ประสิทธิภาพของฟังก์ชันความเหมือนต่อขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดเค สำหรับการจำแนกประเภทข้อมูล

The Efficiency of the Similarity Function to the k-Nearest Neighbors Algorithm for Data Classification

ชนาธิป หมั่นเพียรสุข[1*] และสุพจน์ เฮงพระพรหม[2]

[1]*สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม*
[2]*หน่วยวิจัยอัจฉริยภาพแห่งเครื่องจักร สถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏนครปฐม*
[*]*kan_kan.32@hotmail.com*

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาประสิทธิภาพและเสนอวิธีการปรับปรุงการใช้ฟังก์ชันความเหมือนสำหรับการเพิ่มประสิทธิภาพของการจำแนกประเภทข้อมูลด้วยขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดเค (k-Nearest Neighbor Algorithm: KNN) โดยได้ทำการทดสอบกับชุดข้อมูลเกณฑ์มาตรฐาน (Benchmark) จำนวน 6 ชุดข้อมูล ได้แก่ ชุดข้อมูลแก้ว (Glass) ชุดข้อมูลไวน์ (Wine) ชุดข้อมูลหุบเขา (Hill-Valley) ชุดข้อมูลมะเร็งเต้านม (Wdbc) ชุดข้อมูลมะเร็งต่อมน้ำเหลือง (DLBCL) และชุดข้อมูลมะเร็งลำไส้ (Colon Cancer) การเปรียบเทียบประสิทธิภาพของฟังก์ชันความเหมือนแบบต่าง ๆ ได้ทำการแบ่งฟังก์ชั่นความเหมือนออกเป็น 2 กลุ่ม ได้แก่ 1) ฟังก์ชันการวัดระยะห่าง (Distance Metric) ประกอบด้วย ฟังก์ชั่นระยะห่างยูคลิเดียน (Euclidean) และฟังก์ชันระยะห่างแมนฮัตตัน (Manhattan) และ 2) ฟังก์ชั่นสหสัมพันธ์ (Coefficient) ประกอบด้วย ฟังก์ชันสหสัมพันธ์โคไซน (Cosine) และ ฟังก์ชันสหสัมพันธ์เพียร์สัน (Pearson) จากการทดลอง พบว่า ฟังก์ชันระยะห่างแมนฮัตตันให้ประสิทธิภาพที่ดีในกลุ่มฟังก์ชันการวัดระยะห่าง และ ฟังก์ชันสหสัมพันธ์โคไซน ให้ประสิทธิภาพดีในกลุ่มฟังก์ชันสหสัมพันธ์ ดังนั้นในงานวิจัยนี้จึงได้พัฒนาฟังก์ชั่นใหม่โดยการนำฟังก์ชันระยะห่างแมนฮัตตัน และ ฟังก์ชันสหสัมพันธ์โคไซน มาคำนวณร่วมกัน ซึ่งผลการทดลองพบว่าฟังก์ชันที่นำเสนอให้ประสิทธิภาพในการจำแนกประเภทข้อมูลที่ดีขึ้นสำหรับการจำแนกประเภทข้อมูลด้วยขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดเค

คำสำคัญ: การจำแนกประเภทข้อมูล, วิธีเพื่อนบ้านใกล้ที่สุดเค, ฟังก์ชันความเหมือน

Abstract

This research aims to study and purpose new similarity function to improve the efficiency of data classification using k-Nearest Neighbor Algorithm (KNN). Six benchmark datasets including Glass, Wine, Hill-Valley, Wdbc, DLBCL and Colon Cancer datasets are used to test. To compare the performance, we separate the similarity function into 2 groups: Distance function and Coefficient function. The Distance functions comprise of Euclidean and Manhattan, and the Coefficient functions including Cosine and Pearson. The results show that the Manhattan function yields a good performance for Distance function whereas Cosine yields a good performance for Coefficient function. So, in this research, we purpose the new similarity function which combine Manhattan and Cosine together. The results show that the purpose function gives the better performance for data classification by k-Nearest Neighbor Algorithm.

