افزایش داده و انتخاب مؤثر ویژگی در شبکه‌های مولد متخاصمی جهت تشخیص احساس از گفتار

آرش شیلاندری[۱،*]، حسین مروی[۲] و حسین خسروی[۳]

اطلاعات مقاله

نوع مقاله: پژوهشی
دریافت مقاله: ۱۴۰۰/۰۸/۰۳
بازنگری مقاله: ۱۴۰۰/۱۰/۲۸
پذیرش مقاله: ۱۴۰۰/۱۱/۰۴

واژگان کلیدی:
پردازش گفتار،
انتخاب ویژگی،
افزایش داده،
تشخیص احساس از گفتار،
شبکه‌های مولد متخاصمی.

چکیده

تاکنون، یقینی مبتنی بر موفقیت و یا عدم موفقیت به‌کارگیری روش‌های کاهش ویژگی جهت افزایش کارایی سیستم‌های تشخیص احساس از گفتار حاصل نشده است. این مقاله با هدف افزایش داده‌ها در یک سیستم تشخیص احساس از گفتار، انتخاب ویژگی را مورد بحث و بررسی قرار می‌دهد. آزمایش‌ها بر روی چهار پایگاه‌داده متداول EMO-DB، eNTERFACE05، SAVEE و IEMOCAP در نرم‌افزار پایتون انجام گردیده و علاوه بر این، تجزیه و تحلیل داده‌ها بر روی هر چهار پایگاه‌داده برای چهار احساس غم، عصبانیت، خوشحالی و خنثی ارائه خواهد شد. یک شبکه افزایش داده متخاصمی جهت افزایش نمونه‌ها و دو شبکه انتخاب ویژگی ترکیبی معیار فیشر و الگوریتم جداساز خطی طی دو مرحله و با فیدبکی که از شبکه طبقه‌بند گرفته می‌شود، سیستم تشخیص احساس از گفتار را به نقطه بهینه‌ای از تعداد و ابعاد داده‌ها رسانیده و نشان می‌دهد آنالیز مؤلفه‌های اصلی روی داده‌های همبسته، مؤثرتر و الگوریتم جداساز خطی روی داده‌های با بعد کم بهتر عمل می‌کنند. همچنان که روش فیشر در کاهش سایز بهتر از آنالیز مؤلفه‌های اصلی عمل می‌کند. همچنین ماشین بردار پشتیبان جهت طبقه‌بندی احساسات مورد استفاده قرار گرفته است. نتایج نشان می‌دهد که استفاده از هر دو روش جداساز خطی و معیار فیشر به طور هم‌زمان در سیستم افزایش داده متخاصمی می‌تواند ویژگی‌ها را در ابعاد کمتر فیلتر نموده درحالی‌که اطلاعات احساسی را جهت طبقه‌بندی حفظ نماید. نتایج به‌دست‌آمده با تحقیقات اخیر مقایسه گردیده است که حاکی از دستیابی روش پیشنهادی به صحت ۸۶٫۳۲٪ در پایگاه‌داده برلین می‌باشد.

۱- مقدمه

تشخیص احساس از گفتار تاکنون به دلیل نبود پایگاه‌داده با داده‌های فراوان و در دسترس با مسائل فراوانی روبه‌رو بوده است. عدم توانایی در انتخاب ویژگی‌های مهم و تأثیرگذار در شناسایی احساس، امکان استفاده از این سیستم‌ها را در کاربردهای برخط محدود می‌کند. همچنین وابسته بودن این سیستم به زبان، لهجه، سن، جنسیت و نوع حالت گوینده از عمده مسائل این روش بوده است.

روش‌های انتخاب ویژگی نقش مهمی در عملکرد بهینه مدل‌های یادگیری ماشین دارند. اگر ویژگی‌ها به‌درستی جهت آموزش شبکه عصبی انتخاب شوند این شبکه‌ها می توانند در زمانی کوتاه آموزش‌دیده و به‌درستی پاسخ دهند. تشخیص احساسات از گفتار شامل استخراج ویژگی و تجزیه و تحلیل بردارهای ویژگی استخراج شده از سیگنال گفتار می‌شود. سیستم تشخیص احساس از سیگنال گفتار از دیدگاه تشخیص الگو شامل پنج بخش می‌باشد:

* پست الکترونیک نویسنده مسئول: Shilandari@shahroodut.ac.ir
۱. دانشجوی دکتری تخصصی، دانشکده مهندسی برق، دانشگاه شاهرود
۲. دانشیار، دانشکده مهندسی برق، دانشگاه شاهرود
۳. دانشیار، دانشکده مهندسی برق، دانشگاه شاهرود

https://doi.org/10.22075/jme.2022.24865.2159

۱- پیش‌پردازش ۲- استخراج ویژگی ۳- افزایش داده ۴- انتخاب ویژگی ۵- طبقه‌بندی. تعداد ویژگی‌های استخراج شده بسته به تعیین پارامترهای صوتی در نظر گرفته شده داشته و تغییرات آماری این پارامترها منجر به استخراج بردارهای ویژگی با ابعاد بالا می‌شوند. از این رو، از روش‌های انتخاب ویژگی جهت کاهش ابعاد بردارهای ویژگی استفاده می‌شود. همچنین ویژگی‌های نامرتبط یا تا حدودی مرتبط می‌توانند تأثیر منفی بر عملکرد سیستم داشته باشند. هرچند که معمولاً تعداد نمونه‌ها در پایگاه‌های داده احساسی اندک است ولی ابعاد بردارهای ویژگی استخراج شده در مقایسه با تعداد آنها بالاست. همه این ویژگی‌ها جهت شناخت احساسات مؤثر نیستند و از طرفی احساسات مختلف می‌توانند ویژگی‌های مختلف گفتار را تحت‌تأثیر قرار دهند[۱].

پیاده‌سازی روش‌های افزایش داده و انتخاب ویژگی و پاک‌سازی داده‌ها[1]، اولین و مهم‌ترین گام در طراحی مدل‌های هوشمند یادگیری قلمداد می‌شوند. همچنین، هنگامی که ابعاد فضای ویژگی داده‌ها در مقایسه با تعداد نمونه‌ها بسیار زیاد است، با معضل نفرین ابعاد بالا[2] مواجه خواهیم شد، استفاده از مجموعه ویژگی‌های مناسب، هزینه‌های محاسباتی (نظیر زمان لازم برای آموزش سیستم و همچنین استفاده از آنها در کاربردهای برخط) لازم برای آموزش بهینه سیستم را به‌شدت کاهش می‌دهد. محاسبه درجه اهمیت ویژگی‌ها و استفاده از آن‌ها در مرحله انتخاب ویژگی، گام مهمی در جهت تفسیرپذیری[3] مدل‌های یادگیری ماشین بخصوص پس از افزایش داده خواهد بود [۱].

استفاده از تعداد ویژگی‌های زیاد به‌منظور طراحی سیستم تشخیص الگو باعث افت عملکرد الگوریتم طبقه‌بندی به دلیل تعدد پارامترهای تنظیم، پیچیدگی محاسباتی و آموزش ناکارآمد به دلیل ابعاد بالای مسئله می‌شود. تعداد بهینه ویژگی‌ها را می‌توان به‌عنوان تابعی از تعداد نمونه‌های پایگاه‌داده و همبستگی بین ویژگی‌ها بیان نمود. اگر تعداد ویژگی‌ها بیش از تعداد بهینه انتخاب شود، با پدیده نفرین ابعاد داده و یا بیش برازش مواجه شده و نرخ تشخیص کاهش پیدا می‌کند؛ بنابراین استفاده از الگوریتم‌های انتخاب ویژگی به‌منظور حذف ویژگی‌های اضافی و غیرمرتبط (نویزی) امری ضروری می‌باشد. ویژگی‌های اضافی ویژگی‌هایی هستند که اضافه‌شدن آن‌ها به بردارهای ویژگی پایگاه‌داده، اطلاعات جدیدی به مسئله اضافه نکند و بردارهای ویژگی غیرمرتبط نمونه‌هایی هستند که در طبقه‌بندی اهمیتی نداشته و حاوی اطلاعات مفیدی از داده‌های آن طبقه نباشند[۲].

به منظور انتخاب یک بردار ویژگی بهینه، باید همه زیرمجموعه‌های ممکن را بررسی نمود. اما از آن جا که این کار امری بسیار زمان‌بر و غیرعملی می‌باشد، ناگزیر به‌منظور انتخاب یک بردار ویژگی مناسب از الگوریتم‌های جستجو استفاده می‌شود. همچنین، استفاده از الگوریتم‌های کاهش ویژگی[4] که با استفاده از یک تبدیل مناسب ویژگی را از فضای اصلی به فضایی با ابعاد پایین‌تر منتقل می‌کنند نیز متداول است [۱].

نکته جالب توجه دیگر، پدیدهٔ اوج[5] می‌باشد که تعداد بهینه ویژگی‌ها را به‌صورت تابعی از تعداد نمونه‌ها و همبستگی بین ویژگی‌ها بیان می‌کند [۳]. با افزایش تعداد ویژگی‌های مورد استفاده، ابتدا نرخ تشخیص افزایش پیدا می‌کند اما پس از رسیدن به نقطه اوج، افزایش بیشتر تعداد ویژگی‌ها موجب کاهش نرخ تشخیص می‌شود.

هدف اصلی انتخاب ویژگی، تعیین خصوصیاتی است که بهترین طبقه‌بندی را منجر می‌شوند. به نظر می‌رسد با انتخاب ویژگی‌های بهینه، سایز مجموعه ویژگی‌ها کاهش می‌یابد و عملکرد و دقت طبقه‌بندی افزایش می‌یابد. اگرچه تاکنون تعداد زیادی از روش‌های انتخاب ویژگی برای مطالعات تشخیص احساس از گفتار[6] استفاده شده است ولی برخی از این روش‌ها منجر به کاهش دقت در تشخیص احساس می‌شوند چرا که این روش‌ها کلی بوده و مخصوص استفاده در سیگنال‌های صوت نمی‌باشند.

برای انتخاب بهترین ویژگی‌ها برای یک مدل یادگیری نظارت شده، روش‌های انتخاب ویژگی نظارت شده[7] ارائه شده‌اند. هدف این دسته از الگوریتم‌ها، انتخاب بهترین زیرمجموعه از ویژگی‌ها برای تضمین عملکرد بهینه یک مدل نظارت شده است. این الگوریتم‌ها برای انتخاب بهترین ویژگی‌ها، از داده‌های برچسب زده شده[8] استفاده می‌کنند.

