19th International Computer Science and Engineering Conference (ICSEC)
Chiang Mai, Thailand, 23-26 November, 2015

การพยากรณ์ข้อมูลเกี่ยวกับอันตรายของแผ่นดินไหวในเหมืองถ่านหิน ด้วยตัวแบบเรียนรู้แบบเอ็กซ์ตรีมถ่วงน้ำหนัก

Prediction of Seismic Hazard Data in Coal Mines Using Weighted Extreme Learning Machine

วิรันดา สิงห์ทอง (Wirunda Singthong)[1], ปัญญาพล หอระตะ (Punyaphol Horata)[2], สิรภัทร เชี่ยวชาญวัฒนา (Sirapat Chiewchanwattana)และคำรณ สุนัติ (Khamron Sunat)[4]
[1,2,3,4]ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น
E-mail:bunuch04@gmail.com,punhor1@kku.ac.th, sunkra@kku.ac.th, khamron_sunat@yahoo.com

บทคัดย่อ

เนื่องจากแผ่นดินไหวมักจะก่อให้เกิดผลเสียที่รุนแรง และถ้ายิ่งเป็นการเกิดแผ่นดินไหวในเหมืองถ่านหินแล้วยิ่งจะก่อให้เกิดการสูญเสียมาก ถึงแม้ว่าแผ่นดินไหวนั้นไม่รุนแรง แต่จะทำให้เกิดการถล่มของเหมืองและเกิดการสูญเสียอย่างมาก ทั้งทรัพย์สิน และที่สำคัญคือชีวิตคนงานในเหมือง ในอดีตได้มีการศึกษาการพยากรณ์ในตัวแบบ อาทิ ตัวแบบ q-ModLEM และ MODLEM algorithm แต่อย่างไรก็ตามขั้นตอนวิธีเหล่านั้นไม่ได้แก้ปัญหาความไม่สมดุลของข้อมูล ทำให้ผลของการพยากรณ์ของตัวแบบเกิดความคลาดเคลื่อนได้ ซึ่งวัดได้ด้วยค่าประสิทธิภาพการทำงานที่สมดุล (G-Mean) ดังนั้นในงานวิจัยนี้ได้นำเอา Hybrid Over-sampling and Under-sampling มาใช้ร่วมกับ ตัวแบบเรียนรู้แบบเอ็กซ์ ตรีมถ่วงน้ำหนัก เพื่อแก้ปัญหาการพยากรณ์กับข้อมูลไม่สมดุลของข้อมูลแผ่นดินไหว ข้อมูลกรณีศึกษาที่ใช้ในงานวิจัยนี้ คือข้อมูลระบบตรวจสอบอันตรายของแผ่นดินไหวในเหมืองถ่านหิน (seismic-bumps) จากผลการทดลองพบว่าการนำเอา Hybrid Over-sampling and Under-sampling มาใช้ร่วมกับ ตัวแบบเรียนรู้แบบเอ็กซ์ตรีมถ่วงน้ำหนัก ให้ผลการพยากรณ์วัดด้วยค่า G-Mean เป็น 75.95 ซึ่งสูงกว่าตัวแบบ q-ModLEM และตัวแบบ MODLEM algorithm ที่ให้ค่า G-Mean ที่ 67.93 และ 25.60 ตามลำดับ

คำสำคัญ: ตัวแบบการเรียนรู้แบบเอ็กซ์ตรีมถ่วงน้ำหนัก เหมืองถ่านหิน ข้อมูลที่ไม่สมดุล แผ่นดินไหว ข้อมูลที่ไม่สมดุล

Abstract

Due to seismic phenomena is a cause to hazard effects. Especially in a coal mines, even though a seismic occurs in a low level but it is a course of landslide which it is very dangerous to mine workers and they will be lost. Most of the data is often a problem of information imbalance. The traditional learning methods were proposed to predict of occurrence of seismic. However, the results of the predictions of the models may be much errors when their performance were measured by the G-mean since the imbalance of seismic data is not be addressed. This problem can be mitigated by using the weighted extreme learning machine having its property to handle the imbalance data. Therefore, the weighted extreme learning machine as estimator is combined with hybrid over-sampling and under-sampling techniques to address the imbalance problem of the seismic bumps data set which used a case study in this paper. From experimental results show that the weighted extreme learning machine combined with hybrid over-sampling and under-sampling techniques can give G-Mean 75.95 which its result is better than that of G-mean of q-ModLEM 69.3[7] and MODELM 67.93[8], respectively.