Keywords: data classification, k-nearest neighbor, similarity function

1. บทนำ

ขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดเค (k-Nearest Neighbor Algorithm: KNN) เป็นขั้นตอนวิธีการหนึ่งที่ได้รับความนิยมในศาสตร์ด้านการเรียนรู้ของเครื่องจักร (Machine Learning) และการทำเหมืองข้อมูล (Data Mining) (Xindong Wu, 2008) โดยขั้นตอนวิธีการนี้จะทำการคำนวณหาค่าความคล้ายคลึงโดยใช้ฟังก์ชันความเหมือน (Similarity Function) ระหว่างข้อมูลที่ต้องการทำนายกับชุดข้อมูลสอน (Training Data) ที่มีอยู่ เพื่อดูว่าข้อมูลที่ต้องการทำนายนั้นมีลักษณะคล้ายกับข้อมูลตัวใดมากที่สุด เค ตัว และกำหนดประเภทข้อมูลตามจำนวนเสียงส่วนใหญ่ของข้อมูล เค ตัวนั้น

ประสิทธิภาพของวิธีเพื่อนใกล้ที่สุดเค เกิดจากปัจจัยหลักปัจจัยหนึ่ง คือ ฟังก์ชันความเหมือนสำหรับคำนวณหาค่าความคล้ายคลึงของชุดข้อมูล ซึ่งฟังก์ชันหลักที่เป็นที่นิยมใช้มากที่สุดคือ การคำนวณหาระยะห่างยูคลิเดียน (Euclidean Distance) (Per-Erik Danielsson, 1980) ซึ่งเป็นการวัดค่าความห่างระหว่างจุด 2 จุดในระบบพิกัดคาร์ทีเซียน ที่มาจากทฤษฎีปีทาโกรัส ซึ่งถ้าข้อมูล 2 ตัวมีความคล้ายกันมาก จุด 2 จุด ซึ่งแทนข้อมูลแต่ละตัว จะอยู่ใกล้กันมาก จะทำให้ค่ายูคลิเดียนมีค่าน้อยเข้าใกล้ศูนย์

ได้มีการศึกษาเพื่อหาฟังก์ชันสำหรับการวัดความคล้ายคลึงกันของข้อมูลมาเป็นเวลานานจากหลายหลายวิธีการ (Lillian Lee, 1999) โดยมาจากหลากหลายแนวคิด เช่น การวัดระยะทาง (ระยะห่างยูคลิเดียน, ระยะห่างแมนฮัตตัน ฯลฯ) การวัดความคล้ายคลึงด้วย สหสัมพันธ์ (สหสัมพันธ์แบบโคไซน์: Cosine Coefficient, สหสัมพันธ์เพียร์สัน: Person's Coefficient ฯลฯ) เป็นต้น ซึ่งการวัดความคล้ายคลึงในแต่ละวิธีการก็จะมีข้อดีข้อเสียที่แตกต่างกันไป

ในการวิจัยนี้จะทำการศึกษาหาประสิทธิภาพของฟังก์ชันความเหมือนแบบต่าง ๆ ที่มีผลต่อประสิทธิภาพการจำแนกประเภทข้อมูลของขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดเคและศึกษาหาวิธีการพัฒนาประสิทธิภาพของขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดเคสำหรับปัญหาการจำแนกประเภทข้อมูลด้วยฟังก์ชันความเหมือนที่เหมาะสม

2. ทฤษฎีที่เกี่ยวข้อง

2.1 ขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดเค (k-Nearest Neighbor Algorithm: KNN)

เป็นวิธีการที่ใช้ในการจำแนกประเภทข้อมูล โดยเปรียบเทียบความคล้ายคลึงกับข้อมูลที่มีอยู่มากที่สุด เค ตัว แล้วกำหนดกลุ่มให้กับข้อมูลตัวใหม่ตามเสียงส่วนใหญ่ของสมาชิกเคตัวที่มีความใกล้เคียงที่สุดกับข้อมูลใหม่นี้ ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด เค สรุปได้ดังนี้

1) กำหนดค่าเค (นิยมกำหนดให้เป็นจำนวนคี่)

