کومش- جلد ۲۵، شماره ۶ (پیاپی ۹۸)، آذر و دی ۱۴۰۲

چالش‌ها و راه‌حل‌هایی در جمع‌آوری داده و ارزیابی مدل‌ها در یادگیری ماشین نظارت شده، مطالعه مروری

سعیده علی‌اکبری [۱*](Ph.D)، پیمان حجازی [۲](Ph.D)، زینب هرمزی مقدم [۳و۴](Ph.D)

۱- گروه پرتوشناسی، دانشکده پیراپزشکی، دانشگاه علوم پزشکی سمنان، سمنان، ایران

۲- گروه فیزیک پزشکی، دانشکده پزشکی، دانشگاه علوم پزشکی سمنان، سمنان، ایران

۳- مرکز تحقیقات بیولوژی پرتو، دانشگاه علوم پزشکی ایران، تهران، ایران

۴- گروه علوم پرتویی، دانشکده پیراپزشکی، دانشگاه علوم پزشکی ایران، تهران، ایران

* نویسنده مسئول، تلفن: ۰۹۳۵۴۴۷۰۱۱۲ s.aliakbari@semums.ac.ir

تاریخ دریافت: ۱۴۰۲/۵/۱ تاریخ پذیرش: ۱۴۰۲/۱۲/۷

چکیده

هدف: هدف اصلی یادگیری ماشین یک فرآیند پیچیده است که از طریق تعیین مدل و آموزش آن با استفاده از حجم زیادی از داده‌ها، انجام می‌شود. در گذشته، تمرکز اصلی در این زمینه بیش‌تر بر روی بهبود ساختار مدل‌ها و الگوریتم‌ها بوده است، اما اخیراً تمرکز بهتری به سمت کیفیت و کمیت داده‌ها صورت گرفته است. هدف از این مقاله‌ی مروری بررسی چالش‌ها در جمع‌آوری داده‌ها و ارزیابی مدل در یادگیری ماشین نظارت شده و ارائه‌ی راه حل برای آن است.

مواد و روش‌ها: در این مطالعه چالش‌های پیش روی محققان جهت جمع‌آوری داده و ارزیابی مدل‌های یادگیری ماشین نظارت شده به روش مطالعه‌ی مروری مورد بررسی قرار گرفت، مستندات از پایگاه‌های مطالعاتی PubMed، Scopus، Science Direct و موتور جست‌وجو Google Scholar در بازه‌ی زمانی ۲۰۰۱ الی ۲۰۲۳ بازیابی شد که پس از غربالگری متن کامل ۱۷ مقاله بررسی و به مطالعه وارد شد.

یافته‌ها: در بررسی مطالعات انجام شده چهار چالش عمده در جمع‌آوری داده‌ها در حیطه‌ی یادگیری ماشین نظارت شده که عبارتند از: تعداد ناکافی نمونه، داده‌های آموزشی غیر نماینده، کیفیت پایین داده و ویژگی‌های غیر مرتبط یافت شد. در ارزیابی مدل نیز با چهار چالش که عبارتند از: بیش برازش، کمبود برازش، در دسترس نبودن داده کافی جهت اعتبارسنجی و عدم تطبیق داده‌ها به‌دست آمد.

نتیجه‌گیری: افزایش تعداد نمونه، استفاده از الگوریتم انتخاب تصادفی، پاک‌سازی داده، استفاده از آزمون آماری صحیح، انتخاب ویژگی، استخراج ویژگی، استفاده از مدل ساده‌تر، تکنیک K-fold و پردازش داده‌ها از جمله مواردی است که رعایت آن باعث دست‌یابی به مدلی با عملکرد بهتر می‌شود.

واژه‌های کلیدی: یادگیری ماشین نظارت شده، جمع‌آوری داده، ارزیابی مدل

مقدمه

استفاده از هوش مصنوعی از دهه‌ی ۱۹۵۰ با این سوال مطرح شد که "آیا می‌توان کامپیوترهایی ساخت که فکر کنند؟ پس از آن قوانینی به صورت هزاران خط کد نوشته شد تا در حوزه‌هایی مانند پردازش تصویر و مسیریابی استفاده شود، به این نوع رویکرد، هوش مصنوعی نمادین (Symbolic Artificial Intelligence) گفته می‌شود. مسائلی که در حل آن‌ها از هوش مصنوعی نمادین استفاده می‌شد مسائلی بودند که حل آن‌ها برای انسان دشوار بود و با نوشتن یکسری قوانین ریاضی و رسمی به‌وسیله کدها می‌توان آن‌ها را حل کرد [۱]. چالش هوش مصنوعی زمانی آغاز شد که نوشتن کد، برای حل بعضی مسائل، بسیار دشوار بود لذا برای حل این مشکل، مدل‌هایی ارائه شد تا به‌واسطه‌ی یکسری داده‌ها آموزش ببینند تا بهترین الگوریتم را یاد بگیرند و بر اساس الگوریتمی که یاد گرفته‌اند بتوانند پیش‌بینی‌های درستی داشته باشند. پس از آن یادگیری ماشین (Machin Learning) و یادگیری عمیق (Deep Learning) که زیر مجموعه‌ای از هوش مصنوعی می‌باشند معرفی گردید [۲].

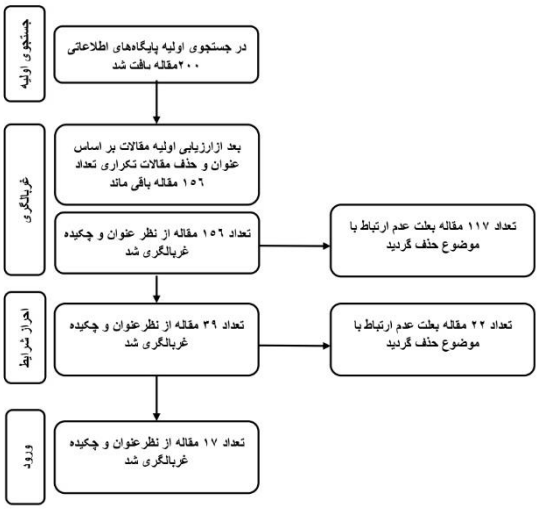
هدف اصلی یادگیری ماشین یک فرآیند پیچیده‌ای است که از طریق تعیین مدل و آموزش آن با استفاده از حجم زیادی از داده‌ها، انجام می‌شود. در گذشته، تمرکز اصلی در این زمینه

بیش‌تر بر روی بهبود ساختار مدل‌ها و الگوریتم‌ها بوده است، اما اخیراً تمرکز بهتری به سمت کیفیت و کمیت داده‌ها صورت گرفته است. این نکته از این جهت ضروری است که داده‌هایی با کیفیت و حجم بیش‌تر، مدل‌های پیچیده‌تر با توانایی‌های یادگیری بیش‌تری را ارائه می‌دهند. از این‌رو، هم‌اکنون به دست آوردن داده‌های بهتر و بیش‌تر، به‌عنوان یکی از اهداف اصلی در یادگیری ماشین مورد توجه قرار گرفته است. با افزایش کمیت و کیفیت داده‌ها، مدل‌های پیشرفته‌تر و دقیق‌تری می‌توانند آموزش ببینند و در نتیجه، عملکرد و کارایی سیستم‌های یادگیری ماشین بهبود می‌یابد [۳-۶].

یادگیری ماشین بر اساس نوع داده‌ها، رویکرد آموزش، و هدف نهایی کاربرد مدل‌ها به چند دسته کلی که عبارتند از: یادگیری نظارت شده، یادگیری نظارت نشده، یادگیری تقویتی، یادگیری نیمه نظارت شده، یادگیری فعال تقسیم‌بندی می‌شود. جمع‌آوری داده برای هر کدام از دسته‌بندی‌های یادگیری ماشین که ذکر شده است می‌تواند تفاوت‌هایی داشته باشد [۷]. محققانی که از یادگیری نظارت شده استفاده می‌کنند با چالش‌هایی مانند تعداد ناکافی داده‌ها، نویز نمونه‌برداری، بایاس نمونه‌برداری، داده‌های آموزشی غیر نماینده، کیفیت داده‌ها، ویژگی‌های غیر مرتبط، انتخاب ویژگی (Feature Selection) و استخراج ویژگی (Feature Extraction) مواجه هستند. در این مطالعه، سعی شده است چالش‌های پیش روی محققان جهت جمع‌آوری داده‌ها و ارزیابی مدل‌ها در حوزه‌ی یادگیری ماشین نظارت شده مورد بررسی قرار گیرد و راه حل‌هایی برای این چالش‌ها معرفی گردد.

