یادگیری ماشین کارآمد برای کلان داده ها: یک مقاله مروری

چکیده :

با فن آوری های در حال ظهور و تمام دستگاه های مرتبط، پیش بینی می شود که مقدار عظیمی از اطلاعات در چند سال آینده ایجاد خواهد شد - در واقع، 90 درصد از داده های کنونی در چند سال گذشته ایجاد شده است که ادامه این روند برای آینده قابل پیش بینی است . فرایند مطالعات و محاسبات پایدار که در مهندسی کامپیوتر و زیر سیستم های مرتبط کارآمد هستند و حداقل تاثیر را بر روی محیط زیست دارند. با این حال، سیستم های یادگیری ماشینی هوشمند فعلی دارای عملکرد محور می باشند - تمرکز بر دقت پیش بینی / و طبقه بندی، بر اساس خواص شناخته شده از نمونه آموزشی به دست می آید. به عنوان مثال، مدل ناپارامتریک مبتنی بر یادگیری ماشینی نیاز به هزینه های محاسباتی بالا در جهت پیدا کردن OPTIMA جهانی دارد. با این کار یادگیری در یک مجموعه داده های بزرگ، تعداد گره های پنهان در داخل شبکه به طور قابل توجهی افزایش می یابد، که در نهایت به افزایش نمایی در پیچیدگی محاسباتی منجر می شود. بنابراین در این مقاله داده مدل سازی نظری و تجربی ، در زمینه های اطلاعات فشرده در مقیاس بزرگ بررسی شده است ، که مربوط به: (1) بهره وری مدل، از جمله نیازهای محاسباتی در یادگیری، و ساختار اطلاعات فشرده مناطق و طراحی (2) روش های الگوریتمی جدید با حداقل حافظه مورد نیاز و پردازش برای به حداقل رساندن هزینه های محاسباتی، در حالی که حفظ / بهبود پیش بینی / دقت طبقه بندی و ثبات آن مد نظر است.

کلمات کلیدی : کلان داده، رایانش سبز، یادگیری ماشینی کارامد، مدل سازی رایانشی

1. مقدمه

امروز، جای تعجب نیست که کاهش هزینه های انرژی یکی از اولویت های اصلی برای بسیاری از تجارت های مربوط به انرژی است. صنعت جهانی اطلاعات و فن آوری ارتباطات (ICT) که حدود 830 تن دی اکسید کربن (CO_2) انتشار داده است که حدود 2 درصد از انتشار گاز CO_2 جهانی می باشد. غول فناوری اطلاعات و ارتباطات به طور مداوم با

نصب سرورهای بیشتر برای گسترش ظرفیت خود اقدام می کند. تعداد کامپیوترهای سرور در مراکز داده 30 میلیون است که شش برابر در دهه گذشته افزایش یافته ، و هر سرور به مراتب بیشتر از مدل های قبلی آن است. استفاده از برق برای سرورها بین سال های 2000 و 2005 ، که تعداد زیادی از سرورهای جدید نصب شد دو برابر شده بود. این افزایش در مصرف انرژی به تبع آن باعث تولیدگازهای گلخانه ای دی اکسید کربن بیشتر می شود، و از این رو باعث تاثیر بر محیط زیست می شود. علاوه بر این، بسیاری از این تجارت ها، به خصوص در شرایط نامشخص اقتصادی به منظور کاهش مصرف انرژی خود را جهت بازار رقابتی تحت فشار قرار داده اند.

با این حال با ظهور فن آوری های جدید و ارتباطات ، پیش بینی می شود که داده های زیادی به عنوان کل در تاریخ سیاره زمین ساخته شده است. با توجه به میزان بی سابقه ای از اطلاعات که تولید خواهد شد، در سال های آینده مرتب سازی و ذخیره ، یکی از چالش های بزرگ تکنولوژی در خدماتش به صنعت برای چگونگی بهرمندی از آن است. در طول دهه گذشته، سیستم های یادگیری ماشینی هوشمند ریاضی به طور گسترده ای در تعدادی از زمینه های داده گسترده و پیچیده از قبیل نجوم، زیست شناسی، اقلیم، پزشکی، امور مالی و اقتصاد به تصویب رسیده اند. با این حال، سیستم های فعلی هوشمند مبتنی بر یادگیری ماشینی ذاتا کارآمد و یا به اندازه کافی برای مقابله با حجم زیادی از داده ها مقیاس پذیر نیستند. به عنوان مثال، برای سال های زیادی، بسیاری از روش های غیر پارامتری و مستقل از مدل نیاز برای هزینه های محاسباتی بالا و برای پیدا کردن روش بهینه جهانی شناخته شده اند. با داده های ابعادی بالا، اطلاعات خوب از ظرفیت اتصالات باعث می شود آن ها بیشتر مستعد مشکل تعمیم شوند اما منجر به افزایش پیچیدگی محاسباتی می شود.

طراحی سیستم های دقیق تر یادگیری ماشینی برای برآوردن نیاز های بازار با توجه به افزایش هزینه های محاسباتی به احتمال بیشتری از اتلاف انرژی منجر خواهد شد.

امروزه، نیاز بیشتری به توسعه مدل هوش کارآمد برای مقابله با خواسته های آینده و طرح های مرتبط با انرژی مشابه وجود دارد. چنین مدل سازی داده گرای کارآمد انرژی برای تعدادی از مناطق اطلاعاتی فشرده مهم است، آن ها بسیاری از صنایع مرتبط را تحت تاثیر قرار می دهند. طراحان باید برای حداکثر کارایی و حداقل مصرف انرژی تمرکز

کنند به طوری که برای معاوضه و در مقابل استفاده سنتی از انرژی ، و افزایش تعداد و تنوع گزینه های موجود برای مدل سازی با انرژی کارآمد استفاده شوند. با این حال، با وجود این واقعیت است که تقاضا برای چنین روش مدل سازی داده ها کارآمد و پایدار برای زمینه های بزرگ و پیچیده اطلاعات فشرده وجود دارد ، تنها تعداد کمی از آن ها در زمینه [6،7] ارائه شده است.

