nccit 2011

THE 7TH NATIONAL CONFERENCE ON COMPUTING AND INFORMATION TECHNOLOGY

PROCEEDINGS OF NCCIT 2011

THE 7TH NATIONAL CONFERENCE ON COMPUTING AND INFORMATION TECHNOLOGY

11-12 MAY 2011

WWW.NCCIT.NET

FACULTY OF INFORMATION TECHNOLOGY
KING MONGKUT'S UNIVERSITY OF TECHNOLOGY NORTH BANGKOK

VOLUME 1

นายเอกรัฐ หล่อพิเชียร

นายนิรุตติ์ พองาม

นางธัญญาภรณ์ บุญยัง

IEEE THAILAND SECTION

บทความวิจัย

การประชุมทางวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 7

11-12 พฤษภาคม 2554

คณะเทคโนโลยีสารสนเทศ
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

WWW.IT.KMUTNB.AC.TH

Table of Contents

Wednesday May 11, 2011	
8:00-9:00	Registration (Nawamintararachinee Building, KMUTNB)
9:00-9.30	Opening Ceremony by Prof. Dr. Teravuti Boonyasopon, *President of King Mongkut's University of Technology North Bangkok*
9:30-10:30	Invited Keynote Speech by Prof.Dr.Serge Tichkiewitch
	Topic : Life Cycle Knowledge Management
10:30-11:00	*Coffee Break*
11:00-12:00	Invited Keynote Speech by Dr.-Ing. Gerd Heinz
	Topic : Interference Networks as a Generalizing Signal Theory between Integral Transformations, Artificial Neural Nets, Optics and Acoustics
12:00-13:00	*Lunch*
13:00-17:00	Parallel Session Presentations
18:00-22:00	Welcome Dinner

Room 1		
Session 1 : Data Mining, Knowledge Discovery, and Knowledge Management		
Time/Paper-ID	**Title/Author**	**Page**
13.00-13.20 **NCCIT2011-03**	**Data Imputation Estimation Using Product Indicator Approaches of Structural Equation Modeling** *Narong Phothi and Somchai Prakancharoen*	1
13.20-13.40 **NCCIT2011-74**	**Agriculture Bibliographic Data Categorization using Feature Selection Techniques** *Pilapan Phonarin, Supot Nitsuwat, and Choochart Haruechaiyasak*	7
13.40-14.00 **NCCIT2011-96**	**An Information Extraction System for Medical Informatics using Text Mining Techniques** *Umaporn Saisangchan, Ekamorn Meesomsak, Natthapong Sripadung, Panida Songram, Kaveepoj Bunluawong, and Phatthanapong Wanchantueng*	13
14.00-14.20 **NCCIT2011-60**	**Comparative Study of Classification Properties between KNN, Naïve Bays, Decision Tree and Rule-Based** *Thanyaporn Boonyoung, Aekkarat Lorphichian and Niroot Porngam*	19
14.20-14.40 **NCCIT2011-100**	**Comparative analysis, data mining techniques to determine the suitability of the technique for prediction the different data sets and different data models** *Sajjaporn Waijanya, Thanarak Limjitsomboon and Chayapol Khoonnaret*	24

คำปรารภ

ศาสตราจารย์ ดร.ธีรวุฒิ บุณยโสภณ
อธิการบดี มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

ปัจจุบันนี้เราไม่สามารถแยกเทคโนโลยีสารสนเทศและการสื่อสาร ออกจากความต้องการในชีวิตประจำวันได้ เนื่องจากมันมีความสำคัญอย่างมากต่อทุกอาชีพ เนื่องจากช่วยอำนวยความสะดวกทางธุรกิจและการพัฒนา ซึ่งไม่เหมือนกับเทคโนโลยีส่วนใหญ่ เนื่องจากเทคโนโลยีสารสนเทศและการสื่อสารเป็นเทคโนโลยีที่ใช้งานได้ทั่วไป ที่จะนำพาไปสู่พัฒนาการ ปรับปรุงและประสิทธิภาพทางด้านเศรษฐกิจ ดังนั้น การใช้เทคโนโลยีสารสนเทศและการสื่อสารสำหรับธุรกิจและการค้า ในการพัฒนาประเทศจึงต้องการความเข้าใจอย่างถ่องแท้

