ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)

ปริญญา

วิศวกรรมคอมพิวเตอร์	วิศวกรรมคอมพิวเตอร์
สาขา	**ภาควิชา**

เรื่อง การผสมผสานเงื่อนไขในการแบ่งกลุ่มกระแสข้อมูลแบบกึ่งมีผู้สอน

Integrating Constraints in Semi-supervised Stream Clustering

นามผู้วิจัย นายทศพร ศิรามพุช

ได้พิจารณาเห็นชอบโดย

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก .....

(รองศาสตราจารย์กฤษณะ ไวยมัย, D.U.)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม .....

(อาจารย์สรรพฤทธิ์ มฤคทัต, Ph.D.)

หัวหน้าภาควิชา .....

(ผู้ช่วยศาสตราจารย์ภุชงค์ อุทโยภาศ, Ph.D.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

.....

(รองศาสตราจารย์กัญจนา ธีระกุล, D.Agr.)

คณบดีบัณฑิตวิทยาลัย

วันที่ ..... **เดือน** ..... **พ.ศ.** .....

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

วิทยานิพนธ์

เรื่อง

การผสมผสานเงื่อนไขในการแบ่งกลุ่มกระแสข้อมูลแบบกึ่งมีผู้สอน

Integrating Constraints in Semi-supervised Stream Clustering

โดย

นายทศพร ศิรามพุช

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์
เพื่อความสมบูรณ์แห่งปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)
พ.ศ. 2556

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

ทศพร ศิรามพุช 2556: การผสมผสานเงื่อนไขในการแบ่งกลุ่มกระแสข้อมูลแบบกึ่งมีผู้สอน ปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์) สาขาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก: รองศาสตราจารย์กฤษณะ ไวยมัย, D.U. 57 หน้า

ในปัจจุบันการแบ่งกลุ่มกระแสข้อมูลมีการศึกษาวิจัยและใช้งานอย่างแพร่หลาย อย่างไรก็ตาม เทคนิคเหล่านี้ยังขาดการใช้ความช่วยเหลือจาก ความรู้ของผู้เชี่ยวชาญ หรือ ประสบการณ์จากในอดีต ซึ่งสามารถประยุกต์ใช้ใน แอพพลิเคชั่นจริง ข้อดีของการใช้เทคนิคเหล่านี้จะช่วยให้ประสิทธิภาพและความแม่นยำของกลุ่มข้อมูลผลลัพธ์ดีขึ้น โดยงานวิจัยนี้นำเสนอการแบ่งกลุ่มกระแสข้อมูลที่ใช้ความรู้มาช่วยในรูปแบบของ เงื่อนไข (Constraints) ชื่อว่า CE-Stream

เงื่อนไขในระดับข้อมูลถูกนำมาใช้เพื่อช่วยเหลือการแบ่งกลุ่มข้อมูลให้ดียิ่งขึ้น ได้แก่ การใช้เงื่อนไขแบบเชื่อมต่อในการแยกตัวของกลุ่มข้อมูล และ การใช้เงื่อนไขแบบไม่เชื่อมต่อในการส่งมอบข้อมูลสมาชิก รวมไปถึง การออกแบบเงื่อนไขให้รองรับกับ พฤติกรรมของกระแสข้อมูลลักษณะต่างๆ อาทิเช่น การพิจารณาการใช้งานของเงื่อนไข (Constraint Activation) การเลือนหายของเงื่อนไข (Constraints Fading) และ การหมดอายุของเงื่อนไข (Constraints Outdating) ยิ่งไปกว่านั้น CE-Stream ยังช่วยลด อัตราการแบ่งแยกกลุ่มข้อมูลที่ไม่จำเป็นให้น้อยลงในระหว่างขั้นตอนการแบ่งกลุ่มข้อมูล และ การส่งมอบสมาชิกข้อมูลที่ผิดอีกด้วย

จากผลการทดลองกับชุดข้อมูลมาตรฐาน Covertype และ KDDcup’99 ซึ่งเป็น ชุดข้อมูลการแบ่งกลุ่มประเภทป่าฝนจากภูมิประเทศ และ ชุดข้อมูลการตรวจจับการบุกรุกเครือข่ายตามลำดับ วัดผลด้วยค่าเอฟเมเชอร์และค่าความบริสุทธิ์ พบว่าผลการทดลองเมื่อเทียบกับเทคนิค E-Stream ดั้งเดิมมีประสิทธิภาพดีขึ้น

_____ _____ ___ / ___ / ___

ลายมือชื่อนิสิต ลายมือชื่ออาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

Tossaporn Sirampuj 2013: Integrating Constraints in Semi-supervised Stream Clustering. Master of Engineering (Computer Engineering), Major Field: Computer Engineering, Department of Computer Engineering. Thesis Advisor: Associate Professor Kitsana Waiyamai, D.U. 57 pages.

Large number of stream clustering techniques has been proposed in recent years. However, these techniques still lack of using background knowledge which is available from domain expert. The advantages of using knowledge improve accuracy and performance of final clusters. In this research work, CE-Stream, an incremental method for stream clustering by using background knowledge as constraints is proposed.

Instance-level constraints have been used to guide better clustering behaviors i.e. using Must-Link constraints on Cluster Splitting and using Cannot-Link constraints on Cluster Assignment. Also, constraint operators are introduced to support evolving characteristics of dynamic constraints i.e. constraint activation, fading and outdating. Constraints operators seamlessly integrate into E-Stream to check active and update constraints and prioritize constraints. Likewise, CE-Stream reduces an excessive splitting during clustering process and misguide of cluster assignment.

Experimental results, using Covertype and KDDCup'99 datasets which are Rain forest typology by geographical and Network Intrusion Detection respectively, show that both F-measure and Purity increased with respect to an original technique E-Stream.

_____ _____ ___ / ___ / ___

Student's signature Thesis Advisor's signature

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

กิตติกรรมประกาศ

ข้าพเจ้าขอกราบขอบพระคุณ รองศาสตราจารย์ ดร.กฤษณะ ไวยมัย อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ได้ประสิทธิ์ประสาทวิชาความรู้และทฤษฎีต่าง ๆ ตลอดจนเป็นที่ปรึกษาในการวางแนวทางวางแผนงาน แก้ปัญหา ตรวจสอบข้อบกพร่องต่าง ๆ และกำลังใจ จนงานวิจัยนี้สำเร็จลุล่วงด้วยดี

ขอขอบคุณ คุณธนภัทร ฆังคะจิตร ที่คอยให้คำปรึกษาและให้ความช่วยเหลือในทุกๆเรื่องๆ และเพื่อนๆ พี่ๆ น้องๆ ในห้องปฏิบัติการวิจัยการวิเคราะห์ข้อมูลและสืบค้นความรู้ (DAKDL) ที่ทำให้มีบรรยากาศดีๆ และช่วยเหลือกัน

ขอขอบคุณ คุณเจฟ (Mr. Jeffery James Dalgliesh) และคุณดั๊ก (Mr. Douglas Marvin Kypfer) จากบริษัท เชฟรอน ที่ให้ความเชื่อมั่นในตัวของข้าพเจ้า และเป็นกำลังใจเสมอมา

ขอขอบคุณ คุณพิมพ์ใจ ทายะติ ที่สนับสนุนในเรื่องสถานที่ ทรัพยากร ในการทำวิจัย และความช่วยเหลือทุกๆเรื่อง และเป็นกำลังใจ คอยรับฟังปัญหา อยู่เสมอ

ขอขอบคุณเจ้าหน้าที่โครงการบัณฑิตศึกษา และเจ้าหน้าที่ธุรการภาควิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเกษตรศาสตร์ที่ช่วยเหลือในการประสานงาน และดำเนินงานด้านเอกสารต่างๆ ให้เป็นไปอย่างสะดวกลุล่วงไปด้วยดี

สุดท้ายนี้ขอขอบพระคุณ คุณพ่อ คุณแม่ และครอบครัวที่มีความเชื่อมั่น สนับสนุน และห่วงใยข้าพเจ้าตลอดมา

ทศพร ศิรามพุช

เมษายน 2556

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

สารบัญ

หน้า

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

สารบัญตาราง

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

สารบัญภาพ

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

สารบัญภาพ (ต่อ)

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

สารบัญภาพ (ต่อ)

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

การผสมผสานเงื่อนไขในการแบ่งกลุ่มกระแสข้อมูลแบบกึ่งมีผู้สอน

Integrating Constraints in Semi-supervised Stream Clustering

คำนำ

การแบ่งกลุ่มข้อมูล (Clustering) เป็นเทคนิคการแบ่งแยกกลุ่มข้อมูล โดยพิจารณาจากความคล้ายคลึงของข้อมูล ให้ข้อมูลที่มีความคล้ายกันอยู่ในกลุ่มข้อมูลเดียวกัน และข้อมูลที่แตกต่างกันอยู่ต่างกลุ่มข้อมูล แต่ในแอพพลิเคชั่นจริง การแบ่งกลุ่มข้อมูลสามารถพัฒนาและเพิ่มประสิทธิภาพได้โดยการใช้องค์ความรู้จากผู้เชี่ยวชาญ หรือประสบการณ์จากอดีต ในรูปแบบของเงื่อนไข (Constraints) การนำเงื่อนไขมาใช้ในการแบ่งกลุ่มข้อมูลสามารถช่วยให้การแบ่งกลุ่มข้อมูลมีประสิทธิภาพและคุณภาพของกลุ่มข้อมูลผลลัพธ์ดีขึ้น

การแบ่งกลุ่มกระแสข้อมูล คือ การพัฒนาเทคนิคการแบ่งกลุ่มข้อมูล ให้รองรับกับ ข้อมูลแบบกระแส ที่เกิดขึ้นอย่างต่อเนื่อง มีจำนวนเป็นอนันต์ หรือเป็นกระแสที่ไม่สิ้นสุด โดย เทคนิคส่วนมากนั้นจะนำเสนอ วิธีการเก็บตัวแทนของกลุ่มข้อมูล (Cluster Representation) เนื่องจากข้อจำกัดของกระแสข้อมูล ในเชิงของเวลา และ หน่วยความจำ รวมไปถึง เทคนิค ในการรองกับการเปลี่ยนแปลงของข้อมูลตามช่วงของเวลา ดังนั้นเทคนิคการแบ่งกลุ่มกระแสข้อมูลจึงต้องมีความรวดเร็วในการประมวลผลข้อมูลอย่างต่อเนื่องและใช้หน่วยความจำได้อย่างมีประสิทธิภาพ

งานวิจัยเกี่ยวกับการผสมผสานเงื่อนไขในการแบ่งกลุ่มกระแสข้อมูลยังมีจำนวนไม่มากและมีจำกัด เนื่องจาก ความแตกต่างกันของโครงสร้าง อาทิเช่น ข้อจำกัดในเรื่องของความเร็วในการประมวลผลเพื่อให้รองรับกับข้อมูลที่เกิดขึ้นอย่างต่อเนื่องที่ไม่สามารถพิจารณาเงื่อนไขทั้งหมดภายในเวลาที่จำกัด และ ข้อจำกัดในเรื่องของพฤติกรรมการเปลี่ยนแปลงของกลุ่มข้อมูลที่ไม่คงที่เปลี่ยนตามกาลเวลาซึ่งทำให้เงื่อนไขแบบข้อมูลคงที่ไม่เหมาะสม จึงทำให้ไม่สามารถประยุกต์ใช้เงื่อนไขในการแบ่งกลุ่มข้อมูลมาใช้ในกระแสข้อมูลได้ อย่างไรก็ตาม แนวทางของงานวิจัยทางด้านนี้ได้นำเสนอ การประยุกต์ใช้เงื่อนไขในระดับต่างๆ เช่น เงื่อนไขระดับข้อมูล (Instance-level Constraints) หรือ เงื่อนไขระดับกลุ่มข้อมูล (Cluster-level Constraints) ที่สามารถรองรับพฤติกรรม

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

การเปลี่ยนแปลงของข้อมูลได้ เช่น การปรับสภาวะของเงื่อนไขตามกาลเวลา การพิจารณาเงื่อนไขกับตัวแทนของกลุ่มข้อมูล และการนำเงื่อนไขไปใช้ในขั้นตอนต่างๆของการแบ่งกลุ่มกระแสข้อมูล

จากการศึกษาเทคนิคการแบ่งกลุ่มกระแสข้อมูลต่างๆ พบว่า เทคนิค E-Stream หรือ การแบ่งกลุ่มกระแสข้อมูลเชิงวิวัฒนาการ เป็นงานวิจัยที่มีความน่าสนใจแตกต่างจากเทคนิคอื่นๆในเรื่องของการรองรับพฤติกรรมการเปลี่ยนแปลงของกระแสข้อมูลได้หลายรูปแบบ อาทิเช่น การรวมกันของกลุ่มข้อมูล และ การแยกตัวของกลุ่มข้อมูล ดังนั้นงานวิจัยนี้จึงต้องการนำเสนอเทคนิค CE-Stream การแบ่งกลุ่มกระแสข้อมูลแบบมีเงื่อนไข ซึ่งพัฒนาต่อยอดจาก เทคนิค E-Stream เพื่อเพิ่มประสิทธิภาพและปรับปรุงให้คุณภาพของกลุ่มข้อมูลผลลัพธ์ดีขึ้น โดยงานวิจัยนี้ได้นำเงื่อนไขระดับข้อมูลมาประยุกต์ใช้ ได้แก่ การพิจารณาเงื่อนไขแบบเชื่อมต่อ (Must-Link Constraints) ในการแบ่งแยกกลุ่มข้อมูล และ การพิจารณาเงื่อนไขแบบไม่เชื่อมต่อ (Cannot-Link Constraints) ในการส่งมอบสมาชิกข้อมูล รวมถึงการออกแบบเงื่อนไขให้เหมาะสมและสอดคล้องกับกระแสข้อมูลนั่นคือ การเลือนหายของเงื่อนไข (Constraints Fading) การหมดอายุของเงื่อนไข (Constraints Outdating) และ การพิจารณาใช้เงื่อนไข (Constraints Activation) โดยการผสมผสานเงื่อนไขในการแบ่งกลุ่มกระแสข้อมูล นอกจากจะช่วยเพิ่มคุณภาพของกลุ่มข้อมูลผลลัพธ์แล้ว ยังสามารถลดอัตราการส่งข้อมูลสมาชิกให้กลุ่มข้อมูลที่ผิดพลาด และการแบ่งแยกกลุ่มข้อมูลที่ไม่จำเป็น นำไปสู่การใช้เวลาประมวลผลที่น้อยลง

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

วัตถุประสงค์

1. ศึกษาเทคนิคการผสมผสานเงื่อนไขในการแบ่งกลุ่มของข้อมูลประเภทต่างๆ เพื่อให้เกิดความรู้ความเข้าใจ และ สามารถนำมาพัฒนาเป็นเทคนิคของตนเองได้ แบ่งออกเป็น

1.1 การผสมผสานเงื่อนไขประเภทต่างๆในการแบ่งกลุ่มข้อมูล

1.2 การผสมผสานเงื่อนไขประเภทต่างๆในการแบ่งกลุ่มกระแสข้อมูล

2. พัฒนาเทคนิคการผสมผสานเงื่อนไขในการแบ่งกลุ่มกระแสข้อมูลแบบ E-Stream เพื่อเพิ่มคุณภาพของกลุ่มข้อมูลผลลัพธ์ โดยการ

2.1 พัฒนาโครงสร้างการทำงานของระบบเพื่อรองรับการแยกตัวของกลุ่มข้อมูล (Cluster Splitting) โดยใช้เงื่อนไขแบบเชื่อมต่อ (Must-Link Constraints)

2.2 พัฒนาโครงสร้างการทำงานของระบบเพื่อรองรับการส่งมอบสมาชิกให้กลุ่มข้อมูล (Cluster Assignment) โดยใช้เงื่อนไขแบบไม่เชื่อมต่อ (Cannot-Link Constraints)

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

ขั้นตอนการวิจัย

1. ศึกษางานวิจัยและทฤษฎีต่างๆ ที่เกี่ยวข้องในการประยุกต์ใช้เงื่อนไขในการแบ่งกลุ่มข้อมูลและ การประยุกต์ใช้เงื่อนไขในการแบ่งกลุ่มกระแสข้อมูล รวมทั้งวิเคราะห์ลักษณะเฉพาะของเงื่อนไขที่รองรับกับกระแสข้อมูลในงานวิจัยต่างๆ เพื่อที่จะนำความรู้ที่ได้มาใช้ในการวิจัยนี้

2. ออกแบบแนวทางการวิจัย วิธีการพัฒนาระบบ และวิธีการวัดผลการทดลอง

3. พัฒนาต่อยอดการแบ่งกลุ่มกระแสข้อมูล E-Stream เพื่อให้รองรับกับเงื่อนไข นวมถึงการออกแบบเงื่อนไขให้สนับสนุนลักษณะเฉพาะของกระแสข้อมูล

4. ทดสอบและวัดผลของระบบที่พัฒนาขึ้น

5. วิเคราะห์ผลและสรุปผลการวิจัยและประโยชน์ที่ได้รับ

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

การตรวจเอกสาร

ในหัวข้อนี้จะกล่าวถึง การแบ่งกลุ่มข้อมูล การแบ่งกลุ่มกระแสข้อมูล การใช้เงื่อนไขในการแบ่งกลุ่มกระแสข้อมูล งานวิจัยต่างๆ ที่เกี่ยวข้อง และสถิติพื้นฐานที่ควรทราบ

1. การแบ่งกลุ่มข้อมูล

การแบ่งกลุ่มข้อมูล (Data Clustering) เป็นการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) เป็นส่วนหนึ่งของ การทำเหมืองข้อมูล เทคนิคนี้ จะทำการแบ่งข้อมูลทั้งหมดเป็นกลุ่มย่อยๆ ที่เรียกว่า กลุ่มข้อมูล (Cluster) โดยใช้ลักษณะความคล้ายคลึงกันของข้อมูลเป็นตัวชี้วัด ข้อมูลที่มีลักษณะใกล้เคียงกันจะถูกจัดให้อยู่ในหมวดหมู่เดียวกัน ในทางตรงกันข้าม ข้อมูลที่มีลักษณะแตกต่างกันจะถูกจัดให้อยู่ในต่างหมวดหมู่กัน (Jain et al., 1999) เทคนิคนี้สามารถนำไปประยุกต์ใช้ได้ในงานต่างๆ เช่น การแบ่งกลุ่มลูกค้าตามลักษณะการใช้งานเพื่อที่จะจัดผลิตภัณฑ์ส่งเสริมการขายที่เหมาะสมกลุ่มลูกค้า

การแบ่งกลุ่มข้อมูลจำแนกออกเป็นหลายประเภท เทคนิคที่เป็นที่นิยมและมีผู้ใช้งานอย่างแพร่หลาย อาทิเช่น การแบ่งกลุ่มข้อมูลแบบอ้างอิงจุดกึ่งกลาง การแบ่งกลุ่มข้อมูลแบบอ้างอิงความหนาแน่น มีรายละเอียดดังนี้

การแบ่งกลุ่มข้อมูลแบบอ้างอิงจุดกึ่งกลาง (Centroid-based Clustering) กลุ่มข้อมูลจะถูกแทนด้วยจุดกึ่งกลางของกลุ่มข้อมูลเพื่อใช้ในการพิจารณาของระบบ เทคนิคที่ใช้อย่างแพร่หลายโดยอ้างอิงจุดกึ่งกลาง คือ K-means clustering (MacQueen et al., 1967) ซึ่ง เป็นการแบ่งกลุ่มข้อมูลโดยแบ่งตามจำนวนค่า k โดยแต่ละจุดข้อมูลจะถูกแบ่งตามจุดกึ่งกลางของแต่ละกลุ่มข้อมูลที่อยู่ใกล้ที่สุด (Closest Centriod)

การแบ่งกลุ่มข้อมูลแบบอ้างอิงความหนาแน่น (Density-based Clustering) กลุ่มข้อมูลจะถูกแทนด้วยพื้นที่ของความหนาแน่น ในส่วนข้อมูลมีความหนาแน่นสูงจะถูกจัดให้เป็นกลุ่มข้อมูล และในส่วนที่มีความหนาแน่นต่ำจะถูกจัดให้เป็นข้อมูลส่วนเกิน (Noise) หรือ ขอบเขต (Border) เทคนิคที่ใช้อย่างแพร่หลายคือ DBSCAN (Ester et al., 1996)