مواد و روش‌ها

استراژدی جست‌وجو: به منظور دسترسی به مطالعات مرتبط با کمیت و کیفیت داده‌ها در یادگیری ماشین نظارت شده جست‌وجو در پایگاه‌های پاب مد (Pubmed)، و اسکوپوس (Scopus) ساینس دایرکت (Science Direct) و موتور جست‌وجو گوگل اسکالر در بازه‌ی زمانی ۲۰۰۱ تا ۲۰۲۳ انجام شد. جست‌وجو با استفاده از کلید واژه‌های "Machine Learning"، "Data"، "Quality" و "Quantity" برای بازیابی مقالات به زبان انگلیسی انجام گردید. کلید واژه‌ها با استفاده از گیومه و یا بدون استفاده از آن جست‌وجو و در صورت نیاز از عملگرهای بولی "AND" و "OR" استفاده شده است. تمام مقالاتی که معیار ورود داشتند مورد ارزیابی قرار گرفتند، غربالگری مطالعات بر اساس معیارهای راهنمای پریسما (PRISMA) انجام شد. (شکل ۱).



شکل ۱. مراحل انتخاب ورود مقاله به مطالعه

معیارهای ورود و خروج و انتخاب مطالعه: از آن‌جا که هدف این مطالعه بررسی چالش داده‌ها در یادگیری ماشین نظارت شده می‌باشد، تمام مقالات منتشر شده حاصل از جست‌وجو در مجلات علمی که تاثیر کمیت و کیفیت داده‌ها در یادگیری ماشین و مدل‌های آن را مورد ارزیابی قرار داده بود جمع‌آوری گردید در مرحله بعد یک نفر از محققین چکیده مقالات را مورد ارزیابی قرار داده و مقالات تکراری و مقالاتی که فقط بر روی داده‌هایی خاص ارزیابی انجام شده بود را شناسایی و از مطالعه خارج کردند در آخر مقالاتی وارد مطالعه شدند که به‌صورت کلی تاثیر کیفیت و کمیت داده‌ها بر یادگیری ماشین نظارت شده را مورد بررسی قرار می‌داد. برای پرهیز از تعصب مقالات توسط محقق سوم نیز مورد ارزیابی قرار گرفت سپس متن تمام مقالاتی که معیار ورود به مطالعه را داشتند مورد ارزیابی قرار گرفت.

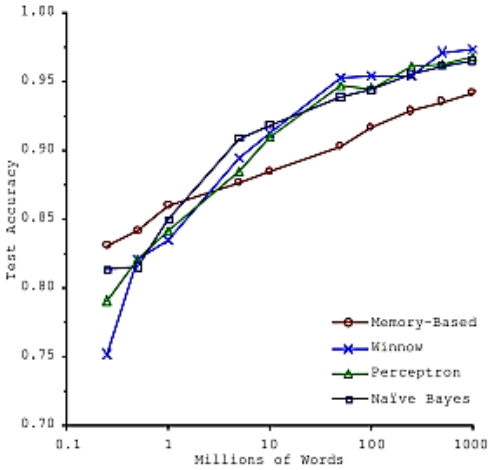
نتایج

طبق مستندات مورد ارزیابی قرار گرفته، چالش‌های جمع‌آوری داده‌ها در حوزه‌ی یادگیری ماشین به صورت زیر دسته‌بندی می‌شوند.

الف) تعداد ناکافی داده‌ها

کیفیت و حجم داده‌ها در یادگیری ماشین نظارت شده نقش بنیادی ایفا می‌کند. داده‌های ناکافی می‌توانند به طور قابل توجهی بر دقت و کارایی مدل‌های یادگیری ماشین تأثیر بگذارند، به ویژه در موقعیت‌هایی که دقت حیاتی می‌باشد، در این عرصه، مدل‌هایی که با مجموعه داده‌های محدود آموزش دیده‌اند، ممکن است نتوانند الگوهای پیچیده را تشخیص دهند یا به نمونه‌هایی که نادرند ولی از اهمیت بالایی برخوردارند، تعمیم یابند. این مشکلات می‌توانند منجر به افزایش خطاهای

نوع اول و دوم شوند، که بیانگر به اشتباه رد کردن یا تأیید یک فرضیه می‌باشد. بنابراین، درک و بهبود چگونگی تأثیرگذاری داده‌های ناکافی بر مدل‌ها می‌تواند کمک کند تا استراتژی‌های جمع‌آوری داده، انتخاب ویژگی، و طراحی مدل‌هایی را که مقاومت بیش‌تری در برابر محدودیت‌های داده‌ای دارند را بهبود بخشید. با مجموعه داده‌های کافی و با کیفیت مدل‌های نسبتاً ساده هم می‌توانند عملکرد قابل قبولی در مسئله‌های پیچیده ارائه دهند. (شکل ۲) [۸]



شکل ۱ اهمیت داده ها بر حسب مدل[۸]

ب) نویز نمونه‌برداری:

در صورتی که تعداد داده‌ها کم باشد، ممکن است داده‌های انتخابی نماینده‌ی جامعه‌ی مورد مطالعه نباشند. این مسئله باعث می‌شود که مدل ارائه شده نتواند پیش‌بینی صحیحی از داده‌های دیگر (داده‌های آزمون) داشته باشد [۹].

ج) بایاس نمونه‌برداری:

هم‌چنین، اگر داده‌ها به صورت غیر تصادفی انتخاب شوند، ممکن است نمونه‌های مشابه یا تکراری در مجموعه داده وجود داشته باشد که باعث افزایش بایاس در مدل می‌شود و نتایج آن ناعادلانه می‌گردد.

بایاس در نمونه‌برداری یک نوع مغایرت یا گرایش سیستماتیک است که در فرآیند جمع‌آوری یا انتخاب نمونه‌هایی از یک جمعیت بزرگ‌تر رخ می‌دهد و موجب می‌شود که نمونه برگزیده شده تمام جنبه‌های جمعیت مورد نظر را به طور دقیق و بی‌طرفانه نمایندگی نکند. این حالت می‌تواند باعث شود که نتایج حاصل از تحلیل داده‌های موجود متأثر از این بایاس‌ها شود و در نتیجه پیش‌بینی‌ها، تعمیم‌ها، یا نتیجه‌گیری‌های نادرستی را در پی داشته باشد. برای جلوگیری از بایاس در نمونه‌برداری، محققان باید از روش‌های نمونه‌برداری احتمالی و تصادفی استفاده نمایند که به هر عضو جمعیت شانس مساوی برای انتخاب شدن بدهد [۱۰]. به عنوان مثال، در یادگیری ماشین، اگر تعداد داده‌های مثبت (مثلاً داده‌هایی که یک حالت معین را نشان می‌دهند) نسبت به تعداد داده‌های منفی (مثلاً داده‌هایی که حالت معین را نشان نمی‌دهند) کم باشد، ممکن است مدل بیش‌تر به انتخاب داده‌های مثبت تمایل داشته باشد و از انتخاب داده‌های منفی کاسته شود که به نتایج نادرست و ناعادلانه منجر می‌شود [۱۱].

د) داده‌های آموزشی غیر نماینده (Nonrepresentative Training Data)

داده‌های غیرنماینده به مجموعه‌ای از داده‌های آموزشی اشاره دارد که دارای ویژگی‌هایی هستند که مستقیماً مرتبط با خروجی (یا هدف) نیستند، به عبارت دیگر، این ویژگی‌ها نباید به صورت مستقیم در فرآیند مدل‌سازی مورد استفاده قرار گیرند. مثالی از این موضوع می‌تواند بررسی رنگ موی زنان با سرطان سینه باشد. در این مثال، رنگ مو به طور مستقیم با سرطان سینه مرتبط نیست و اگر در فرآیند مدل‌سازی این ویژگی‌ها در نظر گرفته شود، مدل ارائه شده نمی‌تواند در آینده به داده‌های دیگر تعمیم داده شود و در ارزیابی با داده‌های آزمون، دقت بسیار پایینی خواهد داشت. به همین دلیل، استفاده از ویژگی‌های غیرنماینده در مدل‌سازی ممکن است باعث ایجاد تبعیض و عملکرد نامناسب مدل گردد. برای جلوگیری از این مشکل، لازم است دقت کافی به جمع‌آوری داده‌های مناسب و نماینده اختصاص داده شود تا مدل بتواند به صورت عمومی و کلی به داده‌های جدید تعمیم یابد و عملکرد مناسبی در مواجهه با داده‌های آزمون ارائه دهد [۱۲].

