مروری بر تکنیک های کلان داده (Big Data)

عطیه زاهد[1*]، محمدرضا سخی[2]
[1] گروه کامپیوتر، دانشگاه آزاد اسلامی کاشان، ایران ، کاشان، a.zahed@iaukashan.ac.ir
[2]کارشناس ارشد IT، دانشگاه آزاد اسلامی کاشان ،ایران ، کاشان ، m.r.sakhi@iaukashan.ac.ir

چکیده – در عصر حاضر کلان داده *(Big Data)* توجه محققان زیادی را به خود جلب کرده ، زیرا رشد بیش از حد داده ها مشکلات زیادی را ایجاد کرده است اما از طرفی اطلاعات پرارزش بالقوه ای نیز در این داده ها نهفته است. در مواجهه با کلان داده چالش های بزرگی مثل: نگهداری، به کارگیری ویافتن راه حل های درست در استفاده از کلان داده ها وجود دارد. تمرکز این مقاله در مرور و بررسی تکنیک هایی است که برای پردازش و به کار گیری ، کم کردن حجم داده های زیاد بدون از بین رفتن داده های مهم، نمایش اطلاعات و تحلیل آنها کشف یا به کارگرفته شده است. همچنین به معرفی چند ابزار کاربردی در این راستا می پردازد.

کلید واژه- کلان داده *(Big Data)* ، داده کاوی *(data mining)* ، یادگیری ماشین *(machine learning)* ، بهینه سازی *(optimization)*.

۱- مقدمه

برای بدست آوردن اطلاعات از کلان داده نیاز داریم که تکنیک ها و تکنولوژی های جدید را برای آنالیز آنها ایجاد کنیم. تاکنون دانشمندان انواع گوناگونی از تکنیک و تکنولوژی ها را برای بدست آوردن، بهینه کردن، آنالیز و مشاهده کلان داده ابداع کرده اند ولی باوجود این هنوز نیازهای زیادی وجود دارد. این تکنیک ها و تکنولوژی ها بسیاری از علوم شامل علم کامپیوتر، اقتصاد، ریاضی، آمار و ... را دربرگرفته اند.

روشهایی از علوم مختلف برای کشف اطلاعات از کلان داده مورد نیاز است. در این مقاله تکنیک های جدید برای بکارگرفتن اطلاعات حجیم مورد بحث قرار میگیرد.

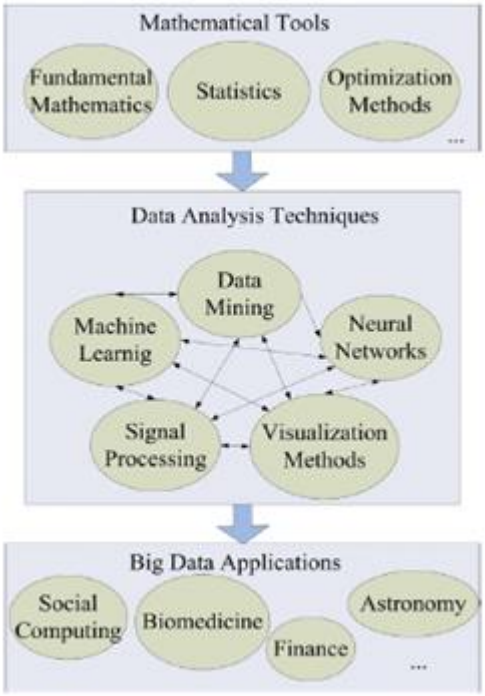
برای درک و به کارگیری کلان داده به ابزارهای خاصی نیاز است. ابزارهای کنونی روی ۳ کلاس تمرکز می کند که شامل ابزارهای در حال پردازش، ابزار پردازش جریان و ابزار تجزیه و تحلیل تعاملی هستند. بیشتر ابزارهای در حال پردازش برپایه ی زیربنای apache hadoop مثل mahout وdryad هستند. مورد آخر به احتمال زیاد برای آنالیز به موقع ابزارهای جریان داده ها لازم هستند Storm و s4 مثال های خوبی برای ابزارهای آنالیز جریان داده ها هستند.

