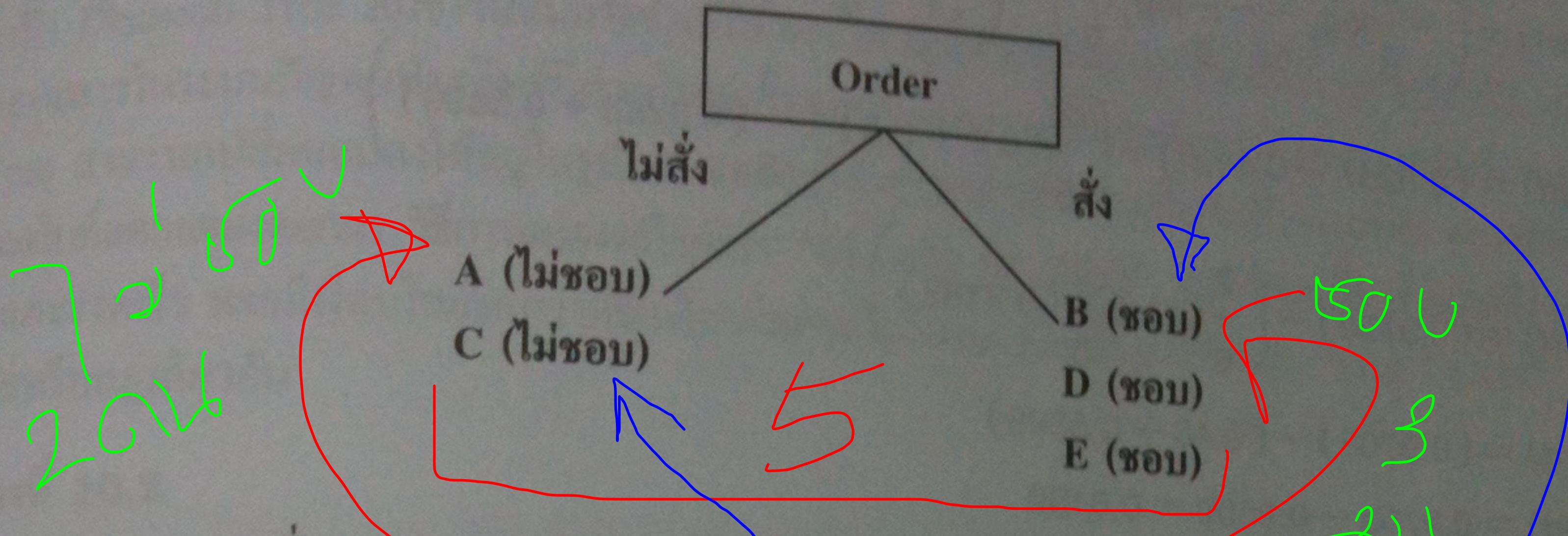


ตารางที่ 10.1 แสดงข้อมูลตัวอย่างการสั่งอาหารลูกค้าที่สั่งของหวาน

ลูกค้า	การสั่งของหวาน	ชนิดของหวาน	ความชอบ
A	ไม่สั่ง	ไอศกรีม	ไม่ชอบ
B	สั่ง	ไอศกรีม	ชอบ
C	ไม่สั่ง	ชนมเด็ก	ไม่ชอบ
D	สั่ง	ไอศกรีม	ชอบ
E	สั่ง	ชนมเด็ก	ชอบ

จากตารางที่ 10.1 ทำให้เราทราบความถี่ของข้อมูลที่สนใจซึ่งแบ่งออกเป็น 2 ส่วน คือ การสั่งของหวานของลูกค้า (Order) และชนิดของหวานที่ลูกค้าสั่ง (Dessert) ดังนั้นจึงต้องพิจารณาว่าข้อมูลใดสมควรเป็นโหนดราก (Root) ของ Decision Tree โดยจะให้ความสนใจว่าความชอบของลูกค้าเป็นอย่างไรและมีผลต่อการสั่งของหวานและชนิดของหวานอย่างไร ซึ่งจะต้องดำเนินการหาค่าของ Gain Function เพื่อนำมาเปรียบเทียบกันว่าข้อมูลใดเหมาะสมที่สุด ในการตัดสินใจนี้

การสั่งของหวานของลูกค้า (Order)



จากแบบจำลองการสั่งของหวานของลูกค้า เมื่อแทนใน Gain Function จะได้ดังนี้

$$E(S) = \left[\frac{2}{5} \log_2 \frac{2}{5} + \frac{3}{5} \log_2 \frac{3}{5} \right] \text{ (ได้จากค่าความน่าจะเป็นของความชอบของลูกค้า)}$$

ดั้งนี้จะได้

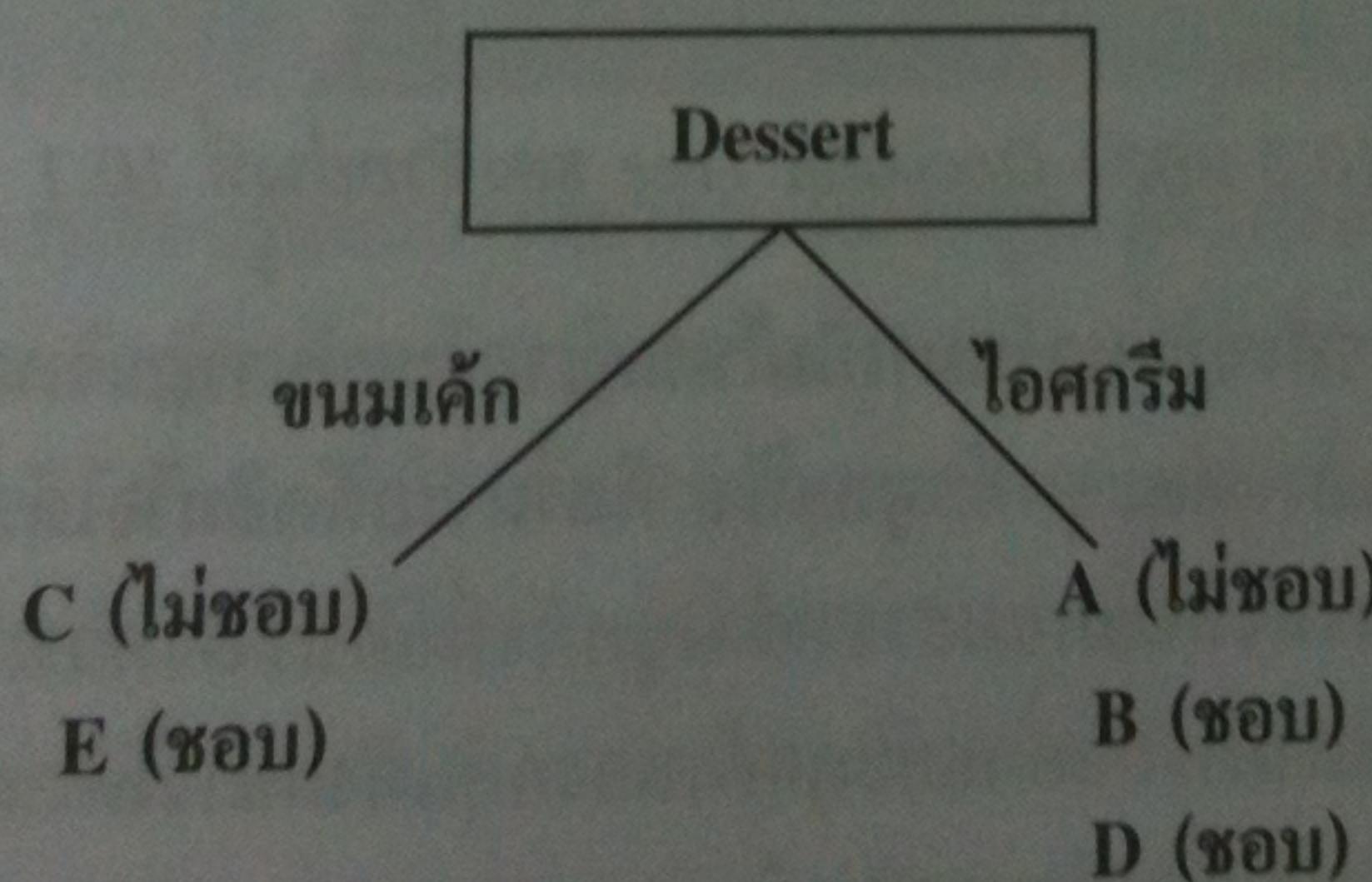
$$Gain(Order) = \left(-\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} \right) - \left(\frac{2}{5} \left(-\frac{2}{2} \log_2 \frac{2}{2} \right) + \frac{3}{5} \left(-\frac{3}{3} \log_2 \frac{3}{3} \right) \right)$$

$$Gain(Order) = \left(\frac{2}{5} \log_2 \frac{2}{5} + \frac{3}{5} \log_2 \frac{3}{5} \right) - \left(\frac{1}{2} \left(-\log_2 1 \right) + \frac{1}{2} \left(-\log_2 1 \right) \right)$$

$$Gain(Order) = (0.5287 + 0.4421) - \left(\frac{1}{2}(0) + \frac{1}{2}(0) \right)$$

$$Gain(Order) = (0.9708) - (0) = 0.9708$$

ชนิดของหวาน (Dessert)



