مجلة جامعة دمشق للعلوم الأساسية
مجلد 41-عدد 2 (2025): 17- 30

Damascus University Journal for the Basic Sciences
Vol 41 No.2 (2025): 17 -30

دراسة مرجعية لخوارزميات تجميع تدفق البيانات

براءة غسان الحمد[1*]، سمير نزيه جعفر[2]

[1] طالبة ماجستير، جامعة دمشق، كلية العلوم، قسم الرياضيات،
baraa.hmd@damascusuniversity.edu.sy

[2] أستاذ مساعد، جامعة دمشق، كلية العلوم، قسم الرياضيات(تصميم نظم وبرمجيات)،
samir.jafar@damascusuniversity.edu.sy

الملخص:

يعد تنقيب تدفق البيانات مجالاً بحثياً نشطاً لأنه يقوم باكتشاف المعرفة من كميات كبيرة من البيانات التي يتم إنشاؤها باستمرار وجمعها في الوقت الفعلي. يشتمل التّعلم غير الخاضع للإشراف على أحد أكثر مهام التنقيب عن البيانات شيوعاً وهو التجميع.

نقدم هذا البحث لتوضيح أهم المفاهيم والخصائص الأساسية المشتركة لخوارزميات تجميع تدفق البيانات مثل تغير المفهوم وهياكل البيانات والنوافذ الزمنية وطرق معالجة البيانات بالإضافة لبعض التحديات التي تواجه الخوارزميات مثل التعامل مع البيانات الشاذة والمتطورة والذاكرة المحدودة والوقت المحدود ومعالجة البيانات متعددة الأبعاد وعدد المجموعات، كما قمنا بأخذ عينة من الخوارزميات لتجميع تدفق البيانات وعرض المفاهيم والتحديات الموضحة ضمن البحث باستخدام هذه العينة وذلك بطريقة إحصائية غرافيكية. لتوضيح ومقارنة المعايير المستخدمة في خوارزميات تجميع تدفق البيانات.

الكلمات المفتاحية: تدفقات البيانات، تجميع تدفق البيانات، التجميع في الوقت الفعلي.

تاريخ الإيداع: 2023/10/09
تاريخ الموافقة: 2023/12/03

ISSN (online): 2789-6366
http://journal.damascusuniversity.edu.sy/index.php/index/index

Survey of Data Stream Clustering Algorithms

Bara'a Ghassan Al-hamad[1*], Samir Nazih Jafar[2]

[1] Master's student, Department of Mathematics, Faculty of Sciences, Damascus University, Syria, baraa.hmd@damascusuniversity.edu.sy.
[2] Professor, Department of Mathematics, Faculty of Sciences, Damascus University, Syria, samir.jafar@damascusuniversity.edu.sy.

Abstract

Data stream mining is an active research field as it discovers knowledgefrom large amounts of data that are constantly being created and collected in real-time. Unsupervised learning is one of the most common tasks in
data stream mining, which is clustering. In this research, we present the main concepts and common characteristics of data stream clustering algorithms, such as concept drift, data structures, time windows, and data processing methods. We also discuss some challenges faced by thesealgorithms, such as handling outliers, evolving data, limited memory andtime, and processing multi-dimensional and multi-group data.
Additionally, we provide a sample of data stream clustering algorithms and illustrate the concepts and challenges discussed in this research using statistical graphics. This is done to clarify and compare the criteria used in data stream clustering algorithms.

Keywords: Data Streams, Data Stream Clustering, Real-Time Clustering.

Received :2023/10/09
Accepted:2023/12/03

1. المقدمة (Introduction):

يولّد تزايد كمية البيانات، وتنوّعها، وسرعتها مفهوم البيانات الضخمة. فهي مجموعة من البيانات التي تستعصي لضخامتها وتعقيدها على التخزين أو المعالجة بإحدى الأدوات المعتادة لإدارة العمليات على البيانات.

يمكن تخزين ومعالجة البيانات العادية باستخدام الأدوات أو التطبيقات المعتادة لإدارة البيانات مثل نظام إدارة قواعد البيانات العلائقية (Relational Database Management System (RDBMS)). وبالمقابل، إن البيانات الضخمة تتطلب أدوات وتقنيات مختلفة لتحليلها واستخراج المعلومات منها مثل(Hadoop)، وهو ما يعد تحدياً للمحللين وعلماء البيانات. وتتميز البيانات الضخمة بتنوعها وسرعتها وحجمها.

تواجه تدفقات البيانات الضخمة العديد من التحديات، ومن أهمها:

1. صعوبة تخزين البيانات الضخمة ومعالجتها باستخدام أنظمة إدارة قواعد البيانات العلائقية، حيث تتطلب تخزيناً كبيراً ومكلفاً، ويمكن أن تتطلب استخدام تقنيات تخزين جديدة ومتطورة لإدارة هذه الكميات الهائلة من البيانات.
2. الحاجة إلى استخدام برامج متوازية واسعة النطاق تعمل على عشرات أو مئات أو حتى آلاف من الخوادم.
3. صعوبة تحليل البيانات الضخمة واستخراج المعلومات المفيدة منها.
4. تحديات الأمان والخصوصية والأخلاقيات المتعلقة باستخدام البيانات الضخمة.
5. صعوبة تحديد الأدوات والتقنيات المناسبة لمعالجة البيانات الضخمة.

