บทที่ 3

การเรียนรู้โดยการสร้างต้นไม้ตัดสินใจ (Decision Tree Learning)

ในระบบปัญญาประดิษฐ์โดยทั่วไป ยังไม่มีความสามารถในการเรียนรู้หรือมีอยู่เพียงเล็กน้อย เนื่องจากข้อจำกัดของระบบ ความรู้ทั้งหมดจะถูกโปรแกรมหรือใส่ไว้ในระบบก่อนที่ระบบจะทำงาน ดังนั้นถ้าความรู้ที่ใส่ให้มีข้อบกพร่องหรือผิดพลาด งานหรือการตัดสินใจก็จะผิดพลาดไปด้วย ระบบไม่สามารถที่จะแก้ไขปรับปรุงข้อบกพร่องหรือแก้ไขความรู้ที่ผิดพลาดได้ด้วยตัวเอง ต้องมีการโปรแกรมหรือใส่ความรู้ใหม่ที่ถูกต้องเข้าไประบบจึงจะสามารถทำงานได้ถูกต้อง ระบบไม่สามารถที่จะปรับปรุงเพิ่มเติมความรู้ จากประสบการณ์หรือตัวอย่างที่มีได้ ไม่สามารถที่จะสร้างขั้นตอนการทำงาน (algorithm) ข้อสรุปหรือวิธีการแก้ปัญหาใหม่ได้ด้วยตัวเอง หรือกล่าวได้ว่าระบบปัญญาประดิษฐ์โดยทั่วไป ยังขาดความสามารถทาง การอนุมานแบบอุปนัย (inductive inference) ซึ่งเป็นการอนุมานความรู้จากสารสนเทศหรือข้อเท็จจริงที่ป้อนให้ เพื่อสรุปออกมาเป็นกฎหรือความรู้ใหม่ขึ้นมา แต่ระบบปัญญาประดิษฐ์ส่วนมากจะเป็น การอนุมานแบบนิรนัย (deductive inference) ซึ่งเป็นการแก้ปัญหาโดยใช้ความรู้หรือกฎเกณฑ์ที่มีอยู่แล้ว ไม่สามารถที่จะคิดหรือสร้างความรู้ใหม่ได้ด้วยตัวเอง (Michalski, 1986)

ระบบปัญญาประดิษฐ์จำเป็นต้องมีฐานความรู้ (knowledge base) ซึ่งบรรจุไปด้วยความจริง หรือกฎที่เฉพาะเกี่ยวกับงานที่สนใจหรือต้องการแก้ปัญหา และอาจจะประกอบด้วยสามัญสำนึก (common sense heuristics) ซึ่งเป็นทฤษฎีหรือความรู้ทั่วไป

ในระบบงานใหม่ที่ยังไม่มีฐานความรู้อยู่ การสร้างฐานความรู้เป็นเรื่องที่ยุ่งยากซับซ้อน ใช้เวลานาน ต้องการผู้เชี่ยวชาญเฉพาะด้าน และฐานความรู้อาจจะมีข้อผิดพลาดแฝงอยู่ อีกทั้งสารสนเทศหรือความรู้ในปัจจุบันมีปริมาณมาก การที่จะเก็บ จัดเรียง ค้นหาสารสนเทศหรือความรู้ที่ต้องการเป็นเรื่องยาก จึงมีการคิดที่จะนำเอาสารสนเทศหรือความรู้เก็บเข้าไปในเครื่องคอมพิวเตอร์ในรูปของฐานความรอบรู้ ซึ่งจะสะดวกต่อการนำไปใช้ โดยมีโปรแกรมที่มีความสามารถในการแปลงข้อมูลเข้าไปเก็บในฐานความรอบรู้ได้อย่างอัตโนมัติ

จากปัญหาที่กล่าวมาข้างต้น เราสามารถที่จะใช้เทคนิควิธีการทางการเรียนรู้ของเครื่อง (Machine Learning) เข้ามาช่วยแก้ปัญหาได้ โดยการสร้างระบบที่มีความสามารถทางการเรียนรู้ สามารถที่จะคิดวิเคราะห์ สร้างกฎในการตัดสินใจ หรือสรุปความรู้ขึ้นมาได้เองจาก ตัวอย่าง ความจริง กฎเกณฑ์ ข้อมูลหรือสารสนเทศที่ให้แก่ระบบได้ แล้วสร้างเป็นฐานความรอบรู้ของตัวเอง เพื่อใช้คิดแก้ปัญหาต่อไป ระบบที่มีความสามารถเช่นนี้ จะเป็นระบบที่ยืดหยุ่น สามารถปรับปรุงข้อผิดพลาดของตัวเองได้ อีกทั้งยังสามารถถ่ายทอดการเรียนรู้หรือความรู้ไปยังอีกเครื่องหนึ่งได้โดยง่าย

วิธีการเรียนรู้แบบต่าง ๆ (Michalski, Carbonell and Mitchell, 1983)

วิธีการเรียนรู้สามารถแบ่งได้หลายวิธี ขึ้นอยู่กับหลักที่ใช้ในการแบ่ง แต่ในที่นี้จะแบ่งวิธีการเรียนรู้โดยใช้ระดับของการอนุมาน (Inferencing) ที่ต้องการของทั้งคอมพิวเตอร์และผู้สอน โดยเรียงจากระดับที่ผู้สอนเป็นคนอนุมาน จนถึงระดับที่เครื่องสามารถอนุมานเองได้ แบ่งออกได้ดังนี้ คือ

1. การเรียนรู้โดยการท่องจำและการปลูกฝังความรู้ใหม่โดยตรง (Rote learning and direct implanting of new knowledge) จะเป็นการใช้ความรู้ใหม่โดยตรง ซึ่งคอมพิวเตอร์ไม่มีการอนุมานหรือเปลี่ยนรูปของความรู้เลย ได้แก่

1.1 การเรียนรู้โดยการโปรแกรม (Learning by being programmed) ความรู้จะถูกวิเคราะห์ สร้างและแก้ไขจากภายนอกโดยผู้สอน ก่อนที่จะป้อนเข้าสู่ระบบคอมพิวเตอร์ เหมือนกับการเขียนโปรแกรมคอมพิวเตอร์

1.2 การเรียนรู้โดยการจำ (Learning by memorization) เป็นการให้ข้อมูลหรือความจริงที่ไม่ต้องอนุมานแก่ระบบ เหมือนการป้อนข้อมูลเข้าสู่ระบบฐานข้อมูลแล้วเรียกขึ้นมาใช้ สิ่งที่ต้องคำนึงถึงคือวิธีการเก็บและเรียกความรู้ขึ้นมาใช้

2. การเรียนรู้โดยการสั่งสอน (Learning from instruction) หรือการเรียนรู้จากการบอก (Learning by being told) การเรียนรู้วิธีนี้ เป็นการค้นหาความรู้จากครูหรือแหล่งข้อมูลอื่น เช่น หนังสือ ต้องการการเปลี่ยนรูปของความรู้จากคำสอนของครูหรือข้อความในหนังสือ ให้อยู่ในรูปที่เข้าใจ โดยรวมเข้ากับความรู้เดิมที่มีอยู่ ผู้เรียนต้องการความสามารถในการอนุมานบ้าง แต่การอนุมานส่วนใหญ่จะมาจากครูผู้สอน ที่จะคอยชี้แนะเพิ่มเติมความรู้ให้มากขึ้น เหมือนกับการเรียนการสอนทั่ว ๆไป ถ้าเป็นการเรียนรู้ของเครื่อง ระบบจะรับคำสั่งหรือคำแนะนำเก็บไว้ แล้วนำไปประยุกต์ใช้ให้เกิดประโยชน์

3. การเรียนรู้โดยการเปรียบเทียบ (Learning by analogy) จะเป็นการหาความจริงหรือความรู้ใหม่จากการเปลี่ยนรูปของข้อมูลที่เหมือนกันหรือมีความสัมพันธ์กัน เพื่อให้ได้ข้อมูลหรือความรู้ใหม่ เช่น คนที่ไม่เคยขับรถบรรทุกเลย แต่ขับรถยนต์เป็น สามารถที่จะเรียนรู้การขับรถบรรทุกได้ จากการเปรียบเทียบอุปกรณ์ที่มีในรถบรรทุกกับที่มีในรถยนต์ทั่วไป ซึ่งมีบางส่วนที่คล้ายกัน วิธีการเรียนรู้แบบนี้ต้องการการอนุมานมากกว่า 2 วิธีข้างต้น แต่ก็ยังต้องการความรู้หรือการแนะนำบางอย่างจากครูผู้สอน โดยความจริงหรือความชำนาญที่ใช้แก้ปัญหาเก่า ซึ่งคล้ายคลึงกับปัญหาใหม่จะถูกนำออกมาจากความจำ แล้วเปลี่ยนรูปเพื่อให้เหมาะสมกับการแก้ปัญหาใหม่ และสามารถจะเก็บไว้ใช้ต่อได้

4. การเรียนรู้จากตัวอย่าง (Learning from example) เป็นการเรียนรู้จากชุดของตัวอย่างเฉพาะในสิ่งที่ต้องการจะเรียนรู้ ซึ่งมีทั้งตัวอย่างที่ถูกและผิด ระบบจะอนุมานข้อมูลตัวอย่างที่ให้ แล้วสรุปเป็นกฎหรือความรู้ ที่สามารถใช้แก้ปัญหากับตัวอย่างใหม่ ๆ ได้ สามารถแบ่งตามแหล่งของตัวอย่าง ได้ดังนี้

4.1 ตัวอย่างมาจากครูผู้สอน ซึ่งเข้าใจปัญหาหรือแนวคิด โดยครูจะสร้างตัวอย่างขึ้นมาเพื่อช่วยให้ผู้เรียนเข้าใจ ถ้าครูรู้ระดับความเข้าใจของผู้เรียนก็สามารถจะสร้างตัวอย่างที่ทำให้ผู้เรียนเข้าใจได้เร็วขึ้น เช่น ระบบของวินสตัน (Winston's system) (Winston, 1992)

4.1 ตัวอย่างมาจากผู้เรียนเอง โดยผู้เรียนจะเข้าใจระดับความรู้ของตัวเอง แล้วลองสร้างตัวอย่างด้วยความเข้าใจของตัวเอง แล้วทดลองหาข้อสรุป เช่น ถ้าผู้เรียนต้องการจะรู้ว่าสารแม่เหล็กคืออะไร แล้วผู้เรียนเข้าใจว่า โลหะทุกชนิดเป็นสารแม่เหล็ก เมื่อผู้เรียนทดสอบตัวอย่าง คือ ทองแดง กับแม่เหล็ก พบ

ว่าทองแดงซึ่งเป็นโลหะชนิดหนึ่งไม่ได้เป็นสารแม่เหล็ก ดังนั้นความเข้าใจเกี่ยวกับสารแม่เหล็กก็จะเปลี่ยนไป ไม่ได้เข้าใจว่าโลหะทุกชนิดเป็นสารแม่เหล็ก

4.3 ตัวอย่างมาจากสภาพแวดล้อมภายนอก ในแบบนี้ ตัวอย่างมีการสร้างขึ้นมาอย่างสุ่มจากสภาพแวดล้อมภายนอก ผู้เรียนต้องเชื่อถือข้อมูลซึ่งควบคุมไม่ได้ เช่น นักดาราศาสตร์พยายามที่จะเรียนรู้ถึงการเกิด supernovas จากตัวอย่างข้อมูลที่ไม่มีรูปแบบที่ปรากฏให้เห็น

การเรียนรู้จากตัวอย่าง จะเป็นเหมือนการทดลอง เพื่อหาข้อสรุปหรือเพิ่มพูนความรู้จากตัวอย่าง ซึ่งในการหาข้อสรุปจากตัวอย่าง ตัวอย่างทั้งหมดจะปรากฏครั้งเดียวเพื่อวิเคราะห์หาข้อสรุป แต่ถ้าเป็นการเพิ่มพูนความรู้ ระบบจะต้องตั้งสมมุติฐานซึ่งสอดคล้องกับความรู้ที่มีอยู่ขึ้นมา แล้วหาข้อสรุปจากตัวอย่างที่มีเพิ่มเติมเรื่อยๆ คล้ายกับการเรียนรู้ของมนุษย์

5. การเรียนรู้จากการสังเกตุและค้นหา (Learning from observation and discovery) เป็นการเรียนรู้ทั่วๆไปที่ไม่ต้องการความช่วยเหลือจากผู้สอน ข้อมูลจะถูกป้อนให้กับระบบ โดยไม่มีการแบ่งหรือจัดกลุ่มของข้อมูล ระบบจะจัดการแบ่งกลุ่มข้อมูล หาความสัมพันธ์ และคิดหาเหตุผลออกมาเอง ซึ่งเป็นระบบที่ยังอยู่ในขั้นของงานวิจัย เพราะเป็นระบบที่ยุ่งยากซับซ้อนมาก โดยเฉพาะในการแบ่งกลุ่มและหาความสัมพันธ์ของข้อมูล

