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

داده کاوی محاسباتی

گزارش ۱ (بررسی روش‌های پایش ماتریس در تسریع الگوریتم‌های شناسایی الگوهای تکراری)

نگارش
محمدامین نعمتی

شماره دانشجویی
۴۰۴۱۲۴۹۰۴

استاد درس
آقای دکتر قطعی

نیمسال اول ۱۴۰۵-۱۴۰۴

چکیده

در این پروژه، یک مجموعه داده‌ی تراکنشی یک فروشگاه مورد تحلیل قرار گرفت. ابتدا داده‌ها بارگذاری شده و مراحل پیش‌پردازش اولیه (شامل حذف ناهنجاری‌ها و نرمال‌سازی ویژگی‌ها) روی آن‌ها اعمال شد. سپس ماتریس اقلام-تراکنش از داده‌های تراکنش‌محور ساخته شد و یک جریان داده‌ی شبیه‌سازی‌شده برای آن تولید گشت. در ادامه، الگوریتم‌های مختلف پایش ماتریس اقلام-تراکنش بر روی داده‌های حاصل اجرا شدند تا تغییرات ساختار داده در طول زمان پایش شوند. برای هر الگوریتم، شاخص‌های تحلیلی مهمی مانند خطای بازسازی، نسبت واریانس توضیح داده شده، نُرم فروبنیوس خطا و زمان اجرا محاسبه شد. این شاخص‌ها با استفاده از نمودارها بر حسب زمان ترسیم شده و عملکرد هر روش با سایر روش‌ها مقایسه گردید. علاوه بر این، الگوهای پرتکرار اقلام در داده‌ها استخراج شده و با روش‌های تشخیص ناهنجاری، نقاط غیرمعمول مشخص گشتند.

واژه‌های کلیدی:

جریان داده، پایش ماتریس، تحلیل مؤلفه‌های اصلی، استخراج الگوهای پرتکرار، تشخیص ناهنجاری

فهرست مطالب

صفحه

چکیده..... أ

فصل اول مقدمه..... ۱

فصل دوم ساخت ماتریس اقلام‌تراکنش و شبیه‌سازی جریان داده..... ۵

۱-۲- ساخت ماتریس اقلام-تراکنش..... ۵

۲-۲- شبیه‌سازی جریان داده ۶

فصل سوم فصل سوم پایش ماتریس با الگوریتم های مختلف..... ۷

۱-۳- روش‌ها و اجرای الگوریتم‌ها..... ۸

۲-۳- محاسبه‌ی شاخص‌ها و نتایج..... ۸

فصل چهارم نتایج..... ۱۰

فصل پنجم تشخیص الگوهای پرتکرار و تشخیص ناهنجاری..... ۱۳

۱-۵- تشخیص الگوهای پرتکرار..... ۱۴

۲-۵- تشخیص ناهنجاری..... ۱۵

فصل ششم نتیجه‌گیری..... ۱۷

Abstract..... ۱۹

فصل اول

مقدمه

مقدمه

در این پروژه هدف بررسی و تحلیل داده‌های فروش یک فروشگاه اینترنتی [1] است تا الگوی خرید مشتریان و روند تغییرات فروش در طول زمان مشخص شود. مجموعه داده شامل اطلاعات فاکتور کالا، تعداد، تاریخ، واحد قیمت، مشتری و کشور است. در این فصل ابتدا داده خام معرفی شده و سپس مراحل ساده پاک‌سازی و خلاصه نتایج اولیه نمایش داده می‌شود.

