การประยุกต์ใช้ซัพพอร์ทเว็กเตอร์แมชชีนในการทำนายการอยู่รอดของผู้ป่วยมะเร็งเต้านม

Support Vector Machine Applied to Survival Prediction in Breast Cancer Patients

สุรเดช บุญลือ[1] ชฎาพร สุขแจ่ม[2] และ ศศิธร สนิทผล[3]
[1, 2, 3]สาขาวิชาวิทยาการคอมพิวเตอร์ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยนอร์ทกรุงเทพ โทรศัพท์ 0-2972-7200
E-mail: [1]tsuradej@gmail.com, [2]chadaphon_num@hotmail.com, [3]sasithon.s@hotmail.com

บทคัดย่อ

บทความวิจัยนี้นำเสนอตัวแบบซัพพอร์ทเว็กเตอร์แมชชีนเพื่อใช้การทำนายการอยู่รอดของผู้ป่วยโรคมะเร็งเต้านม การวิจัยเป็นแบบเชิงทดลองโดยอาศัยข้อมูลอาการช่วยพยากรณ์ จากผู้ป่วยโรคมะเร็งเต้านมจำนวน 198 ระเบียน 34 คุณลักษณะ โดยข้อมูลดังกล่าวจะถูกนำเข้าตัวแบบเรียนรู้ของเครื่องเพื่อการทำนายผลการอยู่รอดด้วยวิธีการฝึกสอนและทดสอบประสิทธิภาพการทำนาย ซึ่งการทดสอบตั้งอยู่บนพื้นฐานวิธีการตรวจสอบแบบไขว้ 3 ส่วน ผลการทดลองที่ได้พบว่าตัวแบบทำนายการอยู่รอดที่ใช้ซัพพอร์ทเว็กเตอร์แมชชีนมีความแม่นยำ 80.30% ขณะที่ตัวแบบทำนายที่ใช้โครงข่ายประสาทเทียมแพร่กลับ และเทคนิคต้นไม้ตัดสินใจมีความแม่นยำ 76.76% และ 72.72% ตามลำดับ

คำสำคัญ: การทำนายการอยู่รอด, มะเร็งเต้านม, ซัพพอร์ทเว็กเตอร์-แมชชีน

Abstract

This research paper presents a Support Vector Machine model for survival prediction in breast cancer patients. The research is an experimental research, which has contributed by cancer prognostic patients with the amount of 138 records 34 features. By doing the research, the dataset of theses patients would transition into machine learning model to survival prediction by training method and testing the efficiency of prediction based on the 3-fold cross validation. The results of experiment show that survival prediction model by Support Vector Machine Technique reaches 80.30%, accuracy, while Back-Propagation Neural Networks and Decision Tree technique approaches 76.76% and 72.72% of measuring the accuracy of survival prediction, respectively

Keywords: Survival Prediction, Brest Cancer, Support Vector Machine

1. คำนำ

ปัจจุบันมะเร็งเป็นโรคที่พบมากและเป็นสาเหตุของการเสียชีวิตที่สำคัญของประชากรในหลายประเทศทั่วโลก มีการคาดการณ์ว่า ในปี พ.ศ. 2558 จะมีผู้ป่วยเป็นโรคมะเร็งเพิ่มขึ้นจาก 9 ล้านคนเป็น 15 ล้านคน สาเหตุมาจากจำนวนประชากร จำนวนผู้สูงอายุและจำนวนผู้สูบบุหรี่ที่เพิ่มขึ้นรวมถึงปัจจัยอื่น ๆ เช่น วิถีการดำเนินชีวิตของคนในปัจจุบัน และสภาวะสิ่งแวดล้อมที่เปลี่ยนแปลงไป ซึ่งเป็นปัจจัยที่ไม่สามารถป้องกันและควบคุมได้อย่างทั่วถึง สำหรับประเทศไทยโรคมะเร็งเป็นสาเหตุการตายอันดับที่สาม รองจากโรคหัวใจและอุบัติเหตุ และมีแนวโน้มว่าจะมีอัตราการตายที่สูงขึ้นทุกปี [1] จากรายงานของสถาบันมะเร็งแห่งชาติประจำปี พ.ศ.2552 [2] พบว่าชนิดของโรคมะเร็งที่พบมากที่สุดในประเทศไทย คือ มะเร็งเต้านม รองลงมาคือ มะเร็งปอด มะเร็งตับ มะเร็งปากมดลูก มะเร็งลำไส้ใหญ่ และมะเร็งช่องปาก ตามลำดับ

