การจำแนกข้อมูลที่ไม่สมบูรณ์ด้วยลำดับเวลาของการตรวจโรค ที่แตกต่างกันจากชุดข้อมูลลำดับเวลาทางการแพทย์

โดย

นางสาวเกล้ากัลยา ศิลาจันทร์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาปรัชญาดุษฎีบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์และสารสนเทศ

ภาควิชาคอมพิวเตอร์

บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร

ปีการศึกษา 2558

การจำแนกข้อมูลที่ไม่สมบูรณ์ด้วยลำดับเวลาของการตรวจโรค ที่แตกต่างกันจากชุดข้อมูลลำดับเวลาทางการแพทย์

โดย

นางสาวเกล้ากัลยา ศิลาจันทร์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาปรัชญาดุษฎีบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์และสารสนเทศ

ภาควิชาคอมพิวเตอร์

บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร

ปีการศึกษา 2558

CLASSIFYING INCOMPLETE DATA WITH DISTINCT DIAGNOSTIC TIME SEQUENCE OF TEMPORAL MEDICAL DATA

By

Miss Klaokanlaya Silachan

A Thesis Submitted in Partial Fulfillment of the Requirement for the Degree

Doctor of Philosophy Program in Computer and Information Science

Department of Computing

Graduate School, Silpakorn University

Academic Year 2015

บัณฑิตวิทยาลัยมหาวิทยาลัยศิลปากร อนุมัติให้วิทยานิพนธ์ เรื่อง “การจำแนกข้อมูลที่ไม่สมบูรณ์ด้วยลำดับเวลาของการตรวจโรคที่แตกต่างกันจากชุดข้อมูลลำดับเวลาทางการแพทย์” เสนอโดย นางสาวเกล้ากัลยา ศิลาจันทร์ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์และสารสนเทศ

.....

(รองศาสตราจารย์ ดร.ปานใจ ธารทัศนวงศ์)

คณบดีบัณฑิตวิทยาลัย

วันที่.....เดือน.....พ.ศ.....

อาจารย์ที่ปรึกษาวิทยานิพนธ์

1. รองศาสตราจารย์ ดร.ปานใจ ธารทัศนวงศ์
2. ศาสตราจารย์ ดร.ชิดชนก เหลือสินทรัพย์

คณะกรรมการตรวจสอบวิทยานิพนธ์

.....ประธานกรรมการ

(อาจารย์ ดร.สุนีย์ พงษ์พินิจภิญโญ

...../...../.....

.....กรรมการ

(ผู้ช่วยศาสตราจารย์ดร.ศรันญา มณีโรจน์)

...../...../.....

.....กรรมการ

(อาจารย์ ดร.ภิญโญ แท้ประสาทสิทธิ์)

...../...../.....

.....กรรมการ

(รองศาสตราจารย์ ดร.ปานใจ ธารทัศนวงศ์)

...../...../.....

.....กรรมการ

(ศาสตราจารย์ ดร.ชิดชนก เหลือสินทรัพย์)

.. ../...../.....

53307803: สาขาวิชาวิทยาการคอมพิวเตอร์และสารสนเทศ

คำสำคัญ: ชุดข้อมูลลำดับเวลา/ ประมาณค่าสูญหาย/ การแปลงค่า/จำแนกข้อมูลเชิงเวลา

เกล้ากัลยา ศิลาจันทร์ : การจำแนกข้อมูลที่ไม่สมบูรณ์ด้วยลำดับเวลาของการตรวจโรคที่แตกต่างกันจากชุดข้อมูลลำดับเวลา อาจารย์ที่ปรึกษาวิทยานิพนธ์ รศ.ดร.ปานใจ ธารทัศนวงศ์ และ ศ.ดร. ชิดชนก เหลือสินทรัพย์. 92 หน้า.

การศึกษาวิจัยในครั้งนี้ มีวัตถุประสงค์ดังนี้ 1. เพื่อพัฒนาวิธีการในการประมาณค่าสูญหายจากชุดข้อมูลทางการแพทย์ลำดับเวลาบนแนวคิดค่าตัวชี้วัดของผู้มารับการรักษาเป็นค่าเฉพาะของแต่ละบุคคลหรือจากค่าที่เหมือนหรือใกล้เคียงของแต่ละบุคคล 2. เพื่อพัฒนาวิธีการในการแปลงรูปแบบของข้อมูลจากชุดข้อมูลทางการแพทย์ลำดับเวลาแต่ยังคงประสิทธิภาพในการจำแนกประเภท ในงานวิจัยนี้ใช้ชุดข้อมูลของผู้ป่วยที่มาทำการตรวจโรคอ้วนที่ศูนย์โรคหลอดเลือดและหัวใจและเมตาลิค โรงพยาบาลรามาธิบดี จำนวน 458 คน รวม 1,215 ระเบียน และ ชุดข้อมูลผู้ป่วยโรคหลอดเลือดสมองชนิดอุดตัน จาก PKDD'02 จำนวน 93 คน รวม 3,010 ระเบียน ผลการวิจัยตามวัตถุประสงค์ที่ 1 ผู้วิจัยได้พัฒนาวิธีการในการประมาณค่าสูญหายในชุดข้อมูลเชิงเวลา 6 วิธี คือ NFDCs-DPimpute, NFDCsSlideW-DPimpute, CB-Extra-DPimpute, S-knn-DPimpute, SLLS-DPimpute และ DPimpute จากนั้นได้ทำการประเมินประสิทธิภาพ และความแม่นยำของการประมาณค่าสูญหายด้วยการประเมินค่าความคลาดเคลื่อนด้วย Normal root mean square error (NRSME) จากผลการวิจัยพบว่าวิธีการ NFDCs-DPimpute จะเป็นวิธีการที่ให้ค่าการประมาณสูญหายที่ให้ผลดีกว่าวิธีอื่น ในส่วนผลการวิจัยตามวัตถุประสงค์ข้อที่ 2 นั้น ผู้วิจัยได้นำเสนอการพัฒนาวิธีหาตัวแทนเพื่อลดมิติข้อมูลคือ วิธีการ Inner distance combination transform (IDCT) เพื่อหาตัวแทนชุดข้อมูลในลักษณะค่าเดียวในชุดข้อมูลลำดับเวลา และนำชุดตัวแทนข้อมูลนำเข้าเพื่อพิจารณาความแม่นยำในการจำแนกประเภทด้วยค่าความแม่นยำ (accuracy) และได้ทำการเปรียบเทียบประสิทธิภาพการจำแนกกับชุดข้อมูลที่เป็นตัวแทนด้วยวิธีการทางสถิติด้วยวิธีหาค่าเฉลี่ย (mean) ค่ามัธยฐาน (median) ค่าเบี่ยงแบนมาตรฐาน (standard deviation) และค่าความแปรปรวน (variance) จากการวิจัยพบว่าวิธี IDTC ให้ค่าความแม่นยำที่ดีกว่าวิธีอื่นในการหาค่าตัวแทนที่ใช้ในการจำแนกด้วยเทคนิคเหมืองข้อมูลซัพพอร์ตเวกเตอร์แมชชีน ผู้วิจัยได้นำวิธีที่ดีได้คัดเลือกไว้มาพัฒนาเป็นขั้นตอนวิธีสำหรับการประมาณค่าสูญหายและลดมิติข้อมูลลำดับเชิงเวลาทางการแพทย์รวมเรียกว่าวิธีการ NFDCs-DPimpute-IDTC

ภาควิชาคอมพิวเตอร์ บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร

ลายมือชื่อนักศึกษา..... ปีการศึกษา 2558

ลายมือชื่ออาจารย์ที่ปรึกษาวิทยานิพนธ์ 1..... 2.....

53307801: MAJOR: COMPUTER AND INFORMATION SCIENCE

KEYWORD : TEMPORAL DATA / IMPUTATION /TRANFORM/ TEMPORAL CLASSIFIER

KLAOKANLAYA SILACHAN : CLASSIFYING INCOMPLETE DATA WITH DISTINCT DIAGNOSTIC TIME SEQUENCE OF TEMPORAL MEDICAL DATA. THESIS ADVISOR: ASSOC.PROF.PANJAI TANTATSANAWONG Ph.D.,PROF.CHIDCHANOK LURSINSAP, Ph.D. 92 pp.

The objectives of this research were 1. to develop methods for data imputation of medical temporal data set based on individual indicators or on similar indicators of each individual and 2. to develop a method to transform data patterns of medical temporal data set but still maintain the performance of classification. The data used in this research are real medical data of patients from the Cardiovascula and Metabolic Center, Ramathibodi Hospital. The data includes 1,215 medical records of 458 patients diagnosed with obesity and 3,010 medical records of 93 patients diagnosed with stroke from PKDD'02. According to the result of the first research objective, the researcher developed six methods to measure the data imputation of time-series data set, namely NFDCs-DPimpute, NFDCsSlideW-DPimpute, CB-Extra-DPimpute, S-knn-DPimpute, SLLS-DPimpute and DPimpute. Then the performance and the accuracy of data imputation were evaluated by measuring the standard error with Normal root mean square error (NRSME). The result showed that among the six methods, the NFDCs-DPimpute contributed the better data imputation as compared to the other methods. According to the result of the second research objective, the researchers proposed the Inner distance combination transform (IDCT) method to find the representation in order to reduce data dimension. This method found the representation of a data set as a single value in temporal medical data and then it was measured to find the classification with accuracy. The performance of the classification of the data set of the representation with statistical analysis method, namely mean, median, standard deviation, and variance. The result showed that the IDTC method contributes the better accuracy as compared to other methods used to study with the support vector machine classification model. As a consequence, the researcher adopted the abovementioned methods and developed an algorithm called NFDCs-DPimpute-IDTC to impute data as well as to reduce the dimensions of data in temporal dataset.

Department of Computing Graduate School, Silpakorn University

Student's signature..... Academic Year 2015

Thesis Advisor's signature 1..... 2 .....

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จลุล่วงได้ด้วยความกรุณาจาก รองศาสตราจารย์ดร.ปานใจ ธารทัศนวงศ์ อาจารย์ประจำภาควิชาคอมพิวเตอร์ มหาวิทยาลัยศิลปากร และศาสตราจารย์ดร.ชิดชนก เหลือสินทรัพย์ อาจารย์ประจำภาควิชาคณนาคณิตศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ซึ่งเป็นอาจารย์ที่ปรึกษาควบคุมวิทยานิพนธ์ของผู้วิจัย ที่กรุณาให้แนวทาง ข้อแนะนำ ติดตามในการทำวิจัยของผู้วิจัยจนสำเร็จลุล่วง ผู้วิจัยรู้สึกสำนึกและกราบขอบพระคุณท่านเป็นอย่างสูงไว้ ณ โอกาสนี้

ขอกราบขอบพระคุณคณะกรรมการสอบทุกท่านที่ให้ข้อคิดและข้อแนะนำอื่นๆเพิ่มเติมที่เกี่ยวข้องประกอบการทำวิทยานิพนธ์นี้เป็นอย่างสูง

ขอกราบขอบพระคุณอาจารย์ทุกท่านในสังกัดภาควิชาคอมพิวเตอร์ ที่ให้ความรู้ในรายวิชาและข้อคิดและข้อแนะนำอื่นๆเพิ่มเติมที่เกี่ยวข้องประกอบการทำวิทยานิพนธ์นี้

ขอกราบขอบพระคุณนพ.ฆนัท ครุฑกูล ศูนย์โรคหัวใจหลอดเลือดและเมตาบอลิก โรงพยาบาลรามาธิบดีที่อนุเคราะห์ข้อมูลเป็นส่วนหนึ่งประกอบการวิจัยนี้

ขอขอบคุณมหาวิทยาลัยราชภัฏนครปฐมและสำนักงานคณะกรรมการการอุดมศึกษา(สกอ) ที่สนับสนุนการศึกษาต่อและสนับสนุนทุนการศึกษาในการศึกษาต่อและการทำวิทยานิพนธ์

ขอขอบคุณเพื่อนนักศึกษาที่เป็นกำลังใจและข้อแนะนำอย่างต่อเนื่อง ขอขอบคุณ ครอบครัว แม่พี่น้อง เพื่อนร่วมงาน ทั้งหลายที่คอยเป็นกำลังใจในการศึกษา และการจัดทำวิทยานิพนธ์ตลอดมาทำให้ผู้วิจัยสำเร็จในครั้งนี้

สารบัญ

บทที่ หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญภาพ

บทที่ 1
บทนำ

การประยุกต์ใช้การจำแนกประเภทเป็นขั้นตอนที่สำคัญสำหรับงานหลายๆ สาขาจำนวนมาก งานด้านหนึ่งคือการนำชุดข้อมูลตรวจรักษาทางการแพทย์ที่เป็นลำดับเวลาของการตรวจรักษาที่สมบูรณ์ในอดีต มาวิเคราะห์ซึ่งมีจุดมุ่งหมายคือพิจารณารูปแบบการจำแนกจากชุดข้อมูลจำนวนมากเพื่อจำแนกข้อมูลออกเป็นกลุ่มข้อมูลให้อยู่ในกลุ่มเดียวกันตามที่กำหนด ซึ่งการจำแนกกลุ่มกับการทำวิจัยด้วยวิธีการการจำแนกกลุ่มจากชุดข้อมูลในลักษณะลำดับข้อมูลเวลา ปัญหาที่น่าสนใจในงานวิจัยทางการแพทย์และทางวิทยาศาสตร์คือ กระบวนการจำแนกกลุ่มเชิงลำดับเวลา รวมทั้งปัญหาการสูญหายของข้อมูลในแต่ละตัวชี้วัด[1] เนื่องจากผู้ป่วยอาจไม่ได้มาตรวจรักษาตามลำดับเวลาโดยเฉพาะตัวแปรชี้วัดจากการตรวจรักษาแต่ละครั้งของผู้ป่วยรวมทั้งด้านข้อมูลคุณลักษณะจากจำนวนการตรวจรักษาของผู้ป่วยแต่ละคนที่มีผลต่อโครงสร้างของชุดข้อมูลและมีผลต่อขั้นตอนวิธีในการจำแนกประเภทจึงเป็นสิ่งจำเป็นที่จะต้องพิจารณาถึงแนวทางที่เหมาะสมจะช่วยเพิ่มความแม่นยำและประสิทธิภาพในการจำแนกประเภทกลุ่มข้อมูลเพื่อนำไปวิเคราะห์ข้อมูลในงานทางด้านการแพทย์ซึ่งเป็นสิ่งที่จะศึกษาในงานวิจัยนี้

1. ความเป็นมาและความสำคัญของปัญหา

ในช่วงสองทศวรรษที่ผ่าน การศึกษาวิจัยเกี่ยวกับการทำเหมืองข้อมูล (Data Mining) ได้รับความสนใจและเป็นที่นิยมอย่างมาก ทั้งวงการวิทยาการคอมพิวเตอร์สารสนเทศ รวมถึงวงการแพทย์และสาขาการวิจัยอื่นๆ อีกทั้งในปัจจุบันยังเป็นพื้นฐานของหลักการจัดการข้อมูลปริมาณมหาศาล (Big Data) อีกด้วย หลักการของการทำเหมืองข้อมูล คือ การเอาข้อมูลต่างๆ ในฐานข้อมูลมาวิเคราะห์และจัดตามกระบวนการของเหมืองข้อมูล

สำหรับการทำเหมืองข้อมูลสำหรับข้อมูลด้านการแพทย์ ทำได้โดยนำชุดข้อมูลตรวจรักษาทางการแพทย์ที่เป็นลำดับเวลา(temporal medical data) ของการตรวจรักษาที่สมบูรณ์ในอดีตมาวิเคราะห์โดยพิจารณารูปแบบการจำแนกจากชุดข้อมูลจำนวนมาก เพื่อจำแนกข้อมูลให้อยู่ในกลุ่มเดียวกันตามที่กำหนด การจำแนกดังกล่าวสามารถทำได้หลากหลายวิธี แต่วิธีที่วงการวิจัยทางการแพทย์ให้ความสนใจมากก็คือ กระบวนการจำแนกกลุ่มเชิงลำดับเวลา(temporal classifier)

ดังจะเห็นได้ชัดในงานวิจัยที่เกี่ยวข้องกับจำแนกกลุ่มกับการพิจารณาค่าข้อมูลกับชุดข้อมูลลำดับเวลา เช่นการศึกษาขั้นตอนการสกัดความรู้บนชุดข้อมูลลำดับเวลา[2] ,ขั้นตอนการจำแนกกลุ่มข้อมูลลำดับเวลาเพื่อปรับปรุงความแม่นยำโดยใช้ฐานข้อมูล meteorological จาก Texas Commission [3] ,ในงานวิจัยนี้นำเสนอเครื่องมือใหม่สำหรับการจำแนกใน machine learning กับชุดข้อมูลอนุกรมเวลา(time series data) เพื่อประโยชน์ในการค้นพบรูปแบบสำหรับการจำแนกประเภทโดยรูปแบบเหล่านี้จะรวมกันเพื่อสร้างกฎการจำแนก interpretable [4]

กระบวนการจำแนกกลุ่มเชิงลำดับเวลาจำเป็นต้องใช้ข้อมูลทางการแพทย์เชิงเวลาที่ผู้ป่วยเข้ามาตรวจรักษา โดยข้อมูลเหล่านั้นจะถูกจัดโครงสร้างของชุดข้อมูลจะเป็นลักษณะค่าข้อมูลของคนไข้เฉพาะรายบุคคลที่พิจารณาตามตัวชี้วัดที่ขึ้นกับเวลาการตรวจรักษาแบ่งเป็นการตรวจระยะสั้น (short-term) และการตรวจระยะยาว (long – term) ซึ่งเป็นคุณสมบัติค่าข้อมูลของผู้ป่วยแต่ละคนซึ่งจะนำมาพิจารณาวิเคราะห์และมีประโยชน์ในการนำไปจำแนกกลุ่มผู้ป่วยตามเงื่อนไขที่เราสนใจ อย่างไรก็ตามประเด็นหนึ่งของวิเคราะห์ข้อมูลลำดับเวลาที่น่าสนใจ คือการทำนายการจำแนกประเภทเพื่อพิจารณาความแม่นยำจากค่าข้อมูลในแต่ละตัวชี้วัดจากข้อมูลลำดับเวลา[5][6] ซึ่งการจำแนกหมวดหมู่กับข้อมูลทางการแพทย์มีบทบาทสำคัญเพื่อจำแนกกลุ่มโรค โดยปกติอัลกอริทึมทางดาต้าไมนิ่งหรือการเรียนรู้ของเครื่องจักรจะเป็นรองรับการจำแนกประเภทกลุ่มข้อมูลเพื่อพิจารณาแอทริบิวต์และรายการในการจำแนกกลุ่มข้อมูลที่จะไม่เกี่ยวข้องกับเรื่องของเวลา เป้าหมายของการจำแนกประเภทซึ่งเป็นขั้นตอนในส่วน post-processing คือการทำนายจะพิจารณาจากชุดข้อมูลที่เป็นตัวแทนของคุณลักษณะในการนำเข้าซึ่งจะต้องเตรียมคุณลักษณะของข้อมูลให้เหมาะสม(pre-processing) ในการนำเข้าข้อมูลเพื่อให้ได้ค่าการเรียนรู้(training data) การเรียนรู้จะเรียนรู้ในนำเข้าในลักษณะเรคอร์ดที่ปรากฏค่าข้อมูลแล้วทำการจับคู่เพื่อใช้ในการทำนายคำตอบหรือจากชุดข้อมูลกลุ่มข้อมูลที่ได้ ซึ่งผลลัพธ์จะได้โมเดลจำแนกประเภทและจะนำข้อมูลอีกส่วนหนึ่งมาใช้สำหรับการทดสอบ(testing) และปรับปรุงโมเดลจนกว่าจะได้ค่าความถูกต้องในระดับพอใจ เมื่อมีข้อมูลใหม่มาก็จะนำข้อมูลใหม่มาผ่านโมเดลเพื่อให้ได้โมเดลสำหรับทำนายกลุ่มการจำแนกนี้ได้ นั่นคือการเรียนรู้ในกระบวนการจำแนกประเภทเพื่อทำนายความแม่นยำของการจำแนกประเภทข้อมูลเชิงเวลาจะนำข้อมูลจริงจากข้อมูลในอดีตที่ผ่านมาจากค่าข้อมูลในแต่ละตัวชี้วัดที่ขึ้นกับเวลาของการตรวจรักษาจากข้อมูลในอดีตที่ผ่านมาเพื่อทำนายค่าที่ปรากฏในอนาคต [7][8][3] ดังนั้นชุดข้อมูลทางด้านการแพทย์จะมีเกี่ยวกับมิติเวลาเพื่อสกัดหรือดึงความสัมพันธ์ที่มีความหมายจากชุดข้อมูลดังกล่าว นั่นคือลักษณะข้อมูลลำดับเวลานี้มีแพร่หลายมากขึ้นและมี

ความสำคัญในการวิเคราะห์ข้อมูลเกี่ยวกับเวลาและลำดับของค่าข้อมูลที่ปรากฏในตัวชี้วัดแต่ละเวลาที่จะมีความสำคัญสำหรับการจำแนกกลุ่มของชุดข้อมูล[8] ดังนั้นการปรากฏหลายๆค่าข้อมูลในแต่ละตัวชี้วัดของผู้ป่วยของแต่ละคนที่ขึ้นกับเวลา[9] ซึ่งเป็นลักษณะค่าลำดับเวลาที่ต่อเนื่องจึงเป็นเรื่องที่น่าสนใจที่จะนำมาพิจารณาวิเคราะห์และจะมีประโยชน์ในการนำไปจำแนกตามเงื่อนไขที่เราสนใจ จึงเป็นสิ่งที่จำเป็นที่จะต้องพิจารณาถึงเทคนิคเฉพาะสำหรับข้อมูลลำดับเวลาในการวิเคราะห์ข้อมูล

ปัญหา

ในงานการจำแนกประเภทจากข้อมูลเชิงเวลาทางการแพทย์กำหนดให้เซตข้อมูลการตรวจรักษาของผู้ป่วยแยกตามรหัสผู้ป่วยคือ $(X_{-PID}) = \{< x_1(t_i), y_i >\}$ หรือ $(X_{-PID}) = [x_1(t), x_2(t), x_3(t), \ldots, x_n(t)]$ เมื่อ X_{-PID} คือ เซตข้อมูลตัวชี้วัดของผู้ป่วยแต่ละคน , $x_1(t)$ คือตัวชี้วัดที่ระบุข้อมูลการตรวจรักษาตามลำดับเวลา(t) และ t เป็นลำดับเวลาการเข้ามารับการตรวจรักษาแต่ละครั้ง , i = 1,2,..n ส่วน y_i แทนค่าคลาสคำตอบในการระบุกลุ่มโรคซึ่งจะสัมพันธ์กับตัวชี้วัด x_i ในเวลา t ตามแนวคิดนี้มีวัตถุประสงค์เพื่อใช้ประโยชน์จากข้อมูลที่มีอยู่ของผู้ป่วยในอดีตที่ผ่านมาสำหรับนำเข้าโมเดลในการเรียนรู้ในการจำแนกประเภทแต่ยังคงความแม่นยำในการจำแนก อย่างไรก็ตามกระบวนการการทำนายการจำแนกประเภทจากข้อมูลเชิงลำดับเวลา ซึ่งมีปัญหาเกี่ยวกับค่าข้อมูลสำหรับนำเข้าบางประการที่จะต้องพิจารณาที่จะมีผลต่อการจำแนกประเภทลักษณะชุดข้อมูลลำดับเวลาดังนี้คือ

ปัญหา 1 : ปัญหาหลักสำหรับข้อมูลนำเข้าเพื่อการจำแนกประเภทคือการปรากฏค่าสูญหาย เป็นปัญหาสำคัญที่มีผลกระทบต่อความแม่นยำในการจำแนกกลุ่มข้อมูล เพราะข้อมูลในการนำเข้าโมเดลการจำแนกประเภทควรจะต้องเป็นค่าข้อมูลที่ครบสมบูรณ์ ดังนั้นการปรากฏค่าสูญหายในแต่ละตัวชี้วัดเมื่อปรากฏค่าข้อมูลที่เราไม่ทราบค่า ทำให้ของชุดข้อมูลนำเข้าไม่สมบูรณ์เป็นปัญหาในการจำแนกประเภททั้งแบบคงที่และแบบลำดับเวลา มีเทคนิคและงานวิจัยที่ใช้ในการประมาณและทดแทนค่าสูญหายมีหลายเทคนิควิธีเพื่อรองรับการให้ความสำคัญของกระบวนการเตรียมข้อมูลก่อนการประมวลผลวิเคราะห์จำแนกข้อมูล โดยทั่วไปจะแบ่งเป็น 2 รูปแบบคือ เทคนิคเชิงสถิติและเทคนิคทางด้านเหมืองข้อมูลเช่น K-nearest neighbors(K-NNimpute),SVD method(SVD impute), least square imputation(LSimpute), Bayesian principle

component analysis(BPCA)[9], และวิธีการอื่นๆ เพื่อวิเคราะห์ปัญหาค่าที่สูญหาย[7] เช่น วัตถุประสงค์งานวิจัยนี้ เพื่อพิจารณาปัญหาการปรากฏค่าสูญหายในรูปแบบของงานการจำแนกประเภท และสรุปหรือเปรียบเทียบกับขั้นตอนที่รู้จักกันดีเพื่อแก้ปัญหาค่าสูญหาย [34] ตามลักษณะของชุดข้อมูลที่แตกต่างกัน แต่หากต้องการประมาณและทดแทนค่าสูญหายในลักษณะข้อมูลในรูปแบบลำดับเวลา ควรจะต้องใช้เทคนิควิธีการที่เหมาะสมเพื่อให้ได้ค่าการประมาณที่ยอมรับได้

ปัญหา 2 : การจำแนกประเภทเราสามารถเลือกนำเฉพาะชุดข้อมูลตัวชี้วัดทั้งหมด โดยไม่สนใจเรื่องของตัวแปรเวลาเพื่อนำเข้าโมเดลทำการจำแนกประเภท แต่อาจเป็นปัญหาหนึ่งที่สำคัญของการทำเหมืองข้อมูลคือหากขนาดของชุดข้อมูลมีขนาดกลางถึงขนาดใหญ่ จะทำให้วิธีการที่ใช้ในการสกัดความรู้เมื่อนำเข้าโมเดลการจำแนกประเภทใช้เวลาในการเรียนรู้นาน

1.1 แนวทางการแก้ปัญหาและขั้นตอนที่นำเสนอ

ในงานวิจัยฉบับนี้ ผู้วิจัยได้ใช้ชุดข้อมูลทดลองจากชุดข้อมูลผู้ป่วยโรคอ้วน(Obesity data) และชุดข้อมูลผู้ป่วยโรคหลอดเลือดสมองชนิดอุดตัน(Thrombosis) มุ่งเน้นพัฒนาวิธีการในส่วนของการเตรียมข้อมูล(pre-processing) ของชุดข้อมูลลำดับเวลา ก่อนนำเข้ากระบวนการจำแนกกลุ่ม (post-processing) ซึ่งเป็นขั้นตอนที่จะสามารถปรับปรุงคุณภาพโดยรวมของรูปแบบของข้อมูลก่อนนำเข้าเพื่อทำนายการจำแนกประเภทต่อไป โดยมีแนวทางการแก้ปัญหาข้างต้นดังนี้ คือ พัฒนาวิธีการเพื่อประมาณค่าสูญหายในชุดข้อมูลลำดับเชิงเวลาจากชุดข้อมูลที่ไม่สมบูรณ์ การประมาณค่าข้อมูลสามารถคำนวณเพื่อทำนายค่าข้อมูลจากโมเดลการประมาณค่าสูญหายของผู้ป่วยแต่ละรายบนแนวคิดวิธีการซึ่งจะเป็นการนำค่าเฉพาะรายบุคคลและจากค่าที่เหมือนหรือใกล้เคียงกันเพื่อประมาณการให้ได้ข้อมูลที่มีค่าใกล้เคียงที่ยอมรับจากการประเมินประสิทธิภาพ เมื่อค่าข้อมูลครบสมบูรณ์แล้วเพื่อลดเวลาในการเรียนรู้ในชุดข้อมูล แต่ยังคงความแม่นยำในการจำแนกประเภท จึงมีแนวคิดคือทำการลดขนาดชุดข้อมูล(Data Reduction) ซึ่งในงานวิจัยนี้มุ่งเน้นในลักษณะ Datasize reduction ด้วยการใช้แถวเป็นหลักในการลดข้อมูล ผสานรวมกับหลักการแปลงชุดข้อมูล(feature transformation) โดยรวมกลุ่มข้อมูลและแปลงข้อมูลเพื่อเปลี่ยนแปลงรูปแบบของข้อมูลจากรูปแบบหนึ่งไปเป็นอีกรูปแบบหนึ่งการลดมิติชุดข้อมูลโดยใช้เทคนิคการแปลงรูปแบบของข้อมูล

หรือเรียกว่าการทรานฟอร์มการเปลี่ยนลักษณะข้อมูลหรือการรวมข้อมูลการแปลงรูปแบบของ (Data Transformation) เป็นการแปลงข้อมูลที่เลือกมาให้อยู่ในรูปแบบที่เหมาะสมสำหรับการนำไปใช้วิเคราะห์ตามอัลกอริทึม(Algorithm) และแบบจำลองที่ใช้ในการทำเหมืองข้อมูลต่อไป เพื่อให้ขั้นตอนการทำเหมืองข้อมูลที่เกิดขึ้นอาจจะมีประสิทธิภาพมากขึ้น ดังนั้นเพื่อแปลงค่าข้อมูลจากลำดับค่าข้อมูลการตรวจรักษาของผู้ป่วยแต่ละคนในแต่ละตัวชี้วัดที่ขึ้นกับเวลาของการตรวจรักษาของผู้ป่วยในลักษณะที่เป็นรายบุคคล(X(t)) จากชุดข้อมูลลำดับเวลาที่สมบูรณ์มีค่าครบให้ได้ค่าเป็นค่าเดี่ยวเฉพาะ(singular value) ผลของการแปลงจะทำให้ได้ข้อมูลชุดใหม่ (X'(t)) ที่จะเป็นตัวแทนของข้อมูลลำดับเวลาทั้งหมดของผู้ป่วยแต่ละคนมาใช้แทนข้อมูลทั้งหมดของแต่ละบุคคล เพื่อนำเข้าทำนายการจำแนกประเภทด้วยอัลกอริทึมสแตติกของการจำแนกประเภทและวัดประสิทธิภาพของจำแนกประเภทต่อไป

ดังนั้นงานวิจัยนี้จึงมุ่งเน้นพัฒนาวิธีการด้วยความคาดหวังว่า จะได้วิธีการประมาณค่าสูญหายในชุดข้อมูลทางการแพทย์เชิงลำดับเวลาและสำหรับการลดมิติข้อมูลในชุดข้อมูลลำดับเวลาในลักษณะข้อมูลรายบุคคลที่ยังคงประสิทธิภาพในการจำแนกประเภทข้อมูล

2. วัตถุประสงค์งานวิจัย

จุดมุ่งหมายหลักสำหรับงานวิจัย มีดังต่อไปนี้

2.1 เพื่อพัฒนาวิธีการในการประมาณค่าสูญหายจากชุดข้อมูลทางการแพทย์ลำดับเวลาบนแนวคิดค่าตัวชี้วัดของผู้มารับการรักษาเป็นค่าเฉพาะของแต่ละบุคคล หรือจากค่าที่เหมือนหรือใกล้เคียงของแต่ละบุคคล

2.2 เพื่อพัฒนาวิธีการในการแปลงรูปแบบของข้อมูลจากชุดข้อมูลทางการแพทย์ลำดับเวลาแต่ยังคงประสิทธิภาพในการจำแนกประเภท

3. ขอบเขตของการศึกษา

3.1 สำหรับการทดลองในทุกวิธีการได้ทดลองกับชุดข้อมูล วิธีการที่ทำการทดลองและประเมินบนชุดข้อมูลลำดับทางการแพทย์เชิงเวลา จำนวน 2 ชุดข้อมูล คือ obesity data จากศูนย์โรคหลอดเลือดและหัวใจฯ โรงพยาบาลรามาธิบดี และ thombiosisdata จาก Padk'dd[40]

3.2 ทำการศึกษาวิจัยการประมาณค่าสูญหายด้วยชุดข้อมูลทางการแพทย์เชิงเวลาดังข้างต้น ด้วยสมมติฐานว่า ค่าตัวชี้วัดของคนไข้เป็นค่าเฉพาะของแต่ละบุคคลและค่าที่เหมือนหรือใกล้เคียงของแต่ละตัวบุคคลเพื่อนำข้อมูลชุดใหม่เป็นชุดข้อมูลนำเข้าที่ได้ไปนำเข้าทำการจำแนกประเภทและการแปลงค่าข้อมูลด้วยการทรานฟอร์มชุดข้อมูลในลักษณะผู้ป่วยรายบุคคล

3.3 ทำนายการจำแนกประเภทด้วยโมเดลการจำแนกแบบคงที่(static classifier) จากชุดข้อมูลในข้อ 3.2

3.4 วัดและเปรียบเทียบประสิทธิภาพความแม่นยำของการจำแนกประเภทจากชุดข้อมูลสมบูรณ์กับชุดข้อมูลจากค่าการประมาณและชุดข้อมูลที่ประมาณค่าสูญหายและจัดการมิติด้วยการทรานฟอร์มค่าข้อมูล

4. ประโยชน์ที่คาดว่าจะได้รับ

ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย ทำให้ได้กระบวนการประมาณค่าสูญหายในชุดข้อมูลทางการแพทย์เชิงลำดับเวลาและสำหรับการแปลงค่าข้อมูลในชุดข้อมูลลำดับเวลาในลักษณะข้อมูลรายบุคคลที่ยังคงประสิทธิภาพในการจำแนกประเภทข้อมูล