جدول ۱-۱ نمونه‌ای از داده‌های خام

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

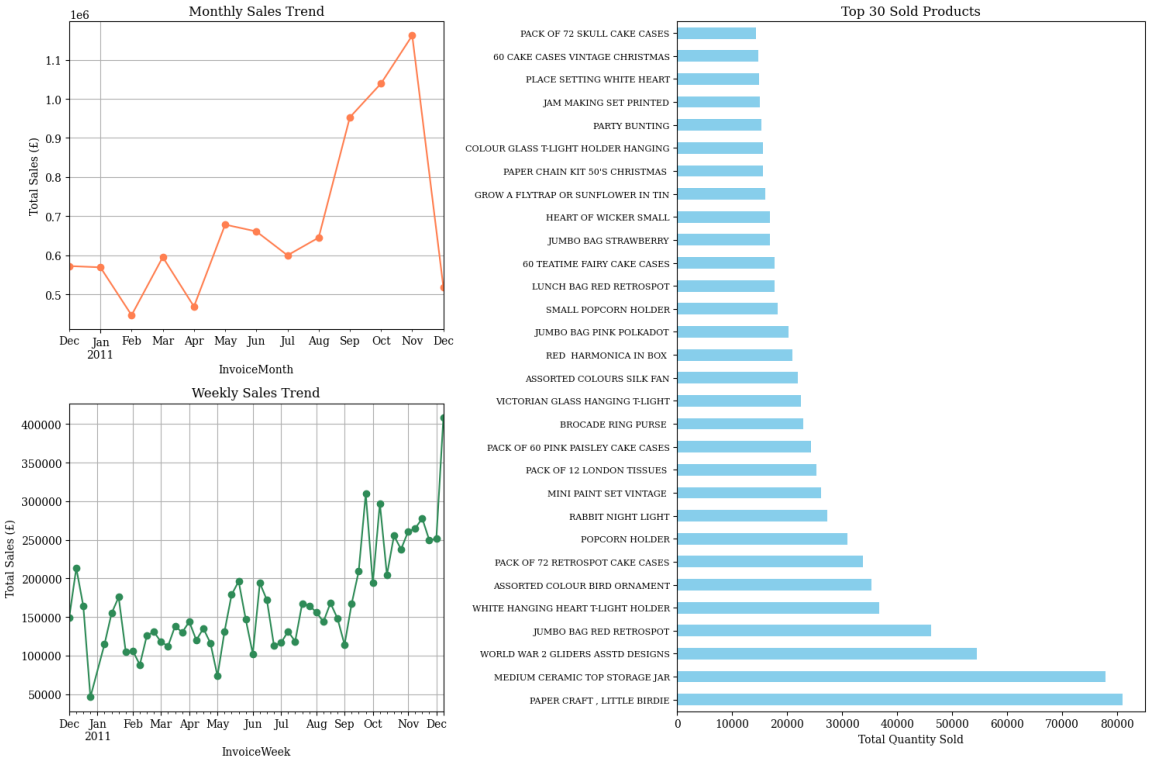
این جدول پنج ردیف از داده اصلی را نشان می‌دهد که در کل شامل **۵۴۱،۹۰۹** ردیف است.

پیش پردازش داده: ستون InvoiceNo از نوع رشته است و برخی فاکتورها با حرف C آغاز می‌شوند که نشان‌دهنده فاکتورهای برگشتی هستند. در سه گام ساده داده پاک‌سازی شد. ابتدا فاکتورهای برگشتی حذف شدند، سپس ردیف‌هایی که تعداد کالا کمتر یا مساوی صفر داشتند کنار گذاشته شدند و در نهایت ردیف‌های دارای مقدار خالی حذف گردید. پس از این مراحل مجموعه داده از **۵۴۱،۹۰۹** ردیف به **۳۹۷،۹۲۴** ردیف کاهش یافت.

جدول ۱-۲ مقایسه داده‌ها پس از پیش‌پردازش

توضیح	تعداد ردیف	وضعیت
داده خام اولیه	541909	قبل از پاک سازی
حذف فاکتورهای برگشتی	532621	حذف فاکتورهای C
حذف مقادیر منفی و صفر	531285	حذف تعداد غیر مثبت
مجموعه نهایی برای تحلیل	397924	حذف مقادیر خالی

شکل نهایی داده دارای **۳۹۷،۹۲۴** ردیف و ۸ ستون است. بازه زمانی از اول دسامبر ۲۰۱۰ تا نهم دسامبر ۲۰۱۱ را دربر می‌گیرد. تعداد مشتریان یکتا **۴۳۳۹**، تعداد محصولات **۳۸۷۷**، تعداد فاکتورها **۱۸۵۳۶** و تعداد کشورها **۳۷** مورد بوده است.



شکل ۱-۱ روند ماهانه و هفتگی فروش و فهرست کالاهای پرفروش

در نمودار روند ماهانه مشاهده می‌شود که فروش از اواسط سال به بعد رشد قابل توجهی دارد و در ماه های پاییز و ابتدای زمستان به بیشترین مقدار خود رسیده است. در نمودار روند هفتگی تغییرات بیشتری وجود دارد ولی مسیر کلی از میانه سال به سمت بالا است و در هفته‌های پایانی بیشترین مقدار فروش ثبت شده است.

