یادگیری آنلاین الگوهای مثبت و منفی به همراه توضیح‌پذیری مبتنی بر بسط هسته

حسین اسمعیلی[۱] **سید کمال‌الدین غیاثی شیرازی**[۲] **احد هراتی**[۳]

[۱] دانشجوی کارشناسی ارشد - دانشکده مهندسی کامپیوتر- دانشگاه فردوسی مشهد- مشهد- ایران
hossein.esmaeli@mail.um.ac.ir

[۲] استادیار- دانشکده مهندسی کامپیوتر- دانشگاه فردوسی مشهد- مشهد- ایران
k.ghiasi@um.ac.ir

[۲] دانشیار- دانشکده مهندسی کامپیوتر- دانشگاه فردوسی مشهد- مشهد- ایران
a.harati@um.ac.ir

چکیده: **مسئله طبقه‌بندی همچنان جزو مسائل مورد بحث خیلی از مقالات روز می‌باشد. اغلب مدل‌های ارائه شده در مقالات، از عدم توضیح دلیلی قابل درک برای انسان رنج می‌برند. یکی از روش‌های ایجاد توضیح‌پذیری، تفکیک وزن‌های شبکه به دو بخش مثبت و منفی مبتنی بر الگو می‌باشد. بخش مثبت نمایانگر وزن‌های مربوط به کلاس درست و بخش منفی نمایانگر وزن‌هایی که به اشتباه به کلاس مذکور نسبت داده شده‌اند. به این شبکه، شبکه‌ی WTA مبتنی بر فاصله اقلیدسی مثبت و منفی یا ED-WTA± گفته می‌شود. در این مقاله با استفاده از بسط هسته علاوه بر دست‌یابی به توضیح‌پذیری محلی، دقت بالاتری به نسبت مقاله‌ی موجود به واسطه‌ی مدل سازی غیرخطی کسب شده است. روش‌هایی در این مقاله به منظور بهبود فضای زمانی و فضای الگوریتم ارائه خواهد شد. همچنین از روش نیستروم برای تقریب هسته به منظور مقیاس پذیر شدن الگوریتم در برابر مجموعه‌دادگان حجیم استفاده شده است. با استفاده از این شبکه تک‌لایه در مجموعه دادگان MNIST دقت ۹۸,۰۱٪ بر روی دادگان آزمون کسب شده است و با استفاده از بسط هسته دلایل استدلال را نیز به خوبی با دادگان ورودی شرح می‌دهد. همچنین توضیح‌پذیری بر روی مجموعه دادگان FERET دو کلاسه بررسی شده است.**

واژه‌های کلیدی: **روش‌های هسته، توضیح‌پذیری، یادگیری مبتنی بر الگو**

نوع مقاله: **پژوهشی**

DOI: 10.52547/jiaeee.20.1.67

تاریخ ارسال مقاله: ۱۴۰۰/۱۱/۲۰
تاریخ پذیرش مشروط مقاله:۱۴۰۱/۰۲/۱۷
تاریخ پذیرش مقاله: ۱۴۰۱/۰۵/۰۳
نام نویسنده‌ی مسئول: دکتر سید کمال‌الدین غیاثی شیرازی
نشانی نویسنده‌ی مسئول : ایران – مشهد – بلوار وکیل آباد – بلوار باهنر – دانشگاه فردوسی – دانشکده‌ی کامپیوتر

۱- مقدمه

به طور معمول، مدل‌های مورد استفاده در شبکه‌های عصبی، یادگیری ماشین، بینایی کامپیوتر و ... به مدل‌های جعبه‌سیاه[۱] معروف هستند. در حالت عادی پردازش داخل مدل‌های جعبه‌سیاه برای انسان‌ها قابل توضیح نیست ولی انسان‌ها تمایل دارند که مدل‌های توضیح‌پذیر داشته باشند. [۱] هر اطلاعات واضحی که داده شود، یک توضیح[۲] می‌باشد. زمانی به توضیح احتیاج پیدا می‌شود که تصمیم گرفته شده توسط شبکه برای انسان قابل درک نباشد. یک مدل توضیح‌پذیر می‌باشد، هرگاه برای تصمیمات گرفته شده توسط مدل در مسئله، دلیلی ارائه دهد. یک توضیح خوب، توضیحی می‌باشد که برای مسائل مختلف منعطف و کلی باشد و بتواند از دانش قبلی ارائه شده توسط طراح استفاده کند. [۲] مدل‌های توضیح‌پذیر در زمینه‌های مختلفی مثل تشخیص شی [۳]، طبقه‌بندی [۴]، رگرسیون [۵] و... با هدف‌گرفتن حوزه‌های متفاوتی از جمله پزشکی [۶]، ماشین‌های خودران [۷] و... استفاده خواهد شد. الگوریتم یادگیری می‌تواند باناظر، بی‌ناظر و یا تقویتی باشد. زمینه مورد بحث این مقاله، طبقه‌بندی با ناظر خواهد بود که با استفاده از توابع هسته، توضیح‌پذیری را بررسی خواهد کرد.

در بخش ۱-۱ به مرور تاریخچه گذشته پرداخته می‌شود، که مروری بر توضیح‌پذیری و انواع روش‌های موجود خواهد بود. در بخش ۱-۲ به بیان بسط هسته و تعاریف آن پرداخته خواهد شد. بخش ۱-۳ به بیان مقاله WTA مبتنی بر فاصله اقلیدسی مثبت و منفی[۳] اختصاص یافته و مبانی آن بیان می‌شود. در بخش ۲ به بازنویسی روابط ذکر شده بر اساس توابع هسته پرداخته شده و ساختار آن تعریف می‌شود. در بخش ۳ در مورد بهبودهایی از جمله بهبود زمانی و بهبود فضایی بحث می‌شود. در بخش ۴ آزمایش‌های انجام شده مطرح می‌شود و به بررسی نتایج بدست آمده از لحاظ توضیح‌پذیری و دقت بر روی داده‌های آزمون پرداخته خواهد شد.

۱-۱- تاریخچه

یکی از روش‌های ساخت مدل توضیح‌پذیر، نمایش نتیجه‌ای تفسیرپذیر توسط مدل برای انسان است. در این مقاله هم‌زمان از کلمه‌های توضیح‌پذیری و تفسیرپذیری استفاده خواهد شد و با توجه به توضیحات ارائه شده در کتاب [۲] بین تفسیرپذیری و توضیح‌پذیری مرز مشخصی وجود ندارد.

از دلایل اهمیت تفسیرپذیری، می‌توان به موارد زیر اشاره کرد:

- یافتن دلیل تصمیم گرفته شده توسط عامل هوش مصنوعی در دنیای واقعی
- افزایش اعتمادپذیری سیستم
- تشخیص میزان صحت
- افزودن تعاملات اجتماعی بین انسان و ماشین.

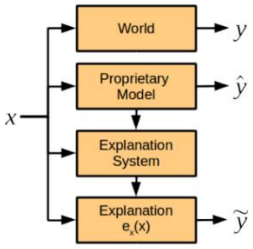
از طرفی، دلایلی وجود دارد که نباید از سیستم‌های تفسیرپذیر استفاده کرد:

- شناسایی مسائل به خوبی انجام شده‌اند و نیازی به مدل تفسیرپذیر نیست
- مدل جعبه‌سیاه به خوبی پاسخگو نباشد
- افزودن تفسیرپذیری، طراح را مجبور به تغییر بیش از اندازه مدل کند.
- باعث کاهش دقت شود.

زمانی از مدل‌های توضیح‌پذیر استفاده می‌شود که در مقابل هزینه‌ی تولید آنها معقول باشد. همچنین می‌تواند باعث تاثیرات منفی از جمله کاهش دقت در مدل شود. با توجه به مسئله می‌تواند این کاهش دقت معقول باشد و برای انسان ارائه توضیح، اهمیت بیشتری داشته باشد و یا غیر معقول باشد و باعث ضررهای جبران ناپذیری شود. [۸,۲]

مدل‌های توضیح‌پذیر به دو دسته داخلی و خارجی تقسیم می‌شوند که هر کدام جداگانه توضیح داده می‌شوند.

بسیاری معتقد هستند که مدل توضیح‌پذیر باید جدا از مدل جعبه‌سیاه طراحی شود. به عنوان مثال در [۹,۸] ، سعی دارند که مدلی کاملا مجزا از مدل جعبه‌سیاه آموزش دهند که بسیار شبیه به مدل جعبه‌سیاه اولیه عمل کند با این تفاوت که خروجی کاملا توضیح‌پذیر برای کاربر دهد. به این‌گونه مدل‌ها، مدل توضیح‌پذیری خارجی می‌گویند.



