รายงานการวิจัย

เรื่อง

การเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์มะเร็งเต้านม

COMPARISON OF FEATURE SELECTION METHODS TO IMPROVE BREAST CANCER PREDICTION

อัจจิมา มณฑาพันธุ์

งานวิจัยนี้ ได้รับทุนอุดหนุนการวิจัยจากมหาวิทยาลัยศรีปทุม

ปีการศึกษา 2560

กิตติกรรมประกาศ

รายงานการวิจัยฉบับนี้สำเร็จและสมบูรณ์เป็นรูปเล่ม ด้วยความกรุณาและเอาใจใส่เป็นอย่างดีจาก ผู้ช่วยศาสตราจารย์ ดร.ชุมพล บุญคุ้มพรภัทร ผู้ทรงคุณวุฒิที่ได้กรุณาให้คำปรึกษาและแนะแนวทางในการดำเนินการทำรายงานในครั้งนี้โดยไม่มีข้อบกพร่อง รวมทั้งข้อเสนอแนะและข้อคิดเห็นต่างๆ ตลอดทั้งการตรวจแก้ไขรายงานฉบับนี้ให้สำเร็จสมบูรณ์ยิ่งขึ้น ผู้วิจัยจึงขอขอบพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้

ขอขอบพระคุณคุณครูอาจารย์ทุกท่านที่ได้ประสิทธิ์ประสาทวิชาความรู้ และประสบการณ์ตลอดจนอำนวยความสำเร็จให้บังเกิดขึ้น

สุดท้ายนี้ขอขอบพระคุณคณาจารย์และเจ้าหน้าที่คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยศรีปทุม ที่เป็นกำลังใจและให้ความช่วยเหลือในการดำเนินงาน และให้คำแนะนำในการทำรายงานการวิจัยครั้งนี้ให้สำเร็จลุล่วงด้วยดีตลอดมา

อัจจิมา มณฑาพันธุ์

ผู้วิจัย

เมษายน 2562

คำนำ

รายงานการวิจัยฉบับนี้จัดทำขึ้นเพื่อเป็นประโยชน์ต่อผู้ที่สนใจศึกษาเกี่ยวกับเรื่องการเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์มะเร็งเต้านม โดยงานวิจัยฉบับนี้ได้นำเสนอเทคนิคต่างๆจำนวน 7 เทคนิคคือ เทคนิค Correlation Based Feature Selection เทคนิค Information Gain เทคนิค Gain Ratio เทคนิค Chi-Square เทคนิค Forward Selection เทคนิค Backward Elimination และเทคนิค Evolutionary Selection ที่นำมาใช้เพื่อคัดเลือกคุณลักษณะที่สำคัญ หลังจากนั้นนำผลการคัดเลือกคุณลักษณะในแต่ละเทคนิคมาคำนวณเปรียบเทียบค่าประสิทธิภาพของการพยากรณ์การเป็นมะเร็งเต้านมด้วยเทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) เพื่อดูว่าเทคนิคไหนสามารถคัดเลือกคุณลักษณะที่สำคัญซึ่งให้ค่าประสิทธิภาพการพยากรณ์การเป็นมะเร็งเต้านมได้ดีที่สุด

ผู้วิจัยหวังเป็นอย่างยิ่งว่ารายงานการวิจัยฉบับนี้จะเป็นประโยชน์ต่อผู้ที่สนใจสามารถนำมาใช้ประโยชน์ในการอ้างอิงได้ต่อไป หากมีข้อบกพร่องประการใด ผู้จัดทำขออภัยไว้ ณ โอกาสนี้ด้วย

อัจจิมา มณฑาพันธุ์

ผู้วิจัย

เมษายน 2562

หัวข้อวิจัย : การเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์มะเร็งเต้านม
ผู้วิจัย : นางอัจจิมา มณฑาพันธุ์

หน่วยงาน : มหาวิทยาลัยศรีปทุม วิทยาเขตบางเขน
ปีที่พิมพ์ : พ.ศ. 2562

บทคัดย่อ

งานวิจัยนี้ได้ศึกษาเกี่ยวกับการเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์มะเร็งเต้านม มีวัตถุประสงค์เพื่อที่จะหาเทคนิคการคัดเลือกคุณลักษณะที่สำคัญเพื่อนำมาใช้พยากรณ์มะเร็งเต้านม โดยใช้วิธีการคัดเลือกคุณลักษณะจากเทคนิคต่างๆจำนวน 7 เทคนิค ได้แก่ เทคนิค Correlation Based Feature Selection เทคนิค Information Gain (IG) เทคนิค Gain Ratio (GR) เทคนิค Chi-Square เทคนิค Forward Selection เทคนิค Backward Elimination และเทคนิค Evolutionary Selection นำผลที่ได้จากแต่ละเทคนิคมาคำนวณหาค่าประสิทธิภาพในการทำนายการเป็นมะเร็งเต้านม ผลการทดลองพบว่าเทคนิค Evolutionary Selection ได้ผลดีที่สุดจากจำนวน 7 เทคนิค จากจำนวนคุณลักษณะของข้อมูลทั้งหมด 30 คุณลักษณะ เทคนิค Evolutionary Selection สามารถลดคุณลักษณะที่สำคัญเหลือเพียง 16 คุณลักษณะ ซึ่งให้ผลการวัดค่าความถูกต้องในการพยากรณ์ได้ดีถึง 95.26%

คำสำคัญ : การคัดเลือกคุณลักษณะ การพยากรณ์มะเร็งเต้านม ซัพพอร์ตเวกเตอร์แมชชีน เทคนิคการคัดเลือกแบบอีโวลูชั่นนารี

Research Title : Comparison of Feature Selection Methods to Improve Breast Cancer Prediction.
Name of Researcher : Mrs. Ajjima Montaphan
Name of Institution : Sripatum University, Bangkhen Campus
Year of Publication : B.E. 2562

ABSTRACT

This study investigated the comparison of feature selection methods to improve breast cancer prediction. The purpose is to find the key feature selection techniques to improve the predictability of breast cancer using 7 techniques such as Correlation Based Feature Selection (CFS) Technique, Information Gain (IG) Technique, Gain Ratio (GR) Technique, Chi-Square Technique, Forward Selection Technique, Backward Elimination Technique, and Evolutionary Selection Technique. The results of each technique were used to calculate the predictive value of breast cancer and show that the Evolutionary Selection technique has the best effect of 7 techniques from a total of 30 data attributes. The Evolutionary Selection technique significantly reduces to 16 significant attributes, which provides a good predictive accuracy of 95.26%.

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

สารบัญ

บทที่ **หน้า**

1 บทนำ 1

1.1 ความเป็นมาและความสำคัญของปัญหา 1

1.2 วัตถุประสงค์ของการวิจัย 2

1.3 คำถามการวิจัย 2

1.4 สมมุติฐานการวิจัย 2

1.5 ขอบเขตของการวิจัย 3

1.6 นิยามศัพท์ 3

2 วรรณกรรมที่เกี่ยวข้อง 5

2.1 ความรู้พื้นฐาน 5

2.2 ทฤษฎีที่รองรับเรื่องที่วิจัย 8

2.3 ผลการวิจัยที่เกี่ยวข้อง 17

3 ระเบียบวิธีวิจัย 18

3.1 ขั้นตอนการดำเนินงานวิจัย 18

3.1.1 จัดหาข้อมูลและเตรียมข้อมูล 20

3.1.2 การวิเคราะห์ข้อมูลด้วยโปรแกรม RapidMiner 25

3.2 ผลการทดลอง 26

4 ผลการวิเคราะห์ข้อมูล 51

4.1 ผลการวิเคราะห์ข้อมูล 51

สารบัญ (ต่อ)

บทที่ **หน้า**

5 สรุป อภิปราย และข้อเสนอแนะ 55

5.1 สรุปผลการวิจัย..... 55

5.2 อภิปรายผล 56

5.3 ข้อเสนอแนะ 56

บรรณานุกรม 57

ประวัติย่อผู้วิจัย..... 60

สารบัญตาราง

ตารางที่ **หน้า**

1 ค่าน้ำหนักของคุณลักษณะจากการใช้เทคนิค CBF 26

2 ค่าน้ำหนักของคุณลักษณะจากการใช้เทคนิค IG..... 31

3 ค่าความถูกต้องของการพยากรณ์การเป็นมะเร็งเต้านมตามจำนวนคุณลักษณะที่ใช้ด้วยเทคนิค IG 33

4 ค่าน้ำหนักของคุณลักษณะจากการใช้เทคนิค GR 36

5 ค่าความถูกต้องของการพยากรณ์การเป็นมะเร็งเต้านมตามจำนวนคุณลักษณะที่ใช้ด้วยเทคนิค GR..... 37

6 ค่าน้ำหนักของคุณลักษณะจากการใช้เทคนิค Chi Squared..... 40

7 ค่าความถูกต้องของการพยากรณ์การเป็นมะเร็งเต้านมตามจำนวนคุณลักษณะที่ใช้ด้วยเทคนิค Chi Squared..... 41

8 ค่าน้ำหนักของคุณลักษณะจากการใช้เทคนิค Forward Selection 45

9 ค่าน้ำหนักของคุณลักษณะจากการใช้เทคนิค Backward Elimination..... 47

10 ค่าน้ำหนักของคุณลักษณะจากการใช้เทคนิค Evolutionary Selection 49

สารบัญภาพประกอบ

ภาพประกอบ **หน้า**

1 The CRISP-DM Reference Model 6

2 การคัดเลือกคุณลักษณะแบบ Filter Approach 8

3 การคัดเลือกคุณลักษณะแบบ Wrapper Approach 9

4 Greedy (heuristic) methods for attribute subset selection 12

5 ตำแหน่งข้อมูลสองกลุ่มในฟีเจอร์สเปซ (Feature Space) 14

6 แสดงขั้นตอนการดำเนินการวิจัย 19

7 แสดงข้อมูลสรุปที่นำมาใช้ในการพยากรณ์ 20

8 การแสดงรายละเอียดและลักษณะของข้อมูลแต่ละคุณลักษณะในกลุ่มค่าเฉลี่ย 23

9 การแสดงรายละเอียดและลักษณะของข้อมูลแต่ละคุณลักษณะในกลุ่มความคลาดเคลื่อนมาตรฐาน 24

10 การแสดงรายละเอียดและลักษณะของข้อมูลแต่ละคุณลักษณะในกลุ่มค่าที่แย่ที่สุด 24

11 ภาพรวมแสดงการพยากรณ์การเป็นมะเร็งเต้านมด้วยเทคนิค SVM ด้วยโปรแกรม RapidMiner 27

12 การเลือกคุณลักษณะที่ได้รับการคัดเลือกเข้าประมวลผล 28

13 กระบวนการในการทำ Validation การประมวลผลด้วย SVM 28

14 กำหนดค่าพารามิเตอร์ให้กับ SVM 29

15 กำหนดค่าพารามิเตอร์ในการทำ Cross validation 29

สารบัญภาพประกอบ (ต่อ)

ภาพประกอบ **หน้า**

16 ผลการวัดประสิทธิภาพจากการเลือกคุณลักษณะของเทคนิค CBF 30

17 ตัวอย่างการเลือกคุณลักษณะโดยการตัดคุณลักษณะที่มีค่าน้ำหนักน้อยที่สุดออกของเทคนิค IG 32

18 ผลการวัดประสิทธิภาพจากการตัดคุณลักษณะที่มีค่าน้ำหนักน้อยที่สุดออกของเทคนิค IG 33

19 การนำ 6 คุณลักษณะที่ได้รับการคัดเลือกเข้าประมวลผลของเทคนิค IG 34

20 ผลการวัดประสิทธิภาพจากการเลือก 6 คุณลักษณะของเทคนิค IG..... 35

21 ตัวอย่างการเลือกคุณลักษณะโดยการตัดคุณลักษณะที่มีค่าน้ำหนักน้อยที่สุดออกของเทคนิค GR..... 38

22 ผลการวัดประสิทธิภาพจากการตัดคุณลักษณะที่มีค่าน้ำหนักน้อยที่สุดออกของเทคนิค GR 39

23 ตัวอย่างการเลือกคุณลักษณะโดยการตัดคุณลักษณะที่มีค่าน้ำหนักน้อยที่สุดออกด้วยเทคนิค Chi Squared..... 42

24 ผลการวัดประสิทธิภาพจากการตัดคุณลักษณะที่มีค่าน้ำหนักน้อยที่สุดออกด้วยเทคนิค Chi Squared..... 43

25 การนำ 9 คุณลักษณะที่ได้รับการคัดเลือกเข้าประมวลผลของเทคนิค Chi Squared..... 44

สารบัญภาพประกอบ (ต่อ)

ภาพประกอบ **หน้า**

26 ผลการวัดประสิทธิภาพจากการเลือก 9 คุณลักษณะด้วยเทคนิค Chi Squared 44

27 ผลการวัดประสิทธิภาพจากการเลือกคุณลักษณะด้วยเทคนิค Forward Selection 46

28 ผลการวัดประสิทธิภาพจากการเลือกคุณลักษณะด้วยเทคนิค Backward Elimination 48

29 ผลการวัดประสิทธิภาพจากการเลือกคุณลักษณะด้วยเทคนิค Evolutionary Selection 50

30 แสดงจำนวนคุณลักษณะที่สำคัญที่ถูกคัดเลือกในแต่ละเทคนิค 7 เทคนิค 52

31 แสดงการเปรียบเทียบการวัดประสิทธิภาพการพยากรณ์การเป็นมะเร็งเต้านมจาก 7 เทคนิคที่เลือกใช้ 53

32 แสดงจำนวนคุณลักษณะและผลการวัดประสิทธิภาพการพยากรณ์การเป็นมะเร็งเต้านมของเทคนิค 7 เทคนิค 54

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การทำเหมืองข้อมูล (Data Mining) เป็นกระบวนการค้นรูปแบบ หรือความสัมพันธ์ที่มีประโยชน์จากข้อมูลที่เก็บรวบรวมมาในอดีต โดยใช้โมเดลต่างๆ เพื่อนำมาใช้ทำนายหรือพยากรณ์การเกิดเหตุการณ์ในอนาคต การทำเหมืองข้อมูลมีกระบวนการตามมาตรฐานอุตสาหกรรม (CRIPS-DM: Cross-Industry Standard Process for Data Mining) ที่ประกอบด้วยขั้นตอนสำคัญๆ 6 ขั้นตอนคือ

1) ทำความเข้าใจปัญหา (Business Understanding) ระบุเป้าหมายที่ต้องการจากการวิเคราะห์ข้อมูลด้วยการทำเหมืองข้อมูล
2) ทำความเข้าใจในข้อมูล (Data Understanding) รวบรวมข้อมูลที่เกี่ยวข้อง ทำความเข้าใจในข้อมูล สำรวจความถูกต้องและจำนวนของข้อมูลให้มีรายละเอียดเพียงพอกับการวิเคราะห์
3) การเตรียมข้อมูล (Data Preparation) ขั้นตอนนี้จะทำการคัดเลือกข้อมูล โดยจะเลือกเฉพาะข้อมูลที่เกี่ยวข้องกับสิ่งที่ต้องการวิเคราะห์ ทำการลบข้อมูลที่ซ้ำซ้อน แก้ไขข้อมูลที่ผิดพลาด แปลงรูปแบบข้อมูลให้อยู่ในรูปแบบที่พร้อมจะนำไปวิเคราะห์
4) การสร้างแบบจำลอง (Modeling) เป็นการสร้างแบบจำลองวิเคราะห์ข้อมูลด้วยเทคนิคเหมืองข้อมูล โดยแบบจำลองจะขึ้นอยู่กับวัตถุประสงค์ของการวิเคราะห์ข้อมูล การเลือกข้อมูลและปรับพารามิเตอร์ของเทคนิคเหมืองข้อมูลเพื่อให้ได้ประสิทธิภาพที่ดี
5) การประเมินวัดประสิทธิภาพของแบบจำลอง (Evaluation) ขั้นตอนนี้จะทำการวัดประสิทธิภาพของแบบจำลองที่สร้างขึ้นเพื่อความถูกต้อง เพื่อที่จะได้เลือกแบบจำลองที่เหมาะสมที่สุดไปใช้งาน
6) การนำแบบจำลองไปใช้งานจริง (Deployment) เป็นการนำแบบจำลองไปใช้งานจริงเพื่อให้บรรลุวัตถุประสงค์ของการทำเหมืองข้อมูล

โดยในงานวิจัยฉบับนี้จะใช้เทคนิคการทำเหมืองข้อมูล มาปรับปรุงการพยากรณ์การเป็นโรคมะเร็งเต้านม โดยเน้นการคัดเลือกคุณลักษณะ (Feature Selection) ซึ่งอยู่ในขั้นตอนที่ 3 (การเตรียมข้อมูล) ของการทำเหมืองข้อมูล การคัดเลือกคุณลักษณะเป็นเทคนิคที่ช่วยลดจำนวนมิติที่ใช้ในแบบจำลองให้ลดลง ซึ่งเป็นขั้นตอนที่สำคัญที่จะทำให้ผลการวิเคราะห์ข้อมูลมีประสิทธิภาพมากขึ้น

1.2 วัตถุประสงค์ของการวิจัย

การวิจัยนี้เป็นการวิจัยองค์ความรู้ เพื่อทำการคัดเลือกคุณลักษณะที่สำคัญในการวิเคราะห์ข้อมูลเพื่อลดจำนวนมิติที่ใช้ในการปรับปรุงการ พยากรณ์การเป็นมะเร็งเต้านม โดยทำการเปรียบเทียบเทคนิคที่ใช้ในการคัดเลือกคุณลักษณะที่มีความสำคัญในการวิเคราะห์ข้อมูล โดยเทคนิคที่ใช้ในการวิจัยครั้งนี้มี 7 แบบคือ 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 และใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) ในการจำแนกข้อมูลของการเป็นมะเร็งเต้านมหรือการไม่เป็นมะเร็งเต้านม

1.3 คำถามการวิจัย

คำถามการวิจัยได้กำหนดขึ้นตามวัตถุประสงค์ของการวิจัยคือ

1) ในการลดจำนวนคุณลักษณะของข้อมูลที่ใช้ในการวิเคราะห์ข้อมูล โดยเทคนิค 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 เทคนิคไหนจะมีความสามารถในการลดจำนวนคุณลักษณะที่ใช้ในการวิเคราะห์ข้อมูลมากที่สุด และมีประสิทธิภาพในการวิเคราะห์ข้อมูลสูงสุด
2) การลดจำนวนคุณลักษณะโดยใช้เทคนิคข้างต้น จะสามารถเพิ่มประสิทธิภาพในการวิเคราะห์ข้อมูล เมื่อเทียบกับการไม่ลดจำนวนคุณลักษณะในการวิเคราะห์ข้อมูลหรือไม่อย่างไร

1.4 สมมุติฐานการวิจัย

การคัดเลือกคุณลักษณะที่สำคัญในการวิเคราะห์ข้อมูล ทำให้จำนวนมิติของข้อมูลที่ใช้ในการวิเคราะห์ข้อมูลลดลง น่าจะมีผลในการเพิ่มประสิทธิภาพการวัดค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านม

1.5 ขอบเขตของการวิจัย

การวิจัยจะทดสอบทำการเปรียบเทียบเทคนิค 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 เพื่อวัดประสิทธิภาพในการพยากรณ์การเป็นมะเร็งเต้านม โดยเปรียบเทียบกับการใช้ข้อมูลทั้งหมดที่ไม่มีการลดจำนวนมิติลง เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) จะถูกนำมาใช้ในการจำแนกข้อมูลของการเป็นมะเร็งเต้านมหรือการไม่เป็นมะเร็งเต้านม โดยกำหนดระยะเวลาไว้ 1 ปี

