

(Special Articles)

Comparison of Feature Selection Methods to Improve Breast Cancer Prediction

Ajjima Montaphan

Department of Applied Computer Science, School of Information Technology, Sripatum University

Correspondence to : ajjima.mo@spu.ac.th

The objective of this research was to study the comparison of the features selection techniques for improving breast cancer prediction. By using features selection from 7 techniques, including Correlation Based Feature Selection, Information Gain, Gain Ratio, Chi-Square, Forward Selection, Backward Elimination and Evolutionary Selection. The leading results from each technique were calculated the effectiveness of the forecast of breast cancer by using Support Vector Machine. The results showed that the percentage of accuracy in breast cancer prediction, based on the number of all 30 attributes, was 91.39 while the Evolutionary Selection gave the best results by reducing the important attributes to only 16 attributes and giving the accuracy of the forecasting results was 95.26 %.

Keywords : Feature Selection, Evolutionary Selection Technique, Support Vector Machine, Breast Cancer Predictions

Royal Thai Air Force Medical Gazette, Vol. 65 No. 2 May - August 2019

(บทความพิเศษ)

การเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุง การพยากรณ์มะเร็งเต้านม

อัจจิมา มณฑาพันธุ์

สาขาวิทยาการคอมพิวเตอร์ประยุกต์, คณะเทคโนโลยีสารสนเทศ, มหาวิทยาลัยศรีปทุม

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาเกี่ยวกับการเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญเพื่อนำมาใช้ในการปรับปรุงการพยากรณ์การเป็นมะเร็งเต้านม โดยใช้วิธีการคัดเลือกคุณลักษณะจากเทคนิคต่าง ๆ จำนวน 7 เทคนิค ได้แก่ เทคนิค Correlation Based Feature Selection เทคนิค Information Gain เทคนิค Gain Ratio เทคนิค Chi-Square เทคนิค Forward Selection เทคนิค Backward Elimination และเทคนิค Evolutionary Selection หลังจากคัดเลือกคุณลักษณะที่สำคัญจึงนำผลที่ได้จากแต่ละเทคนิคมาคำนวณหาค่าประสิทธิภาพในการพยากรณ์การเป็นมะเร็งเต้านมโดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีน ผลการทดลองพบว่าร้อยละของความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านม จากจำนวนคุณลักษณะของข้อมูลทั้งหมด 30 คุณลักษณะเท่ากับ 91.39 ขณะที่เทคนิค Evolutionary Selection ให้ผลดีที่สุดโดยสามารถลดคุณลักษณะที่สำคัญเหลือเพียง 16 คุณลักษณะ และให้ผลการวัดค่าความถูกต้องในการพยากรณ์ได้ถึงร้อยละ 95.26

คำสำคัญ : การคัดเลือกคุณลักษณะ เทคนิคการคัดเลือกแบบอีโวลูชันนารี ซัพพอร์ตเวกเตอร์แมชชีน การพยากรณ์มะเร็งเต้านม

1. บทนำ

โรคมะเร็งเต้านม จัดเป็นโรคร้ายอันดับหนึ่งซึ่งพบในผู้หญิงทั่วโลก จากสถิติของสถาบันมะเร็งแห่งชาติระบุว่า มีผู้หญิงไทยป่วยเป็นมะเร็งเต้านมรายใหม่ประมาณ 13,000 คน

ต่อปีหรือ 35 คนต่อวัน และยังมีแนวโน้มที่จะเพิ่มมากขึ้นทุกปี ปัจจุบันยังไม่ทราบสาเหตุที่แน่ชัด หากแต่มีปัจจัยเสี่ยงหลายประการที่สัมพันธ์กับการเกิดโรค เช่น ผู้หญิงที่มีอายุมากกว่า 50 ปี ปัจจัยทางพันธุกรรมโดยเฉพาะหาก

ครอบครัวมีญาติสายตรง เช่น มารดา พี่สาว น้องสาว เป็นมะเร็งเต้านมหรือมะเร็งรังไข่ มีคนในครอบครัวเป็นโรคมะเร็งเต้านมหลายคน หรือมีญาติเคยเป็นมะเร็งเต้านมเมื่ออายุน้อยหรือเป็นมะเร็งเต้านมพร้อมกันทั้ง 2 ข้าง ปัจจัยด้านฮอร์โมนเพศคือ เริ่มมีประจำเดือนเมื่ออายุน้อยกว่า 12 ปี หรือประจำเดือนหมดช้าหลังอายุ 55 ปี ผู้หญิงที่ไม่มีบุตรหรือมีบุตรคนแรกหลังอายุมากกว่า 30 ปี นอกจากนี้ผู้หญิงที่กินยาฮอร์โมนทดแทนหลังวัยทองเป็นระยะเวลานานเกิน 5 ปี ปัจจัยทั้งหมดนี้ก็มีโอกาสเพิ่มความเสี่ยงของการเกิดมะเร็งเต้านมได้เช่นกัน รวมทั้งการดื่มสุรา การฉายรังสีบริเวณทรวงอกและการกินยาคุมกำเนิด ซึ่งพบว่าเป็นปัจจัยที่มีความสัมพันธ์กับโรคอยู่บ้าง แต่ไม่ชัดเจนมากนัก⁽¹⁾

