ข้อมูลเชิงเวลากับการจำแนกประเภทผู้เป็นโรคเบาหวานในประเทศไทย

Temporal Data and Diabetes Classification in Thailand.

สมภพ ปฐมนพ[1], กฤษฎา ศรีแผ้ว[2] และ ม.ล.กุลธร เกษมสันต์[3]

คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยรังสิต ปทุมธานี

Emails: [1]somphop_pathomnop@hotmail.com, [2]s.kritsada@it.rsu.ac.th, [3]kkasemsan@it.rsu.ac.th

ABSTRACT – Diabetes mellitus is a chronic disease that reduces quality of life since it often causes other complications such as heart disease, high blood pressure, neuropathy and the loss of some organs in the body. This work proposes a temporal features extraction model which extracts the features embedded in historical data such as health examination data for classification. The proposed model can be used with any promising classification methods such as Naïve Bayes, Logistic Regression, C4.5 (J48), Bagging and SVMs. This work evaluates the proposed method on health examination data during 2004-2010 (7 years) of factory employees in Thailand. It consists of 43,523 employees in total where 28,808 employees have only one record and 14,715 employees are examined more than once. Resampling with replacement is applied to the dataset for balancing training instances among the classes before proceeding to training process. Features used for diabetes classification are categorized into three groups: Physical Examination, Urinalysis and Biochemistry. The results of experiments show that the data with temporal feature gains higher classification performance than the data without temporal feature.

KEY WORDS -- temporal model, classification, diabetes, data mining, healthcare.

บทคัดย่อ – โรคเบาหวานเป็นโรคเรื้อรังที่ทำให้ผู้ป่วยมีคุณภาพชีวิตที่ลดลง เนื่องจากโรคเบาหวานมักจะก่อให้เกิดภาวะโรคแทรกซ้อนอื่นๆ ตามมา เช่น โรคหัวใจ, โรคความดันโลหิตสูง, โรคระบบประสาท หรือแม้แต่การสูญเสียอวัยวะบางส่วนในร่างกาย ซึ่งเป็นสาเหตุร่วมของการเสียชีวิตด้วยโรคเบาหวาน งานวิจัยนี้ได้นำเสนอรูปแบบข้อมูลเชิงเวลาด้วยการเพิ่มคุณลักษณะข้อมูลเชิงเวลาจากข้อมูลประวัติการตรวจสุขภาพเพื่อการจำประเภทข้อมูลโดยการใช้อัลกอริทึม ได้แก่ Naïve Bayes, Logistic Regression, C4.5 (J48), Bagging และ SVMs ซึ่งงานวิจัยนี้ได้ทำการทดลองบนข้อมูลการตรวจสุขภาพในระหว่างปี พ.ศ.2547 – 2553 (7 ปี) ของลูกจ้างโรงงงานอุตสาหกรรมในประเทศไทยโดยมีจำนวนลูกจ้างทั้งหมด 43,523 ราย เป็นการตรวจเพียงครั้งเดียว 28,808 ราย และตรวจมากกว่าหนึ่งครั้ง 14,715 ราย โดยได้มีการทำรีแซมพลิ่งแบบแทนที่เพื่อปรับอินสแตนซ์ข้อมูลที่ใช้ในการเรียนในแต่ละคลาสให้สมดุลกันก่อนเข้าสู่กระบวนการเรียนต่อไป จากกลุ่มรายการตรวจสุขภาพ 3 รายการ คือ 1) การตรวจร่างกายทั่วไปโดยแพทย์ 2) การตรวจปัสสาวะ 3) การตรวจสารชีวเคมีในเลือด ผลการทดลองได้แสดงให้เห็นว่าข้อมูลที่เพิ่มคุณลักษณะข้อมูลเชิงเวลาให้ผลประสิทธิภาพการจำแนกประเภทดีกว่าข้อมูลแบบปกติที่ไม่มีคุณลักษณะข้อมูลเชิงเวลา

คำสำคัญ -- ข้อมูลเชิงเวลา; การจำแนกประเภทข้อมูล; โรคเบาหวาน; การทำเหมืองข้อมูล; ข้อมูลการตรวจสุขภาพ

1. บทนำ

ข้อมูลของศูนย์ควบคุมและป้องกันโรคประเทศสหรัฐอเมริกา ได้กล่าวถึงสถิติของโรคเบาหวานไว้ในรายงาน National Diabetes Fact Sheet 2011 [1] ที่ตีพิมพ์เมื่อวันที่ 26 มกราคม 2554 ว่าปัจจุบันมีจำนวนประชากร 25.8 ล้านคน (รวมทั้งเด็กและผู้ใหญ่) ในประเทศสหรัฐอเมริกา เป็นโรคเบาหวานซึ่งคิดเป็นร้อยละ 8.3 ของประชากรทั้งหมด ในจำนวนนี้ได้รับการวินิจฉัยแล้ว 18.8 ล้านคน ส่วนอีก 7 ล้านคนยังไม่ได้รับการวินิจฉัย และเป็นผู้ป่วยรายใหม่ถึง 1.9 ล้านคนที่มีอายุ 20 ขึ้นไป

สำหรับประเทศไทยนั้น สำนักงานสำรวจสุขภาพประชาชนไทย (สสท.)ได้ทำการสำรวจ [2] พบว่าความชุกของโรคเบาหวานในประชากรไทยอายุ 15 ปีขึ้นไป คิดเป็นร้อยละ 6.9 ผู้หญิงมีความชุกสูงกว่าผู้ชาย (ร้อยละ 7.7 และ 6 ตามลำดับ) และความชุกเพิ่มขึ้นตามอายุ จากร้อยละ 0.6 ในกลุ่มอายุ 15-29 ปี เพิ่มขึ้นสูงสุดเป็นร้อยละ19.2 สำหรับอายุ 70-79 ปีในผู้ชาย และร้อยละ 16.7 สำหรับอายุ 60-69 ปีในผู้หญิง

