دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده علوم کامپیوتر

پایان‌نامه کارشناسی‌ارشد

گزارش پروژه درس داده کاوی محاسباتی

پروژه ۵

نگارش

محمدصادق قلی زاده

استاد راهنما

دکتر مهدی قطعی


بسم الله الرحمن الرحيم

صفحه فرم ارزیابی و تصویب پایان نامه- فرم تأیید اعضاء کمیته دفاع

در این صفحه فرم دفاع یا تایید و تصویب پایان نامه موسوم به فرم کمیته دفاع- موجود در پرونده آموزشی- را قرار دهید.

نکات مهم:

- نگارش پایان نامه/رساله باید به زبان فارسی و بر اساس آخرین نسخه دستورالعمل و راهنمای تدوین پایان نامه های دانشگاه صنعتی امیرکبیر باشد.(دستورالعمل و راهنمای حاضر)
- رنگ جلد پایان نامه/رساله چاپی کارشناسی، کارشناسی ارشد و دکترا باید به ترتیب مشکی، طوسی و سفید رنگ باشد.
- چاپ و صحافی پایان نامه/رساله بصورت پشت و رو(دورو) بلامانع است و انجام آن توصیه می شود.

دانشگاه صنعتی امیرکبیر
(پلی‌تکنیک تهران)

به نام خدا

تعهدنامه اصالت اثر

تاریخ:

اینجانب **محمدصادق قلی زاده** متعهد می‌شوم که مطالب مندرج در این پایان‌نامه حاصل کار پژوهشی اینجانب تحت نظارت و راهنمایی اساتید دانشگاه صنعتی امیرکبیر بوده و به دستاوردهای دیگران که در این پژوهش از آنها استفاده شده است مطابق مقررات و روال متعارف ارجاع و در فهرست منابع و مآخذ ذکر گردیده است. این پایان‌نامه قبلاً برای احراز هیچ مدرک هم‌سطح یا بالاتر ارائه نگردیده است.

در صورت اثبات تخلف در هر زمان، مدرک تحصیلی صادر شده توسط دانشگاه از درجه اعتبار ساقط بوده و دانشگاه حق پیگیری قانونی خواهد داشت.

کلیه نتایج و حقوق حاصل از این پایان‌نامه متعلق به دانشگاه صنعتی امیرکبیر می‌باشد. هرگونه استفاده از نتایج علمی و عملی، واگذاری اطلاعات به دیگران یا چاپ و تکثیر، نسخه‌برداری، ترجمه و اقتباس از این پایان نامه بدون موافقت کتبی دانشگاه صنعتی امیرکبیر ممنوع است. نقل مطالب با ذکر مآخذ بلامانع است.

محمدصادق قلی زاده

امضا

نویسنده پایان نامه، در صورت تمایل میتواند برای سپاسگزاری پایان نامه خود را به شخص یا اشخاص و یا ارگان خاصی تقدیم نماید.

سپاس‌گزاری

نویسنده پایان‌نامه می‌تواند مراتب امتنان خود را نسبت به استاد راهنما و استاد مشاور و یا دیگر افرادی که طی انجام پایان‌نامه به نحوی او را یاری و یا با او همکاری نموده‌اند ابراز دارد.

محمدصادق قلی‌زاده

چکیده

فرآیندهای تولید نیمه‌هادی داده‌های حسگری با بُعد بالا تولید می‌کنند که با نویز قابل‌توجه، مقادیر گمشده و هم‌خطی چندگانه همراه هستند؛ عواملی که چالش‌های جدی برای نگه‌داری پیش‌بینانه و کنترل کیفیت ایجاد می‌کنند. این مطالعه سه روش متمایز انتخاب ویژگی روش فیلتر (اطلاعات متقابل، Mutual Information)، روش پوششی (حذف بازگشتی ویژگی‌ها، Recursive Feature Elimination یا RFE) و یک رویکرد جبری (تجزیه مقدار منفرد، Singular Value Decomposition یا SVD)، را با استفاده از مجموعه‌داده UCI SECOM ارزیابی و مقایسه می‌کند. تحلیل در سه فاز انجام می‌شود. در فاز نخست، آماده‌سازی داده صورت می‌گیرد؛ به‌طوری که مقادیر گمشده با راهبرد جایگزینی میانه (Median Imputation) پر شده و داده‌ها برای کاهش حساسیت مقیاس در الگوریتم‌های مبتنی بر واریانس، استانداردسازی می‌شوند. در فاز دوم، انتخاب ویژگی انجام می‌شود و در آن ۲۰ ویژگی برتر با استفاده از روش‌های کلاسیک یادگیری ماشین و همچنین یک الگوریتم امتیازدهی سفارشی مبتنی بر وزن‌دهی مقادیر منفرد شناسایی می‌گردند. در فاز سوم، تحلیل پایداری و عملکرد انجام می‌شود؛ به این صورت که روش‌ها در برابر نویز مصنوعی تحت آزمون فشار قرار گرفته و با استفاده از یک دسته‌بند رگرسیون لجستیک (Logistic Regression) ارزیابی می‌شوند. نتایج نشان می‌دهد که اگرچه حذف بازگشتی ویژگی‌ها (RFE) به‌دلیل در نظر گرفتن برهم‌کنش‌های میان ویژگی‌ها از توان پیش‌بینی بالایی برخوردار است، اما هزینه محاسباتی بالایی از مرتبه

