DOR: 20.1001.1.2538161.1403.9.2.1.2

DOI: 10.22034/csj.2024.203191

تابستان ۱۴۰۳، دوره ۹، شماره ۲، پیاپی ۳۳

دریافت مقاله: ۱۴۰۳/۰۱/۲۹

پذیرش مقاله: ۱۴۰۳/۰۳/۰۱

نوع مقاله: پژوهشی

تشخیص بیماری‌های مزمن به کمک ماشین بردار پشتیبان دوقلو با قیود نرم

حمیده فدیشه‌ای

دانشجوی کارشناسی ارشد، دانشکده علوم ریاضی، دانشگاه فردوسی مشهد، مشهد، ایران

پست الکترونیکی: ha.fadishehei@mail.um.ac.ir

جلال‌الدین نصیری*

استادیار، گروه ریاضی کاربردی، دانشکده علوم ریاضی، دانشگاه فردوسی مشهد، مشهد، ایران

پست الکترونیکی: Jnasiri@um.ac.ir

سهراب عفتی

استاد، گروه ریاضی کاربردی، دانشکده علوم ریاضی، دانشگاه فردوسی مشهد، مشهد، ایران

پست الکترونیکی: s-effati@um.ac.ir

چکیده

بیماری‌های مزمن از دلایل اصلی مرگ‌ومیر در سال‌های اخیر بوده‌اند. پژوهش‌های تجربی بسیاری برای تشخیص بیماری مزمن در کشورهای مختلف گزارش شده‌اند. در مطالعات انجام شده، روش‌های یادگیری ماشین، نقش به‌سزایی در تشخیص بیماری‌های مزمن داشته‌اند. از جملهٔ این روش‌ها، می‌توان به ماشین بردار پشتیبان استاندارد و دوقلو اشاره کرد. داده‌های بیماری‌های مزمن دارای نوفه ذاتی و نمونه‌های پرت می‌باشند. این مهم باعث افت شدید دقّت مدل‌های ماشین بردار پشتیبان استاندارد و دوقلو می‌گردد. در مدل جدید، برای برطرف کردن کاستی بیان شده و رفع حسّاسیت نسبت به داده‌های نوفه‌ای[1]، به جای دسته‌بندی ماشین بردار پشتیبان دوقلو، ماشین بردار پشتیبان دوقلو با قیود نرم پیشنهاد شده است. این مدل فضای شدنی مسئله بهینه‌سازی درجه دوم را گسترش می‌دهد. با گسترش فضای شدنیِ مسئله، به نمونه‌ها اجازهٔ تخطّی از ابَرصفحات داده می‌شود و تأثیر داده‌های نوفه‌ای و دور افتاده در تشخیص نهایی کاهش می‌یابد. الگوریتم پیشنهادی، بر روی پنج مجموعه داده از داده‌های بالینی پزشکی، اجرا و نتایج آن با روش‌های ماشین بردار پشتیبان، درخت تصمیم، رگرسیون لجستیک و نزدیکترین همسایه مقایسه شده‌اند. میانگین دقّت روش پیشنهادی نسبت به بهترین میانگین در روش‌های پیشین، حاکی از بهبود حدود ۷٪ در نتایج تجربی می‌باشد.

واژه‌های کلیدی: یادگیری ماشین، ماشین بردار پشتیبان دوقلو، بهینه‌سازی محدّب، محدودیت‌های فازی، تحمّل، اطمینان.

مقدمه

در این بخش ابتدا ضرورت انجام پژوهش و سپس

* نویسندهٔ مسئول

1-Noise

بعضی از مفاهیم توضیح داده خواهند شد.

ضرورت انجام پژوهش

بیماری‌هـــای مزمـــن به دلیـــل طولانی بـــودن دوران ابتـــلا، عـــوارض آزاردهندهٔ مختلفی دارنـــد. این بیماری‌ها زمینه‌ســـازی برای مبتلا شدن به بیماری‌های عفونی دیگر از جمله کرونا می‌شـــوند. از ایـــن‌رو، بیماری‌های مزمن از مهم‌ترین چالش‌های نظام ســـلامت در جهان هســـتند. در ســـال ۲۰۱۹، حدود ۶۳ ٪ از مرگ و میر جهانی ناشـــی از بیماری‌های مزمن بوده اســـت. رایج‌تریـــن این بیماری‌ها، فشـــار خون، قند خـــون، و بیماری‌های قلبی هســـتند [۱]. عوارض ناشـــی از آن‌ها باعث آسیب رسیدن به قلب، کلیه، مغز و چشم‌ها می‌شود که آثاری مخرّب روی کار و زندگی افـــراد خواهد داشـــت [۲]. افزون بر موارد ذکر شـــده این بیماران در برابر بیماری‌های عفونی (مثلاً ویروس کرونا) ایمنـــی کمتری دارند [۳]. بر پایهٔ گزارش‌های ســـال ۲۰۱۹، ٤٨٪ افراد مبتلا به این ویروس بیماری مزمن نیز داشته‌اند و احتمال بروز علایم شدید در آن‌ها بیشتر بوده‌است [٤]. بنابراین تشخیص زودهنگام بیماری کمک قابل توجّهی به سلامت فرد و جامعه خواهد داشت.

بنابر تعریف ســـازمان بهداشـــت جهانی، بیماری‌های مزمن به آن دســـته از بیماری‌ها گفته می‌شود که منجر به آسیب در ساختار یا کاهش عملکرد بدن شده به گونه‌ای که ســـبب تغییر در زندگی عادی بیمار گردند و طی یک دوره زمانی طولانی ادامه یافته و پایدار باشند [٥].

بیماری‌های مزمن اصلی عبارتند از: بیماری‌های قلبی-عروقی، انواع ســـرطان، انســـداد ریه، نارسایی مزمن کلیه و دیابت. البته بیماری‌هـــای مزمن دیگری از جمله بیماری روانی، اختلالات بینایی، شـــنوایی، ژنتیـــک و بیماری‌های مفصلی نیز وجود دارنـــد. از آن‌جا که عوامل محیطی مثل شرایط اجتماعی، سبک زندگی و آلاینده‌های زیست‌محیطی نقش فزاینده‌ای در پیدایش و پیشرفت این بیماری‌ها دارند از آن‌ها با عنوان بیماری‌های مزمن محیطی نیز یاد می‌شود.

بر اســـاس گزارش سازمان بهداشـــت جهانی ویژگی مشترک بیماری‌های مزمن عبارتند از:

- چند دهه زمان لازم اســـت که همه‌گیری آن‌ها به‌طور کامل تثبیت شود.
- ریشهٔ این بیماری‌ها در سنین جوانی است.
- بـــا توجه به زمـــان طولانی پیـــش از بـــروز آن‌ها، فرصت‌های زیادی برای پیش‌گیری از آن‌ها وجود دارد.
- درمـــان این بیماری‌ها بـــه یک رویکـــرد نظام‌مند و بلندمدّت نیاز دارد.

اگرچه برخی از بیماری‌های مزمن زمینه‌های ارثی دارند یا به دلیل بالا رفتن سن بروز پیدا می‌کنند، پژوهش‌ها نشان داده است که بخش بزرگی از دلایل بروز این بیماری‌ها را می‌توان به تغییر شرایط محیطی، سبک زندگی و رفتار فرد مثل اســـترس، فعالیت بدنی، سوء تغذیه، مصرف دخانیات و الکل نســـبت داد [٦]. تغییرات ذکر شـــده منجر به چهار تغییر متابولیک/ فیزیولوژیک- فشار خون بالا، اضافه وزن و چاقی، گلوکز خون بالا، کلســـترول خون بالا و در نهایت منجر به بیماری‌های مزمن می‌شوند [۷].

با توجه به تأثیر قابل توجّهی که بیماری‌های مزمن بر نظام سلامت دارند، طی سال‌های اخیر مطالعات بسیاری به بررسی تأثیر اقتصادی بیماری‌های مزمن پرداخته‌اند. تأثیر معنادار این بیماری‌ها بر تولید ناخالص داخلی و نرخ رشد درآمد سرانه نشـــان می‌دهد که این اثرات ریشه در تولید اقتصادی در دو سطح فردی و خانواده دارد.

بیماری‌های مزمن به صورت مســـتقیم و غیرمستقیم بـــر اقتصـــاد خانواده‌ها تأثیـــر می‌گذارند. به‌عنـــوان مثال بیماری‌هـــای مزمن ۸٦٪ از هزینه‌های ســـلامت خانوار در ســـال ۲۰۱٦ ایالات متحده را در برگرفته اســـت. امّا بدیهی اســـت که بار[۲] اقتصادی بیماری‌های مزمن تنها هزینه‌های مربوط به خدمات درمانی نیست.

مهمترین هزینه دیگر، درآمد از دســـت رفته است که از کاهش بهره‌وری ناشی از بیماری نشئت می‌گیرد. همچنین گزارش سال ۲۰۰۵ سازمان بهداشت جهانی نشان می‌دهد

2-Burden

که بیماری‌های مزمن حدود ٤٠٪ از سال‌های مولّد را در بازار کار کشورهای نوظهور از بین برده است.

از آن‌جا که بیماری‌های مزمن هزینه‌های بالایی را به فرد تحمیل می‌کنند، ممکن است خانواده‌ها مجبور به افزایش عرضه نیروی کار، کاهش مصرف دیگر کالاها یا دریافت درآمدهای انتقالی از جمله هدایا و کمک‌هایی از دیگر افراد یا نهادهای کشور شوند. شوک هزینه‌ای ناشی از بیماری مزمن باعث کاهش درآمد ناشی از کار و افزایش هزینه‌های پزشکی، افزایش درآمدهای انتقالی و کاهش مصرف خوراکی و غیرخوراکی شده است. بنابراین این بیماری‌ها همچنان چالش مهمی در حوزه بهداشت عمومی در تمام کشورها، از جمله کشورهای با درآمد کم و متوسط، هستند چرا که در این کشورها، بیش از سه چهارم مرگ‌ومیر به دلیل ابتلا به بیماری‌های مزمن رخ می‌دهد.

با توجّه به موارد گفته‌شده، وجود یک سیستم هوشمند که بتواند به پیش‌بینی زودهنگام بیماری‌های مزمن دست یابد مورد نیاز است. تشخیص بهنگام بیماری می‌تواند آثار روحی، روانی، اجتماعی و اقتصادی بسیاری بر فرد و جامعه داشته باشد.

تاریخچه‌ای از تشخیص بیماری‌های مزمن

برای تشخیص بیماری‌های مزمن خاص چندین الگوریتم یادگیری ماشین پیشنهاد شده است [۹].

مَنیرالزمان و همکاران [۱۰] باور داشتند که گسترش مدل‌های پیش‌بینی برای شناسایی دلایل ابتلا به دیابت نوع ۲، می‌تواند به تشخیص بیماری و مداخلهٔ زودهنگام و کاهش هزینه‌های پزشکی کمک کند. در پژوهش مذکور ۱۳۸۱٤٦ فرد شرکت داشتند که ۲۰٤٦۷ نفر از آن‌ها به دیابت نوع ۲ مبتلا بودند. برای پیش‌بینی ابتلا به دیابت نوع۲، روش‌های مختلف یادگیری ماشین شامل ماشین بردار پشتیبان، درخت تصمیم، رگرسیون لجستیک، جنگل تصادفی، دسته‌بند گوسی بیز ساده و از مدل‌های رگرسیون لجستیک وزن‌دار تک متغیره و چند متغیره برای بررسی ارتباط عوامل خطر بالقوه با دیابت نوع ۲ به کار گرفته شدند. شبکه‌های عصبی با دقّت ۸۲/٤ ٪ بهترین نتیجه را داشتند.