در این مقاله یک بررسی جامع از یادگیری ماشینی با حالت صنعت ، پایدار / با انرژی کارآمد انجام شده است، و مطالعات نظری، تجربی و مربوط به توصیه های مختلف را فراهم می کند. هدف ما معرفی یک چشم انداز جدید برای مهندسان، دانشمندان و پژوهشگران در علوم کامپیوتر، و حوزه ICT سبز، و همچنین به عنوان ارائه نقشه راه برای تلاش تحقیقات در آینده است.

این مقاله به شرح زیر است. بخش 2 مناطق اطلاعاتی فشرده مختلف در مقیاس بزرگ را معرفی و ساختار و ماهیتشان ، از جمله رابطه بین مدل های داده و ویژگی های آن ها را مورد بحث قرار می دهد. بخش 3 مسائل در مدل سازی داده های هوشمند فعلی مورد بحث برای پایداری توصیه می شود. بخش 4 نتیجه گیری مقاله.

2. چالش داده های بزرگ

علم الکترونیک به طور معمول اطلاعات فشرده و کیفیت نتایج خود را با کمیت و کیفیت داده های در دسترس بهبود می بخشد. با این حال، سیستم های یادگیری ماشینی هوشمند فعلی ذاتا در بسیاری از موارد به اندازه کافی کارآمد نیستند ، بخش رو به رشد از این داده ها مقدار ناشناخته و تحت بهره دهی هستند. روش های موجود موفق به گرفتن چنین کلان داده بدون کوچکترین مشکل شده اند. هنگامی که مفاهیم قدیمی ، سنت ها و تجربه های گذشته راهنمای ناکافی برای درک عملکرد در آینده باشد. درک موثر و استفاده از این اطلاعات خام یک چالش بزرگ برای مهندسین سبز / محققان امروز در برخواهد داشت. لازم به ذکر است که دامنه این بررسی محدود به جنبه های تحلیلی از مناطق علمی با استفاده از مجموعه داده های بسیار زیاد، و روش برای کاهش پیچیدگی محاسباتی در محیط شبکه-محاسبات توزیع شده و یا حذف شده است.

1.2. جغرافیایی، آب و هوا و محیط زیست

بسیاری از نمونه های اخیر می توانند رشد فوق العاده ای در تولید داده های علمی در یادداشت ها به وجود آورند. تخمین زده می شود که هزاران عدد سنسور بی سیم در حال حاضر وجود دارد ، که هر سنسور یک گیگابایت اطلاعات را در هر روز تولید می کند. اندازه گیری سنسور ها و اطلاعات حسی رکوردی در مورد محیط زیست طبیعی در ابعاد مکانی و زمانی مشترک پدید آورده که هرگز قبلا مشاهده نشده است. این اطلاعات زیست محیطی توسط سنسور از طریق دستگاه های سنجش به، سیستم های کامپیوتری کم قدرت کوچک با ارتباطات رادیویی دیجیتال وصل شده و جمع آوری شده اند. سنسور داده های جمع آوری شده را با یک ایستگاه پایه پردازش می کند، که آن را می توان از طریق اینترنت در دسترس کاربران قرار داد. این سنسورها چند پتابایت داده در هر سال تولید می کنند و تصمیم گیری در زمان واقعی در نظر گرفته می شود ، چه مقدار اطلاعات برای تجزیه و تحلیل چقدر از انتقال تجزیه و تحلیل بیشتر است.

علاوه بر محیط زیست، یک چالش مشابه رو به اقلیم، هواشناسان، و زمین شناسان امروز همچنین در حال ساختن حس قطعی و به طور مستمر افزایش مقدار اطلاعات تولید شده توسط ماهواره زمین، رادار، و شبکه های حسگر بالا استفاده شده است. مرکز اطلاعات جهانی (WDCC) بزرگترین منبع داده های آب و هوا است، و همچنین به عنوان بزرگترین پایگاه داده در جهان شناخته می شود. در آرشیو WDCC 340 ترابایت داده سیستم زمین و مدل و مشاهدات مرتبط وجود دارد، و 220 ترابایت داده در دسترس بر روی وب از جمله اطلاعات در تحقیقات آب و هوا و روند آب و هوایی پیش بینی شده وجود دارد، و همچنین 110 ترابایت (یا 24500 دی وی دی) داده های شبیه سازی آب و هوا م شاهده می شود. داده های WDCC توسط یک رابط استاندارد وب (http://cera.wdc-climate.de) قابل دسترسی است. این داده ها به طور فزاینده ای در بسیاری از فرمت های مختلف در دسترس است و باید به درستی در مدل های مختلف تغییر آب و هوا گنجانیده شود. تفسیر به موقع و دقیق از این داده ها می تواند هشدارهایی زودتر

در زمان تغییرات شدید آب و هوا ارائه دهد، از این رو برای به حداقل رساندن آسیب فاجعه بار ناشی از آن عمل مربوطه به سرعت انجام می شود.