این نمودار نشان می‌دهد که فروش ماهانه در نیمه‌ی دوم سال افزایش چشمگیری داشته است به ویژه در ماه‌های نوامبر و دسامبر که اوج خریدهای فصلی است. همچنین روند هفتگی فروش دارای نوسان‌های منظم است و برخی هفته‌ها با رشد قابل توجه همراه‌اند. در بخش راست شکل ۱-۱، فهرست ۳۰ کالای پرفروش ارائه شده است که نشان می‌دهد اقلام تزیینی و لوازم خانگی بیشترین سهم فروش را دارند.

فصل دوم

ساخت ماتریس اقلام‌تراکنش و شبیه‌سازی جریان داده

ساخت ماتریس اقلام‌تراکنش و شبیه‌سازی جریان داده

در این فصل روش ساخت ماتریس اقلام تراکنش توضیح داده شده و شبیه‌سازی ورود داده به صورت دسته‌ای ارائه می‌گردد.

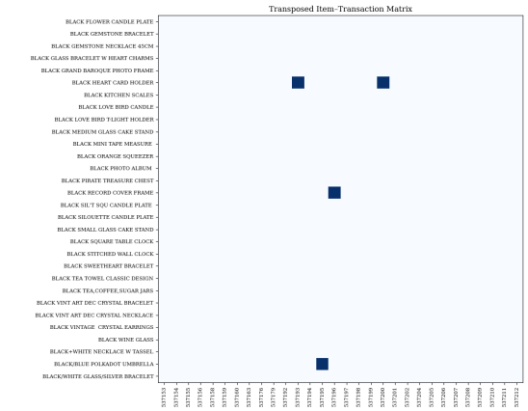
۲-۱-ساخت ماتریس اقلام–تراکنش

در این بخش، داده‌های فروش به‌صورت ماتریس دودویی شامل اقلام و فاکتورها نمایش داده شده است که در آن «۱» نشان‌دهنده حضور کالا در فاکتور است. ماتریس باینری اقلام در برابر فاکتورها ساخته شده و نمونه‌ای از آن در جدول ۲-۱ ملاحظه می‌شود.

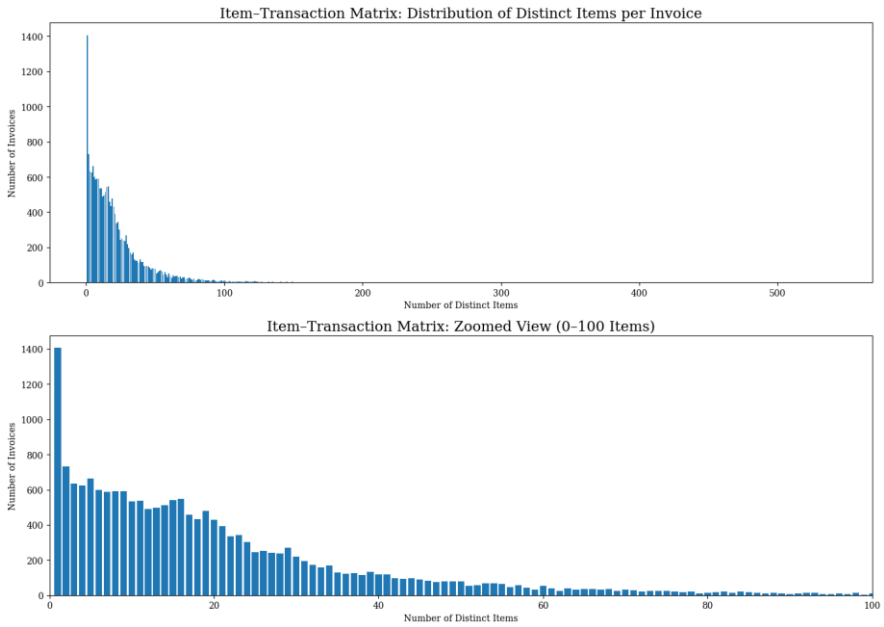
جدول ۲-۱ بخش کوچکی از ماتریس اقلام–تراکنش

شرح کالا	536365	536366	536367	536368	536369	536370	…	581582	581583	581584	581585	581586	581587
4 PURPLE FLOCK DINNER CANDLES	0	0	0	0	0	0	…	0	0	0	0	0	0
50'S CHRISTMAS GIFT BAG LARGE	0	0	0	0	0	0	…	0	0	0	0	0	0
DOLLY GIRL BEAKER	0	0	0	0	0	0	…	0	0	0	0	0	0
I LOVE LONDON MINI BACKPACK	0	0	0	0	0	0	…	0	0	0	0	0	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…
ZINC WILLIE WINKIE CANDLE STICK	0	0	0	0	0	0	…	0	0	0	1	0	0

نمایش گرافیکی بخشی از ماتریس در شکل ۲-۱ قرار دارد و توزیع تعداد اقلام متمایز در هر فاکتور در شکل ۲-۲ نمایش داده شده است.