إن مفهوم تدفق البيانات هو تسلسل مرتب من النقاط $x_1, \ldots, x_n$ التي يجب الوصول إليها بالترتيب ويمكن قراءتها مرة واحدة فقط أو عدد قليل من المرات. كل قراءة لهذا التسلسل تسمى مسح خطي. يتم تحفيز مفهوم التدفق من خلال التطبيقات الحديثة التي تتضمن مجموعات من البيانات الضخمة، على سبيل المثال بيانات الموقع الجغرافي وبيانات أجهزة الاستشعار وبيانات مشاركات العملاء حول منتج معين في مواقع التواصل الاجتماعي والتي تستخدمها الشركات لإجراء تحليلات عليها لمعرفة مدى توافق الآراء حول منتجاتها.

إن هذه المجموعات من البيانات كبيرة جدًا بحيث لا يمكن وضعها في الذاكرة الرئيسية ويتم تخزينها عادةً في أجهزة تخزين ثانوية.

ومن أجل تجاوز هذه الصعوبات، يمكن استخدام تقنيات وأدوات مختلفة مثل تقنيات التخزين السحابي وتقنيات التعلم الآلي والذكاء الاصطناعي وغيرها، هنالك بعض الطرق لتجاوز هذه صعوبات التعامل مع هذا الكم الهائل من البيانات تشمل خوارزميات التجميع. قدمنا في هذا البحث توضيح لأساسيات ومفاهيم خوارزميات تجميع تدفق البيانات وكذلك على بعض التحديات. وقمنا بأخذ عينة عشوائية قدرها 24 خوارزمية تجميع التدفق وعرض المفاهيم الأساسية والتحديات بطريقة إحصائية بالاعتماد على الخوارزميات.

2. فكرة البحث وأهدفه (Idea and objectives):

إن التطور الحاصل في عصرنا الحالي يولد كميات كبيرة من البيانات التي تتزايد باستمرار مثل تحليل بيانات المستشعر وانترنت الأشياء وغيرها. الامر الذي يتطلب فهم كيفية التعامل مع هذا الكم الهائل من البيانات. يتم معالجة هذه البيانات باستخدام طرق مختلفة مثل التصنيف أو التجميع أو تحديد أنماط منها وغيرها من الطرق. من بين الطرق المهمة للعمل على البيانات هي التجميع بحيث يتم تجميع مجموعة البيانات ضمن تجمعات بحيث تمكننا من الحصول على معلومات مهمة من هذه التجمعات. يدفع ذلك للعمل على توضيح الطرق والمفاهيم المستخدمة للتجميع بهدف معرفة أحسن طريقة واختيار أفضل الشروط حسب طبيعة التطبيقات المستخدمة لتحسين وتوضيح هذه التجمعات وجعلها تُفهم بشكل أكثر وتُعطي رؤية أوضح.

3. مواد وطرق البحث(Materials and Methods):

3-1 المفاهيم الأساسية في تجميع تدفق البيانات (Basic concepts in data stream clustering):

التجميع هي مهمة رئيسية من مهام تنقيب البيانات، والتي تعني تصنيف مجموعة بيانات معينة إلى مجموعات(تجمعات أصغر من المجموعة الأساسية) بحيث تكون نقاط البيانات في المجموعة أكثر تشابهاً مع بعضها البعض عكس النقاط الموجودة في مجموعات مختلفة. إن تجميع تدفق البيانات يكون بعكس تجميع نقاط البيانات الثابتة لأنه يطرح العديد من التحديات الجديدة، وذلك لأن تدفقات البيانات تكون بشكل مستمر وكمية البيانات تكون غير محدودة. لذلك يكون من المستحيل الاحتفاظ بكامل دفق البيانات في الذاكرة الرئيسية.

من أجل المعالجة الفعالة والسريعة لتدفقات البيانات المستمرة نتعامل مع دفق البيانات مرة واحدة فقط، فإن عمليات الفحص والمعالجة المتعددة لهذه البيانات غير عملية أو ربما تكون غير ممكنة (أي المرور عدة مرات على نفس مجموعات البيانات). وذلك لأن دفق البيانات يتطلب معالجة سريعة في الوقت الفعلي لوصول البيانات بحيث من المتوقع أن تكون نتائج التنقيب متاحة بغضون وقت استجابة قصير. (Amini, 2014, p. 11)

هناك بعض المفاهيم المتعلقة بتجميع البيانات مثل تلخيص البيانات أو طرق معالجة البيانات وغيرها من الطرق التي سنقدم وصف موجز لها:

3-1-1- **تغير المفهوم (Concept Drift)**: (Zubaroğlu *et al.*, 2020,4) يشير إلى تغير في الخصائص الإحصائية غير المتوقع بمرور الوقت، وهو مشكلة شائعة في تجميع تدفق البيانات. بعض حالات تغير المفهوم هي:

3-1-1-1-**مفاجئ(Sudden)**: يتم بين نقطتي بيانات متتاليتين، يحدث التغيير في وقت واحد ويتم استلام نقاط بيانات من الفئة الجديدة. بحيث لا تصل نقطة بيانات لها خصائص الفئة السابقة مرة أخرى.

$$S = \{\dots, x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, \dots\}$$

3-1-1-2-**تدريجي(Gradual)**: أي عدد نقاط البيانات التي تنتمي إلى الفئة السابقة ينخفض تدريجياً بينما يزداد عدد نقاط البيانات التي تنتمي إلى الفئة الجديدة بمرور الوقت.

$$S = \{\dots, x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, \dots\}$$

3-1-1-3-**متزايد(Incremental)**: الزيادة التدريجية لعدد نقاط البيانات التي تنتمي إلى فئة معينة حيث تتحول هذه الحالات تدريجياً إلى الفئة الجديدة. بعد اكتمال تغير المفهوم تختفي الفئة السابقة، ونقاط البيانات التي تصل خلال تغيير المفهوم هي من الاشكال الانتقالية ولا تنتمي إلى أي من الفئات.