โรคเบาหวาน เป็นโรคเรื้อรังที่ทำให้ผู้ป่วยมีคุณภาพชีวิตที่ลดลง เนื่องจากโรคเบาหวานมักจะก่อให้เกิดภาวะโรคแทรกซ้อนอื่นๆ ตามมา เช่น โรคหัวใจ, โรคหลอดเลือดสมอง, โรคความดันโลหิตสูง, โรคตับ, โรคไต, โรคระบบประสาท หรือแม้แต่การสูญเสียอวัยวะบางส่วนในร่างกาย ซึ่งเป็นสาเหตุร่วมของการเสียชีวิตด้วยโรคเบาหวาน

จากข้อมูลดังกล่าว ผู้วิจัยได้เล็งเห็นถึงภัยร้ายแรงจากโรคเบาหวานที่ก่อให้เกิดความสูญเสียชีวิต หรือแม้แต่ความพิการ ตลอดจนงบประมาณค่าใช้จ่ายในการรักษาทางการแพทย์ จึงได้มีแนวความคิดว่าควรหาแนวทางที่สามารถแจ้งเตือนความเสี่ยงในการเป็นโรคเบาหวานสำหรับกลุ่มคนที่ยังไม่ได้ป่วยเป็นโรคเบาหวาน ค้นหากลุ่มคนที่มีแนวโน้มที่จะเป็นโรคเบาหวานเพื่อการเฝ้าระวัง และกำกับติดตามกลุ่มคนที่เป็นโรคเบาหวานอยู่แล้ว เพื่อลดค่าใช้จ่ายในการตรวจสุขภาพ และการรักษาด้วยวิธีการ “การจำแนกประเภทผู้เป็นโรคเบาหวานโดยใช้ข้อมูลเชิงเวลา” ที่จะนำข้อมูลเชิงเวลามาใช้เพื่อเรียนโมเดลการจำแนกประเภทที่มีประสิทธิภาพที่สูงขึ้น

2. งานวิจัยที่เกี่ยวข้อง

จากการศึกษาประโยชน์ที่ได้รับจากเทคนิคการทำเหมืองข้อมูลในโรคเบาหวาน หลายงานวิจัย [3, 4, 5, 6] เป็นงานเกี่ยวกับการจำแนกประเภทผู้เป็นเบาหวานโดยใช้วิธีการจำแนกประเภทที่แตกต่างกัน เช่น Naive Bayes Logistic Regression, Decision Tree, Support Vector Machines (SVMs) และใช้คุณลักษณะของข้อมูลที่แตกต่างกัน ซึ่งเป็นข้อมูลที่ได้จากการตรวจสุขภาพ

B. Adhi Tama และคณะ [6] ตรวจสอบปัจจัยที่ก่อให้เกิดโรคเบาหวาน โดยได้ศึกษาเกี่ยวกับเวชระเบียนของผู้ป่วยจากโรงพยาบาลของรัฐในประเทศอินโดนีเซียในช่วงปี ค.ศ. 2008-2009 ซึ่งเป็นผู้ป่วยมาอย่างน้อย 10 ปีมีจำนวนทั้งหมด 435 ราย โดย 347 ราย (79.8%) เป็นโรคเบาหวานและผู้ป่วย 88 ราย (20.2%) ไม่ได้เป็นโรคเบาหวาน คุณลักษณะข้อมูลที่ใช้สำหรับการทดลอง ได้แก่ เพศ (Sex), อายุ (Age), ดัชนีมวลกาย (BMI), ความดันโลหิต (BP), ไขมัน (hyperlipidemia), ระดับน้ำตาลในเลือดแบบงดอาหาร (FBS), ระดับน้ำตาลในเลือดแบบไม่งดอาหาร (Instant Blood Sugar), ประวัติครอบครัว, ประวัติเป็นเบาหวานขณะตั้งครรภ์, ประวัติการสูบบุหรี่ และอินซูลินในเลือด (Insulin Plasma) ผลการวิจัยสรุปว่า ปัจจัยเสี่ยงที่มีผลต่อโรคเบาหวานคือ 1) ประวัติการสูบบุหรี่, 2) ประวัติเป็นเบาหวานขณะตั้งครรภ์, 3) อินซูลินในเลือด และยังพบว่าไม่มีความแตกต่างของค่าความถูกต้องในการจำแนกประเภทข้อมูลของอัลกอริธึมต่างๆ

B. H. Cho และคณะ [3] ได้ศึกษาปัจจัยเสี่ยงที่มีผลต่อการเกิดโรคเบาหวานในไตซึ่งส่วนใหญ่จะเป็นสาเหตุของการเสียชีวิตของผู้ป่วย วิธีที่ผู้วิจัยเสนอคือการหาคุณลักษณะที่ดีที่สุดที่จะใช้สำหรับการจำแนกประเภทผู้เป็นเบาหวาน โดยใช้ SVMs ซึ่งได้ค่าความถูกต้องของการทำนายที่ดีกว่าสถิติเดิม (t-test, x2-test, variance) ข้อมูลเป็นข้อมูลทางคลินิกของผู้ป่วย 292 รายที่ป่วยเป็นเบาหวานในไต ระยะเวลา 10 ปี (ค.ศ. 1996-2005) โดยมีคุณลักษณะ 184 รายการ จากทั้งในทางการแพทย์และทางคลินิก เช่น การตรวจร่างกายและทางชีวเคมี ผู้วิจัยสามารถบอกได้ว่าคุณลักษณะที่สำคัญที่มีแนวโน้มที่จะเป็นปัจจัยเสี่ยงเป็นโรคเบาหวานในไตมี 39 รายการที่ทำให้ได้ค่า ROC (Receiver Operating Characteristic) สูงที่สุด

K. Takahashi และคณะ [5] ได้ใช้วิธีการตรวจระดับน้ำตาลเฉลี่ยในเลือด (HbA1C) แทนการทดสอบระดับน้ำตาลในเลือดในพลาสมา (plasma กลูโคส) เพื่อการวินิจฉัยโรคเบาหวาน โดยผลการศึกษาบนข้อมูลที่เก็บในระยะเวลา 4 ปีพบว่าเม็ดเลือดแดง (HbA1C) มีประโยชน์ในการทำนาย

