การปรับแก้วิธีเคเนียร์เรสเนเบอร์โดยใช้ค่าเฉลี่ยเดไซล์ในการประมาณค่าข้อมูลสูญหาย

Adjusted K Nearest Neighbor Method Based on Decile Mean in Missing Data Imputation

พัชนา สุวรรณแสน[1*] ภัทราวดี มากมี[1] และ อาฟีฟี ลาเต๊ะ[2]

Patchana Suwannasaen[1*], Pattrawadee Makmee[1] and Afifi Lateh[2]

Received: 19 November 2019, Revised: 3 February 2020, Accepted: 13 February 2020

บทคัดย่อ

การวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ Decile Mean K-Nearest Neighbor Bhattacharyya Imputation (DKNN-BH) ที่เกิดจากวิธีการประมาณค่าสูญหายด้วยวิธี K-Nearest Neighbor Imputation (KNN) ปรับแก้ด้วยการใช้ค่าเฉลี่ยเดไซล์และการหาระยะทางแบบ Bhattacharyya เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH กับวิธีการประมาณค่าข้อมูลสูญหายค่าเฉลี่ยเลขคณิต วิธีการประมาณค่าข้อมูลสูญหาย K-Nearest Neighbor Imputation (KNN) และวิธีการประมาณค่าข้อมูลสูญหาย Decile Mean K-Nearest Neighbor Imputation (DKNN) ภายใต้การจำลองสถานการณ์ 300 สถานการณ์ จาก 4 เงื่อนไข คือ ขนาดตัวอย่าง ระดับการสูญหายของข้อมูล ขนาดของค่านอกเกณฑ์ และค่าคงที่ k สำหรับการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH วิธีการประมาณค่าข้อมูลสูญหาย KNN และวิธีการประมาณค่าข้อมูลสูญหาย DKNN โดยใช้การจำลองสถานการณ์ วิธีมอนติคาร์โล ทำการทดลองซ้ำจำนวน 500 ครั้ง ผลการศึกษา ปรากฏว่า การพัฒนาวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH เกิดจากการปรับแก้วิธีการประมาณค่าสูญหายวิธี KNN ด้วยการใช้ค่าเฉลี่ยเดไซล์และการหาระยะทางแบบ Bhattacharyya ซึ่งเป็นการปรับแก้ขั้นตอนในการประมาณค่าสูญหายของวิธี KNN ใน 2 ขั้นตอน คือ ขั้นตอนการคำนวณระยะทาง โดยใช้การหาระยะทางแบบ Bhattacharyya และขั้นตอน การประมาณค่าข้อมูลสูญหายด้วยวิธีการใช้ค่าเฉลี่ยเดไซล์ และเมื่อ

---

[1] วิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา มหาวิทยาลัยบูรพา 169 ถนนลงหาดบางแสน ตำบลแสนสุข อำเภอเมืองชลบุรี จังหวัดชลบุรี 20131

[1] College of Research Methodology and Cognitive Science, Burapha University, 169 Longhaad Bangsaen Road, Saensook, Mueang, ChonBuri 20131, Thailand.

[2] คณะศึกษาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตปัตตานี 181 ถนนเจริญประดิษฐ์ ตำบลรูสะมิแล อำเภอเมือง จังหวัดปัตตานี 94000

[2] Faculty of Education, Prince of Songkla University, Pattani Campus, 181 Charoen Pradit Road, Rusamilae, Mueang, Pattani 94000, Thailand.

[*] ผู้นิพนธ์ประสานงาน ไปรษณีย์อิเล็กทรอนิกส์ (Corresponding author, e-mail): patchana.s@nrru.ac.th Tel: 08 9917 1320

พิจารณาผลการเปรียบเทียบประสิทธิภาพของการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH ในการจำลองสถานการณ์ วิธี DKNN-BH ให้ผลดีกว่าการประมาณค่าข้อมูลสูญหายแบบเดิมในทุกกรณี จากการพิจารณาค่าความคลาดเคลื่อนกำลังสองเฉลี่ยที่ให้ค่าต่ำที่สุด เมื่อข้อมูลมีร้อยละการสูญหาย เท่ากับ ร้อยละ 5 10 20 30 และ 40 มีร้อยละค่านอกเกณฑ์ เท่ากับ ร้อยละ 0 5 10 และ 20 และค่า k เท่ากับ 11 13 15 17 และ 19 โดยค่า MSE จะยิ่งมีค่าลดลงเมื่อร้อยละค่านอกเกณฑ์และค่า k ลดลง

คำสำคัญ: การประมาณค่าข้อมูลสูญหาย, ข้อมูลสูญหาย, การปรับแก้วิธีเคเนียร์เรสเนเบอร์

ABSTRACT

The objective of this research was to develop a new method for missing data estimation by using Decile Mean K-Nearest Neighbor Bhattacharyya Imputation (DKNN-BH). This method evaluated the missing data by K-Nearest Neighbor Imputation (KNN) from fine-tuning of the decile mean and Bhattacharyya distance to compare the effectiveness of the new missing data estimation with Mean Imputation (MI), K-Nearest Neighbor Imputation (KNN) and Decile Mean K-Nearest Neighbor Imputation (DKNN) methods. The Monte Carlo simulation was implemented for 300 cases with 4 options : sample size, level of missing data, size of outliers, and k constants for new missing data DKNN-BH, KNN and DKNN methods. Each situation was replicated 500 times. The results showed that the new developed missing data estimation method, DKNN-BH derived from the fine tuning of KNN using Decile Mean and Bhattacharyya distance. There were 2 steps of DKNN-BH: calculation of Bhattacharyya distance and estimation of missing data using Decile Mean method. After comparing the efficacy of both data missing estimation methods from simulation results, the new method (DKNN-BH) was better than the old one in all cases by using the lowest mean square error. The simulation results also revealed that when the percentage of missing data were 5, 10, 20, 30 and 40, the percentage of outliers were 0, 5, 10, 20 and k constant values were 11, 13, 15, 17, and 19 respectively, the lowest mean square error will decrease as the percentage of outliers and k constants decrease.

