**RCALAD: Regularized Complete Adversarially Learned Anomaly detection**

**چکيده**

یکی از مهم‌ترین فعالیت‌های حوزه تحلیل داده تشخیص ناهنجاری می‌باشد که در طیف وسیعی از کاربرد‌ها همچون تشخیص جعل، کاربردهای پزشکی و سیستم‌های امنیتی به کار گرفته می‌شود. علی‌رغم وجود روش‌های آماری و مبتنی بر یادگیری ماشین طراحی مدل‌های موثر در تشخیص ناهنجاری در فضای داده پیچیده با ابعاد بالا همچنان به عنوان یک چالش اساسی باقی‌ مانده است. شبکه‌های مولد تخاصمی قادرند تا بر چالش مورد نظر فائق آمده و توزیع داده‌های دنیای واقعی که دارای پیچیدگی و ابعاد بالا هستند را مدل کنند و همین امر سبب می‌شود تا عملکرد امیدوارکننده‌ای در زمینه تشخیص ناهنجاری از خود نشان دهند. در این کار چارچوب تخاصمی RCALAD با هدف تشخیص ناهنجاری ارائه شده است. اساس کار مدل پیشنهادی بازسازی داده ورودی با استفاده از شبکه مولد و در ادامه محاسبه میزان اختلاف داده اصلی و بازسازی آن به منظور شناسایی نمونه‌های ناهنجار است. لازمه شناسایی موثر نمونه‌های ناهنجار بازسازی ضعیف داده‌های ناهنجار است. مدل پیشنهادی نقاط ضعف کارهای پیشین را پوشش داده و بر دو جنبه تضمین بازسازی ضعیف نمونه‌های ناهنجار و همچنین استفاده از بیشینه اطلاعات موجود در شبکه به صورت توامان برای آموزش بهتر، تمرکز کرده است. تشخیص ناهنجاری با استفاده از خطای بازسازی نیازمند تعریف امتیاز ناهنجاری مناسب است، بنابراین علاوه بر معماری پیشنهادی دو امتیاز ناهنجاری جدید نیز در این کار ارائه شده است. نتایج تجربی بیانگر برتری و قدرت مدل RCALAD در مقایسه با سایر مدل‌های به روز و مطرح در زمینه تشخیص ناهنجاری بوده است.

كليد واژه- تشخیص ناهنجاری، یادگیری ماشین، شبکه مولد تخاصمی، خطای بازسازی، امتیاز ناهنجاری.

**1- مقدمه**

هنگام تجزیه و تحلیل دادگان موجود در دنیای واقعی، شناسایی نمونه‌های غیرمشابه با سایر نمونه‌ها امری ضروری به نظر می‌رسد. چنین نمونه‌هایی با عنوان ناهنجاری شناخته می‌شوند و از عملیات شناسایی چنین نمونه‌هایی با عنوان مسئله تشخیص ناهنجاری یاد می‌شود. این مسئله یک بخش حائز اهمیت از زمینه تحقیقاتی داده‌کاوی است چرا که شامل کشف الگو‌های جذاب و نادر در داده‌هاست[1].

ناهنجاری‌ها جزو پارامترهای مهم هر مجموعه داده‌ای‌ در نظر گرفته مي‌شوند و در دامنه وسیعی از کاربردها تاثیرگذار هستند. به عنوان مثال، الگوی غير معمول ترافيك در یك شبكه کامپیوتری ميتواند به معنای هك شدن رایانه و انتقال داده‌ها به مقصدهای غيرمجاز باشد. رفتار غير عادی در معاملاتی که توسط کارت‌های اعتباری انجام می‌شوند مي‌تواند نشانگر فعاليت‌های اقتصادی با هدف کلاهبرداری باشد[2]، و یا یك ناهنجاری در تصویر MRI ممكن است وجود تومور بدخيم را نشان دهد.[3] علی‌رغم وجود روش‌های آماری و مبتنی بر یادگیری ماشین، طراحی مدل‌های موثر در تشخیص ناهنجاری در فضای داده پیچیده با ابعاد بالا همچنان به عنوان یک چالش اساسی باقی‌ مانده است[4].

شبکه‌های مولد تخاصمی قادرند تا بر چالش مورد نظر فائق آمده و توزیع داده‌های دنیای واقعی که دارای پیچیدگی و ابعاد بالا هستند را مدل کنند و همین امر سبب می‌شود تا عملکرد امیدوارکننده‌ای در زمینه تشخیص ناهنجاری از خود نشان دهند. در شبكه‌های مولد تخاصمی، یک مدل مولد در برابر یك مدل تمایزگر قرار مي‌گيرد، مدل تمایزگر سعی می‌کند میان داده‌های واقعی و داده‌های تولیدی توسط شبکه مولد تمایز ایجاد کند. در شبكه مولد تخاصمی به طور همزمان دو مدل مولد و تمایزگر آموزش داده مي‌شود. مدل مولد G توزیع داده را ضبط مي‌کند و مدل تمایزگر D که احتمال این که نمونه از داده‌های توليد شده توسط G باشد را تخمين مي‌زند. تابع هدف برای شبكه مولد G به حداکثر رساندن احتمال اشتباه شبكه D است. این بستر منجر به یك بازی دو نفره مانند بازی‌های بیشینه-کمینه مي‌شود[5]. توانایی شبکه‌های عصبی تخاصمی در مدل کردن تصاویر طبیعی ثابت شده است[6] [7]، و بر کاربرد‌ آن‌ها در زمینه‌های پردازش گفتار و متن [4]و تصاویر پزشکی روز به روز افزوده می‌شود[8]

در این کار روشی کارآمد و موثر مبتنی بر شبکه‌های مولد تخاصمی که به صورت خاص با هدف تشخیص ناهنجاری، طراحی شده است، پیشنهاد می‌شود. مانند بسیاری از الگوریتم‌های مبتنی بر یادگیری، در این‌ جا دو مرحله اصلی آموزش و آزمایش وجود دارد. در قسمت آموزش همانند دیگر چهارچوب‌های تخاصمی، به نوبت بخش مولد و بخش تمایزگر را آموزش می‌دهیم تا هر دو بخش در عین تناسب به نوبت به‌روزرسانی شود. در اینجا با هدف stabilize کردن آموزش ساختار تخاصمی از توزیع توام پارامترهای موجود در شبکه استفاده شده است و به منظور اعمال نگاشت معکوس نمونه‌های ورودی به فضای نهفته، به طور توام با شبکه تمایزگر و مولد، کدگذار E آموزش داده می‌شود. در بیشتر کارهای قبلی در زمینه تشخیص ناهنجاری، بر تخمین توزیع (density estimation) داده‌های هنجار تمرکز شده است و هیچ لزومی برای بازسازی ضعیف داده های ناهنجار وجود ندارد[8]. در مدل RCALAD با تمرکز بر هدف بازسازی هر چه ضعیف‌تر نمونه‌های ناهنجار، توزیع جریمه به ساختار تخاصمی پیشنهادی اضافه شده است تا کدگذار و مولد به سمت توزیع نمونه‌های هنجار بایاس شوند. آزمایش‌ها روی طیف‌های مختلف دادگان تصویری و جدولی با ابعاد بالا صورت پذیرفته‌اند و نتایج حاصل بیانگر عملکرد بهتر مدل پیشنهادی نسبت به سایر مدل‌های پایه بوده است.