บทที่ 2
แนวคิดทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในงานวิจัยนี้มุ่งเน้นเพื่อการวิเคราะห์ข้อมูลทางการแพทย์จากข้อมูลลำดับเวลาซึ่งต้องใช้วิธีการเฉพาะกับค่าข้อมูลในรูปแบบลำดับเวลา ซึ่งวิธีการจะประยุกต์ใช้เพื่อแก้ปัญหาการจำแนกประเภทกับชุดข้อมูลทางการแพทย์ลำดับเวลาที่มีค่าจำนวนการตรวจที่แตกต่างกันและปรากฏค่าขาดหาย ดังนั้นผู้วิจัยได้ทำการศึกษา เพื่อค้นคว้าทบทวนเอกสารงานวิจัยในอดีตที่เกี่ยวข้อง และได้กล่าวถึงทฤษฎีต่างๆ ที่เกี่ยวข้องกับการพัฒนาขั้นตอนการประมาณและทดแทนค่าสูญหายกับหลักการที่เกี่ยวข้องกับการทรานฟอร์ม

2.1 ชุดข้อมูลทางการแพทย์ลำดับเชิงเวลา(Temporal medical data)

การสกัดความรู้จากชุดข้อมูลลำดับเวลา เป็นการใช้ประโยชน์จากข้อมูลลำดับเวลาในงานทางด้านการแพทย์และสุขภาพให้ความสำคัญกับลักษณะข้อมูลลำดับเวลาเพิ่มมากขึ้น การทำเหมืองข้อมูลเชิงเวลาครอบคลุมการวิเคราะห์ข้อมูลที่เกี่ยวข้องเรื่องของเวลา ในการรวบรวมข้อมูลจำนวนมากข้อมูลจะถูกเก็บรักษาสำหรับหลายจุดเวลา [10][11] ข้อมูลนี้เป็นข้อมูลเชิงเวลาและสัมพันธ์กับจุดของเวลาที่เฉพาะเจาะจง ซึ่งฐานข้อมูลเชิงเวลามักจะอยู่รูปแบบฟอร์ม<รหัสผู้ป่วย, วันที่ตรวจรักษา, ตัวแปรข้อมูล> [12] ลำดับที่ในชุดข้อมูลมักเรียกว่า ลำดับเชิงเวลา ซึ่งในทางการแพทย์และโรงพยาบาล มีข้อมูลที่เกิดขึ้นมาจากระบบรวบรวมข้อมูลที่เกี่ยวกับข้อมูลสุขภาพผู้ป่วยหรือข้อมูลเกี่ยวกับการตรวจรักษาของผู้ป่วยซึ่งเป็นชุดข้อมูลที่ทำการรวบรวมข้อมูลในอดีตซึ่งก็คือข้อมูลที่ผ่านมาถึงปัจจุบันนั่นก็คือเรียกว่าชุดข้อมูลลำดับเวลานั่นเอง [6][10][13] การศึกษาถึงโครงสร้างของตัวแปรหรือตัวชี้วัดเพื่อจะประกอบการพิจารณาการแทนค่าหรือจำแนกกลุ่มโรคในแนวคอลัมน์หรือแนวแถวจึงจะเหมาะสม ดังตัวอย่างตารางที่ 2.1 แสดงรายละเอียดโครงสร้างคุณลักษณะข้อมูลลำดับมิติเวลา [14]

ตารางที่ 2.1 แสดงรายละเอียดโครงสร้างคุณลักษณะข้อมูลลำดับมิติเวลา

	ตัวแปร(Variable)		
Time point	X_1	……..	X_n
1	$X_{1,1}$	……..	$X_{n,1}$
2	$X_{1,2}$	……..	$X_{n,2}$
…….	…….	…….	……...
T	$X_{1,T}$	…….	X_n,T

ลำดับประกอบด้วยชุดของสัญลักษณ์ที่ระบุจากตัวอักษร โดยเฉพาะอย่างยิ่งมักจะเรียกว่าลำดับเวลา(time sequence) และลำดับของอย่างต่อเนื่อง รูปแบบข้อมูลลำดับเวลา $X = \{X_T, t = 1,2, .. ,n\}$ เป้าหมายคือการสกัดความรู้ที่ซ่อนอยู่ที่มีลักษณะของเหตุการณ์ที่เกิดขึ้นใน X โดยกำหนดให้ X_t คือ ตัวชี้วัดจุดเวลา : $(t = 1,2, .., t)$ [14]

ตารางที่ 2 .2 แสดงตัวอย่างชุดเรียนรู้แสดงตัวชี้วัดข้อมูลกับค่าข้อมูลขาดหาย

Case-Patien#	Time - Treatement	Weight	BMI	BMR	SMM	…….	……	Protien (g)	X_N	Class label
1	1	x	X	-	X			x		1
1	2	x	-	-	x			-		1
1	3	-	x	x	-			x		1
….	..									..
2	1	x	-	X	X			x		0

2.2 แนวคิดและงานวิจัยเกี่ยวกับการประมาณค่าสูญหายกับชุดข้อมูลต่างๆ

เทคนิคและงานวิจัยที่ใช้ในการประมาณและทดแทนค่าสูญหายมีหลายเทคนิควิธี เพื่อรองรับการให้ความสำคัญของกระบวนการเตรียมข้อมูลก่อนการประมวลผลวิเคราะห์จำแนกข้อมูล โดยทั่วไปจะแบ่งเป็น 2 รูปแบบ คือ เทคนิคเชิงสถิติและเทคนิคทางด้านเหมืองข้อมูล เช่น K-nearest neighbors(K-nnimpute) [15], SVD method(SVD impute) [15], Least square imputation (LSimpute) [16],Bayesian principle component analysis(BPCA) [17], และขั้นตอนวิธีการอื่นๆ

เพื่อวิเคราะห์ปัญหาค่าที่สูญหายไปในรูปแบบของการจำแนกประเภท [18][17] เพราะจะมีข้อบกพร่องบางอย่างเมื่อนำไปประยุกต์ใช้ในงานการจำแนกประเภท

สำหรับงานวิจัยที่เกี่ยวข้องการประมาณค่าในงานทางด้านการแพทย์ ดังเช่นงานวิจัยดังต่อไปนี้ J.F Rodick et.al [19] ได้ข้อสรุปว่า วิธีการขึ้นอยู่กับเทคนิคการเรียนรู้ซึ่งได้รับพบว่าจะเหมาะสำหรับการใส่ค่าที่หายไปจากและนำไปสู่การเพิ่มประสิทธิภาพอย่างมีนัยสำคัญของความถูกต้องโรคเมื่อเทียบกับวิธีการการประมาณค่าด้วยค่าทางสถิติสำหรับขั้นตอนอื่น Data Analysis and Statistical Software (STATA v.10) [1] จะใช้ในการหาข้อมูลที่หายไปในฐานข้อมูลผู้ป่วยคะแนนต่ำสุด [19], M.N Noraziana [20] นำเสนอขั้นตอนที่มีประสิทธิภาพสำหรับการสกัดข้อมูลในตำแหน่งที่หายไป โดยการแทนที่ค่าแต่ละตัวแปรกับการหาค่ากลางระหว่างสองจุดก่อนและหลังค่าที่สูญหายหาค่ากลางของสองจุดนั่นคือจุดก่อนและหลัง จากข้อมูลชั่วโมงบันทึกการตรวจสอบประจำปีสำหรับชุดทดสอบที่ประกอบด้วยอนุภาคความเข้มข้น PM10 ใน Seberang Pera, ปีนัง มาเลเซีย ด้วยเทคนิค linear, quadralic, cubic และ nearest neighbor, S.Bose et al. [21] นำเสนอผลการทดลองจากชุดข้อมูลไมโครอาเรย์ที่สามารถแยกหลายค่าการแสดงออกที่หายไป มีหลายๆ ขั้นตอนวิธีในการวิเคราะห์การแสดงออกของยีนด้วยเมทริกซ์ค่าข้อมูลนำเข้าที่สมบูรณ์ของค่าอาร์เรย์ยีน ดังนั้นการประมาณค่าที่มีประสิทธิภาพของค่าที่ขาดหายไปเป็นสิ่งจำเป็นเพื่อลดผลกระทบของชุดข้อมูลที่ไม่สมบูรณ์ในการวิเคราะห์และการเพิ่มช่วงของข้อมูลชุด ซึ่งขั้นตอนวิธีการเหล่านี้สามารถนำมาใช้ ขั้นตอนใหม่ในการประมาณค่านำเสนอเพื่อทำนายค่าในตำแหน่งที่สูญหาย นำเสนอขั้นตอนเพื่อเลือกชุดยีนชุดย่อยและตัวอย่างของค่าความเหมือนในแต่ละตำแหน่งและประยุกต์ใช้ขั้นตอนใหม่ในการประมาณค่าในช่วง เพื่อทำนายตำแหน่งที่สูญหาย, Viana et al. [22] มุ่งเน้นไปที่ขั้นตอนการเตรียมข้อมูลของชุดข้อมูลเครื่องโทรศัพท์มือถือดาวเทียม เพื่อการทำนายการย่อยสลายของพลังงานแสงอาทิตย์ โดยเฉพาะอย่างยิ่งปัญหาการสูญเสียข้อมูลเนื่องจากค่าที่ขาดหายไปที่จะต้องทำการแก้ไข, J.M Jerez et al. [15] นำเสนอวิธีการที่ขึ้นอยู่กับเทคนิคการเรียนรู้เครื่องสำหรับประมาณค่าข้อมูลในฐานข้อมูลทางการแพทย์ ผู้วิจัยสรุปว่าวิธีการขึ้นอยู่กับเทคนิคการเรียนรู้เครื่องจักร(machine learning) ที่จะเหมาะสำหรับการประมาณค่าของค่าที่ขาดหายไปที่เหมาะสมตัวอย่างเช่น Multi-layer perceptron, self organizing maps และ k-nearest neighbor (k-NN) ประสิทธิภาพและความถูกต้องเปรียบเทียบกับวิธีการประมาณค่าหลักสถิติ นั่น

คือ mean values, hot-deck และ multiple imputation ขั้นตอนเหล่านี้จะถูกนำมาใช้เพื่อประมาณค่า absent values ใน “El Alamo-I” สำหรับชุดข้อมูลที่เป็นมะเร็งเต้านม(breast cancer data) จำนวน 3679 เรคอร์ดจากหลายๆ โรงพยาบาลที่แตกต่างกัน และ Spanish Breast Cancer Research Group (GEICAM), Eisemann et al. [23] วิจัยบนชุดข้อมูล malignant melanoma และมะเร็งเต้านม(female breast cancer) จาก Schleswig-Holstein Cancer Registry ในประเทศเยอรมัน กรณีที่มีข้อมูลขั้นตอนเนื้องอกที่สมบูรณ์ถูกสกัดและข้อมูลขั้นตอนของพวกเขาบางส่วนจะถูกลบออกตามรูปแบบของ MAR ค่าสูญหายใน tumorstage จะถูกประมาณค่าด้วยวิธีการ multiple imputation ด้วยสมการchained, polynomial regression, predictive mean matching, random forests และ proportional sampling ซึ่งค่าข้อมูลจริงมีความแม่นยำใกล้เคียงมากที่สุดประมาณโดยขั้นตอนวิธี polynomial regression และทำนายค่าด้วยค่า mean matching ส่วน random forests และ proportional sampling มีความแม่นยำน้อยที่สุด, ShusakuTsumoto [9] นำเสนอขั้นตอนที่เรียกว่า CEARI เป็นการรวมระหว่างขั้นตอนวิธี extended moving average และ rule induction เพื่อค้นพบข้อมูลใหม่ในฐานข้อมูลเชิงเวลา(temporal database) CEARI ถูกประยุกต์ใช้ในฐานข้อมูลทางแพทย์ในโรคเซลล์ประสาท(motor neuron disease) ที่ข้อมูลจะปรากฏค่าสูญหาย

2.3 ทฤษฎีเกี่ยวกับพหุนามและการประมาณค่าในช่วงและฟังก์ชัน [10] [18][24] [25][26]

หลักการประมาณค่าในช่วงและฟังก์ชันมีหลายวิธี จึงอธิบายวิธีที่เกี่ยวข้องในงานวิจัยที่ใช้ในการประมาณค่าในงานวิจัยนี้นำเสนอแบ่งเป็น 3 กลุ่มที่เกี่ยวข้องในงานวิจัย คือ

2.3.1 การประมาณค่าในช่วง(Interpolating) กับชุดข้อมูลที่ทราบค่าด้วย Newton’s Interpolation , cubic splines

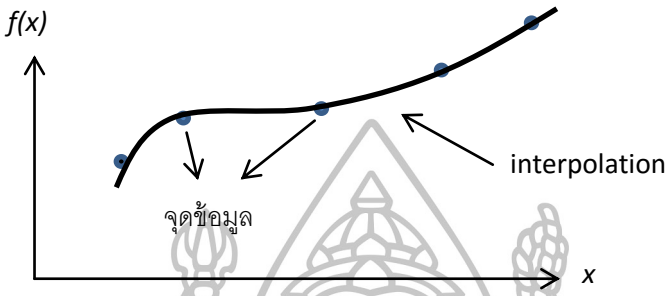
2.3.2 การประมาณค่านอกช่วง(Extrapolating) กับชุดข้อมูลที่อยู่นอกช่วง

2.3.3 หลักการกำลังสองน้อยสุดด้วย least squares

2.3.1 การประมาณค่าในช่วง (Interpolating) กับชุดข้อมูลที่ทราบค่า: [10] [18][24] [25][26]

การประมาณค่าในช่วง(interpolation) คือ การสร้างเส้นโค้งที่ลากผ่านจุดทุกจุดในช่วงข้อมูลและประมาณค่าจุด (x, f(x)) บนเส้นโค้งนั้น การมองข้อมูลเป็นจุดกับค่าข้อมูลที่ทราบค่า

เพื่อประมาณค่าข้อมูลในตำแหน่งหรือเวลาที่เราไม่ทราบค่า โดยจะให้ค่าข้อมูลในแต่ละตัวแปรในชุดของจุด เราสามารถใช้ขั้นตอนในการประมาณค่าในช่วงเพื่อทำนายค่าข้อมูลของแต่ละตัวแปรในตำแหน่งของทุก ๆ จุดดังภาพที่ 2.1



ภาพที่ 2.1 แสดงรูปแบบการประมาณค่าในช่วง(ที่มา : [26])

โครงสร้างของการประมาณค่าในช่วงคือฟังก์ชันที่อยู่ระหว่างจุดที่เราทราบค่า ดังนั้นการประยุกต์ใช้เทคนิคการประมาณค่าในช่วง เพื่อนำมาประมาณค่าสูญหายจะเป็นการหาฟังก์ชั่นที่จะประมาณค่าจากจุดที่เรารู้ ($x_1, x_2, \ldots, x_n$) เรียกว่า จุดหรือตำแหน่งประมาณค่าในช่วง สำหรับระยะระหว่างคู่ของแต่ละจุดเป็นค่าคงที่เรา $x_i=x(t_i)$ สำหรับ i = 1, 2, ... ,n, ที่อยู่, (x (t_1) x (t_2), ...,x (T_N)) จะประมาณบนฟังก์ชัน f (x) บนจุดที่ไม่รู้จัก ในขณะที่ฟังก์ชั่นการประมาณค่าจะช่วยให้การคำนวณของฟังก์ชัน f (x) ที่จุดที่ต้องการในการพัฒนาการประมาณค่าที่ใกล้เคียงที่สุดฟังก์ชัน f (x) ที่คาดว่าจะกำหนดค่าของ x ที่ประมาณเส้นโค้งเรียบบนโดเมนทั้งหมดของฟังก์ชัน f (x) ดังนั้น x มีค่าระหว่าง x_1 และ x_n ในชุดข้อมูล

การประมาณค่าในช่วงข้อมูล หมายถึง เป็นประมาณค่าระหว่างจุดสองจุดและลากกราฟหาฟังก์ชั่นระหว่างจุดนั้น เพื่อประมาณค่าข้อมูลในตำแหน่งที่สามระหว่างจุดสองจุด ซึ่งเรียกว่าการประมาณฟังก์ชันจากชุดข้อมูล ในการประมาณตำแหน่งของจุดที่สามในระหว่างจุดสองจุดที่กำหนด ที่ช่วยให้เราสามารถประมาณการรูปแบบการทำงานจากข้อมูล ซึ่งเป็นกระบวนหาฟังก์ชัน (โดยส่วนใหญ่เป็นฟังก์ชันพหุนาม) ที่กราฟของฟังก์ชันนั้นผ่านจุด (x,y) ซึ่งการลากเส้นกราฟ (ฟังก์ชั่น) จะผ่านจุดทุกจุดของข้อมูล (n+1) จุด การกำหนดชุดของจุดจะง่ายต่อการคำนวณพหุนามที่ผ่านทุกจุดโดยแบ่งช่วงทั้งหมดออกเป็นช่วงย่อยๆแล้วสร้างพหุนามประจำแต่ละช่วงย่อยเหล่านั้น ซึ่งการสร้างฟังก์ชั่นและประมาณค่าจุดข้อมูลลักษณะนี้สิ่งนี้เรียกว่า “การประมาณโดยพหุนามเป็นช่วงๆ” ส่วนฟังก์ชั่นที่ใช้ในกรณีนี้เรียกว่าการประมาณค่าในช่วงเชิงพหุนามหรือพหุนามเสมือนโดยการประมาณระหว่างจุดข้อมูลอยู่ในรูปแบบฟังก์ชันพหุนาม ดังสมการที่ (1)

$$f(x) = b_1 + b_2 x + b_3 x^2 + \cdots + b_n x^n \qquad \text{..... (1)}$$

โดยมีรายละเอียดดังนี้

2.3.1.1 การประมาณค่าในช่วงการประมาณค่าในช่วงฟังก์ชันพหุนามโดยทั่วไป (Polynomial interpolation) ด้วยวิธีผลต่างการแบ่งย่อยของนิวตัน(Newton's Divided Difference interpolating polynomials method)

ขั้นตอนวิธีนี้คือวิธีการที่จะสร้างพหุนามผ่านจุดข้อมูลทั้งหมดที่มี หากมี 1 + n (จาก 0 ถึง n) ค่าข้อมูลพหุนาม P_n ของดีกรี n สามารถที่จะสร้างผ่านทุกจุด ด้วย n+1 จุดข้อมูล

$$(x_1, y_1=f(x_1)), \ldots, (x_n, y_n=f(x_n)),$$

วิธีการของนิวตันประกอบด้วยการแก้สมการพร้อมกันที่เป็นผลจากการคำนวณพหุนามผ่านค่าของข้อมูล มันเป็นเรื่องง่ายที่จะกำหนดพหุนาม interpolating ถ้าสร้างมันในรูปแบบดังต่อไปนี้

$$P_n(x) = b_0 + b_1(x-x_0) + b_2(x-x_0)(x-x_1) + \ldots + b_n(x-x_0)(x-x_1)\cdots(x-x_{n-1}) \qquad \text{.....(2)}$$

สำหรับเงื่อนไขของการประมาณค่าในช่วง $P_n(x_i) = f(x_i)$ ค่าของผลต่างจากการแบ่งย่อยดังแสดงในสมการ (1.5.1) – (1.5.3) นั้น สามารถทำการคำนวณแบบสืบเนื่องไปทีละขั้น(Divide Difference) ดังแสดงโดยตาราง 1 Newton's form.

ตารางที่ 2.3 แสดงการคำนวณแบบผลต่างสืบเนื่องอย่างมีขั้นตอนเพื่อหาค่าผลต่างจากการแบ่งย่อย

x_i	$f(x_i)$	ผลต่างของการแบ่งย่อยครั้งที่1	ผลต่างของการแบ่งย่อยครั้งที่ 2	ผลต่างของการแบ่งย่อยครั้งที่ 3
x_1	$f(x_1)$	$f[x_2,x_1]$	$f[x_3,x_2,x_1]$	$f[x_4, x_3, x_2, x_1]$
x_2	$f(x_2)$	$f[x_3,x_2]$	$f[x_4,x_3,x_2]$	
x_3	$f(x_3)$	$f[x_4,x_3]$		
x_4	$f(x_4)$			

(ที่มา : [26])

ลักษณะของฟังก์ชันในรูปแบบของฟังก์ชันพหุนามที่แสดงในรูป (1) สามารถเขียนได้ในรูปแบบดังนี้

$$f_{n-1}(x) = b_1 + b_2(x-x_1) + \ldots\ldots\ldots + b_3(x-x_1)(x-x_2) + b_4(x-x_1)(x-x_2)(x-x_3) + \cdots + b_n(x-x_1)(x-x_2)\cdots(x-x_{n-1}) \quad \ldots\ldots\ldots(3)$$

โดยค่าสัมประสิทธิ์ b_i , i = 0, 1, 2,……,n สามารถคำนวณได้

$$\begin{aligned} b_1 &= f(x_1) \\ b_2 &= f[x_2, x_1] \\ b_3 &= f[x_3, x_2, x_1] \\ &\vdots \\ b_n &= f[x_n, x_{n-1}, \cdots, x_2, x_1] \end{aligned}$$

โดยวงเล็บสี่เหลี่ยม [ ] แสดงถึงผลต่างการแบ่งย่อย(divided difference) เช่น

ผลต่างจากการแบ่งย่อยครั้งที่ 1 (First Deviation) คือ

$$f[x_i, x_j] = \frac{f(x_i) - f(x_j)}{x_i - x_j} \quad \ldots\ldots\ldots(4)$$

ผลต่างจากการแบ่งย่อยครั้งที่ 2 (Second Deviation) คือ

$$f[x_i, x_j, x_k] = \frac{f[x_i, x_j] - f[x_j, x_k]}{x_i - x_k} \quad \ldots\ldots\ldots(5)$$

ผลต่างของการแบ่งย่อยครั้งที่n (nth finite divided difference) คือ

$$f[x_n, x_{n-1}, \ldots, x_1, x_0] = \frac{f[x_n, x_{n-1}, \ldots, x_1] - f[x_{n-1}, x_{n-2}, \ldots, x_0]}{x_n - x_0} \quad \ldots\ldots\ldots(6)$$

$$f_n(x) = \sum_{i=1}^{n} \left\{ F[x_1, x_2, \ldots, x_i] \prod_{j=1}^{i-1} (x - x_j) \right\}$$

การหาของผลต่างจากการแบ่งย่อยดังแสดงในสมการ(1.5.1) – (1.5.3) นั้น สามารถทำการคำนวณแบบสืบเนื่องไปทีละขั้น

ค่าที่คำนวณได้เมื่อแทนกลับลงไปในสมการ (1) ทำให้ได้ฟังก์ชันพหุนาม

$$f_n(x) = f(x_1) + (x - x_1) f[x_2, x_1] + (x - x_1)(x - x_2) f[x_3, x_2, x_1] \quad \text{.....(7)}$$
$$+ \cdots + (x - x_1)(x - x_2) \cdots (x - x_{n-1}) f[x_n, x_{n-1}, \cdots, x_1]$$

โดยหลักการในภาพรวม คือ การหาฟังก์ชัน *f(x)* ฟังก์ชันหนึ่งจากข้อมูลตามตำแหน่งต่างๆ ทั้งหมดที่กำหนดมาให้โดยแบ่งข้อมูลออกเป็นช่วง ๆ แต่ละช่วงจะได้ดีกรีต่ำสุดซึ่งพหุนามเรียงต่อกัน จะได้ตัวแทนของข้อมูลชุดนี้

2.3.1.2 การประมาณค่าในช่วงเส้นโค้ง (Spline interpolation)

cubic spline เป็นวิธีการที่เหมาะสมขั้นตอนหนึ่งสำหรับประมาณค่าข้อมูลระหว่างที่อยู่ระหว่างกัน โดยสมมติว่ากลุ่มของจุดที่เป็นที่รู้จักคือ (x_0, y_0), (x_1, y_1), ..., (x_n, y_n), โดย cubic spline จะทำการประมาณค่าที่สูญหายด้วยค่าข้อมูลจริงที่ปรากฏเพื่อประมาณค่าระหว่างช่วงด้วยค่าคงที่ โดยแบ่งช่วงทั้งหมดออกเป็นช่วงย่อยๆ แล้วสร้างพหุนามประจำแต่ละช่วงย่อยเรียกว่า “การประมาณโดยพหุนามเป็นช่วงๆ” ซึ่ง cubic spline เป็นการใช้พหุนามกำลังสามระหว่างคู่ของจุดที่แทนค่าด้วยข้อมูล

$$y = f_3(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 \quad \text{.....(8)}$$

พหุนามกำลังสามมีค่าคงตัว 4 ค่า

โดยให้ฟังก์ชัน f นิยามบน [a,b] และมีเซตของจุด $a = x_0 < x_i < \ldots < x_n = b$ ตัวประมาณกำลังสาม S ของ f คือฟังก์ชันที่สอดคล้องตามเงื่อนไขต่อไปนี้

1. S เป็นพหุนามกำลังสาม เขียนแทนด้วย S_j สำหรับช่วงย่อย
   $[x_j, x_{j+1}]$, $j = 0, 1, \ldots, n-1$
2. $S_{j+1}(x_{j+1}) = S_j(x_{j+1})$ *(j = 0,1,.....,n)*
3. $S_{j+1}(x_{j+1}) = S_j(x_{j+1})$ *(j = 0,1,.....,n-2)*
4. $S'_{j+1}(x_{j+1}) = S'_{j+1}(x_{j+1})$ *(j = 0,1,.....,n-2)*
5. $S''_{j+1}(x_{j+1}) = S''_{j+1}(x_{j+1})$ *(j = 0,1,.....,n-2)*

6. เซตของเงื่อนไขขอบเขตข้อใดข้อหนึ่งต่อไปนี้เป็นจริง
   a. $S''(x_0) = S''(x_n) = 0$ (ขอบธรรมชาติหรือขอบอิสระ)
   b. $S'(x_0) = f'(x_0)$ และ $S'(x_n) = f'(x_n)$ (ขอบยึด)

นิยามพหุนามกำลังสามในรูป

$$S_j(x_{j+1}) = a_j+b_j(x-x_j)+c_j(x-x_j)^2+d_j(x-x_j)^3, \; j = 0,1,\ldots,n-1$$

จากเงื่อนไขข้อ 2 จะได้ว่า $S_j(x_j) = a_j = f(x_j)$

จากเงื่อนไขในข้อ 3 จะได้ว่า $a_{j+1} = S_{j+1}(x_{j+1}) = S_j(x_{j+1})$

$$= a_j+b_j(x_{j+1}-x_j)+c_j(x_{j+1}-x_j)^2+d_j(x_{j+1}-x_j)^3, \; j = 0,1,\ldots,n-2$$

ให้ $h_j = (x_{j+1} - x_j)$ สำหรับ $j = 0,1,\ldots,n-1$

ถ้านิยาม $a_n = f(x_n)$ แล้วจะได้ว่า

สมการที่ 1 $a_{j+1} = a_j + b_jh_j + c_jh_j^2 + d_jh_j^3$

สมการที่ 2 $b_{j+1} = b_j + 2c_jh_j + 3d_jh_j^2$

สมการที่ 3 $c_{j+1} = c_j + 3d_jh_j$, $j = 0,1,\ldots,n-2$

จากสมการที่ 3 สามารถหาค่า d_j ได้

$$d_j = \frac{(c_{j+1} - c_j)}{3h_j}$$

แทนค่า d_j ค่ากลับในสมการที่ 1 และสมการที่ 2 จะได้

สมการที่ 4 $a_{j+1} = a_j + b_jh_j + 1/3h_j^2(2c_j+c_{j+1})$ และ

สมการที่ 5 $b_{j+1} = b_j + h_j(c_j+c_{j+1})$, $j = 0,1,\ldots,n-1$

สมการที่ 6 $b_j = \frac{1}{h_j}(a_{j+1} - a_j) - \frac{h_j}{3}(2c_j + c_{j+1})$

สมการที่ 7 $h_{j-1}c_{j-1}+2(h_{j-1}+h_j)c_j+h_jc_{j+1} = \frac{3}{h_j}(a_{j+1} - a_j) - \frac{3}{h_{j-1}}(a_j - a_{j-1})$, $j = 0,1,\ldots,n-1$

งานวิจัยในส่วนประมาณและทดแทนค่าขาดหาย ได้อาศัยแนวคิดของหลักการประมาณค่าในช่วง เพื่อมาพัฒนาวิธีการในการประมาณและทดแทนค่าสูญหายกับชุดข้อมูลลำดับเวลา

2.3.3 ทฤษฎีเกี่ยวกับระเบียบวิธีกำลังสองน้อยสุด (Least-square (LSQ)[27][28][29]

Least-square (LSQ) วิธีกำลังสองน้อยสุดเป็นวิธีประมาณฟังก์ชันแบบหนึ่ง ใช้สำหรับหาสมการที่ใช้เป็นตัวแทนที่ดีที่สุดของชุดข้อมูลที่มีวิธีการคำนวณหาตัวประมาณค่าสัมประสิทธิ์ (Coefficients) ของเส้นถดถอย ซึ่งวิธีกำลังสองน้อยสุดเป็นการนำฟังก์ชันพหุนามมาประมาณค่าในช่วงของข้อมูลแบบจุด(ไม่ต่อเนื่อง) ที่มีจากความคิดที่ว่า จุดข้อมูลที่เรามีทุกๆ จุด (ในที่นี้มี n+1 จุด) จะเสมือนมีเส้นฟังก์ชันที่แท้จริงในอุดมคติ , f(x) ลากผ่านโดยสมมติว่าจุดข้อมูลคือ (x_1, y_1), (x_2, y_2), . . . ,(x_n, y_n) ซึ่งเราสามารถ fit หรือประมาณค่าฟังก์ชัน f(x) นั้นด้วยฟังก์ชันพหุนามอันดับ n ใด ๆ โดยหาผลบวกจากระยะจากจุดทั้งหลายไปยังเส้นถดถอยโดยวัดกันขนานกันกับแกน Y หรือเรียกว่าผลบวกกำลังสองของความคลาดเคลื่อน (Sum of square Error) ดังรูปแบบดังนี้

พิจารณา polynomial ของค่าข้อมูล n ค่า เพื่อคำนวณหาค่า a,b จากจำนวน n จุดหมายถึงจำนวน n ข้อมูล $(x_1,y_1), (x_2,y_2), \ldots\ldots, x_n,y_n)$

$$y = a_0 + a_1x + a_2x^2 + \ldots\ldots\ldots + a_mx^m \quad , \qquad \ldots\ldots\ldots(9)$$

(1) อธิบายได้ด้วยสมการ

$$y = a_0 + a_1x + a_2x^2 + \ldots\ldots\ldots + a_mx_1^m \quad ,$$

$$y = a_0 + a_1x + a_2x^2 + \ldots\ldots\ldots + a_mx_1^m \quad ,$$

$$\ldots\ldots \quad \ldots\ldots \quad \ldots\ldots \quad \ldots\ldots$$

$$y = a_0 + a_1x + a_2x^2 + \ldots\ldots\ldots + a_mx_m^m \quad ,$$

(2) นำเสนอในรูปแบบเมตริกซ์เพื่อแก้ปัญหาแทนสมการ

$y = Xb$ เมื่อ

$$y = \begin{bmatrix} y_1 \\ y_2. \\ y_n \\ \end{bmatrix} \quad , \quad X = \begin{bmatrix} 1 & x_1 & x^2_1 & \ldots\ldots & x^m_1 \\ 1 & x_2 & x^2_2 & \ldots\ldots & x^m_2 \\ . & . & . & .. .. & \\ 1 & x_nx^2_n & \ldots\ldots & x^m_n & \end{bmatrix} \quad , b = \begin{bmatrix} a_0a \\ _1 \\ .a_n \\ \end{bmatrix}$$

ซึ่งสัมประสิทธิ์โพลิโนเมียลสามารถคำนวณได้ดังนี้

$$y = Xb, \quad \text{สมการโพลิโนเมียล}$$

$$Xy = X^tXb$$

$$(X^tx)^{-1}X^ty = (X^tx)^{-1}(X^t x)\, b$$

$$b = (X^t X)^{-1}X^ty$$

จากงานวิจัยนี้นำหลักการสมการถดถอยเพื่อได้ค่าการประมาณมาแทนที่ตำแหน่งที่สูญหายในชุดข้อมูลลำดับเชิงเวลา

2.4 มาตรวัดระยะทางกับการประมาณค่าสูญหาย

2.4.1 แบบยูคลิเดียนหลายมิติ (n Euclidean distance) [10][29]

วิธียูคลิเดียนกรณีหลายมิติ(n Euclidean distance หรือ เรียกว่า Euclidean Distance for Multi-Dimensional Points) เป็นการคำนวณหาระยะห่างระหว่างจุดใดสองจุด ซึ่งระยะทางกำลังสอง(squared distance) จะอยู่ระหว่างสองเวกเตอร์ หรือ สองจุดข้อมูล $x = [\, x_1\ x_2\,]$ และ $y = [\, y_1\ y_2\,]$ ซึ่งคือความแตกต่างของผลรวมกำลังสองระหว่างจุดโดยระยะทาง(distance) ระหว่างเวกเตอร์ x และ y แทนค่าข้อมูลในแต่ละตัวแปรนั่นเอง สามารถใช้ $d_{x,y}$ เขียนเป็นสมการได้ นั่นคือ ผลรวมของจำนวน สำหรับคำนวณระยะทางยูคลิดสามารถคำนวณในลักษณะหลายๆ มิติได้ คือ n มิติ นั่นคือ

- n แทนจำนวนของตัวแปรหรือเรียกว่าปริภูมิ n มิติ
- x แทนข้อมูลชุดที่1 $x = [\, x_{i1}\ \ x_{i2}\ \ldots\ldots\ x_{in}\,]^T$
- y แทนข้อมูลชุดที่2 $y = [\, y_{i1}\ \ y_{i2}\ \ldots\ldots\ y_{in}\,]^T$
- $d_{x,y}$ แทนระยะทางระหว่างข้อมูล x และ y

