دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده علوم کامپیوتر

پایان‌نامه کارشناسی‌ارشد

گزارش پروژه درس داده کاوی محاسباتی

پروژه ۱

نگارش

محمدصادق قلی زاده

استاد راهنما

دکتر مهدی قطعی


بسم الله الرحمن الرحيم

صفحه فرم ارزیابی و تصویب پایان نامه- فرم تأیید اعضاء کمیته دفاع

در این صفحه فرم دفاع یا تایید و تصویب پایان نامه موسوم به فرم کمیته دفاع- موجود در پرونده آموزشی- را قرار دهید.

نکات مهم:

- نگارش پایان نامه/رساله باید به زبان فارسی و بر اساس آخرین نسخه دستورالعمل و راهنمای تدوین پایان نامه های دانشگاه صنعتی امیرکبیر باشد.(دستورالعمل و راهنمای حاضر)
- رنگ جلد پایان نامه/رساله چاپی کارشناسی، کارشناسی ارشد و دکترا باید به ترتیب مشکی، طوسی و سفید رنگ باشد.
- چاپ و صحافی پایان نامه/رساله بصورت پشت و رو(دورو) بلامانع است و انجام آن توصیه می شود.

دانشگاه صنعتی امیرکبیر
(پلی‌تکنیک تهران)

به نام خدا

تعهدنامه اصالت اثر

تاریخ:

اینجانب **محمدصادق قلی زاده** متعهد می‌شوم که مطالب مندرج در این پایان‌نامه حاصل کار پژوهشی اینجانب تحت نظارت و راهنمایی اساتید دانشگاه صنعتی امیرکبیر بوده و به دستاوردهای دیگران که در این پژوهش از آنها استفاده شده است مطابق مقررات و روال متعارف ارجاع و در فهرست منابع و مآخذ ذکر گردیده است. این پایان‌نامه قبلاً برای احراز هیچ مدرک هم‌سطح یا بالاتر ارائه نگردیده است.

در صورت اثبات تخلف در هر زمان، مدرک تحصیلی صادر شده توسط دانشگاه از درجه اعتبار ساقط بوده و دانشگاه حق پیگیری قانونی خواهد داشت.

کلیه نتایج و حقوق حاصل از این پایان‌نامه متعلق به دانشگاه صنعتی امیرکبیر می‌باشد. هرگونه استفاده از نتایج علمی و عملی، واگذاری اطلاعات به دیگران یا چاپ و تکثیر، نسخه‌برداری، ترجمه و اقتباس از این پایان نامه بدون موافقت کتبی دانشگاه صنعتی امیرکبیر ممنوع است. نقل مطالب با ذکر مآخذ بلامانع است.

محمدصادق قلی زاده

امضا

نویسنده پایان نامه، در صورت تمایل میتواند برای سپاسگزاری پایان نامه خود را به شخص یا اشخاص و یا ارگان خاصی تقدیم نماید.

سپاس‌گزاری

نویسنده پایان‌نامه می‌تواند مراتب امتنان خود را نسبت به استاد راهنما و استاد مشاور و یا دیگر افرادی که طی انجام پایان‌نامه به نحوی او را یاری و یا با او همکاری نموده‌اند ابراز دارد.

محمدصادق قلی‌زاده

چکیده

با رشد سریع تجارت الکترونیک، حجم داده‌های تراکنشی که روزانه تولید می‌شود، چالش‌های محاسباتی قابل توجهی را برای روش‌های تحلیل ایستای سنتی ایجاد کرده است. این پروژه به بررسی کاربرد تکنیک‌های داده‌کاوی جریان داده [۱] بر روی مجموعه‌داده UCI Online Retail می‌پردازد و با شبیه‌سازی یک جریان پیوسته از تراکنش‌های فروش، قابلیت‌های پایش بلادرنگ را ارزیابی می‌کند. در این مطالعه، سه الگوریتم کاهش بُعد [۲] پیاده‌سازی و مقایسه شدند: نگاشت تصادفی گاوسی [۳]، تحلیل مؤلفه‌های اصلی افزایشی [۴]، و Frequent Directions (FD). عملکرد این روش‌ها از نظر خطای بازسازی، واریانس توضیح‌داده‌شده و کارایی محاسباتی مورد ارزیابی قرار گرفت. علاوه بر این، سودمندی این «اسکچ»های فشرده برای وظایف پایین‌دستی بررسی شد؛ به‌طور خاص، استخراج الگوهای پرتکرار با استفاده از الگوریتم FP-Growth. نتایج نشان می‌دهد که اگرچه IPCA بالاترین دقت بازسازی را ارائه می‌دهد، روش Fre-quent Directions تعادل مناسبی میان پایداری و کمینه‌سازی خطا در محیط‌های جریانی برقرار می‌کند. همچنین، حساسیت سامانه با تزریق دسته‌هایی از ناهنجاری‌های مصنوعی مورد آزمون قرار گرفت؛ هر دو روش IPCA و Frequent Directions توانستند این انحراف ساختاری را از طریق افزایش‌های چشمگیر در خطای بازسازی با موفقیت شناسایی کنند. این یافته‌ها نشان می‌دهد که اسکچ‌سازی ماتریسی رویکردی کارآمد از نظر حافظه برای حفظ سودمندی جریان‌های داده با بُعد بالا بوده و در عین حال امکان تشخیص ناهنجاری بلادرنگ و کشف الگوها را فراهم می‌سازد. GitHub

واژه‌های کلیدی:

داده‌کاوی جریان داده، طراحی ماتریس، کاهش ابعاد (مسیرهای مکرر)، تشخیص ناهنجاری، داده‌کاوی قوانین وابستگی

---

[۱] (Data Stream Mining)
[۲] (Sketching)
[۳] (Random Gaussian Projection – RGP)
[۴] (Incremental PCA – IPCA)

فهرست مطالب

عنوان صفحه

فهرست اشکال

شکل صفحه

فهرست جداول

جدول صفحه

فهرست نمادها

نماد	مفهوم
$\mathbb{R}^n$	فضای اقلیدسی با بعد n
$\mathbb{S}^n$	کره یکه n بعدی
M^m	خمینه m-بعدی M
$\mathfrak{X}(M)$	جبر میدان‌های برداری هموار روی M
$\mathfrak{X}^1(M)$	مجموعه میدان‌های برداری هموار یکه روی (M,g)
$\Omega^p(M)$	مجموعه p-فرمی‌های روی خمینه M
Q	اپراتور ریچی
$\mathcal{R}$	تانسور انحنای ریمان
ric	تانسور ریچی
L	مشتق لی
Φ	۲-فرم اساسی خمینه تماسی
∇	التصاق لوی-چویتای
Δ	لاپلاسین ناهموار
∇^*	عملگر خودالحاق صوری القا شده از التصاق لوی-چویتای
g_s	متر ساساکی
∇	التصاق لوی-چویتای وابسته به متر ساساکی
Δ	عملگر لاپلاس-بلترامی روی p-فرم‌ها

فصل ۱

مقدمه

۱-۱ مقدمه

۱-۱-۱ پیش‌زمینه

اقتصاد دیجیتال منجر به انفجار تولید داده شده است، به‌ویژه در بخش خرده‌فروشی. هر تراکنش نه‌تنها یک فروش را ثبت می‌کند، بلکه بازتابی از یک تعامل پیچیده میان مشتریان و محصولات است. تحلیل سنتی داده‌ها معمولاً در حالت دسته‌ای [۱] انجام می‌شود؛ یعنی تحلیل تنها پس از ذخیره‌سازی کامل داده‌ها صورت می‌گیرد. اما در تجارت الکترونیک مدرن، داده‌ها به‌صورت جریانی پیوسته و با سرعت بالا وارد سیستم می‌شوند. تأخیر در پردازش این داده‌ها به معنای از دست دادن فرصت‌های پیشنهاددهی بلادرنگ یا تأخیر در شناسایی خطاهای سیستمی و تقلب است.

