Prediction model for stroke severity using data mining techniques from administrative data

Chidapha Traicharoenwong, Surasak Mungsing

School of Information Technology, Sripatum University, Thailand

Abstract

This article presents the introduction of information for administration. (Administrative Data) with 30 independent variables to select the variables that have the power to identify the severity of illness with stroke. The National Institutes of Health Stroke Scale (NIHSS) is divided into 3 levels which are Mild (NIHSS 0-10 points), Moderate (NIHSS 11-20 points) and Severe (NIHSS 21-42 points). Data mining found that there are 5 variables, including periods of high blood lipid levels Atrial Fibrillation (AF), Glasgow Coma Scale (GCS), Barthel Index (BI) and mRS level, can identify the severity of stroke. The accuracy of 85.2 percent, which the classification helps to provide medical resources, including personnel, tools, treatment methods that are different in each group of patients appropriately.

Keywords: Stroke, Severity, Adminitrative Data, Model, Data Mining

Received 23 March 2020; Accepted 25 May 2020

Correspondence: Surasak Mungsing, School of Information Technology, Sripatum University, 2410/2 Phaholyothin Road, Jatujak, Bangkok 10900 (Tel.: +66-2579-1111 ext. 3040; E-mail address: smungsing@gmail.com)

ตัวแบบพยากรณ์ความรุนแรงของอาการป่วยด้วยโรคหลอดเลือดสมอง โดยใช้เทคนิคการทำเหมืองข้อมูลจากข้อมูลเพื่อการบริหาร

จิดาภา ตรัยเจริญวงศ์, สุรศักดิ์ มังสิงห์

คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยศรีปทุม

บทคัดย่อ

บทความนี้ นำเสนอการนำข้อมูลเพื่อการบริหาร (Administrative Data) โดยมีตัวแปรอิสระจำนวน 30 ตัวแปร มาทำการคัดเลือกตัวแปรที่มีอำนาจในการจำแนกความรุนแรงของอาการป่วยด้วยโรคหลอดเลือดสมอง โดยพิจารณาจากค่าคะแนน The National Institutes of Health Stroke Scale (NIHSS) แบ่งเป็น 3 ระดับ ได้แก่ Mild (NIHSS 0-10 คะแนน), Moderate (NIHSS 11-20 คะแนน) และ Severe (NIHSS 21-42 คะแนน) ด้วยวิธีการทำเหมืองข้อมูล (Data Mining) พบว่า มี 5 ตัวแปร ได้แก่ ระยะเวลาที่มีระดับไขมันในเลือดสูง อาการหัวใจเต้นผิดจังหวะ (Atrial Fibrillation: AF) คะแนน Glasgow Coma Scale (GCS), Barthel Index (BI) และระดับ mRS สามารถจำแนกความรุนแรงของอาการป่วยด้วยโรคหลอดเลือดสมอง มีความถูกต้องถึงร้อยละ 85.2 ซึ่งการจัดจำแนกดังกล่าวช่วยให้สามารถจัดเตรียมทรัพยากรทางการแพทย์ ได้แก่ บุคลากร เครื่องมือ แนวทางการรักษาที่มีความแตกต่างกันในแต่ละกลุ่มของผู้ป่วยได้อย่างเหมาะสม

คำสำคัญ: โรคหลอดเลือดสมอง, ความรุนแรงของอาการป่วย, ข้อมูลเพื่อการบริหาร, ตัวแบบพยากรณ์, เหมืองข้อมูล

วันที่รับต้นฉบับ 23 มีนาคม 2563; วันที่ตอบรับ 25 พฤษภาคม 2563

บทนำ

โรคหลอดเลือดสมอง (Cerebrovascular Disease, Stroke) หรือโรคอัมพฤกษ์ อัมพาต เป็นโรคทางระบบประสาทที่พบบ่อย จากข้อมูลองค์กรอนามัยโลก (World Health Organization: WHO) ปี 2016 [1] พบว่า โรคหลอดเลือดสมองเป็นสาเหตุการเสียชีวิตลำดับที่ 2 รองจากโรคหัวใจขาดเลือด มีผู้เสียชีวิตประมาณ 5.8 ล้านคน คิดเป็นอัตรา 77 คนต่อ 100,000 ประชากร สำหรับประเทศไทยตามรายงานตัวชี้วัดกระทรวงสาธารณสุข ปีงบประมาณ 2562 พบว่า อัตราตายของผู้ป่วยในด้วยโรคหลอดเลือดสมอง (I60-I69) คิดเป็นร้อยละ 7.9 [2] ถ้าพิจารณาจำแนกรายกลุ่มโรค ดังนี้ โรคหลอดเลือดสมองแตก (I60-I62) อัตราตายร้อยละ 22.7 โรคหลอดเลือดสมองตีบหรืออุดตัน (I63) อัตราตายร้อยละ 3.8 โรคหลอดเลือดสมองอื่น ๆ (I64-I69) อัตราตายร้อยละ 2.2 และเป็นโรคที่เป็นปัญหาสาธารณสุขที่สำคัญของประเทศไทย

