การประยุกต์ Soft Computing และ *k*-Nearest Neighbor เพื่อใช้ประมาณค่าสูญหายของข้อมูล

นรุตม์ บุตรพลอย

โปรแกรมวิชาเทคโนโลยีคอมพิวเตอร์ คณะเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยราชภัฏกำแพงเพชร

Narut_b@hotmail.com

บทคัดย่อ

บทความนี้เป็นการนำเสนอผลการวิจัยการประมาณค่าสูญหายของข้อมูล โดยเป็นการประยุกต์ใช้อัลกอริทึม *k*-nearest neighbor(knn) ซึ่งจากเดิมวิธีการ *k*-Nearest Neighbour (KNN) เป็นการหาสมาชิกที่มีความใกล้ชิดกับข้อมูลที่สูญหายมา *k* ตัว แล้วต้องนำจำนวน *k* นั้นมาหารเพื่อประมาณค่า แต่ในงานวิจัยนี้ได้ปรับปรุงตัวหารด้วยการนำ หลักการ Soft Computation เข้ามาช่วยในการปรับค่า *k* เรียกวิธีการนี้ว่า Soft K-Nearest Neighbour (SKNN) ในการทดลองครั้งนี้ได้นำข้อมูลมาทดสอบทั้งหมด 4 ชุดข้อมูล แต่ละชุดข้อมูลจำลองขนาดการหายที่แตกต่างกันคือ 10%, 20%, 30%, 40% และ 50% ผลการทดลองพบว่าค่า *k* เท่ากับ 10 นั้นให้ผลลัพธ์การประมาณค่าสูญหายได้ดีที่สุด และค่าพารามิเตอร์ของวิธีการ SKNN ที่ให้ผลลัพธ์ที่ดีที่สุดเท่ากับ 0.3

คำสำคัญ— Soft Computing; Hard Computing; *k*-Nearest Neighbor

1. บทนำ

การสูญหายของข้อมูล(Missing Data) เป็นปัญหาหนึ่งมักพบได้บ่อย โดยเฉพาะกับชุดข้อมูลที่เป็นประเภทข้อมูลสำรวจ สาเหตุการสูญหายเกิดขึ้นด้วยกันหลายสาเหตุ [1] เช่น การสำรวจโดยการตอบแบบสอบถามผู้ตอบอาจไม่ประสงค์ที่จะตอบ หรืออาจจะไม่ทราบคำตอบ ทำให้เว้นว่างไว้ หรือเกิดจากความผิดพลาดของซอฟต์แวร์หรือฮาร์ดแวร์ หรืออาจมีการสูญหายในขั้นตอนการกรอกข้อมูลลงในแหล่งบันทึกข้อมูล เป็นต้น

ข้อเสียของการที่มีการสูญหายของข้อมูล

การสูญหายของข้อมูลเกิดขึ้นเกือบทุกงาน ปัญหาที่สำคัญอย่างหนึ่งคือการไม่เก็บรวบรวมข้อมูลอยู่ตลอดเวลาทำให้ข้อมูลอาจมีการสูญหายไป เมื่อระบบมีข้อมูลสูญหายเกิดขึ้นทำให้เกิดความไม่แน่นอนในการนำข้อมูลไปใช้งาน เมื่อนำชุดข้อมูลที่มีการสูญหายของข้อมูลไปใช้งานอาจทำให้เกิดปัญหาขึ้น [5] นอกจากนี้ข้อมูลสูญหายทำให้เกิดความถูกต้องความแม่นยำในการทำนายด้วยอัลกอริทึมในงานที่เกี่ยวกับการเรียนรู้ของเครื่องลดลง [6] การเกิดข้อมูลสูญหายอาจเกิดขึ้นได้จากหลายปัจจัยเช่น งานวิจัยที่มีลักษณะเป็นแบบสอบถามอาจเป็นไปได้ที่ผู้ตอบแบบสอบถามข้ามหรือไม่ตอบแบบสอบถามในบางคำถาม ผลลัพธ์คือเมื่อนำข้อมูลไปใช้จะข้ามข้อที่มีการสูญหายของข้อมูลทำให้ข้อมูลที่จะนำไปประมวลผลมีขนาดเล็กลงดังนั้นจะเห็นได้ว่า การสูญหายของข้อมูลสร้างความยุ่งยากให้กับงานในลักษณะนี้เป็นส่วนใหญ่เพราะขั้นตอนการทำงานส่วนใหญ่ไม่ได้ถูกออกแบบมาให้รองรับข้อมูลสูญหาย

