TUMJ مجله دانشکده پزشکی، دانشگاه علوم پزشکی تهران، فروردین ۱۳۹۸، دوره ۷۷، شماره ۱، صفحه‌های ۳۳ تا ۴۰ مقاله اصیل

پیش‌بینی بیماری مولتیپل اسکلروزیس با استفاده از رویکردهای داده‌کاوی جنگل تصادفی و ماشین‌بردار پشتیبان بر اساس الگوریتم ژنتیک

امیرحسین هاشمیان[۱و۲]، سارا منوچهری[۳*]
داریوش افشاری[۴]، زهره منوچهری[۵]
نادر سالاری[۶]، سوده شهسواری[۷]

۱- مرکز تحقیقات عوامل محیطی مؤثر بر سلامت، دانشگاه علوم پزشکی کرمانشاه، کرمانشاه، ایران.
۲- گروه آمار زیستی، دانشکده بهداشت، دانشگاه علوم پزشکی کرمانشاه، کرمانشاه، ایران.
۳- گروه آمار زیستی، کمیته تحقیقات دانشجویی، دانشکده بهداشت، دانشگاه علوم پزشکی کرمانشاه، کرمانشاه، ایران.
۴- گروه نورولوژی، دانشکده پزشکی، دانشگاه علوم پزشکی کرمانشاه، کرمانشاه، ایران.
۵- گروه آمار زیستی، کمیته تحقیقات دانشجویی، دانشگاه علوم پزشکی همدان، همدان، ایران.
۶- گروه آمار زیستی، دانشکده بهداشت، دانشگاه علوم پزشکی کرمانشاه، کرمانشاه، ایران.
۷- گروه مدیریت اطلاعات سلامت، دانشکده پیراپزشکی، دانشگاه علوم پزشکی کرمانشاه، کرمانشاه، ایران.

*نویسنده مسئول: کرمانشاه، خیابان شهید باهنر، دانشگاه علوم پزشکی کرمانشاه، دانشکده بهداشت، کمیته تحقیقات دانشجویی.
تلفن: ۰۸۳-۳۸۳۹۸۰۶۹
E-mail: saramanochehri0098@gmail.com

چکیده

دریافت: ۱۳۹۷/۰۷/۲۹ ویرایش: ۱۳۹۷/۰۸/۰۶ پذیرش: ۱۳۹۸/۰۱/۲۰ آنلاین: ۱۳۹۸/۰۱/۳۰

زمینه و هدف: مولتیپل اسکلروزیس (MS) یک بیماری التهابی تخریب‌کننده است که روش مهم تشخیص آن استفاده از دستگاه Magnetic resonance imaging (MRI) است، اما ازآنجاکه MRI از یک میدان مغناطیسی بسیار پرقدرت استفاده می‌کند، در صورت وجود اجسام فلزی در بدن بیماران باعث ایجاد اختلال در وضعیت سلامت بیمار، کارکرد دستگاه و نیز انحراف و تاری در تصاویر می‌شود. با توجه به چنین محدودیتی در استفاده از دستگاه MRI نیاز به‌روش کمک غربالگری احساس می‌شود. بنابراین این مطالعه با هدف مقایسه دو مدل ماشین‌بردار پشتیبان (Support vector machine, SVM) و جنگل تصادفی (Random forest, RF) انجام شد.

روش بررسی: پژوهش کنونی تحلیلی و از نوع مدل‌سازی بود که از اردیبهشت ۱۳۹۶ تا شهریور ۱۳۹۷ در دانشکده بهداشت دانشگاه علوم پزشکی کرمانشاه انجام شد. روش پیشنهادی پژوهش بر روی مجموعه داده‌های بیماری MS که مشخصات آن‌ها در سیستم ثبت سلامت دانشگاه علوم پزشکی کرمانشاه ثبت گردیده، پیاده‌سازی شد. تعداد افراد مورد مطالعه ۳۱۷ نفر بودند (۱۸۸ نفر مبتلابه بیماری MS و ۱۲۸ نفر فاقد آن). به‌منظور برازش مدل SVM، از تابع کرنل شعاع مبنا (Radial basis function, RBF) که پارامترهای آن با الگوریتم ژنتیک (Genetic algorithm, GA) بهینه‌سازی شده‌اند، استفاده شد. سپس مدل SVM با استفاده از معیارهای صحت، حساسیت و ویژگی با مدل RF مقایسه گردید.

یافته‌ها: بر اساس نتایج، صحت، حساسیت و ویژگی مدل SVM به‌ترتیب ۰/۷۹، ۰/۸۰، ۰/۷۸ و برای مدل RF به‌ترتیب ۰/۷۶، ۰/۸۱، ۰/۷۰ به‌دست آمد.

نتیجه‌گیری: دو مدل عملکرد مناسبی داشتند، اما با توجه به صحت به‌عنوان یک معیار مهم برای مقایسه عملکرد مدل‌ها در این حوزه، می‌توان گفت مدل SVM کارایی بهتری نسبت به RF در تشخیص بیماری MS داشت.

کلمات کلیدی: الگوریتم ژنتیک، مولتیپل اسکلروزیس، جنگل تصادفی، ماشین‌بردار پشتیبان.

