การเลือกคุณลักษณะด้วยวิธีสมาชิกที่ใกล้ที่สุด สำหรับการแทนที่ค่าข้อมูลที่ขาดหายด้วยวิธีที่ใกล้ที่สุด

ไก้รุ่ง เฮงพระพรหม* และ พยุง มีสัจ**

บทคัดย่อ

ข้อมูลไมโครอาร์เรย์นิยมใช้ในการศึกษารูปแบบของสิ่งมีชีวิตในระดับโมเลกุล ประกอบด้วยตัวอย่างการทดลองน้อย แต่มีมิติสูง ข้อมูลส่วนใหญ่ที่ได้จะมีค่าข้อมูลที่ขาดหายเป็นจำนวนมาก ทำให้ประสิทธิภาพในการวิเคราะห์ข้อมูลลดลง โดยอัลกอริธึมสำหรับการวิเคราะห์ข้อมูลส่วนใหญ่อัลกอริธึมต้องการข้อมูลที่สมบูรณ์เพื่อให้การวิเคราะห์ข้อมูลมีประสิทธิภาพที่ดีจึงทำให้ต้องการวิธีการแทนค่าข้อมูลที่ขาดหายที่มีประสิทธิภาพ ในงานวิจัยนี้วัตถุประสงค์เพื่อต้องการนำเสนอวิธีการเลือก Feature ที่เหมาะสมในการ Estimate Missing Values สำหรับข้อมูลไมโครอาร์เรย์ด้วยวิธีการ K-Nearest Neighbor ที่ประกอบด้วยสองขั้นตอน คือ การเลือกคุณลักษณะของข้อมูลด้วยวิธีการ K-Nearest Neighbor และการแทนที่ค่าข้อมูลที่ขาดหายด้วยวิธีการ K-Nearest Neighbor จากนั้นทำการเปรียบเทียบประสิทธิภาพกับวิธีการ K-Nearest Neighbor แบบปกติ และ Row Average โดยทำการทดลองกับ 3 Data Set ได้แก่ Lung Tumor, Colon Cancer และ ALL-AML Leukemia dataset จากผลการทดลองพบว่าวิธีการ KNNFS Impute ที่นำเสนอให้ประสิทธิภาพที่ดีกว่าเมื่อเทียบกับวิธีการ K-Nearest Neighbor แบบปกติ และ ROWAverage

คำสำคัญ: ไมโครอาร์เรย์ การแทนที่ข้อมูล ค่าข้อมูลที่ขาดหาย วิธีสมาชิกที่ใกล้ที่สุด

1. บทนำ

เทคโนโลยีไมโครอาร์เรย์ (Microarray Technology) เป็นเทคนิคที่นิยมใช้ในการศึกษารูปแบบของสิ่งมีชีวิตในระดับโมเลกุล ข้อมูลไมโครอาร์เรย์เป็นข้อมูลที่แสดงถึงระดับการแสดงออกของยีน (Gene) หลายพันยีนในเวลาเดียวกัน ซึ่งช่วยให้นักวิจัยในหลากหลายสาขาสามารถศึกษากลไกการทำงานของสิ่งมีชีวิต เช่นการเจริญเติบโตของสิ่งมีชีวิตในช่วงเวลาหรือสภาพแวดล้อมต่าง ๆ ตลอดจนการวินิจฉัยโรคและการค้นพบยารักษาโรค โดยใช้เทคนิคทางการทำเหมืองข้อมูล (Data Mining) และการเรียนรู้ของเครื่องจักร (Machine Learning) ซึ่งการวิเคราะห์ข้อมูลไมโครอาร์เรย์ ได้แก่ การจำแนกประเภท (Classification) [1-3] และการจัดกลุ่ม (Clustering) [4-5] โดยการวิเคราะห์ข้อมูลทางสถิติดังกล่าวหลาย ๆ อัลกอริธึมต้องการข้อมูลที่สมบูรณ์ [6] เพื่อให้การวิเคราะห์ได้ประสิทธิภาพสูงสุด