สาเหตุของการเกิดมะเร็งที่ชัดเจน ในปัจจุบันยังไม่ทราบแน่ชัด แต่อย่างไรก็ตามมีการศึกษา ที่เชื่อได้ว่ามีปัจจัยที่เกี่ยวข้องกับการเกิดโรคอยู่หลายประการ [3], [4] เช่น สาเหตุจากสิ่งแวดล้อมภายนอกร่างกาย เช่น สารเคมีบางชนิด หรือเชื้อจุลินทรีย์ สาเหตุจากภายในร่างกาย เช่น กรรมพันธุ์ที่ผิดปกติ หรือความไม่สมดุลทางฮอร์โมน เป็นต้น การรักษาโรคมะเร็งในปัจจุบันมีหลายวิธี เช่น การผ่าตัดเอาก้อนเนื้อมะเร็งออก รังสีบำบัด เคมีบำบัด การใช้ฮอร์โมน หรือ การฝังแร่ หรืออาจใช้หลายวิธีร่วมกัน ทั้งนี้ขึ้นอยู่กับระยะของโรค พยาธิวิทยาของชิ้นเนื้อ และสภาวะความสมบูรณ์ของผู้ป่วย

ในทางการแพทย์ อัตราการอยู่รอด (Survival Rate) เป็นตัวชี้วัดที่สำคัญ [4] ที่ทำให้ทราบถึงร้อยละของผู้ป่วย ซึ่งอยู่รอดภายหลังจากได้รับการวินิจฉัยโรคและรับการรักษาจนกระทั่งถึงช่วงเวลาหนึ่งที่สนใจพิจารณา เช่น การวัดอัตราการอยู่รอดของผู้ป่วยมะเร็งปากมดลูกที่ 5 ปี (5-year survival rate) ที่เข้ารับการรักษาด้วยการฉายรังสีอย่างเดียวหรือฉายรังสีร่วมกับเคมีบำบัด ทำให้สามารถพิจารณาเปรียบเทียบอัตราการอยู่รอดที่ 5 ปี จำแนกตามวิธีการรักษาได้

เดิมทีการวิเคราะห์อัตราการอยู่รอดโดยทั่วไปได้มาจากวิธีตามรุ่น (Cohort Analysis) ซึ่งสามารถใช้ทำการประมาณอัตราการอยู่รอดย้อนหลังไปในหลาย ๆ ปีได้ [5] แต่ข้อจำกัดที่สำคัญของวิธีการนี้ คือ ไม่สามารถสะท้อนให้เห็นถึงอัตราการอยู่รอดที่นับรวมถึงผู้ป่วยรายหลัง ๆ ที่เพิ่งตรวจพบโรคได้ใหม่ ทำให้ค่าประมาณอัตราการอยู่รอดที่ได้ไม่ใช่ค่าประมาณที่ใกล้เคียงกับค่าประมาณที่เป็นปัจจุบันมากที่สุด จึงได้มีผู้ทำการเสนอวิธีการใหม่ในการวิเคราะห์อัตราการอยู่รอดขึ้นเพื่อที่จะแก้ไขข้อจำกัดของวิธีการเดิมซึ่งเรียกวิธีการนี้ว่า วิธีตามช่วงเวลา (Period Analysis) [6] [7] ซึ่งวิธีการนี้จะทำให้สามารถประมาณอัตราการอยู่รอดได้ใกล้เคียงกับสถานการณ์จริงมากกว่าวิธีเดิม เนื่องจากใช้ข้อมูลที่เป็นปัจจุบันมากยิ่งขึ้น และในการคำนวณไม่ต้องใช้ช่วงเวลาที่ย้อนหลังนานเกินไปเช่นเดียวกับวิธีตามรุ่น [5]

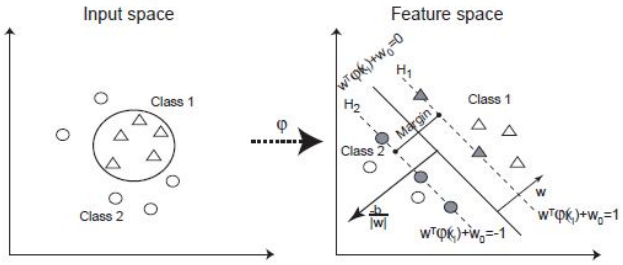
อย่างไรก็ตามปัจจุบันนี้มีการนำศาสตร์ด้านปัญญาประดิษฐ์ โดยเฉพาะเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) มาประยุกต์ใช้กับการทำนายการอยู่รอดของผู้ป่วยมะเร็งเต้านมตามช่วงเวลา อาทิงานวิจัยของ Biganzoli et al. [8] ที่นำโครงข่ายประสาทเทียมแพร่กลับมาวิเคราะห์อัตราการอยู่รอด ส่วนงานวิจัยของ Delen et al. [9] ที่ศึกษาการนำเทคนิคเหมืองข้อมูลมาประยุกต์กับการทำนายการอยู่รอดของผู้ป่วยมะเร็งเต้านม พบว่าเทคนิคต้นไม้ตัดสินใจมีประสิทธิภาพสูงกว่าโครงข่ายประสาทเทียมและการวิเคราะห์การถดถอยโลจิสติกส์ ส่วนเทคนิคซัพพอร์ทเว็กเตอร์แมชชีนมีการนำไปประยุกต์ใช้กับการทำนายโรคมะเร็ง [10] [11] เช่นกัน ซึ่งพบว่ามีประสิทธิภาพสูงและเหมาะสมกับการจำแนกข้อมูลที่มีค่าใกล้เคียงกัน