مقدمه

مولتیپل اسکلروزیس (Multiple sclerosis, MS) یک بیماری التهابی تخریب‌کننده است که با آسیب رساندن به پوشش عایق سلول‌های عصبی، دارای اثرات مخرب روی مغز و نخاع انسان است. در بیماری MS سلول‌هایی که مسئولیت ایجاد و حفظ لایه چربی موسوم به غلاف میلین در مغز را بر عهده دارند از بین می‌رود. لایه چربی (غلاف میلین) در انتقال سیگنال‌های الکتریکی به نورون‌ها کمک می‌کنند. MS شایع‌ترین بیماری دمیلینه‌کننده سیستم عصبی مرکزی و سومین علت اصلی ناتوانی پس از بیماری‌های تروما و روماتیسم است. این بیماری در زنان شایع‌تر از مردان است و به‌طور عمده افراد بین سنین ۴۰-۲۰ سال را تحت تأثیر قرار می‌دهد.[۱] در سال ۲۰۱۳ تعداد بیماران مبتلابه MS ۲/۳ میلیون نفر در سراسر جهان گزارش شده که نسبت به پنج سال پیش از آن ۰/۲ میلیون نفر افزایش یافته است، که از

این تعداد کمابیش ٤۰۰,۰۰۰ نفر مربوط به ایالت متحده امریکا بوده است.[۲] ایران از نظر شیوع بیماری MS، جزو کشورهای با میزان شیوع متوسط به بالا طبقه‌بندی می‌شود که شیوع این بیماری در کشور ایران ٥٤/٥۱ و بروز آن ٥/۸۷ در هر صد هزار نفر برآورد شده است.[۳] میزان شیوع بیماری در استان کرمانشاه ٤۳/۳ در هر ۱۰۰,۰۰۰ نفر گزارش شده است که برای زنان ٦٥/٦ نفر و در مردان ۲۱/۹ در هر ۱۰۰۰۰۰ نفر می‌باشد.[٤] علت اصلی بیماری MS به‌طور کامل قابل درک نیست اما این بیماری در افراد مستعد ژنتیکی که در معرض عوامل محیطی و درعین‌حال تعریف‌نشده قرار گیرند گسترش پیدا می‌کند.[۳] پزشکان به‌منظور تشخیص بیماری MS از روش‌های مختلفی استفاده می‌کنند ازجمله معاینه عصب، معاینه چشم، کشیدن مایع مغزی-نخاعی و استفاده از دستگاه Magnetic resonance imaging (MRI). با وجود آزمون‌های تشخیصی اشاره‌شده، مهم‌ترین روش برای تشخیص بیماری MS استفاده از دستگاه MRI می‌باشد.[٥] اما ازآنجاکه دستگاه MRI از یک میدان مغناطیسی بسیار پرقدرت استفاده می‌کند، در صورت وجود اجسام فلزی در بدن بیماران باعث ایجاد اختلال در وضعیت سلامت بیمار، کارکرد دستگاه و نیز انحراف در تصاویر می‌شود. برخی از این اجسام عبارتند از پیس‌میکر یا ضربان‌ساز مصنوعی قلب، دفیبریلاتور یا دستگاه تنظیم‌کننده ضربان قلب، پروتزهایی که در ناحیه سر و گردن قرار دارند، پیچ و پلاک یا دیگر اجسام فلزی که برای درمان شکستگی استخوان به‌کار می‌روند، وسایل تزریق خودکار دارو مانند تزریق‌کننده خودکار انسولین، اجسام فلزی خارجی که پیش‌تر در بدن وارد شده‌اند به‌ویژه در چشم، تیر یا ترکش در بدن، تاتو، برخی از انواع آی‌یودی، پچ‌های دارویی حاوی فلز.[٦و۷] با توجه به محدودیت بیان‌شده در استفاده از دستگاه MRI هدف از انجام این مطالعه مقایسه کارایی ۲ مدل ماشین‌بردار پشتیبان (Support vector machine, SVM) و جنگل تصادفی (Random forest, RF)، به‌منظور دستیابی به مدلی با پیچیدگی کمتر و عملکرد تشخیصی مناسب جهت تشخیص بیماران مبتلابه MS می‌باشد. درواقع مدل پیشنهادی می‌تواند به‌عنوان روش کمک غربالگری دستگاه MRI، در شرایط حساس مددرسان پزشکان قرار گیرد. لازم به یادآوری است که برخلاف مطالعاتی که تاکنون در این زمینه انجام‌شده و از تصاویر MRI به‌عنوان ورودی ماشین استفاده شده است، در این مطالعه سعی شد که بدون استفاده از تصاویر MRI و بر اساس متغیرهایی که به‌عنوان عوامل خطر بیماری MS شناخته شده است، آموزش ماشین صورت گیرد.

روش بررسی

پژوهش کنونی تحلیلی و از نوع مدل‌سازی می‌باشد که از اردیبهشت ۱۳۹٦ تا شهریور ۱۳۹۷ در دانشکده بهداشت دانشگاه علوم پزشکی کرمانشاه انجام شد. تعداد کل افراد مورد مطالعه ۳۱۷ نفر بودند (۱۸۸ نفر (٥۹/٥٪) مبتلا به بیماری MS و ۱۲۸ نفر (٤۰/٥٪) فاقد آن) که مشخصات مربوط به آن‌ها در سیستم ثبت بیماری مولتیپل اسکلروزیس (MS registry) دانشگاه علوم پزشکی کرمانشاه ثبت گردیده بود. در مسایل مربوط به مدل‌سازی حجم نمونه مورد نیاز در محدوده 5M<n<15 M قرار می‌گیرد که در آن M تعداد متغیرهای مورد بررسی و n حجم نمونه می‌باشد.[۸] ازآنجاکه تعداد متغیرهای مورد مطالعه در این پژوهش، ۲۹ متغیر می‌باشد، بنابراین برای پژوهش کنونی بین ۱٤٥ تا ٤۳٥ نمونه لازم است.