โรคเบาหวาน และคุณลักษณะเพิ่มเติม เช่น Aminotransferase, γ-Glutamyl Transpeptidase ยังมีประโยชน์สำหรับการทำนาย

นอกจากนี้ ยังมีงานวิจัยที่ทำนายโอกาสของผู้ป่วยโรคเบาหวานที่จะเป็นโรคหัวใจ [8] โดยใช้อัลกอริธึม Naive Bayes และการหากลุ่มของคุณลักษณะที่ดีที่สุดเพื่อการสร้างโมเดลการจำแนกประเภท โดยกลุ่มของคุณลักษณะดังกล่าว คือ เพศ (Sex), อายุ (Age), พันธุกรรม (Genetic), น้ำหนัก (Weight), ความดันโลหิต (Blood Pressure), ระดับน้ำตาลในเลือดแบบงดอาหาร (Fasting Blood Sugar), การทดสอบระดับน้ำตาลในเลือดหลังกินอาหาร (Test Blood Sugar Levels after eating) และ ระดับน้ำตาลเฉลี่ยในเลือด (HbA1C)

ซึ่งโดยส่วนใหญ่ งานวิจัยในการจำแนกประเภทข้อมูลจะเน้นที่การสร้างโมเดลการจำแนกประเภทโดยไม่ได้คำนึงถึงความสัมพันธ์ของอินแสตนซ์ข้อมูลที่อาจมีความสัมพันธ์เชิงเวลาได้ อย่างไรก็ตามงานวิจัยของ R. Peter และคณะ [7] ได้นำไปสู่การวิเคราะห์ข้อมูลที่ใช้ความสัมพันธ์ของข้อมูลเชิงเวลาในอดีตจนถึงปัจจุบัน เพื่อที่จะคาดการณ์สิ่งที่จะเกิดขึ้นในอนาคต งานวิจัยนี้เป็นการพยากรณ์อากาศโดยใช้ข้อมูลอุตุนิยมวิทยาของคณะกรรมาธิการคุณภาพสิ่งแวดล้อแห่งรัฐเทกซัส (Texas Commission of Environmental Quality : TCEQ) และข้อมูลไข้หวัดใหญ่ของ Google Flu Trends โดยผลการศึกษาพบว่า การใช้คุณลักษณะ 40 รายการของข้อมูลทั้งหมด 886 อินสแตนซ์ บวกกับความสัมพันธ์ข้อมูลเชิงเวลาให้ค่าความถูกต้องของการจำแนกประเภทข้อมูลที่สูงขึ้นทั้งอัลกอริธึม SVMs และ ID3 ซึ่งสรุปว่า การจำแนกประเภทข้อมูลด้วยรูปแบบข้อมูลเชิงเวลาให้ผลลัพธ์ที่ได้ค่าความถูกต้องมากกว่าข้อมูลแบบปกติบนข้อมูลอุตุนิยมวิทยา

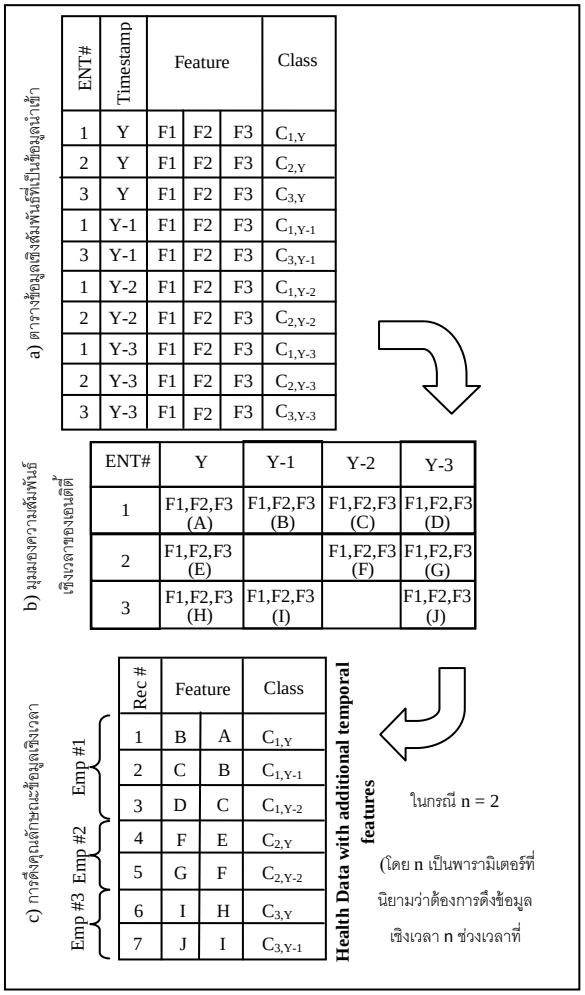
3. วิธีการที่นำเสนอ

3.1 การเพิ่มคุณลักษณะข้อมูลเชิงเวลา

ความสัมพันธ์เชิงเวลาจะเกิดขึ้นระหว่างอินแสตนซ์ของข้อมูล โดยข้อมูลอินแสตนซ์เหล่านั้นเป็นตัวแทนของเอนติตี้ที่มีค่าเหมือนหรือแตกต่างกันตามช่วงเวลา โดยเอนติตี้เป็นแนวคิดเชิงวัตถุ เช่น คน, สัตว์ หรือสิ่งของ ซึ่งข้อมูลการตรวจสุขภาพก็เป็นข้อมูลที่มีความสัมพันธ์เชิงเวลาเช่นกัน เช่น บุคคลหนึ่งจะมีข้อมูลประวัติในอดีตที่ถูกเก็บไว้ตั้งแต่อดีตถึงปัจจุบัน บุคคลแต่ละคนสามารถตรวจสุขภาพของตนเองได้มากกว่าหนึ่งครั้งต่อปี และอาจได้ผลที่ไม่เหมือนกัน การเก็บข้อมูลในลักษณะนี้เกิดขึ้นได้ทุกที่โดยเฉพาะอย่างยิ่งในข้อมูลบางประเภทที่มีการสร้างข้อมูลให้แก่เอนติตี้เดิมซ้ำๆ ตามระยะเวลาที่กำหนด เช่น การจัดตรวจสุขภาพของโรงงานให้กับลูกจ้างเป็นระยะ ๆ

