แบบจำลองทำนายความเสี่ยงการเกิดอุบัติเหตุบนท้องถนน โดยเทคนิคถุงจำแนก
Road Accident Risk Prediction Model using Bagging Techniques.

อมรภัทร์ หาญโคกกรวด[1*], จารี ทองคำ[2], ฉัตรตระกูล สมบัติธีระ[3], ชัยนันท์ สมพงษ์[4]
Amornpatr Hancokkruad[1*], Jaree Thongkam[2], Chattrakul Sombattheera[3],
Chaiyanan Sompong[4]

บทคัดย่อ

การเกิดอุบัติเหตุบนท้องถนนเป็นปัญหาที่มีแนวโน้มที่ทวีความรุนแรงมากขึ้นทุกปี จากการศึกษาพบว่าการเกิดอุบัติเหตุบนท้องถนนส่วนใหญ่เกิดจากมนุษย์ เพื่อช่วยให้คนขับรถตระหนักถึงความเสี่ยงการเกิดอุบัติเหตุบนท้องถนน โดยงานวิจัยนี้ได้สร้างแบบจำลองทำนายความเสี่ยงการเกิดอุบัติเหตุบนท้องถนน ซึ่งเทคนิคถุงจำแนกเป็นวิธีการจำแนกข้อมูลในการทำเหมืองข้อมูล และสามารถเพิ่มประสิทธิภาพในการทำนายให้มากขึ้น ยังมีผู้วิจัยใช้เทคนิคถุงจำแนกน้อย ดังนั้นในงานวิจัยฉบับนี้จึงสร้างแบบจำลองเพื่อการทำนายความเสี่ยงการเกิดอุบัติเหตุบนท้องถนนโดยใช้เทคนิคถุงจำแนก รวบรวมข้อมูลจากสถานีตำรวจภูธรจังหวัดสกลนคร ระหว่างปี พ.ศ.2552 ถึงปี พ.ศ. 2556 จากการทดลองพบว่า เทคนิคถุงจำแนกสามารถสร้างแบบจำลองที่มีประสิทธิภาพมีค่าความแม่นยำมากกว่าแบบจำลองจำแนกเบย์ และแบบจำลองเครื่องมือสนับสนุนเส้นสมมุติ ยิ่งไปกว่านั้นยังพบอีกว่าเทคนิคถุงจำแนกสามารถเพิ่มค่าความแม่นยำให้กับแบบจำลองเครื่องมือสนับสนุน จากเดิมคิดเป็นร้อยละ 0.86

คำสำคัญ: แบบจำลองทำนายความเสี่ยงการเกิดอุบัติเหตุบนท้องถนน, เทคนิคถุงจำแนก, เหมืองข้อมูล

Abstract

The accident that occur on the road as a problem to severe and increase in every years. Of the studies about road accident have found that mostly are caused by humans. So that to help the drivers aware the road accident risk, which in this paper are proposed the road accident risk prediction model. Bagging Technique is approach to classify the data for data mining and can be increasing performance of our prediction which have a few researchers are employed the bagging technique. Therefore, in this paper is proposed the Bagging Technique to build the model for road accident risk prediction. Data are collected from 2552 to 2556 at Sakonnakhon Province Locality Police Station. Experimental results showed that bagging technique is superior to naïve bayes model and support vector machine model. Moreover, also

[1] นิสิตปริญญาโท, [2,3] อาจารย์, สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม อำเภอเมือง จังหวัดมหาสารคาม 4400

[4] อาจารย์, สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏสกลนคร

[1] Graduate student, [2,3]Lecturer, Department Information technology, Faculty of Informatics, Mahasarakham University, Mueang District, Maha Sarakham 44000, Thailand.

[4] Lecturer, Department Information technology, Faculty of Science and Technology, Sakon Nakhon Rajabhat University.

**Corresponding author: Amornpatr Hancokkruad, Department Information technology, Faculty of Informatics, Mahasarakham University, Mueang District, Maha Sarakham 44000, Thailand.*

found that bagging technique can increase the accuracy of support vector machine model from 0.86 percent.

Keyword: Road Accident Risk Prediction Model, Bagging Techniques, Data Mining

บทนำ

การเกิดอุบัติเหตุบนท้องถนนเป็นปัญหาที่มีแนวโน้มที่ทวีความรุนแรงมากขึ้น ในแต่ละปีประเทศต้องสูญเสียทรัพยากรบุคคลที่มีคุณค่า ตลอดถึงทรัพย์สินมูลค่ามหาศาล ส่งผลกระทบต่อสภาพเศรษฐกิจ สังคม คุณภาพชีวิตของประชาชน โดยทุกปีมีรายงานผู้เสียชีวิตจากการเกิดอุบัติเหตุบนท้องถนนกว่า 1.3 ล้านคนทั่วโลกหรือคิดเป็นร้อยละ 46 ของผู้ประสบเหตุ[1] และจากข้อมูลของสำนักงานอำนวยความปลอดภัย กรมทางหลวง พบว่าปี พ.ศ. 2555 ประเทศไทยมีผู้ประสบอุบัติเหตุทั้งหมด 21,621 ราย จำนวนผู้บาดเจ็บ 9,701 ราย และจำนวนผู้เสียชีวิต 1,555 ราย[2] สาเหตุสำคัญของการเกิดอุบัติเหตุบนท้องถนน จากการศึกษาพบว่าการเกิดอุบัติเหตุบนท้องถนนส่วนใหญ่เกิดจากมนุษย์

