به نام حق

درس شناسایی الگو

نام دبیر: استاد سلیمی

گزارش پروژه

امیرحسین قضاتی (۴۰۳۴۴۳۱۳۶)

سهیل حمزه‌بیگی (۴۰۳۴۴۳۰۴۷)

ترم اول سال تحصیلی ۱۴۰۳ – ۱۴۰۴

بخش پیش‌پردازش:

در این بخش به بررسی داده‌ها و پیش‌پردازش‌های لازم پرداختیم. در واقع داده‌های خالی را بررسی کردیم و مشکل آن‌ها را رفع کردیم. سوابق رانندگی‌های مختلف را با هم ترکیب کردیم و یک دیتاست جداگانه از آن ساختیم. هر دیتاست را با لیبل متناظر آن نیز ترکیب کرده و در یک دیتاست قرار دادیم و در کنار دیتاست مرتبط با کیفیف بنزین، از تمامی این دیتاست‌ها، مولفه‌های آماری، رسم هیستوگرام و توزیع‌ها و همبستگی‌های آنان را استخراج کردیم. این رویه با استفاده از کتابخانه plotly صورت گرفت که استفاده از نتایج به صورت تعاملی ممکن و راحت باشد. از ذکر نتایج این بخش به دلیل کثرت، در اینجا خودداری شد و در فایل ipynb قابل مشاهده هستند.

در بخش محاسبه مصرف سوخت لحظه‌ای، ابتدا برای هر رانندگی، تغییرات تجمعی سوخت مصرفی (TPF) و مسافت طی شده (CM) را محاسبه می‌کیم و سپس نسبت این تغییرات را در مقیاس ۱۰۰ کیلومتر به عنوان مصرف سوخت لحظه‌ای محاسبه می‌کنیم.

در اولین بخش آن، تابعی به نام calc_no_shift داده‌های گروه‌بندی شده را بر اساس زمان مرتب می‌کند و اختلاف‌های متوالی بین مقادیر TPF و CM را با استفاده از diff محاسبه می‌کند. در این حالت، تفاوت‌ها بدون شیفت زمانی اضافی محاسبه می‌شوند و مصرف سوخت لحظه‌ای با ضرب نسبت اختلاف‌ها در ۱۰۰ بدست می‌آید.

سپس در تابع calc_forward_shift، داده‌ها نیز بر اساس زمان مرتب شده و از تابع shift با مقدار ۱- استفاده می‌شود تا مقدار بعدی هر رکورد به عنوان مقدار پیش‌بینی شده (forward) برای TPF و CM به دست آید. اختلاف بین مقدار فعلی و مقدار آینده محاسبه شده و مصرف سوخت لحظه‌ای به همین روش به دست می‌آید.

در تابع calc_backward_shift، همان روند با استفاده از shift(1) انجام می‌شود؛ به این ترتیب اختلاف بین مقدار فعلی و مقدار قبلی (backward) گرفته شده و نسبت آن محاسبه می‌شود.