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

2. การแบ่งกลุ่มกระแสข้อมูล

กระแสข้อมูล (Zengyou He et al., 2004) เกิดจากการเกิดขึ้นอย่างต่อเนื่องของข้อมูล ซึ่งมีจำนวนไม่จำกัด และสามารถเปลี่ยนแปลงพฤติกรรมได้ตลอดเวลา มีหลายแอพพลิเคชั่นที่ทำให้เกิดข้อมูลในรูปแบบของกระแส เช่น ข้อมูลการเชื่อมต่อเครือข่าย ข้อมูลในการสืบค้นหาเว็บไซต์ ข้อมูลการใช้บัตรเครดิต เป็นต้น การแบ่งกลุ่มกระแสข้อมูลเกิดจากความต้องการวิเคราะห์ข้อมูลที่มีลักษณะเป็นกระแส การใช้เทคนิคการแบ่งกลุ่มข้อมูลแบบคงที่นั้นไม่สามารถรองรับข้อมูลจำนวนมหาศาลที่เกิดขึ้นอยู่ตลอดเวลานี้ได้ เพราะ เทคนิคการแบ่งกลุ่มข้อมูลแบบคงที่นั้น จะสามารถเรียกใช้ข้อมูลทุกตัวที่ถูกเก็บไว้ในหน่วยความจำได้ตลอดเวลา แต่สำหรับ กระแสข้อมูลนั้น ข้อมูลจะสามารถเก็บในหน่วยความจำได้เพียงชั่วคราว ดังนั้น เทคนิคการแบ่งกลุ่มกระแสข้อมูลจึงต้องออกแบบเพื่อรองรับกับการอ่านข้อมูลเพียงชั่วคราวและมีความรวดเร็วในการประมวลผลรวมไปถึงการใช้ทรัพยากรในการประมวลผลอย่างมีประสิทธิภาพ ที่ผ่านมาได้มีงานวิจัยมากมายที่พัฒนาเทคนิคการแบ่งกลุ่มกระแสข้อมูล หนึ่งในนั้น คือเทคนิค E-Stream (Udommanetanakit. et al., 2007)

2.1 E-Stream: เทคนิคการแบ่งกลุ่มกระแสข้อมูลเชิงวิวัฒนาการ

เทคนิค E-Stream เสนอแนวคิดในการแบ่งกลุ่มกระแสข้อมูล โดยเก็บตัวแทนของกลุ่มข้อมูล (Cluster Representative) แทนการเก็บข้อมูลทั้งหมดเนื่องจากข้อจำกัดทางด้านหน่วยความจำและการให้ความสำคัญกับการเปลี่ยนแปลงของข้อมูลที่เกิดขึ้นอย่างรวดเร็วและต่อเนื่องของกระแสข้อมูล โดยแบ่งประเภทของการเปลี่ยนแปลงพฤติกรรมของกลุ่มข้อมูล ออกเป็น 5 ประเภท ได้แก่ การเกิดขึ้น การหายไป การเปลี่ยนแปลงตัวเอง การรวมตัวกัน และ การแยกตัวกัน

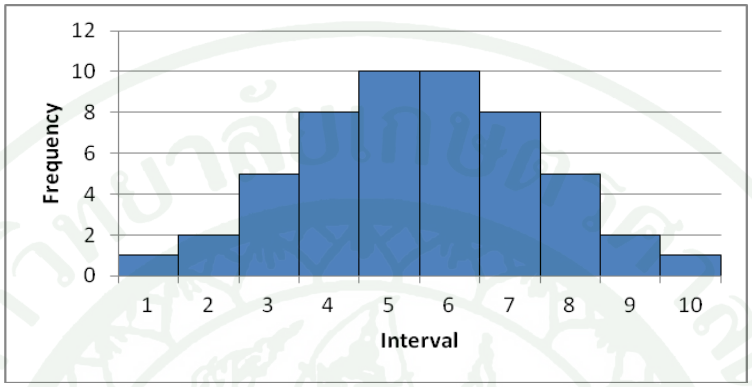
โดยรายละเอียดการทำงานของเทคนิค E-Stream จะถูกนำไปรวมและอธิบายภายใต้หัวข้ออัลกอริทึม CE-Stream

3. สถิติพื้นฐานที่ควรทราบ

ในงานวิจัยนี้ได้อ้างอิงค่าทางสถิติ คือ ฮิสโตแกรม หัวข้อนี้เราจะอธิบายถึงความหมายของฮิสโตแกรมที่ใช้เพื่อเป็นประโยชน์ในการอธิบายหัวข้อถัดไป

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

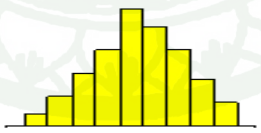
ฮิสโตแกรม เป็นการแสดงผลของชุดข้อมูลในรูปของแท่งความถี่ โดยข้อมูลทั้งหมดจะถูกแบ่งเป็นช่วงๆ แสดงเป็นแผนภูมิแท่งซึ่งแสดงความถี่ของข้อมูลในช่วงนั้น ดังตัวอย่างในภาพที่ 1 ข้อมูลแบ่งออกเป็น 10 ช่วง และมีความถี่ 1, 2, 5, 8, 10, 10, 8, 5, 2, 1 ตามลำดับ



ภาพที่ 1 แสดงลักษณะทั่วไปของฮิสโตแกรม

ลักษณะของฮิสโตแกรมมีหลายแบบดังนี้ (Tague, 2004)

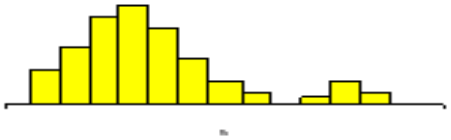
3.1 แบบปกติ (Normal Distribution) การกระจายเป็นไปตามปกติ ค่าเฉลี่ยส่วนใหญ่อยู่ตรงกลาง



ภาพที่ 2 ลักษณะของฮิสโตแกรมแบบปกติ

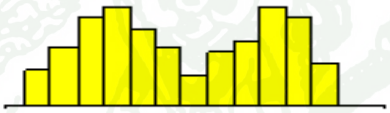
3.2 แบบแยกเป็นเกาะ (Detached Island Type) ข้อมูลที่มีความถี่สูงแยกออกจากข้อมูลที่มีความถี่ต่ำ อาจเกิดจากการเก็บข้อมูลที่มีข้อมูลอื่นมาปะปน

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์



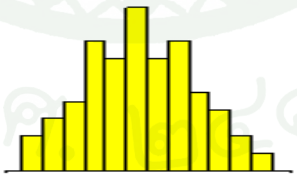
ภาพที่ 3 ลักษณะของฮิสโตแกรมแบบแยกเป็นเกาะ

3.3 แบบระฆังคู่ (Double Peaked Type) ลักษณะความถี่สูงสองยอดห่างกัน โดยตรงกลางเป็นความถี่ต่ำ เกิดจากข้อมูลที่นำมาแจกแจง 2 ชุดมีค่าเฉลี่ยไม่เท่ากัน ซึ่งอาจมาจาก 2 แหล่งข้อมูลที่ต่างกัน



ภาพที่ 4 ลักษณะของฮิสโตแกรมแบบระฆังคู่

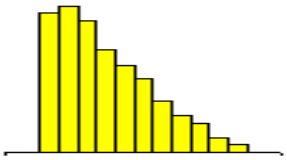
3.4 แบบฟันปลา (Plateau) ในแต่ละช่วงชั้นมีความถี่มากน้อยสลับกันไป เกิดขึ้นได้เมื่อจำนวนข้อมูลมีค่าไม่เท่ากัน และแตกต่างกันมากระหว่างช่วงชั้นที่ติดกัน



ภาพที่ 5 ลักษณะของฮิสโตแกรมแบบฟันปลา

3.5 แบบหน้าผา (Cliff Type, Skewed) ค่าเฉลี่ยอยู่ใกล้กับค่าขอบเขตต่ำ เนื่องจากค่าส่วนใหญ่อยู่ใกล้ค่าขอบเขตต่ำ

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์



ภาพที่ 6 ลักษณะของฮิสโตแกรมแบบหน้าผา

4. การวัดผลคุณภาพของข้อมูล

การวัดผลเพื่อตรวจสอบว่าการแบ่งกลุ่มที่เกิดขึ้นเป็นการแบ่งกลุ่มที่ดีหรือไม่ โดยการเปรียบเทียบกับกลุ่มข้อมูลที่ทราบผลเฉลยแล้ว ตัวอย่างของการวัดผลโดยวิธีนี้คือการวัดค่าเอฟเมเชอร์ (F-Measure) และการวัดค่าความบริสุทธิ์ของกลุ่มข้อมูล (Purity) ซึ่งงานวิจัยนี้ใช้ทั้งสองค่าในการวัดผล

4.1 ค่าเอฟเมเชอร์ (F-Measure)

เป็นการวัดผลแบบภายนอก ซึ่งเป็นการรวมกันระหว่างการวัดผลโดยใช้ค่าความถูกต้อง (Precision) และค่าความครบถ้วน (Recall) การวัดผลนี้จะมองแต่ละกลุ่มข้อมูลผลลัพธ์ที่แบ่งได้ เปรียบเทียบกับกลุ่มข้อมูลผลเฉลย กำหนดให้ i เป็นกลุ่มข้อมูลผลเฉลย มีจำนวน n_i และ j เป็นกลุ่มข้อมูลผลลัพธ์ มีจำนวน n_j ค่าเอฟเมเชอร์สามารถหาได้จากสมการ

$$F(i, j) = \frac{2 * recall(i, j) * precision(i, j)}{recall(i, j) + precision(i, j)} \quad \textbf{(1)}$$

โดยค่าความครบถ้วน ของกลุ่ม i ในกลุ่มข้อมูลผลลัพธ์ j คิดได้จากอัตราส่วนจำนวนข้อมูลของกลุ่ม i ซึ่งอยู่ในกลุ่มข้อมูลผลลัพธ์ j ต่อจำนวนข้อมูลของกลุ่ม i ทั้งหมด ค่าความครบถ้วนสามารถหาได้จากสมการ

$$recall(i, j) = \frac{n_{ij}}{n_i} \quad \textbf{(2)}$$

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

ค่าความถูกต้องของกลุ่ม i ในกลุ่มข้อมูลผลลัพธ์ j คิดได้จากอัตราส่วนจำนวนข้อมูลของกลุ่ม i ซึ่งอยู่ในกลุ่มข้อมูลผลลัพธ์ j ต่อจำนวนข้อมูลของกลุ่มข้อมูลผลลัพธ์ j ทั้งหมด ค่าความถูกต้องสามารถหาได้จากสมการ

$$precision(i, j) = \frac{n_{ij}}{nj} \quad \textbf{(3)}$$

และกำหนดให้ n คือจำนวนข้อมูลทั้งหมด ค่าเอฟเมเชอร์รวมสามารถหาได้จากสมการ

$$F = \sum_i \frac{n_i}{n} \max\{F(i, j)\} \quad \textbf{(4)}$$

4.2 ค่าความบริสุทธิ์ (Purity)

การวัดค่าความบริสุทธิ์ เป็นการวัดผลที่พิจารณาว่าในแต่ละกลุ่ม สามารถแยกข้อมูลได้อย่างชัดเจนมากน้อยเพียงใด โดยการคิดค่ารวมแบบถ่วงน้ำหนักของความถูกต้องที่มากที่สุดของกลุ่มข้อมูลในทุกกลุ่ม (ในกลุ่มข้อมูลผลลัพธ์ j ใดๆ จะเลือกค่าความถูกต้องที่มากที่สุดสำหรับข้อมูลในกลุ่ม i จากนั้นจะนำมาคิดผลรวมแบบถ่วงน้ำหนักของความถูกต้องเหล่านั้น) การหาค่าความบริสุทธิ์มีสมการดังนี้

$$purity = \sum_j \frac{n_j}{n} \max\{precision(i,j)\} \quad \textbf{(5)}$$

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

งานวิจัยที่เกี่ยวข้อง

ในวิทยานิพนธ์ฉบับนี้ ได้ทำการผสมผสานเงื่อนไขลงในกระแสข้อมูลจึงมีการตรวจสอบเอกสารที่เกี่ยวข้องกับเงื่อนไขในรูปแบบต่างๆ และ งานวิจัยที่ผ่านมาที่เกี่ยวข้องกับการผสมผสานเงื่อนไขลงในการแบ่งกลุ่มข้อมูลและการแบ่งกลุ่มกระแสข้อมูล ซึ่งมีรายละเอียดดังต่อไปนี้

1. เงื่อนไข (Constraints)

ในปัจจุบันยังไม่มีการนิยามความหมายของเงื่อนไขได้อย่างชัดเจนใน ศาสตร์ทางด้านการเรียนรู้ของเครื่อง (Machine Learning) อย่างไรก็ตาม ความหมายของเงื่อนไขในเชิงระบบการจัดการฐานข้อมูล (Database Management System) นั้นได้นิยามไว้ว่า ข้อบังคับหรือเงื่อนไขในการอณุญาติให้เก็บเฉพาะข้อมูลที่เหมาะสมลงในฐานข้อมูล (Codd, 1970) โดยเราสามารถใช้นิยามนี้มาดัดแปลงในศาสตร์ทางด้านการเรียนรู้ของเครื่องได้เช่นกัน โดยที่ เงื่อนไขนั้นจะหมายความว่า ข้อบังคับหรือเงื่อนไขเพื่อช่วยเหลือให้ระบบทำงานได้อย่างถูกต้องและมีประสิทธิภาพมากขึ้น เงื่อนไขสามารถแบ่งได้หลายประเภท ดังต่อไปนี้

1.1 เงื่อนไขระดับจุดข้อมูล (Instance-Level Constraints) คือ เงื่อนไขที่ถูกพิจารณาในระดับของจุดข้อมูล ตัวอย่างเช่น เงื่อนไขแบบเชื่อมต่อ (Must-Link Constraints) หมายถึง คู่ของจุดข้อมูลที่ควรอยู่ในกลุ่มข้อมูลเดียวกัน และเงื่อนไขแบบไม่เชื่อมต่อ (Cannot-Link Constraints) หมายถึง คู่ของจุดข้อมูลที่ควรอยู่คนละกลุ่มข้อมูลกัน

1.2 เงื่อนไขระดับกลุ่มข้อมูล (Cluster-Level Constraints) คือ เงื่อนไขที่ถูกพิจารณาในระดับของกลุ่มข้อมูล ตัวอย่างเช่น เงื่อนไขในการกำหนดความกระชับของข้อมูลในแต่ละกลุ่มข้อมูล (Compactness Constraints) เงื่อนไขในการกำหนดจำนวนสมาชิกขั้นต่ำในแต่ละกลุ่มข้อมูล (Minimum Members Constraints) และ เงื่อนไขในการกำหนดระยะห่างระหว่างกลุ่มข้อมูล (Separation Constraints)

2. การผสมผสานเงื่อนไขในการแบ่งกลุ่มข้อมูล

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

ที่ผ่านมา มีการศึกษาและวิจัยเกิดขึ้นอย่างมากมายในการผสมผสานเงื่อนไขในการแบ่งกลุ่มข้อมูล โดยมีรายละเอียดดังต่อไปนี้

2.1 COP-KMeans Clustering

งานวิจัยฉบับนี้ ถือว่าเป็นงานวิจัยฉบับแรกๆ ที่สนใจในการนำเงื่อนไขมาใช้ โดยมีแนวคิดพัฒนาต่อยอดจากเทคนิค K-means Clustering ที่ได้กล่าวไว้ข้างต้น ซึ่ง COP-KMeans Clustering (Wagstaff et al., 2001) เป็นเทคนิคที่นำเงื่อนไขในระดับข้อมูลมาใช้ในการแบ่งกลุ่มข้อมูล โดยเงื่อนไขที่นำมาใช้ ได้แก่ เงื่อนไขแบบเชื่อมต่อ และ เงื่อนไขแบบไม่เชื่อมต่อ จากภาพที่ 7 ระบบจะเริ่มทำการต้นสร้างจุดกลางกลุ่มข้อมูล (บรรทัดที่ 1) และจะพิจารณาการส่งมอบข้อมูลสมาชิกให้แก่กลุ่มข้อมูลควบคู่ไปกับพิจารณาเงื่อนไข หากไม่มีการฝ่าฝืนเงื่อนไขใดๆ ข้อมูลจะถูกส่งให้กับกลุ่มข้อมูลที่ใกล้ที่สุดและจะปรับจุดกลางให้สอดคล้องกับข้อมูลใหม่ที่เข้ามา ดังในบรรทัดที่ 2 และ 3 ระบบจะทำซ้ำไปเรื่อยๆ จนกว่า การแบ่งกลุ่มข้อมูลจะสิ้นสุดลง

COP-KMEANS(data set D, must-link constraints $Con_= \subseteq D \times D$, cannot-link constraints $Con_\neq \subseteq D \times D$)

1. Let $C_1 \ldots C_k$ be the initial cluster centers.
2. For each point d_i in D, assign it to the closest cluster C_j **such that** VIOLATE-CONSTRAINTS(d_i, C_j, $Con_=$, $Con_\neq$) **is false. If no such cluster exists, fail (return {}).**
3. For each cluster C_i, update its center by averaging all of the points d_j that have been assigned to it.
4. Iterate between (2) and (3) until convergence.
5. Return $\{C_1 \ldots C_k\}$.

ภาพที่ 7 แสดงอัลกอริทึม COPK-Means Clustering

ผลการทดลองที่ได้เมื่อทำการทดลองบน 5 กลุ่มตัวอย่างได้แก่ soybean, mushroom, tic-tac-toe, part-of-speech และ iris ซึ่งทั้งหมดเป็นกลุ่มตัวอย่างมาตรฐาน UCI นั้นมีแนวโน้มของประสิทธิภาพที่ดีขึ้นอย่างเห็นได้ชัด รวมไปถึงได้นำไปประยุกต์ใช้ในแอพพลิเคชั่นจริงเพื่อแสดง

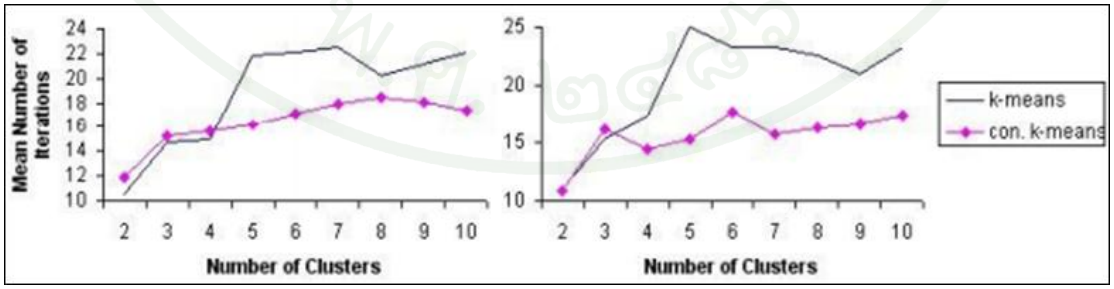
ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

ให้เห็นถึงความสำคัญในการใช้เงื่อนไข นั่นคือ การค้นหาเส้นแบ่งถนนของเครื่องติดตามด้วยระบบกำหนดตำแหน่งบนโลก (GPS) โดยใช้เงื่อนไขที่ได้มาจาก การเก็บข้อมูลของผู้ขับขี่ในการเปลี่ยนเส้นถนน เพื่อช่วยในการค้นหาเส้นแบ่งถนนของระบบ

อย่างไรก็ตาม เทคนิคนี้ยังสามารถปรับปรุงเพื่อเพิ่มประสิทธิภาพได้อีกมากมาย อาทิ เช่น การลำดับการใช้เงื่อนไขในในระบบ นั่นคือ หากระบบทำการเลือกใช้ลำดับของเงื่อนไขได้ไม่ดี อาจทำให้การแบ่งกลุ่มในช่วงแรกนั้นมีคุณภาพต่ำ นำไปสู่การใช้เวลาประมวลผลที่มากขึ้น

2.2 Constrained K-Means

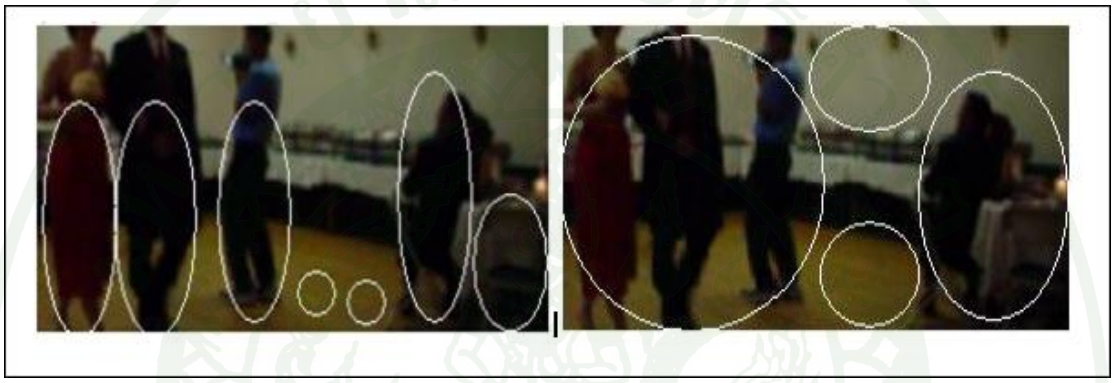
ในงานวิจัยนี้ (Davidson et al., 2005) ยังคงทำการพัฒนาต่อยอดจาก COP-KMeans Clustering โดยยังคงใช้เงื่อนไขระดับจุดข้อมูล และได้เพิ่มเงื่อนไขในระดับกลุ่มข้อมูลในระบบอีกด้วย ได้แก่ เงื่อนไขในการกำหนดความกระชับของข้อมูลในแต่ละกลุ่มข้อมูล (Compactness Constraints) และ เงื่อนไขในการกำหนดระยะห่างระหว่างกลุ่มข้อมูล (Minimum Separation Constraints) โดยงานวิจัยนี้ ได้กล่าวว่า เนื่องจากเงื่อนไขแบบ เชื่อมต่อและไม่เชื่อมต่อ นั้น สามารถก่อให้เกิดปัญหาในเชิงของเวลาการประมวลผลของ K-Means Clustering อาจใช้เวลานาน และระบบต้องสร้างกลุ่มข้อมูลใหม่หลายรอบจนกระทั่งการแบ่งกลุ่มข้อมูลคงที่ ดังที่ได้กล่าวไว้ข้างต้น เมื่อต้องทดลองผลกลุ่มข้อมูลซึ่งมีขนาดใหญ่มากอาจส่งผลให้ประสิทธิภาพลดลง นอกจากนั้น ผู้เขียนยังได้แสดงถึงความสอดคล้องของแต่ละเงื่อนไขในการใช้ร่วมกัน โดยมีการพิสูจน์ทางคณิตศาสตร์และใช้มาตรการวัดความซับซ้อน (Complexity Measurement)