**2- کارهای مرتبط**

تشخیص ناهنجاری به نام‌های شناسایی نوآوری و تشخیص داده پرت نیز شناخته می‌شود، این مسئله همانطور که در[9][10][11] بررسی شد به طور گسترده مورد مطالعه قرار گرفته است. روش‌هایی قبلی که تاکنون در این زمینه مورد استفاده قرار گرفته است به طور کلی به دو دسته representation learning و generative model تقسیم می‌شوند.

روش‌های مبتنی بر representation learning با استخراج وبژگی‌های اصلی و یا یادگیری یک نگاشت از داده‌های نرمال مسئله تشخیص ناهنجاری را حل می‌کند. one-class support vector machine مرز حاشیه‌ای حول داده نرمال را پیدا می‌کند[12]. Deep support vector data description به اختصار DSVDD یک ابر کره به منظور محصور ساختن بازنمایی نمونه‌های نرمال را پیدا می‌کند[13].. liu and gryllias constructed frequency domain features using cyclic spectral analysis and used them svdd frame. This method has been proved robust against outliers and can achieve a high detection rate for bearing anomaly detection.

Odin با استفاده از مقیاس‌بندی دما و اغتشاشات روی یک شبکه عصبی از پیش آموزش دیده شده به شناسایی نمونه‌های ناهنجار روی مجموعه داده‌های تصویری می‌پردازد[14].در [15] محققان رویکرد جدیدی برای شناسایی ناهنجاری‌های تصویری با آموزش مدل روی تصاویر نرمالی که با تبدیل هندسی تغییر یافته‌، ارائه داده‌اند. در این مدل دسته‌بند با استفاده از آماره‌های فعالساز سافتمکس امتیاز ناهنجاری را محاسبه می‌کند .

مدل‌های مولد معمولا تلاش می‌کنند تا بازسازی داده را یاد بگیرند و با استفاده از این بازسازی نمونه‌های ناهنجار را شناسایی می‌کنند.[16]به عنوان مثال خودکدگذارها توزیع داده نرمال را مدل می‌کنند و از خطای بازسازی به عنوان امتیاز ناهنجاری استفاده می‌شود[17][18]. Deep structured energy based models به اختصار DSEBM یک مدل مبتنی بر انرژی یاد میگیرد و هر نمونه را به یک امتیاز انرژی نگاشت می‌کند[19]. deep autoencoding gaussian mixture model به اختصار DAGMM با استفاده از یک کدگذار برای نمونه‌های نرمال یک توزیع گوسی مخلوط تخمین می‌زند. اخیرا از شبکه‌های عصبی تخاصمی در تشخیص ناهنجاری استفاده شده است. به عنوان مثال از این ساختار برای شناسایی ناهنجاری در تصاویر پزشکی استفاده شده است. در [8] کار نگاشت معکوس به فضای نهفته با استفاده از مکانیزم backpropagation بازگشتی صورت می‌پذیرد. در [20]در ادامه کار قبلی انجام شده است، در این کار به منظور کاهش هزینه محاسباتی نگاشت به فضای نهفته به وسیله شبکه کدگذار صورت انجام می‌گیرد.در [21]مدل ارائه شده بر اساس شبکه عصبی تقابلی دوطرفه به اختصار BiGAN بنا نهاده شده است. وظیفه نگاشت معکوس از فضای داده ورودی به فضای نهفته نیز بر عهده کدگذار است. بر خلاف ساختار استاندارد GAN که در آن تمایزگر تنها تصویر واقعی و تصویر تولیدی شبکه مولد را ورودی می‌گیرد، بازنمایی این تصاویر در فضای نهفته هم به عنوان ورودی به شبکه تمایزگر داده می‌شود.

**3- پیش نیازها/Preliminaries**

شبكه‌های مولد تقابلی اولین بار در سال 2014 توسط آقای گودفلو و همکاران ابداع شد[5]، در این شبکه­ها زیرشبکه مولد در برابر زیرشبکه تمایزگر قرار مي‌گيرد، این زیرشبکه‌ها روی مجموعه‌ M نمونه‌ای بدون برچسب آموزش می‌بینند. زیرشبکه مولد نمونه‌های انتخابی از فضای نهفته را به فضای داده ورودی نگاشت می‌کند. زیرشبکه تمایزگر سعی می‌کند میان داده‌ واقعی و داده‌ تولیدی توسط شبکه مولد یعنی تمایز ایجاد کند. این دو زیرشبکه با هم در رقابت هستند، شبکه مولد تلاش می‌کند تا توزیع داده ورودی را تقلید کند در حالی‌که شبکه تمایزگر تلاش می‌کند تا میان نمونه‌های واقعی و داده تولیدی زیرشبکه مولد تمیز دهد. در فاز آموزش شبکه مولد G و شبکه تمایزگر D به صورت متناوب با استفاده از گرادیان کاهشی و به نوبت بهینه می‌شوند.

توزیع روی داده ورودی به صورت نمایش داده می‌شود و به عنوان شبکه مولد در فضای نهفته در نظر گرفته می‌شود. آموزش شبکه GAN با پیدا کردن تمایزگر و مولدی که بتواند مسئله saddle point که به شکل است را حل کند، انجام می‌شود.

تابع به صورت زیر تعریف می‌شود:

حل این مسئله در نهایت به این نتیجه همگرا می شود که توزیع مولد با توزیع داده واقعی برابر باشد. یعنی تمایزگر بهینه سراسری در صورتی به دست خواهد آمد که اگر و تنها اگر PG(x)=q(x) باشد. منظور از PG توزیع یادگرفته شده توسط شبکه مولد است. اثبات این قضیه در مقاله [GAN] آمده است.

در مقاله Adversarially learned inference و یا به اختصار ALI تلاش شده تا با استفاده از کدگذار E(x) توزیع توام کدگذار به صورت و توزیع شبکه مولد به صورت مدل شود[22]. در اینجا توسط شبکه کدگذار یاد گرفته می‌شود. تابع هدف مدل ALI به صورت زیر است:

|  |  |
| --- | --- |
| **(3)** |  |

در معادله بالا Dxz بیانگر شبکه تمایزگر است که x و z را به عنوان ورودی می‌گیرد و مقدار خروجی آن مشخص کننده این است که با چه احتمالی ورودی‌های فعلی از توزیع نشئت گرفته است. کدگذار، شبکه مولد و تمایزگر در حالت بهینه خود قرار می‌گیرند اگر و تنها اگر . اثبات این قضیه در مقاله [ali] ارائه شده است.

علیرغم اینکه توزیع‌های و برای ما مشخص هستند ولی در عمل و حین آموزش مدل لزوما به سمت نقطه بهینه همگرا نمی‌شود. دلیل این اتفاق در مقاله [23]به مسئله پایداری چرخه که به صورت تعریف می‌شود نسبت داده شده است. برای حل این مسئله چارچوب ALICE پیشنهاد داد تا تمایزگر به ساختار شبکه ALI اضافه شود. تابع هدف این مدل به صورت زیر است:

|  |  |
| --- | --- |
| **(4)** |  |

در این کار نشان داده شد که بکارگیری تمایزگر Dxx از نظر تئوری به بهترین بازسازی برای داده ورودی خواهیم رسید[alice] . برای تثبیت آموزش در مدل پایه ALICE، در [4] توزیع­های شرطی را با اضافه­کردن یک تمایزگر دیگر به مدل اعمال کردند و سپس عملیات نرمال­سازی طیفی را انجام دادند. به صورت جزئی‌تر، در این­جا یک شبکه تمایزگر به مدل با هدف تضمین چرخه پایداری در فضای نهفته اضافه شده که وظیفه دارد تا متغیر در فضای نهان و بازسازی آن را تا حد امکان به یکدیگر شبیه کند. با کنار هم قرار دادن بلوک پیشنهاد شده در [alad] و اجزای مدل ALICE ، در نهایت تابع هزینه مدل ALAD به صورت زیر خواهد بود.

|  |  |
| --- | --- |
| **(5)** |  |

در این مقاله نشان داده شده که با افزودن قیود لیپسچیتز[[1]](#footnote-1) به تمایزگر شبکه GAN، فاز آموزش تثبیت خواهد شد همچنین در عمل نشان داده شده که با spectral normalization پارامترهای وزن، روی عملکرد شبکه بهبود خواهیم داشت.