ในการศึกษาที่ผ่านมาพบว่าการคัดเลือกคุณลักษณะที่สำคัญในการวิเคราะห์ข้อมูล เพื่อลดจำนวนมิติที่ใช้ในการทำเหมืองข้อมูลเป็นสิ่งสำคัญ โดยมีการใช้เทคนิคหลาย ๆ แบบเพื่อให้เหมาะสมกับข้อมูลที่ต้องการวิเคราะห์ โดยมีงานวิจัยที่เกี่ยวข้องดังนี้ ภัทราวดี แสงศิริ⁽²⁾ ศึกษาการใช้เทคนิคในการคัดเลือกคุณลักษณะที่สำคัญ เพื่อคัดแยกประเภทของมะเร็งเม็ดเลือดขาวแบบเฉียบพลันซึ่งจำนวนมิติของข้อมูลมีจำนวน 7,129 มิติ สามารถลดมิติของข้อมูลลงเหลือ 36 มิติ และเพิ่มความแม่นยำจากร้อยละ 73.43 % มาเป็นร้อยละ 88.24 นิภาพร ชนะมาร และพรณี สิทธิเดช⁽³⁾ ศึกษาการวิเคราะห์ปัจจัยการเรียนรู้ด้วยการคัดเลือกคุณสมบัติและการพยากรณ์โดยใช้เทคนิคการคัดเลือกคุณสมบัติ 3 วิธี ได้แก่ การคัดเลือกคุณสมบัติแบบ Correlation-based Feature Selection การคัดเลือกคุณสมบัติแบบ Consistency-based Feature Selection และการคัดเลือกคุณสมบัติแบบ Gain Ratio Feature Selection ผลการทดลองทั้งสามเทคนิคสามารถลดจำนวนของคุณสมบัติจาก 22 ตัวแปร เหลือ 9 ตัวแปร 10 ตัวแปร และ 11 ตัวแปร ตามลำดับ พงศพิงศ์ เพ็งศิริ และคณะ⁽⁴⁾ ได้ศึกษาการลดมิติข้อมูลการวิเคราะห์ความสัมพันธ์และการประยุกต์สำหรับวิเคราะห์ข้อมูลพื้นฐานการใช้งานสมาร์ตโฟน โดยทำการศึกษาเทคนิคต่าง ๆ ในการลดมิติโดยทำการเปรียบเทียบเทคนิค Partition data of Association with Aprior CFS, Info Gain และ Relief โดยผ่านการจำแนกข้อมูลด้วย ANN น้ำทิพย์ มากนคร และมาลีรัตน์ โสदानิล⁽⁵⁾ ศึกษาการเปรียบเทียบวิธีการเลือกคุณลักษณะที่เหมาะสมเพื่อการจัดหมวดหมู่เว็บเพจผิดกฎหมายโดยใช้เทคนิคการทำเหมืองข้อมูลโดยใช้เทคนิค Information Gain, Gain Ratio และ

Chi Squared พบว่าวิธีการจัดอันดับแบบ Chi-Square ให้ผลลัพธ์ที่มีประสิทธิภาพดีที่สุดกล่าวคือ เมื่อลดจำนวนคุณลักษณะ 50 % ให้ค่าความถูกต้อง (Accuracy) ร้อยละ 95.98 นิธินันท์ มาตา และคณะ⁽⁶⁾ ได้ศึกษาการสร้างโมเดลสำหรับพยากรณ์การควบคุมประตูละอองน้ำ โดยมีการเปรียบเทียบเทคนิคการลดมิติของข้อมูลคือ Information Gain, Gain Ratio, Correlation based Feature Selection Galavotti L., et al.⁽⁷⁾ ได้ศึกษาเกี่ยวกับการคัดเลือกคุณสมบัติของข้อมูลจากข้อมูลจำนวนมากในการจัดหมวดหมู่ข้อมูลตัวอักษร Bing Xue., et al.⁽⁸⁾ ได้ศึกษาสรุปการคัดเลือกคุณสมบัติของข้อมูลในการลดมิติของข้อมูลและเพิ่มประสิทธิภาพของขั้นตอนวิธี Evolutionary Computation Approaches

จากการศึกษาข้อมูลที่เกี่ยวข้องกับโรคมะเร็งเต้านมพบว่าคุณสมบัติทางกายภาพรูปแบบต่าง ๆ ของผู้หญิงมีผลต่อการเป็นหรือไม่เป็นมะเร็งเต้านมและในทางการแพทย์ได้มีการจัดเก็บข้อมูลคุณสมบัติต่าง ๆ เหล่านี้จำนวนมาก ดังนั้นการวิจัยนี้จึงได้เสนอการคัดเลือกคุณสมบัติที่สำคัญโดยใช้เทคนิคในการลดคุณสมบัติที่เกี่ยวข้องน้อย เพื่อเพิ่มประสิทธิภาพในการพยากรณ์การเป็นมะเร็งเต้านม

2. วัตถุประสงค์การวิจัย

เพื่อศึกษาการคัดเลือกคุณลักษณะที่สำคัญที่ใช้ในการวิเคราะห์ข้อมูลเพื่อปรับปรุงการพยากรณ์การเป็นมะเร็งเต้านม

3. วรรณกรรมที่เกี่ยวข้อง

3.1 การทำเหมืองข้อมูล (Data Mining) หมายถึง กระบวนการที่กระทำกับข้อมูลโดยส่วนใหญ่จะเป็นข้อมูลจำนวนมาก เพื่อค้นหารูปแบบ แนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น โดยอาศัยหลักสถิติ การรู้จำ การเรียนรู้ของเครื่อง และหลักคณิตศาสตร์ โดยมีกระบวนการตามมาตรฐานอุตสาหกรรม (CRIPS-DM: Cross-Industry Standard Process for Data Mining) ประกอบด้วยขั้นตอนสำคัญ ๔ 6 ขั้นตอน ได้แก่

1. ทำความเข้าใจปัญหา (Business Understanding) ระบุเป้าหมายที่ต้องการจากการวิเคราะห์ และวางแผนดำเนินการ ซึ่งในขั้นตอนนี้ถือว่าเป็นการทำความเข้าใจธุรกิจ ทำความเข้าใจวัตถุประสงค์ของโครงการจากมุมมองทางธุรกิจ การแปลงความรู้เป็นข้อมูลเพื่อแก้ปัญหาด้วยการทำเหมืองข้อมูล

