การจัดการข้อมูลสูญหาย: วิธีเคเนียร์เรสเนเบอร์
Management Approach of Missing Data: K-Nearest Neighbor Imputation

*พัชนา สุวรรณแสน**
Patchana Suwannasaen
March 18, 2018 Revised: August 1, 2018 Accepted: September 22, 2018

บทคัดย่อ

ข้อมูลสูญหาย เป็นหนึ่งในปัญหาสำคัญของการวิจัยซึ่งในงานวิจัยหลายประเภทจำเป็นอย่างยิ่งที่ต้องการข้อมูลที่สมบูรณ์ก่อนลงมือดำเนินการในขั้นตอนการวิเคราะห์ข้อมูล เนื่องจากการที่ข้อมูลเกิดการสูญหายส่งผลกระทบถึงประสิทธิภาพและความคลาดเคลื่อนในการวิเคราะห์ข้อมูลทำให้งานวิจัยเหล่านั้นไม่น่าเชื่อถือและได้ผลการวิเคราะห์ที่คลาดเคลื่อนจากความเป็นจริง ดังนั้นนักวิจัยจึงจำเป็นต้องมีแนวทางในการจัดการเพื่อลดการเกิดข้อมูลสูญหายและต้องมีความรู้เกี่ยวกับการจัดการกับปัญหาข้อมูลสูญหายเพื่อเลือกใช้วิธีการจัดการกับข้อมูลสูญหายอย่างเหมาะสม บทความนี้ผู้เขียนได้ทบทวนงานวิจัยและวิธีการจัดการกับปัญหาข้อมูลสูญหาย โดยได้นำเสนอถึง ความหมาย ประเภทและวิธีการจัดการข้อมูลสูญหาย โดยเฉพาะการแทนค่าสูญหายวิธีเคเนียร์เรสเนเบอร์ ซึ่งจะสามารถเป็นแนวทางในการทำให้ได้ข้อมูลที่มีความสมบูรณ์เหมาะสมในการดำเนินงานในขั้นตอนต่อไปของการวิจัย
คำสำคัญ: ข้อมูลสูญหาย, การแทนค่าสูญหาย, ความคลาดเคลื่อน

Abstract

Missing data is one of crucial problems in research. To get a complete set of data before analyzing process is needed in many kinds of research. Due to the missing data imputation, the efficiency and error in data analysis make the research unreliable and inaccurate. Therefore, researchers need to have a management approach to reduce data missing and must have knowledge base on dealing

*สาขาวิชาสถิติประยุกต์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครราชสีมา จ.นครราชสีมา 30000
E-mail: patchana.s@nrru.ac.th

with data missing problems. This article reviews the literatures about data missing and the method how to deal with missing values problems. It presents the meaning, type and method of managing missing data especially in K Nearest Neighbor imputation. This review can be a guideline for getting the complete information to work in the next phase of research.
Keywords: Missing data, Missing data Imputation, Error

บทนำ

ข้อมูลสูญหาย (Missing Data) หมายถึง ข้อมูลที่มีค่าสังเกตที่ไม่มีการระบุค่าหรือค่าว่าง ซึ่งอาจเกิดจากการที่ผู้ให้ข้อมูลไม่ประสงค์จะตอบหรือข้อมูลที่มีการเก็บรวบรวมไว้แล้วแต่เป็นข้อมูลที่ไม่สมบูรณ์ เช่น เกิดการขาดหายของข้อมูล หรือไม่มีข้อมูลในส่วนนั้น โดยที่ค่านั้นสามารถทราบค่าได้ หรือเกิดจากการไม่เข้ากันของข้อมูลทำให้ข้อมูลส่วนนั้นขาดหายไป ข้อมูลสูญหายเป็นกรณีที่พบได้ในงานวิจัยทุกสาขาซึ่งนักวิจัยต้องพิจารณาถึงความเหมาะสมในการจัดการข้อมูลสูญหาย จากการศึกษาของAngela M Wood, Ian R White and Simon G Thompson (2004) พบว่าผลงานวิจัยที่ได้รับการตีพิมพ์ในวารสาร BMJ, JAMA, Lancet and New England Journal of Medicineจำนวน 71 ชิ้น มีงานวิจัยถึงร้อยละ 89 ที่มีปัญหาเรื่องข้อมูลสูญหายและมีเพียง 21% เท่านั้นที่มีการจัดการกับปัญหาข้อมูลที่ไม่สมบูรณ์ โดยระดับความรุนแรงของผลกระทบนี้ขึ้นอยู่กับองค์ประกอบจากหลายส่วน ในบทความนี้จะได้กล่าวถึงข้อมูลสูญหาย ประเภทของข้อมูลสูญหาย วิธีการจัดการข้อมูลสูญหาย วิธีเคเนียร์เรสเนเบอร์ในการจัดการข้อมูลสูญหายและบทสรุป

ข้อมูลสูญหาย

ประเภทของข้อมูลสูญหาย (Type of Missing Data)

