โปรแกรมช่วยวิเคราะห์อัตราเสี่ยงต่อการเป็นโรคหัวใจและหลอดเลือด

บุณยานุช ใหมเง่า[1], ปัญญาวัฒน์ ขามก้อน[1] , ขนิษฐา ศรีแก้ว[1]
[1]สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏเลย

บทคัดย่อ

ปัจจุบันโรคหัวใจและหลอดเลือดมีประชากรกลุ่มเสี่ยงต่อการเกิดโรคจำนวนมาก หากประชาชนทั่วไปสามารถเข้าถึงการตรวจวิเคราะห์ข้อมูลด้วยตนเองในเบื้องต้นได้ จะสามารถเข้ารับการตรวจและรักษาอย่างทันเวลา งานวิจัยครั้งนี้มีวัตถุประสงค์เพื่อพัฒนาโมเดลการวิเคราะห์อัตราเสี่ยงการเกิดโรคหัวใจและหลอดเลือดด้วยอัลกอริทึมโครงข่ายประสาทเทียม (Neural Network) และต้นไม้แห่งการตัดสินใจ (Decision tree) และเพื่อพัฒนาโปรแกรมวิเคราะห์อัตราเสี่ยงการเกิดโรคหัวใจ และหลอดเลือดจากการผลทดลองการสร้างโมเดลที่มีสามารถคาดการณ์ผลการเกิดอัตราเสี่ยงของโรคหัวใจและหลอดเลือดที่มีความถูกต้องมากที่สุด โดยในการทดลองสร้างโมเดลของทั้งสองอัลกอริทึมใช้โปรแกรมเวก้า (WEKA) เป็นเครื่องมือในการทดสอบการวิเคราะห์และจำแนกขอมูลเพื่อหาอัลกอริทึมที่ดีที่สุด นำไปไปใช้ในการพัฒนาโปรแกรมโดยตัวแปรในการจำแนกข้อมูลนั้น สำหรับข้อมูลที่ใช้ในการจำแนกข้อมูลการสร้างโมเดลมีทั้งหมด 5 แอตทริบิวต์คือ ความดันโลหิตการสูบุหรี่การเป็นโรคเบาหวานเพศและช่วงอายุ โดยเก็บข้อมูลจากบุคคลทั่วไปจำนวน 493 ในเขต ต. นาอ้อ อ.เมือง จังหวัดเลย โดยแบบสำรวจข้อมูลผ่านการตรวจสอบจากเจ้าพนักงานสาธารณสุขชำนาญงานโรงพยาบาลส่งเสริมสุขภาพตำบลศรีสองรัก เป็นผู้ตรวจสอบข้อมูลความถูกต้องก่อนนำไปสร้างโมเดล จากการทดลองสร้างโมเดลพบว่าอัลกอริทึมที่สามารถคาดการณ์จำแนกข้อมูลอัตราเสี่ยงของการเป็นโรคหัวใจและหลอดเลือดได้ดีที่สุดคือ อัลกอริทึมโครงข่ายประสาทเทียมซึ่งมีค่าที่โปรแกรมจำแนกได้ถูกต้องสูงสุดคิดเป็นร้อยละ 97.2973 โดยที่อัลกอริทึมต้นไม้เพื่อการตัดสินใจมีค่าที่โปรแกรมจำแนกได้ถูกต้องคิดเป็นร้อยละ95.3347หลังจากนั้นจึงได้นำโมเดลของอัลกอริทึมโครงข่ายประสาทเทียม และต้นไม้เพื่อการตัดสินใจมาพัฒนาเป็นโปรแกรมช่วยวิเคราะห์อัตราเสี่ยงต่อการเป็นโรคหัวใจและหลอดเลือดผลการทดสอบการพัฒนาโปรแกรมพบว่าโมเดลโครงข่ายประสาทเทียมสามารถคาดการณ์อัตราเสี่ยงถูกต้องอยู่ที่ร้อยละ 99.5

คำสำคัญ : โครงข่ายประสาทเทียม ,ต้นไม้เพื่อการตัดสินใจ,โรคหัวใจและหลอดเลือด

Analyzes the risk of cardiovascular disease.

Bunyanuch Maingao[1], Panyawat Khamkon[1], Kanittha Srikaew[1]
[1]Computer Science Faculty of Science and Technology, Loei Rajabhat University

Abstract

Currently, cardiovascular disease has a population risky for many diseases. The general public can analyze preliminary data by yourself can be examination and treatment in a timely.

The research aims develop to a model analyzes the risk of cardiovascular disease with a neural network algorithm and decision tree algorithm and develop program analyzes the risk of cardiovascular disease. The experiment create a model found that the best algorithm can anticipated classification of the risk of cardiovascular disease by the both algorithms use WEKA program is tools in the analysis and Classified information for the best algorithm. It use in developing by the variables in the data classification. Classification of data create model has 5 attribute include Blood pressure, smoking and diabetes, sex and age by save date from general public 493 sets in Na Oa sub-district, Muang district, Loei Province. The survey data was audited by Correctional health experts Si Song Rak Health Promotion Hospital check accurate Data before create a model. The experiment create a model found that the best algorithm can anticipated classification of the risk of cardiovascular disease. It is neural network algorithm which identified correctly the on maximum 97.2973 percent. Decision tree algorithm can identified correctly the program on 95.3347 After that, a model of neural network algorithm and decision tree algorithm develop to a program help analysis the risk of cardiovascular disease. So, the test of develop program found that neural network model can anticipated correctly the risk on percent 99.5

