پنجمین کنفرانس ملی انفورماتیک ایران
پژوهشگاه دانشهای بنیادی، پردیس فرمانیه، تهران
۱۳ و ۱۴ دی ماه ۱۴۰۲

پیش‌بینی ابتلا به بیماریهای مزمن به کمک ماشین بردار پشتیبان دوقلو با قیود نرم

حمیده فدیشه‌ای [۱]، جلال‌الدّین نصیری [۲]، سهراب عفّتی [۳]

[۱] گروه ریاضی کاربردی، دانشکده علوم ریاضی، دانشگاه فردوسی مشهد، مشهد
ha.fadishehei@mail.um.ac.ir
[۲] استادیار، گروه ریاضی کاربردی، دانشکده علوم ریاضی، دانشگاه فردوسی مشهد، مشهد
Jnasiri@um.ac.ir
[۳] استاد، گروه ریاضی کاربردی، دانشکده علوم ریاضی، دانشگاه فردوسی مشهد، مشهد
s-effati@um.ac.ir

چکیده

بیماری‌های مزمن از چالش‌های جدّی مربوط به سلامت انسان می‌باشند. تشخیص بهنگام آن‌ها می‌تواند برای داشتن سبک زندگی سالم مفید باشد. روش‌های یادگیری ماشین از جمله ماشین بردار پشتیبان برای پیش‌بینی این بیماری‌ها قابل استفاده هستند. در این بررسی به کمک یک روش ماشین بردار پشتیبان دوقلو با قیود نرم، سعی شده‌است تا به کمک اطّلاعات پزشکی بیماران، پیش‌بینی شود که آیا آن‌ها به بیماری‌های مزمن، مبتلاخواهند شد یا خیر.

ماشین بردار پشتیبان عادی به دنبال یافتن دو ابرصفحهٔ موازی با بیشترین فاصله است به‌طوری‌که داده‌های دو طبقه در دوطرف آن‌دو قرار گیرند. در ماشین بردار پشتیبان دوقلو، دو ابرصفحه لزوماً موازی نیستند. هریک از دو ابرصفحه نزدیکترین فاصله را به داده‌های کلاس خود داشته و از کلاس مقابل دارای یک فاصلهٔ حداقلی می‌باشند. سرعت این روش بهتر است ولی نسبت به داده‌های نویزی عملکرد خوبی ندارد. در روش پیشنهادی با استفاده از بهینه‌سازی فازی، محدودیت‌های مسأله به‌صورت روابط فازی در نظرگرفته می‌شوند؛ با این‌کار فضای شدنی مسأله بهینه‌سازی درجه دوم گسترش یافته، به نمونه‌ها اجازهٔ تخطی از ابر صفحات داده‌شده و تأثیر داده‌های دورافتاده در تشخیص نهایی کاهش یافته‌است. الگوریتم پیشنهادی، برروی داده‌های بالینی پزشکی، اجرا ونتایج با روش‌های مشابه مقایسه شده‌اند روش پیشنهادی عملکرد بهتری داشته است.

کلمات کلیدی

یادگیری ماشین، ماشین بردار پشتیبان دوقلو، بهینه‌سازی محدّب، محدودیت‌های فازی.

۱- مقدمه

بیماری‌های مزمن به دلیل طولانی بودن دوران ابتلا، عوارض آزاردهنده‌ی مختلف و زمینه‌سازی برای مبتلا شدن به بیماری‌های عفونی دیگر از جمله کرونا، از مهمترین مسأله‌های بهداشتی در جهان هستند. در سال ۲۰۱۹، حدود ۶۳٪ از مرگ و میر جهانی ناشی از بیماریهای مزمن بوده است. رایج‌ترین این بیماری‌ها، فشار خون، قند خون، و بیماریهای قلبی هستند که از نتایج سبک زندگی نادرست می‌باشند [9]. عوارض ناشی از آنها باعث آسیب رسیدن به قلب، کلیه، مغز و چشم‌ها می‌شود که آثاری مخرّب روی کار و زندگی افراد خواهد داشت[3]. افزون بر موارد ذکر شده این بیماران در برابر بیماری‌های عفونی (مثلاً ویروس کرونا) ایمنی کمتری دارند[11]. مطابق گزارش‌های سال ۲۰۱۹، ۴۸٪ افراد مبتلا به این ویروس بیماری مزمن نیز داشته و احتمال بروز علایم شدید در آنها بیشتر است [4]. به همهٔ موارد فوق هزینه‌های گرانی که به دولت‌ها و خانواده‌ها تحمیل می‌شوند نیز، قابل تأمّل است. بنابراین تشخیص زودهنگام بیماری کمک قابل توجّهی به سلامت فرد و جامعه خواهد داشت.

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 [10]. در طی سال‌ها انواع جدیدتر، دقیق‌تر و کاراتری از ماشین بردار پشتیبان ارائه شد. برخی از این روش‌ها در زیر بیان شده‌اند.