จากแบบจำลองชนิดของหวาน เมื่อแทนใน Gain Function จะได้ดังนี้

$$E(S) = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} \quad (\text{ได้จากค่าความน่าจะเป็นของความชอบของลูกค้า})$$

ดังนั้นจะได้

$$Gain(Dessert) = \left(-\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} \right) - \left(\frac{2}{5} \left(-\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} \right) + \frac{3}{5} \left(-\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} \right) \right)$$

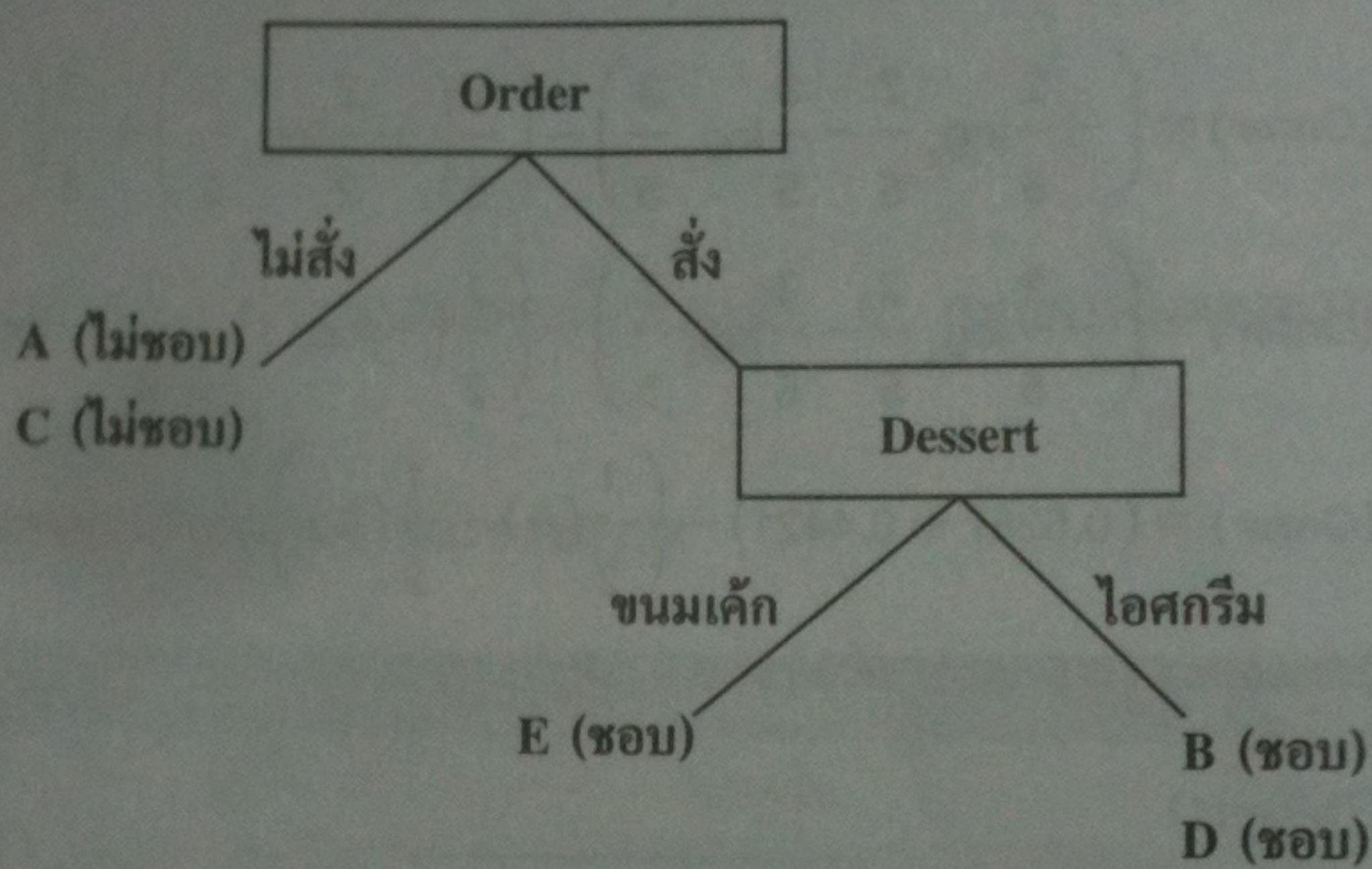
$$Gain(Dessert) = (0.5287 + 0.4421) - \left(\frac{2}{5} (0.5 + 0.5) + \frac{3}{5} (0.5283 + 0.3899) \right)$$

$$Gain(Dessert) = (0.9708) - \left(\frac{2}{5} (1) + \frac{3}{5} (0.9182) \right)$$

$$Gain(Dessert) = (0.9708) - (0.40 + 0.5509)$$

$$Gain(Dessert) = 0.9708 - 0.9509 = 0.0199$$

เมื่อพิจารณาค่าจาก Gain Function จากข้อมูลทั้งสองส่วนแล้ว ทำให้ทราบได้ว่าข้อมูลใดเหมาะสมที่จะเป็นโหนดราก ซึ่งค่า Gain(Order) นิมากกว่า Gain(Dessert) ดังนั้นจึงควรนำชุดข้อมูลของการสั่งของหวานมากำหนดเป็นโหนดราก เมื่อนำมาเขียนเป็น Decision Tree จะได้ดังรูปที่ 10.3



รูปที่ 10.3 แสดง Decision Tree ของตัวอย่างที่ 10.1

จากรูปจะเห็นได้ว่าเมื่อนำชุดข้อมูลการสั่งของหวานมาเป็นโหนดรากจะสามารถแยกแยะข้อมูลได้อย่างเหมาะสม ซึ่งจากตัวอย่างนี้หากพิจารณาการสร้างแบบจำลองของแต่ละชุดข้อมูลแล้ว ก็สามารถตัดสินใจได้ว่าชุดข้อมูลใดมีความเหมาะสม เนื่องจากแบบจำลองของชุดข้อมูลการสั่งของหวานจะแยกแยะระหว่างลูกค้าที่ชอบของหวานได้อย่างชัดเจนอยู่แล้ว แต่สำหรับแบบจำลองของชุดข้อมูลชนิดของหวานยังไม่สามารถแยกแยะลูกค้าสองกลุ่มดังกล่าวได้ ทั้งนี้ขึ้นอยู่กับความซับซ้อนของข้อมูล ด้วยว่ามีมากน้อยเพียงใด ในบางกรณีอาจจำเป็นต้องคำนวณหาค่าของ Gain Function ของโหนดที่อยู่ถัดไปจากโหนดเริ่มต้น เพื่อให้ได้โหนดที่มีความเหมาะสมที่สุด

การเปลี่ยน Decision Tree เป็นกฎ

ในการนำ Decision Tree ไปใช้งานนั้นอาจจำเป็นต้องนำข้อมูลดังกล่าวมาแปลงเป็นอีกรูปแบบหนึ่ง เพื่อให้สะดวกในการใช้งาน โดยเฉพาะในระบบปัญญาประดิษฐ์ที่ส่วนใหญ่การแทนของความรู้นั้นจะต้องอาศัยกฎสำหรับจำแนกข้อมูลของกัน Decision Tree สามารถเปลี่ยนแปลงให้อยู่ในรูปแบบของกฎได้ ซึ่งจะดำเนินการเริ่มต้นที่ไหนดราก และกำหนดกฎข้อเดียวเป็นตัวบ่งบอกว่าเกิดก扬ชนของกัน วิธีการแปลงจะใช้คำว่า “ถ้า...แล้ว (If...then)” มากกำหนดเป็นเงื่อนไขที่ช่วยให้สามารถพิจารณาข้อมูลได้ร่างและรวดเร็ว หากมีเงื่อนไขมากกว่าหนึ่งเงื่อนไขภายในโนดเดียวกันจะใช้คำว่า “และ (AND)” เป็นตัวเชื่อมข้อมูลให้เป็นเงื่อนไขหรือกฎเดียวกัน