ด้วยความสำคัญและผลที่ได้รับจากเทคโนโลยีสารสนเทศข้างต้น คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือจึงได้ร่วมมือทางวิชาการจากมหาวิทยาลัยพันธมิตร ได้แก่ มหาวิทยาลัยแห่งรัฐโอกลาโฮมา ประเทศสหรัฐอเมริกา มหาวิทยาลัยเฟิร์นแห่งเมืองฮาเก้น สหพันธ์สาธารณรัฐเยอรมนี มหาวิทยาลัยเคมนิทซ์ สหพันธ์สาธารณรัฐเยอรมนี มหาวิทยาลัยอีดิธโคแวน ประเทศออสเตรเลีย มหาวิทยาลัยราชภัฏมหาสารคาม มหาวิทยาลัยมหาสารคาม มหาวิทยาลัยโยนก และมหาวิทยาลัยนอร์ทกรุงเทพ เพื่อร่วมกันจัดการประชุมทางวิชาการระดับชาติ ด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 7 (NCCIT 2011) ขึ้นระหว่างวันที่ 11-12 พฤษภาคม 2554 ณ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ โดยมีวัตถุประสงค์เพื่อเผยแพร่และแลกเปลี่ยนข่าวสาร วิทยาการ ความรู้ ทักษะและประสบการณ์ จากการค้นคว้าวิจัยและพัฒนาวิชาการด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศของคณาจารย์ นักศึกษา และประชาชนทั่วไป เพื่อนำเทคโนโลยีด้านคอมพิวเตอร์และสารสนเทศ ไปใช้ประโยชน์อย่างจริงจัง

ขอขอบคุณทุกหน่วยงานและคณะบุคคล ที่มีส่วนสนับสนุนในการจัดการประชุมทางวิชาการในครั้งนี้ ขออวยพรให้การประชุมทางวิชาการ NCCIT 2011 ครั้งนี้ ประสบความสำเร็จบรรลุวัตถุประสงค์ทุกประการ

การเปรียบเทียบการจำแนกข้อมูลของแบบจำลองเคเนียเรสเนเบอร์ นาอีฟเบย์ ต้นไม้ตัดสินใจ และกฎพื้นฐาน

Comparative Study of Classification Properties between KNN, Naïve Bays, Decision Tree and Rule-Based

ธัญญาภรณ์ บุญยัง(Thanyaporn Boonyoung)[1] *เอกรัฐ หล่อพิเชียร(Aekkarat Lorphichian)*[2] *และนิรุตติ์ พองาม(Niroot Porngam)*[3]

ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ

มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

thanya_ttb@hotmail.com, aekkarat@gmail.com, amanak@hotmail.com

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาเปรียบเทียบการจำแนกข้อมูลของแบบจำลองเคเนียเรสเนเบอร์ นาอีฟเบย์ ต้นไม้ตัดสินใจ และกฎพื้นฐาน ใช้ชุดข้อมูลมาตรฐานจาก UCI Machine Learning Repository จำนวน 4 ชุดข้อมูล ที่มีคุณลักษณะแตกต่างกัน คือ ผู้ป่วยโรคหัวใจ (Heart Disease) สถานรับเลี้ยงเด็ก (Nursery) จดหมายขยะ (Spam Mail) และ ยานพาหนะ (Vehicle) ตัวแปรที่ใช้ในการศึกษาคือจำนวนข้อมูลในชุดข้อมูลฝึก ความโน้มเอียงของคลาส และจำนวนคลาส

ผลการศึกษาพบว่า ประสิทธิภาพการจำแนกข้อมูลแบบกฎพื้นฐาน มีประสิทธิภาพสูงที่สุด 99.2 % ในชุดข้อมูลสถานรับเลี้ยงเด็ก สำหรับชุดข้อมูลผู้ป่วยโรคหัวใจ เหมาะสมกับเทคนิคการเรียนรู้แบบเบย์ ชุดข้อมูลสถานรับเลี้ยง จดหมายขยะ และยานพาหนะ เหมาะสมกับเทคนิคต้นไม้ตัดสินใจ โดยให้ค่าประสิทธิภาพความแม่นยำ 84.07%, 99.20%, 93.59% และ72.46% และชุดข้อมูลที่มิติสูงเหมาะกับเทคนิดต้นไม้ตัดสินใจมากที่สุด และจำนวนชุดข้อมูลในแต่ละคลาสที่มีความโน้มเอียงส่งผลให้ประสิทธิภาพความแม่นยำต่ำลง สำหรับการเลือกแอททริบิวต์ที่เหมาะสม จำนวนข้อมูลที่มากพอต่อการเรียนรู้ และมีจำนวนที่ใกล้เคียงกันในแต่ละคลาส ส่งผลให้ประสิทธิภาพการจำแนกข้อมูลดีขึ้น

