

عنوان مقاله :

بررسی روش های انتخاب ویژگی بر اساس الگوریتم های تکاملی (ژنتیک، کلونی مورچه ها، PSO و غیره)

حسن نصرتی ناهوک

مریی دانشکده فنی و مهندسی، دانشگاه سراوان، سراوان، ایران

نرگس خسروی منفرد

دانشجو کارشناسی مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه سراوان، سراوان، ایران

چکیده

انتخاب ویژگی به عنوان یک گام کلیدی در فرآیند یادگیری ماشین، نقش مهمی در کاهش پیچیدگی محاسباتی و بهبود کارایی مدل های یادگیری ایفا می کند. این مقاله به بررسی روش های مختلف انتخاب ویژگی مبتنی بر الگوریتم های تکاملی، از جمله الگوریتم ژنتیک، کلونی مورچگان، و بهینه سازی ازدحام ذرات می پردازد. هدف از این مطالعه، ارائه یک مرور جامع از این الگوریتم ها، ارزیابی عملکرد آن ها در مقایسه با روش های سنتی، و تحلیل مزایا و معایب هر یک در انتخاب ویژگی ها برای مجموعه داده های بزرگ و پیچیده است. یافته های این مقاله نشان می دهد که الگوریتم های تکاملی به دلیل قابلیت های خود در جستجوی فضای ویژگی های بزرگ، ابزارهای قدرتمندی برای انتخاب ویژگی ها به شمار می آیند. در نهایت، این مقاله به بررسی نتایج و پیشنهادات برای تحقیقات آینده در این زمینه می پردازد.

کلید واژه ها

انتخاب ویژگی، الگوریتم های تکاملی، الگوریتم های ژنتیک^۱، بهینه سازی کلونی مورچه ها^۲، بهینه سازی ازدحام ذرات^۳

¹ - Genetic Algorithm - GA

² - Ant Colony Optimization - ACO

³ - Particle Swarm Optimization - PSO

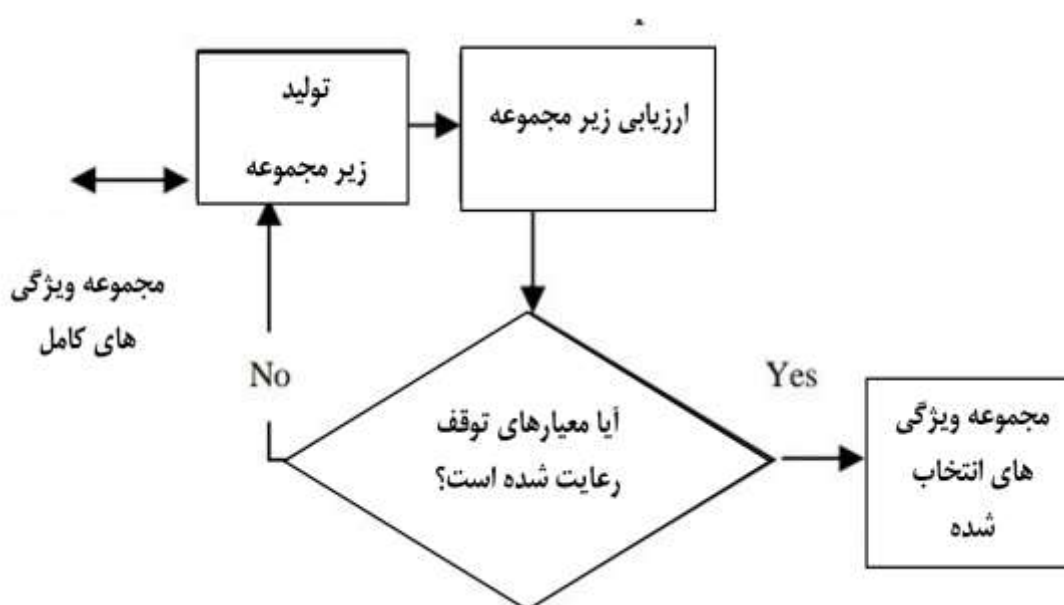
مقدمه

انتخاب ویژگی، فرآیندی است که طی آن مجموعه‌ای از ویژگی‌های مهم و مرتبط از میان ویژگی‌های اولیه یک مجموعه داده انتخاب می‌شود. این فرآیند به طور مستقیم بر دقت و کارایی مدل‌های یادگیری ماشین تأثیر می‌گذارد. در دهه‌های اخیر، با رشد نمایی حجم داده‌ها و افزایش پیچیدگی مدل‌های یادگیری، نیاز به روش‌های مؤثرتری برای انتخاب ویژگی احساس شده است. روش‌های سنتی انتخاب ویژگی، مانند روش‌های مبتنی بر فیلتر و فراابتکار، گرچه در مواردی عملکرد خوبی نشان داده‌اند، اما در مواجهه با داده‌های پیچیده و دارای ابعاد بالا با چالش‌هایی روبرو هستند. این چالش‌ها باعث شده تا محققان به سمت استفاده از الگوریتم‌های تکاملی روی آورند. الگوریتم‌های تکاملی، با استفاده از اصول الهام‌گرفته از طبیعت مانند انتخاب طبیعی، رفتار جمعی، و تکامل، توانسته‌اند بهبودهای قابل توجهی در فرآیند انتخاب ویژگی ایجاد کنند. در این مقاله، ابتدا به بررسی روش‌های سنتی انتخاب ویژگی پرداخته و سپس به معرفی و تحلیل الگوریتم‌های تکاملی خواهیم پرداخت. در نهایت، به مقایسه این روش‌ها و ارائه نتایج پژوهش‌های اخیر در این زمینه خواهیم پرداخت.

با پیشرفت‌های اخیر در زمینه الگوریتم‌های تکاملی، روش‌های جدیدی برای انتخاب ویژگی معرفی شده‌اند که قادر به جستجوی بهینه در فضای بزرگ ویژگی‌ها هستند. این مقاله به بررسی این روش‌های جدید و مقایسه آن‌ها با روش‌های سنتی می‌پردازد. در ادامه، کاربردهای این روش‌ها در حوزه‌های مختلف، از جمله بیوانفورماتیک، تشخیص پزشکی، و تحلیل داده‌های پیچیده مورد بحث قرار می‌گیرد. هدف از این مقاله ارائه یک مرور جامع بر روش‌های انتخاب ویژگی مبتنی بر الگوریتم‌های تکاملی و بررسی تحقیقات و پیشرفت‌های اخیر در این حوزه است.