شکل ۲-۱ نمایش گرافیکی بخشی از ماتریس اقلام–تراکنش



شکل ۲-۲ توزیع تعداد اقلام متمایز در فاکتورها

۲-۲-شبیه سازی جریان داده

در این مرحله، داده‌ها بر اساس تاریخ فاکتورها مرتب و به بازه‌های هفتگی تقسیم شدند تا ورود تدریجی داده‌ها و تغییرات ماتریس اقلام–تراکنش در گذر زمان شبیه‌سازی شود. خلاصه اطلاعات هر دسته هفتگی در جدول ۲-۲ ارائه شده است.

جدول ۲-۲ فهرست بخشی از بچ‌ها شامل ابتدای لیست و پایان لیست

Batch	Period	Rows	Cols
1	2010-11-29/2010-12-05	7753	11
2	2010-12-06/2010-12-12	9568	11
3	2010-12-13/2010-12-19	7060	11
4	2010-12-20/2010-12-26	1779	11
5	2011-01-03/2011-01-09	5230	11
...	...	...	...
49	2011-11-07/2011-11-13	14506	11
50	2011-11-14/2011-11-20	16067	11
51	2011-11-21/2011-11-27	15025	11
52	2011-11-28/2011-12-04	13663	11
53	2011-12-05/2011-12-11	11244	11

فصل سوم

پایش ماتریس با الگوریتم های مختلف و محاسبه شاخص های تحلیلی

پایش ماتریس با الگوریتم های مختلف و محاسبه شاخص های تحلیلی

۳-۱- روش‌ها و اجرای الگوریتم‌ها

در این فصل سه روش فشرده‌سازی و پایش ماتریس شامل پروجکشن تصادفی گوسی، تحلیل مؤلفه‌های اصلی افزایشی و جهت‌های پرتکرار اجرا شدند. در همه روش‌ها بُعد هدف ۵۰، دانه‌ی تصادفی ۴۲ و در روش جهت‌های پرتکرار اندازه‌ی اسکچ نیز ۵۰ در نظر گرفته شد. هر الگوریتم بر روی ماتریس دودویی اقلام–فاکتورها برای هر دسته داده اجرا شد و جدول ۳-۱ زمان اجرای هر روش را برای هر دسته نشان می‌دهد.

جدول ۳-۱ زمان اجرای روش‌ها

batch	rows	cols	rp_time	ipca_time	fd_time
1	402	1712	0.006471	0.246711	0.201381
2	489	1788	0.008622	0.370628	0.319458
3	401	1745	0.007673	0.220380	0.206979
4	108	916	0.002787	0.015390	0.019629
5	224	1446	0.005105	0.115318	0.075407
...	...	...	...	...	...
53	484	1989	0.009481	0.312244	0.356994

۳-۲- محاسبه شاخص ها و نتایج

برای در این بخش برای هر بچ چهار شاخص تحلیلی شامل نُرم فروبنیوس ماتریس اصلی، نسبت واریانس توضیح‌داده‌شده توسط اسکچ، خطای بازسازی نسبی و نُرم فروبنیوس ماتریس بازسازی‌شده محاسبه شد. نسبت واریانس توضیح‌داده‌شده نیز سهم واریانس حفظ‌شده در اسکچ را نسبت به کل واریانس نشان می‌دهد. نتایج در جدول ۳-۲ ارائه شده است.

جدول ۳-۲ شاخص های تحلیلی

batch	rows	cols	frob_X	rp_explained_var	ipca_explained_var	fd_explained_var	rp_recon_err	ipca_recon_err	fd_recon_err
1	402	1712	85.953476	0.977898	0.524941	0.523756	1.014212	0.720188	0.681257
2	489	1788	96.062480	1.012937	0.445661	0.444785	1.014257	0.769287	0.734971
3	401	1745	83.162492	0.986677	0.471117	0.469953	1.014296	0.748842	0.719178
4	108	916	41.952354	1.014410	0.865067	0.863597	1.028104	0.439613	0.364224
5	224	1446	71.791364	0.997345	0.625476	0.624032	1.017876	0.655956	0.603236
...	...	...	...	...	...	...	...	...	...
53	484	1989	104.244904	0.989403	0.480299	0.479425	1.012761	0.754370	0.712524

بر اساس نتایج، پروجکشن تصادفی گوسی واریانس زیادی از داده را حفظ می‌کند اما خطای بازسازی نسبتاً بالایی دارد.

