รายงานวิจัยฉบับสมบูรณ์

เรื่อง

วิธีการที่เหมาะสมสำหรับการแบ่งกลุ่มข้อมูลที่ไม่สมดุลสูง

Optimal Methods for Classification of Highly Imbalanced Datasets

โครงการวิจัยนี้ได้รับการสนับสนุนทุนวิจัย

จาก

สำนักงานคณะกรรมการวิจัยแห่งชาติ

ปีงบประมาณ พ.ศ. ๒๕๕๗

คณะผู้วิจัย

นางสาวเบญจภรณ์ จันทรกองกุล	หัวหน้าโครงการวิจัย*
นางสาวสุวรรณา รัศมีขวัญ	ผู้ร่วมวิจัย*
นางสาวสุนิสา ริมเจริญ	ผู้ร่วมวิจัย*
นายภูสิต กุลเกษม	ผู้ร่วมวิจัย*
นายกฤษณะ ชินสาร	ผู้ร่วมวิจัย*
นายอัณณ์นุพันธ์ รอดทุกข์	ผู้ร่วมวิจัย**
นางสาวปิยนุช วรบุตร	ผู้ช่วยนักวิจัย*
นางสาวจรรยา อ้นปันส์	ผู้ช่วยนักวิจัย*

*Knowledge and Smart Technology Research Center

คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา

**ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยรามคำแหง

กิตติกรรมประกาศ

โครงงานวิจัยนี้ได้รับการสนับสนุนทุนวิจัยสำนักงานคณะกรรมการวิจัยแห่งชาติ ปีงบประมาณ พ.ศ. 2556 และได้รับการอนุเคราะห์ทุนสำหรับผู้ช่วยนักวิจัย (นางสาวปิยนุช วรบุตร) ทุนปริญญาเอกในประเทศ โครงการเครือข่ายเชิงกลยุทธ์เพื่อการผลิตและพัฒนาอาจารย์ในสถาบันอุดมศึกษาสำนักงานคณะกรรมการการอุดมศึกษา กระทรวงศึกษาธิการ

คณะผู้วิจัย

สิงหาคม 2557

บทคัดย่อ

วัตถุประสงค์ของขั้นตอนวิธีในการเรียนรู้คือเพื่อให้เกิดอัตราความผิดพลาดในการเรียนรู้ข้อมูลน้อยที่สุด โดยในงานวิจัยนี้ได้ทำการปรับปรุงฟังก์ชันความผิดพลาดที่ใช้วัดอัตราความผิดพลาดสำหรับชุดข้อมูลที่ไม่สมดุลได้อย่างเหมาะสม ซึ่งฟังก์ชันความผิดพลาดส่วนใหญ่จะใช้ค่าน้ำหนักที่เท่ากันทุกคลาส จากที่ทราบกันโดยทั่วไป ข้อมูลที่ไม่สมดุล หมายถึง ชุดข้อมูลที่มีจำนวนสมาชิกของคลาสส่วนมากและคลาสส่วนน้อยจำนวนไม่เท่ากัน ดังนั้นหากใช้ค่าน้ำหนักเท่ากันทุกคลาสจะทำให้การจัดกลุ่มไม่เหมาะสม และปัญหาของการเรียนรู้รูปแบบข้อมูลของชุดข้อมูลไม่สมดุลส่วนใหญ่พบว่าข้อมูลในคลาสส่วนน้อยถูกครอบงำด้วยข้อมูลของคลาสส่วนมาก จึงเป็นผลให้เกิดความเอนเอียงในการจำแนกข้อมูลทำให้ข้อมูลในคลาสส่วนน้อยเกิดความผิดพลาดในการจำแนกกลุ่มมากกว่าคลาสส่วนมาก จากที่กล่าวมางานวิจัยนี้จึงได้ทำการหาพารามิเตอร์ที่เหมาะสมสำหรับปรับปรุงฟังก์ชั่นความผิดพลาดเฉลี่ยกำลังสอง โดยวิธีการที่นำเสนอได้นำอัตราการซ้อนทับกันของข้อมูลและอัตราความไม่สมดุลของข้อมูลมาใช้ร่วมในการปรับปรุงด้วย สำหรับขั้นตอนวิธีในการเรียนรู้ข้อมูลได้ใช้โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับและฟังก์ชันความผิดพลาดที่ทำการปรับปรุง ขอบเขตของชุดข้อมูลที่ใช้ในงานวิจัยนี้เป็นปัญหาการจำแนกชุดข้อมูลที่มี 2 คลาส จาก UCI ผลการทดลองแสดงให้เห็นว่าฟังก์ชันความผิดพลาดที่ทำการปรับปรุงให้ประสิทธิภาพในการจำแนกข้อมูลดีกว่าฟังก์ชั่นความผิดพลาดเฉลี่ยกำลังสองแบบมาตรฐาน เมื่อเปรียบเทียบกับค่า TPR ค่า G-Mean และ F-measurement

Abstract

The objective of learning is to achieve the least error rate. In this research we proposed a modified cost function as a means to properly measure error rate for imbalanced dataset. Most cost functions apply the same weights to all classes. However, it has been known that for imbalanced problem, the number of instances in the majority class is larger than the minority class. Therefore, the application of equal weight to all classes will significantly lead to improper classification boundary. That is, for most learning model, the minority class would be dominated by majority class which then causes a misclassification on the minority class. The objective of this research is to find the appropriate parameters to improve MSE cost function based on overlap ratio and class distribution ratio. Back-propagation algorithm with the proposed modified cost function is used to solve two-class classification problem. UCI datasets are used for the experimentation. The results show that the modified MSE cost function provides a better result than the standard one, based on True-positive rate, G-Mean, and F-measurement.

สารบัญ

สารบัญ (ต่อ)

นิยามศัพท์

$\mathbf{A}$	ชุดข้อมูลคลาสส่วนมาก *A*
$\mathbf{B}$	ชุดข้อมูลคลาสส่วนน้อย *B*
$\mathbf{X}$	ชุดข้อมูลสำรวจ (set observation data set) $\mathbf{X} = \mathbf{A} \cup \mathbf{B}$
$\mathbf{Z}$	ชุดข้อมูลที่ไม่อยู่บนพื้นที่ซ้อนทับ (set non-overlapping data) $\mathbf{Z} = \mathbf{Z_A} \cup \mathbf{Z_{B\cup}}$
$\mathbf{U}$	ชุดข้อมูลที่ยังไม่ระบุพื้นที่ (set of uncertainty data) $\mathbf{U} = \mathbf{X} - \mathbf{Z}$
$\mathbf{Y}$	ชุดข้อมูลที่ไม่อยู่บริเวณขอบของพื้นที่ซ้อนทับ (set of borderline data)
$\mathbf{O}$	ชุดข้อมูลที่อยู่บนพื้นที่ซ้อนทับ (set of overlapping data) $\mathbf{O} = \mathbf{U} - \mathbf{Y}$
IR	อัตราความไม่สมดุล (Imbalanced Ratio)
$maxF$	ค่าฟิชเชอร์ที่มากสุด (maximum Fisher's discriminant ratio)
KL	Kullback-Leibler Divergence
MD	ระยะทางมหาลาโนบิส (Mahalanobis distance)
HD	ระยะทางฮาวดอร์ฟ (Hausdorff distance)
dDBSCAN	dynamic density based spatial clustering of applications with noise
mHD	ระยะทางฮาวดอร์ฟที่ถูกประยุกต์ (modified Hausdorff distance)
mMD	ระยะทางมหาลาโนบิส ที่ถูกประยุกต์ (modified Mahalanobis distance)
$F - minor$	ตัววัด F บนคลาสส่วนน้อย (F-measure on the minority class)

บทที่ 1 บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ในปัจจุบันมีปัญหาการแบ่งข้อมูลที่กำลังได้รับความสนใจ คือ ปัญหาการแบ่งกลุ่มข้อมูลที่ไม่สมดุล (Imbalanced Datasets) ซึ่งเกิดจากการที่มีข้อมูล 2 กลุ่ม (หรือมากกว่า) โดยข้อมูลที่เป็นกลุ่มหลัก (Majority) จะมีจำนวนของข้อมูลมากกว่า ในขณะเดียวกันข้อมูลกลุ่มรอง (Minority) จะมีจำนวนข้อมูลจำนวนน้อย ทั้งนี้เนื่องจากโดยธรรมชาติของความเป็นจริงการที่เราจะกำหนดให้ขนาดของข้อมูลในกลุ่มหลักและกลุ่มรองให้มีขนาดที่เท่าเทียมกันเพื่อการสอนหรือการจัดกลุ่มข้อมูลนั้นเป็นเรื่องยากหรืออาจจะเป็นไปไม่ได้ ดังนั้น จึงเป็นปัญหาที่ท้าทายและมีความยากมากสำหรับการหาขั้นตอนวิธีที่เหมาะสมสำหรับการแบ่งกลุ่มข้อมูลที่ไม่สมดุล ทั้งนี้เนื่องจาก ถ้านำข้อมูลทั้งสองชุดเข้าสู่ขั้นตอนการเรียนรู้พร้อมกันทั้งหมด จะทำให้ผลการแบ่งกลุ่มข้อมูลเกิดความผิดพลาด กล่าวคือ ข้อมูลที่อยู่ในกลุ่มรองถูกครอบงำหรือจะถูกจัดให้ไปอยู่ในกลุ่มหลักทั้งหมด ซึ่งจะนำไปสู่ปัญหาที่เรียกว่า “ปัญหาการแบ่งกลุ่มข้อมูลผิดกลุ่ม (Misclassification)”

ในธรรมชาติของความเป็นจริง ข้อมูลที่ไม่สมดุลเกิดได้ในโดเมนที่หลากหลาย เช่น ปัญหาการแบ่งกลุ่มข้อมูลทางการแพทย์ (ข้อมูลที่ผู้ป่วยที่เป็นมะเร็งและไม่เป็นมะเร็ง) ปัญหาการจัดการความเสี่ยง การรู้จำใบหน้า และรวมไปถึงปัญหาทางเทคโนโลยีสารสนเทศทั่วไป ประเด็นสำคัญของการวิจัยในปัญหาการแบ่งกลุ่มข้อมูลที่ไม่สมดุล คือ การทำให้ข้อมูลที่ไม่สมดุลสามารถทำงานอย่างมีประสิทธิภาพเมื่อเรียนรู้กับขั้นตอนวิธีมาตรฐานที่มีอยู่ในปัจจุบัน ซึ่งขั้นตอนวิธีมาตรฐานจะทำงานได้ดีและมีประสิทธิภาพข้อมูลที่ใช้ในการเรียนรู้ทั้งสองกลุ่มจะต้องสมดุลกัน

วิธีการที่นิยมใช้ในการแก้ปัญหาการแบ่งกลุ่มข้อมูลที่ไม่สมดุล จะแบ่งออกเป็น 2 ระดับ คือ Data Level และ Algorithmic Level สำหรับการจัดการปัญหาในระดับข้อมูล (Data Level) นั้นสามารถแก้ปัญหาได้โดยใช้วิธีการสุ่มตัวอย่างซ้ำ (Resampling Techniques) ซึ่งมีมิธีการเลือกใช้ที่หลากหลาย เช่น Over-Sampling เป็นวิธีการที่ใช้ในการสุ่มเพิ่มจำนวนข้อมูลในกลุ่มรองให้มีจำนวนใกล้เคียงกับข้อมูลในกลุ่มหลัก และ Under-Sampling เป็นวิธีในการสุ่มเลือกข้อมูลจากกลุ่มหลักให้ได้จำนวนที่ใกล้เคียงกับกลุ่มรอง เป็นต้น สำหรับขั้นตอนวิธีที่น่าสนใจในการสุ่มตัวอย่างซ้ำ ได้แก่ SMOTE (Haibo He, 2009), CE-SMOTE (Si Chen, 2010) และ Borderline-SMOTE (Hui Han, 2005) เป็น

ต้น ซึ่งการจัดการข้อมูลที่ไม่สมดุลในระดับนี้ ต้องใช้ความรู้พื้นฐานทางสถิติโดยเฉพาะเรื่องการเทคนิคการสุ่มข้อมูล

สำหรับการจัดการปัญหาในระดับขั้นตอนวิธี (Algorithmic Level) จะเป็นการปรับปรุงขั้นตอนวิธีสำหรับการเรียนรู้ (Learning Algorithms) เช่น การปรับปรุงโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ และปรับปรุงผลการพยากรณ์ด้วย Particle Swarm Optimization (Asrul Adam, 2010), การใช้ SVM และการแปลง Kernel Function (Zhi-QiangZeng, 2009), การประยุกต์ใช้ขั้นตอนวิธีเชิงพันธุกรรม (VicencSoler, 2006) และ การใช้ต้นไม้สำหรับการตัดสินใจ เป็นต้น จากการศึกษาในบทความวิจัยที่เกี่ยวข้อง พบว่า มีการทำการวิจัยในทั้งสองระดับตามที่กล่าวมาข้างต้น และผลการทดลองในทั้งสองระดับก็จะให้ค่าความถูกต้องในระดับที่ใกล้เคียงกัน ขึ้นกับเทคนิคที่เลือกใช้ แต่ผลการทดลองยังไม่ดีพอและยังคงสามารถปรับปรุงให้ผลการทดลองดีขึ้นได้ ดังนั้น ในงานวิจัยนี้จะนำเสนอการปรับปรุงขั้นตอนวิธีสำหรับการแก้ปัญหาทั้งสองระดับ

1.2 วัตถุประสงค์ของโครงการวิจัย

1.2.1 เพื่อศึกษาธรรมชาติของปัญหาการแบ่งกลุ่มข้อมูลที่ไม่สมดุล และโจทย์ปัญหาการวิจัยในโลกความเป็นจริงที่ข้อมูลไม่สมดุล ตัวอย่างปัญหาการแบ่งกลุ่มข้อมูลทางการแพทย์ (ข้อมูลผู้ป่วยที่เป็นมะเร็งและไม่เป็นมะเร็ง)

1.2.2 เพื่อพัฒนาขั้นตอนวิธีสำหรับการเรียนรู้ที่มีประสิทธิภาพสำหรับปัญหาการแบ่งกลุ่มข้อมูลที่ไม่สมดุลสูง

1.2.3 เพื่อให้ผู้ที่สนใจสามารถนำแนวความคิดที่นำเสนอ ไปศึกษาเพื่อทำการพัฒนาหรือประยุกต์ใช้ในงานวิจัยของตนเองต่อไป

1.3 ขอบเขตของโครงการวิจัย

1.3.1 ข้อมูลที่ใช้ในการทดลองเป็นข้อมูลที่ไม่สมดุลที่ได้จาก UCI Machine Learning Repository

1.3.2 ข้อมูลที่ใช้ในการทดลองไม่มีข้อมูลที่สูญหาย (Non-Missing Values)

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.4.1 ได้ขั้นตอนวิธีสำหรับการเรียนรู้เพื่อการแบ่งกลุ่มข้อมูลสำหรับข้อมูลที่ไม่สมดุลสูง

1.4.2 ได้ผลงานวิจัยที่จะสามารถตีพิมพ์ในวารสารที่คาดว่าจะตีพิมพ์ “ECTI-CIT” หรือ Data and Knowledge Engineering (Science Direct)

1.4.3 สามารถนำไปพัฒนาระบบช่วยตัดสินใจสำหรับปัญหาการเรียนรู้เพื่อการแบ่งกลุ่มข้อมูลสำหรับข้อมูลที่ไม่สมดุลสูง เช่น การวินิจฉัยผู้ป่วยโรคมะเร็ง เป็นต้น

1.4.4 ขั้นตอนวิธีที่นำเสนอจะสามารถใช้เป็นต้นแบบในการศึกษาขั้นสูงต่อไป

1.5 ระยะเวลาทำการวิจัยและแผนการดำเนินงานตลอดโครงการวิจัย

ดำเนินการวิจัยแบบ 2 ปีต่อเนื่อง (ปีงบประมาณ 2556-2557)

แผนการดำเนินงานปีที่ 1 (ปีงบประมาณ 2555-2556)

แผนการดำเนินงานวิจัย	**เดือนที่**											
	1	2	3	4	5	6	7	8	9	10	11	12
➢ การจัดเตรียมเครื่องมือ อุปกรณ์และวัสดุวิจัย	⇨											
➢ การศึกษาข้อมูลที่ไม่สมดุลสูง	→	→	→	→	→	→						
➢ การจัดทำรายงานความก้าวหน้าครั้งที่ 1							⇨					
➢ ขั้นตอนการเรียนรู้ข้อมูลที่ไม่สมดุลสูง							→	→	→	→	→	→
➢ การจัดทำรายงานความก้าวหน้าครั้งที่ 2												⇨

แผนการดำเนินงานปีที่ 2 (ปีงบประมาณ 2556-2557)

แผนการดำเนินงานวิจัย	**เดือนที่**											
	1	2	3	4	5	6	7	8	9	10	11	12
➢ การศึกษาการเรียนรู้แบบผสมสำหรับข้อมูลที่ไม่สมดุลสูง	⇨	⇨	⇨	⇨	⇨	⇨						
➢ การจัดทำรายงานความก้าวหน้าครั้งที่ 3							⇨					
➢ พัฒนาเป็น Web-based Application สำหรับผู้ใช้ทั่วไป							⇨	⇨	⇨	⇨	⇨	⇨
➢ จัดส่งรายงาน												⇨

บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ข้อมูลที่ไม่สมดุล (Imbalanced Datasets)

จากที่ทราบกันโดยทั่วไป ข้อมูลที่ไม่สมดุล อาจจะ หมายถึง ข้อมูลที่มีการกระจายตัวที่ไม่เท่าเทียมกัน หรือ อาจจะ หมายถึง ข้อมูลซึ่งอัตราจำนวนสมาชิกในกลุ่มหลักและกลุ่มรองมีจำนวนไม่เท่ากัน เช่น 100:1, 1000:1 หรือ 10000:1 เป็นต้น ยกตัวอย่าง เช่น ข้อมูลผู้ป่วยโรคมะเร็งชนิดหนึ่ง ปัญหานี้คำตอบที่เราต้องการ คือ ผู้ป่วยรายนั้นเป็น (Positive) หรือ ไม่เป็นมะเร็ง (Negative) ซึ่งข้อมูลผู้ป่วยที่ไม่เป็นโรคมะเร็ง (กลุ่มหลัก) อาจจะมีข้อมูลหลายหมื่นคน แต่ข้อมูลผู้ไม่เป็นโรคมะเร็งอาจมีเพียงประมาณหลักร้อยคน (กลุ่มรอง) ดังนั้น ถ้าเรานำข้อมูลทั้งสองกลุ่มมาสอนเพื่อแบ่งกลุ่มข้อมูลพร้อมกันทั้งหมด เราก็จะพบว่า คำตอบของการพยากรณ์เพื่อการแบ่งกลุ่มข้อมูลทุกๆ ข้อมูลจะถูกจัดเป็นกลุ่มผู้ป่วยที่ไม่เป็นโรคมะเร็งทั้งหมด (Positive Class)

2.2 ขั้นตอนวิธีสำหรับการแบ่งกลุ่มข้อมูลที่ไม่สมดุล

เพื่อให้การดำเนินการศึกษากลุ่มข้อมูลที่ไม่สมดุลมีความสมบูรณ์ยิ่งขึ้น ก่อนจะกล่าวถึงขั้นตอนวิธีสำหรับการแบ่งกลุ่มข้อมูลที่ไม่มีความสมดุล ผู้วิจัยจะกล่าวถึงข้อกำหนดเบื้องต้นสำหรับปัญหาการแบ่งข้อมูลที่ไม่สมดุล ดังนี้

กำหนดให้

S คือ กลุ่มข้อมูลที่จะใช้ในการสอน โดยที่ $|S| = m$ และ $S = \{(x_i, y_i)\}, i = 0,\ldots,m$

$x_i \in X$ และมีมิติข้อมูลเป็น n, $(X = \{f_1, f_2,\ldots, f_n\})$

y_i คือ จำนวนกลุ่มคำตอบที่ต้องการของการแบ่งกลุ่ม โดยที $y_i \in Y = \{1,\ldots,C\}$

ในที่นี้ $C = 2$

S_{ma} คือ เซตย่อยของสมาชิกกลุ่มหลัก โดยที่ $S_{ma} \subset S$

S_{mi} คือ เซตย่อยของสมาชิกกลุ่มรอง โดยที่ $S_{mi} \subset S$

$S_{ma} \cup S_{mi} = S$ และ $S_{ma} \cap S_{mi} = \phi$

สำหรับขั้นตอนวิธีที่ใช้สำหรับการแบ่งกลุ่มข้อมูลที่ไม่สมดุล จะประกอบด้วย 3 วิธีการหลัก ๆ คือ Sampling Methods, Cost-Sensitive Methods และ Kernel-Based Methods

Sampling Methods สำหรับวิธีการนี้จะเป็นการประยุกต์เอาวิธีสุ่มตัวอย่างซึ่งเป็นวิธีการทางสถิติ เพื่อสร้างข้อมูลสำหรับการสอน โดยมีจุดประสงค์เพื่อให้จำนวนสมาชิกในข้อมูลทั้งสองกลุ่มมีความสมดุลกัน ซึ่งประกอบด้วย 2 วิธีการใหญ่ ๆ คือ Oversampling และ Undersampling โดยวิธีการ Oversampling จะทำการสุ่มข้อมูลในกลุ่มรองเพื่อสร้างข้อมูลใหม่ของกลุ่มรองให้มีจำนวนเพิ่มมากขึ้นให้ใกล้เคียงหรือเท่ากับจำนวนข้อมูลในกลุ่มหลัก และในทางตรงข้ามวิธีการ Undersampling จะทำการสุ่มเลือกข้อมูลสำหรับการสอนจากข้อมูลในกลุ่มหลัก ให้ได้จำนวนที่ใกล้เคียงกับจำนวนข้อมูลในกลุ่มรอง

Cost-Sensitive Methods วิธีการนี้จะต่างจากวิธีการแรกที่กล่าวมา โดยวิธีการนี้จะพิจารณาขั้นตอนการเรียนรู้ โดยการสร้างสมมติฐานของการแบ่งกลุ่มข้อมูลที่ไม่สมดุลซึ่งให้ค่าความผิดพลาดจากการสอน (Misclassifying examples) ในการแบ่งกลุ่มข้อมูลให้น้อยที่สุด

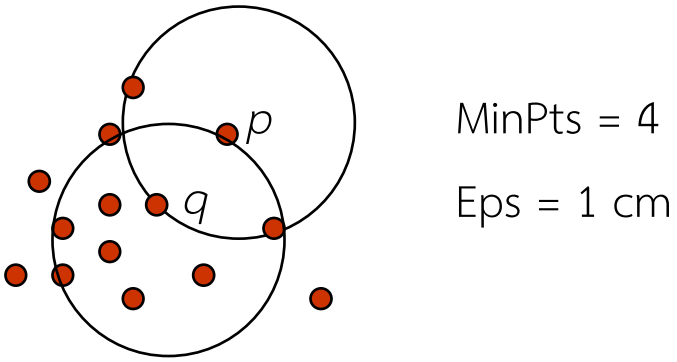
Kernel-based Methods วิธีการนี้เป็นวิธีการใหม่ที่กำลังได้รับความนิยมในการดำเนินการกับกลุ่มข้อมูลที่ไม่สมดุล โดยหลักการแล้วสำหรับวิธีการนี้จะทำการย้ายตำแหน่งของข้อมูล (Map) ที่ไม่สามารถแบ่งกลุ่มได้ในระนาบปกติ โดยการเพิ่มมิติข้อมูลให้สูงขึ้นจนทำให้สามารถแบ่งข้อมูลทั้งสองกลุ่มออกจากกันได้

2.3 การจัดกลุ่มข้อมูลโดยการวัดความหนาแน่น

การจัดกลุ่มข้อมูลด้วยวิธีจะแบ่งกลุ่มตามความหนาแน่นและความต่อเนื่องของข้อมูล พื้นที่ที่ข้อมูลมีความหนาแน่นและต่อเนื่องกันจะถูกเชื่อมต่อกันเป็นพื้นที่ที่ใหญ่ขึ้น เนื่องจากใช้วิธีการเชื่อมต่อกันทำให้รูปร่างของกลุ่มสามารถขยายได้ในทุกทิศทาง และสามารถเกิดเป็นรูปร่างใดๆ ได้ ข้อมูลที่ไม่อยู่ในส่วนที่มีความหนาแน่นจะถูกพิจารณาเป็นข้อมูลผิดปกติ (Outlier) และจะไม่ถูกนำมาพิจารณาในการแบ่งกลุ่ม ข้อดีของวิธีนี้คือ รูปร่างของกลุ่มไม่จำเป็นต้องเป็นทรงกลม และยังสามารถจัดการกับข้อมูลผิดปกติได้ดี ข้อมูลผิดปกติอาจเกิดจากความผิดพลาดของระบบที่สร้างข้อมูล หรืออาจเกิดจากพฤติกรรมที่ผิดปกติของผู้ใช้ ซึ่งถ้าไม่มีวิธีการจัดการกับข้อมูลเหล่านี้ จะทำให้กลุ่มที่แบ่งได้เกิดความผิดเพี้ยน

การแบ่งกลุ่มโดยการวัดความหนาแน่นความน่าจะเป็นมีพารามิเตอร์ที่สำคัญ 2 ตัว คือ

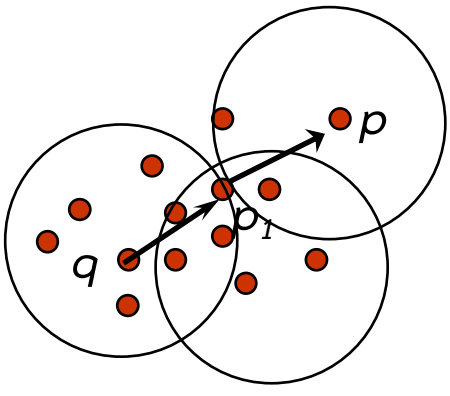
- Eps: รัศมีสูงสุดสำหรับนับจำนวนสมาชิกในกลุ่ม
- MinPts:จำนวนสมาชิกที่น้อยที่สุดภายในรัศมีที่กำหนดให้



รูปที่ 2-1 ตัวอย่างการจัดกลุ่มข้อมูลโดยการวัดความหนาแน่นกำหนดรัศมีเท่ากับ 1 เซนติเมตร และจำนวนสมาชิกที่น้อยที่สุดภายในรัศมีเท่ากับ 4 จุด

ขั้นตอนวิธีสำหรับการจัดกลุ่มข้อมูลโดยการวัดความหนาแน่น

- คำนวณจุดผู้แทนของแต่ละกลุ่ม (Core point) โดยจุดแทนนั้นจะต้องมีจำนวนสมาชิกอย่างน้อยเท่ากับ MinPts ที่กำหนดให้ ดังอย่างอย่างในรูปที่1เราจะได้จุด *p* และ *q* เป็นผู้แทนของสองกลุ่ม
- การคำนวณการเชื่อมต่อองค์ประกอบโดยการวัดความหนาแน่นจากจุด *q* ไปยังจุด *p* จะมีเงื่อนไขดังนี้
    a. *p* belongs to *NEps(q)*
    b. core point condition: *|NEps (q)| >= MinPts*



รูปที่ 2-2 การเชื่อมต่อองค์ประกอบจุด *q* ไปยังจุด *p* โดยมีจุด p_1 เป็นจุดเชื่อม

2.4 ระบบโครงข่ายประสาทเทียม (Neural Networks)

หลายกิจกรรมในชีวิตประจำวันของเราเกี่ยวข้องกับงานปัญญาประดิษฐ์ หรือการรู้จำ ซึ่งปัญหาการรู้จำเป็นปัญหาที่มีความยากและมีความซับซ้อนมากถ้าจะพัฒนาให้เป็นระบบอัตโนมัติในเครื่องคอมพิวเตอร์ แต่ในทางตรงกันข้ามงานดังกล่าวนี้กลับสามารถดำเนินการได้โดยง่ายโดยมนุษย์ กล่าวคือ มนุษย์สามารถรู้จำหรือแยกแยะวัตถุต่างๆ ได้อย่างมากมาย ทั้งที่วัตถุดังกล่าวนั้นกำลังอยู่ในสภาวะแวดล้อมที่มีความหลากหลาย ยกตัวอย่างเช่น นายขาวจะมีความสามารถในการจำเสียงของนายดำ และไม่ว่านายดำจะโทรศัพท์มาคุยกับนายขาวจากสภาพแวดล้อมใดๆ ก็ตาม นายขาวยังคงจำเสียงนายดำได้เสมอ เป็นต้น ดังนั้น จึงเป็นเหตุผลที่สำคัญที่เราต้องพัฒนาระบบการคำนวณ (Computing system) ที่สามารถเข้าใจและเลียนแบบการทำงานของมนุษย์ระบบดังกล่าวนี้ เราเรียกว่าระบบโครงข่ายประสาท (Neural Networks)

ระบบโครงข่ายประสาทเทียม (Artificial Neural Networks: ANN) หรือ บางครั้งอาจสั้นๆ เรียกว่า ระบบโครงข่ายประสาท (Neural Networks) ก็ได้ เกิดมาจากแรงบันดาลใจเกี่ยวกับความต้องการเลียนแบบการทำงานของสมองมนุษย์ด้วยเครื่องคอมพิวเตอร์ เพื่อสร้างระบบคอมพิวเตอร์แบบใหม่ที่มีความสามารถในการเรียนรู้ด้วยตนเอง ซึ่งแตกต่างจากระบบคอมพิวเตอร์ในปัจจุบัน ที่ต้องทำงานตามชุดคำสั่งที่มนุษย์เขียนสั่งไว้ล่วงหน้า (Software/Program) เท่านั้น และ ก็เป็นที่ทราบทั่วไปว่าการประมวลผลของสมองมนุษย์เรานั้น มีความซับซ้อน มีความไม่เป็นเชิงเส้น และเป็นแบบขนานอย่างมาก นอกจากนี้ สมองมนุษย์เรานั้นยังมีความสามารถทางการคำนวณ (เช่น การรู้จำ การรับรู้ และ การควบคุมเครื่องจักร) ได้อย่างรวดเร็วกว่าเครื่องคอมพิวเตอร์ที่เรามีใช้อยู่ในปัจจุบันเราเป็นอย่างมาก ยกตัวอย่างเช่น การมองเห็นของมนุษย์ซึ่งถือว่าเป็นการประมวลผลสารสนเทศแบบหนึ่ง ในระบบการมองเห็นของมนุษย์เรานั้น จะประกอบด้วยขั้นตอนการแทนข้อมูลสภาพแวดล้อมต่างๆ ที่ได้รับ (Data Representation) รวมไปถึงความสามารถในการติดต่อกับสภาพแวดล้อมต่างๆ เหล่านั้นด้วย กล่าวคือ ทันทีที่เรามองเห็น เราจะสามารถแทนข้อมูลที่มองเห็นได้แทบจะทันที และหลังจากนั้นเราจะยังสามารถรู้จำสภาพแวดล้อมต่างๆ นั้นได้ (การรู้จำหมายถึงความสามารถในการจดจำ และ บรรยายสิ่งต่างๆ ที่มองเห็นให้กับผู้อื่นได้ทราบ) กล่าวกันว่า สมองมนุษย์เรานั้น มีความสามารถในการรู้จำสิ่งต่างๆ ที่มองเห็นภายในเวลาประมาณ 100 - 200 มิลลิวินาที ซึ่งไม่มีเครื่องคอมพิวเตอร์ใดสามารถทำได้

แบบจำลองอย่างง่ายของระบบโครงข่ายประสาทเทียมจะมีความคล้ายกับโครงสร้างทางชีววิทยาทางสมองของมนุษย์อย่างมาก ดังแสดงในตารางที่ 1 และ รูปที่ 3โดยมีข้อกำหนดเบื้องต้นดังนี้

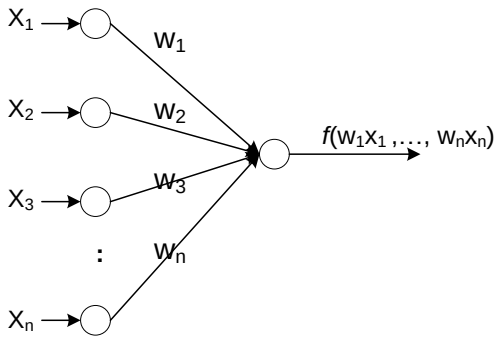
1. ตำแหน่งของค่าน้ำหนักบนโครงข่ายไม่มีความสัมพันธ์กัน

2. คำตอบของแต่ละโหนดจะมีเพียงค่าเดียว ซึ่งจะกระจายไปยังโหนดต่างๆ ที่มีการเชื่อมโยงถึง โดยตำแหน่งของการเชื่อมก็ไม่มีผลเช่นเดียวกัน

3. ข้อมูลที่เข้ามายังแต่ละโหนดในเวลาเดียวกันนั้นจะต้องคงสถานะเดิมไปจนกว่าการคำนวณของฟังก์ชัน $f(w_1x_1,\ldots,w_nx_n)$ จะเสร็จสิ้นลง

ตารางที่ 2-1 คำศัพท์เฉพาะเพื่อเทียบเคียง
(ที่มา : K. Mehrotra et al., Element of Artificial Neural Network)

คำศัพท์เฉพาะชีววิทยาทางสมองของมนุษย์	คำศัพท์เฉพาะของโครงข่ายประสาทเทียม
Neuron	Node/Unit/Cell/Neurode
Synapse	Connection/Edge/Link
Synaptic Efficiency	Connection Strength/Weight
Firing Frequency	Node Output



รูปที่ 2-3 ระบบโครงข่ายประสาทเทียมอย่างง่าย

2.5 ค่าฟิชเชอร์ (Fisher's Discriminant Ratio)

ค่าฟิชเชอร์สามารถคำนวณได้ดังต่อไปนี้

$$f = \frac{(\mu_A - \mu_B)^2}{(\sigma_A^2 + \sigma_B^2)} \quad (2.1)$$

เมื่อ f เป็นค่าฟิชเชอร์ของมิติที่ i, μ_A และ μ_B เป็นค่าเฉลี่ยของคลาสส่วนมาก A และคลาสส่วนน้อย B ของมิติมิติที่ i ตามลำดับ, σ_A^2 และ σ_B^2 เป็นความแปรปรวนของคลาสส่วนมาก A และคลาสส่วนน้อย B ตามลำดับ

2.6 ระยะทางยูคลิด (Euclidean distance)

ระยะทางยูคลิด คือระยะทางปกติระหว่างจุดสองจุดในแนวเส้นตรง สามารถคำนวณได้ดังสมการต่อไปนี้

$$d_{Euclidean}(x, y) = \sqrt{\sum(x_i - y_i)^2} \quad (2.2)$$

เมื่อ $i = 1,2, \dots n$, n คือมิติข้อมูลของข้อมูล x และ y

2.7 ระยะทางมหาลาโนบิส (Mahalanobis distance)

ระยะทางมหาลาโนบิสเป็นการสัดระยะทางระหว่างจุด P และการระจาย D *ถูก*นำเสนอโดย P. C. Mahalanobis เมื่อปี 1936 (Mahalanobis, 1963) ระยะทางมหาลาโนบิสเป็นการวัดที่อาศัยความสัมพันธ์กันระหว่างสองกลุ่มตัวอย่าง กำหนดให้ $\mathrm{g} = (g_1, g_2, \dots, g_d)^T$ เป็นตัวอย่างข้อมูลของ $\mathbf{G}$, $\mu = (\mu_1, \mu_2, \dots, \mu_d)^T$ เป็นค่าเฉลี่ยของ $\mathbf{G}$ ระยะทางมหาลาโนบิสสามารถคำนวณได้ดังสมการต่อไปนี้

$$MD(\mathrm{g}, \mathbf{G}) = \sqrt{(\mathrm{g} - \mu)^T \Sigma^{-1} (\mathrm{g} - \mu)} \quad (2.3)$$

เมื่อ Σ^{-1} เป็นอินเวิร์สโคแวเรียนเมตริกต์ ของ $\mathbf{G}$

2.8 ระยะฮาวดอร์ฟ (Hausdorff distance (*HD*))

ระยะฮาวดอร์ฟถูกนำเสนอโดย by Felix Hausdorff (Hausdorff, 1914) ซึ่งเป็นการวัดระยะระหว่างสมาชิกของ 2 เซต กำหนดให้ $\mathbf{A}$ และ $\mathbf{B}$ be sets of input data, $\mathbf{A} = \{a_1, a_2, \cdots, a_n\}$ และ $\mathbf{B} = \{b_1, b_2, \cdots, b_m\}$ ระยะฮาวดอร์ สามารถคำนวณได้ดังสมการต่อไปนี้

$$H(\mathbf{A}, \mathbf{B}) = Max\big(h(\mathbf{A}, \mathbf{B}), h(\mathbf{B}, \mathbf{A})\big) \tag{2.4}$$

เมื่อ

$$h(\mathbf{A}, \mathbf{B}) = Max_{a \in \mathbf{A}} Min_{b \in \mathbf{B}} ||a - b||$$
$$h(\mathbf{B}, \mathbf{A}) = Max_{b \in \mathbf{B}} Min_{a \in \mathbf{A}} ||b - a||$$

2.9 Kullback-Leibler (*KL*) Divergence

Kullback-Leibler Divergence (Kullback, 1951) แสดงดังสมการต่อไปนี้

$$KL(P||Q) = \sum_i ln \frac{P(i)}{Q(i)} P(i) \tag{2.5}$$

2.10 เครื่องมือในการวัดประสิทธิภาพ

วิธีการวิเคราะห์ความถูกต้อง ในงานวิจัยนี้จะวัดประสิทธิการของผลการทดลองโดยพิจารณาจากค่า accuracy precision recall and F-measure และ G-mean โดยค่าดังกล่าวจะคำนวณได้จากสมการ ต่อไปนี้

ค่าความถูกต้อง (Accuracy)

$$accuracy = \frac{TP + FN}{TP + TN + FP + FN} \tag{2.6}$$

ค่า precision

$$precision = \frac{TP}{TP + FP} \tag{2.7}$$

ค่า recall

$$recall = \frac{TP}{TP + FN} \tag{2.8}$$

ค่า F-measure

$$F - measure = 2 * \frac{precision * recall}{precision + recall} \tag{2.9}$$

ค่า G-mean

$$G - mean = \sqrt{TRN * TPR} \tag{2.10}$$

เมื่อกำหนดให้

$$TRN = \frac{TN}{TN + FP}, \text{ และ } TPR = \frac{TP}{TP + FN},$$

TP = True Positive, TN = True Negative, FN = False Negative, และ FP =False Positive

2.11 การทบทวนวรรณกรรม/สารสนเทศ (Information)ที่เกี่ยวข้อง

Maheshwari et al. (2011) ได้ศึกษาวิจัยเพื่อกำหนดกฎเกณฑ์สำหรับการจัดกลุ่มข้อมูลสำหรับประเภทชุดข้อมูลแบบไม่สมดุล (Imbalanced Datasets) โดยใช้หลักการของวิวัฒนาการพันธุกรรม (Evolutionary Algorithms) เพื่อเพิ่มจำนวนสมาชิกในกลุ่มข้อมูลรองซึ่งเป็นกลุ่มข้อมูลที่เป็นจุดสนใจในการเรียนรู้ ด้วยวิธีการ SMOTE (Synthetic minority over –sampling technique) ก่อนที่ดำเนินการจัดกลุ่มข้อมูลด้วยวิธีการจัดกลุ่ม (Clustering) เพื่อทำความสะอาดข้อมูลโดยการลดความซ้ำซ้อนและข้อมูลที่เป็นส่วนเกิน (Noise) ออกไป โดยประเด็นปัญหาของชุดข้อมูลแบบไม่สมดุลได้แก่ 1. การวัดประสิทธิภาพ ซึ่งต้องใช้เมตริกซ์การประเมินมาช่วยในขบวนการเรียนรู้เพื่อให้ได้คำตอบที่ต้องการ 2. การขาดข้อมูล ซึ่งในกลุ่มข้อมูลรอง จะมีข้อมูลจำนวนค่อนข้างน้อยการจัดโครงสร้างของขอบเขตของกลุ่มเพื่อการตัดสินใจในการแบ่งกลุ่มข้อมูลทำได้ยาก 3. การมีผลกระทบของข้อมูลส่วนเกิน (Noise) มีผลค่อนข้างมากสำหรับกลุ่มข้อมูลรอง 4. การจัดกลุ่มของวิธีการ Machine Leaning แบบทั่วไป มีแนวโน้มที่จะดำเนินการกับข้อมูลในกลุ่มข้อมูลรองว่าเป็นข้อมูลส่วนเกิน จากผลการทดลองพบว่าวิธีการที่นำเสนอสามารถปรับปรุงประสิทธิภาพได้ดี ทั้งในส่วนของค่า F-measure และค่า AUC โดยมีค่าเฉลี่ยของทุกชุดข้อมูลอยู่ที่ 0.784 และ 0.803 ตามลำดับ

Khan et al. (2011) ได้ศึกษาวิจัยเพื่อแก้ปัญหาความไม่สมดุลของชุดข้อมูลสำหรับข้อมูลทางด้านชีววิทยา โดยได้นำหลักการทำงานแบบดั้งเดิมของ Machine Learning ที่เรียกว่า Support Vector Machine มาทำการปรับปรุงสำหรับวิเคราะห์ชุดข้อมูลทางด้าน Eukaryotic genomes จากผลการวิจัยพบว่าจากชุดข้อมูลที่ไม่สมดุลที่มีอัตราส่วน 1:4500 ซึ่งวิธีการทาง Machine Learning แบบดั้งเดิมไม่สามารถดำเนินการได้ดีนักนั้น เมื่อใช้วิธีการที่นำเสนอ กล่าวคือ ได้ใช้หลักการทำ Under Sampling กลุ่มข้อมูลหลัก (Majority Class) และใช้วิธีการแบบ Heuristics เพื่อลดสัดส่วนความไม่สมดุลของชุดข้อมูล หลังจากนั้นจึงได้นำเสนอการทำ SMOTE เพื่อสร้างคุณลักษณะที่ต้องการ (Desired Feature) แล้วทำการเลือกชุดคุณลักษณะที่ดีที่สุดด้วยวิธีการ Feature Selection หลายๆ ตัว กับชุดข้อมูลตัวอย่างของสายพันธุกรรม DNA ซึ่งมีทั้งหมด 11,120 ลักษณะ ให้เหลือเพียง 15 ลักษณะ แล้วใช้หลักการความคล้ายคลึงกัน (Similarity) ในการสร้างชุดทดสอบ ผลการทดลองพบว่าวิธีการที่นำเสนอให้ค่า F-Measure อยู่ที่ระดับ 0.44 ในสัดส่วนของค่า recall และค่า precision ดังนี้ 15%/100% 29%/85% และ 85%/4%

Parvin et al. (2011) ได้ศึกษาวิจัยถึงปัญหาของวิธีการในกลุ่มขั้นตอนวิธีแบบดั้งเดิมของ Machine Learning ที่มีสมมุติฐานว่าการกระจายตัวของชุดข้อมูลที่อยู่ต่างกลุ่มกันมีความสมดุลกันและให้มูลค่าในการจัดกลุ่มผิดพลาดมีความเท่าเทียมกันโดยไม่ได้พิจารณาถึงองค์ประกอบของกลุ่มข้อมูลนั้นๆ ซึ่งในงานวิจัยชิ้นนี้ได้ทำการศึกษาข้อมูลการเป็นโรคมะเร็งซึ่งข้อมูลในบริบทนี้จะเห็นได้ถึงความไม่สมดุลกันของกลุ่มข้อมูลได้อย่างชัดเจน และการจัดกลุ่มผู้ป่วยว่าไม่ป่วยมีนัยสำคัญแตกต่างมากต่อการวินิจฉัยโรค ผู้วิจัยในนำเสนอขั้นตอนวิธีใหม่ Modified Bagging Algorithm ที่ให้ทั้งความรวดเร็วและความมีประสิทธิภาพ โดยใช้รูปแบบขั้นตอนวิธีในการปรับปรุงความไม่สมดุลของข้อมูล โดยใช้ Complex Classification System ที่มีลักษณะคล้ายคลึงกับ Adaboostรวมถึงมีการปรับปรุงวิธีการของ Balance Cascade เพื่อปรับปรุงลดอัตราส่วนการเกิดขึ้นของ False Positive

Sireesha et al. (2011) ได้นำเสนอบทความวิจัยเรื่อง “A Normalized Measure for Estimating Classification Rules for Multi-class Imbalanced Datasets” โดยในงานวิจัยนี้ Prof. Shashiและคณะ ได้นำเสนอวิธีการแบ่งกลุ่มข้อมูลที่ไม่มีความสมดุลแบบหลายกลุ่ม โดยเรียกชื่อว่า “Normalized Strength Score” โดยวิธีการที่เลือกใช้เป็น Association Rules โดยแบ่งกรณีศึกษาออกเป็น ๒ กรณี คือ กรณีที่ itemsetมีค่าความสัมพันธ์ไม่เท่ากันเมื่อพิจารณาเทียบกับข้อมูลสองกลุ่ม และ กรณีที่ itemsetมีความสัมพันธ์เท่ากันเมื่อเทียบกับข้อมูลทั้งสองกลุ่ม วิธีการที่นำเสนอในบทความนี้จะมีข้อดีกว่าวิธีการที่นำเสนอทั่วไป คือ สำหรับข้อมูลไม่สมดุลแบบหลายกลุ่มจะถูกจัดกลุ่ม

เองได้อย่างอัตโนมัติ โดยไม่ต้องมีการแตกเป็นกลุ่มย่อย ๆ สองกลุ่มก่อน แต่วิธีการที่นำเสนอนี้ยังต้องมีการทดลองเพื่อพิสูจน์อีกครั้ง

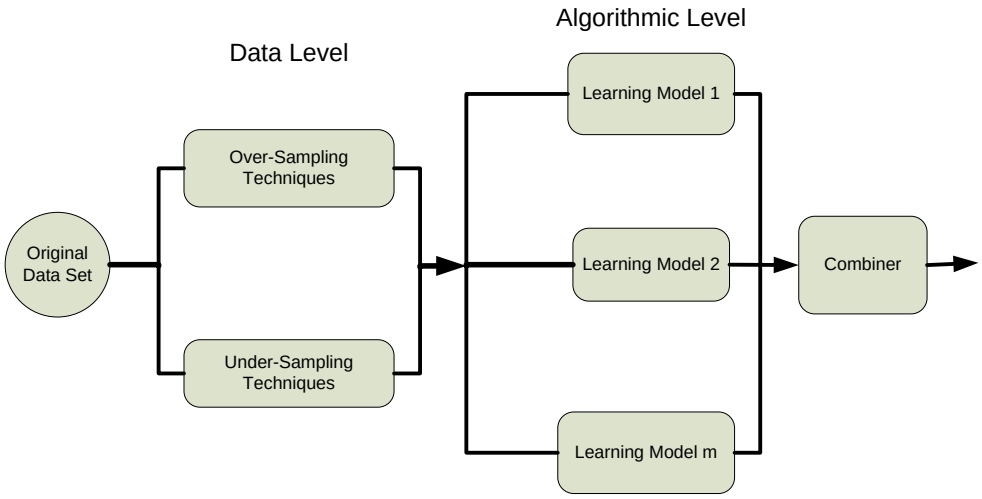
Chumpol et al. (2009) นำเสนอบทความวิจัยเรื่อง “Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem” ในบทความนี้ผู้วิจัยได้นำเสนอวิธีเลือกตัวอย่างของข้อมูลกลุ่มรองที่อยู่บนเส้นขอบเดียวกัน แต่ให้ค่าน้ำหนักที่แตกต่างกัน ซึ่งเรียกว่า “Safe Level” จากผลการทดลองพบว่าวิธีการที่นำเสนอ ให้ค่าความถูกต้อง, ค่า F-Value และ ค่า AUC ที่ดีกว่าผลการทดลองที่ได้จากวิธีการสร้างข้อมูลแบบ SMOTE, แบบ Borderline-SMOTE

