به نام خدا

تمرین دوم درس مبانی داده کاوی

(توضیحات کارهای انجام شده)

جناب استاد دکتر قاسمی

تهیه کننده

امیرعلی خانه عنقا

220797044

1401 – 1400

- **مقدمه**

استخراج جریان می‌تواند به عنوان فرآیند یافتن ساختار پیچیده در یک حجم بزرگ از داده‌ها تعریف شود که در آن داده‌ها در طول زمان تکامل می‌یابند و به یک جریان نامحدود می‌رسند. جریان داده یک توالی از داده‌های پیوسته ورودی است که یک محدودیت یک طرفه را تحمیل می‌کند که در آن دسترسی تصادفی به داده‌ها امکان پذیر نیست.

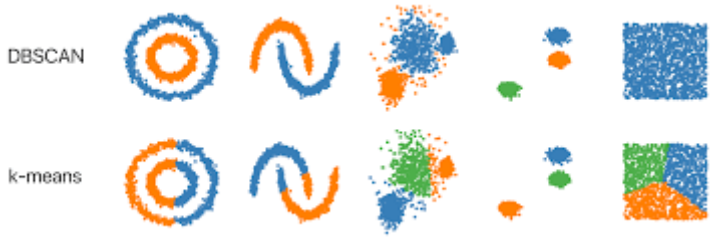
خوشه‌بندی، فرآیندی است که به کمک آن می‌توان مجموعه‌ای از اشیاء را به گروه‌های مجزا افراز کرد. هر افراز یک خوشه نامیده می‌شود. اعضاء هر خوشه با توجه به ویژگی‌هایی که دارند به یکدیگر بسیار شبیه هستند و در عوض میزان شباهت بین خوشه‌ها کمترین مقدار است. در چنین حالتی هدف از خوشه‌بندی، نسبت دادن برچسب‌هایی به اشیاء است که نشان دهنده عضویت هر شیء به خوشه است.

در مقاله داده شده یک الگوریتم خوشه‌بندی جریان داده، به نام چگالی خود سازماندهی مبتنی بر خوشه‌بندی بر روی جریان داده (SOStream) پیشنهاد شده‌است. این الگوریتم چندین ویژگی جدید دارد. به جای استفاده از یک آستانه تشابه تعریف‌شده توسط کاربر یا یک شبکه ثابت، SOStream ساختار در جریان داده‌های در حال تکامل سریع را با تطبیق خودکار آستانه برای خوشه‌بندی مبتنی بر تراکم تشخیص می‌دهد.

الگوریتم های خوشه‌بندی جریان برای گروه‌بندی رویدادها براساس شباهت بین ویژگی‌ها مورد استفاده قرار می‌گیرند. داده‌هایی که به جریان‌ها می‌رسند اغلب حاوی نویز و داده‌های نامرتبط هستند. بنابراین، خوشه‌بندی جریان داده باید قادر به شناسایی، تشخیص و فیلتر کردن این داده‌ها قبل از خوشه‌بندی باشد.

- **خوشه‌بندی برمبنای چگالی(Density-Bases Clustering)**

روش‌های خوشه‌بندی تفکیکی قادر به تشخیص خوشه‌هایی کروی شکل هستند. به این معنی که برای تشخیص خوشه‌ها از مجموعه داده‌هایی به شکل‌های <<کوژ (Convex) >> یا محدب خوب عمل می‌کنند. در عوض برای تشخیص خوشه‌ها برای مجموعه داده‌های <<کاو>> (Concave)یا مقعر دچار خطا می‌شوند. به تصویر 1 توجه کنید که بیانگر شکل‌های کاو است.



