การคัดเลือกปัจจัยที่อธิบายตัวแบบโครงข่ายประสาทเทียมในการพยากรณ์การเกิดอาชญากรรม

Selection of the Attributes Describing Artificial Neural Network Models for Violent Crimes Forecasting

วรากุล กาญจนกัญโห (Varakul Kanchanakanho)* ดร.สิรภัทร เชี่ยวชาญวัฒนา (Dr.Sirapat Chiewchanwattana)**
ดร.คำรณ สุนัติ (Dr.Khamron Sunat)***

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อคัดเลือกปัจจัยสำคัญที่มีอิทธิพลต่ออัตราการเกิดเหตุอาชญากรรม และสามารถอธิบายตัวแบบของสมการพยากรณ์ของการเกิดเหตุอาชญากรรมด้วยแบบจำลองโครงข่ายประสาทเทียม (Artificial Neural Network Model) ได้เป็นอย่างดี โดยเปรียบเทียบค่าประสิทธิภาพความแม่นยำกับวิธีการทางสถิติโดยใช้การวิเคราะห์ถดถอยพหุคูณแบบขั้นตอน ชุดข้อมูล Data Set ที่ใช้ในงานวิจัยนี้เป็นข้อมูลจากการเผยแพร่ด้านสังคมและเศรษฐศาสตร์ เมื่อปี ค.ศ.1990 ข้อมูลทางด้านสถิติที่เกี่ยวข้องกับการบังคับใช้กฎหมายจากการสำรวจของ LEMAS เมื่อปี ค.ศ.1990 และจากรายงานข้อมูลด้านอาชญากรรมของ FBI UCR เมื่อปี ค.ศ.1995 ผลการวิจัยพบว่าปัจจัยที่สามารถอธิบายการเกิดเหตุอาชญากรรมได้เป็นอย่างดีมีทั้งหมด 7 ปัจจัย โดยใช้หลักวิธีเชิงพันธุกรรม (Genetic Algorithm) และการพยากรณ์ด้วยแบบจำลองโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับ (Backpropagation Algorithm) มีประสิทธิภาพความแม่นยำสูงกว่าวิธีการวิเคราะห์ทางสถิติ ซึ่งวัดค่าสัมประสิทธ์การตัดสินใจ (R-Square) ได้เท่ากับ 0.7708 และค่าคลาดเคลื่อนเฉลี่ยสัมบูรณ์น้อยที่สุด (MAE) เท่ากับ 0.1008

ABSTRACT

The objective of this paper was to extract the important attributes through artificial neural network models for forecasting the rate of violent crimes. The data set was collected by communities within the United States. The data combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. The findings showed that, artificial neural network by backpropagation algorithm for forecasting the rate of violent crimes more efficient than using statistics analysis method, and found seven attributes that affect the rate of violent crimes by selected genetic algorithm method. These attributes can explain their outcome of violent crimes at 77.08% (R-Square=0.7708), and the MAE is 0.1008.

คำสำคัญ : ปัจจัยสำคัญ การพยากรณ์ แบบจำลองโครงข่ายประสาทเทียม

Key Words : Important attributes, Forecasting, Artificial Neural Network Model

* *มหาบัณฑิต หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น*

** *ผู้ช่วยศาสตราจารย์ ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น*

*** *ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีมหานคร*

บทนำ

อาชญากรรมเป็นปรากฏการณ์อย่างหนึ่งของสังคมซึ่งอาจพิจารณาได้เป็น 2 นัย คือ พิจารณาทางด้านกฎหมายอย่างหนึ่งและพิจารณาด้านสังคมอย่างหนึ่ง อาชญากรรมทางด้านกฎหมาย คือการกระทำอันเป็นการฝ่าฝืน หรือขัดกับกฎหมายที่วางไว้ และผู้กระทำนั้นจะต้องได้รับโทษตามที่กำหนดไว้ในกฎหมาย อาชญากรรมทางด้านสังคมถือเอาความประพฤติเป็นใหญ่ ซึ่งหมายถึงการประพฤติปฏิบัติหรือพฤติกรรมที่ชั่วร้าย สร้างความเดือนร้อนให้แก่ผู้อื่นให้ได้รับความทุกข์ทรมาน หรือได้รับความเจ็บปวดทั้งร่างกายหรือจิตใจ การไม่เคารพต่อประโยชน์หรือสิทธิของบุคคลอื่น รวมทั้งการเบียดเบียนซึ่งกันและกันก็ถือเป็นอาชญากรรมทางสังคมเช่นกัน (พรชัย และคณะ, 2543)

งานวิจัยนี้ได้นำชุดข้อมูล Data Set ด้านสังคมและเศรษฐกิจเมื่อปี ค.ศ. 1990 ข้อมูลทางด้านสถิติที่เกี่ยวข้องกับการบังคับใช้กฎหมายจากการสำรวจของสถาบัน Law Enforcement Management and Administrative Statistics (LEMAS) เมื่อปี ค.ศ. 1990 และจากรายงานข้อมูลด้านอาชญากรรมของ FBI UCR เมื่อปี ค.ศ. 1995 โดยสำรวจและเก็บรวบรวมข้อมูลอาชญากรรมที่เกิดขึ้นในแต่ละชุมชนและเมืองของประเทศสหรัฐอเมริการวมทั้งหมด 1994 ชุมชน มาใช้ในการวิเคราะห์เพื่อหาปัจจัยที่สามารถอธิบายตัวแบบของสมการพยากรณ์อัตราการเกิดเหตุอาชญากรรม โดยนำเทคนิควิเคราะห์ด้วยแบบจำลองโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับ (Backpropagation Algorithm) มาใช้ในการวิเคราะห์หาตัวแบบพยากรณ์เพื่อคัดเลือกปัจจัยที่มีอิทธิพลต่อการเกิดเหตุอาชญากรรม แล้วเปรียบเทียบประสิทธิภาพความแม่นยำของผลลัพธ์ที่ได้กับเทคนิคทางสถิติ