ในปัจจุบันมีการจัดเก็บข้อมูลเพื่อการบริหาร (Administrative Data) เป็นจำนวนมากในหลายมิติ เพื่อใช้เป็นข้อมูลประกอบการวางแผน การตัดสินใจในการดำเนินงานขององค์กร

ทบทวนวรรณกรรม

การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูลเป็นกระบวนการกลั่นกรองสารสนเทศ (Information) ที่ซ่อนอยู่ในฐานข้อมูลขนาดใหญ่ เพื่อทำนายแนวโน้ม และพฤติกรรม โดยอาศัยข้อมูลในอดีต เพื่อใช้ในการสนับสนุนการตัดสินใจ

เหมืองข้อมูล เป็นขั้นตอนที่สำคัญในกระบวนการค้นพบความรู้แฝงของข้อมูล ที่มีประโยชน์ในฐานข้อมูล จะประกอบด้วยขั้นตอนต่าง ๆ ดังนี้

ผู้นิพนธ์ประสานงาน: สุรศักดิ์ มังสิงห์, คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยศรีปทุม 2410/2 ถนนพหลโยธิน เขตจตุจักร กรุงเทพมหานคร 10900 (โทร.: 0-2579-1111 ต่อ 3040; E-mail address: smungsing@gmail.com)

1. การคัดเลือกข้อมูล (Data Selection) คือ การระบุถึงแหล่งข้อมูลที่จะนำมาใช้ในการทำเหมืองข้อมูล
2. การกรองข้อมูล (Data Cleaning) คือ กระบวนการที่ทำให้เกิดความมั่นใจในคุณภาพของข้อมูลที่จะนำมาใช้วิเคราะห์ว่าถูกต้องโดยการนำข้อมูลที่ไม่ถูกต้องออก
3. การแปลงรูปข้อมูล (Data Transformation) เป็นการแปลงข้อมูลที่เลือกมาให้อยู่ในรูปแบบที่เหมาะสมสำหรับการนำไปใช้ในการวิเคราะห์ตามอัลกอริทึมที่ใช้ในการทำเหมืองข้อมูลต่อไป
4. การทำเหมืองข้อมูล (Data Mining) เป็นการใช้เทคนิคภายในการทำเหมืองข้อมูลโดยทั่วไป ประเภทของงานตามลักษณะของแบบจำลองที่ใช้ในการทำเหมืองข้อมูล สามารถแบ่งออกได้ 2 ประเภทใหญ่คือ

4.1 แบบจำลองเชิงทำนาย (Predictive Data Mining) เป็นการคาดคะเนลักษณะหรือประมาณค่าที่ชัดเจนของข้อมูลที่จะเกิดขึ้นโดยใช้พื้นฐานจากข้อมูลที่ผ่านมาในอดีต

4.2 แบบจำลองเชิงพรรณนา (Descriptive Data Mining) เป็นการหาแบบจำลองเพื่ออธิบายลักษณะบางอย่างของข้อมูลที่มีอยู่ โดยส่วนมากจะเป็นลักษณะการแบบกลุ่มให้กับข้อมูล

5. การวิเคราะห์และประเมินผลลัพธ์ที่ได้ (Result Analysis and Evaluation) เป็นขั้นตอนการแปลความหมายและการประเมิน

ผลลัพธ์ที่ได้ว่ามีความเหมาะสม หรือตรงกับวัตถุ ประสงค์ที่ต้องการหรือไม่ โดยทั่วไปควรมีการแสดงผลอยู่ในรูปแบบที่สามารถเข้าใจได้ง่าย

การจำแนกประเภท (Classification) ใช้สำหรับจัดการข้อมูล การสร้างแบบจำลองของข้อมูล ช่วยในการคาดการณ์เกี่ยวกับข้อมูลใหม่ งานวิจัยนี้ [4] เน้นที่อัลกอริทึม J48 ซึ่งใช้ในการสร้างแผนภูมิการตัดสินใจแบบ Univariate การศึกษาวิจัยยังหารือเกี่ยวกับแนวคิดของต้นไม้ตัดสินใจหลายตัวแปรด้วยกระบวนการจำแนกอินสแตนซ์โดยใช้แอตทริบิวต์มากกว่าหนึ่งรายการในแต่ละโหนดภายใน แนวคิดหลักที่อยู่เบื้องหลังหัวข้อนี้คือ การได้รับความรู้เชิงลึกด้วยการวิจัยในพื้นที่ใหม่โดยการสำรวจเพิ่มเติมเกี่ยวกับข้อมูลข้อมูลความรู้เทคนิคการขุดข้อมูล เครื่องมือ และผลลัพธ์ได้รับการตรวจสอบ ด้วยโปรแกรม Weka

