รายงานวิจัยฉบับสมบูรณ์

โครงการ การพัฒนาตัวแบบทางสถิติเพื่อพยากรณ์ผลผลิตข้าวในเขตภาคเหนือตอนล่าง

โดย ผศ.ดร. อนามัย นาอุดม และคณะ

มีนาคม 2558

สัญญาเลขที่ R2557B062

รายงานวิจัยฉบับสมบูรณ์

โครงการ การพัฒนาตัวแบบทางสถิติเพื่อพยากรณ์ผลผลิตข้าวในเขตภาคเหนือตอนล่าง

คณะผู้วิจัย

1. ผศ.ดร. อนามัย นาอุดม สังกัด ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ ม.นเรศวร
2. ผศ.ดร. จรัสศรี รุ่งรัตนาอุบล สังกัด ภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ ม.นเรศวร

สนับสนุนโดยงบประมาณแผ่นดิน มหาวิทยาลัยนเรศวร

สารบัญ

บทคัดย่อ

งานวิจัยนี้มีจุดมุ่งหมายเพื่อศึกษาปัจจัยที่มีอิทธิพลต่อผลผลิตข้าวในเขตภาคเหนือตอนล่าง และสร้างตัวแบบเพื่อการพยากรณ์ผลผลิตข้าวในเขตภาคเหนือตอนล่างโดยใช้ตัวแบบที่ได้รับความนิยม 2 ตัวแบบ ได้แก่ ตัวแบบ Regression และ และตัวแบบโครงข่ายประสาทเทียม (Artificial neural network: ANN) ข้อมูลที่ใช้เป็นข้อมูลทุติยภูมิซึ่งเก็บรวบรวมโดยสำนักงานเศรษฐกิจการเกษตรในช่วงปี พ.ศ. 2551 - 2553 ทำการวิเคราะห์หาตัวแปรที่มีอิทธิพลต่อผลผลิตข้าวโดยใช้เทคนิคการวิเคราะห์สหสัมพันธ์ (Correlation analysis) จากนั้นนำตัวแปรที่มีนัยสำคัญทางสถิติไปพัฒนาตัวแบบเพื่อการพยากรณ์ โดยแบ่งชุดข้อมูลออกเป็น 2 ชุด คือข้อมูลฝึกสอนและข้อมูลทดสอบ โดยใช้อัตราส่วน 80 : 20 โดยใช้ข้อมูลฝึกสอนในการพัฒนาตัวแบบ ส่วนข้อมูลทดสอบจะถูกนำมาใช้ในการหาค่าความแม่นยำของการพยากรณ์ ผลจากการศึกษาพบว่าตัวแปรที่มีอิทธิพลต่อผลผลิตข้าวได้แก่ ประเภทพันธุ์ข้าว วิธีปลูก ปริมาณพันธุ์ข้าว การใช้ปุ๋ยเคมี ฤดูกาลเพาะปลูก พื้นที่ชลประทาน และเมื่อนำตัวแบบทั้ง 2 แบบมาเปรียบเทียบกันพบว่า ANN ให้ความแม่นยำในการพยากรณ์สูงกว่าตัวแบบ Regression ทั้งนี้เนื่องจากโครงสร้างของ ANN เป็นแบบไม่ใช่เชิงเส้น จึงสามารถสะกัดรูปแบบความสัมพันธ์ระหว่างตัวแปรดังกล่าวข้างต้นที่มีต่อผลผลิตข้าวได้ดีกว่าตัวแบบ Regression นั่นเอง

Abstract

This research aims to identify the factors that influence the production of rice in the lower northern Thailand and develop the models to predict rice yields in the lower northern Thailand by using two popular models namely Regression model and artificial neural network model. The data used in this project was collected by the Department of Agriculture in the period from 2008 – 2010. We first evaluate the variables that influence the production of rice using correlation analysis, and then the significant factors are brought to develop the predictive models. In order to construct the models, the dataset is divided into 2 groups: training set and test set respectively using the ratio 80:20. A training set is used to develop the model while the test set is used to evaluate the accuracy of the models. The results show that the variables that influence the rice production are rice-growing types of seeds, the use of chemical fertilizers, planting season, and irrigable area. The results obtained from the comparative study reveals that ANN provides higher accuracy than Regression model as ANN structure is non-linear and it can capture the pattern of relationships between input variables and rice yield better Regression model.

Executive Summary

ข้าวเป็นพืชเศรษฐกิจที่มีความสำคัญอันดับต้น ๆ ของประเทศไทย เนื่องจากประเทศไทยเป็นผู้นำในการส่งออกข้าวเป็นอันดับ 1 ของโลก ทั้งนี้ประเทศไทยมีศักยภาพในการส่งออกข้าวค่อนข้างสูง เพราะประเทศไทยมีผลผลิตข้าวที่มากพอต่อความต้องการข้าวของตลาดโลก (สำนักงานวิจัยเศรษฐกิจการเกษตร, 2553) การส่งออกข้าวไทยในปัจจุบันได้ขยายไปทุกภูมิภาคของโลก เช่น จีน อินโดนีเซีย อิหร่าน ฮ่องกง มาเลเซีย และสิงคโปร์ เป็นต้น โดยการส่งออกข้าวทำรายได้ให้แก่ประเทศปีละหลายหมื่นล้านบาท ซึ่งคิดเป็นสัดส่วนประมาณร้อยละ 20 ของมูลค่าสินค้าที่ส่งออกทั้งหมด (สำนักงานวิจัยเศรษฐกิจการเกษตร) จึงสามารถสรุปได้ว่าข้าวเป็นพืชเศรษฐกิจที่เป็นที่ต้องการของตลาดทั้งภายในประเทศและต่างประเทศ อย่างไรก็ตามการผลิตข้าวของประเทศไทยเมื่อเทียบต่อหน่วยพื้นที่ปลูกยังคงให้ผลผลิตที่ต่ำ โดยมีค่าเฉลี่ยประมาณ 432 กิโลกรัมต่อไร่ ซึ่งต่ำกว่าเกณฑ์เฉลี่ยของการผลิตข้าวในโลกที่ผลผลิตอยู่ในระดับ 636 กิโลกรัมต่อไร่ (ศูนย์เมล็ดพันธุ์ข้าว, 2550)

เกษตรกรส่วนใหญ่ในประเทศไทยมักจะปลูกข้าวเจ้าเป็นหลักโดยเฉพาะในเขตพื้นที่ภาคเหนือตอนล่างและภาคใต้ (สำนักงานพัฒนาการวิจัยการเกษตร) พื้นที่ปลูกข้าวในเขตภาคเหนือตอนล่างประกอบไปด้วย 8 จังหวัด ได้แก่ กำแพงเพชร พิจิตร พิษณุโลก สุโขทัย อุตรดิตถ์ นครสวรรค์ เพชรบูรณ์ และอุทัยธานี โดยพื้นที่ส่วนใหญ่เป็นที่ราบซึ่งเหมาะแก่การเพาะปลูกข้าว ดังนั้นภาคเหนือตอนล่าง จึงเป็นแหล่งผลิตข้าวและพืชไร่ เช่น อ้อย และข้าวโพด ที่สำคัญของประเทศ

การปลูกข้าวในภาคเหนือตอนล่างสามารถจำแนกเป็น 2 รูปแบบตามฤดูกาล คือข้าวนาปีและข้าวนาปรัง ปัญหาที่พบบ่อยในเกษตรกรที่ปลูกข้าวในเขตพื้นที่ภาคเหนือตอนล่างในปัจจุบัน เนื่องมาจากปัจจัยดังต่อไปนี้

1) ปัญหาด้านจำนวนพื้นที่เพาะปลูกข้าวน้อยลง เนื่องจากพื้นที่ในการเพาะปลูกข้าวถูกนำไปขาย หรือปรับพื้นที่ไปใช้ในงานด้านต่าง ๆ เช่น ทำธุรกิจ สร้างโรงงานอุตสาหกรรม หรือหมู่บ้านจัดสรร ทำให้พื้นที่ที่ใช้ในการเพาะปลูกข้าวมีจำนวนน้อยลง ทำให้เกษตรกรมีพื้นที่ในการเพาะปลูกข้าวไม่เพียงพอ

2) ปัญหาทางด้านผลผลิตต่อไร่ที่ค่อนข้างต่ำ ทำให้เกษตรกรบางส่วนหันมาใช้วิทยาการสมัยใหม่ เช่น การใช้ปุ๋ยเคมี ปุ๋ยอินทรีย์ ซึ่งมีค่าใช้จ่ายสูง ถ้าหากใช้ในปริมาณที่มากเกินไปก็จะเป็นการสิ้นเปลือง อีกทั้งยังมีผลกระทบในระยะยาวต่อความอุดมสมบูรณ์ของดิน มีผลทำให้ดินเสื่อมคุณภาพ

3) ปัญหาการขาดแหล่งข้อมูลหรือสารสนเทศที่สามารถสนับสนุนการเพาะปลูกข้าวให้ได้ผลผลิตที่สูงและเหมาะสมต่อพื้นที่เพาะปลูก เช่น ข้อมูลด้านพันธุ์ข้าว ข้อมูลด้านผลผลิต ข้อมูลด้านวิธีปลูก ข้อมูลด้านการใช้ปุ๋ย เป็นต้น

นอกจากนี้ยังพบว่าปัญหาที่เกิดขึ้นกับเกษตรกรอาจเนื่องจากการขาดอุปกรณ์ที่ทันสมัยในการเข้าถึงข้อมูลจากภาครัฐ และแหล่งที่รวบรวมองค์ความรู้ ถึงแม้ว่าเกษตรกรบางส่วนจะได้รับข้อมูลจากเจ้าหน้าที่ภาครัฐโดยตรง แต่ก็ยังขาดความเข้าใจในข้อมูลหรือองค์ความรู้ที่ชัดเจน เนื่องจากเมื่อเจ้าหน้าที่ภาครัฐเข้าไปถ่ายทอดหรือทำการเผยแพร่ข้อมูลให้กับเกษตรกรในท้องถิ่น เกษตรกรบางส่วนอาจตามไม่ทัน หรือเกษตรกรบางส่วนอาจไม่กล้าถามข้อสงสัย ทำให้ไม่ได้รับความรู้อย่างแท้จริง ดังนั้นเมื่อเกษตรกรต้องการค้นหาความรู้ เกษตรกรต้องเดินทางเพื่อติดต่อกับสำนักงานภาครัฐโดยตรง ซึ่งทำให้เสียเวลาและค่าใช้จ่าย นอกจากนี้ข้อมูลที่ทางสำนักงานภาครัฐเก็บรวบรวมไว้ยังไม่ครบถ้วนและขาดการจัดเก็บที่เป็นระบบระเบียบ รวมไปถึงขาดการจัดการเชิงรุก เช่น ขาดการศึกษาปัจจัยที่มีผลต่อการเพิ่มผลผลิตข้าว หรือตัวแบบเพื่อการพยากรณ์ผลผลิตข้าวโดยพิจารณาปัจจัยต่าง ๆ ที่เกี่ยวข้องไม่ว่าจะเป็น วิธีการปลูก พันธุ์ข้าว และลักษณะดินที่ทำการเพาะปลูก เป็นต้น

การนำเทคโนโลยีสมัยใหม่รวมไปถึงสารสนเทศที่สามารถมาถ่ายทอดองค์ความรู้เพื่อให้ข้อมูลแก่เกษตรกร โดยการนำเทคนิคการพยากรณ์แบบต่าง ๆ มาใช้ในการทำนายผลผลิตข้าวต่อไร่ที่เกษตรกรควรจะได้รับ

โดยพิจารณาปัจจัยต่าง ๆ ที่เกี่ยวข้องกับการเพาะปลูก เพื่อให้เกษตรกรทราบถึงตัวแปรที่สำคัญต่อผลผลิตข้าว จึงเป็นสิ่งที่จำเป็นอย่างยิ่งเพื่อช่วยให้เกษตรกรสามารถเพิ่มผลผลิตข้าวได้มากที่สุด ซึ่งในงานวิจัยนี้ผู้วิจัยได้นำเทคนิคทางสถิติและเทคนิคเหมืองข้อมูลซึ่งเป็นกระบวนการการกลั่นกรองสารสนเทศ (Information) ที่ซ่อนอยู่ในฐานข้อมูลขนาดใหญ่ (Knowledge Discovery in Database, KDD) และทำการสกัดรูปแบบความสัมพันธ์ของตัวแปรที่เกี่ยวข้องออกมาในรูปแบบทางคณิตศาสตร์ โดยเทคนิคนี้จะใช้ข้อมูลในอดีต เพื่อค้นหาความสัมพันธ์จนเป็นองค์ความรู้ที่สำคัญ และใช้สร้างตัวแบบพยากรณ์เพื่อทำนายสิ่งที่สนใจศึกษา ซึ่งผลที่ได้สามารถนำไปใช้ประกอบการตัดสินใจได้