از مدل‌های پیش‌بینی به‌منظور بررسی کیفیت مدیریت مراقبت‌های بهداشتی به‌طور گسترده‌ای استفاده می‌شود. همچنین این مدل‌ها توسط مسئولان به‌منظور تصمیم‌گیری‌های بالینی به‌کار برده می‌شود.[۹] طبقه‌بندی دو حالته (باینری)، ساده‌ترین نوع یک مسئله پیش‌بینی است[۱۰] که در میان مجموعه‌ای از اعضا برای پیدا کردن ویژگی‌های مشترک و ایجاد تمایز بین آن‌ها به جستجو می‌پردازد.[۱۱] تشخیص بیماری MS یک مسئله طبقه‌بندی دوحالته است، یعنی تلاش در جهت پیدا کردن یک طبقه‌بندی‌کننده بهینه برای متمایز کردن افراد دارای MS از افراد فاقد این بیماری. مسئله طبقه‌بندی یک وظیفه اصلی در داده‌کاوی می‌باشد.[۱۲] روش‌های طبقه‌بندی مختلفی در حوزه داده‌کاوی و یادگیری ماشین مانند شبکه‌های عصبی، درختان تصمیم‌گیری، بیز ساده، نزدیک‌ترین همسایگی، RF و SVM وجود دارند که به‌منظور طبقه‌بندی داده‌ها مورد استفاده قرار گیرند. SVM یکی از این روش‌ها است که با توجه به این‌که در سال‌های اخیر توانایی‌های بهتر و عملکرد بالایی در تشخیص انواع بیماری‌ها از خود نشان داده، مورد توجه قرار گرفته است.[۱۳] SVM برای اولین‌بار به‌منظور الگوی شناسایی توسط Vapnik در سال ۱۹۸۲ ارایه شد. سپس توسط Lei برای شناسایی مکان بهینه برای حفاری زمین‌های نفت و گاز در چین، مورد استفاده قرار گرفت. این روش همچنین ابزار محاسباتی قدرتمند برای الگوی طبقه‌بندی و رگرسیون می‌باشد.[۱٤] هدف SVM پیدا

کردن بهترین تابع برای طبقه‌بندی است، به‌گونه‌ای که بتوان اعضای دو کلاس را در مجموعه داده‌ها از هم تشخیص داد.[۱۵] با فرض این‌که کلاس‌ها به‌صورت خطی جداپذیر باشند، ابر صفحه‌ای با بیشینه‌ی حاشیه ایجاد می‌شود که کلاس‌ها را تفکیک کند، اما در مسایلی که مجموعه داده‌های ورودی به‌صورت خطی تفکیک‌پذیر نباشند، داده‌ها با استفاده از تابع کرنل (Kernel function) به فضای ویژگی با بُعد بالا نگاشت پیدا می‌کنند تا بتوان آن‌ها را در این فضای جدید به‌صورت خطی جدا نمود.[۱۶و۱۷] متداول‌ترین توابع کرنل مورد استفاده در SVM عبارتند از تابع کرنل خطی، چندجمله‌ای، شعاع مبنا (Radial basis function, RBF)، سیگموید و غیره. در این مطالعه تابع کرنل RBF به‌کار گرفته شده است، زیرا دارای قدرت تعمیم‌پذیری بالاتر با پیچیدگی کمتر نسبت به تابع کرنل چندجمله‌ای می‌باشد.[۱۸] هنگام به‌کارگیری هر یک از این توابع کرنل در مدل SVM، لازم است به‌منظور افزایش عملکرد در طبقه‌بندی اشیا پارامترهای آن‌ها را به‌درستی تنظیم کنیم. الگوریتم‌های بهینه‌سازی متعدد ازجمله روش جستجوی شبکه، الگوریتم بهینه‌سازی ذرات و الگوریتم ژنتیک (Genetic algorithm, GA)، در این زمینه پیشنهاد شده‌اند.[۱۹] در پژوهش کنونی از GA به‌عنوان روش بهینه‌سازی استفاده شد.

GA توسط Holland در سال ۱۹۷۵ معرفی شد. GA روش جستجوی تصادفی در هوش مصنوعی می‌باشد که به‌منظور بهینه‌سازی جهت توابع تعریف‌شده روی دامنه محدود، معرفی شده است. GA بر اساس نظریه تکامل با انتخاب طبیعی است. این روش از فرآیندهای بیولوژیک وراثتی مانند جهش، انتخاب طبیعی و متقاطع ژنتیکی الهام گرفته شده است.[۲۰] GA از تعدادی راه‌حل ممکن، به‌نام جمعیت، به‌طور هم‌زمان برای حل یک مشکل استفاده می‌کند. هر عضو این جمعیت، رشته‌های کروموزوم نامیده می‌شود. سه مفهوم مهم در GA عبارتند از جمعیت اولیه، تابع هدف و عملگرها. در GA ابتدا چندین راه‌حل ممکن که به‌عنوان کروموزوم کدگذاری شده‌اند، به‌صورت تصادفی از میان جمعیت اولیه انتخاب‌شده و توسط تابع هدف ارزیابی می‌شوند. پس از آن، این کروموزوم‌ها برای تولید نسل بعدی توسط عملگر انتخاب، جهش و متقارن استفاده می‌شود. این روند از یک نسل به دیگری ادامه می‌یابد تا الگوریتم متوقف شود.[۲۱]