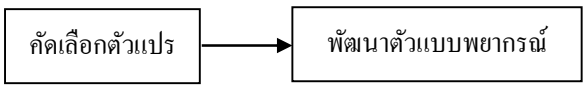
การลดจำนวนตัวแปร (Dimension Reduction) [5]

ในการจับกลุ่มข้อมูล ถ้าข้อมูลมีตัวแปรจำนวนมาก จะประสบปัญหาที่เรียกว่า Curse of Dimensionality คือ การที่ตัวแปรมีจำนวนมาก อาจมีตัวแปรที่ไม่จำเป็นถูกรวมเข้ามาใช้ในการวิเคราะห์ ทำให้ระยะห่างหรือ Dissimilarity ระหว่างทุกคู่ของข้อมูลมีค่าพอ ๆ กัน เนื่องจากข้อมูลถูกทำให้ห่างจากกันหรือกระจายออกไปด้วยค่าของตัวแปรที่ไม่จำเป็น รูปแบบของกลุ่มข้อมูลจึงถูกเปลี่ยนแปลงไปจากเดิม ในการสร้างตัวแบบสำหรับการจำแนกประเภทข้อมูล การนำตัวแปรที่ไม่จำเป็นเข้ามาร่วมในการสร้างตัวแบบ จะทำให้ได้ตัวแบบที่ไม่ถูกต้อง ส่งผลให้ความแม่นยำในการจำแนกประเภทลดน้อยลง การลดจำนวนตัวแปรมีความสำคัญอย่างยิ่งกับข้อมูลที่มีขนาดใหญ่

วัตถุประสงค์การวิจัย

เพื่อพัฒนาตัวแบบพยากรณ์ความรุนแรงของอาการป่วยด้วยโรคหลอดเลือดสมองโดยใช้เทคนิคการทำเหมืองข้อมูลจากข้อมูลเพื่อการบริหาร

กรอบแนวคิดในการวิจัย



รูปที่ 1 ขั้นตอนการวิจัย

ในการศึกษาครั้งนี้ ได้นำข้อมูลเพื่อการบริหารจากฐานข้อมูลคุณภาพการให้บริการโรคหลอดเลือดสมองตีบหรืออุดตัน Registry of Stroke Care Quality จำนวน 31 ตัวแปร [3] ได้แก่ (1) ภาค (2) เขตบริการสุขภาพ (3) จังหวัด (4) ระดับสถานพยาบาล (5) สถานพยาบาล (6) อายุ (7) เพศ (8) การรับส่งต่อ (9) ประวัติโรคเบาหวาน (10) ระยะเวลาการเป็นโรคเบาหวาน (11) ประวัติโรคความดันโลหิตสูง (12) ระยะเวลาการเป็นโรคความดันโลหิตสูง (13) ประวัติโรคไขมันในเลือดสูง (14) ระยะเวลาการเป็นโรคไขมันในเลือดสูง (15) ประวัติโรคหัวใจเต้นผิดจังหวะ (Atrial Fibrillation: AF) (16) ระยะเวลาการเป็นโรคหัวใจเต้นผิดจังหวะ (17) ประวัติการเป็นโรคหลอดเลือดสมอง (18) ชนิดของโรคหลอดเลือดสมอง (19) ประวัติการสูบบุหรี่ (20) ระยะเวลา 6 เดือนที่ผ่านมายังคงสูบบุหรี่หรือไม่ (21) ปริมาณบุหรี่ที่สูบ (มวน/วัน) (22) ประวัติการดื่มแอลกอฮอล์ (23) ระยะเวลา 6 เดือนที่ผ่านมายังคงดื่มแอลกอฮอล์หรือไม่ (24) จำนวนครั้งที่ดื่มต่อเดือน (25) การวินิจฉัยแรกรับ (26) คะแนน Glasgow Coma Scale (GCS) แรกรับ (27) Barthel Index (0-100) แรกรับ (28) ระดับ mRS แรกรับ (29) ระดับน้ำตาลในเลือด (mg%) (30) ระดับไขมัน LDL (mg/dl) (31) คะแนน NIHSS แรกรับ มาพิจารณาว่าตัวแปรใดมีอำนาจในการจำแนกความรุนแรงของอาการป่วยด้วยโรคหลอดเลือดสมอง แบ่งเป็น 3 ระดับ ได้แก่ Mild (NIHSS 0-10 คะแนน), Moderate (NIHSS 11-20 คะแนน) และ Severe (NIHSS 21-42 คะแนน) ทำการคัดเลือกตัวแปรด้วยการ Select attributes โดยใช้วิธี Best First [6] เป็นการตั้งค่าของ AND/OR กราฟได้รับการพิจารณาที่ทั้งค่าต่ำสุดและค่าสูงสุดพร้อมกันนำไปสู่การพัฒนาอัลกอริทึมการค้นหากราฟ AND/OR ที่เรียกว่า GEN-AO Works ซึ่งทำงานกับโหนด OR 2 ประเภท คือ MIN และ MAX และใช้การประมาณทั้งขอบเขตบนและขอบเขตล่างทำให้อัลกอริทึมนี้ได้จำนวนตัวแปรที่เหมาะสม กรณีที่แย่ที่สุดคือมีจำนวน Node เท่ากับจำนวนตัวแปรที่ทดลอง อัลกอริทึมที่เป็นที่รู้จัก คือ การ Pruning ด้วยวิธี Depth-first

