การตัดเล็มอย่างอ่อนสำหรับต้นไม้ตัดสินใจโดยการใช้แบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก

นายก้องศักดิ์ จงเกษมวงศ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2543
ISBN 974-13-0080-8

SOFT PRUNING FOR DECISION TREE USING THE BACKPROPAGATION NEURAL NETWORK

MR. KONGSAK CHONGKASEMWONGSE

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science

Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University

Academic Year 2000

ISBN 974-13-0080-8

หัวข้อวิทยานิพนธ์	การตัดเล็มอย่างอ่อนสำหรับต้นไม้ตัดสินใจโดยการใช้แบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก
โดย	นายก้องศักดิ์ จงเกษมวงศ์
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษา	อาจารย์ ดร.บุญเสริม กิจศิริกุล

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร.สมศักดิ์ ปัญญาแก้ว)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.สมชาย ประสิทธิ์จูตระกูล)

.....อาจารย์ที่ปรึกษา
(อาจารย์ ดร.บุญเสริม กิจศิริกุล)

.....กรรมการ
(อาจารย์ ดร.ฐิต ศิริบูรณ์)

.....กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา)

ก้องศักดิ์ จงเกษมวงศ์ : การตัดเล็มอย่างอ่อนสำหรับต้นไม้ตัดสินใจโดยการใช้แบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก (SOFT PRUNING FOR DECISION TREE USING THE BACKPROPAGATION NEURAL NETWORK) อ. ที่ปรึกษา : อ.ดร.บุญเสริม กิจศิริกุล, 62 หน้า. ISBN 974-13-0080-8.

การเรียนรู้ต้นไม้ตัดสินใจเป็นเทคนิคการเรียนรู้ของเครื่องวิธีการหนึ่งที่ใช้กันอย่างแพร่หลายในปัจจุบัน เมื่อสร้างต้นไม้ตัดสินใจจากข้อมูลสอน โดยเฉพาะข้อมูลที่มีสัญญาณรบกวนนั้น ต้นไม้ตัดสินใจที่ได้อาจอิงกับข้อมูลมากเกินไป ทำให้ต้นไม้ตัดสินใจที่ได้มีขนาดใหญ่มาก และอาจใช้งานได้ไม่ดีกับข้อมูลใหม่ในอนาคต เทคนิคทั่วไปสำหรับแก้ไขการอิงกับข้อมูลมากไป ก็คือ การตัดเล็มต้นไม้ตัดสินใจ วิธีการต่าง ๆ สำหรับการตัดเล็มต้นไม้ตัดสินใจได้ถูกนำเสนอ ทุกวิธีการจะทำโดยการตัดโนดบางโนดออกจากต้นไม้เพื่อลดขนาดของต้นไม้ แต่ในบางครั้งโนดที่ตัดไปนั้นอาจมีความสำคัญอยู่ หรืออาจจะมีประโยชน์ในการแยกแยะข้อมูลในอนาคต

งานวิจัยนี้ได้เสนอวิธีการใหม่ ที่ให้น้ำหนักแก่โนดตามความสำคัญของโนดนั้น ๆ เรียกวิธีการนี้ว่าการตัดเล็มอย่างอ่อน ระดับความสำคัญหรือน้ำหนักของโนดหนึ่ง ๆ จะได้มาจากแบ็กพรอพาเกชันนิวรอล-เน็ตเวิร์ก ผลการทดลองกับข้อมูลทั้งหมด 20 ชุดข้อมูล เพื่อเปรียบเทียบวิธีการที่นำเสนอกับต้นไม้ตัดสินใจที่ไม่ได้ทำการตัดเล็ม และต้นไม้ตัดสินใจที่ทำการตัดเล็ม สรุปได้ว่าผลของงานวิจัยนี้ให้ประสิทธิภาพที่ดีที่สุด

ภาควิชา วิศวกรรมคอมพิวเตอร์	ลายมือชื่อนิสิต
สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์	ลายมือชื่ออาจารย์ที่ปรึกษา
ปีการศึกษา 2543	ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

4170214721 : MAJOR COMPUTER SCIENCE
KEY WORD: DECISION TREE / PRUNING / BACKPROPAGATION NEURAL NETWORK

KONGSAK CHONGKASEMWONGSE: SOFT PRUNING FOR DECISION TREE USING THE BACKPROPAGATION NEURAL NETWORK. THESIS ADVISOR : BOONSERM KIJSIRIKUL, Ph.D, 62 pp. ISBN 974-13-0080-8.

Decision tree learning is a machine learning technique that is in widespread use nowadays. When we construct a decision tree from some data, especially noisy data, the obtained tree may overfit the data and may be very large. This result makes the tree not perform well on new data. The commonly used technique for preventing the overfitting is to prune the decision trees. Many methods for decision tree pruning have been proposed, and all of them remove some nodes from the tree to reduce its size. However, some removed nodes may have a significance level or some contribution in classifying new data.

Instead of absolutely removing nodes, this thesis presents a new method that gives weights to nodes according to their significance. We call this method "soft pruning". The significance level or the weight of a node is determined by a backpropagation neural network. We run experiments on twenty domains to compare our method with error-based pruning that is one of the most effective method for tree pruning. The results demonstrate that our method outperforms error-based pruning.

Department Computer Engineering ..... Student's signature .....
Fields of study Computer Science ..... Advisor's signature .....
Academic year 2000 ..... Co-Advisor's signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างดียิ่งของ อ. ดร.บุญเสริม กิจศิริกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ได้ให้คำแนะนำและข้อคิดเห็นต่าง ๆ ในการวิจัยมาด้วยดีตลอด และขอขอบคุณ ผศ. ดร.สมชาย ประสิทธิ์จูตระกูล อ. ดร.ฐิต ศิริบูรณ์ และ ผศ. ดร.ประภาส จงสถิตย์วัฒนา กรรมการวิทยานิพนธ์ที่กรุณาเสียสละเวลาให้คำแนะนำ ตรวจและแก้ไขวิทยานิพนธ์ฉบับนี้

ขอขอบคุณสมาชิกห้องปฏิบัติการอัจฉริยภาพเครื่องกลและการค้นพบความรู้ (MIND LAB) และเพื่อน ๆ จากห้องปฏิบัติการทั้งหมด ที่ให้คำแนะนำ และความรู้ต่าง ๆ มากมาย ตลอดจนความบันเทิงต่าง ๆ

สุดท้ายนี้ ผู้วิจัยใคร่กราบขอบพระคุณบิดามารดา และครอบครัวที่คอยสนับสนุน และให้กำลังใจแก่ผู้วิจัยเสมอมาจนสำเร็จการศึกษา

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันการเรียนรู้ของเครื่อง (machine learning) ได้เป็นส่วนหนึ่งที่ก่อให้เกิดงานต่าง ๆ มากมายเพื่อใช้ทำงานแทนมนุษย์ หรือทำให้มนุษย์มีความสะดวกสบายขึ้น เช่น การรู้จำตัวอักษร (character recognition) เพื่อความรวดเร็วในการจัดเก็บเอกสาร การรู้จำเสียง (speech recognition) เพื่อใช้ในการจดจำเสียงและออกคำสั่งกับคอมพิวเตอร์ เป็นต้น

ต้นไม้ตัดสินใจ (decision tree) เป็นวิธีการเรียนรู้ของเครื่องวิธีหนึ่งที่มีการใช้งานกันอย่างแพร่หลาย เนื่องจากมีคุณสมบัติในการแยกแยะ (classification) ข้อมูลได้ตั้งแต่สองกลุ่มขึ้นไป การสร้างต้นไม้ตัดสินใจในบางครั้งเมื่อสร้างเสร็จแล้วจะได้ต้นไม้ที่มีขนาดใหญ่มาก ทั้งนี้อาจเป็นผลจากข้อมูลสอน (training data) ที่ใช้ในการสร้างต้นไม้ตัดสินใจมีสัญญาณรบกวน (noise data) จึงทำให้ต้นไม้ที่ได้อิงกับข้อมูลที่ใช้ในการสร้างมากเกินไป หากนำต้นไม้ตัดสินใจนี้ไปใช้งานอาจเกิดความผิดพลาดมากกว่าต้นไม้ที่มีขนาดเล็กได้ ดังนั้นจึงมีวิธีการตัดเล็มต้นไม้ (tree pruning) เพื่อทำให้ต้นไม้มีขนาดเล็กลงและสามารถนำไปใช้งานได้มีประสิทธิภาพมากกว่าต้นไม้ที่มีขนาดใหญ่

วิธีการตัดเล็มต้นไม้ตัดสินใจที่มีอยู่ในปัจจุบัน ส่วนใหญ่เป็นการใช้วิธีทางสถิติเพื่อตรวจสอบว่าควรจะตัดโนดภายใน (internal node) ของต้นไม้โนดใดออกไป โนดในต้นไม้แบ่งออกเป็น 2 ประเภทคือ (1) โนดภายในซึ่งทดสอบคุณสมบัติ (attribute) หนึ่ง ๆ ของข้อมูล (2) โนดใบซึ่งแสดงกลุ่ม (class) ของข้อมูล ในการตัดโนดนั้นจะพิจารณาตัดโนดภายในแล้วแทนที่ด้วยโนดใบ โดยเปรียบเทียบอัตราความผิดพลาด (error rate) ของต้นไม้ตัดสินใจว่า อัตราความผิดพลาดของต้นไม้ซึ่งคงโนดภายในโนดหนึ่งไว้ กับอัตราความผิดพลาดของต้นไม้ที่ได้จากการแทนที่ตำแหน่งของโนดนั้นด้วยกลุ่มที่เป็นไปได้ หากค่าอัตราความผิดพลาดที่ได้ของการแทนที่ด้วยกลุ่มมีค่าน้อยกว่า ก็จะแทนโนดที่ทดสอบคุณสมบัตินั้นด้วยกลุ่มทันที ด้วยวิธีการนี้เองทำให้ต้นไม้ตัดสินใจที่ได้จากการตัดเล็มมีขนาดเล็กลง อย่างไรก็ตาม ในการตัดเล็มอาจทำให้ประสิทธิภาพของต้นไม้ตัดสินใจน้อยลงกว่าต้นไม้ตัดสินใจที่ยังไม่ได้ตัดเล็มก็เป็นได้ เนื่องจากส่วนที่ถูกตัดเล็มออกไปอาจมีความสำคัญอยู่

วิทยานิพนธ์ฉบับนี้นำเสนอวิธีการตัดเล็มอย่างอ่อน โดยนำต้นไม้ตัดสินใจที่ยังไม่ได้ทำการตัดเล็มมาเข้าสู่วิธีการแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก และออกแบบโครงสร้างของนิวรอลเน็ตเวิร์กที่เปรียบเสมือนกับกฎที่ได้จากต้นไม้ตัดสินใจ วิธีการนี้จะปรับเปลี่ยนน้ำหนักของคุณสมบัติในต้นไม้ตัดสินใจเปรียบเสมือนปรับเปลี่ยนความสำคัญของคุณสมบัติต่าง ๆ ของต้นไม้ตัดสินใจ หรืออีกนัยหนึ่งก็คือการตัดเล็มอย่างอ่อนในความหมายที่ว่า การตัดเล็มด้วยวิธีการนี้จะไม่ตัดกิ่งของต้นไม้ทิ้งเลยทีเดียว แต่จะให้น้ำหนักมากกับคุณสมบัติที่สำคัญและให้น้ำหนักน้อยกับคุณสมบัติที่ไม่สำคัญ ค่าของน้ำหนักได้จากการสอนโดยใช้แบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก ผลการทดลองที่ได้จะทำการเปรียบเทียบกับการตัดเล็มโดยใช้ค่าความผิดพลาด (Error-Based Pruning)

1.2 วัตถุประสงค์ของการวิจัย

เพื่อศึกษาและประยุกต์ใช้ต้นไม้ตัดสินใจร่วมกับแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กในการตัดเล็มอย่างอ่อน (soft pruning)

1.3 ขอบเขตของการวิจัย

1. ใช้ระบบ C4.5 รุ่น 8 บนระบบปฏิบัติการยูนิกซ์ (unix) ในการสร้างต้นไม้ตัดสินใจ
2. ใช้ระบบ Aspirin/MIGRAINES Neural Network Software รุ่น 6.0 ในการตัดเล็มอย่างอ่อน
3. ข้อมูลที่ใช้ในการทดสอบเลือกมาจากฐานข้อมูลใน Department of Information and Computer Science, University of California, Irvine [2] โดยใช้ฐานข้อมูลประมาณ 20 ฐานข้อมูล
4. ต้นไม้ตัดสินใจที่นำมาทดสอบมีขนาดไม่เกิน 1000 โนด
5. วิธีการตัดเล็มที่นำมาเปรียบเทียบจะเป็นวิธีการอื่น เช่น การตัดเล็มโดยใช้ค่าความผิดพลาด

1.4 ขั้นตอนการดำเนินการ

1. ศึกษาแนวคิดและทฤษฎีการเรียนรู้ของวิธีการต้นไม้ตัดสินใจ
2. ศึกษาแนวคิดและทฤษฎีการเรียนรู้ของวิธีการแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก
3. เลือกฐานข้อมูลจาก UCI Repository เพื่อนำมาทดสอบ
4. สร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กตามแต่ละฐานข้อมูลจากต้นไม้ตัดสินใจ และทำการทดสอบ
5. วิเคราะห์และเปรียบเทียบผลที่ได้จากต้นไม้ตัดสินใจที่ตัดเล็มแล้ว กับแบ็กพรอพาเกชันนิวรอล-เน็ตเวิร์ก
6. สรุปผลการวิจัย ข้อเสนอแนะ และจัดทำรายงานวิทยานิพนธ์

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. สามารถนำโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กไปใช้งานได้
2. เป็นแนวทางในการทำวิจัยต่อไป

1.6 ผลงานที่ตีพิมพ์จากงานวิจัย

1. วิทยานิพนธ์นี้ได้ตีพิมพ์และนำเสนอในงานประชุมวิชาการวิทยาการคอมพิวเตอร์และวิศวกรรมคอมพิวเตอร์แห่งชาติ 2543 (The National Computer Science and Engineering Conference : NCSEC 2000) เมื่อวันที่ 16-17 พฤศจิกายน พ.ศ.2543 ในบทความเรื่อง "การตัดเล็ม

อย่างอ่อนสำหรับต้นไม้ตัดสินใจโดยการใช้โครงข่ายประสาทเทียมแบบแพร่กระจายย้อนหลัง" โดยผู้นำเสนอ คือ ก้องศักดิ์ จงเกษมวงศ์ และบุญเสริม กิจศิริกุล

2. วิทยานิพนธ์เล่มนี้ได้ตีพิมพ์เป็นส่วนหนึ่งในวารสารวิชาการนานาชาติ Machine Learning Journal ในบทความเรื่อง “Approximate Match of Rules Using Backpropagation Neural Networks” โดย Boonserm Kijsirikul Sukree Sinthupinyo และ Kongsak Chongkasemwongse ซึ่งจะตีพิมพ์ภายในปี 2001 นี้

3. วิทยานิพนธ์เล่มนี้ได้ตีพิมพ์และนำเสนอในงานประชุมวิชาการ Artificial Intelligence and Soft Computing (ASC 2001) ในวันที่ 21-24 พฤษภาคม พ.ศ. 2544 ในบทความเรื่อง “Soft-Pruning Decision Tree” โดยผู้นำเสนอคือ Boonserm Kijsirikul และ Kongsak Chongkasemwongse

1.7 เนื้อหาและรูปแบบการนำเสนอวิทยานิพนธ์

เนื้อหาของวิทยานิพนธ์ฉบับนี้ถูกแบ่งออกเป็น 5 บทดังนี้ คือ บทที่ 1 เป็นบทนำ บทที่ 2 จะกล่าวถึงงานวิจัยและทฤษฎีต่าง ๆ ที่เกี่ยวข้อง เช่น ต้นไม้ตัดสินใจ การตัดเล็มต้นไม้ตัดสินใจโดยใช้ค่าความผิดพลาด วิธีการแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก เป็นต้น บทที่ 3 กล่าวถึงขั้นตอนวิธีการตัดเล็มอย่างอ่อน โดยการนำเทคนิคแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กมาใช้ ส่วนบทที่ 4 จะกล่าวถึงการทดลองและผลการทดลองของขั้นตอนวิธีที่นำเสนอ และบทที่ 5 ซึ่งเป็นบทสุดท้ายจะเป็นบทสรุปของการวิจัยรวมทั้งข้อเสนอแนะต่าง ๆ

บทที่ 2
งานวิจัยและทฤษฎีที่เกี่ยวข้อง

2.1 งานวิจัยที่เกี่ยวข้อง

ในบทนี้กล่าวถึงงานวิจัยที่เกี่ยวกับวิทยานิพนธ์ฉบับนี้ ได้แก่ การประยุกต์การโปรแกรมตรรกะเชิงอุปนัยและแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กในการรู้จำตัวพิมพ์อักษรไทย และ การวิเคราะห์เปรียบเทียบขั้นตอนวิธีสำหรับการตัดเล็มต้นไม้ตัดสินใจ

2.1.1 การประยุกต์การโปรแกรมตรรกะเชิงอุปนัยและแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กในการรู้จำตัวพิมพ์อักษรไทย [1]

สุกรี สินธุภิญโญ ได้วิจัยโดยใช้วิธีการโปรแกรมตรรกะเชิงอุปนัย (Inductive Logic Programming: ILP) ในการสร้างกฎเพื่อรู้จำตัวอักษรภาษาไทย การเรียนรู้ของวิธีการโปรแกรมตรรกะเชิงอุปนัยนี้ เป็นวิธีการเรียนรู้จากตัวอย่างที่แบ่งออกเป็น 2 กลุ่ม คือ ตัวอย่างบวก (positive example) ซึ่งเป็นกลุ่มตัวอย่างที่ตรงแนวคิด (concept) ที่ต้องการ และตัวอย่างลบ (negative example) ซึ่งเป็นกลุ่มตัวอย่างที่ไม่ตรงกับแนวคิดที่ต้องการ ดังนั้นกฎที่ได้จึงเป็นกฎที่เหมาะกับการแบ่งตัวอย่างออกเป็น 2 กลุ่ม คือ ถ้าตัวอย่างที่นำมาทดสอบตรงกับกฎพอดี จะถือว่าตัวอย่างนั้นตรงกับแนวคิด แต่ถ้าไม่ตรงพอดีจะรู้จำตัวอย่างนั้นเป็นตัวอย่างไม่ตรงกับแนวคิด เมื่อนำวิธีการโปรแกรมตรรกะเชิงอุปนัยมาประยุกต์ใช้กับงานที่ต้องการแบ่งตัวอย่างออกเป็นหลายกลุ่ม ดังเช่นการรู้จำตัวอักษรภาษาไทย ซึ่งกฎที่ได้ต้องนำไปใช้แบ่งกลุ่มตัวอย่างเป็น 77 ตัวอักษร หากมีตัวอย่างที่ไม่ตรงพอดีกับกฎข้อหนึ่งข้อใดเลยเช่น ภาพที่มีสัญญาณรบกวน ภาพที่ไม่ชัดเจน เป็นต้น จะไม่สามารถจำแนกตัวอย่างได้

ในงานวิจัยดังกล่าวได้นำวิธีการแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กมาทดลองเลือกกฎที่ใกล้เคียงในกรณีที่ไม่สามารถเลือกกฎที่ตรงพอดีได้ วิธีการแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กนี้สามารถจำแนกตัวอย่างออกเป็นหลายกลุ่มได้ โดยกำหนดค่าความสำคัญให้กับเงื่อนไขภายในกฎ (สัญพจน์) และปรับเปลี่ยนความสำคัญของแต่ละเงื่อนไขโดยใช้แบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก งานวิจัยดังกล่าวได้ทดลองสร้างโครงสร้างของนิวรอลเน็ตเวิร์กสองโครงสร้างคือ (1) แบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กที่ใช้จำนวนสัญพจน์ที่ตรง และจำนวนสัญพจน์ที่ไม่ตรงกับตัวอย่างเป็นอินพุตเวกเตอร์ และ (2) นิวรอลเน็ตเวิร์กที่ใช้ค่าความจริงของแต่ละสัญพจน์เป็นอินพุตเวกเตอร์

ผลการทดลองได้แบ่งกลุ่มตัวอย่างออกเป็นสองกลุ่มคือ กลุ่มที่ใช้เรียนรู้ มีจำนวน 1,078 ตัวอย่าง และกลุ่มที่ใช้ทดสอบ มีจำนวน 2,156 ตัวอย่าง วิธีการที่นำมาทดลองเปรียบเทียบ คือ วิธีการฟัซซีโลจิกและวิธีการซินแทกติก ให้อัตราการรู้จำ 87.62% วิธีการโปรแกรมตรรกะเชิงอุปนัยเพียงอย่างเดียวให้อัตราการรู้จำ 84.97% วิธีการแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กเพียงอย่างเดียว ให้อัตราการรู้จำ 84.25% วิธีการโปรแกรมตรรกะเชิงอุปนัยรวมกับแบ็กพรอพาเกชันเน็ตเวิร์กที่ใช้จำนวนสัญพจน์ที่ตรงและไม่ตรงกับตัวอย่างเป็นอินพุตเวกเตอร์ ให้อัตราการรู้จำ 92.55% และวิธีการโปรแกรมตรรกะเชิง

อุปนัยรวมกับแบ็กพรอพาเกชันเน็ตเวิร์กที่ใช้ค่าความจริงของแต่ละสัญพจน์เป็นอินพุตเวกเตอร์ ให้อัตราการรู้จำ 94.26%

2.1.2 การวิเคราะห์เปรียบเทียบขั้นตอนวิธีสำหรับการตัดเล็มต้นไม้ตัดสินใจ [3]

Esposito, F., Malerba, D., Semeraro, G. และ Kay, J. ได้วิจัยเปรียบเทียบการตัดเล็มต้นไม้ตัดสินใจในวิธีการต่าง ๆ คือ การตัดเล็มแบบความผิดพลาดลดลง (Reduced Error Pruning) การตัดเล็มแบบความผิดพลาดในแง่ร้าย (Pessimistic Error Pruning) การตัดเล็มแบบความผิดพลาดน้อยที่สุด (Minimum Error Pruning) การตัดเล็มแบบค่าวิกฤต (Critical Value Pruning) การตัดเล็มแบบค่าความซับซ้อน (Cost-Complexity Pruning) และ การตัดเล็มโดยใช้ค่าความผิดพลาด (Error-Based Pruning) พร้อมทั้งบอกข้อดีและข้อเสียของการตัดเล็มทั้งหกวิธีนี้ โดยงานวิจัยดังกล่าวได้กล่าวถึงปัจจัยที่เกี่ยวข้องในการตัดเล็มต้นไม้ตัดสินใจ คือ

- การคำนวณอัตราความผิดพลาด ซึ่งในแต่ละวิธีการจะใช้การคำนวณหาค่าความผิดพลาดที่แตกต่างกันออกไป แต่หลักการที่ใช้จะมีส่วนที่เหมือนกันคือ หากค่าอัตราความผิดพลาดที่ได้หลังการตัดเล็มมีค่าน้อยกว่าค่าอัตราความผิดพลาดก่อนการตัดเล็มต้นไม้ตัดสินใจ ก็จะตัดเล็มที่กิ่งนั้น ๆ ได้
- ข้อมูลที่ใช้ในการตัดเล็ม แบ่งเป็นสองแบบคือ (1) ข้อมูลสอน โดยในการตัดเล็มต้นไม้ตัดสินใจของวิธีการนี้ จะใช้ข้อมูลที่ใช้สอนเป็นข้อมูลที่ตัดเล็มด้วย และ (2) ข้อมูลการตัดเล็ม เป็นการแบ่งข้อมูลออกมาส่วนหนึ่งซึ่งไม่ได้เป็นข้อมูลสอนหรือข้อมูลทดสอบ แต่เป็นข้อมูลที่ใช้ในการตัดเล็มโดยเฉพาะ
- วิธีการตัดเล็ม แบ่งออกเป็นสองวิธีการคือ (1) การตัดเล็มก่อนหน้า (pre-pruning) วิธีการตัดเล็มวิธีนี้ จะทำในขณะที่สร้างต้นไม้ตัดสินใจ ทุกครั้งที่จะเพิ่มโนดลงไปในต้นไม้ตัดสินใจ ก็จะคำนวณว่าโนดนั้น ๆ สามารถตัดเล็มออกไปได้หรือไม่ หากทำการตัดเล็มได้ ก็จะไม่เพิ่มโนดนั้น ๆ เข้าไป และ (2) การตัดเล็มภายหลัง (post-pruning) เป็นการตัดเล็มที่ทำหลังจากสร้างต้นไม้ตัดสินใจเสร็จสิ้นแล้ว และค่อยมาเข้าสู่วิธีการตัดเล็มต่อไป

ในงานวิจัยดังกล่าวได้ใช้ฐานข้อมูลใน Department of Information and Computer Science, University of California, Irvine เป็นตัวเปรียบเทียบวิธีการตัดเล็มทั้งหกวิธี พบว่าการตัดเล็มโดยใช้ค่าความผิดพลาดเป็นวิธีการที่ดีมากกว่าหรือใกล้เคียงกับวิธีอื่น ๆ

2.2 ทฤษฎีที่เกี่ยวข้อง

ทฤษฎีต่าง ๆ ที่เกี่ยวข้องกับวิทยานิพนธ์ฉบับนี้ ได้แก่ ต้นไม้ตัดสินใจ การตัดเล็มต้นไม้ตัดสินใจโดยใช้ค่าความผิดพลาด วิธีการแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก และการเปรียบเทียบขั้นตอนวิธี