ซึ่งเราสามารถคำนวณระยะทางระหว่าง x กับ y จะได้สมการคำนวณหาระยะทางแบบยูคลิเดียนหลายมิติดังนี้

$$D_{x,y} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2} \qquad \ldots\ldots\ldots\ldots (10)$$

หรือ

$$D_{x,y} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad \ldots\ldots\ldots (11)$$

ในงานวิจัยนี้ได้นำฟังก์ชันยูคลิเดียนหลายมิติคำนวณเพื่อแสดงถึงความเหมือนหรือความคล้ายกันของชุดข้อมูลของผู้ป่วยที่มาตรวจรักษา ซึ่งระยะห่างระหว่างเวกเตอร์ที่คำนวณได้จะบอกถึงความคล้ายคลึงของชุดข้อมูล ถ้าระยะห่างมากแสดงถึงความคล้ายคลึงก็จะน้อย ถ้าระยะห่างน้อยความคล้ายคลึงก็จะมาก

2.5 ทฤษฎีเกี่ยวกับการลดขนาดข้อมูลและแปลงค่าข้อมูล(feature transformation) [30][31]

ในขั้นตอนที่การเตรียมข้อมูล(pre-processing) เป็นขั้นตอนในตอนแรกก่อนการทำเหมืองข้อมูล ที่จะสามารถปรับปรุงคุณภาพโดยรวมของรูปแบบของข้อมูลก่อนที่จะสกัดความรู้ ซึ่งกระบวนการลดข้อมูล(Data reduction) และแปลงข้อมูล(Data transformation) จะเป็นกระบวนการหนึ่งใน pre-processing เรียกอีกลักษณะหนึ่งคือ แอททริบิวต์ (feature construction) ซึ่งจะเป็นการสร้างแอททริบิวต์ใหม่จากชุดข้อมูลเดิมที่มีอยู่ เพื่อช่วยในการประมวลผลของเหมืองข้อมูล โดย Data Reduction เป็นกระบวนการในการลดขนาดของข้อมูลซึ่งในงานวิจัยนี้เป็นลักษณะ Datasize reduction ด้วยการใช้แถวเป็นหลักในการลดข้อมูล ผสานกับหลักการทรานฟอร์ม โดยลักษณะของแปลงข้อมูล คือโดยรวมกลุ่มข้อมูลและแปรข้อมูลเพื่อเปลี่ยนแปลงรูปแบบของข้อมูลหรือสารสนเทศจากรูปแบบหนึ่งไปเป็นอีกรูปแบบหนึ่ง ประกอบด้วยการลดมิติข้อมูลและประมาณการข้อมูลเพื่อไปพบคุณสมบัติที่เป็นประโยชน์ที่ขึ้นอยู่กับเป้าหมายของงาน จำนวนที่มีประสิทธิภาพของคุณลักษณะภายใต้การพิจารณา สามารถลดหรือเป็นตัวแทนค่าข้อมูลที่สามารถพบได้ ซึ่งเป็นการเปลี่ยนลักษณะข้อมูลหรือการรวมข้อมูลการแปลงรูปแบบของ(Data Transformation) เป็นการแปลงข้อมูลที่เลือกมาให้อยู่ในรูปแบบที่เหมาะสมสำหรับการนำไปใช้วิเคราะห์ตามอัลกอริทึม(Algorithm) และแบบจำลองที่ใช้ในการทำเหมืองข้อมูลต่อไป เพื่อให้ขั้นตอนการทำเหมืองข้อมูลที่เกิดขึ้นอาจจะมีประสิทธิภาพมากขึ้น

2.6 ทฤษฎีเกี่ยวกับหลักการ Inner Products and Norm [18] [29] [32]

แนวคิดของ inner product, ในลักษณะvectors ใน R^2 และ R^3 ปรากฏจุดเริ่มต้นที่จุดกำเนิด ความยาวของเวกเตอร์ x บน R^2 หรือ R^3 เรียกว่า norm ของ x, แสดงด้วย ||x|| ดังนั้นสำหรับ $x = (x_1,x_2)$ R^n, เรามี $||x|| = \sqrt{x_1^2 + x_2^2}$

ดังนั้น, ถ้า $x = (x_1, x_2, x_3)$ ดังนั้น $\| x \| = \sqrt{x_1^2 + x_2^2 + x_3^2}$

R^n เราสามารถกำหนดด้วยนอร์ม norm ของ $x = (x_1, \ldots\ldots, x_n)$ R^nโดย

$$\| x \| = \sqrt{x_1^2 + x_2^2 + \ldots\ldots + x_n^2}$$

2.7 แนวคิดและงานวิจัยที่เกี่ยวข้องกับการจำแนกประเภทกับชุดข้อมูลลำดับเวลา

ในส่วนนี้จะสรุปงานวิจัยต่างๆและเทคนิคเกี่ยวกับการจำแนกประเภทกับข้อมูลลำดับเวลา ดังนี้คือ Ernst, H., and Gert, P., [33] อธิบายขั้นตอนการจำแนกกลุ่มของสัญญาณ Electroencephalograph(EEG) ในส่วนของ brain-computerinterface(BCIs) โดยงานวิจัยนี้เปรียบเทียบกับ 2 ขั้นตอนวิธี คือ โครงสร้างเครือข่ายของนิวรอลเน็ตเวิร์กสำหรับจำแนกข้อมูล single EEG ในส่วนของ BCI โดยจะเปรียบเทียบการจำแนกจากมาตรฐาน multilayer perceptron ร่วมกับการใช้เครือข่าย finite impulse response(FIR) สำหรับการจำแนก singletrail EEG และแสดงประสิทธิภาพการจำแนกแสดง error rates และ ค่าความคลาดเคลื่อนของการจำแนกด้วย FIR, Hirano, S., et al. [17] ได้อธิบายถึงขั้นตอนสำหรับการวิเคราะห์ฐานข้อมูลค่าการตรวจลำดับเวลา(time-series laboratory examination) ซึ่งผลการทดลองแสดงให้เห็นขั้นตอนรวมที่สามารถใช้ค้นพบสิ่งที่น่าสนใจที่ซ่อนอยู่ในฐานข้อมูลลำดับเวลา ด้วยแนวคิดหลักของขั้นตอนนี้คือการจำแนกรูปแบบเชิงเวลาโดยใช้หลักการ multi-scale structure matching ในการหาค่าความเหมือน (similarity) ระหว่างลำดับสองลำดับ ทั้งค่าการตรวจทั้งระยะสั้นและระยะยาว แล้วจัดกลุ่มลำดับข้อมูลด้วยราฟเซต(rough set cluster), Bellazzi , et al. [34] งานวิจัยนี้จะทำการนำเสนอขั้นตอนหลักสำหรับการสกัดความรู้จากลำดับข้อมูลเชิงเวลาของชีวเวชศาสตร์(biomedical) ซึ่งอธิบายถึงการสกัดความรู้จากชุดข้อมูลเชิงเวลาซึ่งเป็นการเตรียมช่วยในการตัดสินใจ ความสามารถในการจัดการและวิเคราะห์ข้อมูลหลายตัวแปรที่ซับซ้อนจะเป็นการให้ข้อมูลที่เป็นประโยชน์ที่จะสกัดจากกิจกรรมด้านการดูแลสุขภาพแบบวันต่อวัน รวมทั้งจากการตรวจสอบของผู้ป่วย โดยในงานวิจัยนี้จะนำเสนอขั้นตอนหลักสำหรับการสกัดความรู้จากชุดข้อมูลลำดับเวลา biomedical สำหรับ Batal, I. ,et al. [16] นำเสนอรูปแบบบนเฟรมเวิร์กการจำแนกประเภทสำหรับข้อมูลอนุกรมเวลาหลายตัวแปร(multivariate timeseries data) วิธีการของพวกเขาขึ้นอยู่กับแนวคิด temporal logic เพื่อสร้างคุณลักษณะการจัดหมวดหมู่และใช้รูปแบบ temporal ที่จะนำเสนอขั้นตอนวิธีการทำเหมืองแร่ที่มีประสิทธิภาพสำหรับรูปแบบการสกัดความรู้เหล่านี้โดยตรง Peter, R. , et al. [35] อธิบายขั้นตอนใหม่ของการจำแนกกลุ่มข้อมูลลำดับเชิงเวลาสำหรับ phenomena นั่นคือ แนวคิดหลักพิจารณาแนวโน้มที่สำคัญที่พัฒนาค่อยๆเมื่อเวลาผ่านไปและมีความผันผวนอย่างรุนแรง ในค่าที่วัด

ได้ในกรณีเวลาที่อยู่ติดกันสภาพอากาศเป็นตัวอย่างที่ดีของเช่น henomenon เพราะความร้อนที่ชัดเจนหรือแนวโน้มการระบายความร้อนในช่วงสัปดาห์หรือเป็นเดือนที่เกิดขึ้นพร้อมกันกับความผันผวนในชีวิตประจำวันอย่างมีนัยสำคัญ ฐานข้อมูลการทดลองที่ใช้ในสำนักงานคณะกรรมการกำกับเท็กซัสที่มีต่อคุณภาพสิ่งแวดล้อม (TCEQ) ฐานข้อมูลซึ่งบันทึกข้อมูลอุตุนิยมวิทยาระหว่างปี 1998 และ 2004 การทดลองบนฐานข้อมูล TCEQ แสดงให้เห็นทั้งสองมีผลที่สำคัญคือ (1) การปรับปรุงความถูกต้องอย่างมีนัยสำคัญที่ได้รับโดยใช้ข้อมูลในอดีตและ (2) การปรับปรุงความถูกต้องโดยใช้คุณสมบัติมากกว่า 3 คุณสมบัติ

บทที่ 3

วิธีดำเนินการวิจัย

ในบทนี้จะเป็นการนำเสนอรายละเอียดของขั้นตอนการดำเนินการวิจัย เพื่อใช้แก้ปัญหาในการนำชุดข้อมูลทางการแพทย์เชิงลำดับเวลานำเข้าเพื่อการจำแนกประเภท ด้วยประมาณค่าสูญหายสำหรับข้อมูลทางการแพทย์เชิงลำดับเวลาและแปลงค่าข้อมูลให้เป็นค่าเดี่ยวเฉพาะ(singular values) ซึ่งจัดเป็นกระบวนการเตรียมชุดข้อมูล(pre-processing) ที่เหมาะสม เพื่อความแม่นยำในการจำแนก

3.1 ขั้นตอนการดำเนินการวิจัย

ในงานวิจัยนี้สามารถแบ่งขั้นตอนในการศึกษาได้ 5 ขั้นตอน ดังนี้

3.1.1 ศึกษาค้นคว้าทางทฤษฎีและวิธีการที่เกี่ยวข้อง โดยการเก็บรวบรวมข้อมูลจากเอกสารและแหล่งข้อมูลที่เกี่ยวข้องทฤษฎีต่างๆที่เกี่ยวข้องกับการจำแนกกลุ่มเชิงเวลาการประมาณค่าสูญหาย การกำหนดขนาดมิติข้อมูลนำเข้าเทคนิควิธีที่เกี่ยวข้อง

3.1.2 การวิเคราะห์และออกแบบการทดลอง

3.1.2.1 รวบรวมเตรียมข้อมูลที่ใช้ในการทดลองจากสองแหล่งคือจากหน่วยงานโรคหลอดเลือดและหัวใจ โรงพยาบาลรามาธิบดีและคัดลอกข้อมูลจาก PKDD'02 Discovery Challenge database ปี2542 ซึ่งเป็นแหล่งข้อมูลทางการแพทย์เชิงลำดับเวลา จากเวบไซต์ http://lisp.vse.cz/pkdd99/Challenge นำข้อมูลทั้งสองชุดมาทำการสุ่มสร้างค่าสูญหายจำนวนชุดข้อมูลละ 3 ชุดข้อมูลแบ่งตามเปอร์เซ็นต์ค่าสูญหาย

3.1.2.2 ออกแบบเทคนิควิธีจัดการมิติข้อมูลเชิงเวลาและการประมาณค่าสูญหายเพื่อความสมบูรณ์ของชุดข้อมูลเพื่อนำมาเข้าในการทำนายการจำแนกประเภท

3.1.3 การพัฒนาและทดสอบประสิทธิภาพ

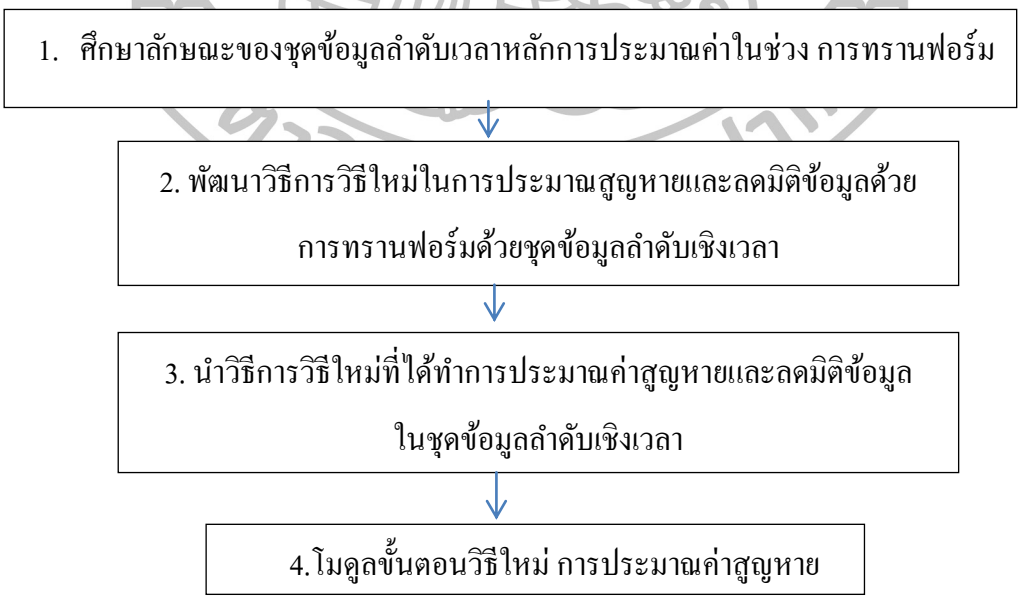
หลังจากออกแบบการทดลอง จะทำการพัฒนาโปรแกรมตามแนวทางที่ได้ออกแบบไว้และทำการเปรียบเทียบประสิทธิภาพ การประมาณค่าสูญหายด้วยเทคนิคขั้นตอนที่ทำการศึกษาและนำมาวัดประสิทธิภาพโดยวิธี NRMSE (NormalRoot Mean Squared Error) กรณีค่าชุดข้อมูลครบสมบูรณ์วัดประสิทธิภาพ การจำแนกข้อมูลจากชุดข้อมูลที่ทรานฟอร์มชุดข้อมูลด้วยวิธีที่พัฒนากรณีชุดข้อมูลที่ไม่สมบูรณ์ ให้นำชุดข้อมูลไปประมาณค่าสูญหายด้วยวิธีข้างต้นแล้วจึงทำการทรานฟอร์ม นำชุดข้อมูลที่ทำการทรานฟอร์มแล้วมาจำแนกประเภทชุดข้อมูลการเปรียบเทียบประสิทธิภาพจะเปรียบเทียบจากการแบ่งชุดข้อมูล และนำมาวัดความแม่นยำของการจำแนกประเภทข้อมูลจากชุดข้อมูลใหม่ที่ได้จากการทรานฟอร์มด้วยค่าความแม่นยำ(accuracy)

3.1.4 การวิเคราะห์ผลและสรุปผลการทดลอง

หลังจากทำการทดลองและทดสอบประสิทธิภาพด้วยวิธีการต่างๆ กับชุดข้อมูลที่กล่าวข้างต้น นำผลการทดลองที่ได้มาทำการวิเคราะห์สรุปผลและข้อเสนอแนะเกี่ยวกับการทดลอง

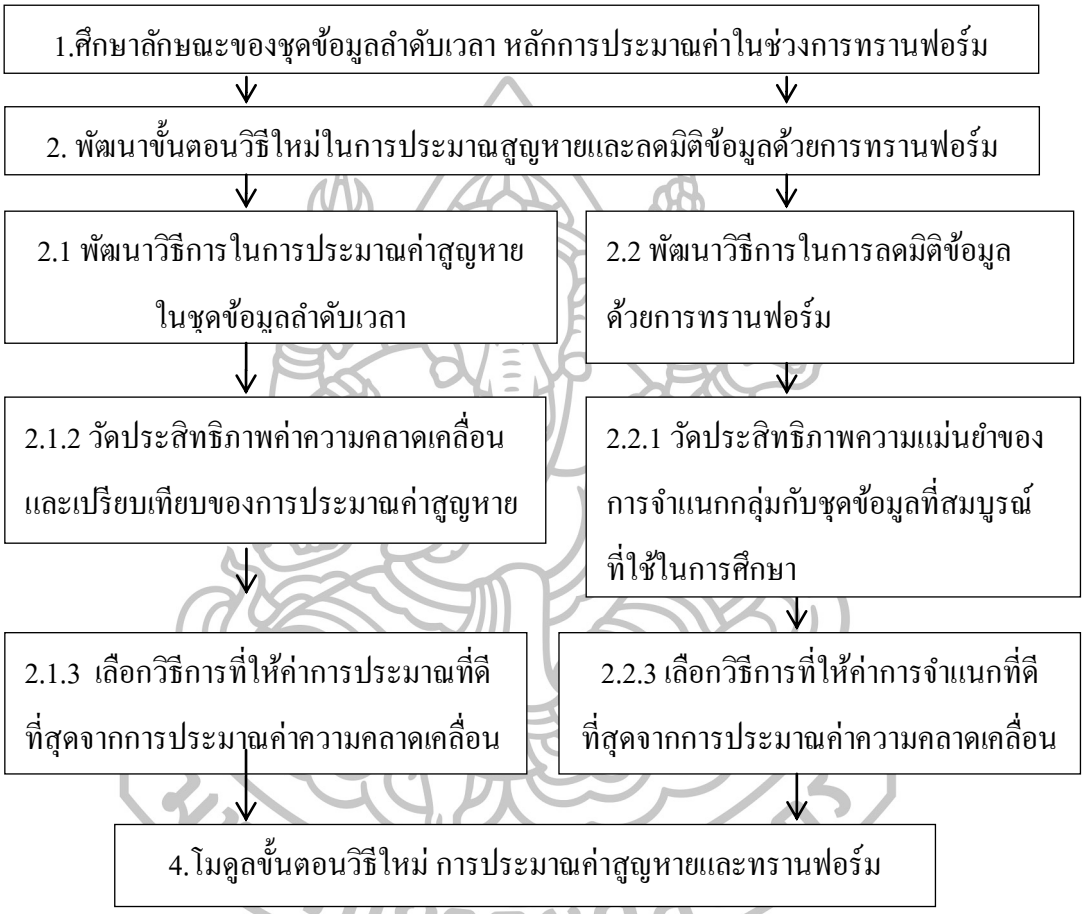
3.2 แผนผังขั้นตอนการวิจัย

ในส่วนของขั้นตอนการพัฒนาขั้นตอนวิธีใหม่ในการประมาณสูญหายและลดมิติข้อมูลในชุดข้อมูลลำดับเชิงเวลาทางการแพทย์ ดังแผนภาพที่ 3.1



แผนผังที่ 3.1 แผนผังแสดงขั้นตอนการวิจัย

ในส่วนของขั้นตอนการพัฒนาขั้นตอนวิธีใหม่ในการประมาณสูญหายและลดมิติข้อมูลในชุดข้อมูลลำดับเชิงเวลาทางการแพทย์ ในงานวิจัยนี้ผู้วิจัยได้แบ่งการพัฒนาออกเป็น 2 ส่วน เพื่อหาวิธีที่ดีที่สุดของแต่ละส่วนจากการวัดประสิทธิภาพเพื่อมาสู่ พัฒนาขั้นตอนวิธีใหม่ในการประมาณค่าสูญหายและลดมิติข้อมูลในชุดข้อมูลลำดับเชิงเวลาทางการแพทย์ ดังแผนผังที่ 3.2



แผนผังที่ 3.2 แผนผังแสดงขั้นตอนการวิจัย

จากแผนผังขั้นตอนการวิจัยข้างต้นมีรายละเอียดในแนวคิดการออกแบบเพื่อพัฒนาแต่ละส่วน ตามขั้นตอนต่อไปนี้

ขั้นตอนที่ 1 ศึกษาการทำงานของชุดข้อมูลลำดับเวลา หลักการประมาณค่าในช่วงการจำแนกประเภทจากชุดข้อมูลเชิงเวลา

ขั้นตอนที่ 2 พัฒนาขั้นตอนวิธีใหม่ในการประมาณสูญหายและลดมิติข้อมูลด้วยการทรานฟอร์มด้วยชุดข้อมูลลำดับเชิงเวลาเพื่อนำเข้าเรียนรู้บนตัวจำแนกประเภทแยกตามวัตถุประสงค์ที่ใช้ในการศึกษาดังนี้

2.1. พัฒนาวิธีการในการประมาณค่าสูญหายในชุดข้อมูลลำดับเวลาชุดข้อมูลนำเข้าไม่สมบูรณ์

2.1.1 ออกแบบและพัฒนาขั้นตอนวิธีในการเรียนรู้โครงสร้างชุดข้อมูลนำเข้าจากการปรากฏค่าสูญหาย บนข้อสมมติฐานแนวคิดในการใช้ค่าเฉพาะรายบุคคล หรือค่าความเหมือนน่าจะให้ค่าการประมาณที่ยอมรับได้ จากการวัดประสิทธิภาพความแม่นยำของค่าที่ประมาณได้

2.1.2 เปรียบเทียบผลการทดลองเพื่อวัดประสิทธิภาพด้วยการประเมินค่าความคลาดเคลื่อน กับขั้นตอนวิธีอื่นที่ใช้ในการศึกษาเพื่อประเมินประสิทธิภาพที่ดีที่สุด

2.1.3 นำผลชุดข้อมูลที่สมบูรณ์ที่ได้จากขั้นตอนวิธีประเมินประสิทธิภาพที่ดีที่สุดมาทำการทรานฟอร์มข้อมูลจากขั้นตอนวิธีในข้อ 2.2

2. 2 พัฒนาวิธีการในการลดมิติข้อมูลด้วยการทรานฟอร์มเมื่อชุดข้อมูลนำเข้าสมบูรณ์

2.2.1 ออกแบบและพัฒนากระบวนการจัดการมิติค่าข้อมูลของผู้ป่วยแต่ละคนแตกต่างกัน

2.2.2 ทรานฟอร์มข้อมูลจากค่าข้อมูลของแต่ละบุคคลในลักษณะลดขนาดของชุดข้อมูล (Datasize reduction) ด้วยการใช้แถวเป็นหลักในการลดข้อมูลเชิงเวลาผสานกับหลักการทรานฟอร์ม ซึ่งลักษณะของการทรานฟอร์มข้อมูล(feature transformation) คือ การรวมกลุ่มข้อมูลและแปรข้อมูลให้อยู่ในรูปแบบที่การประมวลผลเพื่อเปลี่ยนแปลงรูปแบบของข้อมูลหรือสารสนเทศจากรูปแบบหนึ่งไปเป็นอีกรูปแบบหนึ่ง เป็นข้อสมมติฐานเพื่อลดเวลาในการสกัดความรู้(mining) ในกระบวนการจำแนกแต่ยังคงประสิทธิภาพในการจำแนกกลุ่ม

2.2.3 วัดประสิทธิภาพความแม่นยำของการจำแนกกลุ่มกับชุดข้อมูลที่สมบูรณ์จากการประมาณค่าสูญหาย มาทำการเปรียบเทียบขั้นตอนวิธีพัฒนาและที่ใช้ในทดลองที่ให้ประสิทธิภาพที่ดีที่สุด

ขั้นตอนที่ 3 เลือกวิธีการที่ให้ค่าการประมาณที่ดีที่สุดจากการประมาณค่าสูญหายจากการประเมินค่าความคลาด และการจำแนกกลุ่ม

ขั้นตอนที่ 4 โมดูลขั้นตอนวิธีใหม่ การประมาณค่าสูญหายและทรานฟอร์มข้อมูลชุดข้อมูลลำดับเวลาจากขั้นตอนที่ให้ประสิทธิภาพที่ดีที่สุดในขั้นตอนที่ 1 และ 2

บทที่ 4

ผลการดำเนินงานวิจัย

สำหรับบทนี้ในส่วนแรกจะนำเสนออธิบายชุดข้อมูลตามด้วยขั้นตอนที่ใช้ในการศึกษาและผลการทดลอง รวมทั้งผลการเปรียบเทียบประสิทธิภาพความแม่นยำของการประมาณค่าสูญหายและการทรานฟอร์มชุดข้อมูล ดังนี้คือ

4.1 ชุดข้อมูล

งานวิจัยนี้ได้ทำการทดลองกับชุดข้อมูลลำดับลำดับเวลาทางการแพทย์จำนวนสองชุดข้อมูล โดยแต่ละชุดข้อมูลที่ใช้แบ่งเป็นชุดข้อมูลที่สมบูรณ์และชุดข้อมูลที่สุ่มให้ปรากฏค่าสูญหาย ชุดข้อมูลที่วิจัยครั้งนี้ได้แก่

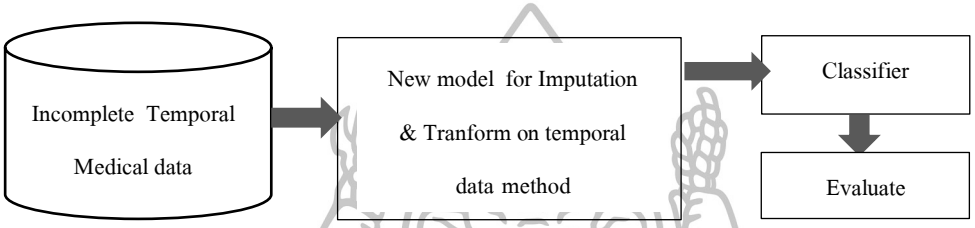
ชุดผู้ป่วยโรคอ้วน (Obesity data) ในงานวิจัยนี้ใช้ชุดค่าข้อมูลการตรวจจริงของผู้ป่วยที่มาทำการตรวจโรคอ้วนที่ศูนย์โรคหลอดเลือดและหัวใจและเมตาลิค โรงพยาบาลรามาธิบดี ประเทศไทย ระหว่างปี พ.ศ. 2523-2542 จำนวน 458 คน จำนวน 1,215 ระเบียน

ชุดข้อมูลผู้ป่วยโรคหลอดเลือดสมองชนิดอุดตัน (Thrombosis-Collagence) ข้อมูลสำหรับโรคหลอดเลือดสมองชนิดอุดตัน ในงานวิจัยนี้ใช้ตาราง TSUM_C ซึ่งเป็นข้อมูลเกี่ยวกับการตรวจทางห้องปฏิบัติการที่จัดเก็บไว้ในระบบสารสนเทศโรงพยาบาล โดยชุดข้อมูลถูกสร้างจัดเป็นส่วนหนึ่ง Discovery Challenge Competition ของ PKDD1999 Discovery Challenge [37] ในการประชุมในยุโรป(3rd European Conference on Principles and Practice of The data) ข้อมูลทั้งหมด ในงานวิจัยนี้ใช้ชุดข้อมูลที่มีผู้ป่วย จำนวน 93 คน และจำนวน 3,010 ระเบียน

ซึ่งได้ทำการเลือกชุดข้อมูลสองชุดนี้เพราะมีลักษณะโครงสร้างเชิงเวลาเหมือนกันซึ่งบอกถึงข้อมูลในอดีตของการตรวจรักษาและระบุโรคของผู้ป่วยในแต่ละครั้ง ประกอบด้วย รหัสผู้ป่วย, วันเวลาการตรวจ และค่าตัวชี้วัด และเป็นข้อมูลการตรวจรักษาในอดีตดังในตารางที่ 2.2

ขั้นตอนที่ใช้ในการดำเนินการวิจัย

ดังนั้นเมื่อศึกษางานวิจัยและลักษณะของข้อมูลลำดับเวลาการพัฒนาขั้นตอนโครงสร้างจากชุดข้อมูลนำเข้าลำดับเวลาเพื่อการจำแนก จึงได้ออกแบบแบ่งการพัฒนาและทดลองเป็น 2 ส่วน คือ กรณีชุดข้อมูลไม่สมบูรณ์ด้วยการประมาณค่าสูญหายและส่วนที่สองการทรานฟอร์มค่าข้อมูลเพื่อลดมิติข้อมูล จากการศึกษาตามขั้นตอน ดังรายละเอียดตามแผนภาพที่ 4.1



แผนภาพที่ 4.1 แสดงกระบวนการเพื่อการจำแนกกลุ่มจากชุดข้อมูลลำดับเวลาที่ไม่สมบูรณ์

ในงานวิจัยนี้ได้พัฒนาตัวแบบบนหลักการประมาณค่าสูญหายและลดมิติข้อมูลด้วยกระบวนการทรานฟอร์มที่เป็นลำดับเวลา ก่อนที่จะได้วิธีการใหม่นี้ ผู้วิจัยได้แบ่งการพัฒนามุ่งเน้นในส่วนของกระบวนการจัดเตรียมข้อมูล(Pre-processing) โดยในการนำเสนอจะแบ่งออกเป็นสองส่วน นั่นคือ ส่วนประมาณค่าสูญหาย(Imputation) และ ส่วนลดมิติข้อมูลด้วยกระบวนการทรานฟอร์ม โดยจะทำการเลือกวิธีการที่ให้ประสิทธิภาพที่ดีที่สุด จากวิธีการที่ใช้ในการศึกษาในแต่ละส่วนเพื่อมาสู่เป็นกระบวนการใหม่เพื่อลดขั้นตอนในเตรียมข้อมูลก่อนจะทำการวิเคราะห์ข้อมูลด้วยเหมืองข้อมูล สำหรับรายละเอียดของวิธีการแต่ละส่วนแยกเป็น 2 ส่วนตามลักษณะวัตถุประสงค์ ดังนี้

4.2.1 วัตถุประสงค์ข้อที่ 1 : การประมาณค่าสูญหายในชุดข้อมูลลำดับเวลา

ในส่วนนี้จะเป็นการนำเสนอรายละเอียดวิธีการประมาณค่าสูญหายในชุดลำดับเวลาทางการแพทย์ เพื่อใช้แก้ปัญหาในกรณีที่ชุดข้อมูลลำดับเวลาปรากฏค่าสูญหายในตัวชี้วัดบางตัว โดยลักษณะข้อมูลลำดับเวลานี้ในส่วนแนวคอลัมน์คือ ตัวชี้วัดและเวลาที่บอกถึงจำนวนครั้งของการมาตรวจรักษา ดังนั้นผู้ป่วยแต่ละคนที่มาตรวจรักษาอาจมีจำนวนครั้งของการมาตรวจที่แตกต่างกัน และมีค่าจากผลการตรวจของแต่ละคน ซึ่งจะต้องเป็นสิ่งที่จะต้องพิจารณาในการประมาณค่าสูญหายให้ครบถ้วนทั้งชุดข้อมูล

4.2.1.1 ขั้นตอนกระบวนการประมาณค่าสูญหาย

1. ทำการนอร์มัลไลซ์ข้อมูล (Normalize data) [14] นอร์มัลไลซ์ข้อมูลในชุดข้อมูลที่สมบูรณ์ด้วย Min-max normalization คือการทรานฟอร์ม ค่าข้อมูลในรูปแบบเชิงเส้น โดยทรานฟอร์มค่าข้อมูลให้อยู่ในช่วงสั้นๆ ในช่วง 0-1 จุดประสงค์ เพื่อให้ค่าตัวชี้วัดเป็นหน่วยวัดเดียวกัน

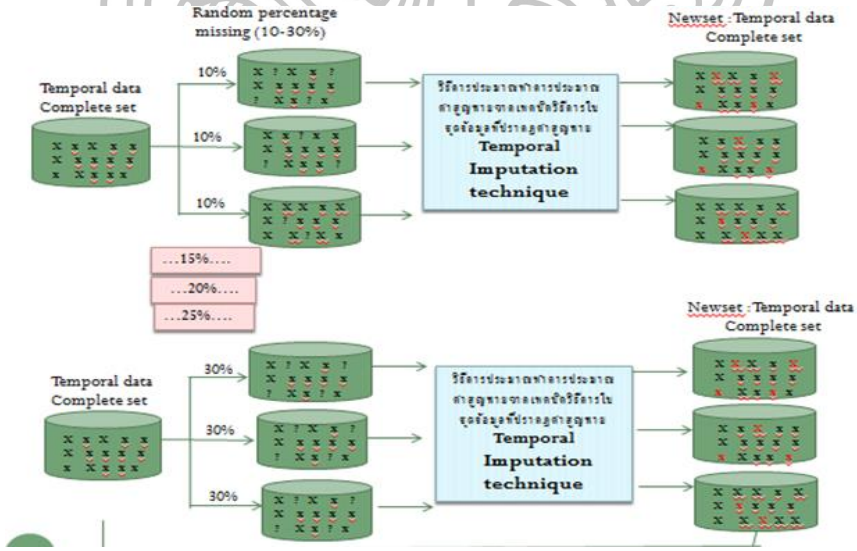
$$v' = \frac{v - \min A}{\max A - \min A}(new\_\max A - new\_\min A) + new\_\min A \qquad \text{.....(8)}$$

โดยกำหนดให้ v คือค่าคุณลักษณะเดิม , v' คือค่าคุณลักษณะใหม่

minA, maxA คือ ค่าต่ำสุดและสูงสุดเดิมของคุณลักษณะ A

new_nimA, new_maxA คือ ค่าต่ำสุดและสูงสุดใหม่ของคุณลักษณะ A

2. ทำการสุ่ม(Random data) สร้างค่าปรากฏการสูญหายในชุดข้อมูลที่สมบูรณ์ จำนวน 3 ชุดข้อมูลที่ตำแหน่งค่าสูญหายไม่ซ้ำกัน ขนาดของข้อมูลในการสร้างค่าขาดหาย (10%, 15%, 20%, 25%, 30%)



