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

پایان‌نامه کارشناسی‌ارشد
گرایش سیستم‌های کامپیوتری

کاهش بعد داده‌های بزرگ مقیاس با استفاده از نگاشت تصادفی

نگارش
سیامک دهبد

استاد راهنما
دکتر عادل محمدپور

استاد مشاور
دکتر هادی زارع

دی ۱۳۹۷

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)

تعهدنامه اصالت اثر

تاریخ: دی ۱۳۹۷

اینجانب **سیامک دهبد** 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.

سیامک دهبد

امضا

تقدیم:

سپاس گزاری

با تشکر از استاد گرامی دکتر محمدپور بابت همراهی و صبر ایشان

سیامک دهبد
دی ۱۳۹۷

چکیده

با ظهور داده‌های بزرگ مقیاس و ناتوانی در نگهداری و پردازش این داده در حافظه، مسئله کاهش بعد اهمیت زیادی پیدا کرده است. یکی از روش‌های کاهش بعد، تصویر تصادفی است که می‌تواند بر روی کلان‌داده‌هایی که بزرگ مقیاس هستند و همچنین بر روی جریان‌های داده، اعمال شود. مبنای این روش ضرب ماتریسی داده‌های اولیه در یک ماتریس تصویرگر است که بعد داده‌های اولیه را کاهش داده ولی اطلاعات آماری مورد نیاز در داده‌های اولیه را با دقت مورد نیاز نگه می‌دارد.

روش تصویر تصادفی برای کاهش بعد داده‌های بزرگ مقیاس مزایای متعددی نسبت به روش‌های دیگر کاهش بعد دارد. در این پایان‌نامه این روش برای داده‌های بزرگ مقیاس با دیگر روش‌های کاهش بعد مقایسه شده است. همچنین توانایی این روش برای داده‌هایی با توزیع پایدار غیر نرمال با دیگر روش‌های کاهش بعد مقایسه شده است.

واژه‌های کلیدی:

کاهش بعد، تصویر تصادفی، توزیع پایدار، داده‌های بزرگ مقیاس

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 اول

مقدمه

عمومیت پیدا کردن داده‌های حجیم مانند داده‌های حجیم تحت وب و جریان‌های داده بزرگ در کاربردهای جدید، موجب به وجود آمدن فرصت‌های و چالش‌هایی برای مهندسین و دانشمندان شده است. [۴۷] برای مثال، زمانی که ماتریس داده $\mathbf{A} \in \mathbb{R}^{n\times D}$ ابعادی در حد وب داشته باشد، عملیات ساده‌ای مانند محاسبه $\mathbf{A}\mathbf{A}^T$ سخت می‌شود. برای ارائه و نگهداری داده‌های حجیم در حافظه‌ای کوچک و برای استخراج اطلاعات آماری اصلی از مجوعه‌ای از بیانی محدود، روش‌های گوناگونی نمونه‌برداری توسعه یافته است. به طور کلی روش تصویر تصادفی پایدار [۱] برای داده‌های با دم سنگین خیلی خوب کار می‌کند.

روش تصویر تصادفی پایدار، ماتریس داده‌های اولیه $\mathbf{A} \in \mathbb{R}^{n\times D}$ را در ماتریس تصادفی $\mathbf{R} \in \mathbb{R}^{D\times k}$ $(k \ll D)$ ضرب می‌کند و نتیجه ماتریس $\mathbf{B} = \mathbf{A}\mathbf{R} \in \mathbb{R}^{n\times k}$ است. معمولا درایه‌های ماتریس تصادفی $\mathbf{R}$ به صورت i.i.d[۲] از یک توزیع α-پایدار متقارن انتخاب می‌شوند ($0 < \alpha \le 2$). ما می‌توانیم مشخصه‌های l_α را در $\mathbf{A}$ بر اساس $\mathbf{B}$ تخمین بزنیم. در مورد حالت l_2 مزیت توزیع تصادفی پایدار توسط لم JL[۳] برجسته شده است. لم JL بیان می‌دارد که کافی است $k = O(\frac{\log n}{\epsilon^2})$ باشد تا هم فاصله دو به دویی با نرم l_α در $\mathbf{A}$ را بتوان با ضریب $1 \pm \epsilon$ از روی ماتریس $\mathbf{B}$ تخمین زد. در تز Ping Li [۴۷] لمی مشابه لم JL برای $0 < \alpha < 2$ اثبات شده است. روش تصویر تصادفی پایدار به یک مسئله تخمین آماری کاهش می‌یابد برای تخمین پارامتر مقیاس برای یک توضیع پایدار α متقارن. این مسئله از این جهت مورد توجه قرار می‌گیرد زیرا ما به دنبال برآوردی می‌گردیم که هم از نظر آماری درست باشند و هم از نظر محاسباتی مقرون به صرفه. برآوردگرهای مختلفی را مطالعه و مقایسه کردیم. شامل میانگین حسابی، میانگین هندسی، میانگین هارمونیک، تقسیم توانی [۴] و برآوردگر حداکثر بزرگنمایی.

در این پایان‌نامه ما به بررسی موارد خاصی از تصویر تصادفی پایدار می‌پردازیم. برای نرم l_2 ارتقایی را با استفاده از اطلاعات حاشیه‌ای پیشنهاد می‌کنیم. همچنین برای حالت l_2 می‌توان ماتریس تصویرگر را از یک توزیع زیرگوسی [۵] بسیار کوچکتر به جای توزیع نرمال انتخاب کرد. با در نظر گرفتن محدودیت‌های معقولی می‌توان، از یک توزیع خاص زیر گوسی استفاده کرد. این توزیع شامل $[-1, 0, 1]$ با احتمالات $\{\frac{1}{s}, 1 - \frac{1}{2s}, \frac{1}{s}\}$ با مقادیر بسیار بزرگی برای s (به عبارتی، تصویر تصادف خیلی گسسته [۶]) می‌تواند به

[1] stable random projection

[2] independent and identically distributed random variables

[3] Johnson-Lindenstrauss

[4] fractional power

[5] sub-Gaussian

[6] very sparse random projections

خوبی تصویر تصادفی نرمال عمل کند. برای حالت نرم l_1 به عبارتی دیگر تصویر تصادفی کوچی [7] انجام تخمین کاری نسبتا جذاب است. برای مثال، محاسبه برآوردگر بیشینه درستنمایی MLE در این حالت از لحاظ محاسباتی ممکن است. و یک توزیع معکوس گاوسی [8] برای مدل‌سازی دقیق توزیع MLE بیان شده است.

روش تصویر تصادفی ار پراکندگی داده‌ها استفاده‌ای نمی‌کند. در حالی که داده‌های بزرگ مقیاس معمولا بسیار پراکنده هستند. از روش تصویر تصادفی می‌توان برای حل مسائل بزرگ مقیاس در علوم و مهندسی در موتورهای جستجو و سیستم‌های اخذ داده، پایگاه‌های داده، سیستم‌های جریان داده جدید، جبر خطی عددی و بسیاری از کارهای یادگیری ماشین و داده کاوی که شامل محاسبه حجیم فاصله‌ها است، استفاده کرد.

[7] Cauchy random projections
[8] inverse Gaussian

فصل دوم
کاهش بعد و داده‌های بزرگ مقیاس

۲-۱ کاهش بعد

۲-۱-۱ تحلیل مولفه‌های اصلی

تحلیل چند متغیره معمولا بر روی داده‌هایی که شامل تعداد زیادی از متغیرهای مرتبط با هم هستند انجام می‌شود.

روش تحلیل مؤلفه‌های اصلی(PCA)[1] یک ابزار کاهش بعد است که می توان از آن برای کاهش یک مجموعه‌ی بزرگ از متغیرها به مجموعه‌ی کوچکتری که غالب اطلاعات مجموعه‌ی بزرگ را دارد استفاده کرد.

روش تحلیل مولفه‌های اصلی یک تابع ریاضی است که تعدادی متغیر (احتمالا) همبسته را به تعداد (کمتر یا مساوی) متغیرهای غیرهمبسته به نام «مولفه‌های اصلی» تبدیل می‌کند.

بیشترین میزان اطلاعات ممکن در داده در اولین مولفه اصلی ثبت می‌شود. بیشترین میزان اطلاعات ممکن باقیمانده به ترتیب در مولفه‌های بعدی ثبت می‌شوند.

تحلیل مولفه‌های اصلی مشابه یک تابع چند متغیره دیگر به نام تحلیل عاملی است. این دو روش در موارد زیادی با یکدیگر اشتباه گرفته می‌شوند، و تفاوت بین این دو، و انواع تحلیل هایی که هر یک برایشان مناسب تر هستند به درستی تشخیص داده نمی‌شود. به طور سنتی، تحلیل مولفه های اصلی بر روی ماتریس های متقارن مربعی انجام می شود. این ماتریس‌ها می تواند یکی از انواع SSCP[2] (مجموع خالص مربعات و ضرب های داخلی)، ماتریس کوواریانس[3] (مجموع مقیاس شده مربعات و ضرب های داخلی)، یا ماتریس همبستگی[4] (مجموع مربعات و ضرب های داخلی داده های استاندارد شده) باشد.

نتایج تحلیل روی ماتریس‌های از نوع SSCP و کوواریانس تغییری ندارند، چرا که تغییرات آنها فقط در یک ضریب مقیاس قابل مشاهده است.

از ماتریس‌های همبستگی زمانی استفاده می‌شود که واریانس متغیرهای منحصر به فرد تفاوت‌های زیادی داشته باشد، و یا واحدهای اندازه‌گیری این متغیرها متفاوت باشد.

[1] principal components analysis
[2] pure sums of squares and cross products
[3] covariance
[4] correlation

اهداف تحلیل مولفه‌های اصلی

تحلیل مولفه‌های اصلی فضای مشخصه‌ها را از تعداد زیادی متغیر به تعداد کمتری عامل کاهش می‌دهد، و یک تابع «غیر وابسته» است (یعنی نیازی وجود ندارد که یک متغیر وابسته تعیین شده باشد).

تحلیل مولفه های اصلی یک روش کاهش یا فشرده سازی ابعاد است. هدف، کاهش بعد است و تضمینی وجود ندارد که این ابعاد قابل تفسیر باشند.

در نهایت، انتظار آن است که زیرمجموعه‌ای از متغیرها از یک مجموعه بزرگتر انتخاب شود، به گونه‌ای که متغیرهای اولیه بیشترین همبستگی را با مولفه اصلی داشته باشند.

تحلیل مولفه‌های اصلی به دنبال رسیدن به ترکیبی خطی از متغیرها است، به گونه ای که بیشینه واریانس از آنها قابل استخراج باشد. پس از آن، این واریانس حذف شده و ترکیب خطی دومی جستجو می‌شود که بیشینه باقی‌مانده واریانس را توصیف می‌کند، و این روند ادامه پیدا می کند. به این روش، روش محور اصلی گفته می شود و عامل های متعامد غیرهمبسته را به دست می‌دهد. تحلیل مولفه‌های اصلی، واریانس (مشترک و یکتای) کل را شرح می‌دهند.

ویژه بردارها: مولفه‌های اصلی، هر دو واریانس مشترک و یکتای متغیرها را منعکس می‌کنند و بنابراین ممکن است که این روش به عنوان یک روش واریانس محور دیده شود که هم به دنبال بازتولید واریانس متغیر کل با تمام مولفه‌ها و هم بازتولید همبستگی‌ها است. مولفه های اصلی، ترکیب های خطی از متغیرهای اولیه هستند که بر اساس میزان سهمشان در به وجود آمدن واریانس در یک بعد متعامد مشخص، وزن دهی می‌شوند. وزن‌های داده شده برای هر یک از مولفه‌های اصلی نسبت به داده‌های اولیه، ویژه بردارها هستند.

ویژه مقدارها: ویژه مقدار یک مولفه، واریانس همه‌ی متغیرهایی را که به آن عامل مرتبط هستند اندازه گیری می کند. نسبت ویژه مقدارها، نسبت اهمیت توصیفی عامل ها با توجه به متغیرها است. اگر یک عامل دارای ویژه مقدار پایین باشد، نشانگر آن است که اثر کمی روی توصیف واریانس در متغیرها دارد، و ممکن است از آن در مقابل عامل های مهمتر چشم پوشی شود. در ویژه مقدارها میزان تغییر در نمونه کل حساب شده است.

ویژه مقدار یک عامل ممکن است حاصل جمع مربعات عامل های تمامی متغیرها باشد. باید توجه شود که ویژه مقدارهای مرتبط با راه حل های دورانی و غیردورانی متفاوت خواهند بود، اگرچه مقدار کل آنها یکسان است.

برای به دست آوردن واریانس همه متغیرها که توسط عامل لحاظ می شود، مجموع مربعات بارگذاری

های عامل برای آن عامل (ستون) را جمع کرده، و بر تعداد متغیرها تقسیم می کنیم. (توجه کنید که تعداد متغیرها برابر با مجموع واریانس آنهاست، چرا که واریانس یک متغیر استاندارد شده مساوی با ۱ است). این کار مشابه تقسیم ویژه مقدار عامل بر تعداد متغیرها است.

