การเลือกลักษณะเด่นด้วยราฟเซตบนพื้นฐานข่าวสารที่เล็กที่สุดของบริเวณขอบเขต

Feature Selection Based on Minimal Boundary and Maximal Lower Approximation

สมบัติ ฝอยทอง[1] ธารารัตน์ พวงสุวรรณ[2] วรวิทย์ พูลสวัสดิ์[2] และไพฑูรย์ ศรีนิล[2]

บทคัดย่อ

การเลือกลักษณะเด่น (FS) นำมาประยุกต์ใช้กับการลดมิติและใช้เลือกซับเซตของคุณลักษณะเริ่มต้นของชุดข้อมูลซึ่งซับเซตนี้ต้องมีประสิทธิภาพการทำนายมากที่สุด วิธีการ FS โดยส่วนใหญ่บนหลักการพื้นฐานของทฤษฎีราฟเซตจะให้ความสำคัญกับฟังก์ชั่นการขึ้นต่อกัน (Dependency function) ใช้เป็นวัดความดีของซับเซตลักษณะเด่น แต่อย่างไรก็ตามการพิจารณาเฉพาะข่าวสารจากบริเวณเชิงบวก (Positive region) และไม่สนใจต่อข่าวสารของบริเวณขอบเขต (Boundary region) จะทำให้ข่าวสารที่มีความสำคัญอย่างมากสูญหายได้ ในงานวิจัยนี้เราได้นำเสนอการเลือกลักษณะเด่นบนหลักการของราฟเซตโดยตัวแปรเที่ยงตรง (VPRS) และมิวชวลอินฟอร์เมชั่น (Mutual information) ด้วยการใช้กฎเกณฑ์ของบริเวณขอบเขตที่เล็กที่สุด กฎเกณฑ์นี้ใช้หาค่า β ที่เหมาะสมอย่างอัตโนมัติแทนที่จะเป็นการรับเข้ามาจากมนุษย์ ซับเซตของลักษณะเด่นเลือกจากค่าความแตกต่างที่มากที่สุดระหว่างข่าวสารการประมาณขอบเขตล่างและข่าวสารที่บรรจุในบริเวณขอบเขต วิธีการที่นำเสนอนี้สามารถให้ค่าความถูกต้องในการจำแนกประเภทสูงกว่าผลลัพธ์ที่ได้รับจากหลักการของบริเวณเชิงบวกอย่างเดียว ผลการทดลองได้แสดงบนข้อมูลแบบไม่ต่อเนื่องและแบบต่อเนื่อง และมีการเปรียบเทียบในส่วนของ ขนาดซับเซต เวลาที่ใช้และความถูกต้องการจำแนกประเภท เมื่อเทียบกับวิธีการ FS อื่นด้วย

คำสำคัญ : บริเวณขอบเขต การจำแนกประเภทข้อมูล การเลือกลักษณะเด่น มิวชวลอินฟอร์เมชั่น ราฟเซต

ABSTRACT

Feature selection (FS) is an important preprocessing step for many applications in artificial intelligence. FS is applied to dimensionality reduction, which is accomplished by selecting a subset of the original features of a data set that possesses the most predictive performance. Most existing FS methods are based on a rough set theory focusing on dependency function, based on lower approximation, for measuring the goodness of the feature subset. However, by determining only information from a positive region but neglecting a boundary region, much of the relevant information could be invisible. This paper, using the

[1] อาจารย์ ดร. คณะวิทยาศาสตร์และศิลปศาสตร์ มหาวิทยาลัยบูรพา วิทยาเขตจันทบุรี

[2] อาจารย์ คณะวิทยาศาสตร์และศิลปศาสตร์ มหาวิทยาลัยบูรพา วิทยาเขตจันทบุรี

maximum lower approximation - minimum boundary region criterion, focuses on feature selection methods based on rough sets and mutual information, which use different values for the lower approximation information and the information contained in the boundary region. The use of this criterion can result in higher predictive accuracy than data obtained using the measure based on the positive region alone. This demonstrates that most of the relevant information can be extracted by using this criterion. Experimental results are illustrated for crisp and real valued data and are compared with other FS methods in terms of subset size, runtime, and classification accuracy.

Keywords : boundary region, classification, feature selection, mutual information, rough set

บทนำ

การเลือกลักษณะเด่น (FS) เป็นเทคนิคเพื่อลดจำนวนของคุณลักษณะโดยการลบคุณลักษณะที่ไม่สัมพันธ์หรือมีความซ้ำซ้อน และทำให้เกิดผลดีที่ตามมา ได้แก่ ความเร็วที่เพิ่มขึ้นของอัลกอริทึมการเรียนรู้ ปรับปรุงความถูกต้องการทำนาย ความสามารถในการทำความเข้าใจผลลัพธ์ FS เป็นกระบวนการเลือกซับเซตของคุณลักษณะของเซตข้อมูลและนอกจากนี้ข่าวสารที่สำคัญที่สุดของเซตข้อมูลก็ยังคงรักษาไว้เหมือนเดิม FS ได้มีการขยายไปสู่หลายสาขาของงานวิจัยได้แก่ การเรียนรู้ของเครื่องจักร (Blum and Langley, 1997; Kohavi and John, 1997) และการทำเหมืองข้อมูล (Dash and Liu, 1997) และประยุกต์ใช้อย่างกว้างขวางในหลายสาขา อย่างเช่น การจำแนกประเภทเอกสาร (Aghdam et al., 2009; Shang et al., 2007) การตรวจจับผู้บุกรุก (Lee et al., 2000, Mun et al., 2009)

ในช่วงหลายสิบปีที่ผ่านมาวิธีการเลือกลักษณะเด่นได้มีการนำเสนอวิธีการที่ใช้กันอย่างกว้างขวางสำหรับการเลือกลักษณะเด่นแบบกรองคือ ราฟเซตและมิวชวลอินฟอร์เมชั่น โดยวิธีการที่มีอยู่โดยส่วนใหญ่บนหลักการของราฟเซตจะเป็นวิธีการประเมินแบบซับเซต ซึ่งเป็นการค้นหาซับเซตที่เล็กที่สุดของคุณลักษณะที่อาศัยข่าวสารจากการประมาณขอบเขตล่างเพียงอย่างเดียว มิวชวลอินฟอร์เมชั่นนั้นเป็นวิธีการที่นำมาใช้กันอย่างกว้างขวางสำหรับการเรียงลำดับความสำคัญของคุณลักษณะ (Feature ranking) ซึ่งเป็นการประเมินแต่ละคุณลักษณะแยกกันและมีการกำหนดน้ำหนักให้กับคุณลักษณะตามระดับความสำคัญ ซับเซตของคุณลักษณะบ่อยครั้งเลือกมาจากตัวบน ๆ ของรายการที่จัดเรียงลำดับ ซึ่งซับเซตนี้จะประมาณว่าเป็นคุณลักษณะที่สำคัญ อย่างไรก็ตามข้อเสียอย่างหนึ่งของการเรียงลำดับความสำคัญของคุณลักษณะคือ มันจะเป็นเรื่องที่ยากที่จะลบคุณลักษณะที่ซ้ำกันเพราะว่าคุณลักษณะที่ซ้ำกันจะมีลำดับที่เหมือนกัน นอกจากนี้เทคนิคนี้จำเป็นที่จะต้องมีการระบุถึงจำนวนของคุณลักษณะที่จะเลือกเสียก่อน ดังนั้นเพื่อที่จะแก้ปัญหาเหล่านี้บางงานวิจัยจึงมีการนำเสนอมิวชวลอินฟอร์เมชั่นสำหรับการหาคุณลักษณะแบบซับเซต

ทฤษฎีราฟเซต (RST) นำเสนอโดย Pawlak (1982, 1991) เป็นโมเดลใหม่ทางคณิตศาสตร์สำหรับจัดการกับความไม่แน่นอนและความไม่สมบูรณ์ของข่าวสาร วิธีการราฟเซตจะทำการวิเคราะห์ข้อมูลบนหลักการ 2 แนวคิดที่สำคัญ ได้แก่ การประมาณขอบเขตล่างและการประมาณขอบเขตบน ราฟเซตได้มีการนำมาประยุกต์ใช้ในหลายสาขาของงานวิจัย เช่น การเลือกลักษณะเด่น (Hassanien, 2004; Jensen and Shen, 2004; Parthal´ain et al., 2010) การจำแนกประเภทเอกสาร (Li et al., 2006; Miao et al., 2009) การเรียนรู้ของเครื่องจักร (Han and Kim, 2008) และในขณะนี้จัดว่าเป็นเทคนิคที่ได้พัฒนาขึ้นอย่างมากในการวิเคราะห์ข้อมูลอย่างชาญฉลาดซึ่งแตกต่างจากวิธีการแบบอื่น ๆ อย่างเช่น ทฤษฎีฟัซซี่เซต วิธีการเชิงสถิติ เนื่องจากการวิเคราะห์ราฟเซตไม่ต้องการอินพุตจากคนหรือขอบเขตความรู้และใช้เฉพาะข่าวสาร

ที่อยู่ในข้อมูลเท่านั้น อย่างไรก็ตามในบางสถานการณ์ RST อาจจะไม่สามารถวิเคราะห์ข้อมูลที่มีสัญญาณรบกวนหรือมีความขัดแย้งกันได้อย่างมีประสิทธิภาพ ดังนั้นจึงมีหลายงานวิจัยพยายามที่แก้ปัญหาเหล่านี้ด้วยการประยุกต์ใช้กับโมเดลของ VPRS (Ziarko, 1993, 2008)

การหารีดัก (Reduct) ของข่าวสาร หรือ ระบบการตัดสินใจ เป็นปัญหาหลักของ RST รีดักเป็นซับเซตของแอตทริบิวต์ที่เล็กที่สุดของข้อมูล และต้องมีการจัดกลุ่มของวัตถุในเซตเอกภพสัมพัทธ์ได้เหมือนกับการจัดกลุ่มวัตถุบนแอตทริบิวต์ทั้งหมดของข้อมูล ชัดเจนว่ารีดักคือ กระบวนการเลือกซับเซตของแอตทริบิวต์ และซับเซตของแอตทริบิวต์ที่เลือกนั้นไม่ใช่เฉพาะข่าวสารที่มีการเก็บรักษาไว้เท่านั้น แต่ซับเซตนั้นต้องมีความซ้ำซ้อนกันน้อยที่สุดด้วย วิธีการที่มีอยู่ของ RST บนหลักการ FS โดยส่วนใหญ่ (Chen et al., 2010; Hedar et al., 2006, Hu et al., 2008; Jensen and Shen, 2004) อาศัยอยู่บนแนวคิดหลักของการประมาณขอบเขตล่างหรือพื้นที่ของความแน่นอนเป็นตัววัดความดีของคุณลักษณะ อย่างเช่น dependency function (Hedar et al., 2006; Jensen and Shen, 2004) significance of attributes (Chen et al., 2010; Hu et al., 2008) ถึงแม้ว่า RST จะประสบความสำเร็จในการประยุกต์ใช้กับปัญหา FS อย่างมาก แต่วิธีการเหล่านี้ก็ไม่ได้สนใจต่อข่าวสารที่เก็บอยู่ในบริเวณขอบเขตหรือพื้นที่ของความไม่แน่นอน ดังนั้นการใช้ข่าวสารจากการประมาณขอบเขตล่างอย่างเดียวไม่เพียงพอสำหรับการเลือกคุณลักษณะได้อย่างมีประสิทธิภาพโดยเฉพาะอย่างยิ่งเมื่อต้องนำมาใช้กับข้อมูลที่มีมิติสูง ๆ หรือ มีสัญญาณรบกวนสูง ๆ ในขณะที่บางงานวิจัยบนหลักการของ RST ที่มีการพิจารณาข่าวสารของบริเวณขอบเขต (Deogun et al., 1995; Inuiguchi and Tsurumi, 2006) แต่วิธีการเหล่านี้พิจารณาเฉพาะความรู้ของการประมาณขอบเขตล่างอย่างเดียว แทนที่จะมีการพิจารณาการประมาณขอบเขตล่างและบริเวณขอบเขตแยกจากกัน ดังนั้นจึงมีบางงานวิจัยที่ประสบผลสำเร็จในการประยุกต์ใช้กับการจำแนกประเภทเอกสาร (Miao et al., 2009) ซึ่งเป็นวิธีการที่พิจารณาการประมาณขอบเขตล่างและบริเวณขอบเขตแยกกัน