1.6 นิยามศัพท์

1.6.1 การทำเหมืองข้อมูล (Data Mining) หมายถึงกระบวนการที่กระทำกับข้อมูล(โดยส่วนใหญ่จะมีจำนวนมาก) เพื่อค้นหารูปแบบ แนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น โดยอาศัยหลักสถิติ การรู้จำ การเรียนรู้ของเครื่อง และหลักคณิตศาสตร์

1.6.2 การคัดเลือกคุณลักษณะ (Feature Selection) หมายถึง การลดขนาดหรือมิติของข้อมูลและยังคงลักษณะสำคัญของข้อมูล

1.6.3 Correlation Based Feature Selection (CFS) หมายถึงเทคนิคที่ใช้การหากลุ่มคุณลักษณะที่ได้รับการประเมินค่าจากความสามารถในการคาดการณ์ โดยคุณลักษณะที่ถูกคัดเลือกใช้สำหรับการจำแนกประเภทของข้อมูล

1.6.4 Information Gain (IG) หมายถึงเทคนิคการพิจารณาจากค่าความน่าจะเป็นของแต่ละคุณลักษณะที่เป็นไปได้แล้ววัดค่าความไร้ระเบียบ (Entropy) เพื่อคัดเลือกคุณลักษณะที่มีความสำคัญในการจำแนกกลุ่มได้ดีที่สุด

1.6.5 Gain Ratio (GR) หมายถึงเทคนิคการวัดจำนวนบิตของข้อมูล เพื่อใช้ในการทำนายคาดเดาจำแนกประเภทของข้อมูล

1.6.6 Chi Square หมายถึงเทคนิคการวัดโดยการใช้สถิติประมาณความสัมพันธ์ร่วมระหว่างคุณลักษณะ เฉพาะกับคลาสของคุณลักษณะเฉพาะ

1.6.7 Forward Selection เป็นเทคนิคที่ใช้โดยการเพิ่มคุณลักษณะทีละ 1 คุณลักษณะ ถ้าคุณลักษณะที่ใส่เพิ่มให้ประสิทธิภาพที่ดีก็จะเก็บไว้และเลือกคุณลักษณะอื่นๆมาเพิ่มต่อไปจนประสิทธิภาพของโมเดลไม่ได้ดีขึ้นก็จะหยุดทำงาน

1.6.8 Backward Elimination เป็นเทคนิคที่ใช้โดยการลดคุณลักษณะ (Backward Elimination) เริ่มต้นด้วยคุณลักษณะทั้งหมดก่อน และตัดคุณลักษณะที่ไม่สำคัญทิ้งไปทีละคุณลักษณะ ถ้าประสิทธิภาพดีขึ้นก็ตัดคุณลักษณะอื่นๆต่อไป จนกว่าจะพบว่าคุณลักษณะนั้นมีผลทำให้ประสิทธิภาพลดลงอย่างมีนัยสำคัญทางสถิติ

1.6.9 Evolutionary Selection เป็นการคัดเลือกคุณลักษณะ โดยวิธีการสุ่มเลือกคุณลักษณะซึ่งเป็นตัวแปรพยากรณ์เข้ามาในสมการทีละตัว และทำการทดสอบหาประสิทธิภาพในการพยากรณ์คำตอบ หากค่าประสิทธิภาพในการคาดการณ์สูงขึ้นจะเก็บคุณลักษณะนั้นไว้ แล้วทำการสุ่มเลือกคุณลักษณะอื่นเข้าไปเพิ่ม หากค่าประสิทธิภาพต่ำลงก็จะถอดคุณลักษณะนั้นออก

1.6.10 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) หมายถึงวิธีการเรียนรู้แบบมีผู้สอนใช้เพื่อการแบ่งประเภทข้อมูล

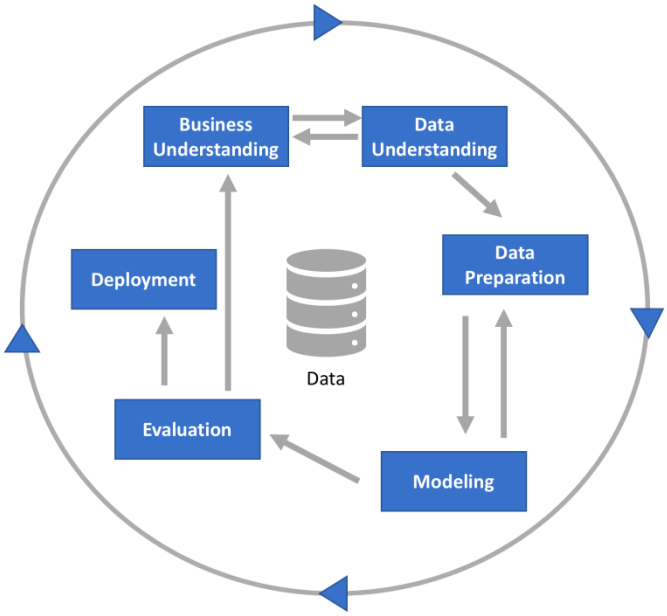
1.6.11 กระบวนการมาตรฐานอุตสาหกรรมสำหรับเหมืองข้อมูล (Cross-Industry Standard Process for Data Mining, CRIPS-DM) หมายถึงกระบวนการมาตรฐานในการวิเคราะห์ข้อมูลด้านเหมืองข้อมูล

บทที่ 2
วรรณกรรมที่เกี่ยวข้อง

2.1 ความรู้พื้นฐาน

ในการทำเหมืองข้อมูลและการเรียนรู้การแก้ปัญหาด้วยคอมพิวเตอร์ในโลกแห่งความเป็นจริงมักเกี่ยวข้องกับข้อมูลที่กำลังเรียนรู้จำนวนมาก อย่างไรก็ตามข้อมูลบางอย่างไม่มีความจำเป็น เนื่องจากมีความซ้ำซ้อนหรือแม้กระทั่งไม่เกี่ยวข้อง ซึ่งอาจลดประสิทธิภาพในการประมวลผลข้อมูลจากการทำเหมืองข้อมูล หรืออาจได้ผลไม่เป็นไปตามวัตถุประสงค์ ดังนั้นการเลือกข้อมูลจึงเป็นสิ่งจำเป็นโดยเฉพาะการเลือกข้อมูลให้ได้ตามคุณลักษณะที่ต้องการ การเลือกคุณลักษณะข้อมูลนอกจากช่วยการแก้ปัญหาการประมวลผลข้อมูลจำนวนมากที่ไม่จำเป็นยังสามารถลดมิติของข้อมูลเป็นการเร่งกระบวนการเรียนรู้ทำให้รูปแบบการเรียนรู้ง่ายขึ้น และเพิ่มประสิทธิภาพการประมวลผลข้อมูลได้เป็นอย่างดี

การทำเหมืองข้อมูล (Data Mining) เป็นกระบวนการค้นรูปแบบหรือความสัมพันธ์ที่มีประโยชน์จากข้อมูลที่เก็บรวบรวมมาในอดีต โดยใช้โมเดลต่างๆ เพื่อนำมาใช้ทำนายหรือพยากรณ์การเกิดเหตุการณ์ในอนาคต โดยมีกระบวนการตามมาตรฐานอุตสาหกรรม (CRIPS-DM: Cross-Industry Standard Process for Data Mining) ประกอบด้วยขั้นตอนสำคัญๆ 6 ขั้นตอน ตามที่แสดงในภาพประกอบ 1



ภาพประกอบ 1 The CRISP-DM Reference Model

ที่มา : https://sharing.luminis.eu/blog/the-forgotten-step-in-crisp-dm-and-asum-dm-methodologies/

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 สำรวจความถูกต้องและจำนวนของข้อมูลให้มีรายละเอียดเพียงพอกับการวิเคราะห์ขั้นตอนนี้นักวิเคราะห์จะทำความเข้าใจกับข้อมูล ด้วยการรวบรวม และอธิบายข้อมูลเบื้องต้น จากนั้นจะทำการสำรวจข้อมูลและการตรวจสอบคุณภาพข้อมูล

ระบุปัญหาคุณภาพข้อมูลเพื่อค้นหาข้อมูลเชิงลึก รวมทั้งการรวบรวมและการบูรณาการข้อมูลจากหลายแหล่งในกรณีที่จำเป็น

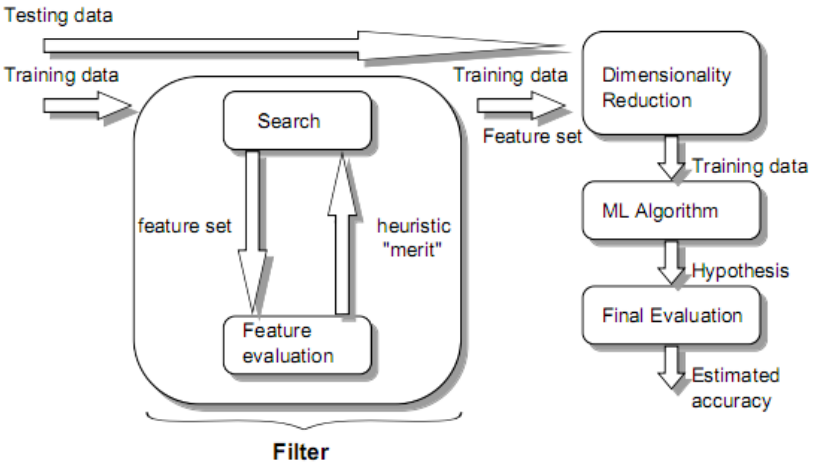
3. การเตรียมข้อมูล (Data Preparation) ขั้นตอนนี้จะทำการคัดเลือกข้อมูล โดยจะเลือกเฉพาะข้อมูลที่เกี่ยวข้องกับสิ่งที่ต้องการวิเคราะห์ ทำการลบข้อมูลที่ซ้ำซ้อน แก้ไขข้อมูลที่ผิดพลาด แปลงรูปแบบข้อมูลให้อยู่ในรูปแบบที่พร้อมจะนำไปวิเคราะห์ การเตรียมข้อมูลจะครอบคลุมกิจกรรมทั้งหมดเพื่อสร้างชุดข้อมูลสุดท้ายหรือข้อมูลที่จะป้อนลงในแบบจำลองจากข้อมูลดิบเริ่มต้น ซึ่งจะรวมถึงการบันทึกข้อมูลและการเลือกคุณลักษณะบิวต์ การแปลงและการทำความสะอาดข้อมูลสำหรับเครื่องมือการสร้างแบบจำลอง โดยมีห้าขั้นตอนในการเตรียมทำข้อมูล ได้แก่ การเลือกข้อมูล การทำความสะอาดข้อมูล การสร้างข้อมูล การรวมข้อมูล และการจัดรูปแบบข้อมูล เลือกข้อมูลเป็นการตัดสินใจเกี่ยวกับข้อมูลที่จะใช้ในการวิเคราะห์ ซึ่งขึ้นอยู่กับเกณฑ์หลายประการ รวมถึงความเกี่ยวข้องกับเป้าหมายการทำเหมืองข้อมูล ข้อจำกัดด้านคุณภาพของข้อมูลที่มีผลโดยตรงกับผลลัพธ์ที่ได้ เทคนิค และปริมาณข้อมูลหรือชนิดข้อมูล ส่วนหนึ่งของกระบวนการคัดเลือกข้อมูลควรเกี่ยวข้องกับการอธิบายว่าเหตุใดข้อมูลบางอย่างถูกรวมหรือควรตัดออก หรือคุณลักษณะไหนมีความสำคัญมากกว่าคุณลักษณะอื่น
4. การสร้างแบบจำลอง (Modeling) เป็นการสร้างแบบจำลองวิเคราะห์ข้อมูลด้วยเทคนิคเหมืองข้อมูล โดยแบบจำลองจะขึ้นอยู่กับวัตถุประสงค์ของการวิเคราะห์ข้อมูล การเลือกข้อมูลและปรับพารามิเตอร์ของเทคนิคเหมืองข้อมูลเพื่อให้ได้ประสิทธิภาพที่ดี ดังนั้นในขั้นตอนนี้จึงเป็นการเลือกเทคนิค เพื่อการสร้างแบบจำลอง
5. การประเมินวัดประสิทธิภาพของแบบจำลอง (Evaluation) ขั้นตอนนี้จะทำการวัดประสิทธิภาพของแบบจำลองที่สร้างขึ้นเพื่อความถูกต้อง เพื่อที่จะได้เลือกแบบจำลองที่เหมาะสมที่สุดไปใช้งาน
6. การนำแบบจำลองไปใช้งานจริง (Deployment) เป็นการนำแบบจำลองไปใช้งานจริงเพื่อให้บรรลุวัตถุประสงค์ของการทำเหมืองข้อมูล

โดยในงานวิจัยชิ้นนี้จะใช้เทคนิคการทำเหมืองข้อมูล มาทำนายการเป็นโรคมะเร็งเต้านม โดยเน้นการคัดเลือกคุณลักษณะ (Feature Selection) ซึ่งอยู่ในขั้นตอนที่ 3 (การเตรียมข้อมูล) ของการทำเหมืองข้อมูล การคัดเลือกคุณลักษณะเป็นเทคนิคที่สำคัญและจำเป็น เนื่องจากสามารถช่วยลดจำนวนมิติที่ใช้ในแบบจำลองให้ลดลง ซึ่งจะเป็นผลทำให้การวิเคราะห์ข้อมูลมีประสิทธิภาพมากขึ้น

2.2 ทฤษฎีที่รองรับเรื่องที่วิจัย

การคัดเลือกคุณลักษณะที่สำคัญ เป็นการคัดเลือกเซตของคุณลักษณะเฉพาะใหม่จากเซตของคุณลักษณะเดิม โดยจะเป็นเซตย่อยของเซตคุณลักษณะเดิม เพื่อเพิ่มประสิทธิภาพในการทำนายและการทำงานให้เร็วขึ้น เนื่องจากบางคุณลักษณะเดิมที่ไม่มีความสำคัญหรือมีความสำคัญน้อยได้ถูกคัดเลือกทิ้งไป การคัดเลือกคุณลักษณะที่สำคัญสามารถแบ่งออกเป็น 2 วิธีได้แก่

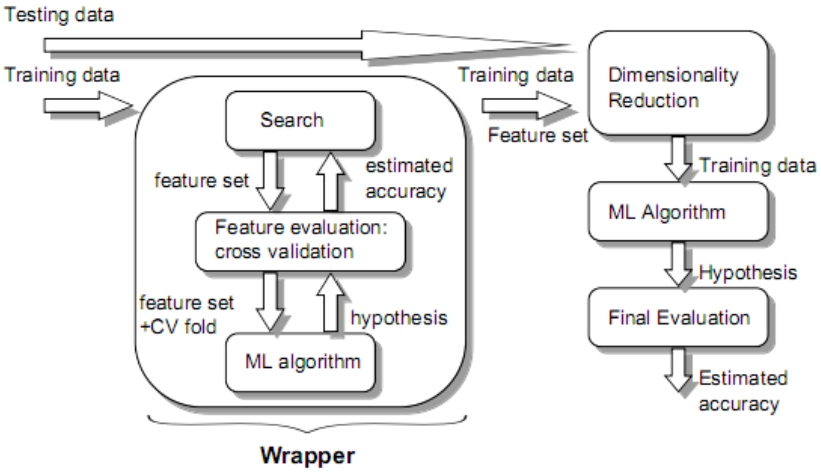
Filter Approach เป็นวิธีที่พยายามคัดเลือกคุณลักษณะที่สำคัญ โดยการคำนวณค่าน้ำหนักหรือค่าความสัมพันธ์ของแต่ละคุณลักษณะ โดยเลือกเฉพาะคุณลักษณะที่มีความสำคัญเก็บเอาไว้ เช่นเทคนิค Correlation Based Feature Selection (CFS) เทคนิค Information Gain (IG) เทคนิค Gain Ratio (GR) เทคนิค Chi-Square



ภาพประกอบ 2 การคัดเลือกคุณลักษณะแบบ Filter Approach

ที่มา : Mark A. Hall, 1999

Wrapper Approach เป็นวิธีคัดเลือกคุณลักษณะที่สำคัญ โดยการคำนวณค่าน้ำหนักการวัดค่าความถูกต้องในการแบ่งกลุ่มข้อมูล มาสร้างเซตของคุณลักษณะใหม่ โดยการเพิ่มหรือลดจำนวนคุณลักษณะจากเซตเดิม เช่นเทคนิค Forward Selection เทคนิค Backward Elimination เทคนิค Evolutionary Selection



ภาพประกอบ 3 การคัดเลือกคุณลักษณะแบบ Wrapper Approach

ที่มา : Mark A. Hall, 1999

2.2.1 เทคนิค Correlation Based Feature Selection (CFS)

เทคนิค Correlation Based Feature Selection (CFS) เป็นเทคนิคในการเลือกคุณสมบัติของคุณลักษณะโดยใช้การพิจารณาบนพื้นฐานความสัมพันธ์ (Correlation-Based Feature Selection : CFS) ของกลุ่มคุณลักษณะที่ได้จากการประเมินค่าจากความสามารถในการคาดการณ์ คุณลักษณะที่ถูกคัดเลือกใช้สำหรับจำแนกประเภทข้อมูล และยังสามารถจัดการกับคุณลักษณะที่ไม่เกี่ยวข้อง CFS จะจัดอันดับกลุ่มย่อยของมิติข้อมูล ทำการคัดเลือกกลุ่มย่อยของมิติข้อมูลที่มีความสัมพันธ์กันสูงกับคลาส และไม่มีความสัมพันธ์กับคลาสอื่นๆ สำหรับมิติข้อมูลที่ไม่เกี่ยวข้องหรือมีความสัมพันธ์ต่ำกับคลาสจะถูกลบทิ้ง มิติข้อมูลที่ซ้ำซ้อนจะถูกขจัดออกไปจากกลุ่มมิติข้อมูลที่มีความสัมพันธ์สูง สมการประเมินกลุ่มย่อยของมิติข้อมูลแบบ CFS แสดงในสมการที่ (1) (ภัทราวุฒิ แสงศิริ, 2553)

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}} \quad (1)$$

โดยที่ M_s คือค่าที่ค้นหาได้ของมิติข้อมูลกลุ่มย่อย S ซึ่งประกอบด้วยมิติข้อมูล K

$\overline{r_{cf}}$ คือค่าเฉลี่ยความสัมพันธ์ของตัวแปรกับคลาส $(f \in s)$

$\overline{r_{ff}}$ คือค่าเฉลี่ยความสัมพันธ์ระหว่างมิติข้อมูล

2.2.2 เทคนิค Information Gain (IG)

เทคนิค Information Gain (IG) เป็นเทคนิคที่ใช้ในการชี้วัดเพื่อเลือกคุณลักษณะข้อมูลที่ต้องการที่น้อยที่สุดที่เหมาะสมในการนำไปใช้ระบุ โดยการคำนวณหาค่า Gain สำหรับแต่ละมิติข้อมูล ซึ่งมิติข้อมูลใดมีค่า Gain สูงสุด จะถูกคัดเลือกเพื่อนำมาใช้ระบุ สมการที่ 2 แสดงการคำนวณค่า Information Gain หรือค่า Entropy ของชุดข้อมูลทั้งหมด สมการที่ 3 แสดงการคำนวณค่า Entropy ของชุดมิติข้อมูลในแต่ละคุณลักษณะ สมการที่ 4 เป็นการคำนวณหาค่า Information Gain สำหรับการพิจารณามิติข้อมูลคุณลักษณะ $\boldsymbol{A}$

$$E(D) = -\sum_{i=1}^{n} p_i \log_2(p_i) \quad (2)$$

$$E_A(D) = \sum_{j=1}^{m} \frac{|D_j|}{D} \times E(D_j) \quad (3)$$

$$\text{Gain}(A) = E(D) - E_A(D) \quad (4)$$

โดยที่ p_i คือ ค่าความน่าจะเป็นที่เรคคอร์ดหนึ่งๆจะมีหมวดหมู่ของข้อมูล

หรือกล่าวว่า การคำนวณค่า Information Gain คือการวัดค่า Entropy ก่อนที่จะมีการแบ่งข้อมูลออกตามมิติข้อมูลและหลังการแบ่งว่ามีประสิทธิภาพดีขึ้นหรือไม่ ถ้ามีประสิทธิภาพดีขึ้นค่า Information Gain จะมีค่าสูง