คำสำคัญ: การจำแนกข้อมูล เหมืองข้อมูล ความโน้มเอียงของคลาส

Abstract

This research aims to study the classification of technique K-Nearest neighbor Na ve Bayes Decision Tree and Rule-Based. Set of reference data from the UCI Machine Learning Repository of four Date Sets with features were difference : Nursery, Heart Disease, Spam Mail and Vehicle . The parameters used in this study are the number of training data,tendency of class and number of classes.

The results showed that the highest accuracy is Rule-Rased 99.2% in the nursery dataset. Heart disease suitable of Na ve Bayes technique and Nursery , Spam Mail Vehicle suitable of decision tree. By the efficiency and accuracy 84.07%, 99.20%, 93.59% and 72.46%. A high dimension dataset suitable of Decision Tree technique most. Tendency of class are naturally affect to lowest performance accuracy. Selection of attributed ,total data per dataset, amount of data in each of training should be similar will effect to performance improve.

Keyword: Data Classification, Data Mining, Imbalance Date Sets

1. บทนำ

เทคนิคการจำแนกข้อมูล (Data Classification) [1] เป็นแขนงหนึ่งในการทำเหมืองข้อมูล (Data Mining) ในการจำแนกข้อมูลจะใช้ชุดข้อมูลฝึกสอน (Training Date Set) ใน

การสร้างโมเดลของคลาสในแต่ละคลาสขึ้น โดยจะนำข้อมูลในแต่ละแอททริบิวท์ของชุดข้อมูลฝึกสอนมาใช้ในการเรียนรู้และสร้างโมเดล โมเดลที่ได้จากการจำแนกประเภทข้อมูลจะทำให้สามารถพิจารณาคลาสในข้อมูลที่ยังมิได้แบ่งประเภทในอนาคตได้ ในส่วนคลาสของข้อมูลนั้นมีทั้งแบบสองค่า (Binary Class) และแบบหลายค่า (Multi-Class) ตัวอย่างเช่น การใช้ข้อมูลผู้ป่วยและการวินิจฉัยของแพทย์ที่เก็บไว้ เพื่อนำมาช่วยวินิจฉัยโรคของผู้ป่วย หรือการวิจัยทางการแพทย์ ในทางธุรกิจจะใช้เพื่อดูคุณสมบัติของผู้ที่จะก่อหนี้ดีหรือหนี้เสีย เพื่อประกอบการพิจารณาการอนุมัติเงินกู้ เป็นต้น

ประสิทธิภาพความแม่นยำในการจำแนกข้อมูลขึ้นอยู่กระบวนการเตรียมข้อมูล (Data Preparation) ไม่ว่าจะเป็นการเลือกแอททริบิวต์ ประเภทของแอททริบิวต์ ชนิดของข้อมูลที่มีหลากหลาย การตรวจสอบค่าผิดปกติของข้อมูล หรือแม้กระทั่งการเลือกข้อมูลสำหรับชุดข้อมูลฝึกสอนที่ในแต่ละคลาสไม่สมดุลกัน (Imbalanced Data sets) [2][3] ซึ่งทำให้เกิดปัญหาการกระจายตัวของคลาสที่ไม่สมดุลกันอันเนื่องมาจากจำนวนชุดข้อมูลคลาสใดคลาสหนึ่งมากกว่าอีกคลาสหนึ่งเป็นจำนวนมาก

