การพัฒนาตัวแบบเพื่อพยากรณ์คุณภาพผลิตภัณฑ์ฮาร์ดดิสก์ด้วยการถดถอยโลจิสติกส์และโครงข่ายประสาทเทียมโดยใช้การวิเคราะห์เหมืองข้อมูล

Model Development to Predict Quality of Hard Disk Drive with Regression Logistics and Neural Network Using Data Mining

จามรี ชูบัวทอง* และสมศรี บัณฑิตวิไล

ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ถนนฉลองกรุง เขตลาดกระบัง กรุงเทพมหานคร 10520

Chammaree Chubuathong* and Somsri Banditvilai

Department of Statistics, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang,

Chalongkrung Road, Ladkrabang, Bangkok 10520

บทคัดย่อ

ฮาร์ดดิสก์เป็นอุปกรณ์จัดเก็บข้อมูลอิเล็กทรอนิกส์ที่มีความสำคัญมากในปัจจุบัน โดยมีหน้าที่ในการจัดเก็บข้อมูลต่าง ๆ ซึ่งมีขนาดพื้นที่ในการจัดเก็บเติบโตแบบทวีคูณและยังได้มีการขยายตลาดจากตลาดคอมพิวเตอร์ไปสู่ตลาดอื่น ๆ ได้แก่ โทรทัศน์ กล้องวงจรปิด เป็นต้น ในยุคที่มีการแข่งขันทางด้านเทคโนโลยีสูง บริษัทผลิตฮาร์ดดิสก์จึงต้องมีการคิดค้นวิธีการหรือกระบวนการผลิตฮาร์ดดิสก์ เพื่อให้ได้ฮาร์ดดิสก์มีคุณภาพดีขึ้น อีกวิธีหนึ่งซึ่งใช้กันอย่างแพร่หลายคือการลดจำนวนของเสีย งานวิจัยนี้เป็นการพัฒนาตัวแบบเพื่อพยากรณ์คุณภาพของผลิตภัณฑ์ฮาร์ดดิสก์ในกระบวนการทดสอบความเชื่อมั่นก่อนการส่งมอบแก่ผู้บริโภค ซึ่งช่วยลดการสูญเสียทรัพยากร เช่น ลดต้นทุนและเวลาในการทดสอบ เนื่องจากจำนวนข้อมูลในการทำงานวิจัยนี้มีขนาดใหญ่จึงใช้การวิเคราะห์เหมืองข้อมูลเข้ามาช่วย และแบบจำลองที่ใช้ในการพยากรณ์คุณภาพของผลิตภัณฑ์ฮาร์ดดิสก์ที่พิจารณามีด้วยกัน 2 แบบ คือ การวิเคราะห์การถดถอยโลจิสติกส์และโครงข่ายประสาทเทียม โดยมีจำนวนตัวแปรอิสระทั้งหมดคือ 197 ตัวแปร ซึ่งเป็นตัวแปรเชิงปริมาณทั้งหมด เพื่อลดจำนวนตัวแปรอิสระลงจึงทดสอบความสัมพันธ์ระหว่างตัวแปรอิสระกับคุณภาพฮาร์ดดิสก์ พบว่ามีตัวแปรอิสระเพียง 25 ตัว ที่มีความสัมพันธ์กับคุณภาพของฮาร์ดดิสก์ จึงนำตัวแปรอิสระทั้ง 25 ตัว ไปสร้างแบบจำลองโครงข่ายประสาทเทียม ส่วนแบบจำลองโลจิสติกส์จะต้องกำจัดปัญหาตัวแปรอิสระที่มีความสัมพันธ์กันเองออกก่อน คงเหลือตัวแปรที่ใช้ในการสร้างแบบจำลอง 19 ตัวแปร จากงานวิจัยพบว่าแบบจำลองที่ได้จาก

***ผู้รับผิดชอบบทความ :** suay-physics@hotmail.com

โครงข่ายประสาทเทียมให้ความถูกต้องในการพยากรณ์คุณภาพของผลิตภัณฑ์ฮาร์ดดิสก์ได้ดีกว่าแบบจำลองจากการวิเคราะห์การถดถอยโลจิสติกส์

คำสำคัญ : เหมืองข้อมูล; โครงข่ายประสาทเทียม; การถดถอยโลจิสติกส์

Abstract

Currently, hard disk drive is an electronic storage device that is very important. It is responsible for storing information which its size is growing exponentially. It has expanded from computers market to other markets, such as television cameras and so on. In a highly competitive technology era, the company that produces hard disk drive must be considered about inventing a new method or process for producing hard disk drive with higher quality. Another method, which is widely used, is to reduce waste. This thesis is developed model to predict the quality of hard disk drive in an outgoing reliability test process. This reduces the loss of resources, such as cost reduction and testing time. The amount of data in this research is huge, therefore 2 data mining techniques which are logistic regression analysis and neural network are employed in this study. There are 197 quantitative independent variables. In order to reduce the independent variables, the relationship between each independent variable and the quality of hard disk drive was tested and found that only 25 variables are related to the quality of hard disk drive. Therefore, 25 independent variables were used to build the neural network. Since logistic regression analysis need to eliminate independent variables that has multicollinearity problem first. Then it was left 19 variables to build logistic regression analysis. The result shows that neural network model performs better prediction than logistic regression analysis model.

Keywords: data mining; neural network; logistic regression

1. บทนำ

ฮาร์ดดิสก์เป็นอุปกรณ์จัดเก็บข้อมูลอิเล็กทรอนิกส์ที่มีความสำคัญมากในปัจจุบัน โดยมีหน้าที่ในการจัดเก็บข้อมูลต่าง ๆ บนปริมาณพื้นที่ที่เพิ่มขึ้นจาก 500 จิกะไบต์ (GB) ไปจนถึง 6 เทระไบต์ (TB) และมีแนวโน้มในการเพิ่มจำนวนความจุอย่างต่อเนื่อง ทำให้อุตสาหกรรมฮาร์ดดิสก์ (hard disk) ไม่ได้หยุดอยู่กับที่บนตลาดคอมพิวเตอร์เพียงอย่างเดียว มีการขยายตลาดฮาร์ดดิสก์ไปใช้ในอุปกรณ์อื่น ๆ อย่างแพร่หลาย ได้แก่ โทรศัพท์มือถือ เครื่องเล่นดีวีดี (DVD) โทรทัศน์และกล้องวงจรปิด เป็นต้น สำหรับฮาร์ดดิสก์ถูกจัดเป็นองค์ประกอบที่สำคัญไม่น้อยไปกว่าระบบคอมพิวเตอร์และเป็นอุปกรณ์ที่ง่ายต่อการอัพเกรดสเป็ค เช่น ความจุ ความเร็วรอบ ขนาดของหน่วยความจำแคช ในยุคที่มีการแข่งขันทางด้านเทคโนโลยีสูง บริษัทผลิตฮาร์ดดิสก์จึงต้องมีการคิดค้นวิธีการ กระบวนการผลิตฮาร์ดดิสก์ เพื่อให้ได้ฮาร์ดดิสก์ที่มีคุณภาพที่ดีที่สุดในราคาถูก อีกวิธีหนึ่งซึ่งใช้กันอย่างแพร่หลายคือการลด

