دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
دانشکده ریاضی و علوم کامپیوتر

گزارش تمرین پنجم درس داده‌کاوی محاسباتی

انتخاب ویژگی با استفاده از مقادیر تکین (SVD) در برابر روش‌های کلاسیک

نگارش
زهرا براتی

استاد درس
دکتر مهدی قطعی

تدریسیار
آقای بهنام یوسفی‌مهر

آذر ۱۴۰۴

چکیده

این پژوهش، مسئله انتخاب ویژگی در داده‌های صنعتی با ابعاد بالا مورد بررسی قرار گرفته است. از دیتاست SECOM به‌عنوان یک نمونه واقعی شامل تعداد زیادی سنسور عددی استفاده شد. پس از انجام پیش‌پردازش شامل حذف ویژگی‌های ثابت، جایگزینی مقادیر گمشده و نرمال‌سازی، سه روش انتخاب ویژگی شامل اطلاعات متقابل[1]، الگوریتم حذفی بازگشت ویژگی‌ها[2] مبتنی‌بر جنگل تصادفی[3] و یک روش جبری مبتنی بر تجزیه مقادیر تکین[4] پیاده‌سازی و مقایسه شدند. عملکرد این روش‌ها با استفاده از مدل Logistic Regression و معیارهای Accuracy و F1-Score ارزیابی شد. نتایج نشان داد که روش RFE بالاترین عملکرد پیش‌بینی را ارائه می‌دهد، درحالی‌که روش جبری مبتنی‌بر SVD از نظر زمان محاسباتی و پایداری در برابر نویز برتری دارد. این مقایسه نشان می‌دهد که انتخاب روش مناسب به ملاحظاتی مانند دقت و سرعت وابسته است.

واژه‌های کلیدی:

تجزیه مقادیر تکین، الگوریتم حذفی بازگشت ویژگی‌ها، اطلاعات متقابل

[1] Mutual Information

[2] Recursive Feature Elimination (RFE)

[3] Random Forest

[4] Singular Value Decomposition (SVD)

فهرست مطالب

صفحه

فصل اول

مقدمه

مقدمه

در مسائل یادگیری ماشین با داده‌های با بعد بالا، انتخاب ویژگی یکی از مراحل کلیدی پیش‌پردازش داده‌ها است. وجود تعداد زیادی ویژگی، به‌ویژه در داده‌های صنعتی، علاوه‌بر افزایش هزینه محاسباتی، می‌تواند باعث کاهش کارایی مدل‌ها به دلیل نویز، افزونگی و هم‌بستگی بین ویژگی‌ها شود. از این رو، استفاده از روش‌های مناسب برای کاهش بعد و انتخاب زیرمجموعه‌ای از ویژگی‌های معنادار اهمیت زیادی دارد.

دیتاست SECOM یکی از نمونه‌های شاخص داده‌های صنعتی با ابعاد بالا است که شامل صدها سنسور عددی و سطح قابل توجهی از نویز و مقادیر گمشده است. این ویژگی‌ها، SECOM را به یک بستر مناسب برای بررسی و مقایسه روش‌های مختلف انتخاب ویژگی تبدیل می‌کند.

در این گزارش، سه رویکرد متفاوت برای انتخاب ویژگی مورد بررسی قرار گرفته‌اند: یک روش فیلتر مبتنی‌بر اطلاعات متقابل، یک روش پوششی مبتنی‌بر مدل Random Forest و الگوریتم RFE و یک روش جبری مبتنی‌بر تجزیه مقادیر تکین. این روش‌ها از نظر دقت طبقه‌بندی، معیار F1، زمان انتخاب ویژگی و پایداری در برابر نویز با یکدیگر مقایسه شده‌اند.

