JIST Journal of Information Science and Technology
Volume 11, NO 1 | JAN – JUN 2021 | 83-91
ISSN: 2651-1053 (Online)
ISSN: 1906-9553 (Print)

การเปรียบเทียบวิธีการคัดเลือกคุณลักษณะด้วยริลีฟเอฟสำหรับหลายป้ายคำศัพท์เพื่อปรับปรุงการจำแนกความหมายภาพ

Comparison of Feature Selection Method with ReliefF base on Multi Label Algorithm to Improve Semantic Image Classification

เตชค์ฐสิณป์ เพียซ้าย[1] และ นัศพ์ชาณัณ ชินปัญช์ธนะ[2]

Tejtasin Phiasai[1] and Nutchanun Chinpanthana[2]

[1]สาขาวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสุโขทัยธรรมาธิราช

[2]วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์

[1]School of Science and Technology, SukhothaiThammathirat Open University, Chaengwattana rd.

[2]College of Innovative Technology and Engineering, Dhurakij Pundit University

Received: April 10, 2021; Revised: May 25, 2021; Accepted: June 19, 2021; Published: June 25, 2021

ABSTRACT – Image classification is developed as part of a framework in digital image processing. The image extraction in relevant features is the most challenging part of classification. The performance of algorithm depends on features considered from the dataset. There are many algorithms attempt to analyze and combine all features but high dimensional dataset degrades the performance of algorithm. To overcome this problem, we propose a ReliefF base on Multi Label Algorithm (ReliefF-ML) to improve semantic image classification. Feature selection technique is used as a first step to analyze in large data set that is to find relevant features and removes redundant features from high dimensional dataset. This paper presents four steps including (1) data preprocessing, (2) feature extraction, (3) ReliefF-ML feature selection model, and (4) efficiency measurement and evaluation of experimental results. The experimental results indicate that our framework offers performance improvements with ReliefF-ML feature selection algorithm. The proposed model can achieve significant improvements for image classification with maximum success rate of 78.87% with 12 features.

KEYWORDS: Image classification, Digital image processing, Feature selection, ReliefF, Semantic images

บทคัดย่อ -- การจำแนกข้อมูลภาพได้ถูกพัฒนามาจากส่วนหนึ่งของการประมวลผลภาพดิจิทัล การนำคุณลักษณะข้อมูลที่มีความเกี่ยวข้องมาใช้งานยังคงเป็นส่วนสำคัญสำหรับการจำแนกข้อมูลภาพ โดยส่วนมาก ประสิทธิภาพของอัลกอริทึมยังคงขึ้นกับข้อมูลที่ถูกเลือกมาใช้จากฐานข้อมูลและอัลกอริทึมส่วนใหญ่จะประกอบด้วยข้อมูลหลายรูปแบบที่นำมาประกอบเข้าด้วยกัน ทำให้ชุดข้อมูลมีขนาดใหญ่จนทำให้ประสิทธิภาพของอัลกอรทึมลดลง สิ่งนี้จึงกลายเป็นปัญหาดังนั้นในงานวิจัยนี้ได้นำเสนอการคัดเลือกคุณลักษณะด้วยวิธีการริลีฟเอฟสำหรับหลายป้ายคำศัพท์เพื่อปรับปรุงการจำแนก

*Corresponding Author: *nutchanun.cha@dpu.ac.th*

ความหมายภาพ ซึ่งการคัดเลือกคุณลักษณะที่เหมาะสมจะถูกใช้เป็นขั้นตอนแรกเพื่อลดทอนคุณลักษณะข้อมูลที่ไม่เกี่ยวข้อง วิธีการดำเนินงาน ประกอบด้วย 4 ขั้นตอนดังนี้ (1) ขั้นตอนการเตรียมข้อมูล (2) การสกัดข้อมูลภาพ (3) การคัดเลือกคุณลักษณะข้อมูลด้วยริลีฟเอฟสำหรับหลายป้ายคำศัพท์และ (4) การวัดประสิทธิภาพและประเมินผล จากการทดลองสามารถระบุได้ว่าการใช้อัลกอริทึมริลีฟเอฟสำหรับหลายป้ายคำศัพท์สามารถเพิ่มประสิทธิภาพของการจำแนกข้อมูลภาพได้ดียิ่งขึ้นถึงร้อยละ 78.87 ด้วย 12 คุณลักษณะ

คำสำคัญ: การจำแนกข้อมูลภาพ, การประมวลผลภาพดิจิทัล, การคัดเลือกคุณลักษณะ, ริลีฟเอฟ, ความหมายภาพ