2) คำนวณความคล้ายคลึงของข้อมูลใหม่กับชุดข้อมูลตัวอย่าง

3) จัดลำดับความคล้ายคลึงและเลือกข้อมูลตัวอย่างที่มีความคล้ายคลึงมากที่สุด เค ตัว

4) พิจารณาข้อมูลตัวอย่างทั้ง เค ตัวเพื่อดูว่าแต่ละตัวถูกจัดอยู่ในกลุ่มใด

5) กำหนดกลุ่มให้กับข้อมูลตัวใหม่ด้วยกลุ่มที่มีจำนวนตัวอย่างมากที่สุดจากค่า เค ในการคำนวณค่าความคล้ายคลึงของตัวอย่าง สามารถใช้สูตรฟังก์ชันความเหมือนต่าง ๆ ที่ได้อธิบายในหัวข้อ 2.2

2.2 ฟังก์ชันความเหมือน (Similarity Function)

เป็นวิธีการวัดความคล้ายคลึงของวัตถุ 2 ตัวใด ๆ โดยทั่วไปจะมีความหมายตรงกันข้ามกับการวัดระยะห่าง (Distance Measure) Brendan J. Frey & Delbert Dueck (2007) ได้นิยามฟังก์ชันความเหมือนดังสมการ (1)

$$s = \sqrt{(X_i - X_k)^2} \qquad (1)$$

โดยที่ X_i คือ ค่าของข้อมูลตัวที่ i

X_k คือ ค่าของข้อมูลตัวที่ k

วิธีการนิยามการวัดความเหมือน มีหลากหลายวิธี ซึ่งวิธีการพื้นฐานที่นิยมใช้ มีดังต่อไปนี้

1) ระยะห่างยูคลิเดียน (Euclidean Distant) เป็นการวัดระยะห่างปกติระหว่างจุด 2 จุดในแนวเส้นตรง ซึ่งอาจวัดได้ด้วยไม้บรรทัด ที่ได้มาจากทฤษฎีพีทาโกรัส ระยะห่างยูคลิเดียน ระหว่างจุด p และ จุด q แสดงด้วย d(p,q) คำนวณได้ดังสมการ (2)

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \quad (2)$$

โดยที่ p_i และ q_i คือ จุด 2 จุดที่ต้องการคำนวณระยะห่าง

ค่า d(p,q) น้อยแสดงว่า 2 จุด p และ q มีความใกล้เคียงกันมาก (หากมีค่าเป็นศูนย์ หมายถึง ทั้ง 2 จุด คือจุดจุดเดียวกัน) แต่หากมีค่ามาก แสดงว่า 2 จุดนี้ มีความหากกัน หรือแตกต่างกันมาก

2) ระยะห่างแมนฮัตตัน (Manhattan Distant) เป็นการวัดระยะทางระหว่างจุดสองจุดตามแกนวัดมุมขวา ชื่อลอกเลียนมาจากตารางเค้าโครงของถนนในแมนฮัตตัน ซึ่งทำให้สามารถใช้เส้นทางที่สั้นที่สุดระหว่างจุดสองจุดในเมือง คำนวณได้ดังสมการ (3)

$$d = \sum_{i}^{n} \left| X_i - Y_i \right| \quad (3)$$

โดยที่ X_i และ Y_i คือจุด 2 จุดที่ต้องการคำนวณระยะห่าง

3) สหสัมพันธ์โคไซน์ (Cosine Coefficient) หรือบางครั้งเรียกว่า ความคล้ายคลึงโคไซน์ (CosineSimilarity) เป็นการวัดความคล้ายคลึงระหว่าง 2 เวคเตอร์ โดยการวัดมุมโคไซน์ของเวคเตอร์ทั้งสอง ซึ่งคำนวณได้จากสมการ (4)

$$\cos(A, B) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum^{n} (A_i)^2} \times \sqrt{\sum^{n} (B_i)^2}} \quad (4)$$

โดยที่ A_i และ B_i คือ 2 เวคเตอร์ที่ต้องการนำมาเปรียบเทียบค่าสหสัมพันธ์โคไซน์จะมีค่าอยู่ระหว่าง -1 ถึง 1 โดยมีความหมายดังนี้

ถ้าค่าเข้าใกล้ 1 หมายถึง ทั้ง 2 เวคเตอร์มีความสัมพันธ์กันมากไปในทิศทางเดียวกัน
ถ้าค่าเข้าใกล้ -1 หมายถึง ทั้ง 2 เวคเตอร์มีความสัมพันธ์กันมากไปในทิศทางตรงข้ามกัน
ถ้าค่าเข้าใกล้ 0 หมายถึง ทั้ง 2 เวคเตอร์ไม่มีความสัมพันธ์กัน