فصل دوم

آماده‌سازی داده‌های صنعتی

آماده‌سازی داده‌های صنعتی

در این پروژه از دیتاست SECOM که از مخزن UCI تهیه شده است، استفاده می‌شود. این دیتاست شامل داده‌های صنعتی حاصل از صدها سنسور در یک فرآیند تولید نیمه‌هادی است. این داده‌ها شامل نویز و مقادیر گمشده هستند. استفاده مستقیم از این داده‌ها بدون پیش‌پردازش مناسب می‌تواند باعث کاهش کیفیت تحلیل و نتایج انتخاب ویژگی شود. هدف این بخش، آماده‌سازی داده‌ها برای مراحل بعدی انتخاب ویژگی و مدل‌سازی است.

۲-۱- بارگذاری دیتاست و نرمال‌سازی داده‌ها

در این بخش، ابتدا داده‌ها بارگذاری شدند و ستون برچسب از ماتریس ویژگی‌ها جدا شد. با توجه به اینکه ستون زمان پیش‌تر از فایل دیتاست حذف شده است و تمامی ویژگی‌ها ماهیت عددی دارند، نیازی به بررسی یا تبدیل نوع داده وجود نداشت. پس از بارگذاری اولیه، ماتریس ویژگی‌ها شامل ۱۵۶۷ نمونه و ۵۹۰ ویژگی بود.

در ادامه، ویژگی‌هایی که واریانس آن‌ها برابر صفر بود شناسایی و حذف شدند. این ویژگی‌ها در تمام نمونه‌ها مقدار ثابتی دارند و هیچ اطلاعاتی برای تفکیک کلاس‌ها فراهم نمی‌کنند. حذف این ستون‌ها باعث کاهش تعداد ویژگی‌ها از ۵۹۰ به ۴۷۴ شد، درحالی‌که تعداد نمونه‌ها بدون تغییر باقی ماند.

پس از حذف ویژگی‌های ثابت، مقادیر گمشده موجود در داده‌ها با استفاده از روش جایگذاری میانه[5] پر شدند. انتخاب میانه به دلیل مقاومت بیشتر آن در برابر داده‌های پرت انجام شد که با توجه به ماهیت پرنویز داده‌ها انتخاب مناسبی است. پس از این مرحله، هیچ مقدار گمشده‌ای در داده‌ها باقی نماند.

در پایان این بخش، نام ویژگی‌ها برای استفاده در مراحل بعدی تحلیل ذخیره شد تا امکان نگاشت نتایج انتخاب ویژگی به سنسورهای اصلی فراهم شود. در این مرحله تنها عملیات ستونی انجام شده و هیچ تغییری در ساختار سطری داده‌ها ایجاد نشده است، بنابراین نگاشت بین نمونه‌ها و برچسب‌ها به‌طور کامل حفظ شده است. داده‌های حاصل از این مرحله به‌صورت نرمال‌سازی‌شده با روش StandardScaler برای مراحل بعدی انتخاب ویژگی و تحلیل مورد استفاده قرار می‌گیرند.

[5] Median Imputation

فصل سوم

روش‌های کلاسیک

روش‌های کلاسیک

در این بخش، دو روش کلاسیک انتخاب ویژگی مورد بررسی قرار می‌گیرند. این دو روش شامل یک روش مبتنی‌بر فیلتر آماری و یک روش پوششی مبتنی‌بر مدل هستند. هدف از این مقایسه، بررسی تفاوت این دو رویکرد از نظر شیوه انتخاب ویژگی، هزینه محاسباتی و نوع اطلاعاتی است که از داده‌ها استخراج می‌کنند.

۳-۱- روش فیلتر

در این مرحله انتخاب ویژگی‌ها با استفاده از اطلاعات متقابل انجام شد.