مدل بعدی مورد استفاده در این مطالعه الگوریتم جنگل تصادفی است که روشی مبتنی بر یادگیری ماشین می‌باشد و نخستین‌بار در سال ۲۰۰۱ توسط Breiman معرفی شد. RF یک روش ناپارامتری و متعلق به خانواده‌ی روش‌های یادگیری گروهی است که شامل گروهی از درختان طبقه‌بندی و رگرسیون می‌باشد. RF متشکل از n درخت $\{T_1(X),T_2(X),...,T_n(X)\}$ است که $X=\{x_1,x_2,...,x_M\}$ یک بردار M بعدی از متغیرها می‌باشد. مجموعه n درخت، تعداد n خروجی به‌صورت $\{\hat{Y}_1=T_1(X),\hat{Y}_2=T_2(X),...,\hat{Y}_n=T_n(X)\}$ تولید می‌کند که $\hat{Y}_b,\ b=1,2,...,n$، مشخص‌کننده کلاس مربوط به هر فرد می‌باشد.

الگوریتم ساخت جنگل به‌صورت زیر انجام می‌شود:

۱- پس از تقسیم کل حجم نمونه به دو گروه آموزش (کمابیش ۷۰٪ مجموعه داده‌ها) و آزمون (کمابیش ۳۰٪ مجموعه داده‌ها).

۲- از مجموعه داده‌های آموزش به حجم N یک نمونه بوت استرپ به حجم n انتخاب می‌کنیم (به‌صورت تصادفی و با جایگذاری).

۳- تعداد m متغیر به‌تصادف از بین M متغیر توضیحی انتخاب می‌گردد ($m < M$). به‌طور معمول تعداد متغیرها با mtry نشان داده می‌شود که باید توسط کاربر انتخاب گردد. پیشنهاد شده است که در مدل رگرسیونی $m=\frac{M}{3}$ و در مدل رده‌بندی $m=\lceil\sqrt{M}\rceil$ انتخاب شود.

٤- برای هر نمونه بوت استرپ انتخاب‌شده یک درخت به‌گونه‌ای ساخته می‌شود که در هر گره، بهترین متغیر انتخاب شود.

درخت تا زمانی رشد می‌کند که به بیش‌ترین اندازه خود برسد و انشعاب بیشتر امکان‌پذیر نباشد. به‌عبارتی قابل هرس کردن نباشد.

گام‌های ۱ تا ٤ را تا زمانی که (ntree) n درخت رشد کند ادامه می‌دهیم. انتخاب رایج برای ntree، ۵۰۰ درخت می‌باشد.

از آنجاکه بوت استرپ، نمونه‌گیری با جایگذاری از داده‌های آموزش می‌باشد، برخی از نمونه‌ها در فرآیند نمونه‌گیری مورد استفاده قرار نمی‌گیرند، نمونه‌های کنار گذاشته‌شده، نمونه‌های خارج از کیسه (Out of bag, OOB)، نامیده می‌شوند، که از آن‌ها به‌منظور برآورد خطا و برآورد متغیر بااهمیت استفاده می‌گردد. در این روش، مقدار بهینه برای دو پارامتر، تعداد درخت و تعداد متغیرهایی که در ساخت درخت به‌کار می‌روند، مورد بررسی قرار می‌گیرد. به‌عبارت دیگر، کمینه کردن خطا در هر درخت معادل با مقدار بهینه برای ntree و mtry است. به‌طور معمول تعداد درخت‌های مورد نیاز برای نیل به ناریبی، ۵۰۰ درخت توصیه شده است. از طرفی به‌صورت پیش‌فرض تعداد متغیرهایی (mtry) که در هر گره وارد الگوریتم درخت تصمیم می‌شود، بزرگ‌ترین عدد صحیح نزدیک به ریشه دوم تعداد کل متغیرها است. طی ساخت درخت‌های RF برای تعیین اینکه گره بر اساس کدام متغیر

افراز شود، از شاخص اهمیت جینی استفاده می‌شود. این مقیاس احتمال هم‌کلاس (هم‌طبقه) بودن دو آیتمی را بیان می‌کند که به‌طور تصادفی از جمعیت مشابه انتخاب می‌شوند.[۲۲] متغیرهای مورد استفاده عبارت بودند از جنسیت، سن، سابقه MS در خانواده، سابقه تروما، بیماری التهابی روده، روماتیسم، بیماری عفونی، استرس، نور اپتیک، دوبینی، لرمیت، مونوپارزی، پاراپارزی، همی پارزی، اسپاسم، سندروم پا، رفلکس هوفمن، پارستزی، بی‌حسی، آتاکسی، لرزش، سرگیجه، افسردگی، اضطراب، اختلال حافظه، خستگی، تکرر ادرار، یبوست و بیماری MS (تمامی متغیرها غیر از سن و جنسیت به‌صورت دوحالته (دارد، ندارد) در نظر گرفته شدند). تجزیه و تحلیل مربوط به روش RF با استفاده از R package random Forest (Liaw and Wiener, 2002) و تجزیه و تحلیل مربوط به روش SVM با استفاده از بسته نرم‌افزاری e1071' package (https://cran.r-project.org/web/packages/e1071/index.html) and caret R packages (https://cran.r- project.org/web/packages/caret/index.html) و ga مربوط به GA در محیط نرم‌افزار R 3.3.3 (R Foundation for Statistical Computing, Vienna, Austria) انجام گردید.

