

ขั้นตอนวิธีในการจัดกลุ่มข้อมูลด้วยเคฮาร์โมนิกมีนแบบทางเลือก

Alternative K-Harmonic Means Clustering Algorithms

กนิษฐา ชื่นตา (Kanittha Chuenta)* สิริภัทร เชื้อชวนวัฒนา (Sirapat Chiewchanwattana)**

คำรณ สุณีดี (Kamron Sunut)***

บทคัดย่อ

การจัดกลุ่มข้อมูลใช้ความคล้ายกันของข้อมูลและขั้นตอนการเรียนรู้แบบไม่มีผู้สอน วิธีพื้นฐานวิธีหนึ่งคือการจัดกลุ่มข้อมูลแบบ K-Mean (KM) ขั้นตอนวิธี K-Harmonic Means (KHM) พัฒนาจาก KM โดยเปลี่ยนการกำหนดกลุ่มจากการพิจารณาค่าระยะห่างน้อยสุดเป็นการหาค่าเฉลี่ยแบบฮาร์โมนิก สามารถเพิ่มประสิทธิภาพ และลดความไวเรื่องค่ากลางเริ่มต้น การได้ค่ากลางที่เหมาะสมที่สุดแบบเฉพาะที่ใน KM และ Fuzzy C-Means (FCM) เป็นอีกหนึ่งขั้นตอนวิธีที่พัฒนาจาก KM จัดกลุ่มข้อมูลจากค่าความเป็นสมาชิกของของกลุ่มของแต่ละข้อมูล ทั้งสามขั้นตอนวิธีที่กล่าวมานั้นมีจุดด้อยคือไม่สามารถจัดการกับข้อมูลรบกวนได้ มีงานวิจัยที่เสนอการเปลี่ยนการวัดความคล้ายจากมาตรวัดยูคลิดเป็นการวัดด้วยฟังก์ชันการเพิ่มขึ้นแบบต่อเนื่องเรียกว่า Alternative ทำให้ Alternative Hard C-Means (AHCM) และ Alternative Fuzzy C-Means (AFCM) แก้ปัญหาข้อมูลรบกวนได้ ดังนั้นในงานวิจัยนี้จึงนำการวัดความใกล้เคียงด้วยฟังก์ชันการเพิ่มขึ้นแบบต่อเนื่องมาใช้กับ KHM เรียกว่า Alternative K-Harmonic Means (AKHM) และให้ผลการจัดกลุ่มข้อมูลที่ดีกว่าเดิม

ABSTRACT

The K-Means (KM) algorithm is the basic method for clustering algorithms using the similarity of data to perform an unsupervised, learning for finding clusters. The K-Harmonic Means (KHM) algorithm change the minimum distance from cluster centers to harmonic means. Therefore, the performance of KHM is better than KM and less sensitive for the initialization of centers. In addition, the local optimum problem is also less than those of KM. The Fuzzy C-Means (FCM) algorithm find clusters by considering the maximum membership values of each data referring to the cluster centers. All three algorithms can not manage with the noisy data or some data which lying far from the other, the outlier data. Some reports suggested that Alternative Hard C Means (AHCM) and Alternative Fuzzy C-Means (AFCM) can solve this problem by replacing the Euclidean distance by a monotone increasing function called Alternative function. Therefore, this study use the combination between function of the KHM clustering and the new method, called the Alternative K-Harmonic Means (AKHM). The result showed AKHM method is a new promising for clustering data.