วัตถุประสงค์ของการวิจัย

1. เพื่อวิเคราะห์หาปัจจัยที่สามารถอธิบายการเกิดอาชญากรรมได้เป็นอย่างดีทั้งวิธีการทางสถิติและโครงข่ายประสาทเทียม

2. เพื่อสร้างตัวแบบสำหรับใช้ในการพยากรณ์การเกิดอาชญากรรมด้วยเทคนิคการวิเคราะห์ที่แตกต่างกัน

3. เพื่อเปรียบเทียบประสิทธิภาพความแม่นยำของตัวแบบพยากรณ์ต่าง ๆ ที่ได้จากการวิเคราะห์ของแต่ละเทคนิค

ทฤษฎีที่เกี่ยวข้อง

การวิเคราะห์ทางสถิติ (กุณฑลี, 2546)

1) การวิเคราะห์สหสัมพันธ์ (Correlation Analysis) คือตัวสถิติที่อธิบายถึงความแข็งแรงของความสัมพันธ์ระหว่างตัวแปร 2 ตัว เช่นการเปลี่ยนแปลงของตัวแปรต้นหรือตัวแปรอิสระ (X) จะมีผลต่อการเปลี่ยนแปลงของตัวแปรตาม (Y) หรือไม่ ค่าสหสัมพันธ์ Pearson (r) สามารถคำนวณได้ดังสมการ ที่ 1

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \qquad (1)$$

การทดสอบนัยสำคัญของความสัมพันธ์ของตัวแปร

ตัวสถิติที่ใช้ในการทดสอบนัยสำคัญของความสัมพันธ์ของความสัมพันธ์ของตัวแปรต้นและตัวแปรตามคือ t-test โดยมีองศาอิสระเท่ากับ n-2 และ r มีความเกี่ยวข้องกับสมการดังนี้

$$t = \sqrt{\frac{n-2}{1-r^2}} \qquad (2)$$

2) การวิเคราะห์การถดถอยพหุคูณ (Multiple Regression Analysis) เป็นการวิเคราะห์หาความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตามเช่นเดียวกับการวิเคราะห์การถดถอยอย่างง่าย แต่มีตัวแปรต้น (X) มากกว่า 1 ตัวขึ้นไป

วิธีการวิเคราะห์ Multiple Regression Analysis มีดังนี้

Enter Method เป็นการนำตัวแปรทั้งหมดทุกตัวเข้าในสมการ (Fit Model) แล้วพิจารณาจากผลการวิเคราะห์ต่อไป

Forward Method เป็นการคัดเลือกตัวแปรอิสระเฉพาะที่มีความสัมพันธ์กับตัวแปรตามมากที่สุดเข้ามาในสมการก่อนแล้ว Fit สมการ

Backward Method เป็นการคัดเลือกตัวแปรทั้งหมดเข้ามาในสมการแล้วตัดตัวแปรที่มีความสำคัญต่อตัวแปร Y น้อยที่สุดออกจากสมการ

Stepwise Method เป็นการคัดเลือกตัวแปรอิสระที่มีความสัมพันธ์ต่อ Y มากที่สุดเข้ามาในสมการพิจารณาจาก Simple Correlation ระหว่างตัวแปรอิสระกับ Y ที่สูงที่สุด ทำการ Fit Model แล้วทำการคัดเลือกตัวแปรอิสระใหม่เข้ามาในสมการอีก

การวิเคราะห์ด้วยโครงข่ายประสาทเทียม

โครงข่ายประสาทเทียม (Artificial Neural Network) คือโมเดลทางคณิตศาสตร์สำหรับประมวลผลสารสนเทศด้วยการคำนวณแบบ Connectionist เพื่อจำลองการทำงานของเครือข่ายประสาทในสมองมนุษย์ด้วยวัตถุประสงค์ที่จะสร้างเครื่องมือซึ่งมีความสามารถในการเรียนรู้การจดจำรูปแบบ (Pattern Recognition) และการอุปมานความรู้ (Knowledge Deduction)

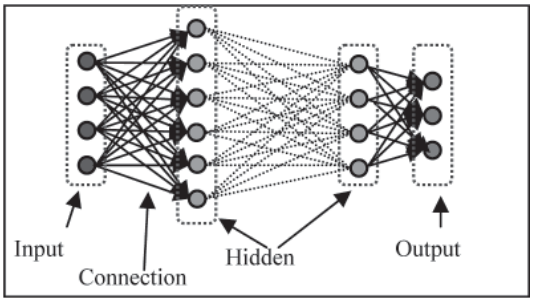
การเรียนรู้สำหรับ Neural Network (นิจรินทร์, 2552)

