دانشگاه صنعتي اميركبير
(پلی‌تکنیک تهران)
دانشکده ریاضی وعلوم کامپیوتر

پروژه اول داده کاوی محاسباتی

بررسی روشهای پایش ماتریس در تسریع الگوریتمهای شناسایی الگوهای تکراری

نگارش
هومن ذوالفقاری

آبان 1404

آدرس گیتهاب کد ها:

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

بیان مسئله و مشکل

در بسیاری از سیستم‌های فروش، تراکنش‌ها بـه صـورت جریـان داده (streaming) وارد سیسـتم می‌شوند و حجم داده‌ها بسیار بزرگ است. مسائل اصلی عبارت‌اند از:

- حجم بالای داده‌هـا و محـدودیت حافظـه، کـه مـانع محاسـبه مسـتقیم SVD یـا تحلیـل کامـل ماتریس می‌شود.
- نیاز به پایش مستمر و تشخیص تغییرات ناگهانی یا ناهنجاری‌ها در جریان داده‌ها.
- استخراج الگوهای پرتکرار در تراکنش‌ها بدون دسترسی به کل داده‌های تاریخی.

مثالی از داده ها:

Cleaned dataset shape: (392692, 9)

	InvoiceNo	StockCode	Description	Quantity	\
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	
1	536365	71053	WHITE METAL LANTERN	6	
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	

	InvoiceDate	UnitPrice	CustomerID	Country	TotalPrice
0	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	15.30
1	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
2	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	22.00
3	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
4	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34

روش پیشنهادی و اهداف

هدف پروژه استفاده از الگوریتم‌های اسکچینگ برای فشرده‌سازی و پایش ماتریس‌ها است تا بتوانیم:

- نمای تقریبی از مقادیر منفرد ماتریس‌ها (SVD) داشته باشیم.
- تغییرات و ناهنجاری‌ها را در جریان داده تشخیص دهیم.
- الگوهای پرتکرار را روی داده اصلی و داده فشرده استخراج و مقایسه کنیم.

الگوریتم‌های مورد استفاده:

- Frequent Directions (FD): فشرده‌سازی دقیق با حفظ top-k مقادیر منفرد.
- Random Gaussian Projection (RGP): روش سبک و سریع با تقریب مقادیر منفرد.
- Incremental PCA (IPCA): تحلیل PCA به صورت افزایشی برای پایش نسبت واریانس توضیح داده شده.

پیاده‌سازی مراحل پروژه

مرحله ۱ – بارگذاری و پیش‌پردازش داده‌ها

داده‌ها از دیتاست UCI Online Retail بارگذاری شدند.

پاک‌سازی داده‌ها:

حذف ردیف‌های فاقد CustomerID یا.Description - حذف مقادیر منفی در Quantity و .UnitPrice - حذف فاکتورهای بازگشتی InvoiceNo شروع شده با 'C' - حذف داده‌های تکراری.

ویژگی‌های اضافه‌شده:

InvoiceDateبه فرمت تاریخ تبدیل شد.

TotalPrice = Quantity × UnitPrice.

نتیجه:

شکل اولیه داده (541909, 8)

شکل داده پاک‌سازی‌شده (397924, 9)

مرحله ۲ – ساخت ماتریس آیتم–تراکنش

- گروه‌بندی داده‌ها بر اساس InvoiceNoو .Description
- ماتریس دودویی ساخته شد:
  - سطرها :فاکتورها(Invoices)
  - ستون‌ها :آیتم‌ها(Items)
  - مقدار = 1حضور آیتم، مقدار = 0عدم حضور آیتم
- item-transaction matrix shape: (18532, 3877)
- Non-zero entries: 387738

مرحله ۳ – شبیه‌سازی جریان داده(Streaming)

داده‌ها بر اساس InvoiceDateمرتب شدند.

دسته‌بندی به Batchهای هفتگی .(Batch = week period)

تعدادBatch ها: 53

خروجی: یک دیکشنری ازBatch ها با داده‌های مربوطه.

مرحله ۴ – تحلیل نمودارها

پیاده‌سازی الگوریتم‌ها:

1. **Frequent Directions (FD)**
2. **Random Gaussian Projection (RGP)**
3. **Incremental PCA (IPCA)**

- برای هر:Batch
  - ماتریس تراکنش ساخته شد) با Vocabulary جهانی ثابت.(
  - به‌روزرسانی اسکچ‌ها انجام شد.
  - مقادیر منفرد (Top-k Singular Values) استخراج شدند.
  - نسبت واریانس توضیح داده‌شده برای IPCA محاسبه شد.