Keyword: Weighted extreme learning machine, Coal mines , imbalance data, seismic, imbalance data

1. บทนำ

การทำเหมืองถ่านหินคือการขุดเจาะหรือเปิดหน้าดินลงไปเพื่อที่จะนำแร่ธาตุที่มนุษย์ต้องการในดินนำมาใช้ โดยปกติแล้วจะมี 2 แบบด้วยกัน คือ การทำเหมืองแบบเหมืองเปิดและการทำเหมืองแบบเหมืองใต้ดิน และจะทำเป็นเหมืองใต้ดินเป็นส่วนใหญ่ เมื่อมีการสั่นไหวของแผ่นดินทำให้เกิดการสั่นสะเทือนของหินมีทั้งเป็นอันตรายและไม่เป็นอันตราย ซึ่งถ้าหากไม่เกิดอันตรายก็จะไม่เกิดการสูญเสีย แต่ถ้าหากเป็นอันตรายจะเกิดการสูญเสียทั้งทรัพย์สินและที่สำคัญคือชีวิตของพนักงานที่ทำงานในเหมืองถ่านหินซึ่งไม่มีใครต้องการให้เกิดการสูญเสียใดๆทั้งนั้น

แต่อย่างไรก็ตาม เราไม่สามารถรู้ได้ว่าการสั่นไหวของหินนั้นจะเกิดอันตรายมากน้อยเพียงใดจึงได้มีการพยากรณ์ข้อมูลการสั่นไหวของหินในเหมืองแร่ว่าจะเกิดอันตรายหรือไม่เกิดอันตรายใช้เทคนิคการเรียนรู้ของเครื่องซึ่งผู้วิจัยได้ทำการเปรียบเทียบการเรียนรู้ของเครื่องทั้ง 3 วิธีดังนี้ คือ ขั้นตอนวิธี MODLEM จะให้ค่าความถูกต้อง (Acc) ที่สูงที่สุด คือ 92.50 แต่ ค่าตัววัดประสิทธิภาพการทำงานที่สมดุล (G-mean) มีค่าต่ำมากคือ 25.60 อีกวิธีคือกฎการสร้างและการลดรูปของข้อมูล (q-ModLEM) [7] ให้ค่าความถูกต้อง ที่ไม่สูงมาก คือ 80.20 แต่ค่า G-mean มีค่าสูงกว่าคือ 67.93 ผู้วิจัยจึงได้เสนอการนำเอา Hybrid Over-sampling and Under-sampling มาใช้ร่วมกับตัวแบบการเรียนรู้แบบเอ็กซ์ตรีมถ่วงน้ำหนัก ให้ค่าความถูกต้อง ที่ไม่สูงมาก คือ 76.48 แต่ G-mean มีค่าสูงที่สุดคือ 75.95

กลุ่มข้อมูลที่ไม่เกิดแผ่นดินไหวแทนด้วยกลุ่มลบ ซึ่งมีข้อมูลจำนวนมาก ในขณะที่กลุ่มข้อมูลที่เกิดแผ่นดินไหวมีปริมาณน้อยมาก ทำให้เกิดปัญหาคลาดเคลื่อนของการฝึกสอน ซึ่งจะทำให้กลุ่มตัวอย่างที่มีปริมาณ

มากทำให้การจัดกลุ่มได้ดีกว่าในกลุ่มตัวอย่างน้อย [2] จึงทำให้เป็นไปได้มากในการเกิดข้อผิดพลาดในการฝึกสอน