۱-۱-۲ بیان مسئله

تحلیل جریان‌های تراکنشی با دو چالش محاسباتی اصلی مواجه است.

بُعد بالای داده‌ها: یک فروشگاه خرده‌فروشی ممکن است هزاران قلم کالای منحصربه‌فرد داشته باشد. ماتریسی که نشان دهد «چه کسی چه چیزی خریده است» ابعادی بسیار بزرگ (با هزاران ستون) دارد و معمولاً تنک [۲] است؛ موضوعی که پردازش بلادرنگ آن را از نظر محاسباتی پرهزینه می‌کند.

---

[۱] (Batch Mode)

[۲] (Sparse)

محدودیت حافظه: در یک سناریوی واقعیِ جریان داده، ذخیره کل تاریخچه تراکنش‌ها در حافظه اصلی [۳] امکان‌پذیر نیست. الگوریتم‌های استانداردی که برای عملکرد صحیح به کل داده‌ها نیاز دارند (مانند PCA کلاسیک)، عملاً با انباشت داده‌ها دچار توقف یا شکست می‌شوند.

۱-۱-۳ اهداف پروژه

این گزارش با پیاده‌سازی تکنیک‌های اسکچ‌سازی ماتریسی [۴] به این چالش‌ها پاسخ می‌دهد. اسکچ‌سازی یک ماتریس عظیم را به نسخه‌ای کوچک‌تر و فشرده تبدیل می‌کند که مهم‌ترین خواص ریاضی آن ،مانند واریانس و هم‌بستگی، را حفظ می‌کند. به‌طور مشخص، این پروژه سه هدف اصلی را دنبال می‌کند.

تحلیل مقایسه‌ای: ارزیابی سه الگوریتم متمایز ،نگاشت تصادفی گاوسی [۵]، تحلیل مؤلفه‌های اصلی افزایشی [۶] و Frequent Directions، از نظر توانایی فشرده‌سازی و بازسازی دقیق داده‌های فروش.

حفظ الگوها: بررسی این‌که آیا این «اسکچ»های فشرده همچنان بینش‌های تجاری مفیدی را حفظ می‌کنند یا خیر؛ به‌ویژه با اعمال استخراج الگوهای پرتکرار [۷] بر داده‌های فشرده برای بازیابی روابط فروش.

تشخیص ناهنجاری: آزمون پایداری این الگوریتم‌ها از طریق تزریق نویز مصنوعی (ناهنجاری) به جریان داده و سنجش این‌که آیا خطای بازسازی می‌تواند به‌عنوان یک سیگنال هشدار قابل اعتماد عمل کند یا خیر.

۱-۱-۴ نمای کلی روش‌شناسی

این مطالعه از مجموعه‌داده UCI Online Retail استفاده می‌کند؛ یک معیار استاندارد برای تحلیل تراکنش‌ها که شامل داده‌های یک خرده‌فروشی آنلاین غیرحضوری مستقر در بریتانیاست. داده‌های خام پیش‌پردازش می‌شوند تا تراکنش‌های لغوشده حذف شده و سپس به یک ماتریس دودویی «کالا تراکنش» تبدیل شوند. این ماتریس به‌صورت یک جریان داده زمان‌مرتب شبیه‌سازی شده و به‌طور ترتیبی به الگوریتم‌های اسکچ‌سازی ورودی داده می‌شود. عملکرد سیستم با استفاده از معیارهای کمی مانند هنجار فروبنیوس [۸] برای سنجش انرژی داده و خطای بازسازی [۹] برای ارزیابی کیفیت فشرده‌سازی

۳ (RAM)
۴ (Matrix Sketching)
۵ (Random Gaussian Projection)
۶ (Incremental PCA)
۷ (FP-Growth)
۸ (Frobenius Norm)
۹ (Reconstruction Error)

اندازه‌گیری می‌شود. در نهایت، با تزریق نویز تصادفی به‌عنوان شبیه‌سازی نفوذ یا اختلال سیستمی، قابلیت‌های تشخیص ناهنجاری سامانه پیشنهادی اعتبارسنجی می‌گردد.

فصل ۲

مروری بر ادبیات

۲-۱ تعاریف مفاهیم پایه

پایش ماتریس [۱] فرایندی است برای نظارت مداوم بر ویژگی‌های آماری و ساختاری ماتریس‌های بزرگ‌داده، به‌گونه‌ای که با ورود داده‌های جدید بتوان تغییرات در ساختار، نُرم یا هم‌بستگی را بدون محاسبه مجدد کل ماتریس شناسایی کرد [۱].

اسکچینگ ماتریس [۲] روشی است برای فشرده‌سازی داده‌های ماتریسی با ابعاد بالا، از طریق نگاشت آن‌ها به فضای کم‌بُعدتر به‌گونه‌ای که اطلاعات آماری اصلی حفظ شود و محاسبات بعدی با سرعت و حافظه کمتر انجام گیرد [۱، ۲].

تقریب کم‌رتبه [۳] به نمایش یک ماتریس در قالب تعداد محدودی از مؤلفه‌های اصلی گفته می‌شود که ساختار غالب داده را حفظ کرده و مؤلفه‌های کم‌اهمیت یا نویزی را حذف می‌کند [۲].

الگوریتم Frequent Directions یک الگوریتم اسکچینگ مؤثر است که با نگهداری خلاصه‌ای از داده‌ها در یک ماتریس کوچک‌تر، تقریبی دقیق از ضرب $A^T A$ را فراهم می‌سازد و برای پایش تغییرات ساختاری داده‌ها بسیار کارآمد است [۳].

طرح‌ریزی تصادفی گوسی [۴] تکنیکی برای کاهش بُعد داده‌ها است که در آن ماتریس اصلی در یک ماتریس تصادفی با توزیع گوسی ضرب می‌شود به‌گونه‌ای که فاصله‌های اقلیدسی بین داده‌ها تقریباً

---

[۱] (Matrix Monitoring)
[۲] (Matrix Sketching)
[۳] (Low-Rank Approximation)
[۴] (Gaussian Random Projection)

حفظ شوند [۴، ۵].

تحلیل مؤلفه‌های اصلی تدریجی [۵] نسخه‌ای از الگوریتم PCA است که به‌صورت آنلاین عمل کرده و با ورود تدریجی داده‌های جدید، مؤلفه‌های اصلی را بدون نیاز به بازآموزی کامل مدل به‌روزرسانی می‌کند [۶].

نُرم فروبنیوس [۶] معیاری از بزرگی یک ماتریس است که برابر با ریشه دوم مجموع مربعات تمامی عناصر آن بوده و برای سنجش انرژی کلی یا تغییرات ساختار ماتریس به‌کار می‌رود [۷].