ปัจจุบันการทำเหมืองข้อมูลได้เข้ามามีบทบาทในการวิเคราะห์ข้อมูลการเกิดอุบัติเหตุบนท้องถนน โดยทำการรวบรวมข้อมูลการเกิดอุบัติเหตุบนท้องถนนจากหน่วยงานต่างๆ ทั้งภาครัฐและภาคเอกชน และนำข้อมูลทั้งหมดมาวิเคราะห์เพื่อให้ได้แบบจำลองสำหรับทำนายความเสี่ยงการเกิดอุบัติเหตุบนท้องถนน ดังเช่น งานวิจัยของ Tian และคณะ[3] ได้วิเคราะห์ปัจจัยที่ก่อให้เกิดอุบัติเหตุบนท้องถนน โดยใช้เทคนิควิธี Rough sets theory และ Association rules พบว่า บุคคล สภาพยานพาหนะ ลักษณะถนน และสิ่งแวดล้อมเป็นปัจจัยที่สำคัญทำให้เกิดอุบัติเหตุบนท้องถนนได้ และถัดมาเป็นงานวิจัยของ Krishnaveni[4] ได้พัฒนาแบบจำลองสำหรับทำนายการบาดเจ็บที่เกิดจากอุบัติเหตุบนท้องถนน โดยใช้เทคนิควิธี Naive bayes bayesian, PART Rule, J48 Decision tree และ Random forest พบว่าการสร้างแบบจำลองด้วยเทคนิควิธี Random forest ให้ความถูกต้องดีกว่าวิธีอื่นๆ คิดเป็นร้อยละ 74.34 จากการทบทวนวรรณกรรมพบว่า ประสิทธิภาพของแบบจำลอง ยังให้ความถูกต้องไม่ดีเท่าที่ควร แต่งานวิจัยของของ Thongkam[5] ได้สร้างแบบจำลองในการทำนายของโรคมะเร็งเต้านม โดยการทำเหมืองข้อมูลด้วยเทคนิควิธี Bagging, Random Tree, Decision Stump, REPTree, J48, Bagging+Random Tree, Bagging+Decision Stump, Bagging+REPTree และ Bagging+J48 จากผลการศึกษาพบว่า Bagging + Random Tree ให้ประสิทธิภาพของแบบจำลองได้เป็นอย่างดี ให้ความถูกต้องดีที่สุดคิดเป็นร้อยละ 98.82 แต่ให้ความถูกต้องมากกว่าเดิมคิดเป็นร้อยละ 1.92 ถัดมางานวิจัยของ Poel[6] ได้ประยุกต์ใช้การทำเหมืองข้อมูลเพื่อการทำนายผลกำไรของลูกค้าระหว่างการขาย โดยวิเคราะห์ข้อมูลจากเว็บและข้อมูลเชิงพานิชย์ ด้วยเทคนิควิธี Bagging, Decision tree และ Bagging + Decision tree จากการศึกษาพบว่า Bagging + Decision tree ให้ประสิทธิภาพของแบบจำลองได้เป็นอย่างดี แต่ให้ความถูกต้องมากกว่าเดิมคิดเป็นร้อยละ 0.56 ซึ่งในทางการทำเหมืองข้อมูลการสร้างแบบจำลองเพื่อทำนายความเสี่ยงการเกิดอุบัติเหตุบนท้องถนน ยังมีผู้วิจัยที่ใช้เทคนิควิธีถุงจำแนก (Bagging) น้อย

ดังนั้นงานวิจัยนี้จึงสร้างแบบจำลองเพื่อทำนายความเสี่ยงการเกิดอุบัติเหตุบนท้องถนนให้มีประสิทธิภาพของแบบจำลองมากยิ่งขึ้น โดยใช้หลักการถุงจำแนกในการเพิ่มประสิทธิภาพของแบบจำลอง เพื่อลดการสูญเสียชีวิตและทรัพย์สินจากการเกิดอุบัติเหตุ และระดับความรุนแรงสำหรับผู้ใช้ยานพาหนะบนท้องถนนได้

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

1. เหมืองข้อมูล

คำว่า “เหมืองข้อมูล” มีนักวิจัยได้ให้นิยามไว้หลายความหมายดังนี้ งานวิจัยของ Dong-xiao[7] ได้

ให้ความหมายเหมืองข้อมูลไว้ว่า เป็นกระบวนการค้นหารูปแบบที่มีความสัมพันธ์ โดยอาศัยการรู้จำแบบ วิธีการทางสถิติ ทางคณิตศาสตร์ และเทคโนโลยี เพื่อให้รูปแบบของข้อมูลที่มีประสิทธิภาพของแบบจำลอง และงานวิจัยของ Poel[6] ได้ให้ความหมายเหมืองข้อมูลไว้ว่า เป็นวิธีการสกัดความรู้ที่ซ่อนอยู่ในฐานข้อมูลขนาดใหญ่ หรือวิเคราะห์ข้อมูลที่ได้ถูกออกแบบมาเพื่อใช้ในกระบวนการตัดสินใจ

