

การจำแนกการมีภาวะเสี่ยงสูงต่อโรคความดันโลหิตสูงด้วยเครื่องจักรการเรียนรู้เอ็กซ์ทรีม

Extreme Learning Machine for Pre-Hypertension Classification

จิราภา โสมชัย สิริภัทร เชี่ยวชาญวัฒนา คำรณ สุนัฒิ และ นิพนธ์พัทธ์ เมืองโคตร

ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น

Email: chirapa_s@kkumail.com, sunkra@kku.ac.th, khamron_sunat@yahoo.com, mnipot@gmail.com

บทคัดย่อ

บทความนี้ได้เสนอวิธีการจำแนกประชากรภาวะเสี่ยงสูงต่อโรคความดันโลหิตสูง (pre-hypertension) ด้วยเครื่องจักรการเรียนรู้เอ็กซ์-ทรีม (Extreme Learning Machine: ELM) ซึ่งเป็นขั้นตอนวิธีที่มีข้อดีในด้านความเร็วในการเรียนรู้ และมีความถูกต้องสูงในการจำแนกประเภท วัตถุประสงค์เพื่อเป็นเครื่องมือช่วยในการวิเคราะห์ข้อมูลทางการแพทย์อีกทางหนึ่ง ในการจำแนกกลุ่มเสี่ยงโรคความดันโลหิตสูง โดยได้ทดลองเปรียบเทียบประสิทธิภาพความถูกต้องกับตัวแบบโครงข่ายประสาทเทียมแบบดั้งเดิมที่นิยมใช้ในปัจจุบัน ได้แก่ โครงข่ายประสาทเทียมแบบมัลติเลเยอร์เพอร์เซ็ปตรอนแบบแพร่ย้อนกลับ (MLP-BP) และโครงข่ายประสาทเทียมแบบมัลติเลเยอร์เพอร์เซ็ปตรอนแบบเลเวนเบิร์ก-มาร์ควอร์ต (MLP-LM) โดยชุดข้อมูลก็นำมาใช้ในการทดลอง ใช้ข้อมูลจากการคัดกรองกลุ่มเสี่ยงของประชากรที่มีอายุตั้งแต่ 15 ปี ขึ้นไป ในเขตพื้นที่ ตำบลเจริญศิลป์ อำเภอเจริญศิลป์ จังหวัดสกลนคร ปีงบประมาณ 2555 ซึ่งมีจำนวนปัจจัยทั้งหมด 12 ปัจจัย ใช้จำนวนข้อมูลทั้งหมด 2,987 ระเบียบ ผลการทดลองพบว่า ค่าความถูกต้องในการจำแนกประชากรที่มีภาวะเสี่ยงสูงต่อโรคความดันโลหิตสูง แบบ ELM ให้ความถูกต้องการจำแนกข้อมูลมากที่สุด (คิดเป็น 90.95 เปอร์เซ็นต์โดยเฉลี่ย)

คำสำคัญ

ความดันโลหิตสูง โครงข่ายประสาทเทียมแบบย้อนกลับ เครื่องจักรการเรียนรู้เอ็กซ์-ทรีม

Abstract

This paper proposes a method to classify the high-risk populations for Pre-hypertension with Extreme Learning Machine (ELM). An important advantage of ELM algorithm is fast learning, and it also provides a higher accuracy to the classification problem. This research aims to study ELM for as a tool to analyze the medical data. We compare the accuracy performance of the classification of Pre-Hypertension with the traditional model such as Multilayer Perceptron-Backpropagation (MLP-BP), and Multilayer Perceptron-Levenberg-Marquardt (MLP-LM). Data sets used in the experiments are the screening risk groups of the population Age 15 years and over and located in Charoensin District, Sakon Nakhon Province, fiscal year 2012 between October and September 2011 and 2012. Those data set have 12 factors, 2,987 samples. The experimental results showed that ELM gave the highest accuracy performance (90.95 percent as average).

Keywords

pre-hypertension, multilayer perceptron-back propagation, extreme learning machine

1. บทนำ

โรคความดันโลหิตสูง (Hypertension) [1] เป็นมหันตภัยเงียบที่มีความรุนแรง ในช่วง 20 ปีที่ผ่านมา ผู้ป่วยจากภาวะแทรกซ้อนที่มีสาเหตุมาจากความดันโลหิตสูงมีจำนวนเพิ่มขึ้นทุกวัน อาทิเช่น โรคภาวะหัวใจวาย ภาวะหัวใจขาดเลือด โรคไตวายเรื้อรัง และโรคหลอดเลือดสมอง ดังนั้นการป้องกันความดันโลหิตสูงสามารถป้องกันอัตราการตายจากโรคหัวใจ และโรคอัมพาตที่เพิ่มขึ้นได้ [2] โรคความดันโลหิตสูงเป็นภัยที่คุกคามชีวิตของทุกคน เนื่องจากไม่มีอาการเตือน โรคความดันโลหิตสูงเป็นปัญหาสำคัญทางด้านสาธารณสุข เป็นกลุ่มโรคที่มีปัจจัยสาเหตุนำสู่โรคที่มีปัจจัยเสี่ยงร่วมและเสียชีวิตก่อนวัยอันควร ซึ่งโรคนี้สามารถป้องกันได้โดยการปรับเปลี่ยนพฤติกรรม เช่น การรับประทานอาหารตามหลักโภชนาการ และมีกิจกรรมทางกายภาพที่เหมาะสม ดังนั้นการจัดกลุ่มผู้ป่วยที่มีโอกาสเกิด

ความเสี่ยงมีความสำคัญในการใช้วางแผนและควบคุมป้องกันโรค และเฝ้าระวังการเกิดอุบัติการณ์ของโรค

โครงข่ายประสาทเทียมเป็นวิธีการที่นิยมนำมาใช้ในการแก้ปัญหาต่างๆ อย่างแพร่หลายในปัจจุบัน อาทิเช่น ปัญหาทางด้านวิศวกรรม ปัญหาทางด้านธุรกิจ หรือปัญหาทางการแพทย์ เป็นต้น สำหรับการแก้ปัญหาทางการแพทย์ Xiuying Dong [3] ได้ประยุกต์ใช้อัลกอริทึมโครงข่ายประสาทเทียม BP neural network ร่วมกับ LM นำมาใช้ในการวิเคราะห์ปัจจัยของโรคความดันโลหิตสูง ปัจจัยหลักที่ก่อให้เกิดโรคความดันโลหิตสูง คือ อายุ พันธุกรรม โรคอ้วน พฤติกรรมการดำรงชีวิต และสภาพแวดล้อม ฯลฯ ผลการทดสอบ พบว่าได้ค่าความผิดพลาดน้อย สามารถนำไปใช้แก้ปัญหาได้เป็นอย่างดี Sumathi [4] ได้นำเทคนิคโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม BP ในการแก้