Chumpol et al. (2011) นำเสนอบทความวิจัยเรื่อง “DBSMOTE: Density-Based Synthetic Minority Over-Sampling” โดยในงานวิจัยนี้ผู้วิจัยได้ประยุกต์ใช้ขั้นตอนวิธี DB-Scan ซึ่งอาศัยหลักการของการวัดความหนาแน่น (Density-Based Algorithm) เป็นตัวสังเคราะห์ข้อมูลในกลุ่มรอง จากผลการทดลองพบว่าวิธีการที่นำเสนอ ให้ค่าความถูกต้อง, ค่า F-Value และ ค่า AUC ที่ดีกว่าผลการทดลองที่ได้จากวิธีการสร้างข้อมูลแบบ SMOTE, แบบ Borderline-SMOTE และ แบบ Safe-Level-SMOTE

Tan.et al. (2007) นำเสนอบทความวิจัยเรื่อง “Complementary Learning Fuzzy Neural Networks: An Approach to Imbalanced Dataset” ในบทความนี้ผู้วิจัยนำเสนอ Neuro-Fuzzy System สำหรับข้อมูลที่ไม่สมดุล โดยวิธีการที่เลือกใช้ในบทความนี้ ได้แก่ MLP และ RBF

บทที่ 3 วิธีดำเนินการวิจัย

ในงานวิจัยนี้จะศึกษาปัญหาการแบ่งข้อมูลที่ไม่สมดุลในระดับขั้นตอนวิธี (Algorithmic Level) ซึ่งจะได้นำเสนอวิธีการแบ่งกลุ่มข้อมูลที่ไม่สมดุล เช่น การปรับปรุงโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ และปรับปรุงผลการพยากรณ์ด้วย Particle Swarm Optimization (Asrul Adam, 2010), การใช้ SVM และการแปลง Kernel Function (Zhi-QiangZeng, 2009), การประยุกต์ใช้ขั้นตอนวิธีเชิงพันธุกรรม (VicencSoler, 2006) และ การใช้ต้นไม้สำหรับการตัดสินใจ จากนั้น เมื่อศึกษาวิธีการทั้งหมดแล้วผู้วิจัยจะพัฒนาขั้นตอนวิธีการเรียนรู้แบบใหม่ที่มีประสิทธิภาพสำหรับการแบ่งกลุ่มข้อมูลที่ไม่สมดุล โดยจะเลือกใช้วิธี Portfolio และ/หรือ Ensemble Learning (รวม Density-Based Algorithm และ ANN) ซึ่งจะทำให้การเลือกขั้นตอนวิธีและผลการทดลองมีความเป็นไปได้ที่จะได้ผลการทดลองที่ดีกว่าวิธีที่เคยได้มีนักวิจัยนำเสนอไว้ก่อนหน้านี้



รูปที่ 3-1 การเรียนรู้แบบผสม

หมายเหตุ สำหรับทฤษฎีการเรียนรู้แบบต่าง ๆ ที่กล่าวมานี้ เช่น โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ, การพยากรณ์ด้วย Particle Swarm Optimization, การแบ่งข้อมูลด้วย SVM และการแปลง Kernel Function, ขั้นตอนวิธีเชิงพันธุกรรม, การใช้ต้นไม้สำหรับการตัดสินใจ และการเรียนรู้แบบรวม (Ensemble Learning) เป็นขั้นตอนการเรียนรู้มาตรฐานซึ่งนักวิจัยทราบกันดีอยู่แล้ว จึงจะไม่ขอกล่าวถึงรายละเอียดในข้อเสนอโครงการวิจัยที่นี้เนื่องจากวัตถุประสงค์หลักของการทำวิจัยในครั้งนี้ คือ การปรับปรุงขั้นตอนวิธีการเรียนรู้ดังกล่าว ให้เหมาะสมกับปัญหาข้อมูลไม่สมดุล

ในงานวิจัยนี้ได้ทำการศึกษาปัญหาการแบ่งกลุ่มข้อมูลที่ไม่สมดุล ในระดับขั้นตอนวิธี (Algorithmic Level) ซึ่งงานวิจัยนี้ได้แบ่งการศึกษาออกเป็น 2 ส่วนด้วยกันคือ

- การเลือกคุณลักษณะข้อมูลสำหรับชุดข้อมูลที่ไม่สมดุล (Feature selection for imbalanced dataset)

- การปรับปรุงฟังก์ชั่นความผิดพลาด (Error function) สำหรับโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับในการแบ่งกลุ่มข้อมูลที่ไม่สมดุล

- วิธีการแบบผสมผสมสำหรับชุดข้อมูลไม่สมดุลตามสภาพพื้นที่

3.1 การเลือกคุณลักษณะข้อมูลสำหรับชุดข้อมูลที่ไม่สมดุล (Feature selection for imbalanced dataset)

คุณลักษณะข้อมูลเป็นสิ่งที่บอกถึงสารสนเทศที่ซ่อนอยู่ของข้อมูลแต่ละชุด โดยทั่วไปข้อมูลที่มีสารสนเทศมากจะดีกว่าข้อมูลที่มีสารสนเทศน้อยนั่นคือสามารถให้รายละเอียดข้อมูลที่ได้มากกว่า แต่ในบางปัญหาของการจำแนกข้อมูลข้อมูลที่มีรายละเอียดมากเกินไปอาจส่งผลให้ประสิทธิภาพในการจำแนกข้อมูลลดลงเนื่องจากสารสนเทศที่ถูกซ่อนอยู่ในบางคุณลักษณะรบกวนต่ออัตราการรู้จำได้

ในงานวิจัยนี้ได้นำเสนอวิธีการแบบผสมในการเลือกคุณลักษณะข้อมูลไม่สมดุลที่ประกอบด้วยคลาสส่วนมาก และคลาสส่วนน้อย ซึ่งปัญหาของข้อมูลไม่สมดุลเกิดจากจำนวนข้อมูลของคลาสส่วนมากมีจำนวนมากกว่าคลาสส่วนน้อย วิธีการที่นำเสนอประกอบไปด้วย 3 ขั้นตอนได้แก่ การเลือกข้อมูลคลาสส่วนมาก การกำหนดจำนวนคุณลักษณะด้วยค่าไอเกน และการเลือกคุณลักษณะด้วยการประยุกต์ใช้ขั้นตอนวิธีเชิงพันธุกรรมร่วมกับโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ รายละเอียดดังต่อไปนี้

3.1.1 การเลือกข้อมูลคลาสส่วนมาก

สำหรับการเลือกข้อมูลบางส่วนของคลาสส่วนมากเนื่องจากลักษณะข้อมูลไม่สมดุลที่จำนวนข้อมูลในแต่ละคลาสแตกต่างกันมากทำให้เกิดความเอนเอียงในการเรียนรู้งานวิจัยนี้จึงได้ทำการเลือกคลาสส่วนมากให้เท่ากับหรือใกล้เคียงกับคลาสส่วนน้อยด้วยวิธีการหาศูนย์กลางของกลุ่มคลาสส่วนมากและเลือกเฉพาะข้อมูลที่อยู่ใกล้ศูนย์กลางของแต่ละกลุ่ม จากนั้นกำหนดจำนวนคุณลักษณะทำให้จำนวนกลุ่มย่อยของคุณลักษณะที่เป็นไปได้มีจำนวนลดลงเพื่อลดเวลาในการค้นหา

จากนั้นจะนำข้อมูลที่ได้ทำการตัดข้อมูลบางส่วนออกไปมาทำการหาคุณลักษณะข้อมูลที่เหมาะสมต่อไป โดยจะเข้าสู่ขั้นตอนการประเมินคุณลักษณะข้อมูลด้วยการคำนวณหาจำนวน

คุณลักษณะข้อมูลที่เหมาะสมโดยใช้ค่าไอเกน และจากนั้นทำการเลือกคุณลักษณะโดยประยุกต์ใช้ขั้นตอนวิธีเชิงพันธุกรรมร่วมกับโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ รายละเอียดอยู่ในหัวข้อถัดไป

3.1.2 การกำหนดจำนวนคุณลักษณะด้วยค่าไอเกน (Eigen value)

ค่าไอเกนเป็นค่าความผันแปรของตัวแปรทั้งหมดในแต่ละองค์ประกอบ ซึ่งสามารถคำนวณได้ดังสมการที่ 3.1

$$AX = \lambda X \tag{3.1}$$

โดย A แทนเมตริกซ์ขนาด $n \times n$

X แทนเวกเตอร์ขนาด $n \times 1$

λ แทนสเกลาร์ เรียกว่าค่าไอเกน

3.1.3 การเลือกคุณลักษณะด้วยการประยุกต์ใช้ขั้นตอนวิธีเชิงพันธุกรรมร่วมกับโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ

ในการเลือกคุณลักษณะได้ประยุกต์ใช้ขั้นตอนวิธีเชิงพันธุกรรมซึ่งเป็นการค้นหาคำตอบแบบสุ่ม ไม่ต้องทำการเปรียบเทียบทุกคำตอบที่เป็นไปได้ทำให้ค้นหาได้เร็วขึ้น ส่วนค่าความเหมาะสมที่ใช้ในงานวิจัยนี้คือค่าเฉลี่ยของความคลาดเคลื่อนยกกำลังสองจากกระบวนการเรียนรู้ของโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ รายละเอียดดังต่อไปนี้

3.1.3.1 ขั้นตอนวิธีเชิงพันธุกรรม

ในการเลือกคุณลักษณะเป็นการหากลุ่มย่อยของคุณลักษณะที่เหมาะสมที่สุด ดังนั้นข้อมูลมีจำนวนคุณลักษณะมากจะมีจำนวนกลุ่มย่อยของคุณลักษณะมากตามไปด้วย ในงานวิจัยนี้ได้ประยุกต์ใช้ขั้นตอนวิธีเชิงพันธุกรรมในการเลือกคุณลักษณะ เนื่องจากขั้นตอนวิธีนี้เป็นการค้นหาคำตอบแบบสุ่มไม่ต้องทำการเปรียบเทียบทุกคำตอบที่เป็นไปได้ จึงช่วยลดเวลาในการทำงานลงได้ ซึ่งขั้นตอนวิธีเชิงพันธุกรรมเป็นวิธีการแก้ปัญหาที่ใช้แนวทางเดียวกับวิธีการที่สิ่งมีชีวิตปรับตัวเองหรือวิวัฒนาการให้เข้ากับสภาพแวดล้อม ในการถ่ายทอดลักษณะทางพันธุกรรมจะมีกระบวนการที่ทำให้เกิดการเปลี่ยนแปลงที่เรียกว่ากระบวนการวิวัฒนาการ ได้แก่ กระบวนการเลือก (Selection) การไขว้เปลี่ยน (Crossover) และกลายพันธุ์ (Mutation) โดยพิจารณาจากค่าความเหมาะสม (Fitness Function) ที่สอดคล้องกับวัตถุประสงค์ของปัญหา (Objective Function) ของโครโมโซมแต่ละตัวเพื่อนำไปสู่กระบวนการ

คัดเลือก ค่าความเหมาะสมที่ใช้ในงานวิจัยนี้คือค่าเฉลี่ยของความคลาดเคลื่อนยกกำลังสอง (Mean Square Error: MSE) จากกระบวนการเรียนรู้ของโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ เป็นค่าที่ได้จากการหาค่าเฉลี่ยของผลต่างระหว่างค่าที่ได้จากการพยากรณ์กับค่าจริงยกกำลังสองตามสมการที่ 3.2

$$E = \frac{1}{2}\sum_{p=1}^{P}\sum_{k=1}^{M}\left(t_k^{(p)} - y_k^{(p)}\right)^2 \quad (3.2)$$

3.1.3.2 โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ

โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ เป็นอัลกอริทึมที่ออกแบบมาโดยใช้เทคนิคการเรียนรู้ของโครงข่ายประสาทเทียมวิธีหนึ่งที่ใช้ในเพอร์เซพตรอนหลายชั้น ซึ่งโดยทั่วไปจะมีรูปแบบการจัดนิวรอนเป็นชั้น (Layer) ประกอบไปด้วย ชั้นอินพุท (Input Layer) การทำงานของอินพุท จะทำหน้าที่รับข้อมูลเข้าสู่โครงข่าย ชั้นซ่อน (Hidden Layer) ทำหน้าที่ช่วยในการประมวลผลซึ่งในการทำงานของแต่ละชั้นจะถูกกำหนดโดยการทำงานของชั้นอินพุท ค่าถ่วงน้ำหนัก (Weight) และค่าไบแอส (Bias) บนความสัมพันธ์ระหว่างชั้นอินพุทและชั้นซ่อน ชั้นเอาท์พุท (Output Layer) ทำหน้าที่ผลิตผลลัพธ์ของโครงข่าย

ข้อมูลนำเข้า $x_1(p)$, $x_2(p)$,..., $x_n(p)$ โดย n เป็นจำนวนมิติข้อมูลที่ได้จากการขั้นตอนการคำนวณค่าลักษณะเฉพาะ และข้อมูลออกได้แก่ $y_{d,1}(p)$, $y_{d,2}(p)$,..., $y_{d,n}(p)$ เมื่อ d เป็นจำนวนคลาส

- คำนวณผลลัพธ์ในชั้นซ่อนตามสมการที่ 3.3

$$y_i(p) = sigmoid\left[\sum_{i=1}^{n} x_i(p) \cdot w_{ij}(p) - \theta_j\right] \quad (3.3)$$

- คำนวณผลลัพธ์ในชั้นเอาท์พุท โดยสมการที่ 3.4

$$y_k(p) = sigmoid\left[\sum_{j=1}^{m} x_{ij}(p) \cdot w_{jk}(p) - \theta_k\right] \quad (3.4)$$

m เป็นจำนวนนิวรอนในชั้นเอาท์พุท

- ทำการปรับปรุงน้ำหนักในชั้นเอาท์พุทโดยสมการที่ 3.5 – 3.8

$$\delta_k(p) = y_k(p) \cdot [1 - y_k(p)] \cdot e_k(p) \quad (3.5)$$

$$e_k(p) = y_{d,k}(p) - y_k(p) \quad (3.6)$$

$$\Delta w_{jk}(p) = \alpha \cdot y_j(p) \cdot \delta_k(p) \quad (3.7)$$

$$w_{jk}(p+1) = w_{jk}(p) + \Delta w_{jk}(p) \quad (3.8)$$

- ทำการปรับปรุงน้ำหนักในชั้นซ่อนโดยสมการที่ 3.9 – 3.11

$$(p) = y_j(p) \cdot \left[1 - y_j(p)\right] \cdot \sum_{k=1}^{l} \delta_k(p) \cdot w_{jk}(p) \quad (3.9)$$

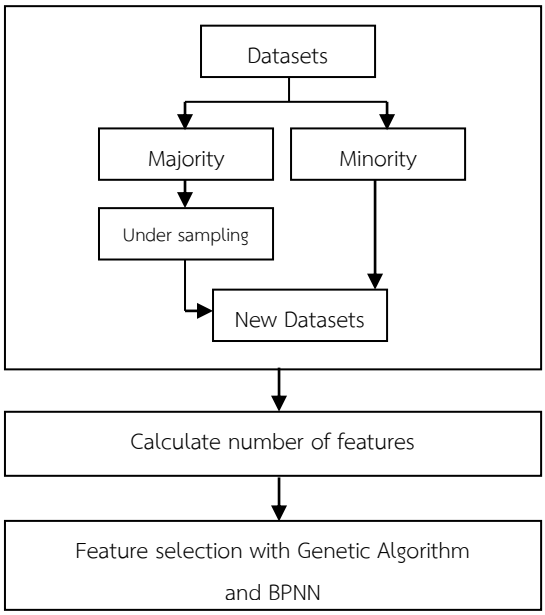
$$\Delta w_{ij}(p) = \alpha \cdot x_j(p) \cdot \delta_j(p) \quad (3.10)$$

$$w_{ij}(p+1) = w_{ij}(p) + \Delta w_{ij}(p) \quad (3.11)$$

3.1.3.3 การเลือกคุณลักษณะข้อมูลไม่สมดุลด้วยวิธีการแบบผสม

ในหัวข้อนี้เป็นรายละเอียดของขั้นตอนของวิธีการที่นำเสนอ ซึ่งได้ถูกนำเสนอในงานประชุมวิชาการระดับชาติ Conference on Knowledge and Smart Technology (KST) ครั้งที่ 4 ปี พ.ศ. 2555 ในหัวข้อเรื่อง การเลือกคุณลักษณะข้อมูลไม่สมดุลด้วยวิธีการแบบผสม ที่มหาวิทยาลัยบูรพา จังหวัดชลบุรี ประเทศไทย รายละเอียดมีดังต่อไปนี้

เนื่องจากข้อมูลที่นำมาใช้เป็นข้อมูลไม่สมดุลที่ประกอบไปด้วยจำนวนข้อมูลในแต่ละคลาสแตกต่างกันมากทำให้เกิดความเอนเอียงในการเรียนรู้ ประกอบกับจำนวนคุณลักษณะข้อมูลที่มีมากบางคุณลักษณะอาจทำให้ประสิทธิภาพในการจำแนกข้อมูลลดลง ดังนั้นงานวิจัยนี้ได้นำเสนอการเลือกคุณลักษณะด้วยวิธีการแบบผสมโดยประกอบไปด้วย 3 ขั้นตอนดังแสดงในรูปที่ 3-2 มีรายละเอียดดังนี้



รูปที่ 3-2 การเลือกคุณลักษณะของข้อมูลไม่สมดุลด้วยวิธีการแบบผสม

1. การเลือกข้อมูลคลาสส่วนมาก ขั้นตอนนี้เป็นการลดจำนวนข้อมูลในกลุ่มคลาสส่วนมากให้มีจำนวนให้เท่ากับหรือใกล้เคียงกับจำนวนข้อมูลของคลาสส่วนน้อย โดยนำชุดข้อมูลที่ไม่สมดุลเฉพาะคลาสส่วนมากมาทำการจัดกลุ่มด้วยวิธีการ k-means ซึ่งเป็นการเรียนรู้แบบไม่มีผู้สอน โดยจะแบ่งข้อมูลออกเป็น k กลุ่ม โดยแทนแต่ละกลุ่มด้วยค่าเฉลี่ยของกลุ่มซึ่งใช้เป็นจุดศูนย์กลางของกลุ่มในการวัดระยะห่างของข้อมูลในกลุ่มเดียวกัน โดยจำนวนกลุ่มที่กำหนดในงานวิจัยนี้เท่ากับอัตราส่วนของจำนวนข้อมูลของคลาสส่วนมากกับจำนวนข้อมูลของคลาสส่วนน้อย จากนั้นทำการเลือกข้อมูลคลาสส่วนมากในแต่ละกลุ่มที่ใกล้ศูนย์กลางของกลุ่ม โดยจำนวนที่เลือกคำนวณจากสมการที่ 4 และ ตัวอย่างการคำนวณแสดงในตารางที่ 1

$$n_i = \frac{N_i}{N_{maj}} \times N_{min} \qquad (4)$$

โดย n_i จำนวนสมาชิกที่เลือกในกลุ่มที่ i

N_i จำนวนสมาชิกในกลุ่มที่ i

N_{maj} จำนวนสมาชิกของคลาสส่วนมาก

N_{min} จำนวนสมาชิกของคลาสส่วนน้อย

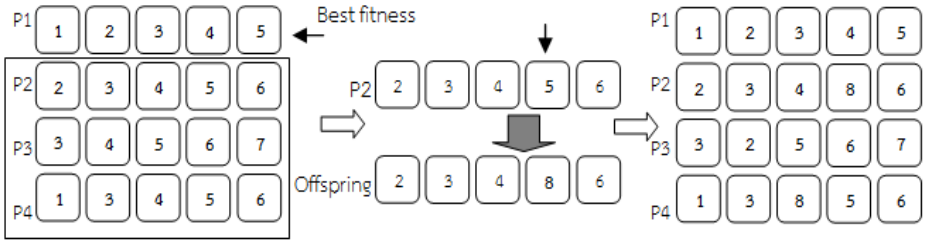
ตารางที่ 3-1 ตัวอย่างการคำนวณจำนวนข้อมูลคลาสส่วนมากที่ถูกเลือกในแต่ละกลุ่ม *(จำนวนคลาสส่วนน้อย=60)*

กลุ่ม	จำนวนข้อมูลคลาสส่วนมาก	จำนวนข้อมูลคลาสส่วนมากที่ถูกเลือก
1	60	18
2	30	9
3	110	33
รวม	200	60

จากนั้นนำข้อมูลคลาสส่วนมากที่เลือกจากทุกกลุ่มมารวมกับข้อมูลคลาสส่วนน้อย จะได้ข้อมูลชุดใหม่ที่มีความสมดุลและนำไปใช้ในขั้นตอนต่อไป

2. การกำหนดจำนวนคุณลักษณะด้วยค่าไอเกน ในการเลือกคุณลักษณะเป็นการเลือกคุณลักษณะย่อยที่เหมาะสมที่สุดจากคุณลักษณะย่อยที่เป็นไปได้ทั้งหมด หากคุณลักษณะย่อยที่เป็นไปได้มีจำนวนมากจะทำให้เวลาในการเลือกมากตามไปด้วย ดังนั้นขั้นตอนนี้จึงได้ทำการกำหนดจำนวนคุณลักษณะ (n) ที่เหมาะสมจากค่าไอเกน ซึ่งคำนวณได้จากสมการที่ 1 โดยพิจารณาจากจำนวนของคุณลักษณะที่มีค่าไอเกนมากกว่าค่าที่กำหนดไว้ในการทดลองนี้กำหนด 0.01 ซึ่งเป็นค่าที่ได้จากการทดลอง
3. การเลือกคุณลักษณะด้วยการประยุกต์ใช้ขั้นตอนวิธีเชิงพันธุกรรมร่วมกับโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ ในขั้นตอนนี้จะทำการค้นหาคุณลักษณะโดยการประยุกต์ใช้ขั้นตอนวิธีเชิงพันธุกรรม โดยกำหนดประชากรเริ่มต้นจากการสุ่มจำนวน 0.1 เท่าของประชากรที่เป็นไปได้ทั้งหมด จำนวนรอบในการทำงาน 300 รอบ รูปแบบของประชากรกำหนดด้วยเลขจำนวนเต็ม 1, 2, 3, .., N เมื่อ N เป็นจำนวนคุณลักษณะทั้งหมด เช่น จำนวนคุณลักษณะที่เลือกเท่ากับ 6 คุณลักษณะ ตัวอย่างรูปแบบประชากรที่เป็นไปได้ 1 2 3 4 5 6, 2 3 4 5 6 7 เป็นต้น
    - ในการคัดเลือก (Selection) ประชากรที่ให้ค่าความเหมาะสมดีสุดจะถูกเก็บไว้ โดยค่าความเหมาะสม (Fitness value) ได้นำเอาค่าเฉลี่ยของความคลาดเคลื่อนยกกำลังสอง ของประชากรแต่ละตัวจากสมการที่ 2 มาประยุกต์ใช้เป็นค่าความเหมาะสม โดยฟังก์ชันกระตุ้นของโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับใช้ฟังก์ชันซิกมอยด์ทั้งชั้นซ่อนและชั้นเอาท์พุท จำนวนนิวรอนในชั้นซ่อนเท่ากับจำนวนอินพุท และชั้นเอาท์พุทเท่ากับ 2 อัตราการเรียนรู้ 0.1 ทำการเรียนรู้ 100 รอบ
    - ประชากรอื่นที่ไม่ถูกคัดเลือกจะถูกนำไปสร้างเป็นประชากรใหม่หรือประชากรลูกหลาน (Offspring) ด้วยกระบวนการกลายพันธุ์ (Mutation) โดยการสุ่ม

เลือกตำแหน่งของประชากรเพื่อทำการเปลี่ยนรหัสดังรูปที่ 3-3 กำหนดให้มีการกลายพันธุ์ในทุกๆ รอบ จำนวนตำแหน่งในการกลายพันธุ์เท่ากับ $0.1 \times (N - n)$ และทำซ้ำขั้นตอนที่ 3.1-3.2 ไปจนกระทั่งครบรอบการทำงานหรือได้ค่าความเหมาะสมตามที่กำหนด และทำการเลือกประชากรที่ให้ค่าความเหมาะสมดีที่สุดไปใช้ในขั้นตอนการทดสอบต่อไป



รูปที่ 3-3 กระบวนการกลายพันธุ์ (Mutation)

3.2 การปรับปรุงฟังก์ชั่นความผิดพลาด (Error function) สำหรับโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับในการแบ่งกลุ่มข้อมูลที่ไม่สมดุล

วิธีการที่นำเสนอได้ประยุกต์ใช้โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับในการแบ่งกลุ่มข้อมูลที่ไม่สมดุล โดยทำการปรับปรุงฟังก์ชั่นความผิดพลาด (Error function) จากฟังก์ชั่นความผิดพลาดเฉลี่ยกำลังสองแบบมาตรฐาน (Standard Mean Square Error: MSE) โดยพิจารณาจากปัจจัยอื่นๆที่เกี่ยวข้อง ได้แก่ อัตราความไม่สมดุล อัตราการซ้อนทับกัน (Overlapping ratio) ของคลาสส่วนมากและคลาสส่วนน้อย ซึ่งวิธีที่ได้นำเสนอนี้ได้ถูกนำเสนอในงานประชุมวิชาการระดับนานาชาติ 2012 7th International Conference on Computing and Convergence Technology (ICCIT, ICEI and ICACT) ที่เมืองโซล ประเทศสาธารณรัฐเกาหลี ในหัวข้อเรื่อง “A Modified Error Function for Imbalanced Dataset Classification Problem” รายละเอียดนำเสนอดังต่อไปนี้

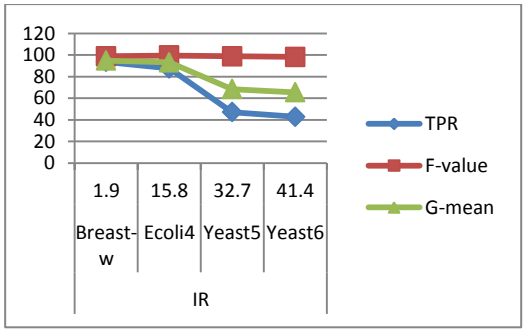
วิธีการที่นำเสนอแบ่งออกเป็น 3 ขั้นตอนดังต่อไปนี้

ขั้นตอนที่ 1 ทำการวัดอัตราความไม่สมดุลและอัตราการซ้อนทับกันของข้อมูล

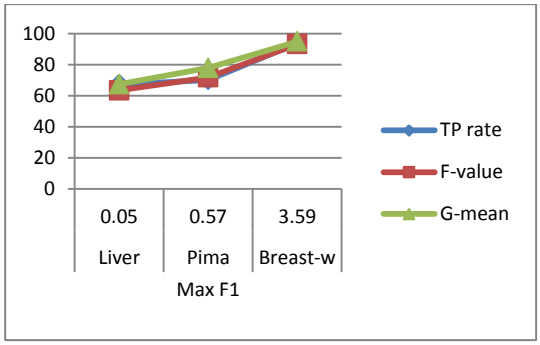
เริ่มต้นนำข้อมูลมาทำการปรับให้อยู่ในช่วง [0..1] จากนั้นทำการวัดอัตราความไม่สมดุลและอัตราการซ้อนทับกันของข้อมูลโดยใช้ Fisher’s discriminant ratio (F1)

ขั้นตอนที่ 2 กำหนดค่าน้ำหนักของคลาสส่วนมาก

ในขั้นตอนนี้ลักษณะของข้อมูลที่ได้จากขั้นตอนที่ 1 นั่นคือ อัตราความไม่สมดุลและอัตราการซ้อนทับกันของข้อมูลมาทำการทดลองเพื่อหาค่าน้ำที่เหมาะสมสำหรับคลาสส่วนมาก รวมถึงศึกษาถึงผลกระทบที่เกิดขึ้นต่อประสิทธิภาพการแบ่งกลุ่มข้อมูล ซึ่งในขั้นตอนนี้ใช้ขั้นตอนวิธีโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับในการจำแนกกลุ่มข้อมูลไม่สมดุล



(ก)



(ข)

ภาพที่ 3-4 (ก) Effect of IRs on the measurements
(ข) Effect of max F1 values on the measurements

ในการทดลองได้แบ่งออกเป็น 2 กรณี คือ กรณีที่ 1 ผลกระทบของอัตราความไม่สมดุลที่มีต่อประสิทธิภาพการแบ่งกลุ่มข้อมูลซึ่งการทดลองนี้ใช้ค่า TPR F และ G-mean เป็นเครื่องมือในการวัดประสิทธิภาพ ภาพที่ 3-4 (ก) แสดงให้เห็นผลการแบ่งกลุ่มข้อมูลของชุดข้อมูล Breast-w, Ecoli4, Yeast5, และ Yeast6, อัตราการซ้อนทับกันของคลาส หรือ ค่าmaximal F1 มีค่า 3.59 3.17 4.17 และ 1.94 ตามลำดับ อัตราความไม่สมดุล (IRs) เท่ากับ 1.86 15.8 32.73 และ 41.41 ตามลำดับ ผลการทดลองแสดงให้เห็นว่าถ้าค่า IRs มีค่าสูงแล้วจะทำให้ค่า TPR ค่า F และค่า G-mean ต่ำ ในทางตรงกันข้าม ถ้า ค่า IRs มีค่าต่ำแล้วจะทำให้ค่า TPR ค่า F และค่า G-mean สูง

ในกรณีที่ 2 ผลกระทบของค่า maximal F1 ต่อประสิทธิภาพการแบ่งกลุ่มข้อมูลซึ่งการทดลองนี้ใช้ค่า TPR F และ G-mean เป็นเครื่องมือในการวัดประสิทธิภาพเช่นเดียวกัน โดยในกรณีที่ 2 นี้ได้

กำหนดอัตราความไม่สมดุลใหล้เคียงกันทุกชุดข้อมูล ภาพที่ 3-4 (ข) แสดงให้เห็นผลการแบ่งกลุ่มข้อมูลของชุดข้อมูล Liver Pima และ Breast-w ซึ่งมีค่า IRs เท่ากับ 1.38 1.87 และ 1.86 ตามลำดับ และค่า maximal F1 เท่ากับ 0.05 0.57 และ 3.59 ตามลำดับ ผลการทดลองแสดงให้เห็นว่า ถ้าค่า maximal F1 มีค่าสูงแล้วค่า TPR ค่า F และค่า G-mean จะสูงด้วย ในทางตรงกันข้าม ถ้าค่า maximal F1 มีค่าต่ำแล้วค่า TPR ค่า F และค่า G-mean จะต่ำด้วย

จากผลการทดลองที่กล่าวมาแล้ว อัตราความไม่สมดุลและอัตราการซ้อนทับกันของคลาสมีผลต่อประสิทธิภาพของการแบ่งกลุ่มข้อมูล ดังนั้นจึงงานวิจัยนี้จึงได้นำทั้งสองปัจจัยที่กล่าวถึงนั่นคืออัตราความไม่สมดุล และอัตราการซ้อนทับกันของคลาสมากทำกำหนดค่าน้ำหนักในการปรับปรุง standard MSE ในขั้นตอนถัดไป โดยค่าน้ำหนักที่กำหนดอยู่ในช่วง 0 ถึง 1 และ r_{maj} เป็นค่าน้ำหนักของคลาสส่วนมาก และ r_{min} ค่าน้ำหนักของคลาสส่วนน้อย จากการทดลองค่าน้ำหนักของคลาสส่วนมากแสดงดังตารางที่ 3-2 ส่วนค่าน้ำหนักของคลาสส่วนน้อยมีค่าเท่ากับ 1

ตารางที่ 3-2 ค่าน้ำหนักของคลาสส่วนมาก $(\boldsymbol{r_{maj}})$

IR	Max F1 values			
	0.00-0.30	0.31-0.70	0.71-2	2.01-5
1.00-10	0.7	0.8	N/A	1
10.01-20	N/A	0.4	N/A	0.8
20.01-30	N/A	0.4	N/A	N/A
30.01-40	N/A	N/A	N/A	0.8
41.01-50	N/A	N/A	0.7	N/A

ขั้นตอนที่ 3 ฝึกสอนและทดสอบ

ในขั้นตอนนี้ข้อมูลจะถูกแบ่งเป็น 2 ชุด คือชุดฝึกสอน และชุดทดสอบในอัตรา 60:40 ในการทดลองใช้ขั้นตอนวิธีโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับในการเรียนรู้ จำนวน 1 ชั้นซ่อน 2 ชั้นผลลัพธ์

$$X^{(P1)} = \left[x_1^{(p1)}, x_2^{(p1)}, \dots, x_N^{(p1)}\right],$$
$$X^{(P2)} = \left[x_1^{(p2)}, x_2^{(p2)}, \dots, x_N^{(p2)}\right]$$

โดยที่ *P1, P2* เป็นข้อมูลของคลาสส่วนมากและคลาสส่วนน้อย N เป็นจำนวนมิติของข้อมูล และ $P = P1 \cup P2$ ดังนั้นข้องมูลนำเข้าจะจัดให้อยู่ในรูปแบบ

$$X^{(P)} = \left[x_1^{(p)}, x_2^{(p)}, \ldots, x_N^{(p)}\right]$$

สำหรับข้อมูลที่เป็นผลลัพธ์ของ X^p จัดให้อยู่ในรูปแบบ

$$t_k^p = \begin{cases} 1 & if\ X^p \text{ is class} \\ 0 & otherwise \end{cases}$$

k = 1,2,..m

ผลลัพธ์ในชั้นซ่อน กำหนดโดย

$$y_j^{(p)} = f\left(\sum_{i=1}^{N} w_{ji}\, x_i^{(p)}\right)$$

w_{ji} เป็นน้ำหนักจาก x_i ไปยังนิวรอล j ในชั้นซ่อน

ผลลัพธ์ในชั้นผลลัพธ์ กำหนดโดย

$$y_k^{(p)} = f\left(\sum_{j=1}^{H} v_{jk} y_j^{(p)}\right)$$

v_{jk} เป็นน้ำหนักจาก $y_j^{(p)}$ ไปยัง $y_k^{(p)}$

ฟังก์ชั่นกระตุ้น (Activation function) ใช้ฟังก์ชั่นซิกมอยด์ กำหนดโดย

$$f(x) = \frac{1}{1+e^{-x}}$$

ที่ชั้นผลลัพธ์ได้ทำการปรับปรุงฟังก์ชัน MSE แบบมาตรฐาน ตามสมการต่อไปนี้

$$E_{prop.}^{(p)} = \begin{cases} \frac{1}{2}\sum_{k=1}^{m} r_{maj}\left(t_k^{(p)} - y_k^{(p)}\right)^2, if\ t_k^{(p)} is\ majority\ class \\ \frac{1}{2}\sum_{k=1}^{m} r_{min}\left(t_k^{(p)} - y_k^{(p)}\right)^2, if\ t_k^{(p)} is\ minority\ class \end{cases}$$

$E_{prop.}^{(p)}$ เป็นผลรวมของค่าความผิดพลาดเฉลี่ยกำลังสอง

m เป็นจำนวนโหนดในชั้นผลลัพธ์

r_{maj}, r_{min} เป็นค่าน้ำหนักของคลาสส่วนมาก และคลาส่วนน้อยตามลำดับ

จากนั้นค่าน้ำหนักจะถูกทำการปรับปรุงตามสมการด้านล่าง

$$\partial_k^{(p)} = -\frac{\partial E_{prop.}^{(p)}}{\partial y_k^{(p)}}$$
$$\partial_k^{(p)} = y_k^{(p)}[1 - y_k^{(p)}]e_k^{(p)}$$

$$e_k^{(p)} = \begin{cases} \mathrm{r_{maj}}(y_k^{(p)} - t_k^{(p)}), if\ t_k^{(p)} is\ majority\ class \\ y_k^{(p)} - t_k^{(p)}, otherwise \end{cases}$$

ปรับปรุงน้ำหนักในชั้นผลลัพธ์

$$\Delta v_{jk}^{(p)} = \propto y_j^{(p)} \partial_k^{(p)}$$
$$v_{jk}^{(p+1)} = v_{jk}^{(p)} + \Delta v_{jk}^{(p)}$$

ปรับปรุงน้ำหนักในชั้นซ่อน

$$\partial_j^{(p)} = y_j^{(p)}[1 - y_j^{(p)}] \sum_{k=1}^{M} \partial_k^{(p)} v_{jk}^{(p)}$$
$$\Delta w_{ji}^{(p)} = \propto X_j^{(p)} \partial_j^{(p)}$$
$$w_{ji}^{(p+1)} = w_{ji}^{(p)} + \Delta w_{ji}^{(p)}$$

โดยทีอัตราการเรียนรู้ $(\propto) = 0.01$ จำนวนรอบ=1000

3.3 วิธีการแบบแบบผสมสำหรับชุดข้อมูลไม่สมดุลตามสภาพพื้นที่

3.3.1 ปัญหาและความสำคัญ

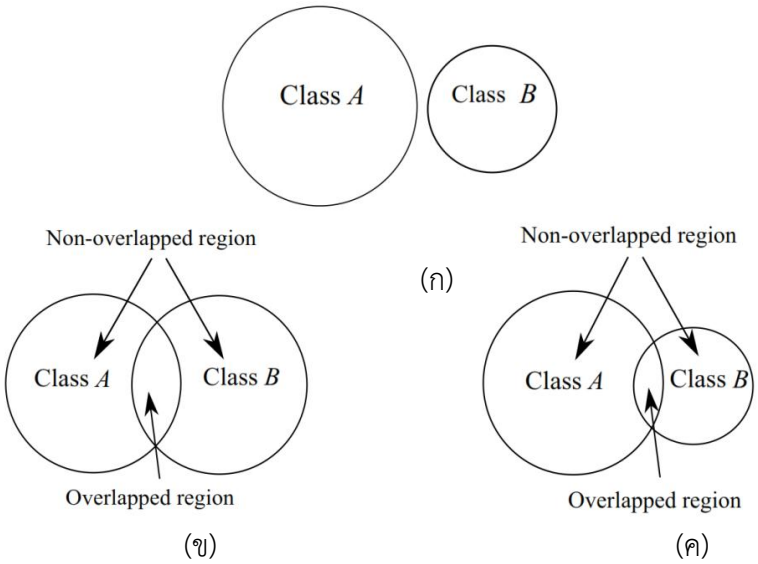
ในงานวิจัยนี้ได้ทำการศึกษาปัญหาการจำแนกข้อมูลที่ไม่สมดุลซึ่งแตกต่างจากปัญหาการจำแนกข้อมูลที่ไม่สมดุลทั่วไปที่คำนึงถึงความแตกต่างของจำนวนข้อมูลที่มีอยู่ในแต่ละคลาสโดยใช้ข้อมูลที่มีอยู่ทั้งหมด แต่ปัญหาที่คณะผู้วิจัยนำเสนอนี้ เป็นปัญหาการจำแนกข้อมูลที่ไม่สมดุลในเฉพาะที่ปรากฏอยู่ในพื้นที่ซ้อนทับ ซึ่งข้อมูลที่อยู่ในบริเวณนี้เป็นข้อมูลที่ยากในการจำแนกว่าอยู่ในกลุ่มไหนและอัตราความไม่สมดุลกันของสมาชิกในแต่ละกลุ่มจะขึ้นอยู่กับข้อมูลที่อยู่ในพื้นที่ทับซ้อน ปัญหาที่กล่าวถึงนี้สามารถอธิบายได้ดังนี้ ในชุดข้อมูลที่มี 2 คลาส แต่ละคลาสมีบางข้อมูลที่อยู่ในบริเวณที่ซ้อนทับกัน จะเห็นว่าข้อมูลทั้งสองคลาสนี้จะถูกแบ่งออกเป็น 3 พื้นที่ ได้แก่ 1) พื้นที่ซ้อนทับ (Overlapped region) กันระหว่างสองคลาส 2) พื้นที่ไม่ซ้อนทับ (Non-overlapped region) กันหรือพื้นที่ที่มีเฉพาะสมาชิกของคลาสดคลาสหนึ่งเท่านั้น ในกรณีนี้จะมีพื้นที่ไม่ซ้อนทับจำนวนเท่ากับจำนวนคลาส และ 3) พื้นที่ที่อยู่บริเวณขอบระหว่างะพื้นที่ซ้อนทับและไม่ซ้อนทับ ซึ่งจะเรียกว่า พื้นที่ขอบ (Borderline region)

โดยทั่วไปแล้วข้อมูลที่อยู่ในพื้นที่ซ้อนทับจะมีจำนวนน้อยกว่าข้อมูลที่อยู่ในพื้นที่ไม่ซ้อนทับ ดังนั้นจึงต้องแยกข้อมูลทั้งสองพื้นที่นี้ออกจากกัน จากรูปที่ 3-5 แสดงตัวอย่างของปัญหาข้อมูลไม่สมดุลทั่วไป กับข้อมูลไม่สมดุลที่ทำการศึกษาในงานวิจัยนี้ จากรูปชุดข้อมูลประกอบไปด้วย 2 คลาส ได้แก่ คลาสส่วนมาก A และคลาสส่วนน้อย B, วงกลมแทนข้อมูลของแต่ละคลาส ซึ่งปัญหาข้อมูลไม่สมดุลทั่วไปแสดงให้เห็นในรูปที่ 3-5 (ก) ขนาดของคลาส A ใหญ่กว่าคลาส B นั่นคือจำนวนสมาชิกของคลาส A มีจำนวนมากกว่า คลาส B รูปที่ 3-5 (ข) และ (ค) แสดงปัญหาข้อมูลไม่สมดุลเมื่อพิจารณาจำนวนของข้อมูลที่อยู่ในพื้นที่ซ้อนทับ และพื้นที่ไม่ซ้อนทับ จากรูปที่ 3-5 (ข) จะเห็นว่าขนาดของคลาส A และคลาส B เท่ากัน แต่ในรูปที่ 3-5 (ค) ขนาดของทั้งสองคลาสแตกต่างกัน ซึ่งความแตกต่างกันของขนาดหรือจำนวนสมาชิกในแต่ละคลาสนี้มีผลกระทบต่อความถูกต้องในการจำแนกข้อมูล

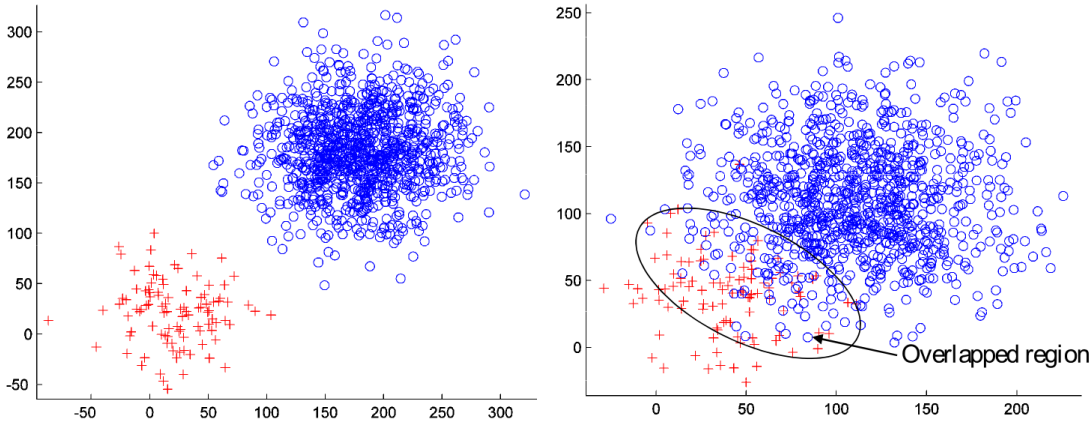
จากรูปที่ 3-5 ประกอบไปด้วยข้อมูล 2 คลาส ได้แก่ คลาส A และคลาส B และข้อมูลบางส่วนของทั้งสองคลาสมีการซ้อนทับกัน กำหนดให้ n_A เป็นจำนวนสมาชิกของคลาส A, n_B เป็นจำนวนสมาชิกของคลาส B ตามลำดับ ซึ่งในการศึกษานี้จะไม่สนใจว่า $n_A \leq n_B$ หรือ $n_A \geq n_B$ แต่สิ่งที่คณะผู้วิจัยให้ความสนใจคือจำนวนสมาชิกของคลาส A และคลาส B ที่อยู่ในพื้นที่ซ้อนทับ แทนด้วย o_A และ o_B ตามลำดับ ซึ่งเมื่อพิจารณาแล้วจะพบว่า$o_A + o_B \ll n_A$ แล$o_A + o_B \ll n_B$ ส่วนพื้นที่ที่อยู่นอกพื้นที่ซ้อนทับเป็นพื้นที่ที่มีสมาชิกเพียงคลาสเดียวเท่านั้นหรือเรียกว่าพื้นที่ที่ไม่ซ้อนทับหรือเรียกอีกอย่างว่า '$clean\ region$' ในการศึกษาเกี่ยวกับการจำแนกข้อมูลไม่สมดุลในพื้นที่ซ้อนทับมีประเด็นที่ต้องทำการศึกษาและหาวิธีในการแก้อยู่ 2 ประเด็น คือ

1. จะมีวิธีการที่ใช้แยกพื้นที่พื้นที่ที่ไม่ซ้อนทับออกจากพื้นที่ซ้อนทับได้อย่างไร
2. จะมีธีการอย่างไรให้การจำแนกข้อมูลที่อยู่ในพื้นที่ซ้อนทับมีความถูกต้องแม่นยำ

จาก $o_A + o_B \ll n_A$ และ $o_A + o_B \ll n_B$ ที่กล่าวมาข้างต้นทำให้เห็นว่าปัญหาข้อมูลไม่สมดุลถูกแยกออกเป็นปัญหาในพื้นที่ซ้อนทับ และปัญหาในพื้นที่ไม่ซ้อนทับในรูปที่ 3-6 แสดงตัวอย่างของปัญหาข้อมูลไม่สมดุลที่ได้ทำการศึกษานี้ ซึ่งปัญหานี้มีความแตกต่างจากปัญหาข้อมูลไม่สมดุลแบบทั่วไปที่ถูกนำเสนอในงานวิจัยของ Napierala และคณะ (2012), Alejo และคณะ (2013), Lin และคณะ (2013)