งานวิจัยนี้ประยุกต์ใช้เทคนิคการพยากรณ์ทางสถิติและเทคนิคเหมืองข้อมูลเข้ามาใช้ในการพยากรณ์ผลผลิตข้าวในเขตภาคเหนือตอนล่าง โดยใช้วิธีโครงข่ายประสาทเทียม (Artificial Neural Network) และวิธีวิเคราะห์การถดถอย (Regression Analysis) โดยพิจารณาจากปัจจัยต่าง ๆ ที่เกี่ยวข้องกับผลผลิตข้าว ผู้วิจัยทำการสร้างตัวแบบพยากรณ์ผลผลิตข้าวและเลือกใช้ตัวแบบที่มีความแม่นยำมากที่สุดไปแนะนำให้บุคลากรทางการเกษตรได้ทราบ และจะขยายผลโดยการนำไปพัฒนาระบบสารสนเทศการสืบค้นข้อมูลและโปรแกรมการพยากรณ์ผลผลิตข้าวในรูปแบบเว็บแอพพลิเคชั่น ซึ่งระบบนี้จะเป็นประโยชน์ต่อเกษตรกร ให้ทราบถึงข้อมูลการพยากรณ์ได้แก่ ด้านพันธุ์ข้าว ด้านผลผลิต ด้านวิธีปลูก ด้านการใช้ปุ๋ย เป็นต้น นอกจากนี้หน่วยงานของภาครัฐสามารถนำผลที่ได้จากการวิจัยไปใช้เป็นแนวทางในการกำหนดนโยบายเพื่อส่งเสริมและสนับสนุนการเพาะปลูกข้าวต่อไป

บทที่ 1

บทนำ

รายงานนี้เป็นส่วนหนึ่งของโครงการวิจัยภายใต้การสนับสนุนโดยงบประมาณแผ่นดิน มหาวิทยาลัยนเรศวร ประจำปีงบประมาณ 2557 โดยคณะผู้วิจัยได้รับการอนุมัติให้ทำงานวิจัยนี้ โดยมีชื่อโครงการและรายละ**เอียด**เกี่ยวกับโครงการวิจัยดังต่อไปนี้

ชื่อโครงการ (ภาษาไทย) การพัฒนาตัวแบบทางสถิติเพื่อพยากรณ์ผลผลิตข้าวในเขตภาคเหนือตอนล่าง

(ภาษาอังกฤษ) Development of statistical models for predicting rice product in the lower northern part of Thailand

คณะผู้วิจัย(ระบุสังกัดภาควิชา) และสัดส่วนที่ทำงานวิจัย (%)

หัวหน้าโครงการวิจัย

ผศ.ดร. อนามัย นาอุดม (50%)
อาจารย์ สังกัดภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์
มหาวิทยาลัยนเรศวร

ผู้วิจัยร่วม ผศ. ดร. จรัสศรี รุ่งรัตนาอุบล (50%)
อาจารย์สังกัดภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ
คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร

สถานที่จัดทำโครงการวิจัย

ภาควิชาคณิตศาสตร์และภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ
คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร

ความสำคัญและที่มาของปัญหาที่ทำการวิจัย

ข้าวเป็นพืชเศรษฐกิจที่มีความสำคัญอันดับต้น ๆ ของประเทศไทย เนื่องจากประเทศไทยเป็นผู้นำในการส่งออกข้าวเป็นอันดับ 1 ของโลก ทั้งนี้ประเทศไทยมีศักยภาพในการส่งออกข้าวค่อนข้างสูง เพราะประเทศไทยมีผลผลิตข้าวที่มากพอต่อความต้องการข้าวของตลาดโลก (สำนักงานวิจัยเศรษฐกิจการเกษตร, 2553) การส่งออกข้าวไทยในปัจจุบันได้ขยายไปทุกภูมิภาคของโลก เช่น จีน อินโดนีเซีย อิหร่าน ฮ่องกง มาเลเซีย และสิงคโปร์ เป็นต้น โดยการส่งออกข้าวทำรายได้ให้แก่ประเทศปีละหลายหมื่นล้านบาท ซึ่งคิดเป็นสัดส่วนประมาณร้อยละ 20 ของมูลค่าสินค้าที่ส่งออกทั้งหมด (สำนักงานวิจัยเศรษฐกิจการเกษตร) จึงสามารถสรุปได้ว่าข้าวเป็นพืชเศรษฐกิจที่เป็นที่ต้องการของตลาดทั้งภายในประเทศและต่างประเทศ อย่างไรก็ตามการผลิตข้าวของประเทศไทยเมื่อเทียบต่อหน่วยพื้นที่ปลูกยังคงให้ผลผลิตที่ต่ำ โดยมีค่าเฉลี่ยประมาณ 432 กิโลกรัมต่อไร่ ซึ่งต่ำกว่าเกณฑ์เฉลี่ยของการผลิตข้าวในโลกที่ผลผลิตอยู่ในระดับ 636 กิโลกรัมต่อไร่ (ศูนย์เมล็ดพันธุ์ข้าว, 2550)

เกษตรกรส่วนใหญ่ในประเทศไทยมักจะปลูกข้าวเจ้าเป็นหลักโดยเฉพาะในเขตพื้นที่ภาคเหนือตอนล่างและภาคใต้ (สำนักงานพัฒนาการวิจัยการเกษตร) พื้นที่ปลูกข้าวในเขตภาคเหนือตอนล่างประกอบไปด้วย 8 จังหวัด ได้แก่ กำแพงเพชร พิจิตร พิษณุโลก สุโขทัย อุตรดิตถ์ นครสวรรค์ เพชรบูรณ์ และอุทัยธานี โดยพื้นที่ส่วนใหญ่เป็นที่ราบซึ่งเหมาะแก่การเพาะปลูกข้าว ดังนั้นภาคเหนือตอนล่าง จึงเป็นแหล่งผลิตข้าวและพืชไร่ เช่น อ้อย และข้าวโพด ที่สำคัญของประเทศ

การปลูกข้าวในภาคเหนือตอนล่างสามารถจำแนกเป็น 2 รูปแบบตามฤดูกาล คือข้าวนาปีและข้าวนาปรัง ปัญหาที่พบบ่อยในเกษตรกรที่ปลูกข้าวในเขตพื้นที่ภาคเหนือตอนล่างในปัจจุบัน เนื่องมาจากปัจจัยดังต่อไปนี้

1) ปัญหาด้านจำนวนพื้นที่เพาะปลูกข้าวน้อยลง เนื่องจากพื้นที่ในการเพาะปลูกข้าวถูกนำไปขาย หรือปรับพื้นที่ไปใช้ในงานด้านต่าง ๆ เช่น ทำธุรกิจ สร้างโรงงานอุตสาหกรรม หรือหมู่บ้านจัดสรร ทำให้พื้นที่ที่ใช้ในการเพาะปลูกข้าวมีจำนวนน้อยลง ทำให้เกษตรกรมีพื้นที่ในการเพาะปลูกข้าวไม่เพียงพอ

2) ปัญหาทางด้านผลผลิตต่อไร่ที่ค่อนข้างต่ำ ทำให้เกษตรกรบางส่วนหันมาใช้วิทยาการสมัยใหม่ เช่น การใช้ปุ๋ยเคมี ปุ๋ยอินทรีย์ ซึ่งมีค่าใช้จ่ายสูง ถ้าหากใช้ในปริมาณที่มากเกินไปก็จะเป็นการสิ้นเปลือง อีกทั้งยังมีผลกระทบในระยะยาวต่อความอุดมสมบูรณ์ของดิน มีผลทำให้ดินเสื่อมคุณภาพ

3) ปัญหาการขาดแหล่งข้อมูลหรือสารสนเทศที่สามารถสนับสนุนการเพาะปลูกข้าวให้ได้ผลผลิตที่สูงและเหมาะสมต่อพื้นที่เพาะปลูก เช่น ข้อมูลด้านพันธุ์ข้าว ข้อมูลด้านผลผลิต ข้อมูลด้านวิธีปลูก ข้อมูลด้านการใช้ปุ๋ย เป็นต้น

นอกจากนี้ยังพบว่าปัญหาที่เกิดขึ้นกับเกษตรกรอาจเนื่องจากการขาดอุปกรณ์ที่ทันสมัยในการเข้าถึงข้อมูลจากภาครัฐ และแหล่งที่รวบรวมองค์ความรู้ ถึงแม้ว่าเกษตรกรบางส่วนจะได้รับข้อมูลจากเจ้าหน้าที่ภาครัฐโดยตรง แต่ก็ยังขาดความเข้าใจในข้อมูลหรือองค์ความรู้ที่ชัดเจน เนื่องจากเมื่อเจ้าหน้าที่ภาครัฐเข้าไปถ่ายทอดหรือทำการเผยแพร่ข้อมูลให้กับเกษตรกรในท้องถิ่น เกษตรกรบางส่วนอาจตามไม่ทัน หรือเกษตรกรบางส่วนอาจไม่กล้าถามข้อสงสัย ทำให้ไม่ได้รับความรู้อย่างแท้จริง ดังนั้นเมื่อเกษตรกรต้องการค้นหาความรู้ เกษตรกรต้องเดินทางเพื่อติดต่อกับสำนักงานภาครัฐโดยตรง ซึ่งทำให้เสียเวลาและค่าใช้จ่าย นอกจากนี้ข้อมูลที่ทางสำนักงานภาครัฐเก็บรวบรวมไว้ยังไม่ครบถ้วนและขาดการจัดเก็บที่เป็นระบบระเบียบ รวมไปถึงขาดการจัดการเชิงรุก เช่น ขาดการศึกษาปัจจัยที่มีผลต่อการเพิ่มผลผลิตข้าว หรือตัวแบบเพื่อการพยากรณ์ผลผลิตข้าวโดยพิจารณาปัจจัยต่าง ๆ ที่เกี่ยวข้องไม่ว่าจะเป็น วิธีการปลูก พันธุ์ข้าว และลักษณะดินที่ทำการเพาะปลูก เป็นต้น

การนำเทคโนโลยีสมัยใหม่รวมไปถึงสารสนเทศที่สามารถมาถ่ายทอดองค์ความรู้เพื่อให้ข้อมูลแก่เกษตรกร โดยการนำเทคนิคการพยากรณ์แบบต่าง ๆ มาใช้ในการทำนายผลผลิตข้าวต่อไร่ที่เกษตรกรควรจะได้รับโดยพิจารณาปัจจัยต่าง ๆ ที่เกี่ยวข้องกับการเพาะปลูก เพื่อให้เกษตรกรทราบถึงตัวแปรที่สำคัญต่อผลผลิตข้าว จึง

เป็นสิ่งที่จำเป็นอย่างยิ่งเพื่อช่วยให้เกษตรกรสามารถเพิ่มผลผลิตข้าวได้มากที่สุด ซึ่งในงานวิจัยนี้ผู้วิจัยได้นำเทคนิคทางสถิติและเทคนิคเหมืองข้อมูลซึ่งเป็นกระบวนการการกลั่นกรองสารสนเทศ (Information) ที่ซ่อนอยู่ในฐานข้อมูลขนาดใหญ่ (Knowledge Discovery in Database, KDD) และทำการสกัดรูปแบบความสัมพันธ์ของตัวแปรที่เกี่ยวข้องออกมาในรูปแบบทางคณิตศาสตร์ โดยเทคนิคนี้จะใช้ข้อมูลในอดีต เพื่อค้นหาความสัมพันธ์จนเป็นองค์ความรู้ที่สำคัญ และใช้สร้างตัวแบบพยากรณ์เพื่อทำนายสิ่งที่สนใจศึกษา ซึ่งผลที่ได้สามารถนำไปใช้ประกอบการตัดสินใจได้

งานวิจัยนี้ประยุกต์ใช้เทคนิคการพยากรณ์ทางสถิติและเทคนิคเหมืองข้อมูลเข้ามาใช้ในการพยากรณ์ผลผลิตข้าวในเขตภาคเหนือตอนล่าง โดยใช้วิธีโครงข่ายประสาทเทียม (Artificial Neural Network) และวิธีวิเคราะห์การถดถอย (Regression Analysis) โดยพิจารณาจากปัจจัยต่าง ๆ ที่เกี่ยวข้องกับผลผลิตข้าว ผู้วิจัยทำการสร้างตัวแบบพยากรณ์ผลผลิตข้าวและเลือกใช้ตัวแบบที่มีความแม่นยำมากที่สุดไปแนะนำให้บุคลากรทางการเกษตรได้ทราบ และจะขยายผลโดยการนำไปพัฒนาระบบสารสนเทศการสืบค้นข้อมูลและโปรแกรมการพยากรณ์ผลผลิตข้าวในรูปแบบเว็บแอพพลิเคชั่น ซึ่งระบบนี้จะเป็นประโยชน์ต่อเกษตรกร ให้ทราบถึงข้อมูลการพยากรณ์ได้แก่ ด้านพันธุ์ข้าว ด้านผลผลิต ด้านวิธีปลูก ด้านการใช้ปุ๋ย เป็นต้น นอกจากนี้หน่วยงานของภาครัฐสามารถนำผลที่ได้จากการวิจัยไปใช้เป็นแนวทางในการกำหนดนโยบายเพื่อส่งเสริมและสนับสนุนการเพาะปลูกข้าวต่อไป