2.2.3 เทคนิค Gain Ratio (GR)

เทคนิค Gain Ratio (GR) เป็นตัวชี้วัดการแบ่งชุดข้อมูลออกเป็นชุดข้อมูลย่อยที่พัฒนามาจาก Information Gain เนื่องจากการใช้ Information Gainในการแบ่งชุดข้อมูลจะมีโอกาสทำให้เกิดความเอนเอียงขึ้น เมื่อคุณลักษณะที่ทำการพิจารณาได้ค่า gain ที่สูงเป็นจำนวนมาก ทำให้คุณลักษณะที่ถูกคัดเลือกไม่ถูกต้อง ตัวอย่างเช่นพิจารณาคุณลักษณะที่ทำหน้าที่เป็นตัวระบุเฉพาะเช่นรหัสผลิตภัณฑ์ การแยกรหัสผลิตภัณฑ์จะส่งผลให้มีชุดข้อมูลย่อยจำนวนมาก แต่ละชุดข้อมูลย่อยมีเพียงหนึ่ง record ซึ่งเมื่อนำมาหาค่า Information Gain จะได้ค่าที่สูงจำนวนมากนั่นเอง

จากความเอนเอียง ทำให้มีการพัฒนาตัวชี้วัดการแบ่งข้อมูลใหม่ที่ชื่อเรียกว่า Gain Ratio โดยการประยุกต์ใช้การทำนอร์มัลไลซ์ค่า Information Gain ด้วยการใช้ค่า "Split Information" ซึ่งสามารถคำนวณได้ดังสมการที่ 5

$$SplitInfo_A(D) = -\sum_{j=1}^{m} \frac{|D_j|}{|D|} \text{ x } \log_2 \frac{|D_j|}{|D|} \quad (5)$$

โดยค่า $SplitInfo_A(D)$ หมายถึงปริมาณข้อมูลที่ถูกพิจารณาโดยการแบ่งข้อมูลในชุดข้อมูล D ออกเป็น m ชุดข้อมูลย่อยตามค่าคุณลักษณะ A โดยหลังจากทำการคำนวณหาค่า $SplitInfo_A(D)$ แล้ว เราจะสามารถคำนวณหาค่า Gain Ratio ได้ดังสมการที่ 6

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)} \quad (6)$$

2.2.4 เทคนิค Chi-Square

Chi Squared (x^2) การประเมินค่าของคุณลักษณะโดยการใช้การคำนวณค่า Chi-Square ทางสถิติ เพื่อศึกษา ว่าการแจกแจงความถี่ของตัวแปรคุณลักษณะเป็นไปตามรูปแบบที่กำหนดไว้หรือไม่ ดังแสดงใน สมการที่ 7

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (7)$$

โดยที่ *O1 , O2 ... On* เป็นความถี่ของตัวแปรที่ได้จากการศึกษา

E1 , E2 ... En เป็นความถี่ที่คาดหวัง (หรือความถี่ที่ควรจะเป็น)

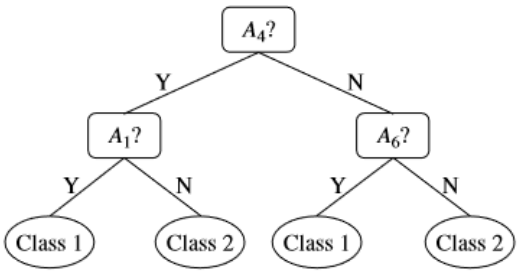
2.2.5 เทคนิค Forward Selection

เทคนิค Forward Selection เป็นเทคนิคที่ใช้วิธีการคัดเลือกคุณลักษณะซึ่งเป็นตัวแปรอิสระเข้ามาในสมการทีละตัว ขั้นแรกพิจารณาจากค่าสัมบูรณ์ของสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) ระหว่างตัวแปรตามกับตัวแปรอิสระที่ให้ค่าสูงสุด หากตัวแปรอิสระตัวนั้นมีคุณสมบัติตามเกณฑ์การนำเข้า ก็จะถูกนำเข้าสมการ หลังจากนั้นเป็นการพิจารณาสัมประสิทธิ์สหสัมพันธ์บางส่วนระหว่างตัวแปรตามกับตัวแปรอิสระที่ไม่อยู่ในสมการถดถอยทีละตัว ตัวแปรอิสระที่มีค่าสัมบูรณ์ของสัมประสิทธิ์สหสัมพันธ์บางส่วนสูงสุดจะถูกนำเข้าสมการ หากตัวแปรนั้นมีคุณสมบัติตามเกณฑ์การนำเข้า ขั้นตอนจะทำซ้ำจนกระทั่งพบว่าไม่สามารถนำตัวแปรอิสระเข้าสมการได้ จึงหยุด

2.2.6 เทคนิค Backward Elimination

เทคนิค Backward Elimination เป็นเทคนิคที่พยายามคัดเลือกตัวแปรที่ดีที่สุดและได้โมเดลประหยัดในการพยากรณ์เช่นเดียวกัน โดยในตอนแรกจะนำตัวแปรพยากรณ์ทุกตัวเข้ามาในสมการ และดำเนินการพิจารณาตัวแปรพยากรณ์ที่มีค่าสัมประสิทธิ์สหสัมพันธ์บางส่วน (Partial Correlation) กับตัวแปรเกณฑ์ และควบคุมอิทธิพลของตัวแปรพยากรณ์อื่นๆ ซึ่งมีค่าต่ำที่สุดออกจากสมการ แล้วจึงดำเนินการทดสอบว่า ค่า R2 ลดลงอย่างมีนัยสำคัญทางสถิติหรือไม่ ถ้าพบว่าลดลงอย่างไม่มีนัยสำคัญทางสถิติแสดงว่าตัวแปรดังกล่าวไม่ได้มีส่วนทำให้การพยากรณ์ตัวแปรเกณฑ์เพิ่มขึ้นเลย แสดงว่าสามารถขจัดออกจากสมการได้ จากนั้นจึงดำเนินการขจัดตัวแปรพยากรณ์ที่มีความสำคัญน้อยรองลงมาออกไปอีก โดยใช้วิธีพิจารณาเช่นเดียวกันซึ่งการขจัดตัวแปรพยากรณ์จะสิ้นสุดเมื่อพบว่ามีผลทำให้ค่า R2 ลดลงอย่างมีนัยสำคัญทางสถิติ หมายความว่า ตัวแปรดังกล่าวมีความสำคัญต่อการพยากรณ์ตัวแปรตาม หากขจัดตัวแปรดังกล่าวออกจากสมการจะทำให้อำนาจการพยากรณ์ตัวแปรเกณฑ์ลดลงจึงต้องคงตัวแปรพยากรณ์ดังกล่าวไว้ในสมการพยากรณ์ต่อไป (ทรงศักดิ์ ภูสีอ่อน. 2554 : 283 ; Brian S. Everitt. 2010 : 93)

Backward elimination คือการนำตัวแปรอิสระทั้งหมดเข้าสู่สมการ หลังจากนั้นจะคัดตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามน้อยที่สุดออกจากสมการ ดำเนินการซ้ำจนไม่สามารถคัดตัวแปรอิสระออกจากสมการถดถอยได้ ก็จะเหลือสมการถดถอยที่ตัวแปรอิสระมีความความสัมพันธ์กับตัวแปรตามเท่านั้น

Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: {} => $\{A_1\}$ => $\{A_1, A_4\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  => Reduced attribute set: $\{A_1, A_4, A_6\}$

ภาพประกอบ 4 Greedy (heuristic) methods for attribute subset selection.

ที่มา : Jaiwei Han, et al., 2012.

2.2.7 เทคนิค Evolutionary Selection

เทคนิค Evolutionary Selection เป็นเทคนิคการคัดเลือกคุณลักษณะโดยนำข้อมูลคุณลักษณะมาหาประสิทธิภาพในการคาดการณ์คำตอบ จากนั้นจึงนำคุณลักษณะมาจับคู่กัน แล้วหาค่าประสิทธิภาพอีกครั้ง หากค่าประสิทธิภาพในการคาดการณ์สูงขึ้นจะเก็บข้อมูลนั้นไว้ แล้วทำการเลือกคุณลักษณะอื่นเข้าไปเพิ่ม หากค่าประสิทธิภาพต่ำลงจะถอดคุณลักษณะนั้นออก แล้วเลือกคุณลักษณะอื่นเข้าไปดังสมการที่ 8

$$IG(\text{parent child}) = \text{Entropy}(\text{parent}) - [p(c_1) \times \text{Entropy}(c_1) + p(c_2) \times \text{Entropy}(c_2)\ldots] \quad (8)$$

โดยที่ $\text{Entropy}(c_1)$ คือ $-p(c_1) \log p(c_1)$

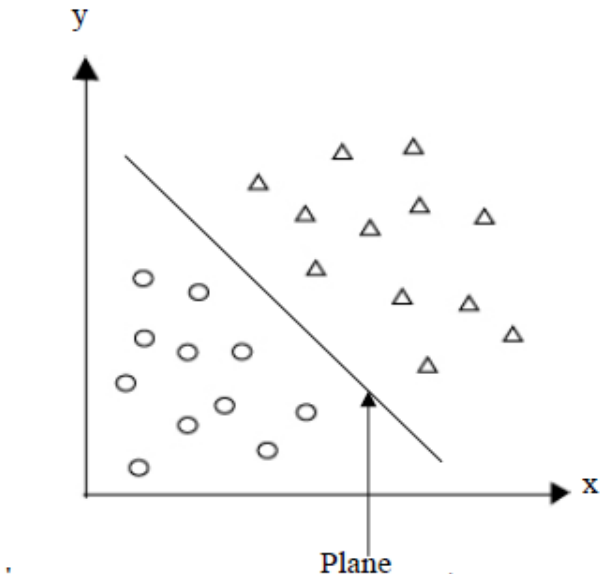
$p(c_1)$ คือ ค่าความน่าจะเป็นของค่า c_1

c คือ ปัจจัยคุณลักษณะแต่ละตัวที่เกี่ยวข้อง

ทั้งนี้ค่า Entropy จะใช้ในการวัดความแตกต่างกันของข้อมูล ซึ่งถ้าข้อมูลมีความแตกต่างกันน้อยจะมีค่า Entropy ต่ำ และถ้าข้อมูลมีความแตกต่างกันมากจะมีค่า Entropy สูง (เอกสิทธิ์ พัชรวงศ์ศักดา, 2557)

2.2.8 เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

เทคนิคซัพพอร์ตเวกเตอร์แมชชีน เป็นเทคนิคที่ใช้ในการแก้ปัญหาทางด้านการรู้จำรูปแบบข้อมูล โดยอาศัยหลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้ โดยเน้นไปยังเส้นแบ่งแยกแยะกลุ่มข้อมูลที่ดีที่สุด (Optimal Separating Hyperplane) สำหรับรากฐานเดิมของซัพพอร์ตเวกเตอร์แมชชีน ถูกนำมาใช้กับข้อมูลที่เป็นเชิงเส้น แต่ในความเป็นจริงแล้วข้อมูลที่นำมาใช้ในระบบเรียนรู้ มีทั้งที่เป็นข้อมูลเชิงเส้นและข้อมูลแบบไม่เป็นเชิงเส้น ซึ่งสามารถแก้ปัญหาดังกล่าวด้วยการนำเคอร์เนลฟังก์ชั่นมาใช้ เพื่อทำการแปลงค่าเวคเตอร์ในสเปซข้อมูลนำเข้าไปสู่สเปซข้อมูลแบบอื่นได้ ทำให้สามารถรองรับข้อมูลที่มีมากกว่าสองมิติ แนวความคิดของซัพพอร์ตเวกเตอร์แมชชีน เกิดจากการนำค่าของกลุ่มข้อมูลมาวางลงในฟีเจอร์สเปซ (Feature Space) จากนั้นจึงหาเส้นที่ใช้แบ่งข้อมูลทั้งสองออกจากกันโดยจะสร้างเส้นแบ่ง (Plane) ที่เป็นเส้นตรงขึ้นมา และเพื่อให้ทราบว่าเส้นตรงที่แบ่งสองกลุ่มออกจากกันนั้น เส้นตรงใดเป็นเส้นที่ดีที่สุดดังภาพประกอบ 5



ภาพประกอบ 5 ตำแหน่งข้อมูลสองกลุ่มในฟีเจอร์สเปซ (Feature Space)

ที่มา: จิรา แก้วสุวรรณ์, 2006.

สามารถนำมาเขียนเป็นสมการเพื่อใช้ในการแก้ไขปัญหา โดยข้อมูลที่นำมาวางลงในพื้นที่คุณลักษณะนั้นเป็นกลุ่มข้อมูลที่อยู่ในรูปของเวกเตอร์จะได้

$$X = ((x_1, y_1), \dots\dots, (x_i, y_i)) \tag{9}$$

โดยที่ X คือ ชุดค่าลักษณะเด่น

จากนั้นทำการกำหนดสมการขึ้นมา ซึ่งสมการดังกล่าวมีลักษณะเป็นสมการเส้นตรง เพื่อนำมาสร้างเป็นเส้นตรงบนไฮเปอร์เพลนซึ่งแบ่งกลุ่มข้อมูลที่มีลักษณะเชิงเส้นสองกลุ่มออกจากกัน โดยมีการกำหนดกลุ่มของข้อมูลทั้งสองฝั่งเป็นเพียงสองค่าซึ่งแทนด้วยค่า y ซึ่งข้อมูลที่เป็นตัวกำหนดความชันและระนาบที่เกิดขึ้นบนไฮเปอร์เพลนเกิดจากคู่ข้อมูล และกำหนดสมการที่เป็นตัวบ่งบอกข้อมูลแต่ละกลุ่มว่าอยู่ส่วนไหนของเส้นแบ่งไฮเปอร์เพลนแสดงได้ ดังสมการที่ 10 และสมการที่ 11

เมื่อนำสมการเงื่อนไขทั้งหมดมาวิเคราะห์เชิงเรขาคณิต โดยพิจารณาในกรณีที่ข้อมูลถูกแบ่งกลุ่มได้สมบูรณ์ตามเงื่อนไขของซัพพอร์ตเวกเตอร์แมชชีนในสมการที่ 12 และข้อมูลการสอนให้ระบบเรียนรู้ต้องอยู่ในรูปแบบของเชิงเส้น สามารถแสดงลักษณะการวางตัวของกลุ่มข้อมูลได้ดังภาพประกอบ 5 เวกเตอร์ของข้อมูลที่ถูกป้อนเข้าสู่ระบบการสอน เพื่อให้ระบบเรียนรู้แทนด้วยสมการ และข้อมูลทั้งสองด้านแบ่งเป็นบวกและลบ สถานะของข้อมูลจึงแทนด้วย ซึ่งมีสองค่า คือ $y = 1$ และ $y = -1$ นั้น แต่ทั้งนี้ก็ยังตัดสินไม่ได้ว่าเส้นแบ่งนั้นควรจะเป็นเส้นใดจึงจะดีที่สุด วิธีการที่ใช้ในการหาเส้นแบ่งที่ดีที่สุดคือการเพิ่มเส้นขอบให้กับเส้นทั้งสองข้าง ทำให้ได้เส้นใหม่ที่จะถือเป็นเส้นขอบของข้อมูลแต่ละฝั่งอีกด้วย เส้นขอบ (Margin) ที่เป็นเส้นขอบของเส้นแบ่งนั้นจะเป็นเส้นที่สัมผัสกับค่าข้อมูลในฟีเจอร์สเปซ (Feature Space) ที่ใกล้ที่สุด เส้นขอบของทั้งสองเส้นที่เพิ่มขึ้นมานี้ถูกแทนด้วยสมการ $w^T x + b \geq y \geq 1$ ถ้าอยู่ด้าน $y = 1$ และ $w^T x + b \leq y \leq -1$ ถ้า $y = -1$ หากเส้นขอบของเส้นแบ่งใด ๆ ที่มีความกว้างมากที่สุด แสดงให้เห็นว่าข้อมูลสองชุดมีการแยกกันชัดเจนมากที่สุด ดังนั้นเส้นแบ่งที่มีเส้นขอบกว้างที่สุดจึงเป็นเส้นแบ่งที่ดีที่สุด และเรียกตำแหน่งการสัมผัสข้อมูลที่ใกล้ที่สุดจากการเพิ่มขอบนี้ว่า “เวกเตอร์สนับสนุน” (Support Vector) โดยเรียกเส้นประที่แบ่งข้อมูลทั้งสองเส้นว่าเส้นขอบ ซึ่งสามารถเขียนเป็นสมการการคำนวณความกว้างของเส้นขอบต้องทำการคำนวณพจน์ให้อยู่ในรูปปกติ (Normalization) โดยการคำนวณจากสมการที่ (15) และ (16) เมื่อแทนค่าลงไปแล้ว (จิรา แก้วสุวรรณ์, 2006)

$w^T x + b \geq y$ เมื่อกำหนดให้ $y = 1$ (10)

$w^T x + b \leq y$ เมื่อกำหนดให้ $y = -1$ (11)

$y(w^T x + b) - 1 \geq 0$ (12)

โดยที่ y คือ ค่ากลุ่มข้อมูล (1,-1)

w คือ ค่าความชัน

x คือ ค่าลักษณะเด่น

b คือ ค่าคงที่ (ค่าตัดแกน y)