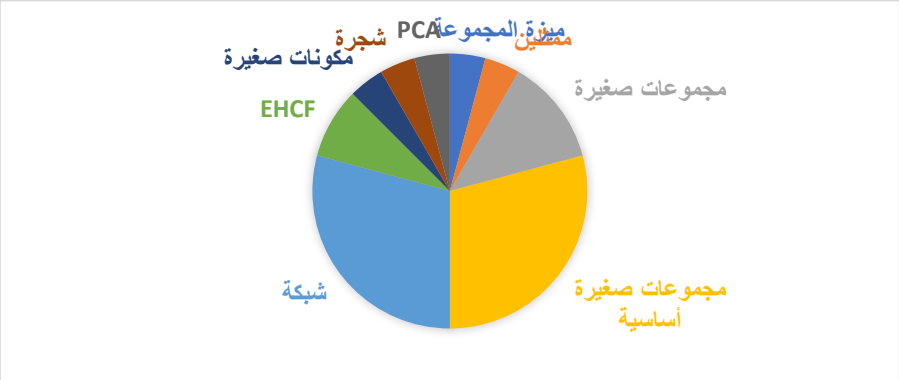
$$S = \{\dots, x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, \dots\}$$

3-1-1-4-**متكرر(Recurring)**: تتغير نقاط البيانات بين خصائص إحصائية مختلفة عدة مرات. لا تختفي أي من الفئات بشكل دائم ولكن كلاها تأتي بالتناوب.

$$S = \{\dots, x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, \dots\}$$

3-1-2- **هياكل البيانات (Data structures)**: (Zubaroğlu *et al.*, 2020,4) تستخدم لتخزين وتنظيم البيانات في تجميع تدفق البيانات، كما أنها مناسبة لتخزين الملخصات الإحصائية لتدفقات البيانات وذلك لأنه لا يمكن تخزين الدفق بشكل كامل في الذاكرة الرئيسية لذلك تستخدم هياكل البيانات لتلخيص الدفق بشكل متزايد. نستعرض بعض الملخصات الأكثر استخداماً وهي ميزة المجموعة (Cluster Feature) والمجموعات الصغيرة(Micro-Clusters) والمجموعات الصغيرة الأساسية(Core-Micro-Clusters) وميزة المجموعة الزمنية (Temporal CF) والشبكات(Grids) وأشجار المجموعة الأساسية(Coreset Trees) والممثلين

(Representatives) والرسم البياني الأسي لميزة المجموعة (EHCF)(Exponential Histogram of Cluster Feature) ومتجهة الميزات (Feature Vector) وتحليل المكون الرئيسي (PCA) (Principal Component Analysis) ومكونات صغيرة (Micro-components) ومصفوفة النماذج (Prototype Array).

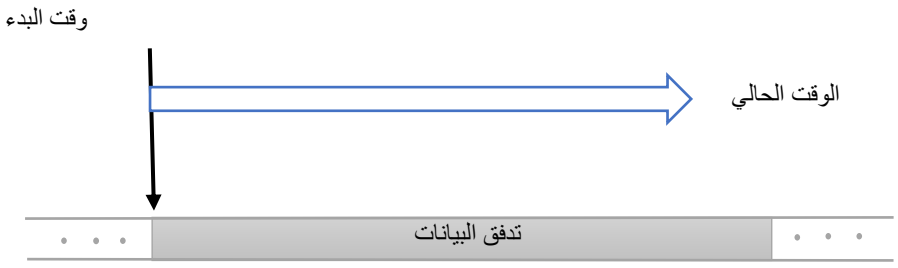


الشكل (1): مخطط يمثل هياكل البيانات لعينة من خوارزميات تجميع تدفق البيانات

3-1-3- النوافذ الزمنية (Time Window Models): (Rastin, 2018, p. 31) تعد النوافذ الزمنية نمط لإجراء عمليات على البيانات ضمن إطار زمني محدد، نظراً لأن تدفقات البيانات تتطور باستمرار مع مرور الوقت وقد تكون غير محدودة. وبالتالي قد تتغير نتائج التجميع بمرور الوقت. تعتبر تدفقات البيانات عملية متطورة بمرور الوقت وتتم معالجتها باستخدام هذا النوافذ، حيث من الممكن أن تعكس المعلومات الحديثة في دفق البيانات تغيرات في توزيع التجمعات المستخرجة، يمكن استخدام هذه المعلومات لشرح تطور العملية. حيث يتم فصل البيانات إلى عدة نوافذ أساسية، أنواع نماذج النوافذ هي كالآتي:

3-1-3-1- النافذة الزمنية المميزة(Landmark window):

يتم تحديد النافذة بنقطة زمنية محددة تسمى نقطة المميزة إلى وقت وصول البيانات الحالي، حيث يتم معالجة أجزاء منفصلة من التدفقات و تكون مفصولة بنقاط مميزة. يتم استخدامها لتجميع كل تدفق البيانات كما يمكن تحديد النقاط المميزة من حيث الوقت (على أساس يومي أو أسبوعي) أو من حيث عدد العناصر. يتم الاحتفاظ بجميع نقاط البيانات التي وصلت بعد النقطة المميزة أو تلخيصها في نافذة البيانات الحديث. عند الوصول إلى نقطة مميزة جديدة، تتم إزالة جميع نقاط البيانات المحفوظة في النافذة ويتم الاحتفاظ بالنقاط الجديدة من النقطة المميزة الحالية في النافذة حتى يتم الوصول إلى نقطة مميزة جديدة. نلاحظ أنه في هذا النوع من نماذج النوافذ، لا يتم النظر في العلاقة بين البيانات في النوافذ المجاورة. تكمن المشكلة في استخدام أي مخطط نافذة بطول محدد



