دانشگاه صنعتی امیرکبیر
(پلی‌تکنیک تهران)
دانشکده ریاضی و علوم کامپیوتر

داده کاوی محاسباتی

گزارش ۵: انتخاب ویژگی با استفاده از مقادیر تکین (SVD) در برابر روش‌های کلاسیک

نگارش
محمدامین نعمتی

شماره دانشجویی
۴۰۴۱۲۴۹۰۴

استاد درس
آقای دکتر قطعی

نیمسال اول ۱۴۰۵-۱۴۰۴

چکیده

در این پروژه روش جبری مبتنی بر تجزیه مقادیر تکین (SVD) برای انتخاب ویژگی در مجموعه‌داده صنعتی SECOM پیاده‌سازی و با روش‌های کلاسیکِ فیلتر مبتنی بر اطلاعات-متقابل و روش پوششی حذف بازگشتی ویژگی‌ها (RFE) مقایسه شد. پس از پیش‌پردازش (حذف ستون‌های ثابت، جایگزینی میانه و استانداردسازی)، امتیازدهی ویژگی‌ها با وزن‌دهی ضرایب بردارهای تکین راست توسط مجذور مقادیر تکین انجام گردید و ۲۰ ویژگی برتر انتخاب شد. مقایسه کارایی با رگرسیون لجستیک نشان داد RFE بالاترین دقت پیش‌بینی را دارد، اما روش SVD بسیار سریع‌تر، تفسیرپذیرتر و در آزمون پایداری در برابر افزودن ۵٪ نویز مقاوم‌تر بود؛ هم‌پوشانی مجموعه‌های منتخب نیز پایین بود که نشان‌دهنده دیدگاهِ متفاوت SVD بر ساختار هندسی داده‌هاست. بنابراین SVD ابزاری مکمل و مناسب برای تحلیل ساختاری و انتخاب ویژگی در مسائل صنعتی با نویز بالا و نیاز به تفسیرپذیری است.

واژه‌های کلیدی

گزینش ویژگی، تجزیه تکین، داده‌های صنعتی، مقاومت در برابر نویز و تفسیرپذیری

فهرست مطالب

صفحه

چکیده ..... أ
فصل اول ۱
فصل دوم ۳
فصل سوم ۵
فصل چهارم ۷
فصل پنجم ۹
فصل ششم نتیجه‌گیری ۱۱
Abstract ۱۳

