دانشگاه صنعتي اميركبير
(پلی‌تکنیک تهران)
دانشکده ریاضی و علوم کامپیوتر

پروژه پنجم درس داده کاوی محاسباتی
علوم کامپیوتر، هوش مصنوعی و محاسبات نرم

انتخاب ویژگی با استفاده از مقادیر تکین در برابر روش های کلاسیک

نگارش
هومن ذوالفقاری

استاد راهنما
دکتر مهدی قطعی

آذر 1404

به نام خدا

تاریخ:

تعهدنامه اصالت اثر

اینجانب هومن ذوالفقاری متعهد می‌شوم که مطالب مندرج در این پایان نامه حاصل کار پژوهشی اینجانب تحت نظارت و راهنمایی اساتید دانشگاه صنعتي اميركبير بوده و به دستاوردهاي ديگران كه در اين پژوهش از آنها استفاده شده است مطابق مقررات و روال متعارف ارجاع و در فهرست منابع و مآخذ ذكر گرديده است. اين پايان نامه قبلاً براي احراز هيچ مدرك هم‌سطح يا بالاتر ارائه نگرديده است.

در صورت اثبات تخلف در هر زمان، مدرك تحصيلي صادر شده توسط دانشگاه از درجه اعتبار ساقط بوده و دانشگاه حق پيگيري قانوني خواهد داشت.

كليه نتايج و حقوق حاصل از اين پايان نامه متعلق به دانشگاه صنعتي اميركبير مي‌باشد. هرگونه استفاده از نتايج علمي و عملي، واگذاري اطلاعات به ديگران يا چاپ و تكثير، نسخه‌برداري، ترجمه و اقتباس از اين پايان نامه بدون موافقت كتبي دانشگاه صنعتي اميركبير ممنوع است.
نقل مطالب با ذكر مآخذ بلامانع است.

در صفحه تعهدنامه اصالت اثر، در قسمت بالا سمت چپ، تاریخ دفاع خود را جایگزین تاریخ نوشته شده کنید.

همچنین در صفحه تعهدنامه اصالت اثر، در خط اول، نام و نام خانوادگی خود را به صورت کامل با نام و نام خانوادگی نمونه، جایگزین کنید. در انتهای متن تعهد، در قسمت امضا نیز باید نام و نام خانوادگی کامل خود را وارد نماید.

هومن ذوالفقاری

امضا

چکیده

هدف از این پروژه، عبور از کاربرد «جعبه سیاه» الگوریتم ها و درک ریاضیاتی این مسئله است که چگونه تجزیه مقادیر تکین (SVD) می تواند ساختار پنهان داده ها را آشکار کند. در این تمرین، ما به جای استفاده صرف از توابع آماده، با تحلیل مستقیم ماتریس های V، مهم ترین ویژگی های یک مجموعه داده صنعتی را شناسایی کرده و آن را با روش های کلاسیک (مانند RFE) مقایسه Tو Σ ، U می کنید.

فهرست مطالب

صفحه

چکیده ..... أ
فصل اول مقدمه 1
فصل دوم آماده سازی داده ها..... 3
آماده سازی داده ها 4
فصل سوم روش های کلاسیک..... 5
فصل چهارم روش جبری 7
فصل پنجم تحلیل هندسی و پایداری..... 10
فصل ششم مقایسه نهایی 13
2-1- نتیجه گیری: 15
منابع و مراجع..... 16

فصل اول
مقدمه (دستور العمل)

مقدمه

لینک کد در گیتهاب:

https://github.com/hoomanzolfaghari84/Data-Mining-Course.git

در فرایندهای صنعتی پیشرفته، به‌ویژه در صنعت ساخت نیمه‌هادی‌ها، حجم بالایی از داده‌های سنسوری به‌صورت پیوسته تولید می‌شود. این داده‌ها معمولاً دارای ابعاد بسیار بالا، نویز قابل توجه و مقادیر گمشده هستند که تحلیل مستقیم آن‌ها را دشوار می‌سازد. از سوی دیگر، استفاده از تمام سنسورها نه‌تنها هزینه‌بر است، بلکه می‌تواند باعث کاهش پایداری و تفسیرپذیری مدل‌های یادگیری ماشین شود.