یافته‌ها

در شروع فرآیند تحلیل، ابتدا فرایند انتخاب ۷۰٪ داده‌ها به‌عنوان مجموعه آموزش و ۳۰٪ به‌عنوان مجموعه آزمون انجام شد. به‌منظور انتخاب بهینه پارامترهای مربوط به RF (mtry و ntree)، الگوریتم با تعداد متغیرها (ویژگی) و تعداد درختان مختلف اجرا و ترکیبی انتخاب شد که نسبت به سایر حالت‌ها دارای OOB کمتری می‌باشد.

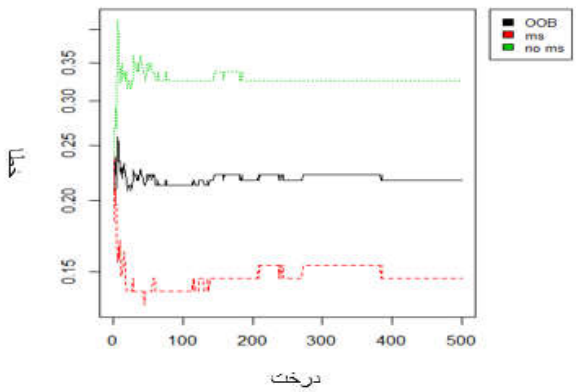
بر اساس نتایج جدول ۱ ستون OOB، mtry=۵ برای ntree=۱۰۰ و ntree=۵۰۰ کاندید پارامترهای بهینه هستند زیرا این دو ترکیب دارای کمترین مقدار خطا می‌باشند، اما ازآنجاکه Breiman در مقاله خود در سال ۲۰۰۱ با استفاده از قانون اعداد بزرگ ثابت کرده است که در تعداد دفعات زیاد اجرای درخت، یعنی جنگل تصادفی با تعداد درخت‌های زیاد، برآورد خطا به‌طور چشمگیری کاهش پیدا خواهد کرد بنابراین ntree=۵۰۰ انتخاب شد.

نمودار ۱ نیز بیانگر این مطلب می‌باشد که افزایش تعداد درخت‌ها، خطاهای OOB را کاهش داده است. با مشاهده نمودار می‌بینیم که اجرای ۵۰۰ درخت در جنگل تصادفی برای داده‌ها کفایت می‌کند زیرا همان‌طور که مشاهده می‌شود از حدود ۴۰۰ درخت به بعد خطاهای OOB مقداری ثابت را نشان می‌دهد. پس از انتخاب بهینه پارامترها، ادامه تحلیل‌های مدل RF با استفاده از ntree=۵۰۰ و mtry=۵ انجام شد.

یکی از ویژگی‌های جنگل تصادفی، انتخاب متغیرهای مهم بر اساس شاخص اهمیت جینی می‌باشد. همان‌طور که در نمودار ۲ دیده می‌شود، با به‌کارگیری شاخص کاهش میانگین جینی، در روش RF، سن با بیشترین کاهش میانگین در ناخالصی گره (و نه کاهش میانگین در صحت) به‌عنوان مهم‌ترین متغیر (۲۲/۳۰)، پس از آن متغیرهای استرس (۲۱/۹۰)، سرگیجه (۱۷/۱۰)، خستگی (۱۶/۲۰) و جنسیت (۸/۶۰) به‌عنوان دیگر متغیرهای اثرگذار مرتب شده‌اند.

جدول ۱: نتایج مربوط به انتخاب بهینه پارامترهای مدل جنگل تصادفی بر اساس مقدار خطای خارج از کیسه

تعداد درخت	خطای خارج از کیسه	
	تعداد متغیرها= ۵	تعداد متغیرها= ۶
۱۰۰	٪۲۲/۲	٪۲۷
۲۰۰	٪۲۲/۶	٪۲۷
۳۰۰	٪۲۲/۵	٪۲۸
۴۰۰	٪۲۲/۵	٪۲۶
۵۰۰	٪۲۲/۲	٪۲۶



نمودار ۱: خطای OOB بر اساس تعداد درختان جنگل

نمودار ۲: تعیین اهمیت متغیرها بر اساس میانگین کاهش جینی

در شروع فرآیند تحلیل SVM، ابتدا فرآیند انتخاب ۷۰٪ داده‌ها به‌عنوان مجموعه آموزش و ۳۰٪ به‌عنوان مجموعه آزمون انجام شد. دو پارامتر c و مربوط به مدل SVM با تابع کرنل RBF با استفاده از الگوریتم ژنتیک بهینه‌سازی شدند پس از برقراری شرط توقف الگوریتم ژنتیک، پارامترهای به‌دست‌آمده را که عبارتند از c=۵۱۶/۳۷۲ و ۱۳۳/۹۱= به‌عنوان ورودی به SVM اعمال نمودیم و سپس معیارهای برازش برای این مدل به‌دست آمدند.