<u>การเรียนรู้จากตัวอย่าง</u> (Learning from example)

การเรียนรู้จากตัวอย่างเป็นการเรียนรู้ที่อาศัยการศึกษาจากตัวอย่าง เพื่อหากฎ ความจริง หรือข้อสรุปที่ใช้ในการแก้ปัญหาหรือการตัดสินใจ โดยอาศัยหลักของการจัดหมวดหมู่ (Classification) ซึ่งเป็นการวิเคราะห์แบ่งจัดกลุ่มตัวอย่าง หาความสัมพันธ์ แล้วสร้างเป็นกฎหรือความจริงที่สามารถให้คำจำกัดความ ที่บ่งบอกถึงกลุ่ม (Class) หรือวัตถุได้ (Rich and Knight, 1991)

งานด้านการจัดหมวดหมู่ เป็นงานที่สำคัญและพบแฝงอยู่ในวิธีการแก้ปัญหาต่าง ๆ เช่น ในระบบที่เราสร้างกฎไว้แล้วว่า ถ้าเราต้องการจะข้ามไปอีกด้านหนึ่งของกำแพง ให้มองหาประตูที่อยู่บนกำแพงนั้น แล้วเดินผ่านไป ซึ่งระบบนี้จะใช้ได้ ถ้าระบบรู้ว่าสิ่งไหนคือกำแพงและสิ่งไหนคือประตู

การที่จะกำหนดว่าสิ่งไหนคือกำแพง สิ่งไหนคือประตู ในการเรียนรู้วิธีอื่น กฎหรือคำนิยามของประตูหรือกำแพงจะถูกกำหนดไว้ล่วงหน้าแล้ว เช่น ประตูมีลักษณะเป็นสี่เหลี่ยม มีกลอน หรือลูกบิด กำแพงจะมีลักษณะเป็นสิ่งที่ขวางระหว่างที่ 2 แห่ง เป็นต้น ซึ่งคำจำกัดความเหล่านี้จะต้องอาศัยผู้ที่มีความเข้าใจเป็นอย่างดี ในการให้คำจำกัดความ แต่ถ้าเป็นการเรียนรู้จากตัวอย่าง คำจำกัดความของประตูหรือกำแพงจะไม่ได้ถูกกำหนดให้กับระบบ แต่ข้อมูลตัวอย่างที่บ่งบอกว่าสิ่งที่มีลักษณะเช่นใดเป็นประตูหรือกำแพง จะถูกป้อนให้กับระบบแทน เช่น สิ่งที่เป็นสี่เหลี่ยมเป็นประตู สิ่งที่ทำจากไม้มีกลอนเป็นประตู จากข้อมูลเหล่านี้เครื่องจะวิเคราะห์หากฎหรือคำจำกัดความที่จะบ่งบอกถึงประตูได้ด้วยตนเอง

การกำหนดกลุ่มหรือคำจำกัดความของสิ่งหนึ่งสิ่งใดจะเป็นเรื่องยาก ถ้าสิ่งนั้นเป็นสิ่งที่เราไม่เข้าใจเป็นอย่างดี หรือมีการเปลี่ยนแปลงอยู่เสมอ

การเรียนรู้จากตัวอย่างมีวิธีหรือเทคนิคในการคิดอยู่หลายอย่าง ซึ่งอาจจะเหมาะสมกับงานที่ต่างกันไป เช่น การเรียนรู้โดยการบันทึกเหตุการณ์ (Learning by recording case) การเรียนรู้โดยการจัดการตัวอย่างหลายแบบ (Learning by managing multiple models) การเรียนรู้โดยการสร้างต้นไม้ตัดสินใจ (Learning by building decision trees) การเรียนรู้โดยการฝึกโครงข่ายประสาท (Learning by training

neural net) เป็นต้น แต่วิธีที่จะนำมาใช้แก้ปัญหา ในวิทยานิพนธ์ฉบับนี้จะใช้วิธีการเรียนรู้โดยการสร้างต้นไม้ตัดสินใจของ J. Rose Quinlan ที่เรียกว่า C4.5 (J. Rose Quinlan, 1993) เพราะลักษณะของตัวอย่างที่ใช้ศึกษาจะเหมาะสมกับวิธีนี้ และวิธีนี้จะแสดงกฎหรือคำจำกัดความของปัญหาในรูปของต้นไม้ที่เป็นลำดับชั้น หรือในรูปของกฎที่เข้าใจง่าย

C4.5

C4.5 เป็นโปรแกรมที่พัฒนาโดย J. Rose Quinlan ซึ่งพัฒนาต่อมาจาก ID3 (J. Rose Quinlan, 1986) เป็นวิธีการเรียนรู้จากตัวอย่างที่อาศัยวิธีการจัดหมวดหมู่ (Classification Model) จากตัวอย่างเฉพาะที่เรียกว่าข้อมูลสอน (Training Data) แล้วสร้างเป็นต้นไม้ตัดสินใจ (Decision Tree) หรือกฎการตัดสินใจ (Rule) ได้โดยอัตโนมัติ

ข้อมูลสอนจะมีลักษณะคล้ายกับข้อมูลในฐานข้อมูลแบบสัมพันธ์ (Relational Database) ที่ประกอบด้วย แถว (Record) หรือในที่นี้เรียกว่าตัวอย่าง (Case) และ สดมภ์ (Column) หรือในที่นี้เรียกว่า ลักษณะ (Attribute) ซึ่งมีด้วยกัน 2 ชนิด ดังตัวอย่างในตารางที่ 3.1 คือ

1. ลักษณะแบ่งพวก (Category Attribute) หรือในที่นี้จะเรียกว่า พวก (Class) เป็นลักษณะที่กำหนดว่าตัวอย่างนั้น ๆ ถูกจัดอยู่ในกลุ่มไหน โดยจะมีเพียง 1 ลักษณะในแต่ละชุดข้อมูล และข้อมูลที่เก็บจะเป็นชนิดไม่ต่อเนื่อง (Discrete Value) เท่านั้น เช่น {ใช่, ไม่ใช่}, {ถูก ,ผิด,} เป็นต้น
2. ลักษณะไม่แบ่งพวก (Non-Category Attribute) เป็นชุดข้อมูลที่บ่งบอกถึงลักษณะต่าง ๆ ของตัวอย่างแต่ละตัวอย่าง โดยแต่ละลักษณะอาจจะเก็บข้อมูลได้ทั้งชนิด ค่าต่อเนื่อง (Continuous Values) เช่น ส่วนสูง, น้ำหนัก เป็นต้น หรือค่าไม่ต่อเนื่อง เช่น สีผม, อาชีพ เป็นต้น

ตัวอย่างลักษณะแบ่งพวก

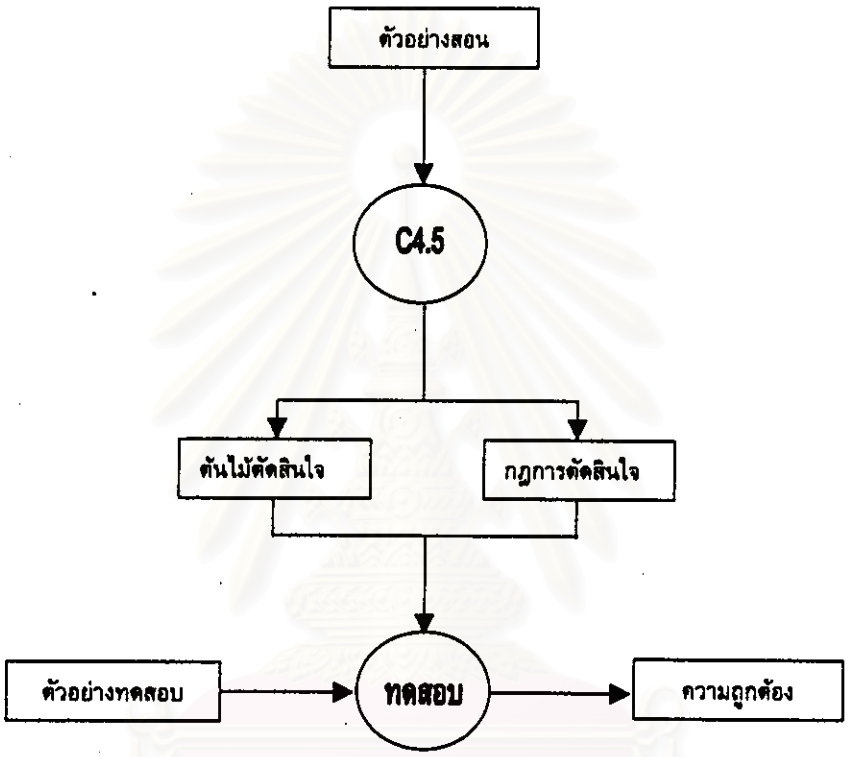
ลักษณะ	ค่าที่เป็นไปได้
การออกรอบ	ออกรอบ, ไม่ออกรอบ

ตัวอย่างลักษณะไม่แบ่งพวก

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

ตารางที่ 3.1 ตัวอย่างลักษณะต่าง ๆ ของการตัดสินใจเล่นกอล์ฟ

จากรูปที่ 3.1 ในการสร้างต้นไม้ตัดสินใจหรือกฎการตัดสินใจ ความสำเร็จไม่ได้อยู่ตรงที่สามารถสร้างต้นไม้ที่สามารถจัดกลุ่มข้อมูลจากตัวอย่างที่ใช้เรียนรู้ได้อย่างถูกต้องเท่านั้น แต่เราหวังว่ามันจะสามารถจัดกลุ่มข้อมูลจากตัวอย่างใหม่ ๆ แบบเดียวกันที่นอกเหนือจากข้อมูลที่ใช้สอนได้อย่างถูกต้องด้วย ดังนั้นในการสร้างต้นไม้ตัดสินใจจึงควรจะมีข้อมูลทดสอบ (Test Data) ที่จะใช้ในการตรวจสอบความถูกต้องของต้นไม้ตัดสินใจที่ได้

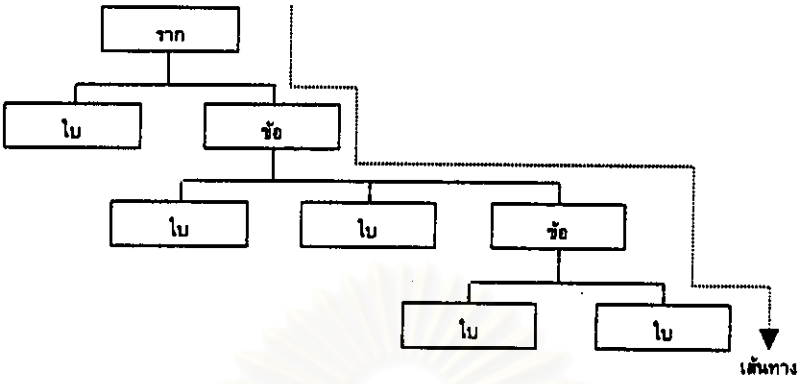


รูปที่ 3.1 แสดงขั้นตอนการทำงานของ C4.5

ต้นไม้ตัดสินใจ (Decision Trees)

ต้นไม้ตัดสินใจเป็นโครงสร้างที่ประกอบขึ้นจาก ราก (Root) ข้อ (Node) กิ่ง (Branch) และ ใบ (Leaf) ช่วยในการตัดสินใจหรือตอบคำถามในเรื่องเฉพาะที่ต้นไม้นี้เก็บไว้ โดยเริ่มจากส่วนราก แล้วไล่ลงไปตาม ข้อ กิ่ง จนกระทั่งถึง ใบ ซึ่งเป็นคำตอบหรือการตัดสินใจ ดังแสดงในรูปที่ 3.2

1. ราก เป็นจุดเริ่มต้นหรือ ข้อแรกของคำถาม โดยคำตอบจะเป็นค่าที่เป็นไปได้ของลักษณะไม่แบ่งพวกบนข้อมูลตัวอย่าง ถ้าคำตอบตรงกับค่าใดบนข้อนี้ ก็จะวิ่งไปสู่กิ่งหรือใบต่อไป
2. ข้อ เป็นจุดของคำถามตามลักษณะไม่แบ่งพวก
3. กิ่ง เป็นค่าที่เป็นไปได้ตามลักษณะไม่แบ่งพวก ซึ่งจะนำไปสู่ข้อหรือใบ
4. ใบ เป็นคำตอบหรือการตัดสินใจ โดยจะเป็นค่าที่เป็นไปได้ของลักษณะแบ่งพวก
5. เส้นทาง (Tree Path) เป็นทางเดินตั้งแต่รากจนถึงใบแต่ละใบ ซึ่งจะนำไปสู่กฎต่อไป



