บทที่ 2
เทคนิคการจัดกลุ่มข้อมูล

การจัดกลุ่มข้อมูล หมายถึงการแบ่งกลุ่มข้อมูลออกเป็นกลุ่ม ๆ โดยที่ข้อมูลในกลุ่มเดียวกันจะมีความคล้ายคลึงกันมากกว่าข้อมูลที่อยู่คนละกลุ่ม ในการศึกษาเทคนิคที่เกี่ยวข้องกับการจัดกลุ่มข้อมูล มีแนวคิดและทฤษฎีต่าง ๆ ที่เกี่ยวข้องอยู่เป็นจำนวนมาก โดยจะกล่าวถึงรายละเอียดของแนวความคิดและทฤษฎีการวัดความคล้ายคลึงกันของข้อมูล การจัดกลุ่มข้อมูลแบบแบ่งแยกออกเป็นส่วน ๆ (Partitioning Method) และการจัดกลุ่มข้อมูลแบบลำดับขั้น (Hierarchical Method)

2.1 การวัดความคล้ายคลึงกันของข้อมูล

ความคล้ายคลึงกันของสองข้อมูลใด ๆ สามารถวัดได้จากระยะทางระหว่างสองข้อมูลนั้น หากมีระยะทางน้อยจะหมายความว่าสองข้อมูลนั้นมีความคล้ายคลึงกันมาก แต่หากมีระยะทางมากจะหมายความว่าสองข้อมูลนั้นมีความคล้ายคลึงกันน้อย ซึ่งมีหลายวิธีการในการคำนวณระยะทางระหว่างข้อมูล เช่น ระยะทางแบบยูคลีเดียน (Euclidean distance) และระยะทางแบบแมนฮัตตัน (Manhattan distance) เป็นต้น

2.1.1 ระยะทางแบบยูคลีเดียน

นิยามของระยะทางแบบยูคลีเดียน : ให้ I และ J เป็นจุดในมิติขนาด D โดยมี X_k เป็นค่าของคุณสมบัติ (Attribute Value) ที่ k ของจุด I และ Y_k เป็นค่าของคุณสมบัติที่ k ของจุด J จะได้ระยะทางระหว่างจุด I และ J คือรากที่สองของผลรวมของผลต่างระหว่างค่าคุณสมบัติของ I และ J ในทุกมิติยกกำลังสอง

$$d(I , J) = (\sum_{k=1}^{D}(X_k - Y_k)^2)^{1/2}$$

(Jiawei Han and Micheline Kamber, 2001)

2.1.2 ระยะทางแบบแมนฮัตตัน

นิยามของระยะทางแบบแมนฮัตตัน : ให้ I และ J เป็นจุดในมิติขนาด D โดยมี X_k เป็นค่าของคุณสมบัติที่ k ของจุด I และ Y_k เป็นค่าของคุณสมบัติที่ k ของจุด J จะได้ระยะทางระหว่างจุด I และ J คือผลรวมของค่าสัมบูรณ์ของผลต่างระหว่างค่าคุณสมบัติของ I และ J ในทุกมิติ

$$d(I , J) = \sum_{k=1}^{D} |X_k - Y_k|$$

(Jiawei Han and Micheline Kamber, 2001)

เนื่องจากระยะทางแบบยูคลีเดียนเป็นวิธีการวัดระยะทางที่ได้รับความนิยมมากที่สุด ในการศึกษาครั้งนี้จึงได้ใช้วิธีการวัดระยะทางแบบยูคลีเดียน

2.2 การจัดกลุ่มข้อมูลแบบแบ่งแยกออกเป็นส่วน ๆ

การจัดกลุ่มข้อมูลแบบแบ่งแยกออกเป็นส่วน ๆ เป็นเทคนิคการจัดกลุ่มข้อมูลโดยจัดกลุ่มข้อมูลออกเป็นกลุ่ม ๆ ตามจำนวนกลุ่มที่ต้องการ ซึ่งในที่นี้จะให้เป็นค่า k โดยกลุ่มที่แบ่งได้แต่ละกลุ่มจะมีสมาชิกอยู่อย่างน้อยหนึ่งข้อมูล และหนึ่งข้อมูลจะถูกแบ่งให้อยู่ในกลุ่มเดียวเท่านั้น เทคนิคนี้เหมาะกับการจัดกลุ่มข้อมูลที่ลักษณะการจัดกลุ่มแบบทรงกลม (Spherical-shaped) และมีขนาดเล็กจนถึงปานกลาง อัลกอริทึมที่ใช้ในการจัดกลุ่มข้อมูลแบบแบ่งแยกออกเป็นส่วน ๆ มีดังต่อไปนี้

