสินใจในรูปที่ 2.1 มาเป็นกฎการตัดสินใจ สังเกตุได้ว่าจำนวนกฎการตัดสินใจจะเท่ากับจำนวนเส้นทางทั้งหมดตั้งแต่โหนดรากลงมาถึงโหนดคำตอบ ในแต่ละเส้นทางจะถูกนับเป็น 1 กฎการตัดสินใจ ดังนั้นจะเห็นได้จากต้นไม้ตัดสินใจในรูปที่ 2.1 มีกฎการ

ตัดสินใจจำนวน 5 กฎ ก่อนที่จะสร้างแบบจำลองต้นไม้ตัดสินใจเพื่อจำแนกผลลัพธ์ได้นั้น จำเป็นต้องมีการเก็บข้อมูลของกรณีตัวอย่างจำนวนหนึ่งที่จะส่งผลให้ผลลัพธ์แตกต่างกันออกไป ยกตัวอย่างเช่น ตารางที่ 2.2 แสดงตัวอย่างชุดข้อมูลในกรณีต่าง ๆ ของทุก ๆ แอตทริบิวต์รวมถึงผลลัพธ์หรือคลาสที่ได้ในแต่ละกรณีตัวอย่าง

ตารางที่ 2.2 แสดงตัวอย่างชุดข้อมูลในการจำแนกผลไม้ระหว่างส้ม กล้วยและผลไม้อื่น ๆ

กรณีตัวอย่าง	แอตทริบิวต์				แอตทริบิวต์เป้าหมาย
	สีของเปลือก	ลักษณะกลม	ลักษณะยาว	รสหวาน	
1	สีส้ม	ใช่	ไม่ใช่	ไม่ใช่	ส้ม
2	สีส้ม	ไม่ใช่	ไม่ใช่	ไม่ใช่	ผลไม้อื่น ๆ
3	สีเหลือง	ไม่ใช่	ใช่	ใช่	กล้วย
4	สีเหลือง	ไม่ใช่	ใช่	ไม่ใช่	ผลไม้อื่น ๆ
5	สีเหลือง	ไม่ใช่	ไม่ใช่	ใช่	ผลไม้อื่น ๆ

จากตารางที่ 2.2 แสดงตัวอย่างชุดข้อมูลที่จะนำไปใช้สร้างแบบจำลองต้นไม้ตัดสินใจเพื่อจำแนกประเภทข้อมูล มีลักษณะเป็นค่าไม่ที่ไม่ต่อเนื่อง (Discrete value)

2.3 อัลกอริทึมไอดีทรี (ID3 หรือ Iterative Dichotomiser 3)

อัลกอริทึม ID3 [1] เป็นหนึ่งในอัลกอริทึมที่ใช้สร้างต้นไม้ตัดสินใจที่ใช้กันอย่างแพร่หลายในทางการจำแนกประเภทข้อมูล ต้นไม้ตัดสินใจผลลัพธ์ที่ได้จากอัลกอริทึม ID3 จะถูกนำไปทดสอบประสิทธิภาพโดยใช้กรณีตัวอย่างอื่น ๆ ที่นอกเหนือจากกรณีตัวอย่างที่ใช้ในการสร้างต้นไม้ตัดสินใจ กรณีตัวอย่างที่ใช้ในการสร้างต้นไม้ตัดสินใจจะถูกเรียกว่าชุดข้อมูลสำหรับฝึกฝน (Training set) ส่วนกรณีตัวอย่างส่วนที่ใช้ในการทดสอบประสิทธิภาพของต้นไม้ตัดสินใจจะเรียกว่าชุดข้อมูลสำหรับทดสอบ (Test set) อัลกอริทึม ID3 ใช้สำหรับจำแนกชุดข้อมูลที่เป็นค่าไม่ต่อเนื่องดังเช่นชุดข้อมูลตัวอย่างในตารางที่ 2.2 และใช้การคำนวณเกนความรู้เป็นเกณฑ์ในการเลือกแอตทริบิวต์ แอตทริบิวต์ใดมีค่าเกนความรู้สูงสุดจะถูกเลือกเพื่อนำมาสร้างโหนดของต้นไม้ตัดสินใจ

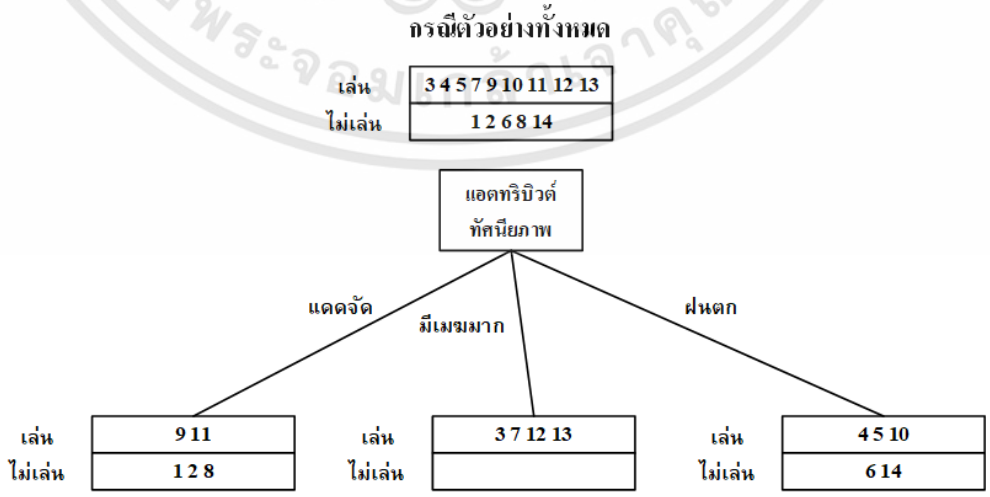
2.3.1 เกนความรู้

เกนความรู้หรือ Information gain คือค่าที่บ่งบอกว่าแอตทริบิวต์ใดสามารถจำแนกประเภทกรณีตัวอย่างหรือการระบุคลาสคำตอบได้มากที่สุด และแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดคือแอตทริบิวต์ที่ดีที่สุดและจะถูกเลือกเพื่อนำไปสร้างโหนด เห็นได้ชัดว่าการคำนวณเกนความรู้เป็นการค้นหาแบบละโมบ (Greedy search) เพราะแอตทริบิวต์ใดสามารถระบุคลาสคำตอบได้มากที่สุดจะถูกเลือกนำมาสร้างโหนดก่อนเสมอ เพื่อความเข้าใจมากขึ้นเราจะพิจารณาแอตทริบิวต์ 2 แอตทริบิวต์ จากชุดข้อมูลการเล่นเทนนิสใน 14 วัน

ตารางที่ 2.3 แสดงชุดข้อมูลการตัดสินใจเล่นเทนนิสใน 14 วัน

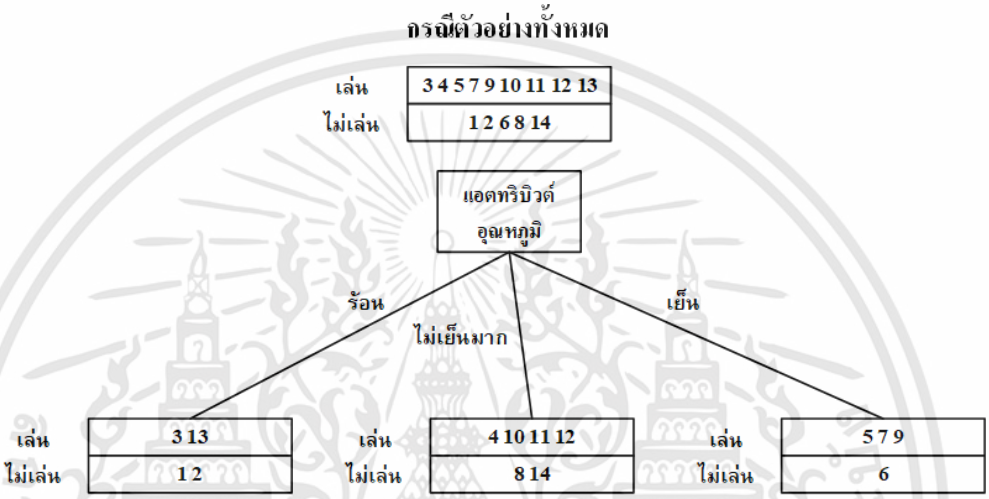
กรณีตัวอย่าง (วัน)	แอตทริบิวต์				แอตทริบิวต์เป้าหมาย
	ทัศนียภาพ	อุณหภูมิ	ความชื้น	ลมแรง	
1	แดดจัด	ร้อน	สูง	ไม่ใช่	ไม่เล่น
2	แดดจัด	ร้อน	สูง	ใช่	ไม่เล่น
3	มีเมฆมาก	ร้อน	สูง	ไม่ใช่	เล่น
4	ฝนตก	ไม่เย็นมาก	สูง	ไม่ใช่	เล่น
5	ฝนตก	เย็น	ปกติ	ไม่ใช่	เล่น
6	ฝนตก	เย็น	ปกติ	ใช่	ไม่เล่น
7	มีเมฆมาก	เย็น	ปกติ	ใช่	เล่น
8	แดดจัด	ไม่เย็นมาก	สูง	ไม่ใช่	ไม่เล่น
9	แดดจัด	เย็น	ปกติ	ไม่ใช่	เล่น
10	ฝนตก	ไม่เย็นมาก	ปกติ	ไม่ใช่	เล่น
11	แดดจัด	ไม่เย็นมาก	ปกติ	ใช่	เล่น
12	มีเมฆมาก	ไม่เย็นมาก	สูง	ใช่	เล่น
13	มีเมฆมาก	ร้อน	ปกติ	ไม่ใช่	เล่น
14	ฝนตก	ไม่เย็นมาก	สูง	ใช่	ไม่เล่น

จากตารางที่ 2.3 เมื่อเราลองนำแอตทริบิวต์ทัศนียภาพและอุณหภูมิ มาทดลองทำเป็นโหนดรากและแบ่งกรณีตัวอย่างออกไปตามแต่ละค่าที่ไม่ซ้ำกันของทั้งสองแอตทริบิวต์ดังในรูปที่ 2.2 และ 2.3 ซึ่งแอตทริบิวต์ทัศนียภาพมีค่าที่ไม่ซ้ำกันได้แก่ “แดดจัด”, “มีเมฆมาก” และ “ฝนตก” แอตทริบิวต์อุณหภูมิมีค่าที่ไม่ซ้ำกันได้แก่ “ร้อน”, “ไม่เย็นมาก” และ “เย็น” แอตทริบิวต์เป้าหมายมี 2 คลาสได้แก่ “เล่น” และ “ไม่เล่น”



รูปที่ 2.2 แสดงการแบ่งกรณีตัวอย่างทั้งหมดของแอตทริบิวต์ทัศนียภาพ

จากรูปที่ 2.2 แอตทริบิวต์ทัศนียภาพมีกรณีตัวอย่างที่เป็นค่าแรกคือ “แดดจัด” ได้แก่กรณีที่ 1, 2, 8, 9 และ 11 ภายในกรณีตัวอย่างที่มีค่า “แดดจัด” นี้เป็นคลาส “เล่น” คือกรณีที่ 9 และ 11 เป็นคลาส “ไม่เล่น” คือกรณีที่ 1, 2 และ 8 กรณีตัวอย่างที่เป็นค่าที่สองคือ “มีเมฆมาก” ได้แก่กรณีที่ 3, 7, 12 และ 13 เป็นคลาส “เล่น” ทั้งหมด และกรณีตัวอย่างที่เป็นค่าสุดท้ายของแอตทริบิวต์ทัศนียภาพคือ “ฝนตก” ได้แก่กรณีที่ 4, 5, 6, 10 และ 14 เป็นคลาส “เล่น” คือกรณีที่ 4, 5 และ 10 เป็นคลาส “ไม่เล่น” คือกรณีที่ 6 และ 14



รูปที่ 2.3 แสดงการแบ่งกรณีตัวอย่างทั้งหมดของแอตทริบิวต์อุณภูมิ

จากรูปที่ 2.3 แอตทริบิวต์อุณภูมิมีกรณีตัวอย่างที่เป็นค่าแรกคือ “ร้อน” ได้แก่กรณีที่ 1, 2, 3 และ 13 เป็นคลาส “เล่น” คือกรณีที่ 3 และ 13 เป็นคลาส “ไม่เล่น” คือกรณีที่ 1 และ 2 กรณีตัวอย่างที่เป็นค่าที่สองคือ “ไม่เย็นมาก” ได้แก่กรณีที่ 4, 8, 10, 11, 12 และ 14 เป็นคลาส “เล่น” คือกรณีที่ 4, 10, 11 และ 12 เป็นคลาส “ไม่เล่น” คือกรณีที่ 8 และ 14 กรณีตัวอย่างที่เป็นค่าสุดท้ายของแอตทริบิวต์อุณหภูมิคือ “เย็น” ได้แก่กรณีที่ 5, 6, 7 และ 9 เป็นคลาส “เล่น” คือกรณีที่ 5, 7 และ 9 เป็นคลาส “ไม่เล่น” คือกรณีที่ 6

เมื่อพิจารณารวมกันทั้งรูปที่ 2.2 และ 2.3 พบว่าเมื่อเราลองนำแอตทริบิวต์ทัศนียภาพและอุณหภูมิมาทดสอบการแบ่งกรณีตัวอย่าง แอตทริบิวต์ทัศนียภาพสามารถจำแนกคลาสคำตอบได้มากกว่าแอตทริบิวต์อุณหภูมิ จากกรณีตัวอย่างทั้งหมดก่อนแบ่งมีกรณีตัวอย่างที่ตอบ “เล่น” 9 กรณี (3, 4, 5, 7, 9, 10, 11, 12, 13) และกรณีตัวอย่างที่ตอบ “ไม่เล่น” 5 กรณี (1, 2, 6, 8, 14) แอตทริบิวต์ทัศนียภาพสามารถระบุคำตอบหรือจำแนกได้ 4 กรณีตัวอย่าง (3, 7, 12, 13) ในค่า “มีเมฆมาก” เพราะกรณีตัวอย่างทั้งหมดตอบเป็นคลาสเดียวกันคือ “เล่น” ทำให้เหลือกรณีตัวอย่างที่ยังไม่สามารถจำแนกได้ที่เป็นคลาส “เล่น” 5 กรณี (9, 11, 4, 5, 10) เป็นคลาส “ไม่เล่น” 5 กรณี (1, 2, 6, 8, 14) จะเห็นว่ากรณีตัวอย่างที่ตอนแรกจำแนกไม่ได้ในคลาส “เล่น” มี 9 กรณี พอนำ

แอตทริบิวต์ทัศนียภาพเข้ามาแบ่งกรณีตัวอย่าง พบว่าเหลือกรณีตัวอย่างที่ไม่สามารถจำแนกได้ในคลาส “เล่น” เพียง 5 กรณี ส่วนคลาส “ไม่เล่น” เหลือ 5 กรณีเท่าเดิมจากตอนก่อนแบ่ง ในส่วนของแอตทริบิวต์อุณหภูมิเมื่อแบ่งกรณีตัวอย่างไปตามค่าที่ไม่ซ้ำกัน พบว่าในแต่ละค่าของมันยังไม่สามารถจำแนกกรณีตัวอย่างได้เลย ทำให้เหลือสัดส่วนของกรณีตัวอย่างที่จำแนกไม่ได้เท่ากับกรณีตัวอย่างในตอนแรก ก็คือยังเหลือกรณีตัวอย่างที่ตอบคลาส “เล่น” 9 กรณีและคลาส “ไม่เล่น” 5 กรณีเท่าเดิมกับตอนก่อนแบ่ง ดังนั้นแอตทริบิวต์ทัศนียภาพจึงถือว่าเป็นแอตทริบิวต์ที่ดี ถึงแม้จะไม่สามารถจำแนกคำตอบได้ทั้งหมด แต่ยังดีกว่าแอตทริบิวต์อุณหภูมิซึ่งไม่สามารถจำแนกคำตอบได้เลย จากเหตุการณ์นี้แอตทริบิวต์ทัศนียภาพจะมีค่าเกนความรู้มากกว่าแอตทริบิวต์อุณหภูมิ เพราะสามารถจำแนกประเภทกรณีตัวอย่างได้มากกว่า และถ้าแอตทริบิวต์ทัศนียภาพสามารถจำแนกกรณีตัวอย่างได้มากที่สุดเมื่อเทียบกับแอตทริบิวต์อีก 2 แอตทริบิวต์ที่เหลือ (แอตทริบิวต์ความชื้น, แอตทริบิวต์ลมแรง) มันจะมีค่าเกนความรู้สูงสุด และจะถูกนำไปสร้างเป็นโหนดรากของต้นไม้ตัดสินใจ เพื่อความชัดเจนให้พิจารณาสมการเกนความรู้ซึ่งอ้างอิงจากสมการเอนโทรปีโดยมีสมการดังนี้

$$IG(I,A) = Entropy(I) - Remainder(I,A) \tag{2.2}$$

IG(I,A) คือค่าเกนความรู้ของ A ซึ่ง A คือแอตทริบิวต์ใด ๆ

I คือเซตของกรณีตัวอย่างที่เหลืออยู่

Entropy(I) คือค่าเอนโทรปีของกรณีตัวอย่างที่เหลืออยู่ก่อนที่จะถูกแบ่งไปตามแอตทริบิวต์ A โดยสมการจะมีลักษณะเหมือนกับสมการที่ 2.1 หรือสมการเอนโทรปีพื้นฐาน

Remainder(I,A) คือผลรวมของค่าเอนโทรปีของแต่ละค่าที่ไม่ซ้ำกันในแอตทริบิวต์ A หลังจากกรณีตัวอย่างถูกแบ่งไปตามค่าที่ไม่ซ้ำกันของ A คือจะเป็นค่าเอนโทรปีโดยรวมหลังการทดสอบแอตทริบิวต์ A เพื่อดูว่าหากแอตทริบิวต์ A ถูกเลือกไปสร้างเป็นโหนดของต้นไม้ตัดสินใจจะเหลือกรณีตัวอย่างที่ยังจำแนกไม่ได้เป็นเท่าใด ซึ่งก็ดูที่ค่า Remainder(I,A) ถ้าค่า Remainder ของแอตทริบิวต์ A มีค่ามาก หมายความว่ามีความไม่แน่นอนมากหรือหลังการทดสอบแอตทริบิวต์ A เหลือกรณีตัวอย่างที่ยังจำแนกไม่ได้อยู่มาก

$$Entropy(I) = -\sum_{c \in C} p(x_c) \log_2 p(x_c) \tag{2.3}$$

C คือเซตของคลาสคำตอบทั้งหมด

$p(x_c)$ คืออัตราส่วนของจำนวนกรณีตัวอย่างที่ตอบคลาส c ต่อจำนวนกรณีตัวอย่างทั้งหมดในเซต I

$$Remainder(I,A) = \sum_{v \in V} p(v) Entropy(v) \tag{2.4}$$

V คือเซตของ value หรือค่าที่ไม่ซ้ำกันของแอตทริบิวต์ A

p(v) คืออัตราส่วนของจำนวนกรณีตัวอย่างที่เป็นค่า v ในแอตทริบิวต์ A ต่อจำนวนกรณีตัวอย่างทั้งหมดในเซต I

Entropy(v) คือค่าเอนโทรปีของกรณีตัวอย่างที่มีค่า v ในแอตทริบิวต์ A

$$Entropy(v) = -\sum_{c \in C} p(x_{vc}) \log_2 p(x_{vc}) \quad \textbf{(2.5)}$$

C คือเซตของคลาสคำตอบทั้งหมด

$p(x_{vc})$ คืออัตราส่วนของจำนวนกรณีตัวอย่างที่เป็นค่า v ในแอตทริบิวต์ A และมีคลาสคำตอบเป็น c ต่อจำนวนกรณีตัวอย่างที่เป็นค่า v ในแอตทริบิวต์ A

จากสมการที่ 2.4 ที่ต้องนำอัตราส่วนของค่านั้น ๆ หรือ p(v) คูณเข้าไปใน Entropy(v) ก่อนที่จะนำค่าเอนโทรปีไปรวมกันนั้น เพราะว่าหากเรานำค่าเอนโทรปีของค่าที่ไม่ซ้ำกันต่าง ๆ ใน A มารวมกันโดยไม่คูณกับ p(v) เราจะได้ค่า Remainder(I,A) ที่มากกว่าค่า Entropy(I) มาก ๆ ดังนั้นเพื่อความถูกต้องจึงจำเป็นต้องคูณอัตราส่วนของค่านั้น ๆ กับค่าเอนโทรปีของมันเข้าไปด้วย แล้วจึงนำมารวมกันจะได้เป็นค่า Remainder(I,A) ตัวอย่างการคำนวณเกนความรู้ในแอตทริบิวต์ทัศนียภาพและแอตทริบิวต์อุณหภูมิจากชุดข้อมูลการเล่นเทนนิสในตารางที่ 2.3

คำนวณค่าเอนโทรปีของกรณีตัวอย่างก่อนที่จะถูกแบ่งตามแอตทริบิวต์ใด ๆ

Entropy(I) = - [(9/14 x $\log_2$9/14) + (5/14 x $\log_2$5/14)] = 0.94

คำนวณค่าเกนความรู้ของแอตทริบิวต์ทัศนียภาพ

Remainder(ทัศนียภาพ) = [p(แดดจัด) x Entropy(แดดจัด)] + [p(มีเมฆมาก) x Entropy(มีเมฆมาก)] + [p(ฝนตก) x Entropy(ฝนตก)]

Entropy(แดดจัด) = - [(2/5 x $\log_2$2/5) + (3/5 x $\log_2$3/5)] = 0.97

Entropy(มีเมฆมาก) = - [(4/4 x $\log_2$4/4) + (0/4 x $\log_2$0/4)] = 0

Entropy(ฝนตก) = - [(3/5 x $\log_2$3/5) + (2/5 x $\log_2$2/5)] = 0.97

ดังนั้น

Remainder(ทัศนียภาพ) = [5/14 x 0.97] + [4/14 x 0] + [5/14 x 0.97] = 0.69

IG (I,ทัศนียภาพ) = 0.94 – 0.69 = 0.25

คำนวณค่าเกนความรู้ของแอตทริบิวต์อุณหภูมิ

Remainder(อุณหภูมิ) = [p(ร้อน) x Entropy(ร้อน)] + [p(ไม่เย็นมาก) x Entropy(ไม่เย็นมาก)] + [p(เย็น) x Entropy(เย็น)]

Entropy(ร้อน) = - [(2/4 x $\log_2$2/4) + (2/4 x $\log_2$2/4)] = 1

Entropy(ไม่เย็นมาก) = - [(4/6 x $\log_2$4/6) + (2/6 x $\log_2$2/6)] = 0.91

Entropy(เย็น) = - [(3/4 x $\log_2$3/4) + (1/4 x $\log_2$1/4)] = 0.81

ดังนั้น

Remainder(อุณหภูมิ) = [4/14 x 1] + [6/14 x 0.91] + [4/14 x 0.81] = 0.91

IG (I,อุณหภูมิ) = 0.94 – 0.91 = 0.03

ค่าเกนความรู้ของแอตทริบิวต์ทั้งหมดจากชุดข้อมูลการเล่นเทนนิสใน 14 วัน

IG (I,ทัศนียภาพ) = 0.25

IG (I,อุณหภูมิ) = 0.03

IG (I,ลมแรง) = 0.05

IG (I,ความชื้น) = 0.03

2.3.2 ขั้นตอนของอัลกอริทึม ID3

ดังที่ได้กล่าวไว้ว่าอัลกอริทึม ID3 จะคำนวณเกนความรู้ของแต่ละแอตทริบิวต์และแอตทริบิวต์ใดมีค่าเกนความรู้มากที่สุดจะถูกเลือกนำมาสร้างโหนดของต้นไม้ตัดสินใจ เมื่อทำการเลือกแอตทริบิวต์ได้แล้วกรณีตัวอย่างปัจจุบันจะถูกแบ่งไปตามค่าที่ไม่ซ้ำกันของแอตทริบิวต์นั้น ๆ และจะทำกระบวนการซ้ำ ๆ ลงไปเรื่อย ๆ จนกว่าจะสามารถจำแนกกรณีตัวอย่างได้ทั้งหมด ขั้นตอนคร่าว ๆ ของอัลกอริทึม ID3 มีดังนี้

ขั้นตอนที่ 1 : คำนวณค่าเกนความรู้ของทุก ๆ แอตทริบิวต์โดยใช้สมการเกนความรู้ดังสมการที่ 2.2

ขั้นตอนที่ 2 : เลือกแอตทริบิวต์ที่มีค่าเกนความรู้มากที่สุด เพื่อสร้างโหนดปัจจุบันของต้นไม้ตัดสินใจ ถ้ามีแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดมากกว่า 1 แอตทริบิวต์ ให้ทำการสุ่มเลือกมา 1 ตัว

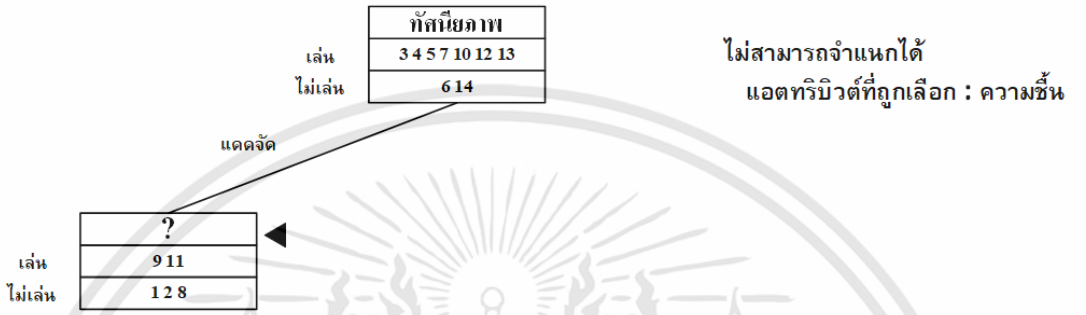
ขั้นตอนที่ 3 : แบ่งกรณีตัวอย่างไปตามค่าที่ไม่ซ้ำกันของแอตทริบิวต์ที่ถูกเลือกจากขั้นตอนที่ 2

ขั้นตอนที่ 4 : ทำซ้ำขั้นตอนที่ 1-3 จนกระทั่งสามารถจำแนกประเภทกรณีตัวอย่างได้ทั้งหมด เพื่อให้ง่ายต่อความเข้าใจให้พิจารณาภาพตัวอย่างของการสร้างต้นไม้ตัดสินใจจากชุดข้อมูลการเล่นเทนนิสในตารางที่ 2.3

รูปที่ 2.4 แสดงภาพที่ 1 ของการสร้างต้นไม้ตัดสินใจจากชุดข้อมูลการเล่นเทนนิส

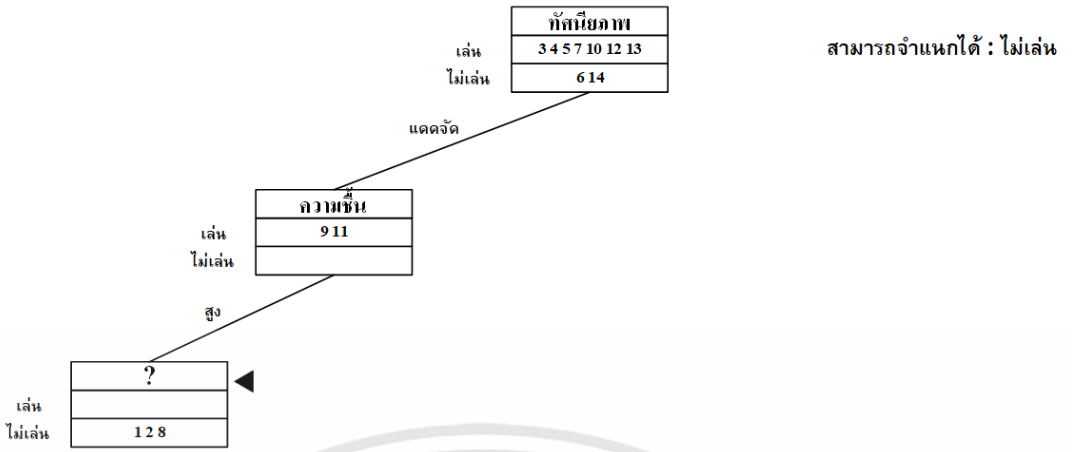
จากรูปที่ 2.4 คือกรณีตัวอย่างทั้งหมดของชุดข้อมูลการเล่นเทนนิสซึ่งถูกจัดกลุ่มตามคลาสคือ คลาส “เล่น” และ คลาส “ไม่เล่น” สามเหลี่ยมทึบสีดำบ่งบอกถึงว่าเรากำลังพิจารณากรณีตัวอย่างที่โหนดใด การพิจารณากรณีตัวอย่างคือการดูว่ากรณีตัวอย่างในโหนดนั้นสามารถจำแนกได้หรือไม่ การที่จะจำแนกได้คือกรณีตัวอย่างทุกตัวในโหนดต้องอยู่กลุ่มคลาสเดียวกันถึงจะสามารถระบุ

เป็นโหนดคำตอบได้ ถ้ายังไม่สามารถจำแนกกรณีตัวอย่างในโหนดได้ ต้องหาแอตทริบิวต์ที่สำคัญที่สุดมาแบ่งกรณีตัวอย่างลงไปในชั้นถัดไปโดยการคำนวณค่าเกนความรู้ จากการคำนวณค่าเกนความรู้ของแอตทริบิวต์ทั้งหมด แอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดหรือแอตทริบิวต์ที่ถูกเลือกได้แก่ แอตทริบิวต์ทัศนียภาพ ดังนั้นในโหนดรากจะถูกระบุด้วยแอตทริบิวต์ทัศนียภาพและหลังจากโหนดปัจจุบันแบ่งกรณีตัวอย่างต่อไปแล้วสามเหลี่ยมทึบก็จะย้ายไปยังโหนดถัดไปดังรูปที่ 2.5



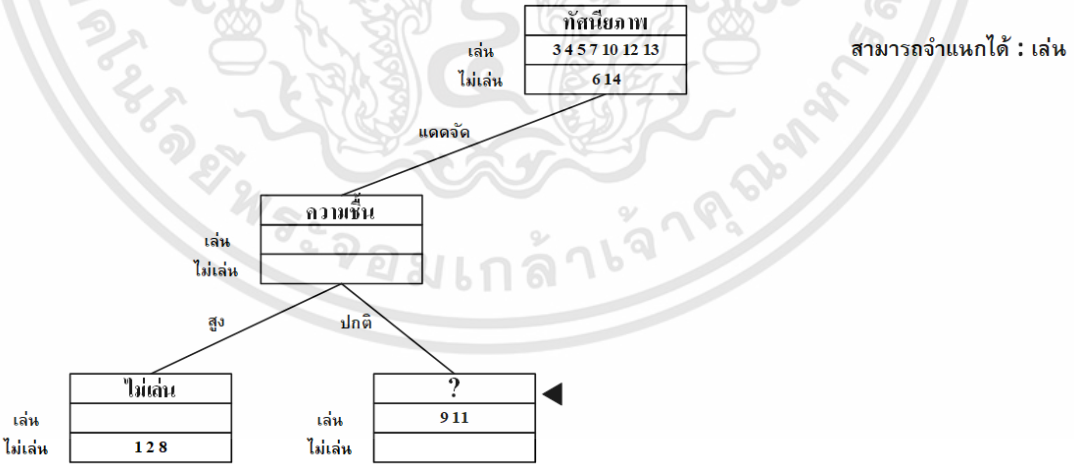
รูปที่ 2.5 แสดงภาพที่ 2 ของการสร้างต้นไม้ตัดสินใจจากชุดข้อมูลการเล่นเทนนิส

จากรูปที่ 2.5 หลังจากโหนดรากถูกระบุด้วยแอตทริบิวต์ทัศนียภาพ ต่อมาจึงแบ่งกรณีตัวอย่างไปตามค่าที่ไม่ซ้ำกันค่าแรกคือ “แดดจัด” และสร้างเป็นโหนดปัจจุบัน (แอตทริบิวต์ทัศนียภาพมีค่าที่ไม่ซ้ำกันคือ 1. “แดดจัด” 2. “มีเมฆมาก” และ 3. “ฝนตก”) โหนดปัจจุบันคือโหนดที่กำลังพิจารณาจะมีสัญญลักษณ์สามเหลี่ยมทึบสีดำอยู่ที่โหนดปัจจุบัน กรณีตัวอย่างที่ 1, 2, 8, 9 และ 11 คือกรณีตัวอย่างที่มีแอตทริบิวต์ทัศนียภาพเป็น “แดดจัด” ดังนั้นกรณีตัวอย่างเหล่านี้จะถูกแบ่งไปตามค่า “แดดจัด” และจะถูกย้ายลงมาจากโหนดราก (โหนดทัศนียภาพ) ลงมาอยู่ตามกลุ่มคลาสในโหนดปัจจุบัน เมื่อพิจารณาโหนดปัจจุบันก็ยังไม่สามารถจำแนกได้เพราะกรณีตัวอย่างกระจายออกเป็น 2 คลาส เมื่อยังไม่สามารถจำแนกได้จึงต้องคำนวณเกนความรู้ของแอตทริบิวต์อุณหภูมิ, แอตทริบิวต์ความชื้น และแอตทริบิวต์ลมแรง เราจะไม่นำแอตทริบิวต์ทัศนียภาพมาคำนวณอีกเพราะมันถูกใช้ไปแล้วในเชิงลึก แอตทริบิวต์ที่ถูกเลือกคือแอตทริบิวต์ความชื้นเพราะมีค่าเกนความรู้มากที่สุด ดังนั้นโหนดปัจจุบันก็จะถูกระบุด้วยแอตทริบิวต์ความชื้นและหลังจากโหนดปัจจุบันแบ่งกรณีตัวอย่างต่อไปแล้วสามเหลี่ยมทึบก็จะย้ายไปยังโหนดถัดไปดังรูปที่ 2.6



รูปที่ 2.6 แสดงภาพที่ 3 ของการสร้างต้นไม้ตัดสินใจจากชุดข้อมูลการเล่นเทนนิส

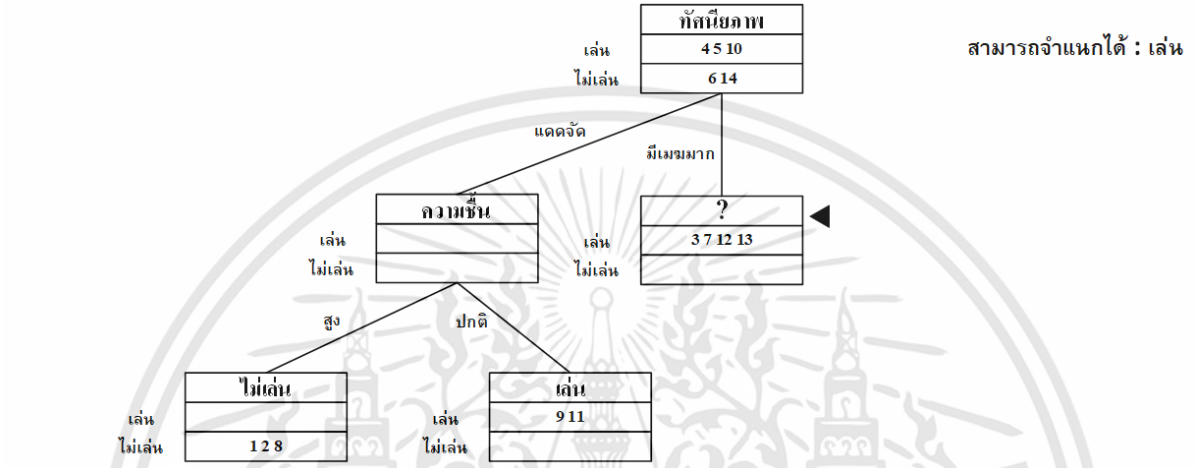
จากรูปที่ 2.6 หลังจากระบุแอตทริบิวต์ความชื้นถัดมาทำการแบ่งกรณีตัวอย่างไปตามค่าที่ไม่ซ้ำกันค่าแรกคือ “สูง” (แอตทริบิวต์ความชื้นมีค่าที่ไม่ซ้ำกันคือ 1. “สูง” และ 2. “ปกติ”) จากโหนดความชื้นกรณีตัวอย่างที่ 1, 2 และ 8 คือกรณีตัวอย่างที่มีแอตทริบิวต์ความชื้นเป็นค่า “สูง” ดังนั้นจึงย้ายจากโหนดความชื้นลงมาอยู่ที่โหนดปัจจุบัน พิจารณาโหนดปัจจุบันพบว่ากรณีตัวอย่างทั้งหมดอยู่ในคลาสเดียวกันจึงสามารถจำแนกเป็นโหนดคำตอบที่ระบุด้วยคลาส “ไม่เล่น” ได้ เมื่อสามารถจำแนกกรณีตัวอย่างในโหนดปัจจุบันได้แล้ว โหนดพ่อแม่ของโหนดปัจจุบันนั้นก็คือโหนดความชื้นก็จะแบ่งกรณีตัวอย่างตามค่าที่ไม่ซ้ำกันค่าต่อไปคือ “ปกติ” แล้วสามเหลี่ยมทึบก็จะย้ายไปยังโหนดถัดไปดังรูปที่ 2.7



รูปที่ 2.7 แสดงภาพที่ 4 ของการสร้างต้นไม้ตัดสินใจจากชุดข้อมูลการเล่นเทนนิส

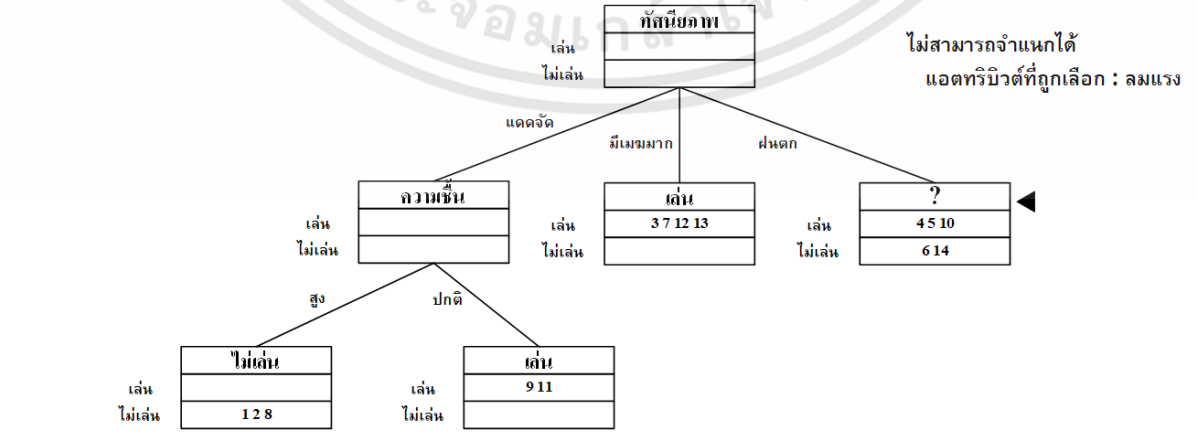
จากรูปที่ 2.7 กรณีตัวอย่างในโหนดความชื้นที่เหลือก็จะถูกแบ่งต่อไปตามค่าที่ไม่ซ้ำกันค่าที่สองคือ “ปกติ” เพราะกรณีตัวอย่างที่แบ่งไปตามค่าแรกคือ “สูง” สามารถจำแนกได้ทั้งหมดแล้ว กรณีตัวอย่างที่ 9 และ 11 ในโหนดความชื้นคือกรณีตัวอย่างที่มีแอตทริบิวต์ความชื้นเป็นค่า