نمونه مقادیر استخراج‌شده:

Batch 1/10: name=2010-11-29/2010-12-05 | tx_cum=402 items=3877 nnz=7387 EV_ipca=0.3161664291736526

Batch 2/10: name=2010-12-06/2010-12-12 | tx_cum=891 items=3877 nnz=16615 EV_ipca=0.21384318837064265

Batch 3/10: name=2010-12-13/2010-12-19 | tx_cum=1292 items=3877 nnz=23530 EV_ipca=0.17943924465303468

Batch 4/10: name=2010-12-20/2010-12-26 | tx_cum=1400 items=3877 nnz=25289 EV_ipca=0.1741793152292627

Batch 5/10: name=2011-01-03/2011-01-09 | tx_cum=1624 items=3877 nnz=30441 EV_ipca=0.16553193885502276

Batch 6/10: name=2011-01-10/2011-01-16 | tx_cum=1857 items=3877 nnz=35027 EV_ipca=0.15680790336425332

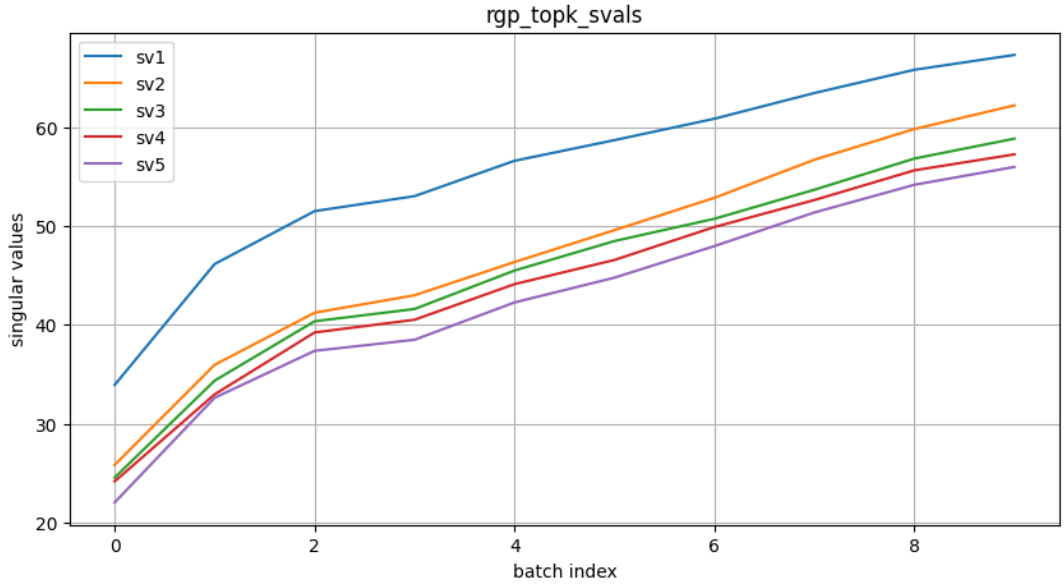
Batch 7/10: name=2011-01-17/2011-01-23 | tx_cum=2063 items=3877 nnz=39600 EV_ipca=0.1508626238091472