2.3 ผลการวิจัยที่เกี่ยวข้อง

ในการศึกษาที่ผ่านมาพบว่า การคัดเลือกคุณลักษณะที่สำคัญในการวิเคราะห์ข้อมูลเพื่อลดจำนวนมิติที่ใช้ในการทำเหมืองข้อมูลเป็นสิ่งสำคัญ โดยมีการใช้เทคนิคหลายๆแบบเพื่อให้เหมาะสมกับข้อมูลที่ต้องการวิเคราะห์ โดยมีงานวิจัยที่เกี่ยวข้องดังนี้ (ภัทราวุฒิ แสงศิริ, 2553) การใช้เทคนิคในการคัดเลือกคุณลักษณะที่สำคัญ เพื่อคัดแยกประเภทของมะเร็งเม็ดเลือดขาวแบบเฉียบพลัน ซึ่งจำนวนมิติของข้อมูลมีจำนวน 7,129 มิติ สามารถลดมิติของข้อมูลลงมาเหลือ 36 มิติ และเพิ่มความแม่นยำจากร้อยละ 73.43% มาเป็นร้อยละ 88.24 (นิภาพร ชนะมาร และพรรณี สิทธิเดช, 2557) ได้ศึกษา การวิเคราะห์ปัจจัยการเรียนรู้ด้วยการคัดเลือกคุณสมบัติและการพยากรณ์ โดยใช้เทคนิคการคัดเลือกคุณสมบัติ 3 วิธี ได้แก่ การคัดเลือกคุณสมบัติแบบ Correlation-based Feature Selection การคัดเลือกคุณสมบัติแบบ Consistency-based Feature Selection และ การคัดเลือกคุณสมบัติแบบ Gain Ratio Feature Selection ผลการทดลองทั้งสามเทคนิคสามารถลดจำนวนของคุณสมบัติจาก 22 ตัวแปร เหลือ 9 ตัวแปร 10 ตัวแปร และ 11 ตัวแปร ตามลำดับ (พฤฒิพงศ์ เพ็งศิริ และคณะ, 2557) ได้ศึกษาการลดมิติข้อมูลการวิเคราะห์ความสัมพันธ์และการประยุกต์สำหรับวิเคราะห์ข้อมูลพื้นฐานการใช้งานสมาร์ทโฟน โดยทำการศึกษาเทคนิคต่าง ๆ ในการลดมิติ โดยทำการเปรียบเทียบเทคนิค Partition data of Association with Aprior CFS, Info Gain และ Relief โดยผ่านการจำแนกข้อมูลด้วย ANN (น้ำทิพย์ มากนคร และมาลีรัตน์ โสดานิล, 2557) ศึกษาการเปรียบเทียบวิธีการเลือกคุณลักษณะที่เหมาะสมเพื่อ การจัดหมวดหมู่เว็บเพจผิดกฎหมายโดยใช้เทคนิคการทำเหมืองข้อมูล โดยใช้เทคนิค Information Gain, Gain Ratio และ Chi Squared พบว่าวิธีการจัดอันดับแบบ Chi-Square ให้ผลลัพธ์ที่มีประสิทธิภาพดีที่สุดกล่าวคือ เมื่อลดจำนวนคุณลักษณะ 50% ให้ค่าความถูกต้อง (Accuracy) ร้อยละ 95.98 (นิธินันท์ มาตา และคณะ, 2558) ได้ศึกษาการสร้างโมเดลสำหรับพยากรณ์การควบคุมประตูระบายน้ำ โดยมีการเปรียบเทียบเทคนิค การลดมิติของข้อมูลคือ Information Gain, Gain Ratio, Correlation based Feature Selection (Galavotti L., et al., 2000) ได้ศึกษาเกี่ยวกับการคัดเลือกคุณสมบัติของข้อมูลจากข้อมูลจำนวนมากในการจัดหมวดหมู่ข้อมูลตัวอักษร (Bing Xue., et al., 2016) ได้ศึกษาสรุปการคัดเลือกคุณสมบัติของข้อมูลในการลดมิติของข้อมูล และเพิ่มประสิทธิภาพของขั้นตอนวิธี Evolutionary Computation Approaches

บทที่ 3

ระเบียบวิธีวิจัย

3.1 ขั้นตอนการดำเนินงานวิจัย

งานวิจัยนี้เป็นงานวิจัยเพื่อศึกษาการคัดเลือกคุณลักษณะที่มีความสำคัญในการปรับปรุงการพยากรณ์โรคมะเร็งเต้านม โดยมีการเปรียบเทียบเทคนิคต่างๆ ที่ใช้ในการคัดเลือกคุณลักษณะที่จะนำมาใช้ในการปรับปรุงการทำนายมะเร็งเต้านม โดยมีขั้นตอนการดำเนินงานการวิจัยประกอบด้วยขั้นตอนหลักดังต่อไปนี้

1. การเตรียมข้อมูล (Data Preparation)

เป็นขั้นตอนของการจัดเตรียมข้อมูลเพื่อการวิเคราะห์ โดยนำข้อมูลที่ได้จากแหล่งข้อมูล UCI Machine Learning Repository ซึ่งประกอบไปด้วยข้อมูลจากคนที่มีชีวิตจำนวน 569 คน มีจำนวนคุณลักษณะ 32 คุณลักษณะ ประกอบด้วย หมายเลขประจำตัวและผลการวินิจฉัย อีก 30 คุณลักษณะเป็นคุณลักษณะทางกายภาพ

2. การคัดเลือกคุณลักษณะที่สำคัญโดยการใช้เทคนิคต่างๆ

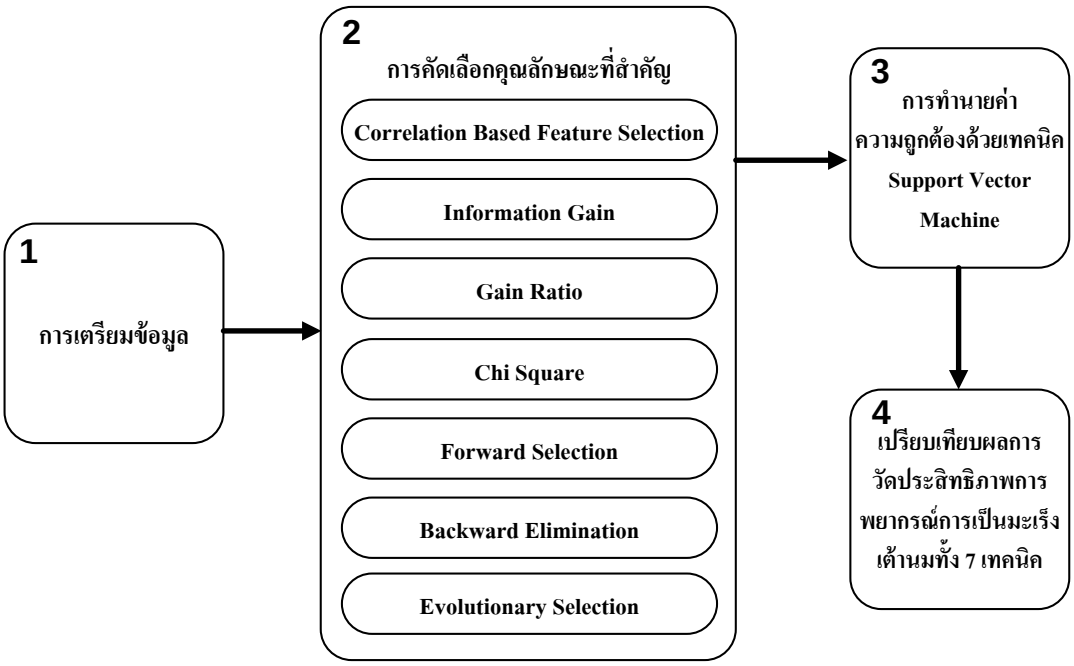
เป็นขั้นตอนการใช้เทคนิคต่างๆจำนวน 7 เทคนิคในการคัดเลือกคุณลักษณะที่สำคัญเพื่อนำมาใช้ในการปรับปรุงการพยากรณ์การเป็นมะเร็งเต้านม

3. การทำนายค่าความถูกต้องด้วยเทคนิค Support Vector Machine

เป็นขั้นตอนการนำคุณลักษณะที่สำคัญที่ถูกคัดเลือกมาแล้วจาก 7 เทคนิค มาวัดประสิทธิภาพการพยากรณ์การเป็นมะเร็งเต้านมโดยการใช้เทคนิค Support Vector Machine (SVM)

4. การเปรียบเทียบผลการวัดประสิทธิภาพการพยากรณ์การเป็นมะเร็งเต้านมทั้ง 7 เทคนิค

เป็นขั้นตอนการนำผลที่ได้จากการวัดประสิทธิภาพการพยากรณ์การเป็นมะเร็งเต้านมจากคุณลักษณะที่สำคัญจากการคัดเลือกด้วยเทคนิคต่างๆ 7 เทคนิคมาเปรียบเทียบ เพื่อดูว่าเทคนิคใดให้ค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมได้เท่าไร



ภาพประกอบ 6 แสดงขั้นตอนการดำเนินการวิจัย

3.1.1 จัดหาข้อมูลและเตรียมข้อมูล

เพื่อการวิเคราะห์เปรียบเทียบเทคนิคต่างๆ ที่ใช้ในการคัดเลือกคุณลักษณะที่จะนำมาใช้ในการปรับปรุงการพยากรณ์มะเร็งเต้านม

ในขั้นตอนนี้ได้นำข้อมูลที่ใช้ในการวิเคราะห์มาจาก UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28diagnostic%29 โดยมีผู้เชี่ยวชาญ 3 ท่านเป็นผู้สร้างข้อมูล

1. Dr. William H. Wolberg, General Surgery Dept.
University of Wisconsin, Clinical Sciences Center, Madison, WI 53792
wolberg '@' eagle.surgery.wisc.edu

2. W. Nick Street, Computer Sciences Dept.
University of Wisconsin, 1210 West Dayton St., Madison, WI 53706
street '@' cs.wisc.edu 608-262-6619

3. Olvi L. Mangasarian, Computer Sciences Dept.
University of Wisconsin, 1210 West Dayton St., Madison, WI 53706
olvi '@' cs.wisc.edu

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

ภาพประกอบ 7 แสดงข้อมูลสรุปที่นำมาใช้ในการพยากรณ์

ที่มา : UCI, “Breast Cancer Wisconsin (Diagnostic) Data Set”, 27 April 2018

เป็นข้อมูลจากคนที่มีชีวิตจริงจำนวน 569 คน จำนวนคุณลักษณะของข้อมูลเท่ากับ 32 คุณลักษณะ โดย 2 คุณลักษณะแรก คือ : หมายเลขประจำตัว และผลการวินิจฉัย อีก 30 คุณลักษณะที่เหลือเป็นคุณลักษณะทางด้านกายภาพถูกนำมาใช้ในการพยากรณ์ ดังนี้

1. id หมายถึง หมายเลขประจำตัว

2. diagnosis (M = malignant, B = benign) หมายถึง การวินิจฉัยโรค แบ่งออกเป็น 2 ประเภท

M = malignant หมายถึง เป็น

B = benign หมายถึง ไม่เป็น

และอีก 30 คุณลักษณะทางด้านกายภาพที่ใช้ในการพยากรณ์ ซึ่งมีรายละเอียดต่อไปนี้

3. radius_mean (mean of distances from center to points on the perimeter) หมายถึง ค่าเฉลี่ยของรัศมี (ระยะทางจากจุดศูนย์กลางไปยังจุดบนเส้นรอบวง)

4. texture_mean (standard deviation of gray-scale values) หมายถึง ค่าเฉลี่ยของลักษณะผิวหนัง

5. perimeter_mean หมายถึง ค่าเฉลี่ยของเส้นรอบรูป

6. area_mean หมายถึง ค่าเฉลี่ยของขนาดพื้นที่

7. smoothness_mean (smoothness : local variation in radius lengths) หมายถึง ค่าเฉลี่ยของความเรียบเนียน

8. compactness_mean (compactness : perimeter^2 / area - 1.0) หมายถึง ค่าเฉลี่ยของความกระทัดรัด

9. concavity_mean (concavity : severity of concave portions of the contour) หมายถึง ค่าเฉลี่ยของส่วนเว้า

10. concave points_mean (concave points : number of concave portions of the contour) หมายถึง ค่าเฉลี่ยจำนวนส่วนเว้า

11. symmetry_mean หมายถึง ค่าเฉลี่ยของความสมมาตร

12. fractal_dimension_mean (fractal_dimension : “coastline approximation” - 1) หมายถึง ค่าเฉลี่ยของสัดส่วนขนาด

13. radius_se (standard error ความคลาดเคลื่อนมาตรฐาน) หมายถึง ความคลาดเคลื่อนมาตรฐานของรัศมี

14. texture_se หมายถึง ความคลาดเคลื่อนมาตรฐานของลักษณะผิวหนัง

15. perimeter_se หมายถึง ความคลาดเคลื่อนมาตรฐานของเส้นรอบรูป

16. area_se หมายถึง ความคลาดเคลื่อนมาตรฐานของขนาดพื้นที่

17. smoothness_se หมายถึง ความคลาดเคลื่อนมาตรฐานของความเรียบเนียน

18. compactness_se หมายถึง ความคลาดเคลื่อนมาตรฐานของความกระทัดรัด

19. concavity_se หมายถึง ความคลาดเคลื่อนมาตรฐานของส่วนเว้า

20. concave points_se หมายถึง ความคลาดเคลื่อนมาตรฐานของจำนวนส่วนเว้า

21. symmetry_se หมายถึง ความคลาดเคลื่อนมาตรฐานของความสมมาตร

22. fractal_dimension_se หมายถึง ความคลาดเคลื่อนมาตรฐานของสัดส่วนขนาด

23. radius_worst หมายถึง ค่าที่แย่ที่สุดของรัศมี

24. texture_ worst หมายถึง ค่าที่แย่ที่สุดของลักษณะผิวหนัง

25. perimeter_ worst หมายถึง ค่าที่แย่ที่สุดของเส้นรอบรูป

26. area_ worst หมายถึง ค่าที่แย่ที่สุดของขนาดพื้นที่

27. smoothness_ worst หมายถึง ค่าที่แย่ที่สุดของความเรียบเนียน

28. compactness_ worst หมายถึง ค่าที่แย่ที่สุดของความกระทัดรัด

29. concavity_ worst หมายถึง ค่าที่แย่ที่สุดของส่วนเว้า

30. concave points_ worst หมายถึง ค่าที่แย่ที่สุดของจำนวนส่วนเว้า

31. symmetry_ worst หมายถึง หมายถึง ค่าที่แย่ที่สุดของความสมมาตร

32. fractal_dimension_ worst หมายถึง ค่าที่แย่ที่สุดของสัดส่วนขนาด

Name	Type	Missing	Statistics		
Filter (32 / 32 attributes): Search for Attributes					
Id **id**	Integer	0	Min 8670	Max 911320502	Average 30371831.432
Label **diagnosis**	Polynominal	0	Least M (212)	Most B (357)	Values B (357), M (212)
radius_mean	Real	0	Min 6.981	Max 28.110	Average 14.127
texture_mean	Real	0	Min 9.710	Max 39.280	Average 19.290
perimeter_mean	Real	0	Min 43.790	Max 188.500	Average 91.969
area_mean	Real	0	Min 143.500	Max 2501	Average 654.889
smoothness_mean	Real	0	Min 0.053	Max 0.163	Average 0.096
compactness_mean	Real	0	Min 0.019	Max 0.345	Average 0.104
concavity_mean	Real	0	Min 0	Max 0.427	Average 0.089
concave points_mean	Real	0	Min 0	Max 0.201	Average 0.049
symmetry_mean	Real	0	Min 0.106	Max 0.304	Average 0.181
fractal_dimension_mean	Real	0	Min 0.050	Max 0.097	Average 0.063

ภาพประกอบ 8 การแสดงรายละเอียดและลักษณะของข้อมูลแต่ละคุณลักษณะในกลุ่มค่าเฉลี่ย

radius_se	Real	0	Min 0.112	Max 2.873	Average 0.405
texture_se	Real	0	Min 0.360	Max 4.885	Average 1.217
perimeter_se	Real	0	Min 0.757	Max 21.980	Average 2.866
area_se	Real	0	Min 6.802	Max 542.200	Average 40.337
smoothness_se	Real	0	Min 0.002	Max 0.031	Average 0.007
compactness_se	Real	0	Min 0.002	Max 0.135	Average 0.025
concavity_se	Real	0	Min 0	Max 0.396	Average 0.032
concave points_se	Real	0	Min 0	Max 0.053	Average 0.012
symmetry_se	Real	0	Min 0.008	Max 0.079	Average 0.021
fractal_dimension_se	Real	0	Min 0.001	Max 0.030	Average 0.004

ภาพประกอบ 9 การแสดงรายละเอียดและลักษณะของข้อมูลแต่ละคุณลักษณะในกลุ่มความคลาดเคลื่อนมาตรฐาน

radius_worst	Real	0	Min 7.930	Max 36.040	Average 16.269
texture_worst	Real	0	Min 12.020	Max 49.540	Average 25.677
perimeter_worst	Real	0	Min 50.410	Max 251.200	Average 107.261
area_worst	Real	0	Min 185.200	Max 4254	Average 880.583
smoothness_worst	Real	0	Min 0.071	Max 0.223	Average 0.132
compactness_worst	Real	0	Min 0.027	Max 1.058	Average 0.254
concavity_worst	Real	0	Min 0	Max 1.252	Average 0.272
concave points_worst	Real	0	Min 0	Max 0.291	Average 0.115
symmetry_worst	Real	0	Min 0.157	Max 0.664	Average 0.290
fractal_dimension_worst	Real	0	Min 0.055	Max 0.207	Average 0.084

ภาพประกอบ 10 การแสดงรายละเอียดและลักษณะของข้อมูลแต่ละคุณลักษณะในกลุ่มค่าที่แย่ที่สุด

3.1.2 การวิเคราะห์ข้อมูลด้วยโปรแกรม RapidMiner

ทำการทดลองโดยใช้เทคนิคต่างๆเพื่อเปรียบเทียบประสิทธิภาพในการคัดเลือกคุณลักษณะที่สำคัญในการพยากรณ์มะเร็งเต้านม และใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) ในการวัดประสิทธิภาพการจำแนกข้อมูลของการเป็นมะเร็งเต้านม ซึ่งมีเทคนิคที่ใช้คัดเลือกคุณลักษณะในการทดลองมีดังนี้

1. เทคนิค Correlation Based Feature Selection (CFS)
2. เทคนิค Information Gain (IG)
3. เทคนิค Gain Ratio (GR)
4. เทคนิค Chi-Square
5. เทคนิค Forward Selection
6. เทคนิค Backward Elimination
7. เทคนิค Evolutionary Selection

3.2 ผลการทดลอง

1. เทคนิค Correlation Based Feature Selection (CFS)

ในการใช้เทคนิค Correlation Based Feature Selection จากคุณลักษณะทั้งหมด 30 คุณลักษณะ ทำให้สามารถเลือกคุณลักษณะที่สำคัญได้จากผลการทดลองตามตารางที่ 1 ดังต่อไปนี้

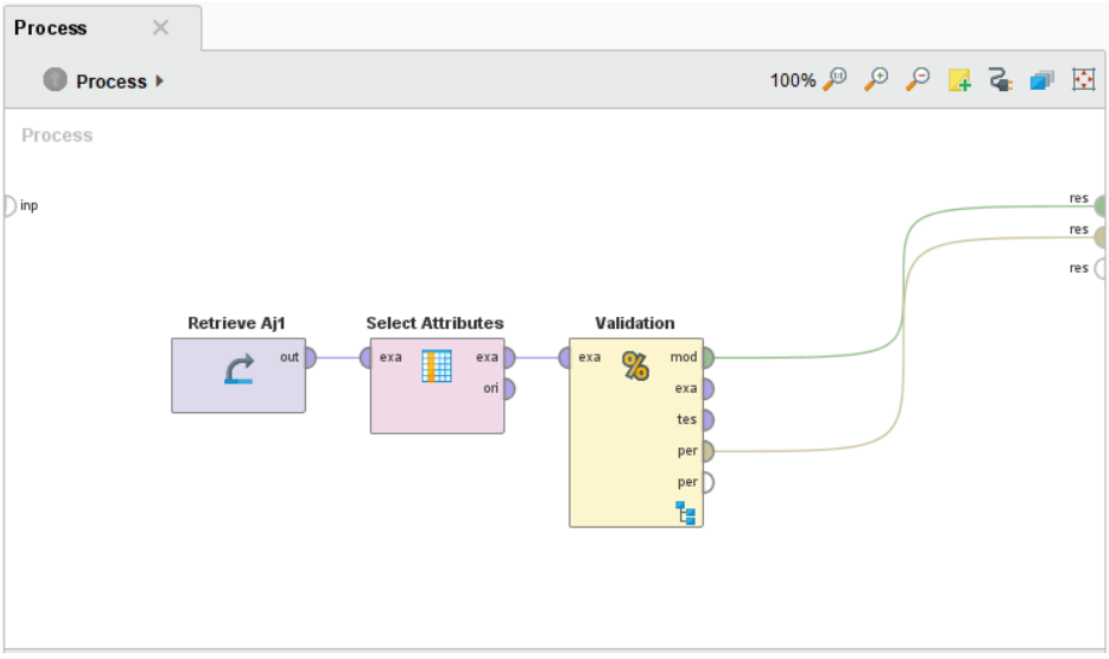
ตารางที่ 1 ค่าน้ำหนักของคุณลักษณะจากการใช้เทคนิค CBF