จากปัญหาดังกล่าว งานวิจัยนี้ได้นำข้อมูลจากชุดข้อมูลมาตรฐานจาก UCI Machine Learning Repository จำนวน 4 ชุดข้อมูล ซึ่งในแต่ละชุดข้อมูลมีความแตกต่างกันทั้งในเรื่องของจำนวนแอททริบิวต์ ชนิดและจำนวนของข้อมูล ประเภทของคลาส รวมถึงชุดข้อมูลที่ไม่สมดุลกันของคลาสในแต่ละคลาส โดยการนำข้อมูลดังกล่าวมาทำการสร้างโมเดลด้วยเทคนิคการจำแนกข้อมูลแบบเนียเรสเนเบอร์ (KNN) นาอีฟเบย์ (Naïve Bayes) ต้นไม้ตัดสินใจ (Decision Tree) และกฎพื้นฐาน (Rule Based) มาสร้างโมเดลเพื่อทำการเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูล

รวมถึงการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลในกรณีที่จำนวนชุดข้อมูลในแต่ละคลาสมีความโน้มเอียง ศึกษาความเหมาะสมระหว่างเทคนิคที่เลือกใช้กับข้อมูลที่มีมิติสูง และสุดท้ายทำการวิเคราะห์หาปัจจัยที่ทำให้ประสิทธิภาพการจำแนกข้อมูลดีขึ้น

2. วัตถุประสงค์ของการวิจัย

2.1 เพื่อทำการเปรียบเทียบประสิทธิภาพ ในการจำแนกข้อมูลด้วยเทคนิคเคเนียเรสเนเบอร์ นาอีฟเบย์ ต้นไม้ตัดสินใจและกฎพื้นฐาน

2.2 เพื่อศึกษาว่าเทคนิคใดเหมาะกับชนิดข้อมูลแบบใดและการจำแนกข้อมูลแบบใดเหมาะสมกับข้อมูลที่มีมิติสูง

2.3 เพื่อเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลในกรณีที่จำนวนชุดข้อมูลในแต่ละคลาสมีความโน้มเอียง

2.4 เพื่อวิเคราะห์หาปัจจัยที่ทำให้ประสิทธิภาพการจำแนกข้อมูลดีขึ้น

3. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

3.1 เหมืองข้อมูล (Data Mining)

เหมืองข้อมูล คือการนำข้อมูลจำนวนมหาศาลมาผ่านขั้นตอนวิธีบางอย่างเพื่อให้ได้มาซึ่งความรู้บางประการที่จะเป็นประโยชน์ต่อมนุษย์ได้เนื่องจากข้อมูลที่ต้องการทำเหมืองข้อมูลมีปริมาณมาก [4] มีกระบวนการในการเตรียมข้อมูลซึ่งประกอบด้วย การทำความสะอาดข้อมูล (Data Cleaning) การรวบรวมข้อมูล (Data Integration) การแปลงข้อมูล (Data Transformation) หลังจากนั้นจะต้องทำการประเมินคุณภาพของรูปแบบ (Pattern Evaluation) และสุดท้ายคือนำเสนอความรู้ใหม่ที่ได้สู่ผู้ใช้งาน (Knowledge Representation) เพื่อให้สามารถนำข้อมูลที่กลั่นกรองได้นำไปใช้ประโยชน์ เป็นข้อมูลสนับสนุนในการตัดสินใจในเรื่องต่าง ๆ ต่อไป ซึ่งมีด้วยกัน 5 รูปแบบ คือ Association Rule เป็นการค้นหากฎความสัมพันธ์ของข้อมูล Classification and Prediction การจำแนกประเภทและการทำนาย การจัดกลุ่มข้อมูล (Cluster Analysis) ต้องมีความคล้ายกันมากที่สุด การหาค่าผิดปกติที่เกิดขึ้น (Outlier Analysis) หรือข้อมูลบางอย่างไม่น่าจะเป็นจริงได้ และการวิเคราะห์แนวโน้ม (Trend and Evolution Analysis)

3.2 เคเรสเนเบอร์ (K-Nearest Neighbor KNN)

K-Nearest Neighbor Pattern Classifier เป็นวิธีการหนึ่งสำหรับ แก้ปัญหา Non Parametric Estimation สำหรับ Classification ซึ่งข้อมูลไม่เป็นรูปร่างที่ดี จะเป็นข้อมูลที่กระจัดกระจาย หลักการก็คือจะคล้ายๆ กับการขยาย Window เพื่อหาจุดได้เท่ากับจำนวน kn ที่ต้องการ เช่น kn = 5 มันก็จะ