1. บทนำ

ปัจจุบันการพัฒนาทางเครื่องมือทางด้านเทคโนโลยีและอุปกรณ์ถ่ายภาพดิจิทัลได้พัฒนาอย่างรวดเร็ว เพื่อตอบสนองการใช้งานทางด้านภาพถ่ายดิจิทัลไม่ว่าจะเป็นอุปกรณ์มือถือ อินเทอร์เน็ท ทำให้การจัดเก็บข้อมูลมัลติมีเดียรวมทั้งการค้นคืนข้อมูลในฐานข้อมูลขนาดใหญ่ยังคงมีการพัฒนาอย่างต่อเนื่องเพื่อให้สามารถนำข้อมูลมาใช้อย่างมีประสิทธิภาพและรวดเร็วสูงสุด เช่นเดียวกันกับงานวิจัยทางด้านการประมวลผลภาพ (digital image processing) ที่พยายามนำเทคนิคการค้นคืนด้วยคุณลักษณะของภาพ [1][2] เช่น รูปร่าง ขนาด สี หรือ พื้นผิว ที่อยู่ในรูปแบบค่าตัวเลขในรูปแบบเวกเตอร์คุณลักษณะเฉพาะเพื่อใช้เป็นดัชนีภาพสำหรับการจัดเก็บภาพลงในฐานข้อมูล วิธีการนี้ยังคงต้องมีการปรับปรุงอย่างต่อเนื่อง แต่วิธีการสร้างป้ายกำกับด้วยคำศัพท์ (keyword annotation) [3] เป็นการนำข้อมูลคำศัพท์ลงในฐานข้อมูลภาพ ซึ่งเป็นอีกหนึ่งวิธีที่นิยมนำมาใช้ในกระบวนการค้นคืนภาพ ทำให้สามารถค้นคืนความหมายวัตถุบนภาพได้อย่างรวดเร็ว รูปแบบของการให้ความหมายจะเป็นการเขียนคำศัพท์หรือข้อความเป็นป้ายกำกับที่มีความสัมพันธ์กันด้วยมนุษย์บนแอพลิเคชันที่รองรับ ข้อมูลเหล่านี้ได้ถูกนำมาผ่านกระบวนการวิเคราะห์ และผสมผสานกันระหว่างคุณลักษณะเพื่อนำมาใช้ในกระบวนการค้นคืนในรูปแบบที่ซับซ้อนขึ้น บางกลุ่มได้มีการแทนข้อมูลในรูปแบบอนุกรมวิธาน WordNet [4] เป็นฐานข้อมูลพจนานุกรมภาษาอังกฤษที่มีโครงสร้างแบบลำดับขั้นตามความสัมพันธ์และความหมายคำศัพท์ และในบางกลุ่มวิจัยได้นำภาพในลักษณะเดียวกัน เพิ่มข้อมูลสาระสำคัญของภาพ เช่น ชื่อสถานที่ ชื่อหัวเรื่อง หรือชื่อเหตุการณ์ [5] แต่อย่างไรก็ตามผลลัพธ์ที่ไม่ถูกต้องตามกลุ่มการจำแนก เนื่องจากการแทนข้อมูลด้วยคำศัพท์นั้นถูกสร้างความสัมพันธ์มีความซับซ้อนจาก โครงสร้างมากจนเกินไป

2. งานวิจัยที่เกี่ยวข้อง

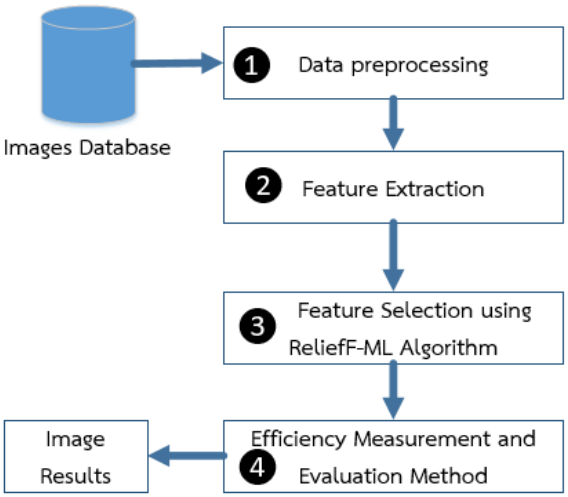
โดยทั่วไปแล้วชุดข้อมูลของภาพดิจิทัล จะมีปริมาณข้อมูลขนาดใหญ่และมีคุณลักษณะภาพที่ถูกสกัดออกมาจำนวนมาก ข้อมูลบางคุณลักษณะจะมีผลที่เป็นเชิงบวกแต่บางคุณลักษณะอาจจะมีผลกระทบที่เป็นเชิงลบกับประสิทธิภาพของการเรียนรู้ในแต่ละกลุ่ม[6] ดังนั้นจึงทำให้เกิดกระบวนการต่างๆเพื่อนำข้อมูลมาวิเคราะห์ประมวลผลเพื่อแปรเปลี่ยนเป็นสารสนเทศให้ตรงตามความต้องการของผู้ใช้งานทำให้การเรียนรู้ของเครื่องเข้ามาช่วยใน การพัฒนากระบวนการวิเคราะห์ความหมายของภาพ รวมทั้งยังสามารถปรับการเรียนรู้และค้นหาข้อมูลใหม่ๆ เพื่อให้เกิดข้อผิดพลาดจากการเรียนรู้น้อยที่สุดการขจัดคุณลักษณะข้อมูลไม่สัมพันธ์กัน ข้อมูลที่มีความซ้ำซ้อนออกไปเพราะกระบวนการเรียนรู้ของเครื่องจากคุณลักษณะข้อมูลจำนวนมากอาจทำให้เกิดความผิดพลาดได้ ถ้าข้อมูลนั้นส่งผลเสียต่อภาพรวม จะทำให้เกิดผลกระทบต่อการเรียนรู้และยังทำให้เปลืองพื้นที่ เปลืองทรัพยากรของเครื่อง [7] โดยทั่วไปวิธีการที่นำมาใช้ในการแก้ไขปัญหา นักวิจัย El-Naggar et al. [10] และ Omar et al. [11] ได้ทดลองการคัดเลือกคุณลักษณะข้อมูลที่มีความสัมพันธ์กันด้วย SVM, Naive Bayes, และ K-nearest neighbor (K-nn) เพื่อวิเคราะห์ข้อมูลที่เป็นรูปแบบของคำศัพท์ ผลที่ได้มีค่าความถูกต้องสูงถึง 90% จากงานวิจัยที่กล่าวมาแล้วนั้นสามารถยืนยันได้ว่าการคัดเลือกคุณลักษณะข้อมูลที่ดีเพื่อใช้ในการจำแนกจะมีส่วนช่วยทำให้ค่าความถูกของการจำแนกต้องดีขึ้นด้วยจะเห็นว่าวิธีการลดจำนวนข้อมูลลง และเหลือเฉพาะข้อมูลที่มีคุณภาพที่เหมาะสมกับกลุ่มข้อมูลมากที่สุด เพื่อเป็นข้อมูลเข้าของระบบ จะเป็นอีกสิ่งหนึ่งที่ช่วยลดปัญหาและทำให้ได้ผลลัพธ์ตรงตามความต้องการมากขึ้น