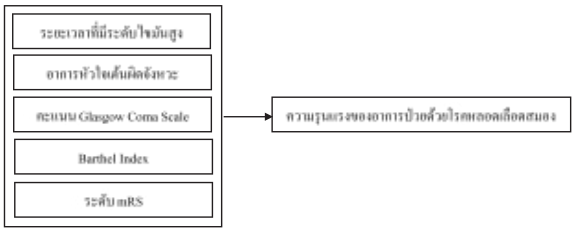
Evaluator: weka.attributeSelection.CfsSubsetEval -P 1 -E 1
Search: weka.attributeSelection.BestFirst -D 1 -N 5
Instances: 106733
Attributes: 31
Evaluation mode: evaluate on all training data
Search Method:
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 218
Merit of best subset found: 0.264
Attribute Subset Evaluator (supervised, Class (nominal): 31 C):
CFS Subset Evaluator
Including locally predictive attributes
Selected attributes: 14, 15, 26, 27, 28: 5
TDLD
AF
GCS
BI
mRS

ตัวแปรที่ได้รับคัดเลือกมี 5 ตัวแปร ได้แก่ ระยะเวลาที่มีระดับไขมันในเลือดสูง อาการหัวใจเต้นผิดจังหวะ (Atrial Fibrillation: AF) คะแนน Glasgow Coma Scale (GCS), Barthel Index (BI) และระดับ mRS

ทำการทดลองเปรียบเทียบความถูกต้องในการจำแนกความรุนแรงของอาการป่วยด้วยโรคหลอดเลือดสมองโดยใช้ตัวแปรพยากรณ์ทั้ง 30 ตัว เปรียบเทียบกับตัวแปรที่ได้รับคัดเลือกเพียง 5 ตัว โดยใช้กฎ ZeroR ให้ผลการพยากรณ์มีความถูกต้องเท่ากันถึงร้อยละ 79.1

Scheme: weka.classifiers.rules.ZeroR
Instances: 106733
Attributes: 31
Test mode: 10-fold cross-validation
=== Summary ===
Correctly Classified Instances 84432 79.1058 %
Incorrectly Classified Instances 22301 20.8942 %

Scheme: weka.classifiers.rules.ZeroR
Instances: 106733
Attributes: 6
Test mode: 10-fold cross-validation
=== Summary ===
Correctly Classified Instances 84432 79.1058 %
Incorrectly Classified Instances 22301 20.8942 %

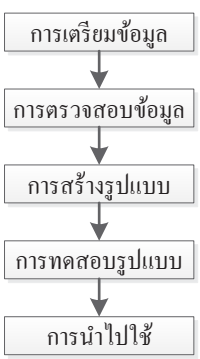


รูปที่ 2 กรอบแนวคิดในการวิจัย (Conceptual Framework)