[1] Data Cleaning
[2] Curse of Dimensionality
[3] Interpretability
[4] Feature Reduction
[5] Peaking Phenomenon
[6] Speech Emotion Recognition (SER)
[7] Supervised Feature Selection
[8] Labelled Data

بااین‌حال، در شرایطی که داده‌های برچسب زده در دسترس نیستند (یادگیری نظارت نشده)، روش‌هایی به نام روش‌های انتخاب ویژگی نظارت نشده[1] پیاده‌سازی شده‌اند که ویژگی‌ها را بر اساس معیارهای مختلفی نظیر واریانس، آنتروپی، قابلیت ویژگی‌ها در حفظ اطلاعات مرتبط با مشابهت‌های محلی[2] و سایر موارد امتیازبندی می‌کنند.
ویژگی‌های مرتبطی که از طریق فرایندهای مکاشفه‌ای نظارت نشده[3] شناسایی شده‌اند، می‌توانند در مدل‌های یادگیری نظارت شده نیز مورداستفاده قرار بگیرند. چنین کاربردهایی از ویژگی‌های شناسایی شده، به سیستم یادگیری نظارت شده اجازه می‌دهد تا علاوه بر شناسایی میزان همبستگی ویژگی‌ها با برچسب طبقه داده‌ها، الگوهای دیگری نیز در داده‌های یادگیری شناسایی کنند.
از دیدگاه طبقه‌بندی، روش‌های انتخاب ویژگی را می‌توان در چهار دسته طبقه‌بندی کرد: ۱- روش‌های فیلتر[4]- روش‌های بسته‌بند[5]- روش‌های تعبیه شده[6]- روش‌های ترکیبی[7]. گزینه دیگر برای انتخاب بهترین ویژگی‌ها، ترکیب روش‌های فیلتر و بسته‌بند است. در این مقاله از یک فرایند دومرحله‌ای برای ترکیب دو روش فیلتر و بسته‌بند استفاده شده است. در مرحله اول ویژگی‌ها بر اساس مشخصه‌های آماری فیلتر می‌شوند. در مرحله بعد، با استفاده از یک روش انتخاب ویژگی بسته‌بند، بهترین ویژگی‌ها برای آموزش یک مدل یادگیری انتخاب می‌شوند و به طور مشخص از ترکیب معیار فیشر به‌عنوان روش انتخاب ویژگی فیلتر و از الگوریتم جداساز خطی به‌عنوان مکمل جهت انتخاب ویژگی‌ها استفاده شده است.
در الگوریتم‌های انتخاب ویژگی بر پایه فیلتر، بسته‌بند و تعبیه شده، با حذف ویژگی‌های اضافی و غیرمرتبط یا انتخاب ویژگی‌های مؤثر، یک زیرمجموعه بهینه از ویژگی‌ها به دست می‌آید بطوریکه ویژگی‌های انتخابی بدون هیچ‌گونه تغییری از بین ویژگی‌های اصلی انتخاب می‌شوند. در مقابل این روش‌ها، الگوریتم‌های کاهش ویژگی نظیر الگوریتم جداسازی خطی[8] و آنالیز مؤلفه‌های اصلی[9] نیز وجود دارند که با اعمال یک تبدیل بر روی همهٔ ویژگی‌ها آن‌ها را به یک فضای جدید با ابعاد کمتر انتقال می‌دهند[۴]. در این روش‌ها، هر یک از ویژگی‌ها که در فضای جدید بیان می‌شوند شاید از نظر فیزیکی مفهوم خاصی نداشته باشند اما حاوی چکیده اطلاعات ویژگی‌های اصلی می‌باشند. بدین منظور برای حل یک مسئله با C ویژگی، ابعاد آن به C-۱ ویژگی (فضای C-۱ بعدی) کاهش می‌یابد.
تا کنون روش‌های مؤثر و متعددی ارائه شده است که با معرفی ویژگی‌های جدید و کارا دقت سیستم‌های بازشناسی احساس را افزایش می‌دهند. در این مقاله، نرم‌افزار openSMILE[۵] جهت استخراج ویژگی‌ها استفاده گردیده و با روش‌های افزایش داده متخاصمی سعی در تولید بردارهای ویژگی جدید جهت افزایش نمونه‌های آموزش و متعادل نمودن پایگاه‌های داده شده است. همچنین، با استفاده از روش‌های انتخاب ویژگی، ویژگی‌های کمتر اثرگذار بر روی کارایی سیستم تشخیص احساس از گفتار حذف و ویژگی‌های مؤثر معرفی گردیدند. کمبود داده‌ها می‌تواند باعث شود که یک مدل یادگیری ماشین قادر به یادگیری توزیع واقعی داده‌ها نباشد و در نهایت منجر به مشکل بیش‌برازش می‌شود. به‌عنوان‌مثال، در هنگام آموزش یک شبکه عصبی عمیق با نمونه‌های آموزشی کم در هر طبقه، بیش‌برازش رخ می‌دهد، درصورتی که هر نمونه هزاران ویژگی دارد. برای حل این مشکل، می‌توان از قاعده‌مندسازی[10] برای ایجاد محدودیت در مدل استفاده کرد [۶]. راه‌حل کلی دیگر، کاهش ابعاد با محدودیت پراکندگی است [۷]. این رویکرد زمانی مؤثر خواهد بود که ویژگی‌های اضافی وجود داشته باشد. در غیر این صورت، اطلاعات مفید را از بین می‌برد و منجر به تخریب عملکرد می‌شود. همچنین، وجود بردارهای ویژگی کم‌اهمیت در آموزش شبکه عصبی، محاسبات را در این شبکه پیچیده و تأثیرگذاری نمونه‌ها را در آموزش بی‌اثر خواهد نمود؛ لذا انتخاب و حذف بردارهای ویژگی کم اهمیت موضوعی است که کمتر به آن در روش‌های افزایش داده به طور هم‌زمان پرداخته شده است. یکی از روش‌های مؤثر برای تقویت و افزایش داده‌ها، استفاده از شبکه‌های مولد متخاصمی است که توسط گودفلو[11] و همکاران در سال ۲۰۱۴ معرفی شده است [۸]. در سال‌های اخیر،

[1] Unsupervised Feature Selection
[2] Local Similarity
[3] Unsupervised Heuristics
[4] Filter
[5] Wrapper
[6] Embedded
[7] Hybrid
[8] Linear Discriminant Analysis
[9] Principal Component Analysis
[10] Regularization
[11] Ian Goodfellow

شبکه‌های مولد متخاصمی به‌عنوان یکی از موفق‌ترین رویکردها برای تولید نمونه شناخته شده‌اند [۳۵]. با استفاده از یک بازی مخالف بین یک شبکه تشخیص‌دهنده و یک شبکه مولد، شبکه‌های مولد متخاصمی آموزش می‌بینند تا نمونه‌هایی تولید کنند که از داده‌های واقعی قابل تشخیص نیستند. علاوه بر این، آنها دارای سه خصوصیت عمده هستند [۸]:۱- شبکه‌های مولد متخاصمی می‌توانند توزیع احتمال در مسائل پیچیده دنیای واقعی را یاد بگیرند.
۲- شبکه‌های مولد متخاصمی می‌توانند با داده‌های ازدست‌رفته (داده‌های نویزی) آموزش داده شوند، حتی زمانی که برچسب‌های بسیاری از نمونه‌ها وجود ندارد و ۳- شبکه‌های مولد متخاصمی دارای خروجی‌های چند مدول هستند به این معنی که آنها می‌توانند چندین جواب صحیح متفاوت تولید کنند و تنوع نمونه‌های تولید شده را افزایش دهند ولی همچنان توجه به انتخاب ویژگی‌های ساخته شده و فیلتر نمودن نمونه‌ها امری ضروری است که در این مقاله به آن پرداخته شده است.

برخی مقالات از روش‌های GMM و HMM استفاده می‌کنند تا توزیع ویژگی‌های آکوستیکی را یاد بگیرند و سپس از طبقه‌بندی بیزین یا اصل ماکزیمم بخت[1] برای تشخیص احساس استفاده می‌کنند. برخی از مدل‌های پس‌زمینه جامع[2] استفاده می‌کنند تا از ویژگی‌های آکوستیکی، بردار ویژگی برای آموزش شبکه ماشین بردار پشتیبان بسازند که این روش بیشتر برای تشخیص گوینده استفاده می‌شود. برخی، روش‌های آماری را روی ویژگی‌های سطح پایین استفاده می‌کنند تا ویژگی‌های آماری سطح بالا را به دست آورند و سپس از ماشین بردار پشتیبان برای طبقه‌بندی ویژگی‌های کلی استفاده می‌شود. برخی از روش‌های KNN و یا درخت تصمیم‌گیری[3] استفاده می‌کنند که نیاز به تعداد ویژگی‌های زیاد و ویژگی‌های ساخته شده به‌صورت دستی دارد.

این مقاله به کاربرد شبکه‌های مولد متخاصمی برای تولید داده‌های مصنوعی و انتخاب بهترین آنها پرداخته است که هدف از آن تولید داده‌هایی است که توزیع داده‌های واقعی را گسترش می‌دهند تا عملکرد سیستم تشخیص احساس از گفتار را بهبود بخشند. به بیان دیگر، فرایند انتخاب ویژگی در داده‌های تولید شده سبب می‌شود که آنها هدفمند به داده‌های اصلی اضافه گردند و سبب آموزش هر چه بهتر شبکه طبقه‌بند احساس شوند و حجم محاسبات را کاهش و کارایی این شبکه را بیفزایند.

یک مشکل اساسی در آموزش شبکه‌های مولد متخاصمی، انتخاب ویژگی‌های مؤثر در سیستم تشخیص احساس از گفتار است. اگر ویژگی‌ها به‌صورت هوشمندانه افزوده نشوند، حجم محاسبات بالا رفته و سیستم تشخیص احساس کند و در استفاده‌های برخط غیر قابل استفاده خواهد شد. همچنین ممکن است بردارهای ویژگی‌های افزوده شده، مرز بین دو طبقه را از بین برده و یا داده‌های مرزی تولید شوند که کار آموزش شبکه طبقه‌بند را مشکل و یا راندمان طبقه‌بندی را پایین آورند. از این رو توجه به انتخاب ویژگی همیشه لازم خواهد بود.

بخش ۲ مروری بر راه‌حل‌های رایج برای مشکل کمبود داده‌ها و روش‌های استفاده شده جهت انتخاب ویژگی را ارائه می‌دهد. بخش ۳ طراحی شبکه پیشنهادی را توصیف می‌کند و تجزیه و تحلیل نظری را ارائه می‌دهد. بخش ۴ جزئیات آزمایش‌ها، از جمله توصیف داده‌ها، ویژگی‌ها، تنظیمات آزمایشی و پروتکل‌های ارزیابی را معرفی می‌کند. بخش ۵ نتایج تجربی را ارائه و تحلیل می‌کند. سرانجام، بخش ۶ نتیجه‌گیری و کارهای آینده را ارائه می‌دهد.

۲- مطالعات مرتبط

در سال ۲۰۱۱، بوژارت و همکارانش با وزن‌دهی به ضرایب مل کپسترول باتوجه‌به موقعیت فرمنت‌ها، ویژگی‌های جدیدی به نام[4] WMFCC را ابداع نموده و با استفاده از این ویژگی‌ها و یک طبقه‌بند مبتنی بر مدل مخفی مارکوف به طبقه‌بندی چهار احساس خشم، آرامش، تأکید و عادی بر روی پایگاه‌داده AIBO پرداختند. آن‌ها به‌منظور ارزیابی روش خود از نرخ تشخیص بدون وزن استفاده نموده و به عدد ۷۰٪ دست یافتند [۹]. در یکی از تحقیقات شاخص صورت‌گرفته در این سال، هو و همکارانش ویژگی‌های طیفی جدیدی را معرفی نموده و از این ویژگی‌ها به‌منظور طبقه‌بندی هفت احساس موجود در پایگاه‌داده برلین[5] و

[1] Maximum Likelihood
[2] Universal Background Models (UBMs)
[3] Decision Tree
[4] Weighted MFCC
[5] A database of German emotional speech

تخمین احساس در فضای پیوسته بر روی پایگاه‌داده VAM استفاده نمودند. آن‌ها با استفاده از الگوریتم انتخاب ویژگی دومرحله‌ای فیلتر - بسته‌بند مبتنی بر معیار فیشر، الگوریتم جداساز خطی و جستجوی روبه‌جلو و طبقه‌بند مبتنی بر ماشین بردار پشتیبان به بهترین نرخ تشخیص ۸۵,۶٪ برای پایگاه‌داده برلین دست پیدا کردند. همچنین با استفاده از رگرسیون مبتنی بر ماشین‌های بردار پشتیان متوسط ضریب همبستگی ۷۳٪ را برای پارامترهای تخمین زده شده در فضای پیوسته احساس به دست آوردند[۱۰]. در این سال لوکا و همکارانش سیستم خود را که به‌منظور بازشناسی احساس بر روی یک پایگاه‌داده طبیعی که با استفاده از مکالمات تلفنی ضبط شده طراحی نموده بودند آزمایش کرده و نتایج را با استفاده از ماتریس‌های تداخل گزارش نمودند [۱۱]. در سال ۲۰۱۲، طیف وسیعی از ویژگی‌های عروضی و طیفی به‌منظور جداسازی چهار احساس خشم، خوشحالی، ناراحتی و عادی بر روی پایگاه‌داده IEMOCAP[1] مورد آزمایش قرار گرفتند. این تحقیقات با استفاده از الگوریتم‌های انتخاب ویژگی جستجوی روبه‌جلو و طبقه‌بندی مبتنی بر ماشین بردار پشتیبان، کارایی ویژگی‌های مختلف را در طبقه‌بندی هر احساس به‌خوبی با یکدیگر مقایسه نموده‌اند[۱۲].