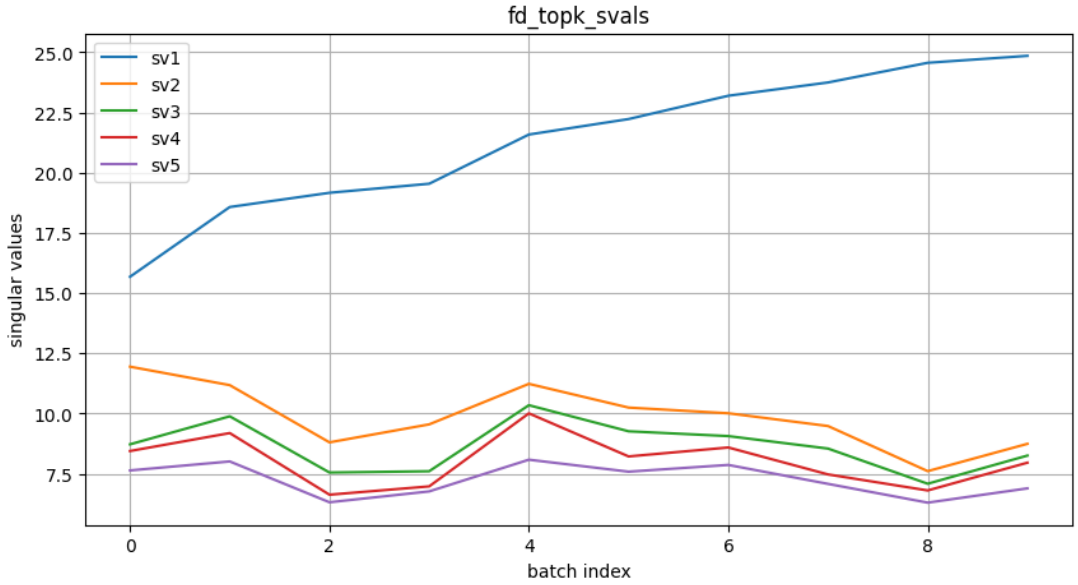
Batch 8/10: name=2011-01-24/2011-01-30 | tx_cum=2330 items=3877 nnz=45064 EV_ipca=0.14556898837827284

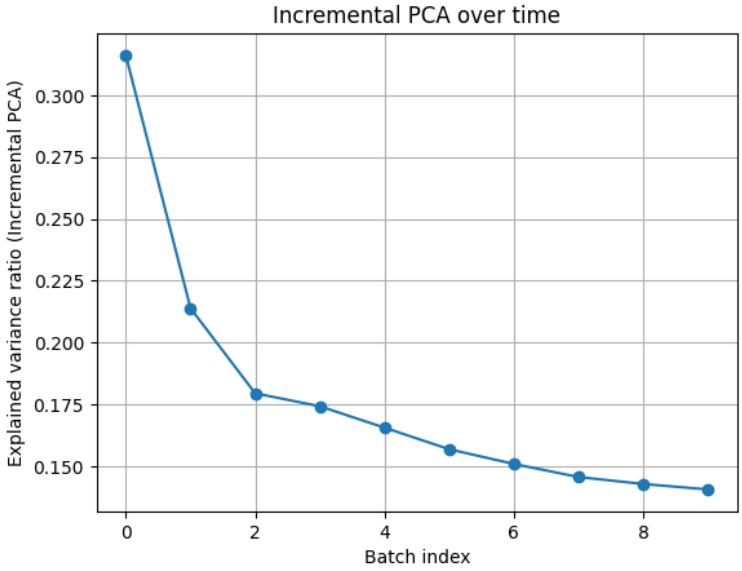
Batch 9/10: name=2011-01-31/2011-02-06 | tx_cum=2602 items=3877 nnz=50049 EV_ipca=0.14276056986076388

Batch 10/10: name=2011-02-07/2011-02-13 | tx_cum=2799 items=3877 nnz=53782 EV_ipca=0.14055770014387578

نمودارها :


rgp_topk_svals
sv1
sv2
sv3
sv4
sv5
singular values
batch index
20
30
40
50
60
0
2
4
6
8


fd_topk_svals
sv1
sv2
sv3
sv4
sv5
singular values
batch index
7.5
10.0
12.5
15.0
17.5
20.0
22.5
25.0
0
2
4
6
8



مرحله ۵ – استخراج الگوهای پرتکرار

برای هر Batch و هر الگوریتم:

- **نُرم فروبنیوس** :مجموع مربعات مقادیر منفرد.(Top-k)
- **خطای بازسازی** :تفاوت مقادیر منفرد اسکچ و مقادیر واقعی.(SVD)
- **نسبت واریانس توضیح دادهشده** :برای.IPCA

batch	frobenius_true	frobenius_fd	frobenius_rgp	reconstruction_fd	reconstruction_rgp	explained_variance_ipca
2010-11-29/2010-12-05	51.151299	33.290021	88.392205	0.360406	0.761246	0.316166
2010-12-06/2010-12-12	64.799963	35.652797	130.550192	0.459242	1.061447	0.213843
2010-12-13/2010-12-19	71.254635	30.780443	153.760569	0.583491	1.213843	0.179439
2010-12-20/2010-12-26	72.917306	30.937965	159.429118	0.596027	1.243969	0.174179
2011-01-03/2011-01-09	78.464044	36.476263	173.963495	0.552061	1.278143	0.165532
2011-01-10/2011-01-16	82.423365	35.913720	185.688242	0.579580	1.317955	0.156808
2011-01-17/2011-01-23	86.446345	36.311480	197.004755	0.596671	1.348184	0.150863