Key words: missing data imputation, missing data, adjusted K Nearest Neighbor method

บทนำ

ข้อมูลที่มีความถูกต้องสมบูรณ์ เป็นสิ่งที่นักวิจัย นักวิเคราะห์ข้อมูล หรือนักจัดการข้อมูลปรารถนา นำไปสู่การใช้ข้อมูลในการคาดคะเนทำนายหรือพยากรณ์สิ่งที่จะเกิดขึ้นในอนาคต เพื่อให้เกิดประโยชน์ในการวางแผนล่วงหน้าได้อย่างใกล้เคียงความเป็นจริง ข้อมูลสูญหาย (Missing data) เป็นปัญหาที่นักวิจัยพบในการนำข้อมูลไปวิเคราะห์ด้วยเทคนิคต่างๆ ในทางปฏิบัติเป็นไปได้ยากที่ข้อมูลจากการเก็บรวบรวมจะมีความสมบูรณ์ถูกต้อง อาจเนื่องมาจากขั้นตอนในการเก็บข้อมูลไม่รัดกุม ผู้ให้ข้อมูลไม่ใส่ใจในรายละเอียดของการตอบ หรือ

เกิดเหตุการณ์ที่ไม่สามารถเก็บข้อมูลได้ตามแผน เหตุการณ์ข้างต้นจึงทำให้เกิดข้อมูลสูญหายซึ่งปัญหาเหล่านี้อาจทำให้เกิดผลกระทบต่อประสิทธิภาพในการวิเคราะห์ข้อมูล

ปัญหาข้อมูลสูญหายอาจไม่ใช่ปัญหาที่รุนแรง ถ้าการวิเคราะห์ข้อมูลเป็นการวิเคราะห์ตัวแปรเดียว (Univariate Data) เช่น การหาค่าเฉลี่ย ร้อยละ หรือสถิติพรรณนาอื่นๆ แต่ถ้าข้อมูลจำเป็นต้องใช้การวิเคราะห์ตัวแปรพหุ (Multivariate Data) เช่น การวิเคราะห์ถดถอยพหุ การวิเคราะห์ปัจจัย การวิเคราะห์กลุ่ม การวิเคราะห์จำแนก เป็นต้น ในกรณีนี้การสูญหายของข้อมูลจะส่งผลกระทบรุนแรง เพราะถ้าหน่วยวิเคราะห์ใดมีตัวแปรขาดหายไปเพียงหนึ่งตัวแปรก็จะตัดหน่วยวิเคราะห์นั้นทิ้งทั้งหน่วยโดยไม่สนใจว่าตัวแปรอื่นมีข้อมูลครบถ้วนหรือไม่ ซึ่งการตัดข้อมูลทิ้ง Kim and Curry (1977) ได้ทำการศึกษาโดยวิธีทดลองพบว่า ถ้าตัวแปรแต่ละตัวหายไปโดยสุ่มเพียงร้อยละ 10 จะมีผลทำให้ต้องตัดหน่วยวิเคราะห์ทิ้งถึงร้อยละ 59 ซึ่งเป็นการสูญเสียข้อมูลที่มีอัตราสูงมาก และข้อมูลที่เหลือหลังจากการตัดค่าสังเกตที่ไม่สมบูรณ์ทิ้งจะทำให้การวิเคราะห์ข้อมูลเกิดความเอนเอียงและไม่ถูกต้อง (Robins and Wang, 2000) ดังนั้นควรพิจารณาหาวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีที่เหมาะสมก่อนการวิเคราะห์ข้อมูล

วิธีการจัดการข้อมูลสูญหาย แบ่งเป็น 2 ประเภทหลักคือ การตัดกลุ่มข้อมูลที่สูญหายทิ้งไป (Ignoring And Discarding Data) และการประมาณค่าข้อมูลสูญหายด้วยค่าสูญหายจากการคำนวณ (Imputation) โดยมีผู้ศึกษาและวิจัยเพื่อพัฒนาวิธีการจัดการค่าสูญหายด้วยวิธีการต่างๆ ที่มีประสิทธิภาพที่สุดมาใช้ในการประมาณค่าสูญหายของข้อมูลให้เหมาะสมกับประเภทของข้อมูลที่เกิดการสูญหาย เริ่มตั้งแต่วิธีพื้นฐาน ได้แก่ วิธีประมาณค่าสูญหายค่าเฉลี่ยเลขคณิต หรือ Mean Imputation: MI ซึ่งเป็นวิธีที่ง่ายในการประมาณค่าข้อมูลสูญหาย แต่วิธีการนี้อาจจะทำให้เกิดการเบี่ยงเบนของข้อมูลและเหมาะสมกับข้อมูลบางประเภทเท่านั้น ไปจนถึงวิธีที่มีความยุ่งยากซับซ้อนมากๆ เช่น Expectation Maximization หรือ Maximum Likelihood Estimation เป็นต้น การประมาณค่าข้อมูลสูญหายโดยใช้เทคนิคทางเหมืองข้อมูล เป็นอีกวิธีการหนึ่งที่ได้รับความสนใจเนื่องจากไม่ยึดติดกับเทคนิคทางสถิติ ตัวอย่างการประมาณค่าข้อมูลสูญหายโดยเทคนิคทางเหมืองข้อมูล เช่น วิธีโครงข่ายประสาทเทียม หรือ Artificial Neural Networks (Bishop, 1995; Schioler and Hartmann, 1992) วิธี Singular Value Decomposition: SVD (Troyanskaya *et al*., 2001; Kim *et al*., 2004) แต่ทั้งสองวิธีนี้เป็นวิธีการประมาณค่าที่มีความซับซ้อน เข้าใจยาก อีกวิธีการหนึ่งในการประมาณค่าด้วยเทคนิคเหมืองข้อมูลที่ได้รับความนิยม คือ วิธีเคเนียร์เรสเนเบอร์ หรือ K Nearest Neighbor: KNN (Troyanskaya *et al*., 2001) ซึ่งเป็นวิธีที่มีความยืดหยุ่นสูง ง่ายต่อการใช้งาน และมีประสิทธิภาพ แต่ยังมีจุดด้อยในการถูกโน้มเอียงโดยค่านอกเกณฑ์ (Hengpraprohm and Meesad, 2008) ขั้นตอนของการประมาณค่าซึ่งใช้การประมาณค่าด้วยค่าเฉลี่ย k จำนวนของข้อมูลที่มีการสูญหายจากงานวิจัยของ Rana *et al*. (2012) ได้ศึกษาการวัดแนวโน้มเข้าสู่ส่วนกลางแบบใหม่ที่มีความแกร่งโดยการใช้ค่าเฉลี่ยเดไซล์ (Decile Mean) ซึ่งสามารถมีความแกร่งกับค่านอกเกณฑ์ได้ และในขั้นตอนของการหาระยะทางซึ่งเดิมใช้การหาระยะทาง Euclidian แต่จากการศึกษาพบว่าการหาระยะทาง Bhattacharyya มีความไวต่อค่านอกเกณฑ์ (Pasunon and Nilakorn, 2007) เช่นกัน ดังนั้นในการศึกษาวิจัยในครั้งนี้ผู้วิจัยจึงสนใจจะพัฒนาวิธีการประมาณค่าข้อมูลสูญหายโดยการปรับแก้ การประมาณค่าสูญหายด้วยวิธีเคเนียร์เรสเนเบอร์ ซึ่งมีความน่าสนใจ

