ผลกระทบของฟังก์ชันความเหมือนต่อการจัดกลุ่มแบบค่าเฉลี่ยเค
The Impact of Similarity Function to K-Means Clustering

ภานุวัฒน์ สุพบุตร[1] และไก้รุ่ง เฮงพระพรหม[2*]

[1]โปรแกรมวิชาวิทยาการคอมพิวเตอร์
[2]สาขาวิชาวิทยาศาสตร์มหาบัณฑิตเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม
[*] kairung2011.heng@gmail.com

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเสนอแนวทางการปรับปรุงการใช้ฟังก์ชันความเหมือนสำหรับการหาผลกระทบของฟังก์ชันความเหมือนต่อการจัดกลุ่มแบบค่าเฉลี่ยเค (K-Means Clustering) โดยได้ทำการทดสอบกับชุดข้อมูลเกณฑ์มาตรฐาน (Benchmark) จำนวน 6 ชุดข้อมูล ได้แก่ ชุดข้อมูลแก้ว (Glass) ชุดข้อมูลไวน์ (wine) ชุดข้อมูลหุบเขา (Hill-Valley) ชุดข้อมูลมะเร็งเต้านม (WDBC) ชุดข้อมูลมะเร็งต่อมน้ำเหลือง (DLBCL) และชุดข้อมูลมะเร็งลำไส้ (Colon) โดยในการทดลองจะทำการทดลองเปรียบเทียบประสิทธิภาพของฟังก์ชันความเหมือนที่มีผลต่อประสิทธิภาพในการจัดกลุ่มแบบค่าเฉลี่ยเค แบ่งออกเป็น 2 ประเภท ได้แก่ 1) ฟังก์ชันความเหมือนที่วัดด้วยการวัดระยะห่าง ประกอบด้วย ฟังก์ชันระยะห่างยูคลิเดียน (Euclidean Distance), ฟังก์ชันระยะห่างแมนฮัตตัน (Manhattan Distance) และ 2) ฟังก์ชันความเหมือนที่วัดด้วยความคล้ายคลึงด้วยสหสัมพันธ์ ประกอบด้วย ฟังก์ชันสหสัมพันธ์แบบโคไซน์ (Cosine Coefficient), ฟังก์ชันสหสัมพันธ์เพียร์สัน (Pearson Coefficient)

จากการทดลองเปรียบเทียบประสิทธิภาพของฟังก์ชันความเหมือนที่มีผลต่อประสิทธิภาพในการจัดกลุ่มแบบค่าเฉลี่ยเค พบว่า ฟังก์ชันความเหมือนที่วัดด้วยการวัดระยะห่าง ฟังก์ชันระยะห่างแมนฮัตตันให้ประสิทธิภาพดีที่สุด และฟังก์ชั่นความเหมือนที่วัดด้วยสหสัมพันธ์ ฟังก์ชันสหสัมพันธ์แบบโคไซน์ให้ประสิทธิภาพดีที่สุด ดังนั้น ฟังก์ชัน CosMan จึงได้ถูกพัฒนาขึ้นจากฟังก์ชันสหสัมพันธ์แบบโคไซน์ และระยะห่างแมนฮัตตัน และเมื่อนำไปทดลองเปรียบเทียบประสิทธิภาพในการจัดกลุ่มแบบค่าเฉลี่ยเค ผลการทดลองที่ได้นั้น ยืนยันได้ว่าวิธีการใหม่ที่นำเสนอให้ประสิทธิภาพที่ดีที่สุด

คำสำคัญ: การจัดกลุ่มแบบค่าเฉลี่ยเค, ฟังก์ชันระยะห่างยูคลิเดียน, ฟังก์ชันระยะห่างแมนฮัตตัน, ฟังก์ชันสหสัมพันธ์แบบโคไซน์, ฟังก์ชันสหสัมพันธ์เพียร์สัน, การจัดกลุ่มข้อมูล

Abstract

The purpose of this study was to present the method to improve the impact of similarity function to K-Means clustering. Sixth data Sets from UCI Machine Learning Repository comprise Glass, Wine, Hill-Valley, WDBC, DLBCL, Colon were used in this research. To discover the similarity functions which impact to K-Means clustering. In the experiment, the similarity functions were separated to two functions to test including to: 1) similarity function with distance comprises of Euclidean Distance and Manhattan Distance. 2) Similarity functions with coefficient comprise Cosine Coefficient and Pearson Coefficient.

The results of experience shown a best similarity function with distance was Manhattan Distance and a best similarity function with coefficient was Cosine Coefficient after that a CosMan was created by Manhattan Distance and Cosine Coefficient method. And, the CosMan was compared efficient with sixth similarity function in above. The result shows that the new method still gave the best performance again.

