دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
دانشکده علوم کامپیوتر

تمرین چهارم

روش‌های پیشرفته در کاوش الگوهای تکراری

نگارش
ثمین مهدی پور
۹۸۳۹۰۳۹

استاد راهنما
دکتر قطعی

استادمشاور
دکتر یوسفی مهر

آذر ۱۴۰۲

چکیده

در این پروژه، با مرور و مقایسه الگوریتم‌های Apriori و FP-Growth بر روی مجموعه‌های داده مختلف با اندازه‌های نمونه متفاوت، تحلیل حساسیت الگوریتم‌ها انجام شد. از رویه‌های پیاده‌سازی برجسته برای مدیریت مجموعه‌های داده بزرگ Growth بهره‌گرفته شد. الگوریتم FP-Growth با سرعت و کارآیی بالاتری نسبت به Apriori عمل کرد. نتایج، الگوهای مفید و ارتباطات داده را به وضوح نشان می‌دهند. تفسیر دقیق و بهینه‌سازی پارامترها، ابزارهای کلیدی در استخراج اطلاعات از داده‌های گسترده و پر رشد می‌باشند.

واژه‌های کلیدی:

الگوریتم Apriori، الگوریتم FP-Growth، کاوش الگوهای تکراری

فهرست مطالب

صفحه

فهرست اشکال

صفحه

فصل اول

مقدمه

مقدمه

در این پژوهش، به بررسی و ارزیابی الگوریتم‌های داده‌کاوی Apriori و FP-Growth بر روی مجموعه‌های داده بزرگ پرداختیم. داده‌کاوی یک حوزه اساسی در علم داده‌ها است که به ما این امکان را می‌دهد تا الگوها و ارتباطات مختلف در داده‌ها را شناسایی و استخراج کنیم.

۱-۱-Apriori Algorithm:

الگوریتم Apriori یکی از الگوریتم‌های مشهور در داده‌کاوی است که برای استخراج الگوهای فراوان در مجموعه‌های داده از اهمیت بالایی برخوردار است. این الگوریتم از ایده "قاعده‌ی Apriori" بهره می‌برد که می‌گوید هر زیرمجموعه‌ای از یک مجموعه داده که فراوانی آن بیشتر از حد تعیین شده (حد حاشیه‌ی حداقل) باشد، خود نیز فراوانی بالایی دارد.

۱-۲-FP-Growth Algorithm:

الگوریتم FP-Growth یکی از الگوریتم‌های قدرتمند در داده‌کاوی مبتنی بر درخت است. این الگوریتم بر اساس ساختار FP-Tree که توالی‌های فراوان و ارتباطات بین آن‌ها را نشان می‌دهد، عمل می‌کند. با ساخت FP-Tree و سپس بهره‌گیری از الگوریتم بازگشتی، این الگوریتم با کمترین تعداد پاسخ‌ها به سوالات داده‌کاوی منجر می‌شود و بدون نیاز به اسکن مجموعه داده، عملکرد بهینه‌ای ارائه می‌دهد.

این تحقیق به مقایسه عملکرد دقیق این دو الگوریتم با استفاده از مجموعه‌های داده با اندازه‌های مختلف، تحلیل حساسیت روی پارامترهای الگوریتم‌ها، و بهره‌گیری از راهبردهای پیاده‌سازی بر مجموعه داده‌های بزرگ Growth پرداخته است. این پژوهش با آشکارسازی و تفسیر الگوها و ارتباطات به‌دست‌آمده، اطلاعات مفیدی را از داده‌ها استخراج کرده و نقش اساسی الگوریتم‌های داده‌کاوی را در تحلیل داده‌های پیچیده و حجیم به روشنی نشان میدهد.

فصل دوم

پیاده‌سازی

۲-۱- مجموعه داده

مجموعه داده شامل ۳۸۷۶۵ ردیف از سفارشات خرید افراد از فروشگاه‌های خواروبار است. این سفارشات می‌توانند تجزیه و تحلیل شده و قوانین انجمنی[۱] با استفاده از تحلیل سبد بازار و الگوریتم‌هایی مانند الگوریتم Fp-growth و Apriori به‌دست آورده شوند.
تجزیه و تحلیل سبد خرید یکی از تکنیک‌های اساسی استفاده‌شده برای کشف ارتباطات میان اقلام است. این تکنیک با جستجوی ترکیب‌هایی از اقلام که به طور مکرر در معاملات اتفاق می‌افتند، عمل می‌کند. به عبارت دیگر، این فرایند تلاش میکند روابط میان اقلامی که افراد خریداری می‌کنند، را شناسایی کند.

۲-۲- قواعد انجمنی

قوانین انجمنی به‌طور گسترده برای تحلیل داده‌های سبد یا تراکنش‌های خرید از فروشگاه‌ها استفاده می‌شوند و هدف آن تشخیص الگوی های بارز در داده‌های تراکنش با استفاده از معیارهای تعریف شده است. سه مفهوم در اینجا تعریف میشود:

۲-۲-۱- پشتیبانی[۲]

نشان می‌دهد چقدر یک مجموعه از اقلام محبوب است، به عنوان اندازه‌گیری نسبت تراکنش‌هایی که یک مجموعه از اقلام در آن ظاهر می‌شود.

---

۱ Association Rules

۲ Support

۲-۲-۲- قاطعیت[۳]

نشان می‌دهد چقدر احتمال دارد زمانی که اقلامی خریداری می‌شوند، اقلام دیگری نیز خریداری شود، به صورت {X -> Y}. این معیار اندازه‌گیری شده توسط نسبت تراکنش‌های دارای اقلام X به تراکنش‌هایی که اقلام Y نیز در آن ظاهر می‌شود.

۲-۲-۳- لیفت[۴]

نشان می‌دهد چقدر احتمال دارد زمانی که اقلام X خریداری می‌شوند، اقلام Y نیز خریداری شود و در عین حال کنترل می‌کند چقدر اقلام Y محبوب هستند.

۲-۳- تجزیه و تحلیل داده های اکتشافی[۵]

پس از بررسی مجموعه داده متوجه میشویم که ۳۸۷۶۵ سطر و سه ستون دارد که نشان دهنده شماره عضویت فرد خریدار، تاریخ خرید و آیتم خریداری شده می باشد. اما این شکل از داده ها در شناسایی الگوها برای ما مفید نخواهد بود بنابراین ساختار دیتاست را به شکلی که در سطرها شماره سفارشات و در ستون ها آیتم های قابل خریداری ذکر شده باشند در می آوریم که درصورت حضور هرآیتم در سبد خرید یک سفارش مشخص مقدار Trueو درصورت عدم حضور False ثبت شود.

	Instant food products	UHT-milk	abrasive cleaner	artif. sweetener	baby cosmetics	bags	baking powder	bathroom cleaner	beef	berries	...	turkey	vinegar	waffles
0	False	False	False	False	False	False	False	False	False	False	...	False	False	False
1	False	False	False	False	False	False	False	False	False	False	...	False	False	False
2	False	False	False	False	False	False	False	False	False	False	...	False	False	False
3	False	False	False	False	False	False	False	False	False	False	...	False	False	False
4	False	False	False	False	False	False	False	False	False	False	...	False	False	False

[۳] Confidence

[۴] Lift

[۵] exploratory data analysis

۲-٤-پیاده‌سازی قواعد انجمنی

همانطور که در بخش قبلی هم صحبت شد، پشتیبانی نشان می‌دهد که یک آیتم بر اساس نسبت تمام تراکنش‌هایی که شامل می‌شود چقدر محبوب است. این محبوبیت در صورتی محقق می شود که با آستانه پشتیبانی مشخص شده توسط کاربر مطابقت داشته باشد. به عنوان مثال، یک آستانه پشتیبانی که روی ۰.۲ (۲۰٪) تنظیم شده است به این معنی است که کاربر تمام مواردی را می خواهد که حداقل در ۲۰٪ از تمام تراکنش ها با هم اتفاق می افتد.

آستانه پشتیبانی بالا ترکیب اقلام بیشتری را ارائه نمی دهد، بنابراین کاهش ارزش ممکن است برای دیدن ترکیب های بسیار بیشتر اقلام برای اهداف بازاریابی مفید باشد.