ه) کیفیت داده‌ها

وجود داده‌های پرت می‌تواند تشخیص الگو را برای سیستم سخت‌تر کند. داده‌های پرت به داده‌هایی اشاره دارند که به‌طور ناخواسته یا اشتباهی از نمونه‌های صحیح دیگر تولید شده‌اند و ممکن است نماینده‌ی درستی از واقعیت نباشند. حضور داده‌های پرت می‌تواند باعث شود که مدل‌ها به الگوهای نادرستی در داده‌ها پاسخ دهند و نتایج غیرقابل اعتمادی را تولید کنند. این امر به خصوص زمانی رخ می‌دهد که داده‌های پرت در مجموعه‌ی داده‌های آموزشی وجود دارند و مدل بر اساس این داده‌ها آموزش می‌بینند. برای پیشگیری و رفع این مشکل، انجام پاک‌سازی داده‌ها از داده‌های پرت پیشنهاد می‌گردد. به‌عبارت دیگر، با حذف داده‌های پرت از مجموعه‌ی داده‌های آموزشی، می‌توان تأثیر مخرب آن‌ها بر روی کیفیت و عملکرد مدل‌ها را کاهش داد. علاوه بر این استفاده از تکنیک‌های پیشرفته‌تری برای پاک‌سازی داده‌ها می‌تواند بهبود چشم‌گیری در دقت و قدرت پیش‌بینی مدل‌ها به همراه داشته باشد. در این راستا، استفاده از سرویس‌ها و

ابزارهای پاک‌سازی داده مانند "Clean Data" می‌تواند به عنوان یک ابزار مفید برای تصفیه و بهبود کیفیت داده‌ها استفاده گردد. این نوع ابزارها می‌توانند به صورت خودکار و با استفاده از الگوریتم‌ها و تکنیک‌های پیشرفته، داده‌های پرت را تشخیص دهند و از مجموعه‌ی داده‌ها حذف کنند [۱۳–۱۵].

و) ویژگی‌های غیر مرتبط (Irrelevant Feature)

گاهی اوقات ویژگی‌هایی که توسط محقق با استدلال و دانش دقیق به هدف مرتبط تعیین می‌شوند، در مدل‌سازی جز ویژگی‌های غیرمرتبط به نظر می‌آیند. این موضوع می‌تواند به دلیل عدم هم‌بستگی دقیق بین ویژگی‌ها و خروجی مورد نظر باشد یا به دلیل وجود اثرات تصادفی و نویز در داده‌ها باشد که باعث تحلیل نادرست ویژگی‌ها شده‌اند [۱۶].

پس از جمع‌آوری داده‌ها و حل مسائل مرتبط با آن، مرحله‌ی بعد تعیین مدل و آموزش آن می‌باشد. چالش‌های ارزیابی مدل عبارتند از:

الف– تعیین مدل:

مدل‌های یادگیری ماشین شامل طبقه‌بندی (Classification)، رگرسیون (Regression)، خوشه‌بندی (Clustering) شبکه‌های عصبی (Neural Network)، ماشین‌های بردار پشتیبان (Support Vector Machines) و غیره می‌باشند. انتخاب مدل صحیح باعث می‌شود که سیستم بهترین عملکرد را در پیش‌بینی و تحلیل داده‌های آزمون ارائه دهد. آموزش مدل با استفاده از داده‌های جمع‌آوری شده بسیار حائز اهمیت است. مرحله‌ی آموزش، به مدل کمک می‌کند تا الگوها و ویژگی‌های مرتبط با داده‌های ورودی و خروجی آموزشی را یاد بگیرد. انتخاب داده‌های آموزشی (Train)، اعتباربخشی (Validation)، آزمون (Test)، استفاده از تکنیک‌های بهینه‌سازی و اعمال روش‌های مناسب برای جلوگیری از بیش برازش (Over Fitting) از جمله موارد مهم در آموزش مدل‌ها می‌باشد. پس از آموزش مدل، عملکرد آن با استفاده از داده‌های آزمون ارزیابی می‌شود تا مشخص شود که مدل به خوبی در مسئله‌ی مورد نظر عمل می‌کند یا نیاز به بهبود دارد [۱۷–۲۱].

در پیش‌بینی و تحلیل داده‌ها چالش مواجه با انتخاب مدل عبارتند از:

الف– ۱) بیش برازش

بیش‌برازش هنگامی اتفاق می‌افتد که مدل به طور غیرمنطقی به داده‌های آموزشی تطبیق یابد، این امر باعث می‌شود دقت مدل در داده‌های آموزشی بسیار بالا باشد اما دقت مدل بر روی داده‌های آزمون کم باشد. برخی از دلایل ایجاد بیش‌برازش عبارتند از:

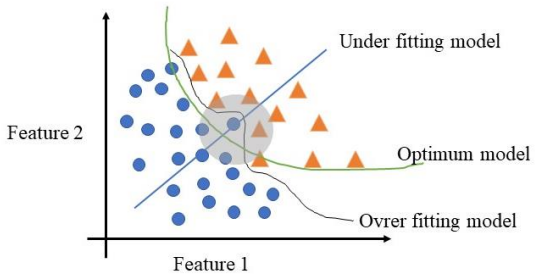
- **پیچیدگی مدل:** استفاده از مدل‌های بسیار پیچیده با تعداد زیادی پارامتر، احتمال بیش‌برازش را افزایش می‌دهد. این مدل‌ها ممکن است بتوانند نویز و اطلاعات تصادفی موجود در داده‌ها را هم دقیقاً یاد بگیرند که منجر به عدم تعمیم‌پذیری به داده‌های جدید می‌شود.
- **کمبود داده‌ها:** هنگامی که داده‌های آموزشی کم باشند، مدل تمایل دارد به‌طور زیادی به این داده‌ها متمایل شود و برازش دقیقی به آن‌ها داشته باشد. به عبارت دیگر، با تعداد کم نمونه‌ها، مدل ممکن است به اطلاعات نمونه‌های خاصی بسنده کند و نتواند الگوهای کلی‌تر را یاد بگیرد [۲۲].

الف– ۲) کمبود برازش (Under Fitting)

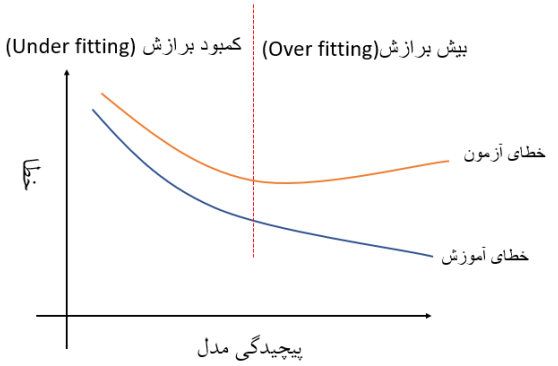
اگر مدل به صورت بسیار ضعیف عمل کند منحنی آبی (شکل ۳) به این معنی است که به داده‌های آموزشی منطبق نیست و نمی‌توان این مدل را به داده‌های دیگر تعمیم داد و مدل مورد استفاده به طور کلی نتوانسته است الگوها و اطلاعات مرتبط با داده‌ها را یاد بگیرد و بنابراین هیچ‌کدام از داده‌های آموزشی و تست را به خوبی پیش‌بینی نکرده است. برای مقابله با این مشکلات، می‌توان اقداماتی انجام داد که به مدل کمک کند بهتر به داده‌ها تطبیق پیدا کند و نتایج بهتری را ارائه دهد. این اقدامات شامل:

- **افزایش داده‌ها:** در صورت امکان، با تولید داده‌های مصنوعی یا استفاده از تکنیک‌های افزایش داده‌ها می‌توان به بهبود کیفیت مدل کمک کرد.
- **انتخاب مدل مناسب:** انتخاب مدل مناسب و متناسب با مسئله‌ی مورد نظر می‌توانند به خوبی به خصوصیات داده‌ها و نوع مسئله انطباق پیدا کند [۲۳].

