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

پیش‌بینی بیماری‌های مزمن با داده‌های نامتوازن توسط ماشین بردار پشتیبان گرانشی

عبدالله محمدی[۱]، جلال‌الدین نصیری[۲] ، سهراب عفتی[۳]
[۱] دانشجوی کارشناسی ارشد علوم داده، ریاضیات کاربردی، دانشگاه فردوسی مشهد ، مشهد
abdhmohammadi@mail.um.ac.ir
[۲] استادیار، گروه ریاضی کاربردی، دانشکده ریاضی دانشگاه فردوسی مشهد، مشهد
jnasiri@um.ac.ir
[۳] استاد، گروه ریاضی کاربردی، دانشکده ریاضی دانشگاه فردوسی مشهد، مشهد
s-effati@um.ac.ir

چکیده

با پیشرفت تکنولوژی، روش‌های مبتنی بر داده برای تشخیص انواع بیماری‌ها، به طور گسترده‌ای مورد توجه قرار گرفته است. در طبقه بندی بیماری‌ها، تشخیص درست فرد "ناسالم" نسبت به تشخیص درست یک فرد سالم از اهمیت بیشتری برخودار است. اغلب داده‌های این بیماری‌ها دارای جامعه‌ی بیمار کوچک و جامعه‌ی سالم بزرگتری است. در این مقاله با تعریف یک تابع وزن ویژه در مدل وزنی الگوریتم twin svm[1] ، نشان داده می شود اختصاص وزن به گروه کوچکتر می تواند در تشخیص طبقه‌ی نمونه‌ها موثرتر باشد، ابتدا مفاهیم پایه‌ای مدل را بیان نموده سپس علاوه بر روال الگوریتم‌های دیگر، برای داده‌های کلاس کوچکتر نیز وزن اختصاص داده می شود. سپس از چندین مجموعه داده‌ی بیماری‌های مزمن مانند سرطان، دیابت و آلزایمر و ...، برای ارزیابی عملکرد روش استفاده نموده با مقایسه‌ی نتایج با چند روش دیگر، نشان داده می شود روش مورد استفاده می‌تواند با دقت بهتری نمونه‌ها را طبقه‌بندی کرده، نمونه‌های کلاس کوچکتر را نیز با دقت بالاتری تشخیص دهد بنابراین می‌توان انتظار داشت بتواند بر روش‌های دیگر برتری داشته باشد.

کلمات کلیدی

ماشین بردار پشتیبان، ماشین بردار پشتیبان دوقلو، مدل وزنی، وزن گرانشی، بیماری مزمن، دیابت، آلزایمر، سرطان

۱- مقدمه

در مطالعه‌ی داده‌های بیماری‌ها یکی از موضوعات با اهمیت، طبقه بندی آن‌هاست در رابطه با بیماری‌هایی مانند سرطان و دیابت، این طبقه‌بندی می‌تواند به صورت "سالم" و "ناسالم" بیان شود. با داشتن مجموعه‌ای از داده‌های ثبت شده، یکی ازروشهای پیش‌بینی طبقه‌ی نمونه‌ی جدید استفاده از ماشین‌های بردار پشتیبان (SVMs) است[1]. SVM یک تکنیک قدرتمند طبقه‌بندی است و سعی می‌کند داده‌ها را با استفاده از دو ابرصفحه‌ی موازی و ایجاد یک صفحه تصمیم‌گیری در بین آن‌ها از هم جدا کند. در بین نسخه‌های مختلف ارائه شده، ماشین‌های بردار پشتیبان دوقلو(Twin SVMs) که قید موازی بودن ابرصفحه ها را نادیده می‌گیرید[2] موفقیت بیشتری به دست آورده است. در کارهای نصیری و همکارن از لحاظ تئوری و هم به صورت کاربرد عملی در تشخیص حرکت انسان مورد استفاده قرار گرفته است[3],[4]. Xiong Si و Jing (2009) برای تشخیص توده در ماموگرافی دیجیتال استفاده کرده‌اند[5]. یکی از کاستی‌هایی که در svm استلندارد و twin svm وجود دارد این است که همه‌ی داده‌ها از اهمیت یکسان برخوردارند که در جهان واقعیت همیشه صادق نیست، برای غلبه بر این مشکل محققان بسیاری سعی کرده اند شیوه‌های مناسبی را ابداع کنند. برای طبقه بندی تومور Duan Hua و همکاران(2022) الگوریتمی با عنوان twin svm هیبریدی فازی را بکار برندند[6] که ترکیبی از مدل twin svm فازی است و با تولید یک ابرکره داده های مثبت و منفی را از هم جدا می‌کند. افزونه‌های بسیاری بر پایه‌ی اختصاص وزن به نمونه‌ها، برای این الگوریتم توسعه داده شده است که به الگوریتم‌های ماشین بردار پشتیبان دوقلوی وزنی(weighted Twin SVMs) مشهورند[7]. در این روش‌ها همچنان به نسبت بین کلاس‌ها اهمیت داده نمی‌شود در حالی که ممکن است داده هایی با نسبت‌های نامتوازن وجود داشته باشد مانند نسبت افراد بیمار به افراد سالم. اخیرا روشهایی برای کنترل نرخ عدم تعادل کلاس[2] ارائه شده است که با تعریف وزن به صورت نسبتی از این نرخ، سعی می‌کنند این مشکل را برطرف کنند اخیرا در یک مطالعه Xiaohan Yuan و دیگران(۲۰۲۲) نیز، یک چارچوب تشخیصی جدید برای بیماری‌های مزمن با داده‌های نامتوازن ارائه کردند[8]. از دیگر روش‌های ارائه شده در این زمینه، الگوریتم LSFLSTSVM-CIL است که در سال ۲۰۲۲ توسط M. A. Ganaie و همکاران توسعه داده شد[9]. این شیوه سعی می کند با استفاده از نرخ عدم تعادل کلاس، یک تابع وزن برای خطای هر نمونه از داده های کلاس بزرگتر تعریف کند در این روش و روش‌های دیگر، میزان اهمیت خطای کلاس کوچکتر برای همه نقاط داده یکسان در نظر گرفته می شود.