Keywords: Euclidean distance, Manhattan distance, Cosine coefficient, Pearson coefficient, clustering

1. บทนำ

การจัดกลุ่มข้อมูล (Data Clustering) เป็นวิธีการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) เป็นงานหนึ่งในศาสตร์ด้านการทำเหมืองข้อมูล (Data Mining) โดยมีวัตถุประสงค์เพื่อพยายามค้นหารูปแบบที่ซ่อนอยู่ในชุดข้อมูลที่ไม่ทราบกลุ่ม มีลักษณะการทำงานคล้ายคลึงกับปัญหาการประมาณการความหนาแน่น (Density Estimation) ในศาสตร์ด้านสถิติ (Shecht,1990) ขั้นวิธีการหนึ่งที่ได้รับความนิยมในการจัดกลุ่มข้อมูล คือ ขั้นตอนวิธีค่าเฉลี่ยเค (K-Means Algorithm) (Wu, Kumar, Quinlan, Ghosh, Yang, Motoda, & Zhou, 2008) โดยขั้นตอนวิธีค่าเฉลี่ยเค ถูกเสนอโดย MacQueen J.B. ในปี ค.ศ. 1967 (MacQueen, 1967) และยังคงได้รับความนิยมเป็นอย่างมากในปัจจุบัน โดยขั้นตอนวิธีการนี้จะทำการแบ่งข้อมูล n ตัวใด ๆ ออกเป็นเคกลุ่ม โดยความเป็นสมาชิกของกลุ่มจะถูกกำหนดด้วยความใกล้เคียงมากที่สุดของค่าเฉลี่ยของกลุ่มนั้น ๆ ซึ่งขั้นตอนวิธีค่าเฉลี่ยเคนี้ จะเริ่มจากการสุ่มจุดศูนย์กลาง (Centroid) ของกลุ่มจำนวน k จุด และทำการกำหนดความเป็นสมาชิกให้กับข้อมูลแต่ละตัวตามความคล้ายคลึงกับจุดทั้ง k จุดที่สุ่มได้ จากนั้นจะทำการคำนวณค่าเฉลี่ยของกลุ่มใหม่ โดยใช้ค่าของสมาชิกของแต่ละกลุ่มแทน และวนซ้ำหาความเป็นสมาชิกใหม่ ซึ่งจะทำซ้ำไปเรื่อย ๆ จนกระทั่งความเป็นสมาชิกของกลุ่มไม่มี การเปลี่ยนแปลงหรือเปลี่ยนแปลงน้อยมากจนพอยอมรับได้

ประสิทธิภาพของขั้นตอนวิธีค่าเฉลี่ยเค เกิดจากปัจจัยหลักปัจจัยหนึ่ง คือ ฟังก์ชันความเหมือนสำหรับคำนวณหาค่าความคล้ายคลึงของชุดข้อมูล ซึ่งฟังก์ชันหลักที่เป็นที่นิยมใช้มากที่สุด คือ การคำนวณหาระยะห่างยูคลิเดียน (Danielsson,1980) ซึ่งเป็นการวัดค่าความห่างระหว่างจุด 2 จุดในระบบพิกัดคาร์ทีเซียน ที่มาจากทฤษฎีปีทาโกรัสซึ่งถ้าข้อมูล 2 ตัวมีความคล้ายกันมาก จุด 2 จุด ซึ่งแทนข้อมูลแต่ละตัว จะอยู่ใกล้กันมาก จะทำให้ค่ายูคลิเดียนมีค่าน้อยเข้าใกล้ศูนย์ได้มีการศึกษาเพื่อหาฟังก์ชันสำหรับการวัดความคล้ายคลึงกันของข้อมูลมาเป็นเวลานานจากหลายหลายวิธีการ (Lee, 1999) โดยมาจากหลากหลายแนวคิด เช่น การวัดระยะทาง (ระยะห่างยูคลิเดียน, ระยะห่างแมนฮัตตัน ฯลฯ) การวัดความคล้ายคลึงด้วย สหสัมพันธ์ (สหสัมพันธ์แบบโคไซน์ และ สหสัมพันธ์เพียร์สัน ฯลฯ) เป็นต้น ซึ่งการวัดความคล้ายคลึงในแต่ละวิธีการก็จะมีข้อดีข้อเสียที่แตกต่างกันไป

ในงานวิจัยนี้ จึงได้ทำการศึกษาผลกระทบของฟังก์ชันความเหมือนแบบต่าง ๆ ที่มีผลต่อการจัดกลุ่มข้อมูลด้วยขั้นตอนวิธีค่าเฉลี่ยเค กับข้อมูลที่มีลักษณะแตกต่างกัน ได้แก่ ความแตกต่างของจำนวนกลุ่มและความแตกต่างของจำนวนคุณลักษณะ จากนั้นจะได้ทำการศึกษาหาวิธีพัฒนาประสิทธิภาพของขั้นตอนวิธีการจัดกลุ่มข้อมูลที่มีประสิทธิภาพต่อไป