บทที่ 2

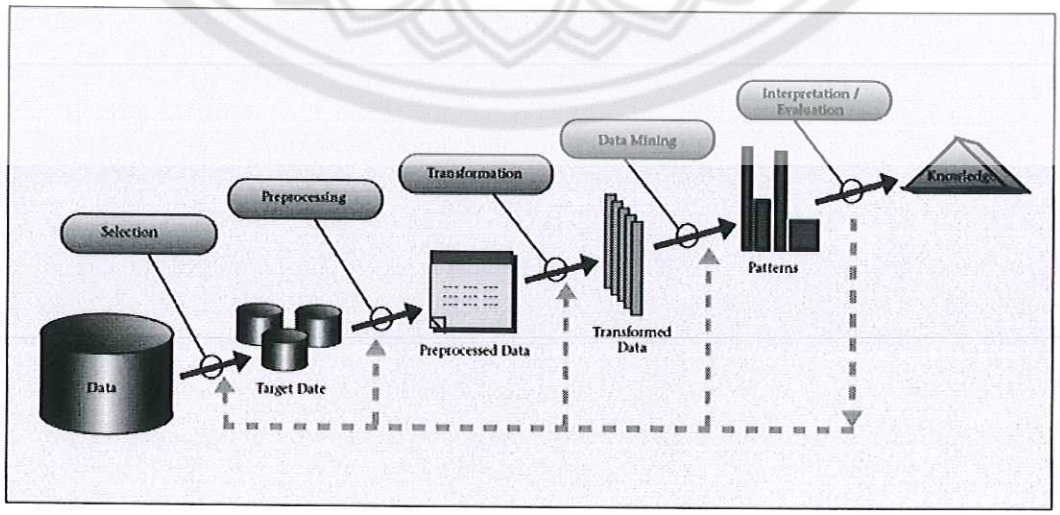
กรอบแนวคิดทฤษฎี และงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงรายละเอียดเกี่ยวกับความสำคัญและที่มาของปัญหาวิจัย รวมไปถึงการทบทวนวรรณกรรมของการศึกษาที่เกี่ยวข้อง และรายละเอียดโดยรวมของกระบวนการพัฒนาตัวแบบเพื่อการพยากรณ์ผลผลิตข้าวในเขตภาคเหนือตอนล่าง ซึ่งรายละเอียดเป็นดังนี้

กรอบแนวคิดในการทำวิจัย

เทคนิคการทำเหมืองข้อมูล (Data mining technique) คือวิทยาการที่รวมศาสตร์หลายๆ สาขา เช่น หลักการรู้จำ การเรียนรู้ของเครื่อง และเทคนิคทางสถิติและคณิตศาสตร์ มาใช้ร่วมกันเพื่อการจัดเก็บ ขุดค้น สารสนเทศ และตีความข้อสนเทศจากข้อมูลขนาดใหญ่ จากเดิมที่มีการจัดเก็บข้อมูลอย่างง่ายๆ มาสู่การจัดเก็บข้อมูลในฐานข้อมูลที่สามารถนำไปใช้ เป็นกระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูล หรือ กลั่นกรองสารสนเทศ (Information) ที่ยังไม่ทราบซึ่งซ่อนอยู่ในฐานข้อมูล เพื่อทำนายแนวโน้มและพฤติกรรม โดยอาศัยข้อมูลในอดีต โดยขั้นตอนในการดำเนินการต้องอาศัยวิธีการต่างๆ มาทำการวิเคราะห์ข้อมูลที่มีอยู่ เพื่อให้ได้ลักษณะของตัวแบบ (Model) ที่นำมาใช้อธิบายสภาพการณ์ที่เกิดขึ้น และนำไปใช้สนับสนุนการตัดสินใจ (Decision support) ในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภท ทั้งในด้านธุรกิจที่ช่วยในการตัดสินใจของผู้บริหาร ในด้านวิทยาศาสตร์และการแพทย์ รวมทั้งในด้านเศรษฐกิจและสังคม (บุญเสริม กิจศิริกุล, 2545 ; Wikipedia, 2555 ; ปาลจิตต์ พันทวาที, 2552 ; เดอะเพาเวอร์ สเตชั่น, 2555 ; Daniel T. Larose, 2005 ; กฤษณะ ไวยมัย, 2544)

การขุดค้น (Mining) เป็นขั้นตอนของการทำเหมืองข้อมูล ซึ่งประสิทธิภาพของการทำเหมืองข้อมูลจะขึ้นอยู่กับปัจจัยและตัวแปรที่นำมาศึกษา ซึ่งปัจจัยที่สำคัญมีอยู่ 2 อย่าง ได้แก่ คุณภาพของชุดข้อมูลและความสามารถของอัลกอริทึมที่นำมาใช้ การทำเหมืองข้อมูลเป็นส่วนหนึ่งในกระบวนการการค้นหาความรู้ในฐานข้อมูล (Knowledge discovery in database : KDD) ซึ่ง KDD เป็นวิธีที่ใช้จัดการข้อมูลที่มีขนาดใหญ่และมีความซับซ้อน เป็นกระบวนการในการค้นหาลักษณะแฝงของข้อมูลที่อยู่ในกลุ่มข้อมูลจำนวนมาก กระบวนการของ KDD ประกอบด้วยขั้นตอนต่างๆ ดังแสดงในภาพ 1 และสามารถเขียนเป็นขั้นตอนได้ดังนี้



ภาพ 1 แสดงกระบวนการของ KDD

1) การคัดเลือกข้อมูล (Data selection) เป็นการระบุถึงแหล่งข้อมูลที่จะนำมาใช้ในการขุดค้น ลักษณะของตัวแปรที่อยู่ในข้อมูล จะมี 2 ชนิดคือ ตัวแปรแบบคุณภาพ (Categorical) และ ตัวแปรแบบปริมาณ (Quantitative)

2) การกรองข้อมูล (Data cleaning) เป็นกระบวนการที่ทำให้เกิดความมั่นใจในคุณภาพของข้อมูลที่จะนำมาวิเคราะห์

3) การแปลงรูปแบบข้อมูล (Data transformation) ข้อมูลที่ผ่านการกรองข้อมูลแล้วจะถูกแปลงให้เป็นรูปแบบของข้อมูลที่เหมาะสมสำหรับนำไปใช้วิเคราะห์

4) การทำเหมืองข้อมูล คือ การเลือกใช้เทคนิคของเหมืองข้อมูลที่เหมาะสมสำหรับงานที่ต้องการ มาประมวลผลข้อมูลที่ผ่านการแปลงรูปแบบข้อมูลแล้วเพื่อดึงความรู้ หรือ ข้อมูลที่น่าสนใจ

5) การวิเคราะห์และประเมินผลลัพธ์ที่ได้ (Result analysis and evaluation) เป็นขั้นตอนการแปลความหมาย และประเมินผลลัพธ์ที่ได้ (Usama Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth, 1996 ; ปาลจิตต์ พันทวาที, 2552 ; Wikipedia, 2555)

โดยทั่วไปประเภทของการทำเหมืองข้อมูลสามารถแบ่งได้เป็น 2 ประเภทใหญ่ๆ คือ

1) การสร้างแบบจำลองในการทำนาย (Predictive modeling, supervised modeling) เป็นการคาดคะเนลักษณะหรือประมาณค่าที่ชัดเจนของข้อมูลที่จะเกิดขึ้นโดยใช้พื้นฐานจากข้อมูลที่ผ่านมาในอดีต ในที่นี้ทุกข้อมูลจะมีคุณสมบัติหนึ่งเรียกว่าคำตอบ (Output) ซึ่งค่าของคุณสมบัตินี้จะเป็นค่าที่ใช้ในการทำนายผลของข้อมูล อัลกอริทึมประเภทนี้จะมุ่งเน้นในการแบ่งแยกข้อมูลออกเป็นกลุ่มตามค่าคุณสมบัติของคำตอบ ซึ่งถ้าค่าคุณสมบัติของคำตอบมีค่าไม่ต่อเนื่อง จะเรียกกระบวนการที่ใช้แบ่งแยกว่า การจำแนก (Classification) ถ้าค่าคุณสมบัติของคำตอบมีค่าต่อเนื่องจะเรียกกระบวนการที่ใช้แบ่งว่าการถดถอย (Regression analysis)

2) การสร้างแบบจำลองในการบรรยาย (Descriptive modeling, unsupervised modeling) เป็นการหาแบบจำลองเพื่ออธิบายลักษณะบางอย่างของข้อมูลที่มีอยู่ ซึ่งโดยส่วนมากจะเป็นลักษณะการแบ่งกลุ่มให้กับข้อมูล ในที่นี้ อาจเป็นการหาความสัมพันธ์ต่างๆ (Association) หรือหาการจัดกลุ่มข้อมูล (Clustering) ซึ่งไม่ได้มีจุดมุ่งหมายเพื่อการทำนาย (บุญเสริม กิจศิริกุล, 2545 ; ปาลจิตต์ พันทวาที, 2552)

จากประเภทหลัก 2 ประเภท เทคนิคเหมืองข้อมูลยังสามารถจำแนกเป็นเทคนิคย่อยได้ 4 แบบ คือ

1) การสร้างกฎความสัมพันธ์ (Association rule Discovery) เป็นเทคนิคเหมืองข้อมูลแบบการบรรยาย โดยหลักการทำงานคือการค้นหาความสัมพันธ์ของข้อมูลจากข้อมูลขนาดใหญ่ที่มีอยู่เพื่อไปวิเคราะห์ หรือทำนายปรากฏการณ์ต่างๆ ตัวอย่างของการประยุกต์ใช้ เช่น การวิเคราะห์ข้อมูลการขายสินค้า โดยเก็บข้อมูลจากระบบณ จุดขาย (POS) หรือร้านค้าออนไลน์ แล้วพิจารณาสินค้าที่ผู้ซื้อมักจะซื้อพร้อมกัน เช่น ถ้าพบว่าคนที่ซื้อเทปวิดีโอมักจะซื้อเทปกาวด้วย ร้านค้าก็อาจจะจัดร้านให้สินค้าสองอย่างอยู่ใกล้กัน เพื่อเพิ่มยอดขาย หรืออาจจะพบว่าหลังจากคนซื้อหนังสือ ก แล้ว มักจะซื้อหนังสือ ข ด้วย ก็สามารถนำความรู้นี้ไปแนะนำผู้ที่กำลังจะซื้อหนังสือ ก ได้

2) การจำแนกและการพยากรณ์ (Classification and prediction) เป็นเทคนิคเหมืองข้อมูลแบบการทำนาย ซึ่ง การจำแนก (Classification) จะเป็นกระบวนการสร้างแบบจำลอง (Model) จำแนกข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ และ การพยาการณ์ (Prediction) จะเป็นการทำนายหาค่าที่ต้องการจากข้อมูลที่มีอยู่ เป็นการหากฎเพื่อระบุประเภทของวัตถุจากคุณสมบัติของวัตถุ เช่น หาความสัมพันธ์ระหว่างผลการตรวจร่างกายต่างๆ กับการเกิดโรค โดยใช้ข้อมูลผู้ป่วยและการวินิจฉัยของแพทย์ที่เก็บไว้ เพื่อนำมาช่วยวินิจฉัยโรคของผู้ป่วย หรือ

การวิจัยทางการแพทย์ ในทางธุรกิจจะใช้เพื่อดูคุณสมบัติของผู้ที่จะก่อหนี้ดีหรือหนี้เสีย เพื่อประกอบการพิจารณาการอนุมัติเงินกู้

3) การจัดกลุ่มข้อมูล (Clustering) เป็นเทคนิคเหมืองข้อมูลแบบการบรรยาย เป็นเทคนิคในการลดขนาดข้อมูลด้วยการรวมกลุ่มตัวแปรที่มีลักษณะเดียวกันไว้ด้วยกัน โดยแบ่งข้อมูลที่มีลักษณะคล้ายกันออกเป็นกลุ่ม แบ่งกลุ่มผู้ป่วยที่เป็นโรคเดียวกันตามลักษณะอาการ เพื่อนำไปใช้ประโยชน์ในการวิเคราะห์หาสาเหตุของโรค โดยพิจารณาจากผู้ป่วยที่มีอาการคล้ายคลึงกัน