โดยใช้วิธีค่าเฉลี่ยเดไซล์และการหาระยะทางแบบ Bhattacharyya (Bhattacharyya, 1943) มาช่วยในการปรับแก้ เพื่อให้ได้วิธีการประมาณค่าสูญหายที่มีประสิทธิภาพในการประมาณค่ามากยิ่งขึ้น เรียกว่า วิธีการประมาณค่าสูญหาย Decile Mean K Nearest Neighbor-Bhattacharyya หรือ DKNN-BH เพื่อแก้ปัญหาค่านอกเกณฑ์ของข้อมูลในการประมาณค่าสูญหาย ที่มีประสิทธิภาพ ถูกต้องและใกล้เคียงกับค่าจริงมากที่สุด โดยมีวัตถุประสงค์เพื่อพัฒนาวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH ที่เกิดจากการปรับแก้วิธีการประมาณค่าข้อมูลสูญหายจากวิธีเคเนียร์เรสเนเบอร์ ด้วยค่าเฉลี่ยเดไซล์และการหาระยะทาง Bhattacharyya และเพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH กับวิธีเดิม คือวิธีการประมาณค่าข้อมูลสูญหายค่าเฉลี่ยเลขคณิต วิธีการประมาณค่าข้อมูลสูญหาย KNN และวิธีการประมาณค่าข้อมูลสูญหาย DKNN ภายใต้การจำลองสถานการณ์

วิธีดำเนินการวิจัย

การวิจัยในครั้งนี้เพื่อพัฒนาวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ ที่เกิดจากการปรับแก้วิธีการประมาณค่าข้อมูลสูญหายจากวิธีเคเนียร์เรสเนเบอร์ ด้วยค่าเฉลี่ยเดไซล์และการหาระยะทางแบบ Bhattacharyya (DKNN-BH) ให้มีประสิทธิภาพในการประมาณค่าสูญหายที่ดี ซึ่งมีวิธีการดำเนินการตามขั้นตอน ดังนี้

1. ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้องกับวิธีการประมาณค่าข้อมูลสูญหาย MI KNN และ DKNN รวบรวมแนวคิดและวิธีการของการประมาณค่าข้อมูลสูญหาย KNN รวมถึงแนวทางแก้ปัญหา

2. ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้องกับการวัดแนวโน้มเข้าสู่ส่วนกลางของข้อมูลโดยใช้ค่าเฉลี่ยเดไซล์

3. ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้องกับวิธีการหาระยะทางแบบ Bhattacharyya

4. เสนอกรอบแนวคิดวิธีประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH ที่เกิดจากการปรับแก้ KNN และ DKNN

5. การจำลองข้อมูลจากการแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Distribution) ซึ่งใช้หลักของการจำลองด้วยเลขสุ่ม (Random Numbers) สุ่มจากประชากรด้วยวิธีการสุ่มอย่างง่าย (Simple Random Sampling) เพื่อให้ได้ขนาดตัวอย่างตามที่กำหนด

6. สร้างข้อมูลให้มีค่านอกเกณฑ์ในตัวแปรตามลักษณะของข้อมูลจริง

7. สร้างข้อมูลสูญหาย ในขั้นตอนนี้จะทำการสร้างข้อมูลสูญหายแบบสุ่มสมบูรณ์ (MCAR) โดยมีร้อยละของข้อมูลสูญหายตามที่กำหนดและข้อมูลที่ถูกตัดออกไปนั้นจะเก็บไว้เพื่อนำมาเปรียบเทียบกับค่าใหม่ที่จะประมาณขึ้นมาทั้ง 4 วิธี

8. ดำเนินการประมาณค่าข้อมูลสูญหาย ด้วยวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH กับวิธีเดิม คือ วิธีการประมาณค่าข้อมูลสูญหายค่าเฉลี่ยเลขคณิต วิธีการประมาณค่าข้อมูลสูญหาย KNN และวิธีการประมาณค่าข้อมูลสูญหาย DKNN ภายใต้การจำลองสถานการณ์ 300 สถานการณ์ จำนวน 4 เงื่อนไข คือ