2. วัตถุประสงค์ของการวิจัย

2.1เพื่อศึกษาประสิทธิภาพของฟังก์ชันความเหมือนแบบต่าง ๆ ที่มีผลต่อการจัดกลุ่มข้อมูลด้วยขั้นตอนวิธีค่าเฉลี่ยเค

2.2เพื่อเสนอแนวทางการปรับปรุงประสิทธิภาพของขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบค่าเฉลี่ยเค ด้วยการใช้ฟังก์ชันความเหมือนที่เหมาะสม

3. ทฤษฎีที่เกี่ยวข้อง

3.1 ขั้นตอนวิธีค่าเฉลี่ยเค (K-Means Algorithm)

การจัดกลุ่มข้อมูลแบบค่าเฉลี่ยเค (Danielsson, 1980) จะเริ่มด้วยการกำหนดจำนวนกลุ่ม k จากนั้นจะทำการกำหนดจุดศูนย์กลางของกลุ่ม (Centroid) แต่ละกลุ่มแบบสุ่ม และทำการจัดข้อมูลแต่ละตัวเข้าไปในกลุ่มที่มีระยะห่างระหว่างจุดศูนย์กลางกลุ่มกับข้อมูลตัวนั้นที่มีค่าระยะห่างน้อยที่สุด โดยใช้ฟังก์ชันการวัดความเหมือนใด ๆ ซึ่งฟังก์ชันที่นิยมใช้เป็นพื้นฐานมากที่สุดตัวหนึ่ง ได้แก่ การวัดระยะห่างยูคลิเดียน ซึ่งคำนวณได้จากสมการ (1)

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

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

d คือ จำนวนคุณลักษณะทั้งหมดของตัวอย่าง

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

หลังจากจัดกลุ่มข้อมูลให้กับตัวอย่างจนครบทุกตัวแล้ว ค่าจุดศูนย์กลางของกลุ่มจะถูกคำนวณใหม่โดยใช้ข้อมูลตัวอย่างในกลุ่มของตัวเอง และจะทำการจัดกลุ่มให้กับข้อมูลตัวอย่างแต่ละตัวใหม่ ซึ่งจะทำเช่นนี้ซ้ำไปเรื่อย ๆ จนกระทั่งพบเงื่อนไขของการสิ้นสุด เช่น สมาชิกของแต่ละกลุ่มไม่มีการเปลี่ยนแปลง หรือครบตามจำนวนรอบสูงสุดที่กำหนด

การจัดกลุ่มข้อมูลแบบค่าเฉลี่ยเค

ข้อมูลนำเข้า : ชุดข้อมูล x(a_1 , a_2 , ... , a_n) , ฟังก์ชันวัดความเหมือน , จำนวนรอบ

1. กำหนดจำนวนกลุ่ม K
2. กำหนดจุดศูนย์กลาง (Centroid) ของแต่ละกลุ่มแบบสุ่ม
3. เริ่มทำซ้ำ
   - 3.1 คำนวณความเหมือนของตัวอย่างแต่ละตัวกับจุดศูนย์กลางของแต่ละกลุ่มด้วยฟังก์ชันวัดความเหมือน
   - 3.2 กำหนดให้ตัวอย่างแต่ละตัวเป็นสมาชิกของกลุ่มที่มีความเหมือนกับจุดศูนย์กลางของกลุ่มนั้นมากที่สุด
   - 3.3 คำนวณหาค่าจุดศูนย์กลางของแต่ละกลุ่มใหม่ด้วยข้อมูลสมาชิกของกลุ่มนั้น ๆ
4. วนซ้ำจนกระทั่งพบเงื่อนไขการสิ้นสุด

ผลลัพธ์ : สมาชิกของแต่ละกลุ่ม

3.2 ฟังก์ชันความเหมือน (Similarity Function)

ฟังก์ชันความเหมือน (Similarity Function) หรือ การวัดความเหมือน (Similarity Measure) เป็นวิธีการวัดความคล้ายคลึงของวัตถุ 2 ตัวใด ๆ โดยทั่วไปจะมีความหมายตรงกันข้ามกับการวัดระยะห่าง (Distance Measure) (Frey & Dueck, 2007)

วิธีการนิยามการวัดความเหมือนมีหลากหลายวิธี ซึ่งวิธีการพื้นฐานที่นิยมใช้มีดังต่อไปนี้

3.2.1 ระยะห่างยูคลิเดียน (Euclidean Distance)

ระยะห่างยูคลิเดียน (Danielsson, 1980) เป็นการวัดระยะห่างปกติระหว่างจุด 2 จุดในแนวเส้นตรง ซึ่งอาจวัดได้ด้วยไม้บรรทัด ที่ได้มาจากทฤษฎีพีทาโกรัส ระยะห่างยูคลิเดียน ระหว่างจุด p และ จุด q แสดงด้วย d(p,q) คำนวณได้ดังสมการ (2)