รูปที่ 3-5 ตัวอย่างข้อมูลไม่สมดุล (ก) คลาส A มีจำนวนสมาชิกมากกว่าคลาส B (ข) คลาส A และคลาส B มีจำนวนสมาชิกเท่ากัน ซึ่งความไม่สมดุลเกิดจากข้อมูลที่อยู่ในพื้นที่ซ้อนทับและไม่ซ้อนทับ (ค) คลาส A และคลาส B มีจำนวนสมาชิกแตกต่างกัน ซึ่งความไม่สมดุลเกิดจากข้อมูลที่อยู่ในพื้นที่ซ้อนทับและไม่ซ้อนทับ



(ก) (ข)

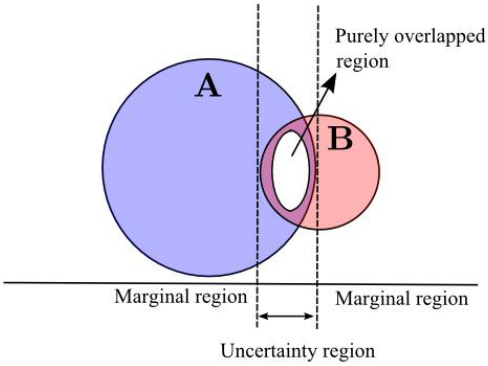
รูปที่ 3-6 ตัวอย่างชุดข้อมูลที่มีความไม่สมดุลสูง (ก) ข้อมูลที่ไม่มีการซ้อนทับ (ข) ข้อมูลที่มีการซ้อนทับ

3.3.2 วิธีการที่นำเสนอ: วิธีการแบบผสมสำหรับข้อมูลไม่สมดุล (Soft-Hybrid Algorithm for Imbalanced Data Set)

วิธีการที่นำเสนอจะทำการแบ่งพื้นที่ของชุดข้อมูลที่ไม่สมดุลออกเป็น 3 พื้นที่ ได้แก่ 1. พื้นที่ไม่มีการซ้อนทับ (non-overlapped region) หรือ พื้นที่ขอบ (marginal region) 2. พื้นที่ที่ยังไม่แน่นอนว่าจะอยู่ในพื้นที่ซ้อนทับ (uncertainty overlapped region) และ 3. พื้นที่ซ้อนทับ (purely overlapped region) แสดงดังรูปที่ 3-7 จากปัญหาการจำแนกข้อมูลไม่สมดุลพบว่ามีความเอนเอียงไปยังคลาสส่วนมาก ดังนั้นในงานวิจัยที่ผ่านมามีการนำเสนอวิธีในการลดความเอนเอียงในการจำแนกข้อมูลมายังคลาสส่วนมากให้น้อยลง ซึ่งเป็นทำได้ยากและวิธีการที่ใช้มีความซับซ้อนส่งผลให้เวลาในการประมวลผลสูงตามไปด้วย ดังนั้นวิธีการที่นำเสนอในงานวิจัยนี้นอกจากลดความเอนเอียงในการจำแนกข้อมูลไปยังคลาสส่วนมาก และทำให้ความถูกต้องในการจำแนกข้อมูลดีขึ้นแล้ว วิธีที่นำเสนอจะต้องเป็นวิธีการที่มีความซับซ้อนต่ำและใช้เวลาในการประมวลผลน้อยด้วย ซึ่งวิธีการที่นำเสนอนี้เรียกว่าวิธีการแบบผสม หรือ Soft-Hybrid

วิธีการที่นำเสนอแบ่งออกเป็น 2 ขั้นตอน ได้แก่ 1. กำหนดพื้นที่ (Boundary region determination) 2. จำแนกประเภทข้อมูล (Classification) ในขั้นตอนกำหนดพื้นที่ ขั้นตอนนี้จำทำการแยกข้อมูลให้อยู่ในพื้นที่ที่เหมาะสมซึ่งประกอบไปด้วย พื้นที่ไม่ซ้อนทับ (Non-overlapped region) พื้นที่บริเวณขอบระหว่างพื้นที่ซ้อนทับและไม่ซ้อนทับ (Borderline region) และพื้นที่ซ้อนทับ (Overlapped region) ส่วนขั้นตอนที่ 2 การจำแนกประเภทข้อมูล เป็นการนำเอาขั้นตอนวิธีที่เหมาะสมสำหรับแต่พื้นที่มาใช้ในการจำแนกประเภทข้อมูล

ลักษณะของข้อมูลที่ใช้ในงานวิจัยนี้เป็นชุดข้อมูลไม่สมดุลที่มีการกระจายตัวแบบเกาส์เซียน และแต่ละคลาสแยกเป็นอิสระจากกัน ข้อมูลทั้งสองคลาสมีการซ้อนทับกันบางส่วน โดยกำหนดให้ $\mathbf{A} = \{(a_i, t_A)|1 \leq i \leq n_A, \mathbf{B} = \{(a_i, t_B)|1 \leq i \leq n_B\}$ และ $\mathbf{X} = \mathbf{A} \cup \mathbf{B}$



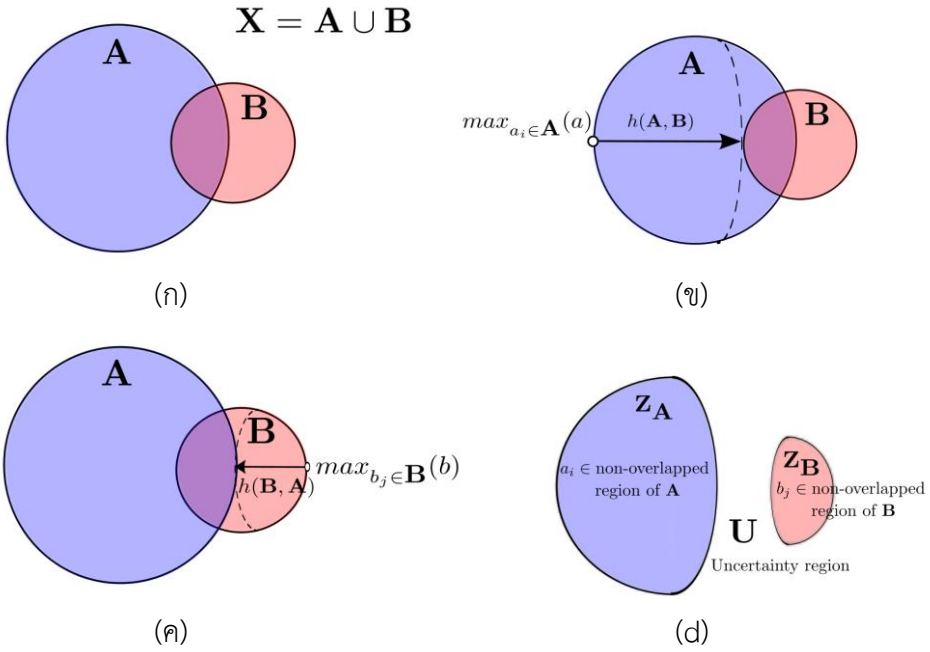
รูปที่ 3-7 พื้นที่ของชุดข้อมูลไม่สมดุล

3.3.2.1 กำหนดพื้นที่ (Boundary region determination)

1. สกัดพื้นที่ที่ไม่ซ้อนทับ (Extracting non-overlapped region)

ในขั้นตอนนี้ได้ประยุกต์ใช้ระยะทางฮาวดอร์ฟ มาทำการแยกข้อมูลที่ไม่ซ้อนทับออกจากข้อมูลอื่นแสดงดังรูปที่ 3-8

จากรูปที่ 3-8 แสดงขั้นตอนการแยกข้อมูลออกจากกันไปยังพื้นที่ที่เหมาะสม รูปที่ 3-8 (ก) แสดงพื้นที่ที่ซ้อนทับกับระหว่างคลาส A และ คลาส B รูปที่ 3-8 (ข) แสดงค่ามากที่สุดของระยะฮาวดอร์ฟของคลาส A มายัง คลาส B แทนด้ว$max_{a \in \mathbf{A}}(a)$ย ส่วนรูปที่ 3-8 (ค) แสดงค่ามากที่สุดของระยะฮาวดอร์ฟของคลาส B มายัง คลาส A แทนด้วย $max_{b \in \mathbf{B}}(b)$ บริเวณที่มีระยะฮาวดอร์ฟไปยังระยะฮาวดอร์ฟ น้อยกว่าระยะฮาวดอร์ฟมากที่สุดของคลาสจะจัดอยู่ในพื้นที่ไม่ทับซ้อน และที่เหลือจัดอยู่ในพื้นที่ไม่แน่นอน ดังรูปที่ 3-8 (d) รายละเอียดขั้นตอนนี้แสดงใน ALGORITHM 1

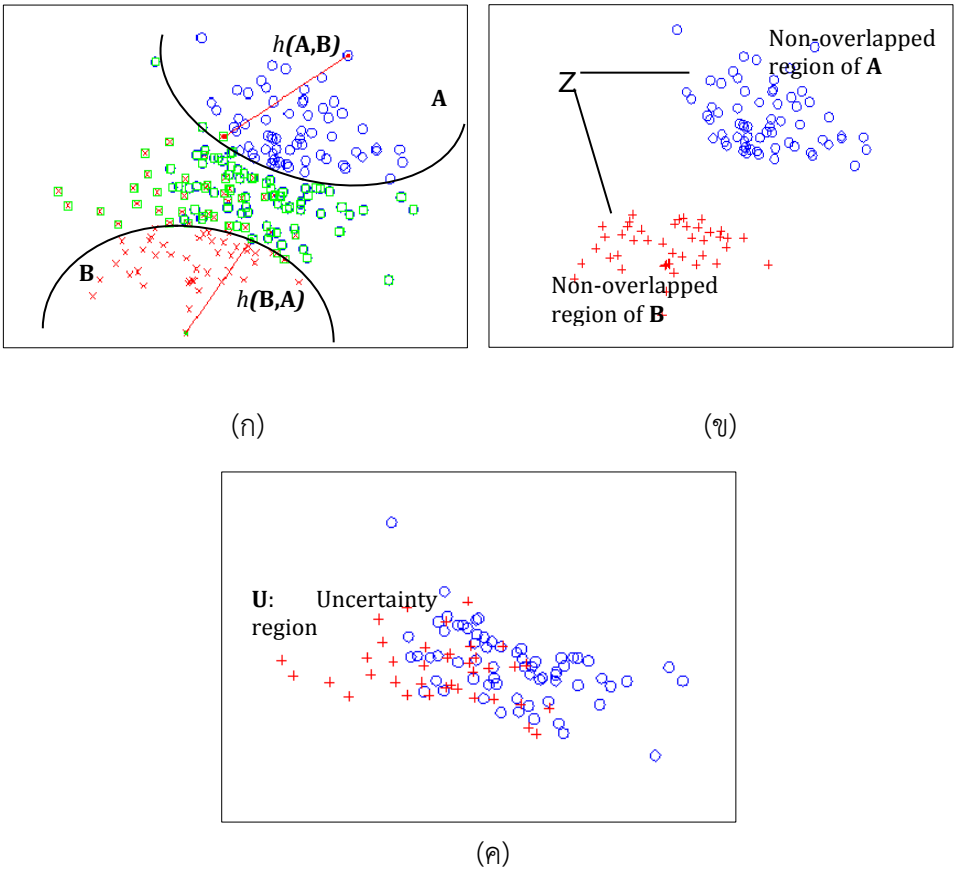


รูปที่ 3-8 แยกข้อมูลที่ไม่ซ้อนทับออกจากข้อมูลอื่น (ก) ขอบเขตของคลาส A และ B (ข) ระยะทางฮาวดอร์ฟ จากคลาส A ไป คลาส B (ค) ระยะทางฮาวดอร์ฟ จากคลาส B ไป class A (d) พื้นที่ไม่ซ้อนทับ และพื้นที่ไม่แน่นอน

ALGORITHM 1. Extracting obviously non-overlapped region.

1. **INPUT:** Set $\mathbf{X} = \mathbf{A} \cup \mathbf{B}$.
2. **OUTPUT:**
    1. **Z**: set of non-overlapping data.
    2. **U**: set of uncertainty data.
3. **Begin**
4. Let $\mathbf{Z_A} = \emptyset$ and $\mathbf{Z_B} = \emptyset$.
5. Compute $h(\mathbf{A}, \mathbf{B})$ and $h(\mathbf{B}, \mathbf{A})$.
6. **For** $a_i \in \mathbf{A}; a \leq i \leq n$ **do**
7. Compute $d_i = \| a_i, max_{a \in \mathbf{A}}(a) \|$.
8. **If** $d_i < h(\mathbf{A}, \mathbf{B})$ **then**
9. $\mathbf{Z_A} = \mathbf{Z_A} \cup \{a_i\}$.
10. **EndIf**
11. **EndFor**
12. **For** $b_j \in \mathbf{B}; b \leq j \leq m$ **do**
13. Compute $d_j = \| b_j, max_{b \in \mathbf{B}}(b) \|$.
14. **If** $d_j < h(\mathbf{B}, \mathbf{A})$ **then**
15. $\mathbf{Z_B} = \mathbf{Z_B} \cup \{b_j\}$.
16. **EndIf**
17. **EndFor**
18. $\mathbf{Z} = \mathbf{Z_A} \cup \mathbf{Z_B}$.
19. $\mathbf{U} = \mathbf{X} - \mathbf{Z}$.
20. **End**

รูปที่ 3-9 แสดงตัวอย่างของการแยกพื้นที่ของข้อมูลคลาส A และ คลาส B ซึ่งแทนด้วยสัญลักษณ์ “+” และ “○” ตามลำดับ



รูปที่ 3-9 ตัวอย่างของการแยกพื้นที่ของข้อมูลคลาส A และ คลาส B (ก) Modified Hausdorff distance (ข) ข้อมูลในพื้นที่ไม่ซ้อนทับ (ค) ข้อมูลในพื้นที่ซ้อนทับ

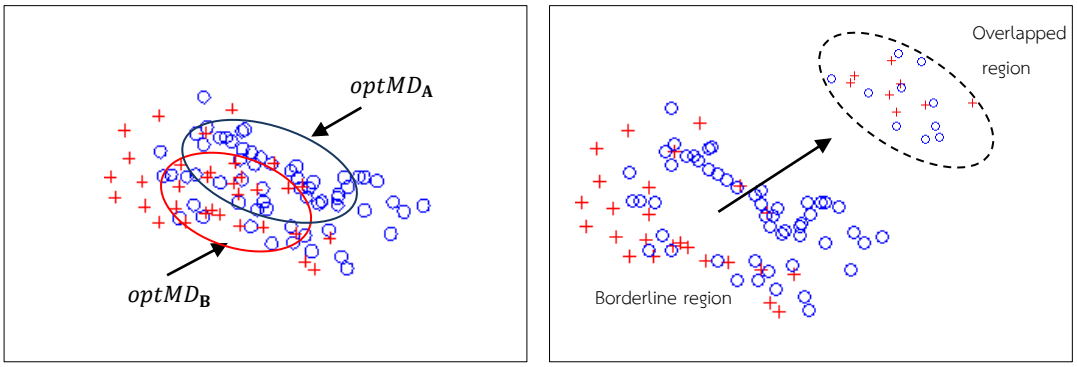
2. ขั้นตอนวิธีเรียนรู้ขอบเขตที่เหมาะสม

ข้อมูลที่อยู่ในพื้นที่ uncertainty จากขั้นตอนที่แล้วจะถูกนำมาหาพื้นที่ที่มีการซ้อนทับกันของข้อมูลและข้อมูลที่อยู่บริเวณขอบ ในงานวิจัยที่ผ่านมาของ (He และ Ghodsi, 2010; Fan และ He,2010; Hwang และและ, 2011) ได้นำเอาวิธีการในการเปลี่ยนรูปข้อมูลให้ไปอยู่ในมิติข้อมูลที่สูงขึ้น อย่างไรก็ตามวิธีการดังกล่าวนี้เมื่อจำนวนข้อมูลมีมากขึ้นจะทำให้เวลาที่ใช้ในการประมวลผลสูง วิธีการที่นำเสนอในหัวข้อนี้แบ่งการทำงานออกเป็น 2 ขั้นตอน ได้แก่ ขั้นตอนการสร้างโมเดลในการหาขอบเขตของพื้นที่ที่เหมาะสม และขั้นตอนการกำหนดพื้นที่ให้กับข้อมูลของแต่ละคลาส ดังรายละเอียดต่อไปนี้

1. ขั้นตอนการสร้างโมเดลในการหาขอบเขตของพื้นที่ที่เหมาะสม ในขั้นตอนนี้ได้ใช้ขั้นตอนวิธีโครงข่ายฟังก์ชั่นรัศมีฐาน (RBFN) โดยใช้เกาส์เซียนเคอเนลฟังก์ชั่น ในการเรียนรู้ข้อมูลเพื่อหาขอบเขตของพื้นที่ ข้อมูลนำเข้าที่ใช้การเรียนรู้ได้แก่ IRs, $maxFs$, KLs และระยะทางระหว่างศูนย์กลางของข้อมูลแต่ละกลุ่ม ผลลัพธ์ที่ได้จากโมเดลนี้คือ ระยะทางมหาลาโนบิสที่เหมาะสมของพื้นที่ซ้อนทับกันของคลาส A และคลาส B หรือ $optMD_A$ และ $optMD_B$ ตามลำดับแสดงดังรูปที่ 3-10 (ก)
2. ขั้นตอนการกำหนดพื้นที่ให้กับข้อมูลของแต่ละคลาส หลังจากที่ได้ระยะทางมหาลาโนบิสที่เหมาะสมของพื้นที่ซ้อนทับกันของแต่ละคลาสแล้ว ขั้นตอนต่อไปจะทำการทำหนดพื้นที่ให้กับข้อมูลของแต่ละคลาสโดยการหาค่าระยะทางมหาลาโนบิสระหว่างข้อมูลแต่ละตัวกับคลาสทั้งสอง กำหนดได้ดังนี้ $MD(a_i, \mathbf{A})$ และ $MD(a_i, \mathbf{B})$ เป็นระยะทางระหว่างข้อมูลใดๆ ของคลาส A กับคลา B ส และคลาส B ลำดับ, $MD(b_j, \mathbf{A})$ และ $MD(b_j, \mathbf{B})$เป็นระยะทางระหว่างข้อมูลใดๆ ของคลาส B กัลคลาส A และคลาส B ลำดับ จากนั้นทำการกำหนดพื้นที่ให้กับข้อมูลใดโดยใช้กฎ รายละเอียดแสดงใน ALGORITHM 2

ALGORITHM 2. Assigning optimal boundary algorithm.

1. **INPUT:** Set of uncertainty data $\mathbf{U} = \mathbf{X} - \mathbf{Z}$.
2. **OUTPUT:**
    1. **O**: set of overlapping data.
    2. **Y**: set of borderline data.
3. **Begin**
4. Let $\mathbf{Y} = \emptyset$ and $\mathbf{O} = \emptyset$.
5. **For** $a_i \in \mathbf{A}; a \leq i \leq n$ **do**
6. Compute $MD(a_i, \mathbf{A})$ and $MD(a_i, \mathbf{B})$.
7. **If** $MD(a_i, \mathbf{A}) > optMD_A$ or $MD(a_i, \mathbf{B}) > optMD_B$ **then**
8. $\mathbf{Y} = \mathbf{Y} \cup \{a_i\}$.
9. **EndIf**
10. **EndFor**
11. **For** $b_j \in \mathbf{A}; b \leq j \leq m$ **do**
12. Compute $MD(b_j, \mathbf{A})$ and $MD(b_j, \mathbf{B})$.
13. **If** $MD(b_j, \mathbf{A}) > optMD_A$ or $MD(b_j, \mathbf{B}) > optMD_B$ **then**
14. $\mathbf{Y} = \mathbf{Y} \cup \{b_j\}$.
15. **EndIf**
16. **EndFor**
17. $\mathbf{O} = \mathbf{U} - \mathbf{Y}$.
18. **End**

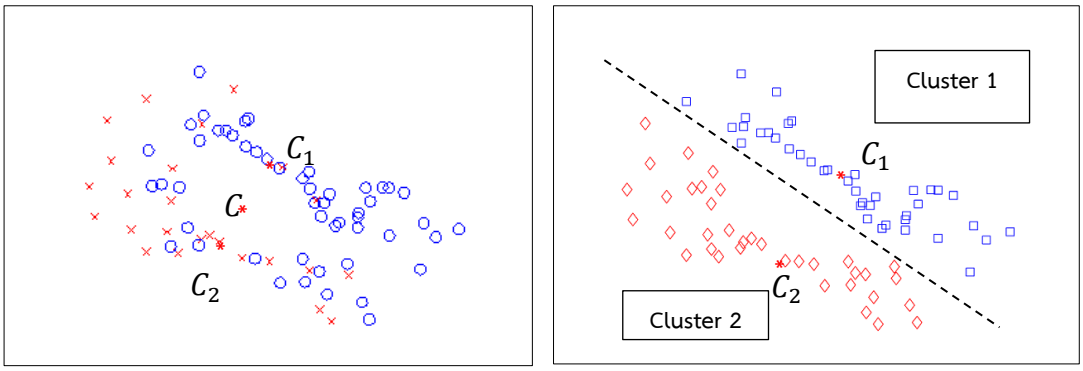


(ก) (ข)

รูปที่ 3-10 ตัวอย่างของการขอบเขตที่เหมาะสม (ก) ระยะทางมหาลาโนบิส
(ข) พื้นที่ซ้อนทับและพื้นที่บริเวณขอบ

3. จัดกลุ่มข้อมูลพื้นที่บริเวณขอบ

ในขั้นตอนนี้ ข้อมูลที่เป็น outlier ที่อยู่ในพื้นที่บริเวณขอบในรูป 3-10 (ข) จะถูกกำจัดออกไป จากนั้นข้อมูลที่เหลือจะถูกนำมาทำการจัดกลุ่มข้อมูลด้วนขั้นตอนวิธี เคมีนส์ (K-Means) ด้วยระยะทางมหาลาโนบิส โดยทำการจัดกลุ่มเป็น 2 กลุ่มตามจำนวนคลาส รายละเอียดแสดงดัง ALGORITHM 3 ผลที่ได้จากขั้นตอนนี้คือข้อมูลกลุ่มที่ 1 (C_1) และกลุ่มที่ 2 (C_2) ดังรูปที่ 6(ข)



(ก) (ข)

รูปที่ 3-10 ตัวอย่างการจัดกลุ่มพื้นที่บริเวณของ (ก) กำหนดจุดศูนย์กลางเริ่มต้น
(ข) จัดกลุ่มข้อมูลออกเน 2 กลุ่ม

ALGORITHM 3. Clustering boundary of borderline region.

1. **INPUT:** Set of borderline data **Y**.
2. **OUTPUT:** left and right clusters.
3. **Begin**
4. Compute mean C from data in **Y**.
5. Let $C_1 = \arg min_{y_i \in \mathbf{Y}} \left(MD(y_i, C)\right)$.
6. Let $C_2 = C + (C - C_1)$.
7. **Repeat**
8. **For** each $y_i \in$ **Y do**
9. **If** $MD(y_i, C_1) < MD(y_i, C_2)$ then
10. Assign y_i to the left cluster.
11. **else**
12. Assign y_i to the right cluster.
13. **EndIf**
14. **EndFor**
15. Let C_1 be the mean of left cluster.
16. Let C_2 be the mean of right cluster.
17. **Until** C_1 and C_2 do not change their values.
18. **End**

3.3.2.2 จำแนกประเภทข้อมูล (Classification)

เมื่อข้อมูลถูกแยกออกเป็น 3 พื้นที่แล้ว จะถูกนำมาจำแนกประเภทข้อมูล ด้วยวิธีการที่เหมาะสมกับแต่ละพื้นที่ โดยมีรายละเอียดดังต่อไปนี้

1. การจำแนกประเภทข้อมูลในพื้นที่ไม่ซ้อนทับ

ข้อมูลที่อยู่ในพื้นที่ไม่ซ้อนทับจะถูกนำมาจำแนกประเภทข้อมูลด้วยขั้นตอนวิธีฟังก์ชั่นรัศมีฐาน (RBF) โดยใช้เกาส์เซียนเคอเนลฟังก์ชั่น

2. การจำแนกประเภทข้อมูลในพื้นที่ขอบ

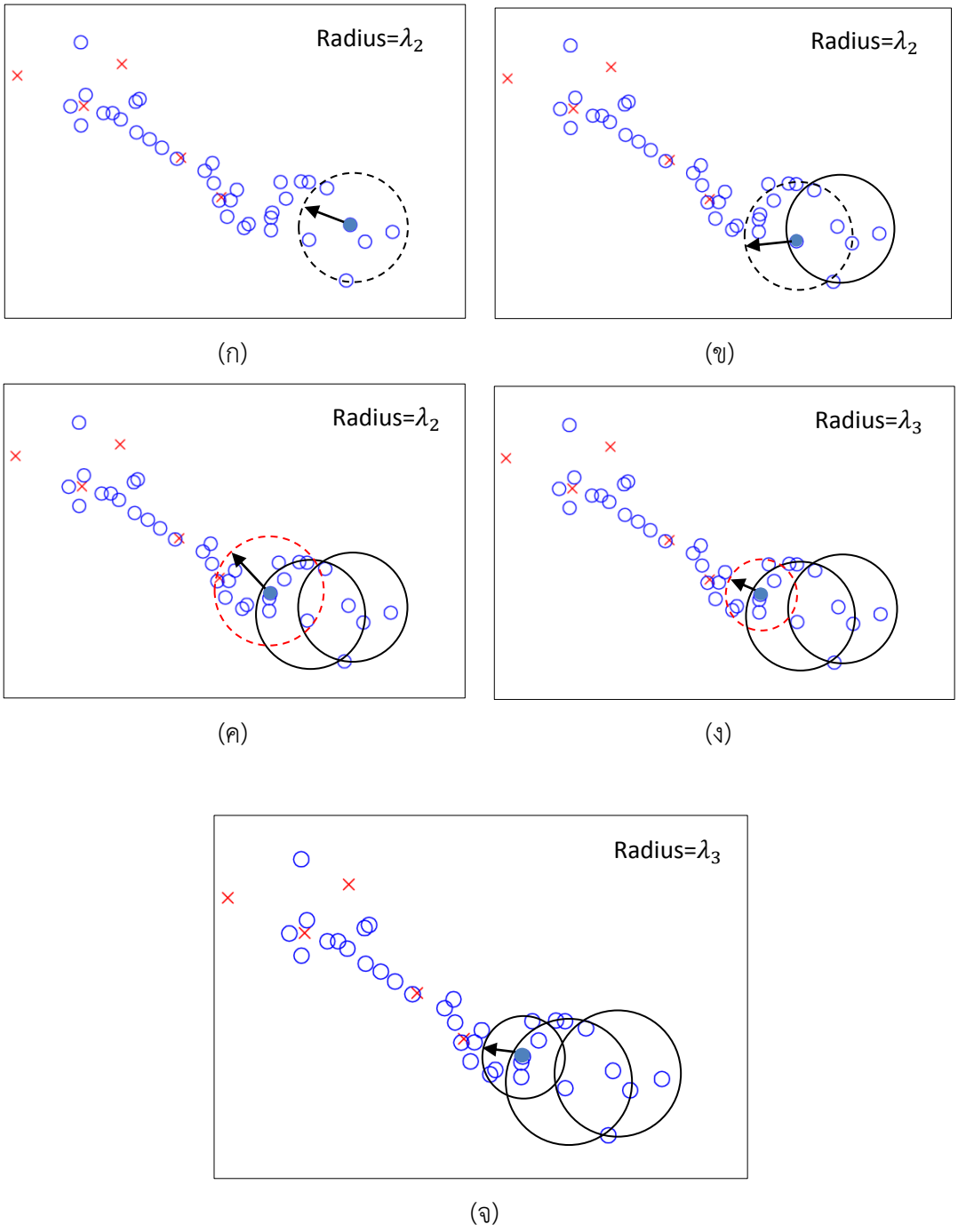
ข้อมูลในพื้นที่บริเวณขอบจะประยุกต์ใช้ Density-based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al., 1996) โดยใช้ค่าไอเก้นลำดับที่ 2 เป็นต้นไปของข้อมูลแต่ละคลาสเป็นรัศมี $\lambda_2, \ldots, \lambda_l, \ldots, \lambda_m$ จำนวนข้อมูลต่ำสุด (Minimum point (minPts)) กำหนดเป็น 3 ซึ่งวิธีที่นำเสนอรัศมี หรือค่า ไอเก้นที่ใช้จะเปลี่ยนไปตามรายละเอียดใน ALGORITHM 4 วิธีการนี้เรียนว่า dynamic DBSCAN

ALGORITHM 4. Classification of borderline region.

1. **INPUT:** left and right clusters of borderline region.
2. **OUTPUT:** optimal sub-clusters.
3. **Begin**
4. **Repeat**
5. Arbitrarily select a point x_i.
6. Retrieve all points in the same class close to x_i with respect to the radius or variance of the 2nd EigenValue.
7. **If** x_i is a *core point* **then**
8. A sun-class is formed.
9. **else**
10. The radius of such sub-class is reduced to the next EigenValue.
11. Retrieve all points in the same class close to x_i with respect to the new radius.
12. **EndIf**
13. **If** x_i is a *border point* **then**
14. Visit the next learning data point.
15. **EndIf**
16. **Until** All data points are processed.
17. **End**

3. การจำแนกประเภทข้อมูลในพื้นที่ซ้อนทับ

ข้อมูลที่อยู่ในพื้นที่ซ้อนทับจะถูกนำมาจำแนกประเภทข้อมูลด้วยขั้นตอนวิธีฟังก์ชั่นรัศมีฐาน (RBF) โดยใช้โพลิโนเมียลเคอเนลฟังก์ชั่นดีกรี 2 โดยข้อมูลที่อยู่ในพื้นที่นี้เหลือจำนวนน้อยลงคือ ชุดข้อมูลที่สังเคราะห์โดยเฉลี่ยเหลือ 37.55 เปอร์เซ็นต์ ชุดข้อมูลมาตรฐานเหลือ 63.41 เปอร์เซ็นต์ ดังนั้นจำนวนที่ลดลงนี้ทำให้เวลาในการการประมวลผลข้อมูลน้อยลงตามไปด้วย



รูปที่ 3-10 การทำงานของ dynamic DBSCAN (ก) เลือกข้อมูลเริ่มต้น (ข) สร้างกลุ่มย่อย (ค) ข้อมูลภายในรัศมีมีข้อมูลคลาสตรงข้าม (ง) รัศมีเปลี่ยนเป็นค่าไอเก้นค่าถัดไป (จ) สร้างกลุ่มย่อยใหม่

บทที่ 4 ผลการทดลอง

4.1. การเลือกคุณลักษณะสำหรับชุดข้อมูลไม่สมดุล

4.1.1 ชุดข้อมูล

ข้อมูลที่ใช้ในการทดลองเป็นข้อมูลจาก UCI Repository Machine Learning จำนวน 4 ชุด ได้แก่ Pakinsons, Sonar, Liver, Pima ซึ่งทั้งหมดเป็นข้อมูลแบบไม่สมดุล รายละเอียดแสดงดังตารางที่ 4-1 ประกอบไปด้วยจำนวนแอททริบิวต์ (Attributes) จำนวนข้อมูล (Instances) จำนวนข้อมูลคลาสส่วนมากกับส่วนน้อย (Maj./Min.) และอัตราส่วนความไม่สมดุล (Imbalanced ratio: IR)

ตารางที่ 4-1 รายละเอียดของชุดข้อมูลก่อนการทดลอง

Datasets	#att	#att	#maj/#min	IR
Pakinsons	22	195	147/48	3.06
Sonar	60	208	111/97	1.14
Liver	6	345	200/145	1.38
Pima	8	768	500/268	1.87

4.1.2 การทดลองและผลการทดลอง

นำข้อมูลแต่ละชุดมาลดจำนวนข้อมูลของคลาสส่วนมากให้มีจำนวนเท่ากับหรือใกล้เคียงคลาสส่วนน้อย จากนั้นนำข้อมูลใหม่ที่ได้มาหาจำนวนที่เหมาะสมจากจำนวนคุณลักษณะที่มีค่าไอเกน มากกว่า 0.01 และทำการเลือกคุณลักษณะด้วยขั้นตอนวิธีทางพันธุกรรมร่วมกับโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ รายละเอียดดังตารางที่ 4-2 และ 4-3 ซึ่งจำนวนคุณลักษณะของข้อมูลทั้ง 4 ชุด มีจำนวน 8 28 3 และ 7 ตามลำดับ

ตารางที่ 4-2 รายละเอียดของชุดข้อมูลเมื่อใช้การเลือก คุณลักษณะข้อมูลไม่สมดุลด้วยวิธีการแบบผสม

Datasets	#att	#att	#maj/#min	IR
Pakinsons	8	96	49/48	1.02
Sonar	28	194	97/97	1
Liver	3	490	245/245	1
Pima	7	536	268/268	1

ตารางที่ 4-3 คุณลักษณะที่ถูกเลือก

Datasets	Attributes
Pakinsons	1 4 5 11 17 18 20 21
Sonar	2 6 7 8 10 11 14 16 20 21 23 24 28 29 31 32 33 38 40 42 44 45 46 47 50 53 59
Liver	1 2 3
Pima	1 2 3 5 6 7 8

จากนั้นนำข้อมูลไปทำการทดสอบแบบ 10-fold cross validation กับตัวจำแนกข้อมูลทั้ง 3 แบบ ทำการเปรียบเทียบระหว่างวิธีการแบบผสมกับข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง โดยกำหนดค่าพารามิเตอร์ดังนี้ 1) โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับฟังก์ชันกระตุ้นใช้ฟังก์ชันซิกมอยด์ ทั้งชั้นซ่อนและชั้นเอาท์พุท อัตราการเรียนรู้ 0.1 โมเมนตัม 0.2 2) โครงข่ายฟังก์ชันรัศมีฐาน ฟังก์ชันกระตุ้นในชั้นซ่อนใช้ฟังก์ชันรัศมีฐาน ส่วนชั้นเอาท์พุทใช้ฟังก์ชันเชิงเส้น 3) ซัพพอร์ตเวกเตอร์แมชชีน ค่า c เท่ากับ 11 ใช้ RBF kernel ผลการทดลองแสดงดังตารางที่ 4-4

ตารางที่ 4-4 ค่าความถูกต้อง (Accuracy) และค่าเฉลี่ย เรขาคณิต (G-Mean) (%)

Datasets	BPNN		RBF		SVM	
	G	Acc	G	Acc	G	Acc
Pakinsons						
- Original (22)	91.03	91.79	70.17	84.10	65.71	85.12
- Proposed (8)	91.75	91.75	85.56	85.56	75.84	76.28
- Diff.	0.72	-0.04	15.39	1.46	10.13	-8.84
Sonar						
- Original (60)	79.70	80.28	72.27	72.11	80.48	80.76
- Proposed (28)	80.40	80.41	75.66	73.71	68.87	69.07
- Diff.	0.7	0.13	3.39	1.6	-11.61	-11.69

ตารางที่ 4-4 (ต่อ)

Datasets	BPNN		RBF		SVM	
	G	Acc	G	Acc	G	Acc
Liver						
- Original (6)	67.22	69.56	61.65	64.34	60.46	67.53
- Proposed (3)	71.83	72.06	66.35	66.55	65.68	66.20
- Diff.	4.61	2.5	4.7	2.21	5.22	-1.33
Pima						
- Original (8)	70.56	72.52	68.52	75.39	65.61	76.43
- Proposed (7)	86.23	86.38	86.14	86.19	83.65	83.95
- Diff.	15.67	13.86	17.62	10.8	18.04	7.52

* Original: ข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง

Proposed: วิธีการแบบผสม

Diff.: ความแตกต่างระหว่างวิธีการแบบผสมกับข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง

จากตารางที่ 4-4 ชุดข้อมูล Pakinsons เมื่อนำมาจำแนกข้อมูลด้วย BPNN พบว่าค่า G-mean เมื่อใช้วิธีการแบบผสมให้ค่ามากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 0.72% ค่า Accuracy น้อยกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 0.04% RBF พบว่าค่า G-mean และค่า Accuracy ของวิธีการแบบผสมให้ค่ามากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 15.39% และ 1.46% SVM พบว่าค่า G-mean ของวิธีการแบบผสมให้ค่ามากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 10.13% ค่า Accuracy น้อยกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 8.84%

ชุดข้อมูล Sonar เมื่อนำมาจำแนกข้อมูลด้วย BPNN พบว่าค่า G-mean และค่า Accuracy เมื่อใช้วิธีการแบบผสมให้ค่ามากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 0.7% และ 0.13% RBF พบว่าค่า G-mean และค่า Accuracy เมื่อใช้วิธีการแบบผสมให้ค่ามากกว่าชุดข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 3.39% และ 1.6% SVM เมื่อใช้วิธีการแบบผสมให้ค่า G-mean และค่า Accuracy น้อยกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 11.61% และ 11.69%

ชุดข้อมูล Liver เมื่อนำมาจำแนกข้อมูลด้วย BPNN พบว่าค่า G-mean และค่า Accuracy เมื่อใช้วิธีการแบบผสมให้ค่ามากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 4.61% และ 2.5% RBF พบว่าค่า G-mean และค่า Accuracy เมื่อใช้วิธีการแบบผสมให้ค่ามากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 4.7% และ 2.21% SVM พบว่าค่า G-mean เมื่อใช้วิธีการแบบผสมให้ค่ามากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 5.22% ค่า Accuracy น้อยกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 1.33%

ชุดข้อมูล Pima เมื่อนำมาจำแนกข้อมูลด้วย BPNN พบว่าค่า G-mean และค่า Accuracy เมื่อใช้วิธีการแบบผสมให้ค่ามากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 15.67% และ 13.86% RBF พบว่าค่า G-mean และค่า Accuracy เมื่อใช้วิธีการแบบผสมให้ค่ามากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 17.62% และ 10.8% SVM พบว่าค่า G-mean และค่า Accuracy เมื่อใช้วิธีการแบบผสมให้ค่ามากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 18.04% และ 7.52%

ตารางที่ 4-5 ค่าความถูกต้องของการจำแนกข้อมูลคลาสส่วนน้อย (%)

Datasets		BPNN	RBF	SVM
Pakinsons	Original	89.58	52.08	43.75
	Proposed	93.75	87.50	68.75
	Diff.	4.17	35.42	25
Sonar	Original	74.22	78.35	77.31
	Proposed	79.38	71.13	74.22
	Diff.	5.16	-7.22	-3.09
Liver	Original	57.93	51.72	42.75
	Proposed	66.20	61.13	57.93
	Diff.	8.27	9.41	15.18
Pima	Original	58.95	54.10	46.64
	Proposed	83.20	83.95	61.30
	Diff.	24.25	29.85	14.66

* Original: ข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง

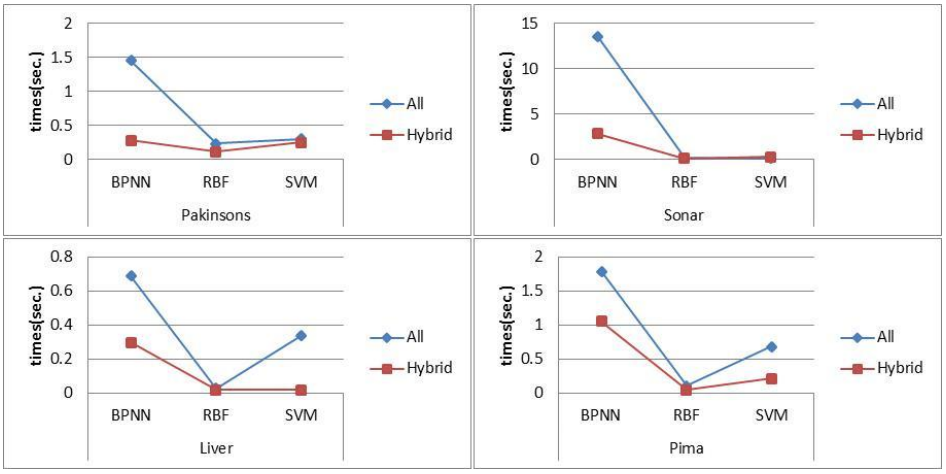
Proposed: วิธีการแบบผสม

Diff.: ความแตกต่างระหว่างวิธีการแบบผสมกับข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง

จากตารางที่ 4-5 เมื่อทำการเปรียบเทียบค่าความถูกต้องของการจำแนกข้อมูลคลาสส่วนน้อย เมื่อวิธีการแบบผสมกับข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง พบว่าชุดข้อมูล Pakinsons เมื่อทำการจำแนกข้อมูลด้วย BPNN RBF SVM ชุดข้อมูลที่ใช้วิธีการแบบผสมให้ค่าความถูกต้องของการจำแนกข้อมูลคลาสส่วนน้อยมากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 4.17% 35.42% และ 25% ชุดข้อมูล Sonar เมื่อทำการจำแนกข้อมูลด้วย BPNN ชุดข้อมูลที่ใช้วิธีการแบบผสมให้ค่าความถูกต้องของการจำแนกข้อมูลคลาสส่วนน้อยมากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 5.16% ส่วน RBF และ SVM ให้ค่าความถูกต้องน้อยกว่า 7.22% และ 3.09% ชุดข้อมูล Liver เมื่อทำการจำแนกข้อมูลด้วย BPNN RBF SVM

ชุดข้อมูลที่ใช้วิธีการแบบผสมให้ค่าความถูกต้องของการจำแนกข้อมูลคลาสส่วนน้อยมากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 8.27% 9.41% และ 15.18% ชุดข้อมูล Pima เมื่อทำการจำแนกข้อมูลด้วย BPNN RBF SVM ชุดข้อมูลที่ใช้วิธีการแบบผสมให้ค่าความถูกต้องของการจำแนกข้อมูลคลาสส่วนน้อยมากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 24.25% 29.85% และ 14.66%

สำหรับเวลาที่ใช้ในการประมวลผลในการจำแนกข้อมูลด้วย BPNN RBF SVM ชุดข้อมูล Pakinsons พบว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุงใช้เวลา 1.45 0.23 และ 0.3 วินาที วิธีการแบบผสมใช้เวลา 0.28 0.11 และ 0.25 วินาที ซึ่งน้อยกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 1.17 0.12 และ 0.05 วินาที ชุดข้อมูล Sonar พบว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุงใช้เวลา 13.49 0.14 และ 0.14 วินาที วิธีการแบบผสมใช้เวลา 2.84 0.08 และ 0.05 วินาที ซึ่งน้อยกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 10.65 0.06 และ 0.09 วินาทีตามลำดับ ชุดข้อมูล Liver พบว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุงใช้เวลา 0.69 0.03 และ 0.34 วินาที วิธีการแบบผสมใช้เวลา 0.3 0.02 และ 0.02 วินาที ซึ่งน้อยกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 0.39 0.01 และ 0.32 วินาที ชุดข้อมูล Pima พบว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุงใช้เวลา 1.79 0.11 และ 0.69 วินาที วิธีการแบบผสมใช้เวลา 1.06 0.05 และ 0.22 วินาที ซึ่งน้อยกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 0.73 0.06 และ 0.47 วินาที ดังรูปที่ 3



รูปที่ 4-1 เวลาที่ใช้ในการประมวลผลระหว่างวิธีการแบบผสมกับข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง

4.2. A Modified Error Function for Imbalanced Classification Problem

4.2.1 ชุดข้อมูล

สำหรับข้อมูลที่ใช้ในการทดลองนำมาจาก UCI Machine Learning Repository อัตราความมาสมดุลอยู่ในช่วง 1.86 และ 15 ชุดข้อมูลบางชุดเป็นปัญหาแบบ multi-classes ซึ่งในการทดลองนี้ได้ทำการปรับให้อยู่ในรูปแบบของปัญหาแบบสองคลาส นั่น 1) Yeast 2) Abalone โดยชุดข้อมูล Yeast ถูกปรับ Yeast4 Yeast5 และ Yeast6 โดยคลาสที่ถูกกำหนดให้เป็นคลาสส่วนน้อยได้แก่ ‘ME2’, ‘ME1’ และ ‘EXC’ ตามลำดับ ส่วน Abalone9-18 คลาส ‘18’ ถูกกำหนดให้เป็นคลาสส่วนน้อย และ and คลาส ‘9’ ถูกกำหนดให้เป็นคลาสส่วนมาก รายละเอียดของชุดข้อมูลแสดงในตารางที่ 4-6.

ตารางที่ 4-6 รายละเอียดของชุดข้อมูลมาตรฐาน

Datasets	#Att.	#Ins.	%Min.	%Maj.	IR	Max F1
Liver	7	345	42.03	57.97	1.38	0.05
Breast-w	10	683	34.99	65.01	1.86	3.59
Pima	9	768	34.9	65.1	1.87	0.57
Haberman	3	306	26.47	73.53	2.78	0.18
Abalone9-18	9	731	5.75	94.25	16.4	0.62
Yeast4	9	1484	3.44	96.56	28.1	1.23
Yeast5	9	1484	2.96	97.04	32.73	4.17
Yeast6	9	1484	2.36	97.64	41.4	1.94

4.2.2 การทดลองและผลการทดลอง

ชุดข้อมูลจากตารางที่ 4-6 จะถูกนำมาแบ่งเป็นสองกลุ่มคือชุดข้อมูลสำหรับฝึกสอน และชุดข้อมูลสำหรับทดสอบ โดยทำการแบ่งด้วยอัตรา 60 : 40 ผลการทดลองแสดงในตารางที่ 4-7

ชุดข้อมูล Liver, ค่า TPR ค่า G-Mean ของวิธีการที่นำเสนอได้ผลดีกว่าวิธีการ standard MSE 5.17% และ 1.21% ตามลำดับ ส่วนค่า F ของวิธีการที่นำเสนอได้ผลน้อยกว่าวิธีการ standard MSE 0.1%

ชุดข้อมูล Breast-w, ค่า TPR ค่า G-Mean และค่า F ของวิธีการที่นำเสนอได้ผลดีกว่าวิธีการ standard MSE 1.06% 0.56% and 0.53% ตามลำดับ

ชุดข้อมูล Pima, ค่า TPR และค่า F ของวิธีการที่นำเสนอได้ผลดีกว่าวิธีการ standard MSE 3.74% and 0.18% ตามลำดับ ส่วนค่า G-Mean ของวิธีการที่นำเสนอได้ผลน้อยกว่าวิธีการ standard MSE 0.28%.

ชุดข้อมูล Haberman, ค่า TPR ค่า G-Mean และค่า F ของวิธีการที่นำเสนอได้ผลดีกว่าวิธีการ standard MSE 6.25% 4.01% และ 4.39% ตามลำดับ

ชุดข้อมูล Abalone9-18, ค่า TPR ค่า G-Mean และค่า F ของวิธีการที่นำเสนอได้ผลดีกว่าวิธีการ standard MSE 12.5% 12.52% and 10.2% ตามลำดับ.