نسبت واریانس توضیح‌داده‌شده [۷] سهم نسبی هر مؤلفه از واریانس کل داده را بازنمایی می‌کند و معیاری برای سنجش کیفیت فشرده‌سازی در روش‌هایی مانند PCA به‌شمار می‌رود [۸].

خطای بازسازی [۸] میزان انحراف میان داده اصلی و داده بازسازی‌شده پس از کاهش بُعد است که نشان‌دهنده دقت یا میزان از دست رفتن اطلاعات در فرایند فشرده‌سازی می‌باشد [۹].

کشف الگوهای پرتکرار [۹] فرایندی در داده‌کاوی است که هدف آن شناسایی مجموعه‌های آیتمی است که به‌صورت مکرر در تراکنش‌ها یا رکوردهای داده ظاهر می‌شوند [۱۰].

الگوریتم آپریوری [۱۰] یکی از الگوریتم‌های کلاسیک در کشف الگوهای پرتکرار است که بر پایه اصل زیرمجموعه پرتکرار عمل کرده و به‌صورت افزایشی مجموعه‌های پرتکرار را شناسایی می‌کند [۱۱].

الگوریتم FP Growth روشی بهینه‌تر نسبت به Apriori است که با استفاده از یک ساختار درختی فشرده به‌نام FP tree مجموعه‌های پرتکرار را بدون نیاز به تولید صریح تمامی زیرمجموعه‌ها استخراج می‌کند [۱۰].

داده‌های جریانی [۱۱] داده‌هایی هستند که به‌صورت پیوسته و زمان‌مند تولید یا دریافت می‌شوند و نیازمند تحلیل بلادرنگ و به‌روزرسانی مستمر مدل‌ها هستند [۱۲].

تشخیص ناهنجاری [۱۲] به شناسایی نمونه‌ها، الگوها یا رفتارهایی گفته می‌شود که با الگوی معمول داده‌ها تفاوت معنادار دارند و می‌توانند نشانه خطا، حمله یا پدیده‌ای جدید باشند [۱۳].

ماتریس طرح‌ریزی تصادفی [۱۳] ماتریسی با مقادیر تصادفی (معمولاً با توزیع گوسی یا یکنواخت)

---

[۵] (Incremental PCA)
[۶] (Frobenius Norm)
[۷] (Explained Variance Ratio)
[۸] (Reconstruction Error)
[۹] (Frequent Pattern Mining)
[۱۰] (Apriori Algorithm)
[۱۱] (Online / Streaming Data)
[۱۲] (Anomaly Detection)
[۱۳] (Random Projection Matrix)

است که برای نگاشت داده‌ها به فضای کم‌بُعدتر به‌کار می‌رود و ضمن کاهش ابعاد، ساختار فاصله‌ها را تقریباً حفظ می‌کند [۵].

تحلیل مؤلفه‌های اصلی [۱۴] روشی آماری برای یافتن جهت‌هایی در فضای داده است که بیشترین واریانس را توضیح می‌دهند و برای کاهش بُعد، حذف هم‌بستگی و فشرده‌سازی داده‌ها استفاده می‌شود [۸].

تجزیه مقادیر منفرد [۱۵] روشی ماتریسی است که هر ماتریس را به ضرب سه ماتریس $U\Sigma V^T$ تجزیه می‌کند و پایه بسیاری از الگوریتم‌های کاهش بُعد و تحلیل ساختار داده است [۷].

ماتریس کوواریانس [۱۶] ماتریسی است که میزان هم‌تغییری میان ویژگی‌های مختلف داده را نشان می‌دهد و مبنای تحلیل‌های آماری مانند PCA و تشخیص الگو محسوب می‌شود [۸].

کاهش بُعد [۱۷] فرایند تبدیل داده‌های پُربُعد به نمایشی کم‌بُعدتر است به‌گونه‌ای که اطلاعات کلیدی حفظ شده و هزینه محاسباتی و نویز کاهش یابد [۹].

[۱۴] (Principal Component Analysis)
[۱۵] (Singular Value Decomposition)
[۱۶] (Covariance Matrix)
[۱۷] (Dimensionality Reduction)

فصل ۳

روش شناسی و نتایج

۳-۱ روش‌شناسی

این مطالعه از یک خط لوله چندمرحله‌ای استفاده می‌کند تا داده‌های خام تراکنشی را با بهره‌گیری از الگوریتم‌های جریانی به بینش‌های عملی تبدیل کند. روش‌شناسی به سه فاز اصلی تقسیم می‌شود: آماده‌سازی داده‌ها، اسکچ‌سازی ماتریسی (کاهش بُعد) و ارزیابی.

۳-۱-۱ آماده‌سازی داده‌ها و شبیه‌سازی جریان

داده‌های خام از مجموعه‌داده UCI Online Retail ابتدا تحت فرایندهای پاک‌سازی قرار گرفتند تا مقادیر تهی (Null)، تراکنش‌های لغوشده (مشخص‌شده با حرف C در شماره فاکتور) و مقادیر منفیِ تعداد کالا حذف شوند.

برای امکان‌پذیر شدن استخراج قوانین انجمنی و اسکچ‌سازی ماتریسی، داده‌ها به یک ماتریس دودویی کالا□تراکنش (A) تبدیل شدند؛ به‌طوری که سطرها نمایانگر فاکتورهای یکتا (n) و ستون‌ها نمایانگر محصولات یکتا (d) هستند:

$$A_{ij} = \begin{cases} ۱ & \text{اگر کالای } j \text{ در فاکتور } i \text{ وجود داشته باشد} \\ ۰ & \text{در غیر این صورت} \end{cases}$$

برای شبیه‌سازی یک محیط بلادرنگ، این ماتریس به‌صورت یک بلوک ایستا پردازش نشد. در عوض،

داده‌ها به‌صورت زمانی مرتب شده و به دسته‌های متوالی

$$(B_1, B_2, \ldots, B_t)$$

برش داده شدند. این کار امکان تقلید از ورود تدریجی داده‌ها مشابه جریان‌های واقعی تراکنش را فراهم کرد.

۳-۱-۲ الگوریتم‌های اسکچ‌سازی ماتریسی

هسته این پروژه، اسکچ‌سازی جریان داده است؛ یعنی نگه‌داری یک تقریب کم‌رتبه از ماتریس که با مصرف حافظه‌ای به‌مراتب کمتر از داده اصلی، خواص آماری آن را حفظ می‌کند. سه رویکرد متمایز مقایسه شدند.

نگاشت تصادفی گاوسی [۱] یک روش مستقل از داده است که ریشه در لم [۲] دارد. این لم بیان می‌کند که می‌توان مجموعه‌ای از نقاط در فضای بُعد بالا را به فضایی با بُعد کمتر نگاشت، به‌گونه‌ای که فاصله‌های اقلیدسی آن‌ها تقریباً حفظ شود. در این روش، یک ماتریس تصادفی ثابت R با درایه‌هایی از توزیع گاوسی تولید شد و هر دسته ورودی B روی آن فرافکنی گردید:

$$B_{\text{sketch}} = B \times R$$

این روش نیازی به آموزش یا به‌روزرسانی وابسته به داده ندارد و از نظر محاسباتی بسیار سریع است.