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

โดยที่ p = {p_1 , p_2 , p_3 , ... , p_n} และ q = {q_1 , q_2 , q_3 , ..., q_n} คือ จุด 2 จุดที่ต้องการคำนวณระยะห่าง

ค่า $d(p,q)$ น้อยแสดงว่า 2 จุด p และ q มีความใกล้เคียงกันมาก (หากมีค่าเป็นศูนย์ หมายถึง ทั้ง 2 จุด คือจุดเดียวกัน) แต่หากมีค่ามาก แสดงว่า 2 จุดนี้ มีความหากันหรือ แตกต่างกันมาก

3.2.2 ระยะห่างแมนฮัตตัน (Manhattan Distance)

ระยะทางแมนฮัตตัน (Chang, 2009) เป็นการวัดระยะทางระหว่างจุดสองจุดตามแนวแกนตั้งฉากสองมิติระหว่างตำแหน่งสองตำแหน่งซึ่งลอกเลียนมาจากตารางเค้าโครงของถนนในแมนฮัตตัน ซึ่งทำให้รถสามารถใช้เส้นทางที่สั้นที่สุดระหว่างจุดสองจุดในเมือง คำนวณได้ดังสมการ (3)

$$d_1(p,q)=\sum_{i}^{n}\left|p_i-q_i\right| \qquad (3)$$

โดยที่ p_i และ q_i คือ จุด 2 จุดที่ต้องการคำนวณระยะห่าง

3.2.3 สหสัมพันธ์โคไซน์ (Cosine Coefficient)

สหสัมพันธ์โคไซน์ (Cho & Won, 2003) หรือบางครั้งเรียกว่า ความคล้ายคลึงโคไซน์ (Cosine Similarity) เป็นการวัดความคล้ายคลึงระหว่าง 2 เวคเตอร์ โดยการวัดมุมโคไซน์ของเวคเตอร์ทั้งสอง ซึ่งคำนวณได้จากสมการ (4)

$$Co\sin e = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum^{n} (A_i)^2} \times \sqrt{\sum^{n} (B_i)^2}} \qquad (4)$$

โดยที่ A = {a_1, a_2, a_3, ... , a_n} และ B = {b_1, b_2, b_3, ..., b_n} คือ 2 เวคเตอร์ที่ต้องการนำมาเปรียบเทียบ
ค่าสหสัมพันธ์โคไซน์จะมีค่าอยู่ระหว่าง -1 ถึง 1 โดยมีความหมายดังนี้
ถ้าค่าเข้าใกล้ 1 หมายถึง ทั้ง 2 เวคเตอร์มีความสัมพันธ์กันมากไปในทิศทางเดียวกัน
ถ้าค่าเข้าใกล้ -1 หมายถึง ทั้ง 2 เวคเตอร์มีความสัมพันธ์กันมากไปในทิศทางตรงข้ามกัน
ถ้าค่าเข้าใกล้ 0 หมายถึง ทั้ง 2 เวคเตอร์ไม่มีความสัมพันธ์กัน

3.2.4 สหสัมพันธ์เพียร์สัน (Pearson Coefficient)

เป็นวิธีที่ใช้วัดความสัมพันธ์ระหว่างตัวแปร (ปารุสก์ บุญพร, พงษ์ศักดิ์ ติยานันท, และ และ สุปิยา เจริญศิริวัฒน์, 2009) หรือข้อมูล 2 ชุด โดยที่ตัวแปร หรือข้อมูล 2 ชุดนั้นจะต้องอยู่ในรูปของข้อมูลในมาตราอันตรภาคหรืออัตราส่วน (Interval or Ratio Scale) ซึ่งคำนวณได้จากสมการ (5)

$$r_{xy} = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{\left[n\sum x_i^{\,2} - (\sum x_i)^2\right]\left[n\sum y_i^{\,2} - (\sum y_i)^2\right]}} \qquad (5)$$

โดยที่	r_{xy}	เป็น ค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน
$\sum x_i$		เป็น ผลรวมของข้อมูลที่วัดได้จากตัวแปรตัวที่ 1 (X)
$\sum y_i$		เป็น ผลรวมของข้อมูลที่วัดได้จากตัวแปรตัวที่ 2 (Y)
$\sum xy$		เป็น ผลรวมของผลคูณระหว่างข้อมูลตัวแปรที่ 1 และ 2
$\sum x_i^{\,2}$		เป็น ผลรวมของกำลังสองของข้อมูลที่วัดได้จากตัวแปรตัวที่ 1
$\sum y_i^{\,2}$		เป็น ผลรวมของกำลังสองของข้อมูลที่วัดได้จากตัวแปรตัวที่ 2
n		เป็น ขนาดของกลุ่มตัวอย่าง