الشكل (2): مثال على النافذة الزمنية المميزة (Rastin, 2018, p. 33)

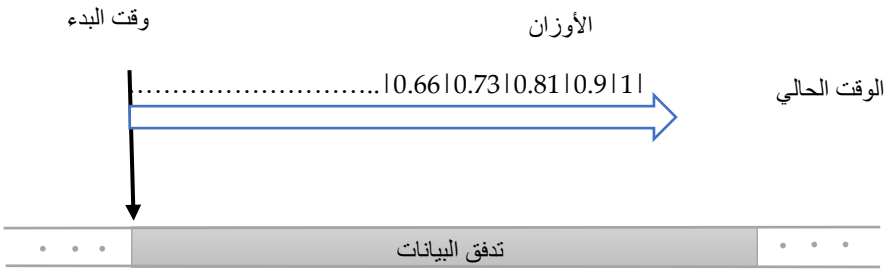
في تحديد حجم النافذة المثالي المطلوب استخدامه. تضمن النافذة الصغيرة أن خوارزميات تدفق البيانات ستكون قادرة على التقاط التغيرات النهائية للمفهوم بسرعة. في الوقت نفسه، في المراحل المستقرة على طول الدفق، قد يؤثر ذلك على أداء خوارزمية التّعلم. من ناحية أخرى، تكون النافذة الكبيرة مرغوبة في المراحل المستقرة، على الرغم من أنها قد لا تستجيب بسرعة لتغيرات المفهوم.

3-1-3-2- النافذة الزمنية المنزلقة (Sliding window):

الشكل (3): مثال على النافذة الزمنية المنزلقة (Rastin, 2018, p. 33)

يتم تخزين أحدث المعلومات فقط من دفق البيانات في بنية بيانات يمكن أن يكون حجمها متغيرًا أو ثابتًا. عادةً ما تكون بنية البيانات هذه عبارة عن طابور(قائمة انتظار)، والذي يأخذ في الاعتبار البيانات من الفترة الزمنية الحالية حتى فترة معينة في الماضي. يعتمد تنظيم العناصر ومعالجتها على مبدأ الطابور. تقوم الخوارزميات التي تعتمد على هذه النوافذ فقط بتحديث الملخصات الإحصائية للبيانات المدرجة في النافذة.

3-1-3-3- النافذة الزمنية المخمدة(Fading window):



عندما $(\lambda = 0.9)$

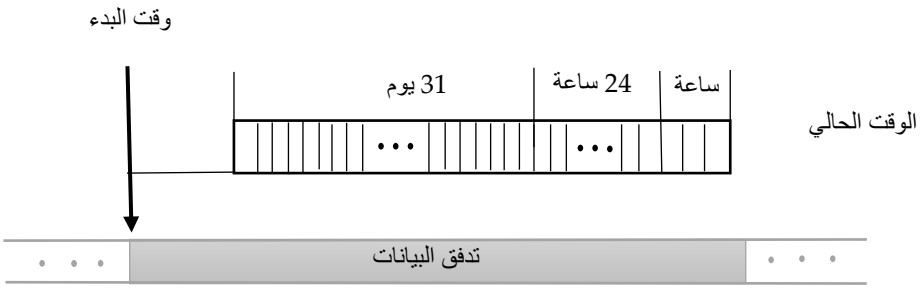
الشكل (4): مثال على النافذة الزمنية المخمدة (Rastin, 2018, p. 33)

إن هذه النافذة تأخذ في الاعتبار أحدث البيانات من خلال ربط الأوزان بكل نقطة بيانات في دفق البيانات، ويتم إعطاء أوزان أكثر للبيانات الحديثة مقارنة بالبيانات القديمة. ويقل وزن البيانات بمرور الوقت، يكون استخدام هذه النافذة لتقليل تأثير البيانات القديمة على نتيجة التجميع، يتم اعتماد هذا النافذة في خوارزميات التجميع المعتمدة على طريقة تجميع الكثافة، نُعطي أوزان لنقاط البيانات باستخدام دالة الأوزان (التضاؤل):

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

حيث λ هو معدل الانحلال و t وقت وصول نقطة البيانات.

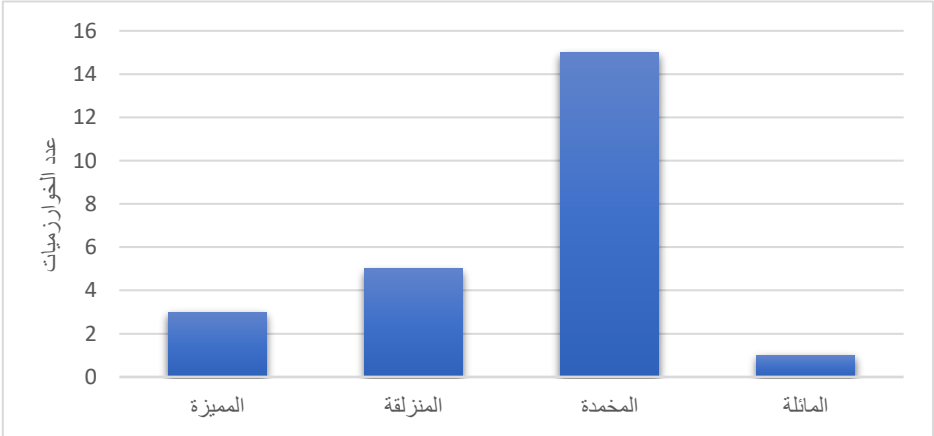
3-1-3-4- النافذة الزمنية المائلة(Tilted-time window):