Keywords : Neural Network, Decision Tree, Cardiovascular

บทนำ

ปัจจุบันประชากรไทยมีแนวโน้มที่จะเป็นโรคหัวใจ และหลอดเลือดเพิ่มมากขึ้น เพื่อให้มีการประเมินความเสี่ยงต่อการเกิดโรคหัวใจและหลอดเลือดที่สามารถประเมินได้ด้วยตนเอง สำนักโรคไม่ติดต่อจึงมีหนังสือแนวทางการประเมินโอกาสเสี่ยงต่อการเกิดโรคหัวใจและหลอดเลือด (สำนักโรคไม่ติดต่อ, ม.ม.ป.) ภายในหนังสือจะมีความรู้เกี่ยวกับโรคหัวใจและหลอดเลือดและยังสามารถประเมินโอกาสเสี่ยงต่อการเกิดโรคของแต่ละบุคคล โดยการประเมินแบ่งออกเป็น 2 กรณี คือ ทราบผล โคเลสเตอรอล (Cholesterol) และกรณีไม่ทราบผลโคเลสเตอรอล และปัจจัยเสี่ยงที่ใช้ในการประเมินมีอยู่ 5 ปัจจัยคือ ช่วงอายุ เพศ ช่วงความดันเลือด การสูบุหรี่ และการเป็นโรคเบาหวาน ปัจจัยเหล่านี้จะถูกนำมาประเมินเพื่อหาอัตราการการเกิดโรคหัวใจและหลอดเลือดและรวมถึงโรคอัมพฤกษ์อัมพาตด้วยใน 10 ปีข้างหน้า โดยมีความเสี่ยงอยู่ 5 ระดับ คือ <10% คือ ความเสี่ยงต่ำ 10-<20% คือความเสี่ยงปานกลาง 20-<30% คือ ความเสี่ยงสูง 30-<40% คือ ความเสี่ยงสูงมาก และ >40%คือ ความเสี่ยงสูงอันตราย

ทางผู้วิจัยทางผู้วิจัยจึงได้ศึกษาและพัฒนาอีกหนึ่งทางเลือกให้กับผู้ใช้ทั่วไป โดยการพัฒนาให้อยู่ในรูปของโปรแกรมวิเคราะห์อัตราเสี่ยงของการเกิดโรคหัวใจและหลอดเลือด สามารถให้ผู้ใช้ได้ ตรวจวิเคราะห์ความเสี่ยงของการเกิดโรคของตนเองได้ โดยโปรแกรมได้อ้างอิงการวิเคราะห์โรคในหนังสือแบบประเมินดังกล่าว ผู้วิจัยจึงได้นำความรู้ทางด้าน ปัญญาประดิษฐ์ มาช่วยในการพัฒนาด้วยโครงข่ายประสาทเทียม (Neural Network) และต้นไม้ตัดสินใจ (Decision Tree) โดยโครงข่ายประสาทเทียมนั้นผู้วิจัยได้เลือกศึกษาอัลกอริทึมการแพร่แบบย้อนกลับ (Back propagation) ส่วนต้นไม้ตัดสินใจได้เลือกศึกษาอัลกอริทึม C4.5 (J48) โดยผู้วิจัยจะทำการศึกษาและเปรียบเทียบทั้ง 2 อัลกอริทึมนี้เพื่อให้ได้อัลกอริทึมที่ดีที่สุดที่จะนำไปช่วยลดกฎและเงื่อนไข เพื่อให้ง่ายต่อการพัฒนาโปรแกรมมากยิ่งขึ้น

วัตถุประสงค์ของการวิจัย

1. เพื่อเปรียบเทียบอัลกอริทึมโครงข่ายประสาทเทียม (Neural Network) และต้นไม้เพื่อการตัดสินใจ (Decision Tree) ในการสร้างโมเดลสำหรับพัฒนาโปรแกรมช่วยวิเคราะห์อัตราเสี่ยงต่อการเป็นโรคหัวใจและหลอดเลือด
2. เพื่อพัฒนาโปรแกรมช่วยวิเคราะห์อัตราเสี่ยงต่อการเป็นโรคหัวใจและหลอดเลือด

วิธีการวิจัย

ขั้นตอนในการวิจัยได้แบ่งออกเป็น 4 ขั้นตอน คือ ศึกษาทฤษฎีและวรรณกรรมที่เกี่ยวข้อง การเก็บรวบรวมข้อมูล การวิเคราะห์ข้อมูลด้วยอัลกอริทึมทางปัญญาประดิษฐ์ และการพัฒนาโปรแกรม

1. ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้อง

การทำเหมืองข้อมูล (Data Mining)