ขยาย Window ไปจนกว่าจะเจอ data จำนวน ตัว จึงจะหยุด 5

จากนั้นก็ทำclassifier ว่ามันอยู่ใกล้กับข้อมูลจุดไหนมากสุด [5]

3.3 การเรียนรู้แบบเบย์ (Naïve Bayes)

การเรียนรู้แบบเบย์สอย่างง่าย จัดเป็นการเรียนรู้แบบมีผลเฉลย (Supervised Learning) ซึ่งกำหนดว่าจะต้องมีการจัดเตรียมข้อมูลฝึกที่มีคำเฉลย ซึ่งหมายถึงประเภทหรือคลาสของข้อมูลที่ต้องการให้เกิดการเรียนรู้เพื่อสร้างคอนเซปต์ของคลาสนั้น การประยุกต์ใช้ทฤษฎีเบย์สกับการจำแนกข้อมูลสามารถทำได้โดยการแทนค่าในสูตร ซึ่งสามารถปรับให้สอดคล้องกับงานการจำแนกข้อมูลได้ดังนี้

$$P(class|attribute) = \frac{P(class)P(attribute|class)}{P(attribute)}$$

จากสูตรข้างต้นสามารถอธิบายได้ว่าในการจำแนกประเภทของข้อมูล (Class) เราใช้การคำนวณค่าความน่าจะเป็นที่ข้อมูลนั้นอยู่ในคลาสใด ๆ เมื่อเราได้เห็นค่าของแอททริบิวท์ของข้อมูลนั้น

อย่างไรก็ตามข้อมูลแต่ละเรคคอร์ดมีแอททริบิวท์มากกว่าหนึ่งตัว ดังนั้น สูตรข้างต้นจึงเขียนใหม่ได้ดังนี้

$$P(class|a_1, a_2, \dots a_n) = \frac{P(class)P(a_1, a_2, \dots a_n|class)}{P(a_1, a_2, \dots a_n)}$$

โดยที่ a_i คือ ค่าของแอททริบิวท์ใด ๆ ที่ ปรากฏในข้อมูลเรคคอร์ดที่ต้องการจำแนก [1]

3.4 ต้นไม้ตัดสินใจ (Decision Tree)

เป็นวิธีการหนึ่งที่สำคัญในการจำแนกข้อมูล โดยมีลักษณะเป็น Flow Chart เหมือนโครงสร้างต้นไม้ ที่แต่ละโหนดแสดงคุณลักษณะ (attribute) ที่ใช้ทดสอบข้อมูล แต่ละกิ่งแสดงผลในการทดสอบและลีฟโหนด (Leaf Node) แสดงกลุ่มหรือคลาสที่กำหนดไว้ซึ่งง่ายต่อการปรับเปลี่ยนเป็นกฎในการจำแนกข้อมูล [1]

3.5 กฎพื้นฐาน (Rule-Based)

เทคนิคสำหรับการแยกประเภทระเบียนโดยการใช้รูปแบบกฎ "if…then…" ซึ่งกฎที่นำมาสร้างโมเดลจะถูกนำเสนอใน รูปแบบการเชื่อมต่อกัน เช่น R= (rl V r2…V rk) โดยที่ R คือ กลุ่มของกฎ และ ri คือ กฎที่ใช้แยกหรือกฎแต่ละอัน แต่ละกฎที่ใช้แยกสามารถถูกอธิบายโดยสมการที่ (1)

$$r_i: \quad (\text{เงื่อนไข}) \rightarrow \; y_i \quad (1)$$

ด้านซ้ายมือของกฎมักเรียกว่า Rule Antecedent หรือ precondition ซึ่งมันจะมีการเชื่อมต่อของ Attribute Set: ดังแสดงในสมการที่ 2

$$\text{Condition}_i = \{(A1\ op\ v1) \wedge (A2\ op\ v2) \wedge (A_k\ op\ V_k)\} (2)$$

ที่ (A_i , v_j) คือ คู่ของ attribute และค่า และ op คือตัวปฏิบัติการแบบตรรกะที่ถูกเลือกจาก set {=, , <, >, >=,<=} ในขณะที่ทางด้านขวามือของกฎ เรียกว่า rule consequent ซึ่งจะใช้ทำนาย class y_i [4]