ดังนั้นจึงสรุปได้ว่าเหมืองข้อมูลหมายถึงกระบวนการค้นหาวิธีการสร้างแบบจำลอง และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลจำนวนมหาศาลโดยอัตโนมัติ ซึ่งใช้ขั้นตอนวิธีการทางสถิติ การเรียนรู้ของเครื่อง และการรู้จำแบบ ปัจจุบันการทำเหมืองข้อมูลได้เข้ามามีบทบาทในการวิเคราะห์ข้อมูลการเกิดอุบัติเหตุบนท้องถนน โดยทำการรวบรวมข้อมูลการเกิดอุบัติเหตุบนท้องถนนจากหน่วยงานต่างๆ ทั้งภาครัฐและภาคเอกชน และนำข้อมูลทั้งหมดมาวิเคราะห์เพื่อให้ได้แบบจำลองทำนายความเสี่ยงการเกิดอุบัติเหตุบนท้องถนน ซึ่งการทำเหมืองข้อมูลในการสร้างแบบจำลองสามารถแบ่งได้เป็น 2 ประเภทใหญ่ๆ ประกอบด้วย 1) การสร้างแบบจำลองเพื่อการทำนาย คือ การนำข้อมูลในอดีตมาสร้างแบบจำลองทำนายอนาคต โดยมีการใช้ข้อมูลสำหรับสอนทุกข้อมูลจะมีคุณสมบัติ เป็นค่าที่ใช้ในการทำนายผลของข้อมูล อัลกอริทึมประเภทนี้จะมุ่งเน้นในการแบ่งแยกข้อมูลออกเป็นกลุ่มตามค่าคุณสมบัติของข้อมูล ซึ่งถ้าค่าคุณสมบัติของข้อมูลมีค่าไม่ต่อเนื่อง จะเรียกกระบวนการที่ใช้แบ่งแยกว่าการจำแนกประเภท ถ้าค่าคุณสมบัติของข้อมูลมีค่าต่อเนื่อง จะเรียกกระบวนการที่ใช้ว่าการถดถอยหรือการทำนาย 2) การสร้างแบบจำลองเพื่อการบรรยาย คือ การนำข้อมูลที่มีอยู่มาดูและศึกษาเพื่อหากฎความสัมพันธ์ต่างๆ หรือหาการจัดกลุ่มข้อมูล ไม่ได้มีจุดมุ่งหมายเพื่อการทำนาย

2. เทคนิคเหมืองข้อมูล

เทคนิคเหมืองข้อมูลในปัจจุบันมีหลายรูปแบบ โดยผู้วิจัยได้ทำการศึกษาและค้นคว้าเกี่ยวกับเทคนิคเหมืองข้อมูลจากแหล่งข้อมูลต่างๆ เทคนิคที่ใช้ทำเหมืองข้อมูลเพื่อการทำนาย ได้แก่ ถุงจำแนก จำแนกเบย์ และเครื่องมือสนับสนุนเส้นสมมุติ ดังต่อไปนี้

1.ถุงจำแนก (Bagging : B)[8] เป็นเทคนิคที่ใช้การคำนวณทางสถิติ และคณิตศาสตร์ ในการเพิ่มประสิทธิภาพให้กับตัวจำแนกหรือแบบจำลองเพื่อการทำนาย โดยที่เทคนิคนี้จะไปทำการลดความผันผวนของผลการทำนายของแบบจำลอง จะเห็นว่าถุงจำแนกจำเป็นต้องมีการทดลองหลายๆ ครั้ง เพื่อค้นหาแบบจำลองที่ดีที่สุด ดัง Figure 1

Input: training set S, number of bagging T

Bagging (T, S)
for $i = 1$ to T {
  S'_1 = sample from class 1 in S (with replacement)
  S'_2 = sample from class 2 in S (with replacement)
  $S' = S'_1 + S'_2$
  train a decision tree C_i from S'
}

Output: T classifiers

Figure 1 The bagging algorithm. T is 50 in our experiments. In each bag, the class distribution is the same as in the original data S.

2. จำแนกเบย์ (Naïve bayes : NB)[13, 14, 15] เป็นเทคนิคที่ใช้ในการจำแนกข้อมูลตามตัวแปร โดยการใช้วิธีการประมาณค่าจากตัวเลขของค่าความแม่นยำ แล้วทำการวิเคราะห์ข้อมูลที่ใช้ในการสร้างแบบจำลอง ทำให้แบบจำลองสามารถเปลี่ยนแปลงได้ โดยข้อมูลในการสร้างแบบจำลองเริ่มจากศูนย์ จากหลักการนี้ทำให้แบบจำลองมีประสิทธิภาพมากขึ้น เทคนิคนี้ไม่ไวต่อจำนวนของตัวแปร และใช้ได้กับตัวแปรทุกชนิด โดยใช้หลักการคำนวณดังสมการที่ 1 ในการเลือกตัวแปรที่ต้องการทำนาย

$$P(X \mid C_i) = P(x_1|C_i)\,x\,P(x_2|C_i)\,x\ldots x\,P(x_n \mid C_i) \quad (1)$$
$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i)$$

เมื่อ P คือ ค่าความน่าจะเป็นของแต่ละตัวแปรตาม

X คือ แถวของข้อมูล ตัวอย่าง $(x_1, x_2, \ldots, x_n)$

n คือ จำนวนของข้อมูล

C_i คือ ตัวแปรตามที่มีค่าเป็น i ในข้อมูล

3. เครื่องมือสนับสนุนเส้นสมมุติ (Support vector machine : SVM)[16, 17, 18] เป็นเทคนิคที่ใช้ในการคัดแยกที่มีการนำมาใช้กันอย่างกว้างขวางในด้านการประมวลผลภาพดิจิตอล หลักของ SVM คือ การให้ปัจจัยนำเข้าที่ใช้ฝึกเป็นเวคเตอร์ในสเปซ N มิติ เช่น ถ้าในกรณีของ 2 มิติ และ 2 มิติ จะเป็นจุดที่อยู่ในระนาบ xy และระนาบ xyz ตามลำดับ จากนั้นทำการสร้างไฮเปอร์เพลน (Hyper plane) ที่จะแยกกลุ่มของเวกเตอร์ปัจจัยนำเข้าออกเป็นประเภทต่างๆ ในกรณีที่ 2 มิติ และ 3 มิติ ไฮเปอร์เพลน คือ เส้นระนาบ จุดเด่นของ SVM จะทำการเก็บจับคู่ของเวคเตอร์ในสเปซปัจจัยนำเข้าให้เข้าสู่ Feature space โดยใช้ฟังก์ชั่นเคอร์นัล ชนิดต่างๆ เช่น เคอร์นัลชิงเส้น เคอร์นัลโพลีโนเมียยล เคอร์นัลเรเดียล และเคอร์นัลสิกมอด เป็นต้น ใน Feature Space ดังกล่าวเวคเตอร์ปัจจัยนำเข้า สามารถแยกประเภทได้โดยไฮเปอร์เพลน ดัง Figure 2