จำนวนของเสีย ซึ่งสามารถลดต้นทุนการผลิตได้เป็นอย่างดี

งานวิจัยนี้ ผู้วิจัยมุ่งเน้นการป้องกันการเกิดของเสียในกระบวนการทดสอบความเชื่อมั่นก่อนการส่งมอบผลิตภัณฑ์ให้แก่ผู้บริโภค (outgoing reliability test, ORT) โดยเก็บรวบรวมข้อมูลอาการผิดปกติของฮาร์ดดิสก์ที่กระบวนการทดสอบความเชื่อมั่นก่อนการส่งมอบ พบว่าลักษณะอาการผิดปกติที่ชุดหัวอ่าน/เขียนมีเป็นจำนวนมาก จึงศึกษาและสร้างแบบจำลองเพื่อพยากรณ์ฮาร์ดดิสก์ว่าชุดหัวอ่าน/เขียนจะมีอาการผิดปกติหรือไม่ ดังนั้นผู้วิจัยจึงศึกษางานวิจัย ขั้นตอนและวิธีการวิเคราะห์ข้อมูล เพื่อที่จะได้มาซึ่งรูปแบบการพยากรณ์ของเสียในกระบวนการทดสอบความเชื่อมั่นก่อนการส่งมอบผลิตภัณฑ์ โดยศึกษาถึงปัจจัยที่ส่งผลกระทบกับคุณภาพฮาร์ดดิสก์ พร้อมทั้งศึกษาระดับความสัมพันธ์ของปัจจัย ในกระบวนการทดสอบผลิตภัณฑ์ฮาร์ดดิสก์ก่อนที่จะเป็นผลิตภัณฑ์ที่เสร็จสมบูรณ์ พิจารณาความสามารถในการพยากรณ์คุณภาพความเชื่อมั่นของผลิตภัณฑ์ฮาร์ดดิสก์ และนำแบบจำลองที่เหมาะสมที่สร้างขึ้นมาใช้ในการพยากรณ์คุณภาพของผลิตภัณฑ์ฮาร์ดดิสก์ก่อนขั้นตอนการทดสอบคุณภาพความเชื่อมั่นก่อนการส่งมอบแก่ผู้บริโภค

เนื่องจากข้อมูลที่จะใช้ในงานวิจัยนั้นมีตัวแปรอิสระจำนวน 197 ตัว และมีปริมาณข้อมูลขนาดใหญ่ จึงเลือกใช้เทคนิคการทำเหมืองข้อมูล (data mining) หรืออาจเรียกว่าการค้นหาความรู้ในฐานข้อมูล (knowledge discovery in databases, KDD) คือ กระบวนการที่กระทำกับข้อมูลที่มีจำนวนมาก เพื่อค้นหารูปแบบ แนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น โดยใช้การวิเคราะห์การถดถอยโลจิสติกส์ (logistic regression analysis) ซึ่งสามารถใช้กับตัวแปรตามที่เป็นตัวแปรเชิงกลุ่มหรือเชิงคุณภาพ ซึ่งมีวัตถุประสงค์และแนวคิดเหมือนกับการวิเคราะห์การถดถอยเชิงเส้น คือ เพื่อศึกษาความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระ และนำสมการที่ได้ไปประมาณค่าหรือพยากรณ์ตัวแปรตาม วิธีที่สอง โครงข่ายประสาทเทียม โครงข่ายประสาทเทียม (neural network) เป็นศาสตร์แขนงหนึ่งทางด้านปัญญาประดิษฐ์ ที่สามารถนำไปประยุกต์ใช้กับงานหลายด้านได้อย่างมีประสิทธิภาพ ได้แก่ การจำแนกรูปแบบ การพยากรณ์ การควบคุม การหาความเหมาะสม และการจัดกลุ่ม เป็นต้น [1]

1.1 การวิเคราะห์การถดถอยโลจิสติกส์ (logistic regression analysis) ถูกนำมาใช้เพื่อทำนายว่าจะเกิดเหตุการณ์หนึ่งขึ้นหรือไม่ โดยมีการกำหนดค่าตัวแปรตัวหนึ่งหรือหลายตัวที่คาดว่าจะส่งผลต่อการเกิดเหตุการณ์นั้น [2]

ข้อตกลงเบื้องต้นการวิเคราะห์การถดถอยโลจิสติกส์

1.1.1 ตัวแปรอิสระ เป็นตัวแปรที่ระดับข้อมูลอยู่ในระดับช่วง (interval scale) เป็นอย่างต่ำ กรณีที่เป็นข้อมูลเชิงกลุ่มให้แปลงเป็นตัวแปรหุ่น (dummy variable) ส่วนตัวแปรตามหรือตัวแปรเกณฑ์ กรณีที่เป็นการวิเคราะห์โลจิสติกส์แบบทวิ (binary logistic regression) จะกำหนด 2 ค่า คือ 0 กับ 1 ส่วนกรณีการวิเคราะห์โลจิสติกส์พหุกลุ่ม (multinomial logistic regression) จะกำหนดตามจำนวนกลุ่มของตัวแปรตาม

1.1.2 ค่าเฉลี่ยความคลาดเคลื่อนเป็นศูนย์และไม่มีความสัมพันธ์กัน นั่นคือ E(e) = 0

1.1.3 ตัวแปรอิสระไม่มีความสัมพันธ์กันเองหรือไม่เกิดปัญหา multicollinearity ซึ่งก็คือตัวแปรอิสระที่มีอยู่อาจมีความสัมพันธ์กันเองได้ โดยจะส่งผลในการเลือกตัวแปรอิสระที่มีความสำคัญต่อตัวแปรตาม หากนักวิจัยเก็บตัวแปรอิสระเหล่านั้นไว้

ทั้งหมดในตัวแบบจะทำให้เกิดการซ้ำซ้อนและตัวแบบจะใหญ่เกินความจำเป็น

1.1.4 การวิเคราะห์การถดถอยโลจิสติกส์จะต้องใช้ขนาดตัวอย่าง n มากกว่าการวิเคราะห์การถดถอยแบบปกติ โดยจะใช้ขนาดตัวอย่างเท่ากับ $n \geq 30p$ โดยที่ p คือ จำนวนตัวแปรทำนาย [3]

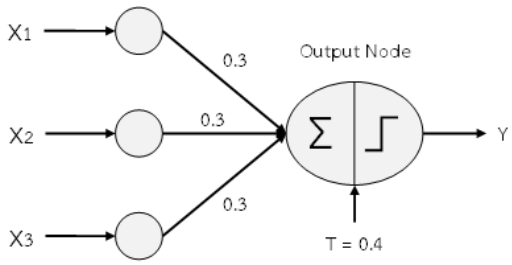
เหตุผลที่ใช้การวิเคราะห์การถดถอยโลจิสติกส์แทนการวิเคราะห์การถดถอยเชิงเส้น คือ เมื่อตัวแปรตามมีค่าได้เพียง 2 ค่า ทำให้ค่าประมาณของตัวแปรตามเป็นโอกาสที่เหตุการณ์ที่สนใจ มีค่าระหว่าง 0 ถึง 1 ถ้าใช้สมการถดถอยเชิงเส้นตรง ค่า Y ที่ได้อาจจะไม่ได้อยู่ในช่วง 0 ถึง 1 หรืออาจมีค่าน้อยกว่า 0 หรือมากกว่า 1 ในการวิเคราะห์การถดถอยเชิงเส้น มีเงื่อนไขว่าค่าความคลาดเคลื่อนต้องมีการแจกแจงแบบปกติ แต่เมื่อตัวแปรตามมีค่าเพียง 2 ค่า คือ 0 กับ 1 จะทำให้ค่าความคลาดเคลื่อน e มีค่าได้เพียง 2 ค่าด้วย ซึ่งเป็นไปไม่ได้ที่ e จะมีการแจกแจงแบบปกติ และค่าความแปรปรวนไม่คงที่ (non-constant error variance) เนื่องจากเงื่อนไขของการวิเคราะห์การถดถอย คือ ค่าความแปรปรวนของค่าความคลาดเคลื่อนหรือ V(e) ต้องคงที่ทุกค่าของ X แต่ในโลจิสติกส์นั้น เมื่อตัวแปรตามมีค่าได้เพียง 2 ค่า และตัวแปรตามมีการแจกแจงแบบเบอร์นูลี่ ซึ่งทำให้ค่าความแปรปรวนและค่าเฉลี่ยมีความสัมพันธ์กัน จึงทำให้เงื่อนไขที่ว่า V(e) คงที่ไม่เป็นจริง ซึ่งทำให้ไม่สามารถใช้การวิเคราะห์การถดถอยเชิงเส้นตรงตามปกติได้ [3]