در سال ۲۰۱۳، مروی و همکارش از ترکیب ویژگی‌های طیفی (ویژگی‌هایی که از طیف سیگنال دست می‌آیند مثل فرمنت‌ها، MFCC و PLP) و ویژگی‌های عروضی (ویژگی‌هایی که از آنالیز سیگنال در حوزه زمان دست می‌آیند و اغلب از منحنی فرکانس گام و انرژی سیگنال استخراج می‌شوند) استفاده کردند. آنها از الگوریتم دومرحله‌ای شامل معیار فیشر و الگوریتم جداساز خطی به‌منظور کاهش ویژگی‌ها و همچنین به‌منظور تشخیص احساس از گفتار استفاده نمودند و نشان دادند که ترکیب ویژگی‌های عروضی و طیفی باعث افزایش متوسط نرخ تشخیص می‌شود و نرخ تشخیص در پایگاه‌داده درام را برای گویندگان مرد به ۴۷,۲۸٪ و نرخ تشخیص در گویندگان زن را به ۵۵,۷۴٪ افزایش داده است. در همین سال، منصور شیخان و همکارانش از شبکه ماشین بردار پشتیبان برای تشخیص احساسات گفتار با استفاده از روش انتخاب ویژگی ANOVA بهره بردند. آنها نشان دادند که حتی با حذف ۲۲٪ از ویژگی‌ها، به طور متوسط دقت تشخیص احساسات را می‌توان تا ۲,۲٪ بهبود داد. همچنین، شبکه ماشین بردار پشتیبان پیشنهادی دقت تشخیص را حداقل به میزان ۸٪ در مقایسه با طبقه‌بندی کننده یکپارچه شبیه‌سازی‌شده بهبود می‌دهد.

در سال ۱۳۹۶، حریمی و همکارش تشخیص احساس از گفتار را با استفاده از مدل و ویژگی‌های دینامیکی غیرخطی بررسی نموده و احساس‌های با سطح برانگیختگی یکسان را با استفاده از ویژگی‌های دینامیکی غیرخطی از یکدیگر جدا نمودند. در روش پیشنهادی که بر روی پایگاه‌داده برلین با استفاده از تکنیک ۱۰ تکه برابر ارزیابی شد، نرخ بازشناسی ۹۶,۳۵٪ و ۸۷,۱۸٪ برای زنان و مردان به دست آمد [۱۳].

در مرجع [۱۴] ابتدا برای هر سگمنت یعنی مجموعه‌ای از فریم‌های پشت‌سرهم، با استفاده از شبکه عصبی عمیق یک توزیع احتمالی حالت احساسی مبتنی بر سگمنت تولید شده و سپس ازروی این توزیع‌های احتمالی، ویژگی‌های در سطح کلی به‌دست‌آمده است. این ویژگی‌های کلی به یک ELM داده شده (یک شبکه عصبی خاص تک‌لایه مخفی ساده و مؤثر) تا احساسات کلی به دست آید. علت استفاده از ELM در گام آخر این است که ELM ساده‌تر از DNN است و داده‌های کمتری برای آموزش نیاز دارد و در اینجا عملکرد خیلی بهتری از ماشین بردار پشتیبان دارد. نتایج عملی نشان می‌دهند که رویکرد پیشنهاد شده اطلاعات احساسی را به طرز مؤثری از ویژگی‌های سطح پایین به دست می‌آورد و منجر به افزایش ۲۰٪ دقت در مقایسه با رویکردهای جدید شده است.

در مرجع [۱۵] از روش اصلاح شده VQ برای کاهش ویژگی‌های آکوستیکی استفاده شده است. به این صورت که روش VQ و همچنین روش VQ تفاضلی روی ویژگی‌های آکوستیکی در سطح فریم استفاده شده است. استفاده از VQ تفاضلی پایداری ویژگی‌ها را با اضافه‌کردن دینامیک زمانی که حین آنالیز آماری از بین می‌رود تقویت می‌کند. مکانیزم استفاده شده در اظهارات مختلف ارزیابی شده است به‌منظور رسیدن به این هدف یک پایگاه‌داده محلی هم‌زمان

[1] Interactive emotional dyadic motion capture database

با پایگاه‌داده EMO-DB با الگوریتم پیشنهاد شده استفاده شده است. همچنین پارامترهای طراحی طبقه‌بندی کننده و ویژگی‌های کاهش داده شده برای عملکرد بهینه فراهم شده است و در نهایت ویژگی‌های دیگری در سطح فریم و در سطح کل سخن اضافه شده تا بسته پیشنهادی تکمیل گردد. در مرجع [۱۶] ابعاد ویژگی را از ۲۷۶ به ۷۵ توسط روش SFFS کاهش دادند و تشخیص احساس به میزان ۲,۷٪ بهتر شد. در یک مطالعه از روش‌های مختلفی برای کاهش ویژگی از ۵۸ به ۱۸، ۲۸، ۳۱ و ۳۳ استفاده شد و دقت تشخیص احساس به ترتیب به میزان ۴,۵٪ و ۳٪ و ۲,۲٪ و ۰,۷٪ افزایش پیدا کرد.

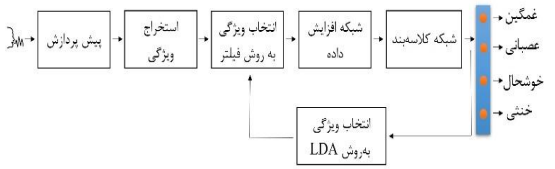
در مرجع [۱۷] عملکرد تشخیص هنگامی که ابعاد ویژگی توسط روش mRMR از ۳۸۰ به ۱۲۱ کاهش پیدا کرد، به میزان ۱,۵٪ کم شد. در مرجع [۱۸] هنگامی که ابعاد ویژگی از ۵۵ به ۴۹ و ۴۵ و ۲۴ و ۸ توسط روش FCBF کاهش پیدا کرد تغییر در میزان تشخیص به ترتیب به میزان ۰,۹٪ و ۱,۱-٪ و ۲,۳-٪ و ۳,۴-٪ به دست آمد. در مرجع [۱۹] نرخ تشخیص با کاهش ویژگی از ۲۰۴ به ۸۷ توسط روش FCBF و طبقه‌بند ماشین بردار پشتیبان به میزان ۱,۵٪ افزایش پیدا کرد. هو و همکاران [۲۰] از شبکه عصبی کانولوشن بسیار عمیق برای تولید ویژگی‌های اضافی برای آموزش مدل‌های صوتی استفاده کردند و دریافتند که افزایش داده به ساخت سیستم‌های تشخیص گفتار کمک بسیاری می‌کند. همچنین مقالاتی در خصوص بهبود نرخ تشخیص احساس با استفاده از تفکیک جنسیتی در سال ۲۰۱۷[۲۱]، استخراج ویژگی‌های مقاوم گفتار در سال ۲۰۲۰ [۲۲] و تولید سیگنال‌های مصنوعی به کمک شبکه-های عصبی مصنوعی در سال ۱۳۹۵ [۲۳] نیز منتشر گردیده‌اند. در سال ۲۰۲۱، چوراسیا و همکارانش یک سیستم تشخیص احساس از گفتار را بر اساس ویژگی‌های استخراج‌شده و به‌دست‌آمده از طیف‌نگارهای فرکانس مل (MFCC) پیشنهاد کردند. آن‌ها از شبکه‌های عصبی عمیق یک‌بعدی و پایگاه‌داده RAVDESS و از شش احساس بر اساس جنسیتشان (مرد و زن) استفاده نموده و توانستند به دقت ۸۲,۳ درصد در طبقه‌بندی برسند. احساسات طبقه بندی شده توسط آن‌ها عبارت‌اند از: شادی، غم، عصبانیت، آرامش، ترس و استرس[۲۴].

در سال‌های اخیر، از شبکه‌های مولد متخاصمی برای تشخیص احساس از گفتار استفاده شده است. به‌عنوان‌مثال، چانگ و شفر از یک شبکه مولد متخاصمی عمیق[1] برای یادگیری نمایش متمایز از گفتار احساسی به روشی نیمه نظارت استفاده کردند [۲۵].

۳- روش پیشنهادی

۳-۱- ساختار کلی پیشنهاد شده

شکل (۱)، ساختار کلی شبکه طبقه‌بندی کننده احساس پیشنهاد شده را برای طبقه‌بندی چهار احساس نمایش می دهد. از روش ترکیبی، جهت انتخاب ویژگی و از روش افزایش داده متخاصمی جهت افزایش داده‌ها و از ماشین بردار پشتیبان جهت طبقه‌بندی استفاده گردیده است.



شکل ۱- ساختار کلی شبکه تشخیص احساس پیشنهادی

تاکنون تحقیقات بسیاری در زمینه افزایش داده‌ها و یا انتخاب ویژگی‌های بهینه در سیستم‌های تشخیص احساس از گفتار انجام شده است ولی به‌ندرت به مسئله انتخاب ویژگی در نمونه‌های مصنوعی تولید شده و یا موضوع افزایش داده و انتخاب ویژگی به‌صورت هم‌زمان پرداخته شده است. همان‌طور که در شکل (۱) و ساختار کلی بررسی شده در این مقاله نمایش داده شده است طی دو مرحله ویژگی‌های مؤثر انتخاب و با فیدبکی که از شبکه طبقه‌بند گرفته می‌شود تکنیک مذکور بهینه خواهد شد.

۳-۲- پیش‌پردازش و استخراج ویژگی

در این مقاله، پس از انجام پیش‌پردازش و حذف نویز از سیگنال گفتار، بردارهای ویژگی توسط نرم‌افزار openSMILE استخراج گردید [۵]. از ویژگی‌های سطح پایین مثل انرژی فریم میانگین مربع[2]، ضرایب کپسترال مل فرکانسی[3]، نرخ عبور از صفر[4]، احتمال صدا[5] و فرکانس‌های اساسی[6] استفاده شد.