شکل ۲ نمونه ای از مدل Under Fitting،Over fitting و m



با توجه به موارد ذکر شده در یادگیری ماشین اگر داده‌ها کم باشد و در نتیجه مدل ساده‌ای انتخاب شود خطای مدل هنگام ارزیابی با داده‌های آموزش و آزمون زیاد خواهد بود بنابراین مدل دچار کمبود برازش شده است و اگر سعی در کاهش مقدار خطای مدل بر اساس داده‌های آموزش باشد ممکن است دچار بیش برازش شده و در نهایت مدل در پیش‌بینی داده‌های آزمون دچار خطای زیادی شود. شکل (۴) [۲۵،۲۴]



شکل ۳. نمودار خطای مدل بر حسب داده های آموزش و آزمون بر حسب پیچیدگی مدل

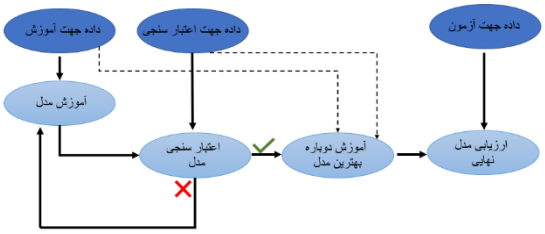
ب) ارزیابی مدل‌های یادگیری ماشین

هدف از ارزیابی مدل‌های یادگیری ماشین، تعمیم دادن مدل به داده‌های آزمون است. برای این هدف داده‌ها را به سه دسته تقسیم می‌شود.

- داده‌های آموزشی (Training Data)
- داده‌های اعتبارسنجی (Validation Data)
- داده‌های آزمون (Test Data)

طبق مطالعات بررسی شده برای تقسیم داده‌ها جهت آموزش و آزمون از روش ۲۰-۸۰ استفاده می‌شود به این صورت که ۸۰ درصد داده‌ها جهت آموزش مدل و ۲۰ درصد جهت آزمون مدل و جهت اعتبارسنجی ۲۰ درصد از داده‌های آموزش در نظر گرفته می‌شود. جهت ارزیابی و انتخاب مدل در یادگیری ماشین، باید مراحل اعتبارسنجی و ارزیابی دقت را انجام داد. این مراحل اجازه می‌دهد تا از جنبه‌ی تعمیم‌پذیری مدل اطمینان و بهترین مدل انتخاب شود.

برای این منظور، داده‌ها به دو بخش داده‌های آموزش و داده‌های اعتبارسنجی تقسیم می‌شود. مدل بر روی داده‌های آموزش، آموزش داده می‌شود و عملکرد آن روی داده‌های اعتبارسنجی ارزیابی می‌شود. این مراحل تا زمانی ادامه می‌یابد که مدل دارای بهترین دقت روی داده‌های اعتبارسنجی را دارا باشد. پس از انتخاب بهترین مدل، می‌توانیم آن را روی مجموعه آزمون اجرا کرده و عملکرد نهایی مدل ارزیابی شود (شکل ۵). لازم به ذکر است که در یک مطالعه بر مبنای یادگیری ماشین باید دقت عملکرد مدل بر اساس داده‌های آزمون در یک مقاله گزارش شود. این روش (اعتبارسنجی مدل) به محقق اجازه می‌دهد تا از بیش‌برازش جلوگیری شود، زیرا مدل با توجه به ارزیابی بر روی مجموعه اعتبارسنجی بهترین تطابق را با داده‌ها نشان می‌دهد و برای داده‌های جدید نیز به خوبی کار می‌کند. در این روش، دقت مدل روی مجموعه اعتبارسنجی یا داده‌های ارزیابی به عنوان معیار انتخاب بهترین مدل استفاده می‌شود و نباید از دقت مدل روی مجموعه آموزش برای گزارش عملکرد مدل استفاده کرد.



شکل ۴. دیاگرامی از یادگیری ماشین بر اساس بررسی چندین مدل

انتخاب داده‌های آزمون به طور تصادفی از مجموعه داده‌ها به عنوان معمول‌ترین روش استفاده می‌شود. هدف اصلی این انتخاب اطمینان حاصل کردن از تعمیم‌پذیری مدل به داده‌های جدید است. با انتخاب داده‌های آزمون به صورت تصادفی از مجموعه داده‌ها، می‌توانیم اطمینان حاصل کنیم که داده‌های آزمون نماینده‌ترین نمونه‌ها از مجموعه داده هستند و به طور عادلانه انتخاب شده‌اند. این انتخاب تصادفی اجازه می‌دهد که مدل را بر روی داده‌هایی آزمایش کرد که مدل هنوز آشنا نیست و در نتیجه، دقت و یا عملکرد نهایی مدل به عنوان معیاری برای تعمیم‌پذیری آن به داده‌های جدید دانست.

به این ترتیب، داده‌های اعتبارسنجی نقش بسیار مهمی در ارزیابی و اعتبارسنجی مدل‌ها دارند و باید با دقت و انصاف انتخاب شوند تا نتایج ارزیابی مدل‌ها را به درستی نشان دهند که مدل چه میزان به داده‌های جدید تعمیم‌پذیر است.

لازم به ذکر است که مرحله‌ای اعتبارسنجی در یک مطالعه هنگامی انجام می‌شود که قرار است از بین چندین مدل بهترین آن انتخاب شود، در صورتی که محقق فقط یک مدل را مورد آموزش قرار داده است دیگر نیازی به مرحله اعتبارسنجی نمی‌باشد [۲۷،۲۶].

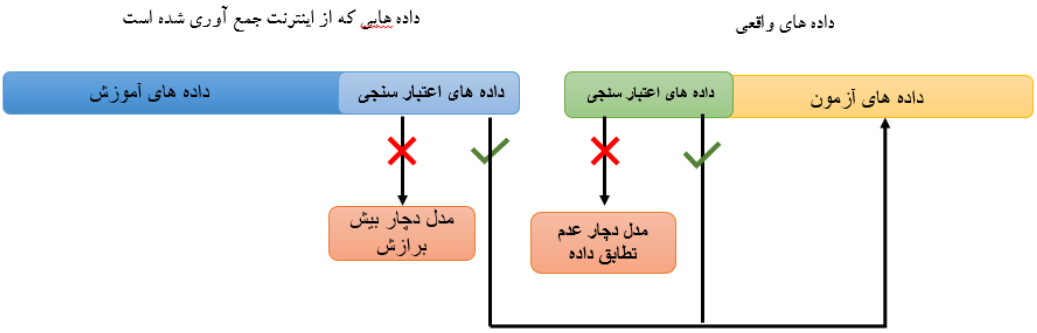
ب-۱) کمبود داده جهت آموزش مدل:

در صورتی که محقق تمایل داشته باشد چندین مدل را در یک مطالعه‌ای که حجم داده‌ها کم است ارزیابی (برای مثال ۱۰۰۰ داده) کند برای اعتبارسنجی نیاز است به روش K-fold

عمل شود. اگر داده‌های کمی در اختیار باشد، در صورتی‌که در چند مرحله به‌طور تصادفی ۸۰درصد داده‌ها برای آموزش و ۲۰درصد برای اعتبارسنجی مدل انتخاب شود، دقت اعتبارسنجی مدل‌ها در هر بار آموزش با داده‌های تصادفی دارای واریانس بالایی می‌باشد که نشان‌دهنده‌ی این است که داده‌ها برای اعتبارسنجی مدل کافی نمی‌باشد، یکی از راه‌حل‌های کاهش اختلاف دقت در ارزیابی مدل‌ها افزایش تعداد داده‌ها است که در بیش‌تر مواقع به دلیل صرف زمان و هزینه‌ی زیاد امکان افزایش داده مقدور نیست برای حل این مشکل از اعتباربخشی k-fold (K-fold cross validation) استفاده می‌شود. در این نوع اعتباربخشی همه‌ی داده‌ها در اعتبارسنجی به روش زیر مشارکت می‌کنند.

K نماینده‌ی این است که داده‌ها به چند دسته تقسیم خواهند شد. برای توضیح این روش به‌طور مثال k=۴ در نظر گرفته می‌شود. بنابراین داده‌های آموزش به ۴ دسته تقسیم می‌شود و یک مدل یکسان چهار مرتبه با داده‌های اعتبارسنجی مختلف ارزیابی می‌شود لازم به ذکر است که کل داده‌ها در اعتبارسنجی شرکت داده شده و به‌عنوان اعتبارسنجی استفاده می‌شود. برای بیان دقت مدل از ۴ دقت به‌دست آمده (x,y,z,t) میانگین‌گیری کرده و میانگین آن را به‌عنوان دقت مدل به‌وسیله‌ی این نوع اعتبارسنجی ذکر می‌شود. (شکل ۶)

شکل ۵ دیاگرام اعتبار سنجی باk-fold برای دو مدل مجزا



شکل ۶ دیاگرامی از چالش های احتمالی هنگام استفاده کردن از داد های واقعی جهت آزمون مدل

حال برای مقایسه این مدل با مدل‌های دیگر (با همین تعداد داده) کافی است همین روش اعتبارسنجی را برای مدل دیگر تکرار شود و با مقایسه‌ی میانگین دقت دو مدل (c , p) بهترین مدل انتخاب می‌شود. سپس دو سوال مطرح است؟ مدل انتخابی را با چه پارامتری گزارش می‌شود و چه دقتی برای داده‌های اعتبارسنجی در نظر گرفته می‌شود؟

بعد از انتخاب بهترین مدل، باید مدل را با کل داده‌ها آموزش داد بعد از آموزش با کل داده‌ها پارامتر مدل انتخابی گزارش می‌شود.