۲-۴-۱- پیاده‌سازی توابع موردنیاز

perform_rule_calculation -۱-۱-۴-۲

این تابع به منظور محاسبه قوانین انجمنی اجرا می‌شود. ورودی‌های اصلی آن شامل ماتریس تراکنش و اقلام، نوع الگوریتم (پیش‌فرض: FP-Growth) و حداقل مقدار پشتیبانی (پیش‌فرض: ۰.۰۰۱) هستند. خروجی این تابع شامل ماتریسی با سه ستون است:

۱. پشتیبانی (Support): مقادیر پشتیبانی برای هر ترکیب اقلام.

۲. مجموعه اقلام (Itemsets): ترکیب‌های اقلام مختلف.

۳. تعداد اقلام (Number_of_items): تعداد اقلام موجود در هر ترکیب.

همچنین، زمان اجرای الگوریتم مورد نظر (Apriori یا FP-Growth) نیز به عنوان خروجی باز می‌گردد. این تابع با استفاده از شرط‌ها، ابتدا مشخص می‌کند کدام الگوریتم باید استفاده شود (Apriori یا FP-Growth) و سپس محاسبات را انجام می‌دهد. در نهایت، تعداد اقلام موجود در هر ترکیب افزوده می‌شود و ماتریس نهایی به عنوان خروجی تابع تولید می‌شود.

compute_association_rule -۲-۱-۴-۲

این تابع برای محاسبه قوانین نهایی انجمنی به کار می‌رود. ورودی‌های این تابع شامل ماتریس قوانین مربوط به الگوریتم‌های داده‌کاوی (Apriori یا FP-Growth)، معیار مورد استفاده برای محاسبه (پیش‌فرض: Lift) و حداقل مقدار آستانه (پیش‌فرض: ۱) می‌باشند.

خروجی این تابع شامل قوانین حاصله بر اساس معیار و آستانه مشخص شده است. این قوانین شامل اطلاعاتی از هر تراکنش هستند که شرایط معیار و آستانه مورد نظر را برآورده می‌کنند.

تابع با استفاده از تابع `association_rules` از کتابخانه‌ی داده‌کاوی، قوانین انجمنی را بر اساس معیار و آستانه ورودی محاسبه می‌کند و در نهایت این قوانین را به عنوان خروجی باز می‌گرداند.

plot_metrics_relationship -۳-۱-۴-۲

این تابع برای رسم نمودار ارتباط بین دو ستون ورودی در ماتریس قوانین انجمنی به کار می‌رود. ورودی‌های این تابع شامل ماتریس قوانین مربوط به الگوریتم‌های داده‌کاوی (Apriori یا FP-Growth) و دو ستون (col1 و col2) مورد نظر برای رسم نمودار هستند.

تابع ابتدا با استفاده از تابع `np.polyfit`، یک خط تطابق بین دو ستون را محاسبه کرده و سپس این خط تطابق را با استفاده از `np.poly1d` ایجاد می‌کند. سپس با استفاده از `plt.plot` نمودار ارتباط بین دو ستون به همراه خط تطابق رسم می‌شود. در نهایت، نمودار با تعیین محورها و عناوین مناسب نمایش داده می‌شود.

compare_time_exec -۴-۱-۴-۲

این تابع برای مقایسه زمان اجرای دو الگوریتم مختلف در داده‌کاوی به کار می‌رود. ورودی‌های این تابع شامل توضیحات دو الگوریتم (algo1 و algo2) می‌باشند.

تابع ابتدا لیستی از زمان اجرا برای هر الگوریتم از ورودی‌ها استخراج می‌کند و سپس با استفاده از `plt.bar` یک نمودار میله‌ای ایجاد می‌کند. رنگ‌های مختلف برای هر الگوریتم در نمودار انتخاب

شده‌اند. سپس با استفاده از `plt.xticks` نام الگوریتم‌ها به عنوان محور x تنظیم می‌شود و با تعیین محورها و عنوان مناسب، نمودار نهایی نمایش داده می‌شود.

۲-۵-اعمال الگوریتم Fp-growth

۲-۵-۱- سایز نمونه : ۲۰ درصد

در این بخش ۲۰ درصد از داده هارا جدا کرده و الگوریتم Fp-growth را روی آنها اعمال کردیم.

- Fp Growth execution took: 0.10614347457885742 seconds

خروجی ماتریسی به صورت زیر بود:

	support	itemsets	number_of_items
0	0.157923	(whole milk)	1
1	0.085879	(yogurt)	1
2	0.060349	(sausage)	1
3	0.009490	(semi-finished bread)	1
4	0.051728	(pastry)	1

اگر یک مجموعه آیتم دارای مقدار پشتیبانی بالایی باشد، به این معنی است که ترکیب آیتم ها اغلب در مجموعه داده اتفاق می افتد. این مجموعه اقلام با پشتیبانی بالا اغلب به عنوان انجمن های قابل اعتماد تری در نظر گرفته می شوند.

در گام بعدی لیفت را انجام دادیم:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(frankfurter)	(citrus fruit)	0.042432	0.056465	0.004009	0.094488	1.673391	0.001613	1.041991	0.420243
1	(citrus fruit)	(frankfurter)	0.056465	0.042432	0.004009	0.071006	1.673391	0.001613	1.030758	0.426493
2	(frankfurter)	(bottled water)	0.042432	0.057133	0.002673	0.062992	1.102546	0.000249	1.006253	0.097130
3	(bottled water)	(frankfurter)	0.057133	0.042432	0.002673	0.046784	1.102546	0.000249	1.004565	0.098645
4	(frankfurter)	(bottled beer)	0.042432	0.043769	0.002339	0.055118	1.259302	0.000482	1.012011	0.215033

تحلیل و بررسی:

- مقدمات و نتایج:

این ستون ها مجموعه ای از موارد درگیر در هر قانون ارتباط را نشان می دهند. به عنوان مثال، ردیف اول یک قانون ارتباطی را نشان می دهد که در آن "frankfurter" مقدم است و "مرکبات" نتیجه آن است.

- پشتیبانی پیشین و پشتیبانی متعاقب آن:

این ستون ها به ترتیب مقادیر پشتیبانی را برای پیشین و متعاقب نشان می دهند. پشتیبانی نسبت تراکنش هایی را که شامل آیتم یا مجموعه اقلام مربوطه می شود اندازه گیری می کند.

- حمایت کردن:

این ستون مقدار پشتیبانی را برای کل قانون نشان می دهد، یعنی نسبت تراکنش هایی که هم مقدمه و هم متعاقب آن را شامل می شود.

- اطمینان:

اطمینان احتمال وقوع نتیجه را با توجه به مقدمه اندازه گیری می کند. این به عنوان پشتیبانی از مقدم محاسبه می شود و در نتیجه تقسیم بر حمایت از مقدم است.

- لیفت:

لیفت اندازه گیری می کند که در مقایسه با مستقل بودن آنها چقدر احتمال بیشتری وجود دارد که مقدم و پیامد با هم اتفاق بیفتند. به صورت (پشتیبانی از پیشین و متعاقب) / (پشتیبانی از پیشین * پشتیبانی از نتیجه) محاسبه می شود.

- قدرت نفوذ:

اهرم تفاوت بین فرکانس مشاهده شده پیشین و متعاقب آن را می سنجد و در صورت مستقل بودن چه چیزی انتظار می رود.

- محکومیت:

قانع بودن معیاری است که نشان می دهد تا چه اندازه نتیجه بر مقدم است. یک ارزش اعتقادی بالا نشان می دهد که نتیجه به شدت به پیشین بستگی دارد.

