بسمه تعالی

تجارت الکترونیک

سید مهدی مهدوی مرتضوی

تمرین سوم: تحلیل سبد خرید و طراحی سیستم پیشنهاد دهنده توسط ARM

۱. مقدمه و هدف تمرین:

هدف اصلی این پروژه، تحلیل سبدهای خرید مشتریان یک فروشگاه آنلاین برای کشف الگوهای پنهان در خریدهای همزمان محصولات است. با استفاده از الگوریتم Apriori، ما به دنبال پاسخ به این سوالات هستیم:

- چه محصولاتی تمایل دارند با هم خریداری شوند؟
- قوی‌ترین ارتباطات بین کدام محصولات وجود دارد؟
- چگونه می‌توان از این الگوها برای بهبود استراتژی‌های فروش استفاده کرد؟

این تحلیلات به کسب‌وکار ها کمک می‌کند تا استراتژی‌های مبتنی بر داده را برای بازاریابی، چیدمان محصولات، پیشنهادات ترکیبی بین کالا های خریداری شده و مدیریت موجودی پیاده‌سازی کنند.

۲. تسک اول: پیش‌پردازش داده‌ها

داده‌های خام فروشگاه‌های آنلاین معمولاً شامل نویز، مقادیر null و تراکنش‌های نامربوط هستند. این تسک با هدف آماده‌سازی داده‌های خام برای تحلیل‌های آتی طراحی شده است.

مراحل کلیدی و منطق پشت آن‌ها:

۱. **بارگذاری داده‌ها با دو روش متفاوت**:
   - داده‌های کوچک (محصولات، دپارتمان‌ها، راهروها): استفاده از pandas برای سرعت بالا
   - داده‌های بزرگ (سفارش‌ها، محصولات سفارش‌ها): استفاده از dask برای مدیریت حافظه
   - منطق: بهینه‌سازی مصرف منابع با توجه به حجم داده‌هاستون‌های Volume وMarket
۲. **حذف داده‌های نامرتبط**:

- حذف سفارشات تک‌کالایی: این سفارش‌ها برای تحلیل association rule ها بی‌معنا هستند
- منطق: association rule ها نیاز به تعامل بین **حداقل دو محصول** دارند
- تأثیر: کاهش حجم داده‌ها با حفظ کیفیت تحلیلی

۳. نمونه‌گیری هوشمند:

- انتخاب ۲۰،۰۰۰ کاربر به صورت تصادفی
- منطق:
  - ایجاد مجموعه داده های قابل مدیریت از نظر محاسبات
  - حفظ تنوع رفتاری کاربران (نمونه گیری به تعداد مناسب)
  - امکان اجرای الگوریتم‌های پیچیده تر در زمان معقول

۴. ذخیره‌سازی ساختاریافته:

- ایجاد ساختار پوشه‌ای منظم برای داده‌های پردازش‌شده (ذخیره سازی نتایج در فولدر مشخص (processed_data))
- امکان بازیابی فایل ها و استفاده از نتایج بدست آمده از تسک اول در مراحل بعدی

۳. تسک دوم: کدگذاری سبدهای خرید

این تسک به منظور تبدیل داده‌های تراکنشی به فرمتی که برای الگوریتم Apriori قابل فهم باشد، استفاده میگردد؛ چالش اصلی در این تسک، مدیریت ابعاد بالا (ده‌ها هزار محصول مختلف) هست:

استراتژی‌های به‌کارگرفته‌شده:

۱. ایجاد ساختار سبد خرید

- گروه‌بندی محصولات بر اساس شماره سفارش (order_id)
- تبدیل داده‌های خطی به ساختار سلسله مراتبی

۲. فیلتر کردن محصولات کم‌تکرار:

- معیار فیلتر: حضور در حداقل 0.5% سبدهای خرید
- منطق:
- کاهش ابعاد ماتریس از ده‌ها هزار نمونه به چند صد ستون
- حذف نویز ها و تمرکز بر الگوهای معنادار
- بهبود کارایی محاسباتی

۳. **کدگذاری یک‌دست (One-Hot Encoding):**
   - ایجاد ماتریس باینری (شامل True/False)
   - مزایا:
     - سادگی تفسیر
     - سازگاری با الگوریتم Apriori
     - کارایی حافظه (memory efficiency)
۴. **حذف سبدهای خرید پراکنده:**
   - حذف سبدهایی با کمتر از ۲ محصول
   - تضمین کیفیت: اطمینان از وجود تعاملات محصولی

بهینه‌سازی‌های انجام‌شده:

۱. **بهینه‌سازی حافظه:**
   - استفاده از dtype=np.bool_ به جای int
   - ذخیره حدود ۲ گیگابایت فضای حافظه
۲. **بهینه‌سازی سرعت:**
   - استفاده از numpy برای عملیات برداری
   - جایگزینی حلقه‌های پایتون با عملیات آرایه‌ای
۳. **بهینه‌سازی کیفیت داده:**
   - تعادل بین حفظ اطلاعات و کاهش ابعاد
   - فیلتر دو مرحله‌ای: ابتدا گروه بندی بر اساس محصولات و بعد حذف سبدهای تک کالایی

۴. تحلیل تسک سوم: استخراج frequent itemset ها

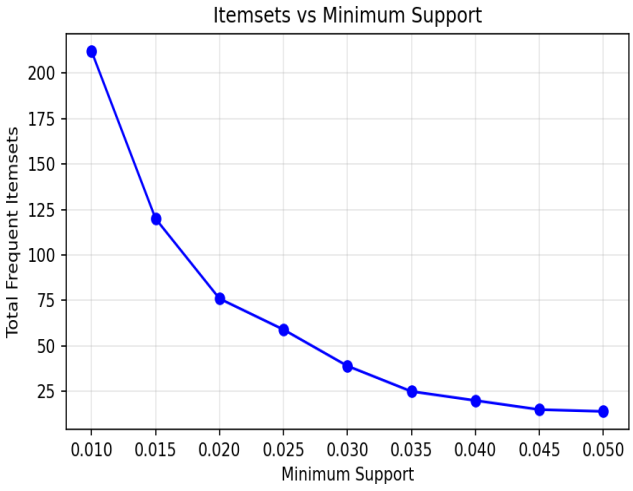
در این مرحله، الگوریتم Apriori بر روی ماتریس کدگذاری‌شده (به صورت True/False) سبدهای خرید اجرا میشود که شامل مراحل زیر است:

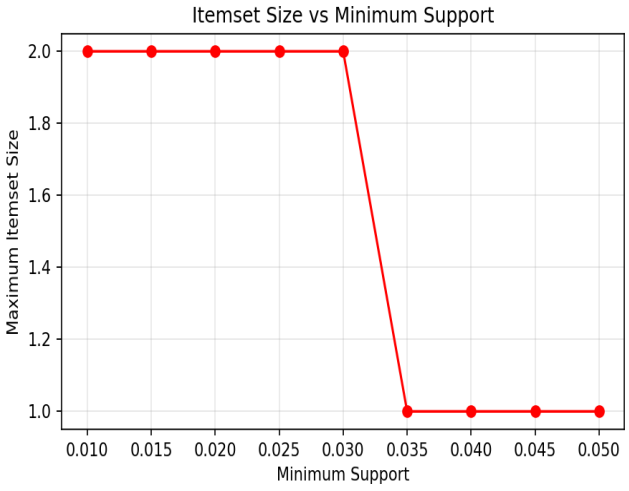
۱. **اکتشاف پارامتر بهینه:**
   - آزمایش ۹ مقدار مختلف برای حداقل پشتیبانی(min_support) از ۰/۰۵ تا ۰/۰۱
   - تحلیل تأثیر هر پارامتر بر تعداد و اندازه مجموعه‌های مکرر (frequent itemset ها)
   - شناسایی نقاط عطف در رفتار الگوریتم

۲. **یافته‌های کلیدی:**
   - آستانه بحرانی ۰/۰۳: اولین نقطه‌ای که مجموعه‌های جفتی ظاهر شدند
   - انتخاب بهینه ۰/۰۲۵: تعادل بین کمیت (۵۹ مجموعه) و کیفیت (۵ مجموعه جفتی)
   - عدم وجود مجموعه‌های بزرگتر از ۲: حتی با min_support=0.01، حداکثر اندازه ۲ باقی ماند
۳. **تحلیل آماری:**
   - ٪۹۱/۵ مجموعه‌های مکرر، تک‌عصره بودند (۵۴ از ۵۹)
   - تنها ٪۸/۵ مجموعه‌ها (۵ مجموعه) رابطه جفتی نشان دادند
   - محصولات برتر دارای پشتیبانی ۱۲-٪۲۱ بودند

بینش‌های کسب‌وکار:

۱. **تمرکز خریدها بر محصولات منفرد:** اکثر مشتریان محصولات را به صورت تکی خریداری می‌کنند
۲. **فرصت‌های محدود برای بسته‌های ترکیبی:** فقط ۵ جفت محصول الگوی تکرار قوی نشان دادند
۳. **محصول برتر:** Product_24852 با ٪۲۰/۹۴ پشتیبانی، پرفروش‌ترین محصول است





۵. تحلیل تسک چهارم: استخراج association rule ها

۱. معیارهای ارزیابی

۱. **پشتیبانی (support): میزان فراوانی هم‌رویدادی محصولات**

- **محدوده:** ۰/۰۲۵۱ تا ۰/۰۳۰۲ (۲/۵-۳٪)
- **تفسیر:** قوانین مبتنی بر خریدهای نسبتاً نادر هستند.