2.2.1 ต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจเป็นโครงสร้างที่ประกอบขึ้นจาก ราก (root) โนด (node) กิ่ง (branch) และ ใบ (leaf) ช่วยในการตัดสินใจหรือตอบคำถามในเรื่องเฉพาะที่ต้นไม้นี้เก็บไว้ โดยเริ่มจากส่วนราก แล้วไล่ลงไปตาม โนด กิ่ง จนกระทั่งถึง ใบ ซึ่งเป็นคำตอบหรือการตัดสินใจ

1. ราก เป็นจุดเริ่มต้นหรือ โนดแรกของคำถาม โดยคำตอบจะเป็นค่าที่เป็นไปได้ของคุณสมบัติไม่แบ่งกลุ่ม (non-category attribute) บนข้อมูลตัวอย่าง ถ้าคำตอบตรงกับค่าใดบนโนดนี้ก็จะวิ่งไปสู่กิ่งหรือใบต่อไป
2. โนด เป็นจุดของคำถามตามคุณสมบัติไม่แบ่งกลุ่ม
3. กิ่ง เป็นค่าที่เป็นไปได้ตามคุณสมบัติไม่แบ่งกลุ่ม ซึ่งจะนำไปสู่โนดหรือใบ
4. ใบ เป็นคำตอบหรือการตัดสินใจ โดยจะเป็นค่าที่เป็นไปได้ของคุณสมบัติแบ่งกลุ่ม (category attribute)
5. เส้นทาง (tree path) เป็นทางเดินตั้งแต่รากจนถึงใบแต่ละใบ ซึ่งจะนำไปสู่กฎต่อไป

ระบบต้นไม้ตัดสินใจที่ใช้ในงานวิทยานิพนธ์ฉบับนี้ เป็นระบบ C4.5 [4][6] ซึ่งเป็นโปรแกรมเรียนรู้ต้นไม้ตัดสินใจที่รู้จักกันอย่างแพร่หลาย พัฒนาโดย J. Ross Quinlan โดยพัฒนาต่อมาจาก ID3 [5] เป็นวิธีการเรียนรู้จากตัวอย่างที่อาศัยวิธีการจัดหมวดหมู่ (classification model) จากตัวอย่างเฉพาะที่เรียกว่าข้อมูลสอนแล้วสร้างเป็นต้นไม้ตัดสินใจ

ข้อมูลสอนจะมีลักษณะคล้ายกับข้อมูลในฐานข้อมูลแบบสัมพันธ์ (relational database) ที่ประกอบด้วย แถว (record) หรือในที่นี้เรียกว่าตัวอย่าง (case) และ สดมภ์ (column) หรือในที่นี้เรียกว่าคุณสมบัติ (attribute) ซึ่งมีด้วยกัน 2 ชนิด ดังตัวอย่างในตารางที่ 2.1 คือ

1. คุณสมบัติแบ่งกลุ่ม (category attribute) หรือในที่นี่จะเรียกว่า กลุ่ม (class) เป็นคุณสมบัติที่กำหนดว่าตัวอย่างนั้น ๆ ถูกจัดอยู่ในกลุ่มไหน โดยจะมีเพียง 1 คุณสมบัติในแต่ละชุดข้อมูล และข้อมูลที่เก็บจะเป็นชนิดไม่ต่อเนื่อง (discrete value) เท่านั้น เช่น {ใช่, ไม่ใช่}, {ถูก, ผิด} เป็นต้น
2. คุณสมบัติไม่แบ่งกลุ่ม (non-category attribute) หรือในที่นี่จะเรียกว่า คุณสมบัติ (attribute) เป็นชุดข้อมูลที่บ่งบอกถึงคุณสมบัติต่าง ๆ ของตัวอย่างแต่ละตัวอย่าง โดยแต่ละคุณสมบัติอาจจะเก็บข้อมูลได้ทั้งชนิด ค่าต่อเนื่อง (continuous values) เช่น ส่วนสูง, น้ำหนัก เป็นต้น หรือค่าไม่ต่อเนื่อง เช่น สีผม, อาชีพ เป็นต้น

ตารางที่ 2.1 ตัวอย่างคุณสมบัติต่าง ๆ ของการตัดสินใจเล่นกอล์ฟ

ตัวอย่างกลุ่มข้อมูล (class)

กลุ่มข้อมูล	**ค่าที่เป็นไปได้**
การออกรอบ	ออกรอบ, ไม่ออกรอบ

ตัวอย่างคุณสมบัติ (attribute)

คุณสมบัติ	**ค่าที่เป็นไปได้**
สภาพแวดล้อม	แดดจ้า, แดดร่ม, ฝนตก
อุณหภูมิ	ค่าต่อเนื่อง
ความชื้น	ค่าต่อเนื่อง
กระแสลม	ลมแรง, ลมปกติ

จากตารางที่ 2.1 ในการสร้างต้นไม้ตัดสินใจ ความสำเร็จไม่ได้อยู่ตรงที่สามารถสร้างต้นไม้ที่สามารถจัดกลุ่มข้อมูลจากตัวอย่างที่ใช้เรียนรู้ได้อย่างถูกต้องเท่านั้น แต่เราหวังว่ามันจะสามารถจัดกลุ่มข้อมูลจากตัวอย่างใหม่ ๆ แบบเดียวกันที่นอกเหนือจากข้อมูลที่ใช้สอนได้อย่างถูกต้องด้วย ดังนั้นการสร้างต้นไม้ตัดสินใจจึงควรจะมีข้อมูลทดสอบ (test data) ที่จะใช้ในการตรวจสอบความถูกต้องของต้นไม้ตัดสินใจที่ได้

การสร้างต้นไม้ตัดสินใจ

ในการสร้างต้นไม้ตัดสินใจจะใช้วิธีแบ่งแยกแล้วเอาชนะ (divide and conquer) โดยการเลือกคุณสมบัติขึ้นมา 1 คุณสมบัติจากคุณสมบัติทั้งหมด เพื่อเป็นรากของต้นไม้ จากนั้นจะแบ่งตัวอย่างออกเป็นกลุ่ม ๆ ตามค่าที่เป็นไปได้ของคุณสมบัติที่เลือกมาของแต่ละตัวอย่าง จากการแบ่งนี้จะทำให้เกิดเหตุการณ์ 3 อย่างคือ

1. กลุ่มตัวอย่างหลักจากแบ่งแล้ว จะประกอบด้วยตัวอย่างที่เป็นกลุ่มเดียวกัน ซึ่งจะกลายเป็นใบต่อไป
2. ไม่มีตัวอย่างตกอยู่ในกลุ่มนี้หลังจากแบ่งแล้ว ที่จุดนี้ก็จะกลายเป็นใบเช่นกัน แต่ถูกจัดอยู่ในกลุ่มใดนั้นต้องตัดสินใจโดยใช้ข้อมูลอื่น โดย C4.5 จะใช้ค่าของกลุ่มที่มีความถี่สูงที่สุดของโนดก่อนหน้าเป็นค่าของใบนี้
3. กลุ่มตัวอย่างที่ได้ประกอบด้วยตัวอย่างหลายกลุ่ม ซึ่งจะต้องทำการแบ่งต่อไป โดยแบ่งตัวอย่างออกเป็นกลุ่ม ๆ ตามค่าที่เป็นไปได้ในแต่ละกิ่งของโนดนี้ จากนั้นจึงเริ่มแบ่งตัวอย่างในแต่ละกิ่งโดยเลือกคุณสมบัติใหม่เพื่อแบ่งตัวอย่างต่อไป

จะเห็นได้ว่าในการสร้างต้นไม้ตัดสินใจ จะเป็นการแบ่งตัวอย่างออกเป็นกลุ่ม ๆ ตามคุณสมบัติจนกระทั่งได้กลุ่มตัวอย่างที่เป็นพวกเดียวกัน จากตัวอย่างการตัดสินใจเล่นกอล์ฟในตารางที่ 2.2 ซึ่งประกอบด้วยคุณสมบัติ 4 คุณสมบัติ (สภาพแวดล้อม อุณหภูมิ ความชื้น และกระแสลม) และกลุ่ม (การออกรอบ) 2 กลุ่ม (ออกรอบ และไม่ออกรอบ)

ตารางที่ 2.2 ตัวอย่างสอนของการตัดสินใจเล่นกอล์ฟ

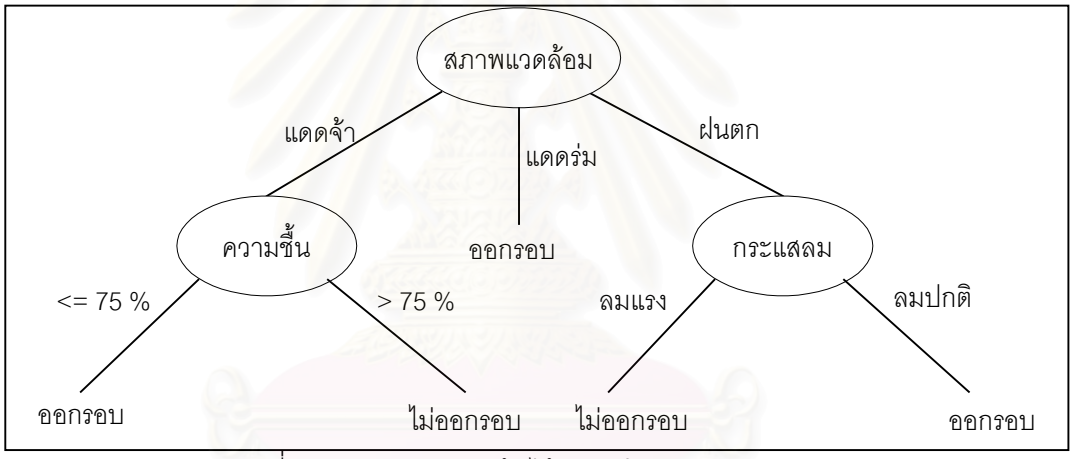
สภาพแวดล้อม	อุณหภูมิ (^{o}F)	ความชื้น (%)	กระแสลม	การออกรอบ
แดดจ้า	75	70	ลมแรง	ออกรอบ
แดดจ้า	80	90	ลมแรง	ไม่ออกรอบ
แดดจ้า	85	85	ลมปกติ	ไม่ออกรอบ
แดดจ้า	72	95	ลมปกติ	ไม่ออกรอบ
แดดจ้า	69	70	ลมปกติ	ออกรอบ
แดดร่ม	72	90	ลมแรง	ออกรอบ
แดดร่ม	83	78	ลมปกติ	ออกรอบ
แดดร่ม	64	65	ลมแรง	ออกรอบ
แดดร่ม	81	75	ลมปกติ	ออกรอบ
ฝนตก	71	80	ลมแรง	ไม่ออกรอบ
ฝนตก	65	70	ลมแรง	ไม่ออกรอบ
ฝนตก	75	80	ลมปกติ	ออกรอบ
ฝนตก	68	80	ลมปกติ	ออกรอบ
ฝนตก	70	96	ลมปกติ	ออกรอบ

เมื่อตัวอย่างไม่ได้ตกอยู่ในกลุ่มเดียวกัน วิธีการแบ่งแยกและจัดกลุ่มจะพยายามแบ่งตัวอย่างออกเป็นกลุ่มย่อย ตามค่าที่เป็นไปได้ของคุณสมบัตินั้นของแต่ละตัวอย่าง และในกลุ่มย่อยแต่ละกลุ่มก็จะมีการแบ่งตัวอย่างต่อไปจนกลุ่มย่อยที่ได้เป็นตัวอย่างเป็นกลุ่มเดียวกัน หรือไม่มีตัวอย่างให้แบ่งต่ออีก จากตัวอย่างสอนในตารางที่ 2.2 เมื่อแบ่งตัวอย่างจากคุณสมบัติสภาพแวดล้อม ในกลุ่มตัวอย่างย่อยที่มีค่าเป็น แดดจ้า และฝนตก จะประกอบด้วยตัวอย่างหลายกลุ่ม ส่วนกลุ่มย่อยที่มีค่าเป็นแดดร่มจะประกอบด้วยตัวอย่างกลุ่มเดียวคือกลุ่มออกรอบ ถ้าในกลุ่มของแดดจ้ามีการแบ่งต่อ โดยเลือกคุณสมบัติความชื้นที่ระดับความชื้นน้อยกว่าหรือเท่ากับ 75 เปอร์เซ็นต์ และความชื้นมากกว่า 75 เปอร์เซ็นต์เป็นระดับที่ใช้แบ่งตัวอย่างต่อ ส่วนในกลุ่มย่อยที่มีค่าของคุณสมบัติสภาพแวดล้อมเป็นฝนตก จะแบ่งต่อโดยใช้คุณสมบัติกระแสลมเป็นตัวแบ่ง โดยหลังจากแบ่งแล้ว กลุ่มย่อยแต่ละกลุ่มจะประกอบด้วยตัวอย่างที่เป็นกลุ่มเดียวกัน เมื่อโปรแกรม C4.5 ทำงานจะได้เป็นต้นไม้ตัดสินใจตามรูปที่ 2.1

```
สภาพแวดล้อม = แดดจ้า:
|   ความชื้น <= 75 %: ออกรอบ
|   ความชื้น > 75 %: ไม่ออกรอบ
สภาพแวดล้อม = แดดร่ม: ออกรอบ
สภาพแวดล้อม = ฝนตก:
|   กระแสลม = ลมแรง: ไม่ออกรอบ
|   กระแสลม = ลมปกติ: ออกรอบ
```

รูปที่ 2.1 ต้นไม้ตัดสินใจของการตัดสินใจเล่นกอล์ฟจากการทำงานของโปรแกรม C4.5

จากรูปที่ 2.1 โปรแกรม C4.5 จะสร้างต้นไม้ตัดสินใจออกมาอยู่ในรูปแบบข้อความ ซึ่งสามารถอ่านให้เข้าใจได้ง่ายขึ้นตามรูปที่ 2.2



รูปที่ 2.2 แผนภาพแสดงต้นไม้ตัดสินใจของการเล่นกอล์ฟ

ในการสร้างต้นไม้ตัดสินใจด้วยวิธีแบ่งแยกแล้วจัดกลุ่มจากตัวอย่างใด ๆ สามารถจะสร้างต้นไม้-ตัดสินใจขึ้นมาได้หลายต้นจากข้อมูลชุดเดียวกัน ขึ้นอยู่กับการเลือกคุณสมบัติที่ใช้แบ่งที่แตกต่างกันไป ยิ่งจำนวนคุณสมบัติที่ใช้แบ่งตัวอย่างและค่าที่เป็นไปได้ในแต่ละคุณสมบัติยิ่งมาก ก็จะทำให้ได้จำนวนต้นไม้ตัดสินใจที่เป็นไปได้มากขึ้น และต้นไม้ที่ได้ก็จะมีขนาดต่าง ๆ กัน บางต้นก็มีขนาดเล็ก บางต้นก็มีขนาดใหญ่ แต่ต้นไม้ที่เราต้องการจะเป็นต้นไม้ที่มีขนาดเล็ก เพราะจะใช้จำนวนครั้งในการแบ่งตัวอย่างน้อย และเป็นต้นไม้ที่เข้าใจง่าย ดังนั้นจึงเป็นการยากที่จะสร้างต้นไม้ทั้งหมดที่เป็นไปได้ก่อน แล้วเลือกต้นไม้ที่ต้องการออกมาเมื่อมีจำนวนคุณสมบัติ หรือค่าที่เป็นไปได้ในแต่ละคุณสมบัติมีจำนวนมาก

จะเห็นได้ว่าจุดสำคัญจะอยู่ที่การเลือกคุณสมบัติที่ใช้แบ่งตัวอย่าง เนื่องจากวิธีนี้จะเป็นการทำไปข้างหน้าไม่มีการย้อนกลับ กล่าวคือเมื่อเลือกคุณสมบัติหนึ่งคุณสมบัติใดขึ้นมาแบ่งตัวอย่างแล้ว จะไม่มีการถอยหรือกลับมาเลือกคุณสมบัติอื่นเพื่อแบ่งใหม่อีก ดังนั้นจึงต้องเลือกคุณสมบัติที่ดีที่สุดในการแบ่ง

ตัวอย่างของแต่ละกิ่ง คุณสมบัติที่ดีที่สุดควรเป็นคุณสมบัติที่เมื่อแบ่งตัวอย่างตามคุณสมบัตินี้แล้ว จะทำให้จำนวนครั้งของการแบ่งต่อหรือจำนวนโนดต่อจากกิ่งนี้น้อยที่สุด ซึ่งจะนำไปสู่ต้นไม้ที่เล็กและเข้าใจง่าย

ค่ามาตรฐานเกน (gain criterion)

วิธีการสร้างต้นไม้ตัดสินใจแบบ ID3 [5] จะใช้ค่ามาตรฐานเกน (gain criteria) ในการตัดสินใจเลือกคุณสมบัติที่จะใช้เป็นรากหรือกิ่งในต้นไม้ โดยการคำนวณค่าเกนของแต่ละคุณสมบัติเมื่อใช้แบ่งตัวอย่าง แล้วเลือกคุณสมบัติที่มีค่าเกนสูงที่สุดมาเป็นรากหรือโนด ค่าเกนนี้คำนวณได้โดยใช้ความรู้จากทฤษฎีสารสนเทศ (information theory) ซึ่งมีสารสำคัญคือ ค่าสารสนเทศของข้อมูลขึ้นอยู่กับค่าความน่าจะเป็นของข้อมูล ซึ่งสามารถวัดอยู่ในรูปของ บิต (bits) จากสูตร

ค่าสารสนเทศของข้อมูล = - $\log_2$(ความน่าจะเป็นของข้อมูล) บิต

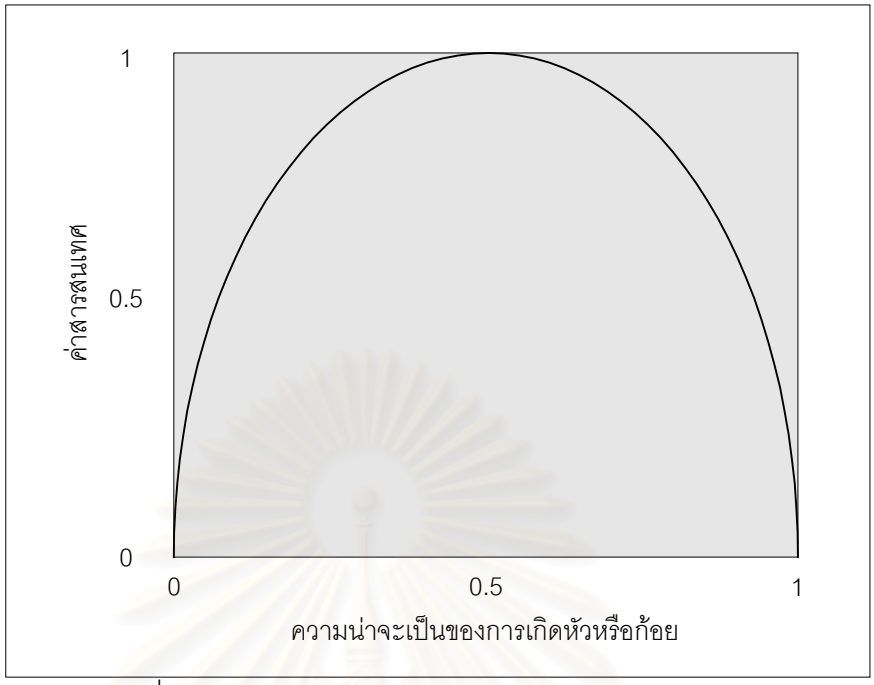
ถ้าให้ชุดของข้อมูล M ประกอบด้วยค่าที่เป็นไปได้ คือ $\{m_1, m_2, \ldots, m_n\}$ และให้ความน่าจะเป็นกับ $P(m_i)$ สำหรับแต่ละค่าที่ปรากฏอยู่ในชุดข้อมูล M ดังนั้นค่าสารสนเทศของ M หรือค่าเอนโทรพี (entropy) ของ M (เขียนแทนด้วย I(M)) จะคำนวณได้จากสูตร

$$I(M) = \sum_{i}^{n} - P(m_i) \times \log_2 (P(m_i)) \text{ บิต}$$

ตัวอย่างเช่น ในการโยนหัวโยนก้อย ชุดข้อมูล M จะประกอบด้วยค่าที่เป็นไปได้ {หัว, ก้อย} และถ้าให้ความน่าจะเป็นที่ออกหัวเท่ากับ P(หัว) และความน่าจะเป็นที่ออกก้อยเท่ากับ P(ก้อย) ดังนั้นค่าสารสนเทศของการโยนหัวโยนก้อย จะคำนวณได้จากสูตร

I(การโยนหัวโยนก้อย) = - P(หัว) x $\log_2$(P(หัว)) - P(ก้อย) x $\log_2$(P(ก้อย)) บิต

เมื่อความน่าจะเป็นของการเกิดหัวหรือก้อยมีค่าต่าง ๆ กันจะสามารถคำนวณค่าสารสนเทศของการโยนหัวโยนก้อยได้ต่าง ๆ กันดังรูปที่ 2.3 ซึ่งจะเห็นได้ว่าเมื่อออกหัวหมดหรือก้อยหมด ค่าสารสนเทศจะเป็น 0 และค่าสารสนเทศจะค่อย ๆ เพิ่มขึ้นจนสูงที่สุดเมื่อความน่าจะเป็นของการเกิดหัวเท่ากับความน่าจะเป็นของการเกิดก้อย แสดงให้เห็นว่าค่าสารสนเทศที่น้อยจะบ่งบอกว่าข้อมูลชุดนั้นมีความแตกต่างกันน้อยหรือเกือบจะเป็นพวกเดียวกัน แต่ถ้าค่าสารสนเทศสูงจะบ่งบอกว่าข้อมูลชุดนั้นมีความแตกต่างกันมาก หรือประกอบด้วยตัวอย่างหลายพวกที่มีจำนวนใกล้เคียงกัน



รูปที่ 2.3 กราฟแสดงค่าสารสนเทศของการโยนหัวโยนก้อย

ในการเลือกคุณสมบัติที่จะมาเป็นรากของโนดใด ๆ จะอาศัยค่ามาตรฐานเกน ซึ่งคำนวณจากค่าสารสนเทศทั้งหมดของชุดข้อมูลนั้นลบด้วยค่าสารสนเทศหลังจากเลือกคุณสมบัติใดคุณสมบัติหนึ่งเป็นราก ค่าสารสนเทศหลังจากแบ่งตามคุณสมบัติที่เลือกแล้วจะคำนวณได้จาก ค่าผลรวมของผลคูณระหว่างค่าสารสนเทศของแต่ละโนดกับอัตราส่วนของตัวอย่างในแต่ละกิ่งต่อตัวอย่างทั้งหมดที่โนดนั้น ๆ หรือความน่าจะเป็นของค่าที่เป็นไปได้ของแต่ละคุณสมบัติ

ถ้าให้ข้อมูลสอนคือ T และคุณสมบัติที่เลือกเป็นราก คือ X และมีค่าทั้งหมดที่เป็นไปได้ n ค่า รากหรือโนดปัจจุบันจะแบ่งตัวอย่าง T ออกเป็นกลุ่มย่อย ๆ $\{t_1, t_2, \ldots, t_n\}$ ตามค่าที่เป็นไปได้ของ X ดังนั้นจึงสามารถคำนวณค่าสารสนเทศหลังจากแบ่งตามคุณสมบัติ X ดังนี้

$$I_x(T) = \sum_{i=1}^{n} \frac{|t_i|}{|T|} \times I(t_i) \quad \text{บิต}$$

ค่ามาตรฐานเกนของคุณสมบัติ X สามารถคำนวณได้จากการลบค่าสารสนเทศทั้งหมดที่โนดนี้กับค่าสารสนเทศที่ได้หลังจากแบ่งด้วยคุณสมบัติ X ดังนี้

ค่ามาตรฐานเกน(X) = $I(T) - I_x(T)$ บิต

ในการตัดสินใจเลือกคุณสมบัติไหนเป็นรากหรือโนดนั้น จะใช้ค่ามาตรฐานเกนที่มีค่าสูงสุดเป็นตัวตัดสิน ถ้าค่ามาตรฐานเกนที่คำนวณจากการแบ่งตัวอย่างตามคุณสมบัติไหนที่มีค่าสูงที่สุดก็จะเลือกคุณสมบัตินั้นเป็นรากหรือโนด จากตัวอย่างการตัดสินใจเล่นกอล์ฟในตารางที่ 2.2 ประกอบด้วยข้อมูล 2 พวก คือ ตัวอย่างที่ตัดสินใจออกรอบ 9 ตัวอย่าง และตัดสินใจไม่ออกรอบ 5 ตัวอย่าง ดังนั้นค่าสารสนเทศทั้งหมดของข้อมูลชุดนี้จะคำนวณได้ดังนี้

$$I(T) = -9/14 \times \log_2(9/14) - 5/14 \times \log_2(5/14)$$
$$= 0.940 \text{ บิต}$$

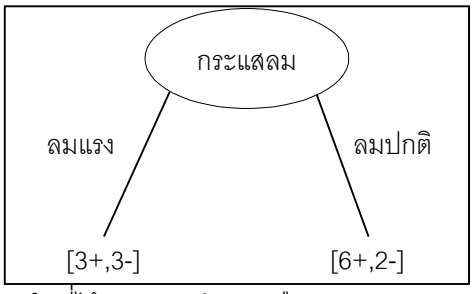
ถ้าแบ่งข้อมูลชุดนี้ตามคุณสมบัติสภาพแวดล้อม จะแบ่งตัวอย่างออกเป็น 3 กลุ่มย่อยดังรูปที่ 2.4 และสามารถคำนวณค่าสารสนเทศหลังจากแบ่งตัวอย่างตามคุณสมบัตินี้ คือ