การพัฒนาแผนขั้นต้น การออกแบบเพื่อให้บรรลุวัตถุประสงค์ และเพื่อให้เข้าใจว่าข้อมูลใดควรได้รับการวิเคราะห์ ข้อมูลใดมีความสำคัญอย่างไร การประเมินสถานการณ์ การกำหนดเป้าหมายการทำเหมืองข้อมูลและการสร้างแผนโครงการ การกำหนดวัตถุประสงค์ทางธุรกิจ การทำความเข้าใจเป้าหมายที่แท้จริงของลูกค้าเป็นสิ่งสำคัญสำหรับการค้นพบปัจจัยสำคัญที่เกี่ยวข้องกับโครงการที่วางแผนไว้ และเพื่อให้มั่นใจว่านักวิเคราะห์ได้ค้นพบวัตถุประสงค์ทางธุรกิจหลัก รวมทั้งคำถามที่เกี่ยวข้องกับธุรกิจที่ต้องการ

2. ทำความเข้าใจในข้อมูล (Data Understanding)

รวบรวมข้อมูลที่เกี่ยวข้อง ทำความเข้าใจในข้อมูล สำนวน ความถูกต้องและจำนวนของข้อมูลให้รายละเอียดเพียงพอกับการวิเคราะห์ ขั้นตอนนี้นักวิเคราะห์จะทำความเข้าใจกับข้อมูลด้วยการรวบรวม และอธิบายข้อมูลเบื้องต้น จากนั้นจะทำการสำรวจข้อมูล และการตรวจสอบคุณภาพข้อมูล ระบุปัญหาคุณภาพข้อมูลเพื่อค้นหาข้อผิดพลาดเชิงลึก รวมทั้งการรวบรวมและการบูรณาการข้อมูลจากหลายแหล่งในกรณีที่เป็น

3. การเตรียมข้อมูล (Data Preparation) ขั้นตอนที่

นี้จะทำการคัดเลือกข้อมูล โดยจะเลือกเฉพาะข้อมูลที่เกี่ยวข้องกับสิ่งที่ต้องการวิเคราะห์ ทำการลบข้อมูลที่ซ้ำซ้อน แก้ไขข้อมูลที่ผิดพลาด แปลงรูปแบบข้อมูลให้อยู่ในรูปแบบที่พร้อมจะนำไปวิเคราะห์ การเตรียมข้อมูลจะครอบคลุมกิจกรรมทั้งหมดเพื่อสร้างชุดข้อมูลสุดท้าย หรือข้อมูลที่จะป้อนลงในแบบจำลองจากข้อมูลดิบเริ่มต้น ซึ่งจะรวมถึงการบันทึกข้อมูล และการเลือกคุณลักษณะ การแปลงและการทำความสะอาดข้อมูลสำหรับเครื่องมือการสร้างแบบจำลอง โดยมีห้าขั้นตอนในการเตรียมข้อมูล ได้แก่ การเลือกข้อมูล การทำความสะอาดข้อมูล การสร้างข้อมูล การรวมข้อมูล และการจัดรูปแบบข้อมูล เลือกข้อมูล เป็นการตัดสินใจเกี่ยวกับข้อมูลที่จะใช้ในการวิเคราะห์ ซึ่งขึ้นอยู่กับเกณฑ์หลายประการ รวมถึงความเกี่ยวข้องกับเป้าหมายการทำเหมืองข้อมูล ข้อจำกัดด้านคุณภาพของข้อมูลที่มีผลโดยตรงกับผลลัพธ์ที่ได้เทคนิค และปริมาณข้อมูลหรือชนิดข้อมูล ส่วนหนึ่งของกระบวนการคัดเลือกข้อมูลควรเกี่ยวข้องกับการอธิบายว่าเหตุใดข้อมูลบางอย่างถูกรวมหรือควรตัดออก หรือคุณลักษณะใดมีความสำคัญมากกว่าคุณลักษณะอื่น

4. การสร้างแบบจำลอง (Modeling) เป็นการสร้าง

แบบจำลองวิเคราะห์ข้อมูลด้วยเทคนิคเหมืองข้อมูล โดย

แบบจำลองจะขึ้นอยู่กับวัตถุประสงค์ของการวิเคราะห์ข้อมูล การเลือกข้อมูลและปรับพารามิเตอร์ของเทคนิคเหมืองข้อมูล เพื่อให้ได้ประสิทธิภาพที่ดี ดังนั้นในขั้นตอนนี้จึงเป็นการเลือกเทคนิค เพื่อการสร้างแบบจำลอง

5. การประเมินวัดประสิทธิภาพของแบบจำลอง (Evaluation) ขั้นตอนนี้จะทำการวัด ประสิทธิภาพของแบบจำลองที่สร้างขึ้นเพื่อความถูกต้อง เพื่อที่จะได้เลือกแบบจำลองที่เหมาะสมที่สุดไปใช้งาน

6. การนำแบบจำลองไปใช้งานจริง (Deployment) เป็นการนำแบบจำลองไปใช้งานจริงเพื่อให้บรรลุวัตถุประสงค์ของการทำเหมืองข้อมูล

3.2 การคัดเลือกคุณลักษณะ (Feature Selection) หมายถึงการลดขนาดหรือมิติของข้อมูลและยังคงคุณลักษณะสำคัญของข้อมูล อยู่ในขั้นตอนการเตรียมข้อมูล (Data Preparation) การคัดเลือกคุณลักษณะที่สำคัญ เป็นการคัดเลือกเซตของคุณลักษณะเฉพาะขึ้นใหม่จากเซตของคุณลักษณะเดิม โดยจะเป็นเซตย่อยของเซตคุณลักษณะเดิม เพื่อเพิ่มประสิทธิภาพในการทำนายและการทำงานให้เร็วขึ้น เนื่องจากบางคุณลักษณะเดิมที่ไม่มีความสำคัญหรือมีความสำคัญน้อยได้ถูกคัดเลือกทิ้งไป