No	Attribute	Weight
1	radius_mean	0
2	texture_mean	0
3	perimeter_mean	0
4	area_mean	0
5	smoothness_mean	0
6	compactness_mean	0
7	concavity_mean	0
8	concave points_mean	0
9	symmetry_mean	0
10	fractal_dimension_mean	0
11	radius_se	1
12	texture_se	0
13	perimeter_se	0
14	area_se	0
15	smoothness_se	0

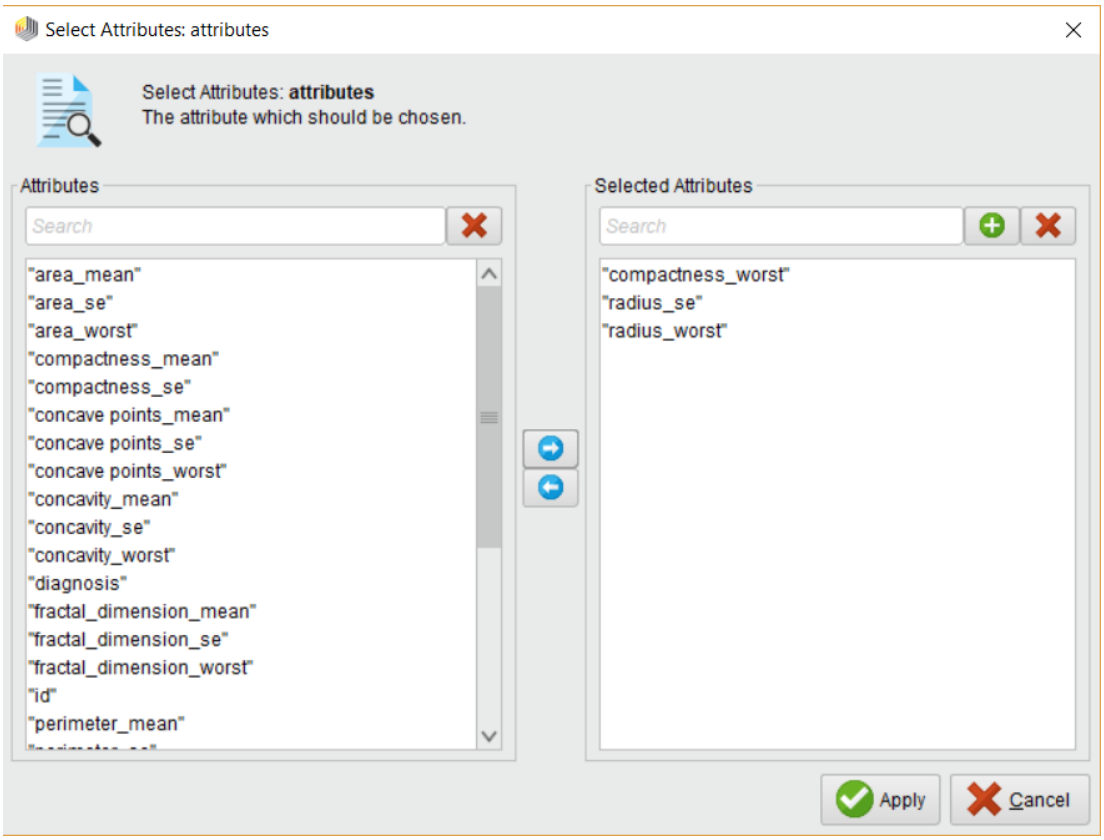
No	Attribute	Weight
16	compactness_se	0
17	concavity_se	0
18	concave points_se	0
19	symmetry_se	0
20	fractal_dimension_se	0
21	radius_worst	1
22	texture_worst	0
23	perimeter_worst	0
24	area_worst	0
25	smoothness_worst	0
26	compactness_worst	1
27	concavity_worst	0
28	concave points_worst	0
29	symmetry_worst	0
30	fractal_dimension_worst	0

จากตารางที่ 1 พบว่าคุณลักษณะที่มีค่าน้ำหนักเป็น 0 จะไม่ได้รับการคัดเลือก จะทำการเลือกเฉพาะคุณลักษณะที่มีค่าน้ำหนักเท่ากับ 1 เท่านั้น ซึ่งคุณลักษณะที่ถูกเลือกจากเทคนิค Correlation Based Feature Selection ได้แก่ radius_se radius_worst และcompactness_worst

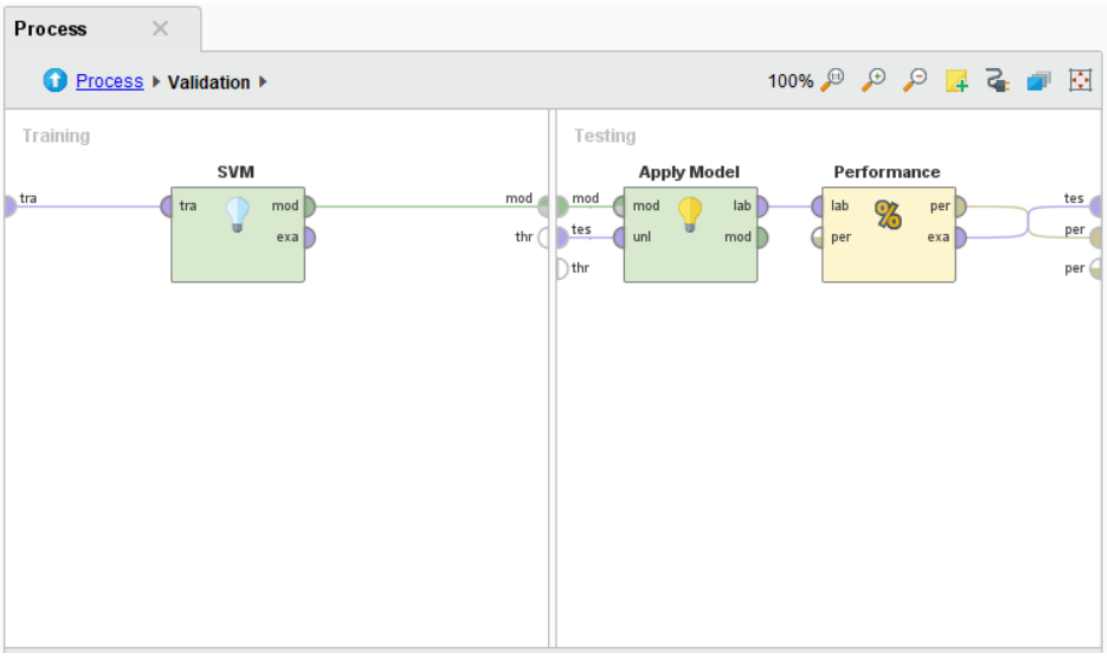
หลังจากนั้นจึงได้นำคุณลักษณะทั้ง 3 มาหาค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมด้วยเทคนิค SVM โดยการใช้โปรแกรม RapidMiner ในการประมวลผลการพยากรณ์ซึ่งประกอบไปด้วยกระบวนการต่างๆ ดังแสดงในภาพประกอบ 11 เริ่มต้นโดยการดึงข้อมูลคุณลักษณะต่างๆด้วยการใช้ Operator Retrieve หลังจากนั้นมีการคัดเลือกคุณลักษณะที่สำคัญที่ได้จากเทคนิค Correlation Based Feature Selection ดังแสดงในภาพประกอบ 12 ต่อจากนั้นเป็นกระบวนการในการทำ Validation การพยากรณ์การเป็นมะเร็งเต้านมด้วยเทคนิค SVM ดังแสดงในภาพประกอบ 13 โดยมีการเซ็ตพารามิเตอร์ให้กับเทคนิค 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 มีการเลือกการทำงานให้เป็นแบบ shrinking และ confidence for multiclass ดังแสดงในภาพประกอบ 14 และมีการกำหนดค่าพารามิเตอร์ในการทำ Cross validation ให้ค่า number of folds เป็น 10 และลักษณะ sampling type เป็นแบบ stratified sampling ดังแสดงในภาพประกอบ 15



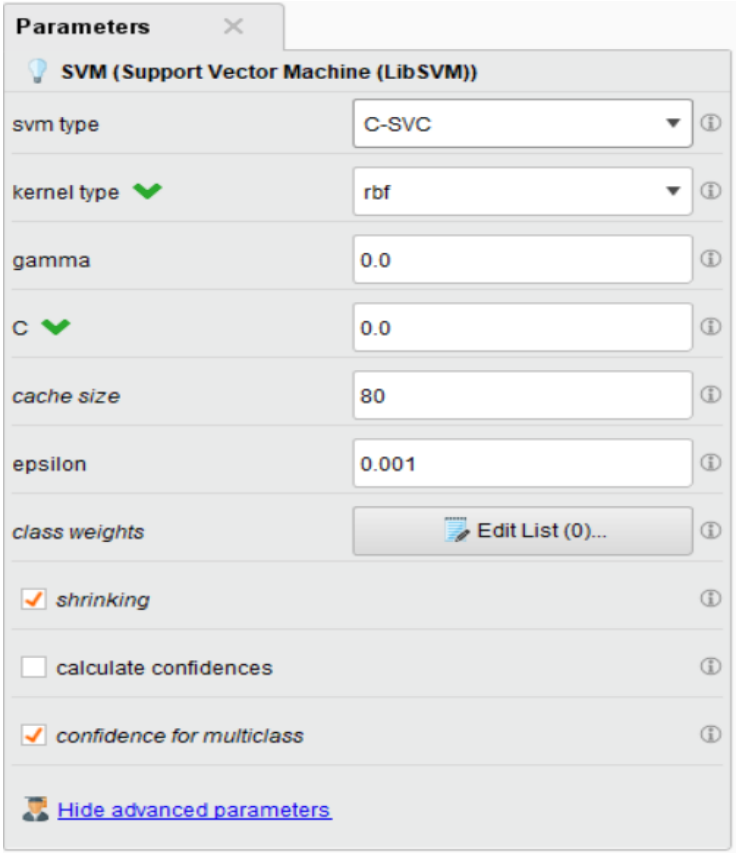
ภาพประกอบ 11 ภาพรวมแสดงการพยากรณ์การเป็นมะเร็งเต้านมด้วยเทคนิค SVM ด้วยโปรแกรม RapidMiner



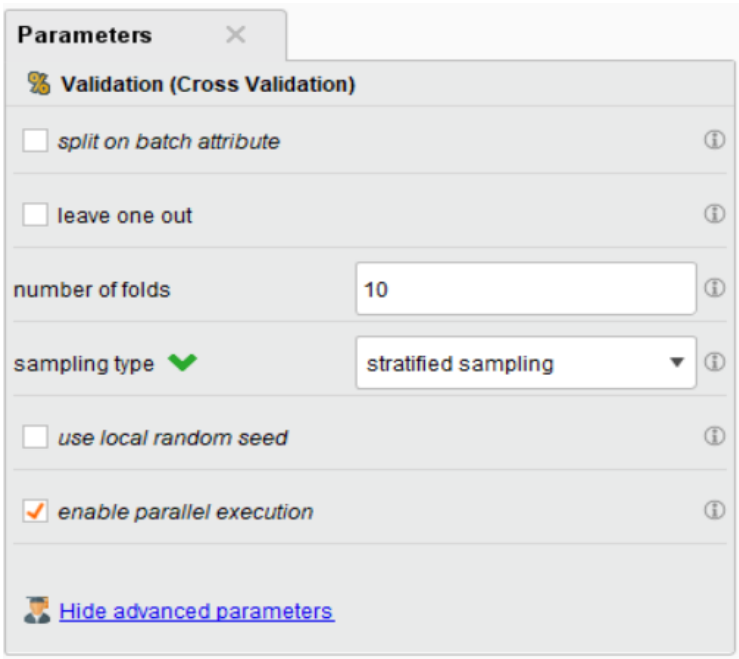
ภาพประกอบ 12 การเลือกคุณลักษณะที่ได้รับการคัดเลือกเข้าประมวลผล



ภาพประกอบ 13 กระบวนการในการทำ Validation การประมวลผลด้วย SVM



ภาพประกอบ 14 กำหนดค่าพารามิเตอร์ให้กับ SVM



ภาพประกอบ 15 กำหนดค่าพารามิเตอร์ในการทำ Cross validation

accuracy: 91.22% +/- 3.99% (mikro: 91.21%)

	true M	true B	class precision
pred. M	168	6	96.55%
pred. B	44	351	88.86%
class recall	79.25%	98.32%	

ภาพประกอบ 16 ผลการวัดประสิทธิภาพจากการเลือกคุณลักษณะของเทคนิค CBF

จากภาพประกอบ 16 แสดงผลการวัดประสิทธิภาพการพยากรณ์โดยการคัดเลือกคุณลักษณะด้วยเทคนิค Correlation Based Feature Selection และการหาค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมด้วยเทคนิค SVM ทำให้ได้ค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมเท่ากับ 91.22%

2. เทคนิค Information Gain (IG)

ในการใช้เทคนิค Information Gain ด้วยคุณลักษณะทั้งหมด 30 คุณลักษณะ ทำให้ได้ค่าน้ำหนักในแต่ละคุณลักษณะจากการทดลองตามตารางที่ 2 ดังต่อไปนี้

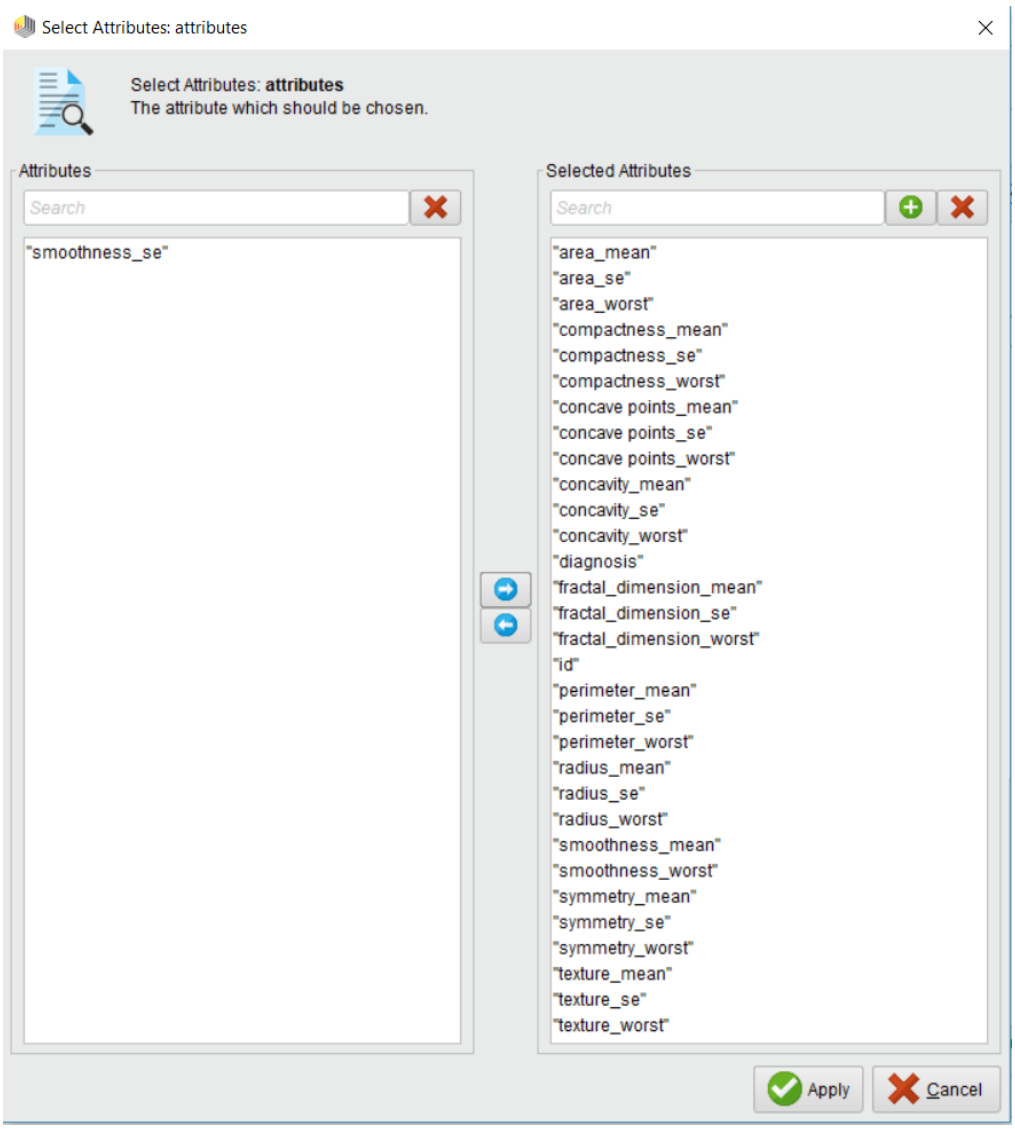
ตารางที่ 2 ค่าน้ำหนักของคุณลักษณะจากการใช้เทคนิค IG

Attribute	Weight
smoothness_se	0.0135
texture_se	0.0163
fractal_dimension_mean	0.0207
symmetry_se	0.0228
fractal_dimension_se	0.0346
symmetry_mean	0.0712
fractal_dimension_worst	0.0747
smoothness_mean	0.0971
compactness_se	0.1098
symmetry_worst	0.1098
smoothness_worst	0.1235
texture_worst	0.1572
texture_mean	0.1593
concave points_se	0.1637
concavity_se	0.1934

Attribute	Weight
compactness_worst	0.2412
compactness_mean	0.2671
radius_se	0.2869
perimeter_se	0.2887
concavity_worst	0.4017
area_se	0.4161
concavity_mean	0.4482
radius_mean	0.4630
perimeter_mean	0.4666
area_mean	0.4748
concave points_mean	0.5458
concave points_worst	0.5491
area_worst	0.5602
radius_worst	0.5619
perimeter_worst	0.5620

เทคนิค Information Gain จะให้ค่าน้ำหนักของคุณลักษณะที่สำคัญออกมา โดยคุณลักษณะที่มีความสำคัญจะมีค่าน้ำหนักมาก คุณลักษณะที่มีความสำคัญน้อยจะมีค่าน้ำหนักน้อย จากผลการทดลองพบว่า คุณลักษณะ perimeter_worst จะมีความสำคัญมากที่สุดซึ่งมีค่าน้ำหนักที่ 0.5620 และคุณลักษณะ smoothness_se จะมีความสำคัญน้อยที่สุดซึ่งมีค่าน้ำหนักที่ 0.0135

เนื่องจากเทคนิคนี้จะให้ค่าน้ำหนักของแต่ละคุณลักษณะออกมา ดังนั้นการตรวจสอบว่าควรจะใช้คุณลักษณะใดบ้างในการหาค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมโดยใช้เทคนิค SVM ในการทำนายความถูกต้อง จึงได้ทำการทดลองโดยการลดคุณลักษณะไปทีละหนึ่งคุณลักษณะเริ่มจากคุณลักษณะที่มีค่าน้ำหนักน้อยที่สุด เพื่อที่จะได้ค่าความถูกต้องที่ดีที่สุดในการทำนายมะเร็งเต้านม ผลที่ได้แสดงในตารางที่ 3



ภาพประกอบ 17 ตัวอย่างการเลือกคุณลักษณะโดยการตัดคุณลักษณะที่มีค่าน้ำหนักน้อยที่สุดออกของเทคนิค IG

accuracy: 91.39% +/- 3.71% (mikro: 91.39%)

	true M	true B	class precision
pred. M	197	34	85.28%
pred. B	15	323	95.56%
class recall	92.92%	90.48%	

ภาพประกอบ 18 ผลการวัดประสิทธิภาพจากการตัดคุณลักษณะที่มีค่าน้ำหนักน้อยที่สุดออกของเทคนิค IG