فصل چهارم

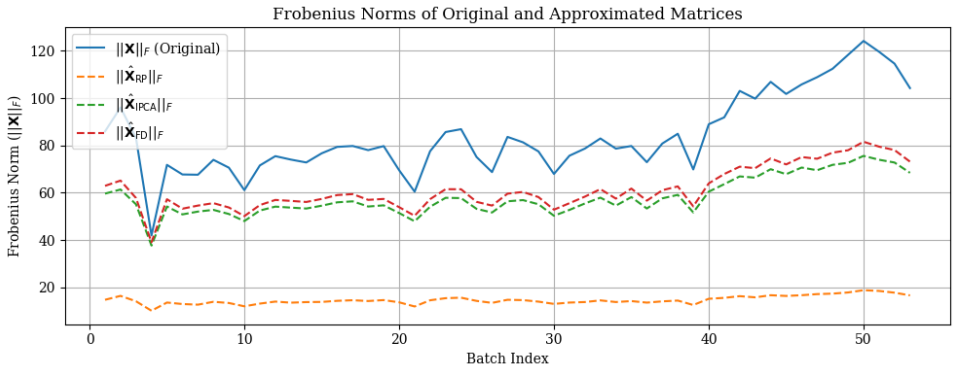
نتایج

نتایج

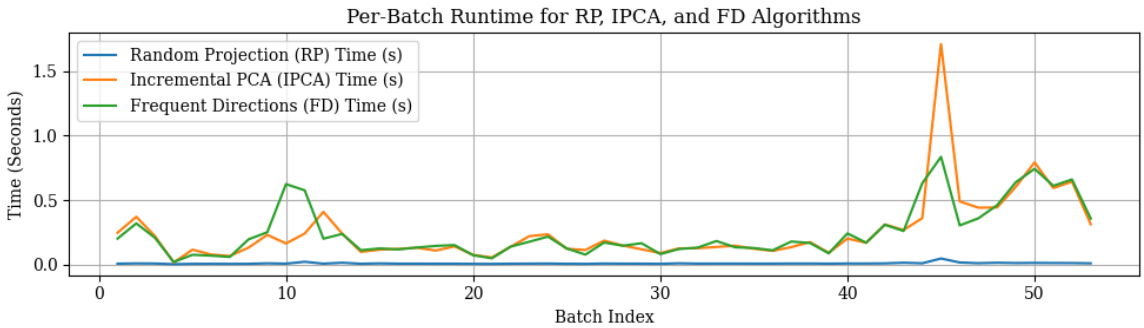
در این فصل، نتایج عددی و تصویری پایش ماتریس با سه روش **پروجکشن تصادفی گوسی**، **تحلیل مؤلفه‌های اصلی افزایشی** و **جهت‌های پرتکرار** ارائه می‌شود. همچنین، مقایسه‌ای از نظر **دقت**، **زمان اجرا** و **پایداری آماری** میان این روش‌ها انجام خواهد شد.

شکل ۴-۱ نمودار خطای بازسازی نسبی در طول بچ ها

شکل ۴-۲ نسبت واریانس توضیح داده شده در طول بچ ها



شکل ۴-۳ نُرم فروبنیوس ماتریس اصلی و ماتریس های تقریبی



شکل ۴-۴ زمان اجرای هر روش در هر بچ

جدول ۴-۱ آمار خلاصه میانگین ها

Metric	Value
rp_mean_err	1.0157
ipca_mean_err	0.7094
fd_mean_err	0.6671
rp_mean_ev	0.9941
ipca_mean_ev	0.5400
fd_mean_ev	0.5388
rp_mean_time	0.0089
ipca_mean_time	0.2499
fd_mean_time	0.2463

فصل پنجم

استخراج الگوهای پرتکرار و تشخیص ناهنجاری

استخراج الگوهای پرتکرار و تشخیص ناهنجاری

۵-۱- استخراج الگوهای پرتکرار

در این بخش، فرایند شناسایی مجموعه‌های پرتکرار بر پایه ماتریس‌های فشرده‌شده حاصل از سه روش پایش ماتریسی انجام شد. نتایج نشان می‌دهد که هر سه روش در حفظ ساختار آماری داده‌ها عملکرد قابل قبولی دارند، اما روش طرح‌ریزی تصادفی به دلیل سادگی و پایداری در نتایج پرتکرار، دقت بالاتری در بازیابی الگوها داشته است. روش مؤلفه‌های اصلی تدریجی با وجود فشرده‌سازی مناسب، اندکی از الگوهای کم‌تکرار را از دست داده است و روش جهت‌های پرتکرار نیز در حفظ روابط قوی‌تر میان آیتم‌ها بهتر عمل کرده است.