4) สหสัมพันธ์เพียร์สัน (Pearson Coefficient) เป็นวิธีที่ใช้วัดความสัมพันธ์ระหว่างตัวแปร หรือข้อมูล 2 ชุด โดยที่ตัวแปร หรือข้อมูล 2 ชุดนั้นจะต้องอยู่ในรูปของข้อมูลในมาตราอันตรภาคหรืออัตราส่วน (Interval or Ratio scale) ซึ่งคำนวณได้จากสมการ (5)

$$r_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{\left(N\sum X^2 - (\sum X)^2\right)\left(N\sum Y^2 - (\sum Y)^2\right)}} \quad (5)$$

โดยที่ r_{xy} เป็น ค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน
$\sum X$ เป็น ผลรวมของข้อมูลที่วัดได้จากตัวแปรตัวที่ 1 (X)
$\sum Y$ เป็น ผลรวมของข้อมูลที่วัดได้จากตัวแปรตัวที่ 2 (Y)
$\sum XY$ เป็น ผลรวมของผลคูณระหว่างข้อมูลตัวแปรที่ 1 และ 2
$\sum X^2$ เป็น ผลรวมของกำลังสองของข้อมูลที่วัดได้จากตัวแปรตัวที่ 1
$\sum Y^2$ เป็น ผลรวมของกำลังสองของข้อมูลที่วัดได้จากตัวแปรตัวที่ 2
N เป็น ขนาดของกลุ่มตัวอย่าง

3. การออกแบบการทดลอง

3.1 ข้อมูลที่ใช้ในการทดลอง

ข้อมูลที่ใช้ในการทดลองเป็นข้อมูลตัวเลขจำนวนจริง ประกอบด้วย 6 ชุดข้อมูล ดังนี้

1) Glass เป็นชุดข้อมูลสำหรับการจำแนกประเภทแก้ว มี 6 ประเภท ประกอบด้วยข้อมูลจำนวน 214 ตัวอย่าง โดยมี 10 คุณลักษณะ

2) wine เป็นชุดข้อมูลสำหรับการจำแนกประเภทชนิดของไวน์ ซึ่งมี 3 ประเภท ประกอบด้วยข้อมูลจำนวน 178 ตัวอย่าง โดยมี 13 คุณลักษณะ

3) wdbc เป็นชุดข้อมูลสำหรับการจำแนกประเภทโรคมะเร็งเต้านม มี 2 ประเภทประกอบด้วยข้อมูลจำนวน 569 ตัวอย่าง มี 32 คุณลักษณะ

4) Hill-Valley เป็นชุดข้อมูลสำหรับการจำแนกประเภทหุบเขา มี 1 ประเภท ประกอบด้วยข้อมูลจำนวน 606 ตัวอย่าง มี 101 คุณลักษณะ

5) Colon ข้อมูลมะเร็งลำไส้ มี 2 ประเภท ประกอบด้วยข้อมูล 62 ตัวอย่าง จากจำนวนคุณลักษณะทั้งหมด 2000 คุณลักษณะ

6) DLBCL ข้อมูลมะเร็งต่อมน้ำเหลืองกลุ่มย่อยของโรคมะเร็งต่อมน้ำเหลือง มี 2 ประเภทประกอบด้วยข้อมูล 47 ตัวอย่าง แบ่งเป็น germinal centre B-like จำนวน 24 ตัวอย่างและ activated B-like จำนวน 23 ตัวอย่าง จากจำนวนคุณลักษณะทั้งหมด 4,026 คุณลักษณะ (Alizadeh et al., 2000)