8.1 ขนาดของกลุ่มตัวอย่าง 3 กลุ่ม คือ จำนวนตัวอย่าง 100 300 และ 500 ตัวอย่าง

8.2 ระดับการสูญหายของข้อมูล 5 ระดับ คือ ร้อยละ 5 10 20 30 และ 40

8.3 ขนาดของค่านอกเกณฑ์ 4 ระดับ คือ ร้อยละ 0 5 10 และ 15

8.4 ค่าคงที่ k 5 ระดับ คือ 11 13 15 17 และ 19 ซึ่งเป็นค่าคงที่ในวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH และวิธีที่เกี่ยวข้องคือ

วิธีการประมาณค่าข้อมูลสูญหาย KNN และวิธีการประมาณค่าข้อมูลสูญหาย DKNN

9. เปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหายพิจารณาจาก ค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error: MSE) ของวิธีการประมาณค่าข้อมูลสูญหายแต่ละวิธี ซึ่งมีสูตรในการคำนวณดังนี้

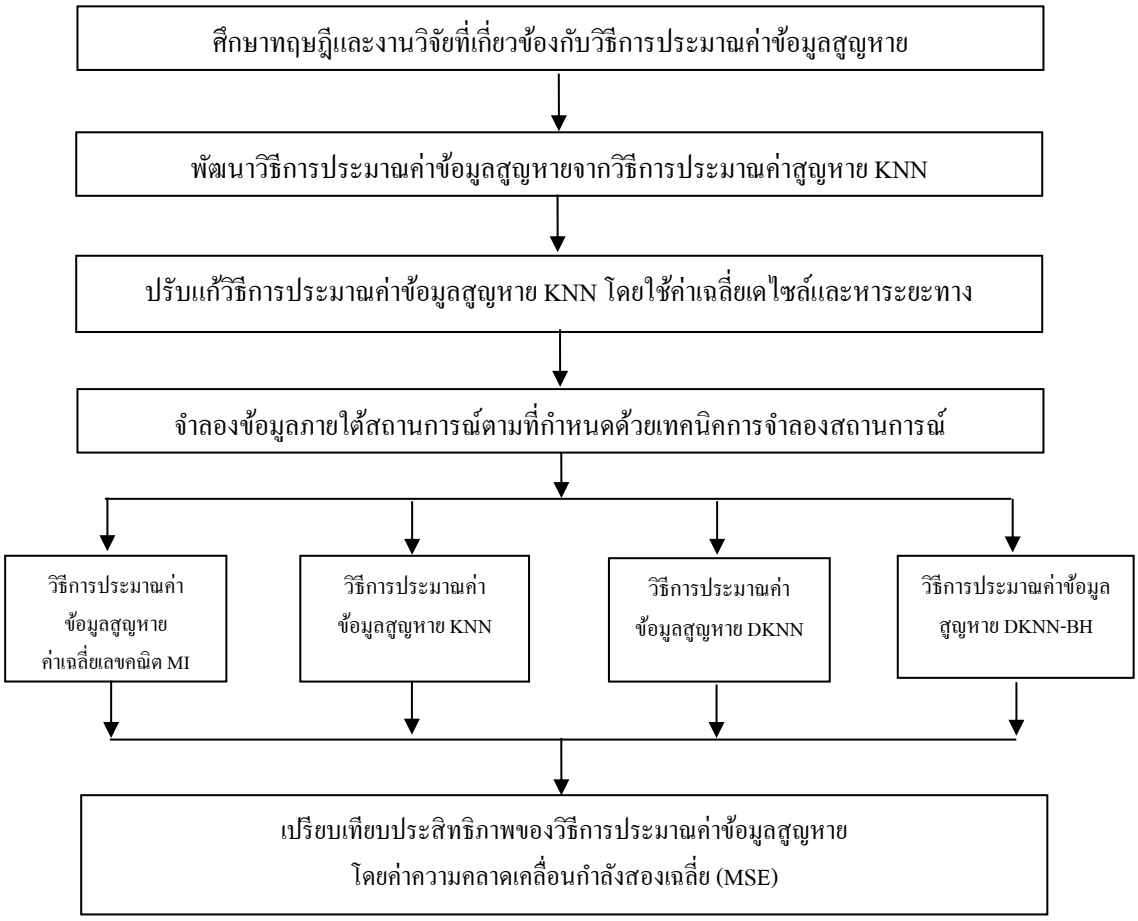
ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error: MSE) โดยมีสูตรในการคำนวณดังนี้

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

เมื่อ y_i คือ ค่าจริงของข้อมูล

$\hat{y}_i$ คือ ค่าทดแทนของข้อมูลที่มีการสูญหาย

n คือ จำนวนค่าที่สูญหาย



ภาพที่ 1 กรอบแนวทางการวิจัยในภาพรวม

ผลการวิจัยและวิจารณ์ผล

การประมาณค่าข้อมูลสูญหายด้วยวิธีใหม่เกิดจากการปรับแก้วิธีการประมาณค่าสูญหายวิธี K-Nearest Neighbor Imputation (KNN) ด้วยการใช้ค่าเฉลี่ยเดไซล์และการหาระยะทางแบบ Bhattacharyya เพื่อประมาณค่าข้อมูลที่สูญหาย โดยมีวัตถุประสงค์เพื่อต้องการปรับปรุงประสิทธิภาพของการประมาณค่าข้อมูลสูญหาย และเพื่อลดผลกระทบจากค่านอกเกณฑ์ (Outlier) เรียกวิธีการนี้ว่า “การประมาณค่าข้อมูลสูญหายวิธี Decile Mean K-Nearest Neighbor Imputation (DKNN)” และเมื่อนำการหาระยะทางแบบ Bhattacharyya มาช่วยในการปรับแก้เรียกวิธีการนี้ว่า “การประมาณค่าข้อมูลสูญหายวิธี Decile Mean K-Nearest Neighbor Bhattacharyya Imputation (DKNN-BH) กำหนดให้ P เป็นประชากรที่มีขนาดเท่ากับ N หน่วยและ S เป็นตัวอย่างที่ได้มาจากการสุ่มแบบไม่ใส่คืนจากประชากร P ขนาดเท่ากับ n หน่วย โดยในแต่ละหน่วย (x_i, y_i) เป็นค่าที่สังเกตได้ โดยที่ $i = 1, 2, \ldots, n$ และ ตัวแปร $x_1, x_2, \ldots, x_n$ เป็นตัวแปรอิสระ