การพิจารณาประเภทของข้อมูลสูญหายเป็นขั้นตอนที่สำคัญ เนื่องจากหากสามารถทราบถึงลักษณะของข้อมูลสูญหายจะช่วยในการพิจารณาแนวทางสำหรับจัดการกับปัญหาความไม่สมบูรณ์หรือการสูญหายของข้อมูลได้อย่างเหมาะสม ซึ่งโดยทั่วไป สามารถจำแนกข้อมูลสูญหายออกเป็น 3 ประเภท (Little and Rubin, 2002) ดังนี้

1. การสูญหายแบบสุ่มอย่างสมบูรณ์ หรือ Missing Completely at Random (MCAR) เป็นลักษณะของข้อมูลสูญหาย ที่เกิดขึ้นอย่างสุ่มจากค่าสังเกตทั้งหมด นั่นคือข้อมูลที่สูญหายเป็นอิสระจากตัวแปรต่างๆ สามารถทำการตรวจสอบลักษณะของข้อมูลสูญหายกลุ่มนี้โดยการแบ่งกลุ่มของค่าสังเกตเป็นกลุ่มข้อมูลปกติและข้อมูลสูญหาย ในกรณีนี้เมื่อทำการทดสอบจะไม่พบความแตกต่างอย่างมีนัยสำคัญระหว่างทั้งสองกลุ่มสำหรับตัวแปรต่าง ๆ ในฐานข้อมูลสำหรับสาเหตุที่ทำให้ข้อมูลเกิดการสูญหายประเภทนี้มีอยู่หลากหลายเหตุผล ซึ่งอาจเกิดจากเครื่องมือเสีย อุปกรณ์เกิดข้อบกพร่อง สภาพอากาศเลวร้าย กลุ่มเป้าหมายที่ศึกษาล้มป่วย หรือการนำเข้าข้อมูลไม่ถูกต้อง

2. การสูญหายแบบสุ่ม หรือ Missing at Random (MAR)เป็นลักษณะของข้อมูลสูญหายซึ่งไม่ได้เกิดขึ้นอย่างสุ่มจากค่าสังเกตทั้งหมด แต่เกิดขึ้นอย่างสุ่มภายในบางส่วนหรือบางกลุ่มของค่าสังเกต นั่นคือค่าของข้อมูลสูญหายขึ้นอยู่กับตัวแปรตัวอื่น ๆ ในฐานข้อมูลซึ่งไม่ได้เป็นตัวแปรที่เกิดข้อมูลสูญหาย ยกตัวอย่าง เช่น หากพบว่าเฉพาะกลุ่มผู้มีรายได้มากที่ไม่ให้ความร่วมมือในการตอบข้อคำถามเกี่ยวกับทัศนคติในการใช้จ่าย ในลักษณะนี้สามารถกล่าวได้ว่าข้อมูลทัศนคติในการใช้จ่ายมีค่าสูญหายแบบ MAR ทั้งนี้เนื่องจากเป็นค่าสูญหายที่เกิดขึ้นเฉพาะในบางส่วนของตัวแปรระดับรายได้

3. การสูญหายแบบไม่สุ่ม หรือ Not Missing at Random (NMAR)เป็นลักษณะของข้อมูลสูญหายซึ่งไม่ได้เกิดขึ้นอย่างสุ่ม โดยค่าของข้อมูลสูญหายขึ้นอยู่กับค่าของข้อมูลสมบูรณ์ในตัวแปรเดียวกัน รวมถึงตัวแปรตัวอื่นด้วย เช่น หากข้อมูลสูญหายของระดับรายได้ขึ้นอยู่ระดับรายได้ในแต่ละช่วงอายุ ข้อมูลสูญหายที่เกิดขึ้นจัดอยู่ในประเภท NMAR หรือในบางกรณีค่าของข้อมูลสูญหายอาจไม่ขึ้นอยู่

กับตัวแปรใด ๆ ในฐานข้อมูลเลย แต่ขึ้นอยู่กับตัวแปรอื่นที่ไม่ได้ถูกเก็บรวบรวมไว้ในการศึกษาครั้งนั้น เช่น ค่าน้ำหนักตัวที่ลดลงขึ้นอยู่กับน้ำหนักตัวตอนเริ่มต้น แต่เนื่องจากตัวแปรน้ำหนักตอนเริ่มต้นไม่ได้ถูกรวบรวมไว้ในฐานข้อมูล ดังนั้นค่าสูญหายของน้ำหนักตัวที่ลดลงจึงขึ้นอยู่กับตัวแปรภายนอกฐานข้อมูล

การจัดการข้อมูลสูญหาย

Rubin และ Little (Malarvizhi R., 2012) ได้กล่าวถึงการแบ่งวิธีการจัดการ ข้อมูลสูญหายของ Roth (1994) เป็น 3 ประเภท คือ