الشكل (5): مثال على النافذة الزمنية المائلة (Rastin, 2018, p. 33)

تستخدم هذه النافذة مستويات مختلفة لتلخيص نقاط البيانات وذلك بناءً على حداثة البيانات، يقوم بتلخيص البيانات الحديثة بشكل دقيق بينما يتم تجميع البيانات القديمة تدريجياً.

يعتمد اختيار النوافذ الزمنية على احتياجات ومتطلبات التطبيق.



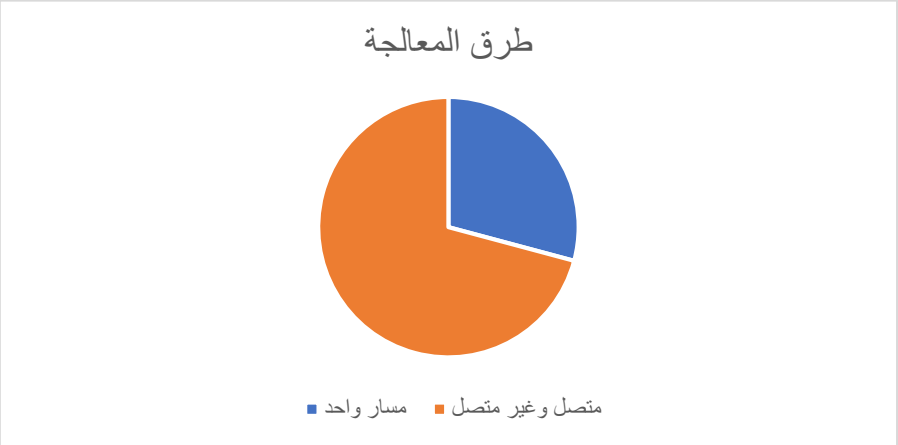
الشكل (6): مخطط يمثل النوافذ الزمنية لعينة من خوارزميات تجميع تدفق البيانات

3-1-4-طرق المعالجة:

لمعالجة تدفقات البيانات المستمرة يكون من الصعب المرور(مسح) كل نقاط البيانات لعدة مرات وذلك بغرض تجميعها ضمن مجموعات لذلك نستخدم طرق المعالجة التالية:

3-1-4-1- **مسار واحد:**(Rastin, 2018, p. 30) يتم تجميع تدفقات البيانات عن طريق مسح تدفقات البيانات مرة واحدة فقط.

3-1-4-2- **متصل وغير متصل:**(Carnein *et al.*, 2019,279) في بعض الأحيان تحتاج خوارزمية تجميع تدفق البيانات إلى التحقق من المجموعات من خلال تقسيم الدفق إلى أجزاء. يتم استخدام نموذج نافذة مختلف لتتبع تطور سلوك تدفقات البيانات. ومع ذلك، لا يمكننا إجراء تجميع ديناميكي عبر جميع الآفاق الزمنية الممكنة لتدفقات البيانات. لذلك، تم تقديم نهج متصل عبر الإنترنت وغير متصل. يحتفظ المكون المتصل عبر الإنترنت بمعلومات موجزة (تلخيصات للبيانات) حول تدفقات البيانات المستمرة والنتيجة هي عدد من المجموعات الصغيرة مما يوفر فهماً للمجموعات. والمكون غير المتصل، يقوم بإعادة التجميع للمجموعات الصغيرة لإيجاد مجموعات نهائية من المجموعات الكبيرة.



الشكل (7): مخطط يمثل طرق معالجة الخوارزميات لعينة من خوارزميات تجميع تدفق البيانات

3-2 التحديات في تجميع تدفق البيانات (Challenges in clustering data streams):

المشكلة المهمة التي تتعلق بتجميع تدفق البيانات هي كيفية معالجة البيانات اللانهائية التي تتدفق باستمرار بمرور الوقت أو كيفية الاحتفاظ بهذه الكمية الضخمة من البيانات من أجل المعالجة اللاحقة.

بملاحظة سلوك البيانات الديناميكي يجب أن يعالج تجميع تدفق البيانات التحديات التالية:

3-2-1- **التعامل مع البيانات الشاذة:** يجب أن تكون أي خوارزمية تجميع قادرة على التعامل مع النقاط الشاذة العشوائية الموجودة في البيانات حيث أن النقاط الشاذة لها تأثير كبير على تكوين المجموعات.