با اینکه ایده alad به پایدار شدن چرخه کمک می‌کند اما همچنان متغیرهای فضای نهفته و ورودی را در دو فضای مستقل از هم بررسی شوند و از وابستگی ذاتی میان متغیرها چشم‌پوشی می­شود. به صورت دقیق‌تر متغیرهای x و در یک روند به صورت جدا از روند بررسی متغیرهای و بررسی می‌شود. در صورتی که روند بازسازی این دو جفت داده در طول یکدیگر قرار دارند و بر یکدیگر اثر مستقیسم می گذارند.

علاوه براین، مشکل دیگر این مدل فرض سهل انگارانه لزوم بازسازی ضعیف برای نمونه‌های ناهنجار است. در واقع در تمامی مدل­­های پیشین این فرض به طور ضمنی در نظر گرفته شده که اگر مدل با داده های هنجار آموزش ببیند ، لزوما برای داده های ناهنجار نگاشت ضعیفی خواهد داشت در حالی که هیچ قیدی به منظور متمایل کردن مدل به سمت تولید بازسازی ضعیف از نمونه‌‌های ناهنجار وجود ندارد. در مدل پیشنهادی RCALAD سعی شده است تا تمامی نقاط ضعف اشاره شده در مدل های پیشین پوشش داده بشود.

**4- روش پیشنهادی**

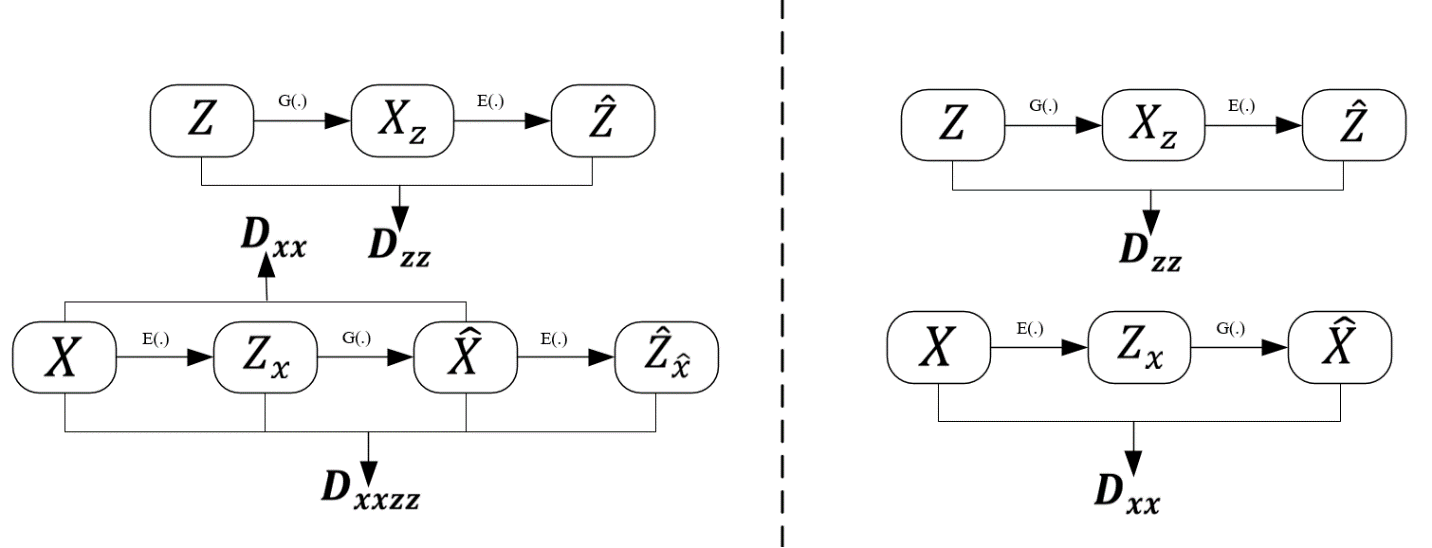
در این بخش به طور دقیق تر به بررسی جزئیات هر یک از مشکلات مورد اشاره در فصل قبل پرداخته می‌شود. ابتدا مسئله complete cycle consistency و روش حل آن شرح داده خواهد شد، سپس به بررسی مسئله استلزام بازسازی ضعیف می پردازیم و در انتها مدل پیشنهادی نهایی که با هدف حل هر دو مسئله ذکر شده طراحی شده است معرفی می‌شود.

**4-1- چرخه پایداری کامل**

همانطور که گفته شد، در مدل ALAD پایداری چرخه برای داده ورودی و متغیر فضای نهفته در دو روند مستقل از هم بررسی می‌شود. در واقع روند نزدیک کردن بازسازی متغیرفضای نهان یعنی به خود متغیر و همچنین نزدیک کردن بازسازی یعنی به خود به صورت جداگانه صورت می‌پذیرد. لازم به ذکر است در این‌جا مقصود از متغیر نمونه‌ای از توزیع گوسی است که به عنوان ورودی به شبکه مولد داده می‌شود و ارتباطی با نگاشت ورودی در فضای نهان ندارد .

مسئلهComplete Cycle Consistency و یا به اختصار CCC بیان می کند که به ازای هر متغیر از فضای ورودی اگر ابتدا کدگذار نگاشت معکوس به فضای نهفته را تخمین زند که معادل می‌باشد. و در مرحله بعد بازنمایی بدست آمده به شبکه مولد وارد ‌شود تا بازسازی شبکه از متغیر ورودی تولید شود. سپس همین بازسازی بار دیگر به شبکه کدگذار داده ­شود تا بازسازی در فضای نهفته نیز محاسبه شود یعنی . در این صورت، انتظار منطقی از هر شبکه مبتنی بر بازسازی این است که دو متغیر و همچنین دو متغیر تا حد امکان کمترین اختلاف را داشته باشند. یعنی تعریف مسئله CCC بدین ترتیب می­شود که، در هر مدل مبتنی بر بازسازی، بایستی به ازای هر داده ورودی و نگاشت آن در فضای نهان، بازسازی ارائه شده توسط شبکه برای هر دو متغیر کمترین خطا و بیشترین شباهت را با آن دو داشته باشد.

در مدل ALAD شباهت میان داده ورودی و بازسازی آن و همچنین شباهت z و بازسازی آن مستقل از هم و در دو چرخه جداگانه بررسی می‌شد و فرض شده بود که مستقل از هم هستند در حالی‌که می‌دانیم این دو چرخه کاملا به یکدیگر وابسته بوده و فرض استقلال این دو مسئله غلط است. در اینجا سعی شده است با بررسی توام متغیر‌های موجود در چرخه CCC در تمایزگر جدید ، عدم استقلال میان متغیرها مدل شود و جریان اطلاعات موجود در این زنجیره برای بهبود آموزش شبکه و تشخیص هر چه بهتر داده‌های ناهنجار به کار گرفته شود. تفاوت میان ورودی تمایزگر و ورودی تمایزگر که در مدل ALAD استفاده شده است در شکل 1 قابل مشاهده است.