คำสำคัญ : การจัดกลุ่ม เค-ฮาร์โมนิกมีน ฟัชชี-ซีมีน

Key Words : Clustering, K-Harmonic Means, Fuzzy C-Means

*นักศึกษานิเทศศาสตร์ สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น

**ผู้ช่วยศาสตราจารย์ ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น

***อาจารย์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีมหานคร

บทนำ

การจัดกลุ่มข้อมูล (Data Clustering) เป็นการเรียนรู้แบบไม่มีผู้สอนคือไม่อาศัยการจัดกลุ่มไว้ล่วงหน้าและไม่ใช้ตัวอย่าง ข้อมูลถูกจัดเข้าด้วยกันโดยพิจารณาจากความคล้ายกันของแต่ละข้อมูล (วิกิพีเดีย, 2552) ซึ่งขั้นตอนวิธีที่นิยมใช้ในการจัดกลุ่มข้อมูลและถือว่าเป็นขั้นตอนวิธีที่ง่ายที่สุดคือ KM ใช้ในการจัดกลุ่มข้อมูลตามจำนวนกลุ่มที่เราต้องการโดยในการจัดกลุ่มจะพิจารณาค่าระยะห่างระหว่างข้อมูลกับค่ากลางด้วยมาตรวัดยูคลิด (Euclidean Distance) เพื่อตัดสินใจว่าข้อมูลแต่ละตัวจะถูกจัดเข้าไปไว้ในกลุ่มไหนนั่นเอง ซึ่งอัลกอริธึมนี้มีจุดอ่อนคือเมื่อข้อมูลมีไม่มาก ค่ากลางแรกของกลุ่มจะถูกตัดสินใจให้เป็นค่ากลางของกลุ่ม, จำนวนกลุ่มต้องกำหนดเองเพราะเราไม่รู้กลุ่มแน่นอน แม้จะใช้ข้อมูลชุดเดียวกันถ้าข้อมูลนำเข้ามีความแตกต่างกันอาจจะต้องแบ่งจำนวนกลุ่มต่างกันด้วย และถ้าจำนวนของข้อมูลน้อย ค่าเริ่มต้นมีผลต่อคำตอบของค่ากลางกลุ่ม ค่าเฉลี่ยที่ใช้ไม่สามารถแก้ปัญหา outliers อาจทำให้มีข้อมูลที่ถูกจัดอยู่เพียงลำพังดังนั้นจึงนำ KHM มาใช้ในการจัดกลุ่มเพื่อแก้ปัญหาเรื่องคำตอบขึ้นกับค่ากลางกลุ่มเริ่มต้นใน KM ขั้นตอนวิธี KHM จะไม่ขึ้นกับค่ากลางเริ่มต้นและไม่เกิดปัญหา outliers เพราะเปลี่ยนการจัดข้อมูลเข้ากลุ่มด้วยการใช้ winner-takes-all ใน KM มาใช้ harmonic means function แทน ซึ่งให้คำตอบการจัดกลุ่มข้อมูลดีกว่า ขั้นตอนวิธี KM (Zhang, 2000) จากความพยายามในการแก้ปัญหาการจัดกลุ่มด้วยขั้นตอนวิธี KM ด้วยวิธีอื่นเช่น FCM เป็นการเพิ่มความน่าเชื่อถือให้กับขั้นตอนวิธี KM (สมชาย, 2549) โดยการคำนวณหาความเป็นสมาชิกมาช่วยในการจัดกลุ่ม โดยการจัดกลุ่มจะดูค่าความเป็นสมาชิกของข้อมูลแต่ละตัว ทำให้ลดค่าความผิดพลาดของการจัดกลุ่มให้น้อยลงด้วย แต่ปัญหาของการได้ค่ากลางที่เหมาะสมที่สุดเฉพาะที่ (Local Optimum) ยังคงมีอยู่ทั้งสามขั้นตอนวิธียังมีจุดด้อยคือไม่สามารถแก้ปัญหาการจัดกลุ่มที่มีบางข้อมูล

อยู่ห่างไกลจากกลุ่มใด ๆ หรือมีข้อมูลรบกวน ทำให้ได้ค่ากลางที่ไม่เหมาะสมจึงใช้การหาระยะห่างหรือวัดความใกล้เคียงมาวัดยูคลิดเป็นการใช้ตัวชี้วัดความใกล้เคียงใหม่ใน KM เรียกว่า Alternative KM (AKM) และ Alternative FCM (AFCM) ทำให้แก้ปัญหานี้ได้ โดยเปลี่ยนการหาระยะห่างด้วยมาตรวัดยูคลิดเป็นฟังก์ชันแบบอื่นคือฟังก์ชันการเพิ่มขึ้นแบบต่อเนื่อง (Monotone Increasing Function) (Wu and Yang, 2002) ดังนั้นในงานวิจัยนี้จึงนำตัวชี้วัดความใกล้เคียงกับขั้นตอนวิธี KHM เรียกว่า Alternative KHM (AKHM)

บทความนี้แบ่งเป็นห้าส่วนดังนี้ทฤษฎีที่เกี่ยวข้องอยู่ในส่วนที่สอง ผลการวิจัยและการอภิปรายผลอยู่ในส่วนที่สาม สรุปผลการวิจัยอยู่ในส่วนที่สี่ และส่วนสุดท้ายคือข้อเสนอแนะ

ทฤษฎีที่เกี่ยวข้อง

1. K-Means

ซึ่งมีขั้นตอนวิธีการจัดกลุ่มดังนี้ (Zhang,

Hsu and Dayal, 1999)

เริ่มจากการกำหนดค่าเริ่มต้นต่อไปนี้

- ค่ากลางกลุ่มเริ่มต้น: m_1
- จำนวนกลุ่ม: K
- ข้อมูล: x_i
- จำนวนข้อมูล: N
- ค่าคงที่สำหรับเงื่อนไขการหยุด: ϵ

ขั้นตอนที่ 1: จัดข้อมูลเข้ากลุ่มด้วยการหาระยะห่างระหว่างทุกข้อมูลกับทุกค่ากลางกลุ่มด้วยสมการที่ 1 ดังนี้

$$d(x, M) = \min\{\|x - m\|^2\} \quad (1)$$

ขั้นตอนที่ 2: ปรับค่ากลางกลุ่มใหม่โดยหาค่าเฉลี่ยของข้อมูลในกลุ่มด้วยสมการที่ 2 ดังนี้

$$m_l = \frac{\sum_{x \in S_l} x}{|S_l|} \quad (2)$$

โดยที่ $S_l \subset X = \{x_i\}_{i=1}^N$ ชับเซตของ x ที่อยู่ใน
กลุ่มของค่ากลาง m_l