กำหนดให้

y เป็นตัวแปรตามที่ขึ้นอยู่กับตัวแปร (ลักษณะ) $x_1, x_2, \ldots, x_m$ โดยที่

$$y = \begin{cases} y_j, j = 1, \ldots, r & \text{เป็นข้อมูลสมบูรณ์} \\ y_j, j = r+1, \ldots, n & \text{เป็นข้อมูลสูญหาย} \end{cases}$$

x_{ij} ; i = 1, 2, … , n; j = 1, 2, … , m เป็นค่าของตัวแปรอิสระที่เก็บรวบรวมข้อมูลได้แบบสมบูรณ์ (ไม่มีการสูญหาย)

ตารางที่ 1 ข้อมูลที่ใช้ในตัวอย่าง

หน่วยที่	y	x_1	x_2	x_3	**...**	x_m
1	y_1	x_{11}	x_{12}	x_{13}	...	x_{1m}
2	y_2	x_{21}	x_{22}	x_{23}	...	x_{2m}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
r	y_r	$x_{(r)1}$	$x_{(r)2}$	$x_{(r)3}$	...	$x_{(r)m}$
$r+1$		$x_{(r+1)1}$	$x_{(r+1)2}$	$x_{(r+1)3}$	...	$x_{(r+1)m}$
⋮	Missing	⋮	⋮	⋮	⋮	⋮
n		x_{n1}	x_{n2}	x_{n3}	...	x_{nm}

การประมาณค่าข้อมูลสูญหายแบบใหม่เกิดจากการปรับแก้วิธีการประมาณค่าสูญหายวิธี K-Nearest Neighbor Imputation (KNN) โดยการใช้ค่าเฉลี่ยเดไซล์และการหาระยะทางแบบ Bhattacharyya เพื่อประมาณค่าข้อมูลที่สูญหาย มีขั้นตอนวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH ดังนี้

ส่วนที่ 1 คำนวณระยะทาง

ขั้นที่ 1 กำหนดจำนวน k โดยค่า k ที่เหมาะสมมีค่าอยู่ระหว่าง 10 ถึง 20 (Cartwright *et*

al., 2003; Ladha and Deepa, 2011) ซึ่งใช้ค่า k ที่ 11 13 15 17 และ 19

ขั้นที่ 2 คำนวณระยะห่างระหว่างตัวอย่างที่ประกอบด้วยข้อมูลสูญหายกับตัวอย่างทั้งหมดด้วยสมการ (1) โดยกำหนดให้ y_j เป็นข้อมูลสูญหาย เมื่อ i = 1, 2,…,r j= 1, 2, … , mและ d_i เป็นค่าระยะห่าง

$$d_i = \text{Bhattacharyya}(y_{i'j}, y_{ij}) = -\log\sum\left(\frac{\sqrt{y_{i'j}\,y_{ij}}}{\sqrt{\sum y_{i'j}\sum y_{ij}}}\right) \quad (1)$$

ขั้นที่ 3 เรียงระยะห่างที่คำนวณได้ทั้งหมดจากน้อยไปหามากและเลือกค่าที่มีระยะห่างกับค่าสูญหายน้อยที่สุด k ตัว

ส่วนที่ 2 การประมาณค่าสูญหาย

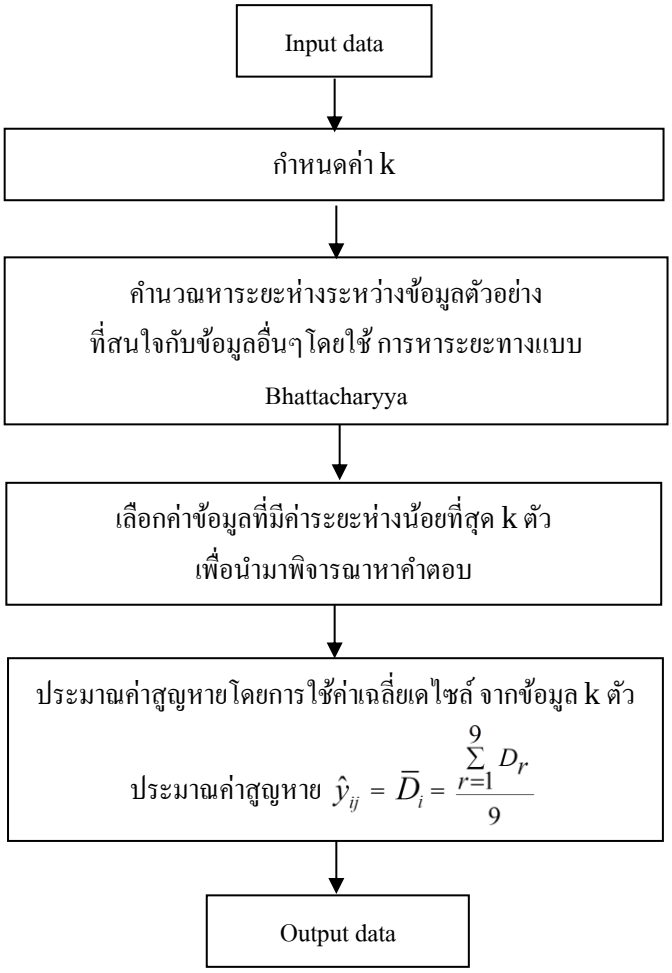
ขั้นที่ 4 คำนวณหาค่าเดไซล์ในทุกตำแหน่ง จากข้อมูล k ตัว

$$\text{ค่าสูญหาย} = \hat{y}_{ij} = \frac{\sum_{r=1}^{9} D_r}{9} \quad (2)$$

เมื่อ D_r คือ ค่าเดไซล์ของข้อมูล ในตำแหน่งที่ r

r คือ ตำแหน่งเดไซล์ของข้อมูล มีค่าเท่ากับ 1, 2, 3, …., 9

ขั้นที่ 5 ประมาณค่าสูญหายโดยนำค่าจากขั้นตอนที่ 4 มาหาค่าเฉลี่ยเดไซล์ และทำการประมาณค่าข้อมูลสูญหาย ซึ่งสามารถอธิบายขั้นตอนวิธีการประมาณค่าสูญหายแบบใหม่ DKNN-BH ดังนี้