به‌منظور ارزیابی و مقایسه عملکرد دو مدل معرفی‌شده برای طبقه‌بندی افراد سالم از افراد بیمار، مدل‌ها بر روی داده‌های آزمون (n=۹۹) اجرا شدند و نتایج در ماتریس اغتشاش (جدول ۲) گنجانده شد. با استفاده از نتایج این ماتریس، معیارهای ارزیابی عملکرد مدل‌ها شامل صحت، حساسیت و ویژگی محاسبه گردید (جدول ۳).

بحث

در عرصه‌های پزشکی، همواره ابزاری برای پیش‌بینی وضعیت افراد در ابتلا به بیماری‌ها بر اساس عوامل خطر هر بیماری لازم است تا بتوان امکان پیشگیری به‌موقع و شناخت هرچه بهتر عوامل خطر را فراهم کرد. از سویی در بیشتر بیماری‌ها، شاهد تعداد زیادی از عوامل خطر هستیم که افزون‌بر داشتن اثرات متقابل با یکدیگر، دارای اثرات غیرخطی نیز هستند.[۲۳] به‌منظور شناسایی، تحلیل و بررسی این‌گونه اثرات و یافتن روابط غیرخطی میان آن‌ها، باید از ابزارهایی ویژه بهره برد. ابزاری که بتواند از انبوه داده‌های موجود الگوهای قابل‌توجه را

جدول ۲: ماتریس اغتشاش توأم برای RF و SVM بر روی داده آزمون (n=۹۹)

کلاس واقعی افراد	کلاس پیش‌بینی‌شده در مدل RF		کلاس پیش‌بینی‌شده در مدل SVM	
	مثبت	منفی	مثبت	منفی
مثبت	٤٥	۱۰	٤٥	۱۱
منفی	۱۳	۳۱	۸	۲۹

جدول ۳: نتایج حاصل از ارزیابی دو مدل طبقه‌بندی‌کننده RF و SVM

مدل	صحت	حساسیت	ویژگی	ارزش اخباری مثبت	ارزش اخباری منفی
RF	۰/۷٦	۰/۸۱	۰/۷۰	۰/۷۷	۰/۷٥
SVM	۰/۷۹	۰/۸۰	۰/۷۸	۰/۸٤	۰/۷۲

کشف و از آن‌ها دانش استخراج کند. در بسیاری از علوم، ازجمله علم پزشکی به‌دلیل ماهیت داده‌ها که مجموعه داده‌های با بعد بالا را شامل می‌شوند، اغلب امکان استفاده از روش‌های کلاسیک آماری در انجام تحلیل‌ها وجود ندارد. داده‌کاوی که همان علم اکتشاف دانش از داده است، به‌خوبی می‌تواند پاسخگوی این نیاز باشد. به‌وسیله داده‌کاوی، داده‌های موجود مورد تحلیل قرار می‌گیرند تا روندهای احتمالی، ارتباطات غیر محسوس و الگوهای مخفی از بین انبوه داده‌ها شناسایی شوند. داده‌کاوی از الگوریتم‌های مختلفی مانند روش جنگل تصادفی، درخت‌های تصمیم، ماشین‌های بردار پشتیبان، شبکه‌های عصبی مصنوعی و غیره بهره می‌گیرد.[۲۴]

با در نظر گرفتن مزایای الگوریتم‌های داده‌کاوی، ازجمله امکان وارد کردن تعداد بسیار زیادی متغیر به مدل، عدم نیاز به فرضیاتی مانند نرمال بودن و توانایی یافتن پارامترها حتی باوجود داده‌های گم‌شده، می‌توان این روش‌ها را ابزاری مناسب برای تحلیل داده‌هایی با حجم بالا در نظر گرفت.[۲۵] در این پژوهش به بررسی و مقایسه معیارهای ارزیابی دو مدل جنگل تصادفی و ماشین بردار پشتیبان به‌منظور دستیابی به مدلی با پیچیدگی کم‌تر و عملکرد تشخیصی مناسب جهت تشخیص بیماران مبتلابه MS، پرداخته شد. بر اساس نتایج حاصل از این مطالعه از بین دو مدل مورد بررسی، مدل SVM با در نظر گرفتن معیارهای صحت (۰/۷۹)، حساسیت (۰/۸۱) و ویژگی (۰/۷۸) نسبت به RF که دارای صحت (۰/۷۶)، حساسیت (۰/۸۰) و ویژگی (۰/۷۰) بود در پیش‌بینی بیماری MS (متغیر وابسته) توان تشخیصی بهتری داشته است که به برتری این مدل در برخی از مطالعات نیز اشاره شده است. ازجمله در مطالعه‌ای که توسط Maroco و همکاران به‌منظور پیش‌بینی بیماری زوال عقل صورت گرفت، عملکرد روش‌های داده‌کاوی تحلیل ممیزی خطی، رگرسیون لجستیک، شبکه عصبی، درخت طبقه‌بندی، SVM و RF مورد مقایسه قرار گرفتند. در بین این روش‌ها مدل SVM دارای بهترین عملکرد با صحت ۰/۷۶ بود.[۲۶] همچنین در مطالعه Nalband و همکارانش که به‌منظور تشخیص اختلالات مفصل زانو به‌کمک روش‌های داده‌کاوی RF و SVM انجام شد، نتایج بیانگر برتری روش SVM بود. همچنین SVM دارای صحتی برابر با ۰/۹۸ و RF دارای صحت ۰/۹۳ بود.[۲۷] شاید بتوان مناسب بودن صحت حاصل از این مطالعه را به خوش‌تعریف بودن داده‌ها نسبت داد. با وجود اینکه مدل SVM در بیشتر مطالعات و ازجمله در مطالعه کنونی دارای توان تشخیصی بیشتری نسبت به RF بوده است، اما مواردی وجود دارد که در آن RF نه‌تنها دقت کمی نداشته بلکه به مدل SVM نیز ترجیح داده شده است، ازجمله در مطالعه‌ای که توسط Shirwaikar به‌منظور پیش‌بینی آپنه در نوزادان انجام شد، عملکرد سه روش SVM، RF و درخت تصمیم مورد مقایسه قرار گرفت. نتایج بیانگر آن است که مدل جنگل تصادفی با صحت ۰/۸۸ نسبت به دو مدل دیگر کارایی بهتری داشته است.[۲۸]