شکل (۱): استفاده از یک سیستم قابل توضیح در کنار سیستم جعبه‌سیاه

این مدل‌ها که ساختار کلی شکل (۱) را دارند، به شبکه‌های معلم– دانش آموزی[۴] معروف هستند و بیشتر در زمینه فرا یادگیری[۵] استفاده می‌شوند. از نمونه‌های این نوع، استفاده از مدلی جداگانه برای یادگیری و تولید مثال بر اساس ویژگی‌های نهفته در نمونه و تشکیل درخت تصمیم و نقشه برجستگی برای نمایش دلیل تصمیم خود می‌باشد. [۱۰] مثال دیگری از این نوع، استفاده از K نزدیک ترین همسایه برای هر لایه از یک شبکه آموزش دیده شده است که بازنمایی هر لایه با داده‌های آموزشی سنجیده می‌شوند. [۱۱]

بسیاری از افراد نیز معتقد هستند که در داخل شبکه، می‌توان مدل‌های تفسیرپذیر و توضیح‌پذیر ایجاد کرد. از این‌گونه مدل‌ها انواع مختلفی وجود دارد و تنوع در این شاخه بالا می‌باشد.

یکی از روش‌ها، ارائه مدلی مبتنی اجزاء[۶] می‌باشد. [۱۲] در این روش با پیدا کردن چندین جزء با بالاترین احتمال که توسط شبکه‌های پیچشی[۷] استخراج می‌شود، سعی می‌شود طبقه‌بندی انجام گردد. در نهایت با تطبیق تمام اجزاء شناسایی شده با داده ورودی، کلاس نهایی تعیین می‌شود. با توجه به اینکه جزءها باعث ایجاد تمایز بین کلاس‌های مختلف می‌شوند بنابراین توضیح قابل استنادی برای انسان وجود خواهد داشت.

روش‌های مبتنی بر الگو[۸] از پرکاربردترین روش ها برای ارائه مدل‌های تفسیرپذیر و توضیح‌پذیر هستند. یکی از آنها الگوهای مثبت و منفی می‌باشد. [۱۳] در این مقاله، با در نظر گرفتن تعدادی الگو برای هرکلاس، سعی می‌شود الگوهای مثبت را به گونه‌ای ایجاد نماید که نمایانگر قسمت‌هایی از داده اصلی باشند و به درستی تشخیص داده شوند. همچنین الگوهای منفی دارای قسمتی از داده‌هایی که متعلق به کلاس ذکر شده نبوده و به اشتباه در کلاس مذکور تشخیص داده شده‌اند، می‌باشند. مقاله ذکر شده، ایده‌ای شبیه به مقاله [۱۴] خواهد داشت که طبقه‌بند عدم شباهت خواهد داشت. در شبکه WTA با الگوهای مثبت و منفی، خروجی یک نورون، توسط محاسبه میزان فاصله هر الگو با ورودی و سپس تفریق آنها از هم به‌دست می‌آید. توضیحات بیشتر در بخش ۱-۳ داده خواهد شد. یکی دیگر از روش‌های مبتنی برالگو، استفاده از الگوهای حاصل از انتقال داده به فضای ویژگی می‌باشد که با استفاده از روش خودکدگذار[۹] [۷]، روش تقریب فوریه [۱۵] و ... انجام می‌شوند.

دیدگاه مشابه [۱۳] وجود دارد که مشتق بازگردانده شده در شبکه‌های عمیق را به دو بخش مثبت و منفی می‌شکند و هر نورون، با مقدار مرجع آن (که تجربی به دست می‌آید) سنجیده می‌شود و با قواعد زنجیره‌ای، مشتق را به ورودی برمی‌گرداند.[۱۶]

همچنین در این زمینه روش‌هایی ترکیبی مبتنی بر توجه[۱۰] و الگو وجود دارد. در [۱۸,۱۷] با کمک شبکه‌های عصبی پیچشی ویژگی‌های نهفته در تصویر را استخراج کرده و به لایه الگوها می‌دهد. با به‌کارگیری الگوهای مختلف، نواحی قابل توجه تصویر را که بیشترین شباهت با الگو را دارند، انتخاب کرده و به آنها امتیاز می‌دهد. لازم به ذکر است طول آموزش قسمتی از تصویرهای دادگان آموزشی به عنوان الگو انتخاب می‌شوند. بنابراین، برای سنجش میزان شباهت، لازم است قسمت‌هایی از تصویر ورودی با الگوها مقایسه شوند که از لحاظ اندازه و ابعاد یکسان هستند. برای این منظور از روش مطرح شده در مقاله [۱۹] استفاده شده است که با استفاده از توابع مبتنی بر سنجش فاصله، میزان شباهت قسمتی از تصویر با الگوی یاد گرفته شده را برای استفاده در لایه الگو می‌دهد. در ادامه ی این کار، مقاله [۶] به لایه الگوها استدلال منفی را افزود. با در نظر گرفتن ماتریس وزن جدید و ضرب در بردار شباهت، با توجه به کلاس صحیح آن الگو، کلاس پیش‌بینی شده را تعیین می‌کند.

از دیگر مقالات مطرح شده توضیح‌پذیری ترکیبی مبتنی بر توجه و الگو، مقاله [۲۰] می‌باشد که بر خلاف مقاله [۱۹] که قسمتی از تصاویر آموزشی را به عنوان الگو انتخاب می‌کرد، کل نمونه را به عنوان الگو انتخاب می‌کند. در این مقاله با پیدا کردن تعداد کمی الگو که توسط کدگذار[۱۱] به فضای ویژگی برده شده‌اند، سعی می‌کند با ترکیبی خطی از الگوهای یافت شده، داده ورودی را استنتاج نماید.

یکی دیگر از روش‌های موجود در زمینه توجه، استفاده از نگاشت فعال سازی کلاس با گرادیان یا Grad-CAM می‌باشد. [۲۱] این الگوریتم با استفاده از گرادیان و دانستن هدف، سیگنالی را به ورودی بر می‌گرداند و مکان قرارگیری هدف را مشخص می‌کند.

روش دیگری مبتنی بر مدل‌های افزودنی[۱۲] به نام مدل‌های افزودنی نورونی یا NAM وجود دارد که با داشتن ورودی x با K ویژگی، به ازای هر ویژگی ورودی، شبکه‌عصبی چند لایه‌ای وجود دارد که هر ویژگی را جداگانه یاد می‌گیرد. [۲۲] تفسیرپذیری روی ویژگی‌ها به دلیل آموزش جداگانه‌ی شبکه‌ها که مستقل از دیگری عمل می‌کنند، با نمایش خروجی هر کدام از شبکه‌ها و تاثیر آنها در پیش‌بینی نهایی، مشخص می‌شود. مشکل این مدل افزایش فضای زمانی لازم برای آموزش این شبکه می‌باشد.

همچنین روش‌هایی وجود دارد که با توجه به حوزه مورد بحث، روشی برای توضیح‌پذیری ارائه می‌دهد. به عنوان مثال در حوزه سری‌های زمانی، یکی از روش‌هایی که توضیح‌پذیری را ایجاد می‌کند استفاده از توابع پایه می‌باشد. در مقاله [۲۳] با استفاده از توابع پایه نورونی و تنها بر اساس دانش یادگیری ماشینی، عملیات پیش‌بینی سری زمانی را به عهده می‌گیرد. یکی دیگر از این روش‌ها در حوزه سری‌های زمانی، استفاده از شیپلت‌ها[۱۳] است. شیپلت‌ها، زیر دنباله‌ای از سری زمانی هستند که نماینده یک کلاس بوده و بیشترین تمایز را بین کلاس خود و کلاس‌های دیگر ایجاد می‌کنند.[۷, ۲۴].

۱-۲- توابع هسته

گاهی مدلی که برای دادگان در نظر گرفته می‌شود با پیچیدگی آنها تناسب ندارند. به عنوان مثال دادگانی که به صورت خطی جداپذیر نباشند ولی مدل به صورت خطی کار کند. یکی از روش‌های موجود برای این منظور، استفاده از توابع هسته است. توابع هسته داده را از فضای ورودی به فضای ویژگی می‌برد و پردازش را با همان مدل در نظر گرفته شده انجام می‌دهد. به فضاهای ضرب داخلی کامل، فضای هیلبرت[۱۴] گفته می‌شود. فضای ویژگی ذکر شده نیز یک فضای هیلبرت خواهد بود.