รูปที่ 2.4 ต้นไม้ตัดสินใจที่ได้หลังจากทำการเลือกคุณสมบัติสภาพแวดล้อมเป็นราก

$$I_{\text{สภาพแวดล้อม}}(T) = 5/14 \times (-2/5 \times \log_2(2/5) - 3/5 \times \log_2(3/5))$$
$$+ 4/14 \times (-4/4 \times \log_2(4/4) - 0/4 \times \log_2(0/4))$$
$$+ 5/14 \times (-3/5 \times \log_2(3/5) - 2/5 \times \log_2(2/5))$$
$$= 0.694 \text{ บิต}$$

$$\text{ค่ามาตรฐานเกน(สภาพแวดล้อม)} = 0.940 - 0.694$$
$$= 0.246 \text{ บิต}$$

แต่ถ้าเราแบ่งข้อมูลชุดนี้ตามคุณสมบัติกระแสลม จะแบ่งตัวอย่างออกเป็น 2 กลุ่มย่อยดังรูปที่ 2.5 และสามารถคำนวณค่าสารสนเทศหลังจากแบ่งตามคุณสมบัตินี้ คือ



รูปที่ 2.5 ต้นไม้ตัดสินใจที่ได้หลังจากทำการเลือกคุณสมบัติกระแสลมเป็นราก

$$I_{\text{กระแสลม}}(T) = 6/14 \times (-3/6 \times \log_2(3/6) - 3/6 \times \log_2(3/6)) + 8/14 \times (-6/8 \times \log_2(6/8) - 2/8 \times \log_2(2/8))$$
$$= 0.892 \text{ บิต}$$

$$\text{ค่ามาตรฐานเกน(กระแสลม)} = 0.940 - 0.892$$
$$= 0.048 \text{ บิต}$$

จากรูปที่ 2.4 และ 2.5 เมื่อทำการเลือกรากแล้ว ค่าที่อยู่ในส่วนของใบจะเป็นจำนวนของตัวอย่างที่ครอบคลุมอยู่ในใบนั้น ๆ ในที่นี้ให้สัญลักษณ์เป็นเครื่องหมายบวก (+) และ ลบ (-) โดยเครื่องหมายบวกแสดงถึงตัวอย่างที่เป็นกลุ่มออกรอบ และเครื่องหมายลบแสดงถึงตัวอย่างที่เป็นกลุ่มไม่ออกรอบ

จะเห็นได้ว่าค่ามาตรฐานเกนของสภาพแวดล้อม จะมากกว่าค่ามาตรฐานเกนของกระแสลม ดังนั้นเราจึงเลือกที่จะแบ่งตัวอย่างตามคุณสมบัติของสภาพแวดล้อม ส่วนในคุณสมบัติอุณหภูมิและความชื้นซึ่งเป็นข้อมูลแบบค่าต่อเนื่องนั้น จะมีวิธีคำนวณค่ามาตรฐานเกนที่แตกต่างจากการคำนวณค่ามาตรฐานเกนของคุณสมบัติที่เป็นค่าไม่ต่อเนื่อง ซึ่งจะแสดงวิธีการคำนวณให้เห็นในหัวข้อการคำนวณสำหรับคุณสมบัติที่เป็นข้อมูลแบบต่อเนื่อง

ค่ามาตรฐานอัตราส่วนเกน (gain ratio criterion)

ใน ID3 จะใช้ค่ามาตรฐานเกนเป็นหลักในการเลือกคุณสมบัติที่จะใช้เป็นรากหรือโนด แต่ใน C4.5 ได้เพิ่มการใช้ค่ามาตรฐานอัตราส่วนเกน (gain ratio criterion) ในการตัดสินใจเลือกคุณสมบัติที่จะใช้เป็นรากหรือโนดอีกอย่างหนึ่ง เนื่องจากค่ามาตรฐานเกนจะมีอคติ (bias) อย่างมากกับข้อมูลที่ประกอบด้วยคุณสมบัติที่มีค่าที่เป็นไปได้จำนวนมาก ๆ เช่นข้อมูลที่ประกอบด้วยคุณสมบัติหมายเลขประจำตัว ซึ่งปกติจะไม่ซ้ำกันในแต่ละตัวอย่าง ถ้าแบ่งข้อมูลตามคุณสมบัตินี้จะทำให้ได้จำนวนตัวอย่างเพียง 1 ตัวอย่างต่อ 1 กิ่งของต้นไม้ และชุดตัวอย่างย่อยที่ได้จะประกอบด้วยข้อมูลกลุ่มเดียว เมื่อคำนวณค่าสารสนเทศจากการแบ่งตัวอย่างบนคุณสมบัตินี้ จะได้เท่ากับ 0 เนื่องจากค่า $\log_2(1) = 0$ ทำให้ค่าเกนที่ได้ในคุณสมบัตินี้จะสูงที่สุดเสมอ

การแก้ไขความอคติของค่ามาตรฐานเกนสามารถทำได้โดยการปรับค่ามาตรฐานเกนให้ถูกต้องโดยใช้ค่าสารสนเทศของการแบ่งแยก (split information) ของแต่ละคุณสมบัติ ซึ่งถ้าให้ T คือชุดของตัวอย่าง เมื่อแบ่งตัวอย่างนี้ตามคุณสมบัติ X จะได้ชุดของตัวอย่างย่อยในแต่ละกิ่ง คือ $\{t_1, t_2, \ldots, t_n\}$ จำนวน n ชุด ตามค่าที่เป็นไปได้ในคุณสมบัติ X เมื่อคำนวณค่าสารสนเทศของการแบ่งแยกได้ ดังนี้

$$\text{ค่าสารสนเทศของการแบ่งแยก} = -\sum_{i=1}^{n} \frac{|t_i|}{|T|} \times \log_2 \left(\frac{|t_i|}{|T|} \right)$$

ค่าสารสนเทศของการแบ่งแยกนี้จะแสดงถึงระดับการกระจายของข้อมูล เมื่อแบ่งข้อมูลตัวอย่าง T เป็น n ชุดย่อยตามคุณสมบัติ X โดยค่านี้จะสูงสุดเมื่อ $|t_i|$ เป็น 1 เท่ากันในทุกกิ่ง และลดลงเมื่อค่า $|t_i|$ เพิ่มขึ้นเมื่อนำค่านี้ไปหารค่ามาตรฐานเกนจะได้ค่ามาตรฐานอัตราส่วนเกน ซึ่งช่วยแก้ไขความอคติของค่ามาตรฐานเกนได้ โดยทำให้ค่ามาตรฐานอัตราส่วนเกนในการแบ่งด้วยคุณสมบัติที่มีการกระจายสูงถูกปรับลดลง ดังนั้นค่ามาตรฐานอัตราส่วนเกนในคุณสมบัติของตัวอย่างที่มีการกระจายตัวของข้อมูลสูงดังที่กล่าวมาแล้วจึงไม่มีค่าสูงที่สุดเสมอ

ค่ามาตรฐานอัตราส่วนเกน = ค่ามาตรฐานเกน / ค่าสารสนเทศของการแบ่งแยก

จากตัวอย่างการตัดสินใจเล่นกอล์ฟในตารางที่ 2.2 เมื่อแบ่งตัวอย่างตามคุณสมบัติสภาพแวดล้อม จะคำนวณค่ามาตรฐานอัตราส่วนเกน ได้ดังนี้

ค่าสารสนเทศของการแบ่งแยก (สภาพแวดล้อม) = $-5/14 \times \log_2(5/14) - 4/14 \times \log_2(4/14)$

$-5/14 \times \log_2(5/14)$

= 1.577 บิต

ค่ามาตรฐานอัตราส่วนเกน(สภาพแวดล้อม) = 0.246/1.577

= 0.156 บิต

และเมื่อแบ่งข้อมูลตัวอย่างตามคุณสมบัติกระแสลม จะคำนวณค่ามาตรฐานอัตราส่วนเกนได้ดังนี้

ค่าสารสนเทศของการแบ่งแยก (กระแสลม) = $-6/14 \times \log_2(6/14) - 8/14 \times \log_2(8/14)$

= 0.985 บิต

ค่ามาตรฐานอัตราส่วนเกน(กระแสลม) = 0.048/0.985

= 0.049 บิต

จากการคำนวณค่ามาตรฐานอัตราส่วนเกนเทียบกันระหว่างการเลือกคุณสมบัติสภาพแวดล้อมกับคุณสมบัติกระแสลม คุณสมบัติสภาพแวดล้อมจะให้ค่าที่มากกว่ากระแสลม แม้ว่าค่าสารสนเทศของการแบ่งแยกของคุณสมบัติกระแสลมจะมีค่ามากกว่าสภาพแวดล้อม แต่ค่ามาตรฐานเกนซึ่งเป็นตัวบอกความเป็นระเบียบของข้อมูลของคุณสมบัติสภาพแวดล้อมมีความเป็นระเบียบมากกว่า

<u>การคำนวณสำหรับคุณสมบัติที่เป็นข้อมูลต่อเนื่อง</u>

ในการคำนวณค่ามาตรฐานเกนหรือค่ามาตรฐานอัตราส่วนเกนสำหรับคุณสมบัติที่ข้อมูลเป็นค่าต่อเนื่อง หรือข้อมูลตัวเลข จะกระทำได้โดยการคำนวณค่ามาตรฐานอัตราส่วนเกนหลังจากแบ่งตัวอย่างตามจุดแบ่งที่เป็นไปได้ในระดับต่าง ๆ ของคุณสมบัติที่เป็นค่าต่อเนื่อง แล้วเลือกจุดแบ่งที่มีค่ามาตรฐานอัตราส่วนเกนสูงที่สุด เป็นระดับที่จะใช้แบ่งตัวอย่าง และใช้ค่ามาตรฐานอัตราส่วนเกนที่สูงที่สุดนี้เป็นตัวแทนในการพิจารณาเลือกคุณสมบัติที่จะใช้แบ่งตัวอย่าง

สมมติว่า ตัวอย่างสอน T ประกอบด้วยคุณสมบัติต่อเนื่อง A เมื่อเรียงข้อมูลตามคุณสมบัติ A จะได้ชุดของค่าที่ไม่ซ้ำกัน m ค่า ตามลำดับ $\{v_1, v_2, \ldots, v_m\}$ จุดที่เป็นระดับที่ใช้แบ่งข้อมูลจะอยู่ระหว่างค่าของ v_i กับ v_{i+1} ดังนั้นจึงมีจุดที่ใช้แบ่งข้อมูลจำนวน m-1 จุดที่เป็นไปได้ ซึ่งโดยปกติจุดที่ใช้แบ่งข้อมูลจะใช้ค่า $(v_i + v_{i+1})/2$ แต่ C4.5 จะใช้ค่าจากตัวอย่างที่สูงที่สุดที่ไม่เกินจุดกึ่งกลางจากการคำนวณในแต่ละช่วงแทนที่จะใช้จุดกึ่งกลางเป็นตัวแบ่ง เพื่อรับประกันว่าค่าที่ปรากฏในต้นไม้ตัดสินใจจะปรากฏอยู่ในตัวอย่างด้วย

จากตัวอย่างการตัดสินใจเล่นกอล์ฟในตารางที่ 2.2 ถ้าเราแบ่งข้อมูลตามคุณสมบัติอุณหภูมิโดยใช้อุณหภูมิระหว่าง 70 ถึง 71 องศาฟาเรนไฮท์ เป็นจุดที่ใช้แบ่ง สามารถจะคำนวณค่ามาตรฐานเกนและค่ามาตรฐานอัตราส่วนเกนได้ดังนี้

$$I_{\text{อุณหภูมิ (ระหว่าง 70 และ 71)}}(T) = 5/14 \times (-4/5 \times \log_2(4/5) - 1/5 \times \log_2(1/5)) + 9/14 \times (-5/9 \times \log_2(5/9) - 4/9 \times \log_2(4/9))$$

= 0.895 บิต

ค่ามาตรฐานเกน(อุณหภูมิ) = 0.940 – 0.895

= 0.045 บิต

ค่าสารสนเทศของการแบ่งแยก (อุณหภูมิ) $= -5/14 \times \log_2(5/14) - 9/14 \times \log_2(9/14)$

= 0.940 บิต

ค่ามาตรฐานอัตราส่วนเกน(อุณหภูมิ) = 0.045/0.940

= 0.0479 บิต

เมื่อคำนวณค่ามาตรฐานเกนและค่ามาตรฐานอัตราส่วนเกนที่แบ่งข้อมูล ณ จุดต่าง ๆ บนคุณสมบัติอุณหภูมิและความชื้นจะได้ค่าดังตารางที่ 2.3 และ 2.4

ตารางที่ 2.3 ค่ามาตรฐานเกนและค่ามาตรฐานอัตราส่วนเกนเมื่อแบ่งตามคุณสมบัติอุณหภูมิ

ระดับอุณหภูมิที่ใช้แบ่ง	ค่ามาตรฐานเกน	ค่ามาตรฐานอัตราส่วนเกน
65	0.010	0.017
68	0.000	0.001
69	0.015	0.017
70	0.045	0.048
71	0.001	0.001
72	0.001	0.001
75	0.025	0.029
80	0.000	0.001
81	0.010	0.017

ตารางที่ 2.4 ค่ามาตรฐานเกนและค่ามาตรฐานอัตราส่วนเกนเมื่อแบ่งตามคุณสมบัติความชื้น

ระดับความชื้นที่ใช้แบ่ง	ค่ามาตรฐานเกน	ค่ามาตรฐานอัตราส่วนเกน
70	0.015	0.017
75	0.045	0.001
78	0.090	0.017
80	0.102	0.048
85	0.025	0.001
90	0.010	0.001

จะเห็นได้ว่าค่ามาตรฐานเกนหรือค่ามาตรฐานอัตราส่วนเกนที่สูงสุดของคุณสมบัติอุณหภูมิหรือความชื้นก็ยังมีค่าน้อยกว่าค่าที่คำนวณได้จากคุณสมบัติของสภาพแวดล้อม ดังนั้นคุณสมบัตินี้จึงไม่ถูกเลือก แต่ถ้าคุณสมบัตินี้ถูกเลือก จุดที่มีค่ามาตรฐานเกนหรือค่ามาตรฐานอัตราส่วนเกนสูงที่สุดจะถูกใช้เป็นจุดแบ่งข้อมูล

<u>การจัดการกับตัวอย่างที่ไม่ทราบค่า</u>

สำหรับข้อมูลในความเป็นจริงนั้น เราอาจจะพบว่าคุณสมบัติบางอย่างของตัวอย่างอาจจะประกอบด้วยข้อมูลที่ไม่ทราบค่า เมื่อจะนำข้อมูลนี้มาเป็นข้อมูลสอนก็จะมีทางเลือก 2 ทาง คือ ทางแรกเป็นการนำเฉพาะตัวอย่างที่ทราบค่าเท่านั้นมาใช้เป็นข้อมูลสอน ซึ่งอาจทำให้ตัวอย่างที่ใช้สอนน้อยลงและอาจสูญเสียความรู้บางอย่างที่จะได้จากข้อมูลชุดนี้ก็ได้ ส่วนอีกทางหนึ่งจะเป็นการรวมเอาตัวอย่างที่มีคุณสมบัติบางอย่างที่ไม่ทราบค่าเข้าไปด้วย แต่จะมีวิธีจัดการเพื่อให้ได้ประโยชน์จากตัวอย่างเหล่านี้ ซึ่งมีอยู่ด้วยกันหลายวิธี เช่น การเติมค่าที่สูญหายด้วยค่าที่มีความถี่สูงที่สุด แต่ C4.5 จะใช้วิธีคำนวณค่ามาตรฐานเกนหรือค่ามาตรฐานอัตราส่วนเกนจากชุดข้อมูลที่ทราบค่าของคุณสมบัตินั้น แล้วปรับลดค่าให้ถูกต้องด้วยความน่าจะเป็นของตัวอย่างที่ทราบค่า ต่อตัวอย่างทั้งหมด

ถ้าให้ T เป็นชุดข้อมูลสอน และ X เป็นคุณสมบัติที่ใช้ทดสอบบนตัวอย่าง A เราสามารถจะคำนวณค่าสารสนเทศทั้งหมดของชุดตัวอย่าง T และค่าสารสนเทศหลังจากแบ่งตัวอย่างบนคุณสมบัติ X เพื่อคำนวณหาค่ามาตรฐานเกนได้จากตัวอย่าง A ที่ทราบค่าเท่านั้น ซึ่งถ้าให้ค่ามาตรฐานเกนของตัวอย่างที่ไม่ทราบค่าเป็น 0 จะได้ว่า

ค่ามาตรฐานเกน (X) = ความน่าจะเป็นที่ A ทราบค่า x ค่ามาตรฐานเกนของตัวอย่างที่ทราบค่า
+ ความน่าจะเป็นที่ A ไม่ทราบค่า x 0
= ความน่าจะเป็นที่ A ทราบค่า x ค่ามาตรฐานเกนของตัวอย่างที่ทราบค่า

ส่วนค่าสารสนเทศของการแบ่งแยก X สามารถปรับได้โดยการเพิ่มชุดตัวอย่างอีก 1 กลุ่ม สมมติว่าคุณสมบัติ X มีค่าที่เป็นไปได้ n ค่า เวลาคำนวณค่าสารสนเทศของการแบ่งกลุ่ม จะมีการแบ่งข้อมูลเป็น n+1 กลุ่มย่อย โดยกลุ่มที่เพิ่มมาจะเป็นกลุ่มของตัวอย่างที่ไม่ทราบค่า

เมื่อมีการแบ่งตัวอย่างตามคุณสมบัติ X เป็นกลุ่มย่อย $t_1, t_2, \ldots, t_n$ ชุด ตามค่าที่เป็นไปได้ $o_1, o_2, \ldots, o_n$ ค่า ตัวอย่างจาก T ซึ่งทราบค่า o_i จะถูกแบ่งอยู่ในชุดย่อย t_i โดยมีความน่าจะเป็นที่ตัวอย่างนี้จะถูกแบ่งอยู่ในกลุ่ม t_i เป็น 1 และความน่าจะเป็นที่ตัวอย่างนี้จะถูกแบ่งอยู่ในกลุ่มอื่นเป็น 0 แต่เมื่อตัวอย่าง T ไม่ทราบค่าจึงเป็นไปได้ว่าตัวอย่างนี้อาจจะมีค่าเป็นค่าใดค่าหนึ่งใน o_i ดังนั้นถ้าให้ w เป็นความน่าจะเป็นที่ตัวอย่างนี้จะถูกแบ่งอยู่ในแต่ละชุดย่อย เมื่อตัวอย่างนี้ทราบค่า ค่าของ w จะมีค่าเป็น 1 และเมื่อตัวอย่างนี้ไม่ทราบค่า ค่าของ w จะมีค่าเป็นความน่าจะเป็นที่จะเกิด o_i ตอนนี้แต่ละชุดย่อย t_i เมื่อต้องการค่า $|t_i|$ จะคำนวณได้จากผลรวมของค่า w ในแต่ละชุดย่อยแทนที่จะเป็นผลรวมของจำนวนตัวอย่างในชุดย่อย t_i

จากตัวอย่างการเล่นกอล์ฟในตารางที่ 2.2 สมมติว่าในตัวอย่างคุณสมบัติสภาพแวดล้อมที่เท่ากับแดดร่ม อุณหภูมิเท่ากับ 72 องศา ความชื้นเท่ากับ 90 % และกระแสลมแรง เราไม่ทราบค่าของสภาพแวดล้อมของตัวอย่างนี้ เมื่อเราสนใจตัวอย่างที่เหลืออีก 13 ตัวอย่าง ซึ่งทราบค่าของคุณสมบัติสภาพแวดล้อมจะสามารถนับจำนวนตัวอย่างในแต่ละกลุ่มได้ดังตารางที่ 2.5

ตารางที่ 2.5 จำนวนตัวอย่างเมื่อแบ่งตามคุณสมบัติสภาพแวดล้อม

สภาพแวดล้อม	ออกรอบ	ไม่ออกรอบ	รวม
แดดจ้า	2	3	5
แดดร่ม	3	0	3
ฝนตก	3	2	5
รวม	8	5	13

เมื่อคำนวณค่าต่าง ๆ ใหม่บนคุณสมบัติสภาพแวดล้อมจะได้ดังนี้

$I(T) = -8/13 \times \log_2(8/13) - 5/13 \times \log_2(5/13)$

$= 0.961$ บิต

$I_{\text{สภาพแวดล้อม}}(T) = 5/13 \times (-2/5 \times \log_2(2/5) - 3/5 \times \log_2(3/5))$
$+ 3/13 \times (-3/3 \times \log_2(3/3) - 0/3 \times \log_2(0/3))$
$+ 5/13 \times (-3/5 \times \log_2(3/5) - 2/5 \times \log_2(2/5))$
$= 0.747$ บิต

ค่ามาตรฐานเกน(สภาพแวดล้อม) $= 13/14 \times (0.961 - 0.747)$
$= 0.199$ บิต

ค่าสารสนเทศของการแบ่งแยก (สภาพแวดล้อม) $= -5/14 \times \log_2(5/14) - 3/14 \times \log_2(3/14)$
$-5/14 \times \log_2(5/14) - 1/14 \times \log_2(1/14)$
$= 1.809$ บิต

ค่ามาตรฐานอัตราส่วนเกน(สภาพแวดล้อม) $= 0.199/1.809$
$= 0.110$ บิต

เมื่อตัวอย่างทั้ง 14 ตัวอย่างถูกแบ่งออกเป็นชุดย่อยตามคุณสมบัติสภาพแวดล้อม ในตัวอย่าง 13 ตัวอย่าง ที่ทราบค่าสภาพแวดล้อมจะถูกแบ่งตามปกติ แต่ตัวอย่างที่เหลือ 1 ตัวอย่างซึ่งไม่ทราบค่าของสภาพแวดล้อมจะถูกแบ่งให้กับทุก ๆ ค่าที่เป็นไปได้ของคุณสมบัติสภาพแวดล้อม คือ แดดจ้า แดดร่ม และ ฝนตก เป็นอัตราส่วน 5/13, 3/13 และ 5/13 ตามลำดับ

ถ้าดูชุดตัวอย่างย่อยหลังจากแบ่งด้วยคุณสมบัติสภาพแวดล้อม เฉพาะในสภาพแวดล้อมที่เป็นแดดจ้าจะประกอบด้วยตัวอย่างดังตารางที่ 2.6

ตารางที่ 2.6 ชุดตัวอย่างหลังจากแบ่งด้วยคุณสมบัติสภาพแวดล้อมเฉพาะที่เป็นแดดจ้า

สภาพแวดล้อม	อุณหภูมิ	ความชื้น	กระแสลม	การตัดสินใจ	น้ำหนัก (w)
แดดจ้า	75	70	ลมแรง	ออกรอบ	1
แดดจ้า	80	90	ลมแรง	ไม่ออกรอบ	1
แดดจ้า	85	85	ลมปกติ	ไม่ออกรอบ	1
แดดจ้า	72	95	ลมปกติ	ไม่ออกรอบ	1
แดดจ้า	69	70	ลมปกติ	ออกรอบ	1
?	72	90	ลมแรง	ออกรอบ	5/13

ถ้าตัวอย่างย่อยชุดนี้ถูกแบ่งต่อบนคุณสมบัติของความชื้นที่ 75 % เหมือนกับการแบ่งเมื่อทราบค่าคุณสมบัติสภาพแวดล้อมของตัวอย่างทั้งหมด จะสามารถแบ่งตัวอย่างออกเป็น 2 ชุดคือ

- ความชื้น $\leq$ 75% ประกอบด้วยตัวอย่างที่ตัดสินใจ ออกรอบ 2 ตัวอย่างและ ไม่ออกรอบ 0 ตัวอย่าง
- ความชื้น $>$ 75% ประกอบด้วยตัวอย่างที่ตัดสินใจ ออกรอบ 5/13 ตัวอย่างและ ไม่ออกรอบ 3 ตัวอย่าง

และเมื่อคำนวณค่ามาตรฐานเกนหรือค่ามาตรฐานอัตราส่วนเกน แล้วสร้างเป็นต้นไม้ตัดสินใจจะได้ต้นไม้ที่มีลักษณะเหมือนเดิมดังนี้

```
สภาพแวดล้อม = แดดจ้า:
|   ความชื้น <= 75 %: ออกรอบ (2.0)
|   ความชื้น > 75 %: ไม่ออกรอบ (3.4/0.4)
สภาพแวดล้อม = แดดร่ม: ออกรอบ (3.2)
สภาพแวดล้อม = ฝนตก:
|   กระแสลม = ลมแรง: ไม่ออกรอบ (2.4/0.4)
|   กระแสลม = ลมปกติ: ออกรอบ (3.0)
```