در نهایت که مدل به‌وسیله‌ی کل داده‌ها آموزش داده شد مدل به‌وسیله داده‌های آزمون مورد ارزیابی قرار می‌گیرد و دقت به‌دست آمده در مطالعه گزارش می‌شود [۲۸-۳۰].

ج) عدم تطبیق داده‌ها: (Mismatch Data)

در حوزه یادگیری ماشین جهت آموزش مدل مورد مطالعه از بانک داده‌های مانند سایت Kaggle استفاده و مدل توسط آن داده‌ها آموزش داده می‌شود اما داده‌های آموزش احتمالاً نمایانگر داده‌های واقعی نمی‌باشند لذا دقت مدل باید بر اساس ارزیابی به‌وسیله‌ی داده‌های واقعی گزارش شود.

اگر دقت اعتبارسنجی روی داده‌های واقعی خوب نباشد دو مشکل برای مدل مطرح می‌شود:

۱) مدل دچار بیش برازش شده است

۲) در داده‌ها عدم تطبیق وجود دارد

برای یافتن این‌که کدام یک از موارد بالا در مطالعه باعث دقت پایین در اعتباربخشی شده است، دو راه‌حل وجود دارد [۳۱].

ابتدا قسمتی از داده‌های آموزشی و واقعی جهت اعتبارسنجی مدل انتخاب می‌شود، اگر دقت اعتبارسنجی مدل با استفاده از داده‌های آموزشی هم‌چنان پایین بود مطمئناً بیش برازش اتفاق افتاده است، و باید از راه‌حل‌های رفع بیش برازش استفاده کرد. اما اگر دقت مدل بر روی داده‌های واقعی که برای اعتبارسنجی استفاده شده بود، پایین بود عدم تطبیق داده اتفاق افتاده است. عدم تطبیق داده یعنی ویژگی‌هایی که داده‌ها با آن آموزش دیدند با ویژگی‌های داده‌های اعتبارسنجی تطابق ندارند.

در حوزه‌ی یادگیری ماشین، جمع‌آوری داده‌هایی با ویژگی‌های با ارزش و هم‌چنین توانایی ایجاد مدل‌های دقیق و مفید، یکی از چالش‌های مهم و اساسی است که محققان با آن روبرو می‌شوند. در این مطالعه چالش‌های پیش روی محققان جهت جمع‌آوری داده‌ها در جدول ۱ و ارزیابی مدل‌ها در حوزه‌ی یادگیری ماشین نظارت شده در جدول ۲ ارائه شده است، در این جدول‌ها چالش‌ها، علت ایجاد و راه‌حل‌هایی برای رفع آن معرفی شده است.

جدول ۱. چالش های جمع آوری داده‌ها در حوزه یادگیری ماشین نظارت شده، علت ایجاد و راه حل هایی جهت رفع آن

چالش	علت ایجاد آن	راه حل
نویز	تعداد ناکافی(کم) نمونه[۳]	افزایش تعداد نمونه در صورت نبود نمونه واقعی استفاده از بانکِ داده های موجود در اینترنت مانند Kaggle [۹]
بایاس	تعداد ناکافی نمونه به شکلی که انتخاب نمونه متمرکز به یک شرایط خاص باشد[۱۱]	استفاده از الگوریتم های انتخاب تصادفی [۱۰]
داده‌های آموزشی غیر نماینده	عدم تسلط محقق به موضوع مورد مطالعه که موجب انتخاب ویژگی هایی می‌شود که نماینده درستی از هدف مورد مطالعه نمی‌باشد[۳۳]	مشاوره با همکاران مسلط به موضوع مورد مطالعه، بعنوان مثال استفاده ازافراد خبره در حوزه پزشکی جهت انتخاب ویژگی های آموزشی نماینده در این حوزه [۱۲]
کیفیت پایین داده ها	وجود داده‌های پرت، عد م دقت در وارد کردن داده‌های بسیار بزرگ می‌تواند باعث ایجاد داده های پرت شود در اکثر تحقیقات مرتبط این نوع از داده‌ها وجود دارند[۱۳]	استفاده از ابزار هایی جهت پاکسازی داده مانند Clean Data [۱۵]
ویژگی های غیر مرتبط	گاهی مواقع با وجود تسلط محقق به موضوع ویژگی هایی هستند که به هدف ارتباط چندانی ندارند[۳۴]	استفاده از آزمون های آماری استفاده از الگوریتم های انتخاب ویژگی استفاده از الگوریتم های استخراج ویژگی[۳۵, ۳۶]

جدول ۲. چالش های ارزیابی مدل در حوزه یادگیری ماشین نظارت شده، علت ایجاد و راه حل هایی جهت رفع آن

چالش	علت ایجاد آن	راه حل
بیش‌برازش	پیچیدگی مدل: استفاده از مدل‌های بسیار پیچیده با تعداد زیادی پارامتر، احتمال بیش‌برازش را افزایش می‌دهد. کمبود داده‌ها: هنگامی که داده‌های آموزشی کم هستند، مدل تمایل دارد به‌طور زیادی به این داده‌ها متمایل شود و برازش دقیقی به آن‌ها داشته باشد.[۲۲]	تغییر مدل و استفاده از مدل های ساده تر افزایش داده و استفاده ار بانکِ داده‌ها در صورت کمبود وقت و هزینه[۲۴]
کمبود برازش	کمبود توانایی مدل: مدل مورد استفاده به طور کلی نتوانسته است الگوها و اطلاعات مرتبط با داده‌ها را یاد بگیرد انتخاب مدل نامناسب: ممکن است مدل انتخابی برای حل مسئله مورد نظر، مناسب نباشد.[۲۲]	افزایش داده‌ها: در صورت امکان، افزایش تعداد نمونه‌ها با استفاده از تکنیک‌های افزایش داده‌ها می‌توان به بهبود کیفیت مدل کمک کرد. انتخاب مدل مناسب: انتخاب مدل مناسب و متناسب با مسئله‌ی مورد نظر، از اهمیت ویژه‌ای برخوردار است. باید مدلی انتخاب شود که به خوبی به خصوصیات داده‌ها و نوع مسئله انطباق پیدا کند.[۲۴, ۲۵]
در دسترس نبودن داده ی کافی جهت اعتبار سنجی	در بعضی از مطالعات جمع آوری داده بسیار زمان بر پر هزینه است که می تواند باعث واریانس زیاد اعتبار سنجی مدل شود، [۲۸]	هنگامی که داده‌ها در حدود ۱۰۰۰ نمونه باشد برای اینکه از تمام داده جهت اعتبار سنجی استفاده شود، روش اعتبار سنجی K-fold بکار گرفته می‌شود [۲۸-۳۰]
عدم تطبیق داد ها	استفاده از داده های موجود در اینترنت باعث افزایش خطا اعتبار سنجی بوسیله داده‌های واقعی می‌شود[۳]	استفاده از داده‌های آموزش جهت اعتبار بخشی و پردازش داده جهت افزایش دقت اعتبار سنجی بوسیله داده‌های واقعی [۳۱]

بحث و نتیجه‌گیری

یادگیری ماشین یک فرآیند پیچیده‌ای است که نیاز به تعیین مدل‌ها و آموزش آن با استفاده از حجم زیادی از داده‌ها دارد. در گذشته، تمرکز اصلی بر روی بهبود ساختار مدل‌ها و الگوریتم‌ها بوده است. اما اخیراً، تمرکز بهتری به سمت کیفیت و کمیت داده‌ها انجام می‌شود. در حوزه تشخیص و درمان، استفاده از یادگیری ماشین نظارت شده بسیار گسترش یافته است. اما چالش اساسی در این حوزه، جمع‌آوری داده‌های صحیح و ارزیابی درست مدل‌ها می‌باشد.