ขอบเขตของการวิจัย

ผู้วิจัยศึกษาจากฐานข้อมูลคุณภาพการให้บริการโรคหลอดเลือดสมอง Registry of Stroke Care Quality ของสถาบันประสาทวิทยา ซึ่งรวบรวมข้อมูลผู้ป่วยโรคหลอดเลือดสมองจากโรงพยาบาลเครือข่ายที่เข้าร่วมโครงการ 125 แห่ง ในประเทศ 124 แห่ง และประเทศเมียนมา 1 แห่ง ตัวแปรที่ศึกษาในงานวิจัยนี้มี (1) ตัวแปรอิสระ ได้แก่ ระยะเวลาที่มีระดับไขมันสูง อาการหัวใจเต้นผิดจังหวะ คะแนน Glasgow Coma Scale (GCS), Barthel Index (BI) และระดับ mRS (2) ตัวแปรตาม คือ ความรุนแรงของอาการป่วยด้วยโรคหลอดเลือดสมองขณะแรกรับ โดยวัดจากคะแนน NIHSS แบ่งเป็น 3 ระดับ ได้แก่ Mild (NIHSS 0-10 คะแนน), Moderate (NIHSS 11-20 คะแนน) และ Severe (NIHSS 21-42 คะแนน) โดยการทำเหมืองข้อมูล ด้วยวิธีจำแนกประเภทข้อมูล (Data Classification) ศึกษาจากข้อมูลผู้ป่วยโรคหลอดเลือดสมอง จำนวน 106,733 ราย จากปี พ.ศ.2554 ถึงปี พ.ศ. 2562

การทดลอง



รูปที่ 3 ขั้นตอนการพัฒนารูปแบบ

การเตรียมข้อมูล ประกอบด้วย การคัดเลือกตัวแปร การแปลงข้อมูลให้มีรูปแบบที่เหมาะสมในการทดลอง ดังตารางที่ 1.

ตารางที่ 1 ตัวอย่างชุดข้อมูล

ID	TDLD	AF	GCS	BI	mRS	Class
1	0	N	15	70	3	1
2	0	N	15	100	1	1
3	0	N	10	40	4	2
4	0	N	15	80	4	1
5	0	N	14	40	4	2
6	0	N	15	100	1	1
7	0	N	15	35	3	1
8	0.1	N	15	65	3	1
9	0	N	15	100	2	1
10	0	N	15	60	4	1

จากข้อมูลผู้ป่วยโรคหลอดเลือดสมอง จำนวน 180,161 ราย เมื่อทำการเตรียมข้อมูล และตรวจสอบข้อมูล มีรายการที่มีข้อมูลครบถ้วนจำนวน 106,733 ราย ดังตารางที่ 2

ตารางที่ 2 ลักษณะผู้ป่วยโรคหลอดเลือดสมอง จำนวน 106,733 ราย

	Mean (S.D.)		ร้อยละ
ระยะเวลาเฉลี่ยที่มีระดับไขมันสูง	1.7 (4.0)		
อาการหัวใจเต้นผิดจังหวัด			8.5
คะแนน Glasgow Coma Scale	14.2 (2.0)		
Barthel Index (BI)	62.9 (30.1)		
ระดับ mRS	2.9 (1.7)		
ความรุนแรงของอาการป่วยด้วย		0-10	79.1
โรคหลอดเลือดสมอง		11-20	14.8
		21-42	6.1

พัฒนาตัวแบบพยากรณ์ความรุนแรงของอาการป่วยด้วยโรคหลอดเลือดสมองโดยใช้เทคนิคการทำเหมืองข้อมูล ด้วยวิธีการจำแนกประเภทของข้อมูล (Classification) โดยใช้ต้นไม้การตัดสินใจ (Decision Tree) การสร้างต้นไม้การตัดสินใจข้อมูลเรียนรู้ จะต้องการได้ต้นไม้ที่ให้ความถูกต้องมากที่สุด ในการจำแนกประเภทของข้อมูลเรียนรู้ และข้อมูลทดสอบ รวมทั้งมีความลึกของต้นไม้ที่น้อยที่สุด (ทำให้การตัดสินใจมีความรวดเร็ว เนื่องจากมีจำนวนครั้งที่ใช้ในการตัดสินใจน้อยที่สุด

การสร้างต้นไม้จะเริ่มจากการสร้างโหนดรากและไล่ไปยังโหนดลูกจนถึงโหนดใบ โดยการเลือกตัวแปรและเงื่อนไขที่จะใช้ในการตัดสินใจสำหรับโหนดใด ๆ ในต้นไม้ การเลือกตัวแปรและเงื่อนไขจะคำนึงถึงการได้มาซึ่งต้นไม้ที่สามารถจำแนกประเภทของข้อมูลได้ถูกต้องมากที่สุด รวมทั้งมีความลึกของต้นไม้น้อยที่สุดด้วย การสร้างเงื่อนไขสำหรับตัวแปรแต่ละตัวจะแตกต่างตามชนิดของตัวแปร และจำนวนลูกของโหนดที่ต้องการสร้าง