1. การตัดกลุ่มข้อมูลที่สูญหายทิ้งไป (Ignoring and Discarding Data) ซึ่งเป็นวิธีที่ง่ายและสะดวกที่สุดและพิจารณาเฉพาะข้อมูลที่มีความสมบูรณ์เท่านั้น โดยวิธีการนี้แบ่งออกเป็นสองแบบ คือ

1.1 วิธี Listwise Deletion ซึ่งเป็นวิธีการตัดแถวที่มีค่าข้อมูลสูญหายทิ้งทั้งหมดซึ่งเป็นวิธีที่ง่ายที่สุดเพื่อให้ได้มาซึ่งกลุ่มข้อมูลที่สมบูรณ์ที่สุด เป็นวิธีการจัดการกับข้อมูลสูญหายที่ง่าย นั่นคือไม่สนใจข้อมูลสูญหายที่เกิดขึ้น โดยจะทำการวิเคราะห์ข้อมูลจากข้อมูลเฉพาะส่วนที่สมบูรณ์ แนวทางนี้จะมีความเหมาะสมในกรณีที่ข้อมูลสูญหายมีจำนวนน้อยมาก และ/ หรือผลจากการวิเคราะห์มีความชัดเจนมาก ซึ่งวิธีการนี้มักถูกกำหนดให้ใช้เป็นหลักสำหรับจัดการกับข้อมูลสูญหายในโปรแกรมคอมพิวเตอร์ทางสถิติทั่ว ๆ ไป หากไม่เจาะจงเลือกใช้วิธีการอื่นในการจัดการกับข้อมูลสูญหาย ข้อเสียของวิธีนี้คือเมื่อตัดข้อมูลที่มีการสูญหายออกไปทำให้ขนาดตัวอย่างลดลง

1.2 วิธี Pairwise Deletion เป็นวิธีการจัดการกับข้อมูลสูญหายสำหรับกรณีที่ทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรคู่ โดยจะทำการวิเคราะห์ข้อมูลจากข้อมูลส่วนที่มีค่าสมบูรณ์ทั้งสองตัวแปร วิธีนี้มีข้อดีในส่วนของการใช้งานข้อมูลอย่างเต็มที่และมีประสิทธิภาพ แต่มีความซับซ้อนกว่าวิธี Listwiseและเสียเวลามากกว่า

2. การประมาณค่าพารามิเตอร์ (Parameter estimation) ตัวอย่างวิธีการประมาณค่า ได้แก่ การประมาณค่าความเป็นไปได้มากที่สุด (Maximum Likelihood Estimation) ซึ่งคำนวณจากข้อมูลที่ทราบทั้งหมดค่าความเป็นไปได้มากที่สุดจะถูกคำนวณโดยแยกคำนวณจากข้อมูลที่สมบูรณ์ของบางตัวแปรและคำนวณจากข้อมูลในทุกตัวแปร ค่าความเป็นไปได้ที่มากที่สุดจากการคำนวณทั้งสอง จะถูกใช้ในการประมาณค่าข้อมูลสูญหายอีกทีหนึ่ง หรือวิธี Expectation maximization (EM) เป็นวิธีที่อาศัยตัวแบบ (Model-base procedures) และใช้หลักการของ Likelihood เพื่อประมาณค่าพารามิเตอร์ ข้อเสียของวิธีเหล่านี้คือมีความซับซ้อนเข้าใจยากในการดำเนินการ

3. การแทนที่ค่าข้อมูลสูญหายด้วยค่าประมาณจากการคำนวณ (Imputation) วิธีการนี้จะทำการประมาณค่าข้อมูลโดยอาศัยความสัมพันธ์ระหว่างกลุ่มข้อมูล เพื่อนำไปแทนค่าข้อมูลสูญหายและเป็นวิธีการที่มีการนิยมใช้ในงานวิจัยหลายแขนง ซึ่งมีวิธีการ และมีนักวิจัยอยู่หลายกลุ่มได้ทำการศึกษาค้นคว้าทดลองโดยตัวอย่างของวิธีการแทนค่าที่เป็นที่นิยม มีดังนี้

3.1 Mean Imputation เป็นวิธีการประมาณค่าข้อมูลสูญหายด้วยค่าเฉลี่ยของข้อมูลที่ทราบค่าในแต่ละตัวแปร ถูกนำเสนอครั้งแรกโดย Wilks (1932) และได้รับการพัฒนาโดยใช้ค่าเฉลี่ยตัวอย่างทดแทนในค่าที่สูญหาย (Little & Robin, 2002) วิธีนี้เป็นวิธีที่ง่ายต่อการทำงาน แต่วิธีนี้อาจถูกโน้มเอียงจากค่าที่อยู่นอกกลุ่มได้จึงให้ประสิทธิภาพในการประมาณค่าไม่ดีนัก (ไก้รุ่ง เฮงพระพรหม, 2551)