ภาพที่ 2 ขั้นตอนวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH

จากการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหายที่พัฒนาขึ้นใหม่หรือวิธี DKNN-BH กับวิธีการประมาณค่าข้อมูลสูญหายอีก 3 วิธี คือ วิธีประมาณค่าด้วยค่าเฉลี่ยเลขคณิต วิธี KNN และวิธี DKNN โดยทำการจำลองข้อมูลด้วยวิธีมอนติคาร์โลทั้งหมด 300 สถานการณ์ และทดลองซ้ำ 500 รอบในแต่ละสถานการณ์ ด้วยโปรแกรม R ในการจำลองข้อมูล ซึ่งมีการแจกแจงปกติ ภายใต้ 4 เงื่อนไข คือ ขนาดตัวอย่าง 3 สถานการณ์ คือ ขนาดตัวอย่าง เท่ากับ 100 300 และ 500 ระดับการสูญหายของข้อมูล 5 ระดับ คือ ร้อยละ 5 10 20 30 และ 40 ขนาดของค่านอกเกณฑ์ 4 ระดับ คือ ร้อยละ 0 5 10 และ 20 และค่า k 5 ระดับ คือ 11 13 15 17 และ 19 ของวิธีการประมาณค่าสูญหายวิธี KNN DKNN และ DKNN-BH โดยใช้เกณฑ์ในการพิจารณาเปรียบเทียบประสิทธิภาพ คือ ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error: MSE) เพื่อสรุปว่าวิธีการประมาณค่าข้อมูลสูญหายวิธีการใดเป็นวิธีที่เหมาะสมที่สุดในแต่ละสถานการณ์ โดยวิธีที่ให้ค่า MSE ต่ำกว่า จะเป็นวิธีที่มีประสิทธิภาพมากกว่า ผลการวิจัยพิจารณาเปรียบเทียบประสิทธิภาพได้ข้อมูลที่ให้ค่า MSE ต่ำที่สุด ซึ่งพบว่าการประมาณค่าข้อมูลสูญหายแบบใหม่ที่พัฒนาขึ้นคือวิธี DKNN-BH ที่มีการใช้ค่าเฉลี่ยเดไซล์และระยะทาง Bhattacharyya ให้ผลดีที่สุดในทุกกรณี ซึ่งสอดคล้องกับงานวิจัยของ Rana *et al.* (2012) ที่ใช้การวัดแนวโน้มเข้าสู่ส่วนกลางโดยวิธีค่าเฉลี่ยเดไซล์มีผลที่ดีกว่าวิธีอื่นๆ เมื่อมีหรือไม่มีค่านอกเกณฑ์ (Outlier) ส่วนค่า k ที่มี

ค่าเหมาะสมให้ผลดีอยู่ในช่วงระหว่าง 10 - 20 ที่อ้างอิงมาและค่า k ที่ให้ผลต่ำที่สุด คือ 11 ใกล้เคียงกับงานของ Troyanskaya *et al*. (2001) ที่ได้ผลค่า k ที่เหมาะสมอยู่ที่ 10 และวิธีการประมาณค่าสูญหายด้วยค่าเฉลี่ยเลขคณิตมีประสิทธิภาพน้อยที่สุดในทุกกรณีเช่นกันซึ่งสอดคล้องกับงานของ Liao *et al*. (2014) โดยในขนาดตัวอย่าง 100 300 และ 500 ให้ค่า MSE ต่ำสุดที่ระดับการสูญหายร้อยละ 5 ในทุกขนาดตัวอย่าง โดยมีรายละเอียดของผลการวิจัยในแต่ละกรณีแสดงในตารางที่ 2 ถึง 4 ดังนี้

ตารางที่ 2 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวประมาณ โดยมีขนาดตัวอย่าง เท่ากับ 100 และระดับการสูญหายร้อยละ 5

ค่า k	ร้อยละค่านอกเกณฑ์	ค่า MSE ของวิธีการประมาณค่าข้อมูลสูญหาย			
		MI	KNN	DKNN	DKNN-BH
11	0	9.67009	9.49846	9.52091	3.76940
	5	9.38047	9.58167	9.59992	3.98445
	10	9.42661	9.19019	9.20404	4.24735
	20	9.58879	9.57781	9.59711	5.06736
13	0	9.67009	9.44458	9.48287	4.18064
	5	9.38047	9.47143	9.49620	4.39178
	10	9.42661	9.21391	9.22958	4.72262
	20	9.58879	9.50967	9.53348	5.59584
15	0	9.67009	9.44153	9.47593	4.58823
	5	9.38047	9.39680	9.42646	4.84497
	10	9.42661	9.22930	9.24258	5.13972
	20	9.58879	9.46702	9.50424	5.97769
17	0	9.67009	9.43029	9.47519	5.01785
	5	9.38047	9.38400	9.42264	5.21839
	10	9.42661	9.26433	9.28139	5.56073
	20	9.58879	9.44313	9.48288	6.33587
19	0	9.67009	9.40654	9.45684	5.41135
	5	9.38047	9.36255	9.41186	5.56356
	10	9.42661	9.25756	9.27788	5.91706
	20	9.58879	9.39747	9.41762	6.65511