- zhangs_metric:

متریک ژانگ یکی دیگر از معیارهای ارتباط بین موارد در یک قانون است.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(frankfurter)	(citrus fruit)	0.042432	0.056465	0.004009	0.094488	1.673391	0.001613	1.041991	0.420243

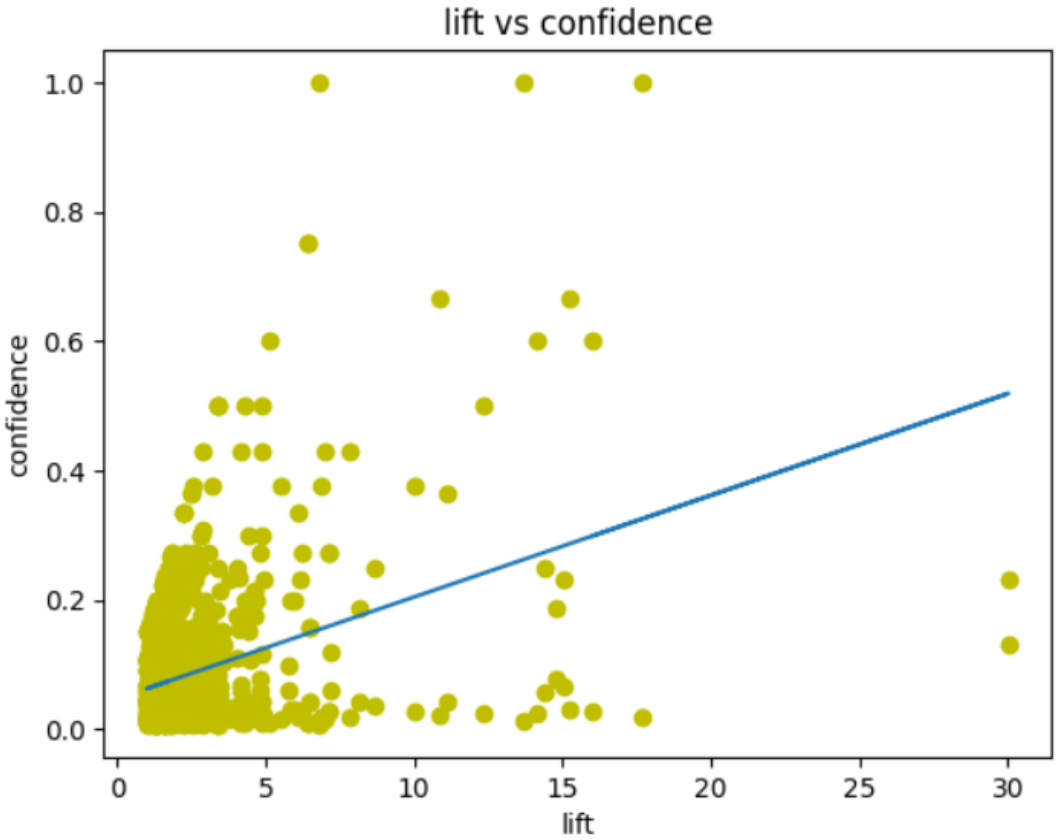
تفسیر یکی از سطرها:

نشان می دهد که اگر تراکنش حاوی "فرانکفورتر" باشد، احتمال ۴.۰۰۹٪ وجود دارد که حاوی "مرکبات" نیز باشد.

اطمینان ۹.۴۴۸۸ درصد نشان می دهد که ۹.۴۴۸۸ درصد از معاملات حاوی «فرانکفورتر» نیز حاوی «مرکبات» است.

افزایش ۱.۶۷۳۳۹۱ نشان می دهد که "مرکبات" ۱.۶۷۳۳۹۱ برابر بیشتر از احتمال خرید "فرانکفورتر" در مقایسه با احتمال خرید آن بیشتر است.

نمودار لیفت در مقایسه با اطمینان:



در نهایت سطرهای که با اطمینان حداقل ۲۰ درصد بودند را بررسی کردیم:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(frankfurter, yogurt)	(citrus fruit)	0.003675	0.056465	0.001002	0.272727	4.830016	0.000795	1.297361	0.795886
1	(frankfurter, citrus fruit)	(yogurt)	0.004009	0.088206	0.001002	0.250000	2.834280	0.000649	1.215726	0.649782
2	(frankfurter, sausage)	(soda)	0.002005	0.102907	0.001002	0.500000	4.858766	0.000796	1.794186	0.795782
3	(frankfurter, soda)	(sausage)	0.004343	0.061477	0.001002	0.230769	3.753763	0.000735	1.220080	0.736801
4	(frankfurter, rolls/buns)	(whole milk)	0.004009	0.147344	0.001336	0.333333	2.262283	0.000746	1.278984	0.560215

۴۰ درصد

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(frankfurter, sausage)	(soda)	0.002005	0.102907	0.001002	0.500000	4.858766	0.000796	1.794186	0.795782
1	(frankfurter, candy)	(citrus fruit)	0.001002	0.056465	0.001002	1.000000	17.710059	0.000946	inf	0.944482
2	(citrus fruit, candy)	(frankfurter)	0.001671	0.042432	0.001002	0.600000	14.140157	0.000931	2.393919	0.930834
3	(butter, sausage)	(bottled beer)	0.002005	0.043769	0.001336	0.666667	15.231552	0.001249	2.868694	0.936224
4	(butter, bottled beer)	(sausage)	0.002005	0.061477	0.001336	0.666667	10.844203	0.001213	2.815570	0.909608

۲-۵-۲- سایر نمونه : ۵۰ درصد

- Fp Growth execution took: 0.32078051567077637 seconds
- ماتریس خروجی:

	support	itemsets	number_of_items
0	0.158514	(whole milk)	1
1	0.109596	(rolls/buns)	1
2	0.070837	(root vegetables)	1
3	0.001738	(nut snack)	1
4	0.041700	(whipped/sour cream)	1

پس از لیفت:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(whipped/sour cream)	(pip fruit)	0.041700	0.046645	0.002005	0.048077	1.030692	0.000060	1.001504	0.031074
1	(pip fruit)	(whipped/sour cream)	0.046645	0.041700	0.002005	0.042980	1.030692	0.000060	1.001337	0.031235
2	(frankfurter)	(citrus fruit)	0.039428	0.056536	0.002807	0.071186	1.259142	0.000578	1.015774	0.214256
3	(citrus fruit)	(frankfurter)	0.056536	0.039428	0.002807	0.049645	1.259142	0.000578	1.010751	0.218141
4	(frankfurter)	(pip fruit)	0.039428	0.046645	0.002272	0.057627	1.235433	0.000433	1.011653	0.198389

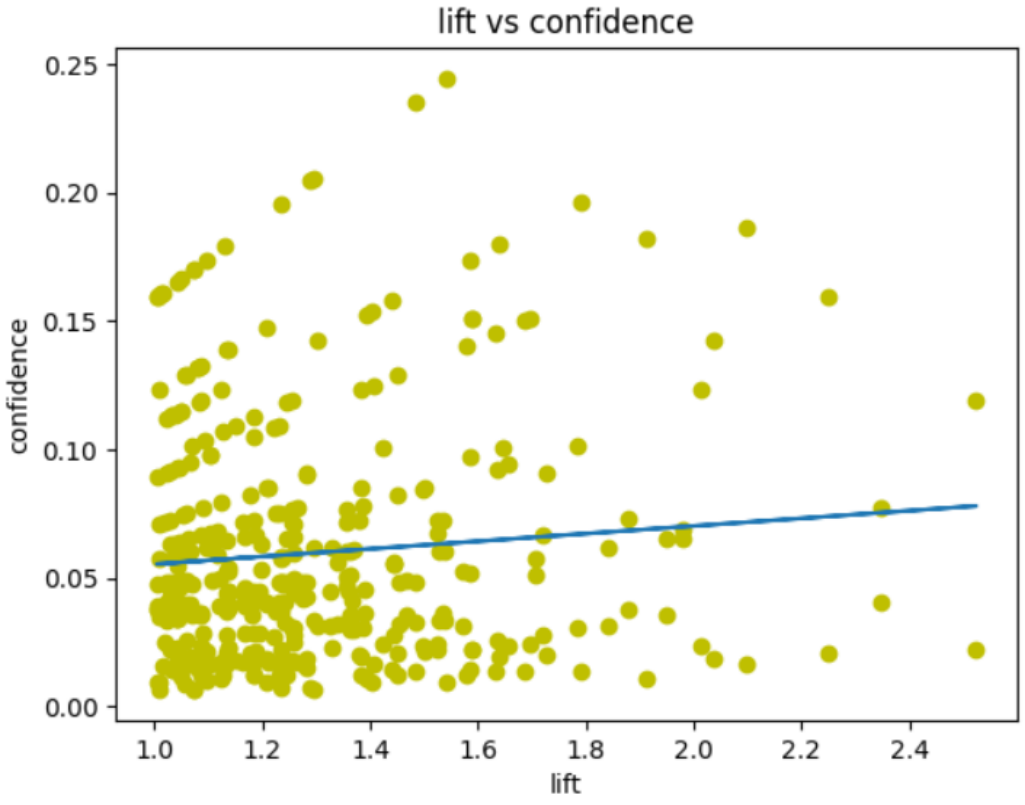
الگوهای شناسایی شده با اطمینان حداقل ۲۰ درصد:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(cat food)	(whole milk)	0.011227	0.158514	0.001871	0.166667	1.051433	0.000092	1.009783	0.049473
1	(fruit/vegetable juice)	(whole milk)	0.033681	0.158514	0.005079	0.150794	0.951297	-0.000260	0.990909	-0.050315
2	(candy)	(whole milk)	0.015771	0.158514	0.002539	0.161017	1.015792	0.000039	1.002984	0.015795
3	(oil)	(soda)	0.014167	0.095028	0.002138	0.150943	1.588409	0.000792	1.065856	0.375763
4	(seasonal products)	(rolls/buns)	0.006950	0.109596	0.001069	0.153846	1.403752	0.000308	1.052295	0.289637