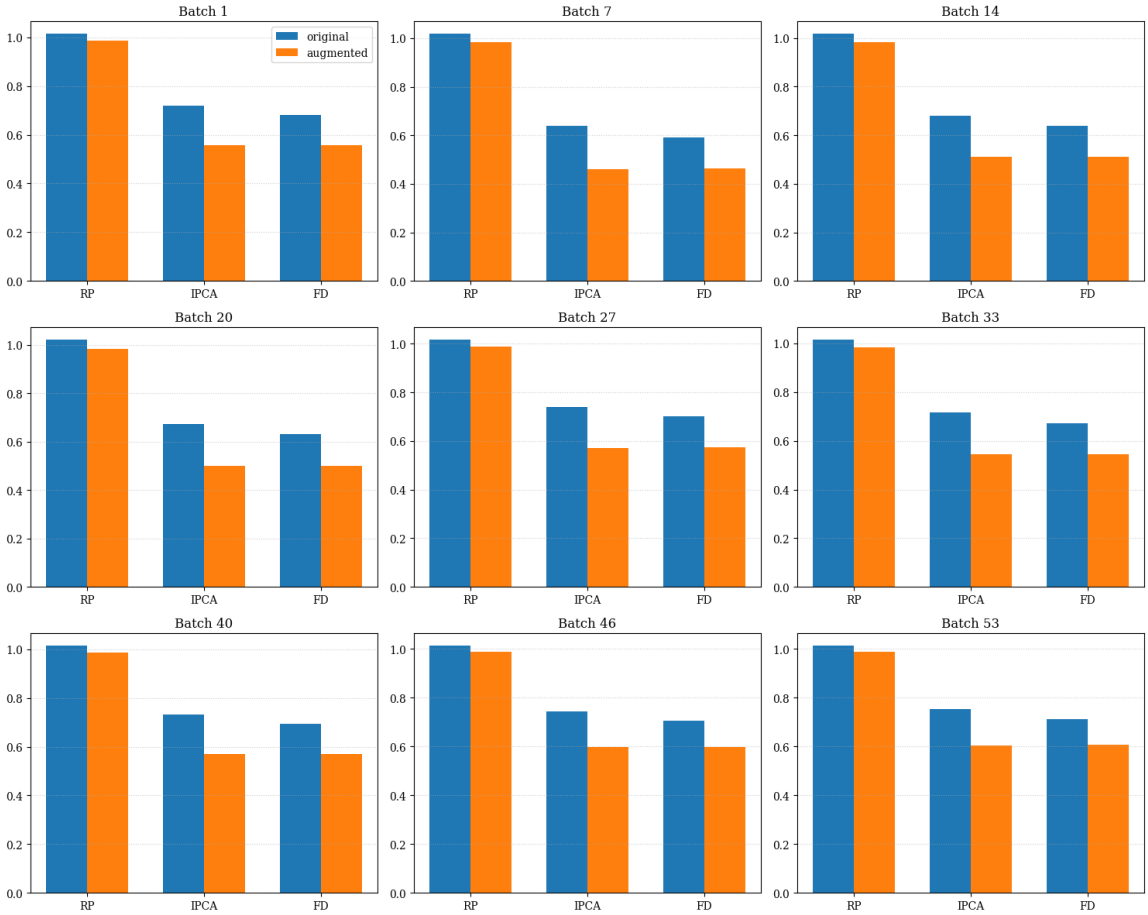
جدول ۵-۱ مقایسه عملکرد الگوریتم‌ها در استخراج الگوهای پرتکرار

Batch	Rows	Cols	Orig_Apriori_Count	Orig_Apriori_Time	FD_Apriori_Count	FD_Apriori_Time	FD_Apriori_Prec	FD_Apriori_Rec	FD_Apriori_F1	FD_Support_Diff
1	402	1712	929	0.591	780	1.251	0.956	0.803	0.873	0.0067
2	489	1788	387	0.087	162	0.018	0.988	0.413	0.583	0.0084
3	401	1745	230	0.036	85	0.017	0.988	0.365	0.533	0.0105
4	108	916	413	0.052	342	0.038	1.000	0.828	0.906	0.0022
5	224	1446	865	0.192	543	0.049	0.958	0.601	0.739	0.0051
…	…	…	…	…	…	…	…	…	…	…
53	484	1989	457	0.122	172	0.019	0.948	0.357	0.518	0.0105