در این پروژه، با استفاده از دیتاست صنعتی SECOM، مسئله انتخاب ویژگی به‌عنوان یک گام کلیدی در تحلیل داده‌های صنعتی بررسی شده است. ابتدا داده‌ها پیش‌پردازش و نرمال‌سازی شدند، سپس روش‌های کلاسیک انتخاب ویژگی شامل Mutual Information و RFE به‌عنوان مبنای مقایسه به‌کار گرفته شدند. در ادامه، یک روش جبری مبتنی بر تجزیه مقادیر تکین (SVD) طراحی و پیاده‌سازی شد که بر ساختار هندسی داده‌ها تکیه دارد.

هدف اصلی این پروژه، مقایسه دقت، پایداری در برابر نویز، سرعت اجرا و تفسیرپذیری روش‌های مختلف انتخاب ویژگی و ارائه یک راهکار مناسب برای کاربردهای صنعتی واقعی است.

فصل دوم
آماده سازی داده ها

آماده سازی داده ها

وضعیت اولیه

- شکل اولیه داده‌ها: **(1567 × 590)**
- بعد از حذف ستون‌های با واریانس صفر: **(1567 × 474)**
- حدود ٪20**سنسورها غیرمفید** بوده‌اند (سنسورهای ثابت و بدون اطلاعات).

مقادیر گمشده (NaN : Not A Number)

- تعداد آنها قبل از Imputation : **41,136**
- بعد از Imputation با median همه مقدار دهی میشوند.
- انتخاب Median کاملاً درست است چون توزیع سنسورها نرمال نیست و نویز صنعتی بالاست.

بعد از نرمال‌سازی:

- میانگین ≈ 0
- انحراف معیار = 1

فصل سوم
روش های کلاسیک

روش های کلاسیک

در این بخش، دو روش کلاسیک انتخاب ویژگی را به‌عنوان مبنای مقایسه (Baseline) روی دیتاست SECOM اجرا می‌کنیم.

روش فیلتر

- **Mutual Information (MI)** میزان وابستگی آماری بین هر ویژگی و برچسب کلاس را می‌سنجد
- مستقل از مدل یادگیری
- سریع و مناسب داده‌های با ابعاد بالا
- فرمول آن:

$$MI(X,Y) = H(X) - H(X \mid Y)$$

هرچه MI بیشتر باشد آنگاه ویژگی اطلاعات بیشتری درباره کلاس دارد.

این روش تعامل بین ویژگی‌ها را در نظر نمی‌گیرد اما برای مقایسه baseline بسیار مناسب است.

روش پوششی RFE + RandomForest

:RFE (Recursive Feature Elimination)

- مدل را آموزش می‌دهد
- کم‌اهمیت‌ترین ویژگی‌ها را حذف می‌کند
- تا رسیدن به تعداد هدف

:RandomForest

- مناسب داده‌های غیرخطی و نویزی صنعتی
- معیار اهمیت ویژگی داخلی دارد

این روش وابستگی بین ویژگی‌ها را لحاظ می‌کند و معمولاً دقت بالاتری نسبت به روش‌های فیلتر دار. اما هزینه محاسباتی بالاتری دارد. **زمان اجرا حدودا 2.49 ثانیه شد** که وابسته به هایپر پارامتر ها نیز میباشد. این نشان دهنده هزینه محاسباتی بالاتر روش‌های پوششی نسبت به روش‌های فیلتر است. به همین دلیل، برای کاهش زمان اجرا، پارامترهای مدل شامل عمق درخت و نرخ حذف ویژگی‌ها تنظیم شد، بدون تغییر ماهیت روش پوششی.

