**تشخیص داده‌های پرت در جریان داده با استفاده از خوشه‌بندی**

چکیده :

کاوش در داده‌های جریانی به دنبال الگوریتم‌هایی است که بتوانند به‌صورت واقعی‌زمانه خوشه‌بندی داده‌های پویا و پرت را انجام دهند. مقاله حاضر بررسی عمیقی را در جهت ارائه‌ی روش DenStream برای خوشه‌بندی در داده‌های جریانی ارائه می‌دهد. این روش قابلیت تطبیق با محیط‌های مختلفی از هواشناسی گرفته تا حوزه‌ی مالی و کشاورزی را داراست. با در نظر گرفتن ویژگی‌های داده‌های جریانی مانند پویایی و ناپایداری، DenStream برآیند یک الگوریتم پویا و واقعی‌زمانه برای خوشه‌بندی این داده‌ها است که امکان شناسایی نقاط ناهنجار و پرت را نیز داراست. این الگوریتم در مقایسه با روش‌های سنتی، از جمله مزایایی نظیر عدم نیاز به فرض تعداد خوشه‌ها یا قابلیت کشف خوشه‌های با اشکال مختلف را داراست. با این وجود، مسائلی نظیر حساسیت به پارامترها و پیچیدگی در تفسیر نیازمند مطالعات بیشتری است.

واژه‌های کلیدی: کاوش داده‌های جریانی، خوشه‌بندی در زمان واقعی، DenStream، نقاط ناهنجار، الگوریتم پویا، پارامترهای خوشه‌بندی، الگوریتم تکاملی

1. مقدمه:

مقدمه‌ای که به چالش‌های موجود در خوشه‌بندی داده‌های جریانی، اهمیت آن‌ها در زمینه‌های مختلف و نیاز به روش‌هایی همچون DenStream اشاره می‌کند. همچنین، معرفی ساختار این بخش‌ها و شرح کوتاهی از محتوای بخش‌های اصلی مقاله.

2. کارهای مرتبط:

در این بخش، علاوه بر CluStream و StreamKM++، الگوریتم‌های دیگری همچون BIRCH و COBWEB که مورد استفاده برای خوشه‌بندی داده‌های جریانی قرار می‌گیرند، مورد بررسی قرار می‌گیرند. این الگوریتم‌ها بر اساس معیارها و ایده‌های مختلفی نسبت به DenStream طراحی شده‌اند. نقاط قوت و ضعف هر یک از این روش‌ها به تفصیل بررسی می‌شوند و ارتباط آن‌ها با DenStream به صورت دقیق مورد بحث قرار می‌گیرد.

3. روش پیشنهادی (DenStream):

الگوریتم DenStream یک الگوریتم خوشه‌بندی تکاملی برای داده‌های جریانی است که به وسیله‌ی میکروخوشه‌ها و خوشه‌های هسته‌ای این امکان را فراهم می‌کند تا داده‌ها به صورت پویا در طول زمان خوشه‌بندی شوند. این الگوریتم به صورت پیوسته میکروخوشه‌ها را به روزرسانی می‌کند و با توجه به ورودی داده‌های جدید، خوشه‌های جدید تشکیل می‌دهد یا میکروخوشه‌های موجود را تغییر می‌دهد.

میکروخوشه‌ها:

میکروخوشه‌ها به عنوان نماینده‌های کوچک از داده‌ها عمل می‌کنند و بر اساس میزان چگالی و ارتباط با داده‌های جدید شکل می‌گیرند. هر میکروخوشه شامل دو بخش است: میکروخوشه هسته‌ای (C-Micro-Cluster) که نقطه‌ای اصلی در خوشه را نمایندگی می‌کند و میکروخوشه بالقوه (O-Micro-Cluster) که شامل داده‌های کمتر ارتباطی با خوشه دارد و به عنوان نقاط پرت مد نظر قرار می‌گیرد.

فرآیند تکامل میکروخوشه‌ها:

1. تشکیل اولیه میکروخوشه‌ها:

- با شروع جریان داده، چند نقطه اولیه به عنوان میکروخوشه‌های هسته‌ای انتخاب می‌شوند.

2. به‌روزرسانی میکروخوشه‌ها با ورود داده جدید:

- هر نقطه داده جدید که وارد جریان می‌شود، با میکروخوشه‌های موجود مقایسه شده و بهترین میکروخوشه برای شامل کردن آن انتخاب می‌شود.

- اگر نقطه‌ای نتواند به میکروخوشه‌های موجود اضافه شود، یک میکروخوشه جدید تشکیل می‌شود یا نقطه به عنوان نقطه پرت در نظر گرفته می‌شود.

3. تشکیل خوشه‌ها از میکروخوشه‌ها:

- با استفاده از میکروخوشه‌های هسته‌ای و بالقوه، خوشه‌های اصلی شناسایی می‌شوند.

4. شناسایی نقاط پرت:

- با بررسی انحراف نقاط از میکروخوشه‌های هسته‌ای، نقاط پرت شناسایی می‌شوند و در نظر گرفته می‌شوند.

5. تطبیق و تنظیم پارامترها:

- پارامترهای DenStream از جمله epsilon، lambda، beta، و mu به طور پویا با تغییرات در جریان داده تنظیم می‌شوند تا الگوریتم بتواند به طور بهینه با داده‌های جدید سازگاری پیدا کند.

این الگوریتم با بهره‌گیری از مفاهیم میکروخوشه و خوشه‌های هسته‌ای، قادر است به طور پویا با داده‌های جدید سازگاری یابد و توانایی خوبی در خوشه‌بندی داده‌های جریانی را داشته باشد.

4. نتایج و آزمایش‌ها:

در این بخش، نتایج آزمایش‌های مختلفی که با استفاده از DenStream انجام شده است، به صورت جزییاتی گزارش می‌شود. معیارهای ارزیابی استفاده شده برای سنجش عملکرد DenStream، مقایسه نتایج با الگوریتم‌های دیگر و تحلیل عملکرد DenStream در مقایسه با آن‌ها در این بخش بیان می‌شود. همچنین، مطالعات موردی از کاربردهای واقعی DenStream مانند بررسی الگوهای مختلف هواشناسی یا مسائل مرتبط با اقتصاد و تجارت در این بخش قرار می‌گیرد.

5. بحث:

تحلیل نتایج به دست آمده، بحث در مورد مزایا و معایب DenStream، ارتباط آن با کاربردهای مختلف و چگونگی تأثیر آن بر این کاربردها. همچنین، پیشنهادات برای بهبود یا توسعه‌ی آینده‌ی این الگوریتم.

6. نتیجه‌گیری:

خلاصه‌ای از نتایج به دست آمده از مقاله، اهمیت این نتایج در حوزه‌ی خوشه‌بندی داده‌های جریانی، و پیشنهادات برای کارهای آینده و توسعه‌ی بیشتر.

7. منابع:

TODO