ขั้นตอนที่ 3: คำนวณค่าประสิทธิภาพจากค่า
ฟังก์ชันวัตถุประสงค์ด้วยสมการที่ 3 ดังนี้

$$OBJ_{KM} = \sum_{l=1}^K \sum_{x \in X} \|x - m_l\|^2 \quad (3)$$

ขั้นตอนที่ 4: นำค่ากลางใหม่คำนวณค่า
ฟังก์ชันวัตถุประสงค์ใหม่ด้วยสมการที่ 3 แล้ว
ตรวจสอบเงื่อนไขการหยุดทำงาน ตามสมการที่ 4
ดังนี้

$$|OBJ_{KM}^t - OBJ_{KM}^{t-1}| \leq \epsilon \quad (4)$$

2. Fuzzy C-Means

ใช้เป็นเครื่องมือที่ช่วยในการตัดสินใจ
ภายใต้ความไม่แน่นอนของข้อมูล การจัดกลุ่มข้อมูลทำ
ให้ข้อมูลสามารถอยู่ได้มากกว่าหนึ่งกลุ่มโดยพิจารณา
จากค่าความเป็นสมาชิกที่มีกับทุกค่ากลางกลุ่มนั้น ๆ
วิธีการหาคำตอบของการจัดกลุ่มด้วย FCM (Hamerly
and Elkan, 2002) เริ่มจากการกำหนดค่าเริ่มต้นต่อไปนี้

- ค่ากลางกลุ่มเริ่มต้น: m_l
- จำนวนกลุ่ม: K
- ค่าชี้กำลังของความเป็นฟัซซี่: γ
- แทนข้อมูลด้วย: x_i
- จำนวนข้อมูลทั้งหมด: N
- ค่าคงที่สำหรับเงื่อนไขการหยุด: ϵ

กำหนดให้ $\mu_{i,l} \in [0, 1]$

ขั้นตอนที่ 1: FCM เป็นขั้นตอนวิธีในการหาค่า
น้อยที่สุดเพื่อพิจารณาประสิทธิภาพของขั้นตอนวิธี
โดยใช้ฟังก์ชันวัตถุประสงค์ (Objective Function) ใน
สมการที่ 5 ดังนี้

$$OBJ_{FCM} = \sum_{i=1}^N \sum_{l=1}^K (x_i - m_l)^\gamma \|x_i - m_l\|^2 \quad (5)$$

จากสมการที่ 5 ค่า $\|x_i - m_l\|^2$ แทนระยะห่าง
ระหว่างข้อมูลที่ i กับค่ากลางกลุ่มที่ l

ขั้นตอนที่ 2: ปรับค่ากลางกลุ่มใหม่ตามสมการ
ที่ 6 ดังนี้

$$m_l = \frac{\sum_{i=1}^N \mu_{i,l}^\gamma x_i}{\sum_{i=1}^N \mu_{i,l}^\gamma} \quad (6)$$

สำหรับการหาค่าความเป็นสมาชิกใหม่คำนวณ
ด้วยสมการที่ 7 ดังนี้

$$\mu_{FCM} = \frac{\|x_i - m_l\|^{-2/(\gamma-1)}}{\sum_{l=1}^K \|x_i - m_l\|^{-2/(\gamma-1)}} \quad (7)$$

ตรวจสอบความถูกต้องของการหาค่าความเป็น
สมาชิกได้ด้วยสมการที่ 8 ดังนี้

$$\sum_{i=1}^K \mu_{i,l}^\gamma = N \quad (8)$$

ขั้นตอนที่ 3: คำนวณค่าฟังก์ชันวัตถุประสงค์ใหม่ด้วยสมการที่ 5 แล้วตรวจสอบเงื่อนไขการหยุดตามสมการที่ 9 ดังนี้

$$\|\mu^{\text{old}} - \mu^{\text{new}}\| \leq \varepsilon$$

(9)

3. K-Harmonic Means

K-Harmonic Means (Zhang, 2000) เป็นขั้นตอนวิธีที่มีวัตถุประสงค์เพื่อแก้ไขข้อด้อยของขั้นตอนวิธี KM วิธีการหาค่าตอบของการจัดกลุ่มด้วย KHM เริ่มจากการกำหนดค่าเริ่มต้นต่อไปนี้

- ค่ากลางกลุ่มเริ่มต้น: m_1
- จำนวนกลุ่ม: K
- ค่าสัมประสิทธิ์ของ L^p -Distance:

$$2 \leq p < \infty$$

- ข้อมูล: x_i
- จำนวนข้อมูล: N
- ค่าคงที่สำหรับเงื่อนไขการหยุด: ε

ขั้นตอนที่ 1: จัดข้อมูลเข้ากลุ่มตามสมการที่ 10 ดังนี้

$$d(x, M) = \frac{|M|}{\sum_{m \in M} \frac{1}{\|x - m\|^p}}$$

(10)