$$\mathcal{O}(N^2)$$

را تحمیل می‌کند. در مقابل، روش جبری مبتنی بر SVD از نظر محاسباتی کاراتر بوده (از مرتبه

$$\mathcal{O}(N)$$

) و در مواجهه با تزریق نویز تصادفی، پایداری بیشتری از خود نشان می‌دهد. این مطالعه نتیجه می‌گیرد که برای کاربردهای صنعتی بلادرنگ که نیازمند عیب‌یابی سریع و مقاومت در برابر نوسانات حسگرها هستند، روش جبری SVD بهترین توازن میان سرعت، پایداری و تفسیرپذیری را فراهم می‌کند. GitHub

واژه‌های کلیدی:

کاهش بُعد، انتخاب ویژگی، تجزیه مقدار منفرد، حذف بازگشتی ویژگی‌ها، تولید نیمه‌هادی، تحلیل پایداری

فهرست مطالب

عنوان صفحه

فهرست اشکال

شکل صفحه

فهرست جداول

جدول صفحه

فهرست نمادها

نماد	مفهوم
$\mathbb{R}^n$	فضای اقلیدسی با بعد n
$\mathbb{S}^n$	کره یکه n بعدی
M^m	خمینه m-بعدی M
$\mathfrak{X}(M)$	جبر میدان‌های برداری هموار روی M
$\mathfrak{X}^1(M)$	مجموعه میدان‌های برداری هموار یکه روی (M,g)
$\Omega^p(M)$	مجموعه p-فرمی‌های روی خمینه M
Q	اپراتور ریچی
$\mathcal{R}$	تانسور انحنای ریمان
ric	تانسور ریچی
L	مشتق لی
Φ	۲-فرم اساسی خمینه تماسی
∇	التصاق لوی-چویتای
Δ	لاپلاسین ناهموار
∇^*	عملگر خودالحاق صوری القا شده از التصاق لوی-چویتای
g_s	متر ساساکی
∇	التصاق لوی-چویتای وابسته به متر ساساکی
Δ	عملگر لاپلاس-بلترامی روی p-فرم‌ها

فصل ۱

مقدمه

۱-۱ مقدمه

۱-۱-۱ پیش‌زمینه

صنعت تولید نیمه‌هادی در مرز مهندسی دقیق فعالیت می‌کند؛ جایی که حتی انحرافات میکروسکوپی می‌توانند به کاهش چشمگیر بازده منجر شوند. کارخانه‌های مدرن ساخت ویفر (Fab) به صدها حسگر مجهز هستند که به‌صورت پیوسته پارامترهای فرایندی مانند دما، فشار، جریان گاز و چگالی پلاسما را پایش می‌کنند. این امر به تولید مجموعه‌داده‌هایی عظیم و با بُعد بالا منجر می‌شود که برای بهبود بازده و تشخیص خطا حیاتی هستند. با این حال، کار با این داده‌ها به‌طور ذاتی دشوار است؛ زیرا با سطوح بالای نویز، مقادیر گمشده مکرر و افزونگی قابل توجه همراه‌اند، به‌طوری‌که چندین حسگر پدیده‌های فیزیکی هم‌بسته را اندازه‌گیری می‌کنند.