امتیاز PC : این امتیازها، امتیازهای هر نمونه (ردیف) در هر عامل (ستون) هستند. امتیاز عامل برای یک نمونه و برای یک عامل داده شده، به صورت مجموع حاصل ضرب امتیاز استاندارد نمونه در هر متغیر با بارگذاری عامل مربوطه برای عامل داده شده محاسبه می‌شود. [۱]

■

۲-۲ خوشه‌بندی و شاخص‌های سنجش

۲-۲-۱ شاخص رند تعدیل شده

برای مقایسه نتایج خوشه‌بندی در کنار شاخص‌های بیرونی، نیازمند معیارهای مورد توافق هستیم. برای مجموعه‌ی n عضوی $S = \{O_1, \ldots, O_n\}$ ، فرض کنید که $U = \{u_1, \ldots, u_R\}$ و $V = \{v_1, \ldots, v_R\}$ نمایانگر دو افراز متفاوت از عضوهای S هستند به گونه ای که $\bigcup_{j=1}^{C} v_j = S = \bigcup_{i=1}^{R} u_i$ و $u_i \cap u_{i'} = \emptyset = v_j \cap v_{j'}$ برای $1 \leq i \neq i' \leq R$ و $1 \leq j \neq j' \leq C$. فرض کنید که U شاخص خارجی و V نتیجه خوشه‌بندی است. فرض کنید a تعداد جفت عضوهایی باشد که در یک کلاس یکسان U و خوشه یکسان V هستند، b تعداد جفت عضوهایی باشد که در یک کلاس یکسان U اما خوشه متفاوت V هستند، c تعداد جفت عضوهایی باشد که در یک کلاس متفاوت U و خوشه یکسان V هستند، و d تعداد جفت عضوهایی باشد که در یک کلاس متفاوت U و خوشه متفاوت V هستند. می‌توان مقادیر a و d را به عنوان موارد توافق، و مقادیر b و c را به عنوان موارد عدم توافق تعریف کرد. شاخص رند[۵] [۵۴] به سادگی به صورت $\frac{a+b}{a+b+c+d}$ تعریف می‌شود. شاخص رند عددی بین ۰ و ۱ است. زمانی که دو خوشه کاملا در حالت توافق باشند، مقدار شاخص رند برابر با ۱ خواهد بود.

یکی از مشکلات شاخص رند آن است که مقدار مورد انتظار شاخص رند دو خوشه تصادفی یک مقدار ثابت (به عنوان مثال صفر) نیست. مقدار شاخص رند تعدیل شده[۶] پیشنهادی توسط هوبرت و اربی[۷] [۳۸] بر پایه این فرض است که توزیع فوق‌هندسی[۸] تعمیم یافته برای مدل تصادفی استفاده می‌شود، به عبارت

[5] Rand index
[6] adjusted Rand index
[7] Hubert and Arabie
[8] hypergeometric

دیگر خوشه های U و V به شکلی تصادفی انتخاب می‌شوند که تعداد عضوهای کلاس‌ها و خوشه‌ها ثابت باشد. فرض کنید n_{ij} تعداد عضوهایی باشد که هم در کلاس u_i و هم در خوشه v_j هستند. در نظر بگیرید $n_{i.}$ و $n_{.j}$ به ترتیب تعداد اعضا در کلاس u_i و خوشه‌ی v_i هستند. تمامی این نشانه‌گزاری‌ها در جدول جدول ۲-۱ بیان شده‌اند.

جدول ۲-۱: نشانه‌گزاری برای جدول پیشایندی مقایسه دو بخش

Class \Cluster	v_1	v_2	$\ldots$	v_C	Sums
u_1	n_{11}	n_{12}	$\ldots$	n_{1C}	$n_{1.}$
u_2	n_{21}	n_{22}	$\ldots$	n_{2C}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
u_R	n_{R1}	n_{R2}	$\ldots$	n_{RC}	$n_{R.}$
Sums	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.C}$	$n_{..} = n$

شکل کلی شاخص با در نظر گرفتن مقادیر انتظاری ثابت بدین شکل است که $\frac{\text{شاخص}-\text{مقدار مورد انتظار شاخص}}{\text{حداکثر شاخص}-\text{مقدار مورد انتظار شاخص}}$ [۵۸]

۲-۳ داده‌های حجیم

عبارات زیر از سایت *Information Week* نقل قول شده‌اند [۲] :

- مقدار داده‌ای که توسط کسب و کارها ذخیره می‌شود تقریبا هر ۱۲ تا ۱۸ ماه دو برابر می‌شود.
- پایگاه داده‌ها بیشتر هم زمان شده‌اند. فروشگاه‌های زنجیره‌ای Wall-Marat داده‌های فروش را هر ساعت به روز می‌کند.
- اضافه شدن یک میلیون خط داده اجازه جستجوهای پیچیده‌تری را می‌دهد. شرکت EBay به کارمندان اجازه می‌دهد برای بدست آوردن درکی عمیق‌تر در خصوص رفتار مشتریان در میان داده‌های حراج در بازه‌های زمانی کوتاه جستجو کنند.
- بزرگترین پایگاه داده‌ها توسط، مرکز شتابدهنده خطی استاندارد، مرکز تحقیقات ناسا، آژانس امنیت ملی و ... در ابعادی در محدوده‌ی پتابایت (هزار ترابایت 10^15 بایت)، اداره می‌شوند.

پدیده نو ظهور مجموعه داده‌ای حجیم، چالش‌های محاسباتی در بسیاری کاربردهای علمی و تجاری

به وجود آورده است. شامل اخترفیزیک، بیوتکنولوزی، جمعیت شناسی[9] ، مالی، سیستم‌های اطلاعات جغرافیایی، دولت، دارو، ارتباطات از راه دور، محیط زیست و اینترنت.

۲-۳-۱ داده‌های حجیم وب

وب چقدر بزرگ است؟ جدول ۲-۲ نشان‌دهنده تعداد بازدید صفحات در موتورهای جستجوی امروزی است. به طور تخمینی حدود $D = 10^10$ صفحه‌ی وب را می‌توان بر اساس بازدید دو واژه‌ی بسیار پر کاربرد « A » و « THE » تخمین زد. جدول ۲-۲ همچنین نشان می‌دهد که حتی کلماتی که به ندرت کاربرد دارند هم تعداد زیادی بازدید دارند.

جدول ۲-۲: تعداد بازدید صفحات برای کلمات با بازخورد بالا و کلمات با بازخورد نادر

Query	Google	Bing
A	25,270,000,000	175,000,000
The	25,270,000,000	101,000,000
Kalevala	7,440,000	939,000
Griseofulvin	1,163,000	332,000
Saccade	1,030,000	388,000

کلماتی با بازخورد معمولی چه میزان بازدید دارند؟ برای جواب این سوال ما به طور تصادفی ۱۵ صفحه از لغتنامه‌ی آموزشی انتخاب می‌کنیم. [۳۷] (لغتنامه‌ای با ۵۷،۱۰۰ کلمه) و اولین کلمه در هر صفحه را مد نظر قرار می‌دهیم. میانه‌ی آماری بر اساس جستجوگر گوگل ۱۰ میلیون صفحه برای کلمه است.

زبان انگلیسی چند کلمه دارد؟ در اینجا عبارتی را از AskOxford.com نقل قول می‌کنیم:

« این بیان میدارد که حداقل یک چهارم میلیون واژه‌ی انگلیسی مستقل وجود دارد. به جز افعال صرفی و کلمات فنی و ناحیه‌ای که توسط OED[10] تحت پوشش قرار نمی‌گیرند یا کلماتی که هنوز به لغتنامه‌های منتشر شده اضافه نشده‌اند. در صورتی که این موارد هم در نظر گرفته شوند تعداد لغات در حدود سه چهارم میلیون لغت خواهد بود »

بنابراین اگر یک ماتریس «عبارت به سند» $\mathbf{A} \in \mathbb{R}^{n\times D}$ در نظر بگیریم. در ابعاد وب این ماتریس در ابعاد $n \approx 10^6$ و $D \approx 10^10$ بزرگ خواهد شد. در اینجا عدد (i, j) در $\mathbf{A}$ تعداد ظهور واژه i در سند j را نشان می‌دهد.

[9] demographics

[10] Oxford english dictionary

کارکردن با ماترسی در این ابعاد بزرگ چالش برانگیز است. برای مثال، شاخص LSI [11] [۲۷] و یک مدل موضوعی فراگیر، از SVD [12] بر روی ماتریس عبارت به سند استفاده می‌کند. که انجام این عملیات در ابعاد وب قطعا غیرممکن است.

یک مشکل اصلی در قبال مجموعه داده‌های سنگین، حافظه کامپیوتر است. به این دلیل که ابعاد و سرعت حافظه فیزیکی بسیار رشد کمتری در مقایسه با پردازنده‌ها (CPU) دارد. این پدیده به عنوان دیوار حافظه شناخته می‌شود [۵۱، ۵۷] . برای مثال، هر چند ممکن است تمامی رخدادهای همزمان دوتایی از پیش محاسبه شوند، ولی نگهداری این حجم از داده در حافظه غیر ممکن است. علاوه بر این، گاهی اوقات تخصیص‌هایی با بیش از دو عامل هم اهمیت پیدا می‌کنند زیرا درخواست‌ها ممکن است شامل بیش از دو واژه هم باشند. یک راه حل ممکن این است که یک «نمونه» از **A** نگهداری شود و همزمانی‌ها بر اساس این نمونه در حین کار تخمین زده شوند. ما حدس می‌زنیم که این روش توسط موتورهای جستجوی امروزی مورد استفاده قرار می‌گیرد، هر چند که روش واقعی قطعا جزو اسرار تجاری آن‌ها است.

هر چند که انتظار می‌رود تخمین‌ها سازگار باشند و فرکانس‌های جفت شده باید با افزایش عبارت به درخواست، کاهش پیدا کنند. جدول ۲-۳ نشان می‌دهد که تخمین‌های بیان شده با موتورهای جستجوی فعلی، همیشه سازگار نیستند.

جدول ۲-۳: با افزایش تعداد عبارات در درخواست، باید فرکانس‌های جفت شده کاهش پیدا کنند. ولی تخمین‌های بیان شده توسط موتورهای جستجو گاهی این موضوع تثبیت شده را نقض می‌کنند.

Query	Hits(Bing)	Hits(Google)
America	150,731,182	393,000,000
America & China	15,240,116	66,000,000
America & China & Britain	235,111	6,090,000
America & CHina & Britain & Japan	154,444	23,300,000

با اینکه، تعداد کل واژه‌های انگلیسی (که به‌طور صحیح نوشته شده‌اند) هم اکنون شگفت‌اور است، در بسیاری کاربردهای متن کاوی، ما باید با ابعاد بسیار بزرگتری سر و کار داشته باشیم. در حالی که یک سند ممکن است بیانگر برداری از تک واژه‌ها باشد (به عبارت دیگر، مدل کیسه لغات [13]). معمولا بهتر است سند به عنوان یک بردار از لغات به صورت l پیوسته [14] [۱۸] بیان شود. برای مثال، با استفاده از

[11] latent semantic indexing
[12] singular value decomposition
[13] bag-of-words
[14] l-shingles

مدل ۳ پیوسته، جمله‌ی ”It is a nice day“ به مجموعه‌ی زیر تجزیه می‌شود. {”it is a“, ”is a nice“, ”a nice day“} این مدل به طور جشمگیری ابعداد داده‌ها را افزایش می‌دهد. به خاطر اینکه، اگر مجموعه‌ی 10^6 تک لغت انگلیسی موجود داشته باشد. مدل ۳ پیوسته تعداد ابعاد را از 10^6 به 10^{18} افزایش می‌دهد.

۲-۳-۲ جریان‌های داده‌ی حجیم

در بسیاری کاربردهای جدید پردازش داده، جریان‌های داده‌ی حجیم نقش بنیادی دارند. جریان‌های داده‌ای که از روترهای اینترنت، سوئیچ‌های تلفن، رصد امسفر، شبکه‌های سنسور، شرایط ترافیکی بزرگراهی، داده‌های مالی و غیره [۵، ۵۲، ۲۶، ۱۱، ۳۹، ۴۴، ۳۶] حاصل می‌شوند.

برخلاف پایگاه داده‌های سنتی، معمول نیست که جریان‌های داده‌ی حجیم (که با سرعت زیادی منتقل می‌شوند) در جای نگهداری شوند. بنابراین پردازش معمولا به طور همزمان انجام می‌شوند. برای مثال، گاهی اوقات «رصد تصویری» داده‌ها با رصد تغییرات زمانی برخی آماره‌ها کفایت می‌کند. برای مثال آماره‌های نظیر: مجموع، تعداد آیتم‌های مجزا، برخی نرم‌های l_α . در برخی کاربردها (برای مثال، طبقه‌بندی صدا/محتوا و جدا سازی) نیاز است یک مدل یادگیری آماری برای کلاسه‌بندی[۱۵] یا خوشه‌بندی[۱۶] جریان داده‌های حجیم توین شود. ولی معمولا فقط می‌توانیم یک‌بار داده‌ها را مورد بررسی قرار دهیم.