3.2 ขั้นตอนการทดลอง

ในการศึกษาประสิทธิภาพของฟังก์ชันความเหมือนต่อขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดเคสำหรับการจำแนกประเภทข้อมูลนั้น ผู้วิจัยได้ทำการใช้ฟังก์ชันความเหมือน 2 กลุ่ม ได้แก่ 1) ฟังก์ชันการวัดระยะห่าง (Distance Metric) ประกอบด้วย ฟังก์ชั่นระยะห่างยูคลิเดียน (Euclidean) และฟังก์ชันระยะห่างแมนฮัตตัน (Manhattan) และ 2) ฟังก์ชั่นสหสัมพันธ์ (Coefficient) ประกอบด้วย ฟังก์ชันสหสัมพันธ์โคไซน (Cosine) และ ฟังก์ชันสหสัมพันธ์เพียร์สัน (Pearson) จากนั้นจะหาวิธีการที่ดีที่สุดของแต่ละกลุ่ม นำมาพัฒนาเป็นฟังก์ชันใหม่เพื่อหาความคล้ายคลึงของตัวอย่าง โดยมีวิธีการดังนี้

1) นำผลของค่าที่ได้จากฟังก์ชันการวัดระยะห่าง (ที่ดีที่สุด) ทั้งหมดมาทำการแปลงให้อยู่ในรูปปกติ 0-1 (0-1 Normalization) เขียนแทนด้วย $Dist_{01}$ ค่าที่ได้ถ้าเข้าใกล้ 0 จะหมายถึงค่าที่ดีที่สุด

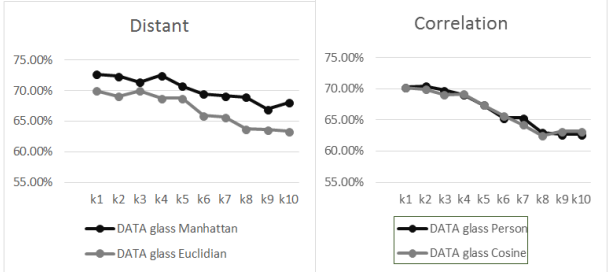
2) นำผลของค่าที่ได้จากฟังก์ชันสหสัมพันธ์ (ที่ดีที่สุด) มาลบออกจาก 1 เขียนแทนด้วย (1 - Coef) ค่าที่ได้ถ้าเข้าใกล้ 0 จะหมายถึงค่าที่ดีที่สุด

3) หาความเหมือนของข้อมูล x และ y แทนด้วย Sim(x,y) โดยการนำผลที่ได้จาก 1) และ 2) มารวมกัน ดังสมการ (6) โดยค่าที่ได้ถ้าเข้าใกล้ 0 จะหมายถึงค่าที่ดีที่สุด

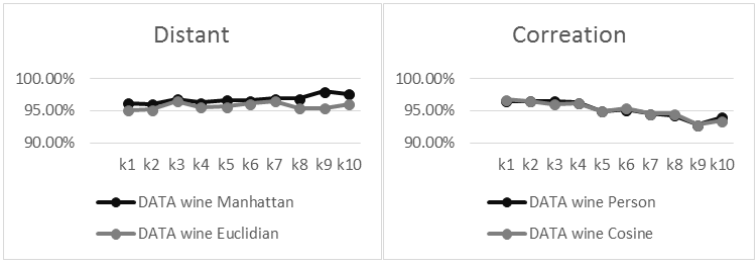
$$Sim(x,y) = Dist_{01} + (1 - Coef) \quad (6)$$

4. ผลการทดลอง

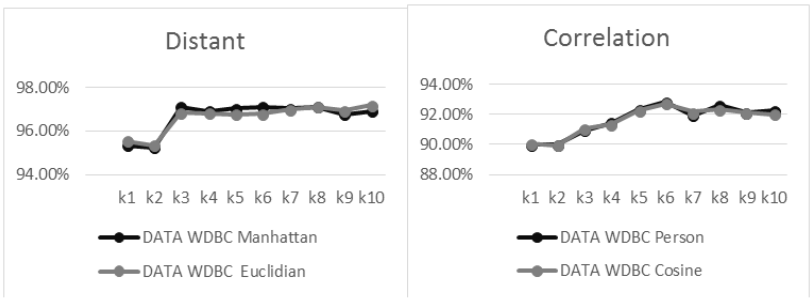
การวัดประสิทธิภาพในการทดลองนี้ ใช้วิธีการ 10-Fold Cross validation และรายงานผลด้วยค่าเฉลี่ย โดยผลการทดลองเปรียบเทียบประสิทธิภาพของฟังก์ชันความเหมือนของทั้ง 2 กลุ่มฟังก์ชัน คือ ฟังก์ชันการวัดระยะห่าง (Distant) และฟังก์ชั่นสหสัมพันธ์ (Correlation) กับทั้ง 6 ชุดข้อมูล แสดงดังภาพที่ 1 – 6