รูปที่ 3.2 ส่วนประกอบของต้นไม้ตัดสินใจ

การสร้างต้นไม้ตัดสินใจ

ในการสร้างต้นไม้ตัดสินใจจะใช้วิธีแบ่งแยกแล้วจัดกลุ่ม (Divide and Conquer) โดยการเลือกลักษณะไม่แบ่งพวกขึ้นมา 1 ลักษณะจากลักษณะไม่แบ่งพวกทั้งหมด เพื่อเป็นรากของต้นไม้ จากนั้นจะแบ่งตัวอย่างออกเป็นกลุ่ม ๆ ตามค่าที่เป็นไปได้ของลักษณะไม่แบ่งพวกที่เลือกมาของแต่ละตัวอย่าง จากการแบ่งนี้จะทำให้เกิดเหตุการณ์ 3 อย่าง คือ

1. กลุ่มตัวอย่างหลังจากแบ่งแล้วจะประกอบด้วยตัวอย่างที่เป็นพวกเดียวกัน ซึ่งจะกลายเป็นใบต่อไป
2. ไม่มีตัวอย่างตกอยู่ในกลุ่มนี้หลังจากแบ่งแล้ว ที่จุดนี้ก็จะกลายเป็นใบเช่นกัน แต่ถูกจัดอยู่ในพวกไหนนั้นต้องตัดสินใจโดยใช้ข้อมูลอื่น โดย C4.5 จะใช้ค่าของพวกที่มีความถี่สูงที่สุดของข้อก่อนหน้าเป็นค่าของใบนี้
3. กลุ่มตัวอย่างที่ได้ประกอบด้วยตัวอย่างหลายพวก ซึ่งจะต้องทำการแบ่งต่อไป โดยแบ่งตัวอย่างออกเป็นกลุ่ม ๆ ตามค่าที่เป็นไปได้ในแต่ละกิ่งของข้อนี้ จากนั้นจึงเริ่มแบ่งตัวอย่างในแต่ละกิ่งโดยเลือกลักษณะไม่แบ่งพวกใหม่เพื่อแบ่งตัวอย่างต่อไป

จะเห็นได้ว่าในการสร้างต้นไม้ตัดสินใจ จะเป็นการแบ่งตัวอย่างออกเป็นกลุ่ม ๆ ตามลักษณะไม่แบ่งพวก จนกระทั่งได้กลุ่มตัวอย่างที่เป็นพวกเดียวกัน จากตัวอย่างการตัดสินใจเล่นกอล์ฟ ซึ่งประกอบด้วยลักษณะไม่แบ่งพวก 4 ลักษณะ และ ลักษณะแบ่งพวกที่แบ่งตัวอย่างออกเป็น 2 พวก ถ้าตัวอย่างถูกเรียงตามลักษณะสภาพแวดล้อม ดังตารางที่ 3.2

เมื่อตัวอย่างไม่ได้ตกอยู่ในพวกเดียวกัน วิธีการแบ่งแยกและจัดกลุ่มจะพยายามแบ่งตัวอย่างออกเป็นกลุ่มย่อย ตามค่าที่เป็นไปได้ของลักษณะนั้นของแต่ละตัวอย่าง และในแต่ละกลุ่มย่อยก็จะมีการแบ่งตัวอย่างต่อไปจนกลุ่มย่อยที่ได้เป็นตัวอย่างเป็นพวกเดียวกัน หรือไม่มีตัวอย่างให้แบ่งต่ออีก จากตัวอย่างสอนในตารางที่ 3.2 เมื่อแบ่งตัวอย่างตามลักษณะสภาพแวดล้อม ในกลุ่มตัวอย่างย่อยที่มีค่าเป็น แดดจ้า และ ฝนตก จะประกอบด้วยตัวอย่างหลายพวก ส่วนกลุ่มย่อยที่มีค่าเป็นแดดร่ม จะประกอบด้วยตัวอย่างพวกเดียว ถ้าใน

กลุ่มของแดดจ้ามีการแบ่งต่อ โดยเลือกลักษณะความชื้นที่ระดับความชื้นน้อยกว่าหรือเท่ากับ 75 เปอร์เซ็นต์ และ ความชื้นมากกว่า 75 เปอร์เซ็นต์ เป็นระดับที่ใช้แบ่งตัวอย่างต่อ ส่วนในกลุ่มย่อยที่มีค่าของลักษณะสภาพแวดล้อมเป็นฝนตก จะแบ่งต่อโดยใช้ลักษณะกระแสลมเป็นตัวแบ่ง โดยหลังจากแบ่งแล้ว แต่ละกลุ่มย่อยจะประกอบด้วยตัวอย่างที่เป็นพวกเดียวกัน ได้เป็นต้นไม้ตัดสินใจตามรูปที่ 3.3 โดยมีการแสดงการแบ่งกลุ่มตัวอย่างดังตารางที่ 3.3

สภาพแวดล้อม	อุณหภูมิ (°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

ตารางที่ 3.2 ตัวอย่างสอนของการตัดสินใจเล่นกอล์ฟ (Quinlan, 1993)

ในการสร้างต้นไม้ตัดสินใจด้วยวิธีแบ่งแยกแล้วจัดกลุ่มจากตัวอย่างใด ๆ สามารถจะสร้างต้นไม้ขึ้นมาได้หลายต้น จากข้อมูลชุดเดียวกันขึ้นอยู่กับการเลือกลักษณะที่ใช้แบ่งที่แตกต่างกันไป ยิ่งจำนวนลักษณะที่ใช้แบ่งตัวอย่างและค่าที่เป็นไปได้ในแต่ละลักษณะยิ่งมาก ก็จะทำให้ได้จำนวนต้นไม้ตัดสินใจที่เป็นไปได้มากขึ้น และต้นไม้ที่ได้ก็จะมีขนาดต่าง ๆ กัน บางต้นก็มีขนาดเล็ก บางต้นก็มีขนาดใหญ่ แต่ต้นไม้ที่เราต้องการจะเป็นต้นไม้ที่มีขนาดเล็ก เพราะจะใช้จำนวนครั้งในการแบ่งตัวอย่างน้อย และเป็นต้นไม้ที่เข้าใจง่าย ดังนั้นจึงเป็นการยากที่จะสร้างต้นไม้ทั้งหมดที่เป็นไปได้ก่อน แล้วเลือกต้นไม้ที่ต้องการออกมาเมื่อมีจำนวนลักษณะ หรือค่าที่เป็นไปได้ในแต่ละลักษณะมีจำนวนมาก

จะเห็นได้ว่าจุดสำคัญจะอยู่ที่การเลือกลักษณะที่ใช้แบ่งตัวอย่าง เนื่องจากวิธีนี้จะเป็นการทำไปข้างหน้าไม่มีการย้อนกลับ คือเมื่อเลือกลักษณะหนึ่งลักษณะใดขึ้นมาแบ่งตัวอย่างแล้ว จะไม่มีการถอยหรือกลับมาเลือกลักษณะอื่นเพื่อแบ่งใหม่อีก ดังนั้นจึงต้องเลือกลักษณะที่ดีที่สุดในการแบ่งตัวอย่างของแต่ละกิ่ง ซึ่งลักษณะที่ดีที่สุดควรเป็นลักษณะที่เมื่อแบ่งตัวอย่างตามลักษณะนี้แล้ว จะทำให้จำนวนครั้งของการแบ่งต่อหรือจำนวนข้อต่อจากกิ่งนี้น้อยที่สุด ซึ่งจะนำไปสู่ต้นไม้ที่เล็กและเข้าใจง่าย

สภาพแวดล้อม = แดดจ้า:

ความชื้น <= 75 %:

สภาพแวดล้อม	อุณหภูมิ (°F)	ความชื้น (%)	กระแสลม	การออกรอบ
แดดจ้า	75	70	ลมแรง	ออกรอบ
แดดจ้า	69	70	ลมปกติ	ออกรอบ

ความชื้น > 75 %:

สภาพแวดล้อม	อุณหภูมิ (°F)	ความชื้น (%)	กระแสลม	การออกรอบ
แดดจ้า	80	90	ลมแรง	ไม่ออกรอบ
แดดจ้า	85	85	ลมปกติ	ไม่ออกรอบ
แดดจ้า	72	95	ลมปกติ	ไม่ออกรอบ

สภาพแวดล้อม = แดดร่ม:

สภาพแวดล้อม	อุณหภูมิ (°F)	ความชื้น (%)	กระแสลม	การออกรอบ
แดดร่ม	72	90	ลมแรง	ออกรอบ
แดดร่ม	83	78	ลมปกติ	ออกรอบ
แดดร่ม	64	65	ลมแรง	ออกรอบ
แดดร่ม	81	75	ลมปกติ	ออกรอบ

สภาพแวดล้อม = ฝนตก:

กระแสลม = ลมแรง:

สภาพแวดล้อม	อุณหภูมิ (°F)	ความชื้น (%)	กระแสลม	การออกรอบ
ฝนตก	71	80	ลมแรง	ไม่ออกรอบ
ฝนตก	65	70	ลมแรง	ไม่ออกรอบ

กระแสลม = ลมปกติ:

สภาพแวดล้อม	อุณหภูมิ (°F)	ความชื้น (%)	กระแสลม	การออกรอบ
ฝนตก	75	80	ลมปกติ	ออกรอบ
ฝนตก	68	80	ลมปกติ	ออกรอบ
ฝนตก	70	96	ลมปกติ	ออกรอบ

ตารางที่ 3.3 การแบ่งตัวอย่างของการตัดสินใจเล่นกอล์ฟ (Quinlan, 1993)

สภาพแวดล้อม = แดดจ้า:
- ความชื้น <= 75 %: ออกรอบ
- ความชื้น > 75 %: ไม่ออกรอบ

สภาพแวดล้อม = แดดร่ม: ออกรอบ

สภาพแวดล้อม = ฝนตก:
- กระแสลม = ลมแรง: ไม่ออกรอบ
- กระแสลม = ลมปกติ: ออกรอบ

รูปที่ 3.3 ต้นไม้ตัดสินใจของการตัดสินใจเล่นกอล์ฟ (Quinlan, 1993)

ค่ามาตรฐานเกน (Gain Criterion)

ในวิธีการสร้างต้นไม้ตัดสินใจแบบ ID3 จะใช้ค่ามาตรฐานเกน (Gain Criteria) ในการตัดสินใจเลือกลักษณะที่จะใช้เป็นรากหรือกิ่งในต้นไม้ โดยการคำนวณค่าเกนของแต่ละลักษณะเมื่อใช้แบ่งตัวอย่าง แล้วเลือกลักษณะที่มีค่าเกนสูงที่สุดมาเป็นรากหรือข้อ ค่าเกนนี้คำนวณได้โดยใช้ความรู้จากทฤษฎีสารสนเทศ (Information Theory) ของ Shannon (Shannon, 1948) ซึ่งมีสารสำคัญคือ ค่าสารสนเทศของข้อมูลขึ้นอยู่กับค่าความน่าจะเป็นของข้อมูล ซึ่งสามารถวัดอยู่ในรูปของ บิต (Bits) จากสูตร

ค่าสารสนเทศของข้อมูล P = - $\log_2$(ความน่าจะเป็นของข้อมูล P) บิต

ถ้าให้ชุดของข้อมูล M ประกอบด้วยค่าที่เป็นไปได้ คือ $\{m_1, m_2, ..., m_n\}$ และให้ความน่าจะเป็นเท่ากับ $P(m_i)$ สำหรับแต่ละค่าที่ปรากฏอยู่ในชุดข้อมูล M ดังนั้นค่าสารสนเทศของ M หรือค่าเอนโทรฟี (Entropy) ของ M จะคำนวณได้จากสูตร