تحلیل مؤلفه‌های اصلی افزایشی [۳] نسخه‌ای جریانی از PCA استاندارد است. در حالی که PCA کلاسیک برای محاسبه ماتریس کوواریانس و استخراج بردارهای ویژه به کل داده نیاز دارد، IPCA این فرایند را از طریق به‌روزرسانی تدریجی ممکن می‌سازد. در این روش، یک برآورد جاری از مؤلفه‌های اصلی و مقادیر منفرد نگه داشته می‌شود و با ورود هر دسته جدید، این مؤلفه‌ها به‌گونه‌ای به‌روزرسانی می‌شوند که واریانس توضیح‌داده‌شده تاریخچه تجمیع‌شده بیشینه گردد. این روش از نظر کمینه‌سازی خطای میانگین مربعات بازسازی، بهینه است.

جهت‌های پرتکرار [۴] یک الگوریتم قطعی است که به‌طور خاص برای اسکچ‌سازی ماتریسی طراحی شده و اغلب به‌عنوان «نسخه جریانی SVD» توصیف می‌شود. این الگوریتم یک بافر کوچک با اندازه

[۱] (Random Gaussian Projection – RGP)
[۲] Johnson–Lindenstrauss
[۳] (Incremental PCA – IPCA)
[۴] (Frequent Directions – FD)

$2k$ نگه می‌دارد. با پر شدن بافر از سطرهای جدید، SVD بافر محاسبه می‌شود، مقادیر منفرد کوچک می‌شوند (کاهش وزن جهت‌های کم‌اهمیت) و تنها k جهت برتر حفظ می‌گردد. این روش برخلاف RGP وابسته به داده و دقیق است و برخلاف IPCA، کران‌های نظری قوی‌تری برای خطا نسبت به تقریب کم‌رتبه بهینه ارائه می‌دهد.

۳-۱-۳ استخراج الگوهای پرتکرار

برای آزمون سودمندی اسکچ‌های فشرده، از الگوریتم FP-Growth استفاده شد. برخلاف الگوریتم Apriori که از راهبرد «تولید و آزمون» بهره می‌برد، FP-Growth یک ساختار درختی فشرده به‌نام FP-tree می‌سازد که امکان استخراج سریع‌تر مجموعه‌اقلام پرتکرار را بدون تولید مجموعه‌های کاندید فراهم می‌کند.

این الگوریتم هم روی داده تنکِ اصلی و هم روی داده بازسازی‌شده از اسکچ‌ها اعمال شد تا Recall اندازه‌گیری شود؛ یعنی درصد الگوهای فروش واقعی که پس از فشرده‌سازی با موفقیت بازیابی شده‌اند.

۳-۱-۴ معیارهای ارزیابی

برای کمی‌سازی عملکرد الگوریتم‌ها، سه معیار اصلی برای هر دسته پایش شد. هنجار فروبنیوس

$$\|A\|_F$$

برای ردیابی بزرگی یا انرژی جریان داده (حجم کل فروش) به‌کار رفت. خطای بازسازی

$$\|A - \hat{A}\|_F$$

فاصله اقلیدسی بین دسته اصلی و بازسازی‌شده از اسکچ را اندازه‌گیری می‌کند که مقادیر کمتر نشان‌دهنده دقت بالاتر است. در نهایت، نسبت واریانس توضیح‌داده‌شده به‌عنوان سهمی از اطلاعات داده که توسط اسکچ حفظ می‌شود، برای ارزیابی کیفیت فشرده‌سازی مورد استفاده قرار گرفت.

فصل ۴

نتایج

۴-۱ نتایج و بحث

۴-۱-۱ بارگذاری داده و پیش‌پردازش

تحلیل با ورود مجموعه‌داده خام UCI Online Retail آغاز شد که شامل ۵۴۱,۹۰۹ رکورد تراکنشی است. برای تضمین کیفیت جریان داده، یک فرایند پاک‌سازی دقیق اعمال شد.

رکوردهای ناقص: سطرهایی که شناسه‌های حیاتی (مانند CustomerID) را نداشتند حذف شدند، زیرا تراکنش‌های ناشناس را نمی‌توان به‌طور قابل اعتماد برای استخراج الگوها ردیابی کرد.

لغوها و خطاها: تراکنش‌هایی که دارای پیشوند C (نشان‌دهنده فاکتورهای برگشتی) بودند و همچنین رکوردهایی با مقادیر منفی یا صفر برای تعداد کالا، فیلتر شدند تا از ایجاد انحراف در تحلیل حجم فروش جلوگیری شود.

تأثیر کمی: این پیش‌پردازش اندازه مجموعه‌داده را از ۵۴۱,۹۰۹ به ۳۹۷,۸۸۴ رکورد معتبر کاهش داد. اگرچه این به معنای کاهش حدود ٪۲۶ از داده‌هاست، اما داده‌های باقی‌مانده پایه‌ای تمیزتر و قابل‌اعتمادتر برای پردازش الگوریتمی فراهم می‌کنند و از نویزی که می‌توانست به‌طور مصنوعی واریانس را افزایش دهد، عاری هستند.

۴-۱-۲ ساخت ماتریس کالا تراکنش

پس از مرحله پاک‌سازی، داده‌ها با موفقیت از فهرست تراکنش‌ها در قالب «طولی» (Long-Format) به یک ماتریس ساخت‌یافته کالا□تراکنش تبدیل شدند.

ساختار: در این ماتریس دودویی، هر سطر متناظر با یک InvoiceNo یکتا و هر ستون نمایانگر یک Description (محصول) یکتا است.

کدگذاری: مقادیر به‌صورت شاخص‌های دودویی (۰ یا ۱) کدگذاری شدند که نشان‌دهنده وجود یا عدم وجود یک کالا در سبد خرید مشخص هستند، مستقل از تعداد خریداری‌شده.

تنکی: (Sparsity) همان‌طور که در داده‌های خرده‌فروشی انتظار می‌رود، ماتریس حاصل به‌شدت تنک است بیشتر مشتریان تنها بخش بسیار کوچکی از هزاران کالای موجود را خریداری می‌کنند. این تنکی یکی از ویژگی‌های کلیدی داده است که ضرورت به‌کارگیری تکنیک‌های کاهش بُعد (Sketching) در مراحل بعدی را توجیه می‌کند.

۴-۱-۳ شبیه‌سازی جریان داده

برای نزدیک شدن به یک محیط تحلیل بلادرنگ، ماتریس ایستای کالا،تراکنش (شامل ۱۸,۵۳۲ فاکتور یکتا) به‌صورت زمانی بازآرایی شد و به چهار دسته متوالی تقسیم گردید.

ساختار دسته‌ها: هر دسته شامل ۴,۶۳۳ تراکنش بود.

توزیع زمانی:

دسته‌های ۱ و ۲ (دسامبر ۲۰۱۰ ، ژوئیه ۲۰۱۱): نمایانگر دوره عملیاتی عادی خرده‌فروش هستند.

دسته‌های ۳ و ۴ (ژوئیه ۲۰۱۱ ، دسامبر ۲۰۱۱): فصل پیش از تعطیلات و خودِ تعطیلات را پوشش می‌دهند؛ دوره‌ای که به‌طور تاریخی با نوسان و حجم فروش بالاتر همراه است.

