นายวิเชษฐ์ พลายมาศ

THE 7TH NATIONAL CONFERENCE ON COMPUTING AND INFORMATION TECHNOLOGY

FACULTY OF TECHNICAL EDUCATION

PROCEEDINGS OF NCCIT 2011
THE 7TH NATIONAL CONFERENCE ON COMPUTING AND INFORMATION TECHNOLOGY
11-12 MAY 2011
WWW.NCCITNET

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

VOLUME 1

IEEE THAILAND SECTION

บทความวิจัย

การประชุมทางวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 7

11-12 พฤษภาคม 2554

คณะเทคโนโลยีสารสนเทศ
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

WWW.IT.KMUTNB.AC.TH

TOHED-094

VOLUME 1

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

FACULTY OF INFORMATION TECHNOLOGY
KING MONGKUT'S UNIVERSITY OF TECHNOLOGY NORTH BANGKOK (KMUTNB)
BANGKOK, THAILAND

WWW. NCCIT.NET

IEEE
THAILAND SECTION

บทความวิจัย

การประชุมทางวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ
ครั้งที่ 7

11-12 พฤษภาคม 2554

คณะเทคโนโลยีสารสนเทศ
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

WWW.IT.KMUTNB.AC.TH

Time/Paper-ID	Title/Author	Page
16.40-17.00 **NCCIT2011-57**	**Terminal Emulator Wrapper Application Programming Interface** *Thaworn Limwattanachai, Atiwong Suchato, and Proadpran Punyabukkana*	481
17.00-17.20 **NCCIT2011-43**	**Application Tools for Prototyping System by Use Case Concept** *Surachat Juttuporn and Suchada Ratanakongnate*	487
18.00-22.00	*Dinner*	
Room Pradoodaeng 3 (Floor 6)		
Time/Paper-ID	**Title/Author**	**Page**
11.00-12.00	**Tele-Center for Education and Development in Rural Area** *Somjet Phusri*	493

Tuesday May 12, 2011	
8:00-9:00	Registration
9:00-10:00	Invited Keynote Speech by Dr.Andreas Riel
	Topic : The Extension of the Existing IT Oriented Innovation Manager to the Manufactory Field
10:00-10:20	*Coffee Break*
11:00-12:00	Invited Keynote Speech by Professor Dr.Craig Valli
	Topic : Honeypots - Machiavellian security countermeasures for the network age
12:00-13:00	*Lunch*
13:00-16:00	*Parallel Session Presentations*
Room 1	
Session 1: Data Mining, Knowledge Discovery, and Knowledge Management	

Time/Paper-ID	Title/Author	Page
13.00-13.20 **NCCIT2011-58**	**Isan Dharma Alphabet to Thai Languages Translation** *Sophan Promsoda and Pusadee Seresangtakul*	503
13.20-13.40 **NCCIT2011-61**	**Measuring Performance of Algorithms for Short Text Classification :A case study of Thai messages on Twitter** *Supatep Satiman and Thippaya Chintakovid*	509

Room 2

Session 1 : Data Mining, Knowledge Discovery, and Knowledge Management

การศึกษาเปรียบเทียบประสิทธิภาพการจำแนกประเภทของอัลกอริธึมในการจำแนกประเภทหลายชนิด

A comparative study of Multiple Classification Algorithms with Classification Performance

ประดิษฐ์ พิทักษ์เสถียรกุล[1] (Pradit Pituksateerakul[1]) และ วิเชษฐ์ พลายมาศ (Wichet Plaimart)[2]

[1,2]ภาควิชาทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

praditp9@gmail.com, wichet.p@gmail.com

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคการจำแนกประเภทของการทำเหมืองข้อมูล 4 แบบ ได้แก่ นาอีฟเบย์ (NB), ต้นไม้การตัดสินใจ (DT), หลักเกณฑ์ (RB) และเคเอ็นเอ็น (KNN) โดยสร้างโมเดลการทดลองที่ทำการจำแนกประเภทบนชุดข้อมูลที่แตกต่างกัน จำนวน 4 ชุด จากการปรับค่ากำหนดที่ส่งผลต่อค่าประสิทธิภาพ ผลการวิจัยพบว่า DT ให้ประสิทธิภาพที่ดีที่สุด ทั้งด้านความถูกต้อง ประสิทธิภาพโดยรวม และเวลาที่ใช้ในการสร้างโมเดล ซึ่งใกล้เคียงกับRB โดย KNN ใช้เวลาต่ำที่สุดในการสร้างโมเดล

คำสำคัญ: อัลกอริธึมการจำแนกประเภท, การวิเคราะห์ประสิทธิภาพการจำแนกประเภท, ประสิทธิภาพอัลกอริธึมการเรียนรู้, เทคนิคการทำเหมืองข้อมูล

Abstract

This study aimed to compare the performance of 4 classification mining techniques including Naïve Bay, Decision Tree, Rule-Based and K-Nearest Neighbor. The model classification was made on 4 different dataset by improve on multiple conditions that affect the performance of the model. The results showed that DT provides the best performance on accuracy, overall efficiency and computation time on modeling as well as RB. The KNN shortest time to generate a model.