รูปที่ 2.6 ต้นไม้ตัดสินใจในการเล่นกอล์ฟเมื่อมีตัวอย่างไม่ทราบค่า

ที่ปลายของใบจะพบตัวเลขที่อยู่ในรูปของ (N) หรือ (N/E) ซึ่ง N ในที่นี้จะกลายเป็นผลรวมของน้ำหนัก w ของตัวอย่างที่ถูกแบ่งอยู่ที่ใบนี้ แทนที่จะเป็นจำนวนตัวอย่าง และ E จะเป็นจำนวนตัวอย่างหรือน้ำหนัก w ของตัวอย่างที่ไม่ใช่พวกเดียวกันกับตัวอย่างที่ใบนี้ ดังนั้นที่ใบซึ่งมีเงื่อนไขว่า ความชื้น $>$ 75: ไม่ออกรอบ (3.4 / 0.4) จะหมายถึงว่าถ้ามีจำนวนตัวอย่าง 3.4 ตัวอย่างที่ถูกแบ่งตกอยู่ที่ใบนี้ จะมีตัวอย่าง 0.4 ตัวอย่างที่ไม่ได้เป็นพวกเดียวกับตัวอย่างที่ใบนี้คือ ถ้าที่ใบนี้ถูกจัดเป็นพวกออกรอบ จะมี 0.4 ตัวอย่างที่เป็นพวกไม่ออกรอบ

ถ้าเราใช้ต้นไม้นี้ในการตัดสินใจเล่นกอล์ฟ เมื่อสภาพแวดล้อมเท่ากับแดดจ้า อุณหภูมิ 70 องศาฟาเรนไฮต์ และกระแสลมปกติ แต่เราไม่ทราบค่าของความชื้น เราสามารถจะวิ่งไปตามทางเดินของต้นไม้ โดยเริ่มจากคุณสมบัติของสภาพแวดล้อม เมื่อทราบว่าสภาพแวดล้อมเป็นแดดจ้า ก็จะวิ่งตามลงไปสู่โนดแรกของต้นไม้ แต่ที่โนดนี้เราไม่ทราบค่าของความชื้น ดังนั้นจะเป็นไปได้ว่าความชื้นอาจจะมากกว่าหรือน้อยกว่า 75 % ก็ได้ จึงเกิดเหตุการณ์ 2 อย่างคือ

1. ถ้าความชื้นน้อยกว่าหรือเท่ากับ 75 % ตัวอย่างนี้จะถูกจัดอยู่ในพวกออกรอบ
2. ถ้าความชื้นมากกว่า 75 % ตัวอย่างนี้จะถูกจัดอยู่ในพวกไม่ออกรอบ ด้วยความน่าจะเป็น 3/3.4 หรือคิดเป็น 88 % และถูกจัดอยู่ในพวกออกรอบ ด้วยความน่าจะเป็น 0.4/3.4 หรือคิดเป็น 12 %

แต่ตอนนี้ ต้นไม้ถูกสร้างขึ้น สำหรับคุณสมบัติสภาพแวดล้อมที่เป็นแดดจ้า จะประกอบด้วยตัวอย่างที่เป็น พวกออกรอบ 2 ตัวอย่าง และพวกไม่ออกรอบ 3.4 ตัวอย่าง ดังนั้นค่าน้ำหนัก w จะเป็น 2/5.4 และ 3.4/5.4 ตามลำดับ เมื่อสรุปเป็นการตัดสินใจจะเป็นดังนี้

ตัดสินใจออกรอบ = (2/5.4 x 100) + (3.4/5.4 x 12)
= 44 %

ตัดสินใจไม่ออกรอบ = (3.4/5.4 x 88)
= 56 %

2.2.2 การตัดเล็มต้นไม้ตัดสินใจโดยใช้ค่าความผิดพลาด

ในการสร้างต้นไม้ตัดสินใจที่ใช้วิธีแบ่งแยกและจัดกลุ่มดังที่กล่าวมาแล้ว จะแบ่งตัวอย่างจนกระทั่งได้ตัวอย่างเป็นพวกเดียวกันหรือไม่มีตัวอย่างให้แบ่งอีกแล้ว ผลที่ได้คือต้นไม้ที่มีความซับซ้อนมาก เนื่องจากมีจำนวนโนด กิ่ง และทางเดินมาก ทำให้เข้าใจยากและการตัดสินใจจะเฉพาะเจาะจงกับข้อมูลตัวอย่างที่ใช้สอนมากเกินไป (overfit the data) ทำให้การตัดสินใจในตัวอย่างใหม่ ๆ ที่นอกเหนือจากตัวอย่างที่ใช้สอนมีความผิดพลาดสูง การตัดเล็มจะทำให้ต้นไม้ตัดสินใจที่ได้ใหม่ไม่เป็นต้นไม้ที่ตัดสินใจได้ถูกต้องเฉพาะกับข้อมูลที่ใช้สอนเท่านั้น แต่ต้นไม้นี้จะเป็นต้นไม้แบบง่าย (simplified tree) ที่สามารถใช้ตัดสินใจในตัวอย่างที่ไม่เคยเห็นได้ถูกต้องพอ ๆ กับตัวอย่างที่ใช้สอนด้วย

C4.5 จะใช้วิธีการตัดเล็มโดยใช้ค่าความผิดพลาด (Error-Based Pruning) คือจะเป็นการรวมกิ่งเป็นใบหรือโนดเป็นใบ โดยที่เมื่อรวมแล้วไม่ทำให้ค่าความผิดพลาดหลังจากรวมแล้วเพิ่มขึ้น ถ้ามีตัวอย่าง N ตัวอย่างที่ใบ และมีตัวอย่าง E ตัวอย่างเป็นตัวอย่างที่ไม่ถูกต้องหรือไม่ใช่พวกเดียวกันกับตัวอย่างที่ใบนี้ ดังนั้นค่าความผิดพลาดที่ใบนี้จะเท่ากับ E/N ซึ่งเป็นค่าความผิดพลาดที่เฉพาะกับตัวอย่างสอนชุดนี้เท่านั้น แต่เราต้องการค่าความผิดพลาดที่เป็นค่าประมาณจากประชากร เพื่อใช้แทนค่าความผิดพลาดที่คาดว่าจะเกิดเมื่อใช้ทดสอบบนข้อมูลที่ไม่เคยเห็น จึงได้ใช้ค่าจำกัดบนของการกระจาย

แบบไบโนเมียล (binomial distribution) ที่ระดับความเป็นอิสระเท่ากับ CF (Confidence Level) เป็นตัวแทนความผิดพลาดของประชากร โดยเขียนอยู่ในรูป $U_{cf}(E,N)$

การประมาณค่าความผิดพลาดสำหรับใบและกิ่งเมื่อใช้กับข้อมูลที่ไม่เคยเห็น จะอยู่บนข้อกำหนดที่ว่าขนาดของตัวอย่างสอนเท่ากับขนาดตัวอย่างของข้อมูลที่ไม่เคยเห็น ดังนั้นถ้าใบประกอบด้วยตัวอย่าง N ตัวอย่าง ค่าความผิดพลาดที่คาดไว้ของตัวอย่างแต่ละตัวอย่างจะเท่ากับ $U_{cf}(E,N)$ ซึ่งสามารถคาดได้ว่าจะมีจำนวนตัวอย่างที่แบ่งผิดพลาดเท่ากับ $N \times U_{cf}(E,N)$ ตัวอย่าง เมื่อทดสอบบนตัวอย่างที่ไม่เคยเห็น และในขณะเดียวกัน จำนวนตัวอย่างที่คาดว่าจะแบ่งผิดพลาดบนโนดใด ๆ จะเท่ากับผลรวมของจำนวนตัวอย่างที่คาดว่าจะแบ่งผิดพลาดในแต่ละกิ่งรวมกัน

```
physician fee freeze = n:
|  adoption of the budget resolution = y: democrat (151.0)
|  adoption of the budget resolution = u: democrat (1.0)
|  adoption of the budget resolution = n:
|  |  education spending = n: democrat (6.0)
|  |  education spending = y: democrat (9.0)
|  |  education spending = u: republican (1.0)
physician fee freeze = y:
|  synfuels corporation cutback = n: republican (97.0/3.0)
|  synfuels corporation cutback = u: republican (4.0)
|  synfuels corporation cutback = y:
|  |  duty free exports = y: democrat (2.0)
|  |  duty free exports = u: republican (1.0)
|  |  duty free exports = n:
|  |  |  education spending = n: democrat (5.0/2.0)
|  |  |  education spending = y: republican (13.0/2.0)
|  |  |  education spending = u: democrat (1.0)
physician fee freeze = u:
|  water project cost sharing = n: democrat (0.0)
|  water project cost sharing = y: democrat (4.0)
|  water project cost sharing = u:
|  |  mx missile = n: republican (0.0)
|  |  mx missile = y: democrat (3.0/1.0)
|  |  mx missile = u: republican (2.0)
```

รูปที่ 2.7 ต้นไม้ตัดสินใจก่อนการตัดเล็ม

จากตัวอย่างของต้นไม้ตัดสินใจก่อนการตัดเล็มกิ่งในรูปที่ 2.7 ในโนดหนึ่งของต้นไม้ที่ประกอบด้วยกิ่งและใบดังนี้

education spending = n: democrat (6.0)
education spending = y: democrat (9.0)
education spending = u: republican (1.0)

จะเห็นได้ว่าไม่มีตัวอย่างที่แบ่งกลุ่มผิดพลาดบนข้อมูลสอนหลังจากการเรียนรู้และสร้างเป็นต้นไม้ตัดสินใจ สำหรับใบแรกที่มีจำนวนตัวอย่างเท่ากับ 6 ตัวอย่าง หรือ N=6 ถูกจัดอยู่ในกลุ่ม democrat โดยที่ไม่มีตัวอย่างผิดพลาด หรือ E=0 เมื่อคำนวณค่าความผิดพลาด $U_{25\%}(0,6)$ ได้เท่ากับ 0.206 (ในที่นี้ C4.5 จะใช้ค่าความเป็นอิสระ CF=25 % เป็นค่าโดยปริยาย ดังนั้นจำนวนตัวอย่างที่คาดว่าจะตอบผิดที่ใบนี้เมื่อใช้ทำนายตัวอย่างที่ไม่เคยเห็น 6 ตัวอย่าง จะเท่ากับ 6 x 0.206 สำหรับใบที่เหลือจะมีค่าความผิดพลาดเป็น $U_{25\%}(0,9)$ หรือเท่ากับ 0.143 และ $U_{25\%}(0,1)$ หรือเท่ากับ 0.750 ตามลำดับ เมื่อคำนวณตัวอย่างที่คาดว่าจะทำนายผิดที่โนดนี้จะเท่ากับ

จำนวนตัวอย่างที่คาดว่าจะทำนายผิดพลาด = 6 x 0.206 + 9 x 0.143 + 1 x 0.750
= 3.273 ตัวอย่าง

ถ้าโนดนี้ถูกแทนที่ด้วยใบที่มีค่าเป็น democrat จะทำให้ใบนี้ประกอบด้วยตัวอย่าง 16 ตัวอย่าง และมีจำนวนตัวอย่างที่อยู่ต่างพวกหรือไม่ใช่ democrat 1 ตัวอย่าง ดังนั้นจะสามารถคำนวณจำนวนตัวอย่างที่คาดว่าจะทำนายผิดพลาด เมื่อใช้ทำนายข้อมูลที่ไม่เคยเห็นได้ดังนี้

จำนวนตัวอย่างที่คาดว่าจะทำนายผิดพลาด = 6 x $U_{25\%}(1,16)$
= 6 x 0.157
= 2.512 ตัวอย่าง

จะเห็นได้ว่า เราสามารถจะแทนโนดนี้ด้วยใบที่มีค่าเป็น democrat ได้ เนื่องจากจำนวนตัวอย่างที่ทำนายผิดพลาดหลังจากแทนโนดนี้ด้วยใบใหม่แล้ว จะน้อยกว่าก่อนที่จะแทนโนดนี้ด้วยใบใหม่ และได้เป็นโนดใหม่ดังนี้

adoption of the budget resolution = y: democrat (151.0)
adoption of the budget resolution = u: democrat (1.0)
adoption of the budget resolution = n: democrat (16.0/1.0)

เมื่อคำนวณจำนวนตัวอย่างที่คาดว่าจะทำนายผิดพลาดที่เกิดสำหรับโนดใหม่ ได้ดังนี้

จำนวนตัวอย่างที่คาดว่าจะทำนายผิดพลาด $= 151 \times U_{25\%}(0,151) + 1 \times U_{25\%}(0,1) + 2.512$

$= 4.642$ ตัวอย่าง

ถ้าโนดนี้ถูกแทนด้วยใบที่มีค่าเป็น democrat อีก จะคำนวณค่าความผิดพลาดได้เท่ากับ $168 \times U_{25\%}(1,168)$ หรือ 2.610 ค่าที่คำนวณได้นี้ก็ยังมีค่าน้อยกว่าผลรวมของตัวอย่างที่คาดว่าจะผิดพลาดของโนดก่อนที่จะรวมเป็นใบ ดังนั้นโนดนี้ก็สามารถที่จะรวมเป็นใบได้อีก เมื่อตัดแต่งเป็นที่เรียบร้อยแล้ว ก็จะได้ต้นไม้ตัดสินใจใหม่ดังรูปที่ 2.8

```
physician fee freeze = n: democrat (168.0/2.6)
physician fee freeze = y: republican (123.0/13.9)
physician fee freeze = u:
|   mx missile = n: democrat (3.0/1.1)
|   mx missile = y: democrat (4.0/2.2)
|   mx missile = u: republican (2.0/1.0)
```

รูปที่ 2.8 ต้นไม้ตัดสินใจหลังการตัดเล็ม

ในต้นไม้ตัดสินใจหลังจากตัดเล็มแล้วค่า (N/E) ที่แต่ละใบ จะมีความหมายดังนี้ โดย N จะเป็นจำนวนตัวอย่างสอนทั้งหมดที่ตกอยู่ที่ใบนี้ ส่วนค่า E จะเป็นจำนวนตัวอย่างที่คาดว่าจะทำนายผิด เมื่อทำนายข้อมูล N ตัวอย่างที่ไม่เคยเห็นบนต้นไม้ต้นนี้

ผลรวมของตัวอย่างที่คาดว่าจะทำนายผิดที่แต่ละใบ เมื่อหารด้วยจำนวนตัวอย่างที่ใช้สอนทั้งหมดจะเป็นค่าประมาณของความน่าจะเป็นของความผิดพลาดของต้นไม้หลังจากตัดเล็มกิ่งแล้ว บนตัวอย่างที่ไม่เคยเห็น จากต้นไม้หลังตัดแต่งแล้วของตัวอย่างในรูปที่ 2.8 มีผลรวมของตัวอย่างที่คาดว่าจะทำนายผิดในแต่ละใบเป็น 20.8 ตัวอย่าง จากตัวอย่างทั้งหมด 300 ตัวอย่าง ดังนั้นค่าประมาณความผิดพลาดของต้นไม้หลังจากตัดแต่งแล้วบนตัวอย่างที่ไม่เคยเห็น จะทำนายผิดประมาณ 6.9 % ดังแสดงในรูปที่ 2.9

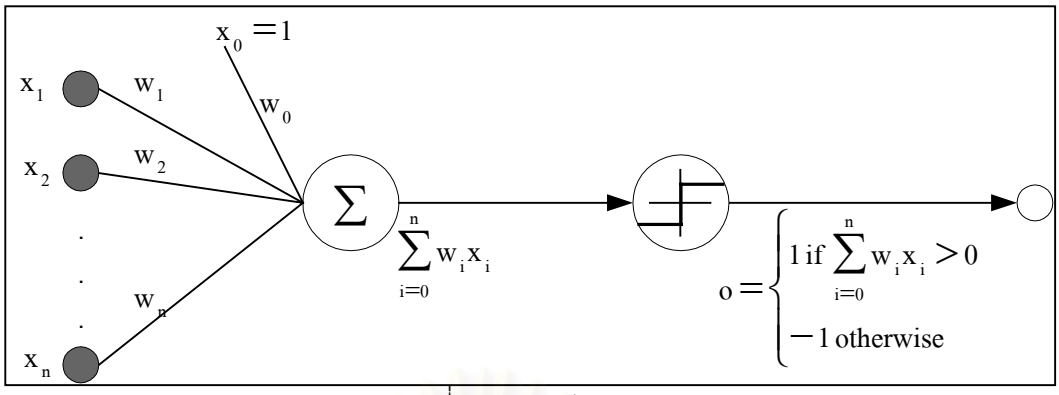
```
Evaluation on training data (300 items):

      Before Pruning          After Pruning
      ----------------        ---------------------------
     Size      Errors   Size      Errors   Estimate

      25    8( 2.7%)      7   13( 4.3%)    ( 6.9%)

Evaluation on test data (135 items):

      Before Pruning          After Pruning
      ----------------        ---------------------------
     Size      Errors   Size      Errors   Estimate

      25    7( 5.2%)      7    4( 3.0%)    ( 6.9%)
```

รูป 2.9 ผลของการตัดเล็มต้นไม้ของตัวอย่าง

2.2.3 วิธีการแบ็กพราพาเกชันนิวรอลเน็ตเวิร์ก

นิวรอลเน็ตเวิร์ก เป็นการเรียนรู้ของเครื่องรูปแบบหนึ่ง ซึ่งมีแนวคิดในการทำงานโดยการจำลองการทำงานบางส่วนของสมองมนุษย์ ที่ประกอบด้วยประสาท (neural) จำนวนมากเชื่อมต่อกัน โดยนิวรอลเน็ตเวิร์กจะจำลองให้มีประสาทจำนวนหนึ่งซึ่งเชื่อมต่อถึงกัน โดยมีค่าน้ำหนัก (weight) ของการเชื่อมต่อแต่ละแห่ง เมื่อมีการให้ตัวอย่างที่ใช้ในการเรียนรู้ นิวรอลเน็ตเวิร์กก็จะปรับค่าน้ำหนักให้เหมาะสม จนได้ผลลัพธ์ที่ถูกต้องหรือมีข้อผิดพลาดน้อยที่สุด และสามารถนำค่าน้ำหนักนี้ไปใช้ในงานที่ต้องการได้

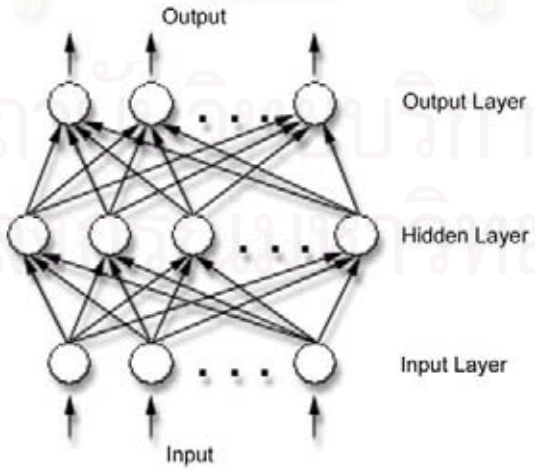
นิวรอลเน็ตเวิร์กแบบที่ง่ายที่สุดเรียกว่า เพอร์เซบตรอน (Perceptron) ซึ่งเป็นนิวรอลเน็ตเวิร์กที่มียูนิตเดียว โดยข้อมูลอินพุตสามารถมีได้หลายโนดและเป็นค่าจำนวนจริง เมื่อเพอร์เซบตรอนทำงานแล้วจะคำนวณค่าของเอาต์พุตออกมาได้ค่าหนึ่ง โดยหลังจากนั้นจะนำมาเปรียบเทียบกับค่าขีดแบ่ง (threshold) โดยหากมีค่ามากกว่าค่าขีดแบ่งแล้วจะกำหนดค่าเอาต์พุตเป็น 1 และเป็น -1 หากค่าน้อยกว่า ในการกำหนดค่าเอาต์พุตให้เป็นค่าระหว่าง 1 และ -1 นี้เรียกกว่า ไบโพลาร์ (bipolar) แต่ในบางครั้งการใช้งานเพอร์เซบตรอน สามารถกำหนดค่าเอาต์พุตเป็นแบบอื่นก็ได้ เช่น ให้เป็นค่าระหว่าง 1 และ 0 เรียกว่า ไบนารี (binary)



รูปที่ 2.10 เพอร์เซบตรอน

รูปที่ 2.10 แสดงโครงสร้างของเพอร์เซบตรอน โดยค่า x_1 ถึง x_n เป็นข้อมูลอินพุต และค่า w_0 ถึง w_n เป็นค่าน้ำหนักที่ใช้สำหรับการคำนวณของเพอร์เซบตรอน โดยให้ค่า w_0 เป็นค่าขีดแบ่ง แต่เนื่องจากต้องการเขียนการคำนวณให้อยู่ในรูปของผลรวม จึงกำหนดให้มีค่า x_0 เป็น 1 เพื่อทำการคูณกับ w_0 ดังแสดงในที่รูป 2.10 ในการใช้งานเพอร์เซบตรอนนี้สามารถจำแนกข้อมูลได้แบบเชิงเส้นเท่านั้น (linearly separable) ทำให้ได้เป็นพื้นผิวการตัดสินใจแบบเชิงเส้น (linear decision surface)

งานวิจัยนี้จะใช้แบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก [6][7][8] (backpropagation neural network) ซึ่งเป็นนิวรอลเน็ตเวิร์กที่ทำการเชื่อมต่อกันของเพอร์เซบตรอน โดยสามารถเชื่อมกันแบบหลายชั้น (multi layer) ได้ และใช้ขั้นตอนวิธีแบ็กพรอพาเกชัน (the backpropagation algorithm) โดยในขั้นตอนการทำงานจะไม่มีการป้อนผลลัพธ์ที่ได้ในแต่ละโนดย้อนกลับไปยังโนดที่ส่งข้อมูลมาให้ โครงสร้างของแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กประกอบด้วยชั้นอินพุต (input layer) ชั้นฮิดเดน (hidden layer) และชั้นเอาต์พุต (output layer) แสดงดังรูปที่ 2.11 โดยจำนวนชั้นฮิดเดนสามารถมีได้มากกว่า 1 ชั้น



รูปที่ 2.11 โครงสร้างแบ็กพรอพาเกชันเน็ตเวิร์ก

ในแต่ละโนดของนิวรอลเน็ตเวิร์กแบบหลายชั้นจะให้ค่าผลลัพธ์ ตามสมการ

$$o = \sigma(\vec{w} \cdot \vec{x})$$

โดย σ เป็นฟังก์ชันกระตุ้น (activation function) ซึ่งนิยมใช้ฟังก์ชันซิกมอยด์ (sigmoid function) ตามสมการ

$$\sigma(y) = \frac{1}{1 + e^{-y}}$$

เมื่อ o คือเอาต์พุต

$\vec{x}$ คืออินพุต

$\vec{w}$ คือค่าน้ำหนักของอินพุตนั้น ๆ

วิธีการแบ็กพรอพาเกชันจะเป็นการเรียนรู้เพื่อปรับค่าน้ำหนักสำหรับนิวรอลเน็ตเวิร์ก โดยที่ค่าน้ำหนักที่ได้จะเป็นค่าน้ำหนักที่ทำให้ค่าผลต่างกำลังสองที่น้อยที่สุด ระหว่างเอาต์พุตที่ได้จากเน็ตเวิร์กและค่าเป้าหมาย โดยมีขั้นตอนสำหรับการปรับเปลี่ยนน้ำหนักดังนี้

กำหนดให้ตัวอย่างที่ใช้ในการเรียนรู้แต่ละตัวอย่างอยู่ในรูป $(\vec{x}, \vec{t})$

เมื่อ $\vec{x}$ เป็นเวกเตอร์ของอินพุตของเน็ตเวิร์ก

$\vec{t}$ เป็นเวกเตอร์ของเป้าหมายของเอาต์พุตของเน็ตเวิร์ก

η เป็นค่าอัตราการเรียนรู้ (learning rate)

x_{ji} เป็นอินพุตขององค์ประกอบ j ซึ่งมาจากองค์ประกอบ i

w_{ji} เป็นค่าน้ำหนักขององค์ประกอบ j ซึ่งมาจากองค์ประกอบ i

1. สร้างนิวรอลเน็ตเวิร์กตามโครงสร้างที่ต้องการ
2. กำหนดจำนวนนิวรอลของแต่ละชั้น
3. กำหนดค่าน้ำหนักเริ่มต้นแบบสุ่มให้มีค่าน้อยๆ (เช่น ระหว่าง -0.05 ถึง 0.05)
4. ทำการปรับค่าน้ำหนักด้วยขั้นตอนวิธีดังนี้ จนกระทั่งลู่เข้าหรือตามเงื่อนไขที่กำหนด

   สำหรับ $(\vec{x}, \vec{t})$ แต่ละตัว ให้ทำดังนี้

   - อินพุต $\vec{x}$ ในเน็ตเวิร์ก และคำนวณเอาต์พุต O_u ในโนด u ทุกโนด

- คำนวณค่าความผิดพลาด δ_k ของโนด *k* ในชั้นเอาต์พุต โดยที่

$$\delta_k = o_k(1-o_k)(t_k - o_k)$$

- คำนวณค่าความผิดพลาด δ_h ของโนด *h* ในชั้นฮิดเดน โดยที่

$$\delta_h = o_h(1-o_h)\sum_{k\in outputs} w_{kh}\delta_k$$

- ทำการปรับค่าน้ำหนัก w_{ji} โดย

$$w_{ji} = w_{ji} + \Delta w_{ji}$$

เมื่อ $$w_{ji} = \eta\delta_j x_{ji}$$

2.2.4 การเปรียบเทียบขั้นตอนวิธี

ในการเปรียบเทียบขั้นตอนวิธีการเรียนรู้สองวิธี ว่าวิธีการใดมีประสิทธิภาพที่ดีกว่า Dietterich [9] ได้เสนอวิธีการเปรียบเทียบสองขั้นตอนวิธี (ในที่นี้เปรียบเทียบระหว่าง ขั้นตอนวิธี A และ ขั้นตอนวิธี B) โดยดูจากค่าระดับความมั่นใจ ซึ่งมีวิธีการดังต่อไปนี้