اطلاعات متقابل یک معیار برگرفته از نظریه اطلاعات است که میزان وابستگی متقابل بین دو متغیر تصادفی را اندازه‌گیری می‌کند. به عبارت دقیق‌تر، این معیار نشان می‌دهد که با دانستن یک متغیر، چقدر می‌توانیم عدم قطعیت[۶] متغیر دیگر را کاهش دهیم. در مسائل طبقه‌بندی، اطلاعات متقابل می‌تواند به‌عنوان معیاری برای سنجش میزان ارتباط بین یک ویژگی و برچسب کلاس مورد استفاده قرار گیرد.

در روش‌های مبتنی‌بر فیلتر، هر ویژگی به‌صورت مستقل و بدون در نظر گرفتن مدل یادگیری ماشین ارزیابی می‌شود. تابع mutual_info_classif در کتابخانه scikit-learn پیاده‌سازی عملی این ایده را ارائه می‌دهد و مقدار اطلاعات متقابل بین هر ویژگی و متغیر هدف را محاسبه می‌کند. یکی از مزیت‌های این معیار، توانایی آن در شناسایی وابستگی‌های غیرخطی است، در حالی که فرضی درباره شکل توزیع داده‌ها اعمال نمی‌کند.

در این پروژه، مقدار اطلاعات متقابل برای تمامی ویژگی‌های موجود پس از پیش‌پردازش محاسبه شد و سپس ۲۰ ویژگی با بیشترین مقدار اطلاعات متقابل انتخاب شدند. نتایج نشان داد که تنها تعداد محدودی از ویژگی‌ها دارای مقدار اطلاعات متقابل نسبتا بالاتری هستند، در حالی که بسیاری از ویژگی‌ها وابستگی بسیار ضعیفی با برچسب کلاس دارند. زمان اجرای این مرحله نسبتا کوتاه و در حدود ۴.۵ ثانیه بود که این موضوع با ماهیت محاسباتی ساده روش‌های فیلتر سازگار است.

[6] Entropy

با وجود مزیت سرعت بالا، این روش تعامل بین ویژگی‌ها را در نظر نمی‌گیرد و هر ویژگی را به‌صورت جداگانه ارزیابی می‌کند. بنابراین، مجموعه ویژگی‌های منتخب لزوما بهترین زیرمجموعه از منظر عملکرد یک مدل یادگیری ماشین نیست.

۳-۲- روش پوششی

در مقابل روش‌های فیلتر، روش‌های پوششی انتخاب ویژگی به‌طور مستقیم از یک مدل یادگیری ماشین برای ارزیابی کیفیت زیرمجموعه‌های مختلف ویژگی استفاده می‌کنند. یکی از روش‌های شناخته‌شده در این دسته، الگوریتم حذف بازگشتی ویژگی‌ها است که توسط Guyon و همکاران معرفی شد.

در RFE، یک مدل پایه آموزش داده می‌شود و ویژگی‌ها براساس معیار اهمیت ارائه‌شده توسط مدل رتبه‌بندی می‌شوند. در هر تکرار، تعدادی از ویژگی‌های کم‌اهمیت حذف می‌شوند و این فرآیند تا رسیدن به تعداد مشخصی از ویژگی‌ها ادامه پیدا می‌کند. در این پروژه، الگوریتم جنگل تصادفی به‌عنوان مدل پایه مورد استفاده قرار گرفت. جنگل تصادفی یک روش مبتنی‌بر مجموعه‌ای از درخت‌های تصمیم است که به دلیل توانایی در مدل‌سازی روابط غیرخطی و مقاومت در برابر نویز، در بسیاری از کاربردهای صنعتی به‌کار می‌رود.

استفاده از جنگل تصادفی در RFE این امکان را فراهم می‌کند که اهمیت ویژگی‌ها براساس رفتار مدل و تعامل آن‌ها با سایر ویژگی‌ها ارزیابی شود. بااین‌حال، این رویکرد مستلزم آموزش مکرر مدل است و در نتیجه هزینه محاسباتی بالاتری نسبت به روش‌های فیلتر دارد. نتایج تجربی نیز نشان داد که زمان اجرای RFE با حدود ۳۳ ثانیه به‌طور قابل توجهی بیشتر از روش مبتنی‌بر اطلاعات متقابل است.