การสร้างต้นไม้การตัดสินใจจนเป็นต้นไม้โตเต็มที่ (Full Grown Tree) ที่ทุกโหนดใบมีข้อมูลประเภทเดียวกันหมดย่อมจะได้ต้นไม้ที่มีความเหมาะสม (Fit) ที่สุดสำหรับข้อมูลเรียนรู้ ทำให้การจำแนกประเภทข้อมูลสำหรับข้อมูลเรียนรู้ไม่มีความผิดพลาดเลย แต่อย่างไรก็ตาม ต้นไม้ที่ได้อาจไม่ใช่ต้นไม้ที่ดีที่สุดสำหรับการจำแนกประเภทข้อมูลทดสอบ หรือข้อมูลที่ต้องการจำแนกประเภทในอนาคต อาจเกิดปัญหา Overfitting ซึ่งสาเหตุอาจเกิดจากต้นไม้ที่สร้างได้ ประกอบด้วยโหนดใบที่ถูกสร้างขึ้นเพื่อรองรับข้อมูลที่เป็น Noise โดยเฉพาะ เมื่อนำข้อมูลที่มีค่าตัวแปรใกล้เคียงกับข้อมูล Noise เหล่านี้มาจำแนกประเภทจะทำให้การจำแนกประเภทไปตกที่โหนดใบเหล่านี้ ทำให้เกิดการจำแนกประเภทผิดพลาด การแก้ปัญหานี้ทำได้โดยการทำ Tree Pruning ทำให้ต้นไม้ที่สร้างได้ไม่เป็นต้นไม้ที่โตเต็มที่ จะไม่ก่อให้เกิดปัญหา Overfitting สามารถทำได้ 2 วิธี คือ

1. Pre Pruning เป็นการหยุดการสร้างโหนดลูกก่อนที่นำไปสู่ต้นไม้เต็มที่ การหยุดทำได้โดยกำหนดเงื่อนไขการหยุดการสร้างโหนดลูกโดยไม่จำเป็นจะต้องได้ชุดข้อมูลที่มีข้อมูลประเภทเดียวกันทั้งหมด โดยกำหนดค่า Threshold ของตัวชี้วัดความหลากหลายของประเภทข้อมูลในโหนดปัจจุบัน ถ้าค่าตัวชี้วัดมีค่าต่ำกว่าค่า Threshold จะหยุดการสร้างโหนดลูก
2. Post Pruning เป็นการสร้างต้นไม้จนเต็มที่และจึงตัดโหนดหรือต้นไม้ย่อยทิ้ง เพื่อให้ต้นไม้มีขนาดเล็กลง สามารถทำได้ 2 ลักษณะ คือ

2.1 แทนต้นไม้ย่อย (Subtree) ทั้งต้นด้วยโหนดใบ

2.2 แทนต้นไม้ย่อยทั้งต้นด้วยต้นไม้ย่อยต้นใดต้นหนึ่งของโหนดรากเดิมของต้นไม้ย่อยที่จะถูกแทนที่ โดยต้นไม้ย่อยของโหนดรากเดิมมีการใช้งานในการตัดสินใจประเภทของข้อมูลเรียนรู้ และข้อมูลทดสอบเป็นจำนวนครั้งมากที่สุด เมื่อเทียบกับการใช้งานของต้นไม้ย่อยต้นอื่น ๆ ของโหนดรากเดิม จึงสามารถตัดทิ้งได้โดยไม่ส่งผลต่อความถูกต้องของการจำแนกประเภทข้อมูลโดยรวม

การ Post Pruning ต้นไม้การตัดสินใจด้วยวิธีความเสี่ยงน้อยที่สุดของ Bayes [7] เป็นการประเมินอัตราความเสี่ยง ด้วยวิธีดำเนินการจากล่างขึ้นบน การแปลง Parent Node ของ Subtree เป็น Leaf Node หากอัตราความเสี่ยงโดยประมาณของ Parent Node ของ Subtree นั้นน้อยกว่า อัตราความเสี่ยงของ Leaf บทความนี้เสนอวิธีการ Pruning ที่พิจารณามาตรฐานการประเมินต่าง ๆ เช่น การเลือกคุณลักษณะความถูกต้อง ความซับซ้อนของ Tree และเวลาที่ใช้ในการ Pruning คะแนนความแม่นยำ/การเรียกคืนค่า (TP/FN) และพื้นที่ภายใต้ ROC ผลการทดลองแสดงให้เห็นว่าวิธีการที่เสนอนั้นสร้างความถูกต้องในการจำแนกประเภทที่ดีขึ้นและความซับซ้อนไม่แตกต่างจากความซับซ้อนของการตัดลดความผิดพลาดที่ลดลง และวิธีการตัดที่ผิด