۲. **اعتماد** (Confidence)**: قدرت پیش‌بینی قانون**
- **محدوده:** ۰/۱۱۹۷ تا ۰/۳۰۵۹ (۱۲-۳۱٪)
- **تفسیر:** قدرت پیش‌بینی متوسط تا ضعیف

۳. **لیفت** (Lift)**: قدرت واقعی ارتباط نسبت به حالت تصادفی**
- **محدوده:** ۱/۰۴۷۳ تا ۱/۷۳۹۸
- **تفسیر:** تمامی ارتباطات مثبت اما نه بسیار قوی

قوانین برتر و تفسیر آنها:

۱. **قانون** Product_47209 → Product_13176
- **لیفت ۱/۷۴:** قوی‌ترین ارتباط کشف شده
- **اعتماد ۱۷/۱۳٪:** اگر مشتری Product_13176 بخرد، ۱۷٪ احتمال خرید Product_47209 وجود دارد
- **تفسیر عملی:** این دو محصول مکمل خوبی برای هم هستند

۲. **قانون:** Product_13176 → Product_47209 **(معکوس):**
- **لیفت:** یکسان ۱/۷۴ اما اعتماد بالاتر ۲۹/۵۹٪
- **نکته مهم:** رابطه نامتقارن - جهت دوم پیش‌بینی قوی‌تری دارد
- **تفسیر:** Product_47209 نشانه‌گر بهتری (نسب به رابطه معکوس) برای خرید Product_13176 است

۳. **قانون ۳:** Product_24852 → Product_47766
- **بالاترین اعتماد:** ۳۰/۵۹٪ در بین همه قوانین
- **لیفت ۱/۴۶:** ارتباط مثبت قابل توجه
- **تفسیر:** خریداران Product_47766 تمایل قوی به خرید محصول پرفروش (Product_24852) دارند

۶. پاسخ به دو سوال کلیدی:

۱. **نقش** Lift **در قوانین انجمنی**

- **معیاری است که قدرت واقعی ارتباط بین دو یا چند محصول را نسبت به حالت تصادفی اندازه‌گیری می‌کند. مقدار Lift بزرگ‌تر از ۱ نشان‌دهنده ارتباط مثبت و معنادار بین محصولات است، به‌طوری که احتمال خرید مشترک آن‌ها بیشتر از احتمال خرید مستقل آن‌هاست. در این پروژه، Lift به ما کمک کرد تا قوی‌ترین ارتباط‌ها (مانند ارتباط بین Product_13176 و Product_47209 با Lift حدود ۱/۷۴) را از میان قوانین کشف‌شده شناسایی کنیم و مطمئن شویم که الگوهای کشف‌شده تصادفی نیستند.**

۲. اهمیت الگوریتم Apriori برای فروشگاه‌های اینترنتی:

- **الگوریتم Apriori به فروشگاه‌های اینترنتی کمک می‌کند تا با تحلیل سبدهای خرید مشتریان، الگوهای پنهان خریدهای همزمان را کشف کنند. این الگوها اساس سیستم‌های پیشنهاددهنده، بازاریابی هدفمند، چیدمان محصولات و طراحی بسته‌های ترکیبی را تشکیل می‌دهند. در این پروژه، Apriori امکان شناسایی محصولات پرتکرار و ارتباطات بین آن‌ها را فراهم کرد و مبنایی داده‌محور برای تصمیم‌گیری‌هایی مانند پیشنهاد محصولات مکمل به مشتریان فراهم نمود.**

نتیجه گیری نهایی: این پروژه نشان داد که اگرچه داده‌های خرید دارای پراکندگی نسبتاً بالایی هستند، اما همچنان می‌توان الگوهای معناداری برای بهبود عملیات کسب‌وکار استخراج کرد. قوی‌ترین ارتباط بین Product_13176 و Product_47209 کشف شد که فرصتی عالی برای ایجاد بسته‌های ترکیبی ارائه می‌دهد.