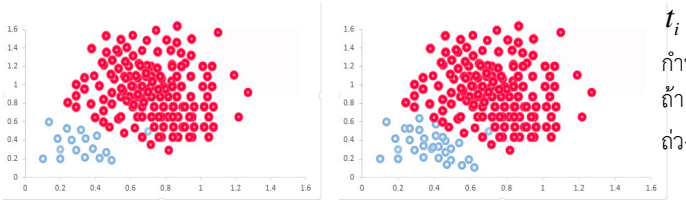
ดังนั้นผู้พัฒนาได้ประยุกต์ใช้วิธีการปรับปริมาณของข้อมูลด้วยวิธี Hybrid Over-sampling and Under-sampling ก่อนส่งเข้าตัวแบบการเรียนรู้แบบเอ็กซ์ตรีมถ่วงน้ำหนักเพื่อนำมาแก้ปัญหาความไม่สมดุลของข้อมูลการเกิดแผ่นดินไหวในเหมืองถ่านหิน เพื่อเพิ่มประสิทธิภาพในการฝึกสอนให้มีความถูกต้องมากขึ้น

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 การจัดการกับข้อมูลที่ไม่สมดุล (Imbalance Data handling)

กรณีที่ต้องการจำแนกข้อมูลไม่สมดุล ซึ่งเป็นจำนวนตัวอย่างที่ปริมาณมีความไม่เท่ากันสูงมากในแต่ละกลุ่ม เมื่อเรานำข้อมูลที่มีจำนวนไม่เท่ากันระหว่างกลุ่ม มาทำการจำแนกประเภท ผลลัพธ์ที่ได้จะทำให้มี การเรียนรู้แต่ข้อมูลกลุ่มมาก และเมื่อทำการจำแนกประเภท ก็จะจำแนกไปในข้อมูลกลุ่มมาก ด้วยเหตุนี้การแก้ปัญหาชุดข้อมูลที่ไม่สมดุลจึงต้องมีการนำขั้นตอนการปรับชุดข้อมูลเข้ามาช่วยในการจำแนกที่แม่นยำ

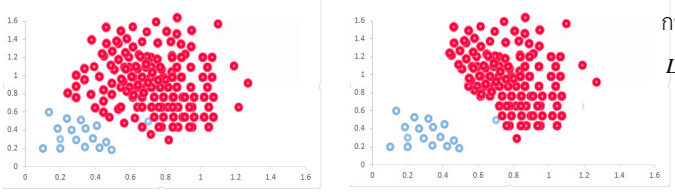
- การสุ่มเพิ่มตัวอย่างกลุ่มน้อย (SMOTE: Synthetic Minority Over-sampling Technique) [4] เป็นการเพิ่มจำนวนตัวอย่าง (Up Sampling) ซึ่งเป็นวิธีการในการเพิ่มจำนวนข้อมูลของกลุ่ม (class) ที่เป็นส่วนน้อย (minority) ให้มีปริมาณข้อมูลใกล้เคียงกับจำนวนตัวอย่างในกลุ่มที่มีจำนวนตัวอย่างมากกว่า (majority) แต่ก็อาจทำให้ข้อมูลที่สร้างขึ้นมาไปอยู่ในบริเวณของข้อมูลที่มีจำนวนมาก แต่อย่างไรก็ตาม ข้อมูลที่มีจำนวนต่างกันมากหลายเท่า ก็ไม่อาจจะทำให้มีความสมดุลของข้อมูลนัก ดังภาพที่ 1



(a) (b)

ภาพที่ 1 (a) ข้อมูลที่ยังไม่ได้ทำ SMOTE (b) ข้อมูลที่ทำการ SMOTE เรียบร้อยแล้ว

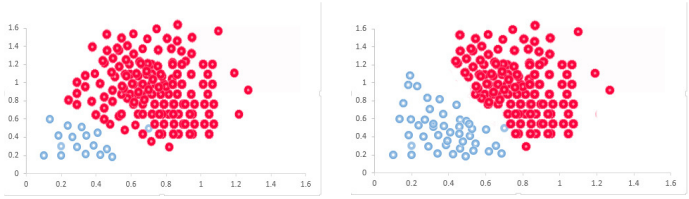
- การสุ่มลดตัวอย่างกลุ่มมาก (under sampling) [9] เป็นการลดปริมาณตัวอย่าง (down) กลุ่มที่มีปริมาณมากให้มีปริมาณที่น้อยลงจนมีปริมาณพอ ๆ กับปริมาณตัวอย่างของกลุ่มที่มีปริมาณน้อย โดยใช้ Tomek Link เป็นการทำงานที่ทำความสะอาดข้อมูลที่อยู่ในบริเวณที่มีการทับซ้อนกันของข้อมูลทั้ง 2 กลุ่ม แต่อย่างไรก็ตามถ้าหากข้อมูลมีความต่างกันมากเกินไปการสุ่มลดตัวอย่างกลุ่มมาก ก็ยังไม่สามารถทำให้ข้อมูลมีความสมดุลกันได้อย่างเหมาะสม ดังภาพที่ 2