ในวงการแพทย์ เป็นที่ทราบกันดีว่าข้อมูลประวัติในอดีตของผู้ป่วยเป็นประโยชน์ในการวินิจฉัยโรคได้ ดังนั้นในงานวิจัยนี้จึงนำเสนอการสร้างโมเดลข้อมูลที่รวมข้อมูลทั้งในอดีตและปัจจุบันในการตรวจสุขภาพของลูกจ้างในโรงงานอุตสาหกรรม เพื่อเรียนโมเดลการจำแนกประเภทข้อมูลด้วยวิธีการเพิ่มคุณลักษณะข้อมูลเชิงเวลาจากความสัมพันธ์เชิงเวลาของข้อมูล และถึงแม้ว่างานวิจัยนี้จะมุ่งเน้นไปที่ข้อมูลการตรวจสุขภาพของลูกจ้างในโรงงานอุตสาหกรรม แต่ก็สามารถนำไปประยุกต์ใช้กับงานอื่นๆ ที่ข้อมูลมีคุณลักษณะที่เกี่ยวข้องกับข้อมูลเชิงเวลาได้

โมเดลสร้างคุณลักษณะข้อมูลเชิงเวลา จะเริ่มโดยการมองความสัมพันธ์ของข้อมูลเชิงเวลาด้วยการรับตารางข้อมูลเชิงสัมพันธ์ที่มีหมายเลขเอนติตี้และเวลาที่ข้อมูลอินสแตนซ์นั้นมีค่าเป็นข้อมูลนำเข้า สมมติว่าตารางข้อมูลเชิงสัมพันธ์เป็นแสดงในรูปที่ 1a), อินสแตนซ์หนึ่งอินสแตนซ์ประกอบด้วยกลุ่มคุณลักษณะ 3 กลุ่มรายการ (F1, F2, F3) และ มีคลาสกำกับประเภทข้อมูล หนึ่งเอนติตี้ (ENT #) สามารถมีค่าคุณลักษณะต่างๆ ในกลุ่มคุณลักษณะทั้งสามที่แตกต่างกันในแต่ละช่วงเวลาที่สะท้อนเป็นอินสแตนซ์ในตารางข้อมูลเชิงสัมพันธ์ ตัวอย่างเช่น ENT #1 เกิดขึ้น ณ เวลา Y, Y-1, Y-2 และ Y-3 ซึ่งสามารถมองสรุปจากความสัมพันธ์เชิงเวลาของเอนติตี้เดียวกันเป็นแต่ละเรคคอร์ดได้ดังรูปที่ 1b) เอนติตี้อาจจะไม่ปรากฏค่าในทุกๆ ช่วงเวลา จึงทำให้ ณ บางช่วงเวลาเป็นค่าว่าง (Null Value) ได้ อย่างไรก็ตามข้อมูลที่จะใช้ในการเรียนโมเดลการจำแนกประเภท แต่ละเรคคอร์ดจะต้องมีจำนวนคุณลักษณะเท่ากัน เราจึงทำการแปลงข้อมูลให้อยู่ในรูปแบบที่นำไปเรียนโมเดลได้ และมีข้อมูลคุณลักษณะเชิงเวลาด้วยได้ดังรูปที่ 1c) ซึ่งสามารถปรับเปลี่ยนพารามิเตอร์ n ได้ เพื่อระบุว่าต้องการสร้างข้อมูลเชิงเวลา n อินแสตนซ์ที่ติดต่อกันสำหรับการจำแนกประเภทข้อมูลเชิงเวลา ซึ่งเรื่องของ n พารามิเตอร์นั้นจะศึกษาต่อในหัวข้อถัดไป สังเกตว่าโมเดลข้อมูลนี้ได้แก้ปัญหาค่าว่าง ณ บางช่วงเวลาโดยการเลื่อนค่า (Shift Value) เป็นช่วงเวลาถัดไปที่มีค่าข้อมูล

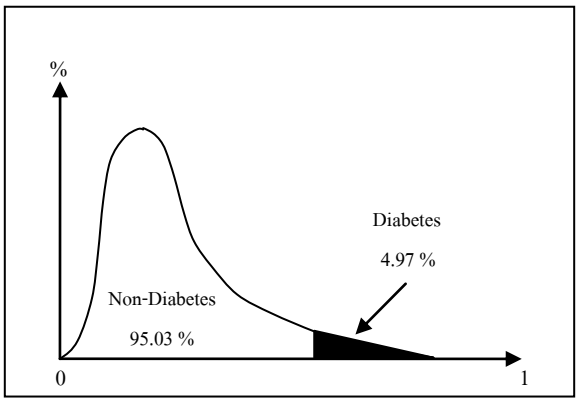


รูปที่ 1. โมเดลสร้างคุณลักษณะข้อมูลเชิงเวลา

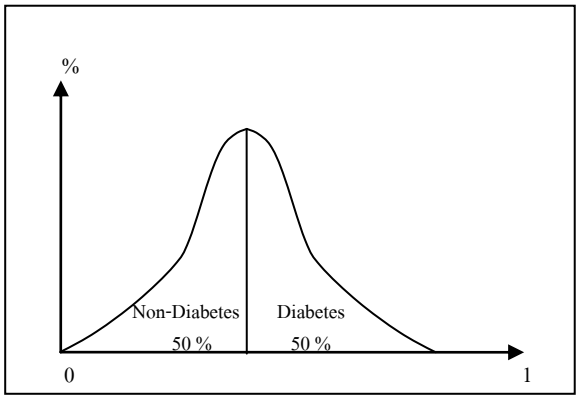
3.2 การทำข้อมูลให้มีความสมดุล

ข้อมูลการตรวจสุขภาพที่ได้นำมาทดลองนั้นประกอบด้วยลูกจ้างจำนวนทั้งสิ้น 43,523 ราย พบว่ามีกลุ่มคนที่ไม่เป็นเบาหวาน (Non-Diabetes Class) 41,359 ราย คิดเป็นร้อยละ 95.03 ของจำนวนทั้งหมด และกลุ่มคนที่เป็นเบาหวาน (Diabetes Class) 2,164 ราย คิดเป็นร้อยละ 4.97 ของจำนวนทั้งหมด