แต่เนื่องจากข้อมูลไมโครอาร์เรย์ส่วนใหญ่มักประกอบด้วยค่าที่ขาดหาย (Missing Value) เป็นจำนวนมาก ซึ่งมีสาเหตุมาจากหลากหลายปัจจัยด้วยกัน ได้แก่ ความละเอียดที่ไม่เพียงพอและความผิดพลาดของรูปภาพ ฝุ่นหรือรอยขีดข่วนบนแผ่น Slide หรือความผิดพลาดในระหว่างกระบวนการทดลอง หากต้องทำการทดลองซ้ำก็จะทำให้เสียเวลาและค่าใช้จ่ายเป็นจำนวนมาก [7] การจัดการกับข้อมูลที่ขาดหายสามารถทำได้หลายวิธี เช่น การตัดเอาระเบียนหรือคุณสมบัติที่มีค่าที่ขาดหายออกหรือโดยการแทนค่าข้อมูลที่ขาดหายด้วยค่าใด ๆ แต่เนื่องจากธรรมชาติของข้อมูลไมโครอาร์เรย์นั้นประกอบด้วยจำนวนตัวอย่างข้อมูลที่มีจำนวนน้อย ถ้ากรณีที่ใช้วิธีการตัดเอาข้อมูลตัวอย่างที่มีค่าที่ขาดหายออก อาจทำให้จำนวนตัวอย่างในการทดสอบมีไม่เพียงพอต่อการวิเคราะห์ข้อมูล ดังนั้นวิธีการที่นิยมใช้สำหรับการจัดการกับค่าข้อมูลที่ขาดหายสำหรับข้อมูลไมโครอาร์เรย์ จึงนิยมใช้การแทนค่าด้วยค่าใด ๆ ซึ่งค่าที่นิยมใช้ในการแทนที่ คือ การแทนค่าข้อมูลที่ขาดหายด้วยศูนย์หรือค่าเฉลี่ย [8] แต่วิธีการทั้งสองนี้เป็นวิธีที่ให้ค่าการประมาณ Missing values ที่ไม่เหมาะสม

มีงานวิจัยจำนวนมากที่ได้ศึกษาและพัฒนาวิธีการแทนค่า

* *ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ*

** *ภาควิชาครุศาสตร์ไฟฟ้า คณะครุศาสตร์อุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ*

ข้อมูลที่ขาดหายที่มีประสิทธิภาพสำหรับปัญหาต่าง ๆ โดยใช้เทคนิคทางสถิติ การทำเหมืองข้อมูล รวมไปถึงเทคนิคการเรียนรู้ของเครื่องจักร เพื่อเพิ่มประสิทธิภาพในการวิเคราะห์ข้อมูล เช่น Troyanskaya et al. [9] ได้ทำการศึกษาเปรียบเทียบวิธีการแทนที่ Missing values แบบ K-nearest neighbor algorithm (KNN impute) Singular Value Decomposition (SVD) และ Row average ผลที่ได้ คือ KNN impute ให้ประสิทธิภาพในการประมาณค่าที่ดีกว่า Oba et al. ได้พัฒนาวิธีการสำหรับ Estimate Missing Values เรียกว่า Bayesian Principal Component Analysis (BPCA impute) [10] ที่สามารถเอาชนะวิธีการ KNN และ SVD ได้ อีกวิธีการหนึ่งที่ให้ประสิทธิภาพ คือ Bayesian Variable Select Algorithm พัฒนาโดย Zhou et al. [11] วิธีการนี้มีความสามารถในการเลือกพารามิเตอร์(ยีน)ที่ใช้ในการ Estimate Missing Values โดยอัตโนมัติ อัลกอริธึมนี้ใช้ทั้ง Linear และ Nonlinear Regression หัวใจสำคัญของอัลกอริธึมคือการทำงานอย่างรวดเร็ว Kim et al. [12] ได้นำเสนอวิธีการ Local least squares (LLS impute) โดยใช้ประโยชน์จากโครงสร้างความเหมือนของข้อมูลเช่นเดียวกับกระบวนการ least squares optimization. Robust Least Squares Estimation with Principal Components (RLSP) ถูกนำเสนอโดย Yoon et al. [13] โดยทำการปรับปรุงประสิทธิภาพของวิธีการ LLS impute เดิม โดยนำวิธีการ Principal Component Analysis มาใช้ในการหาความสัมพันธ์ของข้อมูล วิธีการ RLSP ให้ประสิทธิภาพที่ดีกว่าเมื่อเทียบกับวิธีการ KNN, LLS impute, BPCA