4. วิธีการดำเนินการวิจัย

ในการดำเนินงานวิจัยครั้งนี้มีวิธีดำเนินการวิจัยทั้งหมด 4 ขั้นตอนด้วยกันคือ การเตรียมข้อมูล การสร้างโมเดล และการคำนวณประสิทธิภาพ แสดงรายละเอียดตามขั้นตอนดังนี้

4.1 การเตรียมข้อมูล ในการวิจัยครั้งนี้ใช้ชุดข้อมูลมาตรฐานจาก UCI Machine Learning Repository จำนวน 4 ชุดข้อมูล กลุ่มข้อมูลผู้ป่วยโรคหัวใจ (Heart Disease) สถานรับเลี้ยงเด็กเล็ก (Nursery) จดหมายขยะ (Spam Mail) และยานพาหนะ (Vehicle) แสดงรายละเอียดของข้อมูลทั้ง 4 ชุดข้อมูลดังตารางที่ 1 รายละเอียดชุดข้อมูลที่ใช้ในงานวิจัย

ตารางที่ 1: รายละเอียดชุดข้อมูลที่ใช้ในงานวิจัย

ชุดข้อมูล	แอทริบิวต์	จำนวน	คลาส
Heart Disease	13	270	Absent = 150 Present = 120
Nursery	8	12960	Class 1 = 4320 Class 2 = 2 Class 3 = 328 Class 4 = 4266 Class 5 = 4044
Spam Mail	58	4601	Not Spam = 2788 Spam = 1813

ตารางที่ 1: (ต่อ)

ชุดข้อมูล	แอทริบิวต์	จำนวน	คลาส
Vehicle	19	846	Opel = 212 Saab = 217 Bus = 218 Van = 199

4.2 ออกแบบการทดลอง

เป็นการออกแบบโมเดลสำหรับใช้ในการทดลองตามวัตถุประสงค์ที่กำหนดไว้ ซึ่งแบ่งข้อมูลออกเป็น 3 ชุดการทดลอง โดยมีแสดลรายละเอียดดังต่อไปนี้

1) เลือกทุกแอททริบิวต์ และเลือกจำนวนข้อมูลในแต่ละคลาสมีจำนวนข้อมูลใกล้เคียงกัน

2) เลือกทุกแอททริบิวต์ และเลือกจำนวนข้อมูลในแต่ละคลาสมีจำนวนข้อมูลโน้มเอียงไปคลาสบวก และ คลาสลบ

3) จัดกลุ่มแอททริบิวต์สำหรับข้อมูลที่เป็นค่าต่อเนื่องหรือจัดทำนอร์มอลไลซ์ข้อมูล

4) ปรับลดแอททริบิวต์ในแต่ละชุดข้อมูล

4.3 การคำนวณประสิทธิภาพ

ในการวิจัยครั้งนี้ใช้วิธีวัดประสิทธิภาพของการจำแนกข้อมูลอยู่ 3 วิธี ได้แก่ การหาค่าความแม่นยำ (Precision : P) ค่าความระลึก (Recall : R) และการวัดคุณภาพ (F1 measure)

5. ผลการทดลอง

ตารางที่ 2: แสดงประสิทธิภาพความถูกต้อง

Date Set	KNN	Naïve Bayes	Decision Tree	Rule-Based
Heart Disease	77.40	84.07	79.63	79.63
Nursery	98.38	90.32	97.05	99.20
Spam Mail	90.78	79.29	92.98	93.59
Vehicle	69.86	44.79	72.46	71.51

ตารางที่ 2 เป็นการหาประสิทธิภาพความถูกต้องของการจำแนกข้อมูลทั้ง 4 เทคนิค ในชุดข้อมูลฝึกสอนจะเลือกทุกแอทริบิวต์ และทดสอบข้อมูลแบบข้ามสิบทบ (10-Fold Cross Validation) เพื่อประเมินประสิทธิภาพการจำแนกข้อมูล