هانگ و همکاران [۱۱] با استفاده از ماشین بردار پشتیبان، جنگل تصادفی و تقویت تطبیقی[3]، ابتلای به بیماری مزمن کلیوی را با ۲۱٤۲ فرد شرکت‌کننده در پژوهش، پیش‌بینی کردند. ماشین بردار پشتیبان با ۸۵/۷ ٪ بیشترین دقت را به دست آورد.

از آن‌جا که عکاسی از فوندوس شبکیه یک رویکرد غیرتهاجمی برای پیش‌بینی بیماری‌های مزمن می‌باشد، ژانگ و همکاران [۱۲] با روش ذکرشده، پژوهشی مقطعی در مناطق روستایی در مرکز چین انجام دادند. ۱۲۲۲تصویر برداری شبکیه با بیش از ۵۰ اندازه‌گیری بر روی ٦۲۵ نفر انجام شد. مدل شبکهٔ عصبی پیاده‌سازی شد. پژوهش یادشده، به یک سطح زیر منحنی ROC با AUC = 0/766 برای پیش‌بینی فشارخون و دقت ٦۸/۸ ٪ دست یافت.

العظّم و همکاران [۱۳] به مقایسهٔ عملکرد روش‌های یادگیری نظارت‌شده و نیمه نظارت شده در پیش‌بینی سرطان سینه پرداختند. نُه الگوریتم، رگرسیون لجستیک، دسته‌بند گوسی نایو بیز، ماشین بردار پشتیبان ساده و با هستهٔ گوسی، درخت تصمیم، جنگل تصادفی، گرادیان افزایشی، دورترین گرادیان افزایشی و نزدیک‌ترین همسایه مورد بررسی قرار گرفتند. برای آموزش و آزمایش مدل‌ها، مجموعهٔ داده‌های سرطان ویسکانین استفاده شدند. مدل‌های نزدیک‌ترین همسایگی و افزایش گردایان با دقت ۹۸ ٪ بهترین پیش‌بینی را داشتند.

شی و همکاران [۱٤] مدل‌های نزدیک‌ترین همسایه، ماشین بردار پشتیبان، جنگل تصادفی، شبکهٔ عصبی و دسته بند گوسی بیز ساده و افزایش انطباقی را با اطلاعات پزشکی مربوط به ۱۵۳ بیمار مشکوک به سرطان ریه، مورد آموزش و آزمایش قرار دادند. بیز ساده داده‌های آموزش و آزمایش به دقت ۱۰۰ ٪ دست یافت. نویسندگان عقیده داشتند که ترکیب روش‌های تشخیص با نشانگرهای

3-Adaptive Boosting

زیستی و روش‌های یادگیری ماشین قدرت پیش‌بینی ابتلا به سرطان ریه را بالا می‌برد.

بیمارانی که در مرحلهٔ نهایی بیماری کلیوی هستند، مستعد ابتلا به نوعی بیماری قلبی ویژه نیز می‌باشند. مزاتستا و همکاران [۱۵] دو مجموعه داده متفاوت را مورد بررسی قرار دادند. الگوریتم‌های رگرسیون لجستیک، نزدیک‌ترین همسایه، ماشین بردار پشتیبان خطی، ماشین بردار پشتیبان با هستهٔ گوسی و چندجمله‌ای، درخت تصمیم و دسته‌بند گوسی بیز ساده، مرگ و ابتلای به بیماری‌های قلبی عروقی در بیماران دیالیزی را پیش‌بینی کردند. بهترین نتایج مربوط به دسته‌بند ماشین بردار پشتیبان با کرنل گوسی[4] با دقت ۹۵/۲۵ ٪ و ۹۲/۱۵ ٪ بود.

شیائو هان و همکاران [۱۶] اوّلین چارچوب عمومی و کارآمد را برای تشخیص بیماری‌های مزمن ارائه کردند. روش آن‌ها علاوه بر استخراج ویژگی‌های سطح بالا، مشکل عدم تعادل جدّی موجود در داده‌های بیماری‌های مزمن و برازش بیش از حد را نیز حل کرد. ایشان مدل مبتنی بر شبکهٔ عصبی را معرفی کردند که با در نظر گرفتن وزن، نمونه‌ها را تقویت می‌کرد. داده‌هایی که در شبکهٔ عصبی چندجمله‌ای به اشتباه دسته‌بندی شده بودند به طور ویژه مورد توجه قرار گرفتند. مدل پیشنهادی با نُه مجموعهٔ داده، شامل داده‌های مقایسه‌ای و نمونه‌های واقعی در ابعاد بالا مورد آموزش و آزمایش قرار گرفت. روش جدید در مقایسه با ماشین بردار پشتیبان، رگرسیون لجستیک، درخت تصمیم، نزدیک‌ترین همسایگی و پرسپترون چند لایه، دارای دقّت قابل توجهی بود. مقابله با نوفه ذاتی و عدم تعادل داده‌ها در بیماری‌های مزمن در حال حاضر نیز مورد توجه پژوهشگران می‌باشد.

همان‌طور که ملاحظه گردید، روش ماشین بردار پشتیبان جزو روش‌های تشخیص زودهنگام بیماری‌های مزمن می‌باشد. از طرف دیگر SVM استاندارد، نسبت به نمونه‌های نوفه‌ای و پرت بسیار حسّاس است. تقویت دسته‌بند ماشین بردار پشتیبان به صورتی که برای داده‌های نوفه‌ای دقّت خوبی را کسب کند، می‌تواند در کارایی سیستم هوشمند تشخیص بیماری مزمن مفید باشد. در این پژوهش، برای دستیابی به هدف بیان شده، مدل مسئلهٔ بهینه‌سازی مربوط به ماشین بردار پشتیبان دوقلو، تغییر یافته است و نامساوی‌ها به صورت فازی در نظر گرفته شده‌اند. در نتیجهٔ این تغییر، فضای شدنی مسئله گسترش یافته است. با این روش میزانی از تحمّل خطا به الگوریتم، اعمال شده است. الگوریتم جدید، نسبت به داده‌های نوفه‌ای حساسیت کمتری دارد. برای تشخیص میزان کارایی الگوریتم نسبت به روش‌های پیشین، نرخ دقّت[5] روش نوین، محاسبه شده است. جهت ارزیابی، از روش اعتبار سنجی متقابل با $k = ۱۰$ استفاده شده و در نهایت برای مقایسه با روش‌های مشابه، آزمون آماری فریدمن[6] به کار رفته است.

ادامهٔ ساختار مقاله به این صورت است: بخش دوم مروری بر پیشینهٔ روش‌های ماشین بردار پشتیبان، خواهد بود. در بخش سوم روش پیشنهادی به تفصیل بیان خواهد شد. در بخش چهارم نتایج تجربی روش پیشنهادی، ارزیابی نتایج و در بخش پنجم نتیجه‌گیری و پیشنهادهایی برای پژوهش‌های بعدی ارائه خواهد شد. مراجع مورد استفاده در پژوهش به ترتیبی که در مقاله آمده‌اند در آخرین قسمت مشخص شده‌اند.

مروری بر روش‌های ماشین بردار پشتیبان

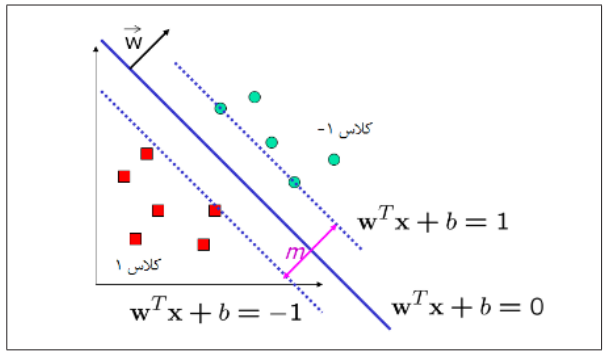
پس از الگوریتم‌های شبکهٔ عصبی، در سال ۱۹۹۵ ایدهٔ جدیدتری به عنوان ماشین بردار پشتیبان توسط وَپنیک و همکاران مطرح شد که در آن از روش‌های بهینه سازی مسایل درجه دوم، برای جداسازی داده‌های دو طبقه از یکدیگر استفاده می‌کرد. مبنای کار یافتن دو ابَرصفحهٔ موازی با بیشترین فاصلهٔ ممکن بود که داده‌های دو طبقه در دوطرف این ابرصفحات قرار گیرند [۱۷]. در طی سال‌ها انواع جدیدتر، دقیق‌تر و کاراتری از ماشین بردار پشتیبان ارائه شد. برخی از این روش‌ها در زیر بیان شده‌اند.

4-Radial Basis Function

5-Accuracy
6-Friedman test

ماشین بردار پشتیبان

فرض کنید که $X = \{(x_i, y_i)\}_{i=1}^{m}$ مجموعهٔ نمونه‌های آموزشی با اندازهٔ m باشد که هر $x_i \in R^n$ یک نمونهٔ یادگیری دارای n ویژگی در فضای ورودی[7] است، $y_i \in \{1, -1\}$ برچسب کلاس دادهٔ x_i و $A_{m_1 \times n}$ مجموعهٔ نمونه‌های با برچسب 1+، $B_{m_2 \times n}$ مجموعهٔ نمونه‌های با برچسب 1- و $m_1 + m_2 = m$ باشند، اگر داده‌ها در فضای ورودی، به‌صورت خطی قابل جداسازی باشند، می‌توان ابرصفحه‌ای جداساز به فرم $w^T x + b = 0$ یافت که در آن w یک بردار n بعدی و b کمیتی عددی است. ماشین بردار پشتیبان به جای یافتن یک ابرصفحهٔ جداکننده، به دنبال یافتن دو ابرصفحهٔ موازی است که تا حد امکان از یکدیگر دور بوده و داده‌های دو طبقه در دو طرف آنها واقع باشند. شکل (۱) تعبیر هندسی ماشین بردار پشتیبان با حاشیهٔ سخت را نمایش می‌دهد.