[1] Twin Support Vector machine

[2] Class Imbalance Rate

در این مقاله تابع وزن جدیدی بر پایه‌ی مفاهیم جاذبه در فیزیک بیان می‌شود، از این تابع در ضرایب خطای هر دو کلاس، برای به حداقل رساندن خطای تشخیص نادرست داده‌ها استفاده می‌شود، بنابراین اختصاص وزن به نمونه‌های کلاس کوچکتر باعث افزایش کارایی مدل می‌شود.

در ادامه، در بخش ۲ مفاهیم اساسی ماشین بردار پشتیبان دوقلو و وزنی معرفی می‌شود. در بخش ۳ روش مورد مطالعه و تابع وزن براساس مفهوم جاذبه‌ی گرانش بین دو جسم بیان می‌شود. در بخش ۴ به ارزیابی نتایج و مقایسه‌ی آن با روش‌های دیگر می‌پردازیم و از چندین مجموعه داده‌ی بیمای‌های مزمن برای آزمایش عملکرد آن استفاده می‌کنیم و نتایج بدست آمده را با چند روش دیگر مقایسه می شود. این مقایسه‌ها نشان می‌دهند که این روش می‌تواند با دقت بالاتری بیماری‌های مزمن را پیش‌بینی کند و انتظار می‌رود به روش‌های دیگر برتری داشته باشد.

۲- پژوهش‌های گذشته

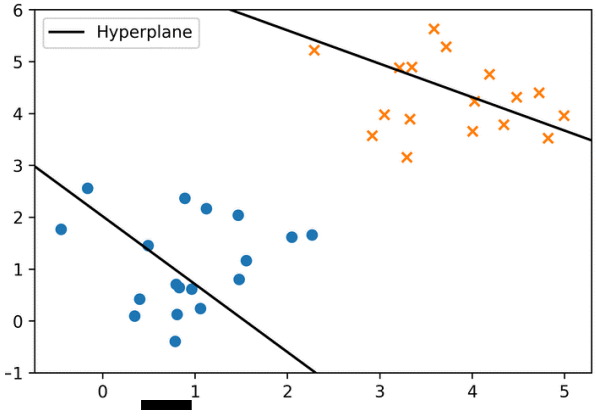
۲-۱- ماشین بردار پشتیبان دوقلو[1]