این تقسیم‌بندی با موفقیت ورود تدریجی داده‌ها را شبیه‌سازی کرد و به ما امکان داد بررسی کنیم الگوریتم‌ها چگونه در طول یک سال و در مواجهه با تغییرات رفتاری بازار سازگار می‌شوند.

۴-۱-۴ پایش ماتریسی و رفتار الگوریتم‌ها

سه الگوریتم اسکچ‌سازی ،نگاشت تصادفی گاوسی [۱]، تحلیل مؤلفه‌های اصلی افزایشی [۲] و Frequent Directions (FD)، بر روی هر دسته اعمال شدند. لاگ‌های پایش، روندهای متمایزی را در ساختار داده‌ها طی زمان نشان می‌دهند.

تغییرپذیری و حجم (واریانس RGP):

واریانس RGP به‌عنوان نماینده‌ای از «انرژی» یا پراکندگی کلی مجموعه‌داده در نظر گرفته می‌شود.

مشاهده: واریانس در نیمه نخست سال پایدار بود (حدود ۱٫۹۴ در دسته ۱ تا ۱٫۸۹ در دسته ۲)، اما در نیمه دوم افزایش محسوسی نشان داد و به ۲٫۱۹ (دسته ۳) و ۲٫۲۶ (دسته ۴) رسید.

تفسیر: این روند با ماهیت فصلی خرده‌فروشی هم‌خوان است. با نزدیک شدن به فصل تعطیلات، تنوع و اندازه سبدهای خرید مشتریان افزایش یافته و به واریانس بالاتر در فضای فرافکنی‌شده منجر شده است.

اشباع مدل (واریانس توضیح‌داده‌شده IPCA):

این معیار نشان می‌دهد تعداد مؤلفه‌های ثابت $(k = ۱۰)$ تا چه حد اطلاعات داده را پوشش می‌دهد.

مشاهده: نسبت واریانس توضیح‌داده‌شده به‌تدریج از ۰٫۰۹۶ (دسته ۱) به ۰٫۰۸۵ (دسته ۴) کاهش یافت.

تفسیر: کاهش این نسبت بیانگر پیچیده‌تر شدن رفتار خرید مشتریان در طول زمان است. با افزایش تنوع محصولات (احتمالاً به‌دلیل تغییرات فصلی موجودی)، یک مدل با ۱۰ مؤلفه ثابت در حفظ همان سهم اطلاعات با دشواری مواجه می‌شود؛ موضوعی که نشان می‌دهد تغییر پویا در اندازه مدل می‌تواند در نسخه‌های آینده مفید باشد.

روندهای غالب (بزرگ‌ترین مقدار منفرد در FD):

بزرگ‌ترین مقدار منفرد در اسکچ Frequent Directions نشان‌دهنده قدرت غالب‌ترین الگوی خرید است.

مشاهده: در نیمه دوم جریان، جهش تندی در قدرت سیگنال رخ داد؛ این مقدار تقریباً از ۹٫۲ (دسته ۱) به ۱۸٫۵ (دسته ۴) دو برابر شد.

تفسیر: این نتیجه نشان می‌دهد که با وجود پیچیده‌تر شدن کلی داده‌ها (طبق IPCA)، اقلام پرفروش اصلی به‌طور قابل توجهی غالب‌تر شدند. این الگو با روندهای تعطیلات سازگار است؛ جایی که چند کالای

[۱] (Random Gaussian Projection – RGP)
[۲] (Incremental PCA – IPCA)

«پرفروش» سهم بزرگی از درآمد را ایجاد می‌کنند.

۴-۱-۵ معیارهای تحلیلی و کیفیت فشرده‌سازی

برای ارزیابی دقیق دقت الگوریتم‌های اسکچ‌سازی، هنجار فروبنیوس (بزرگی کل داده) و خطای بازسازی برای هر دسته محاسبه شد. نتایج، یک سلسله‌مراتب روشن از عملکرد میان سه روش را نشان می‌دهد.

تحول داده (هنجار فروبنیوس اصلی):

هنجار فروبنیوس اصلی

$$\|A\|_F$$

روند فصلی مشاهده‌شده در تحلیل واریانس را تأیید می‌کند.

دسته‌های ۲،۱: هنجار در بازه حدود ۳۰۱،۲۹۵ پایدار بود که نشان‌دهنده حجم فروش یکنواخت است.

دسته‌های ۴،۳: هنجار به‌طور محسوسی به ۳۱۸٫۸ و ۳۲۹٫۱ افزایش یافت. این افزایش در «انرژی ماتریس» جهش حجم تراکنش‌ها و اندازه سبد خرید را در نیمه دوم سال کمی‌سازی می‌کند.

مقایسه دقت (خطای بازسازی):

نزدیکی ماتریس‌های بازسازی‌شده

$$\hat{A}$$

به داده اصلی

$$A$$

مقایسه شد. خطای کمتر نشان‌دهنده فشرده‌سازی بهتر است.

برنده: تحلیل مؤلفه‌های اصلی افزایشی [۳]:

IPCA در همه دسته‌ها کمترین خطای بازسازی را به‌طور پیوسته به‌دست آورد (برای مثال ۲۸۶٫۰۹ در دسته ۱). این نتیجه قابل انتظار است، زیرا PCA از نظر ریاضی برای کمینه‌سازی هنجار فروبنیوس

[۳] (Incremental PCA – IPCA)

اختلاف در رتبه ثابت بهینه است و در این آزمایش به‌عنوان استاندارد طلایی عمل می‌کند.

نفر دوم: Frequent Directions (FD):

FD رقابت‌پذیری بالایی نشان داد و خطاهای آن تنها اندکی بالاتر از IPCA بود (برای مثال ۲۹۲٫۹۴ در دسته ۱). نکته کلیدی این است که با رشد حجم داده در دسته ۴، FD عملکرد نسبی خود را حفظ کرد (۳۱۷٫۷۴) در مقایسه با IPCA (۳۱۱٫۹۲). این موضوع نشان می‌دهد که FD گزینه‌ای مقاوم برای سناریوهای جریانی است که در آن‌ها سربار محاسباتی PCA ممکن است زیاد باشد.

ضعیف‌ترین: نگاشت تصادفی گاوسی [۴]:

RGP بیشترین خطا را نشان داد؛ مقادیر آن تقریباً با هنجار ماتریس اصلی یکسان بودند (برای مثال ۳۲۸٫۷۲ در برابر ۳۲۹٫۱۴ در دسته ۴). این نتیجه نشان می‌دهد که برای این مجموعه‌داده تنک خاص، فرافکنی تصادفیِ بدون یادگیری ویژگی‌ها نتوانست ساختار معنادار تراکنش‌ها را ثبت کند و به بازسازی «محو» انجامید.

حفظ اطلاعات (واریانس توضیح‌داده‌شده):

IPCA نسبت واریانس توضیح‌داده‌شده‌ای در حدود ۸٫۴٪ تا ۸٫۵٪ را حفظ کرد. اگرچه این مقدار در نگاه اول کم به نظر می‌رسد، اما برای داده‌های خرده‌فروشی تنک و با بُعد بالا کاملاً متداول است؛ جایی که واریانس میان هزاران محصول متمایز توزیع شده است. پایداری این معیار (نوسان کمتر از ۰٫۲٪) نشان می‌دهد که IPCA توانسته مدل درونی خود را با جریان داده در حال تغییر سازگار کند، بدون آنکه اطلاعات معنادار به‌طور قابل توجهی از دست برود.