ภาพที่ 4.2 การสุ่ม(Random data) สร้างค่าปรากฏการสูญหาย

Random percentage of missing)

<u>วิธีการ</u>

1. ระบุ percent missing
2. คำนวณหา percent missing

Number miss = ((s*Total)/100) , s is % of missing

3. สร้างแฟ้มจำนวน 3 แฟ้ม ประกอบด้วย
   - แฟ้ม 1 สำหรับการ Random ค่าสูญหายครั้งที่ 1
   - แฟ้ม 2 สำหรับการ Random ค่าสูญหายครั้งที่ 2
   - แฟ้ม 3 สำหรับการ Random ค่าสูญหายครั้งที่ 3
4. Random ครั้งที่ 1

   หาตำแหน่ง missing ในแฟ้มแรก ตามจำนวนเปอร์เซ็นต์ค่าสูญหาย เก็บตำแหน่งในแฟ้ม 1 ไว้
5. Random 2 ในแฟ้ม 2

   หาตำแหน่ง missing เมื่อเจอตรวจสอบกับแฟ้มแรก

   หากตำแหน่งตรงกับตำแหน่งในแฟ้มแรก ขยับไปตำแหน่งถัดไปจนครบตามจำนวน
6. Random 3 ในแฟ้ม 3

   หาตำแหน่ง missing ตรวจสอบเทียบตำแหน่งกับแฟ้ม 1 และแฟ้ม 2 หากตำแหน่งที่ Random ในแฟ้ม 2 ตรงกับตำแหน่งใน แฟ้มแรก ขยับไปตำแหน่งถัดไป จนครบ ตามจำนวน
7. นำตำแหน่งในแฟ้มแรกสร้างค่าว่างในตำแหน่ง random ทั้งหมด

3. ทำการประมาณค่าสูญหาย จากเทคนิคขั้นตอนวิธีในชุดข้อมูลที่ปรากฏค่าสูญหาย ผู้วิจัยได้พัฒนาวิธีการบนหลักการต่างๆ เพื่อประมาณค่าสูญหายในชุดข้อมูลลำดับลำดับเวลา บนแนวคิดในการนำผลค่าการตรวจในแต่ละตัวชี้วัดของตนเองมาทำการประมาณค่าเป็นหลัก ดังวิธีการในหัวข้อ 4.2.1.2

4. ประเมินประสิทธิภาพขั้นตอนวิธีของการประมาณค่า

หลังจากออกแบบการทดลองและพัฒนาขั้นตอนวิธีตามขั้นตอนวิธีที่ได้ จะทำการเปรียบเทียบประสิทธิภาพ โดยจะเปรียบเทียบความคลาดเคลื่อน โดยทำการวัดประสิทธิภาพค่าความคลาดเคลื่อนด้วย (Normal Root Mean Square Error : NRSME) เพื่อทดสอบประสิทธิภาพของชุดข้อมูล เพื่อตรวจสอบประสิทธิภาพของขั้นตอนวิธีเพื่อวัดความน่าเชื่อถือของโมเดลซึ่งเป็นวิธีที่นิยมใช้ โดยเปรียบเทียบกับขั้นตอนวิธีเดิมที่ใช้ในการศึกษาแต่ปรับปรุงขั้นตอนวิธีให้สามารถประมาณค่าในลักษณะรายบุคคลเหมือนขั้นตอนวิธีที่พัฒนา โดยวัดกับเปอร์เซ็นต์ขนาดของข้อมูล

ในการสร้าง ค่าสูญหาย (10% 15% 20% 25% 30%) จำนวนเปอร์เซ็นต์ละ 3 ชุดข้อมูลที่ตำแหน่งสูญหายไม่ซ้ำกัน

Normal Root Mean Square Error : NRMSE

NRMSE ถูกใช้ในการเปรียบเทียบความแม่นยำของการประมาณค่าสูญหายของวิธีการที่แตกต่างกัน NRMSE จะคำนวณข้อผิดพลาดระหว่างค่าจริงและค่าประเมินและประเมินความถูกต้อง โดยกำหนดให้ y_{know} คือ ค่าจริงที่ปรากฏ(real value) , $y_{predict}$ คือ ค่าที่ได้จากการประมาณ (estimated value) , t คือ ลำดับเวลาครั้งที่ในตำแหน่งที่ปรากฏค่าสูญหาย และ n คือ จำนวนค่าสูญหายที่ปรากฏ ค่า NRMSE ที่น้อยจะเป็นเกณฑ์บ่งชี้บอกความแม่นยำที่ดีที่สุด [4][5][24] ดังสมการที่ (9) นี้

$$NRMSE = \sqrt{\frac{\sum_{t=1}^{n}(y_{know} - y_{Predict})^2}{Std_{[yknown]}}} \quad \text{.....(9)}$$

4.2.1.2 เทคนิคหลักการประมาณค่าข้อมูลสูญหาย

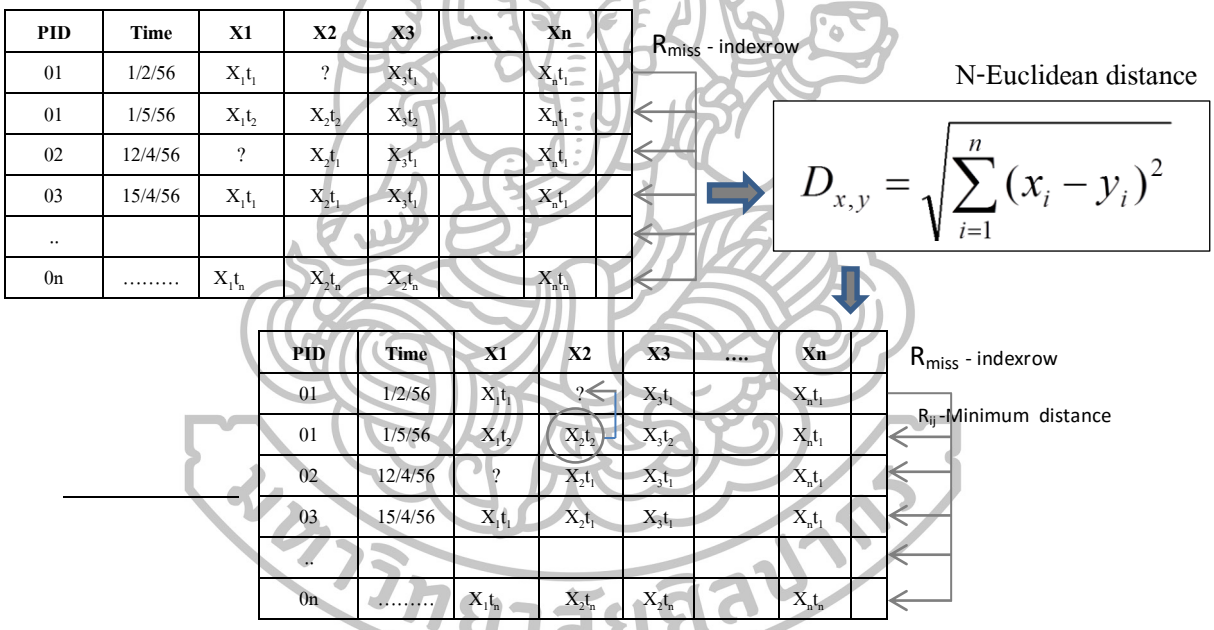
ผู้วิจัยได้ทำการประมาณค่าสูญหายในชุดข้อมูลลำดับเวลาด้วยวิธีการและหลักการดังจากตารางที่ 4.1 ดังกล่าว ผู้วิจัยได้ทำการประมาณค่าสูญหายในชุดข้อมูลลำดับเวลาด้วยวิธีการดังต่อไปนี้

ตารางที่ 4.1 แสดงเทคนิคหลักการประมาณค่าข้อมูลสูญหาย

ขั้นตอน	หลักการประมาณ
DPimpute	วัดค่าความเหมือนและความคล้ายจากระยะทาง
SKnn-DPimpute	การวัดค่าระยะทางที่ใกล้เคียงจำนวน k ค่า (k คือจำนวนครั้ง)
CB & Extra-DPimpute	สไปน์โพลิโนเมียลและวัดค่าความคล้าย
SLLS-DPimpute	สมการเชิงถดถอยและวัดค่าความคล้าย
NFDCslideW-DPimpute	เงื่อนไขดีกรีโพลิโนเมียลและวัดค่าความคล้ายด้วยระยะทาง
NFDCs3-DPimpute	เงื่อนไขดีกรีโพลิโนเมียลและวัดค่าความคล้ายด้วยระยะทาง

1. วิธีการประมาณค่าสูญหาย DPimpute : ใช้หลักการเปรียบเทียบข้อมูลที่สนใจกับข้อมูลอื่นด้วยหลักการวัดระยะทางด้วยยูคิลิดหลายมิติ(Euclidean distance multidimension) หรือ N-Euclidean distance

ขั้นตอนวิธีนี้จะนำค่าข้อมูลที่มีความเหมือนหรือความคล้ายจากระเบียนที่มีรายการปรากฏค่าสูญหายโดยเปรียบเทียบกับระเบียนรายการตรวจทั้งหมดแต่ละรายการ หากทั้งรายการตรวจของตนเองและเทียบความเหมือนความคล้ายกับรายการจากค่าข้อมูลการตรวจของผู้ป่วยคนอื่นที่มีอยู่ในชุดข้อมูลทั้งหมด ซึ่งหากระเบียนใดมีค่าใกล้เคียงก็จะนำค่าข้อมูลของรายการที่ตรงกับตำแหน่งที่ปรากฏค่าสูญหายมาแทนที่ในตำแหน่งดังกล่าว ดังแผนภาพที่ 4.2



แผนภาพที่ 4.2 แสดงแนวคิดการประมาณค่าสูญหายด้วยวิธีการ DPimpute

PID คือ รหัสผู้ป่วย, T ime คือ เวลาของการตรวจรักษา,

$X_1 \ldots X_n$ คือ ตัวแปร/ตัวชี้วัดในชุดข้อมูล,

X_i คือ ค่าการตรวจแต่ละครั้งในแต่ละตัวชี้วัดของแต่ละคนในแถวที่ปรากฏค่าสูญหาย

Y_i คือ ค่าการตรวจแต่ละครั้งในแต่ละตัวชี้วัดของในแถวถัดไป

R_{miss} – indexrow คือ แถวที่ปรากฏค่าสูญหาย,

R_{ij} -Minimum distance คือ แถวที่ได้ค่าระยะทางน้อยที่สุด

รายละเอียดขั้นตอนของวิธีการ DPimpute ดังนี้

Procedure : DPimpute()

Input : Incomplete temporal dataset (D)

Output : Complete temporal dataset(D)

Step 1. Select all temporal medical data

Step 2. Checking missing position

Step3. Pad zero to NaN column

Step4 : Compute N-dimension distance

Step 4.1 : vectorNaN = padZeroData(i,:);

Step 4.2 : sumData = zeros(ndata_row,1)

Step 4.3 : distanceData = padZeroData;

Step 4.4 : Loop and index indentifies the minimum distance row

```
for j=1:ndata_row
    % distanceData(j,:) = (abs(power((padZeroData(j,:) - vectorNaN),[2])));
    distanceData(j,:) =  power((padZeroData(j,:) - vectorNaN),[2]);
    sumData(j, 1) =  sqrt(sum(distanceData(j,:)));
    mx = max(sumData);
    sumData(i,1) = mx;  % Avoid it will be minimum
end
```

Step5 : Replace values in column which is NaN with data in index row, minimum distance.

```
for j=1:c
    resultData(i, colNaN(1, j)) = distanceData(index, colNaN(1,j));
end
```

ผลการทดลอง

ผลการทดลองด้วยตัววัดประสิทธิภาพการประมาณค่าด้วย NRSME ด้วยการแบ่งเป็น % ของการที่ปรากฏค่าสูญหายของวิธีการ DPimpute ดังตารางที่ 4.2-4.4

ตารางที่ 4.2 ค่าความคลาดเคลื่อน(NRMSE) ของชุดข้อมูลลำดับเวลา (DPimpute) ชุดที่ 1

% missing / Dataset	10%	15%	20%	25%	30%	%Average
Obesity	0.10760	0.10824	0.11134	0.11425	0.11484	0.11125
Thrombosis	0.147970	0.152707	0.154287	0.161768	0.168859	0.15712

ตารางที่ 4.3 ค่าความคลาดเคลื่อน(NRMSE) ของชุดข้อมูลลำดับเวลา (DPimpute) ชุดที่ 2

% missing / Dataset	10%	15%	20%	25%	30%	%Average
Obesity	0.10777	0.11093	0.11076	0.11760	0.11456	0.11232
Thrombosis	0.143519	0.146628	0.156782	0.160918	0.170778	0.15573

ตารางที่ 4.4 ค่าความคลาดเคลื่อน(NRMSE) ของชุดข้อมูลลำดับเวลา (DPimpute) ชุดที่ 3

% missing / Dataset	10%	15%	20%	25%	30%	%Average
Obesity	0.10609	0.11130	0.11436	0.11520	0.11184	0.11176
Thrombosis	0.144625	0.149842	0.152767	0.158743	0.166604	0.15452

สรุปและอภิปรายผลการทดลอง

จากผลการทดลองในตารางที่ DPimpute ดังตารางที่ 4.2-4.4 ซึ่งจากชุดข้อมูลทั้งสามชุดที่ตำแหน่งการสูญหายไม่ซ้ำกัน พิจารณาที่ค่าเฉลี่ยเปอร์เซ็นต์ค่าสูญหายของแต่ละชุดข้อมูลให้ค่าความคลาดเคลื่อนเฉลี่ยของชุดข้อมูล obesity ชุดที่ 1-3 ตามลำดับ คือ 0.11125, 0.11232, 0.11176 และ ชุดข้อมูล thrombosis ชุดที่ 1-3 ตามลำดับ คือ 0.15712, 0.15573, 0.154524 ซึ่งจากชุดข้อมูลทั้งสามชุดที่ตำแหน่งการสูญหายไม่ซ้ำกัน ซึ่งให้ค่าเฉลี่ยเปอร์เซ็นต์ความคลาดเคลื่อนที่ใกล้เคียงกัน

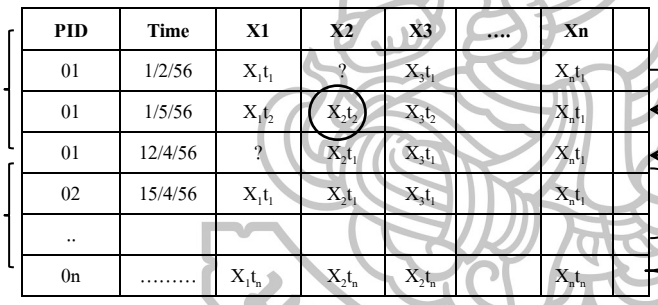
ในตารางผลการทดลองตารางที่ 4.2-4.4 ในส่วนของตำแหน่งเปอร์เซ็นต์ค่าสูญหายจะมีเปอร์เซ็นต์ค่าสูญหายที่มากขึ้น แต่ผลเปอร์เซนต์ความคลาดเคลื่อนบางเปอร์เซนต์ค่าความคลาดเคลื่อนกลับน้อยลง จากการวิเคราะห์นั่นคือ ค่าการประเมินค่าความคลาดเคลื่อนจะต้องประเมินในภาพรวมของชุดข้อมูลทั้งหมดที่ปรากฏ แต่ในการสุ่มสร้างค่าสูญหายตำแหน่งของการปรากฏค่าสูญหายที่ปรากฏ จะมีข้อมูลจริงที่ปรากฏอยู่รอบหรือใกล้กับตำแหน่งที่สูญหาย ซึ่งค่าที่ปรากฏอาจจะเป็นค่าที่มากหรือค่าที่น้อยก็ได้ ซึ่งเมื่อประมาณค่าอาจจะให้ค่าการประมาณที่สูงหรือต่ำผิดปกติได้

จากวิธีการ DPimpute วิธีนี้จะเป็นการหาค่าความเหมือนหรือความคล้ายจากค่าข้อมูลในแต่ละตัวชี้วัดจากรายการที่ปรากฏค่าสูญหาย เปรียบเทียบกับรายการข้อมูลทั้งชุดข้อมูลสามารถนำไปใช้ได้กับชุดข้อมูล การตรวจระยะเวลาสั้นและช่วงระยะยาว ลักษณะข้อมูลที่ใช้ในการประมาณค่าจะเป็นข้อมูลในแต่ละตัวชี้วัดที่นำมาทำการประมาณค่าสูญหายซึ่งจะเป็นค่าข้อมูลของตนเองและรายการของบุคคลอื่นซึ่งมาตรวจแต่ละครั้ง โดยหาค่าของรายการใดที่มีค่าใกล้เคียงกับรายการในตำแหน่งที่สูญหาย แต่ถ้าหากเรานำเฉพาะค่าของตนเองเป็นหลักมาใช้ในการประมาณค่าสูญหายน่าจะได้ค่าการประมาณที่เป็นค่าเฉพาะของคนนั้น ๆ

2. วิธีการประมาณค่าสูญหาย (Sk-NN-DPimpute)

จากวิธีการแรกในการประมาณค่าสูญหายในชุดข้อมูลเชิงเวลาด้วยการหาค่าความเหมือนและความคล้ายบนหลักการวัดระยะทาง(Euclidean distance) ทำให้พบว่ามันเป็นการนำค่าข้อมูลในตัวชี้วัดทั้งของผู้ป่วยที่สนใจและผู้ป่วยคนถัดไปมาประมวลผล เพื่อให้ได้ค่าการประมาณผู้วิจัยจึงได้นำขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด(k-Nearest Neighbour , k-NN) ซึ่งเป็นหลักการที่

ถูกพัฒนาโดย Troyanskaya et al. [7] ใช้หลักการเปรียบเทียบข้อมูลที่สนใจกับข้อมูลอื่นว่ามีความคล้ายคลึงมากน้อยเพียงใดด้วย Euclidean distance และทำนายข้อมูลใหม่โดยอาศัยการเปรียบเทียบกับข้อมูลเรียนรู้จำนวน k ตัวที่อยู่ใกล้ที่สุด โดยจะให้ค่าน้ำหนักโดยการพิจารณาระยะห่างระหว่างข้อมูลที่สนใจกับข้อมูลที่อยู่ใกล้สุด k ตัวร่วมด้วย ในงานวิจัยนี้ผู้วิจัยได้กำหนดค่า k คือตามจำนวนการตรวจรักษาของผู้ป่วยแต่ละคน ซึ่งการหาค่าความเหมือนความคล้ายจะเปรียบเทียบเฉพาะรายการตรวจเฉพาะของบุคคลนั้นๆ ในงานวิจัยนี้จึงเรียกว่า Subspace k-NN แต่ปัญหาหนึ่งของวิธีการดังกล่าวหากข้อมูลผู้ป่วยมาตรวจเพียงครั้งเดียวปรากฏค่าสูญหายจะไม่สามารถประมาณด้วยการเทียบค่ากับรายการอื่นของตนเองในลักษณะรายบุคคลได้ ดังนั้น จึงนำวิธีการ DPimpute มาทำการประมาณค่าในส่วนนี้ ก็จะทำให้สามารถประมาณค่าสูญหายทั้งชุดข้อมูล จึงเรียกวิธีการนี้ว่า Subspace k-nearest neightbor-DPimpute (Sk-NN-DPimpute) มีรายละเอียดดังแผนภาพและขั้นตอนวิธีดังนี้



ภาพที่ 4.3: แสดงแนวคิดการประมาณค่าสูญหายด้วยวิธีการ Sk-NN-DPimpute

PID คือ รหัสผู้ป่วย,T ime คือ เวลาของการตรวจรักษา,

$X_1 \ldots X_n$ คือ ตัวแปร/ตัวชี้วัดในชุดข้อมูล,

$X_i t_i$ คือ ค่าการตรวจแต่ละครั้งในแต่ละตัวชี้วัดของแต่ละคน

R_{miss}–indexrow คือ แถวที่ปรากฏค่าสูญหาย,

R_{ij} -Minimum distance คือ แถวที่ได้ค่าระยะทางน้อยที่สุด k-value คือ จำนวนรายการการตรวจรักษา k รายการ

ขั้นตอนในภาพรวมมีดังนี้คือ

1. นำชุดข้อมูลเชิงเวลาแยกเป็นเมตริกซ์ข้อมูลในลักษณะรายคน

2. กำหนดค่า 0 ลงไปในตำแหน่งที่ปรากฏค่าสูญหายแต่ละคอลัมน์เพื่อสำหรับตรวจสอบตำแหน่งที่ปรากฏค่าสูญหาย

3. ตรวจสอบจำนวนการตรวจรักษาและตำแหน่งที่สูญหายในแต่ละแถวของแต่ละคน

4. ถ้าจำนวนการตรวจของแต่ละคนมากกว่า 2 ครั้ง

5. คำนวณระยะทางด้วย n-Euclidean distance จากเฉพาะข้อมูลในแต่ละแถวของแต่ละคน

6. เลือกแถวที่ปรากฏค่าระยะทางต่ำสุดแล้วนำค่าจริงที่ตรงกับตำแหน่งที่สูญหายไปแทนที่

7. ถ้าจำนวนการตรวจของแต่ละคน ตรวจครั้งเดียวหรือไม่เกิน 2 ครั้ง ประมาณค่าสูญหายด้วยวิธีการ DPimpute

รายละเอียดขั้นตอนของวิธีการ Sk-NN-DPimpute ดังนี้

Procedure : Sk-NN-DPimpute

Input : Incomplete temporal dataset (D)

Output : Complete temporal dataset(D)

Step1 : Transform temporal medical data to low dimension in subspace matrix

1.1 Determine the system's input variables for a temporal data matrix

1.2 Separate the matrix into m x n each of patient subspace dimension.

Step2: Find NaN each column

%Pad zero to NaN column

```
for i=1:ndata_row
  v=isnan(NaNData(i,:));
  colNaN=find(v==1);  % get colum containing NaN
  [r, c] = size(colNaN);  % for loop to fill zero
```

```
    %fill 0 to colum contains NaN
   for j=1:c
       padZeroData(i,colNaN(1,j))=0;
   end
   end
```

Step3 : Checking missing position

```
for i=1:numberOfPatient
  if(i == numberOfPatient)
    startRow = ia(numberOfPatient);
    endRow = ndata_row;  % use for the last patient
    else
    % ia variable contain start row of patient
    startRow = ia(i);
    endRow = ia(i+1);
  end
```

Step4 : Checking k-row value

```
    if(endRow - startRow > 2)
    %This loop is check row which it has NaN data
    for j=startRow:endRow-1
      if (any(isnan(NaNData(j,:))))
        currentRow = (j-startRow)+1;  % currentRow uses for identify row has NaN
        v=isnan(NaNData(j,:));
        colNaN=find(v==1);  % get colum containing NaN
        [r, c] = size(colNaN);  % for loop to fill zero
```

Step5 : compute distance

```
        vectorNaN = padZeroData(j,:);
        sumData = zeros(1:endRow-startRow,1);
        distanceData = padZeroData((startRow:endRow-1),:);
        for k=1:endRow-startRow
          distanceData(k,:) = (abs(power((padZeroData((k+startRow)-1,:) - vectorNaN),[2])));
```

```
    sumData(k, 1) = sqrt(sum(distanceData(k,:)));
   end
  mx = max(sumData);
  sumData(currentRow,1) = mx;
```

Step6 : Replace column whic is NaN with data in index row, minimum distance.

```
  for l=1:c
    resultData(j, colNaN(1, l)) = NaNData(startRow+index-1,colNaN(1,1));
  end
```

Step 7 : Checking time treatment <=2

Step 8. Checking missing position

Step 9. Pad zero to NaN column

Step10. Compute N-dimension distance

Step 10.1 : vectorNaN = padZeroData(i,:);

Step 10.2 : sumData = zeros(ndata_row,1)

Step 10.3 : distanceData = padZeroData;

Step 10.4 : Loop and index indentifies the minimum distance row

```
  for j=1:ndata_row
  distanceData(j,:) = power((padZeroData(j,:) - vectorNaN),[2]);
    sumData(j, 1) = sqrt(sum(distanceData(j,:)));
    mx = max(sumData);
    sumData(i,1) = mx;  % Avoid it will be minimum
  end
```

Step11 : Replace values in column which is NaN with data in index row, minimum distance.

```
  for j=1:c
      resultData(i, colNaN(1, j)) = distanceData(index, colNaN(1,j));
  end
```

ผลการทดลอง

ผลการทดลองด้วยตัววัดประสิทธิภาพการประมาณค่าด้วย NRSME ด้วยการแบ่งเป็น % ของการที่ปรากฏค่าสูญหายคือ

ตารางที่ 4.5 ค่าความคลาดเคลื่อน(NRMSE) ของชุดข้อมูลลำดับเวลา (SkNN-DP) ชุดที่ 1

% missing Dataset	10%	15%	20%	25%	30%	%Average
Obesity	0.35184	0.40588	0.45748	0.48182	0.50208	0.43982
Thrombosis	0.386775	0.474585	0.552964	0.629435	0.714833	0.55172

ตารางที่ 4.6 ค่าความคลาดเคลื่อน(NRMSE) ของชุดข้อมูลลำดับเวลา(Sk**NN-**DP) ชุดที่ 2

% missing Dataset	10%	15%	20%	25%	30%	%Average
Obesity	0.33948	0.45428	0.45845	0.47302	0.49756	0.44456
Thrombosis	0.377631	0.467382	0.563123	0.638664	0.713756	0.55211

ตารางที่ 4.7 ค่าความคลาดเคลื่อน(NRMSE) ของชุดข้อมูลลำดับเวลา (SkNN-DP) ชุดที่ 3

% missing Dataset	10%	15%	20%	25%	30%	%Average
Obesity	0.35783	0.41892	0.46137	0.50439	0.51776	0.45205
Thrombosis	0.375297	0.483222	0.553492	0.640958	0.709619	0.55252

สรุปและอภิปรายผลการทดลอง

จากผลการทดลองในตารางที่ SkNN-DPimpute ดังตารางที่ 4.5-4.7 ซึ่งจากชุดข้อมูลทั้งสามชุดที่ตำแหน่งการสูญหายไม่ซ้ำกัน พิจารณาที่ค่าเฉลี่ยเปอร์เซนต์ค่าสูญหายของแต่ละชุดข้อมูลให้ค่าความคลาดเคลื่อนเฉลี่ยของชุดข้อมูล obesity ชุดที่ 1-3 ตามลำดับ คือ 0.43982, 0.44456, 0.45205 และ ชุดข้อมูล thrombosis ชุดที่ 1-3 ตามลำดับ คือ 0.55172, 0.55211, 0.55252 ซึ่งให้ค่าการประมาณที่ใกล้เคียงกัน

จากวิธีการ SkNN-DPimpute จะทำการประมาณค่าสูญหายด้วยหลักการวัดระยะทางตามจำนวนการตรวจรักษา(k) ด้วยการใช้เฉพาะรายบุคคล แต่จากผลการทดลองจากการวัดประสิทธิภาพค่าความคลาดเคลื่อน วิธีการ DPimpute ให้ประสิทธิภาพความแม่นยำที่ดีกว่า

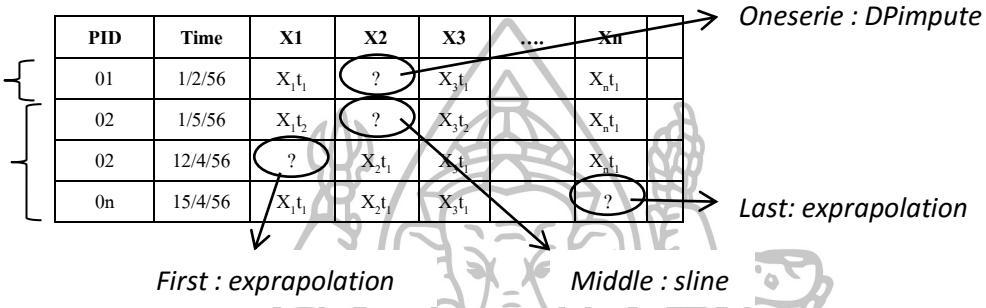
จากการวิเคราะห์การประมาณค่าการตรวจด้วยวิธีการ SkNN-DPimpute ซึ่งใช้ค่าเฉพาะรายบุคคลมาทำการประมาณซึ่งลักษณะข้อมูลการตรวจของผู้ป่วยอาจจะมีความห่างของช่วงการตรวจรักษาในแต่ละครั้ง เช่น ปีละครั้ง สองครั้ง หรือ รายสัปดาห์ ซึ่งทำให้ค่าการตรวจในบางตัวชี้วัดไม่ใกล้เคียงกับค่าการตรวจครั้งแรกๆ เมื่อประมาณค่าด้วยค่าความเหมือนความคล้ายหากตำแหน่งที่ปรากฏค่าสูญหายและมีระยะทางต่ำสุดที่ถูกเลือกมา อาจจะปรากฏค่าจริงที่ไม่ใกล้เคียงกับค่าเดิม สำหรับคนที่มาตรวจครั้งเดียวจะใช้วิธีการ DPimpute มาประมาณค่าซึ่งจะต้องคำนวณกับทุกแถวที่ปรากฏดังนั้นค่าที่เคยถูกประมาณไว้แล้ว สำหรับผู้ป่วยที่มีการตรวจหลายครั้ง ก็จะมีผลหากตำแหน่งที่สูญหายและระยะทางต่ำสุดไปตรงกับตำแหน่งที่เคยประมาณไว้แล้วด้วย **SkNN** หากประมาณได้ค่าไม่ใกล้เคียงจาก **SkNN** ก็มีผลทำการประเมินค่าความคลาดเคลื่อนสูงไปด้วย

จึงได้นำหลักการประมาณค่าในช่วงเพื่อมาประมาณค่าสูญหายในชุดข้อมูลเชิงเวลาดังหัวข้อ 1

3. วิธีการ Cubic Spline & Extra-DPimpute

ขั้นตอนวิธีนี้จะทำการประมาณค่าสูญหายในลักษณะเฉพาะรายบุคคล หากปรากฏค่าสูญหายระหว่างการตรวจแต่ละครั้ง เนื่องจากคิวบิกสไปน์สามารถประมาณค่าสูญหายระหว่างช่วงข้อมูลได้ แต่ขั้นตอนวิธีนี้ไม่สามารถประมาณค่าสูญหายในตำแหน่งแรกและตำแหน่งสุดท้ายได้ เนื่องจากขั้นตอนวิธีนี้จะนำค่าจริงที่ปรากฏที่อยู่ใกล้เคียงตำแหน่งที่ปรากฏค่าสูญหายมาทำการประมาณตามขั้นตอนวิธี ซึ่งไม่ได้นำทุกค่าที่ปรากฏในการตรวจแต่ละครั้งมาทำการประมาณ จึง

ได้นำหลักการประมาณค่านอกช่วงมาใช้ในการประมาณค่าสูญหายในตำแหน่งแรกและตำแหน่งสุดท้าย อีกทั้งสไปน์ไม่สามารถประมาณค่าสำหรับข้อมูลสูญหายที่ปรากฏในกรณีที่ผู้ป่วยมาตรวจ 1 หรือ 2 ครั้งได้ ในส่วนนี้จึงนำหลักการ DPimpute ที่เป็นขั้นตอนในข้อ 1 มาหาค่าความเหมือนและความคล้ายของรายการบุคคลอื่นๆ มาประมาณค่าเฉพาะในส่วนของผู้ป่วยที่มาตรวจ 1 หรือ 2 สองครั้งในชุดข้อมูล ทำให้สามารถประมาณค่าสูญหายได้ทั้งชุดข้อมูลลำดับเวลา จึงเรียกวิธีการนี้ว่า CBE-DPimpute มีรายละเอียด ดังแผนภาพและขั้นตอนวิธีดังนี้



ภาพที่ 4.4: แสดงแนวคิดการประมาณค่าสูญหายด้วยวิธีการ CBE-DPimpute

ดังนั้นการพัฒนาขั้นตอนการประมาณค่าสูญหายด้วยวิธี CBE-DPimpute บนชุดข้อมูลลำดับเวลานี้บน แนวคิดการประมาณค่าจะใช้เฉพาะค่าการตรวจเฉพาะบุคคลมาใช้ในกระบวนการประมาณค่าสูญหาย โดยจะแบ่งการประมาณค่าสูญหายด้วย 2 เงื่อนไขหลัก คือ กรณีผู้ป่วยที่มาตรวจรักษามากกว่า 2 ครั้ง และ ผู้ป่วยที่มาตรวจรักษาเพียงครั้งเดียว

ขั้นตอนในภาพรวมมีดังนี้คือ

1. นำชุดข้อมูลเชิงเวลาแยกเป็นเมตริกซ์ข้อมูลในลักษณะรายคน
2. ทรานโพสเมตริกซ์ และเลือก row vector ของแต่ละคน
3. ตรวจสอบจำนวนการตรวจรักษาและตำแหน่งที่สูญหายในแต่ละแถวของแต่ละคน
4. ถ้าจำนวนการตรวจของแต่ละคนมากกว่า 2 ครั้ง
5. ตรวจสอบตำแหน่งที่ปรากฏค่าสูญหาย

5.1 หากค่าสูญหายปรากฏระหว่างข้อมูล คำนวณค่าที่อยู่ในช่วงด้วย spline

5.2 หากค่าสูญหายปรากฏในตำแหน่งแรกและตำแหน่งสุดท้าย คำนวณค่าที่อยู่นอกช่วงด้วยวิธีการประมาณนอกช่วง

6. แทนที่ในตำแหน่ง

. วนรอบข้อ 2 จนกว่าค่าข้อมูลจะครบสมบูรณ์

8. ถ้าจำนวนการตรวจของแต่ละคน ตรวจครั้งเดียวหรือไม่เกิน 2 ครั้ง ประมาณค่าสูญหายด้วยวิธีการ DPimpute

สำหรับขั้นตอนในรายละเอียดคือ

Procedure Cubic-ExtrapolationDPImpute (CBE-DPimpute)

Input :Incomplete temporal dataset(D)

Output : Complete temporal dataset(D')

Step 1 : Transform temporal medical data to low dimension in subspace matrix

1.1 Determine the system's input variables for a temporal data matrix

1.2 Separate the matrix into m x n each of patient subspace dimension.

1.3 Transpose patient subspaces with missing values ,

$[x(t_1)\ x(t_2)\ x(t_3) \ldots\ldots x(t_n)]$

Step 2 : Compute in transposed patient subspace.