Batch	Orig_Sample_Itemsets (example)	IPCA_Sample_Itemsets (example)	FD_Sample_Itemsets (example)
1	(SET 2 TEA TOWELS I LOVE LONDON)	(60 CAKE CASES VINTAGE CHRISTMAS)	(6 RIBBONS RUSTIC CHARM)
2	(SET 2 TEA TOWELS I LOVE LONDON)	(60 TEATIME FAIRY CAKE CASES)	(3 HEARTS HANGING DECORATION RUSTIC)
3	(12 PENCILS TALL TUBE RED RETROSPOT)	(ALARM CLOCK BAKELIKE PINK)	(72 SWEETHEART FAIRY CAKE CASES)
4	(SET 2 TEA TOWELS I LOVE LONDON)	(3 HOOK PHOTO SHELF ANTIQUE WHITE)	(3 HOOK PHOTO SHELF ANTIQUE WHITE)
5	(10 COLOUR SPACEBOY PEN)	(20 DOLLY PEGS RETROSPOT)	(20 DOLLY PEGS RETROSPOT)
…	…	…	…
53	(12 MESSAGE CARDS WITH ENVELOPES)	(72 SWEETHEART FAIRY CAKE CASES)	(12 PENCIL SMALL TUBE WOODLAND)

۵-۲- تشخیص ناهنجاری

برای بررسی توانایی الگوریتم‌ها در شناسایی تغییرات ناگهانی ساختار داده، تعدادی ردیف غیرعادی به داده افزوده شد و اثر آن در شاخص‌های بازسازی و واریانس بررسی گردید. نتایج نشان داد که روش طرح‌ریزی تصادفی سریع‌تر از سایر روش‌ها به تغییر واکنش نشان می‌دهد، در حالی که مؤلفه‌های اصلی تدریجی در تشخیص ناهنجاری‌های تدریجی عملکرد پایدارتری دارد. روش جهت‌های پرتکرار نیز توانسته است تغییرات ساختاری بزرگ‌تر را با دقت بیشتری آشکار کند.



شکل ۵-۱ تغییرات شاخص‌های پایش ماتریس در دسته‌های زمانی مختلف

جدول ۵-۲ مقایسه عملکرد الگوریتم‌ها در تشخیص ناهنجاری‌ها

Batch	M_Orig	M_Aug	D	Orig_Recon_RP	Orig_Recon_IPCA	Orig_Recon_FD	Aug_Recon_RP	Aug_Recon_IPCA
1	402	412	1712	1.014	0.720	0.681	0.985	0.557
7	206	216	1434	1.018	0.638	0.590	0.983	0.461
14	269	279	1433	1.017	0.680	0.637	0.983	0.510
20	248	258	1362	1.020	0.672	0.631	0.983	0.499
27	373	383	1649	1.015	0.738	0.701	0.986	0.571
33	325	335	1658	1.014	0.715	0.670	0.985	0.543
40	379	389	1795	1.014	0.733	0.694	0.986	0.569
46	429	439	2106	1.012	0.744	0.704	0.989	0.596
53	484	494	1989	1.012	0.754	0.712	0.988	0.604

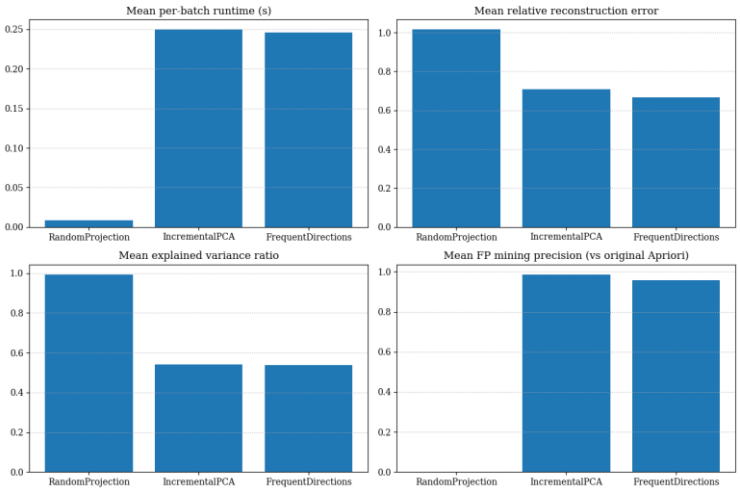
Batch	Aug_Recon_FD	Ratio_RP	Ratio_IPCA	Ratio_FD	EV_RP_Aug	EV_IPCA_Aug	EV_FD_Aug
1	0.558	0.972	0.773	0.819	0.963	0.680	0.680
7	0.463	0.966	0.723	0.784	0.939	0.778	0.778
14	0.511	0.966	0.750	0.803	0.962	0.730	0.729
20	0.500	0.964	0.742	0.793	0.891	0.741	0.740
27	0.572	0.971	0.773	0.816	0.908	0.664	0.663
33	0.545	0.970	0.759	0.812	0.976	0.694	0.693
40	0.570	0.972	0.776	0.821	0.979	0.666	0.665
46	0.597	0.977	0.801	0.848	0.912	0.634	0.633
53	0.605	0.975	0.801	0.850	0.948	0.624	0.623