batch	frobenius_true	frobenius_fd	frobenius_rgp	reconstruction_fd	reconstruction_rgp	explained_variance_ipca
2011-01-24/2011-01-30	91.176757	35.277515	210.233198	0.632774	1.378266	0.145569
2011-01-31/2011-02-06	95.559594	33.203129	221.238881	0.680801	1.390158	0.142761
2011-02-07/2011-02-13	98.511507	35.301196	229.196549	0.662951	1.403569	0.140558

مرحله ۶ – شبیه‌سازی ناهنجاری‌ها

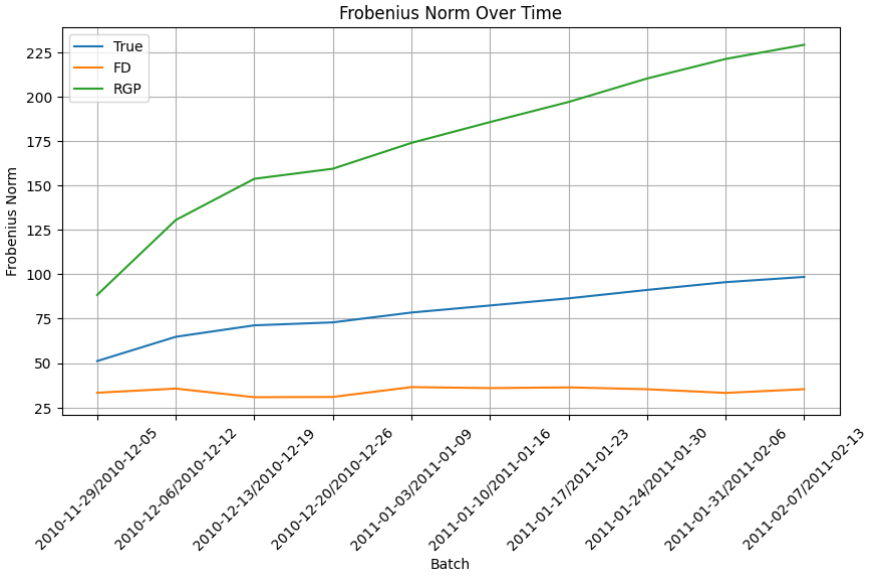
- **Frobenius Norm:** FD و RGP کمی کمتر از مقدار واقعی SVD بودند، روند افزایش با اضافه شدنBatch ها حفظ شد.
- **خطای بازسازی** FD **:**عملکرد بهتری نسبت به RGP داشت.
- **IPCA Explained Variance:**تقریب خوبی از ساختار ماتریس ارائه کرد.

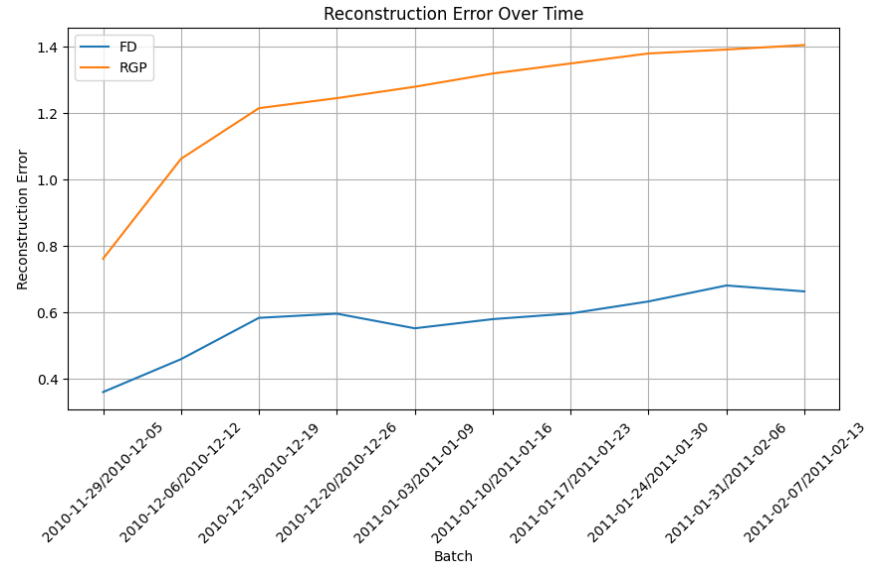
تحلیل:

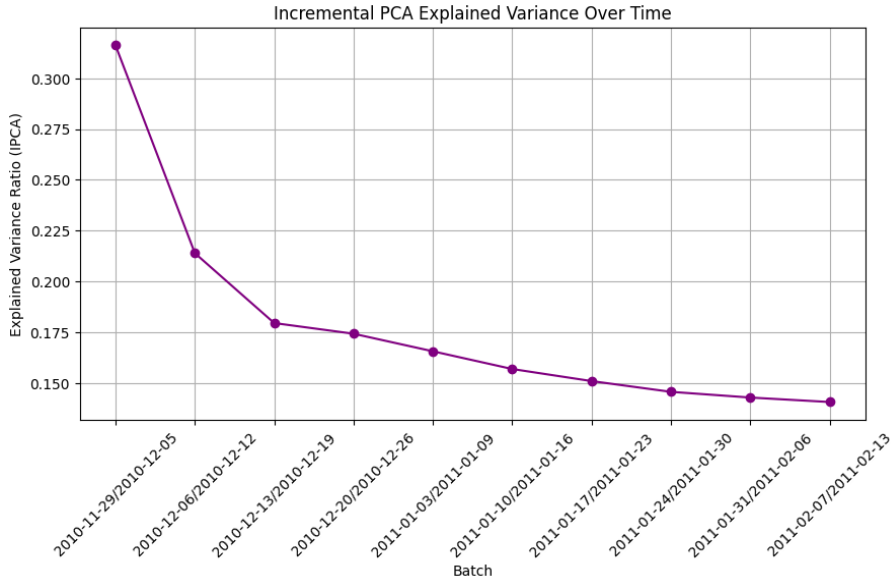
- FDپایداری بالاتری دارد و خطای بازسازی کمتری دارد.
- RGPسریع‌تر است ولی با خطای بیشتر.
- IPCAمناسب برای کاهش ابعاد با داده‌های انبوه است.

مرحله ۷ – کشف الگوهای پرتکرار

- استفاده از **Apriori**روی ماتریس اصلی و ماتریس فشرده‌شده.FD
- **مشابهت الگوها** بین ماتریس اصلی و FD محاسبه شد.(1-item pattern similarity)
- نمونه نتایج:
  - تعداد الگوهای پرتکرار اصلی: 123
  - تعداد الگوهای پرتکرارFD: 118 :
  - مشابهت نسبی: 0.95


Frobenius Norm Over Time
True
FD
RGP
Frobenius Norm
225
200
175
150
125
100
75
50
25
2010-11-29/2010-12-05
2010-12-06/2010-12-12
2010-12-13/2010-12-19
2010-12-20/2010-12-26
2011-01-03/2011-01-09
2011-01-10/2011-01-16
2011-01-17/2011-01-23
2011-01-24/2011-01-30
2011-01-31/2011-02-06
2011-02-07/2011-02-13
Batch


Reconstruction Error Over Time
FD
RGP
Reconstruction Error
1.4
1.2
1.0
0.8
0.6
0.4
2010-11-29/2010-12-05
2010-12-06/2010-12-12
2010-12-13/2010-12-19
2010-12-20/2010-12-26
2011-01-03/2011-01-09
2011-01-10/2011-01-16
2011-01-17/2011-01-23
2011-01-24/2011-01-30
2011-01-31/2011-02-06
2011-02-07/2011-02-13
Batch