(a) (b)

ภาพที่ 2 (a) ข้อมูลที่ยังไม่ได้ทำ Under sampling (b) ข้อมูลที่ทำการ Under sampling เรียบร้อยแล้ว

- Hybrid Over-sampling and Under-sampling [15] เป็นการนำข้อมูลมาทำการสุ่มเพิ่มตัวอย่างกลุ่มน้อย (SMOTE) มีโอกาสเป็นไปได้สูงมากที่อาจจะทำให้เกิดการเพิ่มจำนวนในบริเวณที่ของกลุ่มที่มีจำนวนมาก ดังนั้นจึงนำการลดตัวอย่างของกลุ่มที่มีจำนวนมาก (Tomek Link) เข้ามาทำงานร่วมกันเพื่อช่วยในการทำความสะอาดข้อมูลที่อาจจะเกิดการทับซ้อนกันหรืออยู่ในบริเวณที่ใกล้เคียงกัน และมีการใกล้เคียงกันของจำนวนข้อมูลมากที่สุด เพราะข้อมูลที่นำมา มีความต่างกันมากถึง 1 : 14 ดังภาพที่ 3



(a) (b)

ภาพที่ 3 (a) ข้อมูลที่ยังไม่ได้ทำ Hybrid (b) ข้อมูลที่ทำการ Hybrid เรียบร้อยแล้ว

2.2 ตัวแบบเรียนรู้แบบเอ็กซ์ตรีมถ่วงน้ำหนักในข้อมูลที่เป็นไบนารี

กำหนดให้ตัวอย่างข้อมูลแทนด้วย $[x_i, t_i], i = 1,...,N$ โดยที่ x_i แทนข้อมูลที่มีขนาด m จากตัวอย่างข้อมูลทั้งหมด N ตัวอย่าง ปริมาณ x_i นี้มีปริมาณเป้าหมายคือ t_i ในกรณีของปัญหาการจำแนกกลุ่ม t_i เป็นเวกเตอร์ระบุกลุ่ม ถ้ามีสองกลุ่มจะมีค่ากลุ่ม +1 หรือกลุ่ม -1 กำหนดให้ $\mathbf{W}$ เป็นเมทริกซ์ถ่วงน้ำหนักซึ่งมีขนาด $N \times N$ โดยปกติถ้า x_i มาจาก กลุ่มน้อย (สันนิษฐานว่าจะเป็นกลุ่มบวก) $\mathbf{W}_{ii}$ เป็นตัวถ่วงน้ำหนักที่มีขนาดใหญ่กว่ากลุ่มอื่น ๆ เพื่อถ่วงน้ำหนัก

สมมติฐาน : $\|\mathbf{H}\beta - \mathbf{T}\|^2$ และ $\|\beta\|$ มีค่าน้อย (1)

จะทำให้การทำนายได้ถูกต้องกว่าที่ $\|\beta\|$ มีค่ามาก

$\mathbf{T} = [t_1,...,t_N]$ แทนเมทริกซ์เป้าหมาย มีฟังก์ชันต้นทุนดังนี้

ค่าน้อย : $\mathbf{L}_{P_{ELM}} = \frac{1}{2}\|\beta\|^2 + C\mathbf{W}\frac{1}{2}\sum_{i-1}^{N}\|\xi_i\|^2$

อยู่ในเงื่อนไข : $h(x_i)\beta = t_i^T - \xi_i^T, I = 1,...,N$ (2)

$h(x_i)$ เป็นเวกเตอร์การแมปคุณลักษณะในชั้นที่ซ่อนอยู่ที่เกี่ยวกับ x_i และ β เป็นค่าถ่วงน้ำหนักในชั้นผลลัพธ์ ξ_i แทนข้อผิดพลาดของการฝึกสอน x_i ที่เกิดจากความแตกต่างของค่าเป้าหมาย t_i