الگوریتم Twin SVM دو ابرصفحه غیرموازی را با حل دو مسئله‌ی برنامه ریزی درجه‌ی دو به صورت زیر بدست می آورد:

$$Min\ \frac{1}{2}\left(Aw^{(1)}+e_1b^{(1)}\right)^T\left(Aw^{(1)}+e_1b^{(1)}\right)+c_1e_2^Tq \quad (۱)$$
$$S.t.\ \ -\left(Bw^{(1)}+e_2b^{(1)}\right)+q\geq e_2,\ \ q\geq 0$$

$$Min\ \frac{1}{2}\left(Bw^{(2)}+e_2b^{(2)}\right)^T\left(Bw^{(2)}+e_2b^{(2)}\right)+c_2e_1^Tq \quad (۲)$$
$$S.t.\ \ \left(Aw^{(2)}+e_1b^{(2)}\right)+q\geq e_1,\ \ q\geq 0$$

که در روابط (۱) و (۲) A و B به ترتیب نمونه های کلاس مثبت و منفی هستند $b^{(i)}$ مقادیر اسکالر و $w^{(i)}$ بردارهای پارامترهای ابرصفحه ها، $c_i>0$ پارامترهای مدل، e_i بردارهایی با درایه های ۱ و q مقدار خطای ابرصفحه است. جمله‌ی اول در تابع هدف مجموع مربعات فاصله‌ی نقاط از ابرصفحه‌ی هر کلاس را پیاده سازی می کند و جمله ی دوم مجموع خطای متغیرها را کمینه می‌کند. و نمونه های جدید را براساس نزدیک بودن به ابرصفحه‌ها طبقه بندی می‌کند.



شکل ۱: ابرصفحه‌ی Twin SVM

۲-۲- ماشین بردار پشتیبان دوقلوی وزنی[2]

این مدل با تعریف وزن، اهمیت داده ها را در تعیین ابرصفحه ها در نظر می گیرد، با الهام از نظریه گراف ، یک گراف K نزدیکترین همسایه G را برای مدل سازی ساختار هندسی محلی داده‌ها ایجاد می کند. ماتریس وزن G را به صورت زیر تعریف می کند:

$$W_{ij}=\begin{cases}1 & x_i \text{ در } k-\text{همسایگی } x_j \text{ یا } x_j \text{ در } k-\text{همسایگی } x_i \text{ اگر}\\ 0 & \text{در غیر این صورت}\end{cases} \quad (۳)$$

براساس تعریف فوق برای مدل سازی فشردگی درون کلاسی و تفکیک پذیری بین طبقاتی دو ماتریس گراف برای هریک از جفت TWSVM می سازد، یک گراف درون کلاسی G_s و یک گراف بین کلاسی G_d. این دو زیرگرافهایی از G هستند.

$$W_{s,ij}=\begin{cases}1 & \text{اگر حداقل یکی از } x_i \text{ و } x_j \text{ در همسایگی دیگری در } G_s \text{ باشد}\\ 0 & \text{در غیر این صورت}\end{cases} \quad (۴)$$

$$W_{d,ij}=\begin{cases}1 & \text{اگر حداقل یکی از } x_i \text{ و } x_j \text{ در همسایگی دیگری در } G_d \text{ باشد}\\ 0 & \text{در غیر این صورت}\end{cases} \quad (۵)$$

ایده WLTSVM کشف اطلاعات شباهت ذاتی در نمونه‌های یک کلاس و استخراج بردارهای پشتیبان احتمالی موجود در نمونه‌های کلاس دیگر است. ماتریس وزن را با بازتعریف از روی روابط بالا به صورت زیر می سازد:

$$f_j=\begin{cases}1 & \text{اگر وجود داشته باشد } i \text{ که } W_{d,ij}\neq 0\\ 0 & \text{در غیر این صورت}\end{cases} \quad (۶)$$

مشابه TWSVM سعی می کند دو صفحه غیر موازی بدست آورد که هر کدام با نقاط کلاس مربوطه مطابقت دارند. به دنبال تفسیر هندسی مشابه، یک مسئله بهینه سازی برای تخمین ابرصفحه کلاس +۱ و −۱به صورت زیر بیان می شود:

$$\text{Min}\ \frac{1}{2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_1}W_{s,ij}(w_1^Tx_j^{(1)}+b_1)^2+C\sum_{j=1}^{N_2}\xi_j, \quad (۷)$$
$$\text{s.t.}\ -f_j(w_1^Tx_j^{(2)}+b_1)+\xi_j\geq f_j\cdot 1,\quad \xi_j\geq 0.$$

$$\text{Min}\ \frac{1}{2}\sum_{j=1}^{N_1}d_j(w_1^Tx_j^{(1)}+b_1)^2+C\sum_{j=1}^{N_2}\xi_j, \quad (۸)$$
$$\text{s.t.}\ -f_j(w_1^Tx_j^{(2)}+b_1)+\xi_j\geq f_j\cdot 1,\quad \xi_j\geq 0$$

که در روابط (۷) و (۸) عبارت $f_j.\mathbf{1}$ به معنی ضرب اسکالر $\boldsymbol{f_j}$ در برداری با درایه‌های ۱ است.