ขั้นตอนที่ 2: คำนวณระยะห่างระหว่างจุดข้อมูลกับทุกค่ากลางกลุ่ม (ทำเป็นสัญลักษณ์ง่าย ๆ ได้ดังนี้ $d_{i,1} = \|x_i - m_1\|$ แล้วคำนวณหาค่ากลางกลุ่มใหม่จากสมการที่ 11 ดังนี้

$$m_k = \frac{\sum_{i=1}^N \frac{1}{\left(\sum_{d=1}^{p+2} \left(\sum_{i,k} \frac{1}{d^{p,1}} \right)^2 \right)^{x_i}}}{\sum_{i=1}^N \frac{1}{\left(\sum_{d=1}^{p+2} \left(\sum_{i,k} \frac{1}{d^{p,1}} \right)^2 \right)^2}}$$

(11)

ขั้นตอนที่ 3: คำนวณประสิทธิภาพด้วยฟังก์ชันวัตถุประสงค์ของขั้นตอนวิธี KHM จากสมการที่ 12 ดังนี้

$$OBJ_{KHM_p} = \sum_{i=1}^N \frac{|K|}{\sum_{i=1}^K \frac{1}{\|x_i - m_i\|^p}}$$

(12)

ขั้นตอนที่ 4: ตรวจสอบเงื่อนไขการหยุดรอบการทำงาน ตามสมการที่ 13 ดังนี้

$$\left| OBJ_{KHM_p}^{\text{old}} - OBJ_{KHM_p}^{\text{new}} \right| \leq \varepsilon$$

(13)

4. Alternative K-Means

เรียกว่า Alternative Hard C-Means (AHCM) (Wu and Yang, 2002) การจัดกลุ่มด้วยวิธีนี้จะเป็นการเปลี่ยนการวัดความคล้ายด้วยวิธีใหม่จากเดิมใช้การหาระยะห่างด้วยมาตรวัดยูคลิดมาเป็นการใช้การหาระยะห่างด้วยฟังก์ชันการเพิ่มขึ้นแบบต่อเนื่อง (Monotone Increasing Function) ดังสมการนี้

$$d(x, m) = 1 - \exp(-\beta \|x - m\|^2)$$

(14)

กำหนดให้ β เป็นค่าคงที่บวก จากการแทนการหาระยะห่างด้วยสมการนี้จะได้สมการการปรับค่ากลางดังนี้

$$m = \sum_{i=1}^N \frac{\exp(-\beta \|x - m\|^2)}{\sum_{i=1}^N \exp(-\beta \|x - m\|^2)} x_i \quad (15)$$

ฟังก์ชันวัตถุประสงค์ตามสมการต่อไปนี้

$$OBJ_{AKM} = \sum_{l=1}^K \sum_{i \in S_l} \{1 - \exp(-\beta \|x_i - m_l\|^2)\} \quad (16)$$

ที่กำหนดให้ $S_l \subset X = \{x_i\}_{i=1}^N$

และมีขั้นตอนวิธีการหาคำตอบของการจัดกลุ่มเหมือนกับ KM ทุกขั้นตอน

5. Alternative Fuzzy C-Means

เมื่อนำการวัดความคล้ายด้วยวิธีการใช้ฟังก์ชันการเพิ่มขึ้นแบบต่อเนื่องมาหาค่าระยะทำให้ได้สมการการหาค่าความเป็นสมาชิกดังนี้

$$\mu_{i,k} = \frac{\left[1 / (1 - \exp(-\beta \|x_i - m_k\|^2))\right]^{1/(\gamma-1)}}{\sum_{l=1}^K \left[1 / (1 - \exp(-\beta \|x_i - m_l\|^2))\right]^{1/(\gamma-1)}} \quad (17)$$

สมการในการปรับค่ากลางใหม่คือ

$$m_l = \frac{\sum_{i=1}^N \mu_{i,l}^\gamma \exp(-\beta \|x_i - m_l\|^2) x_i}{\sum_{i=1}^N \mu_{i,l}^\gamma \exp(-\beta \|x_i - m_l\|^2)} \quad (18)$$

ฟังก์ชันวัตถุประสงค์ของ AFCM คือ

$$OBJ_{AFCM} = \sum_{l=1}^K \sum_{i=1}^N (\mu)^\gamma \{1 - \exp(-\beta \|x_i - m_l\|^2)\} \quad (19)$$

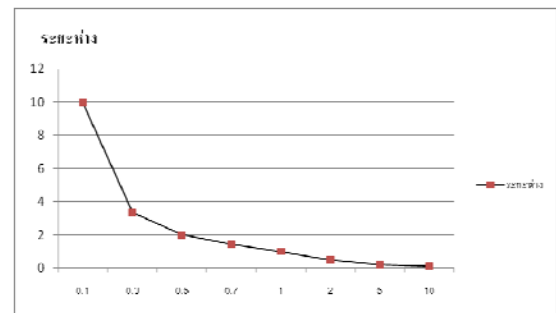
สำหรับวิธีการหาคำตอบจะมีขั้นตอนเหมือนกับขั้นตอนวิธี FCM