วิธีการคัดเลือกคุณลักษณะข้อมูล (feature selection)[8][9] ในรูปแบบของการคัดเลือกข้อมูลจากการวิเคราะห์ความรู้สึก

(sentiment analysis) มีอยู่หลายวิธีตัวอย่างเช่น Document frequency, Information Gain (IG), Fisher Score, Mutual Information, ReliefF, Chi-Square, Support Vector Machine (SVM), Naïve Bayes, Decision Trees, K-Nearest Neighbor (K-NN), Maximum Entropy เป็นต้น และอัลกอริทึมริลีฟเอฟ (ReliefF algorithm) [12] เป็นอีกหนึ่งวิธีที่นิยมใช้สำหรับ การคัดเลือกคุณลักษณะ ซึ่งมีหลายงานวิจัยที่พยายามปรับเปลี่ยนอัลกอริทึมให้สามารถใช้งานได้ในรูปแบบของหลายชุดข้อมูลจากเดิมที่ใช้ได้เพียงแค่ชุดข้อมูลเดียว BR-ReliefF (Binary Relevance ReliefF) [12] ถูกนำเสนอสำหรับข้อมูลป้ายคำศัพท์ที่เพิ่มขึ้น โดยอยู่บนพื้นฐานของวิธีการ Information Gain [13] ด้วยการหาข้อมูลที่มีผลกระทบมากที่สุดกับชุดข้อมูล แต่อย่างไรก็ตามวิธี BR-ReliefF ยังไม่ได้คำนึงถึงความสัมพันธ์ระหว่างข้อมูลภายใน ดังนั้นในงานวิจัยนี้ได้ใช้อัลกอริทึมริลีฟเอฟสำหรับหลายป้ายคำศัพท์ (ReliefF base on Multi Label algorithm : ReliefF-ML)[13] [14] เป็นการคัดเลือกคุณลักษณะแบบหลายชุดข้อมูลด้วยการปรับค่าน้ำหนักเพื่อสร้างความสัมพันธ์ของคุณลักษณะข้อมูลให้เหมาะกับแต่ละป้ายคำศัพท์ และหาความสัมพันธ์แบบคู่ชุดข้อมูลด้วยการพิจารณาจากความถี่และความสัมพันธ์กันระหว่างชุดข้อมูล [15][16] ในงานวิจัยนี้ได้แบ่งเป็น 4 ขั้นตอนดังนี้ (1) การเตรียมข้อมูล (data preprocessing) (2) การสกัดข้อมูล (feature extraction) (3) การคัดเลือกคุณลักษณะข้อมูลด้วย ReliefF-ML (feature selection with ReliefF-ML) และ (4) การวัดประสิทธิภาพและวิธีการประเมินผล (Efficiency Measurement and Evaluation method) ดังแสดงในรูปที่ 1

3. วิธีการวิจัย

ในงานวิจัยได้นำเสนอ วิธีการคัดเลือกคุณลักษณะข้อมูลที่เหมาะสมสำหรับการจำแนกข้อมูลความหมายภาพด้วยอัลกอริทึมริลีฟเอฟสำหรับหลายป้ายคำศัพท์ (ReliefF-ML) มีขั้นตอนการทำงานดังนี้

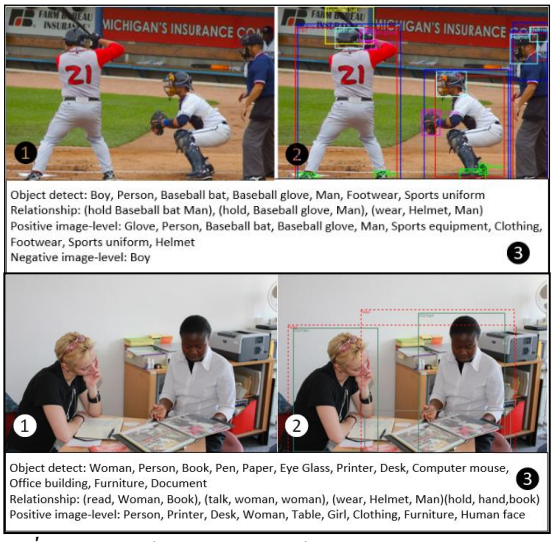


รูปที่ 1. ขั้นตอนการจำแนกความหมายด้วยการคัดเลือกคุณลักษณะด้วยวิธีริลีฟเอฟสำหรับหลายป้ายคำศัพท์

3.1 การเตรียมข้อมูลภาพ (Data Preprocessing)

การเตรียมข้อมูล เริ่มจากคัดเลือกข้อมูลภาพดิจิทัลที่มีวัตถุบนภาพที่เด่นชัด มีวัตถุภาพพื้นหลัง ภาพที่ถูกคัดเลือกเข้ามานั้นเป็นภาพที่มีความหมายภาพชัดเจน ภาพที่ถูกคัดเลือกออกไปไม่นำมาใช้ทดลองจะเป็นภาพที่มีลักษณะผิดปกติคุณลักษณะวัตถุไม่ชัดคลุมเครือ มีขนาดวัตถุขนาดเล็กเกินไปไม่สามารถบ่งชี้ชื่อวัตถุได้ หรือภาพถ่ายระยะใกล้ สำหรับข้อมูลภาพที่ใช้ในการทดลองถูกคัดเลือกจากฐานข้อมูลภาพ Open Images V4 [17][18] ข้อมูลภาพส่วนใหญ่จะนำมาจาก Flickr.comโดยข้อมูลภาพได้มีการสร้างป้ายคำศัพท์ตามกลุ่มที่กำหนดไว้ประกอบด้วยกลุ่มข้อมูลดังนี้ (1) object detection เป็นชื่อวัตถุที่ปรากฏบนภาพประกอบด้วยสูงสุด 8 คำศัพท์ (2) bounding boxes เป็นข้อมูลตำแหน่งของวัตถุที่เด่นบนภาพ ประกอบด้วยกลุ่มข้อมูลทั้งหมด 600 กลุ่ม ที่สอดคล้องกับวัตถุก่อนหน้า (3) visual relationship annotations เป็นการแสดงความสัมพันธ์ระหว่างกลุ่มวัตถุบนภาพ ประกอบด้วย 57 กลุ่ม รูปแบบคำศัพท์ที่ถูกเก็บลงบนฐานข้อมูล และได้มีการจัดเก็บข้อมูลที่เหมาะสมจากป้ายกำกับบนภาพที่มีความหมายใกล้เคียงกัน และกลุ่มคำศัพท์ที่มีความหมายต่างกัน กับกลุ่มคำศัพท์เพื่อช่วยในการวิเคราะห์ข้อมูล แสดงตัวอย่างข้อมูลภาพในรูปที่ 2 เป็นภาพตัวอย่างจากฐานข้อมูลหมายเลข❶เป็นภาพต้นฉบับหมายเลข❷เป็นการแสดงป้ายกำกับbounding boxes บนภาพวัตถุที่ถูกให้ความหมายและความสัมพันธ์ หมายเลข❸เป็นชื่อวัตถุบนภาพ