ดังนั้นงานวิจัยนี้จะนำเทคนิคซัพพอร์ทเว็กเตอร์แมชชีนมาประยุกต์ใช้กับการพัฒนาโปรแกรมทำนายการอยู่รอดของผู้ป่วยมะเร็งเต้านม รวมทั้งมีการเปรียบเทียบและคัดเลือกอัลกอริทึมการเรียนรู้ของเครื่องที่นำมาประยุกต์ใช้สร้างตัวแบบทำนายที่มีประสิทธิภาพสูง เพื่อให้ได้โปรแกรมที่มีประสิทธิผลและมีความเหมาะสมกับการนำไปใช้งานต่อไป

2. ทฤษฎีที่เกี่ยวข้อง

2.1 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines)

วิธีการของซัพพอร์ตเวกเตอร์แมชชีน [12] [13] จัดเป็นเทคนิคที่ใช้ในการแก้ปัญหาทางด้านการรู้จำรูปแบบข้อมูล โดยอาศัยหลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้ โดยเน้นไปยังเส้นแบ่งแยกแยะกลุ่มข้อมูลได้ดีที่สุด (Optimal Separating Hyperplane) ดังรูปที่ 1



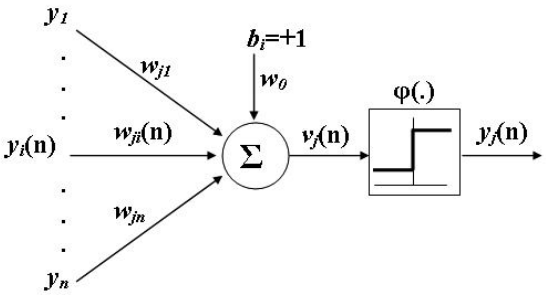
รูปที่ 1 ซัพพอร์ตเวกเตอร์แมชชีน [13]

สำหรับรากฐานเดิมของซัพพอร์ตเวกเตอร์แมชชีน ถูกนำมาใช้กับข้อมูลที่เป็นเชิงเส้น แต่ในความเป็นจริงแล้วข้อมูลที่นำมาใช้ในระบบการสอนให้ระบบเรียนรู้ส่วนใหญ่มักเป็นข้อมูลแบบไม่เป็นเชิงเส้น ซึ่งสามารถแก้ปัญหาดังกล่าวด้วยเคอร์เนล ที่เป็นการเปลี่ยนแปลงมิติของข้อมูลให้สูงขึ้น เพื่อช่วยในการเรียงตัวของข้อมูลเสียใหม่ที่เรียกว่า “พื้นที่มิติสูง” (Higher Dimensional Space) เคอร์เนลที่นิยมใช้มีอยู่ 3 ชนิด [13] ด้วยกันคือ โพลิโนเมียล (Polynomial), เรเดียลเบสิสฟังก์ชัน (Radial Basis Function-RBF) และ ซิกมอยด์ (Sigmoid) โดยงานวิจัยนี้จะใช้เคอร์เนล โพลิโนเมียล เป็นหลัก

2.2 โครงข่ายประสาทเทียม (Neural Network)

การเรียนรู้ของโครงข่ายแบบ Perceptron [14] จะมีลักษณะของโครงสร้างเป็นชั้น ซึ่งข้อมูลที่จะนำเข้ามาเรียนรู้นั้นจะถูกใส่เข้าไปในชั้นแรกสุดซึ่งเป็นชั้นรับข้อมูล (Input Layer) และเมื่อผ่านการคำนวณจากชั้นแรกนี้แล้วผลลัพธ์ ก็จะถูกส่งต่อไปยังชั้นกลางของโครงข่าย (Hidden Layer) ซึ่งในแต่ละหน่วยของชั้นนี้ ก็จะรับข้อมูลจากทุกหน่วยในชั้นก่อนหน้ามาคำนวณ แล้วส่งต่อไปยังชั้นถัดไปและเมื่อข้อมูลถูกส่งต่อกันมาจนถึงชั้นสุดท้าย (Output Layer) ก็จะได้ผลลัพธ์ออกมาจากระบบ ซึ่งการส่งผ่านข้อมูลต่อๆ กันไปแบบนี้จะเรียกว่าเป็นการส่งต่อข้อมูลแบบการป้อนไปข้างหน้า (Feed Forward) หลังจากนั้นจะต้องมีการตรวจสอบผลลัพธ์ที่ได้จากระบบว่ามีความคลาดเคลื่อนจากเป้าหมายมากน้อยเพียงไร