پروسه های آنالیز تعاملی به کاربران اجازه می دهد که آنالیز خودشان را از اطلاعات انجام بدهند. کاربر به طور مستقیم به کامپیوتر وصل می شود و بنابراین می تواند با آن در زمان واقعی تعامل کند Google dremel.و apache drill ابزار کلان داده برپایه آنالیز تعاملی هستند[1] .

۲- تکنیک های کلان داده

کلان داده نیاز به تکنیک های خارق العاده ای برای پردازش موثر حجم زیادی از دیتا در زمان محدود دارند و در نتیجه تکنیک های کلان داده با استفاده از ابزارهای تخصصی بدست می آیند. برای مثال Wal- Mart یادگیری ماشین (machine learning) و تکنیک های آماری را برای کشف الگو از حجم زیادی از تراکنشهای داده ای استفاده کرد. این الگو ها می توانند باعث ایجاد رقابت شدید تری دربین استراژی های مالی و رقابت های تبلیغاتی شود. Taobao (یک شرکت چینی مانند eBay است) تکنیک های داده کاوی جریان عظیم را با داده های ثبت شده ی حاصل از جستجوی کاربران در وبسایتش مقایسه می کند و اطلاعات مطلوبی را برای پشتیبانی از تصمیم گیرندگان(decision-making) حاصل می کند. تکنیک های کلان داده ها (Big Data) مجموعه ای از قواعد را در برمیگیرد که شامل داده های آماری, داده کاوی , یادگیری ماشین , شبکه های عصبی, تحلیل شبکه های اجتماعی, پردازش سیگنال, تشخیص الگو، روشهای بهینه سازی و روش های مجازی سازی

می شود. تکنیک های مخصوص زیادی درباره ی این قواعد وجود دارد که همیشه با هم همپوشانی دارند. این موضوع در شکل ۱ نشان داده شده است.



شکل ۱ : تکنیک های کلان داده

۲-۱- روش های بهینه سازی

روش های بهینه سازی برای برطرف کردن مسائل تا حدودی در زمینه های گوناگونی مانند فیزیک، زیست، مهندسی و اقتصاد به کارگرفته شده اند. در [2]، چندین استراتژی محاسباتی برای نشان دادن مسائل سراسری بهینه سازی مورد بررسی قرار گرفته اند مانند: حرارت دهی شبیه سازی شده (simulated annealing)، حرارت دهی شبیه سازی شده ی تطبیقی، حرارت دهی کوانتوم، که همانند آن، الگوریتم ژنتیک (که به طور طبیعی ساختار خود را به سمت موازی سازی جلو میبرد و میتواند کارایی مفیدی را حاصل کند)نیز مورد بررسی قرار گرفته است.

بهینه سازی تصادفی، شامل برنامه نویسی ژنتیک، برنامه نویسی ارزیابی و بهینه سازی توده ذرات، بهینه سازی هایی هستند که بسیار ویژه و مفیدند و برگرفته از فرایند های طبیعت می باشند. هرچند، اغلب دارای پیچیدگی های حافظه و هزینه ی زمانی هستند. بسیاری از آثار تحقیقاتی [3]، با الگوریتم های تکاملی تعاملی در سدد توسعه ی بهینه سازی هایی با مقیاس بالا برآمده اند.

در برنامه های کاربردی کلان داده ها مانندWSN وITS نیز بهینه سازی های بلادرنگ لازم است. روش های کاهش داده و موازی سازی [4] به ترتیب در برنامه های بهینه سازی به کار رفته اند.

۲-۲- روشهای آماری

علم آمار, علم جمع آوری, سازماندهی و تعامل داده هاست. تکنیک های آماری, برای بدست آوردن همبستگی ها و روابط نسبی بین اهداف گوناگون به کار می رود.

توصیف های عددی نیز توسط علم آمار تامین می شود. هرچند، تکنیک های استاندارد آماری خیلی مناسب مدیریت کلان داده ها نیستند و بسیاری از تحقیقات برای توسعه ی تکنیک های کلاسیک یا روشهای کاملا جدید ارائه شده اند. نویسندگان مرجع [5] الگوریتم تقریبی کارامدی را برای رگراسیون یکنواخت چند متغیره مقیاس بالا ارائه کرده اند که برای توابع ارزیابی که در برابر متغییر های ورودی با نسب یکنواختی عمل میکنند، به کار می رود.