فصل چهارم
روش جبری

روش جبری

تجزیه SVD

برای ماتریس داده‌ها:

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

که:

- $U \in \mathbb{R}^{n\times n}$ ← بردارهای تکین چپ
- $\Sigma \in \mathbb{R}^{n\times p}$ ← مقادیر تکین
- $V \in \mathbb{R}^{p\times p}$ ← بردارهای تکین راست

مستقیماً از numpy.linalg.svd استفاده شد نه PCA و نه Feature Selection آماده.

ستون‌های V نشان می‌دهند هر ویژگی چه سهمی در مؤلفه‌های اصلی دارد. یعنی هر **بردار تکین راست** یک الگوی غالب در داده‌هاست. بردارهای با σ بزرگ‌تر مربوط به ساختار واقعی داده و σ کوچک مربوط به نویزاست. پس یک ویژگی مهم است اگر در **k بردار تکین اول** ضرایب بزرگی داشته باشد و این ضرایب با σ وزن دهی شوند.

فرمول امتیازدهی (طبق صورت سؤال)

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

که:

- V_{ji}: ضریب ویژگی j در بردار تکین iام
- σ_i^2: وزن‌دهی انرژی (واریانس)
- k: تعداد مؤلفه‌های غالب (من 20 گرفتم هم راستا با انتخاب ۲۰ ویژگی در کل تمرین)

در این روش، انتخاب ویژگی بر اساس تجزیه مقدار تکین داده‌های نرمال‌شده انجام شد. هر ویژگی بر اساس میزان مشارکت آن در k بردار تکین اول و با وزن‌دهی به مجذور مقادیر تکین امتیازدهی شد. این وزن‌دهی باعث کاهش اثر مؤلفه‌های نویزی با مقادیر تکین کوچک می‌شود.

روش	نوع	نظارت	وابستگی به کلاس	ماهیت
Mutual Information	Filter	Supervised	بله	آماری
RFE + RF	Wrapper	Supervised	بله	مدل‌محور
SVD-based	Algebraic	Unsupervised	خیر	خطی-جبری