1. แบ่งข้อมูล D_0 เป็น k ส่วน ซึ่งจะได้เซ็ตย่อย $T_1, T_2, \ldots, T_k$ มีขนาดเท่ากัน
2. ให้ i มีค่าเป็น 1 ถึง k
   - $S_i \leftarrow \{D_0 - T_i\}$ หมายความว่า ให้ T_i เป็นข้อมูลในการทดสอบ และข้อมูลที่เหลือเป็นข้อมูลสอน S_i
   - $h_A \leftarrow L_A(S_i)$ หมายความว่า นำข้อมูลสอน S_i มาเข้าสู่ขั้นตอนวิธี A (ในที่นี้แทนด้วย $L_A(S_i)$) หลังจากนั้นจะได้เป็นสมมติฐาน (hypothesis) (ในที่นี้หมายถึงต้นไม้ตัดสินใจ) แทนด้วยสัญลักษณ์ h_A
   - $h_B \leftarrow L_B(S_i)$ หมายความว่า นำข้อมูลสอน S_i มาเข้าสู่ขั้นตอนวิธี B (ในที่นี้แทนด้วย $L_B(S_i)$) หลังจากนั้นจะได้เป็นสมมติฐาน แทนด้วยสัญลักษณ์ h_B
   - $\delta_i \leftarrow error_{Ti}(h_A) - error_{Ti}(h_B)$ หาค่าความผิดพลาดกับข้อมูลทดสอบของขั้นตอนวิธีแต่ละอย่าง และนำค่าความแตกต่างที่ได้เก็บเป็น δ_i
3. คำนวณหาค่า $\overline{\delta}$ โดย

$$\overline{\delta} \equiv \frac{1}{k}\sum_{i=1}^{k}\delta_i$$

เมื่อได้ค่า $\overline{\delta}$ แล้วสามารถนำมาหาค่าระดับความมั่นใจว่ามีนัยสำคัญที่กี่เปอร์เซ็นต์จากสมการ

$$\overline{\delta} \pm t_{N,k-1} S_{\overline{\delta}}$$

$t_{N,k-1}$ เป็นค่าคงที่ซึ่งได้จากตารางแจกแจงแบบ t (t Distribution) (ดูในภาคผนวก ง) โดยค่า N คือเปอร์เซ็นต์ระดับความมั่นใจ และ k-1 คือจำนวน degrees of freedom และ $S_{\overline{\delta}}$ มีค่าเป็น

$$S_{\overline{\delta}} \equiv \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} (\delta_i - \overline{\delta})^2}$$

ในขั้นตอนที่ 1 และ 2 ซึ่งเป็นขั้นตอนที่ใช้ในการแบ่งข้อมูลเป็นส่วน ๆ เพื่อหาสมมติฐานของแต่ละขั้นตอนวิธีนั้น เราเรียกว่าวิธีการครอสวาลิเดชัน (cross validation) [6] ซึ่งวิธีการเปรียบเทียบขั้นตอนวิธี ได้นำวิธีการครอสวาลิเดชันนี้เข้ามาเป็นส่วนหนึ่งของขั้นตอนการทำงาน

บทที่ 3
ขั้นตอนวิธีการตัดเล็มอย่างอ่อน

ในบทนี้จะกล่าวถึงขั้นตอนวิธีการตัดเล็มอย่างอ่อน ซึ่งประกอบไปด้วย 2 ขั้นตอนหลัก คือ (1) การสร้างกฎจากต้นไม้ตัดสินใจ และ (2) การสร้างโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กจากกฎ รวมไปถึงการสร้างข้อมูลสำหรับนิวรอลเน็ตเวิร์ก และการสอนนิวรอลเน็ตเวิร์กที่ได้

3.1 การสร้างกฎจากต้นไม้ตัดสินใจ

หลังจากนำข้อมูลสอนมาสร้างต้นไม้ตัดสินใจแล้ว ผู้ใช้งานสามารถนำต้นไม้ตัดสินใจที่ได้ไปใช้งานกับข้อมูลในอนาคตได้ แต่ในบางครั้งต้นไม้ตัดสินใจที่ได้มีขนาดใหญ่มาก ซึ่งนอกจากต้นไม้จะอิงถึงข้อมูลที่สอนมากเกินไปแล้ว ยังทำให้ผู้ใช้งานที่ต้องการอ่านเพื่อทำความเข้าใจไม่สามารถเข้าใจได้ง่าย ดังนั้นเพื่อทำให้การใช้งานมีความสะดวกมากยิ่งขึ้น เราสามารถเปลี่ยนแปลงรูปแบบต้นไม้ตัดสินใจให้อยู่ในรูปของกฎได้

วิธีการสร้างกฎจากต้นไม้ตัดสินใจ สามารถทำได้โดยการอ่านต้นไม้ตัดสินใจจากรากไปจนถึงใบหนึ่ง ๆ ซึ่งเส้นทางที่ได้จะต้องผ่านโนดที่เก็บคุณสมบัติ และกิ่งซึ่งเก็บค่าที่เป็นไปได้ของโนดนั้น ๆ โดยกฎที่ได้ในต้นไม้ตัดสินใจในแต่ละโนด จะมีจำนวนที่ผ่านโนดและกิ่งต่าง ๆ ไม่เท่ากัน หลังจากนั้นสร้างกฎโดยเปลี่ยนให้อยู่ในรูป “ถ้า...แล้ว” โดยนำโนดและกิ่ง อยู่ในส่วนของ “ถ้า” หากในเส้นทางของกฎที่อ่านมาได้มีโนดและกิ่งอยู่ถัดมาจากโนดที่แล้ว ก็จะเชื่อมต่อกันด้วย “และ” ต่อไป จนเส้นทางที่อ่านมาถึงใบ ซึ่งค่าที่ใบเป็นค่าของกลุ่มที่เป็นไปได้ โดยนำค่าที่ใบอยู่ในส่วนของ “แล้ว”

จากตารางที่ 2.2 ซึ่งเป็นข้อมูลที่ใช้ในการตัดสินใจเล่นกอล์ฟ เมื่อสร้างต้นไม้ตัดสินใจได้เป็นรูปที่ 2.2 สามารถทำเป็นกฎตามวิธีการข้างต้นได้เป็นกฎทั้งหมด 5 ข้อดังรูปที่ 3.1

(1) **ถ้า** สภาพแวดล้อมเป็นแดดจ้า **และ** ความชื้นมีค่าน้อยกว่าเท่ากับ 75 % **แล้ว** ออกรอบ
(2) **ถ้า** สภาพแวดล้อมเป็นแดดจ้า **และ** ความชื้นมีค่ามากกว่า 75 % **แล้ว** ไม่ออกรอบ
(3) **ถ้า** สภาพแวดล้อมเป็นแดดร่ม **แล้ว** ออกรอบ
(4) **ถ้า** สภาพแวดล้อมเป็นฝนตก **และ** กระแสลมเป็นลมแรง **แล้ว** ไม่ออกรอบ
(5) **ถ้า** สภาพแวดล้อมเป็นฝนตก **และ** กระแสลมเป็นลมปกติ **แล้ว** ออกรอบ

รูปที่ 3.1 กฎที่ได้จากต้นไม้ตัดสินใจที่ใช้ทดสอบการเล่นกอล์ฟ

จากกฎทั้งห้าข้อในรูปที่ 3.1 เราสามารถใช้งานได้เช่นเดียวกับต้นไม้ตัดสินใจ โดยเมื่อนำข้อมูลที่ต้องการทดสอบเข้ามาตรวจสอบกับกฎข้อต่าง ๆ ทีละข้อ หากข้อมูลที่นำมาทดสอบนี้มีค่าเป็นคุณสมบัติที่ตรงกับกฎทุกคุณสมบัติ ข้อมูลที่นำมาทดสอบชุดนั้นก็จะมีกลุ่มตามกฎข้อนั้น ๆ หากไม่ตรงทั้งหมด ก็

ทำการเปรียบเทียบกับกฎข้อต่อ ๆ ไป ด้วยวิธีการใช้กฎนี้เอง จะทำให้ผู้ใช้งานสามารถทำความเข้าใจได้ง่ายกว่าในรูปของต้นไม้ตัดสินใจ

การแปลงต้นไม้ตัดสินใจเป็นกฎมีข้อดีที่สำคัญอีกอย่างก็คือ โนดภายใน (internal node) ในต้นไม้ตัดสินใจโนดหนึ่ง ๆ ที่สนใจ จะปรากฏในกฎมากกว่าหนึ่งกฎ ต่างจากต้นไม้ตัดสินใจที่โนดนั้น ๆ จะปรากฏอยู่ครั้งเดียวและแยกตามค่าที่เป็นไปได้ของคุณสมบัติไปแต่ละกิ่ง หากเข้าสู่วิธีการตัดเล็มและโนดนี้ถูกตัดเล็มแล้ว ก็จะทำให้คุณสมบัตินั้นหายไปจากต้นไม้ตัดสินใจทันที แต่จากการแปลงเป็นกฎจะทำให้เกิดความยืดหยุ่นในการตัดเล็มมากกว่าจะกระทำโดยตรงกับต้นไม้ตัดสินใจ กล่าวคือโนดนั้นสามารถจะถูกตัดออกจากกฎหนึ่ง แต่ไม่จำเป็นต้องถูกตัดออกจากอีกกฎหนึ่งได้ ทำให้โนดนั้น ๆ ที่อยู่ในหลาย ๆ กฎสามารถถูกตัดเล็มได้อย่างอิสระไม่ขึ้นต่อกัน เช่น ในบางครั้งโนด x ซึ่งกำลังตรวจสอบคุณสมบัติของข้อมูลเมื่อรวมกับโนดอื่นเช่น โนด y ในกฎข้อหนึ่งแล้วทำให้ความสำคัญของ x ลดลง ก็ควรจะถูกตัดออกได้ แต่ถ้าโนด x ที่อยู่ในกฎอีกข้อหนึ่งไม่ได้รวมกับโนด y แล้วโนด x ก็ยังคงมีความสำคัญอยู่มากซึ่งไม่ควรจะถูกตัดออก กฎแต่ละกฎแทนเส้นทางจากรากถึงใบในต้นไม้ตัดสินใจ หากไม่แปลงต้นไม้ตัดสินใจเป็นกฎ ก็จะไม่สามารถตัดโนดแยกจากกันได้อย่างอิสระอย่างเช่นที่ทำได้กับกฎ อย่างไรก็ตามเพื่อให้การตัดเล็มมีความยืดหยุ่นมากขึ้น วิทยานิพนธ์ฉบับนี้จะไม่ตัดโนดทิ้งเลยทีเดียวแต่ละให้น้ำหนักตามความสำคัญ กล่าวคือหากโนดที่มีความสำคัญมากก็จะให้น้ำหนักมาก และโนดที่มีความสำคัญน้อยก็จะให้น้ำหนักน้อย โดยจะอธิบายวิธีการในหัวข้อต่อไป

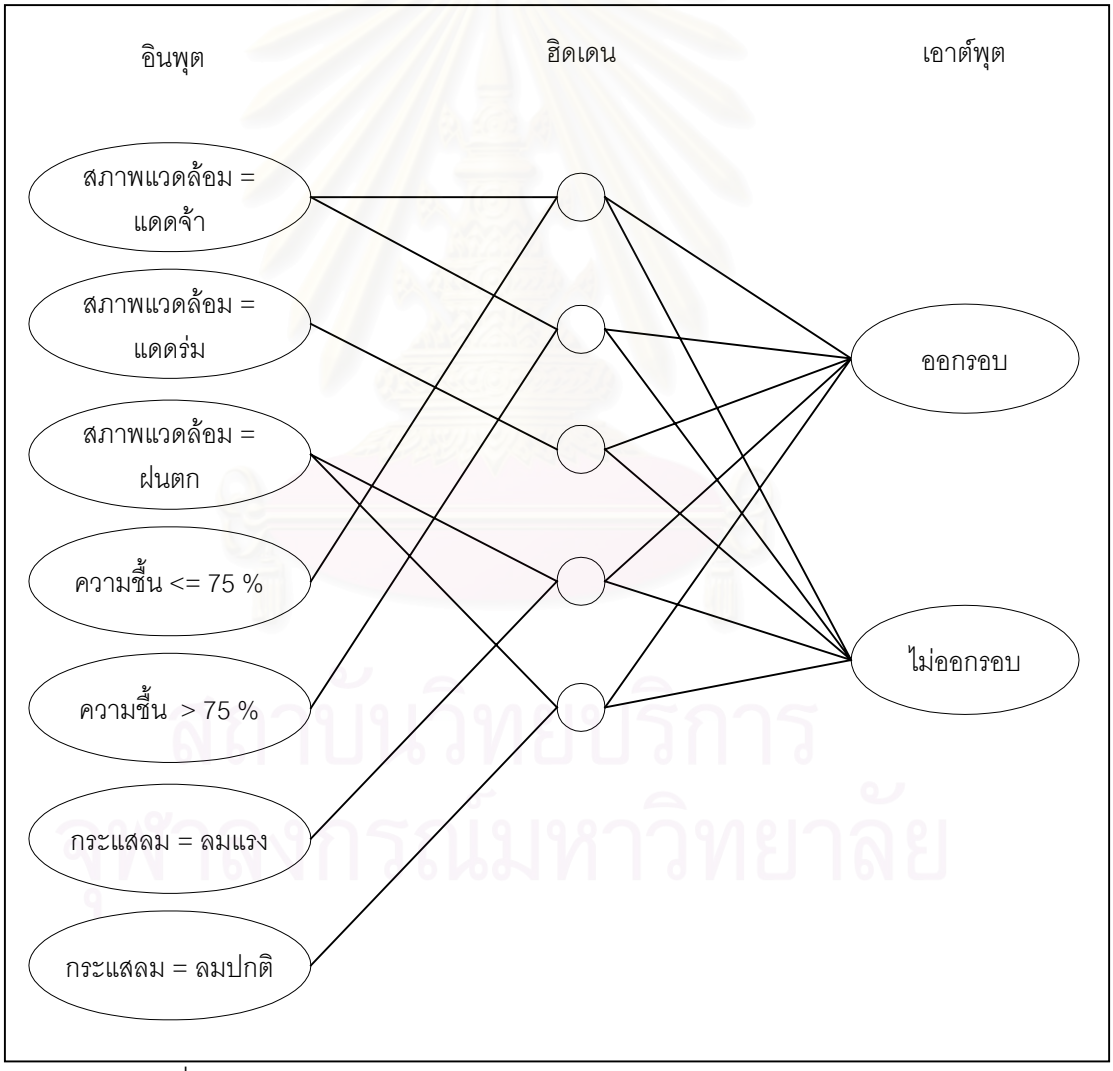
3.2 การสร้างโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กจากกฎ

ในการสร้างโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก จะต้องจำลองโครงสร้างของต้นไม้ตัดสินใจ เพื่อทำให้โครงสร้างที่อยู่ในแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กนี้เปรียบเสมือนต้นไม้ตัดสินใจนั่นเอง วิธีการสร้างโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก ทำได้โดยนำต้นไม้ตัดสินใจที่ไม่มีการตัดเล็มนำมาสร้างเป็นกฎก่อน จากนั้นจึงสร้างโครงสร้างของนิวรอลเน็ตเวิร์กจากกฎที่ได้ ในกฎแต่ละข้อจะนำคุณสมบัติในการทดสอบมาสร้างเป็นอินพุตโนด (input node) ในแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก จำนวนของอินพุตโนดสำหรับกฎข้อหนึ่ง ๆ จะเท่ากับจำนวนการทดสอบคุณสมบัติของกฎข้อนั้น อินพุตโนดของกฎหนึ่ง ๆ ทุกโนดจะเชื่อมต่อฮิดเดนโนด (hidden node) หนึ่งโนดเปรียบเสมือนกับการเชื่อมโยงกันเป็นกฎข้อนั้น กล่าวคือ ฮิดเดนโนดหนึ่งโนดในแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก จะแทนการเชื่อมกันด้วย "และ" ของการทดสอบคุณสมบัติในกฎที่ได้จากต้นไม้ตัดสินใจ ดังนั้นจำนวนฮิดเดนโนดในเน็ตเวิร์กจะมีจำนวนเท่ากับจำนวนกฎจากต้นไม้ตัดสินใจ ส่วนเอาต์พุตโนด (output node) ของเน็ตเวิร์กจะมีจำนวนเท่ากับกลุ่มที่มีอยู่ในต้นไม้ตัดสินใจ โดยการเชื่อมต่อระหว่างฮิดเดนโนดกับเอาท์พุตโนดจะเป็นแบบต่อกันหมด (fully connected)

ต่อไปจะขออธิบายตัวอย่างในการสร้างโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก จากกฎทั้ง 5 ข้อในรูปที่ 3.1 พิจารณากฎข้อที่ 1 คือ

ถ้า สภาพแวดล้อมเป็นแดดจ้า **และ** ความชื้นมีค่าน้อยกว่าเท่ากับ 75 % **แล้ว** ออกรอบ

ในกฎข้อนี้มีคุณสมบัติที่เกี่ยวข้องอยู่ 2 คุณสมบัติคือ สภาพแวดล้อม และความชื้น และมีค่าที่ใช้ในการทดสอบคุณสมบัติเป็น แดดจ้า และมีค่าน้อยกว่าเท่ากับ 75 % ตามลำดับ นั่นคือโครงสร้างของแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กในส่วนอินพุตโนดของกฎข้อนี้จะมีอยู่สองโนดคือ “สภาพแวดล้อม = แดดจ้า” และ “ความชื้น <= 75%” ดังแสดงในรูปที่ 3.2 โครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กที่ได้ของต้นไม้ตัดสินใจในการออกรอบเล่นกอล์ฟโดยรวมแสดงได้ดังรูปที่ 3.2



รูปที่ 3.2 โครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กในการตัดสินใจเล่นกอล์ฟ

3.3 การสร้างข้อมูลสำหรับโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก

เมื่อได้โครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กแล้ว ขั้นตอนต่อมาคือการสร้างข้อมูลสำหรับใช้ในการสอนนิวรอลเน็ตเวิร์กนี้ โดยข้อมูลที่ใช้ในการสอนและข้อมูลที่ใช้ในการทดสอบก็คือข้อมูลที่ใช้ในการสร้างและทดสอบของต้นไม้ตัดสินใจ ข้อมูลเหล่านี้จะถูกแปลงเพื่อสามารถใช้งานได้กับโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก โดยข้อมูลตัวอย่างหนึ่ง ๆ จะแปลงเป็นส่วนชั้นอินพุตและชั้นเอาต์พุต ในวิทยานิพนธ์ฉบับนี้จะใช้ข้อมูลเป็นแบบไบนารีทั้งในส่วนอินพุตและเอาต์พุต ในอินพุตโนดจะพิจารณาตามการทดสอบของอินพุตโนดนั้น หากคุณสมบัติของข้อมูลมีค่าความจริงเป็นเท็จ ให้ค่าของอินพุตโนดนั้นมีค่าเป็น "0" หากมีค่าความจริงเป็นจริงให้ค่าอินพุตโนดนั้นมีค่าเป็น "1" สำหรับเอาต์พุตโนดจะกำหนดโนดที่เป็นกลุ่มของตัวอย่างนั้น ๆ ด้วย “1” และโนดที่เหลือเป็น “0”

ตัวอย่างในการสร้างข้อมูลสำหรับโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กที่ได้จากการตัดสินใจเล่นกอล์ฟ จากตารางที่ 2.2 ในข้อมูลตัวแรก ซึ่งมีข้อมูลคือ “สภาพแวดล้อมเป็นแดดจ้า อุณหภูมิเป็น 75 °F มีความชื้นที่ 70 % สภาพลมมีลมแรง และทำการออกรอบ” เมื่อนำมาสร้างข้อมูลจากโครงสร้างตามรูปที่ 3.2 จะได้อินพุตโนดเป็น “1 0 0 1 0 1 0” ตามลำดับ และเอาต์พุตโนดเป็น “1 0”

<u>การจัดการกับตัวอย่างที่ไม่ทราบค่า</u>

ข้อมูลในการสร้างต้นไม้ตัดสินใจในบางครั้งจะมีข้อมูลที่ไม่ทราบค่าได้ และเมื่อต้องเปลี่ยนข้อมูลมาใช้ในโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กก็จำเป็นจะต้องกำหนดค่าให้ข้อมูลเหล่านั้นเช่นกัน ในการกำหนดค่าให้กับข้อมูลไม่ทราบค่าในวิทยานิพนธ์ฉบับนี้จะใช้วิธีการเดียวกับการสร้างต้นไม้ตัดสินใจ กล่าวคือกำหนดค่าให้โดยอิงตามค่าที่เป็นไปได้ของคุณสมบัตินั้น ๆ ตามข้อมูลที่ใช้สอน กล่าวคือ โนดที่ทดสอบกับคุณสมบัติที่ไม่ทราบค่า ให้มีค่าเท่ากับค่าที่เป็นไปได้ของโนดนั้นหารด้วยจำนวนตัวอย่างทั้งหมดที่ทราบค่า

จากตารางที่ 2.2 ในการตัดสินใจเล่นกอล์ฟ หากข้อมูลตัวแรกคือ “สภาพแวดล้อมเป็นแดดจ้า อุณหภูมิเป็น 75 °F มีความชื้นที่ 70 % สภาพลมมีลมแรง และทำการออกรอบ” ไม่ทราบค่าสภาพแวดล้อม เมื่อเราสนใจตัวอย่างที่เหลืออีก 13 ตัว ซึ่งทราบค่าคุณสมบัติสภาพแวดล้อม โดยแบ่งเป็นสภาพแวดล้อมแดดจ้า 4 ตัวอย่าง สภาพแวดล้อมแดดร่ม 4 ตัวอย่าง และสภาพแวดล้อมฝนตก 5 ตัวอย่าง เมื่อนำมาแปลงข้อมูลเพื่อเข้าสู่โครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กในโนด “สภาพแวดล้อมเท่ากับแดดจ้า” จะมีค่าเท่ากับ 4/13 ในโนด “สภาพแวดล้อมเท่ากับแดดร่ม” มีค่าเท่ากับ 4/13 และโนด “สภาพแวดล้อมเท่ากับฝนตก” มีค่าเป็น 5/13 ดังนั้นเมื่อแปลงข้อมูลทั้งหมดได้อินพุตโนดเป็น “4/13 4/13 5/13 1 0 1 0” ตามลำดับ

บทที่ 4
การทดลองและผลการทดลอง

บทนี้กล่าวถึงการทดลองการตัดเล็มอย่างอ่อนโดยใช้โครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กที่ได้จากชุดข้อมูลทั้งหมด 20 ชุดข้อมูล เปรียบเทียบกับค่าความถูกต้องที่ได้จากต้นไม้ตัดสินใจที่ยังไม่ได้ตัดเล็ม และต้นไม้ตัดสินใจที่ทำการตัดเล็มโดยใช้ค่าความผิดพลาด หลังจากนั้นจะคำนวณเปรียบเทียบขั้นตอนวิธีระหว่างการตัดเล็มอย่างอ่อนกับต้นไม้ตัดสินใจที่ยังไม่ได้ตัดเล็ม และต้นไม้ตัดสินใจที่ตัดเล็มแล้ว

4.1 วิธีการทดลอง

ในการทดลองเพื่อเปรียบเทียบวิธีการตัดเล็มอย่างอ่อนกับต้นไม้ตัดสินใจนั้น เราจะเลือกชุดข้อมูลจาก Department of Information and Computer Science, University of California, Irvine [2] ซึ่งเป็นที่รวบรวมชุดข้อมูลสำหรับเป็นเกณฑ์เปรียบเทียบสมรรถนะ (benchmark) โดยในการเลือกชุดข้อมูลพิจารณาจาก

- จำนวนกลุ่มของชุดข้อมูล โดยในการทดลองจะเลือกชุดข้อมูลที่มีตั้งแต่ 3 กลุ่มขึ้นไป
- ข้อมูลสามารถนำมาใช้งานได้ทันทีหรือเปลี่ยนแปลงเพียงเล็กน้อย โดยข้อมูลสามารถนำไปใช้งานใน ระบบ C4.5 ซึ่งเป็นโปรแกรมที่ใช้ในการสร้างต้นไม้ตัดสินใจที่จะใช้ในการทดลองของวิทยานิพนธ์ฉบับนี้

ชุดข้อมูลที่เลือกมามีทั้งหมด 20 ชุดข้อมูลโดยมีรายละเอียดดังตารางที่ 4.1

ตารางที่ 4.1 ชุดข้อมูลที่ใช้ในการทดลอง

ชื่อชุดข้อมูล	จำนวนคุณสมบัติ	จำนวนกลุ่ม	จำนวนข้อมูลสอน	จำนวนข้อมูลทดสอบ
Anneal	38	6	798	100
Soybean	35	19	307	376
Balance-scale* (625)	4	3	-	-
Iris* (150)	4	3	-	-
Thyroid-disease (allbp)	29	3	2800	972
Thyroid-disease (allhyper)	29	5	2800	972
Thyroid-disease (allhypo)	29	5	2800	972
Thyroid-disease (allrep)	29	4	2800	972

ตารางที่ 4.1 ชุดข้อมูลที่ใช้ในการทดลอง (ต่อ)

ชื่อชุดข้อมูล	จำนวนคุณสมบัติ	จำนวนกลุ่ม	จำนวนข้อมูลสอน	จำนวนข้อมูลทดสอบ
Image	19	7	210	2100
Restricted (lymphography)* (148)	18	4	-	-
Restricted (primary-tumor)* (339)	17	22	-	-
Statlog (satimage)	36	6	4435	2000
Statlog (segment)* (2310)	19	7	-	-
Statlog (shuttle)	9	7	43500	14500
Wine* (178)	13	3	-	-
Waveform* (5000)	21	3	-	-
Waveform + noise* (5000)	40	3	-	-
Glass* (214)	9	6	-	-
Led-display-creator	7	10	2000	500
Led 17	24	10	2000	500