1. Supervised Learning การเรียนแบบมีการสอนเป็นการเรียนแบบที่มีการตรวจคำตอบเพื่อให้วงจรข่ายปรับตัว ชุดข้อมูลที่ใช้สอนวงจรข่ายจะมีคำตอบไว้คอยตรวจดูว่าวงจรข่ายให้คำตอบที่ถูกหรือไม่ ถ้าตอบไม่ถูกวงจรข่ายก็จะปรับตัวเองเพื่อให้ได้คำตอบที่ดีขึ้น

2. Unsupervised Learning การเรียนแบบไม่มีการสอนเป็นการเรียนแบบไม่มีผู้แนะนำ ไม่มีการตรวจคำตอบว่าถูกหรือผิด เป็นการเรียนรู้โดยมีเฉพาะค่า Input เท่านั้นที่ใส่เข้าไปในระบบ โดยจะไม่มีการเรียนรู้จากค่า Output เพื่อนำไปใช้ปรับค่าพารามิเตอร์ต่าง ๆ ของระบบโครงข่ายประสาทเทียม

Backpropagation Algorithm

ขั้นตอนการเรียนรู้แบบแพร่กระจายย้อนกลับเป็นอัลกอริทึมที่ใช้ในการเรียนรู้ของเครือข่ายใยประสาทวิธีหนึ่งที่นิยมใช้ใน Multi-Layer Perceptron (MLP) เพื่อปรับค่าน้ำหนักในเส้นเชื่อมต่อระหว่างโหนดให้เหมาะสม โดยการปรับค่านี้จะขึ้นกับความแตกต่างของค่า Output ที่คำนวณได้กับค่า Output ที่ต้องการ (Dayhoff, 1990)



ภาพที่ 1 โครงสร้าง Backpropagation Neural Network

การทำงานของโครงข่ายประสาทแบบแพร่กระจายย้อนกลับ (Backpropagation Neural Network) เป็นวิธีการฝึกสอนโครงข่ายประสาทหลายชั้นแบบส่งผ่านไปข้างหน้า (Multilayer Feed Forward Neural Network) ที่นิยมใช้กันซึ่งประกอบด้วยชั้นข้อมูลนำเข้า (Input Layer) ชั้นซ่อน (Hidden Layer) และชั้นผลลัพธ์ (Output Layer) โดยมีลักษณะการส่งผ่านข้อมูลแบบส่งผ่านไปข้างหน้าตามลำดับชั้น และเรียนรู้ผ่านฟังก์ชั่นการเรียนรู้จนได้เป็นผลลัพธ์ออกมา จากนั้นจึงนำค่าผลลัพธ์ที่ได้มาเปรียบเทียบกับผลลัพธ์ที่ต้องการเพื่อหาค่าความคลาดเคลื่อน (Error) โดยใช้ค่าคลาดเคลื่อนกำลังสองที่น้อยที่สุด (MSE)

อัลกอริทึมปรับค่าน้ำหนักมี 4 วิธี ดังนี้

1. Momentum (MOM) เป็นขั้นตอนของเริ่มต้นใช้ค่า η มาก ๆ และลดลงในระหว่างการสอน ดังนั้นค่าน้ำหนักจะถูกปรับโดยสมการที่ 3

$$\Delta wji(k+1) = \eta\delta_j a_i + \alpha[\Delta wji(k)] \qquad (3)$$

เมื่อ k คือดัชนีเวลา (Time Index) หรือจำนวนรอบของการปรับปรุงค่าน้ำหนัก

2. Quick Propagation (QP) เป็นขั้นตอนวิธีการเรียนรู้แบบปรับค่าน้ำหนักที่ใช้เวลาน้อยแบบต้องมีผู้สอน แตกต่างจาก Back Propagation ที่วิธีการหาค่าน้ำหนัก ดังสมการที่ 4

$$\Delta w(u) = \frac{f(u)}{f(u-1) - f(u)} \Delta w(u-1) \qquad (4)$$

เมื่อ $f(u)$ แทน ค่าน้ำหนักรอบปัจจุบัน

แทน $f(u-1)$ ค่าน้ำหนักรอบก่อนหน้า

3. Conjugate Gradient Descent (CGD) เป็นการรวมเทคนิคของการหาค่าน้อยที่สุดกับขั้นตอนวิธีการเคลื่อนลงตามความชันเข้าด้วยกัน โดยเคลื่อนที่ไปในทิศทางตรงกันข้ามกับค่าความชัน ดังสมการที่ 5

$$p(0) = -g(0) \qquad (5)$$

โดยการเคลื่อนลงนั้นจะมีลักษณะเป็นเส้นซิกแซก เพื่อต้องการให้เส้นซิกแซกนั้นเรียบขึ้นจึงได้แทนที่ขั้นตอนวิธีการเคลื่อนลงตามความชันด้วยสมการที่ 6

$$p(t+1) = -g(t) + \beta(t)\, p(t) \qquad (6)$$

เมื่อค่าสเกลาร์ $\beta(t)$ ถูกเลือก $p(t+1)$ และ $p(t)$ จับคู่กันโดยใช้การประมาณ Hessian Matrix จะได้ว่า

$$p(t+1).H.p^{T}(t) = 0 \qquad (7)$$

ขั้นตอนวิธีนี้สามารถหลีกเลี่ยงการเกิดการเคลื่อนที่แบบเส้นซิกแซกสำหรับการค้นหาทิศทางในรอบต่อไป