در سال ۲۰۰۱ بانکو مطالعه‌ای منتشر کرد که در آن دقت الگوریتم‌های مختلف یادگیری ماشین، بر حسب تعداد داده‌های آموزشی با فرض ثابت نگه داشتن حجم داده‌های آزمون مورد ارزیابی قرار گرفت و نتیجه گرفته شد که دقت هر مدل یادگیری ماشین با افزایش داده‌های آموزشی بهبود می‌یابد همچنین دقت مدل‌های ساده و پیچیده با افزایش تعداد داده‌های آموزشی به یک‌دیگر نزدیک می‌شوند. نویسندگان این مطالعه با اشاره به تعادل بین دو جنبه مهم در فرآیند یادگیری ماشین (توسعه الگوریتم‌ها و توسعه مجموعه‌های داده) نشان دادند برای مواجهه با مسائل پیچیده اهمیت انتخاب مناسب مجموعه‌های داده بیش‌تر از توسعه الگوریتم‌ها می‌باشد. به عبارت دیگر، اگر مجموعه‌های داده کافی و با کیفیت موجود باشند، حتی الگوریتم‌های نسبتاً ساده هم می‌توانند عملکرد قابل قبولی در مسئله‌های پیچیده ارائه دهند [۸].

در سال ۲۰۲۱ آلتانیان و همکاران، تأثیر اندازه مجموعه داده‌ها را بر عملکرد مدل‌های طبقه‌بندی بررسی کردند. در این مطالعه شش مدل پرکاربرد را در حوزه‌ی پزشکی، روی مجموعه داده‌های کوچک ارزیابی و به بررسی کاهش حجم داده در داده‌های بزرگ‌تر پرداخته‌اند. یافته‌ها نشان می‌دهد عملکرد مدل‌ها بیش‌تر به نمایندگی داده‌ها از توزیع اصلی بستگی دارد تا صرفاً حجم آن‌ها [۳۲].

داده‌ها با کیفیت و حجم بیش‌تر، مدل‌های پیچیده‌تر و با توانایی‌های یادگیری بیش‌تر را ممکن می‌سازد. از این رو، به دست‌آوردن داده‌های کم باعث نویز می‌شود لذا برای حل این چالش استفاده از داده‌های بیش‌تر پیشنهاد می‌گردد اما به‌دلیل پرهزینه و زمان‌بر بودن جمع‌آوری داده استفاده از بانک داده جهت آموزش مدل پیشنهاد می‌شود. هم‌چنین جهت جلوگیری از انتخاب نمونه به گونه‌ای که داده‌ها متمرکز به یک شرایط خاص باشند، استفاده از الگوریتم‌های انتخاب تصادفی توصیه می‌گردد.

جین و همکاران در سال ۲۰۲۲ روشی را ارائه دادند که در آن میزان خطای تعمیم مدل با استفاده از داده‌های کوچک تقریباً ۸۰درصد کم‌تر از روش‌های مرسوم می‌باشد [۹].

کیم و همکاران در سال ۲۰۲۱ مطالعه‌ای انجام داده و در آن نشان دادند چگونه بایاس الگوریتمی منجر به تبعیض در پردازش داده‌ها می‌شود. در این مطالعه سه معیار ارزیابی برای سنجش بایاس ارائه شده است و روش‌هایی در سه دسته پیش‌پردازش، درون‌پردازش، و پس‌پردازش برای کاهش

بایاس توضیح داده شده است. مقاله یک رویکرد پیش‌پردازشی مبتنی بر نظریه اطلاعات معرفی می‌کند تا ویژگی‌های اصلی داده‌ها حفظ شوند و در عین حال از تبعیض جلوگیری گردد. آزمایش‌های انجام‌شده بر روی داده‌های استاندارد و واقعی، اعتبار این رویکرد را در کاهش بایاس‌ها تأیید می‌کند [۱۰]

اهمیت شناسایی داده‌های غیرنماینده در مرحله پیش‌آموزش مدل‌های یادگیری ماشین بسیار حیاتی است، زیرا این داده‌ها می‌توانند باعث ایجاد الگوهای گمراه‌کننده و نتایج ناقص شوند. به گونه‌ای که مدل نه تنها نمی‌تواند دقت لازم را در تشخیص الگوهای درست کسب کند، بلکه ممکن است به نتایجی برسد که ناخواسته باعث تقویت بایاس‌ها گردد.

کاوزگلو و همکاران در یک مطالعه اهمیت استفاده از مجموعه داده‌های آموزشی نماینده جهت طبقه‌بندی شبکه عصبی برای بهبود دقت طبقه‌بندی تصاویر را مورد بحث قرار داده است. در این مطالعه بر لزوم پالایش داده‌های آموزشی برای شناسایی دقیق خروجی‌های مورد مطالعه تأکید دارد. این مطالعه یک رویکرد دو مرحله‌ای برای افزایش بازنمایی داده‌های آموزشی و ارزیابی عملکرد آن در طبقه‌بندی ارائه می‌دهد. نتایج نشان می‌دهد که استفاده از داده‌های آموزشی نماینده منجر به نتایج طبقه‌بندی دقیق‌تر و قابل اطمینان‌تر و با بهبود قابل توجهی در دقت طبقه‌بندی می‌شود [۳۳].

هنگام جمع‌آوری داده گریزی از داده‌های پرت نمی‌باشد لذا اکثر مطالعات حاوی داده‌های پرت می‌باشند بنابراین جهت به حداقل رساند این نوع داده استفاده از ابزارهایی مانند Clean Data جهت بالا بردن کیفیت داده‌ها، لازم می‌گردد تا با داده‌های صحیح مدل عملکرد بهتری داشته باشد.

چائو و همکاران در سال ۲۰۱۶ مطالعه‌ای را بررسی کردند که در آن مطالعه بر چالش‌های تجزیه و تحلیل داده‌ها تمرکز دارند. نادیده گرفتن کیفیت داده‌ها باعث تحلیل‌های نادرست و تصمیمات غیر قابل اعتماد می‌شود. در این مطالعه بر تکنیک‌های اخیر که از محدودیت‌ها، قوانین یا الگوها برای تشخیص خطاها استفاده می‌کند، تأکید می‌شود. این نوع رویکرد را پاک‌سازی داده‌های کیفی می‌نامند. و تکنیک‌های پیشتاز در این زمینه و محدودیت‌های آن را با مثال‌های توضیحی بیان می‌کند [۱۵].

بعضی از ویژگی‌ها با وجود این‌که نماینده‌ی از موضوع می‌باشند اما نمی‌توانند با اهداف مطالعه مرتبط باشند، به‌طور مثال هنگام جمع‌آوری ویژگی‌های تصاویر سونوگرافی جهت تشخیص خوش‌خیمی از بدخیمی ۲۰۰ نوع ویژگی معرفی گشته است که همه‌ی آن‌ها به هدف مطالعه (تشخیص خوش‌خیمی از بدخیمی) مرتبط نمی‌باشد. برای شناسایی ویژگی‌های غیرمرتبط، انجام آزمون‌های آماری معرفی می‌شود. این آزمون‌ها بر روی داده‌ها اجرا می‌شوند تا ارتباط بین ویژگی‌ها و مقادیر خروجی مورد بررسی قرار گیرد. برخی از متدولوژی‌های آماری که می‌توان برای این منظور استفاده کرد عبارتند از:

- **ضریب هم‌بستگی (Correlation Coefficient)**: با استفاده از ضریب هم‌بستگی، می‌توان ارتباط خطی بین هر ویژگی و مقدار خروجی را محاسبه کرد. اگر ضریب هم‌بستگی کم باشد، این نشان‌دهنده ارتباط ضعیف ویژگی با مقدار خروجی است.
- **آزمون فرضیه (Hypothesis Testing)**: این آزمون‌ها برای تعیین این‌که آیا ویژگی و خروجی به طور تصادفی از یک توزیع یکسان می‌آیند یا خیر، استفاده می‌شوند. اگر مقدار P-Value آزمون کم‌تر از آستانه‌ای تعیین شده باشد، نشان‌دهنده وجود ارتباط غیرتصادفی بین ویژگی و خروجی می‌باشند.
- **آنالیز واریانس (ANOVA)**: این آزمون برای مقایسه میانگین گروه‌های مختلف با مقادیر خروجی استفاده می‌شود. اگر میانگین‌ها تفاوت معناداری با یک‌دیگر نداشته باشند، ویژگی‌ها می‌توانند به عنوان غیرمرتبط در نظر گرفته شوند.

استفاده از این آزمون‌ها این امکان را می‌دهد تا ویژگی‌های غیرمرتبط شناسایی شوند و در مدل‌سازی از آن‌ها صرف‌نظر شود تا دقت و کارآیی مدل افزایش یابد. راه حل دیگر برای کاهش تأثیر ویژگی‌های غیرمرتبط و بهبود عملکرد مدل، تکنیک‌های انتخاب و استخراج ویژگی می‌باشد.