انتخاب ویژگی

انتخاب ویژگی برای کاهش تأثیر ابعاد بر روی مجموعه داده‌ها با یافتن زیرمجموعه‌ای از ویژگی‌ها که به طور مؤثر داده‌ها را تعریف می‌کند، استفاده می‌شود. این روش ویژگی‌های مهم و مرتبط با وظیفه معدن کاوی را از داده‌های ورودی انتخاب کرده و ویژگی‌های اضافی و نامربوط را حذف می‌کند. این کار برای شناسایی زیرمجموعه‌ای از ویژگی‌ها که برای مشکل خاص مناسب است، مفید است [۱]. هدف اصلی انتخاب ویژگی، ساختن زیرمجموعه‌ای از ویژگی‌ها به کوچک‌ترین اندازه ممکن است که ویژگی‌های مهم کل داده‌های ورودی را نمایان کند، انتخاب ویژگی مزایای زیادی دارد: کاهش اندازه داده‌ها، کاهش نیاز به ذخیره‌سازی، بهبود دقت پیش‌بینی، جلوگیری از بیش‌برازش، و کاهش زمان اجرای و آموزش از طریق متغیرهای قابل درک آسان. مرحله الگوریتم انتخاب ویژگی به دو مرحله تقسیم می‌شود ۱- تولید زیرمجموعه ۲- ارزیابی زیرمجموعه: در مرحله تولید زیرمجموعه، نیاز است که زیرمجموعه‌ای از مجموعه داده‌های ورودی تولید کنیم و برای استفاده از ارزیابی زیرمجموعه، باید بررسی کنیم که آیا زیرمجموعه تولید شده بهینه است یا خیر؟



شکل ۱. فرآیند انتخاب ویژگی

کاربردهای انتخاب ویژگی

روش‌های انتخاب ویژگی می‌توانند در پیش‌پردازش داده‌ها برای دستیابی به کاهش مؤثر داده‌ها استفاده شوند. این امر به یافتن مدل‌های دقیق داده کمک می‌کند. زیرا در اکثر موارد، جستجوی جامع برای یک زیرمجموعه ویژگی ایده‌آل دشوار است، چندین استراتژی

جستجو پیشنهاد شده است. کاربردهای معمول انتخاب ویژگی شامل خوشه‌بندی، رگرسیون، طبقه‌بندی و کاهش ابعاد است. این کاربردها می‌توانند به مسائل واقعی، مانند بینایی کامپیوتری، پردازش تصویر، زیست‌اطلاعات، داده‌کاوی متنی و کاربردهای صنعتی، اعمال شوند.

مشکلات انتخاب ویژگی

مسائل مختلفی می‌توانند از کاربرد تکنیک‌های انتخاب ویژگی بهره‌مند شوند. داده‌های با ابعاد بالا و اندازه نمونه‌های کوچک در حوزه‌های مختلف به طور فزاینده‌ای محبوب می‌شوند. بسیاری از ویژگی‌های این مشکلات به طور کافی طبقه‌بندی را تسهیل نمی‌کنند. علاوه بر این، مشکل عدم تعادل زمانی رخ می‌دهد که یکی از دو کلاس نمونه‌های بیشتری نسبت به سایر کلاس‌ها داشته باشد. بسیاری از الگوریتم‌ها هنگام تمرکز بر طبقه‌بندی نمونه‌های اصلی، نمونه‌های اقلیت را نادیده می‌گیرند. با این حال، نمونه‌های اقلیت مهم هستند اما به ندرت اتفاق می‌افتند. علاوه بر این، در یادگیری ماشین، تغییر داده‌ها یکی از مشکلات رایج است که زمانی رخ می‌دهد که توزیع مشترک ورودی‌ها و خروجی‌ها در مراحل آموزشی و تست متفاوت است. مورد خاصی از تغییر داده‌ها، که زمانی رخ می‌دهد که تنها توزیع ورودی تغییر می‌کند، به آن "تغییر همبستگی" می‌گویند. علاوه بر این، کاهش ابعاد و به دنبال آن انتخاب ویژگی یکی از متداول‌ترین تکنیک‌های حذف داده‌های نویزی است. در نهایت، هزینه‌های اشتباه طبقه‌بندی و هزینه‌های آزمون دو نوع هزینه مهم در یادگیری حساس به هزینه هستند.

روش‌های پیشین انتخاب ویژگی

۱- روش‌های فیلتر

روش‌های فیلتر یکی از ساده‌ترین و رایج‌ترین روش‌های سنتی انتخاب ویژگی هستند که به دلیل سرعت بالا و پیچیدگی کم، معمولاً در مراحل اولیه انتخاب ویژگی استفاده می‌شوند. در این روش‌ها، ویژگی‌ها به صورت مستقل از مدل یادگیری ارزیابی شده و از معیارهای آماری برای رتبه‌بندی و انتخاب آن‌ها استفاده می‌شود. برخی از معیارهای رایج در این روش‌ها شامل آزمون تست کای دو^۴، اطلاعات متقابل^۵، و ضریب همبستگی پیرسون^۶ هستند. آزمون کای دو برای بررسی همبستگی بین ویژگی‌ها و متغیر هدف به کار می‌رود و ویژگی‌هایی که بالاترین مقدار کای دو را دارند به عنوان مهم‌ترین ویژگی‌ها انتخاب می‌شوند. معیار اطلاعات متقابل به ارزیابی میزان اطلاعاتی که یک ویژگی درباره متغیر هدف ارائه می‌دهد، می‌پردازد و هر چه این مقدار بیشتر باشد، ویژگی مفیدتر تلقی می‌شود. همچنین، ضریب همبستگی پیرسون میزان همبستگی خطی بین یک ویژگی و متغیر هدف را اندازه‌گیری می‌کند، به طوری که ویژگی‌هایی با همبستگی بالاتر احتمالاً تأثیر بیشتری بر خروجی دارند. مزیت اصلی روش‌های فیلتر این است که سریع هستند و نیازی به اجرای مدل‌های پیچیده یادگیری ماشین ندارند، اما معمولاً روابط غیرخطی بین ویژگی‌ها و متغیر هدف را نادیده می‌گیرند [2].

۲- روش‌های پوششی^۷

⁴ - Chi-Square Test

⁵ - Mutual Information

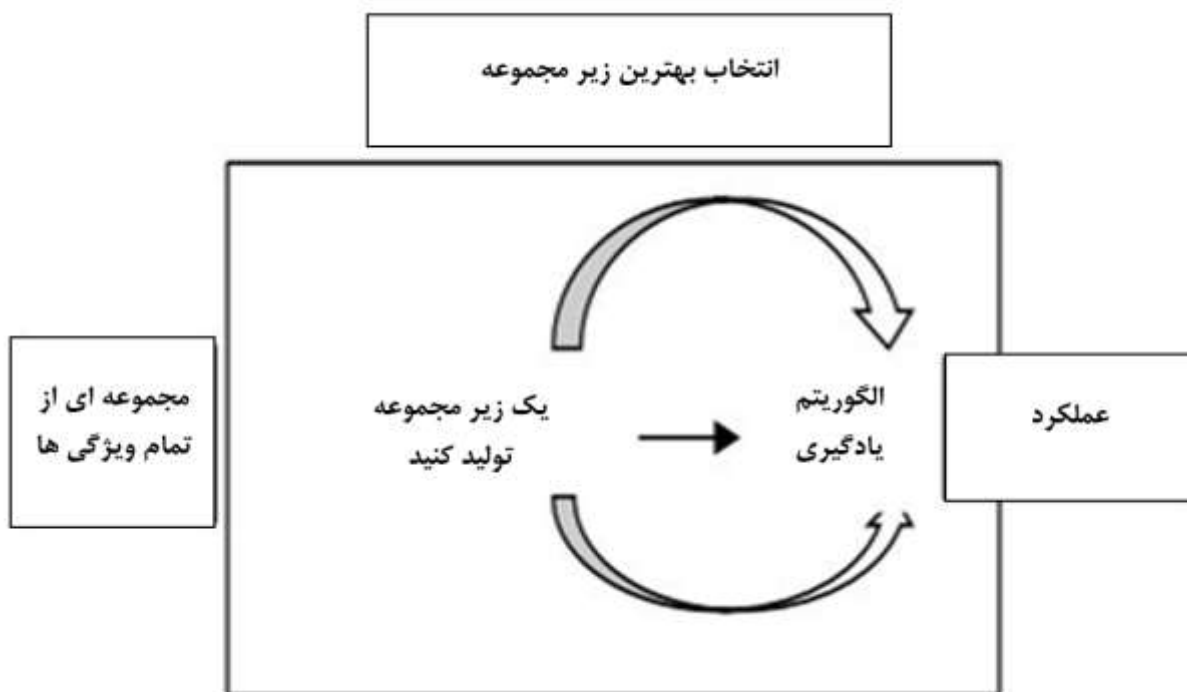
⁶ - Pearson Correlation

⁷ - Wrapper

روشهای پوششی از مدل های یادگیری برای ارزیابی ویژگی ها استفاده می کنند. در این روش ها، ویژگی های مختلف به صورت ترکیبی با هم آزمایش شده و بهترین مجموعه ویژگی ها بر اساس عملکرد مدل انتخاب می شوند. این روش ها به دلیل استفاده از مدل های یادگیری برای ارزیابی، معمولاً دقیق تر از روش های فیلتر هستند، اما به مراتب کندتر و محاسباتی پرهزینه تر هستند.