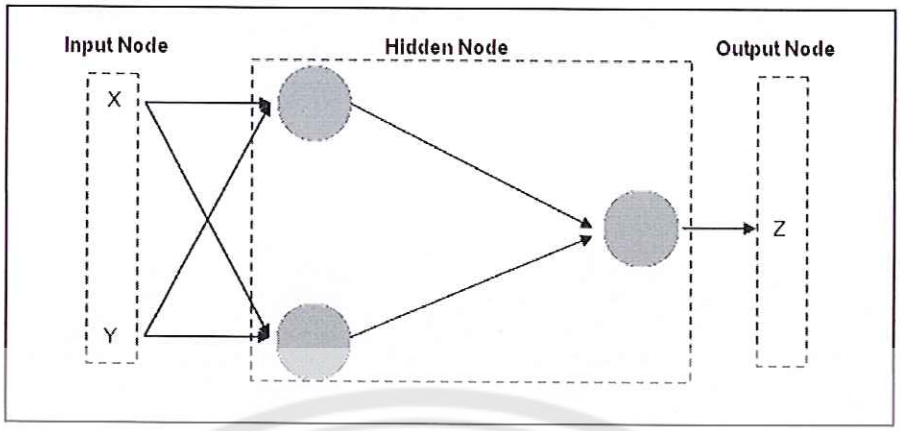
4) การบรรยายและจินตทัศน์ (Description and visualization) เป็นเทคนิคเหมืองข้อมูลแบบการบรรยาย คือ การหาคำอธิบายถึงสิ่งที่จะเกิดขึ้นโดยอาศัยข้อมูลจากฐานข้อมูล และ จินตทัศน์ คือ การนำเสนอข้อมูลในรูปแบบกราฟิก การนำเสนอจะสามารถทำได้มากกว่า 2 มิติ ซึ่งจะสร้างความละเอียดของการนำเสนอและสร้างความเข้าใจให้มากขึ้น ซึ่งเราอาจพบข้อมูลที่ซ้อนเร้นเมื่อดูข้อมูลชุดนั้นด้วยจินตทัศน์ (Wikipedia, 2555 ; อดิศักดิ์ พงษ์พูลผลศักดิ์ และ พิณรัตน์ นุชโพธิ์, 2550 ; เชษฐา จิรไพศาลกุล, 2550)

โครงข่ายประสาทเทียม (Neural Network)

โครงข่ายประสาทเทียม หรือ นิวรอลเน็ต คือ โมเดลทางคณิตศาสตร์ หรือ โมเดลทางคอมพิวเตอร์สำหรับประมวลผลสารสนเทศด้วยการคำนวณแบบคอนเนคชันนิสต์ (connectionist) เป็นเทคโนโลยีที่มาจากงานวิจัยด้านปัญญาประดิษฐ์ (Artificial Intelligence : AI) เพื่อใช้ในการคำนวณค่าฟังก์ชันจากกลุ่มข้อมูล วิธีการของ นิวรอลเน็ต (Artificial neural networks หรือ ANN) เป็นวิธีการที่ให้เครื่องเรียนรู้จากตัวอย่างต้นแบบ แล้วฝึก (train) ให้ระบบได้รู้จักที่จะคิดแก้ปัญหาที่กว้างขึ้นได้ แนวคิดเริ่มต้นของเทคนิคนี้ได้มาจากการศึกษาโครงข่ายไฟฟ้าชีวภาพ (bioelectric network) ในสมอง ซึ่งประกอบด้วยเซลล์ประสาท (neurons) และ จุดประสานประสาท (synapses) ตามโมเดลนี้ ข่ายงานประสาทเกิดจากการเชื่อมต่อระหว่างเซลประสาท จนเป็นเครือข่ายที่ทำงานร่วมกัน

โครงสร้างของนิวรอลเน็ตจะประกอบด้วยโหนดสำหรับ ข้อมูลนำเข้า (Input value) และ ผลลัพธ์ (Output value) การประมวลผลจะกระจายอยู่ในโครงสร้างเป็นชั้น ๆ ได้แก่ ชั้นข้อมูลเข้า (input layer) ชั้นข้อมูลออก (output layer) และ ชั้นข้อมูลซ่อน (hidden layers) มีการกำหนดค่าน้ำหนัก (weight) ให้แก่เส้นทางนำเข้าของข้อมูลนำเข้าแต่ละตัว การประมวลผลของนิวรอลเน็ตจะอาศัยการส่งการทำงานผ่านโหนดต่าง ๆ ในชั้น (layer) เหล่านี้ ในการเรียนรู้ของโครงข่ายประสาทเทียม จะอาศัยอัลกอริทึมการแพร่ย้อนกลับ (Back-propagation Algorithm) ในการสร้างการเรียนรู้สำหรับโครงข่ายประสาทเทียม เพื่อให้มีความคิดเสมือนมนุษย์

ชั้นเครือข่าย (Network layer) เป็นหนึ่งในสถาปัตยกรรมของโครงข่ายประสาทเทียม โดยที่ชั้นเครือข่ายจะประกอบด้วย 3 ชั้น ได้แก่ โหนดข้อมูลเข้า (Input node) โหนดข้อมูลซ่อน (Hidden node) และโหนดข้อมูลออก (Output node)



ภาพ 2 แสดงโครงสร้างของชั้นเครือข่าย (Network Layer)

- การทำงานของโหนดข้อมูลเข้า จะทำหน้าที่แทนส่วนของข้อมูลดิบ ที่จะถูกป้อนเข้าสู่เครือข่าย
- การทำงานของแต่ละโหนดข้อมูลซ่อนจะถูกกำหนดโดยการทำงานของโหนดข้อมูลเข้า และค่าน้ำหนักบนความสัมพันธ์ระหว่าง โหนดข้อมูลเข้า และ โหนดข้อมูลซ่อน
- พฤติกรรมการทำงานของโหนดข้อมูลออก จะขึ้นอยู่กับการทำงานของโหนดข้อมูลซ่อน และ ค่าน้ำหนักระหว่างโหนดข้อมูลซ่อน และ โหนดข้อมูลออก

ประเภทของเครือข่ายนี้เราสามารถกำหนดการแทนค่าให้แก่ โหนดข้อมูลเข้าได้อย่างอิสระ ค่าน้ำหนักระหว่าง โหนดข้อมูลเข้า และ โหนดข้อมูลซ่อนจะถูกกำหนดเมื่อโหนดข้อมูลซ่อนกำลังทำงาน ฉะนั้นเวลาที่แก้ไขค่าน้ำหนักโหนดข้อมูลซ่อนจะสามารถเลือกว่าอะไรคือค่าที่เราแทนเข้ามา

สามารถจำแนกออกเป็น 2 ประเภท คือ โครงข่ายแบบชั้นเดียว (Single-layer perceptron) และโครงข่ายแบบหลายชั้น (Multi-layer perceptron)

1) โครงข่ายแบบชั้นเดียว (Single-layer perceptron) เป็นโครงข่ายประสาทเทียมอย่างง่ายที่มีเพียงชั้นรับข้อมูลป้อนเข้า และชั้นส่งข้อมูลออกเท่านั้น โหนดในชั้นรับข้อมูลป้อนเข้าทำหน้าที่รับข้อมูลเข้า แล้วส่งข้อมูลผ่านเส้นเชื่อโยงต่างๆ ไปให้โหนดในชั้นส่งข้อมูลออก ความเข้มของสัญญาณ หรือ ปริมาณข้อมูลที่นำเข้าสู่โหนดในชั้นส่งข้อมูลออกจะขึ้นอยู่กับค่าน้ำหนักที่อยู่บนเส้นเชื่อมโยง

โหนดในชั้นส่งข้อมูลออกจะนำข้อมูลที่ได้รับมาคำนวณโดยใช้ฟังก์ชันทางคณิตศาสตร์ที่เรียกว่า ฟังก์ชันการแปลง (Transfer function) ที่เหมาะสมกับปัญหา จากนั้นส่งผลลัพธ์ที่ได้ออกมาเป็นข้อมูลส่งออก ถ้า ผลลัพธ์ที่ต้องการเป็น “ใช่” หรือ “ไม่ใช่” จะต้องใช้ Threshold function ดังนี้

$$f(x) = \begin{cases} 1 \ if \ x \geq T \\ 0 \ if \ x < T \end{cases}, T = Threshold\ level \quad (1)$$

หรือ ผลลัพธ์เป็นค่าตัวเลขที่ต่อเนื่อง จะใช้ Continuous function เช่น Sigmoid function

$$f(x) = \frac{1}{1+e^{-ax}} \quad (2)$$

2) โครงข่ายแบบหลายชั้น (Multi-layer perceptron) เป็นโครงข่ายที่มีชั้นแอบแฝง (Hidden layer) ตั้งแต่ 1 ชั้นขึ้นไป โครงข่ายแบบหลายชั้นจะใช้ในกรณีที่ปัญหามีความซับซ้อน ซึ่งโครงข่ายแบบชั้นเดียวไม่สามารถแก้ปัญหาได้ จึงเพิ่มจำนวนโหนดที่มีการคำนวณ หรือชั้นแอบแฝงให้กับโครงข่าย ตัวอย่างของโครงข่ายแบบหลายชั้น เช่น การแพร่ย้อนกลับ (Back propagation), เซลฟ์ออร์แกนไนซิ่งแมปช์ (Self organizing maps)

และ เคาน์เตอร์พรอพะเกชัน (Counter propagation) เป็นต้น(Wikipedia, 2555 ; ธัญรัตน์ สิทธิพล, 2552 ; Artificial Neural Network โครงข่ายประสาทเทียม, 2555 ; ธนาวุฒิ ประกอบผล, 2552)

การวิเคราะห์การถดถอย (Regression analysis)

การวิเคราะห์การถดถอยเป็นเทคนิคการวิเคราะห์ทางสถิติที่ถูกนำมาใช้อย่างแพร่หลายในการศึกษาเพื่อหารูปแบบความสัมพันธ์หรือฟังก์ชันระหว่างกลุ่มตัวแปรอิสระ (Independent variable) มักแทนด้วยตัวแปร X และตัวแปรตาม (Dependent variable) มักแทนด้วยตัวแปร Y จากนั้นนำรูปแบบความสัมพันธ์ที่ประมาณได้ไปพยากรณ์ค่าของตัวแปรตามเมื่อทราบค่ากลุ่มตัวแปรอิสระ ซึ่งการศึกษาหารูปแบบความสัมพันธ์ระหว่างตัวแปรดังกล่าว หากเป็นการศึกษาเฉพาะ 2 ตัวแปร กล่าวคือ ตัวแปรตัวหนึ่งเป็นตัวแปรตามและตัวแปรอีกตัวเป็นตัวแปรอิสระเพียงตัวเดียว เรียกการศึกษานี้ว่า การวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย (Simple linear regression analysis) แต่ถ้ามีตัวแปรอิสระตั้งแต่ 2 ตัวขึ้นไป จะเรียกการศึกษาประเภทนี้ว่าการวิเคราะห์การถดถอยเชิงพหุ (Multiple linear regression analysis)

ตัวแบบการถดถอยหรือฟังก์ชันของ Y เมื่อกำหนดตัวแปรอิสระ $X_1, X_2, ..., X_k$ สามารถเขียนได้ดังสมการ (3)

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon \qquad (3)$$

และเมื่อทำการทดลองหรือเก็บข้อมูลของตัวแปรตามและตัวแปรอิสระแล้ว สามารถเขียนรูปแบบให้อยู่ในรูปของเมทริกซ์ได้ดังนี้

$$Y = X\beta + \varepsilon$$

โดยที่

$$Y_{(n\times 1)} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X_{(n\times p)} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1k} \\ 1 & X_{21} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{bmatrix}$$

$$\beta_{(p\times 1)} = \begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \varepsilon_{(n\times 1)} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

เมื่อ Y แทนเวคเตอร์ของค่าสังเกตขนาด $n\times 1$

X แทน เมทริกซ์ของค่าตัวแปรอิสระขนาด $n\times p$ เมื่อ $p = k+1$

β แทนเวคเตอร์ของค่าสัมประสิทธิ์ถดถอยขนาด $p\times 1$

ε แทนเวคเตอร์ของความคลาดเคลื่อนขนาด $n\times 1$

ในการประมาณค่าพารามิเตอร์ในสมการที่ (3) สามารถทำได้โดยใช้วิธีกำลังสองน้อยที่สุด (Least squares method) ซึ่งสามารถเขียนผลรวมกำลังสองของความคลาดเคลื่อนได้ดังนี้

$$\begin{aligned} \text{SSE} &= \sum_{i=1}^{n} \varepsilon_i^2 \\ &= \varepsilon'\varepsilon \\ &= (Y - X\beta)'(Y - X\beta) \end{aligned}$$

$$= Y'Y - \beta' X'Y - Y'X\beta + \beta' X'X\beta$$
$$= Y'Y - 2\beta' X'Y + \beta' X'X\beta$$