۳- روش پیشنهادی

الگوریتم LSFLSTSVM-CIL[3] مانند مدل پایه‌ی Twin SVM و WLTSVM دو ابرصفحه‌ی غیرموازی را برای جداسازی کلاس‌ها از یکدیگر جستجو می‌کند برای کاهش خطای کلاس بزرگتر نیز در تابع هدف، از یک تابع وزن [10] برای کاهش مجموع مربعات خطای آن کلاس استفاده می کند. در اینجا به بیان خلاصه‌ی از این الگوریتم می‌پردازیم. خواننده می‌توانند برای جزئیات بیشتر به [9] مراجعه کند.

[1] Twin Support Vector Machine (TSVM or Twin SVM)
[2] Weighted Twin Support Vector Machine (WTSVM)
[3] Large Scale Fuzzy Least Squares Twin SVM-Class Imbalance Learning

۳-۱- حالت خطی الگوریتم LSFLSTSVM-CIL

همانطور که گفته شـــد این الگوریتم دو ابرصـــفحه‌ی غیرموازی را برای جداسازی داده‌ها جستجو می‌کند که به صورت زیر داده می شود:

$$\begin{cases} w_1^T x + b_1 = 0 \\ w_2^T x + b_2 = 0 \end{cases} \quad (۹)$$

که در رابطه‌ی فوق $w_1, w_2 \in \mathbb{R}^n$ و $b_1, b_2 \in \mathbb{R}$ و x یک نمونه‌ی دلخواه از داده هاست، تابع هدف مسئله به صورت زیر تعریف می شود:

$$\min_{w_1,b_1,\xi_2,\eta_1} \frac{c_3}{2}(||w_1||^2 + b_1^2) + \frac{1}{2}\eta_1^T\eta_1 + \frac{c_1}{2}(S_2\xi_2)^T(S_2\xi_2)$$

$$\boldsymbol{S.t.}\ \ Aw_1 + eb_1 = \eta_1 \\ -(Bw_1 + eb_1) + \xi_2 = e \quad (۱۰)$$

$$\min_{w_2,b_2,\xi_1,\eta_2} \frac{c_4}{2}(||w_2||^2 + b_2^2) + \frac{1}{2}\eta_2^T\eta_2 + \frac{c_2}{2}(S_1\xi_1)^T(S_1\xi_1) \quad (۱۱)$$

$$\boldsymbol{S.t.}\ \ Bw_2 + eb_2 = \eta_2 \\ (Aw_2 + eb_2) + \xi_1 = e$$

که A و B به ترتیب نمونه‌های کلاس کوچکتر و بزرگتر، w_i, b_i پارامترهای ابرصـــفحه ، ξ_i متغیرهای کمکی e بردار متشـــکل از درایه‌های برابر ۱ ، c_i ابرپارامترهای مدل ، S_2 ماتریس قطری شـــامل وزن خطاهای کلاس بزرگتر و S_1 ماتریس همانی است.

مسئله های بهینه سازی فوق با استفاده از ضرایب لاگرانژ و اعمال شرایط K.K.T به صورت روابط (۱۲) و (۱۳) در می آید[9].

$$\max_{\alpha,\beta} -\frac{1}{2}(\alpha^t \ \ \beta^t)\,\hat{Q}\,(\alpha^t \ \ \beta^t)^t - c_3\beta^t e \quad (۱۲)$$

$$\hat{Q} = \begin{bmatrix} AA^t + c_3 I & AB^t \\ BA^t & BB^t + \frac{c_3}{c_1}(S_2^{-1})^2 \end{bmatrix} + E$$

$$\max_{\lambda,\theta} -\frac{1}{2}(\lambda^t \ \ \theta^t)\,Q'\,(\lambda^t \ \ \theta^t)^t - c_4\theta^t e \quad (۱۳)$$

$$Q' = \begin{bmatrix} BB^t + c_4 I & BA^t \\ AB^t & AA^t + \frac{c_4}{c_2}(S_1^{-1})^2 \end{bmatrix} + E$$

و بردارهای w_i و مقادیر b_i از رابطه‌ی (۱۴) بدست می‌آید که در این رابطه α و β مقادیر ضرایب لاگرانژ هستند.