2.2. بیوگرافی، پزشکی و سلامت

داده های بیولوژیکی با سرعت فوق العاده و با توجه به تلاش های تحقیقاتی بین المللی به نام پروژه ژنوم انسان تولید شده است. تخمین زده می شود که DNA ژنوم انسان شامل حدود 3.2 میلیارد پایه (پایه 3.2 گیگا) جفت در میان بیست و سه کروموزوم است که در مورد یک گیگابایت اطلاعات ترجمه شده توزیع شده است. با این حال، زمان اضافه کردن اطلاعات توالی ژن (داده ها در 100.000 یا پروتئین ترجمه شده و 32000000 اسیدهای آمینه)، حجم داده های مربوطه می تواند به راحتی به منظور حدود 200 گیگابایت گسترش یابد. در حال حاضر، نیز تعیین ساختار طیف سنجی این پروتئین ها، حجم داده ها به طور چشمگیری به چند پتابایت افزایش خواهد یافت، و فرض تنها یک ساختار در پروتئین است.

در دسامبر 2014، مخزن ژنی توالی اسید نوکلئیک شامل 178 میلیون ثبت و پایگاه داده SWISS-PROT (INC. هر دو UniProtKB/SwissProt ، UniProSM و TKB / TrEMBL) از توالی های پروتئینی موجود حدود 18 میلیون ثبت شده است. به طور متوسط، اندازه این پایگاه داده در هر 15 ماه دو برابر می شود. به همین علت بیشتر توسط داده های تولید شده از هزاران پروژه ی مرتبط بیان ژن مطالعه صورت می گیرد، که ساختارهای پروتئین توسط ژن تعیین می شود. از این رو، ما می توانیم تصور کنیم مقدار بسیار زیادی از انواع اطلاعات در هر ماه تولید شده است.

3.2. ستاره ها، کهکشان ها و جهان

حجم داده های دیجیتال از ستاره ها، کهکشان ها و جهان در دهه گذشته با توجه به رشد سریع فن آوری های جدید مانند ماهواره های جدید، تلسکوپ و دیگر ابزار رصدخانه چند برابر شده است. به تازگی، تلسکوپ قابل مشاهده و بررسی مادون قرمز برای نجوم (VISTA) و بررسی انرژی تاریک (DES) بزرگترین پروژه بررسی جهان توسط دو

کنسرسیوم مختلف دانشگاه هایی، از انگلستان، و ایالات متحده آغاز شده ،و انتظار می رود که اندازه پایگاه داده 20-30 ترابایت در دهه آینده شود.

با توجه به DES، رصدخانه آن آنقدر بزرگ است که یک تصویر واحد داده ها از یک منطقه از آسمان 20 برابر اندازه ماه از زمین دیده می شود. در این بررسی تصویر 5000 درجه از آسمان جنوب ایالات متحده و حدود پنج سال طول بکشد. همانطور که برای VISTA، عملکرد مورد نیاز خود را به طوری چالش برانگیز است که آن را در 55 مگابایت / نرخ داده دوم با حداکثر 1.4 ترابایت از اطلاعات بودند. اما، در حال حاضر نسبتا عادی است. بسیاری دیگر از پایگاه داده های علمی ASTRO ، مانند نقشه برداری آسمانی دیجیتال اسلون (SDSS) در حال حاضر در اندازه ترابایت و پاسخ سیستم تلسکوپ-سریع پانوراما بررسی می شوند (Pan-STARRS) انتظار می رود که یک پایگاه داده علم بیش از 100 ترابایت برای پنج سال آینده باشد. به همین ترتیب، تلسکوپ سینوپتیکی بزرگ (LSST) 30 ترابایت داده در هر شب تولید می کند، بازده یک پایگاه داده در مجموع حدود 150 پتابایت بود. انتظار می رود داده های تولید شده توسط تلسکوپ جدید برای قرار گرفتن در اینترنت، کاملا تغییر خواهد کرد.

بسیاری بر این باورند که حجم داده های عظیم و افزایش روزافزون قدرت محاسبات به طور چشمگیری در راه علم و تکنولوژی تغییر خواهد کرد. ما بر این باوریم که این افزایش در داده باعث به چالش کشیدن تحقیقات بیشتر در هر زمینه می شود، از این رو، تحریک جستجو برای روش های جدید صورت می گیرد. به همین ترتیب، چنین چالشی نیاز دارد تا در حوزه علم اطلاعات هوشمند پرداخته شود.

3. مدل سازی داده ها پایدار و یادگیری کارآمد

با در نظر گرفتن هجوم زیادی از داده ها، آن ها قطعا برای بهبود راه در چگونگی مدل های داده های محاسباتی / تحلیلی طراحی شده و توسعه یافته اند. مدل سازی داده ها پایدار می تواند به عنوان یک شکل از تکنولوژی مدل سازی داده ها، با هدف ایجاد داده های مرتبط در این زمینه باشند، با کشف الگوها و همبستگی در یک راه موثر و کارآمد تعریف شده است. مدل سازی داده ها پایدار به طور خاص در 1) دقت یادگیری حداکثر با حداقل هزینه محاسباتی، و

2) پردازش سریع و کارآمد از حجم زیادی از داده متمرکز است. مدل سازی داده ها پایدار به نظر می رسد به دلیل سهولت آن در مقادیر زیادی از داده ها کارآمد ایده آل باشد که کاهش هزینه های مرتبط با آن در بسیاری از موارد به کار گرفته شده مشاهده شده است. در یک چشم انداز گسترده تر، مستلزم یک انقلاب مدل سازی داده ها در علوم الکترونیک است. در واقع، این مدل های داده پایدار به تازگی برای مقابله با مسائل اطلاعات فوق و به عنوان یک نتیجه طراحی شده اند، در مورد منافع مختلف علم الکترونیک . برخی از نمونه های عالی به خوبی در پاتنایک و همکاران، ساندارااوارادان و همکاران، و مقاله مروری مورد بحث قرار گرفته است. از این رو، در این بخش، ما چند توصیه برای مهندسین سبز / محققان در مدل سازی داده ها پایدار ارائه داده ایم.

1.3. مدل گروهی

یکی از عناصر کلیدی موفقیت مدل سازی داده ها پایدار است که برای حفظ و یا بهبود عملکرد آن و به طور قابل توجهی برای کاهش هزینه محاسباتی آن است. آخرین تحقیقات مدل سازی داده ها نشان داده است که روش گروهی محبوبیت زیادی به دست آورده است ، آن ها اغلب بهتر از مدل های انفرادی عمل می کنند. روش گروهی با استفاده از مدل های مختلف برای به دست آوردن عملکرد بهتر می تواند از هر یک از مدل های تشکیل دهنده به دست آید. با این حال، آن می تواند به افزایش قابل توجهی در هزینه های محاسباتی منجر شود. اگر معاملات مدل با داده های در مقیاس بزرگ باشند، پیچیدگی مدل و نیازهای محاسباتی نمایی رشد خواهد کرد. یک مثال از چنین مدل گروهی طبقه بندی بیز است. در طبقه بندی بیز، هر فرضیه بین تناسب به احتمال داده می شود که مجموعه داده یک سیستم نمونه این فرضیه است. به منظور تسهیل در داده ها از اندازه محدود، رای هر فرضیه نیز در احتمال قبلی این فرضیه ضرب خواهد شد. طبقه بندی بیز به شرح زیر بیان شده است :

$$y = \arg\max_{c_j \in C} \sum_{h_i \in H} P(c_j|h_i) P(T|h_i) P(h_i),$$

که در آن y طبقه پیش بینی شده است، C مجموعه ای از تمام طبقات ممکن است،H فضای فرضیه است ، P اشاره به احتمال، و T داده های آموزشی است. به عنوان یک گروه طبقه بندی بیز نشان دهنده یک فرضیه است که نه لزوما

در .H فرضیه ارائه شده توسط طبقه بندی بیز می باشد، با این حال، فرضیه مطلوب در فضای گروه (فضای همه گرووه های احتمالی متشکل از تنها فرضیات در h است) است.

با توجه به مشکل پیش بینی آب و هوا ، پیش بینی های گروه در حال حاضر معمولا در بسیاری از امکانات عمده پیش بینی آب و هوا عملیاتی ساخته شده در سراسر جهان، از جمله مراکز ملی پیش بینی محیط زیست، ایالات متحده، مرکز یورو اروپایی برای پیش بینی میان برد هوا (ECMWF)، انگلستان دفتر مت، مترو فرانسه، محیط زیست کانادا، آژانس هواشناسی ژاپن، اداره هواشناسی، استرالیا، هواشناسی اداره چین، دولت هواشناسی کره و CPTEC برزیل است.

2.3. مشکل پیچیدگی مدل

روش های تخمین بیز به طور کلی در مدل سازی داده ها هوشمند به خوبی اتخاذ شده است چرا که آن ها یک فرمالیسم اساسی برای ترکیب تمام اطلاعات موجود، با توجه به پارامترها برآورد، با پیچیدگی زمانی بهینه شده ارائه داده اند.

یکی از مشکلات جدی در یادگیری مدل های ناپارامتری بیز پیچیدگی و حافظه گسترده مورد نیاز بالای الگوریتمی آن به ویژه برای برنامه نویسی درجه دوم در انجام وظایف در مقیاس بزرگ است. یک طبقه بندی بیز ناپارامتری بدترین حالت می باشد ، به عنوان مثال X با استفاده از تجزیه و تحلیل آماری برای ساخت یک مدل طبقه بندی است، هر الگوریتم یادگیری به بررسی مقادیر ویژگی هر مثال آموزشی باید حداقل پیچیدگی را داشته باشد.

بسیاری از برنامه های کاربردی یادگیری ماشین با مشکلات که در آن تعدادی از ویژگی های استاندارد بین المللی همچنین تعدادی از نمونه X_I بزرگ است وجود دارند. پشتیبانی خطی ماشین بردار در میان برجسته ترین تکنیک یادگیری ماشین برای چنین داده با بعد بالا و پراکنده هستند. در این مقاله، ما با استفاده از دو مدل یادگیری ماشینی نمونه نیمه پارامتری داریم. به عبارت دیگر، این دو مدل کارآمد تر و سریع اصلاح محاسباتی می شوند. پیچیدگی زمانی از Bayes و SVM ها به خوبی به ترتیب در مقاله الکان و جواخیمس مورد بحث است.

3.3. استراتژی یادگیری محلی

یو و همکاران. دو مدل گروهی کارآمد مبتنی بر پشتیبانی بردار مختلف به منظور کاهش هزینه محاسباتی با حفظ عملکرد آن ارائه داده اند. روش یادگیری ثابت کرده است که توسط مطالعات مشابه دیگر موفق است. با یک مدل ناپارامتری، یک انفرادی باید برای هر مجموعه تست شود، که به طور قابل توجهی پیچیدگی محاسباتی و هزینه های آن افزایش خواهد یافت.

به منظور کاهش هزینه های محاسباتی، آن ها پیشنهاد پارتیشن نمونه آموزشی به خوشه را پیشنهاد داده اند، ساخت یک مدل محلی جداگانه برای هر خوشه - روشی به نام یادگیری محلی است. تعدادی از آثار اخیر نشان داده که چنین استراتژی یادگیری محلی از استراتژی یادگیری جهانی برتر هستند، به خصوص در مجموعه داده هایی که به طور مساوی توزیع نشده است. اگر یک روش محلی آموزش در تابع تصمیم یک طبقه بندی ناپارامتری (به عنوان مثال، شبکه رگرسیون عمومی) به تصویب رسد، برای طبقه بندی نیمه پارامتری فرض می شود. تقریب نیمه پارامتری آن را می توان به شرح زیر بیان کرد :

$$Z_i \exp \frac{-(x-c_i)^T (x-c_i)}{2\sigma^2} \approx \sum_{j=1}^{Z_i} \exp \frac{-(x-x_j)^T (x-x_j)}{2\sigma^2},$$

که در آن X_I یک بردار آموزش برای کلاس I در فضای ورودی است، σ یادگیری یک یا صاف کردن پارامتر انتخاب شده در طول آموزش شبکه، و Z_I تعدادی از ورودی بردار آموزش X_I مرتبط با مرکز C_I آن است. در طبقه بندی ناپارامتری، بسیاری از انواع مختلف توابع پایه شعاعی در محل تابع گاوسی انتخاب شده است. تابع پایه شعاعی، در بسیاری از موارد مورد استفاده است، در واقع یک تابع هسته کروی است، که به طور خاص برای برآورد تابع ناپارامتری استفاده می شود. اگر تعداد نمونه های آموزشی نزدیک بی نهایت باشد، تابع ناپارامتری روش وابسته به پارامترهای تابع پایه شعاعی بر آورد شده است، با این حال، برای نمونه های آموزشی محدود، ما همیشه می توانیم برخی از انواع وابستگی در پارامترهای تابع پایه شعاعی را مشاهده کنیم.

استراتژی یادگیری محلی وابستگی بیشتر در پارامترهای تابع پایه شعاعی نسبت به یک مدل ناپارامتریک فراهم می کند به دلیل مدل یادگیری محلی یک تقریب نیمه پارامتری یک مدل یادگیری ناپارامتری / جهانی است. به عبارت دیگر، در مدل سازی نیمه پارامتری، مفروضات مدل قوی تر از مدل های ناپارامتری هستند، اما مدل های پارامتری را با محدودیت کمتری مواجه می کنند. به طور خاص، این تقریب از معایب عملی روش های ناپارامتری در هزینه افزایش خطر ابتلا به خطاهای مشخصات جلوگیری می کند. مدل نیمه پارامتری برای کمک به یادگیری محلی نه تنها در کاهش پیچیدگی مدل بلکه در پیدا کردن تجارت بهینه بین مدل های پارامتری و ناپارامتری بر اساس رسیدن به تعصب مدل پایین و واریانس است. به طور خلاصه، در نتیجه می تواند مزیتی ذاتی برای کاهش نیازهای محاسباتی باشد.

4.3. تقریب نیمه پارامتری

مثال های بالا می تواند به عنوان یک مدل مخلوط تابع کروی با تخصیص بردار مرکز داده دیده شود. دلیلش این است که عرض نسبی توابع کروی در هر مرکز به طور مستقیم به تعداد نسبی بردار آموزشی مرتبط با هر مرکز متناسب است. بسیاری از انواع مختلف مدل های محاسباتی محلی، و روش انتخاب های متنوع از Y_l و گروه بندی بردارهای ورودی مرتبط در هر کلاس l می تواند برای مدل جهانی تقریب نیمه پارامتری استفاده شود.

استراتژی یادگیری محلی تقریب منطقی X_l به اندازه کافی در فضای بردار ورودی نزدیک فراهم می کند. در این صورت، می توان آن ها را به اندازه کافی توسط C_l بردار مرکز در آن فضا محلی نشان داد. در مورد پشتیبانی رگرسیون برداری (SVR)، بردارها C_l را می توان از هر دو k یا نظریه کتاب مشتق کرد. در SVR، که در آن دو خوشه قابل جدا سازی نیستند، آن ها فضای ورودی بر روی نقشه را به یک فضای با ویژگی بالا بعدی تبدیل کرده اند (که در آن کلاس ها به صورت خطی از هم جدا) شده اند،که با استفاده از یک تابع هسته غیر خطی انجام شده است. تابع کرنل محاسبه حاصلضرب عددی تصاویر از دو مثال در فضای ویژگی را بیان کرده است.

با توجه به بردار ورودی n بعدی، $x_i = (x_1, x_2, \ldots, x_n)$ با دو برچسب، $y_i \in \{+1, -1\}$ که در آن $i = 1, 2, \ldots, N$، تصمیم ابرصفحه عملکرد SVR باینری با استفاده از روش هسته است :

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{\ell} y_i a_i \langle \Phi(x), \Phi(x_i) \rangle + b\right)$$

$$= \operatorname{sgn}\left(\sum_{i=1}^{\ell} y_i a_i k(x, x_i) + b\right)$$

و برنامه های درجه دوم به عنوان داده به صورت زیر هستند :

$$\text{maximize } W(a) = \sum_{i=1}^{\ell} a_i - \frac{1}{2} \sum_{i,j=1}^{\ell} a_i a_j y_i y_j k(x_i, x_j),$$

$$\text{subject to } \quad a_i \geq 0, \; i = 1, \ldots, \ell, \quad \text{and} \quad \sum_{i=1}^{\ell} a_i y_i = 0,$$

که در آن ℓ تعداد الگوهای آموزش، Aا پارامترهای SVR است ، $K(.,.)$ یک (ناپارامتری) تابع کروی هسته است، و B مدت تعصب است. در مورد بالا، مدل محلی از K خوشه ساخته شده است. یعنی تابع هدف از خوشه بندی K می توان به صورت زیر بیان شود :

$$\min_{C,Z} \sum_{j=1}^{k} \sum_{i=1}^{n} Z_{i,j} \|X_i - C_j\|_2^2 + R \sum_{j=1}^{k} \left| \sum_{i=1}^{n} Z_{i,j} y_i \right|,$$

که در آن Xا سطر i ام از ماتریس شباهت است ، Cا بردار ردیف $1 \times m$ به نمایندگی از مرکز خوشه j ام است، R یک پارامتر جدا سازی شده غیر منفی است، و $Z_{ij} \in \{0, 1\}$ یک عنصر از خوشه ماتریس عضویت است ، که ارزش آن برابر با یک است اگر بردار منبع i ام متعلق به خوشه j ام و در غیر این صورت اگر صفر باشد. عبارت اول در تابع هدف مربوط به اندازه گیری انسجام خوشه می باشد. با به حداقل رساندن معادله بالا اطمینان حاصل شود که بردار آموزش در یک خوشه همبستگی بردار شباهت هستند. دوره دوم چولگی توزیع کلاس در هر خوشه را می سنجد. با به حداقل

رساندن این مدت اطمینان حاصل شود که هر خوشه شامل تعداد متعادل بردار برآورد مثبت و منفی است. مرکز ثقل خوشه C و Z ماتریس عضویت خوشه تکرار برآورد به شرح زیر است :

- ما یک سری مراکز را در خوشه ایجاد می کنیم و از آن ها برای تعیین ماتریس عضویت خوش استفاده می کنیم.
- ناتریس عضویت خوشه اصلاح شده برای به روز رسانی مراکز استفاده می شود و این عمل تا زمانی ادامه می یابد که الکورینم به یک حداقل محلی، همگرا شود.

برای محاسبه ماتریس Z عضویت خوشه، ما مشکل تبدیل بهینه سازی اصلی، با استفاده از T_1 متغیر را داریم ، به صورت زیر:

$$\min_{Z,t} \sum_{j=1}^{k} \sum_{i=1}^{n} Z_{i,j} \| X_i - C_j \|_2^2 + R \sum_{j=1}^{k} t_j,$$

$$\text{s.t.} \quad -t_j \le \sum_{i=1}^{n} Z_{i,j} y_i \le t_j,$$

$$t_j \ge 0,\ 0 \le Z_{i,j} \le 1,$$

$$\sum_{j=1}^{k} Z_{i,j} = 1,$$

اگر ماتریس عضویت خوشه به دست بیاید، خوشه مرکز و C_j بر اساس موارد زیر به محاسبه می شود :

$$Q_j(X_m) = C_j = \frac{\sum_{i=1}^{n} Z_{i,j} X_i}{\sum_{i=1}^{n} Z_{i,j}}, \quad j = 1, 2, \ldots, N.$$

برای ساخت مدل نیمه پارامتری، ما $Q_I(X)$ برای هر X_I نمونه آموزشی در تابع تصمیم SVR جایگزین مورد استفاده . مدل تقریب نیمه پارامتری جدید است بنابراین به صورت زیر بیان شده است :

$$f(x) = \operatorname{sgn}\left(w \cdot \phi(x) + b\right) = \operatorname{sgn}\left(\sum_{i=1}^{\ell} y_i a_i k(x, c_i) + b\right),$$

و برنامه های درجه دوم به صورت زیر است :

$$\text{maximize } W(a) = \sum_{i=1}^{\ell} a_i - \frac{1}{2} \sum_{i,j=1}^{\ell} a_i a_j y_i y_j k\big(Q_i(x), Q_j(x)\big),$$

$$\text{subject to } \quad a_i \geq 0, \ i = 1, \ldots, \ell, \quad \text{and} \quad \sum_{i=1}^{\ell} a_i y_i = 0.$$

همانطور که گفته شد، همچنین مدل محلی می تواند از کتاب کد ساخته شود. در این مورد، ایده اصلی آن به جای ارزش های کلیدی از یک فضای برداری چند بعدی اصلی با ارزش ها از یک فضا گسسته در ابعاد پایین تر است. بردار پایین بعد نیاز به فضای ذخیره سازی کمتر و در نتیجه داده فشرده دارد.

یک دنباله آموزش متشکل از M بردار منبع، $T = \{x_1, x_2, \ldots, x_m\}$ در نظر بگیرید.

فرض می شود M به اندازه کافی بزرگ است، به طوری که تمام خواص آماری از منبع توسط توالی آموزش دستگیر شده است. ما فرض کنیم که بردار منبع K بعدی هستند ، $X_m = (x_{m,1}, x_{m,2}, \ldots, x_{m,k}), \ m = 1, 2, \ldots, M$

,این بردار ها با انتخاب نزدیکترین بردار تطبیق فشرده، و به شکل یک کد متشکل از مجموعه ای کامل از بردارهای کد.می باشد N تعداد بردارهای کد است، $C = \{c_1, c_2, \ldots, c_n\}$ و هر یک از بردار های کد k بعدی است، $c_n = (c_{n,1}, c_{n,2}, \ldots, c_{n,k}), \ n = 1, 2, \ldots, N$ بردارهای کد نماینده نزدیک ترین فاصله اقلیدسی از بردار منبع می باشد. فاصله اقلیدسی تعریف شده است :

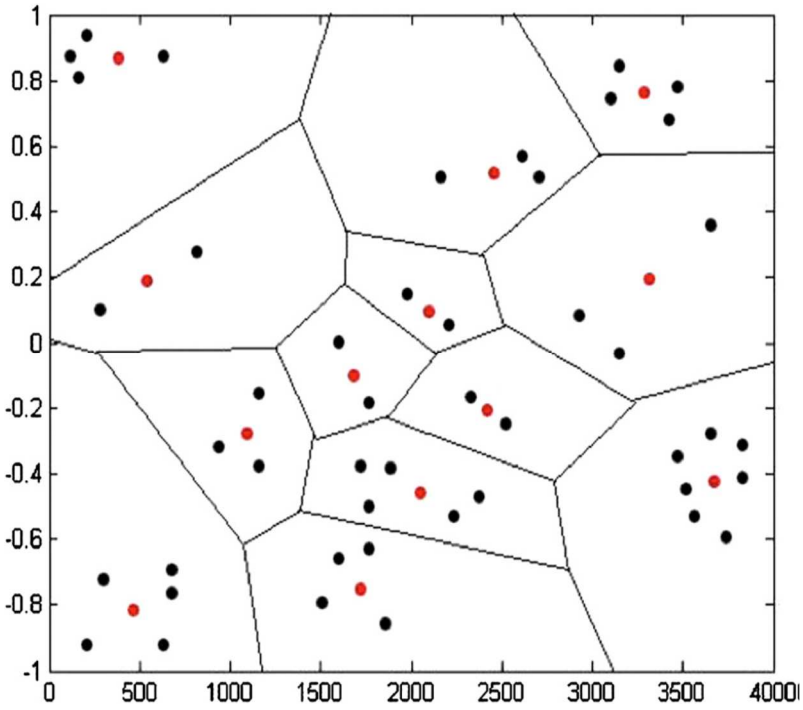
$$d(x, c_i) = \sqrt{\sum_{j=1}^{k} (x_j - c_{ij})^2},$$

که در آن X_J جزء j ام بردار منبع، C_{IJ} است که جزء j ام بردارهای کد، C_I به S_N منطقه نزدیکترین همسایه با بردارهای کد C_N است، و پارتیشن کل منطقه توسط $P = \{S_1, S_2, \ldots, S_N\}$ می باشد. اگر بردار منبع X_M در منطقه C_N

باشد، تقریب آن را می توان توسط $Q(X_m) = c_n, if\ X_m \in S_n$ بدست آورد، اگر $X_m \in S_n$. باشد . منطقه ورونی به صورت زیر تعریف می شود :

$$V_i = \{x \in R^k:\ \|x - c_i\| \leq \|x - c_j\|,\ \text{for all } j \neq i\},$$

بردار های آموزشی در یک منطقه ویژه قرار می گیرند و با یک نقطه قرمز مرتبط با ان نقطه تقریب می شوند



شکل1. دو بعدی (D2) تدریج بردار. (برای تفسیر رجوع به رنگ در این شکل، خواننده به نسخه وب این مقاله اشاره شده است.)

برای پیدا کردن C مطلوب و P، بردار با استفاده از یک اندازه گیری مربعات خطا که دقیقا چگونگی نزدیک به تقریب رت مشخص کرده است. اندازه گیری مربعات به صورت زیر آورده شده است :

$$D_{ave} = \frac{1}{Mk} \sum_{m=1}^{M} \|X_m - Q(X_m)\|^2$$

اگر پارامترهای C و P راه حلی برای به حداقل رساندن مشکل باشند، سپس آن ها باید دو شرط را برآورده سازدند: (1) نزدیکترین همسایه و (2) مرکز . نزدیکترین همسایه نشان می دهد که ناحیه S_N باید از همه بردار که نزدیک تر از هر یک از بردارهای کد دیگر به C_N تشکیل شده است :

$$S_n = \{x: \|x - c_n\|^2 \leq \|x - c_{n'}\|^2, \forall n' = 1, 2, \ldots, N\},$$

در نهایت، شرایط مرکز نشان می دهد که C_N بردارهای کد را می توان از میانگین تمام بردارهای آموزش در منطقه ورونی آن S_N گرفته شده است :

$$c_n = \frac{\sum_{X_m \in S_n} X_m}{\sum_{X_m \in S_n} 1}, \quad n = 1, 2, \ldots, N.$$

همانطور که الکان بحث کرده است، تکنیک های یادگیری محلی با استفاده از وکتورهای C_N برای ایجاد یک مدل محلی - ثابت است که هر یک از مدل های یادگیری هوشمند به بررسی تمام ارزش خواص هر مثال آموزشی باید پیچیدگی است. به عبارت دیگر، چنین استراتژی یادگیری محلی به مراتب بیشتر و کارآمد تر از استراتژی یادگیری جهانی می باشد، به ویژه در حجم زیادی از مشکلات داده.

5.3. یادگیری عمیقی

مدل های یادگیری کم عمق (به عنوان مثال، SVM، MLP، و GMM) به طور گسترده ای در صنعت برای حل مشکلات ساده استفاده شده است و یا به خوبی محدود شده است. با این حال، مدل سازی محدود و توان نمایشی مشکلات پیچیده تر مانند مشکلات زبان طبیعی را پشتیبانی نمی کند. در سال 2006، به اصطلاح یادگیری عمیقی (یادگیری a.k.a. Representation) به عنوان منطقه جدید از تحقیقات ML که سوء استفاده از لایه های متعدد از پردازش اطلاعات در یک معماری سلسله مراتبی برای طبقه بندی الگو و یا نمایندگی یادگیری پدید آمده است (به عنوان مثال، شبکه های عصبی). مزیت اصلی یادگیری عمیقی که به شدت افزایش یافته ، توانایی های پردازش تراشه،کاهش هزینه سخت افزار، و پیشرفت های اخیر در ML نامیده می شود.

شبکه های عصبی عمیق (DNNs) شبکه های چند لایه با لایه های بسیاری پنهان هستند، که وزن به طور کامل متصل و اغلب مقدار دهی اولیه و یا با استفاده از پیش آموزشی انباشته محدود ماشین بولتزمن (RBM) و یا شبکه های باور عمیق (DBMS) می باشند. DBM یک گام بدون نظارت پیش آموزشی که با بهره گیری مقدار زیادی از داده های آموزشی سازمان ملل متحد برای استخراج ساختار و قواعد در ویژگی های ورودی DBN است .نه تنها با استفاده از یک مقدار زیادی از داده های آموزشی بلکه وزن دهی اولیه خوبی برایDNN فراهم می کند . علاوه بر این، اتصالات و مشکلات اتصالات را می توان با استفاده از مرحله قبل از آموزش DNN برطرف کرد. DBN عملکرد عالی در به رسمیت شناختن و طبقه بندی کارها، از جمله پردازش زبان طبیعی، طبقه بندی تصویر، و تشخیص جریان ترافیک نشان داده است. با این حال، DNN هزینه محاسباتی بالا دارد . DSN ، مشکل مقیاس پذیری DNN می باشد، طبقه بندی ساده در بالای هر یک دیگر به منظور ساختار طبقه بندی پیچیده تر انباشته است.

تکنیک های جدید مورد استفاده در بخش 3.3 و 3.4 می تواند با مشکلات DNN به طور طبیعی متناسب باشد. تابع تصمیم DNN به شرح زیر است :

$$P_j = \frac{\exp(x_j)}{\sum_k \exp(x_k)},$$

که P_J نشان دهنده احتمال کلاس و X_J و X_K نشان دهنده ورودی کل به واحدهای J و K است. آنتروپی متقابل است به شرح زیر تعریف می شود :

$$C = \sum_j d_j \log(p_j),$$

که در آن D_J نشان دهنده احتمال هدف برای واحد خروجی J، و P_J خروجی احتمال J پس از استفاده از تابع فعال است. در حال حاضر، تقریب مدل نیمه پارامتری جدید که به عنوان تقریب محاسبه می شود :

$$\frac{\exp(c_j)}{\sum_k \exp(c_k)} \approx \frac{\exp(x_j)}{\sum_k \exp(x_k)},$$

این تقریب بدترین حالت X است و در حال حاضر قادر به کاهش پیچیدگی آن است. همانطور که در استراتژی یادگیری محلی، مفروضات قوی تر از مدل های ناپارامتری هستند، اما کمتر از مدل پارامتری محدود می شوند در حالی که کاهش پیچیدگی محاسباتی به میزان قابل توجهی می باشد.

6.3. محاسبات داده های بزرگ

داده های بزرگ سیستم های کامپیوتری به دو خوشه عمده تقسیم می شوند، بر اساس چگونه داده ها با توجه به محدودیت زمانی مورد تجزیه و تحلیل قرار می گیرند. اول، پردازش خوشه ای حجم زیادی از داده ها روی دیسک با محدودیت های زمانی (به عنوان مثال، MapReduce و GraphLab). دوم، جریان پردازش در حافظه داده ها در زمان واقعی و یا دوره کوتاه زمانی (به عنوان مثال، طوفان، ساموآ). در، هوانگ و لیورگاد که سیستم های محاسباتی نسل بعدی برای تجزیه و تحلیل داده های بزرگ نیاز به طرح های ابتکاری در سخت افزار و نرم افزار است که یک بازی خوب بین الگوریتم های داده بزرگ و اساسی منابع محاسباتی و ذخیره سازی فراهم می کند.

چند چارچوب محاسباتی ، به عنوان مثال، Hadoop، SHadoop، ComMapReduce، Dryad، Piccolo و IBM در یادگیری ماشین موازی وجود دارد ، سیستم دارای این توانایی برای مقیاس یادگیری ماشین است. ترکیبی از یادگیری عمیق و تکنیک های اجرای آموزش های موازی راه های بالقوه برای پردازش داده های بزرگ فراهم می کنند. کواک وی.لی و همکاران مشکل ساخت و ساز سطح بالا، آشکارسازها و ویژگی های کلاس خاص از داده ها بدون برچسب را در نظر گرفته اند. نتایج تجربی نشان می دهد که ممکن است آموزش یک آشکارساز بدون نیاز به تصاویر حاوی یک عنوان باشد.

K.Zhang و X.Chen یک الگوی یادگیری توزیعی برای RBMs و الگوریتم پس انتشار با استفاده از MapReduce ارائه داده اند. DBNs توسط انباشته یک سری از RBMs توزیع شده است و برای پیش آموزشی و توزیع پس انتشار برای ریز تنظیم آموزش داده است. نتایج تجربی نشان می دهد که توزیع RBMs و DBNs با عملکرد خوب از نظر دقت و بهره وری متمایل به داده ها در مقیاس بزرگ هستند.

4. نتیجه گیری

در این بررسی، ما یک نمای کلی از وضعیت فعلی پژوهش در مدل سازی داده ها پایدار ارائه داده ایم. به طور خاص، ما بحث جنبه های نظری و تجربی خود را در زمینه های داده های فشرده در مقیاس بزرگ، مربوط به: (1) بهره وری مدل انرژی ، از جمله نیازهای محاسباتی در یادگیری، و (2) مناطق فشرده اطلاعات ، ساختار و طراحی، از جمله مسائلی دوباره بین مدل های داده ها و ویژگی های آن انجام داده ایم.

با افزایش اطلاعات الکترونیکی ، مدل سازی داده ها پایدار برای ارائه یک راه رو به جلو با توجه به سهولت در دست زدن به مقادیر زیادی از داده ها نشان داده شده است. همچنین پیش بینی شده است که انقلاب مدل سازی داده می تواند به آسانی به مناطق مختلف در علم الکترونیک افزایش یابد. این مدل داده پایدار به تازگی طراحی شده نه تنها قادر به مقابله با پارادایم داده ها در مقیاس بزرگ است، بلکه وسیله ای در به حداکثر رساندن بازگشت آن برای مناطق مختلف علم الکترونیک می باشد.