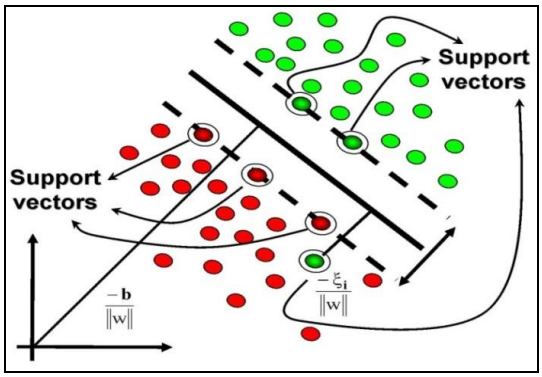

Figure 2 Hyper plane

3. การวัดประสิทธิภาพ

การวัดประสิทธิภาพแบบจำลองทำนายความเสี่ยงการเกิดอุบัติเหตุบนท้องถนน ในด้านเหมืองข้อมูลจำเป็นต้องออกแบบการทดลอง โดยใช้วิธี 10 - Fold cross validation ซึ่งแบ่งชุดข้อมูลออกเป็น 2 ชุด ประกอบด้วย ชุดข้อมูลการเรียนรู้ (Training data) และชุดข้อมูลการทดสอบ (Testing data) สามารถทำโดยแบ่งข้อมูลออก เป็น 10 ส่วนเท่าๆ กัน 1 ส่วนนำไปทดสอบ และ 9 ส่วนนำไปใช้ในการเรียนรู้ ในเบื้องต้นเลือกข้อมูลกลุ่มที่ 1 เป็นชุดข้อมูลการทดสอบ และข้อมูลชุดที่เหลือจะเป็นข้อมูลชุดการเรียนรู้ นำข้อมูลไปสร้างแบบจำลองทำนาย จากนั้นจะสลับข้อมูลกลุ่มที่ 2 มาเป็นชุดข้อมูลการทดสอบ และข้อมูลกลุ่มอื่นๆ ที่เหลือเป็นชุดข้อมูลการทดสอบ สลับอย่างนี้ไปเรื่อยๆ จนครบ 10 กลุ่ม ในแต่ละขั้นตอนหาค่าความแม่นยำ และค่าพื้นที่ใต้เส้นกราฟ ROC (AUC) แล้วจึงหาค่าเฉลี่ยของค่าความแม่นยำและค่าพื้นที่ใต้เส้นกราฟ ROC (AUC) จะเห็นได้ว่าข้อมูลทุกข้อมูลจะถูกทดสอบ ซึ่งผลการทดลองที่ได้สามารถนำมาแสดงใน Figure 3

		Predicted	
		Positive	Negative
True	Positive	TP	FN
	Negative	FP	TN

Figure 3 The Confusion matrix

ค่าที่ได้ในตาราง The Confusion matrix สามารถนำไปหาประสิทธิภาพของแบบจำลองเพื่อการทำนาย เช่น ค่าความแม่นยำ และพื้นที่ใต้เส้นกราฟ ROC (AUC)

1. ค่าความแม่นยำ (Accuracy)[19] เป็นค่าพื้นฐานในการวัดประสิทธิภาพของแบบจำลองในการทำนายผลการวัดที่ได้ค่าเข้าใกล้ค่าจริง ซึ่งค่าจริงแท้ๆ ไม่มีใครทราบ เนื่องจากในการวัดมีความคลาดเคลื่อนเกิดขึ้นเสมอ ในการบ่งบอกถึงความถูกต้องของวิธีวิเคราะห์ จึงต้องเปรียบเทียบค่าที่วัดได้กับค่าจริงที่ยอมรับกัน เพื่อวัดประสิทธิภาพแบบจำลอง โดยค่าข้อมูลเหล่านี้ขึ้นอยู่กับจำนวนของระเบียนข้อมูลที่สามารถทำนาย ดังที่สมการ 2

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (2)$$

จากสมการที่ 2

TP คือ อัตราความถูกต้องเชิงบวก

TN คือ อัตราความถูกต้องเชิงลบ

FP คือ อัตราความผิดพลาดเชิงบวก

FN คือ อัตราความผิดพลาดเชิงลบ

2. ค่าพื้นที่ใต้เส้นกราฟ ROC (Area Under the receiver operating characteristic Curve : AUC) [20, 21] เป็นอีกทางเลือกหนึ่งที่ใช้ในการวัดประสิทธิภาพการทำงานของแบบจำลอง โดยที่เส้นในแกนนอนจะเป็นอัตราบวกจริง (True positive rate) ส่วนในแกนตั้งจะเป็นอัตราบวกเท็จ (False positive rate) ค่าการวัดโดยใช้ AUC นี้จะเริ่มที่ 0 ถึง 1 โดยที่ 0 หมายถึง แบบจำลองนั้นมีประสิทธิภาพต่ำกว่า ส่วน 1 หมายถึง แบบจำลองนั้นมีประสิทธิภาพสูงที่สุด ซึ่งผลการวัดจะง่ายต่อความเข้าใจ แต่ในงานวิจัยนี้ได้นำเอาร้อยละของ AUC มาทำการแสดงผล ซึ่งสามารถที่จะให้ความละเอียดมากกว่า

4. งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้องกับการเกิดอุบัติเหตุบนท้องถนน ดังต่อไปนี้

Tian และคณะ[3] ได้วิเคราะห์ปัจจัยที่ก่อให้เกิดอุบัติเหตุบนท้องถนน โดยใช้เทคนิควิธี Rough sets theory และ Association rules พบว่า บุคคล สภาพยานพาหนะ ลักษณะถนน และสิ่งแวดล้อม เป็นปัจจัยที่สำคัญทำให้เกิดอุบัติเหตุบนท้องถนนได้