ปัญหาหลัก ๆ ของ RST คือ การวิเคราะห์การจัดกลุ่มอย่างสมบูรณ์ของวัตถุที่อยู่ในคลาสที่กำหนด ถึงแม้ว่าจะสามารถที่จะจัดการกับข้อมูลที่มีความขัดแย้งกันในระดับส่วนเล็ก ๆ ได้ แต่ก็ไม่ได้ทนทานต่อสัญญาณรบกวนหรือค่าของแอตทริบิวต์ที่ไม่แน่นอน โดยเฉพาะอย่างยิ่งข่าวสารที่ยอมให้มีการจัดกลุ่มเพียงบางส่วนได้ โดยทั่วไปแล้ว RST นั้นเป็นโมเดลของการจัดกลุ่มที่การจัดกลุ่มนั้นจะถูกต้องทั้งหมด ดังนั้นการจัดกลุ่มโดยการควบคุมระดับของความไม่แน่นอนหรือความผิดพลาดของการจำแนกประเภทนั้นจะอยู่นอกเหนือการทำงานของ RST อย่างไรก็ตาม การยอมรับถึงระดับของความไม่แน่นอนในทางปฏิบัตินั้นสามารถปรับปรุงประสิทธิภาพการทำงานของอัลกอริทึมการเรียนรู้ให้ดีขึ้นได้ เพื่อที่จะแก้ปัญหาข้อบกพร่องเหล่านี้ Ziarko (1993, 2008) ได้มีการแนะนำราฟเซตด้วยตัวแปรเที่ยงตรง (VPRS) ซึ่งพัฒนาต่อมาจาก RST โดยจะมีการพิจารณาบางวัตถุของชุดข้อมูลให้สามารถจำแนกผิดกลุ่มได้ หรือ เป็นวัตถุที่ไม่แน่นอน ดังนั้นในงานวิจัยนี้จะใช้ VPRS ในการแบ่งพาร์ติชันสเปซของคุณลักษณะของชุดของข้อมูล และยิ่งไปกว่านั้นการประมาณค่าขอบเขตล่างและขอบเขตบนของ VPRS สามารถคำนวณได้โดยใช้หลักการของ majority inclusion relation ที่สามารถยอมรับความผิดพลาดที่เกิดขึ้นในระดับที่กำหนดไว้ได้ (β) วิธีการส่วนใหญ่บนพื้นฐานของ VPRS (Miao et al., 2009; Ziarko, 1993, 2008) จะอาศัยการป้อนค่าความผิดพลาดของการจำแนกประเภท (β) ที่ยอมรับได้จากมนุษย์ อย่างไรก็ตามการแนะนำค่าของ β นี้จะทำให้ขัดแย้งกับหลักการของราฟเซตที่ใช้เฉพาะข่าวสารที่ได้รับมาจากข้อมูลเท่านั้น ดังนั้นในงานวิจัยนี้เราจะเสนอวิธีการที่เลือกค่า β แบบอัตโนมัติในระหว่างกระบวนการหารีดักแทนที่จะถูกกำหนดมาก่อนโดยการคำนวณด้วยมนุษย์ ในขณะที่ Foithong et al. (2016) ได้นำเสนอหลักการเลือกลักษณะเด่นบนหลักการของราฟเซตและมิวชวลอินฟอร์เมชั่น นอกจากนี้ได้นำเสนอ

วิธีการเลือกค่า β แบบอัตโนมัติแทนการรับค่ามาจากมนุษย์ โดยการพิจารณาข่าวสารที่อยู่ในบริเวณขอบเขต ซึ่งในงานวิจัยได้ดำเนินการทดลองเพิ่มเติมต่อจาก Foithong et al. (2016) ในการเปรียบเทียบประสิทธิภาพกับวิธีการอื่น ๆ

ส่วนองค์ประกอบที่เหลือของงานวิจัยนี้ประกอบด้วยโครงสร้างดังต่อไปนี้ ส่วนที่ 2 เป็นการสรุปถึงทฤษฎีพื้นฐานของ VPRS และ MI ส่วนที่ 3 วิธีการที่เป็นแนวคิดใหม่สำหรับการเลือกคุณลักษณะบนหลักการของ VPRS และ MI และอัลกอริทึมที่นำเสนอจะมีการแสดงในส่วนนี้ ส่วนที่ 4 จะอธิบายถึงผลลัพธ์ของการประยุกต์ใช้วิธีการ RSAR (Jensen and Shen, 2004), DMRSAR (Parthal´ain et al., 2010), CNS (Liu and Setiono, 1996) และ RelifeF (Kira and Rendell, 1992; Kononenko, 1994) กับชนิดของข้อมูลที่ไม่ต่อเนื่องและแบบต่อเนื่อง โดยผลลัพธ์ของวิธีการที่นำเสนอจะเปรียบเทียบกับวิธีการอื่น ๆ ดังที่กล่าวไปแล้วโดยจะมีการเปรียบเทียบในเทอมของความถูกต้องในการจัดจำแนกประเภทข้อมูล (โดยใช้ 3 classifier ที่แตกต่างกัน) และการลดจำนวนมิติของข้อมูล ส่วนที่ 5 เป็นบทสรุปของงานวิจัย

2. ทฤษฎีพื้นฐาน

ในส่วนนี้จะอธิบายแนวคิดต่าง ๆ ที่เป็นพื้นฐานในทฤษฎีของราฟเซตโดยตัวแปรเที่ยงตรง มิวชวลอินฟอร์เมชั่น และมิวชวลอินฟอร์เมชั่นบนหลักการของราฟเซตโดยตัวแปรเที่ยงตรง

2.1 ราฟเซตโดยตัวแปรเที่ยงตรง (Variable precision rough set)

ปัญหาที่เกิดขึ้นกับ RST คือ เมื่อนำมาประยุกต์ใช้งานกับข้อมูลที่มีสัญญาณรบกวนและข้อมูลที่มีค่าขัดแย้งกันจะทำให้การค้นหาซับเซตของคุณลักษณะนั้นล้มเหลวได้ ในหลาย ๆ เซตข้อมูลจริงที่มีการประยุกต์ใช้การสันนิษฐานว่าข้อมูลมีความถูกต้องแม่นยำเป็นสิ่งที่คาดหวังไม่ได้เลย หรือค่าแอตทริบิวต์เงื่อนไขมีสิ่งเจือปนก็เป็นไปได้ เพื่อที่จะแก้ปัญหาข้อจำกัดเหล่านี้ Ziarko (1993, 2008) ได้มีการแนะนำส่วนขยายของ RST ซึ่งคือ ราฟเซตโดยตัวแปรเที่ยงตรง แนวคิดหลัก ๆ ของ VPRS คือ ยอมให้วัตถุสามารถถูกจัดกลุ่มได้โดยมีค่าความผิดพลาดที่น้อยกว่าค่าระดับที่ถูกกำหนดไว้ก่อนอย่างชัดเจน หลักการพื้นฐานบางอย่างของ VPRS จะแนะนำดังต่อไปนี้

ให้ A และ B เป็น ซับเซตที่ไม่เป็นเซตว่างของ U ระดับสัมพันธ์ของการจัดกลุ่มที่ผิดพลาดของเซต A โดยเทียบกับเซต B ถูกนิยามเป็น

$$c(A,B) = 1 - \frac{|A \cap B|}{|A|} \quad \text{if } |A| > 0 \tag{1}$$

$$c(A,B) = 0 \quad \text{if } |A| = 0 \tag{2}$$

มีข้อสังเกตที่สำคัญ ซึ่ง c(A, B) = 0 ก็ต่อเมื่อ $A \subseteq B$

ความสัมพันธ์แบบรวมกลุ่มเอาส่วนที่มากกว่า (Majority inclusion relation) ซึ่งก็คือ ระดับของการรวมกลุ่มที่ยอมให้มีความผิดพลาดในการจัดกลุ่มที่ยอมรับได้ (Admissible classification error) จะแทนด้วย β นิยามเป็น

$$A \subseteq_\beta B \text{ if and only if } c(A,B) \leq \beta \; 0 \leq \beta < 0\,5 \tag{3}$$

โดยการใช้ $\subseteq_\beta$ แทนที่ของ $\subseteq$ สำหรับการประมาณค่าขอบเขตล่าง p (p-lower approximation) และการประมาณค่าขอบเขตบน (p-upper approximation) บน p ของ A สามารถที่จะถูกนิยามใหม่ตอนนี้ได้เป็น

$$\underline{P}(A)_\beta = \cup\{a_i | [a_i]_P \subseteq_\beta A\} \tag{4}$$

$$\overline{P}(A)_\beta = \cup\{a_i | c([a_i]_P, A) < 1 - \beta\} \tag{5}$$

ดังนั้นนิยามใหม่ของบริเวณเชิงบวก บริเวณเชิง-ลบ และบริเวณขอบเขตบนหลักการของ VPRS เป็นการกำหนดโดย

$$POS_{P\beta}(A) = \underline{P}(A)_\beta \tag{6}$$

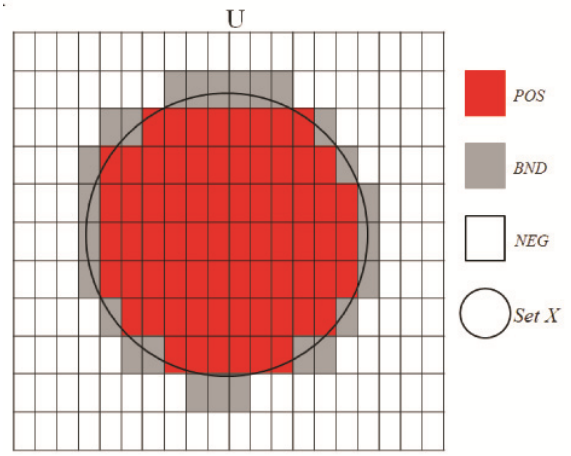
$$NEG_{P\beta}(A) = U - \overline{P}(A)_\beta \tag{7}$$

$$BND_{P\beta}(A) = \overline{P}(A)_\beta - \underline{P}(A)_\beta \tag{8}$$

และในทำนองเดียวกันระดับการขึ้นต่อกันเขียนได้เป็น

$$\gamma_{P\beta}(A) = \frac{|POS_{P\beta}(A)|}{|U|} \quad (9)$$

ในที่นี้การประมาณขอบเขตล่างของเซต A ตีความได้ว่าเป็นการรวบรวมสมาชิกทั้งหมดของเซต (Elementary sets) ซึ่งสามารถจัดกลุ่มใน A ได้ โดยมีผิดพลาดในการจัดกลุ่มไม่มากกว่า β การประมาณขอบเขตบนของเซต A จะประกอบด้วยเซตของสมาชิกทั้งหมดที่ไม่สามารถจัดกลุ่มใน -A ได้ ด้วยความผิดพลาดไม่มากกว่า β สุดท้ายคือ บริเวณขอบเขตของ A ประกอบด้วยเซตสมาชิกทั้งหมดที่ไม่สามารถที่จะจัดกลุ่มเข้าสู่ A หรือ -A ได้ ด้วยค่าความผิดพลาดของการจัดกลุ่มไม่เกิน β