สายชล สินสมบูรณ์ทอง (2558 : 5) Data Mining ถ้าแปลตรง ๆ คือ เหมืองข้อมูล คล้ายกับเหมืองแร่ ที่ขุดดินมาเยอะแต่ได้แร่นิดเดียว Data Mining เป็นศาสตร์แขนงหนึ่งทาง AI (Artificial Intelligence) AI (Artificial Intelligence) ได้นำมาใช้กับData Mining ซึ่งเป็นการค้นหาความสัมพันธ์และรูปแบบทั้งหมด ซึ่งมีอยู่จริงในฐานข้อมูล ซึ่งความสัมพันธ์และรูปแบบเหล่านั้นได้ถูกซ่อนไว้ภายในข้อมูลจำนวนมากที่มีอยู่ Data Mining จะทำการสำรวจและวิเคราะห์ข้อมูลให้อยู่ในรูปแบบที่เต็มไปด้วยความหมายและอยู่ในรูปของกฎ โดยความสัมพันธ์หน่วยนี้แสดงให้เห็นถึงความรู้ต่าง ๆ (Knowledge) ที่มีประโยชน์ในฐานข้อมูลในปัจจุบันองค์กรธุรกิจส่วนใหญ่

เผชิญกับปัญหาของ ข้อมูลดิบจำนวนมาก แต่ข้อมูลที่ประยุกต์ใช้ได้นั้นมีน้อย Data Mining จึงเป็นสาขาที่คาดว่าจะเป็นที่รู้จักและนำมาใช้ประยุกต์ใช้อย่างแพร่หลายเนื่องจาก Data Mining สามารถดึงความรู้ออกมาจากข้อมูลจำนวนมากที่ถูกเก็บสะสมและซ่อนไว้เทคนิคต่าง ๆ ของการทำเหมืองข้อมูล (Techniques in Data Mining)

1.1 ขั้นตอนการสร้างตัวแบบโดยการเรียนรู้จากข้อมูลที่ได้กำหนดคำตอบ (Class) ไว้เรียบร้อยแล้วซึ่งตัวแบ่งที่ได้อาจแสดงในรูปแบบของ 2 ลักษณะคือ

(1)โครงสร้างแผนภาพต้นไม้เพื่อการตัดสินใจ (Decision tree) เป็นที่นิยมกันมากเนื่องจากเป็นลักษณะที่คนจำนวนมากคุ้นเคย ทำให้เข้าใจได้ง่าย มีลักษณะเหมือนแผนภูมิองค์กร โดยที่แต่ละโหนดแสดงคุณลักษณะ (attribute) แต่ละกิ่งแสดงผลในการทดสอบและโหนดของใบ (Leaf node) แสดงคำตอบ (Class) ที่กำหนดไว้

(2) โครงข่ายประสาท (Neural network) เป็นเทคโนโลยีที่มีที่มาจากงานวิจัยด้านปัญญาประดิษฐ์ (Artificial Intelligence : AI) เพื่อใช้ในการคำนวณหาฟังก์ชันจากกลุ่มข้อมูล วิธีการของโครงข่ายประสาท (แท้จริงต้องเรียกให้เต็มว่า Artificial Neural Network หรือ ANN) เป็นวิธีการที่ให้เครื่องเรียนรู้จากตัวอย่างต้นแบบแล้วฝึกหัด (Train) ให้ระบบได้รู้จักละคิดที่จะแก้ปัญหาที่กว้างขวางขึ้นได้ ในโครงสร้างของโครงข่ายประสาทจะประกอบด้วยโหนดข้อมูลเข้า (Input node) และโหนดข้อมูลออกหรือผลลัพธ์ (output node) และการประมวลผลกระจายอยู่ในโครงสร้างเป็นชั้น ๆ ได้แก่ ชั้นข้อมูลเข้า (input layer) ชั้นข้อมูลออกหรือผลลัพธ์ (output layer) และชั้นซ้อน (hidden layer)

1.2 งานวิจัยที่เกี่ยวข้อง

กรรณิการ์ นุชชมภู (2557, บทคัดย่อ) งานวิจัยนี้เป็นการประยุกต์ใช้เทคโนโลยีเหมืองข้อมูลมาใช้ในการพยากรณ์ราคาประเมินที่ดินเพื่อนำไปใช้เป็นแนวทางสำหรับการประเมินราคาที่ดินและช่วยประกอบการตัดสินใจของผู้บริหารหรือการตรวจสอบราคาประเมินที่ดินในปัจจุบัน โดยทำการรวบรวมข้อมูลปัจจัยตัวแปร เช่น ระยะความลึกมาตรฐาน ลักษณะการใช้ ประโยชน์ที่ดิน ลักษณะผิวจราจรข้อมูลราคาตลาด และข้อมูลราคาประเมิน ในพื้นที่อำเภอลาดหลุมแก้ว จังหวัดปทุมธานี โดยใช้เทคนิคการจัดหมวดหมู่ (Classification) มาทำการสร้างโมเดลเพื่อให้ได้ราคาประเมินที่ดินที่ใกล้เคียงกับราคาตลาดมากที่สุด ซึ่งจากผลการทดสอบปรากฏว่า ปัจจัยสำคัญที่มีผลทำให้ ราคาประเมินแตกต่างกันส่วนมากเป็นเรื่องเกี่ยวกับผิวจราจร ไม่ว่าจะเป็นความกว้างของผิวจราจรหรือประเภทของผิวจราจร ซึ่งมีระดับความสำคัญรวมกันถึง 39% และพบว่าทฤษฎีที่ให้ราคาประเมินที่ดีที่สุดคือ Generalized Linear และ Neural Network ซึ่งให้ค่าความแม่นยำถึง 97% ส่วน Decision Tree ก็ให้ค่าความแม่นยำที่สามารถยอมรับได้ คือ 86% ซึ่งสามารถนำไปใช้ประโยชน์สำหรับการสร้างระบบเพื่อการประเมินราคาที่ดินคราวละมากแปลง (Mass Appraisal) หรือระบบแบบจำลองช่วยประเมินราคาคราวละมากแปลง (CAMA: Computer Assisted Mass Appraisal)