شکل (۱): ماشین بردار پشتیبان با حاشیه سخت

با مدل‌سازی مسئلهٔ بالا و حل مسئلهٔ بهینه‌سازی (۱)، می‌توان ابرصفحهٔ جداساز را مشخص کرد.

$$Min \frac{1}{2} \|w\|^2 \quad (1)$$
$$(w, b)$$
$$S.t.\ y_i(w^T x_i + b) \geq 1, i = 1,2, \ldots, m.$$

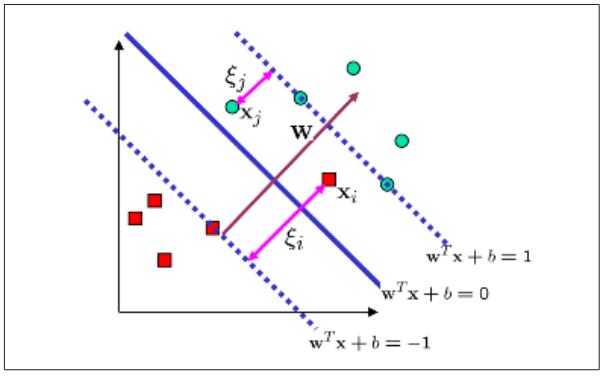
نمایش ماتریسی مسئلهٔ (۱) به صورت (۲) خواهد بود:

$$Min \frac{1}{2} \|w\|^2 \quad (2)$$
$$(w, b)$$
$$S.t.\ Aw + e_1 b \geq e_1,$$
$$Bw + e_2 b \leq -e_2$$

در (۲)، e_1 و e_2 بردارهای با درایه‌های واحد و متناسب با A و B هستند. در حالتی که داده‌ها در فضای ورودی، به‌صورت خطی جداپذیر نباشند، مسئلهٔ SVM با حاشیهٔ نرم به‌صورت (۳) مطرح خواهد شد:

$$Min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} \xi_i \quad (3)$$
$$(w, b)$$
$$S.t.\ y_i(w^T x_i + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0,\ i = 1,2, \ldots, m.$$

که در آن ξ_i متغیر لغزش برای کاهش تأثیر دادهٔ نوفه‌ای x_i است. شکل(۲) نمایانگر ماشین بردار پشتیبان با حاشیهٔ نرم می‌باشد.



شکل (۲): ماشین بردار پشتیبان با حاشیه نرم

پس از نوشتن تابع لاگرانژ و در نظرگرفتن شرایط مکمل زاید، دوگان مسئلهٔ (۳) به صورت زیر است که با حل آن ابرصفحهٔ جداساز معلوم خواهد شد.

$$Min \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_{k=1}^{m} \alpha_k \quad (4)$$
$$\sum_{i=1}^{m} \alpha_i y_i = 0, S.t.$$
$$0 \leq \alpha_i \leq C, i = 1,2, \ldots, m.$$

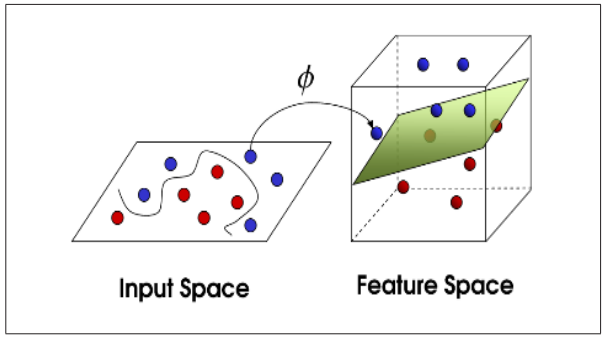
که α_i ضرایب لاگرانژ هستند و تابع تصمیم $D(x)$ طبق

7-Input Space

فرمـــول (۵) موقعیت دادهٔ جدید x را نســـبت به ابرصفحه مشخص خواهد کرد:

$$D(x) = Sign(w^T x + b) = Sign\left(\sum_{i \in S} \alpha_i y_i x_i^T x + b\right) \quad (۵)$$

در رابطـــهٔ (۵)، S نشـــان‌دهندهٔ مجموعـــهٔ اندیس‌های بردارهای پشتیبان اســـت. اگر نمونه‌ها در فضای ورودی جداپذیر خطی نباشند، نمونه‌ها با تکنیک کرنل مطابق شکل (۳) در فضای ویژگی با ابعاد بیشـــتر نگاشته می‌شوند. در ایـــن فضا، ماهیت خود داده، اهمیتـــی ندارد بلکه آنچه مهم اســـت فاصلهٔ داده‌ها از یکدیگر است. با این روش ماشین بردار پشتیبان یک ابرصفحهٔ جداکنندهٔ بهینه برای جداسازی نمونه‌ها پیدا می‌کند.



شکل (۳): نگاشت نمونه‌ها با روش کرنل

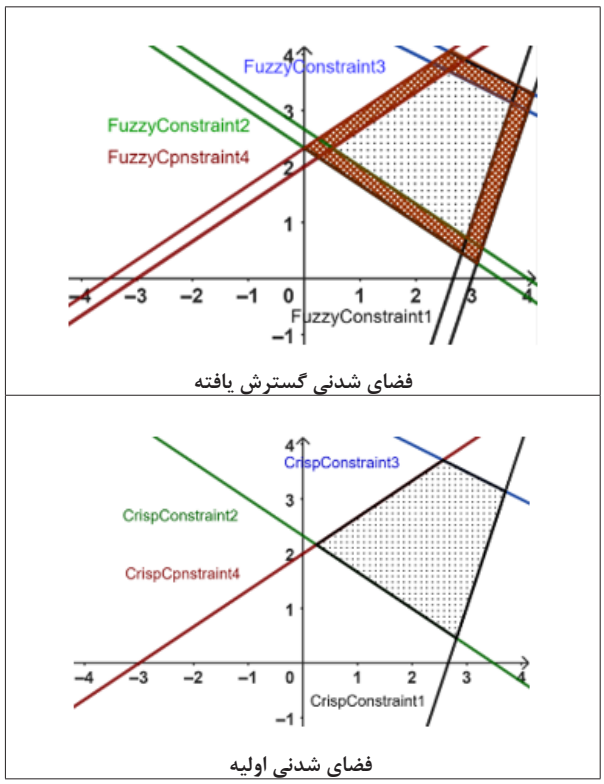
ماشین بردار پشتیبان با قیود نرم

در فرمول (۳) برای مسئله SVM، نظیر هر یک از نمونه‌ها یـــک محدودیـــت به صـــورت $y_i(w^T x_i + b) \geq 1 - \xi_i$ تعریف شد. بنابراین به تعداد نمونه‌ها محدودیت وجود دارد. مجموعهٔ این محدودیت‌ها، یک فضای شـــدنی[8] را به وجود می‌آورد که متغیّرهای مسئلهٔ یعنی $(w, b, \xi_i) \in R^{m+1+n}$ از این فضا انتخاب می‌شوند. قیدهای به‌دست آمده در SVM حالت بدون انعطاف دارند. در روش ماشین بردار پشتیبان با قیود نرم، قیدهای مســـئلهٔ انعطاف بیشـــتری دارند. در محدودیت‌ها به جای نامساوی‌های عادی از نامساوی‌های فازی استفاده شده است. این نوع از قیود جدید در فرمول (٦) بازنویسی شده است:

$$y_i(w^T x_i + b) \gtrsim 1 - \xi_i \,,\ i = 1, \ldots, m. \quad (٦)$$

شـــکل (٤) یک مثال ساده از مســـئلهٔ بهینه‌سازی (۷) با چهار قید منعطف و گسترش فضای شدنی مسئلهٔ را نشان می‌دهد.

$$Minimize\ 4x_1 + 3x_2 \quad (۷)$$
$$S.t.\ \begin{cases} g_1(x_1, x_2) = 3x_1 - x_2 \lesssim 8 \\ g_2(x_1, x_2) = 2x_1 + 3x_2 \gtrsim 7 \\ g_3(x_1, x_2) = x_1 + 2x_2 \lesssim 12 \\ g_4(x_1, x_2) = -x_1 + 2x_2 \lesssim 3 \end{cases}$$



شکل (۴): گسترش فضای شدنی برای مسئلهٔ (۷)

برای به‌دست آمدن جواب‌های مسئله، باید محدودیت‌های مسئلهٔ (۷) برآورده شوند.

علامت $\gtrsim$ به این معنی اســـت که به برخی از نمونه‌های آمـــوزش، اجازه تخطـــی از حدود هم داده می‌شـــود. اگر نمونه‌ای اهمیت کمتری داشته باشـــد، محدودیت نظیر آن

8-Feasible Region

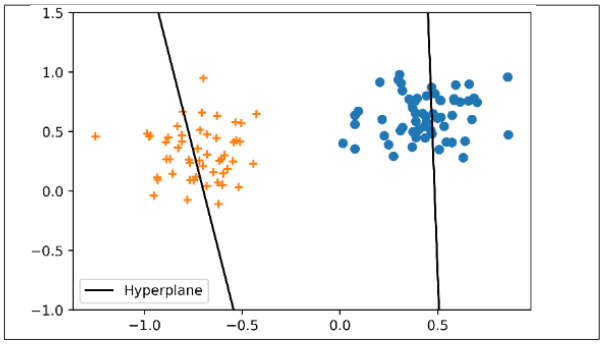
نیز کـم اهمیت‌تر خواهد بود. چنین داده‌ای می‌تواند تخطی بیشـتر از حدی که در رابطهٔ (٦) مشـخص شـده، داشته باشد. حتی می‌توان داده‌های پرت را با یک اشتباه آگاهانه در دسـتهٔ مقابل قرار داد. اما بـرای داده‌های با اهمیت بالا میزان تخطی از حدود کم و ناچیز می‌باشد و جواب مسئله باید در فضای همان قید جسـتجو شود. با این کار فضای شدنی مسئله گسترش می‌یابد و جواب‌های بهینه می‌توانند از فضای وسیع‌تری انتخاب شوند. روش جداسازی جدید و مسـئلهٔ بهینه سـازی نظیرآن به‌صورت زیر بازنویسی می‌شود:

$$Min_{(w,b)} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \quad (۸)$$
$$S.t.\ \ y_i(w^T x_i + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0,\ i = 1,2,\ldots,m.$$

این ساختار، ماشین بردار پشتیبان با قیود نرم نامیده شده است. لازم به ذکر است که در یک مسئلهٔ بهینه‌سازی، اگـر قیود به شـکل فـازی بیان گردنـد، به آن مسـئله، برنامه‌ریـزی فازی نامتقارن گفته می‌شـود. برای حل این گونه مسـئله‌ها روش‌هایی پیشنهاد شده است [۱۸]. اولین گام معرفی یـک تابع عضویت برای تبدیـل کردن روابط فازی به رابطه‌های غیرفازی است. برای توضیحات بیشتر به مرجع [۱۹] مراجعه کنید.

ماشین بردار پشتیبان دوقلو[9]

با گسترش ایدهٔ SVM در سـال ۲۰۰۷ توسط جایادوا [۲۰] و همکاران ایدهٔ یافتن دو ابرصفحهٔ غیرموازی، مطرح شـد که هر یک از آن‌ها تا حـد ممکن به داده‌های یک طبقه نزدیک باشـند و از طبقهٔ مقابل فاصلهٔ مطلوبی بگیرند. این الگوریتم ماشین بردار پشتیبان دوقلو نامیده می‌گردد. شکل (۵) نمایش هندسـی ابرصفحهٔ جداساز در فضای دوبعدی اسـت. در این مسـایل به جای حل یک مسئلهٔ بهینه‌سازی درجه دوم با ابعاد بالا، دو مسئلهٔ درجه دوم جدید با ابعاد نصـف ابعاد قبل و با مرتبهٔ محاسـباتیِ $\frac{1}{4}$ روش قبل حل می‌شوند. این بالاترین مزیّت روش جدید به شمار می‌آید. گسترش‌های قابل توجّهی از ماشین بردار پشتیبان دوقلو مطرح [۲۱] و کاربردهای متعدّدی از این روش ارائه شـده اسـت که از جملهٔ آن‌ها می‌توان به تشخیص رفتار انسانی اشاره کرد [۲۲].