۲- پژوهش‌های پیشین

۲-۱- 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[1] است ، $y_i \in \{1, -1\}$ برچسب کلاس دادهٔ x_i و $A_{m_1 \times n}$ مجموعهٔ نمونه‌های با برچسب ۱-، $B_{m_2 \times n}$ مجموعهٔ نمونه‌های با برچسب ۱+ 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 دوابرصفحهٔ موازی است که تا حدامکان از یکدیگر دور بوده و داده‌های دو طبقه در دوطرف آنها واقع باشند. با مدل‌سازی مسألهٔ بالا و حل مسألهٔ بهینه‌سازی (۱)، میتوان ابرصفحهٔ جداساز را مشخص کرد.

$$\min_{(\omega\,,b)} \frac{1}{2}\|\omega\|^2 \quad \text{(۱)}$$
$$S.t.\quad y_i(\omega^T x_i + b) \geq 1, i = 1,2,\dots,m.$$

نمایش ماتریسی مسألهٔ (۱) به‌فرم (۲) خواهد بود:

$$\min_{(\omega\,,b)} \frac{1}{2}\|\omega\|^2 \quad \text{(۲)}$$
$$S.t.\quad A\omega + e_1 b \geq e_1,$$
$$B\omega + e_2 b \leq -e_2$$

در (۲)، e_1 و e_2 بردارهای با درایه‌های واحد و متناسب با A و B هستند.

در حالتی‌که داده‌ها در فضای ورودی، به‌صورت خطی جداناپذیر هستند، مسألهٔ *svm* با حاشیهٔ نرم به‌صورت (۳) مطرح خواهد شد:

$$\min_{(\omega\,,b)} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m}\xi_i \quad \text{(۳)}$$
$$S.t.\quad y_i(\omega^T x_i + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0,\ i = 1,2,\dots,m.$$

که در آن ξ_i متغیر لغزش برای کاهش تأثیر دادهٔ نویزی x_i است.

پس از نوشتن تابع لاگرانژ و در نظرگرفتن شرایط مکمل زاید و حل مسألهٔ (۳) ابرصفحهٔ جدااساز معلوم خواهد شد.

$$\min \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j - \sum_{k=1}^{m}\alpha_k \quad \text{(۴)}$$
$$S.t.\quad \sum_{i=1}^{m}\alpha_i y_i = 0\,,$$
$$0 \leq \alpha_i \leq C\ ,i = 1,2,\dots,m.$$

که α_i ضرایب لاگرانژ هستند و تابع تصمیم $D(x)$ طبق فرمول (۴) موقعیت دادهٔ جدید x را نسبت به ابرصفحه مشخص خواهد کرد:

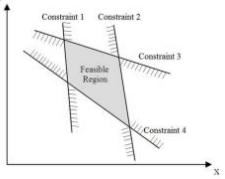
$$D(x) = sign(w^T x + b) \quad \text{(۵)}$$
$$= sign(\sum_{i\in S}\alpha_i y_i x_i^T x + b)$$

در رابطهٔ (۵)، S نشان‌دهندهٔ مجموعهٔ اندیس‌های بردارهای پشتیبان است. اگر نمونه‌ها در فضای ورودی جداپذیر خطی نباشند، نمونه‌ها با تکنیک حقهٔ هسته[۲] در فضای ویژگی با ابعاد بیشتر نگاشته می‌شوند. در این فضا، ماهیت خود داده ، اهمیتی ندارد بلکه آنچه مهم است فاصلهٔ داده ها از یکدیگر است. با این روش ماشین بردار پشتیبان یک ابرصفحهٔ جداکنندهٔ بهینه برای جداسازی نمونه‌ها پیدا می‌کند.

۲-۲- ماشین بردار پشتیبان با قیود نرم

مسألهٔ بهینه‌سازی (۶)، برنامه‌ریزی فازی نامتقارن نامیده می‌شود، اگر محدودیت‌هایش به‌صورت فازی بیان شوند. برای حل این مسأله روش‌هایی پیشنهاد شده است [۲] .

$$Optimize\ \ f(x)$$
$$S.t.\ \begin{cases} constraint\ 1 \\ constraint\ 2 \\ constraint\ 3 \\ constraint\ 4 \end{cases} \quad \text{(۶)}$$



شکل ۱: فضای شدنی برای مسألهٔ (۶)

در دسته‌بندی با قیدهای نرم، محدودیت‌های مسأله به فرم محدودیت فازی درمی‌آیند:

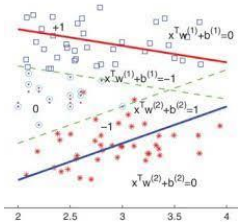
$$y_i(w^T x_i + b) \gtrsim 1 - \xi_i\,, \qquad i = 1,\dots,m. \quad \text{(۷)}$$

ماشین بردار پشتیبان با قیود نرم، انعطاف بیشتری نسبت به ماشین بردار پشتیبان عادی دارد؛ داده‌ها اجازهٔ تخطی از قیدها را دارند، با این راهکار، فضای شدنی مسألهٔ (۶) گسترش می‌یابد که در نتیجه تأثیر داده‌های نویزی کمتر می‌شود. توضیحات بیشتر در مرجع [۱] آمده‌اند.

۲-۳- ماشین بردار پشتیبان دوقلو

با گسترش ایدهٔ SVM در سال ۲۰۰۷ ، توسط جایادوا و همکاران ایدهٔ یافتن دوابرصفحهٔ غیرموازی، مطرح شد که هریک از آن‌ها تاحد ممکن به داده‌های یک طبقه نزدیک باشند و از طبقهٔ مقابل فاصلهٔ مطلوبی بگیرند. این الگوریتم‌ها به ماشین بردار پشتیبان دوقلو[۳] مشهورند. در این مسایل به جای حل یک مسألهٔ بهینه سازی درجه دوم با ابعاد بالا، دو مسألهٔ درجه دوم جدید با ابعاد تقریباً نصف ابعاد قبل(به شرط وجود تعادل بین داده‌های دو طبقه) و با مرتبهٔ محاسباتی $\frac{1}{4}$ روش قبل حل می‌شوند. این بالاترین مزیّت روش جدید به شمار می‌آید. گسترش‌های قابل توجّهی از ماشین بردار پشتیبان دوقلو مطرح [7] و کاربردهای متعدّدی از این روش ارائه شده است که از جملهٔ آن‌ها می‌توان به تشخیص رفتار انسانی [6] اشاره کرد.