۳-۳- جمع‌بندی

به‌طور کلی، روش‌های مبتنی بر فیلتر مانند اطلاعات متقابل به دلیل سادگی و سرعت بالا برای تحلیل اولیه داده‌ها و کاهش سریع ابعاد مناسب هستند. در مقابل، روش‌های پوششی مانند RFE، اگرچه از نظر محاسباتی پرهزینه‌تر هستند، اما انتخاب ویژگی را در چارچوب عملکرد مدل انجام می‌دهند و می‌توانند زیرمجموعه‌ای از ویژگی‌ها را ارائه دهند که برای مدل مورد نظر مناسب‌تر باشد.

فصل چهارم

روش جبری

روش جبری

در این بخش، یک رویکرد جبری برای انتخاب ویژگی ارائه می‌شود که مبتنی‌بر ساختار هندسی داده‌ها است. برخلاف روش‌های کلاسیک مبتنی‌بر فیلتر یا مدل، در این رویکرد هیچ مدل یادگیری ماشین برای انتخاب ویژگی استفاده نمی‌شود و انتخاب ویژگی صرفا بر اساس تجزیه خطی ماتریس داده‌ها انجام می‌شود.

هدف این بخش، استخراج ویژگی‌هایی است که بیشترین نقش را در مؤلفه‌های اصلی داده‌ها ایفا می‌کنند.

۴-۱- تجزیه ماتریس داده‌ها با استفاده از تجزیه مقادیر تکین

پس از پیش‌پردازش و نرمال‌سازی داده‌ها، ماتریس ویژگی‌ها با استفاده از تجزیه مقادیر تکین به سه ماتریس تجزیه شد.

$$X = U\Sigma V^T$$

در این رابطه، ماتریس U شامل بردارهای تکین سمت نمونه‌ها، ماتریس Σ شامل مقادیر تکین مرتب‌شده به‌صورت نزولی، و ماتریس V^T شامل بردارهای تکین سمت ویژگی‌ها است. هر بردار در V^T بیانگر یک جهت اصلی در فضای ویژگی‌ها است و ضرایب آن نشان می‌دهند که هر ویژگی تا چه اندازه در آن جهت نقش دارد.

با توجه به ابعاد داده‌ها، تجزیه به‌صورت کامل انجام شد و ابعاد ماتریس‌ها با تعداد نمونه‌ها و ویژگی‌های باقی‌مانده پس از پیش‌پردازش سازگار بود. زمان اجرای تجزیه SVD نیز در حدود ۰.۲۱ ثانیه بود.

۴-۲- تابع امتیازدهی

برای رتبه‌بندی ویژگی‌ها بر اساس نتایج SVD، یک تابع امتیازدهی برای هر ویژگی تعریف شد. ایده اصلی این است که ویژگی‌هایی که در مؤلفه‌های اصلی اولیه سهم بیشتری دارند، اطلاعات ساختاری مهم‌تری از داده‌ها را نمایش می‌دهند.

تابع امتیازدهی برای ویژگی j به‌صورت زیر تعریف شد:

$$Score_j = \sum_{i=1}^{k} \sigma_i^2 \ . \mid V_{ij} \mid$$

در این فرمول، σ_i مقدار تکین مربوط به مؤلفه iام و V_{ij} ضریب مطلق ویژگی j در بردار تکین iام است. وزن‌دهی ضرایب با مربع مقدار تکین باعث می‌شود مؤلفه‌هایی که واریانس بیشتری از داده‌ها را توضیح می‌دهند، سهم بیشتری در امتیاز نهایی داشته باشند، درحالیBکه مؤلفه‌های با مقدار تکین کوچک‌تر که بیشتر نمایانگر نویز هستند، تاثیر کمتری خواهند داشت.