2.1 Repeat

2.2 Select the transposed subspace in each patient case for imputation.

2.2.1 Repeat

2.2.2 Select row vector in each patient subspace

// $X_n = [x_i(t_1)\ x_i(t_2)\ x_i(t_3) \ldots\ldots x_i(t_n)]$ where $t_i = 1,2,\ldots,n$.

Step 3 : Start fill missing

3.1 Check time- treatment of Patient –ID-$_{0n}$ and missing position

3.2 if time tratement of patient > 2

3.2.1 : Calculate coefficient // S(x) are used to calculate the coefficient of the function

3.2.2 : S(x) is defined as the combination of cubic polynomials $S_i(x)$

For the inner points x_i , $i=2,\ldots\ldots,n-1$

S1.1. S(x) interpolates the point $(x_i , f(x_i))$: $S_i(x_i) = f(x_i)$

S1.2. S(x) is continuous at x_i : $S_{i-1}(x_i) = S_i(x_i)$

S1.3. $S^{'}(x)$ is continuous at x_i : $S^{'}_{i-1}(x_i) = S^{'}_{i}(x_i)$

S1.4. $S^{''}(x)$ is continuous at x_i : $S^{''}_{i-1}(x_i) = S^{''}_{i}(x_i)$

3.2.3 else For the First and Endpoint missing //ค่าที่อยู่นอกช่วง

Call extrapolation for impute()

3.2.4 Repeat Step 2.2.1 for next row in the patient subspace matrix.

3.2.5 Repeat Step 2.1 for next the patient subspace matrix.

Step4 : repeat 2.1-2.7 until there is no missing value in each patient

Step5: Else if time treatment < = 2

// Call DPimpute();

5.1 Checking missing position

5.2 Checking missing position

5.3 Pad zero to NaN column

Step6. Compute N-dimension distance

Step 6.1 : vectorNaN = padZeroData(i,:);

Step 6.2 : sumData = zeros(ndata_row,1)

Step 6.3 : distanceData = padZeroData;

Step 6.4 : Loop and index indentifies the minimum distance row

for j=1:ndata_row

distanceData(j,:) = power((padZeroData(j,:) - vectorNaN),[2]);

sumData(j, 1) = sqrt(sum(distanceData(j,:)));

```
        mx = max(sumData);
        sumData(i,1) = mx;  % Avoid it will be minimum
    end
```

Step7 : Replace values in column which is NaN with data in index row, minimum distance.

```
    for j=1:c
        resultData(i, colNaN(1, j)) = distanceData(index, colNaN(1,j));
    end
```

ผลการทดลอง

ผลการทดลองด้วยตัววัดประสิทธิภาพการประมาณค่าด้วย NRMSE ด้วยการแบ่งเป็น % ของการที่ปรากฏค่าสูญหายคือ

ตารางที่ 4.8 ค่าความคลาดเคลื่อน(NRMSE) ของชุดข้อมูลลำดับเวลา(CBE-DP) ชุดที่ **1**

% missing / Dataset	10%	15%	20%	25%	30%	%Average
Obesity	0.15362	0.17672	0.17966	0.17689	0.19429	0.17624
Thrombosis	0.094873	0.096731	0.109234	0.124589	0.125196	0.11394

ตารางที่ 4.9 ค่าความคลาดเคลื่อน(NRMSE) ของชุดข้อมูลลำดับเวลา(CBE-DP) ชุดที่ 2

% missing / Dataset	10%	15%	20%	25%	30%	%Average
Obesity	0.17599	0.18636	0.19399	0.20148	0.20400	0.19236
Thrombosis	0.096174	0.104513	0.112150	0.117913	0.123097	0.11077

ตารางที่ 4.10 ค่าความคลาดเคลื่อน(NRMSE) ของชุดข้อมูลลำดับเวลา (CBE-DP) ชุดที่ 3

% missing / Dataset	10%	15%	20%	25%	30%	%Average
Obesity	0.16638	0.18788	0.17660	0.18513	0.22564	0.18833
Thrombosis	0.092212	0.104513	0.109459	0.112908	0.119322	0.10768

สรุปและอภิปรายผลการทดลอง

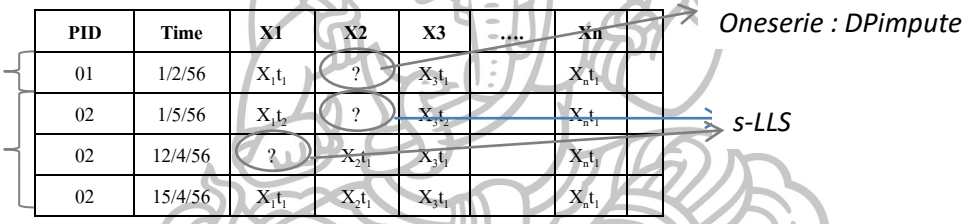
จากผลการทดลองในตารางที่ (CBE-DP) -DPimpute ดังตารางที่ 4.8-4.10 ซึ่งจากชุดข้อมูลทั้งสามชุดที่ตำแหน่งการสูญหายไม่ซ้ำกัน พิจารณาที่ค่าเฉลี่ยเปอร์เซนต์ค่าสูญหายของแต่ละชุดข้อมูลให้ค่าความคลาดเคลื่อนเฉลี่ยของชุดข้อมูล obesity ชุดที่ 1-3 ตามลำดับ คือ 0.17624, 0.19236 , 0.18833 และ ชุดข้อมูล thrombosis ชุดที่ 1-3 ตามลำดับ คือ 0.11394, 0.11077, 0.10768 มีค่าความคลาดเคลื่อนที่ใกล้เคียงกัน

จากการวิเคราะห์การประมาณค่าการตรวจด้วยวิธีการ (CBE-DP)

จากการวัดประสิทธิภาพค่าความคลาดเคลื่อนวิธีการ (CBE-DP) เมื่อเทียบกับวิธีการแรก วิธีการนี้ให้ค่าการประมาณความแม่นยำที่ใกล้เคียงวิธีที่ 1 ส่วนหนึ่งคือค่าจริงที่ปรากฏที่ใช้ในการประมาณค่าจะใช้ค่าที่อยู่ระหว่างตำแหน่งที่ปรากฏค่าสูญหาย ไม่ได้นำทุกค่ามาคำนวณด้วยสมการคิวบิกสไปน์ แต่มีข้อจำกัดของวิธีการคือต้องแยกการประมาณค่าหลายจุด คือ ค่าสูญหายตำแหน่งแรก , ค่าสูญหายตำแหน่งสุดท้าย, ค่าสูญหายตำแหน่งระหว่างกลาง และค่าสูญหายที่ปรากฏในกรณีที่ผู้มาตรวจครั้งเดียว ซึ่งทำให้เกิดเงื่อนไขของการประมาณค่าหลายกรณี ซึ่งจะมีผลต่อค่าการประมาณ สำหรับคนที่มาตรวจครั้งเดียวจะใช้วิธีการ DPimpute มาประมาณค่า ซึ่งจะต้องคำนวณกับทุกแถวที่ปรากฏ ดังนั้นค่าที่เคยถูกประมาณไว้แล้วสำหรับผู้ป่วยที่มีการตรวจหลายครั้ง ก็จะมีผลหากตำแหน่งที่สูญหายและระยะทางต่ำสุดไปตรงกับตำแหน่งที่เคยประมาณไว้แล้วด้วย CBE หากประมาณได้ค่าใกล้เคียง ก็มีผลทำการประเมินค่าความคลาดเคลื่อนต่ำไปด้วย

4. วิธีการประมาณค่าสูญหาย SLLS-DP (Subspace Local Least Squares Distance Imputation)

ขั้นตอนประมวลผลของ Subspace Least Squares Imputation ประมาณค่าข้อมูลสูญหายด้วยการสร้างสมการถดถอยเชิงเส้น เพื่อคำนวณหาค่าสัมประสิทธิ์ซึ่งการประมาณค่าสูญหายโดยอาศัยสมการเส้นตรงที่เหมาะสมใช้วิธีกำลังสองน้อยที่สุดซึ่งทำให้ผลบวกทั้งหมดยกกำลังสองของระยะทางระหว่างจุด (x,y) กับเส้นตรงเส้นหนึ่งมีค่าน้อยที่สุด บนแนวคิดการใช้ค่าข้อมูลของแต่ละคน ซึ่งวิธีการนี้สามารถนำค่าทุกค่าในแต่ละตัวชี้วัดมาทำคำนวณได้ และนำมาประมาณค่าที่สูญหายของข้อมูลแต่ละบุคคล จากแต่ละแถวในผู้ป่วยแต่ละคน ส่วนข้อมูลผู้ที่มาตรวจรักษาเพียงครั้งเดียว หากปรากฏค่าสูญหายใช้แนวคิดการใช้ค่าความเหมือนและความคล้ายของค่าการตรวจจากชุดข้อมูลที่มีอยู่ด้วยหลักการ DP-impute ดังรายละเอียดในแผนภาพที่ 4.5



ภาพที่ 4.5: แสดงแนวคิดการประมาณค่าสูญหายด้วยวิธีการ SLLS-DPimpute

ดังนั้นการพัฒนาขั้นตอนการประมาณค่าสูญหายด้วยวิธี SLLS-DPimpute บนชุดข้อมูลลำดับเวลานี้บนแนวคิดการประมาณค่าจะใช้เฉพาะค่าการตรวจเฉพาะบุคคลมาใช้ในกระบวนการประมาณค่าสูญหาย โดยจะแบ่งการประมาณค่าสูญหายด้วย 2 เงื่อนไขหลัก คือ กรณีผู้ป่วยที่มาตรวจรักษามากกว่า 2 ครั้ง และ ผู้ป่วยที่มาตรวจรักษาเพียงครั้งเดียว ดังรายละเอียดในแผนภาพและขั้นตอนดังนี้คือ

ขั้นตอนในภาพรวมมีดังนี้คือ

1. นำชุดข้อมูลเชิงเวลาแยกเป็นเมตริกซ์ข้อมูลในลักษณะรายคน
2. ถ้าจำนวนการตรวจของแต่ละคนมากกว่า 2 ครั้ง
3. ทรานโพสเมตริกซ์
4. เลือก row vectorของแต่ละคน

5. ตรวจสอบตำแหน่งที่ปรากฏค่าสูญหาย

5.1 แยกค่าจริงที่ปรากฏ (observe value) และค่าตำแหน่งเวลาที่ปรากฏค่าสูญหาย(t_{miss})

5.2 นำค่าจริงที่ปรากฏทั้งหมด (observe values) ประมาณค่าด้วย least square

5.3 แทนที่ค่าที่ได้ในตำแหน่งเวลาที่ปรากฏค่าสูญหาย (t_{miss})

6. วนรอบข้อ 4 จนกว่าค่าข้อมูลจะครบสมบูรณ์

7. ถ้าจำนวนการตรวจของแต่ละคน ตรวจครั้งเดียวหรือไม่เกิน2 ครั้ง

ประมาณค่าสูญหายด้วยวิธีการ DPimpute

ขั้นตอนในรายละเอียดดัง procedure ดังต่อไปนี้

Procedure : SLLS-DPimpute

Input : Incomplete temporal dataset (D)

Output : Complete temporal dataset(D)

Step1 : Transform temporal medical data to low dimension in subspace matrix

1.1 Determine the system's input variables for a temporal data matrix

1.2 Separate the matrix into m x n each of patient subspace dimension.

1.3 Transpose patient subspaces with missing values ,

// column vector are time point in the treatment ,

the row vectors are measurement variables from temporal data.

[$x(t_1)$ $x(t_2)$ $x(t_3)$ …… $x(t_n)$]

Step2 : Compute in transposed patient subspace.

2.1 Repeat

2.2 Select the transposed subspace in each patient case for imputation.

2.2.1 Repeat

2.2.2 Select row vector in each patient subspace

// $X_n = [\, x_i(t_1)\ x_i(t_2)\ x_i(t_3) \ldots\ldots x_i(t_n)]$ where $t_i = 1,2,\ldots,n$

Step3 : Start fill missing

3.1 Check time- treatment of Patient –ID-$_{0n}$ and missing position

// Y_n at $x(t) = [\ x_{i1}\ ?\ x_{i3}\ ?\ \ldots.\ x_{in}\ ]^T$

3.2 if time tratement of patient > 2

3.3 Select a row vector in subspace patient case series for example observe a point

$y_{obs} = x(t_1)\ ,\ x(t_3)\ ,\ x(t_n)\ ,\ y_{t\text{-}miss} = x(t_2) = ?\ ,\ x(t_4) = ?$

3.4 Consider polynomial of fixed n to find values of a,b

$y = a_0 + a_1x + a_2x^2 + \ldots\ldots\ldots + a_mx^m$, n data points

$(x_1,y_1),(x_2,y_2),\ldots\ldots,(x_n,y_n)$

$y = Xb$, polynomial equation

$Xy = X^tX_b$

$(X^tx)^{-1}X^ty = (X^tx)^{-1}(X^tx)\,b$

$b = (X^tX)^{-1}X^ty$

3.5 Select $X(t_{obs})$ in row vector have observe values

$Y(t_{n\text{-}obs}) = y(t_{n\text{-}obs}) = a_0 + (t_{n\text{-}obs})\,a_1 + (t_{n\text{-}obs})^2a_2 + (t_{n\text{-}obs})^3a_3$

3,6 Calculate the equations with lease square matrix and replace with the equation to find b (-1 is mxm inverst matrix)

$b = (x^tx)^{-1}x^ty$

3.7 From 3.4, a_0,a1,a_2 as a polynomial equation

$y = a_0 - a_1x - a_2x^2 - a_3x^3$ //edit a value

3.8 Replace *x* variable in the equation with the time point with missing data (t_{miss})

3.9 Repeat Step 3.1.2 for next sub-set of series until last sub-set

3.10 Repeat Step 2.2.1 for next row in the patient subspace matrix.

3.1.1 Repeat Step 2.1 for next the patient subspace matrix.

Step4 : Else if time treatment < = 2

// Call DPimpute();

Step5: Else if time treatment < = 2

// Call DPimpute();

5.1 Checking missing position

5.2 Pad zero to NaN column

Step6. Compute N-dimension distance

Step 6.1 : vectorNaN = padZeroData(i,:);

Step 6.2 : sumData = zeros(ndata_row,1)

Step 6.3 : distanceData = padZeroData;

Step 6.4 : Loop and index indentifies the minimum distance row

```
for j=1:ndata_row
distanceData(j,:) = power((padZeroData(j,:) - vectorNaN),[2]);
sumData(j, 1) = sqrt(sum(distanceData(j,:)));
mx = max(sumData);
sumData(i,1) = mx; % Avoid it will be minimum
end
```

Step7 : Replace values in column which is NaN with data in index row, minimum distance.

```
for j=1:c
resultData(i, colNaN(1, j)) = distanceData(index, colNaN(1,j));
end
```

Step8 : Repeat 2.1-2.7 until there is no missing value in each patient

ผลการทดลอง

ผลการทดลองด้วยตัววัดประสิทธิภาพการประมาณค่าด้วย NRSME ด้วย การแบ่งเป็น % ของการที่ปรากฏค่าสูญหายคือ

ตารางที่ 4.11 ค่าความคลาดเคลื่อน(NRMSE) ของชุดข้อมูลลำดับเวลา(SLLS-DPimpute) ชุดที่ 1

% missing / Dataset	10%	15%	20%	25%	30%	%Average
Obesity	0.58630	0.59480	0.60225	0.60898	0.59978	0.59842
Thrombosis	2.488174	2.448375	2.467086	2.483993	2.462677	2.47006

ตารางที่ 4.12 ค่าความคลาดเคลื่อน(NRMSE) ของชุดข้อมูลลำดับเวลา(SLLS-DPimpute) ชุดที่ 2

% missing / Dataset	10%	15%	20%	25%	30%	%Average
Obesity	0.57576	0.59875	.59103	0.58903	0.59522	0.58996
Thrombosis	2.453547	2.489672	2.464771	2.468265	2.439718	2.46319

ตารางที่ 4.13 ค่าความคลาดเคลื่อน(NRMSE) ของชุดข้อมูลลำดับเวลา (SLLS-DPimpute) ชุดที่ 3

% missing / Dataset	10%	15%	20%	25%	30%	%Average
Obesity	0.51584	0.58924	0.60225	0.60235	0.58253	0.57844
Thrombosis	2.439168	2.423025	2.453055	2.487622	2.427927	2.44616

สรุปและอภิปรายผลการทดลอง

จากผลการทดลองในตารางที่ (SLLS-DPimpute) ดังตารางที่ 4.11-4.13 ซึ่งจากชุดข้อมูลทั้งสามชุดที่ตำแหน่งการสูญหายไม่ซ้ำกัน พิจารณาที่ค่าเฉลี่ยเปอร์เซนต์ค่าสูญหายของแต่ละชุดข้อมูลให้ค่าความคลาดเคลื่อนเฉลี่ยของชุดข้อมูล obesity ชุดที่1-3 ตามลำดับ คือ 0.59842, 0.58996, 0.57844 และ ชุดข้อมูล thrombosis ชุดที่ 1-3 ตามลำดับ คือ 2.47006, 2.46319, 2.44616 มีค่าความคลาดเคลื่อนที่ใกล้เคียงกัน

จากการวิเคราะห์การประมาณค่าการตรวจด้วยวิธีการ (SLLS-DPimpute)

จากการวัดประสิทธิภาพค่าความคลาดเคลื่อนวิธีการ (SLLS-DPimpute) เมื่อเทียบกับวิธีการแรก วิธีการนี้ให้ค่าการประมาณค่าความคลาดเคลื่อนค่อนข้างสูง ข้อดีของวิธีการนี้จะนำค่าข้อมูลที่ปรากฏจริงทุกค่าในแต่ละตัวชี้วัดของแต่ละคน มาทำคำนวณด้วยสมการเชิงถดถอยด้วยรูปแบบสมการวิธีคำนวณของสมการเชิงถดถอยกับการนำมาคำนวณเพื่อใช้ในการประมาณค่าข้อมูล มีผลทำให้ค่าการประมาณในภาพรวมค่อนข้างสูง ดังนั้นในกรณีที่ผู้มาตรวจครั้งเดียว ซึ่งจะใช้วิธีการ DPimpute มาประมาณค่า จะต้องคำนวณกับทุกแถวที่ปรากฏรวมทั้งค่าที่เคยถูกประมาณไว้แล้วสำหรับผู้ป่วยที่มีการตรวจหลายครั้ง ก็จะมีผลหากตำแหน่งที่สูญหายและระยะทางต่ำสุดไปตรงกับตำแหน่งที่เคยประมาณไว้แล้วด้วย SLLS ซึ่งเมื่อประมาณค่าแล้วส่วนใหญ่ได้ค่าที่ห่างก็จะมีผลทำการประเมินค่าความคลาดเคลื่อนสูงไปด้วย

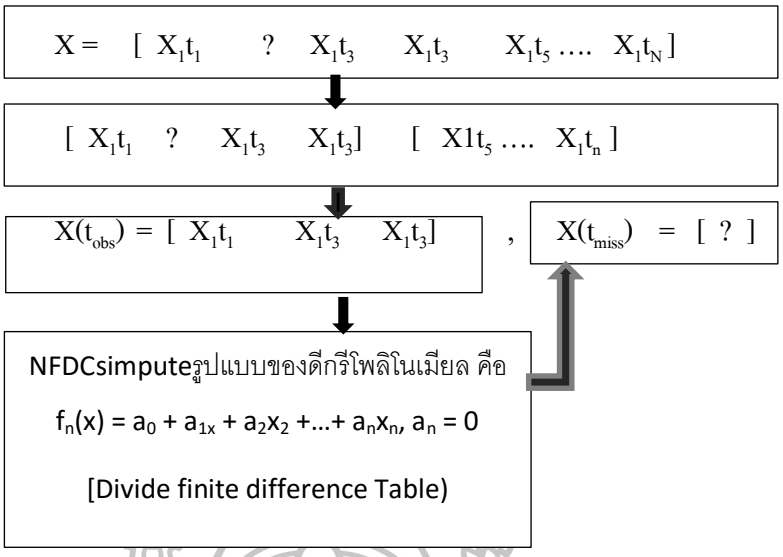
5. วิธีการประมาณค่าสูญหาย NFDCs-DPimpute

วิธีการนี้จะใช้ขั้นตอนวิธีในการประมาณค่าและทดแทนข้อมูลสูญหายในตำแหน่งที่เกิดค่าสูญหายบนหลักการประมาณค่าในช่วง ด้วยโพลิโนเมียลจากตารางการแบ่งย่อยของนิวตัน ด้วยเงื่อนไขของจำนวนครั้งการตรวจรักษาเป็นตัวระบุโพลิโนเมียลดีกรีที่ใช้ในการประมาณค่าสำหรับผู้ป่วยที่ค่าการตรวจมากกว่า 2 ครั้งขึ้นไป

สำหรับรายละเอียดขั้นตอนใน procedure เพื่อประยุกต์ใช้โพลิโนเมียลสำหรับการประมาณค่าในตำแหน่งที่ปรากฏค่าสูญหาย โดยกำหนดให้ชุดของข้อมูลลำดับลำดับเวลา คือ n ข้อมูล $(x_1, y_1), \ldots, (x_n, y_n)$, บนหลักการของนิวตันโพลิโนเมียลที่จะทำการประมาณฟังก์ชัน f(x) ณ ตำแหน่ง x ด้วยการใช้ divided difference table ส่วนของการประมวลผลในการประมาณค่า

ใช้เงื่อนไขของการประมาณด้วยเงื่อนไขลำดับของดีกรี (Condition order degree) เงื่อนไขของการคำนวณเริ่มต้น เพื่อหาโพลิโนเมียลจากค่าข้อมูลจริงในแต่ละตัวแปรของแต่ละเวลาของการตรวจรักษาจากของผู้ป่วยแต่ละคน สิ่งนี้ทำได้ด้วยการใช้ order degree สูงสุดโดยดูจากเวลาการตรวจรักษาทั้งหมดของคนไข้แต่ละคน ซึ่งการประมวลผลจะใช้ค่าข้อมูลที่เป็นค่าต่อเนื่องตามเวลาจำนวนครั้งของการตรวจรักษาเป็นค่าที่ใช้ในการประมาณฟังก์ชัน คือ $x(t) = [xt_1,\ldots,xt_{n+1}]$ รูปแบบของดีกรีโพลิโนเมียล คือ $f_n(x) = a_0 + a_{1x} + a_2x_2 +\ldots+ a_nx_n$, $a_n = 0$. สำหรับ n+1 data point นั่นคือ ฟังก์ชันในการประมาณค่าในช่วงตาม n ดีกรีโพลิโนเมียล ตาม n ตำแหน่ง นั่นคือจำนวนจุดคือจำนวนครั้งของการตรวจรักษาในโครงสร้างการประมาณค่าสูญหายเรียกว่า ลำดับของการประมาณค่าในช่วง นั่นคือ กรณี 1 linear interpolation, $y = a_0 + a_1x$, จำนวน 2 จุด และ quadratic interpolation, $y = a_0 + a_1x + a_2x_2$, ใช้จำนวน 3 จุด และ cubic interpolation, $y = a_0 + a_1x + a_2x_2 + a_3x_3$, ใช้จำนวน 4 จุด หากมากกว่า 4 จุด ในการพบ polynomial $f_n(x)$

และต่อมาจากทดลองพบว่าหากกรณีชุดข้อมูลที่มีการตรวจรักษาระยะเวลานาน นั่นคือ n ดีกรีโพลิโนเมียล และค่าข้อมูลการตรวจที่ไม่ต่อเนื่อง หากนำทุกค่าจริงที่ปรากฏระยะยาวมาประมาณค่า ซึ่งผลการประมาณมีแนวโน้มจะไม่ใกล้เคียงเนื่องจากดีกรีโพลิโนเมียลที่คำนวณออกมา ซึ่งมีผลต่อชุดข้อมูลทั้งหมด เมื่อประเมินค่าความคลาดเคลื่อน จึงมีแนวคิดหากแบ่งข้อมูลเป็นเซตย่อยตามเงื่อนไขดีกรีโพลิโนเมียล โดยได้ทดลองแบ่งเซตย่อยเป็นหลายๆ ค่า พบว่าเซตละประมาณ 3 ค่า คือ ลำดับโพลิโนเมียลดีกรี 2 จะให้ค่าการประมาณที่ดีสำหรับผู้ป่วยที่มาตรวจหลายครั้ง[24] แต่หากค่าสูญหายในเซตไม่ถึง 3 ค่าและสำหรับข้อมูลผู้ที่มาตรวจเพียงครั้งเดียวและผู้ป่วยที่มาทำการตรวจรักษาไม่เกินสองครั้ง นำหลักการวัดค่าความคล้ายด้วยระยะทางบนแนวคิดนำค่าเฉพาะรายบุคคล หรือค่าของบุคคลอื่นที่มีค่าการตรวจที่เหมือนและใกล้เคียงกันมาเป็นตัวประมาณค่าสูญหายบนหลักการ DPimpute ด้วยการคำนวณโดยระยะทางปริภูมิหลายมิติ (n-dimensional Euclidean distance) เพื่อตรวจสอบค่าความเหมือนและความคล้ายของข้อมูลและนำค่าข้อมูลจริงในตำแหน่งแถวและตัวแปรที่มีค่าความเหมือนมาแทนที่ในตำแหน่งที่สูญหาย ดังรายละเอียดในแผนภาพ



ภาพที่ 4.6 แสดงแนวคิดการประมาณค่าสูญหายด้วยวิธีการ NFDCs-DPimpute

ขั้นตอนในภาพรวมมีดังนี้คือ

1. นำชุดข้อมูลเชิงเวลาแยกเป็นเมตริกซ์ข้อมูลในลักษณะรายคน
2. ทรานโพสเมตริกซ์
3. ถ้าจำนวนการตรวจของแต่ละคนมากกว่า 2 ครั้ง
4. เลือก row vectorของแต่ละคน
5. แบ่งเป็นเซตชุดย่อย ชุดย่อยละ 3
    5.1 ตรวจสอบตำแหน่งที่ปรากฏค่าสูญหายและประมาณค่าทีละชุดย่อย
    5.2 แยกค่าจริงที่ปรากฏ (observe value) และค่าตำแหน่งเวลาที่ปรากฏค่าสูญหาย(t_{miss})
    5.3 นำค่าจริงที่ปรากฏทั้งหมด (Xt_{miss}) ประมาณค่าด้วย finite divide difference(NFDC)
    5.4 แทนที่ค่าที่ได้ในตำแหน่งเวลาที่ปรากฏค่าสูญหาย (Xt_{miss})
6. วนรอบข้อ 4 จนกว่าค่าข้อมูลจะครบสมบูรณ์
7. วนรอบข้อ ข้อ 3 จนกว่าจะครบทุกแถว และทุกคน
8. ถ้าจำนวนการตรวจของแต่ละคน ตรวจครั้งเดียวหรือไม่เกิน2 ครั้ง

    ประมาณค่าสูญหายด้วยวิธีการ DPimpute

ขั้นตอนในรายละเอียดดัง procedure ดังต่อไปนี้

Procedure : NFDCs-DPimputation

Input : Incomplete temporal dataset (D)

Output : Complete temporal dataset(D)

Step1 : Transform temporal medical data to low dimension in subspace matrix

1.1 Determine the system's input variables for a temporal data matrix

1.2 Separate the matrix into m x n each of patient subspace dimension.

1.3 Transpose patient subspaces with missing values ,

$[x(t_1)\ x(t_2)\ x(t_3) \ldots\ldots x(t_n)]$

Step2 : Compute in transposed patient subspace.

2.1 Repeat

2.2 Select the transposed subspace in each patient case for imputation.

2.2.1 Repeat

2.2.2 Select row vector in each patient subspace

// $X_n = [x_i(t_1)\ x_i(t_2)\ x_i(t_3) \ldots\ldots x_i(t_n)]$ where $t_i = 1,2,\ldots,n$

Step3 : Check time- treatment of Patient –ID-$_{0n}$ and missing position

// Y at $x(t) = [\ ?\quad x_{i1}\quad x_{i2}\quad x_{i3}\quad \ldots.\ x_{in}\ ]^T$

3.1 if time tratement of patient > 2

3.1.1 Seperate X_n to subset of row vector (4 of set series)

$$FD = X_i(t_i)_{max} - X_i(t_i)_{min}$$

$$FDDV = FD/3$$

3.1.2 Repeat

3.1.2.1 Separate data in two set on $X_{obs}(t_i)$, $X_{miss}(t_i)$

//Y at x(t) = [? x_{i1} x_{i2} x_{i3} …. x_{in}]^T

X_{obs} : This shoulde be contain a set with a occure feature and no missing values.

t_i = [1 2 3]

$x(t_i)_{obs}$ = [$x(t_i)_{obs}$ $x(t_{i+1})_{obs}$ … $x(t_{i+n})_{obs}$]

$y(t_i)_{obs}$ = [$y(t_i)_{obs}$ $y(t_{i+1})_{obs}$ …. $y(t_{i+n})_{obs}$]

,Where $x(t_i)$ is time of treatement , $y(t_i)$ is values at $x(t_i)_{obs}$

X_{miss} : This shoulde be contain a set with a feature and missing values.

t_i = [t_i]

$x(t_i)_{miss}$ = [$(t_i)_{miss}$]

$y(t_i)_{miss}$ = [$y(t_i)_{miss}$ (?)]