شکل ۲: ماشین بردار پشتیبان دوقلو



در توضیح الگوریتم ماشین بردار پشتیبان دوقلو به‌فرض مجموعه‌ای از نمونه ها و اطلاعاتی درمورد آنها موجود است که به هر نمونه یک برچسب اختصاص داده‌شده و به کمک برچسب‌ها (۱ یا ۱−) میتوان کل نمونه ها را به دوگروه دسته‌بندی کرد. بهتر است که توزیع داده‌ها متعادل باشد یعنی با نسبت تقریباً مساوی، با برچسب ۱ یا۱− دسته‌بندی شده باشند. در روش بردار پشتیبان دوقلوبه جای یافتن دو ابرصفحهٔ موازی هم، ابرصفحه‌های جداساز، ما متقاطع اند. (مثلاً h_1 و h_2)

$$h_1 : x^T\omega_1 + b_1 = 0 \quad \text{(۸)}$$
$$h_2 : x^T\omega_2 + b_2 = 0$$

که در رابطه های فوق $\omega_1 \in R^n$ و ω_2 نمایشگر مختصات ابر صفحه b_1 و b_2 بایاس می‌باشند. h_1 تا حدّ امکان به کلاس ۱+ نزدیک و از کلاس ۱− دور است و h_2 تا حدّ ممکن به کلاس ۱− نزدیک و از کلاس ۱+ دور می‌باشد. داده‌های با برچسب منفی بایدفاصله‌ای مطلوب و حداقلی (مثلا یک واحد) از h_1 داشته باشند در غیراین‌صورت خطای دسته‌بندی ایجاد خواهد شد و لازم است که پارامترهای مشخص‌کنندهٔ جداساز طوری تنظیم شوند که داده‌هایی از این دست کمترخ دهند. مشابها در مورد داده‌های دستهٔ دوم هم چنین وضعیتی رخ خواهد داد. از این رو دو مسألهٔ مینی‌موم سازی مطرح خواهد شد و هریک از این دو مسأله 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 شده اند. تمام مباحث

گفته‌شده در مورد جداساز h_2 هم صدق خواهند کرد. با توجه به اینکه مسألهٔ بهینه‌سازی درجه دوم (QPP) به فرم استاندارد زیر است:

$$minimize \quad \left(\frac{1}{2}\right) x^T P x + q^T x \qquad (۹)$$

$$S.t. \quad \begin{cases} Gx \le h \\ Ax = b \end{cases}$$

دو مسألهٔ بهینه‌سازی به صورت زیر تعریف می‌شوند:

$$\min_{(\omega_1 , b_1)} \quad \frac{1}{2} \|A\omega_1 + e_1 b_1\|^2 + C_1 e_2^T q_1 \qquad (۱۰)$$
$$S.t. \quad -(B\omega_1 + e_2 b_1) + q_1 \ge e_2 , q_1 \ge 0$$

و

$$\min_{(\omega_1 , b_1)} \quad \frac{1}{2} \|B\omega_2 + e_2 b_2\|^2 + C_2 e_1^T q_2 \qquad (۱۱)$$
$$S.t. \quad (A\omega_2 + e_1 b_2) + q_2 \ge e_1 , q_2 \ge 0$$

که در آن ، $q_1 \in \mathcal{R}^{m_1}$، $q_2 \in \mathcal{R}^{m_2}$ وC_2 و C_1 پارامترهای خطا با مقادیر مثبت و سایر متغیرها قبلا تعریف شده‌اند. برای حل مسأله‌های فوق، بعد از نوشتن دوگان لاگرانژ و شرایط *مکمل زاید*[۴]، به دو مسألهٔ جدید زیر می‌رسیم:

$$\min_{\alpha} \quad \frac{1}{2} \alpha^T G (H^T H + \epsilon I)^{-1} G^T \alpha - e_2^T \alpha \qquad (۱۲)$$
$$S.t. \quad 0 \le \alpha \le C_1 e_2$$

و

$$\min_{\beta} \quad \frac{1}{2} \beta^t H (G^T G + \epsilon I)^{-1} H^T \beta - e_1^T \beta \qquad (۱۳)$$
$$S.t. \quad 0 \le \beta \le C_2 e_1$$

پس از بهینه‌سازی و یافتن بردارهای α وβ ، ابر صفحه‌های دو کلاس از دستورات زیر به دست می‌آیند:

$$\begin{bmatrix} \omega_1 \\ b_1 \end{bmatrix} = -(H^T H + \epsilon I)^{-1} G^T \alpha \qquad (۱۴)$$

$$\begin{bmatrix} \omega_2 \\ b_2 \end{bmatrix} = (G^T G + \epsilon I)^{-1} H^T \beta \qquad (۱۵)$$

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 قطر اصلی دوماتریس افزوده شده تا ماتریسها وارون‌پذیر باشند. در نهایت با حل مسأله فوق و یافتن ابرصفحات جداکننده، فاصلهٔ نمونه‌های جدید از هر ابر صفحه محاسبه شده و مینی‌موم فاصله، مشخص خواهد کرد که نمونهٔ ذکر شده به کدام دسته تعلق دارد.