شکل (۵): ماشین بردار پشتیبان دوقلو

در توضیح الگوریتم ماشـین بردار پشـتیبان دوقلو، مجموعـه‌ای از نمونه‌ها و اطّلاعاتی در مورد آن‌ها موجود اسـت که به هر نمونه یک برچسـب اختصاص داده شده و بـه کمک برچسـب‌ها (۱ یا ۱-) می‌تـوان کل نمونه‌ها را به دو گروه دسـته‌بندی کرد. بهتر اسـت که توزیع داده‌ها متعادل باشد یعنی با نسبت تقریباً مساوی، با برچسب ۱ یا ۱- دسته‌بندی شده باشند. در روش بردار پشتیبان دوقلو ، ابرصفحه‌های جداسـاز، لزوماً موازی نیستند (مثلاً h_1 و h_2).

$$h_1:\ x^T w_1 + b_1 = 0 \quad (۹)$$
$$h_2:\ x^T w_2 + b_2 = 0$$

کـه در رابطه‌هـای (۹)، $w_1 \in R^n$ و w_2 نمایشـگر مختصـات ابر صفحه و b_1 و b_2 بایاس می‌باشـند. h_1 تا حـدّ امکان به کلاس ۱+ نزدیک و از کلاس ۱- دور اسـت و h_2 تـا حدّ ممکن به کلاس ۱- نزدیک و از کلاس ۱+ دور می‌باشـد. داده‌های با برچسب منفی باید فاصله‌ای مطلوب و حداقلـی (مثلا یک واحد) از h_1 داشـته باشـند در غیر

9-Twin Support Vector Machine

این‌صورت خطای دسته‌بندی ایجاد خواهد شد و لازم است که پارامترهای مشخص‌کنندهٔ جداساز طوری تنظیم شوند که داده‌هایی از این دست کمتر رخ دهند. به طور مشابه در مورد داده‌های دستهٔ دوم هم چنین وضعیتی رخ خواهد داد. از این‌رو دو مسئلهٔ بهینه‌سازی مطرح خواهد شد و هر یک از این دو مسئله نیاز به کمینه کردن دو مقدار متفاوت دارند؛ کم کردن فاصلهٔ ابرصفحهٔ جداساز h_1 از کلاس ۱+ و کم کردن تعداد نمونه‌هایی از کلاس ۱- که به h_1 نزدیک شده‌اند. تمام مباحث گفته شده در مورد جداساز h_2 هم صدق خواهند کرد. دو مسئلهٔ بهینه‌سازی به صورت زیر تعریف می‌شوند:

$$\underset{(w_1,b_1)}{Min}\ \frac{1}{2}\|Aw_1+e_1b_1\|^2+C_1e_2^Tq_1 \qquad (10)$$

و

$$S.t.\ -(Bw_1+e_2b_1)+q_1\geq e_2,q_1\geq 0$$

$$\underset{(w_2,b_2)}{Min}\ \frac{1}{2}\|Bw_2+e_2b_2\|^2+C_2e_1^Tq_2 \qquad (11)$$

$$S.t.\ (Aw_2+e_1b_2)+q_2\geq e_1,q_2\geq 0$$

که در آن $q_1\in\mathcal{R}^{m_2}$، $q_2\in\mathcal{R}^{m_1}$ و C_1 و C_2 پارامترهای پنالتی با مقادیر مثبت و سایر متغیرها قبلا تعریف شده‌اند. برای حل مسئله‌های فوق، بعد از نوشتن دوگان لاگرانژ و شرایط مکمل زاید به دو مسئلهٔ جدید زیر می‌رسیم:

$$\underset{\alpha}{Min}\ \frac{1}{2}\alpha^TG(H^TH+\epsilon I)^{-1}G^T\alpha-e_2^T\alpha \qquad (12)$$

$$S.t.\ \ 0\leq\alpha\leq C_1e_2$$

و

$$\underset{\beta}{Min}\ \frac{1}{2}\beta^tH(G^TG+\epsilon I)^{-1}H^T\beta-e_1^T\beta \qquad (13)$$

$$S.t.\ \ 0\leq\beta\leq C_2e_1$$

پس از حل مسائل و یافتن بردارهای α و β ابر صفحه‌های دو کلاس از روابط زیر به دست می‌آیند:

$$\begin{bmatrix}w_1\\b_1\end{bmatrix}=-(H^TH+\epsilon I)^{-1}G^T\alpha \qquad (14)$$

$$\begin{bmatrix}w_2\\b_2\end{bmatrix}=(G^TG+\epsilon I)^{-1}H^T\beta \qquad (15)$$

در رابطهٔ (۱۵) ماتریس‌های H و G به‌صورت $H=[A\ e_1]$ و $G=[B\ e_2]$ تعریف شده‌اند. مقدار کوچک و مثبت ϵ به مقادیر درایه‌های موجود در قطر اصلی دو ماتریس افزوده شده تا ماتریس‌ها وارون‌پذیر باشند. در نهایت با حل مسئلهٔ فوق و یافتن ابرصفحات جداکننده، فاصلهٔ نمونه‌های جدید از هر ابر صفحه محاسبه می‌گردد. کمترین فاصله، برچسب دادهٔ آزمایش را مشخّص می‌کند. در صورتی‌که داده‌ها به‌صورت خطی قابل تفکیک نباشند، مشابه SVM از کرنل استفاده شده و ابرصفحه‌هایی با معادله‌های زیر تشکیل خواهند شد:

$$h_1:\ K(x^T,C^T)w_1+b_1=0 \qquad (16)$$
$$h_2:\ K(x^T,C^T)w_2+b_2=0$$

توضیحات بیشتر در مورد داده‌هایی که به‌طور خطی جداناپذیرند، در مرجع [۲۳] ذکر شده‌است.

ماشین بردار پشتیبان دو قلو با قیود نرم[10]

داده‌ها در یافته‌های بالینی مربوط به بیماری‌های مزمن دارای نوفه ذاتی هستند. برای پیش‌بینی دقیق‌تر بیماری، ناچار از کاهش دادن تأثیر چنین داده‌هایی هستیم. در این بررسی برای کاهش تأثیر نوفه، با ایده گرفتن از RSVM و TWSVM، یک نوع ماشین بردار پشتیبان دو قلو با محدودیت‌های نرم معرفی خواهد شد.

ماشین بردار پشتیبان دو قلو دارای دو مسئلهٔ بهینه‌سازی درجه دوم است که در آن قیود حالت بدون انعطاف دارند. اشتراک مناطق محدود به قیدهای هر مسئله، فضای شدنی برای هر یک از دو مسئلهٔ بهینه‌سازی را تشکیل می‌دهد. جواب‌های بهینه در این فضا قرار دارند. نوآوری روش پیشنهادی این است که بدون هرگونه تغییر در تابع هدف، فضای شدنی مسئله را گسترش می‌دهد. استفاده از قیود نَرم فازی به جای محدودیت‌های عادی، فضای شدنی را وسیع‌تر می‌کند؛ الگوریتم پیشنهادی در فضای شدنیِ گسترده‌تر، جواب‌های بهینه را می‌یابد. شیوهٔ نوین در مقایسه با روش‌های مشابه عملکرد بهتری داشته است.

10-Relaxed Constraints Twin Support Vector Machine (RTWSVM)

طبقه‌بند خطی

در ماشین بردار پشتیبان دو قلو طبق روابط (۱۰ و ۱۱)، برای هر یک ازنمونه‌های آزمایش موجود در هر کلاس، محدودیتی مخصوص آن کلاس وجود دارد. اشتراک محدودیت‌ها دو فضای شدنی را می‌سازد. جواب‌های بهینهٔ هر مسئله یعنی $(w_1^*, b_1^*, q_1^*) \in R^{m_2+1+n}$ و $(w_2^*, b_2^*, q_2^*) \in R^{m_1+1+n}$ که مشخص‌کنندهٔ ابرصفحات بهینهٔ جداساز هستند، از این دو فضا انتخاب می‌شوند.

روش پیشنهادی جدید با ثابت نگه‌داشتن تابع هدف و ایجاد تغییر در محدودیت‌ها، به دنبال گسترش دادن فضای شدنی است. برای ساختن فضای شدنی بیشینه با اندازهٔ دلخواه، در محدودیت‌ها از نامساوی‌های فازی استفاده شده است. الگوریتم پیشنهادی برای هر یک از داده‌های آزمایش در دو کلاس، میزانی از تحمل و اطمینان در نظر می‌گیرد تا نمونه‌ها اجازهٔ گذر از محدودیت‌های خود را داشته باشند. هر قدر که میزان تحمل الگوریتم برای گذشتن نمونه از محدودیت بیشتر باشد، محدودیت نظیر آن نرم‌تر خواهد بود. داده‌های کم اهمیت تر اجازهٔ تخطی بیشتری دارند اما داده‌های با اهمیت بالاتر، صرفاً در منطقهٔ محدودیت خود قرار می‌گیرند. داده‌های با اهمیت کمتر می‌توانند فاصله‌ای کمتر از واحد با ابرصفحهٔ مربوط به کلاس مقابل داشته‌باشند. حتی داده‌های نوفه‌ای می‌توانند با چشم‌پوشی آگاهانه در طبقهٔ مقابل دسته‌بندی شوند. مسئله‌های بهینه‌سازی جدید به صورت زیر هستند:

$$Min_{(w_1, b_1)} \ \frac{1}{2}\|Aw_1 + e_1 b_1\|^2 + C_1 e_2^T q_1 \qquad (۱۷)$$
$$S.t. \ -(Bw_1 + e_2 b_1) + q_1 \gtrsim e_2, \ q_1 \geq 0$$

و

$$Min_{(w_2, b_2)} \ \frac{1}{2}\|Bw_2 + e_2 b_2\|^2 + C_2 e_1^T q_2 \qquad (۱۸)$$
$$S.t. \ (Aw_2 + e_1 b_2) + q_2 \gtrsim e_1, \ q_2 \geq 0$$

علامت $\gtrsim$ به این معنی است که برای نمونه‌های یادگیری اجازهٔ تخطی از حدود هم داده می‌شود. استفاده از نامساوی فازی به این محدودیت‌ها انعطاف بیشتری می‌بخشد. در واقع با این‌کار فضای شدنی مسئله گسترش می‌یابد و جواب‌های بهینه می‌توانند از فضای بزرگتری انتخاب شوند. ابتدا می‌بایست دو مسئلهٔ بهینه‌سازی (۱۷ و ۱۸) که مسئله‌های بهینه‌سازی غیرخطی درجهٔ دوم با تابع هدف محدب و قیود نامساوی فازی هستند، حل شوند. در گام اوّل برای حل مسئله، محدودیت‌ها را از فرم ماتریسی خارج کرده و برای هر نمونه از هر طبقهٔ منفی، یک محدودیت به صورت (۱۹) در نظر گرفته می‌شود.

$$-(w_1^T x_i + b_1) \gtrsim 1 - q_{1i}, \ q_{1i} \geq 0 \qquad (۱۹)$$
$$i = 1, \ldots, m_2.$$

که x_i، i_اُمین نمونه از داده‌های با برچسب ۱- و q_{1i} خطای مربوط به این نمونه در اثر تخطی از محدودیت نظیر و نزدیک شدن به h_1 می‌باشد. به‌طور مشابه برای نمونهٔ i_ام از طبقهٔ مثبت، محدودیت به صورت (۲۰) خواهد بود:

$$(w_2^T x_i + b_2) \gtrsim 1 - q_{2i}, \ q_{2i} \geq 0 \qquad (۲۰)$$
$$i = 1, \ldots, m_1.$$