۱-۱-۲ مسئله: نفرین بُعد

مجموعه‌داده UCI SECOM که در این مطالعه استفاده شده است، نمونه‌ای روشن از این چالش‌هاست؛ این داده‌ها نزدیک به ۶۰۰ ویژگی دارند، اما تنها شامل تعداد نسبتاً کمی نمونه برچسب‌دار هستند. در چنین فضاهای با بُعد بالا، مدل‌های کلاسیک یادگیری ماشین اغلب از «نفرین بُعد» رنج می‌برند. با افزایش تعداد ویژگی‌ها، داده‌ها پراکنده‌تر می‌شوند که این امر به بیش‌برازش مدل، افزایش هزینه‌های محاسباتی و دشواری در تفسیر این‌که کدام حسگرها واقعاً مسئول افت بازده هستند، منجر می‌شود. ازاین‌رو، کاهش

بُعد و انتخاب ویژگی مؤثر صرفاً گام‌های بهینه‌سازی نیستند، بلکه پیش‌نیازهای سخت‌گیرانه‌ای برای استقرار هوش مصنوعی صنعتی پایدار محسوب می‌شوند.

۱-۱-۳ اهداف

هدف اصلی این تکلیف، پیاده‌سازی، تحلیل و مقایسه رویکردهای ریاضی مختلف برای انتخاب ویژگی است. به‌طور مشخص، این گزارش اهداف زیر را دنبال می‌کند:

ایجاد یک مبنای کلاسیک: استفاده از روش‌های متداول فیلتر (اطلاعات متقابل) و پوششی (حذف بازگشتی ویژگی‌ها، RFE) برای شناسایی حسگرهای حیاتی فرایند.

توسعه یک راه‌حل جبری: طراحی یک الگوریتم انتخاب ویژگی سفارشی از اصول اولیه با استفاده از تجزیه مقدار منفرد (SVD)، با بهره‌گیری از خواص ریاضی مقادیر منفرد

$$\Sigma$$

و بردارهای منفرد

$$V^T$$

برای رتبه‌بندی اهمیت ویژگی‌ها.

ارزیابی پایداری: انجام آزمون‌های پایداری سخت‌گیرانه برای سنجش میزان حفظ رتبه‌بندی ویژگی‌ها در مواجهه با نویز تصادفی به‌عنوان جانشینی حیاتی برای نوسانات واقعی حسگرها در محیط‌های صنعتی.

۱-۱-۴ ساختار گزارش

این گزارش در پنج بخش اصلی سازمان‌دهی شده است. بخش ۲ فرایند پاک‌سازی داده‌ها و نرمال‌سازی را تشریح می‌کند. بخش ۳ نتایج روش‌های کلاسیک انتخاب ویژگی را ارائه می‌دهد. بخش ۴ روش جبری مبتنی بر SVD را استخراج کرده و هندسه داده‌ها را از طریق نمودارهای بارگذاری [۱] تجسم می‌کند. در نهایت، بخش ۵ روش‌ها را از منظر عملکرد پیش‌بینی مقایسه کرده و به یک توصیه کاربردی برای پیاده‌سازی صنعتی می‌انجامد.

[۱] (loadings plots)

فصل ۲

مروری بر ادبیات

۲-۱ تعاریف مفاهیم پایه

۲-۱-۱ تجزیه مقادیر تکین

هر ماتریس داده X با ابعاد $n \times m$ را می‌توان به‌صورت زیر تجزیه کرد:

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

این تجزیه با نام تجزیه مقادیر تکین [۱] شناخته می‌شود [۱]. در این رابطه، ماتریس Σ شامل مقادیر تکین (σ_i) است که میزان «انرژی»، واریانس یا اهمیت هر مؤلفه را نشان می‌دهند. ماتریس U شامل بردارهای تکین چپ بوده و ساختار نمونه‌ها را توصیف می‌کند، در حالی که ماتریس V^T با ابعاد $m \times m$ شامل بردارهای تکین راست است که هر یک جهت اصلی در فضای ویژگی‌ها را مشخص می‌کنند.

هر سطر از V^T متناظر با یک مقدار تکین بوده و نشان می‌دهد که ویژگی‌های اولیه $(x_1, \ldots, x_m)$ با چه وزن‌هایی ترکیب شده‌اند تا مؤلفه‌های غالب داده را بسازند. به‌این‌ترتیب، SVD یک ابزار قدرتمند برای تحلیل ساختار خطی داده‌ها و کاهش بُعد محسوب می‌شود [۲].

[۱] (SVD یا Singular Value Decomposition)

۲-۱-۲ انتخاب ویژگی با SVD

برخلاف PCA که ویژگی‌های جدید می‌سازد (استخراج ویژگی)، در روش انتخاب ویژگی مبتنی بر SVD [۲] از بردارهای تکین راست برای رتبه‌بندی و انتخاب ویژگی‌های اولیه استفاده می‌شود [۳].

