การหารูปแบบของปรากฏการณ์โดยใช้ต้นไม้การตัดสินใจ: กรณีศึกษา อุบัติเหตุการจราจรในจังหวัดนครปฐม

Finding Phenomenon Patterns using Decision Trees: A Case Study of Traffic Accidents in Nakorn Pathom

สุนิดา ศรีสุริยะชัย และ รังสิพรรณ มฤคทัต

คณะวิศวกรรมศาสตร์ สาขาวิชาเทคโนโลยีการจัดการสารสนเทศ มหาวิทยาลัยมหิดล

g4736983@student.mahidol.ac.th, egrmr@mahidol.ac.th

บทคัดย่อ

บทความนี้เสนอการประยุกต์ใช้ต้นไม้การตัดสินใจเพื่อวิเคราะห์ เปรียบเทียบรูปแบบของปรากฏการณ์ที่เกิดขึ้นในกลุ่มตั้งแต่สองกลุ่มขึ้นไป โดยใช้อุบัติเหตุการจราจรในจังหวัดนครปฐมเป็นกรณีศึกษา เทคนิคการสร้างต้นไม้การตัดสินใจที่นำเสนอในบทความนี้คือ C4.5 ต้นไม้การตัดสินใจที่ได้จะนำไปสร้างเป็นโพรไฟล์ของอุบัติเหตุการจราจรที่เกิดขึ้นในจังหวัดนครปฐม สำหรับการวิเคราะห์ปัญหา ตลอดจนการหามาตรการแก้ไขต่อไป

คำสำคัญ ต้นไม้การตัดสินใจ อุบัติเหตุการจราจร

Abstract

This article presents our experience of using decision trees for finding phenomenon patterns in multiple groups. Traffic accidents in Nakorn Pathom is used as a case study. C4.5 is the main focus of this article. The resulting decision trees will be used to construct Nakorn Pathom's traffic accident profiles, which serve as a stepping-stone for problem analysis and suggestion of countermeasures.

Key Words: decision trees, traffic accidents

1. บทนำ

บทความนี้เป็นส่วนหนึ่งของงานวิจัยเชิงประยุกต์ เพื่อสร้างโพรไฟล์ของอุบัติเหตุการจราจร (traffic accident profiles) ที่เกิดขึ้นในจังหวัดนครปฐม โดยนำเทคนิคการทำเหมืองข้อมูลมาประยุกต์ใช้ ข้อมูลอุบัติเหตุการจราจรที่ใช้ในงานวิจัยนี้ได้รับความอนุเคราะห์จากศูนย์นเรนทร กระทรวงสาธารณสุข และจากสถานีตำรวจภูธรในจังหวัดนครปฐม แนวทางการทำวิจัยของเราเริ่มจากการวิเคราะห์ข้อมูลเบื้องต้น เพื่อให้ทราบลักษณะทั่วไปของข้อมูล ตั้งคำถาม แล้วจึงเลือกเทคนิคสำหรับวิเคราะห์ข้อมูลในเชิงลึก

คำถามประเภทหนึ่งที่เราสนใจคือ ปรากฏการณ์ที่เกิดขึ้นในกลุ่มตั้งแต่สองกลุ่มขึ้นไปนั้น มีรูปแบบเหมือนหรือต่างกันหรือไม่ อย่างไร เช่น รูปแบบของอุบัติเหตุที่เกิดขึ้นบริเวณทางตรง ทางโค้ง และทางแยกทางร่วม หรือรูปแบบการประสบอุบัติเหตุในกลุ่มวัยรุ่นและวัยกลางคน เป็นต้น โดยเราต้องการทราบเงื่อนไขหรือคุณสมบัติของแอตทริบิวต์อิสระต่าง ๆ ในกลุ่มเป้าหมายแต่ละกลุ่ม (แอตทริบิวต์เป้าหมายในตัวอย่างคือลักษณะทางและกลุ่มอายุ) งานวิจัยนี้ทดลองนำเทคนิคการจำแนก (classification) มาใช้เพื่อตอบคำถามดังกล่าว โดยเราเน้นเทคนิคที่ให้ผลลัพธ์เป็นต้นแบบการตัดสินใจ คือ C4.5 [3]

