

บทที่ 5

สรุปผลการศึกษา และข้อเสนอแนะการวิจัย

5.1 สรุปผลการวิจัย

จากการดำเนินการวิจัยผู้วิจัยได้ดำเนินการวิจัยตามวัตถุประสงค์การวิจัยคือ การศึกษา การเตรียมข้อมูลเพื่อใช้ในการทำเหมืองข้อมูล ได้แก่ การลดขนาดข้อมูล (Data Reduction), การทำความสะอาดข้อมูล (Data Cleaning), การแปลงข้อมูล (Data Transformation) การศึกษาและเลือกโมเดลที่เหมาะสมในการหาความสัมพันธ์ของการเกิดโรคต่างๆ และการแบ่งกลุ่มของข้อมูลเพื่อเป็นแนวทางในการคัดกรองสุขภาพเบื้องต้น โดยเสนอขั้นตอนวิธีการทำเหมืองข้อมูล (Data Mining) เพื่อจำแนกกลุ่มประชากร (Classification) และนำตัวแบบที่ได้ไปประยุกต์ใช้เป็นแนวทางในการคัดกรองสุขภาพเบื้องต้น ซึ่งการทำเหมืองข้อมูลครั้งนี้มีขั้นตอนหลักๆ 5 ขั้นตอนดังนี้

ขั้นตอน 1 การศึกษาและรวบรวมข้อมูล แบบสำรวจสถานะสุขภาพประชาชนจัดทำโดยศูนย์แพทย์ชุมชน สังกัดโรงพยาบาลบุรีรัมย์ ใช้สำหรับเก็บข้อมูลจากประชาชนทั่วไป โดยในการศึกษานี้เป็นการศึกษาข้อมูลการสำรวจสถานะสุขภาพประชาชนจากประชาชน เขตเทศบาลเมืองบุรีรัมย์ ปี พ.ศ. 2555 ซึ่งมีจำนวน 1,071 คน โดยข้อมูลที่ได้มาจะอยู่ในรูปแบบของแฟ้มข้อมูล Excel โดยประกอบด้วยแอทริบิวต์จำนวน 75 แอทริบิวต์ เช่น ลำดับที่ ชื่อ-สกุล รหัสบัตรประชาชน ที่อยู่ ชุมชน สิทธิการรักษา เพศ สถานภาพ การศึกษา น้ำหนัก ส่วนสูง ดัชนีมวลกาย ขนาดรอบเอว ความดัน ประวัติการเจ็บป่วยของครอบครัว ประวัติการเจ็บป่วย การรักษา การสูบบุหรี่ การดื่มเครื่องดื่มที่มีแอลกอฮอล์ การออกกำลังกาย รสชาติของอาหารที่ชอบ การขับถ่าย เพศสัมพันธ์ การตรวจเต้านม การตรวจมะเร็งปากมดลูก พฤติกรรมสุขภาพด้านอาหาร อารมณ์ ออกกำลังกาย และสรุปผลการคัดกรองสุขภาพ เป็นต้น

ขั้นตอน 2 การเตรียมข้อมูล ทำการลดขนาดข้อมูล เนื่องจากข้อมูลมีแอทริบิวต์อยู่เป็นจำนวนมากจึงจำเป็นต้องมีการลดแอทริบิวต์บางตัวที่ไม่เกี่ยวข้องกับการคัดกรองออกไป เช่น ชื่อ-สกุล รหัสบัตรประชาชน เป็นต้น แอทริบิวต์บางส่วนเป็นค่าต่อเนื่อง เช่น ค่าดัชนีมวลกาย (BMI) ได้ถูกทำให้เป็นค่าไม่ต่อเนื่อง (Discretization) มีค่าที่เป็นไปได้ ได้แก่ 1, 2, 3 ซึ่งมีความหมายดังนี้ 1 หมายถึงผอม (น้อยกว่า 18.5) 2 หมายถึง ปกติ (18.5-22.99) 3 หมายถึง ท้วม (23-24.99), 4 หมายถึง อ้วน (25-29.99) และ 5 หมายถึง อ้วนมาก (มากกว่าหรือเท่ากับ 30) จึงสามารถ