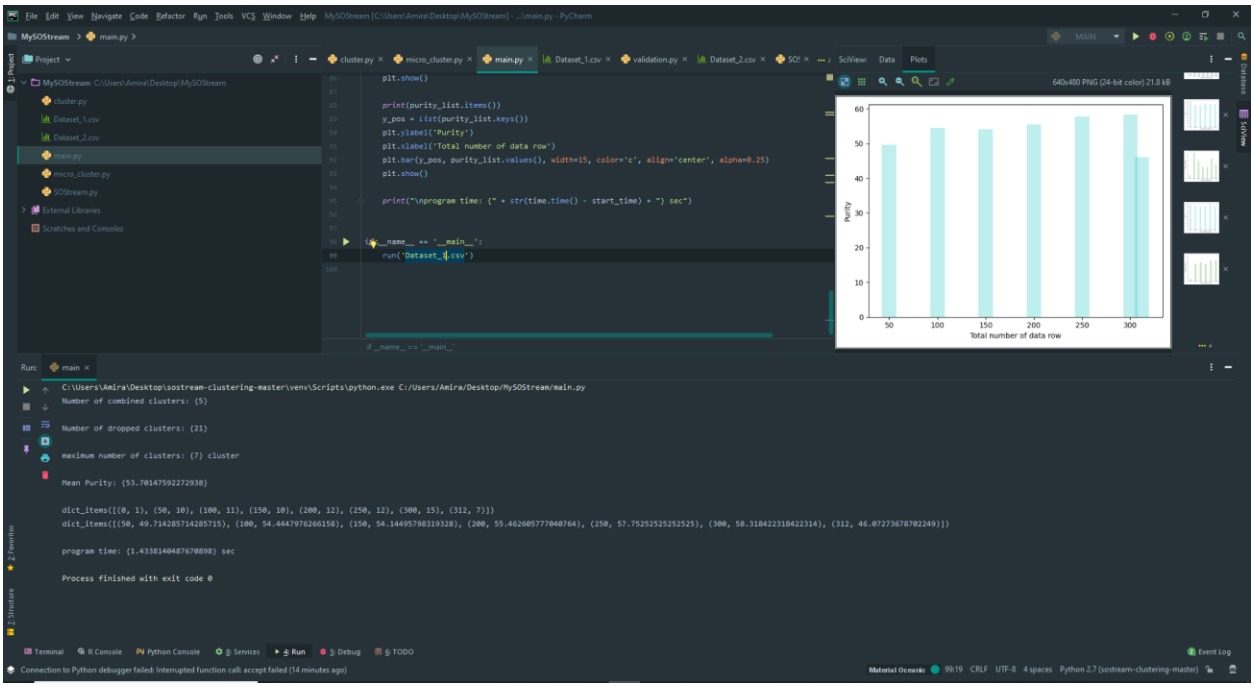
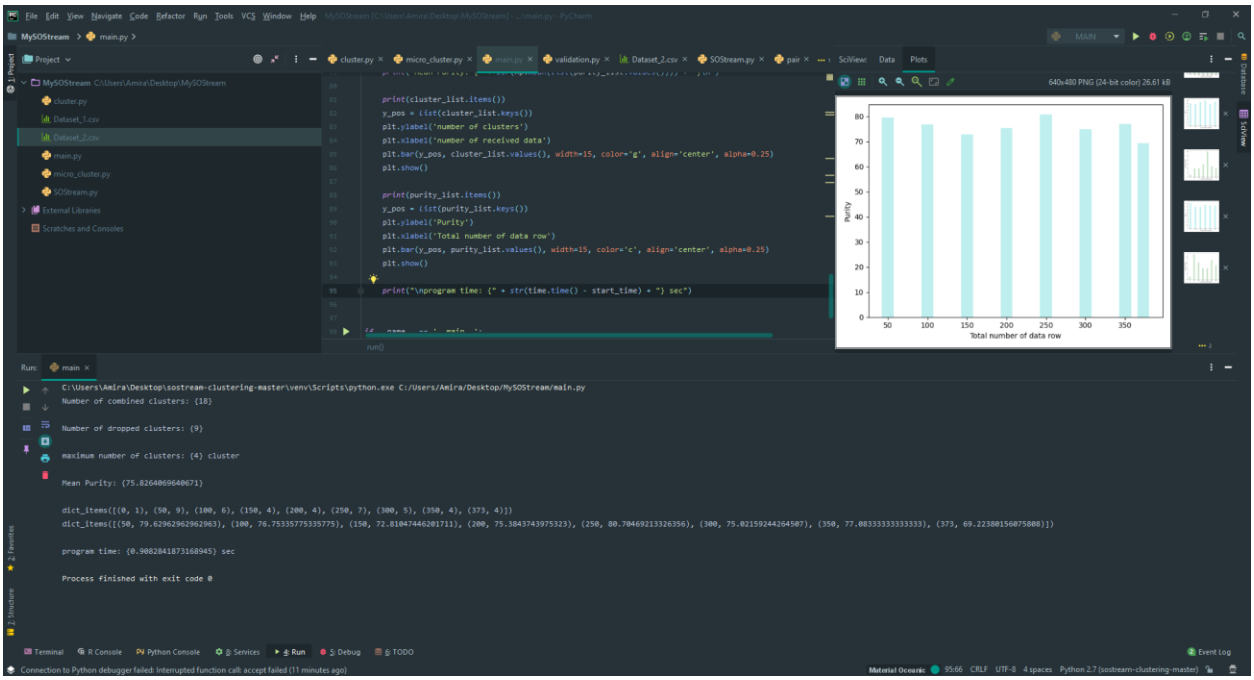
تصویر1: مقایسه خوشه بندی ببرمبنای چگالی و بر مبنای تفکیکی