فصل پنجم
تحلیل هندسی و پایداری

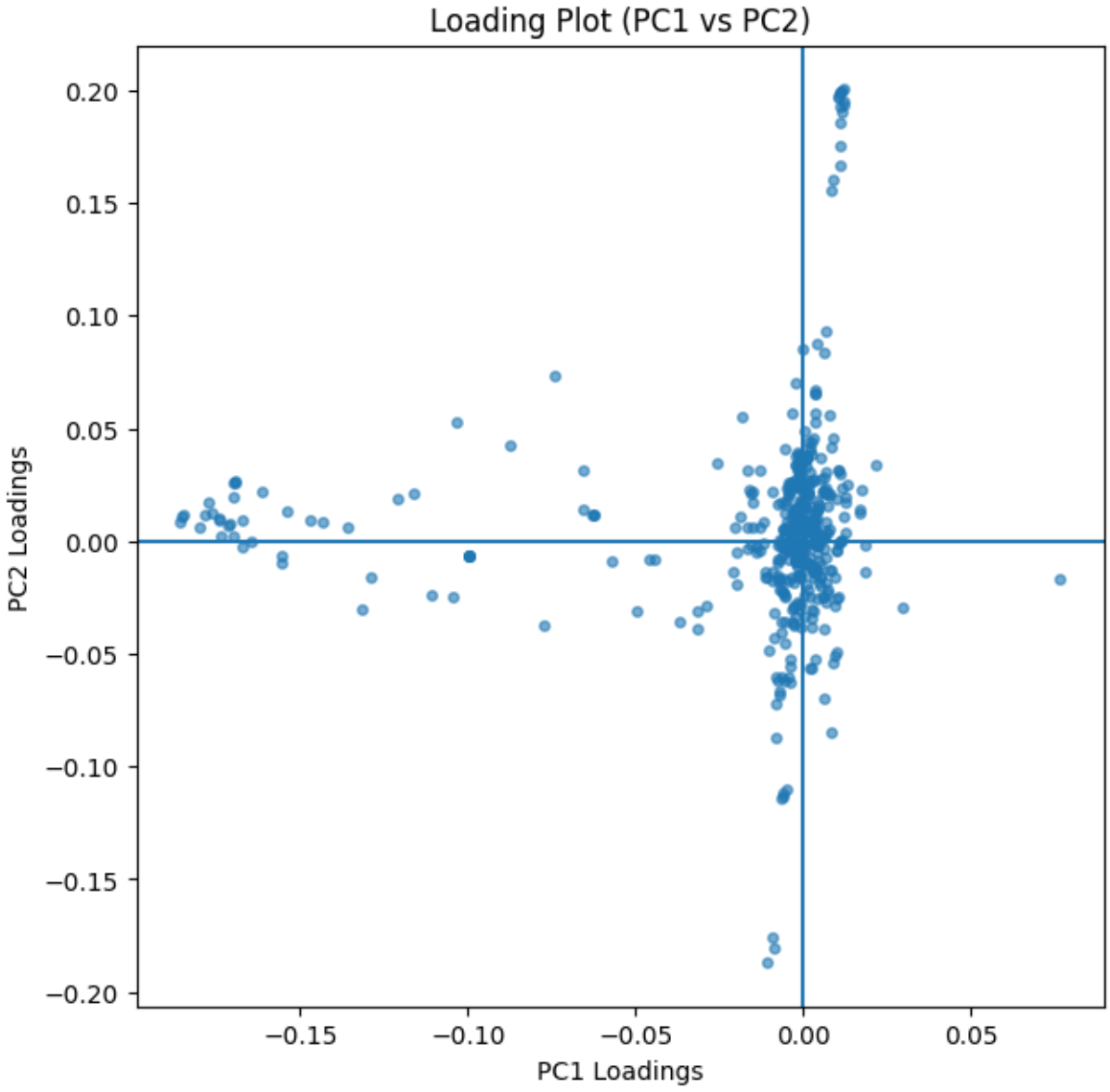
تحلیل هندسی و پایداری

نمودار بارگذاری (Loading Plot)

ستون‌های ماتریس V بردارهای تکین راست هستند. ضرایب هر ویژگی را روی مؤلفه‌های اصلی میگیریم. در نمودار هر نقطه یک سنسور است. اگر دو ویژگی:

- در یک جهت قرار بگیرند ← همبستگی بالا
- در جهت مخالف ← همبستگی منفی
- نزدیک مبدأ ← اهمیت کم

به طور کلی نزدیکی نقاط نشان از رفتار فیزیکی مشابه است. ما برای PC1و PC2 رسم کردیم:



نمودار بارگذاری برای دو مؤلفه اول نشان می‌دهد که تعدادی از سنسورها در جهات مشابه قرار گرفته‌اند که بیانگر رفتار آماری مشابه و همبستگی بالا میان آن‌هاست. این موضوع نشان‌دهنده افزونگی اطلاعات در مجموعه سنسورهاست و ضرورت انتخاب ویژگی را تأیید می‌کند.

تست پایداری(Stability Test)

ابتدا ۵٪ نویز تصادفی اضافه میکنیم. سپس ۲۰ ویژگی برتر را مجدداً با RFE و SVD-based انتخاب میکنیم. و سپس میزان تغییر لیست ویژگی‌ها را مقایسه میکنیم که بیانگر پایداری روش ها است. میبینیم که :

SVD overlap: 20 / 20

RFE overlap: 11 / 20

لیست ویژگی‌های انتخاب‌شده توسط روش SVD-based تغییر کمتری پس از افزودن نویز نشان داد. این موضوع به دلیل ماهیت جبری و استفاده از ساختار کلی داده و وزن‌دهی بر اساس مقادیر تکین بزرگ است، در حالی که روش RFE به دلیل وابستگی به مدل و تصمیمات گسسته درخت‌ها، نسبت به نویز حساس‌تر است.