که x_i، i_اُمین نمونه از داده‌های با برچسب مثبت و q_{2i} خطای مربوط به این نمونه در اثر تخطی از محدودیت نظیر و نزدیک شدن به h_2 می‌باشد. برای نمونه‌های با برچسب مثبت، طرفین محدودیت نظیر در (+۱) و برای نمونه‌های با برچسب منفی، طرفین محدودیت نظیر در (−۱) ضرب شده و در نهایت محدودیت‌ها به صورت (۲۱) ادغام خواهند شد:

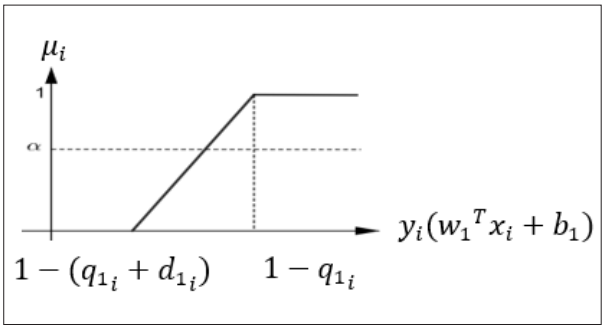
$$y_i(w_1^T x_i + b_1) \gtrsim 1 - q_{1i}, \qquad (۲۱)$$
$$q_{1i} \geq 0, \ i = 1, \ldots, m_2.$$

و

$$q_{2i} \geq 0, i = 1, \ldots, m_1.$$

طبق رابطهٔ (۲۱) هر یک از نمونه‌های آزمایش، یک محدودیت به مسئله تحمیل می‌کنند. مجموعهٔ این محدودیت‌ها فضای شدنی را می‌سازد. متغیرهای مسئله از این فضا انتخاب می‌شوند. هنگام حل مسئله این شرایط

باید برآورده شوند. یک روش برای حل (۲۱) تبدیل مسئله به حالت غیرفازی است. بنابراین گام دوم تعیین توابع عضویت μ_i ،μ'_i برای نامساوی‌های فازی موجود در قیود مسئله است. در روش پیشنهادی هر دو تابع به‌صورت خطی تعریف شده‌اند. شکل (٦) تابع عضویت برای (۱۷) را نشان می‌دهد و فرمول‌های (۲۲ و ۲۳) بیان‌کنندهٔ توابع عضویت μ_i و μ'_i می‌باشند:



شکل (۶): تابع عضویت μ_i

که در آن:

$$\mu_i(w_1,b_1,q_1)=\begin{cases}1, & if\ \ y_i(w_1{}^T x_i+b_1)\geq 1-q_{1_i}\\ \frac{y_i(w_1{}^T x_i+b_1)-1+q_{1_i}+d_{1_i}}{d_{1_i}}, & if\ 1-(q_i+d_{1_i})\leq y_i(w_1{}^T x_i+b_1)\leq 1-q_{1_i}\\ 0, & if\ y_i(w_1{}^T x_i+b_1)\leq 1-(q_{1_i}+d_{1_i})\end{cases} \quad (۲۲)$$

و

$$\mu'_i: R^{m_1+1+n}\to(0,1],\ i=1,2,\dots,m_1.$$

که در آن

$$\mu'_i(w_2,b_2,q_2)=\begin{cases}1, & if\ \ y_i(w_2{}^T x_i+b_2)\geq 1-q_{2_i}\\ \frac{y_i(w_2{}^T x_i+b_2)-1+q_{2_i}+d_{2_i}}{d_{2_i}}, & if\ 1-(q_i+d_{2_i})\leq y_i(w_2{}^T x_i+b_2)\leq 1-q_{2_i}\\ 0, & if\ y_i(w_2{}^T x_i+b_2)\leq 1-(q_i+d_{2_i})\end{cases} \quad (۲۳)$$

از این پس در مورد اولین محدودیت در رابطهٔ (۲۱) و با تمرکز بر داده‌های کلاس منفی و با درنظر گرفتنِ تابع عضویت μ_i، محاسبات مربوط به این طبقه به تفصیل بررسی خواهد شد. μ'_i نیز مشابه خواهد بود.

لازم به توضیح است که اگر q_{1_i} متغیر لغزش و d_{1_i} مقدار تحمل خطا برای نمونهٔ آموزش i–ام و $h_1: x^T w_1+b_1=0$ ابرصفحهٔ جداساز نظیر باشد سه حالت زیر رخ خواهد داد:

اگر $y_i(w_1{}^T x_i+b_1)\geq 1-q_{1_i}$ یعنی فاصلهٔ حداقلی $1-q_{1_i}$ بین نمونهٔ i–ام و ابرصفحهٔ اوّل حفظ شده، این نمونه توسط h_1 به درستی دسته‌بندی شده است و مقدار تعلق فازی برای h_1 برابر ۱ خواهد بود.

نامساوی $y_i(w_1{}^T x_i+b_1)\leq 1-(q_{1_i}+d_{1_i})$ بیان می‌کند که تحمل مُجاز d_{1_i} برای نمونهٔ i–ام در نظر گرفته شده است، ولی این نمونه بیش از تحمل مجاز به ابرصفحهٔ اول نزدیک شده و نتوانسته است محدودیت مسئله را برآورده کند پس h_1 قادر به دسته بندی درست چنین نمونه‌ای نیست و درجه‌ی تعلق فازی برای h_1 صفر خواهد بود.

و معنی نامساوی

$$1-(q_{1_i}+d_{1_i})\leq y_i(w_1{}^T x_i+b_1)\leq 1-q_{1_i}$$

آن است که اگر برای نمونهٔ i–ام بیشینهٔ تحملی به‌اندازهٔ d_{1_i} در نظربگیریم تا بتواند به ابرصفحهٔ کلاس مقابل نزدیک شود، آنگاه رابطهٔ میان مقدار تحمل و درجهٔ عضویت به صورت خطی است و با نزدیک شدن به آستانهٔ تحمل درجهٔ تعلق فازی هم به عدد ۱ نزدیک می‌شود پس h_1 با درجهٔ تعلق μ_i می‌تواند یک ابرصفحهٔ جداساز برای نمونهٔ مزبور باشد.

برای هر یک از اولین محدودیت‌های رابطهٔ (۲۱) تعریف می‌کنیم:

$$X_i = \{(w_1, b_1, q_1) \in R^{m_2+1+n} \mid y_i(w_1{}^T x_i + b_1) \gtrsim 1 - q_{1i}, q_{1i} \geq 0,\ i = 1, \ldots, m_2\} \quad (۲۴)$$

آن‌گاه جواب‌های مسئله باید از مجموعهٔ X با تعریف زیر انتخاب شوند:

$$X = \cap_{i \in I} X_i,\ I = \{1, 2, \ldots, m_2\} \quad (۲۵)$$

بنابراین مسئلهٔ (۱۷) را می‌توان به‌صورت زیر نوشت:

$$Minimize\ \{Q(w_1, b_1, q_1) = C_1 e_2^T q_1 \mid (w_1, b_1, q_1) \in X\} + \frac{1}{2}\|Aw_1 + e_1 b_1\|^2 \quad (۲۶)$$

برای حل (۲۶) از روش آلفا برش استفاده می‌شود. این مسئله در شکل (۶) دیده می‌شود. جواب‌های (۲۶) باید از مجموعهٔ

$$X_\alpha = \{(w_1, b_1, q_1) \in R^{m_2+1+n} \mid \mu_X(w_1, b_1, q_1) \geq \alpha\} \quad (۲۷)$$

انتخاب شوند که:

$$\mu_X(x) = \inf\{\mu_i(x), i \in I\} \quad (۲۸)$$

مجموعهٔ X_α آلفا برش قید i ام است و $\alpha \in (0, 1]$.

جواب بهینهٔ مسئلهٔ (۱۷) با α داده‌شده برابر است با:

$$S(\alpha) = \{(w, b_1, q_1) \in R^{m_2+1+n} \mid C_1 e_2^T q_1 + \frac{1}{2}\|Aw_1 + e_1 b_1\|^2 \leq \{ Min\ \frac{1}{2}\|Aw'_1 + e_1 b'_1\|^2 + C_1 e_1^T q'_1, (w'_1, b'_1, q'_1) \in X_\alpha\} \quad (۲۹)$$

پس می‌توان نوشت:

$$X_\alpha = \cap_{i \in I}\{(w_1, b_1, q_1) \in R^{m_2+1+n} \mid y_i(w_1{}^T x_i + b_1) \geq r_{1i}(\alpha), q_{1i} \geq 0,\ i = 1, \ldots, m_2\} \quad (۳۰)$$

که در آن $r_{1i}(\alpha) = 1 - q_{1i} - d_{1i}(1 - \alpha)$.

زیرا:

$$\frac{y_i(w_1{}^T x_i + b_1) - 1 + q_{1i} + d_{1i}}{d_{1i}} \geq \alpha \Rightarrow y_i(w_1{}^T x_i + b_1) \geq 1 - q_{1i} - d_{1i}(1 - \alpha) \quad (۳۱)$$

پارامترهای α و d_{1i} توسط کاربر تعیین می‌شوند، پارامتر α، که عامل اطمینان نامیده می‌شود، سطحی را مشخص می‌کند که برای سطوح پایین‌تر از آن میزان تعلق فازی نظیر نامساوی موجود در قیود (۲۱)، صفر است. هر چه این مقدار به عدد ۱ نزدیک‌تر باشد، قیود مسئله به TWSVM نزدیک‌تر می‌شود. d_{1i} ها (که در اینجا مقدارشان مساوی d_1 است)، مقدار تحملی هستند که می‌توان به نمونه‌ها نسبت داد. هرچه مقدار تحملی که نسبت می‌دهیم بیشتر باشد، نمونه‌های کلاس منفی، می‌توانند به ابرصفحهٔ جداساز کلاس مثبت، نزدیک‌تر شوند. این بدان معنی است که تعداد بردارهای پشتیبان بیشتر شده و داده‌های بیشتری در تعیین ابرصفحهٔ بهینه نقش خواهند داشت.

در نهایت مدل RTWSVM1 عبارت است از:

$$Min\ \frac{1}{2}\|Aw_1 + e_1 b_1\|^2 + C_1 e_2^T q_1 \quad S.t. -(Bw_1 + e_2 b_1) \geq e_2 - q_1 - d_1(1 - \alpha),\ q_{1i} \geq 0,\ i = 1, \ldots, m_2. \quad (۳۲)$$

گام بعدی حل مسئلهٔ بهینه‌سازی غیرخطی درجه دوم با قیدهای غیرفازی[11] است. تابع هدف اوّلیه همراه با محدودیت‌هایش، به تابع لاگرانژ تبدیل می‌شوند. بردارهای ضرایب نامنفی لاگرانژ در محدودیت‌ها ضرب شده و به تابع هدف افزوده می‌شوند. دوگان مسئله به صورت (۳۳) بازنویسی خواهد شد:

$$Q(w_1, b_1, q_1, \beta_1, \gamma_1) = \frac{1}{2}\|Aw_1 + e_1 b_1\|^2 + C_1 e_2^T q_1 - \beta_1{}^T\{-(Bw_1 + e_2 b_1) + q_1 + d_1(1 - \alpha) - e_2\} - \gamma_1{}^T q_1 \quad (۳۳)$$

که در آن $\beta_1 = (\beta_{11}, \ldots, \beta_{1m_2})^T, \gamma_1 = (\gamma_{11}, \ldots, \gamma_{1m_2})^T$ ضرایب لاگرانژ نامنفی می‌باشند.