[1] Deep Convolutional GAN (DCGAN)
[2] Root-Mean-Square (RMS) Frame Energies
[3] Mel-Frequency Cepstrum Coefficients (MFCCs)
[4] Zero-Crossing Rates
[5] Voice Probabilities
[6] Fundamental Frequencies

برای استخراج این ویژگی‌ها، فریم‌های ۲۵ میلی‌ثانیه با هم‌پوشانی ۱۰ میلی‌ثانیه از سیگنال گفتار در نظر گرفته شدند. سپس ویژگی‌های سطح پایین توسط عملگرهای آماری مانند حداکثر، حداقل، دامنه، انحراف معیار و ... برای استخراج ویژگی‌های سطح بالا پردازش شدند. با استفاده از این عملگرهای آماری به ویژگی‌های سطح پایین، ویژگی‌های فریم، فریم شده به ویژگی‌های بیانی[1] تعمیم داده شد و برای هر چهار مجموعه پایگاه‌داده، ویژگی‌های با واریانس صفر حذف شدند و بقیه ویژگی‌ها به طور مستقل توسط z-norm نرمال گردید. تعداد و نوع این ویژگی‌ها برای هر نمونه، ۲۱۸۵ ویژگی مطابق جدول ۱ می‌باشد.

جدول ۱- تعداد و نوع ویژگی‌های استخراج شده

نوع ویژگی	مشخصات ویژگی
الگوهای طیفی (۲۰۴ ویژگی):	میانگین تعداد الگوها در هر قاب برای هر باند فرکانسی تعداد نسبی هر الگو در باندهای فرکانسی و تعداد نسبی هر الگو در باندهای فرکانسی
انرژی هارمونیک‌ها (۷۸۰ ویژگی):	اعمال ۲۰ تابع آماری به E1 و E2 و مشتقات اول و دوم آن‌ها
ویژگی‌های عروضی (۲۴۱ ویژگی):	اعمال ۲۰ تابع آماری به منحنی فرکانس گام، منحنی انرژی، منحنی نرخ عبور از صفر، منحنی اپراتور انرژی تیگر و همچنین نسبت طول زمانی مصوت‌ها به صامت‌ها و کلیه مشتقات اول و دوم آن‌ها
ویژگی‌های طیفی (۹۶۰ ویژگی):	اعمال ۲۰ تابع آماری به ۱۲ ضریب اول MFCC، ۴ فرمنت اول و مشتقات اول و دوم آن‌ها
۲۰ تابع آماری شامل: مقدار کمینه، بیشینه، برد، میانگین، میانه، ۱۰٪ و ۲۵٪ صدک‌های اول، پنجم، دهم، بیست و پنجم، هفتاد و پنجم، نودم، نود و پنجم و نود و نهم، برد میان چارکی، واریانس، انحراف معیار، چولگی و کشیدگی.	

۳-۳- انتخاب ویژگی به روش فیلتر

پس از استخراج بردارهای ویژگی احساسی و به‌منظور فیلتر نمودن آن‌ها از معیار فیشر جهت انتخاب ویژگی مقدماتی به روش فیلتر مطابق با رابطه ۱ و با چهار طبقه احساسی (C=۴) استفاده گردیده است. در این مرحله، واریانس بین کلاسی و درون کلاسی مطابق رابطه فیشر محاسبه و مقدار به‌دست‌آمده میزان اهمیت ویژگی را مشخص می‌کند.

$$FDR(u) = \frac{2}{c(c-1)} \Sigma_{c1} \Sigma_{c2} \frac{(\mu_{c1,u} - \mu_{c2,u})^2}{\sigma_{c1,u}^2 + \sigma_{c2,u}^2} \quad (۱)$$

$$1 \le c_1 < c_2 \le c$$

دراین‌رابطه $\mu_{ci.u}$ و $\sigma_{ci.u}^2$ مقدار میانگین و واریانس طبقه i–ام $(i=1,2,3,4)$ را برای ویژگی u–ام نشان می‌دهند. امتیاز ویژگی i–ام بوسیله رابطه ۱ محاسبه گردیده و تمامی ویژگی‌ها بر اساس امتیازشان از بیشتر به کمتر مرتب می شوند.

همان‌طور که می‌دانیم، ویژگی‌هایی برای طبقه‌بند احساسی مناسب هستند که واریانس درون کلاسی حداقل و واریانس بین کلاسی حداکثری داشته باشد [۲۵]. این مرحله از انتخاب ویژگی، بسیار سریع خواهد بود و در مدت‌زمان کوتاهی به ویژگی‌ها امتیاز داده می‌شود و امتیازهای پایین حذف می‌گردند؛ این آستانه از قبل مشخص شده است. در این مقاله این آستانه مقادیر ۷۰٪ کمتر از میانگین بالاترین ده امتیاز فیشر در نظر گرفته شد؛ لذا از ۲۱۸۵ ویژگی به ۱۵۸۲ ویژگی برای هر نمونه خواهیم رسید. البته می‌توان به‌جای انتخاب مقدار آستانه، تعداد ویژگی‌های موردنظر را به‌عنوان مقدار آستانه تعیین نمود (مثلاً ۱۵۰۰ ویژگی با بالاترین امتیاز فیشر). اشکال اساسی این مرحله این است که هدف، انتخاب ویژگی‌هایی است که بتوانند عملکرد شبکه طبقه‌بند را بالاتر از قبل برده و راندمان آن را افزایش دهند ولی این خطر همیشه وجود دارد که ویژگی‌های انتخاب شده ترکیب خوبی را تشکیل ندهند. به بیان دیگر ممکن است ویژگی‌هایی انتخاب شوند که به‌تنهایی خوب عمل کنند ولی در کنار هم عالی عمل نکنند.

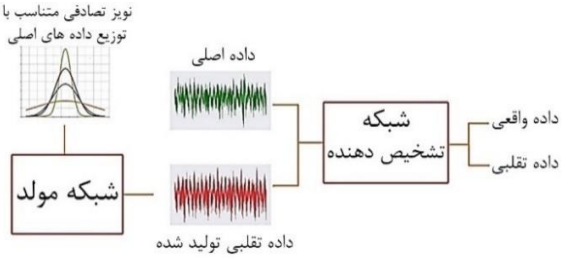
۴-۳- افزایش داده

شکل (۲)، شبکه مولد متخاصمی و اهداف شبکه‌های مولد و تشخیص‌دهنده را نشان می‌دهد. هدف از شبکه مولد، فریب شبکه تشخیص‌دهنده است؛ بنابراین شبکه عصبی مولد آموزش‌داده‌شده است تا خطای طبقه‌بندی نهایی (بین داده‌های واقعی و تولید شده) را به حداکثر برساند.

در مقابل هدف شبکه تشخیص‌دهنده این است که داده‌های تولید شده جعلی را کشف کنند؛ بنابراین شبکه عصبی

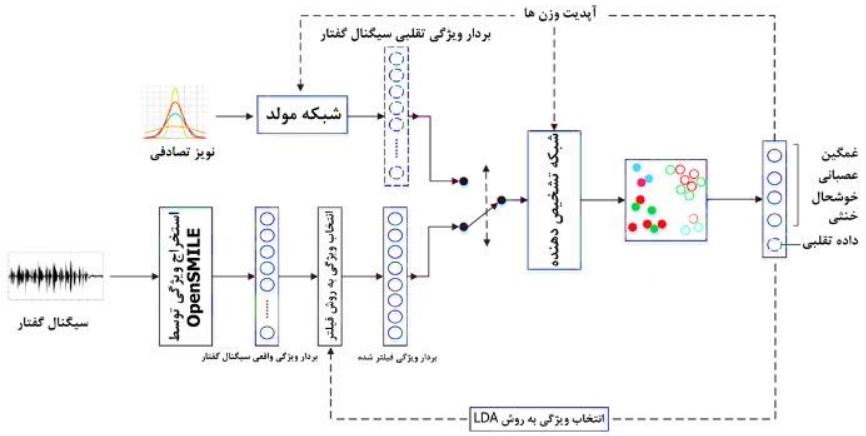
[1] Utterance-Level Features

تشخیص‌دهنده آموزش‌دیده تا خطای طبقه‌بندی نهایی را به حداقل برساند. مطابق شکل (۲)، در هر تکرار از روند آموزش، وزن‌های شبکه مولد به‌منظور افزایش خطای طبقه‌بندی (گرادیان صعودی خطا بر روی پارامترهای مولد) به‌روز می‌شود درحالی‌که وزن‌های شبکه تشخیص‌دهنده برای کاهش این خطا (گرادیان نزولی خطا بر پارامترهای متمایزکننده) به‌روز می‌شود. هر دو شبکه سعی می‌کنند یکدیگر را شکست دهند و با این کار، هر دو بهتر و بهتر می شوند. رقابت بین آنها باعث می‌شود این دو شبکه باتوجه‌به اهداف مربوطه خود پیشرفت کنند. ازنقطه‌نظر تئوری بازی، می‌توان این روش را به‌عنوان یک بازی دونفره در نظر گرفت که وضعیت تعادل در آن وضعیتی است که شبکه مولد، داده‌ها را با توزیع هدفمند دقیق تولید کند درحالی‌که شبکه تشخیص‌دهنده، پیش‌بینی "واقعی" یا "تولید شده" بودن آنها را با احتمال ½ انجام دهد. در واقع بازی زمانی تمام می‌شود که شبکه تشخیص‌دهنده قادر به تشخیص نمونه‌های تولید شده از نمونه‌های اصلی نبوده و با احتمال ۵۰٪ آنها را از هم جدا نماید.



شکل ۲- نمایش شبکه مولد متخاصمی

ازآنجاکه مدل پیشنهادی در شکل (۳)، بردار ویژگی داده‌ها را به‌صورت یک‌به‌یک بین یک منبع نویز و یک نمونه هدف می‌آموزد، برای یک مجموعه‌داده گفتار دارای N برچسب طبقهای احساسی، نیاز به ایجاد انتقال برداری بین هر جفت از آنها است، یعنی $N(N-1)/2$ انتقال، که بسیار زیاد است.



شکل ۳- ساختار شبکه افزایش داده متخاصمی و انتخاب ویژگی پیشنهادی

در روش پیشنهاد شده در این مقاله، از داده‌های برچسب زده شده از هر نوع احساسات به‌عنوان دامنه هدف استفاده شده است درحالی‌که داده‌ها در دامنه منبع، بدون برچسب هستند. بنابراین، یک مجموعه داده هدف مصنوعی تولید می‌شود که به اندازه داده‌های منبع واقعی باشند و دارای همان احساسات باشند. به‌عنوان مجموعه داده هدف واقعی، داده‌های مصنوعی نیز پس از جداسازی توسط روش انتخاب ویژگی، برای آموزش شبکه طبقه‌بند مورداستفاده قرار می‌گیرند. درحالی‌که منبع مصنوعی در دامنه مجموعه داده بدون برچسب قرار دارد. به جای آموزش N شبکه مولد متخاصمی به طور جداگانه، شبکه های افزایش داده متخاصمی در یک چارچوب کامل قرار گرفتند تا نمونه‌های تولید شده هر احساس هدف با یکدیگر مرتبط شوند. علاوه بر این، روش پیشنهاد شده در این مقاله، نمونه‌های مصنوعی را در فضای ویژگی ایجاد می‌کند. به این معنی که ورودی شبکه افزایش داده متخاصمی، یک سیگنال گفتار خام نیست بلکه بردارهای ویژگی مورداستفاده برای شبکه طبقه‌بند است که توسط نرم افزار OpenSMILE که یک نرم افزار منبع باز برای استخراج خودکار ویژگی‌ها از سیگنال‌های صوتی و طبقه‌بندی سیگنال‌های گفتاری و موسیقی است استخراج می‌شوند [۲۶]. دلیل این امر این است که هدف اصلی این مقاله انتخاب ویژگی‌های مهم به جای ترکیب گفتار است. بااین‌حال، تولید نمونه در فضای ویژگی‌ها هم مزایا و هم مضراتی دارد. مزیت آن این است که می‌توان بدون درنظرگرفتن سنتز گفتار، بر تحقیقات شبیه سازی توزیع داده‌ها از طریق شبکه های مولد

متخاصمی متمرکز شد و ایراد آن در این است که احساسات نمونه‌های تولید شده فاقد یک شکل واضح برای ارزیابی ادراکی انسان است اما هنوز هم می‌توان ویژگی‌های عاطفی آنها را با مقایسه شباهت آنها با نمونه‌های داده واقعی آزمایش نمود.