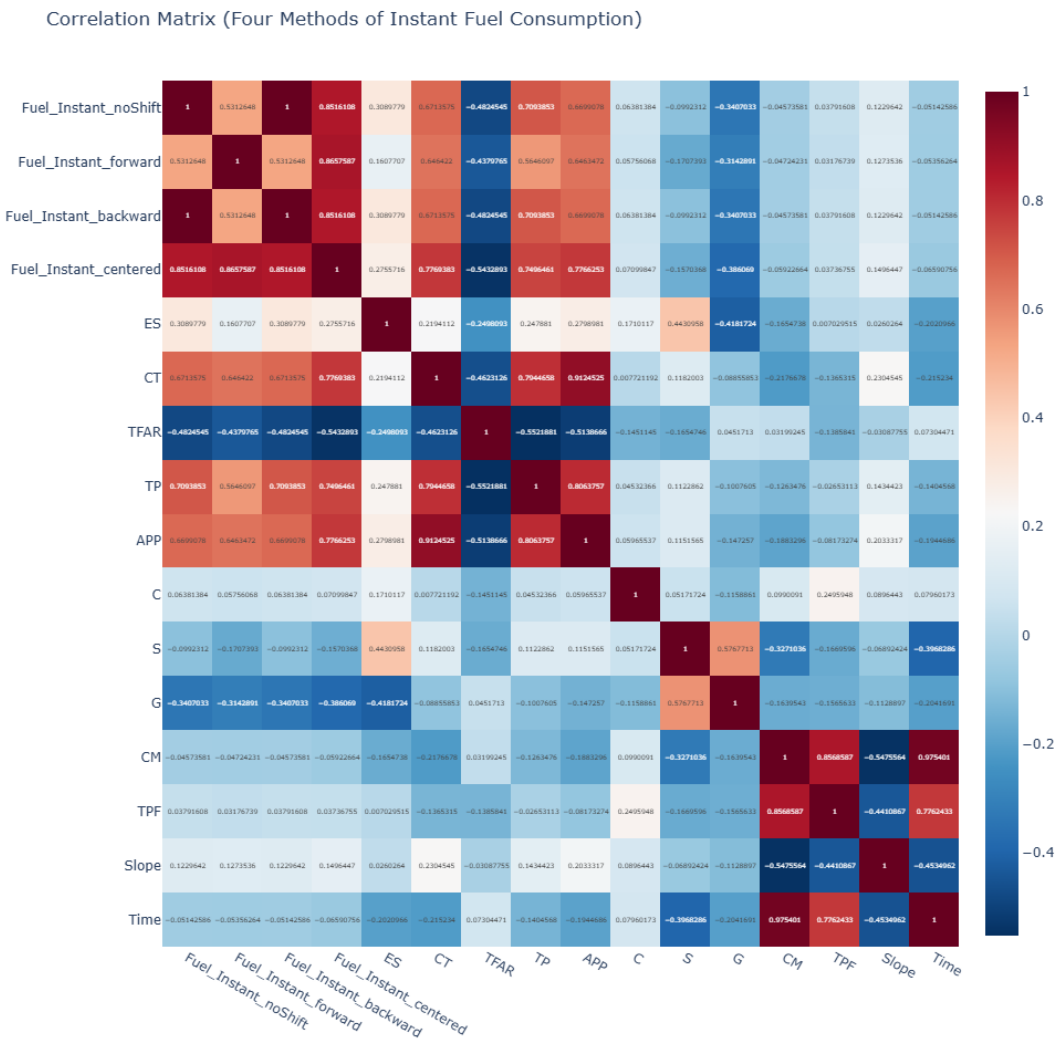
تابع calc_centered_diff از ترکیب داده‌های مقدار قبلی و بعدی استفاده می‌کند. ابتدا با shift(-1) و shift(1) مقدار بعدی و قبلی برای TPF و CM به دست می‌آید. سپس اختلاف بین مقدار بعدی و قبلی به عنوان تفاوت مرکزی (centered) محاسبه شده و نسبت آن به صورت مصرف سوخت لحظه‌ای محاسبه می‌شود.

پس از اجرای این توابع بر روی هر گروه، چهار نوع محاسبه متفاوت برای اختلاف مسافت (DeltaCM) و مصرف سوخت لحظه‌ای در دیتافریم ثبت می‌شود. در ادامه، برای ستون‌هایی که مربوط به تغییرات مسافت هستند، رکوردهایی حذف می‌شوند که مقدار آن‌ها نادرست مثلا NaN یا نزدیک به صفر باشند تا از تقسیم بر صفر جلوگیری شود. همچنین برای ستون‌های مصرف سوخت لحظه‌ای نیز رکوردهای ناصحیح حذف می‌شوند.

در مجموع، این کد با استفاده از چهار روش متفاوت (بدون شیفت، شیفت به جلو، شیفت به عقب و تفاوت مرکزی) به محاسبه مصرف سوخت لحظه‌ای می‌پردازد و سپس با پالایش نتایج، داده‌هایی معتبر برای تحلیل‌های بعدی در اختیار قرار می‌دهد.

بعد از اعمال تابع همبستگی، روش Centered برای متغیرهای کلیدی بالاترین مقدار ضریب همبستگی را داراست. همین‌طور نسبت هوا به سوخت (TFAR) انتظار می‌رود رابطه منفی با مصرف لحظه‌ای داشته باشد (هرچه هوا/سوخت بالاتر باشد، سوخت کم‌تری به ازای هوا مصرف می‌شود، پس مصرف لحظه‌ای پایین‌تر است). پس از نظر منطق فیزیکی هم روش Centered قوی‌ترین همبستگی منفی را نشان می‌دهد.

نتایج بررسی همبستگی:



بخش اول)

ابتدا، داده‌ها بر اساس شناسه رانندگی (Filename) گروه‌بندی شده و با استفاده از تابع build_features، ویژگی‌های آماری و توزیعی از هر گروه استخراج می‌شود. به این صورت که برای هر یک از ستون‌های اصلی مانند ES، CT، TFAR، TP، APP، C، S، G، Slope، Time و CM، مولفه‌های میانگین، انحراف معیار، حداقل و حداکثر محاسبه شده و علاوه بر این، توزیع‌های نسبی (هیستوگرام) برای مقادیر ES و S نیز محاسبه می‌شود. همچنین تعداد تغییرات دنده و تعداد جهش‌های ناگهانی در APP به عنوان ویژگی‌های تکمیلی به دست می‌آید.