۴-۱-۶ تجسم‌سازی و تحلیل عملکرد

برای شناسایی روندها و بررسی پایداری در طول جریان داده، معیارهای عددی به‌صورت بصری نمایش داده شدند. شکل ۱ (در پایین) عملکرد مقایسه‌ای سه الگوریتم را در چهار دسته متوالی نشان می‌دهد.

شکل ۱: تحلیل مقایسه‌ای الگوریتم‌های اسکچ‌سازی ماتریسی (ارجاع به تصویر بارگذاری‌شده شامل سه زیرنمودار)

الف) خطای بازسازی (دقت) نمودار سمت چپ در شکل ۱ هنجار فروبنیوس خطای بازسازی را نمایش می‌دهد.

[۴] (Random Gaussian Projection – RGP)

Incremental PCA (خط آبی): به‌طور پیوسته کمترین خطا را به‌دست می‌آورد و جایگاه خود را به‌عنوان روش بهینه برای کمینه‌سازی واریانس تأیید می‌کند. این منحنی به‌عنوان «کران پایین» یا بهترین فشرده‌سازی ممکن برای این مجموعه‌داده عمل می‌کند.

Frequent Directions (خط سبز): با فاصله‌ای اندک و پایدار، منحنی IPCA را دنبال می‌کند. هرچند خطای آن همواره کمی بیشتر است، اما این شکاف کوچک و ثابت می‌ماند. این شواهد بصری ادعای نظری را تأیید می‌کند که Frequent Directions در محیط‌های جریانی یک تقریب باکیفیت از SVD ارائه می‌دهد.

Random Projection (خط قرمز خط‌چین): بیشترین خطا را نشان می‌دهد و در دسته‌های ۳ و ۴ به‌طور محسوسی از دو روش دیگر فاصله می‌گیرد. این واگرایی نشان می‌دهد که با افزایش پیچیدگی داده‌ها در نیمه دوم سال، ماهیت مستقل از داده نگاشت تصادفی در حفظ جزئیات ساختاری ظریف تراکنش‌ها با مشکل مواجه شده است.

ب) نسبت واریانس توضیح‌داده‌شده (حفظ اطلاعات) نمودار میانی کارایی اسکچ‌ها را در ثبت اطلاعات نشان می‌دهد.

Incremental PCA: نسبت واریانس توضیح‌داده‌شده پایداری (حدود ۰٫۰۹) را در سراسر جریان حفظ می‌کند. تخت‌بودن خط آبی نشان می‌دهد الگوریتم بدون افت عملکرد، خود را با توزیع داده در حال تغییر سازگار کرده است.

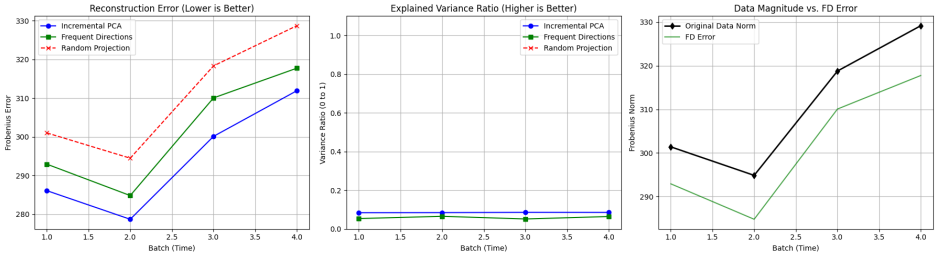
Frequent Directions: اندکی واریانس کمتر (حدود ۰٫۰۶) را ثبت می‌کند، اما همچنان پایدار باقی می‌ماند.

یادداشت درباره Random Projection: همان‌طور که در لاگ‌های عددی دیده می‌شود، معیار واریانس RGP در این مقیاس ۰ تا ۱ قابل مقایسه نیست (و اغلب به‌دلیل ماهیت نگاشت که فاصله‌ها را حفظ می‌کند نه واریانس، از ٪۱۰۰ فراتر می‌رود). ازاین‌رو در این نما حذف یا خارج از مقیاس نمایش داده می‌شود که بار دیگر تفاوت بنیاد ریاضی آن را برجسته می‌کند.

ج) بزرگی داده در برابر روند خطا (پایداری) نمودار سمت راست بررسی می‌کند که آیا افزایش خطا در دسته‌های ۳ و ۴ ناشی از شکست الگوریتم‌ها بوده یا ویژگی‌های خودِ داده‌ها.

هم‌بستگی: خط سیاه (هنجار داده اصلی) و خط سبز (خطای FD) تقریباً همگام حرکت می‌کنند؛ هر دو در دسته ۲ افت کرده و در دسته ۴ به‌طور تند افزایش می‌یابند.

نتیجه‌گیری: افزایش خطا متناسب با افزایش حجم فروش (بزرگی داده) است. این امر نشان می‌دهد الگوریتم‌ها پایدار باقی مانده‌اند؛ تنها «سختی مسئله» افزایش یافته است، زیرا طول بردارهای ورودی در فصل شلوغ تعطیلات بیشتر شده است.



شکل ۴-۱: نتایج بخش ۶

۴-۱-۷ استخراج الگوهای پرتکرار و تحلیل کاربری

برای ارزیابی این‌که آیا اسکچ‌های فشرده همچنان بینش‌های تجاری معنادار را حفظ کرده‌اند یا خیر، از استخراج الگوهای پرتکرار [۵] برای به‌دست‌آوردن قوانین انجمنی استفاده شد (برای مثال: «اگر کاربری نان بخرد، شیر نیز می‌خرد»).

الف) چالش‌های محاسباتی (محدودیت حافظه) در این مرحله با یک مانع محاسباتی جدی مواجه شدیم. اگرچه مجموعه‌داده اصلی به‌شدت تنک است (اکثراً شامل صفرها)، ماتریس بازسازی‌شده از اسکچ‌ها (به‌ویژه IPCA) چگال است و با مقادیر اعشاری غیرصفر پر شده است.

مسئله: تلاش برای پردازش این ماتریس چگال با الگوریتم استاندارد Apriori باعث از کار افتادن نشست Google Colab به‌دلیل اتمام حافظه [۶] شد. ردپای حافظه لازم برای تولید مجموعه‌های کاندید در ماتریس چگال از منابع در دسترس فراتر رفت.

ب) بهینه‌سازی و راه‌حل برای رفع این مشکل، یک رویکرد بهینه‌شده از نظر حافظه پیاده‌سازی شد:

انتخاب الگوریتم: الگوریتم Apriori با FP-Growth جایگزین شد. FP-Growth با ساخت یک ساختار درختی فشرده (FP-Tree) از تولید مجموعه‌های کاندید اجتناب می‌کند و به‌طور قابل توجهی سربار حافظه را کاهش می‌دهد.

[۵] (Frequent Pattern Mining)
[۶] (Out of Memory)

بازگشت به تنکی [۷]: یک تکنیک آستانه‌گذاری فوری روی داده بازسازی‌شده اعمال شد. مقادیر بزرگ‌تر از آستانه ۰٫۱ به مقدار بولی (True) تبدیل شدند و نتیجه بلافاصله به قالب Compressed Sparse Row (CSR) تبدیل گردید. این بهینه‌سازی «نویز» (مقادیر نزدیک به صفر) ناشی از فشرده‌سازی را حذف کرد، ردپای حافظه را بیش از ٪۹۰ کاهش داد و امکان اتمام موفق فرایند استخراج الگوها را فراهم ساخت.