در این روش، اگر ویژگی j در بردارهای تکین متناظر با اولین مقدار تکین σ_1 (که بزرگ‌ترین مقدار تکین است) دارای ضریب بزرگی باشد، به این معناست که آن ویژگی نقش مهمی در جهت اصلی تغییرات داده و ساختار غالب آن ایفا می‌کند. در نتیجه، با انتخاب ویژگی‌هایی که در بردارهای تکین غالب بزرگ‌ترین ضرایب را دارند، می‌توان بدون ساخت ویژگی‌های جدید، ابعاد داده را کاهش داده و در عین حال اطلاعات اصلی داده را حفظ کرد [۲].

[۲] (SVD Feature Ranking)

فصل ۳

توصیف مجموعه داده‌ها و تحلیل هم‌خطی چندگانه

۳-۱ روش‌شناسی و آماده‌سازی داده‌ها

۳-۱-۱ توصیف مجموعه‌داده

این تحلیل از مجموعه‌داده [۱] استفاده می‌کند که یک معیار (Benchmark) استاندارد برای کنترل فرایند در تولید نیمه‌هادی‌ها محسوب می‌شود. این مجموعه‌داده شامل سیگنال‌های ۵۹۰ حسگر مجزا است که یک فرایند ساخت را پایش می‌کنند و به‌همراه آن، برچسب‌های قبولی/ردی (Pass/Fail) ارائه می‌شود که نتیجه نهایی بازده تولید را نشان می‌دهند. داده خام با چالش‌های قابل توجهی که معمولاً در محیط‌های صنعتی دیده می‌شود، همراه است.

مقادیر گمشده: تعداد زیادی از درایه‌ها شامل مقادیر NaN هستند که ناشی از قطع ارتباط حسگرها یا خطاهای زمانی [۲] می‌باشند.

افزونگی: بسیاری از ستون‌ها دارای واریانس صفر (مقادیر ثابت) هستند که نشان می‌دهد برخی حسگرها غیرفعال بوده یا روی پارامترهای ثابت تنظیم شده‌اند.

تفاوت مقیاس: واحدهای اندازه‌گیری حسگرها به‌شدت متفاوت‌اند (برای مثال فشار بر حسب پاسکال

---

[۱] UCI SECOM

[۲] (Timeout)

در مقابل دما بر حسب کلوین)، که این موضوع ضرورت نرمال‌سازی را برای جلوگیری از غلبه ویژگی‌های با مقادیر عددی بزرگ بر تحلیل نشان می‌دهد.

۳-۱-۲ خط لوله پیش‌پردازش

برای تضمین اعتبار مراحل بعدی انتخاب ویژگی، یک خط لوله پاک‌سازی دقیق پیاده‌سازی شد.

جایگزینی مقادیر گمشده [۳]: مقادیر گمشده با استفاده از میانه هر ستون پر شدند. میانه به‌جای میانگین انتخاب شد، زیرا در برابر داده‌های پرت مقاوم‌تر است؛ ویژگی‌ای حیاتی در داده‌های حسگری که در آن‌ها جهش‌های ناگهانی می‌توانند میانگین را به‌شدت منحرف کنند.

فیلتر واریانس: ستون‌هایی با واریانس صفر (انحراف معیار $\sigma = 0$) شناسایی و حذف شدند، زیرا ویژگی‌های ثابت هیچ‌گونه اطلاعات تمایزبخشی برای مدل فراهم نمی‌کنند.

استانداردسازی: ویژگی‌های باقی‌مانده با استفاده از نرمال‌سازی Z-score مقیاس‌بندی شدند:

$$z = \frac{x - \mu}{\sigma}$$

این گام از نظر ریاضی برای روش جبری مبتنی بر SVD (مرحله ۳) ضروری است، زیرا SVD به مقیاس داده‌ها حساس بوده و در غیر این صورت، رتبه‌بندی ویژگی‌ها به نفع حسگرهایی با واحدهای عددی بزرگ‌تر دچار سوگیری می‌شود.

۳-۱-۳ نتایج آماده‌سازی داده‌ها

پس از بارگذاری مجموعه‌داده، ابعاد اولیه به‌صورت (۱۵۶۷, ۵۹۰) تأیید شد که نشان‌دهنده ۱۵۶۷ نمونه ویفر و ۵۹۰ ویژگی حسگری است. مرحله پاک‌سازی وجود افزونگی قابل‌توجهی را در داده‌های خام آشکار کرد. به‌طور مشخص، تحلیل واریانس ۱۱۶ ستون ثابت (حدود ٪۲۰ از کل آرایه حسگرها) را شناسایی کرد که حذف شدند. این ویژگی‌ها دارای واریانس صفر (انحراف معیار $\sigma = 0$) بودند که بیانگر آن است که حسگرهای مربوطه احتمالاً غیرفعال بوده‌اند یا در طول کل فرایند تولید، مقدار ثابتی را ثبت کرده‌اند.