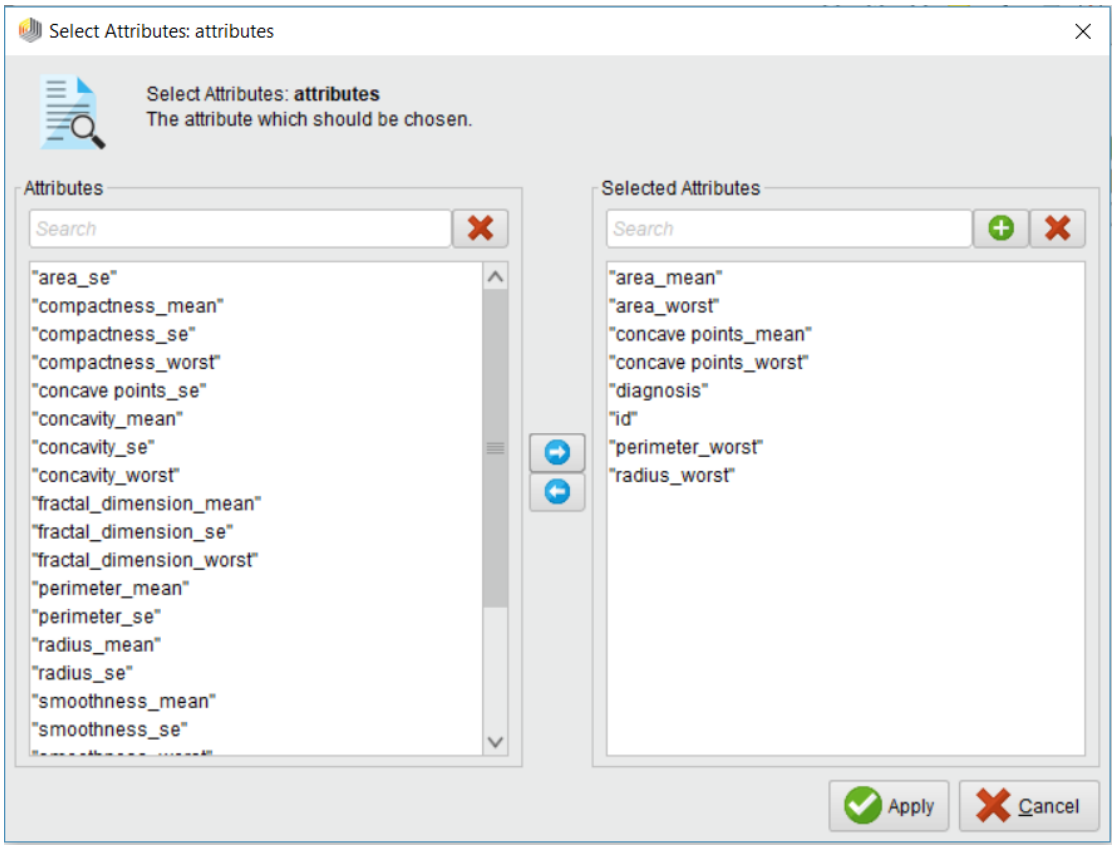
ตารางที่ 3 ค่าความถูกต้องของการพยากรณ์การเป็นมะเร็งเต้านมตามจำนวนคุณลักษณะที่ใช้ด้วยเทคนิค IG

Number of Attribute	Attribute Reduction	Accuracy
30	0	91.39
29	1	91.39
28	2	91.39
27	3	91.39
26	4	91.39
25	5	91.39
24	6	91.39
23	7	91.39
22	8	91.39
21	9	91.39
20	10	91.39
19	11	91.39
18	12	91.39
17	13	91.57
16	14	91.57

Number of Attribute	Attribute Reduction	Accuracy
15	15	91.57
14	16	91.57
13	17	91.57
12	18	91.57
11	19	91.57
10	20	91.57
9	21	92.27
8	22	92.27
7	23	92.27
6	24	92.27
5	25	91.22
4	26	91.22
3	27	91.22
2	28	91.57
1	29	91.74

จากผลการทดลองพบว่าค่าความถูกต้องที่ดีที่สุด โดยมีร้อยละความถูกต้องอยู่ที่ 92.27 ตามตารางที่ 3 จะใช้จำนวนคุณลักษณะอยู่ที่ 6,7,8 หรือ 9 คุณลักษณะ ดังนั้นเราจะเลือกใช้จำนวนคุณลักษณะน้อยสุด 6 คุณลักษณะ ได้แก่

1. area_mean
2. area_worst
3. concave points_mean
4. concave points_worst
5. perimeter_worst
6. radius_worst



ภาพประกอบ 19 การนำ 6 คุณลักษณะที่ได้รับการคัดเลือกเข้าประมวลผลของเทคนิค IG

accuracy: 92.27% +/- 3.15% (mikro: 92.27%)

	true M	true B	class precision
pred. M	192	24	88.89%
pred. B	20	333	94.33%
class recall	90.57%	93.28%	

ภาพประกอบ 20 ผลการวัดประสิทธิภาพจากการเลือก 6 คุณลักษณะของเทคนิค IG

ผลการวัดประสิทธิภาพการพยากรณ์โดยการคัดเลือกคุณลักษณะด้วยเทคนิค Information Gain (IG) และการหาค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมด้วยเทคนิค SVM จากการใช้ 6 คุณลักษณะที่ได้รับการคัดเลือก ได้ค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมเท่ากับ 92.27%

3. เทคนิค Gain Ratio (GR)

ในการใช้เทคนิค Gain Ratio ด้วยคุณลักษณะทั้งหมด 30 คุณลักษณะ ทำให้ได้ค่าน้ำหนักในแต่ละคุณลักษณะจากการทดลองตามตารางที่ 4 ดังต่อไปนี้

ตารางที่ 4 ค่าน้ำหนักของคุณลักษณะจากการใช้เทคนิค GR

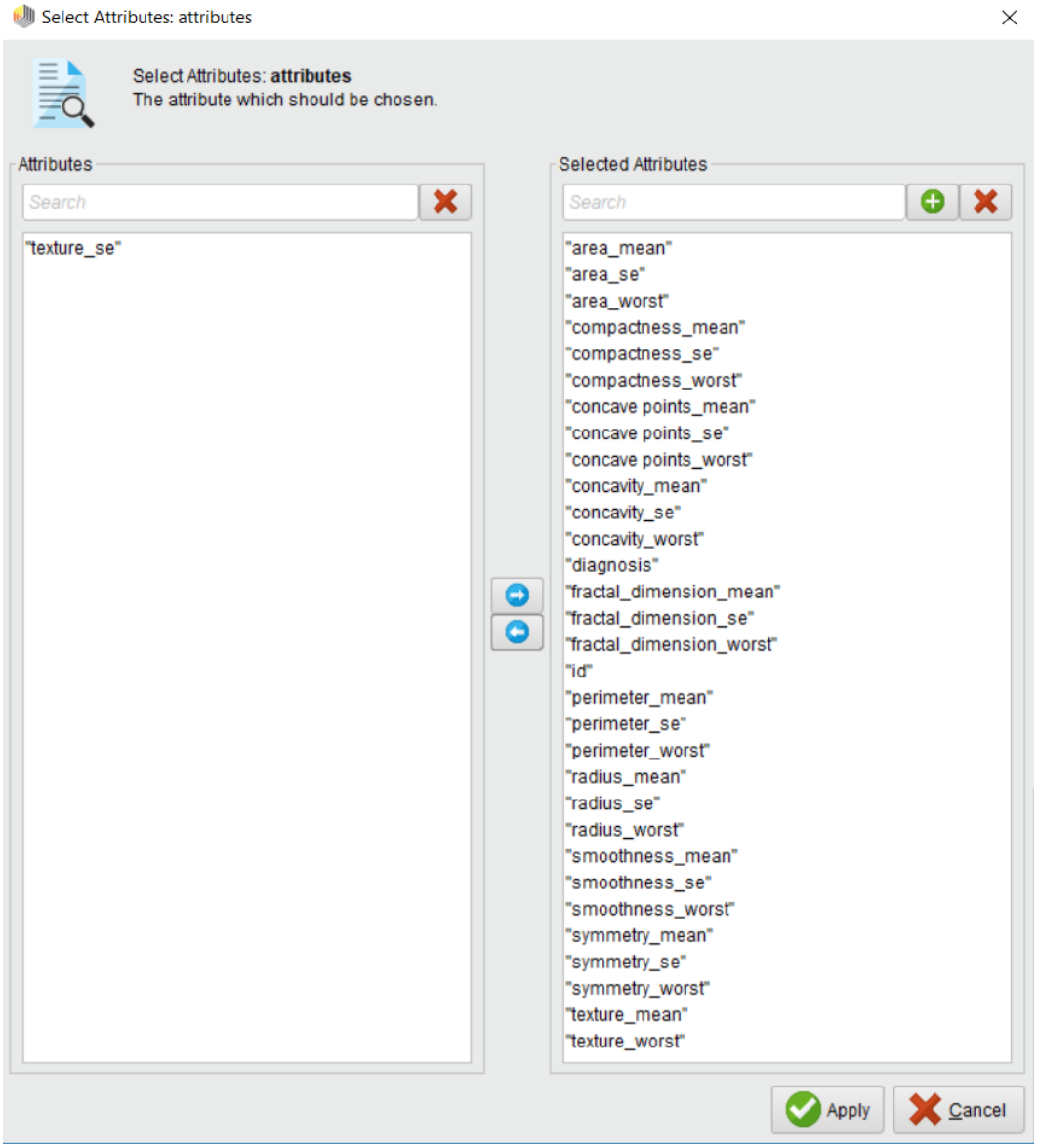
No	Attribute	Weight	No	Attribute	Weight
1	texture_se	0.0749	16	compactness_mean	0.2695
2	fractal_dimension_se	0.0818	17	compactness_worst	0.2807
3	smoothness_mean	0.1166	18	perimeter_se	0.3336
4	compactness_se	0.1349	19	radius_se	0.3604
5	symmetry_se	0.1450	20	concavity_worst	0.4046
6	symmetry_mean	0.1489	21	area_se	0.4708
7	smoothness_se	0.1489	22	concavity_mean	0.4752
8	fractal_dimension_worst	0.1489	23	radius_mean	0.5237
9	smoothness_worst	0.1589	24	perimeter_mean	0.5340
10	texture_mean	0.1601	25	area_mean	0.5371
11	concave points_se	0.1753	26	concave points_mean	0.5670
12	texture_worst	0.1796	27	concave points_worst	0.6035
13	fractal_dimension_mean	0.1901	28	radius_worst	0.6116
14	concavity_se	0.1968	29	area_worst	0.6182
15	symmetry_worst	0.2382	30	perimeter_worst	0.6384

เทคนิค Gain Ratio จะให้ค่าน้ำหนักของคุณลักษณะที่สำคัญออกมา โดยคุณลักษณะที่มีความสำคัญจะมีค่าน้ำหนักมาก คุณลักษณะที่มีความสำคัญน้อยจะมีน้ำหนักน้อย จากผลการทดลองพบว่า คุณลักษณะ perimeter_worst จะมีความสำคัญมากที่สุดซึ่งมีค่าน้ำหนักที่ 0.6384 และคุณลักษณะ texture_se จะมีความสำคัญน้อยที่สุดซึ่งมีค่าน้ำหนักที่ 0.0749

เนื่องจากเทคนิคนี้จะให้ค่าน้ำหนักของแต่ละคุณลักษณะออกมา ดังนั้นการตรวจสอบว่าควรจะใช้คุณลักษณะใดบ้างในการหาค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมโดยใช้เทคนิค SVM จึงได้ทำการทดลองโดยการลดคุณลักษณะไปที่ละหนึ่งคุณลักษณะเริ่มจากคุณลักษณะที่มีค่าน้ำหนักน้อยที่สุด เพื่อที่จะได้ค่าความถูกต้องที่ดีที่สุดในการพยากรณ์การเป็นมะเร็งเต้านม ผลที่ได้แสดงในตารางที่ 5

ตารางที่ 5 ค่าความถูกต้องของการพยากรณ์การเป็นมะเร็งเต้านมตามจำนวนคุณลักษณะที่ใช้ด้วยเทคนิค GR

Number of Attribute	Attribute Reduction	Accuracy	Number of Attribute	Attribute Reduction	Accuracy
30	0	91.39	15	15	91.57
29	1	91.39	14	16	91.57
28	2	91.39	13	17	91.57
27	3	91.39	12	18	91.57
26	4	91.39	11	19	91.57
25	5	91.39	10	20	91.57
24	6	91.39	9	21	92.27
23	7	91.39	8	22	92.27
22	8	91.39	7	23	92.27
21	9	91.39	6	24	92.27
20	10	91.39	5	25	91.22
19	11	91.39	4	26	91.22
18	12	91.57	3	27	91.22
17	13	91.57	2	28	91.22
16	14	91.57	1	29	91.74



ภาพประกอบ 21 ตัวอย่างการเลือกคุณลักษณะโดยการตัดคุณลักษณะที่มีค่าน้ำหนักน้อยที่สุดออกของเทคนิค GR

accuracy: 91.39% +/- 3.71% (mikro: 91.39%)

	true M	true B	class precision
pred. M	197	34	85.28%
pred. B	15	323	95.56%
class recall	92.92%	90.48%	

ภาพประกอบ 22 ผลการวัดประสิทธิภาพจากการตัดคุณลักษณะที่มีค่าน้ำหนักน้อยที่สุดออกของเทคนิค GR

จากผลการทดลองพบว่าค่าความถูกต้องที่ดีที่สุด โดยมีร้อยละความถูกต้องอยู่ที่ 92.27 ใช้จำนวนคุณลักษณะ 6,7,8 หรือ 9 คุณลักษณะ ดังนั้นเราจะเลือกใช้จำนวนคุณลักษณะน้อยสุด 6 คุณลักษณะ ได้แก่

1. area_mean
2. area_worst
3. concave points_mean
4. concave points_worst
5. radius_worst
6. perimeter_worst

จะพบว่าผลการทดลองของ Gain Ratio ค่าความถูกต้องที่ดีที่สุดที่ได้มีลักษณะเหมือนกับ Information Gain

4. เทคนิค Chi Squared

ในการใช้เทคนิค Chi Squared ด้วยคุณลักษณะทั้งหมด 30 คุณลักษณะ ทำให้ได้ค่าน้ำหนักในแต่ละคุณลักษณะจากการทดลองตามตารางที่ 6 ดังต่อไปนี้

ตารางที่ 6 ค่าน้ำหนักของคุณลักษณะจากการใช้เทคนิค Chi Squared

No	Attribute	Weight	No	Attribute	Weight
1	texture_se	13.3265	16	area_se	205.2987
2	smoothness_se	13.5522	17	compactness_worst	208.6836
3	fractal_dimension_mean	24.0610	18	perimeter_se	222.0918
4	fractal_dimension_se	27.9417	19	compactness_mean	222.3855
5	symmetry_se	31.5886	20	radius_se	230.0074
6	concavity_se	51.9918	21	concavity_worst	311.1369
7	fractal_dimension_worst	66.7996	22	area_mean	326.2661
8	symmetry_mean	68.6855	23	concavity_mean	341.0650
9	compactness_se	74.0546	24	radius_mean	344.7614
10	smoothness_mean	81.3973	25	perimeter_mean	359.1691
11	symmetry_worst	100.7488	26	area_worst	364.0997
12	smoothness_worst	109.0351	27	concave points_mean	390.5314
13	concave points_se	126.9527	28	radius_worst	395.6819
14	texture_mean	133.6581	29	perimeter_worst	406.0872
15	texture_worst	134.4097	30	concave points_worst	415.0087

เทคนิค Chi Squared จะให้ค่าน้ำหนักของคุณลักษณะที่สำคัญออกมา โดยคุณลักษณะที่มีความสำคัญจะมีค่าน้ำหนักมาก คุณลักษณะที่มีความสำคัญน้อยจะมีน้ำหนักน้อย จากผลการทดลองพบว่า คุณลักษณะ concave points_worst มีความสำคัญมากที่สุดมีค่าน้ำหนักที่ 415.0087 และคุณลักษณะ texture_se จะมีความสำคัญน้อยที่สุดมีค่าน้ำหนักที่ 13.3265

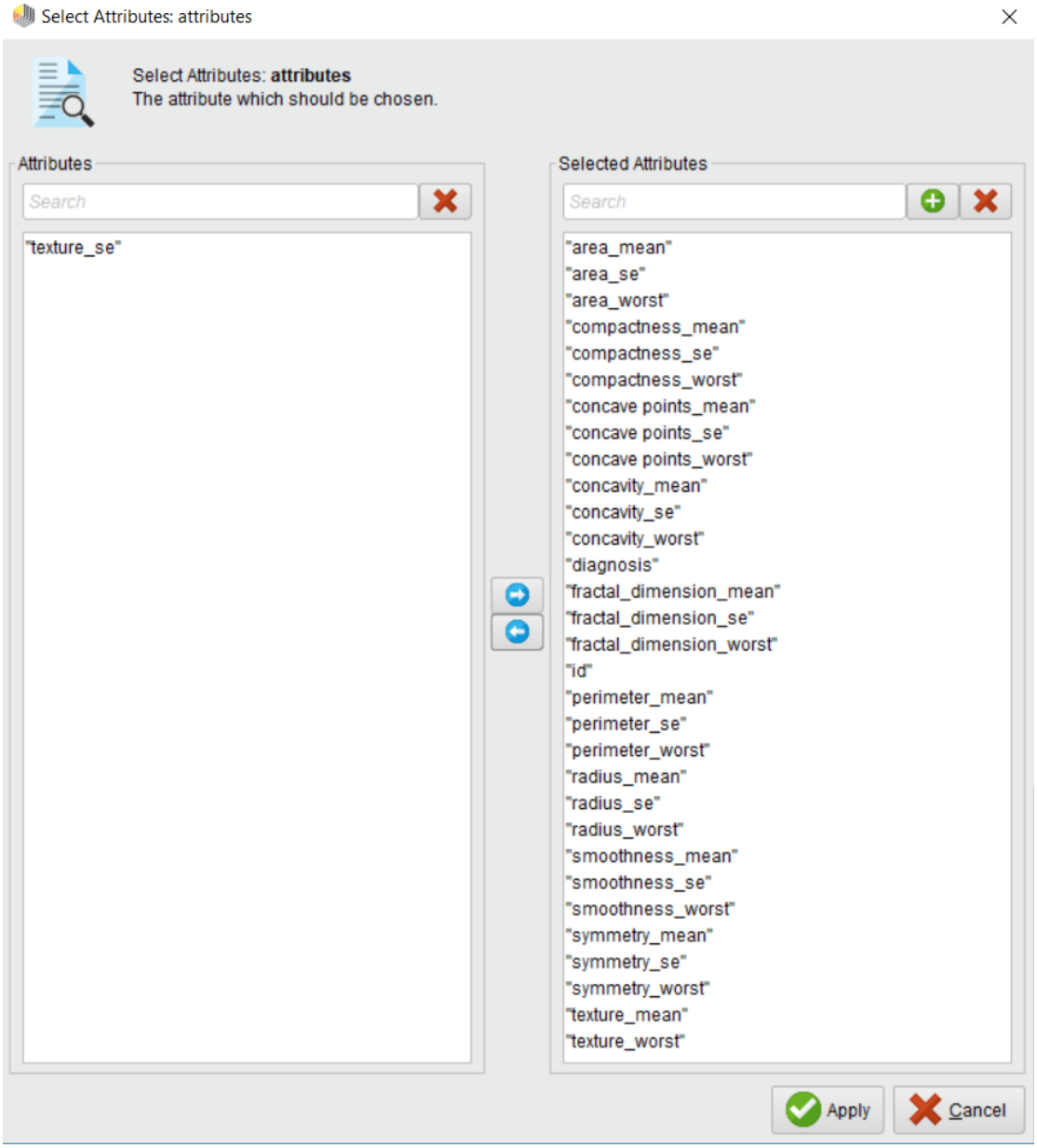
เนื่องจากเทคนิคนี้จะให้ค่าน้ำหนักของแต่ละคุณลักษณะออกมา ดังนั้นการตรวจสอบว่าควรจะใช้คุณลักษณะใดบ้างในการหาค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมโดยใช้เทคนิค SMV จึงได้ทำการทดลองโดยการลดคุณลักษณะไปที่ละหนึ่งคุณลักษณะเริ่มจาก

