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

سپاس‌گزاری

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

سیامک وهبد
دی ۱۳۹۷

چکیده

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 داده‌های 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
$\mathfrak{X}^{\backslash}(M)$	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

۱-۱ مقدمه

۱-۱-۱ مقدمه

فصل دوم

مرور ادبیات

۲-۱ 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. [۳] 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 ($0 < \alpha \leq 2$). 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 [۳] 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 خوبی تصویر تصادفی نرمال عمل کند. برای حالت نرم l_1 به عبارتی دیگر تصویر تصادفی کوچی [۷] 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 [۸] 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.

[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
[7] Cauchy random projections
[8] inverse Gaussian

۲-۲ داده‌های حجیم

عبارات زیر از سایت *Information Week* نقل قول شده‌اند [۹] :

- 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 $۱۰^{۱۵}$ 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 [۱۰] ، مالی، سیستم‌های اطلاعات جغرافیایی، دولت، دارو، ارتباطات از راه دور، محیط زیست و اینترنت.

۲-۲-۱ داده‌های حجیم وب

وب چقدر بزرگ است؟ جدول ۲-۱ نشان‌دهنده تعداد بازدید صفحات در موتورهای جستجوی امروزی است. به طور تخمینی حدود $D = ۱۰^{۱۰}$ صفحه‌ی وب را می‌توان بر اساس بازدید دو واژه‌ی بسیار پر کاربرد « A » و « THE » تخمین زد. جدول ۲-۱ همچنین نشان می‌دهد که حتی کلماتی که به ندرت کاربرد دارند هم تعداد زیادی بازدید دارند.

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

جدول ۲-۱: تعداد بازدید صفحات برای کلمات با بازخورد بالا و کلمات با بازخورد نادر

کلماتی با بازخورد معمولی چه میزان بازدید دارند؟ برای جواب این سوال ما به طور تصادفی ۱۵ صفحه از لغتنامه‌ی آموزشی انتخاب می‌کنیم. [۲] (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 [۱۱] تحت پوشش قرار نمی‌گیرند یا کلماتی که هنوز به لغتنامه‌های

[9] http://www.informationweek.com/news/showArticle.jhtml?articleID=175801775
[10] demographics
[11] Oxford English Dictionary

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 $n \approx ۱۰^۶$ و $D \approx ۱۰^{۱۰}$ بزرگ خواهد شد. در اینجا عدد (i, j) در $\mathbf{A}$ تعداد ظهور واژه i در سند j را نشان می‌دهد.

کارکردن با ماترسی در این ابعاد بزرگ چالش برانگیز است. برای مثال، شاخص LSI [۱۲] [۱]

۲-۳ کاهش بعد

۲-۴ کاهش بعد

۲-۵ کاهش بعد

۲-۵-۱ بارگیری مراجع

[12] latent semantic indexing

منابع و مراجع

[1] 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.

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

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

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

V

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 ...

MSc Thesis

Title of Thesis

By

Name Surname

Supervisor

Dr.

Advisor

Dr.

Month & Year