4. วิธีการดำเนินการวิจัย

4.1 การศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้อง

สำหรับขั้นตอนนี้ผู้วิจัยได้ทำการศึกษา รวบรวมวรรณกรรม บทความวิจัยทั้งในและต่างประเทศ ทำการศึกษาทฤษฎีต่าง ๆ จากตำรา เอกสาร สิ่งพิมพ์ ตลอดจนการค้นหาข้อมูลออนไลน์สำหรับใช้ในการวิจัย รวมถึงเทคนิคการทำเหมืองข้อมูลและการเรียนรู้ของเครื่อง โดยเฉพาะขั้นตอนวิธีการจัดกลุ่มแบบค่าเฉลี่ยเค และฟังก์ชันความเหมือนแบบต่าง ๆ

4.2 รวบรวมข้อมูลและวิเคราะห์ปัญหา

การวิจัยนี้ ได้ทำการรวบรวมข้อมูลเกณฑ์มาตรฐาน (Benchmark) จากเว็ปไซต์ http://archive.ics.uci.edu/ml/ (UCI Machine Learning Repository) สำหรับปัญหาการจำแนกประเภทของมูลที่มีจำนวนกลุ่มและจำนวนคุณลักษณะข้อมูลแตกต่างกัน สำหรับเป็นข้อมูลทดสอบประสิทธิภาพประกอบด้วย 6 ชุดข้อมูล แบ่งตามจำนวนลักษณะข้อมูล 3 แบบดังนี้

4.2.1 ข้อมูล หลักสิบ คุณลักษณะ ได้แก่ ข้อมูล Glass (D1) เป็นชุดข้อมูลสำหรับการจำแนกประเภทแก้ว มี 6 ประเภท ประกอบด้วยข้อมูลจำนวน 214 ตัวอย่าง โดยมี 10 คุณลักษณะ เป็นข้อมูลตัวเลขจำนวนจริง และข้อมูล Wine (D2) เป็นชุดข้อมูลสำหรับการจำแนกประเภทชนิดของไวน์ มี 3 ประเภท ประกอบด้วยข้อมูลจำนวน 178 ตัวอย่าง โดยมี 13 คุณลักษณะ เป็นข้อมูลตัวเลขจำนวนจริง

4.2.2 ข้อมูล หลักร้อย คุณลักษณะ ได้แก่ ข้อมูล Hill-Valley (D3) เป็นชุดข้อมูลสำหรับการจำแนกประเภทหุบเขา มี 2 ประเภทประกอบด้วยข้อมูลจำนวน 606 ตัวอย่าง โดยมี 101 คุณลักษณะ เป็นข้อมูลตัวเลขจำนวนจริง และข้อมูล WDBC (D4) เป็นชุดข้อมูลสำหรับการจำแนกประเภทโรคมะเร็งเต้านม มี 2 ประเภทประกอบด้วยข้อมูลจำนวน 569 ตัวอย่าง โดยมี 32 คุณลักษณะ เป็นข้อมูลตัวเลขจำนวนจริง

4.2.3 ข้อมูล หลักพัน คุณลักษณะ ได้แก่ ข้อมูล Colon (D5) ข้อมูลมะเร็งลำไส้ มี 2 ประเภท ประกอบด้วยข้อมูลจำนวน 62 ตัวอย่าง โดยมี 2000 คุณลักษณะ เป็นข้อมูลตัวเลขจำนวนจริง และข้อมูล DLBCL (D6) ข้อมูลมะเร็งต่อมน้ำเหลืองกลุ่มย่อยของโรคมะเร็งต่อมน้ำเหลือง มี 2 ประเภท ประกอบด้วยข้อมูลจำนวน 47 ตัวอย่าง แบ่งเป็น Germinal Centre B-like จำนวน 24 ตัวอย่างและ Activated B-like จำนวน 23 ตัวอย่าง โดยมี 4,026 คุณลักษณะ เป็นข้อมูลตัวเลขจำนวนจริง

4.3 ออกแบบระบบ

ในการออกแบบการทดลองการศึกษาผลกระทบของฟังก์ชันความเหมือนต่อการจัดกลุ่มแบบค่าเฉลี่ยเค นั้นได้ออกแบบความต้องการสำหรับการพัฒนาระบบไว้ดังนี้

4.3.1 ด้านฮาร์ดแวร์ (Hardware) ของเครื่องที่ใช้ในการพัฒนาฟังก์ชั่นและการพัฒนาระบบ ประกอบด้วย 1) Notebook 1 เครื่อง รุ่น ASUS 5K4LY 2) Memory: 4 GB 3) Processor: Inter ® Core ™ i3 – 2310M CPU @ 2.10 GHz