2. การรวบรวมข้อมูล

ทำการจัดเตรียมข้อมูลเพื่อใช้ในการสร้างโมเดลและทดสอบอัลกอริธึม มีรายละเอียดดังนี้ โดยทำการสร้างแบบสอบถามเพื่อประเมินความเสี่ยงโรคจำนวน 493 ชุดโดยมีปัจจัยเสี่ยงจำนวนห้าปัจจัยได้แก่ เพศ อายุ การเป็นโรคเบาหวาน ค่าความดันเลือด และการสูบบุหรี่ โดยได้นำข้อมูล และความเสี่ยงของการเกิดโรคหลอดเลือดและหัวใจจากเล่มแนวทางการประเมินโอกาสเสี่ยงต่อการเกิดโรคหัวใจ และหลอดเลือด (สำนักโรคไม่ติดต่อ กรมควบคุมโรค กระทรวงสาธารณสุข,

ม.ม.ป. : 12-13) การเก็บข้อมูลทำการเก็บโดยสำรวจตามหมู่บ้านและทำการสอบถามข้อมูลจากชาวบ้าน โดยมีข้อมูลนำเข้าดังนี้

ตารางที่ 1 ข้อมูลนำเข้าปัจจัยที่ก่อให้เกิดอัตราเสี่ยงต่อการเป็นโรคหลอดเลือดและหัวใจ

ลำดับที่	ชื่อข้อมูล	รายละเอียด
1	Sex	เพศ
2	Age	อายุ
3	Diabetes	โรคเบาหวาน
4	Smoking	สูบบุหรี่
5	Systolic blood pressure(SBP)	ค่าความดันเลือด

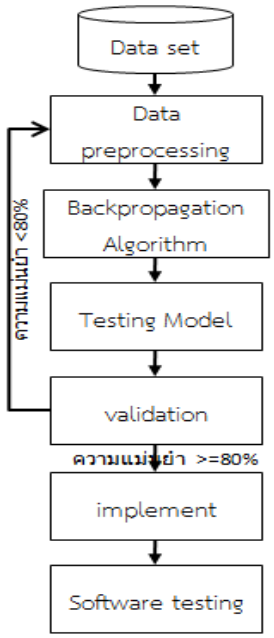
ตารางที่ 2 ข้อมูลผลลัพธ์ (Evaluation Result)

ลำดับ	ชื่อข้อมูล	รายละเอียด
1	Risk1	<10% ระดับต่ำ
2	Risk2	10-<20% ระดับปานกลาง
3	Risk3	20-<30% ระดับสูง
4	Risk4	30-<40% ระดับสูงมาก
5	Risk5	>40% ระดับสูงอันตราย

3. การวิเคราะห์ข้อมูลด้วยอัลกอริทึมทางปัญญาประดิษฐ์

การวิเคราะห์ข้อมูลในครั้งนี้ได้ทำการวิเคราะห์โดยใช้โปรแกรม Weka 3.6 มาใช้ในการวิเคราะห์ข้อมูลและจะแบ่งการวัดประสิทธิภาพออกเป็นสองแบบได้แก่ การวัดประสิทธิภาพแบบ k-fold cross-validation โดยผู้วิจัยได้เลือกเป็น 3-fold cross-validation และการวัดประสิทธิภาพด้วยวิธีการแบ่งข้อมูลแบบสุ่มด้วยการแบ่งร้อยละเพื่อสร้างโมเดลไปพัฒนาโปรแกรมโดยได้แบ่งออกเป็นสองอัลกอริทึมดังนี้

3.1 การสร้างโมเดลโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ



ภาพที่ 1 ขั้นตอนการพัฒนาโปรแกรมด้วยโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ

จากภาพที่ 1 สามารถอธิบายกระบวนการดำเนินงานวิจัยได้ดังนี้

ขั้นตอนที่ 1 Data set ข้อมูลในการดำเนินการวิจัยมีจำนวนทั้งหมด 524 ระเบียน การสอนโครงข่ายจะประกอบไปด้วยขั้นตอนดังนี้

1) กำหนดค่าเริ่มแรกของค่าน้ำหนักหรือเส้นเชื่อม, อัตราการเรียนรู้, ค่าความผิดพาดและค่าไบแอสที่ใช้ในโครงข่ายตามโหนด ต่าง ๆ ของชั้นข้อมูลและชั้นแอบแฝงด้วยวิธีการสุ่ม ซึ่งจะอยู่ในช่วง 0 – 1