ที่ถูกป้ายคำศัพท์ที่มีการกำหนดกลุ่มคำเป็นหมวดหมู่จากฐานข้อมูลภาพ Open Images V4 [17][18]



รูปที่ 2. ตัวอย่างข้อมูลภาพแสดงป้ายกำกับbounding boxes (ภาพบน)และ ตัวอย่างภาพแสดงความสัมพันธ์ภายใน (ภาพล่าง)จากฐานข้อมูลภาพ Open Images V4 [17][18]

3.2 การคัดเลือกข้อมูลด้วยอัลกอริทึมริลีฟเอฟสำหรับหลายป้ายคำศัพท์(Feature Selection using ReliefF-ML Algorithm)

อัลกอริทึม ReliefF-ML [12][13][14] ถูกสร้างขึ้นมาจากกระบวนการให้ค่าน้ำหนักกับฟีเจอร์ (feature weight) โดยมีการกำหนดค่าของฮิต (hit) และมิส (miss) ของอัลกอริทึม ReliefF สามารถเขียนเป็นความสัมพันธ์ในเชิงความน่าจะเป็นได้ดังนี้

$$\rho_l = \frac{\eta_l + \gamma}{n + 2\gamma}, \quad (1)$$

เมื่อกำหนดให้ η_l จำนวนของอินสแตนในชุดข้อมูลที่ขึ้นกับป้ายคำศัพท์ l ให้ n เป็นจำนวนอินสแตนของชุดเรียนรู้ และให้ γ เป็นตัวแปรควบคุมค่าก่อนหน้า โดยให้ $\gamma = 1$เป็นค่าที่เหมาะสมสำหรับการใช้Laplace smoothing กำหนดให้ค่า i และj เป็นค่าระยะห่างระหว่างป้ายคำศัพท์ที่ i และ j ด้วยวิธี HammingDistance D_L แทนค่าความเหมือนกันในการจำแนกข้อมูล

$$D_L(i,j) = \frac{|y_i \Delta y_j|}{k\_nn}, \quad (2)$$

เมื่อกำหนดให้ค่าที่มีเกี่ยวข้อง (relevant) และค่าที่ไม่เกี่ยวข้อง (irrelevant) ของอินสแตน i ในกลุ่มของ k-nearest neighbors จะแบ่งเป็นกลุ่มฮิตH_i^lและกลุ่มมิสM_i^jของค่า i เมื่อกำหนดให้ $\rho_{H_i^l}$ แทนค่าความน่าจะเป็นของสองค่าที่มีความใกล้เคียงกันมากที่สุดที่ใช้ป้ายคำศัพท์ l ร่วมกันกับชุดป้ายคำศัพท์อื่น k_nnจำนวนเพื่อนบ้านใกล้เคียง และ $\rho_{M_i^l}$ แทนค่าความน่าจะเป็นของสองค่าที่มีความใกล้เคียงกันมากที่สุดที่ใช้ป้ายคำศัพท์ที่มีการใช้ร่วมกันกับชุดป้ายคำศัพท์อื่นเขียนเป็นสมการได้ดังนี้

$$\rho_{H_i^l} = \frac{\sum_{j \in H_i^l} D_L(i,j)}{k\_nn}, \quad (3)$$

$$\rho_{M_i^l} = \frac{\sum_{j \in M_i^l} D_L(i,j)}{k\_nn}, \quad (4)$$

อัลกอริทึม ReliefF-ML จะมีความขึ้นต่อกันระหว่างป้ายคำศัพท์ด้วยการคำนวณความน่าจะเป็น$\rho_{H_i^l}$ และ $\rho_{M_i^l}$ ค่าน้ำหนักจะมีผลกระทบต่อข้อมูลในกลุ่ม ดังนั้นกลุ่มข้อมูลจะถูกชี้วัดด้วยค่าน้ำหนักที่มีความต่างกันทำให้เกิดความคล้ายกันของชุดป้ายคำศัพท์ อัลกอริทึมจะทำงานด้วยการวนซ้ำๆและปรับค่าน้ำหนักไปเรื่อยๆด้วยสมการดังนี้

$$\omega_f = \omega_f - \sum_{l \in y_i} (\omega_1 \omega_2 \omega_H) + \sum_{l \in \bar{y}_i} (\omega_3 \rho_{M_i^l} \omega_M), \quad (5)$$