محاسبه امتیازها برای تمامی ویژگی‌ها پس از انجام SVD انجام شد و هزینه محاسباتی این مرحله در مقایسه با خود تجزیه بسیار ناچیز و در حدود ۰.۰۰۰۵ ثانیه بود.

۴-۳- انتخاب ویژگی‌ها براساس امتیازهای محاسبه شده

پس از محاسبه امتیاز برای هر ویژگی، ویژگی‌ها بر اساس مقدار امتیاز به‌صورت نزولی مرتب شدند و ۲۰ ویژگی با بالاترین امتیاز به‌عنوان ویژگی‌های منتخب این روش انتخاب شدند. این ویژگی‌ها بیشترین مشارکت را در مؤلفه‌های اصلی اولیه داشته و بنابراین از دیدگاه ساختار هندسی داده‌ها مهم‌ترین ویژگی‌ها محسوب می‌شوند.

در نتایج مشاهده شد که برخی از ویژگی‌های منتخب امتیازهای بسیار مشابه یا حتی برابر دریافت کرده‌اند. این موضوع نشان‌دهنده آن است که این ویژگی‌ها الگوی مشارکت مشابهی در مؤلفه‌های غالب دارند و از منظر جبری نقش تقریبا یکسانی در بازنمایی داده‌ها ایفا می‌کنند.

۴-۴- جمع‌بندی

در این بخش، یک روش جبری برای انتخاب ویژگی ارائه شد که تنها بر ساختار هندسی داده‌ها متکی است و از هیچ مدل یادگیری ماشین یا اطلاعات برچسب استفاده نمی‌کند. با استفاده از تجزیه مقادیر تکین، میزان مشارکت هر ویژگی در مؤلفه‌های اصلی اولیه استخراج شد و بر اساس یک تابع امتیازدهی ساده، ویژگی‌ها رتبه‌بندی شدند. نتایج نشان داد که این روش با هزینه محاسباتی بسیار کم، قادر به شناسایی ویژگی‌های مهم از دیدگاه ساختاری است. همچنین مشاهده امتیازهای مشابه برای برخی ویژگی‌ها بیانگر

نقش تقریبا یکسان آن‌ها در بازنمایی هندسی داده‌ها است.

فصل پنجم

تحلیل هندسی و پایداری

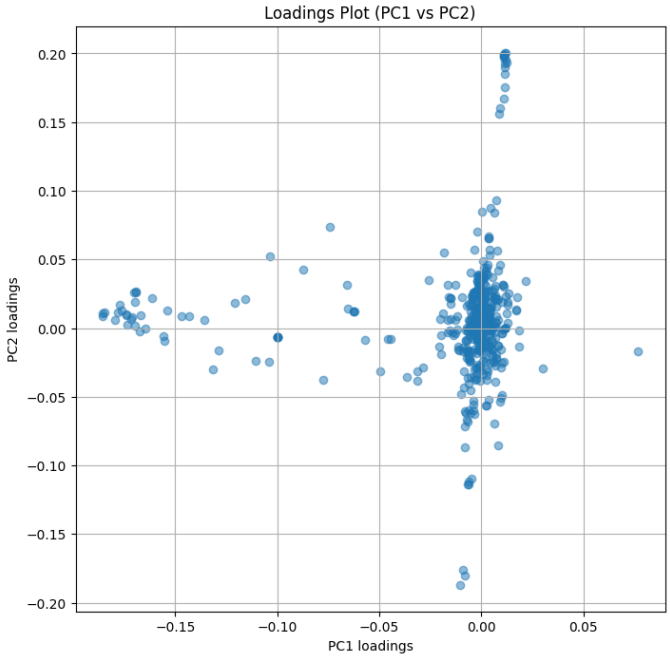
تحلیل هندسی و پایداری

در این بخش به بررسی هندسی ویژگی‌های انتخاب‌شده و ارزیابی پایداری روش‌ها در برابر نویز می‌پردازیم.