ตัวอย่างที่ 10.2

จากตัวอย่างที่ 10.1 ซึ่งมี Decision Tree เป็นดังรูปที่ 10.3 สามารถเปลี่ยนเป็นกฎได้ดังนี้

(1) ถ้า ลูกค้าไม่สั่งของหวาน

แล้ว แสดงว่าลูกค้าไม่ชอบทานของหวาน

(2) ถ้า ลูกค้าสั่งของหวาน และสั่งขนมเค้ก

แล้ว แสดงว่าลูกค้าชอบทานของหวาน

(3) ถ้า ลูกค้าสั่งของหวาน และสั่งไอศกรีม

แล้ว แสดงว่าลูกค้าชอบทานของหวาน

Rule
for Coding

จากที่กล่าวมาจะเห็นได้ว่า Decision Tree สามารถนำเปลี่ยนแปลงเป็นกฎได้ เมื่อมีเหตุการณ์ใดๆ ก็ตามที่เกิดขึ้นในแต่ละช่วงเวลาการจำแนกข้อมูลและทราบผลลัพธ์ได้ ซึ่งเป็นการทั้งนารูปแบบการเรียนรู้และการตัดสินใจของระบบได้เป็นอย่างดี โดยอาศัยจากกฎที่ถูกสร้างขึ้นอย่างสอดคล้องกับ Decision Tree สำหรับกฎที่ได้จะถูกนำไปใช้งานในด้านต่อๆ ตามที่ต้องการท่องไป

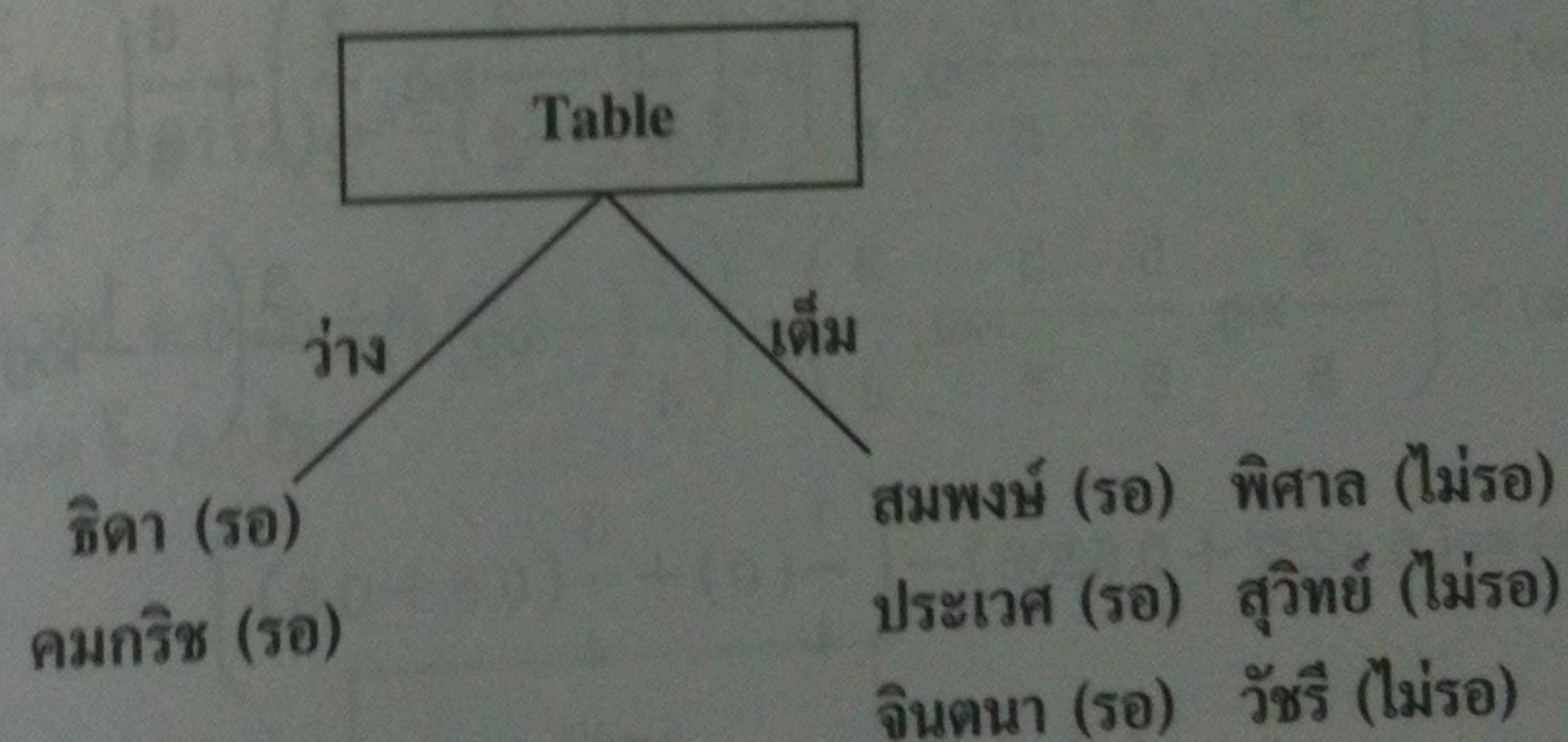
ตัวอย่างที่ 10.3

ตารางที่ 10.2 แสดงข้อมูลดิบของการสำรวจความต้องการของลูกค้าที่รอนั้นใช้บริการ

ลูกค้า	สถานะของโต๊ะ	ราคา	ร้านอาหาร	ระยะเวลา (นาที)	ลูกค้ารอ
สมพงษ์	เต็ม	แพง	ญี่ปุ่น	10-30	รอ
จินดา	เต็ม	ถูก	ไทย	10-30	รอ
พิศาล	เต็ม	แพง	ไทย	10-30	ไม่รอ
สุวิทย์	เต็ม	แพง	ญี่ปุ่น	มากกว่า 30	ไม่รอ
ประเวศ	เต็ม	แพง	อิตาเลี่ยน	0-10	รอ
ธิดา	ว่าง	ถูก	ไทย	10-30	รอ
วัชรี	เต็ม	แพง	อิตาเลี่ยน	10-30	ไม่รอ
คมกริช	ว่าง	แพง	ไทย	0-10	รอ

จากตารางที่ 10.2 ทำให้เราทราบว่าลูกค้าแต่ละคนมีความต้องการและพฤติกรรมในสถานการณ์หนึ่งๆ ที่แตกต่างกัน จะให้ความสนใจกับพฤติกรรมการรอของลูกค้าว่าจะสามารถรอเข้าใช้บริการร้านค้านั้นได้หรือไม่ในปัจจัยที่แตกต่างกัน ซึ่งก่อให้เป็นเชิงข้อมูลที่จำเป็นต้องพิจารณาในการคำนวณหาค่า Gain Function ได้แก่ สถานะของโต๊ะ (Table) ราคา (Price) ประเภทร้านอาหาร (Type) และระยะเวลาในการรอ (Time) และแสดงเป็น Decision Tree ได้ดังนี้

สถานะของโต๊ะ (Table)



จากแบบจำลองสถานะของโต๊ะ เมื่อแทนใน Gain Function จะได้ดังนี้

$$E(S) = -\frac{5}{8} \log_2 \frac{5}{8} - \frac{3}{8} \log_2 \frac{3}{8}$$

7.75 รอบ 30%

ดังนั้นจะได้

รอบ 50%

รอบ 80%

$$Gain(Table) = \left(-\frac{5}{8} \log_2 \frac{5}{8} - \frac{3}{8} \log_2 \frac{3}{8} \right) - \left(\frac{2}{8} \left(-\frac{2}{2} \log_2 \frac{2}{2} \right) + \frac{6}{8} \left(-\frac{3}{6} \log_2 \frac{3}{6} - \frac{3}{6} \log_2 \frac{3}{6} \right) \right)$$