شکل 1: سمت راست نحوه استفاده از متغیرهای فضای داده ورودی و فضای نهفته در چرخه پایداری شبکه ALAD. سمت چپ نحوه استفاده از اطلاعات یک چرخه کامل در مدل پیشنهادی.

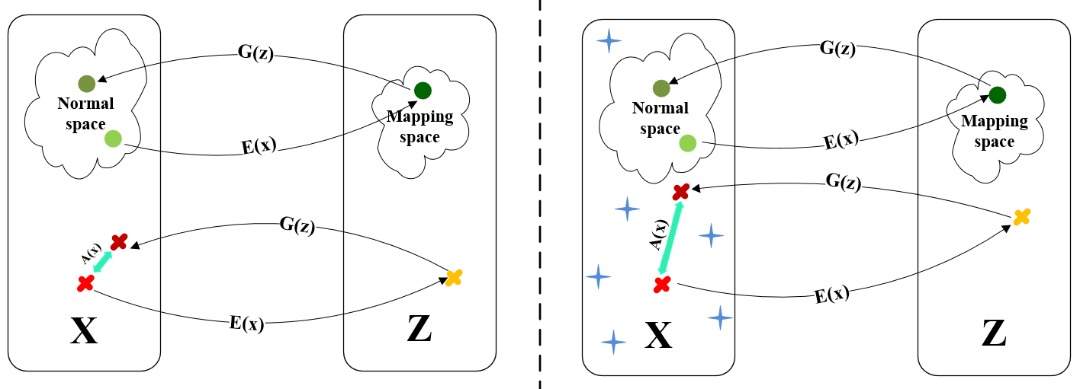
همانطور که در شکل 1 قابل مشاهده است مدل ALAD از اطلاعات یک چرخه کامل استفاده نمی‌کرد. به منظور استفاده از اطلاعات موجود در یک چرخه کامل متغیر جدید معرفی می‌شود. برای محاسبه این متغیر نگاشت معکوس داده ورودی به شبکه مولد داده می‌شود و نگاشت معکوس خروجی حاصل مجددا با استفاده از کدگشا محاسبه می‌شود و با این کار یک چرخه کامل از روند دگردیسی داده ورودی حاصل می‌شود.

به منظور تضمین شرط چرخه پایداری کامل از تمایزگر جدید با ورودی توام استفاده می‌شود، لازم به ذکر است اثربخشی تمایزگر توامان پیش از این یک مرتبه درALIGAN به اثبات رسیده است. این تمایزگر چهارتایی به عنوان داده واقعی و از چهارتایی به عنوان داده تقلبی ورودی می‌گیرد. در واقع این تمایزگر تلاش می‌کند تا ورودی و بازسازی ارائه شده از آن توسط شبکه و همینطور نگاشت معکوس تصویر ورودی در فضای نهان یعنی و بازسازی آن توسط کدگذار تا حد امکان به یکدیگر نزدیک باشد تا یک حلقه کامل پایدار توسط مدل ارائه شود و مدل هر چه بهتر آموزش دیده و stabilize شود.

**4-2- استلزام بازسازی ضعیف**

در مدل‌های مبتنی بر بازسازی تاکنون همیشه فرض بر این بوده است که اگر آموزش و بازسازی برای داده‌های هنجار به خوبی انجام بگیرد، بازسازی داده‌های ناهنجار لزوما ضعیف و متفاوت از داده ورودی خواهد بود. اما این پیش‌فرض در بسیاری از موارد صحیح نیست و نمونه بازسازی شده ناهنجار، میزان اختلاف کمی با نمونه ورودی دارد و به همین سبب، تشخیص آن به عنوان نمونه ناهنجار دشوار خواهد بود. در واقع در هیچ یک از مدل‌های پیشین هیچ استلزام یا قید کنترلی برای متمایل کردن مدل به سمت تولید بازسازی ضعیف برای نمونه‌های ناهنجار ارائه نشده است.

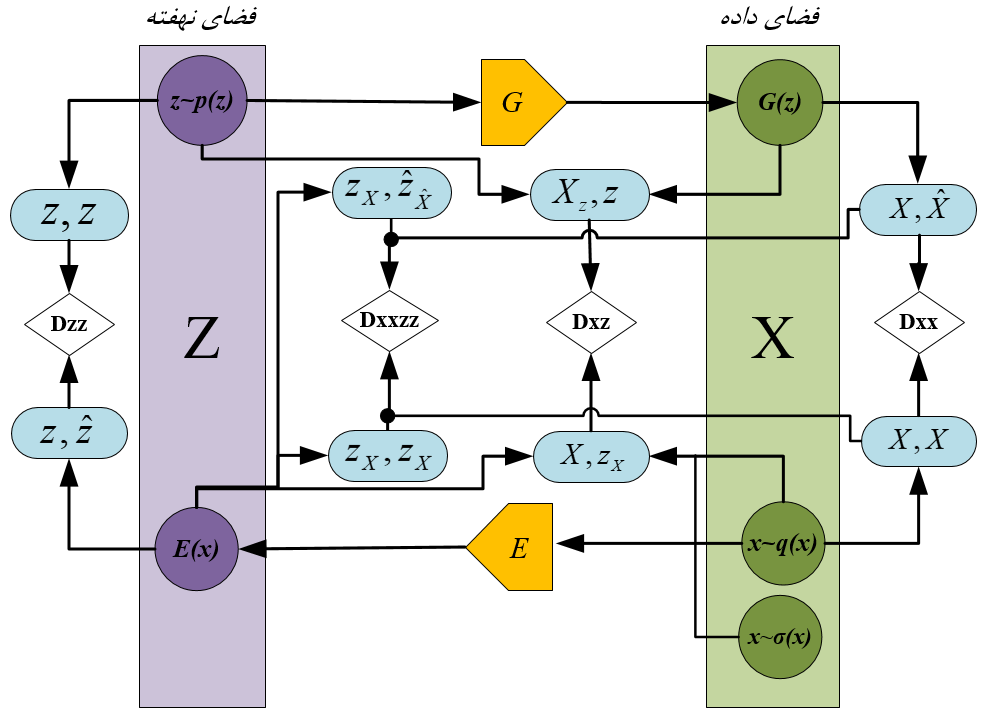
علت وقوع این پدیده نگاشت تنک از فضای داده ورودی به فضای نهفته است. زیرا در فاز آموزش کدگذار تنها نگاشت نمونه‌های هنجار به فضای نهفته را آموزش می‌بیند و در نتیجه فضای متناسب z برای نمونه‌های هنجار به خوبی مدل می‌شود ولی در فاز آزمون با توجه به اینکه مدل تاکنون بقیه فضا از جمله نمونه‌های ناهنجار را ندیده است ممکن است آن را به نقطه‌ای نامعلوم از فضای نهفته نگاشت کند. یعنی در این حالت هیچ اطلاعاتی در خصوص نگاشت برای داده های ناهنجار در دسترس نیست. یک راه­حل برای این موضوع استلزام نگاشت تمامی فضای ورودی به زیرفضای هنجار نهان می باشد. یعنی در اینجا برای پوشش هرچه بهتر فضای داده ورودی از توزیع نویز با نام σ(x) استفاده می‌شود و با نمونه‌گیری از این تابع و متمایل کردن شبکه به سمت تولید بازسازی کلاس داده هنجار شبکه یاد می‌گیرد تا برای طیف به نسبت گسترده‌تری از ورودی‌ها کلاس داده هنجار را بازسازی کند. در این صورت اگر داده ورودی ناهنجار هم باشد مدل آموزش دیده تا بازسازی نزدیک به کلاس داده‌هنجار تولید کند و در نتیجه میان داده ورودی و بازسازی آن فاصله مناسبی ایجاد می‌شود و همین فاصله معیار مناسبی برای تشخیص نمونه‌های ناهنجار خواهد بود. در شکل 2 با نحوه عملکرد این روال آموزشی آشنا می‌شوید.