ชุดข้อมูล Yeast4, ค่า TPR ค่า G-Mean และค่า F ของวิธีการที่นำเสนอได้ผลดีกว่าวิธีการ standard MSE 25% 28.36% และ 32.16% ตามลำดับ

ชุดข้อมูล Yeast5, ค่า TPR ค่า G-Mean และค่า F ของวิธีการที่นำเสนอได้ผลดีกว่าวิธีการ standard MSE 11.76% 7.33% และ 8% ตามลำดับ.

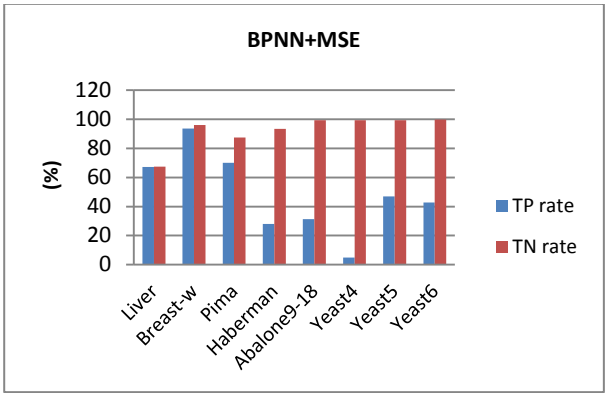
ชุดข้อมูล Yeast6, ค่า TPR ค่า G-Mean และค่า F ของวิธีการที่นำเสนอได้ผลดีกว่าวิธีการ standard MSE 7.16% 6.45% และ 5.21% ตามลำดับ

ตารางที่ 4-7 ประสิทธิภาพการจำแนกข้อมูล (%)

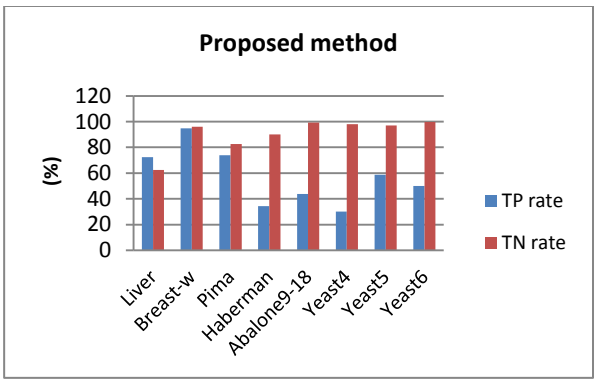
Datasets	Method	TPR	TNR	F	G	Acc
Liver	BPNN+MSE	67.24	67.50	63.41	67.37	67.39
	Proposed	72.41	62.50	64.62	67.27	66.67
Breast-w	BPNN+MSE	93.68	96.05	93.19	94.86	95.22
	Proposed	94.74	96.05	93.75	95.39	95.59
Pima	BPNN+MSE	70.09	87.50	71.77	77.87	80.78
	Proposed	73.83	82.50	71.49	78.05	79.48
Haberman	BPNN+MSE	28.13	93.33	38.30	51.23	76.23
	Proposed	34.38	90.00	42.31	55.62	75.41
Abalone9-18	BPNN+MSE	31.25	99.27	43.48	55.70	95.53
	Proposed	43.75	99.27	56.00	65.90	96.22
Yeast4	BPNN+MSE	5.00	99.30	8.00	22.28	96.12
	Proposed	30.00	98.08	36.36	54.44	96.46
Yeast5	BPNN+MSE	47.06	99.31	55.17	68.36	97.81
	Proposed	58.82	99.13	62.50	76.36	97.98
Yeast6	BPNN+MSE	42.857	99.83	57.143	65.40	98.482
	Proposed	50.02	99.83	63.60	70.62	98.72

รูปที่ 4-2 (ก) แสดงค่า TPR และ ค่า TNR ของ BPNN และ MSE และรูปที่ 4-2 (ข) แสดง TPR และ TNR ของวิธีการที่นำเสนอ ซึ่งจะเห็นว่าวิธีการที่นำเสนอให้ค่า TPR ของวิธีการที่นำเสนอได้ผลดีกว่าวิธีการ BPNN with MSE ส่วนค่าTNR ของวิธีการที่นำเสนอได้ผลน้อยกว่าวิธีการ BPNN with MSE

จากการทดลองจะเห็นว่าชุดข้อมูลส่วนใหญ่ค่า TPR ค่า G-Mean และค่า F ของวิธีการที่นำเสนอได้ผลดีกว่าวิธีการ standard MSE และวิธีการที่นำเสนอยังสามารถเพิ่มประสิทธิภาพความถูกต้องของการแบ่งกลุ่มข้อมูลของคลาสส่วนน้อยในขณะที่ไม่ทำให้คลาสส่วนมากเกิดความผิดพลาดในการแบ่งกลุ่มข้อมูลมากเกินไป



(ก)



(ข)

รูปที่ 4-2 (ก) TPR และ TNR ของ BPNN และ MSE
(b) TPR และ TNR ของวิธีการที่นำเสนอ

4.3. วิธีการแบบผสมผสมสำหรับชุดข้อมูลไม่สมดุลตามสภาพพื้นที่

4.3.1 ชุดข้อมูล

สำหรับวิธีการที่นำเสนอในหัวนี้คือ วิธีการแบบผสม (Soft-Hybrid) สำหรับข้อมูลไม่สมดุลตามสภาพพื้นที่ของข้อมูล โดยตารางที่ 4-8 แสดงรายละเอียดของข้อมูลที่ถูกสังเคราะห์ขึ้น และตารางที่ 4-10 แสดงรายละเอียดของข้อมูลมาจตรฐานจาก UCI (http://archive.ics.uci.edu/ml/) ซึ่งชุดข้อมูลมาตรฐานบางชุดข้อมูลเป็นหลายคลาส ดังนั้นในงานวิจัยนี้ได้นำเอาวิธีของ Keel (http://www.keel.es) มาทำการแปลงข้อมูลจากหลายคลาส เป้นข้อมูลสองคลาส ได้แก่ ชุดข้อมูล Page-blocks-1-3_vs_4 คลาส I, II และ III กำหนดให้เป็นคลาสส่วนมาก ส่วนคลาส IV กำหนดให้เป็นคลาสส่วนน้อย ชุดข้อมูล Vehicle1 คลาส *Saab* กำหนดให้เป็นคลาสส่วนน้อย ข้อมูลที่เหลือกำหนดให้เป็นคลาสส่วนมาก ชุดข้อมูล Vehicle3 คลาส *Opel* กำหนดให้เป็นคลาสส่วนน้อย ข้อมูลที่เหลือกำหนดให้เป็นคลาสส่วนมาก

ตารางที่ 4-8 รายละเอียดของข้อมูลสังเคราะห์

Name	$\#ins$	$\#att$	$\#min$	$\#maj$	IR	KL	$maxF$
A1	2000	2	280	1720	6.14	3.71	1.64
A2	2000	2	300	1700	5.67	3.1	1.74
A3	2000	2	290	1710	5.90	3.04	1.74
A4	2000	2	290	1710	5.90	3.06	1.75
A5	2000	2	313	1687	5.39	2.96	1.79
A6	2000	2	270	1730	6.41	3.97	1.8
A7	2000	2	204	1796	8.80	3.06	1.8
A8	2000	2	322	1678	5.21	3.14	1.87
A9	2000	2	270	1730	6.41	3.56	1.91
A10	2000	2	280	1720	6.14	3.28	1.93
A11	2000	2	200	1800	9.00	3.37	1.95
A12	2000	2	270	1730	6.41	3.34	1.95
A13	2000	2	210	1790	8.52	3.84	2.08

ตารางที่ 4-10 รายละเอียดของข้อมูลมาตรฐาน

Name	$\#ins$	$\#att$	$\#min$	$\#maj$	IR	KL	$maxF$
Breast-w	10	699	239	444	1.86	6.92	3.59
Haberman	306	3	81	225	2.78	0.70	0.18
Liver	345	6	145	200	1.38	0.86	0.05
Pakinsons	195	22	48	147	3.06	24.51	1.50
Pima	768	8	268	500	1.87	1.86	1.50
Page-blocks-1-3_vs_4	472	10	28	444	15.60	149	1.50
Abalone9-18	9	731	42	689	16.4	14.97	0.62
Yeast4	9	1484	51	1433	28.1	Inf	1.23
Yeast5	9	1484	44	144	32.73	Inf	4.17
Yeast6	9	1484	35	1449	41.4	Inf	1.94
Vehicle1	846	18	217	629	2.90	6.68	0.17
Vehicle3	846	18	212	634	2.99	6.01	0.18
German	1000	24	300	700	2.33	3.10	0.66

4.3.2 การทดลองและผลการทดลอง

4.3.2.1 ผลการทดลองของการหาพื้นที่ของข้อมูลไม่สมดุล

ในงานวิจัยนี้ได้ทำการพื้นที่ของชุดข้อมูลออกเป็น 3 พื้นที่ด้วยวิธีการที่นำเสนอในบทที่ 3 รายละเอียดของข้อมูลในแต่ละพื้นที่แสดงในตารางที่ 4-11

ตารางที่ 4-11 รายละเอียดของข้อมูลในแต่ละพื้นที่

Name	Non-overlapping region			Borderline region			Overlapping region		
	$\#ins$	$\#min$	$\#maj$	$\#ins$	$\#min$	$\#maj$	$\#ins$	$\#min$	$\#maj$
A1	829	97	732	225	68	157	946	115	831
A2	1024	59	965	303	91	212	673	150	523
A3	865	47	818	331	65	266	804	178	626
A4	597	46	551	458	117	341	945	127	818
A5	1068	56	1012	285	67	218	647	190	457
A6	1236	74	1162	160	79	81	604	117	487

ตารางที่ 4-11 (ต่อ)

Name	Non-overlapping region			Borderline region			Overlapping region		
	$\#ins$	$\#min$	$\#maj$	$\#ins$	$\#min$	$\#maj$	$\#ins$	$\#min$	$\#maj$
A7	987	37	941	320	57	263	702	110	592
A8	805	82	723	316	42	274	879	198	681
A9	622	68	554	535	78	457	843	124	719
A10	756	55	701	482	68	414	762	157	605
A11	1107	57	1050	295	18	276	599	125	474
A12	1039	66	973	209	40	169	752	164	588
A13	997	47	950	393	60	333	610	103	507
Haberman	100	1	99	38	27	11	168	53	115
Liver	5	3	2	140	53	87	200	89	111
Pakinsons	89	12	77	35	11	24	71	25	46
Pima	14	8	6	153	59	94	601	201	400
Page-	162	12	150	21	10	11	289	6	283
Vehicle1	30	21	9	280	2	278	536	194	342
Vehicle3	27	15	12	196	14	182	623	183	440
German	23	5	18	162	55	107	815	240	575

4.3.2.3 ผลการทดลองของการจำแนกประเภทข้อมูล

ในการทดลองกำหนดการทดลองแบบ 5-fold cross-validation ผลการทดลองของวิธีการที่นำเสนอจะถูกแบ่งออกเป็น 3 พื้นที่ จากนั้นจะนำผลที่ได้มารวมกันดังแสดงในตารางที่ 4-12

ตารางที่ 4-12 ผลการจำแนกประเภทข้อมูลของวิธีการ Soft-Hybrid

Data	RBFN				dDBSCAN				mKernel Learning				Final results (%)		
	TP	FN	TN	FP	TP	FN	TN	FP	TP	FN	TN	FP	TPR	F*	G
A1	97	0	732	0	64	4	137	20	46	69	808	23	73.93	78.11	84.90
A2	59	0	965	0	85	6	202	10	83	66	491	32	75.92	79.93	86.05
A3	47	0	818	0	57	8	257	9	104	74	598	28	71.72	77.76	83.77
A4	46	0	551	0	112	5	331	10	58	69	788	30	74.48	79.12	85.29
A5	56	0	1012	0	52	7	208	18	105	85	415	42	69.84	73.70	82.08
A6	74	0	1162	0	75	4	71	10	54	63	452	35	75.19	78.38	85.57
A7	37	0	941	0	49	8	251	12	50	60	563	29	66.67	71.39	80.71
A8	82	0	723	0	37	5	270	4	129	69	639	42	77.02	80.52	86.55
A9	68	0	554	0	74	4	441	16	62	62	686	33	75.56	78.01	85.68
A10	55	0	701	0	63	5	405	9	91	66	575	30	74.64	79.17	85.41
A11	57	0	1050	0	17	1	272	4	67	58	443	31	70.50	75.00	83.14
A12	66	0	973	0	33	7	162	7	105	57	545	43	76.12	78.16	85.98
A13	47	0	950	0	47	13	326	6	54	49	488	19	70.48	77.28	83.36
Haberman	1	0	99	0	24	3	7	4	19	34	103	12	54.32	62.41	71.03
Liver	3	0	2	0	32	21	60	27	60	29	82	29	65.52	64.19	68.68
Pakinsons	12	0	77	0	8	3	17	7	15	10	41	5	72.92	73.68	81.83
Pima	8	0	6	0	31	28	69	25	119	82	354	46	58.96	63.58	71.12
Page-blocks	12	0	150	0	14	4	10	5	65	59	283	59	59.09	58.90	71.86
Vehicle1	21	0	9	0	2	0	278	0	80	114	274	68	47.47	53.09	65.06
Vehicle3	15	0	12	0	2	12	168	14	79	104	368	72	45.28	48.73	62.56
German	5	0	18	0	41	38	50	33	130	110	495	80	54.32	57.42	67.26

F*: F-value on the minority class

ตารางที่ 4-13 เปรียบเทียบผลการจำแนกประเภทข้อมูลระหว่างขั้นตอนวิธี Soft-Hybrid และขั้นตอนวิธีมาตรฐาน

Data sets	RBFN			SVM1			SVM2			Soft-Hybrid		
	TPR	F*	G	TPR	F*	G	TPR	F*	G	TPR	F*	G
A1	72.29	76.24	83.97	55.71	68.72	74.25	61.07	72.00	77.60	**73.93**	**78.11**	**84.90**
A2	74.00	78.03	84.83	56.00	67.88	74.24	71.33	78.10	83.61	**75.92**	**79.93**	**86.05**
A3	70.13	74.26	80.88	53.45	65.82	72.55	68.62	76.69	82.11	**71.72**	**77.76**	**83.77**
A4	73.45	78.60	84.72	61.38	72.65	77.84	67.93	76.95	81.82	**74.48**	**79.12**	**85.29**
A5	69.65	**74.28**	82.06	52.08	64.17	71.48	66.13	74.33	80.43	**69.84**	73.70	**82.08**
A6	72.96	78.33	84.52	56.67	68.92	74.82	67.41	76.31	81.51	**75.19**	**78.38**	**85.57**
A7	63.24	70.49	78.79	46.08	59.12	67.50	58.82	68.97	76.18	**66.67**	**71.39**	**80.71**
A8	72.98	77.18	84.10	56.21	67.54	74.23	71.12	77.23	83.27	**77.02**	**80.52**	**86.55**
A9	74.81	77.29	85.39	63.70	73.98	79.28	72.22	77.47	84.19	**75.56**	**78.01**	**85.68**
A10	71.79	77.31	83.76	62.86	72.88	78.66	72.14	78.29	84.09	**74.64**	**79.17**	**85.41**
A11	69.00	72.25	82.04	49.50	63.26	70.08	61.00	71.76	77.71	**70.50**	**75.00**	**83.14**
A12	72.67	76.49	82.02	49.63	62.47	69.94	65.93	74.95	80.56	**76.12**	**78.16**	**85.98**
A13	70.05	76.02	82.45	56.19	69.21	74.69	65.24	75.07	80.36	**70.48**	**77.28**	**83.36**
Haberman	18.52	28.04	41.97	2.47	4.76	15.68	0.00	N/A	0.00	**54.32**	**62.41**	**71.03**
Liver	65.50	62.50	66.75	0.00	N/A	0.00	5.52	10.26	23.31	**65.52**	**64.19**	**68.68**
Pakinsons	70.83	72.34	80.65	50.00	65.75	70.47	58.33	72.73	76.12	**72.92**	**73.68**	**81.83**
Pima	58.21	63.29	70.84	50.37	59.87	67.48	54.48	62.26	69.63	**58.96**	**63.58**	**71.12**
Page-blocks-	42.67	49.52	61.03	50.00	58.37	67.40	50.67	**58.80**	67.80	**59.09**	57.90	**71.86**
Vehicle1	42.40	50.69	62.25	16.59	27.91	40.57	18.89	30.15	42.98	**47.47**	**53.09**	**65.06**
Vehicle3	44.34	48.08	61.96	0.94	1.87	9.71	4.25	7.96	20.52	**45.28**	**48.73**	**62.56**
German	52.67	56.23	66.96	52.67	59.74	68.79	49.00	56.21	66.09	**54.32**	**57.42**	**67.26**

F*: F-value on minority class

ตารางที่ 4-13 แสดงการการเปรียบเทียบผลการจำแนกประเภทข้อมูลระหว่างวิธีการที่นำเสนอกับขั้นตอนวิธีมาตรฐานได้แก่ โครงข่ายฟังก์ชันรัศมีฐาน (RBFN), ซัพพอร์ตเวกเตอร์แมชชีนกับโพลิโนเมียลดีกรี2 (SVM1) และซัพพอร์ตเวกเตอร์แมชชีนกับเกาส์เซียนฟังก์ชั่น (SVM2) ผลการจำแนกประเภทข้อมูลของชุดข้อมูลสังเคราะห์พบว่า ค่า TPR ของวิธีการที่นำเสนอให้ผลดีกว่าในทุกชุดข้อมูลเมื่อเทียบกับขั้นตอนวิธีอื่นทั้ง 3 วิธีข้างต้น ส่วนค่า F-minor ของวิธีการที่นำเสนอให้ผลดีว่า SVM1 แต่ชุดข้อมูล A5 วิธีการที่นำเสนอให้ค่าที่น้อยกว่า RBFN และ SVM2 สำหรับค่า G-mean วิธีการที่นำเสนอให้ผลดีกว่าในทุกชุดข้อมูลเมื่อเทียบกับขั้นตอนวิธีอื่นทั้ง 3 วิธีข้างต้น

ผลการจำแนกประเภทข้อมูลของชุดข้อมูลมาตรฐานค่า TPR ของวิธีการที่นำเสนอมากกว่าในทุกชุดข้อมูลเมื่อเทียบกับขั้นตอนวิธีอื่นทั้ง 3 วิธีข้างต้น ส่วนค่า F-minor ของชุดข้อมูล Page-blocks-1-3_vs_4 พบว่าวิธีการที่นำเสนอให้ค่ามากกว่า RBFN แต่น้อยกว่า SVM1 และ SVM2 สำหรับค่า G-mean ของวิธีการที่นำเสนอให้ค่ามากกว่าทั้ง 8 ชุดข้อมูลเมื่อเทียบกับ RBFN และ SVM2 และมีเพียง 1 ชุดข้อมูลคือ German ที่วิธีการที่นำเสนอมีค่าน้อยกว่าเมื่อเทียบกับ SVM1

ตารางที่ 4-14 เปรียบเทียบผลการจำแนกประเภทข้อมูลระหว่างขั้นตอนวิธี Soft-Hybrid และขั้นตอนวิธีมาตรฐานรวมกับ SMOTE

Data sets	SMOTE+RBFN			SMOTE+SVM1			SMOTE+SVM2			Soft-Hybrid		
	TPR	F*	G	TPR	F*	G	TPR	F*	G	TPR	F*	G
A1	65.35	72.41	76.17	63.28	71.71	75.55	63.05	71.56	75.43	**73.93**	**78.11**	**84.90**
A2	70.67	76.19	78.35	70.28	76.08	78.27	70.42	76.23	78.41	**75.92**	**79.93**	**86.05**
A3	67.42	74.16	77.14	66.85	74.05	77.07	66.91	74.11	77.11	**71.72**	**77.76**	**83.77**
A4	59.82	69.15	73.77	59.26	68.98	73.6	59.18	69.05	73.64	**74.48**	**79.12**	**85.29**
A5	**71.00**	76.25	78.14	70.60	*76.26*	78.2	70.21	76.07	78.05	69.84	**73.70**	**82.08**
A6	72.75	77.76	79.29	72.74	77.77	79.31	72.81	77.79	79.32	**75.19**	**78.38**	**85.57**
A7	**69.41**	**75.44**	78.22	68.41	74.97	77.84	67.95	74.79	77.70	66.67	71.39	**80.71**
A8	65.40	72.65	76.13	65.55	72.74	76.21	64.62	72.33	75.87	**77.02**	**80.52**	**86.55**
A9	58.85	68.42	73.30	59.31	68.83	73.63	57.62	67.87	72.79	**75.56**	**78.01**	**85.68**
A10	64.08	72.16	75.90	64.23	72.14	75.89	64.10	72.14	75.89	**74.64**	**79.17**	**85.41**
A11	70.57	76.35	78.75	**70.74**	**76.47**	78.85	70.02	76.20	78.64	70.50	75.00	**83.14**
A12	70.93	76.17	78.46	69.61	75.67	78.08	69.10	75.47	77.94	**76.12**	**78.16**	**85.98**
A13	69.56	75.56	78.29	68.05	75.07	77.91	67.90	75.00	77.86	**70.48**	**77.28**	**83.36**
Haberman	**67.04**	**76.19**	**78.91**	60.34	72.48	75.75	65.36	75.48	78.29	54.3	62.41	71.03
Liver	**66.90**	61.39	65.01	11.41	19.32	32.94	8.28	14.46	28.12	65.5	**64.19**	**68.68**
Pakinsons	**84.07**	82.25	84.21	80.53	87.50	**88.51**	79.65	**87.80**	8.64	72.9	73.68	81.83
Pima	54.85	60.87	68.86	52.24	62.08	69.11	57.09	63.22	70.66	**59**	**63.58**	**71.12**
Page-blocks-1-3_vs_4	**95.41**	**95.57**	96.24	94.75	97.31	**97.34**	94.75	97.31	7.34	59.1	58.9	71.86
Vehicle1	45.16	51.72	63.76	20.28	32.96	44.82	20.28	32.47	44.68	**47.5**	**53.09**	**65.06**
Vehicle3	43.87	48.19	61.88	1.42	2.78	11.89	6.60	11.81	25.47	**45.3**	**48.73**	**62.56**
German	51.12	54.98	65.70	53.67	60.22	69.12	51.44	57.09	66.95	**54.3**	**57.42**	**67.26**

F*: F-value on minority class

ตารางที่ 4-14 แสดงการการเปรียบเทียบผลการจำแนกประเภทข้อมูลระหว่างวิธีการที่นำเสนอกับขั้นตอนวิธีมาตรฐานรวมกับวิธีการ SMOTE ได้แก่ SMOTE+RBFN, SMOTE+SVM1 และ SMOTE+SVM2 ผลการจำแนกประเภทข้อมูลของชุดข้อมูลสังเคราะห์พบว่า ค่า TPR ของวิธีการที่นำเสนอมากกว่าในทุกชุดข้อมูลเมื่อเทียบกับขั้นตอนวิธี SMOTE+SVM2 แต่สำหรับชุดข้อมูล A5, A7 วิธีการที่นำเสนอมีค่าน้อยกว่าขั้นตอนวิธี SMOTE+RBFN ชุดข้อมูล A11 วิธีการที่นำเสนอมีค่าน้อยกว่าขั้นตอนวิธี SMOTE+SVM1 ส่วนค่า F-minor ของวิธีการที่นำเสนอให้ผลดีกว่าขั้นตอนวิธีทั้งสามข้างต้นใน 10 ชุดข้อมูล แต่สำหรับชุดข้อมูล A7 และ A11 ให้ค่าน้อยกว่าขั้นตอนวิธีทั้งสาม สำหรับค่า G-mean วิธีการที่นำเสนอให้ผลดีกว่าในทุกชุดข้อมูลเมื่อเทียบกับขั้นตอนวิธีอื่นทั้ง 3 วิธีข้างต้น

ผลการจำแนกประเภทข้อมูลของชุดข้อมูลมาตรฐานค่า TPR ของวิธีการที่นำเสนอมากกว่าใน 5 ชุดข้อมูลเมื่อเทียบกับขั้นตอนวิธี SMOTE+SVM1 และ SMOTE+SVM2 แต่สำหรับชุดข้อมูล Haberman, Liver, Pakinsons และ Page-blocks-1-3_vs_4 วิธีการที่นำเสนอมีค่าน้อยกว่าขั้นตอนวิธี SMOTE+RBFN และชุดข้อมูล Haberman, Liver และ Pakinsons วิธีการ SMOTE+SVM1 ที่นำเสนอมีค่าน้อยกว่าขั้นตอนวิธี SMOTE+SVM1 and SMOTE+SVM2 ส่วนค่า F-minor พบว่าวิธีการที่นำเสนอให้ค่ามากกว่า RBFN แต่น้อยกว่า SVM1 และ SVM2 ในชุดข้อมูลมาตรฐาน 5 ชุด และมี 3 ชุดข้อมูลคือ Haberman, Pakinsons และ Page-blocks-1-3_vs ที่วิธีการที่นำเสนอให้ผลดีว่าขั้นตอนวิธีทั้งสามที่นำมาเปรียบเทียบข้างต้น สำหรับค่า G-mean ของวิธีการที่นำเสนอให้ค่ามากกว่าทั้ง 8 ชุดข้อมูลเมื่อเทียบกับ RBFN และ SVM2 และมีเพียง 1 ชุดข้อมูลคือ German ที่วิธีการที่นำเสนอมีค่าน้อยกว่าเมื่อเทียบกับ SVM1 สำหรับค่า G-mean ใน 6 ชุดข้อมูลวิธีการที่นำเสนอให้ผลดีว่าขั้นตอนวิธี SMOTE+SVM2 และมี 3 ชุดข้อมูล ได้แก่ Haberman, Pakinsons and Page-blocks-1-3_vs_4 ที่วิธีการที่นำเสนอให้ผลน้อยกว่าขั้นตอนวิธี SMOTE+RBFN และ SMOTE+SVM1 นอกจากนี้มี 2 ชุดข้อมูลได้แก่ Haberman และ Pakinsons ที่วิธีการที่นำเสนอให้ค่าน้อยกว่าขั้นตอนวิธี SMOTE+SVM2

ตารางที่ 4-15 เปรียบเทียบผลการจำแนกประเภทข้อมูลระหว่างขั้นตอนวิธี Soft-Hybrid และขั้นตอนวิธีมาตรฐานรวมกับ ROS

Data sets	ROS+RBFN			ROS+SVM1			ROS+SVM2			Soft-Hybrid		
	TPR	F*	G	TPR	F*	G	TPR	F*	G	TPR	F*	G
A1	65.45	72.45	76.2	63.6	71.92	75.7	63.2	71.74	75.5	**73.9**	**78.1**	**84.90**
A2	70.89	76.17	78.3	70.6	76.21	78.3	70.0	76.08	78.2	**75.9**	**79.9**	**86.05**
A3	68.07	74.40	77.33	66.6	73.94	76.9	66.6	73.97	77.0	**71.7**	**77.7**	**83.77**
A4	60.22	69.45	74.02	59.4	69.11	73.7	59.4	69.19	73.7	**74.4**	**79.1**	**85.29**
A5	**71.27**	**76.45**	78.31	71.0	76.34	78.2	70.6	76.17	78.1	69.8	73.70	**82.08**
A6	73.27	77.87	79.33	72.8	77.74	79.2	72.8	77.84	79.3	**75.1**	**78.3**	**85.57**
A7	**68.24**	74.92	77.8	68.5	*75.04*	77.8	67.9	74.79	77.7	66.6	**71.3**	**80.71**
A8	66.02	72.88	76.32	65.6	72.72	76.2	64.7	72.36	75.9	**77.0**	**80.5**	**86.55**
A9	59.67	68.92	73.73	59.3	68.78	73.6	59.0	68.67	73.4	**75.5**	**78.0**	**85.68**
A10	63.97	72.14	75.88	64.4	72.30	76.0	63.8	72.03	75.7	**74.6**	**79.1**	**85.41**
A11	70.49	**76.35**	78.75	70.4	76.22	78.6	70.2	76.27	78.6	**70.5**	75.00	**83.14**
A12	70.86	76.04	78.34	69.7	75.73	78.1	69.4	75.59	78.0	**76.1**	**78.1**	**85.98**
A13	68.74	75.33	78.11	67.8	75.02	77.8	67.6	75	77.8	**70.4**	**77.2**	**83.36**
Haberman	**66.48**	**75.32**	**78.21**	58.66	70.47	74.17	65.36	75.48	78.29	54.32	62.41	71.03
Liver	51.01	55.27	62.10	2.68	5.23	16.38	17.45	27.66	40.43	**65.52**	**64.19**	**68.68**
Pakinsons	76.99	86.14	87.15	80.53	88.78	89.43	**81.42**	**89.32**	**89.92**	72.92	73.68	81.83
Pima	**59.70**	**64.00**	**71.51**	52.24	61.95	69.04	52.99	61.61	68.99	58.96	63.58	71.12
Page-blocks-1-3_vs_4	**95.74**	**96.05**	**96.63**	94.75	97.14	97.23	94.75	97.14	97.23	59.09	57.90	71.86
Vehicle1	44.70	52.01	63.71	16.13	27.13	39.97	10.14	18.03	31.72	**47.47**	**53.09**	**65.06**
Vehicle3	**57.55**	**54.34**	**68.67**	0.00	NaN	0.00	5.66	10.53	23.72	45.28	48.73	62.56
German	50.16	53.49	64.63	52.72	59.89	68.72	**55.91**	**61.73**	**70.43**	54.32	57.42	67.26

F*: F-value on minority class

ตารางที่ 4-15 แสดงการการเปรียบเทียบผลการจำแนกประเภทข้อมูลระหว่างวิธีการที่นำเสนอกับขั้นตอนวิธีมาตรฐานรวมกับวิธีการสุ่มเพิ่มข้อมูล (Random Over Sampling ROS) ได้แก่ RBFN+ROS, SVM1+ROS และ SVM2+ROS

ผลการจำแนกประเภทข้อมูลของชุดข้อมูลสังเคราะห์พบว่า ค่า TPR ของวิธีการที่นำเสนอมากกว่า 11 ชุดข้อมูลเมื่อเทียบกับขั้นตอนวิธีทั้งสามข้างต้น แต่สำหรับชุดข้อมูล A5,และ A7 วิธีการที่นำเสนอมีค่าน้อยกว่าขั้นตอนวิธีทั้งสาม ส่วนค่า F-minor ของวิธีการที่นำเสนอให้ผลดีกว่าทั้งสามขั้นตอนวิธีข้างต้นอยู่ 10 ชุดข้อมูล แต่สำหรับชุดข้อมูล A5, A7 anและ A11 วิธีการที่นำเสนอให้ค่าน้อย

กว่าขั้นตอนวิธีทั้งสาม สำหรับค่า G-mean วิธีการที่นำเสนอให้ผลดีกว่าในทุกชุดข้อมูลเมื่อเทียบกับขั้นตอนวิธีอื่นทั้ง 3 วิธีข้างต้น

ผลการจำแนกประเภทข้อมูลของชุดข้อมูลมาตรฐานค่า TPR ของวิธีการที่นำเสนอมากกว่าใน 5 ชุดข้อมูลเมื่อเทียบกับขั้นตอนวิธี ROS+RBFN แต่มี 5 ชุดข้อมูล ได้แก่ Haberman, Pakinsons, Pima, Page-blocks-1-3_vs_4 และ Vehicle3 ที่วิธีการที่นำเสนอให้ค่าน้อยกว่าขั้นตอนวิธี ROS+RBFN นอกจากนี้ มี 3 ชุดข้อมูล ได้แก่ Haberman, Pakinsons และ Page-blocks-1-3_vs_4 ที่วิธีการที่นำเสนอให้ค่าน้อยกว่า ROS+SVM1 และมี 4 ชุดข้อมูล ได้แก่ Haberman, Pakinsons, Page-blocks-1-3_vs_4 และ German ที่วิธีการที่นำเสนอให้ค่าน้อยกว่า ROS+SVM2 สำหรับค่า G-mean ของวิธีการที่นำเสนอให้ค่ามากกว่าทั้ง 4 ชุดข้อมูลเมื่อเทียบกับ ROS+SVM1 and ROS+SVM2 และ SVM2 แต่มี 5 ชุดข้อมูลได้แก่ Haberman, Pakinsons, Pima, Page-blocks-1-3_vs_4 and Vehicle3 ที่วิธีการที่นำเสนอให้ค่าน้อยกว่าขั้นตอนวิธี ROS+RBFN และมีอีก 4 ชุดข้อมูล ได้แก่ Haberman, Pakinsons, Page-blocks-1-3_vs_4 และ German ที่วิธีการที่นำเสนอมีค่าน้อยกว่าเมื่อเทียบกับขั้นตอนวิธี ROS+SVM1 และ ROS+SVM2 แต่ 5 ชุดข้อมูล ได้แก่ Haberman, Pakinsons, Pima, Page-blocks-1-3_vs_4 และ Vehicle3 วิธีการที่นำเสนอมีค่าน้อยกว่าเมื่อเทียบกับขั้นตอนวิธี ROS+SVM1 and ROS+SVM2

ตารางที่ 4-16 เปรียบเทียบผลการจำแนกประเภทข้อมูลระหว่างขั้นตอนวิธี Soft-Hybrid และขั้นตอนวิธีมาตรฐานรวมกับ RUS

Data sets	RUS+RBFN			RUS+SVM1			RUS+SVM2			Soft-Hybrid		
	TPR	F*	G	TPR	F*	G	TPR	F*	G	TPR	F*	G
A1	71.43	77.97	83.22	60.71	72.65	77.27	62.5	73.22	78.21	**73.93**	**78.11**	**84.90**
A2	73.00	78.07	83.15	67.33	75.94	80.49	71.33	78.39	82.74	**75.92**	**79.93**	**86.05**
A3	**72.07**	**77.99**	83.20	64.48	74.21	79.14	70.00	77.48	82.27	71.72	77.76	**83.77**
A4	**75.86**	**80.00**	85.64	66.21	75.59	80.49	68.62	77.28	81.94	74.48	79.12	**85.29**
A5	**71.88**	**75.76**	81.47	62.3	71.82	77.02	69.01	75.00	80.36	69.84	73.70	**82.08**
A6	73.70	78.66	83.34	64.81	75.43	79.31	69.63	77.21	81.54	**75.19**	**78.38**	**85.57**
A7	66.67	**72.92**	80.13	49.02	61.35	69.15	59.80	70.32	76.42	**66.67**	71.39	**80.71**
A8	75.16	79.08	84.66	70.81	77.03	82.43	70.50	76.82	82.24	**77.02**	**80.52**	**86.55**
A9	73.70	78.66	84.56	68.89	76.39	81.96	70.37	77.24	82.8	**75.56**	**78.01**	**85.68**
A10	73.93	79.16	84.53	65.36	74.09	79.67	69.64	76.92	82.20	**74.64**	**79.17**	**85.41**
A11	68.50	73.85	81.00	48.50	62.78	69.12	60.50	71.18	76.86	**70.5**	**75.00**	**83.14**
A12	72.59	77.01	82.95	64.07	73.62	78.72	66.67	74.53	80.00	**76.12**	**78.16**	**85.98**
A13	69.05	75.13	81.63	56.67	69.39	74.68	60.48	71.75	77.02	**70.48**	**77.28**	**83.36**

ตารางที่ 4-16 (ต่อ)

Data sets	RUS+RBFN			RUS+SVM1			RUS+SVM2			Soft-Hybrid		
	TPR	F*	G	TPR	F*	G	TPR	F*	G	TPR	F*	G
Haberman	41.98	55.28	62.71	43.21	55.12	62.82	38.27	49.21	58.35	**54.32**	**62.41**	**71.03**
Liver	54.48	59.18	65.40	7.59	13.75	27.27	6.21	11.32	24.60	**65.52**	**64.19**	**68.68**
Pakinsons	**89.58**	**74.14**	78.91	58.33	73.67	76.38	58.33	73.67	76.38	72.92	73.68	**81.83**
Pima	52.99	59.41	67.66	54.10	62.63	69.78	54.85	63.91	70.65	**58.96**	**63.58**	**71.12**
Page-blocks-	**71.43**	**78.43**	**83.75**	42.86	60.00	65.47	50.00	66.67	70.71	59.09	57.90	71.86
Vehicle1	**51.15**	**54.68**	**66.94**	6.91	12.88	26.27	18.43	30.42	42.73	47.47	53.09	65.06
Vehicle3	35.38	42.98	56.50	1.89	3.70	13.74	8.49	15.19	28.98	**45.28**	**48.73**	**62.56**
German	46.67	52.43	63.47	53.33	**59.81**	**68.93**	49.33	55.74	65.86	**54.32**	57.42	67.26

F*: F-value on minority class

ตารางที่ 4-16 แสดงการการเปรียบเทียบผลการจำแนกประเภทข้อมูลระหว่างวิธีการที่นำเสนอกับขั้นตอนวิธีมาตรฐานรวมกับวิธีการสุ่มลดข้อมูล (Random Under Sampling) ได้แก่ RUS+RBFN, RUS+SVM1 และ RUS+SVM2

ผลการจำแนกประเภทข้อมูลของชุดข้อมูลสังเคราะห์พบว่า ค่า TPR ของวิธีการที่นำเสนอมากกว่า 13 ชุดข้อมูลเมื่อเทียบกับขั้นตอนวิธีทั้งสามข้างต้น แต่มี 3 ชุดข้อมูลได้แก่ A3, A4 and A5ที่วิธีการที่นำเสนอมีค่าน้อยกว่าขั้นตอนวิธี RUS+RBFN ส่วนค่า F-minor ของวิธีการที่นำเสนอให้ผลดีกว่าขั้นตอนวิธี RUS+SVM1 ทั้ง 13 ชุดข้อมูล แต่ในขั้นตอนวิธี RUS+RBFN มี 4 ชุดข้อมูล ได้แก่ A3, A4, A5 และ A7 วิธีการที่นำเสนอมีค่าน้อยกว่า ส่วน RUS+SVM2 มีชุดข้อมูล A5 วิธีการที่นำเสนอมีค่าน้อยกว่าสำหรับค่า G-mean วิธีการที่นำเสนอให้ผลดีกว่าในทุกชุดข้อมูลเมื่อเทียบกับขั้นตอนวิธีอื่นทั้ง 3 วิธีข้างต้น

ผลการจำแนกประเภทข้อมูลของชุดข้อมูลมาตรฐานค่า TPR ของวิธีการที่นำเสนอมากกว่าใน 7 ชุดข้อมูลเมื่อเทียบกับขั้นตอนวิธี RUS+SVM1 และมี 3 ชุดข้อมูล ได้แก่ Haberman, Page-blocks-1-3_vs_4 และ Vehicle1 วิธีการที่นำเสนอมีค่าน้อยกว่าขั้นตอนวิธี RUS+RBFN นอกจากนี้ชุดข้อมูล Page-blocks-1-3_vs_4 วิธีการที่นำเสนอมีค่าน้อยกว่าขั้นตอนวิธี RUS+SVM1 และ RUS+SVM2สำหรับค่า F-minor มี 7 ชุดข้อมูล วิธีการที่นำเสนอมีค่ามากกว่าขั้นตอนวิธี RUS+SVM2 และมี 3 ชุดข้อมูลได้แก่ Pakinsons, Page-blocks-1-3_vs_4 และ Vehicle1 ที่วิธีการที่นำเสนอมีค่าน้อยกว่าขั้นตอนวิธี RUS+RBFN และมี 2 ชุดข้อมูลได้แก่ Pakinsons และ German ที่วิธีการที่นำเสนอมีค่าน้อยกว่าขั้นตอนวิธี RUS+SVM1 นอกจากนี้ชุดข้อมูล Page-blocks-1-3_vs_4 วิธีการที่นำเสนอมีค่าน้อยกว่าขั้นตอนวิธี RUS+SVM2 สำหรับค่า G-mean ของวิธีการที่นำเสนอให้ค่ามากกว่า

ทั้ง 8 ชุดข้อมูลเมื่อเทียบกับ RUS+SVM2 แต่มี 2 ชุดข้อมูลได้แก่ Page-blocks-1-3_vs_4 และ Vehicle1 ที่วิธีการที่นำเสนอให้ค่าน้อยกว่าขั้นตอนวิธี RUS+RBFN นอกจากนี้ชุดข้อมูล German วิธีการที่นำเสนอให้ค่าน้อยกว่าขั้นตอนวิธี RUS+SVM1

จากผลการทดลอง ในตารางที่ 4-18 สามารถนำมาหาค่าเฉลี่ยเปอร์เซ็นต์การเพิ่มขึ้นของตัววัดประสิทธิภาพของวิธีการที่นำเสนอกับขั้นตอนวิธีการอื่นทั้ง 12 ขั้นตอนวิธี ผลแสดงดังตารางที่ 4-17 จะเห็นว่าในชุดข้อมูลสังเคราะห์ วิธีการที่นำเสนอให้ค่าเฉลี่ยเปอร์เซนต์การเพิ่มขึ้นมากขึ้นเมื่อเทียบกับขั้นตอนวิธีการอื่นทั้ง 12 ขั้นตอนวิธี สำหรับชุดข้อมูลมาตรฐาน วิธีการที่นำเสนอให้ค่าเฉลี่ยเปอร์เซนต์การเพิ่มขึ้นมากขึ้นเมื่อเทียบกับขั้นตอนวิธีการอื่น 10 ขั้นตอนวิธี มีพียงขั้นตอนวิธี SMOTE+RBFN และ ROS+RBFN ที่ค่า TPR F-minor และ G-mean ในชุดข้อมูลมาตรฐานมีค่าเฉลี่ยเปอร์เซนต์การเพิ่มขึ้นมากขึ้นกว่าวิธีการที่นำเสนอ

ตารางที่ 4-17 ค่าเฉลี่ยเปอร์เซนต์การเพิ่มขึ้นของตัววัดประสิทธิภาพ

Methods	Synthetic data sets			Standard benchmarked data sets		
	TPR	F*	G	TPR	F*	G
RBFN	3.27	2.01	1.77	31.65	19.34	12.21
SVM1	33.14	15.11	14.65	1012.41	546.09	140.52
SVM2	8.75	2.94	4.33	49.25	159.95	66.90
SMOTE+RBFN	9.26	4.76	9.72	**-6.28**	**-6.21**	**-2.96**
SMOTE+SVM1	10.10	4.96	9.88	458.15	234.53	67.53
SMOTE+SVM2	10.80	5.20	10.09	169.67	81.03	36.68
ROS+RBFN	7.57	3.26	8.96	**-5.18**	**-6.96**	**-4.19**
ROS+SVM1	8.03	3.42	9.08	356.62	164.85	49.14
ROS+SVM2	8.69	3.61	9.23	160.88	76.43	36.62
RUS+RBFN	1.55	0.22	1.76	7.78	2.85	3.35
RUS+SVM1	18.73	7.36	9.05	468.71	237.97	85.56
RUS+SVM2	9.80	3.03	5.21	206.14	97.50	47.58

F*: F-value on minority class

4.3.2.4 ผลการเปรียบทียบเวลาที่ใช้ในการประมวลผล

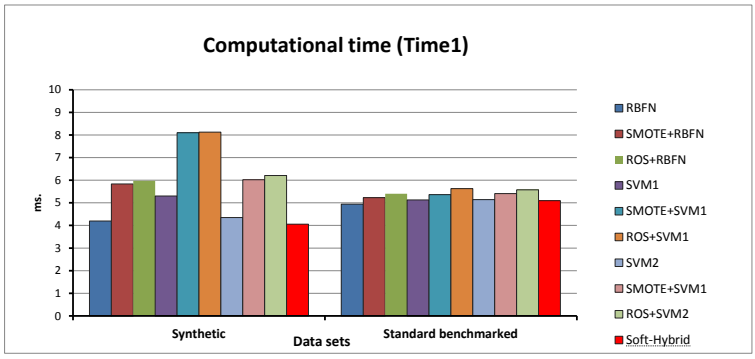
เวลาเฉลี่ยที่ใช้ในการประมวลผลข้อมูลในงานวิจัยนี้จะแบ่งออกเป็น 2 ส่วนคือ Time1 และ Time2 โดย Time1 เป็นการเปรียบเทียบเวลาในการประมวลผลข้อมูลของวิธีการที่นำเนอที่อยู่ในพื้นที่ซ้อนทับ พื้นที่บริเวณขอบ และพื้นที่ไม่ซ้อนทับ กับขั้นตอนวิธี RBFN, SVM1, SVM2, SMOTE

+RBFN, SMOTE+SVM1, SMOTE+SVM2, ROS+RBFN, ROS+SVM1 และ ROS+SVM2 ส่วน Time 2 เป็นการแปรียบเทียบเวลาในการประมวลผลข้อมูลของวิธีการที่นำเนอที่อยู่ในพื้นที่ซ้อนทับ พื้นที่บริเวณขอบกับขั้นตอนวิธี RUS+RBFN, RUS+SVM1 และ RUS+SVM2 ผลการทดลองแสดงดังตารางที่ 4-18

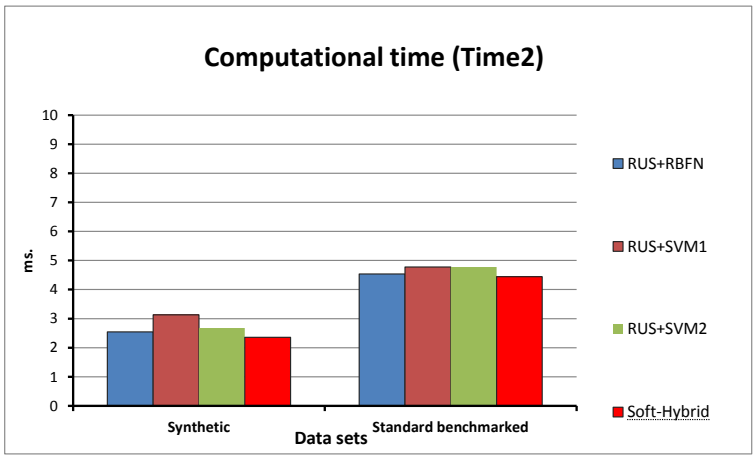
ตารางที่ 4-18 เวลาเฉลี่ยที่ใช้ในการประมวลผล

Methods	Times (ms.)	
	Synthetic	Standard Benchmarked
Time1		
RBFN	4.20	4.94
SVM1	5.30	5.13
SVM2	4.35	5.14
SMOTE+RBFN	5.83	5.23
SMOTE+SVM1	8.10	5.36
SMOTE+SVM2	6.02	5.41
ROS+RBFN	5.97	5.40
ROS+SVM1	8.12	5.63
ROS+SVM2	6.21	5.58
Proposed method	4.06	5.10
Time2		
RUS+RBFN	2.54	4.53
RUS+SVM1	3.13	4.77
RUS+SVM2	2.68	4.77
Proposed method	2.36	4.44