เรานำเสนอบทความนี้ด้วยการอธิบายถึงข้อมูลที่ใช้ในหัวข้อที่ 2 ทบทวนแนวคิดการสร้างต้นไม้การตัดสินใจในหัวข้อที่ 3 อธิบายวิธีดำเนินการวิจัยในหัวข้อที่ 4 และผลการวิจัยเบื้องต้นในหัวข้อที่ 5 ส่วนหัวข้อสุดท้ายเป็นบทสรุปและแนวทางการวิจัยในอนาคต

2. ข้อมูลที่ใช้ในงานวิจัย

ข้อมูลอุบัติเหตุการจราจรที่ใช้ในงานวิจัยนี้มี 3 ชุดคือ

1. ข้อมูลการปฏิบัติงานประจำวันของหน่วยบริการการแพทย์ฉุกเฉิน จากศูนย์นเรนทร จำนวน 1208 ระเบียน
2. ข้อมูลผู้ประสบอุบัติเหตุการจราจร ช่วงเทศกาลปีใหม่ ระหว่างวันที่ 27 ธันวาคม 2547 ถึง 5 มกราคม 2548 จากศูนย์นเรนทร จำนวน 736 ระเบียน
3. ข้อมูลคดีอุบัติเหตุการจราจรที่เกิดขึ้นในจังหวัดนครปฐม จำนวน 1103 คดี

แอตทริบิวต์ที่ปรากฏในข้อมูลสองชุดหลัง ซึ่งใช้ในบทความนี้แสดงในตารางที่ 1 จากฐานข้อมูลของศูนย์นเรนทร เราเลือกทุกระเบียนที่เป็นอุบัติเหตุการจราจรที่เกิดขึ้นในจังหวัดนครปฐม ส่วนข้อมูลชุดที่สามเป็นบันทึกคดีอุบัติเหตุการจราจร ณ สถานีตำรวจภูธรระดับอำเภอทั้งเจ็ดอำเภอในจังหวัดนครปฐม จากการที่เราไม่สามารถเก็บข้อมูลเชิงลึกเช่น ชื่อและทะเบียนยานพาหนะของผู้ประสบเหตุ การบูรณาการข้อมูลทั้งสามชุดจึงไม่สามารถทำได้ (เราไม่สามารถตามรอยได้ว่าระเบียนใดในข้อมูลแต่ละชุดเป็นกรณีเดียวกัน) ในเบื้องต้นเราจะเน้นการวิเคราะห์ข้อมูลเพื่ออธิบายปรากฏการณ์ที่ถูกจับยึด (captured) ด้วยข้อมูลที่มีอยู่ มากกว่าการสร้างต้นแบบเพื่อทำนาย

ตารางที่ 1. แอตทริบิวต์ที่ใช้ในการวิจัย ทุกแอตทริบิวต์เป็นข้อมูลชนิด nominal

ข้อมูลเทศกาลปีใหม่ 2448 (เซตข้อมูล NEW_YEAR)	
Day	วันในสัปดาห์ มี 3 กลุ่ม
Time	ช่วงเวลาในหนึ่งวัน มี 4 กลุ่ม
Gender	เพศของผู้ประสบเหตุ มี 2 กลุ่ม
Age	อายุของผู้ประสบเหตุ มี 4 กลุ่ม
Scene	บริเวณที่เกิดเหตุ มี 4 กลุ่ม
Status	คนเดินเท้า ผู้ขับขี่ หรือผู้โดยสาร
Drunk	อาการมึนเมา (ไม่มี / มี)
Safety	การใช้หมวก เข็มขัดนิรภัย (ไม่ใช้ / ใช้)
H1	ผู้เสียชีวิต (ไม่มี / มี)
H2	ผู้บาดเจ็บ (ไม่มี / มี)
Vehicle_I	ยานพาหนะของผู้ประสบเหตุ มี 6 กลุ่ม
Vehicle_O	ยานพาหนะของคู่กรณี มี 6 กลุ่ม
คดีอุบัติเหตุการจราจร (เซตข้อมูล POLICE)	
Day	วันในสัปดาห์ มี 3 กลุ่ม
Month	เดือน มี 4 กลุ่ม
Time	ช่วงเวลาในหนึ่งวัน มี 4 กลุ่ม
Scene	บริเวณที่เกิดเหตุ มี 3 กลุ่ม
Feature	ลักษณะทาง มี 4 กลุ่ม
Direction	ทิศของการจราจร มี 2 กลุ่ม
{V0, …, V6}	ยานพาหนะที่เกี่ยวข้องกับอุบัติเหตุ แต่ละตัวแปรแทนยานพาหนะแต่ละประเภท (ไม่เกี่ยว / เกี่ยว)
{C0, …, C6}	สาเหตุการเกิดอุบัติเหตุ แต่ละตัวแปรแทนแต่ละสาเหตุ (ใช่ / ไม่ใช่)
H1	ผู้เสียชีวิต (ไม่มี / มี)
H2	ผู้บาดเจ็บสาหัส (ไม่มี / มี)
H3	ผู้บาดเจ็บเล็กน้อย (ไม่มี / มี)

3. การสร้างต้นไม้การตัดสินใจด้วยเทคนิค C4.5

C4.5 เป็นเทคนิคการจำแนกซึ่งให้ผลลัพธ์เป็นต้นไม้การตัดสินใจที่มีคุณสมบัติดังนี้ โหนดใบของต้นไม้เป็นตัวแทนของคลาส หรือค่าที่เป็นไปได้ของแอตทริบิวต์เป้าหมาย โหนดภายในแต่ละโหนดเป็นแอตทริบิวต์อิสระ

และอาร์กที่ออกจากโหนดภายในเป็นทางเลือกในการตัดสินใจซึ่งขึ้นกับค่าของแอตทริบิวต์นั้น ข้อมูลที่ใช้กับ C4.5 จะต้องเป็นชนิด nominal การสร้างต้นไม้การตัดสินใจมีขั้นตอนคร่าว ๆ ดังนี้ ในแต่ละรอบการทำงานแอตทริบิวต์อิสระที่จำแนกอินสแตนซ์ได้ดีที่สุดจะถูกเลือกขึ้นมาเป็นโหนดของต้นไม้ โหนดลูกทั้งหมดจะถูกสร้างขึ้นพร้อมด้วยอาร์กที่แทนการตัดสินใจจากโหนดปัจจุบันไปยังโหนดลูก การทำงานจะสิ้นสุดลงเมื่อโหนดลูกทุกโหนดได้จำแนกอินสแตนซ์ไปตามคลาสของมัน มิฉะนั้นแอตทริบิวต์ใหม่จะถูกเลือกขึ้นมาใส่ในโหนดลูกที่ยังไม่สามารถจำแนกอินสแตนซ์ได้ เกณฑ์การเลือกแอตทริบิวต์ให้กับโหนดของต้นไม้การตัดสินใจคือ แอตทริบิวต์นั้นเมื่อจำแนกอินสแตนซ์แล้วจะให้อัตราส่วนสารสนเทศเพิ่มเติม (information gain ratio) มากที่สุด โดยเราใช้ entropy เป็นตัววัดปริมาณสารสนเทศก่อนและหลังการจำแนก การที่ entropy มีค่าต่ำลงแสดงว่าข้อมูลชุดนั้นมีการกระจัดกระจายน้อยลง หรือข้อมูลถูกจำแนกจนเป็นระเบียบมากขึ้น ทำให้ได้สารสนเทศจากข้อมูลชุดนั้นมากขึ้นด้วย รายละเอียดของอัลกอริทึมสามารถอ่านได้จาก [3] หรือ [6]

ในงานวิจัยนี้ ต้นไม้การตัดสินใจถูกสร้างขึ้นเพื่อศึกษารูปแบบ ค่าของแอตทริบิวต์อิสระต่าง ๆ ในกลุ่มเป้าหมายที่ถูกจำแนกเป็นสองกลุ่มขึ้นไป เราใช้ success rate เป็นตัววัดคุณภาพการจำแนกโดยรวม และใช้ true positive rate, false positive rate, และ precision เป็นตัววัดคุณภาพการจำแนกคลาสแต่ละคลาส ค่าเหล่านี้คำนวณจาก confusion matrix (รูปที่ 1) ดังนี้

$$success\ rate = \left(\sum_{i=1}^{n} f_{ii}\right) / total \quad (1)$$

$$TP(a) = f_{aa} / actual(a) \quad (2)$$

$$FP(a) = (predict(a) - f_{aa}) / (total - actual(a)) \quad (3)$$

$$precision(a) = f_{aa} / predict(a) \quad (4)$$

คลาสที่จำแนกโดยคลาสซิฟายเออร์ (Predicted Classes)

คลาสที่แท้จริง (Actual Classes)	a	b	c	
a	f_{aa}	f_{ab}	f_{ac}	$\sum_{k=1}^{n} f_{ak}$
b	f_{ba}	f_{bb}	f_{bc}	.
c	f_{ca}	f_{cb}	f_{cc}	.
	$\sum_{i=1}^{n} f_{ia}$	.	.	*total*

f_{ac} เป็นจำนวนอินสแตนซ์ที่มีคลาสที่แท้จริงคือคลาส a และถูกจำแนกให้อยู่ในคลาส c

total เป็นจำนวนอินสแตนซ์ทั้งหมด

$\sum_{k=1}^{n} f_{ak}$ หรือ *actual(a)* เป็นจำนวนอินสแตนซ์ทั้งหมดที่มีคลาสที่แท้จริงคือคลาส a

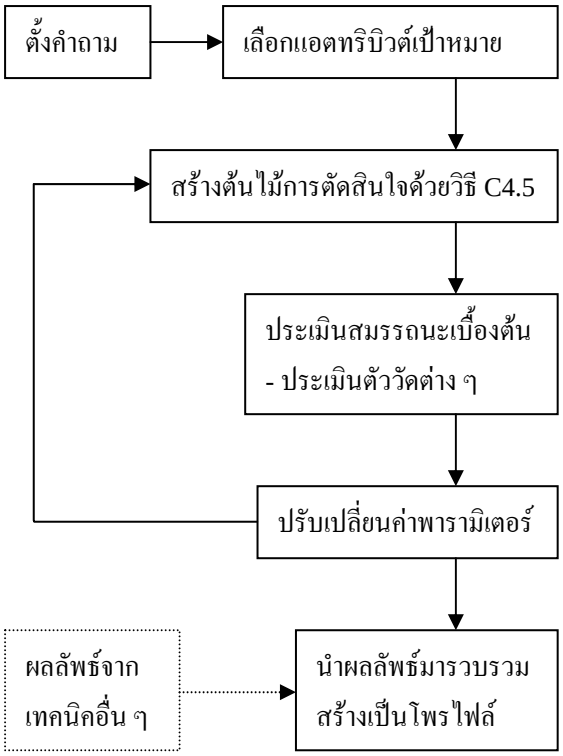
$\sum_{i=1}^{n} f_{ia}$ หรือ *predict(a)* เป็นจำนวนอินสแตนซ์ทั้งหมดที่ถูกจำแนกให้อยู่ในคลาส a

รูปที่ 1. confusion matrix ในการจำแนกคลาส a, b, และ c

TP(a) หรือ true positive rate บอกความถูกต้องในการจำแนกอินสแตนซ์มายังคลาส a ในขณะที่ *FP(a)* หรือ false positive rate บอกความเอนเอียงในการจำแนกอินสแตนซ์มายังคลาส a ซึ่งทำให้เกิดสัญญาณหลอก (false alarm) ส่วน *precision(a)* เป็นความเที่ยงในจำแนกคลาส a คุณสมบัติของคลาสที่เราต้องการคือ มีค่า *TP(a)* และ *precision(a)* สูง แต่มีค่า *FP(a)* ต่ำ ส่วนปัญหา overfitting นั้น เราควบคุมโดยการกำหนดพารามิเตอร์ที่เหมาะสม (เช่น ความถี่ขั้นต่ำของโหนดใบ) ในขั้นตอนการทำ pruning หรือการตัดเล็มต้นไม้การตัดสินใจ

4. วิธีดำเนินการวิจัย

บทความนี้นำเสนอผลการวิเคราะห์ข้อมูลโดยใช้เซตข้อมูล NEW_YEAR และเซตข้อมูล POLICE แอตทริบิวต์ที่ใช้ในการทดลองแสดงไว้ในตารางที่ 1 ส่วนกรอบการวิเคราะห์ข้อมูลแสดงในรูปที่ 2



รูปที่ 2. กรอบการวิจัย

เราใช้โมดูล J48 ในโปรแกรมทำเหมืองข้อมูลเวก้า (Weka, [5]) ซึ่งเป็นโมดูลที่สร้างต้นไม้การตัดสินใจด้วยอัลกอริทึม C4.5 ในแต่ละครั้งที่ได้ต้นไม้การตัดสินใจ หากพบว่าค่าของตัววัดสูงเกินไป (ในกรณีของ FP rate) หรือต่ำเกินไป (ในกรณีของ success rate, TP rate, และ precision) หรือจำนวนโหนดในต้นไม้น้อยเกินไปหรือมากเกินไป พารามิเตอร์ต่าง ๆ จะถูกปรับเปลี่ยนเพื่อให้ได้ต้นไม้การตัดสินใจที่ดีขึ้น ในการจำแนกเราใช้วิธี *n*-fold cross validation ในการแบ่งข้อมูลเป็นส่วนสำหรับฝึกและสำหรับทดสอบ ซึ่ง [4] และ [6] เสนอค่า *n* ที่เหมาะสมคือ 10 ส่วน เกณฑ์ในการตัดเล็ม (pruning) เรากำหนดความถี่ขั้นต่ำของโหนดใบเท่ากับ 10 เช่นกัน

ตัวอย่างคำถามที่เราตั้งขึ้นเพื่อวิเคราะห์ปรากฏการณ์ในกลุ่มเป้าหมายต่าง ๆ ได้แก่

1. ผู้ที่มีอาการมึนเมาและไม่มีอาการมึนเมา มีรูปแบบการเกิดอุบัติเหตุเป็นอย่างไร คำถามข้อนี้เราใช้ Drunk ในเซตข้อมูล NEW_YEAR เป็นแอตทริบิวต์เป้าหมาย
2. อุบัติเหตุที่ทำให้มีผู้เสียชีวิตมีรูปแบบเป็นอย่างไร คำถามข้อนี้เราใช้ H1 ในเซตข้อมูลทั้งสองชุดเป็นแอตทริบิวต์เป้าหมาย

5. ผลการวิจัยเบื้องต้น

เราสร้างต้นไม้การสร้างต้นไม้การตัดสินใจโดยใช้แอตทริบิวต์เป้าหมายดังที่กล่าวมา ผลการวิจัยเบื้องต้นมีดังนี้

5.1. อุบัติเหตุที่เกิดขึ้นในกลุ่มผู้ประสบเหตุที่มีอาการมึนเมาและไม่มีอาการมึนเมา

ตารางที่ 2 แสดงผลลัพธ์จากการสร้างต้นไม้การตัดสินใจเมื่อใช้ drunk เป็นแอตทริบิวต์เป้าหมาย ต้นไม้ที่ได้มีโหนดใบทั้งสิ้น 12 โหนด ในจำนวนนี้มีอยู่ 9 โหนดแสดงรูปแบบการเกิดอุบัติเหตุของผู้ขับขี่ที่ไม่มีอาการมึนเมา อีก 3 โหนดแสดงรูปแบบการเกิดอุบัติเหตุของผู้ขับขี่ที่มีอาการมึนเมา ซึ่งนำมาสรุปได้ดังนี้

คลาส 0 ผู้ประสบเหตุที่ไม่มีอาการมึนเมา

ก. เป็นเพศหญิง

ข. เป็นเพศชาย ซึ่งขับขี่ยานพาหนะ 5 ชนิดคือ รถโดยสารหรือรถตู้, รถปิคอัพ, รถบรรทุกหนักหรือรถพ่วง, จักรยานหรือสามล้อถีบ, และคนเดินเท้า

ค. เป็นเพศชาย ซึ่งขับขี่รถจักรยานยนต์ สวมหมวกนิรภัย เวลาเกิดอุบัติเหตุอยู่ในช่วงเช้าและกลางวัน (06.01–12.00 และ 12.01–18.00)

คลาส 1 ผู้ประสบเหตุที่มีอาการมึนเมา

ก. เป็นเพศชาย ซึ่งขับขี่รถยนต์นั่ง

ข. เป็นเพศชาย ซึ่งขับขี่รถจักรยานยนต์ โดยไม่สวมหมวกนิรภัย เวลาเกิดอุบัติเหตุคือช่วงหลัง 6 โมงเย็น กลางคืน และเช้ามืด (18.01–24.00 และ 00.01–06.00)

การจำแนกในแต่ละกลุ่มมีค่า TP rate และ precision ค่อนข้างสูง ต้นไม้การตัดสินใจค่อนข้างเอนเอียงไปทาง *คลาส 0* (พิจารณาจาก FP rate) เมื่อกลับไปพิจารณาต้นไม้การตัดสินใจเฉพาะในคลาสดังกล่าว เราพบว่าสัญญาณหลอกส่วนใหญ่เกิดขึ้นใน*คลาส 0ค* ในช่วงเวลา 06.01-12.00 โดยมีอัตราการจำแนกผิดเท่ากับ 0.42 ส่วนคลาสย่อยอื่น ๆ มีการจำแนกผิดค่อนข้างต่ำ ดังนั้น หากจะจัดการเฝ้าระวังเพื่อลด / แก้ปัญหาอุบัติเหตุการจราจรในช่วงเทศกาลปีใหม่ (หรือขยายผลให้ครอบคลุมเทศกาลที่มีวันหยุดยาวอื่น ๆ) ก็ควรให้ความสนใจกับผู้ขับขี่รถจักรยานยนต์ ทั้งในคลาส *1ข* และ *0ค* นอกจากนี้ ช่วงกลางคืนถึงเช้ามืดควรกวดขันการสวมหมวกนิรภัยของผู้ขับขี่ด้วย

5.2. อุบัติเหตุที่ทำให้เกิดการเสียชีวิต

ตาราง 3 แสดงผลลัพธ์จากการสร้างต้นไม้การตัดสินใจโดยใช้ H1 เป็นแอตทริบิวต์เป้าหมาย

เนื่องจากในเซตข้อมูล NEW_YEAR มีอุบัติเหตุที่มีผู้เสียชีวิตอยู่เพียง 1.6% ของข้อมูลทั้งหมด ส่วนในเซตข้อมูล POLICE มีคดีที่มีผู้เสียชีวิตอยู่เพียง 15.7% ของข้อมูลทั้งหมด เราสร้างต้นไม้การตัดสินใจจากเซตข้อมูล NEW_YEAR โดยไม่มีการตัดเล็ม และจากเซตข้อมูล POLICE โดยลดความถี่ขั้นต่ำของโหนดใบลง เพื่อไม่ให้การตัดสินใจที่โยงมายัง*คลาส 1* ถูกตัดทิ้ง ต้นไม้ที่ได้ทั้งสองต้นเอนเอียงไปทาง*คลาส 0* ค่อนข้างมาก กล่าวคือ FP rate ของ*คลาส 0* มีค่าสูงมาก ในขณะที่ TP rate ของ*คลาส 1* ก็มีค่าต่ำมากเช่นกัน

อย่างไรก็ตาม ในเซตข้อมูล POLICE มีกฎการตัดสินใจเลือก*คลาส 1* บางข้อที่จำแนกข้อมูลได้บ้าง และมีค่า support ไม่ต่ำจนเกินไปได้แก่

- อุบัติเหตุซึ่งเกิดขึ้นบนทางหลวง รถวิ่งทางเดียว และมีคู่กรณีเป็นคนเดินเท้า (อัตราการจำแนกผิด 0.24, ครอบคลุม 11.6% ของคดีที่มีผู้เสียชีวิต)
- อุบัติเหตุซึ่งมีคู่กรณีเป็นรถจักรยานยนต์และรถบรรทุกหนัก เกิดในช่วงกลางสัปดาห์ หรือระหว่างวันอังคาร–พฤหัสบดี (อัตราการจำแนกผิด 0.4, ครอบคลุม 17.3% ของคดีที่มีผู้เสียชีวิต) เป็นต้น

ข้อสังเกตุดังกล่าวเป็นเพียงข้อสังเกตุเบื้องต้น ซึ่งจะนำไปเสริมกับผลการวิเคราะห์ข้อมูลอื่น ๆ เพื่อให้ได้ข้อสรุปที่สมเหตุสมผลและมีความน่าเชื่อถือต่อไป

ตารางที่ 2. ผลลัพธ์เบื้องต้น เมื่อใช้ drunk เป็นแอตทริบิวต์เป้าหมาย

เซตข้อมูล NEW_YEAR		
สถิติรวม	คลาส 0	คลาส 1
Leaf nodes = 12 Success = 0.73	TP rate = 0.80 FP rate = 0.40 Precision = 0.80	TP rate = 0.61 FP rate = 0.20 Precision = 0.63

ตารางที่ 3. ผลลัพธ์เบื้องต้น เมื่อใช้ H1 เป็นแอตทริบิวต์เป้าหมาย

เซตข้อมูล NEW_YEAR (unpruned)		
สถิติรวม	คลาส 0	คลาส 1
Leaf nodes = 17 Success = 0.98	TP rate = 0.99 FP rate = 0.92 Precision = 0.99	TP rate = 0.08 FP rate = 0.01 Precision = 0.12
เซตข้อมูล POLICE (ใช้ min. leaf size = 3)		
สถิติรวม	คลาส 0	คลาส 1
Leaf nodes = 29 Success = 0.84	TP rate = 0.95 FP rate = 0.78 Precision = 0.87	TP rate = 0.21 FP rate = 0.05 Precision = 0.45