เมื่อกำหนดให้ $\omega_1 = \frac{\rho_l}{\sum_{k \in y_i} \rho_k}$, $\omega_2 = \frac{1 - \rho_{H_i^l}}{\rho_{H_i^l}}$, $\omega_3 = \frac{\rho_l}{\sum_{q \in \bar{y}_i} \rho_q}$, เป็นค่าตัวแปรน้ำหนักที่มีความเกี่ยวข้องกันและไม่เกี่ยวข้องกัน

$$\omega_H = \sum_{j \in H_i^l} \frac{\alpha(x_{if}, x_{jf})}{m}), \quad (6)$$

$$\omega_M = \sum_{l \in M_i^l} \frac{\alpha(x_{if}, x_{jf})}{m}, \quad (7)$$

เมื่อฟังก์ชัน $\alpha(x_{if}, x_{jf})$ เป็นการหาค่าจากความแตกต่างระหว่างค่าของข้อมูลในตำแหน่งที่ f ของข้อมูลiและ j กำหนดให้ mแทนจำนวนกลุ่มตัวอย่างที่ถูกนำไปเรียนรู้ เพื่อทำการประเมินค่าน้ำหนัก โดยที่ ReliefF-ML จะมีความต้องการค้นหาข้อมูลของแต่ละป้ายคำศัพท์ที่มีความสัมพันธ์กันและไม่

สัมพันธ์กันด้วย อัลกอริทึม k-nearest neighbors แสดงรายละเอียดของอัลกอริทึมได้ดังนี้

Algorithm 1. Pseudo-codeReliefF-ML

Require: for each training instance a vector of feature values and the class value
Input: T : data training set of multi label,
m : the number of sample instances,
k_nn : the number of nearest neighbors
Output: weight vector ω for the feature attributes

Begin
for $f \in F$ **do**
$\omega_f = 0;$
end for
for $l \in L$ **do**
Computing label probability ρ_l in equation (1)
end for
for $l = 1$ to m **do**
Random i from T;
for label $l \in \bar{y}_i$ **do**
Updating k-nearest neighbors H_i^l
Computing Probability $\rho_{M_i^l}$ in equation (3)
end for
for label $l \in \bar{y}_i$ **do**
Updating k-nearest neighbors M_i^l;
Computing Probability $\rho_{H_i^l}$ in equation (4)
end for
for $f \in F$ **do**
Updating weight ω_f in equation (5)
end for
end for
return ω;
End

4. การวัดประสิทธิภาพและวิธีการประเมินผล

การวัดประสิทธิภาพและวิธีการประเมินผล สำหรับการทดลองนี้ใช้ในการเปรียบเทียบด้วย วิธีการคัดเลือกคุณลักษณะมาตรฐาน [15][16] ทั้งหมด 4 วิธีดังนี้ (1) Information Gain (2) Chi-square (3) Gini Index และ (4) Relief ดังนี้

4.1 Information Gain (IG)

IGเป็นวิธีการคัดเลือกคุณลักษณะข้อมูลโดยการพิจารณาจากค่าความน่าจะเป็นของแต่ละข้อมูลที่มีความเป็นไปได้ด้วยค่า Entropy เพื่อคัดเลือกคุณลักษณะที่มีความสำคัญในการจำแนกกลุ่มได้ดีที่ดีที่สุดมีสมการดังนี้

IG เป็นวิธีการคัดเลือกคุณลักษณะข้อมูลโดยการพิจารณาจากค่าความน่าจะเป็นของแต่ละข้อมูลที่มีความเป็นไปได้ด้วยค่า Entropy

$$IG(\omega) = -\sum_{j=1}^{k} P(C_j) \log\left(P(C_j)\right) + P(\omega) \sum_{j=1}^{k} P(C_j|\omega) \log(P(C_j|\omega) + P(\overline{\omega}) \sum_{j=1}^{k} P(C_j|\overline{\omega}) \log(P(C_j|\overline{\omega})) \quad (8)$$

เมื่อกำหนดให้ $C = \{C_1, \dots, C_k\}$ แทนคลาสของฟีเจอร์ และมีทั้งหมด K คลาสสามารถเขียนได้ ความน่าจะเป็นของฟีเจอร์เป็น $P(C_j)$ และ $P(\omega)$แทนความน่าจะเป็นที่มีป้ายคำศัพท์ และ $P(C_j|\omega)$ แทนความน่าจะเป็นที่อยู่ในคลาส C_j ที่มีป้ายคำศัพท์ ω และ $\overline{\omega}$ เป็นข้อมูลที่ไม่มีป้ายคำศัพท์ที่ต้องการ

4.2 ไคสแควร์ (Chi-Square)

ไคสแควร์เป็นวิธีการคัดเลือกคุณลักษณะข้อมูลด้วย การใช้วิธีสถิติวัดค่าความต่างกันในรูปแบบการแจกแจงความถี่ของตัวแปรคุณลักษณะระหว่างความถี่ของค่าที่คาดหวังและความถี่ของตัวแปรที่สังเกตได้ดังแสดงในสมการ

$$\chi^2(l_1, l_2) = \frac{N(AD-BC)^2}{(A+C)(B+C)(A+B)(C+D)}, \quad (9)$$

เมื่อกำหนดให้ l_1และ l_2 เป็นป้ายคำศัพท์ของคลาส cและ N เป็นจำนวนของข้อมูลทั้งหมด Aเป็นจำนวนความถี่ที่เกิดป้ายคำศัพท์ของ l_1และ l_2 เกิดขึ้น, B เป็นจำนวนความถี่ที่เกิดป้ายคำศัพท์ของ l_1 แต่ไม่เกิดป้ายคำศัพท์ l_2, Cเป็นจำนวนความถี่ที่เกิดป้ายคำศัพท์ของ l_2แต่ไม่เกิดป้ายคำศัพท์ l_1และ Dเป็นจำนวนความถี่ที่ไม่เกิดป้ายคำศัพท์ของ l_1 และ l_2

4.3 Gini Index

Gini Index เป็นวิธีการที่คัดเลือกคุณลักษณะข้อมูลที่ขึ้นต่อคลาสโดยจะเลือกฟีเจอร์จากระดับความแตกต่างกันของกลุ่มสามารถเขียนสมการดังนี้

$$GI(l_i) = \sum_{j=1}^{k} p\left(l_i|C_j\right)^2 p(C_j|l_i)^2, \quad (10)$$

เมื่อกำหนดให้ $C = \{C_1, \dots, C_k\}$ แทนคลาสของฟีเจอร์มีทั้งหมด K คลาส

4.4 อัลกอริทึมรีลีฟ (Relief Algorithm)