در صورتیکه داده‌ها بصورت خطی قابل تفکیک نباشند، شبیه SVM از حقه‌ی هسته استفاده شده و ابرصفحه هایی با معادله‌های زیر تشکیل خواهند شد:

$$h_1 : K(x^T, C^t)\omega_1 + b_1 = 0 \qquad (۱۶)$$

$$h_2 : K(x^T, C^t)\omega_2 + b_2 = 0$$

توضیحات بیشتر در مورد داده‌هایی که به‌طور خطی جداناپذیرند، در مرجع [8]ذکر شده‌است.

۳- راه حلّ پیشنهادی: ماشین بردار پشتیبان دوقلو با قیود نرم[۵]

۳-۱- طبقه‌بند خطی

در این بررسی تلاش خواهدشد تا به کمک یک نوع ماشین بردار پشتیبان دوقلو با محدودیت‌های نرم، ابتلای فرد به بعضی از انواع بیماری‌های مزمن، پیش‌بینی شود. روش پیشنهادی، در مقایسه با ماشین بردار پشتیبان دوقلو، هنگام مواجهه با داده‌های نویزی عملکرد بهتری دارد.

در روش پیشنهادی سعی شده‌است تا با ایجاد تغییر در محدودیت‌های مسألهٔ بهینه سازی درجه دوم در ماشین بردار پشتیبان دوقلو، به‌جای نامساوی‌های ساده از نامساوی‌های فازی استفاده شود. مسأله‌های بهینه‌سازی جدید از قرار زیر هستند:

$$\min_{(\omega_1 , b_1)} \quad \frac{1}{2} \|A\omega_1 + e_1 b_1\|^2 + C_1 e_2^T q_1 \qquad (۱۷)$$
$$S.t. \quad -(B\omega_1 + e_2 b_1) + q_1 \gtrsim e_2 , q_1 \ge 0$$

و

$$\min_{(\omega_2 , b_2)} \quad \frac{1}{2} \|B\omega_2 + e_2 b_2\|^2 + C_2 e_1^T q_2 \qquad (۱۸)$$
$$S.t. \quad (A\omega_2 + e_1 b_2) + q_2 \gtrsim e_1 , q_2 \ge 0$$

علامت $\gtrsim$ به این معنی است که برای نمونه‌های یادگیری اجازه‌ی تخطی از حدود هم داده می‌شود. در واقع با این‌کار فضای شدنی مسأله گسترش می‌یابد و جواب‌های بهینه از فضای بزرگتری می‌توانند انتخاب شوند.

ابتدا می‌بایست دو مسألهٔ بهینه‌سازی (۱۸) و (۱۷) که مسأله‌های بهینه‌سازی غیرخطی درجهٔ دوم با تابع هدف محدب و قیود نامساوی فازی هستند، حل شوند. در گام اوّل برای حل مسأله، محدودیتها را از فرم ماتریسی خارج کرده و برای هر نمونه از هر طبقه ۱-، یک محدودیت به شکل (۱۹) در نظر گرفته می‌شود.

$$-(\omega_1{}^T x_i + b_1) \gtrsim 1 - q_{1_i} , \ q_{1_i} \ge 0 \qquad (۱۹)$$
$$i = 1, \dots, m_2.$$

که x_i، i_ اُمین نمونه از داده‌های با برچسب ۱- و q_{1_i} خطای مربوط به این نمونه در اثر تخطی از محدودیت نظیر و نزدیک‌شدن به h_1 می‌باشد. به‌طور مشابه برای نمونهٔ i_ام از طبقهٔ ۱+ محدودیت به شکل (۲۰) خواهد بود:

$$(\omega_2{}^T x_i + b_2) \gtrsim 1 - q_{2_i}, \ q_{2_i} \ge 0$$
$$i = 1, \dots, m_1. \qquad (۲۰)$$

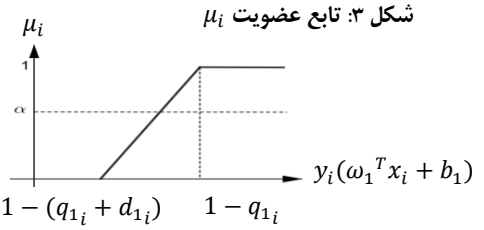
که x_i، i_ اُمین نمونه از داده‌های با برچسب ۱+ و q_{2_i} خطای مربوط به این نمونه در اثر تخطی از محدودیت نظیر و نزدیک‌شدن به h_2 می‌باشد. برای نمونه‌های با برچسب ۱+ طرفین محدودیت نظیر در (۱+) و برای نمونه‌های با برچسب ۱-، طرفین محدودیت نظیر در (۱-) ضرب شده و در نهایت محدودیت‌ها به شکل (۲۱) ادغام خواهندشد.

$$y_i(\omega_1{}^T x_i + b_1) \gtrsim 1 - q_{1_i}, \qquad (۲۱)$$
$$q_{1_i} \ge 0 , \ i = 1, \dots, m_2.$$

و $y_i(\omega_2{}' x_i + b_2) \gtrsim 1 - q_{2_i}$, $q_{2_i} \geq 0\,, i = 1, \dots, m_1.$

گام دوم تعیین توابع عضویت خطی $\mu_i\,, \mu'_i$ برای نامساوی‌های فازی موجود در قیود مسأله است. با توجه به شکل (۲)، در مورد اولین محدودیت در رابطهٔ (۲۱) و با تمرکز بر داده‌های کلاس ۱−، تابع عضویت μ_i مربوط به این طبقه تفصیل بررسی خواهد شد. μ'_i مشابه خواهد بود.

