การเปรียบเทียบประสิทธิภาพการลดมิติข้อมูลและจำแนกข้อมูลโดยวิธีการทางเครือข่ายประสาทเทียม

A Comparative Efficiency of Dimensionality Reduction and Neural Network Classification

ภรัณยา อำมฤครัตน์ (Paranya Ammaruekarat) * ดร.พยุง มีสัจ (Dr.Phayung Meesad) **

บทคัดย่อ

งานวิจัยนี้นำเสนอการเปรียบเทียบประสิทธิภาพของแบบจำลองในการลดมิติข้อมูลและจำแนกข้อมูล ซึ่งใช้ข้อมูลตัวอย่าง จากฐานข้อมูล UCI Machine Learning Database Repository ได้แก่ Ozone , Ionosphere และ Sonar นำมาวิเคราะห์ทำการเปรียบเทียบประสิทธิภาพ โดยใช้วิธีการลดมิติข้อมูล (Dimensionality Reduction) แบบ PCA (Principal Components Analysis) และ CFS (Correlation-based Feature Selection) ร่วมกับวิธีการจำแนกข้อมูลแบบโครงข่ายประสาทเทียมแบบ Multi-Layer Perceptron (MLP) เปรียบเทียบกับซัพพอร์ตเวกเตอร์ แมชชีน (Support Vector Machines : SVM)

การวัดประสิทธิภาพสามารถวัดได้จากความถูกต้องของการจำแนกประเภทของข้อมูลโดยนับจากค่า ความถูกต้องของการจำแนกประเภทข้อมูลที่วัดได้ ซึ่งการทดสอบแบบจำลองที่ได้จะทำการทดสอบผลบนพื้นฐานวิธี 5 - fold Cross Validation โดยผลการทดลองที่ได้ พบว่า วิธีการลดมิติข้อมูลแบบ CFS ร่วมกับวิธีการจำแนกข้อมูลแบบ MLP เข้ามาใช้ในการจำแนกข้อมูลนั้นจะมีประสิทธิภาพที่ดีกว่าการใช้โมเดลแบบอื่นๆ

ABSTRACT

This paper represented comparing efficiency of model dimensionality reduction and classification between PCA (Principal Components Analysis) and CFS (Correlation-based Feature Selection) combine Artificial Neural Network (Multi-layer Perceptron : MLP) classifier comparing Support Vector Machine (SVM) using Ozone data set , Ionosphere data set and Sonar data set from UCI Machine Learning Database Repository. The accuracy rate of classification is used for evaluating efficiency. Moreover, the 5-fold cross validation is used to testing model. The result of experiment shows CFS combine MLP for classification that high efficiency more than PCA combine MLP, CFS combine SVM and PCA combine SVM.

คำสำคัญ : การจำแนกประเภทข้อมูล โครงข่ายประสาทเทียม ซัพพอร์ตเวกเตอร์แมชชีน การลดมิติข้อมูล

Key words : Classification, Multi-Layer Perceptron, Support VectorMachines, Dimensionality Reduction

---

* *นักศึกษาดุษฎีบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ*

** *ผู้ช่วยศาสตราจารย์ สาขาวิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ*

บทนำ

ในสังคมในยุคปัจจุบันมีการแข่งขันกันสูงในทุกๆด้านไม่ว่าจะเป็นการแข่งขันทางด้านเศรษฐกิจ การแข่งขันกันในการเรียนรู้ และแนวโน้มของการนำสารสนเทศมาประกอบการตัดสินใจในงานสาขาต่าง ๆ มีมากขึ้นแต่บางครั้งไม่สามารถสร้างสารสนเทศที่ตรงกับความต้องการขององค์กรได้ ซึ่งในองค์กรต่าง ๆ ส่วนใหญ่ได้มีการเก็บข้อมูลไว้เป็นจำนวนมาก โดยที่ข้อมูลเหล่านี้สามารถนำมาใช้ประโยชน์ได้มากแต่ไม่ค่อยได้ถูกนำมาใช้อย่างจริงจัง การทำเหมืองข้อมูล (Data Mining) เป็นวิธีการหนึ่งที่สามารถนำมาข้อมูลเหล่านั้นมาใช้ให้เกิดประโยชน์

เทคนิคการจำแนกประเภทข้อมูล (Data classification) (Jiawei and Micheline, 2001) เป็นเทคนิคหนึ่งที่สำคัญของการสืบค้นความรู้บนฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from very large Database: KDD) หรือดาต้าไมน์นิง จุดประสงค์ของการจำแนกประเภทข้อมูลคือการสร้างโมเดลการแยกแอทริบิวท์หนึ่งโดยขึ้นกับแอทริบิวท์อื่น โมเดลที่ได้จากการจำแนกประเภทข้อมูลจะทำให้สามารถพิจารณาคลาสในข้อมูลที่ยังมิได้แบ่งกลุ่มในอนาคตได้ เทคนิคการจำแนกประเภทข้อมูลนี้ได้นำไปประยุกต์ใช้ในหลายด้านเช่น การจัดกลุ่มลูกค้าทางการตลาด การตรวจสอบความผิดปกติ และการวิเคราะห์ทางการแพทย์ เป็นต้น