ปัญหาทางด้านการแพทย์ เพื่อทำนายการเป็นโรคความดันโลหิตสูงเพื่อป้องกันวางแผนการรักษาในระยะแรก และการเสียชีวิตจากโรคนี้ โดยใช้ข้อมูลปัจจัยเสี่ยงสาเหตุของโรคความดันโลหิตสูง ได้แก่ อายุ เพศ พันธุกรรม สภาพทางสังคม โรคอ้วน สูบบุหรี่ ดื่มสุรา การออกกำลังกาย ความเครียด โรคเบาหวาน ผลที่ได้สามารถพยากรณ์ความเสี่ยงโรคความดันโลหิตสูง และช่วยให้แพทย์วินิจฉัยโรคได้ในระยะแรก และวางแผนการรักษาต่อไปได้

แต่อย่างไรก็ตาม โครงข่ายประสาทเทียมแบบเดิมนั้นมีข้อเสีย คือ มีการเรียนรู้ที่ช้า Huang [5, 6] ได้นำเสนอขั้นตอนวิธีโครงข่ายประสาทเทียมแบบใหม่ เรียกว่า เครื่องจักรการเรียนรู้เชิงลึกที่มีโครงสร้างแบบ Single-hidden layer feedforward neural networks (SLFNs) เป็นวิธีที่มีความเร็วสูงโดยอาศัยการกำหนดค่าน้ำหนักและค่าเอนเอียงสำหรับ โหนดชั้นซ่อนจากการสุ่มแทนวิธีการวนปรับค่าน้ำหนักแบบเดิม ในขั้นตอนวิธี ELM ได้ประยุกต์วิธีการคำนวณเมทริกซ์ผกผันแบบมอร์-เพนโรส (Moore-Penrose) มาหาค่าน้ำหนักให้ชั้นผลลัพธ์ ซึ่งถือได้ว่าเป็นวิธีที่มีความเร็วสูงในการสอน และให้ค่าความถูกต้องที่ดี

ในบทความนี้ผู้วิจัยจึงได้เสนอโครงข่ายประสาทเทียมแบบ เครื่องจักรการเรียนรู้เชิงลึกที่มี (ELM) มาประยุกต์ใช้กับปัญหาการจำแนกข้อมูลภาวะเสี่ยงสูงต่อโรคความดันโลหิตสูง นอกจากนี้ได้เปรียบเทียบประสิทธิภาพความถูกต้องกับโครงข่ายประสาทเทียมแบบเดิม ประกอบด้วย โครงข่ายประสาทเทียมแบบ MLP-BP และ MLP-LM

ส่วนที่เหลือในบทความนี้ได้แบ่งเนื้อหาออกเป็นสี่ส่วนดังนี้ ส่วนที่สอง ทฤษฎีและงานวิจัยที่เกี่ยวข้อง ส่วนที่สาม วิธีการดำเนินการวิจัย ส่วนที่สี่ อธิบายผลการทดลอง การสรุปผลและข้อเสนอแนะอยู่ในส่วนที่ห้า

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 โรคความดันโลหิตสูง

ความดันโลหิต หมายถึง แรงดันของกระแสเลือดที่กระทบต่อหลอดเลือดแดง อันเกิดจากการสูบฉีดของหัวใจ ความดันเลือดประกอบด้วยสองค่า ได้แก่ 1) ความดันช่วงบน หรือ ความดันซิสโตลิก (systolic blood pressure) คือ แรงดันเลือดขณะที่หัวใจบีบตัวซึ่งจะสูงตามอายุ ความดันช่วงบนในคนคนเดียวกันอาจมีค่าแตกต่างกันบ้างเล็กน้อยตามท่าของร่างกาย การเปลี่ยนแปลงทางอารมณ์ และปริมาณของการออกกำลังกาย 2) ความดันช่วงล่าง หรือความดันไดแอสโตลิก (diastolic blood pressure) คือ แรงดันเลือดขณะที่หัวใจคลายตัว ความดันเลือดปกติขณะพักอยู่ในช่วง 100 - 140 มิลลิเมตรปรอท ในช่วงหัวใจบีบ และ 60 - 90 มิลลิเมตรปรอท ในช่วงหัวใจคลาย ดังนั้น ผู้ที่มีภาวะความดันโลหิตสูงจึงหมายถึง ผู้ที่มีความดันเลือดเท่ากับหรือสูงกว่า 140/90 มิลลิเมตรปรอท ความดันโลหิตสูง แบ่งออกได้เป็นความดันโลหิตสูงปฐมภูมิ มีความดันโลหิตสูงโดยไม่มีสาเหตุชัดเจน ผู้ป่วยส่วนใหญ่ประมาณร้อยละ 90 - 95 จัดเป็นความดันโลหิตสูงปฐมภูมิที่เหลืออีกร้อยละ 5 - 10 เป็นความดันโลหิตสูงแบบทุติยภูมิ มักจะมีสาเหตุจากภาวะอื่นที่มีผลต่อไต หลอดเลือดแดง หัวใจ หรือระบบต่อมไร้ท่อ [1]

ความดันโลหิตสูงเป็นปัจจัยเสี่ยงสำคัญของโรคหลอดเลือดสมอง กล้ามเนื้อหัวใจตายเหตุขาดเลือด หัวใจวาย หลอดเลือดโป่งพอง (เช่น หลอดเลือดแดงใหญ่ เอออร์ตาโป่งพอง) โรคของหลอดเลือดส่วนปลาย ซึ่งเป็นสาเหตุของโรคไตวายเรื้อรัง ความดันโลหิตที่สูงในระดับปานกลางมีความสัมพันธ์กับอายุขัยที่สั้นลง การปรับเปลี่ยนวิถีชีวิตและพฤติกรรม การกินอาหาร สามารถช่วยลดความดันเลือด และลดความ

เสี่ยงจากภาวะแทรกซ้อนต่างๆ ได้ แต่สำหรับผู้ป่วยที่รักษาด้วยการปรับเปลี่ยนวิถีชีวิตแล้วไม่ได้ผลหรือไม่เพียงพอจำเป็นต้องรักษาด้วยยา [7]