ผู้วิจัยจึงได้เสนอวิธีการการรีแซมพลิ่ง (Resampling Method) [12] ซึ่งมีด้วยกัน 2 วิธี คือ แบบแทนที่ (with Replacement) และ แบบไม่แทนที่ (without Replacement) เพื่อทำให้ข้อมูลตัวอย่างที่ใช้สร้างโมเดลจำแนกประเภทมีจำนวนข้อมูลในแต่ละคลาสที่สมดุลกัน ซึ่งจะส่งผลให้ได้ค่าการประเมินประสิทธิภาพที่เหมาะสม โดยการทดลองในครั้งนี้ได้เลือกใช้วิธีรีแซมพลิ่งแบบแทนที่ (Resampling with Replacement) ดังแสดงตัวอย่างในรูปที่ 2 และ 3



รูปที่ 2. การกระจายคลาสก่อนทำรีแซมพลิ่งข้อมูล



รูปที่ 3. การกระจายคลาสหลังทำรีแซมพลิ่งข้อมูล

4. การวางแผนการทดลอง

4.1 ชุดข้อมูล (Dataset)

ชุดข้อมูลที่ใช้ในการวิจัยครั้งนี้เป็นข้อมูลการตรวจสุขภาพในช่วง 2547 - 2553 (7 ปี) ของลูกจ้างโรงงานอุตสาหกรรมในประเทศไทย ซึ่งมีจำนวนทั้งสิ้น 43,523 ราย เป็นการตรวจสุขภาพเพียงครั้งเดียว 28,808 ราย และการตรวจสุขภาพมากกว่าหนึ่งครั้ง 14,715 ราย ซึ่งในจำนวนนี้มีการตรวจสุขภาพทั้งหมด 69,994 ครั้ง โดย 41,186 ครั้งเกิดจากผู้ที่ตรวจสุขภาพมากกว่าหนึ่งครั้ง ข้อมูลมีคุณลักษณะพื้นฐานทั้งหมด 16 รายการจากกลุ่มคุณลักษณะ 3 กลุ่มรายการ คือ การตรวจร่างกายทั่วไปโดยแพทย์ (Physical Examination : F1), การตรวจปัสสาวะ (Urinalysis : F2) และการตรวจสารชีวเคมีใน

เลือด (Biochemistry : F3) โดยคุณลักษณะผลการตรวจระดับน้ำตาลในเลือดเป็นตัวจำแนกประเภทข้อมูล (Fasting Blood Sugar : FBS_dm) ซึ่งรายละเอียดของคุณลักษณะต่างๆ เป็นดังแสดงในตาราง 1

ตารางที่ 1. คุณลักษณะพื้นฐานการตรวจสุขภาพและการแบ่งกลุ่มคุณลักษณะ

Group of Features	Feature Name	Nominal Value
F1	Age *	< 45, 45 – 49, >= 50
	Sex	Male, Female
	Weight *	<= 50, 51 – 99, >= 100
	Height *	< 150, 151 – 169, >= 170
	BMI ** (Body Mass Index)	Thin, Normal, Obesity (<=18) (19-25) (>=26)
	BP_S *** (Systolic Blood Pressure)	Normal,High–Normal, Hypertension (<130) (130-139) (>=140)
	BP_D *** (Diastolic Blood Pressure)	Normal,High–Normal, Hypertension (<85) (85-89) (>=90)
	Pulse *	<= 60, 61 – 79, >= 80
	PE_Nor	Normal, Abnormal
F2	UPr	Negative, Trace, 1+, 2+, 3+, 4+
	USu	Negative, Trace, 1+, 2+, 3+, 4+
F3	CRE * (Creatinine)	Normal, High – Normal, High (<=1.50) (1.51-3.90)(>=3.91)
	GPT *	Normal, High – Normal, High (<=45) (46-89) (>=90)
	CHO * (Cholesterol)	Normal, High – Normal, High (<=200) (201-240) (>=241)
	TG * (Triglyceride)	Normal, High – Normal, High (<=170) (171-400) (>=401)
Class	FBS_dm	Non-Diabetes, Diabetes

* การแบ่งกลุ่มข้อมูลโดยผู้เชี่ยวชาญทางการแพทย์

** การแบ่งกลุ่มข้อมูลโดย WHO BMI Classification

*** การแบ่งกลุ่มข้อมูลโดย WHO-ISH Guideline 2003

เนื่องจากคุณลักษณะบางรายการมีชนิดข้อมูลเป็นตัวเลข การจำแนกประเภทข้อมูลด้วยอัลกอริธึมบางอัลกอริธึม ไม่สามารถรองรับข้อมูลที่เป็นตัวเลขได้ เช่น Naïve Bayes ผู้วิจัยจึงได้ทำดิสครีไตซ์ข้อมูลที่มีชนิดข้อมูลเป็นตัวเลขโดยอาศัยหลักการแบ่งกลุ่มข้อมูลของ WHO [9, 10] และใช้ผู้เชี่ยวชาญทางการแพทย์ ดังแสดงเครื่องหมายไว้ในตาราง 1

อย่างไรก็ตาม ยังไม่มีข้อสรุปที่แน่ชัดจากงานวิจัยที่เกี่ยวข้องว่ากลุ่มคุณลักษณะที่ดีที่สุดสำหรับการจำแนกประเภทข้อมูลโรคเบาหวาน คือคุณลักษณะกลุ่มใดบ้าง ดังนั้นเราจึงใช้กลุ่มคุณลักษณะเหล่านี้ทั้งหมดเป็นชุดข้อมูลหลักสำหรับทดลองและศึกษาประสิทธิภาพการจำแนกประเภทผู้เป็นเบาหวานโดยใช้ข้อมูลเชิงเวลา