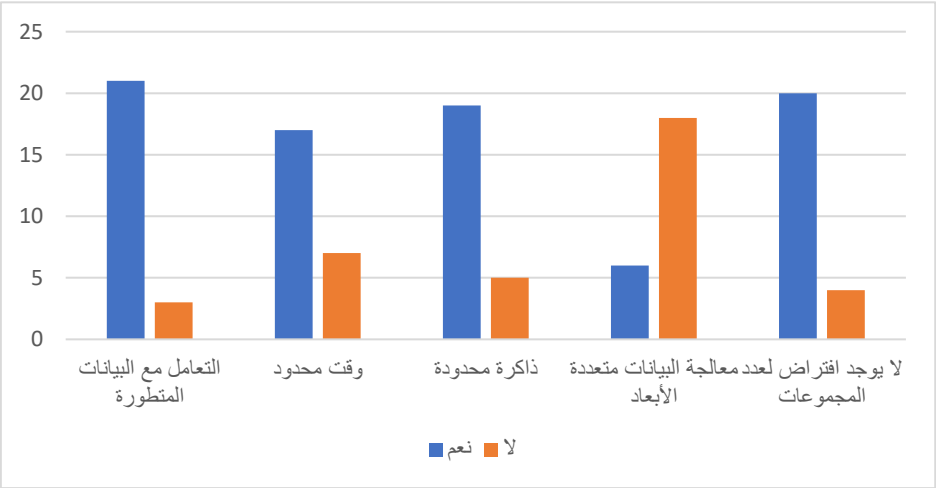
3-2-2- **التعامل مع البيانات المتطورة:** يجب أن تأخذ الخوارزمية في الاعتبار أن تدفقات البيانات تتطور بشكل كبير بمرور الوقت.

3-2-3- **وقت محدود:** تصل تدفقات البيانات باستمرار، الأمر الذي يتطلب استجابة سريعة في الوقت الحقيقي. لذلك، تحتاج خوارزمية التجميع إلى التعامل مع سرعة تدفقات البيانات في الوقت المحدود.

3-2-4- **ذاكرة محدودة:** يتم إنشاء كمية هائلة من تدفقات البيانات بسرعة، الأمر الذي يحتاج إلى ذاكرة غير محدودة. ومع ذلك، يجب أن تعمل خوارزمية التجميع ضمن قيود الذاكرة.

3-2-5- **معالجة البيانات متعددة الأبعاد:** بعض تدفقات البيانات ذات أبعاد متعددة في طبيعتها مثل تجميع المستندات النصية. لذلك، يجب على خوارزمية التجميع التغلب على هذا التحدي في حالة كون بياناتها متعددة الأبعاد.

3-2-6- **عدد المجموعات:** وهو تحديد عدد المجموعات مسبقاً قبل عملية التجميع.



الشكل (8): مخطط يمثل التحديات لعينة من خوارزميات تجميع تدفق البيانات

4. تصنيف خوارزميات تجميع تدفق البيانات (Classification of data stream clustering algorithms):

يوجد عدد كبير من خوارزميات التجميع لمجموعات البيانات الثابتة مثل (K-means, DBSCAN) حيث تم تمديد بعضها لتناسب تدفقات البيانات. بشكل عام يتم تصنيف طرق التجميع إلى خمس فئات رئيسية كما يلي: (Rastin, 2018, p. 34)

نهج التقسيم(Partitioning Approaches): تنظم البيانات في عدد محدد من الأقسام حيث يُمثّل كل قسم بمجموعة، تتشكل هذه الأقسام بناءً على دالة المسافة مما يؤدي إلى تشكل مجموعات كروية وتتأثر النتائج بالنقاط الشاذة (الخارجية)، كما أنه

يجب تحديد عدد المجموعات مسبقاً في هذه الخوارزميات وتتميز بسهولة التنفيذ بشكل عام. من الأمثلة على هذا النوع من الخوارزميات هي: STREAM, cluStream.

مثال: إذا كان لدينا تدفق بيانات من بيانات الطقس، يمكننا تقسيم هذا التدفق إلى أجزاء أصغر حسب المنطقة الجغرافية ثم استخدام الخوارزمية لتجميع كل جزء على حدة.

نهج الهرمي (Hierarchical Approaches): تستخدم هذه الخوارزميات بنية بيانات (Dendrogram) أي أنها تقوم بتجميع البيانات المُعطاة في شجرة من المجموعات، ينقسم هذا النوع من الخوارزميات إلى قسمين: خوارزميات تقسيمية وخوارزميات تكتلية، بمجرد الانتهاء من هذه الخطوة لا يمكن التراجع عنها أبداً. من الأمثلة على هذا النوع من الخوارزميات هي: BIRCH.

مثال: لنفترض أن لدينا تدفق بيانات من بيانات بعض صفحات الويب. يمكننا بناء شجرة من هذا التدفق بناءً على روابط الويب. واستخدام خوارزميات التجميع لتجميع العقد في الشجرة في مجموعات.

نهج القائم على الكثافة (Density-based Approaches): تتشكل المجموعات كمجموعات كثيفة منفصلة عن مجموعات متفرقة. الفكرة الرئيسية هي أن مجموعة معينة تزداد بها البيانات باستمرار طالما أن عدد البيانات في التّجمع لا يتجاوز عدد محدد مسبقاً يسمى العتبة. يمكن استخدام هذه الطريقة لمعرفة الضوضاء أو النقاط الشاذة واكتشاف مجموعات ذات شكل عشوائي. من الأمثلة على هذا النوع من الخوارزميات هي: DenStream.