Krishnaveni[4] ได้พัฒนาแบบจำลองสำหรับทำนายการบาดเจ็บที่เกิดจากอุบัติเหตุบนท้องถนน โดยใช้เทคนิควิธี Naive bayes bayesian, PART Rule, J48 Decision tree และ Random forest พบว่าการสร้างแบบจำลองด้วยเทคนิควิธี Random forest ให้ความถูกต้องดีกว่าวิธีอื่นๆ คิดเป็นร้อยละ 74.34

Chang[22] ได้วิเคราะห์ปัจจัยที่ก่อให้เกิดอุบัติเหตุบนท้องถนน ศึกษาและรวบรวมข้อมูลอุบัติเหตุของประเทศไทเปและประเทศไต้หวัน ในปี 2001 โดยการประยุกต์ใช้เทคนิคเหมืองข้อมูลด้วยวิธี Classification and Regression tree (CART) หาความสัมพันธ์ระดับความรุนแรงของการเกิดอุบัติเหตุการจราจรบนท้องถนน จากการศึกษาพบว่า ปัจจัยด้านสิ่งแวดล้อม มนุษย์และยาพาหนะเป็นปัจจัยเสี่ยงที่ก่อให้เกิดอุบัติเหตุบนท้องถนนได้

งานวิจัยที่เกี่ยวข้องกับเทคนิคถุงจำแนกดังต่อไปนี้

Zhang[23] ได้สร้างแบบจำลองเพื่อการทำนายการให้คะแนนบัตรเครดิต โดยการทำเหมืองข้อมูลด้วยเทคนิควิธีถุงจำแนกร่วมกับต้นไม้ตัดสินใจ ทดสอบจากข้อมูลมหาวิทยาลัยเออร์ไวน์ แคลิฟอร์เนีย (University of California, Irvine : UCI) จากผลการศึกษา ทำให้ประสิทธิภาพของแบบจำลองสูงขึ้นกว่าเดิมคิดเป็นร้อยละ 1.91

Thongkam[5] ได้สร้างแบบจำลองเพื่อการทำนายของโรคมะเร็งเต้านม โดยการทำเหมืองข้อมูลด้วยเทคนิควิธี Bagging, Random tree, Decision stump, REPTree, J48, Bagging+ Random tree, Bagging + Decision stump, Bagging + REPTree และ Bagging + J48 จากผลการศึกษาพบว่า Bagging + Random tree ให้ประสิทธิภาพของแบบจำลองได้เป็นอย่างดี ให้ความถูกต้องดีที่สุดคิดเป็นร้อยละ 98.82 แต่ให้ความถูกต้องมากกว่าเดิมคิดเป็นร้อยละ 1.92

Poel[6] ได้ประยุกต์ใช้การทำเหมืองข้อมูลเพื่อการทำนายผลกำไรของลูกค้าระหว่างการขาย โดยวิเคราะห์ข้อมูลจากเว็บและข้อมูลเชิงพานิชย์ ด้วยเทคนิควิธีถุงจำแนก ต้นไม้ตัดสินใจ และถุงจำแนกร่วมกับต้นไม้ตัดสินใจ จากการศึกษาพบว่าแบบจำลองถุงจำแนกร่วมกับต้นไม้ตัดสินใจให้ประสิทธิภาพของแบบจำลองได้เป็นอย่างดี แต่ให้ความถูกต้องมากกว่าเดิมคิดเป็นร้อยละ 0.56

วิธีการดำเนินการวิจัย

แบบจำลองทำนายความเสี่ยงการเกิดอุบัติเหตุบนท้องถนน โดยใช้เทคนิคถุงจำแนก ซึ่งมีวิธีการดำเนินการวิจัยดังนี้

1. การเตรียมข้อมูล

ข้อมูลในงานวิจัยนี้ได้รับความอนุเคราะห์และรวบรวมจากสถานีตำรวจจังหวัดสกลนคร ซึ่งมีขั้นตอนการเตรียมข้อมูลดังนี้

1. เก็บรวบรวมข้อมูลการเกิดอุบัติเหตุบนท้องถนน เพื่อจัดทำฐานข้อมูลระหว่างเดือนพฤศจิกายน ปี พ.ศ. 2552 ถึงเดือนธันวาคม ปี พ.ศ. 2555 จากสถานีตำรวจจังหวัดสกลนคร ซึ่งอยู่ในรูปแบบเอกสาร ได้ข้อมูลจำนวน 700 รายการ และทำการนำข้อมูลออก 1 รายการเนื่องจากมีข้อมูลไม่ครบถ้วน

2. ทำการแปลงข้อมูลเพื่อให้เข้าสู่โปรแกรม Weka จำนวนข้อมูลที่ได้ทั้งหมด 699 ราย สามารถแบ่งออกได้เป็น 3 Classes คือ เสียชีวิต บาดเจ็บเล็กน้อย และบาดเจ็บสาหัส ซึ่งการเสียชีวิตมีอัตราส่วนเท่ากับ 52.23% บาดเจ็บเล็กน้อยมีอัตราส่วนเท่ากับ 3.72% และบาดเจ็บสาหัสมีอัตราส่วนเท่ากับ 38.05% ต่อจำนวนผู้ประสบอุบัติเหตุบนท้องถนนทั้งหมดสามารถสรุปได้ดัง Table 1

Table 1 Row data

Number			Total	Ratio		
Death	Minor Injuries	Seriously Injuries		Death	Minor Injuries	Seriously Injuries
407	26	266	699	52.23%	3.72%	38.05%