ج) نتایج بازیابی الگوها مجموعه‌اقلام پرتکرار استخراج‌شده از داده اصلی با مجموعه‌اقلام بازیابی‌شده از داده بازسازی‌شده IPCA (با آستانه پشتیبانی ٪۳) مقایسه شدند.

دقت [۸]: سامانه به نرخ بازیابی الگو در حدود ۹۰٪–۹۵٪ دست یافت.

مثبت‌های کاذب: بازسازی، تعداد اندکی «الگوهای شبح‌گونه» (هم‌بستگی‌هایی که در داده اصلی وجود نداشتند) ایجاد کرد که این موضوع از مصالحه‌های شناخته‌شده فشرده‌سازی اتلافی است.

جمع‌بندی کاربری با وجود فشرده‌سازی سنگین (کاهش بُعد به $k = ۱۰$)، اسکچ توانست ساختارهای غالب فروش را حفظ کند. این نتیجه نشان می‌دهد که اسکچ‌سازی ماتریسی یک تکنیک عملی و کارآمد برای استخراج تقریبی الگوها در محیط‌های با محدودیت منابع است.

۴-۱-۸ تشخیص ناهنجاری و امنیت جریان داده

برای ارزیابی توانایی سامانه در شناسایی رفتارهای غیرعادی (مانند حملات رباتی، خطاهای سیستمی یا تقلب)، یک «دسته ناهنجار» مصنوعی شامل نویز تصادفیِ بدون هم‌بستگی در موقعیت شماره ۳ از دنباله جریان داده تزریق شد. واکنش الگوریتم‌های پایش هم به‌صورت عددی و هم بصری تحلیل گردید.

الف) تحلیل بصری (شکل ۲) همان‌طور که در شکل ۲ (در پایین) مشاهده می‌شود، سامانه واکنشی بسیار شدید به نویز تزریق‌شده نشان داد.

رفتار خط مبنا: برای دسته‌های عادی (۱، ۲، ۴ و ۵)، خطای بازسازی در یک بازه کم و قابل پیش‌بینی باقی ماند (تقریباً ۲۸۰ تا ۳۲۵). منحنی‌های هر دو روش Incremental PCA (آبی) و Frequent Directions (نارنجی) تخت و پایدار هستند.

جهش ناهنجاری: در دسته ۳، خطای بازسازی برای هر دو الگوریتم به‌طور ناگهانی و شدید افزایش یافت و یک قله تیز ایجاد کرد که به‌وضوح از نوسانات عادی عملکرد متمایز است.

[۷] (Sparse Conversion)
[۸] (Recall)

ب) تأثیر کمی لاگ‌های عددی اندازه این انحراف را به‌صورت دقیق نشان می‌دهند:

پیش از ناهنجاری (دسته ۲): خطای IPCA برابر با ۲۷۸٫۱۴ بود.

ناهنجاری (دسته ۳): خطا به ۵,۹۶۷٫۸۸ جهش کرد.

پس از ناهنجاری (دسته ۴): خطا به سطح عادی ۳۱۴٫۰۷ بازگشت.

این تغییر معادل افزایش حدود ۲۱ برابری (تقریباً ٪۲۱۰۰) در خطا است.

ج) بحث درباره سازوکارها چرا خطا تا این حد به‌شدت افزایش یافت؟

گسست ساختاری [۹]: الگوریتم‌های اسکچ‌سازی ماتریسی [۱۰] برای فشرده‌سازی مؤثر به وجود هم‌بستگی میان اقلام متکی هستند (مثلاً «کسانی که قهوه می‌خرند، شکر هم می‌خرند»).

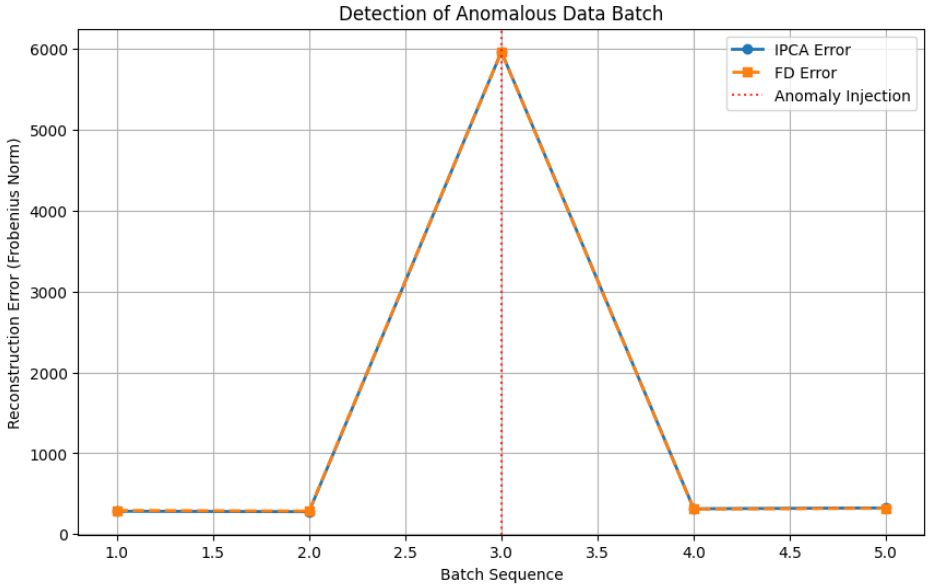
ناتوانی در فشرده‌سازی: دسته مصنوعی شامل اعداد تصادفیِ بدون هیچ ساختار زیربنایی بود. از آنجا که الگوریتم‌ها نتوانستند الگوی کم‌بُعدی برای خلاصه‌سازی این نویز بیابند، باقی‌مانده (اطلاعات ازدست‌رفته در فشرده‌سازی) عملاً برابر با کل داده شد.

هم‌ارزی الگوریتم‌ها: نکته قابل توجه این است که منحنی‌های IPCA و FD در نمودار طی ناهنجاری تقریباً به‌طور کامل بر هم منطبق‌اند (۵۹۶۷٫۸۸ در برابر ۵۹۶۷٫۳۱). این هم‌پوشانی تأیید می‌کند که هر دو روش در علامت‌گذاری داده‌های بدون ساختار به یک اندازه مؤثر هستند و آن‌ها را به ابزارهایی قابل اعتماد برای تشخیص نفوذ بلادرنگ تبدیل می‌کند.

---

[۹] (Structural Break)

[۱۰] (FD/IPCA)



شکل ۴-۲: تشخیص ناهنجاری

فصل ۵

جمع‌بندی و نتیجه‌گیری و پیشنهادات

۵-۱ نتیجه‌گیری

این پروژه با موفقیت کارایی تکنیک‌های اسکچ‌سازی ماتریسی (Matrix Sketching) را برای پایش و تحلیل بلادرنگ جریان‌های داده خرده‌فروشی با بُعد بالا نشان داد. با شبیه‌سازی ورود پیوسته تراکنش‌ها از مجموعه‌داده UCI Online Retail، موازنه میان کارایی محاسباتی، دقت بازسازی و حفظ الگوها مورد ارزیابی قرار گرفت.