เทคนิคที่ใช้ในการประมาณค่าสูญหาย

ปัจจุบันมีนักวิจัยหลายกลุ่มได้พยายามที่จะจัดการข้อมูลที่มีค่าสูญหายเหล่านี้ เทคนิคที่จะนำมาประมาณค่าที่สูญหายของข้อมูลมีหลายวิธี เช่น การหาตัวแทนของข้อมูล หรือการแทนค่าสูญหายด้วยค่ากลาง(Mean) หรือค่าฐานนิยม (Mode) ของกลุ่มนั้นๆ ซึ่งเป็นวิธีการที่ง่ายแต่อาจทำให้เกิดการเบี่ยงเบนของข้อมูลดังนั้นจึงมีนักวิจัยได้พยายามพัฒนาและศึกษาเพื่อหาวิธีการประมาณค่าสูญหายของข้อมูล โดยสามารถแบ่งเทคนิคออกเป็น 3 ประเภท [4] ดังนี้

1. การตัดกลุ่มข้อมูลที่สูญหายทิ้งไป (ignoring and discarding data) วิธีการนี้แบ่งออกเป็นสองวิธีดังนี้

(1.1) Listwise Deletion เป็นวิธีการตัดแถวที่มีค่าข้อมูลสูญหายทิ้งทั้งหมด ซึ่งเป็นวิธีที่ง่ายที่สุดเพื่อให้ได้มาซึ่งกลุ่มข้อมูลที่สมบูรณ์ที่สุด

(1.2) Pairwise Deletion จะไม่ตัดแถวที่ขาดความสมบูรณ์ทิ้งไป แต่จะนำแถวเหล่านั้นมาใช้ในการประมวลผลด้วย จะเป็นการพิจารณาทุกแถวที่มีค่าแอทริบิวที่กำลังสนใจ

2. การประมาณค่าพารามิเตอร์ (Parameter estimation) ตัวอย่างวิธีการประมาณค่าได้แก่ การประมาณค่าความเป็นไปได้มากที่สุด (Maximum Likelihood Estimation) ซึ่งคำนวณจากข้อมูลที่ทราบทั้งหมด ค่าความเป็นไปได้มากที่สุดจะถูกคำนวณโดยแยกคำนวณจากข้อมูลที่สมบูรณ์ของบางตัวแปรและคำนวณจากกากข้อมูลในทุกตัวแปร ค่าความเป็นไปได้ที่มากที่สุดจากการคำนวณทั้งสอง จะถูกใช้ในการประมาณค่าข้อมูลสูญหายอีกทีหนึ่ง

3.การแทนที่ค่าข้อมูลสูญหายด้วยค่าประมาณจากการคำนวณ (Imputation) วิธีการนี้จะทำการประมาณค่าข้อมูลโดยอาศัยความสัมพันธ์ระหว่างกลุ่มข้อมูล เพื่อนำไปประมาณค่าข้อมูลสูญหาย

กลไกการสูญหายของข้อมูล

กลไกการสูญหายของข้อมูลสามารถจำแนกได้ทั้งหมด 3 ประเภท [3] ได้แก่

1. การสูญหายแบบสุ่มอย่างสมบูรณ์(Missing Completely at Random-MCAR) คือ ความน่าจะเป็นที่ข้อมูลสูญหายที่จุดใดๆ ไม่มีความสัมพันธ์กับค่าของข้อมูลอื่นๆ

2. การสูญหายแบบสุ่ม(Missing at Random-MAR) คือ ความน่าจะเป็นที่ข้อมูลสูญหายขึ้นอยู่กับค่าของข้อมูลอื่นๆ ที่ทราบค่าข้อมูล แต่ไม่ขึ้นกับค่าของข้อมูลที่สูญหายของตัวเอง