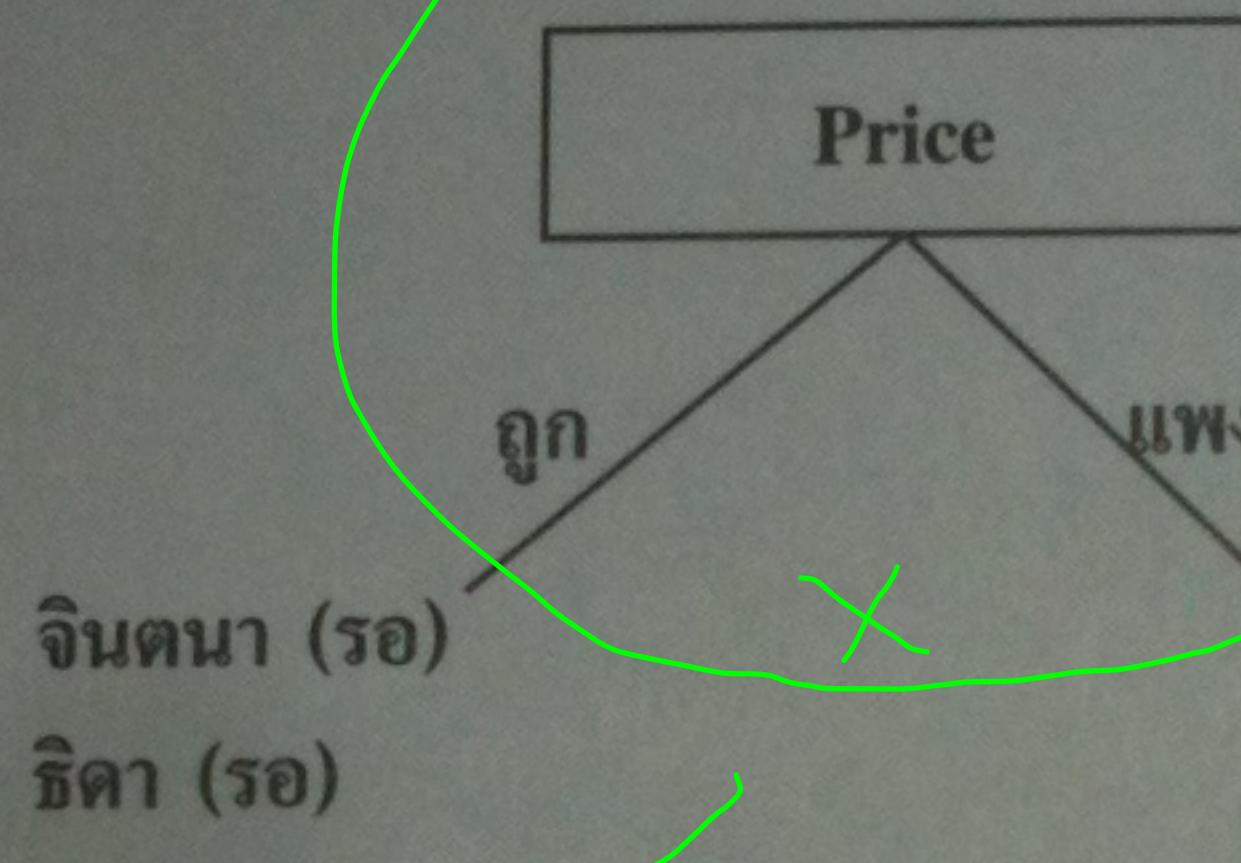
$$Gain(Table) = \left(-\frac{5}{8} \log_2 \frac{5}{8} - \frac{3}{8} \log_2 \frac{3}{8} \right) - \left(\frac{1}{4} (-\log_2 1) + \frac{3}{4} \left(-\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} \right) \right)$$

$$Gain(Table) = (0.4237 + 0.5306) - \left(\frac{1}{4}(0) + \frac{3}{4}(0.5 + 0.5) \right)$$

$$Gain(Table) = (0.9543) - \left(\frac{3}{4}(1) \right)$$

$$Gain(Table) = 0.9543 - 0.75 = 0.2043$$

ราคาอาหาร (Price)



$$[0.4237 - 0.903] / 0.301$$

$$= 0.265 / 0.301$$

$$= 0.881$$

$$\text{ประเวศ (รอ) สุวิทย์ (ไม่รอ)}$$

$$\text{คุณกริช (รอ) วัชรี (ไม่รอ)}$$

จากแบบจำลองราคาอาหาร เมื่อแทนใน Gain Function จะได้ดังนี้

$$E(S) = -\frac{5}{8} \log_2 \frac{5}{8} - \frac{3}{8} \log_2 \frac{3}{8}$$

ดังนั้นจะได้

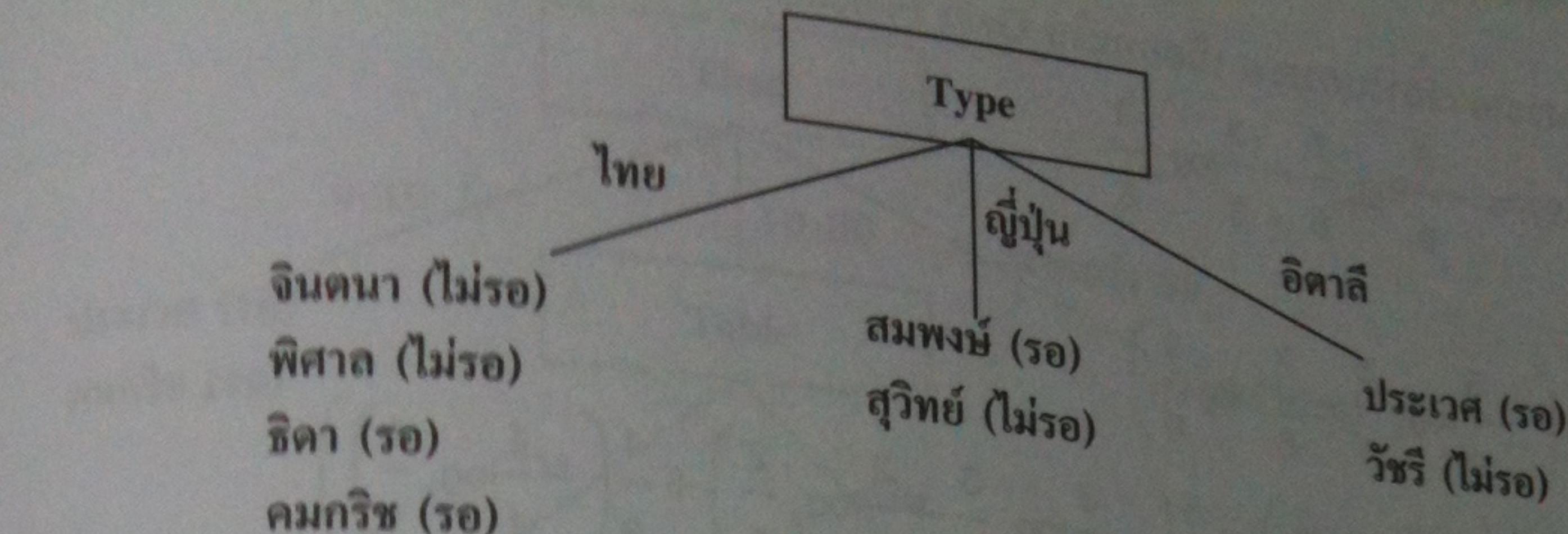
$$Gain(Price) = \left(-\frac{5}{8} \log_2 \frac{5}{8} - \frac{3}{8} \log_2 \frac{3}{8} \right) - \left(\frac{2}{8} \left(-\frac{2}{2} \log_2 \frac{2}{2} \right) + \frac{6}{8} \left(-\frac{3}{6} \log_2 \frac{3}{6} - \frac{3}{6} \log_2 \frac{3}{6} \right) \right)$$

$$Gain(Price) = \left(-\frac{5}{8} \log_2 \frac{5}{8} - \frac{3}{8} \log_2 \frac{3}{8} \right) - \left(\frac{1}{4} (-\log_2 1) + \frac{3}{4} \left(-\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} \right) \right)$$

$$Gain(Price) = (0.4237 + 0.5306) - \left(\frac{1}{4}(0) + \frac{3}{4}(0.5 + 0.5) \right)$$

$$Gain(Price) = (0.9543) - \left(\frac{3}{4}(1) \right)$$

$$Gain(Price) = 0.9543 - 0.75 = 0.2043$$



จากแบบจำลองประเกตว์นอาหาร เมื่อแทนใน Gain Function จะได้ดังนี้

$$E(S) = -\frac{5}{8} \log_2 \frac{5}{8} - \frac{3}{8} \log_2 \frac{3}{8}$$

ตั้งนั้นจะได้

$$Gain(Type) = \left(-\frac{5}{8} \log_2 \frac{5}{8} - \frac{3}{8} \log_2 \frac{3}{8} \right)$$

$$-\left(\frac{4}{8} \left(-\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} \right) + \frac{2}{8} \left(-\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} \right) + \frac{2}{8} \left(-\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} \right) \right)$$

$$Gain(Type) = \left(-\frac{5}{8} \log_2 \frac{5}{8} - \frac{3}{8} \log_2 \frac{3}{8} \right)$$

$$-\left(\frac{1}{2} \left(-\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} \right) + \frac{1}{4} \left(-\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} \right) + \frac{1}{4} \left(-\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} \right) \right)$$