ภาพที่ 8 แสดงผลการทดลองค่าเฉลี่ยจำนวนรอบของกลุ่มข้อมูลทดลอง Pima (ซ้าย) และ Breast Cancer (ขวา)

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

ผลการทดลองนั้นได้มีการเปรียบเทียบกับ K-Means Clustering บน 2 กลุ่มข้อมูลทดลองมาตรฐาน UCI ซึ่งได้แก่ PIMA และ Breast Cancer โดยเปรียบเทียบในเชิงของจำนวนรอบการแบ่งกลุ่มของระบบ ดังภาพที่ 8 นอกจากนั้น ยังได้ทดลองกับแอพพลิเคชั่นจริงกับหุ่นยนตร์สุนัขไอโบะ (Aibo robot) ในการแบ่งกลุ่มวัตถุที่มองเห็นโดยมีการเปรียบเทียบระหว่าง การมองเห็นโดยไม่มีเงื่อนไข และการมองเห็นโดยมีเงื่อนไข ซึ่งสามารถแสดงถึงประสิทธิภาพการแบ่งกลุ่มคนที่ดีขึ้นอย่างเห็นได้ชัด ดังภาพที่ 9



ภาพที่ 9 แสดงการมองเห็นหุ่นยนต์เปรียบเทียบระหว่างไม่มีเงื่อนไข (ซ้าย) และมีเงื่อนไข (ขวา)

2.3 C-DBSCAN

เทคนิค C-DBSCAN (Ruiz et al., 2007) เป็นเทคนิคพัฒนาต่อยอดจากการแบ่งกลุ่มข้อมูลแบบอ้างอิงความหนาแน่น DBSCAN โดยนำเงื่อนไขในระดับข้อมูลมาใช้ ได้แก่ เงื่อนไขแบบเชื่อมต่อ และไม่เชื่อมต่อ มีรายละเอียดขั้นตอนต่างๆที่มีการใช้เงื่อนไขดังต่อไปนี้

2.3.1 การสร้างกลุ่มข้อมูลในพื้นที่ภายใต้เงื่อนไขแบบไม่เชื่อมต่อ (Creating Local Clusters under Cannot-Link Constraints) จะสร้างกลุ่มข้อมูลให้แยกออกจากกันหากมีเงื่อนไขแบบไม่เชื่อมต่ออยู่

2.3.2 การรวมกลุ่มข้อมูลในพื้นที่ภายใต้เงื่อนไขแบบเชื่อมต่อ (Merging Local Clusters under Must-Link Constraints) จะนำกลุ่มข้อมูลที่ได้จาก 2.4.1 มารวมกันหากมีเงื่อนไขแบบเชื่อมต่ออยู่ และนิยามว่า แกนกลุ่มข้อมูลในพื้นที่ (Core Local Cluster)

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

2.3.3 การรวมกลุ่มข้อมูลภายใต้เงื่อนไขแบบไม่เชื่อมต่อ (Merging Cluster under Cannot-Link Constraints) จะนำแกนกลุ่มข้อมูลในพื้นที่ที่ได้จาก 2.4.2 มารวมกลุ่มกับข้อมูลที่เหลือ โดยพิจารณาจาก เงื่อนไขแบบไม่เชื่อมต่อ ที่ยังไม่ได้ใช้ จากขั้นตอน 2.4.1

เทคนิคนี้จะพิจารณาทุกๆเงื่อนไขที่ถูกป้อนในระบบ ซึ่งอาจทำให้เกิดกลุ่มข้อมูลแบบโดดเดี่ยว (Singleton Local Cluster) เป็นจำนวนมากได้

3. การผสมผสานเงื่อนไขในการแบ่งกลุ่มกระแสข้อมูล

กระแสข้อมูลมีลักษณะเฉพาะที่ข้อมูลเกิดขึ้นได้ตลอดและต่อเนื่อง โดย เงื่อนไขที่จะนำมาใช้ต้องสอดคล้องและรองรับต่อลักษณะเฉพาะของกระแสข้อมูลด้วยเช่นกัน จึงทำให้งานวิจัยในเรื่องนี้ยังมีจำนวนไม่มาก อย่างไรก็ตาม มีการศึกษาและวิจัยในการผสมผสานเงื่อนไขในการแบ่งกลุ่มกระแสข้อมูลดังรายละเอียดต่อไปนี้

3.1 User Constraint over Data Stream

ในงานวิจัยนี้ (Ruiz et al., 2006) ได้นำเสนอ วิธีการผสมผสานเงื่อนไขที่ใช้ในการแบ่งกลุ่มข้อมูล มาใช้ในการแบ่งกลุ่มกระแสข้อมูล ได้แก่ เงื่อนไขแบบเชื่อมต่อ และ ไม่เชื่อมต่อ และยังได้เพิ่มเงื่อนไขในระดับ กลุ่มข้อมูล ได้แก่ เงื่อนไขในการกำหนดจำนวนสมาชิกขั้นต่ำในแต่ละกลุ่มข้อมูล อย่างไรก็ตาม ในงานวิจัยนี้ ไม่มีการทดลองและผลการทดลองบนชุดข้อมูลใดๆ เป็นเพียงแค่แนวคิดเท่านั้น

3.2 C-DenStream

C-DenStream (Ruiz et al., 2009) เป็นงานวิจัยต่อยอดจากงานของ (Ruiz et al., 2006) โดยได้นำแนวคิดที่ได้ถูกนำเสนอจากงานวิจัยที่ผ่านมา นำมาใช้ใน เทคนิคการแบ่งกลุ่มกระแสข้อมูล DenStream (Cao et al., 2006)

เทคนิคการแบ่งกลุ่มกระแสข้อมูลแบบอ้างอิงความหนาแน่น DenStream นั้นใช้แนวความคิดของกลุ่มข้อมูลแบบจุลภาค (Micro-cluster) ใช้ในการเก็บตัวแทนของข้อมูลแทนการ

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

เก็บจุดข้อมูลทุกๆจุด จากเทคนิค DBSCAN (Ester et al., 1996) โดยกลุ่มข้อมูลแบบจุลภาคนั้นจะเก็บค่าตัวแปรต่างๆ ได้แก่ w ซึ่งเป็นค่าน้ำหนักของกลุ่มข้อมูลแบบจุลภาค c เป็นค่าศูนย์กลางของกลุ่มข้อมูลแบบจุลภาค และ r เป็นรัศมีของกลุ่มข้อมูลแบบจุลภาค

C-DenStream นั้นได้ทำการเพิ่มเทคนิคเพื่อให้รองรับกับกลุ่มข้อมูลแบบจุลภาคของ DenStream ในเชิงของการใช้หน่วยความจำประมวลผล จึงได้นำเสนอ เงื่อนไขแบบกลุ่มข้อมูลจุลภาค (Micro-cluster Constraints) และในเงื่อนไขแบบกลุ่มข้อมูลจุลภาคจะประกอบด้วย ประเภท (Type) ซึ่งคือประเภทของเงื่อนไขได้แก่ เงื่อนไขแบบเชื่อมต่อ และไม่เชื่อมต่อ น้ำหนัก (Weight) ซึ่งคือน้ำหนักของเงื่อนไข

1. Setting up micro-clusters and micro constraints.
Using *InitPoints* and constraints *CO* from *DS*:
INICIALIATE(InitPoints, CO).

2. Online step: Maintaining micro-clusters and micro constraints.
$T_p = \left\lceil \frac{1}{\lambda} log(\frac{\beta\ \mu}{\beta\ \mu - 1}) \right\rceil$.

Adding new points:
$\forall p \in DS$: *MERGE(p, p-buffer, o-buffer).*

Adding new constraints:
$\forall co \in DS$: *MERGE-CONSTRAINTS(co, CO).*

Updating micro-clusters and micro constraints:
if $(t\ mod\ T_p = 0)$ **then**
  UPDATING MICRO-CLUSTERS (p-buffer, o-buffer).
  UPDATING MICRO CONSTRAINTS (CO-MC).
end

3. Offline step: generating final clusters.
if *user request* **then**
  C-DBSCAN (*p-buffer, CO-MC*).
end

ภาพที่ 10 แสดงอัลกอริทึม C-DenStream

จากภาพที่ 10 เริ่มต้น C-DenStream จะทำการสร้างกลุ่มข้อมูลจากเงื่อนไขทั้งหมด ในขั้นตอนที่ 1 หลังจากนั้น จะแบ่งออกเป็นสองส่วน คือ ส่วนของ ออนไลน์ (Online Step) จะทำการรักษา กลุ่มข้อมูลจุลภาค และ เงื่อนไขแบบจุลภาค และอีกส่วนคือ ออฟไลน์ (Offline Step) เพื่อสร้างกลุ่มข้อมูลสุดท้าย ในกรณีที่ผู้ใช้มีการเรียกดู

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

ในส่วนของ ออนไลน์นั้นจะมีการรวมกลุ่มข้อมูล และ รวมกลุ่มเงื่อนไข ดังเช่นในฟังก์ชั่น MERGE (p, p-buffer, o-buffer) และ MERGE-CONSRAINTS (co, CO) ตามลำดับ ในภาพที่ 10 โดยถูกใช้ต่อเมื่อเวลาในระบบมีค่าเท่ากับ T_p ซึ่งผู้ใช้เป็นคนกำหนด เพื่อรองรับต่อการเปลี่ยนของกระแสข้อมูล และยังคงมีการปรับกลุ่มเงื่อนไขหากว่าค่าน้ำหนักของเงื่อนไขลดลงจนถึงค่าที่กำหนด จะถูกถอดออกจากกลุ่มเงื่อนไข และหากว่า กลุ่มข้อมูลใดมีเงื่อนไขแบบไม่เชื่อมต่ออยู่ภายใน ระบบจะทำการแบ่งแยกกลุ่มข้อมูลออกเป็นสองกลุ่ม

งานวิจัยนี้ยังมีข้อเสียในเชิงของปริมาณของกลุ่มข้อมูลขนาดเล็กอาจมีปริมาณที่เยอะเกินไป เนื่องจาก ระบบจะต้องพิจารณาทุกๆเงื่อนไข

3.3 SemiStream Clustering

งานวิจัยนี้ (Halkaidi et al., 2012) สามารถใช้เงื่อนไขระดับข้อมูล ซึ่งได้แก่ เงื่อนไขแบบเชื่อมต่อและไม่เชื่อมต่อได้อย่างครบถ้วน และยังคงรวมไปถึง เงื่อนไขในระดับกลุ่มข้อมูล ซึ่งคือ เงื่อนไขในการกำหนดความกระชับของข้อมูลในแต่ละกลุ่มข้อมูล (Compactness Constraints) และ เงื่อนไขในการกำหนดระยะห่างระหว่างกลุ่มข้อมูล (Minimum Separation Constraints) เงื่อนไขทั้งหมดนี้ได้มีการออกแบบเงื่อนไขให้เหมาะสมกับกระแสข้อมูล โดยมีนิยามดังต่อไปนี้

นิวเคลียสของกลุ่มข้อมูล s (s-cluster Nucleus) เป็นจุดตัวแทนของกลุ่มข้อมูล หรือ จุดกึ่งกลางในอุดมคติของกลุ่มข้อมูล ซึ่งจะประกอบไปด้วยค่ากลาง (Means) ของทุกๆมิติ

ขอบของกลุ่มข้อมูล s (s-cluster Rim) เป็นช่วงขอบซึ่งประกอบไปด้วยข้อมูลที่ระยะทางอยู่นอกรัศมีของนิวเคลียสแต่ไม่เกินค่าระยะห่างของผลคูณของ T กับค่ารัศมี ซึ่ง T กำหนดโดยผู้ใช้ หากเกินกว่าขอบนี้ถือว่า เป็น ข้อมูลผิดปกติ (Outliers)

การเลื่อมล้ำของกลุ่มข้อมูล s (Overlap between s-clusters) คือค่าประเมินในการรวมกลุ่มข้อมูล s สองกลุ่มเข้าด้วยกัน ก็ต่อเมื่อ ค่าเลื่อมล้ำของกลุ่มใดกลุ่มหนึ่งมีค่ามากกว่าค่า $T_{overlap}$ ซึ่งผู้ใช้กำหนด เมื่อกลุ่มข้อมูล s รวมกันจะถูกเรียกว่า m-cluster

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

อย่างไรก็ตาม งานวิจัยนี้ จะพิจารณาการแบ่งกลุ่มข้อมูลในช่วงถัดไปก็ต่อเมื่อ การแบ่งกลุ่มข้อมูลในช่วงปัจจุบันต้องมีค่าวัดผลประสิทธิภาพที่ดีกว่าการแบ่งกลุ่มข้อมูลก่อนหน้า หากไม่ดีกว่า ระบบจะทำการแบ่งกลุ่มข้อมูลใหม่จนกว่า ค่าวัดผลประสิทธิภาพดีกว่าเดิม ในกรณีนี้อาจทำให้เกิดความล่าช้าได้

จากการอ่านบทความและงานวิจัยที่เกี่ยวทำให้ทราบว่า การผสมผสานเงื่อนไขในการแบ่งกลุ่มกระแสข้อมูลนั้นมีแนวโน้มที่สามารถเป็นไปได้ และ ช่วยเพิ่มประสิทธิภาพกลุ่มข้อมูลผลลัพธ์อีกด้วย ดังนั้นผู้เขียนจึงได้คิดนำวิธีการต่างๆ เช่น การออกแบบเงื่อนไขสำหรับการแบ่งกลุ่มกระแสข้อมูล และ การนำเงื่อนไขไปใช้ในขั้นตอนต่างๆ มาปรับปรุงและประยุกต์ใช้ในงานวิจัยฉบับนี้

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

อุปกรณ์และวิธีการ

อุปกรณ์

ฮาร์ดแวร์

1. เครื่องคอมพิวเตอร์โน๊ตบุ้ค 1 เครื่อง ประกอบด้วยอุปกรณ์ดังต่อไปนี้
   1.1 ซีพียู (CPU) Intel Core i5 ความเร็ว 2.6 GHz
   1.2 หน่วยความจำหลัก 8.0 GB
   1.3 ฮาร์ดดิสก์ขนาด 160 GB

ซอฟต์แวร์

1. ระบบปฏิบัติการ Windows 7 Professional
2. โปรแกรม Weka version 3.6
3. โปรแกรม Microsoft Visual C++ 2010 Professional Edition

วิธีการ

จากการตรวจเอกสารข้างต้นจะเห็นว่างานวิจัยสำหรับการใช้เงื่อนไขในการแบ่งกลุ่มข้อมูลนั้นมีมากมายหลากหลาย แต่ยังมีงานวิจัยไม่มากนักที่ได้พิจารณาถึงการใช้เงื่อนไขในการแบ่งกลุ่มกระแสข้อมูล ซึ่งการใช้เงื่อนไขจะทำให้คุณภาพของกลุ่มข้อมูลที่ได้สูงขึ้น ดังนั้นวิทยานิพนธ์ฉบับนี้จึงต้องการนำเสนอ เทคนิคการแบ่งกลุ่มกระข้อมูลโดยใช้เงื่อนไข CE-Stream ซึ่งเป็นการปรับเพิ่มจาก เทคนิคการแบ่งกลุ่มกระแสข้อมูล E-Stream ในขั้นตอนต่างๆ และมีการปรับปรุงโครงสร้างของเงื่อนไขให้รองรับกับการทำงานของกระแสข้อมูล E-Stream

1. แนวคิดการทำงาน

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

สำหรับกระแสข้อมูลซึ่งมีพฤติกรรมที่แตกต่างไปจากข้อมูลแบบคงที่ได้แก่ การแยกตัวของกลุ่มข้อมูล และการส่งข้อมูลสมาชิกให้กลุ่มข้อมูล และ อื่นๆ งานวิจัยนี้จึงได้นำเงื่อนไขมาใช้เพื่อช่วยเหลือและแนะนำการเปลี่ยนแปลงของกลุ่มข้อมูลให้ถูกต้องมากขึ้นตามความเป็นจริงซึ่งได้จากผู้เชี่ยวชาญหรือความรู้ในอดีต โดยเงื่อนไขจะถูกนำไปใช้ในการแบ่งกลุ่มกระแสข้อมูล E-Stream ในขั้นตอนดังต่อไปนี้

1.1 การแยกตัวของกลุ่มข้อมูล (Cluster Splitting) กลุ่มข้อมูลสามารถแยกตัวออกจากกันได้ ถ้าพิจารณาข้อมูลภายในกลุ่มแล้วพบว่าเกิดเป็นกลุ่มย่อยภายในที่แตกแยกกันอย่างชัดเจน การแยกตัวอาจเกิดได้จากการรวมกลุ่มที่ผิดพลาดในขณะที่ข้อมูลมีจำนวนน้อย หรือการแยกตัวโดยพฤติกรรมของข้อมูลเอง ในขั้นตอนนี้ เงื่อนไขแบบเชื่อมต่อจะถูกนำมาพิจารณา เพื่อป้องกันการแยกตัวที่ไม่จำเป็น

1.2 การส่งมอบข้อมูลสมาชิกให้แต่ละกลุ่มข้อมูล (Cluster Assignment) ข้อมูลจะถูกส่งมอบให้กลุ่มข้อมูลที่อยู่ใกล้ที่สุด แต่หากข้อมูลที่จะถูกส่งมีระยะห่างเกินกว่าค่าที่กำหนดไว้จะถูกปล่อยทิ้งไว้เป็นข้อมูลโดดเดี่ยวเพื่อให้ระบบพิจารณาต่อไป ในขั้นตอนนี้ เงื่อนไขแบบไม่เชื่อมต่อจะถูกนำมาพิจารณา เพื่อป้องกันการส่งมอบข้อมูลสมาชิกไปสู่กลุ่มข้อมูลที่ผิด

2. คำจำกัดความต่างๆ

2.1 ข้อมูลที่เก็บไว้ในระบบ คือ ข้อมูลทั้งหมดที่ระบบเก็บไว้ในหน่วยความจำเพื่อใช้ประมวลผล เราจะแบ่งข้อมูลที่เก็บไว้ในระบบออกเป็น 3 ส่วน คือ ข้อมูลโดดเดี่ยว (Isolate Data Point) กลุ่มข้อมูลที่ไม่ใช้งาน (Inactive Cluster) และกลุ่มข้อมูลใช้งาน (Active Cluster)

2.2 ข้อมูลโดดเดี่ยว (Isolate Data Point) คือ ข้อมูลหนึ่งตัวที่ไม่ถูกจัดเข้าอยู่ในกลุ่มข้อมูลใด ถูกเหลือทิ้งไว้ในระบบเพื่อใช้พิจารณาต่อไป อาจเกาะกลุ่มกันและเกิดเป็นกลุ่มใหม่ หรือถูกลดค่าความสำคัญจนหายไปในที่สุด

2.3 กลุ่มข้อมูลไม่ใช้งาน (Inactive Cluster) คือ ข้อมูลโดดเดี่ยวหรือกลุ่มของข้อมูลโดดเดี่ยวที่เริ่มมีการจับกลุ่มกัน แต่ยังไม่สามารถตั้งตัวเป็นกลุ่มข้อมูลใช้งานได้ เนื่องจากค่าความสำคัญไม่เพียงพอ

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

2.4 กลุ่มข้อมูลใช้งาน (Active Cluster) คือ กลุ่มข้อมูลที่พิจารณาเป็นโมเดลของระบบ สามารถรับข้อมูลใหม่เป็นสมาชิกได้ถ้ามีความคล้ายคลึงกับข้อมูลภายในกลุ่ม

2.5 กลุ่มข้อมูล (Cluster) คือ ข้อมูลโดดเดี่ยวกลุ่มข้อมูลไม่ใช้งาน หรือ กลุ่มข้อมูลใช้งาน ที่มีค่าเก็บไว้ในระบบ ในความเป็นจริงข้อมูลทั้ง 3 แบบถูกเก็บไว้ในรูปแบบตัวแทนกลุ่มข้อมูล (Cluster Representation) เหมือนกัน ดังนั้น เมื่อกล่าวถึงคำว่า กลุ่มข้อมูล จะไม่ได้เจาะจงว่าเป็นข้อมูลโดดเดี่ยว กลุ่มข้อมูลไม่ใช้งาน หรือ กลุ่มข้อมูลใช้งาน