ภาพที่ 1 ราฟเซตโดยตัวแปรเที่ยงตรงในสเปซของคุณลักษณะแบบไม่ต่อเนื่อง

2.2 มิวชวลอินฟอร์เมชั่นบนหลักการราฟเซต (Mutual information based on rough set)

ทฤษฎีข่าวสารถูกเสนอโดย Shannon and Weaver (1949) โดยจะนำมาใช้วัดข่าวสารของข้อมูลด้วยเอนทรอปี (Entropy) และมิวชวลอินฟอร์เมชั่น เอนทรอปีสามารถตีความว่าเป็นการประมาณปริมาณข่าวสารที่แสดงในตัวแปรสุ่ม ส่วนมิวชวลอินฟอร์เมชั่นเป็นการวัดความสัมพันธ์กันระหว่างสองตัวแปรสุ่มและสามารถมองว่าเป็นข่าวสารร่วมกันของสองตัวแปรสุ่มในระบบข่าวสารนั้น เอนทรอปีสามารถเป็นตัววัดข่าวสารสำหรับการเลือกคุณลักษณะบนความรู้ด้วยความน่าจะเป็นที่เกี่ยวข้องกับคุณลักษณะที่กำหนด

ใน RST ความสัมพันธ์แบบชั้นสมมูลจะนำมาสู่พาร์ติชั่นของเอกภพสัมพัทธ์ พาร์ติชั่นสามารถพิจารณาว่าเป็นชนิดของความรู้โดยความหมายของความรู้ในกรอบงานทฤษฎีข่าวสารของราฟเซตสามารถตีความเป็นดังนี้

สำหรับทุก ๆ ซับเซต P $\subseteq$ C ของคุณลักษณะ ให้ U/IND(P) = { O_1, O_2, …, O_n} แทน การแบ่งพาร์ติชั่นของวัตถุของเซต U ด้วยค่าแอตทริบิวต์ของเซต P (IND(P) คือ P-Indiscernibility Relation) โดยข่าวสารเอนทรอปี H(P) ของความรู้ P นิยามเป็น

$$H(P) = -\sum_{i=1}^{n} p(O_i)\log(p(O_i)) \quad (10)$$

โดยที่ $p(O_i) = \frac{|O_i|}{|U|}\ 1 \le i \le n$

ให้ Pและ Q เป็นซับเซตของ C ให้ U/IND(P) = { O_1, O_2, …, O_n} และ U/IND(Q) = { L_1, L_2 ,… , L_n} แทน พาร์ติชั่นที่ชักนำโดยความสัมพันธ์แบบชั้นสมมูล IND(P) และ IND(Q) ตามลำดับ ดังนั้นเอนทรอปีแบบมีเงื่อนไข H(Q|P) ของความรู้ Q ที่กำหนดโดยความรู้ P นิยามเป็น

$$H(Q|P) = -\sum_{i=1}^{n} p(O_i) \sum_{j=1}^{m} p(L_j|O_i)\log(p(L_j|O_i) \quad (11)$$

โดยที่ $p(O_i) = \frac{|O_i|}{|U|}\ p(L_j|O_i) = \frac{|L_j \cap O_i|}{|O_i|}\ 1 \le i \le n, 1 \le j \le m$

มิวชวลอินฟอร์เมชั่นเป็นตัววัดผลรวมของข่าวสารที่ความรู้ P เกี่ยวข้องกับความรู้ Q นิยามเป็น

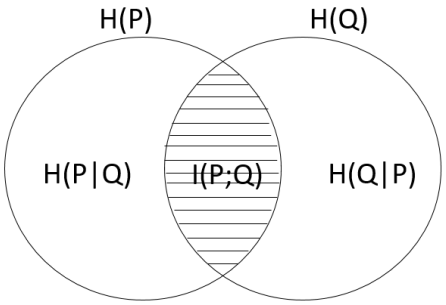
$$I(Q;P) = \sum_{j=1}^{m}\sum_{i=1}^{n} p(L_j, X_i) log \frac{p(L_j, O_i)}{p(L_j)p(O_i)} \quad (12)$$

โดยที่ $p(O_i) = \frac{|O_i|}{|U|}\ p(L_j, O_i) = \frac{|L_j \cap O_i|}{|U|}\ 1 \le i \le n, 1 \le j \le m$

ถ้ามิวชวลอินฟอร์เมชั่นระหว่าง P และ Q เป็นค่ามาก (ค่าน้อย) หมายถึง P และ Q เกี่ยวข้องกันอย่างใกล้ชิด (ไม่ใกล้ชิด) ความสัมพันธ์ระหว่างมิวชวลอินฟอร์เมชั่นและเอนทรอปีสามารถนิยามเป็น

$$I(P;Q) = H(Q) - H(Q|P) \quad (13)$$

ในการเลือกลักษณะเด่น มิวชวลอินฟอร์เมชั่นมีบทบาทในการวัด ความสำคัญและความซ้ำซ้อนระหว่างคุณลักษณะ ข้อดีหลักของ MI คือ มันมีความทนทานต่อสัญญาณรบกวนและการแปลงพิกัด เราจะให้ความสนใจกับวิธีการของมิวชวลอินฟอร์เมชั่นเพื่อที่จะหาซับเซตของลักษณะเด่นที่มีความสำคัญมากที่สุด ในงานวิจัยนี้ มิวชวลอินฟอร์เมชั่นใช้วัดข่าวสารของความสัมพันธ์กันระหว่างการประมาณขอบเขตล่าง $\underline{P}(A)_\beta$ และคลาส A นอกจากนี้ มิวชวลอินฟอร์เมชั่นของบริเวณขอบเขต $BND_{P\beta}(A)$ โดยเทียบกับคลาส A จะวัดออกมาด้วยเช่นกัน โดยความสัมพันธ์ระหว่างเอนทรอปีและมิวชวลอินฟอร์เมชั่นของตัวแปรสุ่ม P และ Q แสดงดังรูปที่ 2



ภาพที่ 2 ความสัมพันธ์ระหว่างเอนทรอปีและมิวชวลอินฟอร์เมชั่น

วิธีดำเนินการวิจัย

ถึงแม้ในปัจจุบันจะมีบางกลไก (Parthal´ain et al., 2010) ที่นำเสนอบนหลักการราฟเซตที่ได้อ้างถึงบริเวณขอบเขต แต่การคำนวณข่าวสารของบริเวณขอบเขตก็ยังคงขึ้นอยู่กับข่าวสารของการประมาณขอบเขตล่างเป็นตัวสำคัญ ในความเป็นจริงแล้วเมื่อการประมาณขอบเขตล่างของคุณลักษณะเป็นเซตว่าง แต่เซตของบริเวณขอบเขตจะไม่เป็นเซตว่าง ซึ่งกลไกนี้ (Parthal´ain et al., 2010) มีประสิทธิภาพที่ไม่เพียงพอสำหรับการเลือกลักษณะเด่นเมื่อนำมาประยุกต์ใช้กับข้อมูลที่มีสัญญาณรบกวน ดังนั้นข่าวสารที่มีประโยชน์ของบริเวณขอบเขตสามารถใช้เพื่อวิเคราะห์ความดีของซับเซตของลักษณะเด่นได้ เมื่อการประมาณขอบเขตล่างเป็นเซตว่าง

วิธีการที่จะอธิบายในส่วนนี้จะใช้ทั้งข่าวสารที่อยู่ในการประมาณขอบเขตล่างและข่าวสารที่อยู่ในบริเวณขอบเขต เพื่อค้นหาซับเซตของคุณลักษณะที่ดีที่สุด การคำนวณเพื่อที่จะประมาณค่าเซตของทั้งการประมาณขอบเขตล่างและบริเวณขอบเขตก็เป็นอิสระกัน ยิ่งไปกว่านั้นมิวชวลอินฟอร์เมชั่นจะใช้เป็นตัววัดข่าวสารของทั้งการประมาณขอบเขตล่างและบริเวณขอบเขตเพื่อที่จะใช้เป็นแนวทางในการค้นหาซับเซตของคุณลักษณะที่ดีที่สุด

3.1 เลือกบริเวณขอบเขตน้อยที่สุด (Min-Boundary Regions)

ดังที่อธิบายไปแล้วข้างต้นว่าปัญหาหลักอย่างหนึ่งของ VPRS คือ การพิจารณาเลือกระดับของความคลาดในการจัดกลุ่มและเกือบทุกเทคนิคบนหลักการ VPRS (Miao et al., 2009; Ziarko, 1993, 2008) จะต้องกำหนดค่า admissible classification error (β) มาก่อน ดังนั้นค่า β ที่เหมาะสมก็จะกำหนดโดยการพิจารณาจากผลลัพธ์ที่ดีที่สุดของของการจัดกลุ่ม อย่างไรก็ตาม ในงานวิจัยนี้จะนำเสนอวิธีการที่เป็นแนวคิดใหม่ที่จะเลือกค่า β แบบอัตโนมัติแทนที่จะมีการกำหนดค่ามาก่อนด้วยคน ซึ่งวิธีการนี้จะคำนวณบนเฉพาะข่าวสารที่บรรจุอยู่ภายในข้อมูลตัวมันเองเท่านั้น

ให้ Y เป็นแอตทริบิวต์การตัดสินใจและกำหนดให้ U ถูกพาร์ติชั่นเป็นคลาสชั้นสมมูล U/IND(Y) = $\{Y_1, Y_2, \ldots, Y_m\}$ แล้วบริเวณขอบเขตของ U/IND(Y) โดยเทียบกับเซตของแอตทริบิวต์ P นิยามเป็น

$$BND_{P\beta}(Y)) = \{\overline{P}(Y_1)_\beta - \underline{P}(Y_1)_\beta\, \overline{P}(Y_2)_\beta - \underline{P}(Y_2)_\beta\, \overline{P}(Y_m)_\beta - \underline{P}(Y_m)_\beta\} \quad (14)$$

ดังนั้น มิวชวลอินฟอร์เมชั่นที่น้อยที่สุดระหว่างความรู้ Y และบริเวณขอบเขต $BND_{P\beta}(Y)$) เป็นกฎเกณฑ์ที่ใช้เพื่อเลือกค่า β ที่เหมาะสมสามารถกำหนดโดย

$$\tilde{\beta} = \min_{0.0 \le \beta < 0.5}\{I(Y; BND_{P\beta}(Y))\} \quad (15)$$

โดย β มีค่าอยู่ในช่วง [0.0, 0.5) โดยจะมีการพิจารณาที่ละ 0.05

ค่า β ที่ให้ค่ามิวชวลอินฟอร์เมชั่นระหว่างความรู้ Y และบริเวณขอบเขต $BND_{P\beta}(Y)$ ต่ำที่สุดจะถูกเลือกเป็นค่า $\tilde{\beta}$ นอกจากนี้ในงานวิจัยนี้เราพบว่ามิวชวลอินฟอร์เมชั่นที่มีค่าต่ำที่สุดนั้นสามารถที่จะพบได้ด้วย β ที่อยู่ในช่วง [0.3, 0.45]

