بسمه تعالی

دانشگاه تهران

پردیس دانشکده های فنی

دانشکده مهندسی برق و کامپیوتر

پیشنهاد و فرم حمایت از پایان‌نامه تحصیلات تکمیلی

☒ کارشناسی ارشد ☐ دکتری

شماره مرجع* :

* شماره مرجع، توسط معاونت پژوهشي پرديس دانشكده هاي

۱- خلاصه اطلاعات پایان نامه

عنوان پایان نامه به زبان فارسی:

یادگیری ماشینی مقیاس‌پذیر با استفاده از چارچوب نگاشت-کاهش

عنوان پایان نامه به زبان انکلیسی:

Scalable Machine Learning Using Map-Reduce Framework

نوع پایان نامه: ☐ **بنیادی** ☒ **کاربردی** ☒ **توسعه‌ای**

پردیس/دانشکده: فنی **دانشکده/گروه:** مهندسی برق و کامپیوتر

مقطع تحصیلی : کارشناسی ارشد **رشته و گرایش تحصیلی:** مهندسی نرم‌افزار

تاریخ پیشنهاد: دی ۹۴ **تاریخ تصویب:**

۲- اطلاعات اساتید راهنما و مشاورین

نوع مسئولیت	نام و نام‌خانوادگی	مرتبه علمی	محل خدمت	امضاء
استاد راهنما (مجری)	احمد خونساری	دانشیار	دانشگاه تهران – دانشکده فنی	
استاد راهنمای دوم (حسب نیاز)				
استاد مشاور				
استاد مشاور دوم (برای دکتری)				

۳- اطّلاعات دانشجو

نام و نام‌خانوادگی: امیر پوران‌بن‌ویسه **شماره دانشجویی:** ۸۱۰۱۹۳۱۲۷

رشته و گرایش تحصیلی: مهندسی نرم‌افزار **دانشکده:** مهندسی برق و کامپیوتر **مقطع تحصیلی:** کارشناسی ارشد

پست الکترونیک: veyseh@ut.ac.ir **تلفن ثابت:** ۰۳۸۱۴۲۴۳۸۰۵ **تلفن همراه:** ۰۹۳۸۵۲۲۶۵۸۷

۴- مشخصات موضوعی پایان نامه

تعریف مسأله، هدف و ضرورت اجرا (حداکثر سه صفحه)

با رشد وگسترش رایانه‌های شخصی و گوشی‌های تلفن همراه هوشمند[۱]، توسعه شبکه‌های حسگر و اینترنت اشیاء[۲]، فراگیر شدن اشتراک‌گذاری مطالب توسط کاربران و توسعه محاسبات ابری[۳] حجم عظیمی از داده‌ها توسط افراد ایجاد می‌شود. نحوه مدیریت این داده‌ها که حجم، سرعت تولید و تنوع زیادی دارند سبب شده است که موضوع داده‌های عظیم[۴] به یکی از چالش برانگیزترین موضوعات روز بدل شود.

در کنار این رشد سریع تولید داده، راه‌حل‌هایی برای مدیریت آن معرفی شده است. Hadoop به عنوان یکی از راه-حل‌های محبوب برای کار با داده‌های عظیم، چارچوبی را برای توزیع داده‌ها و مدیریت منابع توزیع شده فراهم می‌آورد. Hadoop با استفاده از امکانات ذکر شده، چارچوب نگاشت-کاهش را برای کاربر مهیا می‌سازد.