ในงานวิจัยฉบับนี้ข้อมูลประกอบด้วยตัวแปรทั้งหมด 13 ตัวแปร ดัง Table 2

Table 2 Variable

Attributes	Description	Values
Gender	Gender	1 = Male 2 = Female
Age	Age	Actual age.
Time	Accident time.	1 = 00.01-04.00 2 = 04.01-08.00 3 = 08.01-12.00 4 = 12.01-16.00 5 = 16.01-20.00 6 = 20.01-24.00
Vehicle_V	Vehicle accident of victims.	1 = Pedestrian 2 = Bicycle/Tricycle/Motor-tricycle 3 = Motorcycle 4 = Saloon 5 = Pick-Up 6 = Bus/van 7 = Truck/Trailer
Vehicle_P	Vehicles accident of the parties.	1 = Pedestrian 2 = Bicycle/Tricycle/Motor-tricycle 3 = Motorcycle 4 = Saloon 5 = Pick-Up 6 = Bus/van 7 = Truck/Trailer 8 = Other
Scene	Scene of the accident.	1 = Highway 2 = Local road 3 = Municipal road 4 = Household area road
Feature	Feature of the road.	1 = Straight 2 = Intersection 3 = Curve
Weather	Weather	1 = Fine 2 = Rainy 3 = Other
Light	Light	1 = Daylight 2 = Night 3 = With illuminating 4 = Without illuminating 5 = Other
Causes	The cause of the accidents.	1 = Driving over speed limit 2 = Driving in carelessness 3 = Following in close distance 4 = Chopping in close distance 5 = Drink and drive 6 = Illegal overtaking 7 = Driving in the wrong lane 8 = Other
Road_Surface	Road surface.	1 = Dry 2 = Wet 3 = Damaged
Area	Accident area.	1 = Mueang Sakon Nakhon 2 = Kusuman 3 = Kut Bak 4 = Phanna Nikhom 5 = Wa Rit Cha Phum 6 = Song Dao 7 = Sawang Daen Din 8 = Wanon Niwat 9 = Akat Amnuai 10 = Ban Muang 11 = Phang Khon 12 = Kham Ta Kla 13 = Nikhom Nam Un 14 = Tao Ngoi 15 = Khok Si Suphan 16 = Phon Na Kaeo 17 = Phon Na Kaeo 18 = Phu Phan
Severity_Ac	Severity of the accident.	Death Minor_Injuries Seriously_Injuries

2. การสร้างแบบจำลอง

ในการสร้างแบบจำลองงานวิจัยนี้ใช้โปรแกรม Weka 3.6.8 ซึ่งเป็นเครื่องมือที่มีเทคนิคในเหมืองข้อมูลหลายเทคนิค และมีการวัดผลที่มีประสิทธิภาพของแบบจำลอง งานวิจัยนี้ได้เลือกเทคนิควิธีดังต่อไปนี้ 1) ถุงจำแนก 2) เครื่องมือสนับสนุนเส้นสมมุติ 3) การเรียนรู้แบบเบย์ 4) ถุงจำแนกร่วมกับเครื่องมือสนับสนุนเส้นสมมุติ และ 5) ถุงจำแนกร่วมกับการเรียนรู้แบบเบย์ ซึ่งแต่ละกลุ่มจะทำ 10 รอบ

3. การวัดประสิทธิภาพ

การวิเคราะห์ประสิทธิภาพแบบจำลองเพื่อทำนายความเสี่ยงการเกิดอุบัติเหตุบนท้องถนน ในการวัดประสิทธิภาพของแบบจำลอง โดยวัดค่าความแม่นยำ และพื้นที่ใต้เส้นกราฟ ROC (AUC) ทำให้ทราบค่าความถูกต้องของวิธีการที่ทำการวิจัย และพัฒนาตรงตามวัตถุประสงค์ที่ตั้งไว้

ผลการดำเนินงานวิจัย

ผลการดำเนินงานวิจัยฉบับนี้ได้วิเคราะห์ถึงค่าแม่นยำ และค่า AUC ของแบบจำลองทำนายความเสี่ยงของการเกิดอุบัติเหตุทางถนน ในการวิเคราะห์ได้ใช้ขั้นตอนการทดลองจากการทำเหมืองข้อมูล โดยข้อมูลได้จากสถานีตำรวจจังหวัดสกลนคร ระหว่างเดือนพฤศจิกายน ในปี พ.ศ. 2552 ถึงเดือนธันวาคม ในปี พ.ศ. 2555 จากสถานีตำรวจจังหวัดสกลนครสกลนคร ใช้หลักการแยกด้วยวิธีการ 10-fold cross validation เทคนิคที่นำมาใช้ในการสร้างแบบจำลอง คือ 1) ถุงจำแนก (B) 2) เครื่องมือสนับสนุนเส้นสมมุติ (SVM) 3) การเรียนรู้แบบเบย์ (NB) 4) ถุงจำแนกร่วมกับเครื่องมือสนับสนุนเส้นสมมุติ (B+SVM) และ 5) ถุงจำแนกร่วมกับการเรียนรู้แบบเบย์ (B+NB)

1. การวัดประสิทธิภาพโดยวิธีการหาค่าความแม่นยำ ในการทำนายผลการเกิดอุบัติเหตุบนถนนเป็นพื้นฐานการวัดประสิทธิภาพ และประสิทธิผลของแบบจำลอง ซึ่งสามารถแสดงค่าความแม่นยำของการทำนายได้ดัง Table 3

Table 3 Accuracy of road accident risk prediction models.