ตัดแอทริบิวต์ที่เป็นค่าต่อเนื่องเหล่านี้ทิ้งไป คงไว้แต่ค่าที่มีการให้ความหมายแล้ว ทำให้สามารถลดจำนวนแอทริบิวต์ลงไปได้จาก 73 แอทริบิวต์เป็น 58 แอทริบิวต์ ต่อมาข้อมูลที่ไดยังมีความผิดปกติอยู่ต้องมีการแก้ไขข้อมูลให้ถูกต้อง โดยมีข้อมูลที่ได้รับการแก้ไขให้ถูกต้องทั้งหมด 11 แถว และเนื่องจากข้อมูลในบางแอทริบิวต์ผู้ตอบแบบสอบถามสามารถเลือกได้มากกว่า 1 คำตอบ ได้แก่ ประวัติการเจ็บป่วยของบิดามารดา ประวัติการเจ็บป่วยของพี่น้องสายตรง และรสชาติของอาหารที่ชอบ จึงต้องทำการแปลงข้อมูล (Data Transformation) ที่อยู่ในรูปแบบ String ให้กลายเป็นข้อมูลแบบ Binary (มีค่า 0 และ 1) โดยแตกออกเป็นหลายๆ แอทริบิวต์ตามจำนวนค่าของข้อมูลที่เพิ่มไป ได้ทั้งหมด ส่งผลให้มีแอทริบิวต์ทั้งหมดเพิ่มขึ้นเป็น 81 แอทริบิวต์

ขั้นตอน 3 การทดสอบและเลือกเทคนิคที่เหมาะสม ในการหากฎความสัมพันธ์ (Association Rule) ได้ทดสอบเทคนิค 2 เทคนิค คือ Apriori และ FP Growth จากการทดสอบพบว่าข้อมูลที่เกี่ยวข้องกับเรื่องของสุขภาพจะเหมาะสมกับเทคนิค FP Growth มากกว่า Apriori เพราะจะให้ผลลัพธ์เป็นกฎที่มีแต่ค่า positive (มีประวัติการเป็นโรค) เท่านั้น ในขณะที่ Apriori จะให้ผลลัพธ์ที่เป็นกฎที่มีทั้งค่า positive และ negative (ไม่มีประวัติการเป็นโรค) ผสมกัน ซึ่งไม่มีประโยชน์มากนักในการนำมาใช้เพื่อทำนายการเกิดโรคใหม่ในอนาคตของผู้ป่วย ส่วนการจำแนกกลุ่มข้อมูล (Classification) ได้ทดสอบเทคนิคในการจำแนกกลุ่ม 2 เทคนิคที่ให้ผลลัพธ์เป็นกฎ เนื่องจากต้องนำกฎไปพัฒนาเป็นระบบคัดกรองคือ โมเดลการจำแนกกลุ่มแบบ Decision Tree: C4.5 หรือ ในโปรแกรม Weka เรียกว่า โมเดล Classifiers tree J48 และ โมเดลการจำแนกกลุ่มแบบ Decision Tree: Partial Rules หรือในโปรแกรม Weka เรียกว่า โมเดล Classifiers PART จากการทดสอบการจำแนกกลุ่มทั้ง 3 เทคนิคพบว่า เทคนิค Decision Tree: Partial Rules ให้ผลลัพธ์การจำแนกกลุ่มที่ได้ค่า F-measure ที่มากที่สุดคือ 88.80% จึงเลือกเทคนิคนี้ไปพัฒนาระบบ