6. Alternative K-Harmonic Means

ขั้นตอนวิธีนี้เกิดขึ้นจากการใช้ฟังก์ชันเพิ่มขึ้นแบบต่อเนื่องมาใช้ในการหาระยะห่างใน KHM โดยการพิจารณาฟังก์ชันการจัดข้อมูลเข้ากลุ่มด้วย

สมการที่ 10 พิจารณาสมการ $\frac{1}{\|x - m\|^p}$ เมื่อ

กำหนดให้ค่า $p = 2$ และกำหนดให้ข้อมูลเป็น 1 มิติจะได้สมการดังนี้ $\frac{1}{|x - m|^2}$ จะได้กราฟเส้นดังภาพที่ 1 เมื่อระยะห่างที่วัดได้เป็น $\{0.1, 0.3, 0.5, 0.7, 1, 2, 5, 10\}$



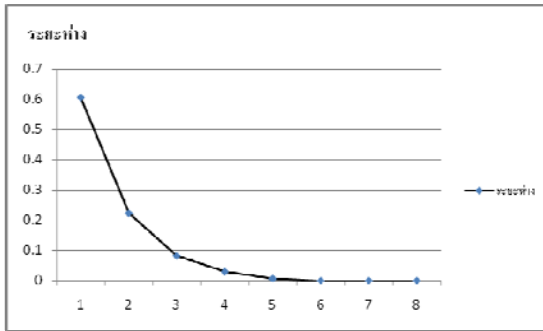
ภาพที่ 1 กราฟระยะห่างวัดจากสมการ $\frac{1}{|x - m|^2}$

เมื่อแทนสมการ $\frac{1}{|x - m|^2}$ เป็นฟังก์ชันการ

เพิ่มขึ้นแบบต่อเนื่องเป็นสมการ $\exp(-\beta \|x - m\|^p)$

$= \frac{1}{\exp(\beta \|x - m\|^p)}$ เมื่อระยะห่างตามเซตข้างต้นและค่า $\beta = 5$, $p = 2$ จะได้กราฟดังภาพที่ 2 ซึ่งลักษณะของกราฟในทั้งสองภาพมีลักษณะคล้ายกันแต่ค่าของระยะห่างที่วัดได้จะมีช่วงแตกต่างกันในภาพที่

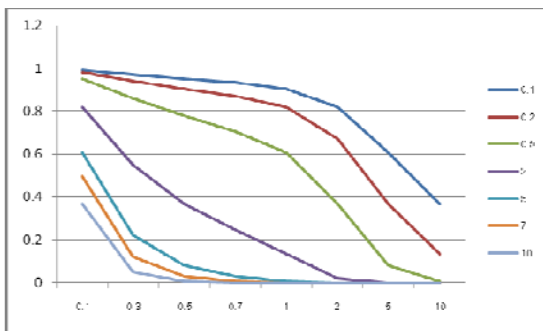
1 ระยะห่างสามารถมีค่ามากกว่า 1 ได้แต่ในภาพที่ 2 ระยะห่างจะมีค่าไม่เกิน 1 ดังนั้นสามารถแทนสมการการจัดข้อมูลเข้ากลุ่มในสมการที่ 10 ด้วยฟังก์ชันการเพิ่มขึ้นแบบต่อเนื่องเป็นสมการใหม่ซึ่งแสดงได้ตามสมการที่ 20



ภาพที่ 2 กราฟระยะห่างวัดจาก

$$\text{สมการ } \frac{1}{\exp(\beta \|x-m\|^p)}$$

สำหรับค่า β เมื่อกำหนดค่าเป็น {0.1, 0.2, 0.5, 2, 5, 7, 10} ในสมการ $\frac{1}{\exp(\beta \|x-m\|^p)}$ เมื่อเป็นข้อมูลหนึ่งมิติ และ $p = 2$ จะได้กราฟดังภาพที่ 3 นี้



ภาพที่ 3 สมการ $\frac{1}{\exp(\beta \|x-m\|^p)}$ เมื่อกำหนดค่า β แตกต่างกัน

เมื่อทำการแทนการหาระยะห่างด้วยฟังก์ชันการเพิ่มขึ้นแบบต่อเนื่องสำหรับการจัดข้อมูลเข้ากลุ่มได้สมการดังต่อไปนี้

$$d(x,m) = \frac{|K|}{\sum_{l=1}^K \exp(-\beta \|x_i - m_l\|^p)} \quad (20)$$

และการปรับค่ากลางใหม่หาจากสมการ

ดังนี้

$$m_k = \frac{\sum_{i=1}^N \frac{1}{\left(1 - \exp(-\beta \|x_i - m_k\|^p)\right)^2} \left(\sum_{l=1}^K \exp(-\beta \|x_i - m_l\|^2)\right)^{2x_i}}{\sum_{i=1}^N \frac{1}{\left(1 - \exp(-\beta \|x_i - m_k\|^p)\right)^2} \left(\sum_{l=1}^K \exp(-\beta \|x_i - m_l\|^2)\right)^2} \quad (21)$$