2) คำนวณโครงข่ายซึ่งแต่ละโหนดที่จะส่งไปยังชั้นแอบแฝงจะประกอบไปด้วย โหนด 5 โหนด ที่มีอินพุตเข้ามาได้แก่ เพศ อายุ ค่าความดันเลือด การสูบบุหรี่ การเป็นโรคเบาหวาน การรวมค่าน้ำหนักของข้อมูลนี้เข้ากับข้อมูลนำเข้า โดยขั้นตอนเป็นดังนี้

$$X_j = \sum_{i=1}^{L}(W_{ij}.O_i + \theta_j) \qquad (1)$$

เมื่อ X_j คือ ผลรวมของน้ำหนักของ 5 โหนด

W_{ij} คือ ค่าน้ำหนักระหว่าง โหนด i กับ โหนด j

O_i คือ ข้อมูลที่ออกจาก โหนด i

θ_j คือ ไบแอสสำหรับ โหนด j

L คือ จำนวน โหนด ในชั้น i

หลังจากได้ผลการคำนวณจาก (2) แล้วค่าจะถูกแปลงรูปไปเป็นเอาพุตด้วยเทคนิค Sigmoid Function แสดงดังต่อไปนี้

$$O_j = \frac{1}{1+e^{-x}} \qquad (2)$$

เมื่อ O_j คือ transformed outputของ โหนด j

3) หลังจากผ่านการคำนวณใน (2) แล้วก็จะส่งค่าไปยังชั้นข้อมูลออก (output layer) แล้วทำการคำนวณในชั้นข้อมูลออกโดยสมการ (1), (2) ตามลำดับ แล้วทำการคำนวณหาค่าความผิดพลาดโดย

$$E_v = \sqrt{\sum_{k=1}^{N}(W_k - O_k)^2} \qquad (3)$$

เมื่อ E_v คือ ค่าความผิดพลาดของข้อมูลชุดที่ v

W_k คือ ผลลัพธ์จริงของข้อมูล
N คือ จำนวนโหนดในชั้นข้อมูลออก

โดยที่ค่าความผิดพลาดที่คำนวณ (E_v) น้อยกว่าค่าความผิดพลาดที่กำหนดไว้ (E_p) ก็ถือว่ายอมรับได้แต่ถ้าหากค่าความผิดพลาดที่คำนวณมากกว่าค่าความผิดพลาดที่กำหนดไว้ จะต้องทำการปรับน้ำหนักและไบแอสใหม่

4) ทำการปรับค่าน้ำหนักและไบแอสใหม่ระหว่างชั้นข้อมูลออกและชั้นซ่อนจากสมการ

$$\delta_k = O_k(1 - O_k)(T_k - O_k) \quad (4)$$
$$\Delta w_{jk} = \mu \delta_k y_j \quad (5)$$
$$\Delta\theta = \mu\delta_k \quad (6)$$

เมื่อ μ คือ อัตราการเรียนรู้
$\Delta\theta$ คือ ไบแอสที่ชั้นเอาพุตและชั้นซ่อนใหม่
Δw_{jk} คือ ค่าน้ำหนักที่ชั้นเอาพุตและชั้นซ่อนใหม่

5) ทำการปรับค่าน้ำหนักและค่าไบแอสที่โหนดชั้นซ่อนและชั้นอินพุต

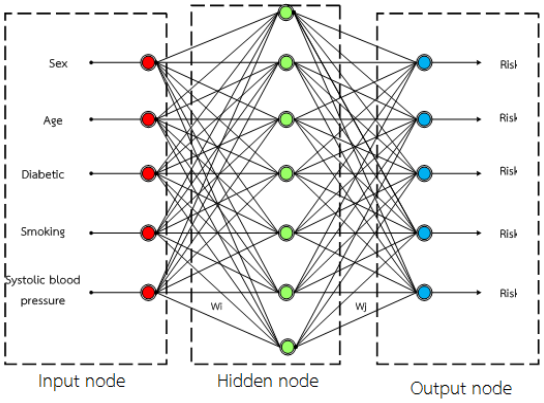
$$\delta_j = (1 - O)O_j \sum_{k=1}^{N} (\delta_k - O_{jk}) \quad (7)$$
$$\Delta w_{jk} = \mu \delta_k y_j \quad (8)$$
$$\Delta\theta = \mu\delta_k \quad (9)$$

เมื่อ μ คือ อัตราการเรียนรู้
$\Delta\theta$ คือ ไบแอสที่ชั้นอินพุตและชั้นซ่อนใหม่
Δw_{jk} คือ ค่าน้ำหนักที่ชั้นอินพุตและชั้นซ่อนใหม่

6) ปรับค่าน้ำหนักใหม่เพื่อใช้ในการคำนวณในรอบถัดไป (n+1) ดังนั้นค่าน้ำหนักใหม่จะได้

$$W_{ij(n+1)} = W_{ij(n)} + \Delta W_{ij(n+1)} \quad (10)$$

7) ทำซ้ำจากข้อ 1 ถึง 6 จนกว่าจะได้ระดับความผิดพลาดที่ต่ำกว่าที่กำหนดแล้วเป็นการจัดกระบวนการเรียนรู้แบบจำลองหากทำการคำนวณซ้ำตามขั้นตอนดังกล่าวนี้แล้ว ค่าระดับความคลาดเคลื่อนยังมากอยู่จะต้องทำการปรับโครงสร้างของโครงข่ายใหม่



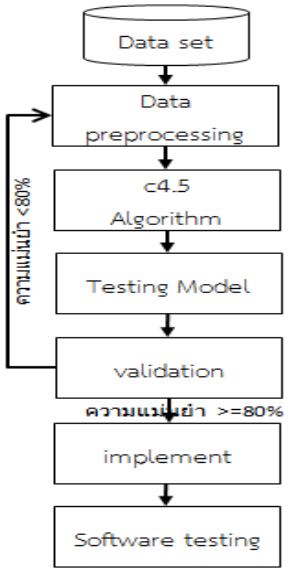
ภาพที่ 2 แบบจำลองโมเดลอัลกอริทึมการเรียนรู้แบบแพร่ย้อนกลับ