จากนั้น ทำการหาอนุพันธ์ย่อยเทียบกับพารามิเตอร์ β แล้วกำหนดให้สมการมีค่าเท่ากับศูนย์ และแทนค่าพารามิเตอร์ด้วยตัวประมาณ b จะได้ว่า

$$\left.\frac{\partial SSE}{\partial \beta}\right|_b = -2X'Y + 2X'Xb = 0$$

จะได้สมการปกติ (Normal equation) ในรูปของเมทริกซ์ดังนี้

$$X'Xb = X'Y$$

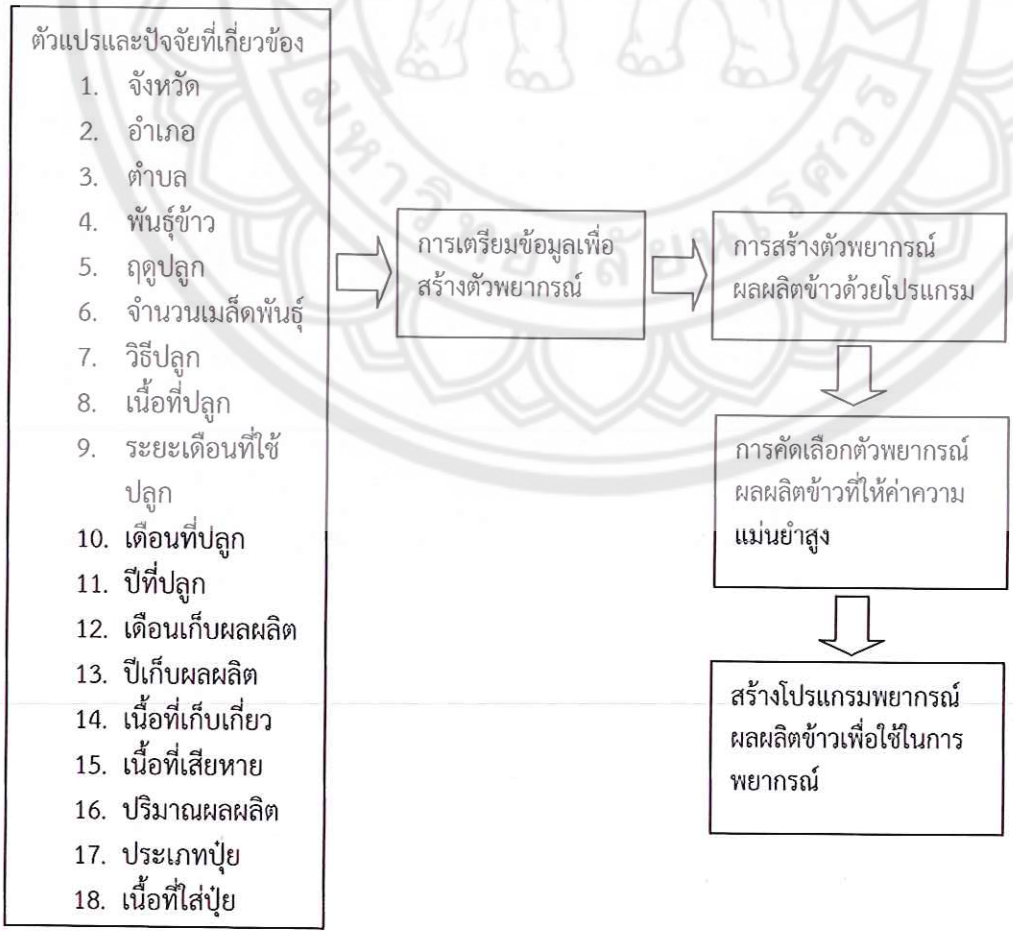
เมื่อแก้สมการจะได้ตัวประมาณกำลังสองน้อยที่สุดของ β คือ

$$b = (X'X)^{-1} X'Y$$

โครงการวิจัยนี้จะนำเทคนิคการวิเคราะห์การถดถอยเข้ามาสร้างตัวแบบเพื่อพยากรณ์ผลผลิตข้าวและนำผลที่ได้ไปเปรียบเทียบกับวิธีโครงข่ายประสาทเทียมว่าตัวแบบใดมีความแม่นยำในการพยากรณ์มากกว่ากัน นอกจากนี้ผู้วิจัยจะเปรียบเทียบเกี่ยวกับความเรียบง่ายในการสร้างตัวแบบ เพื่อหาข้อสรุปเกี่ยวกับข้อเสนอแนะในการเลือกใช้ตัวแบบสำหรับผู้ใช้ต่อไป

กรอบแนวคิดที่ใช้ในการวิจัย

จากที่กล่าวมาข้างต้น สามารถสร้างกรอบแนวคิดที่เกี่ยวข้องกับการพัฒนาตัวแบบเพื่อพยากรณ์ผลผลิตข้าวในเขตภาคเหนือตอนล่างโดยใช้เทคนิคเหมืองข้อมูลและตัวแบบทางสถิติ ผู้วิจัยได้ทำการออกแบบกรอบแนวคิดที่ใช้ในงานการวิจัยครั้งนี้ได้แสดงรายละเอียดดังภาพ



วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษา รวบรวม และคัดกรองข้อมูลที่มีอิทธิพลต่อผลผลิตข้าวในเขตภาคเหนือตอนล่าง
2. เพื่อหาตัวแปรหรือปัจจัยที่มีความสำคัญต่อผลผลิตข้าวในเขตภาคเหนือตอนล่าง
3. เพื่อสร้างตัวแบบในการพยากรณ์ผลผลิตข้าวในเขตภาคเหนือตอนล่างโดยใช้เทคนิคทางสถิติและเทคนิคเหมืองข้อมูล
4. เพื่อเปรียบเทียบความแม่นยำของตัวแบบพยากรณ์ที่สร้างได้
5. เพื่อพัฒนาโปรแกรมการพยากรณ์ผลผลิตข้าวเพื่อหน่วยงานที่เกี่ยวข้องสามารถนำไปใช้ได้

ประโยชน์ที่คาดว่าจะได้รับ

1. ได้ทราบถึงปัจจัยข้าวที่มีผลต่อการพยากรณ์ผลผลิตข้าว
2. ได้ตัวแบบพยากรณ์ผลผลิตข้าวในเขตภาคเหนือตอนล่าง
3. ได้ระบบการพยากรณ์ผลผลิตข้าวและการสืบค้นข้อมูลผ่านเว็บแอพพลิเคชั่น
4. ได้เว็บแอพพลิเคชั่นที่มีความถูกต้องแม่นยำ ด้านการสืบค้นข้อมูลและโปรแกรมการพยากรณ์ผลผลิตข้าว และสามารถนำไปเผยแพร่ให้กับเจ้าหน้าที่และผู้ที่สนใจทั่วไป

ขอบเขตการวิจัย

ทำการวิจัยแบบ Retrospective study โดยใช้ข้อมูลที่เก็บรวบรวมไว้แล้วโดยสำนักงานเศรษฐกิจการเกษตรเขต 2 จากนั้นทำการคัดกรองข้อมูลและแปลข้อมูลเพื่อให้ตรงกับรูปแบบของโมเดลเพื่อการพยากรณ์ที่จะสร้างทำการวิเคราะห์ความสัมพันธ์ของข้อมูลรวมไปถึงการพัฒนาตัวแบบเพื่อการพยากรณ์โดยใช้โปรแกรม R และโปรแกรมทางสำหรับการทำเหมืองข้อมูล WEKA นอกจากนี้เพื่อเป็นการนำผลการวิจัยไปใช้ประโยชน์ ผู้วิจัยจะทำการพัฒนาระบบเพื่อการพยากรณ์ผลผลิตข้าวเพื่อให้นักวิชาการเกษตรสามารถนำไปใช้ได้ และทำการตีพิมพ์เผยแพร่ผลงานวิจัยในวารสารระดับนานาชาติจำนวน 1 เรื่อง โดยขอบเขตของโครงการวิจัยสามารถสรุปได้ดังนี้

1. คัดเลือกตัวแปรที่มีอิทธิพลต่อผลผลิตข้าวในเขตภาคเหนือตอนล่างโดยใช้การวิเคราะห์สหสัมพันธ์ (Correlation analysis) เพื่อคัดเลือกตัวแปรที่มีนัยสำคัญทางสถิติที่ระดับ 0.05 ไปสร้างสมการเพื่อการพยากรณ์ 2 แบบได้แก่ การวิเคราะห์การถดถอยเชิงพหุ และเทคนิคโครงข่ายประสาทเทียม
2. ศึกษาถึงความแม่นยำในการพยากรณ์ของตัวแบบทางสถิติ 2 ตัวแบบ ดังนี้
   2.1 ตัวแบบ Regression โดยใช้ขั้นตอนวิธีการหาค่าเหมาะสมที่สุดในการประมาณค่าพารามิเตอร์ด้วยวิธีกำลังสองน้อยที่สุด (Least square method)
   2.2 ตัวแบบ ANN ที่มีโครงสร้างแบบมัลติเลเยอร์ที่พัฒนาขึ้นโดยโปรแกรม WEKA
3. เกณฑ์ที่ใช้ในการเปรียบเทียบความแม่นยำในการพยากรณ์ได้แก่
   3.1 เกณฑ์ RMSE ซึ่งสามารถคำนวณได้ดังนี้

$$RMSE = \sqrt{\frac{\sum_{i=1}^{k}(y_i - \hat{y}_i)^2}{k}} \qquad (2.9)$$

$$MAE = \frac{1}{k}\sum_{i=1}^{k}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

เมื่อ k คือ จำนวนจุดทดสอบ

y_i คือ ตัวแปรตามที่เป็นจริงสำหรับจุดทดสอบที่ i , $i = 1,2,...k$

$\hat{y}_i$ คือ ผลลัพธ์จากการพยากรณ์สำหรับจุดทดสอบที่ i , $i = 1,2,...k$

วิธีดำเนินการวิจัย

การวิจัยเรื่อง การพัฒนาตัวแบบทางสถิติเพื่อการพยากรณ์ผลผลิตข้าวในเขตภาคเหนือตอนล่าง เป็นวิจัยเชิงพัฒนาซึ่งผู้วิจัยได้แบ่งการศึกษาออกเป็น 3 ขั้นตอน คือ

1. การเตรียมข้อมูลเพื่อสร้างตัวแบบพยากรณ์ผลผลิตข้าวในเขตภาคเหนือตอนล่าง โดยใช้เทคนิคการทำเหมืองข้อมูล โดยทำการศึกษาและรวบรวมข้อมูลข้าวต่าง ๆ ที่เกี่ยวข้องกับผลผลิตข้าว จากเอกสาร หนังสือ วารสาร และงานวิจัยที่เกี่ยวข้อง ซึ่งเป็นการวิจัยจากเอกสาร รวบรวมเพื่อสร้างกรอบแนวคิดที่ใช้ในการวิจัย และเป็นแนวทางในการเพื่อตรวจสอบความถูกต้องของเนื้อหา และมีการเตรียมข้อมูลด้วยการทดสอบหาค่าสหสัมพันธ์ของข้อมูล (Correlations) เพื่อหาความสัมพันธ์ของตัวแปรต้นกับตัวแปรตาม แล้วนำข้อมูลมาวิเคราะห์และสรุปผล
2. การสร้างตัวแบบพยากรณ์ด้วยโปรแกรมเวกา (WEKA) และโปรแกรม R
3. การออกแบบและพัฒนาโปรแกรมสำหรับพยากรณ์ผลผลิตข้าวในเขตภาคเหนือตอนล่าง

บทที่ 3

วิธีดำเนินการวิจัย

การดำเนินงานวิจัยครั้งนี้มุ่งเปรียบเทียบความแม่นยำในการพยากรณ์ของตัวแบบทางสถิติ 2 ตัวแบบ ได้แก่ ตัวแบบ Regression และ ANN สำหรับพยากรณ์ผลผลิตข้าวในเขตภาคเหนือตอนล่างที่เก็บตัวแปรที่เกี่ยวข้องต่าง ๆ ในช่วงปี พ.ศ. 2550 ถึงปี 2553 โดยคัดเลือกตัวแปรที่มีนัยสำคัญทางสถิติเข้าไปสร้างตัวแบบเพื่อการพยากรณ์และทำการเปรียบเทียบความแม่นยำในการพยากรณ์ของแต่ละตัวแบบโดยใช้เกณฑ์ RMSE และเกณฑ์ MAE ซึ่งมีการวางแผนและกำหนดขั้นตอนในการดำเนินงานวิจัย ดังนี้

ขั้นตอนการวิจัย

การดำเนินงานวิจัย มีขั้นตอนดังนี้

1. คัดกรองข้อมูลที่มีอยู่ทั้งหมด โดยตัดข้อมูลที่มีค่าสูญหายออก คงเหลือข้อมูลทั้งสิ้น 9,206 รายการ และแบ่งข้อมูลออกเป็นชุดฝึกสอน (Training set) และชุดทดสอบ (Test set) โดยใช้อัตราส่วน 80:20
2. เลือกตัวแปรอิสระที่มีความสำคัญกับผลผลิตข้าวในเขตภาคเหนือโดยการวิเคราะห์สหสัมพันธ์เพื่อนำเข้าตัวแบบในการพยากรณ์ทั้ง Regression and ANN
3. สร้างตัวแบบทางสถิติ ได้แก่ตัวแบบ Regression และ ตัวแบบ ANN ดังนี้

3.1 สร้างตัวแบบ Regression

นำข้อมูลชุดฝึกสอนและตัวแปรที่มีนัยสำคัญทางสถิติสร้างตัวแบบ Regression โดยวิธีกำลังสองน้อยที่สุด โดยใช้โปรแกรม R เวอร์ชัน 3.1.2 เกณฑ์ในการเลือกตัวแบบที่เหมาะสมที่สุดโดยใช้เกณฑ์ Akaike criterion จะได้สมการเพื่อการพยากรณ์ในรูปแบบของฟังก์ชันพหุนามกำลังสอง (Second order polynomial model) ดังสมการต่อไปนี้

$$\hat{y}(x) = \beta_0 + \sum_{i=1}^{d} \beta_i x_i + \sum_{i=1}^{d} \beta_{ii} x_i^2 + \sum_{i=1}^{d} \sum_{i<j}^{d} \beta_{ij} x_i x_j \qquad (7)$$