- **انتخاب ویژگی**: در این روش، از بین تمام ویژگی‌های موجود تعدادی ویژگی با اهمیت‌تر و مرتبط‌تر انتخاب می‌شوند و بقیه ویژگی‌ها از مطالعه حذف می‌شوند. این انتخاب می‌تواند بر اساس روش‌های مختلفی مانند اطلاعات متقابل (Mutual Information) و یا وزن‌دهی (feature weighting) ماشین‌های یادگیری انجام شود. با حذف ویژگی‌های غیرمرتبط، تعداد زیادی از ویژگی‌ها کاسته و مدل‌ها بر روی ویژگی‌های اصلی و کارآمدتر آموزش داده می‌شوند [۳۴].
- **استخراج ویژگی**: در این روش، از طریق ایجاد روابط و مدل‌سازی میان ویژگی‌ها، ویژگی‌های جدیدی استخراج می‌شود که اطلاعات کلی و مرتبط‌تری از داده‌ها را به مدل ارائه می‌دهد. این ویژگی‌های جدید معمولاً با توجه به

اطلاعات اصلی ویژگی‌ها ایجاد می‌شوند و می‌توانند تعداد کم‌تری از آن‌ها به‌عنوان نماینده وارد مطالعه شوند تا مدل کم‌هزینه‌تر و کارآمدتر گردد [۳۵].

در سال ۲۰۰۶ ویراماچایی الگوریتم نمونه‌برداری فعال را معرفی کرد که می‌تواند تخمین‌های دقیقی از اهمیت ویژگی با هزینه‌های کم‌تر جمع‌آوری داده در مقایسه با نمونه‌برداری تصادفی و الگوریتم‌های نمونه‌برداری پیشنهادی پیشین ارائه دهد [۳۶].

جبار و همکاران در سال ۲۰۱۵ دو چالش مهم تعیین مدل در یادگری ماشین نظارت شده را مورد بررسی قرار دادند. آن‌ها اظهار کردند مدل بیش برازش زمانی رخ می‌دهد که مدل به طور غیرمنطقی و شدید به داده‌های آموزشی تطبیق یافته و دقت مدل در داده‌های آموزشی بسیار بالا است اما دقت مدل بر روی داده‌های آزمون کم می‌باشد و برای جلوگیری از بیش‌برازش از روش‌هایی مانند استفاده از مدل‌های ساده‌تر، اعتبارسنجی مناسب، استفاده از روش‌های کاهش ابعاد و هم‌چنین افزایش تنوع داده‌ها پیشنهاد شده است.

اگر مدل به صورت بسیار ضعیف عمل کند (به این معنی که با دقت خوبی به داده‌های آموزشی منطبق نباشد) و هم‌چنین روی داده‌های آزمون دقت خوبی در پیش‌بینی داده‌ها نداشته باشد. ممکن است با مشکلی مواجه باشیم با عنوان کمبود برازش که به عبارتی به عدم تعمیم‌پذیری مدل اشاره می‌کند. این مشکلات ممکن است به دلیل، کمبود توانایی مدل و انتخاب مدل نامناسب باشد. برای مقابله با این چالش، افزایش داده‌ها و تغییر مدل پیشنهاد می‌گردد [۲۴].

به‌طور مثال در شکل ۳ داده‌هایی نمایش داده شده است که هر کدام معرف دو ویژگی از مجموعه‌ی داده‌ها می‌باشند. در صورتی که مدلی با صد در صد انطباق با داده‌های آموزشی انتخاب شود (منحنی مشکی)، این مدل دچار بیش برازش شده است و در صورتی که داده‌های آزمون در محدوده‌ی دایره خاکستری قرار گیرند مدل آموزش‌دیده آن را در محدوده‌ی دایره تشخیص می‌دهد (با وجود این‌که به احتمال زیاد داده‌های مثلث در این محدوده بیش‌تر خواهند بود).

یکی از چالش‌های تعیین مدل در یادگیری ماشین نظارت شده عدم تطبیق داده‌ها می‌باشد، به این معنی که بین داده‌های آموزش و آزمون تطابقی وجود ندارد، این مشکل معمولاً هنگامی پیش می‌آید که محققین از داده‌های موجود در بانک‌های داده موجود در اینترنت جهت آموزش استفاده می‌کنند، استفاده کردن از این داده‌ها به دلیل حجم بالا برای آموزش مدل‌ها با ارزش می‌باشد اما برای اعتبارسنجی مدل با داده‌های واقعی انجام می‌گردد. گرون و همکاران در کتاب خود که در سال ۲۰۲۲ منتشر کردند اظهار داشتند در صورت عدم تطابق و دقت پایین اعتبارسنجی مدل با داده‌های آزمون، مدل باید توسط قسمتی از داده‌های آموزشی اعتبارسنجی شود تا از عدم بیش برازش مطمئن شد، در صورتی که مدل توسط داده‌های آموزشی (داده‌هایی که از اینترنت جمع‌آوری شده) دارای دقت خوبی بود، مدل با قسمتی از داده‌های واقعی اعتبارسنجی شود، در صورتی که مدل با داده‌های واقعی دقت پایینی در اعتبارسنجی داشت اولین راه حل پردازش داده‌ها می‌باشد، سپس در صورت پایداری دقت پایین باید مدل تغییر یابد [۳]. (شکل ۷)

در بسیاری از مطالعات حجم داده‌ها کم می‌باشد (مثلاً ۱۰۰۰ داده) لذا در صورت ارزیابی چندین مدل برای بررسی اعتبارسنجی مدل‌ها نیاز است به روش اعتبارسنجی K-fold عمل گردد تا بهترین مدلی که با عملکر دقیق‌تری دارد انتخاب شود. دقت اعتبارسنجی مدل در هر بار آموزش با داده‌های تصادفی دارای واریانس بالایی می‌باشد که نشان می‌دهد داده‌ها برای اعتبارسنجی مدل کافی نمی‌باشند لذا یکی از راه حل‌های غلبه بر این مسئله افزایش تعداد داده است، که در بیش‌تر مواقع به دلیل صرف زمان و هزینه‌ی زیاد امکان افزایش داده مقدور نیست برای حل این مشکل از اعتباربخشی k-fold (K-fold cross validation) استفاده می‌شود [۲۹].

تشکر و قدردانی

برای انجام این تحقیق از معاونت پژوهشی و کمیته تحقیقات دانشگاه علوم پزشکی سمنان تشکر و قدردانی می‌گردد.

مشارکت و نقش نویسندگان

تمام نویسندگان در آماده‌سازی این مقاله مشارکت داشته‌اند.

منابع

[1] Flasiński M, Flasiński M. Symbolic artificial intelligence. Introduc Art Intell 2016; 15-22. https://doi.org/10.1007/978-3-319-40022-8_2

[2] Badillo S, Banfai B, Birzele F, Davydov II, Hutchinson L, Kam-Thong T, et al. An introduction to machine learning. Clin Pharmacol Ther 2020; 107: 871-885. https://doi.org/10.1002/cpt.1796 PMid:32128792 PMCid:PMC7189875

[3] Géron A. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: O'Reilly Media, Inc 2022.

[4] Jain A, Patel H, Nagalapatti L, Gupta N, Mehta S, Guttula S, et al, editors. Overview and importance of data quality for machine learning tasks. Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining; 2020. https://doi.org/10.1145/3394486.3406477