الگوسازی شبکه‌های مولد متخاصمی در اصل نیاز به تعریف دو چیز دارد: ۱- معماری ۲- تابع خطا. شبکه‌های مولد متخاصمی همان‌طور که گفته شد از دو شبکه اصلی تشکیل شده‌اند:

یک شبکه مولد که از ورودی تصادفی z با چگالی p(z) استفاده می‌کند و خروجی G(z) که باید (پس از آموزش) دنبال کند، توزیع احتمال هدفمند را تولید می‌کند.

یک شبکه تشخیص‌دهنده که ورودی x را دریافت می‌کند و می‌تواند یک داده واقعی X(t) باشد که چگالی آن با p(x) نشان‌داده‌شده است و احتمال D(x) را به‌عنوان واقعی بودن دادهٔ x گزارش می‌کند.

اگر داده‌های "واقعی" و "تولید شده" به یک نسبت به شبکه تشخیص‌دهنده ارسال شوند، خطای مطلق مورد انتظار از شبکه تشخیص‌دهنده به شرح زیر خواهد بود:

$$E(G,D) = \frac{1}{2}\mathbb{E}_{x\sim p_t}[1-D(x)] + \frac{1}{2}\mathbb{E}_{z\sim p_z}\big[D\big(G(z)\big)\big] = \frac{1}{2}\left(\mathbb{E}_{x\sim p_t}[1-D(x)] + \mathbb{E}_{x\sim p_g}[D(x)]\right) \quad (۲)$$

دراین‌رابطه، $D(x)$ احتمال واقعی بودن نمونه x توسط شبکه تشخیص دهنده، $D\big(G(z)\big)$ احتمال واقعی بودن نمونه x تولید شده توسط شبکه مولد و p_z توزیع نویز ورودی است. هدف شبکه مولد این است که شبکه تشخیص‌دهنده را که هدف آن، توانایی تفکیک داده‌های واقعی و تولید شده است، فریب دهد؛ بنابراین، هنگام آموزش شبکه مولد، هدف این است که این خطا حداکثر شود، درحالی‌که سعی می‌شود این خطا برای شبکه تشخیص‌دهنده به حداقل برسد و نمونه‌های تقلبی به‌خوبی از نمونه‌های اصلی تشخیص داده شوند تا شبکه مولد بتواند نمونه‌های بهتری تولید کند که توسط شبکه تشخیص‌دهنده قابل‌شناسایی نباشد و داریم:

$$\max_G \left(\min_D E(G,D)\right) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \quad (۳)$$

برای هر شبکه مولد با چگالی احتمال p(g)، بهترین شبکه تشخیص‌دهنده آن است که بتواند این تابع را مینیمم کند.

$$\mathbb{E}_{x\sim p_t}[1-D(x)] + \mathbb{E}_{x\sim p_g}[D(x)] = \int_R \big(1-D(x)\big)p_t(x) + D(x)p_g(x)dx \quad (۴)$$

اگر به‌منظور مینیمم کردن نسبت به D در این انتگرال، تابع داخل انتگرال به‌ازای هر مقدار x مینیمم شود به این ترتیب بهترین شبکه تشخیص‌دهنده ممکن برای شبکه مولد مشخص خواهد شد.

$$(p_t(x) \geq p_g(x)) \quad (۵)$$

در حقیقت، یکی از بهترین انتخاب‌ها این است که pt(x) = pg(x) شود. سپس G موردنیاز است که این تابع را ماکزیمم کند.

$$\int_R (1-D_G^*(x))p_t(x) + D_G^*(x)p_g(x)dx = \int_R min(p_t(x),p_g(x))dx \quad (۶)$$

به طور مشابه، برای به حداکثر رساندن این انتگرال نسبت به g، این تابع باید برای هر مقدار x، ماکزیمم شود. ازآنجاکه چگالی p(t) مستقل از مولد G است، G نمی‌تواند به‌گونه‌ای بهتر از حالت زیر تنظیم شود:

$$p_g(x) \geq p_t(x) \quad (۷)$$

البته، ازآنجاکه p(g) یک چگالی احتمال اســت که باید به ۱ برسد، لزوماً برای بهترین G داریم:

$$p_g(x) = p_t(x) \quad (۸)$$

بنابراین، نشان‌داده‌شده است که در یک مورد ایدئال با ظرفیت‌های نامحدود شبکه‌های مولد و تشخیص‌دهنده، نقطه بهینه تنظیم شبکه تخاصمی به‌گونه‌ای است که شبکه مولد چگالی‌ای مانند همان چگالی واقعی را تولید کند و شبکه تشخیص‌دهنده مجبور است در یکی از دو حالت، صحیح بودن را انتخاب کند. در آخر نیز G این معادله را حداکثر می‌کند:

$$\frac{1}{2}\int_R min(p_t(x),p_g(x))dx = \int_R \frac{min(p_t(x),p_g(x))}{p_t(x)+p_g(x)}\frac{p_t(x)+p_g(x)}{2}dx \quad (۹)$$

به این شکل، انتظار می‌رود که G می‌خواهد احتمال شبکه تشخیص‌دهنده را در تشخیص اشتباه به حداکثر برساند. رایانه‌ها اساساً می‌توانند متغیرهای شبه تصادفی ساده تولید

کنند. به‌عنوان‌مثال می‌توانند متغیرهایی تولید کنند که بسیار نزدیک به توزیع یکنواخت عمل می‌کنند. روش‌های مختلفی برای تولید متغیرهای تصادفی پیچیده‌تر وجود دارد از جمله مفهوم "روش تبدیل" که شامل بیان یک متغیر تصادفی به‌عنوان تابعی از متغیرهای تصادفی ساده‌تر است. در یادگیری ماشین، مدل‌های مولد سعی می‌کنند داده‌هایی از توزیع احتمالی خاص (پیچیده) تولید کنند.

مدل‌های مولد شبکه‌های یادگیری عمیق به‌عنوان شبکه‌های عصبی (توابع بسیار پیچیده) مدل شده‌اند که به‌عنوان ورودی یک متغیر تصادفی ساده را گرفته و یک متغیر تصادفی را تولید می‌کنند که از توزیع احتمال هدفمند پیروی می‌کنند. این شبکه‌های مولد را می‌توان "مستقیم" آموزش داد (با مقایسه توزیع داده‌های تولید شده با توزیع داده‌های واقعی). این ایده شبکه‌های مولد تطبیق دهنده است. این شبکه‌های مولد همچنین می‌توانند به‌صورت "غیرمستقیم" با تلاش برای فریب شبکه دیگری که در همان زمان آموزش‌داده‌شده است آموزش داده شوند تا داده‌های تولید شده را از داده‌های واقعی متمایز کند. این ایده شبکه‌های مولد تخاصمی است. این روش برای تغییر دادن تابع خطا از مقایسه مستقیم به غیرمستقیم واقعاً کاری است که می‌تواند برای دستاوردهای بعدی در حوزه یادگیری عمیق بسیار الهام‌بخش باشد.

الگوریتم ۱- نحوه آموزش شبکه مولد متخاصمی با روش نزول گرادیان. تعداد گام‌ها برای آموزش شبکه تشخیص‌دهنده برابر k و به‌عنوان پارامتر اولیه در نظر گرفته شده است.

برای تعداد تکرارهای آموزش تکرار شود:
برای تعداد $\mathbf{k}$ تکرار شود:
به تعداد $\boldsymbol{m}$ از فضای نویز اولیه $\mathbf{p_g(z)}$ نمونه‌برداری می‌شود.
$\boldsymbol{z} = \{\boldsymbol{z^{(1)}}, \dots, \boldsymbol{z^{(m)}}\}$
به تعداد $\boldsymbol{m}$ از توزیع اولیه داده‌ها $\mathbf{p_{data}}$ نمونه‌برداری می‌شود.
$\boldsymbol{x} = \{\boldsymbol{x^{(1)}}, \dots, \boldsymbol{x^{(m)}}\}$
با محاسبه گرادیان زیر خطای شبکه متمایزکننده محاسبه شود:
$\nabla_{\theta_d} \frac{1}{m} \sum_i [\log D(x^{(i)}) + \log(1 - D(G(z^{(i)})))]$
پایان حلقه دوم
به تعداد $\boldsymbol{m}$ از فضای نویز اولیه $\mathbf{p_g(z)}$ نمونه‌برداری می‌شود.
$\boldsymbol{z} = \{\boldsymbol{z^{(1)}}, \dots, \boldsymbol{z^{(m)}}\}$
وزن‌های شبکه مولد با روش کاهش گرادیان به‌صورت زیر به‌روزرسانی می‌شود:
$\nabla_{\theta_g} \frac{1}{m} \sum_i [\log(1 - D(G(z^{(i)})))]$
اتمام حلقه اول

در شبکه‌های مولد متخاصمی فرض بر آن است که یک مجموعه‌داده ورودی $(si,yi)Ni=1$ که نیمی از داده‌ها از نمونه‌های واقعی X و نیمی از نمونه‌های مصنوعی $G(z)$ می‌باشند موجود است. هر نمونه آموزش si با یک برچسب yi مطابقت دارد. تمام نمونه‌های واقعی به‌عنوان یک و همه نمونه‌های مصنوعی به‌عنوان صفر برچسب گذاری شده اند. ازآنجاکه شبکه تشخیص دهنده را می‌توان یک طبقه‌بند دودویی در نظر گرفت، عملکرد این خطا را می‌توان به‌عنوان یک خطای متقاطع باینری تعریف کرد [۲۷].

$$J^{(D)}(D,G) = H((s_i,y_i)_{i=1}^N, D) = -\frac{1}{N}\sum_{i=1}^N [y_i \log D\left(s_i\right) + (1-y_i)\log(1 - D\left(s_i\right))] \quad (۱۰)$$

اگر yi با عدد یک برای $si = x$ و صفر برای $si = G(z)$ جایگزین شود و همچنین میانگین‌ها با انتظارات جایگزین گردند، تابع خطای شبکه تشخیص دهنده به شرح زیر خواهد بود [۲۷]:

$$J^{(D)}(D,G) = -\mathop{\mathbb{E}}_{x\sim p_{data}(x)}[\log D(x)] - \mathop{\mathbb{E}}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \quad (۱۱)$$

که در آن "p" توزیع داده از نمونه واقعی x است. در یک بازی مینیماکس (همچنین به آن جمع صفر نیز گفته می‌شود)، کل خطای بازیکنان همیشه صفر است. به این معنی که خطای شبکه مولد برعکس $J(D)$ است. بااین‌حال، هنگامی که گرادیان نزولی باتوجه‌به G محاسبه می‌شود، تنها قسمت دوم در رابطه ۱۲ مهم است. بنابراین، تابع خطای شبکه مولد در یک بازی مینیماکس می‌تواند به شرح زیر است [۲۷]:

$$J^{(G)}(G)| = \mathop{\mathbb{E}}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \quad (۱۲)$$

کل بازی را می‌توان با یک عبارت اعتبارسنجی خلاصه کرد. در واقع رابطه ۱۴، هدف اصلی در شبکه‌های مولد متخاصمی است و داریم:

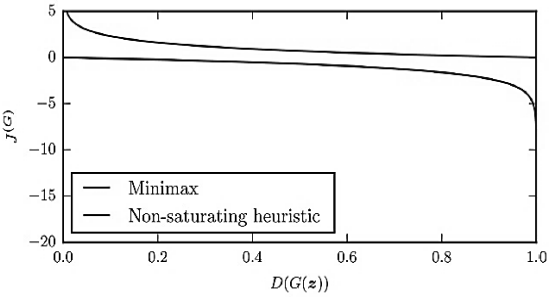
$$\min_G \max_D V(D,G) = \mathop{\mathbb{E}}_{x\sim p_{data}(x)}[\log D(x)] + \mathop{\mathbb{E}}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \quad (۱۳)$$

شبکه مولد سعی می‌کند آن را به حداکثر برساند درحالی‌که شبکه تشخیص‌دهنده سعی در به‌حداقل‌رساندن آن دارد. این بدان معناست که به دنبال کمترین G هستیم که بیشترین D را نتیجه دهد.