เทคนิคการลดขนาดข้อมูล (Data Reduction) (ธรรมศักดิ์, 2548) การลดขนาดข้อมูลเป็นกระบวนการหนึ่งในขั้นตอนการเตรียมข้อมูล นั่นคือการทำให้ข้อมูลตั้งต้นมีขนาดลดลงโดยสูญเสียลักษณะสำคัญของข้อมูลน้อยที่สุดและสูญเสียความถูกต้องของผลลัพธ์น้อยที่สุด เนื่องจากข้อมูลแต่ละตัวจะมีความสำคัญต่อการจัดกลุ่มข้อมูลไม่เท่ากัน ด้วยเทคนิคการเลือกข้อมูลที่ดีจะทำให้สามารถเลือกข้อมูลที่มีความสำคัญและสามารถใช้เป็นตัวแทนของข้อมูลส่วนใหญ่ได้ และในความเป็นจริงมักจะเกิดเหตุการณ์ที่ เรียกกันว่า Curse of dimensionality ขึ้นเสมอ นั้นหมายความว่า จำเป็นต้องลดขนาดมิติของข้อมูลลง (dimensionality reduction) เพื่อให้ classifier สามารถทำงานได้ถูกต้องมากขึ้น

ในงานวิจัยฉบับนี้จะเน้นการนำเสนอวิธีการลดมิติข้อมูลและจำแนกประเภทข้อมูลโดยใช้วิธีการลดมิติข้อมูล แบบ PCA(Principal Components Analysis) และ CFS (Correlation-based Feature Selection) เพื่อลดขนาดมิติของข้อมูลลง ให้เหมาะกับการจำแนกประเภทของข้อมูล และวิธีการที่เป็นที่นิยมในการนำมาประยุกต์ใช้ในการจำแนกข้อมูลวิธีหนึ่งก็คือโครงข่ายประสาทเทียมชนิด Multi-Layer Perceptron (MLP) ซึ่งพบว่าวิธีการดังกล่าวมีความสามารถในการจำแนกข้อมูลอยู่ในเกณฑ์ที่ดี และยังได้นำวิธีการที่มี ความนิยมในการจำแนกข้อมูลและมีประสิทธิภาพสูงอีกวิธีหนึ่ง คือ ซัพพอร์ตเวกเตอร์ แมชชีน เข้ามาเปรียบเทียบประสิทธิภาพโดยนำข้อมูล Ozone , Ionosphere และ Sonar จาก UCI มาทำการทดสอบการวิจัยในครั้งนี้ โดยเนื้อหาในบทความได้แบ่งเป็นส่วนดังนี้ ส่วนที่ 2 กล่าวถึงทฤษฎีที่เกี่ยวข้อง ส่วนที่ 3 วิธีการดำเนินการวิจัย ส่วนที่ 4 ผลการดำเนินงานวิจัย ส่วนที่ 5 กล่าวถึงการสรุปผล การอภิปรายผล และข้อเสนอแนะ และส่วนที่ 6 ได้กล่าวถึงเอกสารอ้างอิงที่ได้ศึกษา

ทฤษฎีที่เกี่ยวข้อง

การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล เป็นกระบวนการเพื่อกลั่นกรองข้อมูลจากฐานข้อมูลขนาดใหญ่ที่มีอยู่ (พนิดา, 2547) โดยมองที่ความสัมพันธ์ของข้อมูล แนวโน้มของข้อมูลต่างๆ เพื่อให้สามารถนำข้อมูลที่กลั่นกรองได้นำไปใช้ประโยชน์ เป็นข้อมูลสนับสนุนในการตัดสินใจในเรื่องต่าง ๆ ต่อไป ซึ่งมีด้วยกัน 5 รูปแบบ คือ

2.1.1 Association Rule เป็นการค้นหากฎความสัมพันธ์ของข้อมูลโดยค้นหาความสัมพันธ์หรือ

ความเชื่อมโยงของข้อมูลทั้งสองชุดหรือมากกว่าสองชุดขึ้นไปไว้ด้วยกัน

2.1.2 Classification and Prediction การจำแนกประเภทและการทำนาย ใช้ค้นหาโมเดลที่อธิบายข้อมูลแต่ละประเภทได้ โดยการนำเสนออาจอยู่ในรูปแบบ Decision-tree, Classification Rule และ Neural Network ซึ่งผู้ใช้ทำนายค่าบางอย่างที่ไม่รู้ หรือค่าที่หายไปในฐานข้อมูล

2.1.3 การจัดกลุ่มข้อมูล (Cluster analysis) ต้องมีความคล้ายกันมากที่สุด

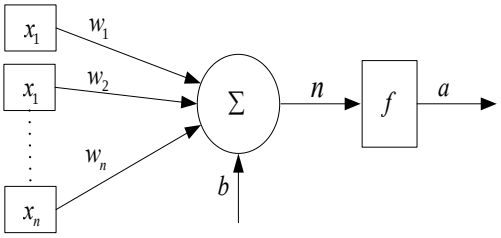
2.1.4 การหาค่าผิดปกติที่เกิดขึ้น (Outlier analysis) หรือข้อมูลบางอย่างไม่น่าจะเป็นจริงได้