1.2 โครงข่ายประสาทเทียม (artificial neural network) โครงข่ายประสาทเทียมได้รับแรงบันดาลใจจากการพยายามที่จะจำลองระบบประสาททางชีวภาพของมนุษย์ สมองมนุษย์โดยส่วนใหญ่จะประกอบด้วยเซลล์ประสาทหลักที่เรียกว่านิวรอน (neuron) การเชื่อมโยงของเซลล์ประสาทผ่านทางเส้นใยที่เรียกว่าแอคซอน (axon) ถูกใช้สำหรับส่งกระแสประสาทจากเซลล์ประสาทหนึ่งไปยังอีกเซลล์ประสาทหนึ่ง เมื่อใดก็ตามที่เซลล์ประสาทถูกกระตุ้น เซลล์ประสาทจะเชื่อมต่อกับแอคซอนของเซลล์ประสาทอื่นๆผ่านทางเดนไดรท์ (dendrite) จุดเชื่อมต่อระหว่างเดนไดรท์และแอคซอนเรียกว่าไซแนปส์ (synapse) นักประสาทวิทยาได้ค้นพบว่าสมองมนุษย์เรียนรู้โดยการเปลี่ยนแปลงความแข็งแรงของไซแนปส์ที่ใช้สำหรับการเชื่อมต่อระหว่างเซลล์ประสาท เมื่อกระตุ้นซ้ำโดยการใช้แรงกระตุ้นเดียวกัน [4]

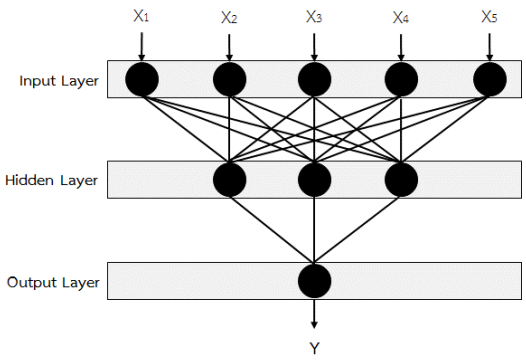
1.2.1 เปอร์เซฟตรอน (perceptron) เป็นรูปแบบที่ง่ายที่สุดประกอบด้วยโหนดสองชนิดคือ อินพุทโหนด (input node) ซึ่งจะใช้เป็นตัวแทนของการป้อนข้อมูลนำเข้า และเอาท์พุทโหนด (output node) ซึ่งจะใช้เป็นตัวแทนแสดงรูปแบบผลลัพธ์ (model output) โหนดในสถาปัตยกรรมโครงข่ายประสาทเทียมเปรียบได้ว่าเป็นเซลล์ประสาทหรือหน่วยงานในเปอร์เซฟตรอน แต่ละอินพุตโหนดจะเชื่อมต่อกันด้วยสายเชื่อมโยงโดยมีค่าน้ำหนักกับเอาท์พุตโหนด การเชื่อมโยงโดยมีค่าน้ำหนักเป็นการจำลองความแข็งแรงของไซแนปส์ ซึ่งเป็นจุดเชื่อมต่อระหว่างเซลล์ประสาท ดังระบบประสาททางชีวภาพ การเรียนรู้ของเปอร์เซฟตรอนจะทำการปรับค่าน้ำหนักการเชื่อมโยง จนกว่าจะได้ค่าน้ำหนักที่พอดีกับความสัมพันธ์ของข้อมูลนำเข้าและผลลัพธ์ เปอร์เซฟตรอนจะคำนวณค่าเอาท์พุต โดยการแสดงผลรวมการถ่วงน้ำหนักของปัจจัยหรือตัวแปรอิสระลบด้วยปัจจัยเอนเอียง (bias factor) T ดังแสดงในรูปที่ 1 [4]

การเรียนรู้ของเปอร์เซฟตรอน พารามิเตอร์ถ่วงน้ำหนัก W จะถูกปรับค่าจนกว่าผลลัพธ์ที่ได้จะมีความสอดคล้องกับผลลัพธ์ที่แท้จริง หลักสำคัญในการคำนวณหาค่าถ่วงน้ำหนักเป็นดังสมการ $w_j^{(k+1)} = w_j^{(k)} + \lambda(y_i - \hat{y}_i^{(k)})x_{ij}$ เมื่อ $w_j^{(k)}$ คือ พารามิเตอร์น้ำหนักที่เกี่ยวข้องกับ x_i เชื่อมต่อข้อมูลนำเข้า

หลังจากการทำซ้ำ k ครั้ง λ คือพารามิเตอร์ที่ทราบค่านั้นก็คืออัตราการเรียนรู้ (learning rate) และ x_{ij} คือค่าครั้งที่ j ของตัวแปรที่ i ในการปรับปรุงค่าน้ำหนักดังสมการ ซึ่งแสดงค่าน้ำหนักใหม่คือ $w_j^{(k+1)}$ เป็นการรวมกันของค่าน้ำหนักเก่า และเทอมของสัดส่วนในการพยากรณ์ผิดพลาด ถ้าค่าพยากรณ์ถูกต้องแล้วค่าถ่วงน้ำหนักจะมีค่าเท่าเดิม [4]



รูปที่ 1 สถาปัตยกรรมโครงข่ายประสาทเทียมแบบเปอร์เซฟตรอน



รูปที่ 2 โครงข่ายประสาทเทียมหลายชั้นแบบ feed-forward

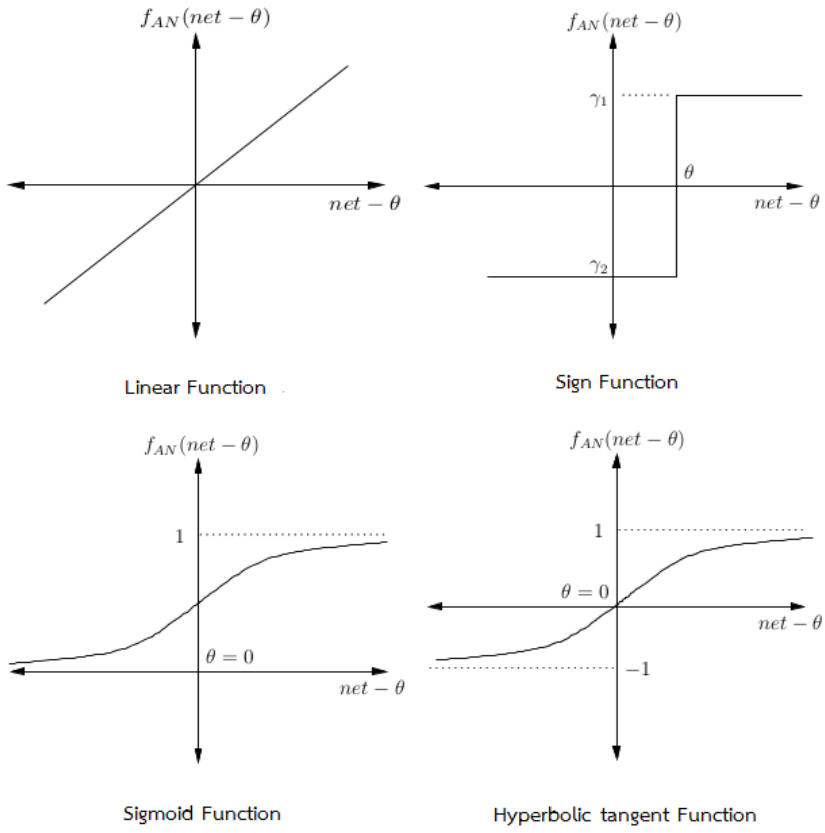
1.2.2 โครงข่ายประสาทเทียมแบบหลายชั้น (multilayer artificial neural network) โครงข่ายมีหลายชั้น ซึ่งชั้นที่อยู่ตรงกลางระหว่างชั้นข้อมูลนำเข้าและชั้นผลลัพธ์ เรียกว่าชั้นซ่อน (hidden layer) และโหนดที่ฝังอยู่ในชั้นนี้เรียกว่าโหนดซ่อน (hidden node) ในโครงข่ายประสาทเทียมแบบ feed-forward โหนดในชั้นหนึ่งจะเชื่อมต่อกับโหนดในชั้นถัดไปเท่านั้น ในโครงข่ายประสาทเทียมหลายชั้น การเชื่อมโยงระหว่างโหนดในแต่ละชั้นจะมีเส้นเชื่อมทั่วถึงกันทุกโหนด (fully connection) ดังแสดงในรูปที่ 2 [5]