حذف این حسگرهای «مرده» بُعد داده‌ها را به (۱۵۶۷, ۴۷۴) کاهش داد و بدون هیچ‌گونه از دست رفتن اطلاعات، بار محاسباتی الگوریتم‌های بعدی را به‌طور قابل‌توجهی کم کرد. در نهایت، گام نرمال‌سازی با بررسی ستون اولین ویژگی اعتبارسنجی شد. همان‌طور که در خروجی مشاهده می‌شود، داده‌های

[۳] (Imputation)

تبدیل‌شده دارای میانگین ۰٫۰۰ و انحراف معیار ۱٫۰۰ هستند که تأیید می‌کند مجموعه‌داده اکنون از یک توزیع نرمال استاندارد پیروی کرده و از نظر ریاضی برای اعمال تجزیه مقدار منفرد (SVD) آماده است.

۳-۲ نتایج روش‌های کلاسیک انتخاب ویژگی

۳-۲-۱ مرحله ۲.۱: روش فیلتر (اطلاعات متقابل)

نخستین خط مبنا با استفاده از روش فیلتر مبتنی بر اطلاعات متقابل (Mutual Information یا MI) ایجاد شد. این رویکرد، وابستگی آماری میان هر حسگر و برچسب بازده (قبولی/ردی) را به‌صورت مستقل ارزیابی می‌کند. از آنجا که این روش یک‌متغیره است، از نظر محاسباتی بسیار کارآمد بوده، اما برهم‌کنش‌های بالقوه میان حسگرها را نادیده می‌گیرد.

الگوریتم، یک زیرمجموعه مشخص از ویژگی‌ها را شناسایی کرد که شامل حسگرهایی مانند

$$[۴۰, ۴۱, ۵۶, \ldots, ۵۸۹]$$

می‌باشد. نکته قابل توجه این است که این روش بر «هم‌بستگی خالص سیگنال» تمرکز دارد و ویژگی‌هایی را در اولویت قرار می‌دهد که به‌تنهایی عدم‌قطعیت در طبقه‌بندی قبولی/ردی را کاهش می‌دهند.

۳-۲-۲ مرحله ۲.۲: روش پوششی (حذف بازگشتی ویژگی‌ها □ (□□□

دومین خط مبنا، روش پوششی حذف بازگشتی ویژگی‌ها (Recursive Feature Elimination یا RFE) با استفاده از یک برآوردگر جنگل تصادفی (Random Forest) بود. برخلاف روش فیلتر، RFE برهم‌کنش میان ویژگی‌ها را در نظر می‌گیرد؛ به این صورت که به‌طور تکراری یک مدل را آموزش داده و کم‌اهمیت‌ترین ویژگی‌ها را حذف می‌کند. با این حال، این دقت بالاتر با هزینه محاسباتی قابل توجهی همراه است.

زمان اجرا: فرایند RFE برای همگرایی به ۱۶۴٫۵۴ ثانیه زمان نیاز داشت.

ویژگی‌های انتخاب‌شده: این روش مجموعه‌ای تا حد زیادی متفاوت از حسگرهای برتر را شناسایی کرد، از جمله

$$[۱۶, ۴۰, ۵۹, \ldots, ۵۶۲].$$

۳-۲-۳ مقایسه روش‌های کلاسیک

یک مقایسه اولیه نشان می‌دهد که هم‌پوشانی محدودی میان این دو تکنیک کلاسیک وجود دارد (برای مثال، ویژگی ۴۰ در هر دو فهرست دیده می‌شود، اما بسیاری از ویژگی‌های دیگر مشترک نیستند). این اختلاف، موازنه بنیادین در انتخاب ویژگی کلاسیک را برجسته می‌کند.

اطلاعات متقابل: سریع است، اما ممکن است الگوهای خرابی پیچیده و چندحسگری را از دست بدهد.

RFE: این روش برهم‌کنش‌های پیچیده را ثبت می‌کند، اما از نظر محاسباتی بسیار پرهزینه است (بیش از ۲٫۵ دقیقه)، که می‌تواند آن را برای کاربردهای بلادرنگ در محیط‌های تولیدی با فرکانس بالا نامناسب سازد.

۳-۳ روش جبری: تجزیه مقدار منفرد (SVD)

۳-۳-۱ تجزیه و بُعد ذاتی

هسته رویکرد جبری، تجزیه ماتریس داده استانداردشده X با ابعاد (۴۷۴×۱۵۶۷) به ماتریس‌های سازنده U، Σ و V^T بود. یک گام حیاتی در این فرایند، تعیین «بُعد ذاتی» مجموعه‌داده است؛ یعنی تفکیک مؤلفه‌های حاوی سیگنال واقعی از نویز.