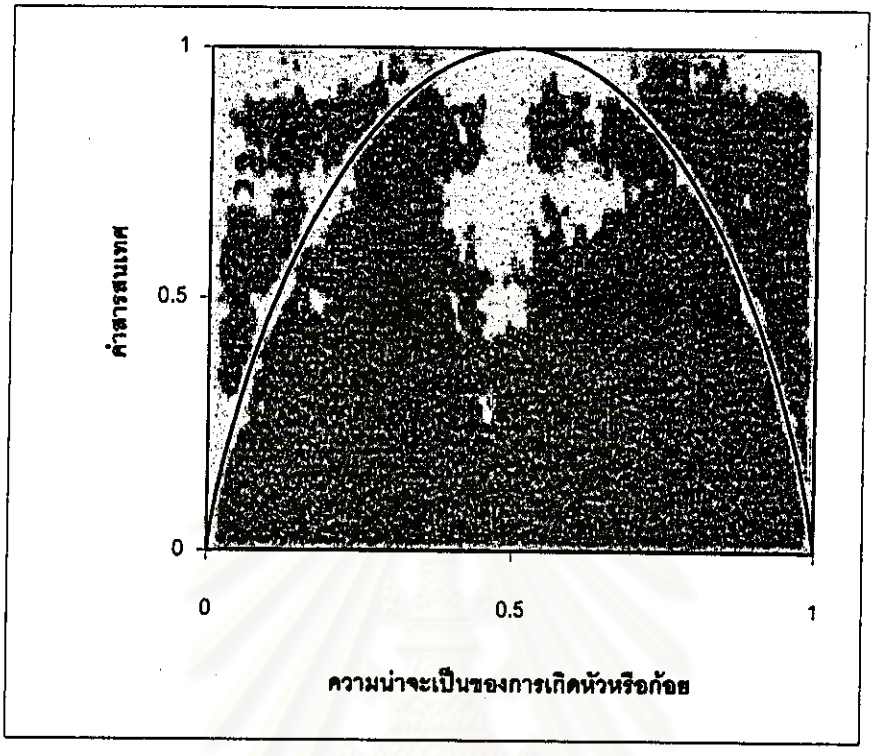
$$I(M) = \sum_{i}^{n} - P(m_i) \times \log_2(P(m_i)) \quad \text{บิต}$$

ตัวอย่างเช่น ในการโยนหัว โยนก้อย 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 3.4 ซึ่งจะเห็นได้ว่าเมื่อออกหัวหมดหรือก้อยหมด ค่าสารสนเทศจะเป็น 0 และค่าสารสนเทศจะค่อย ๆ เพิ่มขึ้นจนสูงที่สุดเมื่อความน่าจะเป็นของการเกิดหัวเท่ากับความน่าจะเป็นของการเกิดก้อย แสดงให้เห็นว่าค่าสารสนเทศที่น้อยจะบ่งบอกว่าข้อมูลชุดนั้นมีความแตกต่างกันน้อยหรือเกือบจะเป็นพวกเดียวกัน แต่ถ้าค่าสารสนเทศสูงจะบ่งบอกว่าข้อมูลชุดนั้นมีความแตกต่างกันมาก หรือประกอบด้วยตัวอย่างหลายพวกที่มีจำนวนใกล้เคียงกัน

ในการเลือกลักษณะที่จะมาเป็นรากของข้อใด ๆ จะอาศัยค่ามาตรฐานเกน ซึ่งคำนวณจากค่าสารสนเทศทั้งหมดของชุดข้อมูลนั้นลบด้วยค่าสารสนเทศหลังจากเลือกลักษณะใดลักษณะหนึ่งเป็นราก ค่าสารสนเทศหลังจากแบ่งตามลักษณะที่เลือกแล้วจะคำนวณได้จาก ค่าผลรวมของผลคูณระหว่างค่าสารสนเทศของแต่ละข้อกับอัตราส่วนของตัวอย่างในแต่ละกิ่งต่อตัวอย่างทั้งหมดที่ข้อนั้น ๆ หรือความน่าจะเป็นของค่าที่เป็นไปได้ของแต่ละลักษณะ



รูปที่ 3.4 กราฟแสดงค่าความรู้ของการโยนหัวโยนก้อย

ถ้าให้ข้อมูลสอนคือ T และลักษณะไม่แบ่งพวกที่เลือกเป็นราก คือ X และมีค่าทั้งหมดที่เป็นไปได้ N ค่า รากหรือข้อปัจจุบันจะแบ่งตัวอย่าง T ออกเป็นกลุ่มย่อย ๆ $\{t_1, t_2, ..., t_n\}$ ตามค่าที่เป็นไปได้ของ X ดังนั้น จึงสามารถคำนวณค่าความรู้หลังจากแบ่งตามลักษณะ X ดังนี้

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

ค่ามาตรฐานเกนของลักษณะไม่แบ่งพวก X สามารถคำนวณได้จากการลบค่าสารสนเทศทั้งหมดที่ข้อนี้กับค่าสารสนเทศที่ได้หลังจากแบ่งด้วยลักษณะ X ดังนี้

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

ในการตัดสินใจเลือกลักษณะไหนเป็นรากหรือข้อนั้น จะใช้ค่ามาตรฐานเกนที่มีค่าสูงสุดเป็นตัวตัดสิน ถ้าค่ามาตรฐานเกนที่คำนวณจากการแบ่งตัวอย่างตามลักษณะไหนที่มีค่าสูงที่สุดก็จะเลือกลักษณะนั้นเป็นรากหรือข้อ จากตัวอย่างการตัดสินใจเล่นกอล์ฟในตารางที่ 3.2 ประกอบด้วยข้อมูล 2 พวก คือ ตัวอย่างที่ตัด

สินใจออกรอบ 9 ตัวอย่าง และตัดสินใจไม่ออกรอบ 5 ตัวอย่าง ดังนั้นค่าสารสนเทศทั้งหมดของข้อมูลชุดนี้จะคำนวณได้ดังนี้

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

ถ้าแบ่งข้อมูลชุดนี้ตามลักษณะสภาพแวดล้อม จะแบ่งตัวอย่างออกเป็น 3 กลุ่มย่อย และสามารถคำนวณค่าสารสนเทศหลังจากแบ่งตัวอย่างตามลักษณะนี้ คือ

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

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

แต่ถ้าเราแบ่งข้อมูลชุดนี้ตามลักษณะกระแสลม จะแบ่งตัวอย่างออกเป็น 2 กลุ่มย่อย และสามารถคำนวณค่าสารสนเทศหลังจากแบ่งตามลักษณะนี้ คือ

$$I_{\text{กระแสลม}}(T) = 6/4 \times (-3/6 \times \log_2(3/6) - 3/6 \times \log_2(3/6))$$
$$+ 8/14 \times (-6/8 \times \log_2(6/8) - 2/8 \times \log_2(2/8))$$
$$= 0.892 \text{ บิต}$$

$$\text{ค่ามาตรฐานเกน(กระแสลม)} = 0.9440 - 0.892$$
$$= 0.048 \text{ บิต}$$

จะเห็นได้ว่าค่ามาตรฐานเกนของสภาพแวดล้อม จะมากกว่าค่ามาตรฐานเกนของกระแสลม ดังนั้นเราจึงเลือกที่จะแบ่งตัวอย่างตามลักษณะของสภาพแวดล้อม ส่วนในลักษณะอุณหภูมิและความชื้นซึ่งเป็นข้อมูลแบบค่าต่อเนื่องนั้น จะมีวิธีคำนวณค่ามาตรฐานเกนที่แตกต่างจากการคำนวณค่ามาตรฐานเกนของลักษณะที่เป็นค่าไม่ต่อเนื่อง ซึ่งค่าที่คำนวณได้มีค่าต่ำกว่าค่ามาตรฐานเกนของสภาพแวดล้อม และจะแสดงวิธีการคำนวณให้เห็นในหัวข้อการคำนวณบนลักษณะที่เป็นข้อมูลแบบต่อเนื่อง

ค่ามาตรฐานอัตราส่วนเกน (Gain Ratio Criterion)

ใน ID3 จะใช้ค่ามาตรฐานเกนเป็นหลักในการเลือกลักษณะที่จะใช้เป็นรากหรือข้อ แต่ใน C4.5 ได้เพิ่มการใช้ค่ามาตรฐานอัตราส่วนเกน (Gain Ratio Criterion) ในการตัดสินใจเลือกลักษณะที่จะใช้เป็นราก

หรือข้ออีกอย่างหนึ่ง เนื่องจากค่ามาตรฐานเกนจะมีอคติ (Bias) อย่างมากกับข้อมูลที่ประกอบด้วยลักษณะไม่แบ่งพวกที่มีค่าที่เป็นไปได้จำนวนมาก ๆ เช่น ข้อมูลที่ประกอบด้วยลักษณะหมายเลขประจำตัว ซึ่งปกติจะไม่ซ้ำกันในแต่ละตัวอย่าง ถ้าแบ่งข้อมูลตามลักษณะนี้จะทำให้ได้จำนวนตัวอย่างเพียง 1 ตัวอย่างต่อ 1 กิ่งของต้นไม้ และชุดตัวอย่างย่อยที่ได้จะประกอบด้วยข้อมูลพวกเดียว เมื่อคำนวณค่าสารสนเทศจากการแบ่งตัวอย่างบนลักษณะนี้ จะได้เท่ากับ 0 เนื่องจากค่า $\log_2(1) = 0$ ทำให้ค่าเกนที่ได้ในลักษณะนี้จะสูงที่สุดเสมอ

การแก้ไขความอคติของค่ามาตรฐานเกนสามารถทำได้โดยการปรับค่ามาตรฐานเกนให้ถูกต้อง โดยใช้ค่าสารสนเทศของการแบ่งแยก (Split Information) ของแต่ละลักษณะ ซึ่งถ้าให้ T คือชุดของตัวอย่าง เมื่อแบ่งตัวอย่างนี้ตามลักษณะ X จะได้ชุดของตัวอย่างย่อยในแต่ละก้าน คือ $\{t_1, t_2, ..., t_n\}$ จำนวน N ชุด ตามค่าที่เป็นไปได้ในลักษณะ 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 ชุดย่อยตามลักษณะ X โดยค่านี้จะสูงสุดเมื่อ $|T_i|$ เป็น 1 เท่ากันในทุกกิ่ง และลดลงเมื่อค่า $|T_i|$ เพิ่มขึ้น เมื่อนำค่านี้ไปหารค่ามาตรฐานเกนจะได้ค่ามาตรฐานอัตราส่วนเกน ซึ่งช่วยแก้ไขความอคติของค่ามาตรฐานเกนได้ โดยทำให้ค่ามาตรฐานอัตราส่วนเกนในการแบ่งด้วยลักษณะที่มีการกระจายสูงถูกปรับลดลง ดังนั้นค่ามาตรฐานอัตราส่วนเกนในลักษณะของตัวอย่างที่มีการกระจายตัวของข้อมูลสูงดังที่กล่าวมาแล้วจึงไม่มีค่าสูงที่สุดเสมอ

ค่ามาตรฐานอัตราส่วนเกน = ค่ามาตรฐานเกน / ค่าสารสนเทศของการแบ่งแยก

จากตัวอย่างการตัดสินใจเล่นกอล์ฟในตารางที่ 3.2 เมื่อแบ่งตัวอย่างตามลักษณะสภาพแวดล้อม จะคำนวณค่ามาตรฐานอัตราส่วนเกน ได้ดังนี้

ค่าสารสนเทศของการแบ่งแยก (สภาพแวดล้อม) = $-5/14 \times \log_2(5/14) - 4/14 \times \log_2(4/14) - 5/14 \times \log_2(5/4)$
= 1.577 บิต

ค่ามาตรฐานอัตราส่วนเกน (สภาพแวดล้อม) = 0.246/1.577
= 0.156 บิต

และเมื่อแบ่งข้อมูลตัวอย่างตามลักษณะของกระแสลม จะคำนวณค่ามาตรฐานอัตราส่วนเกน ได้ดังนี้

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

การคำนวณบนลักษณะที่เป็นข้อมูลต่อเนื่อง

ในการคำนวณค่ามาตรฐานเกนหรือค่ามาตรฐานอัตราส่วนเกนบนลักษณะที่ข้อมูลเป็นค่าต่อเนื่อง (Continuous) หรือข้อมูลตัวเลข จะกระทำได้โดยการคำนวณค่ามาตรฐานอัตราส่วนเกนหลังจากแบ่งตัวอย่างตามจุดแบ่งที่เป็นไปได้ในระดับต่าง ๆ ของลักษณะที่เป็นค่าต่อเนื่อง แล้วเลือกจุดแบ่งที่มีค่ามาตรฐานอัตราส่วนเกนสูงที่สุด เป็นระดับที่จะใช้แบ่งตัวอย่าง และใช้ค่ามาตรฐานอัตราส่วนเกนที่สูงที่สุดนี้เป็นตัวแทนในการพิจารณาเลือกลักษณะที่จะใช้แบ่งตัวอย่าง