โครงข่ายประสาทเทียมอาจใช้ฟังก์ชั่นกระตุ้นอื่น ๆ นอกเหนือจากไซน์ฟังก์ชั่น (sign function) ตัวอย่างของฟังก์ชั่นกระตุ้นอื่น ๆ เช่น ฟังก์ชั่นเชิงเส้น ซิกมอยด์ฟังก์ชั่น (sigmoid) และไฮเปอร์โบลิคเทนเจนท์ฟังก์ชั่น (hyperbolic tangent) ดังแสดงในรูปที่ 3 ฟังก์ชั่นกระตุ้นเหล่านี้ช่วยให้โหนดซ่อนและโหนดผลลัพธ์ดำเนินการผลิตค่าผลลัพธ์ของพารามิเตอร์นำเข้าที่ไม่เป็นเชิงเส้นได้

สำหรับค่าน้ำหนักของโหนดซ่อนนั้นคำนวณได้ยากเทคนิคแบคพรอพพาเกชั่น (back propagation) จึงถูกสร้างขึ้นมาเพื่อแก้ปัญหานี้ โดยที่เทคนิคแบคพรอพพาเกชั่น (back propagation) จะมีสองช่วงในแต่ละการทำซ้ำคือ forward phase และ backward phase ใน forward phase ค่าน้ำหนักซึ่งได้มาจากการทำซ้ำในครั้งก่อนหน้าจะถูกใช้คำนวณค่าผลลัพธ์ การคำนวณจะดำเนินไปด้านหน้าเพียงอย่างเดียว ผลลัพธ์ของโครงข่ายในการทำซ้ำครั้งที่ k จะถูกคำนวณก่อนผลลัพธ์การทำซ้ำครั้งที่ k+1 ส่วน backward phase การปรับค่าน้ำหนักจะถูกนำมาใช้ในทิศทางย้อนกลับ กล่าวคือ ค่าน้ำหนักในครั้งที่ k+1 จะถูกปรับปรุงก่อนที่ค่าน้ำหนักในครั้งที่ k จะถูกปรับปรุง วิธีการ back-propagation ใช้ข้อผิดพลาดของโครงข่ายในครั้งที่ k+1 คาดการณ์ข้อผิดพลาดของโครงข่ายในครั้งที่ k [6]

1.3 การทดสอบไคสแควร์ (Chi-square test) เป็นวิธีการทดสอบเพื่อเปรียบเทียบข้อมูลที่อยู่ในรูปของความถี่หรือในรูปของสัดส่วน ตัวอย่าง เช่น การศึกษาเจตคติความคิดเห็น ความสนใจ หรือการยอมรับเป็นต้น ซึ่งไม่สามารถวัดค่าออกมาเป็นตัวเลขที่

แน่นอนได้ แต่สามารถจำแนกออกเป็นหมวดหมู่ได้ เช่น มากที่สุด มาก ปานกลาง น้อย น้อยที่สุด หรือดี ไม่ดี ถ้าหากต้องการเปรียบเทียบตัวแปร 2 กลุ่ม หรือมากกว่า 2 กลุ่ม ว่ามีความสัมพันธ์กันหรือไม่ การทดสอบไคสแควร์จะเหมาะสมกว่าการทดสอบแบบ Z เนื่องจากการทดสอบแบบ Z เหมาะสำหรับการทดสอบสัดส่วนของประชากรเพียงกลุ่มเดียวหรือการทดสอบความแตกต่างระหว่างสัดส่วนของสิ่งที่สนใจจากประชากร 2 กลุ่ม เท่านั้น การทดสอบไคสแควร์จึงนิยมใช้มากในการเปรียบเทียบหรือทดสอบข้อมูลที่เป็นความถี่หรือข้อมูลที่อยู่ในรูปของสัดส่วน



รูปที่ 3 ชนิดของฟังก์ชันกระตุ้นในโครงข่ายประสาทเทียม

1.3.1 วัตถุประสงค์ของการทดสอบไคส-แควร์ มีวัตถุประสงค์สำคัญ 3 ประการ คือ (1) การทดสอบความกลมกลืน (the goodness of fit test) เป็นการทดสอบไคสแควร์เพื่อศึกษาว่าการแจกแจงความถี่ของตัวแปรเป็นไปตามรูปแบบที่กำหนดไว้หรือไม่ โดยศึกษาจากตัวแปรเพียงตัวเดียวด้วยการเปรียบเทียบระหว่างข้อมูลจากตัวแปรกับข้อมูลที่ได้จากความคาดหมายหรือจากทฤษฎีใด ๆ ว่ามีความสอดคล้องกันหรือไม่ (2) การทดสอบความสัมพันธ์ระหว่างตัวแปร (test of association) หรือเรียกอีกอย่างหนึ่งว่าการทดสอบความเป็นอิสระ (test of independence) เป็นการทดสอบไคสแควร์เพื่อศึกษาว่าตัวแปรต่าง ๆ สัมพันธ์กันหรือไม่โดยการศึกษาความสัมพันธ์ระหว่างตัวแปรทีละคู่ ซึ่งตัวแปรแต่ละตัว

อาจจำแนกออกเป็นหลายกลุ่มหรือหลายพวกที่แจกแจงอยู่ในตารางมิติต่าง ๆ เช่น 2x2, 3x2 หรือ 2x3 เมื่อต้องการทดสอบความสัมพันธ์ระหว่างตัวแปรทีละคู่ จะต้องนำข้อมูลมาใส่ในตารางเพื่อหาความสัมพันธ์ระหว่างตัวแปรทั้งสอง สำหรับการทดสอบสมมติฐานว่าตัวแปรแต่ละคู่จะมีความสัมพันธ์กันหรือไม่ มีหลักการทดสอบความสัมพันธ์ระหว่างตัวแปรเพื่อให้สามารถหาค่าที่คาดหมายได้โดยการกำหนดสมมติฐานเป็นกลางว่าจะไม่มีความสัมพันธ์ระหว่างตัวแปรทั้งสอง (3) การทดสอบความเป็นเอกภาพ (test of homogeneity)การทดสอบความเป็นเอกภาพ หรือเรียกว่าการทดสอบความเป็นเอกพันธ์หรือการทดสอบความคล้ายคลึงกันของตัวแปร (test of homogeneity) เป็นการทดสอบความเหมือนกัน (หรือไม่แตกต่างกัน) ของตัวแปร โดยพิจารณาจากความน่าจะเป็นหรืออัตราส่วนของตัวแปรทั้งสอง ถ้ามีค่าใกล้เคียงกันแสดงว่าตัวแปรมีความเหมือนกัน [11]

1.3.2 วิธีของเพียร์สัน การวัดความสัมพันธ์ระหว่างตัวแปรด้วยวิธีของเพียร์สัน มีสูตรในการคำนวณ คือ $C = \sqrt{\frac{\chi^2}{\chi^2+n}}$ เมื่อ C = ค่าสัมประสิทธิ์ความสัมพันธ์ (มีค่าไม่เกิน 1) และ N = จำนวนสมาชิก (ซึ่งจะต้องมีค่ามากกว่า 0)

จากสูตร ถ้า C = 0 แสดงว่าไม่มีความสัมพันธ์ระหว่างตัวแปรทั้งสอง และถ้า C ยิ่งมีค่ามากแสดงว่าระดับความสัมพันธ์ยิ่งมีค่ามากโดยที่ค่า C สูงสุดสามารถหาได้จากสูตร $C_{max} = \sqrt{\left(\frac{k-1}{k}\right)}$ เมื่อเมื่อ C_{max} ค่าสูงสุดของสัมประสิทธิ์ของความสัมพันธ์ และ k = จำนวนของแถวหรือคอลัมที่มีค่าน้อยที่สุด