Keyword: Classification algorithm, Classification Performance analysis, Learning algorithm performance, Mining technique

1. บทนำ

1.1 สถานการณ์ปัจจุบัน (Current status)

การเติบโตของการนำเทคนิคเหมืองข้อมูลไปประยุกต์ในงานด้านต่างๆ ในปัจจุบันมีอยู่มากมาย ทั้งด้านธุรกิจ การแพทย์ วิทยาศาสตร์ ด้านไบโออินฟอร์เมติกส์ ด้านระบบความปลอดภัยบนเครือข่าย วิทยาการหุ่นยนต์ ฯลฯ เทคนิคต่างๆของเหมืองข้อมูลก่อให้เกิดการค้นพบความรู้ใหม่ที่เสมือนการค้นพบขุมทรัพย์สามารถนำไปใช้ประโยชน์ในเชิงการค้าได้ อย่างไรก็ตาม ปัญหาสำคัญของเทคนิคต่างๆ ที่ได้รับการพัฒนาขึ้นมานั้นก็คือ แต่ละเทคนิคอาจเหมาะสมกับข้อมูลที่มีคุณลักษณะที่แตกต่างกันไปโดยให้ผลที่เป็นค่าความแม่นยำในระดับที่แตกต่างกัน ทำอย่างไร จึงจะตัดสินได้ว่าเทคนิคใดมีความเหมาะสมกับชุดของข้อมูลคุณลักษณะเช่นใด [1] งานวิจัยส่วนใหญ่จึงพยายามหาประสิทธิภาพของอัลกอริธึมเหมืองข้อมูลในด้านความแม่นยำ (Accuracy-Based) [2]

งานวิจัยที่ผ่านมา จึงพยายามประเมินประสิทธิภาพของเทคนิคของการทำเหมืองข้อมูลบนเงื่อนไขต่างๆ หากพิจารณาในแง่การวัดประสิทธิภาพของอัลกอริธึม (algorithm) กับคุณลักษณะของชุดข้อมูล (dataset) อาจสรุปได้ 4 กลุ่ม กลุ่มแรก หนึ่งอัลกอริธึมต่อหนึ่งชุดข้อมูล กลุ่มที่สอง หนึ่งอัลกอริธึมกับหลายชุดข้อมูล กลุ่มที่สาม หลายอัลกอริธึมกับหนึ่งชุดข้อมูล และกลุ่มที่สี่ หลายอัลกอริธึมกับหลายชุดข้อมูล ในหัวข้อถัดไปจะสรุปงานวิจัยของ 4 กลุ่มดังกล่าว ในปัจจุบัน

1.2 แรงจูงใจ (Motivation)

งานวิจัยนี้ เป็นการหาประสิทธิภาพแบบกลุ่มที่ 4 โดยได้ศึกษาเปรียบเทียบประสิทธิภาพ 4 อัลกอริธึม ได้แก่ นาอีฟเบย์ (Naïve Bay), ต้นไม้การตัดสินใจ (Decision Tree), หลักเกณฑ์ (Rule-Based) และ K-Nearest Neighbor (KNN) ที่ทำการจำแนกกลุ่มกับชุดข้อมูลที่ได้จาก UCI Repository จำนวน 4 ชุด โดยสร้างโมเดลการทดลอง จากการปรับค่าพารามิเตอร์ที่ส่งผลต่อค่าประสิทธิภาพ เพื่อทดสอบประสิทธิภาพของโมเดล (Model Performance)

งานวิจัยนี้ต้องการตอบคำถามว่า

1) แต่ละเทคนิคของการจำแนกประเภทข้อมูลมีประสิทธิภาพแตกต่างกันอย่างไรเมื่อนำมาใช้กับชุดข้อมูลจริงจำนวนมาก

2) แต่ละเทคนิคของการจำแนกประเภทข้อมูลมีจุดแข็งและจุดอ่อนอย่างไร หรือเหมาะสมกับข้อมูลในลักษณะใดเมื่อนำมาใช้กับชุดข้อมูลจริง

3) เทคนิคที่มีประสิทธิภาพดีที่สุดจะสามารถปรับให้ดีขึ้นได้อย่างไร

2. ความเป็นมา

2.1 ปริทัศน์วรรณกรรม (Literature review)

งานวิจัยที่ศึกษาเพื่อเปรียบเทียบประสิทธิภาพของเทคนิคการจำแนกประเภท (classification techniques) แบบต่างๆ ของเหมืองข้อมูลเฉพาะในรอบ 10 ปีมานี้ (2003 – 2011) มีจำนวนมาก แต่ละงานวิจัยต่างก็กำหนดเกณฑ์ในการประเมินประสิทธิภาพที่มีรายละเอียดและวิธีการที่หลากหลายกันไปตามวัตถุประสงค์ที่ต้องการจะวัดค่า แต่วิธีการต่างๆ เหล่านั้นแม้มีความแตกต่างกันแต่ก็มีลักษณะร่วมกันประการหนึ่งก็คือ เป็นการประเมินประสิทธิภาพที่มุ่งเน้นด้านความแม่นยำ (Accuracy-approach) ในการทำนายผลของแต่ละเทคนิคเป็นสำคัญ

คำว่า อัลกอริธึม ในงานวิจัยนี้หมายถึงอัลกอริธึมการจำแนกประเภท (classification algorithm) ซึ่งเป็นเทคนิคสำคัญในการทำเหมืองข้อมูล ส่วนชุดข้อมูล หมายถึง ชุดข้อมูล (data set) ที่บรรจุด้วยลักษณะประจำ (attributes) และกรณีตัวอย่าง (instances) โดยจะเป็นชุดข้อมูลที่ผู้วิจัยจัดเตรียมไว้เองหรือนำมาจากฐานข้อมูลสาธารณะสำหรับเครื่องจักรเรียนรู้ก็ได้ ดังนั้น คำว่าอัลกอริธึมการจำแนกประเภทกับเทคนิคการจำแนกประเภทในงานวิจัยนี้สามารถใช้แทนกันได้ บางครั้ง อาจใช้คำว่าเทคนิค หรืออัลกอริธึม เดี่ยวๆ โดยไม่ระบุการจำแนกประเภท ก็ขอให้เข้าใจว่ามีความหมายเดียวกัน

งานวิจัยนี้จึงสรุปเพื่อให้เห็นภาพรวมงานวิจัยด้านการหาประสิทธิภาพของอัลกอริธึมการจำแนกประเภทในการทำเหมืองข้อมูลแต่ละกลุ่ม โดยเรียงตามปีที่ตีพิมพ์งานวิจัย ดังต่อไปนี้

กลุ่มที่หนึ่ง หนึ่งอัลกอริธึมต่อหนึ่งชุดข้อมูล งานวิจัยในกลุ่มนี้ที่พบมีจำนวน 13 เรื่อง ได้แก่ [3,4,5,6,7,8,9,10,11, 12,13,14] และ [15]