จากภาพที่ 2 เป็นแบบจำลองโมเดลอัลกอริทึมการเรียนรู้แบบแพร่ย้อนกลับ สามารถอธิบายในแต่ละชั้นได้ดังนี้

ชั้น Input node คือ ชั้นนำเข้าข้อมูล ประกอบด้วย 5 โหนด ได้แก่ เพศ อายุ โรคเบาหวาน การสูบบุหรี่ และค่าความดันเลือด

ชั้น Hidden node คือ ชั้นซ่อน ใช้ในการประมวลผลจำแนกข้อมูล

ชั้น Output node คือ ชั้นผลลัพธ์ ประกอบด้วย 5 โหนด ตามระดับความเสี่ยงต่อการเป็นโรคหัวใจและหลอดเลือด



ภาพที่ 3 ขั้นตอนการพัฒนาโปรแกรมด้วยต้นไม้เพื่อการตัดสินใจ

3.2 การสร้างโมเดลต้นไม้เพื่อการตัดสินใจ หลักการพื้นฐานของการสร้างแผนภาพต้นไม้เพื่อการตัดสินใจเป็นการสร้างจาก บนลงล่าง คือ เริ่มจากการสร้างรากของต้นไม้ก่อน แล้วจึงแตกกิ่งไปจนถึงใบ โดยแสดงขั้นตอนการสร้างแผนภาพต้นไม้เพื่อการตัดสินใจดังนี้

1) ต้นไม้เริ่มต้นโดยโหนดเดียวแสดงถึงชุดข้อมูลฝึกหัด (training data set)

2) ถ้าข้อมูลทั้งหมดอยู่ในกลุ่มเดียวกัน ให้โหนดนั้นเป็นใบและตั้งชื่อแยกตามกลุ่มของข้อมูลนั้น

3) ถ้าโหนดมีข้อมูลหลายกลุ่มปะปนอยู่ จะต้องวัดค่าผลกำไร (gain) ของแต่ละคุณลักษณะ (attribute) เพื่อที่จะใช้เป็นเกณฑ์ (criterion) ในการคัดเลือกคุณลักษณะที่มี ความสามารถในการแบ่งข้อมูลออกเป็นกลุ่มต่าง ๆ ได้ดีที่สุด โดยคุณลักษณะที่มีผลกำไรมากที่สุดจะถูกเลือกให้เป็นตัวทดสอบหรือคุณลักษณะที่ใช้ในกาตัดสินใจ โดยแสดงในรูปของโหนดบนต้นไม้

4) กิ่งของต้นไม้ถูกสร้างขึ้นจากค่าต่าง ๆ ที่เป็นไปได้ของโหนดทดสอบและข้อมูลจะถูกแบ่งออกตามกิ่งต่าง ๆ ที่สร้างขึ้น

5) ทำการวนซ้ำเพื่อหาคุณลักษณะที่มีผลกำไรมากที่สุด สำหรับข้อมูลที่ถูกแบ่งแยกออกมาในแต่ละกิ่งเพื่อนำคุณลักษณะนี้มาสร้างเป็นโหนดตัดสินใจต่อไป โดยที่คุณลักษณะที่ถูกเลือกมาเป็นโหนดแล้วจะไม่ถูกเลือกมาอีกสำหรับโหนดในระดับต่อ ๆ ไป

6) ทำการวนซ้ำเพื่อแบ่งข้อมูล และแตกกิ่งของต้นไม้ไปเรื่อย ๆ โดยการวนซ้ำจะสิ้นสุดก็ต่อเมื่อเงื่อนไขข้อใดข้อหนึ่งเป็นจริง

3.3 การสร้างโมเดลและการวัดประสิทธิภาพของโมเดล

3.3.1 เทคนิคโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ Backpropagation

1) การสร้างโมเดลด้วยวิธี 3–Fold Cross-Validationได้ค่าCorrectly Classified Instances โมเดลสามารถทำนายข้อมูลถูกต้อง 477ระเบียนหรือคิดเป็น 96.7546 % ของทั้งหมด และ ได้ค่าIncorrectly Classified Instances s โมเดลทำนายข้อมูลไม่ถูกต้อง 16ระเบียนหรือคิดเป็น 3.2454 % ของทั้งหมด และในส่วน Root Mean Squared Error (RMSE) ค่าความคลาดเคลื่อนระหว่างค่าจริง และค่าที่ทำนายได้มีค่าเท่ากับ 0.1063 เพื่อให้ได้ค่า Correctly Classified Instances หรือค่าที่โปรแกรมทำนายได้ถูกต้องมากที่สุดจึงได้ทำการปรับเปลี่ยนโหนดในชั้นซ่อนโดยเพิ่มจากที่โปรแกรมกำหนดไว้ทีละโหนด