شکل ۳: تابع عضویت μ_i



طبق رابطهٔ (۲۱) هریک از نمونه‌های فوق، یک محدودیت به مسأله تحمیل می‌کنند. مجموعهٔ این محدودیت‌ها فضای شدنی را می‌سازد. متغیر−های مسأله یعنی $(\omega_1, b_1, q_1) \in R^{m_2+1+n}$ از این فضا انتخاب می‌شوند. هنگام حل مسأله این شرایط باید برآورده شوند. استفاده از نامساوی فازی به این محدودیت‌ها انعطاف بیشتری می‌بخشد. پارامترهای α و d_{1_i} توسط کاربر تعیین می‌شوند، پارامتر α ، سطحی را مشخص می‌کند که برای سطوح پایین‌تر از آن میزان تعلق فازی برای نامساوی موجود در قیود (۲۱)، صفر است. هر چه این مقدار به عدد ۱ نزدیک‌تر باشد، قیود مسأله به TWSVM نزدیک‌تر می‌شود. d_{1_i} 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 ۱− می‌توانند به ابرصفحهٔ جداساز کلاس ۱+ نزدیک‌تر شوند. این بدان معنی است که تعداد بردارهای پشتیبان بیشتر شده و داده‌های بیشتری در تعیین ابرصفحهٔ بهینه نقش خواهند داشت. تعریف می‌کنیم:

$$\mu_i: R^{m_2+1+n} \to (0,1],\quad i = 1,2,\dots,m_2. \qquad (۲۲)$$

$$\mu_i(\omega_1, b_1, q_1) = \begin{cases} 1, & if \quad y_i(\omega_1{}^T x_i + b_1) \geq 1 - q_{1_i} \\ \frac{y_i(\omega_1{}^T x_i + b_1) - 1 + q_{1_i} + d_{1_i}}{d_{1_i}} & if \quad 1 - (q_{1_i} + d_{1_i}) \leq y_i(\omega_1{}^T x_i + b_1) \leq 1 - q_{1_i} \\ 0, & if \quad y_i(\omega_1{}^T x_i + b_1) \leq 1 - (q_{1_i} + d_{1_i}) \end{cases}$$

و برای هر قید (۲۱) تعریف می‌کنیم:

$$X_i = \{(\omega_1, b_1, q_1) \in R^{m_2+1+n} \mid y_i(\omega_1{}^T x_i + b_1) \gtrsim 1 - q_{1_i}\,, q_{1_i} \geq 0\,,\quad i = 1, \dots, m_2\} \qquad (۲۳)$$

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

$$X = \cap_{i \in I} X_i\,,\quad I = \{1,2,\dots,m_2\} \qquad (۲۴)$$

بنابراین مسألهٔ *rtwsvm1* را می‌توان به‌صورت زیر نوشت:

$$Minimize\ \{Q(\omega_1, b_1, q_1) = \frac{1}{2}\|A\omega_1 + e_1 b_1\|^2 + C_1 e_2^T q_1 \mid (\omega_1, b_1, q_1) \in X\} \qquad (۲۵)$$

برای حل (۲۵) از روش برش آلفا[۶] استفاده می‌شود. این مسأله در شکل ۳ دیده می‌شود. جواب‌های (۲۵) باید از مجموعهٔ

$$X_\alpha = \{(\omega_1, b_1, q_1) \in R^{m+1+n} \mid \mu_X(\omega_1, b_1, q_1) \geq \alpha\} \qquad (۲۶)$$

انتخاب شوند که

$$\mu_X(x) = \inf\{\mu_i(x), i \in I\} \qquad (۲۷)$$

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

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

$$S(\alpha) = \{(\omega_1, b_1, q_1) \in R^{m_2+1+n} \mid \frac{1}{2}\|A\omega_1 + e_1 b_1\|^2 + C_1 e_2^T q_1\} = \{Min \frac{1}{2}\|A\omega'_1 + e_1 b'_1\|^2 + C_1 e_2^T q'_1\,, (\omega'_1, b'_1, q'_1) \in X_\alpha\} \qquad (۲۸)$$

پس می‌توان نوشت:

$$X_\alpha = \cap_{i \in I}\{(\omega_1, b_1, q_1) \in R^{m_2+1+n} \mid y_i(\omega_1{}^T x_i + b_1) \geq r_{1_i}(\alpha), q_{1_i} \geq 0\,, i = 1, \dots, m_2\} \qquad (۲۹)$$

که $r_{1_i}(\alpha) = 1 - q_{1_i} - d_{1_i}(1-\alpha)$.