จากตารางที่ 4-18 จะเห็นว่าค่าเฉลี่ยเวลาที่ใช้ในการประมวลผลของ Time1 ของวิธีการที่นำเสนอกับข้อมูลสังเคราะห์น้อยกว่าวิธีการอื่นทั้ง 9 วิธี ได้แก่ RBFN, SVM1, SVM2, SMOTE +RBFN, SMOTE+SVM1, SMOTE+SVM2, ROS+RBFN, ROS+SVM1 และ ROS+SVM2 ส่วนข้อมูลมาตรฐานขั้นตอนวิธี RBFN ให้เวลาเฉลี่ยในการประมวลผลที่สุดเมือ่เทียบกับ SVM1, SVM2, SMOTE +RBFN, SMOTE+SVM1, SMOTE+SVM2, ROS+RBFN, ROS+SVM, ROS+SVM2 และ วิธีการที่นำเสนอ ส่วน Time2 เวลาเฉลี่ยในการประมวลผลของวิธีการที่นำเสอนกับทั้งข้อมูลสังเคราะห์และข้อมูลมาตรฐานต่ำสุดเมื่อเทียบกับขั้นตอนวิธี RUS+RBFN, RUS+SVM1 และ RUS+SVM2 ผลของเวลาเฉลี่ยในการประมวลผลแสดงดังรูปที่ 4-1



(ก)



(ข)

รูปที่ 4-3 เวลาเฉลี่ยในการประมวลผล (ก) Time1 (ข) Time2

บทที่ 5 สรุปผลการทดลอง

5.1 การเลือกคุณลักษณะสำหรับชุดข้อมูลไม่สมดุล

งานวิจัยนี้ได้นำเสนอวิธีการเลือกคุณลักษณะสำหรับข้อมูลไม่สมดุลที่มีจำนวนมิติข้อมูลที่สูง เพื่อลดจำนวนมิติข้อมูลที่ซ้ำซ้อนและเพิ่มประสิทธิภาพของการจำแนกข้อมูลด้วยวิธีการแบบผสม โดยทำการเลือกข้อมูลคลาสส่วนมากให้มีจำนวนเท่ากับหรือใกล้เคียงกับคลาสส่วนน้อยด้วยวิธีการจัดกลุ่ม และนำค่าไอเกนมากำหนดจำนวนคุณลักษณะที่ต้องการ จากนั้นค้นหาคุณลักษณะย่อยด้วยขั้นตอนวิธีทางพันธุกรรมร่วมกับโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ ทำการเปรียบเทียบประสิทธิภาพของวิธีการที่นำเสนอโดยใช้ค่าความถูกต้อง ค่าเฉลี่ยเรขาคณิต และความถูกต้องของการจำแนกข้อมูลคลาสส่วนน้อย รวมถึงเวลาที่ใช้ในการประมวลผล ทำการทดสอบกับ 3 ตัวจำแนกข้อมูลได้แก่ โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ (BPNN) โครงข่ายฟังก์ชันรัศมีฐาน (RBF Network) และซัพพอร์ตเวกเตอร์แมชชีน (SVM) ผลปรากฏว่าตัวจำแนกข้อมูลทั้ง 3 วิธีให้ค่าความถูกต้อง (Accuracy) ค่าเฉลี่ยเรขาคณิต (G-Mean) โดยส่วนมากของการจำแนกชุดข้อมูลของวิธีการแบบผสมดีกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง รวมถึงค่าความถูกต้องของการจำแนกข้อมูลคลาสส่วนน้อยมีความถูกต้องมากขึ้น โดยแต่ละชุดข้อมูลมีความเหมาะสมกับวิธีการที่ใช้ในการจำแนกข้อมูลที่แตกต่างกัน ในส่วนของเวลาที่ใช้ในการประมวลผลพบว่าชุดข้อมูลทั้งหมดที่ใช้วิธีการแบบผสมใช้เวลาน้อยกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุงในการจำแนกข้อมูลทั้ง 3 วิธี

5.2 A Modified Error Function for Imbalanced Classification Problem

วิธีการที่นำเสนอในหัวข้อนี้คือการปรับปรุงค่าความผิดพลาดเฉลี่ยกำลังสอง (modified MSE) เพื่อใช้เป็นฟังก์ชั่นความผิดพลาด (Error function) สำหรับโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ โดยขึ้นอยู่กับอัตราความไม่สมดุล อัตราการซ้อนทับกันของข้อมูลแต่ละชุดที่แตกต่างกัน ชุดข้อมูลที่นำมาใช้เป็นชุดข้อมูลมาตรฐานดาวน์โหลดจาก UCI Machine Learning Repository เครื่องมือที่ใช้ในการวัดประสิทธิภาพได้แก่ ค่า True Positive Rate (TPR) ค่า G-Mean และค่า F

ผลการทดลองแสดงให้เห็นว่าวิธีการที่นำเสนอให้ประสิทธิภาพในการแบ่งกลุ่มข้อมูลดีกว่าวิธีการที่ใช้ค่าความผิดพลาดเฉลี่ยกำลังสองที่เป็นแบบมาตรฐานในทั้งสามตัววัดนั่นคือค่า TPR ค่า G-Mean และค่า F

5.3 วิธีการแบบผสมผสมสำหรับชุดข้อมูลไม่สมดุลตามสภาพพื้นที่

สำหรับวิธีการที่นำเสนอในหัวนี้คือ วิธีการแบบผสม (Soft-Hybrid) สำหรับข้อมูลไม่สมดุลตามสภาพพื้นที่ โดยวิธีการได้คำนึงถึงลักษณะของข้อมูลที่ไม่สมดุลบนพื้นที่แต่ละบริเวณ เนื่องจากความยากง่ายในการจำแนกข้อมูลและวิธีการที่เหมาะสมในการจำแนกข้อมูลที่ต่างกัน โดยเฉพาะในพื้นที่ที่ซ้อนทับกันของข้อมูลที่มีความยาก การแก้ปัญหาต้องใช้วิธีที่ซับซ้อนและจะส่งผลต่อเวลาที่ใช้ในการทำงานด้วย ดังนั้นวิธีการที่นำเสนอจึงได้ทำการแบ่งข้อมูลของทั้งสองคลาสออกเป็น 3 พื้นที่ ได้แก่ พื้นที่ซ้อนทับ พื้นที่ขอบ และพื้นที่ไม่ซ้อนทับ และในกระบวนการจำแนกประเภทข้อมูลได้ใช้ขั้นตอนวิธีที่เหมาะสมกับข้อมูลในแต่ละพื้นที่ ซึ่งผลการทดลองแสดงให้เห็นว่าวิธีการที่นำเสนอให้ผลการจำแนกข้อมูลในชุดข้อมูลส่วนใหญ่ให้ผลได้ดีกว่าวิธีการอื่นๆ ที่นำมาเปรียบเทียบทั้งตัววัด TPR, ค่า F ของคลาสส่วนน้อย และค่า G-mean อีกทั้งเวลาที่ใช้ในการประมวลผลของวิธีการที่นำเสนอใช้เวลาน้อยกว่าอีกด้วย

อย่างไรก็ตามเนื่องจากขอบเขตที่ได้ทำการศึกษาเป็นปัญหาของข้อมูลไม่สมดุลที่มีเพียงสองคลาส แต่ข้อมูลที่มีอยู่เป็นข้อมูลที่ได้หลายคลาส ซึ่งวิธีการที่นำเสนอหากนำไปประยุกต์ใช้ได้กับข้อมูลที่มีหลายคลาสก็สามารถทำได้ หากจะต้องมีการปรับปรุงพารามิเตอร์บางอย่างเพื่อให้ได้ผลดีขึ้นซึ่งจะได้ทำการศึกษาต่อไป

5.4 ปัญหาและข้อเสนอแนะ

zในงานวิจัยนี้ได้นำเสนอแนวทางในการแก้ปัญหาการจำแนกข้อมูลที่มีความไม่สมดุลสูงภายใต้ข้อกำหนดของขอบเขตความไม่สมดุลของข้อมูลและอัตราการซ้อนทับกันของคลาส แต่อาจมีคุณลักษณะหรือปัจจัยอื่นที่ส่งผลต่อประสิทธิภาพการแบ่งกลุ่มข้อมูลอีกเช่น ลักษณะการกระจายของข้อมูลในแต่ละคลาส หรือจำนวนข้อมูลที่มีขนาดใหญ่ขึ้นอาจนำมาพิจารณาร่วมด้วย

บรรณานุกรม

Alejo, R., Valdovinos, R. M., García, V., & Pacheco-Sanchez, J. H. (2013). “A hybrid method to face class overlap and class imbalance on neural networks and multi-class scenarios”. *Pattern Recognition Letters*, *34*(4), 380-388.

Asrul Adam, Ibrahim Shapiai, Zuwairie Ibrahim, Marzuki Khalid, Lim Chun Chew, Lee Wen Jau, and JunzoWatana (2010), “A Modified Artificial Neural Network Learning Algorithm for Imbalanced Data Set Problem”, Proceedings of the 2nd International Conference on Computational Intelligence, Communication Systems and Networks, pp. 44-48.

Chumphol Bunkhumpornpat, Krung Sinapiromsaran and Chidchanok Lursinsap (2009), “Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem”, ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING, Lecture Notes in Computer Science, 2009, Vol. 5476/2009, pp. 475-482.

Chumphol Bunkhumpornpat, Krung Sinapiromsaran and Chidchanok Lursinsap (2011), “DBSMOTE: Density-Based Synthetic Minority Over-Sampling Technique”, APPLIED INTELLIGENCE, Publish Online: 14 April, 2011. (Springer)

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, pp. 226-231).

Fan, X., & He, Z. (2010, November). A fuzzy support vector machine for imbalanced data classification. In *2010 International Conference on Optoelectronics and Image Processing (ICOIP)* (Vol. 1, pp. 11-14). IEEE. doi:10.1109/ICOIP.2010.61

Haibo He and Edwardo A. Garcia (2009), “Learning from Imbalanced Data”, IEEE Transaction on Knowledge and Data Engineering, Vol. 21, No. 9, pp. 1263-1284.

He, H., & Ghodsi, A. (2010, August). Rare class classification by support vector machine. In *2010 20th International Conference on Pattern Recognition (ICPR),* (pp. 548-551). IEEE.

Hamid Parvin, Behrouz Minaei-Bidgoli, and Hosein Alizadeh (2011), “Iranian Cancer Patient Detection Using a New Method for Learning at Imbalanced Datasets”, INTELLIGENT DATA ENGINEERING AND AUTOMATED LEARNING - IDEAL 2011 Lecture Notes in Computer Science, 2011, Vol. 6936/2011, pp. 299-306.

Hausdorff, F. Grundzge der mengenlehre, Von Veit, Leipzig, 1914.

Hui Han, Wen-Yuan Wang, and Bing-Huan Mao (2005), “Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning”, ADVANCES IN INTELLIGENT COMPUTING Lecture Notes in Computer Science, 2005, Volume 3644/2005, pp.878-887.

Hwang, J. P., Park, S., & Kim, E. (2011). A new weighted approach to imbalanced data classification problem via support vector machine with quadratic cost function. *Expert Systems with Applications*, *38*(7), 8580-8585.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 79-86.

Lin, S. J., Chang, C., & Hsu, M. F. (2013). Multiple extreme learning machines for a two-class imbalance corporate life cycle prediction. *Knowledge-Based Systems*, *39*, 214-223.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. In *Proceedings of the National Institute of Sciences (Calcutta)*, *2*, 49-55.

Mohd.Faheem Khan, Gaurav Chauhan and A. K. Jaitly (2011), "An approach to overcome imbalance datasets of eukaryotic genomes during the analysis by machine learning technique (SVM)", Indian Journal of Science and Technology, Vol. 4, No. 5, pp. 520-524.

Napierala, K., & Stefanowski, J. (2012). BRACID: a comprehensive approach to learning rules from imbalanced data. *Journal of Intelligent Information Systems*, *39*(2), 335-373. doi:10.1007/s10844-011-0193-0

Satyam Maheshwari, Jitendra Agrawal, and Sanjeev Sharma (2011), “A New approach for Classification of Highly Imbalanced Datasets using Evolutionary Algorithms”, International Journal of Scientific & Engineering Research, Vol. 2, Issue 7, pp.1-5

SireeshaRodda and ShashiMogalla (2011), “A Normalized Measure for Estimating Classification Rules for Multi-class Imbalanced Datasets”, International Journal of Engineering Science and Technology, Vol. 3, No. 4, April, 2011, pp. 3216-3220.

Si Chen, GongdeGuo and Lifei Chen (2010), “A New Over-Sampling Method Based on Clustering Ensembles”, Proceedings of the 24th International Conference on Advanced Information Networking and Applications Workshops, pp.599-604.

VicencSoler, Jesus Cerquides, JosepSabria, JordiRoig, Marta Prim (2006), “Imbalanced Datasets Classification by Fuzzy Rule Extraction and Genetic Algorithms”, Proceeding of ICDMW '06 Proceedings of the Sixth IEEE International Conference on Data Mining – Workshops.

Vorraboot, P., Rasmequan, S., Lursinsap, C., & Chinnasarn, K. (2012, July). Feature selection for imbalanced datasets with hybrid approach. In *2012 4 th Knowledge and Smart Technology (KST2012)* (pp. 61-68). (In Thai)

Vorraboot, P., Rasmequan, S., Lursinsap, C., & Chinnasarn, K. (2012, December). A modified error function for imbalanced dataset classification problem. In *2012 7th International Conference on Computing and Convergence Technology (ICCCT)* (pp. 854-859). IEEE.

Vorraboot, P., Rasmequan, S., Chinnasarn, K., & Lursinsap, C., (2014). Improving Classification Rate Constrained to Imbalanced Data Between Overlapped and Non-overlapped Regions by Hybrid Algorithms. *Neurocomputing.* (In Press)

Zhi-QiangZeng and JiGao (2009), “Improving SVM Classification with Imbalance Data Set”, NEURAL INFORMATION PROCESSING, Lecture Notes in Computer Science, 2009, Volume 5863/2009, pp.389-398.

ภาคผนวก

การตีพิมพ์ผลงานวิจัย


KST
Research Center
Knowledge Smart Technologies
Since 2008
มหาวิทยาลัยบูรพา
BURAPHA UNIVERSITY
มหาวิทยาลัยศรีปทุม
SRIPATUM UNIVERSITY
ASSUMPTION UNIVERSITY OF THAILAND
เอกสารประกอบการประชุมวิชาการ
Knowledge and Smart Technology
(KST-2012) ครั้งที่ ๔
July 7-8, 2012
UniNet
NECTEC
a member of NSTDA
IEEE
THAILAND SECTION

2012 4th Conference on Knowledge and Smart Technology (KST)

การเลือกคุณลักษณะข้อมูลไม่สมดุลด้วยวิธีการแบบผสม
Feature Selection for Imbalanced Datasets with Hybrid Approach

ปิยนุช วรบุตร[1] สุวรรณา รัศมีขวัญ[1] กฤษณะ ชินสาร[1] และชิดชนก เหลือสินทรัพย์[2]
[1]คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา ชลบุรี 20130
[2]ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
กรุงเทพมหานคร 10300
E-mail: piyanoot_v@yahoo.com

บทคัดย่อ

คุณลักษณะข้อมูลเป็นสิ่งที่บอกถึงสารสนเทศที่ซ่อนอยู่ของข้อมูลแต่ละชุด โดยทั่วไปข้อมูลที่มีสารสนเทศมากจะดีกว่าข้อมูลที่มีสารสนเทศน้อยนั่นคือสามารถให้รายละเอียดข้อมูลที่เราสนใจได้มากกว่า แต่ในบางปัญหาของการจำแนกข้อมูลข้อมูลที่มีรายละเอียดมากเกินไปอาจส่งผลให้ประสิทธิภาพในการจำแนกข้อมูลลดลงเนื่องจากสารสนเทศที่ถูกซ่อนอยู่ในบางคุณลักษณะรบกวนต่ออัตราการรู้จำได้ ในงานวิจัยนี้ได้นำเสนอวิธีการแบบผสมในการเลือกคุณลักษณะข้อมูลไม่สมดุลที่ประกอบด้วยคลาสส่วนมากและคลาสส่วนน้อย ซึ่งปัญหาของข้อมูลไม่สมดุลเกิดจากจำนวนข้อมูลของคลาสส่วนมากมีจำนวนมากกว่าคลาสส่วนน้อย วิธีการที่นำเสนอประกอบไปด้วย 3 ขั้นตอนได้แก่ การเลือกข้อมูลคลาสส่วนมาก การกำหนดจำนวนคุณลักษณะด้วยค่าไอเกน และการเลือกคุณลักษณะด้วยการประยุกต์ใช้ขั้นตอนวิธีเชิงพันธุกรรมร่วมกับโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ ในการทดลองใช้ชุดข้อมูลไม่สมดุลจาก UCI และทำการเปรียบเทียบประสิทธิภาพด้วยตัวจำแนกข้อมูล 3 วิธี ได้แก่ โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ โครงข่ายฟังก์ชันรัศมีฐาน และซัพพอร์ตเวกเตอร์แมชชีน ผลการทดลองแสดงให้เห็นว่าตัวจำแนกข้อมูลทั้ง 3 วิธี ให้ค่าความถูกต้อง (Accuracy) ค่าเฉลี่ยเรขาคณิต (G-Mean) โดยในการจำแนกชุดข้อมูลที่ใช้วิธีการที่นำเสนอ พบว่าโดยส่วนมากการจำแนกให้ผลลัพธ์ดีกว่าชุดข้อมูลที่ไม่ได้ใช้วิธีการที่นำเสนอ นอกจากนี้ยังทำให้ค่าความถูกต้องของการจำแนกข้อมูลคลาสส่วนน้อยมีความถูกต้องมากขึ้น และเวลาที่ใช้ในการประมวลผลลดลงเนื่องจากจำนวนคุณลักษณะและจำนวนข้อมูลของคลาสส่วนมากลดลง

คำสำคัญ: ข้อมูลไม่สมดุล การจำแนกข้อมูล การเลือกคุณลักษณะ

Abstract

Data features or attributes can be used to identify hidden information of each data set. In general, more information is better. In other words, more details of information can better clarify things. However, in some classification problems, more information may decrease level of accuracy. This is because hidden information in some characteristics will interfere recognition rate. In this paper, hybrid method of feature selection in imbalanced dataset, consisted of majority class and minority class, is proposed. It has been known that in imbalanced problem, number of instants in majority class is larger than minority class. Our proposed method consists of three main steps, which are majority class instant selection (sampling), feature clustering using eigen-value, and feature selection using GA and BPNN. Datasets used in this paper are downloaded from UCI. We compare the selected features obtained from our proposed method by three standard learning algorithms which are BPNN, RBF, and SVM. Experimental results, based on Accuracy and G-Mean values, confirm that our hybrid method provide better results than the recognition using standard dataset. In addition, accuracy of classification in minority class is increase and the processing time is decrease. Because of the number of features and instant of majority are decreases.

Key Words: Imbalanced dataset, Classification, Feature selection

1. บทนำ

การเลือกคุณลักษณะ (Feature Selection) เป็นวิธีการหนึ่งของขั้นตอนการเตรียมข้อมูลก่อนนำไปใช้งานซึ่งทำให้ข้อมูลมีขนาดลดลง แต่สูญเสียลักษณะสำคัญของข้อมูลน้อยที่สุดและยังคงความถูกต้องของผลลัพธ์ [5] ซึ่งในการจำแนกข้อมูลทั่วไปนั้น จำนวนของคุณลักษณะหรือมิติที่มีจำนวนมากนั้นไม่ได้ทำให้ความถูกต้องของการจำแนกข้อมูลสูงตามไปด้วย ซึ่งในบางกรณีทำให้ขั้นตอนวิธีในการเรียนรู้ข้อมูลที่มีประสิทธิภาพทั้งในด้านความเร็วและความถูกต้องนั้นทำงานลดลง ดังนั้นขั้นตอนการเลือกคุณลักษณะนี้จึงเป็นเครื่องมือที่ใช้ในการเตรียมข้อมูลให้มีความถูกต้องและมีความสำคัญ วิธีการเลือกคุณลักษณะที่ดีนั้นจะต้องสามารถเพิ่มประสิทธิภาพของตัวจำแนกข้อมูลและความถูกต้องในการจำแนกข้อมูลด้วย

สำหรับการจำแนกข้อมูลไม่สมดุลนั้นพบว่าความถูกต้องในการจำแนกข้อมูลมีความเอนเอียง เนื่องจากในการเรียนรู้ข้อมูลได้ให้ความสำคัญกับข้อมูลกลุ่มใดกลุ่มหนึ่งมากเกินไป ทำให้ขอบเขตการตัดสินใจของตัวจำแนกข้อมูล (Classifiers) ที่ได้นั้นเกิดความเอนเอียงไปทางกลุ่มของคลาสส่วนมาก (Majority class) ดังนั้นคลาสส่วนน้อย (Minority class) จึงเกิดความผิดพลาดมากกว่า ตัวอย่างของข้อมูลไม่สมดุลสามารถพบได้ทั่วไป ซึ่งสาเหตุของความไม่สมดุลเกิดจากธรรมชาติของข้อมูลเอง เช่น ข้อมูลทางด้านการแพทย์ ข้อมูลตรวจจับความผิดปกติ ข้อมูลที่ไม่ปกติมีจำนวนน้อยมากเมื่อเทียบกับข้อมูลปกติ และอีกสาเหตุหนึ่งคือข้อจำกัดในการจัดเก็บข้อมูล ไม่ว่าจะเป็นค่าใช้จ่ายหรืออันตรายที่เกิดขึ้นในกระบวนการรวบรวมข้อมูล สำหรับปัญหาของข้อมูลไม่สมดุลที่น่าสนใจอย่างหนึ่งคือ ปัญหาการเลือกคุณลักษณะของข้อมูลที่มีคุณลักษณะหรือมิติข้อมูลที่สูง ซึ่งคุณลักษณะที่มีมากนั้นอาจมีบางส่วนที่ซ้ำซ้อน หรือไม่ส่งผลต่อประสิทธิภาพในการทำงานของตัวจำแนกข้อมูล ซึ่งการเลือกคุณลักษณะโดยใช้เทคนิคการเลือกที่ดีจะทำให้ได้ข้อมูลที่สามารถใช้เป็นตัวแทนของข้อมูลส่วนใหญ่ได้ เทคนิคในการเลือกคุณลักษณะเป็นที่นิยมมีหลายวิธีด้วยกัน เช่น เทคนิคการวิเคราะห์สหสัมพันธ์ เทคนิคการหาค่าไคสแคว์ เทคนิคการหาเพื่อนบ้านใกล้สุด [1] [5]

งานวิจัยนี้ได้นำเสนอวิธีการเลือกคุณลักษณะที่มีความสำคัญสำหรับข้อมูลไม่สมดุลที่มีจำนวนมิติข้อมูลที่สูง เพื่อเพิ่มประสิทธิภาพของการจำแนกข้อมูลให้มีความถูกต้องมากขึ้นด้วยวิธีการแบบผสมซึ่งประกอบไปด้วย 3 ขั้นตอน ได้แก่ การเลือกข้อมูลคลาสส่วนมาก การกำหนดจำนวนคุณลักษณะด้วยค่าไอเกน และการเลือกคุณลักษณะด้วยการประยุกต์ใช้ขั้นตอนวิธีเชิงพันธุกรรมร่วมกับโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ สำหรับการเลือกข้อมูลบางส่วนของคลาสส่วนมากเนื่องจากลักษณะข้อมูลไม่สมดุลที่จำนวนข้อมูลในแต่ละคลาสแตกต่างกันมากทำให้เกิดความเอนเอียงในการเรียนรู้งานวิจัยนี้จึงได้ทำการเลือกคลาสส่วนมากให้เท่ากับหรือใกล้เคียงกับคลาสส่วนน้อยด้วยวิธีการหาศูนย์กลางของกลุ่มคลาสส่วนมากและเลือกเฉพาะข้อมูลที่อยู่ใกล้ศูนย์กลางของแต่ละกลุ่ม จากนั้นกำหนดจำนวนคุณลักษณะทำให้จำนวนกลุ่มย่อยของคุณลักษณะที่เป็นไปได้มีจำนวนลดลงเพื่อลดเวลาในการค้นหา และในการเลือกคุณลักษณะได้ประยุกต์ใช้ขั้นตอนวิธีเชิงพันธุกรรมซึ่งเป็นการค้นหาคำตอบแบบสุ่มไม่ต้องทำการเปรียบเทียบทุกคำตอบที่เป็นไปได้ทำให้ค้นหาได้เร็วขึ้น ส่วนค่าความเหมาะสมที่ใช้ในงานวิจัยนี้คือค่าเฉลี่ยของความคลาดเคลื่อนยกกำลังสองจากกระบวนการเรียนรู้ของโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ ในการทดลองได้ทำการเปรียบเทียบประสิทธิภาพการทำงานของตัวจำแนกข้อมูล3 วิธีได้แก่ โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ โครงข่ายฟังก์ชันรัศมีฐาน และซัพพอร์ตเวกเตอร์แมชชีน ระหว่างชุดข้อมูลที่ใช้วิธีการแบบผสมกับข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง พบว่าแต่ละชุดข้อมูลมีความเหมาะสมกับตัวจำแนกข้อมูลที่แตกต่างกัน โดยค่าความถูกต้อง (Accuracy) รวมถึงค่าเฉลี่ยเรขาคณิต (G-Mean) ของวิธีที่นำเสนอโดยโดยส่วนมากให้ผลที่ดีกว่า และทำให้ค่าความถูกต้องของการจำแนกข้อมูลคลาสส่วนน้อยมีความถูกต้องมากขึ้น รวมถึงเวลาที่ใช้ในการจำแนกข้อมูลลดลงด้วย

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ข้อมูลไม่สมดุล (Imbalanced Data)

ข้อมูลไม่สมดุลประกอบไปด้วยจำนวนข้อมูลในแต่ละคลาสแตกต่างกันอย่างมาก โดยในปัญหาการจำแนกข้อมูลแบบสองคลาส (Two-classes classification) จะแบ่งข้อมูลออกเป็นคลาสส่วนมาก และคลาสส่วนน้อย [7][8] ตัวอย่างเช่นข้อมูลทั้งหมดจำนวน 1000 ตัวอย่าง แบ่งเป็นข้อมูลของคลาสส่วนมาก จำนวน 990 ตัวอย่าง คลาสส่วนน้อย จำนวน 10 ตัวอย่าง ซึ่งข้อมูลลักษณะนี้เมื่อนำไปผ่านกระบวนการเรียนรู้และจำแนกข้อมูลมักจะทำให้โมเดลในการเรียนรู้และจำแนกข้อมูลที่ไม่เป็นกลาง และในข้อมูลแบบไม่สมดุลที่มีมิติข้อมูลหรือคุณลักษณะที่สูงเป็นอีกปัญหาหนึ่งที่ส่งผลต่อประสิทธิภาพของการจำแนกข้อมูล เนื่องจาก

บางคุณลักษณะมีความซ้ำซ้อน หรือทำให้ผลที่ได้มีความผิดพลาดเกิดขึ้นได้ มีงานวิจัยที่ทำศึกษาถึงวิธีการเลือกคุณลักษณะของข้อมูลไม่สมดุลไว้หลายงาน เช่น Mina Alibeigi และคณะ [4] ได้นำเสนอวิธีการเลือกคุณลักษณะโดยใช้การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) สำหรับข้อมูลไม่สมดุล โดยทำการตัดคุณลักษณะที่ซ้ำซ้อน (Redundant) ออกบนพื้นฐานของการกระจายตัวของคุณลักษณะ วิธีการที่นำเสนอได้ทำการค้นหาความสัมพันธ์ระหว่างสองคุณลักษณะที่ว่ามีความคล้ายคลึงกันหรือไม่จากค่า PDFs Tian-Yu Liu [9] ได้นำเสนอ MIEE (Mutual Information based feature selection for EasyEnsemble) เป็นการทำ under-sampling โดยใช้ขั้นตอนวิธี EasyEnsemble จากนั้นทำการเลือกคุณลักษณะด้วยการพิจารณาจากค่า Mutual Information German Cuaya และคณะ [3] ได้นำเสนอวิธีการเลือกคุณลักษณะที่เรียกว่า FSMC เป็นการเลือกคุณลักษณะจากการพิจารณาคลาสส่วนน้อยที่มีความแตกต่างจากคลาสส่วนมากจากค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน วิธีนี้เป็นวิธีที่ง่ายและทำงานได้เร็ว และให้ประสิทธิภาพในการวัดสำหรับความถูกต้องของคลาสส่วนน้อย เช่น precision, recall, F-measure และ ROC ได้ดี

2.2 การเลือกคุณลักษณะ (Feature Selection)

คุณลักษณะของข้อมูล หมายถึงลักษณะหรือคุณสมบัติที่ใช้ระบุองค์ประกอบหรือรายละเอียดของชุดข้อมูล ซึ่งคุณลักษณะของข้อมูลที่ดีต้องมีความถูกต้องและเชื่อถือได้ คุณลักษณะของข้อมูลที่พบทั่วไปมีทั้งคุณลักษณะที่เกี่ยวข้อง คุณลักษณะที่ไม่เกี่ยวข้อง รวมถึงคุณลักษณะที่ซ้ำซ้อน หากนำคุณลักษณะทั้งหมดมาใช้งาน อาจส่งผลต่อความถูกต้องในการจำแนกข้อมูลได้ รวมถึงชุดข้อมูลมีการเก็บคุณลักษณะที่มากเกินความจำเป็นจะทำให้สิ้นเปลืองทรัพยากรและเวลาในการประมวลผลอีกด้วย ดังนั้นการเลือกคุณลักษณะที่เหมาะสมจะทำให้ขนาดข้อมูลและเวลาที่ใช้ในการวิเคราะห์ข้อมูลลดลง [2] ได้ข้อมูลที่มีคุณภาพเข้าใจง่าย ส่งผลต่อความถูกต้องความแม่นยำในการทำงาน

วิธีการเลือกคุณลักษณะโดยทั่วไปแบ่งออกเป็นสองวิธี ได้แก่ วิธีการกรอง (Filter selection) ซึ่งการเลือกคุณลักษณะจะเป็นอิสระจากขั้นตอนวิธีในการจำแนกข้อมูล และวิธีการห่อหุ้ม (Wrapper selection) วิธีนี้จะนำขั้นตอนวิธีในการจำแนกข้อมูลมาใช้เป็นส่วนหนึ่งของกระบวนการเลือกคุณลักษณะด้วย

วิธีการประเมินคุณลักษณะมีหลายวิธีด้วยกัน โดยในงานวิจัยนี้ประกอบไปด้วย 2 ขั้นตอนหลัก ได้แก่ การคำนวณหาจำนวนคุณลักษณะโดยใช้ค่าไอเกน และการเลือกคุณลักษณะโดยขั้นตอนวิธีเชิงพันธุกรรมร่วมกับโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ รายละเอียดมีดังนี้

2.2.1 การคำนวณหาจำนวนคุณลักษณะโดยใช้ค่าไอเกน (Eigen values)

ค่าไอเกนเป็นค่าความผันแปรของตัวแปรทั้งหมดในแต่ละองค์ประกอบ ซึ่งสามารถคำนวณได้ดังสมการที่ 1

$$AX = \lambda X \quad (1)$$

โดย A แทนเมตริกซ์ขนาด nxn

X แทนเวกเตอร์ขนาด nx1

λ แทนสเกลาร์ เรียกว่าค่าไอเกน

2.2.2 การเลือกคุณลักษณะโดยขั้นตอนวิธีเชิงพันธุกรรมร่วมกับโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ

- ขั้นตอนวิธีเชิงพันธุกรรม

ในการเลือกคุณลักษณะเป็นการหากลุ่มย่อยของคุณลักษณะที่เหมาะสมที่สุด ดังนั้นข้อมูลมีจำนวนคุณลักษณะมากจะมีจำนวนกลุ่มย่อยของคุณลักษณะมากตามไปด้วย ในงานวิจัยนี้ได้ประยุกต์ใช้ขั้นตอนวิธีเชิงพันธุกรรมในการเลือกคุณลักษณะ เนื่องจากขั้นตอนวิธีนี้เป็นการค้นหาคำตอบแบบสุ่มไม่ต้องทำการเปรียบเทียบทุกคำตอบที่เป็นไปได้ จึงช่วยลดเวลาในการทำงานลงได้ ซึ่งขั้นตอนวิธีเชิงพันธุกรรมเป็นวิธีการแก้ปัญหาที่ใช้แนวทางเดียวกับวิธีการที่สิ่งมีชีวิตปรับตัวเองหรือวิวัฒนาการให้เข้ากับสภาพแวดล้อม ในการถ่ายทอดลักษณะทางพันธุกรรมจะมีกระบวนการที่ทำให้เกิดการเปลี่ยนแปลงที่เรียกว่ากระบวนการวิวัฒนาการ ได้แก่ กระบวนการเลือก (Selection) การไขว้เปลี่ยน (Crossover) และกลายพันธุ์ (Mutation) โดยพิจารณาจากค่าความเหมาะสม (Fitness Function) ที่สอดคล้องกับวัตถุประสงค์ของปัญหา (Objective Function) ของโครโมโซมแต่ละตัวเพื่อนำไปสู่กระบวนการคัดเลือก ค่าความเหมาะสมที่ใช้ในงานวิจัยนี้คือค่าเฉลี่ยของความคลาดเคลื่อนยกกำลังสอง (Mean Square Error: MSE) จากกระบวนการเรียนรู้ของโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ เป็นค่าที่ได้จากการหาค่าเฉลี่ยของผลต่างระหว่างค่าที่ได้จากการพยากรณ์กับค่าจริงยกกำลังสองตามสมการที่ 2

$$E = \frac{1}{2}\sum_{p=1}^{P}\sum_{k=1}^{M}\left(t_k^{(p)} - y_k^{(p)}\right)^2 \quad (2)$$

- โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ

โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ เป็นอัลกอริทึมที่ออกแบบมาโดยใช้เทคนิคการเรียนรู้ของโครงข่ายประสาทเทียมวิธีหนึ่งที่ใช้ในเพอร์เซพตรอนหลายชั้น ซึ่งโดยทั่วไปจะมีรูปแบบการจัดนิวรอนเป็นชั้น (Layer) ประกอบไปด้วย ชั้นอินพุท (Input Layer) การทำงานของอินพุท จะทำหน้าที่รับข้อมูลเข้าสู่โครงข่าย ชั้นซ่อน (Hidden Layer) ทำหน้าที่ช่วยในการประมวลผลซึ่งในการทำงานของแต่ละชั้นจะถูกกำหนดโดยการทำงานของชั้นอินพุท ค่าถ่วงน้ำหนัก (Weight) และค่าไบแอส (Bias) บนความสัมพันธ์ระหว่างชั้นอินพุทและชั้นซ่อน ชั้นเอาท์พุท (Output Layer) ทำหน้าที่ผลิตผลลัพธ์ของโครงข่าย

ข้อมูลนำเข้า $x_1(p)$, $x_2(p)$,..., $x_n(p)$ โดย n เป็นจำนวนมิติข้อมูลที่ได้จากการขั้นตอนการคำนวณค่าลักษณะเฉพาะ และข้อมูลออกได้แก่ $y_{d,1}(p)$, $y_{d,2}(p)$,..., $y_{d,n}(p)$ เมื่อ d เป็นจำนวนคลาส

- คำนวณผลลัพธ์ในชั้นซ่อนตามสมการที่ 3.1

$$y_i(p) = sigmoid\left[\sum_{i=1}^{n} x_i(p)\cdot w_{ij}(p) - \theta_j\right] \quad (3.1)$$

- คำนวณผลลัพธ์ในชั้นเอาท์พุท โดยสมการที่ 3.2

$$y_k(p) = sigmoid\left[\sum_{j=1}^{m} x_{ij}(p)\cdot w_{jk}(p) - \theta_k\right] \quad (3.2)$$

m เป็นจำนวนนิวรอนในชั้นเอาท์พุท

- ทำการปรับปรุงน้ำหนักในชั้นเอาท์พุทโดยสมการที่ 3.3 – 3.6

$$\delta_k(p) = y_k(p)\cdot[1 - y_k(p)]\cdot e_k(p) \quad (3.3)$$

$$e_k(p) = y_{d,k}(p) - y_k(p) \quad (3.4)$$

$$\Delta w_{jk}(p) = \alpha \cdot y_j(p)\cdot \delta_k(p) \quad (3.5)$$

$$w_{jk}(p+1) = w_{jk}(p) + \Delta w_{jk}(p) \quad (3.6)$$

- ทำการปรับปรุงน้ำหนักในชั้นซ่อนโดยสมการที่ 3.7 – 3.9

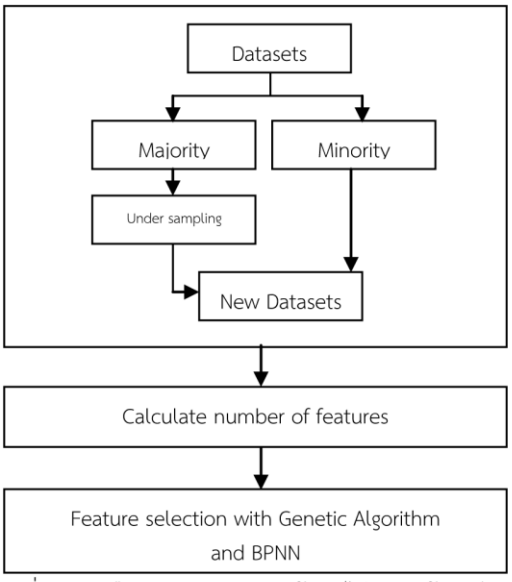
$$(p) = y_j(p)\cdot\left[1 - y_j(p)\right]\cdot\sum_{k=1}^{l}\delta_k(p)\cdot w_{jk}(p) \quad (3.7)$$

$$\Delta w_{ij}(p) = \alpha \cdot x_j(p)\cdot \delta_j(p) \quad (3.8)$$

$$w_{ij}(p+1) = w_{ij}(p) + \Delta w_{ij}(p) \quad (3.9)$$

3. วิธีการที่นำเสนอ

เนื่องจากข้อมูลที่นำมาใช้เป็นข้อมูลไม่สมดุลที่ประกอบไปด้วยจำนวนข้อมูลในแต่ละคลาสแตกต่างกันมากทำให้เกิดความเอนเอียงในการเรียนรู้ ประกอบกับจำนวนคุณลักษณะข้อมูลที่มีมากบางคุณลักษณะอาจทำให้ประสิทธิภาพในการจำแนกข้อมูลลดลง ดังนั้นงานวิจัยนี้ได้นำเสนอการเลือกคุณลักษณะด้วยวิธีการแบบผสมโดยประกอบไปด้วย 3 ขั้นตอนดังแสดงในรูปที่ 1 มีรายละเอียดดังนี้



รูปที่ 1 การเลือกคุณลักษณะของข้อมูลไม่สมดุลด้วยวิธีการแบบผสม

1. การเลือกข้อมูลคลาสส่วนมาก ขั้นตอนนี้เป็นการลดจำนวนข้อมูลในกลุ่มคลาสส่วนมากให้มีจำนวนให้เท่ากับหรือใกล้เคียงกับจำนวนข้อมูลของคลาสส่วนน้อย โดยนำชุดข้อมูลที่ไม่สมดุลเฉพาะคลาสส่วนมากมาทำการจัดกลุ่มด้วยวิธีการ k-means ซึ่งเป็นการเรียนรู้แบบไม่มีผู้สอน โดยจะแบ่งข้อมูลออกเป็น k กลุ่ม โดยแทนแต่ละกลุ่มด้วยค่าเฉลี่ยของกลุ่มซึ่งใช้เป็นจุดศูนย์กลางของกลุ่มในการวัดระยะห่างของข้อมูลในกลุ่มเดียวกัน โดยจำนวนกลุ่มที่กำหนดในงานวิจัยนี้เท่ากับอัตราส่วนของจำนวนข้อมูลของคลาสส่วนมากกับจำนวนข้อมูลของคลาสส่วนน้อย จากนั้นทำการเลือกข้อมูลคลาสส่วนมากในแต่ละกลุ่มที่ใกล้ศูนย์กลางของกลุ่ม โดยจำนวนที่เลือกคำนวณจากสมการที่ 4 และตัวอย่างการคำนวณแสดงในตารางที่ 1

$$n_i = \frac{N_i}{Nmaj} \times Nmin \quad (4)$$

โดย n_i จำนวนสมาชิกที่เลือกในกลุ่มที่ i
N_i จำนวนสมาชิกในกลุ่มที่ i
Nmaj จำนวนสมาชิกของคลาสส่วนมาก
Nmin จำนวนสมาชิกของคลาสส่วนน้อย

ตารางที่ 1 ตัวอย่างการคำนวนจำนวนข้อมูลคลาสส่วนมากที่ถูกเลือกในแต่ละกลุ่ม *(จำนวน Min.=60)*

กลุ่ม	จำนวน Maj.	จำนวน Maj. ที่เลือก
1	60	18
2	30	9
3	110	33
รวม	200	60

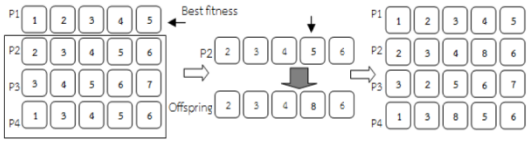
จากนั้นนำข้อมูลคลาสส่วนมากที่เลือกจากทุกกลุ่มมารวมกับข้อมูลคลาสส่วนน้อย จะได้ข้อมูลชุดใหม่ที่มีความสมดุลและนำไปใช้ในขั้นตอนต่อไป

2. การกำหนดจำนวนคุณลักษณะด้วยค่าไอเกน ในการเลือกคุณลักษณะเป็นการเลือกคุณลักษณะย่อยที่เหมาะสมที่สุดจากคุณลักษณะย่อยที่เป็นไปได้ทั้งหมด หากคุณลักษณะย่อยที่เป็นไปได้มีจำนวนมากจะทำให้เวลาในการเลือกมากตามไปด้วย ดังนั้นขั้นตอนนี้จึงได้ทำการกำหนดจำนวนคุณลักษณะ (n) ที่เหมาะสมจากค่าไอเกน ซึ่งคำนวนได้จากสมการที่ 1 โดยพิจารณาจากจำนวนของคุณลักษณะที่มีค่าไอเกนมากกว่าค่าที่กำหนดไว้ ในการทดลองนี้กำหนด 0.01 ซึ่งเป็นค่าที่ได้จากการทดลอง

3. การเลือกคุณลักษณะด้วยการประยุกต์ใช้ขั้นตอนวิธีเชิงพันธุกรรมร่วมกับโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ ในขั้นตอนนี้จะทำการค้นหาคุณลักษณะโดยการประยุกต์ใช้ขั้นตอนวิธีเชิงพันธุกรรม โดยกำหนดประชากรเริ่มต้นจากการสุ่มจำนวน 0.1 เท่าของประชากรที่เป็นไปได้ทั้งหมด จำนวนรอบในการทำงาน 300 รอบ รูปแบบของประชากรกำหนดด้วยเลขจำนวนเต็ม 1,2,3,..,N เมื่อ N เป็นจำนวนคุณลักษณะทั้งหมด เช่น จำนวนคุณลักษณะที่เลือกเท่ากับ 6 คุณลักษณะ ตัวอย่างรูปแบบประชากรที่เป็นไปได้ 1 2 3 4 5 6, 2 3 4 5 6 7 เป็นต้น

3.1 ในการคัดเลือก (Selection) ประชากรที่ให้ค่าความเหมาะสมดีสุดจะถูกเก็บไว้ โดยค่าความเหมาะสม (Fitness value) ได้นำเอาค่าเฉลี่ยของความคลาดเคลื่อนยกกำลังสอง ของประชากรแต่ละตัวจากสมการที่ 2 มาประยุกต์ใช้เป็นค่าความเหมาะสม โดยฟังก์ชันกระตุ้นของโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับใช้ฟังก์ชันซิกมอยด์ทั้งชั้นซ่อนและชั้นเอาท์พุท จำนวนนิวรอนในชั้นซ่อนเท่ากับจำนวนอินพุท และชั้นเอาท์พุทเท่ากับ 2 อัตราการเรียนรู้ 0.1 ทำการเรียนรู้ 100 รอบ

3.2 ประชากรอื่นที่ไม่ถูกคัดเลือกจะถูกนำไปสร้างเป็นประชากรใหม่หรือประชากรลูกหลาน (Offspring) ด้วยกระบวนการกลายพันธุ์ (Mutation) โดยการสุ่มเลือกตำแหน่งของประชากรเพื่อทำการเปลี่ยนรหัสดังรูปที่ 2 กำหนดให้มีการกลายพันธุ์ในทุกๆ รอบ จำนวนตำแหน่งในการกลายพันธุ์เท่ากับ 0.1x(N-n) และทำซ้ำขั้นตอนที่ 3.1-3.2 ไปจนกระทั่งครบรอบการทำงานหรือได้ค่าความเหมาะสมตามที่กำหนด และทำการเลือกประชากรที่ให้ค่าความเหมาะสมดีที่สุดไปใช้ในขั้นตอนการทดสอบต่อไป



รูปที่ 2 กระบวนการกลายพันธุ์ (Mutation)

4. การวัดประสิทธิภาพ

วิธีการวัดประสิทธิภาพการจำแนกข้อมูลโดยทำการเปรียบเทียบระหว่างวิธีการแบบผสมกับข้อมูลเดิมที่ไม่ได้ทำการปรับปรุงใช้วิธีการวัดค่าความถูกต้อง (Accuracy) ดังสมการที่ 5 และค่าเฉลี่ยเรขาคณิต (Geometric Mean: G-Mean) ดังสมการที่ 6 และวัดอัตราส่วนความถูกต้องของคลาสส่วนน้อยดังสมการที่ 7

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (5)$$

$$G-Mean = \sqrt{TPR * TNR} \quad (6)$$

$$TPR = \frac{TP}{TP+FN} \quad (7)$$

$$TNR = \frac{TN}{TN+FP} \quad (8)$$

TPR อัตราส่วนความถูกต้องของคลาสส่วนน้อย
TNR อัตราส่วนความถูกต้องของคลาสส่วนมาก
TP แทนจำนวนข้อมูลคลาสส่วนน้อยที่ถูกต้อง
FP แทนจำนวนข้อมูลคลาสส่วนมากที่ผิด
TN แทนจำนวนข้อมูลคลาสส่วนมากที่ถูกต้อง
FN แทนจำนวนข้อมูลคลาสส่วนน้อยที่ผิด

5. การทดลองและผลการทดลอง

ข้อมูลที่ใช้ในการทดลองเป็นข้อมูลจาก UCI Repository Machine Learning จำนวน 4 ชุด ได้แก่ Pakinsons, Sonar, Liver, Pima ซึ่งทั้งหมดเป็นข้อมูลแบบไม่สมดุล รายละเอียดแสดงดังตารางที่ 2 ประกอบไปด้วยจำนวน

แอททริบิวต์ (Attributes) จำนวนข้อมูล (Instances) จำนวนข้อมูลคลาสส่วนมากกับส่วนน้อย (Maj./Min.) และอัตราส่วนความไม่สมดุล (Imbalanced ratio: IR)

ตารางที่ 2 รายละเอียดของชุดข้อมูลก่อนการทดลอง

Datasets	Attributes	Instances	Maj./Min.	IR
Pakinsons	22	195	147/48	3.06
Sonar	60	208	111/97	1.14
Liver	6	345	200/145	1.38
Pima	8	768	500/268	1.87