یک خاصیت مهم جریان‌های داده‌ای این است که دینامیک هستند. به عنوان یک مدل محبوب، جریان u شامل ورودی‌های (i, u_i) است که $i = 1 to D$. برای مثال، $D = 2^{64}$ زمانی که جریان بیان‌گر IP آدرس‌ها است. [۱۷] ورودی‌ها ممکن است به هر ترتیبی باشند و ممکن است مرتبا به روز شوند. ذات دینامیک جریان داده‌های حجیم فرآیند نمونه‌گیری را بسیار چالش‌برانگیزتر ا زمانی می‌کند که با داده‌های ایستا سر و کار داریم.

[15] classification

[16] clustering

[۱۷] هرچند ما بیشتر اوقات تعداد دقیق ابعاد (D) یک جریان داده را نمی‌دانیم ولی در بیشتر کاربردها کافی است حد بالایی محافظه‌کارانه‌ای را در نظر بگیریم. برای مثال $D = 2^{64}$ زمانی که جریان بیانگر IP های ورودی است. همچنین این یکی از دلایلی است که داده‌ها بسیار پراکنده هستند. به این نکته توجه داشته باشید که ابعاد بسیار بزرک تاثیری در محاسبه‌ی فاصله‌ها و نمونه‌گیری طی الگوریتم‌های معرفی شده در این پایان‌نامه ندارد.

۲-۴ چالش‌های نمونه‌گیری از داده‌های حجیم

در حالی که مسائل جذاب و چالش‌برانگیزی با ورود داده‌های حجیم شکل گرفته‌اند، این پایان‌نامه بر روی توسعه‌ی روش‌های نمونه‌گیری برای محاسبه فاصله در داده‌هایی با ابعاد بسیار بالا با استفاده از حافظه محدود تمرکز دارد.

در کاربردهای مدل‌سازی آماری و یادگیری ماشین، در اغلب موارد به جای داده‌های اصلی به فاصله، به خوصو فاصله‌ی جفتی نیاز داریم. برای مثال، محاسبه ماتریس گرام[۱۸] $\mathbf{AA^T}$ در آمار و یادگیری ماشین معمول است. $\mathbf{AA^T}$ بیانگر همه‌ی ضرب‌های داخلی دوتایی در ماتریس داده‌ی $\mathbf{A}$ است.

دو داده‌ی $u_1, u_2 \in \mathbb{R}^D$ داده شده‌اند. ضرب داخلی آن‌ها (که با a نمایش داده می‌شود) و l_α (که با $d_{(\alpha)}$ نمایش داده می‌شوند با عبارات زیر تعریف می‌شوند: [۱۹]

$$a = u_1^T u_2 = \sum_{i=1}^{D} u_{1,i} u_{2,i} \tag{۲-۱}$$

$$d_{(\alpha)} = \sum_{i=1}^{D} |u_1 - u_2|^\alpha \tag{۲-۲}$$

به این نکته توجه داشته باشید که هم ضرب داخلی و هم فاصله به شکل جمع D جمله تعریف می‌شوند. بنابراین، زمانی که داده‌ها به اندازه‌ای حجیم هستند که نمی‌توان به طور کارا آن‌ها را مدیریت کرد، نمونه‌گیری خیلی عادی به نظر می‌رسد تا بتوان با انتخاب تصادفی k عضو از D جمله تخمینی از مجموع به دست آوریم (با ضریب مقیاس $\frac{D}{k}$). در خصوص ماتریس داده‌ی $\mathbf{A} \in \mathbb{R}^{n\times D}$ انتخاب تصادفی مختصات [۲۰] ، k ستون را از ماتریس داده به طور یکنواخت و تصادفی انتخاب می‌کند.

نمونه‌گیری از این جهت سودمند است که هم سایکل‌های کاری CPU را کاهش می‌دهد و هم در حافظه صرفه‌جویی می‌کند. در کابردهای جدید، در اغلب موارد صرفه‌جویی در حافظه از اهمیت بیشتری برخوردار است. در نیم قرن گذشته گلوگاه محاسباتی حافظه بوده است، نه پردازشگر. سرعت پردازشگرها با نرخ تقریبی ۷۵ درصد در سال رو به افزایش است. در حالی که سرعت حافظه تقریبا سالی ۷ درصد

[18] Gram matrix

[۱۹] ما فاصله l_α را به صورت $d_{(\alpha)} = \sum_{i=1}^{D}|u_1 - u_2|^\alpha$ تعریف کرده‌ایم. به جای اینکه به شکل $(\sum_{i=1}^{D}|u_1 - u_2|^\alpha)^{1/\alpha}$ تعریف کنیم. زیرا شکل اول در کاربردهای عملی عمومیت بیشتری دارد. برای مثال، لم JL ، در ادبیات معمولا به شکل توان دو l_2 بیان می‌شود. $\sum_{i=1}^{D}|u_1 - u_2|^2$ به جای $(\sum_{i=1}^{D}|u_1 - u_2|^2)^{1/2}$. در این پایان‌نامه، ما برای سادگی $\sum_{i=1}^{D}|u_1 - u_2|^2$ را «فاصله l_2» بیان می‌کنیم به جای «مربع فاصله‌ی l_2».

[20] random coordinate sampling

افزایش می‌یابد [۵۱] . این پدیده به عنوان «دیوار حافظه»[21] شناخته می‌شود [۵۱، ۵۷] . بنابراین در کاربردهایی که شامل مجموعه داده‌های حجیم می‌شوند، بحرانی‌ترین کار بیان کردن داده‌ها است. برای مثال، از طریق نمونه‌گیری با فرمی فشرده برای قرارگیری در ابعاد حافظه در دسترس.

۲-۴-۱ مزایای نمونه‌گیری تصادفی مختصات

نمونه‌گیری تصادفی مختصات به دو دلیلی معمولا انتخاب پیش‌فرض است.

- **سادگی** این روش از لحاظ زمانی تنها از مرتبه $O(nk)$ برای نمونه‌گیری k ستون از $\mathbf{A} \in \mathbb{R}^{n \times D}$ طول می‌کشد.
- **انعطاف پذیری** یک مجموعه نمونه را می‌توان برای تخمین بسیاری از شاخص‌های آماری استفاده کرد. شامل: ضرب داخلی، فاصله l_α (برای هر مقداری از α)

۲-۴-۲ معایب نمونه‌گیری تصادفی مختصات

با این حال نمونه‌گیری تصادفی مختصات دو ایراد اساسی دارد.

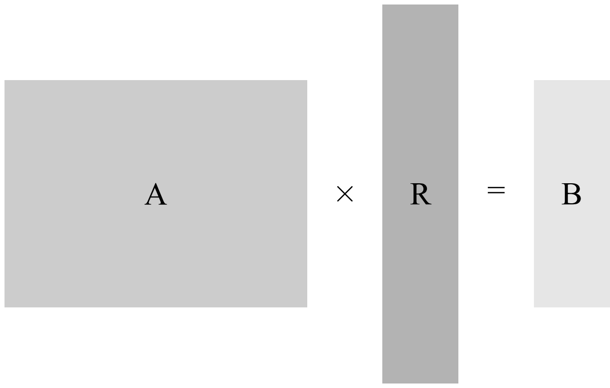
- معمولا دقیق نیست زیرا مقادیری با مقدار زیاد محتمل است که گم شوند. مخصوصا زمانی که داده‌ها دم سنگینی داشته باشند. داده‌های بزرگ مقیاس دنیای واقعی (مخصوصا داده‌های مربوط به اینترنت) همیشه دم سنگین هستند و از قاعده توانی پیروی می‌کنند. [۵۳، ۲۹، ۲۴، ۴۶] زمانی که فاصله l_2 یا ضرب داخلی را تخمین می‌زنیم. واریانس تخمین‌ها بر اساس ممان چهارم داده‌ها تعیین می‌شود. در حالی که در داده‌های دم سنگین، گاهی اوقات حتی ممان اول هم معنی‌دار نیست (محدود نیست) [۵۳] .
- این روش داده‌های پراکنده را به خوبی مدیریت نمی‌کند. بسیاری از داده‌های بزرگ مقیاس به شدت پراکنده هستند، به عنوان مثال، داده‌های متنی [۲۸] و داده‌های بر اساس بازار [۶، ۵۵] . به جز برخی واژه‌های کاربردی مانند "A" و "The" بیشتر لغات با نسبت بسیار کمی در مستندات ظاهر می‌شوند ($< 1\%$) اگر ما داده‌ها را با در نظر گرفتن تعدادی از ستون‌های ثابت نمونه‌گیری کنیم. خیلی محتمل است که بیشتر داده‌های (مقادیر غیر صفر) را از دست بدهیم.به خصوص موارد جذابی که دو مقدار با هم غیر صفر شده‌اند.

[21] memory wall

در این پایان‌نامه ما روش تصویر تصادفی را مورد بررسی قرار می‌دهیم و نشان خواهیم داد که این روش به خوبی قابلیت مدیریت داده‌های دم‌سنگین را دارد.

۲-۵ تصویر تصادفی پایدار

تصویر شکل ۲-۱ ، ایده تصویر تصادفی را نشان می‌دهد. ایده اصلی تصویر تصادفی ضرب ماتریس داده‌ی $\mathbf{A} \in \mathbb{R}^{n\times D}$ در ماتریس تصادفی $\mathbf{R} \in \mathbb{R}^{D\times k}(k \ll D)$ است. که حاصل ماتریس تصویر شده‌ی $\mathbf{B} = \mathbf{A} \times \mathbf{R} \in \mathbb{R}^{n\times k}$ است. $\mathbf{B}$ بسیار کوچکتر از $\mathbf{A}$ است و بنابراین به راحتی قابل ذخیره‌سازی است. (برای مثال: برای حافظه‌های فیزیکی به اندازه‌ی کافی کوچک است)



شکل ۲-۱: تصویر تصادفی پایدار $\mathbf{B} = \mathbf{A} \times \mathbf{R}$ ، $\mathbf{A}$ ماتریس اولیه داده‌ها است.

ماتریس تصویرگر $\mathbf{R} \in \mathbb{R}^{D\times k}$ معمولا از درایه‌های مستقل هم توزیع (i.i.d) یک توزیع متقارن α-پایدار پر شده است. [۵۹] (بنابراین نام این روش «تصویر تصادفی پایدار» است.) بر اساس مشخصات توزیع‌های α-پایدار، داده‌های تصویر شده هم از توزیع α-پایدار پیروی می‌کنند. که بر اساس آن‌ها شاخص‌های l_α و فاصله دودویی l_α در $\mathbf{A}$ تخمین زده می‌شوند و می‌توانیم داده‌های اصلی را دور بریزیم.

موفقیت تصویر پایدار تصادفی توسط لم Johnson-Lindenstrauss (JL) [۴۲] برای کاهش بعد در l_2نشان داده شده است. لم JL بیان می‌کند: رعایت $k = O\left(\frac{\log n}{\epsilon^2}\right)$ تضمین می‌کند هر فاصله l_2 میان n نقطه در هر تعداد بعدی با دقت $1 \pm \epsilon$ با احتمال بالایی تخمین زده شود. (k در اینجا بیانگر تعداد ابعاد کاهش یافته است)

با این حال لم JL برای نرم‌های فاصله با α کوچکتر از ۲ $l_\alpha(\alpha < 2)$ صادق نیست. در صورتی که لازم باشد از برآوردگرهایی استفاده کنیم که متریک باشند (در نامساوی مثلثی صدق کنند). به این

نتیجه «عدم امکان»[۲۲] گفته می‌شود. [۱۹، ۴۵، ۱۶] خوشبختانه شامل برآوردگرهایی که متریک نیستند نمی‌شود. در این پایان‌نامه ما در مورد برآوردگرهای کوناگونی که متریک نیستند صحبت خواهیم کرد. شامل: میانگین هندسی[۲۳] ، میانگین هارمونیک [۲۴] ، توان نسبی[۲۵] و همچنین حداکثر بزرگنمایی.

۲-۶ کاربردها

علاقه‌ی زیادی به تکنیک‌های نمونه برداری وجود دارد که در کاربردهای زیادی مورد استفاده قرار می‌گیرند. مانند: قانون وابستگی[۲۶] [۱۳، ۱۴] ، خوشه‌بندی، بهینه‌سازی درخواست[۲۷] [۴۹، ۲۱] ، تشخیص تکراری[۲۸] [۱۸، ۱۲] و بسیاری موارد دیگر. روش‌های نمونه بردار هر چه بیشتر و بیشتر برای مجموعه‌های بزرگتر اهمیت پیدا می‌کنند.