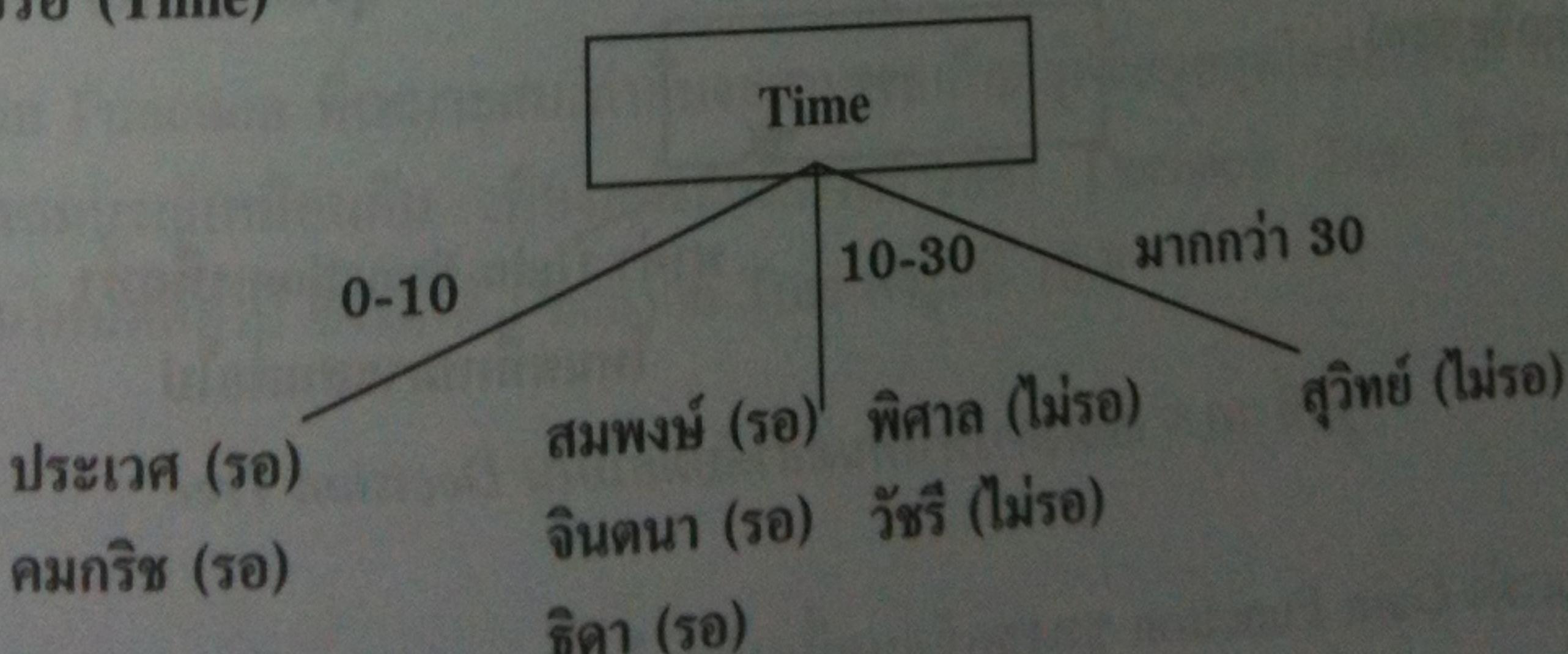
$$Gain(Type) = (0.4237 + 0.5306) - \left(\frac{1}{2}(0.3112 + 0.5) + \frac{1}{4}(0.5 + 0.5) + \frac{1}{4}(0.5 + 0.5) \right)$$

$$Gain(Type) = (0.9543) - \left(\frac{1}{2}(0.8112) + \frac{1}{4}(1) + \frac{1}{4}(1) \right)$$

$$Gain(Type) = 0.9543 - (0.4056 + 0.25 + 0.25)$$

$$Gain(Type) = 0.9543 - 0.9056 = 0.0487$$

ระยะเวลาในการรอ (Time)



หากแทนข้อมูลระยะเวลาในการรอ เมื่อแทนใน Gain Function จะได้ดังนี้

$$E(S) = -\frac{5}{8} \log_2 \frac{5}{8} - \frac{3}{8} \log_2 \frac{3}{8}$$

แล้วนั้นจะได้

$$\text{Gain}(Time) = \left(-\frac{5}{8} \log_2 \frac{5}{8} - \frac{3}{8} \log_2 \frac{3}{8} \right) - \left(\frac{2}{8} \left(-\frac{2}{2} \log_2 \frac{2}{2} \right) + \frac{5}{8} \left(-\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} \right) + \frac{1}{8} \left(-\frac{1}{1} \log_2 \frac{1}{1} \right) \right)$$

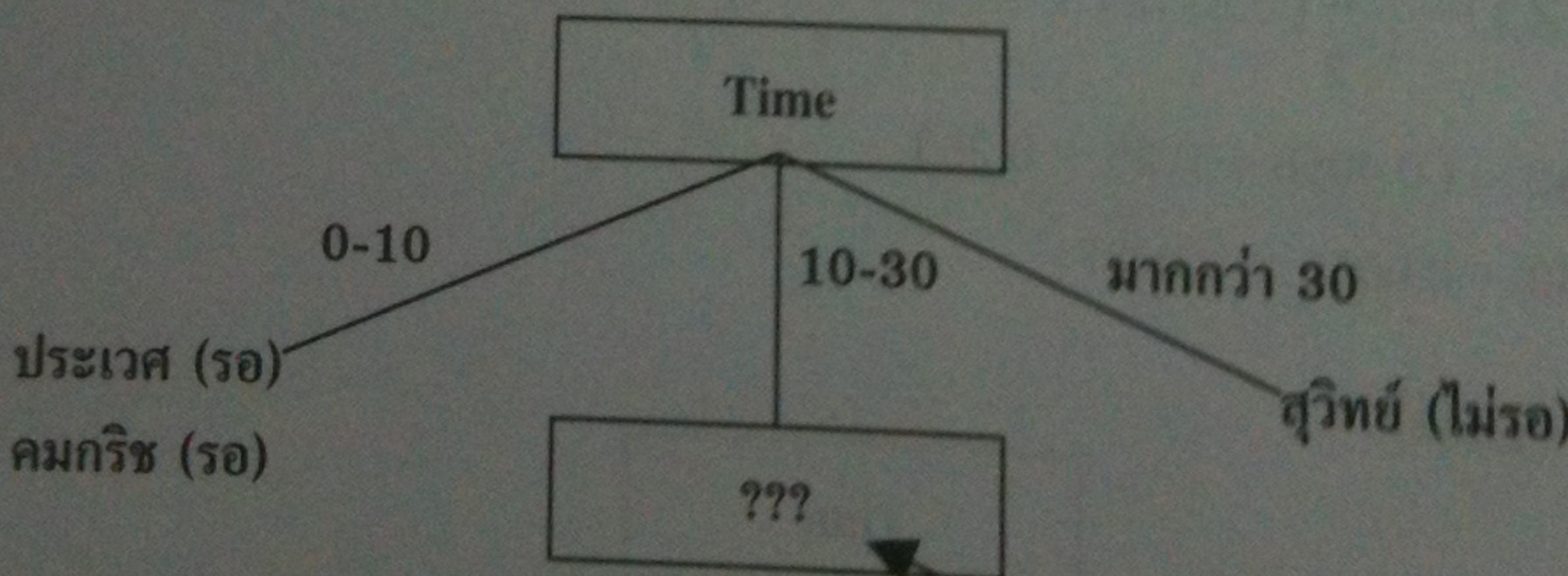
$$\text{Gain}(Time) = \left(-\frac{5}{8} \log_2 \frac{5}{8} - \frac{3}{8} \log_2 \frac{3}{8} \right) - \left(\frac{1}{4} (-\log_2 1) + \frac{5}{8} \left(-\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} \right) + \frac{1}{8} (-\log_2 1) \right)$$

$$\text{Gain}(Time) = (0.4237 + 0.5306) - \left(\frac{1}{4}(0) + \frac{5}{8}(0.4421 + 0.5287) + \frac{1}{8}(0) \right)$$