จากตารางที่ 2 เมื่อพิจารณาค่าความคลาดเคลื่อนกำลังสองเฉลี่ย โดยมีขนาดตัวอย่างเท่ากับ 100 และ ระดับการสูญหาย ร้อยละ 5 ผลการศึกษาพบว่า วิธีการประมาณค่าข้อมูลสูญหายแบบ DKNN-BH มีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุดหรือมีความแม่นยำสูงสุดในทุกกรณี เมื่อข้อมูลมีร้อยละค่านอกเกณฑ์ เท่ากับ ร้อยละ 0 5 10 และ 20 และค่า k เท่ากับ 11 13 15 17 และ 19 โดยค่า

MSE จะยิ่งมีค่าลดลงเมื่อร้อยละค่านอกเกณฑ์และ ค่า k ลดลง

ตารางที่ 3 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวประมาณ โดยมีขนาดตัวอย่าง เท่ากับ 300 และ ระดับการสูญหายร้อยละ 5

ค่า k	ร้อยละ	ค่า MSE ของวิธีการประมาณค่าข้อมูลสูญหาย			
	ค่านอกเกณฑ์	MI	KNN	DKNN	DKNN-BH
11	0	9.48177	9.09297	9.11853	1.52732
	5	9.31272	9.14603	9.16918	2.00627
	10	9.64039	9.15053	9.17719	1.98381
	20	9.47758	9.01873	9.04038	1.71055
13	0	9.48177	8.97301	9.01283	1.67474
	5	9.31272	9.06219	9.10113	2.14105
	10	9.64039	9.04523	9.07977	2.20404
	20	9.47758	8.94787	8.97702	1.91148
15	0	9.48177	8.93509	8.97979	1.80241
	5	9.31272	9.00616	9.04925	2.24997
	10	9.64039	8.96807	9.00945	2.37498
	20	9.47758	8.85502	8.88264	2.13324
17	0	9.48177	8.89182	8.93768	1.92956
	5	9.31272	8.93321	8.98151	2.36595
	10	9.64039	8.93768	8.97003	2.54518
	20	9.47758	8.77235	8.80487	2.34580
19	0	9.48177	8.85982	8.92021	2.05710
	5	9.31272	8.84753	8.89201	2.48599
	10	9.64039	8.94173	8.99940	2.70735
	20	9.47758	8.75574	8.77698	2.57905

จากตารางที่ 3 เมื่อพิจารณาค่าความคลาดเคลื่อนกำลังสองเฉลี่ย โดยมีขนาดตัวอย่างเท่ากับ 300 และระดับการสูญหาย ร้อยละ 5 ผลการศึกษาพบว่า วิธีการประมาณค่าข้อมูลสูญหายแบบ DKNN-BH มีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุดหรือมีความแม่นยำสูงสุดในทุกกรณี เมื่อข้อมูลมีร้อยละค่านอกเกณฑ์ เท่ากับ ร้อยละ 0 5 10 และ 20 และค่า k เท่ากับ 11 13 15 17 และ 19 โดยค่า MSE จะยิ่งมีค่าลดลงเมื่อร้อยละค่านอกเกณฑ์และค่า k ลดลง

ตารางที่ 4 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวประมาณ โดยมีขนาดตัวอย่าง เท่ากับ 500 และ ระดับการสูญหายร้อยละ 5

ค่า k	ร้อยละ ค่านอกเกณฑ์	วิธีการประมาณค่าข้อมูลสูญหาย			
		MI	KNN	DKNN	DKNN-BH
11	0	9.42080	9.18170	9.20606	1.17536
	5	9.36161	8.89337	8.91668	1.43501
	10	9.63915	8.94670	8.96942	1.28910
	20	9.44205	8.83913	8.85792	1.21232
13	0	9.42080	9.10045	9.14724	1.22853
	5	9.36161	8.80469	8.84509	1.52332
	10	9.63915	8.84344	8.88588	1.41543
	20	9.44205	8.78714	8.82435	1.31018
15	0	9.42080	8.99480	9.05247	1.30165
	5	9.36161	8.73206	8.78689	1.62614
	10	9.63915	8.79732	8.85262	1.55660
	20	9.44205	8.72956	8.76819	1.41809
17	0	9.42080	8.93054	8.98601	1.37796
	5	9.36161	8.69288	8.74414	1.72289
	10	9.63915	8.77810	8.82378	1.70418
	20	9.44205	8.67631	8.71967	1.54313
19	0	9.42080	8.88151	8.94994	1.46314
	5	9.36161	8.65568	8.71809	1.81886
	10	9.63915	8.75386	8.80036	1.83927
	20	9.44205	8.66389	8.71890	1.67561

จากตารางที่ 4 เมื่อพิจารณาค่าความคลาดเคลื่อนกำลังสองเฉลี่ย โดยมีขนาดตัวอย่างเท่ากับ 500 และระดับการสูญหาย ร้อยละ 5 ผลการศึกษาพบว่า วิธีการประมาณค่าข้อมูลสูญหายแบบ DKNN-BH มีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุดหรือมีความแม่นยำสูงสุดในทุกกรณี เมื่อข้อมูลมีร้อยละค่านอกเกณฑ์ เท่ากับ ร้อยละ 0 5 10 และ 20 และค่า k เท่ากับ 11 13 15 17 และ 19 โดยค่า MSE จะยิ่งมีค่าลดลงเมื่อร้อยละค่านอกเกณฑ์และค่า k ลดลง

สรุป

1. ผลการพัฒนาวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH

การพัฒนาวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีใหม่เกิดจากการปรับแก้วิธีการประมาณค่าสูญหายวิธี K-Nearest Neighbor Imputation (KNN) ด้วยการใช้ค่าเฉลี่ยเดไซล์และการหาระยะทางแบบ Bhattacharyya เพื่อประมาณค่าข้อมูลที่สูญหาย เรียกวิธีการนี้ว่า “การประมาณค่าข้อมูลสูญหายวิธี Decile