طرح برودر[۲۹] [۱۸] در ابتدا برای تشخیص صفحات وب تکراری معرفی شد. URL های زیادی به HTML های مشابه (یا تقریبا مشابه) اشاره می‌کنند. جواب‌های تخمین زده شده به اندازه‌ی کافی خوب بودند. نیازی نبود تا همه تکراری‌ها پیدا شوند ولی کاربردی بود که تعداد زیدی از آن‌ها پیدا شوند، بدون اینکه بیش از ارزش آن از توان محاسباتی استفاده شود.

در کاربردهای بازیابی اطلاعات (IR)[۳۰] معمولا گلوگاه حافظه‌ی فیزیکی است. زیرا مجموعه‌ی وب برای حافظه (RAM) بسیار بزرگ است و از طرفی ما می‌خواهیم زمان گشتن به دنبال داده‌ها بر روی دیسک را کمینه کنیم. زیرا زمان پاسخ به یک درخواست کلیدی است [۱۵] . به عنوان یک وسیله صرفه‌جویی در فضا، کاهش بعد یک ارائه فشرده از داده‌ها فراهم می‌کند که برای تولید جواب‌های تخمینی در حافظه فیزیکی مورد استفاده قرار می‌گیرند.

ما به بازدید صفحات وب اشاره کردیم. اگر ما یک عبارت جستجوی دو کلمه‌ای داشته باشیم، می‌خواهیم بدانیم چه تعداد از صفحات هر دو کلمه را دارند. فرض می‌کنیم محاسبه‌ی از قبل و نگهداری بازدید صفحات غیر ممکن باشد. حداقل نه برای کلماتی که تکرار زیادی ندارند و سری‌های چند کلمه‌ای. مرسوم است که در بازیابی اطلاعات با یک ماتریس بزرگ عبارت به ازای سند شروع کنیم که در آن

[22] impossibility
[23] geometric mean
[24] harmonic mean
[25] fractional power
[26] association rules
[27] query optimization
[28] duplicate detection
[29] Broder's sketch
[30] information retrieval

مقادیر ورودی نشان‌دهنده‌ی وجود عبارت در متن است. بنا به کاربردهای خاص می‌توانیم یک اندیس معکوس [۳۱] بسازیم و کلیتی از عبارات (برای تخمین ارتباط لغات) یا اسناد (برای تخمین شباهت اسناد) نگهداری کنیم.

۲-۶-۱ کاوش قوانین وابستگی

تحلیل‌های مبتنی بر بازار و قوانین وابستگی [۷، ۸، ۹] ابزارهای مناسبی برای کاوش پایگاه داده‌های تجاری هستند. پایگاه داده‌های تجاری دارند روز به روز بزرگتر و گسسته‌تر می‌شوند. [۶، ۵۵] الگوریتم‌های مختلف نمونه‌برداری پیشنهاد شده است. نمونه برداری این امکان را فراهم می‌کند تا قواعد تخصیص را به صورت آنلاین برآورد کنیم. که می‌تواند مزایایی در کاربردهای خاص داشته باشد.

۲-۶-۲ وابستگی جفتی همه (فاصله‌ها)

در کابردهای مختلفی شامل کلاسه‌بندی بر مبنای فاصله یا خوشه‌بندی و مدل‌سازی زبان با bi-gram[۲۳] ما نیازمند محاسبه‌ی همه‌ی جفت تخصیص‌ها (یا فاصله‌ها) هستیم. ماتریس داده‌ی $\mathbf{A}$ شامل n سطر و D ستون داده شده است. محاسبه‌ی مستقیم $\mathbf{AA}^T$، $O(n^2D)$ هزینه بر است. یا به طور بهینه‌تر $O(n^2\bar{f})$ که $\bar{f}$ تعداد میانگین مقادیر غیر صفر میان تمام سطرهای $\mathbf{A}$ است. محاسبه مستیم می‌تواند به شدت زمان‌بر باشد. همچنین، به طور خاص زمانی که ماتریس داده آنقدر بزرگ است که در حافظه فیزیکی جا نمی‌شود. محاسبه به طور خاص بسیار ناکارآمد خواهد بود.

۲-۶-۳ تخمین فاصله‌ها به طور آنلاین

در حالی که ماتریس داده‌ی اولیه $\mathbf{A} \in \mathbb{R}^{n\times D}$ ممکن است برای حافظه‌ی فیزیکی بسیار بزرگ باشد، نگهداری [۳۲] همه فاصله‌های جفتی و وابستگی‌ها در $\mathbf{A}$، $O(n^2)$ فضا مصرف می‌کند. که می‌تواند برای حافظه‌ی فیزیکی بسیار بزرگتر باشد. در این میان وابستگی‌های چندتایی را کنار می‌گزاریم. در بسیاری از کاربردها نظیر یادگیری برخط، سیستم‌های توصیه آنلاین، تحلیل‌های بازار برخط و موتورهای جستجو، بهتر است که نمونه‌ها (sketches) در حافظه نگهداری شوند و همه‌ی فاصله‌ها به طور آنلاین، زمانی که مورد نیاز باشد، محاسبه شوند.

[31] inverted index
[32] materializing

۲-۶-۴ بهینه‌سازی درخواست از پایگاه داده

در پایگاه داده‌ها یک وظیفه‌ی بسیار مهم تخمین join های چندراهی است، که تاثیر زیادی بر روی کارایی سیستم دارد [۳۵] . بر اساس تخمین دوراهی، سه‌راهی و حتی join هایی از مرتبه‌ی بالاتر، بهینه‌گرهای درخواست یک نقشه برای کمینه کردن تابع هزینه می‌سازند (برای مثال، نوشتن‌های میانی [۳۳]). بهینه بودن اهمیت بسیاری دارد زیرا مثلا نمی‌خواهیم زمان بیشتری برای بهینه‌سازی نقشه نسبت به زمان اجرای آن تلف کنیم.

ما از مثال Governator برای نمایش کاربرد تخمین دو و چند راهه برای بهینه کردن درخواست استفاده می‌کنیم.

جدول ۲-۴: بازدید صفحات گزارش شده توسط Google برای چهار کلمه و وابستگی‌های دو، سه و چهارتایی آن‌ها

	Query	Hits(Google)
One-way	Austria	88,200,000
	Governor	37,300,000
	Schwarzenegger	4,030,000
	Terminator	3,480,000
Two-way	Governor & Schwarzenegger	1,220,000
	Governor & Austria	708,000
	Schwarzenegger & Terminator	504,000
	Terminator & Austria	171,000
	Governor & Terminator	132,000
	Schwarzenegger & Austria	120,000
Tree-way	Governor & Schwarzenegger & Terminator	75,100
	Governor & Schwarzenegger & Austria	46,100
	Schwarzenegger & Terminator & Austria	16,000
	Governor & Terminator & Austria	11,500
Four-way	Governor & Schwarzenegger & Terminator & Austria	6,930

جدول ۲-۴ بازدید صفحات را برای چهار کلمه و ترکیبات دو، سه، چهارتایی آن‌ها نشان می‌دهد. فرض کنیم بهینه‌ساز قصد استخراج نقشه برای درخواست: "Governor, Schwarzenegger, Terminator, Austria" را داشته باشد. راه حل استاندارد این است که با عبارات با کمترین فراوانی شروع کند: $((\text{"Schwarzenegger"} \cap \text{"Terminator"}) \cap \text{"Governor"}) \cap \text{"Austria"}$ این نقشه $579,100$ نوشتن میانی بعد از اولین و دومین join ها دارد. یک بهینه‌سازی می‌تواند $((\text{"Schwarzenegger"} \cap \text{"Austria"}) \cap \text{"Terminator"}) \cap \text{"Governor"}$ باشد که $579,100$ را به $136,000$ کاهش می‌دهد.

[33] intermediate writes

۲-۶-۵ جستجوی نزدیکترین همسایه از مرتبه‌ی زیر خطی

محاسبه‌ی نزدیکترین همسایه در بسیاری کاربردها از اهمیت زیادی برخوردار است. با این حال، به دلیل «نفرین ابعاد»[۳۴] راه حل فعلی برای پیدا کردن بهینه‌ی نزدیکترین همسایه‌ها (حتی به طور تقریبی) اصلا رضایت بخش نیست. [۳۲، ۴۱]

به دلیل ملاحظات محاسباتی، دو شکل اصلی در جستجوی نزدیکترین همسایه‌ها وجود دارد. اول اینکه ماتریس اصلی داده‌ها $\mathbf{A} \in \mathbb{R}^{n\times D}$ ممکن است برای حافظه فیزیکی بسیار بزرگ باشد ولی اسکن کردن دیسک‌های سخت برای پیدا کردن نزدیکترین همسایه‌ها می‌تواند خیلی کند باشد. دوما، پیدا کردن نزدیکترین همسایه‌های یک داده ممکن است $O(nD)$ هزینه بر باشد که می‌تواند به شدت زمان بر شود.

با این حال، روس کاهش ابعادی در این پایان‌نامه می‌تواند در حافظه صرفه‌جویی کند و سرعت محاسبات را افزایش دهد. برای مثال: وقتی ماتریس داده‌ی اولیه $\mathbf{A}$ به ماتریس داده‌ی $\mathbf{B} \in \mathbb{R}^{n\times k}$ کاهش می‌یابد. با این حال، $O(nk)$ و معمولا این درخواست وجود دارد که هزینه‌ی محاسباتی از $O(n)$ به $O(n^\gamma)$ برای $\gamma < 1$ کاهش پیدا کند، حداقل برای کاربردهای خاص.

دو گروه اصلی الگوریتم‌های زیر خطی برای محاسبه عبارتند از KD-Trees (و انواع آن) [۳۳، ۳۴] و Locality-Sensitive Hashing (LSH) [۱۰، ۲۵، ۴۱] این الگوریتم‌ها معمولا با یک فضای متریک کار می‌کنند (که در ان نامساوی مثلثی برقرار است). برای مثال، فضای l_α زمانی که $\alpha \geq 1$ باشد یک متریک است. زمانی که به دنبال نزدیکترین همسایه‌ها در l_α ($\alpha > 1$) میگردیم، می‌توانیم (نسبتا به سادگی) فضای جستجو را به طور کاملا اساسی با استفاده از نامساوی مثلثی کاهش دهیم. به عبارت دیگر، نیازی نیست که همه n نقطه داده‌ها را مورد بررسی قرار دهیم.

در داده‌هایی با ابعاد بسیار بزرگ، الگوریتم‌های زیر خطی موجود شامل KD-trees و LSH ، عملکرد رضایت بخشی ندارند. وقتی حافظه‌ی فیزیکی (به جای CPU) گلوگاه باشد[۳۵] ، یکی از مشکلات اصلی این است که این الگوریتم‌ها برای کاهش هزینه‌ی محاسباتی به حافظه‌ی ابر خطی[۳۶] نیاز دارند که می‌تواند مشکل ساز باشد. [۴۱] به طرح کلی برای LSH توجه کنید که ترکیبی از هش[۳۷] و تصویر تصادفی است. متاسفانه این طرح به دلیل هزینه‌ی زیاد پیش پردازش غیر کاربردی است. [۴۱]

در این پایان‌نامه، موفقیت اصلی کاهش بعد داده $\mathbf{A} \in \mathbb{R}^{n\times D}$ به $\mathbf{B} \in \mathbb{R}^{n\times k}$ و تامین برآوردگرهای

[34] curse of dimensionality
[35] memory wall
[36] super-linear memory
[37] hash

دقیق برای استخراج فاصله‌های اولیه در $\mathbf{A}$ بر اساس $\mathbf{B}$ است. در حالی که سناریوهای مهمی وجود دارند که در آن‌ها نتایج ما رضایت بخش هستند، توسعه‌ی یک الگوریتم زیر-خطی برای تخمین نزدیکترین همسایه‌ها، بر اساس الگوریتم ما یک ایده جذاب برای تحقیقات آینده است. یک مانع اصلی در این راه این است که بیشتر برآوردگرهای ما غیر متریک هستند و بنابراین طراحی یک الگوریتم هوشمند و تحلیل‌های تئوری ممکن است سخت باشد، با این حال غیر ممکن نیست.

فصل سوم
تصویر تصادفی پایدار

روش تصویر تصادفی پایدار(زیرنویس ۱) [۴۸، ۵۶، ۱۱، ۴۰، ۳۹، ۴۳] یک روش پرکاربرد در داده‌کاوی و یادگیری ماشین است. با این روش به طور کارا فاصله $l_\alpha (0 < \alpha \leq 2)$ در داده‌های حجیم (برای مثال: وب یا جریان‌های داده‌ی حجیم) محاسبه می‌شود. در این روش حافظه‌ی کمی استفاده شده و فقط یک بار پایش داده‌ها کافی است.