به k یک تابع هسته می‌گویند، هرگاه متناظر با آن فضای هیلبرت H و نگاشت Φ وجود داشته باشد که داده را از فضای ورودی به فضای ویژگی نگاشت کند، آنگاه رابطه‌ی تابع هسته به صورت رابطه‌ی (۱) قابل بیان است.

$$k(x,y) = \langle \Phi(x), \Phi(y) \rangle_H \qquad (۱)$$

به عنوان تابع هسته، می‌توان از توابعی مبتنی بر ضرب داخلی مثل چندجمله‌ای‌ها و مبتنی بر فاصله مثل گوسی نام برد. دقت کنید که به تابع هسته گوسی، تابع پایه شعاعی[۱۵] نیز گفته می‌شود. در جدول (۱) نمونه‌هایی از تابع هسته آورده شده است.

جدول (۱): نمونه‌هایی از تابع هسته

نام تابع هسته	رابطه
خطی	$K(x,z) = \langle x, z \rangle$
چندجمله‌ای	$K(x,z) = \langle x, z \rangle^P$
چندجمله‌ای اصلاح شده [24]	$K(x,z) = \left(\frac{\langle 2x-1, 2z-1\rangle + 1}{2}\right)^P$
گوسی	$K(x,z) = \exp\left(-\frac{1}{2\sigma^2}\|x-z\|^2\right)$

در حالت کلی دو نوع نگاشت صریح[۱۶] و ضمنی[۱۷] در توابع هسته وجود دارد. طبق حقه هسته[۱۸]، با فرض بیان شدن تمام محاسبات هندسی مسئله بر حسب ضرب داخلی، می‌توان به جای ضرب داخلی، محاسبه تابع هسته را قرار داد. [۲۵] در این حالت از نگاشت ضمنی استفاده شده است و اگر به جای نگاشت ضرب داخلی به فضای ویژگی، خود داده مستقیما به فضای ویژگی نگاشت شود، از نگاشت صریح استفاده شده است. [۳۳]

مقاله [۲۶] توسط توابع هسته و استفاده از یادگیری فیلترهای گابور[۱۹] تنک، توانسته تقریب مناسبی از تصویر ورودی ارائه دهد و همچنین تفسیرپذیری را با توجه به کم بودن تعداد فیلترهای تعیین کننده، اضافه نماید.

مشکلی که در توابع هسته وجود دارد، عدم مقیاس پذیری بالا می‌باشد. به این معنا که با افزایش تعداد داده ورودی، محاسبات حاصل از توابع هسته به شدت افزایش می‌یابد. ضمن اینکه نگهداری بسط هسته در حافظه نیازمند فضای بسیار زیادی خواهد بود. راه‌های مختلفی از جمله تقریب زدن تابع هسته [۲۷] و استفاده از روش‌های بودجه ثابت [۲۸] برای رفع مشکلات ذکر شده، استفاده شده است. همچنین صرف جایگذاری تابع هسته، لزوم افزایش دقت نیست. تابع هسته باید بر اساس مسئله انتخاب شده تا هم مشکل عدم تناسب پیچیدگی توابع هسته با مسئله و هم مشکل بیش‌برازش رخ ندهند. [۲۹]

۱-۳- شبکه‌ی WTA مبتنی بر فاصله اقلیدسی مثبت و منفی

در شبکه‌های WTA، به ازای هر کلاس مسئله، حداقل یک نورون در نظر گرفته می‌شود و با اعمال حاصلضرب وزن‌های یاد گرفته شده در ورودی، نورونی با بیشترین مقدار به عنوان برنده انتخاب می‌شود. با فرض اینکه تعداد نورون‌ها M، تعداد کلاس‌های مسئله K و به ازای هر $k \in \{1,2,\ldots,K\}$ مجموعه O_k نورون‌های در نظر گرفته شده برای کلاس k باشند، خروجی y به صورت زیر است:

$$y = \underset{k \in \{1,2,\ldots,K\}}{argmax} \quad \underset{j \in O_k}{max}\, w_j^T x + b_j \qquad (۲)$$

در این مقاله برای تبدیل خروجی شبکه به توزیع احتمالاتی، از تابع فعال ساز سافت‌مکس[۲۰] استفاده شده که برای هر نورون j توزیع احتمال y_j به صورت زیر است:

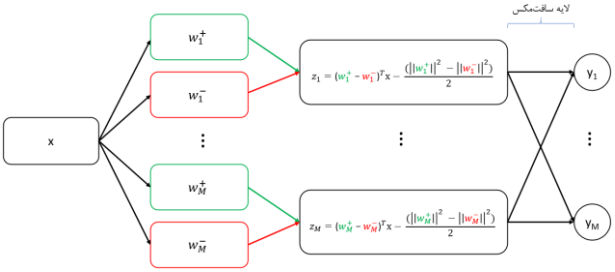
$$y_j = \frac{e^{\beta z_j}}{\sum_{i \in O_k}^{M} e^{\beta z_i}} \qquad (۳)$$

که β یک ضریب قابل یادگیر در طول آموزش است. همچنین با استفاده از تابع زیان آنتروپی متقابل رقابتی[۲۱]، توزیع احتمال هدف برای هر نورون که به کلاس k تعلق دارد به صورت زیر می‌باشد:

$$\tau_j = \begin{cases} \frac{e^{\beta z_j}}{\sum_{i \in O_k} e^{\beta z_i}} & j \in O_k \\ 0 & otherwise \end{cases} \qquad (۴)$$

حال با داشتن توزیع احتمال هدف τ_j تابع زیان آنتروپی متقابل رقابتی برای یک داده آموزشی به صورت زیر تعریف می‌شود:

$$E^{CCE} = -\sum_{j=1}^{M} \tau_j ln(y_j) \qquad (۵)$$



شکل (۲): ساختار مدل شده اصلی [۱۳]

همانطور که در شکل (۲) نشان داده شده است، خروجی هر نورون به صورت زیر قابل محاسبه است که معادل سازی شده با روش ضرب داخلی می‌باشد:

$$z_j = (w_j^+ - w_j^-)^T x_n - \frac{\left(\|w_j^+\|^2 - \|w_j^-\|^2\right)}{2}, \quad j \in \{1,2,\ldots,M\} \qquad (۶)$$

همچنین روابط به روز رسانی وزن ها به صورت زیر می‌باشد:

$$w_j^+ \leftarrow w_j^+ + \mu\beta(\tau_j - y_j)(x_n - w_j^+), \forall j \in O_k \qquad (۷)$$

$$w_j^- \leftarrow w_j^- + \mu\beta(y_j)(x_n - w_j^-), \forall j \notin O_k \qquad (۸)$$

$$\beta \leftarrow \beta + \gamma \sum_{j=1}^{M} (\tau_j - y_j) z_j \qquad (۹)$$

که x_n داده n ام در طول آموزش و μ و γ نرخ یادگیری برای روابط ذکر شده می‌باشند.

۲- نسخه هسته‌ای شبکه ED-WTA±

در مقاله [13] که الگوها، همان وزن‌های شبکه می‌باشند با تفکیک وزن به دو بخش مثبت و منفی، عملکرد نورون را شرح می‌دهد. در این مقاله با باز کردن رابطه‌ی تعریف شده، به رابطه‌ی (۱۰) دست یافته است. لازم به ذکر است $\alpha \in \mathbb{R}^{M\times N}$ ماتریسی از ضرایب قابل یادگیری که M تعداد کل نورون ها و N تعداد کل داده ها هستند.

$$w_j = w_j^+ - w_j^- = \sum_{i=1}^{N} \alpha_j^{i^+} x_i - \sum_{i=1}^{N} \alpha_j^{i^-} x_i \qquad (۱۰)$$

در این مقاله به منظور توضیح‌پذیری بیشتر یک قدم رو به جلو حرکت شده است و بعد از اعمال داده بر روی وزن‌ها، قصد پیدا کردن الگو را دارد. برای این منظور با ضرب داخلی طرفین تساوی رابطه‌ی (۱۰) در داده مشخص x_n رابطه‌ی (۱۱) حاصل می‌شود.

$$\langle w_j , x_n \rangle = \langle \sum_{i=1}^{N} \alpha_j^{i^+} x_i - \sum_{i=1}^{N} \alpha_j^{i^-} x_i \ , x_n \rangle \qquad (۱۱)$$