กลุ่มที่สอง หนึ่งอัลกอริธึมกับหลายชุดข้อมูล งานวิจัยในกลุ่มนี้ที่พบมีจำนวน 11 เรื่อง ได้แก่ [16,17,18,19,20,21,22, 23,24,25] และ [26]

กลุ่มที่สาม หลายอัลกอริธึมกับหนึ่งชุดข้อมูล งานวิจัยในกลุ่มนี้ที่พบมีจำนวน 16 เรื่อง ได้แก่ [27,28,29,30,31,32, 33,34,35, 36,37,38,39,40,41] และ [42]

กลุ่มที่สี่ หลายอัลกอริธึมกับหลายชุดข้อมูล งานวิจัยในกลุ่มนี้ที่พบมีจำนวน 17 เรื่อง ได้แก่ [43,44,45,2,46,47,48,49, 50,51,52,53,1,54,55,56] และ [57]

ข้อสังเกตที่ได้จากการรวบรวมงานวิจัยดังกล่าว มี 3 ประการ

ประการแรก หากย้อนหลังเพียง 3 ปี (2008-2010) พบว่าแนวโน้มของงานวิจัยของกลุ่มที่สามและที่สี่มีจำนวนเพิ่มขึ้นสวนทางกับงานวิจัยของกลุ่มที่หนึ่งและกลุ่มที่สองที่มีการเติบโตลดลง อาจสะท้อนให้เห็นว่า ทิศทางในการเปรียบเทียบประสิทธิภาพของอัลกอริธึมการจำแนกประเภทในการทำเหมืองข้อมูลในปัจจุบันได้พยายามแสวงหาคำตอบว่า เทคนิคต่างๆ ที่ถูกพัฒนาขึ้นมานั้นมีประสิทธิภาพแตกต่างกันอย่างไรเมื่อนำมาใช้กับชุดข้อมูลที่มีอยู่จริง

ประการที่สอง เทคนิคต่างๆที่ถูกพัฒนาขึ้นมานั้นมีจุดอ่อนจุดแข็งหรือเหมาะสมกับกับชุดข้อมูลจริงในลักษณะใด

ประการที่สาม มีการนำเทคนิคหลายชนิดเปรียบเทียบประสิทธิภาพกันบนข้อมูลจริงหลายชุดข้อมูล มากกว่าการเปรียบเทียบเพียงเทคนิคเดียวและกับข้อมูลเพียงชุดเดียว พบได้จากการวิจัยในกลุ่มที่สามและสี่มีการเติบโตเพิ่มขึ้นเรื่อยๆ

นอกจากนั้น ยังพบว่า งานวิจัยด้านการเปรียบเทียบประสิทธิภาพของอัลกอริธึมการจำแนกประเภทในการทำเหมืองข้อมูลนั้นยังมีการเติบโตเพิ่มขึ้นเรื่อยๆ งานวิจัยฉบับนี้เป็นการวิจัยที่จัดอยู่ในกลุ่มที่สี่

2.2 อัลกอริธึมการจำแนกประเภท (Classification Algorithms)

อัลกอริธึมที่ใช้ในการทดลองนี้ประกอบด้วย

2.2.1 อัลกอริธึมต้นไม้การตัดสินใจ (Decision Tree)

เป็นเทคนิคพื้นฐานที่นำไปใช้งานจริงอย่างแพร่หลาย เป็นเทคนิคที่ให้ผลลัพธ์ในลักษณะของโครงสร้างต้นไม้ สามารถรองรับปัญหาทั้งแบบการจำแนกประเภทและแบบถดถอย (Regression) [13] ในการทดลองนี้เราเลือกอัลกอริธึม DT J48

2.2.2 อัลกอริธึมนาอีฟเบย์ (Naïve Bayes Algorithm)

ใช้หลักความน่าจะเป็นซึ่งอยู่บนพื้นฐานของ Bayes' Theorem และสมมติฐาน ที่ให้การเกิดของเหตุการณ์ต่างๆ เป็นอิสระต่อกัน เป็นเทคนิคในการแก้ปัญหาแบบการจำแนกประเภทที่ทั้งสามารถคาดการณ์ผลลัพธ์ได้และสามารถอธิบายได้ [2]

2.2.3 อัลกอริธึมหลักเกณฑ์ (Rule-Based Algorithm)

เป็นอัลกอริธึมที่อาศัยการตั้งกฎเกณฑ์เข้ามาช่วยเช่นเดียวกับอัลกอริธึมแบบต้นไม้การตัดสินใจ[16] เป็นพื้นฐานของเทคนิคการจำแนกที่ถูกนำไปใช้ในอัลกอริธึมอื่นๆ แต่เป็นการสร้างกฎจากใหม่ขึ้นจากกรณีตัวอย่างซึ่งไม่จำเป็นต้องมีโครงสร้างเป็นแบบต้นไม้ ในการทดลองนี้เราเลือกอัลกอริธึม RB PART

2.2.4 อัลกอริธึมเคเอ็นเอ็น (KNN: K-Nearest Neighbor)

เป็นอัลกอริธึมที่ใช้ในการจำแนกประเภทที่อยู่ในกลุ่มการเรียนรู้แบบชี้แนะ (Supervised learning) แตกต่างจากเทคนิคอื่นตรงที่ไม่ได้ใช้ข้อมูลฝึกหัด (training data) ในการสร้าง

แบบจำลอง แต่จะใช้ข้อมูลนั้นมาเป็นตัวแบบจำลอง [30] ในการทดลองนี้เราเลือกอัลกอริธึม KNN IBK ทั้งนี้ การทดลองในรอบแรก กำหนดค่า k=1, 3, 5 แต่ผลการทดลองพบว่า ค่าประสิทธิภาพที่ได้ไม่มีความแตกต่างกันอย่างมีนัยสำคัญ แต่เวลาที่ใช้ในการสร้างโมเดลกลับสูงขึ้น ในงานวิจัยนี้จึงกำหนดให้ค่า k = 1

3. การทดลอง (Experiment)

3.1 เหตุผลการเลือกเทคนิค (Algorithm Selection Criteria)

เหตุผลในการเลือกเทคนิคของการจำแนกประเภทข้อมูลในงานวิจัยนี้ มี 2 ประการ

ประการแรก เทคนิคการจำแนกประเภททั้ง 4 แบบ ได้แก่ ต้นไม้การตัดสินใจ (DT) นาอีฟเบย์ (NB) หลักเกณฑ์ (RB) และเคเอ็นเอ็น (K-NN) ล้วนเป็นเทคนิคพื้นฐานที่ได้รับความนิยมนำไปใช้ทั้งในเชิงทดลองและในสถานการณ์จริง

ประการที่สอง เนื่องจากจากการสำรวจงานวิจัยในหัวข้อ 2.1 พบว่า แม้ความนิยมในการวิจัยด้านการวัดประสิทธิภาพของเทคนิคการจำแนกประเภทในกลุ่มที่สี่ (หลายเทคนิคบนหลายชุดข้อมูล) จะเพิ่มขึ้นเรื่อยๆ แต่การศึกษาเปรียบเทียบทั้ง 4 เทคนิคร่วมกันบนหลายชุดข้อมูลยังไม่มีการวิจัยมาก่อน

ประการที่สาม ชุดข้อมูลที่ใช้งานเพื่อหาประสิทธิภาพ มีความมีความหลากหลายและเหมาะสมกับการทดสอบด้วย 4 เทคนิคดังกล่าว

ด้วยเหตุนี้ จึงเลือก 4 เทคนิคดังกล่าวมาใช้

3.2 ชุดข้อมูล (Dataset) ที่ทดลอง

ชุดข้อมูลที่ใช้ในการทดลองในงานวิจัยนี้ได้นำมาจากฐานข้อมูลสาธารณะเครื่องจักรเรียนรู้ UCI Machine Learning Repository [58] เป็นฐานข้อมูลจากการสำรวจในสภาพการณ์จริง โดยได้คัดเลือกเฉพาะชุดข้อมูลที่คุณลักษณะที่หลากหลาย สามารถเป็นตัวแทนการหาประสิทธิภาพที่ดีได้ มีความสอดคล้องและเหมาะสมกับการทดสอบด้วย 4 เทคนิคดังกล่าว ชุดข้อมูลมี 4 ชุด ประกอบด้วย (1) adult (2) page-blocks (3) hypothyroid (4) soybean มีคุณลักษณะที่ประกอบด้วยจำนวนแอททริบิวต์ กรณีตัวอย่าง (Instances) และคุณลักษณะของแอททริบิวต์ ดังสรุปในตารางที่ 1

ตารางที่ 1: สรุปคุณลักษณะของชุดข้อมูลที่เลือก

Dataset	Attributes	Instances	Attribute Characteristics
adult	14	48,842	Categorical, Integer
page-blocks	30	3,772	Integer, Real
hypothyroid	36	683	Categorical, Integer
soybean	14	48,842	Categorical, Integer

3.3 โมเดล (Model)

สร้างโมเดลการทดลองพื้นฐาน จำนวน 6 โมเดลที่มีคุณลักษณะแตกต่างกันดังตารางที่ 2 โดยใช้ข้อมูลสำหรับฝึก (Training Set) จำนวนร้อยละ 66 ด้วยวิธีการสุ่ม ใช้ข้อมูลในการทดสอบ (Test Set) ชุดเดียวกัน โดยสร้างโมเดลการทดลอง จากการปรับค่าพารามิเตอร์ที่ส่งผลต่อค่าประสิทธิภาพ จึงได้โมเดลทั้งหมดจำนวนเทคนิคละ 6 โมเดล รวมเป็น 24 โมเดล ทำการสุ่มแบ่งข้อมลที่เลือกไว้เป็นสองส่วน คือ 1) ข้อมูลฝึก (training set) และ 2) ข้อมูลทดสอบ (test set) ทำการสร้างโมเดลจำนวน 6 โมเดล ข้อมูลฝึกจะใช้ในขณะฝึกหรือปรับโครงสร้างโมเดล และข้อมูลทดสอบจะใช้สำหรับทดสอบโมเดลหลังจบการกระบวนการเรียนรู้แล้ว เพื่อทดสอบประสิทธิภาพของโมเดล (Model Performance) โดยตัดสินค่าความถูกต้องจากการรันการทดลองที่มีการกำหนดค่าพารามิเตอร์ที่ต่างกัน และกำหนดค่า k เท่ากับ 1 สำหรับเทคนิคแบบเคเอ็นเอ็น ทั้ง 6 โมเดลดังแสดงในตารางที่ 2

ตารางที่ 2: โมเดลการทดลอง

โมเดล	คุณลักษณะ
1	ใช้แอททริบิวต์ทั้งหมด
2	เลือกแอททริบิวต์ที่เหมาะสมโดยแอพพลิเคชันแนะนำ จำนวน 6 แอททริบิวต์
3	ใช้โมเดล 1 ร่วมกับการปรับค่ากำหนด (Preference setting) แบบ useSupervisedDiscretization
4	ใช้โมเดล 2 ร่วมกับโมเดล 3
5	พิจารณาตัดทิ้งแอททริบิวต์ที่มีค่าลักษณะเด่นจำแนก (Distinct) สูงเกินไป และใช้ Error Reduction Prune (DT)
6	ใช้โมเดล 5 ร่วมกับ การปรับค่ากำหนด (Preference setting) แบบ useSupervisedDiscretization

3.4 การวัดประสิทธิภาพการจำแนกประเภท (Classification Performance Evaluation)

ในการวัดประสิทธิภาพการจำแนกประเภทของอัลกอริธึมทั้ง 4 เทคนิคในแต่ละโมเดลกับชุดข้อมูลทั้งหมด มีเป้าหมายเปรียบเทียบความถูกต้อง (Accuracy-Oriented) ค่าประสิทธิภาพและค่าความซับซ้อนของโมเดล (Complexity) ของแต่ละเทคนิคบนข้อมูลชุดเดียวกันเป็นหลัก จึงอาศัยเกณฑ์ (Evaluation criteria) ดังนี้