4.2 การทำข้อมูลให้มีความสมดุล

เนื่องจากจำนวนของคลาสการจำแนกประเภทผู้เป็นเบาหวาน (Non-Diabetes, Diabetes) นั้นมีจำนวนที่แตกต่างกันอย่างมาก ซึ่งผู้วิจัยได้กล่าวไว้ในข้อ 3.2 โดยนำเสนอการทำข้อมูลให้มีความสมดุลของทั้งสองคลาส ผลการทดลองได้ตามตารางที่ 2

ตารางที่ 2. จำนวนของคลาสหลังการรีแซมพลิ่ง

Class	Number of Class (After Resampling)						
	Muti-All	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**
Diabetes	1,762	1,122	532	252	120	58	23
Non-Diabetes	1,854	1,183	596	282	135	62	26
Total	**3,616**	**2,305**	**1,128**	**534**	**255**	**120**	**49**

4.3 วิธีการจำแนกประเภทข้อมูลและเครื่องมือที่ใช้การทดลอง

อัลกอริธึมการจำแนกประเภทข้อมูลหลายตัวได้ถูกประยุกต์ใช้ในการจำแนกประเภทที่เกี่ยวข้องกับโรคเบาหวาน [3, 4, 5, 6] แต่ยังไม่พบว่ามีอัลกอริธึมใดที่แสดงผลว่ามีประสิทธิภาพมากที่สุด ดังนั้นผู้วิจัยจะทำการทดลองการจำแนกประเภทข้อมูลผู้เป็นเบาหวาน โดยใช้อัลกอริธึมในการจำแนกประเภทที่ปรากฏในงานวิจัยต่างๆ ได้แก่ Naive Bayes [8], Logistic Regression [5], C4.5 (J48 สำหรับ WEKA) [6], Bagging [6] และ SVMs [3] ในงานวิจัยนี้ผู้วิจัยได้ใช้ WEKA [11] เป็นเครื่องมือในการทำเหมืองข้อมูล และใช้วิธีการ k-fold cross validation โดยที่ k=10 สำหรับการเรียนและทดสอบโมเดลการจำแนกประเภทข้อมูล เพื่อศึกษาค่าความถูกต้องและประสิทธิภาพของโมเดล

การจำแนกประเภทผู้เป็นเบาหวานที่สร้างจากอัลกอริธึมที่แตกต่างกัน

5. ผลการทดลอง

การประเมินผลการทดลองการจำแนกประเภทข้อมูลได้ใช้ค่าความถูกต้อง (Accuracy) และ F-Measure เพื่อวัดประสิทธิภาพของโมเดลการจำแนกประเภทที่สร้างขึ้นด้วยลักษณะที่แตกต่างกัน

ตารางที่ 3 แสดงผลการทดลองเปรียบเทียบประสิทธิภาพโมเดลการจำแนกประเภทที่เรียนจากข้อมูลด้วยกลุ่มคุณลักษณะการตรวจสุขภาพที่แตกต่างกัน (F1, F2, F3) ผลการทดลองสรุปได้ว่า กลุ่มรายการตรวจปัสสาวะ (Urinalysis: F2) ที่ประกอบไปด้วยคุณลักษณะข้อมูล 2 รายการคือ การตรวจระดับโปรตีนในปัสสาวะ (UPr) และการตรวจระดับน้ำตาลในปัสสาวะได้ให้ค่าความถูกต้อง (Accuracy) และ F-Measure ในการประเมินประสิทธิภาพสูงที่สุดในทุกๆ อัลกอริธึมสำหรับการสร้างโมเดลการจำแนกประเภทข้อมูล และยังสังเกตได้อีกว่าการใช้ข้อมูลกลุ่มคุณลักษณะแบบผนวกเข้าด้วยกัน (กรณี F_i+F_j หรือ $F_1+F_2+F_3$) ก็มิได้ทำให้ค่าความถูกต้อง (Accuracy) และ F-measure เปลี่ยนแปลงมากนัก

ตารางที่ 3. การเปรียบเทียบผลการทดลองกลุ่มคุณลักษณะข้อมูลการตรวจสุขภาพ

Method	Evaluation	F1	F2	F3	F1+F2	F1+F3	F2+F3	F1+F2+F3
N. Bayes	A	94.42	97.17	95.85	96.48	92.70	97.13	94.82
	F	0.13	0.51	0.01	0.46	0.21	0.50	0.41
Logistic R.	A	95.91	97.11	95.91	97.17	95.90	97.15	97.15
	F	0.00	0.49	0.00	0.50	0.01	0.50	0.50
C4.5(J48)	A	95.91	97.17	95.91	97.16	95.91	97.17	97.16
	F	0.00	0.51	0.00	0.51	0.00	0.51	0.50
Bagging	A	95.91	97.16	95.91	97.16	95.90	97.16	97.16
	F	0.00	0.50	0.00	0.50	0.01	0.51	0.50
SVMs	A	95.91	97.06	95.91	96.90	95.91	96.97	96.97
	F	0.00	0.48	0.00	0.43	0.00	0.45	0.45

หมายเหตุ: **A** คือ Accuracy, **F** คือ F-measure

ตารางที่ 4 แสดงการเปรียบเทียบผลการทดลองระหว่างรูปแบบข้อมูลปกติและข้อมูลที่เพิ่มคุณลักษณะข้อมูลเชิงเวลา โดยที่ n คือพารามิเตอร์ของรูปแบบข้อมูลเชิงเวลา ดังที่ได้กล่าวในข้อ 3.1 ตัวอย่างเช่น T3 เป็นกรณีที่ n เท่ากับ 3 ที่สร้างข้อมูลเชิงเวลามาจัดในรูปแบบ 1 เรคคอร์ดประกอบด้วยข้อมูลการตรวจสุขภาพของลูกจ้าง 1 คนที่ติดต่อกัน 3 ครั้ง จากตารางที่ 4 แสดงให้เห็นว่า คุณลักษณะข้อมูลเชิงเวลา (T2-T7) ให้โมเดลการจำแนกประเภทข้อมูลที่มีค่าความถูกต้อง (Accuracy) และ F-Measure สูงกว่าข้อมูลแบบปกติ (Multi-All) ซึ่งไม่มีการเพิ่มคุณลักษณะข้อมูลเชิงเวลา โดยมีค่าสูงที่สุดที่ T5 ในอัลกอริธึม C4.5 (J48) คือ 85.88 และ 0.85 ตามลำดับ และยังพบอีกว่าค่าเฉลี่ยสูงสุดคืออัลกอริธึม Bagging มีค่าความถูกต้อง (Accuracy) และ F-Measure คือ 84.05 และ0.83 ตามลำดับ ซึ่งมีค่ามากกว่าข้อมูลแบบปกติ