که $\langle .,. \rangle$ نمایانگر ضرب داخلی می‌باشد. حال با بسط داده x_n در سمت راست رابطه‌ی (۱۱)، ضرب داخلی بین x_n و هر داده ورودی x_i بوجود خواهد آمد. با توجه به حقه هسته که می‌توان به جای هر ضرب داخلی، تابع هسته مشخص شده‌ای را قرار داد، با جایگذاری تابع هسته k رابطه‌ی (۱۲) بدست خواهد آمد.

$$\begin{aligned} &\langle w_j , x_n \rangle \\ &= \sum_{i=1}^{N} \alpha_j^{i^+} k(x_i , x_n) - \sum_{i=1}^{N} \alpha_j^{i^-} k(x_i , x_n) \\ &= \alpha_j^{+^T} k(x_n,.) - \alpha_j^{-^T} k(x_n,.) \end{aligned} \qquad (۱۲)$$

که $k(x_n,.)$ به معنای محاسبه تابع هسته بین داده x_n و تمام داده‌های مورد بحث خواهد بود که در مقاله، آن را با k_{x_n} نمایش می‌دهد. با بسط رابطه‌ی (۶) و اعمال رابطه‌ی ذکر شده در آن خواهیم داشت:

$$\begin{aligned} Z_j &= (w_j^+ - w_j^-)^T x - \frac{\left(\left\|w_j^+\right\|^2 - \left\|w_j^-\right\|^2\right)}{2} \\ &= w_j^{+^T} x - w_j^{-^T} x - \frac{\left(\left\|w_j^+\right\|^2 - \left\|w_j^-\right\|^2\right)}{2} \\ &= \alpha_j^{+^T} k_{x_n} - \alpha_j^{-^T} k_{x_n} - \\ &\quad \frac{\left(\left\|\alpha_j^{+^T} K\, \alpha_j^+\right\|^2 - \left\|\alpha_j^{-^T} K\, \alpha_j^-\right\|^2\right)}{2} \end{aligned} \qquad (۱۳)$$

با به دست آمدن رابطه‌ی جدید برای خروجی نورون‌ها، نیاز هست که قانون به روز رسانی هم تغییر کند. با توجه به اینکه در نسخه هسته الگوریتم ضرایب بسط دادگان آموزشی یاد گرفته می‌شوند، بنابراین قانون به روز رسانی نیز باید ضرایب بسط را تغییر دهد. با باز کردن w در رابطه‌ی (۷) و (۸) بر اساس رابطه (۱۲)، روابط (۱۴) و (۱۵) بدست می‌آید.

$$\alpha_j^{+^T} X \leftarrow \alpha_j^{+^T} X + \mu\beta(\tau_j - y_j)\left(x_n - \alpha_j^{+^T} X\right) \qquad (۱۴)$$

$$\alpha_j^{-^T} X \leftarrow \alpha_j^{-^T} X + \mu\beta(y_j)\left(x_n - \alpha_j^{-^T} X\right) \qquad (۱۵)$$

با ساده سازی روابط ذکر شده به روابط (۱۶) و (۱۷) خواهیم رسید. در این رابطه، p_n برداری با N درایه که در درایه n ام مقدار یک و در باقی درایه ها صفر خواهد بود و نمایانگر داده آموزشی مورد بررسی در آن دوره آموزش است. در این مقاله به این روش Kernel ED-WTA± گفته می‌شود.

$$\alpha_j^+ \leftarrow \alpha_j^+ + \mu\beta(\tau_j - y_j)(p_n - \alpha_j^+) \qquad (۱۶)$$

$$\alpha_j^- \leftarrow \alpha_j^- + \mu\beta(y_j)(p_n - \alpha_j^-) \qquad (۱۷)$$

دیدگاه مبتنی بر الگو در مقاله [۱۳] با دیدگاه پیشنهاد شده تفاوت خواهد داشت. الگوهای مثبت و منفی حاصل از تجربه و سپس تفکیک روی وزن شبکه حاصل می‌شود؛ اما در مدل پیشنهاد شده این مقاله، صحبت روی حاصل‌ضرب داده و وزن خواهد بود و تفکیک در این قسمت صورت خواهد گرفت. بنابراین الگوهای قابل یادگیری، مثال‌هایی از دادگان خواهند بود. به عبارت دیگر در این قسمت، توضیح‌پذیری با استفاده از بسط داده آموزشی و بر اساس توابع هسته نتیجه می‌شود و برخلاف مقاله [۱۳] میانگین تجربه‌های گذشته به خاطر سپرده نشده؛ بلکه خود دادگان موثر در تصمیم به خاطر سپرده شده است که به آن توضیح‌پذیری مبتنی بر نمونه گفته می‌شود. همچنین برای هر بخش مثبت و منفی چند الگو در نظر گرفته می‌شود که در انتهای آموزش، هر کدام از این الگوها دارای بسط متفاوتی خواهند شد.

۳- بهبودهای ایجاد شده

با توجه به اینکه روش‌های هسته از عدم مقیاس پذیری زمانی و فضایی رنج می‌برند و در حالت عادی برای مجموعه دادگان بزرگ مناسب نیستند، باید روشی داشته باشیم تا هم از لحاظ زمانی و هم از لحاظ فضایی معقول باشند. برای این منظور از روش تقریب نیستروم[۲۲] استفاده می‌شود که هم از لحاظ فضایی و هم از لحاظ زمانی معقول می‌باشد ودقت را حفظ می‌کند. در بخش ۳-۱ در مورد آن صحبت شده است. همچنین روشی برای محاسبه نرم اقلیدسی[۲۳] برای بهبود فضای زمانی در بخش ۳-۲ شرح داده می‌شود. برای بهبود دقت و توضیح‌پذیری، روشی ارائه خواهد شد تا با افزایش دادگان انتخابی، دقت بالا رفته و از توضیح‌پذیری بهتری بهره برده شود. این توضیحات نیز در بخش ۳-۳ ارائه شده است.

۳-۱- تقریب نیستروم

یکی از روش‌های خوب تقریب توابع هسته، تقریب نیستروم می‌باشد. [۲۷] با فرض انتخاب B داده از مجموعه N داده آموزشی که $B << N$ می‌باشد، می‌توان ماتریس هسته اصلی $\mathcal{K}$ را به صورت زیر تقریب زد:

$$\mathcal{K} \approx \widehat{\mathcal{K}}_{N,B}\, \widehat{\mathcal{K}}_{B,B}^{-1}\, \widehat{\mathcal{K}}^T_{B,N} \qquad (۱۸)$$

با این روش، می‌توان بسط α را به جای اینکه بر روی کل دادگان آموزشی تعریف شود، بر روی دادگان انتخاب شده برای تقریب

نیستروم تعریف کرد. این دادگان انتخاب شده، به عنوان الگو برای شبکه به حساب آورده می‌شود و نشان دهنده زیر مجموعه ای دادگان اصلی است که دیدی از مجموعه داده به همراه توضیح‌پذیری بالا می‌دهد.[30] با این کار $\alpha \in \mathbb{R}^{M\times N}$ به $\alpha \in \mathbb{R}^{M\times B}$ تبدیل خواهد شد که حافظه بسیار کمتری مصرف خواهد کرد و به دلیل کمتر شدن تعداد ضرب‌های موجود در طول آموزش، باعث سریعتر شدن روند آموزش نیز خواهد شد. این عمل باعث تغییر پیچیدگی ضرب ماتریس در رابطه‌ی (۱۲) از O(MN) به O(MB) می‌شود. اندیس مجموعه دادگان انتخاب شده برای تقریب در $L = \{l_1,\dots, l_B\}$ ذخیره خواهد شد.