شکل 2: تاثیر حضور توزیع σ(x) در روند آموزش مدل. در این شکل x بیانگر فضای داده ورودی و z بیانگر فضای داده ورودی است. نمونه‌ها توسط مولد از فضای داده ورودی به فضای نهفته نگاشت می‌شوند و وظیفه انجام نگاشت معکوس بر عهده کدگذار است. دایره‌های سبز رنگ نماد نمونه داده‌های هنجار و ضربدر‌های قکد نماد نمونه‌های ناهنجار هستند. بعلاوه آبی رنگ نشانگر نمونه‌های تولید شده توسط توزیع σ(x) هستند که در تنها مرحله آموزش مورد استفاده قرار گرفته‌اند. فلش فیروزه‌ای مقدار امتیاز ناهنجاری را نشان می‌دهد . همانطور که در شکل 3-5 مشاهده می‌شود در صورت عدم حضور σ(x) ( در سمت چپ شکل ) در روند آموزش، امتیاز ناهنجاری برای نمونه­های غیرعادی کمتر از زمانی است که از این توزیع استفاده شده است، در واقع در تصویر سمت راست، توزیع σ(x) مدل را به سمت بازسازی همه نمونه‌ها اعم از ناهنجار و هنجار به سمت توزیع داده‌‌های هنجار متمایل کرده است.

**4-3- مدل RCALAD**

در این بخش با ترکیب هر دو ایده مطرح شده در بخش های قبل یعنی بکارگیری متغیر جدید در تمایزگر و همچنین استفاده از توزیع σ(x) و افزودن آن­ها به مدل پایه [Alad]مدل اصلی پیشنهادی RCALAD معرفی می‌شود. در این شبکه به هر دو مسئله چرخه پایداری کامل و استلزام بازسازی ضعیف به طور همزمان پرداخته شده است و تلاش شده است تا یک چارچوب جامع، کاربردی و سازگار برای تمامی مسائل تشخیص ناهنجاری ارائه شود. شمای کلی مدل پیشنهادی در شکل 3 قابل مشاهده است.



شکل 3: ساختار کلی مدل RCALAD. در این ساختار کدگذار توام با شبکه مولد به منظور انجام عملیات نگاشت معکوس آموزش داده می‌شود. این کدگذار تنها از فضای داده ورودی نمونه می‌گیرد و به طور تقریبی بازنمایی متناسب با آن را در فضای نهفته را تولید می‌کند. در اینجا برای آموزش هم‌زمان هر دو شبکه مولد و کدگذار از یک شبکه تمایزگر توامان با نام استفاده شده است. این تمایزگر بررسی می‌کند که جفت متغیر ورودی متعلق به توزیع داده ورودی و نقطه متناظر با آن در فضای نهفته است و یا توسط شبکه مولد و نمونه‌گیری از فضای نهفته تولید شده است. به منظور ارضای شرط پایداری حلقه در فضای داده ورودی از تمایزگر افزوده شده است، این تمایزگر به صورت توام نمونه داده ورودی x و نمونه بازسازی شده متناظر آن را به عنوان ورودی دریافت می‌کند. همچنین برای تقویت شرط پایداری حلقه در فضای نهفته، تمایزگر به این مجموعه اضافه شده است. این تمایزگر شرط پایداری حلقه را در خلال روند تولید نمونه بازسازی شده چک می‌کند. ورودی این تمایزگر نمونه ورودی شبکه مولد از فضای نهفته و نمونه بازسازی متناظر با آن در فضای نهان است. با اضافه شدن تمایزگر Dxxzz به ساختار موجود، تلاش شده است تا از تمامی اطلاعات موجود در یک چرخه کامل به صورت توام استفاده شود. این تمایزگر تلاش می‌کند تا و بازسازی ارائه شده توسط شبکه و همینطور نگاشت تصویر ورودی در فضای نهان و بازسازی خروجی شبکه مولد توسط کدگذار تا حد امکان به یکدیگر نزدیک کند. بلوک σ(x) به منظور پوشش حداکثری فضای نهفته به ساختار فئق افزوده شده است.

همانطور که در شکل 3 مشاهده می‌شود، با هدف کاهش پیچیدگی زمانی، یک کدگذار توام با شبکه مولد در ساختار کلی شبکه عصبی تقابلی آموزش داده می‌شود. نگاشت معکوس از فضای داده ورودی به فضای نهفته به سادگی با تعبیه کدگذار در معماری پیشنهادی به‌دست می­آید. در اینجا برای آموزش هم‌زمان هر دو شبکه مولد و کدگذار از یک شبکه تمایزگر توامان با نام *Dxz* استفاده شده است. این تمایزگر بررسی می‌کند که جفت متغیر ورودی متعلق به توزیع داده ورودی و نقطه متناظر با آن در فضای نهفته است و یا توسط شبکه مولد و نمونه‌گیری از فضای نهفته z تولید شده است. به منظور ارضای شرط پایداری حلقه در فضای داده ورودی از تمایزگر *Dxx* و *Dzz* استفاده شده است تا هر نمونه و بازسازی متناظر با آن به طور مستقل بهبود یافته وشبیه شوند. تمایزگر *Dxxzz*  با هدف استفاده از تمامی اطلاعات موجود در یک چرخه کامل به صورت توام اضافه شده است. یعنی در کنار بررسی هر دو متغیر و بازسازی آن‌ها در همان فضا، توزیع توامان چهارتایی آن‌ها در روند تشخیص نمونه ناهنجار به کار گرفته شود تا شبکه به وضعیت داده ورودی در حین نگاشت­های متوالی دسترسی داشته باشد و اطلاعات بیشتری برای تمییز داده ها در دسترس داشته باشد. این شبکه وظیفه تمایز بین نمونه‌های چهارتایی  و  را دارد و تلاش می‌کند تا و بازسازی ارائه شده توسط شبکه و همینطور نگاشت تصویر ورودی در فضای نهان و بازسازی خروجی شبکه مولد توسط کدگذار تا حد امکان به طور وابسته و توامان به یکدیگر نزدیک کند. بلوک σ(x) به منظور پوشش حداکثری فضای نهفته به این مدل اضافه شده است. هدف از تعبیه این بلوک تولید نمونه‌های جدید در فضای داده ورودی و سپس نگاشت آن به فضای نهفته متناسب با داده هنجار است. در نهایت تابع هدف مدل پیشنهادی به صورت زیر است.

|  |  |
| --- | --- |
| **(7)** |  |

**4-4- تشخیص ناهنجاری**

هدف اصلی از ارائه مدل پیشنهادی در این مقاله، تشخیص ناهنجاری بر اساس بازسازی داده ورودی است. در این مدل هدف نهایی بازسازی دقیق و شبیه برای داده­های هنجار و بازسازی ضعیف و متفاوت برای نمونه ناهنجار است. یکی از عناصر کلیدی در تشخیص ناهنجاری، تعریف امتیاز ناهنجاری با هدف محاسبه فاصله میان نمونه ورودی و بازسازی ارائه شده توسط شبکه است..[4]

اولین امتیاز ناهنجاری ارائه شده در این مقاله، در این کار نام دارد. در این امتیاز برای محاسبه فاصله میان نمونه‌ها و بازسازی آن­ها، از فضای ویژگی موجود در تمایزگر *Dxxzz* استفاده می‌شود. به این منظور خروجی logit های لایه یکی مانده به آخر، به عنوان ویژگی استفاده می‌شوند. امتیاز ناهنجاری مورد استفاده به صورت زیر و با استفاده از خطای بازسازی نرم یک و مطابق معادله زیر تعریف می‌شود.

|  |  |
| --- | --- |
| **(8)** |  |

در این معادله f(⋅) بیانگرتابع فعالیت لایه یکی مانده به آخر در ساختار تمایزگر *Dxxzz* است. مفهوم بکارگرفته شده پشت تعریف این امتیاز، بکارگیری از میزان اطمینان تمایزگر از کیفیت بازسازی­های ارائه شده توسط شبکه است که اگر خوب انجام شده باشد در واقع نمونه متعلق به داده آموزش دیده شبکه یا همان توزیع داده هنجار است. بدین ترتیب هر چه مقدار این معیار بیشتر باشد، اختلاف بازسازی­ها بیشتر بوده و احتمال ناهنجاری بودن آن داده ورودی بیشتر است.