สมมุติว่า ตัวอย่างสอน T ประกอบด้วยลักษณะต่อเนื่อง A เมื่อเรียงข้อมูลตามลักษณะ A จะได้ชุดของค่าที่ไม่ซ้ำกัน M ค่า ตามลำดับ $\{v_1, v_2, ..., v_m\}$ จุดที่เป็นระดับที่ใช้แบ่งข้อมูลจะอยู่ระหว่างค่าของ v_i กับ v_i+1 ดังนั้นจึงมีจุดที่ใช้แบ่งข้อมูลจำนวน m-1 จุดที่เป็นไปได้ หรือเท่ากับ $\{v_i+1, v_i+2, ..., v_m\}$ ซึ่งโดยปกติจุดที่ใช้แบ่งข้อมูลจะใช้ค่า $(v_i+v_{i+1}) / 2$ แต่ C4.5 จะใช้ค่าจากตัวอย่างที่สูงที่สุดที่ไม่เกินจุดกึ่งกลางจากการคำนวณในแต่ละช่วง แทนที่จะใช้จุดกึ่งกลางเป็นตัวแบ่ง เพื่อรับประกันว่าค่าที่ปรากฏในต้นไม้ตัดสินใจหรือกฎการตัดสินใจจะปรากฏอยู่ในตัวอย่างด้วย

จากตัวอย่างการตัดสินใจเล่นกอล์ฟในตารางที่ 3.2 ถ้าเราแบ่งข้อมูลตามลักษณะอุณหภูมิ โดยใช้อุณหภูมิระหว่าง 70 ถึง 71 องศาฟาเรนไฮท์ เป็นจุดที่ใช้แบ่ง สามารถจะคำนวณค่ามาตรฐานเกนและค่ามาตรฐานอัตราส่วนเกนได้ดังนี้

$I_{\text{อุณหภูมิ}}(T)$ = 5/14 x (-4/5 x $\log_2(4/5)$ – 1/5 x $\log_2(1/5)$)
+ 9/14 x (-5/9 x $\log_2(5/9)$ – 4/9 x $\log_2(4/9)$)
= 0.895 บิต

ค่ามาตรฐานเกน (อุณหภูมิ) = 0.940 - 0.895
= 0.045 บิต

ค่าสารสนเทศของการแบ่งแยก (อุณหภูมิ) = -5/14 x $\log_2(5/14)$ – 9/14 x $\log_2(9/14)$
= 0.940 บิต

ค่ามาตรฐานอัตราส่วนเกน (อุณหภูมิ) = 0.045/0.940
= 0.0479 บิต

เมื่อคำนวณค่ามาตรฐานเกนและค่ามาตรฐานอัตราส่วนเกนที่แบ่งข้อมูล ณ จุดต่าง ๆ บนลักษณะอุณหภูมิและความชื้นจะได้ค่าดังตารางที่ 3.4 และ 3.5

จะเห็นได้ว่าค่ามาตรฐานเกนหรือค่ามาตรฐานอัตราส่วนเกนที่สูงสุดของลักษณะอุณหภูมิหรือความชื้นก็ยังมีค่าน้อยกว่าค่าที่คำนวณได้จากลักษณะของสภาพแวดล้อม ดังนั้นลักษณะนี้จึงไม่ถูกเลือก แต่ถ้าลักษณะนี้ถูกเลือก จุดที่มีค่ามาตรฐานเกนหรือค่ามาตรฐานอัตราส่วนเกนสูงที่สุดจะถูกใช้เป็นจุดแบ่งข้อมูล

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 3.4 ค่ามาตรฐานเกนและค่ามาตรฐานอัตราส่วนเกนเมื่อแบ่งตามลักษณะอุณหภูมิ

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

ตารางที่ 3.5 ค่ามาตรฐานเกนและค่ามาตรฐานอัตราส่วนเกนเมื่อแบ่งตามลักษณะความชื้น

การจัดการกับตัวอย่างที่ไม่ทราบค่า

ในความเป็นจริงของข้อมูล เราอาจจะพบว่าในบางลักษณะของตัวอย่างอาจจะประกอบด้วยข้อมูลที่ไม่ทราบค่า เมื่อจะนำข้อมูลนี้มาเป็นข้อมูลสอนก็จะมีทางเลือก 2 ทาง คือ ทางแรกเป็นการนำเฉพาะตัวอย่างที่ทราบค่าเท่านั้นมาใช้เป็นข้อมูลสอน ซึ่งอาจทำให้ตัวอย่างที่ใช้สอนน้อยลงและอาจสูญเสียความรู้บางอย่างที่จะได้จากข้อมูลชุดนี้ก็ได้ ส่วนอีกทางหนึ่งจะเป็นการรวมเอาตัวอย่างที่มีบางลักษณะที่ไม่ทราบค่าเข้าไปด้วย แต่จะมีวิธีจัดการเพื่อให้ได้ประโยชน์จากตัวอย่างเหล่านี้ ซึ่งมีอยู่ด้วยกันหลายวิธี เช่น การเติมค่าที่สูญหาย

ด้วยค่าที่มีความถี่สูงที่สุด แต่ C4.5 จะใช้วิธีคำนวณค่ามาตรฐานเกนหรือค่ามาตรฐานอัตราส่วนเกนจากชุดข้อมูลที่ทราบค่าของลักษณะนั้น แล้วปรับลดค่าให้ถูกต้องด้วยความน่าจะเป็นของตัวอย่างที่ทราบค่า ต่อตัวอย่างทั้งหมด

ถ้าให้ T เป็นชุดข้อมูลสอน และ X เป็นลักษณะที่ใช้ทดสอบบนตัวอย่าง A เราสามารถจะคำนวณค่าสารสนเทศทั้งหมดของชุดตัวอย่าง T และค่าสารสนเทศหลังจากแบ่งตัวอย่างบนลักษณะ X เพื่อคำนวณหาค่ามาตรฐานเกนได้จากตัวอย่าง A ที่ทราบค่าเท่านั้น ซึ่งถ้าให้ค่ามาตรฐานเกนของตัวอย่างที่ไม่ทราบค่าเป็น 0 จะได้ว่า

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

โดยที่ ความน่าจะเป็นที่ A ทราบค่า = จำนวนตัวอย่างที่ทราบค่า / จำนวนตัวอย่างทั้งหมด

ส่วนค่าสารสนเทศของการแบ่งแยก X สามารถปรับได้โดยการเพิ่มชุดตัวอย่างอีก 1 กลุ่ม สมมุติว่าลักษณะ X มีค่าที่เป็นไปได้ n ค่า เวลาคำนวณค่าความรู้ของการแบ่งแยก จะมีการแบ่งข้อมูลเป็น n+1 กลุ่มย่อย โดยกลุ่มที่เพิ่มมาจะเป็นกลุ่มของตัวอย่างที่ไม่ทราบค่า

เมื่อมีการแบ่งตัวอย่างตามลักษณะ X เป็นชุดย่อย t_1, t_2, ...,t_n ชุด ตามค่าที่เป็นไปได้ o_1, o_2, ...,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 3.2 สมมุติว่าในตัวอย่างลักษณะสภาพแวดล้อมที่เท่ากับแดดร่ม อุณหภูมิเท่ากับ 72 องศา ความชื้นเท่ากับ 90 % และกระแสลมแรง เราไม่ทราบค่าของสภาพแวดล้อมของตัวอย่างนี้ เมื่อเราสนใจตัวอย่างที่เหลืออีก 13 ตัวอย่าง ซึ่งทราบค่าของลักษณะสภาพแวดล้อมจะสามารถนับจำนวนตัวอย่างในแต่ละกลุ่มได้ดังตารางที่ 3.6

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

ตารางที่ 3.6 จำนวนตัวอย่างเมื่อแบ่งตามลักษณะสภาพแวดล้อม (Quinlan, 1993)

เมื่อคำนวณค่าต่าง ๆ บนลักษณะสภาพแวดล้อมจะได้ดังนี้

ค่าสารสนเทศของสภาพแวดล้อม = $- 8/13 \times \log_2(8/13) - 5/13 \times \log_2(5/13)$
= 0.961 บิต

ค่าสารสนเทศหลังจากแบ่งตามสภาพแวดล้อม = $5/13 \times (-2/5 \times \log_2(2/5) - 3/5 \times \log_2(3/5))$
$+ 3/13 \times (-3/3 \times \log_2(3/3) - 0/3 \times \log_2(0/3))$
$+ 5/13 \times (-3/5 \times \log_2(3/5) - 2/5 \times \log_2(3/5))$
= 0.747 บิต

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

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

ค่ามาตรฐานอัตราส่วนเกนของสภาพแวดล้อม = 0.199 / 1.809
= 0.110 บิต

เมื่อตัวอย่างทั้ง 14 ตัวอย่าง ถูกแบ่งออกเป็นชุดย่อยตามลักษณะสภาพแวดล้อม ในตัวอย่าง 13 ตัวอย่าง ที่ทราบค่าของสภาพแวดล้อมจะถูกแบ่งตามปกติ แต่ตัวอย่างที่เหลือ 1 ตัวอย่าง ซึ่งไม่ทราบค่าของสภาพแวดล้อมจะถูกแบ่งให้กับทุก ๆ ค่าที่เป็นไปได้ของลักษณะสภาพแวดล้อม คือ แดดจ้า แดดร่ม และ ฝนตก เป็นอัตราส่วน 5/13, 3/13 และ 5/13 ตามลำดับ

ถ้าดูชุดตัวอย่างย่อยหลังจากแบ่งด้วยลักษณะสภาพแวดล้อม เฉพาะในสภาพแวดล้อมที่เป็นแดดจ้า จะประกอบด้วยตัวอย่างดังตารางที่ 3.7

สภาพแวดล้อม	อุณหภูมิ ($^{\circ}$F)	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

ตารางที่ 3.7 ชุดตัวอย่างย่อยหลังจากแบ่งด้วยลักษณะสภาพแวดล้อมเฉพาะที่เป็นแดดจ้า (Quinlan, 1993)

ถ้าตัวอย่างย่อยชุดนี้ถูกแบ่งต่อบนลักษณะของความชื้นที่ 75 % เหมือนกับการแบ่งเมื่อทราบค่าลักษณะสภาพแวดล้อมของตัวอย่างทั้งหมด จะสามารถแบ่งตัวอย่างออกเป็น 2 ชุด คือ

ความชื้น ≤ 75 % ประกอบด้วยตัวอย่างที่ตัดสินใจ ออกรอบ 2 ตัวอย่าง และ ไม่ออกรอบ 0 ตัวอย่าง

ความชื้น > 75 % ประกอบด้วยตัวอย่างที่ตัดสินใจ ออกรอบ 5/13 ตัวอย่าง และ ไม่ออกรอบ 3 ตัวอย่าง

และเมื่อคำนวณค่ามาตรฐานเกนหรือค่ามาตรฐานอัตราส่วนเกน แล้วสร้างเป็นต้นไม้ตัดสินใจจะได้ต้นไม้ที่มีลักษณะเหมือนเดิม ดังนี้

```
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 = ปกติ  : ออกรอบ (3.0)
```

รูปที่ 3.5 ต้นไม้ตัดสินใจในการเล่นกอล์ฟเมื่อมีตัวอย่างไม่ทราบค่า (Quinlan, 1993)

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 ตัวอย่าง ที่ไม่ได้เป็นพวกเดียวกับตัวอย่างที่ใบนี้คือ ถ้าที่ใบนี้ถูกจัดเป็นพวกออกรอบ จะมี 0.4 ตัวอย่างที่เป็นพวกไม่ออกรอบ

ถ้าเราใช้ต้นไม้นี้ในการตัดสินใจเล่นกอล์ฟ เมื่อสภาพแวดล้อมเท่ากับแดดจ้า อุณหภูมิ 70 องศาฟาเรนไฮต์ และกระแสลมปกติ แต่เราไม่ทราบค่าของความชื้น เราสามารถจะวิ่งไปตามทางเดินของต้นไม้ โดยเริ่มจากลักษณะของสภาพแวดล้อม เมื่อทราบว่าสภาพแวดล้อมเป็นแดดจ้า ก็จะวิ่งตามลงไปสู่ข้อแรกของต้นไม้ แต่ที่ข้อนี้เราไม่ทราบค่าของความชื้น ดังนั้นจึงเป็นไปได้ว่าความชื้นอาจจะมากกว่าหรือน้อยกว่า 75 % ก็ได้ จึงเกิดเหตุการณ์ 2 อย่าง คือ

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

แต่ตอนที่ต้นไม้ถูกสร้างขึ้น สำหรับลักษณะสภาพแวดล้อมที่เป็นแดดจ้า 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 %

<u>การตัดแต่งต้นไม้ตัดสินใจ</u> (Pruning Decision Trees)