$$\text{Gain}(Time) = (0.9543) - \left(\frac{5}{8}(0.9708) \right)$$

$$\text{Gain}(Time) = 0.9543 - 0.6067 = 0.3476$$

จากการคำนวณหาค่า Gain Function ของแต่ละชุดข้อมูลจะได้ค่าที่แตกต่างกันไป โดยค่าที่มากที่สุดจะสามารถแบ่งออกได้ว่าชุดข้อมูลนั้นมีความเหมาะสมในการนำมาเป็นโหนดเริ่มต้นในการสร้าง Decision Tree ซึ่งก็คือ Gain(Time) ที่มีค่ามากที่สุดจากนั้นจึงพิจารณาโหนดที่เหมาะสมในอันดับถัดไปด้วยการคำนวณหา Gain Function เช่นเดียวกับขั้นตอนที่ผ่านมาแล้ว แต่จะพิจารณาเพียงโหนดที่เหลือซึ่งจะเชื่อมต่อกับโหนดเริ่มต้น โดยพิจารณาจากกิ่งก้านที่ยังมีความชันช้อนของข้อมูลหรือยังจำแนกได้ไม่หมด ก่าวคือ โหนดนั้นยังต้องแตกกิ่งก้านต่อไปเพื่อแยกข้อมูลที่แตกต่างกันออกเป็นกลุ่มที่เหมาะสม (ดูรูปที่ 10.5 โหนดตรงกลางยังมีความต้องการขั้งสุดท้าย “รอ” และ “ไม่รอ บันกันอยู่”)

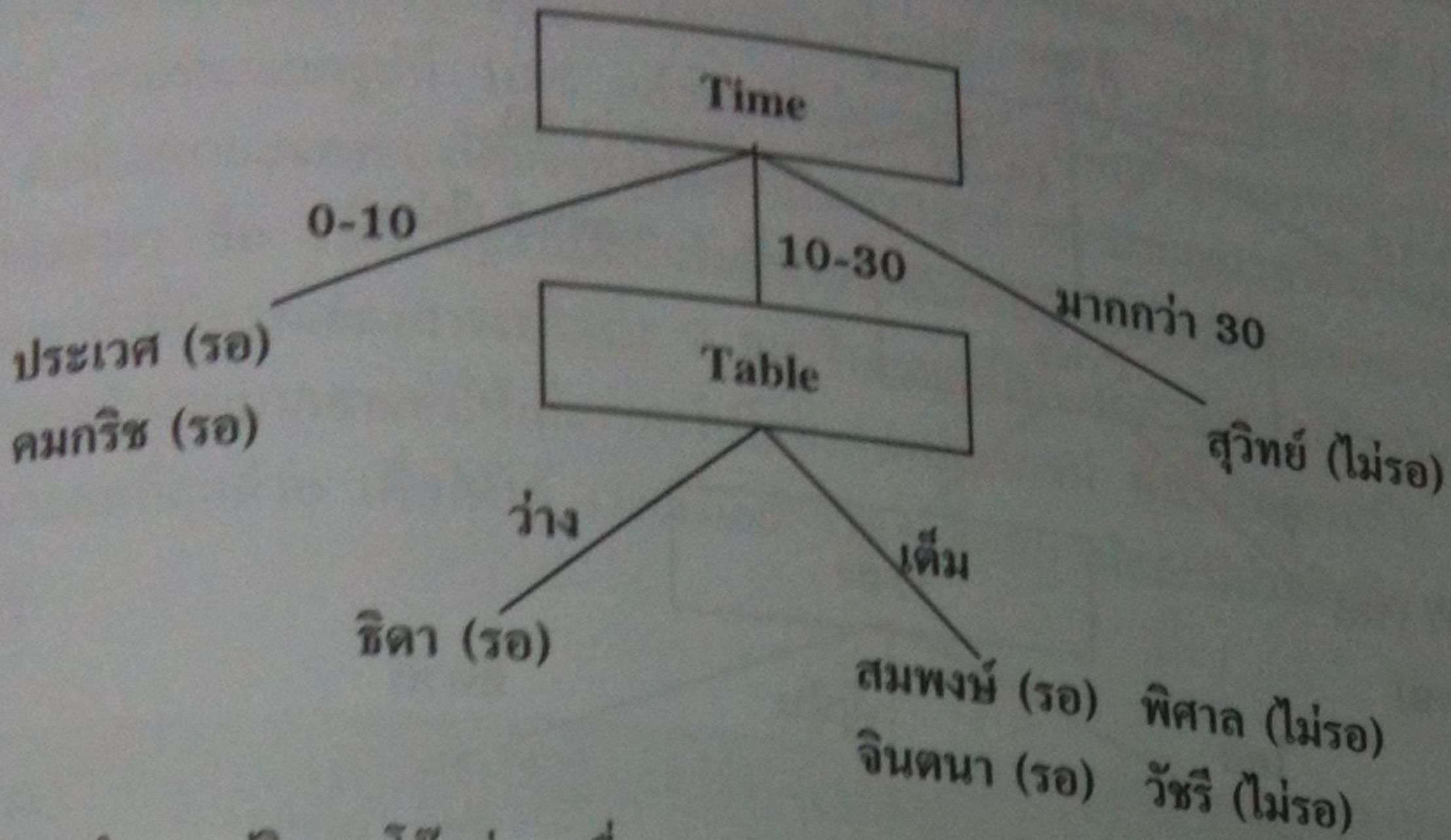


หาค่า Gain Function เพื่อหา
โหนดที่เหมาะสมต่อไป

รูปที่ 10.5 แสดงการหาโหนดลำดับต่อไปใน Decision Tree

ดังนั้นจึงต้องพิจารณาค่า Gain Function ของชุดข้อมูลที่เหลือ ดังนี้

สถานะของโต๊ะ (Table)



จากแบบจำลองปริมาณโต๊ะว่าง เมื่อแทนใน Gain Function จะได้ดังนี้

$$E(S) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5}$$

ดังนั้นจะได้

$$Gain(Table) = \left(-\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} \right) - \left(\frac{1}{5} \left(-\frac{1}{1} \log_2 \frac{1}{1} \right) + \frac{4}{5} \left(-\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} \right) \right)$$

$$Gain(Table) = \left(-\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} \right) - \left(\frac{1}{5} (-\log_2 1) + \frac{4}{5} \left(-\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} \right) \right)$$

$$Gain(Table) = (0.4421 + 0.5257) - \left(\frac{1}{5}(0) + \frac{4}{5}(0.5 + 0.5) \right)$$

$$Gain(Table) = (0.9708) - \left(\frac{4}{5}(1) \right)$$

$$Gain(Table) = 0.9708 - 0.80 = 0.1708$$

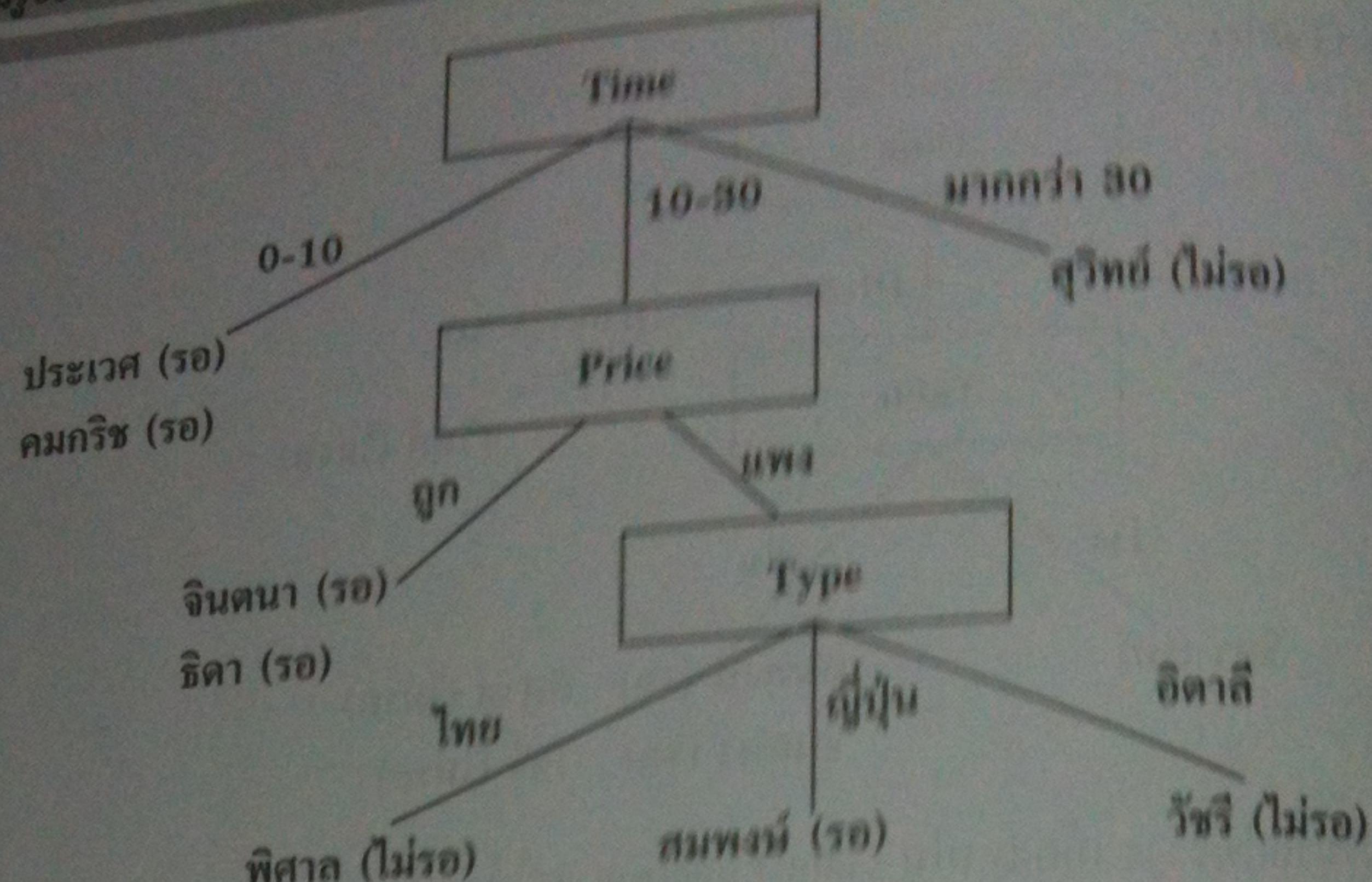
นอกจากชุดข้อมูลปริมาณโต๊ะว่างแล้วจำเป็นต้องคำนวณค่า Gain Function ของชุดข้อมูลอื่นด้วย (Price, Type) ซึ่งได้ค่าดังนี้