2.2 โครงข่ายประสาทเทียมเพอร์เซปตรอนหลายชั้นแบบแพร่ย้อนกลับ (MLP-Back propagation)

การเรียนรู้แบบแพร่ย้อนกลับ (Back propagation Learning Algorithm) ใช้ฟังก์ชันกระตุ้นแบบไม่เชิงเส้น เป็นการส่งผ่านย้อนกลับค่าน้ำหนักการเชื่อมต่อจะถูกเปลี่ยนให้สอดคล้องกับกฎการแก้ข้อผิดพลาด (error correction) คือ ผลต่างของผลตอบที่แท้จริง (actual response) กับผลตอบเป้าหมาย (target response) เกิดเป็นสัญญาณผิดพลาด (error signal) โดยสัญญาณผิดพลาดนี้จะถูกส่งย้อนกลับเข้าสู่โครงข่ายประสาทเทียมในทิศทางตรงกันข้ามกับการเชื่อมต่อ ค่าน้ำหนักจะถูกปรับจนกระทั่งได้คำตอบที่แท้จริงเข้าใกล้ผลตอบเป้าหมายโดยใช้ค่าเฉลี่ยของค่าความผิดพลาดยกกำลังสองน้อยที่สุด Least Mean Square Error ซึ่งใช้กฎเดลต้าในการปรับค่าน้ำหนักขณะฝึกสอนข่ายงาน ซึ่งได้มีการกำหนด สมการวัตถุประสงค์ $E(t)$ ดังนี้

$$E(t) = \sum_{k=1}^K \sum_{i=1}^{N(L)} (e_i(k))^2 = \sum_{k=1}^K \sum_{i=1}^{N(L)} (d_i(k) - z_i(k))^2 \quad (1)$$

เมื่อ K แทน จำนวนฝึกสอนต่อรอบการทำงาน (epoch)

$N(L)$ แทน จำนวนโหนดในชั้น L และในชั้น L แทนชั้นของผลลัพธ์

d_i แทน ค่าข้อมูลจริง (Target Value) ของโหนดที่ i

z_i แทน ค่าเอาต์พุต (Output) ของโหนดที่ i

$E(t)$ แทน ค่าผลรวมความแตกต่างระหว่างค่าเป้าหมายและ

ค่าข้อมูลจริงโดยเป็นผลรวมของโหนดในชั้นเอาต์พุต และทุกข้อมูลฝึกสอนทุกตัว ณ การทำงาน t [8]

2.3 Levenberg-Marquadt BP

เป็นโครงข่ายแบบแพร่ย้อนกลับที่ได้มีการปรับปรุงวิธีการฝึกเครือข่ายเพื่อให้มีความเร็วในการฝึกของแต่ละรอบเร็วขึ้น วิธีแบบ Levenberg-Marquadt เป็นวิธีแบบ Gauss Newton ซึ่งใช้อินเวอร์สเมทริกซ์ของ Hessian ซึ่งเป็นเมทริกซ์จัตุรัสของความผิดพลาด เพื่อทำการปรับน้ำหนักและค่าไบแอสโครงข่ายงานประสาทเทียมเพื่อปรับผลลัพธ์ของแบบจำลองให้ใกล้เคียงกับค่าความจริงมากที่สุด โดยมีสมการการปรับค่าน้ำหนักดังนี้ [9]

$$\nabla x = [J^T(x)J(x) + \mu I]^{-1} J^T(x)e(x) \quad (2)$$

โดยที่ $J(x)$ คือ จาคอบีเยนเมทริกซ์ของอนุพันธ์ค่าผิดพลาดเทียบกับน้ำหนัก

$e(x)$ คือ ค่าเมทริกซ์ของค่าผิดพลาดระหว่างเอาต์พุตจริงกับและเอาต์พุตที่ได้จากโครงข่ายประสาทเทียม

μ คือ ค่าอัตราการเรียนรู้

2.4 ขั้นตอนวิธี ELM

ขั้นตอนวิธีโครงข่ายประสาทเทียมแบบ ELM นำเสนอโดย Huang [4, 5] เป็นขั้นตอนวิธีแบบ SLFNs (Single-Hidden Layer Feed Forward Neural Network) ซึ่งเป็นข่ายงานที่ประกอบด้วยชั้นนำเข้า (Input Layer) ชั้นซ่อน (Hidden Layer) และชั้นผลลัพธ์ (Output Layer) สมมติให้ N เป็นชุดข้อมูล N รายการที่ประกอบด้วยคู่ลำดับ (X_i, t_i) , $i = 1, 2, \dots, N$ โดยที่ $X_i = [x_{i1}, x_{i2}, \dots, x_{in}]^T \in \mathbb{R}^n$ และผลลัพธ์ที่ต้องการ $t_i = [t_{i1}, t_{i2}, \dots, t_{im}]^T \in \mathbb{R}^m$ สำหรับ

ข่ายงานที่มีชั้นซ่อนเป็น $\hat{N}$ โหนด โดยชั้นตอนวิธี ELM จะสุ่มค่าถ่วงน้ำหนัก และค่าเอนเอียงที่เชื่อมระหว่างชั้นนำเข้าและชั้นซ่อน ในชั้นนี้จะได้ผลลัพธ์ที่ผ่านฟังก์ชันกระตุ้นชนิดไม่เชิงเส้น ซึ่งค่าที่ได้จะถูกใช้เป็นข้อมูลนำเข้าของชั้นผลลัพธ์ ซึ่งที่ชั้นผลลัพธ์จะไม่มีฟังก์ชันกระตุ้น จะถูกส่งออกเป็นค่าผลลัพธ์โดยตรง โดยใช้หลักการแปลงจากระบบสมการไม่เชิงเส้นเป็นสมการเชิงเส้น ดังสมการต่อไปนี้

$$H\beta = T \quad (3)$$

โดยที่ $H = \{h_{ij}\}$, $i=1, \dots, N$ และ $j = 1, \dots, \hat{N}$ เป็นเมทริกซ์ผลลัพธ์ โดยผลลัพธ์แต่ละตัวได้มาจากการผ่านฟังก์ชันกระตุ้น $h_{ij} = g(w_j x_i + b_j)$ ซึ่งเป็นฟังก์ชันชนิดไม่เชิงเส้น ผลลัพธ์จะได้มาจากชั้นซ่อนโหนดที่ j ของข้อมูลนำเข้าที่ i แทนด้วย X_i โดย H จะไม่เป็นเมทริกซ์จัตุรัส (nonsquare matrix) สำหรับ $w_j = [w_{j1}, w_{j2}, \dots, w_{jm}]^T$ และ $\beta = [\beta_1, \beta_2, \dots, \beta_{jm}]^T$ b_j เป็นค่าถ่วงน้ำหนักของโหนด j ตามลำดับ