หากความคลาดเคลื่อนจากเป้าหมายมากเกินไปก็จะต้องมีการนำค่าความคลาดเคลื่อนนี้ ไปปรับน้ำหนักการเรียนรู้ใหม่ (Weight) แบบแพร่ย้อนกลับ (Back-Propagation) ซึ่งจะเป็นการปรับน้ำหนักความคลาดเคลื่อนจากชั้นผลลัพธ์ ไปยังชั้นก่อนหน้า และทำการปรับน้ำหนักไปเรื่อยๆ จนกระทั่งถึงชั้นรับข้อมูล ซึ่งกระบวนการเรียนรู้แบบนี้ต้องอาศัยการทำซ้ำหลาย รอบจนกว่าจะได้ผลลัพธ์ตามที่กำหนดหรือได้ค่าความคลาดเคลื่อนที่น้อยจนพอยอมรับได้ ซึ่งจำนวนรอบนี้ก็จะขึ้นอยู่กับ ความยากง่ายของปัญหา ขนาดของข้อมูลรวมไปถึงจำนวนชั้นของโครงสร้างของโครงข่ายที่เราสร้างไว้ด้วย ซึ่งโครงข่าย perceptron แบบชั้นเดียวได้แสดงดังรูปที่ 2



รูปที่ 2 โครงข่าย Perceptron แบบชั้นเดียว [14]

พิจารณาภาพที่ 1 จะเป็นกระบวนการคำนวณการส่งข้อมูลเข้าและกระบวนการแพร่ย้อนกลับของโครงข่าย Perceptron แบบชั้นเดียว โดยจะเป็นการนำข้อมูลเข้ามาคำนวณจนกระทั่งได้ผลลัพธ์ จะเห็นว่าในรูปประกอบด้วยส่วนของ v_j ซึ่งเป็นผลรวมของข้อมูลที่เข้ามาทำการปรับค่าน้ำหนัก ดังสมการที่ (1)

$$v_j = \sum_{j-1}^{n} w_{ij} y_i + w_0 \qquad (1)$$

โดยที่ค่า w_{ij} เป็นค่าน้ำหนักที่ใช้สำหรับปรับค่าข้อมูลและค่า $y_j(n)$) ก็จะเป็นการนำค่าที่ได้มาผ่านฟังก์ชั่นกระตุ้น (activation function) ดังสมการที่ (2)

$$y_j = \varphi(v_j) \qquad (2)$$

โดยที่ $\varphi(v_j)$ คือฟังก์ชั่นกระตุ้นสำหรับการปรับค่า output ของระบบ และจะทำการปรับค่าน้ำหนักค่าใหม่ที่ได้ จากค่าอัตราการเรียนรู้ของค่าอัตราความเปลี่ยนแปลงของค่าความคลาดเคลื่อนเทียบกับอัตราการเปลี่ยนแปลงของค่าน้ำหนัก ตามสมการที่ (3)

$$\Delta w_{ji}(n) = -\eta \frac{\partial E(n)}{\partial w_{ji}(n)} \qquad (3)$$

โดยที่ $\Delta w_{ji}(n)$ คือ ค่าน้ำหนักที่จะนำไปปรับใหม่ให้กับระบบ ส่วน E คือค่าความผิดพลาดรวมของระบบและ η คือ ค่าอัตราการเรียนรู้

2.3 เทคนิคต้นไม้ตัดสินใจ (Decision Tree Technique)

การสร้างต้นไม้ตัดสินใจจะเป็นแบบการค้นหาจากบนลงล่างแบบตะกราม (Top-down Greedy Search) โดยเริ่มจากการเลือกคุณสมบัติที่ดีที่สุดมาสร้างเป็นโนดราก เมื่อข้อมูลผ่านการแบ่งแยกที่โนดรากตามค่าคุณสมบัติของโนดรากแล้ว ก็จะหาคุณสมบัติที่ดีที่สุดของข้อมูลที่ผ่านการแบ่งแยกนั้นมาสร้างเป็นโนดลูกของโนดรากนั้นต่อไป และจะวนสร้างโนดลูกและต้นไม้ย่อยของแต่ละกิ่งไปเรื่อยๆ จนกว่าข้อมูลที่ผ่านการแบ่งแยกนั้นจะจัดอยู่ในกลุ่มเดียวกัน หรือจำนวนข้อมูลที่ผ่านการแบ่งแยกในกิ่งหนึ่งๆ มีค่าน้อยกว่าค่าที่กำหนดไว้ [15]