2.1.5 การวิเคราะห์แนวโน้ม (Trend and evolution analysis)

<u>เทคนิคการจำแนกข้อมูล</u>

เทคนิคการทำเหมืองข้อมูล ที่สำคัญเทคนิคหนึ่งคือ Data Classification (Jiawei and Micheline, 2001) เป็นกระบวนการสร้างโมเดลจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ โดยการสร้างกฎเพื่อช่วยในการตัดสินใจจากข้อมูลที่มีอยู่ เพื่อใช้ทำนายแนวโน้มการเกิดขึ้นของข้อมูลที่ยังไม่เกิดขึ้น โดยการนำเสนอกฎที่ได้จากเทคนิคการจำแนกประเภทข้อมูล

1. โครงข่ายประสาทเทียม (artificial neural network: ANN) (พยุง,2551) มีพื้นฐานมาจากการจำลองการทำงานของสมองมนุษย์ ด้วยโปรแกรมคอมพิวเตอร์ จุดมุ่งหมายของโครงข่ายประสาทเทียมคือต้องการให้คอมพิวเตอร์มีความชาญฉลาดในการเรียนรู้เหมือนที่ มนุษย์มีการเรียนรู้ สามารถฝึกฝนได้ และสามารถนำความรู้และทักษะ รวมทั้งสามารถนำไปประยุกต์ใช้ได้ดีกับปัญหา Classification, Regression และ Clustering เทคนิคนี้ มักถูกเรียกว่า “black box” เนื่องจากการทำงานมีความ ซับซ้อนมากกว่าเทคนิคอื่น ๆ ค่อนข้างมาก การเรียนรู้ของนิวรอลเน็ตเวิร์ก ทำได้โดยการส่งข้อมูลเข้ามายังส่วนที่เรียกว่าเพอร์เซ็ปตรอน (perceptron) สามารถเทียบได้กับเซลสมองของมนุษย์ โดยที่เพอร์เซ็ปตรอนทำการรับข้อมูลที่อยู่ในรูปของเมทริกซ์ซึ่งเป็นตัวเลข เข้ามาคำนวณ ดังภาพที่ 1



ภาพที่ 1 โครงข่ายประสาทเทียมเพอร์เซ็ปตรอน

Function ผลรวม(Summation Function)

$$n = \sum_{i=1}^{z} x_i w_i + b \qquad (1)$$

โดยที่ ตัวแปร n คือ ผลรวมที่ได้จากฟังก์ชันผลรวม

ตัวแปร x_i คือ ค่าข้อมูลเข้าตัวที่ i

ตัวแปร w_i คือ ค่าน้ำหนักของนิวรอนตัวที่ i

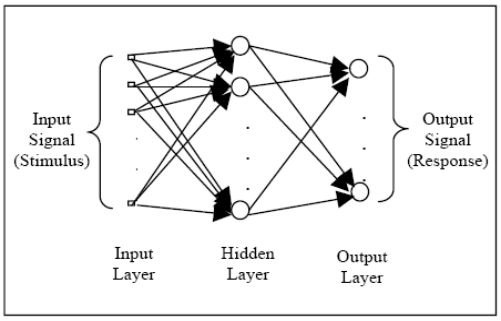
ตัวแปร z คือ จำนวนนิวรอนชั้นข้อมูลเข้า

ตัวแปร b คือ ค่าความโน้มเอียง

ตัวแปร i มีค่าตั้งแต่ 1 ถึง z

โครงข่ายประสาทเทียมแบบ Multilayer Perceptron (MLP) โครงข่ายประสาทเทียมแบบ MLP เป็นรูปแบบหนึ่งของโครงข่ายประสาทเทียมที่มีโครงสร้างเป็นแบบชั้น ใช้สำหรับงานที่มีความซับซ้อนได้ผลเป็นอย่างดี โดยมีกระบวนการฝึกฝนเป็นแบบ Supervise และใช้ขั้นตอนการส่งค่าย้อนกลับ (Backpropagation) สำหรับการฝึกฝนกระบวนการส่งค่าย้อนกลับประกอบด้วย 2 ส่วนย่อยคือ การส่งผ่านไปข้างหน้า (Forward Pass) การส่งผ่านย้อนกลับ (Backward Pass) สำหรับการส่งผ่านไปข้างหน้า ข้อมูลจะผ่านเข้าโครงข่ายประสารทเทียมที่ชั้นของข้อมูลเข้าและจะส่งผ่านจากอีกชั้นหนึ่งไปสู่อีกชั้นหนึ่งจนกระทั่งถึงชั้นข้อมูลออก ส่วนการส่งผ่านย้อนกลับค่าน้ำหนักการเชื่อมต่อจะถูกรับเปลี่ยนให้สอดคล้องกับกฎการแก้ข้อผิดพลาด (error-correction) คือผลต่างของผลตอบที่แท้จริง (actual response) กับผลตอบเป้าหมาย (target response) เกิดเป็นสัญญาณผิดพลาด (error signal) ซึ่งสัญญาณผิดพลาดนี้จะถูกส่งย้อนกลับเข้าสู่โครงข่าย