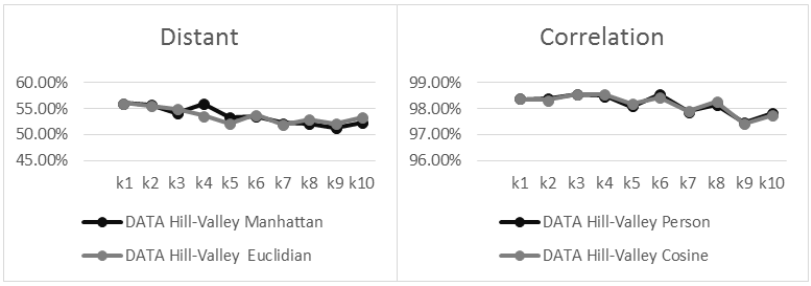
ภาพที่ 1 ผลการเปรียบเทียบความถูกต้องของการจำแนกประเภทข้อมูล Glass



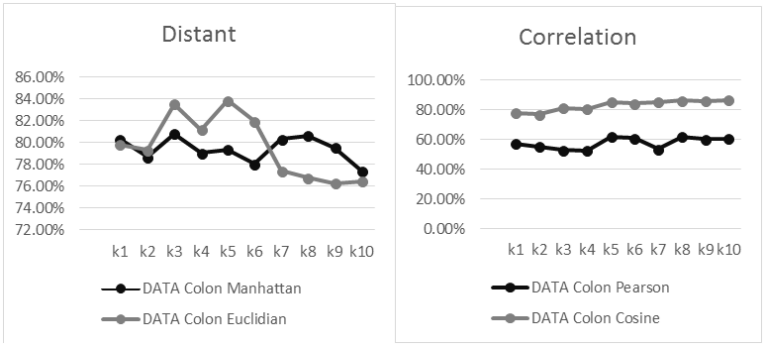
ภาพที่ 2 ผลการเปรียบเทียบความถูกต้องของการจำแนกประเภทข้อมูล Wine



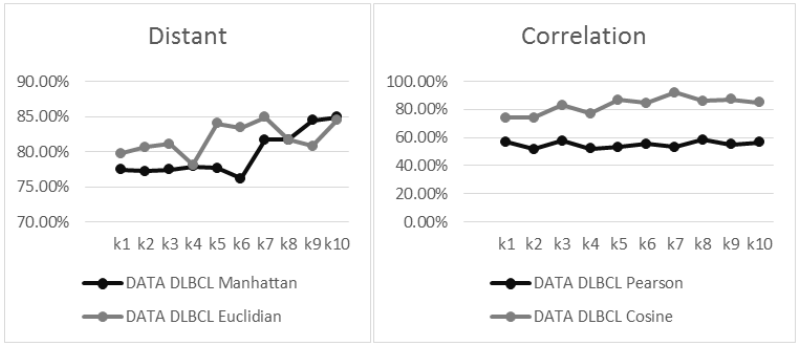
ภาพที่ 3 ผลการเปรียบเทียบความถูกต้องของการจำแนกประเภทข้อมูล WDBC