3.2 Regression Method เป็นวิธีที่ดำเนินการสร้างสมการถดถอยเพื่อทำนายข้อมูลสูญหายจากข้อมูลที่สมบูรณ์ โดยกำหนดให้ตัวแปรตามเป็นตัวแปรที่มีข้อมูลไม่สมบูรณ์ จากนั้นใช้สมการถดถอยที่ได้ทำการประมาณค่าของข้อมูลที่ไม่สมบูรณ์

3.3 Hot Deck Imputation เป็นวิธีการพิจารณาเลือกหน่วยตัวอย่างที่มีลักษณะคล้ายคลึงกันมากที่สุดกับหน่วยตัวอย่างที่เกิดค่าสูญหาย จากนั้นแทนค่าที่สูญหายด้วยค่าของหน่วยตัวอย่างที่คล้ายคลึงนั้น ข้อดีของวิธีนี้คือมีการวัดความคล้ายคลึงกันของข้อมูล

3.4 Multiple Imputation (MI) เป็นวิธีการที่ผสมผสานระหว่างวิธีการ EM และ Raw Maximum Likelihood Methods ร่วมกับความสามารถของคุณสมบัติ Hot Deck เพื่อทำการสร้างชุดจำลองของข้อมูลที่ได้ทำการประมาณค่าข้อมูลสูญหายด้วย Imputed Value แล้วขึ้นมาหลาย ๆ ชุด (ประมาณ 5 ถึง 10 ชุด) จากนั้นทำการวิเคราะห์ข้อมูลจากชุดต่าง ๆ บันทึกผลการวิเคราะห์ที่ได้ โดยผลการวิเคราะห์ที่ได้เหล่านี้จะถูกรวมเข้าด้วยกันเพื่อแทนที่ข้อมูล

3.5 โครงข่ายประสาทเทียม (Artificial Neural Networks หรือ ANN) เป็นการจำลองการทำงานของเซลล์ประสาทในสมองของมนุษย์ การจำลองการคำนวณด้วยคอมพิวเตอร์แทนให้อยู่ในรูปเพอร์เซ็ปตรอน(Ibrahim BerkanAydilek& Ahmet Arslan, 2012)

3.6 วิธีเคเนียร์เรสเนเบอร์ (K-Nearest Neighbors หรือ KNN) เป็นวิธีการประมาณค่าที่มีการใช้อย่างแพร่หลายและมีประสิทธิภาพในการประมาณค่าข้อมูล เป็นเทคนิควิธีการที่ได้รับความนิยมในการใช้งานอย่างมาก สาเหตุเนื่องมาจากเป็นวิธีการที่ง่ายและมีประสิทธิภาพซึ่งสามารถนำไปประยุกต์ใช้กับงานได้อย่างหลากหลายเช่น งานทางด้านการจำแนกข้อมูล (Classification) รวมถึงงานทางด้านการแทนที่ข้อมูลสูญหาย (Missing Values Imputation) จึงเป็นวิธีที่ผู้วิจัยมีความสนใจศึกษาโดยมีรายละเอียดที่จะแสดงในหัวข้อถัดไป

วิธีเคเนียร์เรสเนเบอร์ในการจัดการข้อมูลสูญหาย

วิธีเคเนียร์เรสเนเบอร์หรือ K-Nearest Neighbors: KNN (Troyanskaya et al., 2001) เป็นเทคนิควิธีการที่มีประสิทธิภาพซึ่งสามารถนำไปประยุกต์ใช้กับงานได้อย่างหลากหลายเช่น งานทางด้านการจำแนกข้อมูล (Classification) รวมถึงงานทางด้านการแทนที่ข้อมูลสูญหาย (Missing Values Imputation) มีขั้นตอนการดำเนินการของวิธี KNN สรุปได้ดังนี้

1. กำหนดค่า k

2. คำนวณหาระยะห่างระหว่างข้อมูลตัวอย่างที่สนใจกับข้อมูลอื่นๆ ทุกตัวด้วยวิธีการ Euclidian Distance

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

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

n คือ จำนวนข้อมูล

$x_{i,k}$ คือ คุณสมบัติทั้งหมดของตัวอย่าง

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

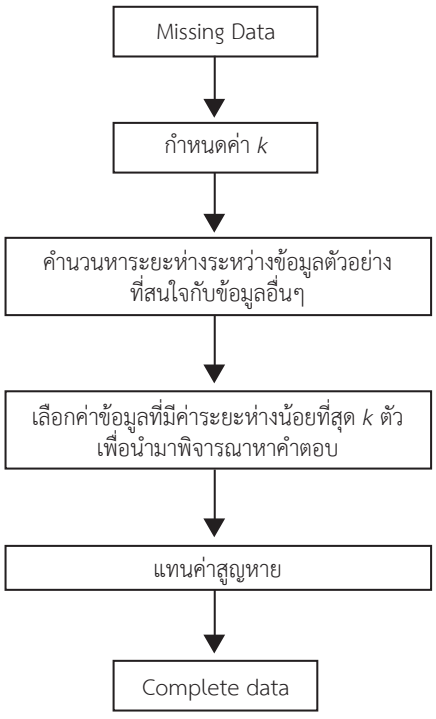
$x_{j,k}$ คือ คุณสมบัติทั้งหมดของตัวอย่าง