,Where $x(t_i)$ is time of treatement , $y(t_i)$ is values at $x(t_i)_{miss}$

3.1.2.2 compute on $X_{obs}(t_i)$ // for polynomial with set of observe values feature are $X_{obs}(t_i)$, $y_{obs}(t_i)$ with condition order degree on Newton's divide difference table

Condition order degree

observe treatment values at the n point is n-th degree using n-th interpolation

$$f_n(x) = b_1 + b_2(x- x(t_{1\text{-}obs}))+b_3(x- x(t_{1\text{-}obs})))(x- x(t_{2\text{-}obs})))+...+b_n(x- x(t_{1\text{-}obs}))+)...(x-x\ (t_{n\text{-}1\text{-}obs}))$$

3.1.2.3 Compute $x(t)_{miss}$ // from set of missing values features .

Compute $f_n(x(t_i))$ with the condition order degree by observer values of n points for using n degree, that is , $f_n(x)$, $x(t_i)$miss

$$f_n(x) = f_n(x(t_i))$$

Compute $f_n(x)$

$$f_n(x) = \sum_{i=1}^{n} \left\{ F[x_1, x_2, ..., x_i] \prod_{j=1}^{i-1} (x - x_j) \right\},$$

$$f_n(x) = f(x_1) + (x - x_1) f[x_2, x_1] + (x - x_1)(x - x_2) f[x_3, x_2, x_1] + \cdots + (x - x_1)(x - x_2) \cdots (x - x_{n-1}) f[x_n, x_{n-1}, \cdots, x_1].$$

3.1.2.4 Replace the values with $f_n(x(t_i))$ in missing position [$X((t_i))_{mis}$]

3.1.3 Repeat Step 3.1.2 for next sub-set of series until last sub-set

3.1.4 Repeat Step 2.2.1 for next row in the patient subspace matrix.

3.1.5 Repeat Step 2.1 for next the patient subspace matrix.

3.2 Else if time treatment < = 2

Call DPimpute()

Setp5 : Loop 3.1 next patient one-has missing Until all patient no missing

ผลการทดลอง

ผลการทดลองด้วยตัววัดประสิทธิภาพการประมาณค่าด้วย NRSME ด้วยการแบ่งเป็น % ของการที่ปรากฏค่าสูญหายคือ

ตารางที่ 4.14 ค่าความคลาดเคลื่อน(NRMSE) ของชุดข้อมูลลำดับเวลา(NFDCs-DP) ชุดที่ 1

% missing / Dataset	10%	15%	20%	25%	30%	%Average
Obesity	0.10937	0.10908	0.10572	0.10790	0.10405	0.10722
Thrombosis	0.088264	0.086666	0.086861	0.086346	0.088185	0.08726

ตารางที่ 4.15 ค่าความคลาดเคลื่อน(NRMSE) ของชุดข้อมูลลำดับเวลา (NFDCs-DP) ชุดที่ 2

% missing / Dataset	10%	15%	20%	25%	30%	%Average
Obesity	0.11005	0.10533	0.10807	0.10764	0.10253	0.10673
Thrombosis	0.087300	0.085698	0.084505	0.086512	0.089347	0.08667

ตารางที่ 4.16 ค่าความคลาดเคลื่อน(NRMSE) ของชุดข้อมูลลำดับเวลา (NFDCs-DP) ชุดที่ 3

% missing / Dataset	10%	15%	20%	25%	30%	%Average
Obesity	0.11068	0.11439	0.10711	0.10661	0.10511	0.1087
Thrombosis	0.088077	0.087218	0.086881	0.086741	0.087557	0.08729

สรุปและอภิปรายผลการทดลอง

จากผลการทดลองในตารางที่ (NFDCs -DPimpute) ดังตารางที่ 4.14-4.16 ซึ่งจากชุดข้อมูลทั้งสามชุดที่ตำแหน่งการสูญหายไม่ซ้ำกัน พิจารณาที่ค่าเฉลี่ยเปอร์เซนต์ค่าสูญหายของแต่ละชุดข้อมูลให้ค่าความคลาดเคลื่อนเฉลี่ยของชุดข้อมูล obesity ชุดที่ 1-3 ตามลำดับ คือ 0.10722, 0.10673, 0.1087 และ ชุดข้อมูล thrombosis ชุดที่ 1-3 ตามลำดับ คือ 0.08726, 0.08667, 0.08729 มีค่าความคลาดเคลื่อนที่ใกล้เคียงกัน

จากการวิเคราะห์การประมาณค่าการตรวจด้วยวิธีการ (NFDCs -DPimpute)

จากการวัดประสิทธิภาพค่าความคลาดเคลื่อนวิธีการ (NFDCs -DPimpute) เมื่อเทียบกับวิธีการ DPimpute, SKnn-DPimpute, SLLS-DPimpute, วิธีการนี้ให้ค่าการประมาณค่าความคลาดเคลื่อนค่อนข้างต่ำ นั่นคือ NFDCimpute นั่นคือพบว่าหากกรณีชุดข้อมูลที่มีการตรวจรักษาระยะเวลานาน นั่นคือ n ดีกรีโพลิโนเมียล และค่าข้อมูลการตรวจที่ไม่ต่อเนื่อง หากนำทุกค่าจริงที่

ปรากฏระยะยาวมาประมาณค่าด้วยสมการในตาราง divide finite difference table ซึ่งผลการประมาณมีแนวโน้มจะไม่ใกล้เคียง เนื่องจากดีกรีโพลิโนเมียลที่คำนวณออกมาให้ค่าตัวเลขค่อนข้างสูง แต่หากจำนวนครั้งการตรวจเป็นการตรวจระยะสั้น(short-term) จากการทดลองวิธีการนี้ให้ค่าการประมาณค่อนข้างใกล้เคียง ส่วนต่อมา NFDCs-DPimpute แบ่งชุดข้อมูลเป็นเซตย่อยเซตละสามค่า ในแต่ละตัวชี้วัดของแต่ละคนด้วยดีกรีโพลิเมียลดีกรี 3 ซึ่งพบว่าให้ค่าการคำนวณในการประมาณค่าสูญหายค่อนข้างใกล้เคียง เมื่อเทียบกับค่าจริงในชุดข้อมูลสมบูรณ์ชุดเดิม ซึ่งเหมาะกับการนำมาใช้กับข้อมูลการตรวจของผู้ป่วยระยะยาว แต่ปรากฏค่าสูญหาย ดังนั้นเมื่อได้ค่าการประมาณที่ใกล้เคียงจากผู้ป่วยที่มาตรวจหลายครั้งแล้ว ดังนั้นเมื่อปรากฏค่าสูญหายในผู้ป่วยที่มาตรวจไม่เกินสองครั้งและเซตข้อมูลย่อยในตัวชี้วัดปรากฏค่าสูญหายมากกว่า 1 ค่าจะใช้วิธีการ DPimpute มาประมาณค่าซึ่งจะเทียบกับค่าที่ได้ทำการประมาณค่าด้วย NFDCs สำหรับที่มาตรวจหลายครั้งและประมาณค่าด้วย NFDCs แล้วก็จะมีผลทำการประเมินค่าความคลาดเคลื่อนต่ำคือมีค่าเหมือนหรือใกล้เคียงไปด้วย

6. วิธีการประมาณค่าสูญหาย NFDCslideW-DPimpute

ขั้นตอน NFDCslideW-DPimpute นำหลักการ Sliding windows มาใช้ในการเลื่อนเป็นหน้าต่างเสมือนที่เลื่อนไปตามแถวของข้อมูลที่มีหน้าต่างเลื่อน (window) สำหรับข้อมูลการตรวจหลายครั้ง กำหนดขนาด w แทนค่าด้วยดีกรีโพลิโนเมียลคือ 3 จากการกำหนดชุดข้อมูลย่อยในวิธีการ NFDCs-impute โดยมีตำแหน่งเริ่มต้นด้านซ้ายสุดของแถวลำดับเวลา

สำหรับข้อมูลผู้ที่มาตรวจเพียงครั้งเดียวและผู้ป่วยที่มาทำการตรวจรักษาไม่เกินสองครั้งและหากในวินโดว์ปรากฏค่าสูญหายทั้งหมดนำหลักการ DPimpute ด้วยการคำนวณโดยระยะทางปริภูมิหลายมิติ เพื่อตรวจสอบค่าความเหมือนและความคล้ายของข้อมูลและนำค่าข้อมูลจริงในตำแหน่งแถวและตัวแปรที่มีค่าความเหมือนมาแทนที่ในตำแหน่งที่สูญหาย

แสดงการทำงานด้วยแผนภาพได้ดังนี้

$X_i = [\ X_1t_1 \quad X_1t_2 \quad ? \quad X_1t_4 \quad X_1t_5 \quad X_1t_6 \quad ? \quad X_1t_N\ ]$

1 .เลื่อนสไลด์จัดการแบ่งข้อมูลในลักษณะวินโดว์(W=3)

จนกว่าจะเจอข้อมูลตัวสุดท้าย

$W1X_i = [\ X_1t_1 \quad X_1t_2 \quad ?\ ]$

$W2X_i = [\ X_1t_2 \quad ? \quad X_1t_4\ ]$

$W3X_i = [\ ? \quad X_1t_4 \quad X_1t_5\ ]$

$W4X_i = [\ X_1t_4 \quad X_1t_5 \quad X_1t_6\ ]$

$W5X_i = [\ X_1t_5 \quad X_1t_6 \quad ?\ ]$

$W6X_i = [\ X_1t_6 \quad ? \quad X_1t_N\ ]$

2. ตรวจสอบแต่ละวินโดว์ที่ปรากฏค่าสูญหาย ในตำแหน่งแรกและในวินโดว์สุดท้าย

3. เลือกวินโดว์ที่ปรากฏค่าสูญหายในตำแหน่งแรก และวินโดว์สุดท้ายที่ปรากฏ

$W3X_i = [\ ? \quad X_1t_4 \quad X_1t_5\ ]$

$W6X_i = [\ X_1t_6 \quad ? \quad X_1t_N\ ]$

4. ประมาณค่าด้วย NFDC

$W3X_i = [\ ? \quad X_1t_4 \quad X_1t_5\ ]$ → NFDCsimpute

$W6X_i = [\ X_1t_6 \quad ? \quad X_1t_N\ ]$ → NFDCsimpute

5. แทนที่ค่า(Replace value)

$W3X_i = [\ X_1t_{3\text{-impute}} \quad X_1t_4 \quad X_1t_5\ ]$

$W6X_i = [\ X_1t_6 \quad X_1t_{7\text{-impute}} \quad X_1t_N\ ]$

แผนภาพที่ 4.7: แสดงแนวคิดการประมาณค่าสูญหายด้วยวิธีการ NFDCslideW-DPimpute

ขั้นตอนในภาพรวมมีดังนี้คือ

1. นำชุดข้อมูลเชิงเวลาแยกเป็นเมตริกซ์ข้อมูลในลักษณะรายคน
2. ทรานโพสเมตริกซ์
3. ตรวจสอบจำนวนการตรวจรักษา ถ้าจำนวนการตรวจของแต่ละคนมากกว่า 2 ครั้ง
4. เลือกทีละ row vector แทนค่าตัวชี้วัดของแต่ละคน

5. เลื่อนตำแหน่งและประมาณค่า

5.1 เลื่อนตำแหน่งข้อมูลเพื่อแยกชุดข้อมูลตามหลัก Slideing window ในแต่ละแถวของค่าตัวชี้วัดของแต่ละคนวินโดว์ละ 3 ค่า

5.2 ตรวจสอบตำแหน่งปรากฏค่าสูญหายในตำแหน่งแรกของแต่ละวินโดว์และวินโดว์สุดท้ายที่ปรากฏค่าสูญหาย

5.3 คำนวณประมาณค่าใน 5.2 ด้วย NFDCsImpute แทนที่ค่าที่ได้ในตำแหน่งเวลาที่ปรากฏค่าสูญหาย (Xt_{miss})

6. วนรอบข้อ 4 จนกว่าค่าข้อมูลจะครบสมบูรณ์

7. วนรอบข้อ ข้อ 3 จนกว่าจะครบทุกแถว และทุกคน

8. ถ้าจำนวนการตรวจของแต่ละคน ตรวจครั้งเดียวหรือไม่เกิน2 ครั้ง หรือตำแหน่งในวินโดว์ที่ปรากฏค่าสูญหายมากกว่า 1 ค่า

ประมาณค่าสูญหายด้วยวิธีการ DPimpute

ขั้นตอนในรายละเอียดดัง procedure ดังต่อไปนี้

Procedure : NFDCsSlideWimpute()

Input : Incomplete temporal dataset (D)

Output : Complete temporal dataset(D)

Step1 : Transform temporal medical data to low dimension in subspace matrix

1.1 Determine the system's input variables for a temporal data matrix

1.2 Separate the matrix into m x n each of patient subspace dimension

1.3 Transpose patient subspaces with missing values , $[x(t_1)\ x(t_2)\ x(t_3) \ldots\ldots x(t_n)]$

Step2 : Compute in transposed patient subspace.

2.1 Repeat

2.2 Select the transposed subspace in each patient case for imputation.

2.2.1 Repeat

2.2.2 Select row vector in each patient subspace

// X_n = [$x_i(t_1)$ $x_i(t_2)$ $x_i(t_3)$ …… $x_i(t_n)$] where $t_i = 1,2,\ldots,n$

Step3 : Check time- treatment of Patient –ID-$_{0n}$ and missing position

// Y at x(t) = [? x_{i1} x_{i2} x_{i3} …. x_{in}]^T

3.1 if time treatement of patient > 2

3.1.1 Separate data with Sliding window on row vector

3.1.2 Check each window have missing position in the first

3.1.3 Select window have missing position in the first and Select the last W

3.1.4 NFDCsImpute

3.1.4.1 Repeat

3.1.4.2 Separate data in set of window from 3.1.3 on $X_{obs}(t_i)$, $X_{miss}(t_i)$

//Y at x(t) = [? x_{i1} x_{i2} x_{i3} …. x_{in}]^T

X_{obs} : This shoulde be contain a set with a occure feature and no missing values.

t_i = [1 2 3]

$x(t_i)_{obs}$ = [$x(t_i)_{obs}$ … $x(t_{i+n})_{obs}$]

$y(t_i)_{obs}$ = [$y(t_i)_{obs}$ …. $y(t_{i+n})_{obs}$]

,Where $x(t_i)$ is time of treatement , $y(t_i)$ is values at $x(t_i)_{obs}$

X_{miss} : This shoulde be contain a set with a feature and missing values.

t_i = [t_i]

$x(t_i)_{miss}$ = [$(t_i)_{miss}$]

$y(t_i)_{miss}$ = [$y(t_i)_{miss}$ (?)]

,Where $x(t_i)$ is time of treatement , $y(t_i)$ is values at $x(t_i)_{miss}$

3.1.4.3 compute on $X_{obs}(t_i)$ // for polynomial with set of observe values feature are $X_{obs}(t_i)$, $y_{obs}(t_i)$ with condition order degree on Newton's divide difference table

Condition order degree

observe treatment values at the n point is n-th degree using n-th interpolation

$$f_n(x) = b_1 + b_2(x - x(t_{1\text{-}obs})) + b_3(x - x(t_{1\text{-}obs}))(x - x(t_{2\text{-}obs})) + \ldots + b_n(x - x(t_{1\text{-}obs})) + \ldots (x - x(t_{n\text{-}1\text{-}obs}))$$

3.1.4.4 Compute $x(t)_{miss}$ // from set of missing values features .

Compute $f_n(x(t_i))$ with the condition order degree by observer values of n points for using n degree, that is , $f_n(x)$, $x(t_i)$miss

$$f_n(x) = f_n(x(t_i))$$

Compute $f_n(x)$

$$f_n(x) = \sum_{i=1}^{n} \left\{ F[x_1, x_2, \ldots, x_i] \prod_{j=1}^{i-1} (x - x_j) \right\},$$

$$f_n(x) = f(x_1) + (x - x_1) f[x_2, x_1] + (x - x_1)(x - x_2) f[x_3, x_2, x_1] + \cdots + (x - x_1)(x - x_2) \cdots (x - x_{n-1}) f[x_n, x_{n-1}, \cdots, x_1].$$

3.1.4.5 Replace the values with $f_n(x(t_i))$ in missing position [$X((t_i))_{mis}$]

3.1.4.6 Repeat Step 2.2.1 for next row in the patient subspace matrix.

3.1.4.7 Repeat Step 2.1 for next the patient subspace matrix.

3.2 Else if (time treatment < = 2) or (missing all position in window)

Call DPimpute()

Setp5 : Loop 2.2 next patient one-has missing Until all patient no missing

ผลการทดลอง

ผลการทดลองด้วยตัววัดประสิทธิภาพการประมาณค่าด้วย NRSME ด้วยการแบ่งเป็น % ของการที่ปรากฏค่าสูญหายคือ

ตารางที่ 4.17 ค่าความคลาดเคลื่อน(NRMSE) ของชุดข้อมูลลำดับเวลา(NFDCslideW-DP) ชุดที่ 1

% missing / Dataset	10%	15%	20%	25%	30%	%Average
Obesity	0.116602	0.124625	0.11337	0.115263	0.109759	0.1159256
Thrombosis	0. 090381	0.098683	0.097055	0.095957	0.096556	0.097195

ตารางที่ 4.18 ค่าความคลาดเคลื่อน(NRMSE) ของชุดข้อมูลลำดับเวลา (NFDCslideW-DP) ชุดที่ 2

% missing / Dataset	10%	15%	20%	25%	30%	%Average
Obesity	0.119815	0.118748	0.117345	0.112039	0.112311	0.1160516
Thrombosis	0.099243	0.096917	0.097547	0.097765	0.09792	0.0978744

ตารางที่ 4.19 ค่าความคลาดเคลื่อน(NRMSE) ของชุดข้อมูลลำดับเวลา (NFDCslideW-DP) ชุดที่ 3

% missing / Dataset	10%	15%	20%	25%	30%	%Average
Obesity	0.125992	0.122046	0.113059	0.114096	0.11125	0.117288
Thrombosis	0.901956	0.098895	0.097888	0.096974	0.094852	0.09576

สรุปและอภิปรายผลการทดลอง

จากผลการทดลอง (NFDCslideW-DPimpute) ดังตารางที่ 4.17-4.19 ซึ่งจากชุดข้อมูลทั้งสามชุดที่ตำแหน่งการสูญหายไม่ซ้ำกัน พิจารณาที่ค่าเฉลี่ยเปอร์เซนต์ค่าสูญหายของแต่ละชุดข้อมูลให้ค่าความคลาดเคลื่อนเฉลี่ยของชุดข้อมูล obesity ชุดที่ 1-3 ตามลำดับ คือ 0.1159256, 0.1160516, 0.117288และ ชุดข้อมูล thrombosis ชุดที่ 1-3 ตามลำดับ คือ 0.097195, 0.0978744, 0.09576 มีค่าความคลาดเคลื่อนที่ใกล้เคียงกัน

จากการวิเคราะห์การประมาณค่าการตรวจด้วยวิธีการ (NFDCslideW-DPimpute)

นั่นคือทำการแบ่งชุดข้อมูล เป็นเซตย่อยเซตละสามค่าด้วยการเลื่อนแบ่งตำแหน่งข้อมูลในลักษณะวินโดว์ในแต่ละตัวชี้วัดของแต่ละคนด้วยดีกรีโพลิเมียลดีกรี 3 และประมาณค่าด้วย NFDCs-impute ซึ่งพบว่าให้ค่าการคำนวณในการประมาณค่าสูญหายค่อนข้างใกล้เคียงกับจากการวัดประสิทธิภาพค่าความคลาดเคลื่อนวิธีการ (NFDCslideW-DPimpute) วิธีการนี้ให้ค่าการประมาณค่าความคลาดเคลื่อนใกล้เคียงกับ NFDCs-DPimpute ดังนั้นเมื่อได้ค่าการประมาณที่ใกล้เคียงจากผู้ป่วยที่มาตรวจหลายครั้งแล้ว เมื่อทำการประมาณค่าสูญหายในผู้ป่วยที่มาตรวจไม่เกินสองครั้งและเซตข้อมูลย่อยปรากฏค่าสูญหายทั้งหมด จะใช้วิธีการ DPimpute มาประมาณค่าซึ่งจะเทียบกับค่าที่ได้ทำการประมาณค่าด้วย NFDCs สำหรับที่มาตรวจหลายครั้งและประมาณค่าด้วย NFDCs แล้วก็จะมีผลทำการประเมินค่าความคลาดเคลื่อนต่ำคือมีใกล้เคียงไปด้วย

7. เปรียบเทียบผลการทดลองของวิธีการประมาณค่าสูญหาย(Experimental Resul

ผลการเปรียบเทียบค่าความแม่นยำด้วย normal root mean standard error (NRMSE) ในรูปแบบ percentage จากชุดข้อมูลที่ใช้ในการทดลองจำนวน 3 ชุดข้อมูลที่ตำแหน่งการสูญหายของข้อมูลไม่ซ้ำกันในแต่ละวิธีการ

ตารางที่ 4.20 ค่าความคลาดเคลื่อน(NRMSE) ของ ชุดข้อมูลลำดับเวลา Obesity ชุดที่ 1

% missing / Impute Methods	10%	15%	20%	25%	30%	Average
DPimpute	0.10760	0.10824	0.11134	0.11425	0.11484	0.11125
SkNN-DP	0.35184	0.40588	0.45748	0.48182	0.50208	0.43982
CBE-DP	0.15362	0.17672	0.17966	0.17689	0.19429	0.17624
SLLS-DP	0.58630	0.59480	0.60225	0.60898	0.59978	0.59842
NFDCslideWindow-DPimpute	0.116602	0.124625	0.11337	0.115263	0.109759	0.1159256
NFDCs-DP	0.10937	0.10908	0.10572	0.10790	0.10405	0.10722

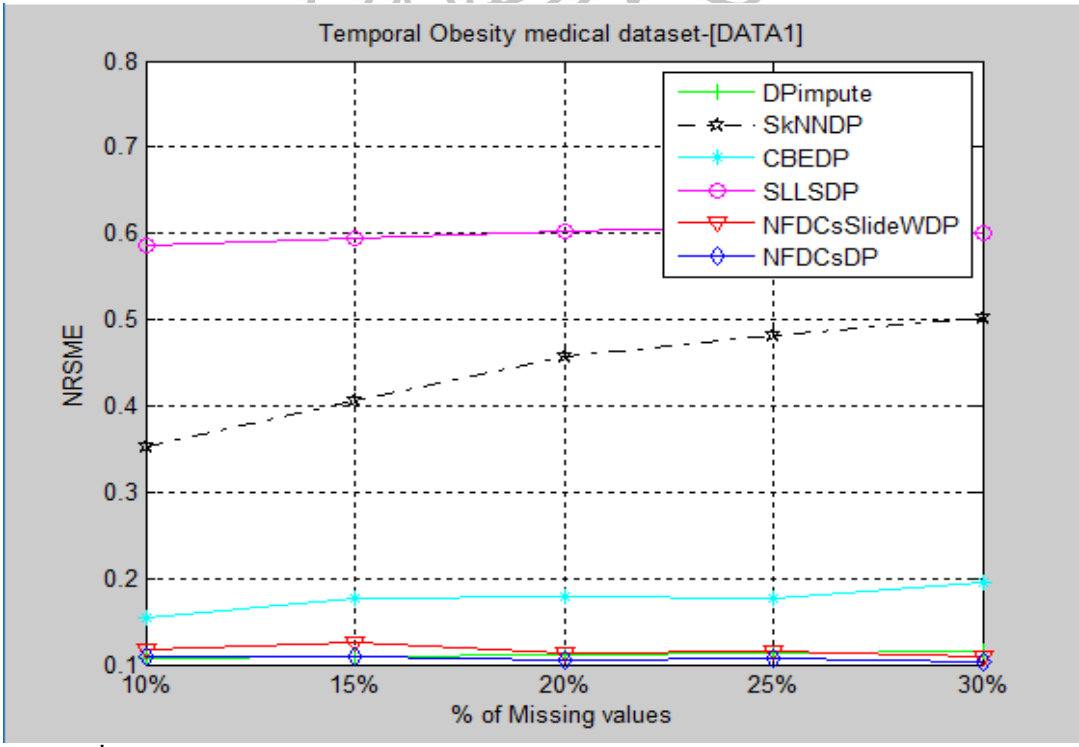
ตารางที่ 4.21 ค่าความคลาดเคลื่อน(NRMSE) ของ ชุดข้อมูลลำดับเวลา Obesity ชุดที่ 2

% missing / Impute Methods	10%	15%	20%	25%	30%	Average
DPimpute	0.10777	0.11093	0.11076	0.11760	0.11456	0.11232
SkNN-DP	0.33948	0.45428	0.45845	0.47302	0.49756	0.44456
CBE-DP	0.17599	0.18636	0.19399	0.20148	0.20400	0.19236
SLLS-DP	0.57576	0.59875	0.59103	0.58903	0.59522	0.58996
NFDCslideWindow-DPimpute	0.119815	0.118748	0.117345	0.112039	0.112311	0.11605
NFDCs-DP	0.11005	0.10533	0.10807	0.10764	0.10253	0.10673

ตารางที่ 4.22 ค่าความคลาดเคลื่อน(NRMSE) ของ ชุดข้อมูลลำดับเวลา Obesity ชุดที่ 3

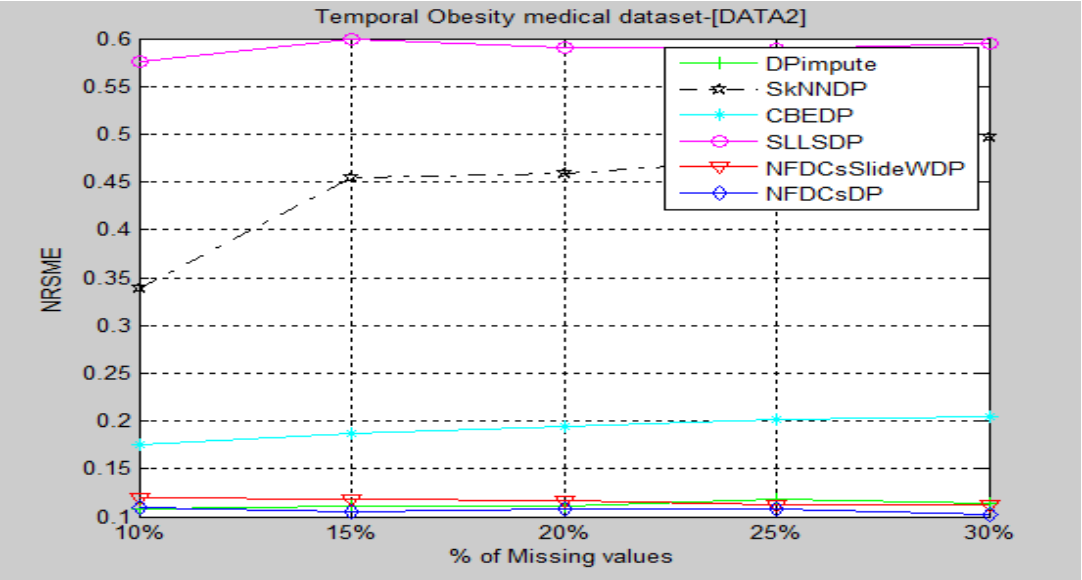
% missing / Impute Methods	10%	15%	20%	25%	30%	Average
DPimpute	0.10609	0.11130	0.11436	0.11520	0.11184	0.11176
SkNN-DP	0.35783	0.41892	0.46137	0.50439	0.51776	0.45205
CBE-DP	0.16638	0.18788	0.17660	0.18513	0.22564	0.18833
SLLS-DP	0.51584	0.58924	0.60225	0.60235	0.58253	0.57844
NFDCslideWindow-DPimpute	0.125992	0.122046	0.113059	0.114096	0.11125	0.117288
NFDCs-DP	0.11068	0.11439	0.10711	0.10661	0.10511	0.10878

Temporal obesity data (ชุด 1)

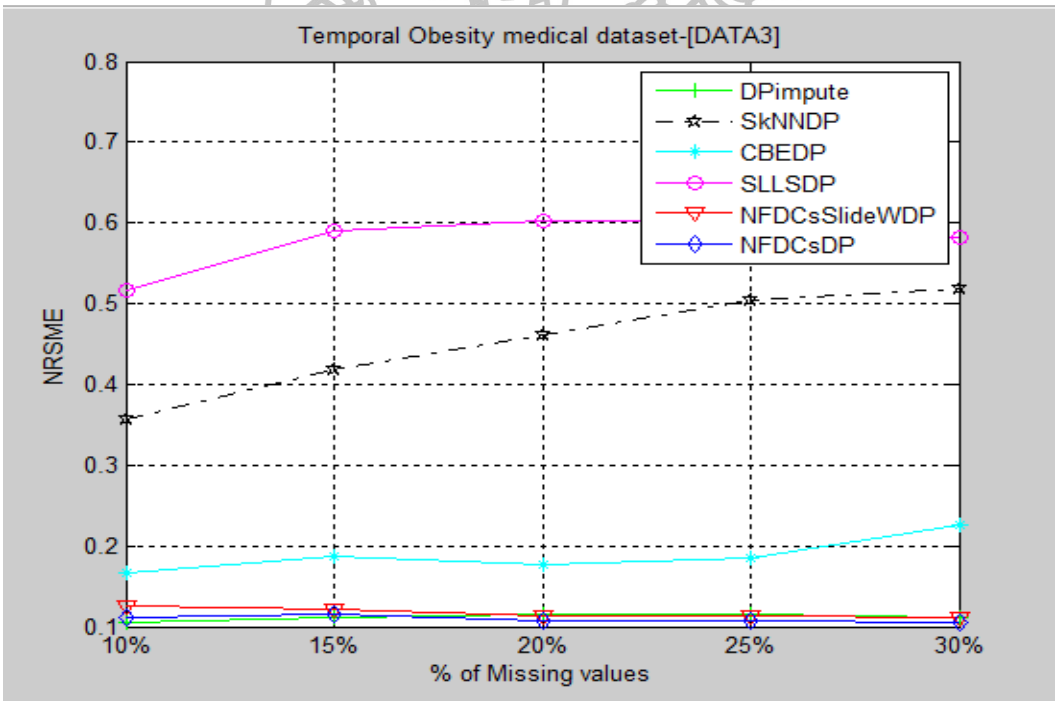


ภาพที่ 4.8 กราฟแสดงภาพรวมของ % of Missing values Temporal obesity data (ชุด 1)

Temporal obesity data (ชุด 2)



ภาพที่ 4.9 กราฟแสดงภาพรวมของ % of Missing values Temporal obesity data (ชุด2)
Temporal obesity data (ชุด 3)



ภาพที่ 4.10 กราฟแสดงภาพรวมของ % of Missing values Temporal obesity data (ชุด2)

นั่นคือจากชุดข้อมูล obesity วิธีการ NFDCs-DPimpute ให้ผลการประมาณค่าที่ดีกว่าวิธีการอื่นๆ ที่ใช้ในการศึกษาจากการวัดประสิทธิภาพค่าความคลาดเคลื่อน(NRMSE)

ข้อมูลผู้ป่วยโรคหลอดเลือดสมองชนิดอุดตัน (Thrombosis)

ตารางที่ 4.23 ค่าความคลาดเคลื่อน(NRMSE) ของ ชุดข้อมูลลำดับเวลา Thrombosis ชุด 1

% missing / Impute Methods	10%	15%	20%	25%	30%	Average
DPimpute	0.147970	0.152707	0.154287	0.161768	0.168859	0.15712
SkNN-DP	0.386775	.474585	0.552964	0.629435	0.714833	0.55172
CP-DP	0.094873	0.096731	0.109234	0.124589	0.125196	0.11394
SLLS-DP	2.488174	2.448375	2.467086	2.483993	2.462677	2.47006
NFDCslideWindow-DPimpute	0. 090381	0.098683	0.097055	0.095957	0.096556	0.097195
NFDCs-DP	0.088264	0.086666	0.086861	0.086346	0.088185	0.08726

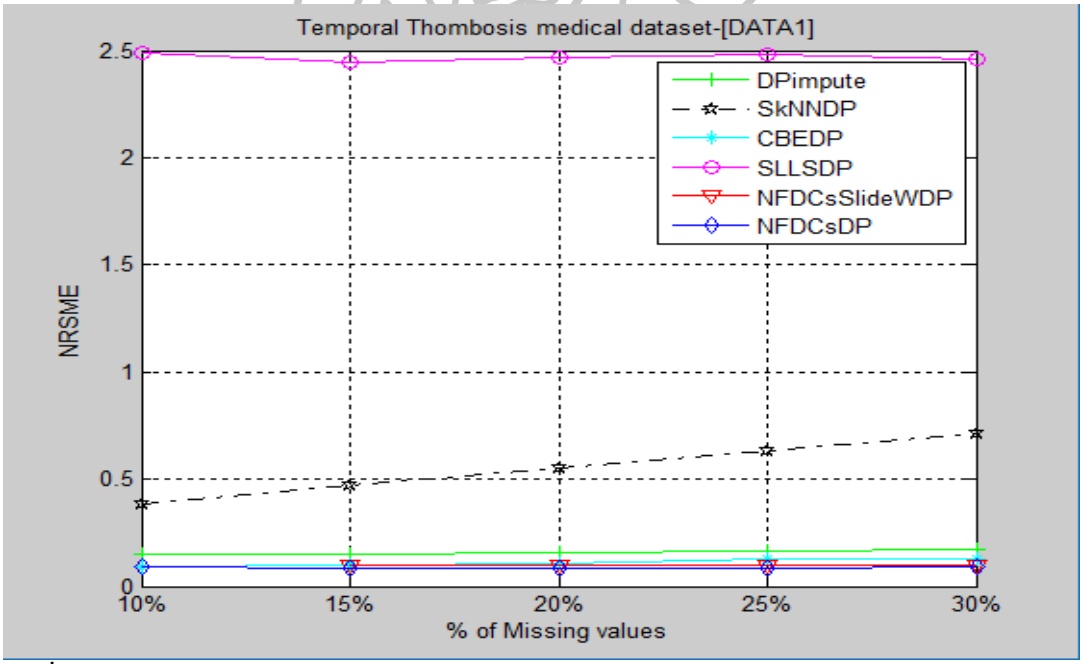
ตารางที่ 4.24 ค่าความคลาดเคลื่อน(NRMSE) ของ ชุดข้อมูลลำดับเวลา Thrombosis ชุด 2

% missing / Impute Methods	10%	15%	20%	25%	30%	Average
DPimpute	0.143519	0.146628	0.156782	0.160918	0.170778	0.15573
SkNN-DP	0.377631	0.467382	0.563123	0.638664	0.713756	0.55211
CP-DP	0.096174	0.104513	0.112150	0.117913	0.123097	0.11077
SLLS-DP	2.453547	2.489672	2.464771	2.468265	2.439718	2.46319
NFDCslideWindow-DPimpute	0.099243	0.096917	0.097547	0.097765	0.09792	0.09787
NFDCs-DP	0.087300	0.085698	0.084505	0.086512	0.089347	0.08667

ตารางที่ 4.25 ค่าความคลาดเคลื่อน(NRMSE) ของ ชุดข้อมูลลำดับเวลา Thrombosis ชุด 3

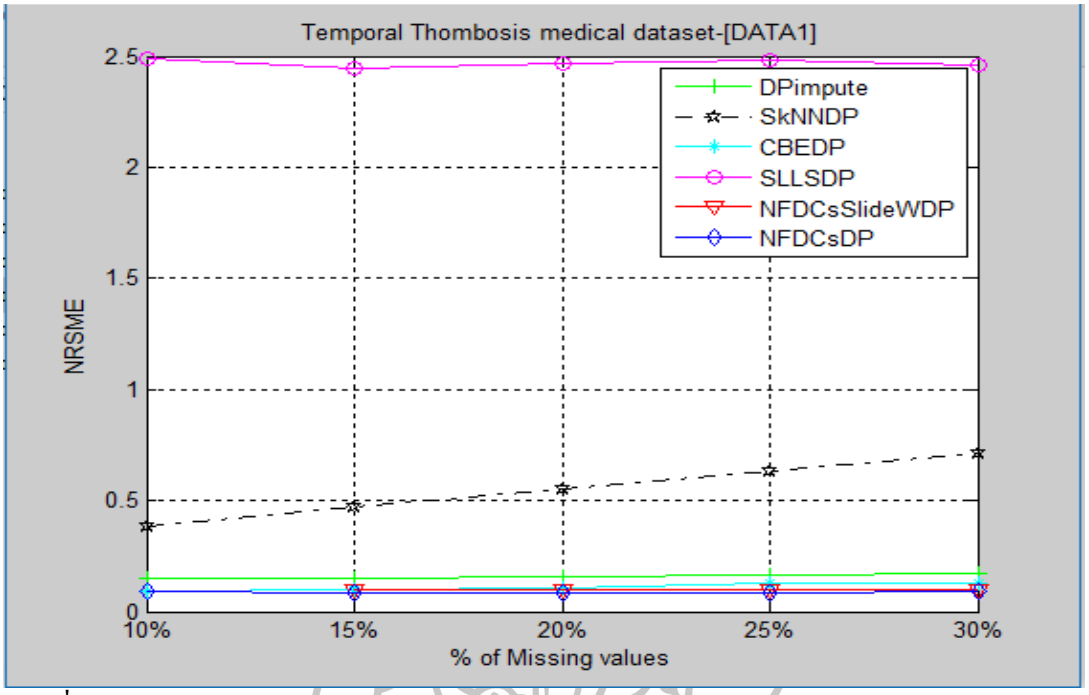
% missing / Impute Methods	10%	15%	20%	25%	30%	Average
DPimpute	0.144625	0.149842	0.152767	0.158743	0.166604	0.15452
SkNN-DP	0.375297	0.483222	0.553492	0.640958	0.709619	0.55252
CP-DP	0.092212	0.104513	0.109459	0.112908	0.119322	0.10768
SLLS-DP	2.439168	2.423025	2.453055	2.487622	2.427927	2.44616
NFDCslideWindow-DPimpute	0.0901956	0.098895	0.097888	0.096974	0.094852	0.09576
NFDCs-DP (*)	0.088077	0.087218	0.086881	0.086741	0.087557	0.08729