ในการประยุกต์สมการที่ (15) เพื่อหาค่าเหมาะสม $\tilde{\beta}$ ในระหว่างกระบวนการหารีดักของทุกซับเซตของคุณลักษณะ P กฎเกณฑ์ในสมการ (15) ต้องมีการคำนวณใหม่ด้วย ดังนั้นในแต่ละซับเซตของคุณลักษณะ P ตัวใหม่จำเป็นที่ต้องหาค่าเหมาะสม $\tilde{\beta}$ ด้วยค่า β ในช่วง 0.3-0.45 ซึ่งค่าเหมาะสมนี้จะถูกนำไปใช้สำหรับหาการประมาณขอบเขตล่างและบริเวณขอบเขตซึ่งจะมีการอธิบายในส่วนถัดไป

3.2 Max-Lower และ Min- Boundary

กฎเกณฑ์ Max-Lower และ Min-Boundary เป็นการหาข่าวสารที่มากที่สุดของบริเวณของความแน่นอนและในขณะเดียวกันจะหาค่าข่าวสารที่น้อยที่สุดในบริเวณของความไม่แน่นอน การประเมินความดีของซับเซตของคุณลักษณะทำได้โดยการคำนวณหาผลรวมของความแตกต่างของข่าวสารที่มากที่สุด โดยการคำนวณนั้นจะทำการลบข่าวสารของบริเวณขอบเขตออกจากข่าวสารของการประมาณขอบเขตล่าง ซึ่งแนวคิดนี้คาดหวังว่าซับเซตของคุณลักษณะที่เป็นผลลัพธ์ที่ได้รับจะเป็นคุณลักษณะที่ตรงกับความต้องการมากที่สุด

การหามิวชวลอินฟอร์เมชั่นที่น้อยที่สุดของบริเวณขอบเขตโดยเทียบกับความรู้ Y สำหรับทุก ๆ ซับเซตของคุณลักษณะ P ด้วย $\tilde{\beta}$ ที่เหมาะสมจะนิยามเป็น

$$BoundInf(P) = \min\{I(Y; BND_{P\tilde{\beta}}(Y))\} \quad (16)$$

โดยผลรวมข่าวสารของมิวชวลอินฟอร์เมชั่นระหว่างการประมาณขอบเขตล่าง $\underline{P}(Y_i)_{\tilde{\beta}}$ และคลาสชั้นสมมูล Y_i ด้วย $\tilde{\beta}$ ที่เหมาะสมถูกแทนด้วย $LowerInf(P)$ สามารถที่จะนิยามเป็นดังนี้

$$LowerInf(P) = \sum_{i=1}^{m} I(Y_i; \underline{P}(Y_i)_{\tilde{\beta}}) \quad (17)$$

ดังนั้นปัญหาของการเลือกซับเซตของคุณลักษณะ P จะสอดคล้องกับการหาค่าสูงสุดของ $LowerInf(P)$ และการหาค่าต่ำสุดของ $BoundInf(P)$ ซึ่งก็คือการหาค่าที่มากสุดของฟังก์ชันเป้าหมาย G(P) โดยที่

$$G(P) = LowerInf(P) - BoundInf(P) \quad (18)$$

จะเห็นได้ชัดว่าถ้า $LowerInf(P) = H(Y)$ แล้วค่าของฟังก์ชันเป้าหมาย G(P) จะเป็นค่ามากที่สุดและข่าวสารที่ถูกประมาณมีความไม่แน่นอนโดยเทียบกับ P ดังนั้นซับเซตของคุณลักษณะ P จะพิจารณาว่าเป็นคุณลักษณะที่ตรงกับความต้องการอย่างแข็งแรง ในทางกลับกันถ้า $BoundInf(P) = H(Y)$ แล้ว P จะนำมาสู่การประมาณค่าของข่าวสารที่มีความไม่แน่นอนสูงที่สุด ดังนั้น P จะเป็นคุณลักษณะที่ไม่ตรงกับความต้องการแสดงว่าจะไม่มีข่าวสารที่มีประโยชน์ที่เกี่ยวกับแอตทริบิวต์ของการตัดสินใจ Y โดยผลรวมของความแตกต่างของทั้งสองค่าที่ได้รับนี้จะอยู่ในช่วง [0, H(Y)] และ $G(P)$ จะมีค่าอยู่ในช่วง [-H(Y), H(Y)] กลไกการเลือกคุณลักษณะใหม่นี้สามารถที่สร้างขึ้นมาได้โดยใช้ผลรวมของความแตกต่างของข่าวสารระหว่างค่าความแน่นอนและค่าของความไม่แน่นอนเพื่อที่จะใช้แนะนำในการค้นหาซับเซตที่ดีที่สุดของคุณลักษณะ

3.3 อัลกอริทึมการเลือกลักษณะเด่น mBML

จากรูปที่ 3 จะแสดงถึงอัลกอริทึม mBMLREDUCT บนหลักการของ VPRS ที่ได้อธิบายไปแล้วก่อนหน้านี้ ในรูปที่ 3 นั้น mBMLREDUCT แต่จะใช้ค่าฟังก์ชันเป้าหมาย G ที่มากที่สุด เพื่อที่จะแนะนำกระบวนการเลือกลักษณะเด่น ถ้าค่า G ของรีดักที่คัดเลือกมาของตัวปัจจุบันมากกว่าตัวก่อนหน้านี้ แสดงว่าซับเซตนี้ (รีดักตัวปัจจุบัน) จะถูกเก็บรักษาไว้และถูกนำไปใช้ในรอบถัดไป ซึ่งกระบวนการเลือกลักษณะเด่นจะสิ้นสุดลงเมื่อการเพิ่มคุณลักษณะที่เหลืออยู่มีผลทำให้ผลลัพธ์ของค่าฟังก์ชั่นข่าวสาร ($LowerInf$) เท่ากับค่าของชุดข้อมูลที่แรกเริ่ม อย่างไรก็ตามในบางสถานการณ์ข้อมูลที่ปราศจากสัญญาณรบกวนเราสามารถใช้ค่าของ H(Y) มาใช้เป็นกฎเกณฑ์การหยุดได้โดยการเปรียบเทียบ $LowerInf$ ของรีดัก

mBMLREDUCT(C, D)
C, the set of all conditional features;
D, the set of decision features.

1) $T \leftarrow \{\}, R \leftarrow \{\}$
2) do
3) $\quad \forall x \in (C - R)$
4) $\quad\quad$ Compute the optimal $\tilde{\beta}$ by formula (15) for $R \cup \{x\}$
5) $\quad\quad$ if $G(R \cup \{x\}) > G(T)$
6) $\quad\quad\quad T \leftarrow R \cup \{x\}$
7) $\quad R \leftarrow T$
8) until $LowerInf(R) == LowerInf(C)$
9) return R

ภาพที่ 3 อัลกอริทึม mBMLREDUCT

อัลกอริทึม mBMLREDUCT จะเริ่มต้นด้วยซับเซต R ที่เป็นเซตว่างและลูป do until จะทำงานโดยการพิจารณาค่า G ของซับเซตและคุณลักษณะที่เป็นเงื่อนไขหนึ่งตัว และค่อย ๆ เพิ่มเข้าไปเข้าไปครั้งละหนึ่งตัวในครั้งถัด ๆ ไป สำหรับแต่ละการทำซ้ำคุณลักษณะที่เป็นเงื่อนไขที่ยังไม่ได้ถูกประเมินจะนำไปรวมชั่วคราวกับซับเซต R และ $\tilde{\beta}$ ที่เหมาะสมของซับเซตนี้ก็จะถูกคำนวณ (บรรทัด 4) ถ้าค่าผลรวมของความแตกต่างของข่าวสารของซับเซตปัจจุบัน ($R \cup \{x\}$) มากกว่าค่าซับเซตก่อนหน้านี้ (T) แล้ว แอตทริบิวต์ที่ถูกเพิ่มเข้าไป (บรรทัด 6) จะถูกเก็บรักษาไว้เป็นส่วนหนึ่งของซับเซต T ตัวใหม่ (บรรทัด 7)

ก่อนที่เราจะศึกษาถึงผลการทดลองของประสิทธิภาพของอัลกอริทึมที่นำเสนอ เราจะแสดงการวิเคราะห์ Time complexity ของ mBMLREDUCT ดังที่เราเห็นในรูปที่ 3 นั้น การคำนวณหลัก ๆ ของอัลกอริทึมจะกี่ยวกับค่า $\tilde{\beta}$ และค่า *G* โดยการประมาณค่าขอบเขตล่างและบริเวณขอบเขตจะมี Time complexity แบบ quadratic ในเทอมของจำนวนแถวข้อมูล (*M*) และในเทอมของมิติ *N* โดยอัลกอริทึมรีดักนี้จะมี Time complexity เป็นแบบ nonlinear ซึ่งก็คือ $O(NM^2)$ และถ้าเราสมมติว่ารีดักมาจากคุณลักษณะเพียงหนึ่งตัวซึ่งกรณีนี้จะเป็นกรณีที่ดีที่สุด (Best-case) ของ Time complexity ซึ่งก็คือ $O(NM^2)$ ส่วนกรณีที่แย่ที่สุด (Worst-case) ของความ time complexity จะเป็น $O(N^2M^2)$ เมื่อทุก ๆ คุณลักษณะมันถูกเลือกเป็นรีดัก

ผลการวิจัย

ในส่วนนี้เราจะแสดงถึงผลลัพธ์ของการทดลองกับข้อมูลที่มีค่าแบบไม่ต่อเนื่องและค่าต่อเนื่อง ซึ่งข้อมูลทั้งสองกลุ่มได้มาจาก UCI Machine Learning Repository (Newman et al., 1998) โดยในขั้นแรกนั้นวิธีการ mBML จะถูกเปรียบเทียบกับวิธีการบนหลักการราฟเซต คือ RSAR (Jensen and Shen, 2004) และ DMRSAR (Parthal´ain et al., 2010) และนอกจากนี้ mBML จะนำไปเปรียบเทียบกับวิธีการเลือกลักษณะเด่นด้วยวิธีการประเมินซับเซตบนหลักการ CNS (Liu and Setiono, 1996) และวิธีการ ReliefF (Kononenko, 1994) โดยที่สำหรับเซตข้อมูลที่บรรจุด้วยคุณลักษณะที่มีค่าแบบต่อเนื่องนั้นเราจะประยุกต์วิธีการทำพาร์ติชั่นแบบ Equidistance (Battiti, 1994) ก่อนการประยุกต์ใช้กับ RSAR DMRSAR CNS ReliefF และ mBML นอกจากนี้ซับเซตของลักษณะเด่นที่ได้จากแต่ละวิธีนั้นเราจะประยุกต์กับตัวจำแนกประเภทข้อมูลซึ่งประกอบด้วย SVM C4.5 และ PART บนแต่ละเซตข้อมูล โดยเราจะใช้หลักการของ 10-fold cross-validation ในการคำนวณหาค่าความถูกต้องเฉลี่ยในการจำแนกประเภทข้อมูลบน

ตัวจำแนกประเภททั้ง 3 ชนิดตามที่ได้ระบุข้างต้น ทั้งนี้ในการทดลองนั้นเราจะทำการเปรียบเทียบทุกเทคนิค FS บนพื้นฐานของขนาดซับเซต ความถูกต้องในการจำแนกประเภท และเวลาที่ใช้ในการค้นหาซับเซตของลักษณะเด่น