โดยมีฟังก์ชันวัตถุประสงค์ของ AKHM

เป็นดังสมการนี้คือ

$$OBJ_{AKHM} = \sum_{i=1}^N \frac{|K|}{\sum_{l=1}^K \exp(-\beta \|x_i - m_l\|^p)} \quad (22)$$

สำหรับวิธีการหาคำตอบของขั้นตอนวิธี

AKHM มีลำดับขั้นตอนเหมือนกับวิธี KHM

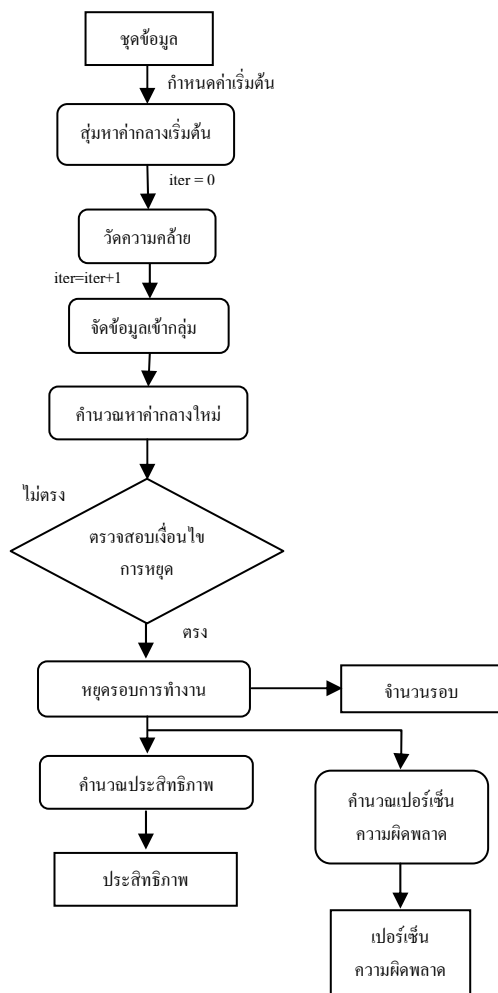
วิธีการทดลองและผลการทดลอง

1. ชุดข้อมูลที่ใช้ในการทดลอง

ใช้ชุดข้อมูลจริงจากฐานข้อมูล UCI Machine Learning Repository ประกอบด้วยชุดข้อมูล 3 ชุดดังนี้ Iris มีขนาด 150 x 4 จำนวน 3 กลุ่ม , Wine มีขนาด 178 x 13 จำนวน 3 กลุ่ม , Soybean (small) มีขนาด 47 x 35 จำนวน 4 กลุ่ม และชุดข้อมูลที่สร้างขึ้นเอง เป็นข้อมูลสองมิติจำนวน 100 ข้อมูล และทำการทดลองเพื่อวัดประสิทธิภาพของอัลกอริทึม ดังนั้นในการทดลองจะมีการใช้ชุดข้อมูลทั้งหมดคือ ข้อมูลจริง 3 ชุดข้อมูลที่สร้างขึ้นเองจำนวน 1 ชุด รวมทั้งสิ้นเป็น 4 ชุดข้อมูล

2. วิธีการทดลอง

นำชุดข้อมูลทั้งสี่ชุดไปเข้ากระบวนการของขั้นตอนวิธีการจัดกลุ่มทั้ง 6 วิธีคือ KM, FCM, KHM, AKM, AFCM, AKHM แล้วคำนวณหาค่าความผิดพลาด เวลาที่ใช้ในการจัดกลุ่ม วัดประสิทธิภาพของแต่ละขั้นตอนวิธี และจำนวนรอบการทำงาน โดยทดลองกับแต่ละชุดข้อมูลจำนวน 100 ครั้งเพื่อหาค่าเฉลี่ยของผลการทดลองในทุก ๆ ขั้นตอนวิธีกับแต่ละชุดข้อมูล โดยการทดลองมีลักษณะการดำเนินการดังแผนภาพด้านล่างนี้



ภาพที่ 4 แผนภาพขั้นตอนการจัดกลุ่มข้อมูล

จากแผนภาพการดำเนินการข้อมูล 4 ชุดถูกนำไปจัดกลุ่มด้วยขั้นตอนวิธีทั้งหมด 6 แบบ จำนวน