4. Levenberg-Marquardt (LM) เป็นขั้นตอนวิธีสำหรับการเรียนรู้แบบปรับค่าน้ำหนัก (Weight) ของโครงข่ายประสาทเทียม โดยการหาค่าพารามิเตอร์ของสมการไม่เชิงเส้น เพื่อลดผลรวมกำลังสองของค่าผิดพลาดลง (Sum of Square Error) โดยการใช้การเรียนรู้แบบปรับค่าน้ำหนักแบบ Linear Descent เพื่อให้ระบบลดค่าผิดพลาดกำลังสองลงให้เหลือน้อยที่สุด

$$H(w) = J^{T}(w)J(w) + \mu I \qquad (8)$$

เมื่อ $H(w)$ แทน Batch Hessian

$J(w)$ แทน Jacobian Matrices

I แทน Identity Matrix

μ แทน ค่าคงที่

การคัดเลือกปัจจัยที่มีอิทธิพลต่อตัวแปรตาม

การคัดเลือกปัจจัยที่มีอิทธิพลต่อตัวแปรตามซึ่งก็คืออัตราการเกิดเหตุอาชญากรรมใช้หลักการของ Generalized Regression Neural Network (GRNN) ซึ่งมีหลักการทำงานคือ การคัดเลือกหรือหาชุดของค่ากลางที่เหมาะสม เพื่อทำให้โครงข่ายให้ค่าผิดพลาดของการประมาณต่ำ เทคนิคการหาค่าที่ดีที่สุด (Optimization Techniques) ซึ่งในงานวิจัยได้เลือกใช้วิธีการคัดเลือกปัจจัยที่อาศัยหลักการของ GRNN มี 4 วิธี ดังนี้

1. Forward Stepwise เป็นวิธีการเลือกชุดข้อมูลครั้งละ 1 ตัว โดยเลือกปัจจัยที่มีความสัมพันธ์กับตัวแปรตามมากที่สุดเข้าสู่โครงข่าย จนกระทั่งไม่มีปัจจัยตัวใดที่มีความสัมพันธ์กับตัวแปรตาม

2. Backward Stepwise เป็นวิธีการเลือกปัจจัยเข้าสู่โครงข่าย โดยที่ขั้นแรกจะนำปัจจัยทุกตัวที่คาดว่าจะมีความสัมพันธ์กับตัวแปรตามเข้าสู่โครงข่าย จากนั้นตัดปัจจัยที่ไม่มีความสัมพันธ์กับตัวแปรตามออกจากสมการครั้งละตัว จนกระทั่งไม่สามารถตัดปัจจัยใดได้อีก

3. Exhaustive Search เป็นวิธีการค้นหาคุณลักษณะที่เป็นไปได้ของชุดข้อมูลนำเข้า แล้วคัดเลือกชุดข้อมูลที่มีคุณลักษณะดีที่สุด

4. Genetic Algorithm : GA การแก้ปัญหาโดยวิธีเชิงพันธุกรรม เพื่อให้ได้จุดที่เหมาะสมที่สุด เป็นเทคนิคการสุ่มค้นปัญหาที่สลับซับซ้อน ซึ่งในปัญหาการหาค่าสูงสุดหรือต่ำสุดของฟังก์ชันนั้น Search Space อาจจะมีค่าที่ไม่ต่อเนื่องและบ่อยครั้งที่มีค่าสูงสุดหรือต่ำสุดหลายค่า GA จะค้นหาค่าสูงสุดหรือต่ำสุดของฟังก์ชันโดยใช้กลุ่มของค่าเริ่มต้นกลุ่มหนึ่งที่สุ่มมาได้แล้วปีนหาค่าสูงสุดหรือต่ำสุดใน Search Space พร้อมกันทั้งกลุ่ม

งานวิจัยที่เกี่ยวข้อง

Redmond and Baveja (2002) ได้ทำการศึกษาและพัฒนาซอฟท์แวร์ระบบ Crime Similarity System (CSS) มาช่วยในการจัดเก็บข้อมูลและ

แลกเปลี่ยนข้อมูลด้านเศรษฐศาสตร์ ด้านอาชญากรรม และด้านเกี่ยวกับการบังคับคดีของเมืองต่างๆ ซึ่งประกอบด้วยข้อมูลของแต่ละชุมชน ข้อมูลที่จัดเก็บจะถูกนำไปใช้ในการวิจัยประกอบการกำหนดแผนแม่บท และสนับสนุนระบบงานด้านฐานข้อมูล

นิจรินทร์ (2552) ศึกษาแบบจำลองโครงข่ายประสาทเทียมสำหรับการวิเคราะห์คุณภาพบัณฑิต โดยใช้โครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับ พบว่ามีประสิทธิภาพดีกว่าการวิเคราะห์ทางสถิติ โดยวัดประสิทธิภาพจากค่าสัมประสิทธิ์การตัดสินใจ (R-Square : R^2) และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE)