در اینجا با قرار دادن حداقل اطمینان ۴۰ درصد خروجی نداشتیم! پس برای مقایسه ۱۵ درصد را به عنوان حد آستانه درنظر گرفتیم:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(cat food)	(whole milk)	0.011227	0.158514	0.001871	0.166667	1.051433	0.000092	1.009783	0.049473
1	(fruit/vegetable juice)	(whole milk)	0.033681	0.158514	0.005079	0.150794	0.951297	-0.000260	0.990909	-0.050315
2	(candy)	(whole milk)	0.015771	0.158514	0.002539	0.161017	1.015792	0.000039	1.002984	0.015795
3	(oil)	(soda)	0.014167	0.095028	0.002138	0.150943	1.588409	0.000792	1.065856	0.375763
4	(seasonal products)	(rolls/buns)	0.006950	0.109596	0.001069	0.153846	1.403752	0.000308	1.052295	0.289637

نمودار لیفت در مقایسه با اطمینان:



۲-۵-۳- اندازه نمونه : ۸۰ درصد

- Fp Growth execution took: 0.5899908542633057 seconds

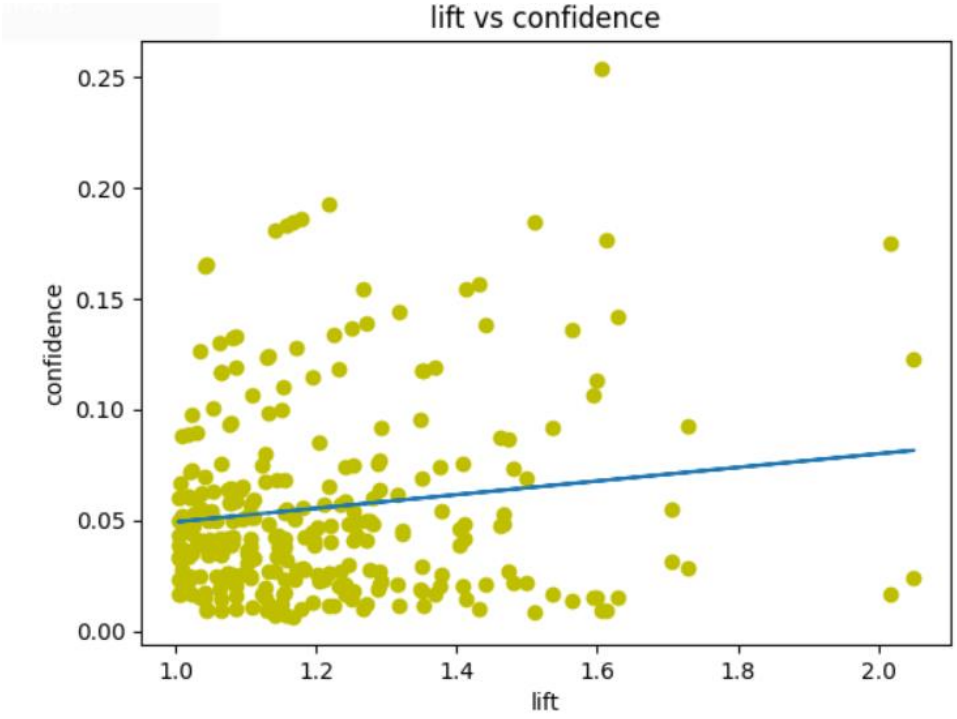
خروجی ماتریس:

	support	itemsets	number_of_items
0	0.158062	(whole milk)	1
1	0.109357	(rolls/buns)	1
2	0.070677	(root vegetables)	1
3	0.001420	(nut snack)	1
4	0.042941	(whipped/sour cream)	1

پس از لیفت:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(frankfurter)	(other vegetables)	0.039181	0.122139	0.005180	0.132196	1.082345	0.000394	1.011590	0.079182
1	(other vegetables)	(frankfurter)	0.122139	0.039181	0.005180	0.042408	1.082345	0.000394	1.003369	0.086665
2	(brown bread)	(pastry)	0.037427	0.051295	0.002005	0.053571	1.044381	0.000085	1.002405	0.044147
3	(pastry)	(brown bread)	0.051295	0.037427	0.002005	0.039088	1.044381	0.000085	1.001729	0.044793
4	(frankfurter)	(brown bread)	0.039181	0.037427	0.001587	0.040512	1.082423	0.000121	1.003215	0.079252

نمودار لیفت در مقایسه با اطمینان:



الگوهای شناسایی شده با اطمینان حداقل ۲۰ درصد:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(yogurt, sausage)	(whole milk)	0.005597	0.158062	0.00142	0.253731	1.605266	0.000535	1.128197	0.379173

الگوهای شناسایی شده با اطمینان حداقل ۱۵درصد:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(candy)	(whole milk)	0.014703	0.158062	0.002423	0.164773	1.042457	0.000099	1.008035	0.041336
1	(seasonal products)	(rolls/buns)	0.006934	0.109357	0.001086	0.156627	1.432253	0.000328	1.056048	0.303907
2	(rolls/buns, soda)	(other vegetables)	0.008104	0.122139	0.001253	0.154639	1.266095	0.000263	1.038446	0.211887
3	(grapes)	(whole milk)	0.014536	0.158062	0.002256	0.155172	0.981720	-0.000042	0.996580	-0.018545
4	(yogurt, soda)	(whole milk)	0.005931	0.158062	0.001086	0.183099	1.158399	0.000149	1.030648	0.137555

۲-۵-٤- کل مجموعه داده:

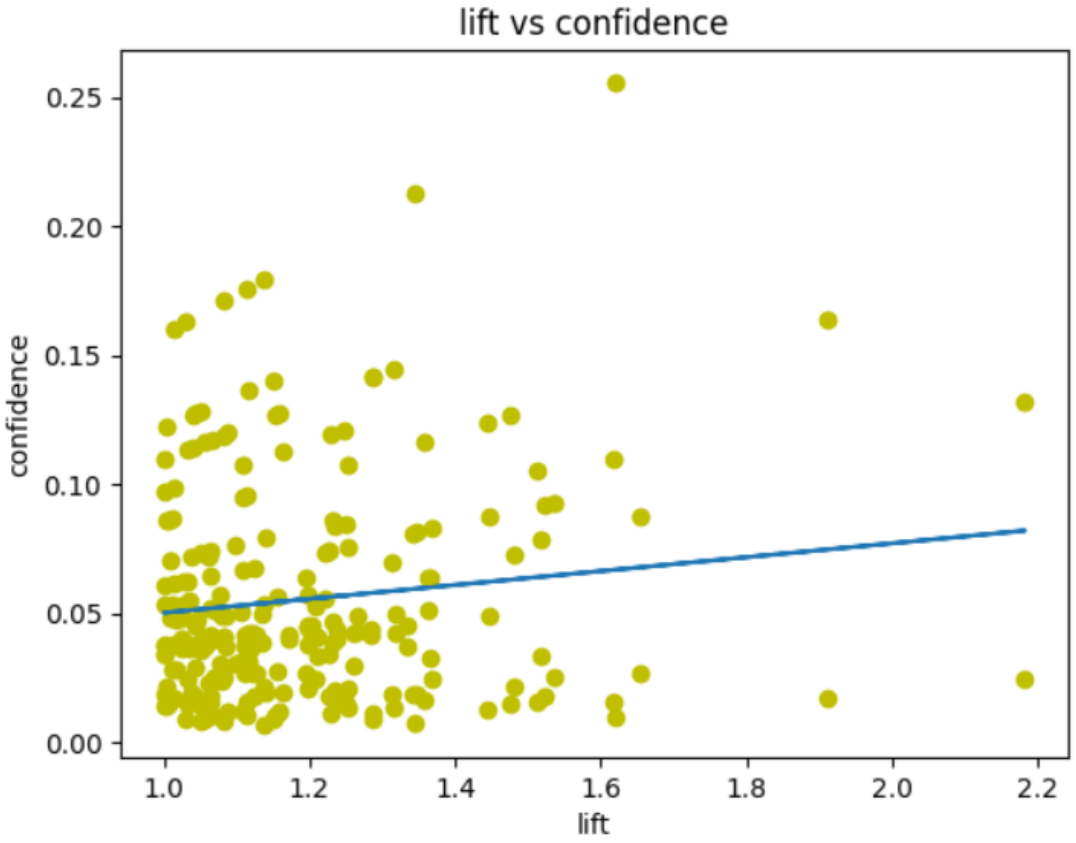
- Fp Growth execution took: 0.5794949531555176 seconds
- ماتریس خروجی:

	support	itemsets	number_of_items
0	0.157923	(whole milk)	1
1	0.110005	(rolls/buns)	1
2	0.069572	(root vegetables)	1
3	0.001470	(nut snack)	1
4	0.043708	(whipped/sour cream)	1

-
- بعد از لیفت:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(frankfurter)	(other vegetables)	0.037760	0.122101	0.005146	0.136283	1.116150	0.000536	1.016420	0.108146
1	(other vegetables)	(frankfurter)	0.122101	0.037760	0.005146	0.042146	1.116150	0.000536	1.004579	0.118536
2	(tropical fruit)	(cat food)	0.067767	0.011829	0.001002	0.014793	1.250543	0.000201	1.003008	0.214911
3	(cat food)	(tropical fruit)	0.011829	0.067767	0.001002	0.084746	1.250543	0.000201	1.018551	0.202746
4	(brown bread)	(pastry)	0.037626	0.051728	0.002005	0.053286	1.030127	0.000059	1.001646	0.030389

نمودار لیفت در مقایسه با اطمینان:



الگوهای شناسایی شده با اطمینان حداقل ۲۰ درصد:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(sausage, yogurt)	(whole milk)	0.005597	0.158062	0.00142	0.253731	1.605266	0.000535	1.128197	0.379173

الگوهای شناسایی شده با اطمینان حداقل ۱۵ درصد:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(candy)	(whole milk)	0.014703	0.158062	0.002423	0.164773	1.042457	0.000099	1.008035	0.041336
1	(seasonal products)	(rolls/buns)	0.006934	0.109357	0.001086	0.156627	1.432253	0.000328	1.056048	0.303907
2	(rolls/buns, soda)	(other vegetables)	0.008104	0.122139	0.001253	0.154639	1.266095	0.000263	1.038446	0.211887
3	(grapes)	(whole milk)	0.014536	0.158062	0.002256	0.155172	0.981720	-0.000042	0.996580	-0.018545
4	(yogurt, soda)	(whole milk)	0.005931	0.158062	0.001086	0.183099	1.158399	0.000149	1.030648	0.137555

۲-۶-اعمال الگوریتم Apriori

۲-۶-۱- اندازه نمونه : ۲۰ درصد

زمان اجرای الگوریتم:

Apriori Execution took: 1.0187361240386963 seconds

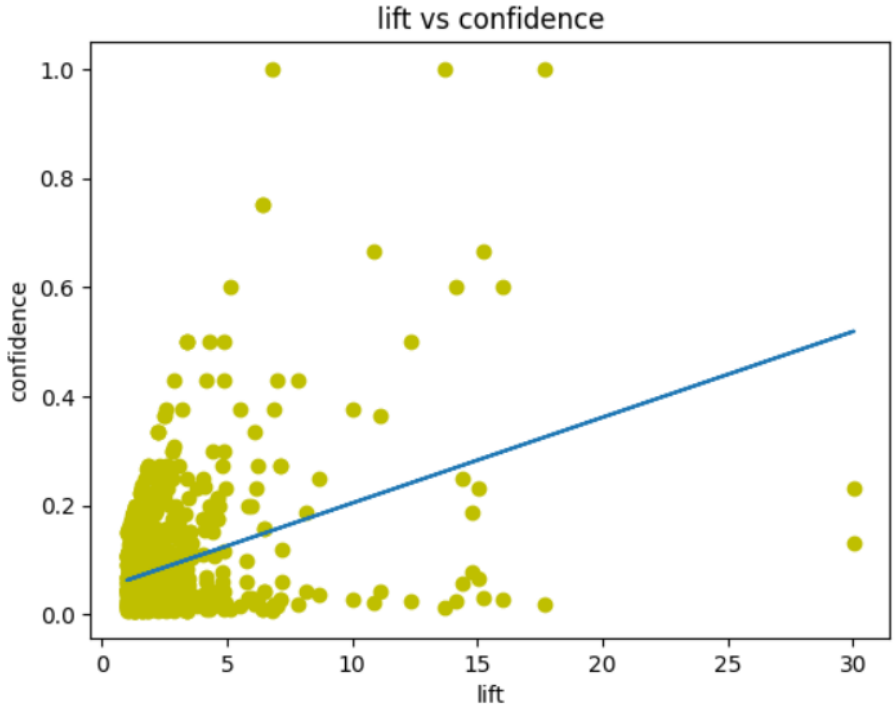
ماتریس خروجی:

	support	itemsets	number_of_items
0	0.004343	(Instant food products)	1
1	0.020047	(UHT-milk)	1
2	0.002673	(abrasive cleaner)	1
3	0.002673	(artif. sweetener)	1
4	0.005680	(baking powder)	1

پس از لیفت:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(pip fruit)	(Instant food products)	0.046442	0.004343	0.001002	0.021583	4.969009	0.000801	1.017620	0.837655
1	(Instant food products)	(pip fruit)	0.004343	0.046442	0.001002	0.230769	4.969009	0.000801	1.239626	0.802237
2	(whole milk)	(Instant food products)	0.147344	0.004343	0.001002	0.006803	1.566196	0.000362	1.002476	0.423981
3	(Instant food products)	(whole milk)	0.004343	0.147344	0.001002	0.230769	1.566196	0.000362	1.108453	0.363087
4	(bottled beer)	(UHT-milk)	0.043769	0.020047	0.001002	0.022901	1.142366	0.000125	1.002921	0.130328

نمودار لیفت در مقایسه با اطمینان



۲-۶-۲- اندازه نمونه : ۵۰ درصد

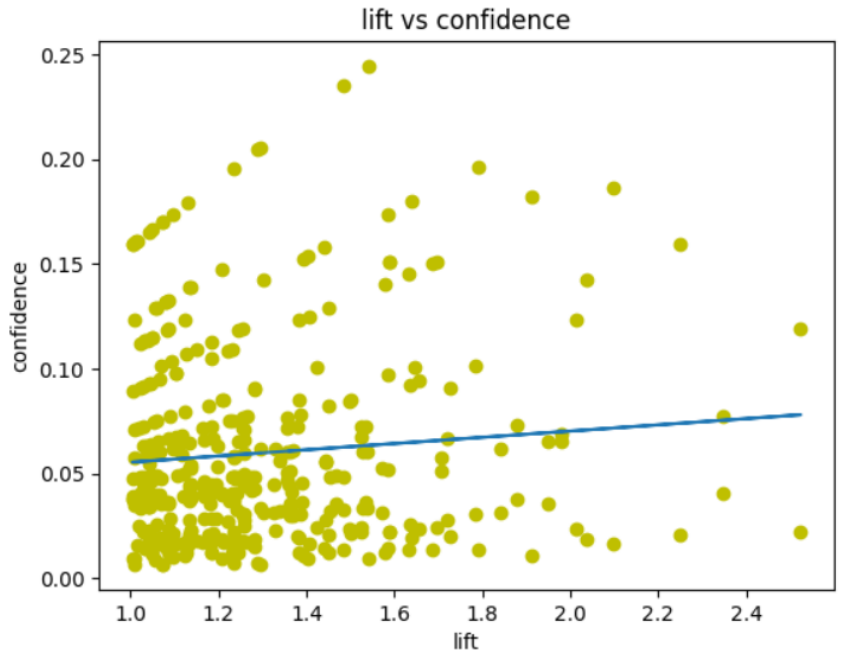
- Apriori Execution took: 1.2805633544921875 seconds
- ماتریس خروجی:

	support	itemsets	number_of_items
0	0.004411	(Instant food products)	1
1	0.022053	(UHT-milk)	1
2	0.002138	(abrasive cleaner)	1
3	0.002272	(artif. sweetener)	1
4	0.007217	(baking powder)	1