ตามที่ทฤษฎีของ Karush-Kuhn-Tucker(KKT) [12] ปัญหาการเพิ่มประสิทธิภาพเทียบเท่า (6) ฟังก์ชันต้นทุนเขียนได้อีกรูป คือ

$$L_{D_{ELM}} = \frac{1}{2}\|\beta\|^2 + CW\ \frac{1}{2}\sum_{i=1}^{N}\|\xi_i\|^2 - \sum_{i=1}^{N}\alpha(\mathbf{h}(\mathbf{x}_i)\beta - t_i + \xi_i) \quad (3)$$

ที่ซึ่งตัวคูณลากรองจ์ α_i เป็นปัจจัยที่คงที่ของตัวอย่าง x_i การหาค่าของ (3) หาโดยการทำอนุพันธ์บางส่วน ของ (3) เทียบกับตัวแปร

(β, ξ_i, α_i) โดยให้อนุพันธ์ของ (3) เท่ากับศูนย์ เพื่อหา ดังสมการ (8)

$$\left.\begin{aligned}
&\frac{\partial L_{D_{ELM}}}{\partial \beta} = 0 \rightarrow \beta = \sum_{i=1}^{N} \alpha_i \mathbf{h}(\mathbf{x}_i)^T = \mathbf{H}^T \alpha \\
&\frac{\partial L_{D_{ELM}}}{\partial \xi_i} = 0 \rightarrow \alpha = CW\xi_i, i = 1,\ldots, N \\
&\frac{\partial L_{D_{ELM}}}{\partial \alpha_i} = 0 \rightarrow \mathbf{h}(\mathbf{x}_i)\beta - t_i + \xi_i = 0, i = 1,\ldots, N
\end{aligned}\right\} \quad (4)$$

การหาคำตอบ เพื่อให้ได้ค่า β ที่เหมาะสม เขียนได้ดังนี้

เมื่อ N มีค่าน้อย : $\beta = \mathbf{H}^T\left(\frac{\mathbf{I}}{C} + \mathbf{WHH}^T\right)^{-1}\mathbf{WT}$

เมื่อ N มีค่ามาก : $\beta = \left(\frac{\mathbf{I}}{C} + \mathbf{H}^T\mathbf{WH}\right)^{-1}\mathbf{H}^T\mathbf{WT}$ (5)

C เป็นค่าคงที่ใด ๆ ของผลการทดลอง เพื่อชดเชยข้อผิดพลาด (regularization)

การพยากรณ์ของตัวอย่างใหม่ $\mathbf{x}$, ฟังก์ชั่นการผลลัพธ์ คือ

$$\begin{aligned}
&f(\mathbf{x}) = sign\, \mathbf{h}(\mathbf{x})\beta : \\
&f(\mathbf{x})_{N\times N} = sign\, h(x)\mathbf{H}^T\left(\frac{\mathbf{I}}{C} + \mathbf{WHH}^T\right)^{-1}\mathbf{WT} \\
&f(\mathbf{x})_{L\times L} = sign\, \mathbf{h}(\mathbf{x})\left(\frac{\mathbf{I}}{C} + \mathbf{H}^T\mathbf{WH}\right)^{-1}\mathbf{H}^T\mathbf{WT}
\end{aligned} \quad (6)$$

เมื่อใช้กับเคอร์เนล [6] $N \times N$

$$\begin{aligned}
f(\mathbf{x})_{\ker nel} &= sign\, \mathbf{h}(\mathbf{x})\mathbf{H}^T\left(\frac{\mathbf{I}}{C} + \mathbf{WHH}^T\right)^{-1}\mathbf{WT} \\
&= sign\begin{bmatrix} K(x, x_1) \\ \vdots \\ K(x, x_N) \end{bmatrix}^T \left(\frac{\mathbf{I}}{C} + \mathbf{W}\Omega_{ELM}\right)^{-1}\mathbf{WT}
\end{aligned} \quad (7)$$