$x_{j,k}$ คือ คุณสมบัติตัวที่ k ของตัวอย่าง x_j

3. เรียงลำดับระยะห่างระหว่างจุด โดยพิจารณาจากข้อมูลที่ใกล้ที่สุด จำนวน k ตัวตามที่กำหนดและในการประมาณค่าข้อมูลที่มีการสูญหาย ทำการประมาณค่าโดยการหาค่าเฉลี่ยจากข้อมูลที่ใกล้ที่สุด k ตัวจากตัวแปรเดียวกันกับตำแหน่งที่มีการสูญหาย มีสูตรในการคำนวณดังนี้

$$\text{ค่าสูญหาย} = \hat{x}_i = \frac{\sum_{i=1}^{k} x_i}{k}$$

ขั้นตอนของการประมาณค่าสูญหายโดยวิธี KNN เป็นดังนี้



ภาพที่ 1ขั้นตอนของการประมาณค่าสูญหายโดยวิธี KNN

เพื่อให้มีความเข้าใจในกระบวนการทำงานของวิธี KNN จึงได้ยกตัวอย่าง การคำนวณด้วยวิธี K-Nearest Neighbor หรือ KNN โดยกำหนดให้ NA คือ ค่าสังเกตที่เป็นค่าสูญหาย จากตัวอย่างนี้อยู่ในตำแหน่ง R_1A_1 และ R_n คือ แถวตามแนวนอนที่ n โดยในตัวอย่างนี้มี 15 แถวซึ่งมีรายละเอียดของการคำนวณดังแสดงในตารางที่ 1และ 2

ตารางที่ 1 ตัวอย่างข้อมูลในการคำนวณ

	A1	A2	A3
R_1	NA	23.92	194.08
R_2	32.40	24.06	140.80
R_3	32.82	24.28	151.38
R_4	31.71	23.93	220.44
R_5	32.25	24.25	159.45
R_6	31.40	24.42	264.88
R_7	32.10	24.06	200.54
R_8	32.00	24.20	286.96
R_9	31.77	23.93	211.39
R_{10}	31.87	24.23	373.52
R_{11}	31.33	24.30	264.18
R_{12}	32.21	24.13	235.42
R_{13}	32.57	24.38	322.98
R_{14}	31.72	24.5	271.83
R_{15}	32.27	24.37	190.22

ขั้นตอนในการคำนวณ มีดังนี้

1. กำหนดค่า $k = 11$
2. คำนวณหาระยะห่างระหว่างข้อมูลตัวอย่างที่สนใจกับข้อมูลอื่น ๆ ทุกตัวด้วยวิธีการ Euclidian Distance

วิธีการคำนวณระยะห่าง		ลำดับ
$dist(R_1,R_2) = \sqrt{(23.92 - 24.06)^2 + (194.08 - 140.08)^2}$	$= 53.28$	8
$dist(R_1,R_2) = \sqrt{(23.92 - 24.28)^2 + (194.08 - 151.38)^2}$	$= 42.70$	7
$dist(R_1,R_2) = \sqrt{(23.92 - 23.93)^2 + (194.08 - 220.44)^2}$	$= 26.36$	4
$dist(R_1,R_2) = \sqrt{(23.92 - 24.25)^2 + (194.08 - 159.45)^2}$	$= 34.63$	5
$dist(R_1,R_2) = \sqrt{(23.92 - 24.42)^2 + (194.08 - 264.88)^2}$	$= 70.80$	10
$dist(R_1,R_2) = \sqrt{(23.92 - 24.06)^2 + (194.08 - 200.54)^2}$	$= 6.46$	2
$dist(R_1,R_2) = \sqrt{(23.92 - 24.20)^2 + (194.08 - 286.96)^2}$	$= 92.88$	12
$dist(R_1,R_2) = \sqrt{(23.92 - 23.93)^2 + (194.08 - 211.39)^2}$	$= 17.31$	3
$dist(R_1,R_2) = \sqrt{(23.92 - 24.23)^2 + (194.08 - 373.52)^2}$	$= 179.44$	14
$dist(R_1,R_2) = \sqrt{(23.92 - 24.30)^2 + (194.08 - 264.18)^2}$	$= 70.10$	9

วิธีการคำนวณระยะห่าง		ลำดับ
$dist(R_1,R_2) = \sqrt{(23.92 - 24.13)^2 + (194.08 - 235.42)^2}$	= 41.34	6
$dist(R_1,R_2) = \sqrt{(23.92 - 24.38)^2 + (194.08 - 322.98)^2}$	= 128.90	13
$dist(R_1,R_2) = \sqrt{(23.92 - 24.50)^2 + (194.08 - 271.83)^2}$	= 77.75	11
$dist(R_1,R_2) = \sqrt{(23.92 - 24.37)^2 + (194.08 - 190.22)^2}$	= 3.89	1