ขั้นตอน 4 การวิเคราะห์ผลลัพธ์ที่ได้และความผิดพลาดที่เกิดขึ้น จากผลการทดสอบหาความสัมพันธ์ด้วยเทคนิค FP Growth โดยพิจารณาจากค่า lift พบกลุ่มโรคที่มีความสัมพันธ์กัน 2 กลุ่ม กลุ่มแรกคือกฎที่เกี่ยวข้องกับโรคความดันโลหิตสูง (p2) โรคหัวใจ (p5) และโรคไขมันในเลือดผิดปกติ (p6) กลุ่มที่สองคือกฎที่เกี่ยวข้องกับโรคเบาหวาน (p1) โรคความดันโลหิตสูง (p2) และโรคหัวใจ (p5) จากความสัมพันธ์นี้ทำให้ทราบว่าเมื่อเป็นโรคดังกล่าวแล้วมักจะมีโรคแทรกซ้อนตามมา เช่น ถ้าคนไข้เป็นโรคหัวใจและโรคความดันโลหิตสูงอยู่แล้ว ก็ต้องระวังจะเป็นโรคเบาหวานและโรคไขมันในเลือดผิดปกติซึ่งสามารถประยุกต์ใช้ความรู้นี้ในการประชาสัมพันธ์ผู้ป่วยเพื่อป้องกันการเกิดโรคแทรกซ้อน เช่น พยายามลดพฤติกรรมเสี่ยงที่จะทำให้เกิดโรคเบาหวานและไขมันในเลือดผิดปกติเสียแต่เนิ่นๆ เพื่อลดความเสี่ยงที่จะเป็นโรคดังกล่าว และจากผลการทดสอบการจำแนกกลุ่มข้อมูลแบบ Decision Rule: Partial Rules พบความผิดพลาดของ

ผลลัพธ์ที่ได้เช่น พบว่าข้อมูลที่อยู่ในกลุ่มเสี่ยงแต่โมเดลทำนายว่าเป็นกลุ่มป่วยมีจำนวน 1 ตัวอย่าง จากกฎ หากบุคคลนั้นมีประวัติการเจ็บป่วยกำลังทำการรักษาอยู่ ปฏิบัติตนตามที่แพทย์แนะนำ และ ไม่มีอาการตาพร่ามัวจะจำแนกว่าเป็นกลุ่มป่วย (196.0/1.0) ซึ่งกฎที่ได้เกิดจากมีข้อมูลที่เหมือนกัน จำนวนมากที่ผู้ป่วยตาไม่พร่ามัว ($p15 = 2$) และอยู่ในกลุ่มป่วย เมื่อพิจารณาข้อมูลตัวอย่างรายนี้ พบว่ามีลักษณะค่าแอทริบิวต์บางตัวที่เหมือนกับคนกลุ่มใหญ่ที่ป่วยทำให้ถูกมองว่าป่วยไปด้วย ซึ่ง ข้อมูลนี้อาจถือได้ว่าเป็นข้อมูลที่ผิดปกติ (outlier) เนื่องจากมีคุณสมบัติแตกต่างไปจากกลุ่มคนทั่วไป นอกจากนี้ยังพบว่าโมเดลจำแนกกลุ่ม Class 1 (กลุ่มปกติ) มีความผิดพลาดสูงคือผิดพลาด 50 ตัวอย่างจาก 94 ตัวอย่าง คิดเป็น 53.2% เนื่องจากมีจำนวนตัวอย่างให้ระบบเรียนรู้มีอยู่เพียง 108 ตัวอย่างเท่านั้น ทำให้มีค่า Precision กับ Recall ที่ต่ำคือ 46.8% และ 40.7% ตามลำดับ วิธีแก้ไขคือควรใส่ข้อมูลที่เป็นกลุ่มปกติเข้าไปให้ระบบได้เรียนรู้เพิ่มอีก เพื่อที่จะได้รูปแบบการจำแนกกลุ่ม ที่ถูกต้องมากยิ่งขึ้น เมื่อให้ผู้เชี่ยวชาญช่วยตรวจสอบพบว่าแอทริบิวต์การปฏิบัติตนเมื่อเจ็บป่วยเป็น ข้อมูลที่ไม่เหมาะสมที่จะใช้จำแนกกฎเนื่องจากเป็นแอทริบิวต์ที่มองภาพรวมของประวัติการเจ็บป่วยและอาการต่างๆทั้งหมด 16 แอทริบิวต์ แต่มีเพียง 6 แอทริบิวต์ที่ถ้าเป็นโรคแล้วผู้เชี่ยวชาญจะถือว่าเป็นกลุ่มป่วยทันที ได้แก่ โรคเบาหวาน โรคความดันโลหิตสูง โรคตับ โรคอัมพาต โรคหัวใจ และ โรคไขข้อในเลือดผิดปกติ ดังนั้นแอทริบิวต์การปฏิบัติตนเมื่อเจ็บป่วยจึงเป็นสาเหตุที่ทำให้เกิดการสร้างกฎที่ผิดพลาดจึงทำการทดลองตัดแอทริบิวต์การปฏิบัติตนเมื่อเจ็บป่วยออก พบว่าการจำแนกกลุ่มปกติ กลุ่มเสี่ยง และกลุ่มป่วยให้ผลลัพธ์ที่ดีขึ้น โดยมีความถูกต้อง (Precision) เท่ากับ 89.20% ค่าระลึก (Recall) เท่ากับ 89.40% และ ค่าความเหวี่ยง (F-measure) เท่ากับ 89.30% ตามลำดับ