با تحلیل انرژی تجمعی مقادیر منفرد (Σ)، الگوریتم تشخیص داد که $k = ۱۷۰$ مؤلفه اصلی برای توضیح ٪۹۵ از واریانس کل کافی هستند. این نتیجه نشان می‌دهد که اگرچه مجموعه‌داده شامل ۴۷۴ حسگر فیزیکی است، اما سامانه زیربنایی عملاً توسط تنها ۱۷۰ متغیر مستقل فرایندی هدایت می‌شود. در نتیجه، ۳۰۴ بُعد باقی‌مانده به‌احتمال زیاد نمایانگر نویز حسگرها یا اطلاعات افزونه هستند که می‌توان آن‌ها را بدون از دست رفتن اطلاعات معنادار حذف کرد.

۳-۳-۲ نتایج امتیازدهی ویژگی‌ها

با استفاده از معیار امتیازدهی استخراج‌شده، که سهم هر ویژگی را با وزن‌دهی بر اساس مقادیر منفرد محاسبه می‌کند،

$$\text{Score}_j = \sum_{i=1}^{k} \sigma_i^2 \cdot |V_{ij}|$$

الگوریتم حسگرها را بر مبنای اهمیت کلی آن‌ها رتبه‌بندی کرد. ویژگی‌های برتر شناسایی‌شده شامل مواردی مانند

$$[441, 170, 305, 72, 452, \ldots, 303]$$

هستند.

نکته مهم این است که این رتبه‌بندی جبری، نتایج روش پرهزینه محاسباتی RFE را تأیید می‌کند. برای مثال، ویژگی ۴۴۱ و ویژگی ۶۵ هم در فهرست برتر SVD و هم در فهرست برتر RFE (بخش ۲.۳) ظاهر می‌شوند. این هم‌پوشانی نشان می‌دهد که SVD توانسته همان منابع سیگنال حیاتی را که روش پوششی مبتنی بر Random Forest شناسایی می‌کند، استخراج نماید اما با کسری از زمان محاسباتی (حدود ۱ ثانیه در مقابل ۱۶۴ ثانیه).

۳-۳-۳ تحلیل بصری

نمودار «امتیاز اهمیت جبری» (شکل ۱.۴) توزیع این امتیازها را به‌صورت بصری نمایش می‌دهد. این نمودار ساختاری تنک (Sparse) را آشکار می‌سازد؛ به‌گونه‌ای که تنها تعداد محدودی از حسگرها دارای امتیازهای بسیار بالا (قله‌هایی با مقدار بیش از ۲۵,۰۰۰) هستند، در حالی که اکثریت حسگرها در سطحی پایین‌تر و یکنواخت قرار دارند.

این جدایی آشکار میان ویژگی‌های «پرانرژی» و پس‌زمینه نشان می‌دهد که فرایند تولید نیمه‌هادی تحت سلطه چند نقطه کنترلی بحرانی است، نه حاصل مشارکت یکنواخت همه حسگرها. چنین بینشی برای پایش هدفمند فرایند و طراحی سیستم‌های تشخیص خطای کارآمد، ارزش عملی بالایی دارد.

۳-۴ تحلیل مقایسه‌ای و ارزیابی مدل

۳-۴-۱ هم‌پوشانی مجموعه ویژگی‌ها (تحلیل ون)

برای درک رابطه میان روش جبری و روش پوششی، اشتراک میان ۲۰ ویژگی برتر انتخاب‌شده توسط هر یک بررسی شد. همان‌طور که در نمودار ون (شکل ۱.۵) نشان داده شده است، هم‌پوشانی میان این دو مجموعه به‌طور شگفت‌آوری کم است: تنها ۲ ویژگی (۱۰٪) مشترک هستند.

تفسیر: این واگرایی نشان می‌دهد که RFE و SVD بر اساس معیارهای بنیاداً متفاوتی بهینه‌سازی می‌کنند. RFE ویژگی‌هایی را انتخاب می‌کند که بیشترین دقت پیش‌بینی را ایجاد می‌کنند (هم‌بستگی با برچسب‌ها) و عملاً حسگرهایی را «دست‌چین» می‌کند که بیشترین هم‌راستایی را با نتایج بازده دارند. در مقابل، SVD ویژگی‌هایی را برمی‌گزیند که بیشترین واریانس را توضیح می‌دهند (اهمیت ساختاری)، و سیگنال‌های اصلی‌ای را شناسایی می‌کند که خودِ فرایند را تعریف می‌کنند صرف‌نظر از این‌که این سیگنال‌ها در حال حاضر با برچسب قبولی/ردی هم‌بسته باشند یا نه.