با این حال، محاسبه p_n دچار تغییر خواهد شد. با توجه به اینکه تعداد دادگان کاهش یافته و بسط دادگان قابل یادگیر دستخوش تغییر شده است، بنابراین داده جدید ممکن است در مجموعه دادگان مورد انتخابی نیستروم حاضر نباشد. بنابراین لازم است تا داده با مجموعه دادگان نیستروم تقریب زده شود. اما تقریب همیشه کامل نیست و دارای خطا می‌باشد. به همین دلیل تابعی به منظور سنجش خطای تقریب $K(x_t, .)$ تعریف می‌شود. اگر فرض شود Pr_B عملگر پروجکشن[۲۴] به زیرفضایی باشد که $K(x_{l_1}, .), \dots, K(x_{l_B}, .)$ بپیماید، رابطه (۱۹) حاصل می‌شود [31] [32].

$$\epsilon_t = \|K(x_t, .) - Pr_B K(x_t, .)\|^2_{\mathcal{H}} \quad (۱۹)$$

اگر فرض شود حاصل پروجکت شده‌ی $K(x_t, .)$ به زیرفضا برابر $\sum_{k=1}^{B} p_k K(x_{l_k}, .)$ باشد، بنابراین با هدف یافتن کمترین خطای تقریب، باید مقادیر بهینه‌ی p را یافت.

$$\epsilon_t = \min_{p} \left\| K(x_t, .) - \sum_{k=1}^{B} p_k K(x_{l_k}, .) \right\|^2_{\mathcal{H}} \quad (۲۰)$$

که با باز کردن رابطه‌ی تابع خطای (۲۰) و گرفتن مشتق نسبت به p و برابر صفر قرار دادن آن، به رابطه‌ی زیر خواهیم رسید:

$$p = \mathcal{K}_{B\times B}^{-1} [K(x_{l_1}, x_n), \dots, K(x_{l_B}, x_n)]^T \quad (۲۱)$$

حال از آنجایی که p نمایانگر وزن‌های بهینه در زیرفضای پیمایش شده است و لازم است که داده‌ی آموزشی بر اساس کلاس خود تقریب زده شود، باید تغییری در رابطه‌ی (۲۱) ایجاد شود. اگر فرض شود از B داده انتخاب شده نیستروم، مجموعه دادگان انتخاب شده برای تقریب هرکلاس با $S_1, S_2, \dots, S_C$ و تعداد داده موجود در هر مجموعه با $Size(S_i)$ نمایش داده شود، آنگاه رابطه عمومی (۲۲) وجود دارد.

$$p = \mathcal{K}_{Size(S_i)\times Size(S_i)}^{-1} [K(x_{S_i^1}, x_n), \dots, K(x_{S_i^{Size(S_i)}}, x_n)]^T \quad (۲۲)$$

پس در این مقاله، با به کارگیری روش نیستروم، از رابطه‌ی (۲۲) برای تقریب داده آموزشی ورودی، توسط مجموعه دادگان انتخاب شده نیستروم استفاده خواهد کرد. به این روش در این مقاله، KerNys ±ED-WTA گفته می‌شود.

انتخاب زیرمجموعه دادگان تقریب نیستروم از دادگان آموزشی موجود، با روش‌های متفاوتی انجام می‌گیرد. در این مقاله برای انتخاب دادگان نیستروم، از خوشه بندی K-Means استفاده می‌شود. برای این کار بر روی مجموعه دادگان هر کلاس خوشه بندی انجام می‌شود. تعداد خوشه‌های انتخابی در ابتدای خوشه بندی تعیین شده و پس از پیدا شدن خوشه‌ها، نزدیک ترین داده متمایز به خوشه را به عنوان نماینده خوشه، انتخاب می‌کند. در نهایت دادگان انتخاب شده هر کلاس را کنار هم گذاشته و مجموعه دادگان انتخابی نیستروم را تشکیل می‌دهد.

این بهبود، باعث کاهش پیچیدگی عمل ضرب موجود در رابطه (۱۲) از O(MN) به O(MB) خواهد شد.

۳-۲- محاسبه تکرارشونده نرم اقلیدسی

با توجه به اینکه در رابطه‌ی (۱۳) نسخه هسته الگوریتم بدست آمده، دو محاسبه نرم اقلدیسی وجود دارد و از آنجایی که نرم اقلیدسی برای محاسبه در هر دوره آموزشی از لحاظ زمانی هزینه زیادی دارد، باید راه حلی برای تسریع این کار ارائه شود. برای این منظور رابطه ای محاسبه می‌شود که مقدار نرم اقلیدسی را در هر دور آموزش تغییر دهد و با پارامترهای موجود، به جای محاسبه مستقیم نرم، آن را بدست می‌آورد. اگر فرض شود در مرحله ابتدایی آموزش $Norm_j^+ = \left\|\alpha_j^{+T} K \alpha_j^+\right\|^2$ و $Norm_j^- = \left\|\alpha_j^{-T} K \alpha_j^-\right\|^2$ باشد، روابط نهایی محاسبه تکرارشونده نرم، در رابطه‌ی (۲۳) و (۲۴) آمده است.

$$\begin{aligned}\left(Norm_j^+\right)_{new} &= \left(\mu\beta(\tau_j - y_j) - 1\right)^2 \left(Norm_j^+\right)_{old} \\ &+2\left(\mu\beta(\tau_j - y_j) - \left(\mu\beta(\tau_j - y_j)\right)^2\right)\alpha_j^{+T} K_{x_n} \\ &+ \left(\mu\beta(\tau_j - y_j)\right)^2 K_{x_n} p\end{aligned} \quad (۲۳)$$

$$\begin{aligned}\left(Norm_j^-\right)_{new} &= \left(\mu\beta(y_j) - 1\right)^2 \left(Norm_j^-\right)_{old} \\ &+2\left(\mu\beta(y_j) - \left(\mu\beta(y_j)\right)^2\right)\alpha_j^{-T} K_{x_n} \\ &+ \left(\mu\beta(y_j)\right)^2 K_{x_n} p\end{aligned} \quad (۲۴)$$

این تغییر، باعث کاهش پیچیدگی محاسبه نرم از $O(MN^2)$ به O(MN) خواهد شد.

۳-۳- افزودن دادگان به‌اشتباه-دسته‌بندی‌شده به مجموعه انتخابی نیستروم

در طول آموزش، تعداد زیادی از دادگان آموزشی به اشتباه به یک کلاس نسبت داده می‌شوند که در مرور زمان با طی شدن روند یادگیری، در صورت داشتن داده انتخابی نیستروم کافی اصلاح می‌شوند. با این حال یکی از روش‌هایی که می‌تواند در دقت سنجش دادگان آزمون موثر باشد، افزودن دادگان به‌اشتباه-دسته‌بندی‌شده که دراین مجموعه حاضر نبودند به مجموعه دادگان انتخابی نیستروم می‌باشد. معمولا این نوع داده‌ها، به دلیل شباهت بیش از اندازه به کلاس‌های دیگر، تقریب خوبی از مجموعه دادگان انتخابی فعلی دریافت نمی‌کنند. این عمل باعث می‌شود که با وارد شدن داده به

مجموعه دادگان انتخابی نیستروم، هم خود داده به درستی بسط داده شود و هم با افزایش دادگان، تعمیم پذیری بهتری جهت تقریب با بسط هسته داشته باشیم. . به این روش Incremental KerNys ±ED-WTA گفته می‌شود.

دقت کنید که مجموعه دادگان آموزشی ثابت می‌ماند و فقط مجموعه دادگان انتخابی نیستروم دچار تغییر می‌شوند. لازم به ذکر است که دادگان به‌اشتباه-دسته‌بندی‌شده در مجموعه دادگان آموزشی حاضر است ولی جهت تقریب زدن در مجموعه دادگان نیستروم وجود نداشتند که به این مجموعه افزوده می‌شوند.

باید این نکته را مورد توجه قرار داد که در ابتدای آموزش، تعداد بسیار زیادی از دادگان اشتباه تشخیص داده می‌شوند و با افزودن دادگان در این مراحل، مزایای مطرح شده در دو بخش قبل از دست می‌رود. بنابراین برای رفع این مشکل، افزودن داده به‌اشتباه-دسته‌بندی‌شده به مجموعه انتخابی نیستروم، به مراحل بعد آموزش موکول شده که شبکه فرآیند یادگیری را تا بخش مناسبی پیش برده باشد. این بخش باعث افزایش سربار زمانی شده؛ اما با این حال، باعث افزایش قابل توجه دقت و همچنین توضیح‌پذیری بسیار بهتر می‌شود.

۴- آزمایشات

در این بخش، نسخه ای از مجموعه دادگان MNIST استفاده خواهد شد که متشکل از ۶۰ هزار داده آموزشی و ۱۰ هزار داده آزمون خواهد بود که دارای ۱۰ کلاس می‌باشند و هرکدام از دادگان دارای ابعاد ۲۸ * ۲۸ خواهند بود. برای هر کلاس ۶ الگو در نظر گرفته می‌شود که در روش ارائه شده این مقاله، هرکدام از الگوها دارای بسط متفاوت خود خواهند بود. برای مجموعه دادگان نیستروم ۵۰۰۰ داده تعیین شده که از این ۵۰۰۰ داده با داشتن ۱۰ کلاس، برای هرکلاس ۵۰۰ داده جدا شده است. مقدار دهی اولیه بسط نیز با K-Means انجام خواهد شد که با مقداردهی میانگین خوشه‌های بدست آمده هر کلاس به عنوان ضریب های مربوطه‌ی اولیه بسط انجام می‌گیرد. توابع هسته مورد استفاده، تابع هسته خطی، چندجمله‌ای اصلاح شده و تابع پایه شعاعی گوسی خواهند بود. همچنین افزودن دادگان به‌اشتباه-دسته‌بندی‌شده به مجموعه دادگان انتخابی نیستروم، بعد از ششمین دوره آموزش انجام می‌گیرد.