นำข้อมูลแต่ละชุดมาลดจำนวนข้อมูลคลาสส่วนมากให้มีจำนวนเท่ากับหรือใกล้เคียงคลาสส่วนน้อย จากนั้นนำข้อมูลใหม่ที่ได้มาหาจำนวนที่เหมาะสมจากจำนวนคุณลักษณะที่มีค่าไอเกน มากกว่า 0.01 และทำการเลือกคุณลักษณะด้วยขั้นตอนวิธีทางพันธุกรรมร่วมกับโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ รายละเอียดดังตารางที่ 3 และ 4 ซึ่งจำนวนคุณลักษณะของข้อมูลทั้ง 4 ชุด มีจำนวน 8 28 3 และ 7 ตามลำดับ

ตารางที่ 3 รายละเอียดของชุดข้อมูลเมื่อใช้การเลือกคุณลักษณะข้อมูลไม่สมดุลด้วยวิธีการแบบผสม

Datasets	Attributes	Instances	Maj./Min.	IR
Pakinsons	8	96	49/48	1.02
Sonar	28	194	97/97	1
Liver	3	490	245/245	1
Pima	7	536	268/268	1

ตารางที่ 4 คุณลักษณะที่ถูกเลือก

Datasets	Attributes
Pakinsons	1 4 5 11 17 18 20 21
Sonar	2 6 7 8 10 11 14 16 20 21 23 24 28 29 31 32 33 38 40 42 44 45 46 47 50 53 59
Liver	1 2 3
Pima	1 2 3 5 6 7 8

จากนั้นนำข้อมูลไปทำการทดสอบแบบ 10-fold cross validation กับตัวจำแนกข้อมูลทั้ง 3 แบบ ทำการเปรียบเทียบระหว่างวิธีการแบบผสมกับข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง โดยกำหนดค่าพารามิเตอร์ดังนี้ 1) โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับฟังก์ชันกระตุ้นใช้ฟังก์ชันซิกมอยด์ ทั้งชั้นซ่อนและชั้นเอาท์พุท อัตราการเรียนรู้ 0.1 โมเมนตัม 0.2 2) โครงข่ายฟังก์ชันรัศมีฐาน ฟังก์ชันกระตุ้นในชั้นซ่อนใช้ฟังก์ชันรัศมีฐาน ส่วนชั้นเอาท์พุทใช้ฟังก์ชันเชิงเส้น 3) ซัพพอร์ตเวกเตอร์แมชชีน ค่า C เท่ากับ 11 ใช้ RBF kernel ผลการทดลองแสดงดังตารางที่ 5

ตารางที่ 5 ค่าความถูกต้อง (Accuracy) และค่าเฉลี่ยเรขาคณิต (G-Mean) (%)

Datasets	BPNN		RBF		SVM	
	G	Acc	G	Acc	G	Acc
Pakinsons						
Original (22)	91.03	91.79	70.17	84.10	65.71	85.12
Proposed (8)	91.75	91.75	85.56	85.56	75.84	76.28
Diff.	0.72	-0.04	15.39	1.46	10.13	-8.84
Sonar						
Original (60)	79.70	80.28	72.27	72.11	80.48	80.76
Proposed (28)	80.40	80.41	75.66	73.71	68.87	69.07
Diff.	0.7	0.13	3.39	1.6	-11.61	-11.69
Liver						
- Original (6)	67.22	69.56	61.65	64.34	60.46	67.53
- Proposed (3)	71.83	72.06	66.35	66.55	65.68	66.20
Diff.	4.61	2.5	4.7	2.21	5.22	-1.33
Pima						
- Original (8)	70.56	72.52	68.52	75.39	65.61	76.43
- Proposed (7)	86.23	86.38	86.14	86.19	83.65	83.95
Diff.	15.67	13.86	17.62	10.8	18.04	7.52

* Original: ข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง
Proposed: วิธีการแบบผสม
Diff.: ความแตกต่างระหว่างวิธีการแบบผสมกับข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง

จากตารางที่ 5 ชุดข้อมูล Pakinsons เมื่อนำมาจำแนกข้อมูลด้วย BPNN พบว่าค่า G-mean เมื่อใช้วิธีการแบบผสมให้ค่ามากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 0.72% ค่า Accuracy น้อยกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 0.04% RBF พบว่าค่า G-mean และค่า Accuracy ของวิธีการแบบผสมให้ค่ามากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 15.39% และ 1.46% SVM พบว่าค่า G-mean ของวิธีการแบบผสมให้ค่ามากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 10.13% ค่า Accuracy น้อยกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 8.84%

ชุดข้อมูล Sonar เมื่อนำมาจำแนกข้อมูลด้วย BPNN พบว่าค่า G-mean และค่า Accuracy เมื่อใช้วิธีการแบบผสมให้ค่ามากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 0.7% และ 0.13% RBF พบว่าค่า G-mean และค่า Accuracy เมื่อใช้วิธีการแบบผสมให้ค่ามากกว่าชุดข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 3.39% และ 1.6% SVM เมื่อใช้วิธีการแบบ

ผสมให้ค่า G-mean และค่า Accuracy น้อยกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 11.61% และ 11.69%

ชุดข้อมูล Liver เมื่อนำมาจำแนกข้อมูลด้วย BPNN พบว่าค่า G-mean และค่า Accuracy เมื่อใช้วิธีการแบบผสมให้ค่ามากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 4.61% และ 2.5% RBF พบว่าค่า G-mean และค่า Accuracy เมื่อใช้วิธีการแบบผสมให้ค่ามากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 4.7% และ 2.21% SVM พบว่าค่า G-mean เมื่อใช้วิธีการแบบผสมให้ค่ามากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 5.22% ค่า Accuracy น้อยกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 1.33%

ชุดข้อมูล Pima เมื่อนำมาจำแนกข้อมูลด้วย BPNN พบว่าค่า G-mean และค่า Accuracy เมื่อใช้วิธีการแบบผสมให้ค่ามากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 15.67% และ 13.86% RBF พบว่าค่า G-mean และค่า Accuracy เมื่อใช้วิธีการแบบผสมให้ค่ามากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 17.62% และ 10.8% SVM พบว่าค่า G-mean และค่า Accuracy เมื่อใช้วิธีการแบบผสมให้ค่ามากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 18.04% และ 7.52%

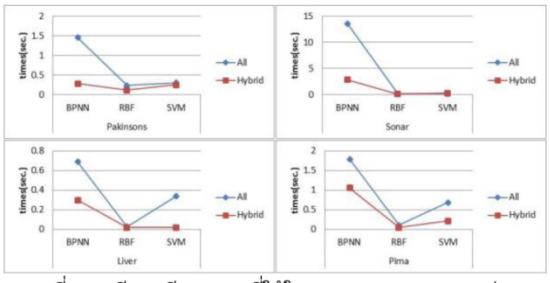
ตารางที่ 6 ค่าความถูกต้องของการจำแนกข้อมูลคลาสส่วนน้อย (%)

Datasets		BPNN	RBF	SVM
Pakinsons	Original	89.58	52.08	43.75
	Proposed	93.75	87.50	68.75
	Diff.	4.17	35.42	25
Sonar	Original	74.22	78.35	77.31
	Proposed	79.38	71.13	74.22
	Diff.	5.16	-7.22	-3.09
Liver	Original	57.93	51.72	42.75
	Proposed	66.20	61.13	57.93
	Diff.	8.27	9.41	15.18
Pima	Original	58.95	54.10	46.64
	Proposed	83.20	83.95	61.30
	Diff.	24.25	29.85	14.66

* Original: ข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง
Proposed: วิธีการแบบผสม
Diff.: ความแตกต่างระหว่างวิธีการแบบผสมกับข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง

จากตารางที่ 6 เมื่อทำการเปรียบเทียบค่าความถูกต้องของการจำแนกข้อมูลคลาสส่วนน้อยเมื่อวิธีการแบบผสมกับข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง พบว่าชุดข้อมูล Pakinsons เมื่อทำการจำแนกข้อมูลด้วย BPNN RBF SVM ชุดข้อมูลที่ใช้วิธีการแบบผสมให้ค่าความถูกต้องของการจำแนกข้อมูลคลาสส่วนน้อยมากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 4.17% 35.42% และ 25% ชุดข้อมูล Sonar เมื่อทำการจำแนกข้อมูลด้วย BPNN ชุดข้อมูลที่ใช้วิธีการแบบผสมให้ค่าความถูกต้องของการจำแนกข้อมูลคลาสส่วนน้อยมากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 5.16% ส่วน RBF และ SVM ให้ค่าความถูกต้องน้อยกว่า 7.22% และ 3.09% ชุดข้อมูล Liver เมื่อทำการจำแนกข้อมูลด้วย BPNN RBF SVM ชุดข้อมูลที่ใช้วิธีการแบบผสมให้ค่าความถูกต้องของการจำแนกข้อมูลคลาสส่วนน้อยมากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 8.27% 9.41% และ 15.18% ชุดข้อมูล Pima เมื่อทำการจำแนกข้อมูลด้วย BPNN RBF SVM ชุดข้อมูลที่ใช้วิธีการแบบผสมให้ค่าความถูกต้องของการจำแนกข้อมูลคลาสส่วนน้อยมากกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 24.25% 29.85% และ 14.66%

สำหรับเวลาที่ใช้ในการประมวลผลในการจำแนกข้อมูลด้วย BPNN RBF SVM ชุดข้อมูล Pakinsons พบว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุงใช้เวลา 1.45 0.23 และ 0.3 วินาที วิธีการแบบผสมใช้เวลา 0.28 0.11 และ 0.25 วินาที ซึ่งน้อยกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 1.17 0.12 และ 0.05 วินาที ชุดข้อมูล Sonar พบว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุงใช้เวลา 13.49 0.14 และ 0.14 วินาที วิธีการแบบผสมใช้เวลา 2.84 0.08 และ 0.05 วินาที ซึ่งน้อยกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 10.65 0.06 และ 0.09 วินาที ตามลำดับ ชุดข้อมูล Liver พบว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุงใช้เวลา 0.69 0.03 และ 0.34 วินาที วิธีการแบบผสมใช้เวลา 0.3 0.02 และ 0.02 วินาที ซึ่งน้อยกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 0.39 0.01 และ 0.32 วินาที ชุดข้อมูล Pima พบว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุงใช้เวลา 1.79 0.11 และ 0.69 วินาที วิธีการแบบผสมใช้เวลา 1.06 0.05 และ 0.22 วินาที ซึ่งน้อยกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง 0.73 0.06 และ 0.47 วินาที ดังรูปที่ 3



รูปที่ 3 เปรียบเทียบเวลาที่ใช้ในการประมวลผลระหว่างวิธีการแบบผสมกับข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง

6. สรุปผลการทดลองและข้อเสนอแนะ

งานวิจัยนี้ได้นำเสนอวิธีการเลือกคุณลักษณะสำหรับข้อมูลไม่สมดุลที่มีจำนวนมิติข้อมูลที่สูง เพื่อลดจำนวนมิติข้อมูลที่ซ้ำซ้อนและเพิ่มประสิทธิภาพของการจำแนกข้อมูลด้วยวิธีการแบบผสม โดยทำการเลือกข้อมูลคลาสส่วนมากให้มีจำนวนเท่ากับหรือใกล้เคียงกับคลาสส่วนน้อยด้วยวิธีการจัดกลุ่ม และนำค่าไอเกนมากำหนดจำนวนคุณลักษณะที่ต้องการ จากนั้นค้นหาคุณลักษณะย่อยด้วยขั้นตอนวิธีทางพันธุกรรมร่วมกับโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ ทำการเปรียบเทียบประสิทธิภาพของวิธีการที่นำเสนอโดยใช้ค่าความถูกต้อง ค่าเฉลี่ยเรขาคณิต และความถูกต้องของการจำแนกข้อมูลคลาสส่วนน้อย รวมถึงเวลาที่ใช้ในการประมวลผล ทำการทดสอบกับ 3 ตัวจำแนกข้อมูลได้แก่ โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ (BPNN) โครงข่ายฟังก์ชันรัศมีฐาน (RBF Network) และซัพพอร์ตเวกเตอร์แมชชีน (SVM) ผลปรากฏว่าตัวจำแนกข้อมูลทั้ง 3 วิธีให้ค่าความถูกต้อง (Accuracy) ค่าเฉลี่ยเรขาคณิต (G-Mean) โดยส่วนมากของการจำแนกชุดข้อมูลของวิธีการแบบผสมดีกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุง รวมถึงค่าความถูกต้องของการจำแนกข้อมูลคลาสส่วนน้อยมีความถูกต้องมากขึ้น โดยแต่ละชุดข้อมูลมีความเหมาะสมกับวิธีการที่ใช้ในการจำแนกข้อมูลที่แตกต่างกัน ในส่วนของเวลาที่ใช้ในการประมวลผลพบว่าชุดข้อมูลทั้งหมดที่ใช้วิธีการแบบผสมใช้เวลาน้อยกว่าข้อมูลเดิมที่ไม่ได้ทำการปรับปรุงในการจำแนกข้อมูลทั้ง 3 วิธี

7. เอกสารอ้างอิง

[1] Abu H. M Kamal, Xingquan Zhu, Abhijit Pandya and Sam Hsu. “Selection with Biased Sample Distributions.” Information Reuse & Integration, 2009. IRI '09. IEEE International Conference on. pp23-28, 2009.

[2] I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection JMLR Special Issue on Variable and Feature Selection”, Kernel Machines Section, Mar, pp. 1157-1182, 2003.

[3] German Cuaya Ang´elica Mu˜noz-Mel´endez, and Eduardo F. Morales, “A Minority Class Feature Selection Method.” Proceeding of CIARP'11 Proceedings of the 16th Iberoamerican Congress conference on Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, pp. 417-424, 2011.

[4] Mina Alibeigi, Sattar Hashemi, and Ali Hamzeh. “Unsupervised Feature Selection Based on the Distribution of Features Attributed to Imbalanced Data Sets.” International Journal of Artificial Intelligence and Expert Systems (IJAE), Vol. 2, Issue 1, pp. 14-22, 2011.

[5] M. Dash and H. Liu, “Feature selection for classification,” Intelligent Data Analysis: An International Journal, 1(3):131-156, 1997.

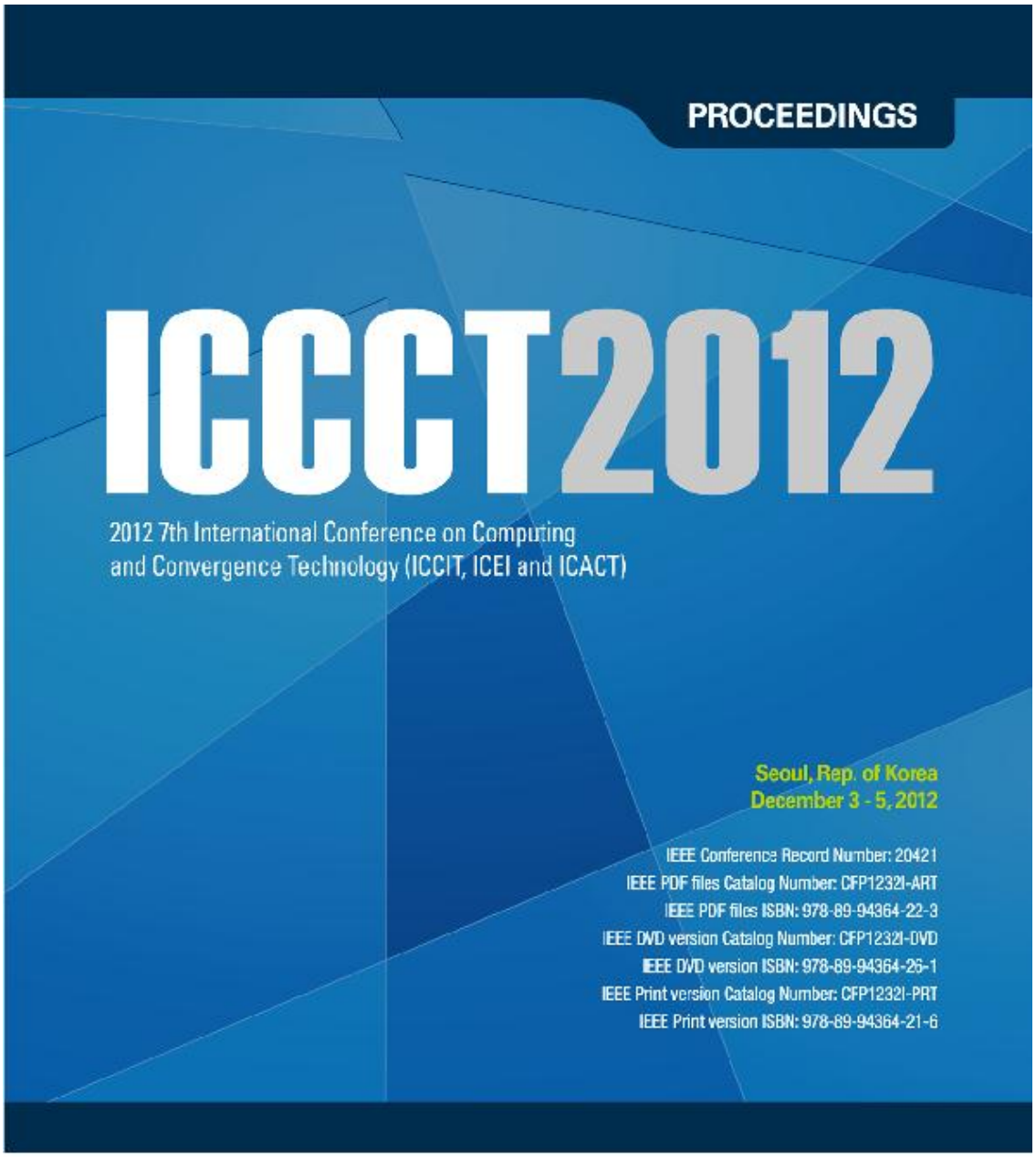
[6] M. Galar, E. Barrenechea, and H. Bustince, “A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches,” IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, ,Issues. 99, pp. 1-22, 2011.

[7] Nitesh V. Chawla, Japkowicz Nathalie, and Kolcz Aleksander, “Editorial: Special Issue on Learning from Imbalanced Data Sets”. Sigkdd Explorations Special Issue on Learning from Imbalanced Datasets, Vol. 6, No.1, pp.1-6, 2004.

[8] N. V. Chawla, K. W. Bowyer, and W. P. Kegelmeyer, “SmoteSynthetic minority over-sampling technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[9] Tian-Yu Liu. “EasyEnsemble and Feature Selection for Imbalance Data Sets.” IJCBS '09 Proceedings of the 2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing, pp. 517-520, 2009.

.



Editors

Dr. Kae Dal Kwack (IEEE Korea Council Chair, Korea)
Prof. Shigeo Kawata (Utsunomiya University, Japan)
Dr. Soonwook Hwang(KISTI, Korea)
Dr. Dongsoo HAN (KAIST, Korea)
Dr. Franz Ko (Dong-A University, Korea/ IBC, UK)

A modified error function for Imbalanced dataset classification problem

Piyanoot Vorraboot
Faculty of Informatics, Burapha University,
Chonburi, Thailand 20131
Email: piyanoot_v@yahoo.com

Suwanna Rasmequan
Faculty of Informatics, Burapha University
Chonburi, Thailand 20131
Email: rsuwanna@buu.ac.th

Chidchanok Lursinsap
Department of mathematics, Faculty of Science,
Chulalongkorn University, Bangkok, Thailand 10330
lchidcha@gmail.com

Krisana Chinnasarn
Faculty of Informatics, Burapha University
Chonburi, Thailand 20131
Email: krisana@buu.ac.th

Abstract **— The objective of learning is to achieve the least error rate. In this paper we proposed a modified cost function as a means to properly measure error rate for imbalanced dataset. Most cost functions apply the same weights to all classes. However, it has been known that for imbalanced problem, the number of instances in the majority class is larger than the minority class. Therefore, the application of equal weight to all classes will significantly lead to improper classification boundary. That is, for most learning model, the minority class would be dominated by majority class which then causes a misclassification on the minority class. The objective of this paper is to find the appropriate parameters to improve MSE cost function based on overlap ratio and class distribution ratio. Back-propagation algorithm with the proposed modified cost function is used to solve two-class classification problem. UCI datasets are used for the experimentation. The results show that the modified MSE cost function provides a better result than the standard one, based on True-positive rate, G-Mean, and F-measurement.**

Index Terms**—imbalanced dataset classification, error function, classification.**

I. Introduction

In recent years, imbalanced dataset classification problems are interested by many researchers who are dealing with the problem of unequal number of data in each class. In two-class classification, there are two classes: majority and minority [1] [2]. The number of instances in majority class is much larger than minority class. This will lead into bias classification. In general, minority class instances will be treated as positive data and majority class instances will be treated as negative data. For example, in case of medical diagnosis, those sick people will be classified in minority class and need attention. However, with the use of traditional classifying algorithm, it has the potential to predict an instance of a positive class as an instance for a negative class. Therefore, in the case of sick people, it may be predicted as normal people. The causes of the imbalanced data can happen in many cases: the nature of the data itself, the high cost of collecting or the high risk in the collection process. Imbalanced dataset can be found in real life, such as an analysis of medical data, a detection of oil spills in satellite radar images, a text categorization, and an intrusion detection.

The solutions to imbalanced data sets can be divided into two levels that are data and algorithmic levels. The main objectives of both levels are to improve the recognition rate of minority class and to reduce bias segmentation. In the data level, the researchers have tried to reduce the imbalanced ratio using different methods. For example, they may implement the techniques on over-sampling, under-sampling or the combination. For the algorithmic level, the researchers have tried to modify the existing algorithms: Support Vector Machine, Neural Network, K-Nearest Neighbor or create new suitable algorithms.

Chawla et al. [2] propose an intelligent oversampling method called "Synthetic Minority Oversampling Technique" (SMOTE). This technique creates new instances up to the amount desired for the minority class using the K-Nearest Neighbors method. Han et al. [3] present a modified version on SMOTE technique which has been called borderline-SMOTE. This method selects minority instances which are on the boundary line and then performs SMOTE to oversample those instances. Si Chen et al. [4] propose a new over-sampling method to manage imbalanced data based on cluster ensembles, named CE-SMOTE. This method wants to provide a better training platform by introducing clustering consistency index to find out the cluster boundary minority samples and then over-sampling these minority samples to increase the original data set. He and Ghodsi [5] propose two modification methods for the soft margin SVM which change or add constraints to the optimization problem. Xiao and Chen [6] propose a new graph classification method based on cost sensitivity to handle imbalance dataset. Boonchuay et al. [7] propose an improved model on decision tree induction algorithm by performing a ternary split on continuous-valued attributes focusing on distribution of minority class instances.

Our research is focused on algorithmic level. Most algorithmic level works attempt to classify the minority class with the lowest error. The measurement method is error or

cost function. If the cost function is unsuitable for dataset, it may cause the inclination learning.

We organize this paper as follows: In section II, we present the methodologies: a) two-class classification with imbalanced dataset, b) imbalanced ratio, c) measure of overlap of individual feature values, d) back-propagation neural network, e) error function and f) measurements. In section III, we present our proposed method: modified MSE function with the consideration of class distribution and overlap ratio. In section IV, we present the experiment and result. In section V, we present the research conclusion.

II. Methodologies

A. Two-class classification with imbalanced dataset

Imbalanced dataset consists of classes of data that have a different number of instances. In the two-classification problem, there are two classes: majority and minority class [1] [2]. For example, a dataset contains 1000 samples, there are 990 instances that belong to majority class and there are 10 instances that belong to minority class. Once this dataset is entered into classifying tools, the classifier model may be bias toward majority class. If the accuracy is 99%, the true predicted instances may be all belong to majority class. In He, and Ghodsi research work [5] note that the effect of imbalanced dataset classification with SVM when the data are highly skew, the number of support vectors of minority class is less than majority class. Furthermore, instances of minority are far from true boundary. These show that the increasing imbalanced ratio have influenced on classification performance.

There are three interested characteristics of imbalanced dataset that effected the classification problem [1]. 1) Class distribution or imbalanced ratio, 2) Lack of data, and 3) Concept complexity or overlap ratio between classes. These characteristics are most considering in the classification task. The standard approaches in machine learning to the classification task perform poorly in the presence of imbalanced dataset. This is due to the fact that most learning algorithms train classifiers by optimizing accuracy. As a result, the classifiers perform well on majority classes and badly on minority ones.

B. Imbalanced ratio

The imbalanced ratio in two-class classification problem is defined as:

$$IR = \frac{number\ of\ majority\ class}{number\ of\ minority\ class} \tag{1}$$

C. Measure of overlap of individual feature values

Fisher's Discriminant Ratio (F1) is one of the overlap measuring tools. The small values of F1 indicate high overlap. It uses the maximum over all feature dimensions to describe a problem. For each feature, the measure f is calculated as:

$$f = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{2}$$

Where μ_1, μ_2, σ_1^2, and σ_2^2 are the means and variances of the two classes, respectively.

D. Back-propagation Neural Network (BPNN)

A network forward propagates activation to produce an output and it backward propagates error to determine weight changes. The weights on the connections between neurons mediate the passed values in both directions. The back-propagation algorithm performs gradient descent to try to minimize the sum square error between the network's output values and the given target values. Fig. 1 shows the back-propagation neural network architecture.

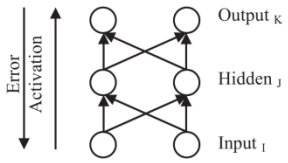


Fig. 1. Back-propagation neural network architecture

E. Error function

In the learning process of BPNN, error or cost function is an evaluating tool. The different of output from learning process with the target output or errors is send back to re-train. Example of error functions are: Cross Entropy (based information), Minkowski, Mean Square Error [8]. In this section, we briefly describe the Mean Square Error in BPNN.

At output nodes, the error between predicted output and desired output of pattern can calculate by eq. 3.

$$E^{(p)} = \frac{1}{2}\sum_{k=1}^{m}\left(t_k^{(p)} - y_k^{(p)}\right)^2 \tag{3}$$

p is pattern
m is number of nodes at output layer
t is desired output
y is predicted output

At the validation stage, the MSE is used to evaluate. The MSE shows the average of every pattern in iteration. An MSE function defined as:

$$\text{MSE} = \frac{1}{P}\sum_{i=1}^{P} E_i \tag{4}$$

P is number of patterns

From equation 3 and 4, when the dataset is imbalanced, then the number of instance that lead to the calculation of error rate are more on the majority class. This will result in misclassification of the minority class.

F. Measurements

There are several measurements for data classification. The most popular metric is the accuracy value on all classes that indicating the accuracy of the overall classes. In the imbalanced dataset classification, the accuracy overall classes could not indicate the true performance. For example, the proportion of majority and minority class is 900:100. The accuracy overall classes is 90%. This accuracy may be true predicted the majority class into 100%, but all instances of minority is misclassification. Therefore, the selecting of measurement tools is significant for Imbalanced Dataset. The ideal tool should be able to ensure the accuracy of all the classes that have a vast different quantity of instances.

A confusion matrix is a visualization tool typically used in supervised and unsupervised learning approaches. In Fig. 2, each row of the matrix represents the instances in an actual class, while each column represents the instances in a predicted class. In this paper, we use the minority class as the positive class and the majority class as the negative class.

		Predicted class	
		Positive	Negative
Actual class	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Fig. 2. Confusion matrix for binary classification problem

- Precision

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \quad (5)$$

- TP rate, Recall, Sensitivity

$$\text{TPR, Recall, Sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (6)$$

- TNR, Specificity

$$\text{TN rate, Specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \quad (7)$$

- F-measure

$$\text{F} - \text{Measure} = \frac{(2 \times \text{precision} \times \text{recall})}{(\text{precision} + \text{recall})} \quad (8)$$

- Geometric mean

In Barandela et al. [9] a metric called the geometric mean (GM) was introduced. With this measure we try to maximize accuracy in order to balance both classes at the same time.

$$G - Mean = \sqrt[n]{\text{sensitivity} \times \text{specificity}} \quad (9)$$

n is number of classes

The performance measures in the experiment are TP rate or sensitivity, F-measure, and Geometric-Mean.

III. Proposed Method

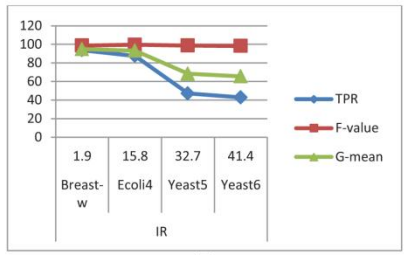
As discuss so far, this paper proposes a modification of traditional cost function, called MSE function. Not only we consider the different between desired output and computing target (result), but also the modified cost function, proposed in this, is considering the degree of class distribution and overlap ratio between two classes. The proposed method involves three main steps as follows:

Step 1: Measuring the degree of class distribution and overlap ratio

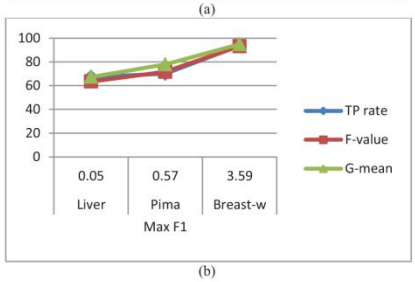
First, the input data in each dimension will be normalized into the range of [0..1]. After that the class distribution or imbalanced ratio for each data dimension is measured using equation 1. Then the Fisher's discriminant ratio (F1) as shown in equation 2 is used to evaluate the degree of overlapping between two classes. As a consequent, we will obtain a set of values of F1 which will be equaled to the number of dimensions. Maximal F1 value is selected to describe an overlapping ratio, that is, the high value indicates a small overlap and the low value indicates a large overlap.

Step 2: Defining the weight of majority class

In this step, the characteristics of the input data as measured in step one are used to consider the degree of weight for the majority class. We have conducted a number of experiments to study the said two characteristics of data: imbalanced ratio and overlap ratio. This experiment is done to see the effect of the classification performance as shown in Fig.3. We have used Back-propagation neural network as a learning algorithm.



(a)



(b)

Fig. 3. (a) Effect of IRs on the measurements (b) Effect of max F1 values on the measurements

The experiment is divided into two cases. The first case is to study the effect of difference in imbalanced ratio to each classification measurement methods: TP-rate, F-value and G-mean. In Fig.3(a) shows 4 datasets which are Breast-w, Ecoli4, Yeast5, and Yeast6, and the value of maximal F1 values of each dataset are 3.59, 3.17, 4.17, and 1.94, respectively. These maximal F1 values indicate small overlap degree. Those dataset got the Imbalance Ratio of 1.86, 15.8, 32.73, and 41.41, respectively. With the above condition, the results show that, if the IRs is high then TP rate, F and G will be low. In contrast, if IRs is low then TP rate, F and G will be high. Second case is to study the effect of maximal F1 values to each measurement methods. In this second case, the value of IRs in each dataset are slightly different. In Fig.3(b) shows 3 datasets Liver, Pima, and Breast-w which have had IRs of 1.38, 1.87, 1.86 and maximal F1 value of 0.05, 0.57, and 3.59, respectively. The experimental results show that, if maximal F1 value is high, then the TP, F and G rate are high too. On the opposite direction, if max F1 value is low then the TP, F and G will be low too.

From the above experiments, the imbalanced ratios and the maximal F1 values have the influence to the performance of imbalanced dataset classification. Therefore, we used the two factors: IRs and maximal F1 to identify weight for the modified standard MSE in the next step. Weights range are between 0 and 1, r_{maj} is weight for majority class and r_{min} is weight for minority class. In our proposed method, we realize the significant of the effect of the majority class to the learning process. Therefore, we set to update the weight only on the majority class, so r_{min} is always given the value of 1. The detail of r_{maj} is shown in table 1.

TABLE 1 WEIGHTS FOR MAJORITY CLASS (r_{maj})

	Max F1 values			
IR	0.00-0.30	0.31-0.70	0.71-2	2.01-5
1.00-10	0.7	0.8	N/A	1
10.01-20	N/A	0.4	N/A	0.8
20.01-30	N/A	0.4	N/A	N/A
30.01-40	N/A	N/A	N/A	0.8
41.01-50	N/A	N/A	0.7	N/A

Step 3: Training and testing

The dataset is divided in proportion of 60% as a training set and 40% as a testing set. This experiment use Back-propagation neural network as a learning algorithm. The setup configuration is based on the trial and error procedure to find a good number of neuron in hidden layer (*H*). Number of nodes in output layer equal to number of class (*m*). The network contains 1 hidden layer.

$$X^{(P1)} = \left[x_1^{(p1)}, x_2^{(p1)}, \dots, x_N^{(p1)}\right],$$
$$X^{(P2)} = \left[x_1^{(p2)}, x_2^{(p2)}, \dots, x_N^{(p2)}\right]$$

P1, P2 are set of majority class, and minority class. *N* is number of nodes in input layer.

We are given $P = P1 \cup P2$, therefore input pattern is:

$$X^{(P)} = \left[x_1^{(p)}, x_2^{(p)}, \dots, x_N^{(p)}\right]$$

Let desire output vector for training pattern X^p show as:

$$t_k^p = \begin{cases} 1 & if\ X^p\ is\ class\ k \\ 0 & otherwise \end{cases}$$

k = 1,2,..m

At hidden nodes, the output signal are define as:

$$y_j^{(p)} = f\left(\sum_{i=1}^{N} w_{ji}\, x_i^{(p)}\right) \quad (10)$$

w_{ji} denotes the weight connecting the x_i to neuron j in hidden layer

At output nodes, the output signal are define as:

$$y_k^{(p)} = f\left(\sum_{j=1}^{H} v_{jk} y_j^{(p)}\right) \quad (11)$$

v_{jk} *denotes the weight connecting* $y_j^{(p)}$ *to* $y_k^{(p)}$

Activation function is sigmoid, show as:

$$f(x) = \frac{1}{1+e^{-x}} \quad (12)$$

At output nodes, the error function is commonly given as the sum of squares of the differences between all predicted and desired output. The proposed are modify by equation 3, define as:

$$E_{prop.}^{(p)} = \begin{cases} \frac{1}{2}\sum_{k=1}^{m} r_{maj}\left(t_k^{(p)} - y_k^{(p)}\right)^2, if\ t_k^{(p)} is\ majority\ class \\ \frac{1}{2}\sum_{k=1}^{m} r_{min}\left(t_k^{(p)} - y_k^{(p)}\right)^2, if\ t_k^{(p)} is\ minority\ class \end{cases} \quad (13)$$

$E_{prop.}^{(p)}$ is total error over the training pattern,

m is Number of nodes in output layer.

r_{maj}, r_{min} are weigths that given by IR and overlap ratio in step 2, $(0 < r_{maj} \leq 1)$, and $r_{min} = 1$.

Next, weights are iteratively updated by the error back-propagation algorithm.

$$\partial_k^{(p)} = -\frac{\partial E_{prop.}^{(p)}}{\partial y_k^{(p)}} \quad (14)$$

$$\partial_k^{(p)} = y_k^{(p)}[1 - y_k^{(p)}]e_k^{(p)} \quad (15)$$

$$e_k^{(p)} = \begin{cases} \mathrm{r_{maj}}(y_k^{(p)} - t_k^{(p)}), if\ t_k^{(p)} is\ majority\ class \\ y_k^{(p)} - t_k^{(p)}, otherwise \end{cases}$$

Updating weights in output layer,

$$\Delta v_{jk}^{(p)} = \propto y_j^{(p)} \partial_k^{(p)} \quad (16)$$

$$v_{jk}^{(p+1)} = v_{jk}^{(p)} + \Delta v_{jk}^{(p)} \quad (17)$$

Updating weights in hidden layer,

$$\partial_j^{(p)} = y_j^{(p)}[1 - y_j^{(p)}] \sum_{k=1}^{M} \partial_k^{(p)} v_{jk}^{(p)} \quad (18)$$

$$\Delta w_{ji}^{(p)} = \propto X_j^{(p)} \partial_j^{(p)} \quad (19)$$

$$w_{ji}^{(p+1)} = w_{ji}^{(p)} + \Delta w_{ji}^{(p)} \quad (20)$$

where, learning rate$(\propto) = 0.01$, iteration=1000

After that, the test set will send to evaluate the model. The predicted output is using argmax(y_k). The measurement tools are described in section 2.

IV. Experiment and Results

The real world imbalanced datasets from UCI Machine Learning Repository are used in our experimentation. Imbalanced ratios are between 1.86 and 15. Some original datasets contained multiple classes which are transformed to two-class problem, they are 1) Yeast and 2) Abalone. After the transformation on dataset "Yeast", we then got Yeast4, Yeast5, and Yeast6, which consist of minority classes as follows: 'ME2', 'ME1' and 'EXC' respectively, the rest are majority classes. For the Abalone9-18, the class label '18' is a minority class and class label '9' is a majority class. The detail of datasets can be found in table 2.

Table 2 characteristics of datasets

Datasets	#Att.	#Ins.	%Min.	%Maj.	IR	Max F1
Liver	7	345	42.03	57.97	1.38	0.05
Breast-w	10	683	34.99	65.01	1.86	3.59
Pima	9	768	34.9	65.1	1.87	0.57
Haberman	3	306	26.47	73.53	2.78	0.18
Abalone9-18	9	731	5.75	94.25	16.4	0.62
Yeast4	9	1484	3.44	96.56	28.1	1.23
Yeast5	9	1484	2.96	97.04	32.73	4.17
Yeast6	9	1484	2.36	97.64	41.4	1.94

Datasets are divided to two sets which are 60% for train, and 40% for test. Back-propagation neural network is used as a learning algorithm. The performance compares with two cost functions, they are: standard MSE and proposed modified MSE.

Table 3 shows the results of 8 imbalanced datasets classification are describe below:

Liver dataset, TP rate, G-Mean of proposed method are better than standard MSE equal 5.17%, and 1.21% respectively. F-measure, proposed method is less than 0.1%.

Breast-w, TP rate, G-Mean, F-measure of proposed method are better than standard MSE equal to 1.06%, 0.56%, and 0.53% respectively.

Pima, TP rate, F-measure of proposed method are better than standard MSE equal to 3.74%, and 0.18%. G-Mean, proposed method is less than standard MSE equal to -0.28%.

Haberman, TP rate, G-Mean, F-measure of proposed method are better than standard MSE equal to 6.25%, 4.01%, and 4.39% respectively.

Abalone9-18, true positive rate, G-Mean, F-measure of proposed method are better than standard MSE equal to 12.5%,12.52%, and 10.2% respectively.

Yeast4, TP rate, G-Mean, F-measure of proposed method are better than standard MSE equal to 25%, 28.36%, and 32.16% respectively.

Yeast5, TP rate, G-Mean, F-measure of proposed method are better than standard MSE equal to 11.76%, 7.33%, and 8% respectively.

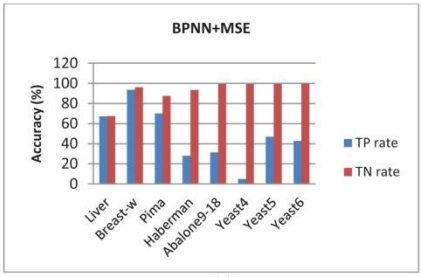
Yeast6, TP rate, G-Mean, F-measure of proposed method are better than standard MSE equal to 7.16%, 6.45%, and 5.21% respectively.

Table 3 TP rate , TN rate, F-measures G-means, accuracy on both classes (%)

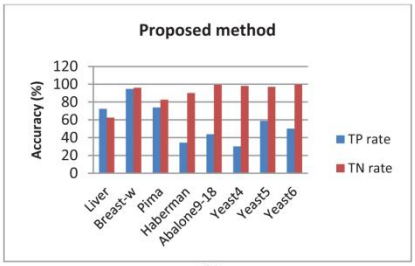
Datasets	Method	TP rate	TN rate	F	G	Acc.
Liver	BPNN+MSE	67.24	67.50	63.41	67.37	67.39
	Proposed	72.41	62.50	64.62	67.27	66.67
Breast-w	BPNN+MSE	93.68	96.05	93.19	94.86	95.22
	Proposed	94.74	96.05	93.75	95.39	95.59
Pima	BPNN+MSE	70.09	87.50	71.77	77.87	80.78
	Proposed	73.83	82.50	71.49	78.05	79.48
Haberman	BPNN+MSE	28.13	93.33	38.30	51.23	76.23
	Proposed	34.38	90.00	42.31	55.62	75.41
Abalone9-18	BPNN+MSE	31.25	99.27	43.48	55.70	95.53
	Proposed	43.75	99.27	56.00	65.90	96.22
Yeast4	BPNN+MSE	5.00	99.30	8.00	22.28	96.12
	Proposed	30.00	98.08	36.36	54.44	96.46
Yeast5	BPNN+MSE	47.06	99.31	55.17	68.36	97.81
	Proposed	58.82	99.13	62.50	76.36	97.98
Yeast6	BPNN+MSE	42.857	99.83	57.143	65.40	98.482
	Proposed	50.02	99.83	63.60	70.62	98.72

Figure 4 (a) shows TP rate and TN rate of BPNN with MSE and (b) TP rate and TN rate of proposed method. TP rate of proposed method are higher than BPNN with MSE, TN rate of proposed method are smaller than BPNN with MSE.

The results of most datasets based on TP rate, G-Mean, and F-measure confirm that modified MSE provide better results than standard MSE. Moreover, the proposed method can improve the accuracy of positive or minority class while accuracy of majority class is not over decreasing.



(a)



(b)

Fig. 4. (a) TP rate and TN rate of BPNN with MSE
(b) TP rate and TN rate of proposed method

V. Conclusion

This paper proposed a modified MSE cost function which considered class distribution and overlap ratio in changing parameter in the cost function. The back-propagation algorithm is used as a training algorithm. Datasets in this paper are the standard one downloaded from UCI Machine Learning Repository. The measurement methods use to measure the performance of the algorithm are: true positive rate, G-Mean, and F value which based mainly on performance of minority class. The results confirm that proposed method have better results than standard MSE on most datasets based on three measurements mentioned above. In the future work, the studying of different characteristics of classes that may affect the performance of minority class will be persued.

Acknowledgment

This work was supported by grant funds from the program Strategic Scholarships for Frontier Research Network for the Ph.D. Program Thai Doctoral degree from the Commission on Higher Education, Thailand.

References

[1] N. V. Chawla, N. Japkowicz and A. Kolcz, "Editorial: Special Issue on Learning from Imbalanced Data Sets," *Sigkdd Explorations Special Issue on Learning from Imbalanced Datasets,* vol. 6, no. 1, pp. 1-6, 2004.

[2] N. V. Chawla, K. W. Bowyer and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research,* vol. 16, pp. 321-357, 2002.

[3] H. Hui, W. Wenyuan and M. Binghuan, "Borderline-smote: A new over-sampling method in imbalanced data sets learning," in *Proceedings of International Conference on Intelligent Computing (ICIC'05)*, 2005, pp.878–887.

[4] S. Chen, G. Guo and L. Chen, "A New Over-Sampling Method Based on Cluster Ensembles," in *Proceedings of 2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, 2010, pp.599-604.

[5] H. He and A. Ghodsi, "Rare Class Classification by Support Vector Machine," in *Proceedings of 2010 20th International Conference on Pattern Recognition (ICPR)*, 2010, pp.548-551 .

[6] G.-S. Xiao and X.-Y. Chen, "Graph classification with imbalanced data sets," in *Proceedings of 2011 First Asian Conference on Pattern Recognition (ACPR)*, 2011, pp.57 - 61.

[7] K. Boonchuay, K. Sinapiromsaran and C. Lursinsap, "Minority split and gain ratio for a class imbalance," in *Proceedings of 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 2011, pp.2060-2064.

[8] C. M. Bishop, Neural Networks for Pattern recognition, Oxford University Press, 1995.

[9] R. Barandela, J. S. Sanchez, V. Garcia and E. Rangel, "Strategies for learning in class imbalance problems," *Pattern Recognition,* vol. 36, no. 3, pp. 849-851, 2003.

TRACK YOUR ACCEPTED ARTICLE

Welcome! **Login** to get personalized options. New user? **Register** | **Why register?**

Your article's details and status are shown in the following table:

Article status

Article title:	Improving Classification Rate Constrained to Imbalanced Data Between Overlapped and Non-overlapped Regions by Hybrid Algorithm
Reference:	NEUCOM14722
Journal title:	Neurocomputing
Corresponding author:	Ms. Piyanoot Vorraboot
First author:	Ms. Piyanoot Vorraboot
Received at Editorial Office:	5 Apr 2014
Article revised:	5 Sep 2014
Article accepted for publication:	2 Oct 2014
Received at Elsevier:	4 Oct 2014
Journal publishing agreement sent to author:	9 Oct 2014
Offprint order form sent to author:	9 Oct 2014
PDF offprint:	Yes
Proofs available:	16 Oct 2014
Proofs returned:	28 Oct 2014
Journal publishing agreement returned:	27 Oct 2014
Offprint order form returned:	13 Oct 2014
Accepted manuscript available online:	view accepted manuscript online
Cited by in Scopus:	0
DOI information:	10.1016/j.neucom.2014.10.007
Status comment:	**No further corrections can now be made.** At this moment it is not yet possible to give you information about the publication date. This depends on the number of articles lined up for publication in the journal. Citation information will be shown when available.

ARTICLE IN PRESS

Neurocomputing ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Improving Classification Rate Constrained to Imbalanced Data Between Overlapped and Non-overlapped Regions by Hybrid Algorithm

Piyanoot Vorraboot [a,*], Suwanna Rasmequan [a], Krisana Chinnasarn [a], Chidchanok Lursinsap [b]

[a] Knowledge and Smart Technology Research Center, Faculty of Informatics, Burapha University, Chonburi 20131, Thailand
[b] Advanced Virtual and Intelligent Computing Research Center, Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Bangkok 10330, Thailand

ARTICLE INFO

Article history:
Received 5 April 2014
Received in revised form
5 September 2014
Accepted 2 October 2014
Communicated by V. Palade

Keywords:
Highly Imbalanced Data with Overlapping area
Soft-Hybrid Algorithms
Responsive Area Mapping Classifications

ABSTRACT

Classification performance on an highly imbalanced with overlapping data set is often found to be inefficient in term of computational time. Because all data were used in the learning step. But it was observed that data could be categorized into proper sub-grouped using data analysis techniques. In this research, a combination technique, called the "Soft-Hybrid" algorithm, was proposed for improving classification performance. The proposed method was divided into two main phases: boundary area determination and responsive classification algorithms for each sub-area. In the first phase, data were grouped as (1) non-overlapping data, (2) borderline data, and (3) overlapping learning data using modified Hausdorff distance, Radial Basis Function Network and K-Means clustering technique with Mahalanobis distance. Then, modified kernel learning method, modified DBSCAN and RBF network were applied to classify the data into proper groups based on statistical values from the classification phase. Finally, the results of all techniques were combined. The experimental results illustrated that the proposed method can significantly improve the effectiveness in classifying imbalanced data having large overlapping sections based on TP rate, *F*-measure and *G*-mean measures. Moreover, the computational times of the proposed method were lower than the standard algorithms used for this type of this problem.