ภาพที่ 4 ผลการเปรียบเทียบความถูกต้องของการจำแนกประเภทข้อมูล Hill-Valley



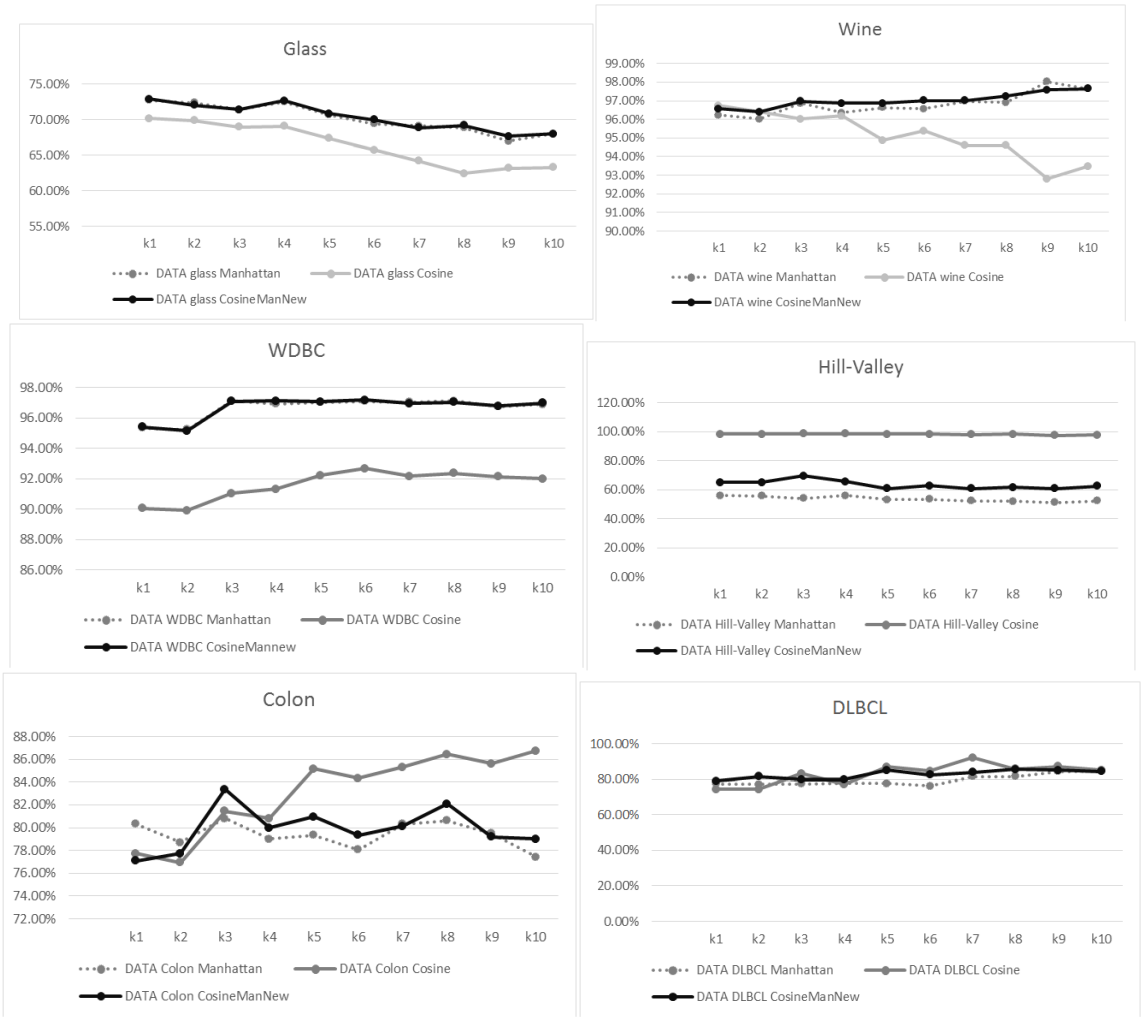
ภาพที่ 5 ผลการเปรียบเทียบความถูกต้องของการจำแนกประเภทข้อมูล Colon



ภาพที่ 6 ผลการเปรียบเทียบความถูกต้องของการจำแนกประเภทข้อมูล DLBCL

จากภาพที่ 1 – 6 แสดงให้เห็นว่า การใช้ฟังก์ชันการวัดระยะห่างนั้น ฟังก์ชันแมนฮัตตันให้ประสิทธิภาพที่ดีสำหรับข้อมูลที่มีจำนวนคุณลักษณะน้อย ๆ แต่เมื่อจำนวนคุณลักษณะเพิ่มมากขึ้น ฟังก์ชันยูคลิเดียนจะให้ประสิทธิภาพที่ดีกว่า แต่โดยภาพรวมแล้วฟังก์ชันแมนฮัตตันให้ประสิทภาพที่ดี ในขณะที่ฟังก์ชั่นสหสัมพันธ์ ถ้าข้อมูลมีจำนวนคุณลักษณะน้อย ๆ ทั้งเพียร์สันและโคไซน์ให้ประสิทธิภาพที่ไม่ต่างกัน แต่เมื่อจำนวนคุณลัณษะเพิ่มมากขึ้น โคไซน์จะให้ประสิทธิภาพที่ดีกว่า ดังนั้นในการทดลองต่อไป จึงเลือกฟังก์ชันแมนฮัตตัน และ โคไซน์มาใช้งานร่วมกัน ตามสมการที่ (6) ผลการทดลองเปรียบเทียบประสิทธิภาพแสดงดังภาพที่ 7