یکی از معروف ترین روش های پوششی، روش حذف به عقب^{۹۸} است. در این روش، مدل با استفاده از تمامی ویژگی ها آموزش داده می شود و سپس به صورت تدریجی ویژگی هایی که کمترین تأثیر را بر عملکرد مدل دارند حذف می شوند تا به مجموعه ای از ویژگی های بهینه برسیم.

روش های پوششی به دلیل وابستگی به مدل های یادگیری، دقت بالاتری دارند و می توانند بهینه سازی مناسبی برای ویژگی ها انجام دهند. با این حال، این روش ها به دلیل تعداد بالای آزمون های مورد نیاز، محاسباتی پرهزینه هستند و برای داده های بزرگ ممکن است بسیار زمان بر باشند [3].



شکل ۲. روش پوششی برای انتخاب ویژگی

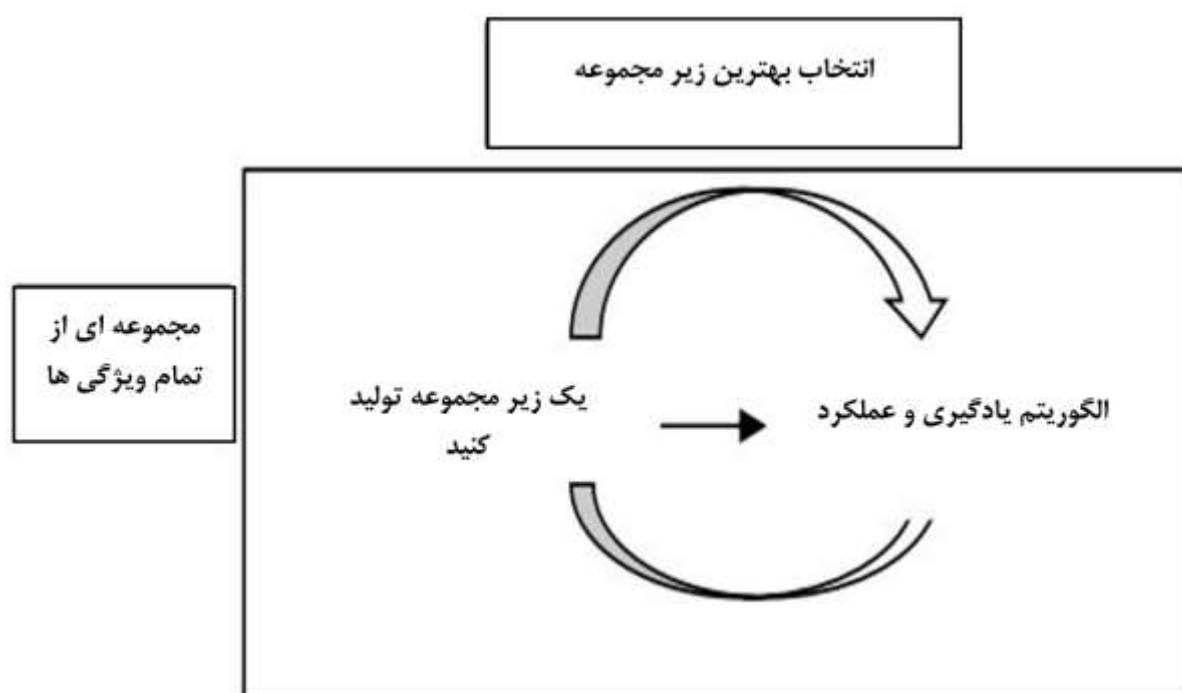
۳- روش های جاسازی شده^{۱۰}

روش های جاسازی شده رویکردهایی در انتخاب ویژگی هستند که به طور هم زمان با فرآیند آموزش مدل یادگیری ماشین انجام می شوند. برخلاف روش های فیلتر و پوششی که قبل یا بعد از آموزش مدل صورت می گیرند، روش های جاسازی شده در حین ساخت

⁸ - Backward Elimination

¹⁰ - Embedded Methods

مدل، ویژگی‌های مهم را شناسایی و انتخاب می‌کند. این رویکرد معمولاً در مدل‌هایی مانند درخت‌های تصمیم‌گیری و رگرسیون‌های خطی با نظم‌دهی مانند لاسو^{۱۱} و رگرسیون ریج^{۱۲} استفاده می‌شود. در درخت‌های تصمیم‌گیری، ویژگی‌هایی که در گره‌های بالاتر قرار دارند و بیشترین اطلاعات را فراهم می‌کنند به عنوان ویژگی‌های مهم‌تر انتخاب می‌شوند. در مقابل، در رگرسیون‌های نظم‌دهی شده، وزن‌هایی که به ویژگی‌های کم‌اهمیت تخصیص داده می‌شوند، به صفر نزدیک شده و در نتیجه این ویژگی‌ها حذف می‌شوند. این روش‌ها به دلیل ادغام انتخاب ویژگی با فرآیند یادگیری، می‌توانند منجر به بهبود عملکرد مدل‌های یادگیری ماشین شوند و هم‌زمان با کاهش پیچیدگی مدل، دقت آن را افزایش دهند. این رویکرد به ویژه در مسائل با تعداد زیادی ویژگی کاربرد دارد، چرا که باعث کاهش بعد و افزایش کارایی مدل‌ها می‌شود. با این حال، روش‌های جاسازی شده ممکن است در مسائل با تعاملات پیچیده بین ویژگی‌ها بهینه‌ترین عملکرد را نداشته باشند و نیاز به تنظیم دقیق پارامترهای مدل دارند[۴].