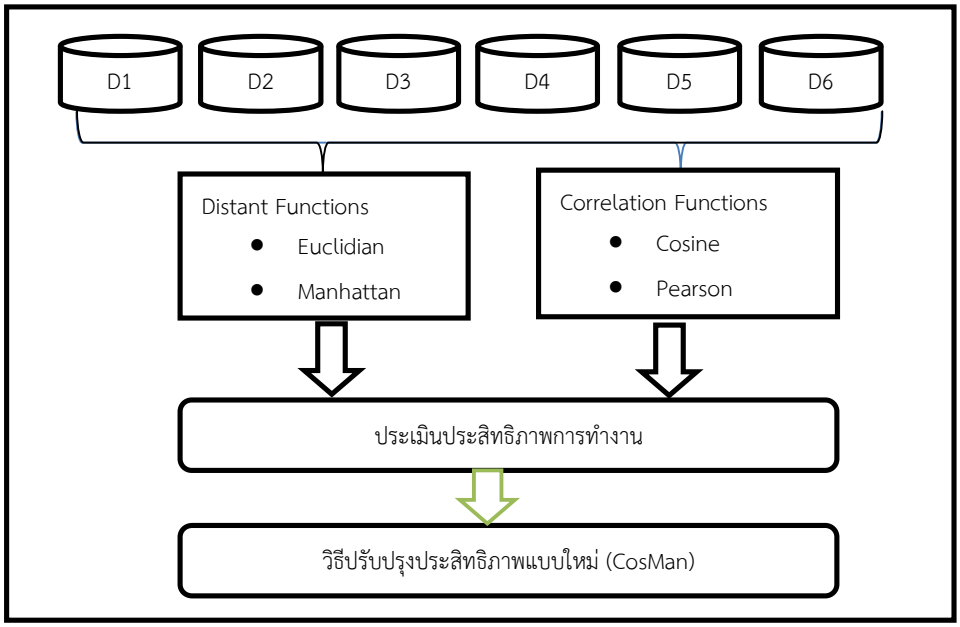
4.3.2 ด้านซอฟต์แวร์ (Software) ของเครื่องที่ใช้ในการพัฒนาฟังก์ชั่น ได้แก่ 1) ระบบปฏิบัติการ Microsoft ® Windows 7 2) โปรแกรม Dev-C++ 5.11 TDM-GCC 4.9.2

4.3.3 การออกแบบการพัฒนาระบบ ขั้นตอนการออกแบบการทำงานของระบบ จะแสดงในภาพที่ 3.1

1) เมื่อทำการวิเคราะห์ประสิทธิภาพของแต่ละฟังก์ชันแล้ว จะนำฟังก์ชันความเหมือนที่วัดด้วยระยะห่าง และ ฟังก์ชันความเหมือนที่วัดด้วยความคล้ายคลึงด้วยสหสัมพันธ์ของแต่ละชนิดที่มีประสิทธิภาพที่ดีที่สุดมาพัฒนาเป็นฟังก์ชันใหม่ขึ้นมา โดยฟังก์ชันที่ดีที่สุดของฟังก์ชันความเหมือนที่วัดด้วยระยะห่าง คือ แมนฮัตตัน และฟังก์ชั่นความเหมือนที่วัดด้วยความคล้ายคลึงด้วยสหสัมพันธ์ คือ โคไซน์

2) ทำการปรับปรุงฟังก์ชันความเหมือนสำหรับการจัดกลุ่มแบบค่าเฉลี่ยเค โดยการนำฟังก์ชันสหสัมพันธ์โคไซน์ และ ฟังก์ชั่นระยะห่างแมนฮัตตัน

3) จากนั้นนำวิธีการใหม่ที่ได้จากการทดลอง ไปทำการทดสอบหาประสิทธิภาพของการทำงาน โดยทำการทดลองเปรียบเทียบประสิทธิภาพการทำงานเทียบกับฟังก์ชันความเหมือนที่วัดด้วยระยะห่าง ได้แก่ ระยะห่างยูคลิเดียนและฟังก์ชันระยะห่างแมนฮัตตัน และ ฟังก์ชันความเหมือนที่วัดด้วยความคล้ายคลึงด้วยสหสัมพันธ์ของแต่ละชนิด ได้แก่ ฟังก์ชันสหสัมพันธ์โคไซน์และสหสัมพันธ์เพียร์สัน



รูปที่ 3.1 แผนผังขั้นตอนการทดสอบประสิทธิภาพของฟังก์ชันความเหมือนต่อการจัดกลุ่มแบบค่าเฉลี่ยเค

4.4 พัฒนาระบบ

การพัฒนาระบบปั่นขั้นตอนการเขียนโปรแกรมเพื่อศึกษาผลกระทบของฟังก์ชันความเหมือนต่อการจัดกลุ่มแบบค่าเฉลี่ยเค ผู้วิจัยได้ใช้เครื่องมือในการพัฒนาโปรแกรม คือ Dev-C++ 5.11 TDM-GCC 4.9.2 และใช้ภาษาซี เป็นเครื่องมือที่ช่วยในการพัฒนาชุดคำสั่ง

4.5 ทดสอบและการประเมินประสิทธิภาพในการทดลอง

ในการทดสอบผู้พัฒนาได้ออกเป็น 2 ขั้นตอนหลัก คือ Black-Box Testing และ White-Box Testing และ ทำการประเมินประสิทธิภาพด้วยความแม่นยำในการจำแนก