Mean K-Nearest Neighbor Bhattacharyya Imputation (DKNN-BH) ซึ่งเป็นการปรับแก้ขั้นตอนในการประมาณค่าสูญหายของวิธี KNN ในขั้นตอนการคำนวณระยะทาง โดยใช้การหาระยะทางแบบ Bhattacharyya และขั้นตอนการประมาณค่าข้อมูลสูญหายด้วยวิธีการใช้ค่าเฉลี่ยเดไซล์

2. ผลการเปรียบเทียบประสิทธิภาพของวิธีประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH กับวิธีแบบเดิม

จากการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหายที่พัฒนาขึ้นใหม่หรือวิธี DKNN-BH กับวิธีการประมาณค่าข้อมูลสูญหายอีก 3 วิธี คือ วิธีประมาณค่าด้วยค่าเฉลี่ยเลขคณิต วิธี KNN และวิธี DKNN โดยทำการจำลองข้อมูลด้วยวิธีมอนติคาร์โลทั้งหมด 300 สถานการณ์ และกระทำซ้ำ 500 รอบในแต่ละสถานการณ์ จำนวน 4 เงื่อนไข คือ ขนาดตัวอย่าง 3 กลุ่ม คือ 100 300 และ 500 ระดับการสูญหายของข้อมูล 5 ระดับ คือ ร้อยละ 5 10 20 30 และ 40 ของขนาดข้อมูล ขนาดของค่านอกเกณฑ์ 4 ระดับ คือ ร้อยละ 0 5 10 และ 15 และค่าคงที่ k 5 ระดับ คือ 11 13 15 17 และ 19 ของวิธีการประมาณค่าสูญหาย KNN DKNN และ DKNN-BH โดยใช้เกณฑ์ในการพิจารณาเปรียบเทียบประสิทธิภาพ คือ ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error: MSE) เพื่อสรุปว่าวิธีการประมาณค่าข้อมูลสูญหายวิธีการใดเป็นวิธีที่เหมาะสมที่สุดในแต่ละสถานการณ์ โดยวิธีใดที่ให้ค่า MSE ต่ำกว่าจะเป็นวิธีที่มีประสิทธิภาพมากกว่า ผลการศึกษาพบว่า วิธีการประมาณค่าข้อมูลสูญหาย DKNN-BH มีค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุดหรือมีความแม่นยำสูงสุดในทุกกรณี เมื่อข้อมูลมีร้อยละค่านอกเกณฑ์ เท่ากับ ร้อยละ 0 5 10 และ 20 และค่า k เท่ากับ 11 13 15 17 และ 19 โดยค่า MSE จะยิ่งมีค่าลดลงเมื่อร้อยละค่านอกเกณฑ์และค่า k ลดลง

กิตติกรรมประกาศ

ขอขอบคุณมหาวิทยาลัยราชภัฏนครราชสีมา และสำนักงานคณะกรรมการการอุดมศึกษา ที่สนับสนุนทุนการศึกษา และขอขอบคุณสำนักงานเศรษฐกิจการเกษตรและกรมอุตุนิยมวิทยาที่ให้ความอนุเคราะห์และอำนวยความสะดวกสำหรับข้อมูลที่สำคัญและเป็นประโยชน์ต่องานวิจัยนี้

เอกสารอ้างอิง

Bhattacharyya, A. 1943. On a measure of divergence between two statistical populations defined by their probability distributions. **Bulletin of the Calcutta Mathematical Society** 35: 99-109.

Bishop, C.M. 1995. **Neural networks for pattern recognition.** Oxford university press, UK.

Cartwright, M.H., Shepperd, M.J. and Song, Q. 2003. Dealing with missing software project data, pp. 154-165. *In* **Proceedings of the 9th IEEE International Software Metrics Symposium (METRICS'03).** IEEE Computer Society, Sydney.

Hengpraprohm, K. and Meesad, P. 2008. Feature selection of K-Nearest Neighbor for missing value imputation using K-Nearest Neighbor. **Information Technology Journal** 4(7): 55-61. (in Thai)

Kim, J.O. and Curry, J. 1977. The treatment of missing data in multivariate analysis. **Sociological Methods & Research** 6(2): 215-240.

Kim, K.Y., Kim, B.J. and Yi, G.S. 2004. Reuse of imputed data in microarray analysis increases imputation efficiency. **BMC Bioinformatics** *5*(1): 160.

Ladha, L. and Deepa, T. 2011. Feature selection methods and algorithms. **International journal on computer science and engineering** 3(5): 1787-1797.

Liao, S.G., Lin, Y., Kang, D.D., Chandra, D., Bon, J., Kaminski, N. and Tseng, G.C. 2014. Missing value imputation in high-dimensional phenomic data: imputable or not, and how?. **BMC Bioinformatics** 15(1): 346.

Pasunon, P. and Nilakorn, P. 2007. Outliers detection in regression analysis by Bhattacharyya Statistics, pp. 11-18. *In* **The Proceeding of 45th Kasetsart University Annual Conference.** Kasetsart University, Bangkok. (in Thai)

Rana, S., Siraj-Ud-Doulah, M., Midi, H. and Imon, A.H.M.R. 2012. Decile mean: A new robust measure of central tendency. **Chiang Mai journal of science** 39(3): 478-485.

Robins, J.M. and Wang, N. 2000. Inference for imputation estimators. **Biometrika** 87: 113-124.

Schioler, H. and Hartmann, U. 1992. Mapping neural network derived from the Parzen window estimator. **Neural Networks** 5(6): 903-909.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R. and Altman, R.B. 2001. Missing value estimation methods for DNA microarrays. **Bioinformatics** 17(6): 520-525.

Vongprasert, J. 2019. Jacknife and Regression Approaches to Missing Data Imputation. **Journal of Applied Statistics and Information Technology** 3(1): 52-61.