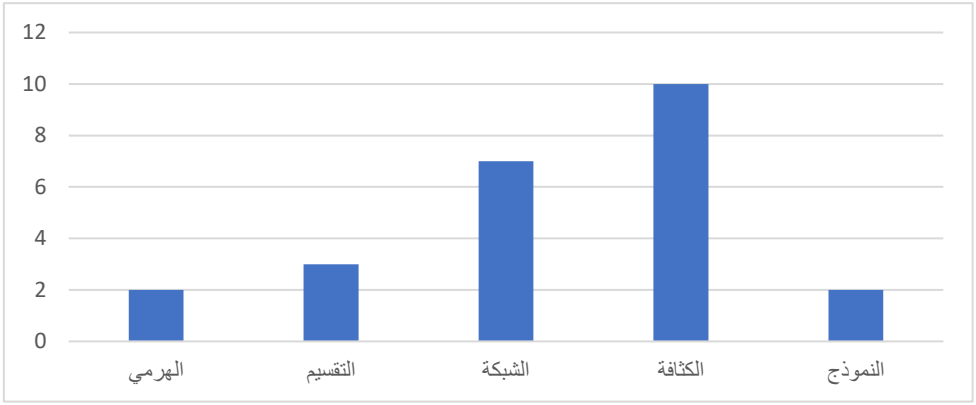
مثال: بفرض أن لدينا تدفق بيانات من بيانات الموقع الجغرافي، يمكننا استخدام طريقة الكثافة لتجميع المجموعات الصغيرة التي تقع في نفس المنطقة في ضمن مجموعات أكبر.

نهج القائم على الشبكة (Grid-based Approaches): يقوم بتقسيم مساحة البيانات إلى عدد من الخلايا التي تشكل الشبكة. يتميز بوقت معالجة سريع لأنه لا يعتمد على عدد البيانات. من الأمثلة على هذا النوع من الخوارزميات هي:D-Stream,DUCStream .

مثال: إذا كان لدينا تدفق بيانات من بيانات الشبكات الاجتماعية. يمكننا إنشاء شبكة من المجموعات بناءً على العلاقات بين المستخدمين. ثم يمكننا استخدام خوارزمية التجميع لتجميع المجموعات في الشبكة.

نهج القائم على النموذج (Model-based Approaches): تحاول تحسين التوافق بين البيانات المقدمة وبعض النماذج الرياضية. من الأمثلة على هذا النوع من الخوارزميات هي: SWEM.

مثال: لدينا تدفق بيانات من بيانات التسويق، يمكن استخدام طريقة النموذج لتعلم خصائص سلوك المستهلك. ثم استخدام النموذج لتجميع العملاء في مجموعات بناءً على تشابه سلوكهم حسب التوزيع الطبيعي.



الشكل (9): مخطط يمثل تصنيف خوارزميات التجميع لعينة من خوارزميات تجميع تدفق البيانات

5. النتائج والمناقشة (Results and Discussion):

تجميع تدفق البيانات هو نوع من التعلم الآلي الذي يتعامل مع البيانات التي يتم إنشاؤها و معالجتها بشكل مستمر. تختلف خوارزميات تجميع التدفق عن خوارزميات التعلم الآلي التقليدية في أنها مصممة للتعامل مع تدفق من البيانات، بدلاً من التعامل مع مجموعة بيانات ثابتة. هذا يعني أنها يجب أن تكون قادرة على التعامل مع البيانات التي تتغير بمرور الوقت. إن كل مسح لمجموعة كبيرة من البيانات يكون صعباً، وبالتالي فإن المعايير التي يتم من خلالها الحكم على أداء خوارزمية التدفق تشمل:

1. مدخلات الخوارزمية وهي المعلومات الأولية التي تؤخذ لعمل الخوارزمية أي أنها الخطوة الاولى والأهم عند العمل على الخوارزمية.
2. تصنيف الخوارزميات ممثلة بطريقة التجميع الأساسية بإحدى الفئات الخمس وتمثل النهج المستخدم أثناء العمل.
3. النوافذ الزمنية نقوم باختيار أحد النوافذ ويتم من خلال هذا المعيار تحديد أهمية نقاط دفق البيانات حيث أن بعضها يتم العمل عليه وبعضها الآخر يتم تجاهلها أو تكون أقل أهمية بحسب حداثة البيانات.
4. طريقة معالجة البيانات المتدفقة باختيار أحد الطرق، تأخذ الخوارزمية سلوك محدد للتعامل مع البيانات إما مسح واحد لنقاط البيانات أو اختيار مكون متصل وغير متصل بالإنترنت حيث أنه باستخدام أحد طرق المعالجة السابقة يتم التعامل مع البيانات بالخطوتين التاليتين:

4.1. تلخيص البيانات وتخزينها ضمن مجموعات صغيرة بطرق محددة ليتم تجميعها ضمن مجموعات تكون أكبر من المجموعات الصغيرة وسيتم تجاهل نقاط تدفق البيانات لأنه لا يمكن تخزينها في الذاكرة الرئيسية ويمكن الاحتفاظ فقط بالملخصات.

4.2. تقنية التجميع الأساسية كطريقة أساسية لتجميع الملخصات ضمن مجموعات كبيرة.

إن هذه المعايير تمثل خطوات سير الخوارزمية، ومن المعايير التي تعتبر ذات أهمية أيضاً وهي تحديات الخوارزميات.

5. التعامل مع البيانات المتطورة.

6. معالجة البيانات متعددة الأبعاد لأنه كلما زاد عدد الابعاد زاد احتمال تجميع البيانات بشكل أحسن لأن رؤية البيانات تصبح أفضل.

7. المعالجة ضمن وقت محدود.

8. المعالجة ضمن ذاكرة محدود، لأن حجم مجموعات البيانات يتجاوز بكثير مقدار الذاكرة الرئيسية المتاحة للخوارزمية فلا يمكن تذكر الكثير من نقاط البيانات التي تم مسحها سابقاً استلزم ذلك تطوير عدد من الخوارزميات لتقوم بتخزين فقط ملخصات للبيانات السابقة، لذلك سيوفر ذاكرة كافية لمعالجة البيانات المستقبلية.