2.2.1 อัลกอริทึมเคมีนส์

อัลกอริทึมเคมีนส์ เป็นอัลกอริทึมในการจัดกลุ่มข้อมูลตามความคล้ายคลึงกันของข้อมูล โดยที่ความคล้ายคลึงกันของข้อมูลจะพิจารณาจากจุดศูนย์กลางของกลุ่มหรือที่เรียกว่าจุดมีน โดยที่จุดมีนคือผลรวมของทุกค่าสมาชิกในกลุ่มหารด้วยจำนวนข้อมูลในกลุ่ม หรือเป็นค่าเฉลี่ยของกลุ่ม (Mean) นั่นเอง

ปัญหาการจัดกลุ่มด้วยอัลกอริทึมเคมีนส์ เป็นการแบ่งกลุ่มข้อมูลเป็นกลุ่ม ๆ จำนวน k กลุ่ม โดยแบ่งกลุ่มให้มีค่าผิดพลาดยกกำลังสองน้อยที่สุด ซึ่งค่าผิดพลาดยกกำลังสอง คือผลรวมของระยะทางระหว่างแต่ละจุดข้อมูล และจุดมีนของกลุ่มตนเอง อัลกอริทึมนี้พยายามที่จะทำให้เกิดข้อมูล k กลุ่มที่มีความหนาแน่นและแยกออกจากกันเท่าที่จะเป็นไปได้

หลักการของอัลกอริทึมเคมีนส์

(1) สุ่มแบ่งกลุ่มข้อมูลออกเป็น k กลุ่ม โดยผู้ใช้เป็นผู้กำหนดค่า k

(2) พิจารณาความคล้ายคลึงกันในลักษณะคุณสมบัติของจุดข้อมูล (Attribute)

(3) ใช้ค่าเฉลี่ย (Mean) ของจุดข้อมูลในกลุ่มเป็นจุดศูนย์กลางของกลุ่มข้อมูล

(4) การกำหนดกลุ่มให้กับจุดข้อมูลจะใช้ระยะทางระหว่างจุดข้อมูลกับค่าเฉลี่ยของกลุ่มข้อมูลนั้น โดยจะกำหนดให้อยู่ในกลุ่มที่มีระยะทางที่สั้นที่สุด

(5) ใช้ค่าผิดพลาดยกกำลังสองที่น้อยสุดเป็นเงื่อนไขในการหยุดคำนวณ

อัลกอริทึมนี้มีความน่าเชื่อถือ และมีประสิทธิภาพในการทำงานกับกลุ่มข้อมูลขนาดค่อนข้างใหญ่ เพราะว่าการคำนวณความซับซ้อนของอัลกอริทึมจะเป็น O(Nkt) ซึ่ง N เป็นจำนวนข้อมูลทั้ง

หมด, k เป็นจำนวนกลุ่ม และ t เป็นจำนวนงานที่ทำซ้ำกัน ปกติแล้ว k จะน้อยกว่า N มาก และ t ก็จะน้อยกว่า N มาก ซึ่งวิธีการนี้ส่วนมากจะสิ้นสุดการทำงานที่ผลลัพธ์ที่ดีที่สุด

อัลกอริทึมเคมีนส์สามารถประยุกต์ไปใช้ได้เมื่อข้อมูลที่ต้องการจะจัดกลุ่มสามารถหาค่าเฉลี่ยได้ ซึ่งอาจจะไม่สามารถนำไปใช้ได้ในบางกรณี เช่น เมื่อข้อมูลมีคุณสมบัติที่สลับซับซ้อนเกินไป และผู้ใช้ยังจำเป็นต้องกำหนดค่า k ซึ่งเป็นจำนวนของกลุ่มที่เฉพาะเจาะจงทำให้เป็นข้อเสียเปรียบ เพราะผู้ใช้มักไม่ทราบค่า k ว่าควรเป็นเท่าไร นอกจากนี้อัลกอริทึมเคมีนส์ไม่เหมาะสมสำหรับการจัดกลุ่มข้อมูลที่มีรูปร่างมีเว้า (nonconvex shape) หรือกลุ่มข้อมูลที่มีขนาดแตกต่างกันมาก ๆ ยิ่งกว่านั้นยังไวต่อสิ่งรบกวน (Noise) และจุดข้อมูลขนาดเล็กที่ไม่เกี่ยวข้องกับข้อมูลหลัก (Outlier) ทำให้มีผลกระทบต่อค่าเฉลี่ย

