به نام خدا

پژوهشکده فناوری اطلاعات

"پروژه تدوین نقشه راه کلان داده‌ها"

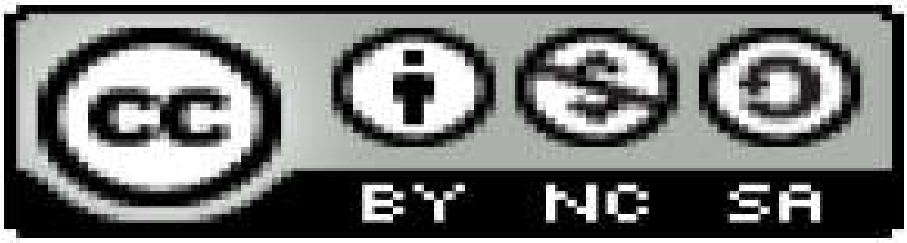
گزارش فاز اول

"تحلیل کلان داده‌ها"

کد پروژه: ۹۰۴۹۵۰۱۰۰

مجری:	دکتر محمدشهرام معین
تهیه‌کننده:	نکیسا برزگر، مهدی جم‌پور، اکبر دارابی و سمیه فتاحی
کد گزارش:	
تاریخ ارائه:	۹۵/۱۲/۲۵
نسخه/وضعیت:	۲/نهایی

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد



نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

شناسنامه گزارش

عنوان: تحلیل کلان داده ها		**شماره نسخه:** ۲
کد:	**نوع گزارش:** راهبردی	**تاریخ ارائه گزارش:** ۹۵/۱۲/۲۵
نام پروژه: تدوین نقشه راه کلان داده‌ها		**نوع پروژه:** راهبردی
تاریخ شروع: ۹۵/۷/۶		**تاریخ پایان:** ۹۶/۴/۶
نام گروه: فناوری اطلاعات / سامانه های چندرسانه ای		
کد پروژه: ۹۰۴۹۵۰۱۰۰		**شماره و تاریخ قرارداد:** حکم شماره ۶۷۰۲/د/۵۰۰/پ در تاریخ ۹۵/۸/۱۰
مجری: محمدشهرام معین		**ناظر/ ناظرین:** آقایان دکتر علیرضا یاری، دکتر روح ا... رحمانی، دکتر مجید رسولی دیسفانی، دکتر امین شکری پور و مهندس فرزاد ابراهیمی
تهیه کننده/ تهیه کنندگان: نکیسا برزگر، مهدی جم‌پور، اکبر دارابی، سمیه فتاحی		
نام و نشانی مجری: تهران، انتهای خیابان کارگر شمالی، پژوهشگاه ارتباطات و فناوری اطلاعات(مرکز تحقیقات مخابرات ایران)-کد پستی: ۱۴۳۹۹۵۵۴۷۱-		
نام و نشانی حمایت کننده: تهران، خیابان دکتر شریعتی، وزارت ارتباطات و فناوری اطلاعات		
ملاحظات:		

چکیده:

امروزه با گسترش روزافزون استفاده از فناوری اطلاعات و نرخ بالای تولید اطلاعات دیجیتال، توسعه فناوری‌ها و روش‌های تحلیل کلان داده‌ها از اهمیت به سزایی برخوردار می‌باشند. در حالت کلی به دلیل خصوصیات ویژه کلان داده‌ها، استفاده از الگوریتم‌های تحلیل اطلاعات اعم از روش‌های آماری، هوش‌مصنوعی و یادگیری ماشینی به شیوه معمول وسنتی قابل انجام و یا مقرون به صرفه نبوده و برای تضمین کارایی قابل قبول باید روش‌های جدیدی مانند Map-Reduce به کار گرفته شوند. در این گزارش روش‌های مورد استفاده در علوم داده برای تحلیل کلان داده‌ها و بسترهای نرم افزاری متناسب با شرایط خاص کلان داده‌ها مورد بررسی قرار گرفته و خصوصیات هر یک بطور خلاصه بیان می‌گردند. همچنین برای مقایسه بهتر این روش‌ها، گونه‌شناسی مرتبط با این روش‌ها نیز ترسیم گردیده است. شایان ذکر است که تمامی روش‌های ذکر شده در این گزارش، به طور عملیاتی پیاده‌سازی گردیده و زیرساخت‌های لازم و روش‌های پیاده‌سازی آن‌ها در بسترهای متناسب فناوری کلان داده‌ها موجود می‌باشند.

کلمات کلیدی:

تحلیل کلان داده‌ها، الگوریتم‌های هوش‌مصنوعی، یادگیری ماشین، علم داده‌ها

وضعیت گزارش: نهایی	**زبان گزارش:** فارسی
وضعیت دسترسی: آزاد	**تعداد صفحات:** ۵۴

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

چکیده

امروزه با گسترش روزافزون استفاده از فناوری اطلاعات و نرخ بالای تولید اطلاعات دیجیتال، توسعه فناوری‌ها و روش‌های تحلیل کلان داده‌ها از اهمیت به سزایی برخوردار می‌باشند. در حالت کلی به دلیل خصوصیات ویژه کلان داده‌ها، استفاده از روشالگوریتم‌های تحلیل اطلاعات اعم از روش‌های آماری، هوش‌مصنوعی و یادگیری ماشینی به شیوه معمول وسنتی قابل انجام و یا مقرون به صرفه نبوده و برای تضمین کارایی قابل قبول باید روش‌های جدیدی مانند Map-Reduce به کار گرفته شوند. در این گزارش روش‌های مورد استفاده در علوم داده برای تحلیل کلان داده‌ها و بسترهای نرم افزاری متناسب با شرایط خاص کلان داده‌ها مورد بررسی قرار گرفته و خصوصیات هر یک بطور خلاصه بیان می‌گردند. همچنین برای مقایسه بهتر این روش‌ها، گونه‌شناسی مرتبط با این روش‌ها نیز ترسیم گردیده است. شایان ذکر است که تمامی روش‌های ذکر شده در این گزارش، به طور عملیاتی پیاده‌سازی گردیده و زیرساخت‌های لازم و روش‌های پیاده‌سازی آن‌ها در بسترهای متناسب فناوری کلان داده‌ها موجود می‌باشند.

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

اطلاعات مرتبط

مستندات مرتبط

شماره مستند	نوع مستند	نام مستند

تغییرات اعمال‌شده در نسخه‌های پیشین

شماره نسخه	تاریخ	تغییرات اعمال شده

تأییدکنندگان

	نام و نام خانوادگی	تاریخ	امضاء	ملاحظات
مجری پروژه	محمدشهرام معین			
تهیه کننده/ تهیه کنندگان	نکیسا برزگر، مهدی جم‌پور، اکبر دارابی و سمیه فتاحی			
ناظر پروژه	دکتر علیرضا یاری، دکتر روح ا... رحمانی، دکتر مجید رسولی دیسفانی، دکتر امین شکری پور و مهندس فرزاد ابراهیمی			
مدیر گروه	مهندس فرزاد ابراهیمی			
مسئول مستندات پژوهشکده	مانا روزی طلب			
رئیس پژوهشکده/ معاون پژوهشی	دکتر علیرضا یاری/ دکتر کامبیز بدیع			

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

تقدیر و تشکر

بدین وسیله از ناظرین و مشاورین محترم پروژه قدردانی می‌شود.

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

سرفصل مطالب

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

فهرست جدول‌ها

فهرست شکل‌ها

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

۱ مقدمه

با فراگیر شدن استفاده از سامانه‌های دیجیتال و علوم کامپیوتر در ۵۰ سال اخیر، روش‌ها و الگوریتم‌های متعددی در رابطه با تحلیل داده‌ها معرفی گردیده‌اند. این روش‌ها، گستره وسیعی را از جمله روش‌های آماری[۱]، مدل‌سازی‌های ریاضی[۲]، استفاده از مقادیر آستانه[۳]، سامانه‌های خبره[۴]، منطق فازی[۵]، شناسایی آماری و ساختاری تشخیص الگو[۶]، سامانه‌های هوشمند[۷] و الگوریتم‌های مبتنی بر یادگیری ماشینی[۸] را در بر می‌گیرند. در دهه‌های اخیر بسیاری از این الگوریتم‌ها، در کاربردهای متعددی مورد استفاده قرار گرفته و استفاده از آن‌ها نتیجه‌بخش بوده است. برای مثال می‌توان به سامانه‌های پیشگویی در بازار بورس، سامانه‌های پردازش تصاویر[۹] و سامانه‌های تشخیص هویت اتوماتیک[۱۰] افراد اشاره کرد. در چند سال اخیر، به دلیل گسترش روزافزون استفاده از فناوری دیجیتال در تمامی کاربردهای مورد نیاز انسان خصوصا" توسعه اینترنت اشیا[۱۱] و رایانش ابری[۱۲]، حجم داده‌های تولید شده با نرخ بسیار بالایی رو به رشد بوده و همه ساله بر این نرخ افزوده می‌گردد. ضمنا" علاوه بر افزایش حجم تولیدی، تنوع و ساختار داده‌ها نیز بسیار متنوع بوده و این خصوصیت باعث عدم امکان استفاده از روش‌های سنتی ذخیره‌سازی و تحلیل سنتی داده‌ها گردیده است. برای رفع این مشکل، متخصصان علوم کامپیوتر چند سالی است که با مطرح کردن ایده‌ی کلان داده‌ها[۱۳]، سعی در معرفی و توسعه حوزه جدیدی از دانش بشری را دارند که در آن بتوان روش‌هایی برای ذخیره‌سازی، تحلیل و استفاده از این حجم زیاد داده تولید شده، ابداع نمود. قابل ذکر است که در حال حاضر زمینه‌های زیرساختی و

---

[۱] Statistical Methods
[۲] Mathematical Modeling
[۳] Thresholding
[۴] Expert Systems
[۵] Fuzzy Systems
[۶] Statistical and Structural Pattern Recognitions
[۷] Intelligence Systems
[۸] Machine Learning
[۹] Image Processing Systems
[۱۰] Identification Recognition systems
[۱۱] Internet of Things
[۱۲] Cloud Computing
[۱۳] Big Data

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

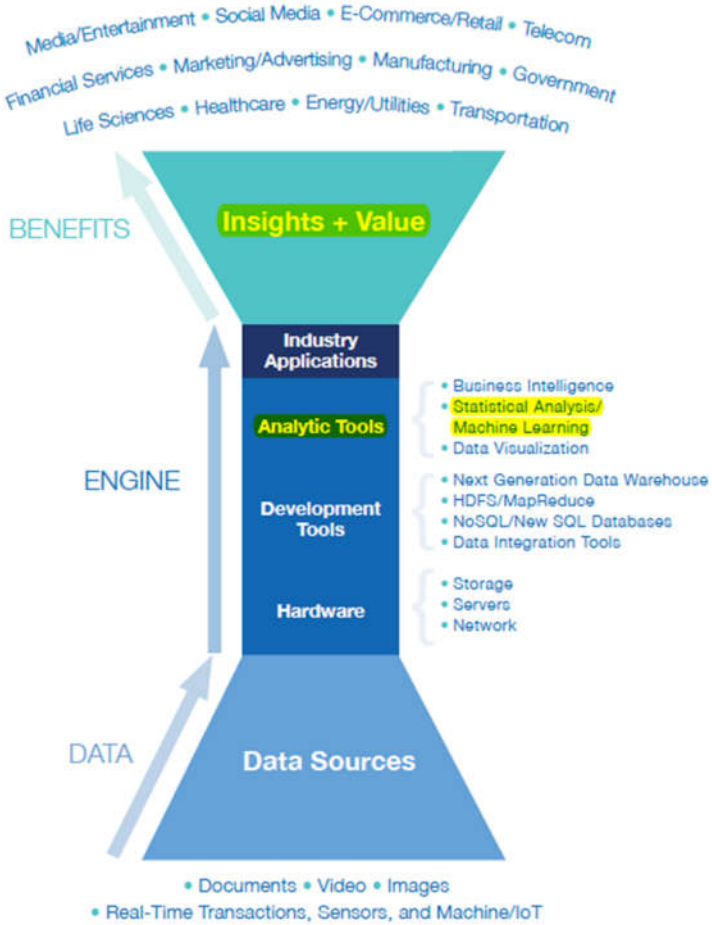
روش‌های ذخیره‌سازی داده‌های عظیم، تحت سلطه شرکت‌های بزرگ فناوری اطلاعات و شرکت‌های اینترنتی ایالات متحده آمریکا می‌باشند. از طرف دیگر بسیاری از ابزارهای پشتیبانی و معماری ذخیره‌سازی در حال حاضر متن باز (Hadoop، Hive، Spark، Shark، HBase، Riak، Titan Flink، و غیره) هستند، که میدان بازی را برای تولیدکنندگان و فروشندگان ابزار در این زمینه را هموار می‌کند. از این رو پیشی گرفتن و یا رقابت با آن‌ها را در این زمینه ها به سادگی با تکرار آن‌چه آن‌ها در حال حاضر به دست آورده‌اند، روش کارآمدی به نظر نمی‌رسد، بلکه باید بر بستر مشخصی که آن‌ها ایجاد کرده‌اند چیزی بنا نهاد [۱].

پیشرفت در فناوری اطلاعات و رشد گسترده آن در زمینه‌های گوناگون مانند کسب و کار، مهندسی، پزشکی و مطالعات علمی موجب تولید داده در مقیاس کلان می شود. کشف دانش و تصمیم گیری مبتنی بر این کلان داده‌های در حال رشد از نظر سازمان‌دهی، سرعت و پردازش یک کار چالش بر انگیز است که یک روند در حال ظهور شناخته شده با نام محاسبات کلان داده می‌باشد. این الگوی جدید، ترکیبی از محاسبه در مقیاس بزرگ، فناوری‌های داده‌های فشرده جدید و روش‌های ریاضی برای تجزیه و تحلیل کلان‌داده هستند. بنابر این الگوریتم و روش‌های تجزیه و تحلیلی که مقیاس پذیر باشند را می‌توان در حوزه کلان داده‌ها بکار گرفت [۱], [۲]. در این گزارش، ابتدا از دیدگاه علوم داده تجارت هوشمند، الگوریتم‌ها و روش‌های تحلیل داده‌ها و گونه‌شناسی مرتبط ، به دلیل اهمیت آنها برای پژوهش، رشد و توسعه در حوزه کلان‌داده‌ها [۱]، [۲] ،. مورد بررسی قرار می‌گیرد. سپس نمونه‌هایی موفق از پیاده‌سازی این الگوریتم‌ها در کاربردهایی مبتنی بر فناوری کلان داده‌ها معرفی گردیده‌اند. شایان ذکر است که الگوریتم‌های معرفی شده در این گزارش همگی بطور عملیاتی در کاربردهای فناوری کلان داده‌ها مورد استفاده قرار گرفته‌اند و قابل اجرا بر روی بسترهای نرم‌افزاری متناسب با خصوصیات ویژه کلان داده‌ها می‌باشند.

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

۲ کلان داده‌ها، هوش تجاری (کسب و کار) و علم داده‌ها

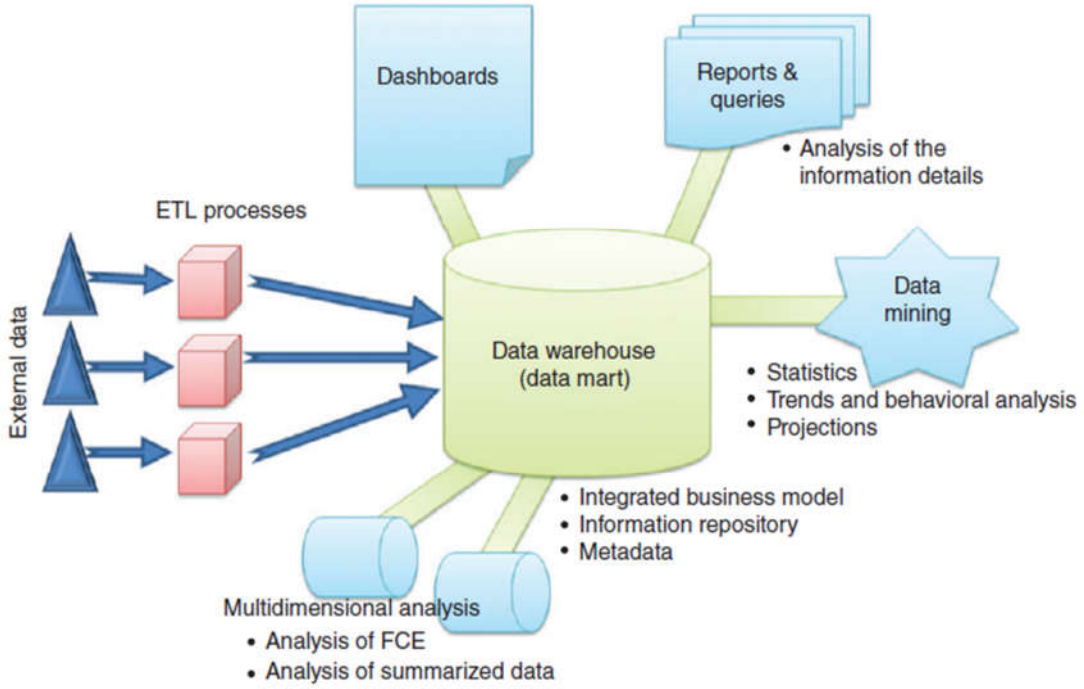
" کلان داده‌ها" داده‌هایی هستند که مقیاس، تنوع و پیچیدگی آن نیاز به معماری، فن، الگوریتم، و روش‌های تجزیه و تحلیل جدید برای مدیریت و استخراج ارزش و دانش پنهان دارد. تجزیه و تحلیل کلان داده‌ها فرصت‌های جدید قابل توجهی برای سازمان‌ها جهت استخراج اطلاعات و ایجاد ارزش‌های نو و مزیت رقابتی از با ارزش‌ترین دارایی خود یعنی داده‌ها (شکل ۱) ایجاد می‌کند. برای کسب و کارها، تجزیه و تحلیل کلان داده‌ها به ایجاد بهره‌وری، کیفیت، محصولات و خدمات شخصی به منظور بالا بردن رضایت و بهره‌مندی مشتری کمک می‌کند. از نظر کاوش و تلاش‌های علمی، تجزیه و تحلیل کلان داده‌ها راه جدیدی از پژوهش‌ها با نتایج بالقوه غنی‌تر و بینش عمیق‌تر از آنچه قبلا" در دسترس است، ترسیم می‌کند. تجزیه و تحلیل کلان داده‌ها در بسیاری از موارد داده‌های ساخت‌یافته و غیر ساخت‌یافته را با تغذیه و پرسش بدون درنگ، گشایش مسیرهای جدید به نوآوری و بینش، ادغام می‌کند.



شکل۱- صنعت کلان داده‌ها [۲]

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

اگر چه به طور کلی در مورد تجزیه و تحلیل داده‌ها نوشته و یا بحث می‌شود، خیلی مهم است که تفاوت بین هوش کسب و کاری[۱۴] و علم داده‌ها[۱۵] را بدانیم. یکی از راه‌های ارزیابی نوع تجزیه و تحلیل در حال انجام، بررسی افق زمانی و روش تحلیل به کار گرفته شده می‌باشد. هوش کسب و کار تمایل به ارائه گزارش، جعبه ابزار، و نمایش داده‌ها در مورد مسائل کنونی و یا در گذشته است. به عبارت دیگر، هوش کسب و کار به خودی خود یک فناوری نیست، بلکه مجموعه‌ای از سامانه‌های اطلاعاتی هستند که به طور هماهنگ کار می‌کنند که شامل سامانه‌هایی از قبیل سامانه‌های انبار داده‌ها، سامانه‌های داده کاوی، سامانه‌های پردازش تحلیلی برخط، سامانه‌های مبتنی بر دانش، ابزار پرس و جو و گزارش، و جعبه ابزار هستند (شکل ۲). علم داده ترکیبی از روش‌های علمی سنتی با توانایی مقایس‌بندی، کشف، یادگیری (پیش‌بینی را هم در بر دارد) و به دست آوردن بینش عمیق از داده‌های عظیم است. تعریف دیگری برای علم داده می‌شود که عبارت است: داده کاوی به عنوان یک پشتیبان از برنامه‌های کاربردی هوش کسب و کار در داده‌های عظیم [۳].



[۱۴] Business intelligence

[۱۵] Data science

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

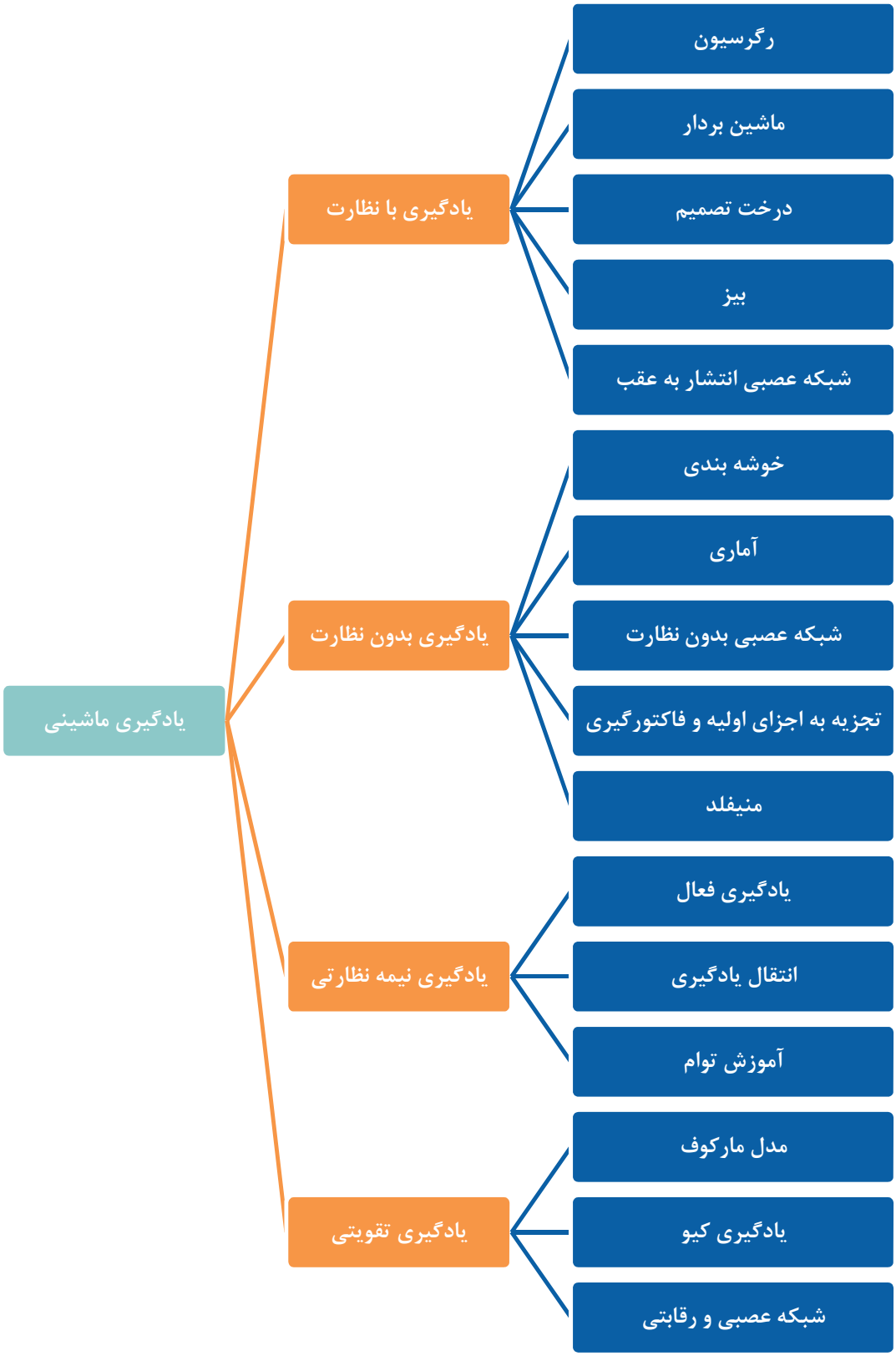
شکل ۲- ساختار هوش تجاری (کسب و کار)[۳]

این بخش یک رویکرد کلی به برخی از فن‌های کلیدی و ابزار مورد استفاده در تجزیه و تحلیل کلان داده‌ها را فراهم می‌کند. دانش روش‌های فراگیر، سازمان‌ها و افراد را به مشارکت و فعالیت در حوزه تجزیه و تحلیل پروژه‌های کلان داده‌ها تشویق خواهد کرد. محتوای این بخش برای کمک به ذینفعان متعدد: کسب و کار و تحلیل گران اطلاعات که به دنبال اضافه کردن تجزیه و تحلیل کلان داده‌ها به مجموعه کارهای خود می‌باشند؛ متخصصان پایگاه داده و مدیران کسب و کار هوشمند، تحلیل‌گران و یا گروه کلان داده‌ها به دنبال غنی‌سازی مهارت‌های تحلیلی خود هستند؛ و فارغ التحصیلان دانشگاهی در حال بررسی "علم داده‌ها" به عنوان یک رشته شغلی می‌باشند، خواهد بود.

۱-۲ گونه شناسی روش‌های تحلیل کلان داده‌ها

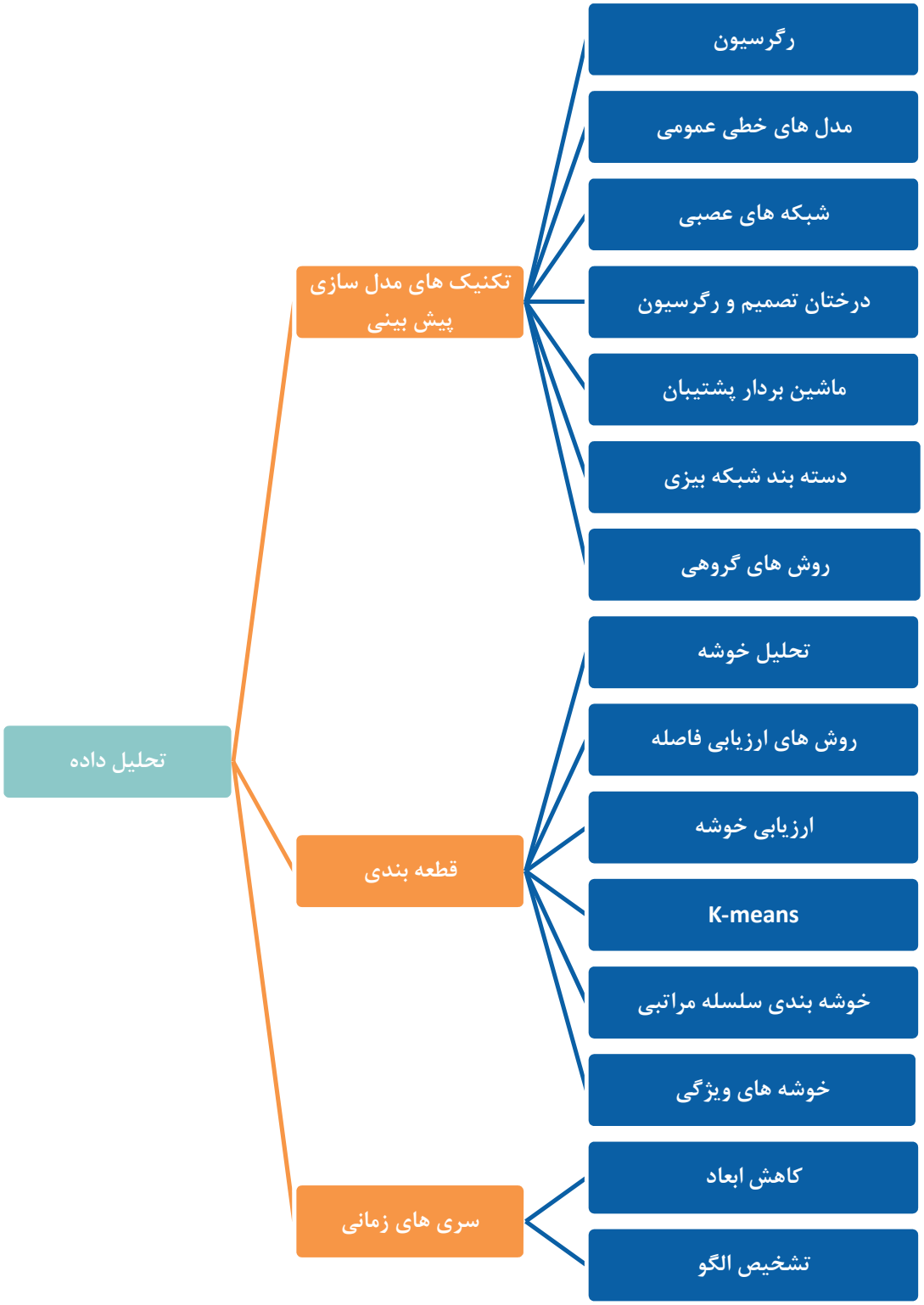
همانطور که ذکر شد روش‌های مختلفی جهت تحلیل کلان داده‌ها مورد استفاده محققان است. یکی از گونه شناسی‌های تحلیل کلان داده‌ها مبتنی بر یادگیری ماشین می‌باشد. به طورکلی، یادگیری ماشین توانایی یادگیری الگوها و استنباط از داده‌ها به صورت خودکار با استفاده از کامپیوتر تعریف می‌شود. روش‌های یادگیری ماشین به صورت خودکار و مقیاس پذیر اجازه بینش داده‌های حجیم و چند بعدی سپس تجمع نتایج را امکان پذیر می‌کنند. همانطور که در شکل۳ می‌بینید، این الگوریتم‌ها را می توان به گونه‌های مختلف دسته‌بندی کرد. همچنین گونه شناسی تحلیل کلان داده‌ها از دید علم داده‌ها در شکل ۴ آورده شده است [۱]، [۲] و [57].

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد



شکل ۳- گونه‌شناسی تحلیل کلان داده‌ها از منظر یادگیری ماشین

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد



شکل۴- گونه‌شناسی تحلیل کلان داده‌ها از منظر علم داده‌ها

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

۲-۱-۱ یادگیری با نظارت

یادگیری با نظارت یا یادگیری تحت نظارت[۱۶] یکی از زیر مجموعه‌های یادگیری ماشینی[۱۷] است. در این روش، به یک سیستم، مجموعه‌ای از زوج‌های ورودی– خروجی ارائه می‌شود و سیستم تلاش می‌کند تا تابعی با نگاشتی بین ورودی و خروجی را یاد بگیرد. هدف سیستم یادگیر، بدست آوردن فرضیه‌ای است که تابع و یا رابطه بین ورودی و یا خروجی را حدس بزند به این روش یادگیری با نظارت گفته می‌شود. یادگیری تحت نظارت نیازمند تعدادی داده ورودی به منظور آموزش[۱۸] سیستم است. قابل توجه است که در این روش، داده‌های آموزش دارای برچسب[۱۹]های انسانی می‌باشند[۴]. روش‌های با نظارت به دسته‌های متفاوتی تقسیم می‌شوند [۵] که در ادامه شرح هریک از روش‌ها با توضیح مختصری آمده است:

۲-۱-۱-۱ رگرسیون

روش تحلیل رگرسیون یک روش آماری است که اغلب برای پیش بینی‌های عددی استفاده می‌شود[۶]. در واقع از این روش برای تخمین روابط بین متغیرها و تحلیل متغیرهای خاص و منحصر بفرد استفاده می‌شود. وقتی که تمرکز بر روی روابط بین متغیر وابسته و یک یا چند متغیر مستقل باشد، تحلیل رگرسیون کمک می کند در فهم اینکه چگونه مقدار متغیر وابسته با تغییر هرکدام از متغیر مستقل و با ثابت بودن دیگر متغیرهای مستقل تغییر می‌کند. قابل توجه است که این روش در تحلیل کلان داده‌ها نیز مورد استفاده قرار می‌گیرد [۷] و توابع مربوط به آن در اسپارک[۲۰] موجود می‌باشد [۸].

---

[16] Supervised learning
[17] Machine Learning
[18] Training
[۱۹] label
[۲۰] Spark

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

۲-۱-۱-۲ طبقه بندی[۲۱]

طبقه‌بندی، مسئله‌ی تخصیص یک مشاهده به یک مجموعه از مشاهدات از قبل دیده شده است. در روش‌های طبقه‌بندی، مجموعه‌ای از داده‌ها جهت آموزش وجود دارند و براساس آن‌ها مشاهدات در دسته‌های مختلف طبقه‌بندی می‌شوند وقتی یک مشاهده جدید وارد می‌شود، براساس ویژگی‌هایش به یکی از دسته‌های از قبل تعریف شده، تخصیص می‌یابد[۶]. روش‌های یادگیری ماشین زیادی هستند که در گروه روش‌های طبقه‌بندی قرار دارند. چندین روش مورد استفاده در کلان داده‌ها در زیر آمده است:

روش بردار پشتیبان[۲۲]

روش بردار پشتیبان یکی از روش‌های طبقه بندی است که توسط وپنیک[۲۳] بر پایه‌ی تئوری یادگیری آماری بنا شد[۹]. این روش از یک نگاشت غیرخطی برای انتقال داده‌های آموزشی به یک فضا با ابعاد بالا استفاده می‌کند. در آن فضا با ابعاد بالا، به دنبال جستجوی ابرصفحه‌هائی با حداکثر حاشیه[۲۴] هست که بتوانند داده‌ها را جداسازی نمایند. از مزایای روش طبقه‌بندی از طریق بردار پشتیبان، می‌توان گفت آموزش آن تقریباً ساده است. در ماکزیمم‌های محلی متوقف نمی‌شود. مصالحه بین پیچیدگی طبقه‌بندی کننده و میزان خطا در این روش به طور واضح کنترل می شود. همچنین برای داده‌ها با ابعاد بالا و همچنین داده‌هایی که دارای توزیع مشخصی نیستند تقریباً خوب جواب می‌دهد. لذا در روش‌های تحلیل کلان داده‌ها مورد توجه محققان قرار گرفته است[۱۰]. همچنین این روش تعمیم-دهی خوبی برای نتایج دارد اما عیبش این است که پیچیدگی تبدیلات و ابرصفحه‌های که به عنوان خروجی می دهد درک و تفسیر بسیار مشکلی دارند. به همین دلیل از این روش همیشه برای طبقه‌بندی استفاده نمی‌شود[۱۱].

روش درخت‌های تصمیم[۲۵]

---

[۲۱] Classification

[22] Support Vector Machine (SVM)
[23] Vapnik
[24] Maximum Margin
[۲۵] Decision Trees

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

روش طبقه‌بندی براساس درخت‌های تصمیم، روش یادگیری از درخت‌های تصمیم توسط داده‌های برچسب‌گذاری شده است. در واقع یک درخت تصمیم یک فلوچارت با ساختار درختی است که هر گره غیر برگی آن نمایانگر یک ویژگی از داده‌ها و هر شاخه درخت نمایانگر مسیر خروجی و هر برگ برچسب کلاس را مشخص می‌کند و بالاترین گره نیز ریشه می‌باشد[۶]. از مزایای درخت‌های تصمیم سادگی، سرعت و دقت مناسب، امکان کارکردن با داده‌های بزرگ و پیچیده، قابلیت ترکیب با روش‌های دیگر و عدم نیاز به تنظیم پارامتر است. همچنین این روش، به طور خودکار عملیات انتخاب ویژگی را از طریق نودهای بالایی درخت انجام می‌دهد. از معایب این روش این است که به صورت نمایی با بزرگ شدن مسئله بزرگ می‌شوند و همچنین حافظه زیادی را نیاز دارند[۶]. با این حال از این روش‌ها در تحلیل کلان داده‌ها استفاده می‌شود [۱۲][۱۳][۱۴].

روش بیز[۲۶]

طبقه‌بندی کننده‌های بیزی، طبقه‌بندی کننده‌های آماری هستند که می‌توانند احتمال عضویت به یک کلاس را براساس داده‌هایی که مشاهده کرده‌اند، پیشگویی کنند. طبقه‌بندی کننده‌های بیزی براساس تئوری بیزی کار می‌کنند. ساده‌ترین نوع آن‌ها که طبقه‌بندی‌کننده‌های naive هستند از جهت کارایی با درخت‌های تصمیم و شبکه‌عصبی قابل رقابت هستند[۶]. بزرگترین ویژگی این روش این است که حجم آموزش اندکی برای شروع کار و تخمین پارامترها نیاز دارد. این روش دارای دقت و سرعت بالا در ارتباط با کار برروی مجموعه کلان داده‌ها است. لذا در تحلیل کلان داده‌ها از این روش استفاده می‌گردد[۱۵][۱۶]. در کتابخانه‌های اسپارک این روش پیاده‌سازی شده و توابع آن به سادگی موجود می‌باشد[۱۷].

شبکه های عصبی الگوریتم انتشار به عقب

یکی از الگوریتم‌های یادگیری شبکه عصبی، الگوریتم انتشار به عقب است. این الگوریتم فرض می‌کند که شبکه ساختاری ثابت و متناسب با یک گراف جهت دار دارد که ممکن است دور (حلقه) نیز داشته باشد. شبکه‌های چندلایه

[۲۶] Bayes

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

که با این الگوریتم آموزش داده می‌شوند می‌توانند انواع مختلفی از سطوح تصمیم‌گیری غیرخطی را نیز یاد بگیرند. به طورکلی استفاده از شبکه‌های عصبی به دلیل یادگیری تطبیقی، خودسازماندهی و تحمل بالای خطا دارای مزیت هستند و از این روش در تحلیل کلان داده‌ها نیز استفاده می‌شود[۱۸]. به عنوان مثال در پژوهشی که اخیرا در سال ۲۰۱۶ توسط Hodge و همکارانش صورت گرفت آن‌ها یک روش مبتنی بر تئوری را برای شبکه عصبی مبتنی بر هدوپ برای انتخاب ویژگی موازی و توزیع شده در مجموعه کلان داده‌ها ارائه دادند [۱۹].

۲-۱-۲ یادگیری بدون نظارت

یادگیری بدون نظارت یک نوع یادگیری ماشینی است که به منظور درک و توصیف ساختار پنهان داده‌های بدون برچسب مورد استفاده قرار می‌گیرد. در این شاخه از یادگیری ماشینی برخلاف روش‌های با نظارت و روش‌های یادگیری تقویتی، از آن‌جا که آموزش گیرنده، نمونه‌های بدون برچسب را مشاهده می‌کند، هیچ‌گونه خطا، جایزه یا سیگنال ارزیابی وجود نخواهد داشت. مرسوم‌ترین روش‌های یادگیری بدون نظارت، روش‌های خوشه‌بندی و تحلیل خوشه هستند که به کشف الگوهای پنهان داده‌ها می‌پردازند. چنانکه خوشه‌ها، بر اساس معیارهای شباهت نظیر فاصله اقلیدسی یا احتمالاتی مدل می‌شوند. یادگیری بدون نظارت در تحلیل کلان داده‌ها بسیار مورد استفاده قرار گرفته است. لذا از آن‌جا که خوشه‌بندی داده‌ها به کوچک‌سازی، مختصرسازی، همگن‌سازی و نهایتا" ساده‌سازی کلان داده‌ها می پردازد، روش‌های بدون نظارت مرسوم در تحلیل کلان داده‌ها در ادامه معرفی شده‌اند.

انواع روش های یادگیری بدون نظارت عبارتند از: روش‌های خوشه بندی، روش‌های آماری، مدل‌های مبتنی بر شبکه عصبی بدون نظارت، روش‌های تجزیه به اجزای اولیه و فاکتورگیری، روش‌های یادگیری بدون نظارت منیفلد. در ادامه الگوریتم‌های مختلفی از روش‌های فوق در حوزه کلان داده‌ها معرفی شده‌اند.

۲-۱-۲-۱ روش‌های خوشه‌بندی[۲۷]

روش K-means

[۲۷] Classification

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

الگوریتم K-means یکی از ساده‌ترین الگوریتم‌های یادگیری بدون نظارت است که مسائل مشهوری را حل می‌کند. ایده اصلی آن بدین شرح است که k مرکز خوشه تعیین شده و بر اساس فاصله سایر اعضاء به مرکز خوشه‌ها، عضویت هر یک از اعضا را به هر خوشه تعیین می‌کند، سپس مجدد با عضویت عناصر جدید مرکز خوشه تعیین می‌شود و عضویت یا عدم عضویت عناصر تعیین می‌شود. این رویه آنقدر تکرار می‌شود که مراکز خوشه‌ها تغییر مکان نداشته باشد یا تغییرات آن بسیار ناچیز باشد. الگوریتم K-means در تحلیل کلان داده‌ها نیز مورد استفاده قرار می‌گیرد. به عنوان مثال [۲۰] و [۲۱] کاربردهایی از این الگوریتم در مسائل کلان داده‌ها را نشان می‌دهد که اولی به توسعه یک روش خوشه بندی پایدار مبتنی بر k-means چندوجهی بر روی کلان داده‌ها پرداخته است و دومی به پیاده‌سازی این روش خوشه‌بندی بر بستر MapReduce پرداخته است که فرآیند توزیع بر روی چندین کامپیوتر و اجرای همزمان را در اختیار کاربران قرار می‌دهد.

روش Mean Shift

الگوریتم خوشه بندی Mean shift یک روش خوشه‌بندی انعطاف پذیر مبتنی بر مرکز خوشه است که براساس انتخاب و بروزرسانی مرکز خوشه از میان اعضای داوطلب انجام می‌پذیرد. این الگوریتم یکی از روش‌های بدون نظارت برای تحلیل کلان داده‌ها می‌باشد که به طور مثال در [۲۲] مورد استفاده قرار گرفته است و با بهبود سرعت اجرای الگوریتم Mean shift با random sampling بر روی کلان داده‌ها، بهبود ایجاد کرده است.

روش خوشه‌بندی طیفی[۲۸]

الگوریتم خوشه‌بندی طیفی یک تکنیک خوشه‌بندی است که با استفاده از طیف مقادیر ویژه ماتریس مشابهت داده‌ها، فرآیند خوشه‌بندی و کاهش ابعاد داده‌ها را انجام می‌دهد. ماتریس مشابهت بعنوان ورودی این الگوریتم می‌باشد که شامل ارزیابی کمی شباهت بین هر جفت عضو داده‌ها می‌باشد. این الگوریتم به دلیل خوشه‌بندی توام با کاهش ابعاد، مورد توجه روش‌های تحلیلی کلان داده‌ها است. چنانکه [۲۳] از ویژگی این الگوریتم برای خوشه‌بندی داده‌های بزرگ بر روی گراف‌ها استفاده کرده است.

[۲۸] Spectral

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

روش خوشه‌بندی سلسله مراتبی[۲۹]

به تکنیک‌هایی که مبتنی بر خوشه‌بندی‌های دارای ترتیب از پیش تعیین شده می‌باشند خوشه‌بندی سلسله مراتبی گفته می‌شود که از تکنیک‌های بسیار متداول در تحلیل کلان داده‌ها به شمار می‌آید. الگوریتم‌های فراوانی مبتنی بر خوشه‌بندی سلسله مراتبی وجود دارد نظیر روش‌های Agglomerative که دارای مرتبه زمانی $On(n^3)$ می‌باشد یا روش‌های مبتنی بر نزدیکترین همسایه نظیر Nearest neighbor chain که می‌تواند تا مرتبه زمانی $O(n^2)$ کاهش زمان مصرفی داشته باشد با این حال [۲۴] روش Nearest neighbor boundary (NNB) را برای خوشه‌بندی کلان داده‌ها مطرح کرد که مرتبه زمانی را می‌تواند تا $O(m \log^2 n)$ کاهش دهد.

روش مدل مخفی مارکف[۳۰]

مدل مخفی مارکف یک روش آماری مارکف است که به منظور خوشه‌بندی نیز مورد استفاده قرار می‌گیرد. به عنوان مثال [۲۵] مجموعه‌ای از داده‌های مدل مارکف را به گروه‌های مارکف خوشه بندی کرده است چنانکه آن‌ها به لحاظ توزیع بازنمایی داده‌ها با یکدیگر مشابه هستند. این روش خوشه‌بندی در [۲۶] به منظور استفاده بر روی کلان داده‌ها فراهم شده است.

روش خوشه‌بندی فازی[۳۱] (خوشه‌بندی نرم[۳۲])

خوشه‌بندی‌های مبتنی بر منطق فازی دارای قابلیت انعطاف پذیری بیشتری هستند و اجازه می‌دهند عضویت هر داده به خوشه‌های مختلف بر اساس مقدار عضویت بیان شوند. مقاله مروری [۲۷] به بیان انواع روش‌های خوشه‌بندی فازی پرداخته است و مقایسه‌ای میان خوشه‌بندی‌های فازی و روش‌های کلاسیک ارائه کرده است همچنین روش خوشه‌بندی مبتنی بر Fuzzy type-2 clustering را معرفی کرده است.

روش DBSCAN

---

[۲۹] Hierarchical
[۳۰] Hidden Markov models
[۳۱] Fuzzy clustering
[۳۲] Soft clustering

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

یکی دیگر از الگوریتم‌های خوشه‌بندی DBSCAN می‌باشد که یک تکنیک مبتنی بر چگالی می‌باشد و در مقایسه با الگوریتم‌هایی نظیر k-means دارای مزیت‌هایی می‌باشد. به عنوان مثال این روش نسبت به ساختار هندسی داده‌ها حساس نمی‌باشد و در نتیجه داده‌ها می‌توانند در ساختاری غیر هندسی در یک خوشه قرار گیرند. این روش خوشه-بندی جزء یکی از روش‌های موفق در تحلیل کلان داده‌ها است. چنانکه [۲۸] از آن برای خوشه‌بندی کلان داده‌ها در اندازه تریلیون استفاده کرده است. همچنین [۲۹] مطالعه مروری بر عملکرد خوشه‌بندها در کلان داده‌ها داشته است که DBSCAN یکی از روش‌های مورد ارزیابی در این مقاله می‌باشد.

روش Brich

یکی از الگوریتم‌های بسیار کارآمد در خوشه‌بندی روش بدون نظارت BIRCH می‌باشد. این روش که یک الگوی تکراری و سلسله مراتبی می‌باشد، برای خوشه‌بندی داده‌ها خصوصا" کلان داده‌ها مورد استفاده قرار می‌گیرد. در این روش، هر نمونه جدید به ابتدای درخت سلسله مراتبی خوشه‌بندی داده‌ها اضافه می‌شود، سپس نمونه جدید به نزدیکترین مرکز خوشه که بیشترین شباهت را با نمونه جدید دارد ملحق می‌شود. ترکیب این الگوریتم با GPU به همراه ویژگی-های پویای موازی پلتفرم CUDA منجر به کسب بهترین نتایج خوشه بندی شده است که توسط [۳۰] گزارش شده است.

۲-۱-۲-۲ روش‌های آماری

Method of moments

روش Method of moments یک روش آماری است برای تخمین مقادیر پارامترهای یک توزیع احتمال که نمونه‌هایی از آن مشاهده شده است. در این روش، تخمین گشتاورهای توزیع احتمال با مقدار نظری گشتاورها (که تابعی از پارامترها هستند) برابر قرار داده شده و مقدار پارامترها تخمین‌زده می‌شوند. متد گشتاورها در [۳۱] به منظور تضمین یادگیری مقادیر زیادی از متغیرهای پنهان مورد استفاده قرار گرفته است.

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

تخمین کواریانس[۳۳]

در بسیاری از مسائل آماری نیاز به تخمین ماتریس کواریانس از طریق پارامترها می‌باشد. با تخمین کواریانس می‌توان استنتاج‌های آماری را بر روی داده‌ها استخراج کرد. به عنوان مثال [۳۲] به طرح چالش‌ها و ارائه تحلیل‌های آماری از جمله تخمین کواریانس بر کلان داده‌ها پرداخته و مثال‌های کاربردی از تخمین کواریانس در کلان داده‌ها ارائه کرده است.

تخمین چگالی[۳۴]

برآورد چگالی عبارت است از ایجاد یک تخمین بر اساس داده‌های مشاهده شده مبتنی بر تابع چگالی احتمال (PDF) غیرقابل مشاهده. [۳۳] با استفاده از برآورد چگالی به ارائه یک مدل جامع مبتنی بر چگالی پرداخته و کلان داده‌ها حوزه تصاویر پزشکی را تحلیل کرده است.

مدل Gaussian Mixture

مدل های مخلوط، مدل‌های احتمالاتی هستند که برای بازنمایش حضور زیرگروه‌های داده‌ها در یک گروه کلی مورد استفاده قرار می‌گیرند بدون نیاز به دانستن آنکه زیر گروه مربوطه متعلق به کدام مشاهده می‌باشد. در [۳۴] از ویژگی‌های مدل گاوسی مخلوط بصورت ترکیبی برای توصیف ویژگی‌های کلان داده‌ها استفاده شده است.

Variational Bayesian Gaussian Mixture

مدل‌های تغییراتی بیزی از روش‌های یادگیری مبتنی بر توزیع پسین (Posterior distribution) است که به دلیل نیاز به محاسبات انتگرالی در توزیع پسین و مشکل بودن آن، روش‌های تغییراتی به جای آن استفاده می‌شود. استفاده از خانواده روش‌های بیزی در تحلیل کلان داده‌ها متداول است. چنانکه [۳۵] با توجه به ضرورت نیاز به توزیع

[۳۳] Covariance estimation

[۳۴] Density Estimation

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

پردازش‌ها و داده‌ها بر روی چندین ماشین، با فراهم کردن تخمین توزیع شده بیزی به تحلیل با کمترین هزینه پرداخته است.

۳-۲-۱-۲ مدل‌های مبتنی بر شبکه عصبی بدون نظارت

ماشین‌های بولتزمن انحصاری

ماشین های انحصاری بولتزمن (RBM) یک شبکه عصبی مصنوعی مولد تصادفی می‌باشد که می‌تواند توزیع احتمال موجود بر روی داده‌های ورودی خود را آموزش ببیند. ماشین های انحصاری بولتزمن یک مدل یادگیری بدون نظارت است که می‌تواند به طور چشمگیری کارآیی برنامه‌های کاربردی را بهبود بخشد. اگر چه پردازش کلان داده‌ها به وسیله ماشین های انحصاری بولتزمن چالش برانگیز است ولی [۳۶] و [۳۷] راهکارهایی را برای به کارگیری ماشین های انحصاری بولتزمن بر روی کلان داده‌ها ارائه کرده‌اند.

نقشه‌های خودسازمانده

نقشه‌های خودسازمانده یکی دیگر از روش‌های یادگیری بدون نظارت مبتنی بر شبکه عصبی است که از نمونه‌های آموزشی ورودی، داده‌هایی با ابعاد کوچکتر و متمایز شده‌تر تولید می‌کند که به آن نقشه (Map) گفته می‌شود. به دلیل کاهش ابعاد، در تحلیل کلان داده‌ها مورد علاقه می‌باشد چنانکه [۳۸] به کمک آن یک مکانیزم پردازش کلان داده‌ها بهینه در هدوپ ارائه کرده است.

نظریه تشدید انطباقی

نظریه تشدید انطباقی[۳۵] (ART) یک مدل یادگیری بدون نظارت می باشد که مبتنی بر چگونگی پردازش اطلاعات توسط مغز مطرح شده است. این سیستم معمولاً از یک فیلد مقایسه‌ای و یک فیلد تشخیص(ساخته شده از تعدادی نورون)، یک پارامتر مراقب (vigilance parameter) و یک واحد بازنشانی (Reset Module) ایجاد شده است. پارامتر مراقب معمولاً تأثیر قابل توجه‌ای بر روی سیستم دارد: چنانچه مقدار این پارامتر بزرگ در نظر گرفته شود حافظه‌هایی

[۳۵] Adaptive resonance theory

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

با جزئیات بالا (طبقه‌های ریز اما زیاد) در اختیار قرار می دهد و برعکس، اگر این مقدار کوچک باشد، حافظه‌هایی معمولی (طبقه‌هایی با اندازه معمولی و تعداد کم) بدست می دهد. فیلد مقایسه ای، یک حامل ورودی (آرایه‌ای یک بعدی از مقادیر) را دریافت کرده و آن را به بهترین همتایش در فیلد تشخیص منتقل می نماید. این بهترین همتا، نورونی تنهاست که مجموعه وزن هایش (Weight Vector)، با حامل ورودی بیشترین تطابق را داشته باشد. هر نورون در فیلد تشخیص یک سیگنال منفی (متناسب با کیفیت تطابق حامل ورودی با نورون دریافت کننده آن) به دیگر نورون های این فیلد ارسال می نماید، در نتیجه از تولید خروجی در آن ها جلوگیری می شود. با این کار، فیلد تشخیص روش منع جانبی را ارائه می کند. به این معنی که نورون‌های داخل آن، به عنوان طبقاتی (categories) عمل می‌کند که حامل‌های ورودی بر اساس آن‌ها دسته بندی می شوند. مرجع [۳۹] با ترکیب این روش و منطق فازی به خوشه بندی کلان داده‌ها پرداخته است.

۴-۲-۱-۲ روش‌های تجزیه به اجزای اولیه و فاکتورگیری

تحلیل مولفه‌ی اصلی[۳۶]

PCA ، یک روش تحلیل برای تجزیه داده‌ها به اجزای اولیه می‌باشد که به منظور کاهش ابعاد داده‌ها مورد کاربرد قرار می‌گیرد. از آنجا که یکی از چالش‌های کلان داده‌ها بزرگی ابعاد داده است روش‌های زیادی مانند [۴۰] و [۴۱] به منظور کاهش ابعاد داده‌ها از این الگوریتم بهره‌مند شده‌اند. چنانکه اولی با تخمین اجزای اولیه تنک (Sparse PCs) یک روش بهینه به منظور پردازش موازی با کمک پردازنده گرافیکی ارائه کرده است و دومی به ارائه روشی بهینه برای مساله ابعاد در کلان داده‌ها به کمک PCA پرداخته است.

تجزیه داده‌ها به مقادیر منفرد

[۳۶] Principal component analysis (PCA)

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

تجزیه داده‌ها به مقادیر منفرد یک روش فاکتورگیری از ماتریس داده‌های حقیقی یا مختلط بوسیله جبر خطی است. بطور کلی روش‌های فاکتورگیری و مختصرسازی مورد علاقه تحلیل کلان داده‌ها می‌باشد که بطور مثال [۴۲] با محاسبه چندین مقدار منفرد و بردارهای منفرد مرتبط‌شان کلان داده‌ها را به اندازه مختصر شده تبدیل رده است.

تحلیل عاملی

تحلیل عاملی[۳۷] یک روش مفید برای ارزیابی ارتباطات متغیرها در یک فضای پیچیده است که به توصیف تنوع میان مشاهدات و همچنین توصیف متغیرهای همبسته از نظر کمینه بودن تعداد متغیرهای مشاهده نشده می‌پردازد[۴۳]. یک روش موازی تحلیل عاملی برای کلان داده‌ها ارائه کرد که قادر است به محاسبات هر زیر تنسور بپردازد و سپس زیرعامل‌ها را با یکدیگر ترکیب کند.

تحلیل مولفه‌های مستقل

تحلیل مولفه‌های مستقل (ICA)، یک روش محاسباتی برای جداسازی یک سیگنال چند متغیره به خرده مولفه‌های ترکیب شدنی است. از آنجا که ICA ماهیت یک مساله پیچیده را به زیرمسائل کوچکتر می‌شکند دارای کاربردهای زیادی در تحلیل کلان داده‌ها است. به عنوان مثال [۴۴] با استفاده از تحلیل مولفه‌های مستقل و ترکیب آن با روش خوشه‌بندی سلسله مراتبی یک ایده جدید برای نمایش کلان داده‌ها با ابعاد کوچکتر و تنک ارائه کرده است.

فاکتورگیری نامنفی ماتریس

فاکتورگیری نامنفی ماتریس (NMF)، روشی است برای تجزیه ماتریس به دو ماتریس که ماتریس‌های تجزیه شده غیرمنفی هستند. از آنجا که مسئله ساده‌سازی در کلان داده‌ها همواره مورد توجه می‌باشد. فاکتورگیری نامنفی ماتریس مورد توجه تحلیل کلان داده‌ها می‌باشد. چنانکه از آن جمله می‌توان پیاده‌سازی آن برای تحلیل کلان داده‌ها در MapReduce اشاره کرد [۴۵].

تخصیص پنهان دیریکله

---

۳۷ Factor Analysis

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

تخصیص پنهان دیریکله (LDA) یک مدل جامع آماری برای مدل‌سازی متغیرهای پنهان است چنان‌که اجازه می‌دهد مجموعه‌های مشاهدات بوسیله گروه‌های پنهان (مشاهده نشده) که توضیح می‌دهند چرا بخش‌هایی از داده‌ها مشابه هستند توصیف شود. مشابه سایر روش‌های تجزیه داده‌ها، این روش نیز مورد توجه تحلیل کلان داده‌ها می‌باشد. به عنوان مثال [۴۶] با استفاده از LDA به قطعه‌بندی کلان داده‌ها یک شهر به P×P قطعه پرداخت چنانکه قطعات از هرگونه همپوشانی اجتناب شده‌اند. سپس هر یک از زیرقطعات بصورت پردازش موازی توسط اسپارک مورد تحلیل و پردازش قرار گرفته‌اند.

۵-۲-۱-۲ روش‌های یادگیری بدون نظارت منیفلد

روش Isomap

مشابه سایر روش‌های کاهش ابعاد، یک روش غیرخطی برای کاهش ابعاد داده است چنانکه یک روش ساده برای تخمین هندسه ذاتی داده‌های بسیار بزرگ بر اساس یک برآورد تقریبی از همسایگان هر داده در کلان داده‌ها فراهم می‌کند [۴۷].

روش Locally Linear Embedding

یک الگوریتم یادگیری بدون نظارت می‌باشد که ابعاد کوچک، با حفظ همسایگی آن در کلان داده‌ها را محاسبه می‌کند از آنجا که اصولا یک روش کاهش ابعاد می‌باشد مورد توجه روش‌های تحلیل کلان داده‌ها بوده و گزارش‌های زیادی نظیر [۴۸] برای آن ارائه شده است چنانکه در این گزارش به ارائه روش‌های کاهش ابعاد بصورت تحلیل خطی و غیرخطی پرداخته است.

روش Multi-dimensional Scaling (MDS)

نمایش کلان داده‌ها یکی از چالش‌های فعال در این حوزه می‌باشد MDS یک روش چند منظوره با کاربرد نمایش کلان داده‌ها می‌باشد که به ناظران اجازه می‌دهد به مشاهده ظریف بین روابط در مجموعه داده بپردازند[۴۹]. با ترکیب MDS و بردار یادگیری به ارائه روشی برای نمایش کلان داده‌هام پرداخته است.

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

روش Stochastic Neighbor Embedding

همسایه تصادفی کدهای جاسازی شده (SNE) تلاش می کند داده‌ها را در فضای با ابعاد کوچک قرار دهد و با توجه به بهینگی حفظ همسایگی داده‌ها، می‌تواند اجازه دهد داده‌ها به چندین زیرمجموعه داده با ابعاد کوچک توسعه یابد. [۵۰] با بکارگیری SNE به ارائه راه حلی برای کلان داده‌ها پرداخته است که به دلیل حجم داده‌ها بطور مرسوم قایل حل نمی‌باشد ولی با بکارگیری SNE داده‌ها را به خوشه‌هایی کوچکتر تقسیم کرده و با پیچیدگی زمانی $O(N \log^2 N)$ آن را حل کرده است.

۲-۱-۳ یادگیری تقویتی

در این بخش ابتدا اصول اولیه یادگیری تقویتی مورد بررسی قرار گرفته و سپس در رابطه با دو روش مهم این حوزه یعنی مدل تصمیم گیری مارکف و یادگیری Q به طور مختصر توضیحاتی داده می‌شود. در ادامه، نمونه‌هایی از استفاده موفق این روش‌ها در حوزه فناوری کلان داده‌ها مطرح می‌گردند. بر اساس مطالعات انجام شده، بطور کلی روش‌های مختلف یادگیری تقویتی با استفاده از بسترهای مناسب توزیع پذیر و انجام تغییرات لازم در الگوریتم‌های موجود، می‌توانند در حوزه کلان داده‌ها مورد استفاده قرار گیرند.

ایده اصلی یادگیری تقویتی[۳۸]، تشویق عامل یادگیری درقبال انجام کارهای صحیح و تنبیه او در قبال انجام کارهای نادرست است [۵۱]. بعبارت دیگر، یادگیری تقویتی، فرایندی یادگیرنده بر اساس انجام سعی و خطا می باشد. عامل یادگیری بر اساس شرایط محیط، تصمیم به انجام عملیاتی می گیرد. بر اساس میزان مطلوبیت نتایج حاصل، عامل یادگیرنده بر اساس میزان موفقیت هر کدام از فعالیت ها با توجه به شرایط محیط بر دانش خود می افزاید. به عبارت دقیق تر، عامل یادگیرنده با تعامل با محیط به آموخته های خود اضافه می‌کند.

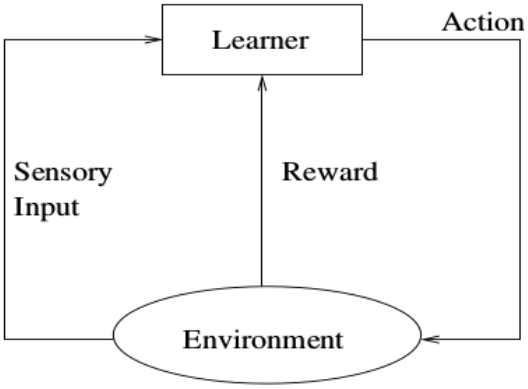
[38] Reinforcement Learning

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

۱-۳-۱-۲ یادگیری بر اساس تشویق و تنبیه[۳۹]

یادگیری تقویتی، یادگیری تابع نگاشتی از شرایط به فعالیت‌هایی است که هدف اصلی آنها افزایش مقدار پاداش‌ها یا سیگنال‌های تقویتی می‌باشد[۵۱]. بعبارت ساده‌تر، یادگیری تقویتی به عنوان روشی آموزشی بر اساس انجام سعی و خطا و بر اساس بازخورد بدست آمده از محیط (یا ارزیابی خارجی) تعریف می‌گردد. عامل یادگیرنده در ابتدا هیچ دانشی در رابطه با کاری که می‌خواهد انجام دهد، نداشته و حتی قادر به انجام پیش بینی در مورد میزان پاداش احتمالی مرتبط با هر عمل خاص نیز نیست.

نمونه‌ای از مسئله یادگیری تقویتی در شکل ۵ نمایش داده شده است. همان‌طور که در شکل نمایش داده شده، عامل یادگیرنده ابتدا سیگنال‌هایی از حسگرهای محیط دریافت کرده که نشان دهنده وضعیت کنونی آن می‌باشند. در ادامه عامل یادگیرنده فعالیتی انجام داده و در قبال آن، سیگنالی تقویتی یا بازخوردی دریافت می‌کند. این سیگنال با توجه به میزان درستی عمل انجام شده، می‌تواند مثبت یا منفی باشد. سیگنال منفی نشانگر تنبیه شدن عامل یادگیرنده بدلیل انجام عملی نادرست است. شایان ذکر است که اعمال انجام شده ممکن است باعث تغییر محیط شده که انتخاب‌ها و عملیات آتی عامل یادگیرنده را تحت تاثیر قرار می‌دهند و در نتیجه، عامل یادگیرنده همواره باید تغییرات محیط اطراف خود را مد نظر قرار دهد.



شکل۵- ساختار کلی روش یادگیری تقویتی

[39] Learning through Awards

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

در هنگام استفاده از روش یادگیری تقویتی، یکی از مهمترین نکاتی که باید در نظر گرفته شود، حفظ تعادل بین جستجوهای به عمل آمده و هزینه آن جستجوها می‌باشد. از این نظر، یادگیری تقویتی دارای دو قسمت مهم زیر است:

- انجام جستجو بر اساس سعی و خطا برای مشخص کردن فعالیت‌های جدید قابل اجرا که مستحق دریافت پاداش باشند. فعالیت‌های انجام شده در این راستا، مرتبط با قسمت اکتشاف الگوریتم[۴۰] هستند.
- استفاده از فعالیت‌هایی که قبلا نتایج مثبتی داشته‌اند و عملکرد مطلوب آن‌ها در حافظه سامانه وجود دارد. این عملیات در راستای قسمت بهره‌برداری از تجربیات گذشته[۴۱] می‌باشند.

استفاده از فعالیت‌هایی که قبلا نتیجه مثبتی در پی داشته و پاداش‌هایی را نسیب عامل یادگیرنده کرده‌اند، از اهمیت زیادی در یادگیری تقویتی برخوردار می‌باشد، ولی به هر ترتیب، عامل یادگیرنده ناگزیر است که با استفاده از انجام جستجوهایی بر اساس سعی و خطا، تنوع فعالیت‌هایی را که قرار است در آینده انجام دهد را بهبود بخشد.

یک عامل یادگیرنده تقویتی دارای اجزا زیر می‌باشد[۵۱]:

- **سیاست اجرایی**[۴۲]: عبارت است از تابع تصمیم گیرنده برای عامل یادگیرنده. این تابع بر اساس شرایط متفاوتی که عامل یادگیرنده ممکن است با آن‌ها روبرو گردد، تصمیم می‌گیرد که چه عملی را انجام دهد. سیاست بکار گرفته شده در حالت کلی شامل مجموعه ای از ارتباط‌های میان شرایط محیط و فعالیت‌های مناسب آن شرایط، یا مجموعه‌ای از قوانین تحریک شونده نسبت به انگیزه‌های به وجود آمده در محیط می‌باشند.
- **تابع پاداش-تنبیه**[۴۳]: این تابع هدف عامل یادگیرنده را تعیین می‌کند. به عبارت دیگر این تابع براساس شرایط خاص موجود، مثبت یا منفی بودن فعالیت‌های انجام شده را مشخص می‌کند. خروجی این تابع

[۴۰] Exploration Component
[۴۱] Exploitation Component
[۴۲] Policy
[۴۳] Reward Function

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

آنی بوده و نشان دهنده شرایط کنونی محیط می‌باشد. هدف نهایی، بیشینه کردن پاداش‌های دریافت شده در زمانی طولانی خواهد بود.

- **تابع ارزش**[۴۴]: هدف اصلی این تابع مشخص کردن هدف در طولانی مدت است. از این تابع برای پیش‌بینی پاداش‌های آینده و نمایش نتیجه بخش بودن فعالیت‌ها در طولانی مدت استفاده می‌گردد.
- **مدل محیط**[۴۵]: ممکن است عامل یادگیرنده تقویتی، مدلی از محیط اطراف خود را نیز داشته باشد که از آن برای شبیه‌سازی محیط استفاده کند. این عمل با استفاده از تابع گذر[۴۶] که شرایط گذر از حالتی با حالت دیگر محیط را توصیف می‌کند، انجام می‌پذیرد.

۲-۳-۱-۲ فرآیندهای تصمیم گیری مارکف

فرایندهای تصمیم‌گیری مارکف یک چارچوب ریاضی بوده که برای مدل‌سازی تصمیم‌گیری در شرایطی که نتایج تا حدودی تصادفی و تا حدودی تحت کنترل عامل تصمیم گیرنده است، مورد استفاده قرار می‌گیرد[۵۱]. از این روش برای مطالعه طیف گسترده‌ای از مسائل که معمولا با استفاده از برنامه‌نویسی پویا و یادگیری رقابتی حل می‌شوند، استفاده می‌گردد. به عبارت دقیق‌تر، فرآیند تصمیم‌گیری مارکف، فرآیند کنترل تصادفی زمان گسسته بوده و در هر گام، هنگامی که فرآیند در یک حالت خاص بوده، عامل یادگیرنده فعالیتی را انجام داده و با توجه به نتایج حاصل از آن فعالیت، فرآیند به حالت جدیدی رفته و پاداشی نیز به عامل یادگیرنده تعلق گیرد.

[۴۴] Value Function
[۴۵] Model of the Environment
[۴۶] Transition Function

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

۲-۱-۳-۳ کیو- یادگیری

کیو-یادگیری[۴۷] یک روش مستقل از مدل[۴۸] یادگیری تقویتی بوده که برای تعیین سیاست بهینه انتخاب فعالیت‌ها برای یک فرایند تصمیم‌گیری مارکف مورد استفاده قرار می‌گیرد[۵۱]. این روش بر اساس آموزش تابع ارزش فعالیت[۴۹] که در نهایت، نتایج انجام فعالیت‌های مشخص را در شرایط محیط با سیاست‌های بهینه وفق می‌دهد، عمل می‌کند.

۲-۱-۳-۴ ترکیب شبکه‌های عصبی و یادگیری رقابتی

شبکه‌های عصبی و یادگیری رقابتی با روش‌های مختلف می‌توانند با یکدیگر ترکیب شوند[۵۱]. یکی از روش‌های ممکن استفاده از شبکه عصبی برای تقریب زدن تابع ارزش می باشد. با استفاده از این تابع مقدار پاداش یا تنبیه مرتبط با عملیات انجام شده، پیشگویی می گردد. برای این منظور می توان از روش دیگری نیز استفاده کرد که در آن از یادگیری رقابتی برای تنظیم وزن های شبکه های عصبی استفاده می شود.

۲-۱-۳-۵ کاربردهای یادگیری تقویتی در حوزه کلان داده‌ها

در حوزه فناوری کلان داده‌ها، از روش‌های یادگیری تقویتی در تحلیل کلان داده‌ها استفاده گردیده است که در ادامه این گزارش به تعدادی از موفق ترین آنها اشاره می گردد:

- در یکی از نمونه‌های موفق استفاده از این روش در حوزه کلان داده‌ها، با موازی‌سازی الگوریتم یادگیری تقویتی با استفاده از زیرساخت رایانش موازی MapReduce، از یادگیری تقویتی برای تحلیل کلان داده‌ها استفاده شده است. در این روش نسخه‌هایی از الگوریتم‌های برنامه نویسی پویا با قابلیت پردازش موازی ارائه گردیده و همچنین الگوریتم یادگیری موازی با قابلیت پردازش موازی مطرح شده است. این الگوریتم، قابلیت تعامل با کلان داده‌ها را با استفاده از تابع تقریب خطی دارد. برای اثبات قابلیت الگوریتم-های پیشنهادی، آنالیزهای زمانی و مکانی (نحوه استفاده الگوریتم از حافظه) مناسبی انجام گرفته که

[۴۷] Q-Learning
[۴۸] Model-Free
[۴۹] Action-Value Function

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

نتایج این ارزیابی‌ها در [۵۲] ارائه گردیده است. با در نظر گرفتن نتایج منتشر شده، امکان استفاده از الگوریتم یادگیری رقابتی با قابلیت پردازش موازی در حوزه کلان داده‌ها امکان‌پذیر به نظر می رسد.

- نمونه دیگری از استفاده الگوریتم‌های یادگیری تقویتی در حوزه کلان داده‌ها در [۵۳] معرفی گردیده است. در این روش، الگوریتم توزیع شده مستقل از مدل یادگیری رقابتی پیاده‌سازی گردیده است. به عبارت دقیق‌تر، در پردازنده مورد استفاده در این سامانه، به طور مجزا با تعامل با محیط، تصمیماتی اتخاذ کرده و مقدار تابع Q را که بین تمام پردازنده‌ها مشترک می‌باشد، به روز می‌کند. بعد از اتمام مرحله به‌روزرسانی، پردازنده‌ها با استفاده از تابع Q نهایی، تصمیمات جدیدی اتخاذ می‌کنند. از آنجایی که هر پردازنده را می‌توان به عنوان یک یادگیرنده مستقل در نظر گرفت که در حال تعامل اطلاعات با دیگر پردازنده‌ها می‌باشد، ساختار توزیع شده ذکر شده را می‌توان به عنوان روش یادگیری با چندین عامل در نظر گرفت. آزمایشات به عمل آمده نشان دهنده توان بالای این روش در حوزه کلان داده‌ها می‌باشد.
- در روش ارائه شده در [۵۴]، مدل یادگیری عمیق توزیع شده‌ای برای یادگیری مستقیم موفقیت‌آمیز سیاست‌های کنترلی مرتبط با داده‌های اخذ شده از حسگرهای متعدد با استفاده از الگوریتم یادگیری تقویتی معرفی گردیده است. این مدل بر پایه شبکه عمیق Q استوار بوده و شبکه عصبی مترادفی با این شبکه عمیق آموزش داده می‌شود. ورودی این شبکه اطلاعات خام دریافتی از حسگرها بوده و خروجی آن، تخمینی از تابع ارزش مرتبط با پاداش‌های آتی متناسب با عملیات انجام شده در محیط می باشد. برای آموزش توزیع شده شبکه عمیق Q ذکر شده، از زیرساخت نرم افزاری DistBelief تغییر یافته برای آموزش کارآمد عاملان یادگیرنده استفاده گردیده است. نتایج نشان می‌دهند که روش معرفی شده قابلیت مقیاس پذیری بالایی دارد و برای استفاده در حوزه کلان داده‌ها مناسب می‌باشد.

در این بخش ابتدا روش‌های یادگیری تقویتی به طور مختصر مورد بررسی قرار گرفته و سپس زیر شاخه‌های اصلی آن که شامل مدل تصمیم گیری مارکف و یادگیری Q می‌باشد، معرفی گردیده‌اند. سپس نمونه‌هایی از کاربردهای

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

موفق این الگوریتم‌ها در حوزه کلان داده‌ها تشریح گردیده‌اند. بر اساس نتایج منتشر شده، با استفاده از زیرساخت‌های مناسب توزیع شده و همچنین انجام تغییراتی در الگوریتم‌های مورد استفاده در راستای ایجاد قابلیت انجام محاسبات توزیع شده، این روش‌ها قابل استفاده در حوزه کلان داده‌ها بوده و می توان حجم بسیار زیادی از داده‌های بدون ساختار را با استفاده از این روش‌ها تحلیل کرد.

۴-۱-۲ روش‌های یادگیری نیمه نظارتی[۵۰]

همان‌طور که نام نشان می‌دهد، یادگیری نیمه نظارتی یک روش یادگیری ما بین یادگیری بدون نظارت و یادگیری با نظارت می‌باشد. در واقع، راهبردهای یادگیری نیمه نظارتی بر پایه گسترش یادگیری بدون نظارت و یا یادگیری با نظارت با شامل کردن اطلاعات اضافی معمولی از دیگر روش یادگیری به انجام می‌رسند. به عبارت دیگر، این دسته مقدار کمی از داده‌های برچسب‌دار استفاده کرده و این اطلاعات را با مجموعه بزرگی از داده‌های بدون برچسب، برای تقریب یک الگوریتم یادگیری مناسب، ترکیب می‌کند [۵۵][۵۶][۵۷]. این الگوریتم‌های به طور خاص برای جامعه داده‌های بزرگ جالب توجه است چون‌که مجموعه بسیار بزرگی از داده‌های بدون برچسب هستند که به خوبی توسط الگوریتم‌های یادگیری با نظارت سنتی مورد بهره‌برداری قرار نگرفته‌اند. این دسته از الگوریتم‌ها به سه گونه یادگیری-فعال[۵۱]، انتقال یادگیری[۵۲] و یا آموزش توام[۵۳] برای تجزیه و تحلیل کلان داده‌ها بکار گرفته شده‌اند [۵۵][۵۸].

یادگیری فعال: در بسیاری از کاربردها در دنیای واقعی، ما با چنین وضعیتی مواجه هستیم که داده‌ها بسیار فراوان اما برچسب کمیاب یا به دست آوردن آن گران است. غالبا، یادگیری از حجم انبوهی از داده‌های بدون برچسب دشوار و وقت‌گیر است. در یادگیری فعال تلاش می‌گردد، با انتخاب یک زیر مجموعه بسیار مهم و حیاتی برای برچسب زدن به این موضوع رسیدگی شود [۵۵][۵۶]. در این راه، یادگیرنده فعال با هدف دستیابی به دقت بالا با استفاده از چند نمونه برچسب ممکن، در نتیجه به حداقل رساندن هزینه به دست آوردن برچسب داده‌ها را دنبال می‌کند. این

---

۵۰ Semisupervised

51 Active Learning
52 Transfer learning
53 Co-training

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

طبقه‌بندی را می‌تواند با چند نمونه برچسب‌دار و از طریق پرس و جو، به جای یادگیری منفعل معمولی، و با عملکرد رضایت بخش به دست آورد. سه شیوه اصلی یادگیری فعال عبارتند از: سنتز پرس و جوی عضویت داشتن[۵۴]، نمونه بردار گزینشی مبتنی بر جریان[۵۵]، نمونه برداری مبتنی بر استخر[۵۶]. روش یادگیری فعال به طور گسترده در زمینه یادگیری ماشین مورد مطالعه قرار گرفته و در بسیاری از کاربردهای پردازش داده‌ها مانند طبقه‌بندی تصویر و شناسایی بیولوژیکی DNA استفاده می‌شود [۵۵].

انتقال یادگیری[۵۷]: در بسیاری از الگوریتم‌های یادگیری ماشین و داده‌کاوی فرض عمده بر این است که داده‌های آموزش و داده‌های مورد بررسی باید دارای فضای ویژگی و همچنین توزیع یکسان داشته باشند. درصورتی‌که، در بسیاری از کاربردهای واقعی، داده‌های انبوه بدست آمده از منابع گوناگون، عدم تجانس بسیار داده‌های جمع‌آوری شده، این فرضیه از بین می‌رود. برای مثال، ما گاهی اوقات یک کار طبقه‌بندی در یک حوزه مورد نظر داریم، ولی داده‌های آموزشی کافی در حوزه مورد علاقه دیگر در دسترس هستند که آن داده‌های دوم ممکن است از یک فضای ویژگی‌های مختلف و یا یک توزیع دیگری پیروی کنند. جهت غلبه بر این مسئله، انتقال یادگیری پیشنهاد شده است که اجازه می‌دهد تا حوزه، وظایف، و توزیع متفاوت باشد. این روش دانش را از یک یا چند منبع وظایف استخراج و به یک وظیفه هدف انتقال می‌دهد. در چنین مواردی، اگر انتقال دانش بدست آمده قبلی با موفقیت و هوشمندانه انجام گیرد، تا حد زیادی یادگیری را تسریع و با دوری جستن از برچسب زدن بسیار گران قیمت داده‌ها، حل مسائل جدید را بهبود می‌بخشد[۵۵][۵۷][۵۹].

بر اساس شرایط مختلف حوزه‌ها و وظایف بین منبع و هدف داده شده در جدول ۱، انتقال یادگیری به سه شیوه: inductive، transductive، و بدون نظارت انجام می‌گیرد. در مورد انتقال یادگیری inductive، وظایف منبع (مبدأ) و هدف (مقصد) متفاوت هستند، بدون توجه به آنکه حوزه‌های مبدا و مقصد همان است یا نه. در مقابل، در

[54] Membership query synthesis
[55] Stream-based selective sampling
[56] Pool-based sampling
[۵۷] Transfer learning

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

انتقال یادگیری transductive، حوزه هدف از حوزه منبع فرق می‌کند، در صورتی که وظایف آن‌ها یکسان هستند. در نهایت، در انتقال یادگیری بدون نظارت، وظیفه هدف با وظیفه منبع فرق دارد، ولی به آن مربوط است [۵۵][۵۹].

جدول ۱- روش‌های انتقال یادگیری و تنظیمات گوناگون مربوط به آن‌ها [۵۹].

Learning Settings		Source and Target Domains	Source and Target Tasks
Traditional Machine Learning		the same	the same
Transfer Learning	*Inductive Transfer Learning* / *Unsupervised Transfer Learning*	the same	different but related
		different but related	different but related
	Transductive Transfer Learning	different but related	the same

آموزش توأم: آموزش توأم یک روش یادگیری نیمه نظارتی است که نیاز به دو نما[۵۸] از داده‌ها دارد. در این روش، هر نمونه داده دو مجموعه از ویژگی‌ها که مختلف و مکمل یکدیگرند را برای توصیف یک واقعیت ارائه می‌دهد. در حالت ایده‌آل، دو مجموعه از ویژگی‌های هر نمونه داده، به طور مشروط از کلاس داده شده به آن مستقل هستند و کلاس هر نمونه داده از هرکدام مجموعه ویژگی به تنهایی و با دقت بالا قابل پیش بینی می‌باشند. آموزش توأم در سنجش کیفیت هوای شهری [۶۰] و دسته‌بندی رایانامه‌ها [۶۱] بکار گرفته شده است. نتایج حاصله از این مطالعات، مزیت بالای این روش را نسبت به دیگر روش طبقه‌بندی نشان می‌دهد.

۲-۱-۵ دیگر روش‌های یادگیری ماشین

۲-۱-۵-۱ الگوریتم های یادگیری عمیق[۵۹]

الگوریتم‌های یادگیری عمیق یک زیرمجموعه از الگوریتم‌های شبکه عصبی مصنوعی هستند، اما آن‌ها را اغلب به طور جداگانه، به دلیل رشد عظیم که در این زمینه وجود دارد، دسته‌بندی کرده‌اند. آن‌ها با ساختار بسیار بزرگتر و پیچیده‌تر به شبکه‌های عصبی و روش‌های یادگیری نیمه نظارتی، جایی که مجموعه داده‌های بزرگ حاوی بسیار کمی از داده‌های با برچسب می‌باشند، مربوط می‌شوند. مشهورترین الگوریتم‌های یادگیری عمیق عبارتند از:[۶۲][۶۳][۶۴].

[۵۸] View

[۵۹] Deep Learning Algorithms

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

- ماشین بولتزمن عمیق[۶۰] (DBM)
- شبکه‌های باور عمیق[۶۱] (DBN)
- شبکه‌های عصبی کانولوشن[۶۲] (CNN)
- انباشته افزارهای تبدیل- خودکار[۶۳]

۲-۵-۱-۲ الگوریتم‌های شبکه عصبی مصنوعی

عملکرد این الگوریتم‌ها الهام گرفته شده از شبکه‌های عصبی بیولوژیک مغز است که در کارهای طبقه‌بندی و رگرسیون استفاده می‌شوند. گونه‌های زیادی از این الگوریتم‌ها وجود دارند. مشهورترین الگوریتم‌ها شبکه‌عصبی مصنوعی عبارتند از:

- Perceptron
- Back-Propagation
- Hopfield Network
- Radial Basis Function Network (RBFN)

۲-۵-۱-۳ الگوریتم‌های مبتنی بر نمونه[۶۴]

در این نوع از الگوریتم ها، مدل یک پایگاه داده از داده‌های نمونه ساخته و از یک معیار مشابهت جهت پیدا کردن بهترین تطبیق و پیش‌بینی استفاده می‌کند. این دسته از الگوریتم گاهی اوقات به عنوان یادگیری مبتنی بر حافظه نیز نامیده می‌شوند. تمرکز اصلی این الگوریتم‌ها بازنمایش داده نمونه ذخیره شده و پیداکردن معیار مشابهت می‌باشد. مشهورترین الگوریتم‌ها از این دسته عبارتند از [۶۴][۶۵]:

- K- نزدیکترین همسایه (KNN)
- Learning Vector Quantization (LVQ)
- نقشه خود-سازمان (SOM)

---

[۶۰] Deep Boltzmann Machine
[۶۱] Deep Belief Networks
[۶۲] Convolutional Neural Network
[۶۳] Stacked Auto-Encoders
[۶۴] Instance Based Algorithms

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

- یادگیری با وزنه-محلی[۶۵] (LWL)

۲-۱-۵-۴ الگوریتم‌های کاهش ابعاد

این الگوریتم‌ها مشابه الگوریتم‌های خوشه‌بندی هستند که به دنبال بدست آوردن ساختار داده‌ها می‌باشند. این الگوریتم‌ها عمدتا بدون نظارت بوده و در پی خلاصه کردن داده‌ها با استفاده از حداقل اطلاعات هستند که می‌تواند برای نمایش داده‌ها ابعادی یا ساده کردن داده‌ها و سپس استفاده از آن‌ها در روش‌های یادگیری با نظارت مفید باشد. بسیاری از این روش‌ها را می توان برای طبقه بندی و رگرسیون استفاده کرد [۶۴][۶۵].

- تجزیه و تحلیل مولفه اصلی[۶۶] (PCA)
- همبستگی مولفه اصلی[۶۷] (PCR)
- همبستگی کمترین مربعات جزئی[۶۸] (PLSR)
- نقشه برداری[۶۹] Sammon
- تغییر مقیاس چند بعدی[۷۰] (MDS)
- Projection Pursuit
- تجزیه و تحلیل تفکیک خطی[۷۱] (LDA)
- تجزیه و تحلیل تفکیک مخلوط[۷۲] (MDA)
- تجزیه و تحلیل تفکیک درجه دوم[۷۳] (QDA)
- تجزیه و تحلیل تفکیک انعطاف پذیر[۷۴] (FDA)

[۶۵] Locally Weighted Learning
[۶۶] Principal Component Analysis
[۶۷] Principal Component Regression
[۶۸] Partial Least Squares Regression
[۶۹] Sammon Mapping
[۷۰] Multidimensional Scaling
[۷۱] Linear Discriminant Analysis
[۷۲] Mixture Discriminant Analysis
[۷۳] Quadratic Discriminant Analysis
[۷۴] Flexible Discriminant Analysis

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

۵-۵-۱-۲ الگوریتم‌های گروهی

روش گروه متشکل از مدل‌های ضعیف متعددی هستند که به طور مستقل آموزش دیده‌اند و برای رسیدن به پیش بینی کلی، پیش بینی‌های آن‌ها با هم ترکیب می‌شوند. این دسته از الگوریتم‌ها که بسیار قدرتمند هستند عبارتند از [۶۴][۶۵]:

- ارتقا[۷۵]
- ادغام راه اندازی‌شده (کیسه)[۷۶]
- AdaBoost
- تعمیم انباشته (آمیخته کردن)[۷۷]
- ماشین ارتقا گرادیان [۷۸] (GBM)
- درختان همبستگی ارتقا یافته گرادیانی[۷۹] (GBRT)
- جنگل تصادفی[۸۰]

[۷۵] Boosting
[۷۶] Bootstrapped Aggregation (Bagging)
[۷۷] Stacked Generalization (blending)
[۷۸] Gradient Boosting Machines
[۷۹] Gradient Boosted Regression Trees
[۸۰] Random Forest

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

۳ ابزارها و سکوهای یادگیری ماشینی

۳-۱ ابزارهای منبع باز برای یادگیری ماشین

SciKit Learn

• SciKit بسته‌های متعدد ریاضی و علوم را برای Python از طریق یک میز کار تعاملی یا جاسازی شده در نرم افزارهای دیگر فراهم می‌کند. این کیت منبع باز و آزاد تحت مجوز BSD در دسترس است [۶۶].

Shogun

• Shogun از قدیمی‌ترین کتابخانه‌های یادگیری ماشین است. اگرچه Shogun در C++ نوشته شده است، اما آن را می‌توان با Java1، Python1، C#، R، Octave، MATLAB و Ruby با استفاده از کتابخانه SWIG به آشکار استفاده کرد [۶۷].

Accord Framework / AForge.net

• Accord یک چارچوب یادگیری ماشین برای Net. است. این ابزار برای کاربردهایی مانند تشخیص تصویر و پردازش صوت استفاده می‌شود. همچنین مجموعه‌ای از الگوریتم‌ها را برای پردازش بینایی، یادگیری ماشین، شبکه‌های عصبی و درخت‌های تصمیم‌گیر فراهم می‌کند [۶۸][۶۹].

Mahout

• Mahout چارچوبی است که به بهترین صورت با Hadoop کار می‌کند، اما بسیاری از الگوریتم‌های آن می‌تواند در بیرون از Hadoop هم اجرا شود. با این حال، بسیاری از الگوریتم‌های آن، به جای چارچوب Spark، Map Reduce را پشتیبانی می‌کند [۷۰].

MLlib

• MLib کتابخانه یادگیری ماشین Apache برای Spark و Hadoop است که برای مقیاس و سرعت طراحی شده است. اگر چه به طور بومی جاوا را پشتیبانی می‌کند، کاربران Python می‌توانند با استفاده از کتابخانه NumPy با آن ارتباط برقرار کنند. MLlib همچنین می‌تواند در بالای Spark بدون Haddop مستقر شود [۷۱].

H2O

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

• الگوریتم‌های H2O از Oxdata بیشتر به سمت فرآیندهای کسب و کار مانند تقلب و یا پیش بینی روند آماده شده است. H2O قادر به تعامل در حالت مستقل با فروشگاه‌های HDFS در MapReduce یا EC2 است. به طور معمول با جاوا استفاده می‌شود اما اتصالات را برای Python، R و Scala نیز فراهم می‌کند [۷۲].

Cloudera Oryx

• Oryx1 از توزیع Cloudera Hadoop، در درجه اول برای Hadoop طراحی شده است. Oryx برای استقرار یادگیری ماشین بر روی داده‌های جریان زنده[۸۱]، توانمندسازی پروژه‌هایی مانند فیلترکردن بلادرنگ هرزنامه[۸۲] و موتورهای توصیه، طراحی شده است. نسخه جدید Oryx2، در حال انجام می‌باشد [۷۳].

GoLearn

• در درجه اول برای Google Go Language طراحی شد، GoLearn مجموعه‌ای جامع از کتابخانه-های یادگیری ماشین را فراهم می‌کند و به طور فزاینده‌ای در حال مشهورشدن است. طبق نظر توسعه دهنده آن،Stephen Witworth، هدف همراه کردن "سادگی" با "سفارش‌پذیری[۸۳]" است. سادگی از چگونگی بار شدن و دستکاری داده در کتابخانه می‌آید، چون که مطابق SciPy و R الگو شده است. سفارش‌پذیری نیز در ذات منبع باز بودن کتابخانه (از MIT مجوز داده شده است) وهم در چگونه برخی از ساختارهای داده می‌توانند به راحتی در یک کاربرد توسعه داده شوند، نهفته است.

Weka

• Weka دارای مجموعه‌ای از الگوریتم‌های یادگیری ماشین طراحی شده برای داده‌کاوی Java می‌باشدکه توسط دانشگاه نیوزیلند ساخته شده است. اگرچه، هدف آن به طور خاص برای Hadoop نیست، اما می‌توان آن را با Hadoop استفاده نمود [۷۴].

---

[۸۱] live streaming data

[۸۲] spam

[۸۳] customizability

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

CUDA-Convnet

• CUDA-Convnet یک کتابخانه یادگیری ماشین برای شبکه‌های عصبی، نوشته شده در C++ برای بهره‌برداری از فناوری پردازش Nvidia CUDA GPU، است. شبکه‌های عصبی بوجود آمده نیز می‌توانند به عنوان اشیاء ترشی‌گذاری شده Python [۸۴] (سلسه مراتب آماده سازی هر شی در پایتون برای تبدیل به جریان بایتی[۸۵]) ذخیره شوند و قابل دسترسی از Python هستند. در حال حاضر روی یک جانشین - CUDA-Convnet2 کار می‌شود[۷۵].

ConvNetJS

• ConvNetJS کتابخانه یادگیری ماشین شبکه‌های عصبی را برای استفاده در JavaScript فراهم می‌کند [۷۶].

۳-۲ سکوهای یادگیری ماشین ابری

سکو ی ابری Google

• یادگیری ماشین ابری گوگل، یک سکوی مدیریت شده است که شما را به آسانی توانمند به ساختن مدل‌های یادگیری ماشین می‌کند که با هر نوع و با هر اندازه داده کار کنید. مدل‌ها می‌توانند با چارچوب TensorFlow ایجاد گردند. مدل آموزش دیده بلافاصله برای استفاده با سکوی پیش‌بینی جهانی گوگل در دسترس است که می‌تواند هزاران نفر از کاربران و ترابایت داده را پشتیبانی کند. سکو با Google Cloud Dataflow برای پیش پردازش، اجازه دادن به شما برای دسترسی به داده‌ها از ذخیره‌ساز ابری گوگل، Google BigQuery یکپارچه است [۷۷].

یادگیری ماشین Amazon

[۸۴] Python pickled objects

[۸۵] Stream byte

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

• یادگیری ماشین آمازون سرویسی است که باعث به کارگیری آسان فناوری یادگیری ماشین برای توسعه‌دهندگان از تمام سطوح مهارت می‌گردد. یادگیری ماشین آمازون، ابزارهای نمایش و راهنمایی[۸۶] را فراهم می‌کند که شما را در فرآیند ایجاد مدل یادگیری ماشین (ML) بدون اجبار به یادگیری الگوریتم‌های پیچیده ML و فن‌آوری هدایت کنند. هنگامی که مدل شما آماده است، یادگیری ماشین آمازون به دست آوردن پیش‌بینی را برای کاربرد شما با استفاده از APIs[۸۷]، بدون نیاز به پیاده‌سازی کد معمولی پیش‌بینی، و یا مدیریت هیچ زیرساختی، آسان می‌کند [۷۸].

BigML

• BigML یک یادگیری ماشین API کاربرپسند و توسعه‌دهنده‌پسند است که به طور عمده روی درخت‌های تصمیم‌گیری تمرکز دارد. انگیزه BigML این است که تجزیه و تحلیل پیش بینی را آسان، قابل درک و جذاب برای کاربران کند. آن روی درک فرآیندهای کسب و کار و ساختن گزارش‌های تحلیلگر کاربر نهایی[۸۸] تمرکز دارد. BigML API، سه خط رابط فرمان- حالت[۸۹] مهم، رابط شبکه و یک Restful API فراهم می‌کند. رابط شبکه BigML ، با ویژگی‌هایی مانند تنها با یک کلیک و گالری قابل توجه است [۷۹].

IBM Watson

• API آی بی ام واتسون یک خدمات شناختی است که فرآیند آماده‌سازی داده را ساده و اجرا تجزیه و تحلیل پیش‌بینی را آسان‌تر می‌کند. همچنین، استفاده ابزار داستان‌سرایی بصری مانند اینفوگرافیک، نقشه‌ها و نمودار برای نشان دادن نتایج تجزیه و تحلیل‌ها را فراهم می‌کند. آی بی ام واتسون برای استفاده عمومی از طریق سکوی خدمات ابری Bluemix آی بی ام در دسترس است[۸۰].

Microsoft Azure

---

[۸۶] Wizards

[۸۷] Application program interface

[۸۸] end-user analyst making reports

[۸۹] modes – Command Line Interface

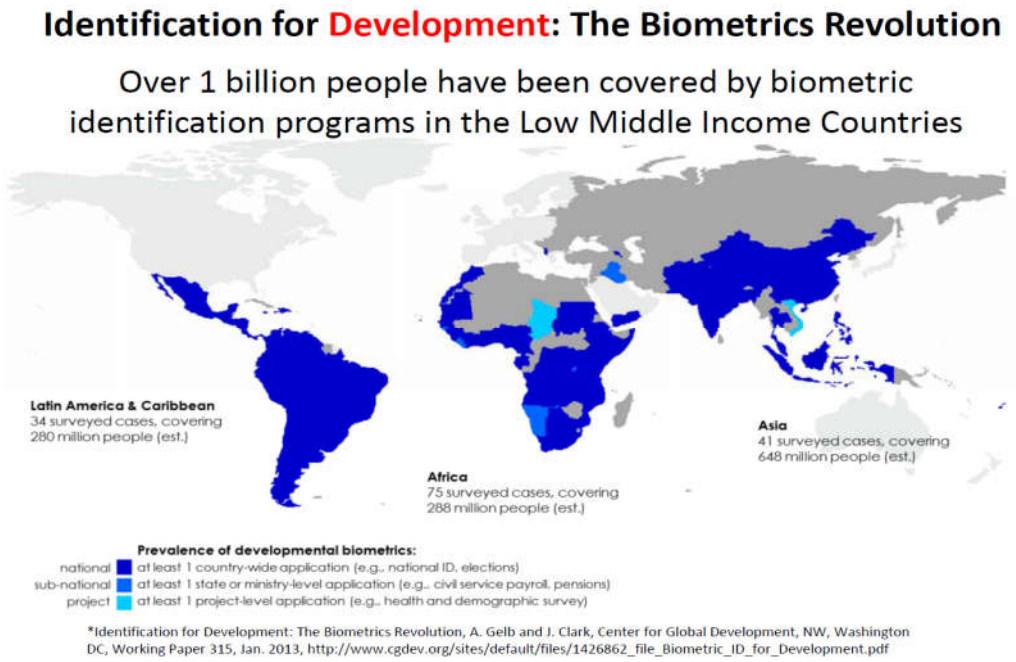
نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

• یادگیری ماشین Azure، استفاده از مدل‌های پیش‌بینی در کاربردهای اینترنت اشیا را با ارائه APIs برای تشخیص تقلب، تجزیه و تحلیل متن، سامانه‌های توصیه و موارد دیگر کسب و کار را برای متخصصین داده‌کاوی آسان می‌سازد. این API بر پایه توانایی‌های یادگیری ماشین ساخته شده است که در محصولات مایکروسافت مانند Bing و Xbox در دسترس هستند [۸۱].

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

۴ بایومتریک در کلان داده‌ها بعنوان یک کاربرد

بیش از یک قرن از توجه به موضوع هویت سنجی یا زیست‌سنجی (بایومتریک) انسان می‌گذرد و امروزه قریب به یک میلیارد نفر در سراسر جهان تحت پوشش هویت سنجی مبتنی بر خصوصیات بایومتریک قرار گرفته‌اند چنانکه می‌توان پراکندگی آنها را در شکل۶ مشاهده کرد. بطور خاص به فرآیند معیارسنجی و شناسایی انسان به کمک خصیصه‌های زیستی اندامها، زیست‌سنجی یا بایومتریک گفته می‌شود. معروفترین اندامهای بایومتریک عبارتند از: اثرانگشت، عنبیه چشم، چهره، دی اِن اِی، صدا، راه رفتن، شبکیه چشم، امضاء، گوش و غیره؛ که هر کدام دارای نقاط قوت و ضعف می‌باشند. بطورکلی اثباتی بر بهینه‌بودن هیچیک از این خصوصیات وجود ندارد ولی برخی از این خصیصه ها نسبت به سایر آنها برتری دارند چنانکه اثرانگشت، عنبیه چشم و چهره جزء مرسوم‌ترین و معتبرترین ویژگی های بایومتریک انسان تلقی شده و به کار می‌روند. عوامل متعددی در انتخاب یک خصیصه به منظور عملیات بایومتریک وجود دارد این عوامل ۸ گانه شامل: یکتایی، پایداری، جامعیت، کلکسیونی، کارآیی، کاربرپسند بودن، آسیب‌پذیری و ادغام-پذیری می باشند که برآیند آنها با توجه به نوع کاربرد در انتخاب بهترین خصیصه بایومتریک موثر است.



شکل۶ – پراکندگی پوشش هویت سنجی مبتنی بر خصیصه های بایومتریک

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

تحلیل بایومتریک بطور ذاتی با کلان داده‌ها گره خورده است چنانکه کارآیی و تاثیر آن در کلان داده‌ها واضح تر و موثرتر نمایانگر می‌شود. مهمترین مسائل مرتبط با بایومتریک خصوصا در کلان داده‌ها، مساله یکتایی[۹۰] و پایداری[۹۱] خصیصه های بایومتریک می‌باشد. بعنوان نمونه، هم اکنون استفاده از بایومتریک در سطح شناسایی افراد یک کشور پرجمعیت نظیر هند با جمعیتی بالغ بر ۱,۲ میلیارد نفر آغاز شده است [۸۲] که به هر سه خصیصه اثرانگشت، عنبیه و تصویر چهره توجه دارد. این موضوع اگرچه نیازمند بستری آماده، برنامه ریزی پیچیده و در عین حال بسیار دقیق است ولی مزایای آن بسیار چشمگیر می‌باشد که از آن جمله می توان به کشف جرم و تقلب، مسائل امنیتی، رسیدگی های مالی، آسایش عمومی، خدمات اجتماعی و غیره اشاره کرد.

مساله یکتایی، خصوصا در کلان داده‌ها بگونه‌ای به تحلیل ویژگی‌های هر یک از خصوصیت‌های بایومتریک می‌پردازد که در یک طیف گسترده (مثلا ۸۰ میلیون نفر جمعیت کشورمان) قادر به تفکیک و شناسایی صحیح نمونه مورد جستجو باشد. از طرف دیگر، مساله پایداری که اساسا نشات گرفته از ماهیت تغییرپذیر اندامهای زیستی انسان است در سیستم های بایومتریک باید چنان مدنظر قرار گیرد که سیستم، توانایی تشخیص و نادیده گرفتن تغییرات اندک را بگونه‌ای اعمال کند که شباهت خصوصیت‌های بایومتریک بیشتر از حد آستانه تمایز هر نمونه با سایر نمونه ها باشد. از این رو اصلی‌ترین چالش‌های بایومتریک در کلان داده‌ها عبارتند از: ۱) بازنمایی (توصیف) اطلاعات[۹۲]، بدین مفهوم که چه اطلاعاتی از خصیصه یا خصوصیات بایومتریک انسان؛ و به چه شکل استخراج و نمایش داده شود که بتواند مسائل بایومتریک در سطح کلان داده‌ها را حل یا بهترین نتیجه را حاصل نماید. ۲) معیار مشابهت[۹۳]، که به نوع سنجش و چگونگی تحلیل اطلاعات به منظور تشابه و تمایز اطلاعات استخراج شده از خصوصیات بایومتریک انسان توجه دارد. چنانکه

[۹۰] Uniqueness
[۹۱] Presistence
[۹۲] Representation
[۹۳] Similarity measure

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

شباهت بین نمونه های اطلاعات افراد مختلف، کم باشد؛ و در مقابل، شباهت نمونه های مختلف اطلاعات یک فرد، بسیار زیاد باشد.

وجود یک سیستم یکپارچه بایومتریک دقیق و با کیفیت بعنوان یک خدمت رسان، نه تنها می تواند از مزایای خدماتی متنوعی برخوردار باشد بلکه می تواند به ایجاد و توسعه کسب و کار، رونق اقتصادی و تجاری، انعطاف پذیری مالی و افزایش اعتماد در فرآیندهای روزمره منجر گردد. به عنوان نمونه، به دو مثال در این زمینه اشاره می شود. ۱) سیستمهای بانکداری که امروزه مبتنی بر هویت‌سنجی از طریق: کارت ملی (حضور فیزیکی)، امضاء (دسته چک، سفته و غیره)، رمز عبور (کارت‌های اعتباری و بانکی) و غیره می‌باشند برای هر یک از مشتریان خود این اطلاعات را ثبت و نگهداری می کنند که مستلزم هزینه‌های فراوانی است در حالیکه در آینده ای نزدیک، این اطلاعات برای هر فرد، در یک مرکزداده[۹۴] متمرکز ثبت و نگهداری خواهد شد و هر سیستم بانکداری بر اساس نیاز خود می‌تواند مشخصات فرد را از مرکز مربوطه پرس‌وجو و تایید یا عدم تایید دریافت کند. به این ترتیب نه تنها افزونگی اطلاعات ناشی از تکرار برای هر فرد به ازای هر حساب بانکی یا هر سیستم بانکداری حذف خواهد شد بلکه هزینه‌های سرسام آور نگهداری این اطلاعات به نسبت تعداد سیستم‌های متقاضی کاهش خواهد یافت؛ ضمن آنکه برقراری امنیت چنین سیستمی به مراتب ساده‌تر و کم هزینه‌تر از سیستم‌های غیرمتمرکز است. ۲) مراکز دولتی در آینده‌ای بسیار نزدیک با تکیه بر یک سیستم متمرکز هویت‌سنجی، به اطلاعات مورد نیاز شهروندان مجهز خواهند شد به این ترتیب خدمات بهداشتی و بیمه، خدمات رفاهی، یارانه، درآمدها و سایر اطلاعات مورد نیاز مراکز دولتی می تواند بصورت یک فرآیند پرسش و پاسخ بایومتریکی از سیستم مربوطه درخواست و دریافت شود بطوریکه هزینه‌های فرآیندهای جمع آوری اطلاعات (مثلا بیمه، وظیفه عمومی، بهداشت و غیره)، نگهداری این اطلاعات (امنیت) و حتی توسعه بستر مربوطه بطور چشمگیری کاهش یافته و سبب بهبود کارآیی می شود.

[۹۴] Data center

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

۴-۱ چالش های بایومتریک در کلان داده‌ها

همانند همه موضوعات قابل بحث در کلان‌داده‌ها، چالش‌های بایومتریک نیز از چهار منظر حجم[۹۵]، سرعت[۹۶]، صحت[۹۷] و تنوع[۹۸] قابل تحلیل و بررسی می باشد. تکنیک‌ها و متدهایی برای این چالش ها وجود دارد؛ متدهای فهرست محور[۹۹] که اجازه می دهند پردازش ها در زمانی سریع و نزدیک به عدد ثابت[۱۰۰] و بدون توجه به اندازه پایگاه اطلاعات بایومتریک انجام پذیرند. متدهایی که با ترکیب چندین خصیصه بایومتریک ریسک‌پذیری و در عین حال دقت را پشتیبانی و با حذف رکوردهای تکراری از صحت و درستی فرآیندهای بایومتریکی اطمینان کسب می‌کنند.

بطور کلی سیستم های بایومتریکی می‌توانند به دو شیوه مورد استفاده قرار گیرند: (الف) ۱ به ۱ که تصدیق هویت[۱۰۱] یا تایید هویت[۱۰۲] نامیده می شود و (ب) ۱ به N که به هویت سنجی[۱۰۳] یا تطابق[۱۰۴] مشهور است. سیستم‌های تصدیق هویت ساده‌تر هستند زیرا مشخصات فرد مورد جستجو با مشخصات شخص ادعا شده موجود است و فقط باید مشابهت یا عدم مشابهت آن دو را بررسی کرد در مقابل سیستم-های هویت‌سنجی کار بسیار پیچیده تری دارند زیرا شانس مشابهت فرد مورد جستجو با افراد دیگر (اشتباها صحیح[۱۰۵]) یا اشتباه در عدم مشابهت با فرد صحیح (اشتباها ناصحیح[۱۰۶]) بسیار زیاد است.

---

[۹۵] Volume
[۹۶] Velocity
[۹۷] Veracity
[۹۸] Variety
[۹۹] Indexing
[۱۰۰] Constant time
[۱۰۱] Verification
[۱۰۲] Authentication
[۱۰۳] Identification
[۱۰۴] Matching
[۱۰۵] False positive
[۱۰۶] False negative

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

امروزه سیستم‌های تصدیق هویت هم در استفاده و هم در دقت، رشد چشمگیری داشته‌اند چنانکه بهره مندی از آنها در پایگاه‌های اطلاعات بایومتریکی کاربردهای دولتی بالغ بر چندین میلیون بلکه چند صد میلیون رکورد نامرسوم نیست. این پایگاه داده‌ها نه تنها دارای تعداد بسیار زیادی از رکوردهای اطلاعاتی می باشند بلکه سیستم های بایومتریکی مربوطه، توانایی پردازش مجموعه متنوعی از ویژگی‌ها نظیر ویدئو، تصویر، صدا، متن یا حتی سیگنال را دارند. بعنوان کلان‌داده‌ها، همچنان چالش‌هایی اعم از دسترسی سریع به اطلاعات، امنیت سیستم، صحت رکوردها و مدیریت تغییرات در این نوع سیستم‌ها مطرح است.

۲-۴ فهرست‌بندی (شاخص‌گذاری)

یک سیستم بزرگ بایومتریکی از حجم بسیار زیادی از انواع داده‌های چندرسانه‌ای شامل تصاویر، ویدئوها، داده‌های متنی ساختار یافته و غیر ساختاری از جمعیت بسیار زیادی از انسان‌ها حتی به اندازه جمعیت یک کشور تشکیل شده است. هنگامیکه چنین سیستمی ایجاد شد، داده‌ها می‌توانند بصورت فرآیندهای تصدیق‌هویت (۱ به ۱) در بسیاری از کاربردهای روزانه نظیر دسترسی به حساب بانکی مورد استفاده قرار گیرد. اگر چه این پایگاه داده همچنین بصورت فرآیندهای هویت‌سنجی (۱ به N) نیز مورد نیاز است برای مواردی همچون جستجوی رکوردهای تکراری، پزشکی قانونی، تشخیص تقلب در رای‌گیری، تعقیبات قانونی و غیره که در همه این موارد به جستجوی همه پایگاه داده اطلاعات نیازمند هستیم. رشد پایگاه داده‌های اطلاعات در تصدیق‌هویت مساله خیلی جدی نیست زیرا پردازش ها بصورت یک فرآیند با زمان مصرفی O(1) انجام می شوند در مقابل پرس‌وجوهای هویت‌سنجی که نیازمند جستجو کل پایگاه داده‌ها هستند بسیار حیاتی هستند که به منظور فراهم کردن یک سیستم کارآ باید هزینه و زمان مصرفی آن از O(n) بهتر باشد. برای جستجوی رکوردهای تکراری نیز مساله فهرست‌بندی بسیار حیاتی است که راه حل ابتدائی برای آن به پیچیدگی زمانی معادل $O(n^2)$ نیاز دارد، در حالیکه تعداد رکوردها در کلان‌داده‌ها چند ده یا چند صد میلیون است ایده‌ی بسیار نامناسبی خواهد بود. با این حال روش‌های فهرست بندی برای اثرانگشت، عنبیه چشم و چهره وجود دارد که می‌توانند زمان مصرفی را تا O(1) بهبود بخشند.

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

اثرانگشت

بطور کلی با استفاده از تکنیک‌های تشخیص الگو به الگوبرداری از هر نمونه اثرانگشت می پردازد بعنوان مثال با مثلث بندی هر سه ریزه[۱۰۷] موجود بر روی اثرانگشت نهایتا الگویی را ایجاد می کند که در فهرست‌بندی (شاخص گذاری) اثرانگشت‌ها می تواند کلان‌داده‌ها را با فاکتور ۱۰۰ کوچک کند ضمن آنکه به اندازه کافی نسبت به سایز، موقعیت و دوران پایدار است.

عنبیه چشم

بطور مشابه در تحلیل بایومتریکی عنبیه چشم که از الگوی غنی‌تری برخوردار است فهرست بندی سبب بهبود پیچیدگی و دقت مکانیزم هویت‌سنجی می‌شود بعنوان مثال Daugman در [۸۳] الگویی را برای تفکیک ۲۰۰ میلیارد عنبیه چشم بر اساس الگوی آن معرفی کرده که با فهرست بندی می‌تواند از کارآیی مطلوبی برخوردار گردد. ایده فهرست‌بندی عنبیه چشم به این ترتیب است که: یک توصیف مختصر از عنبیه برای جستجوی گروهی از نمونه‌ها استفاده می‌شود سپس در مرحله دوم، جستجویی کاملتر به تطابق دقیق عنبیه مورد جستجو با زیرمجموعه داده‌ها می‌پردازد.

چهره

در رهیافت مشابه برای هویت‌سنجی مبتنی بر چهره در کلان داده‌ها، فهرست‌بندی با انتخاب k بهترین چهره مشابه آغاز می شود در حالیکه K<<N می باشد و N تعداد نمونه‌های پایگاه اطلاعات است. و سپس در مرحله دوم k نمونه بطور مجدد با معیارهای دقیقتر رتبه‌بندی می شوند و نهایتا شبیه‌ترین نمونه (یا تعدادی از شبیه‌ترین نمونه‌ها، مثلا ۱۰ شبیه‌ترین) بعنوان نتیجه انتخاب می‌گردند.

[۱۰۷] Minutiae

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

۵ نتیجه‌گیری

با توجه به اهمیت بالای الگوریتم‌های تحلیل کلان داده‌ها و تنوع روش‌های موجود، در این گزارش ابتدا روش‌های تحلیلی موجود در علوم داده مورد بررسی قرار گرفته اند. همان‌طور که در بخش‌های مختلف این گزارش عنوان شده است، روش‌های متنوعی برای تحلیل داده‌ها از قبیل روش‌های آماری، مدل‌سازی‌های ریاضی و روش‌های مبتنی بر یادگیری ماشینی در دهه‌های اخیر برای تحلیل داده‌های موجود، مورد استفاده قرار می‌گیرند. اما با توجه به خصوصیات خاص کلان داده‌ها، روش‌هایی باید مورد استفاده قرار گیرند که قابلیت بالایی در مدیریت حافظه داشته باشند و بتوانند حجم بسیار زیادی از داده‌های بدون ساختار را بررسی کنند. برای این منظور یا باید الگوریتم‌های به شیوه جدیدی مانند Map Reduce و یا ابداعی دیگر اجرا گردد. در این گزارش علاوه بر ترسیم ساختار گونه‌شناسی روش‌های تحلیل کلان داده‌ها، الگوریتم‌های مورد استفاده در این حوزه مورد بررسی قرار گرفته‌اند. ضمنا بسترها و زیرساخت‌های نرم افزاری لازم جهت انجام محاسبات توزیع شده نیز به عنوان یکی از مهم‌ترین اجزا بخش تحلیل کلان داده‌ها مورد مطالعه قرار گرفته اند. بر اساس مطالب ذکر شده در این گزارش، تقریبا اکثر الگوریتم‌های سنتی موجود با اعمال تغییراتی قابل استفاده در حوزه کلان داده‌ها بوده و معمولا در کاربردهای رایج، روش‌های مناسب بر اساس نوع داده و شرایط خاص کاربرد انتخاب می‌گردند.

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

مراجع

[1] Big Data Value Europe, "European Big Data Value Partnership Strategic Research and Innovation Agenda," Big Data Value Europe, BRUSSELS, January 2015.
[2] "The Massachusetts Big Data Report/A Foundation for Global Leadership," MassTech, April 2014.
[3] Alberto Fernández, Sara del Río, Victoria López, Abdullah Bawakid, María J. del Jesus, José M. Benítez and Francisco Herrera, "Big Data with Cloud Computing: an insight on the computing environment, MapReduce, and programming frameworks," John Wiley & Sons, Ltd., Volume 4, September/October 2014.
[4] Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of machine learning*. MIT press.
[5] Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques.
[6] Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.
[7] Jun, S., Lee, S. J., & Ryu, J. B. (2015). A Divided Regression Analysis for Big Data. *Statistics*, *9*(5).
[8] http://spark.apache.org/docs/latest/ml-classification-regression.html
[9] Vapnik, V. (2013). *The nature of statistical learning theory*. Springer Science & Business Media.
[10] Rebentrost, P., Mohseni, M., & Lloyd, S. (2014). Quantum support vector machine for big data classification. *Physical review letters*, *113*(13), 130503.
[11] Auria, L., & Moro, R. A. (2008). Support vector machines (SVM) as a technique for solvency analysis.
[12] Murdopo, A. (2013). Distributed decision tree learning for mining big data streams. *Master of Science Thesis, European Master in Distributed Computing*.
[13] Leong, L. K. (2014). Analyzing big data with decision trees.
[14] Hall, L. O., Chawla, N., & Bowyer, K. W. (1998, October). Decision tree learning on very large data sets. In *Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on* (Vol. 3, pp. 2579-2584). IEEE.
[15] Ayma, V. A., Ferreira, R. S., Happ, P., Oliveira, D., Feitosa, R., Costa, G., & Gamba, P. (2015). Classification Algorithms for Big Data Analysis, a Map Reduce Approach. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *40*(3), 17.
[16] Liu, B., Blasch, E., Chen, Y., Shen, D., & Chen, G. (2013, October). Scalable sentiment classification for big data analysis using Naive Bayes Classifier. In *Big Data, 2013 IEEE International Conference on* (pp. 99-104). IEEE.
[17] http://spark.apache.org/docs/latest/mllib-naive-bayes.html
[18] Sharma, C. (2014). Big Data Analytics Using Neural networks.
[19] Hodge, V. J., O'Keefe, S., & Austin, J. (2016). Hadoop neural network for parallel and distributed feature selection. *Neural Networks*, *78*, 24-35.
[20] Xiao Cai, Feiping Nie, Heng Huang, Multi-View K-Means Clustering on Big Data, IJCAI '13 Proceedings of the Twenty-Third international joint conference on Artificial Intelligence, 2013
[21] Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: simplified data processing on large clusters. Commun. ACM 51, 1 (January 2008), 107-113.
[22] D. Freedman and P. Kisilev, "Fast Mean Shift by compact density representation," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 1818-1825.
[23] Discovery Science: 15th International Conference, DS 2012, Lyon, France, October 29-31, 2012.
[24] W. Zhang, G. Zhang, Y. Wang, Z. Zhu and T. Li, "NNB: An efficient nearest neighbor search method for hierarchical clustering on large datasets," Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015), Anaheim, CA, 2015
[25] Emanuele Coviello, Antoni B. Chan, and Gert R. G. Lanckriet. 2014. Clustering hidden Markov models with variational HEM. J. Mach. Learn. Res. 15, 1, 2014
[26] MapReduce Algorithms for Big Data Analysis, Kyuseok Shim, 8th International Workshop, DNIS 2013, Aizu-Wakamatsu, Japan, March 25-27, 2013

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

[27]A. Ben Ayed, M. Ben Halima and A. M. Alimi, "Survey on clustering methods: Towards fuzzy clustering for big data," 2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR), Tunis, 2014
[28]Md. Mostofa Ali Patwary, Suren Byna, Nadathur Rajagopalan Satish, Narayanan Sundaram, Zarija Lukić, Vadim Roytershteyn, Michael J. Anderson, Yushu Yao, Prabhat, and Pradeep Dubey. 2015. BD-CATS: big data clustering at trillion particle scale. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2015
[29]Ali Seyed Shirkhorshidi, Saeed Aghabozorgi, Teh Ying Wah, Tutut Herawan, Big Data Clustering: A Review, 2014
[30]Jianqiang Dong, Fei Wang, and Bo Yuan. 2013. Accelerating BIRCH for Clustering Large Scale Streaming Data Using CUDA Dynamic Parallelism. In Proceedings of the 14th International Conference on Intelligent Data Engineering and Automated Learning, 2013
[31]Spectral Methods for Unsupervised and Discriminative Learning with Latent Variables, 2015, http://www.ucl.ac.uk/bigdata-theory/prof-animashree-anandkumar/
[32]Franke, B., Plante, J. -F., Roscher, R., Lee, E.-A., Smyth, C., Hatefi, A., Chen, F., Gil, E., Schwing, A., Selvitella, A., Hoffman, M. M., Grosse, R., Hendricks, D., and Reid, N. (2016) Statistical Inference, Learning and Models in Big Data. International Statistical Review, 2016
[33]Matthew Toews , Christian Wachinger, Raul San Jose Estepar, William M. Wells, A Feature-based Approach to Big Data Analysis of Medical Images, 24th International Conference, IPMI 2015
[34]Lizhe Wang, Weijing Song, Peng Liu, Link the remote sensing big data to the image features via wavelet transformation, Cluster Computing, 2016
[35]Steven L. Scott and Alexander W. Blocker and Fernando V. Bonassi, Bayes and big data: the consensus Monte Carlo algorithm, International Journal of Management Science and Engineering Management, 2016
[36]Y. Zhu, Y. Zhang and Y. Pan, "Large-scale restricted boltzmann machines on single GPU," 2013 IEEE International Conference on Big Data, Silicon Valley, CA, 2013
[37]Chun-Yang Zhang, C.L. Philip Chen, Dewang Chen, Kin Tek NG, MapReduce based distributed learning algorithm for Restricted Boltzmann Machine, Neurocomputing, Volume 198, 2016
[38]G. N. Iyer, S. Silas and G. Iyer, "An optimized cloud based big data processing mechanism using Self-Organizing Map in Hadoop environments," 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), Noida, 2015
[39]Lei Meng, Ah-Hwee Tan, Donald C Wunsch, Adaptive Scaling of Cluster Boundaries for Large-Scale Social Media Data Clustering, IEEE Trans Neural Netw Learn Syst. 2015
[40]W. LiuH. ZhangD. TaoEmail authorY. WangK. Lu, Large-scale paralleled sparse principal component analysis, Multimedia Tools and Applications, 2014
[41]Tarek Elgamal, Maysam Yabandeh, Ashraf Aboulnaga, Waleed Mustafa, and Mohamed Hefeeda. sPCA: Scalable Principal Component Analysis for Big Data on Distributed Platforms. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, 2015
[42]Namgil LeeAffiliated, Andrzej Cichocki, Big data matrix singular value decomposition based on low-rank tensor train decomposition, 11th International Symposium on Neural Networks, 2014
[43]d. chen; Y. Hu; L. Wang; A. Zomaya; X. Li, "H-PARAFAC: Hierarchical Parallel Factor Analysis of Multidimensional Big Data," in IEEE Transactions on Parallel and Distributed Systems , 2016
[44]Piercesare Secchi, Simone Vantini, Paolo Zanini, The Virtuous Cycle of Big Data and Big Cities: a Case Study from Milan, 47th Scientific Meeting of the Italian Statistical Society, 2014
[45]Ruiqi Liao, Yifan Zhang, Jihong Guan, Shuigeng Zhou, CloudNMF: A MapReduce Implementation of Nonnegative Matrix Factorization for Large-scale Biological Datasets, Genomics, Proteomics & Bioinformatics, Volume 12, 2014
[46]Zhuolin Qiu, Bin Wu, Bai Wang, Chuan Shi, and Le Yu. 2014. Collapsed Gibbs sampling for latent Dirichlet allocation on spark. In Proceedings of the 3rd International Conference on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications - Vol. 36.

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

[47]P-ISOMAP: An efficient parametric update for ISOMAP for visual analytics, 2010, www.cc.gatech.edu/~hpark/papers/pisomap_sdm10.pdf
[48]L. Cheng and C. You, "Hybrid non-linear dimensionality reduction method framework based on random projections" IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, 2016
[49]Antoine Naud, Włodzisław Duch, Visualization of large data sets using MDS combined with LVQ, Proceedings of the Sixth International Conference on Neural Networks and Soft Computing, Zakopane, Poland, June 11–15, 2003
[50]Jianwei Zheng, Hangke Zhang, Carlo Cattani, Wanliang Wang, Resolution of the big-data problem related to a dimension reduction algorithm based on Multi-Scale similarities in Stochastic Neighbor Embedding, Comput Math Methods Med. 2014
[51]A. *Engelbrecht*, "*Computational* Intelligence, an Introduction", Second Edition, 2007.
[52]Yuxi Li, and Dale Schuurmans, "MapReduce for Parallel Reinforcement Learning", Lecture Notes in Computer Science, vol. 7188, pp 309-320, 2011.
[53]Yi Chun Chen, and Yu Sheng Chen, "A Distributed Implementation for Reinforcement Learning", Institute for Computational and Mathematical Engineering, Stanford University, 2016.
[54]Kevin Chavez, Hao Yi Ong, and Augustus Hong, "Distributed Deep Q-Learning", 2016.
[55]Q. W. G. D. Y. X. a. S. F. Junfei Qiu, "A survey of machine learning for big data processing," EURASIP Journal on Advances in Signal Processing, vol. 67, pp. 1-16, 2016).
[56]X. Z. a. A. B. Goldberg, Introduction to Semi-Supervised Learning, Madison: Morgan &cLaypool publishers, 2009.
[57]Cloud Security Alliance, "Big Data Taxonomy," Cloud Security Alliance, 2014.
[58]X. Wan, "Co-Training for Cross-Lingual Sentiment Classification," in Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the AFNLP, Suntec, Singapore, 2009.
[59]S. J. P. a. Q. Yang, "A Survey on Transfer Learning," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, vol. 22, no. 10, pp. 1345-1359, 2010.
[60]F. L. a. H.-P. H. Yu Zheng, "U-Air: when urban air quality inference meets big data," in KDD '13 Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, Chicago, 2013.
[61]S. K. a. S. Matwin, "Email Classification with Co-Training," in CASCON '01 Proceedings of the 2001 conference of the Centre for Advanced Studies on Collaborative research, Toronto, 2001.
[62]P. K. K. A. A. R. R. C. R. B. Raghavendra Kune, "The anatomy of big data computing," Software—Practice & Experience, vol. 46, no. 1, pp. 79-105, 2016.
[63]C.-Y. Lin, "Big Data Analytics," Columbia University, fall 2016, Available: https://www.ee.columbia.edu/~cylin/course/bigdata/.
[64]"California Data Science," 2016. Available: www.CaliforniaDataScience.com
[65]X. Z. G.-Q. W. a. W. D. Xindong Wu, "Data Mining with Big Data," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, vol. 26, no. 1, pp. 97-106, 2014.
[66]"SciKit Learn," Available: scikit-learn.org.
[67]"Shogun-," Available: http://shogun-toolbox.org/.
[68]"Accord Framework," Available: http://accord-framework.net/.
[69]"Aforge.net," Available: http://aforge.net/.
[70]"Mahout," Available: http://mahout.apache.org/.
[71]"MLlib," Available: http://spark.apache.org/mllib/.
[72]"H2O," Available: http://www.h2o.ai/.
[73]"Oryx," Available: http://oryx.io/.
[74]"Weka," Available: http://www.cs.waikato.ac.nz/ml/weka/.
[75]"CUDA-Convnet," Available: code.google.com/p/cuda-convnet/.
[76]"ConvNetJS," Available: http://cs.stanford.edu/people/karpathy/convnetjs/.

نام گزارش: تحلیل کلان داده‌ها	وضعیت: نهایی	کد

[77]"Google Cloud Platform," Available: https://cloud.google.com/.
[78]"Amazon Machine Learning," Available: https://aws.amazon.com/machine-learning/.
[79]"BigML," Available: https://bigml.com/.
[80]"IBM Watson," Available: https://www.ibm.com/watson/.
[81]"Microsoft Azure," Available: https://azure.microsoft.com/en-us/.
[82] N. K. Ratha, J. H. Connell and S. Pankanti, "Big Data approach to biometric-based identity analytics," in IBM Journal of Research and Development, vol. 59, no. 2/3, pp. 4:1-4:11, March-May 2015.
[83]J. Daugman, "Probing the Uniqueness and Randomness of IrisCodes: Results from 200 Billion Iris Pair Comparisons," in Proceedings of the IEEE, vol. 94, no. 11, pp. 1927-1935, Nov. 2006.