จากภาพ เส้นประคือการใช้ฟังก์ชันแมนฮัตตัน เส้นทึบเทาคือการใช้ฟังก์ชันโคไซน์ และ เส้นทึบดำคือการนำทั้ง 2 ฟังก์ชันมารวมกันตามวิธีการที่นำเสนอ จากผลการทดลองพบว่าให้ประสิทธิภาพที่ดีที่สุดสำหรับ 3 ชุดข้อมูลแรก (ที่มีจำนวนคุณลักษณะน้อย) ได้แก่ Glass, Wine และ WDBC ในขณะที่ชุดข้อมูลที่มีจำนวนคุณลักษณะสูงขึ้น ประสิทธิภาพของวิธีการที่นำเสนอจะไม่ดีที่สุดแต่จะไม่ให้ประสิทธิภาพที่ต่ำที่สุด



ภาพที่ 7 ผลการเปรียบเทียบความถูกต้องของวิธีการที่นำเสนอทั้ง 6 ชุดข้อมูล

5. สรุป

งานวิจัยนี้ได้ทำการศึกษาผลกระทบของการใช้ฟังก์ชันความเหมือนแบบต่าง ๆ ต่อการจำแนกประเภทข้อมูลด้วยขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดเค ประกอบด้วยการใช้ฟังก์ชันการวัดระยะห่าง และ ฟังก์ชั่นสหสัมพันธ์ โดยพบว่า ฟังก์ชันแมนฮัตตันให้ประสิทธิภาพที่ดีในกลุ่มของฟังก์ชันการวัดระยะห่าง และ ฟังก์ชันโคไซน์ให้ประสิทธิภาพที่ดีในกลุ่มของฟังก์ชันสหสัมพันธ์ โดยในการทดลองได้ทำการพัฒนาฟังก์ชันใหม่โดยการนำฟังก์ชันทั้ง 2 กลุ่มมาคำนวณร่วมกัน โดยผลการทดลองพบว่าวิธีการที่นำเสนอให้ประสิทธิภาพที่ดี โดยเฉพาะกับชุดข้อมูลที่มีจำนวนคุณลักษณะไม่มาก

6. กิตติกรรมประกาศ

งานวิจัยนี้ได้รับการสนุบสนุนทุนวิจัยโครงการวิจัยบูรณาการนักศึกษาและอาจารย์เพื่อการพัฒนาท้องถิ่นและความเป็นเลิศทางวิชาการ ปีงบประมาณ 2558 จากสถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏนครปฐม

7. เอกสารอ้างอิง

Alizadeh A.A., Eisen M.B., Davis R.E., Ma C., Lossos I.S., Rosenwald A., Boldrick J.C., Sabet H., Tran T., Yu X., Powell J.I., Yang L., Marti G.E., Moore T., J. Hudson J.JR., Lu L., Lewis D.B., Tibshirani R., Sherlock G., Chan W.C., Greiner T.C., Weisenburger D.D., Armitage J.O., Warnke R., Levy R., Wilson W., Grever M.R., Byrd J.C., Botstein D., Brown P.O. and Staudt L.M. (2000). Distinct type of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403: 503-511.

Brendan J. Frey; Delbert Dueck (2007). "Clustering by passing messages between data points". Science 315: 972–976. doi:10.1126/science.1136800

Lillian Lee. (1999) “Measures of distributional similarity.” Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, pp. 25-32, 1999.

Per-Erik Danielsson. (1980). “Euclidean distance mapping.” Computer Graphics and Image Processing, 14(3), 227–248.

Xindong Wu , Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand and Dan Steinberg. (2008). “Top 10 algorithms in data mining.” Knowledge and Information Systems, 14(1), 1-37.