2 การทบทวนวรรณกรรมที่เกี่ยวข้อง

วราชัย [7] ได้ศึกษาถึงการวิเคราะห์อาการเสียของฮาร์ดดิสก์ โดยการใช้เทคนิคของการทำเหมืองข้อมูลมาใช้คัดแยกรูปแบบของเสีย เพื่อช่วยหาสาเหตุของอาการเสียที่เกิดขึ้นในกระบวนการผลิต ในการพิจารณาลักษณะของตำแหน่งที่เกิดจุดเสียนั้น จะพิจารณาโดยอาศัยเทคนิคการทำเหมืองข้อมูลเพื่อช่วยเพิ่มความสามารถที่จะใช้ช่วยวิเคราะห์หารูปแบบของเสียที่มาจากลักษณะการกระจายตัวของเสียแบบต่างๆ ซึ่งสามารถบอกถึงแนวโน้มของสาเหตุของอาการเสียในลักษณะต่างๆ จากการทดสอบปรากฏว่าเครื่องมือที่พัฒนาขึ้นสามารถทำการคัดแยกข้อมูลและมีความรวดเร็วในการวิเคราะห์สูง โดยการเปรียบเทียบกับการคัดแยกโดยผู้เชี่ยวชาญ ประสิทธิ์ชัย [8] เสนอการสร้างแบบจำลองการทำนายเปอร์เซ็นต์ผลผลิตฮาร์ดดิสก์โดยใช้ทฤษฎีโครงข่ายประสาทเทียมซึ่งสามารถใช้กับข้อมูลที่มีความซับซ้อนสูงและมีความสัมพันธ์ไม่เป็นเชิงเส้น โดยใช้ขั้นตอนวิธีในการเรียนรู้แบบ back-propagation ผลการทำนายที่ได้มีความคลาดเคลื่อนน้อยมาก ซึ่งสามารถสรุปได้ว่าตัวแบบการทำนายที่สร้างโดยวิธีโครงข่ายประสาทเทียมเหมาะสมที่จะใช้เป็นระบบจำลองการทำนายในอุตสาหกรรมฮาร์ดดิสก์ สิริยาภรณ์ [9] ได้ศึกษาการทำนายคุณภาพของผลิตภัณฑ์กาแฟคั่วและบดด้วยการสร้างแบบจำลองเพื่อทำนายคุณภาพเบื้องต้นก่อนการทดสอบโดยผู้เชี่ยวชาญในห้องปฏิบัติการ ซึ่งจะช่วยลดการสูญเสียทรัพยากร ได้แก่ เวลาและแรงงานที่ใช้ในระหว่างการทดสอบในห้องปฏิบัติการโดยผู้เชี่ยวชาญ โดยการตัดการทดลองใช้ปัจจัยในการผลิตที่แบบจำลองทำนายคุณภาพเบื้องต้นของผลิตภัณฑ์ทำนายว่า น่าจะให้ผลิตภัณฑ์สุดท้ายที่ด้อยคุณภาพออกจากการทดลองผลิต แบบจำลองที่นำมาใช้ทำนายคุณภาพของผลิตภัณฑ์กาแฟคั่วบดที่พิจารณามี 4 แบบ ได้แก่ แบบจำลองการถดถอยแบบพหุ แบบจำลองการถดถอยแบบโพลิโนเมียล แบบจำลองโครงข่ายประสาทเทียม และแบบจำลองโครงข่ายประสาทเทียมที่มีการนำเทคนิคพื้นผิวตอบสนองมาช่วยหาโครงสร้างที่

เหมาะสม พบว่าแบบจำลองโครงข่ายประสาทเทียมมีความสามารถในการทำนายค่าคะแนนคุณลักษณะทางประสาทสัมผัสของกาแฟคั่วและบดได้ถูกต้อง และใกล้เคียงกับค่าจริงมากกว่าแบบจำลองการถดถอยแบบพหุและแบบจำลองการถดถอยแบบโพลิโนเมียล และการสร้างแบบจำลองโครงข่ายประสาทเทียมด้วยเทคนิคพื้นผิวตอบสนอง สามารถช่วยลดระยะเวลาในการหาโครงสร้างที่เหมาะสมของแบบจำลองโครงข่ายประสาทเทียมลงได้

3 วิธีการดำเนินงานวิจัย

3.1 ศึกษาข้อมูลเบื้องต้น

ตรวจสอบอาการผิดปกติของฮาร์ดดิสก์ที่เกิดขึ้นในกระบวนการทดสอบความเชื่อมั่นก่อนการส่งมอบแก่ผู้บริโภค โดยตรวจสอบข้อมูลทั้งหมดจำนวน 8 เดือน พบว่า ลักษณะอาการผิดปกติของฮาร์ดดิสก์ในหมวดหมู่ชุดหัวอ่าน/เขียน (HSA) มีจำนวนมากที่สุด ดังนั้นงานวิจัยนี้จึงมุ่งเน้นการสร้างแบบจำลองการทำนายผลการทดสอบฮาร์ดดิสก์ของลักษณะอาการผิดปกติที่ชุดหัวอ่าน/เขียน โดยแบ่งชุดข้อมูลออกเป็นชุดการเรียนรู้จำนวน 60 เปอร์เซ็นต์ ชุดการทวนสอบจำนวน 20 เปอร์เซ็นต์ และชุดการทดสอบจำนวน 20 เปอร์เซ็นต์ ซึ่งข้อมูลในชุดการเรียนรู้ใช้สร้างแบบจำลอง ชุดข้อมูลการทวนสอบใช้เพื่อประเมินความถูกต้อง ปรับปรุง หรือเลือกแบบจำลองที่ดีที่สุด และชุดการทดสอบใช้ในการวัดความถูกต้องของแบบจำลองในตอนสุดท้ายเพื่อให้แน่ใจว่าแบบจำลองที่สร้างขึ้นมีประสิทธิภาพในการจำแนกหรือไม่ โดยแต่ละชุดข้อมูลจะมีสัดส่วนระหว่างฮาร์ดดิสก์ปกติและฮาร์ดดิสก์ผิดปกติเท่ากันดังแสดงในตารางที่ 1 แสดงจำนวนสัดส่วนจำนวนข้อมูลของแต่ละชุดข้อมูล

3.2 การเตรียมข้อมูล

คัดเลือกข้อมูลที่เกี่ยวข้อง ตรวจสอบความสมบูรณ์ของข้อมูล รวมข้อมูลจากหลายแหล่งข้อมูล และเปลี่ยนรูปแบบข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกับเทคนิคการสร้างแบบจำลองที่เลือกใช้

3.3 สร้างตัวแบบด้วยโปรแกรม SAS Enterprise Miner Workstation 13.2

โดยใช้เทคนิคการวิเคราะห์การถดถอยโลจิสติกส์และโครงข่ายประสาทเทียม โดยใช้ข้อมูลตั้งแต่เดือนเมษายนถึงเดือนพฤศจิกายนปี ค.ศ.2015 (จำนวน 8 เดือน)

ตารางที่ 1 สัดส่วนจำนวนข้อมูลของชุดข้อมูลการเรียนรู้ ทวนสอบ และทดสอบ

ชุดข้อมูล (Data Set)	ฮาร์ดดิสก์ผิดปกติ	ฮาร์ดดิสก์ปกติ	รวม	อัตราส่วนของฮาร์ดดิสก์ผิดปกติ
เรียนรู้ (Train)	445	100314	100759	0.44%
ทวนสอบ (Validate)	148	33438	33586	0.44%
ทดสอบ (Test)	149	33440	33589	0.44%
รวม Total	**742**	**167192**	**167934**	**0.44%**

3.3.1 ศึกษาปัจจัยที่ส่งผลกระทบกับคุณภาพฮาร์ดดิสก์ ในกระบวนการทดสอบชุดหัวอ่าน/เขียนและการทดสอบฮาร์ดดิสก์หลังการประกอบ คัดเลือกตัวแปรอิสระที่ส่งผลกระทบหรือมีความสัมพันธ์กับฮาร์ดดิสก์ที่มีอาการผิดปกติที่หัวอ่าน/เขียนขั้นต้นก่อนการสร้างโมเดล โดยใช้วิธีการวิเคราะห์ Chi-square กำจัดปัจจัยหรือตัวแปรอิสระที่มีจำนวนมากและอาจจะส่งผลที่ทำให้เกิดปัญหา multicollinearity