** ชุดข้อมูลที่ไม่มีการแบ่งข้อมูลสอนและข้อมูลทดสอบ*

จากตารางที่ 4.1 จะเห็นว่าชุดข้อมูลที่เลือกมาแบ่งออกเป็นสองประเภทคือ

(1) ชุดข้อมูลที่แบ่งข้อมูลสอนและข้อมูลทดสอบไว้แล้ว สำหรับชุดข้อมูลชนิดนี้การทดลองจะใช้ข้อมูลสอนสร้างต้นไม้ตัดสินใจ และทดสอบความถูกต้องด้วยข้อมูลทดสอบตามที่ได้มีการแบ่งไว้แล้ว หลังจากนั้นนำต้นไม้ตัดสินใจที่ได้นำมาสร้างโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กและเข้าสู่การสอนของนิวรอลเน็ตเวิร์กและหาค่าความถูกต้องของโครงสร้างนี้

(2) ชุดข้อมูลที่ไม่มีการแบ่งข้อมูลสอนและข้อมูลทดสอบ (ชุดข้อมูลที่มีเครื่องหมาย * กำกับไว้และบอกจำนวนตัวอย่างทั้งหมดไว้ หลังชื่อชุดข้อมูล) ในชุดข้อมูลชนิดนี้จะใช้วิธีการครอสวาลิเดชัน (cross validation) [6] (ขั้นตอนที่ 1 และ 2 ในหัวข้อ 2.2.4) โดยแบ่งชุดข้อมูลเป็น 6 โฟลด์ (fold) หลังจากนั้นแต่ละโฟลด์ที่ได้จะสร้างต้นไม้ตัดสินใจและโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กต่อไป

วิธีการทดสอบค่าความถูกต้องบนแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก จะใช้วิธีการเลือกโนดของเอาต์พุตหนึ่งโนดที่มีค่าสูงสุดเป็นโนดที่แสดงถึงกลุ่มที่ได้ในตัวอย่างนั้น ๆ สาเหตุที่ใช้วิธีการนี้เนื่องจากการกำหนดข้อมูลจากโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กที่สร้างไว้ในส่วนของเอาต์พุต เป็นการกำหนดให้แต่ละกลุ่มที่เป็นไปได้ เป็นโนดหนึ่งโนดในชั้นเอาต์พุต ดังนั้นตัวอย่างของข้อมูลหนึ่งตัวอย่างสามารถที่จะเป็นกลุ่มใดกลุ่มหนึ่งได้เพียงกลุ่มเดียวเท่านั้น ดังนั้นวิธีการเลือกโนดที่มีค่าสูงสุดในชั้นเอาต์-พุตจึงสามารถใช้ในการทดลองนี้ได้

4.2 ผลการทดลอง

หลังจากทดสอบค่าความถูกต้องที่ได้จากโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กแล้ว ในงานวิทยานิพนธ์ฉบับนี้จะนำมาเปรียบเทียบขั้นตอนวิธีกับ ต้นไม้ตัดสินใจที่ยังไม่ได้ตัดเล็ม และต้นไม้ตัดสินใจที่ทำการตัดเล็มแล้ว เพื่อเปรียบเทียบว่าวิธีการใดมีประสิทธิภาพที่ดีกว่า ในการทดลองนี้ใช้วิธีการเปรียบเทียบสองขั้นตอนวิธีโดยดูจากค่าระดับความมั่นใจ (confidence level) ซึ่งเป็นการวัดความสำคัญทางสถิติว่าวิธีการหนึ่งดีกว่าอีกวิธีการหนึ่งด้วยความมั่นใจกี่เปอร์เซ็นต์ ผลการทดลองที่ได้แสดงในตารางที่ 4.2

ตารางที่ 4.2 การเปรียบเทียบผลการทดลองด้วยวิธีค่าระดับความมั่นใจกับต้นไม้ตัดสินใจที่ยังไม่ได้ตัดเล็มและต้นไม้ติดสินใจที่ทำการตัดเล็มแล้ว

ชื่อชุดข้อมูล	C4.5 Unpruned	C4.5 Pruned	Soft-Pruning	CL (%) of Soft-Pruning Vs C4.5 Unpruned	CL (%) of Soft-Pruning Vs C4.5 Pruned
Anneal	97	95	97	=	+ 75
Soybean	85.64	86.70	90.96	+ 97.5	+ 95
Balance-scale	69.76 ± 3.81	65.77 ± 5.11	86.56 ± 3.37	+ 99	+ 99
Iris	95.33 ± 3.01	94 ± 4.2	95.33 ± 3.01	=	+ 90
Thyroid-disease (allbp)	96.81	97.84	97.12	=	- 75
Thyroid-disease (allhyper)	98.87	98.56	98.87	=	=
Thyroid-disease (allhypo)	99.49	99.49	99.69	+ 75	+ 75
Thyroid-disease (allrep)	98.77	99.07	98.97	=	=
Image	89.43	91	90.38	+ 84	- 75
Restricted (lymphography)	73.7 ± 6.94	78.38 ± 4.83	77.72 ± 7	=	=
Restricted (primary-tumor)	42.19± 7.34	41.32 ± 8.11	38.66 ± 6.82	=	=
Statlog (satimage)	84.9	85.45	86.80	+ 95	+ 84
Statlog (segment)	96.97 ± 0.9	96.97 ± 0.9	96.75 ± 1.5	=	=
Statlog (shuttle)	99.97	99.95	99.99	+ 84	+ 97.5
Wine	93.30 ± 4.71	93.30 ± 4.71	93.30 ± 4.71	=	=
Waveform	75.76 ± 1.44	75.82 ± 1.47	80.30 ± 1.51	+ 99	+ 99
Waveform + noise	75.22 ± 1.42	75.30 ± 1.37	79.64 ± 0.69	+ 99	+ 99
Glass	66.34 ± 4.46	66.34 ± 4.46	67.22 ± 7.39	=	=
Led-display-creator	75.4	75.2	74.60	=	=
Led 17	66.4	75.2	63.80	- 75	- 99

จากตารางข้างต้น ในหัวข้อ C4.5 Unpruned C4.5 Pruned และ Soft-Pruning เป็นสดมภ์ที่แสดงค่าความถูกต้องของต้นไม้ตัดสินใจที่ไม่ทำการตัดเล็มของ C4.5 ของต้นไม้ตัดสินใจที่ทำการตัดเล็มแล้วของ C4.5 และของวิธีการที่นำเสนอตามลำดับ โดยในชุดข้อมูลที่แบ่งข้อมูลเป็น 6 โฟลด์ ค่าความถูกต้องจะอยู่ในรูป $x \pm y$ โดยค่า x คือ ค่าความถูกต้องเฉลี่ยของ 6 โฟลด์ และค่า y คือค่าเบี่ยงเบนมาตรฐาน (standard deviation)

สำหรับคอลัมน์ CL of Soft-Pruning Vs C4.5 Unpruned และ CL of Soft-Pruning Vs C4.5 Pruned เป็นคอลัมน์ที่แสดงการเปรียบเทียบด้วยค่าระดับความมั่นใจของความถูกต้องของวิธีการตัดเล็มอย่างอ่อนกับของ C4.5 โดยค่าต่าง ๆ ในตารางแสดงความหมายคือ

- +x แสดงกรณีที่วิธีการตัดเล็มอย่างอ่อนดีกว่าด้วยความมั่นใจมากกว่าหรือเท่ากับ x %
- -x แสดงกรณีที่วิธีการตัดเล็มอย่างอ่อนด้อยกว่าด้วยความมั่นใจมากกว่าหรือเท่ากับ x %
- = แสดงกรณีที่วิธีทั้งสองไม่มีความแตกต่างอย่างมีนัยสำคัญทางสถิติ

จากการเปรียบเทียบด้วยค่าระดับความมั่นใจในขั้นตอนวิธีตามตารางที่ 4.2 จะเห็นว่าวิธีของงานวิทยานิพนธ์ฉบับนี้ดีกว่าต้นไม้ตัดสินใจที่ไม่ทำการตัดเล็ม 8 ชุดข้อมูล และด้อยกว่า 1 ชุดข้อมูล ส่วนชุดข้อมูลที่เหลือมีประสิทธิภาพที่ใกล้เคียงกัน และเมื่อเปรียบเทียบกับต้นไม้ตัดสินใจที่ทำการตัดเล็มแล้ววิธีของงานวิจัยนี้ดีกว่า 9 ชุดข้อมูล และด้อยกว่า 3 ชุดข้อมูล ดังนั้นผลการทดลองโดยรวมแสดงให้เห็นว่าวิธีการของงานวิจัยนี้มีประสิทธิภาพดีที่สุดในสามวิธี ตามด้วยต้นไม้ตัดสินใจที่ทำการตัดเล็มแล้ว ส่วนต้นไม้ตัดสินใจที่ยังไม่ได้ตัดเล็มให้ประสิทธิภาพน้อยที่สุด

ในตารางที่ 4.2 เป็นผลที่ได้จากวิธีการที่เสนอเมื่อใช้ค่า 0 และ 1 แทน ค่าความจริงที่เป็น “จริง” และ “เท็จ” สำหรับอินพุตโนด อย่างไรก็ดีการแทนค่าความจริงนั้น เราอาจเลือกใช้เป็นไบโพลาร์ คือ -1 และ 1 ก็ได้ ผลการทดลองที่ใช้ไบโพลาร์สำหรับอินพุตโนดแสดงในภาคผนวก ก. ซึ่งให้ผลในแนวเดียวกันกับการใช้ 0 และ 1

สำหรับการทดลองและผลการทดลองข้างต้นในงานวิจัยนี้ได้ทดลองโดยใช้โปรแกรม C4.5 สำหรับการสร้างต้นไม้ตัดสินใจ รวมไปถึงต้นไม้ตัดสินใจที่ทำการตัดเล็มโดยใช้ค่าความผิดพลาด และใช้โปรแกรม Aspirin/MIGRAINES Neural Network Software สำหรับโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก นอกจากนี้ผู้วิจัยได้สร้างโปรแกรมขึ้นมาสำหรับงานวิจัยนี้ (ดูการใช้งานโปรแกรม ในภาคผนวก ข) โดยนำโมดูล (module) ที่ใช้สำหรับงานวิจัยนี้ของสองโปรแกรมมารวมกัน

บทที่ 5
สรุปผลการวิจัย และข้อเสนอแนะ

5.1 สรุปผลการวิจัย

การตัดเล็มต้นไม้ตัดสินใจในปัจจุบันเป็นการตัดเล็มที่มีวิธีการตัดโนดภายในต้นไม้ออกแล้วแทนที่ด้วยใบ ด้วยวิธีการนี้จะทำให้โนดที่ถูกตัดออกไปซึ่งเก็บคุณสมบัติที่ใช้ทำการทดสอบหายไปจากต้นไม้ตัดสินใจ แต่ในบางครั้งโนดที่ถูกตัดออกไปอาจมีความสำคัญอยู่ แต่เมื่อเข้ากระบวนการตัดเล็มแล้วทำให้โดนตัดออกไป ในงานวิจัยฉบับนี้ได้คิดวิธีการที่ตัดเล็มต้นไม้โดยไม่ใช่เป็นการตัดโนดของต้นไม้ออกไป แต่จะนำวิธีการแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กเข้ามาใช้ในการตัดเล็ม และเรียกวิธีการนี้ว่า การตัดเล็มอย่างอ่อน

งานวิจัยนี้ได้นำเสนอวิธีการตัดเล็มแบบใหม่ ด้วยการใช้โครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ต-เวิร์กเป็นส่วนช่วยในการตัดเล็ม โดยขั้นตอนในการตัดเล็มอย่างอ่อนจะนำข้อมูลสอนมาสร้างต้นไม้ตัดสินใจและทดสอบค่าความถูกต้องของต้นไม้ตัดสินใจด้วยข้อมูลทดสอบ ต้นไม้ตัดสินใจที่ได้นำมาสร้างให้อยู่ในรูปของกฎ หลังจากนั้นทำการสร้างโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กจากกฎที่ได้ โครง-สร้างที่ได้จะเป็นการจำลองต้นไม้ตัดสินใจ ในส่วนของการสอนและการทดสอบโครงสร้างแบ็กพรอพาเก-ชันนิวรอลเน็ตเวิร์กนั้น จะใช้ข้อมูลเดียวกับที่ใช้สร้างต้นไม้ตัดสินใจ

ผลการทดลองกับชุดข้อมูล 20 ชุดข้อมูล และเปรียบเทียบด้วยค่าระดับความมั่นใจ ระหว่างงานวิจัยนี้กับต้นไม้ตัดสินใจที่ไม่ทำการตัดเล็ม ได้ผลดีกว่า 8 ชุดข้อมูล ด้อยกว่า 1 ชุดข้อมูล และเมื่อเปรียบ-เทียบกับต้นไม้ตัดสินใจที่ทำการตัดเล็มแล้ว ได้ผลดีกว่า 9 ชุดข้อมูล และด้อยกว่า 3 ชุดข้อมูล สามารถสรุปได้ว่าผลของงานวิจัยนี้ให้ประสิทธิภาพที่ดีที่สุด

5.2 ข้อเสนอแนะ

1. ขั้นตอนการสอนโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก ในขณะที่ทำการสอน น้ำหนักในการเชื่อมต่อในส่วนต่าง ๆ คือ อินพุตโนดกับฮิดเดนโนด และ ฮิดเดนโนดกับเอาต์พุตโนด จะเปลี่ยนแปลงไปตามข้อมูลสอน ด้วยวิธีการนี้จะเป็นการให้น้ำหนักมากกับคุณสมบัติที่มีความสำคัญมาก และน้ำหนักน้อยกับคุณสมบัติที่มีความสำคัญน้อย หากเพิ่มวิธีการตัดเล็มโนดในแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กเข้าไปในระหว่างการสอน อาจทำให้ค่าความถูกต้องของโครงสร้างดังกล่าวมีความถูกต้องมากยิ่งขึ้น
2. ในการสร้างข้อมูลให้กับโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กจากข้อมูลที่ใช้สร้างต้นไม้ตัดสิน-ใจนั้น งานวิจัยนี้ได้ใช้ข้อมูลแบบไบนารี ซึ่งเป็นเพียงการตรวจสอบค่าความเป็นจริงระหว่างอินพุตโนดและตัวอย่างเท่านั้น หากมีการกำหนดน้ำหนักให้แก่คุณสมบัติแต่ละอย่างได้อาจจะทำให้ค่าความถูกต้องจากโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กดีขึ้นได้ ซึ่งการกำหนดน้ำหนักให้กับคุณสมบัติแต่ละอย่างของข้อมูลนั้นอาจนำเทคนิคฟัซซีเซ็ต (fuzzy set) มาใช้

3. วิธีการแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กที่ใช้ในงานวิจัยนี้ ในขั้นตอนการฝึกโครงสร้างนิวรอลเน็ต-เวิร์กจะใช้ข้อมูลสอนจำนวนหนึ่งเป็นตัวหยุดการฝึกโครงสร้าง ด้วยวิธีการนี้อาจทำให้น้ำหนักต่าง ๆ ที่ได้หลังจากการฝึกไม่ดีพอ เมื่อนำมาตรวจสอบค่าความถูกต้องกับข้อมูลที่ใช้ทดสอบอาจจะได้ผลไม่ดี เนื่องจากข้อมูลสอนถูกนำมาเป็นตัวหยุดการฝึก เราอาจหาวิธีการหยุดฝึกโครงสร้างนิวรอลเน็ตแบบอื่นได้ เช่น การใช้วาลิเดชันเซต (validation set) ซึ่งเป็นวิธีการแบ่งข้อมูลสอนออกมาเฉพาะเพื่อใช้ในการหยุดฝึกโครงสร้าง และข้อมูลที่แบ่งมานี้จะไม่นำไปใช้ในการฝึกโครงสร้าง
4. นอกจากโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กที่ใช้เปรียบเสมือนการตัดเล็มต้นไม้ตัดสินใจในงานวิจัยนี้แล้ว อาจเปลี่ยนแปลงโครงสร้าง หรือนำการเรียนรู้ของเครื่องวิธีอื่นมาใช้เพื่อทำการตัดเล็มต้นไม้ตัดสินใจได้

รายการอ้างอิง

[1] สุกรี สินธุภิญโญ. การประยุกต์การโปรแกรมตรรกะเชิงอุปนัยและแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กในการรู้จำตัวพิมพ์อักษรไทย. วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาวิศวกรรมคอมพิวเตอร์ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2541.

[2] Blake, C.L. & Merz, C.J. UCI Repository of machine learning databases Irvine, CA: University of California, Department of Information and Computer Science, 1998. Available from: http://www.ics.uci.edu/~mlearn/MLRepository.html.

[3] Esposito, F., Malerba, D., Semeraro, G., Kay, J. A Comparative Analysis of Methods for Pruning Decision Trees. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 5, pp. 476-491, 1997.

[4] Quinlan, J. R. C4.5 Programs For Machine Learning, California : Morgan Kaufmann, 1993.

[5] Quinlan, J. R. Induction of decision trees. Machine Learning, pp. 81-106, 1986.

[6] Mitchell, T. M. Machine Learning. The McGraw-Hill Companies, Inc., 1997.

[7] Rich, E., Knight, K. Artificial Intelligence. Singapore: Prentice-Hill, 1991.

[8] Leighton, R. R. The Apirin/MIGRANES Neural Network Software User's manual Release V6.0 [Machine readable data file]. Russell R. Leighton and the MITRE Coporation, 1992.

[9] Dietterich, T. G. Proper statistical tests for comparing supervised classification learning algorithms. (Technical Report). Department of Computer Science, Oregon State University, Corvallis, OR., 1996

ภาคผนวก

ภาคผนวก ก
ผลการทดลองที่ใช้ไบโพลาร์สำหรับอินพุตโนด

ตารางที่ ก1 ผลการทดลองด้วยวิธีค่าระดับความมั่นใจที่ใช้ไบโพลาร์สำหรับอินพุตโนด

ชื่อชุดข้อมูล	C4.5 Unpruned	C4.5 Pruned	Soft-Pruning	CL (%) of Soft-Pruning Vs C4.5 Unpruned	CL (%) of Soft-Pruning Vs C4.5 Pruned
Anneal	97	95	97	=	+ 75
Soybean	85.64	86.70	90.96	+ 97.5	+ 95
Balance-scale	69.76 ± 3.81	65.77 ± 5.11	85.28 ± 2.29	+ 99	+ 99
Iris	95.33 ± 3.01	94 ± 4.2	94.67 ± 2.07	=	=
Thyroid-disease (allbp)	96.81	97.84	96.60	=	- 95
Thyroid-disease (allhyper)	98.87	98.56	98.97	=	+ 75
Thyroid-disease (allhypo)	99.49	99.49	99.79	+ 84	+ 84
Thyroid-disease (allrep)	98.77	99.07	98.97	=	=
Image	89.43	91	89.90	=	- 84
Restricted (lymphography)	73.7 ± 6.94	78.38 ± 4.83	77.69 ± 9.11	=	=
Restricted (primary-tumor)	42.19± 7.34	41.32 ± 8.11	37.49 ± 5.69	=	=
Statlog (satimage)	84.9	85.45	86.45	+ 90	+ 75
Statlog (segment)	96.97 ± 0.9	96.97 ± 0.9	97.06 ± 0.67	=	=
Statlog (shuttle)	99.97	99.95	99.99	+ 90	+ 97.5
Wine	93.30 ± 4.71	93.30 ± 4.71	93.30 ± 4.71	=	=
Waveform	75.76 ± 1.44	75.82 ± 1.47	79.82 ± 1.37	+ 99	+ 99
Waveform + noise	75.22 ± 1.42	75.30 ± 1.37	80.32 ± 0.49	+ 99	+ 99
Glass	66.34 ± 4.46	66.34 ± 4.46	65.87 ± 4.26	=	=
Led-display-creator	75.4	75.2	75.40	=	=
Led 17	66.4	75.2	67.00	=	- 99

ในงานวิทยานิพนธ์ฉบับนี้ได้ทดลองเพิ่มโดยทำการเปลี่ยนข้อมูลอินพุตเป็นแบบไบโพลาร์ กล่าวคือ หากคุณสมบัติของข้อมูลมีค่าความจริงเป็นเท็จ ให้ค่าของอินพุตโนดนั้นมีค่าเป็น "-1" หากมีค่าความจริงเป็นจริงให้ค่าอินพุตโนดนั้นมีค่าเป็น "1" ผลการทดลองที่ได้จากตารางที่ ก1 เปรียบเทียบกับต้นไม้ตัดสินใจที่ไม่ทำการตัดเล็มดีกว่า 7 ชุดข้อมูล ส่วนชุดข้อมูลที่เหลือมีประสิทธิภาพที่ใกล้เคียงกัน และเมื่อเปรียบเทียบกับต้นไม้ตัดสินใจที่ทำการตัดเล็มแล้ว วิธีของงานวิจัยนี้ดีกว่า 9 ชุดข้อมูล และด้อยกว่า 3 ชุดข้อมูล ซึ่งผลที่ได้เป็นไปในทิศทางเดียวกับการใช้อินพุตเป็นแบบไบนารีตามผลการทดลองในบทที่ 4

ภาคผนวก ข
การใช้งานโปรแกรม

ในงานวิทยานิพนธ์ฉบับนี้ได้พัฒนาโปรแกรมโดยครอบคลุมตั้งแต่การใช้งานต้นไม้ตัดสินใจ รวมไปถึงการใช้งานแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก โดยนำฟังก์ชันต่าง ๆ จากโปรแกรมที่ใช้ คือ C4.5 ที่ใช้สำหรับต้นไม้ตัดสินใจ และ Aspirin/MIGRAINES Neural Network Software [8] ที่ใช้สำหรับแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก โปรแกรมดังกล่าวพัฒนาบนระบบปฏิบัติการ Microsoft Windows 98 โดยใช้ Visual C++ รุ่น 6.0

โมดูล (module) หลักที่ใช้ในการพัฒนาโปรแกรมแบ่งออกเป็น 3 ส่วน คือ

1. ต้นไม้ตัดสินใจ

ในส่วนของฟังก์ชันที่เกี่ยวข้องกับต้นไม้ตัดสินใจ จะทำการสร้างต้นไม้ตัดสินใจ พร้อมทั้งตัดเล็มต้นไม้ ตามขั้นตอนในบทที่ 2 ข้อมูลเข้าในฟังก์ชันนี้มี 3 ส่วนด้วยกัน คือ ข้อมูลสอน ข้อมูลทดสอบและรายละเอียดของคุณสมบัติและกลุ่ม หลังจากสร้างต้นไม้ตัดสินใจที่ไม่ได้ตัดเล็ม และต้นไม้ตัดสินใจที่ทำการตัดเล็มแล้ว รวมไปถึง ค่าความถูกต้องของต้นไม้สองต้น จะเก็บข้อมูลทั้งหมดเป็นแบบข้อความ (text file)

2. การแปลงข้อมูลต้นไม้ตัดสินใจเป็นข้อมูลสำหรับแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก

สำหรับโมดูลนี้ เปรียบได้กับขั้นตอนวิธีในบทที่สาม คือ การสร้างกฎจากต้นไม้ตัดสินใจ การสร้างโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก และ การสร้างข้อมูลสำหรับโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กตามบทที่ 3 โดยนำข้อมูลที่ได้จากโมดูลที่หนึ่งมาใช้ หลังจากเสร็จสิ้นการทำงาน จะได้ข้อมูลสอน และข้อมูลทดสอบที่ได้แปลงข้อมูลแล้ว โดยสามารถนำไปใช้งานกับโมดูลแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กต่อไป

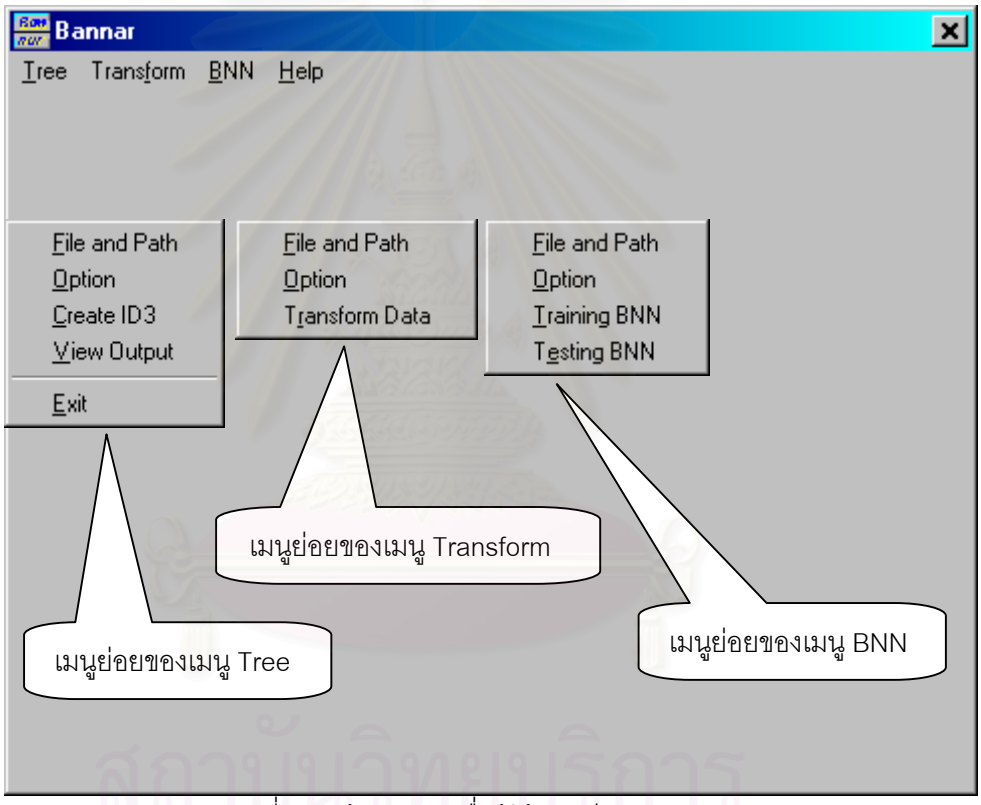
3. แบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก

โมดูลนี้นำข้อมูลสอนมาเข้าสู่กระบวนการแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก ตามขั้นตอนในหัวข้อที่ 2.2.3 โดยเมื่อเสร็จสิ้นการฝึกนิวรอลเน็ตเวิร์กแล้ว จะได้น้ำหนักต่าง ๆ ของโครงสร้างนิวรอลเน็ต-เวิร์ก หลังจากนั้นจะทดสอบด้วยข้อมูลทดสอบเพื่อหาค่าความถูกต้อง และนำค่าความถูกต้องที่ได้ ไปวิเคราะห์ข้อมูลต่อไป

การใช้งานโปรแกรม

เมื่อผู้ใช้งานเรียกโปรแกรมขึ้นมาจะมีเมนูหลักอยู่ 3 เมนูคือ

1. **เมนู** Tree ในส่วนของเมนูนี้จะทำงานที่เกี่ยวกับการสร้างต้นไม้ตัดสินใจ และต้นไม้ตัดสินใจที่ทำการตัดเล็มโดยใช้ค่าความผิดพลาด
2. **เมนู** Transform เป็นเมนูที่ใช้สร้างโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กที่ได้จากต้นไม้ตัดสินใจ รวมถึงการสร้างข้อมูลสอน และข้อมูลทดสอบสำหรับโครงสร้างนิวรอลเน็ตเวิร์กด้วย
3. **เมนู** BNN เมนูนี้ทำงานเกี่ยวกับการฝึกสอนโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก รวมไปถึงการทดสอบค่าความถูกต้องของข้อมูลทดสอบ

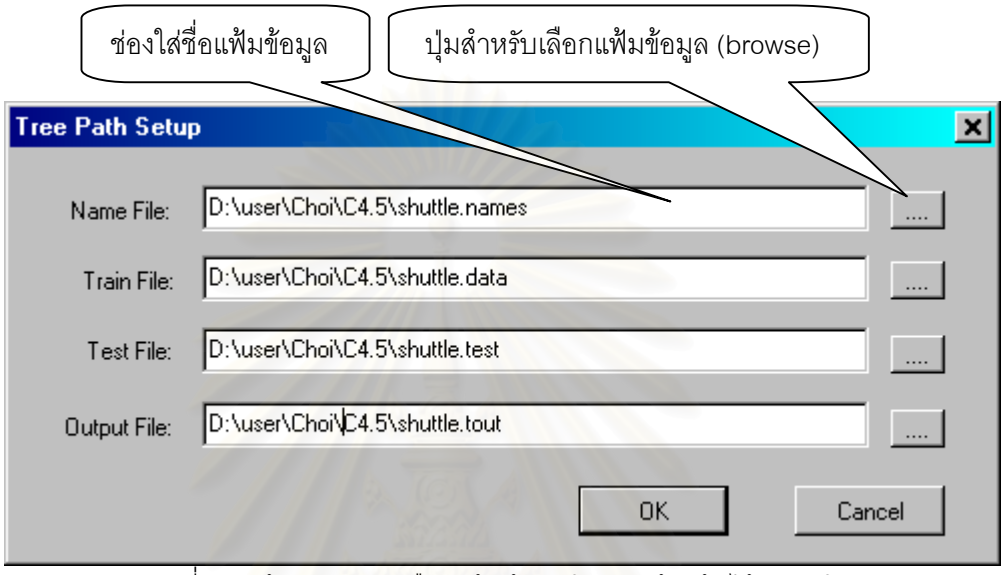


รูปที่ ข1 หน้าจอหลักเมื่อผู้ใช้งานเรียกโปรแกรม

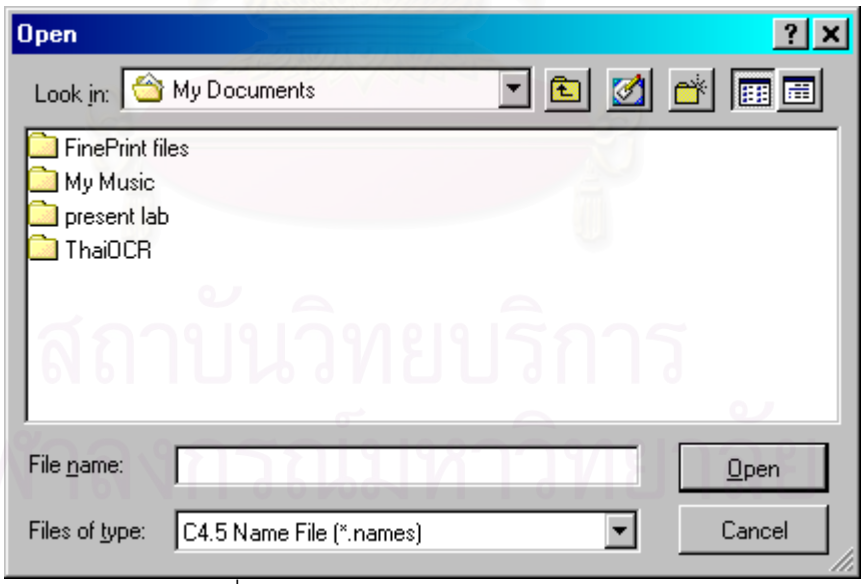
เมนู Tree

ในเมนู Tree นี้มีเมนูย่อยประกอบด้วย

1. **เมนู** File and Path สำหรับเลือกแฟ้มข้อมูลที่ต้องการนำมาสร้างต้นไม้ตัดสินใจ ประกอบด้วยแฟ้มข้อมูลที่เป็นรายละเอียดของชุดข้อมูล (Name File) แฟ้มข้อมูลสอน (Train File) แฟ้มข้อมูลทดสอบ (Test File) และแฟ้มข้อมูลที่เก็บต้นไม้ตัดสินใจ (Output FIle)

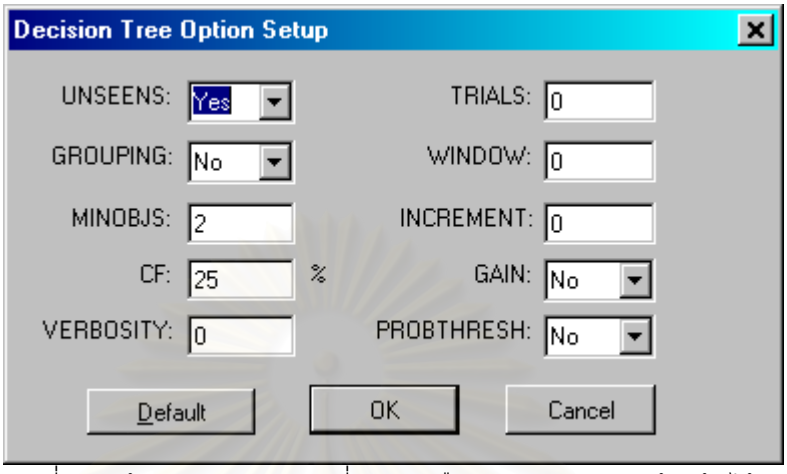


รูปที่ ข2 หน้าจอสำหรับเลือกแฟ้มข้อมูลในการสร้างต้นไม้ตัดสินใจ



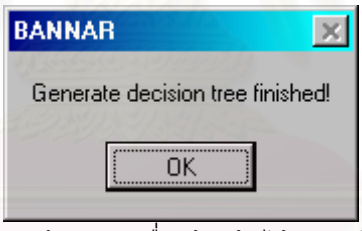
รูปที่ ข3 หน้าจอสำหรับการเลือกแฟ้มข้อมูล

2. **เมนู** Option เป็นทางเลือกต่าง ๆ สำหรับการสร้างต้นไม้ตัดสินใจและต้นไม้ตัดสินใจที่ตัดเล็มแล้ว ทั้งนี้ค่าตั้งต้นที่ให้มาจะเป็นค่าที่ใช้สำหรับงานวิจัยนี้



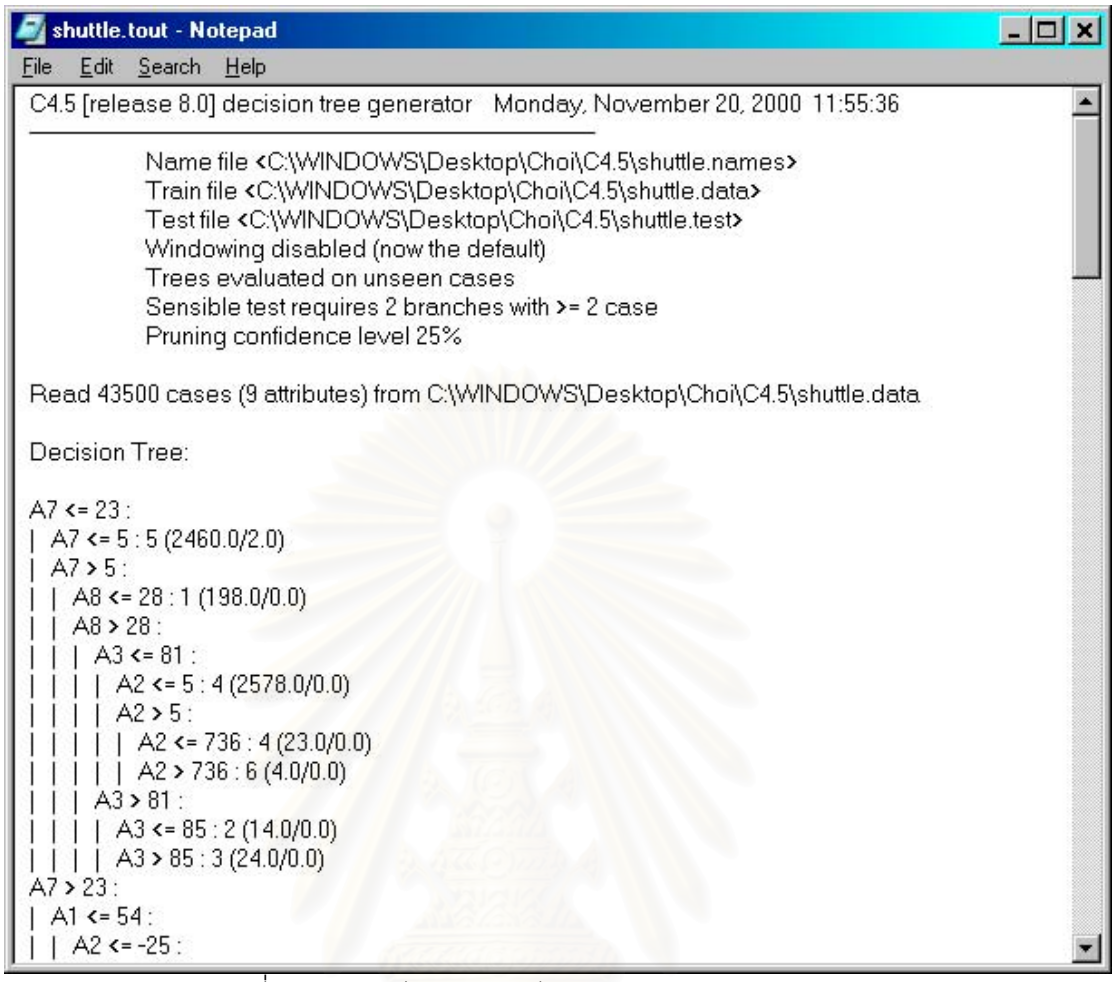
รูปที่ ข4 หน้าจอสำหรับปรับเปลี่ยนทางเลือกต่าง ๆ ของการสร้างต้นไม้

3. **เมนู** Create Tree เป็นเมนูที่ใช้สำหรับสร้างต้นไม้ตัดสินใจที่ได้จากแฟ้มข้อมูลในเมนู File and Path เมื่อสร้างเสร็จจะมีข้อความบอกดังรูปที่ ข5



รูปที่ ข5 ข้อความเมื่อสร้างต้นไม้ตัดสินใจเสร็จ

4. **เมนู** View Output เมื่อสร้างต้นไม้ตัดสินใจเสร็จแล้ว ข้อมูลที่ได้จะเก็บลงแฟ้มข้อมูลที่ระบุไว้ใน เมนู File and Path เราสามารถดูรายละเอียดของต้นไม้ตัดสินใจดังกล่าวได้ โดยเมื่อเลือกเมนู View Output แล้ว จะเรียกใช้งานโปรแกรม notepad สำหรับเปิดแฟ้มข้อมูลที่เก็บต้นไม้ตัดสินใจ ตามรูปที่ ข6
5. **เมนู** Exit ออกจากโปรแกรม

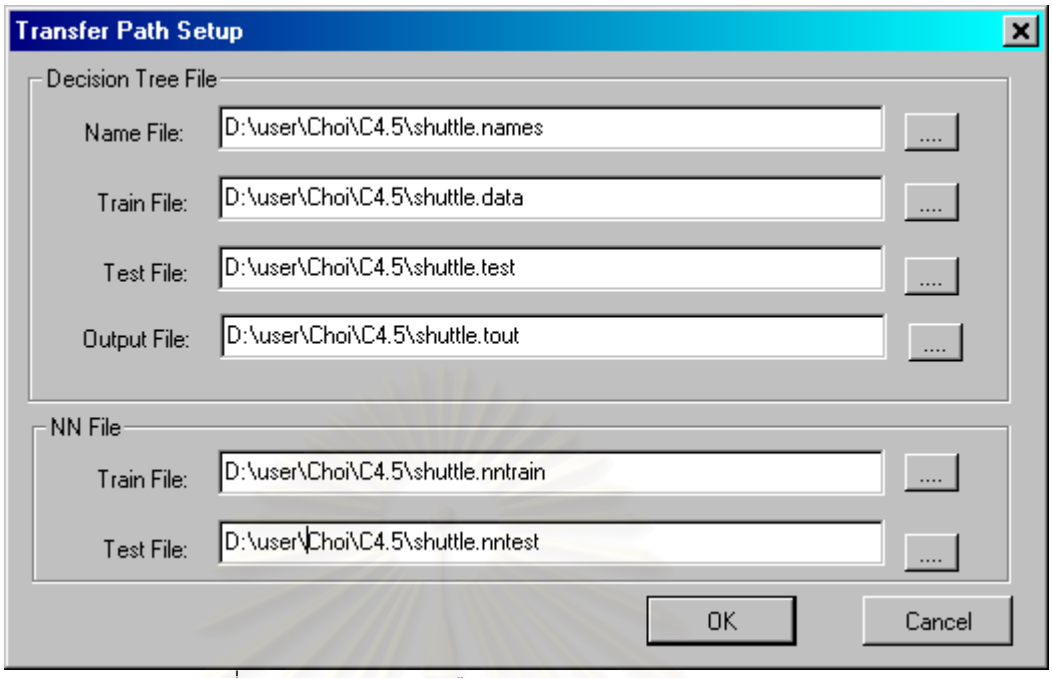


รูปที่ ข6 รายละเอียดข้อมูลต้นไม้ตัดสินใจในโปรแกรม notepad

เมนู Transform

เมนู Transform นี้เป็นเมนูที่ใช้สำหรับการสร้างโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก รวมไปถึงการสร้างข้อมูลสำหรับนิวรอลเน็ตเวิร์ก ประกอบด้วยเมนูย่อยดังนี้

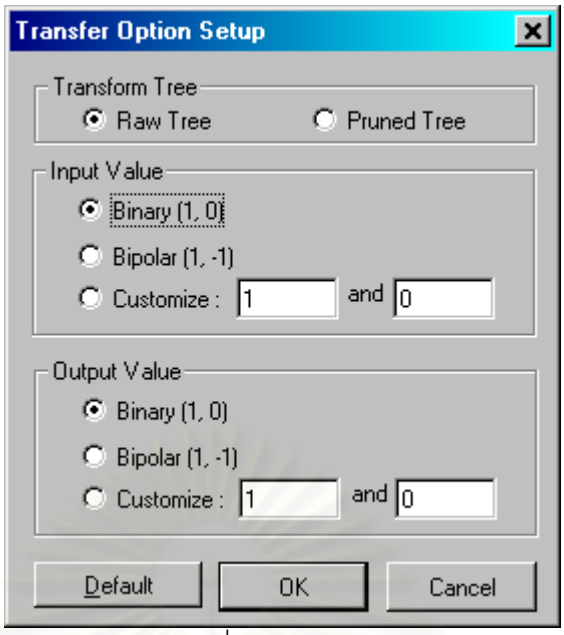
1. **เมนู** File and Path ในเมนูย่อยนี้ จะเลือกแฟ้มข้อมูลที่ใช้สำหรับการสร้างโครงสร้างนิวรอลเน็ตเวิร์ก รวมทั้งข้อมูลสอน และข้อมูลทดสอบของนิวรอลเน็ตเวิร์ก โดยแฟ้มข้อมูลที่เกี่ยวข้องคือ แฟ้มข้อมูลที่เป็นรายละเอียดของชุดข้อมูล (Name File) แฟ้มข้อมูลสอน (Train File) แฟ้มข้อมูลทดสอบ (Test File) และแฟ้มข้อมูลที่เก็บต้นไม้ตัดสินใจ (Output File) ส่วนแฟ้มข้อมูลที่ได้จะเป็นข้อมูลสอน และข้อมูลทดสอบของโครงสร้างแบ็กพรอพาเก-ชันนิวรอลเน็ตเวิร์ก



รูปที่ ข7 หน้าจอสำหรับเลือกแฟ้มข้อมูลในการแปลงข้อมูล

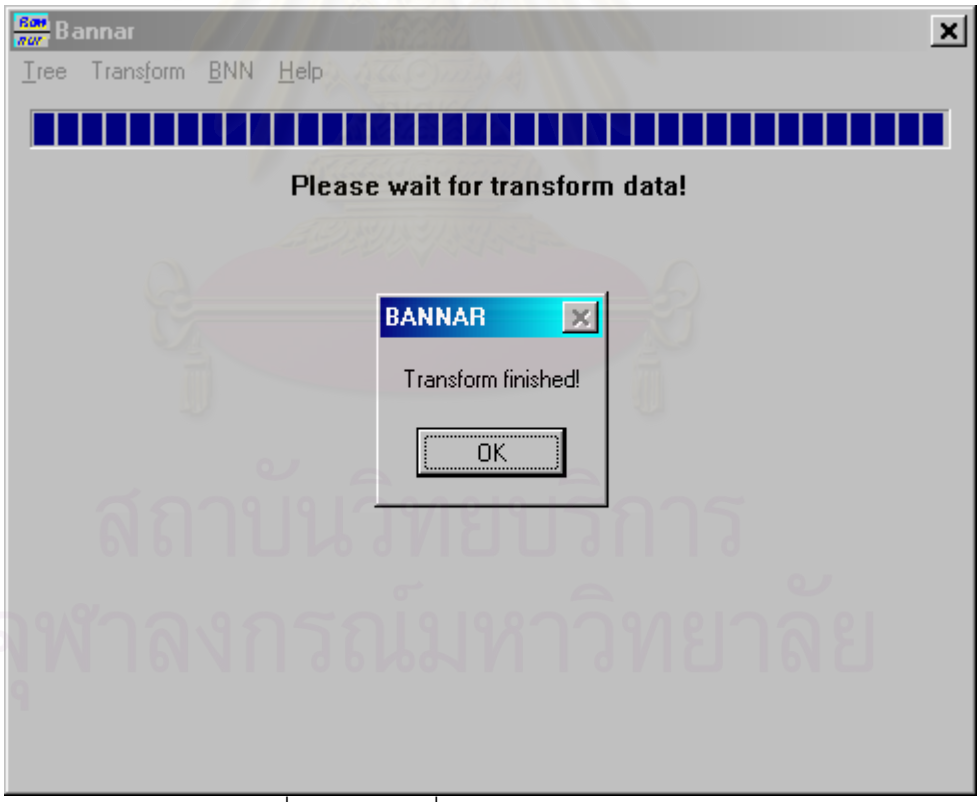
2. **เมนู** Option เป็นทางเลือกต่าง ๆ สำหรับการแปลงข้อมูลจากต้นไม้ตัดสินใจ เป็นข้อมูลที่ใช้กับโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก ทั้งนี้แบ่งออกเป็น 3 ส่วนคือ

- Transform Tree เป็นส่วนที่ใช้สำหรับการเลือกว่าจะใช้ต้นไม้ตัดสินใจที่ยังไม่ตัดเล็ม (Raw Tree) หรือต้นไม้ตัดสินใจที่ทำการตัดเล็มแล้ว (Pruned Tree) ในการสร้างโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก และข้อมูลสอนและข้อมูลทดสอบ
- Input Value เป็นทางเลือกที่ใช้สำหรับการสร้างข้อมูลในส่วนอินพุตโนดของแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก สามารถเลือกเป็นไบนารี ไบโพลาร์ หรือทำการกำหนดเองก็ได้ ทั้งนี้ค่าตั้งต้นจะเป็นไบนารี
- Output Value เป็นทางเลือกที่ใช้สำหรับการสร้างข้อมูลในส่วนเอาต์พุตโนด โดยตัวเลือกจะเป็นเช่นเดียวกันกับ Input Value



รูปที่ ข8 หน้าจอสำหรับปรับเปลี่ยนทางเลือกต่าง ๆ ของการแปลงข้อมูล

3. **เมนู** Transform Data เป็นเมนูที่ใช้ในการแปลงข้อมูล เมื่อแปลงเสร็จสิ้นจะได้ดังรูปที่ ข9

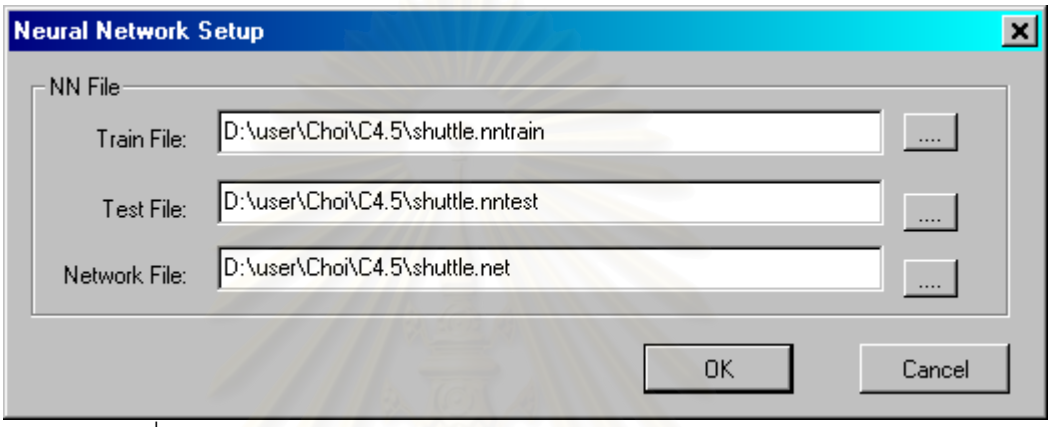


รูปที่ ข9 หน้าจอเมื่อแปลงข้อมูลเสร็จสมบูรณ์

เมนู BNN

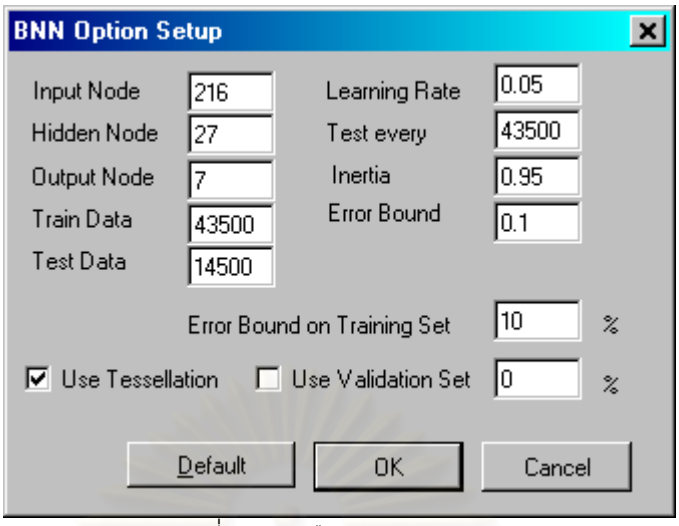
ประกอบด้วยเมนูย่อยดังนี้

1. **เมนู** File and Path สำหรับเมนูย่อยนี้ เป็นการเลือกแฟ้มข้อมูลที่ใช้กับแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก ประกอบไปด้วย แฟ้มข้อมูลสำหรับข้อมูลสอน แฟ้มข้อมูลสำหรับข้อมูลทดสอบ และแฟ้มข้อมูลสำหรับเก็บน้ำหนักต่าง ๆ ของแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก (Network File)