เมื่อ sign แทนฟังก์ชันให้ค่า +1 หรือ -1 และแทน $K(x, x_i)$ แทน เคอร์เนลฟังก์ชัน และ Ω_{ELM} เขียนแทน $\mathbf{HH}^T$

3. ตัวแบบที่นำเสนอ

ในงานวิจัยนี้ได้นำเสนอวิธีแก้ปัญหาการจำแนกสำหรับข้อมูลที่ไม่สมดุล ที่มีจำนวนต่างกันมากถึง 1 ต่อ 14 โดยคิดเป็นถ้าคลาส 1 มี 1 คลาส 0 จะมี 14 ตัว โดยงานวิจัยนี้แบ่งออกแบบสองขั้นตอน คือ การทำให้ข้อมูลสมดุลโดยใช้วิธี Hybrid Over-sampling and Under-sampling ซึ่งข้อมูลที่ได้จะมีความสุมดุลมากที่สุดคือจากลุ่ม ที่เป็น 1 จากเดิมมี 170 ค่า เพิ่มเป็น 433 ค่า คิดเป็น 2.55 เท่าจากของเดิม และจากกลุ่มที่เป็น 0 จากเดิมมี 2414 ลดลงเหลือ 1690 ค่า ข้อมูลมีการลดลง 30 เปอร์เซ็นต์ จากข้อมูลเดิมแต่อย่างไรก็ตาม ข้อมูลที่ได้ทำการ Hybrid แล้วก็ยังมีจำนวนที่ต่างกันมากถึง 1 ต่อ 4 ดังนั้นจึงนำข้อมูลมาฝึกสอนด้วย ตัวแบบเรียนรู้แบบเอ็กซ์ ตรีมถ่วงน้ำหนัก อีกครั้งเพื่อเพิ่มน้ำหนักให้กับข้อมูลอีกรอบ และวัดประสิทธิภาพด้วยตัววัด G-mean สรุปขั้นตอนได้ดังนี้

1. ปรับข้อมูลให้สมดุล ด้วยวิธี Hybrid Over-sampling and Under-sampling
2. จำแนกกลุ่มด้วย ตัวแบบเรียนรู้แบบเอ็กซ์ตรีมถ่วงน้ำหนัก

4. ผลการทดลอง

4.1 การออกแบบการทดลอง

งานวิจัยนี้ใช้ข้อมูลชื่อ Seismic bumps จาก UCI [14] ซึ่งมีข้อมูลทั้งหมด 2584 ข้อมูล มีแอตทริบิวต์ทั้งหมด 18 กลุ่มแบ่งเป็น 2 กลุ่ม คือ กลุ่มที่เกิดอันตราย(class 1) มี 170 ข้อมูล และกลุ่มที่ไม่เกิดอันตราย (class 0) มี 2414 ข้อมูล ข้อมูลที่ได้จาก UCI เป็นข้อมูลชนิด ARFF จึงได้ทำการแปลงข้อมูลให้เป็นชนิดข้อความ และได้มาทำการ Over-sampling , Under sampling และ Hybrid Over-sampling and Under-sampling แล้วปรับข้อมูลให้อยู่ในช่วง 0 ถึง 1 แล้วนำข้อมูลที่ได้ไปทำการครอสวาลิเดชั่นแบ่งเป็น 10 ส่วนเพื่อนำไปใช้ในการจำแนกกลุ่ม

4.2 ตัววัดประสิทธิภาพ

วัดความน่าเชื่อถือ เมื่อสร้างแบบจำลองเสร็จแล้ว จะต้องวัดความน่าเชื่อถือของแบบจำลองที่เราสร้าง โดยพิจารณาจากค่าทางสถิติที่ได้จากการวิเคราะห์และทดสอบ เพื่อใช้ในการคำนวณหาค่าความไว (*Sensitivity*) ความจำเพาะ (*Specificity*) และความถูกต้อง (*Accuracy*) ดังนี้

4.2.1 Confusion Matrix เป็นส่วนแสดงผลให้เห็นผลของการทำนาย โดยที่

- TP คือจำนวนข้อมูลที่ทำนายว่าเกิดแผ่นดินแล้วเกิดแผ่นดินไหวจริง
- FP คือจำนวนที่ทำนายว่าเกิดแผ่นดินไหวแต่ไม่เกิดแผ่นไหว
- TN คือจำนวนที่ทำนายว่าไม่เกิดแผ่นดินไหวแล้วไม่เกิดแผ่นดินไหว
- FN คือจำนวนข้อมูลที่ทำนายว่าไม่เกิดแผ่นดินไหว แต่เกิดแผ่นดินไหว