ในการสร้างต้นไม้ตัดสินใจที่ใช้วิธีแบ่งแยกและจัดกลุ่มดังที่กล่าวมาแล้ว จะแบ่งตัวอย่างจนกระทั่งได้ตัวอย่างเป็นพวกเดียวกันหรือไม่มีตัวอย่างให้แบ่งอีกแล้ว ผลที่ได้คือต้นไม้ที่มีความซับซ้อนมาก เนื่องจากมีจำนวนข้อ กิ่ง และทางเดินมาก ทำให้เข้าใจยากและการตัดสินใจจะเฉพาะเจาะจงกับข้อมูลตัวอย่างที่ใช้สอนมากเกินไป (Overfits The Data) ทำให้การตัดสินใจในตัวอย่างใหม่ ๆ ที่นอกเหนือจากตัวอย่างที่ใช้สอนมีความผิดพลาดสูง การตัดแต่งจะทำให้ต้นไม้ตัดสินใจที่ได้ใหม่ไม่เป็นต้นไม้ที่ตัดสินใจได้ถูกต้องเฉพาะกับข้อมูลที่ใช้สอนเท่านั้น แต่ต้นไม้นี้จะเป็นต้นไม้แบบง่าย (Simplified Trees) ที่สามารถใช้ตัดสินใจในตัวอย่างที่ไม่เคยเห็นได้ถูกต้องพอ ๆ กับตัวอย่างที่ใช้สอนด้วย

C4.5 จะใช้วิธีตัดแต่งโดยใช้ค่าความผิดพลาด (Error-Based Pruning) คือจะเป็นการรวมกิ่งเป็นใบหรือข้อเป็นใบ โดยที่เมื่อรวมแล้วไม่ทำให้ค่าความผิดพลาดหลังจากรวมแล้วเพิ่มขึ้น ถ้ามีตัวอย่าง N ตัวอย่างที่ใบ และมีตัวอย่าง E ตัวอย่าง เป็นตัวอย่างที่ไม่ถูกต้องหรือไม่ใช่พวกเดียวกันกับตัวอย่างที่ใบนี้ ดังนั้นค่าความผิดพลาดที่ใบนี้จะเท่ากับ E/N ซึ่งเป็นค่าความผิดพลาดที่เฉพาะกับตัวอย่างสอนชุดนี้เท่านั้น แต่เราต้องการค่าความผิดพลาดที่เป็นค่าประมาณจากประชากร เพื่อใช้แทนค่าความผิดพลาดที่คาดว่าจะเกิดเมื่อใช้ทดสอบบนข้อมูลที่ไม่เคยเห็น จึงได้ใช้ค่าจำกัดบนของการกระจายแบบไบโนเมียล (Binomial Distribution) ที่ระดับความเป็นอิสระเท่ากับ CF (Confidence Level) เป็นตัวแทนความผิดพลาดของประชากร โดยเขียนอยู่ในรูป $U_{\alpha}(E, N)$

การประมาณค่าความผิดพลาดสำหรับใบและก้าน เมื่อใช้กับข้อมูลที่ไม่เคยเห็น จะอยู่บนข้อกำหนดที่ว่าขนาดของตัวอย่างสอนเท่ากับขนาดตัวอย่างของข้อมูลที่ไม่เคยเห็น ดังนั้นถ้าใบประกอบด้วยตัวอย่าง N ตัวอย่าง ค่าความผิดพลาดที่คาดไว้ของตัวอย่างแต่ละตัวอย่างจะเท่ากับ $U_{\alpha}(E, N)$ ซึ่งสามารถคาดได้ว่าจะมีจำนวนตัวอย่างที่แบ่งผิดพลาดเท่ากับ N x $U_{\alpha}(E, N)$ ตัวอย่าง เมื่อทดสอบบนตัวอย่างที่ไม่เคยเห็น และในขณะเดียวกัน จำนวนตัวอย่างที่คาดว่าจะแบ่งผิดพลาดบนข้อใด ๆ จะเท่ากับผลรวมของจำนวนตัวอย่างที่คาดว่าจะแบ่งผิดพลาดในแต่ละกิ่งรวมกัน

```
physician fee freeze = n:
|   adoption of the budget resolution = y: democrat (151)
|   adoption of the budget resolution = u: democrat (1)
|   adoption of the budget resolution = n:
|   |   education spending = n: democrat (6)
|   |   education spending = y: democrat (9)
|   |   education spending = u: republican (1)
physician fee freeze = y:
|   synfuels corporation cutback = n: republican (97/3)
|   synfuels corporation cutback = u: republican (4)
|   synfuels corporation cutback = y:
|   |   duty free exports = y: democrat (2)
|   |   duty free exports = u: republican (1)
|   |   duty free exports = n:
|   |   |   education spending = n: democrat (5/2)
|   |   |   education spending = y: republican (13/2)
|   |   |   education spending = u: democrat (1)
physician fee freeze = u:
|   water project cost sharing = n: democrat (0)
|   water project cost sharing = y: democrat (4)
|   water project cost sharing = u:
|   |   mx missile = n: republican (0)
|   |   mx missile = y: democrat (3/1)
|   |   mx missile = u: republican (2)
```

<u>รูปที่ 3.6</u> ต้นไม้ตัดสินใจก่อนการตัดแต่ง (Quinlan, 1993)

จากตัวอย่างของต้นไม้ตัดสินใจก่อนการตัดแต่งกิ่งในรูปที่ 3.6 ในข้อหนึ่งของต้นไม้ที่ประกอบด้วยกิ่งและใบดังนี้

```
education spending = n: democrat (6)
education spending = y: republican (9)
education spending = u: democrat (1)
```

จะเห็นได้ว่าไม่มีตัวอย่างที่แบ่งกลุ่มผิดพลาดบนข้อมูลสอนหลังจากการเรียนรู้และสร้างเป็นต้นไม้ตัดสินใจ สำหรับใบแรกที่มีจำนวนตัวอย่างเท่ากับ 6 ตัวอย่าง หรือ N=6 ถูกจัดอยู่ในพวก democrat โดยที่ไม่มี

ตัวอย่างผิดพลาด 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 เมื่อคำนวณตัวอย่างที่คาดว่าจะทำนายผิดที่ข้อนี้จะเท่ากับ

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

ถ้าข้อนี้ถูกแทนที่ด้วยใบที่มีค่าเป็น 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 เราสามารถจะแทนข้อนี้ด้วยใบที่มีค่าเป็น democrat ได้ เนื่องจากจำนวนตัวอย่างที่ทำนายผิดพลาดหลังจากแทนข้อนี้ด้วยใบใหม่ แล้วจะน้อยกว่าก่อนที่จะแทนข้อนี้ด้วยใบใหม่ และได้เป็นข้อใหม่ ดังนี้

adopting of the budget resolution = y: democrate (151)
adopting of the budget resolution = u: democrate (1)
adopting of the budget resolution = n: democrate (16/1)

เมื่อคำนวณจำนวนตัวอย่างที่คาดว่าจะทำนายผิดพลาดที่จะเกิดสำหรับข้อใหม่ ได้ดังนี้

จำนวนตัวอย่างที่คาดว่าจะทำนายผิดพลาด = 151 x $U_{25\%}(0, 151)$ + 1 x $U_{25\%}(0, 1)$ + 2.512
= 4.642 ตัวอย่าง

ถ้าข้อนี้ถูกแทนด้วยใบที่มีค่าเป็น democrat อีก จะคำนวณค่าความผิดพลาดได้ เท่ากับ 168 x $U_{25\%}(1, 168)$ หรือ 2.610 ค่าที่คำนวณได้นี้ก็ยังมีค่าน้อยกว่าผลรวมของตัวอย่างที่คาดว่าจะผิดพลาดของข้อก่อนที่จะรวมเป็นใบ ดังนั้นข้อนี้ก็สามารถที่จะรวมเป็นใบได้อีก เมื่อตัดแต่งเป็นที่เรียบร้อยแล้ว ก็จะได้ต้นไม้ตัดสินใจใหม่ดังรูปที่ 3.7

```
physician fee freeze = n: democrat (168/2.6)
physician fee freeze = y: republican (123/13.9)
physician fee freeze = u:
|   mx missile = n: democrat (3/1.1)
|   mx missile = y: democrat (4/2.2)
|   mx missile = u: republican (2/1)
```

รูปที่ 3.7 ต้นไม้ตัดสินใจหลังการตัดแต่ง (Quinlan, 1993)

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

ผลรวมของตัวอย่างที่คาดว่าจะทำนายผิดที่แต่ละใบ เมื่อหารด้วยจำนวนตัวอย่างที่ใช้สอนทั้งหมด จะเป็นค่าประมาณของความน่าจะเป็นของความผิดพลาดของต้นไม้หลังจากตัดแต่งกิ่งแล้ว บนตัวอย่างที่ไม่เคยเห็น จากต้นไม้หลังตัดแต่งแล้วของตัวอย่างในรูปที่ 3.7 มีผลรวมของตัวอย่างที่คาดว่าจะทำนายผิดในแต่ละใบเป็น 20.8 ตัวอย่าง จากตัวอย่างทั้งหมด 300 ตัวอย่าง ดังนั้นค่าประมาณความผิดพลาดของต้นไม้หลังจากตัดแต่งแล้วบนตัวอย่างที่ไม่เคยเห็น จะทำนายผิดประมาณ 6.9 %

```
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 %)
```

รูปที่ 3.8 ผลของการตัดแต่งต้นไม้ ของตัวอย่างการสำรวจประชามติ (Quinlan, 1993)

การแปลงต้นไม้เป็นกฎ

การตัดแต่งต้นไม้ถึงแม้จะทำให้ต้นไม้ดูง่ายขึ้นกว่าเดิมและตัดสินใจในตัวอย่างที่ไม่เคยเห็นได้ถูกต้อง แต่ต้นไม้ที่ได้ก็ยังคง อุ้ยอ้าย ซับซ้อน และเข้าใจยาก (Quinlan 1993) ดังนั้นถ้าเราต้องการจะลดข้อเสีย

เหล่านี้ ทางหนึ่งที่เป็นไปได้คือ การแปลงต้นไม้เป็นกฎ ซึ่งเข้าใจง่ายกว่าไม่ยุ่งยากซับซ้อน แต่ให้ความถูกต้องบนตัวอย่างที่ไม่เคยเห็นไม่ต่างกัน

สมมติว่าตัวอย่างหนึ่งประกอบด้วยลักษณะ F, G, J และ K ซึ่งมีค่าที่เป็นไปได้ คือ 0 และ 1 กับลักษณะแบ่งพวกที่มีค่าที่เป็นไปได้ คือ Yes และ No ถ้าทุกตัวอย่างที่เป็นพวก Yes ค่าของลักษณะ F และ G จะเท่ากับ 1 เสมอ หรือ ค่าของ J และ K เท่ากับ 1 เสมอ เมื่อนำตัวอย่างนี้ไปสร้างเป็นต้นไม้ตัดสินใจจะได้ดังรูปที่ 3.9 ซึ่งประกอบด้วยกิ่งที่เหมือนกัน 2 กิ่ง ที่ทดสอบค่า J และ K เท่ากับ 1 ซึ่งจะทำให้ต้นไม้เข้าใจยากอุ้ยอ้าย และซับซ้อน การแปลงต้นไม้เป็นกฎจะช่วยลดความซ้ำซ้อนหรือตัดเงื่อนไขที่ไม่จำเป็นออก ทำให้กฎที่ได้กระทัดรัดเข้าใจง่าย โดยที่ความถูกต้องยังคงเดิม

กฎก็คือ ทางเดินจากรากถึงใบในแต่ละใบของต้นไม้ ถ้าเราสร้างกฎโดยการแปลงทางเดินของทุกใบเป็นกฎ กฎที่ได้ก็จะไม่ต่างอะไรจากต้นไม้ซึ่งไม่ได้ทำอะไรให้ดีขึ้น อย่างไรก็ตามเราอาจจะพบว่ากฎแต่ละกฎอาจจะประกอบด้วยเงื่อนไขที่ซ้ำกัน หรือไม่มีความจำเป็น โดยจากตัวอย่างถ้ากฎหนึ่งที่ได้จากต้นไม้นี้คือ

ถ้า F = 1
และ G = 0
และ J = 1
และ K = 1
ดังนั้น การตัดสินใจ คือ Yes

```
F = 0:
|   J = 0: No
|   J = 1:
|   |   K = 0: No
|   |   K = 1: Yes
F = 1:
|   G = 1: Yes
|   G = 0:
|   |   J = 0: No
|   |   J = 1:
|   |   |   K = 0: No
|   |   |   K = 1: Yes
```