สำหรับงานวิจัยนี้ผู้วิจัยมีวัตถุประสงค์เพื่อต้องการนำเสนอวิธีการเลือก Feature ที่เหมาะกับสมการ Estimate Missing Values สำหรับข้อมูลไมโครอาร์เรย์ด้วยวิธีการ K-Nearest Neighbor Feature Selection for K-Nearest Neighbor Imputation (KNNFS Impute) ที่ประกอบด้วยสองขั้นตอน คือ การเลือกคุณลักษณะของข้อมูลด้วยวิธีการ K-Nearest Neighbor และการแทนที่ค่าข้อมูลที่ขาดหายด้วยวิธีการ K-Nearest Neighbor จากนั้นทำการเปรียบเทียบประสิทธิภาพกับวิธีการ K-Nearest Neighbor แบบปกติ และ Row Average ในหัวข้อต่อไปของเอกสารงานวิจัยจะกล่าวถึงความรู้พื้นฐานเกี่ยวกับงานวิจัย หัวข้อที่ 3 จะแสดงรายละเอียดของการออกแบบการทดลอง หัวข้อที่ 4 จะนำเสนอผลการทดลองและจะสรุปผลและอภิปรายผลในหัวข้อที่ 5

2.ทฤษฎีและวรรณกรรมที่เกี่ยวข้อง

2.1 ข้อมูลไมโครอาร์เรย์

ไมโครอาร์เรย์เป็นเทคนิคหนึ่งที่ถูกพัฒนาขึ้นมาเพื่อใช้ในการศึกษาการแสดงออกของยีน โดยตรวจสอบว่าในเวลานั้น ๆ มีเอ็มอาร์เอ็นเอชนิดใดอยู่ในเซลล์ นอกจากใช้ในการตรวจสอบอาร์เอ็นเอแล้วยังใช้ตรวจสอบจีโนมิกดีเอ็นเอ (Genomic DNA) เช่น การตรวจสอบการกลายพันธ์ของดีเอ็นเอ (DNA Mutation) และตรวจการเปลี่ยนแปลงของยีนในโครโมโซมได้

โดยข้อมูลไมโครอาร์เรย์เกิดจากการนำเอาดีเอ็นเอที่ต้องการตรวจสอบ (DNA Sample) กับดีเอ็นเออ้างอิง (DNA Library) มาฉาบด้วยสารเรืองแสดงคนละสี จากนั้นนำมาทำปฏิกิริยาไฮบริไดเซชัน (Hybridization) บนแผ่นสไลด์ จากนั้นทำการตรวจสอบปริมาณความเข้มของแต่ละสีและแปลงค่า (Transformation) ให้เป็นข้อมูลไมโครอาร์เรย์ [14]

2.2 วิธีสมาชิกที่ใกล้ที่สุด (K-Nearest Neighbor: KNN)

วิธีสมาชิกที่ใกล้ที่สุด(K-Nearest Neighbor: KNN) เป็นวิธีการที่ง่ายและมีประสิทธิภาพสำหรับงานการจำแนกประเภทข้อมูล (Classification) (รายละเอียดเพิ่มเติมดูได้จาก [15]) มีขั้นตอนการดำเนินการดังนี้

2.2.1 กำหนดค่า K

2.2.2 คำนวณหาระยะห่างระหว่างข้อมูลตัวอย่างที่สนใจกับข้อมูลอื่นๆ ทุกตัวด้วยวิธีการ Euclidian distance จากสมการดังนี้

$$dist(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{i,k} - x_{j,k})^2} \qquad (1)$$

โดยที่

$dist(x_i, x_j)$ คือ ระยะห่างระหว่างตัวอย่าง x_i กับตัวอย่าง x_j

n คือ จำนวนคุณสมบัติทั้งหมดของตัวอย่าง

$x_{i,k}$ คือ คุณสมบัติตัวที่ k ของตัวอย่าง x_i

2.2.3 เลือกค่าข้อมูลที่มีค่า $dist$ น้อยที่สุด k ตัว เพื่อนำมาพิจารณาหาคำตอบ

3. K-Nearest Neighbor Feature Selection for K-Nearest Neighbor Imputation (KNNFS Impute)

การจัดการกับ Missing Values ด้วยวิธีการ Row average

มักถูกโน้มเอียงด้วย Outlier หรือ ข้อมูลที่ไม่สัมพันธ์กัน เนื่องมาจากการ Estimate Missing Values จะนำ Record ทั้งหมดของข้อมูลมาหาค่าเฉลี่ยเพื่อนำไป Impute ข้อมูล โดยไม่มีการศึกษาถึงความสัมพันธ์ของข้อมูลก่อนซึ่งหากมีบาง Record มีข้อมูลที่เป็น Outlier มีผลให้การ Estimate Missing Value ให้ประสิทธิภาพที่ต่ำ