ประสาทเทียมในทิศทางตรงกันข้ามกับการเชื่อมต่อ ค่าน้ำหนักการเชื่อมต่อจะถูกปรับจนกระทั่งผลตอบที่แท้จริงเข้าใกล้ผลตอบเป้าหมาย ดังภาพที่ 2



ภาพที่ 2 โครงข่ายประสาทเทียม Multilayer Perceptron แบบ 1 hidden layer

2. ซัพพอร์ตเวกเตอร์ แมชชีน (Support Vector Machines : SVM) ตัวแบบของ SVM มีความคล้ายคลึงกับเพอร์เซฟตรอนซึ่งเป็นข่ายงานประสาทเทียมแบบง่ายมีหน่วยเดียวที่จำลองลักษณะของเซลล์ประสาทด้วยการใช้ Kernal Function ในสื่อตีพิมพ์เกี่ยวกับ SVM จะเรียกตัวแปรในการตัดสินใจว่าคุณสมบัติและตัวแปรที่เปลี่ยนแปลงใช้ในการกำหนดระนาบหลายมิติเรียกว่า คุณลักษณะ (feature) ส่วนการเลือกที่มีความเหมาะสมที่สุดเรียกว่า การคัดเลือกคุณลักษณะ (feature selection) จำนวนเซตของคุณลักษณะที่ใช้อธิบายในกรณีหนึ่ง (เช่น แถวของการค่าคาดการณ์) เรียกว่าเวกเตอร์(vector) ดังนั้นจุดมุ่งหมายของตัวแบบ SVM คือการประโยชน์สูงสุดจากระนาบหลายมิติที่แบ่งแยกกลุ่มของเวกเตอร์ในกรณีนี้ด้วยหนึ่งกลุ่มของตัวแปรเป้าหมายที่อยู่ข้างหนึ่งของระนาบ และกรณีของกลุ่มอื่นที่อยู่ทางระนาบต่างกัน ซึ่งเวกเตอร์ที่อยู่ข้างระนาบหลายมิติทั้งหมดเรียกว่า ซัพพอร์ตเวกเตอร์ (Support Vectors)

SVM เป็นวิธีการที่สามารถนำมาใช้ในการจำแนกรูปแบบหรือกลุ่มของข้อมูลได้ โดยจะอาศัยระนาบ มาใช้ในการแบ่งเขตของข้อมูลออกเป็นสองฝั่ง และ support vector machines นี้จะมีคุณลักษณะแบบ inner-product ระหว่างตัว support vector และ input vector

$$\phi(w,\xi) = \frac{1}{2} w^T w + C\sum_{i=1}^{n} \xi_i \qquad (2)$$

จากสมการที่ (2) เป็นการแสดงเวกเตอร์ค่าน้ำหนักของ w โดยจะพยายามลดค่าในเทอมแรกของสมการที่ (2) ให้มีค่าน้อยที่สุด และค่า C จะเป็นค่าคงที่ที่ใช้ สำหรับกำหนดค่าความผิดพลาดในการแยกกลุ่มข้อมูลและ ค่า ξ_i หรือ slack variable ซึ่งจะเป็นการวัดค่า ความผิดพลาดที่คลาดเคลื่อนไปจากตำแหน่งที่เหมาะสม

$$\sum_{i-1}^{n} \alpha_i d_i K(x, x_i) = 0 \qquad (3)$$

จากสมการที่ (3) แสดงค่า decision surface โดย ที่ K(x,xi) เป็น Inner-Product Kernel และ αi คือ ค่า lagrange multipliers และ *di* คือค่า target output สำหรับ kernel ของ SVM ที่นิยมใช้กัน คือ แบบpolynomial เป็นการคำนวณหาเส้นแบ่งโดยใช้สมการเชิงเส้นที่มี degree มากกว่าสองและแบบ RBF ซึ่งเป็นการคำนวณหาขอบเขตข้อมูลโดยอาศัยวิธีการแบบ Radial Basis เข้ามาช่วยในการคำนวณดังแสดงไว้ในสมการที่ (4) และ (5) ตามลำดับ