شکل ۳. روش جدا سازی برای انتخاب ویژگی

۴- همبستگی

انتخاب ویژگی بر اساس همبستگی^{۱۳} یکی از روش‌های سنتی و پرکاربرد در انتخاب ویژگی‌ها است که بر اساس معیارهای آماری و ریاضیاتی برای تعیین ارتباط بین ویژگی‌ها و برچسب هدف عمل می‌کند. این روش به طور خاص به دنبال ویژگی‌هایی است که دارای بیشترین همبستگی با برچسب هدف (متغیر وابسته) باشند و در عین حال کمترین همبستگی را با یکدیگر داشته باشند. اساس کار

¹¹ - LASSO

¹² - Ridge Regression

¹³ - Corelation – based Feature Selection – CFS

انتخاب ویژگی بر اساس همبستگی بر این ایده استوار است که ویژگی‌های مفید آن‌هایی هستند که با برچسب هدف قوی‌ترین رابطه را دارند اما با سایر ویژگی‌ها کمتر مرتبط هستند. این کار به منظور کاهش افزونگی اطلاعات و جلوگیری از هم‌خطی بین ویژگی‌ها انجام می‌شود. این امر می‌تواند به بهبود کارایی مدل و کاهش پیچیدگی محاسباتی کمک کند، زیرا ویژگی‌های انتخاب‌شده اطلاعات غیرضروری یا تکراری کمتری را شامل می‌شوند. برای انجام این کار، انتخاب ویژگی بر اساس همبستگی از معیارهای آماری مانند ضریب همبستگی پیرسون استفاده می‌کند تا میزان همبستگی بین هر ویژگی و برچسب هدف را اندازه‌گیری کند. سپس، این روش ویژگی‌هایی را که بالاترین ضریب همبستگی با برچسب هدف و پایین‌ترین ضریب همبستگی با سایر ویژگی‌ها دارند، انتخاب می‌کند. یکی از مزایای اصلی انتخاب ویژگی بر اساس همبستگی سادگی و سرعت آن است، زیرا نیاز به تنظیم پارامترهای پیچیده ندارد و به راحتی قابل پیاده‌سازی است. همچنین، این روش به دلیل ماهیت مستقل از مدل بودن، می‌تواند به عنوان یک پیش‌پردازش اولیه قبل از استفاده از مدل‌های پیچیده‌تر استفاده شود. با این حال، انتخاب ویژگی بر اساس همبستگی نیز دارای محدودیت‌هایی است. یکی از این محدودیت‌ها این است که تنها همبستگی خطی بین ویژگی‌ها و برچسب هدف را در نظر می‌گیرد و ممکن است تعاملات پیچیده و غیرخطی بین ویژگی‌ها را نادیده بگیرد. همچنین، در مواجهه با داده‌های بسیار پیچیده و با تعداد زیادی ویژگی، انتخاب ویژگی بر اساس همبستگی ممکن است کارایی لازم را نداشته باشد، زیرا محاسبه همبستگی بین همه جفت‌های ویژگی‌ها می‌تواند زمان‌بر باشد [5].

۵- حذف بازگشتی

روش‌های انتخاب ویژگی بر اساس حذف بازگشتی^{۱۴} یکی از روش‌های سنتی و محبوب در انتخاب ویژگی‌ها در یادگیری ماشین است. این روش به صورت بازگشتی و گام به گام، ویژگی‌هایی که تأثیر کمتری بر عملکرد مدل دارند را شناسایی و حذف می‌کند تا در نهایت مجموعه‌ای از ویژگی‌های بهینه باقی بمانند. فرآیند حذف بازگشتی با آموزش یک مدل اولیه بر روی تمام ویژگی‌ها آغاز می‌شود. سپس اهمیت هر ویژگی بر اساس معیارهای مدل ارزیابی می‌شود، و کم‌اهمیت‌ترین ویژگی یا ویژگی‌ها حذف می‌شوند. پس از هر مرحله حذف، مدل دوباره آموزش داده می‌شود و این فرآیند تکرار می‌شود تا تنها ویژگی‌های مهم باقی بمانند. هدف اصلی این روش، بهبود دقت مدل و کاهش پیچیدگی آن از طریق حذف ویژگی‌های غیرضروری است. با اینکه حذف بازگشتی در بسیاری از موارد به خوبی عمل می‌کند، اما به دلیل نیاز به بازآموزی مکرر مدل، می‌تواند زمان‌بر باشد و در مسائل با تعداد زیادی ویژگی یا داده‌های بزرگ، محاسبات سنگینی را به همراه داشته باشد. حذف بازگشتی به ویژه در مسائل پردازش داده‌های با ابعاد بالا مانند بیوانفورماتیک و تشخیص الگو مورد استفاده قرار می‌گیرد [6].

۶- آزمون تحلیل واریانس^{۱۵}

آزمون تحلیل واریانس در انتخاب ویژگی به ارزیابی تأثیر ویژگی‌ها بر متغیر هدف کمک می‌کند. این آزمون با تقسیم داده‌ها به گروه‌های مختلف بر اساس مقادیر متغیر هدف، میزان واریانس (پراکندگی) درون هر گروه و همچنین واریانس بین گروه‌ها را محاسبه می‌کند. هدف این است که بررسی کنیم آیا ویژگی‌های مختلف می‌توانند تفاوت‌های معناداری در میان گروه‌ها ایجاد کنند یا خیر.

¹⁴ -Recursive Feature Elimination - (RFE)

¹⁵ - ANOVA

به عبارت دیگر، اگر واریانس بین گروه‌ها (تفاوت‌های میانگین‌های گروه‌های مختلف) به طور معناداری بیشتر از واریانس درون گروه‌ها (تفاوت‌های درون هر گروه) باشد، این نشان‌دهنده آن است که ویژگی می‌تواند به خوبی گروه‌ها را از یکدیگر تفکیک کند و اطلاعات مفیدی را به مدل ارائه دهد. نتیجه آزمون تحلیل واریانس، که به صورت مقدار پی^{۱۶} ارائه می‌شود، به ما می‌گوید که این تفاوت‌ها از نظر آماری معنادار هستند و می‌توانند به عنوان ویژگی‌های مهم در مدل‌سازی استفاده شوند. [8][7]

۷- تنظیم ویژگی‌های ایمنی

تنظیم ویژگی‌های ایمنی^{۱۷} یک الگوریتم انتخاب ویژگی است که به طور خاص برای شناسایی ویژگی‌های مهم در داده‌های پیچیده و دارای نویز طراحی شده است. این روش به این صورت عمل می‌کند که به طور تصادفی نمونه‌هایی از داده‌ها را انتخاب کرده و برای هر نمونه، نزدیک‌ترین نمونه‌های مشابه (با برچسب مشابه) و نزدیک‌ترین نمونه‌های متفاوت (با برچسب متفاوت) را شناسایی می‌کند. سپس، اهمیت هر ویژگی بر اساس توانایی آن در تفکیک نمونه‌های مشابه از نمونه‌های متفاوت محاسبه می‌شود. تنظیم ویژگی‌های ایمنی به هر ویژگی امتیازی اختصاص می‌دهد که نشان‌دهنده تأثیر آن در تفکیک کلاس‌ها و گروه‌بندی نمونه‌ها است؛ ویژگی‌هایی که توانایی بالاتری در تفکیک گروه‌ها دارند، امتیاز بیشتری دریافت می‌کنند. این روش به دلیل استفاده از نمونه‌های نزدیک و تحلیل مقادیر درون‌گروهی و میان‌گروهی، قادر به شناسایی ویژگی‌های مؤثر حتی در صورت وجود نویز و وابستگی‌های غیرخطی در داده‌ها است و به طور مؤثری ویژگی‌های غیرضروری را کاهش می‌دهد [9].

۸- تست‌های آماری^{۱۸} (الگوریتم ژنتیک)^{۱۹}

تست‌های آماری در انتخاب ویژگی روشی برای ارزیابی اهمیت ویژگی‌ها بر اساس تحلیل‌های آماری هستند که برای تعیین ارتباط بین ویژگی‌ها و متغیر هدف استفاده می‌شوند. به طور معمول، این تست‌ها شامل آزمون‌هایی مانند تست کای‌دو^{۲۰} برای بررسی استقلال بین ویژگی‌های دسته‌ای و متغیر هدف و (تست تی) تی^{۲۱} برای مقایسه میانگین‌های دو گروه مختلف ویژگی‌های عددی هستند. این آزمون‌ها کمک می‌کنند تا ویژگی‌هایی که دارای تأثیر معنادار بر پیش‌بینی متغیر هدف هستند، شناسایی و انتخاب شوند.