همانطور که در شکل ۲-۱ می‌بینید. ایده تصویر تصادفی پایدار، ضرب ماتریس داده‌ها $\mathbf{A} \in \mathbb{R}^{n\times D}$ در ماتریس تصادفی $\mathbf{R} \in \mathbb{R}^{D\times k} (k \ll D)$ است که حاصل یک ماتریس تصویر شده‌ی $\mathbf{B} \in \mathbb{R}^{n\times k}$ است. درایه‌های ماتریس تصادفی $mathbfR$ به طور .i.i.d (مستقل و هم توزیع) از یک توزیع α -پایدار [۱] حاصل می‌شوند. به همین دلیل به این روش «تصویر تصادفی پایدار» گفته می‌شود. به این نکته توجه کنید که توزیع ۲-پایدار معادل توزیع نرمال و توزیع پایدار ۱-پایدار معادل کوچی[۲] است.

حالت خاص تصویر تصادفی نرمال (به عبارت دیگر $\alpha = 2$) نسبتا به خوبی مورد بررسی قرار گرفته است. به رساله [۵۶] مراجعه کنید. بنابراین، بخش اعظم این پایان‌نامه به تصویر تصادفی پایدار $\alpha < 2$ اختصاص یافته است.

پس از مروری بر حالت کلی تصویر تصادفی پایدار $0 < \alpha \leq 2$ ، جزئیات بیشتری در خصوص حالت l_2 مورد بررسی قرار می‌گیرد. سپس ارتقاء روش با استفاده از اطلاعات حاشیه‌ای[۳] بررسی می‌شود. در ادامه، تصویر تصادفی نرمال ساده‌سازی می‌شود. این کار با نمونه‌برداری $\mathbf{R}$ از حالت توزیع گسسته‌ی سه‌نقطه‌ای $[-1, 0, 1]$ انجام می‌شود. این حالت، یک حالت خاص توزیع‌های زیرگوسی[۴] است. سپس نرم l_1[۵] مورد بررسی قرار گرفته و در ادامه حالت کلی $0 < \alpha \leq 2$ مورد بحث قرار می‌گیرد.

۳-۱ مسئله‌ی اصلی در تصویر تصادفی پایدار

مسئله اصلی تصویر تصادفی پایدار یک مسئله‌ی برآورد آماری است. همانطور که بیان شد، ماتریس داده‌ی $\mathbf{A} \in \mathbb{R}^{n\times D}$ را در ماتریس تصادفی $\mathbf{R} \in \mathbb{R}^{D\times k}$ ضرب می‌کنیم تا ماتریس بسیار کوچکتر $\mathbf{B} = \mathbf{A} \times \mathbf{R} \in \mathbb{R}^{n\times k}$ را بدست بیاوریم. هدف این است که مشخصات آماری $\mathbf{A}$ بر اساس ماتریس $\mathbf{B}$ استنتاج شوند. (شامل نرم و فاصله)

بدون از دست دادن کلیت، ما بر ۲ سطر اول $\mathbf{A}$، $u_1, u_2 \in \mathbb{R}^D$ و دو سطر اول در $\mathbf{B}$، $v_1, v_2 \in \mathbb{R}^k$

[1] α-stable distribution
[2] Cauchy
[3] marginal information
[4] sub-Gaussian
[5] Cauchy random projection

تمرکز می‌کنیم. تعریف می‌کنیم $\mathbf{R} = \{r_{ij}\}_{i=1}^{D}\,_{j=1}^{k}$ بنابراین:

$$v_{1,j} = \sum_{i=1}^{D} r_{ij} u_{1,i}, \;\; v_{2,j} = \sum_{i=1}^{D} r_{ij} u_{2,i}, \;\; x_j = v_{1,j} - v_{2,j} = \sum_{i=1}^{D} r_{ij}(u_{1,i} - u_{2,i}). \tag{۳-۱}$$

۳-۱-۱ توزیع‌های پایدار

به طور معمول $r_{ij} \sim S(\alpha, 1)$ و به طور .i.i.d استخراج می‌شود. همچنین در ادامه ما حالت‌های ساده‌تری را هم مورد بررسی قرار می‌دهیم. در اینجا $S(\alpha, 1)$ بیانگر یک توزیع متقارن α-پایدار تصادفی است [۵۹] با پارامتر اندیس α و پارامتر مقیاس ۱.

یک متغییر تصادفی z در صورتی متقارن α-پایدار است که تابع مشخصه‌ی آن به شکل زیر باشد.

$$E\big(\exp\big(\sqrt{-1}zt\big)\big) = \exp\big(-d|t|^{\alpha}\big) \tag{۳-۲}$$

که $d > 0$ پارامتر مقیاس است. ما می‌نویسیم $z \sim S(\alpha, d)$ که به طور کلی شکل بسته‌ای برای تابع چگالی ندارد. به جز حالت $\alpha = 2$ (نرمال) و $\alpha = 1$ (کوچی).

۳-۱-۲ مسئله برآورد آماری

با توجه به خواص تبدیل فوریه، به راحتی می‌توان نشان داد که داده‌های تصویر شده هم از توزیع α-پایدار پیروی می‌کنند که در این حالت پارامتر مقیاس مشخصه‌ی l_α ی (نرم‌ها، فاصله‌ها) داده‌های اصلی در $\mathbf{A}$ است. به طور خاص:

$$v_{1,j} \sim S\left(\alpha, \sum_{i=1}^{D} |u_{1,i}|^{\alpha}\right), \;\; v_{2,j} \sim S\left(\alpha, \sum_{i=1}^{D} |u_{2,i}|^{\alpha}\right), \tag{۳-۳}$$

$$x_j = v_{1,j} - v_{2,j} \sim S\left(\alpha, d_{(\alpha)} = \sum_{i=1}^{D} |u_{1,i} - u_{2,i}|^{\alpha}\right). \tag{۳-۴}$$

بنابراین، کار ما به برآورد پارامتر مقیاس از k نمونه .i.i.d ، $x_j \sim S\big(\alpha, d_{(\alpha)}\big)$ تقلیل پیدا می‌کند. به این خاطر که هیچ شکل بسته‌ای برای تابع چگالی به جز در حالت $\alpha = 1, 2$ وجود ندارد، فرآیند تخمین خود مسئله‌ی جالبی است اگر به دنبال برآوردگرهایی بگردیم که هم به طور آماری دقیق باشند و هم از لحاظ محاسباتی کارا باشند.

یک موضوع مربوط و نزدیک هم تعیین اندازه نمونه k است. روش استاندارد محدود کردن احتمال دم است $\mathbf{Pr}\big(|\hat{d}_{(\alpha)} - d_{(\alpha)}| > \epsilon d_{(\alpha)}\big)$ که $\hat{d}_{(\alpha)}$ برآوردگری برای $d_{(\alpha)}$ است و ϵ دقت مورد نظر است (معمولا $0 < \epsilon < 1$). به طور ایده‌آل امیدوار هستیم نشان دهیم [۶] :

$$\mathbf{Pr}\big(|\hat{d}_{(\alpha)} - d_{(\alpha)}| > \epsilon d_{(\alpha)}\big) \leq 2\exp\left(-k\frac{\epsilon^2}{G}\right), \tag{۳-۵}$$

برای برخی مقادیر ثابت G که می‌تواند تابعی از ϵ هم باشد.

برای ماتریس داده‌ی $\mathbf{A} \in \mathbb{R}^{n\times D}$ ، در مجموع $\frac{n^2}{2} > \frac{n(n-1)}{2}$ جفت فاصله وجود دارد. ما معمولا علاقمندیم که احتمالات دم را به طور همزمان برای همه‌ی جفت‌ها محدود کنیم.

۳-۲ تصویر تصادفی نرمال

برای کاهش بعد در نرم l_2 ، روش تصویر تصادفی نرمال ماتریس داده‌ی اولیه $\mathbf{A} \in \mathbb{R}^{n\times D}$ را در ماتریس تصادفی $\mathbf{R} \in \mathbb{R}^{D\times k} (k \ll D)$ با درایه‌های i.i.d. از $N(0,1)$ ، تا ماتریس تصویر شده‌ی $\mathbf{B} \in \mathbb{R}^{n\times k}$ حاصل شود. تحلیل‌های مربوط به تصویر تصادفی نرمال نسبتا ساده است. برای مثال، به شکل سرراستی می‌توان یک نسخه از لم JL [۷] [۴۲] را برای حالت l_2 استنتاج کرد.

ما در ابتدا برخی خواص اولیه تصویر تصادفی نرمال را بیان می‌کنیم و سپس بر روی اطلاعات حاشیه تمرکز می‌کنیم تا تخمین‌ها را بهینه کنیم. حاشیه‌ها (به عبارت دیگر، نرم l_2 برای هر خط در $\mathbf{A}$) معمولا در ابتدا در دسترس هستند (برای مثال، از طریق نرمال سازی داده‌ها). ولی حتی در حالتی که در دسترس نیستند، محاسبه‌ی نرم l_2 برای تمام سطرهای $\mathbf{A}$ فقط نیازمند یکبار مرور داده‌ها است که هزینه‌ای از $O(nD)$ دارد که قابل صرفنظر است. [۸] از آنجا که اعمال تصویر تصادفی $\mathbf{A} \times \mathbf{R}$ هم اکنون

[۶] بنابر قضیه حدمرکزی برآوردگر $\hat{d}_{(\alpha)}$ بر اساس k نمونه تحت شروط ساده‌ای به حالت نرمال همگرا می‌شود. بنابر محدوده‌ی دم نرمال می‌دانیم که حداقل برای پارامترهای خاصی $\mathbf{Pr}\big(|\hat{d}_{(\alpha)} - d_{(\alpha)}| \geq \epsilon d_{(\alpha)}\big) \leq 2\exp\left(-k\frac{\epsilon^2}{2V}\right)$ باید صادق باشد. در اینجا $\frac{V}{k}$ واریانس مجانبی $\hat{d}_{(\alpha)}$ است. بنابراین، حداقل برای آزمون درستی، می‌توانیم با بررسی این‌که آیا $\lim_{\epsilon\to 0+} G = 2V$ چک کنیم که محدوده‌ی دم نسبت مطلوب را دارا باشد.

[7] Johnson-Lindenstrauss

[۸] این وضعیتی برای زمانی که با جریان داده‌های دینامیک سر و کار داریم اندکی متفاوت است. در جریان‌های داده ما معمولا به دنبال اطلاعات آماری یک جریان داده هستیم تا اختلاف میان دو جریان داده را مد نظر داشته باشیم. به عبارت دیگر، محاسبه نرم l_2 حاشیه‌ای گاهی اوقات هدف اصلی است. به دلیل ذات دینامیک جریان‌های داده (برای مثال، به روز شدن مدام)، محاسبه‌ی حاشیه‌ها می‌تواند پر هزینه باشد.

هزینه‌ای از مرتبه $O(nDk)$ دارد.

در این بخش، ما این قاعده مرسوم تبعیت در ادبیات تصویر تصادفی [۵۶] پیروی می‌کنیم و تعریف می‌کنیم $\mathbf{B} = \frac{1}{\sqrt{k}}\mathbf{AR}$.

۳-۲-۱ مشخصه‌های اصلی

ما فرض می‌کنیم یک ماتریس داده $\mathbf{A} \in \mathbb{R}^{n\times D}$ و یک ماتریس تصویرگر $\mathbf{R} \in \mathbb{R}^{D\times k}$ که به طور i.i.d. از $N(0,1)$ تولید شده است. در نظر می‌گیریم $\mathbf{B} = \frac{1}{\sqrt{k}}\mathbf{AR}$. در نظر بگیرید u_i^T سطر i ام ماتریس $\mathbf{A}$ باشد، و سطر متناظر در $\mathbf{B}$ ، v_i^T باشد. برای راحتی بر روی دو سطر اول $\mathbf{A}$ یعنی u_1 و u_2 همچنین دو سطر اولیه v_1 و v_2 در $\mathbf{B}$ تمرکز می‌کنیم. تعریف می‌کنیم:

$$a = u_1^T u_2,\ m_1 = \|u_1\|^2,\ m_2 = \|u_2\|^2,\ d = \|u_1 - u_2\|^2 = m_1 + m_2 - 2a \qquad (۳-۶)$$

به آسانی می‌توانیم نشان دهیم $\|v_1 - v_2\|$ ، فاصله‌ی l_2 نمونه و $v_1^T v_2$ ضرب داخلی نمونه، برآوردگرهای نااریبی از d و a هستند. لم ۱ واریانس و تابع مشخصه‌ی $v_1^T v_2$ را مشخص می‌کند. اثبات در [۴۷] .