فصل اول

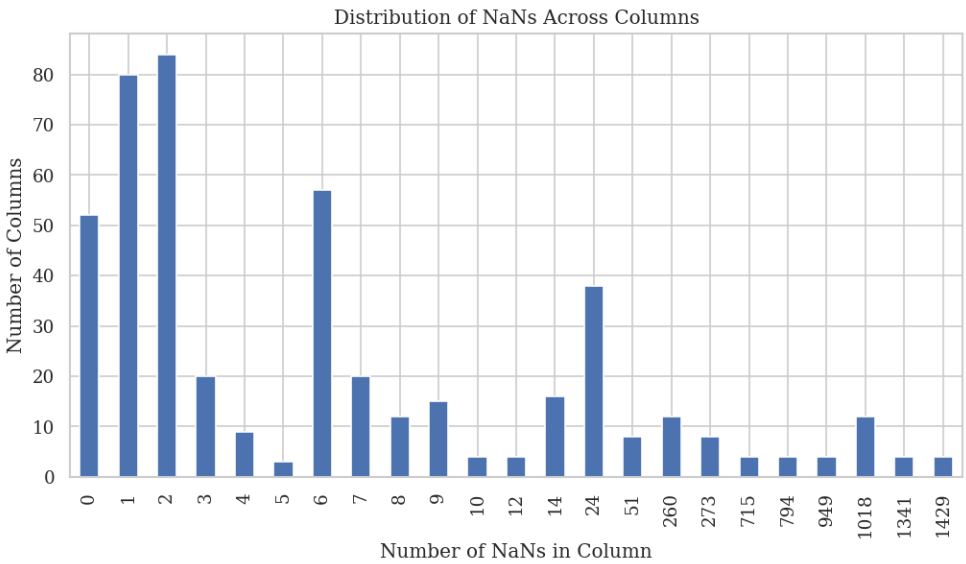
مقدمه

با گسترش روزافزون سامانه‌های صنعتی مبتنی بر حسگرها و تولید حجم عظیمی از داده‌های چندبعدی، مسئله گزینش ویژگی به یکی از چالش‌های اساسی در تحلیل داده‌های صنعتی تبدیل شده است. وجود نویز، مقادیر گمشده، هم‌بستگی شدید بین متغیرها و ابعاد بالای داده‌ها سبب می‌شود که استفاده مستقیم از همه ویژگی‌ها نه‌تنها هزینه محاسباتی بالایی داشته باشد، بلکه تفسیرپذیری و پایداری مدل‌های یادگیری را نیز کاهش دهد. ازاین‌رو، انتخاب زیرمجموعه‌ای مؤثر از ویژگی‌ها که بتواند ساختار اصلی داده را با کمترین اتلاف اطلاعات حفظ کند، اهمیت ویژه‌ای دارد. در این پژوهش، ضمن بررسی روش‌های کلاسیک گزینش ویژگی، تمرکز اصلی بر رویکردهای جبری و به‌ویژه تجزیه مقادیر تکین قرار داده شده است تا نشان داده شود چگونه تحلیل ریاضی ماتریس داده‌ها می‌تواند به درک عمیق‌تر از ساختار هندسی داده‌های صنعتی و انتخاب ویژگی‌های معنادار، پایدار و قابل تفسیر منجر شود.

فصل دوم

آماده‌سازی داده‌های صنعتی دیتاست (SECOM)

در این بخش داده‌ها از فایل‌های مرجع SECOM بارگذاری شد، ستون‌های ثابت با واریانس صفر حذف شدند و توزیع مقادیر گمشده در سطوح ستون و سطر بررسی گردید. ستون‌هایی که بیش از آستانه تعیین‌شده ۵۲ مقدار NaN داشتند حذف و برای بقیه داده‌ها از Median Imputation استفاده شد تا اثر مقادیر گمشده بر تحلیل‌ها کاهش یابد.



توزیع مقادیر گمشده در ستون‌های داده

سپس متغیرها با StandardScaler استاندارد شدند تا سنسورهای با مقیاس بزرگ، تأثیر مصنوعی بر مقادیر تکین نگذارند. حاصلِ پیش‌پردازش مجموعه‌داده‌ای با شکل نهایی ۱۵۶۷ نمونه و ۴۲۳ ویژگی بود که در جدول زیر خلاصه شده است.

خلاصه دیتاست پس از پیش‌پردازش

	مقدار
تعداد نمونه‌ها (ردیف‌ها)	1567
تعداد ویژگی‌ها (ستون‌ها) پس از حذف ستون‌های ثابت	423
روش جایگزینی مقادیر گمشده	Median Imputation
نرمال‌سازی	StandardScaler
یادداشت	ستون‌های با بیش از ۵۲ NaN حذف شدند؛ بقیه با میانه پر شدند.

فصل سوم

روش‌های کلاسیک (مبنای مقایسه)

برای مقایسه معیار پایه، ابتدا از روش فیلتر مبتنی بر اطلاعات-متقابل (mutual_info_classif) استفاده شد و بیست ویژگی برتر بر اساس امتیاز MI استخراج گردیدند.

۲۰ ویژگی برتر بر اساس Information Mutual (MI)

امتیاز(MI)	ویژگی
0.028769	**41**
0.027760	**331**
0.027203	**577**
...	**...**
0.020657	**443**
0.020425	**405**
0.019409	**33**

سپس برای روش پوششی، الگوریتم RFE با الگوریتم پایه RandomForest (n_estimators=50) و پارامتر step=8 اجرا شد تا ۲۰ ویژگی نهایی انتخاب شود؛ زمان اجرای RFE و فهرست ویژگی‌های منتخب در جدول سوم ارائه شده است. این نتایج به‌عنوان مبنایی برای مقایسه با روش جبری مبتنی بر SVD و برای سنجش پایداری و همپوشانی بین روش‌ها به‌کار خواهند رفت.