ความถูกต้องหาได้จากความแม่น (Accuracy) ส่วนค่าประสิทธิภาพวัดได้ความเที่ยง (Precision) ความระลึกได้ (Recall) จากมาตรวัดเอฟ (F-Measure) ซึ่งเป็นค่าเฉลี่ยฮาร์โมนีถ่วงน้ำหนักของค่าความเที่ยงกับค่าความระลึกได้ และค่าคุณลักษณะดำเนินการบริเวณเส้นโค้ง (ROC--Receiver Operating Characteristic Curve) และความซับซ้อนของโมเดลวัดได้จากเวลาที่ใช้สร้างโมเดล (Generated time)

3.5 แอพพลิเคชัน

ใช้ WEKA version 3.7 ซึ่งเป็นซอฟต์แวร์การทำเหมืองข้อมูลแบบเปิดเผยรหัสต้นฉบับ (Open Source) ในการ

ดำเนินงาน (Run) และวิเคราะห์ชุดข้อมูล แล้วนำผลที่ได้ไปเปรียบเทียบตามเงื่อนไขการทดลองตามที่กำหนดไว้

4. ผลการทดลอง (Experimental Result)

ผลการทดลองเปรียบเทียบแบ่งออกเป็น 3 ด้านได้แก่ ด้านประสิทธิภาพที่ดีที่สุดในการจำแนกประเภท ด้านประสิทธิภาพโดยรวม และด้านความถูกต้อง ให้ผลลัพธ์ดังนี้

(1) ประสิทธิภาพของอัลกอริธึมที่ดีที่สุด โดยพิจารณาความถูกต้อง ประสิทธิภาพ และความซับซ้อนของโมเดลในการจำแนกประเภทของอัลกอริธึมทั้ง 4 ได้แก่ DT, RB, KNN และ NB บนชุดข้อมูล พบว่า DT เป็นอัลกอริธึมที่ให้ประสิทธิภาพดีที่สุดบน 3 ชุดข้อมูล ได้แก่ adult, hypothyroid และ soybean ส่วน RB ให้ประสิทธิภาพที่ดีที่สุดบนชุดข้อมูล page-blocks โดยที่ DT บน adult ด้านความถูกต้องมีค่าความแม่นร้อยละ 86.45 ด้านประสิทธิภาพมีค่าความเที่ยง 0.86 ความระลึกได้ 0.86 ค่ามาตรวัดเอฟ 0.86 และ ROC 0.92 ด้านความซับซ้อนมีค่าเวลาที่ใช้สร้างโมเดล 1.20 วินาที บน hypothyroid ด้านความถูกต้องมีค่าความแม่นร้อยละ 99.53 ด้านประสิทธิภาพมีค่าความเที่ยง 0.97 ความระลึกได้ 0.97 ค่ามาตรวัดเอฟ 0.97 และ ROC 0.97 ด้านความซับซ้อนมีค่าเวลาที่ใช้สร้างโมเดล 0.38 วินาที บน soybean ด้านความถูกต้องมีค่าความแม่นร้อยละ 90.94 ด้านประสิทธิภาพมีค่าความเที่ยง 0.92 ความระลึกได้ 0.91 ค่ามาตรวัดเอฟ 0.91 และ ROC 0.98 ด้านความซับซ้อนมีค่าเวลาที่ใช้สร้างโมเดล 0.50 วินาที ส่วน RB ให้ประสิทธิภาพที่ดีที่สุดบนชุดข้อมูล page-blocks ดังแสดงในตารางที่ 3

ตารางที่ 3: เปรียบเทียบประสิทธิภาพของอัลกอริธึมที่ดีที่สุดบนแต่ละชุดข้อมูล

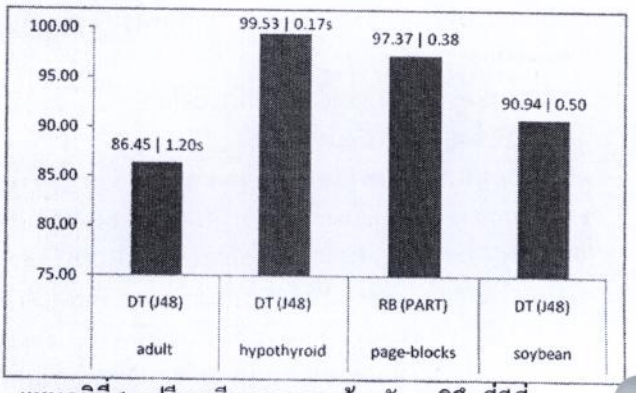
Dataset	Algorithms	Accuracy	Efficiency				Complexity
		Correct	Precision	Recall	F-measure	ROC	Generated Time (sec.)
adult	DT (J48)	86.45	0.86	0.87	0.86	0.88	1.20
hypothyroid	DT (J48)	99.53	0.99	1.00	1.00	0.99	0.17
page-blocks	RB (PART)	97.37	0.97	0.97	0.97	0.97	0.38
soybean	DT (J48)	90.94	0.92	0.91	0.91	0.98	0.50

(2) ด้านประสิทธิภาพโดยรวมของ NB, DT, KNN และ RB บนชุดข้อมูลทั้งหมด พบว่า นอกจากผลลัพธ์ที่อธิบายสรุปในหัวข้อ (1) แล้ว บน hypothyroid ทั้ง DT และ RB ให้ความถูกต้องเท่ากันที่ 99.53 ให้ผลค่าอื่นๆเท่ากัน แต่ RB มีค่า ROC ที่ 0.98 ต่ำกว่า DT เพียงหนึ่งจุด (0.99) ดังแสดงในตารางที่ 4

ตารางที่ 4: เปรียบเทียบประสิทธิภาพโดยรวมของ NB, DT, KNN และ RB

Dataset	Algorithms	Accuracy	Efficiency				Complexity
		Correct	Precision	Recall	F-measure	ROC	Generated Time (sec.)
adult	NB	84.04	0.86	0.84	0.85	0.92	0.49
	DT (J48)	**86.45**	**0.86**	**0.87**	**0.86**	**0.88**	**1.20**
	KNN IBK	85.92	0.85	0.86	0.85	0.89	0.02
	RB (PART)	86.09	0.86	0.86	0.85	0.90	3.19
hypothyroid	NB	98.20	0.98	0.98	0.98	1.00	0.08
	DT (J48)	**99.53**	**0.99**	**1.00**	**1.00**	**0.99**	**0.17**
	KNN IBK	92.75	0.94	0.93	0.93	0.84	0.00
	RB (PART)	**99.53**	**0.99**	**1.00**	**1.00**	**0.98**	**0.17**
page-blocks	NB	95.27	0.96	0.95	0.96	0.99	0.06
	DT (J48)	96.99	0.97	0.97	0.97	0.92	0.47
	KNN IBK	95.91	0.96	0.96	0.96	0.88	0.02
	RB (PART)	**97.37**	**0.97**	**0.97**	**0.97**	**0.97**	**0.38**
soybean	NB	90.08	0.92	0.90	0.90	0.99	0.00
	DT (J48)	**90.94**	**0.92**	**0.91**	**0.91**	**0.98**	**0.50**
	KNN IBK	90.51	0.91	0.91	0.90	0.96	0.00
	RB (PART)	90.08	0.90	0.90	0.90	0.96	0.16

(3) ด้านความถูกต้องของอัลกอริธึมที่ดีที่สุด โดยพิจารณาจากความถูกต้องและเวลาที่ใช้ในหน่วยวินาที บนแต่ละชุดข้อมูล พบว่า อัลกอริธึมที่ดีที่สุดบน adult คือ DT มีความถูกต้องร้อยละ 86.45 ใช้เวลา 1.20 วินาที บน hypothyroid คือ DT มีความถูกต้องร้อยละ 99.53 ใช้เวลา 0.17 วินาที บน page-blocks คือ RB มีความถูกต้องร้อยละ 97.37 ใช้เวลา 0.38 วินาที และบน soybean คือ DT มีความถูกต้องร้อยละ 90.94 ใช้เวลา 0.50 วินาที ดังแสดงในแผนภูมิที่ 1



แผนภูมิที่ 1: เปรียบเทียบความถูกต้องอัลกอริธึมที่ดีที่สุดและเวลาที่ใช้ (วินาที) บนแต่ละชุดข้อมูล

5. สรุป (Conclusion)

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาเปรียบเทียบประสิทธิภาพในด้านความถูกต้อง, ประสิทธิภาพ และความซับซ้อนของโมเดลของอัลกอริธึมการจำแนกประเภทที่ได้รับความนิยม 4 แบบ ได้แก่ ต้นไม้การตัดสินใจ (DT) นาอีฟเบย์ (NB) หลักเกณฑ์ (RB) และเคเอ็นเอ็น (K-NN) บนข้อมูลจริง 4 ชุดที่นำมาจากฐานข้อมูลสาธารณะ

ผลการทดลอง ชี้ให้เห็นว่า DT เป็นอัลกอริธึมที่ดีที่สุดบนชุดข้อมูลที่กำหนดทั้งด้านความถูกต้อง ประสิทธิภาพโดยรวม และเวลาที่ใช้ในการสร้างโมเดล โดยให้ผลลัพธ์ดีที่สุดบนชุดข้อมูลจำนวน 3 ชุดได้แก่ adult, hypothyroid และ soybean ในขณะที่

RB เป็นอัลกอริธึมให้ผลลัพธ์ดีที่สุดบนชุดข้อมูล page-blocks นอกจากนั้น ยังให้ค่าความถูกต้อง ความเที่ยง ความระลึกได้ มาตรวัดเอฟ และเวลาที่ใช้บน hypothyroid ได้เท่ากับ DT แต่ค่า ROC ต่ำกว่า DT เพียงหนึ่งจุดเท่านั้น

มิติด้านความถูกต้อง DT ให้ค่าความถูกต้องสูงที่สุดถึงร้อยละ 99.53 บน hypothyroid ซึ่งมีเพียง 683 กรณีตัวอย่าง แต่ให้ค่าความถูกต้องต่ำที่สุดบน adult ซึ่งมีจำนวน 48,842 กรณีตัวอย่าง ชี้ให้เห็นว่า DT ให้ค่าความถูกต้องต่ำเมื่อทำงานบนจำนวนกรณีตัวอย่างปริมาณมาก

มิติด้านความซับซ้อนของโมเดลซึ่งวัดได้จากเวลาที่ใช้ พบว่า KNN ใช้เวลาต่ำที่สุดในการสร้างโมเดลบนข้อมูลทุกชุด

นอกจากนั้น ผลการทดลองยังชี้ให้เห็นว่า อัลกอริธึมที่มีประสิทธิภาพดีนั้น สามารถปรับแต่งให้มีประสิทธิภาพที่ดีขึ้นได้ โดยการปรับแต่งค่าลักษณะเด่นจำแนกร่วมกับค่ากำหนดให้เหมาะสม

เอกสารอ้างอิง

[1] Ying Yang et al., "To Select or to Weigh: A Comparative Study of Linear Combination Schemes for Superparent-One-Dependence Estimators", *IEEE Transactions On Knowledge and Data Engineering*, vol. 19, no. 12, 2007.

[2] Jin Huang and Jingjing Lu, "Comparing Naive Bayes, Decision Trees, and SVM with AUC and Accuracy", *Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03)*, 2003.

[3] N. Friedman and M. Goldszmidt, "Learning Bayesian Networks with Local Structure," *Proc. 12th Ann. Conf. Uncertainty in Artificial Intelligence*, pp. 252-262, 1996.

[4] Zheng, G.I. Webb, and K.M. Ting, "Lazy Bayesian Rules: A Lazy Semi-Naive Bayesian Learning Technique Competitive to Boosting Decision Trees," *Proc. 16th Int'l Conf. Machine Learning (ICML '99)*, pp. 493-502, 1999.