3. เรียงลำดับระยะห่างระหว่างจุด โดยพิจารณาจากข้อมูลที่ใกล้ที่สุด จำนวน 11 ตัวตามที่กำหนดและในการประมาณค่าข้อมูลที่มีการสูญหาย ทำการประมาณค่าโดยการหาค่าเฉลี่ยจากข้อมูลที่ใกล้ที่สุด 11 ตัวจากตัวแปรเดียวกันกับตำแหน่งที่มีการสูญหาย มีสูตรในการคำนวณดังนี้

$$\text{ค่าสูญหาย} = \hat{x}_i = \frac{\sum_{i=1}^{k} x_i}{k}$$

จะได้

ตารางที่ 2 การหาค่าระยะห่างเพื่อแทนค่าสูญหาย ด้วยวิธี KNN

	A1	ระยะห่าง	ลำดับ
R^{15}	32.27	3.89	1
R^{7}	32.10	6.46	2
R^{9}	31.77	17.31	3
R^{4}	31.71	26.36	4
R^{5}	32.25	34.63	5
R^{12}	32.21	41.34	6
R^{3}	32.82	42.70	7
R^{2}	32.40	53.28	8
R^{11}	31.33	70.10	9
R^{6}	31.40	70.80	10
R^{14}	31.72	77.75	11

ดังนั้น การประมาณค่าข้อมูลสูญหาย X_{11} ซึ่งอยู่ในตำแหน่ง R_1A_1

$$\hat{X}_{11} = \frac{(R_{15}A_1 + R_7A_1 + R_9A_1 + R_4A_1 + R_5A_1 + R_{12}A_1 + R_3A_1 + R_2A_1 + R_{11}A_1 + R_6A_1 + R_{14}A_1)}{k}$$

$$\hat{X}_{11} = \frac{(32.27+32.10+31.77+31.71+32.25+32.21+32.82+32.40+31.33+31.40+31.72)}{11}$$

$$\hat{X}_{11} = 31.998$$

ดังนั้น ค่าสังเกต X_{11} ที่สูญหาย (NA) ซึ่งอยู่ในตำแหน่ง R_1A_1 มีค่าทดแทนเท่ากับ 31.998 ในการคำนวณของตำแหน่งอื่น ๆดำเนินการในขั้นตอนการคำนวณในแบบเดียวกัน

การพิจารณาค่า k ที่เหมาะสมมีหลากหลายงานวิจัยที่เกี่ยวข้อง อาทิเช่น พิจารณาค่า k ที่ $k = \sqrt{m}$ โดย Dudaและ Hart (Jonsson and Wohlin, 2006) โดย k จะเป็นจำนวนคี่ที่มีค่าใกล้เคียงกับ $\sqrt{m}$ มากที่สุด เมื่อ m เป็นจำนวนข้อมูลที่สมบูรณ์ และ k =10 เมื่อข้อมูลเป็นข้อมูลขนาดใหญ่แบบไมโครอาร์เรย์ (Batista and Monard, 2001) จากงานวิจัยของ Troyanskayaet al. (2001) อธิบายว่าค่า k ควรมีช่วงระหว่าง 10 – 20 เนื่องจากถ้าค่า k มีค่ามากเกินไปจะทำให้ความถูกต้องลดลงและใช้เวลาในการคำนวณเพิ่มขึ้น

วิธี KNN มีการพัฒนาปรับปรุงวิธีการนี้ในหลายรูปแบบ ไม่ว่าจะเป็นการใช้วิธีเลือกคุณลักษณะ (Feature Selection) เข้ามาร่วมกับการประมาณค่าสูญหายด้วยวิธี KNN เช่น วิธี Sequential KNN หรือ SKNN โดย Ki-Yeol Kim (2004) และ KNN Feature Selection KNN หรือ KNNFS โดย Meesad and Hengpraprohm (2008) เป็นต้นนอกจากนี้มีการพัฒนาวิธีการ KNN โดยการประมาณค่าสูญหายตามลักษณะของตัวแปรเช่น Grey KNN ซึ่งพัฒนาโดย Huang and Lee (2004) วิธีประมาณค่าแบบนี้ใช้เมื่อข้อมูลมีลักษณะหลายตัวแปรและข้อมูลเป็นแบบผสม นอกจากนี้การพัฒนา วิธี KNN ยังได้มีการพัฒนาโดยการปรับวิธีการของค่าระยะทาง เช่น วิธีประมาณค่าแบบ GKNN หรือ Gray KNN ที่ใช้การหาค่าระยะทางโดยวิธี Gray Distance (Shichao, 2012) เป็นต้น