อัลกอริทึมรีลีฟเป็นวิธีการคัดเลือกคุณลักษณะข้อมูลที่มีลักษณะเด่นจากชุดข้อมูลย่อยที่มีความใกล้เคียงกัน[12][13] [14]เริ่มจากทำการสุ่ม R_iจากกลุ่มและการค้นหากลุ่มใกล้เคียง Kจากกลุ่มเดียวกันที่มีลักษณะข้อมูลใกล้เคียงกันมากที่สุดเป็นค่าฮิต H และ K ที่มีข้อมูลคุณลักษณะใกล้เคียงจากคนละกลุ่มเรียกว่าค่ามิส Mและแต่ละข้อมูลจะมี ω_iสำหรับข้อมูลที่ iที่ประกอบด้วย R_i ,Hและ Mและถ้ากลับกันเมื่อข้อมูลภายใน R_i ใน H มีคุณสมบัติที่ต่างออกไปในข้อมูลที่ i แล้วนั้นค่าประมาณของ ω_iจะถูกลดลงทันทีโดยจะกระทำซ้ำๆกันทั้งหมด nครั้ง สามารถเขียนในรูปแบบของสมการได้ดังนี้

$$\omega_i = \omega_i - \frac{\sum_{k=1}^{K} D_H(k)}{n.k} + \sum_{c=1}^{C-1} p_c . \frac{\sum_{k=1}^{K} D_M(k)}{n.k} \quad (11)$$

เมื่อ $D_H(k)$ และ $D_M(k)$ เป็นผลรวมของระยะทางระหว่างค่าที่ตำแหน่ง k ของ H หรือ M และ p_c คือค่าความน่าจะเป็นของคลาส C

Algorithm 2. Pseudo-code Relief

Require: for each training instance a vector of feature values and the class value
Input:
D : data training set of feature data matrix,
n : the number of repeat times,
K : the number of the neighbors,
Output: weight vector ω for the feature attributes

```
Begin
    for j∈ n do
     Random select an instance R_j;
        Finding K nearest hits H
                and nearest misses M;
          for i=1 to all features do
    Updating estimation ω_i in equation (11);
          end for
    end for
End
```

5. ผลการทดลอง

การทำวิจัยครั้งนี้ได้นำเสนอ วิธีการคัดเลือกคุณ ลักษณะข้อมูลแบบ ReliefF-ML เพื่อการจำแนกข้อมูลภาพที่ดีขึ้น สำหรับชุดข้อมูลภาพที่นำมาใช้ในการทดลองในครั้งนี้ คือ Open Images V4 [17][18] ประกอบด้วย 3,500 ภาพ และ 2,800 ภาพเป็นชุดการเรียนรู้และที่เหลือ 700 ภาพ เป็นชุดภาพทดลอง ภาพทั้งหมดจะลดขนาดและจำนวนกลุ่มคำศัพท์ลงเฉพาะภาพที่มีบุคคลเป็นหลักและเป็นภาพทั่วไปที่ประกอบด้วยภาพภายนอก และภาพภายในบ้าน ข้อมูลประกอบด้วยเซตของกลุ่มข้อมูลดังนี้ (1) object detection เป็นป้ายกำกับประกอบด้วยคำศัพท์ตามกลุ่มวัตถุ 8 คำ (2) bounding boxesประกอบด้วยข้อมูลทั้งหมด 18 คุณลักษณะ (3) visualrelationship annotations ประกอบด้วยข้อมูล 10 คุณลักษณะ [17][18]

การวัดประสิทธิภาพของการจำแนกข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines: SVM) [19] และโครงข่ายประสาทเทียมแบบสังวัตนาการ (Convolutional Neural Network: CNN) [20][21] โดยจะทำการทดสอบประสิทธิภาพของวิธีการคัดเลือกข้อมูลที่เหมาะสมด้วยตาราง Confusion matrix ค่าความถูกต้อง (Accuracy) และทำการเปรียบเทียบวิธีคัดเลือกทั้ง 5 วิธี ดังนี้ IG, Chi-square, Gini Index, Relief และ ReliefF-MLซึ่งทำการจำแนกกลุ่มข้อมูลตามประเภทกิจกรรมประจำวัน ประกอบด้วย office, leisure, housework, sport (game)และ family (time) ด้วยค่าความแม่นยำ(Precision: Prec.) ค่าความระลึก (Recall) และเปรียบเทียบการคัดเลือกคุณลักษณะทั้ง 5 วิธี ผลการทดลองได้แสดงเป็นค่าเฉลี่ยความแม่นยำ (Average Precision: Avg.Prec.) ค่าเฉลี่ยค่าความระลึก (Average Recall: Avg.Recall) และค่าเฉลี่ยความถูกต้อง (Average Accuracy: Avg.Acc.) จากตาราง Confusion matrix ด้วยวิธี SVMและ CNN ดังแสดงผลการทดลองในตารางที่ 1 และตารางที่ 2 ตามลำดับ

จากตารางที่ 1 แสดงผลการทดลองค่าเฉลี่ยของความถูกต้องจากการคัดเลือกข้อมูล 12 คุณลักษณะอยู่ที่ 56.61%, 63.69%, 62.88%, 69.73% และ 72.19% จากวิธี Chi-square, IG, Gini Index, Relief และ ReliefF-ML ตามลำดับจะเห็นว่าการคัดเลือกข้อมูลด้วยวิธี ReliefF-ML ได้ผลของการจำแนกที่สูงที่สุดแต่เมื่อมาเปรียบเทียบกับ ค่าที่สูงขึ้นจากวิธี Relief เพียง 2.46% เท่านั้น แต่ค่าเฉลี่ยความถูกต้องระหว่าง Gini Index กับ Relief นั้นเพิ่มขึ้นสูงถึง 6.85% เมื่อใช้ 12 คุณลักษณะ จำแนกด้วย SVMจากตารางที่ 2แสดงผลการทดลองค่าเฉลี่ยของความถูกต้องจากการคัดเลือกข้อมูล 12 คุณลักษณะอยู่ที่ 68.35%, 68.19%, 69.64%, 69.83% และ 78.78% จากวิธี Chi-square, IG, Gini Index, Reliefและ ReliefF-ML ตามลำดับจะเห็นว่าการ