ตารางที่ 1: Confusion Matrix สำหรับ two-class classification

		True class	
		Positive prediction	Negative prediction
Prediction class	Actual Positive	TP	FN
	Actual Negative	FP	TN

จากตารางที่ 1 สามารถคำนวณค่าต่างๆได้จากสูตร ต่อไปนี้

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (9)$$

$$Specificity = \frac{TN}{TN + FP} \quad (10)$$

Sensitivity, Specificity, และ Geometric mean ตัววัดประสิทธิภาพที่นำมาใช้นี้ เมื่อประสิทธิภาพการทำงานของทั้งสองคลาสที่เกี่ยวข้องและคาดว่าจะสูงพร้อมกัน โดย G-mean [10] ถูกใช้เพื่อวัดประสิทธิภาพของตัวจำแนกข้อมูลที่ไม่สมดุลในหลายงานวิจัย [11] ซึ่งบ่งชี้

ความสมดุลระหว่างประสิทธิภาพการจำแนกของ minority และ majority คลาส กำหนดโดย

$$G-means = \sqrt{Sensitivit\ y \times Specificit\ y} \quad (11)$$

โดยที่ G-mean เป็นตัววัดประสิทธิภาพการทำงานที่สมดุลของขั้นตอนวิธีการเรียนรู้ระหว่างทั้งสองกลุ่มค่า G-mean เข้าใกล้ 1 แสดงว่าตัวแบบทำนายค่า FP และ FN น้อย แสดงว่าตัวแบบดี

4.3 ผลการทดลอง

ตารางที่ 2 :ผลการเปรียบเทียบประสิทธิภาพในการพยากรณ์ข้อมูลอันตรายของแผ่นดินไหวในแหมืองถ่านหิน

Algorithms	balanced data	C	Acc.	BAcc.	G-mean	Acc.1 Specificity	Acc.0 Sensitivity
Weighted ELM	Normal	2^{8}	75.27	70.71	70.34	65.49	75.92
	SMOTE	2^{17}	74.96	74.82	74.77	74.53	75.10
	Under sampling	2^{6}	74.75	71.26	70.98	66.91	75.60
	Hybrid	2^{23}	76.48	75.96	75.95	74.77	77.14
ELM	Normal	-	72.28	65.75	62.09	44.11	87.39
	SMOTE	-	71.47	66.47	63.78	47.76	85.18
	Under sampling	-	71.64	65.155	61.49	43.61	86.70
	Hybrid	-	72.31	67.73	65.41	50.15	85.31
q-ModLEM	-	-	80.20	69.10	67.93	81.90	56.35
MODLEM	-	-	92.50	52.60	25.60	98.58	6.65

ในงานวิจัยนี้ได้ศึกษาเปรียบเทียบกับตัวแบบเรียนรู้แบบเอ็กซ์ตรีมถ่วงน้ำหนัก ข้อมูลแต่ละชุดจะแบ่ง เป็น 4 แบบ คือ แบบปกติ ไม่มีการปรับสมดุล (Normal) แบบปรับด้วยวิธี over sampling (Smote) แบบปรับด้วยวิธี under sampling และ แบบผสม hybrid over and under sampling แล้วส่งให้ตัวจำแนก ซึ่งค่า C คือค่าถ่วงน้ำหนักที่หาตั้งแต่ 2^{-18}ถึง 2^{30} ผลการทดลองแสดงได้ดังตารางที่ 2