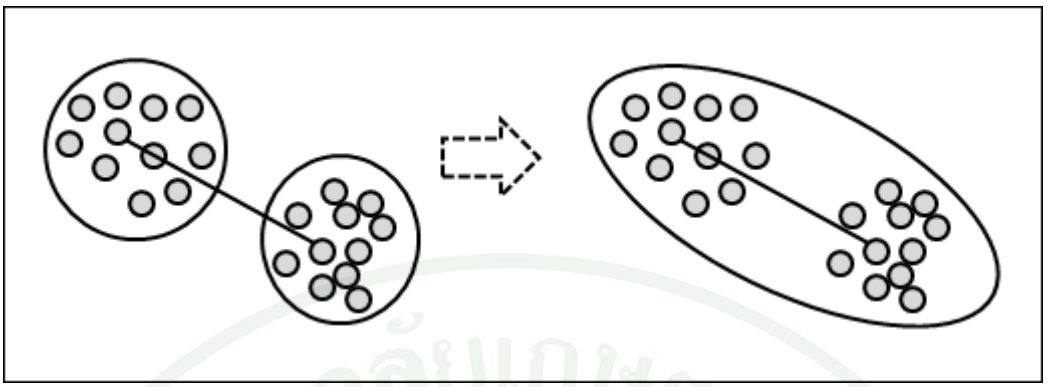
2.6 การเลือนหายของข้อมูล (Fading data point) คือ การลดค่าความสำคัญของข้อมูลเมื่อเวลาผ่านไป เนื่องจากกระแสข้อมูล รูปแบบของกลุ่มข้อมูลสามารถมีการเปลี่ยนแปลงได้ตลอดเวลา เราจึงควรปรับเปลี่ยนความสำคัญให้สอดคล้องกับข้อมูลเก่า โดยยังคงเก็บข้อมูลเดิมแต่ลดความสำคัญของข้อมูลเก่าลงเรื่อยๆตามเวลาเพื่อให้ระบบสามารถปรับตัวเข้ากับการเปลี่ยนแปลงได้ เราใช้ฟังก์ชันการเลือนหายเพื่อลดค่าความสำคัญของข้อมูลให้ λ คืออัตราการเลือนหาย และ t คือเวลาที่ผ่านไปตั้งแต่ข้อมูลปรากฏในระบบ ฟังก์ชันการเลือนหายมีสมการดังนี้

$$f(t) = 2^{-\lambda t} \qquad \textbf{(6)}$$

2.7 ค่าความสำคัญของกลุ่มข้อมูล (Cluster Weight) คือ จำนวนสมาชิกเสมือนของกลุ่มข้อมูลในขณะนั้น เหตุที่เรียกว่าเสมือนเนื่องจากจำนวนข้อมูลจะเสมือนว่ามีการลดลงตามเวลาเนื่องจากมีการเลือนหาย เมื่อเริ่มต้นข้อมูลแต่ละตัวมีค่าความสำคัญเป็น 1 แต่เมื่อเวลาผ่านไปจะมีค่าลดลงเรื่อยๆ กลุ่มข้อมูลใดๆ จะมีความสำคัญเพิ่มขึ้นได้ก็ต่อเมื่อกลุ่มข้อมูลนั้นรับสมาชิกเพิ่ม ซึ่งเป็นไปได้สองแบบ คือ รวมกับกลุ่มข้อมูลอื่น หรือรับข้อมูลเข้าใหม่เป็นสมาชิก

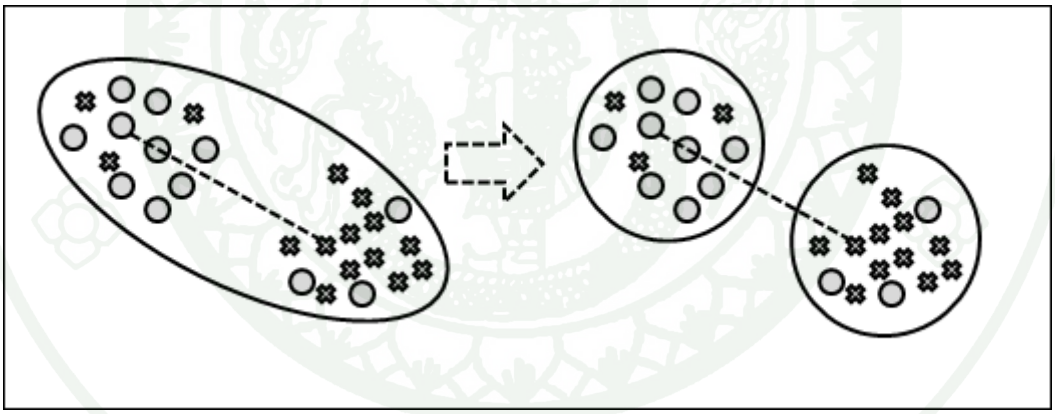
2.8 เงื่อนไขแบบเชื่อมต่อ (Must-Link Constraints) คือ คู่ของจุดข้อมูล ที่ควรอยู่ในกลุ่มข้อมูลเดียวกัน นั้นหมายความว่า หากมีเงื่อนไขแบบเชื่อมต่ออยู่ในคู่ของกลุ่มข้อมูลใดๆ กลุ่มข้อมูลนั้นควรถูกรวมกันให้เหลือกลุ่มข้อมูลเดียวดังภาพที่ 11

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์



ภาพที่ 11 แสดงลักษณะของเงื่อนไขแบบเชื่อมต่อ

2.9 เงื่อนไขแบบไม่เชื่อมต่อ (Cannot-Link Constraints) คือ คู่ของจุดข้อมูล ที่ไม่ควรอยู่ในกลุ่มข้อมูลเดียวกัน นั้นหมายความว่า หากมีเงื่อนไขแบบไม่เชื่อมต่ออยู่ในกลุ่มข้อมูลใดๆ กลุ่มข้อมูลนั้นๆ ควรแยกออกจากกัน ดังภาพที่ 12



ภาพที่ 12 แสดงลักษณะของเงื่อนไขแบบไม่เชื่อมต่อ

2.10 เงื่อนไขที่ใช้งาน (Active Constraints) คือ เงื่อนไขที่จะถูกใช้และพิจารณาในระบบ เนื่องจาก จุดข้อมูลทั้งสองในเงื่อนไขนั้นๆ มีความเหมือนกับข้อมูลที่ผ่านเข้ามาและมีค่าความสำคัญมากกว่าค่า W ที่ผู้ใช้กำหนดขึ้น

2.11 ค่าความสำคัญของเงื่อนไข (Constraints Weight) คือ ค่าน้ำหนักที่น้อยที่สุดของสองจุดข้อมูลในแต่ละเงื่อนไข หมายความว่า หากในเงื่อนไขนั้นๆ มีจุดข้อมูลใดข้อมูลหนึ่งที่มีค่า

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

น้ำหนักของจุดข้อมูลน้อยกว่า ค่า w ที่ผู้ใช้กำหนดไว้ เงื่อนไขนั้นจะไม่นำมาพิจารณาในระบบ ซึ่งมีสมการดังนี้

$$CTweight = \min(weight(CT_x), weight(CT_y)) \quad (7)$$

2.12 การเลือนหายของเงื่อนไข (Constraints Fading) คือ การลดค่าความสำคัญของเงื่อนไขที่ใช้งานเมื่อเวลาผ่านไป เนื่องจากค่าความสำคัญของเงื่อนไขนั้นยาวนานกว่า เงื่อนไขจึงไม่สามารถเลือนหายได้เร็วเท่ากับข้อมูล ให้ r เป็นช่วงระยะเวลา และ t เวลาที่ผ่านไปตั้งแต่จุดข้อมูลในเงื่อนไขปรากฏในระบบ เป็นฟังก์ชั่นการเลือนหายของเงื่อนไขมีสมการดังนี้

$$fad(t) = 2^{-\lambda \frac{t}{r}} \quad \textbf{(8)}$$

2.13 อัตราการใช้เงื่อนไข (Constraints Hit-Rate) คือ จำนวนครั้งที่เงื่อนไขแบบเชื่อมต่อถูกใช้ใน การแยกตัวของกลุ่มข้อมูล (Cluster Splitting) เพื่อที่จะเพิ่มประสิทธิภาพในการใช้เงื่อนไขในการแบ่งกลุ่มกระแสข้อมูล เงื่อนไขจะถูกเรียงลำดับโดยอัตราการใช้เงื่อนไข

3. การจัดการกับฮิสโทแกรม

การเก็บข้อมูลในระบบทั้ง ข้อมูลโดดเดี่ยว กลุ่มข้อมูลไม่ใช้งาน และกลุ่มข้อมูลใช้งาน หรือ กลุ่มข้อมูลผลลัพธ์ ทั้งหมดจะเก็บในรูปแบบของตัวแทนกลุ่มข้อมูลที่เรียกว่า Fading Cluster Structure with Histogram (FCH) (Udommanetanakit et al., 2007)

กระแสข้อมูลคือ ชุดของข้อมูลจำนวนอนันต์ $\overline{X_1}...\overline{X_k}...$ ซึ่งเกิดขึ้นที่เวลา $T_1...T_k...$ ตามลำดับ โดยที่จุดข้อมูลแต่ละตัวประกอบด้วย d มิติ เขียนแทนได้ด้วยสมการ $\overline{X_i} = (x_i^1...x_i^d)$

ให้ t เป็นเวลาปัจจุบัน $C=\{\overline{X_1}...\overline{X_N}\}$ เป็นกลุ่มข้อมูลขนาด d มิติภายในกลุ่มข้อมูล ที่เกิดขึ้นที่เวลา $T_1...T_N$ ตามลำดับ จะได้ตัวแทนกลุ่มข้อมูลที่เกิดจาก C และ t

$$FCH(C,t)=(\overline{FC1(C,t)},\overline{FC2(C,t)},W(t),\overline{H(C,t)},\overline{TF(C,t)}) \quad \textbf{(9)}$$

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

โดยที่

$$\overline{FC1(C,t)}=(FC1^1...FC1^d) \quad \textbf{(10)}$$

และ

$$\overline{FC2(C,t)}=(FC2^1...FC2^d) \quad \textbf{(11)}$$

$FC1^j(C,t)$ เป็นผลรวมถ่วงน้ำหนักด้วยค่าการเลือนหาย ของข้อมูลทั้งหมดในกลุ่มข้อมูล ณ มิติที่ j ดังสมการ

$$FC1^j(C,t)=\sum_{i=1}^{N} f(t-T_i).(x_i^j) \quad \textbf{(11)}$$

$FC2^j(C,t)$ เป็นผลรวมถ่วงน้ำหนักด้วยค่าการเลือนหายของข้อมูลยกกำลังสอง ของข้อมูลทั้งหมดในกลุ่มข้อมูล ณ มิติที่ j ดังสมการ

$$FC2^j(C,t)=\sum_{i=1}^{N} f(t-T_i).(x_i^j)^2 \quad \textbf{(12)}$$

W(t) คือค่าความสำคัญ เป็นผลรวมของค่าการเลือนหายของข้อมูลทั้งหมดในกลุ่มข้อมูล ดังสมการ

$$W(t)=\sum_{i=1}^{N} f(t-T_i) \quad \textbf{(13)}$$

f(t) คือสมการฟังก์ชันการเลือนหาย ให้ λ คืออัตราการเลือนหาย และ t คือเวลาที่เปลี่ยนไป ดังสมการที่ 6

$H^j(C,t)$ เป็นฮิสโตแกรมของข้อมูลทั้งหมดในกลุ่มข้อมูล ณ มิติที่ j โดยกำหนดให้มีจำนวน α ชั้น (bin) ดังสมการ

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

$$H^j(C,t) = (h^{j1}(C,t)...h^{j\alpha}(C,t)) \quad \textbf{(15)}$$

$H_l^j(t)$ เป็นฮิสโตแกรมของข้อมูลทั้งหมดในกลุ่มข้อมูล ณ มิติที่ j ช่วงชั้นที่ l ซึ่งเกิดขึ้นที่เวลา t ดังสมการ

$$H_l^j(t) = \sum_{i=l}^{N} f(t - T_i).(x_i^j).(y_{il}^j) \quad \textbf{(16)}$$

เมื่อ y_{il} เป็นค่าน้ำหนักของข้อมูล x_i ในฮิสโตแกรม ช่วงชั้นที่ l

$$y_{il} = \begin{cases} 1 & if \quad l.b + left \le x_i \le (l+1).b + left \\ 0 & otherwise \end{cases} \quad \textbf{(17)}$$

left เป็นค่าต่ำสุดของข้อมูลภายในกลุ่ม

$$left = \min(x_i^j) \quad \textbf{(18)}$$

right เป็นค่าสูงสุดของข้อมูลภายในกลุ่ม

$$right = \max(x_i^j) \quad \textbf{(19)}$$

b เป็นความกว้างของชั้นหรืออันตรภาคชั้น

$$b = \frac{left + right}{\alpha} \quad \textbf{(20)}$$

โดยสรุปคือ ใน FCH หนึ่งจะประกอบไปด้วย FC1, FC2 และฮิสโตแกรมจำนวน d มิติ ซึ่งในแต่ละฮิสโตแกรมจะมีจำนวน α ชั้น

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

4. อัลกอริทึม CE-Stream

เพื่อที่จะนำเงื่อนไขมาประยุกต์ใช้ใน CE-Stream งานวิจัยนี้จึงได้เพิ่มและปรับปรุงขั้นตอนต่างๆจากเทคนิค E-Stream ได้แก่ CheckActiveAndUpdateConstraints, FadingConstraints, CheckSplit และ ClusterAssignment โดยจะใช้โค้ดจำลอง เป็นตัวอธิบายการทำงานทั้งขั้นตอนของ E-Stream และขั้นตอนของ CE-Stream ที่ได้เพิ่มขึ้นมา ดังต่อไปนี้

Algorithm CE-Stream

1 Retrieve new data X_i
2 CheckActiveAndUpdateConstraints
3 FadingAllClusters
4 PrioritizeConstraints
5 CheckSplit
6 MergeOverlapCluster
7 FadingConstraints
8 FlagActiveCluster
9 (minDistance, index) ← FindClosestCluster
10 if minDistance < radius_factor
11 ClusterAssignment(X_i,FCH_{index})
12 else
13 create new FCH from X_i
13 waiting for new data

ภาพที่ 13 แสดงอัลกอริทึม CE-Stream

4.1 CheckActive&UpdateConstraints

การนำเงื่อนไขมาใช้ใน E-Stream เงื่อนไขจะถูกพิจารณาเมื่อมีข้อมูลที่มีลักษณะเหมือนกับเงื่อนไขในทุกๆมิติ ดังรูปภาพที่ 14 (บรรทัดที่ 2) หากมีเงื่อนไขที่ตรงกับข้อมูลที่เข้ามาทุกมิติ จุดข้อมูลในเงื่อนไขนั้นจะถูกกระตุ้นให้เริ่มใช้งาน เงื่อนไขสามารถใช้งานได้ (Active

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

Constraints) ก็ต่อเมื่อจุดข้อมูลทั้งสองในเงื่อนไขนั้นๆ ได้ถูกพิจารณาให้ใช้งานพร้อมกัน (บรรทัดที่ 4 และ 7)ในกรณีที่ เงื่อนไขนั้นได้เกิดขึ้นมาแล้ว หากยังคงมีข้อมูลใหม่ที่เข้ามามีลักษณะที่ตรงกันอีก เงื่อนไขนั้นจะถูกทำการเพิ่มค่าความสำคัญของเงื่อนไข (บรรทัดที่ 3 และ 6) อนึ่ง หากเงื่อนไขนั้นมีค่าน้ำหนักลดลงจนต่ำกว่าค่าความสำคัญที่ผู้ใช้ได้กำหนดเอาไว้ เงื่อนไขนั้นจะถูกตัดทิ้งออกจากชุดเงื่อนไขที่จะนำไปพิจารณาในระบบ (บรรทัดที่ 10 และ 11) เพื่อให้สอดคล้องกับลักษณะเฉพาะของกระแสข้อมูล โดยที่ อัลกอริทึมนี้สามารถแบ่งได้เป็น สอง ช่วง นั้นคือ ช่วงการตรวจสอบการใช้งานของเงื่อนไข (Check Active Constraints) และ ช่วงปรับเปลี่ยนชุดเงื่อนไข (Update Constraints) โดย กำหนดให้ CT คือชุดของเงื่อนไขทั้งหมดที่ถูกป้อนในระบบ และ ct คือเงื่อนไขหนึ่ง ซึ่งประกอบด้วย จุดข้อมูล ct_x และ ct_y

Algorithm CheckActiveAndUpdateConstraints

1 **for** *ct* **in** *CT*
2 **if** ($X_i = ct_x$) ct_x.status = 1
3 **if** (ct_x exists in *CurrentCT*) setIncreaseWeight(ct_x)
4 **Else** setInitialWeight(ct_x) ; $CurrentCT_t$ add {ct_x}
5 **if** ($X_t = ct_y$) ct_y.status = 1
6 **if** (ct_y exists in CurrentCT) setIncreaseWeight(ct_y)
7 **Else** setInitialWeight(ct_y) ; $CurrentCT_t$ add {ct_y}
8 **endfor**
9 **for** *ct* **in** $CurrentCT_t$
10 **if**(ct_x.status = 1) AND (ct_x.weight < *W*) $CurrentCT_t = CurrentCT_t$.Remove{ ct_x }
11 **if**(ct_y.status = 1) AND (ct_y.weight < *W*) $CurrentCT_t = CurrentCT_t$ .Remove{ ct_y }
12 **endfor**

ภาพที่ 14 แสดงอัลกอริทึม CheckActiveAndUpdateConstraints

4.2 FadingAllClusters

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

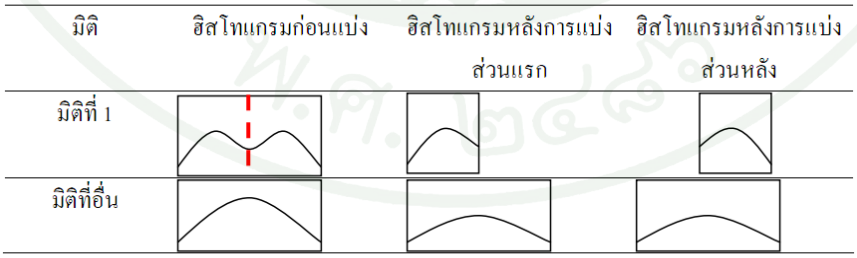
อัลกอริทึมการเลือนค่าความสำคัญของกลุ่มข้อมูลในระบบ เพื่อให้ความสำคัญกับข้อมูลใหม่มากกว่าข้อมูลเดิมในระบบที่ไม่มีการเปลี่ยนแปลงพฤติกรรม โดยจะลดค่าความสำคัญของข้อมูลที่คงอยู่ในระบบ ตามสมการที่ 6

4.3 Prioritize Constraints

อัลกอริทึมนี้ จะทำการจัดลำดับเงื่อนไขที่ใช้งานทั้งหมดโดยใช้ อัตราการใช้เงื่อนไข ที่ได้กล่าวไว้ข้างต้น เพื่อให้ รองรับกับ การเลือนหายและค่าความสำคัญของเงื่อนไข อัลกอริทึมนี้จะช่วยให้ ระบบสามารถทำงานได้เร็วขึ้น และมีประสิทธิภาพ โดยการจัดลำดับ มีสองประเภท นั้นคือ การจัดลำดับแบบขึ้น (Ascending) และการจัดลำดับแบบลง (Descending) การจัดลำดับแบบขึ้นเป็นการนำเงื่อนไขที่ไม่เคยถูกใช้นำมาพิจารณาก่อน และ การจัดลำดับแบบลงเป็นการนำเงื่อนไขที่ถูกใช้มากที่สุดมาพิจารณาก่อน

4.4 CheckSplit

อัลกอริทึมนี้จะทำการแบ่งแยกกลุ่มข้อมูลในทุกๆมิติที่มีการแตกตัวกันของฮิสโตแกรม โดยพิจารณาจากฮิสโตแกรมที่มีลักษณะแบบระฆังคู่ (Double Hump Type) โดยจุดแบ่งของข้อมูล คือ ช่วงที่มีข้อมูลหนาแน่นน้อยระหว่าง จุดที่มีข้อมูลหนาแน่นสูง สองจุดในมิติใดมิติหนึ่งของกลุ่มข้อมูล โดยที่จุดแบ่งนั้น จะต้องมีความแตกต่างจากจุดหนาแน่นสูงอย่างมีนัยสำคัญ ดังภาพที่ 15