ครั้งในการทดลองคือ ชุดข้อมูลละ 100 ครั้ง ในแต่ละขั้นตอนวิธี ดังนั้นการทดลองนี้ได้ทำการทดลองทั้งหมดจำนวน $4 \times 6 \times 100$ คือ 2400 ครั้ง นำผลการทดลองมาหาค่าเฉลี่ยเพื่อประเมินผลต่อไปดังตารางสรุปผลเฉลี่ยของแต่ละขั้นตอนวิธีดังนี้

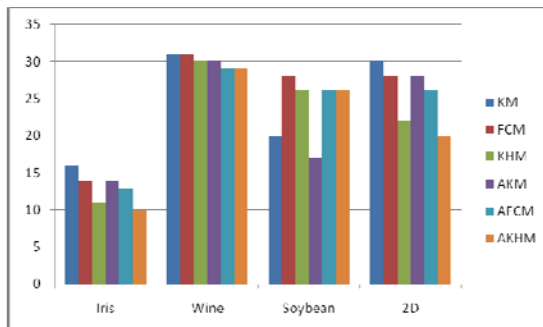
ตารางที่ 1 สรุปผลการจัดกลุ่มข้อมูลจำนวน 4 ชุดด้วย 6 ขั้นตอนวิธี

Algorithm	Task	DATA SET			
		iris	wine	soybean	2D
KM	error	16	31	20	30
	Perf.	98	0.1154	40	358
	Iter	3	4	1	5
FCM	error	14	31	28	28
	Perf.	38.16914	0.92505	27.3766	86.95
	Iter	14	14	43	8
KHM	error	11	30	26	22
	Perf.	755	0.0277	8.6716	1134
	Iter	15	4	10	8
AKM	error	14	30	17	28
	Perf.	10.9818	0.0126	47	1068
	Iter	10	4	4	4
AFCM	error	13	29	26	26
	Perf.	537	178	47	8234
	Iter	18	20	30	8
AKHM	error	10	29	26	20
	Perf.	121	178	47	1076
	Iter	8.5	1	2	2

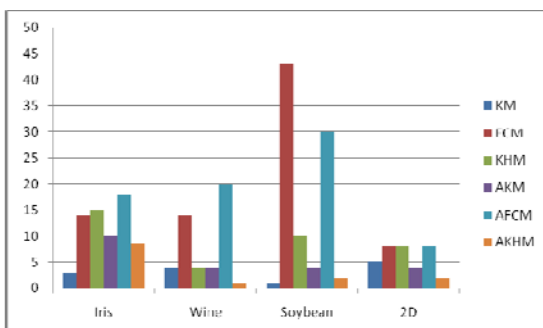
จากตารางที่ 1 error คือ ค่าความผิดพลาดเฉลี่ยในการจัดกลุ่ม Perf. คือ ประสิทธิภาพเฉลี่ยของขั้นตอนวิธีที่วัดด้วยสมการการวัดประสิทธิภาพของวิธี KM ตามสมการที่ 3 ซึ่งหมายถึงการหาผลรวมของระยะห่างของข้อมูลที่ถูกเลือกเข้ากลุ่มทั้งหมดทุกกลุ่ม

นั่นเอง Iter คือ จำนวนรอบเฉลี่ยของการทำงานโดยการทดลองจำนวน 100 ครั้งในแต่ละขั้นตอนวิธีของทุกชุดข้อมูล

จากผลการทดลองตามตารางที่ 1 นำมาแสดงค่าความผิดพลาดเพื่อเปรียบเทียบแต่ละขั้นตอนวิธีดังนี้



ภาพที่ 5 เปรียบเทียบค่าความผิดพลาดจากการ จัดกลุ่มข้อมูลด้วย 6 ขั้นตอนวิธี



ภาพที่ 6 เปรียบเทียบจำนวนรอบการทำงาน

สรุปผลการวิจัย

งานวิจัยนี้ได้นำเสนอผลการทดลองกับข้อมูล 4 ชุด คือ Iris, Wine, Soybean(small) และ 2D กับขั้นตอนวิธีการจัดกลุ่มทั้งหมด 6 แบบ คือ KM, FCM, KHM, AKM, AFKM, AKHM จากผลการทดลองสามารถสรุปโดยจำแนกเป็นกรณีที่น่าสนใจดังนี้ กรณีแรกวิเคราะห์ผลตามค่าความผิดพลาดจากภาพที่ 5 ถ้า