لم ۱: $u_1, u_2 \in \mathbb{R}^D$ داده شده‌اند و یک ماتریس تصادفی $\mathbf{R} \in \mathbb{R}^{D\times k}$ شامل درایه‌های i.i.d. از نرمال استاندارد $N(0,1)$. اگر مقادیر $v_1 = \frac{1}{\sqrt{k}}\mathbf{R}^T u_1$ و $v_2 = \frac{1}{\sqrt{k}}\mathbf{R}^T u_2$ را تعیین کنیم، داریم:

$$E\left(\|v_1 - v_2\|^2\right) = d, \quad Var\left(\|v_1 - v_2\|^2\right) = \frac{2}{k}d^2 \qquad (۳-۷)$$

$$E\left(v_1^T v_2\right) = a, \quad Var\left(v_1^T v_2\right) = \frac{1}{k}\left(m_1 m_2 + a^2\right), \qquad (۳-۸)$$

سومین ممان مرکزی $v_1^T v_2$ عبارت است از:

$$E\left(v_1^T v_2\right)^2 = a, \quad \frac{2a}{k^2}\left(2m_1 m_2 + a^2\right) \qquad (۳-۹)$$

و تابع مولد احتمال برای $v_1^T v_2$ عبارت است از:

$$E\big(\exp\left(v_1^T v_2 t\right)\big) = \left(1 - \frac{2}{k}at - \frac{1}{k^2}(m_1m_2 - a^2)t^2\right)^{-\frac{k}{2}} \qquad (۱۰-۳)$$

که $\frac{-k}{\sqrt{m_1m_2}+a} \le t \le \frac{-k}{\sqrt{m_1m_2}-a}$ است.

بنابراین، برآوردگرهای نااریبی برای فاصله d l_2 و ضرب داخلی a به شکل سر راستی عبارت است از:

$$\hat{d}_{MF} = \|v_1 - v_2\|^2, \quad Var\left(\hat{d}_{MF}\right) = \frac{d^2}{k}, \qquad (۱۱-۳)$$

$$\hat{a}_{MF} = v_1^T v_2, \quad Var\left(\hat{a}_{MF}\right) = \frac{1}{k}\left(m_1m_2 + a^2\right), \qquad (۱۲-۳)$$

که اندیس « MF » به معنی «بدون حاشیه»[۹] نشان دهنده این است که برآوردگرها از اطلاعات حاشیه‌ای $m_1 = \|u_1\|^2$ و $m_2 = \|u_2\|^2$ استفاده نمی‌کنند.

به این نکته توجه کنید که، $k\hat{d}_{MF}/d$ از توزیع χ^2 با k درجه آزادی، پیروی می‌کند، χ_k^2 . بنابراین، به راحتی می‌توان می‌توانیم این محدوده‌های دم را برای لم ۲ اثبات کنیم.

لم ۲:

$$\mathbf{Pr}\left(\hat{d}_{MF} - d > \epsilon d\right) \le \exp\left(-\frac{k}{2}\left(\epsilon - \log(1+\epsilon)\right)\right), \quad \epsilon > 0 \qquad (۱۳-۳)$$

$$\mathbf{Pr}\left(\hat{d}_{MF} - d < -\epsilon d\right) \le \exp\left(-\frac{k}{2}\left(-\epsilon - \log(1-\epsilon)\right)\right), \quad 0 < \epsilon < 1 \qquad (۱۴-۳)$$

اثبات:

از آنجا که $k\hat{d}_{MF}/d \sim \chi_k^2$ ، بر اساس نام مساوی چرنوف [۱۰] [۲۲] ، برای هر $t > 0$ داریم:

[9] margin-free
[10] Chernoff inequality

$$\begin{aligned} &\mathbf{Pr}\big(\hat{d}_{MF} - d > \epsilon d\big) = \mathbf{Pr}\big(k\hat{d}_{MF}/d > k(1+\epsilon)\big) \\ &\le \frac{E\Big(\exp(k\hat{d}_{MF}/dt\Big)}{\exp\big((1+\epsilon)kt\big)} = \exp\Bigg(-\frac{k}{2}\big(\log(1-2t) + 2(1+\epsilon)t\big)\Bigg) \end{aligned} \tag{۱۵-۳}$$

که در $t = t_{NR} = \frac{\epsilon}{2(1+\epsilon)}$ و بنابراین برای هر $\epsilon > 0$ داریم:

$$\mathbf{Pr}\big(\hat{d}_{MF} - d > \epsilon d\big) \le \exp\left(-\frac{k}{2}\left(\epsilon - \log\left(1+\epsilon\right)\right)\right) \tag{۱۶-۳}$$

ما می‌توانیم به طور مشابه برای دیگر محدوده‌ی دم $\mathbf{Pr}\big(\hat{d}_{MF} - d < -\epsilon d\big)$ هم اثبات کنیم. ■

برای راحتی مرسوم است که محدوده دم را در لم ۲ به صورت متقارن $\mathbf{Pr}\left(\left|\hat{d}_{MF} - d\right| > \epsilon d\right)$ نوشته شود. نامساوی‌های ساده‌ای برای $\log(1+\epsilon)$ و $\log(1-\epsilon)$ نتیجه می‌دهد:

$$\mathbf{Pr}\left(\left|\hat{d}_{MF} - d\right| \ge \epsilon d\right) \le 2\exp\left(-\frac{k}{4}\epsilon^2 + \frac{k}{6}\epsilon^3\right), \quad 0 < \epsilon < 1 \tag{۱۷-۳}$$

از آنجا که $\mathbf{A} \in \mathbb{R}^{n\times D}$ تعداد n سطر دارد. به عبارت دیگر $\frac{n(n-1)}{2}$ جفت. ما باید احتمال دم را به طور همزمان برای همه‌ی جفت‌ها محدود کنیم. با استفاده از محدوده تجمیعی بنفرونی[11] کافی است که:

$$\frac{n^2}{2}\mathbf{Pr}\left(\left|\hat{d}_{MF} - d\right| \ge \epsilon d\right) \le \delta \tag{۱۸-۳}$$

به عبارت دیگر کافی است اگر:

$$\frac{n^2}{2} 2\exp\left(-\frac{k}{4}\epsilon^2 + \frac{k}{6}\epsilon^3\right) \le \delta \Rightarrow k \ge \frac{2\log n - \log\delta}{\epsilon^2/4 - \epsilon^3/6} \tag{۱۹-۳}$$

[11] Benferroni union bound

بنابراین ما یک نسخه‌ای از لم JL را نشان داده‌ایم.

لم ۳: اگر $k \geq \frac{2\log n - \log\delta}{\epsilon^2/4 - \epsilon^3/6}$ پس با حداقل احتمال $1-\delta$ ، فاصله l_2 بین هر جفت از داده‌ها (میان n نقطه) می‌تواند با ضریب اطمینان $1 \pm \epsilon$ با استفاده فاصله‌ی l_2 در داده‌های تصویر شده بعد از تصویر تصافی نرمال، تخمین زده شود. $0 < \delta < 1, 0 < \epsilon < 1$. ■

۳-۳ تصویر تصادفی زیر گوسی و بسیار پراکنده

در بخش قبل ما به بررسی تصویر تصادفی نرمال پرداختیم، که در آن ماتریس تصویرگر $\mathbf{R}$ از روی توزیع $N(0,1)$ به طور .i.i.d نمونه‌گیری می‌شود. این انتخاب خاص برای $\mathbf{R}$ ، صرفا برای سهولت تحلیل تئوری است. در واقع می‌توان $\mathbf{R}$ را از هر توزیعی با میانگین صفر و واریانس محدود برای کاهش بعد در نرم l_2 نمونه‌گیری کرد.

نمونه‌گیری $\mathbf{R}$ از یک توزیع زیرگوسی هم از نظر تئوری قابل قبول و هم از جنبه‌ی محاسباتی تسهیل کننده است. برای مثال، محدوده‌ی دم زیر گوسی به سادگی به نسخه‌ای از لم JL منتهی می‌شود.

ما بر روی یک انتخاب معمول از توزیع زیر گوسی تمرکز خواهیم کرد، که درایه‌ها ماتریس $\mathbf{R}$ از مجموعه‌ی $\{-1,0,1\}$ با احتمالات $\{\frac{1}{2s}, 1-\frac{1}{s}, \frac{1}{2s}\}$ ، که $s \geq 1$. به این ترتیب فرآیند نمونه‌گیری ساده‌تر شده و محاسبات سریعتر انجام می‌شوند. در واقع، زمانی که $s < 3$ باشد، واریانس‌های صرحا کوچکتری نسبت به استفاده از تصویر تصادفی نرمال بدست می‌آید.

با در نظر گرفتن قواعد معقول، برای مثال، داده‌های اولیه ممان سوم محدود داشته باشند،می‌توانیم $s \gg 3$ در نظر بگیریم (حتی $s = \sqrt{D}$). تا نتایج s برابر سریعتر بدست بیاوریم؛ و بنابراین، این رویه را تصویر تصادفی بسیار پراکنده می‌نامیم.

۳-۳-۱ تصویر تصادفی زیرگوسی

مشابه قسمت ۳-۲ ماتریس داده را $\mathbf{A} \in \mathbb{R}^{n\times D}$ در نظر می‌گیریم. ماتریس تصویر تصادفی $\mathbf{R} \in \mathbb{R}^{D\times k}$ را تولید کرده و آن را در $\mathbf{A}$ ضرب می‌کنیم تا به یک ماتریس تصویر شده‌ی $\mathbf{B} = \frac{1}{\sqrt{k}}\mathbf{AR} \in \mathbb{R}^{n\times k}$ برسیم. دوباره رو دو ردیف ابتدایی تمرکز می‌کنیم، که یعنی u_1 و u_2 در $\mathbf{A}$ ، و دو ردیف ابتدایی v_1 و v_2 در $\mathbf{B}$ و همچنین تساوی‌های زیر را تعریف می‌کنیم:

$$a = u_1^T u_2, \quad m_1 = \|u_1\|^2, \quad m_2 = \|u_2\|^2, \quad d = \|u_1 - u_2\|^2 = m_1 + m_2 - 2a \qquad (۳-۲۰)$$

$\mathbf{R}$ را به طور i.i.d از یک توزیع زیر گوسی مشخصا پرکاربرد تولید می‌کنیم:($S \geq 1$)

$$r_{ij} = \sqrt{s} \times \begin{cases} 1 & \text{با احتمال } \frac{1}{2s} \\ 0 & \text{با احتمال } 1 - \frac{1}{s} \\ -1 & \text{با احتمال } \frac{1}{2s} \end{cases} \qquad (۳-۲۱)$$

- نمونه‌گیری از معادله ۳-۲۱ ساده‌تر از نمونه‌گیری از $N(0,1)$ است.
- می‌تواند از s برابر افزایش سرعت در ضرب ماتریسی $\mathbf{A} \times \mathbf{R}$ بهره برد، زیرا فقط $\frac{1}{s}$ داده‌های نیازمند پردازش هستند.
- نیازی به عملیات محاسباتی با ممیز شناور نیست و تمامی بار محاسباتی بر روی عملیات تجمیع پایگاه داده است که به خوبی بهینه شده.
- وقتی $s < 3$ باشد می‌توان به تخمین‌هایی با دقت بیشتر (واریانس کمتر) دست پیدا کرد.
- هزینه نگهداری ماتریس $\mathbf{R}$ از $O(Dk)$ به $O(Dk/s)$ کاهش می‌یابد.

[۳، ۴] نشان می‌دهند زمانی که $s = 1$ و $s = 3$ باشد، می‌توان به همان محدوده‌ی JL ای دست پیدا کرد که در تصویر تصادفی نرمال وجود دارد. ما در ادامه به بررسی خواص توزیع زیر گوسی می‌پردازیم، که برای تحلیل محدوده‌ی دم مناسب است. در واقع، آنالیز زیر گوسی نشان می‌دهد که می‌توان حتی در بدترین شرایط از مقادیری اندکی بیشتر از ۳ برای s استفاده کرد.

توزیع زیر گوسی

ما در اینجا مقدمه‌ای کوتاه بر توزیع‌های زیرگوسی بیان می‌کنیم. برای جزئیات و منابع بیشتر می‌توانید به [۲۰] مراجعه کنید. تئوری توزیع‌های زیرگوسی در حدود ۱۹۶۰ آغاز شد.

متغییر تصادفی x زیرگوسی است اگر ثابت $g > 0$ وجود داشته باشد به شکلی که:

$$\mathrm{E}(\exp(xt)) \leq \exp\left(\frac{g^2t^2}{2}\right), \forall t \in \mathbb{R} \tag{۳-۲۲}$$

می‌توان مقدار بهینه‌ی g^2 را از تعریف $T^2(x)$ با استفاده از فرمول زیر بدست آورد.

$$T^2(x) = \sup_{t\neq 0} \frac{2\log \mathrm{E}\left(\exp(xt)\right)}{t^2} \tag{۳-۲۳}$$

توجه کنید که $T^2(x)$ فقط یک نمادگذاری برای مقدار ثابت بهینه‌ی زیرگوسی یک متغیر تصادفی x است (و نه یک نمونه مشخص از x).