ส่วนวิธีการแบบ KNN ปกตินั้นเป็นวิธีการที่ให้ประสิทธิภาพที่ดีกว่า Row Average [9] แต่ประสิทธิภาพยังไม่ดีเท่าที่ควร เนื่องจากการ Estimate Missing Values ด้วยวิธีการ KNN จะทำการคำนวณหา Record (Case) ที่ใกล้เคียงกับ Missing Values มากที่สุด จาก Feature (Column) ทั้งหมดของข้อมูล เพื่อนำมา Impute Missing Value แต่ถ้ามีบาง Feature ที่ไม่สัมพันธ์ (ไม่มีความใกล้เคียงกัน) กับ Missing Values ถูกใช้ในการคำนวณ ผลลัพธ์ คือ Record ที่ได้อาจถูกโน้มเอียงด้วย Feature ที่ไม่เหมาะสมซึ่งจะทำให้ได้ Record ที่ได้ไม่ใกล้เคียงกับ Missing Values ที่ต้องการ Impute จริงๆ จึงทำให้ประสิทธิภาพของการ Estimate Missing Values ที่ได้ไม่ดี

จากเหตุผลดังกล่าวข้างต้น ผู้วิจัยจึงมีแนวความคิดที่ต้องการปรับปรุงประสิทธิภาพของวิธีการ KNN เดิม โดยลดความโน้มเอียงของข้อมูลที่เกิดจาก Feature ที่ไม่เหมาะสมในการหา Record ที่ใกล้กับ Missing Values ที่ต้องการ Impute โดยการนำเสนอวิธีการใหม่เรียกว่า K-Nearest Neighbor Feature Selection for K-Nearest Neighbor Imputation (KNNFS Impute) ซึ่งวิธีการใหม่ที่นำเสนอนี้จะมีขั้นตอนของการเลือก Feature ที่เหมาะสมที่เป็นหัวใจสำคัญสำหรับใช้ในการ Estimate Missing Value โดยวิธีการใหม่นี้ประกอบด้วยขั้นตอนวิธีดังภาพที่ 1

Step 1 Initialize K_F feature

Step 2 Calculate feature distance between X_j j=1,...,col and X_{miss} (Feature with missing values) by using Euclidean Distance...(1)

Step 3 Sort feature distance in an ascending order

Step 4 Select K_F minimums distance

Step 5 Initialize K_C samples

Step 6 Use K_F feature calculate sample distance between R_i, i=1,...,row and R_{miss} (Row with missing values) by using Euclidean Distance...(1)

Step 7 Sort sample distance ascending

Step 8 Select K_C minimums distance

Step 9 Use K_C sample to estimate missing value by average

ภาพที่ 1 KNNFS อัลกอริธึม

ตัวอย่างการคำนวณ KNNFS

ขั้นตอนที่ 1 เลือก K Genes ที่คล้ายกัน

เพื่อที่จะ Estimate Missing Values ต้องอาศัยวิธีการเลือกยีนส์ในการเลือกตัวอย่างที่มีประสิทธิภาพ ขั้นตอนนี้จะลดความโน้มเอียงของข้อมูลที่เกิดจากการเลือก Feature หรือยีนส์ที่ไม่มีความใกล้เคียงกัน ซึ่ง K Genes ที่คล้ายกันจะถูกใช้สำหรับ KNN อีกครั้ง ขณะที่ K เป็นจำนวนที่ถูกกำหนดไว้ล่วงหน้า ใน KNNFS Euclidean distance จะถูกใช้สำหรับการเลือก K Gene ที่คล้ายกัน ดังตัวอย่าง

ตารางที่ 1 ตัวอย่าง Data Set ที่ไม่สมบูรณ์

	A1	A2	A3	A4	A5	A6	A7	A8
1	?	4	5	6	2	3	6	1
2	2	3	4	5	6	5	4	4
3	3	4	5	6	3	2	5	2
4	2	3	2	3	4	1	2	1
5	3	3	1	2	5	3	3	4

กำหนดให้ Aij เป็นสมาชิกของ Dij โดยที่ A11 เป็น Missing Value มีขั้นตอนในการคำนวณดังนี้