3. การสูญหายแบบไม่สุ่ม(Not Missing at Random-NMAR)คือ ความน่าจะเป็นที่ข้อมูลจะสูญหายที่ตำแหน่งใดๆ ขึ้นอยู่กับค่าของตำแหน่งนั้นๆ

2. แนวคิดการแทนที่ข้อมูลด้วย *k*- nearest neighbor

ปัจจุบันมีนักวิจัยหลายกลุ่มได้พยายามที่จัดการข้อมูลที่มีค่าสูญหายเหล่านี้ การแทนค่าข้อมูลที่สูญหายด้วยวิธีต่างๆ วิธี *k*- nearest neighbour :KNN [2] เป็นอัลกอริทึมหนึ่งที่ถูกนำมาประยุกต์ใช้หาความสัมพันธ์ระหว่างกลุ่มข้อมูลที่จะนำมาประมาณค่าที่สูญหาย [8] กับข้อมูลที่มีค่าสูญหายดังสมการ (1)

$$dist\ (X_q, X_i) = \sqrt{\left(\sum_{k=1}^{n}((X_{q,k}) - (X_{i,k}))\right)^2} \quad (1)$$

โดยที่

$dist\ (X_q, X_i)$ คือ ระยะห่างระหว่างตัวอย่าง x_q กับตัวอย่าง x_i คือ คุณสมบัติทั้งหมดของตัวอย่าง x_{pk} คือ คุณสมบัติตัวที่ k ของตัวอย่าง x_i และประมาณค่าข้อมูลสูญหายด้วยสมการที่ (2) ดังนี้

$$\hat{a}_j(X_q) = \frac{\sum_{i=1}^{k} a_j(X_i)}{k} \quad (2)$$

โดยที่

$\hat{a}_j(X_q)$ คือ ค่าประมาณของข้อมูลที่ตำแหน่งแอททริบิว j ของตำแหน่ง X_q เมื่อต้องการประมาณค่าเพื่อแทนข้อมูลที่มีค่าสูญหายจะมีการดำเนินการดังนี้

1. กำหนดค่า k

2. คำนวณหาหาความสัมพันธ์ระหว่างกลุ่มที่จะนำมาประมาณค่าสูญหาย กับกลุ่มข้อมูลสูญหายด้วยวิธีการ Euclidian distance ดังสมการที่ (1)

3. เลือกค่าที่มีระยะห่าง(dist)กับค่าสูญหายน้อยที่สุดมา k ตัว

4. ประมาณค่าข้อมูลสูญหายด้วยการนำ k มาหารดังสมการที่ (2)

ตาราง 1. ตัวอย่าง data set ที่มีค่าสูญหาย

	A_1	A_2	A_3	A_4
R_1	?	2	3	2
R_2	4	1	3	3
R_3	3	4	1	1
R_4	5	2	1	1
R_5	2	2	2	2
R_6	2	4	1	1

ดังตัวอย่างจากตารางที่ 1 ตำแหน่งที่ R_1,A_1 มีค่าสูญหายของข้อมูล เมื่อต้องการแทนค่าสูญหาย ให้ทำตามขั้นตอนโดย ลำดับแรกกำหนดค่า k สมมุติให้มีค่าเท่ากับ 2 จากนั้นคำนวณหาความสัมพันธ์ของตำแหน่งที่หายกับตำแหน่งที่มีอยู่ด้วยสมการที่ (1)

$$dist\ (R_1, R_2) = \sqrt{(2-1)^2 + (3-3)^2 + (2-3)^2}.$$
$$= \sqrt{1^2 + (-1)^2}$$
$$= \sqrt{1+1} = 1.41421$$

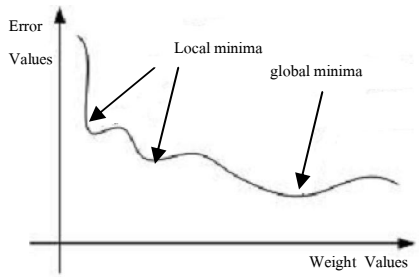
และต้องคำนวณแบบนี้กับทุกตำแหน่ง เลือกเวกเตอร์ที่มีระยะทางสั้นที่สุดของแอททริบิว A_1 ที่สุดมาสองตำแหน่งตามที่กำหนดไว้ที่ค่า k จากนั้นประมาณค่าสูญหายด้วยสมการที่ (2)