แต่ละอัลกอริทึมก็ได้นิยามค่าความดีของคุณสมบัติแตกต่างกันไป อัลกอริทึม CART นิยามความดีของคุณสมบัติโดยใช้ค่าสัมประสิทธิ์จีนี (Gini) แต่ที่แพร่หลายที่สุดคือการใช้ค่ามาตรฐานเกน (Gain criterion) ของอัลกอริทึม ID3 [16] และ C4.5 [17] ซึ่งได้จากการคำนวณโดยอาศัยทฤษฎีสารสนเทศ (Information Theory) และค่าเอนโทรปี (Entropy)

โดยที่ ID3 จะใช้ค่ามาตรฐานเกนเป็นหลักในการเลือกคุณสมบัติที่จะใช้เป็นรากหรือโนด แต่ใน C4.5 ได้เพิ่มการใช้ค่ามาตรฐานอัตราส่วนเกน (Gain Ratio Criterion) ในการตัดสินใจเลือกคุณสมบัติที่จะใช้เป็นรากหรือโนดอีกอย่างหนึ่ง เนื่องจากค่ามาตรฐานเกนจะมีอคติ (Bias) อย่างมากกับข้อมูลที่ประกอบด้วยคุณสมบัติที่มีค่าที่เป็นไปได้จำนวนมากๆ การแก้ไขความอคติของค่ามาตรฐานเกนสามารถทำได้โดยการปรับค่ามาตรฐานเกนให้ถูกต้อง โดยใช้ค่าสารสนเทศของการแบ่งแยก (Split information) ของคุณสมบัติแต่ละตัว ถ้าให้ T คือชุดของตัวอย่าง เมื่อแบ่งตัวอย่างนี้ตามคุณสมบัติ X จะได้ชุดของตัวอย่างย่อยในแต่ละกิ่ง คือ $\{t_1, t_2, \ldots, t_n\}$ จำนวน n ชุด ตามค่าที่เป็นไปได้ในคุณสมบัติ X เมื่อคำนวณค่าสารสนเทศของการแบ่งแยกได้ดังสมการที่ (4) นี้

$$\text{ค่าสารสนเทศของการแบ่งแยก} = \sum_{i=1}^{n} \frac{|t_i|}{|T|} \log_2 I_x \frac{|t_i|}{T} \qquad (4)$$

สารสนเทศของการแบ่งแยกนี้จะแสดงถึงระดับการกระจายของข้อมูลเมื่อแบ่งข้อมูลตัวอย่าง T เป็น n ชุดย่อยตามคุณสมบัติ X โดยค่านี้จะสูงสุดเมื่อ $|t_i|$ เป็น 1 เท่ากันในทุกกิ่ง และลดลงเมื่อค่า $|t_i|$ เพิ่มขึ้น เมื่อนำค่านี้ไปหารค่ามาตรฐานเกนจะได้ค่ามาตรฐานอัตราส่วนเกน ซึ่งช่วยแก้ไขความอคติของค่ามาตรฐานเกนได้ โดยทำให้ค่ามาตรฐานอัตราส่วนเกนในการแบ่งด้วยคุณสมบัติที่มีการกระจายสูงถูกปรับลดลง ดังนั้นค่ามาตรฐานอัตราส่วนเกนในคุณสมบัติของตัวอย่างที่มีการกระจายตัวของข้อมูลสูงดังที่กล่าวมาแล้วจึงไม่มีค่าสูงที่สุดเสมอ จากแนวคิดข้างต้น ตัวแบบทำนายการอยู่รอดมะเร็งเต้านมจะใช้อัลกอริทึม C4.5 ด้วยว่ามีประสิทธิภาพสูงกว่า ID3

3. วิธีการดำเนินการ

3.1 ข้อมูลกลุ่มตัวอย่าง

ชุดข้อมูลของผู้ป่วยมะเร็งเต้านมจากมหาวิทยาลัย Wisconsin ประเทศสหรัฐอเมริกา [18] ซึ่งจะแบ่งออกเป็น 2 กลุ่ม คือ กลุ่มที่ไม่เกิดอาการอีกในเวลา 24 เดือน (non-recurrent) จำนวน 151 คน และกลุ่มที่กลุ่มที่เกิดอาการซ้ำในเวลา 24 เดือน (recurrent) จำนวน 47 คน รวมทั้งหมด 198 ระเบียน 34 คุณลักษณะ