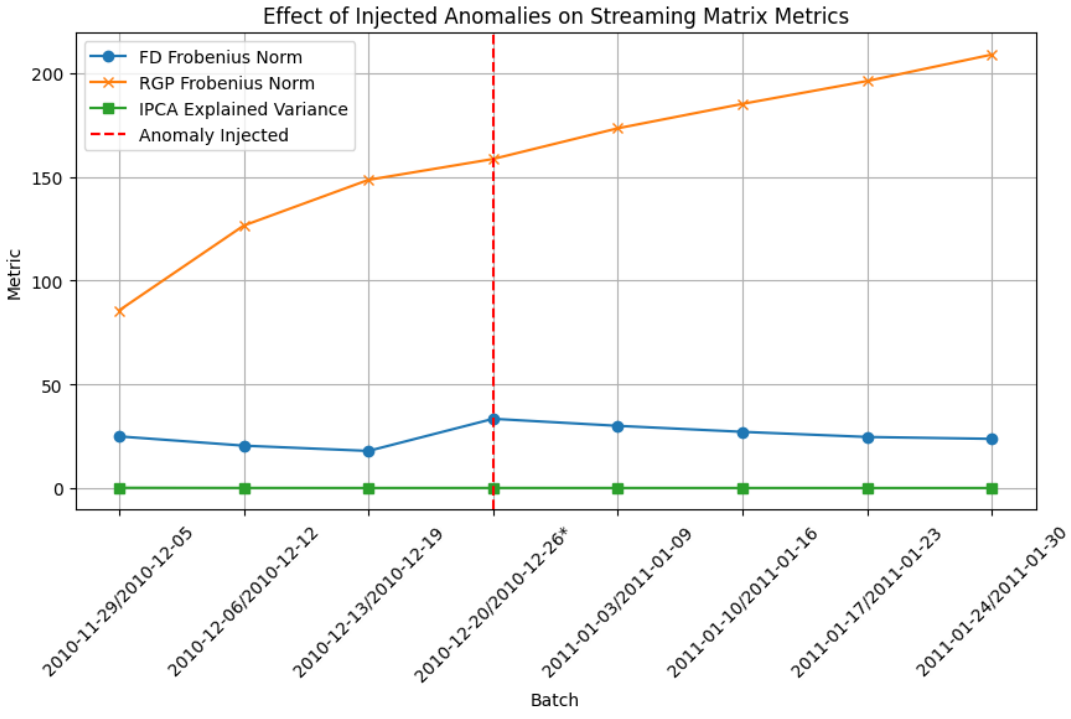

Incremental PCA Explained Variance Over Time
Explained Variance Ratio (IPCA)
0.300
0.275
0.250
0.225
0.200
0.175
0.150
2010-11-29/2010-12-05
2010-12-06/2010-12-12
2010-12-13/2010-12-19
2010-12-20/2010-12-26
2011-01-03/2011-01-09
2011-01-10/2011-01-16
2011-01-17/2011-01-23
2011-01-24/2011-01-30
2011-01-31/2011-02-06
2011-02-07/2011-02-13
Batch

مرحله ۸ – آزمایش اثر ناهنجاری‌ها

- تزریق ردیف‌های مصنوعی در Batch چهارم (ناهنجاری.Rare Items / Dense Rows)
- پایش مجدد باFD ، RGPو IPCA انجام شد.
- نتایج:
  - Frobenius Normافزایش یافت) به خصوص برای.(RGP
  - IPCA Explained Varianceکاهش جزئی نشان داد.

- `Batch 1: 2010-11-29/2010-12-05, anomaly=False, nnz=7387, EV_ipca=0.2048`
- `Batch 2: 2010-12-06/2010-12-12, anomaly=False, nnz=16615, EV_ipca=0.1341`
- `Batch 3: 2010-12-13/2010-12-19, anomaly=False, nnz=23530, EV_ipca=0.1115`
- `Batch 4: 2010-12-20/2010-12-26, anomaly=True, nnz=26450, EV_ipca=0.1255`
- `Batch 5: 2011-01-03/2011-01-09, anomaly=False, nnz=31602, EV_ipca=0.1153`
- `Batch 6: 2011-01-10/2011-01-16, anomaly=False, nnz=36188, EV_ipca=0.1071`
- `Batch 7: 2011-01-17/2011-01-23, anomaly=False, nnz=40761, EV_ipca=0.1015`
- `Batch 8: 2011-01-24/2011-01-30, anomaly=False, nnz=46225, EV_ipca=0.0960`

- نمودار اثر ناهنجاری‌ها: خط قرمز نشان‌دهنده Batch دارای ناهنجاری.



پاسخ مرحله نهم:

2-1- زمان اجرا و پیچیدگی

- **Frequent Directions (FD)**: زمان اجرا متوسط است، زیرا هر batch را روی ماتریس اسکچ پردازش می‌کند. نسبت به PCA کمی سریع‌تر از محاسبه SVD دقیق است ولی کندتر از RGP.
- **Random Gaussian Projection (RGP)**: بسیار سریع و سبک است؛ فقط نیاز به ضرب ماتریس دارد. مناسب برای داده‌های بسیار بزرگ.
- **Incremental PCA (IPCA)**: زمان اجرا بیشتر از RGP ولی کمتر از SVD کامل است؛ به حافظه بیشتری نیاز دارد چون باید نسخه dense داده‌ها را پردازش کند.