“ปกติ” จึงย้ายลงมาอยู่ที่โหนดปัจจุบันและสามารถจำแนกได้เป็นคลาส “เล่น” เมื่อสามารถจำแนกกรณีตัวอย่างในโหนดปัจจุบันได้แล้ว โหนดพ่อแม่ของโหนดปัจจุบันนั้นก็คือโหนดความชื้นแบ่งกรณีตัวอย่างตามค่าที่ไม่ซ้ำกันหมดทุกค่าแล้ว ดังนั้นถึงต้องย้อนขึ้นไปยังโหนดพ่อแม่ของโหนดความชื้นอีกทีนั้นคือโหนดทัศนียภาพ โหนดทัศนียภาพจะแบ่งกรณีตัวอย่างตามค่าที่ไม่ซ้ำกันค่าต่อไปคือ “มีเมฆมาก” แล้วสามเหลี่ยมทึบก็จะย้ายไปยังโหนดถัดไปดังรูปที่ 2.8



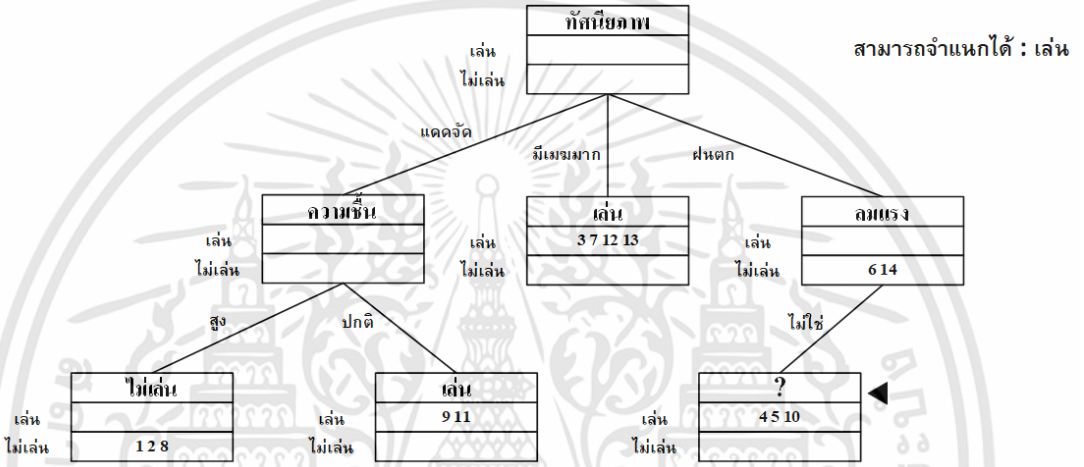
รูปที่ 2.8 แสดงภาพที่ 5 ของการสร้างต้นไม้ตัดสินใจจากชุดข้อมูลการเล่นเทนนิส

จากรูปที่ 2.8 กรณีตัวอย่างที่ถูกแบ่งไปตามค่าแรกของโหนดรากคือ “แดดจัด” มีการแบ่งลงไปเรื่อย ๆ จนจำแนกกรณีตัวอย่างได้ทั้งหมดแล้ว ดังนั้นจึงย้อนกลับมาแบ่งกรณีตัวอย่างตามค่าที่สองคือ “มีเมฆมาก”ของแอตทริบิวต์ทัศนียภาพต่อ เมื่อพิจารณากรณีตัวอย่างที่ถูกแบ่งลงมาที่โหนดปัจจุบัน พบว่าสามารถจำแนกได้เป็นคลาส “เล่น” เมื่อสามารถจำแนกกรณีตัวอย่างในโหนดปัจจุบันได้แล้ว โหนดพ่อแม่ของโหนดปัจจุบันนั้นก็คือโหนดทัศนียภาพจะแบ่งกรณีตัวอย่างตามค่าที่ไม่ซ้ำกันค่าต่อไปคือ “ฝนตก” แล้วสามเหลี่ยมทึบก็จะย้ายไปยังโหนดถัดไปดังรูปที่ 2.9



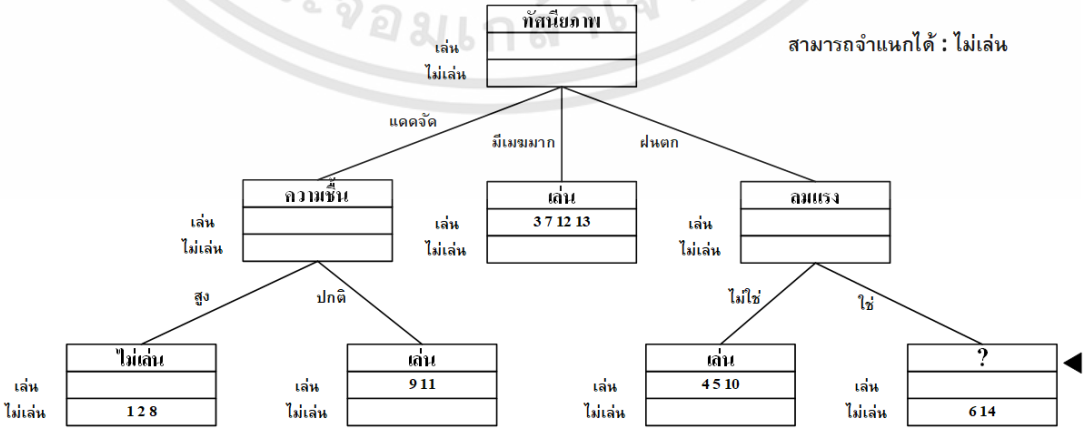
รูปที่ 2.9 แสดงภาพที่ 6 ของการสร้างต้นไม้ตัดสินใจจากชุดข้อมูลการเล่นเทนนิส

จากรูปภาพที่ 2.9 ทำการแบ่งกรณีตัวอย่างในโหนดทัศนียภาพไปตามค่าที่สามคือ “ฝนตก” พบว่าในโหนดปัจจุบันยังไม่สามารถจำแนกกรณีตัวอย่างทั้งหมดได้ จึงต้องคำนวณเกนความรู้ของแอตทริบิวต์อุณหภูมิ, ความชื้น และลมแรง แม้แอตทริบิวต์ความชื้นจะถูกใช้สร้างโหนดไปแล้วในกิ่ง (branch) “แดดจัด” ของแอตทริบิวต์ทัศนียภาพ แต่ก็ถือว่าอยู่คนละกิ่งกันหรือไม่ได้ถูกใช้ในเชิงลึกจึงต้องนำมาคำนวณต่อ แอตทริบิวต์ที่ถูกเลือกคือแอตทริบิวต์ลมแรงเพราะมีค่าเกนความรู้มากที่สุด ดังนั้นโหนดปัจจุบันก็จะถูกระบุด้วยแอตทริบิวต์ลมแรงและหลังจากโหนดปัจจุบันแบ่งกรณีตัวอย่างต่อไปแล้วสามเหลี่ยมทึบก็จะย้ายไปยังโหนดถัดไปดังรูปที่ 2.10



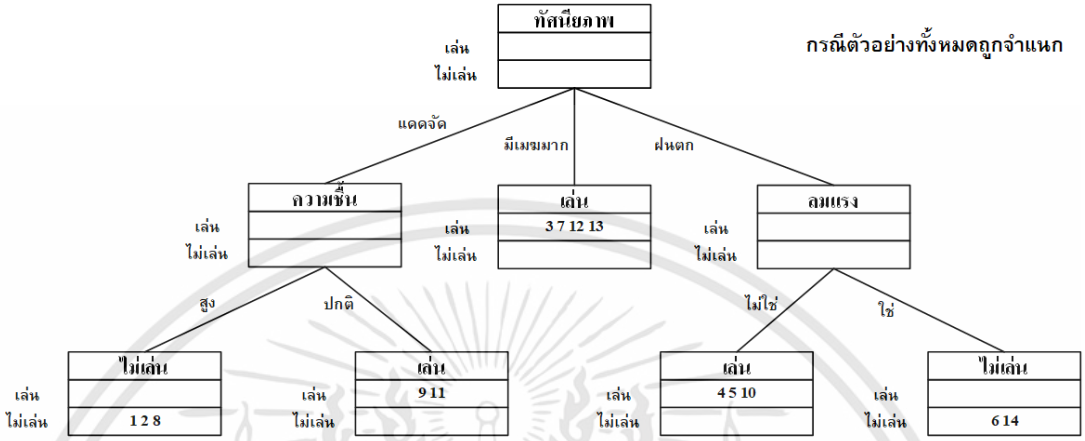
รูปที่ 2.10 แสดงภาพที่ 7 ของการสร้างต้นไม้ตัดสินใจจากชุดข้อมูลการเล่นเทนนิส

จากรูปที่ 2.10 ทำการแบ่งกรณีตัวอย่างไปตามค่าแรกของแอตทริบิวต์ลมแรงคือค่า “ไม่ใช่” พิจารณาโหนดปัจจุบันพบว่าสามารถจำแนกได้เป็นคลาส “เล่น” เมื่อสามารถจำแนกกรณีตัวอย่างในโหนดปัจจุบันได้แล้ว โหนดพ่อแม่ของโหนดปัจจุบันนั้นก็คือโหนดลมแรงจะแบ่งกรณีตัวอย่างตามค่าที่ไม่ซ้ำกันค่าต่อไปคือ “ใช่”แล้วสามเหลี่ยมทึบก็จะย้ายไปยังโหนดถัดไปดังรูปที่ 2.11



รูปที่ 2.11 แสดงภาพที่ 8 ของการสร้างต้นไม้ตัดสินใจจากชุดข้อมูลการเล่นเทนนิส

จากรูปที่ 2.11 แบ่งกรณีตัวอย่างในโหนดลมแรงต่อไปในค่าที่สองคือค่า “ใช่” พิจารณาโหนดปัจจุบันพบว่าสามารถจำแนกได้เป็นคลาส “ไม่เล่น” โหนดปัจจุบันจะถูกระบุด้วยคลาส “ไม่เล่น” ดังรูปที่ 2.12



รูปที่ 2.12 แสดงภาพที่ 9 ของการสร้างต้นไม้ตัดสินใจจากชุดข้อมูลการเล่นเทนนิส

จากรูปที่ 2.12 หลังจากระบุคลาส “ไม่เล่น” ให้กับโหนดในกิ่ง “ใช่” ซึ่งเป็นค่าที่ไม่ซ้ำกันของแอตทริบิวต์ลมแรง เป็นอันเสร็จสิ้นขั้นตอนการสร้างต้นไม้ตัดสินใจจากชุดข้อมูลการเล่นเทนนิส จะสังเกตุว่ากรณีตัวอย่างทั้งหมดจะถูกจำแนกจนหมดสิ้น และการแบ่งกรณีตัวอย่างของต้นไม้ตัดสินใจจะเป็นในลักษณะเดียวกันกับการค้นหาในแนวลึกก่อน (Depth-first search) ก็คือจะทำการค้นหาในส่วนที่ลึกที่สุดก่อนแล้วย้อนขึ้นไป นี้คือขั้นตอนการสร้างต้นไม้ตัดสินใจแบบอธิบายด้วยภาพ ถัดมาให้พิจารณาการเขียนโปรแกรมเพื่อสร้างต้นไม้ตัดสินใจโดยพิจารณารหัสเทียม (Pseudocode) ของอัลกอริทึม ID3 ซึ่งดัดแปลงจากหนังสือ “Artificial Intelligence : A Modern Approach” [7]

```
1   function ID3 (instances, parent_instances, attributes) returns a tree
        if instances is empty,
            then return the leaf node which labeled by the majority of parent_instances
        else if all instances have the same classification,
            then return the leaf node which labeled by its classification
6       else if attributes is empty,
            then return the leaf node which labeled by the majority of instances
        else
            Compute information gains for all attributes using instances and attributes
            A ← Select the attribute which has the best information gain
11          Insert the new node labeled by attribute A to tree
12          for each value v of A do
                ins ← Select instances of value v
                subtree ← ID3 (ins, instances, attribute–A)
                Add a branch to tree with label v and subtree
            return tree
```

รูปที่ 2.13 แสดงรหัสเทียมของอัลกอริทึม ID3

จากรหัสเทียมของอัลกอริทึม ID3 ในรูปที่ 2.13 จะเห็นว่าเป็นฟังก์ชัน ID3 เป็นลักษณะฟังก์ชันเวียนบังเกิด (Recursive function) ที่จะคืนค่า (Return) เป็นโหนดต่อกันเป็นต้นไม้ตัดสินใจไปเรื่อย ๆ มีตัวแปรที่รับค่าดังนี้ ตัวแปร instances คือตัวแปรที่เก็บชุดกรณีตัวอย่างปัจจุบัน ตัวแปร parent_instances คือตัวแปรที่เก็บชุดกรณีตัวอย่างในโหนดพ่อแม่ของโหนดปัจจุบัน ตัวแปร attributes คือตัวแปรที่เก็บแอตทริบิวต์ทั้งหมด เมื่อสิ้นสุดกระบวนการของฟังก์ชันจะคืนค่าต้นไม้ตัดสินใจซึ่งอยู่ในรูปของตัวแปร tree ออกมา เมื่อเริ่มต้นการทำงานของฟังก์ชัน ID3 ตัวแปร instances และ parent_instances จะถูกตั้งค่าให้เท่ากับกรณีตัวอย่างทั้งหมด และจะต้องนำตัวแปรทั้งสามมาตรวจสอบเงื่อนไขดังนี้

เงื่อนไขที่ 1 : ถ้าตัวแปร instances เป็นเซตว่างหรือกรณีตัวอย่างปัจจุบันไม่มีเหลือแล้วในตัวแปร instances ให้คืนค่าโหนดคำตอบที่ระบุด้วยเสียงส่วนมากของคลาสใน parent_instances มิฉะนั้นให้ไปตรวจสอบต่อในเงื่อนไขที่ 2

เงื่อนไขที่ 2 : ถ้ากรณีตัวอย่างปัจจุบันในตัวแปร instances เป็นคลาสเดียวกันทั้งหมด ให้คืนค่าโหนดคำตอบที่ระบุด้วยคลาสนั้น ๆ มิฉะนั้นให้ไปตรวจสอบต่อในเงื่อนไขที่ 3

เงื่อนไขที่ 3 : ถ้าตัวแปร attributes เป็นเซตว่างหรือแอตทริบิวต์ถูกใช้จนหมด ให้คืนค่าโหนดคำตอบที่ระบุด้วยเสียงส่วนมากของคลาสใน instances มิฉะนั้นให้ไปยังเงื่อนไขที่ 4

เงื่อนไขที่ 4 : ถ้าตัวแปรทั้งสามไม่ตรงกับ 3 เงื่อนไขแรกให้กระทำขั้นตอนดังนี้

1. คำนวณค่าเกนความรู้ของแอตทริบิวต์ทั้งหมดโดยใช้ตัวแปร instances และ attributes
2. เลือกแอตทริบิวต์ที่มีค่าเกนความรู้มากที่สุดมาเก็บไว้ในตัวแปร A
3. สร้างโหนดใหม่โดยระบุตามแอตทริบิวต์ในตัวแปร A
4. วนลูป (Loop) ตามค่าที่ไม่ซ้ำกันทั้งหมดของ A โดยเริ่มจากค่าแรกและกระทำดังนี้
    4.1 นำค่าที่ไม่ซ้ำกันปัจจุบันมาเก็บไว้ในตัวแปร v
    4.2 เลือกกรณีตัวอย่างที่มีแอตทริบิวต์ A เป็นค่า v มาเก็บไว้ในตัวแปร ins
    4.3 เรียกทำฟังก์ชัน ID3 ก็คือการ recursion โดยส่งพารามิเตอร์ตามลำดับเป็นตัวแปร ins, instances และ attributes โดยลบแอตทริบิวต์ A คือแอตทริบิวต์ที่ถูกเลือกแล้วออกจาก attributes ก่อนที่จะส่งและคืนค่าฟังก์ชันกลับมาเก็บไว้ในตัวแปร subtree
    4.4 ต่อโหนด subtree ไปยัง tree และระบุกิ่งของ tree เป็น v
5. คืนค่าตัวแปร tree

2.4 อัลกอริทึม C4.5

อัลกอริทึม C4.5 [8] เป็นหนึ่งในอัลกอริทึมที่ใช้สร้างต้นไม้ตัดสินใจสำหรับจำแนกประเภทข้อมูล และเป็นอัลกอริทึมที่พัฒนาจากอัลกอริทึม ID3 โดยพัฒนาดังนี้

1. สามารถจำแนกได้ทั้งชุดข้อมูลที่เป็นค่าต่อเนื่อง (Continuous value) และไม่ต่อเนื่อง
2. สามารถจัดการแอตทริบิวต์ที่มี Missing value
3. สามารถจัดการปัญหาความเข้มงวดในการสร้างกฎการตัดสินใจในอัลกอริทึม ID3 ด้วยการตัดกิ่งออกหรือการทำ Pruning
4. สามารถจัดการปัญหาความลำเอียงในการเลือกแอตทริบิวต์โดยใช้การคำนวณอัตราส่วนเกน (Gain ratio) เป็นเกณฑ์ในการเลือกแอตทริบิวต์

2.4.1 อัตราส่วนเกน

อัตราส่วนเกณหรือ Gain ratio เป็นเกณฑ์ในการเลือกแอตทริบิวต์ที่พัฒนาจากค่าเกนความรู้ เพื่อจัดการปัญหาความลำเอียงในการเลือกแอตทริบิวต์ของเกนความรู้ที่มักจะเลือกแอตทริบิวต์ที่มีจำนวนค่าที่ไม่ซ้ำกันมากกว่า ซึ่งแอตทริบิวต์ที่มีค่าที่เป็นไปได้น้อยกว่าอาจให้ผลลัพธ์ที่ดีกว่า สมการของอัตราส่วนเกนแสดงในสมการที่ 2.6

$$GR(I, A) = \frac{IG(I,A)}{Split\,Entropy(I,A)} \tag{2.6}$$

GR(I,A) คือค่าอัตราส่วนเกนของ A ซึ่ง A คือแอตทริบิวต์ใด ๆ

IG(I,A) คือค่าเกนความรู้ของ A คำนวณได้จากสมการที่ 2.2

I คือเซตของกรณีตัวอย่างที่เหลืออยู่

SplitEntropy(I,A) คือค่าเอนโทรปีของค่าที่ไม่ซ้ำกันใน A

$$Split\,Entropy(I, A) = -\sum_{v \in V} p(v) \log_2 p(v) \tag{2.7}$$

V คือเซตของค่าที่ไม่ซ้ำกันในแอตทริบิวต์ A

p(v) คืออัตราส่วนของจำนวนกรณีตัวอย่างที่เป็นค่า v ในแอตทริบิวต์ A ต่อจำนวนกรณีตัวอย่างทั้งหมดในเซต I

ตัวอย่างการคำนวณอัตราส่วนเกนในแอตทริบิวต์ทัศนียภาพจากชุดข้อมูลการเล่นเทนนิสในตารางที่ 2.3

IG(I,ทัศนียภาพ) = 0.25

SplitEntropy(I,ทัศนียภาพ) = [-p(แดดจัด) x $\log_2$p(แดดจัด)] + [-p(มีเมฆมาก) x $\log_2$p(มีเมฆมาก)] + [-p(ฝนตก) x $\log_2$p(ฝนตก)]

SplitEntropy(I,ทัศนียภาพ) = [-5/14 x $\log_2$5/14] + [-4/14 x $\log_2$4/14] + [-5/14 x $\log_2$5/14]

SplitEntropy(I,ทัศนียภาพ) = 1.577

ดังนั้น

GR(I,ทัศนียภาพ) = 0.25/1.577 = 0.156

2.4.2 ขั้นตอนของอัลกอริทึม C4.5

อัลกอริทึม C4.5 จะคำนวณอัตราส่วนเกนของแต่ละแอตทริบิวต์ แอตทริบิวต์ใดมีค่าอัตราส่วนเกนมากที่สุดจะถูกเลือกนำมาสร้างโหนดของต้นไม้ตัดสินใจ และจะทำการแบ่งกรณีตัวอย่างไปเรื่อย ๆ เช่นเดียวกับอัลกอริทึม ID3 จนกว่ากรณีตัวอย่างทั้งหมดจะถูกจำแนกขั้นตอนของอัลกอริทึม C4.5 มีดังนี้

ขั้นตอนที่ 1 : คำนวณค่าอัตราส่วนเกนของทุก ๆ แอตทริบิวต์โดยใช้สมการที่ 2.6

ขั้นตอนที่ 2 : เลือกแอตทริบิวต์ที่มีค่าอัตราส่วนเกนมากที่สุด เพื่อสร้างโหนดปัจจุบันของต้นไม้ตัดสินใจ

ขั้นตอนที่ 3 : แบ่งกรณีตัวอย่างไปตามค่าที่ไม่ซ้ำกันของแอตทริบิวต์ที่ถูกเลือกจากขั้นตอนที่ 2

ขั้นตอนที่ 4 : ทำซ้ำขั้นตอนที่ 1-3 จนกระทั่งสามารถจำแนกประเภทกรณีตัวอย่างได้ทั้งหมด

ขั้นตอนที่ 5 : ทำการ Pruning ต้นไม้ตัดสินใจที่ได้ โดยอัลกอริทึม C4.5 จะใช้การ Pruning ด้วยวิธี Pessimistic pruning

การทำ Pessimistic pruning คือการลดจำนวนกฎการตัดสินใจที่มากเกินจำเป็น จะคำนวณเพื่อทำนายค่าความผิดพลาด (Predicted error) ก่อนทำ Pruning ของโหนดที่ไม่ใช่โหนดคำตอบ เปรียบเทียบกับค่าความผิดพลาดหลังทำ Pruning ถ้าค่าความผิดพลาดหลังทำ Pruning น้อยกว่าค่าความผิดพลาดก่อนทำ Pruning ให้ทำการ Pruning โหนดโดยแปลงจากโหนดที่ถูกระบุด้วยแอตทริบิวต์ให้เป็นโหนดคำตอบที่ระบุด้วยเสียงส่วนมากของคลาส การทำ Pruning จะเริ่มทำตั้งแต่ด้านล่างของต้นไม้ตัดสินใจไล่ขึ้นไปเรื่อย ๆ ในแต่ละเส้นทาง สมการความผิดพลาดก่อนทำ Pruning แสดงในสมการที่ 2.8 และสมการความผิดพลาดหลังทำ Pruning แสดงในสมการที่ 2.9

$$PredictedError_{BP}(A) = \sum_{v \in V} N \times U_{CF}(E, N) \quad \textbf{(2.8)}$$

$PredictedError_{BP}(A)$ คือค่าความผิดพลาดก่อนทำ Pruning ของ A ซึ่ง A เป็นโหนดที่ถูกระบุด้วยแอตทริบิวต์ใด ๆ

V คือเซตของค่าที่ไม่ซ้ำกันใน A

N คือจำนวนกรณีตัวอย่างปัจจุบันทั้งหมด

E คือจำนวนกรณีตัวอย่างในคลาสที่ไม่ได้เป็นเสียงส่วนมาก

CF คือระดับความเชื่อมั่น (Confidence level) อัลกอริทึม C4.5 ตั้งค่า CF= 25%

U_{CF}(E,N) คือ Upper limit ของความน่าจะเป็น หาได้จากการคำนวณ Confidence limit ในการกระจายแบบ Binomial (Binomial distribution)

$$PredictedError_{AP}(A) = N \times U_{CF}(E, N) \quad \textbf{(2.9)}$$

$PredictedError_{AP}$(A) คือค่าความผิดพลาดหลังทำ Pruning ของ A ซึ่ง A เป็นโหนดที่ถูกระบุด้วยแอตทริบิวต์ใด ๆ

2.5 ทฤษฎีเบย์ (Bayes' Theorem)

ทฤษฎีเบย์คือความสัมพันธ์ระหว่างความน่าจะเป็นแบบมีเงื่อนไข (Conditional probability) ซึ่งความน่าจะเป็นแบบมีเงื่อนไขคือความน่าจะเป็นที่จะเกิดเหตุการณ์หนึ่ง เมื่อมีอีกเหตุการณ์หนึ่งเกิดขึ้นก่อนแล้วโดยมีสมการดังนี้

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \textbf{(2.10)}$$

P(A|B) คือความน่าจะเป็นแบบมีเงื่อนไข คือความน่าจะเป็นที่จะเกิดเหตุการณ์ A เมื่อเหตุการณ์ B เกิดขึ้นก่อน

P(A∩B) คือความน่าจะเป็นที่จะเกิดทั้งเหตุการณ์ A และ B

P(B) คือความน่าจะเป็นที่จะเกิดเหตุการณ์ B

ตัวอย่างการหาความน่าจะเป็นแบบมีเงื่อนไข สมมติว่าความน่าจะเป็นของคนที่จะชอบสุนัขเท่ากับ 0.8 ความน่าจะเป็นของคนที่จะชอบแมวเท่ากับ 0.7 และความน่าจะเป็นที่คนหนึ่งคนจะชอบทั้งสุนัขและแมวเท่ากับ 0.4 นายกอเป็นเพื่อนบ้านของนายขอ บ้านนายกอชอบเลี้ยงสุนัขในขณะที่นายขอชอบเลี้ยงแมว กำหนดให้ A เป็นเหตุการณ์ที่คนจะชอบแมว B เป็นเหตุการณ์ที่คนจะชอบสุนัข

1. นายกอเป็นที่คนชอบเลี้ยงสุนัข โอกาสที่นายกอจะชอบเลี้ยงแมวด้วยคิดเป็นเท่าใด
P(A|B) คือความน่าจะเป็นที่นายกอจะชอบแมว เมื่อเรารู้ว่านายกอเป็นคนที่ชอบเลี้ยงสุนัข
ดังนั้น P(A|B) = 0.4 / 0.8 = 0.5

2. ต่อมาถ้าเราอยากจะหาความน่าจะเป็นที่นายขอจะชอบสุนัข เมื่อเดิมนายขอชอบเลี้ยงแมว
P(B|A) คือความน่าจะเป็นที่นายขอจะชอบสุนัข เมื่อเรารู้ว่านายขอเป็นคนที่ชอบเลี้ยงแมว จะแทนค่าสมการที่ 2.10 ใหม่เป็นดังนี้

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad \textbf{(2.11)}$$

ดังนั้น P(B|A) = 0.4 / 0.7 = 0.57

เมื่อลองพิจารณาสมการที่ 2.10 และ 2.11 จะพบว่ามีจุดที่เหมือนกันคือ P(A∩B) ซึ่งทำให้สามารถนำทั้งสองสมการมาเท่ากันได้ดังนี้

จากสมการที่ 2.10 ย้ายข้าง P(B) ขึ้นไปคูณจะได้เป็นสมการที่ 2.12

$$P(A \cap B) = P(A|B) \times P(B) \quad \textbf{(2.12)}$$

จากสมการที่ 2.11 ย้ายข้าง P(A) ขึ้นไปคูณจะได้เป็นสมการที่ 2.13

$$P(A \cap B) = P(B|A) \times P(A) \quad \textbf{(2.13)}$$

นำสมการที่ 2.12 มาเท่ากันกับสมการที่ 2.13 แล้วย้ายข้างจะได้เป็นสมการของทฤษฎีเบย์คือสมการที่ 2.14

$$P(B|A) = \frac{P(A|B) \times P(B)}{P(A)} \quad \textbf{(2.14)}$$

สมการของทฤษฎีเบย์นั้นทำให้ทราบถึงความสัมพันธ์ระหว่าง P(A|B) และ P(B|A) จากตัวอย่างการหาความน่าจะเป็นแบบมีเงื่อนไขข้างต้น P(A|B) คือความน่าจะเป็นที่นายกอจะชอบแมว เมื่อเรารู้ว่านายกอเป็นคนที่ชอบเลี้ยงสุนัข และคำนวณ P(A|B) จากสมการความน่าเป็นแบบมีเงื่อนไขปกติได้ P(A|B) = 0.5 ต่อมาถ้าเราอยากรู้ค่าของ P(B|A) ก็คือความน่าจะเป็นที่นายขอจะชอบสุนัข เมื่อเรารู้ว่านายขอเป็นคนที่ชอบเลี้ยงแมว เราสามารถใช้ทฤษฎีของเบย์เพื่อคำนวณหาค่า P(B|A) ได้ ดังนั้น P(B|A) = (0.5 x 0.8) / 0.7 = 0.57 ซึ่งเท่ากับ P(B|A) ที่หาด้วยสมการที่ 2.11

2.6 อัลกอริทึมนาอีฟเบย์ (Naive Bayes)

อัลกอริทึมนาอีฟเบย์เป็นวิธีง่าย ๆ ที่ใช้สำหรับการจำแนกประเภทข้อมูล อัลกอริทึมนาอีฟเบย์จะอนุมานแอตทริบิวต์แต่ละตัวเป็นอิสระต่อกัน และใช้ทฤษฎีเบย์มาดัดแปลงเป็นพื้นฐานเพื่อคำนวณความน่าจะเป็นสำหรับจำแนกข้อมูล จากสมการที่ 2.14 หรือสมการของเบย์ ให้เปลี่ยนสัญลักษ์ A และ B เป็น A และ C_k ตามลำดับ โดยที่ A เป็นเซตของแอตทริบิวต์ทั้งหมด C_k คือคลาส และ k คือจำนวนคลาสทั้งหมด จะได้สมการดังนี้

$$P(C_k|A) = \frac{P(A|C_k) \times P(C_k)}{P(A)} \quad \textbf{(2.15)}$$

จากสมการที่ 2.15 เนื่องจาก A คือเซตของแอตทริบิวต์ทั้งหมดและแอตทริบิวต์ทุกตัวเป็นอิสระต่อกันและ A = (a_1, ... , a_n) โดยที่ n คือจำนวนแอตทริบิวต์ทั้งหมด ในทางปฏิบัติเราจะไม่พิจารณาตัวหาร P(A) เพราะตัวหารไม่ได้ขึ้นอยู่กับคลาส ดังนั้นจะได้สมการคำนวณความน่าจะเป็นของแต่ละคลาสเพื่อใช้จำแนกดังนี้

$$P(C_k|A) = \left(\prod_{i=1}^{n} P(a_i|C_k)\right) \times P(C_k) \qquad \textbf{(2.16)}$$

เมื่อนาอีฟเบย์จะทำการทำนายข้อมูล จะคำนวณความน่าจะเป็นที่ตอบเป็นคลาสต่าง ๆ โดยใช้สมการที่ 2.16 ถ้าคลาสไหนมีความน่าจะเป็นมากที่สุด นาอีฟเบย์จะทำนายผลหรือจำแนกประเภทเป็นคลาสนั้น ๆ ตัวอย่างการจำแนกประเภทด้วยนาอีฟเบย์โดยใช้ชุดข้อมูลการเล่นเทนนิสใน 14 วัน ตามตารางที่ 2.3 เป็นชุดข้อมูลสำหรับฝึกฝน ถ้าอยากจะทราบว่าจะเล่นเทนนิสหรือไม่ในกรณีตัวอย่าง วันที่ทัศนียภาพ = ฝนตก, อุณหภูมิ = ไม่เย็นมาก, ความชื้น = ปกติ และ ลมแรง = ใช่ มีวิธีการคำนวณดังนี้

1. คำนวณความน่าจะเป็นที่จะเล่นเทนนิส

P(เล่น | A) = P(ทัศนียภาพ=ฝนตก | เล่น) x P(อุณหภูมิ=ไม่เย็นมาก | เล่น) x
P(ความชื้น = ปกติ | เล่น) x P(ลมแรง = ใช่ | เล่น) x P(เล่น)

ดังนั้น P(เล่น | A) = 3/9 x 4/9 x 6/9 x 3/9 x 9/14 = 0.0212

2. คำนวณความน่าจะเป็นที่ไม่เล่นเทนนิส

P(ไม่เล่น | A) = P(ทัศนียภาพ=ฝนตก | ไม่เล่น) x P(อุณหภูมิ=ไม่เย็นมาก | ไม่เล่น) x
P(ความชื้น = ปกติ | ไม่เล่น) x P(ลมแรง = ใช่ | ไม่เล่น) x P(ไม่เล่น)

ดังนั้น P(ไม่เล่น | A) = 2/5 x 2/5 x 1/5 x 3/5 x 5/14 = 0.0069

3. เปรียบเทียบความน่าจะเป็นของคลาสทั้งหมด

P(เล่น | A) = 0.0212

P(ไม่เล่น | A) = 0.0069

ซึ่ง P(เล่น | A) > P(ไม่เล่น | A) ดังนั้นกรณีตัวอย่างนี้จึงจำแนกเป็นคลาส “เล่น”

บทที่ 3
งานวิจัยที่เกี่ยวข้อง

3.1 ปัญหาของอัลกอริทึม ID3 ที่งานวิจัยที่เกี่ยวข้องพยายามแก้ไข

3.1.1 ปัญหาการลำเอียงในการเลือกแอตทริบิวต์

ปัญหานี้เกิดจากการคัดเลือกแอตทริบิวต์ด้วยการคำนวณเกนความรู้ของอัลกอริทึม ID3 แอตทริบิวต์ที่มีค่าที่ไม่ซ้ำกันหลายค่ามักจะมีค่าเกนความรู้ที่มากกว่าแอตทริบิวต์ที่มีค่าที่ไม่ซ้ำกันน้อยค่า ซึ่งอัลกอริทึม ID3 จะเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดมาสร้างโหนดของต้นไม้ตัดสินใจก่อนเสมอ ดังนั้นแอตทริบิวต์ที่มีจำนวนค่าที่ไม่ซ้ำกันมากก็มักจะถูกเลือกก่อนแอตทริบิวต์ที่มีจำนวนค่าที่ไม่ซ้ำกันน้อย ๆ นี้คือการลำเอียงในการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 ที่เกิดขึ้น ซึ่งแอตทริบิวต์ที่มีจำนวนค่าที่ไม่ซ้ำกันน้อย ๆ อาจส่งผลให้ต้นไม้ตัดสินใจผลลัพธ์มีความแม่นยำในการจำแนกที่ดีกว่าต้นไม้ตัดสินใจผลลัพธ์ที่สร้างจากแอตทริบิวต์ที่มีจำนวนค่าที่ไม่ซ้ำกันมาก

3.1.2 ปัญหาแอตทริบิวต์ที่มีความสำคัญเท่ากัน

ปัญหานี้เกิดในระหว่างการสร้างต้นไม้ตัดสินใจ เมื่อเหตุการณ์ที่มีแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดมากกว่า 1 แอตทริบิวต์เกิดขึ้นในขณะการคัดเลือกแอตทริบิวต์ ซึ่งแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดและเท่ากันเหล่านี้ก็คือแอตทริบิวต์ที่มีความสำคัญเท่ากัน อัลกอริทึม ID3 จะไม่สามารถตัดสินใจได้ว่าแอตทริบิวต์ที่มีความสำคัญเท่ากันตัวใดจะดีที่สุด นำไปสู่การสุ่มเลือกแอตทริบิวต์ที่มีความสำคัญเท่ากันเหล่านี้เพื่อนำมาสร้างโหนดในต้นไม้ตัดสินใจ พฤติกรรมนี้นำไปสู่ความไม่มั่นคงของความลึกสูงสุดในต้นไม้ตัดสินใจผลลัพธ์ เพราะแต่ละทางเลือกของแอตทริบิวต์ที่มีความสำคัญเท่ากันที่ถูกสุ่มนำไปสู่ต้นไม้ตัดสินใจผลลัพธ์ที่อาจมีระดับความลึกสูงสุดแตกต่างกัน

3.2 งานวิจัยที่เกี่ยวข้อง

3.2.1 การปรับปรุงเกนความรู้ของอัลกอริทึม ID3 โดยการถ่วงน้ำหนัก

การปรับปรุงเกนความรู้ของอัลกอริทึม ID3 โดยการถ่วงน้ำหนัก (Weighted Modified Information Gain) หรืออัลกอริทึม ωID3 [3] พัฒนาขึ้นเพื่อขจัดปัญหาการลำเอียงในการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 หลักการของอัลกอริทึม ωID3 คือ เมื่อแอตทริบิวต์ใด ๆ มีจำนวนค่าที่ไม่ซ้ำกันเท่ากับจำนวนค่าที่ไม่ซ้ำกันสูงสุดของแอตทริบิวต์ทั้งหมดและมีค่าเกนความรู้มากที่สุด จะต้องทำการปรับปรุงค่าเกนความรู้ของแอตทริบิวต์นั้น ๆ ด้วยสมการที่ 3.1 – 3.4 และจะใช้ค่าเกนความรู้ที่ปรับปรุงใหม่เป็นเกณฑ์ในการเลือกแอตทริบิวต์ ขั้นตอนอื่น ๆ ของอัลกอริทึม ωID3 จะ

เหมือนกันกับอัลกอริทึม ID3 แบบดั้งเดิมทุกอย่าง แต่จะแตกต่างกันตรงที่การคำนวณเกณฑ์ในการเลือกแอตทริบิวต์

$$Gain'(A) = \omega_A \times Gain(A) \quad (3.1)$$

Gain'(A) คือค่าเกนความรู้ที่ถูกปรับปรุงแล้วของ A ซึ่ง A เป็น attribute ตัวหนึ่ง

Gain(A) คือค่าเกนความรู้เดิมของ A ซึ่งคำนวณได้จากสมการที่ 2.2

ω_A คือค่าถ่วงน้ำหนักของ A

$$\omega_A = \frac{1}{\alpha}\omega'_A \quad (3.2)$$

โดยที่ α=1 เมื่อค่าเกนความรู้และจำนวนค่าที่ไม่ซ้ำกันของ A คือค่าสูงสุด มิฉะนั้น α=ω'_A

$$\omega'_A = \frac{AF_D(A)}{\sum_{i\in n} AF_D(i)} \quad (3.3)$$

$AF_D(A)$ คือค่าความสัมพันธ์ระหว่าง A และคลาสของข้อมูล

n คือเซตของแอตทริบิวต์ทั้งหมด

$$AF_D(A) = \frac{\sum_{v\in V}|A_{vd} - (\sum_{c\in C \wedge c\neq d} A_{vc})|}{nv} \quad (3.4)$$

V คือเซตของค่าที่ไม่ซ้ำกันทั้งหมดของ A

d คือคลาสแรกของชุดข้อมูล

C คือเซตของคลาสทั้งหมด

A_{vd} คือจำนวนกรณีตัวอย่างที่มีค่า v ใน A และเป็นคลาส d

A_{vc} คือจำนวนกรณีตัวอย่างที่มีค่า v ใน A และเป็นคลาส c (c คือคลาสอื่น ๆ ที่ไม่ใช่คลาสแรกตามลำดับในชุดข้อมูล)

nv คือจำนวนค่าที่ไม่ซ้ำกันทั้งหมดของ A

อัลกอริทึม ωID3 สามารถขจัดปัญหาการลำเอียงในการเลือกแอตทริบิวต์ที่มีจำนวนค่าที่ไม่ซ้ำกันมาก ๆ ได้ ซึ่งการปรับปรุงเกนความรู้ของแอตทริบิวต์ที่มีค่าที่ไม่ซ้ำกันสูงสุดและมีค่าเกนความรู้สูงสุดส่งผลให้แอตทริบิวต์ที่มีค่าที่ไม่ซ้ำกันน้อยกว่ารองลงมามีโอกาสถูกเลือกบ่อยครั้ง ซึ่งในงานวิจัยนี้ระบุว่าอัลกอริทึม ωID3 สามารถเพิ่มความแม่นยำในการจำแนกจากอัลกอริทึม ID3

