การผสานกฎจากป่าแบบสุ่ม

นางสาวนภาพร ศิริกุลวิริยะ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2554

INTEGRATION OF RULES FROM RANDOM FORESTS

Ms. Naphaporn Sirikulviriya

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computer Science
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2011

หัวข้อวิทยานิพนธ์	การผสานกฎจากป่าแบบสุ่ม
โดย	นางสาวนภาพร ศิริกุลวิริยะ
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.บุญสม เลิศหิรัญวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ)

..... กรรมการ
(รองศาสตราจารย์ ดร.ญาใจ ลิ่มปิยะกรณ์)

..... กรรมการภายนอกมหาวิทยาลัย
(ดร.เด่นดวง ประดับสุวรรณ)

นภาพร ศิริกุลวิริยะ : การผสานกฎจากป่าแบบสุ่ม. (INTEGRATION OF RULES FROM RANDOM FORESTS) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผู้ช่วยศาสตราจารย์ ดร. สุกรี สิญธุภิญโญ, 61 หน้า.

วิธีการป่าแบบสุ่ม เป็นเครื่องมือการทำนายที่มีประสิทธิภาพในการทำเหมืองข้อมูล อย่างไรก็ตาม การใช้กฎที่ได้จากป่าเป็นเรื่องที่ยากมากเพราะมีกฎจำนวนมาก ซึ่งอยู่ในรูปแบบของข้อมูลที่ได้จากจำนวนของต้นไม้ งานวิจัยนี้ได้นำเสนอวิธีการใหม่ที่สามารถผสานกฎจากต้นไม้หลายต้นในป่าแบบสุ่ม การทดลองแสดงให้เห็นว่ากฎที่ได้จากวิธีการที่นำเสนอให้ผลลัพธ์ที่สามารถเปรียบเทียบและเข้าใจได้จากค่าเฉลี่ยของจำนวนกฎที่ลดลง

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....
สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....
ปีการศึกษา 2554

5171419521 : MAJOR COMPUTER SCIENCE
KEYWORDS : RANDOM FORESTS / INTEGRATION OF RULES / DECISION TREE

NAPHAPORN SIRIKULVIRIYA : INTEGRATION OF RULES FROM RANDOM FORESTS. ADVISOR : ASST. PROF. SUKREE SINTHUPINYO, Ph.D., 61 pp.

Random forest method is an effective prediction tool in data mining. However, to use the rules obtained from the forest is a strenuous task because there are a lot of rules, which are patterns of the data, from a number of trees. This paper proposes a new method which can integrate rules from multiple trees in a forest. The experiments show that the rules obtained from our method yields the comparable results and are understandable by means of the decreased number of rules.

Department Computer Engineering Student's Signature .....
Field of Study Computer Science Advisor's Signature .....
Academic Year 2011

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้เป็นการศึกษาเรื่อง การผสานกฎจากป่าแบบสุ่ม สามารถสำเร็จลุล่วงไปได้ด้วยความอนุเคราะห์อย่างยิ่งของผู้ช่วยศาสตราจารย์ ดร.สุกรี สิญธุภิญโญ (อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก) ที่ได้ให้ความรู้ คำแนะนำและแนวทางการวิจัย ตลอดจนการตรวจสอบและแก้ไขข้อบกพร่องต่างๆ จนกระทั่งเสร็จสมบูรณ์ไปได้ด้วยดี ผู้เสนอวิทยานิพนธ์จึงขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้

ขอขอบพระคุณ ศาสตราจารย์ ดร. บุญเสริม กิจศิริกุล รองศาสตราจารย์ ดร.ญาใจ ลิ่มปิยะกรณ์ และอาจารย์ ดร.เด่นดวง ประดับสุวรรณ กรรมการสอบวิทยานิพนธ์ ที่กรุณาเสียสละเวลาให้คำแนะนำ ตรวจสอบ และแก้ไขวิทยานิพนธ์ฉบับนี้

ขอขอบคุณครอบครัว เพื่อน ๆ พี่ ๆ ทุกคน รวมทั้งหัวหน้างานที่มีส่วนช่วยเหลือในการทำวิทยานิพนธ์ครั้งนี้ให้เสร็จสมบูรณ์ได้ด้วยดี ซึ่งมิได้กล่าวนามไว้ ณ ที่นี้ทั้งหมด

ท้ายสุดนี้ หากมีสิ่งใดขาดตกบกพร่องหรือข้อผิดพลาดประการใด ผู้เสนอวิทยานิพนธ์ขออภัยเป็นอย่างสูงในข้อบกพร่องและความผิดพลาดนั้น และหวังว่าวิทยานิพนธ์ฉบับนี้จะเป็นประโยชน์บ้างไม่มากก็น้อยสำหรับผู้ที่สนใจจะศึกษารายละเอียด

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

วิธีการเอนเซมเบิล (Ensemble Method) เป็นวิธีการที่ได้รับความสนใจอย่างมากและเป็นที่ยอมรับกันอย่างกว้างขวาง ในวงการเกี่ยวกับการเรียนรู้ของเครื่อง (Machine Learning) และการทำเหมืองข้อมูล (Data Mining) ว่าให้ผลได้ดีกว่าตัวจำแนกเดียว (Single Classifier) โดยในช่วง 10 กว่าปีที่ผ่านมามีการพัฒนาอัลกอริทึมแบบเอนเซมเบิล อย่างมาก เช่น Bagging (Breiman-1996), Boosting (Freund and Schapire-1996), Arching (Breiman-1998) และ Random Forests (Breiman-2001)

อย่างไรก็ตามเอนเซมเบิลที่สร้างขึ้นโดยใช้เทคนิคในปัจจุบัน บางครั้งก็เกินความจำเป็น เพราะมันต้องใช้หน่วยความจำจำนวนมากในการเก็บแบบจำลองการเรียนรู้ และยังใช้เวลาในการคำนวณมาก เพื่อที่จะให้ได้ผลทำนาย อย่างไรก็ตามมันก็ไม่จริงเสมอไปที่ว่ายิ่งเอนเซมเบิล มีขนาดใหญ่ขึ้นจะให้ผลทำนายที่ดีขึ้น

ป่าแบบสุ่ม (Random Forests) เป็นเทคนิคหนึ่งของการเรียนรู้ของเครื่องที่พัฒนาไปอย่างมากสำหรับการทำเหมืองข้อมูลและการค้นหาความรู้ (Knowledge Discovery) วิจัยโดย Leo Breiman [1] มีคุณสมบัติในการจำแนกตัวอย่างที่ประกอบไปด้วยชุดของตัวจำแนกที่มีโครงสร้างแบบต้นไม้ สามารถทำนายประเภทของข้อมูลที่มีรูป แบบที่ไม่สามารถคาดเดาได้ ในการสร้างแบบจำลองเพื่อนำไปใช้ในการทำนายผลของข้อมูล เป็นวิธีการวิเคราะห์ที่เหมาะสมที่สุด สำหรับโครงสร้างของข้อมูลที่มีความซับซ้อนที่รวมอยู่ในชุดข้อมูลขนาดเล็กจนถึงปานกลางที่มีข้อมูลน้อยกว่า 10,000 แถว แต่ยอมให้มีข้อมูลมากกว่า 1,000,000 คอลัมน์ ดังนั้นป่าแบบสุ่มได้ถูกรับรองโดยนักวิจัยด้าน ชีวการแพทย์ และ เภสัชกรรม จำนวนมาก ว่ามีประสิทธิภาพในการทำนายสูงและแม่นยำ เนื่องจากกฎที่ว่าด้วยจำนวนมากหรือการเพิ่มจำนวนของต้นไม้ในป่า ไม่ได้ทำให้มันเกิดความเฉพาะเจาะจง (overfit) เมื่อเราทำการสุ่มที่ถูกประเภทไปแล้วจะทำให้ตัวจำแนก (classifier) และตัวถดถอย (regressor) นี้มีความแม่นยำ

ถึงแม้ว่าป่าแบบสุ่ม จะมีประสิทธิภาพในการทำนายสูงและแม่นยำแล้ว แต่ข้อจำกัดในด้านที่ว่า ต้นไม้ที่สร้างขึ้นจำเป็นต้องโตเต็มต้นโดยไม่มีการตัดเล็มกิ่ง (Pruning) ดังนั้นหากเราสามารถลดเวลาในการจำแนกตัวอย่างหรือลดเวลาในการทำนายได้ ก็จะเป็นผลให้วิธีการป่าแบบสุ่ม นี้มีประสิทธิภาพมากขึ้น

ปัจจุบันมีอัลกอริทึมที่ช่วยในการรวมผลลัพธ์ที่ได้จากการทำนายของต้นไม้ตัดสินใจที่วิจัยโดย Zhi-Hua Zhou และ Wei Tang [2] และมีนักวิจัยจำนวนมากได้พัฒนาอัลกอริทึมในการตัดเล็มกิ่ง (pruning ensemble) นำหลายเทคนิคในการปรับปรุงชุดข้อมูลสอน หรือปรับแต่งต้นไม้ตัดสินใจ ตัวอย่างเช่น งานวิจัยของ Yi Zhang, Samuel Burer และ W. Nick Street [3] ที่ได้นำเสนอวิธีการตัดเล็มกิ่งแบบใหม่เพื่อปรับปรุงประสิทธิภาพและประสิทธิผลของเอนเซมเบิล โดยใช้เทคนิค SDP relaxation เพื่อให้ได้มาซึ่งการประมาณการที่ดีและให้ผลว่าอัลกอริทึมในการตัดเล็มกิ่งโดยวิธีการแบบ SDP นี้ให้ประสิทธิภาพที่ดีกว่าอัลกอริทึมในการตัดเล็มกิ่ง แบบ Metric มีงานวิจัยเกี่ยวกับการกำจัดข้อมูลรบกวน (Noise Reduction) ซึ่งทำให้ได้ต้นไม้ตัดสินใจที่มีประสิทธิภาพในการทำนายดีขึ้น โดยให้น้ำหนักกับแต่ละหน่วยที่กำลังพิจารณา เพื่อปรับปรุงหรือตัดกิ่งออก ซึ่งเป็นการปรับปรุงต้นไม้หรือชุดข้อมูลสอน มีงานวิจัยของ Anneleen Van Assche และ Hendrik Blockeel [4] ที่นำเสนอวิธีการสร้างต้นไม้ต้นเดียวต้นใหม่จากกลุ่มของต้นไม้ในเอนเซมเบิล ซึ่งในหลายๆ งานวิจัยยังไม่มีวิธีการปรับปรุงกฎและรวมกฎจากเอนเซมเบิลหรือกลุ่มของต้นไม้ตัดสินใจหลายต้นเป็นต้นไม้ต้นดียว เพื่อให้ได้เซตของต้นไม้ตัดสินใจที่มีขนาดเล็กที่สุด โดยที่เซตของกฎที่เล็กที่สุดนี้มีประสิทธิภาพในการทำนายสูงอีกทั้งยังช่วยลดเวลาในการทำนายได้อีกด้วย

ดังนั้นในงานวิจัยนี้จึงได้นำเสนอวิธีจากวิธีการเรียนรู้ของต้นไม้ตัดสินใจต้นเดียวที่ได้จากการผสานกฎจากป่าแบบสุ่ม โดยจะใช้วิธีการเลือกเซตของกฎจากกลุ่มของต้นไม้ตัดสินใจในป่าแบบสุ่มมาผสานกัน หากได้เปอร์เซ็นต์ความถูกต้องของผลทำนายเพิ่มขึ้นก็จะนำผลลัพธ์ที่เกิดจากเซตของกฎใหม่นี้มาใช้ในการผสานกฎครั้ง

1.2 วัตถุประสงค์ของการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อนำเสนอวิธีการผสานกฎจากป่าแบบสุ่มและวิเคราะห์ผลลัพธ์เพื่อเปรียบเทียบประสิทธิภาพการทำนายของป่าแบบสุ่มที่ไม่ได้ผสานกฎและป่าแบบสุ่มที่ทำการผสานกฎว่าวิธีการที่นำเสนอให้ผลออกมาดีกว่าหรือไม่

1.3 ขอบเขตของการวิจัย

1. พัฒนาอัลกอริทึมที่ใช้ผสานกฎจากป่าแบบสุ่มโดยทำการผสานต้นไม้ตัดสินใจจากป่าแบบสุ่มครั้งละ 2 ต้น
2. ใช้ชุดข้อมูลจาก UCI Repository เป็นข้อมูลสำหรับการทดสอบ

1.4 ประโยชน์ที่คาดว่าจะได้รับ

การผสานกฎจากป่าแบบสุ่มที่นำเสนอ ทำให้ได้เซตของกฎใหม่ที่ให้ผลทำนายที่ดีขึ้น สามารถลดจำนวนกฎและเพิ่มความถูกต้องในการทำนายให้ดีขึ้น และสามารถนำวิธีการที่นำเสนอในงานวิจัยนี้ไปเป็นแนวทางในการวิจัยต่อไป

1.5 วิธีดำเนินการวิจัย

1. ศึกษาทฤษฎีพื้นฐานของป่าแบบสุ่ม
2. ศึกษาทฤษฎีพื้นฐานของการแปลงต้นไม้เป็นกฎ
3. ศึกษาทฤษฎีพื้นฐานของการสร้างต้นไม้ต้นเดียวเพื่อแทนต้นไม้หลายๆ ต้น
4. ศึกษาเครื่องมือที่ช่วยในการสร้างต้นไม้ตัดสินใจที่พัฒนาใช้กันอยู่ในปัจจุบัน
5. จัดเตรียมข้อมูลตัวอย่างที่จะใช้ในการทดสอบ
6. พัฒนาอัลกอริทึมการผสานกฎที่ได้จากป่าแบบสุ่ม
7. ออกแบบวิธีการทดลอง
8. ทดสอบวิธีการที่นำเสนอ
9. วิเคราะห์และเปรียบเทียบผลลัพธ์ของป่าแบบสุ่มก่อนการผสานกฎกับผลลัพธ์ของป่าแบบสุ่มหลังการผสานกฎ
10. สรุปผลการดำเนินงานและเรียบเรียงวิทยานิพนธ์

1.6 ลำดับขั้นตอนในการเสนอผลการวิจัย

วิทยานิพนธ์นี้แบ่งเนื้อหาทั้งหมดออกเป็น 5 บทดังต่อไปนี้ บทที่ 1 เป็นบทนำซึ่งกล่าวถึงความเป็นมาและความสำคัญของปัญหา วัตถุประสงค์ของการวิจัย รวมถึงประโยชน์ที่คาดว่าจะได้รับ บทที่ 2 กล่าวถึงแนวคิดทฤษฎีและงานวิจัยที่เกี่ยวข้อง บทที่ 3 กล่าวถึงวิธีการออกแบบขั้นตอนการดำเนินงาน บทที่ 4 กล่าวถึงวิธีการทดลองและผลการทดลอง บทที่ 5 กล่าวถึงการสรุปผลการวิจัยและข้อเสนอแนะ

1.7 ผลงานที่ตีพิมพ์จากวิทยานิพนธ์

ส่วนหนึ่งของวิทยานิพนธ์นี้ ได้รับการตอบรับให้ตีพิมพ์เป็นบทความทางวิชาการในหัวข้อเรื่อง “Integration of Rules from a Random Forest” โดย นางสาวนภาพร ศิริกุลวิริยะ และผู้ช่วยศาสตราจารย์ ดร. สุกรี สินธุภิญโญ ในงานประชุมวิชาการ 2011 International Conference on Information and Electronics Engineering (ICIEE 2011) ณ กรุงเทพมหานคร ประเทศไทย วันที่ 28-29 พฤษภาคม พ.ศ. 2554

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 แนวคิดและทฤษฎีที่เกี่ยวข้อง

2.1.1 ป่าแบบสุ่ม (Random Forests)

ป่าแบบสุ่ม (Random Forests) คำนี้มาจาก Random Decision Forests ซึ่งถูกเสนอครั้งแรก โดย Tin Kam Ho จาก Bell Labs ในปี 1995 ต่อมาภายหลังวิธีการนี้ได้ถูกขยายและจัดทำให้เป็นรูปแบบทั่วไปมากขึ้นโดย Leo Breiman ซึ่งวิธีการนี้จะรวมเอาความคิดของวิธีการ bagging ของ Leo Breiman และ การเลือกแบบสุ่มของคุณลักษณะ ของ Tim Kam Ho และ Yali Amit และ Donald Geman เพื่อสร้างกลุ่มของต้นไม้ตัดสินใจที่มีการควบคุมความแปรปรวน การเลือกเซตย่อยแบบสุ่มของคุณลักษณะเป็น ตัวอย่างของวิธีการย่อยแบบสุ่ม

ป่าแบบสุ่มเป็นตัวจำแนกเอนเซมเบิล ที่ประกอบไปด้วยต้นไม้ตัดสินใจหลายต้น และให้ผลลัพธ์เป็นคลาส ที่เป็นผลลัพธ์ของคลาสจากต้นไม้แต่ละต้น อัลกอริทึมสำหรับป่าแบบสุ่มถูกพัฒนาโดย Leo Breiman และ Adele Cutler และ "Random Forests" เป็นเครื่องหมายการค้าของพวกเขา

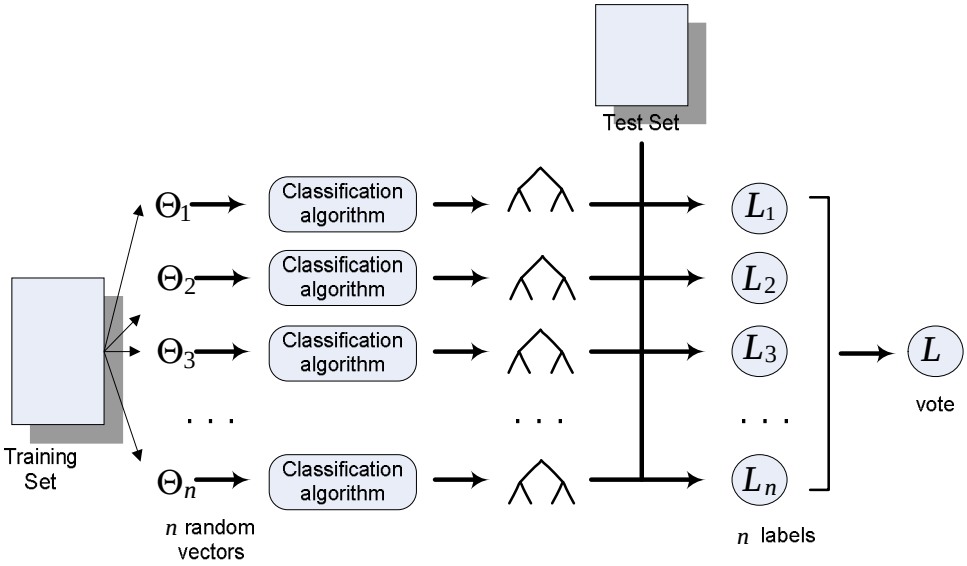
ป่าแบบสุ่ม เป็นส่วนประกอบของตัวทำนายแบบต้นไม้ซึ่งแต่ละต้นนั้นจะขึ้นอยู่กับค่าของเวคเตอร์ที่สุ่มขึ้นมาอย่างอิสระต่อกันด้วยการกระจายแบบเดียวกันของต้นไม้ทั้งหมดที่อยู่ในป่า ความผิดพลาดโดยทั่วไปของป่าจะมีค่าเข้าสู่ลิมิตเมื่อจำนวนของต้นไม้ในป่ามีจำนวนเยอะขึ้น ความผิดพลาดโดยทั่วไปของป่าในการจำแนกแบบต้นไม้ขึ้นอยู่กับความแข็งแรง(strength)ของต้นไม้แต่ละต้นในป่า และความสัมพันธ์ (correlation) ระหว่างกัน

การใช้การเลือกสุ่มเพื่อที่จะแบ่งโหนดที่มีอัตราความผิดพลาดที่น่าพอใจกว่าเทคนิค Adaboost แต่ก็มีความเป็นไปได้ที่จะพบข้อมูลรบกวน ได้มากกว่าการประมาณการภายในควบคุมความผิดพลาดที่เกิดขึ้น ความแข็งแรงและความสัมพันธ์ สิ่งเหล่านี้ใช้ในการแสดงการตอบสนองในการเพิ่มขึ้นของจำนวนคุณลักษณะที่ใช้ในการประมาณการภายในในการแบ่งยังถูกนำมาใช้ในการวัดความสำคัญของตัวแปรอีกด้วย ความคิดเหล่านี้ต่างก็สามารถนำมาใช้ได้กับการถดถอย (regression) ด้วยเหมือนกัน

Leo Breiman แสดงให้เห็นว่า ไม่เพียงแต่ว่าป่าแบบสุ่มที่จะเป็นตัวจำแนกที่มีประสิทธิภาพสูงเท่านั้นแต่ยังได้กล่าวถึงปัญหาจำนวนมากที่มีความซับซ้อนและกระทบกับประสิทธิภาพของวิธีการจำแนกแบบอื่นที่ขัดแย้งกับขอบเขตของแอพพลิเคชันแบบต่างๆ โดยเฉพาะไม่ต้องการการทำให้สมมติฐานง่ายขึ้น ในเรื่องแบบจำลองการกระจายตัวของข้อมูล และใน

เรื่องขบวนการผิดพลาด ยิ่งไปกว่านั้นมัน ก็ง่ายต่อการรองรับประเภทของข้อมูลต่างๆ ได้อย่างง่ายดาย และมีความคงทนอย่างสูงในการฝึกฝนด้วยขนาดของป่า ในขณะที่จำนวนของต้นไม้ในป่าแบบสุ่มเพิ่มค่าความผิดพลาดทั่วไป ที่ถูกแสดงให้เห็นในงานวิจัย ได้เบนเข้าหากัน และ ถูกกำหนดขอบเขต