برخی از ویژگی‌های اولیه‌ی توزیع‌های زیرگوسی:

- اگر x زیر گوسی باشد آنگاه $\mathrm{E}(x) = 0$ و $\mathrm{E}(x^2) \leq T^2(x)$. برای هر مقدار ثابت c ، $T^2(cx) = c^2T^2(x)$. و

$$\mathbf{Pr}(|x| > t) \leq 2\exp\left(-\frac{t^2}{2T^2(x)}\right) \tag{۳-۲۴}$$

- اگر $x_1, x_2, \ldots, x_D$ زیرگوسی مستقل باشند، آنگاه $\sum_{i=1}^{D} x_i$ زیرگوسی است.

$$T^2\left(\sum_{i=1}^{D} x_i\right) \leq \sum_{i=1}^{D} T^2(x_i) \tag{۳-۲۵}$$

- اگر x زیرگوسی باشد، آنگاه برای همه‌ی $t \in [0,1)$ ،

$$\mathbf{E}\left(\exp\left(\frac{x^2t}{2T^2(x)}\right)\right) \leq (1-t)^{-\frac{1}{2}} \tag{۳-۲۶}$$

[۳، ۴] همچنین معادله ۳-۲۶ را برای توزیع ویژه‌ی معادله ۳-۲۱ بدست آورده‌اند. یک متغیر تصادفی زیرگوسی x صریحا زیرگوسی است اگر $\mathrm{E}(x^2) = T^2(x)$

- اگر x صریحا زیر گوسی باشد، آنگاه $\mathrm{E}(x^3) = 0$ و کشیدگی [۱۲] غیر مثبت خواهد بود، به عبارت دیگر $\frac{\mathrm{E}(x^4)}{\mathrm{E}^2(x^2)} - 3 \leq 0$.
- اگر $x_1, x_2, \ldots, x_D$ صریحا زیرگوسی مستقل باشند، آنگاه $\sum_{i=1}^{D} x_i$ صریحا زیرگوسی است.

$$T^2\left(\sum_{i=1}^{D} x_i\right) = \sum_{i=1}^{D} T^2(x_i) = \sum_{i=1}^{D} \mathbf{E}\left(x_i^2\right) \tag{۳-۲۷}$$

■

۳-۴ تصویر تصادفی کوچی برای l_1

در بخش‌های قبلی به تصویر تصادفی برای کاهش بعد در نرم l_2 پرداخته شد. در این بخش به کاهش بعد در نرم l_1 پرداخته خواهد شد.

در اینجا هم با یک ماتریس داده‌ی $\mathbf{A} \in \mathbb{R}^{n\times D}$ کار خواهیم کرد. و یک ماتریس تصویرگر تصادفی $\mathbf{R} \in \mathbb{R}^{D\times k}$ که به طور .i.i.d از توزیع کوچی استاندارد $C(0, 1)$ نمونه‌گیری شده است، تولید خواهیم کرد. ما اجازه خواهیم داد که ماتریس تصویرشده $\mathbf{B} = \mathbf{A} \times \mathbf{R} \in \mathbb{R}^{n\times k}$ باشد. بدون آنکه ضریب نرمال‌سازی $\frac{1}{\sqrt{k}}$ که در بخش‌های قبلی مشاهده کردیم، حضور داشته باشد. ضمن آنکه این کار به یک تخمین آماری منجر خواهد شد که پارامتر مقیاس‌دهی را از تعداد k متغییر تصادفی کوچی به طور .i.i.d برآورد می‌کند.

از آنجا که کوچی میانگین محدود ندارد. نمی‌توانیم از یک برآوردگر خطی آنطور که در تصویر تصادفی نرمال استفاده کردیم، استفاده کنیم. علاوه بر این، نتیجه‌ی عدم امکان بیان شده در [۱۷، ۴۵، ۱۶] اثبات کرده است که وقتی از یک تصویرگر خطی استفاده شود، نمی‌توان از برآوردگرهای خطی بدون رخ دادن خطاهای بزرگ استفاده کرد. به عبارت دیگر، لم JL برای l_1 صدق نمی‌کند.

در این بخش سه برآوردگر غیرخطی ارائه و یک معادل برای لم JL برای l_1 استنتاج می‌شود. از آنجا که برآوردگرهای ما، متریک نیستند، این معادل لم JL از حالت کلاسیک لم JL برای l_2 ضعیفتر است.

[12] kurtosis

۳-۴-۱ خلاصه نتایج اصلی

ما دوباره مانند بخش‌های قبلی دو سطر اول A ، u_1 و u_2 و دو سطر اول B ، v_1 و v_2 را در نظر می‌گیریم. فاصله‌ی l_1 را با $d = \sum_{i=1}^{D} |u_{1,i} - u_{2,i}|$ تعریف می‌کنیم.

در تصویر تصادفی کوچی، فعالیت اصلی آن است که پارامتر مقیاس‌دهی کوچی از k نمونه‌ی $x_j \sim C(0,d)$ به طور .i.i.d استخراج شود. برخلاف تصویر تصادفی نرمال، نمی‌توان d را از میانگین نمونه برآورد کرد (به عبارت دیگر، $\frac{1}{k}\sum_{j=1}^{k}|x_j|$ زیرا $\mathrm{E}(x_j) = \infty$) .

سه نوع برآوردگر غیر خطی مورد بررسی قرار خواهند گرفت: برآوردگرهای میانه‌ی نمونه، برآوردگرهای میانگین هندسی و برآوردگرهای حداکثر درستنمایی.

- **برآوردگرهای میانه نمونه**

برآوردگر میانه‌ی نمونه $\hat{d}_{me}$ و نسخه‌ی بدون انحراف $\hat{d}_{me,c}$ به شکل زیر هستند.

$$\hat{d}_{me} = \text{median}\left(|x_j|, j = 1, 2, \ldots, k\right) \tag{۳-۲۸}$$

$$\hat{d}_{me,c} = \frac{\hat{d}_{me}}{b_{me}} \tag{۳-۲۹}$$

$$b_{me} = \int_0^1 \frac{(2m+1)!}{(m!)^2} \tan\left(\frac{\pi}{2}t\right)\left(t - t^2\right)^2 dt, \quad k = 2m+1 \tag{۳-۳۰}$$

برای سهولت، ما فقط $k = 2m+1, m = 1, 2, \ldots$ را در نظر می‌گیریم.

در بین تمامی برآوردگرهای چندکی، $\hat{d}_{me}$ (و $\hat{d}_{me,c}$) کوچکترین مقدار واریانس مجانبی را بدست می‌دهد.

- **برآوردگرهای میانگین هندسی**

برآوردگر میانگین هندسی، $\hat{d}_{gm}$ و نسخه‌ی بدون انحراف $\hat{d}_{gm,c}$ به شکل زیر هستند:

$$\hat{d}_{gm} = \prod_{j=1}^{k} |x_j|^{1/k} \tag{۳-۳۱}$$

$$\hat{d}_{gm,c} = \cos^k\left(\frac{\pi}{2k}\right) \prod_{j=1}^{k} |x_j|^{1/k} \tag{۳-۳۲}$$

از نظر واریانس‌های مجانبی، برآوردگرهای میانگین هندسی به صورت مجانبی متناظر با برآوردگرهای میانه‌ی نمونه هستند. اگر چه از نظر محدوده‌ی دم، برآوردگرهای میانه‌ی نمونه ممکن است نیازمند نمونه‌ای به اندازه‌ی تا دو برابر بزرگتر باشند.

- **برآوردگر حداکثر درستنمایی**

این برآوردگر که به صورت $\hat{d}_{MLE,c}$ تعریف می‌شود. برآوردگر بدون انحراف حداکثر درستمانی (MLE) عبارت است از:

$$\hat{d}_{MLE,c} = \hat{d}_{MLE}\left(1 - \frac{1}{k}\right) \tag{۳-۳۳}$$

که $\hat{d}_{MLE}$ یک معادله‌ی غیر خطی MLE را حل می‌کند.

$$-\frac{k}{\hat{d}_{MLE}} + \sum_{j=1}^{k} \frac{2\hat{d}_{MLE}}{x_j^2 + \hat{d}_{MLE}^2} = 0 \tag{۳-۳۴}$$

برآوردگرهای میانه‌ی نمونه و میانگین هندسی از نظر واریانس مجانبی، دقتی معادل 80% MLE دارند. در حالی که استنتاج محدوده‌های دمی فرم-بسته دشوار است. نشان خواهیم داد که توزیع $\hat{d}_{MLE,c}$ را می‌توان به وسیله‌ی یک معکوس گوسی[۱۳] تخمین زد.

۳-۵ تصویر تصادفی α -پایدار

توضیحات در بخش‌های قبلی، در مورد تصویر تصادفی نرم l_2 و تصویر تصادفی نرم l_1 صحبت کردیم. در این بخش، کاهش بعد در نرم l_α ، برای $0 < \alpha \leq 2$ مورد بررسی قرار خواهد گرفت. و نرم‌های l_1 و l_2 به عنوان حالت خاص بررسی می‌شوند.

مسئله اساسی در تصویر تصادفی پایدار، انجام برآورد آماری است. به عبارت دیگر، برآورد پارامتر مقیاس‌دهی توزیع پایدار متقارن. از آنجا که چگالی احتمال توزیع پایدار جز برای $\alpha = 1, 2$ فرم بسته ندارد. تولید برآوردگرهایی که از نظر آماری دقیق و از نظر محاسباتی بهینه هستند، جذاب است.

[13] Inverse Gaussian

برآوردگرهایی که بر اساس میانه‌های نمونه (با به طور کلی بر اساس چندک‌های نمونه) تولید شده‌اند، در علم آمار شناخته‌شده‌اند، اما خیلی دقیق نیستند به خصوص در مورد نمونه‌های کوچک، و برای تحلیل نظری از جمله محدوده‌های دم زمانی که $\alpha \neq 1, 2$ راحت نیستند.

ما در اینجا برآوردگرهای مختلفی را بر اساس میانگین هندسی، میانگین هارمونیک و توان نسبی بررسی خواهیم کرد.

۳-۵-۱ نتایج اصلی

گفته شد که، اگر دو بردار $u_1, u_2 \in \mathbb{R}^D$ (برای مثال، u_1 و u_2 دو ردیف اول در ماتریس داده‌ی $\mathbf{A}$ باشند) ، اگر $v_1 = \mathbf{R}^T u_1$ و $v_2 = \mathbf{R}^T u_2$ باشند که $\mathbf{R} = \mathbb{R}^{D\times k}$ شامل نمونه‌های i.i.d. در $S(\alpha, 1)$ باشد، آنگاه $x_j = v_{1,j} - v_{2,j}, j = 1, 2, \ldots, k$ ، همگی $S(\alpha, d_{(\alpha)})$ های i.i.d. هستند. که $d_{(\alpha)} = \sum_{i=1}^{D} |u_{1,i} - u_{2,i}|^\alpha$. بنابراین مسئله اصلی برآورد پارامتر مقیاس‌دهی $d_{(\alpha)}$ از تعداد k نمونه‌ی i.i.d. توزیع $S(\alpha, d_{(\alpha)})$ است. در بخش‌های قبلی به طور خلاصه توزیع‌های پایدار را مرور کردیم.

برآوردگری پرکاربرد در آمار بر اساس نمونه میان چندکی [14] [۳۰، ۳۱، ۵۰] است که به دلیل تقارن $S(\alpha, d_{(\alpha)})$ می‌توان آن را به صورت برآوردگر میانه‌ی نمونه، ساده‌سازی کرد.

$$\hat{d}_{(\alpha),me} = \frac{\text{median}\left\{|x_j|^\alpha, j = 1, 2, \ldots, k\right\}}{\text{median}\left\{S(\alpha, 1)\right\}^\alpha} \qquad (۳-۳۵)$$

علی‌رغم سادگی، مسائل بسیاری براساس برآوردگر میانه‌ی نمونه $\hat{d}_{(\alpha),me}$ وجود دارند. این برآوردگر، علی الخصوص برای نمونه‌های کوچک یا α ی کوچک دقیق نیست. همچنین برای تحلیل نظری دقیق از جمله تحلیل محدوده‌ی دم دشوار است.

ما برآوردگرهای زیادی را بر اساس میانگین هندسی، میانگین هارمونیک و توان کسری ارائه خواهیم کرد.

- برآوردگر میانگین هندسی (بدون انحراف) $\hat{d}_{(\alpha),gm}$:
- برآوردگر میانگین هندسی (دارای انحراف) $\hat{d}_{(\alpha),gm,b}$:

[14] inter-quantiles

این معادله به طور مجانبی معادل $\hat{d}_{(\alpha),gm}$ است. اگرچه برای $0.25 \leq \alpha \leq 1$ ، توزیع $\hat{d}_{(\alpha),gm,b}$ خطای میانگین مربعات کوچکتری در مقایسه با $\hat{d}_{(\alpha),gm}$ دارد.

- برآوردگر میانگین هارمونیک $\hat{d}_{(\alpha),hm}$: این معادله زمانی که $\alpha \to 0+$ به صورت مجانبی بهینه است و در مقایسه با برآوردگرهای میانگین هندسی، برای $\alpha \leq 0.344$ واریانس مجانبی کوچکتری دارد.

- برآوردگر میانگین ریاضی

برای $\alpha = 2$ بهترین روش استفاده از برآوردگر میانگین ریاضی $\frac{1}{k}\sum_{j=1}^{k}|x_j|^2$ است. این برآوردگر را می‌توان با استفاده از برآوردگر بیشینه درستنمایی، به شکلی که در بخش تصویر تصادفی نرمال توضیح داده شد، با استفاده از اطلاعات حاشیه‌آی ارتقاع داد.