การประยุกต์หลักการของ Soft Computing **มาปรับค่าการประมาณค่าสูญหาย**

Soft Computing [7] เป็นเทคนิคที่จำลองแบบความคิดมาจากมนุษย์ เป็นวิธีการคำนวณสภาพภาพแวดล้อมที่ไม่แน่นอน และปรากฏการณ์ที่ซับซ้อน แนวคิดของ Soft Computing เป็นการแนวคิดการคำนวณแบบปัญญาประดิษฐ์ มีลักษณะเป็นการประมวลผลแบบอัจฉริยะ มักจะมีการดูองค์ประกอบรอบข้างเป็นพื้นฐานสำหรับการเรียนรู้ใหม่ Soft Computing จะมาแทนที่การคำนวณแบบดังเดิมนั้น Hard Computing โดยเป็นรูปแบบการประมวลผลด้วยคอมพิวเตอร์ในปัจจุบันมีการประมวลผลเป็นขั้นตอน การแก้ปัญหาที่ชัดเจน ได้ผลลัพธ์ที่มีความถูกต้อง มีความแน่นอน แต่ในบางสถานะการรูปแบบการคำนวณจะอยู่ภายใต้สภาวะที่มีความไม่แน่นอน ความคลุมเครือ ซึ่งการคำนวณแบบ Hard Computing ไม่สามารถแก้ไขปัญหาเหล่านี้ได้

จากสมการที่สองของงานวิจัยได้กล่าวถึงการหาค่าเฉลี่ยที่จะนำมาแทนค่าสูญหายของข้อมูล โดยการนำค่า k ที่กำหนดมาหาร แต่ปัญหาที่พบคือ ค่าสูญหายบางตำแหน่งอาจไม่ต้องการหารด้วยค่า k แต่อาจต้องการตัวหารที่เพิ่มขึ้น หรือลดลง ดังตัวอย่างตารางที่ 1 มีการสูญหายของข้อมูลตำแหน่ง R_1,A_1 เมื่อมีการหาความสัมพันธ์ และเลือกข้อมูลที่มีระยะทางสั้นที่สุดมา 2 ตัวแล้วนั้นต้องถูกหารด้วยค่า k ที่กำหนด ซึ่งผลลัพธ์ที่ได้อาจไม่ใช่คำตอบที่ต้องการ หรือคำตอบที่ถูกต้องที่สุด คำตอบที่ได้อาจไปตกอยู่ปัญหาที่เรียกว่า local minima ได้ หมายความว่าคำตอบที่ได้คิดว่าเป็นจุด

ต่ำสุด หรือเหมาะสมที่สุด แต่ในความเป็นจริงอาจมีคำตอบที่เป็นจุดต่ำหรือเหมาะสมกว่านี้ที่เรียกว่า global minima ดังรูปที่ 1



รูปที่ 1. ปัญหา local minima

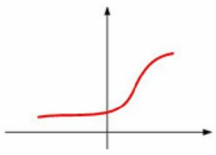
ดังนั้นถ้าขยายขอบเขตของคำตอบโดยการกระตุ้นหรือปรับค่าตัวหาร ซึ่งในที่นี้คือ k อาจเป็นไปได้ที่คำตอบจะตกลงไปที่ global minima นั้นคือถ้ามีการปรับค่า k ก่อนนำไปหาร โดยการนำแนวคิดแบบ Soft Computing มาช่วยอาจทำให้ได้ผลลัพธ์ที่ดีขึ้น โดยงานวิจัยที่[7] ได้ทดลองนำ logistic sigmoid function ดังสมการที่ (3) มาช่วยปรับค่าการหารแล้วเรียกวิธีการใหม่นี้ว่า Soft k-Nearest Neighbor (SKNN) เข้ามาช่วยปรับค่า k ดังสมการที่ (4)

$$x = \frac{1}{(1+e^{-n})} \tag{3}$$

โดยที่

ค่า n คือค่าสุ่มตั้งแต่ 0.1 ถึง 1.0

และเนื่องจากสมการนี้ค่า n ไม่ได้เป็นค่าคงที่ แต่เป็นการสุ่มค่าขึ้นมาดังนั้นกราฟที่ได้จะไม่เป็นเส้นตรง แต่จะมีลักษณะเป็นเส้นโค้งคล้ายกับ S-shape ดังรูปที่ 2