دومین امتیاز ارائه­شده در این مقاله با هدف بیشینه کردن بهره­گیری از اطلاعات موجود در مدل برای تشخیص ناهنجاری می­باشد. در این بخش معیار تعریف شده است. این امتیاز از جمع خروجی هر سه تمایزگر‌ ، و تشکیل شده است. در واقع چون در اینجا تمام تمایزگر‌های موجود در مدل پیشنهادی تنها روی نمونه‌های هنجار آموزش دیده‌اند و بازسازی برای تمامی فضای داده ورودی به سمت فضای داده هنجار متمایل شده است، پس انتظار می‌رود تصویر بازسازی شده نمونه ناهنجار و همچنین بازنمایی آن در فضای نهفته که توسط کدگذار تولید می‌شود، بسیار متفاوت از ورودی باشد و تمایزگرهای موجود در مدل به راحتی این ورودی‌های ناهنجار را شناسایی کنند. بیان ریاضی این معیار در معادله زیر آورده شده است.

|  |  |
| --- | --- |
| **(9)** |  |

حال مسئله قابل بررسی این موضوع می‌باشد که آیا معیار حاوی اطلاعات کافی برای تشخیص داده‌های هنجار از ناهنجار می‌باشد یا خیر. پاسخ به این سوال در حالت کلی بله می‌باشد زیرا این تمایزگرها در طی فرایند آموزش یاد می‌گیرند که به اختلاف میان دوتایی (x,x) و و همچنین دوتایی و توجه کنند یعنی هر چه از و یا از فاصله بگیرد، تشخیص آن برای تمایزگرها ساده­تر می­شود. در مدل پیشنهادی با افزودن توزیع σ(x) و متمایل سازی تمامی بازسازی به سمت توزیع داده‌های هنجار، اختلاف بازسازی برای داده‌های ناهنجار زیادتر هم شده و عملاً خروجی تمایزگرها یک معیار قابل اتکا برای تشخیص ناهنجاری می‌شود. در نهایت الگوریتم محاسبه امتیازهای ناهنجاری پیشنهادی مطابق الگوریتم زیر قابل مشاهده است.

|  |
| --- |
| Algorithm 2 Regularized Complete Adverarially Learned Anomaly Detection |
| Input where is the feature layer of  Output  1: procedure INFERENCE  2: Encode samples, Construct latent Embedding  3: Reconstruct samples  4: Reconstruct latent Embedding  5:  6:  7: return  8: end procedure |

الگوریتم 1: روند محاسبه­ی امتیاز ناهنجاری.

**5- آزمایش‌ها**

**5-1- دادگان**

برای سنجش عملکرد مدل پیشنهادی و بررسی کارایی آن از جنبه‌های مختلفاز مجموعه داده‌هایی با ویژگی متفاوت استفاده می‌شود. روش پیشنهادی روی مجموعه داده‌های تصویری و جدولی موجود آزمایش شده است. برای دادگان جدولی از چهار مجموعه داده kddcup99، arrhythmia، thyroid و musk استفاده شده است. دادگان kddcup99 دادگان مرتبط با نفوذ در شبکه است. arrhythmia که یک مجموعه پزشکی مرتبط با آریتمی قلبی با 16 کلاس است. thyroid یک دادگان سه کلاسی مرتبط با بیماری تیرویید است. دادگان musk با هدف دسته‌بندی شش کلاسی روی مشک مولکولی ایجاد شده است. در مجموعه داده‌های معرفی شده به ترتیب20، 15، 2.5 و 3.2 درصد از داده‌ها نمونه‌های ناهنجار هستند بنابراین در مرحله آزمون پس از محاسبه امتیاز ناهنجاری همین نسبت از دادگان که بیشترین مقدار امتیاز ناهنجاری را دارا می‌باشند به عنوان ناهنجاری دسته‌بندی می‌شوند. ارزیابی مدل پیشنهادی روی این مجموعه داده‌ها با محاسبه معیارهای F1، Recall و Precision صورت می‌پذیرد. برای مجموعه داده تصویری دو دادگان CIFAR10 و SVHN در نظر گرفته شده‌اند. هر دو این دادگان ده کلاسی هستند و هربار یک کلاس به عنوان کلاس هنجار و سایر 9 کلاس به عنوان کلاس ناهنجار در نظر گرفته می‌شود. معیار مورد استفاده برای ارزیابی مدل روی دادگان تصویری area under the receiver operating curve به اختصار AUROC است.

برای تمامی دادگان مورد استفاده 80 درصد دادگان به عنوان داده آموزشی و 20 درصد به عنوان داده آزمون انتخاب می‌شود. 25 درصد از داده آموزشی به عنوان داده ارزیابی (validation) انتخاب می‌شود. لازم به ذکر است در مرحله آموزش همه نمونه‌های ناهنجار از داده آموزشی حذف می‌شود.

**5-2- مدل‌های پایه**

در این قسمت چارچوب پیشنهادی با تعداد زیادی از روش‌های تشخیص ناهنجاری مقایسه می‌شود، در ادامه به طور مختصر مدل‌هایی که پیش از بررسی نشده، شرح داده می‌شود. در مدل OC-SVM یک مرز حول نمونه‌های هنجار یاد گرفته می­شود و نمونه‌هایی که خارج از این مرز قرارگیرند به عنوان ناهنجار در نظر گرفته می‌شوند. این مرز توسط روش ماشین بردار پشتیبان ایجاد می‌شود و هسته مورد استفاده در این روش RBF است.[19][12] روش IF یک روش از دسته روش‌های کلاسیک یادگیری ماشین است و به جای مدل کردن توزیع داده هنجار، داده ناهنجار را از سایر نمونه‌ها جدا می‌کند. در این روش ابتدا تعدادی ویژگی انتخاب می‌شود و در ادامه میانگین فاصله هر نمونه تا ریشه به عنوان امتیاز ناهنجاری در نظر گرفته می‌شود.[24] اساس کار مدل DSEBM بر استفاده از انرژی لایه‌هایی که در خودگذارهای حذف نویز به کار برده می‌شوند بنا نهاده شده است. از خطای بازسازی و همچنین خود انرژی به عنوان امتیاز ناهنجاری در این مدل استفاده شده است. DSEBM-r بیانگر تشخیص ناهنجاری با خطای بازسازی و DSEBM-e بیانگر تشخیص ناهنجاری با امتیاز انرژی است.[19]