پس از مشتق‌گیری نسبت به متغیرها و یافتن ریشه‌ها خواهیم داشت:

11-Crisp

$$\frac{\partial Q(w_1,b_1,q_1,\beta_1,\gamma_1)}{\partial \omega_1} = 0 \Rightarrow \quad (۳۴)$$
$$A^T(Aw_1 + e_1 b_1) + B^T\beta_1 = 0 \quad \text{و}$$

$$\frac{\partial Q(w_1,b_1,q_1,\beta_1,\gamma_1)}{\partial b_1} = 0 \Rightarrow \quad (۳۵)$$
$$e_1{}^T(Aw_1 + e_1 b_1) + e_2{}^T\beta_1 = 0 \quad \text{و}$$

$$\frac{\partial Q(w_1,b_1,q_1,\beta_1,\gamma_1)}{\partial q_1} = 0 \Rightarrow \quad (۳۶)$$
$$C_1 e_2^T - \beta_1{}^T - \gamma_1{}^T = 0$$

به‌علاوه باید شرایط مکمل زاید نیز برقرار شوند:

$$\beta_1{}^T\{-(Bw_1 + e_2 b_1) + q_1 + d_1(1-\alpha) - e_2\} = 0 \text{ و } \gamma_1{}^T q_1 = 0 \quad (۳۷)$$

از ترکیب روابط (۳۴ و ۳۵) خواهیم داشت:

$$U_1 = -(H^T H + \epsilon I)^{-1} G^T \beta_1 \quad (۳۸)$$

که در آن: $U_1 = [w_1\ b_1]^T\ H = [A\ e_1]$ و $G = [B\ e_2]$.

با جای‌گذاری (۳۸) در (۳۳) و ساده کردن روابط و استفاده از رابطهٔ (۳۶)، مسئلهٔ نهایی به صورت زیر خواهد بود:

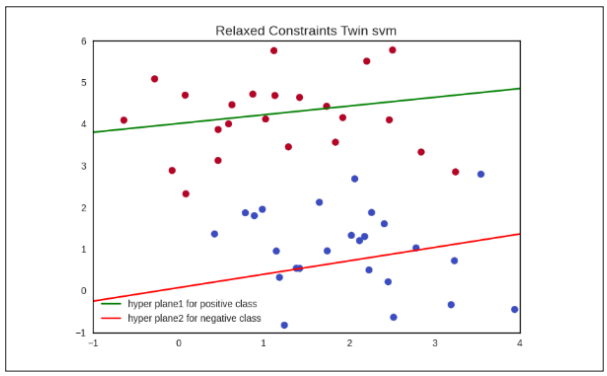
$$Min \frac{1}{2}\beta_1{}^T G(H^T H + \epsilon I)^{-1} G^T \beta_1 - \beta_1{}^T\{d_1(1-\alpha) - e_2\} \quad (۳۹)$$
$$S.t. \quad 0 \le \beta_1 \le C_1 e_2$$

مسئله مشابه تمام روابط (۲۴ تا ۳۹) برای داده‌های کلاس دیگر نیز برقرار است. با حل (۳۹) ضرایب لاگرانژ به دست می‌آیند. دو ابرصفحهٔ غیرموازی، مشخص می‌شوند. فاصلهٔ عمودی نمونه‌های جدید از هر ابر صفحه محاسبه می‌گردد. کمترین فاصله، برچسب دادهٔ آزمایش را مشخّص می‌کند:

$$class(x) = \arg\underset{j=1,2}{Min}\left(d_j(x)\right),$$

که در آن $d_j(x)$ از دستور (۴۰) محاسبه می‌شود:

$$d_j(x) = \frac{|x^T w_j + b_j|}{\|w_j\|} \quad (۴۰)$$



شکل (۷): نمایش هندسی بردار پشتیبان دوقلو با قیود نرم.

طبقه‌بند غیرخطی

مسایل یادگیری ماشین در دنیای واقعی غالباً به صورت خطی جداپذیر نیستند. روش پیشنهادی RT-WSVM برای جداسازی مسایل غیرخطی، نمونه‌ها را به فضای ویژگی با ابعاد بالاتر نگاشت می‌کند. بدین‌منظور دو ابرسطح[12] به صورت روابط (۱۶) تعریف می‌شوند، که در آن $C = [A\ B]$ و K تابع کرنل دلخواه می‌باشد. در نسخهٔ غیرخطی، مسایل بهینه‌سازی اصلی برای کلاس مثبت و منفی به ترتیب در روابط (۴۱ و ۴۲) تعریف شده است.

$$\underset{(w_1,b_1)}{Min}\ \frac{1}{2}\|K(A,C^T)w_1 + e_1 b_1\|^2 + C_1 e_2^T q_1 \quad (۴۱)$$
$$S.t. \quad -(K(B,C^T)w_1 + e_2 b_1) + q_1 \gtrsim e_2, \quad q_1 \ge 0 \quad \text{و}$$

$$\underset{(w_2,b_2)}{Min}\ \frac{1}{2}\|K(B,C^T)w_2 + e_2 b_2\|^2 + C_2 e_1^T q_2 \quad (۴۲)$$
$$S.t. \quad (K(A,C^T)w_2 + e_1 b_2) + q_2 \gtrsim e_1,\ q_2 \ge 0.$$

برای مسئلهٔ (۴۱)، مشابه حالت خطی، تابع عضویت فازی μ_i تعریف شده و به کمک روش آلفا برش، نامساوی‌های فازی به غیرفازی تبدیل می‌شوند.

رابطهٔ (۴۳) مسئله را در حالت غیرخطی و با قیود غیر فازی نشان می‌دهد:

12-Hypersurface

جدول (۱): مشخصات مجموعهٔ داده برای ارزیابی روش RTWSVM

مجموعهٔ داده	تعداد نمونه‌ها و ویژگی	تعداد نمونه‌های مثبت	تعداد نمونه‌های منفی	نسبت منفی/مثبت	دادهٔ از دست‌رفته
CKD	۲۴×۴۰۰	۲۵۰	۱۵۰	۱/۷	ندارد
Hep	۱۹×۱۵۵	۱۲۳	۳۲	۳/۸	دارد
Heart	۱۳×۲۷۰	۱۲۰	۱۵۰	۰/۸۰	ندارد
PIDD	۸×۷۶۸	۲۶۸	۵۰۰	۰/۵۴	ندارد
WDBC	۳۰×۵۶۹	۲۱۲	۳۵۷	۰/۵۹	دارد

$$Min\ \frac{1}{2}\|K(A,C^T)w_1+e_1b_1\|^2+C_1e_2^Tq_1 \quad (43)$$
$$S.t.\ -(K(B,C^T)w_1+e_2b_1)\geq e_2-q_1-d_1(1-\alpha)$$
$$q_{1_i}\geq 0,\ i=1,\dots,m_2.$$

مسئلهٔ دوگان (۴۳) به صورت (۴۴) خواهد بود:

$$Q(w_1,b_1,q_1,\beta_1,\gamma_1)=\frac{1}{2}\|K(A,C^T)w_1+e_1b_1\|^2+C_1e_2^Tq_1-\beta_1^T\{-(K(B,C^T)w_1+e_2b_1)+q_1+d_1(1-\alpha)-e_2\}-\gamma_1^Tq_1 \quad (44)$$

پس از مشتق‌گیری و برقراری شرایط مکمل زاید:

$$U_1=-(S^TS+\epsilon I)^{-1}R^T\beta_1 \quad (45)$$

که در آن: $S=[K(A,C^T)\ e_1]\ R=[K(B,C^T)\ e_2],$
$U_1=[w_1\ b_1]^T$

و مشابه نسخهٔ خطی کلاس نمونهٔ جدید x با محاسبهٔ فاصلهٔ عمودی آن از دو ابرسطح مشخص می‌شود، به‌طوری‌که تابع تصمیم برای نسخهٔ غیرخطی به‌صورت زیر تعریف می‌گردد:

$$class(x)=\arg\underset{j=1,2}{Min}\left(d_j(x)\right),$$

که در آن $d_j(x)$ از دستور (۴۶) محاسبه می‌شود:

$$d_j(x)=\frac{|K(x^T,C^T)w_j+b_j|}{\|w_j\|} \quad (46)$$

نتایج و ارزیابی

پایگاه داده

داده‌های مورد استفاده در این تحقیق، از پایگاه‌های دادهٔ Kaggle و UCI استخراج شده‌اند. به دلیل اهمیت موضوع پنج مجموعه دادهٔ PIDD، Heart-statlog، Hep، CKD و WDBC که به ترتیب مرتبط با بیماری کلیه، هپاتیت، بیماری قلبی، دیابت و سرطان سینه هستند، مورد بررسی قرار گرفته‌اند. داده‌های توصیفی[13] از دست رفته با اولین مُد در ستون مزبور و داده‌های عددی با صفر جایگزین شده‌اند. عدم تعادل بین نمونه‌ها در مجموعه داده‌های CKD و Hep مشهود است. جدول (۱) مجموعهٔ داده را توصیف می‌کند.

نحوهٔ پیاده‌سازی و اجرای الگوریتم‌ها

تمام روش‌ها در زبان برنامه‌نویسی پایتون و در محیط رایگان برخط google colab پیاده‌سازی شده‌اند. از کتابخانهٔ cvxopt برای بهینه‌سازی و از کتابخانه‌های pandas، numpy و matplotlib جهت محاسبات ریاضی، ساختن چارچوبی برای داده‌ها و رسم نمودارها استفاده شده است.

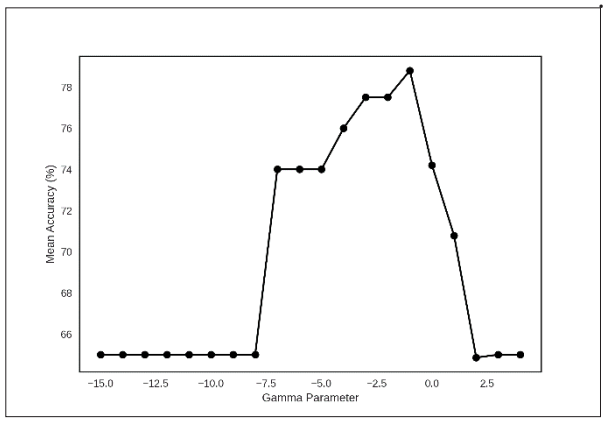
از آن‌جا که در دنیای واقعی داده‌ها جداپذیر خطی نیستند، تابع هستهٔ گوسی که قدرت تعمیم‌پذیری بهتری دارد به کار رفته است:

$$K(x,x')=\exp\left(-\frac{\|x-x'\|^2}{2\gamma^2}\right) \quad (47)$$

دقت روش RTWSVM به انتخاب پارامترهای بهینه بسیار وابسته است. بدین‌منظور از روش جستجوی شبکه‌ای برای پیدا کردن پارامتر بهینه استفاده شده است. پارامترهای پنالتی c_1 وc_2از مجموعهٔ $\{2^i|i=-7,-6,\dots,7\}$ و پارامترهای α_1، α_2، d_1 و d_2که به ترتیب، برای سطح اطمینان برش آلفا و ضریب تحمل الگوریتم، مورد استفاده بوده‌اند از مجموعهٔ [۰/۹۵ و ۰/۷۵ و ۰/۲۵] و پارامتر