3.2 สร้างตัวแบบทำนาย

ออกแบบและสร้างตัวแบบทำนายโดยคัดเลือกตัวแบบทำนายที่มีประสิทธิภาพความแม่นยำสูงสุดระหว่างตัวแบบทำนายที่ใช้เทคนิคซัพพอร์ทเว็กเตอร์แมชชีน โครงข่ายประสาทเทียมแพร่กลับ และเทคนิคต้นไม้ตัดสินใจ เพื่อนำไปใช้ในส่วนการทำนายของโปรแกรมที่พัฒนาขึ้น ซึ่งข้อมูลที่นำเข้าที่จะฝึกสอนและทดสอบจะมี 33 คุณลักษณะ ส่วนข้อมูลนำออกจะเป็นกลุ่มที่ไม่เกิดอาการอีกและกลุ่มผู้ป่วยที่เกิดอาการซ้ำ รวม 2 คลาส

3.3 วิธีวัดประสิทธิภาพของตัวแบบทำนาย

การประเมินผลใช้วิธีวัดค่าความถูกต้อง ค่าความแม่นตรง (Precision) ค่าความระลึก (Recall) และการวัดประสิทธิภาพโดยรวม (F-measure) ซึ่งมีค่าอยู่ระหว่าง 0 - 1 ซึ่ง 1 หมายถึงประสิทธิภาพดี [19] ดังสมการที่ (5) (6) และ (7) ตามลำดับ

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

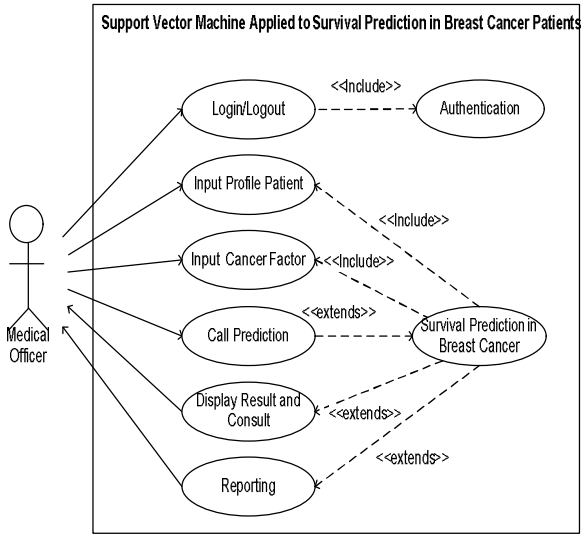
$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \qquad (6)$$

$$\text{F-measure} = \frac{2*(\text{Recall}*\text{Precision})}{\text{Recall}+\text{Precision}} \qquad (7)$$

3.4 ขั้นตอนการพัฒนาแอพพลิเคชั่น

3.4.1 การออกแบบโปรแกรม

ใช้การวิเคราะห์และออกแบบระบบเชิงวัตถุ และใช้โปรแกรม Microsoft Visio 2003 เป็นเครื่องมือช่วยพัฒนา ซึ่งภาพรวมของระบบสามารถแสดงได้ดังรูปที่ 3 ต่อไปนี้



รูปที่ 3 ภาพรวมระบบงาน

3.4.2 การพัฒนาโปรแกรม

โปรแกรมถูกพัฒนาขึ้นด้วย Microsoft C#.NET 2008 ในลักษณะ Windows Base Application และใช้ MS SQL Server 2008 express จัดเก็บฐานข้อมูล โดยจะมีส่วนคอมโพเน็นต์ของซัพพอร์ท-เว็กเตอร์แมชชีนที่สร้างขึ้นด้วยโปรแกรม Peltarion ไว้สำหรับทำนายการอยู่รอดของผู้ป่วยมะเร็งเต้านม

3.4.3 การทดสอบโปรแกรม

ใช้เทคนิค Black Box Testing ในการทดสอบโปรแกรม ซึ่งผู้ประเมินเป็นกลุ่มเชี่ยวชาญด้านการพัฒนาซอฟต์แวร์ จำนวน 5 คน มาทำแบบทดสอบประเมินประสิทธิภาพและความพึงพอใจของผู้ใช้ โดยแบ่งเกณฑ์ระดับ ออกเป็น 5 ระดับ ซึ่งต้องมีคะแนนเฉลี่ยตั้งแต่ 4 ขึ้นไปจึงจะยอมรับว่าระบบมีประสิทธิภาพในการใช้งานได้ในสภาพการทำงานจริง