คือ ปัญหาที่ตัวแปรอิสระมีความสัมพันธ์กันเองเกินกว่าระดับที่ยอมรับได้ ในงานวิจัยนี้ ถ้าค่า Pearson correlation มีค่ามากกว่า 0.8 ถือว่าตัวแปรอิสระคู่นั้นมีความสัมพันธ์กันมาก เนื่องจากตัวแปรอิสระของงานวิจัยนี้มีจำนวนมาก จึงแก้ไขโดยการตัดตัวแปรอิสระตัวแปรใดตัวแปรหนึ่งออก โดยจะพิจารณาจากความรู้ทางเทคนิคของผลิตภัณฑ์ว่าตัวแปรใดมีความสำคัญมากกวาก็เลือกตัวแปรนั้นไว้ และตัดตัวแปรอิสระอีกตัวทิ้งไป เพราะถ้าตัวแปรอิสระมีความสัมพันธ์กันเองมากเกินไปจะละเมิดสมมติฐานขอที่วาตัวแปรอิสระตองไมมีความสัมพันธเชิงเสนตอกัน การที่เกิดปญหานี้ทำให้คาพารามิเตอรของตัวแปรอิสระเกิดการผิดพลาดและไมมีนัยสำคัญทางสถิติ ทำให้ขอสรุปของตัวแปรตามที่เกิดจากตัวแปรอิสระตัวนั้น ๆ เกิดความผิดพลาดได ตัวแปรตามในที่นี้คือฮาร์ดดิสก์ที่มีอาการผิดปกติกำหนดให้เท่ากับ 1 และฮาร์ดดิสก์ที่ไม่มีอาการผิดปกติกำหนดให้เท่ากับ 0

3.3.2 สร้างแบบจำลองการวิเคราะห์การถดถอยโลจิสติกส์และโครงข่ายประสาทเทียมด้วยโปรแกรม SAS Enterprise Miner Workstation 13.2

3.4 ประเมินความถูกต้องของแบบจำลองที่ได้จากการวิเคราะห์การถดถอยแบบโลจิสติกส์และโครงข่ายประสาทเทียม

โดยนำมาเปรียบเทียบกันทั้งผลการทดลองของชุดข้อมูลการเรียนรู้ ชุดการทดสอบ และชุดการทวนสอบ เพื่อคัดเลือกตัวแบบที่ให้ผลการพยากรณ์ที่ดีที่สุด ซึ่งสามารถพิจารณาได้จากค่า confusion matrix เป็นการเก็บข้อมูลที่เกี่ยวกับการแบ่งแยกข้อมูลจริง กับข้อมูลที่เกิดจากการพยากรณ์

4. ผลการวิจัยและสรุป

งานวิจัยนี้ใช้เทคนิคการวิเคราะห์เหมืองข้อมูล ซึ่งสามารถใช้กับข้อมูลที่มีขนาดใหญ่ (big data) โดยการสร้างตัวแบบในการพยากรณ์ คือ การวิเคราะห์การถดถอยโลจิสติกส์และโครงข่ายประสาทเทียม ทดสอบความสัมพันธ์ระหว่างตัวแปรอิสระกับคุณภาพของฮาร์ดดิสก์ พบว่ามีตัวแปรอิสระจำนวน 25 ตัวแปร ที่มีความสัมพันธ์กับคุณภาพของฮาร์ดดิสก์ และจากเงื่อนไขของการวิเคราะห์การถดถอยโลจิสติกส์จะต้องกำจัดตัวแปรอิสระที่มีความสัมพันธ์กันเองก่อน ดังนั้นจึงคงเหลือตัวแปรอิสระที่นำไปสร้างแบบจำลองการวิเคราะห์การถดถอยโลจิสติกส์จำนวน 19 ตัวแปร ดังตารางที่ 2 แสดงจำนวนตัวแปรที่นำไปสร้างตัวแบบในการพยากรณ์คุณภาพของฮาร์ดดิสก์แต่ละวิธี โดยเครื่องหมายถูกหมายถึงตัวแปรที่นำไปสร้างแบบตัวแบบ เครื่องหมายผิดหมายถึงตัวแปรที่ไม่ได้นำไปสร้างตัวแบบ

การวิเคราะห์การถดถอยโลจิสติกส์พบว่า มีตัวแปรอิสระที่ส่งผลกระทบกับคุณภาพของผลิตภัณฑ์ฮาร์ดดิสก์คือ X_12, X_91, X_94, X_97, X_110, X_119, X_124, X_133 และ X_168 รวมทั้งสิ้น 9 ตัวแปร จากตัวแปรอิสระที่ใช้สร้างตัวแบบทั้งหมด 19 ตัวแปร โดยได้ทำการตรวจสอบปัจจัยดังกล่าวกับผู้เชี่ยวชาญด้านเทคนิคของผลิตภัณฑ์เป็นที่เรียบร้อยแล้วว่า ปัจจัยเหล่านี้มีความสำคัญกับคุณภาพของฮาร์ดดิสก์จริง และเมื่อพิจารณาจากค่าการทดสอบทางสถิติ P-value ของการวิเคราะห์การถดถอยโลจิสติกส์แบบเป็นขั้นตอน (stepwise) จะเห็นได้ว่าตัวแปรทั้ง 9 ตัวแปรมีค่า P-value น้อยกว่า 0.05 แสดงว่าตัวแปรเหล่านี้มีผลกระทบหรือมีความสัมพันธ์กับคุณภาพของฮาร์ดดิสก์อย่างมีนัยสำคัญดังตารางที่ 3

ค่าสัมประสิทธิ์ที่ได้สามารถเขียนสมการการวิเคราะห์การถดถอยโลจิสติกส์เพื่อพยากรณ์ฮาร์ดดิสก์ผิดปกติได้ดังนี้

$$P(\text{ฮาร์ดดิสก์ผิดปกติ}) = \frac{1}{1+e^{-f(x)}}$$

โดย f(x) = -0.05570X_12 - 0.00958X_91 + 0.00908X_94 - 0.57620X_97 + 0.13360X_110 - 0.16050X_119 + 0.07250X_124 - 0.20050X_133 - 0.14310X_168 เมื่อ f(x) คือ ฟังก์ชันเชิงเส้นของตัวแปรอิสระ P (ฮาร์ดดิสก์ผิดปกติ) คือ ความน่าจะเป็นของการเกิดเหตุการณ์หรือความน่าจะเป็นของฮาร์ดดิสก์ผิดปกติ e คือ exponential function = 2.71828 และ X คือ ตัวแปรอิสระ