คุณลักษณะที่มีค่าน้ำหนักน้อยที่สุด เพื่อที่จะได้ค่าความถูกต้องที่ดีที่สุดในการพยากรณ์มะเร็งเต้านม ผลที่ได้แสดงในตารางที่ 7

ตารางที่ 7 ค่าความถูกต้องของการพยากรณ์การเป็นมะเร็งเต้านมตามจำนวนคุณลักษณะที่ใช้ด้วยเทคนิค Chi Squared

Number of Attribute	Attribute Reduction	Accuracy
30	0	91.39
29	1	91.39
28	2	91.39
27	3	91.39
26	4	91.39
25	5	91.39
24	6	91.39
23	7	91.39
22	8	91.39
21	9	91.39
20	10	91.39
19	11	91.39
18	12	91.39
17	13	91.39
16	14	91.39

Number of Attribute	Attribute Reduction	Accuracy
15	15	91.57
14	16	92.27
13	17	92.27
12	18	92.27
11	19	92.27
10	20	92.27
9	21	92.27
8	22	91.57
7	23	91.57
6	24	91.57
5	25	91.22
4	26	91.57
3	27	91.57
2	28	91.74
1	29	62.74



ภาพประกอบ 23 ตัวอย่างการเลือกคุณลักษณะโดยการตัดคุณลักษณะที่มีค่าน้ำหนักน้อยที่สุดออกด้วยเทคนิค Chi Squared

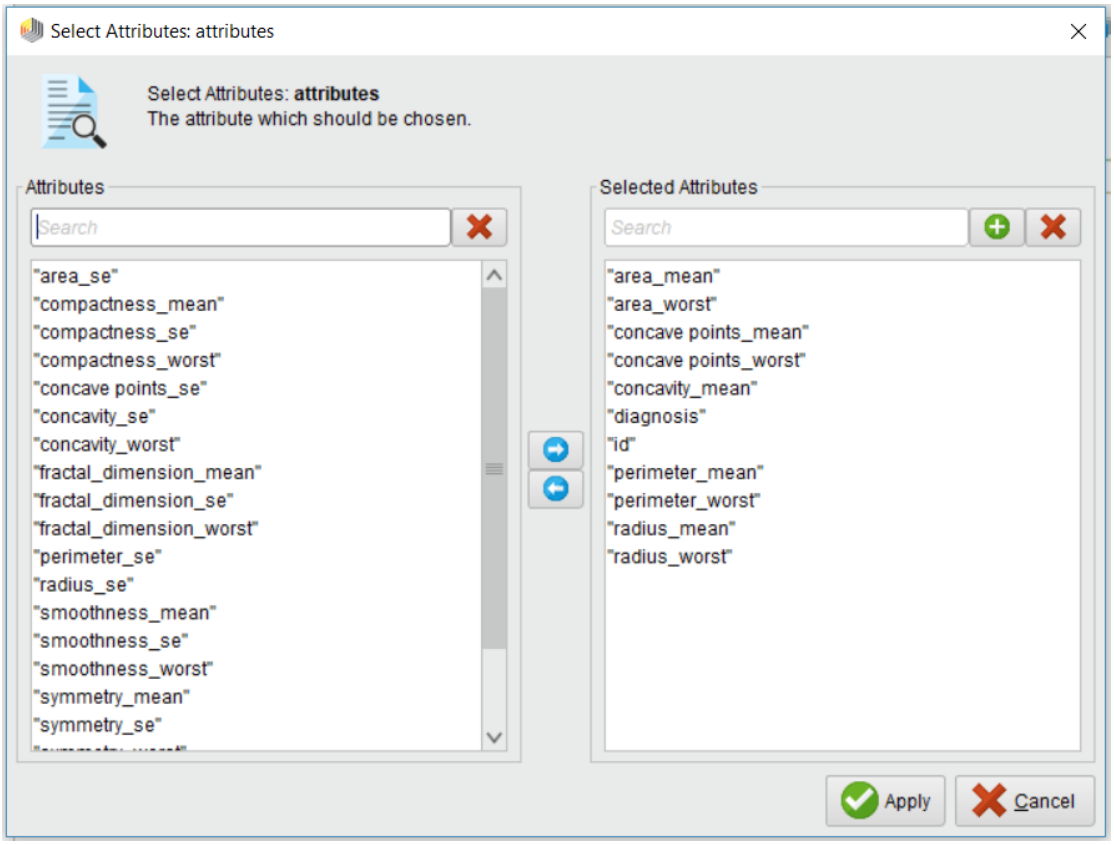
accuracy: 91.39% +/- 3.71% (mikro: 91.39%)

	true M	true B	class precision
pred. M	197	34	85.28%
pred. B	15	323	95.56%
class recall	92.92%	90.48%	

ภาพประกอบ 24 ผลการวัดประสิทธิภาพจากการตัดคุณลักษณะที่มีค่าน้ำหนักน้อยที่สุดออกด้วยเทคนิค Chi Squared

จากผลการทดลองพบว่าค่าความถูกต้องที่ดีที่สุด โดยมีร้อยละความถูกต้องอยู่ที่ 92.27 ใช้จำนวนคุณลักษณะ 9,10,11,12,13 หรือ 14 คุณลักษณะ ดังนั้นเราจะเลือกใช้จำนวนคุณลักษณะน้อยสุด 9 คุณลักษณะ ได้แก่

1. area_mean
2. area_worst
3. concavity_mean
4. concave points_mean
5. concave points_worst
6. perimeter_mean
7. perimeter_worst
8. radius_mean
9. radius_worst



ภาพประกอบ 25 การนำ 9 คุณลักษณะที่ได้รับการคัดเลือกเข้าประมวลผลของ

เทคนิค Chi Squared

accuracy: 92.27% +/- 3.15% (mikro: 92.27%)

	true M	true B	class precision
pred. M	192	24	88.89%
pred. B	20	333	94.33%
class recall	90.57%	93.28%	

ภาพประกอบ 26 ผลการวัดประสิทธิภาพจากการเลือก 9 คุณลักษณะด้วย

เทคนิค Chi Squared

ผลการวัดประสิทธิภาพการพยากรณ์โดยการคัดเลือกคุณลักษณะด้วยเทคนิค Chi Squared และการหาค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมด้วยเทคนิค SVM จากการใช้ 9 คุณลักษณะที่ได้รับการคัดเลือก ให้ค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมเท่ากับ 92.27% ดังแสดงในภาพประกอบ 26

5. เทคนิค Forward Selection

ในการใช้เทคนิค Forward Selection ด้วยคุณลักษณะทั้งหมด 30 คุณลักษณะ ทำให้ได้คุณลักษณะที่ถูกเลือกจากผลการทดลองตามตารางที่ 8 ดังต่อไปนี้

ตารางที่ 8 ค่าน้ำหนักของคุณลักษณะจากการใช้เทคนิค Forward Selection

No	Attribute	Weight	No	Attribute	Weight
1	radius_mean	0	16	compactness_se	0
2	texture_mean	0	17	concavity_se	0
3	perimeter_mean	0	18	concave points_se	0
4	area_mean	0	19	symmetry_se	0
5	smoothness_mean	0	20	fractal_dimension_se	0
6	compactness_mean	0	21	radius_worst	0
7	concavity_mean	0	22	texture_worst	1
8	concave points_mean	0	23	perimeter_worst	1
9	symmetry_mean	0	24	area_worst	0
10	fractal_dimension_mean	0	25	smoothness_worst	0
11	radius_se	0	26	compactness_worst	0
12	texture_se	0	27	concavity_worst	1
13	perimeter_se	0	28	concave points_worst	0
14	area_se	0	29	symmetry_worst	0
15	smoothness_se	0	30	fractal_dimension_worst	0

จากตารางที่ 8 พบว่าคุณลักษณะที่มีค่าน้ำหนักเป็น 0 จะไม่ได้รับการคัดเลือก และจะเลือกคุณลักษณะที่มีค่าน้ำหนักเท่ากับ 1 เท่านั้น ซึ่งคุณลักษณะที่ถูกเลือกจากเทคนิค Forward Selection ได้แก่ texture_worst, perimeter_worst และ concavity_worst หลังจากนั้นจึงได้นำคุณลักษณะทั้ง 3 มาหาค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมด้วยเทคนิค SVM โดยการใช้โปรแกรม RapidMiner ในการประมวลผลการพยากรณ์

accuracy: 95.07% +/- 2.60% (mikro: 95.08%)

	true M	true B	class precision
pred. M	193	9	95.54%
pred. B	19	348	94.82%
class recall	91.04%	97.48%	

ภาพประกอบ 27 ผลการวัดประสิทธิภาพจากการเลือกคุณลักษณะด้วยเทคนิค Forward Selection

จากผลการทดลองด้วย 3 คุณลักษณะที่ถูกเลือกมาโดยเทคนิค Forward Selection ค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านม โดยใช้เทคนิค SVM ในการทำนายความถูกต้องอยู่ที่ร้อยละ 95.07 ดังแสดงในภาพประกอบ 27

6. เทคนิค Backward Elimination

ในการใช้เทคนิค Backward Elimination ด้วยคุณลักษณะทั้งหมด 30 คุณลักษณะ ทำให้ได้คุณลักษณะที่ถูกเลือกจากผลการทดลองตามตารางที่ 9 ดังต่อไปนี้

ตารางที่ 9 ค่าน้ำหนักของคุณลักษณะจากการใช้เทคนิค Backward Elimination

No	Attribute	Weight	No	Attribute	Weight
1	radius_mean	1	16	compactness_se	1
2	texture_mean	1	17	concavity_se	1
3	perimeter_mean	1	18	concave points_se	0
4	area_mean	0	19	symmetry_se	1
5	smoothness_mean	1	20	fractal_dimension_se	1
6	compactness_mean	1	21	radius_worst	1
7	concavity_mean	1	22	texture_worst	1
8	concave points_mean	1	23	perimeter_worst	1
9	symmetry_mean	1	24	area_worst	0
10	fractal_dimension_mean	1	25	smoothness_worst	1
11	radius_se	1	26	compactness_worst	1
12	texture_se	1	27	concavity_worst	1
13	perimeter_se	1	28	concave points_worst	1
14	area_se	0	29	symmetry_worst	1
15	smoothness_se	1	30	fractal_dimension_worst	1

จากผลการทดลองพบว่ามี 26 คุณลักษณะที่ถูกเลือกมาโดยเทคนิค Backward Elimination ซึ่งมีคุณลักษณะที่ไม่ถูกเลือกเพียง 4 คุณลักษณะคือ area_mean, area_se, concave points_se และ area_worst หลังจากนั้นจึงได้นำคุณลักษณะทั้ง 26 คุณลักษณะมาหาค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมด้วยเทคนิค SVM โดยการใช้โปรแกรม RapidMiner ในการประมวลผลการพยากรณ์

accuracy: 95.08% +/- 2.32% (mikro: 95.08%)

	true M	true B	class precision
pred. M	193	9	95.54%
pred. B	19	348	94.82%
class recall	91.04%	97.48%	

ภาพประกอบ 28 ผลการวัดประสิทธิภาพจากการเลือกคุณลักษณะด้วยเทคนิค Backward Elimination

ผลการทดลองด้วย 26 คุณลักษณะที่ถูกเลือกมาโดยเทคนิค Backward Elimination ค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านม โดยใช้เทคนิค SVM ในการทำนายความถูกต้องอยู่ที่ร้อยละ 95.08 ดังแสดงในภาพประกอบ 28

7. เทคนิค Evolutionary Selection

ในการใช้เทคนิค Evolutionary Selection ด้วยคุณลักษณะทั้งหมด 30 คุณลักษณะ ทำให้ได้คุณลักษณะที่ถูกเลือกจากผลการทดลองตามตารางที่ 10 ดังต่อไปนี้

ตารางที่ 10 ค่าน้ำหนักของคุณลักษณะจากการใช้เทคนิค Evolutionary Selection

No	Attribute	Weight
1	radius_mean	0
2	texture_mean	1
3	perimeter_mean	1
4	area_mean	0
5	smoothness_mean	1
6	compactness_mean	0
7	concavity_mean	1
8	concave points_mean	1
9	symmetry_mean	1
10	fractal_dimension_mean	1
11	radius_se	0
12	texture_se	1
13	perimeter_se	1
14	area_se	0
15	smoothness_se	0

No	Attribute	Weight
16	compactness_se	1
17	concavity_se	0
18	concave points_se	0
19	symmetry_se	0
20	fractal_dimension_se	1
21	radius_worst	1
22	texture_worst	1
23	perimeter_worst	1
24	area_worst	0
25	smoothness_worst	0
26	compactness_worst	1
27	concavity_worst	0
28	concave points_worst	0
29	symmetry_worst	1
30	fractal_dimension_worst	0

จากตารางที่ 10 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 และsymmetry_worst หลังจากนั้นจึงได้นำคุณลักษณะทั้ง 16 คุณลักษณะมาหาค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมด้วยเทคนิค SVM โดยการใช้โปรแกรม RapidMiner ในการประมวลผลการพยากรณ์

accuracy: 95.26% +/- 1.36% (mikro: 95.25%)

	true M	true B	class precision
pred. M	194	9	95.57%
pred. B	18	348	95.08%
class recall	91.51%	97.48%	

ภาพประกอบ 29 ผลการวัดประสิทธิภาพจากการเลือกคุณลักษณะด้วย

เทคนิค Evolutionary Selection

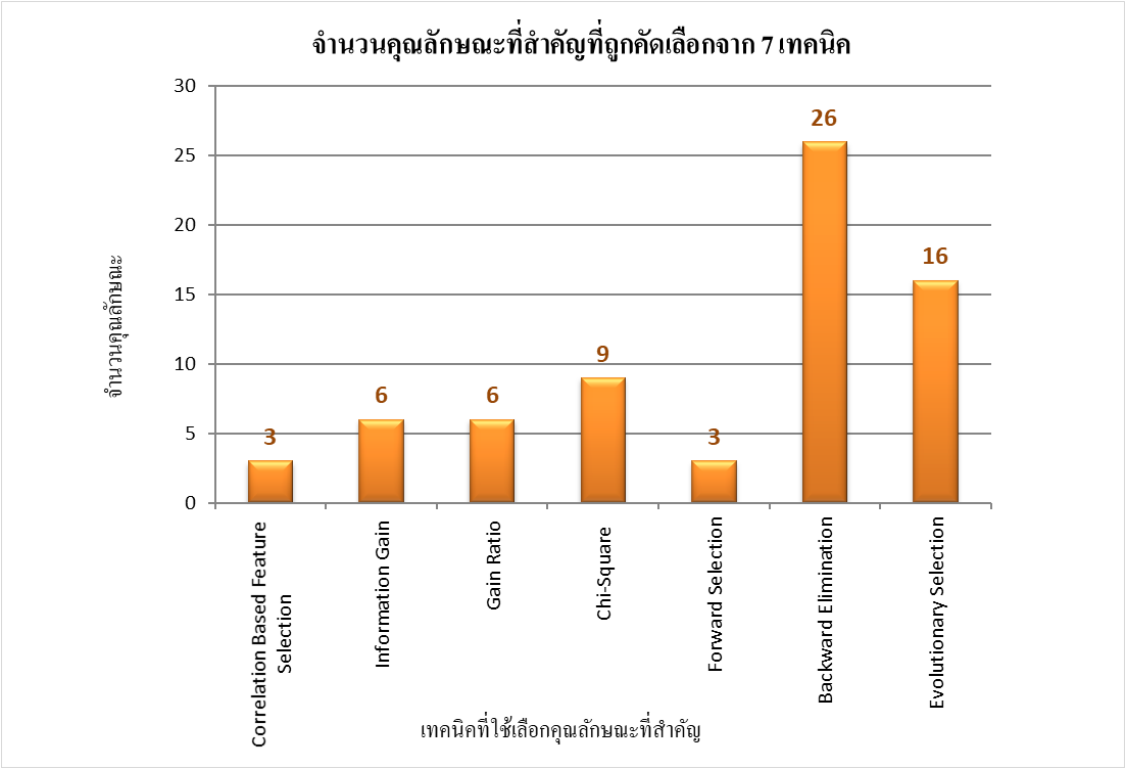
จากผลการวัดประสิทธิภาพการพยากรณ์โดยการคัดเลือกคุณลักษณะ 16 คุณลักษณะด้วยเทคนิค Evolutionary Selection และการหาค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมด้วยเทคนิค SVM จะได้ค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมเท่ากับ 95.26% แสดงในภาพประกอบ 29

บทที่ 4

ผลการวิเคราะห์ข้อมูล

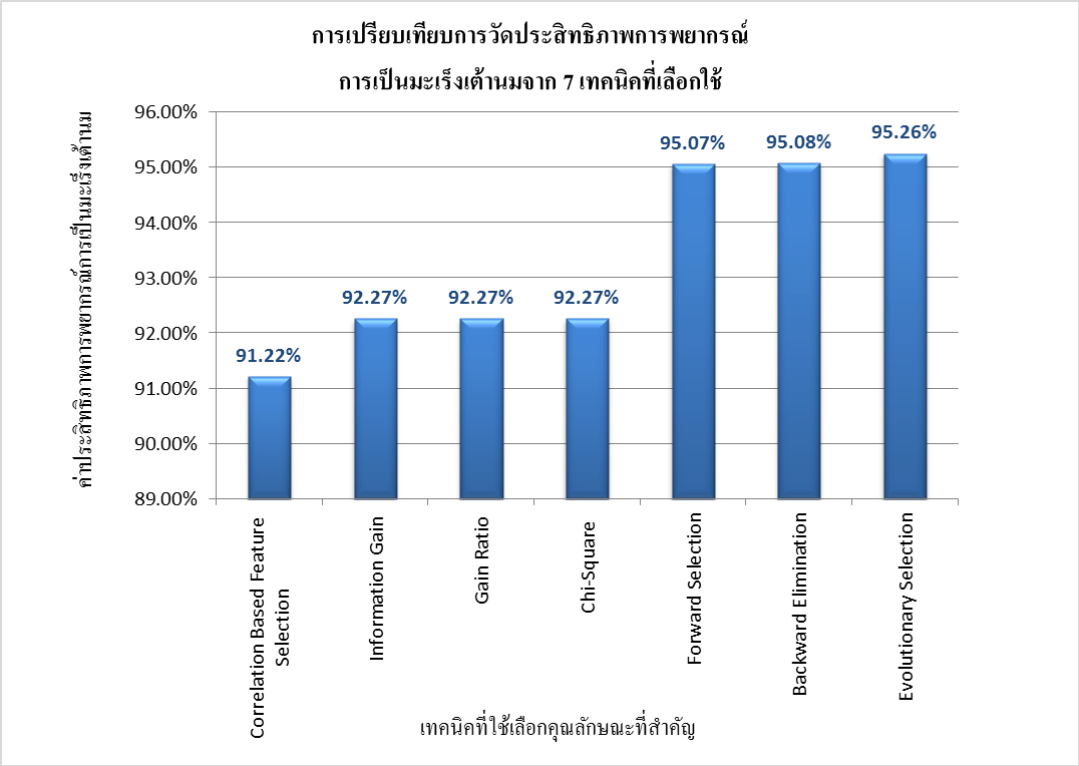
4.1 ผลการวิเคราะห์ข้อมูล

จากผลการทดลองเพื่อคัดเลือกคุณลักษณะที่เหมาะสมเพื่อปรับปรุงการพยากรณ์การเป็นมะเร็งเต้านมจากการใช้ข้อมูลผู้ทดลอง 569 คน จำนวนคุณลักษณะของข้อมูล 32 คุณลักษณะ โดย 2 คุณลักษณะแรก คือ : หมายเลขประจำตัว และผลการวินิจฉัย อีก 30 คุณลักษณะที่เหลือเป็นคุณลักษณะทางด้านกายภาพ จากนั้นใช้เทคนิคต่างๆในการคัดเลือกคุณลักษณะที่สำคัญจำนวน 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 เพื่อนำมาเปรียบเทียบการวัดประสิทธิภาพในการทำนายมะเร็งเต้านม สรุปผลการทดลองดังแสดงในภาพประกอบ 30 และ ภาพประกอบ 31