ขั้นตอน 5 การพัฒนาระบบคัดกรองสุขภาพเบื้องต้น เป็นการนำกฎที่ได้จากการจำแนกกลุ่มด้วยเทคนิค Decision Rule: Partial Rules ไปพัฒนาเป็นเงื่อนไขในการคัดกรองสุขภาพประชาชนเพื่อช่วยลดภาระงานของเจ้าหน้าที่พยาบาลผู้มีหน้าที่เก็บบันทึกและวิเคราะห์ข้อมูลของประชาชน และเปิดให้ประชาชนทั่วไปที่สนใจในเรื่องสุขภาพสามารถเข้ามาตรวจคัดกรองสุขภาพเบื้องต้นของตัวเองได้

5.2 ข้อเสนอแนะ

1. ค่าความถูกต้องของการจำแนกข้อมูลบางกลุ่ม เช่น กลุ่มปกติ มีความถูกต้องน้อยเพราะว่ามีจำนวนตัวอย่างที่เป็นกลุ่มปกติอยู่น้อย ดังนั้นควรจัดเตรียมข้อมูลสำหรับให้ระบบได้เรียนรู้ที่มีจำนวนในแต่ละกลุ่มใกล้เคียงกัน

2. หากมีการปรับปรุงแก้ไขแบบฟอร์มการสอบถามพฤติกรรมสุขภาพของประชาชนในอนาคต ยังสามารถประยุกต์ใช้เทคนิควิธีการต่างๆของงานวิจัยนี้ได้ หรือสามารถใช้กับเอกสารแบบฟอร์มอื่นๆได้

3. ในอนาคตเมื่อระบบคัดกรองสุขภาพเบื้องต้นสามารถเก็บบันทึกข้อมูลของประชาชนทั้งประเทศได้เป็นจำนวนมากแล้ว และต้องการที่จะนำข้อมูลดังกล่าวมาหารูปแบบการจำแนกใหม่ด้วยโปรแกรม Weka อาจจำเป็นต้องแยกข้อมูลออกเป็นภูมิภาคต่างๆเช่น ภาคเหนือ ภาคกลาง ภาคอีสาน ภาคใต้ เป็นต้น เพราะลักษณะของคนในแต่ละภูมิภาคมีความแตกต่างกัน ถ้านำมาจำแนกกลุ่มร่วมกันอาจได้ผลลัพธ์ที่มีความถูกต้องน้อย หรือถ้าต้องการความถูกต้องมากยิ่งขึ้นไปอีกก็อาจจำแนกกลุ่ม โดยใช้ข้อมูลแยกเป็นแต่ละจังหวัด