จากตารางผลการเปรียบเทียบประสิทธิภาพในการพยากรณ์ข้อมูลอันตรายของแผ่นดินไหวในแหมืองถ่านหิน จะเห็นได้ว่าค่าความถูกต้อง (Acc.) ของตัวแบบเรียนรู้แบบเอ็กซ์ตรีมถ่วงน้ำหนักหลังจากมีการปรับข้อมูลให้สมดุลด้วยวิธี *Hybrid over-sampling and under-sampling* ให้ พบว่าในกรณีการปรับด้วยวิธีผสม มีค่าต่ำ 76.48 ค่าความสมดุลของความถูกต้อง (BAcc.) ของตัวแบบเรียนรู้แบบเอ็กซ์ตรีมถ่วงน้ำหนัก มีค่าสูงที่สุด 75.96 เมื่อเทียบกับค่านี้ของตัวแบบอื่น ตัววัดประสิทธิภาพการทำงานที่สมดุล (G-means) ของตัวแบบเรียนรู้แบบเอ็กซ์ตรีมถ่วงน้ำหนัก มีค่าสูงที่สุด 75.95 เมื่อปรับเทียบกับตัวแบบดั่งเดิม q-ModeELM และ MODLEM

5. สรุปการวิจัย

ในการศึกษาการพยากรณ์การเกิดแผ่นดินไหวในเหมืองถ่านหิน เราใช้ข้อมูล Seismic bumps จาก UCI โดยใช้ตัวแบบการเรียนรู้แบบเอ็กซ์ตรีมถ่วงน้ำหนักเปรียบเทียบกับ 2 ตัวพยากรณ์ คือ q-ModLEM และ MODLEM ผลการทอลองแสดงให้เห็นว่า ตัวแบบการเรียนรู้แบบเอ็กซ์ตรีมถ่วงน้ำหนัก ผสมกับการปรับข้อมูลให้สมดุลด้วยวิธี *Hybrid over sampling and under-sampling* ช่วยเพิ่มประสิทธิภาพการพยากรณ์ในกรณีข้อมูลไม่สมดุล โดยวัดจากค่าประสิทธิภาพการทำงานที่สมดุล (G-mean) สูงกว่าตัวแบบที่นำมาเปรียบเทียบ

เอกสารอ้างอิง

[1] Sikora M., Wrobel L.: Application of rule induction algorithms for analysis of data collected by seismic hazard monitoring systems in coal mines. Archives of Mining Sciences, 55(1), 2010, 91-114

[2] ZONG, Weiwei; HUANG, Guang-Bin; CHEN, Yiqiang. Weighted extreme learning machine for imbalance learning. Neurocomputing, 2013, 101: 229-242

[3] Huang, GB., Zhu, QY.&Siew CK. 2006. Extreame Learning Machine: Theory and applications. Neurocomputing, 70(1-3): 489-501

[4] Kalton, G. (2009). Methods for oversampling rare subpopulations in social surveys. Survey Methodology, 35(2), 125-141

[5] Fayyad, U. M. and K. B.Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In Bajcsy, R. (ed.). Machine Learning. Proc. Of 13th. IJCAI; 1-Sep-1993. Morgan-Kaufmann: San Francisco; 1993. pp. 1022-1027

[6] G.-B. Huang, H. Zhou, X. Ding, R. Zhang, Extreme learning machine for regression and multi-class classification, IEEE Trans. Syst. Man Cybern., 42 (2) (2012) 513–529

[7] Sikora M.: Rule quality measures in creation and reduction of data rule models. Lecture Notes in Artificial Intelligence 4259, 2006, 716-725

[8] Stefanowski J.: On combined classifiers, rule induction and rough sets. Transactions on Rough Sets VI (LNCS 4374) Springer-Verlag, 2007, s. 329 350..

[9] Tomek, I. (1976). A generalization of the k-NN rule. Systems, Man and Cybernetics, IEEE Transactions on, (2), 121-126.

[10] Kubat, M., & Matwin, S. (1997, July). Addressing the curse of imbalanced training sets: one-sided selection. In ICML (Vol. 97, pp. 179-186)..

[11] Ertekin, S., Huang, J., & Giles, C. L. (2007, July). Active learning for class imbalance problem. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 823-824). ACM.

[12] R. Fletcher, Practical Methods of Optimization: Constrained Optimization, vol. 2, Wiley, New York, 1981

[13] T. Fawcett, An introduction to roc analysis, Pattern Recognition Lett. 27 (June) (2006) 861–874.

[14] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[15] Ramentol, Enislay, et al. "SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory." Knowledge and information systems 33.2 (2012): 245-265.