سپس اطلاعات مربوط به برچسب‌های واقعی مانند 3_Label و 5_Label از همان داده‌ها استخراج شده و با داده‌های ویژگی استخراج‌شده ادغام می‌شود تا یک دیتافریم نهایی حاصل شود. این دیتافریم شامل ردیف‌هایی است که هر کدام نمایانگر یک رانندگی با ویژگی‌های استخراج شده به همراه برچسب مربوط به آن هستند.

بعد از این مرحله، داده‌های به دست آمده از ویژگی‌ها انتخاب و استانداردسازی می‌شوند تا مقیاس‌بندی یکسان برای الگوریتم‌های خوشه‌بندی فراهم شود. سپس برای دو حالت تعداد خوشه (۳ و ۵) فرآیند خوشه‌بندی انجام می‌شود. در اینجا چند روش خوشه‌بندی متفاوت مانند KMeans، AgglomerativeClustering، SpectralClustering، FCM، GMM و DBSCAN روی داده‌های استاندارد شده اعمال می‌شوند و نتایج هر روش در یک دیکشنری ذخیره می‌شود.

برای بهبود دقت و پایداری خوشه‌بندی، یک روش Ensemble پیاده‌سازی شده است. در این بخش، با استفاده از وزن‌دهی سفارشی برای هر روش، به طوری که به عنوان مثال به روش‌های نامطمئن وزن کمتری داده شود، یک Confusion matrix ساخته می‌شود. در این ماتریس اگر دو نمونه در خروجی یک روش در یک خوشه باشند، مقدار وزنی به عنصر مربوطه اضافه می‌شود. پس از نرمال‌سازی این ماتریس، فاصله بین نمونه‌ها به عنوان ۱ منهای Confusion matrix محاسبه می‌شود و سپس با استفاده از خوشه‌بندی AgglomerativeClustering بر روی این ماتریس فاصله، خوشه‌بندی نهایی حاصل می‌شود.

در ادامه، دقت هر یک از روش‌های خوشه‌بندی و همچنین خوشه‌بندی Ensemble با استفاده از تابع calc_accuracy اندازه‌گیری می‌شود. این تابع با استفاده از الگوریتم Hungarian بهترین تطبیق بین برچسب‌های واقعی و برچسب‌های تولیدشده را برقرار می‌کند تا دقت نهایی را به دست آورد.

در انتها، به منظور نمایش بصری نتایج خوشه‌بندی نهایی، داده‌های استاندارد شده با استفاده از تکنیک PCA به دو بُعد کاهش می‌یابند و سپس بر روی یک نمودار پراکندگی رسم می‌شوند. رنگ هر نقطه نشان‌دهنده خوشه‌ای است که به آن تعلق دارد، که این امر تفکیک و بررسی بصری نتایج خوشه‌بندی را آسان می‌کند.

نتایج:

K=3, km: 0.79

K=3, agg: 0.56

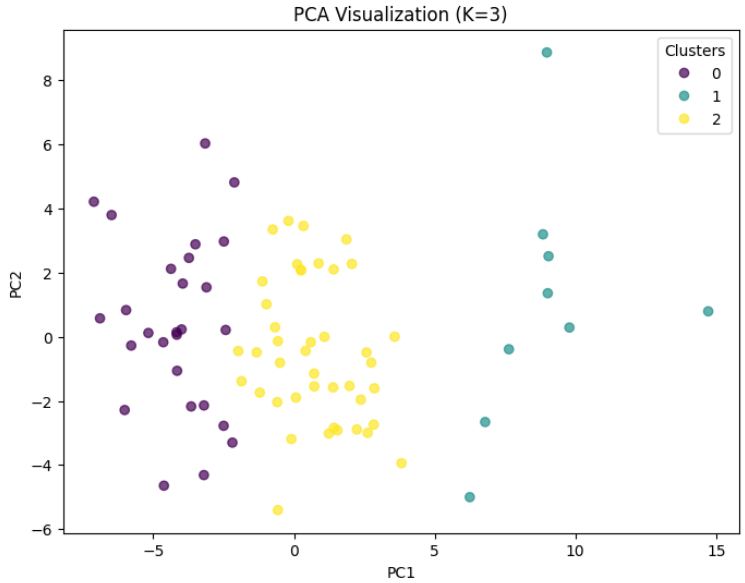
K=3, sp: 0.49

K=3, fcm: 0.58

K=3, gm: 0.78

K=3, db: 0.46

K=3, Ensemble: 0.79



با توجه به نتایج و نمودارهای ارائه‌شده، می‌توان دریافت که داده‌ها در حالت سه خوشه، ساختار منسجم‌تری دارند و روش‌هایی مانند K-Means یا GMM توانسته‌اند در این سناریو حدود ۰.۷۹ دقت کسب کنند. این وضعیت حاکی از آن است که داده‌ها حداقل از لحاظ هندسی در یک فضای کاهش‌یافته با PCA، به شکل سه توده‌ی عمده توزیع شده‌اند و تحمیل تعداد بیشتر خوشه‌ها سبب می‌شود برخی نمونه‌ها به شکل مصنوعی در زیرخوشه‌های اضافی قرار گیرند و برچسب‌بندی واقعی بر اساس سه یا پنج نوع رانندگی با هم‌پوشانی بیشتری همراه شود؛ ازاین‌رو معیار دقت در حالت پنج خوشه کاهش می‌یابد.