3.2.1 Correlation Based Feature Selection (CFS) หมายถึงเทคนิคที่ใช้การหาคุณสมบัติคุณลักษณะที่ได้รับการประเมินค่าจากความสามารถในการคาดการณ์โดยคุณลักษณะที่ถูกคัดเลือกใช้สำหรับการจำแนกประเภทของข้อมูล

3.2.2 Information Gain (IG) หมายถึงเทคนิคการพิจารณาจากค่าความน่าจะเป็นของแต่ละคุณลักษณะที่เป็นไปได้แล้ววัดค่าความไร้ระเบียบ (Entropy) เพื่อคัดเลือกคุณลักษณะที่มีความสำคัญในการจำแนกกลุ่มได้ดีที่สุด

3.2.3 Gain Ratio (GR) หมายถึงเทคนิคการวัดจำนวนบิตของข้อมูล เพื่อใช้ในการทำนายค่าความน่าจะเป็นของประเภทของข้อมูล

3.2.4 Chi Square หมายถึงเทคนิคการวัดโดยใช้สถิติประมาณความสัมพันธ์ร่วมระหว่างคุณลักษณะเฉพาะกับคลาสของคุณลักษณะเฉพาะ

3.2.5 Forward Selection เป็นเทคนิคที่ใช้โดยการเพิ่มคุณลักษณะทีละ 1 คุณลักษณะ ถ้าคุณลักษณะที่ใส่เพิ่มให้ประสิทธิภาพที่ดีก็จะเก็บไว้และเลือกคุณลักษณะอื่น ๆ มาเพิ่มต่อไปจนประสิทธิภาพของโมเดลไม่ได้ดีขึ้นก็จะหยุดทำงาน

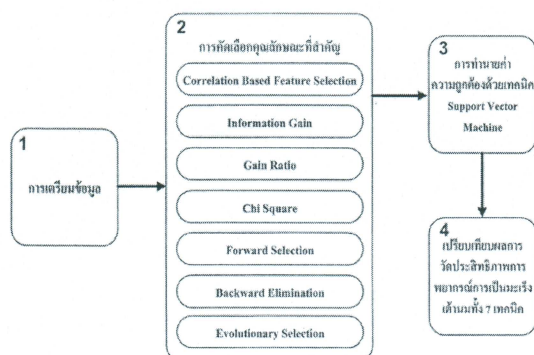
3.2.6 Backward Elimination เป็นเทคนิคที่ใช้โดยการลดคุณลักษณะ (Backward Elimination) เริ่มต้นด้วยคุณลักษณะทั้งหมดก่อนและตัดคุณลักษณะที่ไม่สำคัญทิ้งไปทีละคุณลักษณะ ถ้าประสิทธิภาพดีขึ้นก็ตัดคุณลักษณะอื่น ๆ ต่อไป จนกว่าจะพบว่าคุณลักษณะนั้นมีผลทำให้ประสิทธิภาพลดลงอย่างมีนัยสำคัญทางสถิติ

3.2.7 Evolutionary Selection เป็นการคัดเลือกคุณลักษณะ โดยวิธีการสุ่มเลือกคุณลักษณะซึ่งเป็นตัวแปรพยากรณ์เข้ามาในสมการที่เลวตัว และทำการทดสอบหาประสิทธิภาพในการพยากรณ์คำตอบ หากค่าประสิทธิภาพในการคาดการณ์สูงขึ้นจะเก็บคุณลักษณะนั้นไว้ แล้วทำการสุ่มเลือกคุณลักษณะอื่นเข้าไปเพิ่ม หากค่าประสิทธิภาพต่ำลงก็จะถอดคุณลักษณะนั้นออก

3.3 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)⁽¹⁰⁾ เป็นเทคนิคการจำแนกประเภทของข้อมูลที่มีพื้นฐานมาจากทฤษฎีการเรียนรู้จากสถิติ คล้ายโครงข่ายประสาทเทียม แต่ SVM ใช้หลักการลดค่าความเล็งเชิงโครงสร้างให้ต่ำสุด (Structural Risk Minimization) เพื่อลดค่าความผิดพลาดของการทำนาย (Minimized Error) พร้อมกับเพิ่มระยะการแบ่งแยกให้มากที่สุด (Maximized Margin) หลักการของ SVM คือการหาสมการของสมการเพื่อสร้างเส้นแบ่งแยกข้อมูล โดยการเลือกเส้นหรือระนาบแบ่งแยกประเภทข้อมูลที่เหมาะสมที่สุด

4. วิธีการวิจัยและผลการวิจัย

งานวิจัยนี้เป็นงานวิจัยเพื่อศึกษาการคัดเลือกคุณลักษณะที่มีความสำคัญในการปรับปรุงการพยากรณ์โรคมะเร็งเต้านม โดยมีการเปรียบเทียบเทคนิคต่าง ๆ ที่ใช้ในการคัดเลือกคุณลักษณะที่จะนำมาใช้ในการปรับปรุงการทำนายมะเร็งเต้านม โดยมีขั้นตอนการดำเนินงานการวิจัยประกอบด้วยขั้นตอนหลักดังต่อไปนี้



รูปที่ 1 แสดงขั้นตอนการดำเนินการวิจัย

4.1 การเตรียมข้อมูล

เป็นขั้นตอนของการจัดเตรียมข้อมูลเพื่อการวิเคราะห์ โดยนำข้อมูลที่ได้จากแหล่งข้อมูล UCI Machine Learning Repository ซึ่งเป็นข้อมูลจากคนที่เสียชีวิตจริงจำนวน 569 คน จำนวนคุณลักษณะของข้อมูลเท่ากับ 32 คุณลักษณะ โดย 2 คุณลักษณะแรก คือ : หมายเลขประจำตัว และผลการวินิจฉัย อีก 30 คุณลักษณะที่เหลือเป็นคุณลักษณะทางด้านกายภาพที่ถูกนำมาใช้ในการพยากรณ์