ค่าความผิดพลาดต่ำย่อมแสดงถึงความถูกต้องที่สูงนั่นเอง ผลการทดลองในงานวิจัยนี้แสดงให้เห็นว่าวิธี AKHM สามารถจัดกลุ่มข้อมูลที่ทำให้ค่าความถูกต้องสูงที่สุดเมื่อเปรียบเทียบกับวิธีอื่นเมื่อใช้กับข้อมูล Wine, Iris และ 2D วิธี AKM ให้ความถูกต้องสูงสำหรับชุดข้อมูล Soybean(small) กรณีที่สองพิจารณาจากจำนวนรอบการทำงานที่น้อยที่สุดตามภาพที่ 6 AKHM ใช้รอบการทำงานต่ำสุดกับข้อมูล Wine และ 2D สำหรับ Iris และ Soybean (small) มีรอบการทำงานต่ำสุดเมื่อใช้ วิธี KM แต่จำนวนรอบที่ได้เมื่อใช้วิธี AKHM ก็ถือว่าใกล้เคียงกับจำนวนรอบจากวิธี KM ด้วยเช่นกัน กรณีสุดท้ายคือ การพิจารณาจากค่าประสิทธิภาพของขั้นตอนวิธีเมื่อใช้วิธีการวัดประสิทธิภาพตามการวัดประสิทธิภาพของ KM ดังตารางเปรียบเทียบผลการวัดประสิทธิภาพเฉลี่ยแต่ละขั้นตอนวิธีดังนี้

ตารางที่ 2 เปรียบเทียบประสิทธิภาพเฉลี่ยของขั้นตอนวิธีที่วัดด้วยฟังก์ชันวัดประสิทธิภาพของวิธี KM

Algorithm	DATA SET			
	iris	wine	soybean	2D
KM	98	0.1154	40	358
FCM	38.16914	0.92505	27.3766	86.95
KHM	755	0.0277	8.6716	1134
AKM	10.9818	0.0126	47	1068
AFKM	537	178	47	8234
AKHM	121	178	47	1076

กรณีสุดท้ายคือ การพิจารณาจากประสิทธิภาพของขั้นตอนวิธีค่าที่ต่ำที่สุดจะเป็นค่าที่ดีที่สุดนั่นเอง จากตารางที่ 2 วิธี AKM ให้ประสิทธิภาพที่ดีกว่าที่สุดสำหรับชุดข้อมูล Iris และ Wine วิธี KHM ให้ประสิทธิภาพดีที่สุดสำหรับชุดข้อมูล Soybean

(Small) และวิธี FCM ให้ประสิทธิภาพดีที่สุดสำหรับชุดข้อมูล 2D จากการพิจารณาผลการทดลองที่ดีที่สุด ในทั้งสามกรณีแสดงให้เห็นว่าในแต่ละกรณีจะมีขั้นตอนวิธีแบบทางเลือกที่ให้ค่าที่ดีที่สุดรวมอยู่ด้วยเป็นส่วนใหญ่ และวิธีที่ใช้การหาความคล้ายด้วยฟังก์ชันการเพิ่มขึ้นแบบต่อเนื่องจะให้ผลการทดลองที่ดีกว่าการหาความคล้ายด้วยมาตรวัดยูคลิด โดยเฉพาะวิธี AKHM มีประสิทธิภาพสูงสุดในด้านของความถูกต้องที่สูงสุดและจำนวนรอบที่น้อยสุดในเกือบทุกชุดข้อมูลที่ทำการศึกษาทดลอง ซึ่งตรงตามวัตถุประสงค์การทดลอง

ข้อเสนอแนะ

ถ้าจะนำขั้นตอนวิธีการจัดกลุ่มด้วย AKHM ไปใช้ในการจัดกลุ่มควรรศึกษาลักษณะของชุดข้อมูลจำนวนกลุ่มที่เหมาะสมเพราะงานแต่ละอย่างมีวัตถุประสงค์ที่ต้องการต่างกันแม้จะใช้ชุดข้อมูลเดียวกันดังนั้นจึงต้องทำความเข้าใจถึงปัญหาให้ดีก่อนการจัดกลุ่มยังมีอีกหลายกลยุทธ์ที่น่าสนใจเพื่อนำมาประยุกต์ใช้การจัดกลุ่มให้มีประสิทธิภาพสูงขึ้น

เอกสารอ้างอิง

วิกิพีเดีย. 2552. การแบ่งกลุ่มข้อมูล. ค้นเมื่อ

12 มิถุนายน 2552, จาก

<http://th.wikipedia.org/wiki>

สมชาย จำปาทอง. 2549. การจัดกลุ่มข้อมูลแบบ

ปรับตัวได้โดยใช้ฟuzzyเซต. วิทยานิพนธ์วิทยา

ศาสตรมหาบัณฑิต สาขาวิชาวิทยาการ

คอมพิวเตอร์ บัณฑิตวิทยาลัย

มหาวิทยาลัยขอนแก่น.

Zhang, B. 2000. Generalized k-harmonic means-

boosting in unsupervised learning. Technical

Report HPL-2000-137, HP Laboratories,

Palo Alto.

Zhang, B., Hsu, M. and Dayal, U. 1999. K-harmonic means-a data clustering algorithm. Technical Report HPL-1999-124, HP Laboratories, Palo Alto.

Hamerly, G. and Elkan, C. 2002. Alternative to the k-means algorithm that find better clusterings. Conference on Information and Knowledge Management Proceedings of the eleventh international, McLean, Virginia, USA, pp 600-607.

Wu, K. and Yang, M. 2002. Alternative c-means clustering algorithms. Pattern Recognition. 35, pp 2267-2278.