اما بروز رسانی خطا در شبکه مولد در یک بازی مین ماکس در عمل به‌خوبی انجام نمی‌شود و هنگامی‌که وزن‌های شبکه مولد هنوز به‌صورت مناسب انتخاب نشده‌اند، شبکه تشخیص‌دهنده به‌راحتی می‌تواند تشخیص دهد که نمونه‌های تولید شده جعلی هستند و با اطمینان بالا آنها را رد می‌کند [۲۷]. در نتیجه، شیب شبکه مولد جهت اصلاح وزن از بین می‌رود و گرادیان آن کوچک می‌شود. تابع خطا در شبکه مولد در یک بازی اکتشافی غیراشباع به شرح زیر است [۲۷]:

$$J^{(G)}(G) = -\mathop{\mathbb{E}}_{z\sim p_z(z)}[\log(D(G(z))] \qquad (۱۴)$$

شکل (۴)، تفاوت عملکرد خطای شبکه مولد در بازی اکتشافی مین ماکس و غیراشباع را نشان می‌دهد. محور افقی احتمال توصیف یک نمونه مصنوعی از نظر شبکه تشخیص‌دهنده را نشان می‌دهد. هرچه این مقدار بالاتر باشد، خطا در شبکه مولد کمتر است. به‌عبارت‌دیگر، هر چه شبکه تشخیص‌دهنده به عدد یک نزدیک می‌شود، نمونه تولید شده توسط شبکه مولد واقعی‌تر است. در قسمت سمت چپ نمودار، $D(G(z))$ نزدیک به صفر است و اغلب در شروع یک فرایند آموزش ظاهر می‌شود. در این زمان، شبکه تشخیص دهنده می‌تواند به راحتی تشخیص دهد که یک نمونه واقعی یا ساختگی است، زیرا شبکه مولد شروع به نمونه گیری از توزیع نویز تصادفی $pz(z)$ با پارامترهای تصادفی می‌کند.



شکل ۴- مقایسه خطای شبکه مولد در یک بازی مین ماکس و یک بازی اکتشافی غیراشباع[۲۸]

شبکه مولد $J(G)$ برای تولید نمونه $G(z)$ به احتمال $D(G(z))$که شبکه تشخیص دهنده به نمونه اختصاص می‌دهد بستگی دارد. هر چه احتمال واقعی بودن بالاتر باشد، خطای کمتری نیز در شبکه مولد دریافت می‌کند. مشاهده می‌شود که منحنی مینماکس در انتهای سمت چپ بسیار صاف است، به این معنی که شبکه مولد دارای شیب بسیار کمی است. با استفاده از گرادیان نزولی شیب شبکه مولد قبلاً متوقف شده است تا در مرحله اولیه عملکرد خود را بهبود بخشد و وزن‌های آن آپدیت گردند. با استفاده از محدودیت های تابع خطای در نظر گرفته شده، منحنی بازی اکتشافی غیر اشباع، شیب خود را در انتهای کار از دست می‌دهد. در این زمان، فرایند آموزش به حد مطلوب رسیده و نمونه‌های تولید شده قادر به فریب دادن شبکه تشخیص دهنده هستند. بنابراین، تابع خطای شبکه مولد در بازی غیر اشباع اغلب در عمل مورداستفاده قرار می‌گیرد درحالی‌که رابطه ۱۴ برای تحلیل نظری استفاده می‌گردد.

۳-۵- انتخاب ویژگی به روش ترکیبی

برای مواجهه با مشکل ابعاد بیش از حد داده‌ها، کاهش ابعاد با ترکیب ویژگی‌ها انجام شد. ترکیبات خطی به جهت محاسباتی و تحلیل ساده‌تر می‌باشند. در واقع روش خطی، داده‌های با ابعاد زیاد را بر روی فضای با ابعاد کمتر منعکس می‌کند. این روش بر اساس آنالیز تفکیک‌پذیری خطی می‌باشد و ماتریس انتقال بهینه را به‌عنوان ماتریسی که کوواریانس بین طبقها را حداکثر و کوواریانس درون طبقها را حداقل می‌کند، تعریف می‌نماید. در این روش تعداد پایه‌های خطی به‌وسیله تعداد طبقه‌ها محدود می‌شود (۴ طبقه) و پایه‌ها همیشه متعامد نیستند؛ لذا از چهار پایه خطی جهت پیاده‌سازی استفاده گردیده است. همچنین، ماتریس‌های کوواریانس درون کلاسی و برون کلاسی به فرم زیر محاسبه می‌شوند[۲۹].

$$S_w = \frac{1}{N}\sum_{i-1}^{I}\sum_{ni=1}^{N_i}(X_{n_i}-\mu_i)(X_{n_I}-\mu_i)^T \qquad (۱۵)$$

$$S_b = \frac{1}{N}\sum_{I=1}^{1} N_1\,(\mu_1-\mu\ )(\mu_1-\mu_I\ )^T \qquad (۱۶)$$

که در آن N تعداد کل نمونه‌ها، N_i تعداد نمونه‌های طبقه i–ام، μ_i میانگین طبقه i–ام، I تعداد طبقه ها و X_{n_I} نمونه n_i–ام می‌باشند. S_w همبستگی درون کلاسی و S_b تمایز بین کلاسی را اندازه گیری می‌کنند. هدف به دست آوردن ماتریس تبدیلW است به نحوی که رابطه فیشر را ماکزیمم کند.

$$J(W) = \frac{W^T trace(S_b)W}{W^T trace(S_w)W} \qquad (۱۷)$$

شبکه طبقه‌بند از نوع ماشین بردار پشتیبان و از نوع چهار طبقه می‌باشد. همچنین از کرنل تابع پایه شعاعی لاپلاس[1] استفاده گردیده است.

$$k(x,y) = exp(-\frac{\|x-y\|}{\sigma}) \quad (۱۸)$$

در این مرحله از انتخاب ویژگی، شبکه طبقه‌بند (ماشین بردار پشتیبان) در جریان انتخاب ویژگی است و عملکرد آن را به‌صورت مستقیم کنترل می‌نماید. ترکیب این دو مرحله انتخاب ویژگی و همچنین افزایش هوشمندانه داده‌ها به روش افزایش داده متخاصمی، سیستم تشخیص احساس از گفتار را به نقطه بهینه‌ای از تعداد داده و ابعاد داده‌ها رسانید.

استفاده از روش‌های انتخاب ویژگی همیشه موفقیت طبقه‌بندی را افزایش نمی‌دهد. نتایج در ادامه نشان می‌دهد که استفاده از هر دو روش جداساز خطی و آنالیز مؤلفه‌های اصلی به طور هم‌زمان، منجر به نتیجه بهتری نسبت به هرکدام از آنها می‌شود. زیرا آنالیز مؤلفه‌های اصلی روی داده‌های همبسته مؤثرتر عمل می‌کند و الگوریتم جداساز خطی روی داده‌های با بعد کم بهتر عمل می‌کند و روش فیشر در کاهش سایز بهتر از آنالیز مؤلفه‌های اصلی عمل می‌کند.

۴- آزمایش‌ها

۴-۱- پایگاه‌داده[2]

آزمایش‌ها این مقاله بر روی چهار پایگاه‌داده متداول EMO-DB، eNTERFACE05، SAVEE و IEMOCAP انجام گردیده و علاوه بر این، تجزیه‌وتحلیل داده‌ها بر روی هر چهار پایگاه‌داده برای چهار احساس غمگین، عصبانی، خوشحال و خنثی ارائه شده است.

معروف‌ترین این پایگاه‌داده‌ها، پایگاه‌داده گفتار احساسی برلین[3] [۳۰] است. پایگاه‌داده گفتار احساسی برلین، یک مجموعه‌داده کوچک است که شامل ۸۰۰ جمله است که به هفت طبقه احساس تقسیم شده‌اند. تمام گفته‌ها توسط ده بازیگر حرفه‌ای ضبط شده است.

۴-۲- معیار ارزیابی داده‌ها و محیط شبیه‌سازی

دقت[4]، صحت[5] و امتیاز اف یک[6] سه سنجه متداولی هستند که علاوه بر صحت طبقه‌بندی برای مسائل طبقه‌بندی یک مدل شبکه عصبی مورداستفاده قرار می‌گیرند. سه سنجه اضافه دیگری که کمتر متداول اما محبوب هستند عبارت‌اند از: ضریب کاپای کوهن[7]، MCC[8] و ماتریس درهم ریختگی[9]. از معیار دقت و صحت جهت معیار ارزیابی آزمایش‌ها استفاده گردیده است. شبیه‌سازی‌ها با استفاده از Keras و Tensorflow2.2 در نرم‌افزار پایتون Python 3.8 (64-bit) انجام گردیدند.

۵- نتایج

به‌منظور مقایسه توزیع‌های پیچیده در فضای با ابعاد بالا، ابتدا اندازه‌گیری هم‌پوشانی مقادیر ویژگی‌های منحصربه‌فرد را معرفی کرده و داریم:

$$f = \frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2} \quad (۱۹)$$

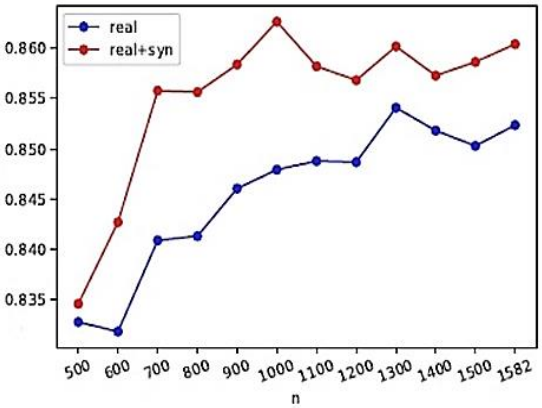
که در آن μ_1، μ_2، σ_1، σ_2 میانگین و انحراف معیار از دو مجموعه داده برای یک بعد از ویژگی خاص هستند. هرچه این مقدار بزرگتر باشد، منطقه همپوشانی کمتری بین دو مجموعه داده وجود دارد.