รูปที่ 2. กราฟที่ได้จากสมการ logistic sigmoid function

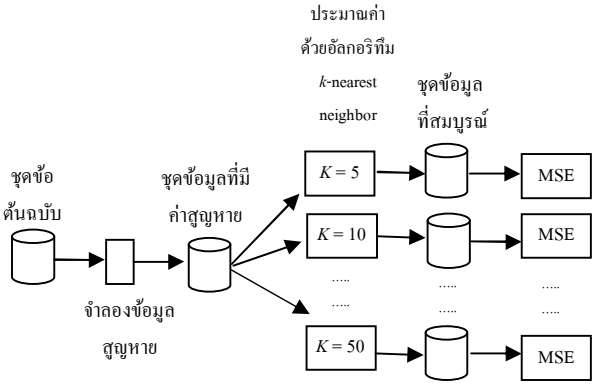
จากสมการที่ (4) ได้นำ logistic sigmoid function เข้ามาช่วยปรับค่าโดยการนำไปหารกับค่า k ที่กำหนดไว้ ทำให้ค่า k เพิ่มขึ้นไม่คงที่ในลักษณะสุ่ม ซึ่งผลลัพธ์ที่ได้อาจมีความใกล้เคียงกับค่าสูญหายมากขึ้น

$$\hat{a}_j(X_q) = \frac{\sum_{i=1}^{k} a_j(X_i)}{k/(1/(1+e^{(-n*k)}))} \tag{4}$$

3. อุปกรณ์และวิธีการทดลอง

การวิจัยครั้งนี้ได้นำข้อมูลมาทดสอบจำนวน 4 ชุดข้อมูล โดยเป็นข้อมูลประเภท Numerical แต่ละชุดข้อมูลจำนวน ได้ทำการจำลองข้อมูลสูญหายเป็นแบบ MCAR และจำลองขนาดการสูญหายที่แตกต่างกันคือ 10%, 20%, 30%, 40% และ 50% เมื่อได้ข้อมูลที่ต้องการได้พัฒนาโปรแกรมเพื่อแทนค่าสูญหาย โดยใช้ภาษาซีในการพัฒนาโปรแกรม และทำการทดลองโดยแบ่งเป็น 2 ระยะได้แก่

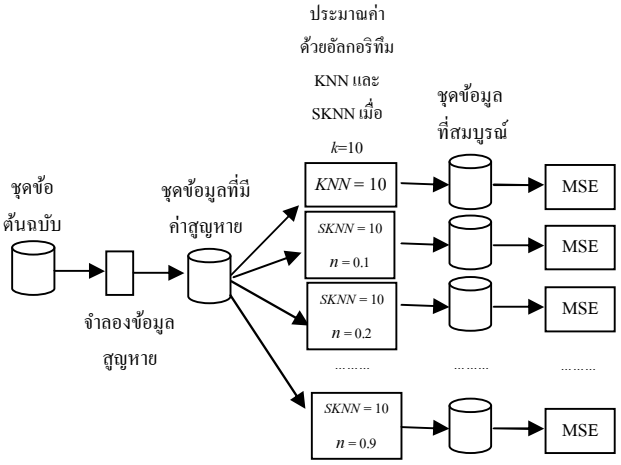
ระยะที่ 1 หาค่า k ที่ดีที่สุดสำหรับประมาณค่าสูญหายซึ่งสามารถแสดงได้ดังรูปที่ 3



รูปที่ 3. จำลองขั้นตอนการทดลองการประมาณค่าสูญหาย
ด้วย k-nearest neighbor

รูปที่ 3 เป็นการแสดงขั้นตอนการทดลองการประมาณค่าสูญหายด้วย k-nearest neighbor โดยในส่วนนี้ผู้วิจัยต้องการหาค่า k ที่ให้ประสิทธิภาพที่ดีที่สุด นั้นคือให้ค่า Mean Square Error น้อยที่สุดนั้นเอง โดยกำหนดค่า k เท่ากับ 5 10 15 20 25 30 35 40 45 และ 50 ตามลำดับ ซึ่งค่า k ที่ให้ประสิทธิภาพที่ดีที่สุดได้แก่ค่า k เท่ากับ 10