วิธีการดำเนินการวิจัย

ข้อมูลประชากรและกลุ่มตัวอย่าง

ข้อมูลที่ใช้ในการวิจัยนี้เป็นข้อมูลที่ได้จากการสำรวจจาก Community ซึ่งประกอบไปด้วยชุมชน เมือง และรัฐ ต่างๆ ในสหรัฐอเมริกา รวมจำนวนทั้งสิ้น 1994 Community และปัจจัยต่างๆ ที่มีอิทธิพลต่อการเกิดเหตุอาชญากรรมรุนแรงรวมทั้งสิ้น 123 ปัจจัย โดยเป็นปัจจัยที่นำมาเป็นตัวแปรอิสระจำนวน 122 ปัจจัย และเป็นตัวแปรตามอีก 1 ปัจจัย ข้อมูลที่จะนำไปใช้ในการวิเคราะห์ได้นำมาผ่านกระบวนการของ Data Cleaning ซึ่งมี 2 วิธี รายละเอียดดังนี้

วิธีที่ 1 ตรวจสอบข้อมูลในแต่ละ Community ถ้าหากพบว่าแถวใดมีค่า Missing Value จะตัด Community นั้นออกไปทั้งแถวจึงทำให้เหลือ Community เพียง 319 ส่วนจำนวนปัจจัยยังคงเท่าเดิม

วิธีที่ 2 ตรวจสอบข้อมูลในแต่ละปัจจัย ถ้าหากพบว่าปัจจัยใดมีค่า Missing Value จะตัดปัจจัยนั้นออกไปทั้งคอลัมน์จึงทำให้เหลือปัจจัยเพียง 101 ปัจจัย โดยเป็นปัจจัยที่ใช้เป็นตัวแปรอิสระ 100 ปัจจัย และตัวแปรตามอีก 1 ปัจจัย ส่วน Community มีจำนวนเท่าเดิม

ขั้นตอนการวิเคราะห์ข้อมูล

ขั้นตอนที่ 1 คัดเลือกกลุ่มปัจจัยที่มีค่าสหสัมพันธ์ต่อการเกิดอาชญากรรมสูง แล้วทำการทดสอบนัยสำคัญทางสถิติด้วย T-Test เพื่อหาทิศทางของความสัมพันธ์

ขั้นตอนที่ 2 คัดเลือกกลุ่มปัจจัยที่มีค่าสหสัมพันธ์สูงมาแบ่งชุดข้อมูลสำหรับการวิเคราะห์ออกเป็น 2 ชุด โดยชุดที่ 1 ประกอบด้วยข้อมูล Train: Validation:Test เป็น 80 % : 5% : 15% และชุดที่ 2 เป็น 90% : 5% : 5%

ขั้นตอนที่ 3 คัดเลือกปัจจัยที่อธิบายการเกิดเหตุอาชญากรรมด้วยวิธีทางสถิติโดยวิธีการวิเคราะห์ถดถอยพหุคูณแบบขั้นตอน (Stepwise Multiple Regression Analysis) และโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับ (Backpropagation Algorithm)

ขั้นตอนที่ 4 สร้างตัวแบบที่จะใช้เป็นสมการพยากรณ์ด้วยข้อมูลเริ่มต้น

ขั้นตอนที่ 5 ทำการพยากรณ์โดยใช้วิธีต่างๆ เพื่อพยากรณ์ค่าจากข้อมูลทดสอบ

การวัดประสิทธิภาพความแม่นยำของการพยากรณ์

1. ค่าสัมบูรณ์ของค่าคลาดเคลื่อนเฉลี่ย (Mean Absolute Error : MAE) ของชุดข้อมูลทดสอบ

$$MAE = \frac{\sum_{i=1}^{n} |e_i|}{n} \qquad (9)$$

เมื่อ e_i แทน ผลต่างระหว่างค่าข้อมูลจริงและค่าพยากรณ์

n แทน ข้อมูลในการพยากรณ์

2. ความแม่นยำของการพยากรณ์วัดจากค่าสัมประสิทธิ์การตัดสินใจ (R-Square : R^2)

ผลการทดลอง

การคัดเลือกปัจจัยโดยโครงข่ายประสาทเทียม

ผลการคัดเลือกปัจจัยที่เหมาะสมที่สามารถอธิบายการเกิดเหตุอาชญากรรมได้เป็นอย่างดี ปรากฏผลการทดลองดังตารางที่ 1

ตารางที่ 1 ผลการคัดเลือกปัจจัยที่มีผลต่ออัตราการเกิดเหตุอาชญากรรมด้วยหลักการของ GRNN

ผลการคัดเลือกปัจจัย	วิธีการคัดเลือก	ค่าความเหมาะสม (Fitness)
ชุดที่ 1 PctPopUnderPov, PctKids2Par, PctHousNoPhone	Forward Stepwise	45.9190
ชุดที่ 2 PctPopUnderPov, FemalePctDiv, TotalPctDiv, PctKids2Par, PctYoungKids2Par, PctTeen2Par, PctHousNoPhone	Backward Stepwise	61.9512
ชุดที่ 3 medIncome, medFamInc, perCapInc, PctUnemployed, PctYoungKids2Par, PctPersOwnOccup, RacialMatchCommPol	Genetic Algorithm	61.9915