[5] I. Dhillon, S. Mallela, and R. Kumar, "A Divisive Information-Theoretic Feature Clustering Algorithm for Text Classification," *J.Machine Learning Research (JMLR)*, special issue on variable and feature selection, vol.3, pp.1265-1287, 2003.

[6] G.R. Hjaltason and H. Samet, "Improved Bulk-Loading Algorithms for Quadtrees," *Proc. ACM Int'l Symp. Advances in Geographic Information Systems (GIS)*, pp. 110-115, 2003.

[7] Y. Yang and J.O. Pedersen, "A Comparison Study on Feature Selection in Text Categorization," *Proc. Int'l Conf. Machine Learning*, pp.412-420, 2003.

[8] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter, "Distributional Word Clusters vs. Words for Text Categorization," *J.Machine Learning Research*, vol. 3, pp. 1183-1208, 2003.

[9] T. Silander and P. Myllymaki, "A Simple Approach for Finding the Globally Optimal Bayesian Network Structure," *Proc. 22nd Ann. Conf. Uncertainty in Artificial Intelligence (UAI '06)*, 2006.

[10] H. Zhang, L. Jiang, and J. Su, "Augmenting Naive Bayes for Ranking," *Proc. 22nd Int'l Conf. Machine Learning (ICML '05)*, pp. 1020-1027, 2005.

[11] J. Abellan, "Application of Uncertainty Measures on Credal Sets on the Naive Bayesian Classifier," *Int'l J. General Systems*, vol. 35, no. 6, pp. 675-686, 2006.

[12] R.AGRAWAL, R.SRIKANT. "Fast Algorithms for Mining Association Rules", *Proceedings of 2006 International Conference on Very Large Data Bases*, Santiago, Chile,: 11-16, 2006

[13] Jia Wenguang and Huang LiJing, "Improved C4.5 Decision Tree", *Proceedings of the ACM SIGMOD conference on Management of Data*, : 487-499, 2007

[14] A. McCallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," *Proc. Assoc. for the Advancement of Artificial Intelligence (AAAI '08)* Workshop Learning for Text Categorization, 2008.

[15] Tian Lan, Catherine, Huang and Deniz Erdogmus. "A Comparison of Temporal Windowing Schemes for Single-trial ERP Detection", *Proceedings of the 4th International IEEE EMBS Conference on Neural Engineering*, Antalya, Turkey, 2009: 12-18

[16] C. Hsu and C. Lin. "A comparison on methods for multiclass support vectormachines". *Technical report, Department of Computer Science and Information Engineering*, National Taiwan University, Taipei, Taiwan, 2001.

[17] C. X. Ling and H. Zhang. "Toward Bayesian classifiers with accurate probabilities". *In Proceedings of the Sixth Pacific-Asia Conference on KDD*, 123–134. Springer, 2002.

[18] P. Langley,W. Iba, and K. Thomas. "An analysis of Bayesian classifiers". *In Proceedings of the Tenth National Conference of Artificial Intelligence*, p 223–228.AAAI Press, 2002.

[19] Z. Xie, W. Hsu, Z. Liu, and M.L. Lee, "SNNB: A Selective Neighborhood Based Naive Bayes for Lazy Learning," *Proc. Sixth Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD '02)*, pp. 104-114, 2002.

[20] D. Meyer, F. Leisch, and K. Hornik. "Benchmarking support vector machines". *Technical report,* Vienna University of Economics and Business Administration, 2003.

[21] U. Fayyad and K. Irani. "Multi-interval discretization of continuous-valued attributes for classification learning". *In Proceedings of Thirteenth International Joint Conference on Artificial Intelligence*, p 1022–1027. Morgan Kaufmann, 2003.

[22] L. De Ferrari, "Mining Housekeeping Genes with a Naive Bayes Classifier," *MSc thesis, School of Informatics, Univ. of Edinburgh*, 2005.

[23] Liangxiao Jiang, Harry Zhang, "Learning Instance Greedily Cloning Naive Bayes for Ranking," icdm, pp.202-209, *Fifth IEEE International Conference on Data Mining (ICDM'05)*, 2005

[24] F. Zheng and G.I. Webb, "A Comparative Study of Semi-Naïve Bayes Methods in Classification Learning," *Proc. Fourth Australasian Data Mining Conf. (AusDM '05)*, pp. 141-156, 2005.

[25] M. Ceci, "Naive Bayesian Learning from Structural Data," *PhD dissertation, Dipartimento di Informatica, Univ. of Bari*, Italy, 2005

[26] F. Zheng and G.I. Webb, "Efficient Lazy Elimination for Averaged One-Dependence Estimators," *Proc. 23rd Int'l Conf. Machine Learning (ICML '06)*, pp. 1113-1120, 2006.

[27] R. Kothuri, S. Ravada, and D. Abugov, "Quadtree and R-Tree indices in Oracle Spatial: *A Comparison Using GIS Data," Proc. ACM SIGMOD*,pp. 546-556, 2002.

[28] Wenmin Li, Jiawei Han, Jian Pei, "CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules," icdm, pp.369, *First IEEE International Conference on Data Mining (ICDM'03)*, 2003

[29] Tadashi Nomoto, Yuji Matsumoto, "An Experimental Comparison of Supervised and Unsupervised Approaches to Text Summarization," icdm, pp.630, *First IEEE International Conference on Data Mining (ICDM'01)*, 2003

[30] Yahia, M.E. "K-nearest neighbor and C4.5 algorithms as data mining methods: advantages and difficulties Computer Systems and Applications", *ACS/IEEE International Conference on Mining and Intelligent System*. 2003

[31] K. Torkkola, "Linear Discriminant Analysis in Document Classification," *Proc. IEEE Int'l Conf. Data Mining (ICDM-2003) Workshop Text Mining (TextDM '03)*, 2003.

[32] F. Debole and F. Sebastiani, "An Analysis of the Relative Hardness of Reuters-21578 Subsets," *Proc. Fourth Int'l Conf. Language Resources and Evaluation (LREC '04)*, pp.971-974, 2004.