ป่าแบบสุ่มเป็นตัวจำแนกที่ประกอบไปด้วยชุดของตัวจำแนกที่มีโครงสร้างแบบต้นไม้ $\{h(x,\Theta k), k=1, ..., k=n\}$ โดยที่ $\{\Theta k\}$ เป็นเวกเตอร์แบบสุ่มที่มีการกระจายที่เป็นอิสระต่อกันโดยสิ้นเชิง และต้นไม้ทุกต้นจะมีหนึ่งโหวตสำหรับคลาสที่นิยมมากที่สุด ที่นำเข้าเวกเตอร์จำนวน x ครั้ง โดยที่ h คือ x คือ เวคเตอร์นำเข้า n คือจำนวนของตัวอย่างในชุดข้อมูลสอน



ภาพที่ 1 โครงสร้างโดยทั่วไปของป่าแบบสุ่ม

กฎที่ว่าด้วยจำนวนมากหรือการเพิ่มขึ้นของต้นไม้ในป่าแสดงให้เห็นว่าการเกิดความเฉพาะเจาะจงไม่ได้เป็นปัญหาแต่อย่างใด

มีงานวิจัยบางงานได้บอกว่าป่าแบบสุ่มนั้น เกิดความผิดพลาดแบบทั่วไปน้อยกว่าวิธีการอื่น อย่างเช่น วิธีการเลือกแบ่งแบบสุ่มทำได้ดีกว่าวิธีการแบกกิ้งการทำให้ความถูกต้องดีขึ้นจะต้องมีการลดค่าความสัมพันธ์ ในขณะเดียวกันก็ต้องรักษาความเข็งแรง เอาไว้ด้วยป่าที่นำมาศึกษาประกอบไปด้วยการใช้วิธีการเลือกอินพุตแบบสุ่ม หรือ ใช้ส่วนประกอบจากอินพุตที่ทุกโหนด

ในการสร้างต้นไม้ผลลัพธ์ของป่า ที่ได้ให้ความถูกต้องแม่นยำที่เปรียบเทียบได้กับเอดาบูท (Adaboost) กระบวนการนี้มีคุณลักษณะที่น่าพอใจดังนี้

1. ความถูกต้องดีเทียบเท่ากับเอดาบูท หรือดีกว่าในบางครั้ง
2. ค่อนข้างมีความคงทนต่อข้อมูลที่มีค่าผิดปกติ และ ข้อมูลรบกวน
3. เร็วกว่าแบกกิง หรือ บูสต์ติง
4. ให้ประโยชน์ในการประมาณภายในของความผิดพลาด, ความแข็งแรง, ความสัมพันธ์ และ ความสำคัญของตัวแปร
5. ง่ายและทำให้ขนานได้อย่างง่าย

Algorithm 1: Pseudo code for the random forest algorithm

1. To generate c classifiers:
2. **for** i = 1 to c **do**
3. Randomly sample the training data D with replacement to produce D_i
4. Create a root node, N_i containing D_i
5. Call BuildTree(N_i)
6. **end for**
7.
8. **BuildTree(N):**
9. **if** N contains instances of only one class **then**
10. **return**
11. **else**
12. Randomly select x% of the possible splitting features in N
13. Select the feature F with the highest information gain to split on
14. Create f child nodes of N , N_1,..., N_f , where F has f possible values ($F_1, \ldots, F_f$)
15. **for** i = 1 to f **do**
16. Set the contents of N_i to D_i , where D_i is all instances in N that match
17. F_i
18. Call BuildTree(N_i)
19. **end for**
20. **end if**

ภาพที่ 2 รหัสจำลองของอัลกอริทึมป่าแบบสุ่ม [5]

อัลกอริทึมในการเรียนรู้ป่าแบบสุ่ม [1],[6] ต้นไม้แต่ละต้นจะสร้างขึ้นโดยใช้ขั้นตอนดังต่อไปนี้

1. ให้ N เป็นจำนวนของกรณีฝึกฝนและ M เป็นจำนวนของตัวแปรในการจำแนก

2. ให้ m เป็นจำนวนของตัวแปรอินพุตที่ใช้กำหนดการตัดสินใจที่โหนดของต้นไม้ โดยที่ m ควรจะน้อยกว่า M มาก
3. เลือกชุดของการเรียนรู้สำหรับต้นไม้โดยการเลือก N ครั้งด้วยการแทนที่จาก N กรณีฝึกฝนทั้งหมด (ใช้ตัวอย่าง บูตสแตรป) ใช้ส่วนที่เหลือของกรณีเพื่อประเมินความผิดพลาดของต้นไม้ โดยการทำนาย
4. สำหรับแต่ละโหนดของต้นไม้ ให้สุ่มเลือกตัวแปร m บนพื้นฐานการตัดสินใจที่โหนดนั้น คำนวนการแบ่งที่ดีที่สุดบนพื้นฐานของตัวแปร m นั้นในชุดข้อมูลสอน
5. ต้นไม้แต่ละต้นจะต้องเติบโตอย่างเต็มที่ และต้องไม่มีการตัดกิ่ง

ประโยชน์ของวิธีการป่าแบบสุ่ม

1. สำหรับชุดข้อมูลจำนวนมาก มันสร้างตัวจำแนกที่มีความแม่นยำสูง
2. มันสามารถทำงานกับตัวแปรอินพุทจำนวนมากได้
3. มันประมาณความสำคัญของตัวแปรในการกำหนดการจำแนก
4. มันสร้างการประมาณการที่ไม่บิดเบือนภายใน ของความผิดพลาดโดยทั่วไปในระหว่างขบวนการสร้างป่า
5. มันรวมวิธีที่ดีสำหรับการประมาณข้อมูลที่สูญหายและยังคงไว้ซึ่งความแม่นยำเมื่อข้อมูลส่วนใหญ่ได้สูญหายไป
6. มันให้วิธีการทดลองเพื่อระบุความสัมพันธ์ของตัวแปร
7. มันสามารถถ่วงดุลย์ความผิดพลาดในการสร้างคลาส จากชุดข้อมูลที่ไม่สมดุลย์
8. มันคำนวณหาค่าความใกล้เคียงกันระหว่างกรณี ซึ่งมีประโยชน์สำหรับการรวมเป็นกลุ่มของข้อมูล, การระบุขอบเขตของข้อมูล และ การแสดงภาพข้อมูล
9. โดยวิธีข้างต้น มันสามารถถูกขยายไปถึงข้อมูลแบบไม่มีการกำหนดประเภทของข้อมูลมาก่อนซึ่งจะนำไปสู่ การรวมกลุ่มที่ไม่ได้ดูแล การระบุขอบเขต และภาพรวมของข้อมูล
10. การเรียนรู้อย่างเร็ว

ข้อเสียของวิธีการป่าแบบสุ่ม

1. ป่าแบบสุ่มถูกตัดเล็มกิ่งเพื่อให้เหมาะกับชุดข้อมูลบางชุด วิธีนี้จะทำให้เกิดการจำแนกที่มีข้อมูลรบกวน และงานแบบถดถอย

2. ป่าแบบสุ่มไม่สามารถทำงานได้กับชุดข้อมูลขนาดใหญ่ที่ไม่มีความสัมพันธ์กัน และ เอนเซมเบิลของต้นไม้ตัดสินใจแบบการลดค่าเอนโทรปี
3. มันจะมีประสิทธิภาพที่จะเลือกขอบเขตการตัดสินใจแบบสุ่ม มากกว่าขอบเขตการตัดสินใจแบบการลดค่าเอนโทรปี ดังนั้นการสร้างข้อมูลเอนเซมเบิลขนาดใหญ่จะยิ่งเป็นไปได้มากขึ้น แม้ว่า มันดูเหมือนจะเป็นประโยชน์ในช่วงแรก แต่มันมีผลกระทบของ การคลาดเคลื่อนของการคำนวณจากช่วงเวลาฝึกฝนไปยังช่วงเวลาทดสอบ ซึ่ง ถือว่าเป็นข้อเสียในแอพพลิเคชั่นส่วนใหญ่

ป่าแบบสุ่ม เป็นเครื่องมือที่มีประสิทธิผลในการทำนาย เนื่องจากกฎจำนวนมากไม่ได้ทำให้มันเกิดความเฉพาะเจาจง เมื่อเราทำการสุ่มที่ถูกประเภทไปแล้วจะทำให้มันเป็นตัวจำแนกและตัวถดถอยที่มีความแม่นยำ นอกจากนี้โครงสร้างในเรื่องของความแข็งแรงของตัวทำนายแต่ละตัวและความสัมพันธ์กันยังทำให้เข้าใจถึงความสามารถในการทำนายของป่าแบบสุ่มได้อีกด้วย การใช้งานการประมาณค่าแบบเอาท์ออฟแบก ทำให้ค่าในทางทฤษฎีอื่นๆ ของความแข็งแรงและความสัมพันธ์กันชัดเจนขึ้น

ในระยะหนึ่ง แนวคิดเดิมๆ ที่ว่าป่าแบบสุ่ม ไม่สามารถแข่งขันกับ อาร์คชิงอัลกอริทึม ในแง่ของความถูกต้องแม่นยำได้ เพราะผลลัพธ์ที่ได้ ได้ทำให้ความเชื่อเหล่านั้นหายไป แต่ก็นำไปสู่คำถามที่น่าสนใจบูสต์ติงและอาร์คชิง ว่ามีความสามารถในการลดความคลาดเคลื่อน ได้ดีเช่นเดียวกับ variance (Schapire et al [1998]). อะแดปทีฟแบกกิงอัลกอริทึม ในการถดถอยถูกออกแบบมาเพื่อลดความคลาดเคลื่อนและทำงานได้อย่างมีประสิทธิภาพในการจำแนกเช่นเดียวกันกับการถดถอยและอาร์คชิง มันสามารถเปลี่ยนชุดข้อมูลสอนในขณะที่มีการทำงานไปด้วยได้

ป่าแบบสุ่มให้ผลลัพธ์ที่สามารถเปรียบเทียบได้กับบูสต์ติงและอะแดปทีฟแบกกิง แต่ไม่ได้มีการเปลี่ยนแปลงชุดข้อมูลสอนเพิ่มขึ้น ความแม่นยำของป่าแบบสุ่มบ่งชี้ว่าพวกมันสามารถลดความคลาดเคลื่อนได้ กลไกของมันไม่ได้ปรากฏให้เห็นได้ชัด ป่าแบบสุ่มอาจจะมองเป็นกระบวนการแบบเบเซียน (Bayesian)

การนำเข้าข้อมูลแบบสุ่มและคุณลักษณะแบบสุ่มต่างก็ให้ผลที่ดีในการจำแนก ในการศึกษานี้ชนิดของการสุ่มที่ใช้มีเพียงแบกกิงและคุณลักษณะแบบสุ่มเท่านั้น อย่างเช่น หนึ่งในผู้ที่ถูกอ้างถึงแนะนำให้ใช้การรวมกันของบูลีนแบบสุ่มของคุณลักษณะ

การเพิ่มขึ้นของความแม่นยำนั้นได้มาจากการรวมคุณลักษณะแบบสุ่มกับบูสต์ติง หรือไม่นั้น สำหรับชุดข้อมูลที่มีขนาดใหญ่ขึ้น ดูเหมือนว่าอัตราความผิดพลาดที่น้อยลงอย่างมีนัยสำคัญจะมีความเป็นไปได้การเปลี่ยนแปลงจะน้อยลงบนชุดข้อมูลที่มีขนาดเล็กลงซึ่งต้องการการทดลองเพิ่มเติมในส่วนนี้มากขึ้น แต่ก็มีการแนะนำว่าการนำเข้าแบบสุ่ม ที่ต่างกันก็จะให้ผลที่ดีขึ้นได้

บทความก่อนหน้านี้ของ Breiman แสดงให้เห็นว่าในที่ว่างของการกระจายตัวสำหรับสองกลุ่มปัญหาป่าแบบสุ่มเหมือนกันกับแสดงเคอร์เนล ที่กระทำกับตำแหน่งขอบที่แท้จริง ข้อขัดแย้งมีว่าการสุ่มแบบที่มีความสัมพันธ์ต่ำบังคับให้เกิดการสมมาตรของเคอร์เนล ขณะที่ความแข็งแกร่งเสริมค่าความเบ้ (skewness) ที่ต้องการ ณ แนวขอบเส้นโค้ง หวังเป็นอย่างยิ่งว่า จะช่วยนำทางให้กับบทบาทของความสัมพันธ์ระหว่างกัน และความแข็งแรง โครงร่างทฤษฎีที่ให้ไว้โดย Klienberg สำหรับการจำแนกแบบมีสถิติ อาจช่วยในการทำความเข้าใจได้

2.1.2 **ต้นไม้ตัดสินใจ** (Decision Tree)

ต้นไม้ตัดสิน (Decision Tree) [7] นับเป็นวิธีการเรียนรู้ที่นิยมใช้ในกระบวนการเรียนรู้ของเครื่อง ต้นไม้ตัดสินใจสามารถเรียนรู้โดยการแยกแยะข้อมูลในกลุ่มตัวอย่างออกเป็นกลุ่มย่อยต่างๆ การแยกแยะข้อมูลออกเป็นกลุ่มย่อยนี้ทำได้โดยใช้คุณลักษณะ (Attribute) ของข้อมูลเป็นตัวแยกแยะ การสร้างต้นไม้ตัดสินใจจะทำจากบนลงล่าง (Top-Down) โดยเริ่มจากการเลือกคุณลักษณะ และพิจารณาคุณลักษณะที่มีค่าสารสนเทศที่สูงสุด (Information Gain) หรือค่าเอนโทรปี (Entropy) ที่น้อยที่สุด ในการคัดเลือกคุณลักษณะที่จะขึ้นมาเป็นโหนดราก และโหนดต่อไปตามลำดับ

โดยหลักการพื้นฐานของการสร้างต้นไม้ตัดสินใจ เป็นการสร้างในลักษณะจากบนลงล่าง คือเริ่มจากการสร้างรากของต้นไม้ก่อนแล้วจึงแตกกิ่งไปจนถึงใบโดยแสดงขั้นตอนการสร้างต้นไม้ตัดสินใจได้ดังนี้

1) ต้นไม้เริ่มต้นโดยมีโหนดเพียงโหนดเดียวแสดงถึงชุดข้อมูลสอน
2) ถ้าข้อมูลทั้งหมดอยู่ในกลุ่มเดียวกันแล้ว ให้โหนดนั้นเป็นใบและตั้งชื่อแยกตามกลุ่มของข้อมูลนั้น
3) ถ้าในโหนดมีข้อมูลหลายกลุ่มปะปนอยู่จะต้องวัดค่าเกน (Gain) ของแต่ละคุณลักษณะเพื่อที่จะใช้เป็นเกณฑ์ในการคัดเลือกคุณลักษณะ ที่มีความสามารถในการแบ่งแยกข้อมูลออกเป็นกลุ่มต่างๆ ได้ดีที่สุดโดยคุณลักษณะที่มีค่าเกนมากที่สุดจะถูกเลือกให้เป็นตัวทดสอบหรือคุณลักษณะใช้ในการตัดสินใจ โดยแสดงในรูปของโหนดบนต้นไม้

4) กิ่งของต้นไม้ ถูกสร้างขึ้นจากค่าต่างๆ ที่เป็นไปได้ของโหนดทดสอบ และข้อมูลจะถูกแบ่งออกตามกิ่งต่างๆที่สร้างขึ้น
5) ทำการวนซ้ำเพื่อหาคุณลักษณะที่มีค่าเกนมากที่สุด สำหรับข้อมูลที่ถูกแบ่งแยกออกมาในแต่ละกิ่งเพื่อนำคุณลักษณะนี้มาสร้างเป็นโหนดตัดสินใจต่อไป โดยที่คุณลักษณะที่ถูกเลือกมาเป็นโหนดแล้วจะไม่ถูกเลือกมาอีก สำหรับโหนดในระดับต่อๆ ไป
6) ทำการวนซ้ำเพื่อแบ่งข้อมูลและแตกกิ่งของต้นไม้ไปเรื่อยๆ โดยการวนซ้ำจะสิ้นสุดก็ต่อเมื่อเงื่อนไขข้อใดข้อหนึ่งต่อไปนี้เป็นจริง

ต้นไม้ตัดสินใจเป็นโครงสร้างที่ใช้แสดงกฎที่ได้จากเทคนิคการจำแนกประเภทข้อมูล โดยต้นไม้ตัดสินใจจะมีลักษณะคล้ายโครงสร้างต้นไม้ที่แต่ละโหนดแสดงคุณลักษณะ ในการสร้างต้นไม้ตัดสินใจปัญหาสำคัญที่ต้องพิจารณาคือควรจะตัดสินใจเลือกคุณลักษณะใดมาทำหน้าที่เป็นโหนดรากในแต่ละขั้นตอนของการสร้างต้นไม้และต้นไม้ย่อยของต้นไม้ตัดสินใจ เกณฑ์ที่ใช่ช่วยประกอบการเลือกคุณลักษณะคือการคำนวณค่ามาตรฐานเกน (Gain Criterion) ซึ่งเป็นค่าที่บ่งบอกว่าคุณลักษณะนั้นสามารถจำแนกกลุ่มของข้อมูลได้ดีเพียงใด โดยทดลองเลือกแต่ละคุณลักษระที่เป็นไปได้จากชุดข้อมูลมาทำหน้าที่เป็นโหนดราก ถ้าคุณวมบัติใดให้ค่าเกนสูงที่สุด แสดงว่าคุณลักษณะนั้นสามารถจำแนกกลุ่มของข้อมูลได้ดีที่สุด การใช้ค่าเกนสารสนเทศจะช่วยลดจำนวนครั้งของการทดสอบในการแยกแยะข้อมูลอีกทั้งยังรับประกันว่าต้นไม้ตัดสินใจที่ได้ไม่มีความซับซ้อนมากจนเกินไป

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} Entropy(S_v) \qquad (1)$$

เมื่อ

A คือ คุณลักษณะที่นำมาพิจารณา

S คือ ข้อมูลทั้งหมดที่นำมาพิจารณา

$Gain(S,A)$ คือ ค่าการเพิ่มขึ้นของสารสนเทศซึ่งแสดงถึงความสามารถในการจำแนก ข้อมูลของคุณลักษณะ A บนข้อมูล S

$v \in Value(A)$ คือจำนวนข้อมูลที่มีค่าของคุณลักษณะ A บนข้อมูล v

$$Entropy(S) \equiv -p_+ \log_2 p_+ - p_- \log_2 p_- \qquad (2)$$

$Entropy(S)$ คือค่าความสับสนของข้อมูล S

เมื่อ

p_+ คือ จำนวนตัวอย่างที่อยู่ในคลาสบวก

p_- คือ จำนวนตัวอย่างที่อยู่ในคลาสลบ

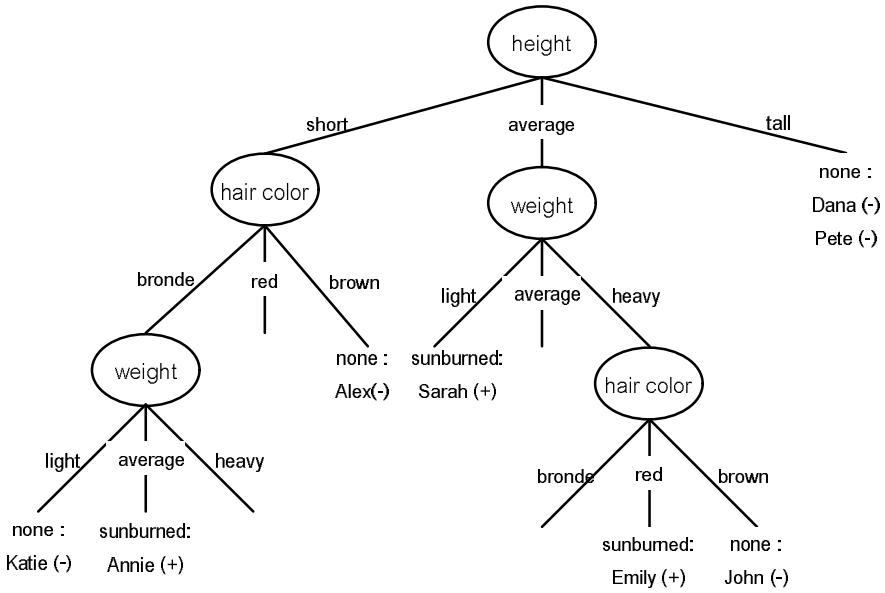
การแทนข้อมูลต้นไม้ตัดสินใจ

วิธีการแทนค่าข้อมูลผลลัพธ์จากการเรียนรู้ของต้นไม้ ประกอบด้วย

1) โหนดภายใน (Internal Node) คือ คุณลักษณะต่างๆ ของข้อมูล เมื่อข้อมูลใดตกลงมาที่โหนดจะใช้คุณลักษณะข้อมูลนี้เป็นตัวตัดสินใจว่าข้อมูลจะไปในทิศทางใด โดยโหนดภายในที่เป็นจุดเริ่มต้นของต้นไม้เรียกว่า "โหนดราก" (Root Node)
2) กิ่ง (Banch หรือ Link) คือ ค่าคุณลักษณะของคุณลักษณะข้อมูลของโหนดภายในที่แตกกิ่งนี้ออกมาโหนดภายในจะแตกกิ่งเป็นจำนวนเท่ากับจำนวนคุณลักษณะของโหนดภายในนั้น
3) โหนดใบ (Leaf Node) คือ กลุ่มต่างๆ ซึ่งเป็นผลลัพธ์สำหรับการจำแนกข้อมูลหรือการแบ่งกลุ่มข้อมูล

ตารางที่ 1 ข้อมูลตัวอย่างสอนของปัญหาการอาบแดด

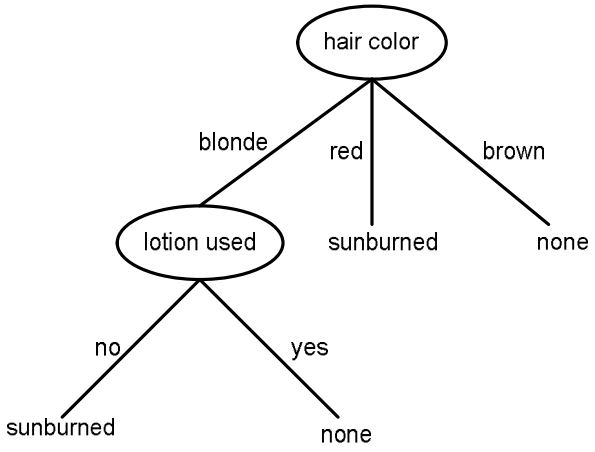
Name	Hair	Height	Weight	Lotion	Result
Sarah	blonde	average	light	no	sunburned
Dana	blonde	tall	average	yes	none
Alex	brown	short	average	yes	none
Annie	blonde	short	average	no	sunburned
Emily	red	average	heavy	no	sunburned
Pete	brown	tall	heavy	no	none
John	brown	average	heavy	no	none
Katie	blonde	short	light	yes	none



ภาพที่ 3 ต้นไม้ตัดสินใจที่สอดคล้องกับข้อมูลตัวอย่างสอนของปัญหาการอาบแดด

2.1.3 การเปลี่ยนต้นไม้เป็นกฎ

ระบบปัญญาประดิษฐ์ส่วนใหญ่ใช้การแทนความรู้ในรูปของกฎ เราสามารถเปลี่ยนต้นไม้ให้อยู่ในรูปของกฎเพื่อใช้ในกรณีที่ระบบของเราใช้การแทนความรู้ของกฎเป็นหลัก วิธีการเปลี่ยนต้นไม้เป็นกฎ [8] เป็นการแทนกฎจากต้นไม้ตัดสินใจในรูปของกฎ "ถ้า..แล้ว.." หรือ “IF THEN” เป็นการแสดงทุกเส้นทางเริ่มต้นจากโหนดรากไปยังโหนดใบ และนำค่าของแต่ละคุณลักษณะเชื่อมกันด้วย AND



ภาพที่ 4 ตัวอย่างต้นไม้ตัดสินใจ

จากตัวอย่างต้นไม้ตัดสินใจ ใน ภาพที่ 3 ต้นไม้ตัดสินใจดังกล่าวจำแนกผู้คนที่เกิด และไม่เกิดอาการผิวไหม้จากการอาบแดด สามารถเปลี่ยนเป็นกฎ IF THEN ได้ดังนี้

1) IF hair color is blonde AND lotion used is no THEN person gets sunburned.
2) IF hair color is blonde AND lotion used is yes THEN nothing happens.
3) IF hair color is red THEN the person gets sunburned.
4) IF hair color is brown THEN nothing happens.