۵-۱- نمودار بارگذاری

برای تحلیل هندسی داده‌ها، ضرایب بارگذاری ویژگی‌ها بر روی دو مؤلفه اصلی اول (PC1 و PC2) استخراج و در قالب نمودار بارگذاری ترسیم شد. در این نمودار، هر نقطه نمایانگر یک ویژگی (سنسور) است و مختصات آن نشان‌دهنده میزان مشارکت آن ویژگی در دو جهت اصلی تغییرات داده‌ها است.

بررسی نمودار نشان می‌دهد که بخش عمده‌ای از ویژگی‌ها در نزدیکی مبدأ قرار گرفته‌اند که بیانگر سهم محدود آن‌ها در مؤلفه‌های اصلی است. در مقابل، تعدادی از ویژگی‌ها دارای ضرایب بزرگ‌تر هستند و در نواحی خاصی از فضا تجمع یافته‌اند. این الگو نشان‌دهنده وجود گروه‌هایی از سنسورها با رفتار مشابه است که تغییرات آن‌ها در داده‌ها هم‌راستا بوده است. چنین خوشه‌بندی‌ای بیانگر افزونگی اطلاعاتی میان برخی ویژگی‌ها است و نشان می‌دهد که SVD قادر است ساختار هندسی پنهان داده‌ها را به‌خوبی آشکار کند.



۵-۲- تست پایداری

برای ارزیابی پایداری روش‌های انتخاب ویژگی، ٪۵ نویز تصادفی گاوسی به داده‌های نرمال‌شده اضافه شد و فرآیند انتخاب ویژگی مجدداً برای دو روش SVD و RFE اجرا و سپس میزان اشتراک بین ویژگی‌های منتخب قبل و بعد از افزودن نویز محاسبه شد.

نتایج نشان داد که روش مبتنی‌بر SVD پایداری بسیار بالایی دارد، به‌طوری که ۱۹ ویژگی از ۲۰ ویژگی منتخب اولیه پس از افزودن نویز نیز بدون تغییر باقی ماندند. در مقابل، روش RFE تنها ۱۰ ویژگی مشترک با حالت بدون نویز داشت. این اختلاف قابل توجه نشان می‌دهد که روش‌های جبری که بر ساختار کلی داده‌ها تکیه دارند، نسبت به روش‌های پوششی حساسیت کمتری به نویز دارند.

۵-۳- جمع‌بندی

براساس نتایج حاصل از آزمون پایداری، مشاهده شد که روش مبتنی‌بر SVD در مقایسه با روش RFE ثبات بیشتری در انتخاب ویژگی‌ها نشان می‌دهد. بخش عمده‌ای از ویژگی‌های منتخب توسط SVD پس از افزودن نویز نیز بدون تغییر باقی ماندند، درحالی‌که در روش RFE تغییرات قابل توجه‌تری در مجموعه ویژگی‌های منتخب مشاهده شد. این رفتار نشان می‌دهد که روش‌های جبری که بر ساختار کلی داده‌ها تکیه دارند، حساسیت کمتری نسبت به نویزهای تصادفی دارند.

فصل ششم

مقایسه نهایی

مقایسه نهایی

در این بخش، عملکرد سه روش انتخاب ویژگی شامل Mutual Information، RFE و SVD با استفاده از یک مدل ساده Logistic Regression مقایسه شد. برای هر روش، مدل روی زیرمجموعه‌ای شامل ۲۰ ویژگی منتخب آموزش داده شد و معیارهای Accuracy، F1-Score و زمان انتخاب ویژگی مورد بررسی قرار گرفتند. سپس با رسم نمودار ون درصد اشتراک بین روش RFE و SVD مشخص شد.