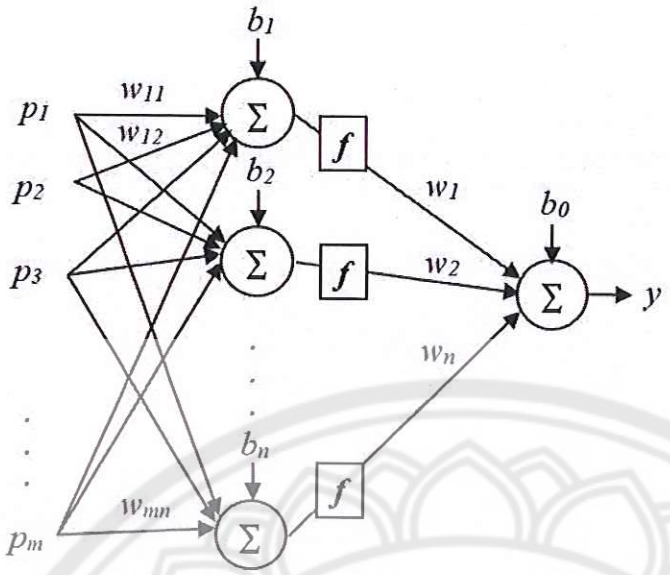
3.2 สร้างตัวแบบ ANN

สร้างตัวแบบตามรูปแบบความสัมพันธ์ดังสมการ

$$y = f(wp + b)$$

เมื่อ w เป็นค่าน้ำหนักถ่วงของตัวแปรเข้าแต่ละตัว

โดยทั่วไปการหาตัวแบบ ANN มักจะใช้โหนดหลายค่า (ในกรณีคือค่า $\sum$) ดังแสดงในภาพ 7



ภาพ 7 แสดงโครงสร้างของตัวแบบ ANN แบบหลายโหนด

ซึ่งกระบวนการทำงานของตัวแบบ ANN สามารถเขียนได้ดังนี้

$$y^1 = f^1(w^1 p + b^1)$$
$$y^2 = f^2(w^2 p + b^2)$$
$$\vdots$$
$$y^N = f^N(w^N p + b^N)$$

เมื่อ N แทนจำนวนโหนด

4. คำนวณหาค่า RMSE และ MAE
5. สรุปผลการวิจัยในแต่ละมิติปัญหาทดสอบ

บทที่ 4

ผลการวิจัย

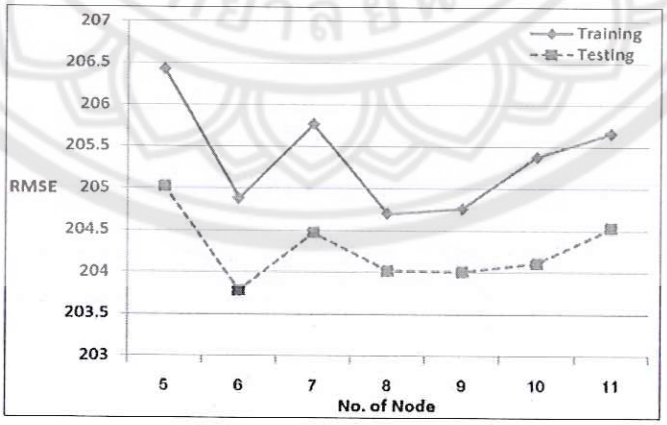
จากข้อมูลที่ได้จากการคัดกรอง และวิเคราะห์ทางสถิติเพื่อหาตัวแปรที่มีอิทธิพลต่อผลผลิตข้าวเฉลี่ยต่อไร่ ในเขตภาคเหนือตอนล่าง เมื่อนำมาสร้างตัวแบบทางสถิติ และคำนวณค่า RMSE และ MAE สำหรับเปรียบเทียบความแม่นยำในการพยากรณ์ของตัวแบบในการพยากรณ์ ทั้ง 2 ตัวแบบ ได้แก่ Regression และ ANN ผลการวิจัยที่ได้จะนำเสนอในรูปแบบตารางแยกตามวิธีการวิเคราะห์ที่แตกต่างกัน

ตาราง 1 แสดงตัวแปรที่มีนัยสำคัญทางสถิติและตัวแปรตาม

Variable name	Meaning	Range	*r*	P-value
RiceType	Rice seed type	{1,2,3,4,5,6}	0.243	<0.001
Method	Cultivating method	{1,2,3}	0.258	<0.001
SeedPerRai	Seed quantity, kg per rai	Min=7, Max=160	0.194	<0.001
PuiPerRai	Fertilizer used, kg per rai	Min=2.5, Max=116.7	0.287	<0.001
Period	Period of cultivation	Min=3, Max=8	-0.255	<0.001
InChon	Irrigation area	{0,1}	0.128	<0.001
ProdPerRai	**Rice yield, kg per rai**	**Min=50, Max=1900, Avg=562.22**		

จากตาราง 1 จะเห็นได้ว่าค่าตัวแปรที่มีความสำคัญต่อผลผลิตข้าวอย่างมีนัยสำคัญทางสถิติได้แก่ พันธุ์ข้าว วิธีการปลูก ปริมาณเมล็ดข้าวต่อไร่ ปริมาณปุ๋ยที่ใช้ ช่วงเวลาในการเก็บเกี่ยว และเขตชลประทาน ตามลำดับ

เมื่อนำตัวแปรที่มีอิทธิพลมาสร้างตัวแบบในการพยากรณ์ ขั้นตอนต่อไปเป็นการประมาณค่าพารามิเตอร์ที่เหมาะสมสำหรับตัวแบบแต่ละประเภท กรณีตัวแบบ ANN จะมีการหาค่าพารามิเตอร์ที่เหมาะสมโดยการทดลองค่าพารามิเตอร์ที่ระดับแตกต่างกัน และระดับที่ก่อให้เกิดค่า RMSE ที่น้อยที่สุดจะถือว่าเป็นชุดพารามิเตอร์ที่เหมาะสมที่สุด ดังรูป



ภาพ 8 แสดงวิธีการหาค่าพารามิเตอร์ที่เหมาะสมสำหรับตัวแบบ ANN

จากนั้นนำตัวแบบ Regression และตัวแบบ ANN มาเปรียบเทียบความสามารถในการพยากรณ์ทั้งในชุดฝึกสอนและชุดทดสอบ จะได้ผลดังตาราง

ตาราง 1 แสดงค่า RMSE และ MAE สำหรับตัวแบบ Regression และ ANN

Methods	Training data		Test data	
	RMSE	MAE	RMSE	MAE
Regression	207.05172	0.450052	204.2636	0.443842
ANN	204.88462	0.293767	203.7862	0.297208

จากตารางจะได้ว่า ตัวแบบ ANN มีความแม่นยำในการพยากรณ์มากกว่าตัวแบบ Regression เนื่องจากให้ค่า RMSE และค่า MAE ที่ต่ำกว่าทั้งในกรณีชุดฝึกสอนและชุดทดสอบ ดังนั้นจึงสามารถสรุปได้ว่าตัวแบบ ANN มีความเหมาะสมที่จะถูกนำไปใช้ในการพยากรณ์ผลผลิตเฉลี่ยต่อไร่ของข้าวในเขตภาคเหนือตอนล่าง

บทที่ 5

สรุปผลการวิจัย

การวิจัยครั้งนี้มีจุดมุ่งหมายเพื่อเปรียบเทียบความแม่นยำในการพยากรณ์ของตัวแบบทางสถิติ 2 ตัวแบบ ได้แก่ ตัวแบบ Regression และตัวแบบ ANN สำหรับพยากรณ์ผลผลิตข้าวในเขตภาคเหนือตอนล่าง โดยจะทำการประมาณค่า RMSE และ MAE และเปรียบเทียบความแม่นยำในการพยากรณ์ของตัวแบบทางสถิติทั้ง 2 ตัวแบบ ทั้งในชุดฝึกสอนและชุดทดสอบ

สรุปผลการวิจัย

จากการเปรียบเทียบความแม่นยำในการพยากรณ์ของตัวแบบพยากรณ์ผลผลิตเฉลี่ยต่อไร่ของข้าวในเขตภาคเหนือตอนล่าง พบว่า ตัวแบบ ANN ให้ความแม่นยำในการพยากรณ์มากกว่าตัวแบบ Regression ทั้งนี้เนื่องจากโครงสร้างของตัวแบบ ANN มีความยืดหยุ่นมากกว่า และสามารถค้นหารูปแบบความสัมพันธ์ระหว่างกลุ่มตัวแปรได้ดีกว่า จึงส่งผลให้ตัวแบบที่มีความแม่นยำในการพยากรณ์ นอกจากนี้ยังพบว่าตัวแบบ Regression มีความแม่นยำน้อยกว่าเนื่องจากมีตัวแปรเชิงคุณภาพในตัวแบบการพยากรณ์ค่อนข้างมาก จึงส่งผลให้ความแม่นยำในภาพรวมด้อยกว่า ANN ดังนั้นจึงสามารถสรุปได้ว่าตัวแบบ ANN ควรถูกนำมาใช้การพยากรณ์ผลผลิตข้าว เนื่องจากให้ความแม่นยำสูง รวมไปถึงเป็นตัวแบบที่ค่อนข้างยืดหยุ่นและไม่จำเป็นต้องทราบข้อมูลภูมิหลังของตัวแปรต่าง ๆ

อภิปรายผล

จากการศึกษาวิจัยเรื่องการสร้างตัวแบบเพื่อพยากรณ์ผลผลิตข้าวในเขตภาคเหนือตอนล่าง มีประเด็นที่จะนำมาอภิปรายผลได้ดังนี้

พันธุ์ข้าว วิธีการปลูกข้าว จำนวนเมล็ดพันธุ์ ปริมาณปุ๋ยเคมี และปริมาณปุ๋ยหมัก เป็นปัจจัยที่มีผลกระทบต่อผลิตข้าว เนื่องจากการเลือกใช้เมล็ดพันธุ์ข้าวที่เหมาะสมจะสามารถให้ผลผลิตข้าวที่เพิ่มมากขึ้น การเลือกพันธุ์ข้าวในการเพราะปลูกจะต้องคำนึงถึงความเหมาะสมกับสภาพภูมิประเทศ ภูมิอากาศ ความต้านทานต่อโรคและแมลง วิธีการปลูกข้าว เป็นปัจจัยที่มีผลต่อผลผลิตข้าว เพราะการปลูกข้าวมีหลายวิธี ทั้งนาดำ นาหว่านน้ำตม และนาหว่านน้ำสำรวย ซึ่งให้ผลผลิตข้าวที่ต่างกันในเขตภาคเหนือตอนล่าง ปุ๋ยเป็นอาหารพืชที่ต้นข้าวต้องการมากสำหรับการเจริญเติบโต โดยเฉพาะดินนาที่มีความไม่อุดมสมบูรณ์ของดิน จะต้องมีการใส่ปุ๋ยในดินเพื่อต้นข้าวจะได้แข็งแรง แตกกอมาก และให้ผลิตผลสูง ธาตุอาหารที่ต้นข้าวต้องการปุ๋ยมาก ได้แก่ ไนโตรเจน ฟอสฟอรัส และโพแทสเซียม ส่วนมากมักจะมีในปุ๋ยเคมี เพราะฉะนั้น ปุ๋ยข้าวจะต้องมีธาตุเหล่านี้จำนวนมาก การใส่ปุ๋ยในปริมาณที่พอเหมาะจะได้ผลผลิตข้าวที่ดี ซึ่งสอดคล้องกับการศึกษาของ ทิพวรรณ สาระกุล และประภัสรา ศิริกาญจน์ พบว่า เนื้อที่การเพาะปลูกข้าวเจ้า ราคาน้ำมัน ความเร็วลม ปริมาณน้ำฝน ความยาวนานแสงแดด อุณหภูมิ มีผลต่อผลผลิตข้าว การใช้พันธุ์ข้าวที่ดี บริสุทธิ์ ไม่มีเมล็ดพันธุ์ข้าวอื่นหรือเมล็ดวัชพืชปลอมปน ย่อมส่งผลให้ได้ผลผลิตที่ดี และช่วยลดต้นทุนการผลิตได้อย่างมาก การเลือกข้าวพันธุ์ดี ต้องเริ่มจากข้าวที่ตรงตามพันธุ์ที่ต้องการจะปลูกไม่มีพันธุ์อื่นปะปน มีความแข็งแรง และมีเปอร์เซ็นต์ความงอกสูง เหมาะสมกับสภาพแวดล้อมในพื้นที่ปลูกของตนเอง ซึ่งนอกจากจะช่วยลดปริมาณวัชพืช ซึ่งเป็นศัตรูตัวสำคัญที่ทำให้ทั้งปริมาณและคุณภาพผลผลิตลดลง

ข้อเสนอแนะ

1. ควรศึกษาตัวแปรที่เกี่ยวข้องหรือปัจจัยอื่นๆ เช่นสภาพอากาศ ความชื้น สภาพดินและการดูแลรักษาระดับน้ำในนา ที่มีอิทธิพลต่อผลผลิตข้าวเพื่อนำมาใช้เป็นตัวแปรนำเข้าในการสร้างตัวแบบพยากรณ์

2. เพื่อเป็นแนวทางในการศึกษาและเปรียบเทียบความแม่นยำในการพยากรณ์ของตัวแบบ ควรศึกษาตัวแบบอื่น ๆ เช่น Multivariate adaptive regression splines (MARS) รวมไปถึงการปรับปรุงโครงข่ายประสาทเทียม (Artificial neural network: ANN) ที่มีการปรับปรุงโครงสร้างที่ดียิ่งขึ้น

3. เพื่อเป็นแนวทางในการศึกษาและเปรียบเทียบความแม่นยำในการพยากรณ์ของตัวแบบ โดยใช้เกณฑ์ในการเปรียบเทียบเกณฑ์อื่น เช่น เกณฑ์ค่าสัมบูรณ์ของความสัมพัทธ์ของความคลาดเคลื่อนเฉลี่ย (Relative average absolute error) เกณฑ์ค่าสัมบูรณ์ของความสัมพัทธ์ของความคลาดเคลื่อนที่มากที่สุด (Relative maximum absolute error) เป็นต้น

บรรณานุกรม

กฤษณะ ไวยมัย, ชิดชนก ส่งศิริ และธนาวินท์ รักธรรมมานนท์. (2544). การใช้เทคนิคดาต้าไมน์นิงเพื่อพัฒนาคุณภาพทางการศึกษาคณะวิศวกรรมศาสตร์. **NECTEC Technical Juornal**, 3(11), 134 – 142.