ตารางที่ 2 ตัวแปรอิสระที่นำไปสร้างตัวแบบในการพยากรณ์คุณภาพของฮาร์ดดิสก์แต่ละวิธี

ตัวแปรอิสระ	ความหมายของตัวแปรอิสระ	หน่วย	ตัวแปรอิสระที่ใช้สร้างแบบจำลอง	
			การวิเคราะห์การถดถอยโลจิสติกส์	โครงข่ายประสาทเทียม
X_4	อัตราความถูกต้องเทียบกับข้อผิดพลาดหลังการเขียนสัญญาณ	เดซิเบล (dB)	✓	✓
X_{12}	การเขียนข้อมูลซ้ำ ครั้งที่1	เดซิเบล (dB)	✓	✓
X_{15}	การเขียนข้อมูลซ้ำ ครั้งที่2	เดซิเบล (dB)	✗	✓
X_{31}	ค่าต่ำสุดของการเขียนข้อมูลซ้ำ ครั้งที่ 1 พื้นที่ A	เดซิเบล (dB)	✗	✓
X_{34}	ค่าต่ำสุดของการเขียนข้อมูลซ้ำ ครั้งที่ 2 พื้นที่ A	เดซิเบล (dB)	✗	✓
X_{48}	ค่าการเคลื่อนย้ายหัวอ่าน ทดสอบครั้งที่ 1	พิโคมิเตอร์ต่อโวลต์ (Picometers per volt)	✓	✓
X_{49}	ค่าการเคลื่อนย้ายหัวอ่าน ทดสอบครั้งที่ 2	พิโคมิเตอร์ต่อโวลต์ (Picometers per volt)	✗	✓
X_{50}	ค่าการเคลื่อนย้ายหัวอ่าน ทดสอบครั้งที่ 3	พิโคมิเตอร์ต่อโวลต์ (Picometers per volt)	✗	✓
X_{91}	การสั่นของหัวอ่าน แบบที่2 การทดสอบครั้งที่ 1 พื้นที่ A	เดซิเบล (dB)	✓	✓
X_{93}	การสั่นของหัวอ่าน แบบที่2 การทดสอบครั้งที่ 1 พื้นที่ C	เดซิเบล (dB)	✓	✓
X_{94}	การสั่นของหัวอ่าน แบบที่2 การทดสอบครั้งที่ 2 พื้นที่ A	เดซิเบล (dB)	✓	✓
X_{97}	การสั่นของหัวอ่าน แบบที่1 การทดสอบครั้งที่ 1 พื้นที่ A	เดซิเบล (dB)	✓	✓
X_{100}	การสั่นของหัวอ่าน แบบที่1 การทดสอบครั้งที่ 2 พื้นที่ A	เดซิเบล (dB)	✗	✓
X_{104}	แรงต้านการบินหัวอ่านระหว่างค่าเดิม (Original) และค่าที่ทำให้ปกติ (Normalize) พื้นที่ด้านใน ทดสอบครั้งที่ 1 พื้นที่ A	เมกะวัตต์ (MW)	✓	✓
X_{110}	แรงต้านการบินหัวอ่านระหว่างค่าเดิม (Original) และค่าที่ทำให้ปกติ (Normalize) พื้นที่ด้านใน ทดสอบครั้งที่ 2 พื้นที่ B	เมกะวัตต์ (MW)	✓	✓
X_{114}	แรงต้านการบินหัวอ่านระหว่างค่าเดิม (Original) และค่าที่ทำให้ปกติ (Normalize) พื้นที่ด้านใน ทดสอบครั้งที่ 3 พื้นที่ A	เมกะวัตต์ (MW)	✓	✓
X_{119}	แรงต้านการบินหัวอ่านระหว่างค่าเดิม (Original) และค่าที่ทำให้ปกติ (Normalize) พื้นที่ด้านนอก ทดสอบครั้งที่ 1 พื้นที่ A	เมกะวัตต์ (MW)	✓	✓
X_{120}	แรงต้านการบินหัวอ่านระหว่างค่าเดิม (Original) และค่าที่ทำให้ปกติ (Normalize) พื้นที่ด้านนอก ทดสอบครั้งที่ 1 พื้นที่ B	เมกะวัตต์ (MW)	✓	✓
X_{121}	แรงต้านการบินหัวอ่านระหว่างค่าเดิม (Original) และค่าที่ทำให้ปกติ (Normalize) พื้นที่ด้านนอก ทดสอบครั้งที่ 1 พื้นที่ C	เมกะวัตต์ (MW)	✓	✓
X_{122}	แรงต้านการบินหัวอ่านระหว่างค่าเดิม (Original) และค่าที่ทำให้ปกติ (Normalize) พื้นที่ด้านนอก ทดสอบครั้งที่ 1 พื้นที่ D	เมกะวัตต์ (MW)	✓	✓
X_{124}	แรงต้านการบินหัวอ่านระหว่างค่าเดิม (Original) และค่าที่ทำให้ปกติ (Normalize) พื้นที่ด้านนอก ทดสอบครั้งที่ 2 พื้นที่ A	เมกะวัตต์ (MW)	✓	✓
X_{129}	แรงต้านการบินหัวอ่านระหว่างค่าเดิม (Original) และค่าที่ทำให้ปกติ (Normalize) พื้นที่ด้านนอก ทดสอบครั้งที่ 3 พื้นที่ A	เมกะวัตต์ (MW)	✓	✓
X_{133}	แรงต้านการบินหัวอ่านระหว่างค่าเดิม (Original) และค่าที่ทำให้ปกติ (Normalize) พื้นที่ด้านนอก ทดสอบครั้งที่ 3 พื้นที่ E	เมกะวัตต์ (MW)	✓	✓
X_{157}	ค่าความต้านทานของเซ็นเซอร์หัวอ่าน	เมกะวัตต์ (MW)	✓	✓
X_{168}	ขนาดความผิดพลาดหลังจากการเขียนสัญญาณครั้งที่ 2 พื้นที่ A	เดซิเบล (dB)	✓	✓

ตารางที่ 3 ค่าพารามิเตอร์ที่ประมาณได้จากการวิเคราะห์การถดถอยโลจิสติกส์

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
X_{12}	1	-0.05570	0.02270	6.02	0.0141	-0.0608	0.946
X_{91}	1	-0.00958	0.00243	15.60	<.0001	-0.1320	0.990
X_{94}	1	0.00908	0.00339	7.18	0.0074	0.0968	1.009
X_{97}	1	-0.57620	0.03630	252.17	<.0001	-0.4365	0.562
X_{110}	1	0.13360	0.02930	20.77	<.0001	0.1087	1.143
X_{119}	1	-0.16050	0.04570	12.34	0.0004	-0.0920	0.852
X_{124}	1	0.07250	0.02010	12.94	0.0003	0.0797	1.075
X_{133}	1	-0.20050	0.03720	29.10	<.0001	-0.1161	0.818
X_{168}	1	-0.14310	0.02590	30.55	<.0001	-0.0945	0.867

การวิเคราะห์โครงข่ายประสาทเทียมแบบหลายชั้น (multilayer artificial neural network) โดยกำหนดให้มีชั้นอินพุตจำนวน 25 โหนด หรือเท่ากับจำนวนตัวแปรอิสระ กำหนดชั้นซ่อน 1 ชั้น จำนวน 5 หน่วยซ่อน และชั้นเอาพุตจำนวน 1 โหนด ที่อัตราการเรียนรู้เท่ากับ 0.1 ใช้เทคนิค back-propagation

งานวิจัยนี้ใช้ค่าความถูกต้องของการพยากรณ์ (accuracy) ในการเปรียบเทียบประสิทธิภาพของทั้ง 2 วิธี โดยใช้สูตร $Accurary = \frac{TN+TP}{FN+TN+FP+TP}$ โดยที่ true positive (TP) คือ ผลการพยากรณ์ว่าจะเกิดเหตุการณ์และผลลัพธ์บอกว่าเกิดเหตุการณ์ true negative (TN) คือ ผลการพยากรณ์ว่าจะไม่เกิดเหตุการณ์และผลลัพธ์บอกว่าไม่เกิดเหตุการณ์ false positive (FP) คือ ผลการพยากรณ์ว่าจะเกิดเหตุการณ์และผลลัพธ์บอกว่าไม่เกิดเหตุการณ์ false negative (FN) คือ ผลการพยากรณ์ว่าจะไม่เกิดเหตุการณ์และผลลัพธ์บอกว่าเกิดเหตุการณ์ true positive rate (TPR) คือ ร้อยละของการพยากรณ์การเกิดเหตุการณ์ที่ถูกต้อง (TP/TP+FN) และ true negative rate (TNR) คือ ร้อยละของการพยากรณ์การไม่เกิดเหตุการณ์ที่ถูกต้อง (TN/TN+FP) [10]

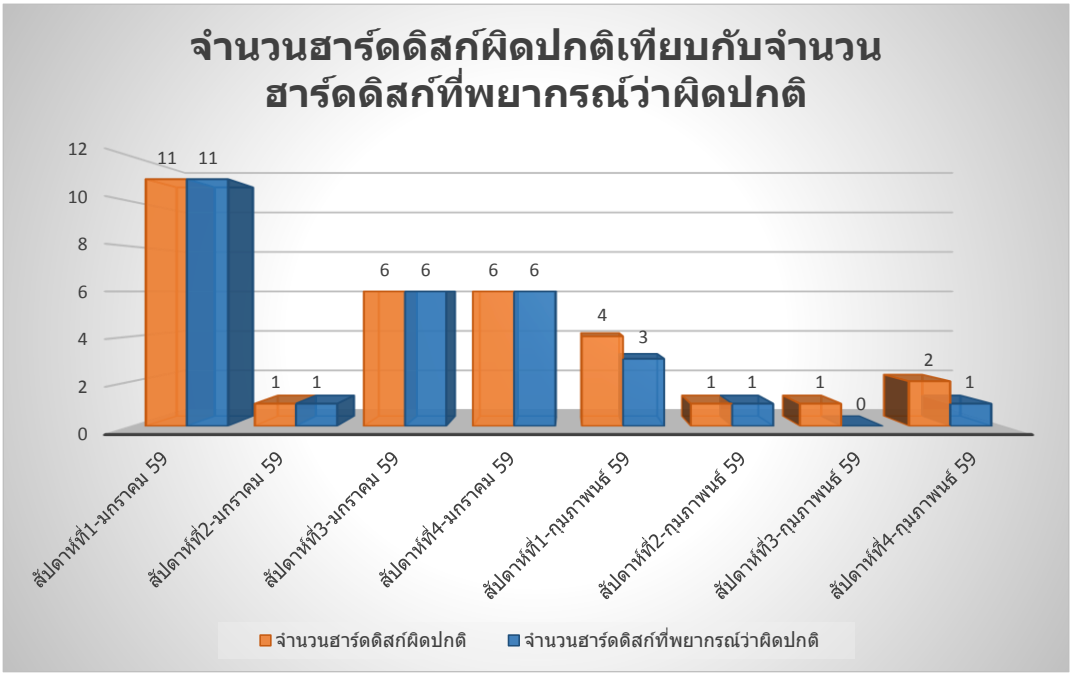
ตารางที่ 4ก ผลการวิเคราะห์การถดถอยโลจิสติกส์