[5] Budach L, Feuerpfeil M, Ihde N, Nathansen A, Noack N, Patzlaff H, et al. The effects of data quality on machine learning performance. ArXiv (preprint) 2022.
[6] Kariluoto A, Kultanen J, Soininen J, Pärnänen A, Abrahamsson P, editors. Quality of data in machine learning. 2021 IEEE 21st international conference on software quality, reliability and security companion (QRS-C); 2021: IEEE. https://doi.org/10.1109/QRS-C55045.2021.00040
[7] Sarker IH. Machine learning: Algorithms, real-world applications and research directions. SN Comput Sci 2021; 2: 160. https://doi.org/10.1007/s42979-021-00592-x PMid:33778771 PMCid:PMC7983091
[8] Banko M, Brill E, editors. Mitigating the paucity-of-data problem: Exploring the effect of training corpus size on classifier performance for natural language processing. Proceedings of the first international conference on Human language technology research; 2001. https://doi.org/10.3115/1072133.1072204 PMid:22478353 PMCid:PMC2731358
[9] Jin J, Yin F, Xu Y, Zhang J, editors. Learning a model with the most generality for small-sample problems. proceedings of the 2022 5th international conference on algorithms, computing and artificial intelligence; 2022. https://doi.org/10.1145/3579731.3579814
[10] Kim JY, Cho SB. An information theoretic approach to reducing algorithmic bias for machine learning. Neurocomputing 2022; 500: 26-38. https://doi.org/10.1016/j.neucom.2021.09.081
[11] Chen M, Cheng H, Du Y, Xu M, Jiang W, Wang C, editors. Two wrongs don't make a right: Combating confirmation bias in learning with label noise. Proceedings of the AAAI Conference on Artificial Intelligence; 2023. https://doi.org/10.1609/aaai.v37i12.26725
[12] Akhatov A, Ulugmurodov SA. Training data selection and labeling for machine learning braille recognition models. Int J Contem Sci Tech Res 2023; 15-21.
[13] Whang SE, Roh Y, Song H, Lee JG. Data collection and quality challenges in deep learning: A data-centric ai perspective. VLDB J 2023; 32: 791-813. https://doi.org/10.1007/s00778-022-00775-9
[14] Angloher G, Banik S, Bartolot D, Benato G, Bento A, Bertolini A, Breier R, Bucci C, Burkhart J, Canonica L. Towards an automated data cleaning with deep learning in CRESST. Eur Phys J Plus 2023; 138: 1-11. https://doi.org/10.1140/epjp/s13360-023-03674-2 PMid:36741916 PMCid:PMC9886615
[15] Chu X, Ilyas IF, Krishnan S, Wang J, editors. Data cleaning: Overview and emerging challenges. Proceedings of the 2016 international conference on management of data; 2016. https://doi.org/10.1145/2882903.2912574
[16] Singh A, Thakur N, Sharma A, editors. A review of supervised machine learning algorithms. 2016 3rd international conference on computing for sustainable global development (INDIACom); 2016: IEEE.
[17] Eminaga O, Abbas M, Shen J, Laurie M, Brooks JD, Liao JC, Rubin DL. PlexusNet: A neural network architectural concept for medical image classification. Comput Biol Med 2023; 154: 106594. https://doi.org/10.1016/j.compbiomed.2023.106594 PMid:36753979
[18] Gupta A, Chaithra N, Jha J, Sayal A, Gupta V, Memoria M, editors. Machine learning algorithms for disease diagnosis using medical records: a comparative analysis. 2023 4th International Conference on Intelligent Engineering and Management (ICIEM); 2023: IEEE. https://doi.org/10.1109/ICIEM59379.2023.10165850
[19] Kaur P, Singh RK. A review on optimization techniques for medical image analysis. Concur Comput Pract Exp 2023; 35: e7443. https://doi.org/10.1002/cpe.7443
[20] Shanbehzadeh M, Valinejadi A, Afrah R, Kazemi AH, Orooji A, Kaffashian MR. Comparison of machine-learning algorithms efficiency to build a predictive model for mortality risk in COVID-19 hospitalized patients. Koomesh 2021; 24: 128-138. (Persian)
[21] Tanhapour M KL, Maghooli K, Rostam Niakan Kalhori S. Determining the progression stages of liver fibrosis in patients with chronic hepatitis B. Koomesh 2022; 24: 639-647. (Persian)
[22] Ying X, editor. An overview of overfitting and its solutions. Journal of physics: Conference series; 2019: IOP Publishing. https://doi.org/10.1088/1742-6596/1168/2/022022
[23] Nazha A, Elemento O, McWeeney SK, Miles M, Haferlach T. How I read an article that uses machine learning methods. Blood Adv 2023; 2023010140. https://doi.org/10.1182/bloodadvances.2023010140 PMid:37276509 PMCid:PMC10425665
[24] Jabbar H, Khan RZ. Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). Comput Sci Commun Instru Devic 2015; 70: 978-981. https://doi.org/10.3850/978-981-09-5247-1_017
[25]Swathi P. Analysis on solutions for over-fitting and under-fitting in machine learning algorithms. Int J Innov Res Sci Eng Technol 2018; 7: 10.15680.
[26] Uçar MK, Nour M, Sindi H, Polat K. The effect of training and testing process on machine learning in biomedical datasets. Mathem Prob Engin 2020; 2020. https://doi.org/10.1155/2020/2836236
[27] Avuçlu E, Elen A. Evaluation of train and test performance of machine learning algorithms and Parkinson diagnosis with statistical measurements. Med Biol Eng Comput 2020; 58: 2775-2788. https://doi.org/10.1007/s11517-020-02260-3 PMid:32920727
[28] Anguita D, Ghelardoni L, Ghio A, Oneto L, Ridella S, editors. The 'K' in K-fold Cross Validation. ESANN 2012.
[29] Wong TT, Yeh PY. Reliable accuracy estimates from k-fold cross validation. IEEE Trans Knowledge Data Eng 2019; 32: 1586-1594. https://doi.org/10.1109/TKDE.2019.2912815
[30] Fushiki T. Estimation of prediction error by using K-fold cross-validation. Stat Comput 2011; 21: 137-146. https://doi.org/10.1007/s11222-009-9153-8
[31] Lewis GA, Bellomo S, Ozkaya I, editors. Characterizing and detecting mismatch in machine-learning-enabled systems. 2021 IEEE/ACM 1st Workshop on AI Engineering-Software Engineering for AI (WAIN); 2021: IEEE. https://doi.org/10.1109/WAIN52551.2021.00028
[32] Althnian A, AlSaeed D, Al-Baity H, Samha A, Dris AB, Alzakari N, et al. Impact of dataset size on classification performance: an empirical evaluation in the medical domain. Appl Sci 2021; 11: 796. https://doi.org/10.3390/app11020796
[33] Kavzoglu T. Increasing the accuracy of neural network classification using refined training data. Environ Model Software 2009; 24: 850-858. https://doi.org/10.1016/j.envsoft.2008.11.012
[34] Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. Neurocomputing 2018; 300: 70-79. https://doi.org/10.1016/j.neucom.2017.11.077
[35] Guyon I, Elisseeff A. An introduction to feature extraction. Feature extraction: foundations and applications: Springer; 2006. p. 1-25. https://doi.org/10.1007/978-3-540-35488-8_1
[36] Veeramachaneni S, Olivetti E, Avesani P, editors. Active sampling for detecting irrelevant features. Proceedings of the 23rd international conference on machine learning; 2006. https://doi.org/10.1145/1143844.1143965

Koomesh -Volume 25, Issue 6 November and December 2023, 551-561

Challenges and solutions in data collection and model evaluation in supervised machine learning: a review article

Saeedeh Aliakbari (Ph.D) [*1], Peyman Hejazi (Ph.D) [2], Zeinab Hormozi-Moghaddam (Ph.D)[3,4]
1- Dept. of Radiology, Allied Health Sciences Faculty, Semnan University of Medical Sciences, Semnan, Iran
2 – Dept. of Medical Physics, Semnan University of Medical Sciences, Semnan, Iran
3- Radiation Biology Research Center, Iran University of Medical Sciences (IUMS), Tehran, Iran
4- Dept. of Radiation Sciences, Allied Medicine Faculty, Iran University of Medical Sciences (IUMS), Tehran, Iran

* Corresponding author. +098 9354470112 s.aliakbari@semums.ac.ir
Received: 2023 Jul 23; Accepted: 2024 Feb 26

Introduction: The main purpose of machine learning is a complex process that is carried out by determining the model and training it using a large volume of data. In the past, the main focus in this field was more on improving the structures of models and algorithms, but recently more emphasis has been placed on the quality and quantity of data. This article aims to provide an overview of the problems in data collection and offer a solution for them.

Materials and Methods: In this study, the challenges faced by researchers in collecting data and evaluating supervised machine-learning models were examined through a review method. Documentation from PubMed, Scopus, Science Direct databases, and Google Scholar search engine from 2001 to 2023 was retrieved. After screening, a total of 17 full articles were reviewed and included in the study.

Results: The findings indicate that researchers in supervised machine learning studies face four challenges in data collection, which are: insufficient number of samples, unrepresentative training data, poor data quality, and irrelevant features, and in model evaluation, they face four challenges: overfitting, lack of generalizability, lack of sufficient data for validation, and mismatched data.

Conclusion: Increasing the sample size, utilizing a random selection algorithm, data cleansing, using the correct statistical test, feature selection, feature extraction, using a simpler model, the K-fold technique, and data processing are among the factors that contribute to achieving a model with better performance.

Keywords: Supervised Machine Learning, Data Collection, Model Evaluation