به عبارت دیگر، تست‌های آماری به شناسایی ویژگی‌هایی که تفاوت‌های آماری قابل توجهی با متغیر هدف دارند، کمک می‌کنند و ویژگی‌های کم‌اهمیت یا غیرمفید را از مجموعه داده‌ها حذف می‌کنند. این روش‌ها به ساده‌سازی مدل‌ها و بهبود عملکرد آن‌ها از طریق انتخاب ویژگی‌های معنادار و کاهش پیچیدگی داده‌ها کمک می‌کنند [10].

۹- تحلیل مؤلفه‌های اصلی

¹⁶ - p-value

¹⁷ - Relief F

¹⁸ - Statistical tests

¹⁹ - Genetic algorithm

²⁰ - Chi-Square Test

²¹ - t (t-Test)

تحلیل مؤلفه های اصلی^{۲۲} روشی برای انتخاب ویژگی ها و کاهش ابعاد داده ها است که به تحلیل ویژگی های اصلی و شناسایی مؤلفه های اصلی جدید که بیشترین واریانس داده ها را توضیح می دهند، می پردازد. در تحلیل مؤلفه های اصلی، ویژگی های اصلی به مجموعه ای از مؤلفه های اصلی خطی تبدیل می شوند که هر مؤلفه جدید به طور عمده ترین واریانس موجود در داده ها را نمایان می کند و از ویژگی های اصلی به دست می آید. این مؤلفه ها به گونه ای انتخاب می شوند که کمترین همبستگی با یکدیگر داشته باشند و اطلاعات زیادی را از داده های اصلی حفظ کنند. به این ترتیب، تحلیل مؤلفه های اصلی ویژگی های مهم و مرتبط را شناسایی کرده و ویژگی های اضافی و غیرضروری را کاهش می دهد، که به ساده سازی مدل، کاهش پیچیدگی و بهبود کارایی پیش بینی کمک می کند [11].

روش های جدید : الگوریتم های تکاملی

۱- الگوریتم های ژنتیک^{۲۳}

الگوریتم ژنتیک یکی از روش های تکاملی است که بر اساس اصول انتخاب طبیعی و ژنتیک عمل می کند. در این روش، هر مجموعه ویژگی به عنوان یک کروموزوم در نظر گرفته می شود و هر ویژگی به عنوان یک ژن در کروموزوم مدل می شود. الگوریتم ژنتیک با ایجاد یک جمعیت اولیه از کروموزوم ها و اعمال عملیات های ژنتیکی نظیر انتخاب، تقاطع^{۲۴} و جهش^{۲۵} به تدریج کروموزوم های بهینه تر را تولید می کند. هدف از این فرآیند، یافتن مجموعه ای از ویژگی ها است که بتوانند عملکرد مدل یادگیری را به حداکثر برسانند. عملیات انتخاب در الگوریتم ژنتیک، براساس عملکرد هر کروموزوم در مدل یادگیری انجام می شود، به این صورت که کروموزوم های با عملکرد بهتر احتمال بیشتری برای انتخاب و تولید نسل بعد دارند. عملیات تقاطع باعث ترکیب دو کروموزوم والد و تولید فرزندان جدید می شود که ممکن است ویژگی های بهتری را نسبت به والدین داشته باشند. عملیات جهش نیز با تغییر تصادفی بخشی از کروموزوم، تنوع در جمعیت را افزایش می دهد و از گیر افتادن الگوریتم در بهینه های محلی جلوگیری می کند. الگوریتم ژنتیک به دلیل قابلیت جستجوی گسترده و مؤثر خود در فضای ویژگی ها، به عنوان یکی از روش های محبوب انتخاب ویژگی در مسائل پیچیده و با ابعاد بالا شناخته می شود [12].

۲- بهینه سازی کلونی مورچه ها^{۲۶}

الگوریتم کلونی مورچه ها یک روش الهام گرفته از رفتار اجتماعی مورچه ها است که در طبیعت از طریق گذاشتن فرومون^{۲۷} برای یافتن کوتاه ترین مسیر به غذا عمل می کنند. در انتخاب ویژگی، هر مورچه به عنوان یک عامل مستقل عمل می کند که به طور تدریجی و با توجه به اطلاعات فرومون ها، مسیرهای مختلفی را در فضای ویژگی ها طی می کند تا بهترین مجموعه ویژگی ها را پیدا کند.

²² - Principal Component Analysis - PCA

²³ - Genetic Algorithms

²⁴ - Crossover

²⁵ - Mutation

²⁶ - Ant Colony Optimization- ACO

²⁷ - Pheromone

فرومون‌ها به عنوان یک حافظه جمعی عمل می‌کنند که اطلاعاتی درباره‌ی مسیرهای موفق و ناکام قبلی را ذخیره می‌کنند. در این روش، هر بار که یک مورچه مسیر موفق را طی می‌کند، میزان فرومون روی آن مسیر افزایش می‌یابد، که باعث جذب بیشتر مورچه‌های دیگر به همان مسیر می‌شود.

از سوی دیگر، مسیرهایی که کمترین میزان موفقیت را داشته‌اند، به تدریج فرومون‌های خود را از دست می‌دهند و کمتر مورد استفاده قرار می‌گیرند. این فرآیند به تدریج منجر به کشف بهترین مجموعه ویژگی‌ها می‌شود که بالاترین عملکرد را در مدل یادگیری دارند. الگوریتم کلونی مورچگان به دلیل طبیعت پویا و انعطاف‌پذیر خود، به عنوان یکی از روش‌های مؤثر در حل مسائل انتخاب ویژگی در داده‌های پیچیده و بزرگ شناخته می‌شود [13].

۳- بهینه‌سازی ازدحام ذرات²⁸

الگوریتم بهینه‌سازی ازدحام ذرات از رفتار جمعی پرندگان الهام گرفته است. در این الگوریتم، هر ذره (که نمایانگر یک مجموعه ویژگی است) در فضای جستجو حرکت می‌کند و براساس موقعیت خود و بهترین موقعیتی که تاکنون دیده است، موقعیت خود را به‌روزرسانی می‌کند. هر ذره دارای دو ویژگی مهم است: سرعت²⁹ و موقعیت³⁰. سرعت هر ذره تعیین می‌کند که در مرحله بعدی به کدام سمت حرکت کند، و موقعیت ذره نشان‌دهنده مجموعه ویژگی‌هایی است که در آن لحظه در حال بررسی است. الگوریتم بهینه‌سازی ازدحام ذرات از دو نوع اطلاعات برای هدایت ذرات استفاده می‌کند: بهترین موقعیت فردی³¹ و بهترین موقعیت گروهی³². بهترین موقعیت فردی، بهترین مجموعه ویژگی‌هایی است که هر ذره تاکنون پیدا کرده است، و بهترین موقعیت گروهی، بهترین مجموعه ویژگی‌هایی است که تمام ذرات تاکنون یافته‌اند. ترکیب این دو اطلاعات به ذرات کمک می‌کند تا به تدریج به سمت مناطق بهتر در فضای جستجو حرکت کنند و بهترین مجموعه ویژگی‌ها را پیدا کنند. الگوریتم بهینه‌سازی ازدحام ذرات به دلیل سادگی، کارایی بالا و قابلیت جستجوی همزمان در چندین منطقه از فضای ویژگی‌ها، به عنوان یکی از روش‌های موفق در انتخاب ویژگی شناخته می‌شود [14].