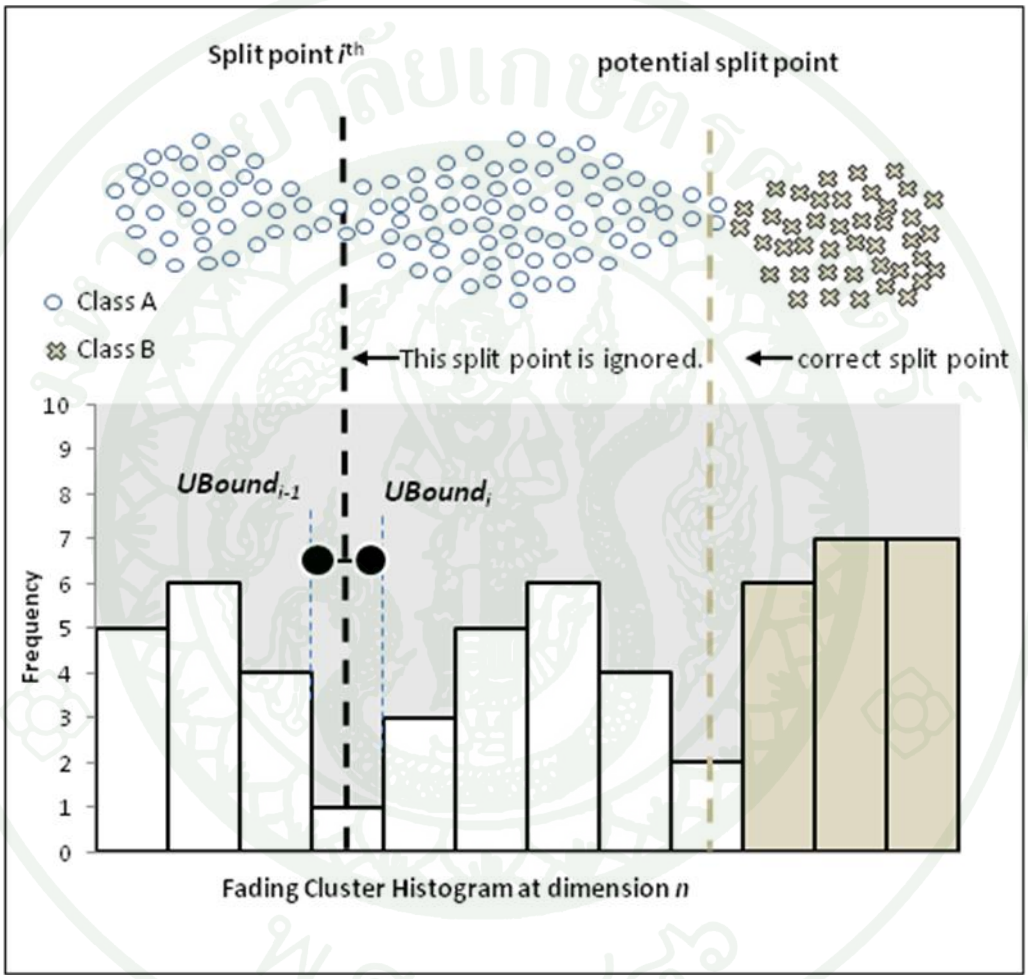


ภาพที่ 15 ลักษณะการแบ่งฮิสโตแกรมในมิติต่างๆ

อย่างไรก็ตาม ในขั้นตอนการแบ่งแยกนี้ จะมีการนำเงื่อนไขแบบเชื่อมต่อมาใช้เพื่อช่วยให้ การแบ่งแยกกลุ่มข้อมูลนั้นมีความถูกต้องมากขึ้น ตัวอย่างเช่น ในบางกรณีที่ฮิสโตแกรมมี

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

ลักษณะ แบบฟันปลา (Plateau) ซึ่งมีความถี่ที่สลับกัน อาจทำให้การแบ่งแยกกลุ่มข้อมูล มีการคลาดเคลื่อนและทำให้ความแม่นยำลดน้อยลง โดยที่ เงื่อนไขแบบเชื่อมต่อ จะช่วยตรวจสอบการแบ่งแยกกลุ่มข้อมูลที่ไม่ถูกต้อง และนำไปสู่การแบ่งแยกกลุ่มข้อมูลที่ถูกต้องและแม่นยำมากขึ้น ดังภาพที่ 16



ภาพที่ 16 แสดงการแบ่งกลุ่มข้อมูลโดยมีเงื่อนไขช่วยเหลือ

จากภาพ จะเห็นว่า มีจุดที่มีความหนาแน่นต่ำอยู่สองจุดซึ่ง หากแบ่งกลุ่มข้อมูลโดยใช้ข้อมูลในเชิงสถิติ จุดแรกทางซ้ายจะเป็นจุดที่ถูกแบ่ง เนื่องจากมีความหนาแน่นต่ำกว่าจุดที่สองทางด้านขวา อย่างไรก็ตามในจุดที่ถูกแบ่งนี้ มีเงื่อนไขแบบเชื่อมต่อคร่อมอยู่ หมายความว่า มีเงื่อนไขอยู่ในระหว่างค่าขอบเขตบนของแท่งฮิสโตแกรมที่ถูกแบ่ง (Upper $bound_i$) และ แท่งฮิสโตแกรมที่ถูก

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

แบ่งลำดับก่อนหน้า (Upper $bound_{i-1}$) ซึ่งจะเป็นการบังคับให้ระบบเพิกเฉยจากการแบ่งกลุ่มข้อมูลที่จุดนี้ ซึ่งอาจมีความเป็นไปได้ที่จุดที่สองจะเป็นจุดที่ถูกแบ่งต่อไปเมื่อเวลาผ่านไป

```
Algorithm ClusterSplitting
1  For fch in FCHs
2     if (fch.CheckSplit(S_bin,S_atr))
3     For ct in CurrentCT
4        if( fch[S_atr].Ubound[S_bin-1], fch[S_atr].Ubound[S_bin] cover  [ct_x,ct_y] Then
5           IGNORE_SPLIT; ct.HITRATE++
```

ภาพที่ 17 แสดงอัลกอริทึม ClusterSplitting

เนื่องจากกลุ่มข้อมูลใน E-Stream นั้นถูกเก็บในรูปแบบของฮิสโตแกรม แต่เงื่อนไขถูกเก็บในระดับจุดข้อมูล ดังนั้น การใช้เงื่อนไขในการเปรียบเทียบกับฮิสโตแกรม จึงต้องอาศัยขอบเขตบน (upper bound) ของฮิสโตแกรม ในการพิจารณาเทียบกับจุดข้อมูล ซึ่งงานวิจัยนี้ได้ใช้จุดแบ่ง (S_bin) และ มิติที่ถูกแบ่ง (S_atr) จาก E-Stream มาใช้ ดังภาพที่ 17 หากมีมิติใดมิติหนึ่งที่ค่าของเงื่อนไขอยู่ระหว่าง ค่าขอบเขตบนของจุดที่ถูกแบ่ง ระบบจะทำการเพิกเฉยการแบ่งแยกกลุ่มข้อมูล และ เพิ่มอัตราการใช้เงื่อนไขทันที

4.5 MergeOverlapCluster

เป็นการรวมกลุ่มข้อมูลที่มีพื้นที่ร่วมกันเข้าด้วยกัน โดยการตรวจสอบทุกคู่ของข้อมูลภายในระบบ แบ่งเป็น 2 ประเภท คือ

4.5.1 การรวมกลุ่มของข้อมูลโดดเดี่ยว โดยการตรวจสอบทุกคู่ของข้อมูลโดดเดี่ยว โดยจะรวมสองข้อมูลโดดเดี่ยวเข้าด้วยกัน ต่อเมื่อค่าระยะห่างระหว่างสองข้อมูลโดดเดี่ยวใดๆ น้อยกว่าค่าระยะห่างระหว่างศูนย์กลางของสองกลุ่มข้อมูลใช้งานในระบบที่มีค่าน้อยโดยกำหนดให้ C คือกลุ่มข้อมูล d คือจำนวนมิติ และ center คือจุดศูนย์กลาง ซึ่งหาได้จากสมการ

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

$$dist(C_a, C_b) = \frac{1}{d} \cdot \sum_{j=1}^{d} \left| center_{C_a}^{j} - center_{C_b}^{j} \right| \quad \textbf{(21)}$$

Algorithm Merge_pp

1 for i ← 1 to #isolate
2 for j ← i+1 to #isolate
3 if ($dist_{num}[i,j] < minDistance_{num}(FCH_a, FCH_b)$)
4 Merge($isolate_i$, $isolate_j$)

ภาพที่ 18 แสดงอัลกอริทึม Merge_pp

4.5.2 การรวมกลุ่มของกลุ่มข้อมูล โดยการตรวจสอบทุกคู่ของกลุ่มข้อมูลในระบบทั้งกลุ่มข้อมูลใช้งาน และกลุ่มข้อมูลไม่ใช้งาน โดยจะรวมสองกลุ่มข้อมูลเข้าด้วยกัน ต่อเมื่อค่าระยะห่างระหว่างศูนย์กลางน้อยกว่าค่าเกณฑ์การรวม (merge_threshold)

Algorithm Merge_gg

1 for i ← 1 to |FCH|
2 for j ← i+1 to |FCH|
3 $overlap_{num}[i,j]$ ← dist(FCH_i, FCH_j) - merge_threshold*(FCH_i.sd, FCH_j.sd)
4 if (overlap[i,j] > 0)
5 if (i,j) not in S
6 Merge(FCH_i, FCH_j)

ภาพที่ 19 แสดงอัลกอริทึม Merge_gg

4.6 FadingConstraints

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

เนื่องจากเงื่อนไขต้องมีการรองรับการเลือนหายของกระแสข้อมูลจากลักษณะเฉพาะของกลุ่มกระแสข้อมูลที่อาจมีการเกิดใหม่ หรือ หายไป เมื่อเวลาผ่านไป อัลกอริทึมนี้จึงมีหน้าที่ในการลดค่าความสำคัญของเงื่อนไขโดยอ้างอิงจากสมการที่ 8

4.7 Limit Maximum Cluster

เป็นอัลกอริทึมในการจำกัดจำนวนกลุ่มข้อมูลภายในระบบ โดยเทียบกับค่าเกณฑ์จำนวนกลุ่มมากที่สุด (maximum_cluster) หากจำนวนกลุ่มข้อมูลมากกว่าเกณฑ์ จะดำเนินการวนรอบกลุ่มข้อมูลทั้งหมดในระบบเพื่อหากลุ่มข้อมูลที่ใกล้กันมากที่สุดแล้วนำมารวมกัน โดยจะทำไปเรื่อยๆ จนกว่าจำนวนกลุ่มข้อมูลจะไม่เกินค่าเกณฑ์ที่กำหนดไว้

Algorithm LimitMaximumCluster

1 while |FCH| > maximum_cluster or #isolate > maximum_isolate

2 for i ← 1 to |FCH|

3 for j ← i+1 to |FCH|

4 dist[i,j] ← dist(FCH_i, FCH_j)

5 (first, second) ← $argmin_{(i,j)}$(dist[i, j])

6 Merge(FC_{first}, FC_{second})

ภาพที่ 20 แสดงอัลกอริทึม LimitMaximumCluster

4.8 FlagActiveCluster

อัลกอริทึมเพื่อเปลี่ยนสถานะของกลุ่มข้อมูลไม่ใช้งาน เป็นกลุ่มข้อมูลใช้งาน เมื่อกลุ่มข้อมูลนั้นมีค่าความสำคัญมากกว่าเกณฑ์ที่กำหนด

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

```
Algorithm FlagActiveCluster
1       for i ←1 to |FCH|
2               if FCH_i.W > active_threshold
3                       flag FCH_i as active cluster
4               else
5                       remove active flag from FCH_i
```

ภาพที่ 21 แสดงอัลกอริทึม FlagActiveCluster

4.9 FindClosestCluster

การหาค่าระยะห่างระหว่างข้อมูลใหม่กับกลุ่มข้อมูลใช้งาน นำมาใช้เพื่อหาว่าข้อมูลใหม่ควรเป็นสมาชิกของกลุ่มข้อมูลใช้งานใดภายในระบบ โดยเทคนิค CE-Stream จะจัดลำดับระยะห่างที่ใกล้ที่สุดพิจารณาก่อน หากว่าระหว่างข้อมูลใหม่และกลุ่มข้อมูลที่ใกล้ที่สุด มี เงื่อนไขแบบไม่เชื่อมต่อเกิดอยู่ ระบบจะทำการข้ามไปสู่กลุ่มข้อมูลที่อยู่ใกล้ในลำดับถัดไป ท้ายที่สุดหากไม่สามารถหากลุ่ม ที่ข้อมูลใหม่ได้เนื่องจากมีเงื่อนไขเกิดขึ้นในทุกๆคู่ ข้อมูลใหม่นั้นจะถูกแยกไว้ เพื่อรอรับสมาชิกใหม่ หรือถูกลบออกจากระบบเมื่อมีความสำคัญน้อยกว่าที่กำหนด

```
Algorithm FindClosestCluster
1       for i ←1 to |FCH|
2               dist[i] > dist(FCH_i, x_i)
3       (minDistance, i) ← min(dist[i])
4       return (minDistance, i)
```

ภาพที่ 22 แสดงอัลกอริทึม FindClosestCluster

เนื่องจากเหตุผลเดียวกับ เงื่อนไขแบบเชื่อมต่อ นั่นคือ เงื่อนไขแบบไม่เชื่อมต่อ ไม่สามารถเทียบโดยตรงกับกลุ่มข้อมูลได้ งานวิจัยนี้จึงได้นำค่าที่มากที่สุด (max) และ ค่าที่น้อยที่สุด

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

(min) ของ ฮิสโตแกรมในทุกๆมิติ ของ FCH นั้นๆ มาเปรียบเทียบกับเงื่อนไข (บรรทัดที่ 3) หากว่าในทุกๆมิติ ของเงื่อนไขนั้นอยู่ในระหว่าง ค่ามากที่สุดและค่าน้อยที่สุด ดังภาพที่ 24

```
Algorithm ClusterAssignment
1    for each closest cluster fch in FCH
2        for each dimension d in fch
3        if ( X_i BETWEEN fch[d].min AND fch[d].max )   bool_flag++
4        end for
5    if (bool_flag != dimension)   fch.add(X_i)
6    next closest fch
```

ภาพที่ 23 แสดงอัลกอริทึม ClusterSplitting

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

ผลและวิจารณ์

1. ข้อมูลสำหรับการทดลอง

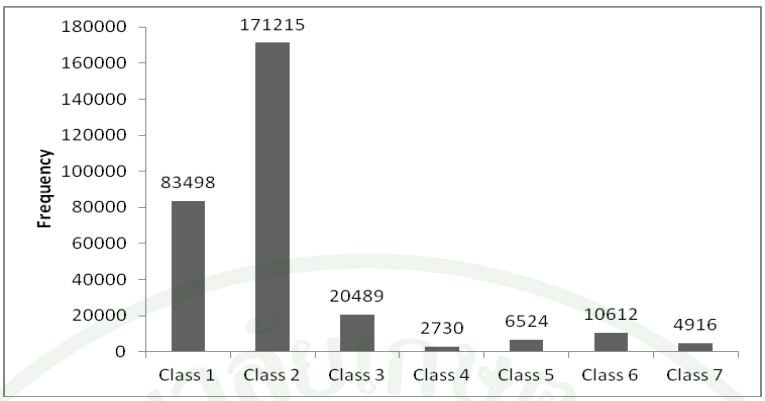
การทดลองของงานวิจัยนี้ใช้ข้อมูลชุดจริงซึ่งมีสองชุดข้อมูลได้แก่ Covertype และ KDDCup'99 ซึ่งมีรายละเอียดดังต่อไปนี้

1.1 Covertype เป็นชุดข้อมูลการเปลี่ยนแปลงของชนิดป่าฝน มีขนาด 54 มิติ แบ่งออกเป็นข้อมูลแบบตัวเลข 10 มิติ มีจำนวนข้อมูล 581,012 ตัว ทั้งหมด 7 คลาส ซึ่งสามารถแสดงความแตกต่างทางคุณภาพของ CE-Stream ได้อย่างชัดเจนเมื่อเปรียบเทียบกับเทคนิค E-Stream

1.2 KDDCup'99 เป็นชุดข้อมูลมาตรฐาน UCI เกี่ยวกับการบุกรุกของเครือข่าย ชุดข้อมูลนี้เป็นชุดข้อมูลที่ถูกใช้อย่างแพร่หลายในงานวิจัยทางด้านการแบ่งกลุ่มกระแสข้อมูล มีขนาด 43 มิติ แบ่งออกเป็นข้อมูลแบบตัวเลข 34 มิติ มีจำนวนข้อมูล 494,200 ตัว ทั้งหมด 23 คลาส

วิทยานิพนธ์ฉบับนี้นำเสนอผลการทดลองของชุดข้อมูล Covertype เป็นหลัก เนื่องจากผลการทดลองของ E-Stream บนชุดข้อมูล KDDCup'99 มีผลการทดลองที่มีประสิทธิภาพสูงอยู่แล้ว แต่ในทางตรงกันข้าม บนชุดข้อมูล Covertype นั้น ผลการทดลองนั้นมีประสิทธิภาพที่ต่ำ เพราะมีข้อมูลแบบตัวเลขมีเพียง 10 มิติ และมีการกระจายตัวของข้อมูลที่ไม่มาตรฐาน ดังภาพที่ 24 เห็นได้ว่า การกระจายตัวของข้อมูลนั้นส่วนใหญ่จะอยู่ที่ คลาสสีแดง และคลาสสีน้ำเงิน เป็นหลัก ซึ่ง CE-Stream สามารถช่วยปรับปรุงและพัฒนาประสิทธิภาพให้สูงขึ้นอย่างเห็นได้ชัด และ ในการทดลองนี้ ผู้เขียนใช้จำนวนชุดข้อมูลในการทดลอง ครึ่งหนึ่ง ของข้อมูลทั้งหมด เพื่อจุดประสงค์ในการลดระยะเวลาในการทดลอง

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์



ภาพที่ 24 แสดงการกระจายตัวของชุดข้อมูล Covertype

2. การวัดคุณภาพของกลุ่มข้อมูลที่ได้ด้วยค่าความบริสุทธิ์และค่าเอฟเมเชอร์

การทดลองนี้ได้แบ่งการวัดคุณภาพออกเป็นสามส่วน แบ่งออกเป็น

2.1 คุณภาพของกลุ่มข้อมูลผลลัพธ์เปรียบเทียบกับจำนวนของเงื่อนไขแบบสุ่ม

2.2 คุณภาพของกลุ่มข้อมูลผลลัพธ์เมื่อใช้เงื่อนไขแบบเลือกพิจารณาประเภทต่างๆ ได้แก่ การใช้เงื่อนไขแบบเชื่อมต่อ การใช้เงื่อนไขแบบไม่เชื่อมต่อ และ การใช้เงื่อนไขทั้งแบบเชื่อมต่อและไม่เชื่อมต่อร่วมกัน

2.3 คุณภาพของกลุ่มข้อมูลผลลัพธ์เมื่อมีการปรับค่าการเลือนหายของเงื่อนไข

สำหรับการทดลองนี้ เราได้ทำการวัดคุณภาพโดยเปรียบเทียบกับ เทคนิค E-Stream เดิม โดยงานวิจัยนี้ได้ปรับค่าพารามิเตอร์ต่างๆดังตารางที่ 1 อ้างอิงตามการปรับค่าของ E-Stream

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

ตารางที่ 1 ค่าพารามิเตอร์ต่างๆ ของทั้งสองอัลกอริทึม

เทคนิค E-Stream	เทคนิค CE-Stream
Decay_rate 0.1	Decay_rate 0.1
Radius_factor 3	Radius_factor 3
Stream_speed 500	Stream_speed 500
Fading_threshold 0.1	Fading_threshold 0.1
Merge_threshold 1.25	Merge_threshold 1.25
Active_threshold 5	Active_threshold 5
Maximum_isolate 10	Maximum_isolate 10
	Round_fading 13

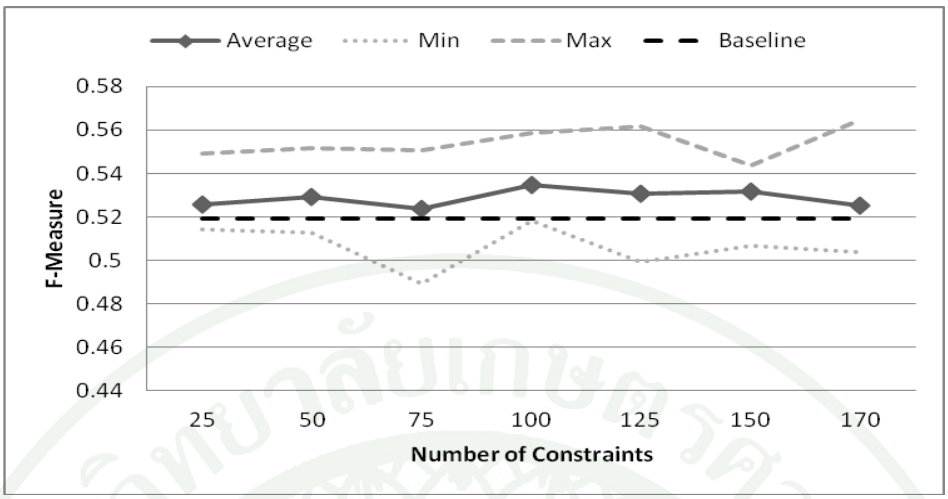
2.1 คุณภาพของกลุ่มข้อมูลเมื่อเปรียบเทียบกับจำนวนของเงื่อนไข