4.1 การเปรียบเทียบผลการทดลองของ RSAR DMRSAR และ mBML

ในส่วนนี้ mBML จะมีการเปรียบกับ RSAR และ DMRSAR ซึ่งเป็น FS บนหลักการ RST ผลลัพธ์ต่าง ๆ จะถูกแสดงในเทอมของขนาดของซับเซต เวลาที่ใช้ และความถูกต้องการจำแนกประเภท ทั้งบนเซตข้อมูลแบบไม่ต่อเนื่องและข้อมูลที่มีค่าต่อเนื่อง โดยเซตข้อมูลแบบไม่ต่อเนื่องใช้ในช่วงขนาดจาก 106 ถึง 3,190 แถวข้อมูลและข้อมูลจะมี 14 ถึง 61 แอตทริบิวต์ ในขณะที่เซตข้อมูลที่มีค่าต่อเนื่องใช้ในช่วงขนาดจาก 178 ถึง 5,000 แถวข้อมูลและระหว่าง 14 ถึง 217 แอตทริบิวต์ซึ่งผลลัพธ์ต่าง ๆ จะแสดงในตารางที่ 1 2 3 และ 4 ตามลำดับ

4.1.1 ความถูกต้องการจำแนกประเภทของข้อมูล (Classification Accuracy)

ผลลัพธ์การทดลองได้ทดสอบบนข้อมูลแบบไม่ต่อเนื่องซึ่งได้สรุปในตารางที่ 1 ผลลัพธ์เหล่านี้แสดงความถูกต้องในการจำแนกประเภทโดยค่าเฉลี่ยเป็นเปอร์เซ็นต์ โดยการใช้กับแต่ละตัวจำแนกประเภททั้ง 3 ตัวที่กล่าวไปแล้วข้างต้น การคำนวณค่าความถูกต้องในการจำแนกประเภทจะดำเนินการบนเซตข้อมูลที่มีการลดมิติแล้ว ซึ่งได้รับมาจากเทคนิคการลดขนาดมิติ RSAR DMRSAR และ mBML ตามลำดับ

ผลลัพธ์ค่าความถูกต้องในการจำแนกประเภทข้อมูลที่แสดงในตารางที่ 1 นั้น เราสามารถเห็นถึงวิธีการ mBML ดำเนินการได้ดีมาก ๆ และแสดงการเพิ่มขึ้นในความถูกต้องของการจำแนกประเภทได้อย่างโดดเด่นอย่างน้อย 1 ตัวจำแนกประเภท (credit heart soybean promoters splice และ dermatology) โดยบนเซตข้อมูล credit และ soybean วิธีการ mBML ได้มีการเพิ่มขึ้นสูงสุดเป็น 5% และ 9.1% เมื่อเปรียบเทียบกับ RSAR และ DMRSAR ตามลำดับ มันเป็นสิ่งที่น่าประทับใจที่การเพิ่มขึ้นในค่าความถูกต้องในการจำแนกประเภทข้อมูลบน splice เพิ่มขึ้นสูงสุด 14.7% เมื่อเปรียบเทียบกับ DMRSAR และเพิ่มขึ้นสูงสุด 16.3% เมื่อเปรียบเทียบกับ RSAR สำหรับบางเซตข้อมูลเมื่อพิจารณาซับเซตของแอตทริบิวต์ที่มีความไม่สอดคล้องกันในระดับที่สูงของวัตถุ (เช่น credit heart soybean promoters splice และ dermatology) วิธีการ mBML สามารถที่จะจัดการกับความไม่สอดคล้องกันในข้อมูลที่สูงได้ดีกว่าทั้ง RSAR และ DMRSAR เมื่อพิจารณาจากความถูกต้องการจำแนกประเภท mBML บางครั้งอาจจะค้นพบซับเซตของขนาดที่เหมือน (แต่ประกอบด้วยคุณลักษณะที่แตกต่างกัน) กับ RSAR และ DMRSAR แต่ถึงอย่างไรก็ตามก็แสดงถึงการเพิ่มขึ้นในความถูกต้องของการจำแนกประเภทข้อมูลสำหรับอย่างน้อย 1 ตัวจำแนกประเภท (เช่น vote dermatology wq และ dna) และบางครั้งมีการเพิ่มขึ้นอย่างมีนัยสำคัญ ดังนั้นเป็นสิ่งชัดเจนจากค่าความถูกต้องการจำแนกประเภทข้อมูลที่ mBML สามารถเลือกซับเซตของแอตทริบิวต์ที่มีข่าวสารที่มีประโยชน์มากกว่าวิธีการ RSAR และ DMRSAR และนอกจากนี้ยังแสดงให้เห็นว่า mBML สามารถเลือกซับเซตมีข่าวสารที่มีประโยชน์มากกว่า วิธีการ RSAR และ DMRSAR

ตารางที่ 1 ค่าความถูกต้องเฉลี่ยในการจำแนกประเภทกับชุดข้อมูลแบบไม่ต่อเนื่อง

ชุดข้อมูล	SVM			J48			PART		
	RSAR	DMRSAR	mBML	RSAR	DMRSAR	mBML	RSAR	DMRSAR	mBML
credit	71.48	70.59	**74.90**	70.00	69.59	**72.00**	71.10	71.50	**72.00**
heart	77.21	77.50	**80.61**	**81.97**	**81.97**	80.25	80.61	**81.97**	80.95
vote	95.00	95.00	**95.66**	93.66	93.66	**94.50**	93.33	93.33	**94.00**
soybean	83.06	85.34	**89.57**	81.75	83.38	**86.64**	78.50	78.50	**85.66**
lymphography	81.75	81.75	**83.10**	73.64	**75.35**	73.64	**77.70**	77.02	77.02
promoters	85.84	85.84	**94.33**	84.90	84.90	**84.96**	**93.39**	**93.39**	**93.39**
splice	81.66	83.00	**94.98**	81.91	82.35	**94.48**	81.84	81.92	**93.79**
dermatology	81.00	83.00	**87.43**	69.83	71.30	**82.40**	73.46	75.58	**77.93**
wq	71.40	64.87	**72.20**	**68.52**	61.80	66.02	64.78	63.72	**69.09**
dna	**93.71**	**93.71**	93.08	83.64	83.64	**84.27**	**85.84**	**85.84**	**85.84**

เมื่อพิจารณาค่าความถูกต้องการจำแนกประเภทข้อมูลบนเซตข้อมูลที่เป็นค่าต่อเนื่อง ดังแสดงในตารางที่ 3 ผลลัพธ์ต่าง ๆ ได้แสดงให้เห็นถึงวิธี mBML สามารถดำเนินการได้ดีกว่า RSAR และ DMRSAR ในเทอมของค่าความถูกต้องในการจำแนกประเภทข้อมูล ในขณะที่ผลลัพธ์การลดขนาดของมิติที่ได้รับจะเหมือนกับ RSAR และ DMRSAR สิ่งนี้เป็นเครื่องยืนยันที่แสดงให้เห็นว่า mBML ทนทานต่อสัญญาณรบกวนในข้อมูลและดำเนินการกับสัญญาณรบกวนได้ดีกว่าทั้งวิธีการ RSAR และ DMRSAR เราสังเกตเห็นว่าจากผลลัพธ์การจำแนกประเภทข้อมูลนั้น วิธีการ mBML ดำเนินการได้ดีกว่าและแสดงถึงการเพิ่มขึ้นในค่าความถูกต้องการจำแนกประเภทข้อมูลสำหรับทุก ๆ ตัวจำแนกประเภท (ตัวอย่างเช่น wine mfeat-zer waveform2) ซึ่งเป็นการเพิ่มขึ้นอย่างมาก และ mBML ดำเนินการได้ดีกว่าทั้ง RSAR และ DMRSAR ในทุกกรณี ถึงแม้ว่าในบางข้อมูลค่าความถูกต้องการจำแนกประเภทข้อมูลมีการลดลงบ้าง แต่การลดลงนี้ไม่มีนัยสำคัญ ในบางกรณี DMRSAR ล้มเหลวบนเซตข้อมูล clean1 และ water2 เนื่องจากวิธีการนี้เลือกซับเซตของคุณลักษณะที่เป็นค่าสูงสุดเฉพาะที่ (Local maxima) และบางครั้งอาจจะตกอยู่ในกับหลุมดัก

ดังที่ถูกอธิบายมาแล้วข้างต้นว่าวิธีการที่นำเสนอนี้สามารถดำเนินการกับความไม่สอดคล้องกันได้อย่างเข้มแข็ง (เช่น splice mfeact-zer เป็นต้น) และทนทานต่อสัญญาณรบกวน (เซตข้อมูลที่มีค่าแบบต่อเนื่อง โดยส่วนใหญ่จะมีสัญญาณรบกวน) ถึงแม้ว่าวิธีการ RSAR และ DMRSAR สามารถจัดการกับความไม่สอดคล้องกันในข้อมูลได้ด้วยการเพิ่มแอตทริบิวต์ของซับเซต แต่ทั้งสองวิธีก็ไม่ทนทานต่อสัญญาณรบกวนและดำเนินการได้ไม่ดีกับความไม่สอดคล้องกันสูง ๆ ในข้อมูล อย่างไรก็ตามก็เป็นไปไม่ได้ที่จะหลีกเลี่ยงสัญญาณรบกวนในข้อมูล เมื่อเราต้องประยุกต์ใช้เทคนิค FS กับเซตข้อมูลที่มีค่าแบบต่อเนื่อง

4.1.2 ขนาดของซับเซตและเวลาที่ใช้ (Subset Size and Runtimes)

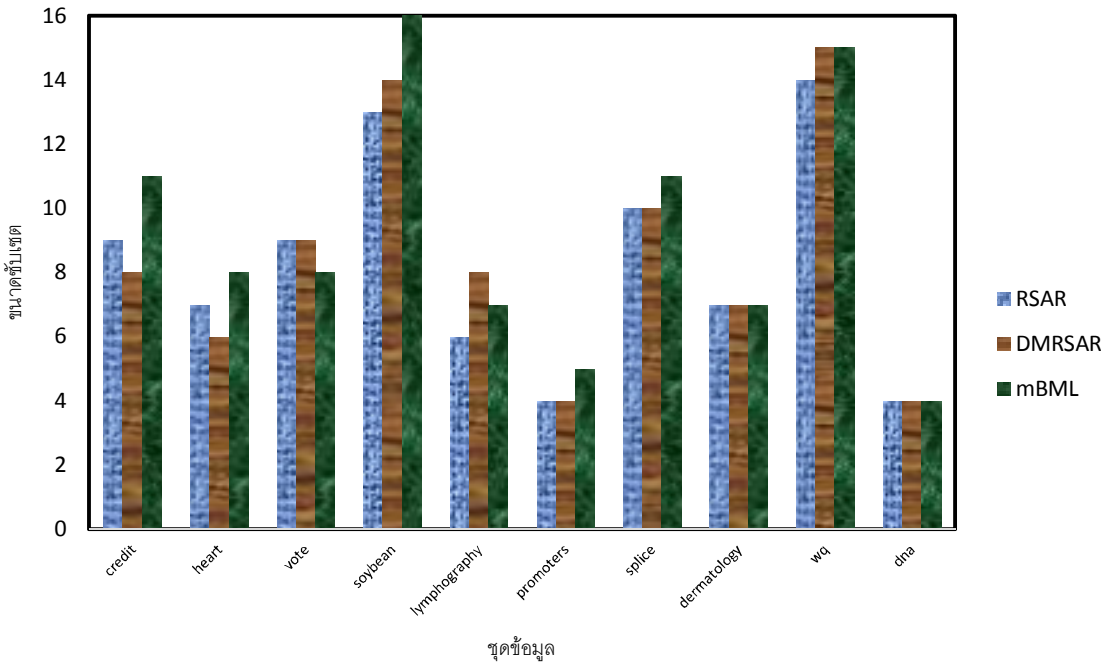
การเปรียบเทียบของขนาดของซับเซตและเวลาที่ใช้กับวิธีการ RSAR DMRSAR และ mBML ได้แสดงในตารางที่ 2 ถึงแม้ว่าซับเซตจะยังคงใหญ่กว่า RSAR และ DMRSAR ในเทอมของการลดขนาดมิติ แต่วิธีที่นำเสนอได้แสดงถึงการเพิ่มขึ้นอย่างมีนัยสำคัญในความถูกต้องการจำแนกประเภทข้อมูลอย่างน้อย 1 ตัวจำแนกประเภทข้อมูล (ได้แก่ credit heart soybean promoters และ splice) ยิ่งไปกว่านั้นผลลัพธ์การจำแนกประเภทข้อมูลบนเซตข้อมูล soybean และ splice เมื่อพิจารณาบนตัวจำแนกประเภทข้อมูลแล้วเราพบว่าสมรรถนะของ mBML ดีกว่าทั้งวิธีการ RSAR และ DMRSAR อย่างมีนัยสำคัญ วิธีการ mBML ได้แสดงให้เห็นถึงซับเซตของแอตทริบิวต์ที่เลือกมีมูลค่า