จากตารางที่ 1 พบว่า ปัจจัยชุดที่ 3 ให้ค่าความเหมาะสมมากที่สุดเท่ากับ 61.9915 ด้วยวิธี Genetic Algorithm เมื่อกำหนดความน่าจะเป็นของการข้ามสายพันธุ์ (Crossover rate)ที่ 80% จำนวนปัจจัยที่ถูกคัดเลือกทั้งหมด 7 ปัจจัย ได้แก่ ปัจจัยครอบครัว (ไม่มีเงื่อนไขทางสายเลือด) ที่มีฐานะปานกลาง (medIncome) ปัจจัยจำนวนครอบครัว (มีเงื่อนไขสายเลือดเดียวกัน) ที่มีฐานะปานกลาง (medFamInc) ปัจจัยจำนวนรายได้ต่อเมือง (perCapInc) ปัจจัยจำนวนเปอร์เซ็นต์ของประชากรที่ใช้แรงงานและว่างงานซึ่งมีอายุตั้งแต่ 16 ปี ขึ้นไป (PctUnemployed) ปัจจัยจำนวนเปอร์เซ็นต์ของเด็กที่มีอายุ 4 ปี และอยู่ภายใต้การดูแลของ 2 ครอบครัว (PctYoungKids-2Par) ปัจจัยเปอร์เซ็นต์ของประชากรที่พักอาศัยอยู่บ้านของตนเอง (PctPersOwnOccup) และปัจจัยที่แสดงถึงสัดส่วนความสัมพันธ์ระหว่างจำนวนเจ้าหน้าที่ตำรวจที่ปฏิบัติหน้าที่ต่อขนาดชุมชนแต่ละแห่ง (RacialMatchCommPol)

การวัดประสิทธิภาพความแม่นยำของการพยากรณ์

ปัจจัยที่คัดเลือกได้ทั้ง 3 ชุด จะนำเข้าสู่กระบวนการพยากรณ์อัตราการเกิดเหตุอาชญากรรมโดยใช้โครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับ (Backpropagation Algorithm) เพื่อเลือกจำนวนชั้นซ่อนและจำนวนโหนดในชั้นซ่อนที่เหมาะสมกับชุดข้อมูล แล้วสร้างแบบจำลองของทั้ง 3 ชุดปัจจัย โดยวัดจากค่าความเหมาะสม (Fitness) ผลการทดลองหาจำนวนโหนดในชั้นซ่อนและแบบจำลองของชุดปัจจัยที่เหมาะสมสำหรับการเรียนรู้ปรับค่าน้ำหนักระหว่างหน่วยประมวลผลทั้ง 4 วิธี ผลการทดลองดังตารางที่ 2

ตารางที่ 2 แบบจำลองของชุดปัจจัยซึ่งวัดจากค่าความเหมาะสม (Fitness)

ปัจจัย	Hidden	Input	Node	Output	แบบจำลอง	Fitness
ชุดที่ 1	1	3	4	1	[3-4-1]	4.4183
ชุดที่ 2	1	7	13	1	[7-13-1]	5.8105
ชุดที่ 3	1	7	18	1	[7-18-1]	5.4901

จากตารางที่ 2 เป็นการหาแบบจำลองที่ให้ค่าความเหมาะสมมากที่สุดของแต่ละชุดปัจจัย โดยคัดเลือกจำนวนชั้นซ่อน และจำนวนโหนดในชั้นซ่อน พบว่าปัจจัยชุดที่ 1 มีแบบจำลองโครงข่ายเป็น [3-4-1] ปัจจัยชุดที่ 2 มีแบบจำลองโครงข่ายเป็น [7-13-1] และปัจจัยชุดที่ 3 มีแบบจำลองโครงข่ายเป็น [7-18-1] จากนั้นนำแบบจำลองทั้งหมดไปทำการทดลองปรับค่าน้ำหนักระหว่างหน่วยประมวลผลด้วยอัลกอริทึมทั้ง 4 วิธี ได้แก่ Quick Propagation, Conjugate Gradient Descent, Levenberg-Marquardt และ Momentum โดยแบ่งสัดส่วนชุดข้อมูล Train: Validation:Test เป็น 2 แบบ ได้แก่ 80:5:15 และ 90:5:5 ได้ผลการคัดเลือกดังนี้

ตารางที่ 3 แสดงผลการคัดเลือกแบบจำลองที่มีค่าสัมประสิทธิ์การตัดสินใจมากที่สุด

แบบจำลอง	R-Square	MAE
[3-4-1]	0.6397	0.1344
[7-13-1]	0.6783	0.1279
[7-18-1]	0.7708	0.1008

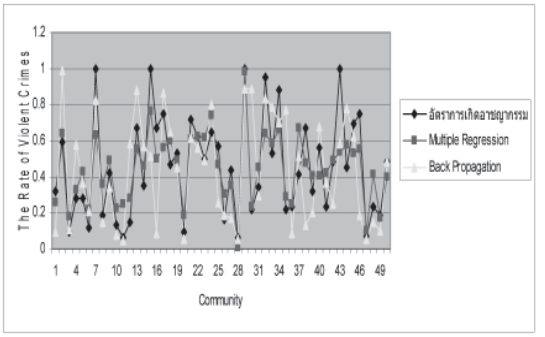
จากตารางที่ 3 ผลการพยากรณ์จากชุดข้อมูลปัจจัยชุดที่ 3 พบว่าแบบจำลองโครงข่าย [7-18-1] ซึ่งใช้การปรับค่าน้ำหนักด้วยวิธี Quick Propagation (QP) และแบ่งสัดส่วนเปอร์เซ็นต์ของข้อมูล Train:Validation:Test เป็น 80:5:15 สามารถวัดค่าสัมบูรณ์ของค่าคลาดเคลื่อนเฉลี่ยน้อยที่สุดเท่ากับ 0.1008 และค่าสัมประสิทธิ์การตัดสินใจ (R-Square) มากที่สุดเท่ากับ 0.7708