2.1.1 คุณภาพของกลุ่มข้อมูลเมื่อเปรียบเทียบกับจำนวนของเงื่อนไขแบบสุ่ม

โดยแนวคิดการผสมผสานเงื่อนไขในการแบ่งกลุ่มกระแสข้อมูลนั้น การทดลองจะมีการสุ่มชุดเงื่อนไขจำนวน 10 ชุด และทำการทดลองบนชุดเงื่อนไขทั้ง 10 ชุด และนำผลการทดลองของทุกๆชุดมาทำการหาค่าเฉลี่ย พร้อมกับเพิ่มจำนวนของเงื่อนไขครั้งละ 25 คู่ เพื่อให้ผลการทดลองมีประสิทธิภาพและความแม่นยำเพิ่มขึ้น บนชุดข้อมูล Covertype

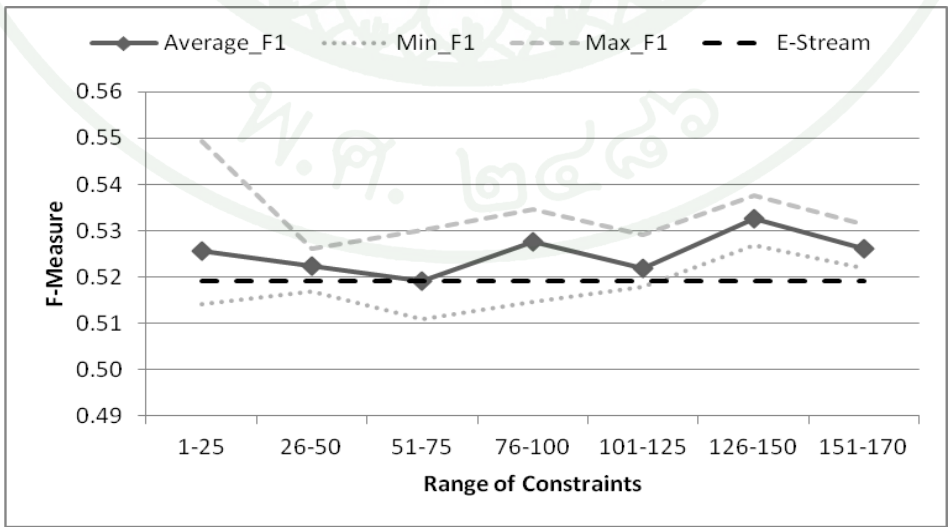
จากผลการทดลองและศึกษาพบว่า หากเลือกใช้เงื่อนไขแบบสุ่ม และเพิ่มขึ้นตามลำดับนั้น ถึงแม้ว่า คุณภาพของผลการทดลองเฉลี่ยนั้นมีแนวโน้มที่ดีขึ้น แต่ไม่สามารถเพิ่มประสิทธิภาพหรือความแม่นยำได้เสมอไป ดังภาพที่ 25 แสดงให้เห็นถึงค่าเอฟเมเชอร์ที่น้อยที่สุดนั้นต่ำกว่า E-Stream อย่างเห็นได้ชัด

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์



ภาพที่ 25 แสดงผลของกลุ่มข้อมูลผลลัพธ์เมื่อใช้เงื่อนไขแบบสุ่มบนชุดข้อมูล Covertype

หลังจากนั้น เราได้นำชุดเงื่อนไขจากภาพที่ 25 มาทำการแบ่งออกเป็นช่วงละ 25 คู่ เพื่อแสดงถึงประสิทธิภาพของเงื่อนไขในแต่ละช่วง สังเกตได้ว่า มีบางช่วงและบางชุดของเงื่อนไขที่ช่วยเพิ่มประสิทธิภาพของการแบ่งกลุ่มกระแสข้อมูลได้ดีมากขึ้น ในทางตรงกันข้ามมีบางช่วงและบางชุดของเงื่อนไขที่ลดประสิทธิภาพลง จากการทดลองนี้สรุปได้ว่า การนำเงื่อนไขผสมผสานในการแบ่งกลุ่มกระแสข้อมูลนั้น ต้องมีการเลือกเงื่อนไขอย่างมีนัยยะ เนื่องจากว่า การแบ่งกลุ่มกระแสข้อมูลนั้นมีการเปลี่ยนแปลงตลอดเวลา นั่นหมายความว่า เงื่อนไขบางชุดอาจไม่สามารถช่วยเพิ่มประสิทธิภาพ และอาจลดประสิทธิภาพลงด้วยซ้ำดังภาพที่ 26



ภาพที่ 26 แสดงการแปรผันกลุ่มข้อมูลผลลัพธ์ในช่วงต่างๆของเงื่อนไขบนชุดข้อมูล Covertype

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

2.1.2 การสร้างชุดเงื่อนไขที่มีประสิทธิภาพ

จากเหตุผลข้างต้น งานวิจัยนี้จึงได้มีการนำเทคนิค Associative Classification Rule มาประยุกต์ใช้ในการเลือกชุดเงื่อนไขเพื่อให้สอดคล้องกับชุดข้อมูลทดลองและมีประสิทธิภาพ โดยจะนำชุดข้อมูลทดลองหากฎที่มีค่าความเชื่อมั่นสูง (Confidence) เมื่อเปรียบเทียบกับจำนวนข้อมูลทั้งหมด โดยมีสูตรดังนี้

$$Conf(X \Rightarrow Y) = Supp(X \cup Y)_{Class} / Supp(X \cup Y)_{All} \quad \textbf{(22)}$$

โดยที่ค่าสนับสนุน (Support) หมายถึง จำนวนนับของข้อมูลที่เกิดขึ้นในชุดข้อมูล ในงานนี้ ผู้เขียนได้กำหนดให้ X เป็น ค่าตัวแปรมิติ และ Y เป็นคลาส ส่วน $Supp_{Class}$ คือ จำนวนนับที่เกิดขึ้นเฉพาะคลาสนั้นๆ ส่วน $Supp_{All}$ คือ จำนวนนับที่เกิดขึ้นในชุดข้อมูลทั้งหมด

ตารางที่ 2 แสดงตัวอย่างของกฎที่ได้จาก Associative Classification

Rules	$Supp_{cls}$	$Supp_{All}$	Conf
1. Hillshade_Noon='(221.8-237.9]' ==> class=3	4106	124268	0.033
2. Horizontal_Distance_To_Roadway='(636.2-954.3]'==>class=3	4082	22708	0.180
3. Elevation='(2354.5-2453.6]' ==> class=3	4047	7103	0.570
4. Horizontal_Distance_To_Roadways='(-inf-318.1]' ==> class=3	4042	11067	0.365

จากตัวอย่างกฎที่ 3 ตารางที่ 2 แสดงให้เห็นถึงข้อมูลที่มีค่า Elevation อยู่ในช่วงที่ 2354.5 ถึง 2453.6 มีจำนวน 4047 ตัว เฉพาะชุดข้อมูลที่อยู่ในคลาส 3 และ มีจำนวน 7103 ตัวในชุดข้อมูลทั้งหมด ทำให้มีค่าความเชื่อมั่นสูงถึง 0.570

หลังจากที่เราได้กฎจากเทคนิคข้างต้น ผู้เขียนจะทำการเลือก จุดข้อมูลโดยอ้างอิงจากกฎทั้งหมดเพื่อนำมาสร้างเป็นเงื่อนไขแบบเชื่อมต่อ ในส่วนของเงื่อนไขแบบไม่เชื่อมต่อ จะถูกสร้างโดยใช้เงื่อนไขแบบเชื่อมต่อเป็นพื้นฐานเพื่อให้ชุดเงื่อนไขทั้งสองมีความสอดคล้องกัน

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

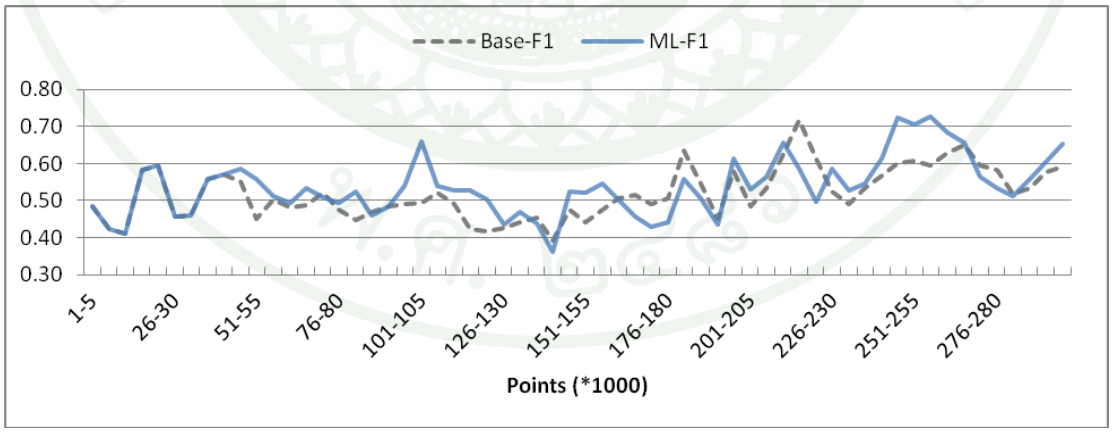
การนำเทคนิคนี้มาใช้เปรียบเสมือนการได้เงื่อนไขมาจากผู้เชี่ยวชาญ เพราะ สามารถได้เงื่อนไขที่มีความสามารถในการบ่งชี้ลักษณะเฉพาะของแต่ละคลาสได้อย่างชัดเจน และมีความหมายในเชิงข้อมูล อีกด้วย

2.2 คุณภาพของกลุ่มเงื่อนไขเมื่อใช้เงื่อนไขประเภทต่างๆ

ในการทดลองหัวข้อนี้ ใช้ชุดข้อมูล Covertype เป็นหลัก ยกเว้นในส่วนของการใช้เงื่อนไขแบบเชื่อมต่อและไม่เชื่อมต่อร่วมกัน จะทำการทดลองบนชุดข้อมูล KDDCup'99 ด้วย เพื่อให้เห็นว่า CE-Stream สามารถเพิ่มประสิทธิภาพโดยรวมให้สูงขึ้นได้

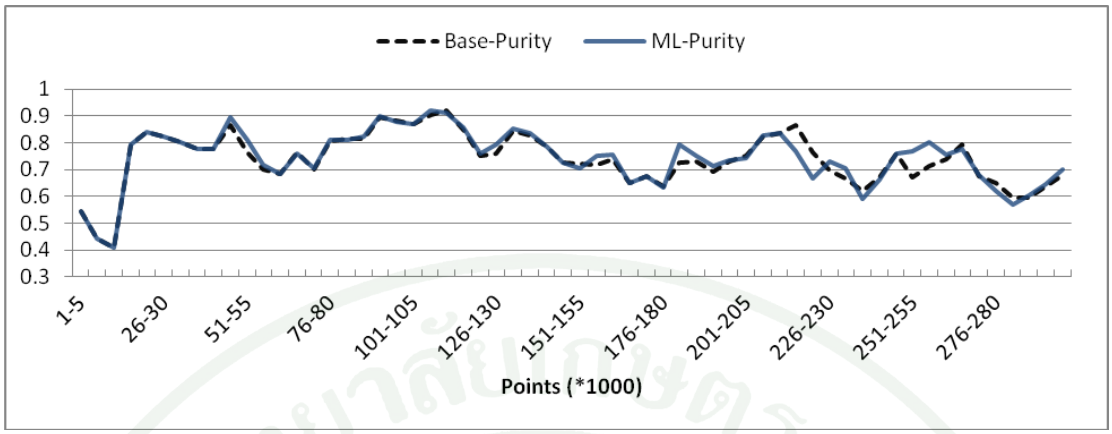
2.2.1 การใช้เงื่อนไขแบบเชื่อมต่อ

การใช้เงื่อนไขแบบเชื่อมต่อในช่วงระยะการแยกตัวกันของกลุ่มข้อมูลนั้นแสดงให้เห็นถึง คุณภาพของกลุ่มข้อมูลที่ดีขึ้นเมื่อเปรียบเทียบกับ E-Stream ทั้งในเชิงของ ค่าความบริสุทธิ์ และค่าเอฟเมเชอร์ โดยที่ เอฟเมเชอร์นั้นเพิ่มขึ้นจาก 51.92 ไปจนถึง 53.77 และ ค่าความบริสุทธิ์เพิ่มขึ้นจาก 73.95 ไปจนถึง 74.53 ดังภาพที่ 27 และภาพที่ 28 ซึ่งเงื่อนไขในที่นี้ใช้เพียงแค่ 20 คู่เท่านั้น

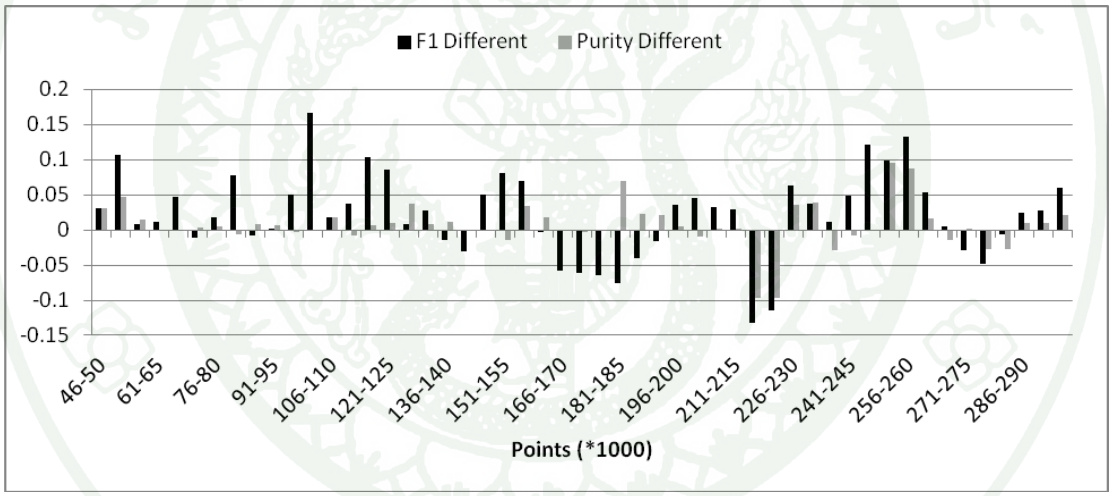


ภาพที่ 27 แสดงผลการทดลองค่าเอฟเมเชอร์ของการใช้เงื่อนไขแบบเชื่อมต่อ เมื่อเทียบกับ E-Stream บนชุดข้อมูล Covertype

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์



ภาพที่ 28 แสดงผลการทดลองค่าบริสุทธิ์ของการใช้เงื่อนไขแบบเชื่อมต่อ เมื่อเทียบกับ E-Stream บนชุดข้อมูล Covertype



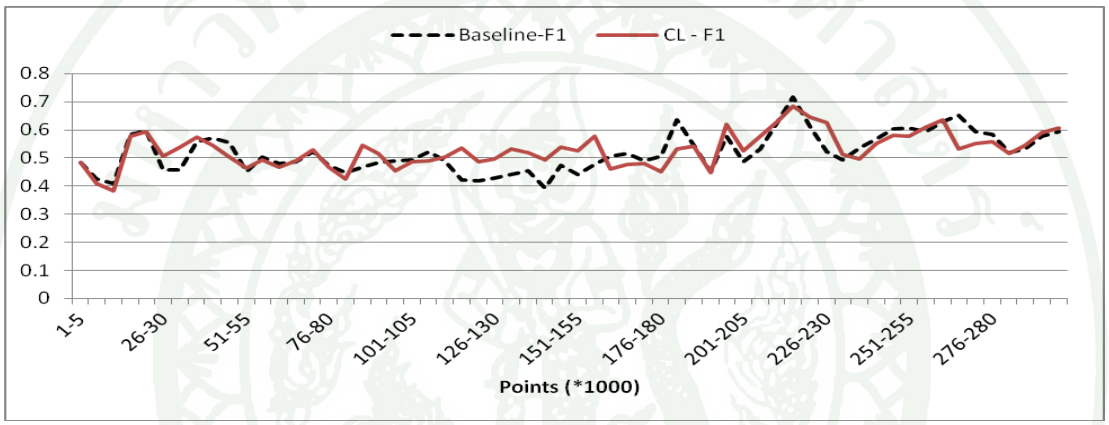
ภาพที่ 29 แสดงค่าเอฟเมเชอร์และค่าความบริสุทธิ์ที่เพิ่มขึ้น/ลดลงเมื่อมีการใช้เงื่อนไขแบบเชื่อมต่อ เมื่อเทียบกับ E-Stream

จากภาพที่ 29 แสดงถึงค่าความแตกต่างของผลการทดลองการใช้เงื่อนไขแบบเชื่อมต่อเทียบกับเทคนิค E-Stream โดย แกนศูนย์คือคุณภาพข้อมูลผลลัพธ์ของ E-Stream แท่งสีฟ้าคือผลต่างของเอฟเมเชอร์ และ แท่งสีแดงคือผลต่างของค่าบริสุทธิ์ เมื่อเปรียบเทียบกับ E-Stream ตามลำดับ เห็นได้ว่า ผลต่างที่เป็นบวกนั้นมีค่อนข้างมากกว่าเมื่อเทียบกับผลต่างที่เป็นลบทั้งค่าเอฟเมเชอร์และ ค่าบริสุทธิ์

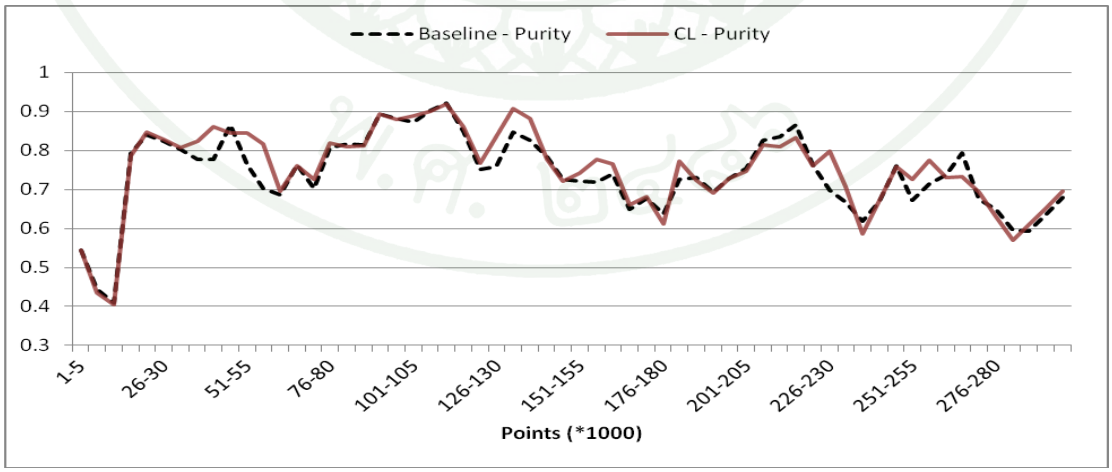
ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

2.2.2 การใช้เงื่อนไขแบบไม่เชื่อมต่อ

การใช้เงื่อนไขแบบไม่เชื่อมต่อในการส่งมอบข้อมูลสมาชิกให้กลุ่มข้อมูลนั้นแสดงให้เห็นถึง คุณภาพของกลุ่มข้อมูลที่ดีขึ้นเมื่อเปรียบเทียบกับ E-Stream ทั้งในเชิงของ ค่าความบริสุทธิ์และค่าเอฟเมเชอร์ โดยที่ เอฟเมเชอร์นั้นเพิ่มขึ้นจาก 51.92 ไปจนถึง 52.84 และ ค่าความบริสุทธิ์เพิ่มขึ้นจาก 73.95 ไปจนถึง 75.24 ดังภาพที่ 30 และภาพที่ 31 ซึ่งเงื่อนไขในที่นี้ใช้เพียงแค่ 14 คู่เท่านั้น

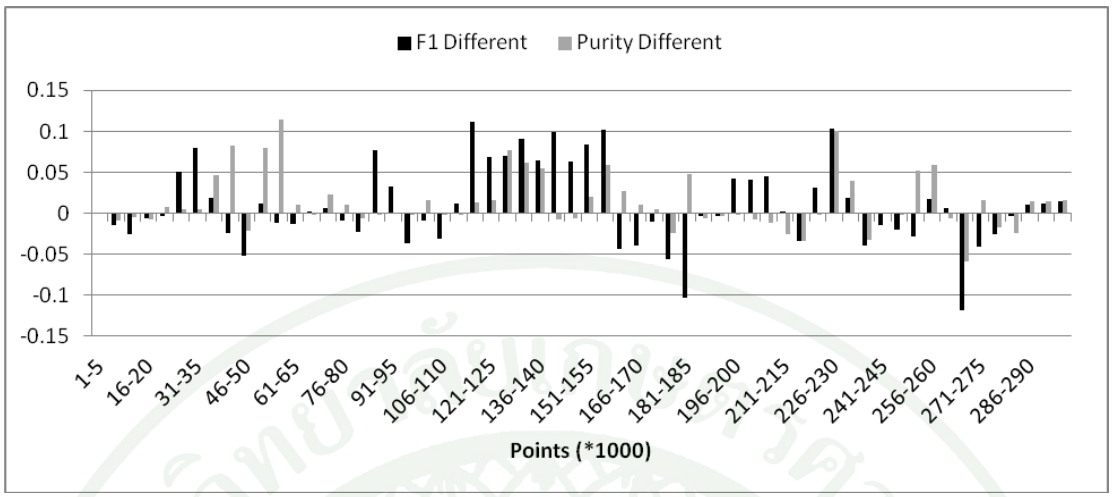


ภาพที่ 30 แสดงผลการทดลองค่าเอฟเมเชอร์ของการใช้เงื่อนไขแบบไม่เชื่อมต่อ เมื่อเทียบกับ E-Stream บนชุดข้อมูล Covertype