งานวิจัยที่เกี่ยวข้องกับการประมาณค่า สูญหายที่มีการนำวิธี KNN ไปใช้กับข้อมูลมีอยู่หลากหลายงาน เช่น Gustavo E.A.P.A Batista (2003) ได้ทดสอบประสิทธิภาพของการแทนที่ข้อมูลด้วย KNN กับ Mean และ Mode imputation ด้วยการจำลองข้อมูล 4 ชุดข้อมูลจากฐานข้อมูลกลาง คือ Bupa, Cmc Pima และ Breast Cancer โดยจำลองข้อมูลสูญหายแบบ MCAR และเมื่อตรวจสอบประสิทธิภาพของการประมาณค่าด้วยการจำแนกข้อมูลแบบ C4.5 และ CN2 พบว่าวิธี KNN ให้ผลในการประมาณค่าสูญหายที่ดีกว่า และยังพัฒนาวิธี KNN ให้มีประสิทธิภาพโดยการถ่วงด้วยน้ำหนัก กลายเป็น weight KNN อีกด้วย การพัฒนาของวิธี KNN ยังถูก Hassani et al. (2004) ศึกษาเปรียบเทียบ โดยใช้สัดส่วนข้อมูลสูญหายร้อยละ 20 เปรียบเทียบวิธีการแทนค่าสูญหาย 2 วิธี คือ NN Imputation และวิธีประมาณค่าสูญหายด้วยการวิเคราะห์ถดถอย พบว่าวิธีการประมาณค่าสูญหายด้วยวิธี NN Imputation มีประสิทธิภาพดีกว่า นอกจากนี้ Joseet al. (2010) ได้ศึกษาวิธีการประมาณค่าข้อมูลด้วยเทคนิคทางสถิติซึ่งประกอบด้วย วิธี Mean Hot-deck และ Multiple Imputation และเทคนิคทางเหมืองข้อมูลซึ่งประกอบด้วยวิธี Multilayer Perceptron (MLP) K-Nearest Neighbor (KNN) และ Self Organizing Maps (SOM) จากข้อมูลจริงของปัญหามะเร็งเต้านม พบว่าวิธีการประมาณค่าข้อมูลสูญหายโดยวิธีการ K-Nearest Neighbor (KNN) เป็นวิธีการที่เหมาะสมในการประมาณค่าข้อมูลสูญหายจากข้อมูลจริงประเภทนี้ จากงานวิจัยข้างต้นจะเห็นได้ว่าการประมาณค่าสูญหายด้วยวิธี KNN ให้ผลการประมาณค่าที่ดีกว่าวิธีอื่นและสามารถใช้ได้กับข้อมูลจริงได้อีกด้วย เมื่อทำการศึกษางานวิจัยในประเทศไทยที่ได้นำเสนอวิธีการประมาณค่าสูญหายด้วย วิธี KNN พบงานวิจัยของ อุมาภรณ์ สายแสงจันทร์ (2548) ซึ่งได้ศึกษาวิธีการประมาณค่าสูญหายของข้อมูลโดยอาศัยหลักการของวิธี KNN โดยนำเสนอวิธีการให้น้ำหนักแบบ Fuzzy-weight เมื่อทำการตรวจสอบประสิทธิภาพ ระหว่างวิธี KNN ที่ให้น้ำหนักแบบ Fuzzy-weight กับ วิธี KNN ที่ใช้Distance-weight พบว่าวิธี Fuzzy weight KNN มีความเหมาะสมกับข้อมูลที่มีระดับการสูญหายที่ 30% ขึ้นไป แต่ถ้าต่ำกว่านั้น ทั้งสองวิธีให้ผลที่ไม่แตกต่างกันนอกจากนี้ยังพบงานวิจัยของ นรุตม์ บุตรพลอย (2554) ที่ได้ประยุกต์ใช้ Soft Computing มาช่วยในการปรับค่า ของวิธีการประมาณค่าสูญหายแบบ K-Nearest Neighbor โดยเรียกวิธีนี้ว่า Soft K–Nearest Neighbor หรือ SKNN โดยทดลองกับ ชุดข้อมูลจำลองการสูญหาย 5 ระดับ พบว่า ค่า ที่เหมาะสมคือ 10 และวิธี SKNN ให้ผลลัพธ์ดีที่สุดที่พารามิเตอร์เท่ากับ 0.3 จะเห็นได้ว่ามีนักวิจัยที่ใช้วิธีการประมาณค่าสูญหายจากวิธี KNN เป็นพื้นฐานในการพัฒนาวิธีการประมาณค่าสูญหายแบบอื่นให้มีประสิทธิภาพในการประมาณค่ายิ่งขึ้น

บทสรุป

การจัดการเพื่อแก้ปัญหาข้อมูลสูญหายจำเป็นต้องศึกษาเกี่ยวกับวิธีการจัดการกับข้อมูลสูญหายอย่างเหมาะสมเพื่อจะได้ข้อมูลที่มีความสมบูรณ์พร้อมสำหรับกระบวนการในการวิเคราะห์ข้อมูลต่อไป การประมาณค่าข้อมูลสูญหายเป็นการจัดการข้อมูลสูญหายที่ได้รับความนิยมและวิธีเคเนียร์เรสเนเบอร์ (K-Nearest Neighbor หรือ KNN) ก็เป็นวิธีที่น่าสนใจในการศึกษาเพิ่มเติม ซึ่งผู้วิจัยมีความสนใจที่จะพัฒนาวิธีการประมาณค่าข้อมูลสูญหายวิธีเคเนียร์เรสเนเบอร์ในการเพิ่มประสิทธิภาพในการประมาณค่าข้อมูลสูญหายในงานวิจัยครั้งต่อไป เพื่อประโยชน์ในการประมาณค่าข้อมูลสูญหายให้มีความสมบูรณ์ก่อนการนำข้อมูลนั้นไปวิเคราะห์ในงานด้านต่างๆต่อไป