แบบจำลอง	Accuracy (%)
B	67.67
SVM	66.81
NB	65.95
B+SVM	67.67
B+NB	67.10

จาก Table 3 แสดงถึงค่าความแม่นยำในการทำนายการเกิดอุบัติเหตุบนท้องถนนเป็นพื้นฐานการวัดประสิทธิภาพ จากผลการทดลองปรากฏว่าความแม่นยำจากแบบจำลองถุงจำแนกคิดเป็นร้อยละ 67.67 ซึ่งมีค่าความถูกต้องสูงสุด ตามด้วยแบบจำลองเครื่องมือสนับสนุนเส้นสมมุติมีความถูกต้องคิดเป็นร้อยละ 66.81 ส่วนแบบจำลองจำแนกเบย์ให้มีความถูกต้องน้อยที่สุดคิดเป็นร้อยละ 65.95 ตามลำดับ แต่เมื่อนำเทคนิคถุงจำแนกเข้ามาผสมร่วมกันเพื่อช่วยในการเพิ่มประสิทธิภาพของแบบจำลอง ทำให้แบบจำลองถุงจำแนกร่วมกับแบบจำลองจำแนกเบย์ให้ผลลัพธ์ความถูกต้องคิดเป็นร้อยละ 67.10 ซึ่งเพิ่มจากเดิมคิดเป็นร้อยละ 1.15 และแบบจำลองถุงจำแนกร่วมกับเครื่องมือสนับสนุนเส้นสมมุติให้ผลลัพธ์ความถูกต้องคิดเป็นร้อยละ 67.67 ซึ่งเพิ่มจากเดิมคิดเป็นร้อยละ 0.86

2. การวัดประสิทธิภาพโดยวิธีการหาค่าพื้นที่ใต้เส้นกราฟ ROC (AUC) ในการทำนายผลการเกิดอุบัติเหตุบนถนนเป็นพื้นฐานการวัดประสิทธิภาพ และประสิทธิผลของแบบจำลอง สามารถแสดงผลค่า AUC ได้ Table 4

Table 4 AUC of road accident risk prediction models.

แบบจำลอง	AUC (%)
B	71.00
SVM	66.40
NB	71.30
B+SVM	71.40
B+NB	71.60

จาก Table 4 แสดงค่า AUC ในการทำนายการเกิดอุบัติเหตุบนท้องถนนเป็นพื้นฐานการวัดประสิทธิภาพ จากผลการทดลอง ปรากฏว่าแบบจำลองจำแนกเบย์ให้ผลลัพธ์สูงสุดคิดเป็นร้อยละ 71.30 ซึ่งมีประสิทธิภาพดีที่สุด ตามด้วยแบบจำลองถุงจำแนกให้ผลลัพธ์คิดเป็นร้อยละ 71.00 ส่วนแบบจำลองเครื่องมือสนับสนุนเส้นสมมุติให้ผลลัพธ์น้อยที่สุดคิดเป็นร้อยละ 66.40 ตามลำดับ แต่เมื่อนำเทคนิคถุงจำแนกเข้ามาผสมร่วมกันเพื่อช่วยในการเพิ่มประสิทธิภาพของแบบจำลอง ทำให้แบบจำลองถุงจำแนกร่วมกับแบบจำลองจำแนกเบย์ให้ผลลัพธ์คิดเป็นร้อยละ 71.60 ซึ่งเพิ่มจากเดิมคิดเป็นร้อยละ 0.30 และแบบจำลองถุงจำแนกร่วมกับเครื่องมือสนับสนุนเส้นสมมุติให้ผลลัพธ์คิดเป็นร้อยละ 71.40 ซึ่งเพิ่มจากเดิมคิดเป็นร้อยละ 5.00

สรุปผลการวิจัย

จากจุดประสงค์ในการทำวิจัยครั้งนี้ ผู้วิจัยได้ทำการทดลองแบบจำลองให้เหมาะสมกับชุดข้อมูลการเกิดอุบัติเหตุบนท้องถนน โดยนำเอาเทคนิคถุงจำแนกเพิ่มประสิทธิภาพความถูกต้องให้กับเทคนิคจำแนกเบย์ และเทคนิคเครื่องมือสนับสนุนเส้นสมมุติ สามารถที่จะอภิปรายผลได้ดังนี้ แบบจำลองถุงจำแนกให้ค่าความถูกคิดเป็นร้อยละ 67.67 ซึ่งมีค่าความถูกต้องดีที่สุด แต่เมื่อนำเทคนิคถุงจำแนกเข้ามาผสมร่วมกันเพื่อช่วยในการเพิ่มประสิทธิภาพของแบบจำลอง ทำให้แบบจำลองถุงจำแนกร่วมกับแบบจำลองจำแนกเบย์ให้ผลลัพธ์ความถูกต้องคิดเป็นร้อยละ 67.10 ซึ่งเพิ่มจากเดิมคิดเป็นร้อยละ 1.15 และแบบจำลองถุงจำแนกร่วมกับเครื่องมือสนับสนุนเส้นสมมุติให้ผลลัพธ์ความถูกต้องคิดเป็นร้อยละ 67.67 ซึ่งเพิ่มจากเดิมคิดเป็นร้อยละ 0.86 และ AUC เป็นอีกทางเลือกหนึ่งที่ใช้ในการวัดประสิทธิภาพการทำงานของแบบจำลอง ปรากฏว่าแบบจำลองจำแนกเบย์ให้ผลลัพธ์คิดเป็นร้อยละ 71.30 ซึ่งให้ประสิทธิภาพของแบบจำลองดีที่สุด แต่เมื่อนำเทคนิคถุงจำแนกเข้ามาผสมร่วมกันเพื่อช่วยในการเพิ่มประสิทธิภาพของแบบจำลอง ทำให้แบบจำลองถุงจำแนกร่วมกับแบบจำลองจำแนกเบย์ให้ผลลัพธ์คิดเป็นร้อยละ 71.60 ซึ่งเพิ่มจากเดิมคิดเป็นร้อยละ 0.30 และแบบจำลองถุงจำแนกร่วมกับเครื่องมือสนับสนุนเส้นสมมุติให้ผลลัพธ์คิดเป็นร้อยละ 71.40 ซึ่งเพิ่มจากเดิมคิดเป็นร้อยละ 5.00 ซึ่งจะเห็นว่าเทคนิคถุงจำแนกสามารถช่วยให้เทคนิคพื้นฐานสร้างแบบจำลองที่มีประสิทธิภาพมากขึ้นได้เป็นอย่างดี