1. Introduction

In the past decade, binary classification with imbalanced had been proposed and implemented in many real world applications such as in the case of medical diagnosis. Sick people should be classified into a minority or positive class and put into an attention list. However, using the traditional classifying algorithm has the potential to predict an instance of a positive class as an instance of a negative class. This was because of the occurrence of the dominant tendency. That is, the number of normal people was much larger than the number of sick people. Therefore any sick person might have a tendency to be classified as normal or majority class [1–3]. In case of accuracy evaluation, many proposed methods in the literature produced efficient or acceptable results. However, in some proposed methods, they worked well with high computational complexity. In this paper, Soft-Hybrid algorithm with low computational time complexity for the large data sets which are highly imbalanced due to overlapping areas of data were considered. These were done by reducing the domination or the influence of the majority class over the minority class. Extracting all obviously non-overlapping areas was prior applied. Then, RBF learning for identifying optimal boundary of purely overlapping area was trained. Complement points from the RBF learning were clustered into two groups. Each group was extracted into sub-clusters based on their Eigenvalues. Unlike standard algorithms based on the brute force algorithms, the proposed method expended more effort classifying the more difficult overlapping subset of data. In most cases of our study our method was less computationally intensive.

In the literature, researchers proposed a number of methods for classifying an imbalanced data set. Those methods can be divided into two main areas: the data level and the algorithm level [4,5]. In the data level, the number of instances in each class was changed. Most methods attempted to solve the imbalanced data set at the data level mainly with over-sampling techniques, under-sampling techniques, and hybrid techniques. With the over-sampling technique, a number of instances in the minority class was added until the number of its members was equal to the number of instances in majority class. With an under-sampling technique, the size of

* Corresponding author.
E-mail address: piyanoot_v@yahoo.com (P. Vorraboot).

http://dx.doi.org/10.1016/j.neucom.2014.10.007
0925-2312/

Nomenclature

A	set of majority class
B	set of minority class
X	set of observed input data, $\mathbf{X} = \mathbf{A} \cup \mathbf{B}$
Z	set of non-overlapping data
U	set of uncertainty data $\mathbf{U} = \mathbf{X} - \mathbf{Z}$
Y	set of borderline data
O	set of overlapping data $\mathbf{O} = \mathbf{U} - \mathbf{Y}$
IR	imbalanced Ratio
maxF	maximum Fisher's discriminant ratio
KL	Kullback–Leibler divergence
MD	Mahalanobis distance
HD	Hausdorff distance
dDBSCAN	dynamic density based spatial clustering of applications with noise
mHD	modified Hausdorff distance
mMD	modified Mahalanobis distance
F-minor	*F*-measure on the minority class

the majority class was downsized. When using with a hybrid method, the membership of both classes was readjusted until the sizes of classes are all equal [6–9]. However, these sampling techniques were likely to cause an over-fitting issue. This consequence was a result of sampling with a replacement condition. Another possible issue was the discarding of useful data in the majority class when an under-sampling technique is implemented. In response to these issues numerous methods have been proposed to attempt to solve the imbalanced data set problem at the data level. SMOTE was a technique proposed by Chawla et al. [2] that created new artificial minority instances by extrapolating values between existing minority instances. Han et al. [10] proposed the modification algorithms of the SMOTE, called the borderline-SMOTE. It was an over-sampling technique which applied the data points at the border between classes. Furthermore, Batuwita and Palade [11] proposed the re-sampling techniques called '*closet under*' and '*closet over*' methods. These methods found the elements located closer the boundary region by the SVM learning method, then used those elements in re-sampling.

At the algorithmic level, the researchers tried to apply a variety of existing traditional algorithms such as Support Vector Machine, Neural Network and K-Nearest Neighbor. However, some researchers proposed modifications to those methods such as assigning a different weight to each class depending on the number of instances [12–18]. Those attempts were to reduce the effect of bias-classification. The bias was caused by the fact that an equal weight assigned to all classes often leads to an improper sum-squared error of the learning process which is mainly derived from the instances of the majority class. In that case, the separation hyperplane was shifted to the minority class area. In addition, in the research work of Batuwita and Palade [19], the modification of the SVM algorithm for the class imbalanced learning called the FSVM-CIL was proposed. This method was used to handle the problem of imbalanced classes with the outliers and noise.

The rest of this paper is organized as follows. In Section 2, the problem formulation of this work will be discussed. In Section 3, related concepts and background will be shown. In Section 4, the proposed method using soft-hybrid algorithms or responsive classification algorithms for an improved classification rate in a highly imbalanced large data set will be discussed. In Section 5, the experiments and results will be discussed. The final section will be shown in the conclusion.

2. Problem formulation

There are a number of applications for classification algorithms on highly imbalanced overlapping data such as the cancer cell classification. Cancer and normal cells have been found in the same location in the human body. It is known that the number of normal cells is much greater than the number of cancer cells. In other words, the two sets of cells are highly imbalanced with an overlapping area. We then assume that two Gaussian classes majority class *A* and minority class *B* exist with different parameters as illustrated in Fig. 1(b), which shows the imbalanced data set that consists of overlapped classes. The two classes in an imbalanced data set *X* can be defined in the following equation:

$$\mathbf{X} = \mathbf{A} \cup \mathbf{B} \tag{1}$$

The objective of the imbalanced data set classification is to improve the recognition rate of highly uneven instances between classes that always leads to a bias in classification. The standard learning algorithms proposed in the literature provided an inappropriate accuracy because of the cost function [20],

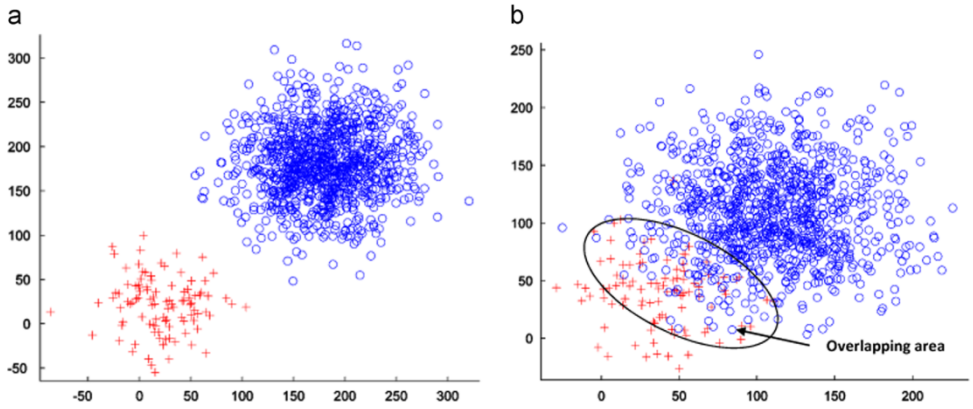


Fig. 1. Example of imbalanced large data set. (a) Non-overlapping data set. (b) Overlapping data set.

Please cite this article as: P. Vorraboot, et al., Improving Classification Rate Constrained to Imbalanced Data Between Overlapped and Non-overlapped Regions by Hybrid Algorithm, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.10.007

$E = \sum_{i=1}^{n}(o_i - t_i)^2$ where $n = n_A + n_B$ is the number of instances in majority class A and minority class B. Because n_A is much larger than n_B, the cost function E is dominated by the error from the majority class A. Hence, misclassification of the minority class B occurs, as shown in Fig. 2. But there are many successful algorithms for classifying imbalanced data sets with overlapping data such as those in [21–23]. However, they required a high computational time complexity. In this paper, a soft-hybrid learning algorithm with low computational complexity for highly imbalanced large data sets was proposed. The learning algorithm groups the data points into their proper areas: non-overlapping, borderline, and overlapping, and classifies them by the proper classification algorithm as presented in Section 4.

3. Related concepts and background

Since this study concerns the problem caused by imbalanced data, the boundaries of classes and the overlapped region between two classes must be identified prior to the learning process. Thus several distribution measures and a distance between sets are briefly summarized in this section.

3.1. Highly overlapping imbalanced large data

For imbalanced data sets in dichotomy classification [24,25], the number of instances in the majority class A is much larger than the number of instances in the minority class B or $n_A \gg n_B$. The degree of imbalanced data set or the imbalanced ratio (IR) is defined as Eq. (2). The high value of IR can lead to a high misclassification rate

$$IR = \frac{n_A}{n_B} \tag{2}$$

3.2. Fisher's Discriminant Ratio

Fisher's Discriminant Ratio is an overlap measuring tool [26,27]. A small value of Fisher's Discriminant Ratio indicates a high overlap. It uses the maximum Fisher's Discriminant Ratio ($maxF$) over all dimensions to describe a problem. The Fisher's Discriminant Ratio of each dimension can be described in the following equation:

$$f = \frac{(\mu_A - \mu_B)^2}{(\sigma_A^2 + \sigma_B^2)} \tag{3}$$

where f is Fisher's Discriminant Ratio of dimension i, μ_A and μ_B are means of majority class A and minority class B of dimension i, respectively. σ_A^2 and σ_B^2 are the variances of majority class A and minority class B of dimension i, respectively.

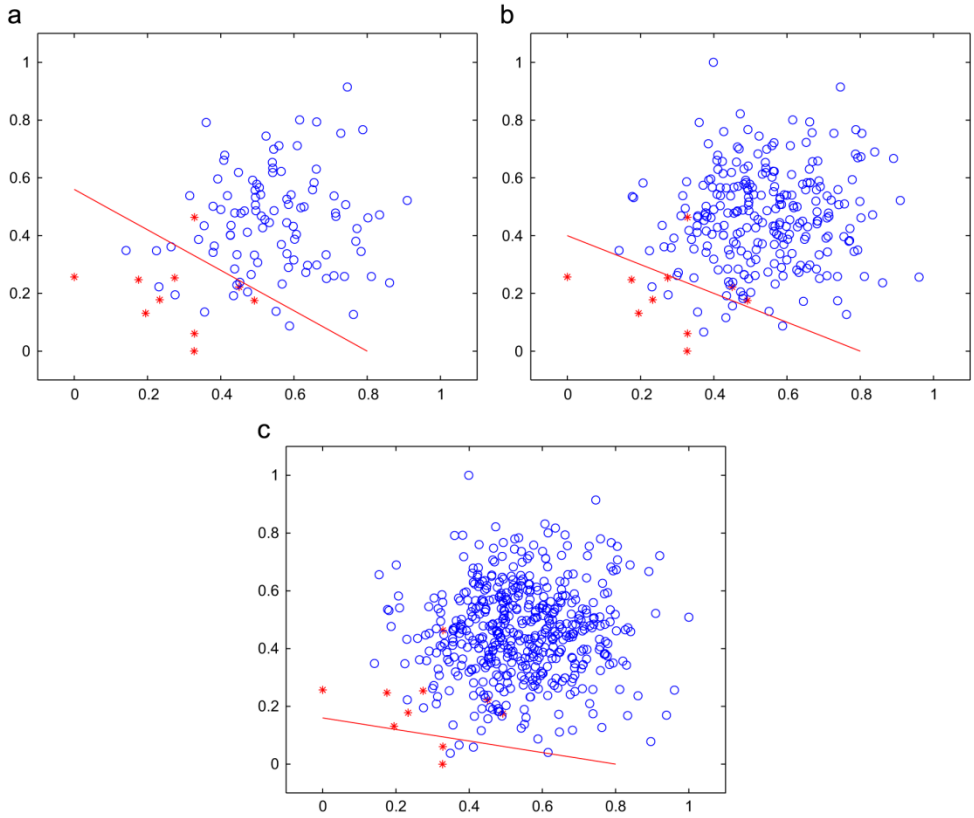


Fig. 2. Example of imbalanced large data set. (a) $IR = 10$. (b) $IR = 25$. (c) $IR = 50$.

Please cite this article as: P. Vorraboot, et al., Improving Classification Rate Constrained to Imbalanced Data Between Overlapped and Non-overlapped Regions by Hybrid Algorithm, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.10.007

3.3. Hausdorff distance (HD)

Hausdorff distance was introduced by Felix Hausdorff [28] to measure the distance between sets by comparing the distance between all members of a set. For example, let **A** and **B** be sets of input data, $\mathbf{A}=\{a_1,a_2,\ldots,a_n\}$ and $\mathbf{B}=\{b_1,b_2,\ldots,b_m\}$. First, the minimum Euclidean distance of each point on set **A** to all points on set **B** is computed. Second, the maximum of the minimum distances from set **A** to set **B** is filtered. The same process is applied to the data in set **B**. The Hausdorff distance between set **A** and set **B**, denoted as $H(\mathbf{A},\mathbf{B})$, is the maximum distance between the maximum of minimum distance from set **A** and set **B**. The Hausdorff distance between set **A** and set **B** is defined as

$$H(\mathbf{A},\mathbf{B})=\max(h(\mathbf{A},\mathbf{B}),h(\mathbf{B},\mathbf{A})) \tag{4}$$

where

$$h(\mathbf{A},\mathbf{B})=\max_{a\in\mathbf{A}}\min_{b\in\mathbf{B}}\|a-b\|$$
$$h(\mathbf{B},\mathbf{A})=\max_{b\in\mathbf{B}}\min_{a\in\mathbf{A}}\|b-a\| \tag{5}$$

3.4. Kullback–Leibler (KL) divergence

Kullback–Leibler divergence is an asymmetric measure of the difference between two probability distributions P and Q. For discrete probability distributions P and Q, Kullback–Leibler divergence of Q from P can be denoted as $KL(P\|Q)$ where

$$KL(P\|Q)=\sum_i \ln\frac{P(i)}{Q(i)}P(i) \tag{6}$$

is referred to a loss of information when Q is used to estimate P [29].

3.5. Mahalanobis distance (MD)

Mahalanobis distance is a measure of the distance between a point P and a distribution D, introduced by Mahalanobis in 1936 [30]. It is based on correlations between variables by which different patterns can be identified and analyzed. The Mahalanobis distance of vector $\mathbf{g}$, $\mathbf{g}=(g_1,g_2,\ldots,g_d)^T$ from an observation group **G** with $\mu=(\mu_1,\mu_2,\ldots,\mu_d)^T$ is defined as

$$MD(\mathbf{g},\mathbf{G})=\sqrt{(\mathbf{g}-\mu)^T\Sigma^{-1}(\mathbf{g}-\mu)} \tag{7}$$

where Σ^{-1} is the inverse of the covariance matrix of **G**.

3.6. Kernel functions

Kernel functions can be used in many applications as they provide a simple bridge from linearity to non-linearity for learning algorithms which can be expressed in terms of dot products. The polynomial kernel function as shown in Eq. (8) is a function commonly used in machine learning

$$K(x_i,y_j)=(\alpha\cdot x_i^T\cdot x_j+c)^d \tag{8}$$

The adjustable parameters are the slope α, the constant term c, and the polynomial degree d.

4. Proposed methods

It has been observed that imbalanced data with overlapping area can be separated into three sub-regions: marginal area or obviously non-overlapped area, uncertainty overlapped area, and purely overlapped area, as illustrated in Fig. 3. But the difficulty in reducing the domination of the majority class was in how to find the real boundary, for placing a set of separating hyperplanes between classes. Note that the previously proposed learning algorithms must employ all data points of each data set. This leads to a high computational complexity. Hence, the low computational time complexity method called the "Soft-Hybrid" for imbalanced data was described in this section. The proposed method consisted of two main steps: Boundary Area Determination and Classification as described in Sections 4.1 and 4.2, respectively. In the boundary area determination phase, data were separated into its proper constituent areas which were non-overlapping, borderline, and overlapping areas. Then, in the classification phase the optimal classification algorithms for each area were used.

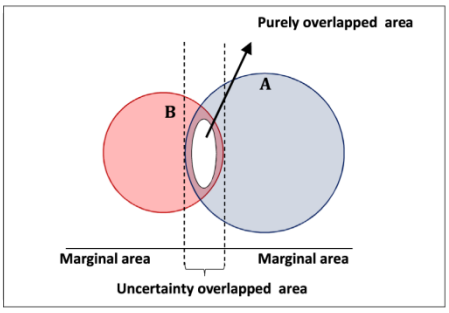


Fig. 3. Three sub-regions of imbalanced data with overlapping area.

As previously described, a large imbalanced data set can be considered as a union of independent sources of the data set. Our study focused only on two Gaussian classes with independent and identical distributions (iid). With these two distributions, there is no relation among the members of the classes and they are independent from each other. Let $\mathbf{A}=\{a_i,t_A\}$, $\mathbf{B}=\{b_i,t_B\}$ and $\mathbf{X}=\mathbf{A}\cup\mathbf{B}$ be majority class, minority class, and the observed overlapping imbalanced data, respectively.

4.1. Boundary area determination

To reduce learning time complexity, input data must be preprocessed. As shown in Fig. 3, the marginal area or obviously non-overlapped area could be extracted or removed from the observed input data before proceeds the further steps, as detailed in Section 4.1.1. Then, identification boundary between uncertainty overlapped area and purely overlapped area was proposed by the RBF learning, as explained in Section 4.1.2. Finally, binary clustering of the uncertainty overlapped area or borderline area using K-Means with Mahalanobis distance was used, as shown in Section 4.1.3. It was believed that the proposed method for determining boundary area provided a lower computational time algorithm with acceptable results.

4.1.1. Extracting obviously non-overlapping areas

Hausdorff distances as shown in Eq. (5) were applied for extracting obviously non-overlapping areas. $h(\mathbf{A},\mathbf{B})$ and $h(\mathbf{B},\mathbf{A})$ were calculated. Fig. 4(a) shows boundaries of observed overlapping input data belong to class A and class B. Let $\max_{a\in\mathbf{A}}(a)$ and $\max_{b\in\mathbf{B}}(b)$ be the maximum Hausdorff distance point of classes A and B as shown in Fig. 4(b) and (c), respectively. Therefore, data points in Hausdorff distance of its class were assigned into obvious non-overlapping area as illustrated in Fig. 4(d).

Please cite this article as: P. Vorraboot, et al., Improving Classification Rate Constrained to Imbalanced Data Between Overlapped and Non-overlapped Regions by Hybrid Algorithm, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.10.007

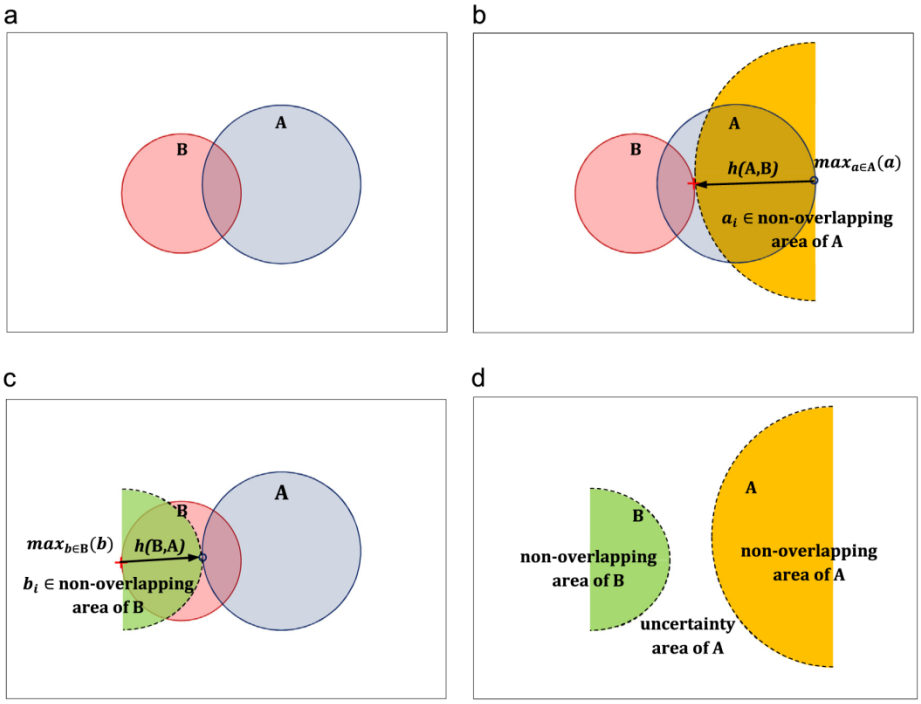


Fig. 4. The extracting obviously non-overlapping area. (a) Boundaries of observed input data. (b) Hausdorff distance from class *A* to class *B*. (c) Hausdorff distance from class *B* to class *A*. (d) Non-overlapping area and uncertainty area.

Algorithm 1. Extracting obviously non-overlapping area.

1. **INPUT**: Training set **X**, $\mathbf{X} = \mathbf{A} \cup \mathbf{B}$
2. **BEGIN**
3. Compute $h(\mathbf{A},\mathbf{B})$ and $h(\mathbf{B},\mathbf{A})$
4. Compute $d_i = \|a_i, \max_{a\in\mathbf{A}}(a)\|$ and $d_j = \|b_i, \max_{b\in\mathbf{B}}(b)\|$
5. **IF** $d_i < h(\mathbf{A},\mathbf{B})$ **THEN**
6. $a_i \in$ non - overlapping area of class *A*
7. **ELSE**
8. $a_i \in$ uncertainty area
9. **ENDIF**
10. **IF** $d_j < h(\mathbf{B},\mathbf{A})$ **THEN**
11. $b_j \in$ non-overlapping area of class *B*
12. **ELSE**
13. $b_i \in$ uncertainty area
14. **ENDIF**
15. **END**

From the proposed idea described above, the result of the extracting obviously non-overlapping areas with real data is shown in Fig. 5. Fig. 5(a) shows the Hausdorff distance between two classes. The obviously non-overlapping areas **Z** are shown in Fig. 5(b). Hence, subtraction points $\mathbf{X}-\mathbf{Z}$ were determined as an uncertainty area as shown in Fig. 5(c).

4.1.2. Optimal boundary learning algorithm

Data in uncertainty area obtained from the previous step could be considered as purely overlapping data and borderline data. In the previous literature [12,31,32], the researchers suggested the mapping method for mapping overlapping data into the higher dimensional space. It led to the high number of data and high computational time in the learning process. Therefore in this section, the method for finding the optimal boundary of sub-areas of the uncertainty area was proposed. The proposed method consists of two steps: constructing optimal boundary model and assigning the optimal boundary steps.

1. *Constructing optimal boundary model*: To be able to find the proper model for constructing the boundary, the training and testing data were generated. Then, statistic values for identifying characteristic of each group *IRs*, *maxFs*, *KLs* and the distance between center of two classes were computed. All four parameters were set as an input vector for the RBFN learning with Guassian kernel. Output vector was the optimal Mahalanobis distance for each class. Then, the optimal Mahalanobis distances of each class $optMD_A$ and $optMD_B$ for determining possibilistic boundary of class *A* and class *B* were obtained from the testing step as shown in Fig. 6(a).
2. *Assigning the optimal boundary*: Since, the optimal Mahalanobis distances $optMD_A$ and $optMD_B$ are achieved from the previous step. These two parameters were not enough to identify the optimal boundary between the pure overlapping and borderline areas. Therefore, the Mahalanobis distances between each data point and two classes were computed, for data points in class *A* called $MD(a_i,\mathbf{A})$ and $MD(a_i,\mathbf{B})$, or data points in class *B* called $MD(b_j,\mathbf{A})$ and $MD(b_j,\mathbf{B})$. Consequently, four rule-based classification algorithms for identifying optimal boundary in the uncertainty area are established in Algorithm 2.

Please cite this article as: P. Vorraboot, et al., Improving Classification Rate Constrained to Imbalanced Data Between Overlapped and Non-overlapped Regions by Hybrid Algorithm, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.10.007

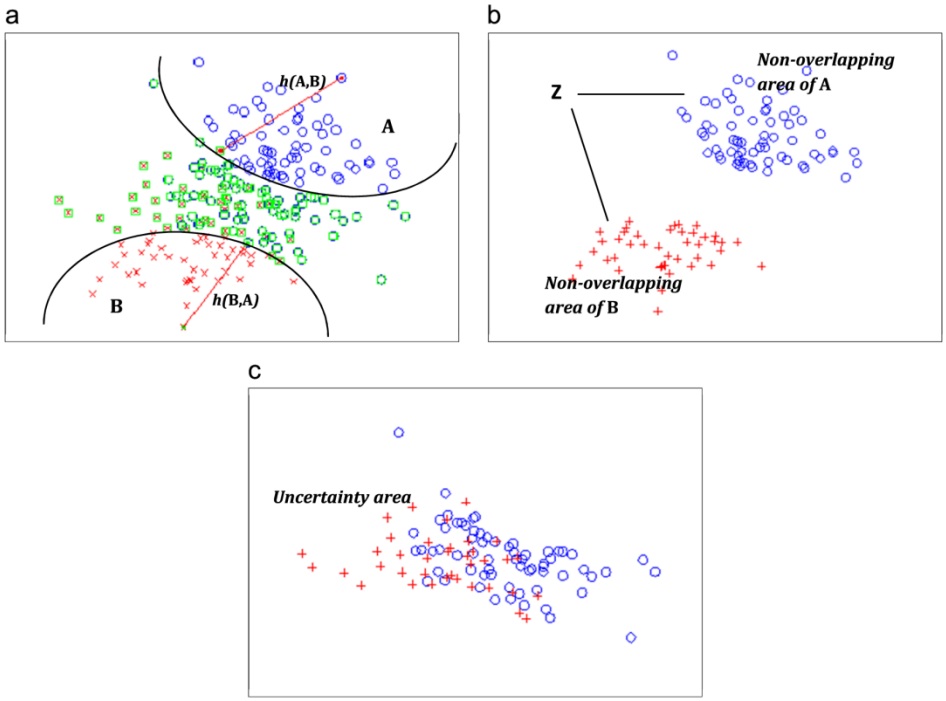


Fig. 5. Example of extracting obviously non-overlapping area. (a) Modified Hausdorff distance. (b) Data in obviously non-overlapping area. (c) Data in uncertainty area.

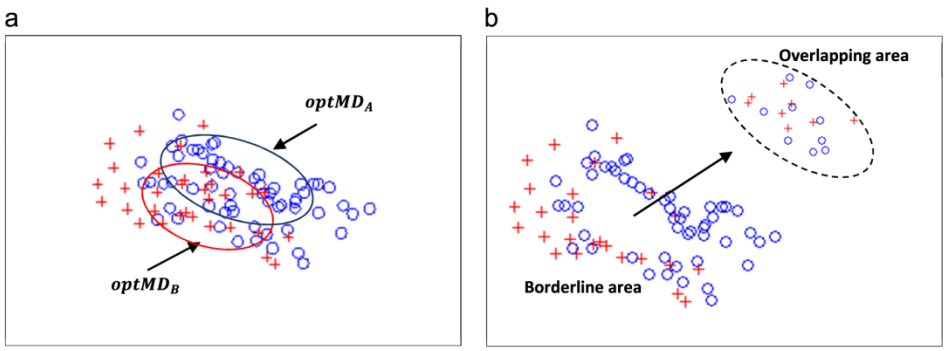


Fig. 6. Example of the optimal boundary learning algorithm. (a) The optimal Mahalanobis distances of class *A* and class *B*. (b) The overlapping and borderline areas.

Algorithm 2. Assigning optimal boundary algorithm.

1. **INPUT:** uncertainty data set, $\mathbf{U} = \mathbf{X} - \mathbf{Z}$
2. **OUTPUT:**
   1: **O:** set of overlapping data.
   2: **Y:** set of borderline data.
3. **Begin**
4. Let $\mathbf{Y} = \phi$ and $\mathbf{O} = \phi$.
5. For $a_i \in \mathbf{A}$; $1 \le i \le n$ **do**
6. Compute $MD(a_i, \mathbf{A})$ and $MD(a_i, \mathbf{B})$.
7. **If** $MD(a_i, \mathbf{A}) > optMD_{\mathbf{A}}$ or $MD(a_i, \mathbf{B}) > optMD_{\mathbf{B}}$ **then**
8. $\mathbf{Y} = \mathbf{Y} \cup \{a_i\}$.
9. **EndIf**
10. **EndFor**
11. **For** $b_j \in \mathbf{B}$; $1 \le j \le m$ **do**
12. Compute $MD(b_j, \mathbf{A})$ and $MD(b_j, \mathbf{B})$.
13. **If** $MD(b_j, \mathbf{A}) > optMD_{\mathbf{A}}$ or $MD(b_j, \mathbf{B}) > optMD_{\mathbf{B}}$ **then**
14. $\mathbf{Y} = \mathbf{Y} \cup \{b_j\}$.
15. **EndIf**
11. **EndFor**
12. $\mathbf{O} = \mathbf{U} - \mathbf{Y}$
13. **End**

4.1.3. Clustering boundary of borderline area

In this process, the outlier points of the borderline area as shown in Fig. 6(b) were detected and removed by Box-plot statistical value measurement based on Mahalanobis distance with

Please cite this article as: P. Vorraboot, et al., Improving Classification Rate Constrained to Imbalanced Data Between Overlapped and Non-overlapped Regions by Hybrid Algorithm, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.10.007

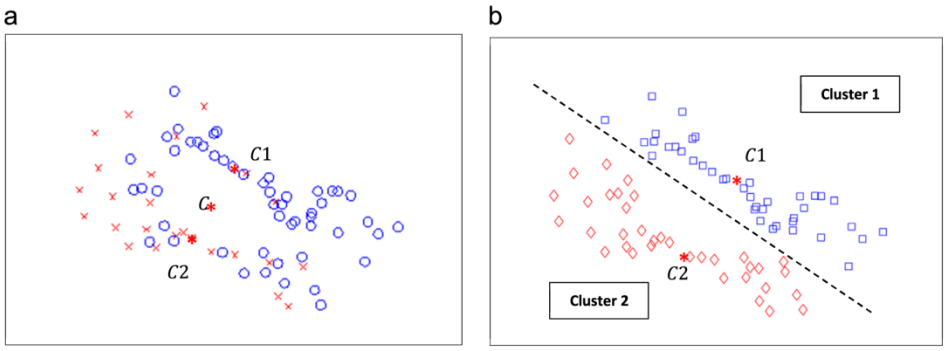


Fig. 7. Example of the clustering boundary of borderline area. (a) Initial cluster centroids. (b) Clusters of data in borderline area.

their center. K-Means clustering based on Mahalanobis distance was used to clustering boundary of borderline area. The minimum Mahalanobis distance point from mean C of all data was assigned to be an initial centroid point of cluster 1, called C_1. Then, the Miller point $C2$, opposite to $C1$, was assigned to be an initial centroid point of cluster 2, as shown in Fig. 7(a). After that, data in borderline area were clustered using K-Means clustering based on the Mahalanobis distance, as detailed in Algorithm 3. The clustering result is shown in Fig. 7(b).

Algorithm 3. Clustering boundary of borderline area.

1. **INPUT:** Training set $\mathbf{X} \in$ borderline area, $\mathbf{X} = \mathbf{A} \cup \mathbf{B}$
2. **BEGIN**
3. Compute C from mean of $\mathbf{X}$
4. C_1 is minimum point of $MD(x_i, C)$
5. $C_2 = C + (C - C_1)$
6. **REPEAT**
7. **IF** $MD(x_i, C_1) < MD(x_i, C_2)$ **THEN**
8. $x_i \in$ cluster1
9. **ELSE**
10. $x_i \in$ cluster2
11. **ENDIF**
12. Update C_1 by compute mean of cluster1
13. Update C_2 by compute mean of cluster2
14. **UNTIL** C_1 and C_2 are not change
15. **END**

4.2. Responsive classification algorithms

After boundary of each region was determined in the previous section, responsive learning algorithm for each extracted region was proposed. The RBF network (RBFN) was applied and learned for classifying the obviously non-overlapped region as shown in Section 4.2.1. For data classification in the borderline region, a learning method called the dDBSCAN with dynamic radius was proposed as illustrated in Section 4.2.2. Finally, a RBF network with polynomial kernel function of degree 2 was deployed to capture the data located in the purely overlapped region as detailed in Section 4.2.3.

4.2.1. Classification of obviously non-overlapping areas

A supervised neural network, the RBFN, has been used for classifying data points of obviously non-overlapping areas. As described in subsection an identifying obviously non-overlapping areas, two obviously non-overlapping areas were obtained. One belongs to class A and one belongs to class B as shown in Fig. 4(c). RBFN used classification algorithm for non-overlapping data set. The optimal parameters obtained from the learning process for the RBFN use two Gaussian kernels.

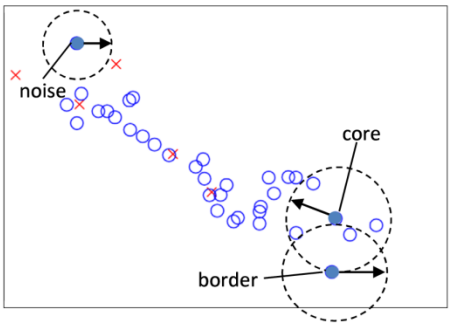


Fig. 8. The parameters of dynamic DBSCAN.

4.2.2. Classification of borderline area

Data points located in the borderline area consist of data points from two classes. The shape of this area will normally be chestnut-like. That is why we propose to classify that data using Density-based Spatial Clustering of Applications with Noise (DBSCAN) [33]. Because of its multivariate nature, the standard DBSCAN algorithm was modified by fixing the radius and the specified number of points (minPts) for each sub-density area. From the experiments, the minimum point was set to 3, the proper radius to use as an initial estimate for each sub-density area was the λ_2 of all the Eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_l, \ldots, \lambda_m$, where m be the number of data dimensions. During the classification process, the radius was automatically shrinking down to $\lambda_3, \lambda_4, \ldots, \lambda_m$ to fit its coverage range. This method called the dynamic DBSCAN (dDBSCAN).

From Fig. 8 shows the parameters of dDBSCAN algorithm. For example, let *minPts* be 3.

Definition 1. A point is called a *core point* if it has more neighbors than *minPts* within the same radius.

Please cite this article as: P. Vorraboot, et al., Improving Classification Rate Constrained to Imbalanced Data Between Overlapped and Non-overlapped Regions by Hybrid Algorithm, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.10.007

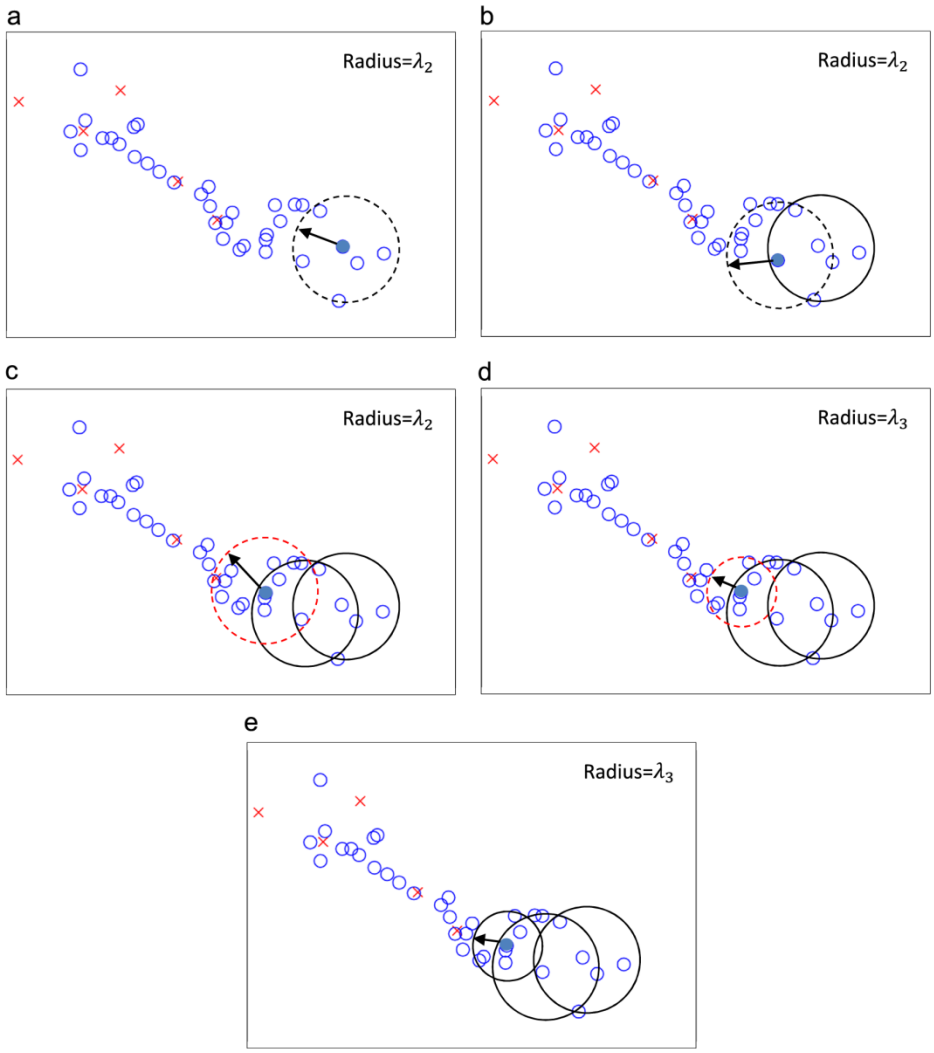


Fig. 9. The process of the dynamic DBSCAN.

Definition 2. A *border point* is a point having fewer neighbors than *MinPts* within the same radius and it is also in the neighborhood of a core point.

Definition 3. A *noise point* is a point which is neither a *core point* nor a *border point*.

Fig. 9 shows the dDBSCAN computation steps. For each class, while x_i was unvisited, start to find the neighbors of x_i. In Fig. 9(a), the learning process with an arbitrary data point was started. Then, λ_2 was applied as the first radius. The members of this group were counted as long as the condition of minimum point equal to 3 was met. Then each member of the neighborhood was used to also repeatedly find its neighbor to cover most members. Once a member of a different class was found, as shown in Fig. 9(c), the radius of such a sub-class was reduced to the next eigenvalue, as shown in Fig. 9(d). The algorithm for classifying data in borderline area is as shown in Algorithm 4.

Please cite this article as: P. Vorraboot, et al., Improving Classification Rate Constrained to Imbalanced Data Between Overlapped and Non-overlapped Regions by Hybrid Algorithm, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.10.007

Algorithm 4. Classification of borderline area.

1. **INPUT**: Training set $\mathbf{X} \in$ borderline area, $\mathbf{X} = \mathbf{A} \cup \mathbf{B}$
2. **BEGIN**
3. **REPEAT**
4. Arbitrarily select a point of x_i
5. Retrieve all points in the same class which closed to x_i with respect to the radius or variance of the 2nd EigenValue (λ_2)
6. **IF** x_i is *core point* **THEN**
7. A sub-class is formed
8. **ELSE**
9. The radius of such sub-class is reduced to the next EigenValue
10. Retrieve all points in the same class which closed to x_i with respect to the new radius
11. **ENDIF**
12. **IF** x_i is a *border point* **THEN**
13. Visit the next learning data point
14. **ENDIF**
15. **UNTIL** All the data points have been processed
16. **END**

4.2.3. Classification of overlapping area

According to the main objectives of the proposed method was to reduce the computational complexity in large data sets. Therefore, determination processes for reducing sizes of input data was proposed. As a result, the computational time for this step was much lower than other algorithms proposed in the literatures [12,31,32]. The average amount of data removed was 62.45% of the synthetic data set and 36.59% of the standard benchmarked data set. Data points located in the overlapped region were then directed towards the optimal polynomial parameters of the Kernel function. In this study, the polynomial Kernel function of degree 2, $K(X,Y) = \phi(X)^T \phi(Y)$, was a suitable parameter to apply. The classification algorithm used in this region was the RBFN.

5. Experiment and results

5.1. Data set

The data sets in this research consisted of synthetic data sets and the standard benchmarking data sets. These data sets were two classes with a normal distribution and independence from each other. The standard benchmarking data sets were downloaded from the UCI repository of machine learning databases [34]. The summary descriptions of both the synthetic data sets and the standard benchmarking data sets are shown in Table 1.

From Table 1, let, *#ins* be the numbers of instances of two classes, *#att* be the numbers of attributes, *#min* be the numbers of elements in minority class, *#maj* be the numbers of elements in the majority class, *IR* be an imbalanced ratio between two classes, *KL* be the Kullback–Leibler divergence between two classes and *maxF* be the maximum Fisher's value of two classes. From the research goals were to improve classification performance for highly imbalanced data with overlapping and to reduce the computational time in learning process. To indicate these points, the range of four properties *IR*, *KL*, and *maxF* was set to 1.38–1.56, 0.7–24.51, and 0.1–2, respectively.

It was known that some of the standard benchmarking data sets, including the Page-blocks-1-3_vs_4, Vehicle1, and Vehicle3 data sets, were multi-class data sets. In this regard, a method suggested by KEEL [35] was used for mapping each multi-class data set into a binary class data set. In this experiment, the Page-blocks-1-3_vs_4 classes I, II and III were combined and labeled as the majority class while class IV was labeled as the minority class. For the Vehicle1 data set with four classes, the class labeled as "Saab" was chosen as the minority class while the remaining classes were combined and labeled as the majority class. In the Vehicle3 data set with four classes, the class labeled as "Opel" was set as the minority class while the remaining classes were combined and labeled as the majority class.

Table 1
Summary description of data sets.

Name	#ins	#att	#min	#maj	IR	KL	maxF
Synthetic data sets							
A1	2000	2	280	1720	6.14	3.71	1.64
A2	2000	2	300	1700	5.67	3.1	1.74
A3	2000	2	290	1710	5.90	3.04	1.74
A4	2000	2	290	1710	5.90	3.06	1.75
A5	2000	2	313	1687	5.39	2.96	1.79
A6	2000	2	270	1730	6.41	3.97	1.8
A7	2000	2	204	1796	8.80	3.06	1.8
A8	2000	2	322	1678	5.21	3.14	1.87
A9	2000	2	270	1730	6.41	3.56	1.91
A10	2000	2	280	1720	6.14	3.28	1.93
A11	2000	2	200	1800	9.00	3.37	1.95
A12	2000	2	270	1730	6.41	3.34	1.95
A13	2000	2	210	1790	8.52	3.84	2.08
Standard benchmarking data sets							
Haberman	306	3	81	225	2.78	0.7	0.18
Liver	345	6	145	200	1.38	0.86	0.05
Pakinsons	195	22	48	147	3.06	24.51	1.5
Pima	768	8	268	500	1.87	1.86	1.5
Page-blocks-1-3_vs_4	472	10	28	444	15.6	149	1.5
Vehicle1	846	18	217	629	2.9	6.68	0.17
Vehicle3	846	18	212	634	2.99	6.01	0.18
German	1000	24	300	700	2.33	3.10	0.66

Table 2
Summary description of data sets in the three areas.

Data sets	Non-overlapping area			Borderline area			Overlapping area		
	#ins	#min	#maj	#ins	#min	#maj	#ins	#min	#maj
A1	829	97	732	225	68	157	946	115	831
A2	1024	59	965	303	91	212	673	150	523
A3	865	47	818	331	65	266	804	178	626
A4	597	46	551	458	117	341	945	127	818
A5	1068	56	1012	285	67	218	647	190	457
A6	1236	74	1162	160	79	81	604	117	487
A7	987	37	941	320	57	263	702	110	592
A8	805	82	723	316	42	274	879	198	681
A9	622	68	554	535	78	457	843	124	719
A10	756	55	701	482	68	414	762	157	605
A11	1107	57	1050	295	18	276	599	125	474
A12	1039	66	973	209	40	169	752	164	588
A13	997	47	950	393	60	333	610	103	507
Haberman	100	1	99	38	27	11	168	53	115
Liver	5	3	2	140	53	87	200	89	111
Pakinsons	89	12	77	35	11	24	71	25	46
Pima	14	8	6	153	59	94	601	201	400
Page-blocks-1-3_vs_4	162	12	150	21	10	11	289	6	283
Vehicle1	30	21	9	280	2	278	536	194	342
Vehicle3	27	15	12	196	14	182	623	183	440
German	23	5	18	162	55	107	815	240	575

Please cite this article as: P. Vorraboot, et al., Improving Classification Rate Constrained to Imbalanced Data Between Overlapped and Non-overlapped Regions by Hybrid Algorithm, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.10.007

5.2. Experimental results

Experimental results consist of two parts. The first part is boundary region determination results indicating the sizes of data in each region for classification phase. The sizes of data in each region effected computational time in classification phase. The second part was classification results. In this part, the results show both classification results and computational times of the classification algorithms.

5.2.1. Boundary area determination results

For the boundary region determination step, data set was partitioned into three regions: obviously non-overlapped, borderline, and overlapped regions. Modified Hausdorff distance, RBFN, and K-means with Mahalanobis distance were responded and applied to identify obviously non-overlapped, borderline, and overlapped regions, respectively. We did this in order to separate data into three regions suitable to classification methods for each group were used for classifying data in each region. These identified regions of data helped reduce the polynomial computational time of the mapping function. The results of the boundary region determination are shown in Table 2. The results indicated that average amount of data removed was 62.45% of the synthetic data set and 36.59% of the standard benchmarked data set. It was seen that the remaining data or data in overlapped region were mapped into a higher dimensional space. In the other word, the smaller number of data points was used for mapping, the less time was required for computation.

5.2.2. Classification results

To be able to find the proper model for constructing the boundary, the training data of 795 groups were generated. All generated data groups were two classes with two dimensions. *IR*, *KL*, *maxF* and distance between center of classes were set as 1–13, 0.04–5.3, 0.6–4.6, and 0.04–0.4, respectively. The number of elements in each group was different sizes varying between 100 and 1000 patterns.

The test sets in each area illustrated in Table 2 were classified by the optimal Soft-Hybrid algorithm. The RBFN was applied to recognize data points in the obviously non-overlapping area. For the borderline area, a density based learning algorithm named dDBSCAN was used. The proposed dDBSCAN method provides the optimal radius and the density for identifying the shape of each subcircle in the borderline area, based on their Eigenvalues $\lambda_2, \lambda_3, \ldots, \lambda_m$.

Finally, data points in overlapping area consisted of two classes such that the location of data points in each class were located near the data points in another class. Therefore it was hard to classify them in their original dimensional space. For this reason, the algorithms in the literatures [12,32,31] proposed kernel learning methods for mapping all data points into a higher dimensional space. This mapping process resulted in the high computational time of those learning algorithms. The proposed algorithm in the classification phase tried to extract data points only in the overlapping area. Then a polynomial kernel function was applied for mapping data points in the overlapping area into a proper position in the higher dimensional space. After that, the RBFN algorithm was used as a learning algorithm. These methods were called the mkernel learning method.

The experiments were conducted by using five-fold cross-validation set to the parameter values. The accuracy of minority class (TPR), *F*-value of minority class (*F*-minor) and *G*-mean were measured as the performance indice of the experiments. The results are shown in Table 3. The results from the RBFN showed that the classification performance, including TPR, *F*-minor and *G*-mean values, in the obvious non-overlapping area, was 100% accurate. The results for those data points in the obviously non-overlapping area were 100% perfectly classifying into the proper classes. This was because of the properties of each data point that pass through the extracting obvious non-overlapping areas which were linearly separable as shown in Fig. 5(b).