[33] I.H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, *second ed. Morgan Kaufmann*, 2005.

[34] H. Schutze, D. Hull, and J. Pedersen, "A Comparison of Classifiers and Document Representations for the Routing Problem," *Proc. 18th Ann. Int'l ACM Special Interest Group on Information Retrieval (SIGIR '05) Conf. Research and Development in Information Retrieval*, pp.229-237, 2005.

[35] Y. Huang, D. Erdogmus, S. Mathan, M. Pavel, "Comparison of linear and nonlinear approaches in single trial ERP detection in rapid serial visual presentation tasks," *in Proc. of IEEE International Joint Conference on Neural Networks*, Vancouver, Canada, 2006.

[36] Y. Huang, D. Erdogmus, S. Mathan, M. Pavel, "Boosting linear logistic regression for single trial ERP detection in rapid serial visual presentation tasks," *Proc. of the 28th International Conf. of IEEE EMBS, New York*, 2006.

[37] T. Li, S. Zhu, and M. Ogihara, "Using Discriminant Analysis for Multi-Class Classification: An Experimental Investigation," *Knowledge and Information Systems*, vol. 10, no. 4, pp.453-472, 2006.

[38] Daniele Soria, Jonathan M. Garibaldi, Elia Biganzoli, Ian O. Ellis, "A Comparison of Three Different Methods for Classification of Breast Cancer Data," icmla, pp.619-624, *Seventh International Conference on Machine Learning and Applications*, 2008

[39] Elias Oliveira, Patrick Marques Ciarelli, Claudine Badue, Alberto Ferreira De Souza, "A Comparison between a KNN Based Approach and a PNN Algorithm for a Multi-label Classification Problem," isda, vol. 2, pp.628-633, *Eighth International Conference on Intelligent Systems Design and Applications*, 2008

[40] Ayse Cufoglu, Mahi Lohi, Kambiz Madani, "A Comparative Study of Selected Classifiers with Classification Accuracy in User Profiling," csie, vol. 3, pp.708-712, *WRI World Congress on Computer Science and Information Engineering*, 2009

[41] Marta Capdevila and W. Márquez Flórez, A Communication Perspective on Automatic Text Categorization, *IEEE Transaction on Knowledge Engineering*, (vol. 21 no. 7) pp. 1027-1041, 2009

[42] You Jung Kim, Jignesh M. Patel, "Performance Comparison of the R-Tree and the Quadtree for kNN and Distance Join Queries,"*IEEE Trans. on Knowledge and Data Engineering*, vol. 22, no. 7, pp. 1014-1027, 2010,

[43] L. Kuncheva, "Switching between Selection and Fusion in Combining Classifiers: An Experiment," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 32, no. 2, pp. 146-156, 2002.

[44] Zhihai Wang and Geoffrey I. Webb, "Comparison of Lazy Bayesian Rule and Tree-Augmented Bayesian Learning", *IEEE*. 2002: 490-497

[45] Lian Yan and David J. Miller, "Approximate Maximum Entropy Learning for Classification: Comparison with Other Methods", *IEEE Neural Computation*, vol. 12 no. 9 p243-252, 2003

[46] F. Provost and T. Fawcett. Analysis and visualization of classifier performance: comparison under imprecise class and cost distribution. *In Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, pages 43–48. AAAI Press, 2003

[47] Bernard Zenko, Ljup co Todorovski and Sa so D zeroski, A comparison of stacking with meta decision trees to bagging, boosting, and stacking with other methods, *Proceedings of the 2003 IEEE International Conference on Data Mining* (ICDMi03), 2003

[48] Lina Arbach D. Lee Bennett, Mammographic Masses Classification: Comparison Between Backpropagation Neural Network (B"), K Nearest Neighbors (K"), And Human Readers, *Proc. on Computer Engineering and Communication Engineering (CECE2003*, 2003

[49] R. Caruana, A. Niculescu, G. Crew, and A. Ksikes, "Ensemble Selection from Libraries of Models," *Proc. 21st Int'l Conf. Machine Learning*, 2004.

[50] Y. Yang, K. Korb, K.M. Ting, and G.I. Webb, "Ensemble Selection for Superparent-One-Dependence Estimators," *Proc. 18th Australian Joint Conf. Artificial Intelligence*, 2005.

[51] F. Zheng and G.I. Webb, "A Comparative Study of Semi-NaiveBayes Methods in Classification Learning," *Proc. Fourth Australa-sian Data Mining Conf. (AusDM '05)*, pp. 141-156, 2005.

[52] J. Demsar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *J. Machine Learning Research*, vol. 7, pp. 1-30, 2006.

[53] R.E. Banfield, L.O. Hall, K.W. Bowyer, and W.P. Kegelmeyer, "A Comparison of Decision Tree Ensemble Creation Techniques," *IEEE Trans. Pattern Analysis and Machine Learning*, vol. 29, no. 1, pp. 173-180, Jan. 2007.

[54] Yazhene Krishnaraj, Chandan K. Reddy, "Boosting Methods for Protein Fold Recognition: An Empirical Comparison," bibm, pp.393-396, *IEEE International Conference on Bioinformatics and Biomedicine*, 2008

[55] Ayse Cufoglu, Mahi Lohi and Kambiz Madani, A Comparative Study of Selected Classifiers with Classification Accuracy in User Profiling, *IEEE World Congress on Computer Science and Information Engineering*, 2009.

[56] Mohammed M Mazid, A B M Shawkat Ali and Kevin S Tickle, A Comparison Between Rule Based and Association Rule Mining Algorithms, *IEEE Third International Conference on Network and System Security* 2009.

[57] Mourad Abbas, Kamel Smaili, and Daoud Berkani, TR-Classifier and kNN Evaluation for Topic Identification tasks, *International Journal on Information and Communication Technologies*, Vol. 3, No. 3, June 2010

[58] Frank, A. & Asuncion, A. "UCI Machine Learning Repository" [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2010.