برای انجام آزمایشات، از کامپیوتری با پردازنده مرکزی intel core i7 6700hq به همراه حافظه 16GB استفاده شده است. همچنین تمام پردازش‌های انجام شده بر روی پردازنده مرکزی انجام گرفته است. با مشخصات ذکر شده کامپیوتر، در روش اولیه Kernel ±ED-WTA توانایی ذخیره ماتریس هسته کلی بین دادگان، به دلیل محدودیت حافظه نیست و نیاز است تا در هر دور آموزش، k_{x_n} برای استفاده در رابطه‌ی (۱۲) محاسبه شود. اما در روش KerNys ±ED-WTA به دلیل کاهش یافتن ابعاد لازم برای محاسبه ماتریس هسته، در ابتدای آموزش محاسبه ماتریس هسته بین مجموعه دادگان نیستروم انجام شده و در حافظ ذخیره می‌شوند. این کار باعث بهبود قابل توجه زمان هر دور آموزش در نتیجه نهایی می‌شود. در این مقاله برای انتخاب تابع هسته، زیرمجموعه‌ای مشخص از دادگان آموزشی تحت عنوان دادگان اعتبارسنجی انتخاب شده و توابع هسته امتحان و پارامترهای هسته انتخاب می‌شوند.

دقت روش‌های ارائه شده در این مقاله با نتایج مقاله [۱۳] مقایسه می‌شود. در جدول (۲) نتایج بدست آمده بر روی دادگان آزمون با توابع مختلف هسته و همچین حالت بدون افزودن دادگان اشتباه و حالت افزودن دادگان اشتباه به مجموعه دادگان انتخابی نیستروم، نیز سنجیده شده است.

جدول (۲): نتایج بدست آمده بر اساس توابع هسته مختلف به همراه زمان هر دور آموزش

مدل	تابع هسته			دقت برتر	زمان هر دور آموزش (دقیقه)
	خطی	چند جمله‌ای اصلاح شده	گوسی		
±ED-WTA	-	-	-	٪۹۶,۵۳	۰,۱
Kernel ±ED-WTA	٪۹۶,۵	**٪۹۸,۰۷**	-	٪۹۸,۰۷	۱۶۱
KerNys ±ED-WTA	٪۹۶,۴	٪۹۷,۴۴	٪۹۷,۵۸	٪۹۷,۵۸	**۳,۱۸**
Incremental KerNys ±ED-WTA	٪۹۶,۳۱	٪۹۷,۷۵	**٪۹۸,۰۱**	٪۹۸,۰۱	± ۰,۷۵ ۵,۷۵

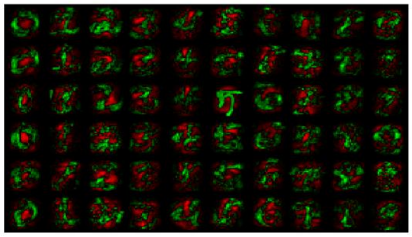
برای تولید نتایج توضیح‌پذیر قابل مقایسه با مقاله [۱۳] نیاز هست تا با بسط‌های بدست آمده در طول آموزش، تصویری برای نمایش تولید شود. با توجه به اینکه بسط‌ها میزان مشارکت یک داده را می‌گویند، می‌توان با ضرایب بسط، جمع وزن دار روی دادگان ورودی زد.

$$w_j^+ = \sum_{i=1}^{Size(S_c)} \alpha_j^{i^+} x_{S_c^i}$$
$$w_j^- = \sum_{i=1}^{Size(S_c)} \alpha_j^{i^-} x_{S_c^i} \qquad (۲۵)$$

بنابراین می‌توان نتایج شکل (۳) و شکل (۴) را بدست آورد و تفاضل این دو وزن مثبت و منفی نیز در تصویر شکل (۵) نمایش داده شده است. همانطور که مشاهده می‌شود، نتایجی مشابه به مقاله [۱۳] گرفته شده است و همچنان به دلیل شباهت بالای الگوهای مثبت و منفی و همچنین محو بودن الگوها، توضیح‌پذیری خوبی وجود نخواهد داشت. بنابراین در این مقاله روش جدیدی برای این کار ارائه خواهد شد.

شکل (۳): میانگین الگوهای مثبت KerNys ±ED-WTA. از ستون سمت چپ شش الگوی مثبت کلاس عدد ۰ تا سمت راست شش الگوی مثبت کلاس عدد ۹

شکل (۴): میانگین الگوهای منفی KerNys ±ED-WTA. از ستون سمت چپ شش الگوی منفی کلاس عدد ۰ تا سمت راست شش الگوی منفی کلاس عدد ۹



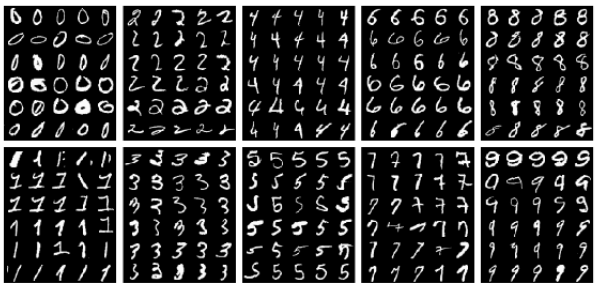
شکل (۵): میانگین حاصل تفریق الگوهای مثبت از منفی KerNys ±ED-WTA. از ستون سمت چپ شش الگوی نهایی مرتبط به کلاس عدد ۰ تا سمت راست شش الگوی نهایی مرتبط به کلاس عدد ۹

با توجه به اینکه در این مقاله از بسط مبتنی بر توابع هسته استفاده شده است، می‌توان تجربه کسب شده در طول آموزش را به ازای داده‌های مجزا به خاطر سپرد. اگر بیشترین تاثیر یک داده در بسط خواسته شود، کافی است بزرگترین ضریب هر بسط را جداسازی و نمایش داد. لازم به ذکر است که نتایج نشان داده شده مربوط به آزمایش KerNys ±ED-WTA بدون اضافه شدن دادگان گرفته شده می‌باشند.

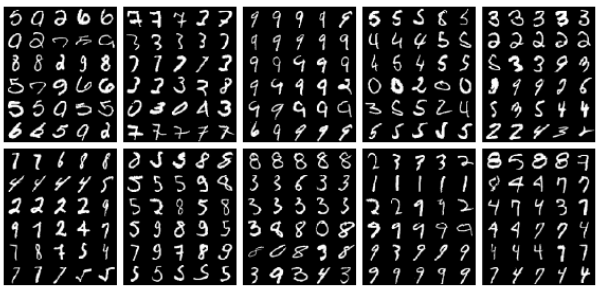
شکل (۶): داده با بزرگترین ضریب در بسط هر الگوی مثبت در KerNys ±ED-WTA. از ستون سمت چپ شش الگوی مثبت کلاس عدد ۰ تا سمت راست شش الگوی مثبت کلاس عدد ۹

شکل (۷): داده با بزرگترین ضریب در بسط هر الگوی منفی در KerNys ±ED-WTA. از ستون سمت چپ شش الگوی منفی کلاس عدد ۰ تا سمت راست شش الگوی منفی کلاس عدد ۹

نتایج شکل (۶) که نمایانگر الگوهای مثبت می‌باشند، طبق انتظار، مدل تمامی الگوها به درستی دادگان کلاس خود را یاد گرفته است. اما برخلاف مقاله [۱۳] که الگوهای منفی شباهت بسیار زیادی به الگوهای مثبت داشتند، می‌توان در الگوهای منفی به صورت دقیق دادگان به‌اشتباه-دسته‌بندی‌شده را مشاهده کرد. مثلا در ستون اول سمت چپ شکل (۶) و شکل (۷)، که الگوهای مثبت و منفی کلاس ۰ به نمایش گذاشته شده است، مشاهده می‌شود در الگوهای منفی تمامی دادگان به‌اشتباه-دسته‌بندی‌شده دارای انحنا می‌باشند و یا برای کلاس ۱، دارای کشیدگی می‌باشند.