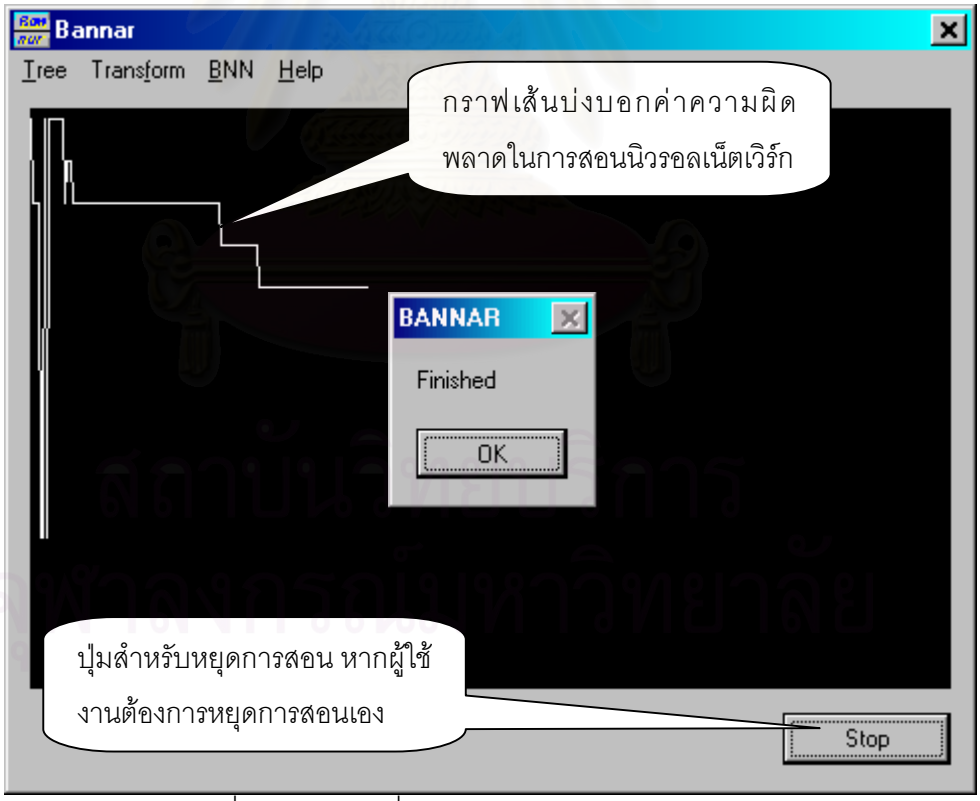
รูปที่ ข10 หน้าจอสำหรับเลือกแฟ้มข้อมูลในแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก

2. **เมนู** Option เป็นทางเลือกต่าง ๆ สำหรับแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก ทั้งนี้หากผู้ใช้งานแปลงข้อมูลจากเมนู Transform มาก่อน ข้อมูลต่าง ๆ คือจำนวนอินพุตโนด (Input Node) จำนวนฮิดเดนโนด (Hidden Node) จำนวนเอาต์พุตโนด (Output Node) จำนวนข้อมูลสอน (Training Data) และจำนวนข้อมูลทดสอบ (Test Data) จะใส่ค่าให้อัตโนมัติ หากผู้ใช้งานไม่ได้ผ่านโมดูลแปลงข้อมูลมาก่อน จะต้องใส่ข้อมูลข้างต้นเอง สำหรับค่าต่าง ๆ ที่เหลือเป็นค่าที่เกี่ยวข้องกับการสอนนิวรอลเน็ตเวิร์ก ซึ่งค่าตั้งต้นที่ให้ไว้เป็นค่าที่ใช้สำหรับงานวิจัยนี้



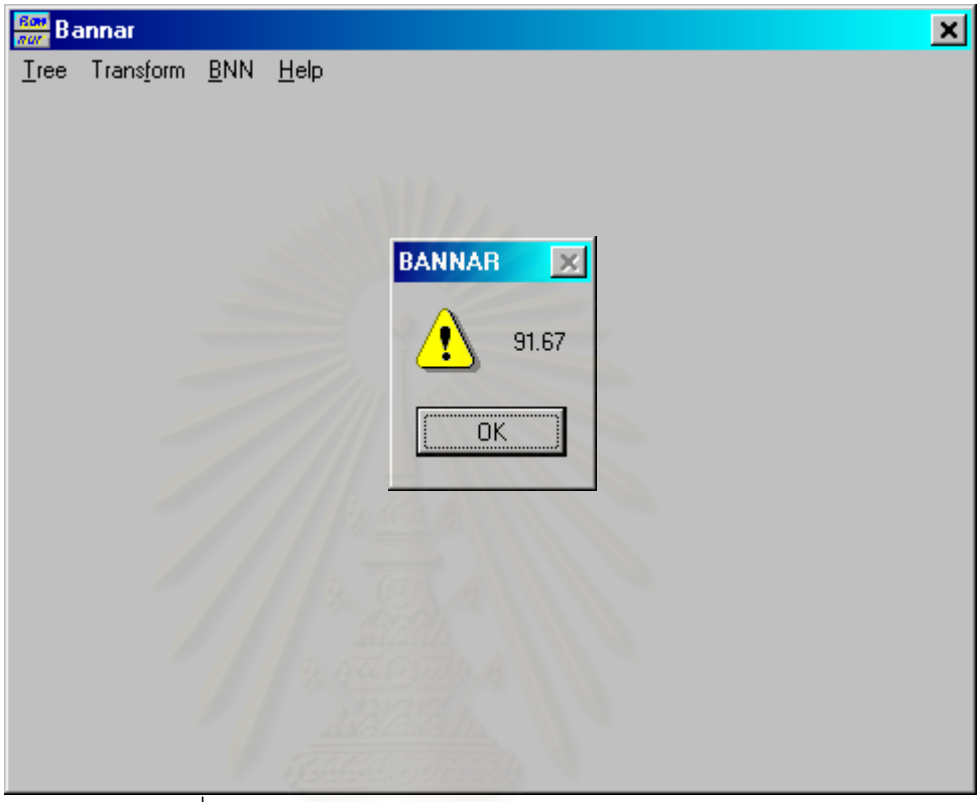
รูปที่ ข11 หน้าจอสำหรับปรับเปลี่ยนทางเลือกต่าง ๆ ของแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก

3. **เมนู** Training BNN เป็นเมนูที่ใช้สำหรับสอนแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก ตามเงื่อนไขต่าง ๆ จากเมนู Option โดยค่าน้ำหนักต่าง ๆ ของโครงสร้างนิวรอลเน็ตเวิร์ก จะถูกเก็บลงแฟ้มข้อมูล เมื่อสอนนิวรอลเน็ตเวิร์กเสร็จแล้ว จะขึ้นข้อความบอกดังรูปที่ ข12



รูปที่ ข12 หน้าจอเมื่อสอนนิวรอลเน็ตเวิร์กเสร็จสมบูรณ์

4. **เมนู** Testing BNN สำหรับเมนูนี้จะทดสอบโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก โดยข้อมูลทดสอบ เมื่อทดสอบเสร็จแล้วจะขึ้นข้อความที่บอกถึงเปอร์เซ็นต์ความถูกต้องที่ได้กับข้อมูลทดสอบดังรูปที่ ข13



รูปที่ ข13 หน้าจอแสดงค่าความถูกต้องของข้อมูลทดสอบ

ภาคผนวก ค
รายละเอียดข้อมูลที่ใช้ในการทดลอง

ข้อมูลต่าง ๆ ที่ใช้ในการทดลองของวิทยานิพนธ์ฉบับนี้ได้เลือกมาจาก Department of Information and Computer Science, University of California, Irvine [2] โดยมีทั้งหมด 20 ชุดข้อมูล ค่าที่เป็นไปได้ของคุณสมบัติต่าง ๆ หากเป็นค่าต่อเนื่องจะแทนด้วยคำว่า “continuous”

ชื่อชุดข้อมูล Anneal

จำนวนคุณสมบัติ 38

จำนวนกลุ่ม 6 กลุ่มคือ 1,2,3,4,5,U

ชื่อคุณสมบัติ	ค่าที่เป็นไปได้	ชื่อคุณสมบัติ	ค่าที่เป็นไปได้
family	--,GB,GK,GS,TN,ZA,ZF,ZH,ZM,ZS	chrom	C,-
product-type	C, H, G	phos	P,-
steel	-,R,A,U,K,M,S,W,V	cbond	Y,-
carbon	continuous	marvi	Y,-
hardness	continuous	exptl	Y,-
temper_rolling	-,T	ferro	Y,-
condition	-,S,A,X	corr	Y,-
formability	-,1,2,3,4,5	blue/bright/varn/clean	B,R,V,C,-
strength	continuous	lustre	Y,-
non-ageing	-,N	jurofm	Y,-
surface-finish	P,M,-	s	Y,-
surface-quality	-,D,E,F,G	p	Y,-
enamelability	-,1,2,3,4,5	shape	COIL, SHEET
bc	Y,-	thick	continuous
bf	Y,-	width	continuous
bt	Y,-	len	continuous
bw/me	B,M,-	oil	-,Y,N
bl	Y,-	bore	0000,0500,0600,0760
m	Y,-	packing	-,1,2,3

ชื่อชุดข้อมูล Soybean

จำนวนคุณสมบัติ 35

จำนวนกลุ่ม 19 กลุ่มคือ diaporthe-stem-canker, charcoal-rot, rhizoctonia-root-rot,phytophthora-rot, brown-stem-rot, powdery-mildew, downy-mildew, brown-spot, bacterial-blight, bacterial-pustule, purple-seed-stain, anthracnose, phyllosticta-leaf-spot, alternarialeaf-spot, frog-eye-leaf-spot, diaporthe-pod-&-stem-blight, cyst-nematode, 2-4-d-injury, herbicide-injury

ชื่อคุณสมบัติ	ค่าที่เป็นไปได้	ชื่อคุณสมบัติ	ค่าที่เป็นไปได้
date	0,1,2,3,4,5,6	stem	0,1
plant-stand	0,1	lodging	0,1
precip	0,1,2	stem-cankers	0,1,2,3
temp	0,1,2	canker-lesion	0,1,2,3
hail	0,1	fruiting-bodies	0,1
crop-hist	0,1,2,3	external-decay	0,1,2
area-damaged	0,1,2,3	mycelium	0,1
severity	0,1,2	int-discolor	0,1,2
seed-tmt	0,1,2	sclerotia	0,1
germination	0,1,2	fruit-pods	0,1,2,3
plant-growth	0,1	fruit-spots	0,1,2,3,4
leaves	0,1	seed	0,1
leafspots-halo	0,1,2	mold-growth	0,1
leafspots-marg	0,1,2	seed-discolor	0,1
leafspot-size	0,1,2	seed-size	0,1
leaf-shread	0,1	shriveling	0,1
leaf-malf	0,1	roots	0,1,2
leaf-mild	0,1,2		

ชื่อชุดข้อมูล Balance-scale

จำนวนคุณสมบัติ 4

จำนวนกลุ่ม 3 กลุ่มคือ L,B,R

ชื่อคุณสมบัติ	ค่าที่เป็นไปได้	ชื่อคุณสมบัติ	ค่าที่เป็นไปได้
Left-Weight	1,2,3,4,5	Right-Weight	1,2,3,4,5
Left-Distance	1,2,3,4,5	Right-Distance	1,2,3,4,5

ชื่อชุดข้อมูล Iris

จำนวนคุณสมบัติ 4

จำนวนกลุ่ม 3 กลุ่มคือ Iris-setosa, Iris-versicolor, Iris-virginica

ชื่อคุณสมบัติ	ค่าที่เป็นไปได้	ชื่อคุณสมบัติ	ค่าที่เป็นไปได้
sepal-length	continuous	petal-length	continuous
sepal-width	continuous	petal-width	continuous

ชื่อชุดข้อมูล allbp

จำนวนคุณสมบัติ 29

จำนวนกลุ่ม 3 กลุ่ม คือ increased binding protein, decreased binding protein,negative

ชื่อคุณสมบัติ	ค่าที่เป็นไปได้	ชื่อคุณสมบัติ	ค่าที่เป็นไปได้
age	continuous	psych	f, t
sex	M, F	TSH measured	f, t
on thyroxine	f, t	TSH	continuous
query on thyroxine	f, t	T3 measured	f, t
on antithyroid medication	f, t	T3	continuous
sick	f, t	TT4 measured	f, t
pregnant	f, t	TT4	continuous
thyroid surgery	f, t	T4U measured	f, t
I131 treatment	f, t	T4U	continuous
query hypothyroid	f, t	FTI measured	f, t
query hyperthyroid	f, t	FTI	continuous
lithium	f, t	TBG measured	f, t
goitre	f, t	TBG	continuous
tumor	f, t	referral source	WEST, STMW, SVHC, SVI, SVHD, other
hypopituitary	f, t		

ชื่อชุดข้อมูล allhyper

จำนวนคุณสมบัติ 29

จำนวนกลุ่ม 5 กลุ่ม คือ hyperthyroid, T3 toxic, goitre, secondary toxic,negative

ชื่อคุณสมบัติ	ค่าที่เป็นไปได้	ชื่อคุณสมบัติ	ค่าที่เป็นไปได้
age	continuous	psych	f, t
sex	M, F	TSH measured	f, t
on thyroxine	f, t	TSH	continuous
query on thyroxine	f, t	T3 measured	f, t
on antithyroid medication	f, t	T3	continuous
sick	f, t	TT4 measured	f, t
pregnant	f, t	TT4	continuous
thyroid surgery	f, t	T4U measured	f, t
I131 treatment	f, t	T4U	continuous
query hypothyroid	f, t	FTI measured	f, t
query hyperthyroid	f, t	FTI	continuous
lithium	f, t	TBG measured	f, t
goitre	f, t	TBG	continuous
tumor	f, t	referral source	WEST, STMW, SVHC, SVI, SVHD, other
hypopituitary	f, t		

ชื่อชุดข้อมูล allhypo

จำนวนคุณสมบัติ 29

จำนวนกลุ่ม 5 กลุ่ม คือ hypothyroid, primary hypothyroid, compensated hypothyroid,secondary hypothyroid,negative

ชื่อคุณสมบัติ	ค่าที่เป็นไปได้	ชื่อคุณสมบัติ	ค่าที่เป็นไปได้
age	continuous	psych	f, t
sex	M, F	TSH measured	f, t
on thyroxine	f, t	TSH	continuous
query on thyroxine	f, t	T3 measured	f, t
on antithyroid medication	f, t	T3	continuous
sick	f, t	TT4 measured	f, t
pregnant	f, t	TT4	continuous
thyroid surgery	f, t	T4U measured	f, t
I131 treatment	f, t	T4U	continuous
query hypothyroid	f, t	FTI measured	f, t
query hyperthyroid	f, t	FTI	continuous
lithium	f, t	TBG measured	f, t
goitre	f, t	TBG	continuous
tumor	f, t	referral source	WEST, STMW, SVHC, SVI, SVHD, other
hypopituitary	f, t		

ชื่อชุดข้อมูล allrep

จำนวนคุณสมบัติ 29

จำนวนกลุ่ม 4 กลุ่ม คือ replacement therapy, underreplacement, overreplacement,negative

ชื่อคุณสมบัติ	ค่าที่เป็นไปได้	ชื่อคุณสมบัติ	ค่าที่เป็นไปได้
age	continuous	psych	f, t
sex	M, F	TSH measured	f, t
on thyroxine	f, t	TSH	continuous
query on thyroxine	f, t	T3 measured	f, t
on antithyroid medication	f, t	T3	continuous
sick	f, t	TT4 measured	f, t
pregnant	f, t	TT4	continuous
thyroid surgery	f, t	T4U measured	f, t
I131 treatment	f, t	T4U	continuous
query hypothyroid	f, t	FTI measured	f, t
query hyperthyroid	f, t	FTI	continuous
lithium	f, t	TBG measured	f, t
goitre	f, t	TBG	continuous
tumor	f, t	referral source	WEST, STMW, SVHC, SVI, SVHD, other
hypopituitary	f, t		

ชื่อชุดข้อมูล Image

จำนวนคุณสมบัติ 19

จำนวนกลุ่ม 7 กลุ่มคือ BRICKFACE, SKY, FOLIAGE, CEMENT, WINDOW, PATH, GRASS

ชื่อคุณสมบัติ	ค่าที่เป็นไปได้	ชื่อคุณสมบัติ	ค่าที่เป็นไปได้
region-centroid-col	continuous	rawred-mean	continuous
region-centroid-row	continuous	rawblue-mean	continuous
region-pixel-count	continuous	rawgreen-mean	continuous
short-line-density-5	continuous	exred-mean	continuous
short-line-density-2	continuous	exblue-mean	continuous
vedge-mean	continuous	exgreen-mean	continuous
vegde-sd	continuous	value-mean	continuous
hedge-mean	continuous	saturatoin-mean	continuous
hedge-sd	continuous	hue-mean	continuous
intensity-mean	continuous		

ชื่อชุดข้อมูล lymphography

จำนวนคุณสมบัติ 18

จำนวนกลุ่ม 4 กลุ่มคือ 1, 2, 3, 4

ชื่อคุณสมบัติ	ค่าที่เป็นไปได้	ชื่อคุณสมบัติ	ค่าที่เป็นไปได้
lymphatics	1,2,3,4	lym_nodes_enlar	1,2,3,4
block_of_affere	1,2	changes_in_lym	1,2,3
bl_of_lymph_c	1,2	defect_in_node	1,2,3,4
bl_of_lymph_s	1,2	changes_in_node	1,2,3,4
by_pass	1,2	changes_in_stru	1,2,3,4,5,6,7,8
extravasates	1,2	special_forms	1,2,3
regenerationof	1,2	dislocation_of	1,2
early_uptake_in	1,2	exclusion_of_no	1,2
lym_nodes_dimin	1,2,3	no_of_nodes_in	1,2,3,4,5,6,7,8

ชื่อชุดข้อมูล primary-tumor

จำนวนคุณสมบัติ 17

จำนวนกลุ่ม 22 กลุ่มคือ 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22

ชื่อคุณสมบัติ	ค่าที่เป็นไปได้	ชื่อคุณสมบัติ	ค่าที่เป็นไปได้
age	1,2,3	liver	1,2
sex	1,2	brain	1,2
histologic-type	1,2,3	skin	1,2
degree-of-diffe	1,2,3	neck	1,2
bone	1,2	supraclavicular	1,2
bone-marrow	1,2	axillar	1,2
lung	1,2	mediastinum	1,2
pleura	1,2	abdominal	1,2
peritoneum	1,2		

ชื่อชุดข้อมูล satimage

จำนวนคุณสมบัติ 36

จำนวนกลุ่ม 6 กลุ่มคือ 1, 2, 3, 4, 5, 7

ชื่อคุณสมบัติ	ค่าที่เป็นไปได้	ชื่อคุณสมบัติ	ค่าที่เป็นไปได้
A1	continuous	A19	continuous
A2	continuous	A20	continuous
A3	continuous	A21	continuous
A4	continuous	A22	continuous
A5	continuous	A23	continuous
A6	continuous	A24	continuous
A7	continuous	A25	continuous
A8	continuous	A26	continuous
A9	continuous	A27	continuous
A10	continuous	A28	continuous
A11	continuous	A29	continuous
A12	continuous	A30	continuous
A13	continuous	A31	continuous
A14	continuous	A32	continuous
A15	continuous	A33	continuous
A16	continuous	A34	continuous
A17	continuous	A35	continuous
A18	continuous	A36	continuous

ชื่อชุดข้อมูล segment

จำนวนคุณสมบัติ 19

จำนวนกลุ่ม 7 กลุ่มคือ 1, 2, 3, 4, 5, 6, 7

ชื่อคุณสมบัติ	ค่าที่เป็นไปได้	ชื่อคุณสมบัติ	ค่าที่เป็นไปได้
A1	continuous	A11	continuous
A2	continuous	A12	continuous
A3	continuous	A13	continuous
A4	continuous	A14	continuous
A5	continuous	A15	continuous
A6	continuous	A16	continuous
A7	continuous	A17	continuous
A8	continuous	A18	continuous
A9	continuous	A19	continuous
A10	continuous		

ชื่อชุดข้อมูล shuttle

จำนวนคุณสมบัติ 9

จำนวนกลุ่ม 7 กลุ่มคือ 1, 2, 3, 4, 5, 6, 7

ชื่อคุณสมบัติ	ค่าที่เป็นไปได้	ชื่อคุณสมบัติ	ค่าที่เป็นไปได้
A1	continuous	A6	continuous
A2	continuous	A7	continuous
A3	continuous	A8	continuous
A4	continuous	A9	continuous
A5	continuous		

ชื่อชุดข้อมูล Wine

จำนวนคุณสมบัติ 13

จำนวนกลุ่ม 3 กลุ่มคือ 1, 2, 3

ชื่อคุณสมบัติ	ค่าที่เป็นไปได้	ชื่อคุณสมบัติ	ค่าที่เป็นไปได้
A1	continuous	A8	continuous
A2	continuous	A9	continuous
A3	continuous	A10	continuous
A4	continuous	A11	continuous
A5	continuous	A12	continuous
A6	continuous	A13	continuous
A7	continuous		

ชื่อชุดข้อมูล Waveform

จำนวนคุณสมบัติ 21

จำนวนกลุ่ม 3 กลุ่มคือ 0, 1, 2

ชื่อคุณสมบัติ	ค่าที่เป็นไปได้	ชื่อคุณสมบัติ	ค่าที่เป็นไปได้
A1	continuous	A12	continuous
A2	continuous	A13	continuous
A3	continuous	A14	continuous
A4	continuous	A15	continuous
A5	continuous	A16	continuous
A6	continuous	A17	continuous
A7	continuous	A18	continuous
A8	continuous	A19	continuous
A9	continuous	A20	continuous
A10	continuous	A21	continuous
A11	continuous		

ชื่อชุดข้อมูล Waveform + noise

จำนวนคุณสมบัติ 40

จำนวนกลุ่ม 3 คือ 0, 1, 2

ชื่อคุณสมบัติ	ค่าที่เป็นไปได้	ชื่อคุณสมบัติ	ค่าที่เป็นไปได้
A1	continuous	A21	continuous
A2	continuous	A22	continuous
A3	continuous	A23	continuous
A4	continuous	A24	continuous
A5	continuous	A25	continuous
A6	continuous	A26	continuous
A7	continuous	A27	continuous
A8	continuous	A28	continuous
A9	continuous	A29	continuous
A10	continuous	A30	continuous
A11	continuous	A31	continuous
A12	continuous	A32	continuous
A13	continuous	A33	continuous
A14	continuous	A34	continuous
A15	continuous	A35	continuous
A16	continuous	A36	continuous
A17	continuous	A37	continuous
A18	continuous	A38	continuous
A19	continuous	A39	continuous
A20	continuous	A40	continuous

ชื่อชุดข้อมูล Glass

จำนวนคุณสมบัติ 9

จำนวนกลุ่ม 6 กลุ่มคือ 1, 2, 3, 5, 6, 7

ชื่อคุณสมบัติ	ค่าที่เป็นไปได้	ชื่อคุณสมบัติ	ค่าที่เป็นไปได้
RI	continuous	K	continuous
Na	continuous	Ca	continuous
Mg	continuous	Ba	continuous
Al	continuous	Fe	continuous
Si	continuous		

ชื่อชุดข้อมูล Led-display-creator

จำนวนคุณสมบัติ 7

จำนวนกลุ่ม 10 กลุ่มคือ 0, 1, 2, 3, 4, 5, 6, 7, 8, 9

ชื่อคุณสมบัติ	ค่าที่เป็นไปได้	ชื่อคุณสมบัติ	ค่าที่เป็นไปได้
A1	continuous	A5	continuous
A2	continuous	A6	continuous
A3	continuous	A7	continuous
A4	continuous		

ชื่อชุดข้อมูล Led 17

จำนวนคุณสมบัติ 24

จำนวนกลุ่ม 10 คือ 0, 1, 2, 3, 4, 5, 6, 7, 8, 9

ชื่อคุณสมบัติ	ค่าที่เป็นไปได้	ชื่อคุณสมบัติ	ค่าที่เป็นไปได้
A1	continuous	A13	continuous
A2	continuous	A14	continuous
A3	continuous	A15	continuous
A4	continuous	A16	continuous
A5	continuous	A17	continuous
A6	continuous	A18	continuous
A7	continuous	A19	continuous
A8	continuous	A20	continuous
A9	continuous	A21	continuous
A10	continuous	A22	continuous
A11	continuous	A23	continuous
A12	continuous	A24	continuous

ภาคผนวก ง

ตารางแจกแจงแบบ t (t Distribution)

ตารางแจกแจงแบบ t นี้เป็นตารางที่ใช้ในการหาระดับความมั่นใจในหัวข้อที่ 2.2.4 ซึ่งในงานวิทยานิพนธ์ฉบับนี้ใช้ degrees of freedom (df) มีค่าเท่ากับ 5

df	ระดับความมั่นใจที่ N (%)											
	50	60	70	80	90	95	96	98	99	99.5	99.8	99.9
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.663.	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.15	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.295	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
inf.	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291

ประวัติผู้เขียน

นายก้องศักดิ์ จงเกษมวงศ์ เกิดเมื่อวันที่ 16 กรกฎาคม 2519 ที่จังหวัดกรุงเทพมหานคร สำเร็จการศึกษาหลักสูตรวิทยาศาสตรบัณฑิต (วท.บ.) สาขาวิชาวิทยาการคอมพิวเตอร์ จากภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย เมื่อปีการศึกษา 2540 และเข้าศึกษาต่อหลักสูตรวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2541