۶-۱- مقایسه

نتایج نشان می‌دهند که روش RFE بالاترین دقت و F1-Score را در میان سه روش به دست آورده است. این رفتار قابل انتظار است، زیرا RFE یک روش پوششی است و انتخاب ویژگی را مستقیماً با در نظر گرفتن عملکرد مدل طبقه‌بندی انجام می‌دهد. بااین‌حال، این بهبود عملکرد با هزینه محاسباتی قابل توجهی همراه بوده است و زمان انتخاب ویژگی در این روش به‌مراتب بیشتر از دو روش دیگر است.

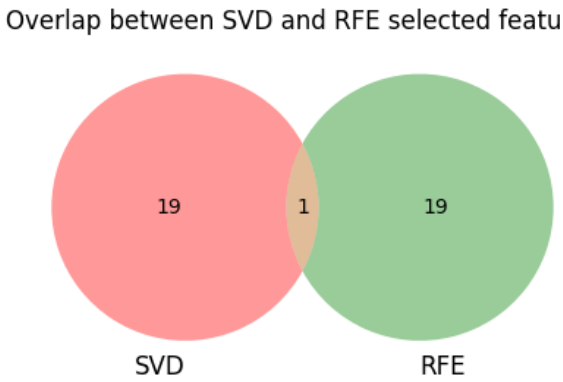
در مقابل، روش مبتنی‌بر SVD کمترین زمان انتخاب ویژگی را دارد و از نظر محاسباتی بسیار سریع است. بااین‌وجود، عملکرد طبقه‌بندی حاصل از ویژگی‌های منتخب این روش ضعیف‌تر از دو روش دیگر ارزیابی شد. این نتیجه نشان می‌دهد که هرچند SVD ساختار کلی داده‌ها را به‌خوبی استخراج می‌کند، اما معیار جبری استفاده‌شده برای رتبه‌بندی ویژگی‌ها الزاماً بهینه‌ترین مجموعه را برای وظیفه طبقه‌بندی فراهم نمی‌کند.

روش Mutual Information عملکردی بینابینی دارد. این روش نسبت به SVD دقت و F1-Score بالاتری ارائه می‌دهد و در عین حال هزینه محاسباتی آن به‌مراتب کمتر از RFE است. بااین‌حال، از آنجا که Mutual Information یک روش فیلتر است و وابستگی مستقیمی به مدل نهایی ندارد، عملکرد آن از RFE ضعیف‌تر باقی می‌ماند.

Method	Accuracy	F1-Score	Feature Selection Time (s)
MI	0.651805	0.171717	4.539981
RFE	0.785563	0.273381	32.630737
SVD	0.592357	0.135135	0.000508

۶-۲- تحلیل همپوشانی

تحلیل همپوشانی بین ویژگی‌های منتخب SVD و RFE نشان داد که تنها یک ویژگی مشترک میان این دو مجموعه وجود دارد که معادل ۵ درصد همپوشانی است. این نتیجه بیانگر آن است که این دو روش از دیدگاه‌های کاملاً متفاوتی به مسئله انتخاب ویژگی نگاه می‌کنند؛ SVD بر ساختار جبری و واریانس کلی داده‌ها تکیه دارد، در حالی که RFE مستقیماً بر قدرت تفکیک ویژگی‌ها در فرآیند طبقه‌بندی متمرکز است.



۶-۳- جمع‌بندی

در مجموع، برای کاربردهای صنعتی که نیازمند تفسیرپذیری و عملکرد پیش‌بینی بالا هستند و محدودیت زمانی شدیدی وجود ندارد، روش RFE گزینه مناسب‌تری است. در مقابل، در سناریوهایی که سرعت پردازش و مقیاس‌پذیری اهمیت بیشتری دارند، روش SVD به دلیل هزینه محاسباتی بسیار پایین می‌تواند انتخاب قابل دفاعی باشد. روش Mutual Information نیز به‌عنوان یک راهکار میانی، توازن مناسبی میان سرعت و عملکرد ارائه می‌دهد.