$$K(x, x_i) = (x^T x_i + 1)^P \qquad (4)$$

$$K(x, x_i) = \exp(-y \| x - x_i \|^2) \qquad (5)$$

<u>เทคนิคการลดขนาดข้อมูล (Data Reduction)</u>

การลดขนาดข้อมูล (ธรรมศักดิ์,2548) เป็นกระบวนการหนึ่งในขั้นตอนการเตรียมข้อมูล นั่นคือการทำให้ข้อมูลตั้งต้นมีขนาดลดลงโดยสูญเสียลักษณะสำคัญของข้อมูลน้อยที่สุดและสูญเสียความถูกต้องของผลลัพธ์น้อยที่สุด เนื่องจากข้อมูลแต่ละตัวจะมีความสำคัญต่อการจัดกลุ่มข้อมูลไม่เท่ากัน ด้วยเทคนิคการเลือกข้อมูลที่ดีจะทำให้สามารถเลือกข้อมูลที่มีความสำคัญและสามารถใช้เป็นตัวแทนของข้อมูลส่วนใหญ่ได้ โดยนำเสนอ อัลกอริทึมในการลดมิติข้อมูล ได้แก่ การวิเคราะห์ด้วย PCA (principal component analysis) (Lindsay, 2002) เป็นเทคนิคที่ใช้ในการลดมิติของเวกเตอร์ลักษณะ โดยการฉาย (project) เวกเตอร์ไปบนแกนใหม่ที่เรียกว่าแกนองค์ประกอบหลัก (principal

component) ซึ่งแกนเหล่านี้มีความสำคัญแตกต่างกันลงไปตามค่าความแปรปรวน (variance) บนแต่ละแกน และ CFS (Correlation-based Feature Selection) (Mark and Geoffrey ,2003) ซึ่งมีหลักของการทำงานคือการหากลุ่มของแอทริบิวต์ที่ถูกประเมินค่าจาก heuristic ที่ซึ่งพิจารณาดูจากกลุ่มของแอทริบิวต์ที่ถูกคัดเลือกสำหรับการจำแนกประเภทของข้อมูลกับระดับของความสัมพันธ์ภายในที่เกี่ยวข้องกัน

วิธีการวิเคราะห์ความแม่นตรงของโมเดล k-fold cross-validation

การตรวจสอบไขว้กัน (Cross Validation) (Ron ,1995) เป็นวิธีการในตรวจสอบค่าความผิดพลาดในการคาดการณ์ของโมเดล โดยพื้นฐานของวิธีการการตรวจสอบไขว้กันคือการสุ่มตัวอย่าง (resampling) โดยเริ่มจากแบ่งชุดข้อมูลออกเป็นส่วน ๆ และนำบางส่วนจากชุดข้อมูลนั้นมาตรวจสอบ ผลลัพธ์จากการทำการตรวจสอบไขว้กันมักถูกใช้เป็นตัวเลือกในการกำหนดโมเดล เช่น สถาปัตยกรรมเครือข่ายการสื่อสาร (network architecture) โมเดลในการคัดแยกประเภท (classification model)

ในกรณีการทำ K - fold cross-validation เราจะแบ่งข้อมูลออกเป็น K ชุดเท่าๆกัน และทำการคำนวณค่าความผิดพลาด K รอบ โดยแต่ละรอบการคำนวณข้อมูลชุดหนึ่งจากข้อมูล K ชุดจะถูกเลือกออกมาเพื่อเป็นข้อมูลทดสอบ และข้อมูลอีก K - 1 ชุดจะถูกใช้เป็นข้อมูลสำหรับการเรียนรู้

K - fold Cross Validation (K = 5) ชุดข้อมูลหลังจากทำการแบ่งออกเป็น 5 ชุดข้อมูลย่อยเท่าๆกัน โดยแต่ละกล่องคือชุดข้อมูลย่อย 1 ชุดตัวอย่างภาพที่ 3

Iteration 1: train on	2	3	4	5	, test on	1
Iteration 2: train on	1	3	4	5	, test on	2
Iteration 3: train on	1	2	4	5	, test on	3
Iteration 4: train on	1	2	3	5	, test on	4
Iteration 5: train on	1	2	3	4	, test on	5

ภาพที่ 3 *5* - fold Cross Validation

วิธีการดำเนินการวิจัย

ศึกษาปัญหาและความต้องการของระบบ

เพื่อนำมาเป็นข้อมูลในการวิเคราะห์และออกแบบพัฒนาในขั้นต่อไปผู้พัฒนาจึงได้ทำการวิเคราะห์รูปแบบ ข้อมูลโดยใช้อัลกอริทึมของโครงข่ายประสาทเทียมแบบ MLP และซัพพอร์ตเวกเตอร์ แมชชีน มาทำการเทียบเคียงหาประสิทธิภาพเพื่อความแม่นยำในการทำนายค่าโดยวิเคราะห์ลักษณะของชุดข้อมูล (Datasets) ซึ่งชุดข้อมูลที่ใช้เป็นข้อมูลที่ได้จาก UCI จำนวน 3 ชุดข้อมูล โดยมีรายละเอียดดังนี้

- Ozone มีรายละเอียดข้อมูล คือ จำนวนข้อมูลทั้งหมด 2536 จำนวน Attribute ทั้งหมด 73 Attribute มี 2 classes คือ คลาส 1 = Ozone day และ คลาส 0 = Normal day