زیرا:

$$\frac{y_i(\omega_1{}^T x_i + b_1) - 1 + q_{1_i} + d_{1_i}}{d_{1_i}} \geq \alpha \Rightarrow y_i(\omega_1{}^T x_i + b_1) \geq 1 - q_{1_i} - d_{1_i}(1-\alpha) \qquad (۳۰)$$

در نهایت مدل RTWSVM عبارت‌است از:

$$Minimize \quad \frac{1}{2}\|A\omega_1 + e_1 b_1\|^2 + C_1 e_2^T q_1 \qquad (۳۱)$$
$$S.t. -(B\omega_1 + e_2 b_1) \geq e_2 - q_1 - d_1(1-\alpha)$$
$$q_{1_i} \geq 0\,,\quad i = 1, \dots, m_2.$$

گام بعدی حل مسألهٔ بهینه‌سازی غیرخطی درجه دوم با قیدهای غیرفازی۷ است. تابع هدف اولّیه همراه با محدودیت‌هایش، به تابع لاگرانژ تبدیل می‌شوند:

$$Q(\omega_1, b_1, q_1, \beta_1, \gamma_1) = \frac{1}{2}\|A\omega_1 + e_1 b_1\|^2 + C_1 e_2^T q_1 - \beta_1{}^T\{-(B\omega_1 + e_2 b_1) + q_1 + d_1(1-\alpha) - e_2\} - \gamma_1{}^T q_1 \qquad (۳۲)$$

که در آن $\gamma_1 = (\gamma_{1_1}, \dots, \gamma_{1_{m_2}})^T$ ، $\beta_1 = (\beta_{1_1}, \dots, \beta_{1_{m_2}})^T$ ضرایب لاگرانژ و نامنفی می‌باشند.

$$\frac{\partial Q(\omega_1, b_1, q_1, \beta_1, \gamma_1)}{\partial \omega_1} = 0 \Rightarrow A^T(A\omega_1 + e_1 b_1) + B^T \beta_1 = 0 \qquad (۳۳)$$

$$\frac{\partial Q(\omega_1, b_1, q_1, \beta_1, \gamma_1)}{\partial b_1} = 0 \Rightarrow e_1{}^T(A\omega_1 + e_1 b_1) + e_2{}^T \beta_1 = 0 \qquad (۳۴)$$

$$\frac{\partial Q(\omega_1, b_1, q_1, \beta_1, \gamma_1)}{\partial q_1} = 0 \Rightarrow C_1 e_2^T - \beta_1{}^T -, \gamma_1{}^T = 0 \qquad (۳۵)$$

به‌علاوه باید شرایط مکمل زاید نیز برقرار شوند:

$$\beta_1{}^T\{-(B\omega_1 + e_2 b_1) + q_1 + d_1(1-\alpha) - e_2\} = 0 \quad , \gamma_1{}^T q_1 = 0 \qquad (۳۶)$$

از ترکیب روابط (۳۳) و (۳۴) :

$$U_1 = -(H^t H + \epsilon I)^{-1} G^t \beta_1 \quad (۳۷)$$

که در آن: $G = [B\ e_2], H = [A\ e_1]$ و $U_1 = [\omega_1\ b_1]^T$. با جایگذاری (۳۷) در (۳۲) و ساده کردن روابط و استفاده از رابطهٔ (۳۵)، مسألهٔ نهایی به فرم زیر خواهد بود:

$$Min \frac{1}{2} \beta_1{}^T G(H^T H + \epsilon I)^{-1} G^T \beta_1 - \beta_1{}^T \{d_1(1-\alpha) - e_2\} \quad (۳۸)$$

$$S.t. \qquad 0 \le \beta_1 \le C_1 e_2$$

مشابه تمام روابط (۲۲) تا (۳۶) برای داده‌های کلاس دیگر نیز برقرار است. با حل (۳۸) و به‌دست‌آمدن ضرایب لاگرانژ و مشخص شدن دو ابرصفحهٔ غیرموازی، تابع تصمیم (۳۹) مشخص خواهد کرد که نمونهٔ جدید به کدام طبقه تعلق می‌گیرد.

$$\underset{j=1,2}{argmin} \ |x^T \omega_j + b_j| \quad (۳۹)$$

۳-۲- طبقه‌بند غیرخطی

توسعه‌ی روش پیشنهادی RTWSVM برای حالت غیرخطی مشابه *SVM* استاندارد است. برای این حالت، داده‌های آموزش با استفاده از تابع φبه فضای ویژگی نگاشت می‌شوند. با تعریف تابع هستهٔ:

$$K(x_i, x_j) = \varphi^T(x_i)\varphi(x_j) = z_i . z_j$$

ابرسطح‌های جداکننده به فرم (۱۶) تبدیل می‌شوند. در فضای ویژگی مسألهٔ اولیه به صورت (۴۰) و (۴۱)

$$\underset{(\omega_1, b_1)}{min} \ \frac{1}{2} \|K(A, C^T)\omega_1 + e_1 b_1\|^2 + C_1 e_2^T q_1 \quad (۴۰)$$
$$S.t. \quad -(k(B, C^T)\omega_1 + e_2 b_1) + q_1 \gtrsim e_2, \quad q_1 \ge 0$$

$$\underset{(\omega_2, b_2)}{min} \ \frac{1}{2} \|K(B, C^T)\omega_2 + e_2 b_2\|^2 + C_2 e_1^T q_2 \quad (۴۱)$$
$$S.t. \quad (K(A, C^T)\omega_2 + e_1 b_2) + q_2 \gtrsim e_1, \quad q_2 \ge 0$$

که در آن $C = [A\ B]$ و K تابع هستهٔ دلخواه می‌باشد.