ตารางที่ 4. การเปรียบเทียบผลการทดลองระหว่างชุดข้อมูลปกติและชุดข้อมูลเชิงเวลา

Algorithm	Evaluation	Multi-All	T2	T3	T4	T5	T6	T7	Avg.
N. Bayes	A	74.97	77.74	79.61	78.65	79.22	75.83	79.59	78.44
	F	0.74	0.78	0.79	0.78	0.78	0.74	0.77	0.77
Logistic R.	A	76.22	79.22	80.50	78.28	76.47	74.17	73.47	77.02
	F	0.74	0.78	0.78	0.77	0.77	0.74	0.74	0.76
C4.5 (J48)	A	76.19	79.13	81.65	83.90	85.88	78.33	79.59	81.41
	F	0.74	0.78	0.80	0.82	0.85	0.76	0.77	0.80
Bagging	A	77.38	80.65	83.07	88.20	85.10	77.50	89.80	84.05
	F	0.76	0.790	0.812	0.87	0.83	0.748	0.90	0.83
SVMs	A	73.04	77.27	81.65	81.09	85.49	71.67	79.59	79.46
	F	0.69	0.74	0.80	0.79	0.85	0.71	0.79	0.78

หมายเหตุ: **A** คือ Accuracy, **F** คือ F-measure

จากตารางที่ 4 เป็นการทดลองบนชุดข้อมูลที่ได้มาจากการเพิ่มคุณลักษณะข้อมูลเชิงเวลาที่แตกต่างกัน ส่วนชุดข้อมูล Multi-All เป็นข้อมูลที่ไม่มีคุณลักษณะข้อมูลเชิงเวลาหรือเรียกว่าชุดข้อมูลปกติ ซึ่งในกรณีของชุดข้อมูล Multi-All นั้นสร้างโดยกำหนดให้ 1 เรคคอร์ด คือข้อมูลการตรวจสุขภาพของลูกจ้าง ณ เวลาใดเวลาหนึ่งเท่านั้น แต่ในกรณีชุดข้อมูลเชิงเวลา T2 จะเป็นการสร้างโดยกำหนดให้ 1 เรคคอร์ดประกอบไปด้วยข้อมูลการตรวจสุขภาพที่ต่อเนื่องกัน 2 ครั้งของคนหนึ่งคน หรือในกรณีชุดข้อมูลเชิงเวลา T5 จะเป็นการสร้างโดยกำหนดให้ 1 เรคคอร์ดประกอบไปด้วยข้อมูลการตรวจสุขภาพที่ต่อเนื่องกัน 5 ครั้งของคนหนึ่งคน(คนๆนั้นต้องมีการตรวจสุขภาพอย่างน้อย 5 ครั้งด้วย) ซึ่งจะทำลักษณะเดียวกันนี้จนถึงการสร้างชุดข้อมูลเชิงเวลา T7 ซึ่งเป็นจำนวนการตรวจสุขภาพสูงสุด (7 ปี) จากชุดข้อมูลที่มีทั้งหมด

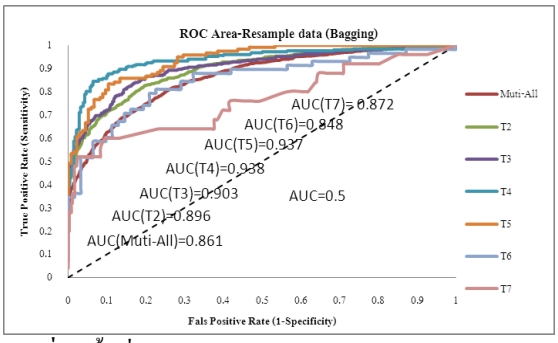
ผลการทดลองในตารางที่ 4 นั้นยังแสดงให้เห็นว่าการเพิ่มคุณลักษณะข้อมูลเชิงเวลาให้ค่าความถูกต้อง (Accuracy) และค่า F-Measure ที่สูงกว่าชุดข้อมูลแบบปกติ (Multi-All) ที่ไม่ได้อาศัยการเพิ่มคุณลักษณะข้อมูลเชิงเวลาในทุกๆ อัลกอริธึมโดยมีค่าสูงสุด คือ อัลกอริธึม Naïve Bayes ที่ T3 (79.61, 0.79), Logistic Regression ที่ T3 (80.50, 0.78), C4.5 (J48) ที่ T5 (85.88, 0.85), Bagging ที่ T4 (88.20, 0.87) และ SVMs ที่ T5 (85.49, 0.85)

จำนวนเรคคอร์ดของชุดข้อมูลแบบปกติและชุดข้อมูลเชิงเวลาแต่ละชุดข้อมูลเป็นดังตารางที่ 5 โดยจะเห็นเปอร์เซ็นต์ที่ลดลงของจำนวนเรคคอร์ดที่ T สูงขึ้นเมื่อเปรียบเทียบกับข้อมูลแบบปกติ นั่นหมายถึงจำนวนอินสแตนซ์ที่ใช้เรียนโมเดลลดลง ตัวอย่างเช่น ในอัลกอริธึม C4.5 (J48) ที่ได้ค่าการประเมินประสิทธิภาพสูงสุด คือ Accuracy เท่ากับ 85.88 และ F-measure เท่ากับ 0.85 ที่ T5 พบว่าใช้จำนวนอินสแตนซ์เพียงแค่ 255 อินสแตนซ์ หรือคิดเป็นร้อยละ 7.05 ของจำนวนอินสแตนซ์ทั้งหมดของข้อมูลแบบปกติ