K=5, km: 0.54

K=5, agg: 0.40

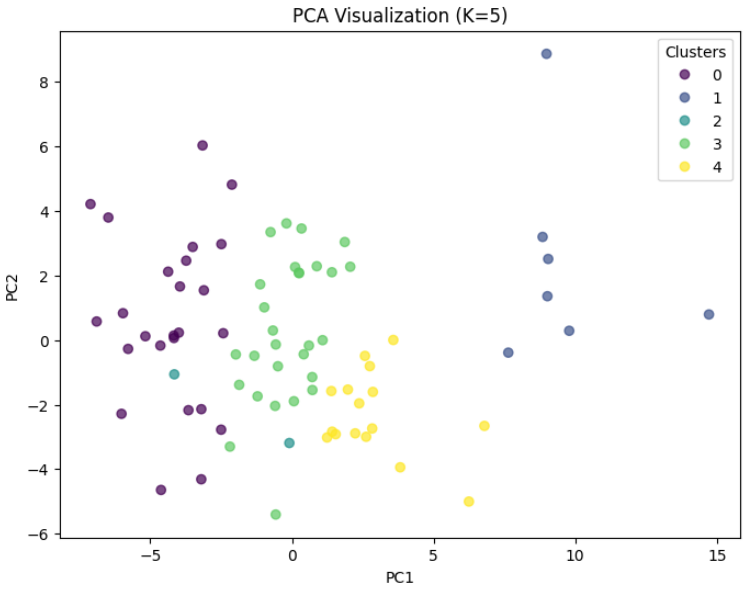
K=5, sp: 0.35

K=5, fcm: 0.44

K=5, gm: 0.55

K=5, db: 0.32

K=5, Ensemble: 0.56



مجموعه نتایج نشان می‌دهد که در وضعیت سه خوشه، خروجی الگوریتم‌های K-Means و GMM نزدیک به هم و در بهترین حالت برابر ۰.۷۸ تا ۰.۷۹ است، درحالی‌که روش‌های دیگر نظیر Agglomerative ، DBSCAN یا Spectral Clustering با پارامترهای کنونی به دقت پایین‌تری دست یافته‌اند. در حالت پنج خوشه نیز K-Means و GMM همچنان عملکرد بهتری نسبت به سایر روش‌ها دارند، اما به طور کلی دقت کمتری نسبت به سناریوی سه خوشه به دست می‌آید و این امر در نمودار PCA نیز مشهود است که نقاط مجبور به توزیع در پنج رنگ شده‌اند و توده‌های کلی داده به صورت افقی یا عمودی شکسته شده‌اند. Ensemble Clustering نهایتاً به اندازه روش‌های تک‌الگوریتمی مثل K-Means یا GMM عمل می‌کند و در حالت سه خوشه به همان حدود ۰.۷۹ رسیده است.

مشاهده نمودار با کاهش ابعاد توسط PCA نیز تأیید می‌کند که در حالت سه خوشه، سه توده متمایز در صفحه (PC1, PC2) پدیدار می‌شود. یکی در سمت راست با فاصله قابل ملاحظه از دو توده دیگر، و دو توده دیگر در ناحیه مرکزی و چپ نمودار که مرز میانشان از پراکندگی نقطه‌ها در محورهای اصلی قابل تشخیص است. در وضعیت پنج خوشه، همان سه توده به اجبار به بخش‌های بیشتری خرد می‌شوند و تداخل رنگ در ناحیه‌های نزدیک مرزها، کاهش دقت را توجیه می‌کند.

در مجموع نتایج آزمایش حاکی از آن است که داده‌ها به شکل ذاتی سه گروه مجزا را بهتر منعکس می‌کنند و روش‌هایی با مفروضات کروی یا توزیع گاوسی، مانند K-Means و GMM، برای این ساختار سودمندتر بوده‌اند. عملکرد ضعیف‌تر روش‌های دیگری مثل Agglomerative، DBSCAN و Spectral Clustering احتمالاً ناشی از پارامترهای پیش‌فرض یا شکل واقعی داده‌هاست که با مفروضات این الگوریتم‌ها مانند پیوستگی در DBSCAN یا ساختار گرافی در Spectral، سازگاری کمتری دارد. بهبود آن دسته از روش‌ها ممکن است با تنظیم دقیق‌تر پارامترها یا بازی بیشتری با ویژگی‌ها صورت گیرد.