فصل ششم

جمع‌بندی و نتیجه‌گیری

جمع‌بندی و نتیجه‌گیری

روش RandomProjection سریع‌ترین و پایدارترین روش است (میانگین زمان ۰.۰۰۹ ثانیه) اما با وجود حفظ بالاترین واریانس توضیح‌داده‌شده (۰.۹۹۴)، بیشترین خطای بازسازی (۱.۰۱۵) را دارد و در تشخیص ناهنجاری‌ها کاملاً ناتوان است. در مقابل، دو روش IncrementalPCA و FrequentDirections حدود ۲۵ تا ۳۰ برابر نسبت به روش اول کندتر بوده‌اند، با این حال در بازسازی داده و تشخیص ناهنجاری عملکرد بهتری دارند. در میان آن‌ها، FrequentDirections کمترین خطای بازسازی (۰.۶۶۷) و بهترین تعادل بین دقت (۰.۹۶) و بازخوانی (۰.۴۹) را ارائه می‌دهد، در حالی که IncrementalPCA دقت بالاتری دارد اما بازخوانی آن پایین‌تر است. بنابراین، برای سرعت و پایداری می‌توان از RandomProjection استفاده کرد، ولی برای دقت و تشخیص متعادل ناهنجاری‌ها FrequentDirections گزینه مناسب‌تری می‌باشد.



شکل ۶-۱ نمودار مقایسه‌ای کارایی الگوریتم‌ها

جدول ۶-۱ مقایسه کمی عملکرد الگوریتم‌ها بر اساس معیارهای مختلف

Method	mean_time_s	median_time_s	mean_recon_err	mean_explained_var	stab_ev_std	stab_err_std	fp_mean_prec	fp_mean_rec	anom_mean_ratio
Random Projection	0.008861	0.007080	1.015732	0.994124	0.016638	0.002924	0.000000	0.000000	0.970783
Incremental PCA	0.249857	0.147232	0.709351	0.540048	0.076214	0.056026	0.986617	0.312415	0.767032
Frequent Directions	0.246259	0.170670	0.667082	0.538821	0.076046	0.061663	0.956272	0.487601	0.816660

Refrences

[1] Chen D. Online Retail [dataset]. 2015. UCI Machine Learning Repository. Available from: https://doi.org/10.24432/C5BW33.

Abstract

In this project, a transactional dataset from a store was analyzed. Initially, the data was loaded, and initial preprocessing steps (including outlier removal and feature normalization) were applied. Subsequently, an item-transaction matrix was constructed from the transaction-based data, and a simulated data stream was generated for it. Following this, various item-transaction matrix monitoring algorithms were executed on the resulting data to track changes in the data structure over time. For each algorithm, key analytical metrics such as reconstruction error, explained variance ratio, Frobenius norm of error, and execution time were calculated. These metrics were plotted over time, and the performance of each method was compared with the others. Furthermore, frequent item patterns were extracted from the data, and outliers were identified using anomaly detection methods.

Key Words: Data stream, Matrix monitoring, Principal Component Analysis, Frequent pattern mining, Anomaly detection.

GITHUB LINK

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

Amirkabir University of Technology
(Tehran Polytechnic)

Department of computer science

Computational Data Mining

Report 1: Investigation of Matrix Monitoring Methods for Accelerating Repetitive Pattern Identification Algorithms

By
Mohammadamin Nemati

Student ID
404124904

Lecturer
Dr. Ghatee

First Semester of the 2025–2026 Academic Year