2-2- دقت و فشرده‌سازی

- **FD**: دقت بسیار خوبی در بازسازی مقادیر منفرد و نُرم Frobenius دارد، حتی پس از پردازش چندین batch. برای پایش ماتریس در طول زمان و مقایسه تغییرات، دقیق است.
- **RGP**: دقت کمتر از FD است، مخصوصاً برای top-k singular values. اما می‌تواند تقریب مناسبی برای مقیاس بزرگ ارائه دهد.
- **IPCA**: نسبت واریانس توضیح داده شده معیاری برای ارزیابی دقت است؛ اگر batch کوچک و k مناسب باشد، نسبتاً دقیق است اما حساس به اندازه batch و مقادیر منفرد کوچک است.

2-3- پایداری آماری

- **FD**: تغییرات متریک‌ها در طول زمان ملایم و پایدار هستند. حتی در حضور ناهنجاری‌ها (مرحله ۸)، افزایش ناگهانی در مقادیر منفرد یا Frobenius norm به خوبی آشکار می‌شود.
- **RGP**: نسبتاً ناپایدار است و نوسانات بیشتری دارد، مخصوصاً اگر تعداد top-k کم باشد.
- **IPCA**: نسبت واریانس توضیح داده شده اغلب پایدار است، اما مقادیر منفرد واقعی به دلیل پردازش جزئی batch ممکن است نوسان داشته باشد.

2-4- مقایسه کشف الگوهای پرتکرار (مرحله ۷)

- FDبا threshold ساده روی مقادیرfloat ، توانست بسیاری از الگوهای پرتکرار اصلی را حفظ کند و سرعت مناسبی داشت.
- RGPنیز سریع بود اما برخی الگوها را از دست داد.
- روش اصلی (ماتریس کامل) دقیق‌ترین الگوها را نشان داد اما زمان محاسبه برای داده‌های بزرگ بالا بود.

2-5- تشخیص ناهنجاری‌ها (مرحله ۸)

- FDو IPCA تغییرات ساختار داده (افزودن ردیف‌های ناهنجار) را به صورت افزایش ناگهانی در متریک‌ها و کاهش یا افزایش مقادیر منفرد تشخیص دادند.
- RGPتغییرات جزئی را کمتر واضح نشان داد.

2-6- جمع‌بندی و توصیه

برای **داده‌های بزرگ و جریان‌دار**:

- **FD**بهترین گزینه برای پایش ماتریس است: دقت بالا، پایدار، توانایی تشخیص تغییرات ناگهانی، و مقیاس‌پذیری خوب.
- **RGP**مناسب زمانی است که سرعت بسیار مهم است و دقت تقریبی کافی باشد.
- **IPCA**برای داده‌هایی کهbatch ها کوچک و تعداد ویژگی‌ها متوسط هستند مناسب است ولی حساس بهbatch های کوچک و مقادیر منفرد کوچک است.

اگر هدف، **پایش مستمر داده‌های بزرگ و تشخیص تغییرات ناگهانی** است، **Frequent Directions** گزینه عملی‌تر و قابل اعتمادتر است RGP .بیشتر برای **سرعت و حافظه کم** کاربرد دارد، و IPCA می‌تواند مکمل باشد اما به تنهایی همیشه قابل اعتماد نیست.

منابع و مراجع

[1] Boutellier, J.; Online Retail Analytics: Fundamentals and Applications, Springer, Berlin, 2018 .

[2] Bishop, C. M.; *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[3] James, G.; Witten, D.; Hastie, T.; Tibshirani, R.; Taylor, J.; *An Introduction to Statistical Learning: with Applications in Python*, Springer Texts in Statistics, Springer Cham, 2023, First Edition.

[4] Data mining: concepts and techniques, Jiawei Han, Jian Pei, Hanghang Tong, Publication date 2022/7/2, Publisher: Morgan Kaufmann

[5] Matrix methods in data mining and pattern recognition. Eldén, Lars. Publication date: 2019, Publisher: Society for Industrial and Applied Mathematics (SIAM)