- Ionosphere มีรายละเอียดข้อมูล คือ จำนวนข้อมูลทั้งหมด 351 จำนวน Attribute ทั้งหมด 34 Attribute มี 2 classes คือ คลาส g = Good และ คลาส b = Bad

- Sonar มีรายละเอียดข้อมูล คือ จำนวนข้อมูลทั้งหมด 208 จำนวน Attribute ทั้งหมด 60 Attribute มี 2 classes คือ คลาส R = Rock และ คลาส M = Mine

การเตรียมข้อมูลสำหรับทำดาต้าไมน์นิง (Data preparation)

นำไปกำจัด Missing value โดยใช้การแทนค่าแบบ Series Mean และ ทำการวิเคราะห์พฤติกรรมของข้อมูลโดยได้เลือกวิธีการ พล็อตข้อมูลแบบ BoxPlot จะพบว่า ข้อมูลส่วนใหญ่กระจายตัวแบบเกาะกลุ่มกันจะมีบาง Attribute ที่มีข้อมูลมีการกระจายมาก และข้อมูลส่วนใหญ่จะมีค่า Outlier น้อย

โปรแกรมที่ใช้ในงานวิจัยครั้งนี้ ผู้จัดทำได้เลือกใช้โปรแกรม Weka เวอร์ชั่น 3.7 และ Matlab ซึ่งเป็นซอฟต์แวร์ด้านการทำเหมืองข้อมูลที่ได้รับการยอมรับอย่างแพร่หลายมาทำการวิจัย

ทำการลดขนาดมิติของข้อมูล โดยอัลกอริทึมในการลดขนาดมิติของข้อมูล แบบ PCA และ CFS มาทำการเปรียบเทียบผลกับการเลือกใช้แอทริบิวต์ทั้งหมด สามารถสรุปผลจากการลดมิติข้อมูลได้ดังตารางที่ 1

ตารางที่ 1 ผลของจำนวน Attribute จากการลดมิติข้อมูลโดยอัลกอริทึม แบบ PCA และCFS

Data	Original	PCA	CFS
Ozone	73	20	18
Ionosphere	34	24	14
Sonar	60	30	19

การสร้างโมเดลระบบและการสอนข้อมูล

การเรียนรู้แบบมีการควบคุม (Supervised Learning) เป็นการเรียนรู้ซึ่งต้องมีชุดข้อมูลสำหรับการเรียนรู้ (Training Data)

3.3.1 สร้างโมเดลโครงข่ายประสาทเทียม (Neural Network) แบบ Multi-Layer Perceptron (MLP)

3.3.2 สร้างโมเดลซัพพอร์ตเวกเตอร์ แมชชีน (Support Vector Machines -SVM) ที่ใช้ kernel ด้วย rbf

การวัดประสิทธิภาพ

จำนวนข้อมูลที่เลือกมาทดสอบทั้งหมด 146 ตัวอย่าง โดยใช้การทดสอบแบบ 5-fold Cross-Validation ในงานวิจัยนี้ใช้การแบ่งข้อมูลสำหรับการทดสอบเป็น 5 ชุดย่อย (fold) แต่ละชุดย่อยมีจำนวนข้อมูลตามแต่ละวงจรการเดินที่บันทึกได้ ชุดข้อมูลที่ได้แบ่งเป็น 5 ชุดย่อย ฝึกสอนด้วยชุดข้อมูล 4 ชุด ส่วนที่เหลืออีก 1 ชุดเก็บไว้สำหรับการทดสอบ ทำการทดลองซ้ำ 5 ครั้งแต่เปลี่ยนชุดข้อมูลสำหรับฝึกสอนและทดสอบใหม่ โดยการวัดประสิทธิภาพของความถูกต้องของข้อมูลในงานวิจัยนี้ วัดได้จากค่าความถูกต้องของการจัดกลุ่มของข้อมูล ซึ่งการทดสอบประสิทธิภาพจะแบ่งออกเป็นสองแนวทางด้วยกัน คือ ส่วนแรกจะเป็นการทดสอบ Multi-layer perceptron และส่วนที่สองจะใช้อัลกอริทึมของ Support Vector Machines

ผลการดำเนินงาน

การทำนายโดยใช้เทคนิคเหมืองข้อมูล โดยจะแบ่งเป็นอัลกอริทึม โครงข่ายประสาทเทียม แบบ MLP และซัพพอร์ตเวกเตอร์แมชชีน เป็นตัวคัดแยก

1. ผลการทดสอบข้อมูลด้วยอัลกอลิทึม MLP ในส่วนนี้จะเป็นการนำโครงสร้างของ multi-layer perceptron เข้ามาทดสอบ และทำการปรับเปลี่ยนจำนวน node ใน hidden layer และเลือกประสิทธิภาพที่ดีที่สุด ซึ่งผลลัพธ์ที่ได้สามารถสรุปได้ดังตารางที่ 2

ตารางที่ 2 เปรียบเทียบประสิทธิภาพของ MLP โดยใช้การเปลี่ยนโนด ในชั้นซ่อนและเลือกประสิทธิภาพที่ดีที่สุด