Data Set Characteristics:	Multivariate	Number of Instances:	569	Area:	Life
Attribute Characteristics:	Real	Number of Attributes:	32	Date Donated:	1995-11-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	607257

รูปที่ 2 แสดงรายละเอียดข้อมูลที่ถูกนำมาใช้ในการพยากรณ์จาก UCI Machine Learning Repository⁽⁹⁾

ตารางที่ 1 แสดงคุณลักษณะทางด้านกายภาพที่ถูกนำมาใช้ในการพยากรณ์

ลำดับ	คุณลักษณะ	ลำดับ	คุณลักษณะ
1	radius_mean	16	compactness_se
2	texture_mean	17	concavity_se
3	perimeter_mean	18	concave points_se
4	area_mean	19	symmetry_se
5	smoothness_mean	20	fractal_dimension_se
6	compactness_mean	21	radius_worst
7	concavity_mean	22	texture_worst
8	concave points_mean	23	perimeter_worst
9	symmetry_mean	24	area_worst
10	fractal_dimension_mean	25	smoothness_worst
11	radius_se	26	compactness_worst
12	texture_se	27	concavity_worst
13	perimeter_se	28	concave points_worst
14	area_se	29	symmetry_worst
15	smoothness_se	30	fractal_dimension_worst

4.2 การคัดเลือกคุณลักษณะที่สำคัญโดยการใช้ เทคนิคต่าง ๆ เป็นขั้นตอนการใช้เทคนิคต่าง ๆ จำนวน 7 เทคนิคในการคัดเลือกคุณลักษณะที่สำคัญเพื่อนำมาใช้ในการปรับปรุงการพยากรณ์การเป็นมะเร็งเต้านม ซึ่งเทคนิคที่ใช้คัดเลือกคุณลักษณะในการทดลองมีดังนี้ เทคนิค Correlation Based Feature Selection (CFS) เทคนิค Information Gain (IG) เทคนิค Gain Ratio (GR) เทคนิค Chi-Square เทคนิค Forward Selection เทคนิค Backward Elimination และเทคนิค Evolutionary Selection

4.3 การทำนายค่าความถูกต้องด้วยเทคนิค Support Vector Machine เป็นขั้นตอนการนำคุณลักษณะที่สำคัญที่ถูกคัดเลือกมาแล้วจาก 7 เทคนิค มาวัดประสิทธิภาพการพยากรณ์การเป็นมะเร็งเต้านมโดยใช้เทคนิค Support Vector Machine (SVM)

4.3.1 เทคนิค Correlation Based Feature Selection (CFS) การใช้เทคนิค Correlation Based Feature Selection จากคุณลักษณะทั้งหมด 30 คุณลักษณะ พบว่าคุณลักษณะที่มีค่าน้ำหนักเป็น 0 จะไม่ได้รับการคัดเลือก จะทำการเลือกเฉพาะคุณลักษณะที่มีค่าน้ำหนักเท่ากับ 1 เท่านั้น ซึ่งคุณลักษณะที่ถูกเลือกจากเทคนิค Correlation Based Feature Selection ได้แก่ radius_se radius_worst และ compactness_worst

หลังจากนั้นจึงได้นำคุณลักษณะทั้ง 3 มาหาค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมด้วยเทคนิค SVM โดยการใช้โปรแกรม RapidMiner ในการประมวลผลการพยากรณ์ซึ่งประกอบไปด้วยกระบวนการต่าง ๆ เริ่มต้นด้วยการใช้ Operator Retrieve เพื่อการคัดเลือกคุณลักษณะที่สำคัญจากเทคนิค Correlation Based Feature Selection ต่อจากนั้นนำคุณลักษณะที่สำคัญที่ได้รับการคัดเลือกแล้วเข้าสู่กระบวนการการทำ Validation เพื่อการพยากรณ์การเป็นมะเร็งเต้านมด้วยเทคนิค SVM โดยมีการกำหนดพารามิเตอร์ให้กับเทคนิค SVM ด้วยการกำหนดค่า svm type เป็น C-SVC ค่า kernel type เป็น rbf ค่า gamma เป็น 0.0 ค่า C เป็น 0.0 ค่า cache size เป็น 80 ค่า epsilon เป็น 0.001 เลือกการทำงานให้เป็นแบบ shrinking และ confidence for multiclass และมีการกำหนดค่าพารามิเตอร์ในการทำ Cross validation ให้ค่า number of folds เป็น 10 และลักษณะ sampling type เป็นแบบ stratified sampling ผลการวัดประสิทธิภาพการพยากรณ์โดยการคัดเลือกคุณลักษณะด้วยเทคนิค Correlation Based Feature Selection และการหาค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมด้วยเทคนิค SVM ทำให้ได้ค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมเท่ากับร้อยละ 91.22

4.3.2. เทคนิค Information Gain (IG) การใช้เทคนิค Information Gain ด้วยคุณลักษณะทั้งหมด 30 คุณลักษณะ จะให้ค่าน้ำหนักของคุณลักษณะที่สำคัญออกมาโดยคุณลักษณะ