แบบดั้งเดิมได้อย่างชัดเจน แต่อย่างไรก็ตามผลการทดลองแสดงให้เห็นว่า อัลกอริทึม ωID3 สร้างต้นไม้ตัดสินใจที่มีความซับซ้อนทั้งจำนวนโหนดทั้งหมดและจำนวนของกิ่งที่เพิ่มขึ้นมากกว่าอัลกอริทึม ID3 แบบดั้งเดิมส่งผลให้ใช้พื้นที่ในการเก็บต้นไม้ตัดสินใจผลลัพธ์ในหน่วยความจำมากขึ้น

3.2.2 การปรับปรุงอัลกอริทึม ID3 โดยการประยุกต์ทฤษฎีเซตอย่างหยาบเพื่อคำนวณค่าความมั่นคงของแอตทริบิวต์

งานวิจัยนี้พัฒนาขึ้นเพื่อขจัดปัญหาการลำเอียงในการเลือกแอตทริบิวต์ของเกนความรู้ที่มักจะเลือกแอตทริบิวต์ที่มีจำนวนค่าที่ไม่ซ้ำกันมาก ดังที่ได้กล่าวไว้ว่าแอตทริบิวต์ที่มีจำนวนค่าที่ไม่ซ้ำกันน้อย ๆ อาจจะส่งผลให้ต้นไม้ตัดสินใจผลลัพธ์มีความแม่นจำในการจำแนกที่ดีกว่าแอตทริบิวต์ที่มีจำนวนค่าที่ไม่ซ้ำกันมากที่สุด งานวิจัยนี้ [2] ได้ประยุกต์ทฤษฎีเซตอย่างหยาบเพื่อคำนวณค่าความมั่นคง (Consistency) และใช้การคำนวณ Consistency เป็นเกณฑ์ในการเลือกแอตทริบิวต์แทนที่การคำนวณเกนความรู้ของอัลกอริทึม ID3 แบบดั้งเดิม ส่วนขั้นตอนอื่น ๆ จะเหมือนกับอัลกอริทึม ID3 แบบดั้งเดิมปกติทุกอย่าง แต่จะแตกต่างกันแค่การคำนวณเกณฑ์ในการเลือกแอตทริบิวต์ การคำนวณ Consistency ใช้สมการดังต่อไปนี้

$$C(A) = \frac{\sum_{v \in V} x_{v1}^2 - x_{v2}^2}{ins^2} \quad \textbf{(3.5)}$$

C(A) คือค่า Consistency ของ A ซึ่ง A เป็นแอตทริบิวต์ตัวหนึ่ง

V คือเซตของค่าที่ไม่ซ้ำกันทั้งหมดของ A

X_{v1} คือจำนวนกรณีตัวอย่างที่มีค่าที่ไม่ซ้ำกันของ A เป็น v และเป็นคลาสที่มีจำนวนมากที่สุด

X_{v2} คือจำนวนกรณีตัวอย่างที่มีค่าที่ไม่ซ้ำกันของ A เป็น v และเป็นคลาสที่มีจำนวนรองลงมา

ins คือจำนวนกรณีตัวอย่างปัจจุบันทั้งหมด

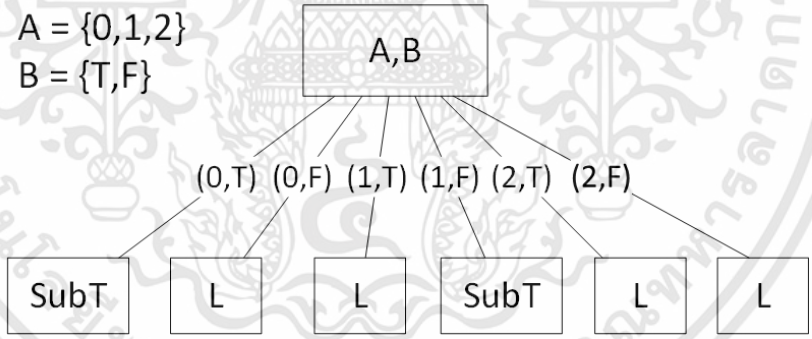
วิธีการนี้จะเลือกแอตทริบิวต์ที่มีค่า Consistency สูงสุดมาสร้างโหนดของต้นไม้ตัดสินใจ แต่ถ้ามีแอตทริบิวต์มากกว่าหนึ่งแอตทริบิวต์ที่มีค่า Consistency สูงสุดเท่ากัน จะดำเนินการหาค่า Consistency ของแอตทริบิวต์ที่เท่ากันเหล่านี้ต่อโดยใช้สมการที่ 3.6 หลังจากหาค่า Consistency อีกรอบแล้ว จะทำการเลือกแอตทริบิวต์ที่มีค่า Consistency มากที่สุดจากสมการ 3.6 มาสร้างโหนดการตัดสินใจ

$$C(A) = \frac{\sum_{v \in V} x_{v1}^2}{ins^2} \quad \textbf{(3.6)}$$

วิธีการนี้สามารถขจัดปัญหาการลำเอียงในแอตทริบิวต์ที่มีจำนวนค่าที่ไม่ซ้ำกันมาก ๆ ได้ เนื่องจากเกณฑ์ในการเลือกแอตทริบิวต์ถูกแทนที่ด้วยการประยุกต์ทฤษฎีเซตอย่างหยาบเพื่อคำนวณค่าความมั่นคง ซึ่งมีผลการทดลองที่ชัดเจนว่าความแม่นยำในการจำแนกประเภทเพิ่มมากขึ้นอย่างชัดเจนจากอัลกอริทึม ID3 แบบดั้งเดิม อย่างไรก็ตามข้อจำกัดของงานวิจัยนี้คือจำแนกประเภทได้เฉพาะกับชุดข้อมูลที่มีเพียง 2 คลาสเท่านั้น ซึ่งในความเป็นจริงมีชุดข้อมูลอยู่มากมายที่มีจำนวนคลาสมากกว่า 2 ขึ้นไป

3.2.3 การปรับปรุงอัลกอริทึม ID3 โดยการรวมค่าระหว่างแอตทริบิวต์ที่มีความสำคัญเท่ากัน

ในวิทยานิพนธ์เล่มนี้จะใช้ชื่อเรียกงานวิจัยนี้ว่าอัลกอริทึม EVC-ID3 เพื่อความสะดวกในการเขียนผลการทดลอง ซึ่งผู้เขียนดัดแปลงชื่อย่อของงานวิจัยนี้ [5] จากชื่องานวิจัยต้นฉบับภาษาอังกฤษจาก “Improving ID3 Algorithm by Combining Values from Equally Important Attributes” เป็น “Equally important attribute Values Combining ID3” และจากนั้นจะได้ชื่อย่อเป็น EVC-ID3 งานวิจัยนี้พัฒนาขึ้นเพื่อแก้ปัญหาแอตทริบิวต์ที่มีความสำคัญเท่ากัน วิธีการแก้ปัญหาของงานวิจัยนี้คือ สุ่มแอตทริบิวต์ที่มีความสำคัญสูงสุดและเท่ากันมา 2 แอตทริบิวต์ เพื่อนำมารวมค่าที่ไม่ซ้ำกันและสร้างเป็นโหนดเดียวกันดังรูปที่ 3.1



รูปที่ 3.1 แสดงการรวมค่าที่ไม่ซ้ำกันระหว่างแอตทริบิวต์ A และ B

จากรูปสมมติว่า A และ B คือแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดและมีค่าเท่ากัน แอตทริบิวต์ A มีค่าที่ไม่ซ้ำกันเป็น 0,1 และ 2 แอตทริบิวต์ B มีค่าที่ไม่ซ้ำกันเป็น T และ F โหนด L คือโหนดคำตอบ ส่วน SubT คือต้นไม้ย่อยภายในต้นไม้ตัดสินใจผลลัพธ์ จะเห็นได้ว่ายิ่งมีจำนวนแอตทริบิวต์ที่จะนำไปรวมกันมากเท่าไหร่ จำนวนกิ่งของโหนดก็จะยิ่งเพิ่มมากขึ้น เพราะฉะนั้นงานวิจัยนี้จึงจำกัดการรวมโหนดที่มีความสำคัญเท่ากันไว้เพียง 2 แอตทริบิวต์ หลังรวมแอตทริบิวต์แล้วก็จะทำการแบ่งกรณีตัวอย่างไปเรื่อย ๆ เช่นเดียวกับอัลกอริทึม ID3 แบบดั้งเดิม ขั้นตอนการสร้างต้นไม้ตัดสินใจของงานวิจัยนี้เหมือนกันกับอัลกอริทึม ID3 แบบดั้งเดิมแทบจะทุกอย่าง แต่แตกต่างกันที่เมื่อเกิดเหตุการณ์ที่มีแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดมากกว่า 1 แอตทริบิวต์ งานวิจัยนี้จะสุ่ม

เลือกมา 2 แอตทริบิวต์และนำค่าที่ไม่ซ้ำกันมารวมกันและสร้างเป็นโหนดเดียวกัน ในขณะที่อัลกอริทึม ID3 แบบดั้งเดิมจะสุ่มเลือกแอตทริบิวต์เหล่านั้นมา 1 แอตทริบิวต์และนำมาสร้างโหนดของต้นไม้ตัดสินใจ ผลการทดลองพบว่างานวิจัยนี้สามารถลดจำนวนความลึกสูงสุดของต้นไม้ตัดสินใจผลลัพธ์และยังสามารถรักษาความแม่นยำในการจำแนกไว้ในระดับเดียวกันกับอัลกอริทึม ID3 อย่างไรก็ตาม งานวิจัยนี้มีข้อจำกัดคือต้องใช้ชุดข้อมูลที่มีจำนวนกรณีตัวอย่างมาก ๆ มิฉะนั้นความแม่นยำอาจจะลดลงได้ และผลการทดลองแสดงให้เห็นชัดเจนว่าต้นไม้ตัดสินใจผลลัพธ์ที่ได้มีความซับซ้อนหรือมีจำนวนกฎการตัดสินใจเพิ่มขึ้นมากกว่าต้นไม้ตัดสินใจที่สร้างจากอัลกอริทึม ID3 แบบดั้งเดิมอย่างมาก ซึ่งส่งผลต่อปริมาณพื้นที่ในหน่วยความจำที่ใช้ในการเก็บต้นไม้ตัดสินใจผลลัพธ์ ต้องใช้พื้นที่ในการเก็บต้นไม้ตัดสินใจมากขึ้นกว่าอัลกอริทึม ID3 แบบดั้งเดิมอย่างมาก

3.2.4 การปรับปรุงอัลกอริทึม ID3 โดยการประยุกต์ใช้การค้นหาแบบเอสตาร์

การปรับปรุงอัลกอริทึม ID3 โดยการประยุกต์ใช้ A* search หรืออัลกอริทึม ID3-A* [4] มีวัตถุประสงค์เพื่อแก้ไขปัญหาแอตทริบิวต์ที่มีความสำคัญเท่ากัน ดังที่ได้กล่าวไว้เนื่องจากอัลกอริทึม ID3 แบบดั้งเดิมไม่สามารถตัดสินใจได้ว่าแอตทริบิวต์ที่มีความสำคัญเท่ากันแอตทริบิวต์ใดจะดีที่สุด นำไปสู่พฤติกรรมการสุ่มแอตทริบิวต์ที่มีความสำคัญสูงสุดเท่ากัน ซึ่งส่งผลให้ต้นไม้ตัดสินใจผลลัพธ์มีความลึกสูงสุดที่ไม่แน่นอน A* search ถูกนำมาใช้เพื่อที่จะค้นหาแอตทริบิวต์ที่ส่งผลให้ความลึกสูงสุดของต้นไม้ตัดสินใจมีจำนวนน้อยที่สุด หลักการของอัลกอริทึม ID3-A* จะเหมือนกับอัลกอริทึม ID3 ปกติ แต่สิ่งที่เพิ่มมาคือเมื่อเกิดเหตุการณ์ที่มีแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดมากกว่า 1 แอตทริบิวต์ อัลกอริทึม ID3-A* จะทำการค้นหาแอตทริบิวต์ที่จะทำให้ต้นไม้ตัดสินใจสุดท้ายมีความลึกที่น้อยที่สุดโดยการประยุกต์ใช้ A* search เมื่อได้แอตทริบิวต์ที่ดีที่สุดแล้ว ก็จะเลือกแอตทริบิวต์นั้น ๆ นำไปสร้างโหนดของต้นไม้ตัดสินใจ สำหรับขั้นตอนของอัลกอริทึม ID3-A* ให้พิจารณาขั้นตอนที่เพิ่มขึ้นมาจากอัลกอริทึม ID3 เพื่อความเข้าใจง่ายจะอธิบายโดยสมมุติว่าเกิดเหตุการณ์ที่แอตทริบิวต์ B และ C มีค่าเกนความรู้สูงสุดเท่ากัน เมื่อเกิดเหตุการณ์ที่มีแอตทริบิวต์ที่มีความสำคัญเท่ากันขึ้นอัลกอริทึม ID3-A* จะมีขั้นตอนดังนี้

ขั้นตอนที่ 1 คำนวณค่า f(n) ระหว่างโหนดที่ระบุด้วยแอตทริบิวต์ B และ C โดยใช้สมการที่ 3.7 และ 3.8

$$f(n) = norm\big(depth(n)\big) + \big(1 - IG(n)\big) \qquad \textbf{(3.7)}$$

norm(depth(n)) คือความลึกปัจจุบันของโหนดนั้น ๆ ที่ถูกแปลงให้อยู่ระหว่าง 0.1 ถึง 0.9

IG(n) คือค่าเกนความรู้ของโหนดนั้น ๆ คำนวณได้จากสมการที่ 2.2

$$norm\big(depth(n)\big) = 0.1 + \frac{(depth(n) - minDepth)(0.9 - 0.1)}{maxDepth - minDepth} \quad \textbf{(3.8)}$$

depth(n) คือความลึกปัจจุบันของโหนดนั้น ๆ

minDepth คือความลึกปัจจุบันที่น้อยที่สุดของต้นไม้ตัดสินใจ

maxDepth คือความลึกปัจจุบันที่มากที่สุดของต้นไม้ตัดสินใจ

ขั้นตอนที่ 2 แบ่งกรณีตัวอย่างของโหนดที่มีค่า f(n) ที่น้อยสุดและทำการเลือกโหนดในลักษณะนี้ต่อไปเรื่อย ๆ จนกว่ากรณีตัวอย่างทั้งหมดจะถูกจำแนก

ขั้นตอนที่ 3 เลือกโหนดที่มีค่า f(n) สุดท้ายที่น้อยที่สุด (เลือกระหว่าง B และ C) มาต่อเป็นโหนดจริงในต้นไม้ตัดสินใจ

ผลการทดลองแสดงให้เห็นว่างานวิจัยนี้สามารถลดความลึกของต้นไม้ตัดสินใจได้อย่างชัดเจนในชุดข้อมูลเดียวเท่านั้น ในขณะที่ความแม่นยำในการจำแนกถูกรักษาไว้ในระดับเดียวกับอัลกอริทึม ID3 แบบดั้งเดิม ส่วนข้อจำกัดของงานวิจัยนี้คือชุดข้อมูลต้องมีขนาดใหญ่และซับซ้อน จึงจะเห็นประสิทธิภาพของ ID3-A* ได้อย่างชัดเจน เพราะว่าชุดข้อมูลขนาดใหญ่มีโอกาสที่จะเกิดปัญหาแอตทริบิวต์ที่มีความสำคัญเท่ากันได้บ่อยกว่าชุดข้อมูลขนาดเล็กอย่างมากและใช้เวลาในการฝึกฝนนานกว่าอัลกอริทึม ID3 แบบดั้งเดิมอย่างชัดเจน ยิ่งชุดข้อมูลมีขนาดใหญ่มาก ๆ จะยิ่งใช้เวลาในการสร้างต้นไม้ตัดสินใจมาก

บทที่ 4
งานวิจัยที่เสนอ

จากปัญหาความเข้มงวดในการสร้างกฎการตัดสินใจของอัลกอริทึม ID3 ผู้เขียนได้อธิบายแนวคิดพื้นฐานการแก้ปัญหานี้ไว้ในบทที่ 1 โดยการปรับปรุงอัลกอริทึม ID3 ด้วยการละเลยกรณีตัวอย่างส่วนน้อยเพื่อขจัดปัญหาความเข้มงวดในการสร้างกฎการตัดสินใจ ในบทนี้เราจะมาอธิบายถึงอัลกอริทึมใหม่ที่เสนอคืออัลกอริทึม MII-ID3 (Minor Instances Ignoring ID3) เป้าหมายของอัลกอริทึม MII-ID3 คือเพื่อลดจำนวนของกฎการตัดสินใจและความลึกสูงสุดของต้นไม้ตัดสินใจแบบดั้งเดิม ในขณะที่ความแม่นยำลดลงเพียงเล็กน้อย จากแนวคิดในการพัฒนาที่อธิบายไว้ในบทที่ 1 หัวเรื่องที่ 1.4 ถ้าเราจะทำตามนั้นเลยมันก็เหมือนกับว่าเราจำเป็นต้องกำหนดพารามิเตอร์ในการละเลยกรณีตัวอย่างเช่น โหนดปัจจุบันจะทำการละเลยกรณีตัวอย่างเมื่อจำนวนกรณีตัวอย่างในคลาสแรกห่างจากจำนวนกรณีตัวอย่างในคลาสอื่นรวมกันเกิน 80 เปอร์เซ็นต์ หรืออาจจะกำหนดเป็น 90 เปอร์เซ็นต์ นี้ก็คือพารามิเตอร์ ในการพัฒนาจริงเราไม่ต้องการให้มีพารามิเตอร์เพราะมันจะดูเป็นการจำกัดจนเกินไปและจะทำให้การเปรียบเทียบซับซ้อนเพราะต้องลองทดสอบด้วยหลาย ๆ ค่าของพารามิเตอร์ ดังนั้นในงานวิจัยนี้เราจะทำการละเลยกรณีตัวอย่างส่วนน้อยนี้ด้วยการคำนวณ Threshold ของความน่าจะเป็นที่จะสอดคล้องกับเงื่อนไขต่าง ๆ สำหรับการสร้างต้นไม้ตัดสินใจ ดังที่ได้กล่าวไว้ว่าการสร้างต้นไม้ตัดสินใจจะต้องใช้กรณีตัวอย่างที่หามาจำนวนหนึ่ง และการทดสอบประสิทธิภาพของต้นไม้ตัดสินใจที่ได้นั้นจะต้องใช้กรณีตัวอย่างที่ต้นไม้ตัดสินใจไม่เคยพบมาก่อนเพื่อเป็นการวัดว่าต้นไม้ตัดสินใจที่ได้สามารถจำแนกข้อมูลที่ไม่เคยพบมาก่อนได้ดีแค่ไหน ในทางปฏิบัติจริงเมื่อเรามีชุดข้อมูลที่จะนำไปสร้างต้นไม้ตัดสินใจและทดสอบประสิทธิภาพแล้วนั้น เราจะนำชุดข้อมูลนั้น ๆ ไปแบ่งเป็น 2 ส่วนคือชุดข้อมูลสำหรับฝึกฝนหรือกรณีตัวอย่างส่วนที่จะนำไปสร้างต้นไม้ตัดสินใจและชุดข้อมูลสำหรับทดสอบหรือกรณีตัวอย่างส่วนที่จะนำไปทดสอบประสิทธิภาพซึ่งผู้เขียนจะอธิบายถึงการแบ่งชุดข้อมูลของงานวิจัยนี้ในบทถัดไป การสร้างต้นไม้ตัดสินใจด้วยกระบวนการของอัลกอริทึม MII-ID3 มีดังนี้

ขั้นตอนที่ 1 คำนวณค่าเกนความรู้ในแต่ละแอตทริบิวต์โดยใช้สมการที่ 2.2 หลังจากนั้นเลือกแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุด ถ้าแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดมีมากกว่า 1 แอตทริบิวต์ให้ทำการสุ่มเลือกแอตทริบิวต์ที่ดีที่สุดมาหนึ่งแอตทริบิวต์

ขั้นตอนที่ 2 สร้างโหนดและแบ่งกรณีตัวอย่างไปตามค่าที่ไม่ซ้ำกันของแอตทริบิวต์ที่ถูกเลือกจากขั้นตอนก่อนหน้า

ขั้นตอนที่ 3 คำนวณ Threshold ของความน่าจะเป็นในการละเลยกรณีตัวอย่างส่วนน้อยหรือความน่าจะเป็นในการแปลงโหนดให้เป็นโหนดคำตอบด้วยการนับเสียงส่วนมากโดยใช้สมการดังนี้

$$Prop = e^{-\left(\frac{CurDepth}{Remainder(I,A)*RemAtts}\right)} \quad \textbf{(4.1)}$$

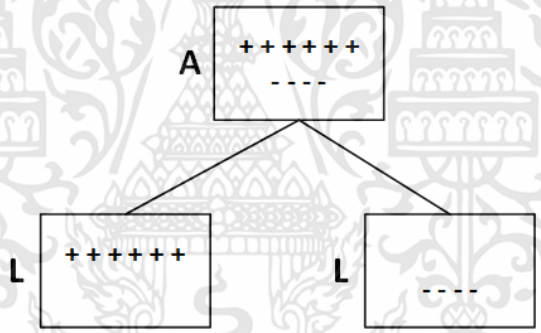
Prop คือ Threshold ของความน่าจะเป็นในการแปลงโหนดให้เป็นโหนดคำตอบ

Remainder(I,A) คือผลรวมของค่าเอนโทรปีของแต่ละค่าที่ไม่ซ้ำกันในแอตทริบิวต์ A หลังจากกรณีตัวอย่างถูกแบ่งไปตามค่าที่ไม่ซ้ำกันต่าง ๆ ของ A คำนวณจากสมการที่ 2.4

CurDepth คือความลึกปัจจุบันของต้นไม้ตัดสินใจ

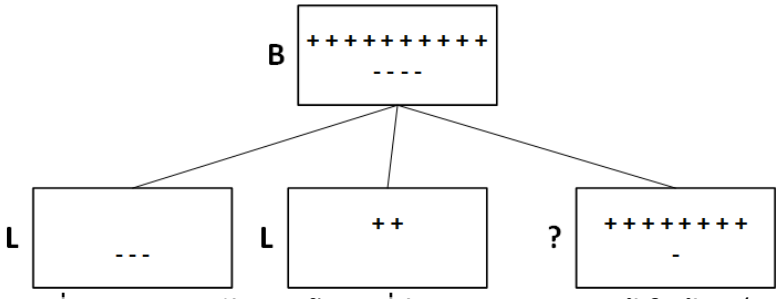
RemAtts คือจำนวนของแอตทริบิวต์ที่เหลืออยู่และไม่เป็นศูนย์

การละเลยกรณีตัวอย่างส่วนน้อยจะขึ้นอยู่กับค่า Prop ของโหนดนั้น ๆ ยิ่งค่า Prop มาก โอกาสในการละเลยกรณีตัวอย่างหรือแปลงโหนดจะยิ่งน้อยก็คือมันจะแปรผกผันต่อกัน ในสมการ Prop มีตัวแปรที่สำคัญอยู่สามตัว ตัวแรกคือ Remainder ถ้าค่า Remainder ของโหนดที่ระบุด้วยแอตทริบิวต์ใดเท่ากับศูนย์ นั้นหมายความว่าไม่มีความไม่แน่นอนหรือไม่เหลือกรณีตัวอย่างที่จำแนกไม่ได้ แสดงว่าโหนดที่ระบุด้วยแอตทริบิวต์นั้น ๆ เป็นโหนดการจำแนกประเภทหรือโหนดที่มีโหนดลูกทั้งหมดเป็นโหนดคำตอบ ดังนั้นการละเลยกรณีตัวอย่างจะไม่เกิดขึ้น



รูปที่ 4.1 แสดงตัวอย่างโหนดการจำแนกประเภทที่มีค่า Remainder เป็นศูนย์

จากรูปที่ 4.1 โหนด A เป็นโหนดการจำแนกประเภทหรือโหนดที่มีโหนดลูกเป็นโหนดคำตอบทั้งหมด (โหนด L) เครื่องหมาย + และ – คือกรณีตัวอย่างที่เป็นคลาสบวกและลบตามลำดับ ดังที่ได้กล่าวไว้ค่า Remainder คือค่าเอนโทรปีหรือค่าความไม่แน่นอนโดยรวมหลังจากทดสอบการแบ่งกรณีตัวอย่างของโหนดนั้น ๆ ดังนั้นค่า Remainder ของโหนด A จะหาได้จากค่าเอนโทรปีของโหนดลูกทั้งหมดของ A รวมกัน จากตัวอย่างรูปที่ 4.1 เนื่องจากโหนดลูกของ A เป็นโหนดคำตอบทั้งหมด ในแต่ละโหนดคำตอบจะมีเพียง 1 คลาส ดังนั้นค่าความไม่แน่นอนหรือค่าเอนโทรปีจึงเท่ากับศูนย์ทั้งสองโหนดคำตอบ จึงส่งผลให้ค่าเอนโทรปีโดยรวมหรือ Remainder ของโหนด A เท่ากับศูนย์ เมื่อ Remainder เท่ากับศูนย์การละเลยกรณีตัวอย่างจึงไม่ควรกระทำ กรณีที่ค่า Remainder มีค่าที่น้อยมาก ๆ หรือเข้าใกล้ศูนย์หมายความว่าโหนดนั้น ๆ มีโหนดลูกที่มีกรณีตัวอย่างส่วนน้อย ดังนั้นโหนดลูกของมันควรที่จะทำการละเลยกรณีตัวอย่างส่วนน้อยโดยการแปลงเป็นโหนดคำตอบและระบุด้วยคลาสที่เป็นเสียงส่วนมาก



รูปที่ 4.2 แสดงตัวอย่างโหนดที่มีค่า Remainder เข้าใกล้ศูนย์

จากรูปที่ 4.2 ค่า Remainder ของโหนด B จะเท่ากับค่าเอนโทรปีของโหนดลูกทั้งสามโหนดของมันรวมกัน โหนดลูกสองโหนดแรกจากทางซ้ายเป็นโหนดคำตอบ ค่าเอนโทรปีในโหนดคำตอบจะเท่ากับศูนย์ดังนั้นค่า Remainder ของโหนด B จะขึ้นอยู่กับโหนด ? เพียงอย่างเดียว เมื่อพิจารณาโหนด ? จะเห็นว่าค่าเอนโทรปีหรือค่าความไม่แน่นอนของ ? มีค่าที่น้อยมากหรือเข้าใกล้ศูนย์เพราะว่ามันเกือบจะเป็นระเบียบหรือแน่นอนอยู่แล้ว มีกรณีตัวอย่างเพียงตัวเดียวที่เป็นคลาสอื่น ดังนั้นโหนด ? เป็นโหนดที่มีกรณีตัวอย่างส่วนน้อยอยู่และค่า Remainder ของโหนด B จะเข้าใกล้ศูนย์ ดังนั้นโหนดลูกของโหนด B ซึ่งก็คือโหนด ? ควรจะถูกแปลงเป็นโหนดคำตอบและถูกระบุคลาสคำตอบด้วยเสียงส่วนมากของคลาสนั้นคือคลาสบวกนั้นเอง นี้คือเหตุผลที่เราใช้ค่า Remainder เข้ามาเป็นตัวแปรในสมการ Prop เพราะมันบ่งบอกถึงการมีกรณีตัวอย่างส่วนน้อยอยู่ ถ้าค่า Remainder เข้าใกล้ศูนย์หรือน้อยมาก ๆ จะส่งผลให้ค่า Prop มีค่าน้อย เมื่อ Prop มีค่าน้อยโอกาสในการแปลงโหนดหรือละเลยกรณีตัวอย่างส่วนน้อยจะมีมาก ซึ่งตรงตามเหตุผลที่ว่าโหนดที่มีค่า Remainder น้อย ๆ โหนดลูกของมันควรจะถูกละเลยกรณีตัวอย่างส่วนน้อย อีกสองตัวแปรที่สำคัญได้แก่ CurDepth และ RemAtts ดังที่ได้กล่าวไว้ว่าเราไม่ควรละเลยกรณีตัวอย่างส่วนน้อยที่ตอนเริ่มต้นในการสร้างต้นไม้ตัดสินใจ เพราะโหนดด้านบนมีความสำคัญมาก ๆ ถ้าหากเราแปลงโหนดตั้งแต่ด้านบนให้เป็นโหนดคำตอบเลย ความแม่นยำในการจำแนกจะลดลงอย่างมาก ตัวแปร CurDepth คือจำนวนความลึกปัจจุบันของต้นไม้ตัดสินใจ ในตอนเริ่มแรกของการสร้างต้นไม้ตัดสินใจจะยังมีระดับความลึกที่น้อย ๆ อยู่ ดังนั้นเมื่อ CurDepth มีค่าน้อยจะส่งผลให้ Prop มีค่ามากและโอกาสในการละเลยกรณีตัวอย่างส่วนน้อยจะมีน้อย ส่วนตัวแปร RemAtts คือจำนวนของแอตทริบิวต์ที่เหลืออยู่ ในช่วงเริ่มแรกของการสร้างต้นไม้ตัดสินใจจะยังมีแอตทริบิวต์ที่ยังไม่ได้ถูกใช้ในการสร้างโหนดเป็นจำนวนมาก ดังนั้น RemAtts จะมีค่ามากในช่วงเริ่มแรกจะส่งผลให้ Prop มีค่ามาก เมื่อ Prop มีค่ามากทำให้โอกาสในการละเลยกรณีตัวอย่างส่วนน้อยจะมีค่าน้อย สรุปได้ว่าตัวแปรทั้งสองนี้มีไว้เพื่อป้องกันการละเลยกรณีตัวอย่างในช่วงเริ่มแรกของการสร้างต้นไม้ตัดสินใจ เพื่อป้องกันไม่ให้ความแม่นยำในการจำแนกประเภทลดลง

ขั้นตอนที่ 4 สุ่มตัวเลขระหว่าง 0 และ 1 ถ้าตัวเลขที่สุ่มมากกว่าค่า Prop ในสมการที่ 4.1 และค่า Remainder(I,A) ไม่เท่ากับศูนย์ ให้แปลงโหนดให้เป็นโหนดคำตอบที่ระบุด้วยเสียงส่วนมากของคลาสในกรณีตัวอย่างที่เหลืออยู่

ขั้นตอนที่ 5 ทำซ้ำขั้นตอนที่ 1-4 โดยใช้แอตทริบิวต์และกรณีตัวอย่างที่เหลืออยู่จนกระทั่งสามารถจำแนกประเภทกรณีตัวอย่างได้ทั้งหมด

จากขั้นตอนทั้งหมดของอัลกอริทึม MII-ID3 ขั้นตอนที่ 3 และ 4 คือขั้นตอนที่ถูกเพิ่มเข้ามา ขั้นตอนอื่น ๆ จะเหมือนกับขั้นตอนของอัลกอริทึม ID3 แบบดั้งเดิม เราสามารถดัดแปลงรหัสเทียมของอัลกอริทึม ID3 แบบดั้งเดิมในรูปที่ 2.13 โดยการเปลี่ยนบรรทัดที่ 1 เป็น “function MII-ID3 (instances, parent_instances, attributes, state) returns tree” ต่อมาเปลี่ยนบรรทัดที่ 6 เป็น “else if attributes is empty or state equal to true” และแทรกคำสั่งทั้งหมดในรูปที่ 4.3 ระหว่างบรรทัด 11 และ 12 ของรูปที่ 2.13 รหัสเทียมแบบสมบูรณ์ของอัลกอริทึม MII-ID3 แสดงในรูปที่ 4.4

```
r ← Randomize decimal between 0 and 1
prop ← Compute threshold of probability value
if r is greater than prop and Remainder(A) is not zero,
    then set value of state to true
else
    then set value of state to false
```

รูปที่ 4.3 แสดงบางส่วนของคำสั่งในอัลกอริทึม MII-ID3

```
1   function MII-ID3 (instances, parent_instances, attributes, state) returns a tree
      if instances is empty,
        then return the leaf node which labeled by the majority of parent_instances
      else if all instances have the same classification,
        then return the leaf node which labeled by its classification
6     else if attributes is empty or state equal to true,
        then return the leaf node which labeled by the majority of instances
      else
        Compute information gains for all attributes using instances and attributes
        A ← Select the attribute which has the best information gain
11      Insert the new node labeled by attribute A to tree
        r ← Randomize decimal between 0 and 1
        prop ← Compute threshold of probability value
        if r is greater than prop and Remainder(A) is not zero,
          then set value of state to true
        else
          then set value of state to false
18      for each value v of A do
          ins ← Select instances of value v
          subtree ← ID3 (ins, instances, attribute–A)
          Add a branch to tree with label v and subtree
        return tree
```

รูปที่ 4.4 แสดงรหัสเทียมของอัลกอริทึม MII-ID3 เมื่อตัวแปร state ถูกตั้งค่าเริ่มต้นเป็น false

จากรหัสเทียมในรูปที่ 4.4 มีตัวแปรที่รับค่าคือตัวแปร instances, parent_instances และ attributes โดยจะตั้งค่าเริ่มต้นเหมือนกับอัลกอริทึม ID3 ปกติ มีตัวแปรสำคัญที่เพิ่มเข้ามาคือตัวแปร state ก่อนที่จะเริ่มตัวแปร state จะถูกตั้งค่าเริ่มต้นเป็น false ตัวแปร state ก็คือตัวแปรประเภทบูลีน (Boolean) ที่เอาไว้ระบุสถานะว่าหลังจากการคำนวณค่า Prop แล้วโหนดนี้จะละเลยกรณีตัวอย่างส่วนน้อยหรือไม่ ถ้า state = true หมายความว่าต้องแปลงโหนดให้เป็นโหนดคำตอบหรือต้องละเลยกรณีตัวอย่างส่วนน้อย และถ้า state = false หมายถึงไม่ต้องทำการละเลยกรณีตัวอย่างส่วนน้อย เมื่ออัลกอริทึม MII-ID3 เริ่มทำงานจะทำการตรวจสอบเงื่อนไขดังนี้

เงื่อนไขที่ 1 : ถ้าตัวแปร instances เป็นเซตว่างหรือกรณีตัวอย่างปัจจุบันไม่มีเหลือแล้ว ให้คืนค่าโหนดคำตอบที่ระบุด้วยเสียงส่วนมากของคลาสใน parent_instances มิฉะนั้นให้ไปตรวจสอบต่อในเงื่อนไขต่อไป

เงื่อนไขที่ 2 : ถ้ากรณีตัวอย่างปัจจุบันในตัวแปร instances เป็นคลาสเดียวกันทั้งหมด ให้คืนค่าโหนดคำตอบที่ระบุด้วยคลาสนั้น ๆ มิฉะนั้นให้ไปตรวจสอบต่อในเงื่อนไขต่อไป

เงื่อนไขที่ 3 : ถ้าตัวแปร attributes เป็นเซตว่างหรือแอตทริบิวต์ถูกใช้จนหมด หรือตัวแปร state เท่ากับ true ให้คืนค่าโหนดคำตอบที่ระบุด้วยเสียงส่วนมากของคลาสใน instances มิฉะนั้นให้ไปยังเงื่อนไขต่อไป

เงื่อนไขที่ 4 : ถ้าตัวแปรทั้งสี่ไม่ตรงกับ 3 เงื่อนไขแรกให้กระทำขั้นตอนดังนี้

1. คำนวณค่าเกนความรู้ของแอตทริบิวต์ทั้งหมดโดยใช้ตัวแปร instances และ attributes
2. เลือกแอตทริบิวต์ที่มีค่าเกนความรู้มากที่สุดมาเก็บไว้ในตัวแปร A
3. สร้างโหนดใหม่โดยระบุตามแอตทริบิวต์ในตัวแปร A
4. สุ่มเลขทศนิยมระหว่าง 0 และ 1 เก็บค่าเลขที่สุ่มได้ไว้ในตัวแปร r
5. คำนวณค่า Threshold ของความน่าจะเป็นในการละเลยกรณีตัวอย่างส่วนน้อยแล้วเก็บค่าไว้ในตัวแปร Prop
6. ถ้า r มากกว่า Prop และค่า Remainder ของแอตทริบิวต์ A ไม่เท่ากับศูนย์แล้วค่าของตัวแปร state จะเท่ากับ true มิฉะนั้นจะเท่ากับ false
7. วนลูป (Loop) ตามค่าที่ไม่ซ้ำกันทั้งหมดของ A โดยเริ่มจากค่าแรกและกระทำดังนี้
    7.1 นำค่าที่ไม่ซ้ำกันปัจจุบันมาเก็บไว้ในตัวแปร v
    7.2 เลือกกรณีตัวอย่างที่มีแอตทริบิวต์ A เป็นค่า v มาเก็บไว้ในตัวแปร ins
    7.3 เรียกทำฟังก์ชัน MII-ID3 ก็คือการ recursion โดยส่งพารามิเตอร์ตามลำดับเป็นตัวแปร ins, instances, attributes และ state โดยลบแอตทริบิวต์ A คือแอตทริบิวต์ที่ถูกเลือกแล้วออกจากตัวแปร attributes และคืนค่าฟังก์ชันกลับมาเก็บไว้ในตัวแปร subtree
    7.4 ต่อโหนด subtree ไปยัง tree และระบุกิ่งของ tree เป็น v
8. คืนค่าตัวแปร tree

เมื่อได้ต้นไม้ตัดสินใจที่สร้างจากอัลกอริทึม MII-ID3 เรียบร้อยแล้ว ต่อไปจะเป็นการทดสอบประสิทธิภาพการจำแนกประเภทของต้นไม้ตัดสินใจที่ได้โดยใช้ชุดข้อมูลสำหรับทดสอบ กรณีตัวอย่างแต่ละกรณีที่อยู่ในชุดข้อมูลสำหรับทดสอบจะถูกนำไปจำแนกประเภทผ่านต้นไม้ตัดสินใจทีละกรณีตัวอย่างและจะจำแนกประเภทโดยวิ่งลงไปตามต้นไม้ตัดสินใจโดยลงไปในเส้นทางตามค่าที่มีในกรณีตัวอย่างจนถึงโหนดคำตอบ เมื่อถึงโหนดคำตอบแล้วจะทำการตรวจสอบว่าคลาสคำตอบที่ได้จากโหนดคำตอบในต้นไม้ตัดสินใจนั้นตรงกับคลาสของกรณีตัวอย่างนั้น ๆ ในชุดข้อมูลสำหรับทดสอบหรือไม่ ถ้าตรงกันหมายความว่าต้นไม้ตัดสินใจที่ได้สามารถจำแนกประเภทกรณีตัวอย่างนั้นได้ถูกต้อง ทำแบบนี้กับทุก ๆ กรณีตัวอย่างแล้วจึงวัดร้อยละความแม่นยำในการจำแนกประเภทออกมา ซึ่งร้อยละความแม่นยำในการจำแนกจะเท่ากับจำนวนกรณีตัวอย่างที่ตอบถูกทั้งหมดหารด้วยจำนวนกรณีตัวอย่างทั้งหมดคูณด้วยหนึ่งร้อย

บทที่ 5
ผลการทดลอง

5.1 ชุดข้อมูลที่ใช้ในการทดลอง

สำหรับการทดลองนี้เราได้เลือก 17 ชุดข้อมูลจากฐานข้อมูล UCI [6] เพื่อทดสอบประสิทธิภาพของอัลกอริทึมที่เสนอ โดยชุดข้อมูลที่คัดเลือกมานั้นมีทั้งขนาดเล็กไปจนถึงขนาดใหญ่ ตารางที่ 5.1 แสดงจำนวนกรณีตัวอย่าง, จำนวนแอตทริบิวต์ และจำนวนคลาสในแอตทริบิวต์เป้าหมายของชุดข้อมูล 17 ชุด