ภาพที่ 31 แสดงผลการทดลองค่าบริสุทธิ์ของการใช้เงื่อนไขแบบไม่เชื่อมต่อ เมื่อเทียบกับ E-Stream บนชุดข้อมูล Covertype

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์



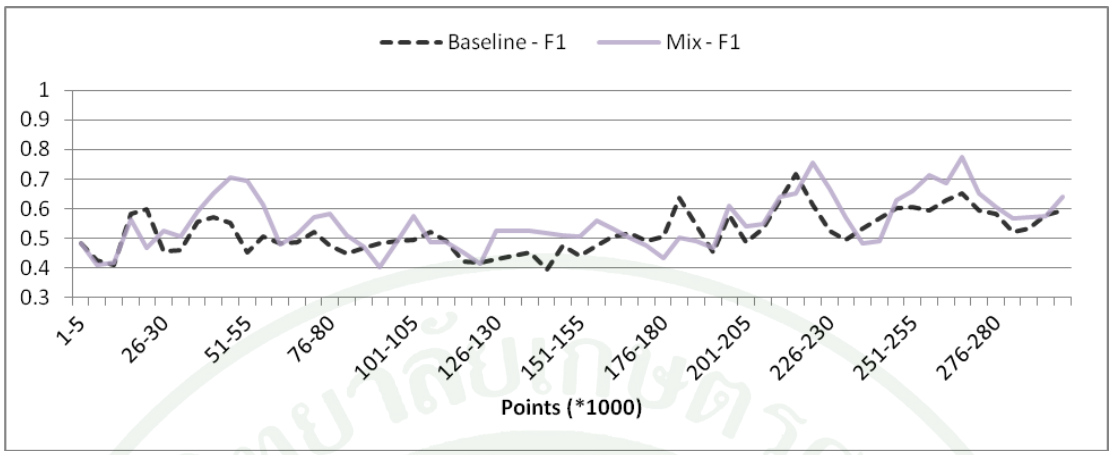
ภาพที่ 32 แสดงค่าเอฟเมเชอร์และค่าความบริสุทธิ์ที่เพิ่มขึ้น/ลดลงเมื่อมีการใช้เงื่อนไขแบบไม่เชื่อมต่อ เมื่อเทียบกับ E-Stream

จากภาพที่ 32 แสดงถึงค่าความแตกต่างของผลการทดลองการใช้เงื่อนไขแบบไม่เชื่อมต่อเทียบกับเทคนิค E-Stream โดย แกนศูนย์คือคุณภาพข้อมูลผลลัพธ์ของ E-Stream แท่งสีฟ้าคือผลต่างของเอฟเมเชอร์ และ แท่งสีแดงคือผลต่างของค่าบริสุทธิ์ เมื่อเปรียบเทียบกับ E-Stream ตามลำดับ เห็นได้ว่า ผลต่างที่เป็นบวกของค่าบริสุทธิ์นั้นมีค่อนข้างมากกว่าเมื่อเทียบกับผลต่างที่เป็นลบ

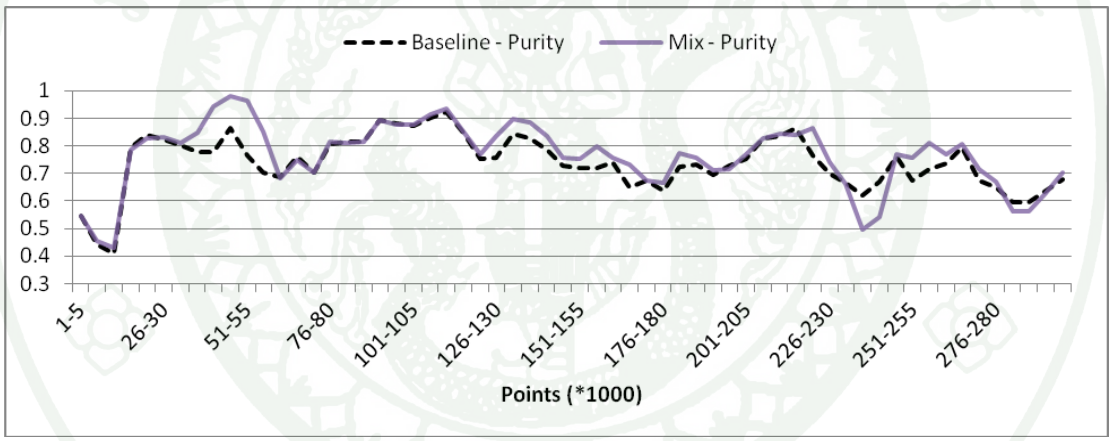
2.2.3 การใช้เงื่อนไขแบบเชื่อมต่อและไม่เชื่อมต่อร่วมกัน

เมื่อเรานำเงื่อนไขทั้งสองแบบมาใช้ร่วมกัน สามารถทำให้คุณภาพของกลุ่มข้อมูลนั้นดีขึ้นอย่างชัดเจน ค่าเอฟเมเชอร์ และค่าความบริสุทธิ์ นั้นเพิ่มขึ้นถึง 55.20 และ 76.43 จากเดิมที่ 51.92 และ 73.95 บนชุดข้อมูล Covertype ตามลำดับ ดังภาพที่ 33 และ ภาพที่ 34

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

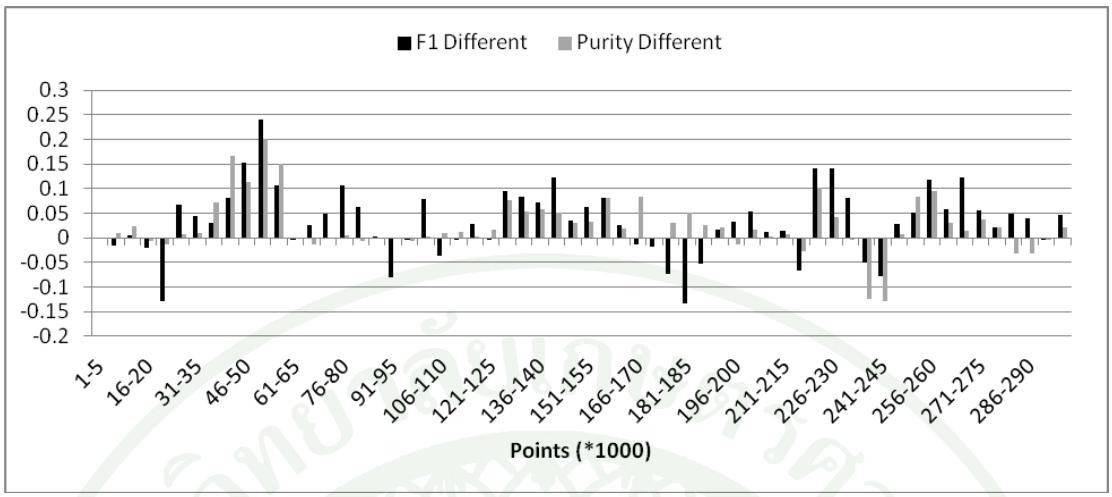


ภาพที่ 33 แสดงผลการทดลองค่าเอฟเมเชอร์ของการใช้เงื่อนไขแบบเชื่อมต่อและไม่เชื่อมต่อร่วมกัน เมื่อเทียบกับ E-Stream บนชุดข้อมูล Covertype



ภาพที่ 34 แสดงผลการทดลองค่าบริสุทธิ์ของการใช้เงื่อนไขแบบเชื่อมต่อและไม่เชื่อมต่อร่วมกัน เมื่อเทียบกับ E-Stream บนชุดข้อมูล Covertype

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

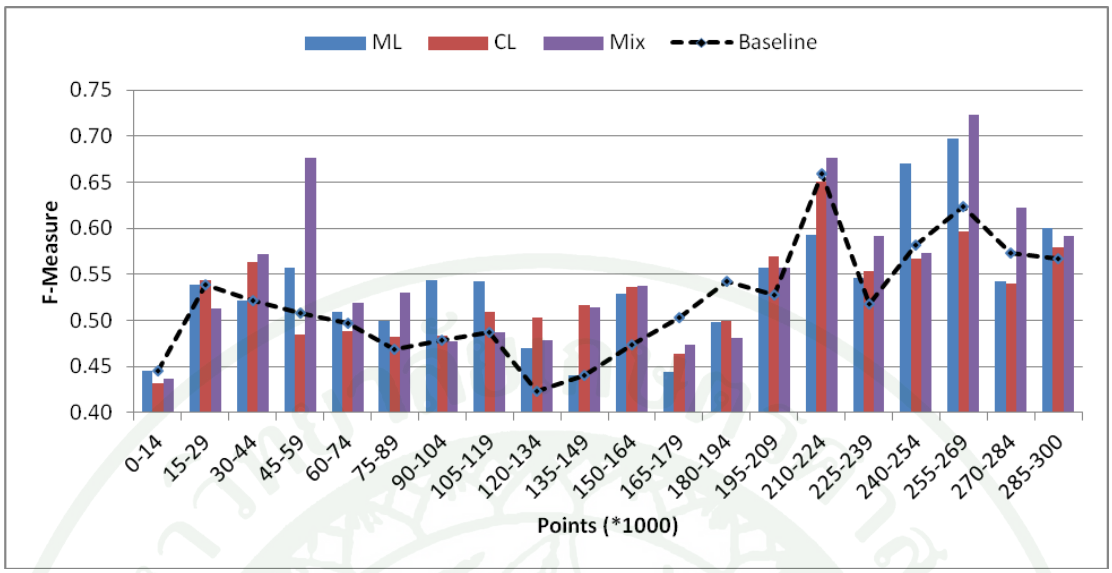


ภาพที่ 35 แสดงค่าเอฟเมเชอร์และค่าความบริสุทธิ์ที่เพิ่มขึ้น/ลดลงเมื่อมีการใช้เงื่อนไขแบบเชื่อมต่อและไม่เชื่อมต่อ เมื่อเทียบกับ E-Stream

จากภาพที่ 35 แสดงถึงค่าความแตกต่างของผลการทดลองการใช้เงื่อนไขแบบเชื่อมต่อและไมเชื่อมต่อร่วมกันเทียบกับเทคนิค E-Stream โดย แกนศูนย์คือคุณภาพข้อมูลผลลัพธ์ของ E-Stream แท่งสีฟ้าคือผลต่างของเอฟเมเชอร์ และ แท่งสีแดงคือผลต่างของค่าบริสุทธิ์ เมื่อเปรียบเทียบกับ E-Stream ตามลำดับ เห็นได้ว่า ผลต่างที่เป็นบวกของทั้งค่าบริสุทธิ์และเอฟเมเชอร์นั้นมีมากกว่าผลต่างที่เป็นลบอย่างเห็นได้ชัด ซึ่งมีจำนวนที่น้อยมาก

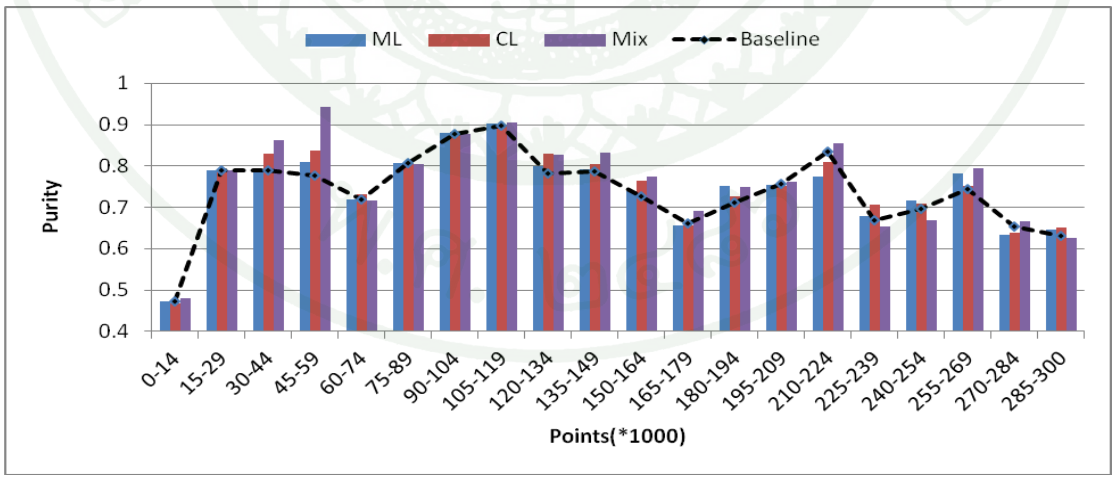
อย่างไรก็ตาม จากภาพที่ 36 และภาพที่ 37 แสดงถึงแนวโน้มของการใช้เงื่อนไขร่วมกันซึ่งในบางช่วงข้อมูลการนำเงื่อนไขทั้งสองประเภทมาใช้ร่วมกัน อาจไม่ส่งเสริมกันได้ หากเงื่อนไขนั้นไม่สอดคล้องกัน โดยแท่งสีม่วงคือผลการทดลองคุณภาพเมื่อนำเงื่อนไขทั้งสองแบบมารวมกัน สีฟ้าเป็นเงื่อนไขแบบเชื่อมต่อ และ สีแดงเป็นเงื่อนไขแบบไม่เชื่อมต่อ

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์



ภาพที่ 36 แสดงผลการทดลองเปรียบเทียบค่าเอฟเมเชอร์ของเงื่อนไขแต่ละประเภท บนชุดข้อมูล Covertype

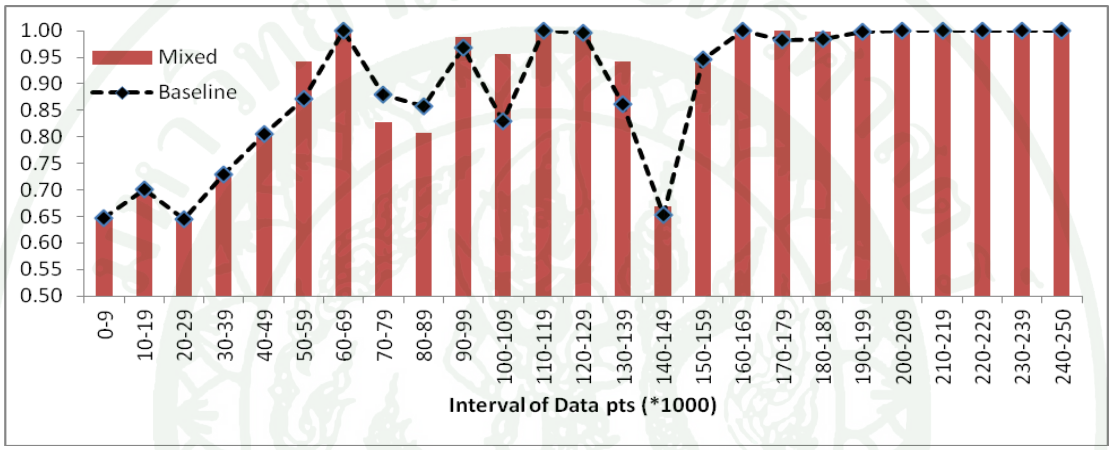
ผลการทดลองของ CE-Stream นั้น มีบางช่วงที่คุณภาพของกลุ่มข้อมูลผลลัพธ์ต่ำกว่า ผลการทดลองเดิม แต่โดยรวมเห็นได้ว่าทั้งค่าเอฟเมเชอร์ และค่าบริสุทธิ์ โดยเฉลี่ยนั้น มีค่าสูงขึ้นหรือเทียบเท่ากับของเดิม



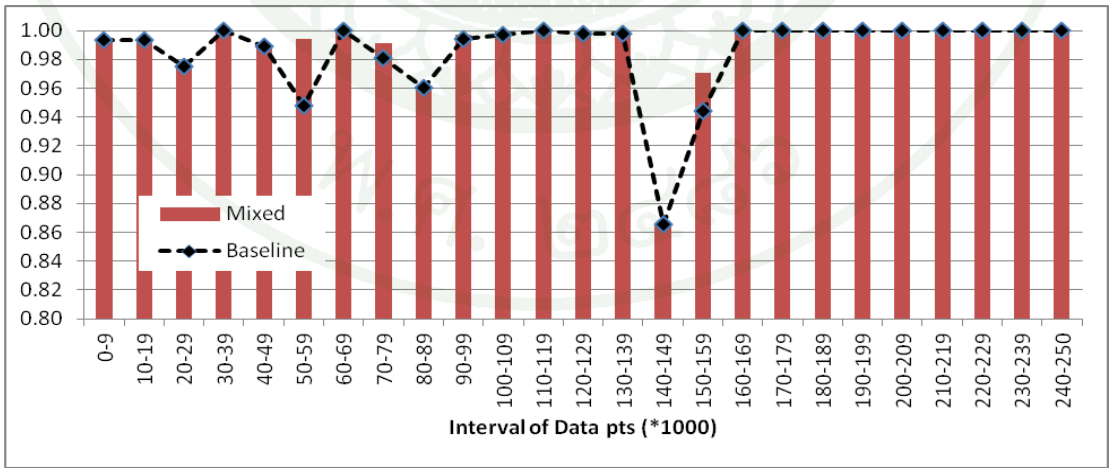
ภาพที่ 37 แสดงผลการทดลองเปรียบเทียบค่าบริสุทธิ์ของเงื่อนไขแต่ละประเภท บนชุดข้อมูล Covertype

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

เมื่อนำเงื่อนไขใช้ในชุดข้อมูล KDDCup'99 แม้ว่า ประสิทธิภาพของ E-Stream นั้นสูงมากอยู่แล้ว แต่ CE-Stream สามารถเพิ่มประสิทธิภาพให้ดีขึ้นเล็กน้อย ดังภาพที่ 38 และ 39 จากเดิมค่าเอฟเมเชอร์ 89.56 ปรับเพิ่มขึ้นเป็น 90.52 และค่าบริสุทธิ์จากเดิม 98.55 เป็น 98.93 ในภาพที่ 35 และ ภาพที่ 36 แสดงให้เห็นถึงคุณภาพของกลุ่มข้อมูลที่ดีขึ้นหรือเทียบเท่ากับเทคนิคเดิมโดยเปรียบเทียบในช่วงของข้อมูล ช่วงละ 10000 ตัว พิจารณาจากเงื่อนไขทั้งหมด 50 คู่ จำแนกเป็นเงื่อนไขเชื่อมต่อ 30 คู่ และ เงื่อนไขไม่เชื่อมต่อ 20 คู่ สร้างจากวิธีที่ได้กล่าวไว้ข้างต้น



ภาพที่ 38 แสดงผลการทดลองเปรียบเทียบค่าเอฟเมเชอร์ในการใช้เงื่อนไขแบบเชื่อมต่อและไม่เชื่อมต่อร่วมกัน เมื่อเทียบกับ E-Stream บนชุดข้อมูล KDDCup'99



ภาพที่ 39 แสดงผลการทดลองเปรียบเทียบค่าความบริสุทธิ์ในการใช้เงื่อนไขแบบเชื่อมต่อและไม่เชื่อมต่อร่วมกัน เมื่อเทียบกับ E-Stream บนชุดข้อมูล KDDCup'99

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

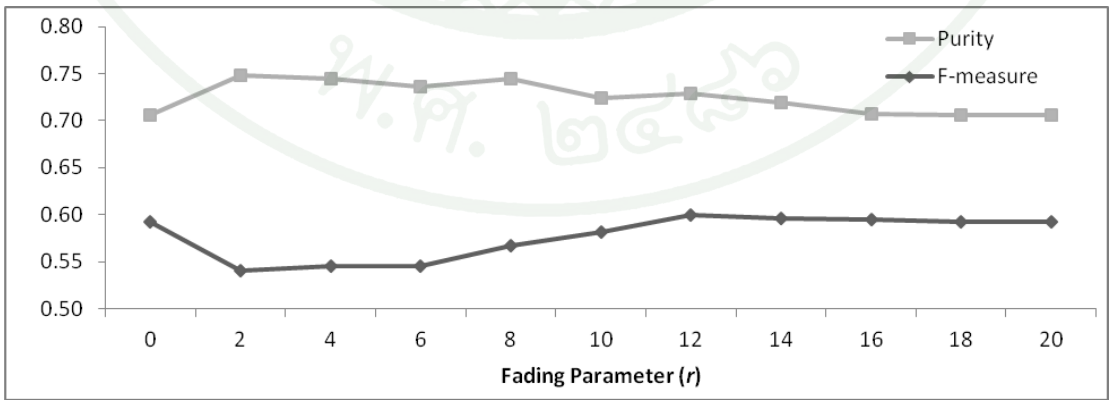
ตารางที่ 3 แสดงถึงผลสรุปการทดลองบนชุดข้อมูลทั้งสองโดยการใช้เงื่อนไขประเภทต่างๆ Baseline คือ การใช้ E-Stream ที่ไม่มีเงื่อนไขช่วย จะเห็นได้ว่า E-Stream นั้นมีค่าเอฟเมเชอร์ และค่าบริสุทธิ์ ต่ำกว่า โดยส่วนมาก