در روش‌های یادگیری ماشینی[۵]، به طور ویژه یادگیری عمیق[۶]، استفاده از داده‌های بیشتر به کیفیت بهتر ماشین کمک خواهد کرد. یکی از چالش‌های پیش‌رو در این حوزه، مقیاس‌پذیری روش یادگیری با افزایش حجم دادگان است. با این حال الگوریتم‌های مقیاس‌پذیر اندکی تا بحال معرفی شده‌اند و غالب کارهای موجود به کمک روش‌های تقریبی[۷] به مسئله مقیاس‌پذیر کردن روش‌های موجود یادگیری ماشینی پرداخته‌اند. هرچند چنین رویکردی از باب این که در یادگیری ماشینی کیفیت نهایی کار اهمیت دارد، مورد قبول است لیکن چنین دیدگاهی تنها به تسریع عمل یادگیری ماشین می‌پردازد و از بهبود کیفیت آن غافل است. به طور مثال در شبکه حافظه‌ای[۸] [1] ماشین نیازمند جستجو در حافظه‌ای عظیم از بردارهاست. در چنین شرایطی با رشد حجم داده این شبکه کارایی خود را عملا از دست خواهد داد. به طور مثال در کاربرد پاسخ‌گویی به پرسش‌های طبیعی با استفاده از شبکه حافظه‌ای، هر حقیقت[۹] به صورت یک بردار در حافظه ذخیره می‌شود. در صورتی که از معماری توزیع شده استفاده نکنیم، عملا امکان نگهداری حجم عظیمی از حقایق که می‌تواند حقایق موجود در یک پایگاه دانش[۱۰] باشد، وجود ندارد. در چنین شرایطی نیازمند نگهداری کل دادگان و توزیع آنها در بین گره‌های[۱۱] مختلف هستیم. به عنوان مثالی دیگر، راه‌حل‌هایی برای یادگیری شبکه‌های عصبی عمیق[۱۲] به کمک پردازش موازی معرفی شده است[2]، اما این روش‌ها از هزینه محاسباتی و پیچیدگی مدل رنج می‌برند.

Spark چارچوبی بر مبنای Hadoop است که کتابخانه‌ای را برای یادگیری ماشینی فراهم می‌کند. در این کتابخانه بسیاری از روش‌ها همچون SVM بدون کرنل پیاده‌سازی شده است. علی‌رغم مقیاس‌پذیری مناسب این پیاده‌سازی‌ها، اغلب به دلیل سادگی مدل‌های آن، در حل مسائل پیچیده کارکرد مناسبی ندارند.

در گذشته بسیاری از دادگان آموزشی، به طور ویژه نمونه‌های برچسب‌گذاری شده آن، عمدتا به دلیل کمبود منابع برای جمع‌آوری داده و هزینه بالای برچسب‌گذاری آن، حجم محدودی داشتند. با رشد روزافزون اشتراک‌گذاری داده توسط کاربران و امکاناتی نظیر جمع‌سپاری[۱۳] دو محدودیت مذکور برطرف گردیده است. به طور مثال در [3] با

[1] Smart phone
[2] Internet of Things (IoT)
[3] Cloud Computing
[4] Big Data
[5] Machine Learning
[6] Deep Learning
[7] Approximation
[8] Memory Network
[9] Fact
[10] Knowledge bae
[11] Node
[12] Deep Neural Network
[13] Crowd-Sourcing

استفاده از مطالب به اشتراک گذاشته شده کاربران تارنماهای[14] پرسش‌وپاسخ و امکان جمع‌سپاری AMT[15] مجموعه-ای مشتمل بر بیش از ۶۰۰۰ پرسش‌وپاسخ برچسب‌گذاری شده جمع‌آوری شد.

در نتیجه با پیدایش مفهوم داده‌های عظیم و در اختیار قرار گرفتن توان مدیریت و پردازش آن از طریق راه‌حل-هایی همچون Hadoop پژوهشی برای تطبیق دادن روش‌های یادگیری ماشینی به طور خاص یادگیری عمیق برای استفاده از این دادگان عظیم و رسیدن به روشی مقیاس‌پذیر ضروری است.

با توجه به توضیحات داده شده اهداف این پژوهش شامل موارد زیر می‌باشد:

- پیاده‌سازی طبقه‌بند[16] به صورت مقیاس‌پذیر بر بستر Hadoop
- بررسی توانایی طبقه‌بند برای مقیاس‌پذیری و مقایسه نتایج آن با روش‌های پیشگام[17] موجود