ประสิทธิภาพความแม่นยำระหว่างการวิเคราะห์ด้วยโครงข่ายประสาทเทียมและวิธีการทางสถิติแสดงการเปรียบเทียบดังปรากฏในตารางที่ 4

ตารางที่ 4 เปรียบเทียบประสิทธิภาพผลของการคัดเลือกปัจจัยด้วย GRNN และวิธีทางสถิติ

วิธีการพยากรณ์	R-Square	MAE
Stepwise Multiple Regression Analysis	0.6520	0.1010
Artificial Neural Network (Back Propagation Algorithm)	0.7708	0.1008

ประสิทธิภาพความแม่นยำจากวิธีการทางสถิติและโครงข่ายประสาทเทียม สามารถพล็อตเป็นกราฟเปรียบเทียบแสดงดังปรากฏในภาพที่ 2



ภาพที่ 2 กราฟแสดงการเปรียบเทียบประสิทธิภาพของการพยากรณ์อัตราการเกิดเหตุอาชญากรรม

จากตารางที่ 4 และภาพที่ 2 เป็นผลการทดลองจากการนำข้อมูลที่ผ่านการคัดเลือกเบื้องต้นด้วยวิธีที่ 1 มาทำการวิเคราะห์ พบว่าผลการพยากรณ์ชุดข้อมูลทดสอบด้วยแบบจำลองโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับให้ค่าสัมประสิทธิ์การตัดสินใจ (R-Square) เท่ากับ 0.7708 และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยเท่ากับ 0.1008 ส่วนวิธีการวิเคราะห์ถดถอยพหุคูณแบบขั้นตอน ให้ค่าสัมประสิทธิ์การตัดสินใจ (R-Square) เท่ากับ 0.6520 และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) เท่ากับ 0.1010

ปัจจัยที่ถูกคัดเลือกโดยใช้หลักการของ GRNN มีทั้งหมด 7 ปัจจัย ส่วนการคัดเลือกโดยใช้หลักสถิติมีทั้งหมด 9 ปัจจัย และเปรียบเทียบความสัมพันธ์ของแต่ละปัจจัยรวมทั้งประสิทธิภาพของการพยากรณ์ดังปรากฏในตารางที่ 5

ตารางที่ 5 เปรียบเทียบปัจจัยที่ถูกคัดเลือกโดย GA และวิธีการวิเคราะห์ถดถอยพหุคูณแบบขั้นตอน

	วิธีที่ใช้ในการพยากรณ์	
	Genetic Algorithm	Stepwise Multiple Regression Analysis
ปัจจัยที่ถูกคัดเลือก	1. medIncome	1. PctKids2Par
	2. medFamInc	2. HousVacant
	3. perCapInc	3. PctPersDenseHous
	4. PctUnemployed	4. medIncome
	5. PctYoungKids2Par	5. racePctBlack
	6. PctPersOwnOccup	6. MalePctDivorce
	7. RacialMatchCommPol	7. PctIlleg
		8. PctVacantBoarded
		9. NumUnderPov
RSquare	0.7708	0.6520
MAE	0.1008	0.1010

จากตารางที่ 5 พบว่าประสิทธิภาพความแม่นยำของการพยากรณ์ด้วยวิธีโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับมีประสิทธิภาพในการพยากรณ์อัตราการเกิดเกิดอาชญากรรมได้ถึง 77.08 %

สรุปผลการวิจัย

งานวิจัยนี้ได้ศึกษาและคัดเลือกปัจจัยที่มีผลอิทธิพลต่อการเกิดอาชญากรรม ด้วยหลักการของ GRNN ซึ่งมีทั้งหมด 7 ปัจจัย และคัดเลือกจำนวนชั้นซ่อนที่เหมาะสมได้ 1 ชั้นซ่อน 18 โหนดในชั้นซ่อน ด้วยอัลกอริทึม Quick Propagation ปรับค่าน้ำหนักระหว่างประมวลผล ตัวแบบจำลอง คือ [7-18-1] และสามารถสรุปได้ 2 ประเด็นดังนี้

1. การเปรียบเทียบความสามารถในการพยากรณ์การเกิดอาชญากรรมด้วยโครงข่ายประสาทเทียม และวิธีการวิเคราะห์ถดถอยพหุคูณ พบว่าการพยากรณ์ด้วยโครงข่ายประสาทเทียมมีความสามารถในการพยากรณ์การเกิดเหตุอาชญากรรมได้ประมาณ 77.08% มีค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยเท่ากับ 0.1008 ซึ่งถือว่ามีประสิทธิภาพในการอธิบายการเกิดอาชญากรรมได้สูงกว่าวิธีทางด้านสถิติ เนื่องจากวิธีการวิเคราะห์ถดถอยพหุคูณมีข้อกำหนดเบื้องต้นที่เคร่งครัด แต่โครงข่ายประสาทเทียมไม่มีข้อกำหนดเบื้องต้น มีเพียงชั้นซ่อนที่สามารถเรียนรู้และปรับค่าในหน่วยประมวลผลของข้อมูลนำเข้าก่อนที่จะส่งผลลัพธ์ออกมา