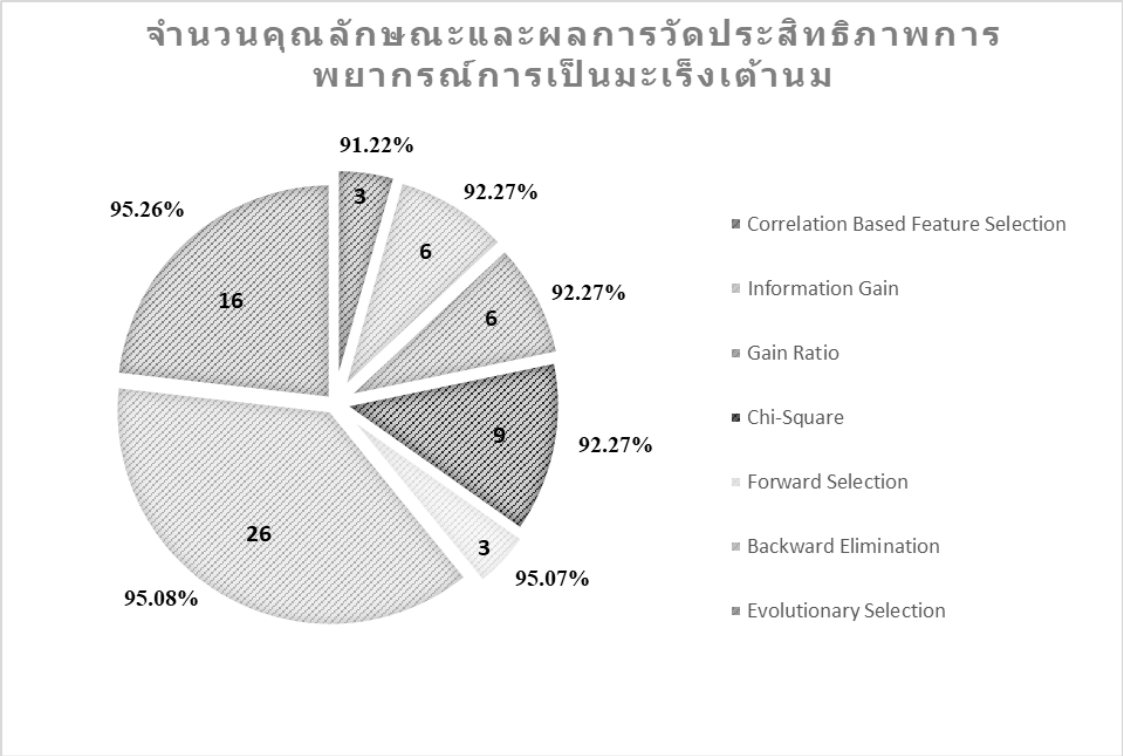
ภาพประกอบ 30 แสดงจำนวนคุณลักษณะที่สำคัญที่ถูกคัดเลือกในแต่ละเทคนิค 7 เทคนิค

จากภาพประกอบ 30 แสดงการเปรียบเทียบจำนวนคุณลักษณะจากการคัดเลือกด้วยเทคนิคต่างๆทั้ง 7 เทคนิค พบว่าเทคนิคที่คัดเลือกจำนวนคุณลักษณะน้อยที่สุดเพียง 3 คุณลักษณะมีอยู่ 2 เทคนิค ได้แก่ เทคนิค Correlation Based Feature Selection และ เทคนิค Forward Selection เทคนิคที่คัดเลือกคุณลักษณะที่สำคัญออกมาจำนวนคุณลักษณะน้อยรองลงมา 6 คุณลักษณะจะมีอยู่ 2 เทคนิค คือ เทคนิค Information Gain และเทคนิค Gain Ratio ถัดมาเป็นเทคนิค Chi-Square ที่คัดเลือกคุณลักษณะที่สำคัญออกมาจำนวนคุณลักษณะ 9 คุณลักษณะ และเทคนิค Evolutionary Selection สามารถคัดเลือกคุณลักษณะที่สำคัญออกมาจำนวนคุณลักษณะ 16 คุณลักษณะ สำหรับเทคนิคที่คัดเลือกคุณลักษณะที่สำคัญออกมาจำนวนคุณลักษณะมากที่สุด 26 คุณลักษณะได้แก่ Backward Elimination



ภาพประกอบ 31 แสดงการเปรียบเทียบการวัดประสิทธิภาพการพยากรณ์การเป็นมะเร็งเต้านมจาก 7 เทคนิคที่เลือกใช้

จากภาพประกอบ 31 แสดงการเปรียบเทียบการวัดประสิทธิภาพการพยากรณ์การเป็นมะเร็งเต้านมจากเทคนิคต่างๆจำนวน 7 เทคนิค โดยจากผลการเปรียบเทียบเทคนิคที่วัดประสิทธิภาพได้สูงสุดคือเทคนิค Evolutionary Selection ซึ่งวัดค่าประสิทธิภาพได้ถึง 95.26% รองลงมาได้แก่ เทคนิค Backward Elimination วัดค่าประสิทธิภาพได้ 95.08% เทคนิค Forward Selection วัดค่าประสิทธิภาพได้ 95.07% เทคนิค Information Gain เทคนิค Gain Ratio และเทคนิค Chi-Square ทั้ง 3 เทคนิควัดค่าประสิทธิภาพได้ 92.27% เท่ากัน สำหรับเทคนิค Correlation Based Feature Selection วัดค่าประสิทธิภาพได้ต่ำสุดเพียง 91.22%



ภาพประกอบ 32 แสดงจำนวนคุณลักษณะและผลการวัดประสิทธิภาพการพยากรณ์การเป็นมะเร็งเต้านมของเทคนิค 7 เทคนิค

จากภาพประกอบ 32 แสดงจำนวนคุณลักษณะและผลการวัดค่าความถูกต้องการพยากรณ์การเป็นมะเร็งเต้านมจากเทคนิคต่างๆที่เลือกจำนวน 7 เทคนิค โดยเทคนิค Correlation Based Feature Selection คัดเลือกคุณลักษณะที่สำคัญออกมาจำนวน 3 คุณลักษณะและนำคุณลักษณะที่เลือกมาวัดค่าความถูกต้องการพยากรณ์ได้ 91.22% เทคนิค Information Gain คัดเลือกคุณลักษณะที่สำคัญออกมาจำนวน 6 คุณลักษณะและนำคุณลักษณะที่เลือกมาวัดค่าความถูกต้องการพยากรณ์ได้ 92.27% เทคนิค Gain Ratio คัดเลือกคุณลักษณะที่สำคัญออกมาจำนวน 6 คุณลักษณะและนำคุณลักษณะที่เลือกมาวัดค่าความถูกต้องการพยากรณ์ได้ 92.27% เทคนิค Chi-Square คัดเลือกคุณลักษณะที่สำคัญออกมาจำนวน 9 คุณลักษณะและนำคุณลักษณะที่เลือกมาวัดค่าความถูกต้องการพยากรณ์ได้ 92.27% เทคนิค Forward Selection คัดเลือกคุณลักษณะที่สำคัญออกมาจำนวน 3 คุณลักษณะและนำคุณลักษณะที่เลือกมาวัดค่าความถูกต้องการพยากรณ์ได้ 95.07% เทคนิค Backward Elimination คัดเลือกคุณลักษณะที่สำคัญออกมาจำนวน 26 คุณลักษณะและนำคุณลักษณะที่เลือกมาวัดค่าความถูกต้องการพยากรณ์ได้ 95.08% และเทคนิค Evolutionary Selection คัดเลือกคุณลักษณะที่สำคัญออกมาจำนวน 16 คุณลักษณะและนำคุณลักษณะที่เลือกมาวัดค่าความถูกต้องการพยากรณ์ได้ 95.26%

บทที่ 5
สรุป อภิปราย และข้อเสนอแนะ

5.1 สรุปผลการวิจัย

บทความวิจัยนี้ได้ศึกษาเกี่ยวกับการเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์มะเร็งเต้านม โดยใช้วิธีการคัดเลือกคุณลักษณะที่สำคัญจากเทคนิคต่างๆ จำนวน 7 เทคนิค ได้แก่ เทคนิค Correlation Based Feature Selection เทคนิค Information Gain (IG) เทคนิค Gain Ratio (GR) เทคนิค Chi-Square เทคนิค Forward Selection เทคนิค Backward Elimination และเทคนิค Evolutionary Selection หลังจากนั้นนำผลการคัดเลือกในแต่ละเทคนิคมาคำนวณหาค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมด้วยเทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

จากผลการทดลองในการใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนวัดค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมกับจำนวนคุณลักษณะของข้อมูลที่ได้จากการเก็บมาโดยไม่มีการลดคุณลักษณะเลยจำนวน 30 คุณลักษณะ ได้ค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมเท่ากับ 91.39% และจะเห็นว่าเมื่อใช้เทคนิคต่างๆในการคัดเลือกคุณลักษณะที่สำคัญนำมาวัดค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านมด้วยเทคนิคซัพพอร์ตเวกเตอร์แมชชีนสามารถลดจำนวนคุณลักษณะและเพิ่มประสิทธิภาพในการจำแนกข้อมูลดังนี้ เทคนิค Correlation Based Feature Selection ลดคุณลักษณะที่สำคัญออกมาเหลือเพียงจำนวน 3 คุณลักษณะและให้ผลการวัดค่าความถูกต้องของการพยากรณ์ได้ 91.22% เทคนิค Information Gain ลดคุณลักษณะที่สำคัญออกมาเหลือเพียงจำนวน 9 คุณลักษณะและให้ผลการวัดค่าความถูกต้องของการพยากรณ์ได้ 92.27% เทคนิค Gain Ratio ลดคุณลักษณะที่สำคัญออกมาเหลือเพียงจำนวน 9 คุณลักษณะและให้ผลการวัดค่าความถูกต้องของการพยากรณ์ได้ 92.27% เทคนิค Chi-Square ลดคุณลักษณะที่สำคัญออกมาเหลือเพียงจำนวน 14 คุณลักษณะและให้ผลการวัดค่าความถูกต้องของการพยากรณ์ได้ 92.27% เทคนิค Forward Selection ลดคุณลักษณะที่สำคัญออกมาเหลือเพียงจำนวน 3 คุณลักษณะและให้ผลการวัดค่าความถูกต้องของการพยากรณ์ได้ 95.07% เทคนิค Backward Elimination ลดคุณลักษณะที่สำคัญออกมาเหลือเพียงจำนวน 26 คุณลักษณะและให้ผลการวัดค่าความถูกต้องของการพยากรณ์ได้ 95.08% และเทคนิค Evolutionary Selection สามารถลดคุณลักษณะที่สำคัญ

ออกมาเหลือเพียงจำนวน 16 คุณลักษณะและให้ผลการวัดค่าความถูกต้องของการพยากรณ์ได้ 95.26%

5.2 อภิปรายผล

การวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาการคัดเลือกคุณลักษณะที่สำคัญที่ใช้ในการวิเคราะห์ข้อมูลเพื่อปรับปรุงการพยากรณ์การเป็นมะเร็งเต้านม โดยมีสมมุติฐานการวิจัยว่า การคัดเลือกคุณลักษณะที่สำคัญจะทำให้จำนวนมิติของข้อมูลที่ใช้ในการวิเคราะห์ข้อมูลลดลง และน่าจะมีผลในการเพิ่มประสิทธิภาพการวัดค่าความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านม จึงได้ทำการศึกษาเปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะที่สำคัญจำนวน 7 เทคนิคได้แก่ เทคนิค Correlation Based Feature Selection เทคนิค Information Gain (IG) เทคนิค Gain Ratio (GR) เทคนิค Chi-Square เทคนิค Forward Selection เทคนิค Backward Elimination และ เทคนิค Evolutionary Selection และใช้ข้อมูลในการวิเคราะห์จาก UCI Machine Learning Repository เป็นคนที่ยังมีชีวิตจำนวน 569 คน จากผลการทดลองถ้าเราใช้ข้อมูลทั้งหมดโดยไม่มีการลดจำนวนมิติข้อมูลจะได้ค่าความถูกต้องการพยากรณ์การเป็นมะเร็งเต้านมเท่ากับ 91.39% ขณะที่เมื่อใช้เทคนิค Evolutionary Selection สามารถคัดเลือกคุณลักษณะที่สำคัญเพื่อใช้ในการวิเคราะห์ข้อมูลเพียง 16 คุณลักษณะและให้ค่าความถูกต้องการพยากรณ์การเป็นมะเร็งเต้านมได้ดีที่สุดเท่ากับ 95.26% จะเห็นว่าการคัดเลือกคุณลักษณะที่สำคัญด้วยเทคนิค Evolutionary Selection สามารถลดมิติของข้อมูลจาก 30 คุณลักษณะลงมาเหลือเพียง 16 คุณลักษณะ และสามารถเพิ่มความแม่นยำในการวัดค่าความถูกต้องการพยากรณ์การเป็นมะเร็งเต้านมจาก 91.39% มาเป็น 95.26%

5.3 ข้อเสนอแนะ

เนื่องจากเทคนิคที่ใช้ในการวัดค่าความถูกต้องการพยากรณ์การเป็นมะเร็งเต้านมนอกเหนือจากเทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) ยังมีอีกหลายเทคนิค เช่น เทคนิคต้นไม้การตัดสินใจ (Decision Tree) เทคนิคแบบเครือข่ายใยประสาท (Neural Network) หรือเทคนิคแบบเบย์ส (Bayes Classifier) ดังนั้นจึงเป็นสิ่งที่น่าสนใจในการทดลองเพื่อดูว่าเทคนิคต่างๆจะได้ค่าความถูกต้องในการพยากรณ์เป็นอย่างไรเมื่อเทียบกับเทคนิคซัพพอร์ตเวกเตอร์แมชชีน

บรรณานุกรม

บรรณานุกรม

จิรา แก้วสุวรรณ์. 2006. “การตรวจจับและการแก้ไขการวางตัวของภาพโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน.”, วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาเทคโนโลยีคอมพิวเตอร์ คณะครุศาสตร์อุตสาหกรรม สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

ทรงศักดิ์ ภูสีอ่อน. 2554. **การประยุกต์ใช้ SPSS วิเคราะห์ข้อมูลงานวิจัย**. มหาสารคาม : มหาวิทยาลัยมหาสารคาม.

นิธินันท์ มาตา และคณะ. 2558. “การค้นหาปัจจัยเพื่อสร้างโมเดลสำหรับพยากรณ์การควบคุมประตูระบายน้ำ.”, **การประชุมวิชาการระดับประเทศด้านเทคโนโลยีสารสนเทศ ครั้งที่ 7**. ระหว่างวันที่ 29-30 ตุลาคม 2558 ณ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง หน้า 352-357.

นิภาพร ชนะมาร และพรรณี สิทธิเดช. 2557. “การวิเคราะห์ปัจจัยการเรียนรู้ด้วยการคัดเลือกคุณสมบัติและการพยากรณ์.”, **วารสารมหาวิทยาลัยราชภัฏสกลนคร** 6(12), 31-45.

น้ำทิพย์ มากนคร และมาลีรัตน์ โสดานิล. 2557. “การเปรียบเทียบวิธีการเลือกคุณลักษณะที่เหมาะสมเพื่อ การจัดหมวดหมู่เว็บเพจผิดกฎหมายโดยใช้เทคนิคการทำเหมืองข้อมูล.”, **ประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 10**. 8 - 9 พฤษภาคม 2557 ณ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ภูเก็ต, หน้า 168-173.

พฤฒิพงศ์ เพ็งศิริ และคณะ. 2557. “การลดมิติข้อมูลการวิเคราะห์ความสัมพันธ์และการประยุกต์สำหรับวิเคราะห์ข้อมูลพื้นฐานการใช้งานสมาร์ทโฟน.”, **ประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 10**. 8 - 9 พฤษภาคม 2557 ณ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ภูเก็ต, หน้า 528–534.

ภัทราวุฒิ แสงศิริ. 2553. “การคัดแยกประเภทของมะเร็งเม็ดเลือดขาวโดยใช้วิธีการจัดอันดับร่วมกับเทคนิคซัพพอร์ตเวกเตอร์แมชชีน.”, **วารสารวิจัย มข.(บศ.)** 10 (2) : เม.ย.-มิ.ย. 2553, หน้า 10-17.

เอกสิทธิ์ พัชรวงศ์ศักดา. 2557. **การวิเคราะห์ข้อมูลด้วยเทคนิค ดาต้า ไมน์นิ่ง เบื้องต้น (An Introduction to Data Mining Techniques).** กรุงเทพฯ, หน้า 53-57

เอกสิทธิ์ พัชรวงศ์ศักดา. 2559. **Advanced Predictive Modeling with R & RapidMiner Studio 7.** พิมพ์ครั้งที่ 6. กรุงเทพฯ: เอเชีย ดิจิตอลการพิมพ์.

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

Brian S. Everitt. 2010. **Multivariable Modeling and Multivariate Analysis for The Behavioral Sciences.** Taylor & Francis Group, LLC.

Colin Shearer. 2000. "The CRISP-DM Model : The New Blueprint for Data Mining." **JOURNAL OF DATA WAREHOUSING**, Volume 5 Number 4, Fall 2000 : p13

Galavotti, L., Sebastiani, F. and Simi, M. 2000. "Feature Selection and Negative Evidence in Automated Text Categorization.", **Proceedings of KDD-00**, 2000.

Jaiwei Han, Micheline Kamber, and Jian Pei. 2012. **Data Mining Concepts and Techniques**. New York: Elsevier Inc., page 104.

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]

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.

P.-N. Tan, M. Steinbach, and V. Kumar. 2006. **Introduction to data mining.** vol. 1: Pearson Addison Wesley Boston, 2006.

ประวัติย่อผู้วิจัย

ชื่อ	นางอัจจิมา มณฑาพันธุ์
วัน เดือน ปีเกิด	วันที่ 30 สิงหาคม 2501
สถานที่เกิด	กรุงเทพมหานคร
สถานที่อยู่ปัจจุบัน	บ้านเลขที่ 1486 ซอยปาริชาติ ถนนสุทธิสาร เขตห้วยขวาง กทม. 10310
ตำแหน่งหน้าที่การงานปัจจุบัน	อาจารย์ประจำสาขาวิชาวิทยาการคอมพิวเตอร์ประยุกต์
สถานที่ทำงานปัจจุบัน	คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยศรีปทุม บางเขน
ประวัติการศึกษา	พ.ศ. 2523 วท.บ. (สถิติ) จาก มหาวิทยาลัยศิลปากร ค.ศ. 1982 M.S. (Computer Science) Florida Institute of Technology, U.S.A.