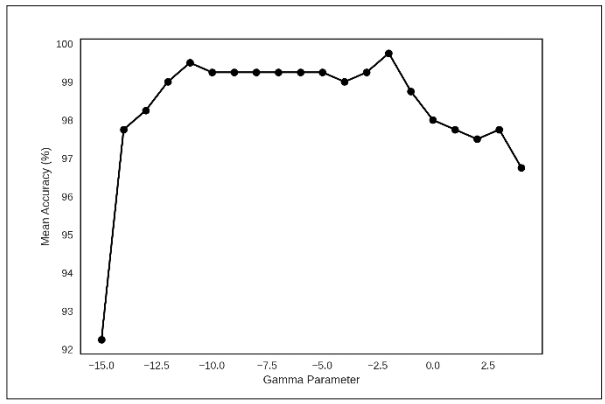
13-Categorical

تابع هسته γ نیز از $\{2^i | i = -15, -14, \ldots, 5\}$ انتخاب شده‌اند. برای تعیین میزان دقّت روش، اعتبارسنجی متقابل به کار رفته است؛ الگوریتم ارائه شده با ۹۰٪ داده‌ها (k=10) آموزش دیده و با ۱۰٪ باقیمانده آزمایش شده، این مراحل ده بار تکرار شده‌اند. پس از ۱۰ بار تکرار، میانگین و انحراف معیار دقت روش پیشنهادی به‌دست آمده‌اند. شکل‌های (۸- ۱۱) بیان‌گر تأثیر تغییرات پارامتر گاما بر میزان دقت روش پیشنهادی در پایگاه‌های دادهٔ PIDD، Heart-statlog ، CKD و WDBC می‌باشند.

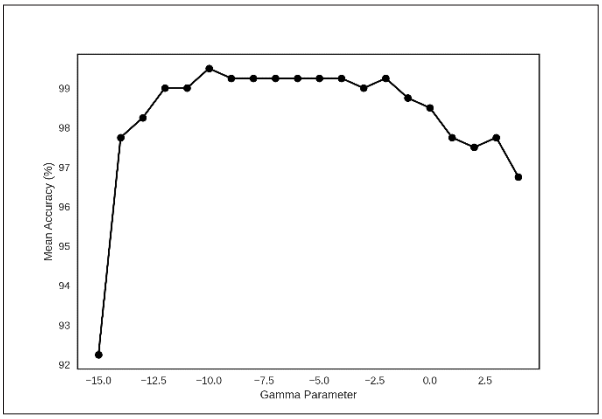


شکل (۸): تاثیر پارامتر گاما بر دقت الگوریتم درپایگاه دادهٔ PIDD

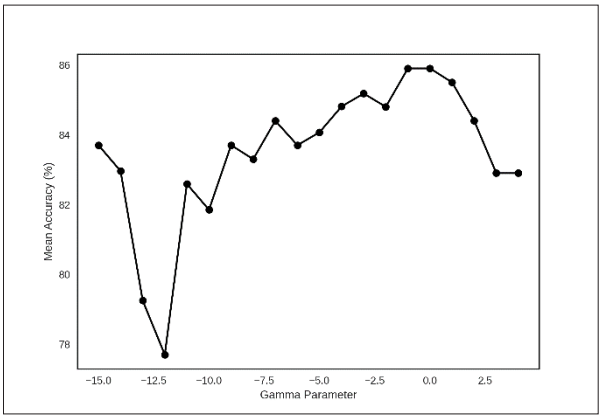
به منظور بزرگنماییِ نمودارِ رابطهٔ بین گاما و دقّت الگوریتم، نمودار دقت، بر اساس لگاریتم گاما در مبنای ۲ رسم شده است. بیشترین حسّاسیت نسبت به تغییرات گاما در مجموعه دادهٔ Heart-statlog قابل مشاهده است اما WDBC و CKD نسبت به تغییرات مقاوم‌تر بوده‌اند.



شکل (۹): تاثیر پارامتر گاما بر دقت الگوریتم درپایگاه دادهٔ CKD



شکل (۱۰): تاثیر پارامتر گاما بر دقت الگوریتم درپایگاه دادهٔ WDBC



شکل (۱۱): تاثیر پارامتر گاما بر دقت الگوریتم درپایگاه دادهٔ Heart-statlog

نتایج تجربی و تجزیه و تحلیل آماری بر روی داده‌های دنیای واقعی

با هدف تحلیل عملکرد، روش پیشنهادی، همراه با SVM ساده، رگرسیون لجستیک، درخت تصمیم و نزدیک‌ترین همسایه، را بر روی پنج مجموعهٔ داده‌های جدول (۱)، مقایسه شده‌اند. نتایج روش جدید هم با نتایج به‌دست آمده از مرجع [۲۴] و هم با پیاده‌سازی دوبارهٔ مدل‌های ذکر شده، بر روی داده‌های بیماری‌های مزمن، گزارش شده‌اند. به جز نتایج مجموعه دادهٔ هپاتیت که دچار عدم تعادل محسوس می‌باشد، اعداد به دست آمده از پیاده‌سازی مجدد و جدول گزارش شده در مرجع [۲۴] تقریبا یکسان هستد که اعتبار پژوهش را تایید می‌کند. نتایج مقایسهٔ دقت و رتبهٔ الگوریتم پیشنهادی با روش‌های پیشین در جدول‌های (۲- ۵) قابل

جدول (۲): مقایسهٔ دقت دسته‌بندهای پیشین و روش پیشنهادی برگرفته از مرجع [۲۴]

دقت	SVM	LR	DT	KNN	RTWSVM
CKD	۹۸/۷۵	۹۸/۷۵	۹۷/۵۰	۹۵/۰۰	۹۹/۵۰
Hep	۸۰/۰۰	۸۳/۰۰	۸۳/۰۰	۸۰/۰۰	۱۰۰/۰۰
PIDD	۷۷/۹۲	۷۹/۸۷	۷۴/۶۸	۷۶/۶۲	۷۸/۸۱
WDBC	۹۷/۱۴	۹۷/۱۴	۹۵/۰۰	۹۷/۱۴	۹۸/۳۹
میانگین	۸۸/۲۲	۸۹/۶۹	۸۷/۵۴	۸۷/۱۹	۹۷/۱۴

جدول (۳): مقایسهٔ رتبهٔ دسته‌بندهای پیشین و روش پیشنهادی برگرفته از مرجع [۲۴]

رتبه	SVM	LR	DT	KNN	RTWSVM
CKD	۳	۲	۴	۵	۱
Hep	۴	۲	۲	۴	۱
PIDD	۳	۱	۵	۴	۲
WDBC	۳	۳	۵	۳	۱
میانگین	۳/۳۷۵	۲/۱۲۵	۴/۱۲۵	۴/۱۲۵	۱/۲۵

جدول (۴): مقایسهٔ دقت دسته‌بندهای پیشین و روش پیشنهادی

دقت	SVM	LR	DT	KNN	RTWSVM
CKD	۹۷/۰۰∓ ۱/۸۷	۹۸/۰۰± ۱/۵۰	۹۵/۲۵± ۳/۰۵	۹۷/۲۵± ۲/۰۷	۹۹/۵۰± ۱/۰۰
Heart	۸۳/۳۳± ۸/۹۵	۸۳/۷۰± ۸/۶۳	۷۴/۴۴± ۷/۴۹	۸۱/۱۱± ۷/۴۹	۸۵/۹۲± ۷/۵۵
Hep	۰±۱۰۰	۰±۱۰۰	۰±۱۰۰	۸۸/۴۱± ۷/۷۱	۰±۱۰۰
PIDD	۷۷/۵۹± ۴/۱۷	۷۷/۱۹± ۴/۶۸	۶۹/۵۳± ۵/۱۶	۷۳/۸۳± ۳/۷۵	۷۸/۸۱± ۵/۵۹
WDBC	۹۷/۵۴± ۲/۹۵	۹۵/۷۸± ۳/۴۳	۹۱/۷۵± ۴/۳۶	۹۶/۶۶± ۲/۴۱	۹۸/۳۹± ۱/۴۸
میانگین	۹۱/۰۹۲	۹۰/۹۳۴	۸۶/۱۹۴	۸۷/۴۵۲	۹۲/۵۲۴

مشاهده‌اند؛ RTWSVM عملکرد بهتری دارد.

آزمون آماری ساده و ناپارامتریک فریدمن از توزیع χ^2 با $(k-1)$ درجهٔ آزادی که در آن k تعداد مدل‌ها و N مجموعهٔ داده مورد استفاده قرار گرفته است. به منظور انجام آزمون آماری، میانگین رتبهٔ چهار روش بر اساس دقت محاسبه شده و در جدول (۲ و ۳) نشان داده شده است. ابتدا بر اساس فرضیهٔ صفر، تمام روش‌ها را یکسان در نظر می‌گیریم و سپس آزمون فریدمن از رابطهٔ (۴۸) محاسبه می‌شود.

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right] \qquad (۴۸)$$

در رابطهٔ (۴۸)، $R_j = \frac{1}{N}\sum_j r_i^j$، میانگین رتبهٔ روش j ام و r_i^j نشان‌دهندهٔ رتبهٔ j امین روش بر روی i امین مجموعهٔ داده از N مجموعهٔ داده، است:

$$\chi_F^2 = \frac{12 \times 5}{5(5+1)}\left[2/7^2 + 7 + 4/5^2 + 3/8^2 + 1/3^2 - \frac{5(5+1)^2}{4}\right] = 11/92$$

سپس مقدار F_F براساس فرمول (۴۹) محاسبه می‌شود.

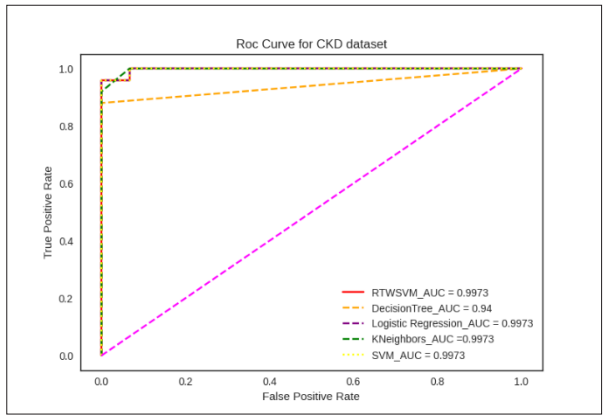
$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2} = \frac{4 \times 11/92}{5 \times 4 - 11/92} = 5/9 \qquad (۴۹)$$

که در آن $k-1=4$، $(k-1)(N-1)=16$ و F_F از توزیع F با $(k-1)(N-1)$ درجه آزادی پیروی می‌کند. با در نظرگرفتن سطح اهمیت $\alpha = 0/05$، مقدار بحرانی $F(4,16) = 3/01$ به‌دست می‌آید و از آن‌جا که مقدار F_F از آن بیشتر است، فرضیهٔ صفر رد می‌شود.