ระยะที่ 2 เป็นการหาค่าพารามิเตอร์ที่ดีที่สุดของวิธีการ SKNN เมื่อกำหนดค่า k เท่ากับ 10 ดังแสดงดังรูปที่ 4

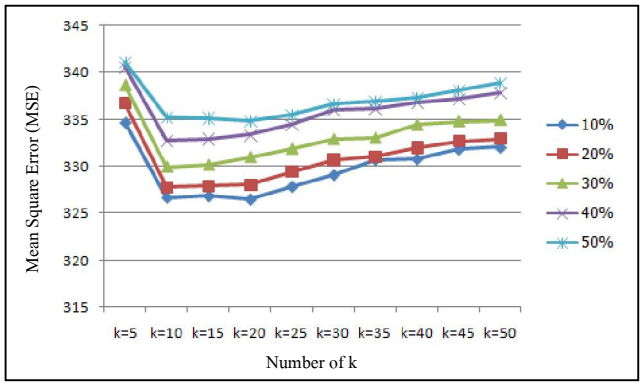


รูปที่ 4. จำลองขั้นตอนการทดลองการประมาณค่าสูญหายด้วย SKNN

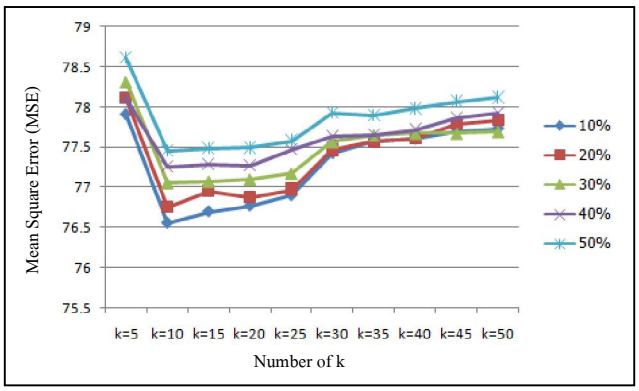
4. ผลการทดลอง

การทดลองนี้แบ่งออกเป็น 2 ระยะ โดยระยะแรกผู้วิจัยได้ทดสอบหาประสิทธิภาพของอัลกอริทึม KNN โดยหาว่าค่า k ที่ให้ผลลัพธ์ที่ดี คือให้ค่า MSE ที่ลดลงมากที่สุดควรมีค่า k เท่ากับเท่าไหร่ ในระยะที่ 2 ผู้วิจัยได้ทดลองประมาณค่าสูญหายด้วย SKNN โดยกำหนดค่า k เท่ากับ 10 และทดสอบหาค่าพารามิเตอร์ n ที่ดีที่สุดว่าควรเป็นเท่าไหร่

จากผลการทดลองในระยะที่ 1 ดังรูปที่ 5 และ 6 พบว่าค่า k ที่ให้ประสิทธิภาพที่ดีที่สุดได้แก่ค่า k เท่ากับ 10 ซึ่งจะเห็นได้ว่าค่า MSE ณ ตำแหน่งดังกล่าวลดลงมากที่สุด ดังนั้นผู้วิจัยจึงเลือกที่จะนำค่า k ดังกล่าวมาทดลองต่อในระยะที่ 2



รูปที่ 5. ผล MSE จากการประมาณค่าด้วย KNN ของชุดข้อมูล cloud



รูปที่ 6. ผล MSE จากการประมาณค่าด้วย KNN ของชุดข้อมูล concrete

ระยะที่ 2 เป็นการประมาณค่าการสูญหายด้วย SKNN โดยระยะนี้ผู้วิจัยทำการเปรียบเทียบประสิทธิภาพระหว่างวิธีการ KNN ที่เลือกค่า k เท่ากับ 10 กับวิธีการ SKNN ที่เลือกค่า k เท่ากับ 10 พร้อมกับปรับค่าตัวหารใหม่ดังสมการที่ (4) โดยทดลองใส่ค่าพารามิเตอร์ n ตั้งแต่ 0.1 ถึง 0.9 ซึ่งจากการทดลองพบว่าวิธีการประมาณค่าการสูญหายด้วย SKNN เมื่อให้ค่าพารามิเตอร์เท่ากับ 0.3 ให้ผลลัพธ์ที่ดีกว่า KNN