รูปที่ 3.9 แสดงต้นไม้ตัดสินใจที่ตัดสินใจเป็น Yes เมื่อ F=G=1 หรือ J=K=1 (Quinlan, 1993)

เราพบว่าการตัดสินใจจะไม่มีผลกระทบเนื่องจากค่า F หรือค่า G ในกฎนี้ เนื่องจากเราทราบว่าการตัดสินใจเป็น Yes จะเกิดขึ้นเมื่อ J=K=1 เท่านั้นก็เพียงพอ ดังนั้น 2 เงื่อนไขนี้จึงสามารถตัดออกจากกฎนี้ได้ และเขียนเป็นกฎใหม่ได้ คือ

ถ้า J = 1
และ K = 1
ดังนั้น การตัดสินใจ คือ Yes

อย่างไรก็ตามในตัวอย่างจริงซึ่งเราไม่ทราบเงื่อนไขล่วงหน้า เราจะทราบได้อย่างไรว่าเงื่อนไขอันใดสมควรที่จะตัดออกจากกฎหรือไม่ สมมติ ถ้าให้กฎ R อยู่ในรูป

ถ้า A แล้ว ตัดสินใจ C

และ กฎหลังจากปรับแล้ว คือ R^{-} อยู่ในรูป

ถ้า A^{-} แล้ว ตัดสินใจ C

โดยที่ A^{-} เกิดจากการลบเงื่อนไข X จากกฎ R เมื่อคำนึงถึงจำนวนตัวอย่างที่สอดคล้องกับเงื่อนไข X และไม่สอดคล้องกับเงื่อนไข X กับการเข้าพวก C และการไม่เข้าพวก C หลังจากลบเงื่อนไข X แล้วจะเป็นดังตารางที่ 3.8

	เข้าพวก C	ไม่เข้าพวก C
สอดคล้องกับเงื่อนไข X	Y_1	E_1
ไม่สอดคล้องกับเงื่อนไข X	Y_2	E_2

ตารางที่ 3.8 แสดงจำนวนตัวอย่างเมื่อลบกฎ R (Quinlan, 1993)

ถ้ามองถึงจำนวนตัวอย่างที่สอดคล้องกับเงื่อนไข X เดิมของกฎ R จะประกอบด้วยตัวอย่าง Y_1 ตัวอย่างที่เข้าพวก C และ E_1 ตัวอย่าง ที่ไม่เข้าพวก C หรือคิดเป็นจำนวนตัวอย่างของทั้งหมด $Y_1 + E_1$ ตัวอย่าง และจำนวนตัวอย่างที่ไม่สอดคล้องกับเงื่อนไข X ของกฎใหม่ R^{-} จะประกอบด้วยตัวอย่าง Y_2 ตัวอย่างที่เข้าพวก C และ E_2 ตัวอย่างที่ไม่เข้าพวก C หรือคิดเป็นจำนวนตัวอย่างทั้งหมด $Y_2 + E_2$ ตัวอย่าง แต่เมื่อเงื่อนไข X ถูกลบออกจากกฎ R จะทำให้ตัวอย่างที่สอดคล้องกับเงื่อนไข X ของกฎ R ถูกรวมเข้ากับจำนวนตัวอย่างที่สอดคล้องกับกฎ R^{-} ดังนั้นจำนวนตัวอย่างทั้งหมดที่สอดคล้องกับกฎ R^{-} จะเท่ากับ $Y_1 + Y_2 + E_1 + E_2$

จากการตัดแต่งต้นไม้ตัดสินใจ เราสามารถจะประมาณค่าความผิดพลาดของแต่ละกิ่ง ได้จากค่าจำกัดบน $U_{\alpha}(E, N)$ เมื่อใช้ทดสอบกับข้อมูลที่ไม่เคยเห็นมาก่อน โดยที่ E คือจำนวนตัวอย่างที่ทำนายผิดจากจำนวนตัวอย่างทั้งหมด N ตัวอย่าง ที่ค่าระดับความเป็นอิสระเท่ากับ CF การประมาณค่าความผิดพลาดนี้สามารถนำมาใช้กับการลบเงื่อนไขบางเงื่อนไขออกจากกฎได้ โดยที่ค่าความผิดพลาดของกฎ R สามารถคำนวณได้จาก $U_{\alpha}(E_1, Y_1 + E_1)$ และความผิดพลาดของกฎ R^{-} สามารถคำนวณได้จาก $U_{\alpha}(E_1 + E_2, Y_1 + Y_2 + E_1 + E_2)$ ถ้าค่าความผิดพลาดของกฎ R^{-} น้อยกว่าค่าความผิดพลาดของกฎ R ดังนั้นการลบเงื่อนไข X ออกจากกฎ R ก็สมควรที่จะทำได้

```
TSH ≤ 6: negative
TSH > 6:
|  FTI ≤ 64:
|  |  TSH measured = f: negative
|  |  TSH measured = t:
|  |  |  T4U measured = f:
|  |  |  |  TSH ≤ 17: compensated hypothyroid
|  |  |  |  TSH > 17: primary hypothyroid
|  |  |  T4U measured = t:
|  |  |  |  thyroid surgery = f: primary hypothyroid
|  |  |  |  thyroid surgery = t: negative
|  FTI > 64:
|  |  on thyroxine = t: negative
|  |  on thyroxine = f:
|  |  |  TSH measured = f: negative
|  |  |  TSH measured = t:
|  |  |  |  thyroid surgery = t: negative
|  |  |  |  thyroid surgery = f:
|  |  |  |  |  TT4 > 150: negative
|  |  |  |  |  TT4 ≤ 150:
|  |  |  |  |  |  TT4 measured = f: primary hypothyroid
|  |  |  |  |  |  TT4 measured = t: compensated hypothyroid
```

รูปที่ 3.10 แสดงต้นไม้ตัดสินใจสำหรับเงื่อนไขการเป็นโรคไฮโปรไธรอยด์ (Quinlan, 1993)

จากตัวอย่างต้นไม้ตัดสินใจที่แสดงเงื่อนไขของการเป็นโรคไฮโปรไธรอยด์ (Hypothyroid) ดังในรูปที่ 3.10 ถ้ากฎหนึ่งในต้นไม้นี้ คือ

ถ้า TSH > 6
และ FTI < 64
และ TSH measured = t
และ T4U measured = t
และ thyroid surgery = t
แล้ว จัดอยู่ในพวก negative

เมื่อกฎนี้ครอบคลุมตัวอย่างสอน 3 ตัวอย่าง โดยถูกจัดอยู่ในพวก negative 2 ตัวอย่าง และเป็นพวกอื่น 1 ตัวอย่าง ดังนั้นค่า $Y_1=2$ และ $E_1=1$ เมื่อคำนวณค่า $U_{cf}(E_1, Y_1 + E_1)$ ที่ CF เท่ากับ 25 % จะได้ค่า

$U_{25\%}(1, 3)$ เท่ากับ 69 % ถ้าเราทดลองลบเงื่อนไขแต่ละเงื่อนไขออกจากกฎนี้ และคำนวณค่าความผิดพลาดหลังจากลบเงื่อนไขแต่ละเงื่อนไขออกแล้วจะได้ดังตารางที่ 3.9

เงื่อนไขที่ลบ	$Y_1 + Y_2$	$E_1 + E_2$	ค่าความผิดพลาด (%)
TSH > 6	3	1	55
FTI ≤ 64	6	1	34
TSH measured = t	2	1	69
T4U measured = t	2	1	69
Thyroid surgery = t	3	59	97

ตารางที่ 3.9 แสดงค่าความผิดพลาดหลังจากลบเงื่อนไขแต่ละเงื่อนไขในกฎ ขั้นตอนที่ 1 (Quinlan, 1993)

จากตารางที่ 3.9 จะเห็นได้ว่าค่าความผิดพลาดหลังจากลบเงื่อนไข FTI ≤ 64 จะมีค่าน้อยที่สุดคือ 34 % และน้อยกว่าค่าความผิดพลาดก่อนลบที่ 69 % ดังนั้นจึงสามารถลบเงื่อนไขนี้ออกจากกฎได้ จากนั้นก็ทดลองลบเงื่อนไขอื่นต่อไป โดยคำนวณค่าความผิดพลาดของกฎก่อนลบเงื่อนไข และหลังจากลบเงื่อนไขต่าง ๆ ได้ค่าความผิดพลาดก่อนลบเงื่อนไขเป็น $U_{25\%}(1, 6)$ หรือเท่ากับ 34 % และค่าความผิดพลาดหลังจากลบเงื่อนไขแล้ว ดังตารางที่ 3.10

เงื่อนไขที่ลบ	$Y_1 + Y_2$	$E_1 + E_2$	ค่าความผิดพลาด (%)
TSH > 6	31	1	8
TSH measured = t	6	1	34
T4U measured = t	7	1	30
Thyroid surgery = t	44	179	82

ตารางที่ 3.10 แสดงค่าความผิดพลาดหลังจากลบเงื่อนไขแต่ละเงื่อนไขในกฎ ขั้นตอนที่ 2 (Quinlan, 1993)

จากตารางที่ 3.10 จะเห็นได้ว่าค่าความผิดพลาดหลังจากลบเงื่อนไข TSH > 6 จะมีค่าน้อยที่สุดในการทดลองครั้งนี้ คือ 8 % และน้อยกว่าค่าความผิดพลาดก่อนลบเงื่อนไขที่ 34 % ดังนั้นจึงสามารถลบเงื่อนไขนี้ออกจากกฎได้อีก 1 เงื่อนไข เมื่อทำการลบเงื่อนไขต่าง ๆ จนไม่มีค่าความผิดพลาดของกฎหลังจากลบเงื่อนไขอันใดแล้วน้อยกว่าค่าความผิดพลาดของกฎก่อนลบเงื่อนไข ก็จะได้กฎที่ปรับปรุงให้ง่ายขึ้น โดยมีความผิดพลาดไม่ต่างจากกฎเริ่มต้น จากตัวอย่างเมื่อลบเงื่อนไขที่ไม่จำเป็นแล้วจะได้กฎซึ่งครอบคลุมตัวอย่าง 35 ตัวอย่าง ซึ่งมีค่าความผิดพลาด 7 % ดังนี้

ถ้า thyroid surgery = t
แล้ว จัดอยู่ในพวก negative

หลังจากที่ได้ปรับปรุงกฎต่าง ๆ ให้อยู่ในรูปที่ง่ายขึ้น โดยการตัดเงื่อนไขบางอย่างภายในกฎที่ไม่มีความจำเป็นหรือไม่มีผลต่อความถูกต้องของกฎนั้น ๆ แล้ว จะพบว่ากฎที่ได้บางกฎอาจจะซ้ำกัน หรือบางกฎมีค่าความผิดพลาดสูงมาก หรือมีตัวอย่างที่เข้าเงื่อนไขมากกว่า 1 กฎ ดังนั้นจึงต้องมีการกลั่นกรองเลือกเฉพาะกฎที่เหมาะสมจากชุดของกฎที่ได้ทั้งหมด และหาพวกให้กับตัวอย่างที่ไม่เข้ากับกฎใด ๆ เนื่องจากมีการลบบางกฎทิ้งไป โดยไม่ทำให้ความถูกต้องในการตัดสินใจลดลง

สมมติว่าจากกฎที่ได้ทั้งหมดหลังจากปรับปรุงแล้ว เราเลือกชุดของกฎ S ซึ่งครอบคลุมพวก C คุณค่าหรือประสิทธิภาพของชุดของกฎนี้สามารถสรุปหรือวัดได้จากจำนวนตัวอย่างซึ่งครอบคลุมโดย S ซึ่งไม่เข้าพวกกับ C ในที่นี้เรียกว่า ความผิดพลาดด้านบวก (False Positives) และจำนวนตัวอย่างที่เข้าพวก C แต่ไม่เข้ากฎใดกฎหนึ่งใน S หรือเรียกว่าความผิดพลาดด้านลบ (False Negatives) โดย C4.5 จะใช้วิธี มินิมั่ม เดสคริบชั่นเรนท์ปริ้นซิเปิล (Minimum Description Length Principle) (Quinlan และ Rivest , 1987) หรือ MDL ในการวัดประสิทธิภาพของชุดของกฎ S จากค่าระหว่างความถูกต้อง (Accuracy) ของกฎกับความซับซ้อน (Complexity) ของกฎซึ่งสามารถอธิบายได้ ดังนี้