بعد از لیفت کردن:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(bottled beer)	(UHT-milk)	0.044507	0.022053	0.001337	0.030030	1.361725	0.000355	1.008224	0.278011
1	(UHT-milk)	(bottled beer)	0.022053	0.044507	0.001337	0.060606	1.361725	0.000355	1.017138	0.271628
2	(frankfurter)	(UHT-milk)	0.039428	0.022053	0.001337	0.033898	1.537134	0.000467	1.012261	0.363782
3	(UHT-milk)	(frankfurter)	0.022053	0.039428	0.001337	0.060606	1.537134	0.000467	1.022544	0.357319
4	(sausage)	(UHT-milk)	0.061347	0.022053	0.001470	0.023965	1.086710	0.000117	1.001959	0.085006

مقایسه اطمینان با لیفت:



الگوی شناسایی شده با اطمینان حداقل ۲۰ درصد:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(spread cheese)	(whole milk)	0.005881	0.158514	0.001203	0.204545	1.290396	0.000271	1.057868	0.226375
1	(rolls/buns, yogurt)	(whole milk)	0.009088	0.158514	0.002138	0.235294	1.484377	0.000698	1.100405	0.329309
2	(sausage, yogurt)	(whole milk)	0.006014	0.158514	0.001470	0.244444	1.542102	0.000517	1.113732	0.353662
3	(yogurt, tropical fruit)	(whole milk)	0.005213	0.158514	0.001069	0.205128	1.294072	0.000243	1.058644	0.228436

الگوی شناسایی شده با اطمینان حداقل ۱۵ درصد:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(candy)	(whole milk)	0.015771	0.158514	0.002539	0.161017	1.015792	0.000039	1.002984	0.015795
1	(cat food)	(whole milk)	0.011227	0.158514	0.001871	0.166667	1.051433	0.000092	1.009783	0.049473
2	(chocolate)	(whole milk)	0.024325	0.158514	0.003742	0.153846	0.970554	-0.000114	0.994484	-0.030158
3	(condensed milk)	(whole milk)	0.006282	0.158514	0.001069	0.170213	1.073804	0.000073	1.014099	0.069166
4	(detergent)	(whole milk)	0.008019	0.158514	0.001203	0.150000	0.946290	-0.000068	0.989984	-0.054121

۲-۶-۳- اندازه نمونه : ۸۰ درصد

- Apriori Execution took: 4.373556852340698 seconds

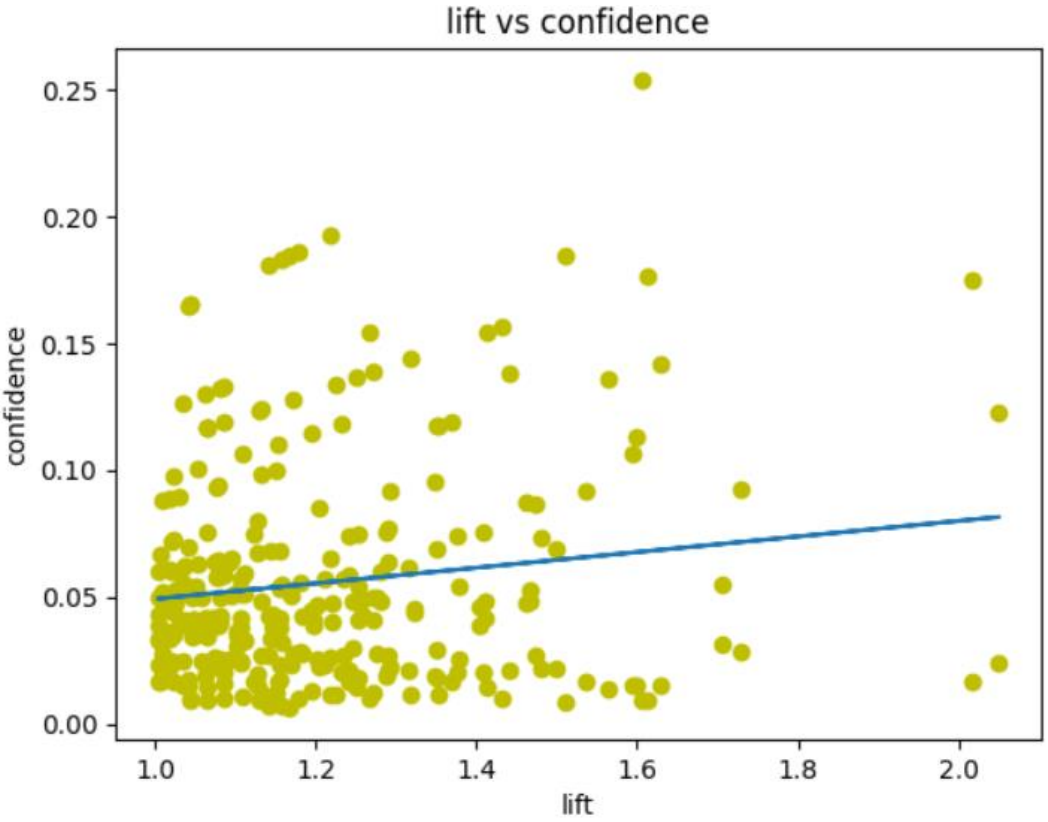
ماتریس خروجی:

	support	itemsets	number_of_items
0	0.004177	(Instant food products)	1
1	0.021303	(UHT-milk)	1
2	0.001504	(abrasive cleaner)	1
3	0.002005	(artif. sweetener)	1
4	0.007352	(baking powder)	1

پس از لیفت:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(bottled beer)	(UHT-milk)	0.046032	0.021303	0.001086	0.023593	1.107505	0.000105	1.002346	0.101753
1	(UHT-milk)	(bottled beer)	0.021303	0.046032	0.001086	0.050980	1.107505	0.000105	1.005214	0.099183
2	(frankfurter)	(UHT-milk)	0.039181	0.021303	0.001003	0.025586	1.201054	0.000168	1.004396	0.174224
3	(UHT-milk)	(frankfurter)	0.021303	0.039181	0.001003	0.047059	1.201054	0.000168	1.008267	0.171041
4	(brown bread)	(beef)	0.037427	0.034252	0.001420	0.037946	1.107851	0.000138	1.003840	0.101136

نمودار مقایسه اطمینان با لیفت:



الگوی شناسایی شده با اطمینان حداقل ۲۰ درصد:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(sausage, yogurt)	(whole milk)	0.005597	0.158062	0.00142	0.253731	1.605266	0.000535	1.128197	0.379173

الگوی شناسایی شده با اطمینان حداقل ۱۵ درصد:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(bottled beer)	(whole milk)	0.046032	0.158062	0.007268	0.157895	0.998943	-0.000008	0.999802	-0.001108
1	(candy)	(whole milk)	0.014703	0.158062	0.002423	0.164773	1.042457	0.000099	1.008035	0.041336
2	(chewing gum)	(whole milk)	0.011947	0.158062	0.001838	0.153846	0.973329	-0.000050	0.995018	-0.026985
3	(condensed milk)	(whole milk)	0.006015	0.158062	0.001086	0.180556	1.142310	0.000135	1.027450	0.125335
4	(detergent)	(whole milk)	0.008605	0.158062	0.001420	0.165049	1.044202	0.000060	1.008368	0.042699