ตารางที่ 5.1 แสดงลักษณะชุดข้อมูลที่ใช้ในการทดลอง

ชุดข้อมูล	จำนวนกรณีตัวอย่าง	จำนวนแอตทริบิวต์	จำนวนคลาส
Connect-4	67557	42	3
Phishing Websites	11055	30	2
Insurance Company Benchmark (COIL2000)	9822	85	2
Molecular Biology (Splice-junction Gene Sequences)	3190	60	3
Soybean (Large)	683	35	19
Audiology (Standardized)	226	69	24
Balance Scale	625	4	3
Car Evaluation	1728	6	4
Chess (King-Rook vs. King-Pawn)	3196	36	2
Congressional Voting Records	435	16	2
Hayes-Roth	160	4	3
MONK's Problems (monks-1)	556	6	2
Mushroom	8124	22	2
Nursery	12960	8	5
SPECT Heart	267	22	2
Tic-Tac-Toe Endgame	958	9	2
Firm-Teacher_Clave-Direction_Classification	10800	16	7

5.2 เงื่อนไขในการทดลอง

ก่อนทำการทดลองชุดข้อมูลทั้งหมดในตารางที่ 5.1 จะถูกแบ่งเป็น 2 ส่วนคือชุดข้อมูลสำหรับฝึกฝนและชุดข้อมูลสำหรับทดสอบ การแบ่งข้อมูลของงานวิจัยนี้นั้นจะต้องกระจายกรณีตัวอย่างในแต่ละคลาสให้เท่า ๆ กันไปในแต่ละส่วนเพื่อให้สัดส่วนของคลาสเป็นสัดส่วนที่ใกล้เคียงกันทั้งชุดข้อมูลสำหรับฝึกฝนและชุดข้อมูลสำหรับทดสอบ ยกตัวอย่างเช่นสมมุติชุดข้อมูลที่จะทำการทดลองคือชุดข้อมูลการเล่นเทนนิสใน 14 วันจากตารางที่ 2.3 ชุดข้อมูลนี้มีกรณีตัวอย่างทั้งหมด 14 กรณีตัวอย่าง และใน 14 กรณีตัวอย่างนี้มีจำนวนกรณีตัวอย่างที่ตอบคลาสเล่น 9 กรณี และมีจำนวนกรณีตัวอย่างที่ตอบคลาสไม่เล่น 5 กรณี งานวิจัยนี้จะทำการสุ่มแบ่งกรณีตัวอย่างของชุดข้อมูลนี้ให้เป็น 2 ส่วน ส่วนแรกคือชุดข้อมูลสำหรับฝึกฝนจะทำการสุ่มกรณีตัวอย่างที่ตอบคลาสเล่น 4 กรณีและกรณีตัวอย่างที่ตอบคลาสไม่เล่น 2 กรณี โดยรวมแล้วชุดข้อมูลสำหรับฝึกฝนจะมีกรณีตัวอย่างทั้งหมด 6 กรณี ส่วนที่สองคือชุดข้อมูลสำหรับทดสอบจะมีจำนวนกรณีตัวอย่างที่เหลือและตอบคลาสเล่น 5 กรณีและจำนวนกรณีตัวอย่างที่เหลือและตอบคลาสไม่เล่น 3 กรณี โดยรวมแล้วชุดข้อมูลสำหรับทดสอบจะมีกรณีตัวอย่างทั้งหมด 8 กรณี จะเห็นว่าเมื่อเราแบ่งข้อมูลจริง ๆ จะไม่ได้เท่ากันเพราะจำนวนของกรณีตัวอย่างในแต่ละคลาสอาจจะเป็นจำนวนคี่ซึ่งเมื่อพยายามแบ่งเป็น 2 ส่วนก็จะหารไม่ลงตัว ดังนั้นส่วนที่เกินออกมาจึงนำไปเพิ่มให้กับชุดข้อมูลสำหรับทดสอบหรือส่วนที่สอง เพื่อความเข้าใจมากขึ้นให้พิจารณาตารางที่ 5.2 และ 5.3 แสดงกรณีตัวอย่างที่ถูกแบ่งจากชุดข้อมูลการเล่นเทนนิสใน 14 วัน

ตารางที่ 5.2 แสดงชุดข้อมูลสำหรับฝึกฝนที่แบ่งจากชุดข้อมูลการเล่นเทนนิสใน 14 วัน

กรณีตัวอย่าง	แอตทริบิวต์				แอตทริบิวต์
	ทัศนียภาพ	อุณหภูมิ	ความชื้น	ลมแรง	เป้าหมาย
2	แดดจัด	ร้อน	สูง	ใช่	ไม่เล่น
4	ฝนตก	ไม่เย็นมาก	สูง	ไม่ใช่	เล่น
6	ฝนตก	เย็น	ปกติ	ใช่	ไม่เล่น
7	มีเมฆมาก	เย็น	ปกติ	ใช่	เล่น
11	แดดจัด	ไม่เย็นมาก	ปกติ	ใช่	เล่น
13	มีเมฆมาก	ร้อน	ปกติ	ไม่ใช่	เล่น

ตารางที่ 5.3 แสดงชุดข้อมูลสำหรับทดสอบที่แบ่งจากชุดข้อมูลการเล่นเทนนิสใน 14 วัน

กรณีตัวอย่าง	แอตทริบิวต์				แอตทริบิวต์
	ทัศนียภาพ	อุณหภูมิ	ความชื้น	ลมแรง	เป้าหมาย
1	แดดจัด	ร้อน	สูง	ไม่ใช่	ไม่เล่น
3	มีเมฆมาก	ร้อน	สูง	ไม่ใช่	เล่น
5	ฝนตก	เย็น	ปกติ	ไม่ใช่	เล่น
8	แดดจัด	ไม่เย็นมาก	สูง	ไม่ใช่	ไม่เล่น
9	แดดจัด	เย็น	ปกติ	ไม่ใช่	เล่น
10	ฝนตก	ไม่เย็นมาก	ปกติ	ไม่ใช่	เล่น
12	มีเมฆมาก	ไม่เย็นมาก	สูง	ใช่	เล่น
14	ฝนตก	ไม่เย็นมาก	สูง	ใช่	ไม่เล่น

การทดลองของอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม MII-ID3 จะทำการทดลอง 1000 ครั้งต่อ 1 ชุดข้อมูลโดยใช้ชุดข้อมูลสำหรับฝึกฝนและชุดข้อมูลสำหรับทดสอบเหมือนกันทั้งสองอัลกอริทึมแล้วหาค่าเฉลี่ยออกมา ที่ต้องทดลองถึง 1000 ครั้งและหาค่าเฉลี่ยเพราะว่าในแต่ละรอบของการทดลองของอัลกอริทึม ID3 และ MII-ID3 จะได้ต้นไม้ตัดสินใจผลลัพธ์ที่แตกต่างกันไปในแต่ละรอบ เพราะอัลกอริทึมทั้งสองมีระบบการสุ่มเลือกอยู่ ในอัลกอริทึม ID3 จะสุ่มเลือกแอตทริบิวต์ที่จะนำมาสร้างโหนดเมื่อมีแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดมากกว่า 1 แอตทริบิวต์ ในส่วนของอัลกอริทึม MII-ID3 จะทำการสุ่มแอตทริบิวต์เช่นเดียวกับอัลกอริทึม ID3 และยังมีการหา Threshold ของความน่าจะเป็นที่จะละเลยกรณีตัวอย่างอีก ดังนั้นในแต่ละรอบจะสุ่มได้ไม่เหมือนกัน เพื่อการวัดประสิทธิภาพที่สเถียรเราควรทดลองหลาย ๆ ครั้งและหาค่าเฉลี่ยของตัววัดนั้น ๆ ออกมาซึ่ง 1000 รอบนั้นเพียงพอต่อการวัด

5.3 การจัดการค่าสูญหาย (Missing value)

เนื่องจากชุดข้อมูลที่ใช้ในการทดสอบมีชุดข้อมูลที่มีค่าสูญหายได้แก่ Soybean, Audiology, Congressional Voting Records และ Mushroom จึงจำเป็นต้องจัดการค่าสูญหายในชุดข้อมูลเหล่านี้ก่อนจะนำไปทดลองเพื่อความถูกต้องในการจำแนกข้อมูล ยกตัวอย่างชุดข้อมูลที่มีค่าสูญหายดังรูปที่ 5.1

Instance	Attribute 1	Class
1	T	Y
2	T	N
3	F	Y
4	F	N
5	F	N
6	F	N
7	?	N
8	?	Y
9	?	Y
10	?	Y

รูปที่ 5.1 แสดงตัวอย่างชุดข้อมูลที่มีค่าสูญหาย

จากรูปที่ 5.1 คือตัวอย่างชุดข้อมูลที่มีแอตทริบิวต์เดียวคือ Attribute 1 ที่มีค่าที่ไม่ซ้ำกันคือ T และ F ส่วนค่า ? ใน Attribute 1 คือค่าสูญหาย ชุดข้อมูลในรูปที่ 5.1 มี Class คือ Y และ N วิธีจัดการค่าสูญหายในงานวิจัยนี้คือ เราอยากจะทราบว่าค่า ? ใน Attribute 1 ควรจะเป็นค่า T หรือค่า F ดังนั้นขั้นตอนแรกให้แยกกรณีตัวอย่างที่มี ? ใน Attribute 1 ออกจากกรณีตัวอย่างที่ไม่มีค่า ? ใน Attribute 1 ดังแสดงในรูปที่ 5.2

Instance	Attribute 1	Class
1	T	Y
2	T	N
3	F	Y
4	F	N
5	F	N
6	F	N

(a)

Instance	Attribute 1	Class
7	?	N
8	?	Y
9	?	Y
10	?	Y

(b)

รูปที่ 5.2 แสดงการแยกกรณีตัวอย่างที่มีค่า ? ใน Attribute 1 ออกจากกรณีตัวอย่างปกติ

จากรูปที่ 5.2 (a) คือกรณีตัวอย่างปกติที่ไม่มีค่าสูญหายหรือค่า ? และ (b) คือกรณีตัวอย่างที่มีค่าสูญหายใน Attribute 1 พิจารณาที่ (b) มีกรณีตัวอย่างที่เป็นคลาส N คือกรณีตัวอย่างที่ 7 และมีกรณีตัวอย่างที่เป็นคลาส Y คือกรณีตัวอย่างที่ 8, 9 และ 10 คำถามคือเราจะแทนค่า ? ในกรณีตัวอย่างที่เป็นคลาส N และ Y ใน (b) ได้อย่างไร

พิจารณากรณีตัวอย่างที่เป็นคลาส N ใน (a) พบว่ามีกรณีตัวอย่างที่มีค่า T ใน Attribute 1 จำนวน 1 กรณีตัวอย่างคือกรณีตัวอย่างที่ 2 และมีกรณีตัวอย่างที่มีค่า F ใน Attribute 1 จำนวน 3 กรณีตัวอย่างคือกรณีตัวอย่างที่ 4, 5 และ 6 กรณีตัวอย่างที่เป็นค่า F และมีคลาส N ใน (a) มีจำนวนมากกว่ากรณีตัวอย่างที่เป็นค่า T และมีคลาส N ดังนั้นเราจะแทนค่า ? ของกรณีตัวอย่างที่มีคลาส N ใน (b) ก็คือกรณีตัวอย่างที่ 7 ด้วยค่า F

พิจารณากรณีตัวอย่างที่เป็นคลาส Y ใน (a) พบว่ามีกรณีตัวอย่างที่มีค่า T ใน Attribute 1 จำนวน 1 กรณีตัวอย่างคือกรณีตัวอย่างที่ 1 และมีกรณีตัวอย่างที่มีค่า F ใน Attribute 1 จำนวน 1 กรณีตัวอย่างคือกรณีตัวอย่างที่ 3 จะเห็นว่ากรณีตัวอย่างที่เป็นค่า T และมีคลาส Y ใน (a) มีจำนวนเท่ากับกรณีตัวอย่างที่เป็นค่า F และมีคลาส Y ดังนั้นเราจะแทนค่า ? ของกรณีตัวอย่างที่มีคลาส Y ใน (b) ก็คือกรณีตัวอย่างที่ 8, 9 และ 10 ด้วยการสุ่มระหว่างค่า T และ F เนื่องจากมีจำนวนเท่ากัน ดังนั้นชุดข้อมูลที่ได้จากการจัดการค่าสูญหายของชุดข้อมูลในรูปที่ 5.1 จะเป็นไปได้ 2 แบบ แบบที่ 1 แสดงในรูปที่ 5.3 เมื่อแทนค่า ? ของกรณีตัวอย่างที่มีคลาส N ด้วยค่า F และแทนค่า ? ของกรณีตัวอย่างที่มีคลาส Y ด้วย T (จากการสุ่มระหว่าง T และ F) ต่อมาแบบที่ 2 แสดงในรูปที่ 5.4 เมื่อแทนค่า ? ของกรณีตัวอย่างที่มีคลาส N ด้วยค่า F และแทนค่า ? ของกรณีตัวอย่างที่มีคลาส Y ด้วย F (จากการสุ่มระหว่าง T และ F)

Instance	Attribute 1	Class
1	T	Y
2	T	N
3	F	Y
4	F	N
5	F	N
6	F	N
7	F	N
8	T	Y
9	T	Y
10	T	Y

รูปที่ 5.3 แสดงชุดข้อมูลแบบที่ 1 ที่ได้จากจากการจัดการค่าสูญหาย

Instance	Attribute 1	Class
1	T	Y
2	T	N
3	F	Y
4	F	N
5	F	N
6	F	N
7	F	N
8	F	Y
9	F	Y
10	F	Y

รูปที่ 5.4 แสดงชุดข้อมูลแบบที่ 2 ที่ได้จากการจัดการค่าสูญหาย

5.4 การจัดกลุ่มชุดข้อมูล

เพื่อให้ง่ายต่อการวิเคราะห์ผลการทดลองผู้เขียนจะทำการแบ่งกลุ่มชุดข้อมูลเป็น 3 กลุ่มตามขนาดและความซับซ้อนได้แก่ กลุ่มชุดข้อมูลขนาดใหญ่, กลุ่มชุดข้อมูลขนาดกลาง และกลุ่มชุดข้อมูลขนาดเล็ก ซึ่งขนาดและความซับซ้อนของชุดข้อมูลพิจารณาจากจำนวนของกรณีตัวอย่างและจำนวนแอตทริบิวต์ที่ใช้ในการสร้างต้นไม้ตัดสินใจ โดยชุดข้อมูลที่มีขนาดใหญ่และซับซ้อนจริง ๆ นั้นจะส่งผลให้ต้นไม้ตัดสินใจที่ได้มีขนาดใหญ่และมีจำนวนกฎการตัดสินใจมาก เพื่อให้การแบ่งกลุ่มเป็นไปอย่างสมเหตุสมผลที่สุดเราจึงต้องพิจารณาจำนวนแอตทริบิวต์ที่ใช้จริงเท่านั้น เพราะเมื่อนำชุดข้อมูลทั้ง 17 ชุดไปทดลองกับอัลกอริทึม ID3 แบบดั้งเดิมพบว่ามีชุดข้อมูลบางชุดที่มีจำนวนแอตทริบิวต์มาก แต่เมื่อนำไปทดลองสร้างต้นไม้ตัดสินใจโดยใช้เกนความรู้เป็นเกณฑ์ในการเลือกแอตทริบิวต์พบว่าจำนวนของแอตทริบิวต์ที่ถูกใช้ในการสร้างต้นไม้ตัดสินใจจริง ๆ นั้นน้อยกว่าจำนวนแอตทริบิวต์ทั้งหมด ดังนั้นการแบ่งกลุ่มชุดข้อมูลจะพิจารณาจากจำนวนกรณีตัวอย่างและจำนวนแอตทริบิวต์ที่ถูกใช้จริงโดยอิงจากการทดลองอัลกอริทึม ID3 แบบดั้งเดิม 1000 รอบและหาค่าเฉลี่ยของแอตทริบิวต์ที่ใช้จริงในแต่ละชุดข้อมูลออกมา

ตารางที่ 5.4 แสดงจำนวนแอตทริบิวต์ที่ใช้จริงโดยเฉลี่ยจากการทดลองอัลกอริทึม ID3 เป็นจำนวน 1000 รอบในแต่ละชุดข้อมูล

ชุดข้อมูล	จำนวนกรณีตัวอย่าง	จำนวนแอตทริบิวต์	จำนวนแอตทริบิวต์ที่ถูกใช้โดยเฉลี่ย
Connect-4	67557	42	42
Phishing Websites	11055	30	30
Insurance Company Benchmark	9822	85	85
Molecular Biology	3190	60	60
Soybean	683	35	35
Audiology	226	69	**15.79**
Balance Scale	625	4	4
Car Evaluation	1728	6	6
Chess	3196	36	**26.38**
Congressional Voting Records	435	16	16
Hayes-Roth	160	4	4
MONK's Problems	556	6	**5**
Mushroom	8124	22	**4.92**
Nursery	12960	8	8
SPECT Heart	267	22	22
Tic-Tac-Toe Endgame	958	9	9
Firm-Teacher_Clave-Direction_Classification	10800	16	16

จากตารางที่ 5.4 ชุดข้อมูล Audiology, Chess, MONK's Problems และ Mushroom มีจำนวนแอตทริบิวต์ที่ใช้จริงน้อยกว่าจำนวนแอตทริบิวต์ทั้งหมด โดยเฉพาะชุดข้อมูล Audiology, Chess และ Mushroom ที่จำนวนแอตทริบิวต์ที่ใช้จริงนั้นแตกต่างจากจำนวนแอตทริบิวต์ทั้งหมดอย่างชัดเจน

การแบ่งกลุ่มของชุดข้อมูลจะใช้เงื่อนไขดังนี้ เงื่อนไขที่ 1 ถ้าชุดข้อมูลใด ๆ มีจำนวนกรณีตัวอย่างตั้งแต่ 3000 กรณีขึ้นไปและมีจำนวนแอตทริบิวต์ที่ถูกใช้ในการสร้างต้นไม้ตัดสินใจตั้งแต่ 30 ขึ้นไปให้อยู่กลุ่มชุดข้อมูลขนาดใหญ่ เงื่อนไขที่ 2 ถ้าชุดข้อมูลใด ๆ มีจำนวนกรณีตัวอย่างน้อยกว่า 1001 กรณีและมีจำนวนแอตทริบิวต์ที่ถูกใช้น้อยกว่า 11 ให้อยู่กลุ่มชุดข้อมูลขนาดเล็ก เงื่อนไขที่ 3 ถ้าชุดข้อมูลใด ๆ ไม่ตรงตามเงื่อนไขที่ 1 และ 2 ให้อยู่กลุ่มชุดข้อมูลขนาดกลาง

ตารางที่ 5.5 แสดงการจัดกลุ่มของชุดข้อมูล

กลุ่ม	ชุดข้อมูล	จำนวนกรณีตัวอย่าง	จำนวนแอตทริบิวต์/จำนวนแอตทริบิวต์ที่ถูกใช้โดยเฉลี่ย
ขนาดใหญ่	Connect-4	67557	42/42
	Phishing Websites	11055	30/30
	Insurance Company Benchmark	9822	85/85
	Molecular Biology	3190	60/60
ขนาดกลาง	Soybean	683	35/35
	Audiology	226	69/15.79
	Car Evaluation	1728	6/6
	Chess	3196	36/26.38
	Congressional Voting Records	435	16/16
	Mushroom	8124	22/4.92
	Nursery	12960	8/8
	SPECT Heart	267	22/22
	Firm-Teacher_Clave-Direction_Classification	10800	16/16
ขนาดเล็ก	Balance Scale	625	4/4
	Hayes-Roth	160	4/4
	MONK's Problems	556	6/5
	Tic-Tac-Toe Endgame	958	9/9

5.5 ผลการทดลองระหว่างอัลกอริทึม ID3 และ MII-ID3

5.5.1 ความแม่นยำในการจำแนกและความลึกสูงสุดโดยเฉลี่ย

ตารางที่ 5.6 แสดงร้อยละของความแม่นยำในการจำแนกและความลึกสูงสุดโดยเฉลี่ยในการทดลอง 1000 รอบของกลุ่มชุดข้อมูลขนาดใหญ่ระหว่าง ID3 และ MII-ID3

ชุดข้อมูล	ร้อยละของความแม่นยำโดยเฉลี่ย		ความลึกสูงสุดโดยเฉลี่ย		ร้อยละในการลดลงของความลึกสูงสุด
	อัลกอริทึม ID3	อัลกอริทึม MII-ID3	อัลกอริทึม ID3	อัลกอริทึม MII-ID3	
Connect-4	**73.65**	73.59	22.73	**15.73**	30.81
Phishing Websites	**94.91**	93.65	30.00	**10.94**	63.54
Insurance Company Benchmark	90.28	**90.91**	85.00	**24.01**	71.75
Molecular Biology	**88.99**	88.75	60.00	**8.79**	85.35

จากผลการทดลองในตารางที่ 5.6 แสดงให้เห็นว่าจำนวนความลึกสูงสุดโดยเฉลี่ยของต้นไม้ตัดสินใจที่สร้างจากอัลกอริทึม MII-ID3 ลดลงจากอัลกอริทึม ID3 อย่างเห็นได้ชัดโดยอัลกอริทึม MII-ID3 สามารถลดจำนวนความลึกสูงสุดโดยเฉลี่ยได้ร้อยละ 30.81, 63.54, 71.75 และ 85.35 ภายในชุดข้อมูล Connect-4, Phishing Websites, Insurance Company Benchmark และ Molecular Biology ตามลำดับ ในขณะที่ความแม่นยำโดยเฉลี่ยในอัลกอริทึม MII-ID3 ลดลงเพียงเล็กน้อยจากอัลกอริทึม ID3 ซึ่งถือว่าคุ้มค่าที่จะทำการละเลยกรณีตัวอย่างในชุดข้อมูลขนาดใหญ่ เพราะสามารถลดระดับความลึกสูงสุดได้มากแลกกับการที่ความแม่นยำจะลดลงเล็กน้อย ดังเช่นในชุดข้อมูล Connect-4, Phishing Websites และ Molecular Biology ในส่วนของชุดข้อมูล Insurance Company Benchmark อัลกอริทึม MII-ID3 มีความแม่นยำโดยเฉลี่ยเพิ่มขึ้นกว่าอัลกอริทึม ID3 แบบดั้งเดิมเล็กน้อย

ตารางที่ 5.7 แสดงร้อยละของความแม่นยำในการจำแนกและความลึกสูงสุดโดยเฉลี่ยในการทดลอง 1000 รอบของกลุ่มชุดข้อมูลขนาดกลางระหว่าง ID3 และ MII-ID3

ชุดข้อมูล	ร้อยละของความแม่นยำโดยเฉลี่ย		ความลึกสูงสุดโดยเฉลี่ย		ร้อยละในการลดลงของความลึกสูงสุด
	อัลกอริทึม ID3	อัลกอริทึม MII-ID3	อัลกอริทึม ID3	อัลกอริทึม MII-ID3	
Soybean	**86.99**	86.21	35.00	**8.08**	76.93
Audiology	**75.64**	75.01	6.00	**5.80**	3.35
Car Evaluation	**91.58**	83.42	6.00	**4.58**	23.60
Chess	**99.29**	94.82	13.00	**7.41**	43.03
Congressional Voting Records	94.45	**95.39**	16.00	**4.35**	72.83
Mushroom	**100.00**	99.64	4.00	**2.76**	30.90
Nursery	**96.86**	89.16	8.00	**5.37**	32.88
SPECT Heart	**80.77**	79.99	22.00	**7.29**	66.89
Firm-Teacher_Clave-Direction_Classification	**71.71**	67.78	15.34	**10.97**	28.46

จากตารางที่ 5.7 แสดงให้เห็นว่าอัลกอริทึม MII-ID3 สามารถสร้างต้นไม้ตัดสินใจที่มีความลึกสูงสุดน้อยกว่าอัลกอริทึม ID3 อย่างชัดเจนในชุดข้อมูลขนาดกลางเกือบทุกตัวมีเพียงชุดข้อมูล Audiology เท่านั้นที่ความลึกสูงสุดโดยเฉลี่ยลดลงไม่มาก เมื่อพิจารณาที่ความแม่นยำโดยเฉลี่ยในชุดข้อมูล Soybean, Audiology, SPECT Heart และ Mushroom มีความแม่นยำในการจำแนกโดยเฉลี่ยจากอัลกอริทึม MII-ID3 ลดลงจากอัลกอริทึม ID3 แบบดั้งเดิมเพียงเล็กน้อย ในชุดข้อมูล Congressional Voting Records อัลกอริทึม MII-ID3 ไม่เพียงแต่ลดจำนวนความลึกสูงสุด แต่ยังสามารถเพิ่มความแม่นยำโดยเฉลี่ยให้มากขึ้นจากอัลกอริทึม ID3 แบบดั้งเดิมเล็กน้อย ดังนั้นในชุดข้อมูล Soybean, Audiology, SPECT Heart, Congressional Voting Records และ Mushroom ถือว่าคุ้มค่าสำหรับการละเลยกรณีตัวอย่างส่วนน้อยของอัลกอริทึม MII-ID3 ในส่วนของชุดข้อมูล Car Evaluation, Chess, Nursery และ Firm-Teacher_Clave-Direction_Classification แม้ว่าต้นไม้ตัดสินใจที่ได้จากอัลกอริทึม MII-ID3 จะมีความลึกสูงสุดลดลงน้อยกว่าอัลกอริทึม ID3 แบบดั้งเดิมอย่างเห็นได้ชัด แต่ความแม่นยำโดยเฉลี่ยนั้นน้อยกว่าอัลกอริทึม ID3 แบบดั้งเดิมอย่างชัดเจนเช่นกัน ซึ่งไม่คุ้มค่าที่จะละเลยกรณีตัวอย่างส่วนน้อยในชุดข้อมูลเหล่านี้

ตารางที่ 5.8 แสดงร้อยละของความแม่นยำในการจำแนกและความลึกสูงสุดโดยเฉลี่ยในการทดลอง 1000 รอบของกลุ่มชุดข้อมูลขนาดเล็กระหว่าง ID3 และ MII-ID3

ชุดข้อมูล	ร้อยละของความแม่นยำโดยเฉลี่ย		ความลึกสูงสุดโดยเฉลี่ย		ร้อยละในการลดลงของความลึกสูงสุด
	อัลกอริทึม ID3	อัลกอริทึม MII-ID3	อัลกอริทึม ID3	อัลกอริทึม MII-ID3	
Balance Scale	62.61	**64.38**	4.00	**3.78**	5.40
Hayes-Roth	**79.01**	73.69	4.00	**3.20**	19.95
MONK's Problems	**98.19**	86.96	4.00	**3.90**	2.50
Tic-Tac-Toe Endgame	**86.01**	77.68	7.00	**5.16**	26.36

จากตารางที่ 5.8 ชุดข้อมูล Hayes-Roth และ Tic-Tac-Toe Endgame มีความลึกสูงสุดโดยเฉลี่ยลดลงอย่างเห็นได้ชัดโดยลดลงร้อยละ 19.95 และ 26.36 ตามลำดับ แต่เมื่อพิจารณาความแม่นยำโดยเฉลี่ยแล้วลดลงอย่างมากเมื่อเปรียบเทียบกับอัลกอริทึม ID3 แบบดั้งเดิม ในชุดข้อมูล Balance Scale และ MONK's Problems มีความลึกสูงสุดลดลงเพียงเล็กน้อยโดยลดลงร้อยละ 5.4 และ 2.5 ตามลำดับ เมื่อพิจารณาความแม่นยำโดยเฉลี่ยในชุดข้อมูล Balance Scale มีความแม่นยำในอัลกอริทึม MII-ID3 มากกว่าความแม่นยำในอัลกอริทึม ID3 แบบดั้งเดิม แต่สำหรับชุดข้อมูล MONK's Problems ความแม่นยำโดยเฉลี่ยลดลงอย่างมากเมื่อเปรียบเทียบกับอัลกอริทึม ID3 แบบดั้งเดิม ดังนั้นในชุดข้อมูลขนาดเล็กมีเพียงชุดข้อมูล Balance Scale เท่านั้นที่คุ้มค่าในการละเลยกรณีตัวอย่างส่วนน้อย การใช้อัลกอริทึม MII-ID3 ในชุดข้อมูลขนาดเล็กตัวอื่น ๆ ได้แก่ Hayes-Roth, Tic-Tac-Toe Endgame และ MONK's Problems ถือว่าไม่คุ้มค่าแม้ความลึกสูงสุดโดยเฉลี่ยจะลดลงแต่ความแม่นยำโดยเฉลี่ยก็ลดลงอย่างมากเช่นกัน

จากผลการทดลองในตารางที่ 5.6-5.8 ทำให้สังเกตได้ว่าอัลกอริทึม MII-ID3 นั้นเหมาะกับชุดข้อมูลที่มีความซับซ้อนมากหรือชุดข้อมูลขนาดใหญ่ เพราะผลการทดลองที่ได้จะสามารถเห็นความแตกต่างของการลดลงของความลึกสูงสุดโดยเฉลี่ยได้อย่างชัดเจนในขณะที่ความแม่นยำโดยเฉลี่ยลดลงเพียงเล็กน้อย ซึ่งถือว่าคุ้มค่าอย่างมากในการประยุกต์ใช้อัลกอริทึม MII-ID3 กับชุดข้อมูลขนาดใหญ่ ในส่วนของชุดข้อมูลขนาดกลางมีทั้งชุดข้อมูลที่คุ้มค่าเมื่อใช้อัลกอริทึม MII-ID3 และในบางชุดข้อมูลก็ไม่คุ้มค่าเพราะความแม่นยำโดยเฉลี่ยลดลงปะปนกันไป กลุ่มสุดท้ายคือชุดข้อมูลขนาดเล็กอัลกอริทึม MII-ID3 นั้นไม่เหมาะสมกับชุดข้อมูลขนาดเล็ก ชุดข้อมูลขนาดเล็กส่วนใหญ่ไม่คุ้มค่าที่จะใช้อัลกอริทึม MII-ID3 มีเพียงชุดข้อมูลขนาดเล็กส่วนน้อยเท่านั้นที่ใช้อัลกอริทึม MII-ID3 ได้คุ้มค่าและความแม่นยำไม่ตกลง เพื่อความชัดเจนในเรื่องความแม่นยำในการจำแนกของชุดข้อมูลทั้งหมด ให้พิจารณาที่กระบวนการละเลยกรณีตัวอย่างส่วนน้อยของอัลกอริทึม MII-ID3 จากเหตุการณ์ที่เกิดขึ้นจริงในขณะสร้างต้นไม้ตัดสินใจในชุดข้อมูล MONK's Problems ซึ่งในขณะการตัดสินใจที่จะละเลยกรณีตัวอย่างของโหนดตัวหนึ่งที่มีความลึกปัจจุบัน (CurDepth) เป็น 2 มีจำนวนแอตทริบิวต์ที่

เหลืออยู่ (RemAtts) 4 และมีค่า Remainder เท่ากับ 0.78 ดังที่ได้กล่าวไว้ว่าค่า Remainder คือผลรวมของค่าเอนโทรปีของแต่ละค่าที่ไม่ซ้ำกันในแอตทริบิวต์ที่นำมาระบุในโหนดหลังจากกรณีตัวอย่างถูกแบ่งไปตามค่าที่ไม่ซ้ำกันต่าง ๆ ของแอตทริบิวต์นั้น ถ้าโหนดใด ๆ มีค่า Remainder ที่น้อยมาก ๆ หมายความว่าหลังจากโหนดนั้นแบ่งกรณีตัวอย่างไปตามค่าที่ไม่ซ้ำกันแล้วมีโหนดลูกบางตัวของมันที่ได้หลังถูกแบ่งกรณีตัวอย่างมีกรณีตัวอย่างส่วนน้อยอยู่ ดังนั้นโหนดลูกของมันควรจะถูกละเลยกรณีตัวอย่างส่วนน้อย จากกรณีการตัดสินใจในการละเลยกรณีตัวอย่างส่วนน้อยในโหนดตัวหนึ่งในชุดข้อมูล MONK's Problems อัลกอริทึม MII-ID3 จะทำการคำนวณค่า Prop จากสมการที่ 4.1 โดยใช้ค่า CurDepth, RemAtts และ Remainder ของโหนดที่กำลังพิจารณา จากเหตุการณ์ดังกล่าวเมื่อนำค่า Remainder, CurDepth และ RemAtts ไปคำนวณจะได้ค่า Prop เท่ากับ 0.53 ต่อมาอัลกอริทึม MII-ID3 จะทำการสุ่มตัวเลขทศนิยมระหว่าง 0-1 ถ้าตัวเลขที่สุ่มมากกว่าค่า Prop ก็จะทำการละเลยกรณีตัวอย่างในโหนดลูกทุกตัวของโหนดปัจจุบัน จากกรณีนี้พบว่าอัลกอริทึม MII-ID3 สุ่มตัวเลขได้เท่ากับ 0.94 ซึ่งมากกว่าค่า Prop ดังนั้นโหนดลูกของโหนดที่กำลังพิจารณานี้จะถูกละเลยกรณีตัวอย่างส่วนน้อยทั้งหมดโดยการแปลงเป็นโหนดคำตอบที่ระบุด้วยคลาสคำตอบที่มากที่สุด เมื่อพิจารณาย้อนกลับไปที่ค่า Remainder ของโหนดแล้วยังมีค่าสูงอยู่ ซึ่งหมายความว่ายังไม่ควรถูกละเลยกรณีตัวอย่างส่วนน้อยที่โหนดลูกของมัน เพราะจะทำให้ความแม่นยำในการจำแนกลดลง ซึ่งจะเห็นได้ว่าอัลกอริทึม MII-ID3 ใช้การสุ่มตัวเลข ถ้าสุ่มตัวเลขแล้วมากกว่าค่า Prop ก็จะทำการละเลยกรณีตัวอย่าง ดังนั้นแม้ว่าค่า Prop จะมีค่ามากก็ตาม ซึ่งไม่ควรละเลยกรณีตัวอย่างส่วนน้อยอย่างยิ่ง แต่ก็อาจจะมีบางครั้งที่จะเกิดกรณีนี้ขึ้น คือสุ่มได้ตัวเลขที่มากกว่าค่า Prop แม้ค่า Prop จะมีค่าสูงแล้วก็ตามเพราะมันคือการสุ่ม ซึ่งเกิดกรณีนี้บางครั้งหรือบ่อยครั้งไม่เท่ากันในทุก ๆ ชุดข้อมูล ถ้าชุดข้อมูลใดมีกรณีนี้เกิดขึ้นบ่อยครั้งในขณะสร้างต้นไม้ตัดสินใจและยังเป็นชุดข้อมูลขนาดเล็กที่มีจำนวนกรณีตัวอย่างน้อย ๆ อีกโอกาสที่ความแม่นยำลดลงจะมากกว่าชุดข้อมูลขนาดใหญ่ที่มีจำนวนกรณีตัวอย่างมาก นี้ก็เป็นข้อเสียของอัลกอริทึม MII-ID3 ที่เป็นสาเหตุที่ทำให้ความแม่นยำลดลง ซึ่งจะเห็นได้จากผลการทดลองที่ได้ ชุดข้อมูลขนาดใหญ่มีความแม่นยำในอัลกอริทึม MII-ID3 ทั้งน้อยกว่าเพียงเล็กน้อยและมากกว่าอัลกอริทึม ID3 ทุกชุดข้อมูล ชุดข้อมูลขนาดกลางมีทั้งความแม่นยำลดลงอย่างมาก มีทั้งความแม่นยำลดลงเพียงเล็กน้อยคละกันไป ในขณะที่ชุดข้อมูลขนาดเล็กส่วนมากความแม่นยำในอัลกอริทึม MII-ID3 ลดลงอย่างมาก ที่เป็นอย่างนี้นั้นนอกจากความแม่นยำจะขึ้นอยู่กับการเกิดเหตุการณ์ดังกล่าวบ่อยครั้งแล้ว อย่างไรก็ตามความแม่นยำในอัลกอริทึม MII-ID3 จะลดหรือเพิ่มก็ขึ้นอยู่กับชุดข้อมูลสำหรับทดสอบในชุดข้อมูลแต่ละชุดด้วย ดังนั้นเราจะพิจารณาที่การทดสอบประสิทธิภาพของอัลกอริทึม MII-ID3 โดยการใช้ชุดข้อมูลสำหรับทดสอบในชุดข้อมูลแต่ละชุด เนื่องจากความแม่นยำของอัลกอริทึม MII-ID3 ในแต่ละชุดข้อมูลนั้นวัดได้จากการทดสอบประสิทธิภาพ ดั้งนั้นเราจึงต้องพิจารณาเกณฑ์ในการวัดจากการทดสอบประสิทธิภาพของอัลกอริทึม MII-ID3 ในแต่ละชุดข้อมูลได้แก่ ร้อยละโดยเฉลี่ยของจำนวนกรณีตัวอย่างที่ถูกจำแนกโดยโหนด

คำตอบปกติ (%L), ร้อยละโดยเฉลี่ยของจำนวนกรณีตัวอย่างที่ถูกจำแนกโดยโหนดคำตอบที่เกิดจากการละเลยกรณีตัวอย่างส่วนน้อย (%T), ร้อยละของความแม่นยำโดยเฉลี่ยของจำนวนกรณีตัวอย่างที่ถูกจำแนกโดยโหนดคำตอบปกติ (%Acc_L) และร้อยละของความแม่นยำโดยเฉลี่ยของจำนวนกรณีตัวอย่างที่ถูกจำแนกโดยโหนดคำตอบที่เกิดจากการละเลยกรณีตัวอย่างส่วนน้อย (%Acc_T) ซึ่งในอัลกอริทึม MII-ID3 จะมีโหนดคำตอบ 2 ประเภท ประเภทแรกคือโหนดคำตอบปกติที่สร้างจากเงื่อนไขตามเดิมของอัลกอริทึม ID3 และประเภทที่สองคือโหนดคำตอบที่เกิดจากการละเลยกรณีตัวอย่างส่วนน้อยก็คือโหนดที่มีกรณีตัวอย่างภายในมากกว่า 1 คลาสและเป็นกรณีตัวอย่างส่วนน้อยซึ่งถูกทำให้เป็นโหนดคำตอบโดยระบุคลาสคำตอบตามคลาสที่มีกรณีตัวอย่างมากที่สุดโดยไม่สนใจกรณีตัวอย่างในคลาสอื่น ๆ

ตารางที่ 5.9 แสดง %L, %T, %Acc_L และ %Acc_T ในชุดข้อมูลที่มีความแม่นยำโดยเฉลี่ยในอัลกอริทึม MII-ID3 มากกว่าความแม่นยำโดยเฉลี่ยในอัลกอริทึม ID3

กลุ่ม	ชุดข้อมูล	ร้อยละของความแม่นยำโดยเฉลี่ย		%L	%T	%Acc_L	%Acc_T
		อัลกอริทึม ID3	อัลกอริทึม MII-ID3				
ขนาดใหญ่	Insurance Company Benchmark	90.28	**90.91**	90.11	9.89	90.67	92.39
ขนาดกลาง	Congressional Voting Records	94.45	**95.39**	81.50	18.50	96.84	88.10
ขนาดเล็ก	Balance Scale	62.61	**64.38**	58.08	41.92	71.31	52.93

จากตารางที่ 5.9 พิจารณาที่ชุดข้อมูล Insurance Company Benchmark และ Congressional Voting Records นั้นมีความแม่นยำในอัลกอริทึม MII-ID3 สูงทั้งในโหนดคำตอบปกติ (%Acc_L) และโหนดคำตอบที่เกิดจากการละเลยกรณีตัวอย่างส่วนน้อย (%Acc_T) จึงส่งผลให้ความแม่นยำโดยรวมมากกว่าอัลกอริทึม ID3