Then, the data points in the borderline area were classified by dDBSCAN while the data points in the overlapping areas were classified by a modified Kernel Learning Method. Lastly, the final results were calculated based on the percentage of the sum of the values in each term of measurement in each area. For example, the final TPR value of data set A1 in Table 3 column 14 was computed as the percentage of the sum of the values of RBFN, dDBSCAN, and the mKernel learning methods.

Table 3
The classification results of the Soft-Hybrid.

Data sets	RBFN non-overlapping area				dDBSCAN borderline area				mKernel learning overlapping area				Final results (%)		
	TP	FN	TN	FP	TP	FN	TN	FP	TP	FN	TN	FP	TPR	*F**	*G**
A1	97	0	732	0	64	4	137	20	46	69	808	23	73.93	78.11	84.90
A2	59	0	965	0	85	6	202	10	83	66	491	32	75.92	79.93	86.05
A3	47	0	818	0	57	8	257	9	104	74	598	28	71.72	77.76	83.77
A4	46	0	551	0	112	5	331	10	58	69	788	30	74.48	79.12	85.29
A5	56	0	1012	0	52	7	208	18	105	85	415	42	69.84	73.70	82.08
A6	74	0	1162	0	75	4	71	10	54	63	452	35	75.19	78.38	85.57
A7	37	0	941	0	49	8	251	12	50	60	563	29	66.67	71.39	80.71
A8	82	0	723	0	37	5	270	4	129	69	639	42	77.02	80.52	86.55
A9	68	0	554	0	74	4	441	16	62	62	686	33	75.56	78.01	85.68
A10	55	0	701	0	63	5	405	9	91	66	575	30	74.64	79.17	85.41
A11	57	0	1050	0	17	1	272	4	67	58	443	31	70.50	75.00	83.14
A12	66	0	973	0	33	7	162	7	105	57	545	43	76.12	78.16	85.98
A13	47	0	950	0	47	13	326	6	54	49	488	19	70.48	77.28	83.36
Haberman	1	0	99	0	24	3	7	4	19	34	103	12	54.32	62.41	71.03
Liver	3	0	2	0	32	21	60	27	60	29	82	29	65.52	64.19	68.68
Pakinsons	12	0	77	0	8	3	17	7	15	10	41	5	72.92	73.68	81.83
Pima	8	0	6	0	31	28	69	25	119	82	354	46	58.96	63.58	71.12
Page-blocks-1-3_vs_4	12	0	150	0	14	4	10	5	65	59	283	59	59.09	58.90	71.86
Vehicle1	21	0	9	0	2	0	278	0	80	114	274	68	47.47	53.09	65.06
Vehicle3	15	0	12	0	2	12	168	14	79	104	368	72	45.28	48.73	62.56
German	5	0	18	0	41	38	50	33	130	110	495	80	54.32	57.42	67.26

*G**: *G*-mean. *F**: *F*-value of the minority class.

Please cite this article as: P. Vorraboot, et al., Improving Classification Rate Constrained to Imbalanced Data Between Overlapped and Non-overlapped Regions by Hybrid Algorithm, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.10.007

Table 4
The classification results of the standard classification algorithms.

Data sets	RBFN			SVM1			SVM2			Soft-Hybrid		
	TPR	F^*	G^*	TPR	F^*	G^*	TPR	F^*	G^*	TPR	F^*	G^*
A1	72.29	76.24	83.97	55.71	68.72	74.25	61.07	72.00	77.60	**73.93**	**78.11**	**84.90**
A2	74.00	78.03	84.83	56.00	67.88	74.24	71.33	78.10	83.61	**75.92**	**79.93**	**86.05**
A3	70.13	74.26	80.88	53.45	65.82	72.55	68.62	76.69	82.11	**71.72**	**77.76**	**83.77**
A4	73.45	78.60	84.72	61.38	72.65	77.84	67.93	76.95	81.82	**74.58**	**79.12**	**85.29**
A5	69.65	**74.28**	82.06	52.08	64.17	71.48	66.13	74.33	80.43	**69.84**	73.70	**82.08**
A6	72.96	78.33	84.52	56.67	68.92	74.82	67.41	76.31	81.51	**75.29**	**78.38**	**85.57**
A7	63.24	70.49	78.79	46.08	59.12	67.50	58.82	68.97	76.18	**66.77**	**71.39**	**80.71**
A8	72.98	77.18	84.10	56.21	67.54	74.23	71.12	77.23	83.27	**77.02**	**80.52**	**86.55**
A9	74.81	77.29	85.39	63.70	73.98	79.28	72.22	77.47	84.19	**75.66**	**78.01**	**85.68**
A10	71.79	77.31	83.76	62.86	72.88	78.66	72.14	78.29	84.09	**74.64**	**79.17**	**85.41**
A11	69.00	72.25	82.04	49.50	63.26	70.08	61.00	71.76	77.71	**70.50**	**75.00**	**83.14**
A12	72.67	76.49	82.02	49.63	62.47	69.94	65.93	74.95	80.56	**76.12**	**78.16**	**85.98**
A13	70.05	76.02	82.45	56.19	69.21	74.69	65.24	75.07	80.36	**70.58**	**77.28**	**83.36**
Haberman	18.52	28.04	41.97	2.47	4.76	15.68	0	N/A	0	**54.32**	**62.41**	**71.03**
Liver	65.50	62.50	66.75	0	N/A	0	5.52	10.26	23.31	**65.52**	**64.19**	**68.68**
Pakinsons	70.83	72.34	80.65	50.00	65.75	70.47	58.33	72.73	76.12	**72.92**	**73.68**	**81.83**
Pima	58.21	63.29	70.84	50.37	59.87	67.48	54.48	62.26	69.63	**59.06**	**63.58**	**71.12**
Page-blocks-1-3_vs_4	42.67	49.52	61.03	50.00	58.37	67.40	50.67	**58.89**	67.80	**59.10**	57.90	**71.86**
Vehicle1	42.40	50.69	62.25	16.59	27.91	40.57	18.89	30.15	42.98	**47.57**	**53.09**	**65.06**
Vehicle3	44.34	48.08	61.96	0.94	1.87	9.71	4.25	7.96	20.52	**45.38**	**48.73**	**62.56**
German	52.67	56.23	66.96	52.67	59.74	**68.79**	49.00	56.21	66.09	**54.32**	**57.42**	67.26

G^*: G-mean, F^*: F-value of the minority class.

Table 5
The classification results of the combination of SMOTE and standard classification algorithms.

Data sets	SMOTE+RBFN			SMOTE+SVM1			SMOTE+SVM2			Soft-Hybrid		
	TPR	F^*	G^*	TPR	F^*	G^*	TPR	F^*	G^*	TPR	F^*	G^*
A1	65.35	72.41	76.17	63.28	71.71	75.55	63.05	71.56	75.43	**73.93**	**78.11**	**84.90**
A2	70.67	76.19	78.35	70.28	76.08	78.27	70.42	76.23	78.41	**75.92**	**79.93**	**86.05**
A3	67.42	74.16	77.14	66.85	74.05	77.07	66.91	74.11	77.11	**71.72**	**77.76**	**83.77**
A4	59.82	69.15	73.77	59.26	68.98	73.60	59.18	69.05	73.64	**74.58**	**79.12**	**85.29**
A5	**71.00**	76.25	78.14	70.60	**76.26**	78.20	70.21	76.07	78.05	69.84	73.70	**82.08**
A6	72.75	77.76	79.29	72.74	77.77	79.31	72.81	77.79	79.32	**75.29**	**78.38**	**85.57**
A7	**69.41**	**75.44**	78.22	68.41	74.97	77.84	67.95	74.79	77.70	66.77	71.39	**80.71**
A8	65.40	72.65	76.13	65.55	72.74	76.21	64.62	72.33	75.87	**77.02**	**80.52**	**86.55**
A9	58.85	68.42	73.30	59.31	68.83	73.63	57.62	67.87	72.79	**75.66**	**78.01**	**85.68**
A10	64.08	72.16	75.90	64.23	72.14	75.89	64.10	72.14	75.89	**74.64**	**79.17**	**85.41**
A11	70.57	76.35	78.75	**70.74**	**76.47**	78.85	70.02	76.20	78.64	70.50	75.00	**83.14**
A12	70.93	76.17	78.46	69.61	75.67	78.08	69.10	75.47	77.94	**76.12**	**78.16**	**85.98**
A13	69.56	75.56	78.29	68.05	75.07	77.91	67.90	75.00	77.86	**70.58**	**77.28**	**83.36**
Haberman	**67.04**	**76.19**	**78.91**	60.34	72.48	75.75	65.36	75.48	78.29	54.32	62.41	71.03
Liver	**66.90**	61.39	65.01	11.41	19.32	32.94	8.28	14.46	28.12	65.52	**64.19**	**68.68**
Pakinsons	**84.07**	82.25	84.21	80.53	87.50	**88.51**	79.65	**87.82**	8.64	72.90	73.68	81.83
Pima	54.85	60.87	68.86	52.24	62.08	69.11	57.09	63.22	70.66	**59.06**	**63.58**	**71.12**
Page-blocks-1-3_vs_4	**95.41**	**95.57**	96.24	94.75	97.31	**97.34**	94.75	97.31	7.34	59.19	58.90	71.86
Vehicle1	45.16	51.72	63.76	20.28	32.96	44.82	20.28	32.47	44.68	**47.57**	**53.09**	**65.06**
Vehicle3	43.87	48.19	61.88	1.42	2.78	11.89	6.60	11.81	25.47	**45.38**	**48.73**	**62.56**
German	51.12	54.98	65.70	53.67	60.22	69.12	51.44	57.09	66.95	**54.32**	**57.42**	**67.26**

G^*: G-mean, F^*: F-value of the minority class.

After obtaining the final results as shown in Table 3 columns 11–13, we compared those figures with the classification performance of the three standard learning algorithms: Radial Basis Function Network (RBFN), SVM with Polynomials kernel function degree-2 (SVM1) and SVM with RBF kernel function (SVM2). Moreover, the methods at the data level were combined with these three standard learning algorithms in the experiments, which were the Synthetic Minority Over-sampling Technique (SMOTE), the Random Over Sampling (ROS) and the Random Under Sampling (RUS). The results as shown in Tables 4–7 columns 2–10 are the classification results of the comparative algorithms. Columns 11–13 show the classification results of the proposed method.

Table 4 shows the classification results of the proposed method and those of the three standard classification algorithms: RBFN, SVM1 and SVM2. From the classification results on the synthetic data sets, it was clearly seen that the TPR values of the proposed method for all 13 synthetic data sets were higher than those of the three standard classification algorithms. However, for all 13 synthetic data sets the F-minor values of the proposed method also had higher F-minor values than SVM1 algorithm. But for the RBFN and SVM2 algorithms there was one data set, A5, in which our proposed method offered lower values. For the last performance measurement G-mean, the results showed that the G-mean values of the proposed method for all 13 synthetic data sets had higher G-mean values than those of the three standard

Please cite this article as: P. Vorraboot, et al., Improving Classification Rate Constrained to Imbalanced Data Between Overlapped and Non-overlapped Regions by Hybrid Algorithm, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.10.007

classification algorithms. For the classification results on the standard benchmarking data sets, it was seen that the TPR values of the proposed method for all eight standard benchmarking data sets had higher TPR values than those of the three standard classification algorithms. However, for all eight standard benchmarking data sets the *F*-minor values of the proposed method were higher than those of the RBFN algorithm. But for the SVM1 and SVM2 algorithms there was one standard benchmarking data set, page-blocks-1-3_vs_4, in which our proposed method offered smaller values. For the last performance measurement *G*-mean, the results showed that for all eight standard benchmarking data sets, the proposed method had higher *G*-mean values than the RBFN and SVM2 algorithms. But for the SVM1 algorithm there was one data set, German, in which our proposed method offered a lower value.

Table 5 shows the classification results of the proposed method and the combination of SMOTE and the three standard classification algorithms: SMOTE+RBFN, SMOTE+SVM1 and SMOTE+SVM2. From the classification results on the synthetic data sets, the TPR values of the proposed method for 11 of the synthetic data sets result in higher TPR values than the SMOTE+SVM3 algorithm. But for the SMOTE+RBFN and SMOTE+SVM1 algorithms there were three synthetic data sets, A5, A7 and A11, in which our proposed method offered lesser values. For the SMOTE+SVM2 algorithm there were two data sets, A5 and A7, in which our proposed method offered the smaller values. For the *F*-minor measurement, the *F*-minor values of the proposed method for 10 synthetic data sets were higher than those of the three combinations of SMOTE and the standard classification algorithms. But for those three combinations of SMOTE and the standard classification algorithms, there were three synthetic data sets, A5, A7 and A11, in which our proposed method offered lower values. For the last performance measurement *G*-mean, the results showed that the *G*-mean values of the proposed method for all 13 synthetic data sets were higher than those of the three combination of SMOTE and the standard classification algorithms. For the classification results on standard benchmarking data sets, it was seen that the TPR values of the proposed method for five of the standard benchmarking data sets were higher than those of the SMOTE+SVM1 and SMOTE+SVM2 algorithms, but for SMOTE+RBFN there were four standard benchmarking data sets, Haberman, Liver, Pakinsons and Page-blocks-1-3_vs_4, in which our proposed method gave lower values. For the SMOTE+SVM1 and SMOTE+SVM2 algorithms, there were three standard benchmarking data sets, Haberman, Liver and Pakinsons, in which our proposed method yields smaller values. For the *F*-minor measurement, the *F*-minor values of the proposed method for five of the standard benchmarking data sets were higher than those of the three combinations of SMOTE and the standard classification algorithms. But for those three combinations of SMOTE and the standard classification algorithms there were three standard benchmarking data sets, Haberman, Pakinsons and Page-blocks-1-3_vs_4, in which our proposed method offered lesser values. For the last performance measurement *G*-mean, the *G*-mean values of the proposed method for six of the standard benchmarking data sets were higher than those of the SMOTE+SVM2 algorithm. For the SMOTE+RBFN and SMOTE+SVM1 algorithms there are three standard benchmarking data sets, Haberman, Pakinsons and Page-blocks-1-3_vs_4, in which our proposed method produces lesser values. For the SMOTE+SVM2 algorithm there were two standard benchmarking data sets, Haberman and Pakinsons, in which our proposed method resulted in smaller values.

Table 6 shows the classification results of our proposed method (Soft-Hybrid) and the combination of ROS and the three standard classification algorithms, i.e. ROS+RBFN, ROS+SVM1, and ROS +SVM2. For the classification results on the synthetic data sets, the TPR values of the proposed method for 11 synthetic data sets were higher than those three combinations of ROS and the three standard classification algorithms. But, for those three combinations of ROS and the three standard classification algorithms, there were two data sets, A5 and A7, in which our proposed method offered lesser values. For the F-minor measurement, the F-minor values of the proposed method for 10 synthetic data sets were higher than those of the three combinations of ROS and the three standard classification algorithms. But, for three combinations of ROS and the three standard classification algorithms there were three data sets, A5, A7 and A11, in which our proposed method yielded lower values. For the last performance measurement

Table 6
The classification results of the combination of ROS and standard classification algorithms.

Data sets	ROS+RBFN			ROS+SVM1			ROS+SVM2			Soft-Hybrid		
	TPR	*F**	*G**	TPR	*F**	*G**	TPR	*F**	*G**	TPR	*F**	*G**
A1	65.45	72.45	76.20	63.69	71.92	75.73	63.24	71.74	75.57	**73.93**	**78.11**	**84.90**
A2	70.89	76.17	78.30	70.65	76.21	78.36	70.08	76.08	78.29	**75.92**	**79.93**	**86.05**
A3	68.07	74.4	77.33	66.63	73.94	76.97	66.60	73.97	77.00	**71.72**	**77.76**	**83.77**
A4	60.22	69.45	74.02	59.41	69.11	73.70	59.47	69.19	73.76	**74.58**	**79.12**	**85.29**
A5	**71.27**	**76.45**	78.31	71.00	76.34	78.23	70.60	76.17	78.05	69.84	73.70	**82.08**
A6	73.27	77.87	79.33	72.86	77.74	79.25	72.85	77.84	79.36	**75.29**	**78.38**	**85.57**
A7	**68.24**	74.92	77.80	68.55	**75.04**	77.89	67.95	74.79	77.77	66.77	71.39	**80.71**
A8	66.02	72.88	76.32	65.69	72.72	76.20	64.71	72.36	75.90	**77.02**	**80.52**	**86.55**
A9	59.67	68.92	73.73	59.37	68.78	73.60	59.03	68.67	73.49	**75.66**	**78.01**	**85.68**
A10	63.97	72.14	75.88	64.44	72.30	76.02	63.84	72.03	75.79	**74.64**	**79.17**	**85.41**
A11	70.49	**76.35**	78.75	70.40	76.22	78.64	70.25	76.27	78.69	**70.50**	75.00	**83.14**
A12	70.86	76.04	78.34	69.75	75.73	78.13	69.49	75.59	78.02	**76.12**	**78.16**	**85.98**
A13	68.74	75.33	78.11	67.88	75.02	77.87	67.69	75.00	77.85	**70.58**	**77.28**	**83.36**
Haberman	**66.48**	**75.32**	**78.21**	58.66	70.47	74.17	65.36	75.48	78.29	54.32	62.41	71.03
Liver	51.01	55.27	62.10	2.68	5.23	16.38	17.45	27.66	40.43	**65.52**	**64.19**	**68.68**
Pakinsons	76.99	86.14	87.15	80.53	88.78	89.43	**81.42**	**89.32**	**89.92**	72.92	73.68	81.83
Pima	**59.70**	**64.00**	**71.51**	52.24	61.95	69.04	52.99	61.61	68.99	59.06	63.58	71.12
Page-blocks-1-3_vs_4	**95.74**	**96.05**	**96.63**	94.75	97.14	97.23	94.75	97.14	97.23	59.19	57.90	71.86
Vehicle1	44.70	52.01	63.71	16.13	27.13	39.97	10.14	18.03	31.72	**47.57**	**53.09**	**65.06**
Vehicle3	**57.55**	**54.34**	**68.67**	0	NaN	0	5.66	10.53	23.72	45.38	48.73	62.56
German	50.16	53.49	64.63	52.72	59.89	68.72	**55.91**	**61.73**	**70.43**	54.32	57.42	67.26

*G**: *G*-mean, *F**: *F*-value of the minority class.

Please cite this article as: P. Vorraboot, et al., Improving Classification Rate Constrained to Imbalanced Data Between Overlapped and Non-overlapped Regions by Hybrid Algorithm, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.10.007

Table 7
The classification results of the combination of RUS and standard classification algorithms.

Data sets	RUS+RBFN			RUS+SVM1			RUS+SVM2			Soft-Hybrid		
	TPR	F*	G*	TPR	F*	G*	TPR	F*	G*	TPR	F*	G*
A1	71.43	77.97	83.22	60.71	72.65	77.27	62.50	73.22	78.21	**73.93**	**78.11**	**84.90**
A2	73.00	78.07	83.15	67.33	75.94	80.49	71.33	78.39	82.74	**75.92**	**79.93**	**86.05**
A3	**72.07**	**77.99**	83.20	64.48	74.21	79.14	70.00	77.48	82.27	71.72	77.76	**83.77**
A4	**75.86**	**80.00**	85.64	66.21	75.59	80.49	68.62	77.28	81.94	74.58	79.12	**85.29**
A5	**71.88**	**75.76**	81.47	62.30	71.82	77.02	69.01	75.00	80.36	69.84	73.70	**82.08**
A6	73.70	78.66	83.34	64.81	75.43	79.31	69.63	77.21	81.54	**75.29**	**78.38**	**85.57**
A7	66.67	**72.92**	80.13	49.02	61.35	69.15	59.80	70.32	76.42	**66.77**	71.39	**80.71**
A8	75.16	79.08	84.66	70.81	77.03	82.43	70.50	76.82	82.24	**77.02**	**80.52**	**86.55**
A9	73.70	78.66	84.56	68.89	76.39	81.96	70.37	77.24	82.80	**75.66**	**78.01**	**85.68**
A10	73.93	79.16	84.53	65.36	74.09	79.67	69.64	76.92	82.20	**74.64**	**79.17**	**85.41**
A11	68.50	73.85	81.00	48.5	62.78	69.12	60.50	71.18	76.86	**70.50**	**75.00**	**83.14**
A12	72.59	77.01	82.95	64.07	73.62	78.72	66.67	74.53	80.00	**76.12**	**78.16**	**85.98**
A13	69.05	75.13	81.63	56.67	69.39	74.68	60.48	71.75	77.02	**70.58**	**77.28**	**83.36**
Haberman	41.98	55.28	62.71	43.21	55.12	62.82	38.27	49.21	58.35	**54.32**	**62.41**	**71.03**
Liver	54.48	59.18	65.40	7.59	13.75	27.27	6.21	11.32	24.60	**65.50**	**64.19**	**68.68**
Pakinsons	**89.58**	**74.14**	78.91	58.33	73.67	76.38	58.33	73.67	76.38	72.92	73.68	**81.83**
Pima	52.99	59.41	67.66	54.10	62.63	69.78	54.85	63.91	70.65	**59.06**	**63.58**	**71.12**
Page-blocks-1-3_vs_4	**71.43**	**78.43**	**83.75**	42.86	60.00	65.47	50.00	66.67	70.71	59.19	57.90	71.86
Vehicle1	**51.15**	**54.68**	**66.94**	6.91	12.88	26.27	18.43	30.42	42.73	47.57	53.09	65.06
Vehicle3	35.38	42.98	56.50	1.89	3.70	13.74	8.49	15.19	28.98	**45.38**	**48.73**	**62.56**
German	46.67	52.43	63.47	53.33	**59.81**	**68.93**	49.33	55.74	65.86	**54.32**	57.42	67.26

G*: G-mean, F*: F-value of the minority class.

G-mean, the results show that the G-mean values of the proposed method for all 13 synthetic data sets were higher than those of the three combinations of ROS and the three standard classification algorithms. For the classification results on standard benchmarked data sets, it was that the TPR values of the proposed method for five of the standard benchmarked data sets had higher TPR values than those of the ROS+RBFN. But, for the ROS+RBFN algorithm there were five standard benchmarked data sets, Haberman, Pakinsons, Pima, Page-blocks-1-3 vs 4 and Vehicle3, in which our proposed method gave lesser values, and for the ROS+SVM1 algorithms there were three standard benchmarked data sets, Haberman, Pakinsons and Page-blocks-1-3 vs 4, in which our proposed method offered smaller values. For the ROS+SVM2 there were four standard benchmarked data sets, Haberman, Pakinsons, Page-blocks-1-3 vs 4 and German, in which our proposed method resulted in lower values. For the F-minor measurement, the F-minor values of the proposed method for four standard benchmarked data sets were higher than those of the ROS+SVM1 and ROS+SVM2 algorithms. But, for the ROS+RBFN algorithm there were five standard benchmarked data sets, Haberman, Pakinsons, Pima, Page-blocks-1-3 vs 4 and Vehicle3, in which our proposed method offered lesser values. For the ROS+SVM1 and ROS+SVM2 there were four standard benchmarked data sets, Haberman, Pakinsons, Page-blocks-1-3 vs 4 and German, in which our proposed method yields smaller values. For the last performance measurement G-mean, the G-mean values of the proposed method for four standard benchmarked data sets were higher than those of the ROS+SVM1 and ROS+SVM2 algorithms. But, for the ROS+RBFN algorithm there were five standard benchmarked data sets, Haberman, Pakinsons, Pima, Page-blocks-1-3 vs 4 and Vehicle3, in which our proposed method offered lower values, and for ROS+SVM1and ROS+SVM2 there were four standard benchmarked data sets, Haberman, Pakinsons, Page-blocks-1-3 vs 4 and German, in which our proposed method gave smaller values.

Table 7 shows the classification results of the proposed method and the combination of RUS and the three standard classification algorithms: RUS+RBFN, RUS+SVM1 and RUS+SVM2. From the classification results on the synthetic data sets, the TPR values of the proposed method for all 13 synthetic data sets were higher than the TPR values of the RUS+SVM1 and RUS+SVM2 algorithms. But for the RUS+RBFN algorithm there are three data sets, A3, A4 and A5, in which our proposed method offered lesser values. For the F-minor measurement, the F-minor values of the proposed method for 13 synthetic data sets were higher than the F-minor values of the RUS+SVM1 algorithms. But for the RUS+RBFN algorithm there were four synthetic data sets, A3, A4, A5 and A7, in which our proposed method produces lower values and for RUS+SVM2 there was one synthetic data set, A5, in which our proposed method offered a smaller value. For the last performance measurement G-mean, the results show that the G-mean values of the proposed method for all 13 synthetic data sets were higher than those of the three combinations of RUS and the three standard classification algorithms. For the classification results on standard benchmarking data sets, it was seen that the TPR values of the proposed method for seven of the standard benchmarking data sets were higher than the TPR values of the RUS+SVM1 and RUS+SVM2 algorithms. But for RUS+RBFN there

Table 8
The average percentage of the improvement.

Methods	Synthetic data sets			Standard Benchmarking data sets		
	TPR	F*	G*	TPR	F*	G*
RBFN	3.27	2.01	1.77	31.65	19.34	12.21
SVM1	33.14	15.11	14.65	1012.41	546.09	140.52
SVM2	8.75	2.94	4.33	49.25	159.95	66.90
SMOTE+RBFN	9.26	4.76	9.72	−6.28	−6.21	−2.96
SMOTE+SVM1	10.10	4.96	9.88	458.15	234.53	67.53
SMOTE+SVM2	10.80	5.20	10.09	169.67	81.03	36.68
ROS+RBFN	7.57	3.26	8.96	−5.18	−6.96	−4.19
ROS+SVM1	8.03	3.42	9.08	356.62	164.85	49.14
ROS+SVM2	8.69	3.61	9.23	160.88	76.43	36.62
RUS+RBFN	1.55	0.22	1.76	7.78	2.85	3.35
RUS+SVM1	18.73	7.36	9.05	468.71	237.97	85.56
RUS+SVM2	9.80	3.03	5.21	206.14	97.50	47.58

G*: G-mean, F*: F-value of the minority class.

Please cite this article as: P. Vorraboot, et al., Improving Classification Rate Constrained to Imbalanced Data Between Overlapped and Non-overlapped Regions by Hybrid Algorithm, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.10.007

were three standard benchmarking data sets, Haberman, Page-blocks-1-3_vs_4 and Vehicle1, in which our proposed method gave lower values. For the RUS+SVM1 and RUS+SVM2 algorithms there was one standard benchmarking data set, Page-blocks-1-3_vs_4, in which our proposed method offered a smaller value. For the *F*-minor measurement, the *F*-minor values of the proposed method for seven of the standard benchmarking data sets were higher than the *F*-minor values of the RUS+SVM2 algorithm. But for the RUS+RBFN algorithm there were three standard benchmarking data sets, Pakinsons, Page-blocks-1-3_vs_4 and Vehicle1, in which our proposed method offered lesser values. For the RUS+SVM1 there were two standard benchmarking data sets, Pakinsons and German, in which our proposed method yields lower values, and for RUS+SVM2 there was one standard benchmarking data set, Page-blocks-1-3_vs_4, in which our proposed method generated smaller values. For the last performance measurement *G*-mean, the *G*-mean values of the proposed method for all eight of the standard benchmarking data sets were higher than those from the RUS+SVM2 algorithm. But for the RUS+RBFN algorithm there were two standard benchmarking data sets, Page-blocks-1-3_vs_4 and Vehicle1, in which our proposed method offered lesser values. For RUS+SVM1 there was one standard benchmarking data set, German, in which our proposed method produces a smaller value.

Table 8 shows the average percentage of the improvement of the resulting values of the proposed method compared with the 12 classification algorithms: RBFN, SVM1, SVM2, SMOTE+RBFN, SMOTE+SVM1, SMOTE+SVM2, ROS+RBFN, ROS+SVM1, ROS+SVM2, RUS+RBFN, RUS+SVM1 and RUS+SVM2.

From the average percentage of the improvements of the results of the synthetic data sets as shown in Table 8, it was seen that the average improvements in the accuracy percentage of TPR values of the proposed method were 3.27%, 33.14%, 8.75%, 9.26%, 10.10%, 10.80%, 7.57%, 8.03%, 8.69%, 1.55%, 18.73% and 9.80% as compared with those 12 standard classification algorithms. These figures indicated that the proposed method improved the accuracy rate of classification on minority classes. In other words, it reduced the bias on classification by the majority class. The domination of instances of the majority class leads to a bias in the learning process that was geared towards the error value that mostly was gathered from members of the majority class. Such domination by the majority class resulted in a lower accuracy in the rate of classification for the minority classes. Likewise, the average percentage incremental improvements of results of *F*-minor values of the proposed method were 2.01%, 15.11%, 2.94%, 4.76%, 4.96%, 5.20%, 3.26%, 3.42%, 3.61%, 0.22%, 7.36% and 3.03% as compared with those 12 standard classification algorithms. The better values of *F*-minor indicate that the performance of the precision and recall measurements of the minority classes were improved. For the last performance measurement, the average percentage of the improvement of the results for the *G*-mean values of the proposed method was 1.77%, 14.65%, 4.33%, 9.72%, 9.88%, 10.09%, 8.96%, 9.08%, 9.23%, 1.76%, 9.05% and 5.21% as compared with those 12 standard classification algorithms. *G*-mean values measure the classification performance of both the majority and the minority class. The proposed method offered better performance.

For the standard benchmarking data sets, the average percentage of the improvement of the results of the TPR values of the proposed method was 31.65%, 1012.41%, 49.25%, −6.28%, 458.15%, 169.67%, −5.18%, 356.62%, 160.88%, 7.78%, 468.71% and 206.14% as compared with those 12 classification algorithms. These figures indicated that the proposed method improved the accuracy rate of classification for the minority classes. In other words, it reduced the bias for classification by the majority class. The domination of elements of the majority class led to a bias in the learning process that was geared towards the error value that was mostly gathered from members of the majority class. Such domination by the majority class resulted in a lower accuracy rate of classification for the minority class. The average percentage of the improvement of the results of the *F*-minor values of the proposed method was 19.34%, 546.09%, 159.95%, −6.21%, 234.53%, 81.03%, −6.96%, 164.85%, 76.43%, 2.85%, 237.97% and 97.50% as compared with those 12 classification algorithms. The better *F*-minor values indicated that the performance of the precision and recall measurements of the minority classes was improved. For the last performance measurement, the average percentage of the improvement in the results for the *G*-mean values of the proposed method was 12.21%, 140.52%, 66.90%, −2.96%, 67.53%, 36.68%, −4.19%, 49.14%, 36.62%, 3.35%, 85.56% and 47.58% as compared with those 12 classification algorithms. *G*-mean values measure the classification performance of both the majority and

Table 9
The average computational time.

Methods	Times (ms.)	
	Synthetic data sets	Standard benchmarking data sets
Time1		
SMOTE+RBFN	4.20	**4.94**
SMOTE+SVM1	5.30	5.13
SMOTE+SVM2	4.35	5.14
SMOTE+RBFN	5.83	5.23
SMOTE+SVM1	8.10	5.36
SMOTE+SVM2	6.02	5.41
ROS+RBFN	5.97	5.40
ROS+SVM1	8.12	5.63
ROS+SVM2	6.21	5.58
Soft-Hybrid	**4.06**	5.10
Time2		
RUS+RBFN	2.54	4.53
RUS+SVM1	3.13	4.77
RUS+SVM2	2.68	4.77
Soft-Hybrid	**2.36**	**4.44**

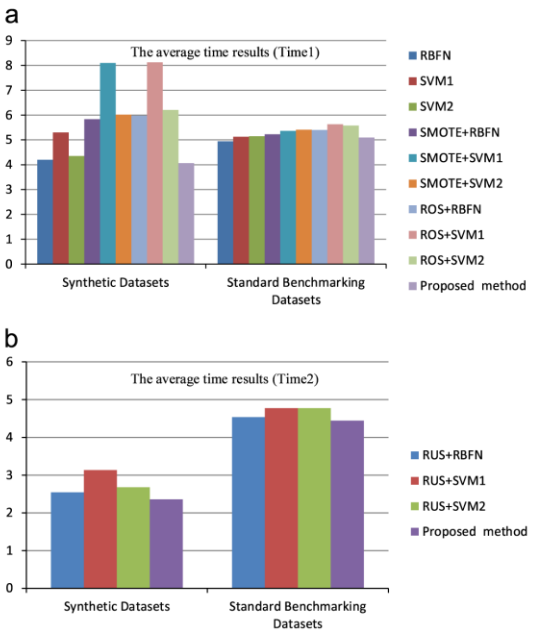


Fig. 10. The average time results. (a) Time1. (b) Time2.

Please cite this article as: P. Vorraboot, et al., Improving Classification Rate Constrained to Imbalanced Data Between Overlapped and Non-overlapped Regions by Hybrid Algorithm, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.10.007

minority classes. The proposed method was proven to offer better performance.

The average computational time of the proposed method was divided into two groups: the computational time for all data points in three areas (borderline, overlapping and obviously non-overlapping areas), called the Time1, and the computational time for data points in the borderline and overlapping areas, called the Time2. The computational time for Time1 of the proposed method was compared with times for nine other classification algorithms: RBFN, SVM1, SVM2, SMOTE +RBFN, SMOTE+SVM1, SMOTE+SVM2, ROS+RBFN, ROS+SVM1 and ROS+SVM2. The computational time for the Time2 of the proposed method was compared with times required for three other algorithms: RUS+RBFN, RUS+SVM1 and RUS+SVM2.

Table 9 shows the average computational time. For Time1, it was seen that the average times of the proposed method on synthetic data sets were lower than nine other classification algorithms, as shown in Fig. 10(a). For the standard benchmarking data sets, the average computational time of the proposed method for Time1 was lower than that of all eight other classification algorithms. But there was one classification algorithm, RBFN, where our proposed method required a higher average computational time.

For the Time2, it was seen that the average time of the proposed method on synthetic data sets and standard benchmarking data sets had a lower average computational time than all three other classification algorithms, as shown in Fig. 10(b).

6. Conclusions

This research work, proposed modified hybrid algorithms for improving the classification rate of highly imbalanced large data sets with overlapping areas. The proposed methods have been divided into two main phases. The first phase is the boundary area determination. In this phase, we study and develop the solution using three different techniques. The first technique uses a modified Hausdorff distance to find elements that are in the obvious non-overlapping area. The second technique uses RBFN to find elements that are in the borderline area. The third technique uses K-means clustering with Mahalanobis distance to find elements in the overlapping area. For the second phase of responsive data classification, we performed an experimental study and developed three different methods for classifying each data area. These methods are RBFN, dynamic DBSCAN, and the modified kernel Learning method.

The experimental results illustrate that the proposed method significantly improves the effectiveness in classifying imbalanced data with overlaps using TPR, *F*-minor and *G*-mean as performance measurement techniques. Most of the average percentage results of all the performance measurement are improved compared with the 12 other classification algorithms. For RBFN, SVM1, SVM2, SMOTE+RBFN, SMOTE+SVM1, SMOTE+SVM2, ROS+RBFN, ROS+SVM1, ROS+SVM2, RUS+RBFN, RUS+SVM1 and RUS+SVM2, for the synthetics data sets, the TPR values were improved by 3.27%, 33.14%, 8.75%, 9.26%, 10.10%, 10.80%, 7.57%, 8.03%, 8.69%, 1.55%, 18.73% and 9.80%, respectively. The *F*-minor values were improved by 2.01%, 15.11%, 2.94%, 4.76%, 4.96%, 5.20%, 3.26%, 3.42%, 3.61%, 0.22%, 7.36% and 3.03%. The *G*-mean values were improved by 1.77%, 14.65%, 4.33%, 9.72%, 9.88%, 10.09%, 8.96%, 9.08%, 9.23%, 1.76%, 9.05% and 5.21%. The average percentage results of all the performance measurements compared with those of the 12 other classification algorithms for the eight standard benchmarking data sets are also improved. The TPR values were improved by 31.65%, 1012.41%, 49.25%, −6.28%, 458.15%, 169.67%, −5.18%, 356.62%, 160.88%, 7.78%, 468.71% and 206.14%. The *F*-minor values were improved by 19.34%, 546.09%, 159.95%, −6.21%, 234.53%, 81.03%, −6.96%, 164.85%, 76.43%, 2.85%, 237.97% and 97.50%. The *G*-mean values were improved by 12.21%, 140.52%, 66.90%, −2.96%, 67.53%, 36.68%, −4.19%, 49.14%, 36.62%, 3.35%, 85.56% and 47.58%.

In addition, as a result of boundary areas determination and the responsive classification algorithms consume less computational time for mapping overlapping data points into their higher dimensional space. The number of mapping points of synthetic data and standard benchmarking data was 37.55% and 63.41% on average of all points, respectively.

Moreover, the results of the average computational time show that the proposed method has a lower average computational time compared to most of the other classification algorithms on synthetic data sets and standard benchmarking data sets. There is only one classification algorithm of Time1, RBFN, in which our proposed method has a higher average computational time on the standard benchmarking data set. The improvement of Time1 was measured to be 29.21% on average for synthetic data and 3.87% on average for standard benchmarking data. And the improvement of Time2 was measured to be 27.06% on average for synthetic data and 6.18% on average for standard benchmarking data.

However, when using our Soft-Hybrid Algorithm on non-synthetic data, one must ensure that it actually has a Gaussian distribution using the well-known Kurtosis test. For this research work, only two class classification problems were studied, but there are also some related interesting problems that we should study in the future. These include multiple class problems with higher imbalanced ratios, overlapping ratios and Kullback–Leibler divergence.

Acknowledgments

This research was supported by grant funds from the Office of the Higher Education Commission, Ministry of Education (Thailand), as part of its program called the Program Strategic Scholarships for Frontier Research Network for the Ph.D. Program Thai Doctoral degree and the Research Grant of Burapha University through National Research Council of Thailand (Grant no. 80/2556).

References

[1] N.V. Chawla, N. Japkowicz, A. Kotcz, Editorial: special issue on learning from imbalanced data sets, ACM Sigkdd Explor. Newslett. 6 (1) (2004) 1–6.
[2] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, Artif. Intell. Res. 16 (2002) 321–357.
[3] P. Soda, A multi-objective optimisation approach for class imbalance learning, Pattern Recognit. 44 (8) (2011) 1801–1810.
[4] T. Liu, Y. Liang, W. Ni, Minority identification for imbalanced dataset, in: Control Conference (CCC), 2012 31st Chinese, IEEE, 2012, pp. 3897–3902.
[5] S. Kotsiantis, D. Kanellopoulos, P. Pintelas, et al., Handling imbalanced datasets: a review, GESTS Int. Trans. Comput. Sci. Eng. 30 (1) (2006) 25–36.
[6] X. Guo, Y. Yin, C. Dong, G. Yang, G. Zhou, On the class imbalance problem, in: Fourth International Conference on Natural Computation, 2008. ICNC'08, vol. 4, IEEE, 2008, pp. 192–201.
[7] R. Barandela, J.S. Sánchez, V. Garcıa, E. Rangel, Strategies for learning in class imbalance problems, Pattern Recognit. 36 (3) (2003) 849–851.
[8] K. Li, X. Kong, Z. Lu, L. Wenyin, J. Yin, Boosting weighted ELM for imbalanced learning, Neurocomputing 128 (2014) 15–21.
[9] S. Chen, G. Guo, L. Chen, A new over-sampling method based on cluster ensembles, in: 2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops (WAINA), IEEE, 2010, pp. 599–604.
[10] H. Han, W.-Y. Wang, B.-H. Mao, Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning, in: Advances in Intelligent Computing, Springer, 2005, pp. 878–887.
[11] R. Batuwita, V. Palade, Efficient resampling methods for training support vector machines with imbalanced datasets, in: The 2010 International Joint Conference on Neural Networks (IJCNN, 2010), IEEE, 2010, pp. 1–8.
[12] H. He, A. Ghodsi, Rare class classification by support vector machine, in: 2010 20th International Conference on Pattern Recognition (ICPR), IEEE, 2010, pp. 548–551.

Please cite this article as: P. Vorraboot, et al., Improving Classification Rate Constrained to Imbalanced Data Between Overlapped and Non-overlapped Regions by Hybrid Algorithm, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.10.007

[13] G.-S. Xiao, X.-Y. Chen, Graph classification with imbalanced data sets, in: 2011 First Asian Conference on Pattern Recognition (ACPR), IEEE, 2011, pp. 57–61.
[14] K. Boonchuay, K. Sinapiromsaran, C. Lursinsap, Minority split and gain ratio for a class imbalance, in: 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), vol. 3, IEEE, 2011, pp. 2060–2064.
[15] R. Alejo, J.M. Sotoca, V. García, R.M. Valdovinos, Back propagation with balanced MSE cost function and nearest neighbor editing for handling class overlap and class imbalance, in: Advances in Computational Intelligence, Springer, 2011, pp. 199–206.
[16] S.-H. Oh, Error back-propagation algorithm for classification of imbalanced data, Neurocomputing 74 (6) (2011) 1058–1061.
[17] W. Zong, G.-B. Huang, Y. Chen, Weighted extreme learning machine for imbalance learning, Neurocomputing 101 (2013) 229–242.
[18] R. Batuwita, V. Palade, Class Imbalance Learning Methods for Support Vector Machines, John Wiley & Sons, Inc. (2013) http://dx.doi.org/10.1002/9781118646106.ch5, pp. 83–99.
[19] R. Batuwita, V. Palade, Fsvm-cil: fuzzy support vector machines for class imbalance learning, IEEE Trans. Fuzzy Syst. 18 (3) (2010) 558–571.
[20] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995.
[21] R. Alejo, R.M. Valdovinos, V. García, J. Pacheco-Sanchez, A hybrid method to face class overlap and class imbalance on neural networks and multi-class scenarios, Pattern Recognit. Lett. 34 (4) (2013) 380–388.
[22] S.-J. Lin, C. Chang, M.-F. Hsu, Multiple extreme learning machines for a two-class imbalance corporate life cycle prediction, Knowledge-Based Syst. 39 (2013) 214–223.
[23] K. Napierala, J. Stefanowski, BRACID: a comprehensive approach to learning rules from imbalanced data, J. Intell. Inf. Syst. 39 (2) (2012) 335–373.
[24] M. Farquad, I. Bose, Preprocessing unbalanced data using support vector machine, Dec. Support Syst. 53 (1) (2012) 226–233.
[25] M. Gao, X. Hong, S. Chen, C.J. Harris, A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems, Neurocomputing 74 (17) (2011) 3456–3466.
[26] V. García, R.A. Mollineda, J.S. Sánchez, On the k-NN performance in a challenging scenario of imbalance and overlapping, Pattern Anal. Appl. 11 (3-4) (2008) 269–280.
[27] J. Luengo, A. Fernández, F. Herrera, Addressing data-complexity for imbalanced data-sets: a preliminary study on the use of preprocessing for C4. 5, in: Ninth International Conference on Intelligent Systems Design and Applications, 2009. ISDA'09, IEEE, 2009, pp. 523–528.
[28] F.F. Hausdorff, Grundzge der mengenlehre, Von Veit, Leipzig, 1914.
[29] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Stat. (1951) 79–86.
[30] P.C. Mahalanobis, On the generalized distance in statistics, Proc. Natl. Inst. Sci. (Calcutta) 2 (1936) 49–55.
[31] J.P. Hwang, S. Park, E. Kim, A new weighted approach to imbalanced data classification problem via support vector machine with quadratic cost function, Expert Syst. Appl. 38 (7) (2011) 8580–8585.
[32] X. Fan, Z. He, A fuzzy support vector machine for imbalanced data classification, in: 2010 International Conference on Optoelectronics and Image Processing (ICOIP), vol. 1, IEEE, 2010, pp. 11–14.
[33] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: KDD, vol. 96, 1996, pp. 226–231.
[34] UCI Machine Learning Repository, ⟨http://www.ics.uci.edu/ mlearn/MLRepository.html⟩.
[35] KEEL (Knowledge Extraction based on Evolutionary Learning), KEEL-Dataset, ⟨http://www.keel.es⟩.

Piyanoot Vorraboot received the B.Sc. degree in management information systems from Mahasarakham University, Mahasarakham, Thailand, in 1999 and the M.Sc. degree in information technology from King Mongkut's Institute of Technology North Bangkok, Bangkok, Thailand, in 2004. Currently she is a Ph.D. student in computer science at Burapha University, Chonburi, Thailand.

Suwanna Rasmequan received the B.B.A. degree in finance and banking and the M.Sc. degree in computer information system from Assumption University, Bangkok, Thailand, in 1992 and 1994, respectively. She received Ph.D. degrees in computer science in 2002 from University of Warwick, Coventry, United Kingdom. She worked in the business sector from 1984 until 1997 in two major businesses namely Packaging Industry and Satellite Communication Provider. Her work responsibilities in those businesses was starting with the beginning post of Executive Secretary and ending with the post of Section Manager. She was a Lecturer at the Department of Computer Science, Burapha University, Chonburi, Thailand during 1997–2006. She has been working as an Assistant Professor from year 2006 up to present at the Faculty of Informatics, Burapha University which is the former the Department of Computer Science, Faculty of Science, Burapha University. Her major research interests include empirical modeling, decision support system, machine learning and their applications to other science and business areas.

Krisana Chinnasarn received the B.Sc. in statistics from Srinakharinwirot University Mahasarakham Campus, Thailand, in 1992. He then received M.Sc. in computer science and information technology from King Mongkut's Institute of Technology Ladkrabang, Thailand, in 1997. He finally received Ph.D. in computer science from Chulalongkorn University, Thailand, in 2004. During 1996–1997, he worked as a Research Assistant in the Computer Center, King Mongkut Institute of Technology Ladkrabang, Thailand. During 2002–2003, he visited Oxford University Computing Laboratory, Oxford, England, as a Ph.D. Visiting student. From 1997 to presents, he is a lecturer in Faculty of Informatics, Burapha University which is a former Department of Computer Science, Faculty of Science, Burapha University. He has been appointed as an Assistant Professor in Computer Science since 2006. His major research interests include machine learning and digital image processing and their applications to other science and engineering areas.

Chidchanok Lursinsap received the B.Eng. degree (honors) in computer engineering from Chulalongkorn University, Bangkok, Patumwan, Thailand, in 1978 and the M.S. and Ph.D. degrees in computer science from the University of Illinois at Urbana-Champaign, Urbana, in 1982 and 1986, respectively. He was a Lecturer at the Department of Computer Engineering, Chulalongkorn University, in 1979. In 1986, he was a Visiting Assistant Professor at the Department of Computer Science, University of Illinois at Urbana-Champaign. From 1987 to 1996, he worked at The Center for Advanced Computer Studies, University of Louisiana at Lafayette, as an Assistant and Associate Professor. After that, he came back to Thailand to establish Ph.D. program in computer science at Chulalongkorn University and became a Full Professor. His major research interests include neural learning and its applications to other science and engineering areas.

Please cite this article as: P. Vorraboot, et al., Improving Classification Rate Constrained to Imbalanced Data Between Overlapped and Non-overlapped Regions by Hybrid Algorithm, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.10.007