6. บทสรุปและแนวทางการทำวิจัยต่อ

บทความนี้นำเสนอการวิเคราะห์ รูปแบบของอุบัติเหตุการจราจรที่เกิดขึ้นในกลุ่มตั้งแต่สองกลุ่มขึ้นไป โดยใช้ต้นไม้การตัดสินใจ ผลการวิจัยเบื้องต้นแสดงให้เห็นว่าต้นไม้การตัดสินใจสะท้อนภาพของปรากฏการณ์ที่เกิดขึ้นในแต่ละกลุ่มได้พอสมควร แต่การใช้ต้นไม้การตัดสินใจยังมีข้อจำกัดในกรณีต่อไปนี้

- ข้อมูลเทไปอยู่ในกลุ่มเพียงกลุ่มเดียว แอตทริบิวต์อิสระที่มีอยู่ไม่สามารถจำแนกกลุ่มในแอตทริบิวต์เป้าหมาย ทำให้ต้นไม้การตัดสินใจเป็นเพียงต้นไม้โหนดเดียว
- ข้อมูลกระจัดกระจายไปอยู่ในกลุ่มหลายกลุ่มมากเกินไป ทำให้ต้นไม้การตัดสินใจเป็นต้นไม้ขนาดใหญ่ แต่มีลักษณะ overfitting

อย่างไรก็ตาม การสร้างโพรไฟล์ของอุบัติเหตุการจราจรยังต้องอาศัยการทำเหมืองข้อมูลหลาย ๆ แบบประกอบกัน โดยเราจะนำผลจำแนกด้วยวิธีอื่น ๆ เช่น Naïve Bayes มาวิเคราะห์เปรียบเทียบด้วย นอกจากนี้ เรายังจะศึกษาแนวทางการสกัดหารูปแบบของปรากฏการณ์ที่เกิดขึ้นในกลุ่มขนาดเล็ก (ดังตัวอย่าง หัวข้อ 5.2) เนื่องจากกลุ่มขนาดเล็กเหล่านี้อาจเป็นกลุ่มเป้าหมายที่ควรจับตา เฝ้าระวังเป็นกรณีพิเศษ งานที่เรากำลังศึกษาอยู่ได้แก่ [1] และ [2] ซึ่งเสนอแนวทางการวิเคราะห์และการจัดเตรียมข้อมูลสำหรับการจำแนก ในกรณีที่ขนาดของคลาสไม่สมดุลกัน (imbalanced classes)

การเฝ้าระวังกลุ่มขนาดเล็กนั้น จะต้องศึกษา ชั่งน้ำหนักระหว่างต้นทุน (cost) และผลกำไร (benefit) นอกจากนี้ การจัดทำข้อสรุปหรือข้อเสนอแนะที่มีพื้นฐานจากการวิเคราะห์กลุ่มขนาดเล็ก จะต้องทำด้วยความระมัดระวัง เพื่อให้เกิดความน่าเชื่อถือและสามารถนำไปใช้ประโยชน์ได้จริง

7. กิตติกรรมประกาศ

งานวิจัยนี้เป็นส่วนหนึ่งของโครงการวิจัย “เหมืองข้อมูลการจราจรสำหรับจังหวัดนครปฐม” ซึ่งได้รับทุนวิจัยจากสำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ (สวทช.)

8. เอกสารอ้างอิง

[1] Batista, G., Prati R. C., Monard, M. C., “A Study of the Bahavior of Several Methods for Balancing Machine Learning Training Data”, SIGKDD Exploration (6), 2004.
[2] Japkowicz, N., “The Class Imbalance Problem: Significance and Strategies”, Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI 2000).
[3] Quinlan, J. R., “C4.5: Programs for Machine Learning”, Morgan Kaufmann, 1993.
[4] Roiger, R. J., Geatz, M., W.: Data Mining: A Tutorial-Based Primer. Addison-Wesley, 2003.
[5] Weka: Data Mining Software in Java. University of Waikato, New Zealand. http://www.waikato.ac.nz/ml/weka
[6] Witten, I. H., Frank, E.: Data Mining: Practical Machine Learning Tools and Technique. 2nd edition. Elsevier Inc., 2005.