ตารางที่ 3 แสดงสรุปตารางผลการทดลองของชุดข้อมูลทั้งสองและค่าวัดผลต่างๆ

	ชุดข้อมูล Covertype		ชุดข้อมูล KDDCup'99	
	ค่าเอฟเมเชอร์	ค่าบริสุทธิ์	ค่าเอฟเมเชอร์	ค่าบริสุทธิ์
E-Stream	51.92	73.95	89.56	98.55
E-Stream + ML	53.77	74.53	90.16	99.09
E-Stream + CL	52.84	75.24	88.66	98.53
E-Stream + Mix	55.20	76.43	90.52	98.93

2.3 คุณภาพของกลุ่มข้อมูลผลลัพธ์เมื่อมีการปรับค่าการเลือนหายของเงื่อนไข

การทดลอง CE-Stream นั้นมีการทดลองปรับค่าพารามิเตอร์จำนวนรอบของการเลือนหายของเงื่อนไขเพื่อสังเกตผลกระทบต่อคุณภาพของกลุ่มข้อมูลผลลัพธ์ โดยการทดลองนี้ได้ใช้ ชุดเงื่อนไขแบบเชื่อมต่อและไม่เชื่อมต่อ จำนวน 20 คู่ และ 15 คู่ ตามลำดับ บนชุดข้อมูล Covertype ตามที่ได้ทดลองไว้ข้างต้นก่อนหน้านี้



ภาพที่ 40 แสดงผลการทดลองเมื่อปรับค่าจำนวนรอบการเลือนหายของเงื่อนไขโดยเปรียบเทียบค่าบริสุทธิ์ และค่าเอฟเมเชอร์บนชุดข้อมูล Covertype

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

จากผลการทดลองดังภาพที่ 40 แสดงให้เห็นถึงผลกระทบต่อคุณภาพของกลุ่มข้อมูลผลลัพธ์เมื่อมีการปรับเปลี่ยนค่าจำนวนรอบการเลือนหายของเงื่อนไข โดยหากปรับค่าในช่วง 2 ถึง 12 ค่าเอฟเมเชอร์จะลดลงและค่อยๆเพิ่มขึ้นตามลำดับ จนกระทั่งค่อนข้างคงที่เมื่อ ค่าการเลือนหายมีค่ามากกว่า 13 ขึ้นไป ในทางตรงกันข้าม ค่าบริสุทธิ์จะมีค่าเพิ่มขึ้นและค่อยๆลดลงตามลำดับในช่วง 2 ถึง 12 และจะค่อนข้างคงที่เมื่อ ค่าการเลือนหายมีค่ามากกว่า 13 ขึ้นไป

อย่างไรก็ตาม การปรับจำนวนรอบการเลือนหายนั้นอาจขึ้นอยู่กับชุดของเงื่อนไขที่ถูกใช้ในระบบด้วยเช่นกัน

3. การเปรียบเทียบเวลาในการทำงานเทียบกับ E-Stream

เนื่องจาก CE-Stream มีการเพิ่มฟังก์ชัน เพื่อรองรับการทำงานร่วมกับเงื่อนไขประเภทต่างๆ ดังนั้นการเปรียบเวลาในการทำงานระหว่าง CE-Stream และ E-Stream จึงสามารถวัดได้ 2 แบบ คือ การวิเคราะห์ความซับซ้อนด้านเวลา และ อัตราการประมวลผลข้อมูลในเวลา 1 วินาที

การวิเคราะห์ความซับซ้อนด้านเวลา (Time Complexity) ของ CE-Stream ได้มีแสดงถึงขั้นตอนต่างๆ ดังภาพที่ 41 โดยกำหนดให้ K คือจำนวน K กลุ่มข้อมูล, N_c คือจำนวนเงื่อนไขที่ปรากฏในระบบ และ J คือจำนวนมิติของจุดข้อมูล

อัตราการเติบโตของฟังก์ชัน CheckActiveAndUpdateConstraints นั้นมีค่าเป็น $O(JN_c)$ เนื่องจาก ณ เวลาหนึ่ง ระบบจะทำการวนรอบเพื่อเปรียบเทียบจุดข้อมูลใหม่ที่เข้ามากับชุดของเงื่อนไขที่ถูกป้อนในระบบ N จุด ในทุกๆ J มิติ

อัตราการเติบโตของฟังก์ชัน CheckSplit นั้นมีค่าเป็น $O(KJN_c)$ เนื่องจาก ณ เวลาหนึ่ง ระบบจะทำการหาจุดแบ่งในมิติของข้อมูล J โดยวนรอบกลุ่มข้อมูลทั้งหมด K กลุ่มในระบบ เมื่อพบกลุ่มข้อมูลที่จะพิจารณา ระบบจะทำการตรวจสอบเงื่อนไขที่ถูกใช้งานในช่วงเวลาดังกล่าวเป็นจำนวน N_c จุด เพื่อหาเงื่อนไขที่เกี่ยวข้องกับจุดแบ่งของกลุ่มข้อมูลนั้นๆ

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

อัตราการเติบโตของฟังก์ชัน PrioritizeConstraints นั้นมีค่าเป็น $O(N_c \log N_c)$ เนื่องจากเงื่อนไขถูกเก็บในรูปแบบของเวกเตอร์ (vector) ซึ่งการจัดลำดับของเวกเตอร์นั้นอัตราการเติบโตเป็น O(N log N) นั่นเอง (C++ Reference, 2013)

อัตราการเติบโตของฟังก์ชัน FadingConstraints นั้นมีค่าเป็น $O(N_c)$ เนื่องจากระบบจะทำการปรับปรุงค่าเลือนหายของเงื่อนไขเป็นจำนวน N_c และอัตราการเติบโตของฟังก์ชัน ClusterAssignment มีค่าเป็น $O(KN_c)$ เนื่องจาก ระบบจะทำการส่งมอบสมาชิกโดยตรวจสอบจากเงื่อนไขจำนวน N_c เปรียบเทียบกับจำนวนของกลุ่มข้อมูล K กลุ่ม

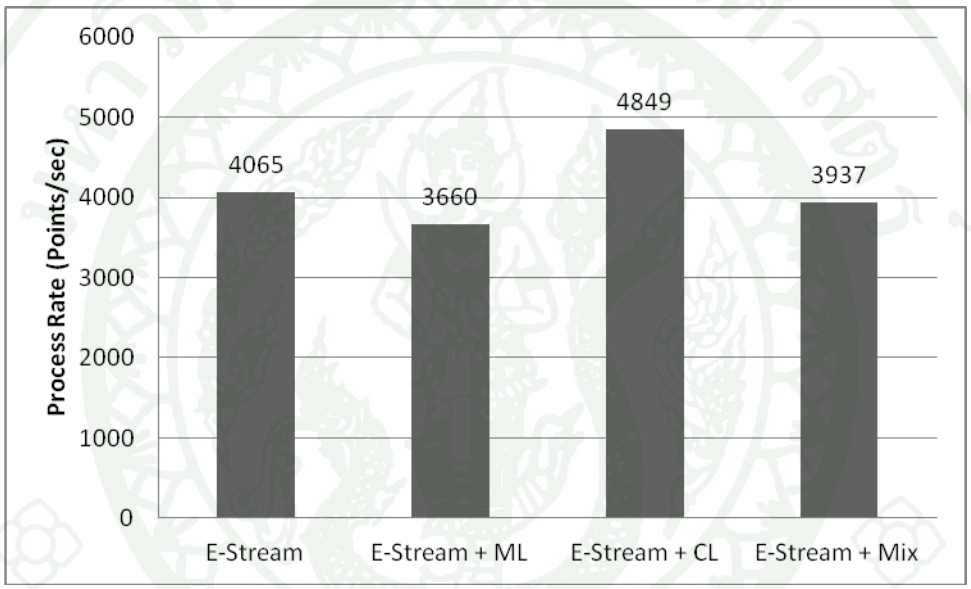
ในกรณีที่มีการป้อนชุดเงื่อนไขเป็นจำนวนมาก อาจทำให้ค่า N_c นั้นมีค่าสูงจนทำให้ฟังก์ชัน CheckSplit ใช้เวลาในการประมวลผลมาก ซึ่งอาจก่อให้เกิดการประมวลผลของระบบที่ช้าลงได้

Algorithm CE-Stream		
1	Retrieve new data X_i	
2	**CheckActiveAndUpdateConstraints**	**$O(JN_c)$**
3	FadingAllClusters	O(K)
4	**PrioritizeConstraints**	**$O(N_c \log N_c)$**
5	**CheckSplit**	**$O(KJN_c)$**
6	**MergeOverlapCluster**	**$O(K^2J)$**
7	**FadingConstraints**	**$O(N_c)$**
8	FlagActiveCluster	O(K)
9	(minDistance, index) ← FindClosestCluster	O(KJ)
10	if minDistance < radius_factor	
11	**ClusterAssignment(X_i,FCH_{index})**	**$O(KN_c)$**
12	else	
13	create new FCH from X_i	
13	waiting for new data	

ภาพที่ 41 แสดงอัตราการเติบโตในแต่ละฟังก์ชันของ CE-Stream

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

ถึงแม้ว่าจากการแสดงถึงความซับซ้อนของเวลา ประสิทธิภาพของ CE-Stream นั้นจะมีอัตราการเติบโตของฟังก์ชันโดยรวม สูงกว่า E-Stream แต่จากการทดลองดังภาพที่ 42 หากมีการนำเงื่อนไขที่มีประสิทธิภาพ และมีจำนวนไม่มากจนเกินไปนั้น อาทิเช่น เงื่อนไขแบบไม่เชื่อมต่อ หรือเงื่อนไขแบบเชื่อมต่อและไม่เชื่อมต่อร่วมกัน สามารถทำให้ CE-Stream นั้นมีความเร็วในการประมวลผลที่เร็วกว่า หรือเทียบเท่า เนื่องจาก เงื่อนไขที่มีประสิทธิภาพจะช่วยลดอัตราการแยกตัวของกลุ่มข้อมูลที่ไม่จำเป็น รวมไปถึงลดอัตราการส่งมอบข้อมูลสมาชิกที่ผิด ซึ่งสามารถนำไปสู่ การพิจารณาการแบ่งกลุ่มที่อาจใช้เวลาประมวลผลมากขึ้นนั่นเอง



ภาพที่ 42 แสดงอัตราการประมวลผลข้อมูลใน 1 วินาทีของระบบ ในการใช้เงื่อนไขประเภทต่างๆ

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

สรุปและข้อเสนอแนะ

สรุป

งานวิจัยนี้ได้นำเสนอเทคนิคการแบ่งกลุ่มกระแสข้อมูลแบบมีเงื่อนไข CE-Stream ซึ่งพัฒนาต่อจาก E-Stream โดยเทคนิค CE-Stream ได้มีการนำเงื่อนไขระดับข้อมูลได้แก่ เงื่อนไขแบบเชื่อมต่อ (Must-Link Constraints) และเงื่อนไขแบบไม่เชื่อมต่อ (Cannot-Link Constraints) มาใช้ในการแยกตัวของกลุ่มข้อมูล และการส่งมอบข้อมูลสมาชิกตามลำดับ รวมถึงการออกแบบเงื่อนไขให้รองรับกับลักษณะเฉพาะของกระแสข้อมูล ได้แก่ การพิจารณาการใช้งานเงื่อนไข (Constraints Activation) การเลือนหายของเงื่อนไข (Constraints Fading) และ การหมดอายุของเงื่อนไข (Constriants Outdating) เมื่อเปรียบเทียบคุณภาพของกลุ่มข้อมูลผลลัพธ์ระหว่างเทคนิค CE-Stream และเทคนิค E-Stream สรุปได้ว่า

1. คุณภาพของข้อมูลผลลัพธ์ เมื่อเปรียบเทียบด้วยค่าบริสุทธิ์และค่าเอฟเมเชอร์นั้น เทคนิค CE-Stream โดยใช้เงื่อนไขประเภทต่างๆ ให้ประสิทธิภาพที่ดีกว่า เทคนิค E-Stream เกือบทุกกรณี

2. การเลือกใช้ชุดเงื่อนไขอย่างมีนัยยะมีผลต่อประสิทธิภาพและคุณภาพของกลุ่มข้อมูลผลลัพธ์ มากกว่า จำนวนของชุดเงื่อนไขที่พิจารณา อย่างเห็นได้ชัด

3. เทคนิค CE-Stream ใช้เวลาในการประมวลผลน้อยกว่า เทคนิค E-Stream หากมีการพิจารณาชุดเงื่อนไขที่มีจำนวนไม่มากจนเกินไป

ข้อเสนอแนะ

1. การใช้เงื่อนไขในขั้นตอนต่างๆของ CE-Stream ยังไม่สามารถรองรับถึงการใส่เงื่อนไขในระบบแบบอิสระ นั่นหมายถึง หากผู้ใช้มีเงื่อนไขใหม่ที่เพิ่มขึ้นจากตอนเริ่มของระบบ ผู้ใช้จะไม่สามารถป้อนเงื่อนไขใหม่เข้าไปในระบบได้ เนื่องจาก กระแสข้อมูลอาจมีการเปลี่ยนแปลงต่อเนื่องเมื่อเวลาผ่านไปซึ่ง เงื่อนไขเดิมไม่อาจช่วยให้กลุ่มข้อมูลผลลัพธ์มีประสิทธิภาพที่ดีขึ้น

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

2. จากผลการทดลองจะเห็นได้ว่า ยังมีบางช่วงที่มีการใช้เงื่อนไขต่างประเภทร่วมกัน อาจก่อให้เกิดคุณภาพของกลุ่มผลลัพธ์ที่มีคุณภาพต่ำลง เนื่องจาก เทคนิค CE-Stream ยังขาดการพิจารณาความสอดคล้องระหว่างเงื่อนไข ในเวลาที่ระบบทำงานอยู่ ซึ่งหากมีการพิจารณาเงื่อนไขในการใช้ร่วมกันส่วนนี้ จะสามารถทำให้ กลุ่มข้อมูลผลลัพธ์นั้น มีประสิทธิภาพได้ดียิ่งขึ้น

3. เงื่อนไขสามารถประยุกต์ใช้ในขั้นตอนการรวมกันของกลุ่มข้อมูล ทั้งการรวมกลุ่มของข้อมูลโดดเดี่ยว (Isolate Data Point Merging) และ การรวมกลุ่มของกลุ่มข้อมูล (Clusters Mergning) เพื่อช่วยเพิ่มค่าบริสุทธิ์ของกลุ่มข้อมูลผลลัพธ์ นำไปสู่ประสิทธิภาพของการแบ่งกลุ่มข้อมูลที่มีประสิทธิภาพสูงขึ้น

4. การสร้างเงื่อนไขเพื่อใช้ในระบบควรมีวิธีการสร้างที่สามารถพิจารณารูปแบบการเปลี่ยนแปลงของข้อมูลด้วย เพื่อให้มีความสอดคล้องกับรูปแบบของข้อมูลนั้นๆด้วย และ เงื่อนไขบางตัวควรมีความสามารถที่จะคงอยู่ได้ตลอดเวลา หากเงื่อนไขนั้นๆมีความสำคัญที่เพียงพอเพื่อช่วยปรับปรุงระบบให้มีประสิทธิภาพมากขึ้น

5. การนำเงื่อนไขในระดับจุดข้อมูลมาใช้ในงานวิจัยนี้ โอกาศที่เงื่อนไขถูกนำมาใช้ในแอพพลิเคชั่นจริงยังค่อนข้างเป็นไปได้ยาก เนื่องจากเงื่อนไขในระดับนี้มีความละเอียดมากเกินไป ซึ่งผู้เชี่ยวชาญอาจไม่สามารถให้ความรู้เพื่อมาช่วยเหลือได้ในระดับนี้ ซึ่งหากมีการพัฒนาระดับของเงื่อนไขให้ละเอียดน้อยลง หรือ เปลี่ยนจากระดับจุดข้อมูลเป็นระดับกฎ (Rule-based Level Constraints) อาจทำให้สามารถนำความรู้จากผู้เชี่ยวชาญมาประยุกต์ใช้ได้ง่ายและเหมาะสมกับแอพพลิเคชั่นจริงมากขึ้นด้วยเช่นกัน

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

เอกสารและสิ่งอ้างอิง

Aggarwal, C., J. Han and P.S. Yu. 2004. A Framework for Projected Clustering of High Dimensional Data Streams, pp. 852-863. *In* **Proceeding of the 30th International Conference on Very Large Database (VLDB 2004).** 29 August -3 September 2004. Toronto, Canada.

Bilenko, M., S. Basu and R.J. Mooney. 2004. Active semi-supervision for pairwise constrained clustering, pp. 333-344. *In* **Proceedings of the 2004 SIAM international conference on data mining (SDM-04).** 22-24 April 2004. Florida, USA.

Bradley, P. S., K. P. Bennett and A. Demiriz. 2000. **Constrained k-means clustering. Technical Report**, MSR-TR-2000-65, Microsoft Research, Redmond, WA

C++ Reference - Sorting. 2013. **C++ Sourcecode Reference**. Available Source: http://www.cplusplus.com/reference/algorithm/sort/. January, 2013.

Cao, F., M. Ester, W. Qian and A. Zhou. 2006. Density-based clustering over an evolving data stream with noise, pp. 328-339. *In* Proceedings of the 12th **SIAM International Conference on Data Mining.** 26-28 April 2012. California, USA.

Codd, E.F. 1970. A Relational Model of Data for Large Shared Data Banks. *In* **Communications of the ACM 13.** Vol. 12 Issue 6: 377-387.

Halkidi, M., M. Spiliopoulou and A. Pavlou. 2012. A Semi-supervised Incremental Clustering Algorithm for Streaming Data, pp.578-590. *In* **Proceedings of the 16th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD'12).** 29 May – 1 June 2012. Kuala Lumpur, Malaysia.

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

He, Z., X. Xu, S. Deng and J.Z. Huang. 2004. Clustering Categorical Data Streams. *In* **Journal of Computational Methods on Science and Engineering** 11(4): 185-192.

Jain, A.K., M. N. Murty and P. J. Flynn. 1999. Data Clustering: A Review. *In* **ACM Computing Surveys**, vol. 31, no. 3: 264-323.

Udommanetanakit, K., T. Rakthanmanon and K. Waiyamai. 2007. E-stream: Evolution-based technique for stream clustering, pp. 605-615. I*n* **Proceedings of the 3rd International Conference of Advanced Data Mining and Applications (ADMA 2007)**. 6-8 August 2007. Harbin, China.

Milenova, B.L. and M.M. Campos. 2003. **Clustering Large Databases with Numeric and Nominal Valuees Using Orthogonal Proections**, Oracle research. USA.

Ruiz, C., M. Spiliopoulou and E. Menasalvas. 2006. User Constraints Over Data Streams. *In* **Proceedings of Workshop Notes of IWKDDS Workshop (PKDD'06).** 18-22 September 2006. Berlin, Germany.

Ruiz, C., M. Spiliopoulou and E. Menasalvas. 2009. C-DenStream: Using Domain Knowledge on a Data Stream, pp. 287-301. *In* **Proceedings of the 12th International Conference Discovery Science (DS 2009).** 3-5 October 2009. Porto, Portugal**.**

Sirampuj, T., T. Kangkachit and K. Waiyamai. 2013. CE-Stream: Evaluation-Based Technique for Stream Clustering with Constraints. *In* **Proceedings of the 10th International Joint Conference on Computer Science and Software Engineering (JCSSE'13).** 29-31 May 2013**.** Khon Kean**,** Thailand.

Tague, N.R. 2004. **The Quality Toolbox, Second Edition**, ASQ Quality Press. USA.

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

Wagstaff, K., C. Cardie, S. Rogers and S. Schroedl. 2001. Constrained K-means clustering with Background Knowledge, pp.577-584. *In* **Proceedings of the 18th International Conference on Machine Learning (ICML).** MA, USA.

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

ประวัติการศึกษา และการทำงาน

ชื่อ –นามสกุล	นายทศพร ศิรามพุช
วัน เดือน ปี ที่เกิด	วันที่ 8 สิงหาคม 2528
สถานที่เกิด	กรุงเทพมหานคร
ประวัติการศึกษา	วศ.บ. (วิศวกรรมซอฟแวร์และความรู้) คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์
ตำแหน่งหน้าที่การงานปัจจุบัน	IT Technologist
สถานที่ทำงานปัจจุบัน	บริษัท เชฟรอนเอเชียเซ้าท์ จำกัด
ผลงานดีเด่นและรางวัลทางวิชาการ	-
ทุนการศึกษาที่ได้รับ	-

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์