مدل ALAD با استفاده از چارچوب شبکه‌های عصبی تقابلی که در (مقاله آلیس) معرفی شد و آموزش همزمان یک کدگذار با هدف دستیابی به نگاشت معکوس از فضای داده ورودی ، به تشخیص ناهنجاری پرداخته است. در این کار با استفاده از محاسبه خطای نگاشت ویژگی در تمایزگر ، امتیاز ناهنجاری محاسبه می‌شود.[4] در مدل DAGMM یک خودکدگذار و یک شبکه تمایزگر برای تولید فضای نهفته معقول و بازسازی ویژگی‌ها آموزش می‌دهد و در مرحله آزمون میزان درست­نمایی بازنمایی مدل در فضای نهفته و ویژگی‌های بازسازی شده توسط GMM محاسبه شده و به عنوان امتیاز ناهنجاری در نظر گرفته می‌شود.[25]اولین مدل مبتنی بر شبکه‌های عصبی تقابلی که با منظوز تشخیص ناهنجاری طراحی شد مدل AnoGAN است. در این مدل بعد از آموزش DCGAN روی داده‌های هنجار، با استفاده از بهینه کردن جمع وزن‌دار خطای بازسازی به وسیله گرادیان نزولی نمونه‌های آزمایشی به فضای نهفته برده می‌شوند.[8]در روش DSVDD یک شبکه عصبی که حجم ابر کره محیط بر داده هنجار را کمینه می‌کند آموزش داده می‌شود. امتیاز ناهنجاری در این مدل فاصله اقلیدسی میان مرکز این ابرکره تا داده ورودی است.[13]مدل DCAE یک مدل کلاسیک خودکدگذار است که در آن کدگذار و کدگشا دارای ساختار کانولوشنی هستند. امتیاز ناهنجاری در این مدل نرم دو خطای بازسازی است.[26]

**5-3- آزمایش‌ها روی دادگان جدولی**

نتایج ارزیابی مدل پیشنهادی RCALAD روی دادگان جدولی kddcup99، arrhythmia، thyroid و musk در جدول 1 خلاصه شده است. به جز مدل ALAD که بر روی هر مجموعه داده اجرا و نتایجش گزارش شده، برای سایر مدل‌ها از نتایج موجود **در [alad v deep autoencoding guassian mixture model for unsupervised anomaly detection]** استفاده شده است. ساختار‌های مورد استفاده در شبکه مولد، تمایزگر و کدگذار همگی لایه‌های کاملا متصل استاندارد با توابع فعالساز غیرخطی هستند. لازم به ذکر است در این مرحله از توزیع N(0,I)به عنوان σ(x) استفاده می‌شود.

برای مقایسه شفاف میان مدل‌های مختلف از error bar در ردیف آخر جدول 1 استفاده شده است. همانطور که در این جدول قابل مشاهده است مدل پیشنهادی روی دادگان arrhythmia و musk نسبت به سایر مدل‌ها بسیار موفق عمل کرده است. روی دادگان KDD بر اساس معیار F1 بهترین مدل است ولی روی دادگان thyroid با توجه به عملکرد کم‌نظیر مدل IF در رتبه دوم قرار می‌گیرد.

جدول 4-1 اینجا باید باشه

**5-4- آزمایش‌ها روی دادگان تصویری**

در این قسمت عملکرد مدل پیشنهادی روی دادگان تصویری در دو جدول مستقل از هم بررسی می‌شود. همانطور که در جدول 2 و 3 قابل مشاهده است، مدل پیشنهادی روی دادگان CIFAR10 بهبود قابل توجهی ایجاد کرده است و روی دادگان SVHN با اختلاف کم نسبت به مدل پایه ALAD بهترین عملکرد را داشته است.

جدول مربوط به دادگان تصاویر اینجا بیاد

**5-5- مطالعات فرسایشی**

در این قسمت کارکرد هر یک از اجزای اضافه شده به مدل پایه را روی دادگان cifar10 و SVHN بررسی می‌کنیم. آزمایش‌ها در این قسمت در زمان حضور و عدم حضور تمایزگر و همچنین حضور و عدم حضور σ(x) تکرار می‌شود.

جدول تصاویر مطالعات فرسایشی

همانطور که در جدول مشخص است اضافه کردن تمایزگر بهبود قابل توجهی روی دادگان SVHN ایجاد نکرده است ولی روی دادگان cifar10 دقت را تا اندازه خوبی بهبود داده است. توزیع σ(x) روی دادگان svhn معیار AUROC را به میزان کمی کاهش داده است. این می‌تواند بدین دلیل باشد که در برخی نمونه‌های این مدل با اعمال σ(x) فاصله میان داده ناهنجار و بازسازی آن به جای آن که زیاد شود کم شده باشد، به بیان دیگر در زمان عدم حضور این توزیع شبکه مولد بازنمایی به نسبت دورتری از داده ورودی ایجاد می‌کرده است و با اعمال توزیع σ(x) بازسازی داده ناهنجار به خود آن نزدیک تر شده است. شهود این استدلال در شکل فلان قابل مشاهده است. توزیع σ(x) روی دادگان CIFAR10 عملکرد مناسبی داشته و معیار AUROC را بهبود داده است.

**5-6- ارزیابی کفایت تمایزگر**

با اضافه کردن تمایزگر Dxxzz آیا نیازی به تمایزگرDxx و Dzz است یا خیر؟ برای پاسخگویی مناسب به این سوال باید جدول زیر را مشاهده کرد.

شکل کفایت اینجا

مطابق نتایج تئوری، افزودن تمایزگر Dxxzz به چارچوب کلی و در کنار دیگر تمایزگرها بالاترین کارایی را داشته است. پس از آن حذف Dxx ضربه کمتری به مدل می¬زند زیرا بخشی از اطلاعاتی که استخراج می¬کند، توسط تمایزگر Dxxzz پوشش داده می¬شود. اما با توجه به اینکه Dzz در یک چرخه مستقل میزان شباهت z و بازسازی آن را مورد بررسی قرار می‌دهد طبیعی است که حذف آن میزان دقت را کاهش دهد. همان¬طور که نتایج نشان می¬دهد، نتیجه این بخش این است که تمایزگر Dxxzz به تنهایی کافی نیست و این سه تمایزگر در کنار هم بیشترین کارایی را دارند و تمایزگر Dxxzz به تنهایی نمی¬تواند تمامی جنبه¬ها را دیده و اطلاعات مورد نیاز را استخراج کند.

**5-7- تشخیص ناهنجاری**

در این قسمت امتیاز ناهنجاری معرفی شده در این مقاله مورد ارزیابی قرار گرفته و با دیگر امتیازهای ناهنجاری مقایسه می‌شود. در این‌جا خروجی خام تمایزگر‌ها با عنوان لاجیت نامگذاری می‌شود، همچنین خروجی لایه پنهان قبل از لایه لاجیت، ویژگی نامیده می‌شود. در اینجا برای محاسبه امتیاز ناهنجاری از متغیرهای موجود در ساختار تمایزگر استفاده می‌شود. بیان ریاضی امتیاز ناهنجاری پیشنهادی در این مقاله و همچنین سایر امتیازهای ناهنجاری مورد مقایسه در معادله فلان آورده شده است.

|  |  |
| --- | --- |
| **(10)** | Logits  Features |

جدول روی دادگان جدولی

همانطور که در جدول بالا مشاهده می‌شود روی دادگان جدولی خروجی‌ تمایزگر دارای بهترین نتایج به نسبت سایر امتیاز‌های ناهنجاری است. با توجه به اینکه تعداد ویژگی‌ها روی دادگان جدولی به نسبت دادگان تصویری کمتر است تمایزگر قادر به تشخیص مناسب نمونه‌های ناهنجار است. نتایج روی دادگان تصویری در جدول زیر قابل مشاهده است.