در الگوریتمِ DBSCAN دو پارامتر وجود دارد. یکی از آن‌ها شعاع است که به آن Epsilon نیز می‌گویند و دومی حداقل نقاط موجود در یک خوشه است که به آن MinPoints می‌گویند. نحوه‌ی کار الگوریتم ساده است. این الگوریتم ابتدا یک نمونه (که همان یک نقطه در فضای برداری می‌شود) را انتخاب می‌کند و با توجه به شعاعِ Epsilon به دنبال همسابه برای این نقطه در فضا می‌گردد. اگر الگوریتم در آن شعاعِ مشخصِ Epsilon حداقل توانست به تعدادِ MinPoint نقطه پیدا کند، آن‌گاه همه‌ی آن نقطه‌ها با هم به یک خوشه تعلق می‌گیرند. الگوریتمْ سپس به دنبال یکی از نقطه‌های همجوار نقطه فعلی می‌رود تا دوباره با شعاعِ Epsilon در آن نقطه به دنبالِ نقاط همسایه دیگر بگردد و اگر تعدادِ نقاطِ همسایه‌ی جدید بازهم پیدا شوند، این الگوریتم دوباره همه آن نقاطِ جدید را با نقاط قبلی به یک خوشه متعلق می‌کند و اگر نقطه‌ی جدیدی در همسایگی پیدا نکرد این خوشه تمام شده است و برای پیدا کردنِ خوشه‌های دیگر در نقاط دیگر، به صورت تصادفی یک نقطه دیگر را انتخاب کرده و شروع به یافتن همسایه و تشکیلِ خوشه‌ی جدید برای آن نقطه می‌کند. این کار آنقدر ادامه پیدا می کند تا تمامی نقاط بررسی شوند.

SOStream یک الگوریتم خوشه‌بندی مبتنی بر چگالی است که می‌تواند آستانه خود را با جریان داده تطبیق دهد. این روش از یک تابع محوشدگی نمایی برای کاهش تاثیر داده‌های

قدیمی استفاده می‌کند که ارتباط آن‌ها در طول زمان کاهش می‌یابد. **SOStream** ویژگی‌های جدید زیر را دارد:

- تنظیم یک آستانه به صورت دستی برای خوشه‌بندی مبتنی بر چگالی (آستانه تشابه، اندازه شبکه و غیره) دشوار است و اگر این پارامتر بر روی یک مقدار نامناسب تنظیم شود، آنگاه الگوریتم از بیش برازش رنج می‌برد، در حالی که در نهایت خوشه‌بندی ناپایدار است. **SOStream** این مساله را با استفاده از یک مقدار آستانه یادگیری پویا برای هر خوشه براساس ایده ساخت فضاهایی با حداقل تعداد نقاط، مورد بررسی قرار می‌دهد.
- **SOStream** از یک استراتژی به روزرسانی خوشه‌ای جدید استفاده می‌کند که از تکنیک‌های یادگیری رقابتی که برای خود سازماندهی نقشه‌ها (**SOM** ها) و آینده (خوشه‌بندی با استفاده از نمایندگان) توسعه داده شده‌اند، الهام‌گرفته شده‌است . در آینده از یک استراتژی کوچک کردن منحصر به فرد استفاده می‌شود که ما را تشویق به پیاده‌سازی روش مشابه برای **SOStream** می‌کند. خوشه‌های کوچکی که پس از کوچک شدن شکل می‌گیرند، به عنوان نماینده خوشه جهانی مورد استفاده قرار می‌گیرند. روند کوچک کردن همچنین به شناسایی درست خوشه‌های دارای همپوشانی بالا کمک می‌کند. در نتیجه، خوشه‌ها نسبت به داده‌های نامربوط کم‌تر حساس می‌شوند.
- تمام جنبه‌های **SOStream** (از جمله حذف، اضافه کردن، ادغام و محوشدگی خوشه‌ها) به صورت لحظه‌ای انجام می‌شوند.

در پیاده سازی داده شده الگوریتم‌ها و چگونگی پیاده سازی **SOStream** به صورت کامل قابل مشاهده است.

خروجی های کد براساس دو دیتاست داده شده :