คัดเลือกข้อมูลด้วยวิธี ReliefF-ML ได้ผลของการจำแนกที่สูงที่สุดแต่เมื่อมาเปรียบเทียบกับ ค่าที่สูงขึ้นจากวิธี Relief มากถึง9% แต่ค่าเฉลี่ยความถูกต้องระหว่าง Gini Index กับ Relief นั้นเพิ่มขึ้นน้อยมากไม่ถึง 1% เมื่อมีการจำแนกด้วย CNN

ตารางที่ 1. เปรียบเทียบประสิทธิภาพของการคัดเลือกคุณลักษณะข้อมูลด้วยการจำแนกแบบSVM

Method	No. of features	Performance (%)		
		Avg.Prec.	Avg.Recall	Avg. Acc.
all features		58.01	57.53	57.53
Chi-Square	25	55.78	56.30	50.85
	18	57.42	57.14	57.14
	12	56.32	56.43	56.61
IG	25	55.78	56.30	55.71
	18	60.77	60.96	60.83
	12	63.88	63.69	63.69
Gini Index	25	59.51	58.52	59.35
	18	62.81	61.99	61.98
	12	63.64	62.91	62.88
Relief	25	61.36	61.50	61.18
	18	66.25	66.42	66.23
	12	69.84	69.88	69.73
ReliefF-ML	25	65.90	65.47	65.44
	18	71.39	71.43	71.31
	12	72.22	72.22	72.19

ตารางที่ 2. เปรียบเทียบประสิทธิภาพของการคัดเลือกคุณลักษณะข้อมูลด้วยการจำแนกแบบCNN

Method	No. of features	Performance (%)		
		Avg.Prec.	Avg.Recall	Avg. Acc.
all features		60.75	60.76	60.56
Chi-Square	25	59.28	59.31	50.85
	18	64.71	64.81	57.14
	12	68.39	68.35	56.61
IG	25	59.60	59.35	55.71
	18	64.97	64.58	60.83
	12	68.20	68.19	63.69
Gini Index	25	65.92	65.38	59.35
	18	66.25	65.46	61.98
	12	70.01	69.64	62.88
Relief	25	67.29	66.53	61.18
	18	67.46	77.26	66.23
	12	69.94	69.83	69.73
ReliefF-ML	25	66.50	66.46	65.44
	18	73.88	73.82	71.31
	12	78.93	78.87	72.19

จากตารางที่ 3 – 5 แสดงผลการทดลองการจำแนกข้อมูลภาพด้วยตาราง Confusion matrix ด้วย CNN โดยมีการใช้อัลกอริทึม ReliefF-ML คัดเลือกคุณลักษณะข้อมูล 25, 18 และ 12 คุณลักษณะ แสดงผลการจำแนกข้อมูลตามประเภทกิจกรรม 5 ประกอบด้วย office, leisure, housework, sport และ family ด้วยค่าความแม่นยำ (Prec.) ค่าความระลึก (Recall) และ ค่าความถูกต้อง (Acc.) จะเห็นว่าในตารางที่ 3 การจำแนกข้อมูลในกลุ่มของ office และ family ได้ค่าความแม่นยำ 66% และ 64.8% ตามลำดับ มีค่าความถูกต้องเฉลี่ยรวมเป็น 66.46% แต่เมื่อมีการคัดเลือกคุณลักษณะเป็น 18 จะได้ค่าความแม่นยำ family สูงถึง 75.5% และ sport 75% ได้ค่าความถูกต้องเฉลี่ยรวมเป็น 73.82% แสดงผลการทดลองในตารางที่ 4 และในตารางที่ 5 แสดงผลการทดลองการจำแนกด้วย 12 คุณลักษณะได้ค่าความแม่นยำในกลุ่ม sport สูงถึง 81.6% และได้ค่าเฉลี่ยความถูกต้องรวมเป็น 78.87%

ตารางที่ 3. ตารางจำแนกความหมายภาพด้วย ReliefF-ML ชุดข้อมูล 25 คุณลักษณะ

Method	office	leisure	housework	sport	family	Prec.	Recall
office	68	12	8	7	9	66.0	65.4
leisure	13	65	6	8	9	64.4	64.4
housework	8	7	65	7	11	69.1	66.3
sport	9	8	7	67	8	68.4	67.7
family	5	9	8	9	68	64.8	68.7
Avg. Acc.				**66.46**			

ตารางที่ 4. ตารางจำแนกความหมายภาพด้วยReliefF-ML ชุดข้อมูล 18 คุณลักษณะ

Method	office	leisure	housework	sport	family	Prec.	Recall
office	78	5	8	5	9	72.2	74.3
leisure	8	75	3	5	5	73.5	78.1
housework	6	8	73	9	4	73.0	73.0
sport	5	6	9	78	5	75.0	75.7
family	11	8	7	7	71	75.5	68.3
Avg. Acc.				**73.82**			

ตารางที่ 5. ตารางจำแนกความหมายภาพด้วยReliefF-ML ชุดข้อมูล 12 คุณลักษณะ

Method	office	leisure	housework	sport	family	Prec.	Recall
office	81	9	5	4	7	81.0	76.4
leisure	3	76	4	5	4	75.2	82.6
housework	8	4	72	6	7	80.0	74.2
sport	5	7	2	84	6	81.6	80.8
family	3	5	7	4	79	76.7	80.6
Avg. Acc.				**78.87**			