อย่างมากที่ถูกสกัดมาจากข้อมูล splice โดยการพิจารณาข่าวสารที่บรรจุอยู่ในทั้งบริเวณแน่นอน (Certainty region) และบริเวณของความไม่แน่นอน (Uncertainty region) ดังนั้นการหาค่าสูงสุดของความแตกต่างของข่าวสารระหว่างการประมาณขอบเขตล่างและบริเวณขอบเขตด้วยข้อมูลที่มีความไม่สอดคล้องกันสูง ๆ นั้นบางครั้งนำมาสู่การค้นพบซับเซตที่มีขนาดใหญ่กว่า RSAR และ DMRSAR อย่างไรก็ตามเมื่อพิจารณาถึงประสิทธิภาพการทำงานบน FS ด้วยการพิจารณาเฉพาะขนาดของซับเซตเพียงอย่างเดียวเป็นสิ่งที่ไม่เพียงพอ ในขณะที่สมรรถนะความถูกต้องการจำแนกประเภทข้อมูลก็ยังคงมีความสำคัญอย่างมากเช่นเดียวกัน

ตารางที่ 2 การเปรียบเทียบขนาดของซับเซตและเวลาที่ใช้กับชุดข้อมูลแบบไม่ต่อเนื่อง

ชุดข้อมูล	จำนวนของคุณลักษณะ	จำนวนแถวข้อมูล	ขนาดของซับเซต			เวลาที่ใช้ในการหารีดัก (นาที)		
			RSAR	DMRSAR	mBML	RSAR	DMRSAR	mBML
credit	21	1000	9	**8**	11	0.61	1.26	1.31
heart	14	294	7	**6**	8	0.01	0.03	0.05
vote	17	300	9	9	**8**	0.03	0.08	0.08
soybean	36	307	**13**	14	16	0.18	0.56	0.33
lymphography	19	148	**6**	8	7	0.01	0.03	0.01
promoters	58	106	**4**	**4**	5	0.01	0.01	0.03
splice	61	3190	**10**	**10**	11	18.35	39.36	46.61
dermatology	35	358	**7**	**7**	**7**	0.08	0.18	0.20
wq	39	521	**14**	15	15	0.68	1.43	1.06
dna	58	318	**4**	**4**	**4**	0.06	0.11	0.16

ตารางที่ 4 แสดงการเปรียบเทียบของจำนวนคุณลักษณะที่เลือกโดยแต่ละอัลกอริทึม FS โดยดำเนินการกับเซตข้อมูลที่มีค่าเป็นแบบค่าต่อเนื่อง ผลลัพธ์โดยส่วนใหญ่แสดงให้เห็นถึงวิธีการ mBML ได้รับขนาดเหมือน (แต่ประกอบด้วยคุณลักษณะที่แตกต่างกัน) กับวิธีการ RSAR และ DMRSAR อย่างไรก็ตาม mBML ได้แสดงให้เห็นถึงการเพิ่มของสมรรถนะอย่างน้อย 1 ตัวจำแนกประเภทข้อมูลและบางครั้งดีกว่าอย่างมีนัยสำคัญ อย่างไรก็ตามทุกวิธีการที่ถูกใช้ในงานวิจัยนี้จำเป็นที่ต้องมีการทำการแปลงข้อมูลแบบต่อเนื่องไปเป็นแบบไม่ต่อเนื่อง (Discretizing) กับเซตข้อมูลที่มีค่าต่อเนื่อง และแทนที่ค่าข้อมูลเริ่มต้นด้วยค่าแบบไม่ต่อเนื่อง บางครั้งกระบวนการนี้อาจจะนำมาสู่การสูญหายของข่าวสาร และบางครั้งเป็นสาเหตุของสัญญาณรบกวนในข้อมูล ดังนั้นแสดงว่าวิธีการ mBML ทนทานต่อสัญญาณรบกวนได้ดีกว่าทั้งวิธี RSAR และ DMRSAR จากการพิจารณาจากการลดขนาดมิติและค่าความถูกต้องที่เพิ่มขึ้นในการจำแนกประเภทข้อมูล



ภาพที่ 4 กราฟแสดงการเปรียบขนาดของซับเซตบนข้อมูลแบบไม่ต่อเนื่องกับวิธีการราฟเซต

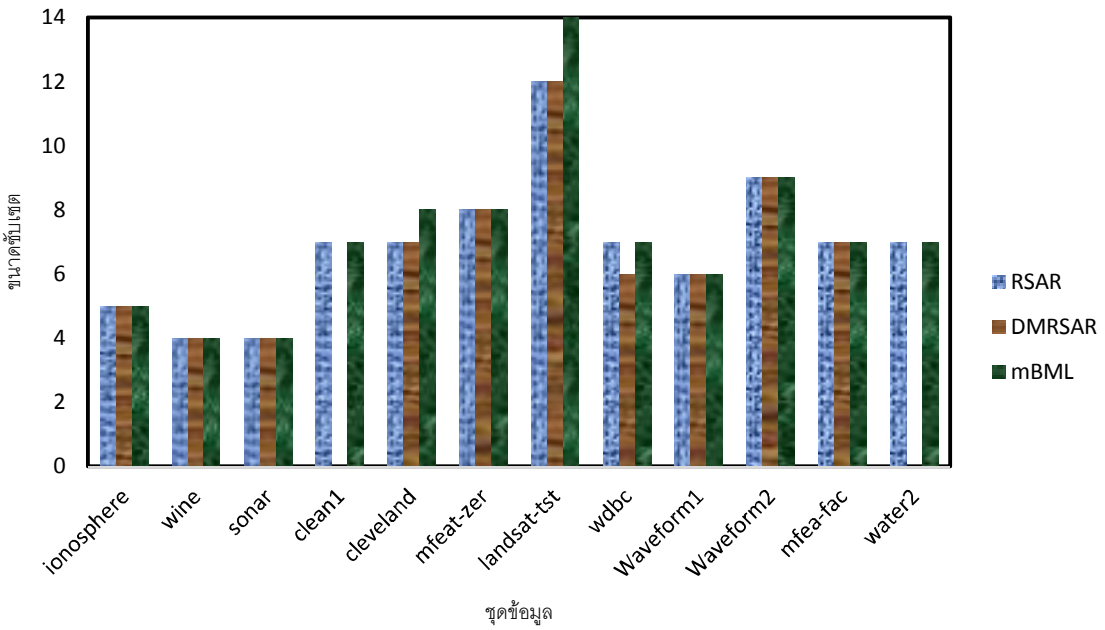
ตารางที่ 3 ค่าความถูกต้องเฉลี่ยในการจำแนกประเภทกับชุดข้อมูลแบบต่อเนื่อง

ชุดข้อมูล	SVM			J48			PART		
	RSAR	DMRSAR	mBML	RSAR	DMRSAR	mBML	RSAR	DMRSAR	mBML
ionosphere	**83.19**	83.13	81.48	89.74	89.45	**91.31**	86.60	87.74	**88.88**
wine	93.82	93.82	**95.5**	89.88	89.88	**96.06**	90.44	90.44	**94.38**
sonar	**76.44**	**76.44**	74.51	70.67	70.67	**75.48**	75.00	75.00	**77.88**
clean1	**69.11**	N/A	**69.11**	76.05	N/A	**77.73**	**75.84**	N/A	75.63
cleveland	81.48	81.48	**83.83**	**79.79**	**79.79**	79.12	74.74	74.74	**77.44**
mfeat-zer	67.00	67.00	**72.00**	60.90	60.35	**67.40**	63.30	63.30	**68.70**
landsat-tst	83.70	83.70	**84.90**	**85.75**	**85.75**	83.70	84.20	84.20	**85.20**
wdbc	95.43	94.90	**96.30**	**95.25**	92.72	93.97	94.20	93.84	**94.80**
Waveform1	76.40	76.40	**77.40**	69.20	69.20	**70.00**	70.00	70.00	**72.00**
Waveform2	76.76	76.76	**79.86**	73.70	73.70	**75.62**	72.98	72.98	**75.02**
mfea-fac	84.75	84.75	**86.50**	80.90	80.90	**83.40**	81.55	81.55	**83.20**
water2	79.65	N/A	**80.40**	**81.95**	N/A	**81.95**	**81.38**	N/A	80.99

ตารางที่ 2 และ 4 เป็นการบันทึกเวลาที่ใช้สำหรับวิธีการ RSAR DMRSAR และ mBML โดยการทดสอบบนเซตข้อมูลแบบไม่ต่อเนื่องและข้อมูลที่มีค่าต่อเนื่อง ซึ่งเวลาที่ใช้ของ mBML โดยประมาณเพิ่มขึ้น 2 เท่าเมื่อเปรียบเทียบกับRSAR แต่เพิ่มขึ้นเพียงเล็กน้อยเมื่อเปรียบเทียบ mBML กับ DMRSAR การเพิ่มขึ้นของเวลาที่ใช้ของวิธีการ mBML จะเกี่ยวข้องกับเวลาที่ถูกใช้ในการคำนวณค่าเทรสโฮลด์ (Threshold) ที่เหมาะสมของแต่ละซับเซตของแอตทริบิวต์

ตารางที่ 4 การเปรียบเทียบขนาดของซับเซตและเวลาที่ใช้กับชุดข้อมูลแบบต่อเนื่อง

ชุดข้อมูล	จำนวนของคุณลักษณะ	จำนวนแถวข้อมูล	ขนาดของซับเซต			เวลาที่ใช้ในการหารีดัก (นาที)		
			RSAR	DMRSAR	mBML	RSAR	DMRSAR	mBML
ionosphere	34	351	**5**	**5**	**5**	0.11	0.25	0.25
wine	14	178	**4**	**4**	**4**	0.01	0.01	0.01
sonar	61	208	**4**	**4**	**4**	0.08	0.15	0.18
clean1	167	476	**7**	N/A	**7**	0.86	N/A	2.20
cleveland	14	297	**7**	**7**	8	0.01	0.03	0.03
mfeat-zer	48	2000	**8**	**8**	**8**	4.98	10.03	8.90
landsat-tst	37	2000	**12**	**12**	14	8.00	10.13	15.56
wdbc	31	569	7	**6**	7	0.16	0.31	0.45
Waveform1	22	500	**6**	**6**	**6**	0.15	0.30	0.31
Waveform2	41	5000	**9**	**9**	**9**	24.30	47.68	58.61
mfea-fac	217	2000	**7**	**7**	**7**	36.20	37.01	34.45
water2	39	521	**7**	N/A	**7**	0.20	N/A	0.63