ที่มีความสำคัญจะมีค่าน้ำหนักมาก คุณลักษณะที่มีความสำคัญน้อยจะมีค่าน้ำหนักน้อย จากผลการทดลองพบว่าคุณลักษณะ perimeter_worst จะมีความสำคัญมากที่สุดซึ่งมีค่าน้ำหนักที่ 0.5620 และคุณลักษณะ smoothness_se จะมีความสำคัญน้อยที่สุดซึ่งมีค่าน้ำหนักที่ 0.0135 ดังนั้นในการหาค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมโดยใช้เทคนิค SVM ในการทำนายความถูกต้อง จึงได้ทำการทดลองโดยการลดคุณลักษณะไปที่ละหนึ่งคุณลักษณะเริ่มจากคุณลักษณะที่มีค่าน้ำหนักน้อยที่สุด เพื่อที่จะได้ค่าความถูกต้องที่ดีที่สุดในการทำนายมะเร็งเต้านม จากผลการทดลองพบว่า 6 คุณลักษณะที่มีค่าความถูกต้องที่ดีที่สุดได้แก่ area_mean area_worst concave points_mean concave points_worst perimeter_worst และ radius_worst และเมื่อนำมาวัดผลการวัดประสิทธิภาพการพยากรณ์ด้วยเทคนิค SVM ได้ค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมเท่ากับร้อยละ 92.27

4.3.3. เทคนิค Gain Ratio (GR) ในการใช้เทคนิค Gain Ratio ด้วยคุณลักษณะทั้งหมด 30 คุณลักษณะ จะให้ค่าน้ำหนักของคุณลักษณะที่สำคัญออกมา โดยคุณลักษณะที่มีความสำคัญจะมีค่าน้ำหนักมาก คุณลักษณะที่มีความสำคัญน้อยจะมีค่าน้ำหนักน้อย จากผลการทดลองพบว่าคุณลักษณะ perimeter_worst จะมีความสำคัญมากที่สุดซึ่งมีค่าน้ำหนักที่ 0.6384 และคุณลักษณะ texture_se จะมีความสำคัญน้อยที่สุดซึ่งมีค่าน้ำหนักที่ 0.0749 ดังนั้นในการหาค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมโดยใช้เทคนิค SVM ในการทำนายความถูกต้อง จึงได้ทำการทดลองโดยการลดคุณลักษณะไปที่ละหนึ่งคุณลักษณะเริ่มจากคุณลักษณะที่มีค่าน้ำหนักน้อยที่สุด เพื่อที่จะได้ค่าความถูกต้องที่ดีที่สุดในการทำนายมะเร็งเต้านม จากผลการทดลองพบว่า 6 คุณลักษณะที่มีค่าความถูกต้องที่ดีที่สุดได้แก่ area_mean area_worst concave points_mean concave points_worst radius_worst และ perimeter_worst และเมื่อนำมาวัดผลการวัดประสิทธิภาพการพยากรณ์ด้วยเทคนิค SVM ได้ค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมเท่ากับร้อยละ 92.27

4.3.4. เทคนิค Chi Squared ในการใช้เทคนิค Chi Squared ด้วยคุณลักษณะทั้งหมด 30 คุณลักษณะ จะให้ค่าน้ำหนักของคุณลักษณะที่สำคัญออกมา โดยคุณลักษณะ

ที่มีความสำคัญมากจะมีค่าน้ำหนักมาก คุณลักษณะที่มีความสำคัญน้อยจะมีน้ำหนักน้อย จากผลการทดลองพบว่า คุณลักษณะ concave points_worst มีความสำคัญมากที่สุด มีค่าน้ำหนักที่ 415.0087 และคุณลักษณะ texture_se จะมีความสำคัญน้อยที่สุดมีค่าน้ำหนักที่ 13.3265 ดังนั้นในการหาค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมโดยใช้เทคนิค SVM ในการทำนายความถูกต้อง จึงได้ทำการทดลองโดยการลดคุณลักษณะไปที่ละหนึ่งคุณลักษณะเริ่มจากคุณลักษณะที่มีค่าน้ำหนักน้อยที่สุด เพื่อที่จะได้ค่าความถูกต้องที่ดีที่สุดในการทำนายมะเร็งเต้านม จากผลการทดลองพบว่ามี 9 คุณลักษณะที่มีค่าความถูกต้องที่ดีที่สุดได้แก่ area_mean area_worst concavity_mean concave points_mean concave points_worst perimeter_mean perimeter_worst radius_mean และ radius_worst และเมื่อนำมาวัดผลการวัดประสิทธิภาพการพยากรณ์ด้วยเทคนิค SVM ได้ค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมเท่ากับร้อยละ 92.27

4.3.5. เทคนิค Forward Selection ในการใช้เทคนิค Forward Selection ด้วยคุณลักษณะทั้งหมด 30 คุณลักษณะ พบว่าคุณลักษณะที่มีค่าน้ำหนักเป็น 0 จะไม่ได้รับการคัดเลือก และจะเลือกคุณลักษณะที่มีค่าน้ำหนักเท่ากับ 1 เท่านั้น ซึ่งคุณลักษณะที่ถูกเลือกจากเทคนิค Forward Selection ได้แก่ texture_worst, perimeter_worst และ concavity_worst หลังจากนั้นจึงได้นำคุณลักษณะทั้ง 3 มาหาค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมด้วยเทคนิค SVM ได้ค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมเท่ากับร้อยละ 95.07

4.3.6. เทคนิค Backward Elimination ในการใช้เทคนิค Backward Elimination ด้วยคุณลักษณะทั้งหมด 30 คุณลักษณะ จากผลการทดลองพบว่ามี 26 คุณลักษณะที่ถูกเลือกมาโดยเทคนิค Backward Elimination ซึ่งมีคุณลักษณะที่ไม่ถูกเลือกเพียง 4 คุณลักษณะคือ area_mean area_se concave points_se และ area_worst หลังจากนั้นจึงได้นำคุณลักษณะทั้ง 26 คุณลักษณะมาหาค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมด้วยเทคนิค SVM ได้ค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมเท่ากับร้อยละ 95.08