بر اساس نتایج حاصل از مطالعه کنونی می‌توان چنین نتیجه گرفت که SVM نسبت به RF عملکرد بهتری در تشخیص بیماری MS از خود به نمایش گذاشته است. همچنین برآیند بالینی مطالعه کنونی این است که به پزشکان متخصص این حوزه پیشنهاد می‌شود از مدل SVM به‌عنوان یک روش کمک غربالگری MRI در شرایط حساس جهت تشخیص افراد مبتلا به MS استفاده شود. از طرفی با توجه به مطالعات انجام‌شده در حوزه بالینی، مشخص شد شناسه‌های انتخاب‌شده با استفاده از روش RF شامل سن، استرس، سرگیجه، خستگی و جنسیت، جزو شناسه‌های تأثیرگذار در تشخیص بیماری MS می‌باشند. در پایان پیشنهاد می‌شود مدل SVM را بر روی حجم بیشتری از نمونه‌ها به‌منظور کسب صحت بیشتر اعمال نمود.

سپاسگزاری: این مقاله حاصل پایان‌نامه تحت عنوان "بررسی مقایسه‌ای میزان کارایی مدل‌های مبتنی بر ۲ روش داده‌کاوی جنگل‌های تصادفی و ماشین‌بردار پشتیبان در پیش‌بینی بیماری مولتیپل اسکلروزیس" در مقطع کارشناسی ارشد در سال ۱۳۹۷ و کد ۹٦٤۱٦ می‌باشد که با حمایت دانشگاه علوم پزشکی و خدمات بهداشتی درمانی کرمانشاه اجرا شده است.

References

1. Ion-Mărgineanu A, Kocevar G, Stamile C, Sima DM, Durand-Dubief F, Van Huffel S, et al. Machine learning approach for classifying multiple sclerosis courses by combining clinical data with lesion loads and magnetic resonance metabolic features. *Front Neurosci* 2017;11:398.
2. Browne P, Chandraratna D, Angood C, Tremlett H, Baker C, Taylor BV, et al. Atlas of Multiple Sclerosis 2013: A growing global problem with widespread inequity. *Neurology* 2014:83(11):1022-4.
3. Etemadifar M. Izadi S. Nikseresht A. Sharifian M. Sahraian MA. Nasr Z. Estimated prevalence and incidence of multiple sclerosis in Iran. *Eur Neurol* 2014;72(5-6):370-4.

4. Saman-Nezhad B, Rezaee T, Bostani A, Najafi F, Aghaei A. Epidemiological characteristics of patients with multiple sclerosis in Kermanshah, Iran in 2012. *J Mazandaran Univ Med Sci* 2013;23(104):97-101.
5. Beer A, Biberacher V, Schmidt P, Righart R, Buck D, Berthele A, et al. Tissue damage within normal appearing white matter in early multiple sclerosis: assessment by the ratio of T1- and T2-weighted MR image intensity. *J Neurol* 2016;263(8):1495-502.
6. Dempsey MF, Condon B, Hadley DM. MRI safety review. *Semin Ultrasound CT MR* 2002;23:392-401.
7. von Smekal A, Seelos KC, Küper CR, Reiser M. Patient monitoring and safety during MRI examinations. *Eur Radiol* 1995;5(3):302-5.
8. Hooman HA. Structural Equation Modeling Using LISREL Software. Tehran: SAMT Publication; 2005.
9. Stylianou N, Akbarov A, Kontopantelis E, Buchan I, Dunn KW. Mortality risk prediction in burn injury: Comparison of logistic regression with machine learning approaches. *Burns* 2015;41(5):925-34.
10. Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G. Support vector machines and kernels for computational biology. *PLoS Comput Biol* 2008;4(10):e1000173.
11. Chen M-S, Han J, Yu PS. Data mining: an overview from a database perspective. Knowledge and data engineering. *IEEE Trans Knowl Data Eng* 1996;8(6):866-83.
12. Zheng B, Yoon SW, Lam SS. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Syst Appl* 2014;41(4):1476-82.
13. Tomar D, Agarwal S. A survey on data mining approaches for healthcare. *Int J Bio-Sci Bio-Technol* 2013;5(5):241-66.
14. Savu-Krohn C, Rantitsch G, Auer P, Melcher F, Graupner T. Geochemical fingerprinting of coltan ores by machine learning on uneven datasets. *Nat Resour Res* 2011;20(3):177-91.
15. Tamura H, Tanno K. Midpoint validation method for support vector machines with margin adjustment technique. *Int J Innov Comput Inf Control* 2009;5:4025-32.
16. Lin CF, Wang SD. Fuzzy support vector machines. *IEEE Trans Neural Netw* 2002;13(2):464-71.
17. Javed I, Ayyaz MN, Mehmood W. Efficient Training Data Reduction for SVM Based Handwritten Digits Recognition. International Conference on Electrical Engineering. Lahore: Pakistan; 2007.
18. Kapp MN, Sabourin R, Maupin P. A dynamic model selection strategy for support vector machine classifiers. *Appl Soft Comput* 2012;12(8):2550-65.
19. Hsu CW, Chang CC, Lin CJ. A Practical Guide to Support Vector Classification. Technical Report, Department of Computer Science and Information Engineering, University of National Taiwan: Taipei; 2003.
20. Michalewicz Z. Genetic Algorithms + Data Structures = Evolution Programs. 2nd ed. New York, NY: AI Series, Springer-Verlag; 1994.
21. Huang J, Hu X, Yang F. Support vector machine with genetic algorithm for machinery fault diagnosis of high voltage circuit breaker. *Measurement* 2011;44(6):1018-27.
22. Breiman L. Random forests. *Mach Learn* 2001;45(1):5-32.
23. Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, Tourassi GD. Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance. *Neural Netw* 2008;21(2-3):427-36.
24. Guleria P, Sood M. Data mining in education: a review on the knowledge discovery perspective. *Int J Data Min Knowl Manag Process* 2014;4(5):47.
25. Basheer IA, Hajmeer M. Artificial neural networks: fundamentals, computing, design, and application. *J Microbiol Methods* 2000;43(1):3-31.
26. Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I, de Mendonça A. Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Res Notes* 2011;4:299.
27. Nalband S, Sundar A, Prince AA, Agarwal A. Feature selection and classification methodology for the detection of knee-joint disorders. *Comput Methods Programs Biomed* 2016;127:94-104.
28. Shirwaikar RD, Acharya DU, Makkithaya K, Surulivelrajan M, Lewis LES. Machine learning techniques for neonatal apnea prediction. *J Artif Intell* 2016;9(1-3);33-8.