ภาพที่ 5 กราฟแสดงการเปรียบขนาดของซับเซตบนข้อมูลแบบต่อเนื่องกับวิธีการราฟเซต

อย่างไรก็ตามในบางกรณี mBML สามารถทำงานได้เร็วกว่า DMRSAR เมื่อบริเวณขอบเขตซับเซตของคุณลักษณะประกอบด้วยวัตถุจำนวนมาก ซึ่งเวลาที่ถูกใช้ไปในการคำนวณระยะทางระหว่างการประมาณขอบเขตล่าง และบริเวณขอบเขตจะเพิ่มขึ้น (ตัวอย่าง mfeat-zer)

4.2 การเปรียบเทียบผลการทดลองของ CNS ReliefF และ mBML

ในส่วนนี้ mBML จะเปรียบเทียบกับ 2 อัลกอริทึมการเลือกคุณลักษณะที่เป็นที่นิยมกัน โดยเราใช้โปรแกรม ReliefF และ CNS จากซอฟต์แวร์ Weka โดยตัวประเมินซับเซตบนพื้นฐานความสอดคล้อง (CNS) เป็นการเลือกซับเซตแอตทริบิวต์ซึ่งใช้ความสอดคล้องคลาสเป็นตัววัดการประเมินความดีของ

ซับเซต ในงานวิจัยนี้วิธีการ CNS ใช้การวัดความสอดคล้องแบบ Liu และ Setiono's (Liu and Setinono, 1996) และวิธีสุดท้าย คือ ReliefF (Kira and Rendell, 1997; Kononenko, 1994) เป็นวิธีการจัดเรียงแอตทริบิวต์บนหลักการของแถวข้อมูลซึ่งถูกนำเสนอโดย Kira และ Rendell (1997) และต่อมามีการปรับปรุงขึ้น โดย Kononenko (1994) โดยที่วิธีการ ReliefF มีการทำงานโดยการสุ่มเลือกตัวอย่างข้อมูลและทำการสร้างย่านจุดที่อยู่ใกล้ที่สุดกับแถวข้อมูลนั้นจากคลาสเดียวกันและคลาสตรงกันข้ามกัน

เพื่อให้แน่ใจถึงการเปรียบเทียบของ mBML และ ReliefF มีความสมเหตุสมผลกัน เราจะเลือกขนาดของซับของวิธีการ REliefF ที่เท่ากันกับวิธี mBML โดยการเปรียบเทียบผลลัพธ์ของ mBML กับสองวิธีจะแสดงทั้งในเทอมของขนาดของซับเซตและความถูกต้องการจำแนกประเภทข้อมูล

4.2.1 ความถูกต้องการจำแนกประเภทของข้อมูล (Classification Accuracy)

ตารางที่ 5 และ 7 แสดงถึงการเปรียบเทียบของความถูกต้องการจำแนกประเภทข้อมูลบนข้อมูลแบบไม่ต่อเนื่อง และแบบต่อเนื่อง สำหรับวิธีการ CNS ReliefF และ mBML ผลลัพธ์จากสองตารางนี้เป็นสิ่งที่แสดงถึงความโดดเด่นของวิธีการ mBML เกือบทุกข้อมูลที่ดำเนินการได้ดีมาก ๆ และแสดงถึงความถูกต้องการจำแนกประเภทข้อมูลที่สูงกว่า CNS ถึงแม้ผลลัพธ์ที่ได้รับของ mBML มีค่าความถูกต้องการจำแนกประเภทข้อมูลที่ลดลงบางกรณีของตัวจำแนกประเภทข้อมูลและบางข้อมูลที่เปรียบเทียบกับ CNS แต่การลดลงนี้เป็นแบบน้อย ๆ ในทางตรงกันข้ามในกรณีส่วนใหญ่ของข้อมูลแบบไม่ต่อเนื่องและค่าข้อมูลแบบทศนิยม mBML ได้แสดงถึงจำนวนที่เพิ่มขึ้นอย่างมากในสมรรถนะของค่าความถูกต้องการจำแนกประเภทข้อมูลเมื่อเปรียบเทียบกับ CNS และนอกจากนี้ mBML ยังให้ผลดีกว่า CNS ในกรณีโดยส่วนใหญ่และบางครั้งดีกว่าอย่างมีนัยสำคัญ

ตารางที่ 5 ค่าความถูกต้องเฉลี่ยในการจำแนกประเภทกับชุดข้อมูลแบบไม่ต่อเนื่อง

เซตข้อมูล	SVM			J48			PART		
	CNS	ReliefF	mBML	CNS	ReliefF	mBML	CNS	ReliefF	mBML
credit	74.20	**75.30**	74.90	72.20	**72.40**	72.00	**74.10**	73.00	72.00
heart	80.61	**83.33**	80.61	**79.25**	**79.25**	**79.25**	80.95	78.57	**80.95**
vote	93.33	93.66	**95.66**	93.66	94.00	**94.50**	93.33	93.00	**94.00**
soybean	84.36	88.9	**89.57**	80.45	85.66	**86.64**	76.22	82.41	**85.66**
lymphography	79.72	**83.10**	**83.10**	74.32	**76.35**	73.64	76.35	**77.02**	**77.02**
promoters	85.84	**95.22**	94.33	84.90	83.01	**84.96**	**93.39**	86.79	**93.39**
splice	94.26	94.29	**94.98**	93.82	94.04	**94.48**	93.13	92.97	**93.79**
dermatology	75.69	76.25	**87.43**	72.06	75.69	**82.40**	63.40	75.13	**77.93**
wq	**75.04**	73.12	72.20	66.02	**67.90**	66.02	67.56	66.79	**69.09**
dna	93.71	**94.65**	93.08	83.64	83.64	**84.27**	85.84	**86.16**	85.84

จากผลลัพธ์ในตารางที่ 5 และ 7 ที่ได้รับโดย mBML มีการลดลงของมิติข้อมูลในกรณีของบางตัวจำแนกประเภทข้อมูลและบางเซตข้อมูลที่ซึ่งมันมีขนาดเท่ากันกับของ ReliefF การลดลงนี้ลดลงอย่างไม่มีนัยสำคัญ อย่างไรก็ตามมันเป็นความน่าประทับใจที่ประสิทธิภาพและความสามารถของ mBML สามารถให้ความถูกต้องสูงกว่าเมื่อเปรียบเทียบกับ ReliefF ในทุก ๆ ตัวจำแนกประเภทข้อมูล เห็นได้ชัดว่าค่าความถูกต้องการจำแนกประเภทข้อมูลบนข้อมูลแบบไม่ต่อเนื่อง (ตัวอย่าง soybean dermatology เป็นต้น) และข้อมูลมีค่าแบบต่อเนื่อง (เช่น ionosphere mfeat-zer landsat wdbc mfeat-fac เป็นต้น) mBML แสดงให้เห็นถึงผลที่

ดีกว่า ReliefF ในทุก ๆ ตัวจำแนกประเภทข้อมูล ผลลัพธ์ในตารางที่ 7 ยังได้แสดงถึงเซต ข้อมูล 12 ข้อมูล โดยเฉพาะอย่างยิ่ง waveform1 และ waveform2 ที่แสดงให้เห็นถึงสมรรถนะค่าความถูกต้องการจำแนกประเภทที่ค่าลดลงไม่มากเมื่อเปรียบเทียบกับ ReliefF ค่าที่ลดลงนี้ไม่มากและ mBML ยังให้ผลที่ดีกว่า CNS ในกรณีอื่นทั้งหมดบางครั้งก็ดีกว่าอย่างมีนัยสำคัญ

ตารางที่ 6 การเปรียบเทียบขนาดของซับเซตบนชุดข้อมูลแบบไม่ต่อเนื่อง

ชุดข้อมูล	จำนวนของคุณลักษณะ	จำนวนแถวข้อมูล	ขนาดของซับเซต		
			CNS	RelifeF	mBML
credit	21	1000	**8**	11	11
heart	14	294	**8**	**8**	**8**
vote	17	300	10	**8**	9
soybean	36	307	**11**	16	16
lymphography	19	148	**7**	**7**	**7**
promoters	58	106	**4**	5	5
splice	61	3190	**10**	11	11
dermatology	35	358	**6**	7	7
wq	39	521	**14**	15	15
dna	58	318	**4**	**4**	**4**

4.2.2 ขนาดของซับเซต

ถึงแม้ว่า mBML จะให้ขนาดซับเซตที่ใหญ่กว่า CNS ในบางเซตข้อมูลแต่ mBML ได้แสดงให้เห็นถึงค่าความถูกต้องในการจำแนกประเภทข้อมูลที่ดีกว่า CNS โดยส่วนใหญ่ทั้งเซตข้อมูลแบบไม่ต่อเนื่องและค่าแบบต่อเนื่อง ในทุกกรณี mBML จะมีขนาดซับเซตเท่ากันกับ ReliefF โดยขนาดซับเซตที่เท่ากันที่ถูกค้นพบโดย mBML จะถูกใช้เป็นจำนวนลักษณะเด่นของการจัดเรียงโดย ReliefF มีเพียงจำนวนเล็กน้อยเท่านั้นบนข้อมูลแบบไม่ต่อเนื่องที่ ReliefF แสดงถึงจำนวนที่เพิ่มขึ้นเพียงเล็กน้อยในแง่ของความถูกต้องการจำแนกประเภทข้อมูล อย่างไรก็ตาม mBML แสดงให้เห็นถึงจำนวนที่เพิ่มขึ้นในความถูกต้องการจำแนกประเภทข้อมูลและ ดีกว่า ReliefF ในกรณีอื่น ๆ ทุกกรณีบางครั้งก็ดีกว่าอย่างมีนัยสำคัญ

ในทำนองเดียวกันสำหรับทั้งเซตข้อมูลค่าแบบต่อเนื่องซึ่ง mBML สามารถปรับปรุงความถูกต้องการจำแนกประเภทข้อมูลของทุก ๆ ตัวจำแนกประเภทข้อมูล และการปรับปรุงที่เห็นได้ชัดอย่างมากบนเซตข้อมูล ionosphere mfeat-zer landsat wdbc และ mfeat-fac ดังนั้นข้อมูลเหล่านี้แสดงให้เห็นถึงซับเซตของแอตทริบิวต์ประกอบด้วยข่าวสารที่มีมูลค่าอย่างมากที่ถูกสกัดออกมาจากข้อมูลเหล่านี้ด้วยวิธีการ mBML

ตารางที่ 7 ค่าความถูกต้องเฉลี่ยในการจำแนกประเภทกับชุดข้อมูลแบบต่อเนื่อง

ชุดข้อมูล	SVM			J48			PART		
	CNS	RelifeF	mBML	CNS	RelifeF	mBML	CNS	RelifeF	mBML
ionosphere	81.48	**83.19**	81.48	90.31	88.31	**91.31**	**88.88**	**88.88**	**88.88**
wine	93.25	94.94	**95.50**	**96.62**	89.88	96.06	**94.38**	93.82	**94.38**
sonar	72.11	**75.96**	74.51	75.00	71.63	**75.48**	77.40	71.15	**77.88**
clean1	65.96	61.13	**69.11**	75.63	68.69	**77.73**	75.42	68.48	**75.63**
cleveland	82.15	**83.83**	**83.83**	78.11	**81.48**	79.12	77.77	**79.13**	77.44
mfeat-zer	63.95	71.15	**72.00**	59.50	66.70	**67.40**	60.70	67.75	**68.70**
landsat-tst	83.50	81.85	**84.90**	81.30	82.10	**83.70**	83.55	82.65	**85.20**
wdbc	**96.48**	96.30	96.30	**94.20**	**94.20**	92.97	94.20	94.43	**94.80**
Waveform1	77.40	**80.40**	77.40	70.00	**72.60**	70.00	67.60	71.40	**72.00**
Waveform2	78.14	**82.36**	79.86	74.16	**77.34**	74.62	73.90	**78.00**	75.02
mfea-fac	84.40	83.19	**86.50**	81.15	71.60	**83.40**	81.45	69.90	**83.20**
water2	78.31	**94.94**	80.40	80.23	**82.34**	81.95	78.88	**82.72**	80.99