ชุดข้อมูล Balance Scale แม้ว่าจะมีความแม่นยำในโหนดคำตอบที่เกิดจากการละเลยกรณีตัวอย่างส่วนน้อยต่ำลง (%Acc_T = 52.93) แต่มีความแม่นยำที่โหนดคำตอบปกติ (%Acc_L = 71.31) สูงกว่าความแม่นยำในอัลกอริทึม ID3 (62.61%) ที่จำแนกโดยโหนดคำตอบปกติ 100 เปอร์เซ็นต์อย่างชัดเจน ความแม่นยำในอัลกอริทึม MII-ID3 จึงมากกว่าความแม่นยำในอัลกอริทึม ID3 แบบดั้งเดิม

การวิเคราะห์ให้ชัดเจนจะทราบว่า %L, %T, %Acc_L และ %Acc_T ที่วัดออกมาจากการทดสอบด้วยอัลกอริทึม MII-ID3 นั้นส่งผลต่อความแม่นยำโดยรวมของอัลกอริทึม MII-ID3 อย่างไรในเชิงตัวเลข ตัวอย่างการวิเคราะห์ความแม่นยำในเชิงตัวเลขจากชุดข้อมูล Insurance

Company Benchmark โดยวิเคราะห์จากสัดส่วนของ %L, %T, %Acc_L และ %Acc_T ที่ได้จากอัลกอริทึม MII-ID3

1. หาร้อยละความแม่นยำที่ได้จากโหนดคำตอบปกติ

90.11 (%L) * 90.67 (%Acc_L) / 100 = 81.70

2. หาร้อยละความแม่นยำที่ได้จากโหนดคำตอบที่เกิดจากการละเลยกรณีตัวอย่างส่วนน้อย

9.89 (%T) * 92.39 (%Acc_T) / 100 = 9.14

3. หาความแม่นยำโดยเฉลี่ยของอัลกอริทึม MII-ID3

81.70 + 9.14 = 90.84 ซึ่งใกล้เคียงกับร้อยละความแม่นยำโดยเฉลี่ยในอัลกอริทึม MII-ID3 ที่วัดได้ (90.91)

จากการแสดงขั้นตอนการวิเคราะห์ความแม่นยำของชุดข้อมูล Insurance Company Benchmark จาก %L, %T, %Acc_L และ %Acc_T ก็ชัดเจนแล้วว่าสัดส่วนในการจำแนกกรณีตัวอย่างจากโหนดคำตอบปกติและโหนดคำตอบที่เกิดจากการละเลยกรณีตัวอย่างส่วนน้อยรวมถึงความแม่นยำจากโหนดคำตอบทั้ง 2 ประเภทนั้นส่งผลถึงความแม่นยำโดยเฉลี่ยในอัลกอริทึม MII-ID3 ซึ่งสัดส่วนดังกล่าวจากชุดข้อมูล Insurance Company Benchmark ส่งผลให้ความแม่นยำโดยเฉลี่ยในอัลกอริทึม MII-ID3 มากกว่าความแม่นยำในอัลกอริทึม ID3 แบบดั้งเดิม ซึ่งการที่ความแม่นยำในชุดข้อมูลทั้งหมดในตารางที่ 5.9 เพิ่มขึ้นจากอัลกอริทึม ID3 แบบดั้งเดิมเล็กน้อย เพราะว่ากรณีตัวอย่างในชุดข้อมูลสำหรับทดสอบในแต่ละชุดข้อมูลทั้งหมดในตารางที่ 5.9 มีกรณีตัวอย่างที่ถูกจำแนกที่โหนดคำตอบที่เกิดจากการละเลยกรณีตัวอย่างส่วนน้อยแล้วสามารถจำแนกได้ถูกต้องหลายกรณี หมายความว่ากรณีตัวอย่างที่จำแนกได้ถูกต้องเหล่านี้ตอบเป็นคลาสที่ตรงกับคลาสเสียงส่วนมากที่ถูกระบุในโหนดคำตอบที่เกิดจากการละเลยกรณีตัวอย่างส่วนน้อยนั้น ๆ จึงทำให้ความแม่นยำในการจำแนกในอัลกอริทึม MII-ID3 เพิ่มขึ้นเล็กน้อย เนื่องจากมีกรณีตัวอย่างในชุดข้อมูลสำหรับทดสอบที่ถูกจำแนกที่โหนดคำตอบที่เกิดจากการละเลยกรณีตัวอย่างส่วนน้อยและมีคลาสคำตอบตรงกับคลาสเสียงส่วนมากที่ถูกระบุในโหนดคำตอบนั้น ๆ เป็นจำนวนที่เพียงพอที่จะทำให้ความแม่นยำเพิ่มขึ้นจากเดิมเล็กน้อย

ตารางที่ 5.10 แสดง %L, %T, %Acc_L และ %Acc_T ในชุดข้อมูลที่มีความแม่นยำโดยเฉลี่ยในอัลกอริทึม MII-ID3 น้อยกว่าความแม่นยำโดยเฉลี่ยในอัลกอริทึม ID3 เพียงเล็กน้อย

กลุ่ม	ชุดข้อมูล	ร้อยละของความแม่นยำโดยเฉลี่ย		%L	%T	%Acc_L	%Acc_T
		อัลกอริทึม ID3	อัลกอริทึม MII-ID3				
ขนาดใหญ่	Connect-4	**73.65**	73.59	35.53	64.47	77.98	71.23
	Phishing Websites	**94.91**	93.65	51.48	48.52	98.60	88.33
	Molecular Biology	**88.99**	88.75	84.78	15.22	90.59	78.15
ขนาดกลาง	Soybean	**86.99**	86.21	88.15	11.85	89.52	62.41
	Audiology	**75.64**	75.01	92.74	7.26	75.34	29.82
	SPECT Heart	**80.77**	79.99	60.08	39.92	90.54	63.65
	Mushroom	**100.00**	99.64	95.90	4.10	100.00	66.94

จากตารางที่ 5.10 ชุดข้อมูล Connect-4 เมื่อทดสอบประสิทธิภาพด้วยอัลกอริทึม MII-ID3 แล้ว มีความแม่นยำในการจำแนกที่โหนดคำตอบปกติ (%Acc_L) 77.98 เปอร์เซ็นต์ซึ่งมากกว่าความแม่นยำในอัลกอริทึม ID3 4.34 (77.98-73.65) เปอร์เซ็นต์ ในขณะที่ความแม่นยำในการจำแนกที่โหนดคำตอบที่เกิดจากการละเลยกรณีตัวอย่างส่วนน้อยคือ (%Acc_T) 71.23 เปอร์เซ็นต์ ซึ่งน้อยกว่าความแม่นยำในอัลกอริทึม ID3 2.42 (73.65-71.23) เปอร์เซ็นต์และมีกรณีตัวอย่างที่จำแนกโดยโหนดคำตอบที่เกิดจากการละเลยกรณีตัวอย่างส่วนน้อยถึง (%T) 64.47 เปอร์เซ็นต์ ดังนั้นความแม่นยำในอัลกอริทึม MII-ID3 จึงลดลงจากความแม่นยำในอัลกอริทึม ID3 แบบดั้งเดิมเพียงเล็กน้อย

ในชุดข้อมูล Phishing Websites, Molecular Biology, Soybean, Audiology, SPECT Heart และ Mushroom มีความแม่นยำในการจำแนกโดยโหนดคำตอบที่เกิดจากการละเลยกรณีตัวอย่างส่วนน้อยในอัลกอริทึม MII-ID3 (%Acc_T) น้อยกว่าความแม่นยำในอัลกอริทึม ID3 แบบดั้งเดิมอย่างเห็นให้ชัด แต่อย่างไรก็ตามความแม่นยำในการจำแนกโดยโหนดคำตอบปกตินั้น (%Acc_L) มากกว่าความแม่นยำในอัลกอริทึม ID3 แบบดั้งเดิมประกอบกับชุดข้อมูลเหล่านี้มีจำนวนกรณีตัวอย่างที่ถูกจำแนกโดยโหนดคำตอบปกติ (%L) มากกว่า 50 เปอร์เซ็นต์ในทุกชุดข้อมูล จึงส่งผลให้ความแม่นยำในอัลกอริทึม MII-ID3 โดยรวมลดลงจากอัลกอริทึม ID3 เพียงเล็กน้อย

ตัวอย่างการวิเคราะห์ความแม่นยำในเชิงตัวเลขจากชุดข้อมูล Phishing Websites โดยวิเคราะห์จากสัดส่วนของ %L, %T, %Acc_L และ %Acc_T ที่ได้จากอัลกอริทึม MII-ID3

1. หาร้อยละความแม่นยำที่ได้จากโหนดคำตอบปกติ

51.48 (%L) * 98.60 (%Acc_L) / 100 = 50.76

2. หาร้อยละความแม่นยำที่ได้จากโหนดคำตอบที่เกิดจากการละเลยกรณีตัวอย่างส่วนน้อย

48.52 (%T) * 88.33 (%Acc_T) / 100 = 42.86

3. หาความแม่นยำโดยเฉลี่ยของอัลกอริทึม MII-ID3

50.76 + 42.86 = 93.62 ซึ่งใกล้เคียงกับร้อยละความแม่นยำโดยเฉลี่ยในอัลกอริทึม MII-ID3 ที่วัดได้ (93.65) ในชุดข้อมูล Phishing Websites

ดังนั้นการที่ความแม่นยำของชุดข้อมูลเหล่านี้ลดลงเพียงเล็กน้อย เพราะเนื่องจากมีกรณีตัวอย่างในชุดข้อมูลสำหรับทดสอบที่ถูกจำแนกที่โหนดคำตอบที่เกิดจากการละเลยกรณีตัวอย่างส่วนน้อยและมีคลาสคำตอบตรงกับคลาสเสียงส่วนมากที่ถูกระบุในโหนดคำตอบนั้น ๆ ไม่เพียงพอที่จะทำให้ความแม่นยำเพิ่มขึ้นหรือเท่ากับอัลกอริทึม ID3 แบบดั้งเดิม กล่าวอีกอย่างได้ว่าชุดข้อมูลทั้งหมดในตารางที่ 5.10 มีกรณีตัวอย่างในชุดข้อมูลสำหรับทดสอบที่ถูกจำแนกที่โหนดคำตอบที่เกิดจากการละเลยกรณีตัวอย่างส่วนน้อยและมีคลาสคำตอบที่ไม่ตรงกับคลาสเสียงส่วนมากที่ถูกระบุในโหนดคำตอบนั้น ๆ ก็คือกรณีตัวอย่างบางส่วนมีคลาสคำตอบที่ไปตรงกับคลาสคำตอบที่เป็นกรณีตัวอย่างส่วนน้อยที่ถูกละเลยไปบ้างแต่ยังไม่มากเท่าไหร่ จึงทำให้ความแม่นยำในชุดข้อมูลเหล่านี้ลดลงจากอัลกอริทึม MII-ID3 เพียงเล็กน้อย

ตารางที่ 5.11 แสดง %L, %T, %Acc_L และ %Acc_T ในชุดข้อมูลที่มีความแม่นยำโดยเฉลี่ยในอัลกอริทึม MII-ID3 น้อยกว่าความแม่นยำโดยเฉลี่ยในอัลกอริทึม ID3 อย่างเห็นได้ชัด

กลุ่ม	ชุดข้อมูล	ร้อยละของความแม่นยำโดยเฉลี่ย		%L	%T	%Acc_L	%Acc_T
		อัลกอริทึม ID3	อัลกอริทึม MII-ID3				
ขนาดกลาง	Car Evaluation	**91.58**	83.42	68.16	31.84	98.18	53.59
	Chess	**99.29**	94.82	73.00	27.00	99.88	81.00
	Nursery	**96.86**	89.16	57.42	42.58	99.70	74.67
	Firm-Teacher_Clave-Direction_Classification	**71.71**	67.78	6.75	93.25	86.50	66.18
ขนาดเล็ก	Hayes-Roth	**79.01**	73.69	64.93	35.07	86.26	45.58
	MONK's Problems	**98.19**	86.96	67.63	32.37	99.08	58.78
	Tic-Tac-Toe Endgame	**86.01**	77.68	47.37	52.63	92.62	63.57

จากตารางที่ 5.11 พบว่าชุดข้อมูลทุกตัวในตารางมีความแม่นยำในอัลกอริทึม ID3 แบบดั้งเดิมหรือความแม่นยำจากกรณีตัวอย่างที่ถูกจำแนกด้วยโหนดคำตอบปกติ 100 เปอร์เซ็นต์มีมากกว่า 70 เปอร์เซ็นต์ในชุดข้อมูลทุกตัว นั้นหมายความว่าแต่เดิมชุดข้อมูลทุกตัวในตารางที่ 5.11 มีความแม่นยำที่สูงอยู่แล้วทุกตัว เมื่อนำชุดข้อมูลทุกตัวไปทดสอบประสิทธิภาพด้วยอัลกอริทึม MII-ID3

พบว่าในแต่ละชุดข้อมูลมีกรณีตัวอย่างที่ถูกแบ่งไปจำแนกในโหนดคำตอบที่เกิดจากการละเลยกรณีตัวอย่างส่วนน้อยในอัลกอริทึม MII-ID3 (%T) นั้นมีจำนวนมากกว่า 27 เปอร์เซ็นต์ในชุดข้อมูลทุกตัว และในชุดข้อมูล Firm-Teacher_Clave-Direction_Classification มีกรณีตัวอย่างที่จำแนกโดยโหนดคำตอบที่เกิดจากการละเลยกรณีตัวอย่างส่วนน้อยถึง 93.25 เปอร์เซ็นต์ และความแม่นยำในโหนดคำตอบที่เกิดจากการละเลยกรณีตัวอย่างส่วนน้อยนั้นมีค่าต่ำอย่างเห็นได้ชัดในทุก ๆ ชุดข้อมูล ดังนั้นจึงส่งผลให้ความแม่นยำโดยรวมในอัลกอริทึม MII-ID3 ของชุดข้อมูล Car Evaluation, Chess, Nursery, Firm-Teacher_Clave-Direction_Classification, Hayes-Roth, MONK's Problems และ Tic-Tac-Toe Endgame ลดลงน้อยกว่าความแม่นยำในอัลกอริทึม ID3 แบบดั้งเดิมอย่างเห็นได้ชัด ซึ่งไม่คุ้มค่าที่จะใช้อัลกอริทึม MII-ID3

พิจารณาตัวอย่างการวิเคราะห์ความแม่นยำในเชิงตัวเลขจากชุดข้อมูล Firm-Teacher_Clave-Direction_Classification โดยวิเคราะห์จากสัดส่วนของ %L, %T, %Acc_L และ %Acc_T ที่ได้จากอัลกอริทึม MII-ID3

1. หาร้อยละความแม่นยำที่ได้จากโหนดคำตอบปกติ

6.75 (%L) * 86.50 (%Acc_L) / 100 = 5.84

2. หาร้อยละความแม่นยำที่ได้จากโหนดคำตอบที่เกิดจากการละเลยกรณีตัวอย่างส่วนน้อย

93.25 (%T) * 66.18 (%Acc_T) / 100 = 61.71

3. หาความแม่นยำโดยเฉลี่ยของอัลกอริทึม MII-ID3

5.84 + 61.71 = 67.55 ซึ่งใกล้เคียงกับร้อยละความแม่นยำโดยเฉลี่ยในอัลกอริทึม MII-ID3 ที่วัดได้ (67.78) ในชุดข้อมูล Firm-Teacher_Clave-Direction_Classification

ดังนั้นการที่ชุดข้อมูลทั้งหมดในตารางที่ 5.11 มีความแม่นยำในอัลกอริทึม MII-ID3 ที่ลดลงจากอัลกอริทึม ID3 แบบดั้งเดิมอย่างชัดเจน เนื่องจากมีกรณีตัวอย่างในชุดข้อมูลสำหรับทดสอบที่ถูกจำแนกที่โหนดคำตอบที่เกิดจากการละเลยกรณีตัวอย่างส่วนน้อยและมีคลาสคำตอบที่ตรงกับคลาสเสียงส่วนน้อยที่อยู่ในกรณีตัวอย่างส่วนน้อยที่ถูกละเลยไปเป็นจำนวนที่มากพอที่จะส่งผลให้ความแม่นยำลดลงอย่างมาก เพราะจำแนกประเภทของกรณีตัวอย่างในชุดข้อมูลสำหรับทดสอบผิดเป็นจำนวนมากนั้นเอง

ตั้งแต่ตารางที่ 5.9-5.11 เป็นการวัดโดยใช้เกณฑ์ในการวัดที่เกี่ยวข้องกับความแม่นยำในการจำแนกเพื่ออธิบายและแสดงเหตุผลการเพิ่มขึ้นหรือลดลงของแม่นยำในการจำแนกได้อย่างชัดเจน ถัดมาให้พิจารณาร้อยละในการลดลงของความลึกสูงสุดของอัลกอริทึม MII-ID3 เมื่อเปรียบเทียบกับอัลกอริทึม ID3 แบบดั้งเดิม จากร้อยละในการลดลงของความลึกสูงสุดในตารางที่ 5.6-5.8 จะเห็นว่าบางชุดข้อมูลนั้นสามารถลดความลึกสูงสุดได้มากกว่า 60 เปอร์เซ็นต์ แต่ก็มีบางชุดข้อมูลที่ลดความลึกสูงสุดได้น้อยเนื่องจากในแต่ละชุดข้อมูลมีลักษณะบางอย่างที่ต่างกันได้แก่

1. จำนวนแอตทริบิวต์ทั้งหมด

2. จำนวนความลึกสูงสุดโดยเฉลี่ยเมื่อทดสอบกับอัลกอริทึม ID3

3. ร้อยละของจำนวนโหนดที่ถูกสร้างจากกรณีแอตทริบิวต์ถูกใช้จนหมดเมื่อทดสอบกับอัลกอริทึม ID3 โดยปกติการสร้างโหนดคำตอบของอัลกอริทึม ID3 แบบดั้งเดิมมีทั้งหมด 3 กรณีดังที่ได้อธิบายไว้ในบทที่ 2 หนึ่งในสามกรณีการสร้างโหนดคำตอบคือเมื่อเกิดกรณีที่แอตทริบิวต์ถูกใช้จนหมดในเชิงลึกแล้ว แต่ยังไม่สามารถจำแนกประเภทกรณีตัวอย่างปัจจุบันได้ดังนั้นอัลกอริทึม ID3 จะสร้างโหนดคำตอบโดยการจำแนกประเภทโดยระบุคลาสเป็นเสียงส่วนมากของกรณีตัวอย่างปัจจุบัน คือถ้ากรณีตัวอย่างปัจจุบันอยู่ในคลาสใดมากที่สุดจะระบุเป็นคลาสนั้น

ตารางที่ 5.12 แสดงร้อยละในการลดลงของความลึกสูงสุดในอัลกอริทึม MII-ID3 จากอัลกอริทึม ID3, จำนวนแอตทริบิวต์ทั้งหมด, ความลึกสูงสุดโดยเฉลี่ยเมื่อทดสอบกับอัลกอริทึม ID3 และร้อยละของจำนวนโหนดที่ถูกสร้างจากกรณีแอตทริบิวต์ถูกใช้จนหมดเมื่อทดสอบกับอัลกอริทึม ID3 ในทุกชุดข้อมูล

ชุดข้อมูล	ร้อยละในการลดลงของความลึกสูงสุดในอัลกอริทึม MII-ID3 จากอัลกอริทึม ID3	จำนวนแอตทริบิวต์ทั้งหมด	ความลึกสูงสุดโดยเฉลี่ยของอัลกอริทึม ID3	ร้อยละของจำนวนโหนดที่ถูกสร้างจากกรณีแอตทริบิวต์ถูกใช้จนหมด
Connect-4	30.81	42	22.73	0.00
Phishing Websites	63.54	30	30.00	3.17
Insurance Company Benchmark	71.75	85	85.00	0.18
Molecular Biology	85.35	60	60.00	0.18
Soybean	76.93	35	35.00	0.74
Audiology	3.35	69	6.00	0.00
Balance Scale	5.40	4	4.00	0.00
Car Evaluation	23.60	6	6.00	0.00
Chess	43.03	36	13.00	0.00
Congressional Voting Records	72.83	16	16.00	4.26
Hayes-Roth	19.95	4	4.00	8.82
MONK's Problems	2.50	6	4.00	0.00
Mushroom	30.90	22	4.00	0.00
Nursery	32.88	8	8.00	0.00
SPECT Heart	66.89	22	22.00	5.06
Tic-Tac-Toe Endgame	26.36	9	7.00	0.00
Firm-Teacher_Clave-Direction_Classification	28.46	16	15.34	0.00

จากตารางที่ 5.12 แสดงให้เห็นว่าชุดข้อมูล Phishing Websites, Congressional Voting Records, Soybean, SPECT Heart, Insurance Company Benchmark และ Molecular Biology มีการลดลงของความลึกสูงสุดมากกว่า 60 เปอร์เซ็นต์ เพราะชุดข้อมูลเหล่านี้มีจำนวนแอตทริบิวต์ที่มาก และความลึกสูงสุดของชุดข้อมูลเหล่านี้เมื่อทดสอบกับอัลกอริทึม ID3 พบว่าชุดข้อมูลเหล่านี้มีความลึกสูงสุดเท่ากับจำนวนแอตทริบิวต์ทั้งหมดในแต่ละชุดข้อมูล และยังมีโหนดคำตอบที่สร้างจากกรณีที่แอตทริบิวต์ถูกใช้จนหมดในอัลกอริทึม ID3 อีกด้วย หมายความว่าชุดข้อมูลเหล่านี้พยายามจำแนกประเภทกรณีตัวอย่างลงไปเรื่อย ๆ จนกระทั่งใช้แอตทริบิวต์จนหมดก็ยังไม่สามารถจำแนกกรณีตัวอย่างได้ทั้งหมด ร้อยละของจำนวนโหนดที่ถูกสร้างจากกรณีแอตทริบิวต์ถูกใช้จนหมดในอัลกอริทึม ID3 ของชุดข้อมูลเหล่านี้จึงไม่เท่ากับ 0 ชุดข้อมูลเหล่านี้จึงมีความลึกสูงสุดที่มากที่สุดเท่าที่เป็นไปได้และการใช้แอตทริบิวต์จนหมดแล้วยังไม่สามารถจำแนกกรณีตัวอย่างทั้งหมดได้ หมายความว่าชุดข้อมูลเหล่านี้มีโหนดที่มีกรณีตัวอย่างส่วนน้อยอยู่มากซึ่งจะตรงกับเป้าหมายของอัลกอริทึม MII-ID3 ที่จะละเลยกรณีตัวอย่างส่วนน้อยนี้เพื่อลดความเข้มงวดในกฎการตัดสินใจของอัลกอริทึม ID3 เมื่อชุดข้อมูลเหล่านี้มีกรณีตัวอย่างส่วนน้อยมากและอีกทั้งยังมีความลึกสูงสุดที่มากเมื่อถูกละเลยกรณีตัวอย่างส่วนน้อยในอัลกอริทึม MII-ID3 ความลึกสูงสุดของชุดข้อมูลเหล่านี้จึงลดลงจากอัลกอริทึม ID3 ได้มากกว่า 60 เปอร์เซ็นต์ พิจารณาชุดข้อมูล MONK’s Problems เป็นชุดข้อมูลที่มีร้อยละการลดลงของความลึกสูงสุดในอัลกอริทึม MII-ID3 จากอัลกอริทึม ID3 น้อยที่สุดเพียง 2.50 เปอร์เซ็นต์ เพราะว่าชุดข้อมูลนี้ปกติมีแอตทริบิวต์ทั้งหมดที่น้อยมากมีเพียง 6 แอตทริบิวต์เท่านั้น หมายความว่าความลึกสูงสุดในอัลกอริทึม ID3 ที่เป็นไปได้มากที่สุดของชุดข้อมูลนี้คือ 6 เท่านั้น จากตารางที่ 5.12 พบว่าชุดข้อมูล MONK’s Problems มีความลึกสูงสุดแค่ 4 และชุดข้อมูลที่มีขนาดเล็กนั้นแทบจะไม่มีกรณีตัวอย่างส่วนน้อยหรือไม่มีเลย ดังนั้นอัลกอริทึม MII-ID3 จึงลดความลึกสูงสุดของชุดข้อมูล MONK’s Problem ได้น้อยมาก

5.5.2 จำนวนกฎการตัดสินใจโดยเฉลี่ย

เพื่อความชัดเจนว่าอัลกอริทึม MII-ID3 สามารถลดความเข้มงวดในการสร้างกฎการตัดสินใจในอัลกอริทึม ID3 แบบดั้งเดิม ให้พิจารณาผลการทดลองดังตารางต่อไปนี้ จำนวนกฎการตัดสินใจของต้นไม้ตัดสินใจผลลัพธ์นั้นจะเท่ากับจำนวนโหนดคำตอบทั้งหมดดังที่ได้เคยกล่าวไว้ ซึ่งถ้าสามารถลดจำนวนกฎการตัดสินใจในต้นไม้ตัดสินใจลงได้จะส่งผลให้ต้นไม้ตัดสินใจมีขนาดที่เล็กลงและเป็นการช่วยประหยัดพื้นที่ในการเก็บในหน่วยความจำ

ตารางที่ 5.13 แสดงจำนวนกฎการตัดสินใจโดยเฉลี่ยในการทดลอง 1000 รอบระหว่าง ID3 และ MII-ID3 ในชุดข้อมูลที่มีความแม่นยำโดยเฉลี่ยในอัลกอริทึม MII-ID3 มากกว่าความแม่นยำโดยเฉลี่ยในอัลกอริทึม ID3 และชุดข้อมูลที่มีความแม่นยำโดยเฉลี่ยในอัลกอริทึม MII-ID3 น้อยกว่าความแม่นยำโดยเฉลี่ยในอัลกอริทึม ID3 เพียงเล็กน้อย

กลุ่ม	ชุดข้อมูล	จำนวนกฎการตัดสินใจโดยเฉลี่ย		ร้อยละในการลดลงของจำนวนกฎการตัดสินใจ
		อัลกอริทึม ID3	อัลกอริทึม MII-ID3	
ขนาดใหญ่	Connect-4	14743.47	**4646.37**	68.49
	Phishing Websites	1009.06	**100.77**	90.01
	Insurance Company Benchmark	22784.01	**3421.84**	84.98
	Molecular Biology	568.53	**288.36**	49.28
ขนาดกลาง	Soybean	134.55	**75.63**	43.79
	Audiology	42.49	**40.89**	3.76
	Congressional Voting Records	23.49	**7.42**	68.41
	Mushroom	34.32	**22.80**	33.56
	SPECT Heart	79.00	**15.48**	80.40
ขนาดเล็ก	Balance Scale	257.00	**97.82**	61.94

จากผลการทดลองในตารางที่ 5.13 แสดงให้เห็นว่าอัลกอริทึม MII-ID3 สามารถลดจำนวนกฎการตัดสินใจจากอัลกอริทึม ID3 แบบดั้งเดิมได้มากกว่า 30 เปอร์เซ็นต์ในชุดข้อมูลส่วนใหญ่ มีเพียงชุดข้อมูล Audiology เท่านั้นที่สามารถลดจำนวนกฎการตัดสินใจจากอัลกอริทึม ID3 แบบดั้งเดิมได้เพียง 3.76 เปอร์เซ็นต์ซึ่งถือว่าคุ้มค่าอยู่ดี เพราะชุดข้อมูลทั้งหมดในตารางที่ 5.13 เป็นชุดข้อมูลที่มีความแม่นยำโดยเฉลี่ยมากกว่าหรือน้อยกว่าอัลกอริทึม ID3 เพียงเล็กน้อย เมื่อความแม่นยำโดยเฉลี่ยในอัลกอริทึม MII-ID3 ของชุดข้อมูลทั้งหมดอยู่ในระดับที่น่าพอใจ อีกทั้งความลึกสูงสุดโดยเฉลี่ย (ตารางที่ 5.6-5.8) และจำนวนกฎการตัดสินใจโดยเฉลี่ยยังลดลงอีกด้วย ดังนั้นชุดข้อมูลทั้งหมดในตารางที่ 5.13 จึงคุ้มค่าอย่างมากที่จะใช้อัลกอริทึม MII-ID3 หรือทำการละเลยกรณีตัวอย่างส่วนน้อย

ตารางที่ 5.14 แสดงจำนวนกฎการตัดสินใจโดยเฉลี่ยในการทดลอง 1000 รอบระหว่าง ID3 และ MII-ID3 ในชุดข้อมูลที่มีความแม่นยำโดยเฉลี่ยในอัลกอริทึม MII-ID3 น้อยกว่าความแม่นยำโดยเฉลี่ยในอัลกอริทึม ID3 อย่างชัดเจน

กลุ่ม	ชุดข้อมูล	จำนวนกฎการตัดสินใจโดยเฉลี่ย		ร้อยละในการลดลงของจำนวนกฎการตัดสินใจ
		อัลกอริทึม ID3	อัลกอริทึม MII-ID3	
ขนาดกลาง	Car Evaluation	221.49	**50.34**	77.27
	Chess	37.00	**14.17**	61.69
	Nursery	636.48	**71.88**	88.71
	Firm-Teacher_Clave-Direction_Classification	1580.76	**120.02**	92.41
ขนาดเล็ก	Hayes-Roth	34.00	**23.47**	30.96
	MONK's Problems	55.00	**35.84**	34.84
	Tic-Tac-Toe Endgame	135.00	**48.86**	63.81

จากตารางที่ 5.14 แสดงให้เห็นว่าอัลกอริทึม MII-ID3 สามารถลดจำนวนกฎการตัดสินใจได้มากกว่า 30 เปอร์เซ็นต์ในชุดข้อมูลทั้งหมดในตาราง แต่อย่างไรก็ตามชุดข้อมูลทั้งหมดในตารางที่ 5.14 นั้นมีความแม่นยำโดยเฉลี่ยในอัลกอริทึม MII-ID3 น้อยกว่าความแม่นยำโดยเฉลี่ยในอัลกอริทึม ID3 อย่างชัดเจน ดังนั้นชุดข้อมูลเหล่านี้จึงไม่คุ้มค่าที่จะใช้อัลกอริทึม MII-ID3 เพราะความแม่นยำลดลงอย่างมากนั้นเอง

5.5.3 จำนวนโหนดทั้งหมดในต้นไม้ตัดสินใจผลลัพธ์โดยเฉลี่ย

จำนวนโหนดทั้งหมดในต้นไม้ตัดสินใจคือผลรวมของจำนวนโหนดการตัดสินใจ, โหนดการจำประเภทและโหนดคำตอบ หรือก็คือโหนดทั้งหมดในต้นไม้ตัดสินใจผลลัพธ์ ซึ่งจำนวนโหนดทั้งหมดจะส่งผลต่อขนาดของต้นไม้ตัดสินใจ

ตารางที่ 5.15 แสดงจำนวนโหนดทั้งหมดโดยเฉลี่ยในการทดลอง 1000 รอบระหว่าง ID3 และ MII-ID3 ในชุดข้อมูลที่มีความแม่นยำโดยเฉลี่ยในอัลกอริทึม MII-ID3 มากกว่าความแม่นยำโดยเฉลี่ยในอัลกอริทึม ID3 และชุดข้อมูลที่มีความแม่นยำโดยเฉลี่ยในอัลกอริทึม MII-ID3 น้อยกว่าความแม่นยำโดยเฉลี่ยในอัลกอริทึม ID3 เพียงเล็กน้อย

กลุ่ม	ชุดข้อมูล	จำนวนโหนดทั้งหมดโดยเฉลี่ย		ร้อยละในการลดลงของจำนวนโหนดทั้งหมด
		อัลกอริทึม ID3	อัลกอริทึม MII-ID3	
ขนาดใหญ่	Connect-4	22114.71	**6969.06**	68.49
	Phishing Websites	1869.85	**169.28**	90.95
	Insurance Company Benchmark	26385.54	**3908.88**	85.19
	Molecular Biology	716.56	**362.11**	49.46
ขนาดกลาง	Soybean	202.55	**109.03**	46.17
	Audiology	64.49	**61.67**	4.37
	Congressional Voting Records	45.98	**13.84**	69.90
	Mushroom	39.32	**25.80**	34.37
	SPECT Heart	157.00	**29.97**	80.91
ขนาดเล็ก	Balance Scale	321.00	**122.02**	61.99

จากผลการทดลองในตารางที่ 5.15 แสดงให้เห็นว่าอัลกอริทึม MII-ID3 สามารถลดจำนวนโหนดทั้งหมดในต้นไม้ตัดสินใจผลลัพธ์จากอัลกอริทึม ID3 แบบดั้งเดิมได้มากกว่า 30 เปอร์เซ็นต์ มีเพียงชุดข้อมูล Audiology เท่านั้นที่สามารถลดจำนวนโหนดทั้งหมดในต้นไม้ตัดสินใจผลลัพธ์จากอัลกอริทึม ID3 แบบดั้งเดิมได้เพียง 4.37 เปอร์เซ็นต์ ชุดข้อมูลทุกตัวในตารางที่ 5.15 นั้นถือว่าคุ้มค่าต่อการใช้อัลกอริทึม MII-ID3 เพราะเป็นชุดข้อมูลที่มีความแม่นยำโดยเฉลี่ยมากกว่าหรือน้อยกว่าอัลกอริทึม ID3 เพียงเล็กน้อย อีกทั้งจำนวนความลึกสูงสุดโดยเฉลี่ย (ตารางที่ 5.6-5.8), จำนวนกฎการตัดสินใจ (ตารางที่ 5.13-5.14) และจำนวนโหนดทั้งหมดโดยเฉลี่ยของต้นไม้ตัดสินใจผลลัพธ์ยังลดลงอีกด้วย

ตารางที่ 5.16 แสดงจำนวนโหนดทั้งหมดโดยเฉลี่ยในการทดลอง 1000 รอบระหว่าง ID3 และ MII-ID3 ในชุดข้อมูลที่มีความแม่นยำโดยเฉลี่ยในอัลกอริทึม MII-ID3 น้อยกว่าความแม่นยำโดยเฉลี่ยในอัลกอริทึม ID3 อย่างชัดเจน

กลุ่ม	ชุดข้อมูล	จำนวนกฎการตัดสินใจโดยเฉลี่ย		ร้อยละในการลดลงของจำนวนกฎการตัดสินใจ
		อัลกอริทึม ID3	อัลกอริทึม MII-ID3	
ขนาดกลาง	Car Evaluation	303.49	68.86	77.31
	Chess	71.00	26.71	62.38
	Nursery	898.48	102.59	88.58
	Firm-Teacher_Clave-Direction_Classification	3160.52	239.04	92.44
ขนาดเล็ก	Hayes-Roth	46.00	31.05	32.49
	MONK's Problems	82.00	53.25	35.06
	Tic-Tac-Toe Endgame	202.00	72.79	63.97

จากตารางที่ 5.16 แสดงให้เห็นว่าอัลกอริทึม MII-ID3 สามารถลดจำนวนโหนดทั้งหมดในต้นไม้ตัดสินใจผลลัพธ์ได้มากกว่า 30 เปอร์เซ็นต์ในชุดข้อมูลทั้งหมดในตาราง และแม้ว่าชุดข้อมูลทั้งหมดในตารางนั้นอัลกอริทึม MII-ID3 จะสามารถลดได้ทั้งจำนวนความลึกสูงสุด (ตารางที่ 5.6-5.8), จำนวนกฎการตัดสินใจ (ตารางที่ 5.13-5.14) และจำนวนโหนดทั้งหมดของต้นไม้ตัดสินใจผลลัพธ์ อย่างไรก็ตามถือว่าไม่คุ้มค่าเพราะว่าความแม่นยำในการจำแนกโดยเฉลี่ยของชุดข้อมูลเหล่านี้ลดลงจากอัลกอริทึม ID3 แบบดั้งเดิมอย่างเห็นได้ชัด

5.5.4 เวลาในการฝึกฝนและเวลาในการทดสอบต้นไม้ตัดสินใจผลลัพธ์โดยเฉลี่ย

เมื่อจำนวนความลึกสูงสุดและจำนวนกฎการตัดสินใจของต้นไม้ตัดสินใจผลลัพธ์ที่ได้จากอัลกอริทึม MII-ID3 ลดลงจากอัลกอริทึม ID3 อย่างเห็นได้ชัดในหลาย ๆ ชุดข้อมูลจากผลการทดลองก่อนหน้า จึงน่าสนใจอย่างยิ่งที่จะทำการวัดเวลาในการฝึกฝน (เวลาในการสร้างต้นไม้ตัดสินใจ) และเวลาในการทดสอบต้นไม้ตัดสินใจผลลัพธ์เพื่อจะดูว่าความลึกสูงสุดและจำนวนกฎการตัดสินใจที่ลดลงอย่างเห็นได้ชัดนั้นจะส่งผลต่อเวลาในการฝึกฝนและเวลาในการทดสอบต้นไม้ตัดสินใจผลลัพธ์มากน้อยเพียงใด

ตารางที่ 5.17 แสดงเวลาในการฝึกฝนและเวลาในการทดสอบต้นไม้ตัดสินใจผลลัพธ์โดยเฉลี่ยระหว่าง ID3 และ MII-ID3

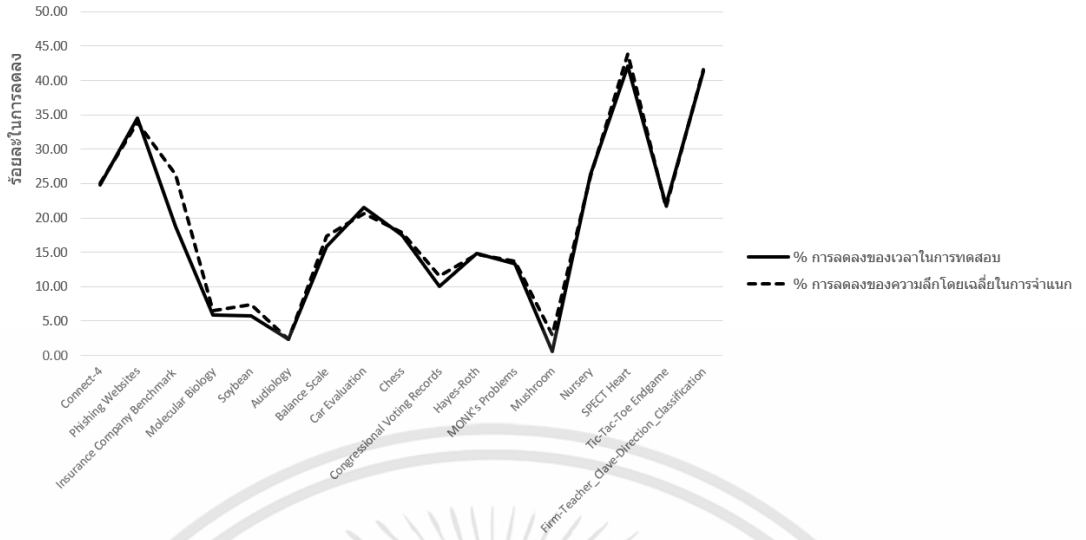
การเปรียบเทียบความแม่นยำโดยเฉลี่ยระหว่าง MII-ID3 และ ID3	ชุดข้อมูล	เวลาในการฝึกฝนโดยเฉลี่ย (ms)		ร้อยละในการลดลงของเวลาในการฝึกฝน	เวลาในการทดสอบโดยเฉลี่ย (ms)		ร้อยละในการลดลงของเวลาในการทดสอบ
		อัลกอริทึม ID3	อัลกอริทึม MII-ID3		อัลกอริทึม ID3	อัลกอริทึม MII-ID3	
MII-ID3 มากกว่า ID3	Insurance Company Benchmark	417.209	**234.966**	43.68	0.015912	**0.012937**	18.69
	Congressional Voting Records	0.805	**0.630**	21.69	0.000461	**0.000415**	10.07
	Balance Scale	0.787	**0.572**	27.27	0.000778	**0.000655**	15.88
MII-ID3 น้อยกว่า ID3 เพียงเล็กน้อย	Connect-4	1441.300	**1040.922**	27.78	0.254854	**0.191490**	24.86
	Phishing Websites	64.018	**40.600**	36.58	0.024674	**0.016163**	34.49
	Molecular Biology	55.042	**48.923**	11.12	0.005304	**0.004993**	5.87
	Soybean	14.872	**12.835**	13.70	0.001007	**0.000949**	5.77
	Audiology	10.363	**10.146**	2.10	0.000378	**0.000369**	2.32
	SPECT Heart	1.673	**0.950**	43.20	0.000696	**0.000403**	42.05
	Mushroom	14.403	**14.211**	1.33	0.005154	**0.005124**	0.57
MII-ID3 น้อยกว่า ID3 อย่างเห็นได้ชัด	Car Evaluation	2.027	**1.454**	28.28	0.002030	**0.001594**	21.48
	Chess	15.839	**12.913**	18.47	0.005573	**0.004598**	17.49
	Nursery	20.193	**15.182**	24.81	0.017814	**0.013118**	26.36
	Firm-Teacher_Clave-Direction_Classification	83.916	**46.897**	44.12	0.041603	**0.024379**	41.40
	Hayes-Roth	0.166	**0.136**	18.06	0.000182	**0.000155**	14.79
	MONK's Problems	0.531	**0.452**	14.83	0.000677	**0.000586**	13.35
	Tic-Tac-Toe Endgame	1.889	**1.373**	27.33	0.001632	**0.001276**	21.84