۴- رویکردهای ترکیبی³³

رویکرد ترکیبی یک روش پیشرفته در انتخاب ویژگی است که سعی دارد بهترین‌های هر دو دنیای روش‌های سنتی و تکاملی را ترکیب کند تا نتایج بهینه‌ای را در مسائل پیچیده انتخاب ویژگی ارائه دهد. این رویکرد به این صورت عمل می‌کند که ابتدا از یک روش سنتی، مانند فیلتر یا فراابتکار، برای کاهش تعداد ویژگی‌ها استفاده می‌کند و سپس از یک الگوریتم تکاملی برای بهینه‌سازی مجموعه ویژگی‌های انتخاب شده بهره می‌برد. هدف از این ترکیب، دستیابی به سرعت بالای روش‌های سنتی در مراحل اولیه و دقت و کارایی بالای روش‌های تکاملی در مراحل بعدی است. به عنوان مثال، ممکن است ابتدا از تحلیل مؤلفه‌های اصلی برای کاهش بعد و انتخاب ویژگی‌های مهم‌تر استفاده شود، سپس این ویژگی‌ها به عنوان ورودی به یک الگوریتم ژنتیک یا بهینه‌سازی ازدحام ذرات داده شوند تا

²⁸ - Particle Swarm Optimization - PSO

²⁹ - Velocity

³⁰ - Position

³¹ - Personal Best

³² - Global Best

³³ - Hybrid Approach

بهترین ترکیب از ویژگی‌ها انتخاب شود. این روش باعث می‌شود که فضای جستجوی الگوریتم تکاملی کاهش یابد و در نتیجه فرآیند بهینه‌سازی سریع‌تر و کارآمدتر انجام شود.

همچنین، استفاده از رویکرد ترکیبی می‌تواند خطر گیر افتادن در بهینه‌های محلی را که در الگوریتم‌های تکاملی به تنهایی وجود دارد، کاهش دهد. رویکردهای ترکیبی در بسیاری از مطالعات اخیر نشان داده‌اند که می‌توانند به طور قابل توجهی عملکرد مدل‌های یادگیری ماشین را بهبود بخشند، به خصوص در مسائل با داده‌های پیچیده و بزرگ. با این حال، انتخاب ترکیب مناسب از روش‌های سنتی و تکاملی نیازمند دانش و تجربه کافی است، و تحقیقات بیشتری برای توسعه و بهینه‌سازی این رویکردها لازم است.

۵- الگوریتم‌های دیفرانسیلی^{۳۴}

الگوریتم‌های دیفرانسیلی یکی از روش‌های تکاملی موثر در انتخاب ویژگی‌ها است که به طور خاص برای بهینه‌سازی مسائل پیچیده و پیوسته طراحی شده است. در این روش، ابتدا یک جمعیت اولیه از راه‌حل‌ها که هر کدام نمایانگر یک مجموعه از ویژگی‌ها هستند، به صورت تصادفی ایجاد می‌شود. سپس، هر راه‌حل براساس عملکرد آن در مدل یادگیری ماشین ارزیابی می‌شود. الگوریتم‌های دیفرانسیلی از سه عملگر اصلی جهش، ترکیب، و انتخاب برای تولید نسل‌های جدید استفاده می‌کند. در مرحله جهش، تفاوت بین چند بردار به یک بردار دیگر اضافه می‌شود تا یک بردار جهش‌یافته جدید ایجاد شود. سپس، در مرحله ترکیب، این بردار جدید با بردار اصلی ترکیب می‌شود تا یک راه‌حل فرزند تولید شود. در نهایت، در مرحله انتخاب، بهترین راه‌حل‌ها از میان راه‌حل‌های والد و فرزند انتخاب می‌شوند تا نسل بعدی را تشکیل دهند. این فرآیند تکرار می‌شود تا زمانی که مجموعه‌ای از ویژگی‌های بهینه برای مدل به دست آید. الگوریتم‌های دیفرانسیلی به دلیل توانایی در جستجوی فضای گسترده ویژگی‌ها و پیدا کردن ترکیبات بهینه از میان آن‌ها، به ویژه در مسائل پیچیده و داده‌های بزرگ، بسیار موثر است.

۶- الگوریتم‌های برنامه‌نویسی ژنتیکی

الگوریتم‌های برنامه‌نویسی ژنتیکی^{۳۵} یکی از روش‌های تکاملی پیشرفته است که برای انتخاب ویژگی‌ها نیز کاربرد دارد. در الگوریتم‌های برنامه‌نویسی ژنتیکی، به جای استفاده از بردارهای عددی یا باینری، هر راه‌حل به صورت یک درخت برنامه‌نویسی نمایش داده می‌شود که می‌تواند عملیات ریاضی، منطقی، یا تابع‌های پیچیده‌تر را شامل شود. این درخت‌ها به عنوان مدل‌های پیشنهادی عمل می‌کنند که بهترین ترکیب ویژگی‌ها را شناسایی می‌کنند. فرآیند الگوریتم‌های برنامه‌نویسی ژنتیکی شامل مراحل انتخاب، جهش، و ترکیب است که به صورت مکرر اعمال می‌شود تا درخت‌هایی تولید شوند که با استفاده از مجموعه ویژگی‌های بهینه، دقت پیش‌بینی مدل یادگیری ماشین را افزایش دهند. یکی از مزایای الگوریتم‌های برنامه‌نویسی ژنتیکی این است که قادر به کشف روابط غیرخطی پیچیده بین

³⁴ - Differential Evolution - DE

³⁵ - Genetic Programming - GP

ویژگی‌ها است که ممکن است در روش‌های سنتی نادیده گرفته شوند. با این حال، اجرای الگوریتم‌های برنامه‌نویسی ژنتیکی به دلیل ساختار پیچیده درختی و نیاز به محاسبات فراوان، زمان‌بر و محاسباتی سنگین است.

۷- یادگیری عمیق

انتخاب ویژگی با استفاده از یادگیری عمیق^{۳۶} یک رویکرد نوآورانه و موثر است که از قدرت شبکه‌های عصبی عمیق برای شناسایی ویژگی‌های مهم و تاثیرگذار در مجموعه داده‌های بزرگ و پیچیده بهره می‌برد. در این روش، شبکه‌های عصبی چندلایه با ساختارهای پیچیده‌ای مانند شبکه‌های عصبی کانولوشنی^{۳۷} و شبکه‌های عصبی بازگشتی^{۳۸} به طور خودکار و بدون نیاز به مداخله انسانی، ویژگی‌های مناسب را از داده‌ها استخراج می‌کنند. فرآیند انتخاب ویژگی در یادگیری عمیق به طور معمول با آموزش شبکه‌ای بزرگ بر روی داده‌های خام آغاز می‌شود. لایه‌های اولیه شبکه عصبی معمولاً به شناسایی ویژگی‌های ساده و سطح پایین مانند لبه‌ها در تصاویر یا الگوهای اولیه در داده‌های سری زمانی می‌پردازند.