เพื่อให้การอธิบายดูกระชับ เข้าใจง่าย จึงขอกำหนดรูปแบบของการอธิบายอัลกอริทึมในเอกสาร โดยตัวเข้มจะแสดงถึงหัวข้อหรือคำหลักที่ใช้ในการอธิบาย และตัวเอนจะแสดงชื่อเฉพาะหรือความหมาย

ชื่ออัลกอริทึม : *เคมีนส์*

หน้าที่ : *อัลกอริทึมสำหรับแบ่งส่วนโดยใช้ค่าเฉลี่ยของข้อมูลในกลุ่ม*

ข้อมูลเข้า : *จำนวนกลุ่มที่ต้องการแบ่ง และจุดข้อมูลจำนวน N ข้อมูล*

ผลลัพธ์ : *กลุ่มของข้อมูลที่มีความคล้ายคลึงกันจำนวนตามที่ต้องการ โดยที่มีค่าผิดพลาดยกกำลังสองน้อยที่สุด*

วิธีการ :

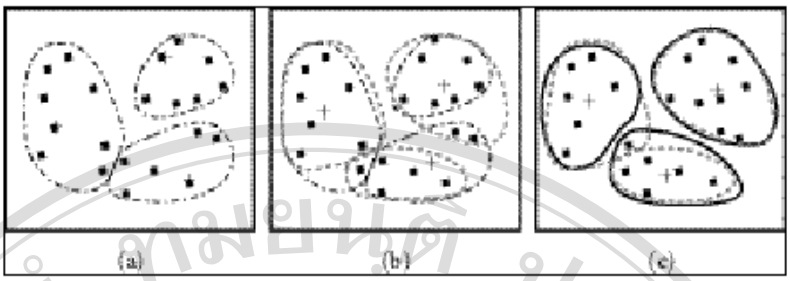
1) *ทำการสุ่มค่าจุดข้อมูลจำนวน k ข้อมูล จากจุดข้อมูลทั้งหมด ซึ่งใช้เป็นตัวแทนของค่าเฉลี่ยกลุ่ม หรือเป็นจุดศูนย์กลางของกลุ่ม*
2) **ทำซ้ำข้อต่อไปนี้**

2.1) กำหนดจุดข้อมูลที่เหลืออยู่ให้กับแต่ละกลุ่มที่มีความคล้ายกันบนพื้นฐานของระยะทางระหว่างจุดข้อมูลกับค่าเฉลี่ยของกลุ่ม

2.2) ทำการคำนวณค่าเฉลี่ยของแต่ละกลุ่มใหม่

จนกระทั่ง *ไม่มีการเปลี่ยนกลุ่มของข้อมูล ซึ่งหมายถึงค่าผิดพลาดยกกำลังสองน้อยที่สุดแล้ว*

ซึ่งวิธีการของอัลกอริทึมเคมีนส์สามารถแสดงผลที่ได้ในแต่ละขั้นตอนออกมาเป็นรูปภาพ ดังรูปที่ 2.1



รูป 2.1 วิธีการของอัลกอริทึมเคมีนส์

(Jiawei Han and Micheline Kamber, 2001)

2.2.2 อัลกอริทึมเคมีเดียนส์

อัลกอริทึมเคมีเดียนส์ เป็นอัลกอริทึมในการจัดกลุ่มข้อมูลตามความคล้ายคลึงกันของข้อมูล เหมือนอัลกอริทึมเคมีนส์ แต่ความคล้ายคลึงกันของข้อมูลจะพิจารณาจากจุดศูนย์กลางของกลุ่ม หรือที่เรียกว่าจุดมีเดียน โดยที่จุดมีเดียนของกลุ่มข้อมูลจะเป็นสมาชิกในกลุ่มข้อมูลนั้นที่มีผลรวมของระยะทางระหว่างจุดข้อมูลในกลุ่มกับจุดมีเดียนน้อยที่สุด หรือกล่าวได้ว่า จุดมีเดียนของกลุ่มเป็นจุดที่อยู่ใกล้กับจุดข้อมูลอื่น ๆ ภายในกลุ่มมากที่สุด