จากผลการทดลองในตารางที่ 5.17 แสดงให้เห็นว่าเวลาในการฝึกฝนและเวลาในการทดสอบต้นไม้ตัดสินใจในอัลกอริทึม MII-ID3 น้อยกว่าอัลกอริทึม ID3 ในทุก ๆ ชุดข้อมูล อย่างไรก็ตามชุดข้อมูลที่คุ้มค่าในการใช้อัลกอริทึม MII-ID3 คือชุดข้อมูลที่มีความแม่นยำโดยเฉลี่ยในอัลกอริทึม MII-ID3 มากกว่าความแม่นยำโดยเฉลี่ยในอัลกอริทึม ID3 และชุดข้อมูลที่มีความแม่นยำโดยเฉลี่ยในอัลกอริทึม MII-ID3 น้อยกว่าความแม่นยำโดยเฉลี่ยในอัลกอริทึม ID3 เพียงเล็กน้อย และเนื่องจากกรณีตัวอย่างส่วนมากในชุดข้อมูลสำหรับทดสอบอาจจะถูกจำแนกที่ระดับความลึกที่น้อยกว่าความลึกสูงสุดของต้นไม้ตัดสินใจ ดังนั้นจึงทำการวัดความลึกโดยเฉลี่ยในการจำแนกกรณีตัวอย่างหรือค่าเฉลี่ยของความลึกที่กรณีตัวอย่างถูกจำแนกในต้นไม้ตัดสินใจ

ตารางที่ 5.18 แสดงความลึกโดยเฉลี่ยในการจำแนกกรณีตัวอย่างระหว่าง ID3 และ MII-ID3

ชุดข้อมูล	ความลึกโดยเฉลี่ยในการจำแนก		ร้อยละในการลดลงของความลึกโดยเฉลี่ยในการจำแนก
	อัลกอริทึม ID3	อัลกอริทึม MII-ID3	
Connect-4	9.36	**7.02**	25.00
Phishing Websites	5.46	**3.61**	33.86
Insurance Company Benchmark	4.44	**3.27**	26.35
Molecular Biology	4.13	**3.86**	6.51
Soybean	3.70	**3.43**	7.44
Audiology	4.04	**3.94**	2.34
Balance Scale	3.14	**2.59**	17.32
Car Evaluation	2.89	**2.29**	20.63
Chess	4.35	**3.57**	17.88
Congressional Voting Records	2.70	**2.39**	11.52
Hayes-Roth	2.78	**2.37**	14.76
MONK's Problems	3.01	**2.60**	13.70
Mushroom	1.52	**1.48**	3.03
Nursery	3.39	**2.50**	26.06
SPECT Heart	6.73	**3.78**	43.81
Tic-Tac-Toe Endgame	4.21	**3.30**	21.58
Firm-Teacher_Clave-Direction_Classification	9.58	**5.59**	41.65

จากตารางที่ 5.18 แสดงให้เห็นว่าร้อยละในการลดลงของความลึกโดยเฉลี่ยในการจำแนกกรณีตัวอย่างเป็นไปในทางเดียวกันกับร้อยละในการลดลงของเวลาในการทดสอบ (ตารางที่ 5.17) เพราะว่าความลึกโดยเฉลี่ยในการจำแนกนั้นจะส่งผลต่อเวลาในการทดสอบโดยตรง ถ้ากรณีตัวอย่างในชุดข้อมูลสำหรับทดสอบส่วนมากถูกจำแนกที่ระดับความลึกน้อย ๆ เวลาในการทดสอบก็จะน้อย ในทางกลับกันถ้ากรณีตัวอย่างส่วนมากถูกจำแนกที่ระดับความลึกสูง เวลาในการทดสอบก็จะมาก ดังนั้นเมื่ออัลกอริทึม MII-ID3 สามารถลดเวลาในการทดสอบจากอัลกอริทึม ID3 เป็นร้อยละเท่าใด การลดลงในระดับความลึกเฉลี่ยในการจำแนกในอัลกอริทึม MII-ID3 จะลดลงเป็นร้อยละที่สัมพันธ์กันกับการลดลงของเวลาในการทดสอบดังในรูปที่ 5.5



รูปที่ 5.5 แสดงกราฟความสัมพันธ์ระหว่างร้อยละการลดลงของเวลาในการทดสอบและร้อยละการลดลงของความลึกโดยเฉลี่ยในการจำแนก

จากรูปที่ 5.5 เส้นทึบแสดงร้อยละการลดลงของเวลาในการทดสอบ เส้นประแสดงร้อยละการลดลงของความลึกโดยเฉลี่ยในการจำแนก จะเห็นว่าเส้นทั้งสองของชุดข้อมูลส่วนมากค่อนข้างจะเป็นไปในทิศทางเดียวกัน

สรุปผลการทดลองระหว่างอัลกอริทึม ID3 และอัลกอริทึม MII-ID3 พบว่าอัลกอริทึม MII-ID3 สามารถลดจำนวนกฎการตัดสินใจ, ลดจำนวนความลึกสูงสุด, ลดจำนวนโหนดทั้งหมด, ลดเวลาในการฝึกฝน, ลดเวลาในการทดสอบ และลดความลึกโดยเฉลี่ยในการจำแนกจากอัลกอริทึม ID3 ได้อย่างชัดเจน ในขณะที่ความแม่นยำเพิ่มขึ้นจากอัลกอริทึม ID3 และลดลงเพียงเล็กน้อยในชุดข้อมูล 10 ชุดข้อมูล จากชุดข้อมูลที่ใช้ในการทดลองทั้งหมด 17 ชุดข้อมูล ใน 10 ชุดข้อมูลที่ใช้อัลกอริทึม MII-ID3 ได้อย่างคุ้มค่านี้ประกอบไปด้วย ชุดข้อมูลขนาดใหญ่ทั้งหมด 4 ชุดข้อมูล ชุดข้อมูลขนาดกลาง 5 ชุดข้อมูล จากชุดข้อมูลขนาดกลางทั้งหมด 9 ชุดข้อมูล และชุดข้อมูลขนาดเล็ก 1 ชุดข้อมูล จากชุดข้อมูลขนาดเล็กทั้งหมด 4 ชุดข้อมูล ซึ่งสรุปได้ว่าอัลกอริทึม MII-ID3 นั้นเหมาะสมและคุ้มค่าในการใช้ลดความเข้มงวดในการสร้างกฎการตัดสินใจและลดจำนวนความลึกสูงสุดในอัลกอริทึม ID3 ในขณะที่ความแม่นยำลดลงเพียงเล็กน้อยในชุดข้อมูลขนาดใหญ่ที่มีทั้งจำนวนแอตทริบิวต์และกรณีตัวอย่างเป็นจำนวนมาก ทั้งยังเหมาะสมกับชุดข้อมูลขนาดกลางบางชุดข้อมูล และไม่เหมาะสมกับชุดข้อมูลขนาดเล็กส่วนมาก

5.6 ผลการทดลองระหว่างอัลกอริทึม MII-ID3 และ ID3-A*

ดังที่ได้กล่าวไว้ว่าอัลกอริทึม ID3-A* จะทำการค้นหาแอตทริบิวต์ที่มีความสำคัญทั้งหมด และแอตทริบิวต์ใดที่ทำให้ความลึกสูงสุดของต้นไม้ตัดสินใจผลลัพธ์มีความลึกน้อยที่สุดจะถูกเลือกนำมาสร้างโหนดในต้นไม้ตัดสินใจเพื่อเป็นการขจัดปัญหาแอตทริบิวต์ที่มีความสำคัญเท่าเทียมกัน ดังนั้นอัลกอริทึม ID3-A* จะต้องค้นหาในทุก ๆ กรณีของต้นไม้ตัดสินใจที่เป็นไปได้จากจำนวนแอตทริบิวต์ที่มีความสำคัญเท่ากัน แล้วเลือกต้นไม้ตัดสินใจที่มีจำนวนความลึกสูงสุดต่ำที่สุดออกมา จึงส่งผลให้ชุดข้อมูลที่มีจำนวนแอตทริบิวต์ที่มีความสำคัญเท่ากันเป็นจำนวนมากไม่สามารถทดลองได้สำเร็จในระยะเวลาที่มี ชุดข้อมูลดังกล่าวได้แก่ Phishing Websites, Congressional Voting Records, Insurance Company Benchmark, Molecular Biology, Soybean และ SPECT Heart ดังนั้นสำหรับผลการทดลองในการเปรียบเทียบกับอัลกอริทึม ID3-A* จะแสดงเฉพาะชุดข้อมูลที่สามารถทดลองจนสำเร็จ

5.6.1 ความแม่นยำในการจำแนกและความลึกสูงสุดโดยเฉลี่ย

ตารางที่ 5.19 แสดงร้อยละความแม่นยำในการจำแนกระหว่าง MII-ID3 และ ID3-A*

กลุ่ม	ชุดข้อมูล	ร้อยละของความแม่นยำโดยเฉลี่ย	
		อัลกอริทึม MII-ID3	อัลกอริทึม ID3-A*
ขนาดใหญ่	Connect-4	**73.59**	73.58
ขนาดกลาง	Audiology	75.01	**76.92**
	Car Evaluation	83.42	**90.29**
	Chess	94.82	**99.25**
	Mushroom	99.64	**100.00**
	Nursery	89.16	**96.98**
	Firm-Teacher_Clave-Direction_Classification	67.78	**71.65**
ขนาดเล็ก	Balance Scale	**64.38**	62.94
	Hayes-Roth	73.69	**79.01**
	MONK's Problems	86.96	**100.00**
	Tic-Tac-Toe Endgame	77.68	**85.59**

จากการทดลองในตารางที่ 5.19 แสดงให้เห็นว่าความแม่นยำในการจำแนกในอัลกอริทึม MII-ID3 มีค่ามากกว่าความแม่นยำในอัลกอริทึม ID3-A* เพียงเล็กน้อยหรือแทบจะไม่ต่างกันแค่ในชุดข้อมูล Connect-4 และ Balance Scale ในชุดข้อมูลอื่น ๆ ในตารางที่ 5.19 นั้นอัลกอริทึม MII-ID3 มีความแม่นยำในการจำแนกน้อยกว่าความแม่นยำในอัลกอริทึม ID3-A* ซึ่งในชุดข้อมูลได้แก่ Car Evaluation, Chess, Nursery, Hayes-Roth, MONK's Problems, Tic-Tac-Toe Endgame และ Firm-Teacher_Clave-Direction_Classification มีความแม่นยำในการจำแนกใน

อัลกอริทึม MII-ID3 น้อยกว่าความแม่นยำในอัลกอริทึม ID3-A* อย่างเห็นได้ชัดและมีเพียงในชุดข้อมูล Audiology และ Mushroom เท่านั้นที่มีความแม่นยำในอัลกอริทึม MII-ID3 น้อยกว่าความแม่นยำในอัลกอริทึม ID3-A* เพียงเล็กน้อยหรือแทบจะไม่ต่างกัน ดังนั้นจะเห็นว่าอัลกอริทึม ID3-A* มีความแม่นยำในชุดข้อมูลส่วนใหญ่มากกว่าอัลกอริทึม MII-ID3 และยังสามารถรักษาความแม่นยำในการจำแนกจากอัลกอริทึม ID3 แบบดั้งเดิมให้อยู่ในระดับที่ใกล้เคียงทุก ๆ ชุดข้อมูลในตารางที่ 5.19 (ความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมในทุก ๆ ชุดข้อมูลดูได้จากตารางที่ 5.6-5.8) ซึ่งไม่เหมือนกับอัลกอริทึม MII-ID3 ที่มีความแม่นยำในการจำแนกลดลงจากอัลกอริทึม ID3 แบบดั้งเดิมอย่างเห็นได้ชัดในบางชุดข้อมูลขนาดกลางและส่วนมากในชุดข้อมูลขนาดเล็ก ซึ่งได้แก่ชุดข้อมูลดังนี้ Car Evaluation, Chess, Nursery, Firm-Teacher_Clave-Direction_Classification, Hayes-Roth, MONK's Problems และ Tic-Tac-Toe Endgame ส่วนชุดข้อมูลที่เหลืออื่น ๆ ในตารางที่ 5.19 นั้นอัลกอริทึม MII-ID3 มีความแม่นยำมากกว่าอัลกอริทึม ID3 ในชุดข้อมูล Balance Scale และมีความแม่นยำที่น้อยกว่าอัลกอริทึม ID3 เพียงเล็กน้อยในชุดข้อมูล Connect-4, Audiology และ Mushroom แต่อย่างไรก็ตามยังมีชุดข้อมูลที่อัลกอริทึม MII-ID3 มีความแม่นยำในการจำแนกมากกว่าอัลกอริทึม ID3 หรือน้อยกว่าเพียงเล็กน้อยอีก 6 ชุดข้อมูลได้แก่ Phishing Websites, Congressional Voting Records, Insurance Company Benchmark, Molecular Biology, Soybean และ SPECT Heart ดังที่ได้กล่าวไว้ว่าอัลกอริทึม ID3-A* ไม่สามารถทำการทดลองในชุดข้อมูลเหล่านี้ได้สำเร็จ หรืออาจจะสำเร็จแต่ต้องใช้ระยะเวลาในการทดลองที่นานมาก ๆ เนื่องจากชุดข้อมูลเหล่านี้มีความถี่ในการเกิดกรณีที่มีแอตทริบิวต์ที่มีความสำคัญเท่ากันบ่อยมาก ๆ และมีจำนวนแอตทริบิวต์ที่มีความสำคัญเท่ากันเป็นจำนวนมาก

ตารางที่ 5.20 แสดงความลึกสูงสุดโดยเฉลี่ยระหว่าง MII-ID3 และ ID3-A*

ชุดข้อมูล	ความลึกสูงสุดโดยเฉลี่ย	
	อัลกอริทึม MII-ID3	อัลกอริทึม ID3-A*
Connect-4	**15.73**	19.00
Audiology	**5.80**	6.00
Car Evaluation	**4.58**	6.00
Chess	**7.41**	13.00
Mushroom	**2.76**	4.00
Nursery	**5.37**	8.00
Firm-Teacher_Clave-Direction_Classification	**10.97**	15.00
Balance Scale	**3.78**	4.00
Hayes-Roth	**3.20**	4.00
MONK's Problems	**3.90**	4.00
Tic-Tac-Toe Endgame	**5.16**	7.00

จากผลการทดลองในตาราง 5.20 แสดงให้เห็นว่าอัลกอริทึม MII-ID3 มีจำนวนความลึกสูงสุดของต้นไม้ตัดสินใจผลลัพธ์น้อยกว่าจำนวนความลึกสูงสุดของต้นไม้ตัดสินใจผลลัพธ์ในอัลกอริทึม ID3-A* ในทุก ๆ ชุดข้อมูล จะเห็นว่าอัลกอริทึม MII-ID3 สามารถลดจำนวนความลึกสูงสุดจากอัลกอริทึม ID3 แบบดั้งเดิมได้มากกว่าอัลกอริทึม ID3-A* ส่วนอัลกอริทึม ID3-A* สามารถลดความลึกสูงสุดจากอัลกอริทึม ID3 ในต้นไม้ตัดสินใจผลลัพธ์เพียงในชุดข้อมูล Connect-4 และ Firm-Teacher_Clave-Direction_Classification คิดเป็น 16.41 และ 2.18 เปอร์เซ็นต์ตามลำดับ (ความลึกสูงสุดของอัลกอริทึม ID3 แบบดั้งเดิมในทุกชุดข้อมูลดูได้จากตารางที่ 5.6-5.8) ส่วนชุดข้อมูลอื่น ๆ ในตารางที่ 5.20 อัลกอริทึม ID3-A* ไม่สามารถลดจำนวนความลึกได้หรือมีจำนวนความลึกสูงสุดโดยเฉลี่ยในอัลกอริทึม ID3-A* เท่ากับอัลกอริทึม ID3 ที่อัลกอริทึม ID3-A* สามารถลดความลึกสูงสุดของต้นไม้ตัดสินใจได้แค่ในชุดข้อมูล Connect-4 และ Firm-Teacher_Clave-Direction_Classification ก็เพราะว่าชุดข้อมูลทั้งสองนั้นไม่เพียงแต่มีแอตทริบิวต์ที่มีความสำคัญเท่ากัน แต่ในแอตทริบิวต์ที่มีความสำคัญเท่ากันนั้นนำไปสู่ความลึกของต้นไม้ตัดสินใจผลลัพธ์ที่แตกต่างกัน ดังที่ได้กล่าวไว้ว่าอัลกอริทึม ID3-A* จะค้นหาต้นไม้ตัดสินใจผลลัพธ์ในแต่ละกรณีที่เป็นไปได้จากแอตทริบิวต์ที่มีความสำคัญเท่ากัน แล้วจะคัดเลือกต้นไม้ตัดสินใจผลลัพธ์ที่มีความลึกสูงสุดต่ำที่สุดออกมา แม้ว่าในตารางที่ 5.20 มีชุดข้อมูลที่มีแอตทริบิวต์ที่มีความสำคัญเท่ากันได้หลายชุดข้อมูล แต่มีเพียงในชุดข้อมูล Connect-4 และ Firm-Teacher_Clave-Direction_Classification เท่านั้นที่อัลกอริทึม ID3-A* สามารถลดความลึกสูงสุดได้นั้นก็เพราะว่าชุดข้อมูลอื่น ๆ นั้นอาจจะมีแอตทริบิวต์ที่มีความสำคัญเท่ากันจริง แต่แอตทริบิวต์ที่มีความสำคัญเท่ากันนั้นไม่ได้นำไปสู่ต้นไม้ตัดสินใจผลลัพธ์ที่มีความลึกสูงสุดแตกต่างกัน กล่าวง่าย ๆ คือไม่ว่าจะเลือกแอตทริบิวต์ที่มีความสำคัญตัวใดก็ตาม จะได้ผลลัพธ์เป็นต้นไม้ตัดสินใจที่มีความลึกสูงสุดเท่ากันทุกกรณีและบางชุดข้อมูลอาจจะไม่มีแอตทริบิวต์ที่มีความสำคัญเท่ากันเลยส่งผลให้ต้นไม้ตัดสินใจผลลัพธ์ที่เป็นไปได้จึงมีเพียงแค่กรณีเดียวดังนั้นความลึกสูงสุดของชุดข้อมูลจำพวกนี้จึงไม่ลดลงในอัลกอริทึม ID3-A* เพราะได้ต้นไม้ตัดสินใจผลลัพธ์เป็นแบบเดียวกับอัลกอริทึม ID3 แบบดั้งเดิมนั้นเอง ส่วนอัลกอริทึม MII-ID3 สามารถลดความลึกสูงสุดได้ในชุดข้อมูลทุกตัว อย่างไรก็ตามดังที่ได้กล่าวไว้แม้อัลกอริทึม MII-ID3 จะลดความลึกสูงสุดในต้นไม้ตัดสินใจผลลัพธ์ได้ในทุก ๆ ชุดข้อมูล แต่ชุดข้อมูลที่คุ้มค่าในการใช้ MII-ID3 คือชุดข้อมูลที่มีความแม่นยำในการจำแนกมากกว่าหรือน้อยกว่าความแม่นยำในอัลกอริทึม ID3 เพียงเล็กน้อย

5.6.2 จำนวนกฎการตัดสินใจโดยเฉลี่ย

ตารางที่ 5.21 แสดงจำนวนกฎการตัดสินใจโดยเฉลี่ยระหว่าง MII-ID3 และ ID3-A*

ชุดข้อมูล	จำนวนกฎการตัดสินใจโดยเฉลี่ย	
	อัลกอริทึม MII-ID3	อัลกอริทึม ID3-A*
Connect-4	**4646.37**	14649.00
Audiology	**40.89**	45.00
Car Evaluation	**50.34**	222.00
Chess	**14.17**	37.00
Mushroom	**22.80**	33.00
Nursery	**71.88**	641.00
Firm-Teacher_Clave-Direction_Classification	**120.02**	1555.00
Balance Scale	**97.82**	257.00
Hayes-Roth	**23.47**	34.00
MONK's Problems	**35.84**	55.00
Tic-Tac-Toe Endgame	**48.86**	135.00

จากผลการทดลองในตารางที่ 5.21 แสดงให้เห็นว่าต้นไม้ตัดสินใจผลลัพธ์ของอัลกอริทึม MII-ID3 มีจำนวนกฎการตัดสินใจที่น้อยกว่าอัลกอริทึม ID3-A* อย่างเห็นได้ชัดเจนในทุก ๆ ชุดข้อมูล อัลกอริทึม ID3-A* ในชุดข้อมูลส่วนใหญ่ได้แก่ชุดข้อมูล Chess, Balance Scale, Hayes-Roth, MONK's Problem และ Tic-Tac_Toe Endgame มีจำนวนกฎการตัดสินใจเท่ากับอัลกอริทึม ID3 แบบดั้งเดิมก็คือไม่สามารถลดกฎการตัดสินใจได้ (จำนวนกฎการตัดสินใจของอัลกอริทึม ID3 แบบดั้งเดิมในทุกชุดข้อมูลดูได้จากตารางที่ 5.13-5.14) ในชุดข้อมูล Audiology, Car Evaluation และ Nursery มีจำนวนกฎการตัดสินในอัลกอริทึม ID3-A* มากกว่าอัลกอริทึม ID3 แบบดั้งเดิม ในชุดข้อมูล Connect-4, Mushroom, Firm-Teacher_Clave-Direction_Classification อัลกอริทึม ID3-A* สามารถลดจำนวนกฎการตัดสินใจได้เพียงเล็กน้อยคิดเป็น 0.64, 3.84 และ 1.63 เปอร์เซ็นต์ตามลำดับในขณะที่อัลกอริทึม MII-ID3 สามารถลดจำนวนกฎการตัดสินใจจากอัลกอริทึม ID3 ในชุดข้อมูลส่วนใหญ่อย่างเห็นได้ชัดมีเพียงชุดข้อมูล Audiology เท่านั้นที่อัลกอริทึม MII-ID3 ลดจำนวนกฎการตัดสินใจได้เพียงเล็กน้อย

5.6.3 จำนวนโหนดทั้งหมดในต้นไม้ตัดสินใจผลลัพธ์โดยเฉลี่ย

ตารางที่ 5.22 แสดงจำนวนโหนดทั้งหมดโดยเฉลี่ยระหว่าง MII-ID3 และ ID3-A*

ชุดข้อมูล	จำนวนโหนดทั้งหมดโดยเฉลี่ย	
	อัลกอริทึม MII-ID3	อัลกอริทึม ID3-A*
Connect-4	**6969.06**	21973.00
Audiology	**61.67**	67.00
Car Evaluation	**68.86**	304.00
Chess	**26.71**	71.00
Mushroom	**25.80**	38.00
Nursery	**102.59**	903.00
Firm-Teacher_Clave-Direction_Classification	**239.04**	3109.00
Balance Scale	**122.02**	321.00
Hayes-Roth	**31.05**	46.00
MONK's Problems	**53.25**	82.00
Tic-Tac-Toe Endgame	**72.79**	202.00

จากผลการทดลองในตารางที่ 5.22 แสดงจำนวนโหนดทั้งหมดในต้นไม้ตัดสินใจผลลัพธ์ ซึ่งจำนวนโหนดทั้งหมดในตารางที่ 5.22 จะมีลักษณะที่เพิ่มขึ้นลดลงหรือเป็นไปในทางเดียวกันกับจำนวนกฎการตัดสินใจในตารางที่ 5.21 จากตารางที่ 5.22 แสดงให้เห็นว่าต้นไม้ตัดสินใจผลลัพธ์ของอัลกอริทึม MII-ID3 มีจำนวนโหนดทั้งหมดน้อยกว่าอัลกอริทึม ID3-A* อย่างเห็นได้ชัดเจนในทุก ๆ ชุดข้อมูล ดังนั้นอัลกอริทึม MII-ID3 สามารถที่จะลดจำนวนโหนดทั้งหมดจากอัลกอริทึม ID3 แบบดั้งเดิมได้มากกว่าและดีกว่าอัลกอริทึม ID3-A* ในอัลกอริทึม ID3-A* นั้นชุดข้อมูลส่วนใหญ่ซึ่งได้แก่ชุดข้อมูล Chess, Balance Scale, Hayes-Roth, MONK's Problem และ Tic-Tac_Toe Endgame มีจำนวนโหนดทั้งหมดเท่ากับอัลกอริทึม ID3 แบบดั้งเดิม (จำนวนโหนดทั้งหมดของอัลกอริทึม ID3 แบบดั้งเดิมในทุกชุดข้อมูลดูได้จากตารางที่ 5.15-5.16) ในชุดข้อมูล Audiology, Car Evaluation และ Nursery มีจำนวนโหนดทั้งหมดในอัลกอริทึม ID3-A*มากกว่าอัลกอริทึม ID3 แบบดั้งเดิม ในชุดข้อมูล Connect-4, Mushroom, Firm-Teacher_Clave-Direction_Classification อัลกอริทึม ID3-A* สามารถลดจำนวนโหนดทั้งหมดจากอัลกอริทึม ID3 แบบดั้งเดิมเพียงเล็กน้อยในขณะที่อัลกอริทึม MII-ID3 สามารถลดจำนวนโหนดทั้งหมดจากอัลกอริทึม ID3 ในชุดข้อมูลส่วนใหญ่อย่างเห็นได้ชัดในชุดข้อมูลเกือบจะทั้งหมด มีเพียงชุดข้อมูล Audiology เท่านั้นที่อัลกอริทึม MII-ID3 ลดจำนวนโหนดทั้งหมดได้เพียงเล็กน้อย

5.6.4 เวลาในการฝึกฝนและเวลาในการทดสอบต้นไม้ตัดสินใจผลลัพธ์โดยเฉลี่ย

ตารางที่ 5.23 แสดงเวลาในการฝึกฝนต้นไม้ตัดสินใจผลลัพธ์โดยเฉลี่ยระหว่าง MII-ID3 และ ID3-A*

ชุดข้อมูล	เวลาในการฝึกฝนโดยเฉลี่ย (ms)	
	อัลกอริทึม MII-ID3	อัลกอริทึม ID3-A*
Connect-4	**1040.922**	5875.259
Audiology	**10.146**	51.838
Car Evaluation	**1.454**	22.389
Chess	**12.913**	39.148
Mushroom	**14.211**	50.420
Nursery	**15.182**	50.867
Firm-Teacher_Clave-Direction_Classification	**46.897**	427.909
Balance Scale	**0.572**	25.378
Hayes-Roth	**0.136**	19.083
MONK's Problems	**0.452**	20.445
Tic-Tac-Toe Endgame	**1.373**	21.564

จากผลการทดลองในตารางที่ 5.23 แสดงให้เห็นว่าอัลกอริทึม MII-ID3 ใช้เวลาในการฝึกฝนหรือเวลาในการสร้างต้นไม้ตัดสินใจผลลัพธ์น้อยกว่าอัลกอริทึม ID3-A* อย่างเห็นได้ชัดในชุดข้อมูลทั้งหมด หมายความว่าอัลกอริทึม MII-ID3 สามารถลดเวลาในการฝึกฝนจากอัลกอริทึม ID3 แบบดั้งเดิมลงได้ดีกว่าอัลกอริทึม ID3-A* พิจารณาที่อัลกอริทึม ID3-A* ใช้เวลาในการฝึกฝนเพื่อให้ได้ซึ่งต้นไม้ตัดสินใจผลลัพธ์มากกว่าอัลกอริทึม ID3 แบบดั้งเดิมอย่างมาก (เวลาในการฝึกฝนของอัลกอริทึม ID3 แบบดั้งเดิมในทุกชุดข้อมูลดูได้จากตารางที่ 5.17) โดยอัลกอริทึม ID3-A* ใช้เวลาในการฝึกฝนเพิ่มขึ้นจากอัลกอริทึม ID3 แบบดั้งเดิมมากกว่า 59 เปอร์เซ็นต์ในชุดข้อมูลทั้งหมด ส่วนในชุดข้อมูล Car Evaluation, Balance Scale, Hayes-Roth, MONK's Problems และ Tic-Tac_Toe Endgame ใช้เวลาในการฝึกฝนในอัลกอริทึม ID3-A* เพิ่มขึ้นจากเวลาในการฝึกฝนในอัลกอริทึม ID3 มากกว่า 90 เปอร์เซ็นต์ขึ้นไป ดังนั้นถ้าชุดข้อมูลมีความซับซ้อนมากหรือมีจำนวนแอตทริบิวต์ที่เท่ากันเป็นจำนวนมากอัลกอริทึม ID3-A* จะใช้เวลาในการฝึกฝนนานมาก ๆ

ตารางที่ 5.24 แสดงเวลาในการทดสอบต้นไม้ตัดสินใจผลลัพธ์โดยเฉลี่ยระหว่าง MII-ID3 และ ID3-A*

ชุดข้อมูล	เวลาในการทดสอบโดยเฉลี่ย (ms)	
	อัลกอริทึม MII-ID3	อัลกอริทึม ID3-A*
Connect-4	**0.191490**	0.253690
Audiology	**0.000369**	0.000381
Car Evaluation	**0.001594**	0.001981
Chess	**0.004598**	0.005587
Mushroom	**0.005124**	0.006752
Nursery	**0.013118**	0.017707
Firm-Teacher_Clave-Direction_Classification	**0.024379**	0.041302
Balance Scale	**0.000655**	0.000777
Hayes-Roth	**0.000155**	0.000182
MONK's Problems	**0.000586**	0.000677
Tic-Tac-Toe Endgame	**0.001276**	0.001629

จากผลการทดลองในตารางที่ 5.24 แสดงให้เห็นว่าอัลกอริทึม MII-ID3 ใช้เวลาในการทดสอบต้นไม้ตัดสินใจผลลัพธ์น้อยกว่าอัลกอริทึม ID3-A* อย่างเห็นได้ชัดในชุดข้อมูลทั้งหมด ดังนั้นอัลกอริทึม MII-ID3 สามารถลดเวลาในการทดสอบลงได้ดีกว่าอัลกอริทึม ID3-A* ส่วนอัลกอริทึม ID3-A* ใช้เวลาในการทดสอบต้นไม้ตัดสินใจผลลัพธ์ใกล้เคียงหรือแทบจะไม่แตกต่างจากอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูลทั้งหมด (เวลาในการทดสอบของอัลกอริทึม ID3 แบบดั้งเดิมในทุกชุดข้อมูลดูได้จากตารางที่ 5.17) ดังนั้นจึงต้องไปพิจารณาที่ความลึกโดยเฉลี่ย

ตารางที่ 5.25 แสดงความลึกโดยเฉลี่ยในการจำแนกกรณีตัวอย่างระหว่าง MII-ID3 และ ID3-A*

ชุดข้อมูล	ความลึกโดยเฉลี่ยในการจำแนก	
	อัลกอริทึม MII-ID3	อัลกอริทึม ID3-A*
Connect-4	**7.02**	9.36
Audiology	**3.94**	4.04
Car Evaluation	**2.29**	2.89
Chess	**3.57**	4.35
Mushroom	**1.48**	1.52
Nursery	**2.50**	3.38
Firm-Teacher_Clave-Direction_Classification	**5.59**	9.58
Balance Scale	**2.59**	3.14
Hayes-Roth	**2.37**	2.78
MONK's Problems	**2.60**	3.01
Tic-Tac-Toe Endgame	**3.30**	4.21

จากผลการทดลองในตารางที่ 5.25 แสดงให้เห็นว่าอัลกอริทึม MII-ID3 มีความลึกโดยเฉลี่ยในการจำแนกกรณีตัวอย่างน้อยกว่าความลึกโดยเฉลี่ยในการจำแนกกรณีตัวอย่างในอัลกอริทึม ID3-A* อย่างชัดเจนในชุดข้อมูลทั้งหมดจึงส่งผลให้เวลาในการทดสอบของอัลกอริทึม MII-ID3 นั้นลดลงในทุก ๆ ชุดข้อมูล ในส่วนของอัลกอริทึม ID3-A* นั้นมีความลึกโดยเฉลี่ยในการจำแนกกรณีตัวอย่างใกล้เคียงหรือแทบจะไม่แตกต่างจากอัลกอริทึม ID3 แบบดั้งเดิมในชุดข้อมูลทั้งหมด (ความลึกโดยเฉลี่ยในการจำแนกกรณีตัวอย่างของอัลกอริทึม ID3 แบบดั้งเดิมในทุกชุดข้อมูลดูได้จากตารางที่ 5.18) ดังนั้นเมื่อความลึกโดยเฉลี่ยในการจำแนกกรณีตัวอย่างแทบจะไม่แตกต่างกันจึงส่งผลให้เวลาในการทดสอบต้นไม้ตัดสินใจผลลัพธ์ของอัลกอริทึม ID3 และ ID3-A* แทบจะไม่แตกต่างกันด้วย (เวลาในการทดสอบของอัลกอริทึม ID3-A* ในทุกชุดข้อมูลดูได้จากตารางที่ 5.24)

สรุปผลการทดลองระหว่างอัลกอริทึม MII-ID3 และอัลกอริทึม ID3-A* พบว่าอัลกอริทึม MII-ID3 มีจำนวนความลึกสูงสุด, จำนวนกฎการตัดสินใจ, จำนวนโหนดทั้งหมด, เวลาในการฝึกฝน, เวลาในการทดสอบ และความลึกโดยเฉลี่ยในการจำแนกน้อยกว่าอัลกอริทึม ID3-A* ในชุดข้อมูลทั้งหมด โดยเฉพาะอย่างยิ่งเวลาในการฝึกฝนของอัลกอริทึม ID3-A* ใช้เวลามากกว่าอัลกอริทึม ID3 แบบดั้งเดิมอย่างมาก ในขณะที่อัลกอริทึม MII-ID3 ใช้เวลาในการฝึกฝนน้อยกว่าอัลกอริทึม ID3 แบบดั้งเดิมอย่างชัดเจน อย่างไรก็ตามอัลกอริทึม ID3-A* มีความแม่นยำในการจำแนกมากกว่าอัลกอริทึม MII-ID3 อย่างชัดเจนในชุดข้อมูลส่วนมาก

5.7 ผลการทดลองระหว่างอัลกอริทึม MII-ID3 และ EVC-ID3

ดังที่ได้อธิบายอัลกอริทึม EVC-ID3 ไว้ในบทที่ 3 เมื่อเกิดเหตุการณ์ที่มีแอตทริบิวต์ที่มีความสำคัญเท่ากันในขณะสร้างต้นไม้ตัดสินใจ อัลกอริทึม EVC-ID3 จะนำแอตทริบิวต์ที่เท่ากันมารวมกันเป็นโหนดเดียวกันแล้วจะขยายกิ่งไปตามค่าที่ไม่ซ้ำกันของแอตทริบิวต์ที่นำมาจับคู่กัน โดยงานวิจัยนี้กำหนดจำนวนการจับคู่ของแอตทริบิวต์ที่มีความสำคัญเท่ากันไว้ 2 แอตทริบิวต์ เนื่องจากถ้าหากจำนวนแอตทริบิวต์ที่มีความสำคัญเท่ากันที่จะนำมาสร้างโหนดมีจำนวนมาก จำนวนของกิ่งที่จะถูกขยายจากโหนดที่รวมแอตทริบิวต์ที่สำคัญนั้นจะยิ่งเพิ่มมากขึ้น ดังนั้นผลการทดลองในอัลกอริทึม EVC-ID3 ดังต่อไปนี้เมื่อเกิดกรณีที่มีแอตทริบิวต์ที่มีความสำคัญเท่ากัน จะทำการสุ่มแอตทริบิวต์ที่มีความสำคัญเท่ากันมา 2 แอตทริบิวต์แล้วนำมาสร้างเป็นโหนดเดียวกัน

5.7.1 ความแม่นยำในการจำแนกและความลึกสูงสุดโดยเฉลี่ย

ตารางที่ 5.26 แสดงร้อยละความแม่นยำในการจำแนกระหว่าง MII-ID3 และ EVC-ID3

กลุ่ม	ชุดข้อมูล	ร้อยละของความแม่นยำโดยเฉลี่ย	
		อัลกอริทึม MII-ID3	อัลกอริทึม EVC-ID3
ขนาดใหญ่	Connect-4	73.59	**73.98**
	Phishing Websites	93.65	**94.94**
	Insurance Company Benchmark	**90.914**	90.913
	Molecular Biology	88.75	**89.87**
ขนาดกลาง	Soybean	86.21	**86.57**
	Audiology	75.01	**76.29**
	Car Evaluation	83.42	**90.52**
	Chess	94.82	**99.38**
	Congressional Voting Records	**95.39**	93.70
	Mushroom	99.64	**100.00**
	Nursery	89.16	**96.93**
	SPECT Heart	79.99	**82.25**
	Firm-Teacher_Clave-Direction_Classification	67.78	**72.27**
ขนาดเล็ก	Balance Scale	**64.38**	61.98
	Hayes-Roth	73.69	**79.01**
	MONK's Problems	86.96	**98.20**
	Tic-Tac-Toe Endgame	77.68	**86.23**

จากการทดลองในตารางที่ 5.26 แสดงให้เห็นว่าความแม่นยำในการจำแนกในอัลกอริทึม MII-ID3 มีค่ามากกว่าความแม่นยำในอัลกอริทึม EVC-ID3 เพียงเล็กน้อยหรือแทบจะไม่แตกต่างต่างกันเลยในชุดข้อมูล Insurance Company Benchmark, Congressional Voting Records และ Balance Scale ส่วนชุดข้อมูลอื่น ๆ ในตารางที่ 5.26 นั้นอัลกอริทึม MII-ID3 มีความ

แม่นยำในการจำแนกน้อยกว่าอัลกอริทึม EVC-ID3 โดยอัลกอริทึม MII-ID3 มีความแม่นยำในการจำแนกน้อยกว่าอัลกอริทึม EVC-ID3 เพียงเล็กน้อยหรือแทบจะไม่ต่างกันในชุดข้อมูล Connect-4, Phishing Websites, Molecular Biology, Soybean, Audiology และ Mushroom และอัลกอริทึม MII-ID3 มีความแม่นยำในการจำแนกน้อยกว่าอัลกอริทึม EVC-ID3 อย่างเห็นได้ชัดเจนในชุดข้อมูล Car Evaluation, Chess, Nursery, SPECT Heart, Hayes-Roth, MONK's Problems, Firm-Teacher_Clave-Direction_Classification และ Tic-Tac-Toe Endgame ซึ่งชุดข้อมูลทั้งเจ็ดชุดนี้ (ยกเว้นชุดข้อมูล SPECT Heart) ล้วนเป็นชุดข้อมูลที่ไม่คุ้มค่าที่จะใช้อัลอริทึม MII-ID3 เพราะความแม่นนำในอัลกอริทึม MII-ID3 ของชุดข้อมูลเหล่านี้ลดลงจากอัลกอริทึม ID3 แบบดั้งเดิมอย่างชัดเจน ดังนั้นอัลกอริทึม EVC-ID3 สามารถรักษาระดับความแม่นยำในการจำแนกจากอัลกอริทึม ID3 แบบดั้งเดิมให้อยู่ในระดับที่น่าพอใจในชุดข้อมูลทุกขนาด (ความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมในทุก ๆ ชุดข้อมูลดูได้จากตารางที่ 5.6-5.8) ในขณะที่อัลกอริทึม MII-ID3 สามารถรักษาความแม่นยำให้ลดลงเพียงเล็กน้อยหรือมากกว่าอัลกอริทึม ID3 ในชุดข้อมูลขนาดใหญ่ทุกชุดข้อมูล ส่วนขนาดกลางมีทั้งสามารถรักษาความแม่นยำได้และไม่ได้ปะปนกันไปเกือบจะเท่า ๆ กัน ในชุดข้อมูลขนาดเล็กมีเพียงส่วนน้อยเท่านั้นที่จะใช้อัลกอริทึม MII-ID3 แล้วความแม่นยำไม่ลดลงอย่างชัดเจนจากอัลกอริทึม ID3

ตารางที่ 5.27 แสดงความลึกสูงสุดโดยเฉลี่ยระหว่าง MII-ID3 และ EVC-ID3

ชุดข้อมูล	ความลึกสูงสุดโดยเฉลี่ย	
	อัลกอริทึม MII-ID3	อัลกอริทึม EVC-ID3
Connect-4	**15.73**	18.31
Phishing Websites	**10.94**	23.00
Insurance Company Benchmark	**24.01**	45.00
Molecular Biology	**8.79**	33.00
Soybean	**8.08**	20.00
Audiology	5.80	**5.35**
Car Evaluation	**4.58**	6.00
Chess	**7.41**	12.00
Congressional Voting Records	**4.35**	11.00
Mushroom	**2.76**	4.00
Nursery	**5.37**	8.00
SPECT Heart	**7.29**	15.36
Firm-Teacher_Clave-Direction_Classification	**10.97**	14.65
Balance Scale	**3.78**	4.00
Hayes-Roth	**3.20**	4.00
MONK's Problems	**3.90**	4.00
Tic-Tac-Toe Endgame	**5.16**	7.00

จากผลการทดลองในตารางที่ 5.27 แสดงให้เห็นว่าอัลกอริทึม MII-ID3 มีความลึกสูงสุดของต้นไม้ตัดสินใจผลลัพธ์น้อยกว่าอัลกอริทึม EVC-ID3 ในชุดข้อมูลส่วนมากอย่างเห็นได้ชัด มีเพียงชุดข้อมูล Audiology เท่านั้นที่มีความลึกสูงสุดในอัลกอริทึม EVC-ID3 น้อยกว่า MII-ID3 สำหรับอัลกอริทึม EVC-ID3 มีชุดข้อมูลที่ความลึกสูงสุดโดยเฉลี่ยไม่ลดลงจากอัลกอริทึม ID3 แบบดั้งเดิม (ความลึกสูงสุดของอัลกอริทึม ID3 แบบดั้งเดิมในทุกชุดข้อมูลดูได้จากตารางที่ 5.6-5.8) ได้แก่ชุดข้อมูลดังต่อไปนี้ Car Evaluation, Mushroom, Nursery, Balance Scale, Hayes-Roth, MONK's Problems และ Tic-Tac-Toe Endgame ที่ความลึกสูงสุดไม่ลดลงในข้อมูลเหล่านี้เพราะการที่จะลดความลึกสูงสุดลงได้อย่างชัดเจนในอัลกอริทึม EVC-ID3 นั้นชุดข้อมูลใด ๆ จะต้องมีความถี่ในการเกิดกรณีที่มีแอตทริบิวต์ที่มีความสำคัญเท่ากันบ่อยครั้งในขณะสร้างต้นไม้ ตัดสินใจ ดังนั้นอัลกอริทึม MII-ID3 สามารถลดความลึกสูงสุดจากอัลกอริทึม ID3 แบบดั้งเดิมได้มากกว่าอัลกอริทึม EVC-ID3 อย่างเห็นได้ชัด

5.7.2 จำนวนกฎการตัดสินใจโดยเฉลี่ย

ตารางที่ 5.28 แสดงจำนวนกฎการตัดสินใจโดยเฉลี่ยระหว่าง MII-ID3 และ EVC-ID3

ชุดข้อมูล	จำนวนกฎการตัดสินใจโดยเฉลี่ย	
	อัลกอริทึม MII-ID3	อัลกอริทึม EVC-ID3
Connect-4	**4646.37**	25833.80
Phishing Websites	**100.77**	1462.76
Insurance Company Benchmark	**3421.84**	94954.50
Molecular Biology	**288.36**	1639.35
Soybean	**75.63**	245.24
Audiology	**40.89**	73.06
Car Evaluation	**50.34**	272.00
Chess	**14.17**	47.34
Congressional Voting Records	**7.42**	36.00
Mushroom	**22.80**	112.00
Nursery	**71.88**	776.00
SPECT Heart	**15.48**	112.26
Firm-Teacher_Clave-Direction_Classification	**120.02**	2245.37
Balance Scale	**97.82**	285.00
Hayes-Roth	**23.47**	34.00
MONK's Problems	**35.84**	61.00
Tic-Tac-Toe Endgame	**48.86**	223.62

จากผลการทดลองในตารางที่ 5.28 แสดงให้เห็นว่าอัลกอริทึม MII-ID3 มีจำนวนกฎการตัดสินใจโดยเฉลี่ยน้อยกว่าจำนวนกฎการตัดสินใจในอัลกอริทึม EVC-ID3 ในชุดข้อมูลทั้งหมด

ดังนั้นอัลกอริทึม MII-ID3 จึงสามารถลดจำนวนกฎการตัดสินใจจากอัลกอริทึม ID3 แบบดั้งเดิมได้มากกว่าอัลกอริทึม EVC-ID3 อย่างเห็นได้ชัดในทุก ๆ ชุดข้อมูล (จำนวนกฎการตัดสินใจของอัลกอริทึม ID3 แบบดั้งเดิมในทุกชุดข้อมูลดูได้จากตารางที่ 5.13-5.14) อัลกอริทึม EVC-ID3 มีจำนวนกฎการตัดสินใจที่เพิ่มขึ้นจากอัลกอริทึม ID3 แบบดั้งเดิมอย่างชัดเจนในชุดข้อมูลเกือบจะทั้งหมด มีเพียงชุดข้อมูล Hayes-Roth เท่านั้นที่จำนวนกฎการตัดสินใจในอัลกอริทึม EVC-ID3 เท่ากับอัลกอริทึม ID3 แบบดั้งเดิม ซึ่งจะเห็นได้ว่าอัลกอริทึม EVC-ID3 ไม่สามารถลดจำนวนกฎการตัดสินใจในต้นไม้ตัดสินใจผลลัพธ์ลงได้แล้วยังเพิ่มจำนวนกฎการตัดสินใจขึ้นอีกอย่างเห็นได้ชัด ในขณะที่อัลกอริทึม MII-ID3 นั้นสามารถลดจำนวนกฎการตัดสินใจได้ในทุก ๆ ชุดข้อมูล

5.7.3 จำนวนโหนดทั้งหมดในต้นไม้ตัดสินใจผลลัพธ์โดยเฉลี่ย

ตารางที่ 5.29 แสดงจำนวนโหนดทั้งหมดโดยเฉลี่ยระหว่าง MII-ID3 และ EVC-ID3

ชุดข้อมูล	จำนวนโหนดทั้งหมดโดยเฉลี่ย	
	อัลกอริทึม MII-ID3	อัลกอริทึม EVC-ID3
Connect-4	**6969.06**	32607.92
Phishing Websites	**169.28**	2023.65
Insurance Company Benchmark	**3908.88**	96890.24
Molecular Biology	**362.11**	1755.66
Soybean	**109.03**	296.24
Audiology	**61.67**	93.40
Car Evaluation	**68.86**	347.00
Chess	**26.71**	79.67
Congressional Voting Records	**13.84**	53.00
Mushroom	**25.80**	117.00
Nursery	**102.59**	1024.00
SPECT Heart	**29.97**	161.28
Firm-Teacher_Clave-Direction_Classification	**239.04**	3572.31
Balance Scale	**122.02**	346.00
Hayes-Roth	**31.05**	46.00
MONK's Problems	**53.25**	88.00
Tic-Tac-Toe Endgame	**72.79**	288.95

จากผลการทดลองในตารางที่ 5.29 แสดงให้เห็นว่าอัลกอริทึม MII-ID3 มีจำนวนโหนดทั้งหมดในต้นไม้ตัดสินใจผลลัพธ์โดยเฉลี่ยน้อยกว่าอัลกอริทึม EVC-ID3 อย่างเห็นได้ชัดในชุดข้อมูลทั้งหมด ในส่วนของอัลกอริทึม EVC-ID3 มีจำนวนโหนดทั้งหมดในต้นไม้ตัดสินใจผลลัพธ์โดยเฉลี่ยเพิ่มขึ้นจากอัลกอริทึม ID3 แบบดั้งเดิมอย่างชัดเจนในหลาย ๆ ชุดข้อมูล (จำนวนโหนดทั้งหมดของอัลกอริทึม ID3 แบบดั้งเดิมในทุกชุดข้อมูลดูได้จากตารางที่ 5.15-5.16) มีเพียงชุดข้อมูล Hayes-

Roth เท่านั้นที่จำนวนโหนดทั้งหมดเท่ากับอัลกอริทึม ID3 แบบดั้งเดิม ดังนั้นอัลกอริทึม EVC-ID3 ไม่สามารถลดจำนวนโหนดทั้งหมดในต้นไม้ตัดสินใจผลลัพธ์ลงได้แล้วยังเพิ่มจำนวนโหนดทั้งหมดขึ้นอีกอย่างเห็นได้ชัดเช่นเดียวกับจำนวนกฎการตัดสินใจในตารางที่ 5.28 ในขณะที่อัลกอริทึม MII-ID3 นั้นสามารถลดจำนวนโหนดทั้งหมดได้ในทุก ๆ ชุดข้อมูล

5.7.4 เวลาในการฝึกฝนและเวลาในการทดสอบต้นไม้ตัดสินใจผลลัพธ์โดยเฉลี่ย

ตารางที่ 5.30 แสดงเวลาในการฝึกฝนต้นไม้ตัดสินใจผลลัพธ์โดยเฉลี่ยระหว่าง MII-ID3 และ EVC-ID3

ชุดข้อมูล	เวลาในการฝึกฝนโดยเฉลี่ย (ms)	
	อัลกอริทึม MII-ID3	อัลกอริทึม EVC-ID3
Connect-4	**1040.922**	1706.940
Phishing Websites	**40.600**	64.600
Insurance Company Benchmark	**234.966**	432.224
Molecular Biology	**48.923**	57.961
Soybean	**12.835**	13.863
Audiology	**10.146**	11.135
Car Evaluation	**1.454**	1.933
Chess	**12.913**	15.050
Congressional Voting Records	**0.630**	0.814
Mushroom	**14.211**	14.858
Nursery	**15.182**	19.412
SPECT Heart	**0.950**	1.387
Firm-Teacher_Clave-Direction_Classification	**46.897**	103.747
Balance Scale	**0.572**	0.774
Hayes-Roth	**0.136**	0.166
MONK's Problems	**0.452**	0.533
Tic-Tac-Toe Endgame	**1.373**	1.919

จากผลการทดลองในตารางที่ 5.30 แสดงให้เห็นว่าอัลกอริทึม MII-ID3 ใช้เวลาในการฝึกฝนน้อยกว่าอัลกอริทึม EVC-ID3 อย่างเห็นได้ชัดในชุดข้อมูลทั้งหมด ในส่วนของอัลกอริทึม EVC-ID3 นั้นมีเวลาในการฝึกฝนมากกว่าอัลกอริทึม ID3 แบบดั้งเดิมอย่างชัดเจน (เวลาในการฝึกฝนของอัลกอริทึม ID3 แบบดั้งเดิมในทุกชุดข้อมูลดูได้จากตารางที่ 5.17) ในชุดข้อมูล Connect-4, Audiology และ Firm-Teacher_Clave-Direction_Classification ถัดมาอัลกอริทึม EVC-ID3 ใช้เวลาในการฝึกฝนมากกว่าอัลกอริทึม ID3 แบบดั้งเดิมเพียงเล็กน้อยหรือแทบจะไม่แตกต่างในชุดข้อมูล Phishing Websites, Insurance Company Benchmark, Molecular Biology, Hayes-Roth,

MONK's Problems และ Tic-Tac-Toe Endgame ในส่วนของชุดข้อมูลที่เหลือนั้นได้แก่ชุดข้อมูล Soybean, Car Evaluation, Congressional Voting Records, Chess, Mushroom, Nursery, SPECT Heart และ Balance Scale ใช้เวลาในการฝึกฝนในอัลกอริทึม EVC-ID3 น้อยกว่าอัลกอริทึม ID3 แบบดั้งเดิมเพียงเล็กน้อยหรือแทบจะไม่ต่างกัน ดังนั้นจะเห็นว่าอัลกอริทึม MII-ID3 สามารถลดเวลาในการฝึกฝนลงจากอัลกอริทึม ID3 แบบดั้งเดิมได้ชัดเจนในชุดข้อมูลทั้งหมด ในขณะที่อัลกอริทึม EVC-ID3 ไม่สามารถลดเวลาในการฝึกฝนจากอัลกอริทึม ID3 แบบดั้งเดิมได้อย่างชัดเจนและยังมีบางชุดข้อมูลใช้เวลาในการฝึกฝนเพิ่มขึ้นอย่างชัดเจนอีกด้วย

ตารางที่ 5.31 แสดงเวลาในการทดสอบต้นไม้ตัดสินใจผลลัพธ์โดยเฉลี่ยระหว่าง MII-ID3 และ EVC-ID3

ชุดข้อมูล	เวลาในการทดสอบโดยเฉลี่ย (ms)	
	อัลกอริทึม MII-ID3	อัลกอริทึม EVC-ID3
Connect-4	**0.191490**	0.252507
Phishing Websites	**0.016163**	0.023362
Insurance Company Benchmark	**0.012937**	0.014287
Molecular Biology	**0.004993**	0.005239
Soybean	**0.000949**	0.001037
Audiology	**0.000369**	0.000383
Car Evaluation	**0.001594**	0.001973
Chess	**0.004598**	0.005535
Congressional Voting Records	**0.000415**	0.000488
Mushroom	**0.005124**	0.005349
Nursery	**0.013118**	0.017640
SPECT Heart	**0.000403**	0.000563
Firm-Teacher_Clave-Direction_Classification	**0.024379**	0.040468
Balance Scale	**0.000655**	0.000773
Hayes-Roth	**0.000155**	0.000181
MONK's Problems	**0.000586**	0.000675
Tic-Tac-Toe Endgame	**0.001276**	0.001624

จากผลการทดลองในตารางที่ 5.31 แสดงให้เห็นว่าอัลกอริทึม MII-ID3 ใช้เวลาในการทดสอบต้นไม้ตัดสินใจผลลัพธ์น้อยกว่าอัลกอริทึมอัลกอริทึม EVC-ID3 อย่างเห็นได้ชัดในทุก ๆ ชุดข้อมูล พิจารณาที่อัลกอริทึม EVC-ID3 เวลาในการทดสอบต้นไม้ตัดสินใจผลลัพธ์ลดลงจากอัลกอริทึม ID3 แบบดั้งเดิมเพียงเล็กน้อยหรือแทบจะไม่แตกต่าง (เวลาในการทดสอบของอัลกอริทึม ID3 แบบดั้งเดิมในทุกชุดข้อมูลดูได้จากตารางที่ 5.17) ในชุดข้อมูลเกือบทั้งหมด มีเพียงในชุดข้อมูล Insurance

Company Benchmark และ SPECT Heart ที่เวลาในการทดสอบในอัลกอริทึม EVC-ID3 ลดลงจากอัลกอริทึม ID3 อย่างเห็นได้ชัดโดยลดลงถึง 10.21 เปอร์เซ็นต์และ 19.02 เปอร์เซ็นต์ตามลำดับ

ตารางที่ 5.32 แสดงความลึกโดยเฉลี่ยในการจำแนกกรณีตัวอย่างระหว่าง MII-ID3 และ EVC-ID3

ชุดข้อมูล	ความลึกโดยเฉลี่ยในการจำแนก	
	อัลกอริทึม MII-ID3	อัลกอริทึม EVC-ID3
Connect-4	**7.02**	9.31
Phishing Websites	**3.61**	5.22
Insurance Company Benchmark	**3.27**	3.86
Molecular Biology	**3.86**	4.10
Soybean	**3.43**	3.59
Audiology	**3.94**	3.97
Car Evaluation	**2.29**	2.87
Chess	**3.57**	4.33
Congressional Voting Records	**2.39**	2.66
Mushroom	**1.48**	1.52
Nursery	**2.50**	3.37
SPECT Heart	**3.78**	5.47
Firm-Teacher_Clave-Direction_Classification	**5.59**	9.35
Balance Scale	**2.59**	3.11
Hayes-Roth	**2.37**	2.78
MONK's Problems	**2.60**	3.01
Tic-Tac-Toe Endgame	**3.30**	4.20

จากผลการทดลองในตารางที่ 5.32 แสดงให้เห็นว่าอัลกอริทึม MII-ID3 มีความลึกโดยเฉลี่ยในการจำแนกกรณีตัวอย่างน้อยกว่าอัลกอริทึม EVC-ID3 อย่างเห็นได้ชัดในทุก ๆ ชุดข้อมูลพิจารณาที่อัลกอริทึม EVC-ID3 มีความลึกโดยเฉลี่ยในการจำแนกกรณีตัวอย่างลดลงจากอัลกอริทึม ID3 แบบดั้งเดิม (ความลึกโดยเฉลี่ยในการจำแนกกรณีตัวอย่างของอัลกอริทึม ID3 แบบดั้งเดิมในทุกชุดข้อมูลดูได้จากตารางที่ 5.18) เพียงเล็กน้อยหรือแทบจะไม่ต่างกันในชุดข้อมูลเกือบทั้งหมด มีเพียงชุดข้อมูล Insurance Company Benchmark และ SPECT Heart ที่มีความลึกโดยเฉลี่ยในการจำแนกกรณีตัวอย่างในอัลกอริทึม EVC-ID3 ลดลงจากอัลกอริทึม ID3 อย่างเห็นได้ชัดโดยลดลงถึง 13.22 เปอร์เซ็นต์และ 18.71 เปอร์เซ็นต์ตามลำดับซึ่งสอดคล้องกับการลดลงของเวลาในการทดสอบในอัลกอริทึม EVC-ID3 ในตารางที่ 5.31

สรุปผลการทดลองระหว่างอัลกอริทึม MII-ID3 และอัลกอริทึม EVC-ID3 พบว่าอัลกอริทึม MII-ID3 มีจำนวนความลึกสูงสุด, จำนวนกฎการตัดสินใจ, จำนวนโหนดทั้งหมด, เวลาใน

การฝึกฝน, เวลาในการทดสอบ และความลึกโดยเฉลี่ยในการจำแนกน้อยกว่าอัลกอริทึม ID3-A* ในชุดข้อมูลทั้งหมด โดยเฉพาะจำนวนกฎการตัดสินใจและจำนวนโหนดทั้งหมดของอัลกอริทึม EVC-ID3 มีจำนวนมากกว่าอัลกอริทึม ID3 แบบดั้งเดิมอย่างมากในชุดข้อมูลเกือบทั้งหมด ในขณะที่อัลกอริทึม MII-ID3 สามารถลดจำนวนกฎการตัดสินใจและจำนวนโหนดทั้งหมดลงได้น้อยกว่าอัลกอริทึม ID3 แบบดั้งเดิมอย่างชัดเจน อย่างไรก็ตามอัลกอริทึม EVC-ID3 มีความแม่นยำในการจำแนกมากกว่าอัลกอริทึม MII-ID3 อย่างชัดเจนในชุดข้อมูลส่วนมาก

5.8 ผลการทดลองระหว่างอัลกอริทึม MII-ID3 และอัลกอริทึม C4.5

อัลกอริทึม C4.5 เป็นอัลกอริทึมที่ถูกพัฒนาจากอัลกอริทึม ID3 เพื่อแก้ปัญหาในหลาย ๆ ด้านดังที่ได้อธิบายไว้ในบทที่ 2 การพัฒนาหลักคือใช้การคำนวณอัตราส่วนเกนเป็นเกณฑ์ในการเลือกแอตทริบิวต์เพื่อจัดการปัญหาความลำเอียงในการเลือกแอตทริบิวต์ของอัลกอริทึม ID3 และมีการทำ Prunning กับต้นไม้ตัดสินใจที่ได้เพื่อจัดการปัญหาความเข้มงวดในการสร้างกฎการตัดสินใจของอัลกอริทึม ID3

5.8.1 ความแม่นยำในการจำแนกและความลึกสูงสุดโดยเฉลี่ย

ตารางที่ 5.33 แสดงร้อยละความแม่นยำในการจำแนกระหว่าง MII-ID3 และ C4.5

ชุดข้อมูล	ร้อยละของความแม่นยำโดยเฉลี่ย	
	อัลกอริทึม MII-ID3	อัลกอริทึม C4.5
Connect-4	73.59	**77.76**
Phishing Websites	93.65	**95.91**
Insurance Company Benchmark	90.91	**94.03**
Molecular Biology	88.75	**92.86**
Soybean	86.21	**90.38**
Audiology	75.01	**82.14**
Balance Scale	64.38	**66.45**
Car Evaluation	83.42	**86.13**
Chess	94.82	**99.19**
Congressional Voting Records	95.39	**95.41**
Hayes-Roth	73.69	**79.01**
MONK's Problems	**86.96**	84.53
Mushroom	99.64	**100.00**
Nursery	89.16	**95.19**
SPECT Heart	79.99	**82.09**
Tic-Tac-Toe Endgame	77.68	**82.88**
Firm-Teacher_Clave-Direction_Classification	67.78	**74.13**

จากตารางที่ 5.33 แสดงให้เห็นว่าอัลกอริทึม C4.5 มีความแม่นยำในการจำแนกมากกว่าอัลกอริทึม MII-ID3 อย่างเห็นได้ชัดในชุดข้อมูลส่วนมาก มีเพียงชุดข้อมูล MONK's Problems เท่านั้นที่มีความแม่นยำในอัลกอริทึม MII-ID3 มากกว่าความแม่นยำในอัลกอริทึม C4.5 อัลกอริทึม C4.5 สามารถเพิ่มความแม่นยำในการจำแนกจากอัลกอริทึม ID3 แบบดั้งเดิมให้มากขึ้นในชุดข้อมูลส่วนมาก (ความแม่นยำของอัลกอริทึม ID3 แบบดั้งเดิมในทุก ๆ ชุดข้อมูลดูได้จากตารางที่ 5.6-5.8) ในขณะที่อัลกอริทึม MII-ID3 สามารถรักษาความแม่นยำให้ลดลงเพียงเล็กน้อยหรือดีกว่าอัลกอริทึม ID3 แค่ในชุดข้อมูลขนาดใหญ่ทุกชุดข้อมูลและขนาดกลางบางชุดข้อมูลเท่านั้น

ตารางที่ 5.34 แสดงความลึกสูงสุดโดยเฉลี่ยระหว่าง MII-ID3 และ C4.5

ชุดข้อมูล	ความลึกสูงสุดโดยเฉลี่ย	
	อัลกอริทึม MII-ID3	อัลกอริทึม C4.5
Connect-4	**15.73**	22
Phishing Websites	**10.94**	17
Insurance Company Benchmark	24.01	**0**
Molecular Biology	8.79	**8**
Soybean	8.08	**8**
Audiology	**5.80**	8
Balance Scale	3.78	**3**
Car Evaluation	**4.58**	5
Chess	**7.41**	12
Congressional Voting Records	4.35	**3**
Hayes-Roth	3.20	**3**
MONK's Problems	**3.90**	4
Mushroom	**2.76**	5
Nursery	**5.37**	7
SPECT Heart	7.29	**5**
Tic-Tac-Toe Endgame	5.16	**5**
Firm-Teacher_Clave-Direction_Classification	**10.97**	12

จากตารางที่ 5.34 แสดงให้เห็นว่าอัลกอริทึม MII-ID3 สามารถลดจำนวนความลึกสูงสุดได้มากกว่าอัลกอริทึม C4.5 ในชุดข้อมูล 9 ชุดจากทั้งหมด 17 ชุดข้อมูล และอัลกอริทึม C4.5 สามารถลดจำนวนความลึกสูงสุดได้มากกว่าอัลกอริทึม MII-ID3 ในอีก 8 ชุดข้อมูล ทั้งสองอัลกอริทึมสามารถลดจำนวนความลึกสูงสุดของต้นไม้ตัดสินใจผลลัพธ์ได้พอ ๆ กัน แต่อัลกอริทึม C4.5 มีความแม่นยำในการจำแนกที่มากกว่าอัลกอริทึม MII-ID3 ในชุดข้อมูลส่วนมากอย่างเห็นได้ชัด ชุดข้อมูล Insurance Company Benchmark มีความลึกสูงสุดในอัลกอริทึม C4.5 เป็น 0 เพราะมี 2 คลาสโดยมีกรณีตัวอย่างที่เป็นคลาสเสียงส่วนมากถึง 94.03 เปอร์เซ็นต์ ในขณะที่กรณีตัวอย่างในอีกคลาสมีแค่

5.97 เปอร์เซ็นต์ ดังนั้นเมื่ออัลกอริทึม C4.5 คำนวณค่าความผิดพลาดหลังทำ Pruning ที่โหนดรากจึงพบว่ามีค่าน้อยกว่าความผิดพลาดก่อนทำ Pruning ที่โหนดราก อัลกอริทึม C4.5 จึงทำการ Pruning ที่โหนดรากของต้นไม้ตัดสินใจ หมายความว่ากรณีตัวอย่างในชุดข้อมูลสำหรับทดสอบของ Insurance Company Benchmark จะถูกจำแนกประเภทเป็นคลาสเสียงส่วนมากทั้งหมดในโหนดรากที่ถูก Pruning ให้เป็นโหนดคำตอบ ความแม่นยำที่ได้จากการ Pruning จนเหลือเพียงแค่โหนดรากของอัลกอริทึม C4.5 มีความแม่นยำมากกว่าอัลกอรึม ID3 แบบดั้งเดิมและอัลกอริทึม MII-ID3 ในชุดข้อมูล Insurance Company Benchmark อีกด้วย

5.8.2 จำนวนกฎการตัดสินใจโดยเฉลี่ย

ตารางที่ 5.35 แสดงจำนวนกฎการตัดสินใจโดยเฉลี่ยระหว่าง MII-ID3 และ C4.5

ชุดข้อมูล	จำนวนกฎการตัดสินใจโดยเฉลี่ย	
	อัลกอริทึม MII-ID3	อัลกอริทึม C4.5
Connect-4	4646.37	**2493**
Phishing Websites	**100.77**	126
Insurance Company Benchmark	3421.84	**1**
Molecular Biology	288.36	**129**
Soybean	75.63	**50**
Audiology	40.89	**19**
Balance Scale	97.82	**25**
Car Evaluation	**50.34**	62
Chess	**14.17**	28
Congressional Voting Records	7.42	**7**
Hayes-Roth	23.47	**19**
MONK's Problems	35.84	**19**
Mushroom	**22.80**	24
Nursery	**71.88**	213
SPECT Heart	15.48	**6**
Tic-Tac-Toe Endgame	**48.86**	49
Firm-Teacher_Clave-Direction_Classification	**120.02**	526

จากตารางที่ 5.35 แสดงให้เห็นว่าอัลกอริทึม C4.5 มีจำนวนกฎการตัดสินใจของต้นไม้ตัดสินใจผลลัพธ์ที่น้อยกว่าอัลกอริทึม MII-ID3 อย่างเห็นได้ชัดในชุดข้อมูลส่วนมาก ดังนั้นอัลกอริทึม C4.5 จึงสามารถลดจำนวนกฎการตัดสินใจจากอัลกอริทึม ID3 แบบดั้งเดิมได้ดีกว่าอัลกอริทึม MII-ID3 (จำนวนกฎการตัดสินใจของอัลกอริทึม ID3 แบบดั้งเดิมในทุกชุดข้อมูลดูได้จากตารางที่ 5.13-5.14)

5.8.3 จำนวนโหนดทั้งหมดในต้นไม้ตัดสินใจผลลัพธ์โดยเฉลี่ย

ตารางที่ 5.36 แสดงจำนวนโหนดทั้งหมดโดยเฉลี่ยระหว่าง MII-ID3 และ C4.5

ชุดข้อมูล	จำนวนโหนดทั้งหมดโดยเฉลี่ย	
	อัลกอริทึม MII-ID3	อัลกอริทึม C4.5
Connect-4	6969.06	**3739**
Phishing Websites	**169.28**	218
Insurance Company Benchmark	3908.88	**1**
Molecular Biology	362.11	**160**
Soybean	136.14	**84**
Audiology	109.03	**72**
Car Evaluation	61.67	**30**
Chess	**68.86**	86
Congressional Voting Records	13.84	**13**
Mushroom	**25.80**	29
Nursery	31.05	**25**
SPECT Heart	53.25	**28**
Firm-Teacher_Clave-Direction_Classification	**25.62**	29
Balance Scale	**102.59**	312
Hayes-Roth	29.97	**11**
MONK's Problems	**72.79**	73
Tic-Tac-Toe Endgame	**239.04**	1051

จากตารางที่ 5.36 แสดงให้เห็นว่าอัลกอริทึม C4.5 มีจำนวนโหนดทั้งหมดของต้นไม้ตัดสินใจผลลัพธ์ที่น้อยกว่าอัลกอริทึม MII-ID3 อย่างเห็นได้ชัดในชุดข้อมูลส่วนมาก ซึ่งเป็นไปในทางเดียวกันกับจำนวนกฎการตัดสินใจในตารางที่ 5.35 ดังนั้นอัลกอริทึม C4.5 จึงสามารถลดจำนวนโหนดทั้งหมดจากอัลกอริทึม ID3 แบบดั้งเดิมได้ดีกว่าอัลกอริทึม MII-ID3 (จำนวนโหนดทั้งหมดของอัลกอริทึม ID3 แบบดั้งเดิมในทุกชุดข้อมูลดูได้จากตารางที่ 5.15-5.16)

5.9 ผลการทดลองระหว่างอัลกอริทึม MII-ID3 และอัลกอริทึมนาอีฟเบย์

เนื่องจากนาอีฟเบย์นั้นมีลักษณะในการฝึกฝนและการทดสอบที่แตกต่างกับอัลกอริทึม ID3 แบบดั้งเดิม ซึ่งอัลกอริทึมนาอีฟเบย์จะจำแนกประเภทข้อมูลโดยการคำนวณความน่าจะเป็นที่อยู่บนพื้นฐานของทฤษฎีของเบย์โดยปราศจากการสร้างต้นไม้ตัดสินใจสำหรับจำแนกดังเช่นในอัลกอริทึม ID3, MII-ID3, ID3-A*, EVC-ID3 และ C4.5 ดังนั้นเกณฑ์ในการวัดที่จะเปรียบเทียบระหว่างอัลกอริทึม MII-ID3 และอัลกอริทึมนาอีฟเบย์ได้นั้นจึงมีเพียงแค่ความแม่นยำโดยเฉลี่ยในการจำแนกเท่านั้น

ตารางที่ 5.37 แสดงร้อยละความแม่นยำในการจำแนกระหว่าง MII-ID3 และนาอีฟเบย์

ชุดข้อมูล	ร้อยละของความแม่นยำโดยเฉลี่ย	
	อัลกอริทึม MII-ID3	อัลกอริทึม นาอีฟเบย์
Connect-4	**73.59**	72.13
Phishing Websites	**93.65**	92.95
Insurance Company Benchmark	**90.91**	80.98
Molecular Biology	88.75	**94.61**
Soybean	**86.21**	84.84
Audiology	**75.01**	54.70
Car Evaluation	83.42	**86.13**
Chess	**94.82**	88.93
Congressional Voting Records	**95.39**	91.28
Mushroom	**99.64**	99.56
Nursery	89.16	**90.37**
SPECT Heart	79.99	**85.07**
Firm-Teacher_Clave-Direction_Classification	67.78	**77.59**
Balance Scale	64.38	**89.46**
Hayes-Roth	73.69	**82.72**
MONK's Problems	**86.96**	72.66
Tic-Tac-Toe Endgame	**77.68**	69.73

จากผลการทดลองในตารางที่ 5.37 แสดงให้เห็นว่าอัลกอริทึม MII-ID3 มีความแม่นยำในการจำแนกมากกว่าอัลกอริทึมนาอีฟเบย์ในชุดข้อมูล 10 ชุดใน 17 ชุดข้อมูลทั้งหมด ในขณะที่อัลกอริทึมนาอีฟเบย์มีความแม่นยำในการจำแนกมากกว่าอัลกอริทึม MII-ID3 ในอีก 7 ชุดข้อมูล อัลกอรีทึม MII-ID3 มีความแม่นยำในการจำแนกมากกว่าอัลกอริทึมนาอีฟเบย์ในชุดข้อมูล Connect-4, Phishing Websites, Insurance Company Benchmark, Soybean, Audiology, Chess, Congressional Voting Records, Mushroom, MONK's Problems และ Tic-Tac-Toe Endgame ในส่วนของอัลกอริทึมนาอีฟเบย์มีความแม่นยำในการจำแนกมากกว่าอัลกอริทึม MII-ID3 ในชุดข้อมูล Nursery,

Molecular Biology, Car Evaluation, SPECT Heart, Balance Scale, Firm-Teacher_Clave-Direction_Classification และ Hayes-Roth ความแม่นยำในการจำแนกของอัลกอริทึม MII-ID3 จะมีค่าเพิ่มขึ้นมากหรือลดลงมากจากอัลกอริทึม ID3 แบบดั้งเดิมในแต่ละชุดข้อมูลนั้นขึ้นอยู่กับปัจจัยที่ได้กล่าวไว้ในผลการทดลองระหว่างอัลกอริทึม MII-ID3 และอัลกอริทึม ID3 แล้ว ซึ่งจะแตกต่างกับอัลกอริทึมนาอีฟเบย์เพราะอัลกอริทึมนาอีฟเบย์นั้นจะใช้การคำนวณความน่าจะเป็นในการจำแนกประเภท ดังนั้นถ้าหากชุดข้อมูลใด ๆ มีการกระจายของค่าที่ไม่ซ้ำกันในทุก ๆ แอตทริบิวต์และมีการกระจายของจำนวนคลาสเท่ากันในแต่ละค่าที่ไม่ซ้ำกันในชุดข้อมูลสำหรับทดสอบและชุดข้อมูลสำหรับฝึกฝนก็จะส่งผลให้ชุดข้อมูลนั้น ๆ มีความแม่นยำในการจำแนกในอัลกอริทึมนาอีฟเบย์สูง ซึ่งในการแบ่งข้อมูลในการจำแนกชุดข้อมูลที่มีลักษณะดังกล่าวมีน้อยมาก ๆ ในเมื่อการกระจายในแต่ละค่าที่ไม่ซ้ำกันของแอตทริบิวต์ทั้งในชุดข้อมูลสำหรับฝึกฝนและทดสอบไม่เป็นไปในอัตราส่วนที่เท่ากันหรือใกล้เคียงกันนั้น ความแม่นยำในอัลกอริทึมนาอีฟเบย์จะขึ้นอยู่กับว่าถ้าหากมีกรณีตัวอย่างที่มีค่าค่าหนึ่งในแอตทริบิวต์ใด ๆ ในชุดข้อมูลสำหรับฝึกฝนน้อยมาก ๆ หรือไม่มีเลย แต่ในชุดข้อมูลสำหรับทดสอบกลับมีกรณีตัวอย่างที่มีค่านั้นหลายกรณีก็จะส่งผลให้ความแม่นยำต่ำอย่างมาก แต่ถ้าในชุดข้อมูลสำหรับทดสอบไม่มีหรือมีกรณีตัวอย่างที่มีค่านั้นน้อยมาก ๆ เหมือนกับในชุดข้อมูลสำหรับฝึกฝนความแม่นยำก็อาจยังอยู่ในเกณฑ์ที่ดี สรุปได้ว่าอัลกอริทึม MII-ID3 มีความแม่นยำในการจำแนกมากกว่าอัลกอริทึมนาอีฟเบย์ในชุดข้อมูลส่วนมาก

5.10 การทดลองเพื่อปรับปรุงอัลกอริทึม MII-ID3 ด้วยการ Normalization ตัวแปร CurDepth และ RemAtts เพื่อความเท่าเทียมในชุดข้อมูลทั้งหมด

เนื่องจากอัลกอริทึม MII-ID3 ใช้ตัวแปร CurDepth (ความลึกปัจจุบัน) และ RemAtts (จำนวนแอตทริบิวต์ที่เหลืออยู่) มาคำนวณหาค่า Prop ซึ่งเป็นค่า Threshold ของความน่าจะเป็นในการละเลยกรณีตัวอย่างส่วนน้อย ซึ่งชุดข้อมูลที่มีขนาดเล็กจะมีความลึกปัจจุบันและจำนวนแอตทริบิวต์ที่เหลืออยู่ที่เป็นไปได้น้อยกว่าชุดข้อมูลขนาดใหญ่ จึงอาจเกิดความไม่เท่าเทียมในการละเลยกรณีตัวอย่างส่วนน้อย ดังนั้นก่อนจะนำค่า RemAtts ไปคำนวณจะทำการ Normalization ค่า RemAtts ให้อยู่ระหว่าง 0-1 ก่อนเพื่อความเท่าเทียมในแต่ละชุดข้อมูล ส่วนตัวแปร CurDepth นั้นไม่สามารถ Normalization ได้เนื่องจากการทำ Normalization จะต้องรู้ค่าสูงสุดและต่ำสุดของความลึกทั้งหมดของต้นไม้ตัดสินใจ ซึ่งจะรู้ค่าสูงสุดของความลึกได้จะต้องสร้างต้นไม้ตัดสินใจให้เสร็จก่อน ดังนั้นตัวแปร CurDepth เราจะหารด้วย 10 ก่อนนำไปคำนวณค่า Prop โดยจะเรียกการปรับปรุงอัลกอริทึม MII-ID3 ด้วยการ Normalization ตัวแปร CurDepth และ RemAtts ว่า “MII-ID3-N” เพื่อความสะดวกในการเขียนผลการทดลอง