فصل هفتم

جمع‌بندی و نتیجه‌گیری

جمع‌بندی و نتیجه‌گیری

در این پژوهش، مسئله انتخاب ویژگی در داده‌های صنعتی با ابعاد بالا با تمرکز بر دیتاست SECOM مورد بررسی قرار گرفت. پس از انجام مراحل پیش‌پردازش شامل حذف ویژگی‌های ثابت، جایگزینی مقادیر گمشده و نرمال‌سازی داده‌ها، سه رویکرد متفاوت برای انتخاب ویژگی شامل روش فیلتر مبتنی‌بر Mutual Information، روش پوششی RFE مبتنی‌بر Random Forest و یک روش جبری مبتنی‌بر SVD پیاده‌سازی و با یکدیگر مقایسه شدند.

نتایج تجربی نشان داد که روش RFE بالاترین عملکرد را از نظر Accuracy و F1-Score ارائه می‌دهد. این موضوع قابل انتظار است، زیرا این روش انتخاب ویژگی را مستقیماً با در نظر گرفتن عملکرد مدل طبقه‌بندی انجام می‌دهد. با این حال، هزینه محاسباتی بالای این روش و حساسیت بیشتر آن نسبت به نویز از محدودیت‌های اصلی آن محسوب می‌شود.

در مقابل، روش فیلتر مبتنی‌بر Mutual Information با هزینه محاسباتی بسیار کمتر، عملکردی متوسط ارائه داد. این روش به دلیل سادگی و سرعت بالا، گزینه مناسبی برای کاهش اولیه ابعاد داده‌ها است، اما به دلیل عدم در نظر گرفتن تعامل بین ویژگی‌ها، نمی‌تواند به‌تنهایی بهترین زیرمجموعه ویژگی‌ها را برای مدل نهایی تضمین کند.

روش جبری مبتنی‌بر SVD رویکردی مستقل از مدل ارائه داد که بر ساختار هندسی داده‌ها تکیه دارد. اگرچه عملکرد طبقه‌بندی این روش نسبت به RFE پایین‌تر بود، اما از نظر سرعت اجرا و پایداری در برابر نویز برتری قابل توجهی نشان داد. نتایج آزمون پایداری نیز نشان داد که ویژگی‌های منتخب توسط این روش در مواجهه با نویز تغییرات بسیار محدودی دارند که این موضوع در کاربردهای صنعتی اهمیت ویژه‌ای دارد.

در مجموع، نتایج این مطالعه نشان می‌دهد که هیچ روش واحدی به‌طور مطلق برتر نیست و انتخاب روش مناسب به اهداف مسئله بستگی دارد. در کاربردهایی که دقت پیش‌بینی اولویت اصلی است و محدودیت زمانی وجود ندارد، روش‌های پوششی مانند RFE انتخاب مناسبی هستند. در مقابل، در سناریوهای صنعتی با داده‌های پرنویز و نیاز به سرعت و پایداری بالا، روش‌های جبری مبتنی‌بر SVD می‌توانند گزینه‌ای کارآمد و قابل اتکا باشند. روش Mutual Information نیز به‌عنوان یک راهکار میانی،

توازنی مناسب میان سادگی، سرعت و عملکرد فراهم می‌کند.

منابع و مراجع

[1] R. F. G.V. Haines, "Modeling by singular value decomposition and the elimination of statistically insignificant coefficients," *Computers & Geosciences,* pp. 19-28, 2013.

[2] J. W. S. B. &. V. V. Isabelle Guyon, "Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning," *Machine Learning,* 2002.

[3] I. Guyon, "AnIntroduction to Variable and Feature Selection," *Machine Learning Research 3,* 2003.

https://scikit-learn.org

https://tutorial24.ir/2025/10/11/mutual-information

پیوست‌ها

لینک گیت‌هاب پروژه:

https://github.com/ZahraBarati99/Feature-Selection-SVD-vs.-Classical