ปัญหาการจัดกลุ่มด้วยอัลกอริทึมเคมีเดียนส์ เป็นการแบ่งกลุ่มข้อมูลเป็นกลุ่ม ๆ ตามจำนวนกลุ่มที่ต้องการ โดยแบ่งกลุ่มให้มีค่าผิดพลาดยกกำลังสองน้อยที่สุด เหมือนกับอัลกอริทึมเคมีนส์

หลักการของอัลกอริทึมเคมีเดียนส์

(1) สุ่มแบ่งกลุ่มข้อมูลออกเป็น k กลุ่ม โดยผู้ใช้เป็นผู้กำหนดค่า k

(2) พิจารณาความคล้ายคลึงกันในลักษณะคุณสมบัติของจุดข้อมูล (Attribute)

(3) ใช้จุดข้อมูลที่อยู่ใกล้กับจุดข้อมูลอื่น ๆ ภายในกลุ่มมากที่สุด เป็นจุดศูนย์กลางของกลุ่ม

(4) การกำหนดกลุ่มให้กับจุดข้อมูลจะใช้ระยะทางระหว่างจุดข้อมูลกับจุดมีเดียนของกลุ่มข้อมูลนั้น โดยจะกำหนดให้อยู่ในกลุ่มที่มีระยะทางที่สั้นที่สุด

(6) ใช้ค่าผิดพลาดยกกำลังสองที่น้อยสุดเป็นเงื่อนไขในการหยุดคำนวณ

จากอัลกอริทึมเคมีเดียนส์ สิ่งรบกวน และจุดข้อมูลขนาดเล็กที่ไม่เกี่ยวข้องกับข้อมูลหลักจะมีผลกระทบต่อจุดมีเดียนของกลุ่มน้อยกว่าจุดมีนของอัลกอริทึมเคมีนส์ นอกจากนี้อัลกอริทึมเคมีเดียนส์ยังสามารถนำไปใช้สำหรับการจัดกลุ่มข้อมูลที่มีรูปร่างมีเว้า หรือกลุ่มข้อมูลที่มีขนาดแตกต่างกันมาก ๆ ได้ แต่อย่างไรก็ตาม การคำนวณความซับซ้อนของอัลกอริทึมจะเป็น $O(N^2kt)$ ซึ่งมากกว่าอัลกอริทึมเคมีนส์ และผู้ใช้ยังจำเป็นต้องกำหนดค่า k เหมือนอัลกอริทึมเคมีนส์

ชื่ออัลกอริทึม : *เคมีเดียนส์*

หน้าที่ : *อัลกอริทึมสำหรับแบ่งส่วนโดยใช้ค่าศูนย์กลางของข้อมูลในกลุ่ม*

ข้อมูลเข้า : *จำนวนกลุ่มที่ต้องการแบ่ง และจุดข้อมูลจำนวน N ข้อมูล*

ผลลัพธ์ : *กลุ่มของข้อมูลที่มีความคล้ายคลึงกันจำนวนตามที่ต้องการ โดยที่มีค่าผิดพลาดยกกำลังสองน้อยที่สุด*

วิธีการ :

1) ทำการสุ่มค่าจุดข้อมูลจำนวน k ข้อมูล จากจุดข้อมูลทั้งหมด ซึ่งใช้เป็นตัวแทนของค่าจุดศูนย์กลางของกลุ่ม

2) **ทำซ้ำข้อต่อไปนี้**

2.1) กำหนดจุดข้อมูลที่เหลืออยู่ให้กับแต่ละกลุ่มที่มีความคล้ายกันบนพื้นฐานของระยะทางระหว่างจุดข้อมูลกับจุดศูนย์กลางของกลุ่ม

2.2) ทำการคำนวณหาจุดศูนย์กลางของแต่ละกลุ่มใหม่

จนกระทั่ง *ไม่มีการเปลี่ยนกลุ่มของข้อมูล ซึ่งหมายถึงค่าผิดพลาดยกกำลังสองน้อยที่สุดแล้ว*

(Sanpawat Kantabutra, 2001)

2.3 การจัดกลุ่มข้อมูลแบบลำดับขั้น

การจัดกลุ่มข้อมูลแบบลำดับขั้น เป็นเทคนิคการจัดกลุ่มข้อมูลโดยจัดกลุ่มข้อมูลออกเป็นลำดับขั้น ซึ่งมีขนาดกลุ่มหลายระดับ ตั้งแต่กลุ่มขนาดเล็กจนถึงใหญ่ จึงทำให้สามารถดูการจัดกลุ่มของข้อมูลได้ทั้งแบบคร่าว ๆ และแบบละเอียด ซึ่งผลลัพธ์สุดท้ายของการจัดกลุ่มจะได้เป็นต้นไม้ของการจัดกลุ่ม (Clustering Tree) วิธีการที่ใช้ในการจัดกลุ่มข้อมูลแบบลำดับขั้น ได้แก่ วิธีการรวมกันเข้าเป็นกลุ่ม (Agglomerative Method) และวิธีการแยกย่อย (Divisive Method)