جدول دادگان تصویری

همانطور که در جدول 4-8 مشخص است عملکرد امتیاز مبتنی بر ویژگی‌ها روی دادگان تصویری بسیار مناسب است، این می‌تواند به این دلیل باشد بردار ویژگی‌ها برای هر عکس نسبت به بردار ویژگی‌های موجود در دادگان جدولی بزرگتر است و استفاده از پارامترهای بیشتر به منظور تشخیص ناهنجاری سبب بهبود عملکرد امتیاز مورد نظر شده است.

**6- جمع‌بندی**

در این کار روش مبتنی بر شبکه‌های عصبی تقابلی به منظور تشخیص ناهنجاری معرفی شد. در مدل پیشنهادی از یک کدگذار برای نگاشت معکوس از فضای داده ورودی استفاده می‌شود و برای ارضای شرط چرخه پایداری از تمایزگر *Dxx* کمک گرفته می‌شود. به منظور پایدارسازی روند آموزش شبکه مولد تقابلی تمایزگر *Dzz* در ساختار تقابلی مدل بهره گرفته می‌شود. با هدف استفاده از اطلاعات یک چرخه کامل در مدل پیشنهادی متغیر معرفی و در نتیجه تمایزگر *Dxxzz* در مدل پیشنهادی جای داده شد. علاوه بر این برای متمایل کردن شبکه به سمت توزیع داده نرمال توزیع σ(x) استفاده می‌شود. نتایج حاصل از آزمایش‌ها بیانگر اثربخشی مدل پیشنهادی در زمینه تشخیص ناهنجاری و همچنین برتری آن نسبت به سایر مدل‌های state of the art روی دادگان جدولی و تصویری است. استفاده از چرخه پایداری کامل در سایر مدل‌های مبتنی بر بازسازی و همچنین پوشش فضای داده وروی این مدل‌ها به وسیله توزیع σ(x) به عنوان کارهای آتی این مقاله در نظر گرفته شده‌اند.

**مراجع**

[1] X. Shu, L. Cheng, and S. J. Stolfo, “Anomaly Detection as a Service.”

[2] C. Jiang, J. Song, G. Liu, L. Zheng, and W. Luan, “Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism,” *IEEE Internet Things J.*, vol. 5, no. 5, pp. 3637–3647, 2018, doi: 10.1109/JIOT.2018.2816007.

[3] X. Dai and M. Bikdash, “Distance-based outliers method for detecting disease outbreaks using social media,” *Conf. Proc. - IEEE SOUTHEASTCON*, vol. 2016-July, 2016, doi: 10.1109/SECON.2016.7506752.

[4] H. Zenati, M. Romain, C. S. Foo, B. Lecouat, and V. Chandrasekhar, “Adversarially Learned Anomaly Detection,” *Proc. - IEEE Int. Conf. Data Mining, ICDM*, vol. 2018-Novem, pp. 727–736, 2018, doi: 10.1109/ICDM.2018.00088.

[5] I. Goodfellow *et al.*, “Generative Adversarial Nets,” in *Advances in Neural Information Processing Systems*, Oct. 2014, pp. 2672–2680, doi: 10.1109/ICCVW.2019.00369.

[6] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” *4th Int. Conf. Learn. Represent. ICLR 2016 - Conf. Track Proc.*, pp. 1–16, 2016.

[7] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, “Generative Adversarial Networks: An Overview,” *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, 2018, doi: 10.1109/MSP.2017.2765202.

[8] T. Schlegl, P. Seeb, S. Waldstein, U. Schmidt-Erfurth, and G. Langs, “Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery,” *Int. Confrence Inf. Process. Med. Imaging*, vol. 2, pp. 146–157, 2017, doi: 10.1007/978-3-319-59050-9.

[9] R. Kaur and S. Singh, “A survey of data mining and social network analysis based anomaly detection techniques,” *Egypt. Informatics J.*, vol. 17, no. 2, pp. 199–216, 2016, doi: 10.1016/j.eij.2015.11.004.

[10] A. Zimek, E. Schubert, and H. Kriegel, “REVIEW A Survey on Unsupervised Outlier Detection in High-Dimensional Numerical Data,” *Signal Processing*, vol. 99, pp. 215–249, 2012, doi: 10.1002/sam.

[11] M. A. F. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, “A review of novelty detection,” *Signal Processing*, vol. 99, pp. 215–249, 2014, doi: 10.1016/j.sigpro.2013.12.026.

[12] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Piatt, “Support vector method for novelty detection,” *Adv. Neural Inf. Process. Syst.*, no. January, pp. 582–588, 2000.

[13] L. Ruff *et al.*, “Deep one-class classification,” *35th Int. Conf. Mach. Learn. ICML 2018*, vol. 10, pp. 6981–6996, 2018.

[14] S. Liang, Y. Li, and R. Srikant, “Enhancing the reliability of out-of-distribution image detection in neural networks,” *6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc.*, pp. 1–15, 2018.

[15] I. Golan and R. El-Yaniv, “Deep anomaly detection using geometric transformations,” *Adv. Neural Inf. Process. Syst.*, vol. 2018-Decem, no. NeurIPS, pp. 9758–9769, 2018.

[16] Z. Yang, I. S. Bozchalooi, and E. Darve, “Regularized Cycle Consistent Generative Adversarial Network for Anomaly Detection.”

[17] D. T. Nguyen, Z. Lou, M. Klar, and T. Brox, “Anomaly detection with multiple-hypotheses predictions,” *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 8418–8432, 2019.

[18] S. Pidhorskyi, R. Almohsen, D. A. Adjeroh, and G. Doretto, “Generative probabilistic novelty detection with adversarial autoencoders,” *Adv. Neural Inf. Process. Syst.*, vol. 2018-Decem, no. Nips, pp. 6822–6833, 2018.

[19] S. Zhai, Y. Cheng, W. Lu, and Z. Zhang, “Deep structured energy based models for anomaly detection,” *33rd Int. Conf. Mach. Learn. ICML 2016*, vol. 3, pp. 1742–1751, 2016.

[20] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, “f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks,” *Med. Image Anal.*, vol. 54, pp. 30–44, 2019, doi: 10.1016/j.media.2019.01.010.

[21] H. Zenati, C. S. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar, “Efficient GAN-Based Anomaly Detection,” 2018, [Online]. Available: http://arxiv.org/abs/1802.06222.

[22] V. Dumoulin *et al.*, “Adversarially learned inference,” *5th Int. Conf. Learn. Represent. ICLR 2017 - Conf. Track Proc.*, pp. 1–18, 2017.

[23] C. Li *et al.*, “ALICE : Towards Understanding Adversarial Learning for Joint Distribution Matching arXiv : 1709 . 01215v2 [ stat . ML ] 5 Nov 2017,” no. Nips, pp. 1–22, 2017.

[24] F. Tony Liu, K. Ming Ting, and Z.-H. Zhou, “Isolation Forest ICDM08,” *Icdm*, 2008, [Online]. Available: https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf%0Ahttps://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf?q=isolation-forest.

[25] B. Zong *et al.*, “Deep autoencoding Gaussian mixture model for unsupervised anomaly detection,” *6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc.*, pp. 1–19, 2018.

[26] A. Makhzani and B. Frey, “Winner-take-all autoencoders,” *Adv. Neural Inf. Process. Syst.*, vol. 2015-Janua, pp. 2791–2799, 2015.

1. Lipschitz Constraints [↑](#footnote-ref-1)