ตาราง 2. เปอร์เซ็นของค่า MSE ที่ลดลงของการประมาณค่าการสูญหายด้วย SKNN เมื่อเทียบกับ KNN ของชุดข้อมูล cloud

ขนาดการสูญหาย	เปอร์เซ็นต์การลดลงของ MSE ของ SKNN เมื่อแทนค่า n ตั้งแต่ 0.1 – 0.9								
	0.1	*0.2*	*03*	*0.4*	*0.5*	*0.6*	*0.7*	*0.8*	*0.9*
10%			2%	1%	1%				
20%		7%	7%	3%	1%				
30%		4%	6%	3%	1%				
40%									
50%									

ตาราง 3. ตารางสรุปเพื่อแสดงตำแหน่งและค่าพารามิเตอร์ n ของวิธีการ SKNN ที่ให้ผลลัพธ์ที่ดีกว่า 10-KNN

Data Set	% of missing	ค่าพารามิเตอร์ n								
		0.	0.2	0.3	0.	0.	0.	0.	0.	0.
Cloud	10%			X	X	X				
	20%		X	X	X	X				
	30%		X	X	X	X				
	40%									
	50%									
Concrete										
	10%			X	X	X	X			
	20%			X		X				
	30%		X	X		X				
	40%									
	50%									
Wine										
	10%	X	X	X						
	20%	X	X	X						
	30%									
	40%									
	50%									
Yeast										
	10%	X	X	X						
	20%			X	X					
	30%			X	X					
	40%									
	50%									

ตารางที่ 2 เป็นการแสดงการเปรียบเทียบเปอร์เซ็นต์ที่ลดลงของค่า MSE ของวิธีการ SKNN เมื่อเทียบกับ KNN ของชุดข้อมูล cloud และตารางที่ 3 เป็นตารางสรุปรวมตำแหน่งของวิธีการ SKNN ที่ลดลงกว่าวิธีการ KNN โดยใช้สัญลักษณ์ “ x ” เพื่อระบุถึงตำแหน่งดังกล่าว ซึ่งจากตารางจะเห็นว่าค่าพารามิเตอร์ที่ 0.3 ให้ผลลัพธ์ที่ดีในทุกชุดข้อมูล

5. วิเคราะห์ผลการทดลอง

ปัจจัยที่ทำให้ค่าพารามิเตอร์ 0.3 ให้ผลลัพธ์ดีกว่าค่าพารามิเตอร์อื่นอาจเนื่องมาจากเมื่อมีการแทนค่าพารามิเตอร์ในสมการที่ (4) ค่าของตัวหารที่เกิดขึ้นมาใหม่นั้นเหมาะสมหรืออาจจะใกล้เคียงกับตัวหารที่กลุ่ม k ต้องการ เพราะเมื่อลองแทนค่าพารามิเตอร์น้อยกว่า 0.3 ค่าตัวหารที่ได้มาใหม่จะเพิ่มสูงกว่าค่า k อยู่มาก เมื่อแทนค่าพารามิเตอร์มากกว่า 0.3 ค่าตัวหารจะเข้าใกล้ค่า k มากเกินไป จึงให้ผลลัพธ์ในการประมาณค่าการสูญหายของข้อมูลไม่แตกต่างกับการประมาณค่าแบบ KNN

อย่างไรก็ตามจากตารางที่ 2 จะสังเกตเห็นว่าการให้ค่าพารามิเตอร์อยู่ในช่วง 0.2 ถึง 0.5 ผลลัพธ์ของการประมาณค่าการสูญหายของข้อมูลยังสามารถประมาณค่าได้ดีกว่า KNN แต่ค่าพารามิเตอร์ดังกล่าวไม่สามารถใช้ได้กับทุกชุดข้อมูลที่นำมาทดลอง มีเพียงพารามิเตอร์ที่เท่ากับ 0.3 เพียงค่าเดียวที่สามารถให้ค่า MSE ลดลงเมื่อเทียบกับการประมาณค่าการสูญหายแบบ KNN ในทุกชุดข้อมูล