در این مقاله، تلاش شده است ویژگی‌هایی که بیشترین ارتباط با احساسات دارند انتخاب و سپس بر اساس آن ویژگی‌های انتخاب شده، شبکه طبقه‌بند جهت جدا نمودن چهار احساس آموزش داده شود. اگر اطلاعات اضافی که با عواطف مرتبط نیست فیلتر شوند، می‌توان عملکرد شبکه طبقه‌بند را برای تشخیص احساس از سیگنال گفتار بهبود بخشید و یا دست‌کم سرعت آن را بالا برد. داده‌های مصنوعی تولید شده توسط روش پیشنهادی احتمالاً در مناطقی رخ می‌دهند که مطمئناً می‌توانند به احساساتشان اختصاص یابند؛ بنابراین، تفاوت بین داده‌های مصنوعی و داده‌های هدف نشانگر تغییرات موردنیاز برای تقویت بخشیدن عواطف در داده‌ها است.

بردار تفاوت بین مجموعه‌داده مصنوعی تولید شده و داده‌های موجود در پایگاه‌داده برلین محاسبه گردیدند، سپس بزرگ‌ترین ابعاد ویژگی n بردار تفاوت استخراج شد. این اختلاف به‌عنوان مقدار همپوشانی تعریف شده در رابطه ۱۹ محاسبه گردید. علاوه بر این، به ترتیب دو شبکه ماشین بردار پشتیبان بر اساس ویژگی‌های انتخاب شده ۱- فقط

[1] Laplacian Radial Basis Function
[2] Dataset
[3] Berlin Database of Emotional Speech (EmoDB)
[4] Precision
[5] Recall
[6] F1-Score
[7] Cohen's Kappa
[8] Matthews Correlation Coefficient
[9] Confusion Matrix

داده‌های واقعی و ۲- ترکیب داده‌های واقعی و مصنوعی آموزش داده شد. مقادیر مختلف n برای مشاهده رابطه بین عملکرد شبکه طبقه‌بند و تعداد ویژگی‌های انتخاب شده بررسی گردید.



شکل ۵- نتایج طبقه‌بندی بر اساس تعداد ویژگی‌ها

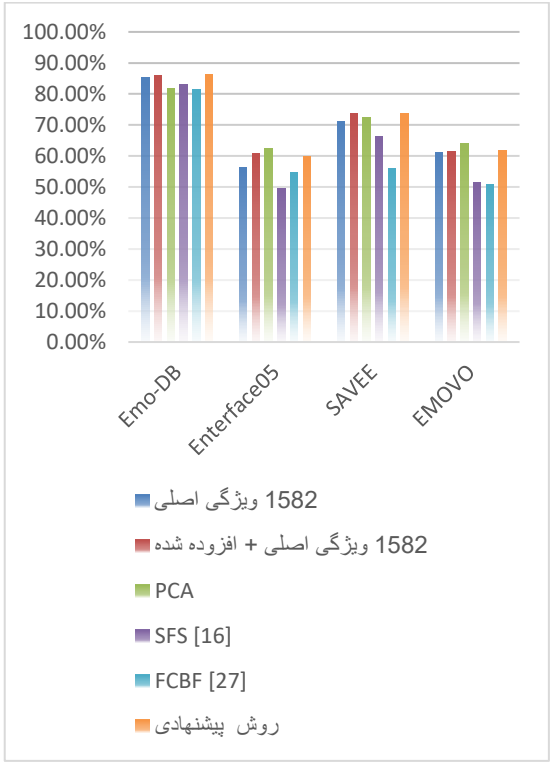
شکل (۵)، معیار صحت هر دو شبکه طبقه‌بند را نشان می‌دهد. همان‌طور که ملاحظه می‌شود، شبکه طبقه‌بند بر اساس ترکیبی از داده‌های واقعی و مصنوعی وقتی $n = 1000$ است به حداکثر خود یعنی ۸۶٫۳۲٪ می‌رسد و شبکه طبقه‌بند مبتنی بر داده‌های واقعی وقتی $n = 1300$ می‌باشد به حداکثر خود یعنی ۸۵٫۲۰٪ خواهد رسید. می‌توان استنباط کرد که آخرین ۵۸۲ ویژگی طبقه‌بند با داده‌های "واقعی + مصنوعی" و ۲۸۲ ویژگی آخر برای طبقه‌بندی‌کننده با داده‌های "واقعی" حاوی اطلاعات بسیار کمی در مورد احساسات هستند. بنابراین، داده‌های مصنوعی نه تنها می‌توانند جهت افزایش داده‌ها بلکه برای انتخاب ویژگیها به‌منظور بهبود عملکرد شبکه طبقه‌بند مورداستفاده قرار گیرند و دقت و سرعت این شبکه را در طبقه‌بندی افزایش دهد. همچنین، پدیده اوج که پیشتر توضیح داده شد به وضوح در شکل (۵) قابل ملاحظه است و همان‌طور که مشاهده می‌شود با افزایش تعداد ویژگی‌های مورداستفاده، ابتدا نرخ تشخیص افزایش پیدا می‌کند اما پس از رسیدن به نقطه اوج، افزایش بیشتر تعداد ویژگی‌ها موجب کاهش نرخ تشخیص می‌شود. علت اصلی این پدیده افت کارایی الگوریتم طبقه‌بند به دلیل پیچیدگی و ابعاد زیاد فضای ویژگی در مقابل تعداد داده‌های آموزشی می‌باشد. به‌عبارت‌دیگر، در چنین شرایطی به دلیل تعدد پارامترهای تنظیم طبقه‌بند (به دلیل ابعاد بالای فضا) و محدود بودن تعداد داده‌های آموزشی (که بستگی به پایگاه‌داده و روش ارزیابی دارد) عمومیت مسئله ازدست‌رفته و الگوریتم به‌جای یادگیری قاعده طبقه‌بندی، به‌صورت خاص داده‌های آموزشی را فرامی‌گیرد که این مسئله کاهش دقت طبقه‌بندی را در مواجهه با داده‌های جدید (داده‌های تست) در پی دارد.

جدول ۲، نتایج طبقه‌بندی چهار احساس از سیگنال گفتار به‌وسیله ماشین بردار پشتیبان و برای چهار پایگاه‌داده نشان می‌دهد. هنگامی که فقط از انتخاب ویژگی به روش فیلتر (معیار فیشر) و فقط بر روی پایگاه‌داده اصلی کار کنیم، نتایج مطابق ستون سوم خواهد بود. تفاوت ستون سوم و چهارم افزودن داده توسط شبکه افزایش داده متخاصمی به پایگاه‌داده اصلی است. همان‌طور که مشاهده می‌شود در تمامی پایگاه‌داده‌های بکار گرفته شده معیار صحت افزایش یافته است چرا که داده‌های متنوع‌تری در اختیار طبقه‌بند جهت آموزش بوده است. ستون پنج، شش و هفتم نتایج همین روند را به ترتیب به‌وسیله سه روش انتخاب ویژگی PCA، SFS[۳۱] و FCBF[۳۲] نشان می‌دهد. همان‌طور که مشاهده می‌شود در اکثر داده‌ها کاهش معیار صحت گزارش گردیده است. علت اصلی آن این است که این روش‌های انتخاب ویژگی نتوانسته‌اند تمامی ویژگی‌های اصلی را جهت آموزش ماشین بردار پشتیبان در خود جای دهند و با کاهش ویژگی‌ها سبب کاهش معیار صحت گردیده‌اند. در ستون آخر، نتایج روش پیشنهادی گزارش گردیده است. همان‌طور که ملاحظه می‌شود، در پایگاه‌داده برلین صحت ۸۶٫۳۲٪ گزارش گردیده است که در مقایسه با سایر روش‌های بکار گرفته شده در [۳۱] و [۳۲] بالاتر است. نمودار میله‌ای ۱ همچنین نمایشی از مقایسه هر شش روش و موفقیت هرکدام به‌صورت مقایسه با دیگری می باشد. بالاترین نرخ دقت مربوط به پایگاه‌داده برلین و روش پیشنهادی است. در سایر پایگاه‌داده‌های مورد آزمایش بیشترین نرخ دقت مربوط به پایگاه‌داده SAVEE و با استفاده از روش پیشنهادی است اگرچه که در دو پایگاه‌داده دیگر روش PCA به موفقیت بالاتری دست‌یافته است چرا که به نظر می‌رسد آنالیز مؤلفه‌های اصلی روی داده‌های همبسته مؤثرتر عمل می‌کند و الگوریتم جداساز خطی روی داده‌های با بعد کم بهتر عمل می‌کند. همچنان که روش فیشر در کاهش سایز بهتر از آنالیز مؤلفه‌های اصلی عمل می‌کند.

جدول ۲- مقایسه نرخ موفقیت طبقه‌بندی در پایگاه‌های داده متفاوت و به‌وسیله روش‌های انتخاب ویژگی گوناگون

پایگاه‌داده	طبقه‌بند	۱۵۸۲ ویژگی اصلی	۱۵۸۲ ویژگی اصلی+ افزوده	PCA	SFS[۳۱]	FCBF[۳۲]	روش پیشنهادی
EMO-DB	SVM	۸۵,۲۰	۸۶,۰۰	۸۱,۷۱	۸۲,۹۳	۸۱,۳۲	۸۶,۳۲
Enterface05	SVM	۵۶,۲۳	۶۰,۶۵	۶۲,۳۱	۴۹,۴۹	۵۴,۶۲	۵۹,۷۷
SAVEE	SVM	۷۰,۹۸	۷۳,۶۵	۷۲,۳۹	۶۶,۲۶	۵۵,۸۳	۷۳,۸۲
IEMOCAP	SVM	۶۱,۱۰	۶۱,۳۸	۶۳,۹۱	۵۱,۴۸	۵۰,۸۹	۶۱,۶۷

نمودار ۱- مقایسه نرخ موفقیت طبقه‌بندی به‌وسیله روش‌های انتخاب ویژگی گوناگون



برای ارزیابی عملکرد روش پیشنهاد شده، نتایج به‌دست‌آمده با یک مطالعه اخیر [۳۳] که در آن نویسندگان از الگوریتم‌های مختلف کاهش ابعاد غیرخطی برای استخراج نمایش ویژگی‌های سطح پایین برای تشخیص احساس از سیگنال گفتار استفاده کرده‌اند مقایسه شده است. سه الگوریتم کاهش ابعاد با عملکرد بالا در [۳۳] انتخاب گردیده که این روش‌ها شامل مقیاس‌بندی چندبعدی[1]، تجزیه‌وتحلیل مؤلفه‌های اصلی[2] و رمزگذار خودکار است. نتایج در جدول ۳ ارائه شده است.

جدول ۳- مقایسه نتایج با استفاده از الگوریتم‌های مختلف کاهش ابعاد ویژگی در پایگاه‌داده IEMOCAP

روش	تکنیک افزایش داده	UAR (%)
SMACOF MDS [33]	ندارد	۵۸,۸
PCA [33]	ندارد	۵۷,۷
AUTOENCODER [33]	ندارد	۵۷,۸
SMACOF MDS [33]	دارد	۵۸,۹
PCA [33]	دارد	۵۸,۳
AUTOENCODER [33]	دارد	۵۸,۵
روش پیشنهادی	دارد	۶۱,۶۷

در [۳۳]، نویسندگان از ماشین بردار پشتیبان برای طبقه‌بندی ویژگی‌هایی که توسط هر الگوریتم کاهش ابعاد آموخته شده است، استفاده کردند. از جدول ۳ می‌توان دریافت که مدل پیشنهادی عملکرد بهتری نسبت به سایر تکنیک‌های کاهش ابعاد غیرخطی دارد و نشان می‌دهد که مدل پیشنهادی به طور کارآمد ویژگی‌ها را در ابعاد کمتر فیلتر می‌کند درحالی‌که اطلاعات احساسی را جهت طبقه‌بندی حفظ می‌نماید.