6. การวัดประสิทธิภาพและวิธีการประเมินผล

การทำวิจัยครั้งนี้ได้นำเสนอวิธีการคัดเลือกคุณลักษณะข้อมูลที่เหมาะสมสำหรับการจำแนกข้อมูลภาพในส่วนของการแปลความหมายภาพการลดจำนวนมิติของการจัดเก็บข้อมูล ด้วยวิธีการคัดเลือกคุณลักษณะเด่นและเหมาะสมเพื่อนำมาใช้งานได้อย่างมีประสิทธิภาพ อัลกอริทึมริลีฟเอฟสำหรับหลายป้ายคำศัพท์เป็นอัลกอริทึมสำหรับคัดเลือกคุณลักษณะข้อมูลที่มีลักษณะเด่นระหว่างฟีเจอร์ออกมาเพื่อใช้งาน การทำงานของอัลกอริทึมเป็นการคัดเลือกชุดข้อมูลย่อยเพื่อค้นหาฟีเจอร์ที่มีความใกล้เคียงกัน แต่ถูกจำกัดด้วยการประมาณค่าสำหรับฟีเจอร์เพียงแค่ป้ายคำศัพท์เดียวด้วยการหาค่าเฉลี่ยคะแนนของทุกๆ ข้อมูล ดังนั้นทำให้มีการปรับอัลกอริทึมรีลีฟใหม่เป็นReliefF-MLเพื่อแก้ไขปัญหาของวิธีริลีฟที่ไม่สมบูรณ์และยังสามารถจำแนกได้เป็นหลายกลุ่ม ด้วยการสุ่มเลือกข้อมูลตัวอย่าง และอัลกอริทึมจะคำนวณค่าที่มีความใกล้เคียงกันจากกลุ่มเดียวกันและกลุ่มที่ตรงกันข้ามกัน เพื่อให้เหมาะสมกับข้อมูลหลายตัวคุณภาพของแต่ละคุณลักษณะที่ได้จะขึ้นกับข้อมูลนำเข้าและความแตกต่างของกลุ่มที่กำหนดว่ามีความใกล้เคียงกันมากน้อย เพียงใดพบว่าการใช้วิธีการคัดเลือกสามารถจำแนกข้อมูลความหมายของภาพได้ถูกต้องกว่าวิธีอื่นๆถึงร้อยละ78.87

เอกสารอ้างอิง

[1] Galleguillos C. and Belongie S., “Context Based Object Categorization: A Critical Survey”, Computer Vision and Image Understanding, 114. pp. 712-722, 2010.

[2] N. Chinpanthana, “A Study of Feature Extraction Techniques used for Content Based Image Retrieval System”, Christian University Journal, Vol.23, No.1pp. 130-139, 2017.

[3] Russell, B.C., Torralba, A., Murphy, K.P. et al., “LabelMe: A Database and Web-Based Tool for Image Annotation”, Int J Comput Vision, 77. pp. 157–173, 2008.

[4] Junshi Huang, Rogerio S. Feris, Qiang Chen, Shuicheng Yan, “Cross-domain Image Retrieval with a Dual Attribute-aware Ranking Network”, Computer Vision and Pattern Recognition. pp. 1062-1070, 2015.

[5] I. Simon, N. Snavely, and S. Seitz., “Scene summarization for online image collections”, Computer Vision. pp. 1–8, 2007.

[6] Tang, J., S. Alelyani, and H. Liu, “Feature selection for classification: A review”, Data classification: Algorithms and applications. p. 37, 2014.

[7] Janecek, A., Gansterer, W., Demel, M. and Ecker, G., “On the relationship between feature selection and classification accuracy”, In New Challenges for Feature Selection in Data Mining and Knowledge Discovery, pp. 90-105, 2008.

[8] Parikh R. B., Obermeyer Z., and Bates D. W., “Making Predictive Analytics a Routine Part of patient Care.” 2016.

[9] Liu, H. and H. Motoda, “Feature selection for knowledge discovery and data mining,” Springer Science & Business Media, Vol. 454, 2012.

[10] El-Naggar, N., Y. El-Sonbaty, and M.A. El-Nasr, “Sentiment analysis of modern standard Arabic and Egyptian dialectal Arabic tweets,” IEEE Computing Conference, 2017.

[11] Omar, N., et al., “A comparative study of feature selection and machine learning algorithms for arabic sentiment classification,” Asia Information Retrieval Symposium. Springer, 2014.

[12] N. Spolaor, E. A. Cherman, M. C. Monard, “Using ReliefF for multi-label feature selection,” Proceedings of

the Conferencia Latinoamericana de Informatica, Brazil, pp. 960–975, 2011.

[13] N. Spolar, E. Cherman, M. Monard, H. Lee, "Filter approach feature selection methods to support multi-label learning based on ReliefF and Information Gain", Proceedings of the Advances in Artificial Intelligence - SBIA2012, LNCS, Springer Berlin Heidelberg, pp. 72–81, 2012.

[14] D. Kong, C. Ding, H. Huang, and H. Zhao, "Multi-label ReliefF and F-statistic Feature Selections for Image Annotation," Proceedings of Computer Vision and Pattern Recognition, pp. 2352–2359, 2012.

[15] Liu, H. and H. Motoda, "Feature selection for knowledge discovery and data mining," Springer Science & Business Media, Vol. 454, 2012.

[16] Tang, J., S. Alelyani, and H. Liu, "Feature selection for classification: A review," Data classification: Algorithms and applications, p. 37, 2014.

[17] A. Kuznetsova, et al., "The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale," IJCV, 2020.

[18] Papadopolous et al., "Extreme clicking for efficient object annotation," Computer Vision and Pattern Recognition ICCV, 2017.

[19] T. Joachims, "Text categorization with support vector machines learning with many relevant features," In Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany. Berlin: Springer. pp.137–42, 1998.

[20] J. Long, E. Shelhamer, and T. Darrell. "Fully convolutional networks for semantic segmentation," In CVPR, 2015.

[21] Ren S., He K., Girshick R. and Sun J., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE International Conference on Computer Vision, 2015.