2.1.4 ชนิดของข้อมูล

ข้อมูลถูกอธิบายด้วยคุณลักษณะต่างๆ [9] ที่เป็นลักษณะทางกายภาพ หรืออธิบายเวลาที่วัตถุหรือข้อมูลนั้น ทั้งนี้ในการทำเหมืองข้อมูลมีชื่อเรียกคุณลักษณะหลากหลายแบบ อาทิเช่น ตัวแปร (Variable) คาเรคเตอริสติก (Characteristic) ฟิลด์ (Field) คุณลักษณะ (Feature) หรือ มิติ (Dimension) เป็นต้น

คุณลักษณะข้อมูลหมายถึงคุณสมบัติเพื่อใช้ในการอธิบายข้อมูลหรือวัตถุหนึ่งๆ ว่ามีความแตกต่างจากวัตถุอื่น เช่นสีตาของแต่ละคน

ก่อนที่จะกล่าวถึงชนิดของคุณลักษณะ ในที่นี้จะอธิบายคุณสมบัติหรือตัวดำเนินการ (Operations) ในเชิงจำนวนของคุณลักษณะข้อมูล ซึ่งแบ่งออกเป็น 4 ลักษณะ

- ความแตกต่างกัน (Distinctness) ตัวดำเนินการที่สอดคล้องคือ เท่ากับ (=) และไม่เท่ากับ (≠)
- เรียงลำดับ (Order) ตัวดำเนินการของคุณสมบัตินี้ได้แก่ มากกว่า (>) มากกว่าเท่ากับ (≥) น้อยกว่า (<) และน้อยกว่าเท่ากับ (≤)
- การเพิ่มหรือลดค่า (Addition) ตัวดำเนินการที่สอดคล้องกับคุณสมบัตินี้ได้แก่ การบวก (+) และการลบ (–)
- ทวีคูณ (Multiplication) ตัวดำเนินการที่เกี่ยวข้องกับคุณสมบัตินี้ คือ การคูณ (*) และการหาร (/)

คุณลักษณะของข้อมูลมีหลายรูปแบบ และแต่ละแบบสามารถใช้ตัวดำเนินการต่างชนิดกัน การรู้จักชนิดของคุณลักษณะเป็นสิ่งจำเป็นในการทำเหมืองข้อมูล ชนิดของคุณลักษณะข้อมูลแบ่งออกเป็นสองชนิดหลักๆ ได้แก่ เชิงคุณภาพ (Qualitative) และเชิงปริมาณ (Quantitative) ซึ่งแต่ละแบบก็สามารถจำแนกออกเป็นแบบย่อยๆ เช่นคุณลักษณะเชิงคุณภาพ อาจเป็นเลขนอร์มินอล (Nominal) ตัวอย่างข้อมูลเช่นรหัสนิสิต รหัสไปรษณีย์ หรือเป็นตัวเลขที่เป็นลำดับ (Ordinal)

เช่นขนาดของเสื้อผ้า ใหญ่ กลาง เล็ก ส่วนคุณลักษณะข้อมูลที่เป็นเชิงปริมาณ จำแนกเป็นข้อมูลที่จัดแบ่งเป็นช่วง (Interval) และเป็นสัดส่วน (Ratio)

นอกจากนี้คุณลักษณะข้อมูลอาจถูกอธิบายได้ด้วยจำนวนของข้อมูลที่เป็นไปได้ โดยแบ่งออกเป็นสองแบบดังนี้

- คุณลักษณะข้อมูลแบบต่อเนื่อง (Continuous attribute) ประกอบด้วยข้อมูลซึ่งเป็นเลขจำนวนจริง แทนได้ด้วยเลขทศนิยม เช่นอุณหภูมิของภูมิอากาศ ความสูง หรือน้ำหนัก เป็นต้น
- คุณลักษณะข้อมูลแบบไม่ต่อเนื่อง (Discrete attribute) ประกอบด้วยข้อมูลที่นับได้ สามารถจัดเป็นกลุ่มหรือหมวดหมู่ได้ แทนได้ด้วยเลขจำนวนเต็ม เช่น รหัสของบุคคล (เลข 13 หลัก) คุณลักษณะข้อมูลแบบนี้อาจเป็นแบบไบนารีก็ได้ (Binary attribute) ซึ่งมีค่าที่เป็นไปได้เพียงสองค่า เช่น 0 หรือ 1 จริงหรือเท็จ เพศชายหรือเพศหญิง เป็นต้น

2.2 เอกสารและงานวิจัยที่เกี่ยวข้อง

2.2.1 งานวิจัยที่เกี่ยวกับการลดจำนวนกฎและการเพิ่มความเสถียรภาพ

Lemuel R. Waitman, Douglas H. Fisher และ Paul H. King [10] ได้อธิบายวิธีการรวมกฎบนบูทแสตรปจำนวนมากที่ซ้ำกันของกฎการอุปนัย (Induction Rule) เพื่อลดจำนวนของกฎที่นำเสนอแก่นักวิเคราะห์ เพื่อวัดและเพิ่มความเสถียรภาพของกระบวนการของกฎการอุปนัยและกำหนดการวัดความแปรปรวนของขอบเขต สำหรับขอบเขตการตัดสินใจของคุณลักษณะที่ต่อเนื่องกันและความแม่นยำในการประมาณค่าแบบจุดอย่างแม่นยำ การวัดความคล้ายกันระหว่างกฎหลาย ๆ กฎถูกนำเสนอเป็นพื้นฐานของการวัดขนาดแบบหลายมิติ ไปยังการแสดงภาพของกฎที่คล้ายกัน ในงานวิจัยนี้ได้นำไปปรับใช้กับข้อมูลการผ่าตัดและชุดข้อมูลไทรอยด์ (thyroid) จาก UCI

กฎการอุปนัยตรงกับวัตถุประสงค์ของงานนี้เพราะว่ากฎการเหนี่ยวนำ จะมุ่งเน้นไปที่ตัวอย่างบวกซึ่งใช้แทนเหตุการณ์ที่น่าแปลกใจบางอย่างที่ต้องการสังเกตุ ซึ่งเป็นสิ่งตรงกันข้ามกับการจำแนกของตัวอย่างบวกและตัวอย่างลบด้วยตัวจำแนก เช่น ต้นไม้ตัดสินใจ ซึ่งการจำแนกข้อมูลทั้งหมด โดยทั่วไปแล้วจะพิจารณาผลลัพธ์ทีทดแทนกันได้ หากเสียงส่วนใหญ่ของตัวอย่างเป็นลบ ตัวจำแนกอาจจะสร้างเพื่อเพิ่มประสิทธิภาพการจำแนกทั้งหมดที่ค่าใช้จ่ายของการสร้างกิ่งที่แยกความผิดปกติ แม้ว่ากฎการอุปนัยนำเสนอจุดเริ่มต้นที่ดีสำหรับการวิเคราะห์ข้อมูลก่อนการผ่าตัดและเหมือนโดเมนที่คล้ายกัน ระบบในปัจจุบันยังมีข้อบกพร่อง

งานวิจัยนี้พยายามที่จะลดจำนวนของกฎที่จำเป็นจะต้องตรวจสอบโดยนักวิเคราะห์ เพื่อจะกำหนดค่าความแปรปรวนให้กับขอบเขตการตัดสินใจและการประมาณค่าแบบจุด และเพื่อวัดและปรับปรุงเสถียรภาพของกระบวนการอุปนัยกฎ เพื่อให้บรรลุเป้าหมาย มีการใช้กฎการอุปนัยซ้ำๆกัน ด้วยการใช้วิธีการบรูท (Brute) กับเซตย่อยของข้อมูลที่ต่างกันแต่ซ้อนทับกัน และนามธรรมที่เกิดขึ้นกับการทดลองการอุปนัย หลายๆการทดลอง ใช้บูทแสตรป เป็นพื้นฐานของการทดลองการอุปนัยกฎ หลายๆ การทดลองแม้ว่ารูปแบบการตรวจสอบไขว้ (cross validation) สามารถนำไปปรับใช้เพื่อวัตถุประสงค์นี้เช่นกัน

สำหรับคุณลักษณะที่ไม่ต่อเนื่องกัน ค่าของคุณลักษณะจำเป็นจะต้องเหมือนกัน สำหรับคุณลักษณะสามารถมีค่าที่แตกต่างกันได้ ยกตัวอย่างเช่น

IF CPTCode = 29 AND Height < 164 AND HeartRateVariability >= 28.6
THEN NauseaGreaterThanMild = yes

ที่มีกฎเกือบจะเหมือนกับกฎที่สอง

IF CPTCode = 29 AND Height < 158 AND HeartRateVariability >= 25.7
THEN NauseaGreaterThanMild = yes

แต่ไม่เหมือนกับกฎที่สาม

IF CPTCode = 29 AND Height >= 140 AND HeartRateVariability >= 21.3
THEN NauseaGreaterThanMild = yes

และไม่เหมือนกับกฎที่สี่

IF CPTCode <> 27 AND Height < 164 AND HeartRateVariability >= 31.5
THEN NauseaGreaterThanMild = yes

ถ้ากฎที่เกือบจะเหมือนกันนี้มีซ้ำกันหลาย ๆ กฎ กฎสรุปที่ถูกสร้างขึ้นก็จะเห็นค่าเบี่ยงเบนเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของความถูกต้องและความครอบคลุมของการขยายลาปาสบนบูตสแตรปของกฎสรุป (เปอร์เซ็นต์ของกรณีที่มีความพึงพอใจของกฎก่อนหน้านั้น) และจำนวนกฎที่ซ้ำกันที่ประกอบด้วยกฎที่เกือบจะเหมือนกันที่สนับสนุนกฎสรุป กฎสรุปจะประกอบไปด้วย สถิติพื้นฐานเกี่ยวกับความแปรปรวนของคุณลักษณะที่ต่อเนื่องกันที่รวมอยู่ในกฎ

การตีความของกฎที่คล้ายคลึงกันสามารถระบุโดยใช้การวัดความคล้ายคลึงกัน ซึ่งถูกใช้ในการวิเคราะห์จัดกลุ่มวิศวกรรมและการปรับขนาดหลายมิติที่ใช้กันทั่วไปในด้านจิตวิทยา การวิเคราะห์แบบหลายมิติ จะใช้ในการแสดงผลที่ซับซ้อน N มิติ เป็นกราฟ 1,2 3 มิติ พิจารณากฎต่อไปนี้ และสมมุติช่วงอายุจาก 10-93 ปี

A. Weight < 70 THEN NauseaVomit = Significant

B. Weight < 70 AND 10 < AGE < 15 THEN NauseaVomit = Significant

C. Weight < 70 AND 10 < AGE < 50 THEN NauseaVomit = Significant

ความคล้ายคลึงกัน ระหว่าง A และ C มีค่ามากกว่าความคล้ายคลึงกันระหว่าง A และ B ทั้งนี้เป็นเพราะว่าไม่มีข้อจำกัดเรื่องอายุในกฎ A ดังนั้นช่วงอายุของกฎ C จาก 10 ถึง 50 ใกล้เคียงกับช่วงของกฎ A จาก 10 ถึง 93กว่า ช่วงของกฎ B จาก 10 ถึง 15

หลังจากที่ กฎสรุปได้ถูกสร้างขึ้น จะถูกดึงขึ้นมาเพื่อตรวจสอบระดับที่ต่ำที่สุดของการสนับสนุน ซึ่งถูกระบุไว้เพื่อจำกัดจำนวนของกฎสรุปที่ปรากฏ ระดับของการสนับสนุนเป็นจำนวนซ้ำกันที่เก็บไว้เพื่อสนับสนุนกฎพื้นฐาน ทั้งนี้จะขึ้นอยู่กับชุดของข้อมูลและจำนวนกฎที่เจอนักวิเคราะห์อาจจะต้องยกตัวอย่างเช่น ให้ความสำคัญกับกฎที่เกิดขึ้นซ้ำกันทั้งหมด 10 ครั้ง กฎสรุปตัวอย่าง ที่ถูกแสดงไว้ด้านล่าง คุณลักษณะที่ต่อเนื่องกันจะตามด้วยส่วนเบี่ยงเบนมาตรฐานที่อยู่ในวงเล็บ อัตราส่วนออด (odds-ratio) ซึ่งเป็นอัตราส่วนของกฎของความถูกต้องขยายลาปาสบนบูตสแตรป กับความแพร่หลายของผลที่ตามมาสำหรับชุดข้อมูลทั้งหมดถูกแสดงไว้ในวงเล็บในบรรทัดต่อมา พร้อมกับ ความถูกต้องและความครอบคลุมของกการขยายลาปาสบนบูตสแตรป อัตราส่วนออด บ่งชี้ว่า มีจำนวนครั้งเท่าไหร่ ที่จะเกิดผลลัพธ์สำหรับกลุ่มตัวอย่างซึ่งสอดคล้องกับการเรียงลำดับของกฎที่มาก่อนหลังเมื่อเทียบกับข้อมูลทั้งหมด

IF CPT = 29 AND HeartRateVariability >= 28.6 (4.5) AND Height < 164

(6.0) THEN NauseaGreaterThanMild = yes

(2.87x as likely) Accuracy: 51.2 (10.9), Coverage: 5.9 (1.1), 10/10

กฎสรุปที่เก็บคุณลักษณะที่ต่อเนื่องกันที่เหมือนกัน 2 ครั้งถูกทำการทดสอบทางสถิติโดยการกรองกฎที่ด้อยกว่าออก จากตัวอย่าง ทำการพิจารณากฎที่ระบุช่วงตัวแรก Height โดยระบุขอบเขตบนและล่าง

IF Height >= 158.6 (1.2) AND Height < 162.9 (0.7) AND Phase2Recovery < 29.4

(6.24) THEN NauseaGreaterThanMild = Yes

(2.89x as likely) Accuracy: 51.7 (8.0), Coverage: 6.6 (2.6), 10/10

จากตัวอย่าง Height เป็นข้อจำกัดไปที่ช่วงแคบซึ่งไม่คาดหวังในตัวอย่างนี้ กฎที่เกือบจะเหมือนกันที่เกิดขึ้นในการทำซ้ำทั้งหมด 10 ครั้ง และกฎสรุปที่สร้างได้เป็นดังนี้

IF Height >= 158.6 (1.2) AND Height < 162.9 (0.7) AND Phase2Recovery < 29.4

(6.24)

THEN NauseaGreaterThanMild = Yes

(2.89x as likely) Accuracy: 51.7 (8.0), Coverage: 6.6 (2.6), 10/10

ในกรณีนี้ อาจจะมีความกังวลว่าระยะทางระหว่าง 2 ขอบเขตที่มีขนาดเล็กแต่ค่าเบี่ยงเบนมาตรฐานขนาดเล็กแสดงให้เห็นว่ากฎสรุปนั้นสำคัญ ในกรณีนี้พวกเขาจะเลือกที่จะรักษากฎ แต่โดยทั่วไปแล้วพวกเขาต้องการที่กำจัดกฎที่ช่วงที่ระบุไว้มีความหมายเนื่องจาก ค่าเบี่ยงเบนมาตรฐานขนาดใหญ่สัมพันธ์ไปยังระยะทางระหว่างขอบบนและขอบล่าง

หลังจากการสร้างกฎการอุปนัย และกฎสรุป นักวิเคราะห์ต้องการที่จะเปรียบเทียบกฎเพื่อตรวจสอบว่า ถ้ากฎการอุปนัยได้ค้นพบกฎจากพื้นที่ที่แตกต่างของพื้นที่ของปัญหาหรือบางรูปแบบย่อยจากส่วนเล็กๆของพื้นที่ของปัญหาจากตัวอย่างกฎที่เกี่ยวข้องกับผลลัพธ์ที่อาจจะเกี่ยว กับอายุและน้ำหนัก นักวิเคราะห์อาจจะเปรียบเทียบกฎที่มีผลลัพธ์ต่างกันเพื่อดูว่ามันมีคุณลักษณะเดี่ยวกันเกี่ยวข้องหรือไม่

งานวิจัยนี้สามารถสรุปได้ว่า กฎการอุปนัยนั้นเหมาะสมกับโดเมนของปัญหาที่เกิดขึ้นจากหลายความเสี่ยงและเหตุการณ์จำนวนมาก การเชื่อมกฎสามารถเข้าใจได้ง่ายแต่ว่าขอบเขตที่แน่นอนของกฎ ความไม่แน่นอนของอัลกอริทึมที่เกี่ยวข้องกับขอบเขตเหล่านี้และจำนวนกฎที่ไม่เกี่ยวข้องทำให้เกิดข้อสงสัยสงสัยจากผู้เชี่ยวชาญโดเมน กฎสรุปแสดงให้เห็นถึงความแปรปรวนของขอบเขตที่เกี่ยวข้องที่ถูกคาดหวังโดยผู้เชี่ยวชาญโดเมน และการลดจำนวนกฎที่จะต้องถูกวิเคราะห์ การวัดแบบหลายมิติเป็นเครื่องมือที่มีค่าสำหรับการตรวจสอบ เซตของกฎการอุปนัย

2.2.2 การสร้างต้นไม้ตัดสินใจต้นเดียวแทนกลุ่มของต้นไม้ตัดสินใจ

Anneleen Van Assche และ Hendrik Blockeel ได้นำเสนอวิธีการเรียนรู้แบบจำลองที่ตีความได้แบบเดียวจากเอนเซมเบิล โดยไม่จำเป็นต้องสร้างจากข้อมูลจำลองเป็นแบบจำลองใหม่ที่สามารถทำให้เข้าใจกระบวนการตัดสินใจและให้ความแม่นยำมากกว่าการเรียนรู้โดยตรงกับข้อมูลเพียงอย่างเดียว โดยมุ่งหวังที่จะผ่านพ้นปัญหาความเข้าใจโดยใช้การเรียนรู้จากต้นไม้ตัดสินใจต้นเดียวจากเอนเซมเบิล ที่สร้างด้วยแบกกิง ซึ่งคำนวณค่าฮิวริสติก จากเอนเซมเบิล เพื่อตัดสินว่าการทดสอบไหนจะถูกใช้ในต้นไม้ใหม่ มีการกำหนดเงื่อนไขการหยุด เพื่อหลีกเลี่ยงต้นไม้ที่มีขนาดใหญ่ โดยกำหนดเงื่อนไขการหยุดที่ไม่สมมูลย์กัน เพื่อสร้างต้นไม้ให้เข้าใจได้มากขึ้น มาประยุกต์ใช้เพื่อหลีกเลี่ยงการแบ่งกิ่งที่ซ้ำซ้อนกันที่นอกเหนือจากวิธีการตัดกิ่งขณะเรียนรู้ และการตัดแต่งกิ่งหลังการเรียนรู้ ซึ่งประเมินการทดลองด้วยชุดข้อมูล UCI จำนวนมาก และเปรียบเทียบกับวิธีการเดิมด้วย (Bagging ที่ใช้ J48 (Weka's C4.5)) ที่สร้างด้วยข้อมูลจำลอง เพื่อ