1) กำหนด K_F= 3

2) คำนวณหา Distance ของ Feature A11 กับทุก Attribute โดยใช้ Euclidean distance สมการที่ (1) ดังนี้

$$dist(A_1, A_2) = \sqrt{(2-3)^2 + (3-4)^2 + (2-3)^2 + (3-3)^2} = 1.732$$
$$dist(A_1, A_3) = \sqrt{(2-4)^2 + (3-5)^2 + (2-2)^2 + (3-1)^2} = 3.464$$
$$dist(A_1, A_4) = \sqrt{(2-5)^2 + (3-6)^2 + (2-3)^2 + (3-2)^2} = 4.472$$
$$dist(A_1, A_5) = \sqrt{(2-6)^2 + (3-3)^2 + (2-4)^2 + (3-5)^2} = 4.123$$
$$dist(A_1, A_6) = \sqrt{(2-5)^2 + (3-2)^2 + (2-1)^2 + (3-3)^2} = 3.317$$
$$dist(A_1, A_7) = \sqrt{(2-4)^2 + (3-5)^2 + (2-2)^2 + (3-3)^2} = 2.828$$
$$dist(A_1, A_8) = \sqrt{(2-4)^2 + (3-2)^2 + (2-1)^2 + (3-4)^2} = 2.646$$

3) จากนั้นเรียงลำดับ Attribute ที่มี Distance โดยเรียงจากน้อยไปหามาก ดังนี้ A2, A8, A7, A6, A3, A5, A4

4) เลือก K_F Attribute ไปใช้คำนวณในขั้นตอนที่ 2 ซึ่งก็คือ A2, A8, A7

ขั้นตอนที่ 2 การเลือก Samples

หลังจากได้ K_F Attribute ที่เป็นตัวแทนของข้อมูลทั้งหมดที่ผ่านการลดความโน้มเอียงของ Attribute เรียบร้อยแล้วนำมาคำนวณหา Sample ที่ใกล้ที่สุด K_C Sample ดังตัวอย่าง

ตารางที่ 2 ตัวอย่าง Data Set ที่ได้จากการเลือก Attribute

	A1	A2	A7	A8
1	?	4	6	1
2	2	3	4	4
3	3	4	5	2
4	2	3	2	1
5	3	3	3	4

5) กำหนด K_C =3

6) คำนวณหา Distance ของ Sample ด้วย Feature ที่เป็นตัวแทนจากขั้นตอนที่ 1 A11 กับ ทุก Record โดยใช้ Euclidean distance สมการที่ (1) ดังนี้

$$dist(R_1, R_2) = \sqrt{(4-3)^2 + (6-4)^2 + (1-4)^2} = 3.741$$
$$dist(R_1, R_3) = \sqrt{(4-4)^2 + (6-5)^2 + (1-2)^2} = 1.414$$
$$dist(R_1, R_4) = \sqrt{(4-3)^2 + (6-2)^2 + (1-1)^2} = 4.123$$
$$dist(R_1, R_5) = \sqrt{(4-3)^2 + (6-3)^2 + (1-4)^2} = 4.359$$

7) จากนั้นเรียงลำดับ Sample ที่มี Distance โดยเรียงจากน้อยไปหามาก ดังนี้ R3, R2, R4, R5

8) เลือก K_C Sample ไปคำนวณในขั้นตอนที่ 3 คือ R3, R2, R4

ขั้นตอนที่ 3 การ Estimate Missing Values

หลังจากที่ได้ข้อมูล Samples จำนวน K_C Samples ในขั้นตอนที่สองแล้วจากนั้น นำ Samples ที่ได้มาหาค่าเฉลี่ยเพื่อนำไป Impute Missing Values ดังตัวอย่าง A11= (R13+R12+R14)/3= 2.33

4. ผลการทดลอง

ในการคำนวณหาประสิทธิภาพของการแทนค่าข้อมูลที่ขาดหายจะคำนวณ Normalized Root Mean Squared Error (NRMSE) ซึ่งเป็นค่า Root Mean Squared Error ระหว่าง Imputed Matrix และ Original Matrix [12] ดังสมการ

$$NRMSE = \sqrt{mean[(y_{guess} - y_{ans})^2]} / std[y_{ans}] \quad (2)$$

โดยที่

y_{quess} คือ ค่าที่ได้จากการ Impute

y_{ans} คือ ค่าการแสดงออกของยีนต้นฉบับ

$std[y_{ans}]$ คือ ค่าส่วนเบี่ยงเบนมาตรฐานของยีนต้นฉบับ

ในการทดลองนี้ผู้วิจัยได้ทำการศึกษาข้อมูลไมโครอาร์เรย์ซึ่งเป็นข้อมูลที่สมบูรณ์ไม่มี Missing Values โดยได้ทำการคัดลอกข้อมูลจากเว็บไซต์ http://sdmc.lit.org.sg/GEDatasets จำนวน 3 Dataset คือ