6. الاستنتاجات والتوصيات (Discussion and Further Work):

التزايد الكبير لاستخدام أجهزة التكنولوجيا الحديثة يتزايد معها عدد الأجهزة المترابطة. حيث تقوم الأجهزة باستمرار بإنشاء بيانات وهذه البيانات تكون كبيرة الحجم وتتشكل بسرعة عالية والتي تسمى تدفقات البيانات. لذلك، فإن معالجة هذه تدفقات البيانات ضمن وقت مقبول لتنفيذ التطبيق تثير الاهتمام ويبدو أن التجميع هو أنسب طريقة لمعالجة تدفقات البيانات. في هذا البحث قمنا بتوضيح أهم المفاهيم لتجميع تدفق البيانات مثل تغيير المفهوم والنوافذ الزمنية وهياكل البيانات،وكذلك على بعض التحديات التي تواجه الخوارزميات مثل طرق الكشف عن البيانات الشاذة والوقت والذاكرة المحدودة. كما عرضنا هذه المفاهيم بطريقة إحصائية وذلك بأخذ عينة من خوازميات التجميع وتمثيلها باستخدام المفاهيم الأساسية المستخدمة لهذه الخوارزميات وتحديات الخوارزميات. لذلك نحاول انطلاقاً من هذا البحث تحسين بعض صفات الخوارزميات لجعلها تناسب شروط هذه التدفقات المتطورة وتكون أكثر ملائمة وفاعلية في التطبيقات من خلال تطوير بعض الخوارزميات الهجينة.

المراجع (References):

1. Agrahari, S and Singh, A.K.(2021).Concept Drift Detection in Data Stream Mining : A literature review.
2. Amini, A.,(2014). *An Adaptive Density-Based Method For Clustering Evolving Data Streams*.(Doctoral dissertation, University of Malaya).
3. Amini, A. and Wah, T.,(2011). Density Micro-Clustering Algorithms on Data Streams: A Review.
4. Carnein, M and Trautmann,H.(2019). Optimizing Data Stream Representation: An Extensive Survey on Stream Clustering Algorithms, *Springer Fachmedien Wiesbaden GmbH*.
5. Chen, D. , Du, T. , Zhou, J. , Wu, Y. and Wang, X. (2022). DWDP-Stream: A Dynamic Weight and Density Peaks Clustering Algorithm for Data Stream.
6. Hu, S. , Pang, Y. ,He, Y. ,Yang, Y. , Zhang, H. , Zhang, L. , Zheng, B. , Hu, C. and Wang, Q. (2023). An Enhanced Version of MDDB-GC Algorithm: Multi-Density DBSCAN Based on Grid and Contribution for Data Stream.
7. Krleža, D. ,Vrdoljak, B. and Brčić, M.(2020).Statistical hierarchical clustering algorithm for outlier detection in evolving data streams.
8. Liu, H. , Wu, A. Wei, M. and Chang, C. (2022). SKDStream: a dynamic clustering algorithm on time-decaying data stream.
9. Liu, J. , Peng, Y. and Zhang, D. (2021). Anomaly Detection Based on Multiple Streams Clustering for Train Real-Time Ethernet.
10. Ma, B. , Yang, C. , Li, A. , Chi, Y. and Chen, L. (2023). A Faster DBSCAN Algorithm Based on Self-Adaptive Determination of Parameters.
11. Mansalis, S. , Ntoutsi, E. , Pelekis, N. and Theodoridis,Y. (2018). An evaluation of data stream clustering algorithms, *wileyonlinelibrary*.
12. N, V. and Sakkarapani, K. (2022).Improvisation of DBSCAN Clustering Algorithm Using Various Metrics.
13. Neto, J.M. ,Severiano, C.A. ,Guimaraes, F.G. ,Castro, C.L. ,Lemos, A.P. ,Galindo, J.C. and Cohen, M.W.(2020).Evolving clustering algorithm based on mixture of typicalities for stream data mining.
14. Nordahl, C. ,Boeva, V. ,Grahn, H. and Netz, M.P.(2021). EvolveCluster: an evolutionary clustering algorithm for streaming data.
15. Rastin, p., (2018). *Automatic and Adaptive Learning for Relational Data Stream Clustering*.(Doctoral dissertation, University of Paris).
16. Silva, J. ,Faria, E. , Barros, R. , Hruschka, E. And Gama, J. (2014).Data Stream Clustering: A Survey.
17. Wang, L. and Li, H. (2017).Clustering Algorithm Based on Grid and Density for Data Stream.
18. Wang, X. and Wang, L. (2018).Research on Data Stream Clustering Algorithm Based on Decay Time Window.
19. Wang, Z. ,Ye, Z. , Du, Y. , Mao, Y. , Liu, Y. , Wu, Z. and Wang, J. (2023).AMD-DBSCAN: An Adaptive Multi-density DBSCAN for datasets of extremely variable density.
20. Wang, H. ,Yu, Y. ,Wang, Q. and Wan, Y. (2012).A Density-Based Clustering Structure Mining Algorithm for Data Streams.
21. Yin, L. , Hu, H. , Li, K. ,Zheng, G. ,Qu, Y. and Chen, H. (2023).Improvement of DBSCAN Algorithm Based on K-Dist Graph for Adaptive Determining Parameters.
22. Zubaroğlu, A and Atalay,V.(2020). Data Stream Clustering: A Review, cornell university.