- برآوردگر توان کسری $\hat{d}_{(\alpha),fp}$:

برای $\alpha = 2, \hat{d}_{(\alpha),fp}$ معادل برآوردگر میانگین ریاضی و برای $\alpha \to 0+$ معدل برآوردگر میانگین هارمونیک است. علاوه بر این، برای $\alpha \to 1$ ، $\hat{d}_{(\alpha),fp}$ دارای واریانس مجانبی برابر با برآوردگر میانگین هندسی است.

■

منابع و مراجع

[1] ftp://statgen.ncsu.edu/pub/thorne/molevoclass/atchleyoct19.pdf.

[2] http://www.informationweek.com/news/showarticle.jhtml?articleid=175801775.

[3] Achlioptas, Dimitris. Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 274–281. ACM, 2001.

[4] Achlioptas, Dimitris. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of computer and System Sciences*, 66(4):671–687, 2003.

[5] Aggarwal, Charu C. *Data streams: models and algorithms*, volume 31. Springer Science & Business Media, 2007.

[6] Aggarwal, Charu C, Wolf, Joel L, and Yu, Philip S. *A new method for similarity indexing of market basket data*. ACM, 1999.

[7] Agrawal, Rakesh, Imieliński, Tomasz, and Swami, Arun. Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM, 1993.

[8] Agrawal, Rakesh, Mannila, Heikki, Srikant, Ramakrishnan, Toivonen, Hannu, Verkamo, A Inkeri, et al. Fast discovery of association rules. *Advances in knowledge discovery and data mining*, 12(1):307–328, 1996.

[9] Agrawal, Rakesh, Srikant, Ramakrishnan, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[10] Andoni, Alexandr and Indyk, Piotr. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 459–468. IEEE, 2006.

[11] Babcock, Brian, Babu, Shivnath, Datar, Mayur, Motwani, Rajeev, and Widom, Jennifer. Models and issues in data stream systems. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 1–16. ACM, 2002.

[12] Brin, Sergey, Davis, James, and Garcia-Molina, Hector. Copy detection mechanisms for digital documents. In *ACM SIGMOD Record*, volume 24, pages 398–409. ACM, 1995.

[13] Brin, Sergey, Motwani, Rajeev, and Silverstein, Craig. Beyond market baskets: Generalizing association rules to correlations. In *Acm Sigmod Record*, volume 26, pages 265–276. ACM, 1997.

[14] Brin, Sergey, Motwani, Rajeev, Ullman, Jeffrey D, and Tsur, Shalom. Dynamic itemset counting and implication rules for market basket data. *Acm Sigmod Record*, 26(2):255–264, 1997.

[15] Brin, Sergey and Page, Lawrence. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998.

[16] Brinkman, Bo and Charikar, Moses. On the impossibility of dimension reduction in l 1. *Journal of the ACM (JACM)*, 52(5):766–788, 2005.

[17] Brinkman, Bo and Charikar, Moses. On the impossibility of dimension reduction in l 1. *Journal of the ACM (JACM)*, 52(5):766–788, 2005.

[18] Broder, Andrei Z. On the resemblance and containment of documents. In *Compression and complexity of sequences 1997. proceedings*, pages 21–29. IEEE, 1997.

[19] Buhler, Jeremy and Tompa, Martin. Finding motifs using random projections. *Journal of computational biology*, 9(2):225–242, 2002.

[20] Buldygin, Valeri Vladimirovich and Kozachenko, IU V. *Metric characterization of random variables and random processes*, volume 188. American Mathematical Soc., 2000.

[21] Chaudhuri, Surajit, Motwani, Rajeev, and Narasayya, Vivek. On random sampling over joins. In *ACM SIGMOD Record*, volume 28, pages 263–274. ACM, 1999.

[22] Chernoff, Herman et al. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952.

[23] Church, Kenneth Ward and Hanks, Patrick. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

[24] Crovella, Mark E and Bestavros, Azer. Self-similarity in world wide web traffic: evidence and possible causes. *IEEE/ACM Transactions on networking*, 5(6):835–846, 1997.

[25] Datar, Mayur, Immorlica, Nicole, Indyk, Piotr, and Mirrokni, Vahab S. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry*, pages 253–262. ACM, 2004.

[26] Datar, Mayur and Indyk, Piotr. Comparing data streams using hamming norms. In *Proceedings 2002 VLDB Conference: 28th International Conference on Very Large Databases (VLDB)*, page 335. Elsevier, 2002.

[27] Deerwester, Scott, Dumais, Susan T, Furnas, George W, Landauer, Thomas K, and Harshman, Richard. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[28] Dhillon, Inderjit S and Modha, Dharmendra S. Concept decompositions for large sparse text data using clustering. *Machine learning*, 42(1-2):143–175, 2001.

[29] Faloutsos, Michalis, Faloutsos, Petros, and Faloutsos, Christos. On power-law relationships of the internet topology. In *ACM SIGCOMM computer communication review*, pages 251–262. ACM, 1999.

[30] Fama, Eugene F and Roll, Richard. Some properties of symmetric stable distributions. *Journal of the American Statistical Association*, 63(323):817–836, 1968.

[31] Fama, Eugene F and Roll, Richard. Parameter estimates for symmetric stable distributions. *Journal of the American Statistical Association*, 66(334):331–338, 1971.

[32] Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. *The elements of statistical learning*, volume 10. Springer series in statistics New York, NY, USA:, 2001.

[33] Friedman, Jerome H, Baskett, Forest, and Shustek, Leonard J. An algorithm for finding nearest neighbors. *IEEE Transactions on computers*, 100(10):1000–1006, 1975.

[34] Friedman, Jerome H, Bentley, Jon Louis, and Finkel, Raphael Ari. An algorithm for finding best matches in logarithmic time. *ACM Trans. Math. Software*, 3(SLAC-PUB-1549-REV. 2):209–226, 1976.

[35] Garcia-Molina, Hector. Database systems: the complete book/hector garcia, molina jeffrey d. ullman, jennifer widom, 2002.

[36] Henzinger, Monika Rauch, Raghavan, Prabhakar, and Rajagopalan, Sridhar. Computing on data streams. *External memory algorithms*, 50:107–118, 1998.

[37] Hornby, Albert Sydney, editor. *Oxford Advanced Learner's Dictionary of Current English*. Oxford University Press, Oxford, UK, fourth edition, 1989.

[38] Hubert, Lawrence and Arabie, Phipps. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[39] Indyk, Piotr. Stable distributions, pseudorandom generators, embeddings and data stream computation. In *focs*, page 189. IEEE, 2000.

[40] Indyk, Piotr. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *Journal of the ACM (JACM)*, 53(3):307–323, 2006.

[41] Indyk, Piotr and Motwani, Rajeev. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613. ACM, 1998.

[42] Johnson, William B and Lindenstrauss, Joram. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

[43] Johnson, William B and Schechtman, Gideon. Embeddingl p m intol 1 n. *Acta Mathematica*, 149(1):71–85, 1982.

[44] Kannan, J Feigenbaum S, Strauss, M, and Viswanathan, M. An approximate l1-difference algorithm for massive data streams. *Unknown*, Unknown.

[45] Lee, James R and Naor, Assaf. Embedding the diamond graph in l p and dimension reduction in l 1. *Geometric & Functional Analysis GAFA*, 14(4):745–747, 2004.

[46] Leland, Will E, Willinger, Walter, Taqqu, Murad S, and Wilson, Daniel V. On the self-similar nature of ethernet traffic. *ACM SIGCOMM Computer Communication Review*, 25(1):202–213, 1995.

[47] Li, Ping. *Stable random projections and conditional random sampling, two sampling techniques for modern massive datasets*. Stanford, 2007.

[48] Li, Ping. Estimators and tail bounds for dimension reduction in l α (0< $\alpha\leq$ 2) using stable random projections. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 10–19. Society for Industrial and Applied Mathematics, 2008.

[49] Matias, Yossi, Vitter, Jeffrey Scott, and Wang, Min. Wavelet-based histograms for selectivity estimation. In *ACM SIGMoD Record*, volume 27, pages 448–459. ACM, 1998.

[50] McCulloch, J Huston. Simple consistent estimators of stable distribution parameters. *Communications in Statistics-Simulation and Computation*, 15(4):1109–1136, 1986.

[51] McKee, Sally A. Reflections on the memory wall. In *CF'04: Proceedings of the 1st conference on Computing frontiers*, page 162, 2004.

[52] Muthukrishnan, S. Data streams: Algorithms and applications (foundations and trends in theoretical computer science). *Hanover, MA: Now Publishers Inc*, 2005.

[53] Newman, Mark EJ. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

[54] Rand, William M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[55] Strehl, Alexander and Ghosh, Joydeep. A scalable approach to balanced, high-dimensional clustering of market-baskets. In *International Conference on High-Performance Computing*, pages 525–536. Springer, 2000.

[56] Vempala, Santosh S. *The random projection method*, volume 65. American Mathematical Soc., 2005.

[57] Wulf, Wm A and McKee, Sally A. Hitting the memory wall: implications of the obvious. *ACM SIGARCH computer architecture news*, 23(1):20–24, 1995.

[58] Yeung, Ka Yee and Ruzzo, Walter L. Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data. *Bioinformatics*, 17:763–774, 2001.

[59] Zolotarev, VM. One-dimensional stable distributions. translated from the russian by hh mcfaden. translation edited by ben silver. translations of mathematical monographs, 65. *American Mathematical Society, Providence, RI*, 1986.

پیوست

موضوعات مرتبط با متن گزارش پایان نامه که در یکی از گروه‌های زیر قرار می‌گیرد، در بخش پیوست‌ها آورده شوند:

۱. اثبات های ریاضی یا عملیات ریاضی طولانی.

۲. داده و اطلاعات نمونه (های) مورد مطالعه (Case Study) چنانچه طولانی باشد.

۳. نتایج کارهای دیگران چنانچه نیاز به تفصیل باشد.

۴. مجموعه تعاریف متغیرها و پارامترها، چنانچه طولانی بوده و در متن به انجام نرسیده باشد.

کد میپل

```
with(DifferentialGeometry):
with(Tensor):
DGsetup([x, y, z], M)
frame name: M
a := evalDG(D_x)
D_x
b := evalDG(-2 y z D_x+2 x D_y/z^3-D_z/z^2)
```

واژه‌نامه‌ی فارسی به انگلیسی

آ

اسکالر . . . . Scalar

ب

بالابر . . . . Lift

پ

پایا . . . . Invariant

ت

تناظر . . . . Correspondence

ث

ثابت‌ساز . . . . Stabilizer

ج

جایگشت . . . . Permutation

چ

چند جمله‌ای . . . . Polynomial

ح

حاصل‌ضرب دکارتی . . . . Cartesian product

خ

خودریختی . . . . Automorphism

د

درجه . . . . Degree

ر

ریزپردازنده . . . . microprocessor

ز

زیرمدول . . . . Submodule

س

سرشت . . . . Character

ص

صادقانه . . . . Faithful

ض

ضرب داخلی . . . . Inner product

ط

طوقه . . . . Loop

ظ

ظرفیت . . . . Valency

ع

عدم مجاورت . . . . Nonadjacency

ف

فضای برداری . . . . Vector space

ک

کاملاً تحویل‌پذیر . . . Complete reducibility

گ

گراف . . . . Graph

م

ماتریس جایگشتی . . . Permutation matrix

ن

ناهمبند . . . . Disconnected

و

وارون‌پذیر . . . . Invertible

ه

همبند . . . . Connected

ی

یال . . . . Edge

واژه‌نامه‌ی انگلیسی به فارسی

A

Automorphism . . . . خودریختی

B

Bijection . . . . دوسویی

C

Cycle group . . . . گروه دوری

D

Degree . . . . درجه

E

Edge . . . . یال

F

Function . . . . تابع

G

Group . . . . گروه

H

Homomorphism . . . . همریختی

I

Invariant . . . . پایا

L

Lift . . . . بالابر

M

Module . . . . مدول

N

Natural map . . . . نگاشت طبیعی

O

One to One . . . . یک به یک

P

Permutation group . . . . گروه جایگشتی

Q

Quotient graph . . . . گراف خارج‌قسمتی

R

Reducible . . . . تحویل پذیر

S

Sequence . . . . دنباله

T

Trivial character . . . . سرشت بدیهی

U

Unique . . . . منحصربفرد

V

Vector space . . . . فضای برداری

Abstract

This page is accurate translation from Persian abstract into English.

Key Words:

Write a 3 to 5 KeyWords is essential. Example: AUT, M.Sc., Ph. D,..

Amirkabir University of Technology
(Tehran Polytechnic)

Department of ...

M. Sc. Thesis

Title of Thesis

By

Name Surname

Supervisor

Dr.

Advisor

Dr.

Month & Year