4.3.7. เทคนิค Evolutionary Selection ในการใช้เทคนิค Evolutionary Selection ด้วยคุณลักษณะทั้งหมด 30 คุณลักษณะ พบว่าคุณลักษณะที่มีค่าน้ำหนักเป็น 0 จะไม่ได้รับการคัดเลือก และจะเลือกคุณลักษณะที่มีค่าน้ำหนักเท่ากับ 1 เท่านั้น ซึ่งคุณลักษณะที่ถูกเลือกจากเทคนิค Evolutionary Selection ได้แก่ texture_mean perimeter_mean smoothness_mean concavity_mean concave points_mean symmetry_mean fractal_dimension_mean texture_se perimeter_se compactness_se fractal_dimension_se radius_worst texture_worst perimeter_worst compactness_worst และ symmetry_worst หลังจากนั้นจึงได้นำคุณลักษณะทั้ง 16 คุณลักษณะมาหาค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมด้วยเทคนิค SVM จากผลการวัดประสิทธิภาพการพยากรณ์โดยการคัดเลือกคุณลักษณะ 16 คุณลักษณะด้วยเทคนิค Evolutionary Selection และการหาค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมด้วยเทคนิค SVM จะได้ค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมเท่ากับร้อยละ 95.26

4.4 การเปรียบเทียบผลการวัดประสิทธิภาพการพยากรณ์การเป็นมะเร็งเต้านมทั้ง 7 เทคนิค เป็นขั้นตอนการนำผลที่ได้จากการวัดประสิทธิภาพการพยากรณ์การเป็นมะเร็งเต้านมจากคุณลักษณะที่สำคัญจากการคัดเลือกด้วยเทคนิคต่าง ๆ 7 เทคนิคมาเปรียบเทียบ เพื่อดูว่าเทคนิคใดให้ค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมได้เท่าไร

จากผลการทดลองเพื่อคัดเลือกคุณลักษณะที่เหมาะสมเพื่อปรับปรุงการพยากรณ์การเป็นมะเร็งเต้านมจากการใช้ข้อมูลผู้ทดลอง 569 คน จำนวน 30 คุณลักษณะด้านกายภาพ และใช้เทคนิคต่าง ๆ ในการคัดเลือกคุณลักษณะที่สำคัญจำนวน 7 เทคนิค ได้แก่ เทคนิค Correlation Based Feature Selection (CFS) เทคนิค Information Gain (IG) เทคนิค Gain Ratio (GR) เทคนิค Chi-Square เทคนิค Forward Selection เทคนิค Backward Elimination และเทคนิค Evolutionary Selection เพื่อนำมาเปรียบเทียบการวัดประสิทธิภาพในการพยากรณ์มะเร็งเต้านมด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน สรุปผลการทดลองดังแสดงในตารางที่ 2

ตารางที่ 2 แสดงเทคนิคที่ใช้ จำนวนคุณลักษณะที่ได้รับ การคัดเลือก และผลการวัดประสิทธิภาพการพยากรณ์

เทคนิคที่ใช้ในการคัดเลือกคุณลักษณะ	จำนวน คุณลักษณะที่ใช้ใน การพยากรณ์	ค่าร้อยละของการวัด ประสิทธิภาพ
ไม่มีการใช้เทคนิคใดๆ	30	91.39
Correlation Based Feature Selection	3	91.22
Information Gain	6	92.27
Gain Ratio	6	92.27
Chi-Square	9	92.27
Forward Selection	3	95.07
Backward Elimination	26	95.08
Evolutionary Selection	16	95.26

5. บทสรุป

การวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาการคัดเลือกคุณลักษณะที่สำคัญที่ใช้ในการวิเคราะห์ข้อมูลเพื่อปรับปรุงการพยากรณ์การเป็นมะเร็งเต้านม โดยมีสมมุติฐานการวิจัยว่าการคัดเลือกคุณลักษณะที่สำคัญจะทำให้จำนวนมิติของข้อมูลที่ใช้ในการวิเคราะห์ข้อมูลลดลง และอาจมีผลในการเพิ่มประสิทธิภาพการวัดค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านม จึงได้ทำการศึกษาเปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะที่สำคัญจำนวน 7 เทคนิค จากผลการทดลองในการนี้ใช้คุณลักษณะทั้งหมดจำนวน 30 คุณลักษณะ โดยไม่มีการลดจำนวนมิติข้อมูลจะพบค่าความถูกต้องการพยากรณ์มะเร็งเต้านมร้อยละ 91.39 ขณะที่ใช้เทคนิค Evolutionary Selection ซึ่งเป็นเทคนิคที่ได้ผลดีที่สุดสามารถลดคุณลักษณะเพื่อใช้ในการวิเคราะห์ข้อมูลลงไปได้ถึง 14 คุณลักษณะเหลือเพียง 16 คุณลักษณะและให้ค่าความถูกต้องการพยากรณ์การเป็นมะเร็งเต้านมเพิ่มขึ้นเป็นร้อยละ 95.26

6. ข้อเสนอแนะการวิจัย

เทคนิคการพยากรณ์ นอกเหนือจากเทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) ที่ใช้ในงานวิจัยนี้ ผู้วิจัยพบว่ามีเทคนิคอื่น ๆ ที่น่าสนใจ เช่น เทคนิคต้นไม้การตัดสินใจ (Decision Tree) เทคนิคแบบเครือข่ายประสาท (Neural Network) หรือเทคนิค

แบบเบย์ส (Bayes Classifier) ที่สามารถนำมาประเมินประสิทธิภาพการวัดค่าความถูกต้องของการพยากรณ์ในการนำมาเปรียบเทียบกับเทคนิคซัพพอร์ตเวกเตอร์แมชชีน เพื่อให้ได้ค่าประสิทธิภาพที่สูงขึ้น

