

การเปรียบเทียบประสิทธิภาพและวิเคราะห์การจำแนกข้อมูลโดยใช้โครงข่ายประสาทเทียม

ซัพพอร์ตเวกเตอร์แมชชีน นาอ์ฟเบย์ และ เคเนียร์เรสต์เนเบอร์

A Comparative Efficiency and Analysis of Classification by Neural Network,

Support Vector Machines, Naïve Bayes and K-Nearest Neighbor

วาทีณี นุ้ยเพียร (Vatinee Nuipian) * ดร.สมชาย ปรากการเจริญ (Dr.Somchai Prackanchaen)**

ดร.พยุ่ง มีสัจ (Dr.Phayung Meesad)***

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการจำแนกข้อมูล โดยเลือกใช้อัลกอริทึม โครงข่ายประสาทเทียม แบบ มัลติเลเยอร์เพอร์เซ็ปตรอน ซัพพอร์ตเวกเตอร์แมชชีน นาอ์ฟเบย์ และ เคเนียร์เรสต์เนเบอร์ เพื่อประเมิน ประสิทธิภาพค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความถ่วงดุล (F-Measure) ใช้ข้อมูลจาก UCI ประกอบด้วย Ozone Days และ Adult เลือกกลุ่มข้อมูลโดยมีจำนวนคลาสเท่ากันในข้อมูล แต่ละชุด เป็นการทดลองแบบมีการเรียนรู้ จากผลการวิจัยอัลกอริทึมที่ดีที่สุดของข้อมูล Ozone Days คือ ซัพพอร์ตเวกเตอร์แมชชีนฟังก์ชันเคอร์เนลแบบ Rbf มีค่าความถูกต้อง 95.83 เปอร์เซ็นต์ ค่าความแม่นยำ 96 เปอร์เซ็นต์ ค่าความระลึก 96 เปอร์เซ็นต์และค่าความถ่วงดุล 96 เปอร์เซ็นต์ ส่วนข้อมูล Adult คือ ซัพพอร์ตเวกเตอร์แมชชีนฟังก์ชันเคอร์เนลแบบ พอลลิโนเมียล มีค่าความถูกต้อง 79.66 เปอร์เซ็นต์ ค่าความแม่นยำ 80 เปอร์เซ็นต์ ค่าความระลึก 80 เปอร์เซ็นต์ และค่าความถ่วงดุล 80เปอร์เซ็นต์ อัลกอริทึมที่สามารถเลือกใช้ได้ดีคือ ซัพพอร์ตเวกเตอร์แมชชีน สามารถใช้กับ ลักษณะข้อมูลที่เป็นตัวเลขหรือข้อมูลเชิงกลุ่มแบบข้อความ ซึ่งเทคนิคเหล่านี้สามารถประยุกต์ใช้กับการสร้าง เทคโนโลยีการจัดเก็บและนำเสนอเนื้อหาแบบมีโครงสร้าง โดยสามารถวิเคราะห์ จำแนกหรือจัดแบ่งข้อมูลที่มีความสัมพันธ์กับข้อมูลอื่นๆ แบบเชิงความหมายได้ต่อไป

ABSTRACT

The objective this research is to compare efficiency of four classification model that are Artificial Neural Network (Multi-layer Perceptron : MLP), Support Vector Machine (SVM) , Naive Bayes (NB) and K-Nearest Neighbor(K-NN) using correct classification Accuracy, Precision, Recall and F-Measure on Ozone days and Adult data set from UCI Machine Learning Database Repository . The results of experiment on efficiency of Support Vector Machine (Rbf) algorithms are Accuracy 95.83 %, Precision 96 %, Recall 96% and F-Measure 96% on Ozone data set and Accuracy 79.66 %, Precision 80 %, Recall 80% and F-Measure 80% on Adult data set on Support Vector

*คณบดีบัณฑิต ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

**ผู้ช่วยศาสตราจารย์ ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

***ผู้ช่วยศาสตราจารย์ ภาควิชาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

Machine (Polynomial). The result of this research represent Support Vector Machine (SVM) kernel function Radial Basis and Polynomial are suitable for grouping member and categorical data type. In addition algorithms can apply to Semantic Information Technology by analyzing, classifying or partitioning data that relate to other data.