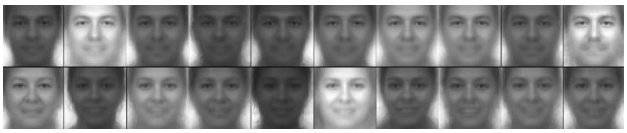
شکل (۸): پنج داده با بزرگترین ضریب در بسط هر الگوی مثبت در KerNys ±ED-WTA که هر سطر نشان‌دهنده‌ی یک الگو می‌باشد. از سمت چپ بالا بلوک الگوهای مثبت کلاس عدد ۰ تا سمت راست پایین بلوک الگوهای مثبت کلاس عدد ۹



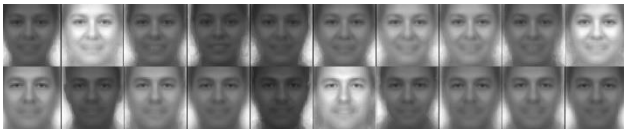
شکل (۹): پنج داده با بزرگترین ضریب در بسط هر الگوی منفی در KerNys ±ED-WTA که هر سطر نشان‌دهنده‌ی یک الگو می‌باشد. از سمت چپ بالا بلوک الگوهای منفی کلاس عدد ۰ تا سمت راست پایین بلوک الگوهای منفی کلاس عدد ۹

با توجه به اینکه روش ارائه شده بر اساس بسط یاد گرفته شده دادگان می‌باشد، علاوه بر نمایش داده با بیشترین تاثیر، می‌توان دادگان بیشتری در هر الگو را به نمایش گذاشت. به عنوان مثال، در شکل (۸) و شکل (۹) برای هر الگو ۵ داده با بزرگترین ضریب در بسط انتخاب شده و به نمایش گذاشته شده است. از این نمونه‌ها به عنوان استنباطی برای یادگیری استفاده شده که باعث توضیح‌پذیری قابل توجهی خواهد شد.

مجموعه دادگان دیگری با نام FERET موجود است. با در نظر گرفتن مجموعه دوکلاسه جنسیت مذکر و مونث، تعداد ۸۲۲ نمونه جنسیت مذکر و ۶۵۴ نمونه جنسیت مونث که از هر شخص حداقل دو نمونه در مجموعه دادگان وجود دارد. همانند مقاله [۱۳]، به دقت ۱۰۰ درصد با تابع هسته چندجمله‌ای با درجه دو رسیده، اما توضیح‌پذیری بیشتری در این مقاله حاصل شده است. در این مجموعه دادگان با توجه به دو کلاسه بودن، ۱۰ خوشه در نظر گرفته شده است. در شکل (۱۰) و شکل (۱۱) میانگین الگوها برای مقایسه با مقاله [۱۳] آورده شده است.

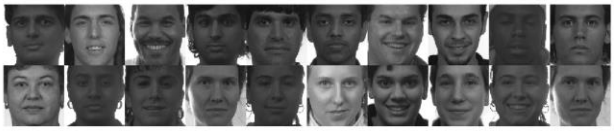


شکل (۱۰): میانگین الگوهای مثبت در Kernel ±ED-WTA. سطر بالا مربوط به الگوهای مثبت مردان و سطر پایین مربوط به الگوهای مثبت زنان است.

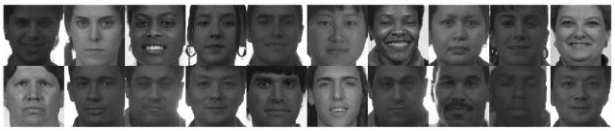


شکل (۱۱): میانگین الگوهای منفی در Kernel ±ED-WTA. سطر بالا مربوط به الگوهای منفی مردان (که زنان هستند) و سطر پایین مربوط به الگوهای منفی زنان (که مردان هستند) است.

در شکل (۱۲) و شکل (۱۳) داده با بزرگترین ضریب در بسط مرتبط انتخاب شده است. همانطور که مشاهده می‌شود، دادگان به درستی تفکیک شده که برای هر کلاس و به ازای هر الگو، دلیل تصمیم با نمونه‌ها توضیح داده شده است. توجه کنید که در مجموعه دادگان کوچک به واسطه مقداردهی اولیه، ممکن است مدل به الگوهای تکراری برسد.

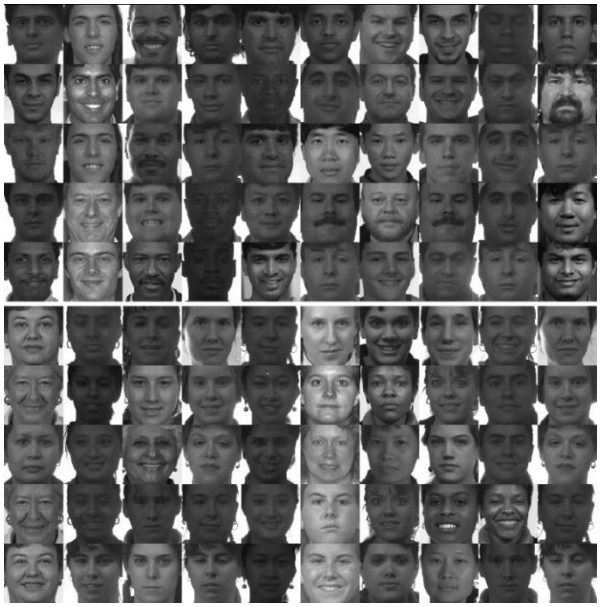


شکل (۱۲): داده با بزرگترین ضریب در بسط هر الگوی مثبت در Kernel ±ED-WTA. سطر بالا مربوط به الگوهای مثبت مردان و سطر پایین مربوط به الگوهای مثبت زنان است که به‌جای خود الگو (که میانگین دادگان است)، داده‌ی با بیشترین ضریب را نشان داده‌ایم.

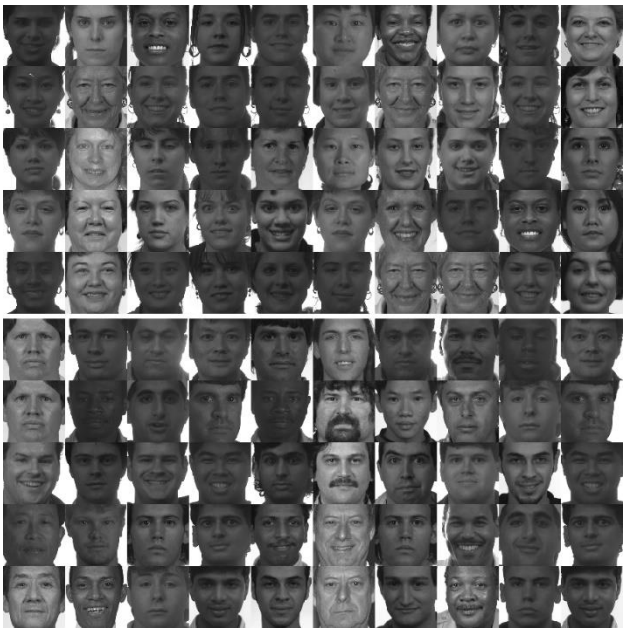


شکل (۱۳): داده با بزرگترین ضریب در بسط هر الگوی منفی در Kernel ±ED-WTA. سطر بالا مربوط به الگوهای منفی مردان (که زن هستند) و سطر پایین مربوط به الگوهای منفی زنان (که مرد هستند) می‌باشد که به‌جای خود الگو (که میانگین دادگان است)، داده‌ی با بیشترین ضریب را نشان داده‌ایم.

همچنین در شکل (۱۴) و شکل (۱۵) پنج داده با بزرگترین ضریب در بسط آمده است. لازم به ذکر است که در این مجموعه دادگان به دلیل وجود دادگان محدود، احتمال بیشتری برای انتخاب شدن تکراری یک داده برای یک الگو وجود دارد.



شکل (۱۴): پنج داده با بزرگترین ضریب در بسط هر الگوی مثبت در Kernel ±ED-WTA که هر ستون نماینده یک الگو می‌باشد. پنج سطر بالا مربوط به الگوهای مثبت مردان و پنج سطر پایین مربوط به الگوهای مثبت زنان است.



شکل (۱۵): پنج داده با بزرگترین ضریب در بسط هر الگوی منفی در Kernel ±ED-WTA که هر ستون نماینده یک الگو می‌باشد. پنج سطر بالا مربوط به الگوهای منفی مردان (که زنان هستند) و پنج سطر پایین مربوط به الگوهای منفی زنان (که مردان هستند) می‌باشد. توجه کنید که الگوی منفی زنان سمت چپ، دو داده شبیه به هم بیشترین تاثیر را داشته اند که با توجه به گردن شخص، می‌توان فهمید که دو تصویر مجزا از یک شخص هستند.