گرایش دیگری از تحلیل های آماری داده محور، متمرکز بر روی پیاده سازی الگوریتم های آماری موازی و مقیاسی می باشد. یک مقاله ی مروری تکنیک آماری موازی در مرجع [6]یافت می شود و چندین الگوریتم آماری موازی در [7] بررسی شده است. محاسبات آماری و یادگیری آماری دو موضوع شاخص از زیرمجموعه ی تنکیک های آماری به حساب می آیند.

۲-۳- داده کاوی

داده کاوی، مجموعه ای از تکنیک ها است که اطلاعات ارزشمندی(که الگو نامیده میشوند)را از داده ها اقتباس میکنند که شامل تحلیل خوشه بندی، دسته بندی، رگراسون و یادگیری قانون وابستگی می شود. داده کاوی شامل روش هایی از یادگیری ماشین و علم آمار می باشد. کاوش کلان داده ها نسبت به سایر الگوریتم های داده کاوی چالش بر انگیز تر است.

اگر خوشه بندی را به عنوان مثالی از داده کاوی در نظر بگیریم, یک روش بدیهی خوشه بندی در کلان داده ها توسعه ی روشهای فعلی(مانند خوشه بندی ساسله مراتبی, K-Means , فازی وCMeans) به نحوی می باشد که بتوانند از پس حجم های کاری عظیم نیز برآیند[8]. اکثر توسعه هایی که از چنین الگوریتم هایی گرفته شده است وابسته به تحلیل مقدار مشخصی از کلان داده به عنوان نمونه از کل میشود و در زمینه ی چگونگی نمونه برداری از میان کل جامعه ی آماری، با یکدیگر متفاوتند. از جمله این تکنیک ها می توان به الگوریتم CLARA(برنامه های کاربردی عظیم خوشه بندی), CLARANS (برنامه های کاربردی عظیم خوشه بندی مبتنی بر جستجوی تصادفی),BIRCH (کاهش تکرار متعادل سلسله بندی خوشه) و.. اشاره کرد. الگوریتم های ژنتیک نیز در خوشه بندی به عنوان معیار بهینه سازی به کار می رود تا نقاط مثبت مختلف را منعکس کند.

خوشه بندی داده های بزرگ همچنین برای اجرای توزیع و موازی سازی در حال توسعه است . با انتخاب تجزیه و تحلیل مجزا به عنوان مثالی دیگر، محققان سعی می کنند [9] تا الگوریتم کارآمدی را برای تجزیه و تحلیل مجزا در مقیاس بزرگ توسعه دهند . تاکید بر کاهش پیچیدگی محاسباتی است. با انتخاب داده های اطلاعاتی زیست شناسی به عنوان مثال دیگر ، به طور فزاینده ای داده محور می شود که منجر به تغییر پارادایم از زیست شناسی تک ژنی سنتی به رویکردی می گردد که تجزیه و تحلیل پایگاه داده یکپارچه و داده کاوی را ترکیب میکند . این پارادایم جدید سنتز پرتره های بزرگ عملکرد ژنومی را قادر می سازد .

۲-۴- یادگیری ماشین

یادگیری ماشینی یک استیلای مهم هوش مصنوعی است که هدفش طراحی الگوریتم هایی است که اجازه می دهند به کامپیوترها تا رفتارهایشان را بر اساس داده های تجربی بروز دهند .

بارز ترین مشخصه یادگیری ماشینی کشف دانش و تصمیم گیری هوشمندانه ی خودکار است . هنگامی که کلان داده مورد نظر است ، لازم است که مقیاس الگوریتم های یادگیری ماشین، هم در یادگیری تحت نظارت و هم در یادگیری بدون نظارت ، افزایش داده شود .

یادگیری ماشین عمیق به یک محدوده تحقیقات جدید در هوش مصنوعی تبدیل شده است . علاوه بر این ، چندین فریمورک مانند Map/Reduce ، DryadLINQ و جعبه ابزار یادگیری ماشینی موازی IBM وجود دارد که قابلیت هایی را برای افزایش مقیاس یادگیری ماشینی دارد [10]. به عنوان مثال، ماشین بردار پشتیبانی (SVM) ، که یک الگوریتم بسیار اساسی مورد استفاده در طبقه بندی و رگرسیون مسایل است ، هم از نظر کاربرد و هم از نظر زمان محاسبات به شدت دچار مساله ی مقیاس پذیری است.

SVM موازی (PSVM) اخیرا به منظور کاهش حافظه و زمان مصرف معرفی شده است . بسیاری از الگوریتم های یادگیری ماشینی رشته ای وجود دارد اما بسیاری از رشته های فرعی خاص مهم در یادگیری ماشینی در مقیاس بزرگ مانند سیستم های بزرگ پیشنهاد دهنده ، پردازش زبان طبیعی، یادگیری قانون مشارکتی، کل آموزی ، هنوز هم با مشکلات مقیاس پذیری روبرو هستند[11].

شبکه های عصبی مصنوعی (ANN) از تکنیک های بالغ است و طیف گسترده ای از برنامه های کاربردی را پوشش می دهد. برنامه های کاربردی موفق آن را میتوان در تشخیص الگو، تجزیه و تحلیل تصویر، کنترل تطبیقی و حوزه های دیگر پیدا کرد .

بسیاری از شبکه های عصبی مصنوعی که در حال حاضر برای هوش مصنوعی به کار گرفته شده بر اساس برآوردهای آماری ، بهینه سازی طبقه بندی و تئوری کنترل می باشند .

به طور کلی میتوان اذعان نمود که هر چه ، لایه های پنهان و گره ها در یک شبکه عصبی بیشتر باشند ، با دقت بالاتری می توانند الگوریتم را فراهم کنند . با این حال، پیچیدگی در شبکه های عصبی زمان یادگیری را افزایش می دهد.

بنابراین، فرایند یادگیری در یک شبکه عصبی با حجم بالای داده ها به شدت زمان و حافظه مصرف میکند .

پردازش عصبی از مجموعه داده ها در مقیاس بزرگ اغلب به شبکه های بسیار بزرگ منتهی میشود . پس ، دو چالش عمده در این وضعیت وجود دارد . یکی این است که اجرای الگوریتم های متداول آموزشی بسیار ضعیف اجرا می شود و دیگری این

است که زمان آموزش و محدودیت های حافظه به طور فزاینده ای بهبود پذیر نیستند .

به طور طبیعی، دو رویکرد معمول را می توان دراین وضعیت به کار برد : یکی کاهش حجم داده ها با برخی از روش های نمونه گیری و ساختار شبکه عصبی که شاید یکسان باقی بماند . و رویکرد دیگر افزایش مقیاس شبکه عصبی به روش های موازی و توزیعی است[12] .

به عنوان مثال، ترکیبی از یادگیری عمیق و تکنیک های استنباطی آموزش موازی راه های بالقوه ای را برای پردازش داده های بزرگ فراهم می کند .

۲-۵- روش تجسم

روش تجسم، تکنیک های مورد استفاده برای ایجاد جداول، تصاویر، نمودارها و دیگر روش های نمایش بصری برای درک داده ها هستند.

تجسم داده های بزرگ مانند مجموعه ی نسبی داده های کوچک سنتی به دلیل پیچیدگی در 3Vs از 4Vs آسان نیست .

روش های تجسم دارای پیشرفتهایی بوده ولی این پیشرفت ها هنوز کافی نیست . هنگامی که تجسم داده ها در مقیاس بزرگ مورد نظر است ، بسیاری از محققان با استفاده از استخراج ویژگی و مدل سازی هندسی به طور قابل ملاحظه اندازه داده ها را قبل از ارائه داده ها ی واقعی کاهش می دهد .

برای دقت بیشتر و موثرتر تفسیر داده ها ، برخی از محققان سعی می کنند تفسیر داده های نرم افزارهای حالت دسته ای را در بالاترین رزولوشن ممکن و موازی اجرا کنند . در[13] ، مولف سعی میکند داده ها را جمع و جور کند و داده ها را در مقیاس بزرگ به یک تقریب خوب برساند .

۲-۶- تحلیل شبکه های اجتماعی

تحلیل شبکه اجتماعی (SNA) که به عنوان یک تکنیک کلیدی در جامعه شناسی مدرن ظهور کرده است ، به روابط اجتماعی بر حسب نظر تئوری شبکه می نگرد و شامل گره ها و روابطه ها است . همچنین موارد قابل توجه ذیل را در زمینه انسان شناسی، زیست شناسی، مطالعات ارتباطات، اقتصاد، جغرافیا، تاریخ، علوم اطلاع رسانی، مطالعات سازمانی، روانشناسی اجتماعی، مطالعات توسعه ، و زبانشناسی اجتماعی به دست آورده است و در حال حاضر به عنوان یک ابزار معمول مصرف کننده در دسترس است .

SNA شامل طراحی سیستم اجتماعی ، مدل سازی رفتار انسان ، تجسم شبکه اجتماعی ، تجزیه و تحلیل تکامل شبکه های اجتماعی ، و نمودار تحقیق و کاوش است .

به تازگی ، شبکه های اجتماعی آنلاین و تجزیه و تحلیل رسانه های اجتماعی محبوب شده اند . یکی از موانع اصلی در ارتباط با SNA وسعت داده های بزرگ است . تجزیه و تحلیل شبکه ای متشکل از میلیون ها و یا میلیاردها موضوع و از نظر محاسباتی پرهزینه است. دو محدوده تحقیقاتی پرشور، محاسبات اجتماعی و محاسبات ابری، تا حدودی تاکید کننده SNA می باشند .

سطح بالاتر فن آوری های کلان داده شامل سیستم های فایل های توزیعی ، سیستم های محاسباتی توزیعی ، سیستم پردازش موازی انبوه (MPP) ، داده کاوی مبتنی بر محاسبات شبکه ، ذخیره سازی مبتنی بر ابر و منابع محاسباتی ، و همچنین محاسبات دانه ای و محاسبات بیولوژیکی میباشد .

بسیاری از محققان بدی ابعاد پذیری را به عنوان یک جنبه از مشکلات کلان داده در نظر می گیرند . در واقع، کلان داده نباید در حجم داده های معمولی فشرده شوند بلکه همه ویژگی ابعاد بزرگ داده ها باید درنظر گرفته شود . در واقع ، پردازش داده ها با ابعاد بالا در حال حاضر یک کار دشوار در پژوهش های علمی است .

پیشرفته ترین تکنیک ها برای اداره داده های با ابعاد بسیاربزرگ به طور مستقیم باعث کاهش بعد داده میشود . یعنی سعی می شود نقشه فضای داده با ابعاد بالا با حداقل فقدان اطلاعات به فضایی با ابعاد کم تر تبدیل شود .

تعداد زیادی از روش ها به منظور کاهش ابعاد وجود دارد . روش نگاشت خطی، مانند تجزیه و تحلیل مولفه های اصلی (PCA) و تجزیه و تحلیل عوامل، تکنیک های کاهش ابعاد خطی محبوب هستند[14].

تکنیک های غیر خطی شامل هسته PCA ، تکنیک های یادگیری چندگانه : مانند Isomap ، تعبیه خطی محلی (LLE)، چشایی LLE، لاپلاس ، و LTSA می باشد. به

تازگی، یک شبکه عمیق مولد ، به نام خود رمزگذار ، به خوبی به عنوان کاهش ابعاد غیر خطی به کار می رود . طرح ریزی تصادفی در کاهش ابعاد نیز توسعه خوبی یافته است [15].

۳- نتیجه گیری

افزایش روز افزون داده ها از یک طرف چالش های جدیدی را ایجاد کرده است ولی از طرفی دیگر فرصتهای جدیدی نیز فراهم کرده که در صورت کشف و استفاده صحیح از آنها، افق های زیادی را پیش روی محققان،پزشکان، بازرگانان، دولتمردان و... می گشاید و چه بسا می تواند در حل مسائلی که تاکنون غیر قابل حل به نظر میرسیده اند کارگشا باشد. اما قبل از بهره برداری از مزایای آن ابتدا باید راه حل های استفاده از آن را بشناسیم.

این مقاله سعی میکند به تکنیک هایی که تا کنون برای کلان داده استفاده شده است، بپردازد . تکنیکهایی مانند : روشهای بهینه سازی، یادگیری ماشین، داده کاوی ، روش های تجسم و تحلیل شبکه های اجتماعی. همچنین در بعضی موارد به ارائه پیشنهادهایی در راستای بهبود هر یک از این روشها می پردازد.

مراجع

[1] C.L. Philip Chen, C.-Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Inform. Sci.(2014),http://dx.doi.org/10.1016/j.ins.2014.01.015

[2] Muhammad Sahimi, Hossein Hamzehpour, Efficient computational strategies for solving global optimization problems, Comput. Sci. Eng. 12 (4)(2010) 74–83.

[3] XYamille del Valle, Ganesh Kumar Venayagamoorthy, Salman Mohagheghi, Jean-Carlos Hernandez, Ronald G. Harley, Particle swarm optimization:basic concepts, variants and applications in power systems, IEEE Trans. Evol. Comput. 12 (2) (2008) 171–195.

[4] Yi Cao, Dengfeng Sun, A parallel computing framework for large-scale air traffic flow optimization, IEEE Trans. Intell. Trans. Syst. 13 (4) (2012) 1855–1864.

[5] Sysoev Oleg, Oleg Burdakovb, A. Grimvalla, A segmentation-based algorithm for large-scale partially ordered monotonic regression, Comput. Stat.Data Anal. 55 (8) (2011) 2463–2476.

[6] Philippe Pebay, David Thompson, Janine Bennett, Ajith Mascarenhas, Design and performance of a scalable, parallel statistics toolkit, in: 2011 IEEE International Symposium on Parallel and Distributed Processing Workshops and Phd Forum (IPDPSW), 2011, pp. 1475–1484.

[7] Janine Bennett, Ray Grout, Philippe Pebay, Diana Roe, David Thompson, Numerically stable, single-pass, parallel statistics algorithms, in: IEEE International Conference on Cluster Computing and Workshops, 2009, CLUSTER '09, 2009, pp. 1–8.

[8] Jin Zhou, C.L. Philip Chen, Long Chen, Hong-Xing Li, Wei Zhao, A collaborative fuzzy clustering algorithm in distributed network environments, IEEE Trans. Fuzzy Syst. PP (99) (2013) 1.

[9] Weiya Shi, Yue-Fei Guo, Cheng Jin, Xiangyang Xue, An improved generalized discriminant analysis for large-scale data set, in: Seventh InternationalConference on Machine Learning and Applications, 2008, 2008, pp. 769–772.

[10] Deng Cai, Xiaofei He, Jiawei Han, Srda: an efficient algorithm for large-scale discriminant analysis, IEEE Trans. Knowl. Data Eng. 20 (1) (2008) 1–12.

[11] Jian xiong Dong, Adam Krzy_ zak, Ching Y. Suen, Fast svm training algorithm with decomposition on very large data sets, IEEE Trans. Pattern Anal.Mach. Intell. 27 (4) (2005) 603–618.

[12] Wei-Hua Lin Shuang Shuang Li Cheng Chen, Zhong Liu, Kai Wang, Distributed

modeling in a mapreduce framework for data-driven traffic flow forecasting, IEEE Trans. Intell. Trans. Syst. 14 (1) (2013) 22–33.
[13] David Thompson, Joshua A. Levine, Janine C. Bennett, Peer-Timo Bremer, Attila Gyulassy, Valerio Pascucci, Philippe P. Pebay, Analysis of large-scale scalar data using hixels, in: 2011 IEEE Symposium on Large Data Analysis and Visualization (LDAV), 2011, pp. 23–30.
[14] Naiyang Guan, Dacheng Tao, Zhigang Luo, Bo Yuan, Online nonnegative matrix factorization with robust stochastic approximation, IEEE Trans. Neural Networks Learning Syst. 23 (7) (2012) 1087–1099.
[15] Bjon Bringmann, Michele Berlingerio, Francesco Bonchi, Aristides Gionis, Learning and predicting the evolution of social networks, IEEE Intell. Syst. 25 (4) (2010) 26–35.