روشها و فنون اجرایی طرح

چارچوب نگاشت-کاهش چارچوبی است که در آن با استفاده از زوج مقادیر و تعریف توابع نگاشت و کاهش عملیات به صورت توزیع شده انجام می‌شود. از این چارچوب برای انجام عملیاتی نظیر محاسبات ماتریسی می‌توان بهره گرفت. در بسیاری از روش-های یادگیری نیازمند محاسبات سنگینی از قبیل ضرب ماتریسی می‌باشیم. این موضوع زمانی که تعداد دادگان آموزشی بسیار زیاد باشد می‌تواند فرایند یادگیری و نیز تست مدل را به شدت کند نماید. همچنین از سویی دیگر در برخی از مدل‌ها نیازمند ذخیره‌سازی حجم عظیمی از دادگان در حافظه می‌باشیم. در چنین شرایطی می‌بایست دادگان مورد نیاز مدل را بر روی چندین گره[18] جدا نگهداری کنیم. Hadoop با استفاده از سامانه مدیریت پرونده[19] HDFS[20] می‌تواند در توزیع دادگان مورد نیاز مدل مورد استفاده قرار گیرد. مزیت Hadopp نسبت به Spark در مقیاس‌پذیری بهتر آن است.

روش‌های یادگیری با در اختیار داشتن دادگان بیشتر می‌توانند کارایی مناسب‌تری داشته باشند اما استفاده از دادگان زیاد به دلیل کند شدن فرایند آموزش و یا تست عملی نیست. در پژوهش جاری قصد داریم با استفاده از چارچوب نگاشت-کاهش این مانع بر سر راه روش‌های یادگیری ماشینی را مطالعه کنیم و در صورت امکان راه‌حلی ارائه دهیم. به منظور بررسی توانایی این راه-حل در مقیاس‌پذیری، بر روی دادگان با اندازه‌های گوناگون عملیات یادگیری و تست انجام می‌شود و راه‌حلی مطلوب است که زمان اجرای الگوریتم رابطه خطی با افزایش حجم دادگان استفاده شده برای آموزش و تست داشته باشد.

پیشینه تحقیق (همراه با ذکر منابع اساسی)

پس از معرفی چارچوب نگاشت-کاهش در [4]، [5] با استفاده از این چارچوب به سرعت بخشیدن به روش‌های مختلف یادگیری ماشینی از قبیل naïve Bayes، SVM و Neural Network پرداخت. مشابه این پیاده‌سازی‌ها در Spark انجام شده‌است. هرچند در مقاله مذکور از توزیع دادگان و موازی‌سازی[21] به منظور افزایش سرعت یادگیری استفاده شده است توجهی به افزایش کارایی مدل به کمک این سازوکار نشده است. همانطور که در بخش تشریح مسئله آورده شد، بسیاری از مدل‌های یادگیری عمیق همچون شبکه‌های حافظه‌ای نیازمند به دسترسی به حجم عظیمی از حافظه دارند که این حافظه لزوما در یک

[14] Website
[15] Apache Mechanical Turk
[16] Classifier
[17] State of the art
[18] Node
[19] File System
[20] Hadoop Distributed File System
[21] Parallelizing

گره جا نمی‌گیرد. روش‌هایی برای جایگزینی حافظه به منظور غلبه بر این مشکل و تسریع عملیات جست‌وجو در حافظه معرفی شده است[6] اما این روش‌ها به کاهش کارایی مدل منجر می‌شود.

در حالی که در کار [5] هدف رسیدن به تقریبی از کارایی مدل نبود، بسیاری از کارهای انجام شده در این حوزه به کمک روش‌های تقریبی به مقیاس‌پذیر کردن مدل پرداخته‌اند. به طور مثال در [7] به کمک روش گرادیان نزولی تصادفی[۲۲] به تقریب محاسبات لازم بر روی کل دادگان آموزشی پرداخت. در [8] به بررسی کاربرد روش گرادیان نزولی تصادفی در رگرسیون لجستیک برای استفاده در شبکه توئیتر[۲۳] پرداخته شده است. استفاده موازی‌سازی شده از تقریب به کمک گرادیان نزولی تصادفی در شبکه‌های عصبی نیز کارایی مناسبی از خود نشان داده است[10][9].

با وجود کارهای انجام شده در حوزه افزایش سرعت یادگیری و تقریب زدن کیفیت نهایی مدل، پژوهشی برای آنکه بتوان با در نظر گرفتن تمام دادگان همچنان از سرعت و کیفیت مناسبی برخودار بود، مشاهده نشده است. خصوصا اینکه در برخی از مدل‌ها همچون شبکه‌های حافظه‌ای، استفاده از کل دادگان در برخی از کاربردها (همچون استخراج اطلاعات از منابع دانش) ضروری است.