5. ผลการทดลอง

5.1 ผลการทดสอบประสิทธิภาพกับข้อมูลคุณลักษณะหลักสิบ

ผลการทดลองจากตารางที่ 1 แสดงให้เห็นว่าวิธีการ CosMan ให้ผลลัพธ์ที่ดีกว่าในทุกชุดข้อมูล คือ ชุดข้อมูล Glass ให้ประสิทธิภาพในการจัดกลุ่มที่ 60.60$\pm$1.97 ซึ่งดีกว่าวิธีการอื่นๆ เมื่อเทียบในชุดข้อมูลเดียวกัน ส่วนชุดข้อมูล Wine วิธีการใหม่ที่นำเสนอ (CosMan) ยังให้ประสิทธิภาพที่ดีกว่าทุกวิธีที่ทดสอบ คือ ให้ตัวเลขประสิทธิภาพที่ 91.85$\pm$2.89 ในชุดข้อมูลเดียวกัน และชุดข้อมูลของ WDBC ยังยืนยันผลการทดลองว่า วิธีการใหม่ที่นำเสนอให้ประสิทธิภาพที่ดีที่สุด คือ 92.72$\pm$1.30 เมื่อเทียบการทดลองกับทุกวิธีในชุดข้อมูลเดียวกัน

ตารางที่ 1 การทดสอบประสิทธิภาพกับข้อมูลคุณลักษณะหลักสิบ

ชุดข้อมูล	Distance		Correlation		New Technique
	Euclidian	Manhattan	Cosine	Pearson	CosMan
Glass	54.95$\pm$4.20	55.19$\pm$3.22	55.47$\pm$2.57	54.97$\pm$3.43	60.60$\pm$1.97
Wine	89.21$\pm$10.92	95.06$\pm$1.89	87.98$\pm$2.82	82.92$\pm$7.97	91.85$\pm$2.89
WDBC	92.65$\pm$0.39	92.06$\pm$0.71	78.82$\pm$1.43	78.40$\pm$1.10	92.72$\pm$1.30

5.2 การทดสอบประสิทธิภาพกับข้อมูลคุณลักษณะหลักร้อย

ส่วนตารางที่ 2 เป็นผลการทดลองที่ได้ทำการทดสอบประสิทธิภาพกับข้อมูลคุณลักษณะหลักร้อย คือ ชุดข้อมูล Hill-Valley โดยมี 101 คุณลักษณะ หลังจากการทดลองพบว่า ประสิทธิภาพในการทดลองเมื่อเทียบกับวิธีการอื่นๆ ที่ทำการทดลองในชุดข้อมูลเดียวกัน วิธีการใหม่ที่นำเสนอ (CosMan) ให้ประสิทธิภาพที่น้อยกว่าฟังก์ชั่นความเหมือนที่วัดด้วยความคล้ายคลึงเชิงสหสัมพันธ์ แต่มากกว่าฟังก์ชันความเหมือนที่วัดด้วยการวัดระยะห่าง

ตารางที่ 2 การทดสอบประสิทธิภาพกับข้อมูลคุณลักษณะหลักร้อย

ชุดข้อมูล	Distance		Correlation		New Technique
	Euclidian	Manhattan	Cosine	Pearson	CosMan
Hill-Valley	51.06±0.37	51.25±0.40	56.19±4.18	57.62±4.17	51.92±0.39

5.3 การทดสอบประสิทธิภาพกับข้อมูลคุณลักษณะหลักพัน

ตารางที่ 3 ซึ่งเป็นผลการทดลองที่ได้ทำการทดสอบประสิทธิภาพกับข้อมูลคุณลักษณะหลักพัน ประกอบด้วย ชุดข้อมูล Colon และ ชุดข้อมูล DLBCL หลังจากการทดลองพบว่า ในชุดข้อมูล Colon ประสิทธิภาพในการทำงานของฟังก์ชั่นความเหมือนที่วัดด้วยความคล้ายคลึงด้วยสหสัมพันธ์ให้ประสิทธิภาพที่ดีที่สุด รองลงมาคือ วิธีการใหม่ที่นำเสนอ ส่วนฟังก์ชันความเหมือนที่วัดด้วยระยะห่างเป็นวิธีการที่ให้ประสิทธิภาพที่น้อยที่สุด ส่วนการทดลองในชุดข้อมูล DLBCL พบว่า ฟังก์ชั่นความเหมือนที่วัดด้วยความคล้ายคลึงด้วยสหสัมพันธ์ ให้ประสิทธิภาพที่ดีที่สุด (Cosine) และในขณะเดียวกันก็ให้ประสิทธิภาพที่แย่ที่สุด (Pearson) เช่นกัน