نتایج RFE با RandomForest (n_estimators=50, step=8)

	مقدار
فهرست ۲۰ ویژگی منتخب (شاخص‌ها)	[2, 16, 25, 26, 38, 59, 64, 65, 71, 102, 103, 132, 153, 301, 348, 426, 473, 476, 477, 574]
زمان اجرای RFE (ثانیه)	106.84781009900001

فصل چهارم

روش جبری (رتبه‌بندی با مقادیر تکین)

در این بخش، روش جبری مبتنی بر تجزیه مقادیر تکین به‌عنوان هسته ریاضی تحلیل به‌کار گرفته شد تا بدون استفاده از توابع آماده انتخاب ویژگی، ساختار خطی پنهان در داده‌های صنعتی SECOM استخراج شود؛ بدین منظور، ماتریس داده‌ها پس از حذف ستون برچسب به‌صورت صریح تجزیه شد و با تمرکز بر مقادیر تکین بزرگ‌تر که بیانگر مؤلفه‌های با انرژی بالاتر هستند، نقش هر ویژگی در شکل‌دهی ساختار اصلی داده مورد ارزیابی قرار گرفت. امتیاز هر ویژگی با در نظر گرفتن ضرایب آن در بردارهای تکین راست و وزن‌دهی این ضرایب توسط مربع مقادیر تکین ده مؤلفه اول محاسبه شد تا سهم مؤلفه‌های قوی‌تر برجسته گردد. نتایج نشان داد که تنها تعداد محدودی از ویژگی‌ها دارای امتیازهای بسیار بزرگ‌تری نسبت به سایرین هستند که این امر بیانگر نقش غالب آن‌ها در مؤلفه‌های اصلی و ساختار هندسی داده است؛ بر همین اساس، بیست ویژگی با بالاترین امتیاز به‌عنوان مجموعه نهایی منتخب روش SVD در نظر گرفته شدند.

ده ویژگی برتر بر اساس رتبه‌بندی جبری مبتنی برSVD

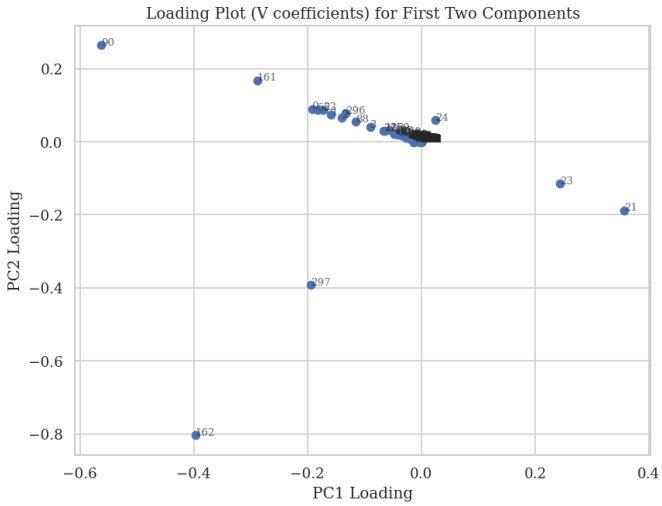
رتبه	شاخص ویژگی	مقدار امتیاز
1	90	2.322158×10^{11}
2	162	2.030112×10^{11}
3	21	1.494461×10^{11}
4	161	1.431010×10^{11}
5	23	1.054765×10^{11}
6	297	9.940024×10^{10}
7	0	7.914316×10^{10}
8	55	7.502060×10^{10}
9	22	7.146835×10^{10}
10	296	6.623470×10^{10}

فصل پنجم

تحلیل هندسی و ارزیابی پایداری

در این بخش، تحلیل هندسی با رسم نمودار بارگذاری برای دو مؤلفه اول انجام شد تا موقعیت هر ویژگی در فضای مؤلفه‌ها و خوشه‌بندی احتمالی سنسورها بررسی شود.

سپس آزمون پایداری با افزودن نویز تصادفی سطح ۵٪ به داده‌ها انجام شد و عملکرد انتخاب ویژگی با روش پوششی (حذف بازگشتی ویژگی‌ها) و روش جبری (مبتنی بر تجزیه مقادیر تکین) بر روی داده نویزی ارزیابی شد تا تعداد ویژگی‌های تغییر یافته برای هر روش محاسبه گردد.