พลาดน้อยที่สุด การทดลองยังแสดงให้เห็นว่าวิธีการที่นำเสนอแสดงให้เห็นถึงผลการดำเนินงานที่น่าพอใจในแง่ของคะแนนความแม่นยำ/การเรียกคืนค่า (TP/FN) และพื้นที่ภายใต้ ROC

การทดลองนี้จะใช้อัลกอริทึม

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: ReductionAtt1
Attributes: 6
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree

Number of Leaves : 214
Size of the tree : 427
Time taken to build model: 5.94 seconds
=== Stratified cross-validation ===
=== Summary ===

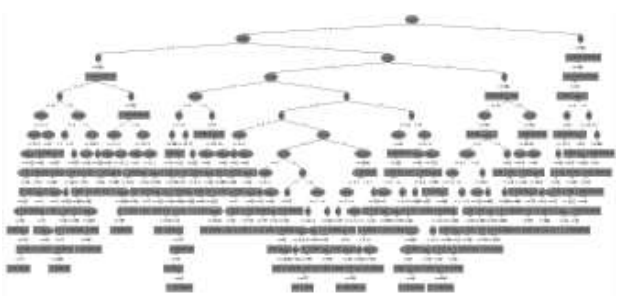
Correctly Classified Instances	90929	85.193 %
Incorrectly Classified Instances	15804	14.807 %
Kappa statistic	0.5407	
Mean absolute error	0.1386	
Root mean squared error	0.265	
Relative absolute error	59.6278 %	
Root relative squared error	77.7393 %	
Total Number of Instances	106733	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.956	0.350	0.912	0.956	0.933	0.654	0.909	0.962	1
	0.471	0.069	0.543	0.471	0.504	0.426	0.854	0.480	2
	0.429	0.017	0.617	0.429	0.506	0.489	0.925	0.543	3
Weighted Avg.	0.852	0.288	0.839	0.852	0.844	0.611	0.901	0.865	

=== Confusion Matrix ===

a	b	c	<– classified as
80699	3361	372 |	a = 1
7036 7457	1350 |	b = 2	
7702915	2773|	c = 3	



รูปที่ 4 ต้นไม้การตัดสินใจ

จากรูปที่ 4 พบว่าคะแนน GCS เป็น Root คะแนน ≤ 14 แตกมาเป็น Node แรก ด้วยตัวแปร ระดับ mRS ≤ 3 แตกเป็น Node ถัดไปด้วยตัวแปร BI ≤ 63 แตกต่อไปเรื่อย ๆ ดังตัวอย่างเส้นทางต่อไปนี้

1. GCS <= 14, mRS <= 3, BI <= 63, GCS <= 11, BI <= 32, mRS <= 2, mRS <= 0, TDLD <= 1.5, TDLD <= 0, BI <= 21, GCS <= 6: 3 (Severe)
2. GCS <= 14, mRS <= 3, BI <= 63, GCS <= 11, BI <= 32, mRS <= 2, mRS <= 0, TDLD <= 1.5, TDLD <= 0, BI <= 21, GCS > 6, BI <= 12, BI <= 7: 1 (Mild)
3. GCS <= 14, mRS <= 3, BI <= 63, GCS <= 11, BI <= 32, mRS <= 2, mRS <= 0, TDLD <= 1.5, TDLD <= 0, BI <= 21, GCS > 6, BI <= 12, BI > 7: 3 (Severe)
4. GCS <= 14, mRS <= 3, BI <= 63, GCS <= 11, BI <= 32, mRS <= 2, mRS <= 0, TDLD <= 1.5, TDLD <= 0, BI <= 21, GCS > 6, BI > 12: 1 (Mild)
5. GCS <= 14, mRS <= 3, BI <= 63, GCS <= 11, BI <= 32, mRS <= 2, mRS <= 0, TDLD <= 1.5, TDLD <= 0, BI <= 21: 1 (Mild)

…

จากการทดลองนี้มีจำนวน 214 เส้นทาง และเงื่อนไขการจำแนก 427 เงื่อนไข ตัวแบบการพยากรณ์นี้ได้มีการทำ Pruning มีความสามารถในการจำแนกได้ถูกต้องถึงร้อยละ 85.2