مسأله برای (۴۰) ادامه خواهد یافت. برای (۴۱) روش‌ها مشابه خواهد بود.

$$\underset{}{min} \ \frac{1}{2} \|K(A, C^T)\omega_1 + e_1 b_1\|^2 + C_1 e_2^T q_1 \quad (۴۲)$$
$$S.t. \quad -(K(B, C^T)\omega_1 + e_2 b_1) \ge e_2 - q_1 - d_1(1-\alpha)$$
$$q_{1_i} \ge 0, \quad i = 1, \dots, m_2.$$

و مسأله‌ی دوگان بصورت (۴۳) درمی‌آید.

$$Q(\omega_1, b_1, q_1, \beta_1, \gamma_1) = \frac{1}{2} \|K(A, C^T)\omega_1 + e_1 b_1\|^2 + C_1 e_2^T q_1 - \beta_1{}^T \{-(K(B, C^T)\omega_1 + e_2 b_1) + q_1 + d_1(1-\alpha) - e_2\} - \gamma_1{}^T q_1 \quad (۴۳)$$

پس از مشتق‌گیری و برقراری شرایط مکمل زاید:

$$U_1 = -(S^T S + \epsilon I)^{-1} R^T \beta_1 \quad (۴۴)$$

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

$U_1 = [\omega_1\ b_1]^T$ و

مشابه نسخهٔ خطی کلاس نمونهٔ جدید x با محاسبهٔ فاصلهٔ عمودی آن از دو ابرسطح مشخص می‌شود ؛ به طوری که تابع تصمیم برای نسخهٔ غیرخطی به‌صورت زیر تعریف می‌گردد:

$$\underset{j=1,2}{argmin} \ |K(x^T, C^T)\omega_j + b_j| \quad (۴۵)$$

۴- نتایج و ارزیابی

۴-۱- پایگاه داده

داده‌های مورد استفاده در این تحقیق، از پایگاه‌های دادهٔ Kaggle و UCI استخراج شده‌اند. به دلیل اهمیت موضوع چهار مجموعه دادهٔ :CKD، Hep، PIDD و WDBC که به ترتیب مرتبط با بیماری کلیه، هپاتیت، دیابت و سرطان سینه هستند، مورد بررسی قرارگرفته‌اند. داده‌های توصیفی از دست رفته با اولین مُد در ستون مزبور و داده‌های عددی با صفر جایگزین شده‌اند. جدول (۱) مجموعهٔ داده را توصیف می‌کند.

مجموعهٔ داده	تعداد نمونه‌ها	تعداد نمونه‌های مثبت	تعداد نمونه‌های منفی	تعداد ویژگی	دادهٔ از دست‌رفته
CKD	۴۰۰	۲۴۸	۱۵۲	۲۴	ندارد
Hep	۱۵۵	۳۲	۱۲۳	۱۹	دارد
PIDD	۷۶۸	۲۶۸	۵۰۰	۸	ندارد
WDBC	۵۶۹	۲۱۲	۳۵۷	۳۰	دارد

جدول ۱: مشخصات مجموعهٔ داده برای ارزیابی روش RTWSVM

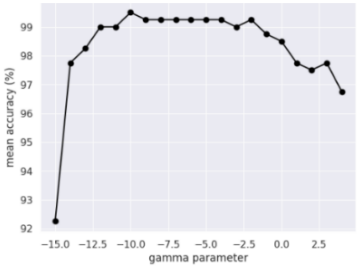
۴-۲- نحوهٔ پیاده‌سازی و اجرای الگوریتم‌ها

تمام روش‌ها در زبان برنامه‌نویسی پایتون در محیط آنلاین google colab پیاده‌سازی شده‌است. آزمایش‌ها روی یک کامپیوتر شخصی با پردازندهٔ core i5، سیستم عامل ویندوز ۱۰ و ۸ گیگابایت حافظه صورت گرفته است. از کتابخانهٔ cvxopt برای بهینه‌سازی استفاده شده است.

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

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\gamma^2}\right) \quad (۴۶)$$

دقت روش RTWSVM به انتخاب پارامترهای بهینه بسیار وابسته است. بدین منظور از روش جستجوی شبکه‌ای برای پیداکردن پارامتر بهینه استفاده شده است. پارامترهای خطا از مجموعهٔ $\{2^i | i = -7, -6, \dots, 7\}$ وپارامترهای α_1 ، α_2 ، d_1 و d_2که به ترتیب، برای سطح برش آلفا و ضریب تحمل الگوریتم، مورد استفاده بوده‌اند از مجموعهٔ [۰.۹۵ و ۰.۷۵ و ۰.۲۵] و پارامتر تابع هسته γ نیز از $\{2^i | i = -15, -6, \dots, 5\}$ انتخاب شده‌اند. برای تعیین میزان دقّت روش، از اعتبارسنجی متقابل استفاده شده‌است؛ الگوریتم ارائه شده با ۹۰٪ داده‌ها (k=10) آموزش دیده و با ۱۰٪ باقیمانده آزمایش شده، این مراحل ده بار تکرار شده و پس از ۱۰ بار تکرار، میانگین و انحراف معیار دقت روش پیشنهادی به‌دست آمده‌اند. شکل ۵ بیان‌گر تأثیر تغییرات پارامتر گاما بر میزان دقت روش پیشنهادی در پایگاه دادهٔ CKD است.