เอกสารอ้างอิง

- ธีรวิทย์ กุหลาบประม. (9 เมษายน 2561). "BDMS เนตรจ้อง "เต้านม" ด้วยตัวเอง คัดกรองมะเร็งเบื้องต้นได้", MGR Online, สืบค้นเมื่อ 14 เมษายน 2562, <https://www.dmh.go.th/news-dmh/view.asp?id=27795>
- ภัทราวุธ แสงศิริ. 2553. "การคัดแยกประเภทของมะเร็งเม็ดเลือดขาวโดยใช้วิธีการจัดอันดับร่วมกับเทคนิคซัพพอร์ตเวกเตอร์แมชชีน.", วารสารวิจัย มช. (บศ.) 10(2):เม.ย.-มิ.ย.2553, หน้า 10-7.
- นิภาพร ชนะมาร และพรณี สิทธิเดช. 2557. "การวิเคราะห์ปัจจัยการเรียนรู้ด้วยการคัดเลือกคุณสมบัติและการพยากรณ์.", วารสารมหาวิทยาลัยราชภัฏสกลนคร 6(12):31-45.
- พญพิพงค์ เพ็งศิริ และคณะ. 2557. "การลดมิติข้อมูลการวิเคราะห์ความสัมพันธ์และการประยุกต์สำหรับวิเคราะห์ข้อมูลพื้นฐานการใช้งานสมาร์ตโฟน.", ประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 10. 8 - 9 พฤษภาคม 2557 ณ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ภูเก็ต, หน้า 528-34.
- นำทิพย์ มากนคร และมาลีรัตน์ โสตานิล. 2557. "การเปรียบเทียบวิธีการเลือกคุณลักษณะที่เหมาะสมเพื่อการจัดหมวดหมู่เว็บเพจผิดกฎหมายโดยใช้เทคนิคการทำเหมืองข้อมูล.", ประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 10. 8 - 9 พฤษภาคม 2557 ณ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ภูเก็ต, หน้า 168-73.
- นิธินันท์ มาตา และคณะ. 2558. "การค้นหาลำดับปัจจัยเพื่อสร้างโมเดลสำหรับพยากรณ์การควบคุมประตูละขานน้ำ.", การประชุมวิชาการระดับประเทศด้านเทคโนโลยีสารสนเทศ ครั้งที่ 7. ระหว่างวันที่ 29-30 ตุลาคม 2558 ณ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบังหน้า 352-7.

7. Galavotti, L., Sebastiani, F. and Simi, M. 2000. "Feature Selection and Negative Evidence in Automated Text Categorization.", Proceedings of KDD-00, 2000.
8. Bing Xue, Mengjie Zhang, Will N. Browne. 2016. "A Survey on Evolutionary Computation Approaches to Feature Selection." IEEE Transactions on Evolutionary Computation, Volume: 20 Issue: 4, Aug. 2016
9. William H. Wolberg, W. Nick Street, and Olvi L. Mangasarian. (1995-11-01). "Breast Cancer Wisconsin (Diagnostic) Data Set." UCI Machine Learning Repository. [online] Available: <http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28diagnostic%29>, [Accessed : 14/04/2017]
10. ชัชวาล วรวิทย์รัตนกุล และสุรศักดิ์ มั่งสิงห์. 2557. "การยืนยันลายเซ็นด้วยวิธีซอฟต์แวร์ตรวจจับเอกสาร (National Conference on Information Technology: NCIT) ครั้งที่ 6, 27-28 กุมภาพันธ์ 2557.
11. จิรา แก้วสุวรรณ. 2006. "การตรวจจับและการแก้ไขการวางตัวของภาพโดยใช้ซอฟต์แวร์ตรวจจับเอกสาร.", วิทยานิพนธ์ปริญญาโท สาขาเทคโนโลยีคอมพิวเตอร์ คณะครุศาสตร์อุตสาหกรรม สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ.
12. ทรงศักดิ์ ภูสีอ่อน. 2554. การประยุกต์ใช้ SPSS วิเคราะห์ข้อมูลงานวิจัย. มหาสารคาม : มหาวิทยาลัยมหาสารคาม. เอกสิทธิ์ พัทธวงค์ศักดิ์. 2557. การวิเคราะห์ข้อมูลด้วยเทคนิค ดาต้า ไมนิ่ง เบื้องต้น (An Introduction to Data Mining Techniques). กรุงเทพฯ, หน้า 53-7.
13. เอกสิทธิ์ พัทธวงค์ศักดิ์. 2559. Advanced Predictive Modeling with R & RapidMiner Studio 7. พิมพ์ครั้งที่ 6. กรุงเทพฯ: เอเชีย ดิจิตอลการพิมพ์.
14. Brian S. Everitt. 2010. Multivariable Modeling and Multivariate Analysis for The Behavioral Sciences. Taylor & Francis Group, LLC.
15. Colin Shearer. 2000. "The CRISP-DM Model : The New Blueprint for Data Mining." JOURNAL OF DATA WAREHOUSING, Volume 5 Number 4, Fall 2000 : p13
16. Jaiwei Han, Micheline Kamber, and Jian Pei. 2012. Data Mining Concepts and Techniques. New York: Elsevier Inc., page 104.
17. Majid Bahrepour. 2018. "The Forgotten Step in CRISP-DM and ASUM-DM Methodologies." [online] Available: <https://sharing.luminis.eu/blog/the-forgotten-step-in-crisp-dm-and-asum-dm-methodologies/>, [Accessed : 14/01/2019]
18. Mark A. Hall. 1999. "Correlation-based Feature Selection for Machine Learning." Doctor of Philosophy. Department of Computer Science, The University of Waikato, Hamilton, New Zealand.
19. P.-N. Tan, M. Steinbach, and V. Kumar. 2006. Introduction to data mining. vol. 1: Pearson Addison Wesley Boston, 2006.