ตารางที่ 3 การทดสอบประสิทธิภาพกับข้อมูลคุณลักษณะหลักพัน

ชุดข้อมูล	Distance		Correlation		New Technique
	Euclidian	Manhattan	Cosine	Pearson	CosMan
Colon	64.52±0	64.52±0	79.35±8.46	76.13±8.29	69.35±3.80
DLBCL	64.25±8.90	61.70±9.78	78.30±7.29	51.06±0	62.55±5.13

6. สรุปผลการทดลอง

งานวิจัยนี้ได้ทำการศึกษาหาประสิทธิภาพของฟังก์ชันความเหมือนแบบต่าง ๆ ที่มีผลต่อการจัดกลุ่มแบบค่าเฉลี่ยเค และศึกษาหาวิธีการพัฒนาประสิทธิภาพของการจัดกลุ่มแบบค่าเฉลี่ยเค โดยในการทดลองได้ทำการแบ่งฟังก์ชันความเหมือนออกเป็น 2 ประเภท ได้แก่ 1) ฟังก์ชันความเหมือนที่วัดด้วยการวัดระยะห่าง ประกอบด้วย ฟังก์ชั่นระยะห่างยูคลิเดียน, ฟังก์ชันระยะห่างแมนฮัตตัน และ 2) ฟังก์ชั่นความเหมือนที่วัดด้วยความคล้ายคลึงด้วยสหสัมพันธ์ ประกอบด้วย ฟังก์ชันสหสัมพันธ์แบบโคไซน์ ฟังก์ชันสหสัมพันธ์เพียร์สัน กับข้อมูลเกณฑ์มาตรฐาน (Benchmark) จำนวน 6 ชุด ได้แก่ Glass, wine, WDBC, Hill-Valley , Colon และ DLBCL ซึ่งคัดลอกมาจากเว็บไซต์ของ UCI Machine Learning Repository

จากการทดลองเปรียบเทียบประสิทธิภาพของฟังก์ชันความเหมือนที่มีผลต่อประสิทธิภาพในการจัดกลุ่มแบบค่าเฉลี่ยเค พบว่า ฟังก์ชันความเหมือนที่วัดด้วยการวัดระยะห่าง ฟังก์ชันระยะห่างแมนฮัตตัน ให้ประสิทธิภาพดีที่สุด และฟังก์ชั่นความเหมือนที่วัดด้วยความคล้ายคลึงด้วยสหสัมพันธ์ ฟังก์ชันสหสัมพันธ์แบบโคไซน์ ให้ประสิทธิภาพดีที่สุด ดังนั้น ฟังก์ชันใหม่จึงได้ถูกพัฒนาขึ้นร่วมกันระหว่างฟังก์ชันสหสัมพันธ์แบบโคไซน์ และฟังก์ชันระยะห่างแมนฮัตตัน ซึ่งเมื่อนำไปทดลองเปรียบเทียบประสิทธิภาพในการจัดกลุ่มแบบค่าเฉลี่ยเค ผลการทดลองพบว่า วิธีการใหม่ที่นำเสนอให้ประสิทธิภาพที่ดีกับข้อมูลที่มีจำนวนคุณลักษณะน้อย ๆ คือ หลักสิบ แต่เมื่อทำการทดลองกับชุดข้อมูลที่มีจำนวนคุณลักษณะมากขึ้น พบว่า ประสิทธิภาพของการทำงานจะลดลง

7. เอกสารอ้างอิง

Chang, D. J., Desoky, A. H., Ouyang, M., & Rouchka, E. C. (2009, May). Compute pairwise manhattan distance and pearson correlation coefficient of data points with gpu. **In Software Engineering, Artificial Intelligences, Networking and Parallel/Distributed Computing, 2009. SNPD'09. 10th ACIS International Conference on IEEE (pp. 501-506).**

Cho, S. B., & Won, H. H. (2003, January). Machine learning in DNA microarray analysis for cancer classification. **In Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003-Volume 19 (pp. 189-198)**. Australian Computer Society, Inc..

Danielsson, P. E. (1980). Euclidean distance mapping. **Computer Graphics and image processing**, 14(3), 227-248.

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. **science**, 315(5814), 972-976.

Lee L. (1999). Measures of distributional similarity. **In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (ACL '99)** (pp. 25-32). Association for Computational Linguistics.

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. **In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability** (Vol. 1, No. 14, pp. 281-297).

Shecht D.F. (1990). Probabilistic neural networks. **Neural Networks**, *3*(1), 109-118.

Wu, X., et al. (2008). Top 10 algorithms in data mining. **Knowledge and information systems**, 14(1), pp. 1-37.

ปารุสก์ บุญพร, พงษ์ศักดิ์ ติยานันท์, และ และ สุปิยา เจริญศิริวัฒน์. (2009). การจำแนกกลุ่มขนาดรูปร่างหญิงไทยโดยใช้เทคนิค K-means Clustering. **The 6th International Joint Conference on Computer Science and Software Engineering (JCSSE 2009).** ภูเก็ต, ประเทศไทย.