$$Gain(Table) = 0.1708$$

$$Gain(Price) = 0.4199$$

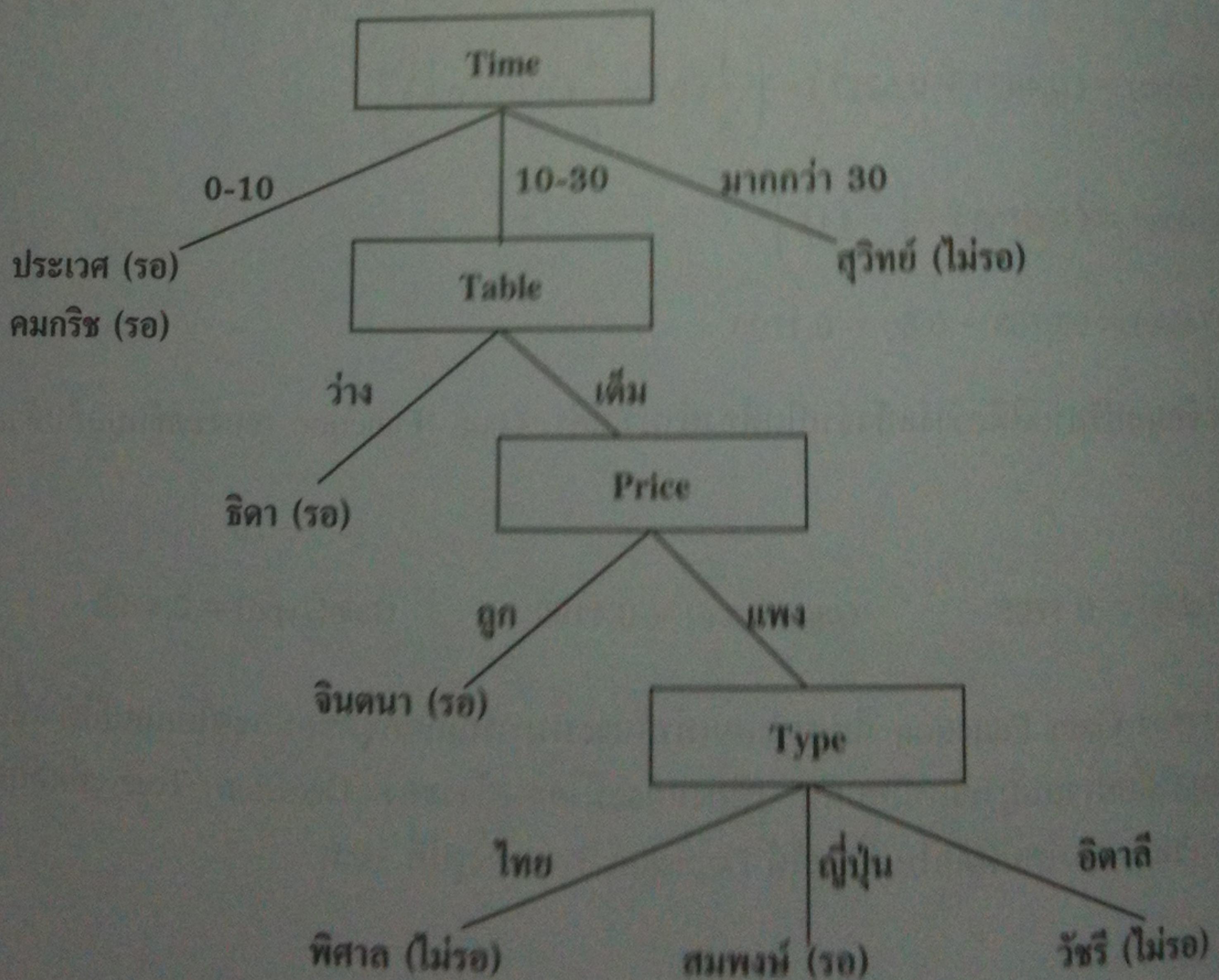
$$Gain(Type) = 0.4199$$

ในการนี้ที่มีค่าของ Gain Function ที่เหมาะสมเท่ากันจะสามารถเลือกชุดข้อมูลชุดใดก่อนก็ได้ เนื่องจากชุดข้อมูลทั้งสองสามารถจำแนกข้อมูลได้อย่างสมบูรณ์เหมือนกัน เพียงแต่จะมีโครงสร้างของ Decision Tree ที่แตกต่างกันเท่านั้น สำหรับทั้งยังนี้จะเลือก Price เป็นโหนดต่อไป ซึ่งจะได้ Decision Tree ดังรูปที่ 10.6



Section 10.6 uses Decision Tree learning method 10.3

จากรูปที่ 10.6 จะเป็น Decision Tree ที่มีความซับซ้อนนิ่ว เป็นรูปการสอนยกตัวอย่างข้อมูลที่สนใจของคนหุ้น
ทั้งหมดได้อย่างชัดเจน ถึงแม้ว่าจะมีหน่วยไม่ครบถ้วนก็ตาม เนื่องจาก Decision Tree ได้ไม่จำเป็นต้องนำหน่วย Table มาพิจารณาด้วย ด้วยเหตุนี้จึงมีความซับซ้อนน้อยกว่า Decision Tree ที่นำหน่วย Table มาร่วมพิจารณา ในความเป็นจริงอาจมี Decision Tree ที่สอดคล้องกับทุกข้อมูลทั้งหมดได้มากกว่ารูปแบบ รูปที่ 10.6
Decision Tree ที่ง่ายต่อการพิจารณาข้อมูลหรืออาจชี้บันทึกว่า Decision Tree เดิมที่ได้ ดังรูปที่ 10.7



รูปที่ 10.7 แสดงตัวอย่าง Decision Tree สำหรับปัญหาที่ต้องตัดสินใจซื้อบ้านหรือไม่ ตามตัวอย่างที่ 10.3

ผู้อย่างที่ 10.4

กำหนดให้มีการจำแนกสัตว์ 5 ชนิด จำแนกจากคุณสมบัติ ที่สามารถแบ่งออกถึงสายพันธุ์ได้ ดังนี้

	ออกฤกษ์เป็นนา	ออกฤกษ์เป็นไฟ	เลี้ยงฤกษ์หวานรำ	รำหวานหวานรำ	ผิวหนานหิรือเมล็ด	สัตว์เดือดอุ่น	สัตว์เดือดเย็น	อยู่ทึ่ในน้ำและบนก	หายใจด้วยเหงือก	สายพันธุ์
กบ	1					1	1			สัตว์ครึ่งบกครึ่งน้ำ
เป็ด	1	1	1	1	1		1	1		สัตว์จำพวกนก
ด้วงคาว	1	1	1	1						สัตว์เลี้ยงลูกด้วยนม
ช้าง					1		1			สัตว์เลือดคลาน
ปลาแซลมอน	1				1		1		1	สัตว์จำพวกปลา

นิชุดข้อมูลของสัตว์ชนิดหนึ่งที่ทราบเพียงข้อมูลด้านคุณสมบัติ ดังนี้

	ออกฤกษ์เป็นนา	ออกฤกษ์เป็นไฟ	เลี้ยงฤกษ์หวานรำ	รำหวานหวานรำ	ผิวหนานหิรือเมล็ด	สัตว์เดือดอุ่น	สัตว์เดือดเย็น	อยู่ทึ่ในน้ำและบนก	หายใจด้วยเหงือก	สายพันธุ์
สัตว์ปริศนา		1						1		???