ตารางที่ 8 การเปรียบเทียบขนาดของซับเซตบนเซตข้อมูลแบบต่อเนื่อง

ชุดข้อมูล	จำนวนของคุณลักษณะ	จำนวนแถวข้อมูล	ขนาดของซับเซต		
			CNS	RelifeF	mBML
ionosphere	34	351	**5**	**5**	**5**
wine	14	178	**4**	**4**	**4**
sonar	61	208	**4**	**4**	**4**
clean1	167	476	8	**7**	**7**
cleveland	14	297	**7**	8	8
mfeat-zer	48	2000	**7**	8	8
landsat-tst	37	2000	**13**	14	14
wdbc	31	569	**6**	7	7
Waveform1	22	500	**6**	**6**	**6**
Waveform2	41	5000	**9**	**9**	**9**
mfea-fac	217	2000	**6**	7	7
water2	39	521	**6**	7	7

อย่างไรก็ตาม mBML เป็นวิธีการเลือกลักษณะเด่นที่ตรงกันข้ามกับวิธีการจัดเรียงคุณลักษณะโดยที่ค่าเทรสโฮลด์ไม่ต้องกำหนดมาก่อนและการเลือกจะเสร็จสมบูรณ์ทันทีที่กฎเกณฑ์การหยุดเป็นจริง แต่สำหรับวิธีการ ReliefF นั้นจำเป็นที่ต้องมีการกำหนดจำนวนของย่านจุดที่ใกล้ที่สุด (K) มาก่อน ซึ่งในงานวิจัยนี้ค่า K กำหนดให้มีค่าเป็น 10 โดยข้อเสียที่สำคัญของวิธีการจัดเรียงลักษณะเด่นคือ มันอาจจะล้มเหลวในการประยุกต์กับฟังก์ชันการจำแนกประเภทข้อมูลที่ขึ้นกับคุณลักษณะที่เกิดขึ้นพร้อมกันตั้งแต่สองลักษณะเด่นขึ้นไป (เช่น ปัญหา XOR) อย่างไรก็ตามปัญหานี้สามารถถูกจัดการได้ด้วยวิธีการเลือกซับเซตลักษณะเด่น

โดยทั่วไปแล้วเวลาที่ใช้ในการหารีดักของเทคนิค FS บนหลักการของการพาติชั่นข้อมูล (ได้แก่ วิธีการบนหลักการราฟเซต) โดยส่วนใหญ่แล้วจะช้ากว่าทั้งวิธีการเลือกซับเซตบนหลักการค่าความสอดคล้องกัน (Consistency) และการจัดเรียงแอตทริบิวต์บน

หลักการของแถวข้อมูล (Instance) อย่างไรก็ตาม การพิจารณาถึงเฉพาะเวลาที่ใช้ในการหาซับเซตของลักษณะเด่นเพียงอย่างเดียวอาจจะไม่เพียงพอกับงาน FS ในขณะที่สมรรถนะในเรื่องความถูกต้องการจำแนกประเภทข้อมูลก็สำคัญอย่างมากเช่นเดียวกัน

สรุปและวิจารณ์ผล

การวัดปริมาณข่าวสารทั้งในบริเวณของความแน่นอนและบริเวณขอบเขตโดยใช้หลักการของ VPRS ได้ค่าความถูกต้องการจำแนกประเภทได้ผลลัพธ์ที่ดีมาก ๆ ซึ่งวิธีการที่นำเสนอได้ผลที่ดีกว่า RSAR และ DMRSAR โดยเฉพาะอย่างยิ่งกับข้อมูลที่มีสัญญาณรบกวนและข้อมูลที่ซึ่งส่วนเล็ก ๆ ของแต่ละแอตทริบิวต์มีความขัดแย้งกัน การเลือกซับเซตด้วยวิธีการที่นำเสนอจะเลือกข่าวสารของบริเวณที่มีความแน่นอนมีค่าสูงสุดและในเวลาเดียวกันข่าวสารของบริเวณขอบเขตมีค่าต่ำสุด ทำให้เกิดประสิทธิภาพที่ดีขึ้นมากเมื่อเปรียบเทียบกับวิธีการที่ใช้หลักการของ dependency function ดังนั้น เป็นสิ่งชัดเจนว่าซับเซตของลักษณะเด่นที่ได้รับจากวิธีการที่นำเสนอนั้นจะบรรจุข่าวสารที่มีความสำคัญมากกว่า ที่ได้รับโดยวิธีการ RSAR และ DMRSAR

ในงานวิจัยนี้ เราได้นำมิวชวลอินฟอร์เมชั่นมาคำนวณความสำคัญของซับเซตบนชุดข้อมูลที่ถูกพาร์ติชั่นโดยการใช้ VPRS โดยงานวิจัยนี้ได้ทำการเลือกค่า β ที่เหมาะสมจากค่า β ที่ทำให้บริเวณขอบเขตของซับเซตคุณลักษณะมีปริมาณข่าวสารน้อยที่สุดแทนที่จะเป็นการรับมาจากมนุษย์ แล้วจากนั้นนำ β ที่เลือกนั้นมาคำนวณหาการประมาณขอบเขตล่างและการประมาณขอบเขตบน แน่นอนว่าค่า β ที่เหมาะสมนั้นจะทำให้ซับเซตที่เลือกมานั้นจะมีข่าวสารของความแน่นอนมากที่สุดนั่นเอง ผลการทดลองได้เน้นย้ำถึงข่าวสารที่มีความสำคัญสามารถสกัดออกมาโดยการทำให้ข่าวสารของการประมาณขอบเขตล่างมีค่ามากที่สุดและในเวลาเดียวกันข่าวสารที่บรรจุอยู่ในบริเวณขอบเขตมีค่าน้อยที่สุด จากผลการทดลองวิธีการที่นำเสนอสามารถให้ค่าความถูกต้องในการจำแนกประเภทที่เพิ่มขึ้นกว่าวิธีการอื่น ๆ บนหลักการราฟเซตในขณะที่ประสิทธิภาพการลดขนาดมิติข้อมูลก็ให้ผลเป็นที่น่าพอใจ ไม่ได้แตกต่างมากมายกว่าวิธีการอื่น ๆ

เอกสารอ้างอิง

Aghdam, M.H., Ghasem-Aghaee, N., and Basiri, M.E. 2009. Text feature selection using ant colony optimization. Expert Systems with Applications. 36: 6843-6853.

Battiti, R. 1994. Using mutual information for selecting features in supervised neural net learning. IEEE Trans. Neural Networks. 5(4): 537 - 550.

Blum, A.L., and Langley, P. 1997. Selection of relevant features and examples in machine learning. Artificial Intelligence. 97: 245-271.

Chen, Y., Miao, D. and Wang, R. 2010. A rough set approach to feature selection based on ant colony Optimization. Pattern Recognition Letters. 31(3): 226-233.

Dash, M., and Liu, H. 1997. Feature selection for classification. Intelligent Data Analysis: An Int'l J. 1(3): 131-156.

Deogun, J.S., Raghavan, V.V. and Sever, H. 1995. Exploiting upper approximation in the rough set Methodology. Proc. First Int'l Conf. Knowledge Discovery and Data Mining: 69-74.

Foithong, S., Srinil, P., Tawsopar Yangyuen, K. and Phattaraworamet, T. 2016. Rough-Mutual Feature Selection Based-on Minimal-Boundary and Maximal-Lower. Int'l Conf. Management and Innovation Technology International Conference (MITicon): 137-141.

Han, S.W. and Kim, J.Y. 2008. A new decision tree algorithm based on rough set theory. Int'l J. Innovative Computing, Information and Control. 4(10): 2749-2757.

Hassanien, A. 2004. Rough set approach for attribute reduction and rule generation: A case of patients with suspected breast cancer. J. Am. Soc. Information Science and Technology. 55(11): 954-962.

Hedar, A. Wang, J. and Fukushima, M. 2006. Tabu search for attribute reduction in rough set theory. Technical Report 2006-008. Dept. of Applied Mathematics and Physics. Kyoto Univ.

Hu, Q., Yu, D., Liu, J. and Wu, C. 2008. Neighborhood rough set based heterogeneous feature subset selection. Information Sciences. 178(15): 3577-3594.

Inuiguchi, M. and Tsurumi, M. 2006. Measures based on upper approximations of rough sets for analysis of attribute importance and interaction. Int'l J. Innovative Computing, Information and Control. 2(1): 1-12.

Jensen, R. and Shen, Q. 2004. Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based Approaches. IEEE Trans. Knowledge and Data Eng. 16(12): 1457-1471.

Kim, Y., Street W. and Menczer, F. 2000. Feature selection for unsupervised learning via evolutionary. Proc. Sixth ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining: 365-369.

Kira, K. and Rendell, L. 1992. A practical approach to feature selection. Proc. Ninth Int'l Conf. Machine Learning: 249-256.

Kohavi, R. and John, G. H. 1997. Wrappers for feature subset selection. Artificial Intelligence. 97(1-2): 273-324.

Kononenko, I. 1994. Estimating attributes: analysis and extensions of Relief. Proc. Seventh European Conf. Machine Learning: 171-182.

Lee, W., Stolfo, S. J. and Mok, K.W. 2000. Adaptive intrusion detection: a data mining approach. AI Rev. 14(6): 533-567.

Li, Y. Shiu, S.C.K., Pal, S.K. and Liu, J.N.K. 2006. A rough set-based case-based reasoner for text Categorization. International Journal of Approximate Reasoning. 41(2): 229-255.

Liu, H. and Setiono, R. 1996. A probabilistic approach to feature selection: a filter solution. Proc. 13th Int'l Conf. Machine Learning: 319-327.

Miao, D., Duan, Q., Zhang, H. and Jiao, N. 2009. Rough set based hybrid algorithm for text classification. Expert Systems with Applications. 36: 9168-9174.

Mun, G.J., Noh, B.N. and Kim, Y.M. 2009. Enhanced stochastic learning for feature selection in intrusion Classification. Int'l J. Innovative Computing, Information and Control. 5(11): 3625 - 3635.

Newman, D.J., Hettich, S., Blake, C.L. and Merz, C.J. 1998. UCI repository of machine learning Databases. dept. of information and computer science. Univ. of California-Irvine. (http://www.ics.uci.edu/mlearn/MLRepository.html).

Parthal´ain, N.M., Jensen, R. and Shen, Q. 2010. A distance measure approach to exploring the rough set boundary region for attribute reduction. IEEE Transactions on Knowledge and Data Engineering. 22(3): 305-317.

Pawlak, Z. 1991. Rough sets: Theoretical aspects of reasoning about data. Kluwer Academic Publishing.

Pawlak, Z.1982. Rough sets. Int. J. Inf. Comput. Sci. 11: 314-356.

Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y. and Wang, Z. 2007. A novel feature selection algorithm for text categorization. Expert Systems with Applications. 33: 1-5.

Shannon, C.E. and Weaver, W. 1949. The mathematical theory of communication. University of Illinois Press. Urbana, Illinois.

Ziarko, W. 1993. Variable precision rough set model. J. Comput. Syst. Sci. 46(1): 44-54.

Ziarko, W. 2008. Probabilistic approach to rough sets. International Journal of Approximate Reasoning. 49(2): 272-284.