۵- نتیجه‌گیری

در این مقاله نسخه هسته الگوریتم آموزش شبکه ED-WTA± معرفی شد. نشان داده شد که اعمال توابع هسته بر روی الگوریتم، باعث بهبود دقت و همچنین بهبود توضیح‌پذیری می‌شود. مشاهده شد که مشکل زمانی و فضایی در نسخه هسته بوجود می‌آید و برای این منظور، مقاله از روش تقریب نیستروم و همچنین محاسبه تکرار شونده نرم اقلیدسی استفاده کرد. همچنین توضیح‌پذیری مبتنی بر نمونه ارائه شد که به‌جای آنکه همانند شبکه‌ی ED-WTA± از الگوهای میانگین داده‌ها استفاده کند، از خود نمونه‌های آموزشی به عنوان نماینده‌ی الگوها استفاده می‌کند. برخلاف الگوهای میانگین که حالتی محو دارند، خود داده‌ها بسیار واضح هستند.

مراجع

[1] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nat. Mach. Intell., vol. 1, no. 5, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.

[2] C. Molnar, "Interpretable Machine Learning Products. A Guide for Making Black Box Models Explainable." 2019. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[3] A. Kortylewski, Q. Liu, A. Wang, Y. Sun, and A. Yuille, "Compositional Convolutional Neural Networks: A Robust and Interpretable Model for Object Recognition Under Occlusion," Int. J. Comput. Vis., vol. 129, no. 3, pp. 736–760, 2021, doi: 10.1007/s11263-020-01401-3.

[4] Q. Zhang, X. Wang, Y. N. Wu, H. Zhou, and S. C. Zhu, "Interpretable CNNs for Object Classification," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 10, pp. 3416–3431, 2021, doi: 10.1109/TPAMI.2020.2982882.

[5] Y. Lou, R. Caruana, and J. Gehrke, "Intelligible models for classification and regression," Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 150–158, 2012, doi: 10.1145/2339530.2339556.

[6] G. Singh and K. C. Yow, "These do not Look like Those: An Interpretable Deep Learning Model for Image Recognition," IEEE Access, vol. 9, pp. 41482–41493, 2021, doi: 10.1109/ACCESS.2021.3064838.

[7] A. H. Gee, D. Garcia-Olano, J. Ghosh, and D. Paydarfar, "Explaining deep classification of time-series data with learned prototypes," CEUR Workshop Proc., vol. 2429, pp. 15–22, 2019.

[8] F. Doshi-Velez et al., "Accountability of AI under the law: The role of explanation," arXiv, 2017, doi: 10.2139/ssrn.3064761.

[9] J. V. Jeyakumar, J. Noor, Y.-H. Cheng, L. Garcia, and M. Srivastava, "How Can I Explain This to You? An Empirical Study of Deep Neural Network Explanation Methods," Adv. Neural Inf. Process. Syst. 33 (NeurIPS 2020), no. NeurIPS, p. 12, 2020, [Online]. Available: files/485/Jeyakumar et al. - How Can I Explain This to You An Empirical Study .pdf

[10] R. Guidotti, A. Monreale, S. Matwin, and D. Pedreschi, "Black Box Explanation by Learning Image Exemplars in the Latent Feature Space," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11906 LNAI, pp. 189–205, 2020, doi: 10.1007/978-3-030-46150-8_12.

[11] N. Papernot and P. Mcdaniel, "Deep k-Nearest Neighbors : Towards Confident , Interpretable and Robust Deep Learning," no. c.

[12] S. Saralajew, L. Holdijk, M. Rees, E. Asan, and T. Villmann, "Classification-by-components: Probabilistic modeling of reasoning over a set of components," Adv. Neural Inf. Process. Syst., vol. 32, no. NeurIPS, pp. 1–12, 2019.

[13] R. Zarei-Sabzevar, K. Ghiasi-Shirazi and A. Harati, "Prototype-Based Interpretation of the Functionality of Neurons in Winner-Take-All Neural Networks," in IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2022.3155174.

[14] E. Pękalska, R. P. W. Duin, and P. Paclík, "Prototype selection for dissimilarity-based classifiers," Pattern Recognit., vol. 39, no. 2, pp. 189–208, 2006, doi: 10.1016/j.patcog.2005.06.012.

[15] W. Tang, L. Liu, and G. Long, "Interpretable time-series classification on few-shot samples," arXiv, 2020.

[16] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," arXiv, 2017.

[17] O. Li, H. Liu, C. Chen, and C. Rudin, "Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions," arXiv, pp. 3530–3537, 2017.

[18] C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin, "This looks like that: Deep learning for interpretable image recognition," arXiv, no. NeurIPS, pp. 1–12, 2018.

[19] K. Ghiasi-Shirazi, "Generalizing the Convolution Operator in Convolutional Neural Networks," Neural Process. Lett., vol. 50, no. 3, pp. 2627–2646, 2019, doi: 10.1007/s11063-019-10043-7.

[20] S. O. Arik and T. Pfister, "Protoattend: Attention-based prototypical learning," J. Mach. Learn. Res., vol. 21, pp. 1–35, 2020.

[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization," Rev. do Hosp. das Cl??nicas, vol. 17, pp. 331–336, 2016, [Online]. Available: http://arxiv.org/abs/1610.02391

[22] R. Agarwal, N. Frosst, X. Zhang, R. Caruana, and G. E. Hinton, "Neural additive models: Interpretable machine learning with neural nets," arXiv, 2020.

[23] B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio, "N-BEATS: Neural basis expansion analysis for interpretable time series forecasting," arXiv, pp. 1–31, 2019.

[24] F. B. Schölkopf, Bernhard, Alexander J. Smola, Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2002.

[25] Bishop, Pattern Recognition and Machine Learning, no. 8. 2006. doi: 10.1088/1751-8113/44/8/085201.

[26] K. Ghiasi-Shirazi, "Learning 2D Gabor filters by infinite kernel learning regression," J. Comput. Math. Data Sci., vol. 1, p. 100016, Sep. 2021, doi: 10.1016/J.JCMDS.2021.100016.

[27] Williams; Christopher; and Matthias Seeger, "Using the Nyström method to speed up kernel machines," Proc. 14th Annu. Conf. neural Inf. Process. Syst., pp. 3–9, 2001.

[28] J. Lu, S. C. H. Hoi, J. Wang, P. Zhao, and Z. Y. Liu, "Large scale online kernel learning," J. Mach. Learn. Res., vol. 17, pp. 1–43, 2016.

[29] H. Xiong, M. N. S. Swamy, and M. O. Ahmad, "Optimizing the kernel in the empirical feature space," IEEE Trans. Neural Networks, vol. 16, no. 2, pp. 460–474, 2005, doi: 10.1109/TNN.2004.841784.

[30] J. Bien and R. Tibshirani, "Prototype selection for interpretable classification," Ann. Appl. Stat., vol. 5, no. 4, pp. 2403–2424, 2011, doi: 10.1214/11-AOAS495.

[31] P. Honeine, "Online kernel principal component analysis: A reduced-order model," IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 9, pp. 1814–1826, 2012, doi: 10.1109/TPAMI.2011.270.

[32] Z. Xu, Q. Song, F. Haijin, and D. Wang, "Online prediction of time series data with recurrent kernels," Proc. Int. Jt. Conf. Neural Networks, 2012, doi: 10.1109/IJCNN.2012.6252747.

[۳۳] غیاثی راد, حسین علی, علیاری شوره دلی و فریور. "تحلیل و مقایسه ۲۱ قید محدودسازی در الگوریتم گرادیان نزولی اتفاقی بـه روش کرنـل." نشـریه مهندسی برق و الکترونیک ایران: ۱۴۰۰.

زیرنویس‌ها

[1] Black Box Models

[2] Explanation

[3] Positive / Negative Euclidean Distance Winner Takes All (±ED-WTA)

[4] Teacher - Student

[5] Meta Learning

[6] Components

[7] Convolutional Neural Networks

[8] Prototype

[9] Autoencoder

[10] Attention

[11] Encoder

[12] Additive Models

[13] Shapelets

[14] Hilbert

[15] Radial Basis Function (RBF)

[16] Explicit

[17] Implicit

[18] Kernel Trick

[19] Gabor

[20] SoftMax

[21] competitive cross entropy (CCE)

[22] Nyström

[23] Euclidean norm

[24] Projection operator