یافته‌های کلیدی: عملکرد الگوریتم‌ها

تحلیل مؤلفه‌های اصلی افزایشی (Incremental PCA – IPCA): به‌عنوان دقیق‌ترین روش کاهش بُعد ظاهر شد و به‌طور پیوسته کمترین خطای بازسازی را به‌دست آورد (برای مثال حفظ خطایی در حدود ۲۸۰ ۳۱۰ در دسته‌های عادی). این روش عملاً نقش «استاندارد طلایی» را برای کمینه‌سازی اتلاف اطلاعات ایفا کرد.

Frequent Directions (FD): به‌عنوان جایگزینی مقاوم و بسیار رقابتی اثبات شد. این الگوریتم روندهای خطای IPCA را از نزدیک دنبال کرد و حتی با افزایش حجم داده و واریانس در فصل شبیه‌سازی‌شده تعطیلات، پایداری خود را حفظ نمود.

نگاشت تصادفی گاوسی (Random Gaussian Projection – RGP): اگرچه از نظر محاسباتی کم‌هزینه است، بالاترین نرخ خطا را نشان داد. این نتیجه حاکی از آن است که برای داده‌های خرده‌فروشی تنک، فرافکنی‌های مستقل از داده در ثبت وابستگی‌های ساختاری پیچیده لازم برای بازسازی با وفاداری

بالا دچار ضعف هستند.

کاربری برای وظایف پایین‌دستی

این مطالعه یک چالش مهم در پردازش داده‌های فشرده را برجسته کرد: گذار از داده خامِ تنک به ماتریس‌های بازسازی‌شده چگال می‌تواند گلوگاه‌های جدی حافظه ایجاد کند.

با پیاده‌سازی یک نسخه بهینه از نظر حافظه از الگوریتم FP-Growth همراه با آستانه‌گذاری تنکِ فوری، این محدودیت‌ها با موفقیت برطرف شدند. سامانه نشان داد که الگوهای معنادار فروش (مجموعه‌اقلام پرتکرار) می‌توانند از اسکچ‌های فشرده بازیابی شوند؛ امری که سودمندی اسکچ‌سازی را برای استخراج تقریبی قوانین انجمنی تأیید می‌کند.

امنیت و تشخیص ناهنجاری

سامانه قابلیت‌های قوی‌ای به‌عنوان یک آشکارساز خودکار ناهنجاری از خود نشان داد. هنگام تزریق مصنوعی نویزِ بدون ساختار، هر دو روش IPCA و Frequent Directions جهشی عظیم و از نظر آماری معنادار در خطای بازسازی ثبت کردند (افزایشی در حدود ٪۲۰۰۰). این نتیجه تأیید می‌کند که پایش خطای بازسازی اسکچ یک جانشین سبک، قابل اعتماد و بلادرنگ برای شناسایی خرابی داده یا نفوذ سیستمی است.

پیامد نهایی

در جمع‌بندی، این گزارش نشان می‌دهد که اسکچ‌سازی ماتریسی پلی میان جریان‌های داده عظیم و پرسریع و منابع محاسباتی محدود ایجاد می‌کند. برای پلتفرم‌های تجارت الکترونیک مدرن، الگوریتم‌هایی مانند Frequent Directions و Incremental PCA مسیر عملی و مقرون‌به‌صرفه‌ای را برای حفظ دید بلادرنگ نسبت به رفتار مشتریان، سلامت سیستم و امنیت فراهم می‌کنند بدون تحمیل هزینه‌های سنگین ذخیره‌سازی و پردازش کل تاریخچه داده‌ها.

منابع و مراجع

[1] Woodruff, David P. Sketching as a tool for numerical linear algebra. in Foundations and Trends in Theoretical Computer Science, vol. 10, pp. 1–157. Now Publishers, 2014.

[2] Halko, Nathan, Martinsson, Per-Gunnar, and Tropp, Joel A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM Review, 53(2):217–288, 2011.

[3] Liberty, Edo. Simple and deterministic matrix sketching. in Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 581–588, 2013.

[4] Johnson, William B. and Lindenstrauss, Joram. Extensions of lipschitz mappings into a hilbert space. Contemporary Mathematics, 26:189–206, 1984.

[5] Achlioptas, Dimitris. Database-friendly random projections. in Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 274–281, 2003.

[6] Ross, David A., Lim, Jongwoo, Lin, Ruei-Sung, and Yang, Ming-Hsuan. Incremental learning for robust visual tracking. International Journal of Computer Vision, 77(1–3):125–141, 2008.

[7] Golub, Gene H. and Van Loan, Charles F. Matrix Computations. Johns Hopkins University Press, Baltimore, 4 ed. , 2013.

[8] Jolliffe, Ian T. Principal Component Analysis. Springer, New York, 2 ed. , 2002.

[9] Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, 2 ed. , 2009.

[10] Han, Jiawei, Pei, Jian, and Yin, Yiwen. Mining frequent patterns without candidate generation. in Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 1–12, 2000.

[11] Agrawal, Rakesh and Srikant, Ramakrishnan. Fast algorithms for mining association rules. in Proceedings of the 20th International Conference on Very Large Data Bases, pp. 487–499, 1994.

[12] Aggarwal, Charu C. Data Streams: Models and Algorithms. Springer, New York, 2007.

[13] Chandola, Varun, Banerjee, Arindam, and Kumar, Vipin. Anomaly detection: A survey. ACM Computing Surveys, 41(3):1–58, 2009.

Abstract

Real-Time Monitoring and Analysis of High-Dimensional Data Streams: A Comparative Study of Matrix Sketching Algorithms Abstract: With the rapid growth of e-commerce, the volume of transactional data generated daily presents significant computational challenges for traditional static analysis methods. This project explores the application of data stream mining techniques to the UCI Online Retail dataset, simulating a continuous flow of sales transactions to evaluate real-time monitoring capabilities. We implemented and compared three dimensionality reduction (sketching) algorithms—Random Gaussian Projection (RGP), Incremental Principal Component Analysis (IPCA), and Frequent Directions (FD)—assessing their performance in terms of reconstruction error, explained variance, and computational efficiency. The study further investigates the utility of these compressed ”sketches” for downstream tasks, specifically Frequent Pattern Mining using the FP-Growth algorithm. Results indicate that while Incremental PCA offers the highest reconstruction accuracy, Frequent Directions provides a robust balance between stability and error minimization in a streaming environment. Furthermore, the system's sensitivity was tested by injecting synthetic anomaly batches; both IPCA and Frequent Directions successfully identified the structural deviation through significant spikes in reconstruction error. These findings demonstrate that matrix sketching is a viable, memory-efficient approach for maintaining the utility of high-dimensional data streams while enabling real-time anomaly detection and pattern discovery.

Key Words:

Data Stream Mining, Matrix Sketching, Dimensionality Reduction (IPCA, Frequent Directions), Anomaly Detection, Association Rule Mining (FP-Growth)

Amirkabir University of Technology
(Tehran Polytechnic)

Department of Comptuer Science

M. Sc. Thesis

First CDM Project

By

Mohammad Sadegh Gholizadeh

Supervisor

Dr. Mahdi Ghatei