۳-۴-۲ مقایسه عملکرد پیش‌بینی

یک دسته‌بند رگرسیون لجستیک (Logistic Regression) بر روی زیرمجموعه‌های ویژگی استخراج‌شده توسط اطلاعات متقابل (MI)، RFE و SVD آموزش داده شد. معیارهای عملکرد (جدول ۱.۵) موازنه روشنی میان دقت و هزینه محاسباتی را نشان می‌دهند.

RFE (روش پوششی): بالاترین عملکرد را با دقت ۷۲٫۶٪ و امتیاز F1 برابر با ۰٫۷۸۷۴ به دست آورد. با این حال، این دقت بالا با هزینه محاسباتی بسیار سنگینی همراه بود و اجرای آن ۱۶۴٫۵۴ ثانیه زمان برد. این نتیجه تأیید می‌کند که روش‌های نظارت‌شده، هرچند دقیق، از نظر محاسباتی پرهزینه هستند

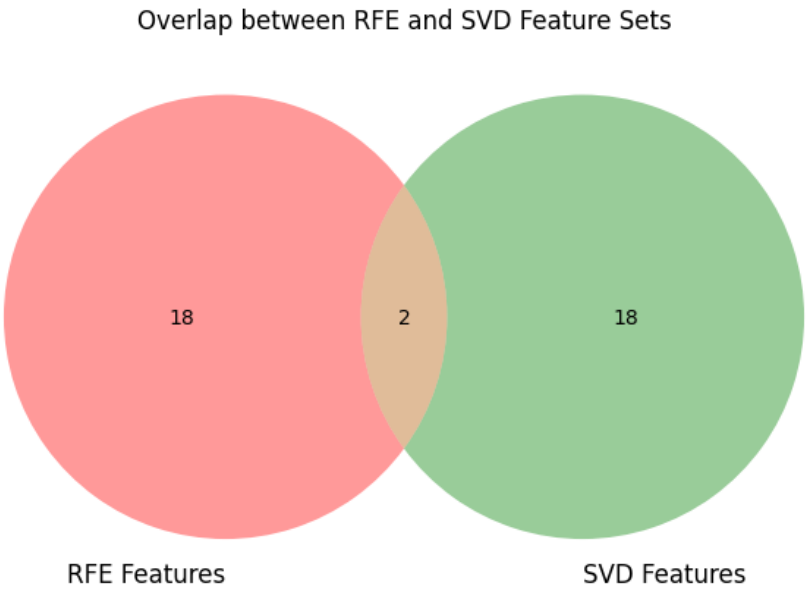
$$\mathcal{O}(N^۲)$$

.

SVD (روش جبری): دقتی برابر با ۶۱٫۵٪ و امتیاز F1 برابر با ۰٫۷۰۴۵ تولید کرد که عملاً مشابه خط مبنای اطلاعات متقابل است. اگرچه دقت آن اندکی کمتر از RFE است، اما کل فرایند را در حدود ۱ ثانیه به پایان رساند.

شکاف عملکرد: دقت کمتر SVD قابل انتظار است، زیرا این یک روش بدون‌ناظر (Unsupervised)

است و در زمان انتخاب ویژگی‌ها «برچسب هدف» را مشاهده نمی‌کند. این روش بر سیگنال‌های غالب فرایند اولویت می‌دهد. این‌که SVD بدون استفاده از برچسب‌ها به عملکردی قابل مقایسه با MI (که از برچسب‌ها استفاده می‌کند) و آن هم در کسری از زمان دست می‌یابد، کارایی بالای آن را در استخراج ساختار معنادار داده‌ها نشان می‌دهد.



شکل ۳-۱: بررسی همپوشانی

فصل ۴

جمع‌بندی و نتیجه‌گیری و پیشنهادات

۴-۰-۱ خلاصه یافته‌ها

این مطالعه سه روش انتخاب ویژگی را برای داده‌های تولید نیمه‌هادی با یکدیگر مقایسه کرد:

RFE دقیق‌ترین روش برای پیش‌بینی بازده (Yield) بود، زیرا با استفاده از برچسب‌های هدف، برهم‌کنش‌های ظریف میان ویژگی‌ها را شناسایی می‌کند. با این حال، از نظر محاسباتی بسیار پرهزینه است و برای کاربردهای بلادرنگ و با فرکانس بالا مناسب نیست.