การวิเคราะห์การถดถอยโลจิสติกส์							
ชุดข้อมูล Data Set	ผลการพยากรณ์(Prediction Result)				เกณฑ์การวัด (Measurements)		
	False Negative	True Negative	False Positive	True Positive	TPR	TNR	Accuracy
เรียนรู้ (Train)	113	73549	26765	332	74.61%	73.32%	73.32%
ทวนสอบ (Validation)	43	24528	8910	105	70.95%	73.35%	73.34%
ทดสอบ (Test)	43	24516	8924	106	71.14%	73.31%	73.30%
รวม Total	199	122593	44599	543	73.18%	73.32%	73.32%

ตารางที่ 4ก พบว่าการวิเคราะห์การถดถอยโลจิสติกส์ให้ค่าความถูกต้องของการพยากรณ์เท่ากับ 73 % ส่วนการวิเคราะห์โครงข่ายประสาทเทียมให้ค่าความถูกต้องของการพยากรณ์เท่ากับ 92 % ซึ่งถ้าแยกเป็นแต่ละชุดการทดสอบทั้งสามชุดข้อมูล ค่าความถูกต้องของการวิเคราะห์โครงข่ายประสาทเทียมให้ค่าสูงกว่าการวิเคราะห์การถดถอยโลจิสติกส์ทั้งสามชุดข้อมูล จึงสรุปได้ว่าวิธีการวิเคราะห์โครงข่ายประสาทเทียมให้ผลการพยากรณ์ที่ดีกว่าการวิเคราะห์การถดถอยโลจิสติกส์ ดังแสดงในตารางที่ 4ก และ 4ข

ตารางที่ 4ข ผลการวิเคราะห์โครงข่ายประสาทเทียม

โครงข่ายประสาทเทียม							
ชุดข้อมูล	ผลการพยากรณ์(Prediction Result)				เกณฑ์การวัด (Measurements)		
Data Set	False Negative	True Negative	False Positive	True Positive	TPR	TNR	Accuracy
เรียนรู้ (Train)	4	92789	7525	441	99.10%	92.50%	92.53%
ทดสอบ (Validation)	6	30962	2476	142	95.95%	92.60%	92.61%
ทวนสอบ (Test)	6	30944	2496	143	95.97%	92.54%	92.55%
รวม Total	16	154695	12497	726	97.84%	92.53%	92.55%



รูปที่ 4 จำนวนฮาร์ดดิสก์ที่ผิดปกติเทียบกับจำนวนฮาร์ดดิสก์ที่พยากรณ์ว่าผิดปกติ

เนื่องจากวิธีโครงข่ายประสาทเทียมให้ค่าความถูกต้องสูงกว่าการวิเคราะห์การถดถอยโลจิสติกส์ จึงนำไปใช้ในการพยากรณ์คุณภาพของฮาร์ดดิสก์ก่อนกระบวนการทดสอบความเชื่อมั่น พบว่าจำนวนฮาร์ดดิสก์ผิดปกติมีค่าใกล้เคียงกับจำนวนฮาร์ดดิสก์ที่พยากรณ์ว่าผิดปกติ จะเห็นได้ว่าแบบจำลองสามารถ

พยากรณ์ฮาร์ดดิสก์ผิดปกติได้ถูกต้อง และสามารถป้องกันไม่ให้เกิดฮาร์ดดิสก์ผิดปกติก่อนกระบวนการทดสอบฮาร์ดดิสก์จะเสร็จสิ้นลงถึง 92 เปอร์เซ็นต์ ดังนั้นจึงสามารถสรุปได้ว่าแบบจำลองที่ได้จากการวิเคราะห์โครงข่ายประสาทเทียมมีความสามารถในการพยากรณ์คุณภาพของฮาร์ดดิสก์ที่เหมาะสมและสามารถนำมาใช้งานได้จริง (รูปที่ 4)

5. ข้อเสนอแนะ

แบบจำลองทั้งสองชนิดที่ได้ศึกษานั้นสามารถปรับเปลี่ยนและแก้ไขได้ เนื่องจากในแต่ละช่วงเวลาอาจมีการปรับปรุงกระบวนการผลิตหรือเปลี่ยนแปลงเทคโนโลยีในการผลิตทำให้ข้อมูลต่าง ๆ มีการเปลี่ยนแปลง ดังนั้นจึงต้องมีการเรียนรู้แบบจำลองหรือที่เรียกว่า retraining model แต่ถ้าหากมีการเปลี่ยนแปลงของข้อมูลมากเกินไปจะต้องสร้างแบบจำลองใหม่หรือที่เรียกว่า rebuilding model โดยยังคงใช้กระบวนการสร้างแบบจำลองเช่นเดิมได้ [1]

6. กิตติกรรมประกาศ

ขอขอบคุณ โรงงานกรณีศึกษา ที่ให้ข้อมูลในการวิจัยครั้งนี้

7. รายการอ้างอิง

[1] Tan, P.N., Steinbach, M. and Kumar, V., 2005, Introduction to Data Mining, Pearson International Edition, Inc., Addison Wesley.

[2] นำชัย ศุภฤกษ์ชัยสกุล, 2553, การวิเคราะห์ Logistic Regression, สถาบันวิจัยพฤติกรรมศาสตร์มหาวิทยาลัยศรีนครินทรวิโรฒ,กรุงเทพฯ.

[3] กัลยา วานิชย์บัญชา, 2546, การวิเคราะห์สถิติขั้นสูงด้วย SPSS for Windows, พิมพ์ครั้งที่ 3, ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ.

[4] Fausett, L. 1994, Fundamentals of Neural Networks: Architectures, Algorithms and Application, Prentice Hall, Inc., New Jersey.

[5] Krose, B. and van der Smagt, P., 1996, An Introduction to Neural Networks, 8th Ed., Amsterdam, Oberpfaffenhofen.

[6] Kumar, S., 2012, Neural Networks: A Classroom Approach, 2nd Ed., McGraw Hill Education, New Delhi.

[7] วราชัย สิงห์จังหรีด, 2550, การวิเคราะห์อาการเสียของฮาร์ดดิสก์โดยวิธีการเหมืองข้อมูล, วิทยานิพนธ์ปริญญาโท, บัณฑิตวิทยาลัย มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 73 น.

[8] ประสิทธิ์ชัย บุญเสริม, 2552, การสร้างโมเดลเพื่อทำนายจากเครือข่ายใยประสาทแบบสุ่มสำหรับการผลิตเฮชจีเอ, วิทยานิพนธ์ปริญญาโท, บัณฑิตวิทยาลัย มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี, 68 น.

[9] สิริยาภรณ์ ไกรมาก, 2552, แบบจำลองการทำนายเพื่อการประเมินคุณภาพของผลิตภัณฑ์กาแฟคั่วและบด, วิทยานิพนธ์ปริญญาโท, บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์, 154 น.

[10] Chen, C., Liaw, A. and Breiman, L., 2004, Using Random Forest to Learn Imbalanced Data, University of California, Report ID: 666.

[11] Campbell, I., 2007, Chi-squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations, Stat. Med. 26: 3661-3675.