2.3.1 วิธีการรวมกันเข้าเป็นกลุ่ม

วิธีการรวมกันเข้าเป็นกลุ่ม เป็นวิธีการจัดกลุ่มข้อมูลจากกลุ่มข้อมูลขนาดเล็กไปจนถึงขนาดใหญ่ โดยการค่อย ๆ จัดกลุ่มให้มีขนาดใหญ่ขึ้นเรื่อย ๆ ซึ่งเริ่มแรกจะให้แต่ละข้อมูลเป็นหนึ่งกลุ่ม จากนั้นทำการคำนวณหาความคล้ายคลึงกันของข้อมูลจากระยะทาง หากสองข้อมูลใดมีระยะทางใกล้กันก็ทำการจับคู่ให้อยู่ในกลุ่มเดียวกัน ทำให้จำนวนกลุ่มที่ได้ลดลงไปทีละครึ่ง ทำซ้ำไปเรื่อย ๆ จนได้ข้อมูลรวมกันเป็นหนึ่งกลุ่มข้อมูล ก็จะได้ต้นไม้ของการจัดกลุ่มออกมา

2.3.2 วิธีการแยกย่อย

วิธีการแยกย่อย เป็นวิธีการจัดกลุ่มข้อมูลจากกลุ่มข้อมูลขนาดใหญ่ไปจนถึงขนาดเล็ก โดยการแบ่งกลุ่มให้มีขนาดเล็กลงเรื่อย ๆ ซึ่งเริ่มต้นจากให้ข้อมูลทั้งหมดเป็นหนึ่งกลุ่มข้อมูล จากนั้นจะพิจารณาระยะทางระหว่างข้อมูล หากมีระยะห่างอยู่ในระดับหนึ่ง ก็จะทำการจัดให้อยู่ในกลุ่มเดียวกัน ถ้ามีระยะทางห่างกันมากกว่าระดับที่ได้กำหนดไว้ ก็จะไม่จัดให้อยู่กลุ่มเดียวกัน ทำซ้ำไปเรื่อยๆ จะทำให้กลุ่มที่แบ่งได้ค่อย ๆ เล็กลง ในขณะเดียวกันก็จะได้จำนวนกลุ่มที่มากขึ้นเรื่อย ๆ เช่นกัน สุดท้ายก็จะได้เป็นต้นไม้ของการจัดกลุ่มออกมา

(Jiawei Han and Micheline Kamber, 2001)

ดังที่ได้กล่าวมาแล้วว่าเทคนิคที่เกี่ยวข้องกับการจัดกลุ่มข้อมูลมีอยู่เป็นจำนวนมาก ไม่ว่าจะเป็นเทคนิคการจัดกลุ่มข้อมูลแบบแบ่งออกเป็นส่วน หรือการจัดกลุ่มข้อมูลแบบลำดับขั้น ซึ่งในการเลือกใช้เทคนิคในการจัดกลุ่มแบบใดจะขึ้นอยู่กับชนิดของข้อมูลที่จะทำการจัดกลุ่ม และวัตถุประสงค์เฉพาะในการใช้งาน จะเห็นว่าในการจัดกลุ่มบทคัดย่อทางวิชาการ ต้องการจัดกลุ่มบทคัดย่อออกเป็นกลุ่ม ๆ ตามจำนวนกลุ่มที่ต้องการ และในการจัดกลุ่มนั้นก็ต้องการข้อมูลศูนย์กลางที่เป็นตัวแทนแสดงลักษณะสำคัญของกลุ่มได้ ดังนั้นจึงต้องใช้จุดศูนย์กลางที่เป็นข้อมูลภายในกลุ่มนั้น ๆ ซึ่งเป็นข้อมูลที่มีอยู่จริง อีกทั้งข้อมูลบทคัดย่อก็มีลักษณะที่หลากหลาย อัลกอริทึมเคมีเดียนส์จึงมีความเหมาะสมที่จะนำมาใช้เพื่อทำการจัดกลุ่มบทคัดย่อทางวิชาการ