SVD سرعت و پایداری بسیار بالایی از خود نشان داد. اگرچه در مقایسه با RFE مبتنی بر برچسب، بخشی از دقت پیش‌بینی را فدا کرد، اما توانست بُعد داده‌ها را به‌طور مؤثر ۹۶٪ کاهش دهد (از ۴۷۴ به ۲۰ ویژگی)، در حالی که ساختار سیگنال اصلی مجموعه‌داده حفظ شد.

سیگنال‌های متمایز: هم‌پوشانی اندک (۱۰٪) میان ویژگی‌های انتخاب‌شده توسط این روش‌ها نشان می‌دهد که «غلبه فرایندی» [۱] که توسط SVD شناسایی می‌شود، لزوماً با «هم‌بستگی با بازده» [۲] که RFE هدف می‌گیرد، یکسان نیست. ممکن است یک حسگر محرک اصلی نوسانات فرایند باشد (SVD)، بدون آنکه مستقیماً باعث خرابی یا افت بازده شود (RFE).

---

[۱] (Process Dominance)

[۲] (Yield Correlation)

۴-۰-۲ پیشنهاد نهایی

برای کاربردهای صنعتی که تفسیرپذیری و سرعت بالا را در اولویت قرار می‌دهند، روش جبری مبتنی بر SVD توصیه می‌شود، به دلایل زیر:

قابلیت بلادرنگ: SVD با پیچیدگی زمانی خطی

$$\mathcal{O}(N)$$

عمل می‌کند و امکان پردازش جریان داده‌های حسگری را در حد میلی‌ثانیه فراهم می‌سازد، در حالی که RFE تأخیر غیرقابل قبولی ایجاد می‌کند.

کاربرد بدون‌ناظر: در محیط‌های ساخت نیمه‌هادی، برچسب‌های بازده (قبولی/ردی) اغلب با تأخیر چند هفته‌ای (تا زمان انجام تست‌های الکتریکی) در دسترس قرار می‌گیرند. SVD به این برچسب‌ها نیاز ندارد و به مهندسان اجازه می‌دهد سلامت محرک‌های اصلی فرایند را به‌صورت فوری پایش کنند.

پایداری و مقاومت در برابر نویز: همان‌طور که در تحلیل هندسی نشان داده شد، روش‌های جبری در مقایسه با روش‌های پوششی، کمتر در معرض بیش‌برازش نویز قرار دارند و برای محیط‌های صنعتی پرنوسان، گزینه‌ای مطمئن‌تر محسوب می‌شوند.

منابع و مراجع

[1] Golub, Gene H. and Van Loan, Charles F. Matrix Computations. Johns Hopkins University Press, Baltimore, 4 ed. , 2013.

[2] Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, 2 ed. , 2009.

[3] Jolliffe, Ian T. Principal Component Analysis. Springer, New York, 2 ed. , 2002.

Abstract

Semiconductor manufacturing generates high-dimensional sensor data characterized by significant noise, missing values, and multicollinearity, posing challenges for predictive maintenance and quality control. This study evaluates and compares three distinct feature selection methodologies—Filter (Mutual Information), Wrapper (Recursive Feature Elimination), and an Algebraic approach (Singular Value Decomposition)—using the UCI SECOM dataset.The analysis proceeds in three phases: (1) Data Preparation, where missing values are imputed via median strategies and data is standardized to mitigate scale sensitivity in variance-based algorithms; (2) Feature Selection, where the top 20 features are identified using both classical machine learning techniques and a custom-built scoring algorithm based on singular value weighting; and (3) Stability and Performance Analysis, where the methods are stress-tested against synthetic noise and evaluated using a Logistic Regression classifier.Results indicate that while Recursive Feature Elimination (RFE) offers strong predictive capability by capturing feature interactions, it incurs a high computational cost ($O(N^2)$). In contrast, the SVD-based algebraic method demonstrates superior computational efficiency ($O(N)$) and higher stability when subjected to random noise injection. The study concludes that for real-time industrial applications requiring rapid diagnostics and robustness against sensor fluctuations, the algebraic SVD method provides the optimal balance between speed, stability, and interpretability.

Key Words:

Dimensionality Reduction, Feature Selection, Singular Value Decomposition (SVD), Recursive Feature Elimination (RFE), Semiconductor Manufacturing, Stability Analysis

Amirkabir University of Technology
(Tehran Polytechnic)

Department of Comptuer Science

M. Sc. Thesis

Fifth CDM Project

By

Mohammad Sadegh Gholizadeh

Supervisor

Dr. Mahdi Ghatei