เอกสารอ้างอิง

ไก้รุ่ง เฮงพระพรหม. (2551). การเลือกคุณลักษณะด้วยวิธีสมาชิกที่ใกล้ที่สุดสำหรับการแทนที่ค่าข้อมูล ที่ขาดหายด้วยวิธีการที่ใกล้ที่สุดแบบให้น้ำ หนักของข้อมูลไมโครอาร์เรย์. เอกสารประกอบการประชุมวิชาการระดับชาติด้าน คอมพิวเตอร์และเทคโนโลยีสารสนเทศครั้งที่ 6.

นรุตม์ บุตรพลอย. (2553). การประยุกต์ Soft Computing และ k-Nearest Neighbor เพื่อใช้ประมาณค่าสูญหายของข้อมูล. เอกสารประกอบการประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศครั้งที่ 8.

อุมาภรณ์ สายแสงจันทร์. (2548). การประมาณค่าของข้อมูลที่สูญหาย โดยใช้ตัวแบบฟัซซี่เนียเรส มหาวิทยาลัยขอนแก่น.

Batista, G. E. A. P. A., and Monard, M. C. (2001). A Study of K-Nearest Neighbour as a Model-Based Method to Treat Missing Data. The Argentine Symposium on Artificial Intelligence (ASAI'03), Buenos Aires, Argentine, 30, 1–9.

Batista, G.E.A.P.A., and Monard, M.C. (2003). Experimental Comparison of k-Nearest Neighbour and Mean or Mode imputation Methods with the Internal Strategies used by C4.5 and CN2 to Treat Missing Data. ICMC-USP. 186.

Hassani, B.T., V. Le May, P.L. Marshall, H. Temesgen, and A. Zumrawi. (2004). Regeneration imputation models for complex stands of southeastern British Columbia. The Forestry Chronicle. 80 (2): 271-278.

Huang, Ch., and Lee, H. (2004). A Grey-Base Nearest Neighbor Approach for Missing Attribute Value Prediction, Applied Intelligence, 20, 239-252.

Ibrahim, B.A., and Ahmet, A. (2012).A Novel Hybrid Approach to Estimating Missing Values in Database Using K-Nearest Neighbors and Neural Networks. International Journal of Innovative Computing Information and Control. 8(7A), 4705-4717.

Jönsson, P., and Wohlin, C. (2006). Benchmarking k-Nearest Neighbour Imputation with Homogeneous Likert Data, Empirical Software Engineering: An International Journal, 11(3), 463-489.

Jose, M.J., Molina, I., Pedro, J., and Leonardo F. (2010). Missing data Imputation Using Statistical and Machine Learning Methods in a Real Breast Cancer Problem. Artificial Intelligence in Medicine, 50, 105-115.

Ki-Yeon Kim, Byoung- Jin Kim, and Gwan-Su Yi. (2004). Reuse of imputed data in microarray analysis increase imputation efficiency. BMC bioinformatics, 5(1): 160.
Little, R. J. A.,and Robin, D. B. (2002). Statistical analysis with missing data. Hoboken, N.J: John Wiley & Sons.
Malarvizhi, R., and Thanamani, A. S. (2012). K - Nearest Neighbor in Missing Data Imputation. International Journal of Engineering Research and Development, 5(1), 5–7.
Meesad P. and Hengpraprohm K., (2008). Combination of KNN Based Feature Selection and KNN Based Missing Value Imputation of Microarray Data, Presented at 3rd International Conference on Innovative Computing Information and Control, 341.
Shichao Zhang. (2012). Nearest Neighbor Selection for Iteratively KNN Imputation, Journal of Systems and Software, 85(11), 2541-2552.
Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., Botstein D., and Altman R. B. (2001). Missing values estimation methods for DNA microarrays. Bioinformatics, 17(6), 520-525.
Wilks, S.S. (1932). Moments and Distributions of Estimates of Population Parameters from Fragmentary Samples, Annals of Mathematical Statistics, 3(2), 163–195.
Wood, A. M., White I. R., & Thompson S. G. (2004). Are Missing Outcome Data Adequately Handled? A Review of Published Randomized Controlled Trials in Major Medical Journals. Clinical Trial, 13, 68-76.
Roth, P. L. (1994). Missing Data: A Conceptual Review for Applied Psychologists. Personnel Psychology, 47(3), 537-560.