เปรียบเทียบความแม่นยำ ความซับซ้อนและความเสถียรภาพกับต้นไม้ต้นเดียวต้นฉบับและตัวจำแนกของวิธีการ lsm (lsm_t, lsm_td, lsm_d) ที่ใช้เงื่อนไขการหยุดที่ใช้แค่ชุดข้อมูลสอนเท่านั้น และเปรียบเทียบกับต้นไม้ต้นเดียวที่ได้มากจากการใช้วิธี CMM ตามที่ Domingos ได้อธิบายไว้ สร้างต้นไม้ที่ตัดเล็มกิ่ง CMM(p) และต้นไม้ที่ไม่ตัดกิ่ง CMM(up) ที่ใช้ข้อมูลจำลองที่สร้างขึ้น และสามารถสรุปได้ว่าความแม่นยำในการทำนายของ lsm สามารถให้การตัดสินใจที่จำเป็นสำหรับกำหนดป้ายกำกับ (label) ตัวอย่างได้เหมือนกันที่แบกกิง ทำ ในส่วนของความเข้าใจของแบบจำลอง สามารถสรุปได้ว่าสามารถตีความต้นไม้ด้วย lsm เช่นที่สามารถตีความต้นไม้ที่ได้จาก J48 และในส่วนของความเสถียรภาพ CMM มีค่าเกน เสถียรภาพมากที่สุด

การสร้างต้นไม้ต้นเดียวเพื่อแทนกลุ่มของต้นไม้ สมมุติว่า E เป็นกลุ่มของต้นไม้ตัดสินใจ N ต้น ที่ต้องการแทนด้วยต้นไม้ตัดสินใจต้นเดียว เอนเซมเบิล E ให้ค่าทำนายไปยังตัวอย่างใหม่ x ขึ้นอยู่กับการรวมการทำนายของต้นไม้พื้นฐานแต่ละต้น ในกรณีนี้ค่าเฉลี่ย

$$L_E(x) = argmax_{C_i}\left(\frac{1}{N}\sum_{k}^{N} P_k(C_i \mid x)\right) \tag{3}$$

สำหรับการคำนวณค่าฮิวริสติก จากเอนเซมเบิล เพื่อให้ความเข้าใจกับเอนเซมเบิล ตัวทดสอบที่ถูกใส่ในโหนดของต้นไม้ใหม่ควรจะเป็นโหนดที่มีความรู้(ข้อมูล)มากที่สุด ณ ขณะนั้น ขึ้นอยู่กับเอนเซมเบิล จุดประสงค์คือ การประมาณแนวคิดในเอนเซมเบิล เพราะฉะนั้นการกระจายตัวของคลาสที่ถูกทำนายโดยเอนเซมเบิล จะถูกใช้และไม่ใช่ การกระจายตัวที่มีอยู่โดยตรงในข้อมูลสอน ซึ่งต้องการที่จะคำนวณ ว่าการทดสอบไหนที่มีค่าเกนสารสนเทศสูงที่สุด ที่ขึ้นอยู่กับเอนเซมเบิลในโหนด n ของต้นไม้ต้นใหม่ สมมุติว่า B เป็นตัวเชื่อมของการทดสอบ ที่เกิดขึ้นตามเส้นทางจากโหนดราก จนถึงโหนด n ดังนั้น สูตรปกติของค่าเกนสารสนเทศ IG สำหรับการทดสอบ T ใน n คือ

$$IG(T \mid B) = entropy(B) - P(T \mid B)entropy(T \wedge B) - P(\neg T \mid B)entropy(\neg T \wedge B) \tag{4}$$

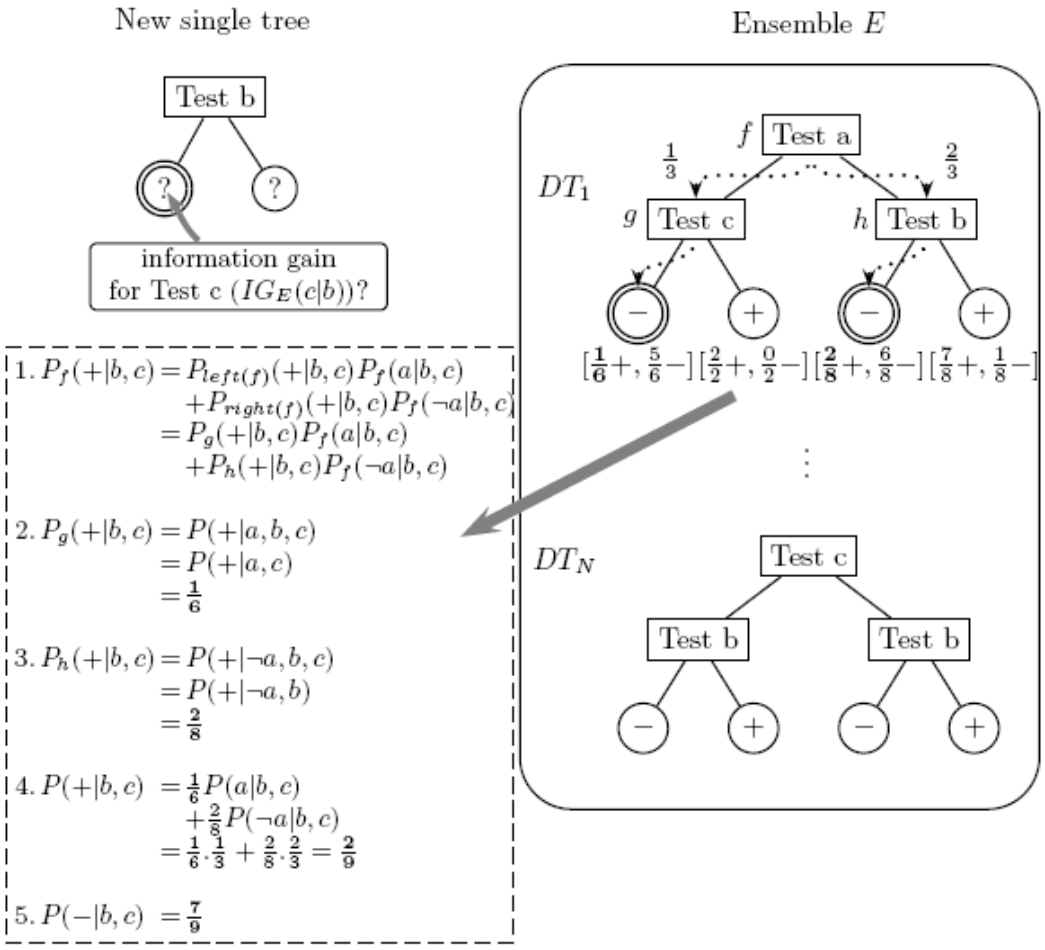
เมื่อ

$$entropy(A) = \sum_{i=1}^{c} - P(C_i \mid A)\log_2 P(C_i \mid A) \tag{5}$$

โดยที่ c คือจำนวนรวมของคลาสทั้งหมด คือคลาสที่ i และ A เป็นเซตใดๆ ของเงื่อนไข

ต้นไม้ตัดสินใจที่สร้างแบบจำลองประเภทการกระจายในข้อมูล สามารถที่จะใช้การประมาณ สมมุติว่ามีต้นไม้ตัดสินใจ เราสามารถประมาณ ด้วยการเพิ่มจำนวนผ่านต้นไม้ และประยุกต์ใช้กฎของความน่าจะเป็นทั้งหมดในแต่ละโหนดจนกระทั่งถึงใบ ดังนั้นจะได้

$$P_k(C_i|A) = \sum_{leaves\, l_{kj}\, in\, DT_k} P(C_i|Y_{kj} \wedge A)P(Y_{kj}|A) \tag{6}$$



ภาพที่ 5 การคำนวณค่าฮิวริสติกจากต้นไม้ในเอนเซมเบิลเพื่อสร้างเป็นต้นไม้ใหม่ต้นเดียว

จากงานวิจัยที่ได้กล่าวมาข้างต้นนี้ ได้แสดงให้เห็นประสิทธิภาพและประสิทธิผลของการเรียนรู้จากต้นไม้ตัดสินใจต้นเดียวได้ว่ามีความแม่นยำน้อยกว่าการเรียนรู้จากข้อมูลจำลอง แต่ให้ความเสถียรภาพมากกว่าการเรียนรู้จากข้อมูลจำลอง

2.2.3 งานวิจัยที่เกี่ยวกับการรวมเอาต์พุตจากต้นไม้ตัดสินใจหลายต้นแทนที่ด้วยต้นไม้เพียงต้นเดียว

Giovanni Seni, Edword Yang และ Said Akar [11] ได้นำเสนอ แอปพลิเคชันของอัลกอริทึมการจำลองทางสถิติที่เรียกว่า Rule Ensemble ให้กับปัญหาการแสดงลักษณะการสูญหายของข้อมูล โดยแสดงให้เห็นว่าสามารถทำนายได้ดีพอๆ กับวิธีที่ดีที่สุดเพื่อให้เห็นภาพของ

ประโยชน์จากวิธีการนี้ จึงได้ทำการวิเคราะห์ข้อมูลของอุตสาหกรรมเซมิคอนดัคเตอร์เป็นตัวอย่าง ได้นำเสนอวิธีการค้นหาข้อมูลแบบไม่เชิงเส้น (non-linear) ของต้นไม้ตัดสินใจที่ได้รับการพิสูจน์แล้วว่ามีประโยชน์กับปัญหาการแสดงลักษณะการสูญหายของข้อมูลที่สามารถจัดการข้อมูลที่ผสมกันและข้อมูลที่สูญหายได้และนำเอาต้นไม้ตัดสินใจแบบบูสต์ที่มีความคิดพื้นฐานการรวมเอาต์พุต จากต้นไม้หลายๆ ต้นเพื่อสร้างการตัดสินใจที่ดี โดยรวมต้นไม้เล็กๆ จำนวนมากเข้าด้วยกันแบบเชิงเส้น และแทนที่ด้วยต้นไม้เพียงต้นเดียว เสนอการประยุกต์ใช้การรวมกฎที่ได้จากแบบเอนเซมเบิล กับเครื่องมือในการตีความถูกนำมาใช้ในการพัฒนากับเซตของข้อมูล ใช้ RuleFit ในการสร้างการทดลอง หาว่าเครื่องจักรตัวใดในขั้นตอนการผลิตที่เป็นผลให้เกิดการสูญหายของผลิตภัณฑ์มากที่สุด สรุปความผิดพลาดของการทดสอบโดยการประมาณผลลัพธ์ของข้อมูลในระดับ ล็อต/เวเฟอร์ โดยใช้ 3 วิธีคือเอเซมเบิล, การใช้กฎตัวแปรเดียวในเอนเซมเบิล และต้นไม้ต้นเดียวที่สร้างด้วยวิธีแบบ CART ใช้วิธีบูทแสตรป ในการคำนวนการกระจายของข้อมูล และให้ผลลัพธ์ออกมาว่า ความผิดพลาดโดยเฉลี่ยในการทำนายของแบบจำลองเอนเซมเบิล ของข้อมูลในระดับล็อต ที่มีข้อมูลในลักษณะกว้างมีค่าน้อยกว่าแบบจำลองอื่นมาก ส่วนข้อมูลในระดับเวเฟอร์แบบจำลองเอนเซมเบิลแสดงให้เห็นว่าพัฒนาได้ดีขึ้นอีกทั้งยังช่วยพัฒนาความผิดพลาดของแบบจำลองต้นไม้ต้นเดียวให้ดีขึ้นอีกด้วย

2.2.4 งานวิจัยที่เกี่ยวกับการรวมผลลัพธ์ของต้นไม้ตัดสินใจหลายต้นและวิธีการโหวตผลลัพธ์ของต้นไม้ตัดสินใจ

Zhi-Hua Zhou และ Wei Tang ได้ทำการรวมผลลัพธ์ของต้นไม้ตัดสินใจหลายต้นโดยได้ยกตัวอย่างอัลกอริทึม Gasen-b มาอธิบายพร้อมทั้งเปรียบเทียบผลลัพธ์ของอัลกอริทึมนี้กับการทำนายของต้นไม้ตัดสินใจหลายต้นที่ถูกรวมด้วยวิธีอื่น โดยอธิบายว่า การรวมผลลัพธ์เป็นกระบวนการที่ใช้ตัวเรียนรู้หลายตัวซึ่งมีข้อมูลสอนและอัลกอริทึมการเรียนรู้เหมือนกัน ผลทำนายที่ได้จะถูกรวบรวมเพื่อจัดการในรอบต่อไป วิธีการรวมผลลัพธ์นี้จัดว่ามีความถูกต้องแม่นยำกว่าการใช้ตัวเรียนรู้เพียงตัวเดียว และได้รับการยอมรับกันอย่างกว้างขวาง ตัวอย่างงานที่ใช้ในการรวมผลลัพธ์ เช่น การสแกนตัวอักษรแบบ Optical Character Recognition (OCR), การจดจำใบหน้า (Face Recognition), การวิเคราะห์ตัวยาทางการแพทย์ต่างๆ

แต่ผลงานวิจัยในปัจจุบันได้อธิบายไว้ว่าหากตัวเรียนรู้เป็นนิวรอลเน็ตเวิร์กและมีโครงสร้างเป็นต้นไม้ตัดสินใจแล้ว การรวมผลลัพธ์จากตัวเรียนรู้เพียงบางตัว ทำให้ได้ผลลัพธ์ที่แม่นยำกว่าการรวมผลลัพธ์จากตัวเรียนรู้ทุกตัวที่มีอยู่ การทดสอบอัลกอริทึมที่เลือกรวมผลลัพธ์เฉพาะบางตัว

ได้ข้อสรุปว่า ไม่เพียงแต่เซตของผลลัพธ์ที่ได้จะมีขนาดเล็กกว่าเท่านั้น แต่ยังมีความแม่นยำมากกว่าอีกด้วย

อย่างไรก็ตาม การที่จะเลือกว่าตัวเรียนรู้ไหนควรจะเก็บมาคำนวณหรือตัวเรียนรู้ไหนควรจะตัดออกไปยังเป็นเรื่องที่ยากอยู่นัก ดังนั้น Zhi-Hua Zhou และ Wei Tang จึงได้ทำการทดลองโดยใช้อัลกอริทึมแบบ GASEN นำมาพัฒนาอัลกอริทึมนี้ให้มีประสิทธิภาพยิ่งขึ้น และตั้งชื่ออัลกอริทึมใหม่นี้ว่า GASEN-b ซึ่งจากผลการทดลองประสิทธิภาพของอัลกอริทึมนี้ ก็แสดงให้เห็นว่ามีประสิทธิภาพในการทำนายและมีความแม่นยำที่สุดวิธีหนึ่ง

Robert E. Banfield, Lawrence O. Hall, Kevin W. Bowyer และ W. Philip Kegelmeyer [12] ได้นำเสนอการเปรียบเทียบเทคนิคการสร้างวิธีโหวตผลลัพธ์ของต้นไม้ตัดสินใจโดยนำเอาอัลกอริทึมแบบแบกกิง และอัลกอริทึมอื่นๆ ซึ่งใช้การสุ่มเป็นพื้นฐานมาเปรียบเทียบกัน ได้ข้อสรุปที่น่าสนใจหลายอย่างเช่น วิธีการสับสเปซ ที่ใช้ได้ดีมากกับชุดข้อมูลขนาดใหญ่ แต่กับชุดข้อมูลเล็กๆ จะทำงานได้ไม่ดี หรือการสุ่มต้นไม้ หรือการสุ่มป่า เป็นเทคนิคที่ใช้ในการรวมผลลัพธ์ที่ได้โดยตรงเท่านั้นส่วน บูสต์ติง แบกกิง และการสุ่มสับสเปซสามารถใช้ในการเรียนรู้ของระบบโครงข่ายประสาทเทียมได้ด้วย

Thomas G. Dietterich [13] ได้นำเสนอการเปรียบเทียบการโหวตผลลัพธ์ของต้นไม้ตัดสิน ใจ ที่นำเอาวิธีการบูสต์ติง แบกกิงและการสุ่มมาเปรียบเทียบกันและให้ผลลัพธ์ว่า วิธีการแบบบูสต์ติง ได้ผลลัพธ์ที่ดีกว่าอัลกอริทึมอื่น

บทที่ 3
การออกแบบขั้นตอนการดำเนินงาน

ในบทนี้จะกล่าวถึงขั้นตอนการผสานกฎจากป่าแบบสุ่ม โดยมีขั้นตอนและหลักการดังต่อไปนี้

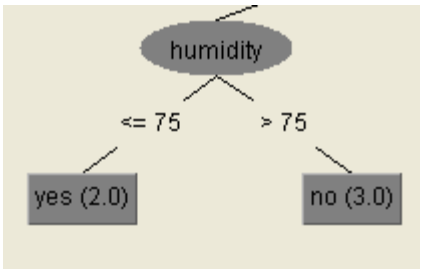
3.1 การจัดการกับรูปแบบของคุณลักษณะของข้อมูล

ค่าของข้อมูลในแต่ละคุณลักษณะแบ่งออกได้เป็น 2 แบบ ได้แก่

1. ค่าต่อเนื่อง (Continuose) ประกอบด้วยข้อมูลซึ่งเป็นเลขจำนวนจริง แทนได้ด้วยเลขทศนิยม เช่น ความสูง
2. ค่าไม่ต่อเนื่อง (Discrete)) ประกอบด้วยข้อมูลที่นับได้ สามารถจัดเป็นกลุ่มหรือหมวดหมู่ได้ แทนได้ด้วยเลขจำนวนเต็ม สามารถแบ่งออกเป็น
   2.1 แบบมีลำดับ (Ordinal) เช่น ขนาดของเสื้อผ้า ใหญ่ กลาง เล็ก
   2.2 แบบไม่มีลำดับ (Norminal) เช่น รหัสนิสิต รหัสไปรษณีย์

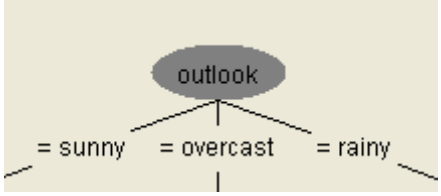
การสร้างกฎจากต้นไม้ตัดสินใจในงานวิจยนี้ สามารถเป็นไปได้ใน 2 รูปแบบ

1. แยกแยะด้วยช่วงของค่าข้อมุล (เฉพาะข้อมูลที่เป็นตัวเลข)



ภาพที่ 6 ต้นไม้ตัดสินใจบางส่วนที่แยกแยะด้วยช่วงของค่าข้อมูล

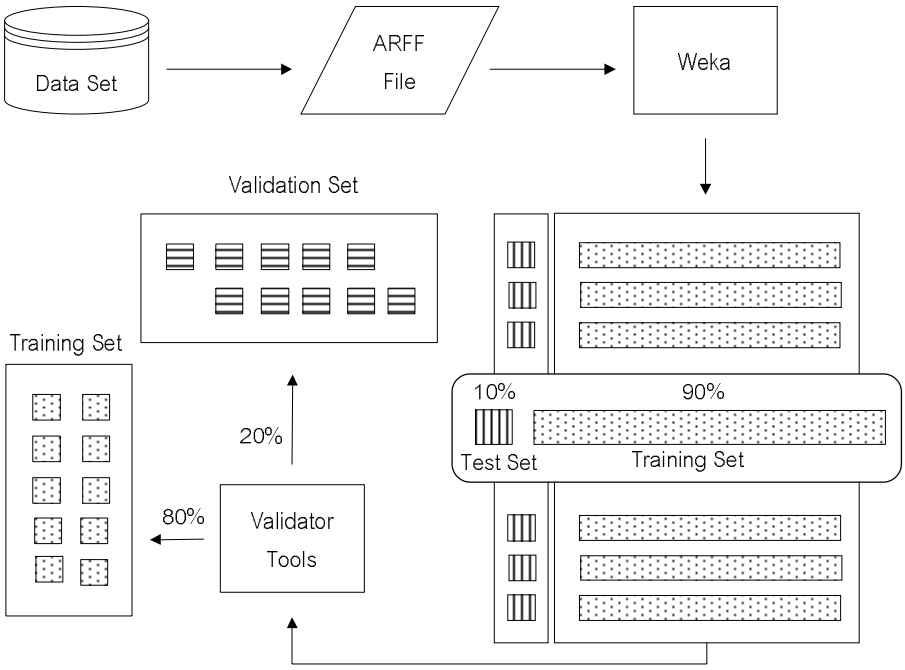
2. แยกแยะด้วยค่าเฉพาะของข้อมูลที่เป็นตัวเลขหรือประเภทที่ไม่สามารถวัดแบ่งเป็นช่วงได้



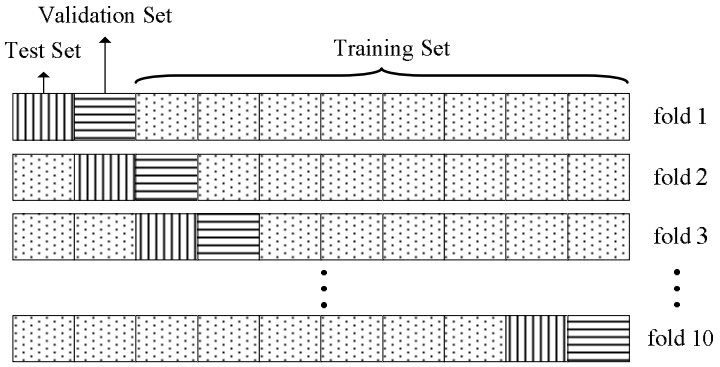
ภาพที่ 7 ต้นไม้ตัดสินใจบางส่วนที่แยกแยะด้วยค่าเฉพาะของข้อมูลที่ไม่สามารถแบ่งเป็นช่วงได้

3.2 การจัดเตรียมข้อมูลสำหรับทดลอง

ชุดข้อมูลที่ใช้ในการทดลองเป็นชุดข้อมูลจากฐานข้อมูล UCI Machine Learning Repository [14] จำนวน 7 ชุด ได้แก่ Balance Scale, Blood Transfusion, Haberman's Survival, Iris, Liver Disorders, Pima Indians Diabetes Database, Statlog และแบ่งข้อมูลสำหรับทดสอบด้วยโปรแกรมเวกา (WEKA) [15] แบบการตรวจสอบแบบไขว้กัน 10 ชุด (10-fold Cross Validation) ข้อมูลที่ใช้จะถูกแบ่งออกเป็น ชุดข้อมูลสอน (Training Set) 90% และชุดข้อมูลทดสอบ (Test Set) 10% ในชุดข้อมูลสอน 90% เราจะนำมาแบ่งอีกครั้งด้วยโปรแกรมพัฒนาขึ้นมาเอง เพื่อใช้เป็นชุดข้อมูลสอน 80% และอีก 20% ใช้เป็นชุดข้อมูลตรวจสอบ (Validation Set) ในการหาเซตของกฎใหม่



ภาพที่ 8 แผนผังการแบ่งชุดข้อมูลเพื่อใช้ทดสอบ

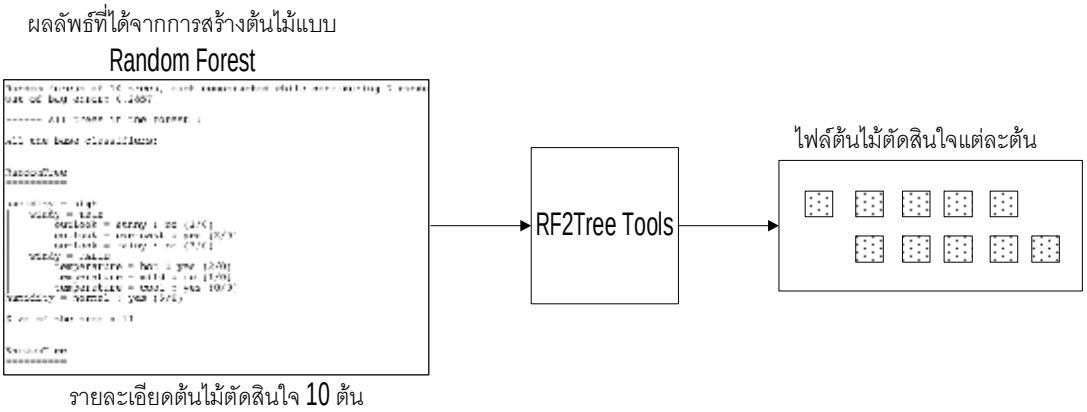


ภาพที่ 9 ตำแหน่งของชุดข้อมูลสอน. ชุดข้อมูลทดสอบ และชุดข้อมูลตรวจสอบ ด้วยวิธีการตรวจสอบแบบไขว้กัน 10 ชุด

3.3 การจัดเตรียมต้นไม้ตัดสินใจในการผสาน

เซตของต้นไม้ตัดสินใจที่นำมาผสานกฎ เป็นเซตของต้นไม้ตัดสินใจที่ได้จากการทำ ป่าแบบสุ่ม แต่เนื่องจากโปรแกรมเวกา ไม่แสดงผลลัพธ์ของการทำป่าแบบสุ่มออกมาเป็นข้อมูลต้นไม้ตัดสินใจ จึงจำเป็นต้องแก้ไขโปรแกรมให้สามารถ Visualize ต้นไม้แบบป่าแบบสุ่มได้

นำรายละเอียดของต้นไม้ตัดสินใจที่ได้จากการตัวจำแนกแบบป่าแบบสุ่มมาสร้างข้อมูลนำเข้าโดยใช้โปรแกรมที่พัฒนาขึ้นมาเอง (RF2Tree Tools : ภาคผนวก ค.) ให้มีรูปแบบตามที่ได้กล่าวไว้ข้างต้น



ภาพที่ 10 แผนผังการนำต้นไม้ตัดสินใจมากสร้างเป็นไฟล์ข้อมูลนำเข้า

3.4 การจัดเตรียมข้อมูลนำเข้า

ไฟล์ข้อมูลนำเข้าแต่ละบรรทัดจะระบุตำแหน่งของไฟล์ของต้นไม้ตัดสินใจแต่ละต้น ที่สร้างด้วยวิธีป่าแบบสุ่มด้วยโปรแกรมเวกา ก่อนบรรทัดสุดท้ายระบุตำแหน่งไฟล์ของชุดข้อมูลตรวจสอบ ซึ่งเป็นไฟล์ข้อมูลที่มีโครงสร้างแบบ ARFF และบรรทัดสุดท้ายของไฟล์ข้อมูลนำเข้าให้กำหนดไฟล์พาทของชุดข้อมูลทดสอบ

ก่อนทำการผสานกฎจากป่าแบบสุ่ม โปรแกรมจะทดสอบคำนวณหาเปอร์เซ็นต์ความถูกต้องของผลทำนายจากต้นไม้ตัดสินใจแต่ละต้น และจัดเรียงลำดับการผสานก่อนหลังตามวิธีที่กำหนดดังต่อไปนี้

1. เรียงลำดับจากต้นไม้ตัดสินใจที่มีเปอร์เซ็นต์ความถูกต้องของผลทำนายมาก
2. เรียงลำดับจากต้นไม้ตัดสินใจที่มีเปอร์เซ็นต์ความถูกต้องของผลทำนายน้อยที่สุด

3.5 ข้อมูลนำออก

ข้อมูลนำออกจะแสดงตำแหน่งของไฟล์ต้นไม้ตัดสินใจที่ทำการผสานกฎ จำนวนกฎ เปอร์เซ็นต์ความถูกต้องของผลทำนาย และแสดงรายละเอียดของกฎแต่ละข้อที่ได้หลังการผสานกฎ

3.6 การผสานกฎจากป่าแบบสุ่ม

ในการผสานกฎจากป่าแบบสุ่ม จะทำการพิจารณาต้นไม้ตัดสินใจครั้งละ 2 ต้น ซึ่งมีขั้นตอนดังต่อไปนี้

3.6.1 การลดเงื่อนไขที่ไม่จำเป็นออก

ในขั้นตอนนี้ เราจะทำการตัดเงื่อนไขที่ไม่จำเป็นออก เพื่อลดความซ้ำซ้อนของกฎทั้งหมด โดยจำกัดขอบเขตช่วงของข้อมูลให้มีความจำเพาะ (specific) มากที่สุดยกตัวอย่าง เช่น

```
IF weight>40 AND weight>70 AND weight>80 AND weight<150
AND height<180 THEN figure=fat
```

เราจะเห็นว่าเงื่อนไข weight>80 นั้น จำเพาะมากกว่า weight>40 และ weight>70 ดังนั้น weight>40 และ weight>70 จะถูกตัดออก ซึ่งสุดท้ายแล้วเราก็จะได้เซตของกฎใหม่ดังต่อไปนี้

IF weight>80 AND weight<150 AND height<180 THEN figure=fat

3.6.2 การจัดการกฎกับต้นไม้ตัดสินใจ

เนื่องจากเราทำการพิจารณาต้นไม้ทีละคู่ ดังนั้นเราจึงแยกวิธีการจัดการกฎไว้ดังนี้

1) กฎที่มีเงื่อนไขและผลลัพธ์เหมือนกันทุกประการ สำหรับกฎทุกคู่ที่เหมือนกันทุกประการ เราจะจัดการกฎโดยการตัดกฎให้เหลือไว้เพียงกฎเดียว ยกตัวอย่างเช่น

Rule 1: IF thickness=thin AND lace=glue THEN report=minor
Rule 2: IF thickness=thin AND lace=glue THEN report=minor
New Rule: IF thickness=thin AND lace=glue THEN report=minor

2) กฎที่มีเงื่อนไขเหมือนกันทุกประการ แต่ให้ผลลัพธ์ต่างกัน สำหรับกฎทุกคู่ที่มีเงื่อนไขเหมือนกันทุกประการแต่ให้ผลลัพธ์ต่างกัน เราจะจัดการกฎโดยการตัดกฎคู่นั้นออก ยกตัวอย่างเช่น

Rule 1: IF face=soft AND age>3 THEN toy=doll
Rule 2: IF face=soft AND age>3 THEN toy=elastic
New Rule: -

3) กฎที่มีเงื่อนไขคล้ายกันบางส่วนแต่ให้ผลลัพธ์เหมือนกัน สำหรับกฎทุกคู่ที่มีเงื่อนไขคล้ายกันบางส่วน แต่ให้ผลลัพธ์เหมือนกัน เราจะจัดการกฎโดยทำการตรวจสอบบนชุดข้อมูลตรวจสอนว่าการทำให้กฎมีนัยทั่วไปมากขึ้น (more general) หรือ ทำให้กฎมีความจำเพาะมากขึ้น (specific) แบบไหนที่ทำให้เปอร์เซ็นต์ความถูกต้องของผลทำนายเพิ่มขึ้น ยกตัวอย่าง การจัดการกฎหลังจากตรวจสอบการทำให้กฎมีนัยทั่วไปมากขึ้นนั้นให้เปอร์เซ็นต์ความถูกต้องของผลทำนายดีกว่าการทำให้กฎจำเพาะมากขึ้น เช่น

Rule 1: IF fur=short AND nose=yes AND tail=yes THEN type=bear
Rule 2: IF fur=short AND ear=yes AND nose=yes AND tail=yes THEN type=bear
New Rule: IF fur=short AND nose=yes AND tail=yes THEN type=bear

4) กฎที่มีเงื่อนไขแบบสามารถขยายช่วงของค่าข้อมูลได้ซึ่งมีค่าของข้อมูลที่พิจารณาตรงกันแต่เครื่องหมายชี้วัดทำให้ช่วงของค่าข้อมูลต่างกัน เราจะจัดการกฎโดยทำการตรวจสอบบนชุดข้อมูลตรวจสอบว่า ช่วงของค่าข้อมูลใดในเงื่อนไขของกฎทั้งคู่ที่เมื่อทำการปรับแล้วให้เปอร์เซ็นต์ความถูกต้องของผลทำนายดีกว่า ยกตัวอย่างเช่น

Rule 1: IF duty=recording AND period<3 AND period>1.5 THEN wage=1500
Rule 2: IF duty=recording AND period<2 AND period>1 THEN wage=1500
New Rule: IF duty=recording AND period<3 AND period>1 THEN wage=1500

5) กฎที่มีเงื่อนไขที่ช่วงของค่าข้อมูลซ้อนทับกัน ซึ่งให้ผลลัพธ์ต่างกัน เราจะจัดการกฎโดยการนำเอาช่วงของค่าข้อมูลที่ทับซ้อนกันออกจากกฎแต่ละข้อ ยกตัวอย่างเช่น

Rule 1: IF credit=yes AND money>20,000 THEN allow=yes
Rule 2: IF credit=yes AND money<40,000 THEN allow=no
New Rule 1: IF credit=yes AND money>=40,000 THEN allow=yes
New Rule 2: IF credit=yes AND money<=20,000 THEN allow=no
Rule 3: IF usage>100 AND payment=paid THEN promotion=false
Rule 4: IF usage>=200 AND usage<400 AND payment=paid THEN
promotion=true
New Rule 3: IF usage>100 AND usage<200 AND payment=paid THEN
promotion=false
New Rule 4: IF usage>=200 AND usage<400 AND payment=paid THEN
promotion=true
New Rule 5: IF usage>=400 AND payment=paid THEN promotion=false

6) ถ้าเปอร์เซ็นต์ความถูกต้องจากเซตของกฎใหม่ มากกว่าเซตของกฎเดิมให้ทำการพิจารณากฎคู่ถัดไปตามเงื่อนไขจากข้อ 1 - 5

บทที่ 4
วิธีการทดลองและผลการทดลอง

ในบทนี้จะกล่าวถึงวิธีการดำเนินการวิจัยโดยประกอบไปด้วย เครื่องมือที่ใช้ในการวิจัย ข้อมูลที่ใช้ในการทดลอง วิธีการทดลอง ผลการทดลองและปัญหาและข้อจำกัดในการดำเนินงานวิจัย ดังนี้

4.1 เครื่องมือที่ใช้ในการวิจัย

4.1.1 ฮาร์ดแวร์

- หน่วยประมวลผลกลาง Intel® Core™ 2 Duo 1.20 GHz
- หน่วยความจำ 1.5 กิกะไบต์ (RAM 1.5 GB)
- ฮาร์ดดิสก์ 60 กิกะไบต์ (Hard Disk 60 GB)

4.2.1 ซอฟต์แวร์

- โปรแกรมเวกา เวอร์ชัน 3.6.0 เพื่อใช้ในการแบ่งชุดข้อมูลสอนและข้อมูลทดสอบด้วยการตรวจสอบแบบไขว้กัน 10 ชุด
- โปรแกรม Eclipse SDK เวอร์ชัน 3.1.2 ในการพัฒนาด้วยภาษาจาวา (Java)

4.2 ข้อมูลที่ใช้ในการทดลอง

ใช้ชุดข้อมูล (Data Set) จาก UCI Machine Learning Repository จำนวน 7 ชุด ได้แก่ Balance Scale, Blood Transfusion, Haberman's Survival, Iris, Liver Disorders, Pima Indians Diabetes Database and Statlog โดยในแต่ละชุดจะนำมาแบ่งเป็นชุดข้อมูลสอนและชุดข้อมูลทดสอบด้วยโปรแกรมเวกา ด้วยวิธีการตรวจสอบแบบไขว้กัน 10 ชุด ข้อมูล 10% จะถูกใช้เป็นชุดข้อมูลทดสอบและอีก 90% จะนำมาแบ่งอีกครั้งด้วยโปรแกรมพัฒนาขึ้นมาเอง เพื่อใช้เป็นชุดข้อมูลสอน 80% และใช้เป็นเป็นชุดข้อมูลตรวจสอบ ในการหาเซตของกฎใหม่ อีก 20% และทำการเปรียบเทียบผลการทดลองจากเปอร์เซ็นต์ความถูกต้องของการทำนาย ระหว่างป่าแบบสุ่มที่ทำการผสานกฎกับ Random Forests และต้นไม้ตัดสินใจแบบ J48 (C4.5)

ตารางที่ 2 ชุดข้อมูลที่ใช้ทดสอบ

ชื่อชุดข้อมูล (Data Set)	จำนวนข้อมูล (Number of Instances)	จำนวนกลุ่ม (Number of Classes)	จำนวนคุณลักษณะ (Number of Attributes)
Balance Scale	625	3	4
Blood Transfusion	748	2	5
Haberman's Survival	306	2	3
Iris	150	3	4
Liver Disorders	345	2	6
Pima Indians Diabetes	768	2	8
Statlog	690	2	14

4.3 วิธีการทดลอง

จัดเตรียมไฟล์ข้อมูลต้นไม้ตัดสินใจและไฟล์นำเข้าวางไว้ในโฟลเดอร์เดียวกัน แล้วรันโปรแกรม (ภาคผนวก ค.)

เราแบ่งวิธีการผสานกฎออกเป็น 2 แบบ ซึ่งให้ผลลัพธ์ตามที่แสดงไว้ในตารางที่ 3

1. เลือกต้นไม้ที่มีเปอร์เซ็นต์ความถูกต้องของผลทำนายมากที่สุดมาผสานก่อน (RFh)
2. เลือกต้นไม้ที่มีเปอร์เซ็นต์ความถูกต้องของผลทำนายน้อยที่สุดมาผสานก่อน (RFl)

4.4 ผลการทดลอง

จากการทดลองจะทำให้ได้เปอร์เซ็นต์ความถูกต้องของผลการทำนายจากการผสานกฎด้วยวิธีการผสานทั้ง 2 แบบ (RFh, RFl) เปอร์เซ็นต์ความถูกต้องของผลการทำนายป่าแบบสุ่ม และเปอร์เซ็นต์ความถูกต้องของผลการทำนายต้นไม้ตัดสินใจแบบ J48 ดังที่แสดงไว้ในตารางที่ 3 และจำนวนของกฎหลังจากการผสานกฎทั้ง 2 แบบ (RFh, RFl) จำนวนของกฎของป่าแบบสุ่มและจำนวนของกฎของต้นไม้ตัดสินใจแบบ J48 ดังที่แสดงไว้ในตารางที่ 4

ตารางที่ 3 ค่าเฉลี่ยของเปอร์เซ็นต์ความถูกต้องของผลการทำนายที่ได้จากการผสานกฎจากป่าแบบสุ่มจากเปรียบเทียบกับป่าแบบสุ่มและต้นไม้ตัดสินใจแบบ J48

ชุดข้อมูล	ค่าเฉลี่ยความถูกต้องในการทำนาย (%)			
	RFh	RFl	Random Forests	J48
Balance Scale	90.88582	91.67947	79.35230	78.70200
Blood Transfusion	94.78559	97.60001	70.04324	77.54234
Haberman's Survival	92.17204	94.7957	59.83869	72.51611
Iris	96.00000	98.66667	95.33332	94.66667
Liver Disorders	97.99159	98.85714	69.65546	62.02523
Pima Indians Diabetes	97.39576	97.26759	71.49182	72.91356
Statlog	98.11594	98.11594	85.55216	85.36231

ตารางที่ 4 ค่าเฉลี่ยของจำนวนกฎที่ได้จากการผสานกฎจากป่าแบบสุ่มจากเปรียบเทียบกับป่าแบบสุ่มและต้นไม้ตัดสินใจแบบ J48

ชุดข้อมูล	ค่าเฉลี่ยของจำนวนกฎ			
	RFh	RFl	Random Forests	J48
Balance Scale	432.3	530.5	898.1	34.7
Blood Transfusion	633	595.3	1149.9	6.8
Haberman's Survival	268.2	353.6	529.1	3.1
Iris	9	14.9	59.4	3.7
Liver Disorders	224	304.3	531.1	18
Pima Indians Diabetes	584.4	661.4	861.6	22
Statlog	377.2	428.7	735.8	22.9

การผสานกฎที่เริ่มผสานจากกฎที่มีเปอร์เซ็นต์ความถูกต้องของผลทำนายน้อยที่สุดมาผสานก่อน ให้เปอร์เซ็นต์ความถูกต้องของผลทำนายออกมาสูงกว่าการเลือกผสานกฎจากกฎที่มีเปอร์เซ็นต์ความถูกต้องของผลทำนายสูงสุดมาผสานก่อน แต่จำนวนกฎที่ได้จากการผสานของการเลือกกฎที่มีเปอร์เซ็นต์ความถูกต้องของผลทำนายสูงมาผสานก่อนมีจำนวนกฎน้อยกว่า

4.5 ปัญหาและข้อจำกัด

โครงสร้างของไฟล์ชุดข้อมูลที่นำมาทดสอบมีรูปแบบของไฟล์ (ARFF) ที่ต่างจากที่กำหนดไว้ ไม่สามารถนำมาใช้ทดสอบได้ทันที ดังนั้นจึงจำเป็นจะต้องบันทึกเป็นเท็กซ์ไฟล์ (text file) และทำการตัดแบ่งข้อมูลต้นไม้ตัดสินใจ 1 ต้นต่อ 1 ไฟล์

อัลกอริทึมในการผสานกฎนี้เหมาะสำหรับข้อมูลที่มีคุณสมบัติไม่มาก และค่าของข้อมูลของคุณลักษณะสามารถแบ่งช่วงได้

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

ผลการวิจัยจากการทดลองด้วยชุดข้อมูลจาก UCI Machine Learning Repository จำนวน 7 ชุด ทำให้พบว่าวิธีการผสานกฎจากป่าแบบสุ่มที่นำเสนอนั้นนอกจากจำนวนกฎจะลดลงแล้ว ยังให้เปอร์เซ็นต์ความถูกต้องของผลทำนายได้ดีกว่า Random Forests และต้นไม้ตัดสินใจแบบ C4.5(J48) จากผลการทดลองทำให้เข้าใจได้ว่า การเลือกต้นไม้ที่มีเปอร์เซ็นต์ความถูกต้องของผลทำนายต่ำมาผสานกฎก่อน ให้เปอร์เซ็นต์ความถูกต้องของผลทำนายดีกว่าการเลือกต้นไม้ที่มีเปอร์เซ็นต์ความถูกต้องทำนายสูงมาผสานก่อน

การผสานกฎจากป่าแบบสุ่มถึงแม้จะให้ผลลัพธ์ที่ดีขึ้น แต่ก็ใช้เวลาในการผสานกฎนานขึ้นเช่นกัน โดยเฉพาะต้นไม้ติดสินใจที่มีความซับซ้อนมาก เมื่อทำการแปลงออกมาเป็นกฎก็จะได้เซตของกฎที่มีขนาดใหญ่ ในการผสานกฎแต่ละขั้นตอนก็จะนานขึ้นแปรผันไปตามจำนวนของกฎ อัลกอริทึมในการผสานกฎที่พัฒนาขึ้นมาเหมาะสำหรับข้อมูลที่มีคุณลักษณะของข้อมูลไม่มาก และค่าของข้อมูลแบบแบ่งช่วงได้เท่านั้น

5.2 แนวทางในการพัฒนาต่อ

วิธีการผสานกฎยังมีอีกหลายจุดที่สามารถปรับปรุงและพัฒนาต่อให้ดีขึ้น เช่นการจัดการกฎที่คล้ายกัน การจัดการผสานกฎในครั้งเดียว และการผสานกฎที่ครอบคลุมข้อมูลทุกประเภท

การจัดการกฎที่คล้ายกันแบบสามารถขยายช่วงของข้อมูลที่พิจารณาได้ การรวมกฎที่คล้ายกันแบบนี้ อาจทำให้ประสิทธิภาพลดลง เพราะหลังจากการผสานกฎหลายครั้ง อาจจะนำกฎที่ไม่มีความคล้ายกันเลยมารวมเข้าด้วยกัน ดังนั้นในการทดลองจึงทำการรวมกฎที่คล้ายกันในลักษณะนี้ไม่เกิน 1 ครั้งในแต่ละรอบ ซึ่งควรต้องปรับปรุงในส่วนของวิธีการนี้ เนื่องจากลำดับของกฎที่นำมาพิจารณา มีผลทำให้กฎที่มีลำดับใกล้กันมากที่สุดถูกผสานก่อน ดังนั้นผลลัพธ์ที่ได้อาจจะไม่ได้ดีขึ้นเสมอไป

สำหรับกฎที่ได้จากต้นไม้ติดสินใจที่สร้างโดยชุดข้อมูลสอนที่มีข้อมูลรบกวนมาก เซตของกฎอาจจะมีขนาดใหญ่มาก ดังนั้นการทำให้กฎมีขนาดเล็กลงด้วยวิธีการตัดเล็มกิ่งหรือตัดข้อมูลรบกวนออกไป อาจจะช่วยให้ได้เซตของกฎที่มีประสิทธิภาพมากขึ้น

รายการอ้างอิง

[1] Breiman, L. Random Forests. In Journal of Machine Learning, 45,1(2001) : 5-32.

[2] Zhou, Z.-H., and Tang, W. Selective Ensemble of Decision Trees. Nanjing 210093, National Laboratory for Novel Software Technology, China, 2003.

[3] Zhang, Y., Burer, S., and Street, W.N. Ensemble Pruning via Semi-definite Programming. In Journal of Machine Learning Research, 1315-13387, 2006.

[4] Assche, A.V., and Blockeel, H. Seeing the Forest through the Trees: Learning a Comprehensible Model from an Ensemble. In Proceedings of ECML, 418-429, 2007.

[5] Anderson, G. PhD thesis: Random Relational Rules. 2009.

[6] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. Classification and Regression Trees. Belmont, California: Wadsworth, 1984.

[7] Quinlan, J. R. Learning Decision Tree Classifiers. In ACM Computing Surveys, vol. 28 no. 1 (March 1996): 71-72.

[8] บุญเสริม กิจศิริกุล, ปัญญาประดิษฐ์ในเอกสารคำสอนวิชา 2110654 ภาควิชาวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย,หน้า 153-163, 2548.

[9] Tan ,P.N., Steinbach, M., and Kumar, V. Introduction to Data Mining, Pearson International Edition, 2006.

[10] Waitman, L.R., Fisher, D.H., King, P.H. Bootstrapping rule induction to achieve rule stability and reduction. In Journal of Intelligent Information Systems, 49-77, 2006.

[11] Seni, G., Yang, E., and Akar, S. Yield Modeling with Rule Ensembles. In 18th Annual IEEE/SEMI Advanced Semiconductor Manufacturing Conference, 228-233, 2007.

[12] Banfield, R.E., Lawrence, O., Bowyer, K.W., and Kegelmeyer, W.P. A Comparison of Decision Tree Ensemble Creation Techniques. In IEEE Transaction on Pattern Analysis and Machine Intelligence, 173-180, 2007.

[13] Margineantu, D.D., and Dietterich, T.G. Improved Class Probability Estimates from Decision Tree Models. In Nonlinear Estimation and Classification; Lecture Notes in Statistics, pp.169-184. New York : Springer-Verlag, 2002.

[14] Asuncion, A., and Newman, D.J. UCI Machine Learning Repository [Online]. 2007. Available from: http://archive.ics.uci.edu/ml/ [2009, May 25]

[15] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and I. Witten, H. The WEKA Data Mining Software: An Update. SIGKDD Explorations, 11, 1(2009) : 10-18.

[16] Seni, G., and Elder, J. From Trees to Forest and Rule Sets, A Unified Overview of Ensemble Methods. In 13th International Conference on Knowledge Discovery and Data Mining (KDD), 2007.

[17] Quinlan, J. R. Generating Production Rules from Decision Trees. In Proceedings of the 10th International Joint Conference on Artificial Intelligence, Milan, Italy: Morgan Kaufmann, 304-307, 1987.

[18] Opitz, D., and Maclin, R. Popular Ensemble Methods: An Empirical Study.In Journal of Artificial Intelligence Research,169–198, 1999.

[19] Witten, I.H., and Frank, E. Attribute-Relational File Format. University of Waikato, New Zealand, 2002.

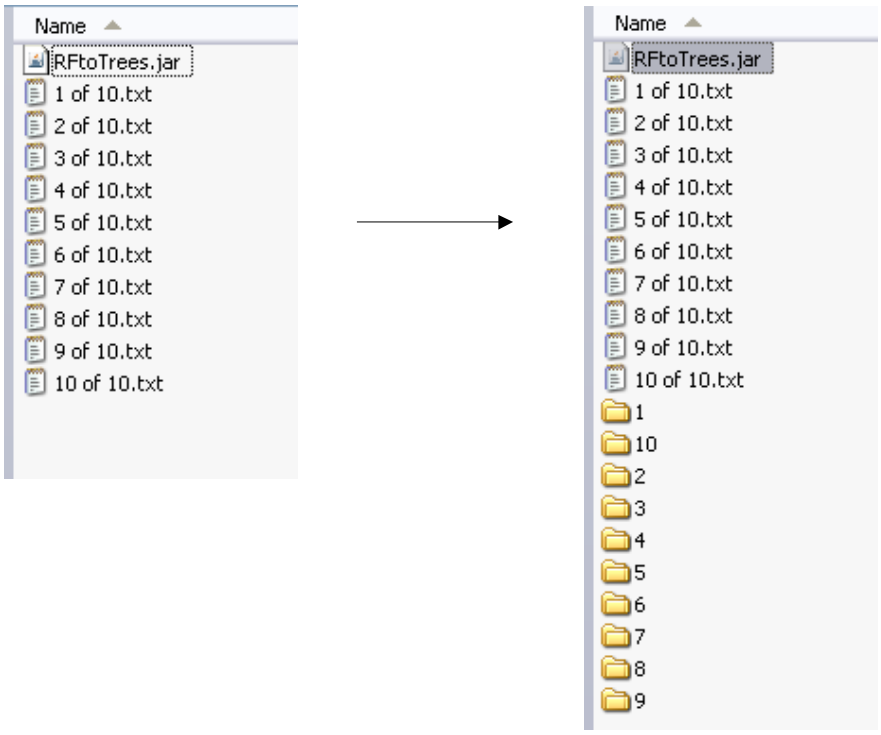
[20] Breiman, L., and Cutler, A. Random Forests [Online]. 2005. Available from : http://www.stat.berkeley.edu/~breiman/ [2009, August 8]

ภาคผนวก

ภาคผนวก ก
วิธีการใช้งาน RF2Tree Tools

RF2Tree Tools (Random Forest to Trees Tools) เป็นโปรแกรมที่พัฒนาขึ้นเพื่อจัดรูปแบบรายละเอียดของต้นไม้ตัดสินใจที่ได้จากการสร้างแบบ Random Forest ด้วยโปรแกรมเวกาให้มีโครงสร้างเป็นไฟล์ข้อมูลนำเข้าที่ต้องการ คือตัดแบ่งและบันทึกให้มีต้นไม้ตัดสินใจ 1 ต้นต่อ 1 ไฟล์

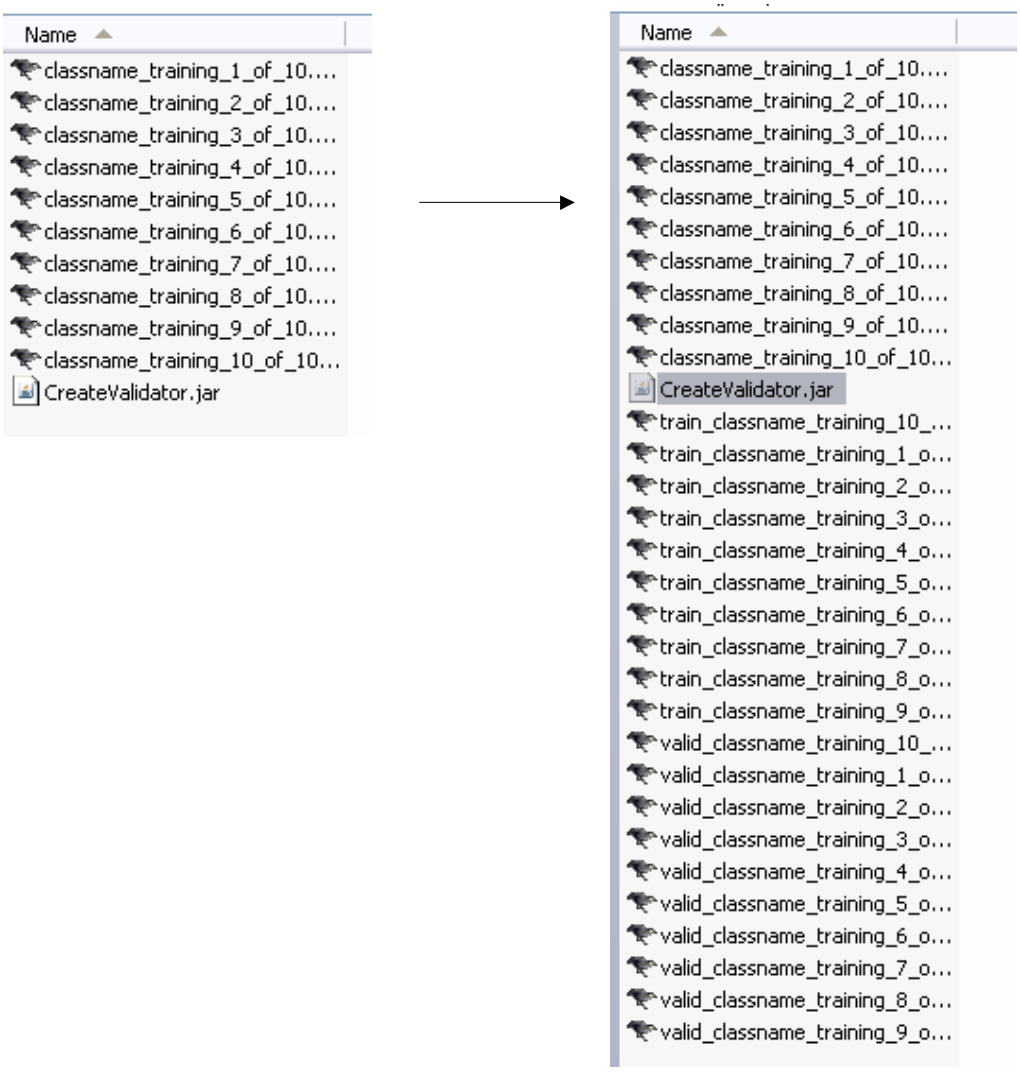
วิธีการใช้งาน ให้นำรายละเอียดข้อมูลที่ได้จากการสร้างต้นไม้ตัดสินใจป่าแบบสุ่ม ด้วยโปรแกรมเวกา บันทึกเป็น Text file (.txt) จากนั้นนำ RF2Tree Tools (RF2Trees.jar) มาวางไว้กับในโฟลเดอร์เดียวกับต้นไม้ตัดสินใจที่บันทึกเป็นไฟล์ตัวอักษร (.txt) จากนั้นดับเบิ้ลคลิก RF2Trees.jar เพื่อรันโปรแกรม โปรแกรมจะทำการแบ่งต้นไม้ตัดสินใจออกเป็นไฟล์ละ 1 ต้น



ภาคผนวก ข
วิธีใช้งาน Validator Tools

Validator Tools เป็นโปรแกรมที่พัฒนาขึ้นเพื่อแบ่งชุดข้อมูลสร้างออกเป็นชุดข้อมูลสอน 80% และชุดข้อมูลตรวจสอบ 20%

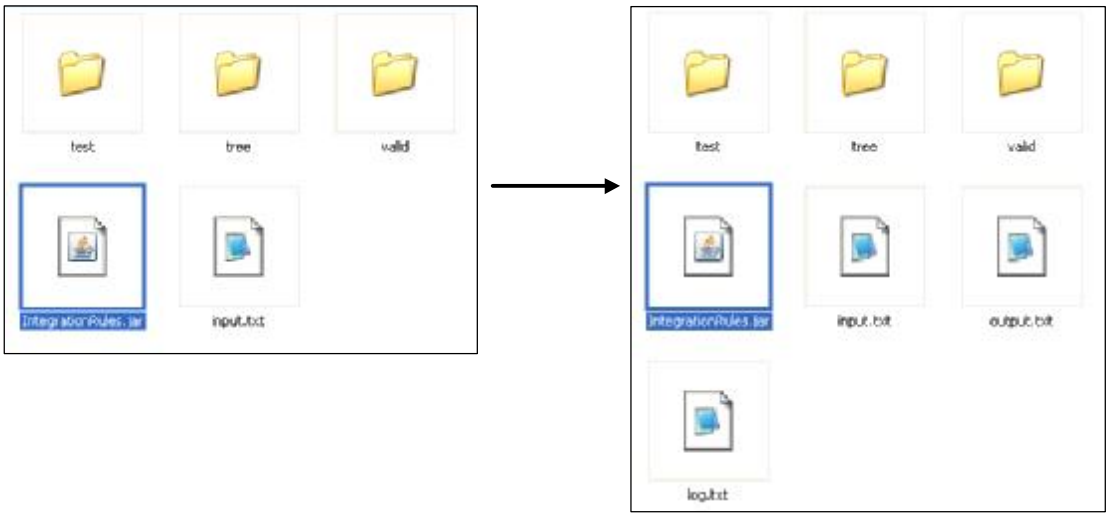
วิธีการใช้งานให้นำ Validator Tools (CreateValidator.jar) มาวางไว้ที่โฟลเดอร์เดียวกับไฟล์ข้อมูลสร้าง ที่ได้จากการทำการตรวจสอบแบบไว้กัน 10 ชุด ด้วยโปรแกรมเวกา จากนั้นดับเบิลคลิก CreateValidator.jar เพื่อรันโปรแกรม จะได้ไฟล์ใหม่จำนวน 20 ไฟล์ ซึ่งถูกแบ่งเป็นชุดข้อมูลสอนและชุดข้อมูลตรวจสอบ



ภาคผนวก ค
วิธีการใช้ Integration Rule Tools

วิธีการใช้โปรแกรมในการผสานกฎ (IntegrationRules.jar) หลังจากที่เราได้จัดเตรียมข้อมูลของต้นไม้ที่มีการแบ่งต้นไม้ 1 ต้นต่อ 1 ไฟล์แล้ว ให้เรากำหนดไฟล์นำเข้า (Input.txt) โดยนำวางไว้โฟลเดอร์เดียวกับโปรแกรมในการผสานกฎ (integrationRules.jar) จากนั้นดับเบิ้ลคลิก integrationRules.jarเพื่อรันโปรแกรมทดสอบ ซึ่งจะได้ไฟล์ทดสอบออกมา 2 ไฟล์ ดังนี้

1. ไฟล์นำออก (output.txt) เป็นไฟล์ของต้นไม้ตัดสินใจที่ทำการผสานกฎแล้วทำให้ได้เซวของกฎที่มีเปอร์เว็นต์ความถูกต้องของผลทำนายสูงขึ้น เปอร์เซ็นต์ความถูกต้องของผลทำนาย และแสดงกฎแต่ละข้อที่ได้หลังการผสานกฎ
2. ไฟล์รายละเอียดการผสานกฎ (log.txt) เป็นไฟล์ที่แสดรายละเอียดการทำงานตั้งแต่รายละเอียดต้นไม้ที่นำเข้า กฎที่ได้จากการแปลงกฎ และเปอร์เซ็นต์ความถูกต้องของผลทำนายหลังการผสานกฎแต่ละครั้ง



รูปแบบไฟล์ข้อมูนำเข้า

```
tree_1_file_path
tree_2_file_path
tree_3_file_path
    :       :
tree_n_file_path
validation_set_file_path
test_set_file_path
```

รูปแบบไฟล์ข้อมูลนำออก

```
Trees those integrated in this new rule set -
tree_a_file_path
tree_b_file_path
    :       :
Average Number of Rules from Trees in Random Forest : m

Total Rules : n
Correctly Classified Instances (Validator Set) : k %
Correctly Classified Instances (Test Set) : k %
===Rule after integrating ===
IF .... THEN ....
IF .... THEN ....
IF .... THEN ....
    :    :
```

ภาคผนวก ง
ตัวอย่างไฟล์ต้นไม้ตัดสินใจ

```
petallength < 2.35 : Iris-setosa (37/0)
petallength >= 2.35
|   petallength < 4.95
|   |   petalwidth < 1.65 : Iris-versicolor (35/0)
|   |   petalwidth >= 1.65
|   |   |   sepalwidth < 3.1 : Iris-virginica (2/0)
|   |   |   sepalwidth >= 3.1 : Iris-versicolor (1/0)
|   petallength >= 4.95
|   |   petallength < 5.05
|   |   |   sepallength < 6.35 : Iris-virginica (2/0)
|   |   |   sepallength >= 6.35 : Iris-versicolor (1/0)
          |   |   petallength >= 5.05 : Iris-virginica (30/0)
```

ภาคผนวก จ
ตัวอย่างไฟล์ข้อมูลนำเข้า ไฟล์ข้อมูลนำออก

ตัวอย่างไฟล์ข้อมูลนำเข้า

```
D:\Integration_Rules_RF\Iris\tree\1\1.txt
D:\Integration_Rules_RF\Iris\tree\1\2.txt
D:\Integration_Rules_RF\Iris\tree\1\3.txt
D:\Integration_Rules_RF\Iris\tree\1\4.txt
D:\Integration_Rules_RF\Iris\tree\1\5.txt
D:\Integration_Rules_RF\Iris\tree\1\6.txt
D:\Integration_Rules_RF\Iris\tree\1\7.txt
D:\Integration_Rules_RF\Iris\tree\1\8.txt
D:\Integration_Rules_RF\Iris\tree\1\9.txt
D:\Integration_Rules_RF\Iris\tree\1\10.txt
D:\Integration_Rules_RF\Iris\valid\valid_classname_training_1_of_10.arff
D:\Integration_Rules_RF\Iris\test\classname_test_1_of_10.arff
```

ตัวอย่างไฟล์ข้อมูลนำออก

```
Trees those integrated in this new rule set -
D:\Integration_Rules_RF\Iris\tree\1\1.txt
Average Number of Rules from Trees in Random Forest : 6.6

Total Rules : 7
Correctly Classified Instances (Validator Set) : 100.0 %
Correctly Classified Instances (Test Set) : 93.333336 %
===Rule after integrating ===
IF petallength < 2.35 THEN Iris-setosa
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth < 1.65 THEN
Iris-versicolor
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65 AND
sepalwidth < 3.1 THEN Iris-virginica
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65 AND
sepalwidth >= 3.1 THEN Iris-versicolor
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength < 6.35 THEN
Iris-virginica
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength >= 6.35 THEN
Iris-versicolor
IF petallength >= 5.05 THEN Iris-virginica
```

ตัวอย่างไฟล์รายละเอียดการผสานกฎของชุดข้องมูล Iris

```
@@read tree from D:\Integration_Rules_RF\Iris\tree\1\1.txt
Number of Rules : 7
Correctly Classified Instances : 100.0 %
@@read tree from D:\Integration_Rules_RF\Iris\tree\1\2.txt
Number of Rules : 6
Correctly Classified Instances : 100.0 %
@@read tree from D:\Integration_Rules_RF\Iris\tree\1\3.txt
Number of Rules : 8
Correctly Classified Instances : 100.0 %
@@read tree from D:\Integration_Rules_RF\Iris\tree\1\4.txt
Number of Rules : 6
Correctly Classified Instances : 96.296295 %
@@read tree from D:\Integration_Rules_RF\Iris\tree\1\5.txt
Number of Rules : 5
Correctly Classified Instances : 100.0 %
@@read tree from D:\Integration_Rules_RF\Iris\tree\1\6.txt
Number of Rules : 7
Correctly Classified Instances : 92.59259 %
@@read tree from D:\Integration_Rules_RF\Iris\tree\1\7.txt
Number of Rules : 8
Correctly Classified Instances : 85.18519 %
@@read tree from D:\Integration_Rules_RF\Iris\tree\1\8.txt
Number of Rules : 6
Correctly Classified Instances : 100.0 %
@@read tree from D:\Integration_Rules_RF\Iris\tree\1\9.txt
Number of Rules : 8
Correctly Classified Instances : 100.0 %
@@read tree from D:\Integration_Rules_RF\Iris\tree\1\10.txt
Number of Rules : 5
Correctly Classified Instances : 100.0 %
---order of trees which will be integrated ---
1. 100.0%   D:\Integration_Rules_RF\Iris\tree\1\1.txt
2. 100.0%   D:\Integration_Rules_RF\Iris\tree\1\2.txt
3. 100.0%   D:\Integration_Rules_RF\Iris\tree\1\3.txt
4. 100.0%   D:\Integration_Rules_RF\Iris\tree\1\5.txt
5. 100.0%   D:\Integration_Rules_RF\Iris\tree\1\8.txt
6. 100.0%   D:\Integration_Rules_RF\Iris\tree\1\9.txt
7. 100.0%   D:\Integration_Rules_RF\Iris\tree\1\10.txt
8. 96.296295%     D:\Integration_Rules_RF\Iris\tree\1\4.txt
9. 92.59259%      D:\Integration_Rules_RF\Iris\tree\1\6.txt
10. 85.18519%     D:\Integration_Rules_RF\Iris\tree\1\7.txt
@@read tree from D:\Integration_Rules_RF\Iris\tree\1\1.txt
@@initial tree is D:\Integration_Rules_RF\Iris\tree\1\1.txt
tree is -----
petallength < 2.35 : Iris-setosa (37/0)
petallength >= 2.35
|   petallength < 4.95
|   |   petalwidth < 1.65 : Iris-versicolor (35/0)
|   |   petalwidth >= 1.65
|   |   |   sepalwidth < 3.1 : Iris-virginica (2/0)
|   |   |   sepalwidth >= 3.1 : Iris-versicolor (1/0)
|   petallength >= 4.95
|   |   petallength < 5.05
|   |   |   sepallength < 6.35 : Iris-virginica (2/0)
|   |   |   sepallength >= 6.35 : Iris-versicolor (1/0)
|   |   petallength >= 5.05 : Iris-virginica (30/0)
```

```
---tree to rules ---
IF petallength < 2.35 THEN Iris-setosa
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth < 1.65
THEN Iris-versicolor
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth < 3.1 THEN Iris-virginica
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth >= 3.1 THEN Iris-versicolor
IF petallength >= 2.35 AND petallength >= 4.95 AND petallength < 5.05
AND sepallength < 6.35 THEN Iris-virginica
IF petallength >= 2.35 AND petallength >= 4.95 AND petallength < 5.05
AND sepallength >= 6.35 THEN Iris-versicolor
IF petallength >= 2.35 AND petallength >= 4.95 AND petallength >=
5.05 THEN Iris-virginica
---clean the initial rules for integration ---
---Initial rules for this iteration ---
IF petallength < 2.35 THEN Iris-setosa
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth < 1.65
THEN Iris-versicolor
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth < 3.1 THEN Iris-virginica
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth >= 3.1 THEN Iris-versicolor
IF petallength >= 2.35 AND petallength >= 4.95 AND petallength < 5.05
AND sepallength < 6.35 THEN Iris-virginica
IF petallength >= 2.35 AND petallength >= 4.95 AND petallength < 5.05
AND sepallength >= 6.35 THEN Iris-versicolor
IF petallength >= 2.35 AND petallength >= 4.95 AND petallength >=
5.05 THEN Iris-virginica
---Initial rules after cleaning ---
IF petallength < 2.35 THEN Iris-setosa
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth < 1.65
THEN Iris-versicolor
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth < 3.1 THEN Iris-virginica
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth >= 3.1 THEN Iris-versicolor
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength < 6.35
THEN Iris-virginica
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength >= 6.35
THEN Iris-versicolor
IF petallength >= 5.05 THEN Iris-virginica
this initial can validate 100.0%
@@read tree from D:\Integration_Rules_RF\Iris\tree\1\2.txt
@@integrate with D:\Integration_Rules_RF\Iris\tree\1\2.txt
tree is -----
petalwidth < 0.8 : Iris-setosa (39/0)
petalwidth >= 0.8
|   petallength < 5.05
|   |   sepallength < 4.95
|   |   |   petallength < 3.9 : Iris-versicolor (1/0)
|   |   |   petallength >= 3.9 : Iris-virginica (1/0)
|   |   sepallength >= 4.95
|   |   |   petalwidth < 1.75 : Iris-versicolor (32/0)
|   |   |   petalwidth >= 1.75 : Iris-virginica (1/0)
|   petallength >= 5.05 : Iris-virginica (34/0)
---tree to rules ---
IF petalwidth < 0.8 THEN Iris-setosa
```

```
IF petalwidth >= 0.8 AND petallength < 5.05 AND sepallength < 4.95
AND petallength < 3.9 THEN Iris-versicolor
IF petalwidth >= 0.8 AND petallength < 5.05 AND sepallength < 4.95
AND petallength >= 3.9 THEN Iris-virginica
IF petalwidth >= 0.8 AND petallength < 5.05 AND sepallength >= 4.95
AND petalwidth < 1.75 THEN Iris-versicolor
IF petalwidth >= 0.8 AND petallength < 5.05 AND sepallength >= 4.95
AND petalwidth >= 1.75 THEN Iris-virginica
IF petalwidth >= 0.8 AND petallength >= 5.05 THEN Iris-virginica
this new tree can validate 100.0%
integrate with rule set #1
D:\Integration_Rules_RF\Iris\tree\1\2.txt
---Initial rules for this iteration ---
IF petallength < 2.35 THEN Iris-setosa
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth < 1.65
THEN Iris-versicolor
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth < 3.1 THEN Iris-virginica
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth >= 3.1 THEN Iris-versicolor
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength < 6.35
THEN Iris-virginica
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength >= 6.35
THEN Iris-versicolor
IF petallength >= 5.05 THEN Iris-virginica
IF petalwidth < 0.8 THEN Iris-setosa
IF petalwidth >= 0.8 AND petallength < 5.05 AND sepallength < 4.95
AND petallength < 3.9 THEN Iris-versicolor
IF petalwidth >= 0.8 AND petallength < 5.05 AND sepallength < 4.95
AND petallength >= 3.9 THEN Iris-virginica
IF petalwidth >= 0.8 AND petallength < 5.05 AND sepallength >= 4.95
AND petalwidth < 1.75 THEN Iris-versicolor
IF petalwidth >= 0.8 AND petallength < 5.05 AND sepallength >= 4.95
AND petalwidth >= 1.75 THEN Iris-virginica
IF petalwidth >= 0.8 AND petallength >= 5.05 THEN Iris-virginica
Correctly Classified Instances : 100.0%
Roll back the rule set
@@read tree from D:\Integration_Rules_RF\Iris\tree\1\3.txt
@@integrate with D:\Integration_Rules_RF\Iris\tree\1\3.txt
tree is -----
petalwidth < 0.8 : Iris-setosa (37/0)
petalwidth >= 0.8
|   petalwidth < 1.75
|   |   petallength < 5.3
|   |   |   petalwidth < 1.65 : Iris-versicolor (33/0)
|   |   |   petalwidth >= 1.65
|   |   |   |   sepallength < 5.8 : Iris-virginica (1/0)
|   |   |   |   sepallength >= 5.8 : Iris-versicolor (1/0)
|   |   petallength >= 5.3 : Iris-virginica (4/0)
|   petalwidth >= 1.75
|   |   petallength < 4.95
|   |   |   sepalwidth < 3.1 : Iris-virginica (1/0)
|   |   |   sepalwidth >= 3.1 : Iris-versicolor (1/0)
|   |   petallength >= 4.95 : Iris-virginica (30/0)
---tree to rules ---
IF petalwidth < 0.8 THEN Iris-setosa
IF petalwidth >= 0.8 AND petalwidth < 1.75 AND petallength < 5.3 AND
petalwidth < 1.65 THEN Iris-versicolor
```

```
IF petalwidth >= 0.8 AND petalwidth < 1.75 AND petallength < 5.3 AND
petalwidth >= 1.65 AND sepallength < 5.8 THEN Iris-virginica
IF petalwidth >= 0.8 AND petalwidth < 1.75 AND petallength < 5.3 AND
petalwidth >= 1.65 AND sepallength >= 5.8 THEN Iris-versicolor
IF petalwidth >= 0.8 AND petalwidth < 1.75 AND petallength >= 5.3
THEN Iris-virginica
IF petalwidth >= 0.8 AND petalwidth >= 1.75 AND petallength < 4.95
AND sepalwidth < 3.1 THEN Iris-virginica
IF petalwidth >= 0.8 AND petalwidth >= 1.75 AND petallength < 4.95
AND sepalwidth >= 3.1 THEN Iris-versicolor
IF petalwidth >= 0.8 AND petalwidth >= 1.75 AND petallength >= 4.95
THEN Iris-virginica
this new tree can validate 100.0%
integrate with rule set #1
D:\Integration_Rules_RF\Iris\tree\1\3.txt
---Initial rules for this iteration ---
IF petallength < 2.35 THEN Iris-setosa
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth < 1.65
THEN Iris-versicolor
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth < 3.1 THEN Iris-virginica
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth >= 3.1 THEN Iris-versicolor
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength < 6.35
THEN Iris-virginica
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength >= 6.35
THEN Iris-versicolor
IF petallength >= 5.05 THEN Iris-virginica
IF petalwidth < 0.8 THEN Iris-setosa
IF petalwidth >= 0.8 AND petalwidth < 1.75 AND petallength < 5.3 AND
petalwidth < 1.65 THEN Iris-versicolor
IF petalwidth >= 0.8 AND petalwidth < 1.75 AND petallength < 5.3 AND
petalwidth >= 1.65 AND sepallength < 5.8 THEN Iris-virginica
IF petalwidth >= 0.8 AND petalwidth < 1.75 AND petallength < 5.3 AND
petalwidth >= 1.65 AND sepallength >= 5.8 THEN Iris-versicolor
IF petalwidth >= 0.8 AND petalwidth < 1.75 AND petallength >= 5.3
THEN Iris-virginica
IF petalwidth >= 0.8 AND petalwidth >= 1.75 AND petallength < 4.95
AND sepalwidth < 3.1 THEN Iris-virginica
IF petalwidth >= 0.8 AND petalwidth >= 1.75 AND petallength < 4.95
AND sepalwidth >= 3.1 THEN Iris-versicolor
IF petalwidth >= 0.8 AND petalwidth >= 1.75 AND petallength >= 4.95
THEN Iris-virginica
Correctly Classified Instances : 100.0%
Roll back the rule set
@@read tree from D:\Integration_Rules_RF\Iris\tree\1\5.txt
@@integrate with D:\Integration_Rules_RF\Iris\tree\1\5.txt
tree is -----
petallength < 2.45 : Iris-setosa (33/0)
petallength >= 2.45
|   petallength < 5
|   |   petalwidth < 1.65 : Iris-versicolor (35/0)
|   |   petalwidth >= 1.65
|   |   |   sepalwidth < 3.1 : Iris-virginica (2/0)
|   |   |   sepalwidth >= 3.1 : Iris-versicolor (1/0)
|   petallength >= 5 : Iris-virginica (37/0)
---tree to rules ---
IF petallength < 2.45 THEN Iris-setosa
```

```
IF petallength >= 2.45 AND petallength < 5 AND petalwidth < 1.65 THEN
Iris-versicolor
IF petallength >= 2.45 AND petallength < 5 AND petalwidth >= 1.65 AND
sepalwidth < 3.1 THEN Iris-virginica
IF petallength >= 2.45 AND petallength < 5 AND petalwidth >= 1.65 AND
sepalwidth >= 3.1 THEN Iris-versicolor
IF petallength >= 2.45 AND petallength >= 5 THEN Iris-virginica
this new tree can validate 100.0%
integrate with rule set #1
D:\Integration_Rules_RF\Iris\tree\1\5.txt
---Initial rules for this iteration ---
IF petallength < 2.35 THEN Iris-setosa
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth < 1.65
THEN Iris-versicolor
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth < 3.1 THEN Iris-virginica
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth >= 3.1 THEN Iris-versicolor
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength < 6.35
THEN Iris-virginica
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength >= 6.35
THEN Iris-versicolor
IF petallength >= 5.05 THEN Iris-virginica
IF petallength < 2.45 THEN Iris-setosa
IF petallength >= 2.45 AND petallength < 5 AND petalwidth < 1.65 THEN
Iris-versicolor
IF petallength >= 2.45 AND petallength < 5 AND petalwidth >= 1.65 AND
sepalwidth < 3.1 THEN Iris-virginica
IF petallength >= 2.45 AND petallength < 5 AND petalwidth >= 1.65 AND
sepalwidth >= 3.1 THEN Iris-versicolor
IF petallength >= 2.45 AND petallength >= 5 THEN Iris-virginica
Correctly Classified Instances : 100.0%
Roll back the rule set
@@read tree from D:\Integration_Rules_RF\Iris\tree\1\8.txt
@@integrate with D:\Integration_Rules_RF\Iris\tree\1\8.txt
tree is -----
petallength < 2.45 : Iris-setosa (35/0)
petallength >= 2.45
|   petalwidth < 1.7
|   |   petallength < 4.95 : Iris-versicolor (38/0)
|   |   petallength >= 4.95 : Iris-virginica (2/0)
|   petalwidth >= 1.7
|   |   sepallength < 5.95
|   |   |   sepalwidth < 3.1 : Iris-virginica (3/0)
|   |   |   sepalwidth >= 3.1 : Iris-versicolor (3/0)
|   |   sepallength >= 5.95 : Iris-virginica (27/0)
---tree to rules ---
IF petallength < 2.45 THEN Iris-setosa
IF petallength >= 2.45 AND petalwidth < 1.7 AND petallength < 4.95
THEN Iris-versicolor
IF petallength >= 2.45 AND petalwidth < 1.7 AND petallength >= 4.95
THEN Iris-virginica
IF petallength >= 2.45 AND petalwidth >= 1.7 AND sepallength < 5.95
AND sepalwidth < 3.1 THEN Iris-virginica
IF petallength >= 2.45 AND petalwidth >= 1.7 AND sepallength < 5.95
AND sepalwidth >= 3.1 THEN Iris-versicolor
IF petallength >= 2.45 AND petalwidth >= 1.7 AND sepallength >= 5.95
THEN Iris-virginica
```

```
this new tree can validate 100.0%
integrate with rule set #1
D:\Integration_Rules_RF\Iris\tree\1\8.txt
---Initial rules for this iteration ---
IF petallength < 2.35 THEN Iris-setosa
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth < 1.65
THEN Iris-versicolor
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth < 3.1 THEN Iris-virginica
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth >= 3.1 THEN Iris-versicolor
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength < 6.35
THEN Iris-virginica
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength >= 6.35
THEN Iris-versicolor
IF petallength >= 5.05 THEN Iris-virginica
IF petallength < 2.45 THEN Iris-setosa
IF petallength >= 2.45 AND petalwidth < 1.7 AND petallength < 4.95
THEN Iris-versicolor
IF petallength >= 2.45 AND petalwidth < 1.7 AND petallength >= 4.95
THEN Iris-virginica
IF petallength >= 2.45 AND petalwidth >= 1.7 AND sepallength < 5.95
AND sepalwidth < 3.1 THEN Iris-virginica
IF petallength >= 2.45 AND petalwidth >= 1.7 AND sepallength < 5.95
AND sepalwidth >= 3.1 THEN Iris-versicolor
IF petallength >= 2.45 AND petalwidth >= 1.7 AND sepallength >= 5.95
THEN Iris-virginica
Correctly Classified Instances : 100.0%
Roll back the rule set
@@read tree from D:\Integration_Rules_RF\Iris\tree\1\9.txt
@@integrate with D:\Integration_Rules_RF\Iris\tree\1\9.txt
tree is -----
petallength < 2.45 : Iris-setosa (35/0)
petallength >= 2.45
|   petallength < 4.95
|   |   sepallength < 4.95
|   |   |   sepalwidth < 2.45 : Iris-versicolor (1/0)
|   |   |   sepalwidth >= 2.45 : Iris-virginica (2/0)
|   |   sepallength >= 4.95
|   |   |   petalwidth < 1.7 : Iris-versicolor (37/0)
|   |   |   petalwidth >= 1.7 : Iris-virginica (1/0)
|   petallength >= 4.95
|   |   petallength < 5.05
|   |   |   sepallength < 6.35 : Iris-virginica (2/0)
|   |   |   sepallength >= 6.35 : Iris-versicolor (1/0)
|   |   petallength >= 5.05 : Iris-virginica (29/0)
---tree to rules ---
IF petallength < 2.45 THEN Iris-setosa
IF petallength >= 2.45 AND petallength < 4.95 AND sepallength < 4.95
AND sepalwidth < 2.45 THEN Iris-versicolor
IF petallength >= 2.45 AND petallength < 4.95 AND sepallength < 4.95
AND sepalwidth >= 2.45 THEN Iris-virginica
IF petallength >= 2.45 AND petallength < 4.95 AND sepallength >= 4.95
AND petalwidth < 1.7 THEN Iris-versicolor
IF petallength >= 2.45 AND petallength < 4.95 AND sepallength >= 4.95
AND petalwidth >= 1.7 THEN Iris-virginica
IF petallength >= 2.45 AND petallength >= 4.95 AND petallength < 5.05
AND sepallength < 6.35 THEN Iris-virginica
```

```
IF petallength >= 2.45 AND petallength >= 4.95 AND petallength < 5.05
AND sepallength >= 6.35 THEN Iris-versicolor
IF petallength >= 2.45 AND petallength >= 4.95 AND petallength >=
5.05 THEN Iris-virginica
this new tree can validate 100.0%
integrate with rule set #1
D:\Integration_Rules_RF\Iris\tree\1\9.txt
---Initial rules for this iteration ---
IF petallength < 2.35 THEN Iris-setosa
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth < 1.65
THEN Iris-versicolor
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth < 3.1 THEN Iris-virginica
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth >= 3.1 THEN Iris-versicolor
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength < 6.35
THEN Iris-virginica
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength >= 6.35
THEN Iris-versicolor
IF petallength >= 5.05 THEN Iris-virginica
IF petallength < 2.45 THEN Iris-setosa
IF petallength >= 2.45 AND petallength < 4.95 AND sepallength < 4.95
AND sepalwidth < 2.45 THEN Iris-versicolor
IF petallength >= 2.45 AND petallength < 4.95 AND sepallength < 4.95
AND sepalwidth >= 2.45 THEN Iris-virginica
IF petallength >= 2.45 AND petallength < 4.95 AND sepallength >= 4.95
AND petalwidth < 1.7 THEN Iris-versicolor
IF petallength >= 2.45 AND petallength < 4.95 AND sepallength >= 4.95
AND petalwidth >= 1.7 THEN Iris-virginica
IF petallength >= 2.45 AND petallength >= 4.95 AND petallength < 5.05
AND sepallength < 6.35 THEN Iris-virginica
IF petallength >= 2.45 AND petallength >= 4.95 AND petallength < 5.05
AND sepallength >= 6.35 THEN Iris-versicolor
IF petallength >= 2.45 AND petallength >= 4.95 AND petallength >=
5.05 THEN Iris-virginica
Correctly Classified Instances : 100.0%
Roll back the rule set
@@read tree from D:\Integration_Rules_RF\Iris\tree\1\10.txt
@@integrate with D:\Integration_Rules_RF\Iris\tree\1\10.txt
tree is -----
petalwidth < 0.8 : Iris-setosa (37/0)
petalwidth >= 0.8
|   petallength < 5.05
|   |   petalwidth < 1.65 : Iris-versicolor (38/0)
|   |   petalwidth >= 1.65
|   |   |   sepallength < 6.35 : Iris-virginica (3/0)
|   |   |   sepallength >= 6.35 : Iris-versicolor (3/0)
|   petallength >= 5.05 : Iris-virginica (27/0)
---tree to rules ---
IF petalwidth < 0.8 THEN Iris-setosa
IF petalwidth >= 0.8 AND petallength < 5.05 AND petalwidth < 1.65
THEN Iris-versicolor
IF petalwidth >= 0.8 AND petallength < 5.05 AND petalwidth >= 1.65
AND sepallength < 6.35 THEN Iris-virginica
IF petalwidth >= 0.8 AND petallength < 5.05 AND petalwidth >= 1.65
AND sepallength >= 6.35 THEN Iris-versicolor
IF petalwidth >= 0.8 AND petallength >= 5.05 THEN Iris-virginica
this new tree can validate 100.0%
```

```
integrate with rule set #1
D:\Integration_Rules_RF\Iris\tree\1\10.txt
---Initial rules for this iteration ---
IF petallength < 2.35 THEN Iris-setosa
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth < 1.65
THEN Iris-versicolor
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth < 3.1 THEN Iris-virginica
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth >= 3.1 THEN Iris-versicolor
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength < 6.35
THEN Iris-virginica
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength >= 6.35
THEN Iris-versicolor
IF petallength >= 5.05 THEN Iris-virginica
IF petalwidth < 0.8 THEN Iris-setosa
IF petalwidth >= 0.8 AND petallength < 5.05 AND petalwidth < 1.65
THEN Iris-versicolor
IF petalwidth >= 0.8 AND petallength < 5.05 AND petalwidth >= 1.65
AND sepallength < 6.35 THEN Iris-virginica
IF petalwidth >= 0.8 AND petallength < 5.05 AND petalwidth >= 1.65
AND sepallength >= 6.35 THEN Iris-versicolor
IF petalwidth >= 0.8 AND petallength >= 5.05 THEN Iris-virginica
Correctly Classified Instances : 100.0%
Roll back the rule set
@@read tree from D:\Integration_Rules_RF\Iris\tree\1\4.txt
@@integrate with D:\Integration_Rules_RF\Iris\tree\1\4.txt
tree is -----
petallength < 2.45 : Iris-setosa (34/0)
petallength >= 2.45
|   petalwidth < 1.55
|   |   petallength < 4.95 : Iris-versicolor (39/0)
|   |   petallength >= 4.95 : Iris-virginica (2/0)
|   petalwidth >= 1.55
|   |   sepalwidth < 3.15 : Iris-virginica (26/0)
|   |   sepalwidth >= 3.15
|   |   |   sepallength < 6.35 : Iris-versicolor (3/0)
|   |   |   sepallength >= 6.35 : Iris-virginica (4/0)
---tree to rules ---
IF petallength < 2.45 THEN Iris-setosa
IF petallength >= 2.45 AND petalwidth < 1.55 AND petallength < 4.95
THEN Iris-versicolor
IF petallength >= 2.45 AND petalwidth < 1.55 AND petallength >= 4.95
THEN Iris-virginica
IF petallength >= 2.45 AND petalwidth >= 1.55 AND sepalwidth < 3.15
THEN Iris-virginica
IF petallength >= 2.45 AND petalwidth >= 1.55 AND sepalwidth >= 3.15
AND sepallength < 6.35 THEN Iris-versicolor
IF petallength >= 2.45 AND petalwidth >= 1.55 AND sepalwidth >= 3.15
AND sepallength >= 6.35 THEN Iris-virginica
this new tree can validate 100.0%
integrate with rule set #1
D:\Integration_Rules_RF\Iris\tree\1\4.txt
---Initial rules for this iteration ---
IF petallength < 2.35 THEN Iris-setosa
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth < 1.65
THEN Iris-versicolor
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
```

```
AND sepalwidth < 3.1 THEN Iris-virginica
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth >= 3.1 THEN Iris-versicolor
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength < 6.35
THEN Iris-virginica
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength >= 6.35
THEN Iris-versicolor
IF petallength >= 5.05 THEN Iris-virginica
IF petallength < 2.45 THEN Iris-setosa
IF petallength >= 2.45 AND petalwidth < 1.55 AND petallength < 4.95
THEN Iris-versicolor
IF petallength >= 2.45 AND petalwidth < 1.55 AND petallength >= 4.95
THEN Iris-virginica
IF petallength >= 2.45 AND petalwidth >= 1.55 AND sepalwidth < 3.15
THEN Iris-virginica
IF petallength >= 2.45 AND petalwidth >= 1.55 AND sepalwidth >= 3.15
AND sepallength < 6.35 THEN Iris-versicolor
IF petallength >= 2.45 AND petalwidth >= 1.55 AND sepalwidth >= 3.15
AND sepallength >= 6.35 THEN Iris-virginica
Correctly Classified Instances : 100.0%
Roll back the rule set
@@read tree from D:\Integration_Rules_RF\Iris\tree\1\6.txt
@@integrate with D:\Integration_Rules_RF\Iris\tree\1\6.txt
tree is -----
petallength < 2.45 : Iris-setosa (37/0)
petallength >= 2.45
|   petallength < 5.05
|   |   petalwidth < 1.65 : Iris-versicolor (36/0)
|   |   petalwidth >= 1.65
|   |   |   sepallength < 5.4 : Iris-virginica (1/0)
|   |   |   sepallength >= 5.4
|   |   |   |   sepallength < 5.95 : Iris-versicolor (1/0)
|   |   |   |   sepallength >= 5.95
|   |   |   |   |   sepallength < 6.35 : Iris-virginica (1/0)
|   |   |   |   |   sepallength >= 6.35 : Iris-versicolor (1/0)
|   petallength >= 5.05 : Iris-virginica (31/0)
---tree to rules ---
IF petallength < 2.45 THEN Iris-setosa
IF petallength >= 2.45 AND petallength < 5.05 AND petalwidth < 1.65
THEN Iris-versicolor
IF petallength >= 2.45 AND petallength < 5.05 AND petalwidth >= 1.65
AND sepallength < 5.4 THEN Iris-virginica
IF petallength >= 2.45 AND petallength < 5.05 AND petalwidth >= 1.65
AND sepallength >= 5.4 AND sepallength < 5.95 THEN Iris-versicolor
IF petallength >= 2.45 AND petallength < 5.05 AND petalwidth >= 1.65
AND sepallength >= 5.4 AND sepallength >= 5.95 AND sepallength < 6.35
THEN Iris-virginica
IF petallength >= 2.45 AND petallength < 5.05 AND petalwidth >= 1.65
AND sepallength >= 5.4 AND sepallength >= 5.95 AND sepallength >=
6.35 THEN Iris-versicolor
IF petallength >= 2.45 AND petallength >= 5.05 THEN Iris-virginica
this new tree can validate 100.0%
integrate with rule set #1
D:\Integration_Rules_RF\Iris\tree\1\6.txt
---Initial rules for this iteration ---
IF petallength < 2.35 THEN Iris-setosa
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth < 1.65
THEN Iris-versicolor
```

```
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth < 3.1 THEN Iris-virginica
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth >= 3.1 THEN Iris-versicolor
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength < 6.35
THEN Iris-virginica
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength >= 6.35
THEN Iris-versicolor
IF petallength >= 5.05 THEN Iris-virginica
IF petallength < 2.45 THEN Iris-setosa
IF petallength >= 2.45 AND petallength < 5.05 AND petalwidth < 1.65
THEN Iris-versicolor
IF petallength >= 2.45 AND petallength < 5.05 AND petalwidth >= 1.65
AND sepallength < 5.4 THEN Iris-virginica
IF petallength >= 2.45 AND petallength < 5.05 AND petalwidth >= 1.65
AND sepallength >= 5.4 AND sepallength < 5.95 THEN Iris-versicolor
IF petallength >= 2.45 AND petallength < 5.05 AND petalwidth >= 1.65
AND sepallength >= 5.4 AND sepallength >= 5.95 AND sepallength < 6.35
THEN Iris-virginica
IF petallength >= 2.45 AND petallength < 5.05 AND petalwidth >= 1.65
AND sepallength >= 5.4 AND sepallength >= 5.95 AND sepallength >=
6.35 THEN Iris-versicolor
IF petallength >= 2.45 AND petallength >= 5.05 THEN Iris-virginica
Correctly Classified Instances : 100.0%
Roll back the rule set
@@read tree from D:\Integration_Rules_RF\Iris\tree\1\7.txt
@@integrate with D:\Integration_Rules_RF\Iris\tree\1\7.txt
tree is -----
petalwidth < 0.8 : Iris-setosa (34/0)
petalwidth >= 0.8
|   petallength < 5.05
|   |   petalwidth < 1.45 : Iris-versicolor (28/0)
|   |   petalwidth >= 1.45
|   |   |   petallength < 4.75 : Iris-versicolor (5/0)
|   |   |   petallength >= 4.75
|   |   |   |   sepalwidth < 2.35 : Iris-virginica (1/0)
|   |   |   |   sepalwidth >= 2.35
|   |   |   |   |   sepallength < 5.95 : Iris-versicolor (2/0)
|   |   |   |   |   sepallength >= 5.95
|   |   |   |   |   |   sepallength < 6.15 : Iris-virginica (1/0)
|   |   |   |   |   |   sepallength >= 6.15 : Iris-versicolor (2/0)
|   petallength >= 5.05 : Iris-virginica (35/0)
---tree to rules ---
IF petalwidth < 0.8 THEN Iris-setosa
IF petalwidth >= 0.8 AND petallength < 5.05 AND petalwidth < 1.45
THEN Iris-versicolor
IF petalwidth >= 0.8 AND petallength < 5.05 AND petalwidth >= 1.45
AND petallength < 4.75 THEN Iris-versicolor
IF petalwidth >= 0.8 AND petallength < 5.05 AND petalwidth >= 1.45
AND petallength >= 4.75 AND sepalwidth < 2.35 THEN Iris-virginica
IF petalwidth >= 0.8 AND petallength < 5.05 AND petalwidth >= 1.45
AND petallength >= 4.75 AND sepalwidth >= 2.35 AND sepallength < 5.95
THEN Iris-versicolor
IF petalwidth >= 0.8 AND petallength < 5.05 AND petalwidth >= 1.45
AND petallength >= 4.75 AND sepalwidth >= 2.35 AND sepallength >=
5.95 AND sepallength < 6.15 THEN Iris-virginica
IF petalwidth >= 0.8 AND petallength < 5.05 AND petalwidth >= 1.45
AND petallength >= 4.75 AND sepalwidth >= 2.35 AND sepallength >=
```