فصل ششم
مقایسه نهایی

مقایسه نهایی

در این بخش از Logistic Regression ساده و interpretable استفاده شد. از Accuracy و F1-score برای مقایسه عملکرد روش‌ها استفاده شد و زمان انتخاب ویژگی بررسی شد. (execution time برای RFE و زمان محاسبه SVD). MI و SVD سریع، و RFE کند است.

Method	Accuracy	F1-score	Feature Selection Time (s)
Mutual Information	0.9299	0.9003	**0.02 s**
RFE + RF	**0.9363**	**0.9111**	**2.49 s**
SVD-based	**0.9363**	0.9075	**≈ 0 s**

بررسی نتایج:

- RFE و SVD از نظر Accuracy برابر هستند (0.9363)
- RFE کمی F1 بالاتر دارد (≈ +0.0035)
- زمان انتخاب ویژگی:
  - RFE تقریبا 100 برابر کندتر
  - SVDقریبا آنی
  - MI بسیار سریع ولی دقت کمتر
- بهبود F1 در RFE بسیار ناچیز است و در مقابل هزینه محاسباتی بالایی دارد.

همچنین میبینیم که فقط ۱ ویژگی مشترک از ۲۰ ویژگی داریم. یعنی SVDو RFE از دو دید کاملاً متفاوت به داده نگاه می‌کنند. اما با وجود این، به عملکرد طبقه بندی تقریباً یکسان میرسند.

:RFE

- مدل محور
- حساس به تصمیمات درخت و نویز

:SVD

- ساختار محور
- مبتنی بر انرژی و زیرفضای غالب داده

برای کاربرد صنعتی با تفسیرپذیری و سرعت بالا کدام روش پیشنهاد می‌شود؟

روش SVD-based feature selection

زیرا:

- ✓ سرعت بسیار بالا
- ✓ پایداری بهتر در برابر نویز (طبق تست بخش ۴)
- ✓ تفسیر پذیری هندسی (Loading plot، زیرفضای داده)
- ✓ عملکرد برابر با RFE، بدون هزینه محاسباتی

RFE فقط زمانی پیشنهاد می‌شود که:

- داده کم‌نویز باشد
- زمان محاسباتی اهمیتی نداشته باشد
- فقط دقت نهایی مهم باشد

2-1-نتیجه گیری:

نتایج نشان می‌دهد که روش SVD-based با وجود عدم استفاده از اطلاعات برچسب، به دقتی برابر با روش پوششی RFE دست می‌یابد، در حالی که زمان انتخاب ویژگی آن به‌طور چشمگیری کمتر است. همچنین، همپوشانی بسیار کم ویژگی‌های منتخب توسط SVD و RFE نشان می‌دهد که این دو روش از دیدگاه‌های متفاوتی ساختار داده را تحلیل می‌کنند. با توجه به نیازهای صنعتی شامل سرعت، پایداری و تفسیرپذیری، روش SVD-based انتخاب مناسب‌تری محسوب می‌شود.

.

منابع و مراجع

[1] https://archive.ics.uci.edu/ml/datasets/SECOM

[2] Guyon, I., & Elisseeff, A.(2003) .
An introduction to variable and feature selection.
Journal of Machine Learning Research, 3, 1157–1182.

[3] Machine Learning: A Probabilistic Perspective — Kevin P. Murphy

[4] Pattern Recognition and Machine Learning — Christopher Bishop

[5] منابع اصلی درس

Amirkabir University of Technology
(Tehran Polytechnic)

… Department …

MSc or PhD Thesis

Title of Thesis

By
Name

Supervisor
Dr.

Advisor
Dr.

Month & Year