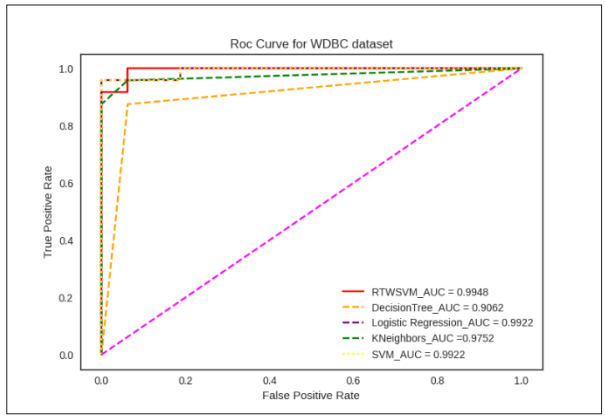
یکی از روش‌های بررسی و ارزیابی عملکرد دسته‌بندی دودویی، نمودار مشخصه عملکرد یا به اختصار منحنی ROC است. کارایی الگوریتم‌های دسته‌بندهای دودویی معمولا به وسیله شاخص‌هایی مثل حسّاسیت[14] و صراحت[15] سنجیده می‌شوند. اما در نمودار ROC هر دوی این شاخص‌ها ترکیب شده و به صورت یک منحنی نمایش داده می‌شوند. منطقه زیر منحنی[16] اندازه‌گیری، بیان‌گر توانایی طبقه‌بندی‌کننده برای تمایز بین کلاس‌ها است و به عنوان خلاصه منحنی ROC استفاده می‌شود. هر چه AUC بالاتر باشد، عملکرد مدل در تشخیص کلاس‌های مثبت و

14-Sensitivity
15-Specificity
16-Area Under the Curve

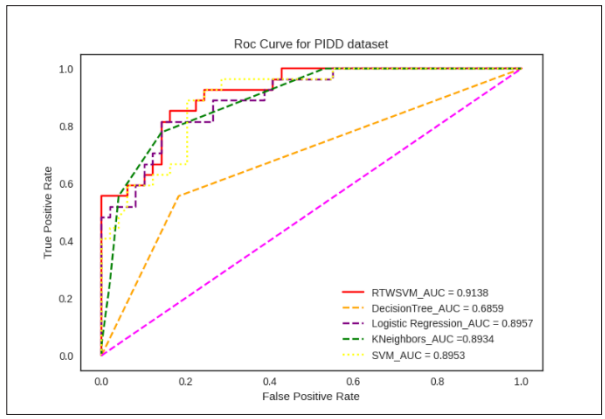
منفی بهتر اســت. وقتی AUC = ۱، طبقه‌بندی‌کننده می‌تواند کاملاً بین تمام نقاط کلاس مثبت و منفی به درســتی تمایز قائل شــود و اگر AUC = 0، طبقه‌بندی‌کننده همه منفی‌ها را به‌عنوان مثبت و همه مثبت‌هــا را به‌عنوان منفی پیش‌بینی خواهد کرد. عملکرد روش پیشــنهادی در شکل‌های (۱۲-۱۶) در قالب نمودارهای ROC آمده است.



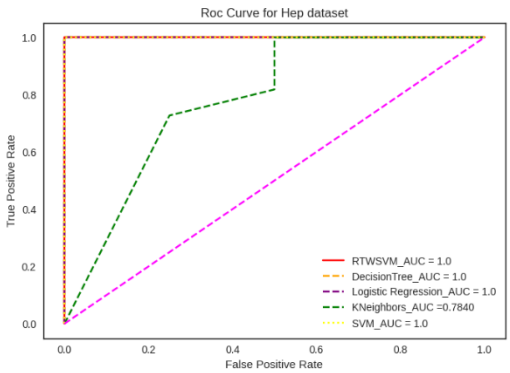
شکل (۱۲): نمودار مشخصه عملکرد بر مجموعه دادهٔ CKD



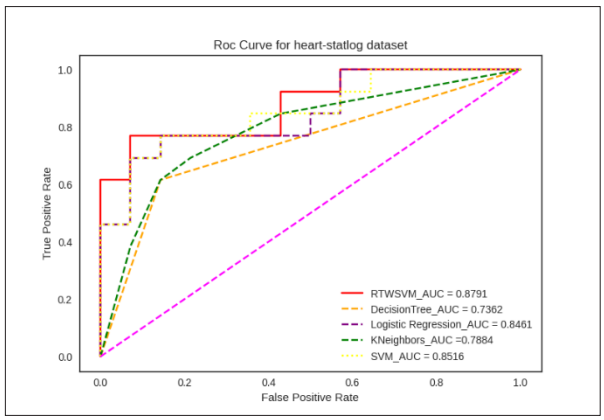
شکل (۱۳): نمودار مشخصه عملکرد بر مجموعه دادهٔ WDBC



شکل (۱۴): نمودار مشخصه عملکرد بر مجموعه دادهٔ PIDD



شکل (۱۵): نمودار مشخصه عملکرد بر مجموعه دادهٔ Hep



شکل (۱۶): نمودار مشخصه عملکرد بر مجموعه دادهٔ Heart-statlog

نتیجه‌گیری و پیشنهادهای آتی

گســترش فضای شــدنی، عــدم تأثیر منفی بــر تابع هدف، کاهش اثر داده‌های پرت، ســرعت بیشتر و کارایی بالاتر از ویژگی‌های مثبت روش جدید اســت. به‌کارگیری الگوریتم‌های جستجو مانند الگوریتم‌های تکاملی، می‌تواند یافتن مقدار بهینه برای پارامترهای زیاد موجود را سرعت بخشد. همچنین می‌توان با در نظرگرفتن یک ماتریس وزنی ضرایب تحمل، به هر یک از نمونه‌ها توجه ویژه داشــت. با این کار تأثیر داده‌های نوفهٔ کاهش خواهد یافت.

مراجع

[1] Souza-Pereira, L., Pombo, N., Ouhbi, S., Felizardo, V. and Garcia, N., 2020. «Clinical decision support systems for chronic diseases: A systematic literature review», Computer

Methods and Programs in Biomedicine, Vol. 195, pp. 105565.
[2] Alkenani, A. H., Li, Y., Xu, Y. & Zhang, Q., 2021. "Predicting Alzheimer's disease from spoken and written language using fusion-based stacked generalization", Journal of Biomedical Informatics, Vol. 118, pp. 103803.
[3] Yuan, X., Chen, S., Yuwen, L., An, S., Mei, S. & Chen, T., 2020. "An improved SEIR model for reconstructing the dynamic transmission of COVID-19", In Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2320 – 2327.
[4] Guo, Y. et al., 2021. "A review of wearable and unobtrusive sensing technologies for chronic disease management", Computers in Biology and Medicine, Vol. 129, pp. 104163.
[5] Wang, Haidong, et al., 2016. "Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015", The lancet 388, Vol. 10053, pp. 1459-1544.
[6] Beyer, K.M., Namin, S., 2023. "Chronic environmental diseases: burdens, causes, and response», In Biological and environmental hazards, risks, and disasters, Elsevier, pp. 223-249.
[7] Organization, W. H. WHO Noncommunicable diseases https://www.who.int/westernpacific/health-topics/noncommunicable-diseases
[۸] نورالدّینی، صدرالدّین.، ۱۴۰۰. «اثر اقتصادی بیمار یها ی مزمن بر هزینه و درآمد خانوار ایرانی»، نشریه علمی (فصلنامه)، پژوهشها و سیاستهای اقتصادی شماره ۹ ، صفحه ۲۰۷ – ۲۴۱.
[9] Delpino, F.M., Costa, Â.K., Farias, S.R., Chiavegatto Filho, A.D.P., Arcêncio, R.A. and Nunes, B.P., 2022. "Machine learning for predicting chronic diseases: a systematic review», Public Health, Vol. 205, pp.14-25.
[10] Maniruzzaman, M., Rahman, M.J., Ahammed, B. and Abedin, M.M., 2020. "Classification and prediction of diabetes disease using machine learning paradigm», Health information science and systems, Vol. 8, pp. 1-14.
[11] Huang, J., Huth, C., Covic, M., Troll, M., Adam, J., Zukunft, S., Prehn, C., Wang, L., Nano, J., Scheerer, M.F. and Neschen, S., 2020. "Machine learning approaches reveal metabolic signatures of incident chronic kidney disease in individuals with prediabetes and type 2 diabetes», Diabetes, Vol. 69(12), pp.2756-2765.
[12] Zhang, L., Yuan, M., An, Z., Zhao, X., Wu, H., Li, H., Wang, Y., Sun, B., Li, H., Ding, S. and Zeng, X., 2020. "Prediction of hypertension, hyperglycemia and dyslipidemia from retinal fundus photographs via deep learning: A cross-sectional study of chronic diseases in central China», PloS one, Vol. 15(5), p.e0233166.
[13] Al-Azzam, N. and Shatnawi, I., 2021. "Comparing supervised and semi-supervised machine learning models on diagnosing breast cancer», Annals of Medicine and Surgery, V0l. 62, pp.53-64.
[14] Xie, Y., Meng, W.Y., Li, R.Z., Wang, Y.W., Qian, X., Chan, C., Yu, Z.F., Fan, X.X., Pan, H.D., Xie, C. and Wu, Q.B., 2021. "Early lung cancer diagnostic biomarker discovery by machine learning methods», Translational oncology, V0l. 14(1), pp. 100907.
[15] Mezzatesta, S., Torino, C., De Meo, P., Fiumara, G. and Vilasi, A., 2019. "A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis», Computer methods and programs in biomedicine, Vol. 177, pp. 9-15.
[16] Yuan, X., Chen, S., Sun, C. and Yuwen, L., 2022. "A novel early diagnostic framework for chronic diseases with class imbalance», Scientific Reports, Vol. 12(1), pp. 8614.
[17] Tanveer, M., Rajani, T., Rastogi, R., Shao, Y.H. and Ganaie, M.A., 2022. Comprehensive review on twin support vector machines. Annals of Operations Research, pp.1-46.
[۱۸] ناجی عظیمی. زهرا، تابستان ۹۵، آشنایی با برنامه‌ریزی خطی فازی، ویراستهٔ وحیدیان کامیاد، علی، ویرایش اوّل، مشهد، انتشارات دانشگاه فردوسی مشهد.
[۱۹] سبزه‌کار، مصطفی، آذر ۸۸، بررسی تأثیر توجه مضاعف به نمونه‌های یادگیری با استفاده از قیود بهینه‌سازی طبقه‌بندهای ماشین بردار پشتیبان، پایان‌نامهٔ کارشناسی ارشد، گروه مهندسی کامپیوتر، دانشکده مهندسی، دانشگاه فردوسی مشهد، صفحات ۸۲-۹۰.
[20] Jayadeva, Khemchandani, R. and Chandra, S., 2018. "Twin support vector machines: models, extensions and applications», Springer Publishing Company, Incorporated.
[21] Nasiri, J.A. and Mir, A.M., 2020. "An enhanced KNN-based twin support vector machine with stable learning rules», Neural computing and applications, Vol. 32(16), pp.12949-12969.
[22] Mozafari, K., Nasiri, J.A., Charkari, N.M. and Jalili, S., 2011, December. Action recognition by local space-time features and least square twin SVM (LS-TSVM). In 2011 first international conference on informatics and computational intelligence (pp. 287-292). IEEE.
[23] Raschka, S. and Mirjalili, V. 2017, Python machine learning second edition. Birmingham, England: Packt Publishing.
[24] Yuan, X., Chen, S., Sun, C. and Yuwen, L., 2022, "A novel early diagnostic framework for chronic diseases with class imbalance», Scientific Reports, Vol. 12(1), pp. 8614.