ตารางที่ 5. จำนวนเรคคอร์ดชุดข้อมูลปกติและชุดข้อมูลเชิงเวลา

Type of Data	Extending Features with Resampled Data	
	Number of Instances	%
Multi-All	3,616	100.00
T2	2,305	63.74
T3	1,128	31.19
T4	534	14.77
T5	255	7.05
T6	120	3.32
T7	49	1.36

รูปที่ 4 แสดงพื้นที่ใต้กราฟ ROC ระหว่างชุดข้อมูลที่เพิ่มคุณลักษณะข้อมูลเชิงเวลาตั้งแต่ T2 ถึง T7 กับชุดข้อมูลแบบปกติที่ไม่ใช้คุณลักษณะข้อมูลเชิงเวลาจากการใช้อัลกอริธึม Bagging โดยพบว่าชุดข้อมูลที่เพิ่มคุณลักษณะข้อมูลเชิงเวลาได้ค่าพื้นที่ใต้กราฟ ROC สูงสุดที่ T4 คือ 0.938 และมีค่ามากกว่าชุดข้อมูลแบบปกติ (Multi-All) คือ 0.861 โดยมีจำนวนอินสแตนซ์สำหรับการเรียนโมเดลในการจำแนกประเภทข้อมูล 534 อินสแตนซ์ หรือคิดเป็นร้อยละ 5.34 ของจำนวนอินสแตนซ์ทั้งหมดของข้อมูลแบบปกติ



รูปที่ 4. พื้นที่ใต้กราฟ ROC โดยการใช้อัลกอริธึม Bagging

6. ข้อสรุป

กลุ่มคุณลักษณะของรายการตรวจสุขภาพที่เป็นประโยชน์สำหรับการจำแนกประเภทข้อมูลผู้เป็นเบาหวานคือ รายการตรวจปัสสาวะ (Urinalyis : F2) ที่ประกอบไปด้วยคุณลักษณะระดับโปรตีนในปัสสาวะ (UPr) และระดับน้ำตาลในปัสสาวะ (USu) การประเมินประสิทธิภาพโมเดลการจำแนกประเภทที่สร้างจากอัลกอริทึม Naive Bayes, Logistic Regression, C4.5 (J48), Bagging และ SVMs พบว่าไม่มีความแตกต่างอย่างมีนัยสำคัญในการจำแนกประเภทผู้เป็นเบาหวาน โมเดลการสร้างคุณลักษณะข้อมูลเชิงเวลาเป็นประโยชน์อย่างมากสำหรับการจำแนกประเภทผู้เป็นเบาหวาน โดยให้โมเดลการจำแนกที่มีประสิทธิภาพสูงกว่าการใช้ข้อมูลแบบปกติที่ไม่มีข้อมูลเชิงเวลา แม้งานนี้จะเป็นการจำแนกประเภทผู้เป็นเบาหวาน แต่โมเดลสร้างคุณลักษณะข้อมูลเชิงเวลาที่เสนอมีความเป็นสามัญสามารถนำไปประยุกต์ใช้กับข้อมูลอื่นๆ ที่มีลักษณะข้อมูลเชิง

เวลาแฝงอยู่ในตัวข้อมูลได้ เช่น ข้อมูลพยากรณ์อากาศ, ข้อมูลทางการวินิจฉัยโรค, ข้อมูลการพยากรณ์ความล้มเหลวของธุรกิจ เป็นต้น

7. เอกสารอ้างอิง

[1] National Center for Chronic Disease Prevention and Health Promotion. National Diabetes Fact Sheet. [Online]. Available:http://www.cdc.gov/diabetes/pubs/pdf/ndfs_2011.pdf, 2011.

[2] สำนักงานสำรวจสุขภาพประชาชนไทย. “รายงานการสำรวจสุขภาพประชาชนไทยโดยการตรวจร่างกาย.” [ออนไลน์]. เข้าถึงได้จาก : http://nheso.or.th/loadfile/diabetes_mellitus.pdf, 2554.

[3] B. H. Cho, H. Yu, K. Kim, T. H. Kim, I. Y. Kim and S. I. Kim. Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods. Journal Artificial Intelligence in Medicine. 2008, 42 : 37-53.

[4] H. N. A. Pham and E. Triantaphyllou. Prediction of Diabetes by Employing a New Data Mining Approach Which Balance Fitting and Generalization. Computer and Information Science. 2008, 131:11-26.

[5] K. Takahashi, H. Uchiyama, S. Yanagisawa and I. Kamae. The Logistic Regression and ROC Analysis of Group-based Screening for Predicting Diabetes Incidence in Four Years. The Kobe journal of medical science. 2006, 52 (6): 171-180.

[6] B. A. Tama, Rodiyatul F.S. and Hermansyah. An Early Detection Method of Type-2 Diabetes Mellitus in Public Hospital. Proceeding of The International Conference on Informatics, Cybernetic,and Computer Applications. Bangalore. 2010, 9 (2): 287-294.

[7] R. Peter and T. Thomas. Temporal Data Classification using Linear Classifiers. Journal Information Systems. 2011, 36 (1): 30-41.

[8] G. Parthiban, A. Rajesh, and S. K. Srivatsa. Diagnosis of Heart Disease for Diabetic Patients using Naïve Bayes Method. International Journal of Computer Applications. 24 (2011) : 7-11.

[9] World Health Organization. BMI Classification. [Online]. Available : http://apps.who.int/bmi/index.jsp?introPage=intro_3.html, 2011.

[10] World Health Organization. 2003 World Health Organization (WHO)/International Society of Hypertension (ISH) statement on management of hypertension. [Online]. Available : http://www.who.int/cardiovascular_diseases/guidelines/hypertension/en/, 2011.

[11] I. H. Witten, E. Frank. Data mining: Practical Machine Learning Tools and Techniques, 3rd Edition. San Francisco: Morgan Kaufmann, 2011.

[12] A. T. Arnholt. Resample with R. Teaching Statistics, 2007, 29(1), 21-26.