Data	**Original**	**PCA**	**CFS**
Ozone	96.5753	96.8454	97.0032
Ionosphere	88.4921	91.6667	94.5869
Sonar	81.2500	83.1731	79.3269

2. ผลการทดสอบข้อมูลด้วยอัลกอลิทึม SVM ในส่วนนี้จะเป็นการนำโครงสร้างของ SVM มาทดสอบ โดยใช้ kernel แบบ polynomial ซึ่งจะทำการปรับเปลี่ยนค่า C เพื่อหาค่าที่เหมาะสมที่ให้ประสิทธิภาพสูงสุด

ปัญหาหลายๆอย่างของการสร้าง model SVM ที่ดี ปัญหาอย่างหนึ่งนั้นก็คือ การหาค่าพารามิเตอร์ที่เหมาะสม (parameter tuning) โดยปกติแล้ว SVM ก็จะมี parameter ตัวหนึ่งนั่นก็คือ C (อาจจะมีparameter ตัวอื่นๆ สำหรับ kernel ที่ต่างกันไป) ซึ่งค่า C ที่แตกต่างกันก็จะได้ model ของ SVM ที่ให้ผลไม่เหมือนกันด้วย โดยปกติแล้วนั้น ค่า C เป็นตัวกำหนด tradeoff ระหว่าง error บน training set กับขนาดของ margin หรืออีกนัยหนึ่งก็คือจะให้ความสำคัญกับ error บน training set มากน้อยแค่ไหน จากงานวิจัยนี้ได้พยายามหาค่า C ที่เหมาะสม ที่ให้ประสิทธิภาพสูงสุด โดยกำหนดค่า degree = 2 ผลลัพธ์ที่ได้สามารถสรุปได้ดังตารางที่ 3

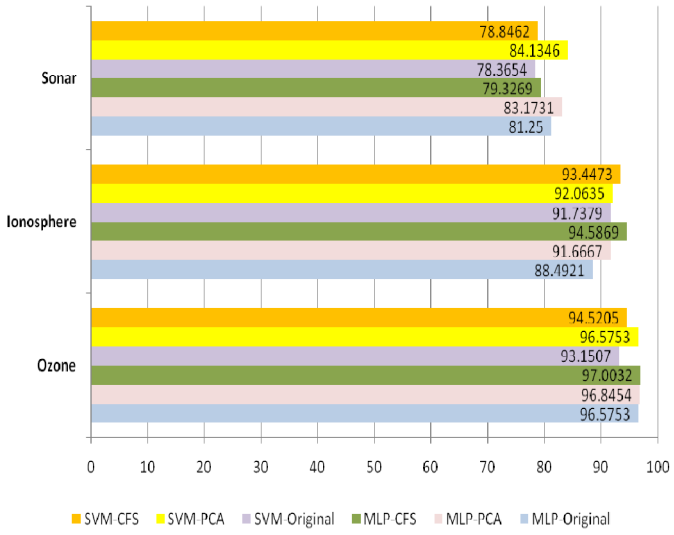
ตารางที่ 3 ผลการเปรียบเทียบค่าความถูกต้องของ SVM kernel แบบ polynomial

Data	Original		PCA		CFS	
	C	Correctly	C	Correctly	C	Correctly
Ozone	1	93.1507	5	96.5753	1	94.5205
Ionosphere	7	91.7379	7	92.0635	9	93.4473
Sonar	45	78.3654	40	84.1346	23	78.8462

นำมาทำการเปรียบเทียบค่าความถูกต้องระหว่าง MLP และ SVM สรุปได้ดังตารางที่ 4 และภาพที่ 4

ตารางที่ 4 การเปรียบเทียบค่าความถูกต้องระหว่าง MLP และ SVM

Type		Ozone	Ionosphere	Sonar
MLP	Original	96.5753	88.4921	81.2500
	PCA	96.8454	91.6667	83.1731
	CFS	97.0032	94.5869	79.3269
SVM	original	93.1507	91.7379	78.3654
	PCA	96.5753	92.0635	84.1346
	CFS	94.5205	93.4473	78.8462



ภาพที่ 4 การเปรียบเทียบของการทดสอบในแต่ละอัลกอลิทึม

จากผลการทดสอบจะเห็นได้ว่า การจำแนกโดยการใช้ PCA และ CFS มาช่วย ในการจำแนกทั้งแบบ MLP และ SVM ให้ผลการจำแนกได้ดีขึ้นกว่าการจำแนกแบบปกติ ซึ่งสามารถสรุปผลความถูกต้องที่เพิ่มขึ้น ถึง 11 ผลการทดลอง จาก 12 ผลการทดลอง ดังตารางที่ 5

จากตารางที่ 5 ผลลัพธ์ที่ได้จากการทำเหมืองข้อมูล โดยใช้โมเดลในการเรียนรู้ด้วยอัลกอริทึม CFS ร่วมกับวิธีการจำแนกข้อมูลแบบ MLP มีค่าความถูกต้อง เมื่อนำมาทดสอบ สูงมากกว่าโมเดลแบบอื่นๆ

จากผลการทดลองนี้สรุปได้ว่า

- เมื่อนำเทคนิคการลดแอทริบิวต์และจำแนกข้อมูลโดยใช้ CFS ร่วมกับ MLP เมื่อนำมาทดสอบกับข้อมูลทั้ง 3 ชุดข้อมูล ให้ค่าความถูกต้องในการจำแนกข้อมูลสูงที่สุด ซึ่งจะเห็นได้จาก 2 ชุดข้อมูล คือ Ozone ให้ค่าความถูกต้อง 97.0032 และ Ionoshere ให้ค่าความถูกต้อง 94.5869 ซึ่งมากกว่าโมเดลแบบอื่นๆ
- เมื่อพิจารณาจากความถูกต้องในการจำแนกแต่ละประเภท จะเห็นว่า CFS ร่วมกับ MLP จะให้ค่าความถูกต้องในการจำแนกข้อมูลสูงกว่า CFS ร่วมกับ SVM และ PCA ร่วมกับ SVM จะให้ค่าความถูกต้องในการจำแนกข้อมูลสูงกว่า CFS ร่วมกับ SVM

สรุป อภิปรายผล และข้อเสนอแนะ

จากการที่ได้นำเสนอการนำเทคนิคการลดมิติข้อมูลที่เหมาะสม เพื่อเพิ่มประสิทธิภาพการทำนายเพื่อการสังเคราะห์โมเดลได้อย่างรวดเร็ว และเพื่อลด

ความซับซ้อนของรูปแบบโมเดลนั้นจะพบว่า ผลของประสิทธิภาพการจำแนกถูกต้องมากขึ้นเมื่อเทียบกับการเลือกใช้ แอทริบิวต์ ทั้งหมด โดยเฉพาะในการทดสอบแบบ CFS ร่วมกับวิธีการจำแนกข้อมูลแบบ MLP จะเห็นว่าประสิทธิภาพในการจำแนกข้อมูล ได้เปอร์เซ็นต์ค่าความถูกต้องที่เพิ่มขึ้นสูงกว่าการเลือกใช้แอทริบิวต์ทั้งหมด และแบบอื่นๆ

ในการศึกษาครั้งต่อไปผู้วิจัยมุ่งศึกษาที่จะหาว่าปัจจัยใดบ้างที่มีผลทำให้การวิเคราะห์การจำแนกข้อมูลให้มีความถูกต้องมากขึ้น รวมไปถึงการศึกษาปรับปรุงขั้นตอนวิธีให้มีประสิทธิภาพมากขึ้นด้วยและจากการที่

ได้นำเสนอการนำเทคนิคการลดมิติข้อมูล น่าจะเป็นประโยชน์กับเทคนิคการทำเหมืองข้อมูลเทคนิคอื่นๆ ในการจำแนกประเภทข้อมูล ซึ่งอาจจะทำให้ประสิทธิภาพของการทำเหมืองข้อมูลเพิ่มขึ้น

ตารางที่ 5 เปอร์เซ็นต์ค่าความถูกต้องที่เพิ่มขึ้นจากการทดสอบระหว่าง MLP และ SVM

Data	MLP-Original	MLP-PCA		MLP-CFS		SVM-Original	SVM-PCA		SVM-CFS	
	Correctly	Correctly	Increase	Correctly	Increase	Correctly	Correctly	Increase	Correctly	Increase
Ozone	96.5753	96.8454	***0.2701***	**97.0032**	***0.4279***	93.1507	96.5753	***3.4246***	94.5205	***1.3698***
Ionosphere	88.4921	91.6667	***3.1746***	**94.5869**	***6.0948***	91.7379	92.0635	***0.3256***	93.4473	***1.7094***
Sonar	81.25	83.1731	***1.9231***	79.3269	-	78.3654	**84.1346**	***5.7692***	78.8462	***0.4808***

เอกสารอ้างอิง

พนิดา ยืนยงสวัสดิ์. การพยากรณ์ปริมาณการใช้ยาโดยใช้โครงข่ายประสาทเทียม, สารนิพนธ์ภาควิชา เทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้า พระนครเหนือ, 2547.

พยุง มีสัจ,ระบบฟัซซีและโครงข่ายประสาทเทียม, เอกสารประกอบการสอน, คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ 2551.

ธรรมศักดิ์ เธียรนิเวศน์. การลดขนาดข้อมูลด้วยน้ำหนักความหนาแน่นเพื่อการจัดกลุ่มข้อมูลขนาดใหญ่, วิทยานิพนธ์ สาขาวิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี ปีการศึกษา 2548.

Jiawei Han and Micheline Kamber., Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, 2001.

Lindsay I. Smith. , A tutorial on principal components analysis, February 2002, pp1-26

Mark A.Hall and Geoffrey Holmes. “Benchmarking Attribute Selection Techniques for Discrete Class Data Mining”, IEEE ,2003.

Ron Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, vol. 2, no. 12, pp. 1137–1143, 1995.