ข้อเสนอแนะ

ในอนาคตผู้วิจัยมีแนวคิดพัฒนาแบบจำลองเพื่อเพิ่มประสิทธิภาพให้กับแบบจำลองในการทำนาย โดยใช้หลักการอื่นๆ เช่น การกรองข้อมูลที่ไม่สำคัญกับการทำนาย ซึ่งเป็นวิธีการหนึ่งในการเพิ่มประสิทธิภาพของแบบจำลอง

กิตติกรรมประกาศ

ขอขอบคุณสถานีตำรวจภูธรจังหวัดสกลนคร สกลนคร ที่ให้ความอนุเคราะห์ข้อมูลในการศึกษาครั้งนี้

เอกสารอ้างอิง

1. สถาบันนิติเวชวิทยา. อุบัติเหตุจราจรทางบก. สืบค้นเมื่อ 19 พฤศจิกายน 2554]; Available from: http://www.ifm.go.th/2010/index.php.
2. สำนักงานอำนวยความสะดวกปลอดภัย กรมทางหลวง. ผู้ประสบภัย. 2556. สืบค้นเมื่อ 25 กุมภาพันธ์ 2556]; Available from: http://bhs.doh.go.th/statistic/casualty.
3. Tian R, Yang Z, Zhang M. Method of Road Traffic Accidents Causes Analysis Based on Data Mining. International Conference Computational Intelligence and Software Engineering. 2010:p.1- 4
4. Krishnaveni S, Hemalatha M. A Perspective Analysis of Traffic Accident using Data Mining Techniques. International Journal of Computer Applications. 2011;Vol. 23 (7) :pp.40-8.

5. Thongkam J, Sukmak V. Bagging Random Tree for Analyzing Breast Cancer Survival. KKU Research Journal 2012;Vol.17(1):pp.1-13.
6. Poel DVd, D'Haen J, Thorleuchter D. Predicting customer profitability during acquisition : Finding the optimalcombination of data source and data mining technique. Expert Systems with Applications. 2013;Vol. 40:pp.2007-12.
7. Dong-xiao N, Yong-li W. Support Vector Machines Based on Data Mining Technology in Power Load Forecasting. International Conference on Wireless Communications, Networking and Mobile Computing. 2007:pp. 5373 - 6
8. Breiman L. Bagging predictors. Mach Learn. 1996;24:pp. 123–40.
9. Christophe C, Kristel J, Aurélie L. Trimmed bagging. Journal Computational Statistics & Data Analysis. 2007;Vol. 52:pp. 362 – 8.
10. Secchi P, Vantini S, Vitelli V. Bagging Voronoi classifiers for clustering spatial functional data. International Journal of Applied Earth Observation and Geoinformation 2013;Vol. 22:pp. 53–64.
11. Peter C. Austina, Jack V. Tua, Jennifer E. Hoe, Daniel Levye, Douglas S. Leea. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. Journal of Clinical Epidemiology. 2013 Vol.66:pp. 398-407.
12. Crouxa C, Joossensa K, Lemmensb A. Trimmed bagging. Computational Statistics and Data Analysis 2007;Vol. 52:pp. 362 – 8.
13. Zhang J, Chen C, Xiang Y, Zhou W, Xiang Y. Internet Traffic Classification by Aggregating Correlated Naive Bayes Predictions. IEEE Transactions on Information Forensics and Security. 2013;Vol. 8(1):pp. 5-15.
14. Jiang L, Zhang H, Cai Z. A Novel Bayes Model : Hidden Naive Bayes. IEEE Transactions on Knowledge and Data Engineering. 2009;Vol. 21(10):pp. 1361-71.
15. Kim S-B, Han K-S, Rim H-C, Myaeng SH. Some Effective Techniques for Naive Bayes Text Classification. IEEE Transactions on Knowledge and Data Engineering. 2006;Vol. 8(11):pp. 1457-66.
16. Nagi J, Yap KS, Tiong SK. Nontechnical Loss Detection for Metered Customers in Power Utility Using Support Vector Machines. IEEE Transactions on Power Delivery. 2010;Vol. 25(2):pp. 1162-70.
17. Xue Z, Sun X, Liang Y. Application of Data Mining Technology Based on FRS and SVM for Fault Identification of Power Transformer. International Conference on Artificial Intelligence and Computational Intelligence. 2009 Vol. 2:pp. 452- 5.
18. Mavroforakis ME, Theodoridis S. A Geometric Approach to Support Vector Machine (SVM) Classification. IEEE Transactions on Neural Neteorks. 2006;Vol. 17(3):pp. 671- 82
19. Dodek PM, Wiggs BR. Logistic regression model to predict outcome after in-hospital cardiac arrest: validation, accuracy, sensitivity and specificity. Resuscitation. 1998;Vol. 36(3):pp. 201-8.
20. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters 2006 Vol. 27:pp. 861–74.

21. Smoot BJ, Wong JF, Dodd MJ. Comparison of Diagnostic Accuracy of Clinical Measures of Breast Cancer–Related Lymphedema : Area Under the Curve. Archives of Physical Medicine and Rehabilitation 2011;Vol. 92(4):pp. 603-10.
22. Chang L-Y, Wang H-W. Analysis of traffic injury severity : An application of non-parametric classification tree techniques. Journal Accident Analysis and Prevention 2006;Vol. 38:pp. 1019–27.
23. Zhang D, Zhou X, Leung SCH, Zheng J. Vertical bagging decision trees model for credit scoring. Expert Systems with Applications. 2010;Vol. 37:pp. 7838 – 43.