ชุดข้อมูลลำดับเวลา Thrombosis (Temporal Thombosis dataset) (ชุด 1)



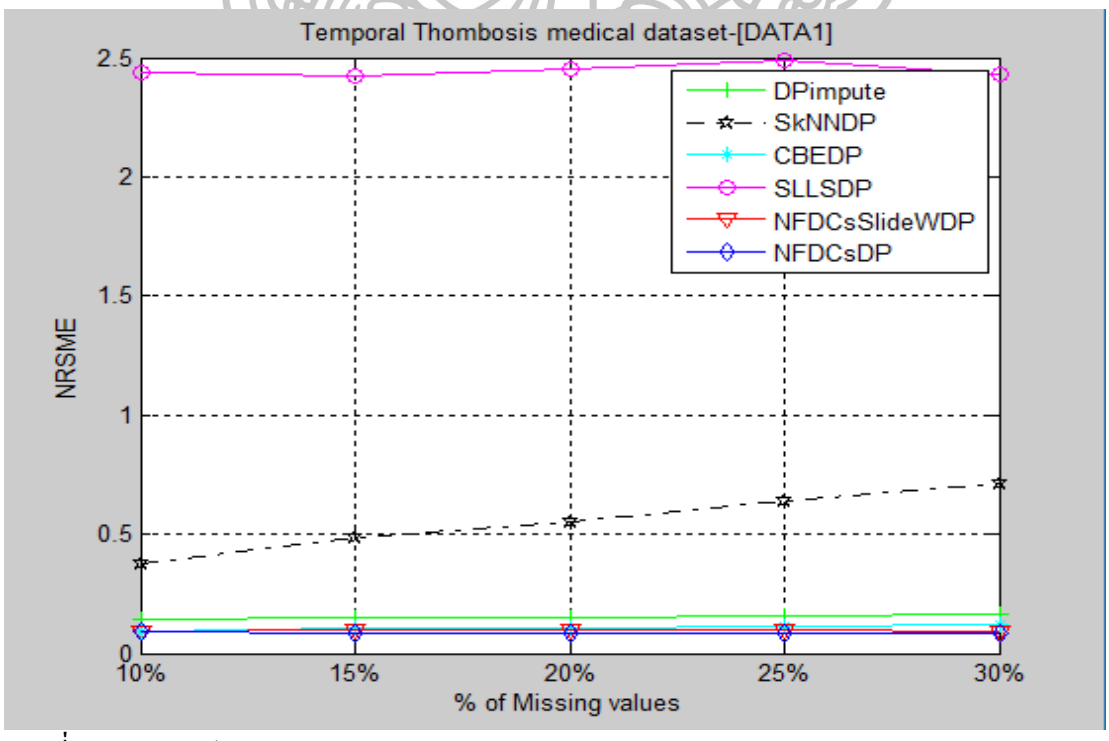
ภาพที่ 4.11 : กราฟแสดงภาพรวมของ % of Missing values(Temporal Thombosis dataset) (ชุด 1)

ชุดข้อมูลลำดับเวลา Thrombosis (Temporal Thombosis dataset) (ชุด 2)



ภาพที่ 4.12 : กราฟแสดงภาพรวมของ % of Missing values(Temporal Thombosis dataset) (ชุด2)

ชุดข้อมูลลำดับเวลา Thrombosis (Temporal Thombosis dataset) (ชุด 3)



ภาพที่ 4.12 : กราฟแสดงภาพรวมของ % of Missing values(Temporal Thombosis dataset) (ชุด 3)

สรุปและอภิปรายผลในตารางรวม

1. จากผลการทดลองของตารางที่ 4.20-4.2ถ ภาพที่ 4.8-4.12 แสดงผลการทดลองประสิทธิภาพความแม่นยำของขั้นตอนวิธีกับการประมาณค่าสูญหายในชุดข้อมูลในรูปแบบเปอร์เซ็นต์ ด้วย NRMSE นั่นคือประกอบด้วย 6 ขั้นตอนวิธีที่ใช้ในการประมาณค่าสูญหายจากชุดข้อมูลลำดับเวลาผู้ตรวจโรคอ้วน โดยการเปรียบเทียบแบ่งชุดข้อมูลเป็น 10%, 15%, 20%, 25%, และ 30% จากค่าข้อมูลที่ปรากฏการสูญหาย โดยผลการประมาณค่าที่ดีที่สุดคือ NFDCs-DP ซึ่ง NFDCs-DP ให้ค่าเฉลี่ย NRMSE ที่ต่ำที่สุด เมื่อเปรียบเทียบกับอีก 5 ขั้นตอนคือ DPimpute, SkNN-DP , CBE-DP , SLLS-DP และ NFDCsSlideW-DPimpute ตามลำดับ

2. ในตารางผลการทดลองตารางที่ 4.2-4.4 ในส่วนของตำแหน่งเปอร์เซ็นต์ค่าสูญหายจะมีเปอร์เซ็นต์ค่าสูญหายที่มากขึ้น แต่ผลเปอร์เซนต์ความคลาดเคลื่อนบางเปอร์เซนต์ค่าความคลาดเคลื่อนกลับน้อยลงจากการวิเคราะห์นั่นคือ

ค่าการประเมินค่าความคลาดเคลื่อนจะต้องประเมินในภาพรวมของชุดข้อมูลทั้งหมดที่ปรากฏจะไม่มองที่ค่าการประมาณค่าเดียว แต่ในการสุ่มสร้างค่าสูญหายเราไม่ทราบว่าตำแหน่งของการสูญหายจะไปปรากฏที่ตำแหน่งใดตามเปอร์เซนต์ค่าสูญหายที่เราระบุ ดังนั้นตำแหน่งของการปรากฏค่าสูญหายที่ปรากฏ จะมีข้อมูลจริงที่ปรากฏอยู่รอบหรือใกล้กับตำแหน่งที่สูญหาย เช่น 30% หากค่าที่ปรากฏอาจจะเป็นค่าที่มากหรือค่าที่น้อยก็ได้ ซึ่งปัจจัยคือค่าข้อมูลจริงที่นำมาประมาณค่าในตำแหน่งที่สูญหายกับการคำนวณในสมการวิธี ซึ่งเมื่อประมาณค่าอาจจะให้ค่าการประมาณที่สูงหรือต่ำผิดปกติได้ รวมทั้งหากตำแหน่งเปอร์เซนต์ค่าสูญหายสูงๆ แต่การสุ่มสร้างค่าขาดไปไปปรากฏในตำแหน่งของข้อมูลที่มี ค่าข้อมูลที่นำมาคำนวณแล้วได้ค่าการประมาณที่ใกล้เคียง รวมทั้ง DPimpute ที่ใช้ในการประมาณผู้ที่มาตรวจไม่เกินสองครั้ง หรือในบางตำแหน่งที่ปรากฏค่าสูญหายที่บางวิธีไม่สามารถประมาณค่านำตำแหน่งนั้นได้ ที่หากตำแหน่งของการประมาณหาค่าความเหมือนที่ได้ ได้ค่าการประมาณค่อนข้างใกล้เคียงในเปอร์เซนต์ค่าสูญหายที่สูงกว่าก็มีผลทำให้ค่าความคลาดเคลื่อนต่ำกว่าเปอร์เซ็นต์ค่าสูญหายที่น้อยกว่าก็เป็นได้

8. สรุปวิเคราะห์ผลการทดลองการประมาณค่าสูญหายในชุดข้อมูลลำดับเวลา

8.1 จากการวัดประสิทธิภาพค่าความคลาดเคลื่อน

จากชุดข้อมูลที่ใช้ในการทดลอง ลักษณะชุดข้อมูลผู้ป่วยโรคอ้วน(obesity data) จะเป็นลักษณะค่าข้อมูลของการมาตรวจรักษาทั้งในลักษณะระยะสั้นและระยะยาว ส่วนชุดข้อมูลผู้ป่วยโรคหลอดเลือดสมองชนิดอุดตัน(Thrombosis) ส่วนมากจะเป็นค่าข้อมูลของผู้ป่วยที่มาทำการรักษาระยะยาว

จากการวัดประสิทธิภาพค่าความคลาดเคลื่อนการประมาณค่าที่สูญหายด้วย NRSME จะเห็นได้ว่าทั้ง 5 ขั้นตอนวิธี และจากทั้งสองชุดข้อมูลคือ ชุดข้อมูลผู้ป่วยโรคอ้วน (obesity data) และชุดข้อมูลผู้ป่วยโรคหลอดเลือดสมองชนิดอุดตัน(Thrombosis) ชุดข้อมูลละ 3 ชุดย่อยที่ตำแหน่งการสูญหายของข้อมูลไม่ซ้ำกันแบ่งตามเปอร์เซนต์การสูญหายในแต่ละชุดข้อมูลย่อย จะเห็นว่าขั้นตอน NFDCs-DP จะเป็นขั้นตอนที่ให้ค่าการประมาณสูญหายที่ดีกว่าขั้นตอนเทคนิควิธีอื่นที่ใช้ในการศึกษาจากข้อมูลชุดดังกล่าว

8.2 พัฒนาโมเดลตามวัตถุประสงค์ข้อ 2 : การทรานฟอร์มค่าข้อมูล(feature transform)

เมื่อค่าประมาณค่าและทดแทนค่าข้อมูลที่สูญหายได้ชุดข้อมูลสมบูรณ์แล้ว นั่นคืออย่างแรก จะนำชุดข้อมูลลำดับเวลาที่เป็นชุดข้อมูลที่ผ่านการประมาณค่าสูญหายและได้เป็นชุดข้อมูลสมบูรณ์เพื่อทรานฟอร์มรายการค่าข้อมูลบนหลักการทรานฟอร์ม ซึ่งจะทำการลดมิติเพื่อใช้ประโยชน์จากชุดข้อมูล ในงานวิจัยนี้จะทำการทรานฟอร์มข้อมูลเพื่อให้ได้ค่าข้อมูลชุดใหม่ในลักษณะค่าข้อมูลเดี่ยว(singular values) ซึ่งจะทำให้ได้ชุดข้อมูลขนาดเล็กลง

โครงสร้างผลลัพธ์ของชุดข้อมูลใหม่ที่ผ่านกระบวนการทรานฟอร์มชุดข้อมูลลำดับเวลาจะได้โครงสร้างดังตารางที่ 4.26

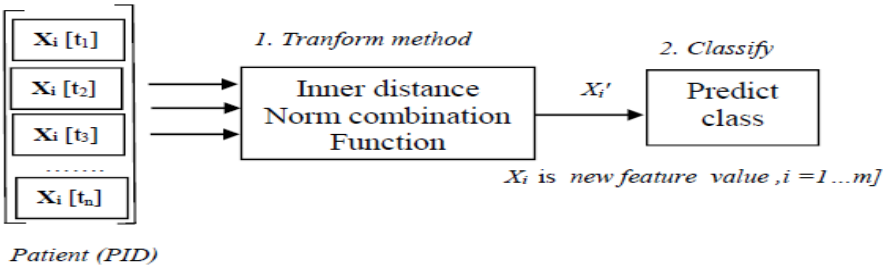
ตารางที่ 4.26 โครงสร้างผลลัพธ์ของการทรานฟอร์ม

PID	IDCT			Mean			F_n			C
	X_1'	X_2'	…	X_1'	..	X_n'	X_1'	.	X_n'	
01	..	..							..	..
02	.	..							..	..
…	..	..							..	..
N	..	..							..	..

8.3 วิธีการการทรานฟอร์มค่าข้อมูลที่พัฒนา

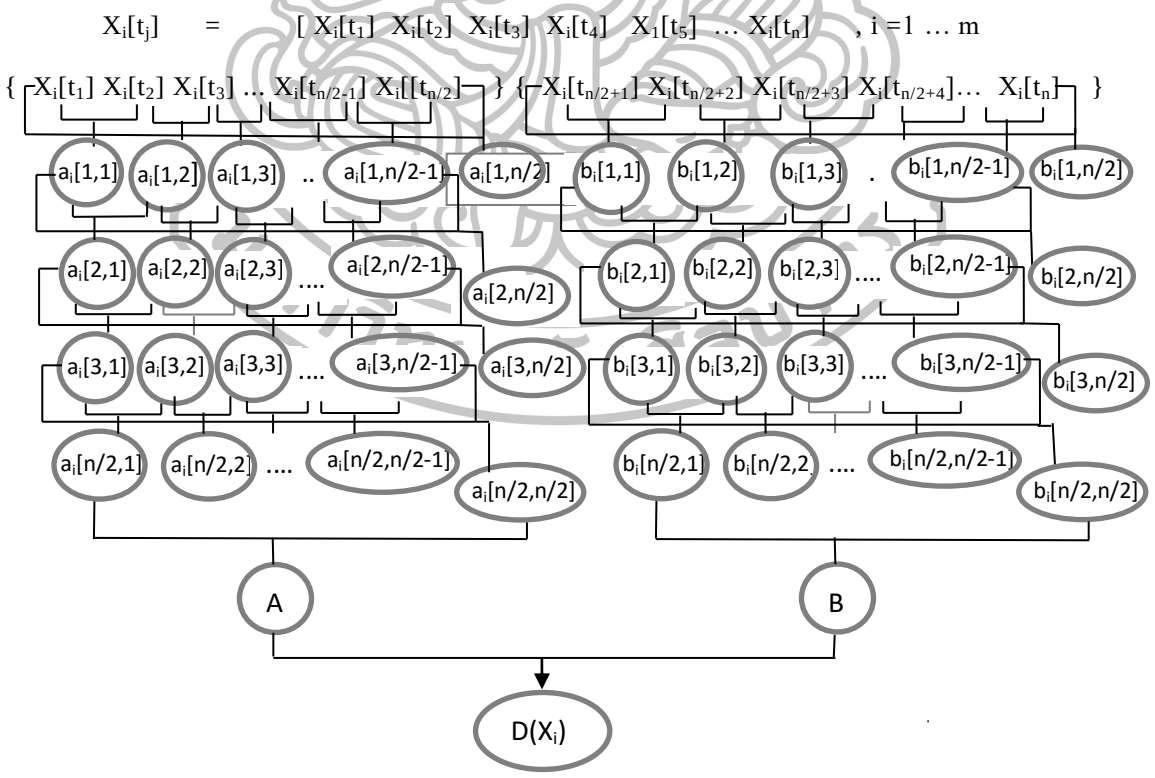
1) Inner distance combination transform (IDCT)

คือกระบวนการ Inner distance combination transform (IDCT) จะทำให้ได้ค่าข้อมูลชุดใหม่ในลักษณะ singular values และลำดับต่อนำมานำข้อมูลชุดใหม่นำเข้ากระบวนการจำแนกประเภทเพื่อพิจารณาความแม่นยำของการจำแนกจากการวัดประสิทธิภาพ ดังแผนภาพที่ 4.1.4



ภาพที่ 4.14 : Block diagram วิธีการทรานฟอร์มชุดข้อมูลลำดับเวลาเพื่อการจำแนก

ก่อนการพัฒนาขั้นตอนและทดลองผู้วิจัยคาดว่าหลักการการประมวลผลด้วยการใช้ฟังก์ชัน inner distance นี้ เมื่อคำนวณมาแล้วน่าจะให้ค่าที่ยอมรับได้จึงนำมาทำการรวม (combine) กับค่าข้อมูลจริง(real values) ด้วยหลักการรวม(combination) ภายใต้การคำนวณด้วย inner distance function ของแต่ละชุดลำดับค่าข้อมูลในตัวชี้วัดของผู้ป่วยแต่ละคน จากค่าการตรวจที่อยู่ใกล้กันซึ่งมันน่าจะเป็นที่ใกล้เคียงกันตามระยะเวลาของการตรวจรักษาและให้ครบทุกค่าจากชุดข้อมูลในแต่ละตัวชี้วัดของแต่ละคน



ภาพที่ 4.15 : การ combination กับ Inner distance function ของชุดข้อมูล.

จากแผนภาพที่ 4.15 แนวคิดนี้จะทำการลดขนาดของชุดข้อมูลด้วยการทรานฟอร์มข้อมูลตามตัวชี้วัดของผู้ป่วยรายบุคคล(*PID-X*$_i$) โดยกำหนดให้ $X_i[t_j]$ แทนลำดับค่าข้อมูลตามตัวแปรเวลาตามลำดับเวลาการตรวจรักษาให้ได้เป็นค่าเฉพาะค่าเดียว แล้วข้อมูลของผู้ป่วยทั้งหมดก็จะเป็นชุดข้อมูลใหม่ (*Xi'*) , *Xi'* จะถูกใช้แทน $X_i[t_j]$, *j= 1..n,* เพื่อเป็นชุดข้อมูลนำเข้าในการจำแนกกลุ่ม ดังแผนภาพแสดงรูปแบบภายใต้การคำนวณด้วย inner distance function และ หลักการรวม (combination) ของแต่ละชุดลำดับค่าข้อมูลในตัวชี้วัดของผู้ป่วยแต่ละคน

ตัวอย่าง

1. Set

$$X_i[t_j] = [\, X_i[t_1]\ X_i[t_2]\ X_i[t_3]\ X_i[t_4]\ X_1[t_5]\ \dots\ X_i[t_n] \,] \quad , i = 1 \dots m$$

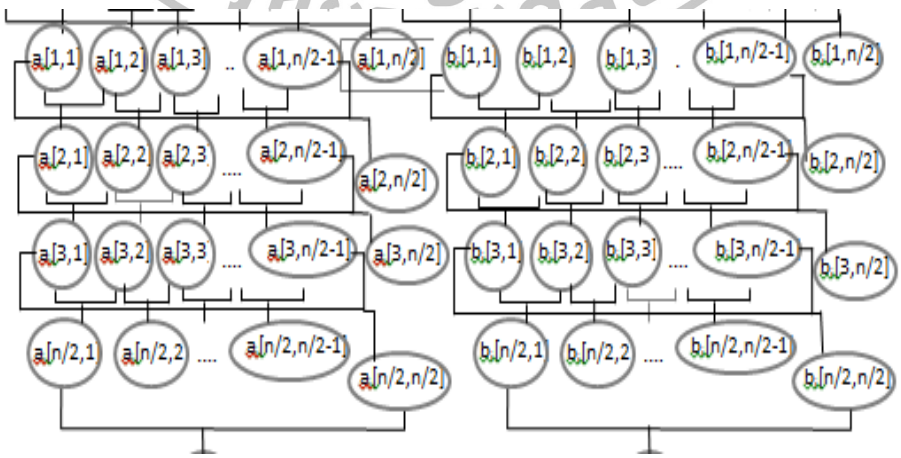
2. Separate to two set (n/2) // ถ้า set เป็นเลขคี่เป็นทศนิยมปัดไปเป็น 1 ในเซตแรก

$$\{\ \lceil X_i[t_1]\ X_i[t_2]\ X_i[t_3]\ \dots\ X_i[t_{n/2-1}]\ X_i[\lceil t_{n/2} \rceil]\ \rceil\ \}\ \{\ \lceil X_i[t_{n/2+1}]\ X_i[t_{n/2+2}]\ X_i[t_{n/2+3}]\ X_i[t_{n/2+4}] \dots\ X_i[t_n]\ \rceil\ \}$$

3. คำนวณจำนวน level (a[n/2,1) ที่จะจับคู่ combination

4. จับคู่คำนวณจากค่าการตรวจที่อยู่ใกล้กัน (ทีละ Set) จนครบตามจำนวน levelในข้อ 3 ด้วยสมการ Inner distance norm

$$\| x \| = \sqrt{x_1^2 + x_2^2 + \dots\dots + x_n^2}$$



$$\| a11\| = \sqrt{(x_i t_1)^2 + (x_i t_2)^2} \qquad \| b11\| = \sqrt{(x_i t_{n//2-1})^2 + (x_i t_{n//2-2})^2}$$

$$\| a11\| = \sqrt{(x_i t_1)^2 + (x_i t_2)^2} \qquad \| b11\| = \sqrt{(x_i t_{n//2-1})^2 + (x_i t_{n//2-2})^2}$$

……

……

5. คำนวณหาค่า Max ใน level สุดท้ายในแต่ละ SET (A) , (B)

$$Max(A) = max[(a11),(a12),\ldots)$$

$$Max(B) = max[(b11),(b12),\ldots)$$

6. คำนวณหาค่า Max

$$D(X_i) = Max \|A\| ,\|B\|$$

7. วนรอบไปทำตัวแปรตัวต่อไป

8. วนรอบไปทำคนต่อไป

จากแนวคิดดังตัวอย่างและแผนภาพที่ 4.15 จึงพัฒนาขั้นตอนและทำการทดลองกับชุดข้อมูลดังลำดับขั้นตอนวิธีต่อไปนี้

Procedure: Inner distance combination transformation method
Input: a complete temporal dataset (D)
Output: singular new feature values

Step1. Separate data in the patient subspace matrix.
Step2. Repeat
Step3. Select each patient in the subspace matrix.
3.1 Transpose the patient subspace matrix.
3.2 Repeat
3.2.1 X_i ; $i = i+1$, $i=1,2,3,\ldots,m$
3.2.2 Select a row value in the subspace matrix.
$[X_i[t_1]\ X_i[t_2]\ X_i[t_3]\ X_i[t_4] \ldots X_i[t_n]]$, $i=1 \ldots m$
3.2.3 Separate a set in the series into two sets.
$[X_i[t_1]\ X_i[t_2]\ X_i[t_3] \ldots X_i[t_{n/2}]]$ and $[X_i[t_{n/2+1}]\ X_i[t_{n/2+2}]\ X[tn_{/2+3}] \ldots X_i[t_n]]$
3.2.4 Compute A and B from the first set in 3.4
In each level j ; $j = 1,2,3, \ldots,n/2$.

$$j = 1;\ a_i[1,k] \quad \begin{cases} \sqrt{x_i[t_k]^2 + x_i[t_{k+1}]^2} & , \ k = 1,2,3,\ldots,n/2-1 \\ \sqrt{x_i[t_k]^2 + x_i[t_1]^2} & , \ k = n/2 \end{cases} \qquad \ldots\ldots\ldots(1)$$

$$b_i[1,k] = \begin{cases} \sqrt{x_i[t_k]^2+x_i[t_{k+1}]^2} & , \ k = n/2+1, n/2+2,\ldots,n-1 \\ \sqrt{x_i[t_k]^2+x_i[t_{n/2+1}]^2} & , \ k = n \end{cases} \quad \ldots\ldots\ldots(2)$$

when $a_i[j,k]$ and $b_i[j,k]$ where $j = 2,3,4,\ldots,n/2$

$$a_i[j,k] = \begin{cases} \sqrt{a_i[j,k]^2+ a_i[j,k+1]^2} & , \ k = 1,2,3,\ldots,n/2 \\ \sqrt{a_i[j,k]^2+ a_i[j,1]^2} & , \ k = n/2 \end{cases} \quad \ldots\ldots\ldots(3)$$

$$b_i[j,k] = \begin{cases} \sqrt{b_i[j,k]^2+ b_i[j,k+1]^2} & , \ k = 1,2,3,\ldots,n/2 \\ \sqrt{b_i[j,k]^2+ b_i[j,1]^2} & , \ k = n/2 \end{cases} \quad \ldots\ldots\ldots(4)$$

3.2.5 Compute A and B

$$\|A\| = \text{Max}(a_i[j,k]) \quad , j = n/2 \ , \ k = 1,2,3,\ldots,n/2 \quad \ldots\ldots\ldots (5)$$

$$\|B\| = \text{Max}(b_i[j,k]) \quad , j = n/2 \ , \ k = 1,2,3,\ldots,n/2 \quad \ldots\ldots\ldots (6)$$

3.2.6 Compute D

$$D(X_i) = \text{Max}(\|A\|, \|B\|) \quad \ldots\ldots\ldots (7)$$

3.2.7 Until i <= m.

Step4. $X_i = [(x_1')\ (x_2')\ldots..\ (x_n')]$ //new feature

Step5. Repeat step 2 for patient subspace *n*+1.

Step6. $X_i' = [(x_1')\ (x_2')\ldots..\ (x_n')]$ //new feature set

อธิบายแนวคิดขั้นตอนวิธี :

สำหรับขั้นตอนวิธี Inner distance combination transformation method โดยกำหนดให้ PID-0n- $X_i[t_j]$ คือ ชุดของคนไข้แต่คนที่ระบุถึงจำนวนของการตรวจรักษาเพื่อทำการนำสู่การจำแนกประเภท, กำหนดให้ X_i คือ ตัวชี้วัดแต่ละตัว และ t_j คือเวลาของการตรวจรักษาของคนไข้แต่ละคน โดยในชุดของข้อมูลผู้ป่วยแต่ละคนจะประกอบด้วยเวลาการตรวจรักษาของแต่ละตัวชี้วัด (T_j), $T_j = \{t_1, t_2, t_3, \ldots, t_n\}$, เมื่อ j คือ จำนวนครั้งการตรวจของแต่ละคนและ *n* คือจำนวนครั้งทั้งหมดที่ตรวจรักษาของแต่ละคน ซึ่งจำนวนครั้งของการมาตรวจรักษาของแต่ละคนอาจจะไม่เท่ากัน ดังนั้นจะทำการแปลงข้อมูลเพื่อให้ได้ชุดข้อมูลชุดตัวแทน ค่าตัวชี้วัด X_i เป็นค่าต่อเนื่องที่ขึ้นกับเวลาลำดับของการตรวจวัด ดังนั้นการทรานฟอร์มจะทำการทรานฟอร์มค่าข้อมูลไปยังข้อมูลใหม่ด้วยกระบวนการเหมืองข้อมูล ซึ่ง $X_i \longrightarrow X_i'$ จะทำการทรานฟอร์มแต่ละลำดับของ X_i เพื่อแทนด้วยข้อมูลเดี่ยว X_i'. กำหนดให้ a_it_i และ b_it_i คือชุดของค่าข้อมูลแต่ละตัวแปร, D คือผลลัพธ์ของการหาค่าสูงสุดของการ(combination) ระหว่าง A และ B, นำเสนอโดย X_i จากค่าข้อมูลคนไข้แต่ละคนผลของการคำนวณของ IDCT คือ ชุดค่าข้อมูลชุดใหม่นำเสนอโดย X_i' จากชุด

ข้อมูลผู้ป่วยทั้งหมด เพื่อให้ง่ายและสามารถนำเข้าสู่ การทำนายกลุ่มของการจำแนกประเภท

สำหรับการทรานฟอร์มชุดข้อมูลรายบุคคลก็จะได้ข้อมูลชุดใหม่ที่เป็นตัวแทนชุดข้อมูลในลักษณะ singular values ประสิทธิภาพของขั้นตอนใหม่วิธีการนี้ได้นำมาเปรียบเทียบกับหลักการทางสถิติที่นำมาทำการทรานฟอร์มข้อมูลคือ mean tranform, median tranform, stdev tranform, variance tranform ด้วยการประเมินประสิทธิภาพความแม่นยำจากการจำแนกกลุ่มด้วยค่าความแม่นยำ Accuracy ดังรายละเอียดและผลการประเมินประสิทธิภาพในหัวข้อ 4.2

4.2 การจำแนกกลุ่มจากชุดข้อมูลผ่านการประมาณค่าและทรานฟอร์ม

จากจุดประสงค์ของการทรานฟอร์ม เมื่อทำการทรานฟอร์มได้ชุดข้อมูลที่เป็นตัวแทนแล้วควรจะยังคงมีประสิทธิภาพในการจำแนกกลุ่ม ดังนั้นในส่วนนี้จะทำการศึกษาประสิทธิภาพความแม่นยำของการจำแนกกลุ่มจากชุดข้อมูลที่ประมาณค่าสูญหายและทรานฟอร์ม

ผู้วิจัยได้ทำการนำชุดข้อมูลที่ทำการทรานฟอร์มจากชุดข้อมูลที่ประมาณค่าสูญหายด้วยวิธีการ NFDCs-DPimpute นำเข้าโมเดลการจำแนกประเภท ที่นิยมใช้ในการศึกษาเพื่อประเมินประสิทธิภาพของการจำแนกข้อมูล ในการเลือกโมเดลจำแนกประเภทปัจจัยหนึ่งของความแม่นยำของการจำแนกประเภท คือ โมเดลในการจำแนกที่เหมาะสมซึ่งจะทำให้การเรียนรู้ชุดข้อมูลมีความแม่นยำ จึงได้พิจารณาเทคนิคการวิเคราะห์ที่ให้ค่าความแม่นยำที่เหมาะสมกับลักษณะชุดข้อมูล โดย

1. พิจารณาจากงานวิจัยที่เกี่ยวข้องกับงาน Temporal classification, time series classification จากการอ้างอิงที่เป็นโมเดลที่นิยมใช้ในการทดสอบการจำแนกประเภท เช่น

งานวิจัยเรื่อง Medical data mining Issue and Experimental ใช้โมเดล NN , NB [36]

งานวิจัยเรื่อง Temporal data classifier using linear classifier ใช้โมเดล SVM [35]

งานวิจัยนี้ Classification of thombosis collagence disease base on C4.5 algorithm

ใช้โมเดล SVM , C4.5, DT [37]

2. ลักษณะชุดข้อมูลเป็น binary class : เนื่องจากลักษณะค่าข้อมูลในชุดข้อมูลที่ใช้ในการทดลอง เป็นชุดข้อมูลลักษณะ binary class โมเดลที่ใช้รองรับการจำแนกประเภทแบบไบนารีคลาสได้

3. พิจารณาจากคุณลักษณะของชุดข้อมูล

เนื่องจากลักษณะค่าข้อมูลในชุดข้อมูลที่ใช้ในการทดลองเป็นชุดข้อมูลประเภทตัวเลข ค่าข้อมูลของชุดข้อมูลในการทรานฟอร์มเป็นค่าลักษณะตัวเลข โมเดลดังกล่าควรสามารถรับข้อมูลนำเข้าในลักษณะค่าตัวเลขได้ จึงได้นำโมเดลในข้อ 1 นี้มาทำการทดลองกับชุดข้อมูลดังกล่าวเพื่อพิจารณาเทคนิคการวิเคราะห์ให้เหมาะสมกับลักษณะชุดข้อมูล และ เนื่องจากลักษณะคลาสคำตอบเป็น binary class มีจำนวน class 1 และ 0 ที่ระบุการเป็นโรคอ้วนหรือไม่เป็น ซึ่งจำนวนของคลาสที่ระบุการเป็นโรคของแต่ละคนปรากฏไม่เท่ากัน เพื่อเพิ่มประสิทธิภาพในการจำแนกวิธีหนึ่งคือทำความไม่สมดุลของคลาสในชุดข้อมูลด้วยวิธีการ Resample คือ การเพิ่มจำนวนคลาสให้เท่ากันในการจำแนกกลุ่มเพื่อค่าความแม่นยำกำหนดรูปแบบการจำแนกเพื่อเรียนรู้และทดสอบชุดข้อมูลในลักษณะ fold cross validation ซึ่งงานวิจัยส่วนใหญ่จะใช้ 5 , 10 fold เนื่องจากได้ค่าความถูกต้องเป็นที่น่าพอใจ สำหรับงานวิจัยนี้ได้เลือกใช้ 10 fold cross validation ได้ทำการทดลองความแม่นยำของการจำแนกจาก 10 fold cross validation โดยกำหนดจำนวนกลุ่ม หรือ k = 10 โดยจะแบ่งข้อมูลออกเป็น 10 ส่วนและแต่ละส่วนมีค่าข้อมูลเท่ากัน ซึ่งผลการทดลองจะอยู่ในรูปแบบของตารางคอนฟิวชั่นเมตริกซ์(confusion matrix) ด้วยวิธีการวิเคราะห์ความถูกต้อง ในงานวิจัยนี้จะวัดประสิทธิการของผลการทดลองโดยพิจารณา จากค่า precision, Recall, F-measure, Accuracy

ประเมินประสิทธิภาพความแม่นยำของการจำแนกกลุ่มลำดับเวลาจากค่าข้อมูลที่สมบูรณ์ การตรวจสอบประสิทธิภาพของการเรียนรู้ด้วยชุดเรียนรู้และจากค่าความแม่นยำของการจำแนกกลุ่ม(Accuracy) [8]

$$\text{F-measure} = \frac{2 \times \text{Recall} \times \text{precision}}{\text{Recall} + \text{Precision}}$$

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

ผลการทดลอง

ผลการเปรียบเทียบค่าความแม่นยำของการจำแนกกลุ่มชุดข้อมูลที่ได้ ทรานฟอร์มด้วยวิธีการ IDTC obesity ด้วย 10-fold cross validation

ตารางที่ 4.27 ผลการเปรียบเทียบค่าความแม่นยำของการจำแนกกลุ่มชุดข้อมูล obesity ที่ได้ทรานฟอร์มด้วยวิธีการ IDTC

วิธีการ	Inner distance combination -IDTC(Proposed)			
	Decision Tree	**Naïve Bays**	**SVM**	**Neural Network**
ค่าความแม่นยำ(Accuracy)	63.2385	80.3063	85.3392	84.4639

จากผลการทดลองข้างต้น จึงนำหลักการของ SVM ซึ่งเป็นโมเดลที่นิยมใช้ในงานวิจัย เพราะให้ประสิทธิภาพการจำแนกสูงสุดในชุดข้อมูลที่ศึกษา มาทดลองกับชุดข้อมูลที่เป็นตัวแทนจากการที่ประมาณค่าสูญหายและทรานฟอร์มจากวิธีการ IDTC และได้ทำการเปรียบเทียบกับวิธีการอื่นด้วยหลักสถิติ คือ mean, median, stdev, variance ในที่ใช้ในการหาตัวแทนค่าของชุดข้อมูลใน ชุดหรือกลุ่มข้อมูลในสิ่งที่สนใจในการศึกษา ดังผลการทดลองในตาราง