2. ปัจจัยที่มีอิทธิพลต่อการเกิดอาชญากรรม ได้แก่ ปัจจัยทางด้านครอบครัวที่มีฐานะปานกลาง (ซึ่งแบ่งเป็นครอบครัวโดยสายเลือดและไม่มีเงื่อนไขทางสายเลือด) ปัจจัยด้านจำนวนรายได้ต่อเมือง ปัจจัยด้านเปอร์เซ็นต์ของประชากรที่ใช้แรงงานและว่างงานซึ่งมีอายุตั้งแต่ 16 ปีขึ้นไป ปัจจัยด้านเปอร์เซ็นต์ของเด็กที่มีอายุ 4 ปี และอยู่ภายใต้การดูแลของ 2 ครอบครัว ปัจจัยด้านเปอร์เซ็นต์ของประชากรที่พักอาศัยอยู่บ้านของตนเอง และปัจจัยด้านสัดส่วนความสัมพันธ์ระหว่างจำนวนเจ้าหน้าที่ตำรวจต่อขนาดชุมชนแต่ละแห่ง ซึ่งจากการวิเคราะห์ปัจจัยทั้งหมดด้วยเส้นแนวโน้มของการเกิดอาชญากรรมต่อแต่ละปัจจัยแล้วพบว่ามีเพียง 1 ปัจจัยคือปัจจัยด้านเปอร์เซ็นต์ของประชากรที่ใช้แรงงานและว่างงานซึ่งมีอายุตั้งแต่ 16 ปี ขึ้นไป มีความสัมพันธ์เชิงบวกต่อการเกิดเหตุอาชญากรรมนั่นคือ ยิ่งปัจจัยนี้มีปริมาณเพิ่มขึ้นเท่าไหร่แล้วจะส่งผลให้อัตราการเกิดเหตุอาชญากรรมเพิ่มมากขึ้นด้วย ส่วนอีก 6 ปัจจัยที่เหลือพบว่าเส้นแนวโน้มมีความสัมพันธ์เชิงลบต่อการเกิดอาชญากรรมนั่นคือ ยิ่งปัจจัยเหล่านี้มีจำนวนเพิ่มมากขึ้นเท่าไหร่แล้วจะทำให้การเกิดอาชญากรรมลดน้อยลงเท่านั้น

ข้อเสนอแนะ

1. ในการคัดเลือกปัจจัยเห็นได้ว่าประสิทธิภาพของการพยากรณ์อยู่ในเกณฑ์ค่อนข้างสูงซึ่งวัดค่าสัมประสิทธิ์การตัดสินใจได้เท่ากับ 0.7708 ซึ่งยังไม่อยู่ในเกณฑ์ที่สูงพอ ทั้งนี้เนื่องมาจากปัจจัยที่นำมาวิเคราะห์มีเฉพาะปัจจัยที่เกี่ยวข้องกับเฉพาะด้านสังคมและเศรษฐกิจ เท่านั้น ดังนั้นข้อมูลที่นำมาทำการวิจัยในครั้งต่อไป หากมีการพิจารณานำปัจจัยทางด้านศาสนา ปัจจัยด้านการเมือง หรือปัจจัยอื่น ๆ ที่เกี่ยวข้องมาพิจารณาร่วมด้วย จะทำให้ผลลัพธ์น่าเชื่อถือและเพิ่มประสิทธิภาพของแบบจำลองมากยิ่งขึ้น

2. รูปแบบการจัดเก็บฐานข้อมูลอาชญากรรมและกระบวนการวิจัย สามารถนำมาประยุกต์ใช้เป็นกรณีศึกษาเพื่อให้บรรลุเป้าหมายตามแผนยุทธศาสตร์ของแผนแม่บทเทคโนโลยีสารสนเทศและการสื่อสาร (ฉบับที่ 2) สำนักงานตำรวจแห่งชาติ พ.ศ. 2552 - 2556

เอกสารอ้างอิง

กุณฑลี เวชสาร. 2546. การวิจัยการตลาด. พิมพ์ครั้งที่ 4. กรุงเทพฯ: จุฬาลงกรณ์ มหาวิทยาลัย.

นิจรินทร์ ชูราษี. 2552. แบบจำลองโครงข่ายประสาทเทียมสำหรับการวิเคราะห์คุณภาพบัณฑิต. วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ. บัณฑิตวิทยาลัย มหาวิทยาลัยขอนแก่น.

พรชัย ขันตี, ธัชชัย ปิตะนิละบุตร และอัศวิน วัฒนวิบูลย์. 2543. ทฤษฎีและงานวิจัยทางอาชญาวิทยา. กรุงเทพฯ: บริษัทบุ๊คเน็ท จำกัด.

Campolo, M., Andreussi, P., Sodati, A. 1997. River flood forecasting with a neural network model. Water Resources Research. 35(4): 1191-1197.

Dayhoff, JE. 1991. Neural Network Architectures: An Introduction. New York: Vaan Nostrand.

Gorr, W., Olligschlaeger, A., Thompson, Y. 2003. Short-term forecasting of crime. International Journal of Forecasting. 19(4): 579-594.

Maier, HR., Dandy, GC. 2000. Neural networks for the prediction and forecasting of water resources variables. Environmental Modelling & Software. 15: 101-124.

Redmond, M., Baveja, A. 2002. A data-driven software tool for enabling cooperative. European Journal of Operational Research. 141: 660-678.