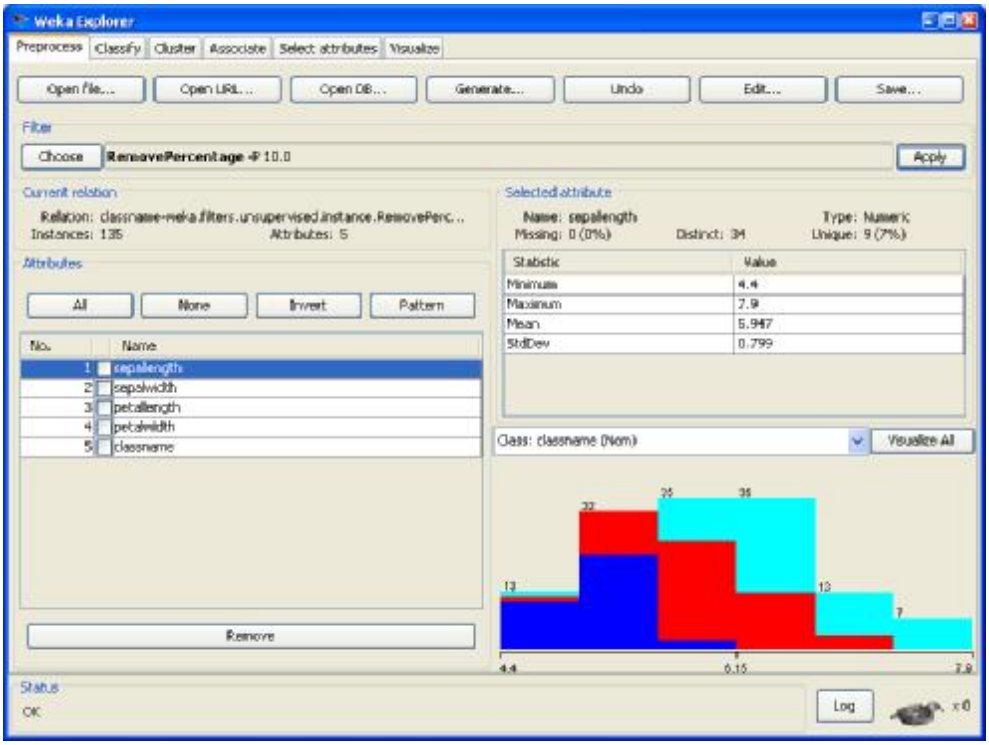
```
5.95 AND sepallength >= 6.15 THEN Iris-versicolor
IF petalwidth >= 0.8 AND petallength >= 5.05 THEN Iris-virginica
this new tree can validate 100.0%
integrate with rule set #1
D:\Integration_Rules_RF\Iris\tree\1\7.txt
---Initial rules for this iteration ---
IF petallength < 2.35 THEN Iris-setosa
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth < 1.65
THEN Iris-versicolor
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth < 3.1 THEN Iris-virginica
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth >= 3.1 THEN Iris-versicolor
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength < 6.35
THEN Iris-virginica
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength >= 6.35
THEN Iris-versicolor
IF petallength >= 5.05 THEN Iris-virginica
IF petalwidth < 0.8 THEN Iris-setosa
IF petalwidth >= 0.8 AND petallength < 5.05 AND petalwidth < 1.45
THEN Iris-versicolor
IF petalwidth >= 0.8 AND petallength < 5.05 AND petalwidth >= 1.45
AND petallength < 4.75 THEN Iris-versicolor
IF petalwidth >= 0.8 AND petallength < 5.05 AND petalwidth >= 1.45
AND petallength >= 4.75 AND sepalwidth < 2.35 THEN Iris-virginica
IF petalwidth >= 0.8 AND petallength < 5.05 AND petalwidth >= 1.45
AND petallength >= 4.75 AND sepalwidth >= 2.35 AND sepallength < 5.95
THEN Iris-versicolor
IF petalwidth >= 0.8 AND petallength < 5.05 AND petalwidth >= 1.45
AND petallength >= 4.75 AND sepalwidth >= 2.35 AND sepallength >=
5.95 AND sepallength < 6.15 THEN Iris-virginica
IF petalwidth >= 0.8 AND petallength < 5.05 AND petalwidth >= 1.45
AND petallength >= 4.75 AND sepalwidth >= 2.35 AND sepallength >=
5.95 AND sepallength >= 6.15 THEN Iris-versicolor
IF petalwidth >= 0.8 AND petallength >= 5.05 THEN Iris-virginica
Correctly Classified Instances : 100.0%
Roll back the rule set
---After trim rules ---
IF petallength < 2.35 THEN Iris-setosa
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth < 1.65
THEN Iris-versicolor
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth < 3.1 THEN Iris-virginica
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth >= 3.1 THEN Iris-versicolor
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength < 6.35
THEN Iris-virginica
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength >= 6.35
THEN Iris-versicolor
IF petallength >= 5.05 THEN Iris-virginica
---After eliminate redundant rule ---
IF petallength < 2.35 THEN Iris-setosa
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth < 1.65
THEN Iris-versicolor
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth < 3.1 THEN Iris-virginica
IF petallength >= 2.35 AND petallength < 4.95 AND petalwidth >= 1.65
AND sepalwidth >= 3.1 THEN Iris-versicolor
```

```
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength < 6.35
THEN Iris-virginica
IF petallength >= 4.95 AND petallength < 5.05 AND sepallength >= 6.35
THEN Iris-versicolor
IF petallength >= 5.05 THEN Iris-virginica
```

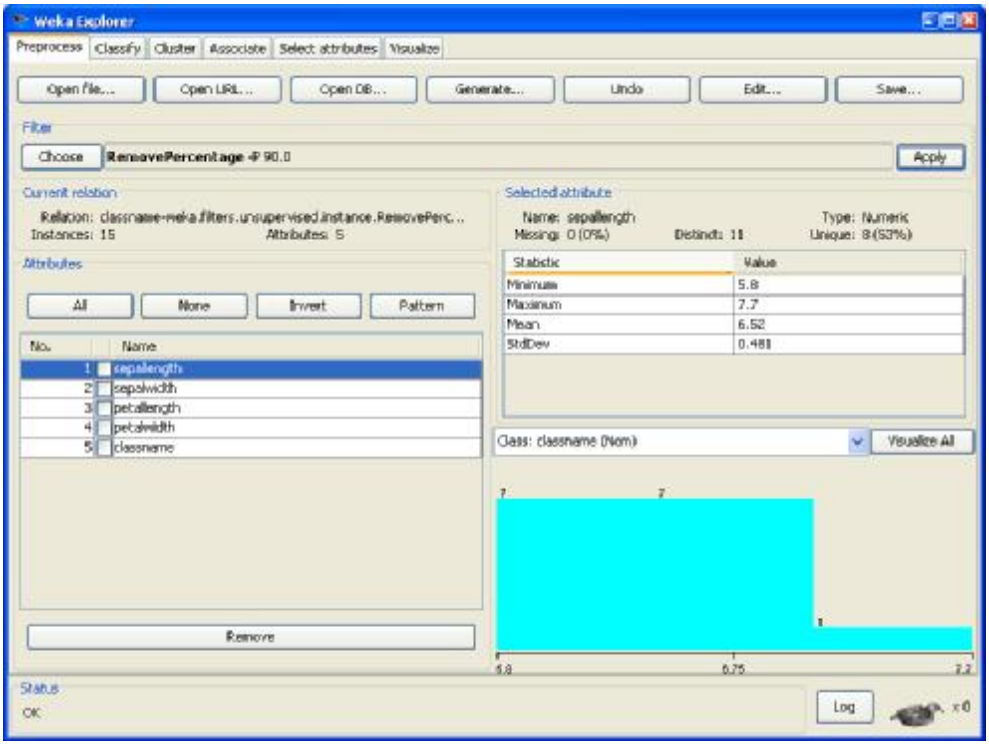
ภาคผนวก ฉ
วิธีการแบ่งชุดข้อมูลสอนและชุดข้อมูลตรวจสอบ

การแบ่งข้อมูลที่ใช้สำหรับการทดลอง เราจะทำการแบ่งชุดข้อมูลสอน 90% และชุดข้อมูลทดสอบ 10% ด้วยโปรแกรมเวกา โดยจะทำการเลือกชุดข้อมูลที่ต้องการแบ่ง จากรูปเราเลือกชุดข้อมูล Iris หลังจากนั้นเลือกตัวกรองที่ผู้ใช้กำหนดเอง (Unsupervised) และเลือกลักษณะระเบียน (Instance) แบบ RemovePercentage

รูปการแบ่งข้อมูลสอน 90 เปอร์เซ็นต์



รูปการแบ่งข้อมูลทดสอบ 10 เปอร์เซ็นต์



ภาคผนวก ซ

รายละเอียดผลการทดสอบ

Balance Scale

Fold#	RF (%)	No of Rule	J48 (%)	No of Rule	RFh (%)	No of Rule	RFl (%)	No of Rule
1	79.3651	945	79.3651	30	87.30159	426	92.06349	598
2	85.7143	766	82.5397	38	87.30159	346	92.06349	443
3	74.6032	880	79.3651	39	96.82539	372	93.65079	565
4	79.3651	879	80.9524	29	88.88888	495	90.47619	426
5	82.5397	879	82.5397	37	90.47619	435	90.47619	516
6	82.2581	920	83.871	47	91.93549	488	91.93549	479
7	79.0323	955	74.1935	34	95.16129	524	93.54839	543
8	82.2581	956	77.4194	28	96.77419	553	98.38710	733
9	69.3548	875	62.9032	32	82.25806	360	83.87096	476
10	79.0323	926	83.871	33	91.93549	324	90.32258	526
Average	79.3523	898.1	78.70201	34.7	90.88582	432.3	91.67947	530.5

Blood Transfusion

Fold#	RF (%)	No of Rule	J48 (%)	No of Rule	RFh (%)	No of Rule	RFl (%)	No of Rule
1	65.3333	1,194	80.00000	5	100.00000	959	93.33336	362
2	65.3333	1,094	70.66670	8	81.33333	668	96.00000	920
3	74.6667	1,203	76.00000	1	92.00000	802	97.33336	630
4	74.6667	1,178	84.00000	10	94.66667	379	98.66664	349
5	70.6667	1,113	78.66670	5	98.66667	612	97.33336	856
6	69.3333	1,161	73.33330	5	100.00000	443	96.00000	484
7	73.3333	1,159	82.66670	10	100.00000	502	100.00000	575
8	74.6667	1,142	73.33330	13	92.00000	884	97.33336	772
9	63.5135	1,120	77.02700	7	98.64865	321	100.00000	285
10	68.9189	1,135	79.72970	4	90.54054	760	100.00000	720
Average	70.04324	1149.9	77.54234	6.8	94.78559	633	97.60001	595.3

Haberman's Survival

Fold#	RF (%)	No of Rule	J48 (%)	No of Rule	RFh (%)	No of Rule	RFl (%)	No of Rule
1	64.5161	520	70.9677	2	93.54839	291	96.77419	345
2	67.7419	531	74.1935	1	96.77419	381	93.54839	356
3	70.9677	530	74.1935	1	87.09677	31	90.32258	305
4	70.9677	489	74.1935	5	87.09677	327	87.09677	414
5	70.96.77	558	74.1935	1	93.54839	329	100.0000	309
6	74.1935	527	77.4194	4	90.32258	364	93.54839	434
7	70.0000	562	66.6667	3	93.33334	227	100.0000	372
8	56.6667	542	73.3333	1	86.66666	188	100.0000	375
9	63.3333	503	70.0000	4	100.0000	280	93.33334	259
10	60.0000	529	70.0000	9	93.33334	264	93.33334	367
Average	59.83869	529.1	72.51611	3.1	92.17204	268.2	94.7957	353.6

Iris

Fold#	RF (%)	No of Rule	J48 (%)	No of Rule	RFh (%)	No of Rule	RFl (%)	No of Rule
1	93.3333	66	86.6667	3	93.33334	7	93.33334	20
2	100.0000	55	100.0000	4	100.00000	17	100.00000	19
3	100.0000	87	100.0000	5	100.00000	10	100.00000	19
4	100.0000	49	100.0000	4	100.00000	4	100.00000	10
5	93.3333	64	93.3333	4	93.33334	8	93.33334	13
6	93.3333	39	93.3333	3	100.00000	8	100.00000	13
7	93.3333	53	86.6667	3	100.00000	6	100.00000	6
8	86.6667	49	86.6667	4	86.66666	9	100.00000	14
9	93.3333	60	100.0000	4	86.66666	6	100.00000	16
10	100.0000	72	100.0000	3	100.00000	15	100.00000	19
Average	95.33332	59.4	94.66667	3.7	96	9	98.66667	14.9

Liver Disorders

Fold#	RF (%)	No of Rule	J48 (%)	No of Rule	RFh (%)	No of Rule	RFl (%)	No of Rule
1	60.0000	538	57.1429	21	85.71429	309	97.14286	327
2	57.1429	541	57.1429	19	100.00000	160	100.00000	266
3	71.4286	529	65.7143	12	100.00000	229	100.00000	301
4	62.8571	505	57.1429	27	100.00000	159	97.14286	232
5	65.7143	528	74.2857	18	97.14286	315	94.28571	342
6	88.2353	508	64.7059	7	100.00000	225	100.00000	297
7	73.5294	533	55.8824	18	97.05882	256	100.00000	288
8	73.5294	513	70.5882	17	100.00000	259	100.00000	359
9	61.7647	578	61.7647	18	100.00000	168	100.00000	333
10	82.3529	538	55.8824	23	100.00000	160	100.00000	298
Average	69.65546	531.1	62.02523	18	97.99159	224	98.85714	304.3

Pima Indians Diabetes

Fold#	RF (%)	No of Rule	J48 (%)	No of Rule	RFh (%)	No of Rule	RFl (%)	No of Rule
1	79.2208	842	79.2208	11	96.10390	664	97.40259	750
2	70.1299	845	74.0260	35	100.00000	456	96.10390	671
3	64.9351	829	71.4286	28	98.70130	587	97.40259	657
4	75.3247	888	74.0260	18	100.00000	524	98.70130	606
5	74.0260	859	75.3247	21	100.00000	592	100.00000	736
6	68.8312	853	64.9351	10	94.80519	517	96.10390	764
7	72.7273	887	74.0260	36	97.40259	699	97.40259	616
8	61.0390	841	72.7273	25	92.20779	587	93.50649	416
9	75.0000	866	65.7895	15	97.36842	609	97.36842	676
10	73.6842	906	77.6316	21	97.36842	609	98.68421	722
Average	71.49182	861.6	72.91356	22	97.39576	584.4	97.26759	661.4

Statlog (Austrlian Credit Approval)

Fold#	RF (%)	No of Rule	J48 (%)	No of Rule	RFh (%)	No of Rule	RFl (%)	No of Rule
1	89.8551	746	89.8551	18	100.00000	447	97.10145	426
2	91.3043	755	89.8551	23	100.00000	385	100.00000	477
3	88.8551	812	88.4058	17	98.55073	305	100.00000	392
4	82.6087	712	82.6087	27	100.00000	341	98.55073	505
5	79.7101	695	82.6087	17	98.55073	465	95.65218	352
6	85.5072	777	85.5072	20	100.00000	503	97.10145	492
7	79.7101	647	81.1594	27	94.20289	383	98.55073	463
8	85.5072	773	79.7101	23	94.20289	323	97.10145	322
9	84.0580	716	85.5072	30	100.00000	328	97.10145	633
10	88.4058	725	88.4058	27	95.65218	292	100.00000	225
Average	85.55216	735.8	85.36231	22.9	98.11594	377.2	98.11594	428.7

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวนภาพร ศิริกุลวิริยะ เกิดวันที่ 6 กันยายน พ.ศ.2523 สถานที่เกิดจังหวัดเลย สำเร็จการศึกษาระดับปริญญาวิทยาศาสตรบัณฑิต (วท.บ.) สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยสยาม ในปีการศึกษา 2545 และเข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2551