۲-۶-۴- کل داده ها

- Apriori Execution took: 2.9995524883270264 seconds

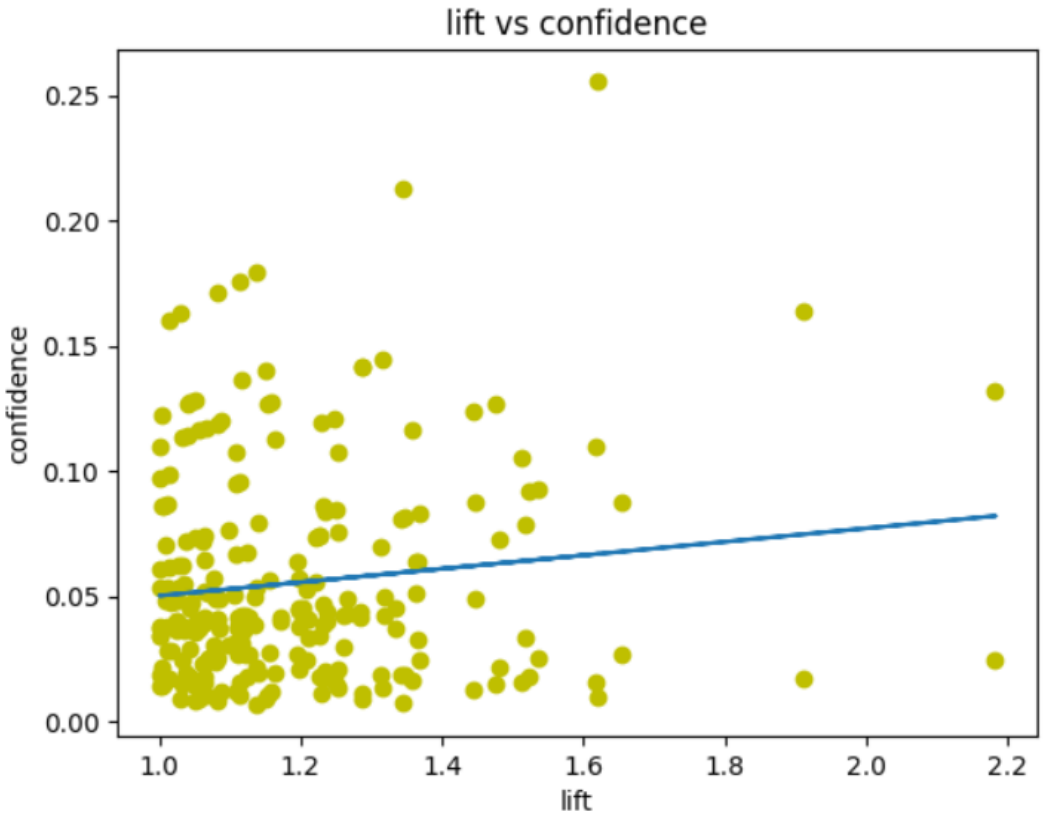
ماتریس خروجی:

	support	itemsets	number_of_items
0	0.004010	(Instant food products)	1
1	0.021386	(UHT-milk)	1
2	0.001470	(abrasive cleaner)	1
3	0.001938	(artif. sweetener)	1
4	0.008087	(baking powder)	1

بعد از لیفت:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(tropical fruit)	(UHT-milk)	0.067767	0.021386	0.001537	0.022682	1.060617	8.785064e-05	1.001326	0.061307
1	(UHT-milk)	(tropical fruit)	0.021386	0.067767	0.001537	0.071875	1.060617	8.785064e-05	1.004426	0.058402
2	(brown bread)	(beef)	0.037626	0.033950	0.001537	0.040853	1.203301	2.597018e-04	1.007196	0.175559
3	(beef)	(brown bread)	0.033950	0.037626	0.001537	0.045276	1.203301	2.597018e-04	1.008012	0.174891
4	(citrus fruit)	(beef)	0.053131	0.033950	0.001804	0.033962	1.000349	6.297697e-07	1.000012	0.000369

نمودار مقایسه اطمنیان با لیفت:



الگوی شناسایی شده با اطمینان حداقل ۲۰ درصد:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(sausage, rolls/buns)	(whole milk)	0.005347	0.157923	0.001136	0.212500	1.345594	0.000292	1.069304	0.258214
1	(sausage, yogurt)	(whole milk)	0.005748	0.157923	0.001470	0.255814	1.619866	0.000563	1.131541	0.384877

الگوی شناسایی شده با اطمینان حداقل ۱۵ درصد:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(bottled beer)	(whole milk)	0.045312	0.157923	0.007151	0.157817	0.999330	-0.000005	0.999874	-0.000702
1	(detergent)	(whole milk)	0.008621	0.157923	0.001403	0.162791	1.030824	0.000042	1.005814	0.030162
2	(frozen fish)	(whole milk)	0.006817	0.157923	0.001069	0.156863	0.993287	-0.000007	0.998743	-0.006759
3	(ham)	(whole milk)	0.017109	0.157923	0.002740	0.160156	1.014142	0.000038	1.002659	0.014188
4	(semi-finished bread)	(whole milk)	0.009490	0.157923	0.001671	0.176056	1.114825	0.000172	1.022008	0.103985

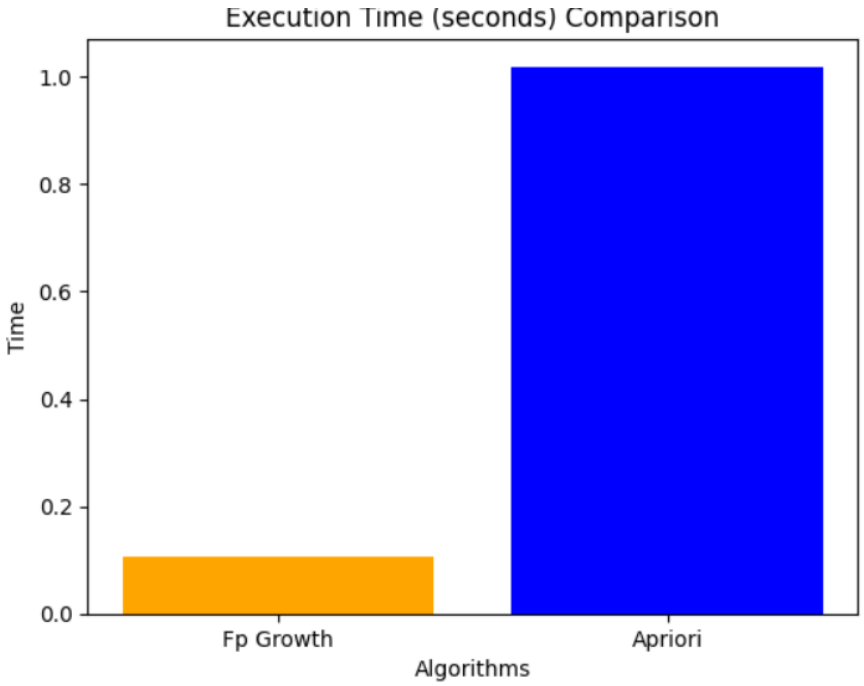
۲ـ۷ـ

فصل سوم

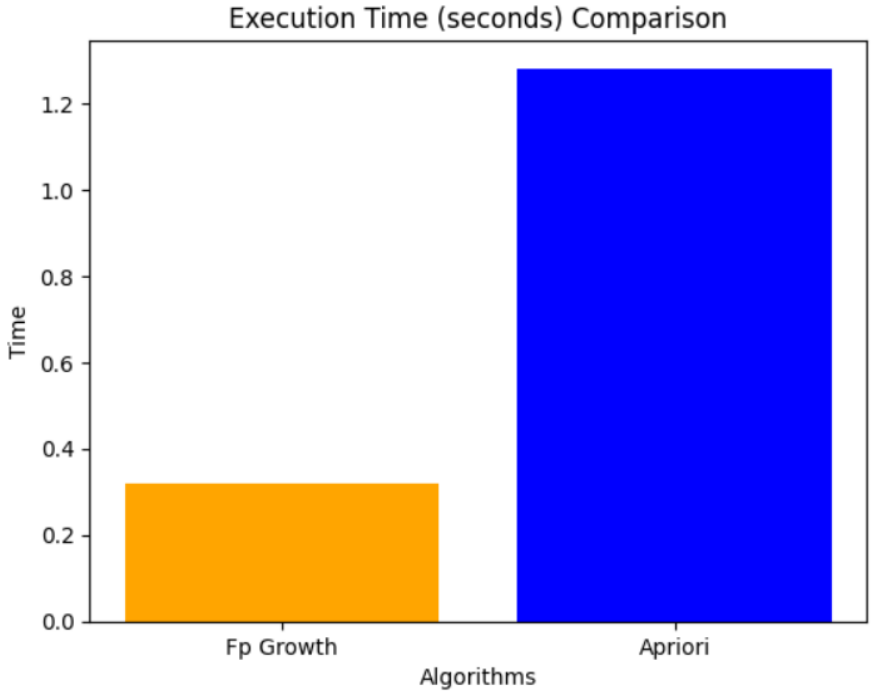
مقایسه و جمع‌بندی

۳-۱- مقایسه زمانی:

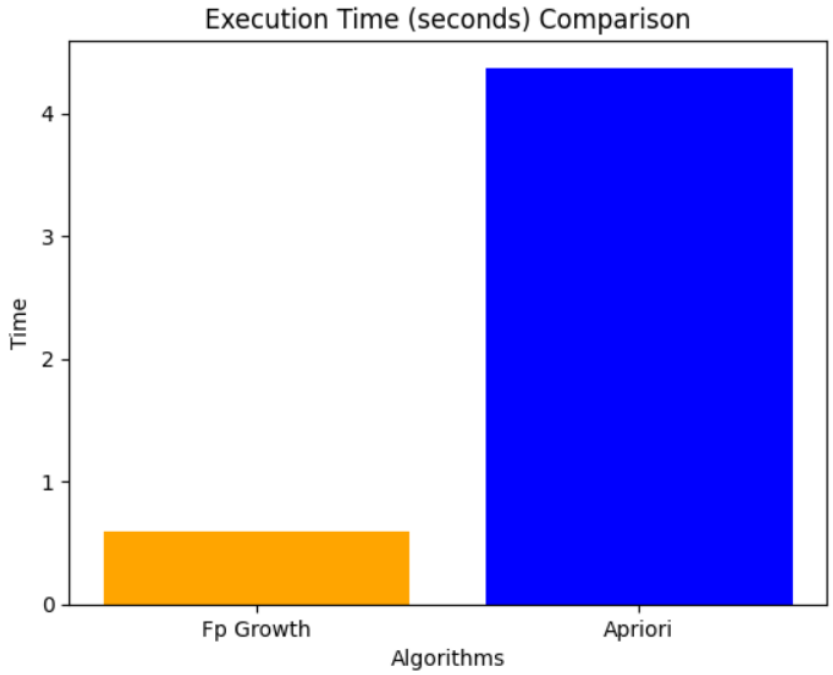
۳-۱-۱- سایز نمونه : ۲۰ درصد



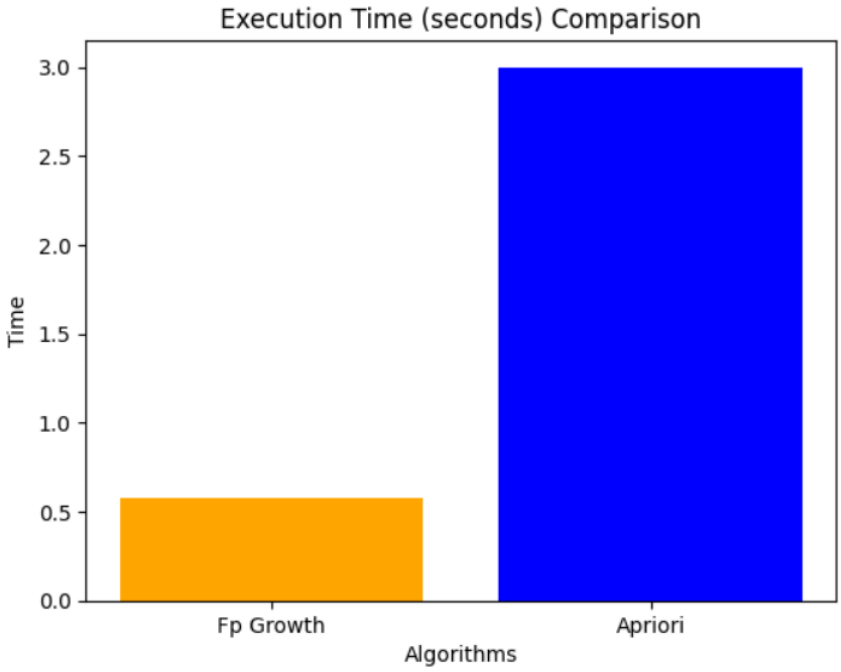
۳-۱-۲- سایز نمونه : ۵۰ درصد



۳-۱-۳- سایز نمونه : ۸۰ درصد



۳-۱-۱- سایز نمونه : ۱۰۰ درصد



پس بصورت کلی مشاهده میشود عملکرد الگوریتم Fp-growth به مراتب در سایز نمونه مشابه بهتر بوده است.

۳-۲- مقایسه الگوی شناسایی شده

۳-۲-۱- سایز نمونه : ۲۰ درصد با اطمینان ۲۰ درصد

	pattern
Fp-growth	(frankfurter, yogurt)
Apriori	(pip fruit) (Instant food products)

۳-۲-۲- سایز نمونه : ۲۰ درصد با اطمینان ۱۵ درصد

	pattern
Fp-growth	(frankfurter) (yogurt, citrus fruit)
Apriori	(newspapers) (cream)

۳-۲-۳- سایز نمونه : ۵۰ درصد با اطمینان ۲۰ درصد

	pattern
Fp-growth	(rolls/buns, yogurt)
Apriori	(whole milk) (spread cheese)

۳-۲-٤- سایز نمونه : ۵۰ درصد با اطمینان ۱۵ درصد

	pattern
Fp-growth	(fruit/vegetable juice)
Apriori	(whole milk) (candy)

۳-۲-۵- سایز نمونه : ۸۰ درصد با اطمینان ۲۰ درصد

	pattern
Fp-growth	(whole milk) (yogurt, sausage)
Apriori	(whole milk) (sausage, yogurt)

۳-۲-٦- سایز نمونه : ۸۰ درصد با اطمینان ۱۵ درصد

	pattern
Fp-growth	(whole milk) (candy)
Apriori	(whole milk) (bottled beer)

۳-۲-۷- سایز نمونه : ۱۰۰ درصد با اطمینان ۲۰ درصد

	pattern
Fp-growth	(whole milk) (sausage, yogurt)
Apriori	(whole milk) (sausage, rolls/buns)

۳-۲-۸- سایز نمونه : ۱۰۰ درصد با اطمینان ۱۵ درصد

	pattern
Fp-growth	(whole milk) (candy)
Apriori	(whole milk) (bottled beer)

منابع و مراجع

[1] https://www.kaggle.com/datasets/heeraldedhia/groceries-dataset

پیوست‌ها

- **https://colab.research.google.com/drive/1R6M0JSQpC6jNE8NZNzxMbbfmTPAltaeN?usp=drive_link**

Abstract

In this project, by reviewing and comparing Apriori and FP-Growth algorithms on different data sets with different sample sizes, the sensitivity analysis of the algorithms was performed. Employed outstanding implementation practices for handling large Growth datasets. FP-Growth algorithm performed faster and more efficiently than Apriori. The results clearly show useful patterns and relationships in the data. Accurate interpretation and optimization of parameters are key tools in extracting information from extensive and growing data.

Key Words: Apriori algorithm, FP-Growth algorithm, Frequency patterns detection

Amirkabir University of Technology
(Tehran Polytechnic)

Department of Computer Science

Project 4

Advanced methods in exploring repetitive patterns

By
Samin Mahdipour

Supervisor
Dr.Ghatee

Advisor
Dr.Yousofi Mehr

November 2023