4. ผลของการดำเนินงาน

4.1 ส่วนตัวแบบการทำนาย

จากการนำข้อมูลอาการผู้ป่วยมาพยากรณ์ จากผู้ป่วยโรคมะเร็งเต้านมจำนวน 198 ระเบียน 34 คุณลักษณะ ไปทดสอบกับตัวแบบทำนายด้วยวิธีทดสอบ 3 Fold Cross-Validation พบว่าผลการทดสอบประสิทธิภาพของตัวแบบทำนายโดยใช้เทคนิคซัพพอร์ทเว็ก-เตอร์แมชชีน โครงข่ายประสาทเทียมแพร่กลับ และเทคนิคต้นไม้ตัดสินใจ ให้ค่าดังตารางที่ 1 ข้างล่างต่อไปนี้

ตารางที่ 1 ผลการทดสอบประสิทธิภาพของตัวแบบทำนาย

Algorithm	TP Rate	FP Rate	Precision	Recall	F-Measure
SVM	0.803	0.457	0.787	0.803	0.789
BP-ANN	0.768	0.453	0.758	0.768	0.762
J48	0.724	0.744	0.622	0.727	0.657

ส่วนการหาความแม่นยำของการทำนายจากตัวแบบพบว่าตัวแบบที่ใช้เทคนิคซัพพอร์ทเว็กเตอร์แมชชีน มีความแม่นยำ 80.30% ขณะที่ตัวแบบทำนายที่มีใช้โครงข่ายประสาทเทียมแพร่กลับ และเทคนิคต้นไม้ตัดสินใจ มีความแม่นยำ 76.76% และ 72.72% ตามลำดับ

ด้วยว่าเทคนิคซัพพอร์ทเว็กเตอร์แมชชีน มีพฤติกรรมที่จะแยกแยะข้อมูล โดยใช้สมการระนาบหลายมิติที่จะพยายามหาจุดข้อมูลที่ทำให้ได้สมการระนาบหลายมิติที่ใช้แบ่งแยกดีที่สุด (Optimal Hyperplane) ความถูกต้องที่สุด โดยพิจารณาจากระยะห่าง (Margin) ระหว่างคลาส ซึ่งเส้นระนาบที่ดีที่สุดนี้จะสามารถจำแนกกลุ่มผู้ป่วยกลุ่มที่ไม่เกิดอาการอีกในเวลา 24 เดือน กับเกลุ่มที่เกิดอาการซ้ำ

ดังนั้นตัวแบบทำนายที่นำมาประยุกต์ใช้กับการพัฒนาโปรแกรมที่ใช้ทำนายการอยู่รอดของผู้ป่วยมะเร็งเต้านมนี้จะเป็นตัวแบบที่ใช้เทคนิคซัพพอร์ทเว็กเตอร์แมชชีน

4.2 ส่วนประเมินหาประสิทธิภาพของโปรแกรม

กลุ่มผู้เชี่ยวชาญเต้านการพัฒนาซอฟต์แวร์ จำนวน 5 คนได้ประเมินโปรแกรม ซึ่งสามารถสรุปได้ดังตารางที่ 2 ข้างล่างนี้

ตารางที่ 2: การประเมินประสิทธิภาพโดยผู้เชี่ยวชาญ

รายการประเมิน	$\overline{X}$	SD	ระดับ
1. ผลการประเมินด้านความสามารถในการทำงาน	4.17	0.27	ดี
2. ผลการประเมินด้านความต้องการของผู้ใช้	4.14	0.20	ดี
3. ผลการประเมินด้านการใช้งานของโปรแกรม	4.12	0.23	ดี
4. ผลการประเมินด้านผลลัพธ์ที่ได้จากโปรแกรม	4.15	0.45	ดี
5. ผลการประเมินด้านความปลอดภัย	4.28	0.39	ดี
สรุปประเมินประสิทธิภาพโดยผู้เชี่ยวชาญ	4.17	0.11	ดี

6. บทสรุป

งานวิจัยนี้ได้นำความรู้ด้านปัญญาประดิษฐ์มาประยุกต์ใช้กับการทำนายการอยู่รอดของผู้ป่วยมะเร็งเต้านม โดยจะมีการนำอัลกอริทึมการเรียนรู้ของเครื่องที่ประกอบด้วยเทคนิคซัพพอร์ทเว็กเตอร์แมชชีน โครงข่ายประสาทเทียมและเทคนิคต้นไม้ตัดสินใจมาสร้างตัวแบบทำนายการอยู่รอดของผู้ป่วยมะเร็งเต้านม แล้วเปรียบเทียบหาประสิทธิภาพความแม่นยำในการทำนาย เพื่อหาตัวแบบทำนายที่ให้ความแม่นยำสูงสุดที่จะเหมาะสมกับการนำไปพัฒนาโปรแกรมต่อไป