ตารางที่ 5.38 แสดงร้อยละความแม่นยำในการจำแนกระหว่าง MII-ID3 และ MII-ID3-N

กลุ่ม	ชุดข้อมูล	ร้อยละของความแม่นยำโดยเฉลี่ย	
		อัลกอริทึม MII-ID3	อัลกอริทึม MII-ID3-N
ขนาดใหญ่	Connect-4	**73.59**	71.52
	Phishing Websites	**93.65**	92.74
	Insurance Company Benchmark	90.91	**92.71**
	Molecular Biology	**88.75**	87.54
ขนาดกลาง	Soybean	**86.21**	83.78
	Audiology	**75.01**	72.09
	Car Evaluation	83.42	**85.96**
	Chess	**94.82**	91.43
	Congressional Voting Records	95.39	**95.60**
	Mushroom	**99.64**	99.51
	Nursery	89.16	**89.71**
	SPECT Heart	**79.99**	79.64
	Firm-Teacher_Clave-Direction_Classification	**67.78**	66.26
ขนาดเล็ก	Balance Scale	**64.38**	63.46
	Hayes-Roth	73.69	**76.50**
	MONK's Problems	86.96	**90.20**
	Tic-Tac-Toe Endgame	77.68	**77.89**

จากผลการทดลองในตารางที่ 5.38 แสดงให้เห็นว่าอัลกอริทึม MII-ID3 มีความแม่นยำมากกว่า MII-ID3-N 10 ชุดข้อมูลจากทั้งหมด 17 ชุดข้อมูล ในขณะที่ MII-ID3-N มีความแม่นยำมากกว่าอัลกอริทึม MII-ID3 ในอีก 7 ชุดข้อมูล ดังนั้นอัลกอริทึม MII-ID3 มีความแม่นยำมากกว่าอัลกอริทึม MII-ID3-N อย่างชัดเจนในชุดข้อมูลส่วนมาก เพราะการ Normalize ตัวแปร RemAtts ให้มีค่าน้อยลงและอยู่ในระหว่าง 0-1 ของอัลกอริทึม MII-ID3-N ทำให้โอกาสในการละเลยกรณีตัวอย่างส่วนน้อยมีมากขึ้นซึ่งมีความเสี่ยงในการลดลงของความแม่นยำในชุดข้อมูลส่วนมาก ผลการทดลองแสดงให้เห็นแล้วว่าการไม่ทำ Normalization ในตัวแปร RemAtts และ CurDepth ของอัลกอริทึม MII-ID3 ให้ผลลัพธ์ในด้านความแม่นยำที่ดีกว่าอัลกอริทึม MII-ID3-N ในชุดข้อมูลส่วนมาก

ตารางที่ 5.39 แสดงความลึกสูงสุดโดยเฉลี่ยระหว่าง MII-ID3 และ MII-ID3-N

ชุดข้อมูล	ความลึกสูงสุดโดยเฉลี่ย	
	อัลกอริทึม MII-ID3	อัลกอริทึม MII-ID3-N
Connect-4	15.73	**11.25**
Phishing Websites	10.94	**6.13**
Insurance Company Benchmark	24.01	**7.28**
Molecular Biology	8.79	**5.62**
Soybean	8.08	**5.61**
Audiology	5.80	**4.92**
Car Evaluation	**4.58**	5.19
Chess	7.41	**4.23**
Congressional Voting Records	4.35	**3.65**
Mushroom	2.76	**2.28**
Nursery	**5.37**	5.76
SPECT Heart	7.29	**5.13**
Firm-Teacher_Clave-Direction_Classification	10.97	**9.48**
Balance Scale	**3.78**	4.00
Hayes-Roth	**3.20**	3.76
MONK's Problems	**3.90**	3.98
Tic-Tac-Toe Endgame	**5.16**	5.28

จากผลการทดลองในตารางที่ 5.39 แสดงให้เห็นว่าอัลกอริทึม MII-ID3-N สามารถลดจำนวนความลึกสูงสุดได้มากกว่าอัลกอริทึม MII-ID3 ในชุดข้อมูลส่วนมาก โดยอัลกอริทึม MII-ID3-N มีความลึกสูงสุดน้อยกว่าอัลกอริทึม MII-ID3 11 ชุดข้อมูลจากทั้งหมด 17 ชุดข้อมูล และอัลกอริทึม MII-ID3 มีความลึกสูงสุดน้อยกว่าอัลกอริทึม MII-ID3-N ในอีก 6 ชุดข้อมูล

ตารางที่ 5.40 แสดงจำนวนกฎการตัดสินใจโดยเฉลี่ยระหว่าง MII-ID3 และ MII-ID3-N

ชุดข้อมูล	จำนวนกฎการตัดสินใจโดยเฉลี่ย	
	อัลกอริทึม MII-ID3	อัลกอริทึม MII-ID3-N
Connect-4	4646.37	**668.48**
Phishing Websites	100.77	**39.43**
Insurance Company Benchmark	3421.84	**915.73**
Molecular Biology	288.36	**133.77**
Soybean	75.63	**53.85**
Audiology	40.89	**32.29**
Car Evaluation	**50.34**	75.71
Chess	14.17	**6.65**
Congressional Voting Records	7.42	**6.37**
Mushroom	22.80	**18.87**
Nursery	**71.88**	86.05
SPECT Heart	15.48	**10.12**
Firm-Teacher_Clave-Direction_Classification	120.02	**63.38**
Balance Scale	**97.82**	152.58
Hayes-Roth	**23.47**	28.14
MONK's Problems	**35.84**	41.63
Tic-Tac-Toe Endgame	**48.86**	50.87

จากผลการทดลองในตารางที่ 5.40 แสดงให้เห็นว่าอัลกอริทึม MII-ID3-N สามารถลดจำนวนกฎการตัดสินใจได้มากกว่าอัลกอริทึม MII-ID3 ในชุดข้อมูลส่วนมาก โดยอัลกอริทึม MII-ID3-N มีจำนวนกฎการตัดสินใจน้อยกว่าอัลกอริทึม MII-ID3 11 ชุดข้อมูลจากทั้งหมด 17 ชุดข้อมูล และอัลกอริทึม MII-ID3 มีจำนวนกฎการตัดสินใจน้อยกว่าอัลกอริทึม MII-ID3-N ในอีก 6 ชุดข้อมูล เช่นเดียวกันกับผลการทดลองในตารางที่ 5.39

บทที่ 6
บทสรุปและข้อเสนอแนะ

6.1 สรุป

อัลกอริทึม ID3 เป็นหนึ่งในอัลกอริทึมที่ใช้สร้างต้นไม้ตัดสินใจที่ใช้กันอย่างแพร่หลายในทางการจำแนกประเภทข้อมูล หนึ่งในข้อเสียของอัลกอริทึม ID3 คือมีความเข้มงวดในการสร้างกฎการตัดสินใจ จำนวนกฎการตัดสินใจที่มากเกินไปไม่เป็นผลดีต่อการทดสอบชุดข้อมูลส่วนที่ไม่เคยพบ ความแม่นยำในการจำแนกอาจจะต่ำลงได้ และยังต้องใช้พื้นที่ในการเก็บต้นไม้ตัดสินใจในหน่วยความจำมากขึ้นอีกด้วย ซึ่งถ้าชุดข้อมูลสำหรับเรียนรู้มีจำนวนแอตทริบิวต์หรือกรณีตัวอย่างที่มากเกินไป ต้นไม้ตัดสินใจที่ได้อาจจะมีจำนวนกฎการตัดสินใจที่มากเกินจำเป็น บางส่วนของกฎการตัดสินใจเหล่านี้อาจจะมีบางโหนดที่มีจำนวนกรณีตัวอย่างที่น้อยมาก ๆ แนวคิดของงานวิจัยนี้คือถ้าเราตัดกรณีตัวอย่างส่วนน้อยนี้ออกไปความแม่นยำในการจำแนกประเภทจะลดลงเพียงเล็กน้อย ทั้งยังเป็นการลดจำนวนกฎการตัดสินใจและจำนวนความลึกสูงสุดอีกด้วย ดังนั้นงานวิจัยนี้จึงพัฒนาอัลกอริทึม MII-ID3 ขึ้นเพื่อแก้ปัญหาความเข้มงวดในการสร้างกฎการตัดสินใจในอัลกอริทึม ID3 โดยการละเลยกรณีตัวอย่างส่วนน้อย วัตถุประสงค์ของอัลกอริทึม MII-ID3 คือเพื่อลดความเข้มงวดในการสร้างกฎการตัดสินใจและความลึกสูงสุดของอัลกอริทึม ID3 แบบดั้งเดิม ในขณะที่ความแม่นยำลดลงเพียงเล็กน้อย

จากผลการเปรียบเทียบประสิทธิภาพระหว่างอัลกอริทึม MII-ID3, อัลกอริทึม ID3, อัลกอริทึม ID3-A* และอัลกอริทึม EVC-ID3 พบว่าอัลกอริทึม MII-ID3 มีข้อดีคือสามารถสร้างต้นไม้ตัดสินใจผลลัพธ์ที่มีจำนวนความลึกสูงสุด จำนวนกฎการตัดสินใจ จำนวนโหนดทั้งหมด เวลาในการฝึกฝนและเวลาในการทดสอบน้อยกว่าอัลกอริทึม ID3, อัลกอริทึม ID3-A* และอัลกอริทึม EVC-ID3 ได้อย่างเห็นได้ชัดในชุดข้อมูลส่วนมาก อย่างไรก็ตามอัลกอริทึม MII-ID3 มีข้อเสียคือไม่สามารถรักษาความแม่นยำในการจำแนกจากอัลกอริทึม ID3 แบบดั้งเดิมให้ลดลงเพียงเล็กน้อยหรือมากขึ้นในชุดข้อมูลทั้งหมด อัลกอริทึม MII-ID3 สามารถรักษาความแม่นยำในการจำแนกจากอัลกอริทึม ID3 แบบดั้งเดิมให้ลดลงเพียงเล็กน้อยหรือมากขึ้นในชุดข้อมูลขนาดใหญ่ทั้งหมดและชุดข้อมูลขนาดกลางบางชุดข้อมูลเท่านั้น ในชุดข้อมูลขนาดเล็กส่วนใหญ่และชุดขนาดกลางบางชุดข้อมูลมีความแม่นยำในอัลกอริทึม MII-ID3 ที่น้อยกว่าความแม่นยำในอัลกอริทึม ID3, อัลกอริทึม ID3-A* และอัลกอริทึม EVC-ID3 อย่างชัดเจน เนื่องจากอัลกอริทึม MII-ID3 จะละเลยกรณีตัวอย่างส่วนน้อยโดยขึ้นอยู่กับ Threshold ของความน่าจะเป็นหรือค่า Prop ในชุดข้อมูลทั้งหมดระหว่างสร้างต้นไม้ตัดสินใจในอัลกอริทึม MII-ID3 จะมีเหตุการณ์ที่ละเลยกรณีตัวอย่างส่วนน้อยในขณะที่มีค่า Prop และ Remainder มาก ซึ่งเมื่อเกิดเหตุการณ์นี้ขึ้น ไม่ควรจะละเลยกรณีตัวอย่างส่วนน้อยเพราะจะส่งผลให้ความแม่นยำในชุดข้อมูลที่มี

ขนาดเล็กมีความเสี่ยงที่ความแม่นยำในการจำแนกจะลดลงอย่างชัดเจนสูงกว่าชุดข้อมูลขนาดใหญ่ และนี้คือข้อจำกัดของอัลกอริทึม MII-ID3

จากการเปรียบเทียบประสิทธิภาพระหว่างอัลกอริทึม MII-ID3 และอัลกอริทึม C4.5 พบว่า อัลกอริทึม C4.5 มีจำนวนกฎการตัดสินใจและจำนวนโหนดทั้งหมดของต้นไม้ตัดสินใจผลลัพธ์น้อยกว่า อัลกอริทึม MII-ID3 ในชุดข้อมูลส่วนมาก ส่วนการเปรียบเทียบความลึกสูงสุดทั้งสองอัลกอริทึม สามารถลดจำนวนความลึกสูงสุดของต้นไม้ตัดสินใจผลลัพธ์ได้พอ ๆ กัน แต่อัลกอริทึม C4.5 มีความแม่นยำในการจำแนกที่มากกว่าอัลกอริทึม MII-ID3 ในชุดข้อมูลส่วนมากอย่างเห็นได้ชัด

ส่วนการเปรียบเทียบความแม่นยำในการจำแนกระหว่างอัลกอริทึม MII-ID3 และอัลกอริทึมนาอีฟเบย์พบว่า อัลกอริทึม MII-ID3 มีความแม่นยำมากกว่าอัลกอริทึมนาอีฟเบย์ในชุดข้อมูลส่วนมาก ดังนั้นสรุปได้ว่าอัลกอริทึม MII-ID3 เหมาะสมกับชุดข้อมูลขนาดใหญ่ที่มีจำนวนกรณีตัวอย่างตั้งแต่ 3000 กรณีขึ้นไปและมีจำนวนแอตทริบิวต์ที่ถูกใช้ในการสร้างต้นไม้ตัดสินใจในอัลกอริทึม ID3 แบบดั้งเดิมตั้งแต่ 30 แอตทริบิวต์ขึ้นไป เนื่องจากผลการทดลองแสดงให้เห็นว่าอัลกอริทึม MII-ID3 สามารถลดความเข้มงวดในการสร้างกฎการตัดสินใจและความลึกสูงสุดของอัลกอริทึม ID3 แบบดั้งเดิม และสามารถรักษาความแม่นยำในการจำแนกให้ลดลงเพียงเล็กน้อยหรือมากขึ้นได้ในชุดข้อมูลส่วนมาก ซึ่งเป็นชุดข้อมูลขนาดใหญ่ทั้งหมด ชุดข้อมูลขนาดกลางบางตัว และชุดข้อมูลขนาดเล็กเป็นส่วนน้อย

6.2 ข้อเสนอแนะ

การปรับปรุงอัลกอริทึม MII-ID3 ให้สามารถใช้ได้กับชุดข้อมูลขนาดเล็ก และชุดข้อมูลขนาดกลางบางตัวแล้วความแม่นยำในการจำแนกลดลงเพียงเล็กน้อยนั้น อาจลองเปลี่ยนจากการละเลยกรณีตัวอย่างส่วนน้อยที่ขึ้นอยู่กับ Threshold ของความน่าจะเป็น ให้เป็นการละเลยกรณีตัวอย่างส่วนน้อยโดยการกำหนดพารามิเตอร์ในการละเลยกรณีตัวอย่างส่วนน้อยโดยตรง เพื่อหลีกเลี่ยงเหตุการณ์ที่เกิดการละเลยกรณีตัวอย่างส่วนน้อยในขณะที่โหนดปัจจุบันมีค่า Prop และ Remainder มาก ซึ่งนำไปสู่การลดลงของความแม่นยำในการจำแนก จากเดิมที่อัลกอริทึม MII-ID3 ทำการคำนวณค่า Prop ในโหนดปัจจุบัน แล้วสุ่มตัวเลขทศนิยมระหว่าง 0-1 ขึ้นมา ถ้าตัวเลขทศนิยมที่สุ่มได้มีค่ามากกว่าค่า Prop ของโหนดปัจจุบัน จะทำการละเลยกรณีตัวอย่างส่วนน้อยในโหนดลูกของโหนดปัจจุบันทั้งหมด เพื่อปรับปรุงอัลกอริทึม MII-ID3 ให้มีความแม่นยำที่ดีในชุดข้อมูลขนาดกลางและขนาดเล็ก วิธีที่แก้ไขได้คืออาจยกเลิกการใช้การคำนวณค่า Prop และการสุ่มเป็นตัวกำหนดการละเลยกรณีตัวอย่างส่วนน้อยในโหนดลูกของโหนดปัจจุบัน แต่ให้พิจารณาที่โหนดลูกที่จะละเลยกรณีตัวอย่างส่วนน้อยเลยว่าในโหนดลูกตัวนี้มีกรณีตัวอย่างส่วนน้อยอยู่จริงหรือไม่ ซึ่งเงื่อนไขการตัดสินใจอาจกำหนดเป็นพารามิเตอร์เช่น ถ้าจำนวนกรณีตัวอย่างในคลาสที่มากสุดของโหนดใดห่างจากจำนวนกรณีตัวอย่างในคลาสอื่นของโหนดนั้นรวมกันแล้วห่างมากกว่า 90 เปอร์เซ็นต์ ถือว่าในโหนดมีกรณี

ตัวอย่างส่วนน้อย ให้ละเลยกรณีตัวอย่างส่วนน้อยในโหนดนั้น ซึ่งอาจปรับค่าพารามิเตอร์จาก 90 เปอร์เซ็นต์ เป็น 80 เปอร์เซ็นต์ หรืออื่น ๆ เพื่อหาค่าพารามิเตอร์ที่เหมาะสมที่สุดสำหรับชุดข้อมูลทุกชุด ซึ่งการกำหนดพารามิเตอร์ในการละเลยกรณีตัวอย่างส่วนน้อยของโหนดใดโดยตรง สามารถลดความเสี่ยงในการลดลงของความแม่นยำได้มากอย่างแน่นอน เพราะเป็นการพิจารณาที่โหนดที่จะละเลยกรณีตัวอย่างโดยตรง ว่าโหนดนั้นมีกรณีตัวอย่างส่วนน้อยตามพารามิเตอร์ที่ตั้งไว้จริงหรือไม่ ถ้าไม่มีตามเงื่อนไขของพารามิเตอร์ที่ตั้งไว้ จะไม่เกิดการละเลยกรณีตัวอย่างส่วนน้อยอย่างแน่นอน ซึ่งต่างจากการละเลยกรณีตัวอย่างส่วนน้อยโดยใช้ค่า Threshold ของความน่าจะเป็นแบบเดิม เพราะแม้จะมีค่า Threshold ของความน่าจะเป็นในการละเลยกรณีตัวอย่างส่วนน้อยที่มาก อย่างไรก็ตามมันยังมีโอกาสเกิดขึ้นอยู่และเกิดเหตุการณ์นี้ขึ้นในชุดข้อมูลทั้งหมด ส่งผลให้ความแม่นยำในอัลกอริทึม MII-ID3 ของชุดข้อมูลบางตัวลดลงจากอัลกอริทึม ID3 อย่างเห็นได้ชัด

เอกสารอ้างอิง

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

[2] Z. Wang, Y. Liu, and L. Liu. "A new way to choose splitting attribute in ID3 algorithm" IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conf., Chengdu, China, 2017. pp. 659-663.

[3] C. Guan and X. Zeng. "An Improved ID3 Base on Weighted Modified Information gain" 7th Int. Conf. Computational Intelligence and Security, Hainan, China, 2011. pp. 1283-1285.

[4] N. Kaewrod and K. Jearanaitanakij. "Improving ID3 Algorithm by Using A* Search" 21st Int. Computer Science and Engineering Conf., Bangkok, Thailand, 2017. pp. 237-240.

[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.

[7] S. Russell and P. Norvig. Artificial Intelligence: A Modern Approach. 3rd ED. Essex, England : Pearson. 2014.

[8] J.R. Quinlan. C4.5 : Programs for Machine Learning. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc. 1993.

ภาคผนวก ก.

งานวิจัยที่ได้รับการตีพิมพ์

22nd International Computer Science and Engineering Conference (ICSEC2018)

Improving ID3 Algorithm by Ignoring Minor Instances

Nicha Kaewrod, Kietikul Jearanaitanakij
Department of Computer Engineering, Faculty of Engineering
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
Email: 59601087@kmitl.ac.th, kietikul.je@kmitl.ac.th

Abstract—**Among various classification algorithms, ID3 is one of the most widely used and well-known tools that generates an efficient decision tree. Nevertheless, ID3 is too rigorous in generating the decision rules. As a result, the final decision tree may carry too many decision rules. Some of these decision rules may have very low number of instances which do not make significant change to the classification accuracy. The aim of this paper is to propose an approach to relax the rigorousness of the conventional ID3 algorithm by ignoring minor instances so that the resulting decision tree will have the lower number of depths yet produce promising accuracy. The proposed algorithm is examined on six datasets from UCI repository and Weka. The experimental results indicate that the proposed algorithm not only significantly reduces the maximum number of depths of the decision tree, but also retains the classification accuracy in the satisfying level. Moreover, the training time, the classification time, and the number of decision rules of the proposed algorithm are lower than those of the conventional ID3.**

Keywords—Decision tree; ID3 algorithm; Decision rule; Classification

I. INTRODUCTION

ID3 algorithm, proposed by Quinlan in 1986 [1], was developed for solving the classification problem. It is one of the most popular and simple algorithms in decision tree learning. The conventional ID3 algorithm keeps generating a node and splitting the training instances until there is no more instance left. The attribute selection criterion is choosing the largest value of the information gain among the remaining attributes. However, the bias on the many-value attribute is an obvious drawback of the information gain measure, i.e., the conventional ID3 tends to favor the attribute which has many values [1]. Consequently, Quinlan invented the gain ratio [1] to handle this disadvantage. Although gain ratio increases a chance of selecting the few-value attribute, the many-value attribute still has the tendency to be selected if it has the number of remaining instances which has the same attribute value approaching the number of remaining instances [2]. There have been some attempts to address the attribute selection in ID3 algorithm. Wang and Yu Liu and Lu Liu [3] applied rough set theory to calculate the consistency of attributes. This method chooses an attribute that has the highest consistency. The experimental results indicated that it can improve the classification accuracy, but it is appropriate for the problem which have only two classes. Guan and Zeng [4] proposed a new way to choose an attribute of the ID3 algorithm based on Weighted Modified Information gain for increasing the accuracy. However, the generated decision tree was more complex than that of the conventional ID3. Wang and Jiang [2] calculated an average gain inspired by the gain ratio's weakness. They compared the experimental results between the average gain and the gain ratio using 36 UCI datasets. The average gain was able to maintain the high accuracy, while the training time and the size of the decision tree are less than those of the gain ratio. Kraidech and Jearanaitanakij [5] modified ID3 algorithm by combining values between two equally important attributes into a single node of the decision tree. Although this method can significantly decrease the maximum number of depths from the conventional ID3 and preserve the classification accuracy, the number of decision rules at the combining node significantly increases. Therefore, the resulting decision tree is very rigorous, i.e., it contains too many decision rules.

We propose the algorithm to relax the rigorousness of the conventional ID3 algorithm by ignoring minor instances. Consequently, the proposed algorithm can reduce the number of decision rules and the maximum depth of the decision tree while the accuracy rate of classification is still preserved. Moreover, the proposed algorithm can reduce both the training time and the testing time in most datasets. The proposed algorithm is examined by six datasets. There are five datasets from the UCI repository [6], e.g., Connect-4, Phishing Websites, Insurance Company Benchmark (COIL 2000), Molecular Biology (Splice-junction Gene Sequences) and Soybean (Large), and Supermarket dataset from Weka [7].

The rest of this paper is organized as the following sections. Section II describes the concept of the conventional ID3 algorithm. Section III illustrates the problem when the conventional ID3 produces the rigorous decision tree. Section IV presents the proposed algorithm along with the pseudocode. In section V, we explain the datasets and the analysis of the experimental results. Finally, section VI summarizes the performance of the proposed algorithm.

II. THE CONVENTIONAL ID3 ALGORITHM

The ID3 algorithm is a well-known classification tool that uses the training instances of the dataset to generate a decision tree. The classification performance of the resulting decision tree is evaluated by using a set of unseen instances. The basic concept of ID3 algorithm is briefed below [8].

Step 1: Use Eq. (1) and (2) to compute the information gain of every attribute.

$$IG(I,A) = -\sum_{d \in D} p(d)\log_2 p(d) - R(A) \quad (1)$$

Where

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

I is a set of remaining instances,

D is a set of target classes in *I*,

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

p(d) is the proportion of the number of instances in class *d* to the total number of instances in *I*.

$$R(A) = \sum_{v \in V} - p(v) \sum_{x \in D} p(x) \log_2 p(x) \qquad (2)$$

Where

V is a set of class values in the attribute *A*,

p(v) is the proportion of the total value *v* in *A* to the total elements in *I*,

p(x) is the proportion of the number of instances in class *x*, which have value *v*, to the total number of instances in *A*, which its value is *v*.

Step 2: Select the attribute which has the maximum information gain to build the current node. In case that there are at least two attributes whose information gains are the greatest value, randomly choose one of those attributes as the current decision node.

Step 3: Classify the remaining training instances by the attribute chosen in step 2.

Step 4: Repeat steps 1-3 on the remaining training instances and attributes until all training instances are classified. The following Fig.1 shows the pseudocode of the conventional ID3 which is simplified from the version presented in [8].

```
1   function ID3 (instances, parent_instances, attributes) returns a tree
      if instances is empty,
        then return the leaf node which labeled by the majority of parent_instances
      else if all instances have the same classification,
        then return the leaf node which labeled by its classification
6     else if attributes is empty,
        then return the leaf node which labeled by the majority of instances
      else
        Compute information gains for all attributes using instances and attributes
        A ← Select the attribute which has the best information gain
11      Insert the new node labeled by attribute A to tree
12      for each value v of A do
          ins ← Select instances of value v
          subtree ← ID3 (ins, instances, attribute–A)
          Add a branch to tree with label v and subtree
        return tree
```

Fig. 1. The pseudocode of conventional ID3

III. Rigorous Decision Rule Problem in the Conventional ID3

In this section, we introduce the problem when the conventional ID3 algorithm is too rigorous in generating the decision rule. Generally, if the training set has a lot of attributes and instances, the final decision tree may contain too many decision rules. Some of these decision rules may have very low number of instances. If we remove them, it tends not to make significant change to the classification accuracy. Fig. 2 shows the example of the partial decision tree when there are too many decision rules. Nodes L and E represent a leaf node and a classified node, respectively. Here we define a classified node as a node which all its children are leaf nodes. In addition, A, B, C and D are decision nodes which still have instances to be further classified so their children compose of non-leaf node(s) and leaf node(s). The symbols + and – are positive and negative instances, respectively.

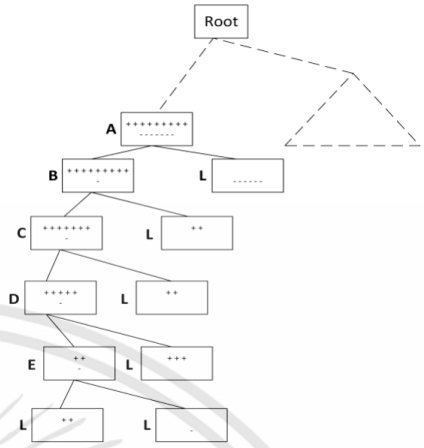


Fig. 2. Example of the partial decision tree with too many decision rules

Let us focus on the left sub tree in Fig. 2. Assume that some attributes have been used as the decision nodes in the upper part of the tree (dash line). Node B contains one negative and nine positive instances. Since the number of negative instances is so small comparing to the number of positive instances, it is worth to ignore those minor instances, i.e., negative instances in this scenario, to prevent the decision tree from generating unnecessary decision rule. As a result, node B is transformed into a leaf node and labeled it with the majority class of instances in node B. Fig. 3 shows the partial decision tree when node B is transformed into a leaf node.

Fig. 3. The final left partial decision tree when node B is transformed into a leaf node

IV. The Proposed Algorithm

According to the rigorous decision rule problem presented in previous section, if we transform a node which contains a few minor instances we can significantly save the depths of the decision tree with a slight change of an accuracy. For example, in Fig.2 (before transforming node B) and Fig.3 (after transforming node B), the maximum number of depths and decision rules are reduced from [depth=5, rules=6] to [depth=1, rules=2]. It is worth to note that we should not ignore the minor instances at the beginning of the algorithm since most attributes at the top of the decision tree are very important and the classification accuracy might be significantly reduced if we transform them into the leaf nodes. Therefore, we modify the ID3 algorithm, called Minor Instances Ignoring ID3 (MII-ID3), to reduce the number of decision rules and the maximum depth of the decision tree

while its accuracy rate is still preserved. The process of the proposed algorithm is described below.

Step 1 : Compute the information gain for each attribute using Eq.(1) and (2). Afterwards, choose the best attribute which has the largest information gain. If there are two or more best attributes, randomly choose one among them.

Step 2 : Create a new node by splitting the training instances depending on the possible values of the best attribute chosen in the previous step.

Step 3 : Calculate the probability of transformation by using Eq.(3).

$$Prop = e^{-(\frac{CurDepth}{R(A)*RemAtts})} \quad (3)$$

Where

Prop is the probability of transformation,

R(A) is the remainder value calculated by Eq.(2),

C*urDepth* is the current depth of the decision tree,

RemAtts is the number of remaining attributes; cannot be zero.

The transformation depends on the probability value, *Prop*. The chance of transformation will be high if the value of Prop is low. There are three important elements of *Prop*. The first one is *R(A)* or the remainder value. The remainder of an attribute A is zero if A is a classified node, i.e., its children are all leaves. The node transformation is not applied to a classified node. In case of *R(A)* is very small (approaching zero), the children of an attribute A should be transformed into leaf nodes. The reason is that the number of instances in one class will be so small comparing to those of the other classes. Therefore, it is worth to ignore those minor instances, as previously described in Fig. 2 and 3. Another two factors in Eq.(3) are *CurDepth* and *RemAtts.* The node transformation is unlikely to occur if the current depth is too shallow, i.e., attributes at the top of the decision tree are very important, and the number of remaining attributes is large. During the beginning of the decision tree construction, there will be a lot of remaining attributes. Therefore, the chance of attribute transformation should be diminished.

Step 4 : Randomly pick a random value *r* between 0 and 1. If *r* is greater than the probability *Prob* in Eq. (3) and R(A) is not zero, transform a node by creating the new leaf node and labeling it with the majority class of the remaining instances.

Sep 5 : Repeat steps 1-4 on the remaining instances and attributes until all instances are classified.

We can simply modify the pseudocode of the conventional ID3 algorithm in Fig. 1 by changing line 6 to "**else if** *attributes* is empty or *state* equal to true" and inserting statements in Fig. 4 between lines 11 and 12 of Fig. 1. The complete pseudocode of the proposed algorithm is shown in Fig. 5.

```
r ← Randomize decimal between 0 and 1
prop ← Compute probability value
if r is greater than prop and R(a) is not zero,
    then set value of state to true
else
    then set value of state to false
```

Fig. 4. Portion of statements of the proposed algorithm

```
1   function ID3 (instances, parent_instances, attributes) returns a tree
      if instances is empty,
        then return the leaf node which labeled by the majority of parent_instances
      else if all instances have the same classification,
        then return the leaf node which labeled by its classification
6     else if attributes is empty or state equal to true,
        then return the leaf node which labeled by the majority of instances
      else
        Compute information gains for all attributes using instances and attributes
        A ← Select the attribute which has the best information gain
11      Insert the new node labeled by attribute A to tree
        r ← Randomize decimal between 0 and 1
        prop ← Compute probability value
        if r is greater than prop and R(a) is not zero,
          then set value of state to true
        else
          then set value of state to false
18      for each value v of A do
          ins ← Select instances of value v
          subtree ← ID3 (ins, instances, attribute–A)
          Add a branch to tree with label v and subtree
        return tree
```

Fig. 5. The pseudocode of MII-ID3 algorithm when variable 'state' is initialized to false

V. EXPERIMENTS

A. Datasets

We consider the standard datasets which contain many attributes and instances because those datasets tend to encounter the nodes with minor instances more frequently than simple datasets. Table I shows characteristics of six complex datasets selected from UCI repository and Weka.

TABLE I. CHARACTERISTICS OF SIX DATASETS

Datasets	No. of instances	No. of attributes
Connect-4	67557	42
Phishing Websites	11055	30
Supermarket	4627	216
Insurance Company Benchmark	9822	85
Molecular Biology	3190	60
Soybean	683	35

B. Experimental conditions and results

Every dataset in Table I is randomly divided into two halves for the training set and the test set. We also keep the balance of instances among all classes within each half. To derive the stable results, we conduct the experiments for one thousand times with the same training and test sets for both the conventional ID3 and the proposed algorithm (MII-ID3). The experimental results between the conventional ID3 and MII-ID3 is shown in Table II.

TABLE II. THE EXPERIMENTAL COMPARISONS ON THE AVERAGE OF 1000 RUNS BETWEEN THE CONVENTIONAL ID3 AND MII-ID3

Datasets	Measurement			
	Avg. Accuracy		*Avg. Max Depth*	
	ID3	*MII-ID3*	*ID3*	*MII-ID3*
Connect-4	**73.65**	73.59	22.73	**15.73**
Phishing Websites	**94.91**	93.65	30.00	**10.94**
Supermarket	68.68	**70.47**	216.00	**21.51**
Insurance Company Benchmark	90.28	**90.91**	85.00	**24.01**
Molecular Biology	**88.99**	88.75	60.00	**8.79**
Soybean	**88.03**	87.33	35.00	**8.13**

The comparison in Table II indicates that the MII-ID3 algorithm produces the average maximum depth significantly less than that of the conventional ID3 algorithm. MII-ID3 algorithm can reduce the average maximum depth in Connect-4, Phishing Websites, Supermarket, Insurance Company Benchmark (COIL 2000), Molecular Biology (Splice-junction Gene Sequences) and Soybean (Large) by 30.81%, 63.54%, 90.04%, 71.75%, 85.35%, and 76.78%, respectively. Moreover, its average accuracy is still preserved in the satisfactory level. Besides two measurements, it is interesting to measure the execution time of two methods. Table III shows the execution time comparisons between the conventional ID3 and MII-ID3 algorithm. Subsequently, Table IV shows the percentage of improvement of both training time and testing time.

TABLE III. THE EXECUTION TIME COMPARISON BETWEEN THE CONVENTIONAL ID3 AND THE MII-ID3

Datasets	Measurement			
	Avg. Training Time (ms)		*Avg. Testing Time (ms)*	
	ID3	*MII-ID3*	*ID3*	*MII-ID3*
Connect-4	1441.30	**1040.92**	0.25485	**0.19149**
Phishing Websites	64.02	**40.60**	0.02467	**0.01616**
Supermarket	374.50	**299.00**	0.02030	**0.01693**
Insurance Company Benchmark	417.21	**234.97**	0.01591	**0.01294**
Molecular Biology	55.04	**48.92**	0.00530	**0.00499**
Soybean	15.29	**13.18**	0.00099	**0.00093**

TABLE IV. THE PERCENTAGE OF IMPROVEMENT OF TRAINING TIME AND TESTING TIME

Datasets	% Improvement	
	Training Time	*Testing Time*
Connect-4	27.78	24.86
Phishing Websites	36.58	34.49
Supermarket	20.16	16.60
Insurance Company Benchmark	43.68	18.69
Molecular Biology	11.12	5.87
Soybean	13.80	5.64

The experimental results in Table III indicate that the averages of both training time and testing time of the MII-ID3 algorithm are less than those of the conventional ID3 in all datasets. Since most instances may be classified at more shallow levels than the maximum depth, we report the average depth of all instances, when they are classified, in Table V.

TABLE V. THE COMPARISON OF THE AVERAGE DEPTH FOR THE TEST SET BETWEEN THE CONVENTIONAL ID3 AND MII-ID3

Datasets	Measurement		
	Avg. Depth		*% Improvement*
	ID3	*MII-ID3*	
Connect-4	9.36	**7.02**	25.00
Phishing Websites	5.46	**3.61**	33.86
Supermarket	10.92	**9.04**	17.20
Insurance Company Benchmark	4.44	**3.27**	26.35
Molecular Biology	4.13	**3.86**	6.51
Soybean	3.62	**3.36**	7.36

The comparison results in Table V indicate that the improvement of the average depth is consistent with the improvement of the test time in Table IV. Fig. 6 illustrates the relation between the reduction of testing time and the reduction of average depth for all datasets.

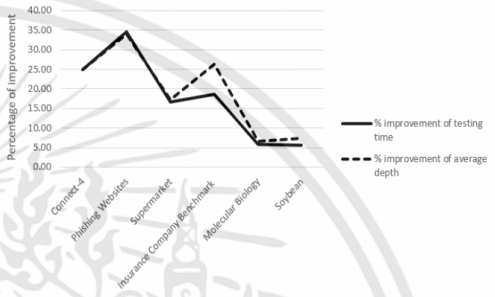


Fig. 6. The relation between the reduction of test time and the reduction of average depth

In order to confirm that MII-ID3 can relax the rigorousness of the conventional ID3 algorithm in generating a set of decision rules, we measure the number of decision rules, i.e., the number of leaf nodes, between both algorithms and illustrate in Table VI.

TABLE VI. THE NUMBER OF DECISION RULES BETWEEN THE CONVENTIONAL ID3 AND THE MII-ID3

Datasets	Measurement		
	Avg. No. of Decision Rules		*% Improvement*
	ID3	*MII-ID3*	
Connect-4	14743.47	**4646.37**	68.49
Phishing Websites	1009.06	**100.77**	90.01
Supermarket	596.05	**286.15**	51.99
Insurance Company Benchmark	22784.01	**3421.84**	84.98
Molecular Biology	568.53	**288.36**	49.28
Soybean	193.70	**103.93**	46.35

From Table VI, MII-ID3 can significantly reduce the number of decision rules from the conventional ID3 in all datasets. Therefore, the rigorousness of the conventional ID3 is drastically relaxed as the MII-ID3 algorithm cuts off the number of decision rules more than 45% in every dataset.

The crucial reason behind this experimental result is that the minor instances ignoring in the MII-ID3 algorithm significantly decreases the number of maximum depths during a training period. Since the number of maximum depths of the decision tree decreases, the average depth of all instances when they are classified also decreases. As a result, the classification time during the test phase is also reduced. In conclusion, MII-ID3 algorithm not only reduces the number of depths in the decision tree, but also reduces the training time and the test time, while its accuracy is still acceptable. Moreover, it also significantly reduces the number of decision rules, compared to the conventional ID3 algorithm.

VI. Conclusion

This paper proposes the MII-ID3 algorithm to relax the rigorous decision rule problem of the conventional ID3 algorithm by ignoring minor instances. The proposed algorithm is evaluated by using six datasets from UCI repository and Weka. The experimental results indicate that the proposed algorithm significantly decreases the average maximum number of depths of the decision tree while the average classification accuracy is still preserved. MII-ID3 also produces fewer number of decision rules than the conventional ID3 algorithm. As a result, the cost of space for keeping nodes in the decision tree also decreases. In addition, the proposed algorithm also reduces the classification time and the training time in every dataset.

References

[1] J. R. Quinlan,“Induction of decision trees,” *Mach. Leaning*, Vol. 1, pp.81-106, 1986.

[2] D. Wang and L. Jiang, “An Improved Attribute Selection Measure for Decision Tree Induction,” in *4th Int. Conf. Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, Haikou, China, pp.1-5.

[3] Z. Wang, Y. Liu, and L. Liu, "A new way to choose splitting attribute in ID3 algorithm," in *2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conf. (ITNEC)*, Chengdu, China, pp.659-663.

[4] C. Guan and X. Zeng, "An Improved ID3 Based on Weighted Modified Information Gain," in *2011 7th Int. Conf. Computational Intelligence and Security*, Hainan, China, pp.1283-1285.

[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. (ICSEC)*, 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/datasets.html

[7] E. Frank, M. Hall, P. Reutemann, and L. Trigg. *WEKA*. (2017) [online]. Available: https://www.cs.waikato.ac.nz/ml/weka/downloading.html

[8] S. Russell and P. Norvig, "Learning from Examples," in *Artificial Intelligence A Modern Approach*, 3rd ed. Essex, England: Pearson, 2014, ch. 18, sec. 3, pp. 708-718.

ประวัติผู้เขียน

ชื่อ-นามสกุล	นางสาวณิชา แก้วรอด
วัน เดือน ปีเกิด	11 มีนาคม 2537 ที่อุตรดิตถ์
ที่อยู่	47/4 หมู่ 2 ต.ป่าเซ่า อ.เมือง จ.อุตรดิตถ์ 53000 โทร 080-5065856
ประวัติการศึกษา	2559 เทคโนโลยีสารสนเทศ สาขาวิชาวิศวกรรมคอมพิวเตอร์ (เกียรตินิยมอันดับ2) มหาวิทยาลัยแม่ฟ้าหลวง