6. สรุปผลการทดลอง

จากการทดลอง ในการประมาณค่าที่สูญหายของข้อมูลพบว่าวิธีการ KNN ที่กำหนดค่า k เท่ากับ 10 สามารถให้ประสิทธิภาพที่ดีที่สุด นั้นคือค่า MSE ลดลงมากที่สุด หลังจากนั้นผู้วิจัยได้นำหลักการของ soft k-nearest neighbor (SKNN) มาช่วยเพื่อทดลองกับการประมาณค่าโดยกำหนดค่า k เท่ากับ 10 ผลลัพธ์ที่ได้พบว่าค่าพารามิเตอร์ที่ให้ผลลัพธ์ที่ดีที่สุด ที่สามารถใช้ได้กับทุกชุดข้อมูลได้แก่ 0.3

โดยชุดข้อมูล cloud เปอร์เซ็นต์ที่ค่า MSE ของ SKNN ลดลงกว่า KNN สูงสุดเท่ากับ 7% ณ ขนาดการสูญหายที่ 20% ชุดข้อมูล concrete เปอร์เซ็นต์ที่ค่า MSE ของ SKNN ลดลงกว่า KNN สูงสุดเท่ากับ 11% ณ ขนาดการสูญหายที่ 10% ชุดข้อมูล wine เปอร์เซ็นต์ที่ค่า MSE ของ SKNN ลดลงกว่า KNN สูงสุดเท่ากับ 11 % ณ ขนาดการสูญหายที่ 10% และ ชุดข้อมูล yeast เปอร์เซ็นต์ที่ค่า MSE ของ SKNN ลดลงกว่า KNN สูงสุดเท่ากับ 8% ณ ขนาดการสูญหายที่ 10%

แต่อย่างไรก็ตามจากตารางที่ 3 พบว่าวิธีการ SKNN ให้ผลลัพธ์ที่ดีกว่าเฉพาะขนาดการสูญหายของข้อมูลที่มีขนาดน้อย แต่เมื่อข้อมูลมีขนาดใหญ่ 40% ขึ้นไป วิธีการ KNN ให้ผลลัพธ์ที่ดีกว่า

เอกสารอ้างอิง

[1] Gustavo E.A.P.A Batista and Maria Carolina Monard. *Experimental Comparison of K-Nearest Neighbour and MEAN or MODE Imputation Methods with then Internal Strategies used by C4.5 and CN2 to Treat Missing Data*. University of Sao Paulo at Sao Carlos. Brazil. 2003

[2] James E, S. Macleod, Andrew luk and D. Michael Titerington. *A Re-Examination of the DistanceWeighted k-Nearest Neighbor Classification Rule*. IEEE Trans Man Cybernetics. 1987

[3] JosephL.Schafer and John W. Graham. 2002. *Missing Data : Our View of the State of the Art*. Pennsylvania State University Psychological Methods,vol.7,no.2, pp 147–177

[4] Little RJ,Rubin DB. *Statistical Analysis with Missing Data*. New York: John Wiley and Sons, 1987

[5] Maret Pakkunainen, Jarmo Kilpelainen, Satu-Pia Reinikainen and Pentti Minkkinen. *Effect of missing values in estimation of mean of auto-correlated measurement series*. Department of Chemical Technology, Lappeenranta University of Technology, 2007

[6] Scar Ortega Lobo and Masayuki Numao. *Ordered Estimation of Missing Values*. Department of Computer Science Tokyo Institute of Technology data sets. The Journal of Systems and Software vol. 80, pp 51-62.

[7] Erel Avineri. 2005. Soft Computing Methodologies and Application. Springer-Verlag.

[8] อุมาภรณ์ สายแสงจันทร์. *การประมาณค่าสูญหายของข้อมูลที่สูญหายโดยตัวแบบฟัซซี่เนียเรส*.วิทยานิพนธ์สาขาวิทยาการคอมพิวเตอร์. มหาวิทยาลัยขอนแก่น, 2548.