ตารางที่ 4.28 ผลการเปรียบเทียบความแม่นยำของการจำแนกกลุ่มจากชุดข้อมูลที่เป็นตัวแทน

ชุดข้อมูลตัวแทนจากการทรานฟอร์ม	SVM			
	Accuracy	**Precision**	**Recall**	**F-measure**
Inner distance combination IDTC(*)data	88.8889	0.893	0.889	0.885
Mean transform data	84.3434	0.854	0.843	0.834
Median transform data	72.2222	0.711	0.722	0.705
Stdev transform data	85.8586	0.862	0.859	0.853
Variance transform data	85.8586	0.862	0.859	0.853

จากการพัฒนาวิธีการประมาณค่าสูญหายและทรานฟอร์มข้อมูล เมื่อนำชุดข้อมูลเริ่มต้นชุดข้อมูลเดียวกันที่ประมาณค่าสูญหายและทรานฟอร์มข้อมูล จากชุดข้อมูลลำดับเวลาเฉพาะคนที่มาตรวจหลายครั้งเพื่อนำเข้าไปทรานฟอร์มในแต่ละวิธีด้วยวิธีการ IDTC transform, Mean transform, Median transtorm, Stdev transform และ Variance transform ก็จะได้ชุดข้อมูลที่เป็นตัวแทน เมื่อนำเข้าโมเดลการจำแนกกลุ่ม SVM (Support vector machine) และ วัดประสิทธิภาพการจำแนกกลุ่มจากค่าความแม่นยำด้วยค่า Accuracy พบว่าชุดข้อมูลที่ประมาณค่าสูญหายและทรานฟอร์มข้อมูลด้วยวิธีการ IDTC transform มีความแม่นยำในการจำแนกกลุ่มได้ดีกว่าเมื่อเปรียบเทียบกับวิธี Mean transform, Median transtorm, Stdev transform และ Variance transform ดังตารางที่ 4.28

อภิปราย

จากกระบวนการในการทรานฟอร์มเพื่อหาตัวแทนชุดข้อมูล การที่จะบอกว่าชุดข้อมูลที่เราได้ดีหรือไม่ ขึ้นอยู่กับการนำไปใช้ในสิ่งที่เราสนใจ ในงานวิจัยนี้สนใจพิจารณาความแม่นยำจากชุดข้อมูลชุดนี้ว่ายังคงค่าความแม่นยำในการจำแนกทำนายประเภทกลุ่มด้วยกระบวนการจำแนกกลุ่มและวัดประสิทธิภาพการจำแนกกลุ่ม

ดังนั้นจากผลการทดลอง การที่ค่าตัวแทนจากชุดข้อมูลของวิธีการ IDTC ให้ผลการจำแนกดีกว่าชุดข้อมูลอื่นๆ พิจารณาจากคุณสมบัติต่อไปนี้

1. ในสถานการณ์จริงทั่วไปข้อมูลที่ที่สนใจมักมีค่ามากหรือน้อยปะปนอยู่ ค่าเฉลี่ยจะนำทุกค่ามาพิจารณารวมทำให้ข้อมูลที่มีค่ามากหรือน้อยเฉลี่ยรวมกันไป ส่วนมัธยฐานเป็นวิธีที่พิจารณาเพียงข้อมูลที่มีค่าอยู่ ณ ตำแหน่งตรงกลางซึ่งไม่พิจารณาข้อมูลอื่นๆ การใช้วิธี Median transform จะทำให้ได้ตัวแทนชุดข้อมูลที่พิจาณาบางค่าเท่านั้น ส่วนความแปรปรวนและส่วนเบี่ยงเบนมาตรฐานเป็นการวัดระยะห่างของข้อมูลจากค่าเฉลี่ย ในขณะที่วิธีการ IDTC พิจารณาจากการจับคู่ค่าที่ติดกันด้วยการหา Inner distance ทุกคู่ที่อยู่ติดกัน ไม่ว่าข้อมูลจะมีค่ามากหรือน้อยจะพิจารณาค่าข้อมูลทุกตัว จึงสามารถใช้กับข้อมูลลักษณะใดก็ได้ ดังนั้นจึงบ่งบอกได้ว่าเป็นชุดข้อมูลจากวิธีการทรานฟอร์มที่มีประสิทธิภาพจากการวัดประสิทธิภาพในการจำแนกกลุ่มได้ดีกว่าวิธีอื่นๆ

2. ในการหาตัวแทนชุดข้อมูลจากข้อมูลเชิงเวลาด้วยกระบวนการในการทรานฟอร์มซึ่งจะได้ชุดข้อมูลที่เป็นตัวแทนการจะบอกว่าชุดข้อมูลที่เราได้จากวิธีการทรานฟอร์มแต่ละชุดดีหรือไม่ ปัจจัยหนึ่งขึ้นอยู่กับการนำไปใช้ในสิ่งที่เราสนใจ ในงานวิจัยนี้สนใจพิจารณาความแม่นยำจากชุดข้อมูลชุดนี้ว่ายังคงค่าความแม่นยำในการจำแนกทำนายประเภทกลุ่มด้วยกระบวนการจำแนกกลุ่มและวัดประสิทธิภาพ ดังนั้นเมื่อชุดข้อมูลของเราผ่านการจำแนกประเภทด้วยวิธีการเดียวกัน เราก็สามารถเปรียบเทียบประสิทธิภาพการจำแนกชุดข้อมูลที่ต่างๆ กันโดยดูจากค่าความแม่นยำ (accuracy) ถ้าค่าความแม่นยำสูงแสดงว่าเป็นวิธีการที่มีประสิทธิภาพในขบวนการทำนายการจำแนกกลุ่มจากชุดข้อมูลที่วัดจากค่าความแม่นยำก็จะบ่งบอกได้ว่าค่าตัวแทนชุดข้อมูลนี้เป็นค่าที่ดีด้วย

4.3 สรุปวิธีการ Imputation&Tranform

NFDC-DPimpute& Inner distance combination transformation method (NFDCs-DPimpute-IDTC)

จากการทดลองผู้วิจัยจึงได้นำขั้นตอนวิธีที่ใช้ในการศึกษาที่ให้การประเมินประสิทธิภาพที่ดีที่สุด จากสองกระบวนการรวมกระบวนการขั้นตอนเดียว ในกรณีที่ชุดข้อมูลปรากฏค่าสูญหายและทำการทรานฟอร์มเพื่อให้สามารถจัดการเตรียมชุดข้อมูลครั้งเดียวเพื่อให้ได้ข้อมูลชุดใหม่พร้อมในการนำเข้าโมเดลการจำแนกประเภท

Procedure NFDCs-DPimpute-IDTC method

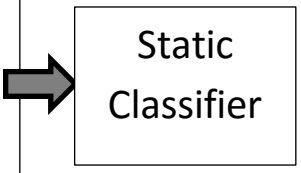
//NFDCs-DPimpute

……..

// Inner distance combination transform

……..

End.



Input : Incomplete temporal dataset (D)

Output : Complete temporal dataset(D)

Step1 : Transform temporal medical data to low dimension in subspace matrix

1.1 Determine the system's input variables for a temporal data matrix

1.2 Separate the matrix into m x n each of patient subspace dimension.

1.3 Transpose patient subspaces with missing values ,

Step2 : Compute in transposed patient subspace.

2.1 Repeat

2.2 Select the transposed subspace in each patient case for imputation.

2.2.1 Repeat

2.2.2 Select row vector in each patient subspace

Step3 : Check time- treatment of Patient –ID-$_{0n}$ and missing position

3.1 if time tratement of patient > 2

3.1.1 Seperate X_n to subset of row vector (4 of set series)

$FD = X_i(t_i)_{max} - X_i(t_i)_{min}$

$FDDV = FD/3$

3.1.2 Repeat

3.1.2.1 Separate data in two set on $X_{obs}(t_i)$, $X_{miss}(t_i)$

3.1.2.2 compute on $X_{obs}(t_i)$ // for polynomial with set of observe values

feature are $X_{obs}(t_i)$, $y_{obs}(t_i)$ with condition order degree on Newton's divide difference table

Condition order degree

observe treatment values at the second time point is second degree

using linear interpolation

$$f_n(x) = b_1 + b_2(x - x(t_{1\text{-}obs}))$$

observe treatment values at the three point is third degree

using quadralic interpolation.

$$f_n(x) = b_1 + b_2(x - x(t_{1\text{-}obs})) + b_3(x - x(t_{1\text{-}obs}))(x - x(t_{2\text{-}obs}))$$

observe treatment values at the fourth point is fourth degree

using cubic interpolation

$$f_n(x) = b_1 + b_2(x - x(t_{1\text{-}obs})) + b_3(x - x(t_{1\text{-}obs}))(x - x(t_{2\text{-}obs})) + b_4(x - x(t_{1\text{-}obs})(x - x(t_{2\text{-}obs}))(x - x(t_{3\text{-}obs}))$$

observe treatment values at the n point is n-th degree using n-th interpolation

$$f_n(x) = b_1 + b_2(x - x(t_{1\text{-}obs})) + b_3(x - x(t_{1\text{-}obs}))(x - x(t_{2\text{-}obs}))) + \ldots + b_n(x - x(t_{1\text{-}obs})) + \ldots)(x - x(t_{n\text{-}1\text{-}obs}))$$

3.1.2.3 Compute $x(t)_{miss}$ // from set of missing values features .

- Compute $f_n(x(t_i))$ with the condition order degree by observer values of n points for using n degree, that is , $f_n(x)$, $x(t_i)$miss

$$f_n(x) = f_n(x(t_i))$$

Compute $f_n(x)$

$$f_n(x) = \sum_{i=1}^{n} \left\{ F[x_1, x_2, \ldots, x_i] \prod_{j=1}^{i-1} (x - x_j) \right\},$$

$$f_n(x) = f(x_1) + (x - x_1) f[x_2, x_1] + (x - x_1)(x - x_2) f[x_3, x_2, x_1] + \cdots + (x - x_1)(x - x_2) \cdots (x - x_{n-1}) f[x_n, x_{n-1}, \cdots, x_1].$$

3.1.2.4 Replace the values with $f_n(x(t_i))$ in missing position [$X((t_i))_{mis}$]

3.1.3 Repeat Step 3.1.2 for next sub-set of series until last sub-set

3.1.4 Repeat Step 2.2.1 for next row in the patient subspace matrix.

3.1.5 Repeat Step 2.1 for next the patient subspace matrix.

3.2 Else if time treatment < = 2

Call DPimpute()

Setp4 : Loop 3.1 next patient one-has missing Until all patient no missing

// **Inner distance combination transformation method**

*S*tep5. Select each patient in the subspace matrix.

5.1 Transpose the patient subspace matrix.

5.2 Repeat

5.2.1 X_i ; i = i+1 , i=1,2,3,…,m

5.2.2 Select a row value in the subspace matrix.

$[X_i[t_1]\ X_i[t_2]\ X_i[t_3]\ X_i[t_4] \ldots X_i[t_n]\ ]$, i=1 … m

5.2.3 Separate a set in the series into two sets.

$[X_i[t_1]\ X_i[t_2]\ X_i[t_3] \ldots X_i[t_{n/2}]]$ and $[X_i[t_{n/2+1}]\ X_i[t_{n/2+2}]\ X[tn_{/2+3}] \ldots\ X_i[t_n]]$

5.2.4 Compute A and B from the first set in 5.2.3

In each level j ; j = 1,2,3, ...,n/2 .

$$j = 1;\quad a_i[1,k] = \begin{cases} \sqrt{x_i[t_k]^2+x_i[t_{k+1}]^2} & , \ k = 1,2,3,\ldots,n/2\text{-}1 \\ \sqrt{x_i[t_k]^2+x_i[t_1]^2} & , \ k = n/2 \end{cases} \quad \ldots\ldots\ldots(1)$$

$$b_i[1,k] = \begin{cases} \sqrt{x_i[t_k]^2+x_i[t_{k+1}]^2} & , \ k = n/2+1, n/2+2,\ldots,n\text{-}1 \\ \sqrt{x_i[t_k]^2+x_i[t_{n/2+1}]^2} & , \ k = n \end{cases} \quad \ldots\ldots\ldots\ldots(2)$$

when $a_i[j,k]$ and $b_i[j,k]$ where j = 2,3,4,…,n/2

$$a_i[j,k] = \begin{cases} \sqrt{a_i[j,k]^2+ a_i[j,k+1]^2} & , \ k = 1,2,3,\ldots,n/2 \\ \sqrt{a_i[j,k]^2+ a_i[j,1]^2} & , \ k = n/2 \end{cases} \quad \ldots\ldots\ldots(3)$$

$$b_i[j,k] = \begin{cases} \sqrt{b_i[j,k]^2+ b_i[j,k+1]^2} & , \ k = 1,2,3,\ldots,n/2 \\ \sqrt{b_i[j,k]^2+ b_i[j,1]^2} & , \ k = n/2 \end{cases} \quad \ldots\ldots\ldots(4)$$

5.2.5 Compute A and B

$$||A|| = \text{Max}\,(a_i[j,k]) \quad , j = n/2 \ , \ k = 1,2,3,\ldots,n/2 \qquad \ldots\ldots\ldots (5)$$

$$||B|| = \text{Max}\,(b_i[j,k]) \quad , j = n/2 \ , \ k = 1,2,3,\ldots,n/2 \qquad \ldots\ldots\ldots (6)$$

5.2.6 Compute D

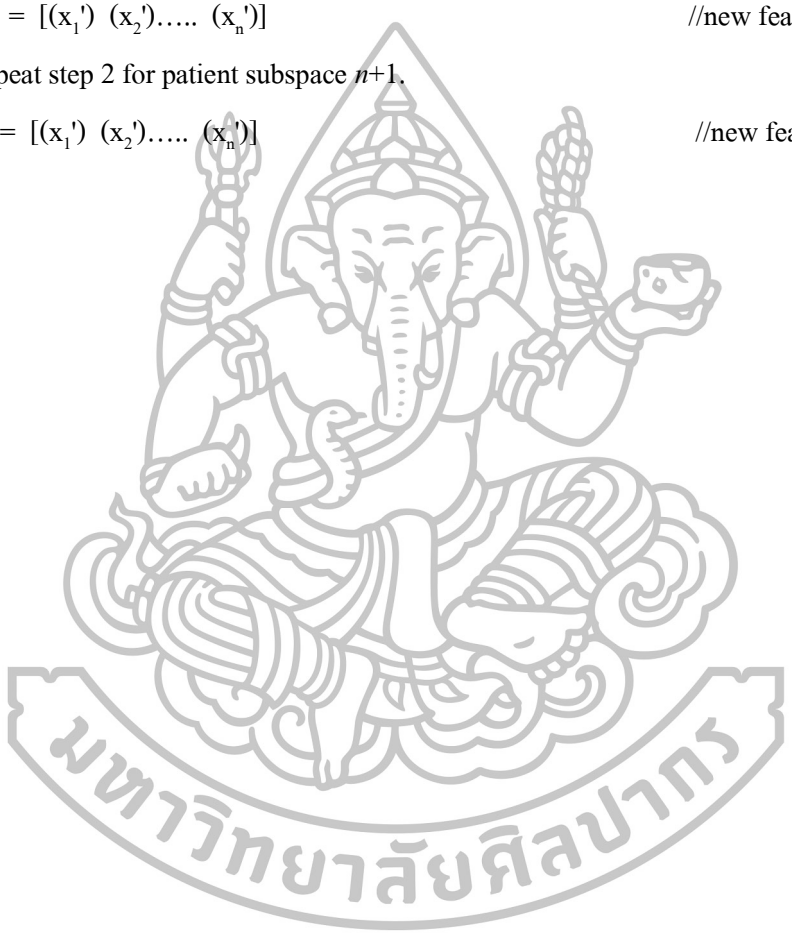
$$D\,(X_i) = \text{Max}\,(||A||, ||B||] \,) \qquad \ldots\ldots\ldots (7)$$

5.2.7 Until i <= m.

Step6. $X_i = [(x_1')\ (x_2')\ldots..\ (x_n')]$ //new feature

Step7. Repeat step 2 for patient subspace n+1.

Step8. $X_i' = [(x_1')\ (x_2')\ldots..\ (x_n')]$ //new feature set



บทที่ 5
สรุปผลการดำเนินงานวิจัย

ในงานวิจัยนี้แสดงให้เห็นปัญหาของการจำแนกกลุ่มจากชุดข้อมูลลำดับเวลาทางการแพทย์ที่มีจำนวนครั้งการตรวจที่ไม่เท่ากันและไม่สมบูรณ์ และแก้ปัญหาด้วยจัดเตรียมข้อมูลในกระบวนการวิธีประมาณค่าสูญหายและทรานฟอร์มเพื่อลดมิติข้อมูลจากชุดข้อมูลลำดับเวลาก่อนนำเข้าสู่กระบวนการจำแนกประเภท

5.1 สรุปผลการวิจัย

งานวิจัยนี้เน้นประกอบวัตถุประสงค์เพื่อเป็นจัดเตรียมข้อมูล (Pre-processing) ให้ไม่เกิดปัญหาในการนำเข้าข้อมูลพร้อมประมวลผลในกระบวนการจำแนก แต่ยังคงความแม่นยำในการจำแนกประเภทกลุ่ม ซึ่งได้แบ่งงานวิจัยโดยมีสองวัตถุประสงค์ที่กำหนดไว้เพื่อหาวิธีการที่ดีที่สุดจากการวัดประสิทธิภาพในแต่ละส่วนสำหรับกระบวนการวิธีใหม่ที่พัฒนา นั่นคือ

1. ส่วนของทำนายค่าด้วยการประมาณค่าสูญหายในชุดข้อมูลทางการแพทย์เชิงลำดับเวลา
2. ส่วนของการทรานฟอร์มเพื่อลดมิติข้อมูลจากผลของการประมาณค่าจากชุดข้อมูลสมบูรณ์

สำหรับในส่วนของวัตถุประสงค์ข้อ 1 การทำนายค่าเพื่อประมาณค่าสูญหายในชุดข้อมูลลำดับเวลาที่นำเสนอได้นำเสนอวิธีการในการประมาณค่าสูญหายเพื่อที่จะหาวิธีการที่สามารถให้ค่าการประมาณครบถ้วนทั้งชุดข้อมูลลำดับเวลา โดยมีแนวคิดหลักของขั้นตอนที่นำเสนอในงานวิจัยอยู่บนแนวคิดการใช้ค่าเฉพาะรายบุคคล รวมกับการหาค่าความเหมือนหรือความคล้ายมาเป็นค่าในการประมาณเพื่อทำนายค่า ด้วยวิธีการ NFDCs-DPimpute , DPimpute , CBE-DPimpute, SLLs-DPimpute , Sknn-DPimpute, NFDCsSlideW-DPimpute จากการศึกษากับชุดข้อมูลสองชุด คือ ชุดข้อมูลลำดับเวลาผู้มาตรวจรักษาโรคอ้วน(Obesity) และ ชุดข้อมูลลำดับเวลาผู้ป่วยโรคหลอดเลือดหัวใจอุดตัน(Thrombosis) จากการวัดประสิทธิภาพค่าความคลาดเคลื่อน เพื่อประเมินความแม่นยำของการประมาณค่าสูญหายจากการแบ่งเปอร์เซนต์ของค่าสูญหายเพื่อการประมาณค่า

ผลการทดลองจะแสดงให้เห็นถึงความแม่นยำของวิธีการ NFDCs-DPimpute จะให้ความแม่นยำที่ดีที่สุดจากวิธีการอื่นๆ สำหรับในส่วนวัตถุประสงค์ที่สองเมื่อค่าประมาณค่าและทดแทนค่าข้อมูลที่สูญหายได้ชุดข้อมูลสมบูรณ์แล้ว นั่นคือ อย่างแรก จะนำชุดข้อมูลลำดับเวลาที่เป็นชุดข้อมูลที่ผ่านการประมาณค่าสูญหายและได้เป็นชุดข้อมูลสมบูรณ์เพื่อทรานฟอร์มรายการค่าข้อมูลบนหลักการทรานฟอร์ม ซึ่งจะทำการลดมิติข้อมูลรายบุคคลที่มีค่าการตรวจหลายครั้งเพื่อหาตัวแทนชุดข้อมูลได้ค่าข้อมูลชุดใหม่ของแต่ละบุคคลในลักษณะค่าข้อมูลเดี่ยว(singular values) ซึ่งจะทำให้ได้ชุดข้อมูลขนาดเล็กลง จึงนำเสนอขั้นตอนใหม่ที่พัฒนา คือ วิธีการ Inner distance combination transform (IDCT) ซึ่งแนวคิดหลักรูปแบบภายใต้การคำนวณด้วย inner distance function และหลักการรวม(combination) ของแต่ละชุด โดยนำมาเปรียบเทียบกับหลักการทางสถิติคือ mean, median, stdev, variance สำหรับข้อมูลชุดใหม่ที่เป็นตัวแทนชุดข้อมูลเมื่อนำเข้าในโมเดลในการจำแนกกลุ่ม ด้วยโมเดล Support vector machine จากตัววัดการประเมินประสิทธิภาพความแม่นยำของการจำแนกกลุ่ม โดยขั้นตอนวิธี IDCT จะให้ค่าความแม่นยำในการจำแนกกลุ่มซึ่งถือได้ว่าได้ค่าตัวแทนชุดข้อมูลที่ดีในการนำเข้าโมเดลในการเรียนรู้การจำแนกประเภทที่ดีกว่าขั้นตอนอื่นที่ใช้ในการเปรียบเทียบหลังจากนั้นจะทำการเลือกขั้นตอนวิธีที่ดีที่สุดมารวมเพื่อให้สามารถจัดการข้อมูลสูญหายและแปลงค่าข้อมูลให้เป็นกระบวนการเดียว เพื่อลดขั้นตอนในการประมวลผล ขั้นตอนใหม่นี้เรียกว่า NFDCs-DP-IDTC

6.2 ปัญหาและข้อเสนอแนะในงานวิจัย

สำหรับงานงานวิจัยนี้คือ เราจะได้วิธีการในการประมาณค่าสูญหายและลดมิติข้อมูลด้วยกระบวนการทรานฟอร์มในชุดข้อมูลเชิงเวลาทางด้านการแพทย์ที่ได้ค่าครบทั้งรายการและชุดข้อมูลสำหรับผู้ที่มาตรวจครั้งเดียวและผู้มาตรวจหลายครั้งบนหลักการสมมติฐาน คือการใช้ค่าเฉพาะรายบุคคลและค่าความเหมือนความคล้ายของทั้งหมดเป็นค่าในการประมาณข้อมูลเพื่อให้ได้ค่าการประมาณที่ยอมรับได้และยังคงประสิทธิภาพในการจำแนกประเภท ซึ่งสามารถประยุกต์ใช้กับค่าข้อมูลในงานจริงที่มีโครงสร้างลักษณะเดียวกัน

ข้อเสนอแนะ

1. สำหรับการประมาณค่าสูญหายและการทรานฟอร์มข้อมุลในการหาตัวแทนชุดข้อมูล หากนำโมเดลนี้ไปใช้ในการประมาณค่าสูญหายในชุดข้อมูลทางแพทย์ หรือประยุกต์ใช้กับลักษณะชุดข้อมูลที่มีลักษณะคล้ายๆ กัน ควรจัดการโครงสร้างให้อยู่ในลักษณะชุดข้อมูลเชิงเวลาตามโครงสร้างในงานวิจัยซึ่งสามารถใช้ได้กับจำนวนคุณลักษณะตัวชี้วัดที่มีจำนวนมากได้

2. สำหรับวิธีการที่ใช้ในงานวิจัย ในการประมาณค่าสูญหายได้นำหลายๆ หลักการมาใช้เพื่อให้ได้ค่าการประมาณที่ใกล้เคียงและครบถ้วน เช่น n-euclidean distance, k-nearest neightbor หลักการประมาณค่าในช่วง สมการถดถอย, Sliding windows และอื่นๆ ที่เกี่ยวข้องใช้ในการทดลอง หลังจากนี้หากต้องการเพิ่มประสิทธิภาพการประเมินค่าความคาดเคลื่อนของการประมาณค่าสูญหาย อาจศึกษาถึงเทคนิควิธีการอื่นๆ ที่จะทำการประมาณค่าสูญหายตามโครงสร้างของชุดข้อมูลในงานวิจัยให้มีประสิทธิภาพมากขึ้น

3. แนวทางพัฒนาต่อ หากนำหลักการประมาณค่าสูญหายหรือทรานฟอร์มข้อมูลที่พัฒนาไปประยุกต์เข้ากับโมเดลสำหรับการจำแนกในลักษณะแบบฝังติด(Embeded classifier) จะลดขั้นตอนการประมวลผลกระบวนการจำแนกประเภทเมื่อปรากฏชุดข้อมูลที่มีค่าสูญหายหรือลักษณะข้อมูลที่อยู่ในรูปแบบเชิงเวลา

รายการอ้างอิง

[1] A. Patel.(2004) "Impact of missing data in trainingartificial neural networks for computer-aided Diagnosis." **Inter-national Conference on Machine Learning and Applications2004.** pp.351 – 354.

[2] C. M. Antunes and A. L. Oliveira(2001) " Temporal data mining: an overview." **KDD 2001 Workshop on Temporal Data Mining**. San Francisco, CA, August 26.

[3] Pierre Geurts. (2001) " Pattern extraction time series classification.PKDD 2001. " LNAI 2168, pp. 115- 127.

[4] S.Thrun. (1996) " Is learning the n-thing any easier than learning the first. " **Advance in Neural Information Processing Systems (NIPS)**. p.640-646.

[5] Mohd. Shahnawaz, AshishRanjan, Mohd Danish.(2011). "Temporal Data Mining: An Overview." **International Journal of Engineering and Advanced Technology (IJEAT)** ISSN:2249 – 8958, Volume-1, Issue-1.

[6] Zhang, C.Q., et al.(2007) " An Imputation Method for Missing Values." **PAKDD, LNAI 4426**, pp. 1080 – 1087.

[7] Ahmed Y. Tawfik and Krista Stricklan.(2000) " Mining Medical Data for Causal and Temporal Patterns." **The Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases**- *PKDD-2000.*

[8] LyadBatal, SmitriyFradkin , James Harrison. (2012) " **Mining Recent Temporal Patterns for Event Detection in Multivariate Time series data**. KDD'12**.**

[9] S. Tsumoto. (1999) "Rule discovery in large time-series medical databases," **Principles of Data Mining and Knowledge Discovery, Lecture Notes in Computer Science Volume 1704**, Heidelberg: Springer Berlin, pp. 23-31.

[10] McCurdy A, K. C. Ng, B. N. Parlett. (1998). **Mathematics of Computation**, 168p.

[11] H. Kim, G.H. Golub and H. Park.(2005) "Missing value estimation for DNA microarray gene expression data: local least squares imputation." Bioinformatics 21 . 187-198.

[12] C.Balasubramanian, Dr.K.Duraiswamy. (2009) "An Application of Bayesian classification to Interval Encoded Temporal mining withprioritized items." **International Journal of Computer Science and Information Security(IJCSIS)**. Vol. 3, No. 1.

[13] Juthaphorn Sinsomboonthong. Kasetsart J.(2011) " Estimation of the Correlation Coefficient for a Bivariate Normal Distribution with Missing Data." **Natural Science 45** : 736 – 742p.

[14] Klaokanlaya Silachan and Panjai Tantatsanawong. (2014) " Imputation of Medical Data Using Subspace Condition Order Degree Polynomials." **Journal of Information Processing(JIPS)**. Vol10, No3, 2014.

[15] J. M. Jerez, I. Molina, J. L. Subirats, and L. Franco.(2006) " Missing data imputation in breast Cancer prognosis" **Proceedings of the 24th IASTED International Conference on Biomedical Engineering**, Innsbruck, Austria, pp. 323-328.

[16] Batal, I. ; Valizadegan, H. ; Cooper, G.F. ; Hauskrecht, M.(2011) " A Pattern Mining Approach for Classifying Multivariate Temporal Data**." IEEE International Conference on Bioinformatics and Biomedicine (BIBM2011)**.

[17] Hirano,S.Xiaoguang Sun,Tsumoto,S. (2001) " Using time-dependent neural networks for EEG Classification", **In Proceeding of the 10th IEEE International Conference on Fuzzy Systems**,Vol 3, pp. 1547 - 1550 .

[18] A. A. Al-Hussainan**,** B. M. Al-Eideh, and Y. S. H. Al-Zalzalah. (2001) "The Adjusted Empirical Lorenz Curve Using a Newton's Divided Difference Formula." **International Journal of Applied Mathematics**. Volume 7, No. 3.

[19] J. F. Roddick and M. Spiliopoulou (2002). “A survey of temporal knowledge discovery paradigms and methods.” **IEEE Transactions on Knowledge and Data Engineering**, vol. 14, no. 4, pp.750-767.

[20] M. N. Noraziana, Y. A. Shukric, R. N. Azamc, and A. M. M. Al Bakrib (2008) “Estimation of Missing Values in air pollution data using single imputation techniques. “ ***ScienceAsia*.** vol. 34, no. 3, pp. 341-345.

[21] S. Bose, C. Das, S. Dutta, and S. Chattopadhyay(2012) “ A novel interpolation based missing value estimation method to predict missing values in microarray gene expression data." ***International Conference on Communications, Devices and Intelligent Systems*,** December 28-29, 2012, pp. 318-321.

[22] N. Viana, A. Pereira, R. Ribeiro, and A. Donati(2004). “ Handling missing values in solar array performance degradation forecasting.” **Proceedings of the 15th Mini-EURO Conference on Managing Uncertainty in Decision Support Models**.September 22-24.

[23] N. Eisemann, A. Waldmann, and A. Katalinic. (2011) “Imputation of missing values of tumour stage in population-based cancer registration." **BMC Medical Research Methodology**. vol. 11, p. 129, Sep.2011.

[24] ธีระพล สลีวงศ์,ชื่นชม พงษ์ชวลิต. (2542) “ระเบียบวิธีเชิงตัวเลข.” **สำนักพิมพ์มหาวิทยาลัยธนบุรี** ครั้งที่ 2.2542.

[25] Froberg,C.E (1969) “Introduction to Nemurical Analysis. **“ Addition_wesley, Reading,Mass.**

[26] ปราโมทย์ เดชะอำไพ,ศ.ดร.(2549). “ ระเบียบวิธีเชิงตัวเลขในงานวิศวกรรม.“ สำนักพิมพ์จุฬาลงกรณ์ มหาวิทยาลัย,ครั้งที่5.2549.

[27] Nasrin AkterRipa.(2010) " Analysis of Newton's Forward Interpolation Formula.**"** **International Journal of Computer Science & Emerging Technologies (E-ISSN: 2044-6004)**, Volume 1, Issue 4.

[28] D. Moore and G. McCabe.(2003) " Introduction to the Practice of Statistics." **W. H. Freeman and Co.** London.

[29] Peter J. Olver.(2008) " Applied Linear Algebra." **Prentice-Hall**.

[30] E. Acuña and C. Rodriguez (2004). "The Treatment of Missing Values and its Effect on Classifier Accuracy," **Classification, Clustering, and Data Mining Applications**. :Springer Berlin Heidelberg, pp. 639-647.

[31] Klaokanlaya Silachan and Panjai Tantatsanawong. (2015) "An Inner Distance Combination Transform for Classification of Temporal Medical Data." **Journal of Coverage Information System(JCIT)**. Vol10, No4.

[32] Akanksha Singh Thakur, NamrataSahayam , ME IV Sem .(2013) " Speech Recognition Using Euclidean Distance." **International Journal of Emerging Technology and Advanced Engineering**.Volume 3, Issue 3.

[33] Ernst, H., and Gert, P.(2000) **"**Using Time-Dependent Neural Networks for EEG Classification." **IEEE Trans. Rehabil. Eng,** Vol. 8, no. 4, pp. 457–463.

[34] Bellazzi, R., Sacchi, L., and Concaro, S.(2009) " Methods and tools for mining multivariate temporaldata in clinical and biomedical applications." **Proceeding of IEEE Engineering Medicineand Biology Society.** pp. 5629-5632.

[35] Peter, R., Thomas, T.(2011) " Temporal data classification using linear classifiers ", **Journal of information Systems, Vol 36, pp. 30–41.**

[36] Veen, Ranjit S. (2008). Medical data mining issues and experiments. . **AMERICAN UNIVERSITY.**

[37] Sarah A. Soliman; Safia Abbas; Abdel-Badeeh M. Salem .(2015) " Classification of thrombosis collagen diseases based on C4.5 algorithm " **IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS).**

ประวัติผู้วิจัย

ชื่อ – สกุล	เกล้ากัลยา ศิลาจันทร์
ที่อยู่	85/37 ต.นครปฐม อ.เมือง จ.นครปฐม
ที่ทำงาน	คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม

ประวัติการศึกษา

พ.ศ. 2534	สำเร็จการศึกษาอนุปริญญาวิทยาศาสตร์ สาขาวิชาคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี วิทยาลัยครูมหาสารคาม
พ.ศ. 2537	สำเร็จการศึกษาวิทยาศาสตรบัณฑิต สาขาวิชาคอมพิวเตอร์ศึกษา เกียรตินิยมอันดับ 2 คณะวิทยาศาสตร์และเทคโนโลยี วิทยาลัยครูสุรินทร์
พ.ศ. 2542	สำเร็จการศึกษาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีการจัดการระบบสารสนเทศ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยมหิดล

ประวัติการทำงาน

พ.ศ. 2537	ปฏิบัติงานสอน ประจำภาควิชาคอมพิวเตอร์ วิทยาลัยครูมหาสารคาม
พ.ศ. 2537 –	รับราชการปฏิบัติงานสอน ประจำภาควิชาคอมพิวเตอร์ สถาบันราชภัฏนครปฐม
ปัจจุบัน	รับราชการปฏิบัติงานสอน ประจำสาขาเทคโนโลยีคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม