روش های مدل سازی تنک به کمک یادگیری ماشین

محمد حسن شماخی[1]

[1]دانشگاه صنعتی امیرکبیر، mh.shammakhi@aut.ac.ir

چکیده- در این گزارش قصد داریم به نحوه استفاده از مدل های یادگیری ماشین در حل مسائل مختلف پرداخته و سپس به طراحی و توسعه چندین روش جدید در یادگیری تنک اشاره کنیم. در یادگیری تنک قصد داریم مدل هایی ارائه کنیم تا از میان ویژگی های ممکن در مسئله و مشاهدات صورت گرفته تعدادی از آنها را طوری انتخاب کند که توانایی پیش بینی مشاهدات مسئله را داشته باشد به طوری که با قرار دادن ورودی جدید در مدل، پیش بینی مناسبی برایمان به همراه داشته باشد.

همچنین این مدل ها را در دو طیف مختلف مسئله یعنی مسائل دسته بندی و رگرسیون بررسی خواهیم نمود و به توضیح دو روش پیاده سازی شده در یک سال اخیر خواهیم پرداخت. در آخر دو روش از مجموعه روش های مدل سازی تنک را توضیح خواهیم داد.

کلمات کلیدی- یادگیری ماشین، یادگیری تنک، پیش بینی

۱- مقدمه

اکثر مسائل موجود در عالم طبیعی به عوامل بسیار زیادی وابسته هستند که برخی از این عوامل تاثیر زیادی بر روی هدف مورد نظر ما دارد درحالی که برخی تاثیر کمتر برروی خروجی مسئله مورد نظر دارد.

یادگیری ماشین، علم مطالعه الگوریتم هایی است که عملکرد آن ها با ارائه داده به آن ها بهبود می یابد. در یادگیری ماشین، نرم افزار می آموزد که خروجی خود را با داده های آموزشی که خروجی درست آن ها برای الگوریتم فراهم شده است، تطبیق دهد. در این روش ها تلاش می شود با دانستن تعدادی از این پارامتر های مشاهده شده از داده ها به تخمینی مناسب از خروجی برسیم که این علاقه ناشی از این است که در خیل عظیمی از داده اطلاعات مهمی نهفته می باشد که بشر قادر به تشخیص آن نیست و یا اینکه ممکن است موقع طراحی یک سیستم تمامی ویژگیهای آن شناخته شده نباشد در حالیکه ماشین می تواند حین کار آنها را یاد بگیرد، همچنین ممکن است محیط در طول زمان تغییر کند. ماشین می تواند با یادگیری این تغییرات خود را با آنها وفق دهد.

روش های یادگیری را بر اساس معیار های مختلف می توان دسته بندی کرد.

یکی از این معیار ها دسته بندی براساس هدف خروجی می باشد که به دسته های یادگیری با ناظر ، بدون ناظر و نیمه نظارتی تقسیم کرد. در یادگیری با ناظر هدف پیدا کردن یک برچسب یا مقدار معین به ازای هریک از داده های ورودی می باشد مانند مسئله جستجوی مکان با استفاده از آنتن های بدون سیم. در مسائل بدون ناظر، هدف پیدا کردن تشابه موجود بین داده ها است، تا بتوان از میان مجموعه داده ها تعدادی از آنها را به عنوان نماینده آنها انتخاب کرد مانند سرویس های الهام گیرنده در درک میزان تشابه بین تصاویر یا تشخیص مطالب مورد علاقه شما در شبکه های اجتماعی نظیر فیس بوک، اینستوگرام و غیره. در مسائل نیمه نظارتی مانند یادگیری نظارتی هدف پیدا کردن برچسب یا حل مسئله ای رگرسیون می باشد در حالی که به دلایل مختلف از جمله نبود امکان جمع آوری داده ها به صورت کامل، برخی از داده ها ناقص بوده که نیاز به بازسازی آنها می باشد

از طرفی دیگر در بکارگیری روش های یادگیری می توان روش بکارگرفته شده را به دسته های یادگیری استنتاجی، یادگیری ارتباطی، یادگیری بیزین، یادگیری تقویتی و یادگیری تکاملی تقسیم کرد.

این روش ها به سه دسته پارامتریک غیر پارامتری و همچنین محلی تقسیم می شوند.

روش های پارامتری مانند مدل های خطی به دنبال بدست آوردن چند پارامتر و متغیر به منظور رسیدن به یک مدل تعیین کننده هستیم تا از آن برای تخمین خروجی مورد نظر بر اساس ورودی ها استفاده کنیم

فرض کنید که می خواهیم یک مدل برای درک میزان فروش یک محصول بر اساس میزان هزینه تبلیغاتی در رادیو بدست آوریم پس از استفاده از داده هایی از موارد مشاهده شده از میزان فروش بر این اساس تصمیم به آموزش مدلی داریم که ما را به تخمینی از میزان خروج برساند.

در این نمودار محور افقی میزان هزینه انجام شده جهت تبلیغات در رادیو و محور عمودی میزان فروش کالای مورد نظر است.

روش های غیرپارامتری مانند روش ماشین بردار پشتیبان[۱] و ماشین بردار مرتبط[۲] هستند که با استفاده از خود داده ها به مدلی تعیین کننده برای رسیدن به خروجی مورد نظر می باشد وظیفه این روش ها استفاده از داده های مناسب از بین تمام مجموعه داده ها به منظور استفاده دقیقتر از داده های مشاهده شده می باشد که ضعف اصلی این روش ها مشکل بیش تطبیقی[۳] و یا عجز روش در موقعیت هایی است که داده های مشاهده شده در آن ناحیه وجود ندارد.

روش ماشین بردار پشتیبان از جمله روش‌های نسبتاً جدیدی است که در سال‌های اخیر کارایی خوبی نسبت به روش‌های قدیمی‌تر برای طبقه‌بندی از جمله شبکه‌های عصبی پرسپترون[۴] نشان داده است. مبنای کاری آن دسته‌بندی خطی داده‌ها است که سعی در تقسیم خطی داده‌ها دارد بطوریکه خط انتخاب شده حاشیه اطمینان بیشتری داشته باشد. حل معادلهٔ پیدا کردن خط بهینه برای داده‌ها به وسیله روش‌های برنامه ریزی درجه دوم[۵] که روش‌های شناخته شده‌ای در حل مسائل محدودیت‌دار هستند صورت می‌گیرد. قبل از تقسیم خطی برای اینکه ماشین بتواند داده‌های با پیچیدگی بالا را دسته‌بندی کند داده‌ها را به وسیله ی تابع φ به فضای با ابعاد خیلی بالاتر می‌بریم. برای اینکه بتوانیم مساله ابعاد خیلی بالا را با استفاده از این روش‌ها حل کنیم از قضیه دوگانی لاگرانژ برای تبدیل مساله ی مینیمم سازی مورد نظر به فرم دوگانی آن که در آن به جای تابع پیچیده ی φ که ما را به فضایی با ابعاد بالا می‌برد، تابع ساده‌تری به نام تابع هسته که ضرب برداری تابع φ است ظاهر می‌شود استفاده می‌کنیم. از توابع هسته مختلفی از جمله هسته‌های نمایی، چندجمله‌ای و سیگموید می‌توان استفاده نمود.

۲- یادگیری ماشین و کاربرد های آن

از یادگیری ماشین در مسائل مختلف روز بسیار استفاده می گردد که از جمله کاربرد های مخابراتی آن می توان تشخیص محل افراد با توجه اطلاعاتی که انتن های اطراف از فرد می گیرند؛ تشخیص متن از درون تصویر و یا مسایل پردازش گفتار از جمله تغییر صدا[۶]، شناسایی صدا[۷] ، تایید صدا[۸] را نام برد.

۳- روشی ابتکاری در تشخیص عدد در تصویر

در این بخش به شرح روشی برای تشخیص اعداد دستنویس خواهیم پرداخت که با حفظ دقت، ابعاد و پیچیدگی مدل یادگرفته شده را تا حد امکان کاهش می دهد. روش ارائه شده در ابتدا با استفاده از الگوریتم انتخاب ویژگی بیشترین ارتباط-کمترین حشو[۹] و شبکه پرسپترون[۱۰] که محاسبات نسبتا کمی دارد، ویژگی های مناسب برای دسته بندی حروف را می یابد و در انتها با ویژگی های به دست آمده و ترکیب دو دسته بند پرسپترون و ماشین بردار پشتیبان، یک مدل دقیق و نسبتا ساده آموزش داده می شود. الگوریتم بر روی دو دادگان واقعی ORHD و MNIST تست شده است و با کاهش قابل ملاحظه پیچیدگی مدل، نسبت به روش های مشابه به ترتیب به دقت های قابل قبول ۹۶,۱ و ۹۸,۱۴ رسیده است.

در بسیاری از مقالات از تمامی ویژگی های خام تصاویر برای دسته بندی استفاده شده است. یک عیب مهم اینگونه روش ها پیچیدگی دسته بند می باشد، زیرا ابعاد تصاویر غالبا بزرگ بوده و باعث می شود که مدل آموزش داده شده دارای ابعاد زیاد و پیچیدگی بالایی باشد. علاوه بر این با افزایش بعد، قدرت تعمیم دهندگی روش در مواجهه با داده های تست جدید کاهش می یابد.

در این مقاله روشی ارائه شده است که در عین دارا بودن دقت بالا از تعداد محدودی از ویژگی های خام تصویر به طور کارآمد استفاده می کند.

حال به نحوه گزینش ویژگی ها در تصویر با معیار mRMR و همچنین کاهش پیچیدگی محاسبات خواهیم پرداخت. سپس با بکارگیری روش ترکیبی MLPماشین بردار پشتیبان و نحوه ترکیب mRMR و روش ترکیبی MLPماشین بردار پشتیبان روش ابتکاری مد نظر را پیادسازی کرده و نتایج بدست آمده از ترکیب متد

---

۱ SVM

۲ RVM

۳ Over fit

۴ Perceptron

۵ Quadratic Programing

۶ Voice conversion

۷ Voice identification

۸ Voice verification

۹ mRMR

۱۰ Perceptron

گزینش ویژگی مطرح شده با روش MLPماشین بردار پشتیبان را خواهیم گفت.

۳-۱- روش گزینش ویژگی mRMR

در مسائل کلاس بندی اغلب به دنبال بهترین مدل با بکارگیری بهترین زیرمجموعه از ویژگی ها هستیم که بتوانند مدلی مناسب به منظور کاهش خطا ارائه دهند. در ایتدا به بیان روشی می پردازیم که با معیاری متفاوت، بهترین زیر مجموعه از ویژگی ها را انتخاب می کند. این روش که با نام mRMR می باشد کارآمدی خود را در مقاله [۱] به اثبات رسانده است در این روش به دنبال m ویژگی ای هستیم که بیشترین وابستگی بین آن زیر مجموعه از ویژگی ها و کلاس خروجی وجود داشته باشد. به ازای m برابر ۱ معرف بهترین ویژگی و برای m بزرگتر از ۱ بهترین زیرمجموعه m تایی از ویژگی ها را معرفی کند. در این مقاله با توجه به اینکه داده مورد نظر ما داده های خام تصاویری از دست خط هایی است که می خواهیم بازشناسایی کنیم، بنابراین هر ویژگی معرف یک پیکسل می باشد لذا به دنبال روشی هستیم که بهترین زیر مجموعه از پیکسل ها را انتخاب کرده و از آنها برای تشخیص عدد پیش رو استفاده کند.

معیاری که برای توضیح میزان وابستگی در نظر گرفتیم اطلاعاتی است که بین مجموعه پیکسل انتخاب شده با کلاس خروجی وجود دارد یعنی به دنبال زیرمجموعه ای از پیکسل ها با بیشترین وابستگی به کلاس خروجی هستیم:

$$\max D(S,c)\ ,\ D = I(\{x_i, i=1\ ,\dots,m\};c) \qquad (1)$$

که اطلاعات متقابل موجود بین دو مقدار تصادفی x و y به صورت زیر می باشد.

$$I(x;y) = p(x,y)\log\iint\left(\frac{p(x,y)}{p(x)p(y)}\right)dx\,dy \qquad (2)$$

حال برای رسیدن به بیشترین وابستگی به تحلیل فرمول ۱ می پردازیم.

$$I(S_m;c) = \iint p(S_m,c)log\left(\frac{p(S_m,c)}{p(S_m)p(c)}\right)dS_m dc$$
$$= \iint p(S_{m-1},x_m,c)log\left(\frac{p(S_{m-1},x_m,c)}{p(S_{m-1},x_m)p(c)}\right)dS_{m-1}dx_m dc$$
$$= \int\dots\int p(x_1,x_2,\dots,x_m,c)\log\left(\frac{p(x_1,x_2,\dots,x_m,c)}{p(x_1,x_2,\dots,x_m)p(c)}\right) \qquad (3)$$
$$dx_1dx_2\dots dx_m dc$$

مشکلی که وجود دارد محاسبه چگالی احتمال مشترک بین مجموعه پیکسل ها و برچسب های خروجی است اول اینکه تعداد نمونه ها برای محاسبه تابع چگالی مشترک کم است و دوم اینکه بدست آوردن تابع چگالی مشترک نیاز به معکوس کردن ماتریس کواریانس دارد که اگر فرض کنیم k ویژگی انتخاب شده داشته باشیم که هرکدام a حالت مختلف بتواند داشته باشد در صورت $\min(a^k, N)$ حالت مختلف خواهیم داشت. که N تعداد کل نمونه ها می باشد. به همین خاطر در نظر می گیریم.

$$\max D(S,c)\ ,\qquad D = \frac{1}{|S|}\sum_{i=1}^{m} I(x_i,c) \qquad (4)$$

این در حالی است که باید در نظر داشت که بدنبال پیکسل هایی هستیم که هم بیشترین اطلاعات بین آنها و برچسب های خروجی وجود داشته باشد و هم اطلاعات و محتوای هر پیکسل کمترین شباهت را با پیکسل های دیگر داشته باشد. یعنی:

$$\min R(S)\quad ,\quad R = \frac{1}{|S|^2}\sum_{x_i,x_j\in S} I(x_i,x_j) \qquad (5)$$

به همین منظور مقدار $\Phi = D - R$ را حداکثر می سازیم.

در مقاله [۱] ثابت می شود که D-R بهتر از دیگر توابع جهت ماکسیم کردن ترکیبی از D و R در راستای تحقق هدف مورد بحث می باشد و بهترین پارامتر برای بهینه سازی را به ما می دهد.

برای رسیدن به این مهم فرض می کنیم بهترین m-1 پیکسل که بیشترین وابستگی با کلاس خروجی را طبق معیار Φ دارد انتخاب کرده باشیم می توان ثابت کرد برای بدست آوردن m امین ویژگی برای ساخت m پیکسل حیاتی باید به دنبال پیکسلی با بیشترین مقدار پارامتر زیر در بین پیکسل های کاندید باشیم.

$$\max_{x_j\in X-S_{m-1}}[\,I\,(x_j;c) - \frac{1}{m-1}\sum\nolimits_{x_i\in S_{m-1}} I(x_j,x_i)] \qquad (6)$$

این مدل را می توان با بکارگیری روش های wrapper مختلف اعمال کرد که ما در روش خود از مدل انتخاب پیش رو[۱۱] استفاده خواهیم کرد که موجب ایجاد رابطه ۷ بین زیر پیکسل های انتخاب شده به ازای m های مختلف می شود.

$$S_1 \subset S_2 \subset S_3 \subset \dots \subset S_n \qquad (7)$$

۳-۲- روش MLPماشین بردار پشتیبان و نحوه بکارگیری آن

همانطور که در بخش مقدمه شرح داده شد یکی از مشکلات الگوریتم های بازشناسایی متن و اعداد این است که متد بکارگرفته شده برای تشخیص عدد برای تصاویر نامطلوب عملکرد مناسبی ندارند این در حالی است که بیشتر اشتباهات الگوریتم ها در

[۱۱] forward selection

تشخیص این دسته از داده ها بین زوج اعداد مشابه نظیر (7 , 1) ، (8 , 6) و غیره می باشد. در این مقاله می خواهیم این مسئله بسیار حیاتی را حل کنیم به این صورت که ابتدا با استفاده از روش MLP (که در [۲] به طور کامل معرفی شده است) که به تعداد پیکسل های ورودی گره[۱۲] ورودی ، یک لایه پنهان[۱۳] ، ۵۰ گره در آن لایه پنهان و ۱۰ گره خروجی که نشان دهنده امتیاز هریک از گره ها که نماینده یکی از برچسب های ۰ تا ۹ می باشد شبکه را آموزش می دهیم سپس با توجه به مقادیر امتیازات گره های خروجی دو برچسبی که گره هایی آنها بیشینه امتیازات را دارند استفاده می کنیم اگر برای هر کدام از داده های تست اختلاف بیشینه اول و دوم امتیازات بیش از یک مقدار آستانه ای (۰,۴ در آزمایشات استفاده شده است) بود با قطعیت می گوییم برچسب مربوط به گره‌ی با بیشترین امتیاز برچسب خروجی است اما اگر برای داده تستی این اختلاف کمتر از میزان آستانه بود برای تشخیص بین برچسب بیشینه اول امتیازها و بیشینه دوم از روش ماشین بردار پشتیبان [۳] استفاده می کنیم که این روش ترکیبی در [۴] به طور مفصل پرداخته شده است در این حالت تفاوت بسیار زیادی در خطای نهایی ایجاد می شود و به میزان مناسبی خطا کاهش می یابد.

۳-۳- *بکارگیری روش ترکیبی MLPماشین بردار پشتیبان با استفاده از متد mRMR*

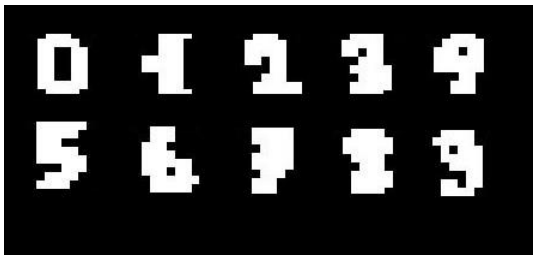
پس از بیان میزان کارآمدی مطلوب روش MLPماشین بردار پشتیبان برای شناسایی تصاویر عددی به بیان روش ترکیبی بین مدل انتخاب ویژگی (پیکسل) mRMR و ترکیب آن با روش MLPماشین بردار پشتیبان خواهیم پرداخت به این صورت که از یک انتخاب بهترین پیکسل با معیار mRMR شروع به انتخاب بهترین زیرمجموعه از پیکسل ها کرده و بر آن زیرمجموع از پیکسل ها از مجموعه تمام پیکسل ها روش MLPماشین بردار پشتیبان را آموزش می دهیم. سپس خطای ناشی از این مدل بر روی داده های آموزش را محاسبه کرده و چنانچه این خطا از یک خطای آستانه ای کمتر شد عمل انتخاب پیشرو را قطع کرده و آن زیرمجموعه m عضوی از پیکسل ها را به عنوان بهترین زیرمجموعه از پیکسل ها در نظر می گیریم

۳-۴- *پیاده سازی روش ابتکاری*

روش ابتکاری خود را بر روی دو دادگان دست خط اعداد با ویژگی های کاملا متفاوت امتحان می کنیم.

۳-۴-۱- *دادگان ORHD*

این دادگان که شامل ۵۶۲۰ تصویر ۸در۸ از اعداد است شامل ۳۹۷۶ داده آموزش و ۱۷۹۷ داده تست می شود. نمونه هایی از این دادگان شامل اعداد ۰ تا ۹ در شکل زیر آمده است.[۷]



شکل ۱: نمونه ای از تصاویر دادگان ORHD

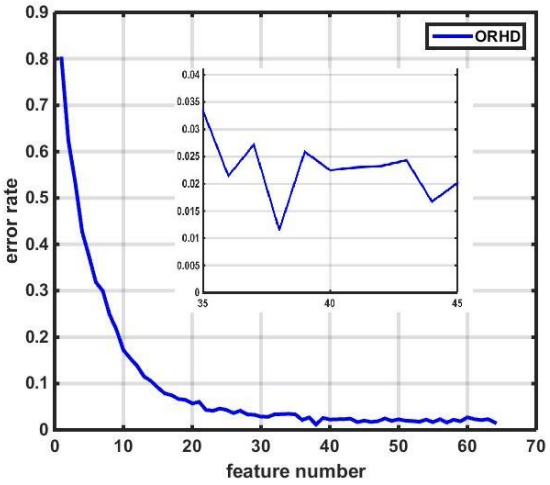
برای پیاده سازی الگوریتم خود بر این دادگان ابتدا سطح خاکستری[۱۴] پیکسل ها را که مقداری بین ۰ تا ۱۵ می تواند داشته باشد را به صورت logical به ۰ و ۱ تغییر می دهیم به این صورت که هر بک از پیکسل ها با سطح خاکستری کمتر ۸ مقدار ۰ و به پیکسل هایی که سطح خاکستری بیش از ۸ داشته باشند مقدار ۱ می دهیم سپس حال به انتخاب زیر پیکسل ها با معیار mRMR می پردازیم و با درنظر گرفتن ۲ درصد خطای آستانه برای داده آموزش از روی نمودار ۱ می توان فهمید که الگوریتم با انتخاب ۳۸ پیکسل از ۶۴ پیکسل به این خطای آستانه می رسد حال با استفاده از همین ۳۸ پیکسل شبکه MLPماشین بردار پشتیبان را برروی داده های تست اعمال می کنیم. در مقایسه نتایج بدست آمده با دیگر مقالات می توان دید دقت روش ارائه شده مطلوبیت کافی را داراست چرا که دقت روش ارائه شده بر داده های تست ۹۶,۱ ٪ می باشد که در مقایسه با روش های بیان شده در مقاله [۵] یعنی EWLDR (۹۳,۸۸) WLDR (۹۳,۹۳) ، ePAC LDA (۹۳,۸۸) (۹۳,۸۸) مناسب می باشد.

---

[۱۲] node

[۱۳] Hidden layer

[۱۴] graylevel



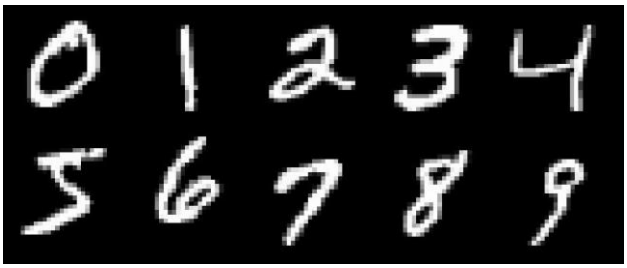
شکل ۲: نمونه ای از تصاویر دادگان ORHD

این در حالی است که ما فقط از ۳۸ پیکسل از ۶۴ پیکسل موجود استفاده کرده ایم در حالی که دیگر روش ها از تمام پیکسل ها استفاده کرده اند و همچنین در روشی مانند LDA از استخراج ویژگی[۱۵] نیز استفاده شده است.

علت این اختلاف قابل توجه این است که پس از بکارگیری شبکه عصبی MLP از ماشین بردار پشتیبان بر روی برخی داده های تست (با تصویر نامطلوب از عدد که تشخیص آن دشوار است) اعمال می کنیم تا تمییز مناسبی بین انتخاب دو برچسب با احتمال بیشتر به عنوان خروجی ایجاد کنیم که در پی اجرای این روند بر روی داده ها ۴ درصد خطا بهبود پیدا کرده است که تقریبا می توان گفت با این ایده خطا را به نصف کاهش داده ایم

۳-۲-۴- *دادگان MNIST*

این دادگان که در مقالات زیادی در زمینه بازشناسایی دست خط اعداد به کار گرفته شده است شامل ۷۰۰۰۰ تصویر ۲۸در۲۸ از اعداد است که تعداد ۶۰۰۰۰ داده آموزش و ۱۰۰۰۰ داده تست در این دادگان وجود دارد. نمونه ای از تصاویر ۰ تا ۹ این دادگان در شکل زیر آمده است.[۶]



شکل ۳: نمونه ای از تصاویر دادگان MNIST

برای این دادگان نکته قابل توجه وجود ۷۸۴ برای ۶۰۰۰۰ داده آموزش و ۱۰۰۰۰ داده تست است که حجم محاسبات بسیار زیادی را ایجاد می کند به همین منظور بدنبال راهی جهت کاهش این حجم بالا از محاسبات باید بود که این مهم به وسیله روش ارائه شده در تحقق خواهد یافت.

با توجه به وضوح و مطلوبیت بیشتر دست خط موجود در این دادگان مقدار آستانه در نظر گرفته شده برای این دادگان بازای مقادیر مختلف در جدول ۱ آمده است.

جدول ۱: **دقت الگوریتم ارائه شده بر روی مجموعه دادگان MNIST به ازای آستانه های مختلف**

خطای آستانه (٪)	تعداد پیکسل انتخاب شده	خطای داده تست
1	472	2.83
0.8	487	2.56
0.5	524	**1.86**

این در حالی است که در روش های مقالات دیگر استفاده از روشهایی نظیر شبکه عصبی با ۱۰۰ گره خروجی [۷] از تمام ۷۸۴ پیکسل برای تشخیص عدد استفاده شده است. و با این حال به خطایی بهتر از ۱,۴ درصد نرسیده اند. این در حالی است که پیچیدگی و حجم محاسبات آنها بسیار بیشتر بوده و در عین حال در روش ارائه شده فقط از ۵۲۴ از تمام ۷۸۴ پیکسل برای تشخیص عدد داخل تصویر استفاده کرده ایم.

۴- نمونه ای از روش های مدل سازی تنک

در دهه‌ی گذشته الگوریتم‌های آموزش ماشین بر مبنای روش-های بیزین به شدت مورد علاقه و توجه قرار گرفته است. در پژوهش‌های صورت گرفته، روش‌های مبتنی بر هسته تنک (نظیر ماشین بردار مرتبط و ماشین بردار پشتیبان) در حل مسائل رگرسیون و طبقه‌بندی کارایی بالایی از خود نشان داده‌اند. خاصیت تنک بودن که به عنوان یک دانش پیشین به این الگوریتم‌ها اعمال می‌شود علاوه بر کاهش پیچیدگی مدل و هزینه محاسباتی در مرحله‌ی آزمایش و پیش‌بینی، از بیش‌تطبیقی جلوگیری می‌کند و قابلیت تعمیم‌پذیری این الگوریتم‌ها را افزایش می‌دهد. در سال‌های اخیر از روش‌های آموزش بیزین تنک برای بازیابی سیگنال‌های تنک از روی اندازه‌گیری‌های آن‌ها استفاده شده است. از مزایای این روش‌ها می‌توان به تخمین همزمان سیگنال مورد نظر به همراه خطا

[۱۵] Feature extraction

که معیاری برای مورد اطمینان بودن اندازه‌گیری در بازیابی سیگنال در اختیار ما قرار می‌دهد، اشاره نمود.

الگوریتم آموزش ماشین ماشین بردار مرتبط که در سال ۲۰۰۱ توسط تیپینگ[۱۶] ارائه شد [۸]، یک چارچوب بیزین کلی برای به دست آوردن یک پاسخ تنک در مسائل رگرسیون و طبقه بندی با استفاده از مدل های خطی نسبت به پارامترها را بیان می کند. الگوریتم ماشین بردار مرتبط را می توان با ماشین بردار پشتیبان [۹] مقایسه نمود. آنچنانکه در پژوهش ها گزارش شده است ، ماشین بردار مرتبط می تواند نتایجی مشابه یا بهتر از ماشین بردار پشتیبان ارائه دهد در حالی که از تعداد بسیار کمتری توابع پایه استفاده می کند. به عبارت دیگر پاسخ ماشین بردار مرتبط بسیار تنک تر از پاسخ ماشین بردار پشتیبان است. علاوه بر این می توان مزایای دیگری نظیر پیش بینی احتمالاتی، تخمین پارامترهای نویز و امکان استفاده از توابع پایه دلخواه (که لزوماً شرط Mercer را تامین نمی کنند) برای ماشین بردار مرتبط در مقایسه با ماشین بردار پشتیبان برشمرد. همچنین در ماشین بردار مرتبط بر خلاف ماشین بردار پشتیبان نیازی به تنظیم پارامتر C برای مصالحه خطا-حاشیه در مرحله آموزش نیست.

در این بخش ابتدا مدل بیزین ماشین بردار مرتبط برای مسائل رگرسیون بیان می شود و مراحل به دست آوردن هایپرپارامترها و سپس وزن ها با استفاده از آن ها ذکر می شود. در ادامه نحوه استفاده از ماشین بردار مرتبط برای مسائل طبقه بندی بیان خواهد شد.

۴-۱- *آموزش بیزین تنک برای رگرسیون*

یک مجموعه داده ورودی-خروجی را در نظر بگیرید که x_n بردار ورودی و t_n مقدار خروجی متناظر با آن را نشان می دهد. در مسائل استاندارد احتمالاتی فرض بر آن است که خروجی-های t_n نمونه هایی از مدل زیر هستند:

$$t_n = y(x_n, \mathbf{W}) + \varepsilon_n$$

که در این فرمول مدل مسئله مورد نظر را نشان می دهد و نویز جمع شونده است و نمونه های آن مستقل بوده و دارای توزیع گوسی با میانگین صفر و واریانس هستند.

در ماشین بردار مرتبط مدل پایه به صورت زیر تعریف می شود:

$$y(x, \mathbf{W}) = \sum_{i=1}^{N} w_i k(x, x_i) + w_0$$

در رابطه فوق یک تابع هسته بوده و برای هر نمونه آموزشی یک تابع پایه را تعریف می-کند. پارامترهای قابل تنظیمی هستند که وزن نامیده می شوند و هدف تخمین مقدار مناسب برای این پارامترها است. در ماشین بردار مرتبط نحوه تعیین وزن ها یک چارچوب کاملاً احتمالاتی دارد.

آنچنانکه گفته شد در مدل بیان شده نمونه های نویز مستقل و دارای توزیع هستند، لذا می توان توزیع احتمال شرطی خروجی متناظر با یک نمونه خاص را به صورت زیر نوشت:

$$p(t_n \mid x_n) = N(y(x_n, \mathbf{W}), \sigma^2)$$

با فرض استقلال زوج های ورودی-خروجی، احتمال فوق را برای همه ی نمونه های آموزشی (N نمونه) می توان از حاصلضرب رابطه ی فوق برای هر نمونه به صورت زیر به دست آورد:

$$p(\mathbf{t} \mid \mathbf{W}, \sigma^2) = \prod_{i=1}^{N} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \exp\left\{-\frac{1}{2\sigma^2}(t_i - \phi(x_i)\mathbf{W}^T)^2\right\}$$

$$= (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(t_i - \phi(x_i)\mathbf{W}^T)^2\right\}$$

$$= (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left\|(\mathbf{t} - \boldsymbol{\varphi}\mathbf{W}^T)\right\|^2\right\}$$

در رابطه فوق $\mathbf{t} = [t_1, t_2, ..., t_N]^T$ و $\boldsymbol{\varphi}$ یک ماتریس N×N+1 است که به صورت زیر تعریف می شود:

$$\boldsymbol{\varphi} = [\phi(x_1), \phi(x_2), ..., \phi(x_N)]^T$$

$$\phi(x_n) = [1, k(x_n, x_1), k(x_n, x_2), ..., k(x_n, x_N)]^T, n = 1, ..., N$$

از آنجایی که در رابطه فوق تعداد پارامترها با تعداد نمونه ها برابر است، بنابراین تخمین بیشترین شباهت w و σ^2 منجر به بیش تطبیق می شود. لذا برای مقابله با این مشکل از دیدگاه بیزین استفاده می شود و یک توزیع احتمال پیشین روی وزن ها به عنوان یک محدودیت اضافی روی پارامترها در نظر گرفته می شود. ترجیح بر آن است که مدلی با پیچیدگی کمتر یا تعداد توابع پایه کمتر به دست آید، این ویژگی با تنک بودن w مدل می شود. لذا برای w توزیعی به عنوان توزیع پیشین در نظر گرفته می-شود که در اطراف

---

[۱۶] Tipping

مقدار صفر به صورت تیز ماکزیمم شود.

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=0}^{N} N(w_i|0,\alpha_i^{-1}) = \prod_{i=0}^{N} (2\pi\alpha_i^{-1})^{-\frac{1}{2}} \exp\left\{-\frac{\alpha_i}{2}(w_i-0)^2\right\}$$
$$= (2\pi)^{-\frac{N+1}{2}} \left\{\prod_{i=0}^{N} \alpha_i^{\frac{1}{2}}\right\} \exp\left\{-\frac{1}{2}\mathbf{w}^T A\mathbf{w}\right\}$$

که $\boldsymbol{\alpha} = [\alpha_0, \alpha_1, ..., \alpha_N]^T$ بردار هایپرپارامترها بوده و به هر وزن w_i مستقلاً هایپرپارامتر α_i (دقت یا معکوس واریانس) اختصاص می یابد و $A = diag(\alpha_0, \alpha_1, ..., \alpha_N)$.

پس از تعریف احتمالات پیشین، با استفاده از قاعده بیز توزیع پسین روی تمام پارامترهای مجهول با فرض داشتن داده ها به دست می آید:

$$p(\mathbf{w}, \boldsymbol{\alpha}, \sigma^2 | \mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{w}, \boldsymbol{\alpha}, \sigma^2)\, p(\mathbf{w}, \boldsymbol{\alpha}, \sigma^2)}{p(\mathbf{t})}$$

سپس چنانچه داده تست x^* داده شده باشد، توزیع خروجی متناظر با آن یعنی t^* را می توان به صورت زیر پیش بینی نمود.

$$p(t^*|\mathbf{t}) = \int p(t^*|\mathbf{w}, \boldsymbol{\alpha}, \sigma^2)\, p(\mathbf{w}, \boldsymbol{\alpha}, \sigma^2|\mathbf{t})\, d\mathbf{w}\, d\boldsymbol{\alpha}\, d\sigma^2$$

اما متاسفانه رابطه فوق را نمی توان مستقیما محاسبه نمود. زیرا محاسبه انتگرال عامل نرمالیزه کننده زیر در سمت راست معادله ممکن نیست:

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w}, \boldsymbol{\alpha}, \sigma^2)\, p(\mathbf{w}, \boldsymbol{\alpha}, \sigma^2)\; d\mathbf{w}\, d\boldsymbol{\alpha}\, d\sigma$$

اما می توان توزیع پسین را به صورت زیر تجزیه نمود:

$$p(\mathbf{w}, \boldsymbol{\alpha}, \sigma^2 | \mathbf{t}) = p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}, \sigma^2)\; p(\boldsymbol{\alpha}, \sigma^2 | \mathbf{t})$$

عبارت اول سمت راست رابطه فوق را می توان به صورت زیر نوشت:

$$p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}, \sigma^2) = \frac{p(\mathbf{t}|\mathbf{w}, \sigma^2)\; p(\mathbf{w}|\boldsymbol{\alpha})}{p(\mathbf{t}|\boldsymbol{\alpha}, \sigma^2)}$$

در عبارت سمت راست رابطه فوق صورت کسر به راحتی از روابط گفته شده قابل محاسبه است. مخرج نیز از محاسبه انتگرال عبارت صورت که به صورت کانولوشن گوسی است به راحتی محاسبه می شود. بنابراین عبارت زیر قابل محاسبه است:

$$p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}, \sigma^2) = (2\pi)^{-\frac{N+1}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{w}-\mu)^T \Sigma^{-1}(\mathbf{w}-\mu)\right\}$$

و کواریانس و میانگین پسین عبارتند از:

$$\Sigma = (\sigma^{-2}\boldsymbol{\varphi}^T\boldsymbol{\varphi} + A)^{-1}$$

$$\mu = \sigma^{-2} \Sigma \boldsymbol{\varphi}^T \mathbf{t}$$

و اما در مورد عبارت دوم سمت راست رابطه فوق باید یک تقریب را پذیرفت و آن تقریب توزیع پسین روی هایپرپارامترها ($p(\boldsymbol{\alpha}, \sigma^2|\mathbf{t})$) با یک تابع ضربه در مقادیر با بیشترین احتمال آن است یعنی $p(\boldsymbol{\alpha}, \sigma^2|\mathbf{t}) \approx \delta(\boldsymbol{\alpha}_{MP}, \sigma^2_{MP})$

به این ترتیب ماشین بردار مرتبط، جستجو برای مقادیر با بیشترین احتمال پسین برای هایپرپارامترها است یعنی ماکزیمم کردن $p(\boldsymbol{\alpha}, \sigma^2|\mathbf{t})$ روی $\boldsymbol{\alpha}, \sigma^2$ از طرفی می دانیم:

$$p(\boldsymbol{\alpha}, \sigma^2|\mathbf{t}) \propto p(\mathbf{t}|\boldsymbol{\alpha}, \sigma^2)\, p(\boldsymbol{\alpha})\, p(\sigma^2)$$

با فرض توزیع یکنواخت برای کافی است که عبارت ماکزیمم شود.

$$p(\mathbf{t}|\boldsymbol{\alpha}, \sigma^2) = \int p(\mathbf{t}|\mathbf{w}, \sigma^2)\, p(\mathbf{w}|\boldsymbol{\alpha})\, d\mathbf{w}$$
$$= (2\pi)^{-\frac{N}{2}} \left|\sigma^2 I + \boldsymbol{\varphi} A^{-1} \boldsymbol{\varphi}^T\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{t}^T(\sigma^2 I + \boldsymbol{\varphi} A^{-1}\boldsymbol{\varphi}^T)^{-1}\mathbf{t}\right\}$$

در مدل بیزین به کمیت فوق شباهت حاشیه ای گفته می شود و ماکزیمم کردن آن را روش بیشترین شباهت نوع دوم می گویند.

مقادیر که رابطه فوق را ماکزیمم می کند به صورت بسته به دست نمی آید لذا از یک روش تخمین تکراری بیشترین امید ریاضی[۱۷] استفاده می شود.

با مشتق گیری از رابطه قبل نسبت به و برابر با صفر قرار دادن آن و بازنویسی مجدد با استفاده از روش ماکای[۱۸] رابطه زیر به دست می آید:

$$\alpha_i^{new} = \frac{\gamma_i}{\mu_i^2}$$

به طوری که میانگین پسین i امین وزن بوده و می توان آن را محاسبه کرد و کمیت γ_i به صورت زیر تعریف می شود:

$$\gamma_i \equiv 1 - \alpha_i \Sigma_{ii}$$

و α_i ، i امین عنصر قطری کواریانس پسین وزن ها است که از رابطه فوق و مقادیر فعلی محاسبه می شود.

برای واریانس نویز، مشتق گیری به معادله زیر برای به روز رسانی منجر می شود:

---

[۱۷] EM

[۱۸] Mackay

$$(\sigma^2)^{new} = \frac{\|\mathbf{t} - \boldsymbol{\varphi}\mu\|^2}{N - \sum_i \gamma_i}$$

پارامتر N در مخرج عبارت فوق تعداد نمونه ها رانشان می دهد و نه تعداد توابع پایه.

پس از همگرا شدن مرحله تخمین هایپرپارامترها، پیش بینی بر اساس توزیع پسین روی وزن ها انجام می شود. به این ترتیب برای داده جدید x^* توزیع به صورت زیر به دست می آید:

$$p(t^* \mid \mathbf{t}, \boldsymbol{\alpha}_{MP}, \sigma^2_{MP}) = \int p(t^* \mid \mathbf{w}, \sigma^2_{MP})\, p(\mathbf{w} \mid \mathbf{t}, \boldsymbol{\alpha}_{MP}, \sigma^2_{MP})\, d\mathbf{w}$$

از آنجایی که هر دو عبارت داخل انتگرال گوسی هستند، داریم:

$$p(t^* \mid \mathbf{t}, \boldsymbol{\alpha}_{MP}, \sigma^2_{MP}) = N(t^* \mid y^*, (\sigma^2)^*)$$

که در آن:

$$y^* = \mu^T \boldsymbol{\varphi}(x^*)$$

$$(\sigma^2)^* = \sigma^2_{MP} + \boldsymbol{\varphi}(x^*)^T \sum \boldsymbol{\varphi}(x^*)$$

بنابراین میانگین خروجی پیش بینی شده برای داده x^* از یعنی توابع پایه ای که با میانگین پسین وزن دار شده است، به دست می آید. توجه شود که بیشتر المان های صفر هستند. واریانس پیش بینی شده در رابطه فوق شامل دو بخش است. بخش اول تخمین نویز در داده است و بخش دوم ناشی از ابهام در پیش بینی وزن ها است.

آموزش بیزین تنک برای طبقه بندی نیز همانند آموزش این روش برای مسایل رگرسیون می باشد.

مراجع

[1] Ciresan, Dan, Ueli Meier, and Jürgen Schmidhuber. "Multi-column deep neural networks for image classification." Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012

[2] Poultney, Christopher, Sumit Chopra, and Yann L. Cun. "Efficient learning of sparse representations with an energy-based model." Advances in neural information processing systems. 2006.

[3] Hanchuan Peng et al.," Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy" IEEE Trans. Pattern Anal. Mach. Intell, vol. 27, NO.8, pp. 1126-1238, 2005

[4] Bishop CM, Pattern Recognition and Machine Learning, Springer, USA, 2006

[5] CORINNA CORTES, VLADIMIR VAPNIK," Support-Vector Networks" Machine Learning, pp. 273-297, 1995

[6] A. Bellili et al.,"An MLP-SVM combination architecture for offline handwritten digit recognition "international journal on document and recognition(IJDAR),pp. 244-252, 2003

[7] "Optical Recognition of Handwritten Digits Data Set". Internet: https://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits , [Aug.7, 2015]

[8] E.K. Tang et al., "Linear dimensionality reduction using relevance weighted LDA" Pattern Recognition, pp. 485-493, 2005

[9]M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine", Journal of Machine Learning Res. 1, 211–244, 2001.

V. N. Vapnik, "Statistical Learning Theory", Wiley. 1998.

[9] S. K. Majumder, N. Ghosh, P. K. Gupta, "Relevance vector machine for optical diagnosis of cancer", Lasers Surg. Med. 36 (4), 323–333, 2005.

[10] C.M.Bishop, M. E. Tipping, "Variational relevance vector machine", In Proc. 16th Conf. on Uncertainty in Artificial Intelligence. Morgan Kaufman Publishers, 2000.

[11] D. J. C. Mackay, "Bayesian Interpolation", Neural Computation, 4(3):415-447, 1992a.

[12] D. J. C. Mackay, "The evidence framework applied to classification networks", Neural Computation, 4(5):720-736, 1992b.