ต้องการทราบว่าสัตว์ชนิดนี้เป็นสัตว์ที่ใกล้เคียงกับสายพันธุ์ใดมากที่สุด (1NN)

จากชุดข้อมูลของสัตว์ปริศนา จะนำมาใช้หาค่าระยะทางของชุดข้อมูลสัตว์ทั้ง 5 ชนิด ด้วยสมการ Euclidean เมื่อแทนค่าในแต่ละชุดข้อมูลจะได้ดังนี้

ระยะทางของชุดข้อมูลกบ (d_1)

$$d_1 = \sqrt{(0)^2 + (1-1)^2 + (0)^2 + (0)^2 + (0)^2 + (0)^2 + (0-1)^2 + (1-1)^2 + (0)^2} \\ = \sqrt{1} = 1$$

ระยะทางของชุดข้อมูลเป็ด (d_2)

$$d_2 = \sqrt{(0)^2 + (1-1)^2 + (0)^2 + (0-1)^2 + (0)^2 + (0-1)^2 + (0)^2 + (1-1)^2 + (0)^2} \\ = \sqrt{1+1} = \sqrt{2} = 1.414$$

ระยะทางของชุดข้อมูลค้างคาว (d_3)

$$d_3 = \sqrt{(0-1)^2 + (1-0)^2 + (0-1)^2 + (0-1)^2 + (0)^2 + (0)^2 + (0)^2 + (0)^2 + (1-0)^2 + (0)^2} \\ = \sqrt{1+1+1+1+1} = \sqrt{5} = 2.236$$

ระยะทางของชุดข้อมูล (d_4)

$$d_4 = \sqrt{(0)^2 + (1-1)^2 + (0)^2 + (0)^2 + (0-1)^2 + (0)^2 + (0-1)^2 + (1-0)^2 + (0)^2} \\ = \sqrt{1+1+1} = \sqrt{3} = 1.732$$

ระยะทางของชุดข้อมูลปลาแซลมอน (d_5)

$$d_5 = \sqrt{(0)^2 + (1-1)^2 + (0)^2 + (0)^2 + (0-1)^2 + (0)^2 + (0-1)^2 + (1-0)^2 + (0-1)^2} \\ = \sqrt{1+1+1+1} = \sqrt{4} = 2$$

จากค่าของระยะทางที่ได้จากชุดข้อมูลของสัตว์ทั้ง 5 ชนิด ระยะทางที่มีค่าน้อยที่สุด คือ d_4
แสดงว่าสัตว์ปริศนามีข้อมูลที่ใกล้เคียงกับชุดข้อมูลของกบ
จึงสรุปได้ว่าสัตว์ปริศนาชนิดนี้มีสายพันธุ์เป็น สัตว์ครึ่งบกครึ่งน้ำ

จากตัวอย่างที่ 10.4 เป็นการคำนวณหาระยะทางที่ใกล้มากที่สุด ก็รู้คือ จะมีชุดข้อมูลซึ่งเดียวกันที่มีระยะห่างน้อยที่สุด
เพิ่มมากขึ้นชุดข้อมูลเท่านั้น ในกรณีที่ต้องการชุดข้อมูลที่มีค่าใกล้เคียงมากกว่าหนึ่งชุดข้อมูล หรือ k ชุด การจำแนกทางชุดข้อมูล
ประเภทนี้จะเรียกว่า “ k -NN” ซึ่งแสดงดังตัวอย่างต่อไปนี้

ตัวอย่างที่ 10.5

กำหนดให้มีการจำแนกสัตว์ 5 ชนิด จำแนกจากคุณสมบัติ ที่สามารถอภิปรายได้ ดังนี้

	สายพันธุ์	หมายเหตุความชื้น	อยู่ในน้ำ	สัตว์เลื้อยคลาน	สัตว์เดินบนดิน	ผิวหนังแห้ง	มนุษย์	เลี้ยงลูกด้วยนม	ออกลูกเป็นไข่	กบ
เป็ด	สัตว์ครึ่งบกครึ่งน้ำ				1	1			1	1
ค้างคาว	สัตว์จำพวกกบ		1		1					1
ปลาแซลมอน	สัตว์เลี้ยงถูกด้วยนม		1							1
ธรรมชาติ	สัตว์เลื้อยคลาน			1		1				1
แมว	สัตว์จำพวกปลา			1		1				1
นกกระจองเทศ	สัตว์เลื้อยคลาน		1		1					1
	สัตว์เลี้ยงถูกด้วยนม									1
	สัตว์จำพวกกบ									1

มีชุดข้อมูลของสัตว์ชนิดหนึ่งที่ทราบเพียงข้อมูลด้านคุณสมบัติ ดังนี้

สัตว์ปริศนา	ออกฤกษ์เป็นตัว	ออกฤกษ์เป็นไข่	เลี้ยงลูกด้วยนม	มนุษย์ตามธรรมชาติ	ผิวหายานหรือไม่เกิด	สัตว์เลือดอ่อน	สัตว์เลือดเย็น	อยู่ทึ่งในน้ำและบนบก	หายใจด้วยเหงือก	สายพันธุ์
		1	1		1					???

ต้องการทราบว่าสัตว์ชนิดนี้เป็นสัตว์ที่ใกล้เคียงกับสายพันธุ์ใดบ้างอย่างน้อย 3 ชนิด (3NN) หากค่าระยะทางของแต่ละชุดข้อมูล (เหมือนตัวอย่างที่ 10.4) จะได้ผลลัพธ์ดังนี้

$$\text{กบ} (d_1) = 2.449$$

$$\text{ปลาแซลมอน} (d_5) = 2.645$$

$$\text{เป็ด} (d_2) = 1.732$$

$$\text{จะเขี้ยว} (d_6) = 2.236$$

$$\text{ค้างคาว} (d_3) = 1.414$$

$$\text{แมว} (d_7) = 1$$

$$\text{งู} (d_4) = 2.236$$

$$\text{นกกระจองเทศ} (d_8) = 1.414$$

จากค่าระยะทางที่ได้ของทั้ง 8 ชุดข้อมูล จะเลือกค่าระยะทางที่น้อยที่สุด 3 อันดับ

ได้แก่ แมว (d_7) = 1, ค้างคาว (d_3) = 1.414 และ นกกระจองเทศ (d_8) = 1.414

ดังนั้นสัตว์ปริศนาดังกล่าวมีความใกล้เคียงกับสายพันธุ์ สัตว์เลี้ยงลูกด้วยนมและสัตว์จำพวกนก แต่ว่า 3 ชุดข้อมูลที่ได้มามี 2 ชุดข้อมูลที่มีสายพันธุ์ตรงกัน จึงสามารถสรุปจากเสียงข้างมากได้ว่า สัตว์ปริศนาชนิดนี้มีสายพันธุ์เป็น สัตว์เลี้ยงลูกด้วยนม

จะเห็นว่า Decision Tree และ Nearest Neighbor Classification เป็นเทคนิคการเรียนรู้ของเครื่องจักรที่แตกต่างกัน โดย Decision Tree จะเป็นวิธีการจำแนกเพื่อหาความถูกต้องของข้อมูลที่สนใจ โดยที่ทราบค่าของชุดข้อมูลทั้งหมด ซึ่งเป็นไปตามหลักการของ Supervised Learning ส่วน Nearest Neighbor จะพิจารณาจากการนำค่าของชุดข้อมูล ซึ่งเป็นวิธีการแบบ Unsupervised Learning โดยให้ความสำคัญกับความสัมพันธ์ของชุดข้อมูลแทนการคำนวณหาค่าความน่าจะเป็นจากข้อมูลทั้งหมด ดังนั้นจึงต้องพิจารณาชุดข้อมูลที่สนใจว่าควรเลือกใช้วิธีการจำแนกกลุ่มแบบใดจึงจะเหมาะสมที่สุด