Colon Tumor รวบรวมจากกลุ่มตัวอย่าง (ผู้ป่วย) จำนวน 62 คน ประกอบด้วยผู้ที่ป่วยเป็นโรคมะเร็งจำนวน 40 คน และผู้ที่ไม่ป่วยเป็นโรคมะเร็งจำนวน 22 คน โดยทำการเลือกข้อมูลการแสดงออกของยีนมา 2,000 ยีน

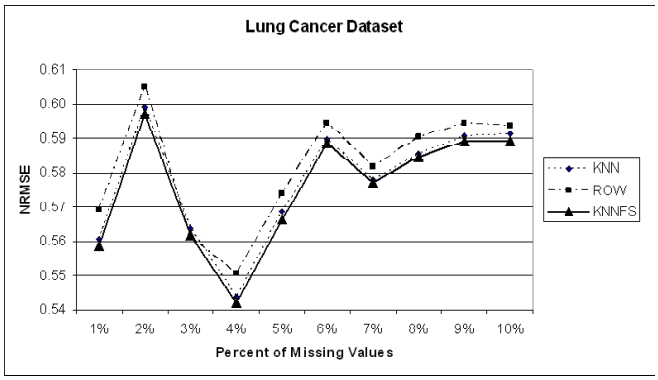
Lung Cancer ข้อมูลทั้งหมดมีจำนวน 181 กลุ่มตัวอย่าง โดยจำแนกออกเป็น 2 Class คือ Malignant Pleural Mesothelioma (MPM) และ Adenocarcinoma (ADCA) แบ่งเป็น 31 MPM และ 150 ADCA ในการทดลองจะใช้ข้อมูลสำหรับการสอน 32 ตัวอย่าง คือ 16 MPM 16 ADCA ซึ่งข้อมูลประกอบด้วยยีน (Attribute) 12,533 ยีน

ALL-AML Leukemia เป็นข้อมูลไขกระดูกของผู้ป่วย จำนวน 38 ตัวอย่าง จำแนกออกเป็น 2 Class คือ 27 ALL และ 11 AML ในการทดลองจะใช้ข้อมูลสำหรับการทดสอบ 34 ตัวอย่าง ประกอบด้วย 20 ALL และ 14 AML ซึ่งข้อมูลประกอบด้วยยีน 7129 ยีน

จากการทดลองการแทนค่าข้อมูลที่ขาดหายด้วยเทคนิค KNNFS Impute โดยทำการสุ่มสร้าง Missing Values ตั้งแต่ 1%-10% กับทั้ง 3 Dataset ผลการทดลองพบว่า การใช้เทคนิค KNNFS Impute ในการประมาณค่า Missing Values ให้ประสิทธิภาพที่ดีกว่า KNN ปกติ และ Row Average แสดงให้เห็นจากการวัดประสิทธิภาพด้วยค่า NRMSE ของ Missing values ตั้งแต่ 1%-10% ในการแทนค่าข้อมูลที่ขาดหายด้วยวิธี KNN, Row average, KNNFS Impute ในข้อมูล Lung Cancer Dataset ดังแสดงในตารางที่ 3 และภาพที่ 2 ALL-AML Leukemia Dataset ดังแสดงในตารางที่ 4 และภาพที่ 3 และ Colon Tumor Dataset ดังแสดงในตารางที่ 5 และภาพที่ 4 ดังต่อไปนี้

ตารางที่ 3 แสดงค่า NRMSE ของ Missing values ตั้งแต่ 1%-10% ในการ Estimate Missing Values ด้วยวิธี KNN, Row average, KNNFS Impute ในข้อมูล Lung Cancer Dataset

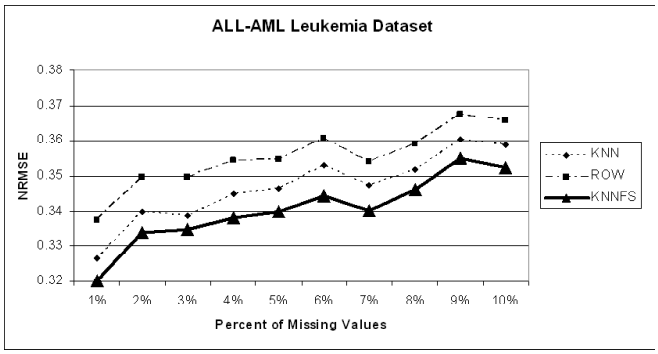
Lung Cancer Dataset										
Missing	**1%**	**2%**	**3%**	**4%**	**5%**	**6%**	**7%**	**8%**	**9%**	**10%**
KNN	0.56	0.60	0.56	0.54	0.57	0.59	0.58	0.59	0.59	0.59
ROW	0.57	0.61	0.56	0.55	0.57	0.59	0.58	0.59	0.59	0.59
KNNFS	0.56	0.60	0.56	0.54	0.57	0.59	0.58	0.58	0.59	0.59