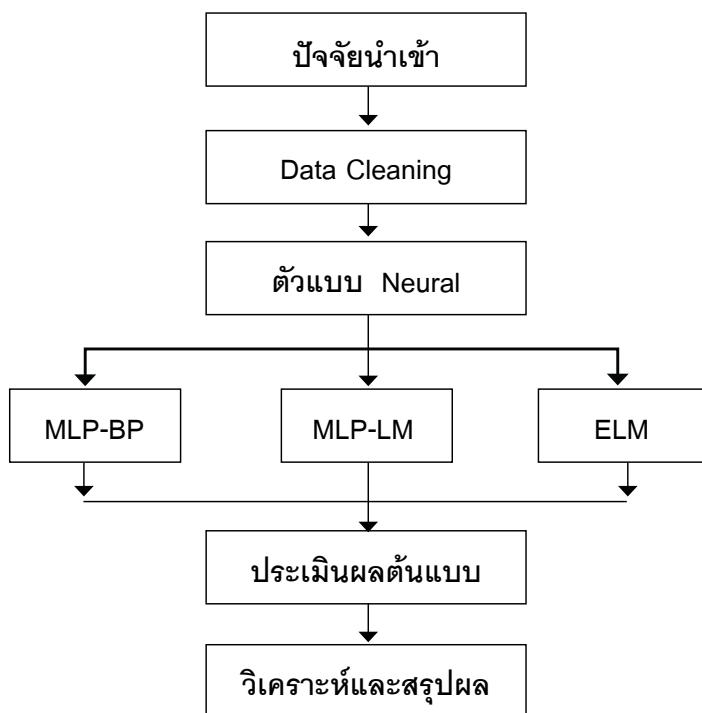
$\beta = [\beta_1, \beta_2, \dots, \beta_{jm}]^T$ $j = 1, \dots, \hat{N}$ เป็นเมทริกซ์ค่าถ่วงน้ำหนักของชั้นผลลัพธ์เป็นเวกเตอร์ของค่าถ่วงน้ำหนักที่อยู่ระหว่างเซลล์ประสาทที่ j ของชั้นซ่อนและชั้นผลลัพธ์ $T = [t_1, t_2, \dots, t_N]^T$ เป็นเมทริกซ์ผลลัพธ์นำ T ไปใช้หาค่าน้ำหนักที่ชั้นผลลัพธ์ได้ด้วยวิธีกำลังสองน้อยสุด สามารถเขียนได้ดังนี้

$$\hat{\beta} = H^+ T \quad (4)$$

โดยที่ H^+ แทนเมทริกซ์ผกผันแบบมอร์-เพนโรส ซึ่ง $\hat{\beta}$ เป็นคำตอบประมาณของสมการที่มีค่าอนอร์มต่ำสุด (Minimum norm least-squares) ดังสมการต่อไปนี้

$$\|H\hat{\beta} - T\| = \|HH^+ T - T\| = \min_{\beta} \|H\beta - T\| \quad (5)$$

ผู้ที่สนใจสามารถศึกษารายละเอียดของขั้นตอนวิธี ELM เพิ่มเติมได้จาก [4, 5]



ภาพที่ 1 ขั้นตอนการดำเนินการวิจัย

3. วิธีการดำเนินการวิจัย

บทความนี้ได้นำเสนอวิธีการจำแนกข้อมูลประชากรที่มีภาวะเสี่ยงสูงต่อโรคความดันโลหิตสูงด้วยโครงข่ายประสาทเทียมแบบเครื่องจักรการเรียนรู้เอ็กซ์-ทรีม (ELM) และเพื่อแสดงให้เห็นถึงความสามารถของขั้นตอนวิธี ELM ผู้วิจัยได้เปรียบเทียบประสิทธิภาพของตัวแบบกับโครงข่ายประสาทเทียมแบบเดิม ได้แก่ MLP-BP และ MLP-LM ซึ่งมีวิธีการดำเนินการวิจัยดังภาพที่ 1

3.1 การวิเคราะห์ข้อมูล

สำหรับข้อมูลที่ใช้ในบทความนี้ ผู้วิจัยได้รวบรวมข้อมูลจากผลการคัดกรองความเสี่ยงประชากรที่มีอายุตั้งแต่ 15 ปีขึ้นไป ในเขตพื้นที่ ตำบลเจริญศิลป์ อ.เจริญศิลป์ จ.สกลนคร ปีงบประมาณ 2555 จำนวน

2,987 คน ซึ่งผู้วิจัยได้ทำการคัดเลือกเฉพาะปัจจัยที่มีผลต่อการเกิดโรคความดันโลหิตสูงที่จะส่งผลกระทบต่อกลุ่มความเสี่ยงในระดับต่างๆ ของโรคนี้ โดยศึกษาข้อมูลปัจจัยที่ก่อให้เกิดภาวะโรคความดันโลหิตสูงเฉพาะปัจจัยด้านกรรมพันธุ์ ด้านสภาพร่างกาย และด้านการปฏิบัติตน ซึ่งประกอบด้วยปัจจัยจากงานวิจัยของสมาคมความดันโลหิตสูงแห่งประเทศไทย [10] และสถาบัน Framingham Heart Study ในโครงการหัวใจ ปอด และหลอดเลือดแห่งชาติแห่งมหาวิทยาลัยบอสตัน [11] ที่ได้ศึกษาปัจจัยเสี่ยงการเป็นโรคความดันโลหิตสูงดังนี้ อายุ เพศ ค่าความดันโลหิต ค่าดัชนีมวลกาย เส้นรอบเอว มีประวัติพ่อแม่พี่น้องญาติสายตรงเป็นโรคความดันโลหิตสูง โรคเบาหวาน การสูบบุหรี่ การดื่มแอลกอฮอล์ การกินเค็ม และการออกกำลังกาย ดังแสดงในตารางที่ 1

ตารางที่ 1 ปัจจัยเบื้องต้นและรูปแบบการเก็บรวบรวมข้อมูล

ที่	ข้อมูล	รายละเอียดข้อมูล
1	อายุ (Age)	ค่าตัวเลขอายุ
2	เพศ (Sex)	เพศชาย = 1 (1) เพศหญิง = 2 (0)
3	ค่าความดันขณะหัวใจบีบตัว (BPS)	ค่าตัวเลขความดันโลหิตขณะหัวใจบีบตัว
4	ค่าความดันขณะหัวใจคลายตัว (BPD)	ค่าตัวเลขความดันขณะหัวใจคลายตัว
5	ดัชนีมวลกาย (BMI)	ค่าตัวเลขดัชนีมวลกาย
6	เส้นรอบเอว (Waist)	ค่าตัวเลขความยาวเส้นรอบเอว
7	ประวัติครอบครัวที่เป็นโรคความดันโลหิตสูง (HX_HT)	มี = Y (1) ไม่มี = N (0)
8	โรคเบาหวาน (DM)	มี = Y (1) ไม่มี = N (0)
9	การสูบบุหรี่ (Smoking)	สูบ = Y (1) ไม่สูบ = N (0)
10	ดื่มสุรา (Drinking)	ดื่ม = Y (1) ไม่ดื่ม = N (0)
11	กินเค็ม (Salt)	กินเค็ม = Y (1) ไม่กินเค็ม = N (0)
12	การออกกำลังกาย (Exercise)	1 = ออกกำลังกายทุกวัน ครั้งละ 30 นาที (4) 2 = ออกกำลังกาย > สัปดาห์ละ 3 ครั้ง ครั้งละ 30 นาที (3) 3 = ออกกำลังกายสัปดาห์ละ 3 ครั้ง ครั้ง ละ 30 นาที (2) 4 = ออกกำลังกาย < สัปดาห์ละ 3 ครั้ง (1) 5 = ไม่ออกกำลังกายเลย (0)

ค่าตัวเลขในวงเล็บแสดงการแทนค่าข้อมูลเชิงลักษณะ (Categorical data) ด้วยตัวเลข

เมื่อพิจารณาข้อมูลที่ได้รวบรวมมาแล้ว พบว่าข้อมูลมีความยุ่งยากในการจำแนกกลุ่ม เนื่องจากลักษณะของข้อมูลเป็นข้อมูลแบบไม่สมดุล (Imbalanced data sets) และเป็นข้อมูลแบบหลายกลุ่ม (Multiclass data sets) รายละเอียดแสดงในตารางที่ 2 ดังนั้นเพื่อให้เกิดความเท่าเทียมกันในการแบ่งข้อมูลสำหรับการใช้ในการสอนและทดสอบ ข้อมูลจะถูกแยกออกไปทำการแบ่งชุดข้อมูลสอนและทดสอบทีละกลุ่ม ในแต่ละกลุ่มจะถูกแบ่งโดยใช้วิธีการเลือกสุ่มแบบความเที่ยงตรง K กลุ่ม (K-Fold Cross Validation) ซึ่งในบทความนี้ทำการทดลองโดยกำหนดการเลือกสุ่ม K เป็น 10 Fold รายละเอียดแสดงในตารางที่ 3

ตารางที่ 2 จำนวนข้อมูลของแต่ละกลุ่มที่ใช้ในการวิจัย

กลุ่มตัวอย่าง	จำนวน (คน)
1. กลุ่มปกติ (ไม่พบความเสี่ยง)	329
2. กลุ่มปกติ (มีปัจจัยเสี่ยง)	2,388
3. กลุ่มเสี่ยงสูง	253
4. สงสัยรายใหม่	17
รวม	2,987

จากตารางที่ 2 ข้อมูลจากกลุ่มตัวอย่างทั้งสิ้น 2,987 คน พบว่าเป็นกลุ่มปกติ (ไม่พบความเสี่ยง) ร้อยละ 11.01 กลุ่มปกติ (มีปัจจัยเสี่ยง) คิดเป็นร้อยละ 79.95 กลุ่มเสี่ยงสูงร้อยละ 8.47 และกลุ่มสงสัยรายใหม่ ร้อยละ 0.75 ตามลำดับ

3.2 การเตรียมข้อมูล

3.2.1 การเตรียมข้อมูล เนื่องจากข้อมูลในงานวิจัยนี้มีหลายรูปแบบทั้งเป็นข้อมูลเชิงจำนวน (Numerical data) และ ข้อมูลเชิงลักษณะ (Categorical data) ซึ่งต้องนำชุดข้อมูลมาแปลงเป็นค่าตัวเลข แต่ละตัวแปรของชุดข้อมูลนำเข้าดังนี้ ข้อมูลรูปแบบ Numerical ได้แก่ Age, BPS, BPD, BMI และ Waist แทนด้วยตัวเลขจริงที่ได้จากการเก็บ

ข้อมูล ข้อมูลรูปแบบ Categorical ได้แก่ Sex, HX_HT, DM, Smoking, Drinking, Salt และ Exercise แทนข้อมูลด้วยค่าตัวเลขแสดงไว้ในวงเล็บ ดังตารางที่ 1

3.2.2 การทำนอร์มอลไลซ์ (Normalization) ทำการปรับให้อยู่ในรูปแบบการแจกแจงปกติในช่วง $[-1, 1]$ เพื่อให้ข้อมูลสามารถนำไปใช้ประมวลผลได้ โดยใช้วิธีการทำนอร์มอลไลซ์ข้อมูลชนิด Min-Max Normalization ดังนี้

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \min_A) + \text{new_max}_A \quad (6)$$

เมื่อ v คือคุณลักษณะเดิม และ v' คือค่าคุณลักษณะใหม่ $\max_A, \min_A$ คือ ค่าต่ำสุดและสูงสุดของคุณลักษณะข้อมูลตามลำดับ และ $\text{new_max}_A, \text{new_min}_A$ คือ ค่าต่ำสุด และสูงสุดใหม่ของคุณลักษณะข้อมูล ตามลำดับ

3.3 การวัดประสิทธิภาพ

ในการวัดประสิทธิภาพเปรียบเทียบระหว่างขั้นตอนวิธี ELM กับโครงข่ายประสาทเทียมแบบอื่นๆ จะใช้ค่าดังต่อไปนี้

1. ค่าเฉลี่ยความถูกต้อง (Average accuracy) ซึ่งจะบอกถึงความสามารถในการจำแนกข้อมูลของตัวจำแนกข้อมูล ค่าเฉลี่ยของความถูกต้องสูง แสดงว่าตัวจำแนกข้อมูลสามารถจำแนกข้อมูลได้ดี ดังสมการต่อไปนี้

$$\text{Average_accuracy} = \frac{1}{N} \sum_{i=1}^N \left[\frac{1}{L} \sum_{j=1}^L \text{Accuracy}_{ij} \right] \quad (7)$$

$$\text{Accuracy} = \left[\frac{\text{TotalValue} - \text{MissclassificationValue}}{\text{TotalValue}} \right] \times 100 \quad (8)$$

โดยที่ N เป็นจำนวนโหนดชั้นซ่อน และ L คือจำนวนรอบในการทดลอง ในการทดลองนี้กำหนดรอบเป็น 30 รอบ

2. ค่าเฉลี่ยของส่วนเบี่ยงเบนมาตรฐาน (Average SD) ซึ่งจะบอกถึงการกระจายตัวของกลุ่มข้อมูล หากค่าเฉลี่ยส่วนเบี่ยงเบนมาตรฐานน้อย แสดงว่าข้อมูลนั้นมีความน่าเชื่อถือสูง ดังสมการต่อไปนี้

$$Mean_{Accuracy} = \frac{1}{N} \sum_{j=1}^N Accuracy_j \quad (9)$$

$$SD = \left[\frac{1}{N} \sum_{j=1}^N (Accuracy_j - Mean_{Accuracy})^2 \right]^{\frac{1}{2}} \quad (10)$$

$$Average_SD = \frac{1}{L} \sum_{i=1}^L SD_i \quad (11)$$

โดยที่ N เป็นจำนวนโหนดชั้นซ่อน และ L คือจำนวนรอบในการทดลอง

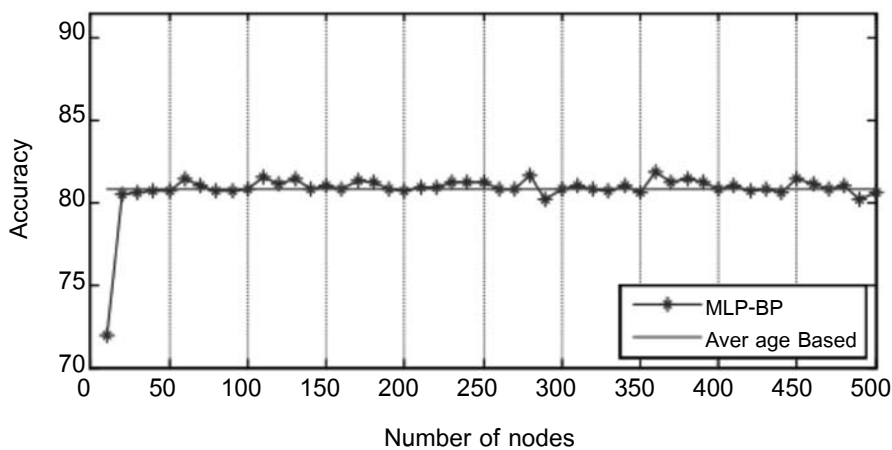
4. ผลการทดลอง

บทความนี้ได้ทดลองจำแนกข้อมูลประชากรที่มีภาวะเสี่ยงสูงต่อโรคความดันโลหิตสูง ด้วยโครงข่ายประสาทเทียมแบบเครื่องจักรการเรียนรู้

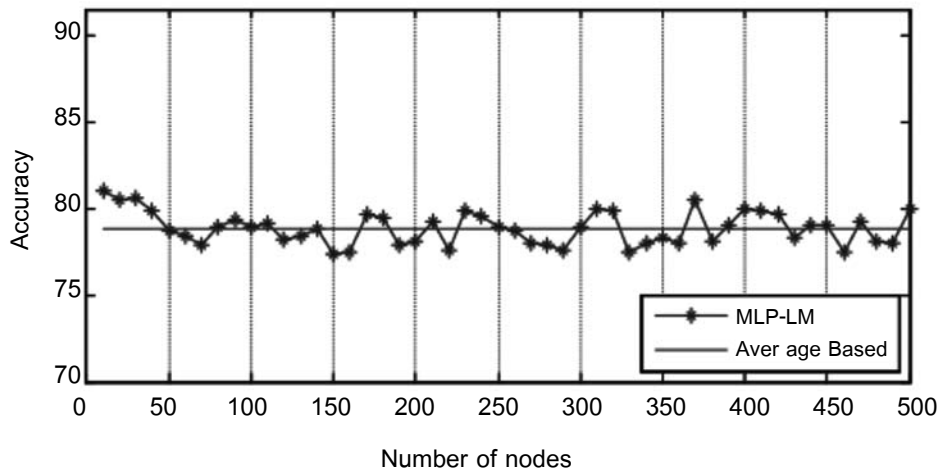
เอ็กซ์ทรีม (ELM) โดยทดลองในสภาพแวดล้อมระบบปฏิบัติการ Windows 7, หน่วยประมวลผล Intel® Core(TM) i5-3470, 3.20GHz แรม 8 GB โปรแกรม MATLAB R2012b

4.1 จำนวนโหนดชั้นซ่อนที่เหมาะสมของโครงข่ายประสาทเทียมแบบ MLP-BP, MLP-LM และ ELM

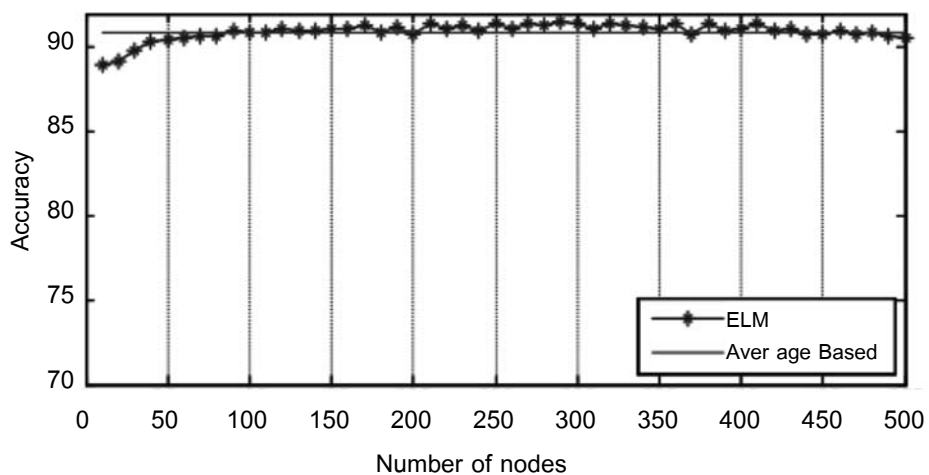
ในการหาจำนวนโหนดชั้นซ่อนที่เหมาะสมของขั้นตอนวิธี ELM ทดลองโดยกำหนดจำนวนโหนดเริ่มต้นที่ 10 โหนด เพิ่มจำนวนโหนดชั้นที่ละ 10 โหนดต่อรอบ จนถึง 500 โหนด จากนั้นหาจำนวนโหนดของชั้นซ่อนที่เหมาะสม โดยหาค่าเฉลี่ยความถูกต้องของจำนวน 10 โหนด ถึง 500 โหนด มาสร้างกราฟเปรียบเทียบค่าความถูกต้องแต่ละตัวของจำนวน 10 โหนด ถึง 500 โหนด แสดงเป็นกราฟได้ดังภาพที่ 2, 3 และ 4



ภาพที่ 2 จำนวนโหนดชั้นซ่อนที่เหมาะสมของขั้นตอนวิธี MLP-BP



ภาพที่ 3 จำนวนโหนดชั้นซ่อนที่เหมาะสมของขั้นตอนวิธี MLP-LM



ภาพที่ 4 จำนวนโหนดชั้นซ่อนที่เหมาะสมของขั้นตอนวิธี ELM

จากการทดลองพบว่าจำนวนโหนดของชั้นซ่อนที่เริ่มให้ค่าความถูกต้องมากกว่าระดับค่าเฉลี่ยของ MLP-BP ที่จำนวนโหนดที่ 60 โหนด และ MLP-LM จำนวนโหนดที่ 10 โหนด ตามลำดับ โดยมีเงื่อนไขการหยุดจำนวนรอบสูงสุดในการฝึกสอน

(epochs) เป็น 250 รอบ และค่าความผิดพลาดเป้าหมาย (goal) เท่ากับ 0.001

จากการทดลองขั้นตอน ELM พบว่าจำนวนโหนดของชั้นซ่อนที่เริ่มให้ค่าความถูกต้องมากกว่าระดับค่าเฉลี่ย คือ จำนวนโหนดที่ 90 โหนด

4.2 การเปรียบเทียบประสิทธิภาพการจำแนก

จากการทดลองเปรียบเทียบประสิทธิภาพการจำแนกการมีภาวะเสี่ยงสูงต่อความดันโลหิตสูงระหว่างขั้นตอนวิธีแบบ ELM โดยเปรียบเทียบประสิทธิภาพกับความถูกต้องโครงข่ายประสาทเทียมแบบ MLP-BP และ MLP-LM ใช้จำนวนโหนดในชั้นซ่อน 90 โหนด สำหรับขั้นตอนวิธีแบบ ELM และสำหรับ MLP-BP และ MLP-LM ใช้จำนวนโหนดในชั้นซ่อน 60 โหนด และ 10 โหนด ตามลำดับ ซึ่งเป็นจำนวนโหนดที่เหมาะสม หาได้จากการทดลองในหัวข้อที่ 4.1 หาค่าเฉลี่ยความถูกต้อง ค่าเฉลี่ยส่วนเบี่ยงเบนมาตรฐาน และเวลาเฉลี่ย ทั้งฝึกสอนและทดสอบ ทดลอง จำนวน 30 รอบ ได้ผลการทดลองดังตารางที่ 4 พบว่าขั้นตอนวิธี ELM ให้ความถูกต้องเป็น 90.95% ค่าเฉลี่ยส่วนเบี่ยงเบนมาตรฐาน 0.18 และเวลาเฉลี่ย 0.00125 วินาที (น้อยกว่า 0.01) ซึ่งเมื่อพิจารณาจากค่าเฉลี่ยความถูกต้อง ค่าเฉลี่ยส่วนเบี่ยงเบนมาตรฐาน และเวลาเฉลี่ยเทียบกับโครงข่ายประสาทเทียมแบบ MLP-BP และ MLP-LM พบว่าขั้นตอนวิธี ELM ให้ผลต่างค่าเฉลี่ยความถูกต้องการทดสอบดีกว่าวิธี MLP-BP และ MLP-LM อยู่ 8.64% และ 10.11% ให้ผลต่างค่าเฉลี่ยส่วนเบี่ยงเบนมาตรฐานการทดสอบอยู่ 0.11 และ 0.54 ตามลำดับ และพิจารณาจากผลต่างเวลาเฉลี่ยพบว่าขั้นตอนวิธี ELM ให้ผลต่างเวลาเฉลี่ยเร็วกว่าวิธี MLP-BP และ MLP-LM อยู่ 0.04 และ 0.02 วินาที ตามลำดับ แสดงให้เห็นว่าขั้นตอนวิธี ELM ให้ประสิทธิภาพการจำแนกดีกว่าโครงข่ายประสาทเทียมแบบอื่น ทั้งในด้านความถูกต้อง ด้านการกระจายตัวของความถูกต้องพิจารณาจากค่าเฉลี่ยส่วนเบี่ยงเบนมาตรฐาน และด้านความเร็ว พิจารณาจากเวลาเฉลี่ยที่น้อยที่สุด เมื่อเทียบกับโครงข่ายประสาทเทียมแบบอื่น