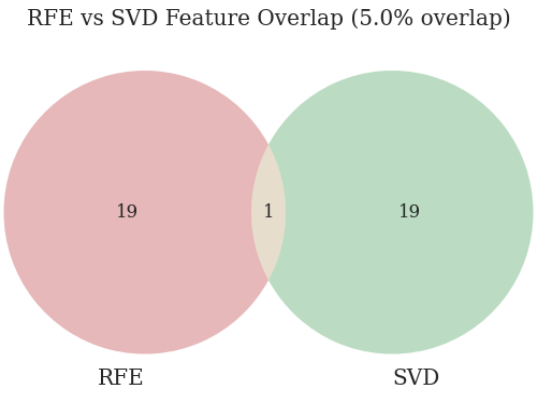


تعداد ویژگی‌های تغییر یافته	روش
15	حذف بازگشتی ویژگی‌ها (RFE)
0	مبتنی بر تجزیه مقادیر تکین (SVD)

فصل ششم

مقایسه نهایی و ارزیابی

در این بخش سه مجموعه ویژگی منتخب از سه روش با یک مدل رگرسیون لجستیک ساده بر روی هر زیرداده آموزش و ارزیابی شده‌اند و معیارهای دقت، معیار F1 و زمان انتخاب ویژگی گزارش گردیده‌اند. همچنین میزان همپوشانی بین مجموعه‌های منتخَب بررسی شده و نمودار ون تهیه شده است. نتایج عددی استخراج‌شده از اجرای کد در جداول زیر آورده شده‌اند. انتخاب روش به نیاز صنعتی بستگی دارد: برای سرعت و تفسیرپذیری، روش فیلتر مناسب است؛ برای بالاترین دقت، روش پوششی بهتر عمل می‌کند؛ و روش جبری مبتنی بر SVD برای تحلیل ساختار هندسی و پایداری در برابر نویز ارزشمند است.



خلاصه نهایی روش‌های انتخاب ویژگی، زمان، عملکرد و هم‌پوشانی

روش انتخاب ویژگی	نوع روش	زمان انتخاب	دقت	معیار F1	اشتراک با SVD	اشتراک
اطلاعات-متقابل	فیلتر	7.502350	0.668790	0.161290	0	0.0٪
حذف بازگشتی ویژگی‌ها(RFE)	پوششی	106.847810	0.770701	0.289474	1	5.0٪
مبتنی بر تجزیه مقادیر تکین(SVD)	جبری	0.140207	0.613588	0.133333	—	—

جمع‌بندی عملکرد و کاربردپذیری روش‌ها

معیار	روش فیلتر	روش پوششی	روش جبری(SVD)
سرعت انتخاب	سریع	کند	بسیار سریع
تفسیرپذیری	بالا	متوسط	خیلی خوب
پایداری در نویز	معمولاً خوب	حساس	مقاوم‌تر
مناسب برای صنعت	سرعت و تفسیر	بالاترین دقت	تحلیل ساختاری

Abstract

In this project, an algebraic method based on Singular Value Decomposition (SVD) was implemented for feature selection on the industrial SECOM dataset and compared with classical filter methods based on mutual information and the wrapper method Recursive Feature Elimination (RFE). After preprocessing (removal of constant columns, median imputation, and standardization), feature scoring was performed by weighting the coefficients of the right singular vectors by the squared singular values, and the top 20 features were selected. Performance comparison using logistic regression showed that RFE achieved the highest predictive accuracy; however, the SVD-based method was significantly faster, more interpretable, and more robust in stability tests under the addition of 5% noise. The low overlap among the selected feature sets indicates that SVD provides a distinct perspective on the geometric structure of the data. Therefore, SVD serves as a complementary and suitable tool for structural analysis and feature selection in industrial problems characterized by high noise levels and a need for interpretability.

Key Words Feature Selection, Noise Robustness, Interpretability

GITHUB LINK

https://github.com/nematiamin57-pixel/data_mining1404

Amirkabir University of Technology
(Tehran Polytechnic)

Department of Computer Science

Computational Data Mining

Report 5: Feature selection using singular values (SVD) versus classical methods

By
Mohammadamin Nemati

Student ID
404124904

Lecturer
Dr. Ghatee

First Semester of the 2025–2026 Academic Year