سپس، لایه‌های میانی و بالایی شبکه این ویژگی‌های ساده را ترکیب کرده و به ساخت ویژگی‌های سطح بالا و پیچیده‌تر مانند اشیاء یا روابط بلندمدت در داده‌ها ادامه می‌دهند. یکی از تکنیک‌های رایج در این زمینه استفاده از رگولایزاسیون است که به کاهش وزن ویژگی‌های غیرمهم یا کم‌اهمیت در طول فرآیند آموزش کمک می‌کند. این کار به جلوگیری از پیچیدگی بیش از حد مدل و بهینه‌سازی عملکرد آن منجر می‌شود. همچنین، تکنیک‌های توجه^{۳۹} به شبکه‌ها امکان می‌دهند تا بر ویژگی‌های خاص و مهم تمرکز کرده و تاثیر بیشتری به آن‌ها بدهند، که این امر به بهبود دقت و کارایی مدل کمک می‌کند. مزیت اصلی استفاده از یادگیری عمیق در انتخاب ویژگی این است که این روش قادر است روابط غیرخطی پیچیده و تعاملات میان ویژگی‌ها را که در روش‌های سنتی ممکن است نادیده گرفته شوند، شناسایی کند. این ویژگی به خصوص در مسائل با داده‌های حجیم و پیچیده مانند پردازش تصویر، پردازش زبان طبیعی، و تحلیل داده‌های پزشکی، که ویژگی‌های مهم ممکن است در داده‌های پرنویز یا با ابعاد بالا مخفی شده باشند، بسیار مهم است. با این حال، انتخاب ویژگی مبتنی بر یادگیری عمیق نیاز به منابع محاسباتی قوی و داده‌های زیادی برای آموزش دارد.

همچنین، تنظیم و بهینه‌سازی مدل‌های یادگیری عمیق به دلیل پیچیدگی بالای آن‌ها نیازمند دانش فنی و تجربه است. به رغم این چالش‌ها، این روش به دلیل دقت و قابلیت اطمینان بالا در مسائل پیچیده و داده‌های بزرگ، به یک ابزار قدرتمند و مؤثر در انتخاب ویژگی تبدیل شده است [15].

۷- روش‌های مبتنی بر شبکه‌های عصبی پیچیده^{۴۰} (رمزگذارهای خودکار)^{۴۱}

انتخاب ویژگی با استفاده از روش‌های مبتنی بر شبکه‌های عصبی پیچیده، مانند رمزگذارهای خودکار، به تحلیل و استخراج ویژگی‌های اصلی از داده‌های پیچیده کمک می‌کند. رمزگذارهای خودکار شبکه‌های عصبی هستند که برای یادگیری نمایش فشرده و کم‌بعد از داده‌ها طراحی شده‌اند.

³⁶ - Deep Learning-based Feature Selection

³⁷ - CNNs

³⁸ - RNNs

³⁹ - Attention

⁴⁰ - Methods based on complex neural networks

⁴¹ - Autoencoder

این شبکه‌ها شامل دو بخش اصلی هستند :

رمزگذار^{۴۲} : که ورودی‌های داده را به یک فضای ویژگی کم‌بعدتر نقشه می‌کند و رمزگشا^{۴۳} : که سعی می‌کند داده‌ها را از فضای فشرده به حالت اولیه خود بازسازی کند. هدف اصلی این است که ویژگی‌های اصلی و غیرضروری را شناسایی کرده و به یک نمایش فشرده از داده‌ها دست یابند .

با تحلیل لایه‌های میانی رمزگذارهای خودکار ، می‌توان ویژگی‌های مهم و برجسته را شناسایی کرد، چرا که این لایه‌ها نمایانگر ویژگی‌های کلیدی و غیرخطی داده‌ها هستند. این روش به کاهش ابعاد داده‌ها، کاهش پیچیدگی مدل و بهبود کارایی پیش‌بینی با تمرکز بر ویژگی‌های مهم و معنادار کمک می‌کند[16].

۹- شبکه‌های بیزین

انتخاب ویژگی با استفاده از شبکه‌ها^{۴۴} به تحلیل روابط احتمالاتی پیچیده بین ویژگی‌ها و متغیر هدف می‌پردازد تا ویژگی‌های مهم و تأثیرگذار را شناسایی کند. شبکه‌های بیزین مدل‌های گرافیکی هستند که وابستگی‌های احتمالاتی بین متغیرها را با استفاده از گراف‌های جهت‌دار و احتمال‌های شرطی نمایش می‌دهند. در این روش، با ساخت و تحلیل شبکه بیزین، می‌توان تأثیر هر ویژگی بر متغیر هدف را ارزیابی کرد و ویژگی‌هایی که بیشترین تأثیر را دارند، شناسایی نمود .

شبکه‌های بیزین به دلیل توانایی‌شان در مدل‌سازی تعاملات پیچیده و عدم قطعیت‌ها، می‌توانند ویژگی‌های کلیدی را که اطلاعات زیادی درباره پیش‌بینی متغیر هدف ارائه می‌دهند، مشخص کنند. به این ترتیب، این روش به انتخاب ویژگی‌های موثر و کاهش ویژگی‌های غیرضروری کمک می‌کند، که به بهبود دقت مدل و ساده‌سازی آن منجر می‌شود.

مقایسه روش‌های سنتی و جدید

(۱) مقایسه عملکرد

روش‌های سنتی انتخاب ویژگی مانند فیلترها و روش‌های پوششی به دلیل سرعت بالا و سادگی در بسیاری از مسائل کاربرد دارند. با این حال، این روش‌ها در مواجهه با داده‌های با ابعاد بالا و پیچیدگی‌های زیاد دچار محدودیت‌هایی می‌شوند. از سوی دیگر، الگوریتم‌های تکاملی، که بر اصول تکامل طبیعی و رفتارهای اجتماعی استوارند، قادر به کاوش در فضای جستجوی بسیار بزرگی هستند و می‌توانند به نتایج بهینه دست یابند. این الگوریتم‌ها با ترکیب ویژگی‌ها و تعاملات بین آن‌ها می‌توانند عملکرد مدل‌های یادگیری ماشین را به طور قابل توجهی بهبود بخشند.

(۲) مقیاس‌پذیری

⁴² - Encoder

⁴³ - Decoder

⁴⁴ -Bayesian Networks

روش‌های سنتی، به ویژه در روش‌های پوششی و جاسازی شده، نیاز به ارزیابی تمامی ترکیبات ممکن از ویژگی‌ها دارند، که با افزایش تعداد ویژگی‌ها به مشکل برمی‌خورند.

در مقابل، الگوریتم‌های تکاملی مانند الگوریتم ژنتیک، الگوریتم بهینه‌سازی کلونی مورچگان، و الگوریتم بهینه‌سازی ازدحام ذرات به دلیل قابلیت جستجوی همزمان در چندین مسیر، به طور موثری مقیاس‌پذیری را افزایش داده و برای داده‌های بزرگ مناسب‌تر هستند.

۳) تفسیرپذیری

روش‌های سنتی انتخاب ویژگی به دلیل استفاده از مدل‌های ساده‌تر و قابل تفسیرتر مانند رگرسیون لاسو و درخت‌های تصمیم، معمولاً تفسیرپذیری بیشتری دارند. در مقابل، الگوریتم‌های تکاملی به دلیل فرآیندهای پیچیده‌تر و جستجوی در فضای بزرگ‌تر، ممکن است تفسیرپذیری کمتری داشته باشند. با این حال، می‌توان با ترکیب این الگوریتم‌ها با روش‌های تفسیرپذیر، این مشکل را تا حدی کاهش داد. این رویکرد ترکیبی می‌تواند توازن بین بهینه‌سازی عملکرد و حفظ تفسیرپذیری در مدل‌های یادگیری ماشین ایجاد کند.