ตัวแบบการพยากรณ์นี้ สามารถนำไปใช้ประโยชน์ในการจำแนกกลุ่มผู้ป่วยโรคหลอดเลือดสมองในการวินิจฉัยครั้งแรกว่าอาการรุนแรงระดับใด เพื่อใช้ในการจัดเตรียมทรัพยากรทางการแพทย์ ได้แก่ สหวิชาชีพที่เกี่ยวข้องในการดูแลรักษาผู้ป่วย เครื่องมือทางการแพทย์ การวางแผนการรักษาผู้ป่วยในแต่ละกลุ่มอาการ ส่วนผู้ป่วยและญาติสามารถใช้ประโยชน์ในทางป้องกัน เช่น ควบคุมระดับไขมันในเลือด รับประทานยาต้านการแข็งตัวของเลือดตามคำแนะนำของแพทย์ หรือเภสัชกร เพื่อป้องกันภาวะหัวใจเต้นผิดจังหวะ เป็นต้น

การทดสอบรูปแบบ เพื่อประเมินคุณภาพของความรู้ที่ได้หลังจากการทำเหมืองข้อมูลจากฐานข้อมูลผู้ป่วยเรียบร้อยแล้ว ต้องประเมินว่าความรู้ที่ได้สามารถนำไปจำแนกข้อมูลผู้ป่วยได้ถูกต้องแม่นยำเพียงใด การประเมินคุณภาพผลลัพธ์จะใช้ผู้เชี่ยวชาญ ความสำเร็จของการทำเหมืองข้อมูลอยู่ที่ผลลัพธ์คือ องค์ความรู้ที่ได้จากการประมวลผลเป็นที่น่าสนใจ หรือเป็นสิ่งที่เหนือความคาดหมายของผู้เชี่ยวชาญในสาขาหรือไม่

การนำไปใช้ เป็นการนำเสนอความรู้ที่เป็นผลลัพธ์จากการทำเหมืองข้อมูลให้แก่ผู้ใช้ หรือผู้เชี่ยวชาญเฉพาะด้าน

การวิจัยในอนาคตจะพัฒนาตัวแบบโดยการศึกษาตัวแปรบางตัวที่ไม่ได้รับคัดเลือกให้เป็นตัวแปรพยากรณ์ ด้วยการกำหนดความคลาดเคลื่อนของข้อมูล และมีการทดลองในหลากหลายอัลกอริทึม เพื่อจะได้ตัวแบบที่เหมาะสมมากยิ่งขึ้น

กิตติกรรมประกาศ

ขอขอบคุณ รองศาสตราจารย์ ดร.อนงค์นาฏ ศรีวิหค ที่อนุเคราะห์ให้ความรู้เรื่องการทำเหมืองข้อมูล นายแพทย์สมชาย โตวณะบุตร ที่ให้คำแนะนำและข้อคิดเห็นทางการแพทย์ ศูนย์ข้อมูลข่าวสารระบบประสาท สถาบันประสาทวิทยาอนุเคราะห์ข้อมูลสำหรับการศึกษาในครั้งนี้

เอกสารอ้างอิง

[1] World Health Organization. “Top 10 causes of death,” https://www.who.int/gho/mortality_burden_disease/causes_death/top_10/en/, 2018, [Aug, 2019].

[2] Health Data Center (HDC), กระทรวงสาธารณสุข. “รายงานตามตัวชี้วัดในระดับกระทรวง ปี 2562,” https://hdcservice.moph.go.th/hdc/reports/report_kpi.php?flag_kpi_level=1&flag_kpi_year=2019&source=pformated/format1.php&id=7ac059f4e4e3d08750d2ee23600556af, 2562, [ส.ค. 2562].

[3] ศูนย์ข้อมูลข่าวสารระบบประสาท สถาบันประสาทวิทยา. “ข้อมูลคุณภาพการให้บริการ: โรคหลอดเลือดสมองตีบหรืออุดตัน Registry of Stroke Care Quality,” http://neuronetworks.org/stroke [ส.ค. 2562].

[4] M. Mathuria, “Decision Tree Analysis on J48 Algorithm for Data Mining”, Intrenational Journal of Advanced Research in Computer Science and Software Engineering.

[5] สุรพงค์ เอื้อวัฒนามงคล. “การทำเหมืองข้อมูล Data Mining,” กรุงเทพฯ: สำนักพิมพ์สถาบันบัณฑิตพัฒนบริหารศาสตร์, 2559, หน้า 20, 49.

[6] P. P. Chakrabarti and S. Ghose, “A general best first search algorithm in AND/OR graphs,” Journal of Algorithms, vol. 13, no. 2, pp. 177–187, Jun. 1992.

[7] Y. Peng, Z. Wu, and J. Jiang, “A novel feature selection approach for biomedical data classification,” Journal of Biomedical Informatics, vol. 43, no. 1, pp. 15–23, Feb. 2010.