พบว่าเทคนิคการจำแนกข้อมูลแบบกฎพื้นฐาน มีประสิทธิภาพการจำแนกข้อมูลสูงที่สุด รองลงมาคือเทคนิคเคเนียเรสเนเบอร์ (KNN) ซึ่งเป็นการจำแนกข้อมูลในชุดข้อมูล Nursery มีประสิทธิภาพการจำแนกที่ 99.2% และ 98.38% ตามลำดับ

ตารางที่ 3: แสดงประสิทธิภาพความถูกต้องของข้อมูลที่ในแต่ละคลาสมีจำนวนข้อมูลโน้มเอียง

Date Set	KNN	Naïve Bayes	Decision Tree	Rule-Based
Heart				
1	75.56	82.61	79.35	76.09
2	73.91	88.04	83.69	84.78
3	69.56	75.00	77.17	67.39
Nursery				
1	86.34	85.11	85.49	89.97
2	79.55	76.39	78.16	82.87
3	70.14	76.08	79.94	75.15
Spam				
1	75.65	70.00	71.96	72.61
2	73.48	70.87	80.87	63.04
3	73.48	70.87	80.87	63.04
Vehicle				
1	67.05	47.73	80.68	77.27
2	67.05	44.32	69.32	70.45
3	67.05	43.18	75.00	73.86

ตารางที่ 3 เป็นการหาประสิทธิภาพความถูกต้องในการจำแนกข้อมูลของเทคนิคเคเนียเรสเนเบอร์ (KNN) การเรียนรู้แบบเบย์ (Naïve Bayes) ต้นไม้ตัดสินใจ (Decision Tree) และกฎพื้นฐาน (Rule Based) ทั้ง 4 ชุดข้อมูล โดยทำการแบ่งข้อมูลออกเป็นเป็น 2 ส่วนคือ ชุดข้อมูลฝึกสอนและชุดข้อมูลทดสอบ ซึ่งจะแบ่งเป็น 90% และ 10% ตามลำดับ จากนั้นนำข้อมูลในชุดข้อมูลฝึกสอนมาทำการสร้างแบบทดลองออกเป็น 3 โมเดล คือ โมเดล 1 เป็นการแบ่งข้อมูลในแต่ละคลาสมีจำนวนใกล้เคียงกัน โมเดล 2 เป็นการแบ่งชุดข้อมูลในแต่ละคลาสไม่เท่ากันโดยมีความโน้มเอียงไปในทางลบ และ โมเดล 3 เป็นการแบ่งชุดข้อมูลในแต่ละคลาสไม่เท่ากัน โดยมีความโน้มเอียงไปในทางบวก

6. สรุปและอภิปรายผล

1. จากการทดลองพบว่าเทคนิคการจำแนกข้อมูลแบบกฎพื้นฐาน มีประสิทธิภาพการจำแนกค่อนข้างสูง เนื่องจากกฎ

พื้นฐานจะมีการคัดเลือกเฉพาะแอททริบิวต์ที่มีความสำคัญต่อชุดข้อมูล และตัดแอททริบิวต์เหลือออกไป

2. จากการทดลองพบว่าเทคนิคแต่ละเทคนิคมีความเหมาะสมกับข้อมูลแตกต่างกันสามารถสรุปได้ดังต่อไปนี้

2.1 ชุดข้อมูลผู้ป่วยโรคหัวใจ (Heart Disease) เหมาะกับเทคนิคการเรียนรู้แบบเบย์มากที่สุด เนื่องจากการเรียนรู้แบบเบย์เป็นการหาค่าความน่าจะเป็นของทุกแต่ละแอททริบริต์ทำให้การจำแนกข้อมูลผู้ป่วยโรคหัวใจมีประสิทธิภาพความแม่นยำค่อนข้างสูงคือ 84.07%

2.2 ชุดข้อมูลสถานรับเลี้ยงเด็ก (Nursery) มีความประสิทธิภาพความแม่นยำสูงในทุกเทคนิค เนื่องจากจำนวนข้อมูลในชุดข้อมูลฝึกสอนมีจำนวนมาก สำหรับเทคนิคที่เหมาะสมต่อการจำแนกข้อมูลมากที่สุดคือกฎพื้นฐาน ให้ค่าความแม่นยำสูงถึง 99.20% เพราะแอททริบิวต์ทุกตัวเป็นข้อมูลแบบกลุ่มจึงทำให้การนำข้อมูลดังกล่าวไปสร้างกฎทำได้ง่ายและมีความถูกต้องสูง