روش‌های سنتی و جدید در انتخاب ویژگی‌ها از نظر عملکرد و کاربرد تفاوت‌های قابل توجهی دارند. روش‌های سنتی مانند فیلتر و پوششی در انتخاب ویژگی به شدت وابسته به تکنیک‌های آماری و مدل‌های پایه هستند، به طوری که فیلترها بر اساس معیارهای آماری ساده و بدون توجه به تعاملات پیچیده ویژگی‌ها عمل می‌کنند، در حالی که پوششی‌ها با ارزیابی ترکیب‌های مختلف ویژگی‌ها از طریق یک مدل یادگیری ماشین، به انتخاب بهینه‌تری می‌پردازند، اما این روش‌ها زمان‌بر و محاسباتی سنگین هستند. در مقابل، روش‌های جدید مانند انتخاب ویژگی بر اساس تکنیک‌های یادگیری عمیق و الگوریتم‌های پیشرفته‌تر، توانایی بیشتری در شناسایی ویژگی‌های پیچیده و تعاملات غیرخطی دارند و می‌توانند در داده‌های بزرگ و پیچیده عملکرد بهتری داشته باشند. این روش‌ها ممکن است نیاز به منابع محاسباتی بیشتری داشته باشند و پیچیدگی بیشتری در پیاده‌سازی داشته باشند، اما به طور کلی توانایی بیشتری در بهبود دقت مدل‌های یادگیری ماشین ارائه می‌دهند.

در این پروژه انتخاب ویژگی‌ها با استفاده از الگوریتم‌های مختلف و دو روش مختلف انتخاب ویژگی، یعنی انتخاب ویژگی مبتنی بر فیلتر و انتخاب ویژگی مبتنی بر پوشش توصیف شده است. با توجه به سال‌های گذشته، می‌توان مشاهده کرد که استفاده از الگوریتم‌های مختلف در زمینه‌های مختلف، با نوسانات قابل توجهی همراه بوده است. استفاده از الگوریتم‌های الهام‌گرفته از طبیعت برای انتخاب ویژگی‌ها نقش مهمی را ایفا می‌کند. هر دو روش فیلتر و پوشش در طول سال‌ها برای انتخاب ویژگی‌ها استفاده شده است.

نتیجه‌گیری

در این مقاله، روش‌های سنتی و جدید انتخاب ویژگی مورد بررسی قرار گرفتند. روش‌های سنتی هرچند که در بسیاری از مسائل کارآمد هستند، اما در مواجهه با داده‌های پیچیده و با ابعاد بالا دچار محدودیت‌هایی می‌شوند. الگوریتم‌های تکاملی با قابلیت جستجوی همزمان و کاوش در فضای بزرگ‌تر، به طور مؤثری می‌توانند به بهبود عملکرد انتخاب ویژگی کمک کنند.

روش‌های جدید، به ویژه الگوریتم‌های تکاملی و ترکیبی، چشم‌اندازهای جذابی برای انتخاب ویژگی در مسائل پیچیده فراهم کرده‌اند. با این حال، هنوز چالش‌هایی مانند بهبود مقیاس‌پذیری و کاهش پیچیدگی محاسباتی وجود دارد که نیازمند تحقیقات بیشتری هستند. پژوهش‌های آینده می‌توانند به توسعه الگوریتم‌های کارآمدتر و ترکیب بهتر روش‌های مختلف برای بهبود انتخاب ویژگی کمک کنند.

روش‌های انتخاب ویژگی مبتنی بر الگوریتم‌های تکاملی، به دلیل توانایی بالای خود در جستجوی بهینه و مدیریت فضای بزرگ ویژگی‌ها، ابزاری قدرتمند و کارآمد برای بهبود مدل‌های یادگیری ماشین هستند. الگوریتم‌هایی نظیر ژنتیک، کلونی مورچه‌ها، و بهینه‌سازی ازدحام ذرات توانسته‌اند در مواجهه با داده‌های پیچیده و بزرگ، عملکرد بهتری نسبت به روش‌های سنتی ارائه دهند. این الگوریتم‌ها با الهام از طبیعت و تکامل، قابلیت انعطاف‌پذیری و سازگاری با مسائل مختلف را دارند، که منجر به کاهش ابعاد داده‌ها و بهبود کارایی مدل‌های یادگیری می‌شود. علاوه بر این، با توجه به مزایا و محدودیت‌های هر کدام از این روش‌ها، استفاده از ترکیبی از آن‌ها یا به کارگیری تکنیک‌های هیبریدی می‌تواند راه‌حل‌های بهینه‌تری را برای مسائل پیچیده فراهم کند. تحقیقات آینده می‌تواند به توسعه و بهبود این الگوریتم‌ها و کاربردهای آن‌ها در زمینه‌های مختلف از جمله بیوانفورماتیک و داده‌کاوی متنی تمرکز کند.

منابع

۱. S. Ayesha, M. K. Hanif, and R. Talib, "Overview and comparative study of dimensionality reduction techniques for high dimensional data," *Information Fusion*, vol. 59, pp. 44-58, 2020.
۲. Palamidessi, Catuscia, and Marco Romanelli. 2020. "Feature selection in machine learning: R\'enyi min-entropy vs Shannon entropy." *arXiv preprint arXiv:2001.09654*. (۲۰۲۰)
۳. Rathee, Seema, and Saroj Ratnoo. 2020. "Feature selection using multi-objective CHC genetic algorithm." *Procedia Computer Science*, 167: 1656-1664
۴. A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction
۵. Optimized multi correlation-based feature selection in software defect prediction 2023
۶. Botnet attacks classification in AMI networks with recursive feature elimination (RFE) and machine learning algorithms 2023
۷. Novel Meta-Heuristic Algorithm for Feature Selection, Unconstrained Functions and Engineering Problems 2022

۸. Feature selection in wind speed forecasting systems based on meta-heuristic optimization2023
۹. A feature selection approach based on NSGA-II with ReliefF2023
۱۰. Fusion of statistical importance for feature selection in Deep Neural Network-based Intrusion Detection System2023
۱۱. A Tutorial on Principal Component Analysis for Dimensionality Reduction in Machine Learning2023
۱۲. Estrela, Gustavo, Marco Dimas Gubitoso, Carlos Eduardo Ferreira, Junior Barrera, and Marcelo S.Reis. 2020. "An efficient, parallelized algorithm for optimal conditional entropy-based feature
۱۳. Exploring Ant Colony Optimization for Feature Selection: A Comprehensive Review2024
۱۴. An enhanced particle swarm optimization with position update for optimal feature selection2024
۱۵. Enhancing intrusion detection in IoT networks using machine learning-based feature selection and ensemble models2024
۱۶. Waste classification using AutoEncoder network with integrated feature selection method in convolutional neural network models2020

Investigation of feature selection methods based on evolutionary algorithms (genetics, ant colony, PSO, etc.)

Hassan Nosrati Nahok

Instructor of Technical and Engineering Faculty, Saravan University, Saravan, Iran

Narges Khosravi Monfared

Bachelor student of Computer Engineering, Technical and Engineering Faculty, Saravan University, Saravan, Iran

abstract

Feature selection as a key step in the machine learning process plays an important role in reducing computational complexity and improving the efficiency of learning models. This paper reviews different feature selection methods based on evolutionary algorithms, including genetic algorithm, ant colony, and particle swarm optimization. The purpose of this study is to provide a comprehensive overview of these algorithms, evaluate their performance compared to traditional methods, and analyze the advantages and disadvantages of each in feature selection for large and complex datasets.

The findings of this paper show that evolutionary algorithms are powerful tools for feature selection due to their capabilities in searching large feature spaces. Finally, this paper discusses the results and suggestions for future research in this field.

Keywords

Feature selection, evolutionary algorithms, genetic algorithms, ant colony optimization, particle swarm optimization