شکل ۴: اثر افزایش پارامتر γ روی دقت دسته‌بند *rtwsvm*

٤-۳- نتایج تجربی

روش پیشنهادی با SVM ساده، رگرسیون خطی، درخت تصمیم و نزدیک‌ترین همسایه، که از مرجع [12] به‌دست‌آمده، مقایسه شده‌اند.

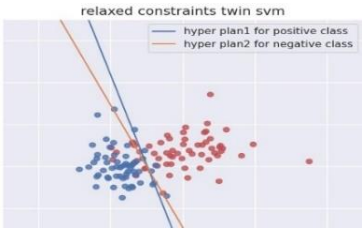
نتایج مقایسهٔ دقت و رتبهٔ الگوریتم پیشنهادی در جدول شماره ۲ و۳ قابل مشاهده اند؛ RTWSVM عملکرد بهتری دارد.

دقت	SVM	LR	DT	KNN	RTWSVM
CKD	۹۷.۸۵±۱.۵۵	۹۸.۷۵	۹۷.۵۰	۹۵.۰۰	۹۹.۵۰±۱
Hep	۸۰.۰۰	۸۳.۰۰	۸۳.۰۰	۸۰.۰۰	۱۰۰±۰
PIDD	۷۷.۹۲	۷۹.۸۷	۷۴.۶۸	۷۶.۶۲	۷۸.۸۱±۵.۵۹
WDBC	۹۷.۱۴	۹۷.۱۴	۹۵.۰۰	۹۷.۱۴	۹۸.۳۹±۱.۴۸
میانگین	۸۸.۲۲	۸۹.۶۹	۸۷.۵۴	۸۷.۱۹	**۹۴.۱۷**

جدول ۲: مقایسهٔ دقت دسته‌بندهای پیشین و روش پیشنهادی

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 ۳: مقایسهٔ رتبهٔ دسته‌بندهای پیشین و روش پیشنهادی



شکل ۵: نمایش هندسی بردار پشتیبان دوقلو با قیود نرم.

۵- نتیجه‌گیری و پژوهش‌های آینده

گسترش فضای شدنی، عدم تأثیر منفی بر تابع هدف، کاهش اثر داده‌های پرت، سرعت بیشتر و کارایی بالاتر از ویژگی‌های مثبت روش جدید است. بکارگیری الگوریتم‌های جستجو مانند الگوریتم‌های تکاملی، می‌تواند یافتن مقدار بهینه برای پارامترهای زیاد موجود را سرعت بخشد. نیز می‌توان با در نظرگرفتن یک ماتریس وزنی ضرایب تحمل، به هریک از نمونه‌ها توجه ویژه داشت. با این کار تأثیر داده‌های نویزی کاهش خواهد یافت.

مراجع

[۱] سبزه‌کار، مصطفی، **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 ۹۰-۸۲ ، آذر ۸۸.

[۲] ناجی عظیمی، زهرا، **آشنایی با برنامه‌ریزی خطی فازی،** ویراستهٔ وحیدیان کامیاد، علی، ویرایش اوّل، مشهد، انتشارات دانشگاه فردوسی مشهد، تابستان ۹۵.

[3] Alkenani, A. H., Li, Y., Xu, Y. & Zhang, Q. "Predicting Alzheimer's disease from spoken and written language using fusion-based stacked generalization", J. Biomed. Inform. 118, 103803, 2021.

[4] Guo, Y. et al. "A review of wearable and unobtrusive sensing technologies for chronic disease management", Comput. Biol. Med. 129, 104163, 2020.

[5] Higgins, V., Sohaei, D., Diamandis, E. P. & Prassas, I. "COVID-19: From an acute to chronic disease? Potential long-term health consequences", Crit. Rev. Clin. Lab. Sci. 58(5), 297–310, 2021.

[6] Mozafari, Kourosh, Jalal A. Nasiri, Nasrollah Moghadam Charkari, and Saeed Jalili. "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, 2011.

[7] Nasiri, Jalal A., and Amir M. Mir. "An enhanced KNN-based twin support vector machine with stable learning rules." *Neural computing and applications* 32, no. 16 : 12949-12969, 2020.

[8] Raschka S, Mirjalili V**.** Python machine learning second edition, Packt Publishing, BIRMINGHAM – MUMBA, 2017.

[9] Souza-Pereira, L., Pombo, N., Ouhbi, S., Felizardo, V. & Garcia, N. "Clinical decision support systems for chronic diseases: A systematic literature review", Comput. Methods Progr, Biomed, 195, 105565, 2020.

[10] Tanveer M, Rajani T, Rastogi R, Shao YH, Ganaie MA. "Comprehensive review on twin support vector machines", Annals of Operations Research, 8:1-46, 2022.

[11] Yuan, X., Chen, S., Yuwen, L., An, S., Mei, S. & Chen, T. "An improved SEIR model for reconstructing the dynamic transmission of COVID-19", In Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2320–2327 , 2020.

[12] Yuan X, Chen S, Sun C, Yuwen L. A novel early diagnostic framework for chronic diseases with class imbalance, Scientific Reports, 21;12(1):8614, 2022.

پانویس ها

۱ Input space

۲ Kernel Trick

۳ Twin Support Vector Machine (TWSVM)

۴ Karush–Kuhn–Tucker (KKT)

۵ Relaxed Constraints Twin Support Vector Machine (RTWSVM)

۶ $\alpha - cut$

۷ Crisp