เมื่อผู้ส่ง (Sender) กับผู้รับ (Receiver) มีตัวอย่างข้อมูลสอนเหมือนกันทั้งคู่ แต่ตัวอย่างของฝ่ายส่งจะมีข้อมูลของการแบ่งพวก ส่วนฝ่ายรับจะไม่มีข้อมูลนี้ ผู้ส่งต้องติดต่อกับผู้รับเพื่อส่งข้อมูลเกี่ยวกับการแบ่งพวกให้กับผู้รับ โดยการส่งข้อมูลการแบ่งพวกทางทฤษฎี (Classification Theory) ไปพร้อมกับข้อยกเว้น (Exceptions) ของทฤษฎีนี้ ผู้ส่งสามารถจะเลือกระดับความซับซ้อนของทฤษฎีที่จะส่งได้ โดยการส่งทฤษฎีที่ง่ายซึ่งมีข้อยกเว้นมาก หรือส่งทฤษฎีที่ยากซึ่งมีข้อยกเว้นน้อย แต่สถานะของ MDL ที่ดีที่สุดจะใช้จำนวนบิตในการเข้ารหัส (Encode) ข้อมูลทั้งหมดของทฤษฎี และข้อยกเว้นรวมกันน้อยที่สุด

ข่าวสารซึ่งส่งให้กันในที่นี้เป็นข่าวสารที่บ่งบอกถึงข้อมูลของตัวอย่างสอนในแต่ละพวก C ที่ครอบคลุมโดยชุดของกฎ S หรือในที่นี้เรียกว่าต้นทุนของทฤษฎี (Theory Cost) กับความผิดพลาดในการระบุพวกของตัวอย่างโดยชุดของกฎ S หรือในที่นี้เรียกว่าต้นทุนความผิดพลาด (Exception Cost) โดยทั้งสองค่านี้จะถูกเข้ารหัสให้อยู่ในรูปของบิต ซึ่งชุดของกฎ S จะดีที่สุด เมื่อผลรวมของต้นทุนทางทฤษฎีกับต้นทุนความผิดพลาดมีค่าน้อยที่สุด

ในการคำนวณต้นทุนจะกระทำทีละพวกของ C โดยต้นทุนทางทฤษฎีของชุดของกฎ S จะเป็นผลรวมของต้นทุนทางทฤษฎีของกฎแต่ละกฎภายในชุดของกฎ S แล้วปรับลดด้วยจำนวนบิตที่ใช้ในการเข้ารหัสลำดับที่ในการเรียงของแต่ละกฎ โดยถ้ามีกฎอยู่ X กฎ จะมีการเรียงลำดับของกฎที่เป็นไปได้ X! กฎ ดังนั้นค่าที่ปรับลดจะเท่ากับ $\log_2(X!)$

ส่วนการคำนวณค่าต้นทุนความผิดพลาดของชุดของกฎ S จะอาศัยจำนวนตัวอย่างที่ผิดพลาดซึ่งครอบคลุมโดยกฎ S หรือเรียกว่าความผิดพลาดด้านบวก กับตัวอย่างที่ผิดพลาดซึ่งไม่ครอบคลุมโดยกฎ S หรือเรียกว่าความผิดพลาดด้านลบ โดยถ้าให้กฎ S ครอบคลุมตัวอย่าง r ตัวอย่าง จากตัวอย่างที่ใช้สอนทั้งหมด n ตัวอย่าง และมีจำนวนตัวอย่างที่ผิดพลาดด้านบวกเป็น fp และมีจำนวนตัวอย่างที่ผิดพลาดด้านลบเป็น fn ดังนั้นจำนวนบิตที่ใช้ในการเข้ารหัสความผิดพลาด จะเป็นดังนี้

$$\text{ค่าต้นทุนความผิดพลาด} = \log_2\left(\binom{r}{fp}\right) + \log_2\left(\binom{n-r}{fn}\right)$$

ในทางปฏิบัติค่าต้นทุนทางทฤษฎีมักจะมีค่าสูงกว่าความเป็นจริง เมื่อเทียบกับค่าต้นทุนความผิดพลาด เนื่องจากลักษณะต่าง ๆ ที่ใช้สอนแต่ละลักษณะจะมีความเหมือนหรือใกล้เคียงกัน ดังนั้นจึงเป็นไปได้ว่ากฎที่ต่างกันอาจจะให้ชุดตัวอย่างที่เหมือนกันได้ ทั้ง ๆ ที่ค่าต้นทุนทางทฤษฎี ต่างกัน (Quinlan และ Rivest, 1989) ดังนั้นจึงได้มีการปรับลดค่าต้นทุนทางทฤษฎี โดยใช้ค่า W ซึ่งเป็นค่าที่น้อยกว่า 1 ดังนี้

ค่าผลรวมของต้นทุน = ค่าต้นทุนความผิดพลาด + (W x ค่าต้นทุนทางทฤษฎี) บิต

ค่า W ที่เหมาะสมจะขึ้นอยู่กับความน่าจะเป็นที่กฎ 2 กฎ ครอบคลุมตัวอย่างเดียวกัน ซึ่งก็ขึ้นอยู่กับความซ้ำซ้อนของแต่ละลักษณะ โดยปกติ C4.5 จะใช้ค่า W เท่ากับ 0.5 ในการปรับลดค่าความซ้ำซ้อนนี้

จากตัวอย่างเงื่อนไขของการเป็นโรคไฮโปรไธรอยด์ ซึ่งประกอบด้วยตัวอย่างสอน 2514 ตัวอย่าง ถ้ามองถึงกฎที่แบ่งตัวอย่างออกเป็นพวก primary hypothyroid พบว่ามีด้วยกัน 3 กฎ ซึ่งถ้าให้หมายเลขประจำกฎเป็น 4, 5 และ 7 จะสามารถจัดกฎเหล่านี้ได้เป็น 8 ชุดที่เป็นไปได้ และคำนวณค่าต้นทุนต่าง ๆ ได้ดังตารางที่ 3.11

ทฤษฎี		ความผิดพลาด			ต้นทุนรวม
ชุดของกฎ	ต้นทุน	ด้านบวก	ด้านลบ	ต้นทุน	
-	0.0	0	64	425.8	425.8
{4}	17.1	2	12	116.8	125.3
{5}	19.8	1	6	63.9	73.8
{7}	15.7	1	61	411.8	419.6
{4, 5}	35.8	3	5	64.6	82.5
{4, 7}	31.7	3	9	97.8	113.6
{5, 7}	34.4	2	3	42.1	59.3
{4, 5, 7}	49.9	4	2	41.0	65.9

ตารางที่ 3.11 ค่าต้นทุนเมื่อจัดชุดของกฎในแบบต่าง ๆ (Quinlan, 1993)

จากตาราง 3.11 ในบรรทัดสุดท้ายซึ่งเป็นการเลือกกฎทั้ง 3 กฎพร้อมกัน ค่าต้นทุนทางทฤษฎีของทั้ง 3 กฎ จะหาได้จากผลรวมของต้นทุนทางทฤษฎีในแต่ละกฎรวมกัน แล้วลบด้วยจำนวนบิตที่ใช้ในการจัดลำดับของกฎ ดังนี้

ค่าต้นทุนทางทฤษฎีของกฎ {4, 5, 7} = 17.1 + 19.8 + 15.7 - $\log_2(3!)$

= 49.9 บิต

ส่วนต้นทุนความผิดพลาดคำนวณจากจำนวนตัวอย่างที่ผิดพลาดด้านบวก 4 ตัวอย่างจาก 66 ตัวอย่างที่ครอบคลุมโดยชุดกฎนี้ และจำนวนตัวอย่างที่ผิดพลาดด้านลบ 2 ตัวอย่าง จาก 2,448 ตัวอย่างที่ไม่ครอบคลุมโดยชุดของกฎนี้ ดังนี้

$$\text{ค่าต้นทุนความผิดพลาดของกฎ } \{4, 5, 7\} = \log_2\left(\binom{66}{4}\right) + \log_2\left(\binom{2448}{2}\right)$$

= 41.0 บิต

ค่าผลรวมต้นทุน = 41.0 + (0.5 x 49.9)

= 65.9 บิต

จากตารางที่ 3.11 จะเห็นได้ว่า การเลือกชุดของกฎ 5 และ 7 จะให้ต้นทุนรวมต่ำที่สุด คือ 59.3 บิต ดังนั้นในกฎที่แบ่งตัวอย่างเป็น primary hypothyroid จึงสามารถเลือกเพียงกฎที่ 5 และกฎที่ 7 เท่านั้นก็เพียงพอ

หลังจากเลือกชุดของกฎในแต่ละคลาสหรือพวกเรียบร้อยแล้ว ขั้นต่อไปคือการเรียงลำดับพวกของชุดของกฎที่ได้และการเลือกค่าโดยปริยายของพวก (Default Value) เพื่อกำหนดค่าให้กับตัวอย่างที่ไม่ถูกครอบคลุมโดยกฎใดกฎหนึ่งในชุดของกฎที่ได้

ชุดของกฎในแต่ละพวกที่ได้จะพบว่า มีค่าความผิดพลาดด้านบวก (False Positive Error) หรือมีจำนวนตัวอย่างที่ครอบคลุมโดยชุดของกฎนี้ แต่ไม่ใช่พวกเดียวกันกับชุดของกฎนี้ ซึ่งในการเรียงลำดับของพวกนั้นจะใช้ค่านี้ในการตัดสินใจ โดยชุดของกฎในพวกที่มีค่าความผิดพลาดด้านบวกน้อยที่สุดจะถูกเลือกก่อน และเลือกชุดของกฎในพวกที่มีค่าความผิดพลาดด้านบวกมากขึ้นเรียงลำดับไป

ในการเลือกค่าโดยปริยายของพวกให้กับตัวอย่างเมื่อตัวอย่างนี้ไม่ถูกครอบคลุมโดยกฎที่เลือกไว้ จะเลือกจากพวกที่มีจำนวนตัวอย่างสอนที่ไม่ครอบคลุมโดยกฎที่เลือกมามากที่สุด หรือในที่นี้จะดูได้จากค่าความผิดพลาดด้านลบ (False Negative Error) ของชุดกฎในแต่ละพวกที่มีค่ามากที่สุด

พวก	จำนวนกฎทั้งหมด	จำนวนกฎที่เลือก	จำนวนตัวอย่างที่ครอบคลุม	จำนวนตัวอย่างผิดพลาดด้านบวก	จำนวนตัวอย่างผิดพลาดด้านลบ
Negative	6	5	2319	2	3
Primary	3	2	66	2	3
Compensated	2	1	120	0	9

ตาราง 3.12 สรุปผลการเลือกชุดของกฎในแต่ละพวก (Quinlan, 1993)

จากตารางที่ 3.12 เป็นการสรุปผลจากการเลือกชุดของกฎในแต่ละพวกพบว่า พวก Compensated Hypothyroid มีค่าความผิดพลาดด้านบวกเท่ากับ 0 ซึ่งน้อยที่สุด ดังนั้นจึงถูกเลือกไว้เป็นอันดับแรกตามด้วย พวก Primary Hypothyroid และพวก Negative ตามลำดับ และยังพบว่าในพวก Compensated Hypothyroid มีจำนวนตัวอย่างสอน 9 ตัวอย่าง ซึ่งไม่ถูกครอบคลุมโดยกฎใด ๆ มากกว่าพวกอื่น ๆ ดังนั้น พวก Compensated Hypothyroid จึงถูกเลือกเป็นพวกปกติสำหรับชุดของกฎนี้ และได้เป็นชุดของกฎดังรูปที่ 3.11 เมื่อนำกฎนี้ไปทดสอบความถูกต้องบนตัวอย่างที่ไม่เคยเห็น 1,258 ตัวอย่าง พบว่ามีตัวอย่างที่จัดกลุ่มผิด 8 ตัวอย่าง หรือคิดเป็น (0.6 %) ซึ่งใกล้เคียงกับการทดสอบของต้นไม้ตัดสินใจ

```
    if  on thyroxine = f
        thyroid surgery = f
        TSH > 6
        TT4 ≤ 150
        FTI > 64
Then class compensated hypothyroid [98.9 %]
    If  thyroid surgery = f
        TSH > 6
        FTI ≤ 64
Then class primary hypothyroid [95.6 %]
    If  on thyroxine = f
        TT4 measured = f
        TSH > 6
Then class primary hypothyroid [45.3 %]
    If  TSH ≤ 6
Then class negative [99.9 %]
    If  on thyroxine = t
        FTI > 64
Then class negative [99.5 %]
    If  TSH measured = f
Then class negative [99.5 %]
    If  TT4 > 150
Then class negative [99.4 %]
    If  thyroid surgery = t
Then class negative [92.7 %]
    If  none of the above
Then class compensated hypothyroid
```

รูปที่ 3.11 สรุปกฎที่ได้จากตัวอย่างการเป็นโรคไฮโปรไธรอยด์ (Quinlan, 1993)