برای بررسی بهتر مدل پیشنهادی در یک محیط متقاطع، آزمایش‌هایی با استفاده از داده‌های واقعی، مصنوعی، واقعی+ مصنوعی، واقعی+ مصنوعی (کاهش‌یافته) انجام شد و نتایج آنها با دو منبع [۳۴] و [۲۸] مقایسه گردید. همان‌طور که در جدول ۴ گزارش شده است، روش پیشنهادی به نتایج بهتری در مقایسه با دو منبع ذکر شده دست‌یافت. همچنین، انتخاب ویژگی به روش پیشنهادی توانست نتایج را اندکی بهبود دهد. این نشان می‌دهد که مدل پیشنهادی عملکرد سیستم تشخیص احساس از گفتار را در یک محیط متقابل با استفاده از ترکیب داده‌های واقعی و مصنوعی و همچنین هنگامی که داده‌های آموزشی با این نمونه‌های مصنوعی افزایش می‌یابد و سپس فیلتر می‌گردند،

[1] SMACOF (MDS)

[2] PCA

بهبود می‌بخشد و استفاده از روش انتخاب ویژگی آن را بهینه می‌نماید.

جدول ۴- نتایج ارزیابی بین کلاسی

روش	داده‌های واقعی	داده‌های مصنوعی	داده‌های واقعی+ مصنوعی	داده‌های واقعی+ مصنوعی (کاهش‌یافته)
Sahu et al. [34]	۴۵,۱۴	۳۳,۹۶	۴۵,۴۰	-
Bao et al. [28]	۴۵,۵۸	۴۱,۵۸	۴۶,۵۲	-
روش پیشنهادی	۴۶,۱	۴۲,۲۱	۴۶,۶۵	۴۶,۹۵

۶- نتیجه‌گیری

انتخاب ویژگی‌های مؤثر و افزایش داده‌ها از طریق شبکه‌های مولد متخاصمی در پایگاه‌داده‌های کوچک یک موضوع مهم در تشخیص احساس از سیگنال گفتار است. مدل پیشنهادی در این مقاله از ترکیب دو شبکه انتخاب ویژگی، یک شبکه افزایش داده متخاصمی و یک ماشین بردار پشتیبان تشکیل شد. نشان داده شد هنگامی‌که شبکه طبقه‌بند بر اساس ترکیبی از داده‌های واقعی و مصنوعی آموزش می‌بینند، معیار صحت وقتی n=۱۰۰۰ است به حداکثر خود یعنی٪۸۶,۳۲ رسیده و شبکه طبقه‌بند آموزش داده شده مبتنی بر داده‌های واقعی وقتی n =۱۳۰۰ می‌باشد به حداکثر خود یعنی ٪۸۵,۲۰ خواهد رسید. می‌توان استنباط کرد که آخرین ۵۸۲ ویژگی طبقه-بند با داده‌های "واقعی + مصنوعی" و ۲۸۲ ویژگی آخر برای طبقه‌بندی‌کننده با داده‌های "واقعی" حاوی اطلاعات بسیار کمی در مورد احساسات هستند و می‌توان آن‌ها را حذف نمود. از ترکیب دو روش معیار فیشر و الگوریتم جداساز خطی جهت کاهش این ویژگی‌ها استفاده و معیار صحت ٪۸۶,۳۲ در پایگاه‌داده احساسی برلین گزارش شد. بنابراین، داده‌های مصنوعی نه تنها می‌توانند جهت افزایش داده‌ها بلکه برای انتخاب ویژگیها به‌منظور بهبود عملکرد شبکه طبقه‌بند مورداستفاده قرار گیرند و دقت و سرعت این شبکه را در طبقه‌بندی افزایش دهد. باتوجه‌به تحقیقات انجام شده می‌توان گفت که انتخاب ویژگی عملکرد تشخیص احساس را بهتر می‌کند و به طبقه‌بند، داده، نسبت کاهش سایز و حتی به نحوه جمع‌آوری داده‌ها بستگی دارد.

مراجع

[1] Eva Lieskovská, Maroš Jakubec, Roman Jarina and Michal Chmulík, “A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism,” in Electronics, May. 2021.

[2] Akçay, M.B.; O˘guz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication, 2020, pp.116, 56–76.

[3] C. Sima and E. R. Dougherty, “The peaking phenomenon in the presence of feature-selection,” Pattern Recognition Letter , vol. 29, no. 11, 2008, pp. 1667–1674.

[4] J. Rong, G. Li, and Y.-P. P. Chen, “Acoustic feature selection for automatic emotion recognition from speech,” Information Processing & Management, vol. 45, no. 3, 2009, pp. 315–328.

[5] F. Eyben, F. Weninger, F. Gross, and B. Schuller, “Recent developments in openSMILE, the Munich open-source multimedia feature extractor,” in Proc. 21st ACM International Conference on Multimedia, 2013, pp. 835–838.

[6] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, “A unified framework for high-dimensional analysis of M-estimators with decomposable regularizes,” statistical science, vol. 27, no. 4, Nov 2012, pp. 538–557.

[7] H. Zou, T. Hastie, and R. Tibshirani, “Sparse principal component analysis,” Journal of Computational and Graphical Statistics, vol. 15, no. 2, 2006, pp. 265–286.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. “Generative adversarial nets”. In: Advances in neural information processing systems, 2014.

[9] E. Bozkurt, E. Erzin, Ç. E. Erdem, and A. T. Erdem, “Formant position based weighted spectral features for emotion recognition,” Speech Communication., vol. 53, no. 9, 2011, pp. 1186–1197.

[10] S. Wu, T. H. Falk, and W.-Y. Chan, “Automatic speech emotion recognition using modulation spectral features,” Speech Communication, vol. 53, no. 5, 2011, pp. 768–785.

[11] P. Laukka, D. Neiberg, M. Forsell, I. Karlsson, and K. Elenius, “Expression of effect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation,” Computer Speech & Language, vol. 25, no. 1, 2011, pp. 84–104.

[12] H. Pérez-Espinosa, C. A. Reyes-García, and L. Villaseñor-Pineda, “Acoustic feature selection and classification of emotions in speech using a 3D continuous emotion model,” Biomed. Signal Process. Control, vol. 7, no. 1, 2012, pp. 79–87.

[۱۳] علی حریمی، علیرضا احمدی فرد، علی شهزادی و خشایار یغمایی، "تشخیص احساس از روی گفتار با استفاده از طبقه‌بند مبتنی بر مدل و ویژگی های دینامیکی غیر خطی"، نشریه مهندسی برق و مهندسی کامپیوتر ایران، دوره ۱۵، شماره ۲، تابستان ۱۳۹۶، صفحه ۱۵۲-۱۴۵.

[14] K. Han, D. Yu, and I. Tashev, Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine., vol. 3, 2014, pp. 232–243.

[15] H. Palo and M. Mohanty, “Modified-VQ Features for Speech Emotion Recognition,” Journal of Mathematical Sciences, vol. 16,Sep 2016, pp. 406–418.

[16] B. Schuller, R. Müller, M. Lang, and G. Rigoll, Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles., vol. 2, 2005, pp. 565–572.

[17] I. Luengo, E. Navas, and I. Hernáez, “Feature Analysis and Evaluation for Automatic Emotion Identification in Speech,” Multimedia, IEEE transaction, vol. 12, Nov 2010, pp. 490–501.

[18] D. Gharavian, M. Sheikhan, and F. Ashoftedel, “Emotion recognition improvement using normalized formant supplementary features by a hybrid of DTW-MLP-GMM model,” Neural Computing & Applications, vol. 22, no. 6, 2013, pp. 1181–1191.

[19] X. Zhao, S. Zhang, and B. Lei, “Robust emotion recognition in a noisy speech via sparse representation,” Neural Computing & Applications, vol. 24, Jun. 2013.

[20] H. Hu, T. Tan, and Y. Qian, “Generative adversarial network-based data augmentation for noise-robust speech recognition,” in Proc. The international Conference on Acoustics, Speech, & Signal Processing (ICASSP), Apr 2018, pp. 5044–5048.

[21] A. Harimi and Kh. Yaghmaie, “improving speech emotion recognition via gender classification,” in Journal of Modeling in Engineering., vol.48, 2017, pp. 184–200.

[22] M. Sadeghi, H. Marvi and A. Ahmadifard “A New and Efficient Feature Extraction Method for Robust Speech Recognition Based on Fractional Fourier Transform and Differential Evolution Optimizer,” in Journal of Modeling in Engineering., vol.61, 2020, pp. 86–96.

[۲۳] سیدعلی سلیمانی ایوری، محمدرضا فدوی امیری و حسین مروی، "تولید سیگنال مصنوعی زلزله به‌کمک مدلی جدید در فشرده‌سازی و آموزش شبکه‌های عصبی مصنوعی"، نشریه مدلسازی در مهندسی، دوره ۱۴، شماره ۴۶، پائیز ۱۳۹۵، صفحه ۸۵-۷۵.

[24] M. Chourasia, S. Haral, , S. Bhatkar, and S. Kulkarni, “Emotion Recognition from Speech Signal Using Deep Learning.” In Intelligent Data Communication Technologies and Internet of Things, 2021, pp. 471–481.

[25] J. Chang, S. Scherer. “Learning representations of emotional speech with deep convolutional generative adversarial networks”. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.

[26] F. Eyben, F. Weninger, F. Gross and B. Schuller. “Recent developments in openSMILE, the Munich open-source multimedia feature extractor”. In: Proc. 21st ACM international conference on Multimedia. ACM, vol. 5, 2013, pp. 232–240.

[27] I. Goodfellow. “NIPS 2016 tutorial: Generative adversarial networks”. In: arXiv preprint arXiv:1701.00160, 2016.

[28] F. Bao, M. Neumann, and N. T. Vu, “Cyclegan-based emotion style transfer as data augmentation for speech emotion recognition.” In Interspeech, 2019, pp. 2828–2832.

[29] W. Y. Zhao, “Discriminant component analysis for face recognition,” in Proceeding’s 15th International Conference on Pattern Recognition. ICPR-2000, vol. 2, 2000, pp. 818–821.

[30] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of German emotional speech,” in Proc. 9th European Conference on Speech Communication and Technology , 2005, pp 1–4.

[31] B. Yang and M. Lugger, “Emotion recognition from speech signals using new harmony features,” Signal Processing, vol. 90, no. 5, 2010, pp. 1415–1423.

[32] I. Luengo, E. Navas, and I. Hernáez, “Feature Analysis and Evaluation for Automatic Emotion Identification in Speech,” Multimedia, IEEE transaction, vol. 12, Nov 2010, pp. 490–501.

[33] G. Paraskevopoulos, E. Tzinis, N. Ellinas, T. Giannakopoulos, and A. Potamianos, “Unsupervised low-rank representations for speech emotion recognition,” Proc. Interspeech 2019, 2019, pp. 939–943.

[34] S. Sahu, R. Gupta, and C. Espy-Wilson, “On enhancing speech emotion recognition using generative adversarial networks,” Proc. Interspeech 2018, 2018, pp. 3693–3697.

[35] S. Latif , M. Asim, R. Rana, S. Khalifa, R. Jurdak and B.W. Schuller, “Augmenting Generative Adversarial Networks for Speech Emotion Recognition, ” Proc. Interspeech, 521-525, doi: 10.21437/Interspeech.2020-3194, 2020.