ภาพที่ 2 แสดงค่า NRMSE ของ Missing values ตั้งแต่ 1%-10% ในการ Estimate Missing Values ด้วยวิธี KNN, Row average, KNNFS Impute ในข้อมูล Lung Cancer Dataset

ตารางที่ 4 แสดงค่า NRMSE ของ Missing values ตั้งแต่ 1%-10% ในการ Estimate Missing Values ด้วยวิธี KNN, Row average, KNNFS Impute ในข้อมูล ALL-AML Leukemia Dataset

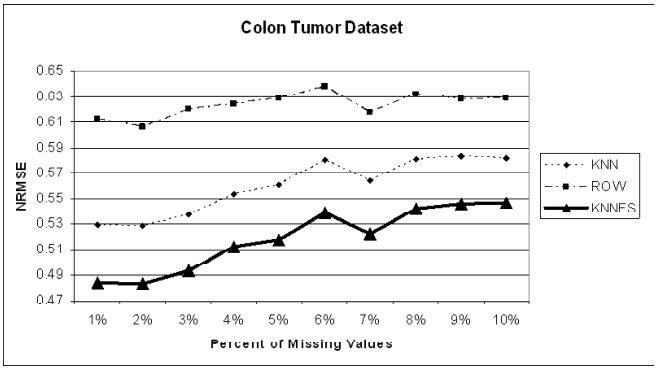
ALL-AML Leukemia Dataset										
Missing	**1%**	**2%**	**3%**	**4%**	**5%**	**6%**	**7%**	**8%**	**9%**	**10%**
KNN	0.33	0.34	0.34	0.34	0.35	0.35	0.35	0.35	0.36	0.36
ROW	0.34	0.35	0.35	0.35	0.35	0.36	0.35	0.36	0.37	0.37
KNNFS	0.32	0.33	0.33	0.34	0.34	0.34	0.34	0.35	0.36	0.35



ภาพที่ 3 แสดงค่า NRMSE ของ Missing values ตั้งแต่ 1%-10% ในการ Estimate Missing Values ด้วยวิธี KNN, Row average, KNNFS Impute ในข้อมูล ALL-AML Leukemia Dataset

ตารางที่ 5 แสดงค่า NRMSE ของ Missing values ตั้งแต่ 1%-10% ในการ Estimate Missing Values ด้วยวิธี KNN, Row average, KNNFS Impute ในข้อมูล Colon Tumor Dataset

Colon Tumor Dataset										
Missing	**1%**	**2%**	**3%**	**4%**	**5%**	**6%**	**7%**	**8%**	**9%**	**10%**
KNN	0.53	0.53	0.54	0.55	0.56	0.58	0.56	0.58	0.58	0.58
ROW	0.61	0.61	0.62	0.62	0.63	0.64	0.62	0.63	0.63	0.63
KNNFS	0.48	0.48	0.49	0.51	0.52	0.54	0.52	0.54	0.55	0.55



ภาพที่ 4 แสดงค่า NRMSE ของ Missing values ตั้งแต่ 1%-10% ในการ Estimate Missing Values ด้วยวิธี KNN, Row average, KNNFS Impute ในข้อมูล Colon Tumor Dataset

5. สรุปผลและอภิปรายผล

ผลการทดลองการแทนค่าข้อมูลที่ขาดหายด้วยเทคนิค KNNFS Impute โดยทำการสุ่มสร้าง Missing Values ตั้งแต่ 1%-10% พบว่า เทคนิค KNNFS Impute สามารถให้ประสิทธิภาพในการประมาณค่า Missing Values ที่ดีกว่า KNN ปกติ และ Row Average ด้วยค่า NRMSE ที่น้อยที่สุดในทั้ง 3 Dataset คือ Lung Cancer Dataset, ALL-AML Leukemia Dataset และ Colon Tumor Dataset