การทำเหมืองข้อมูล. (2555). สืบค้นเมื่อ 5 ธันวาคม 2554. จาก th.wikipedia.org/wiki/การทำเหมืองข้อมูล.

การป้องกันอุบัติเหตุจราจร. (2555). สืบค้นเมื่อ 5 ธันวาคม 2554, จาก http://www.benchama.ac.th/stdweb/sukabunyat/transport.html.

โครงข่ายประสาทเทียม. (2555). สืบค้นเมื่อ 5 ธันวาคม 2554 จาก th.wikipedia.org/wiki/โครงข่ายประสาทเทียม.

เชษฐา จิรไพศาลกุล. (2550). **การทำเหมืองข้อมูลสำหรับการวิเคราะห์การขาย.** วิทยานิพนธ์, มหาวิทยาลัยศรีปทุม. กรุงเทพฯ. สืบค้นเมื่อ 9 มกราคม 2555, จาก http://dllibrary.spu.ac.th:8080/dspace/handle/123456789/96.

ณพงศ์ วาณิชยพงศ์. (26 สิงหาคม 2553). **Confusion Matrix.** สืบค้นเมื่อ 20 ธันวาคม 2555, จาก http://plagad.wordpress.com/2010/08/26/confusion-matrix/.

ต้นไม้การตัดสินใจ. (2555). สืบค้นเมื่อ 5 ธันวาคม 2554. จาก th.wikipedia.org/wiki/ต้นไม้การตัดสินใจ.

ธนาวุฒิ ประกอบผล. (2552). โครงข่ายประสาทเทียม Artificial Neural Networks. **วารสาร มฉก.วิชาการ.** 12(24), 73 – 87.

นฤพนต์ ว่องประชานุกูล. (2548). **วิธีที่เหมาะสมสำหรับการตัดกิ่งต้นไม้ตัดสินใจของการทำเหมืองข้อมูลทางด้านวิทยาศาสตร์.** วิทยานิพนธ์ วศ.ม., มหาวิทยาลัยเทคโนโลยีสุรนารี. สืบค้นเมื่อ 9 มกราคม 2555, จาก http://dcms.thailis.or.th/tdc//index.php.

บุญเสริม กิจศิริกุล. (2545). **รายงานวิจัยฉบับสมบูรณ์ โครงการย่อยที่ 7 อัลกอริทึมการทำเหมืองข้อมูล.** กรุงเทพฯ: จุฬาลงกรณ์มหาวิทยาลัย.

ปาลจิตต์ พันทวาที. (2552). **บทความ DATAMINING.** สืบค้นเมื่อ 9 มกราคม 2555, จาก http://panjitpan.blogspot.com/2009/03/datamining.html.

พยุง มีสัจ และ ณัฏฐา ห่อประชุม. (2548). ระบบแบ่งกลุ่มแบบฟัซซีเบเซียนและการประยุกต์ใช้ในการอนุมัติสินเชื่อเบื้องต้น Fuzzy Bayesian Classification System and Its Application for Basic Credit Approval. **วารสารวิชาการพระจอมเกล้าพระนครเหนือ.** 15(3), 32 – 40.

สิทธิโชค มุกดาสกุลภิบาล. (2551). **การวัดประสิทธิภาพของขั้นตอนวิธี ตัวจำแนก C4.5, ADTree และ Naive Bayes ในการจำแนกข้อมูลการชุกซ่อนสิ่งเสพติดสำหรับไปรษณีย์ระหว่างประเทศ.** วิทยานิพนธ์ วท.ม., มหาวิทยาลัยเกษตรศาสตร์, กรุงเทพฯ.

โสภิตา ธูปเพ็ง. (6 สิงหาคม 2550) **เปรียบเทียบเทคนิค Data Mining กับ เทคนิคอื่นๆ.** สืบค้นเมื่อ 30 ธันวาคม 2554, จาก http://system5.multiply.com/journal/item/53?&show_interstitial=1&u=%2Fjournal%2Fitem.

สำนักงานเศรษฐกิจการเกษตร. (2552). ข้อมูลพื้นฐานเศรษฐกิจการเกษตร. สืบค้นเมื่อ 4 สิงหาคม 2553, จาก http://www.oae.go.th

สำนักงานเศรษฐกิจการเกษตร. (2553) สถานการณ์สินค้าเกษตรที่สำคัญและแนวโน้มปี 2553. สืบค้นเมื่อ 10 สิงหาคม 2553, จาก http://www.oae.go.th

Artificial Neural Network โครงข่ายประสาทเทียม. (2555). สืบค้นเมื่อ 22 กุมภาพันธ์ 2554, จาก. http://alaska.reru.ac.th/text/NN.pdf.

Daniel T. Larose. (2005). *Discovering Knowledge in Data An Introduction to Data Mining.* United States of America

DANIEL T. LAROSE. (2006). **Data Mining Methods and Models.** United States of America

Data Mining. (2552). สืบค้นเมื่อ 22 กุมภาพันธ์ 2554, จาก http://wekathai.blogspot.com/2009/11/data-mining.html.

Henry C. C. and Boosarawongse, R. (2007). Forecasting Thailand's rice export: Statistical techniques vs. artificial neural networks. Computers & Industrial Engineering, 53, pages 610-627.

Knowledge Discovery in Databases. (2552). สืบค้นเมื่อ 22 กุมภาพันธ์ 2554, จาก http://wekathai.blogspot.com/2009/11/data-mining.html.

Lertworasirikul, S. and Tipsuwan, Y. (2008). *Moisture content and water activity prediction of semi-finished cassava crackers from drying process with artificial neural network.* Journal of Food Engineering, 84, pages 65-74.

Mohamed A. Abdel-Aty and Hassan T. Abdelwahab. (2004). *Predicting Injury Severity Levels in Traffic Crashes: A Modeling Comparison*. Journal of Transporting Engineering. pages 204 – 210.

Movagharnejad, K. and Nikzad, M. (2007). *Modeling of tomato drying using artificial neural network.* Computers and Electronics in Agriculture, 59, pages 78-85.

Topuz, A. (2010). *Predicting moisture content of agricultural products using artificial neural networks*. Advanced in Engineering Software, 41, pages 464-470.

Usama Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth. (1996). *From Data Mining to Knowledge Discovery in Database*. **AI Magazine.** 17(3). 37 – 54.

ภาคผนวก

1. บทความที่ได้รับการตีพิมพ์ในวารสาร Lecture Notes in Electrical Engineering 339, DOI 10.1007/978-3-662-46578-3_88

A Comparison of Artificial Neural Network and Regression Model for Predicting the Rice Production in Lower Northern Thailand

Anamai Na-udom[1,*] and Jaratsri Rungrattanaubol[2]

[1]Department of Mathematics, Faculty of Science, Naresuan University, Phitsanulok, Thailand.
anamain@nu.ac.th
[2]Department of Computer Science and Information Technology, Faculty of Science, Naresuan University, Phitsanulok, Thailand.
jaratsrir@nu.ac.th

Abstract. Lower Northern Thailand is one of the main regions which can produce the highest rice yield. If the emphasis is on producing the rice yield in order to meet the standard yield, then the key factors, such as characteristics of rice farm, rice seed types, cultivation period, quantity of fertilizer usage, number of seeds, must be clearly studied and understood. This paper studies factors influencing the rice products and develops a model to predict rice yield per rai that can support farmers to plan their rice farming in Lower Northern Thailand. The aim of this paper is to compare the prediction accuracy between two popular predictive techniques for modelling rice yield namely, artificial neural network (ANN) and Regression. Root mean square of error (RMSE) and mean absolute error (MAE) values are used to compare prediction accuracy of the predictive models. The result shows that ANN is superior over regression model in terms of prediction accuracy and it is flexible to develop.

Keywords: Predictive Model; Artificial Neural Network; Regression model; Rice product in Lower Northern Thailand.

1 Introduction

Over the past three decades, Thailand has been recognized as the largest rice exporter in the world. In 2013, the department of agricultural extension reported that over 65 million hectares of land have been used to grow the rice field. There were 3.1 million households with farmers cultivating rice crops. According to the export records, Thailand earned more than a billion baht from rice export. Hence it is very clear that rice productivity is the major source of income of the agricultural sector and the industrial sector which contributes the employment for several million households.

Normally Thai farmers cultivate rice twice a year, in rainy season and in summer time, respectively. Hence rice production in Thailand can be classified into 2 groups according to the season of cultivation. The first group is called major rice which cultivated during June to December and another group is called second rice which normally grown during summer period. According to the empirical studies, it has been observed that rice yield has been affected by two aspects: the environment and the

K.J. Kim (ed.), *Information Science and Applications*,
Lecture Notes in Electrical Engineering 339, DOI 10.1007/978-3-662-46578-3_88

farmer's practice [1, 2]. Environmental factors include characteristics of cultivation area, amount of water, climate etc. The farmer's practice consists of selection of rice grain, selection of the appropriate method of rice cultivation area, and choice of fertilizer.

The lower northern Thailand consists of 8 provinces and it is suitable for rice cultivation as the landscape is very rich and moisture. Approximately 16.8 of cultivation are situated in the irrigation areas. It was reported that lower northern Thailand can produce the most major rice comparing to other parts in Thailand. Though the farmers in this part can cultivate rice through the year but it has been observed that the current yield of rice per rai is well below the standard yield (698 kg per rai). Hence the challenging is to investigate the factors influencing the rice yield so a suitable plan can be made prior to the cultivation time. This will benefit the farmers to increase their yields and income.

The development of accurate prediction models of the rice yields is important for the government organization to maximize the value to the farmer income [1, 3]. In the past, mathematical model such as linear regression model was used to predict to crop yield. However, the weakness of this method is that it relies on linear relationship assumption. Hence, non-linear approaches such as artificial neural network (ANN) and Bayesian classification are used to overcome the complex situation [1, 4]. Various modelling methods have been used to find an accurate predictive model. For instance Ji et al. [5] compared the performance of ANN and Regression models for rice yield prediction in mountainous regions and the results showed that ANN is superior over Regression. Shabri et al. [6] used time series forecasting technique to predict rice yield in Malaysia and compared the prediction accuracy with ANN and the results showed ANN performs better than forecasting technique. Paswan and Begum [7] discussed the performance of ANN and regression models in predicting the crop production. Raorane et al. [4] claimed that reliable and accurate forecasting techniques are required for decision making in the government office prior to pre-harvest crop. Uno et al. [8] did a comparison between ANN and stepwise multiple linear regression models in predicting corn yield and the results revealed that there was no clear difference between the two methods in terms of prediction accuracy.

Hence the aim of this paper is to compare the prediction accuracy between the two popular modelling methods including Regression and artificial neural network models. The selection of input factors will be presented and then the selected factors will be taken to the predictive model. The prediction accuracy of each model is implemented by using root mean square error (RMSE). In the next section, we present the research method including details of statistical models used in this study. The results based on prediction accuracy will be given in section 3 and the conclusion is summarized in section 4 respectively.

2 Research Method

In order to compare the prediction accuracy of Regression model and ANN, a data set of rice production in Lower Northern Thailand collected during 2007 to 2010 is used [9]. We first screen the important factors that influence the rice product using correlation analysis when nonparametric correlation is used. Then the key factors are applied to fit Regression and ANN models. The total number of 9,206 records is split into training set and test set with 80:20 proportions; hence, 7,380 records for training set and 1,826 for test set. The training set is used to construct a predictive model and the test set is applied as an unseen data for testing. The prediction accuracy is validated through RMSE and MAE values by using the training and test set. In this section, we present the details of statistical models.

2.1 Regression Model

Regression analysis is one of the most effective methods that have been successfully used in the context of yield prediction since it is simple to construct and provides information on input variables sensitivity [10]. This method is based on the assumption of random error arising from a large number of insignificant input factors. Given an output response, y, and input variables $= (x_1, \dots, x_d)$, the relationship between y and x can be mathematically written as

$$y = f(x) + \varepsilon \tag{1}$$

where ε is a random error which is assumed to be normally distributed with mean zero and variance σ^2. Since the true response surface function $f(x)$ is unknown, a response surface $g(x)$ is created to approximate $f(x)$. Therefore the predicted values are obtained by using $\hat{y} = g(x)$, which $g(x)$ can be treated as a polynomial function of $(X_1, X_2, \dots, X_d)$. The observed data set can be expressed in the matrix form using the data matrix X as

$$y_0 = X\beta + \varepsilon \tag{2}$$

where $y_0 = (y_1, y_2, \dots, y_n)^T$, x is a $n \times \alpha$ design matrix, β is a $(\alpha \times 1)$ vector of the regression coefficients, and ε is a $(n \times 1)$ vector of random error. The number of unknown parameters in equation (2) is determined by α, where $\alpha = 2d + \binom{d}{2} + 1$. The vector of least squares estimators, $\hat{\beta}$, can be determined subject to the minimization of

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = (y_0 - X\beta)^T (y_0 - X\beta) \tag{3}$$

Minimization of equation (3) yields

$$X^T X\hat{\beta} = X^T y_0 \tag{4}$$

Hence, the least squares estimator of β is

$$\hat{\beta} = (X^T X)^{-1} X^T y_0 \quad (5)$$

, provided that $(X^T X)$ is invertible.

Once β is estimated, equation (5) can be used to predict the rice yield value at any untried settings of input variables.

2.2 Artificial Neural Network Model

Artificial neural network (ANN) is commonly used in complex decision making problems [11]. Unlike a usual statistical approximation model, ANN does not require any assumptions of the model, making it easy to use in many applications such as science, engineering and health science [12]. The inspiration for neural networks was the recognition that complex learning systems in animal brains consisted of closely interconnected sets of neurons. A particular neural may be relatively simple in structure but dense networks of interconnected neurons could perform complex learning tasks such as pattern recognitions and approximation models. ANN consists of input (p), a data set, which is combined through a combination function such as summation ($\sum$) then pass such information into an activation function (f) to produce an output response (y) and b is a bias as shown in Fig. 1.

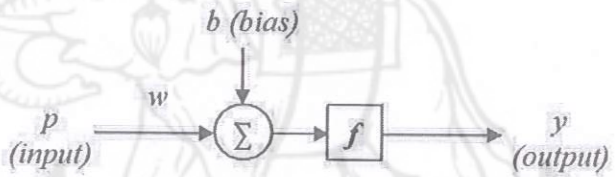


Fig. 1. A basic layout of ANN

The summary of ANN process can be rewritten as

$$y = f(wp + b) \quad (6)$$

, where w is the weight of each input variable.

Typically ANN is formed by multiple nodes (and probably multiple layers) as depicted in Fig. 2. Each node is symbolized by $\sum$. An activation function (f) can be the same or different. Examples of activation function are a linear, sigmoid and symmetrical hard limit. In this research, a sigmoid function is used. The weights between each node are adjusted by back propagation method. Fig. 2 displays ANN structure, which contains m input variables, one hidden n node layer and one output.

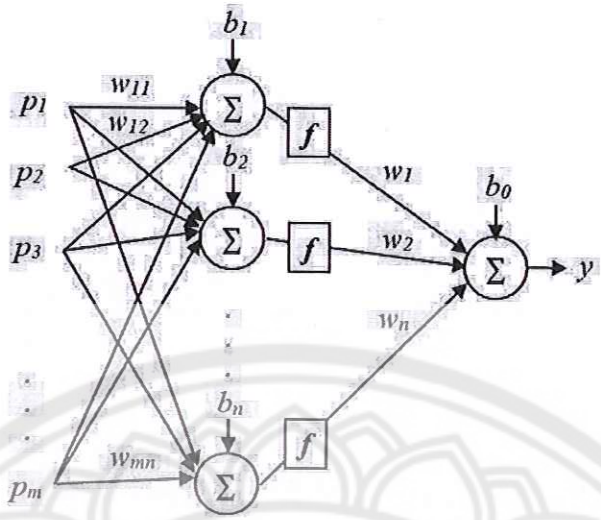


Fig. 2. ANN with multiple nodes

The entire process can be rewritten as,

$$y = \left[\sum_{i=1}^{n} w_i \times \left(f\left(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}p_i + b_i\right)\right)\right] + b_0 \quad (7)$$

, where n is the number of nodes and m is the number of inputs.

2.3 Statistical Data Analysis

The data set used in this study is secondary data of rice yield collected from 2007 to 2010 in 8 provinces in the lower northern part of Thailand. The data set consist of 12 variables, in order to conduct a reliable model, only important variables would be included in the model. In this paper the spearman rank correlation coefficient (r) was used as a criterion to select the variable for the model. The input variables that are statistically related to the rice yield at the significance level of 0.05 are presented in Table 1. There are only 6 input variables included in the development of the prediction models, plus one output, which is a rice yield measured in terms of kilogram per rai.

Table 1. The selected significant input variables and output.

Variable name	Meaning	Range	r	P-value
RiceType	Rice seed type	{1,2,3,4,5,6}	0.243	<0.001
Method	Cultivating method	{1,2,3}	0.258	<0.001
SeedPerRai	Seed quantity, kg per rai	Min=7, Max=160	0.194	<0.001
PuiPerRai	Fertilizer used, kg per rai	Min=2.5, Max=116.7	0.287	<0.001
Period	Period of cultivation	Min=3, Max=8	-0.255	<0.001
InChon	Irrigation area	{0,1}	0.128	<0.001
ProdPerRai	**Rice yield, kg per rai**	**Min=50, Max=1900, Avg=562.22**		

To build a predictive ANN model, the key parameters to be considered are number of inputs, a number of layers and number of nodes for each layer, activation function, learning rate, momentum rate for weight calculation with back propagation method and learning iterations. In this research, Weka [13] is used as a tool to create the ANN model, by varying those parameters as shown in Fig. 3. The most optimized ANN model is obtained with 13 inputs, 1 layer 6 nodes, a sigmoid function, 0.1 learning rate and 0.1 momentum rate, and 500 learning iterations. The 13 inputs are formed by 6 RiceType, 3 Method, SeedPerRai, PuiPerRai, Period and InChon.

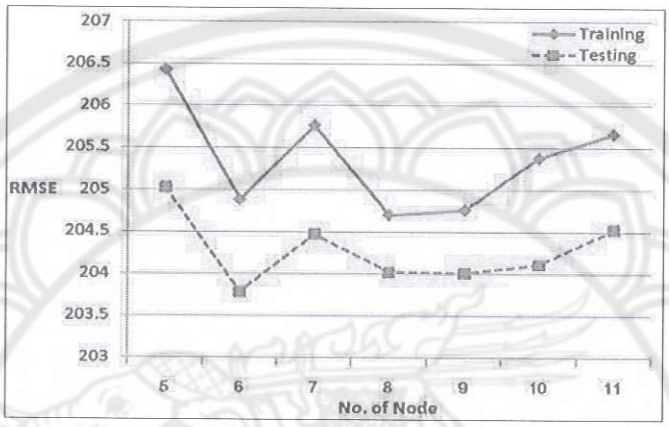


Fig. 3. RMSE values of ANN models with different number of node.

The stepwise multiple regression models have been fitted through the least squares method as described in section 2.1. The model that minimizes the RMSE is considered as the best searched model for predicting rice production.

3 Results

In this section the regression and ANN models are compared on the basis of RMSE and MAE values. The performance of each model for both of training set and test set are calculated using RMSE and MAE, defined as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{k}(y_i - \hat{y}_i)^2}{k}} \tag{8}$$

$$MAE = \frac{1}{k}\sum_{i=1}^{k}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{9}$$

, where k is the number of test points, y_i is the actual response of the i^{th} test point and $\hat{y}_i$ is the predicted response from statistical models for the i^{th} test point. Lower values for RMSE and MAE imply a more accurate prediction model.

In order to calculate the prediction accuracy of ANN and regression models, the dataset is randomly split into training set and test set with a proportion of 80:20. After the best model of each method is found, the prediction accuracy is validated and the result is presented in Table 2.

Table 2. The RMSE and MAE for both methods on training and test data.

Methods	Training data		Test data	
	RMSE	MAE	RMSE	MAE
Regression	207.05172	0.450052	204.2636	0.443842
ANN	204.88462	0.293767	203.7862	0.297208

It can be clearly seen from Table 2 that ANN performs better than regression as the RMSE and MAE values obtained from both of training and test set are lower than that of regression model. As the structure this data set is quite complex and most of the input factors are qualitative data, hence the assumption free approach like ANN is more suitable comparing to regression model.

4 Conclusions

This paper presents the performance of the two popular predictive models, regression and ANN for predicting rice production in the lower northern Thailand. The results on training and predicting the rice yield reveal that ANN performs better than regression. The ANN model is also flexible in order to set all related parameters. Furthermore ANN seems to be robust to different structure of complex data. Hence ANN model would be recommended to use for modelling rice yield especially when the information of factors is not known.

References

1. Khairunniza-Bejo, S., Mustaffha, S. and Ismail, W. I. W.: Application of Artificial Neural Network in Predicting Crop Yield: A Review. Journal of Food Science and Engineering 4: 1-9 (2014).
2. Zhange, G. et al.: Predicting with artificial neural network. International Journal of Predicting, 14, 35-62 (1998)
3. Kaul, M., Hill, R. L., Walthall, C.: Artificial neural network for corn and soybean prediction. Agricultural System, 85, 1-18 (2005)
4. Raorane, A. A. and Kulkarni, R. V.: Review-Role of Data Mining in Agriculture. International Journal of Computer Science and Information Technology. 4(2), 270-272 (2013)
5. Ji, B., Sun, Y., Yang, S., and Wan, J.: Artificial neural networks for rice yield prediction in mountainous regions. The Journal of Agricultural Science 145(03), 249-261 (2007).
6. Shabri, A., R. Samsudin, et al.: Forecasting of the rice yields time series forecasting using artificial neural network and statistical model. Journal of Applied Sciences 9(23): 4168-4173 (2009).

7. Paswan, R. P. and S. A. Begum.: Regression and Neural Networks Models for Prediction of Crop Production. International Journal of Scientific & Engineering Research 4(9), 98-108 (2013).
8. Uno, Y., Prasher, R., Laeroix, R., Goel, P.K., Karimi, A., and Viau, et al.: Artificial neural network to predict corn yield from compact airborne spectrographic imager data. Computers and Electronics in Agriculture, 47, 149-161 (2005)
9. Office of Agricultural Economics, http://www.oae.go.th
10. Montgomery, D. C., Peck, E. A. and Vining, G. G. Introduction to Linear Regression Analysis. Fifth Edition, John Wiley & Sons, New Jersey (2012).
11. Bozdogan, H. : Statistical Data Mining and Knowledge discovery, Chapman & Hall/CRC, New York (2003).
12. Ripley, B.D., Statistical aspects of neural networks. In: Barndoff-Nielsen, O.E., Jensen, J.L., Kendall, W.S., editors. Networks and chaos-statistical and probabilistic aspects, Chapman & Hall, New York, 1993, 40-123.
13. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, 11(1), (2009)

เลขทะเบียน 21

หนังสือยินยอมการเผยแพร่ผลงานทางวิชาการบนเว็บไซต์ ฐานข้อมูล NU Digital Repository (http://obj.lib.nu.ac.th/media/) สำนักหอสมุด มหาวิทยาลัยนเรศวร

ตามที่ข้าพเจ้า ผศ.ดร.อนามัย นาอุดม (ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์) ได้ส่งผลงานทางวิชาการการรายงานการวิจัย (เรื่อง) รายงานวิจัยฉบับสมบูรณ์โครงการการพัฒนาตัวแบบทางสถิติเพื่อพยากรณ์ผลผลิตข้าวในเขตภาคเหนือตอนล่าง

ปีที่พิมพ์ 2558

ข้าพเจ้าขอรับรองว่า ผลงานทางวิชาการเป็นลิขสิทธิ์ของข้าพเจ้า ผศ.ดร.อนามัย นาอุดม (ผู้วิจัยร่วม) ผศ.ดร.จรัสศรี รุ่งรัตนาอุบล เป็นเจ้าของลิขสิทธิ์ร่วม และเพื่อให้ผลงานทางวิชาการของข้าพเจ้าเป็นประโยชน์ต่อการศึกษาและสาธารณชน จึงอนุญาตให้เผยแพร่ผลงาน ดังนี้

☑ อนุญาตให้เผยแพร่

☐ ไม่อนุญาตให้เผยแพร่ เนื่องจาก.....

.....

.....

ลงชื่อ .....

(ผศ.ดร.อนามัย นาอุดม)

วันที่ 2 พฤศจิกายน 2558

หมายเหตุ ลิขสิทธิ์ใดๆ ที่ปรากฏอยู่ในผลงานนี้เป็นความรับผิดชอบของเจ้าของผลงาน ไม่ใช่ของสำนักหอสมุด