2) การสร้างโมเดลด้วยวิธี Percentage split 70% ได้ค่าCorrectly Classified Instances โมเดลสามารถทำนายข้อมูลถูกต้อง 144 ระเบียนหรือคิดเป็น 97.2973% ของทั้งหมด และค่า Incorrectly Classified Instances จากโมเดลทำนายข้อมูลไม่ถูกต้อง 4 ระเบียนหรือคิดเป็น 2.7027 % ของทั้งหมด และในส่วนของค่า Root Mean Squared Error (RMSE) ค่าความคลาดเคลื่อนระหว่างค่าจริงและค่าที่ทำนายได้ มีค่าเท่ากับ 0.0884

3.3.2 เทคนิคต้นไม้ตัดสินใจ (Decision Tree) C4.5 (J48)

1) การสร้างโมเดลด้วยวิธี 3–Fold Cross-Validationได้ค่าCorrectly Classified Instances โมเดลสามารถทำนายข้อมูลถูกต้อง 470 ระเบียนหรือคิดเป็น 95.3347 % ของทั้งหมด และได้ค่าIncorrectly Classified Instances s โมเดลทำนายข้อมูลไม่ถูกต้อง 23 ระเบียนหรือคิดเป็น 4.6653 % ของทั้งหมด และในส่วน Root Mean Squared Error (RMSE) ค่าความคลาดเคลื่อนระหว่างค่าจริงและค่าที่ทำนายได้ มีค่าเท่ากับ 0.1308

2) การสร้างโมเดลด้วยวิธี Percentage split 70% ได้ค่า Correctly Classified Instances โมเดลสามารถทำนายข้อมูลถูกต้อง 139 ระเบียนหรือคิดเป็น 93.9189 % ของทั้งหมดและค่า Incorrectly Classified Instances จากโมเดลทำนายข้อมูลไม่ถูกต้อง 9 ระเบียนหรือคิดเป็น 6.0811% ของทั้งหมด และในส่วนของค่า Root Mean Squared Error (RMSE) ค่าความคลาดเคลื่อนระหว่างค่าจริงและค่าที่ทำนายได้ มีค่าเท่ากับ 0.145

3.4 การเปรียบเทียบประสิทธิภาพของโมเดล

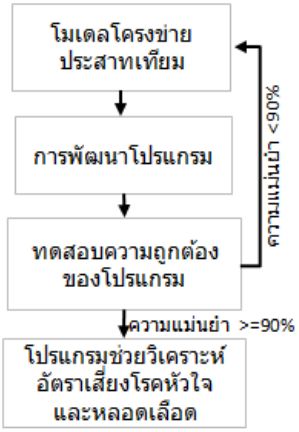
3.4.1 เปรียบเทียบโมเดลที่สร้างขึ้นจากทั้ง 2 อัลกอริทึม คือ C4.5 (J48) และ Backpropagation ด้วยวิธี 3–Fold Cross-Validation

จากผลการเปรียบเทียบด้วยวิธี 3–Fold Cross-Validation จะเห็นผลลัพธ์ว่าอัลกอริทึม Backpropagation ให้มีความถูกต้องสูงถึง 96.7546% และอัลกอริทึม C4.5 (J48) มีค่าทำนายถูกต้องอยู่ที่ 95.3347%

ดังนั้นจึงสรุปได้ว่า อัลกอริทึม Backpropagation เหมาะกับข้อมูลชุดนี้ และจะใช้อัลกอริทึม Backpropagation ในการพัฒนาระบบต่อไป

4. ขั้นตอนการพัฒนาโปรแกรม

4.1 เมื่อได้โมเดลโครงข่ายประสาทเทียมเรียบร้อยแล้ว ก็จะโมเดลไปพัฒนาโปรแกรมเป็นเว็บแอพลิเคชั่น



ภาพที่ 4 การพัฒนาโปรแกรมช่วยวิเคราะห์อัตราเสี่ยงโรคหัวใจและหลอดเลือดด้วยอัลกอริทึม Backpropagation

โดยข้อมูลในโมเดลก็จะประกอบไปด้วยค่าน้ำหนักของแต่ละโหนด ค่า threshold และโหนดผลลัพธ์ เมื่อพัฒนาโปรแกรมเสร็จแล้วก็นำโปรแกรมไปทดสอบความถูกต้องผลทดสอบพบว่าโปรแกรมช่วยวิเคราะห์อัตราเสี่ยงโรคหัวใจและหลอดเลือดด้วยอัลกอริทึม Backpropagation ได้มีค่าที่โปรแกรมทำนายได้ถูกต้องอยู่ที่ร้อยละ 99.5 จากข้อมูลทั้งหมด 200 ระเบียนคิดเป็น 199 ระเบียน และทำนายไม่ถูกต้องอยู่ที่ร้อยละ 0.5 จากข้อมูลทั้งหมด 200 ระเบียนคิดเป็น 1 ระเบียน

4.2 เมื่อได้โมเดลต้นไม้เพื่อการตัดสินใจเรียบร้อยแล้ว ก็จะโมเดลไปพัฒนาโปรแกรมเป็นวินโดว์แอพลิเคชั่น

ผลการวิจัย

ในการสร้างโมเดลจะใช้ 2 เทคนิค โดยจะมีเทคนิคโครงข่ายประสาทเทียมโดยเลือกใช้อัลกอริทึม Backpropagation และใช้เทคนิคต้นไม้ตัดสินใจ โดยเลือกใช้อัลกอริทึม C4.5 (J48) โดยมีข้อมูลที่ใช้ในการสร้างโมเดล 493 ระเบียน จากการทดสอบข้อมูลด้วยทั้ง 2 อัลกอริทึม พบอัลกอริทึม Backpropagation สามารถจำแนกข้อมูลได้ดีกว่าอัลกอริทึมต้นไม้ตัดสินใจ

สรุปและอภิปรายผล

ในการจัดทำวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบอัลกอริทึม C4.5 (J48) และ Backpropagation และหากอัลกอริทึมใดมีประสิทธิภาพมาที่ดีเหมาะสำหรับ การพัฒนาโปรแกรมช่วยวิเคราะห์อัตราเสี่ยงต่อการเป็นโรคหัวใจและหลอดเลือด เนื่องจากเป็นอัลกอริธึมที่นิยมนำมาใช้ในการจำแนกข้อมูล โดยการนำอัลกอริธึมดังกล่าวมาสร้างโมเดลและวัดประสิทธิภาพกับกลุ่มข้อมูลจำนวน 493 ชุด เพื่อสร้างโมเดลในโปรแกรม Weka 3.6.2 โมเดลที่ให้ผลค่าความถูกต้องมากที่สุดไปใช้ในการพัฒนาโปรแกรม

1. อัลกอริธึม Backpropagation โมเดลที่ได้มีปรับโหนดในชั้นซ่อนมีจำนวน 5 โหนด ใช้วิธี 3–Fold Cross-Validation มีค่าความถูกต้องจากข้อมูล 493 ระเบียนโมเดลสามารถทำนายข้อมูลถูกต้อง 477 ระเบียน หรือคิดเป็น 96.7546% ของทั้งหมด มีค่าความไม่ถูกต้อง โมเดลทำนายข้อมูลไม่ถูกต้อง 16 ระเบียน หรือคิด 2.027% ของทั้งหมด และมีค่าความคลาดเคลื่อนระหว่าง

ค่าจริงและค่าที่พยากรณ์ได้ (Root Mean Squared Error : RMSE) เท่ากับ 0.1063 หลังจากนำโปรแกรมที่พัฒนาด้วยอัลกอริธึม Backpropagation ใช้วิธี 3–Fold Cross-Validation มาทดสอบความถูกต้อง ค่าความถูกต้องที่ได้อยู่ที่ร้อยละ 99.5 จากข้อมูล 200 ระเบียนทำนายถูกต้อง 199 ระเบียน และทำนายไม่ถูกต้องอยู่ที่ร้อยละ 0.5 จากข้อมูล 200 ระเบียนทำนายไม่ถูกต้อง 1 ระเบียน

2. อัลกอริธึม C4.5 (J48) โมเดลที่ได้มีโหนดราก คือ ค่าความดันเลือด ใช้วิธีวิธี 3–Fold Cross-Validation มีค่าความถูกต้อง จากข้อมูล 493 ระเบียนโมเดลสามารถทำนายข้อมูล ถูกต้อง 470 ระเบียนหรือคิดเป็น 95.2703% ของทั้งหมด มีค่าความไม่ถูกต้องโมเดลทำนายข้อมูลไม่ถูกต้อง 23 ระเบียน หรือคิด 4.6653% ของทั้งหมด และมีค่าความคลาดเคลื่อนระหว่างค่าจริงและค่าที่พยากรณ์ได้ เท่ากับ 0.1308 หลังจากนำโปรแกรมด้วยอัลกอริธึม C4.5 (J48) ใช้วิธี 3–Fold Cross-Validation มาทดสอบความถูกต้องที่พัฒนามาทดสอบความถูกต้อง ค่าความถูกต้องที่ได้อยู่ที่ร้อยละ 98 จากข้อมูล 200 ระเบียนทำนายถูกต้อง 196 ระเบียน และทำนายไม่ถูกต้องอยู่ที่ร้อยละ 2จากข้อมูล 200 ระเบียนทำนายไม่ถูกต้อง 4 ระเบียน

ข้อเสนอแนะ

เพื่อให้โปรแกรมช่วยวิเคราะห์อัตราเสี่ยงต่อการเป็นโรคหัวใจและหลอดเลือด โดยใช้โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับมีประสิทธิภาพมากขึ้น ทางผู้วิจัยเห็นว่าระบบควรได้รับการพัฒนา ในเรื่องของการเก็บรวบรวมข้อมูลที่ใช้ในการวิเคราะห์ให้มีจำนวนมาก เพื่อให้ได้ผลของการวิเคราะห์ข้อมูลมีความสมบูรณ์มากขึ้น

References

Jurarat Tangkittiwat, Nalinpat Porrawatpreyakorn. (2557). **A Model for Swine Disease Analysis using Neural Network.** The Tenth National Conference on Computing and Information Technology (NCCIT2014). pp. 26-31, Bangkok.

Phanthipha Petchboonmee, Duangkamol Phonak and Monchai Tiantong. (2556). The Forecastion of David Kolb's Experiential Learning Style Using the Classification Rules with Decision Tree Technique. **Thammasat Journal of Science and Technology,** 21(6), 547-557.

Bureau of Non Communicable Diseases. **(n.d). Guidelines for Assessment of Cardiovascular Risk**. Bangkok: The war veterans organization of thailand.

Kannika Nutchomphu, Maleerat Sodanil. (2557). **Land Price Forecasting using Data Mining Techniques.** The Tenth National Conference on Computing and Information Technology (NCCIT2014), pp. 671-676, Bangkok.