สาเหตุที่วิธีการ Row average ให้ประสิทธิภาพในการ Estimate Missing Values ที่ต่ำนั้น เนื่องมาจากการนำข้อมูลที่ไม่ได้ทำการหาความใกล้เคียงของข้อมูลก่อนทำการคำนวณค่าเฉลี่ยซึ่งทำให้ผลที่ได้จากการ Estimate อาจถูกโน้มเอียงด้วย Outlier หรือข้อมูลที่ไม่สัมพันธ์กัน ส่วนวิธีการแบบ KNN ปกติที่ให้ประสิทธิภาพรองลงมา ซึ่งในการ Estimate Missing Values ด้วย KNN นั้น จะทำการคำนวณหาตัวอย่าง (Case หรือ Row) ที่ใกล้เคียงกับ Missing Values มากที่สุดโดยคำนวณจาก Feature(Column) ทั้งหมดแต่เนื่องจาก Feature บาง Feature อาจไม่มีความสัมพันธ์กัน (ไม่มีความใกล้เคียงกัน) กับข้อมูลที่ต้องการ Impute จริง จึงทำให้ประสิทธิภาพของการ Estimate ต่ำ แต่วิธีการ KNNFS Impute เป็นวิธีการที่ทำการเลือกคุณลักษณะ (Feature) ของข้อมูลทำให้ได้คุณลักษณะที่สัมพันธ์กับข้อมูลที่ต้องการ Impute เพื่อเป็นตัวแทนในการนำไปคัดเลือกตัวอย่างที่เหมาะสมสำหรับ Impute Missing values ของข้อมูลแต่ละ Dataset จริงจึงทำให้ได้ประสิทธิภาพที่ดีที่สุด

เนื่องจากในงานวิจัยนี้ได้ทำการศึกษากับเทคนิค K-Nearest Neighbor เท่านั้น ซึ่งเป็นเทคนิคที่ให้ประสิทธิภาพยังไม่ดีมากนัก ดังนั้นงานวิจัยที่จะทำการศึกษาในอนาคตคือจะนำเอาเทคนิคนี้ไปประยุกต์ใช้กับเทคนิคที่มีประสิทธิภาพมากขึ้น

6. เอกสารอ้างอิง

[1] Brown M.P.S. Grundy W.N. Lin D. Cristianini N. Sugnet CW, Furey TS, Ares M Jr, Haussler D. "**Knowledge-based analysis of microarray gene expression data by using support vector machines**". Proc Natl Acad Sci USA 2000, 97:262-267.

[2] Ji X.L. Ling J.L. Sun Z.R. "**Mining gene expression data using a novel approach based on hidden Markov models**", FEBS Letters 2003, 542:125-131.

[3] Alter O, Brown PO,Botstein D. "**Singular Value decomposition for genome-wide expression data processing and modeling**", Proc Natl Acad Sci USA 2000, 97:10101-10106.

[4] Eisen MB, Spellman PT, Brown PO, Botstein D. "**Cluster analysis and display of genome-wide expression patterns**", Proc Natl Acad Sci USA 1998, 97:262-267.

[5] Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR, "**Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation**", Proc Natl Acad Sci USA 1999, 96:2907-2912.

[6] Wit E. and McClure J, **Statistics for Microarrays:**

Design, Analysis and Inference. John Wiley and Sons Ltd, West Sussex, England, 2004, pp.65-69.

[7] Sehgal MS, Gondal L, Dooley LS, "**Collateral Missing value imputation: a new robust missing value estimation algorithm for microarray data**", Bioinformatics 2005, 21:2417-2423.

[8] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Staudt LM, et al, "**Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling**", Nature 2000, 403:503-511.

[9] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB, "**Missing values estimation methods for DNA microarrays**", Bioinformatics 2001, 17:520-525.

[10] Oba S, Sato MA, Takemasa I, Monden M, Matsubara KI, Ishii S, "**A Bayesian missing value estimation method for gene expression profile data**", Bioinformatics 2003, 19:2088-2096.

[11] Xhou XB, Wang XD, Dougherty ER, "**Missing - value estimation using linear and non-linear regression with Bayesian gene selection**", Bioinformatics, 2003, 19:2302-2307.

[12] Kim H, Golub GH, Park H, "**Missing value estimation for DNA microarray gene expression data: local least squares imputation**", Bioinformatics, 2005, 21:187-198.

[13] Yoon D, Lee EK, Park T, "**Robust imputation method for missing values in microarray data**", BMC Bioinformatics, 2007, 8(2):S6.