คำสำคัญ : การจำแนกข้อมูล โครงข่ายประสาทเทียม ซัพพอร์ตเวกเตอร์แมชชีน นาอ์ฟเบย์ และ เคเนียร์เรสต์เนเบอร์

Key Words : Classification, Neural Network, Support Vector Machine (SVM) , Naive Bayes (NB) and K-Nearest Neighbor (K-NN)

บทนำ

ข้อมูลในรูปแบบอิเล็กทรอนิกส์ มีปริมาณมากขึ้นทุกวัน และมีเนื้อหาหลากหลายจึงเป็นเรื่องที่ยากกับการจัดกลุ่มอีกทั้งต้องใช้เวลาาน เช่น ชนิดของข้อมูลที่เป็นข้อความ และมีความสัมพันธ์กัน โครงสร้างแตกต่างกัน เทคนิคการจำแนกข้อมูลนี้จึงมีความสำคัญกับการสืบค้นความรู้บนฐานข้อมูลขนาดใหญ่ หรือใช้การสร้างโมเดลแยกแยะทวิวิธโดยขึ้นอยู่กับแอทริบิวต์อื่น ทำให้สามารถพิจารณาคลาสใหม่ได้ ซึ่งมีการนำเสนอเทคนิควิธีการจำแนกกลุ่มหรือหมวดหมู่ เช่น Neural Network, (Holmes, et al 2002) Support Vector Machine (SVM), (Baoli, et al 2004) Naïve Bayes, (Canasai, et al 2001) K-Nearest Neighbor (K-NN) (Huma , et al 2002) และ Decision Tree (Youngjoong , et al 2002) และเป็นเทคนิคที่นิยมใช้งานดังนี้ผู้วิจัยจึงได้นำเทคนิคเหล่านี้มาประยุกต์เข้ากับงานวิจัย ที่มีชนิดข้อมูล เป็นสองประเภทคือตัวเลข (Number) และข้อมูลเชิงกลุ่มแบบข้อความ (Categorical) เพื่อเลือกใช้อัลกอริทึมที่ดีที่สุด และนำไปใช้ในการจำแนกข้อมูลแบบเชิงความหมายต่อไป

วัตถุประสงค์ของการวิจัย

เพื่อประเมินประสิทธิภาพ ค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล และเลือกใช้อัลกอริทึมที่ดีที่สุด สำหรับการจำแนกข้อมูล

เอกสารและงานวิจัยที่เกี่ยวข้อง

1. การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล เป็นกระบวนการค้นกรองข้อมูลจากฐานข้อมูลขนาดใหญ่โดยมองที่ความสัมพันธ์ของข้อมูล แนวโน้มของข้อมูลต่าง ๆ เพื่อให้สามารถนำข้อมูลที่กลั่นกรองได้มาใช้ประโยชน์ และเป็นข้อมูลสนับสนุนในการตัดสินใจในเรื่องต่าง ๆ ซึ่งมีด้วยกัน 5 รูปแบบ คือ

1.1 Association Rule เป็นการค้นหากฎความสัมพันธ์ของข้อมูลโดยค้นหาความสัมพันธ์หรือ

ความเชื่อมโยงของข้อมูลทั้งสองชุดหรือมากกว่าสองชุดขึ้นไปไว้ด้วยกัน

1.2 Classification and Prediction การจำแนกประเภท และการทำนายใช้ค้นหาโมเดลที่อธิบายข้อมูลแต่ละประเภทได้ โดยการนำเสนออาจอยู่ในรูปแบบ Decision-tree, Classification Rule, Neural Network ซึ่งผู้ใช้งานบางอย่างไม่รู้ หรือค่าที่หายไปบนฐานข้อมูล

1.3 Cluster Analysis การจัดกลุ่มข้อมูล ต้องมีความคล้ายกันมากที่สุด

1.4 Outlier Analysis การหาค่าผิดปกติที่เกิดขึ้นหรือข้อมูลบางอย่างไม่น่าจะเป็นจริงได้

1.5 Trend and Evolution Analysis การวิเคราะห์แนวโน้ม (Han, et al 2001)

2. เทคนิคการจำแนกข้อมูล (Classification) เป็นกระบวนการสร้างโมเดลการจำแนกข้อมูล ซึ่งสร้างกฎการตัดสินใจจากข้อมูลที่มีอยู่โดยใช้เทคนิคดังต่อไปนี้

2.1 โครงข่ายประสาทเทียม (Artificial Neural Network: ANN) คือแบบจำลองทางคณิตศาสตร์สำหรับประมวลผลสารสนเทศ ด้วยการคำนวณแบบคอนเนชันนิสต์ (Connectionist) ได้มาจากการศึกษาข่ายงานไฟฟ้าชีวภาพ (Bioelectric Network) เกิดจากการเชื่อมต่อระหว่างเซลล์ประสาทเป็นการเลียนแบบวิธีการทำงานของสมองมนุษย์ ซึ่งเป็นรูปแบบการคำนวณที่ค่อนข้างซับซ้อน โดยนำข้อมูลต่าง ๆ ไปใช้ในการวิเคราะห์ ตีความหรือคาดคะเน (ซูฟีย์, 2551)

โครงข่ายประสาทเทียมแบบ มัลติเลเยอร์เพอร์เซ็ปตรอน เป็นรูปแบบหนึ่งของโครงข่ายประสาทเทียมที่มีโครงสร้างเป็นแบบหลายๆ ชั้น ใช้สำหรับงานที่มีความซับซ้อน ใช้ขั้นตอนการส่งค่าย้อนกลับ (Back Propagation) สำหรับการฝึกฝนกระบวนการส่งค่าย้อนกลับ ประกอบด้วย 2 ส่วนย่อยคือ การส่งผ่านไปข้างหน้า (Forward Pass) การส่งผ่านย้อนกลับ (Backward Pass) สำหรับการส่งผ่านไปข้างหน้า ข้อมูลผ่านเข้าโครงข่ายประสาทเทียมที่ชั้นข้อมูลเข้า และส่งผ่าน จากอีกชั้นหนึ่งไปสู่อีกชั้นหนึ่งจนกระทั่งถึงชั้น

ข้อมูลออก ส่วนการส่งผ่านย้อนกลับค่าน้ำหนักการเชื่อมต่อจะถูกปรับเปลี่ยนให้สอดคล้องกับกฎการแก้ไขผิดพลาด (Error-Correction) คือผลต่างของผลตอบที่แท้จริง (Actual Response) กับผลตอบเป้าหมาย (Target Response) เกิดเป็นสัญญาณผิดพลาด (Error Signal) ซึ่งสัญญาณผิดพลาดนี้ถูกส่งย้อนกลับเข้าสู่โครงข่ายประสาทเทียมในทิศทางตรงกันข้ามกับการเชื่อมต่อ และค่าน้ำหนักของการเชื่อมต่อจะถูกปรับจนกระทั่งผลตอบที่แท้จริงเข้าใกล้ผลตอบเป้าหมาย

หลักการทำงานของมัลติเลเยอร์เพอร์เซปตรอน คือในแต่ละชั้นของชั้นซ่อนตัว (Hidden Layer) มีฟังก์ชันสำหรับคำนวณเมื่อได้รับสัญญาณ (Output) จากโหนดในชั้นก่อนหน้านี้ เรียกว่า Activation Function โดยในแต่ละชั้นไม่จำเป็นต้องเป็นฟังก์ชันเดียวกัน ชั้นซ่อนตัวมีหน้าที่สำคัญคือ พยายามแปลงข้อมูลที่เข้ามาในชั้น (Layer) นั้นๆ ให้สามารถแยกแยะความแตกต่างโดยใช้เส้นตรงเส้นเดียว (Linearly Separable) และก่อนที่ข้อมูลถูกส่งไปถึงชั้นข้อมูลออก (Output Layer) ในบางครั้งอาจจำเป็นต้องใช้ชั้นซ่อนตัวมากกว่า 1 ชั้นในการแปลงข้อมูลให้อยู่ในรูป Linearly Separable (พยุ่ง, 2551)

2.2 เทคนิคซ์พอร์ตเวกเตอร์เมชชีน คือการสร้างสมการเส้นตรงเพื่อแบ่งเขตข้อมูล 2 กลุ่มออกจากกัน โดย SVM พยายามสร้างเส้นแบ่งตรงกึ่งกลางระหว่างกลุ่มให้มีระยะห่างระหว่างขอบเขตของทั้งสองกลุ่มให้มากที่สุด SVM ใช้ฟังก์ชันแม่สำหรับย้ายข้อมูลจาก Input Space ไปยัง Feature Space และสร้างฟังก์ชันวัดความคล้ายที่เรียกว่า (Kernel Function) บน Feature Space

ข้อดีของวิธีการนี้คือ รองรับจำนวนคุณลักษณะได้มาก และมีความถูกต้องสูง ข้อเสียคือ ต้องเลือก Kernel Function ที่เหมาะสม

2.3 นาอิวเบย์ คือโมเดลการจำแนกกลุ่มที่ใช้หลักความน่าจะเป็นซึ่งอยู่บนพื้นฐานของ Bayes' Theorem เช่นกำหนดให้การเกิดของเหตุการณ์ต่างๆที่ใช้ในการจำแนกกลุ่มนั้นเป็นอิสระต่อกันโดยกำหนดให้ $P(H)$

ความน่าจะเป็นที่จะเกิดเหตุการณ์ H และ $P(H|E)$ คือความน่าจะเป็นที่จะเกิดเหตุการณ์ H เมื่อเกิดเหตุการณ์ E จากตัวแปรที่กำหนดและแนวคิดของ Bayes' Theorem นั้นเราสามารถทำนายเหตุการณ์ที่พิจารณาได้จากการเกิดของเหตุการณ์ต่างๆ ได้ดังสมการ

$$P(H|E) = [P(E|H) \times P(H)] / P(E)$$

เช่นการทำนายว่าฝนจะตกเมื่อมีเหตุการณ์มีเมฆดำกำหนดให้ H คือเหตุการณ์ที่ฝนตก และ E คือเหตุการณ์มีเมฆดำ และสามารถทำนายสภาพอากาศได้ดังสมการต่อไปนี้

$$P(\text{ฝนตก} | \text{เมฆดำ}) = [P(\text{เมฆดำ} | \text{ฝนตก}) \times P(\text{ฝนตก})] / P(\text{เมฆดำ})$$

กำหนดให้

- $P(\text{เมฆดำ} | \text{ฝนตก})$ คือความน่าจะเป็นที่มีเมฆดำเมื่อฝนตก ซึ่งจะพิจารณาการเกิดเมฆดำเมื่อมีเหตุการณ์ฝนตกเท่านั้น ความน่าจะเป็นนี้สามารถเก็บรวบรวมโดยใช้หลักการทางสถิติ
- $P(\text{ฝนตก})$ คือความน่าจะเป็นที่ฝนจะตก
- $P(\text{เมฆดำ})$ คือความน่าจะเป็นที่มีเมฆดำ

การคำนวณการจำแนกกลุ่มของเหตุการณ์ที่มีการเกิดของเหตุการณ์ต่างๆที่ใช้ในการจำแนกกลุ่มมากกว่า 1 ชนิด ได้ดังสมการต่อไปนี้

$$P(H|E_1, E_2, \dots, E_n) = [P(E_1, E_2, \dots, E_n|H) \times P(H)] / P(E_1, E_2, \dots, E_n)$$

เมื่อกำหนดให้เหตุการณ์ $E_1, E_2, \dots, E_n$ คือเหตุการณ์ n เหตุการณ์ที่ใช้ในการจำแนกกลุ่ม และแต่ละเหตุการณ์ต่างๆที่ใช้ในการจำแนกกลุ่มเป็นอิสระต่อกัน

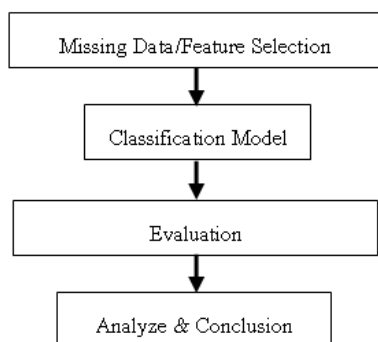
2.4 เคเนียร์เรสต์เนเบอร์ คือการตัดสินใจของคลาสสำหรับแทนเงื่อนไขหรือกรณีใหม่ โดยการตรวจสอบจำนวนบางจำนวน หรือเงื่อนไขที่เหมือนกันหรือใกล้เคียงกันมากที่สุด ใช้เวลาในการคำนวณสูงเพราะ

การคำนวณเป็นการเพิ่มขึ้นแบบแฟกทอเรียล ตามจุดทั้งหมด ขณะที่ Decision Tree หรือโครงข่ายประสาทเทียม ประมวลผลเพื่อสร้างเงื่อนไขได้เร็วกว่า เพราะเคเนียร์เรสต์เนเบอร์ มีการคำนวณทุกครั้งที่มีการฝึกใหม่ ดังนั้นเพื่อความรวดเร็วสำหรับเทคนิคนี้ข้อมูลทั้งหมดที่ใช้บ่อยจะต้องถูกเก็บไว้ในหน่วยความจำ
ชื่อว่า Memory-Base Reasoning

ศึกษาปัญหาและความต้องการของระบบ

วิธีการดำเนินการวิจัย

เพื่อนำข้อมูลมาใช้ในการวิเคราะห์และ ออกแบบพัฒนาในขั้นต่อไปผู้พัฒนาจึงได้ทำการวิเคราะห์ รูปแบบข้อมูลด้วยการใช้เทคนิคการทำเหมืองข้อมูล โดยใช้ อัลกอริทึมของโครงข่ายประสาทเทียม แบบมัลติเลเยอร์เพอร์เซ็ปตรอน ซัพพอร์ตเวกเตอร์แมชชีน นาอิวเบย์ และ เคเนียร์เรสต์เนเบอร์ ทำการเปรียบเทียบหาประสิทธิภาพ ค่าความถูกต้อง ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความถ่วงดุล (F-Measure) ใช้ข้อมูลจาก UCI คือ Ozone Days ลักษณะข้อมูลเป็นตัวเลข และ Adult ลักษณะข้อมูลเป็นเชิงกลุ่มแบบข้อความและตัวเลข มีขั้นตอนการดำเนินการวิจัยดังภาพที่ 1 และเลือกใช้แบบจำนวนคลาสเท่ากันในแต่ละคลาส ดังตารางที่ 1



ภาพที่ 1 ขั้นตอนการดำเนินการวิจัย

ตารางที่ 1 ลักษณะข้อมูลทั้ง 2 ชุด

Data	Instance	Attributes	Attributes test	Instance test
ozone days	2,536	76	24	72
Adult	32,561	15	15	7,508

การเตรียมข้อมูลสำหรับการทำเหมืองข้อมูล

ขั้นตอนที่ 1 เตรียมข้อมูล (Data Preparation) นำไปกำจัด Missing Value โดยใช้การแทนค่า แบบ Series Mean ทำการวิเคราะห์พฤติกรรมข้อมูลโดยเลือกวิธีการพล็อตข้อมูลแบบ BoxPlot และ Visualize พบว่า ข้อมูลส่วนใหญ่กระจายตัวแบบเกาะกลุ่มมีบาง Attribute ที่ข้อมูลมีการกระจายมาก และข้อมูลส่วนใหญ่มีค่าสังเกตผิดปกติ (Outlier) น้อย

ขั้นตอนที่ 2 ลดขนาดข้อมูล (Feature Selection) คือการหาโมเดลหรือแพทเทิร์นที่ข้อมูลส่วนใหญ่แสดง ลักษณะเหล่านั้นออกมาเหมือนกัน การลดขนาดข้อมูลให้กับ Ozone days เนื่องจากมีข้อมูลจำนวนมากและได้ทำการทดสอบค่าความถูกต้องของข้อมูล (วาทีนี้, และ คณะ 2552) จึงเลือกใช้แบบผู้วิจัยพิจารณาเอง จำนวน 24 แอททริบิวต์ ส่วนข้อมูล Adult มีแอททริบิวต์ไม่มากและเมื่อทดลองลดขนาดข้อมูลแล้วไม่สามารถลดจำนวนแอททริบิวต์ได้ หลังจากนั้นนำข้อมูลที่ผ่านการลดขนาดแบ่งออกเป็นสองส่วนคือ ส่วนแรกใช้ในกระบวนการค้นหา แพทเทิร์น (Training Set) ส่วนที่สองใช้ตรวจสอบความถูกต้องของแพทเทิร์น (Test Set) โดยเลือกจำนวน Class เท่ากันในแต่ละชุด ผู้วิจัยเลือกใช้โปรแกรม Weka เวอร์ชัน 3.6.0 ซึ่งเป็นซอฟต์แวร์ที่ใช้งานด้านการทำเหมืองข้อมูล และใช้โปรแกรม Weka ในการคัดเลือกข้อมูล (Select Attribute)

ขั้นตอนการสร้างโมเดลการจำแนกกลุ่มข้อมูล

สร้างโมเดลการเรียนรู้แบบมีการควบคุม (Supervised Learning) ทำการสร้างโมเดล โครงข่ายประสาทเทียมแบบ มัลติเลเยอร์เพอร์เซ็ปตรอน

ซัพพอร์ทเวกเตอร์แมชชีน ใช้ Kernel แบบ Rbf และ Polynomial นาอ์ฟเบย์ และ เคเนียร์เรสต์เนเบอร์

การประเมินผล

การประเมินค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล ซึ่ง ค่าความระลึกเป็นการวัดความสามารถของระบบในการดึงข้อมูลที่เกี่ยวข้องออกมาในขณะที่ ค่าความแม่นยำเป็นการวัดความสามารถในการที่จัดข้อมูลที่ไม่เกี่ยวข้องออกไป มีค่าอยู่ระหว่าง 0-1 ถ้า 1 หมายถึง ประสิทธิภาพดี

$$\text{Precision} = \frac{\text{จำนวนข้อมูลที่เกี่ยวข้องทั้งหมดและถูกดึงออกมา}}{\text{จำนวนข้อมูลที่ถูกดึงออกมาทั้งหมด}}$$

$$\text{Recall} = \frac{\text{จำนวนข้อมูลที่เกี่ยวข้องทั้งหมดและถูกดึงออกมา}}{\text{จำนวนข้อมูลที่เกี่ยวข้องทั้งหมด}}$$

$$\text{F-Measure} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

ผลการวิจัย

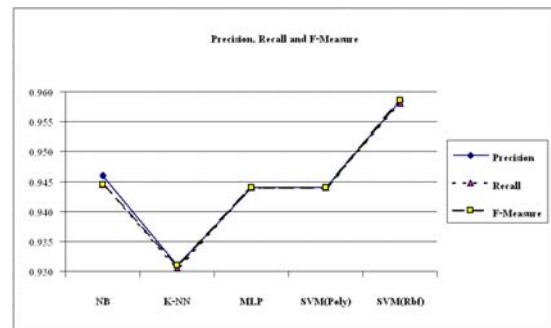
การจำแนกกลุ่มโดยใช้เทคนิคการทำเหมืองข้อมูล นำข้อมูลมาทำการเรียนรู้และทดสอบโดยใช้อัลกอริธึมตามที่กล่าวมาข้างต้นโดยนำมาเปรียบเทียบ และสรุปผลแยกตามกลุ่มข้อมูลทั้ง 2 ชุด

ตารางที่ 2 เปรียบเทียบค่าความถูกต้องของกลุ่มข้อมูล Ozone days โดยเรียงจากค่าความถูกต้องมากไปน้อย

Type (Ozone)	Correctly	Incorrect
Support Vector Machines (Rbfkernel)	95.83	4.16
Support Vector Machines (Polykernel)	94.44	5.55
Naive Bayes	94.44	5.55
Multi-Layer Perceptron	94.44	5.55
K-Nearest Neighbour	93.05	6.94

ตารางที่ 3 เปรียบเทียบค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุลของกลุ่มข้อมูล Ozone days

Type	Precision	Recall	F-Measure
NB	0.946	0.945	0.945
K-NN	0.931	0.931	0.931
MLP	0.944	0.944	0.944
SVM(Poly)	0.944	0.944	0.944
SVM(Rbf)	0.959	0.958	0.959



ภาพที่ 2 เปรียบเทียบค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล ของข้อมูล Ozone days

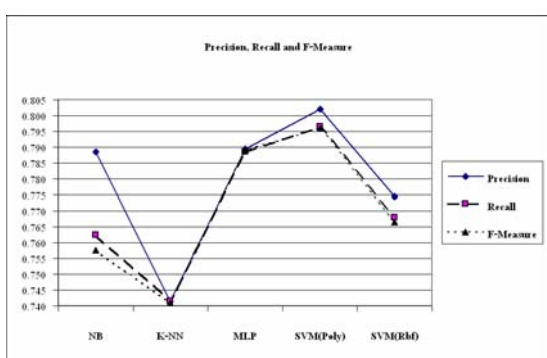
จากภาพที่ 2 แสดงให้เห็นถึงการประเมินผลค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล ของข้อมูล Ozone Days ตามอัลกอริธึมที่เลือกใช้มีค่าใกล้เคียงกัน

ตารางที่ 4 เปรียบเทียบค่าถูกต้องของกลุ่มข้อมูล Adult โดยเรียงจากค่าความถูกต้องมากไปน้อย

Type (Adult)	Correctly	Incorrect
Support Vector Machines (Polykernel)	79.66	20.33
Multi-Layer Perceptron	78.87	21.12
Support Vector Machines (Rbfkernel)	77.25	22.74
Naive Bayes	76.25	23.73
K-Nearest Neighbour	74.12	25.87

ตารางที่ 5 เปรียบเทียบค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุลของกลุ่มข้อมูล Adult

Type	Precision	Recall	F-Measure
NB	0.789	0.763	0.758
K-NN	0.742	0.742	0.741
MLP	0.790	0.789	0.789
SVM(Poly)	0.802	0.797	0.796
SVM(Rbf)	0.775	0.768	0.767



ภาพที่ 3 เปรียบเทียบค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล ของข้อมูล Adult

จากภาพที่ 3 แสดงให้เห็นถึงการประเมินผลค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล ของข้อมูล Adult โดยมีอัลกอริทึมของ ซัพพอร์ตเวกเตอร์แมชชีนฟังก์ชันเคอร์เนลแบบพอลิโนเมียล มีค่าสูงที่สุด รองลงมาคือโครงข่ายประสาทเทียม แบบมัลติเลเยอร์เปอร์เซปตรอน ซัพพอร์ตเวกเตอร์แมชชีนฟังก์ชันเคอร์เนลแบบ Rbf นาอูฟเบย์ และ เคเนียร์เรสต์เนเบอร์ ตามลำดับ

ตารางที่ 6 เปรียบเทียบประสิทธิภาพ ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล ของข้อมูล 2 ชุด

Type	Ozone Precision	Adult Precision	Ozone Recall	Adult Recall	Ozone F-Measure	Adult F-Measure
NB	95%	79%	94%	76%	94%	76%
K-NN	93%	74%	93%	74%	93%	74%
MLP	94%	79%	94%	79%	94%	79%
SVM(Poly)	94%	80%	94%	80%	94%	80%
SVM(Rbf)	96%	77%	96%	77%	96%	77%

จากตารางที่ 6 อัลกอริทึมที่ดีที่สุดของข้อมูล Ozone Days คือ ซัพพอร์ตเวกเตอร์แมชชีนฟังก์ชันเคอร์เนลแบบ Rbf ค่าความแม่นยำ 96 เปอร์เซ็นต์ ค่าความระลึก 96 เปอร์เซ็นต์ และค่าความถ่วงดุล 96 เปอร์เซ็นต์ ส่วนข้อมูล Adult คือ ซัพพอร์ตเวกเตอร์แมชชีนฟังก์ชันเคอร์เนลแบบ พอลิโนเมียล ค่าความแม่นยำ 80 เปอร์เซ็นต์ ค่าความระลึก 80 เปอร์เซ็นต์ และค่าความถ่วงดุล 80 เปอร์เซ็นต์ อัลกอริทึมที่สามารถเลือกใช้ได้ดีคือ ซัพพอร์ตเวกเตอร์แมชชีน สามารถใช้กับลักษณะข้อมูลที่เป็นตัวเลขหรือข้อมูลเชิงกลุ่มแบบข้อความ

สรุปอภิปรายผล

การจัดทำวิจัยครั้งนี้มีวัตถุประสงค์เพื่อประเมินค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล และเลือกใช้อัลกอริทึมที่ดีที่สุดโดยใช้ข้อมูล Ozone days มีชนิดข้อมูลเป็นจำนวนจริงและ Adult มีชนิดข้อมูลเป็นเชิงกลุ่มและจำนวนเต็ม จากผลการวิจัย อัลกอริทึมที่ดีที่สุด ของข้อมูลทั้งสองแบบคือ ซัพพอร์ตเวกเตอร์แมชชีน ฟังก์ชันเคอร์เนลแบบ Radial Basis และ Polynomial เป็นแบบเชิงเส้นซึ่งตรงกับงานวิจัยของ (Fubo, et al 2008)(Leslie, et al 2005) และสามารถประยุกต์ใช้กับการสร้างเทคโนโลยีการจับเก็บและนำเสนอเนื้อหาแบบมีโครงสร้าง โดยสามารถวิเคราะห์ จำแนกหรือจัดแบ่งข้อมูลที่มีความสัมพันธ์กับข้อมูลอื่นๆ ซึ่งเป็นเชิงความหมาย ได้ต่อไป

เอกสารอ้างอิง

- ชูชีพย์ ดาษะ. 2551. รายงาน Neural Network. รายงานประกอบวิชา Seminar in Computer Engineering. มหาวิทยาลัยรังสิต.
- พวง มีสัจ. 2551. ระบบฟัซซี่และโครงข่ายประสาทเทียม. เอกสารประกอบการสอน. คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ.
- วาทีน นุ้ยเพียร ภรณ์ชา อำมฤตรัตน์ เดช ธรรมศิริ ณรงค์ โปธิ และพวง มีสัจ. 2552. Ozone day Comparison Clustering with Data Mining Technical. การประชุมทางวิชาการ NCIT2009 ครั้งที่ 5:911-916
- Baoli, Li, Qin, Lu and Shiwen, Yu. 2004. An adaptive k -nearest neighbor text categorization strategy. ACM Transactions on Asian Language Information Processing (TALIP). 3 (December 2004) : 215-226.
- Canasai, K., and Chuleerat, J., 2001. Thai Text Classification based on NaïveBayes. Technical Report. Department of Computer Science, Kasetsart University
- Fubo Shao, Guoping He and Xin Zhang. 2008. An Improved Algorithm for Multiclass Text Categorization with Support Vector Machine. International Symposium on Computational Intelligence and Design.
- Han, J. and Kamber, M. 2001. Data Mining Concepts and Techniques. San Francisco: Morgan Kaufmann Polisher.
- Holmes, Geoffrey, Pfahringer, Bernhard and Kirkby, Richard. 2002. Multiclass alternating decision trees. Proceedings of the 13th European Conference on Machine Learning (ECML). 161-172. Helsinki, Finland
- Huma, LH., Craig, SD., John, ST., Nello, CTN and Chris, WK. 2002. Text Classification using String Kernels. Journal of Machine Learning Research 2: 419-444
- Leslie A. Notman, Paul T. Sowden, Emre Ozgen. 2005. The nature of learned categorical perception effects :a psychophysical approach. from www.sciencedirect.com
- Youngjoong, Ko. and Jungyun, Seo. 2002. Using the Feature Projection Technique Based on a Normalized Voting Method for Text Classification Information Processing & Management. 40 : 191-208