5. สรุปผลและข้อเสนอแนะ

การแก้ปัญหาการจำแนกประชากรที่มีภาวะเสี่ยงสูงต่อโรคความดันโลหิตสูงสามารถใช้เป็นระบบวิเคราะห์ข้อมูลทางด้านสุขภาพเชิงรุกให้กับหน่วยงานที่มหาวิทยาลัยใช้ข้อมูลกลุ่มผู้มีความเสี่ยงเพื่อจัดโปรแกรมการปรับพฤติกรรม ดูแลคนไข้แต่ละประเภทกลุ่มเสี่ยงได้อย่างมีประสิทธิภาพ และเฝ้าระวังภาวะความเสี่ยงของโรคความดันโลหิตสูงที่จะเกิดขึ้นให้เหมาะสมกับความเสี่ยงของแต่ละกลุ่มและตรงเป้าหมาย จากผลการศึกษาพบว่าขั้นตอนวิธี ELM มีความสามารถในการจำแนกกลุ่มประชากรที่มีภาวะเสี่ยงต่อโรคความดันโลหิตสูง ได้ดีกว่าขั้นตอนวิธีดั้งเดิม คือ MLP-BP และ MLP-LM เมื่อเปรียบเทียบกันทั้งในด้านเวลา และความถูกต้องในการแก้ปัญหา แต่อย่างไรก็ตาม จากการศึกษานี้ก็ยังมีประเด็นปัญหาอื่นๆ เพิ่มเติม เช่น ปัญหาการจำแนกกลุ่มเนื่องจากลักษณะของข้อมูลเป็นข้อมูลแบบไม่สมดุล (Imbalanced data sets) ซึ่งควรจะได้มีการศึกษาวิจัยต่อไปในอนาคต อีกทั้งในปัจจุบันขั้นตอนวิธี ELM ก็ยังได้มีการพัฒนาปรับปรุงให้มีประสิทธิภาพไปมากกว่าขั้นตอนวิธี ELM ขั้นพื้นฐานจากนักวิจัยหลายคน [12,13] ดังนั้นในการศึกษาการประยุกต์ใช้ขั้นตอนวิธี ELM ใหม่ๆ ในอนาคตอาจช่วยเพิ่มประสิทธิภาพของการแก้ปัญหาได้เพิ่มมากขึ้น

ตารางที่ 3 การแบ่งชุดข้อมูลสอนและทดสอบที่ละกลุ่ม โดยใช้วิธีการเลือกสุ่มแบบความเที่ยงตรง 10 กลุ่ม

กลุ่ม	10 - Fold Cross Validation																			
	1		2		3		4		5		6		7		8		9		10	
	Tr	Te	Tr	Te	Tr	Te	Tr	Te	Tr	Te	Tr	Te	Tr	Te	Tr	Te	Tr	Te	Tr	Te
1	297	32	296	33	296	33	296	33	296	33	296	33	296	33	296	33	296	33	296	33
2	2150	238	2149	239	2149	239	2149	239	2149	239	2149	239	2149	239	2149	239	2149	239	2150	238
3	228	25	227	26	227	26	227	26	228	25	228	25	228	25	228	25	228	25	228	25
4	16	1	15	2	15	2	15	2	15	2	15	2	15	2	15	2	16	1	16	1
รวม	269	29	2687	30	2687	30	2688	29	2688	29	2688	29	2688	29	2688	29	2689	29	2690	29

Tr คือ ข้อมูลการสอน Te คือ ข้อมูลทดสอบ

ตารางที่ 4 ผลการทดลองของตัวแบบโครงข่ายประสาทเทียม

ขั้นตอนวิธี	จำนวน โหนด	% ค่าเฉลี่ยความถูกต้อง		ค่าเฉลี่ยส่วนเบี่ยงเบน มาตรฐาน		เวลาเฉลี่ย (วินาที)	
		ฝึกสอน	ทดสอบ	ฝึกสอน	ทดสอบ	ฝึกสอน	ทดสอบ
MLP-BP	60	82.51	82.31	0.18	0.29	8.40	0.04
MLP-LM	10	90.48	80.84	0.84	0.72	52.98	0.02
ELM	90	91.42	90.95	0.05	0.18	0.06	น้อยกว่า 0.01

เอกสารอ้างอิง

- [1] สุรเกียรติ อาชานานุภาพ, “ตำราการตรวจรักษาโรคทั่วไป 2:350 โรคกับการดูแลรักษาป้องกัน”, พิมพ์ครั้งที่ 5, โฮลิสติก แพบลิชชี, 2553.
- [2] สำนักงานโรคไม่ติดต่อ, “ข้อมูลสถิติโรคไม่ติดต่อเรื้อรัง”, กรมควบคุมโรค กระทรวงสาธารณสุข (ออนไลน์). สืบค้นจาก <http://thaincd.com/information-statistic/non-communicable-disease-data.php>
- [3] Xiuying D. and Wang P., “Study on the causes of hypertension with improved BP neural network”, 2010 International Conference on E-Health Networking, Digital Ecosystems and Technologies, pp. 21 - 24.
- [4] Sumathi B. and Santhakumaran, A. “Pre-Diagnosis of Hypertension Using Artificial NeuralNetwork”, Journal of Computer Science and Technology Volume 11 Issue 2 Version 1.0 February 2011, pp. 42 - 48.

- [5] Huang, G.-B., Zhu, Q.-Y. and Siew, C.-K. "Extream Learning Machine Theory and applications", Neurocomputing. Vol.70, pp. 489 - 501, 2006.
- [6] Huang, G.-B. Zhu Q.-Y. and Siew, C.-K. "Extream Learning Machine a survey", International Journal of Machine Learning and Cybernetics, Vol.2, No.2, pp. 107 - 122, 2011.
- [7] น้อย อุบลเดชประธารักษ์ณ์. "แนวทางการรักษาโรคความดันโลหิตสูงในเวชปฏิบัติทั่วไป", สำนักงานหลักประกันสุขภาพแห่งชาติ, 2551.
- [8] สิริภัทร เชี่ยวชาญวัฒนา, "เอกสารประกอบการสอนวิชาข่ายงาน ประสาท", ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น, 2552.
- [9] T.Hagan Martin and B.Menhaj, "Brieft Paper Training Feedforward Networks with the Marquardt Algorithm," IEEE Trans.Neural Net., Vol. 3, No, pp. 989 - 993, 1994.
- [10] Hypertension [online]. Available:<http://www.thaihyperten-sion.org>, Access 2013.
- [11] Hypertension, Rick factor.[online]. Available: www.framing-hamheartstudy.org, Access 2013.
- [12] Horata, P., Chiewchanwattana S. and Sunat, K. "Robuts extream learning machines", Neurocomputing, Vol. 102, pp. 31 - 44, 2013.
- [13] ภักราช มุสิกะวัน, สิริภัทร เชี่ยวชาญวัฒนา และ คำรณ สุนิติ, "เครื่องจักรเรียนรู้เอ็กซ์ตรีมเชิงวัฒนาการโดยใช้พื้นฐานของขั้นตอนการค้นหาค่าที่เหมาะสมแบบสแต็ปไชส์แรนคอมและขั้นตอนวิธีแบบหึ่งห้อย", in NCCIT 2012, May. 9 - 10, 2012, pp. 202 - 209.