2.3 ชุดข้อมูลจดหมายขยะ (Spam Mail) มีความประสิทธิภาพความแม่นยำสูงในทุกเทคนิค เนื่องจากจำนวนข้อมูลในชุดข้อมูลฝึกสอนมีจำนวนมาก สำหรับเทคนิคที่เหมาะสมต่อการจำแนกข้อมูลมากที่สุดคือกฎพื้นฐาน ให้ค่าความแม่นยำสูงถึง 93.59%

2.4 ชุดข้อมูลยานพาหนะ (Vehicle) มีประสิทธิภาพความแม่นยำต่ำในทุกเทคนิค เนื่องมาจากจำนวนคลาสมีหลายคลาสแต่จำนวนข้อมูลในแต่ละคลาสมีจำนวนน้อยทำให้ชุดข้อมูลฝึกสอนมีข้อมูลในการเรียนรู้น้อยส่งผลให้ประสิทธิภาพการจำแนกต่ำ สำหรับเทคนิคที่ให้ความแม่นยำสูงที่สุดคือต้นไม้ตัดสินใจให้ค่าความแม่นยำ 72.46%

3. จำนวนชุดข้อมูลในแต่ละคลาสมีความโน้มเอียงส่งผลให้ประสิทธิภาพความแม่นยำต่ำลง เพราะเทคนิคการจำแนกข้อมูลไม่ได้คำนึงถึงการกระจายตัวของข้อมูล ซึ่งเมื่อศึกษาเพิ่มเติมพบว่าคลาสที่มีจำนวนข้อมูลมากจะถูกจำแนกได้อย่างถูกต้องในขณะที่คลาสที่มีจำนวนข้อมูลน้อยจะถูกจำแนกผิด

4. ปัจจัยที่ทำให้ประสิทธิภาพการจำแนกข้อมูลดีขึ้น คือการเลือกแอทริบิวท์ที่เหมาะสมในแต่ละชุดข้อมูลฝึก จำนวนข้อมูลในชุดข้อมูลฝึกควรมีจำนวนที่มากพอต่อการเรียนรู้ และจำนวนข้อมูลในแต่ละคลาสของชุดข้อมูลฝึกควรจะมีจำนวนที่ใกล้เคียงกัน

เอกสารอ้างอิง

[1] นวลวรรณ สุนทรภิษัช ."ปัญญาประดิษฐ์" ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ .2553.

[2] X. Benjamin, W. Nathalie, "Boosting support vector machines for imbalanced Date Sets", *Knowledge-Based and Intelligent Information and Engineering Systems 2010*, pp. 1-20, 2010.

[3] Mladeni_c D. ,Grobelnik M., Feature Selection for Unbalanced Class Distribution and Naive Bayes, Proc.

[4] Jiawei Han and Micheline Kamber., *Data Mining Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.

[5] Shakhnarovish, Darrell, and Indyk, ed (2005). *Nearest-Neighbor Methods in Learning and Vision.* MIT Press. ISBN 0-262-19547-X.

[6] ภัทราวุฒิ แสงศิริ, ศจีมาจ ณ วิเชียร และพยุง มีสัจ. "การเปรียบเทียบประสิทธิภาพการลดตัวแปร ข้อมูลเข้าที่เหมาะสมสำหรับโครงข่ายประสาทเทียม ระหว่างเทคนิคการเลือกตัวแปรถอยหลังที ละขั้น และการวิเคราะห์องค์ประกอบเพื่อพยากรณ์กลุ่มข้อมูลโรคมะเร็ง" in *The 6th National Conference on Computing and Information Technology. 2010* : 851-858

[7] วิภาดา เวทย์ประสิทธิ์ และพรพิมล ณ นคร. "การลดตัวแปรข้อมูลเข้าสำหรับการพยากรณ์ฝนโดยใช้โครงข่ายประสาทเทียม" in *The Joint Conference on Computer Science and Software Engineering*. November 17-18 2005 : 7-9

[8] Florian Buettner, Sarah Gulliford, Steve Webb and Mike Partridge. Using Bayesian logistic regression with high-order interactions to model radiation-induced toxicities following radiotherapy. *IEEE*,2009