TUMJ Tehran University Medical Journal, April 2019; Vol. 77, No. 1: 33-40 Original Article

Prognosis of multiple sclerosis disease using data mining approaches random forest and support vector machine based on genetic algorithm

Amir Hossein Hashemian Ph.D.[1,2]
Sara Manochehri M.Sc.[3*]
Daryoush Afshari M.D.[4]
Zohreh Manochehri Ph.D.[5]
Nader Salari Ph.D.[6]
Soodeh Shahsavari Ph.D.[7]

1- Research Center for Environmental Determinants of Health (RCEDH), Kermanshah University of Medical Sciences, Kermanshah, Iran.
2- Department of Biostatistics, School of Health, Kermanshah University of Medical Sciences, Kermanshah, Iran.
3- Department of Biostatistics, Students Research Committee, School of Health, Kermanshah University of Medical Sciences, Kermanshah, Iran.
4- Department of Neurology, Faculty of Medicine, Kermanshah University of Medical Sciences, Kermanshah, Iran.
5- Department of Biostatistics, Student Research Committee, Hamadan University of Medical Sciences, Hamadan, Iran.
6- Department of Biostatistics, School of Health, Kermanshah University of Medical Sciences, Kermanshah, Iran.
7- Department of Health Information Management, School of Paramedical, Kermanshah University of Medical Sciences, Kermanshah, Iran.

[*] Corresponding author: Student Research Committee, School of Health, Kermanshah University of Medical Sciences, Shahid Bahonar St., Kermanshah, Iran.
Tel: +98 83 38398069
E-mail: saramanochehri0098@gmail.com

Abstract

Received: 21 Oct. 2018 Revised: 28 Oct. 2018 Accepted: 09 Apr. 2019 Available online: 19 Apr. 2019

Background: Multiple sclerosis (MS) is a degenerative inflammatory disease which is most commonly diagnosed by magnetic resonance imaging (MRI). But, since the MRI device uses of a magnetic field, if there are metal objects in the patient's body, it can disrupt the health of the patient, the functioning of the MRI, and distortion in the images. Due to limitations of using MRI device, screening seems necessary for those patients who have metal objects in their bodies. Therefore, this study is carried out to compare two models: support vector machine and random forest.

Methods: This analytical-modelling research was implemented on MS data collection, the specifications of which are recorded in health registry system in School of Public Health, Kermanshah University of Medical Sciences, Iran, from May 2017 to August 2018. For the purpose of this study, a total of 317 subjects were selected as a sample; 188 subjects were diagnosed with MS and 128 subjects showed no symptoms of MS. In order to fit the support vector machine (SVM) model, radial basis kernel function was used. The parameters of this machine were optimized with genetic algorithm. After this step, the support vector machine and random forest (RF) were compared with respect to three factors: accuracy, sensitivity, and specificity.

Results: Based upon the obtained results of study, accuracy, sensitivity, and specificity of SVM were 0.79, 0.80, and 0.78, respectively. In comparison, accuracy, sensitivity, and specificity of RF were found to be 0.76, 0.81, and 0.70, respectively.

Conclusion: In general, both models which were compared in current study showed desirable performance; however, in term of accuracy, as an important criteria for performance comparison in this area of research, it can be argued that support vector machine can do better than random forest in diagnosing multiple sclerosis.

Keywords: genetic algorithm, multiple sclerosis, random forest, support vector machine.