ผลการทดสอบพบว่าตัวแบบทำนายที่ใช้เทคนิคซัพพอร์ท - เว็กเตอร์แมชชีน มีความแม่นยำ 80.30% ขณะที่ตัวแบบทำนายที่มีใช้โครงข่ายประสาทเทียบแพร่กลับ และเทคนิคต้นไม้ตัดสินใจ มีความแม่นยำ 76.76% และ 72.72% ตามลำดับ นอกจากนั้นเมื่อนำตัวแบบทำนายไปใช้ในการพัฒนาโปรแกรมและผ่านการประเมินประสิทธิภาพโดยผู้เชียวชาญด้วยแบบสอบถาม จะได้คะแนนเฉลี่ยอยู่ในระดับ 4.17 (SD = 0.11) ซึ่งสามารถนำไปใช้งานได้ต่อไป

7. กิตติกรรมประกาศ

ขอขอบพระคุณมหาวิทยาลัยนอร์ทกรุงเทพที่ให้ทุนสนับสนุนงานวิจัยนี้

เอกสารอ้างอิง

[1] สายพิณ โชติวิเชียร. “มะเร็ง”. 2549. [ระบบออนไลน์]. แหล่งที่มา http://nutrition.anamai.moph.go.th/cancer1.

[2] สถาบันมะเร็งแห่งชาติ. “รายงานทะเบียนมะเร็งระดับโรงพยาบาล”. กลุ่มงานเทคโนโลยีสารสนเทศ สถาบันมะเร็งแห่งชาติ. 2553

[3] Cancer Research UK. “What causes cancer”. 2006. [ระบบออนไลน์]. แหล่งที่มา http://www.cancerhelp.org.uk.

[4] Cancer survival rate. “A tool to understand your prognosis”. 2007. [ระบบออนไลน์]. แหล่งที่มา http://www.MayoClinic.com.

[5] Brenner H, Gefeller O, Hakulinen T. “Period analysis for ‘up-to-date’ cancer survival data: theory, empirical evaluation, computational realization and applications”. *Eur J Cancer* 2004; 40: pp 326-335

[6] Brenner H and Gefeller O. “An Alternative Approach to Monitoring Cancer Patient Survival”. *Cancer* 1996; 78: pp 2004-2010

[7] Brenner H and Gefeller O. “Deriving More Up-to-Date Estimates of Long-Term Patient Survival”. J Clin Epidemiol 1997; 50: pp211-216

[8] Biganzoli, E., Boracchi, P., Mariani, L., and Marubini, E. “Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. Statistics in Medicine”, 1998; 17, pp 1169–1186.

[9] Delen D, Walker G, Kadam A. “Predicting breast cancer survivability: a comparison of three data mining methods”. Artificial Intelligence in Medicine. 2005 Jun; 34(2): pp113-27.

[10] Mu, T., Nandi, A.K., “Detection of breast cancer using v-SVM and RBF networks with self organized selection of centers”. Medical Applications of Signal Processing, 2005, pp 47– 52.

[11] Mihir Sewak Priyanka Vaidya Chien-Chung Chan Zhong-Hui Duan. “SVM Approach to Breast Cancer Classification”. Proceeding IMSCCS '07 Proceedings of the Second International Multi-Symposiums on Computer and Computational Sciences. 2007, pp 32-37

[12] C.J.C. Burges. “A Tutorial on Support Vector Machines for Pattern Recognition”. Data Mining and Knowledge Discovery, 2(2), 1998. pp 121-167

[13] C.W. Hsu, C.C. Chang, C.J. Lin, et al. “A practical guide to support vector classification”, 2003.

[14] Haykin S. “Neural Network a comprehensive foundation”. 2nd edition, USA: Prentice hall, 1999.

[15] บุญเสริม กิจศิริกุล. “อัลกอริทึมการทำเหมืองข้อมูล”. รายงานการวิจัย. วิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์ มหาวิทยาลัย, 2545.

[16] Quinlan, J. R. “Induction of decision tree”. Machine Learning, Vol. 1, 1986, pp 81-106,.

[17] Quinlan, J. R, “C4.5: Programs for Machine Learning”. San Mateo, CA: Morgan Kaufmann. 1993

[18] W. N. Street, O. L. Mangasarian, and W.H. Wolberg. “An inductive learning approach to prognostic prediction”. Proceedings of the Twelfth International Conference on Machine Learning, pages522--530, San Francisco, 1995. Morgan Kaufmann.

[19] Max Brame. “Principles of Data Mining”. Springer-Verlag London Limited 2007, pp 173-176