مراجع

[1] S. Sukhbaatar, J. Weston, and R. Fergus. "End-to-end memory networks." Advances in Neural Information Processing Systems. 2015.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet classication with deep convolutional neural networks" In Advances in Neural Information Processing Systems, pp. 1097-1105, 2012.

[3] B. Jonathan, A. Chou, R. Frostig, and P. Liang. "Semantic Parsing on Freebase from Question-Answer Pairs." In EMNLP, pp. 1533-1544. 2013.

[4] J. Dean and S. Ghemawat. "Mapreduce: Simplied data processing on large clusters" In OSDIn'04, pp. 137-150, 2005.

[5] C. Chu, S. Kyun Kim, Y. Lin, Y. Yu, G. Bradski, A. Y Ng, and K. Olukotun. "Map-reduce for machine learning on multicore" Advances in Neural Information Processing Systems, pp. 19-281, 2007.

[6] J. Weston, S. Chopra, and A. Bordes. "Memory networks." arXiv preprint arXiv, pp. 1410-3916 2014.

[7] L. Bottou. "Large-scale machine learning with stochastic gradient descent" In Proceedings of COMPSTAT'2010, pp. 177-186. Springer, 2010.

[8] J. Lin and A. Kolcz. "Large-scale machine learning at twitter" In Proceedings of the 2012 ACM SIGMOD Inter- national Conference on Management of Data, pp. 793-804. ACM, 2012.

[9] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le. "Large scale distributed deep networks" In Advances in Neural Information Processing Systems, pp. 1223-1231, 2012.

[10] M. Li, D. G. Andersen, A. J. Smola, and K. Yu. "Communication efficient distributed machine learning with the parameter server" In Advances in Neural Information Processing Systems, pp. 19-27, 2014.

[22] Stochastic Gradient Descent
[23] Twitter

۵- مصوبه شورای پژوهشی و تحصیلات تکمیلی دانشکده مهندسی برق و کامپیوتر

۵-۱- فرم پیشنهاد و حمایت از پایان‌نامه در تاریخ ..... در شورای پژوهشی و تحصیلات تکمیلی دانشکده /گروه مطرح و نظر شورا به شرح زیر اعلام می‌شود:

☐ تصویب شد ☐ نیاز به اصلاح دارد ☐ به تصویب نرسید

۵-۲- عنوان طرح جامع تحقیقات استاد راهنما:

۵-۳- آیا پایان نامه پیشنهادی مرتبط با طرح جامع تحقیقات استاد راهنما/مشاور/گروه آموزشی/ دانشکده می باشد:

☐ بلی ☐ خیر

امضا استاد راهنما

امضاء رئیس / معاون پژوهشی و تحصیلات تکمیلی دانشکده مهندسی .....

شماره:

تاریخ:

معاون محترم آموزشی و تحصیلات تکمیلی پردیس دانشکده های فنی

با سلام و احترام,

فرم پیشنهاد و حمایت از پایان‌نامه کارشناسی ارشد / رساله دکتری آقای / خانم .....

با عنوان .....

به راهنمایی آقای / خانم دکتر .....

در شورای پژوهشی و تحصیلات تکمیلی دانشکده مهندسی ..... مورخ ..... به تصویب رسید.

خواهشمند است دستور فرمایید اقدامات مقتضی انجام شود.

امضاء رئیس / معاون پژوهشی و تحصیلات تکمیلی دانشکده مهندسی .....

شماره:

تاریخ:

معاون محترم پژوهشی پردیس دانشکده های فنی

با سلام و احترام ,

به پیوست فرم پیشنهاد و حمایت از پایان‌نامه تحصیلات تکمیلی با مشخصات مذکور که به تصویب شورای پژوهشی وتحصیلات تکمیلی دانشکده مهندسی ..... رسیده است، جهت دستور اقدام مقتضی تقدیم می‌شود.

امضاء معاون آموزشی و تحصیلات تکمیلی پردیس دانشکده های فنّی

رونوشت: معاون محترم پژوهشی و تحصیلات تکمیلی دانشکده مهندسی ..... : جهت اطلاع و پیگیری