$$\begin{bmatrix} w_1 \\ b_1 \end{bmatrix} = \frac{1}{c_3}\begin{bmatrix} A^t & B^t \\ e^t & e^t \end{bmatrix}\begin{bmatrix} \alpha \\ \beta \end{bmatrix} \quad (۱۴)$$

پس از حل مسئله‌ی فوق تابع تصمیم گیری برای پیش‌بینی طبقه‌ی نمونه‌ی جدید از رابطه‌ی زیر استفاده می شود:

$$class(x) = \underset{i=1,2}{argmin}(|x^T w_i + b_i|) \quad (۱۵)$$

۳-۲- حالت غیرخطی الگوریتم LSFLSTSVM-CIL

استفاده از توابع هسته[۱] در داده‌هایی که به صورت خطی قابل تفکیک نیستند یکی از ترفندهای مفید اســـت[12]. برای حالت غیر خطی با انتخاب یک هســـته‌ی مناســـب، ماتریس A با $K(A, A^T)$ و ماتریس B با $K(B, B^T)$ ماتریس AB^T با $K(A, B^T)$ و ماتریس BA^T با $K(B, A^T)$ جایگزین می شود[9]. با جایگزینی موارد ذکر شده تابع تصمیم به صورت زیر اصلاح می شود:

$$f_1(x) = \frac{1}{c_3}\left[K(x, A^t),\ K(x, B^t)\right]\begin{bmatrix}\alpha \\ \beta\end{bmatrix} + b_1 \quad (۱۶)$$

$$b_1 = \tfrac{1}{c_3}(e^t\alpha + e^t\beta)$$

$$f_2(x) = \frac{1}{c_4}\left[K(x, B^t),\ K(x, A^t)\right]\begin{bmatrix}\lambda \\ \theta\end{bmatrix} + b_2 \quad (۱۷)$$

$$b_2 = -\tfrac{1}{c_4}(e^t\lambda + e^t\theta)$$

$$class(x) = \underset{i=1,2}{argmin}(|f_i(x)|) \quad (۱۸)$$

و کلاس نمونه‌ی جدید مطابق رابطه‌ی (۱۸) پیش بینی می‌شود.

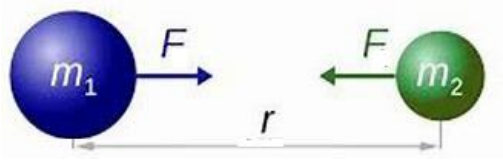
۳-۳- تابع وزن

در مدل‌های پیشـین چون مقادیر وزن به صـورت ضـرایب خطای نمونه های کلاس بزرگتر درنظر گرفته شـــده لند فقط منجر به کاهش خطای نمونه‌های کلاس بزرگتر می‌شـود و برای داده‌های کلاس کوچکتر وزنی اختصـاص داده نمی‌شود، در داده‌های بیماری‌های مزمن مثل سـرطان و دیابت و بیماری‌های ویروســـی همه‌گیر مانند COVID-19 اهمیت بســـیار زیادی وجود دارد که خطای خیلی کمتری در تشـــخیص کلاس کوچکتر اتفاق بیافتد در این داده‌ها تشخیص اشتباه یک نمونه به معنای این است که یک شخص ناسالم را سالم تشخیص دهیم. و این هزینه‌ی گزافی خواهد داشت. ما در اینجا ابتدا به اصلاح تابع وزن استفاده شده در مدل فوق اقدام کردیم و سپس با تخصیص وزن به داده‌های کلاس کوچکتر خطای این داده‌ها را کاهش دادیم.

۳-۴- تعریف تابع وزن

در فیزیک نیروی گرانش F بین دوجسم به جرم m_1 و m_2 با فاصله‌ی r از یکدیگر و با ثابت گرانش G به صورت زیر تعریف می شود:

$$F = G\frac{m_1 \times m_2}{r^2} \quad (۱۹)$$



شکل ۲- برهم کنش دو جرم در فیزیک

[۱] Kernel Functions

با بهره گیری از این مفهوم و اندکی تغییر در آن، برای نمونه‌ی x وزن خطای آن را به صورت زیر تعریف می کنیم:

$$weight(x) = G\frac{N_1 \times N_2}{r(x)+1} \qquad (۲۰)$$

که در رابطه‌ی فوق N_1 تعداد نمونه‌ها در همسایگی نمونه‌ی x به همراه خود آن، و N_2 نیز همین مفهوم برای مرکز کلاس به شعاع ε است ، در رابطه‌ی فوق $r(x)$ فاصله‌ی نمونه‌ی x از مرکز کلاس است. این رابطه برای محاسبه‌ی عناصر قطری ماتریس های S_1 و S_2 استفاده می‌شود، این درحالی است که در مرجع [9]، [10] و [12] فقط به ماتریس S_2 وزن داده می‌شود و S_1 یک ماتریس همانی در نظر گرفته شده است. مقدار IR برای نقاط کلاس کوچکتر برابر یک و برای نقاط کلاس بزرگتر به صورت رابطه‌ی زیر تعریف می‌شود:

$$IR = \frac{number\ of\ samples\ in\ class\ B}{number\ of\ samples\ in\ class\ A} \qquad (۲۱)$$

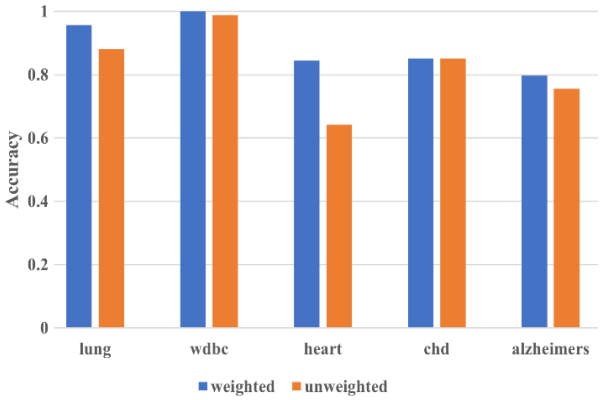
در اینجا نمونه‌ی x مانند یک جسم با جرم N_1 در نظر گرفته می‌شود که عضویتش در کلاس تحت تاثیر مرکز کلاس با جرم N_2 می باشد. و از این رو متناسب با وزنش در تعیین ابرصفحه نقش ایفا می‌کند. در رابطه‌ی (۲۰) اهمیت نمونه‌ی x با توجه به تعداد نقاط اطراف و همچنین فاصله‌ی آن از مرکز کلاس سنجیده می‌شود. هر چه تعداد نقاط اطراف یک نمونه بیشتر و به مرکز کلاس نزدیک‌تر باشد دارای وزن بیشتری در این کلاس خواهد بود، می‌دانیم فاصله‌ی یک داده‌ی پرت از مرکز کلاس زیاد است علاوه بر این به دلیل دورافتادگی، در اطراف خود نقاط بسیار اندکی دارد و یا یک نقطه‌ی کاملا تنهاست. با این توصیفات چنین نقطه‌ای وزن بسیار اندکی خواهد داشت و بنابراین از اهمیت کمتری برخوردار است.

جدول ۱: نسبت عدم تعادل کلاس

Dataset	Lung	wdbc	Heart	chd	Alzheimer
Imbalance ratio	7.0	1.6	1.23	5.52	2.33

۴- نتایج داده‌های آزمایشی

در این بخش عملکرد روش خود را روی چند نمونه از داده‌های بیمای‌های مزمن مورد ارزیابی قرار می‌دهیم، برای حالت غیرخطی از کرنل گوسی $K(x_1,x_2) = e^{-\mu||x_1-x_2||^2}$ و برای حالت خطی از کرنل خطی $K(x_1,x_2) = x_1.x_2^T$ استفاده کردیم. مقادیر پارامترها ازجدول ۱ و با استفاده از جستجوی تصادفی بدست آمده است، برای ارزیابی مدل از روابط (۲۲) تا (۲۶) استفاده شده است. در شکل ۳ عملکرد مدل در دو حالت مورد بررسی قرار گرفته است. یک بار به داده‌های هر دو کلاس وزن تخصیص داده شده و در حالت دوم وزن داده‌های کلاس کوچکتر برابر یک در نظر گرفته شده است. نسبت عدم تعادل این داده‌ها در جدول ۲ نشان داده شده است. در



شکل ۳: تفاوت دقت در دو حالت وزن دهی

شکل ۳ دیده می‌شود مدل مورد مطالعه در ۴ مورد از ۵ مورد نتایج بهتر و در یک مورد دارای نتیجه‌ی برابر بدست آورده است. می توان انتظار داشت وقتی به کلاس کوچکتر وزن داده شود نتایج بهتری کسب شود.

جدول ۲: پارامترهای مدل

پارامترهای مدل	محدوده‌ی مقادیر
c, c_0	$0.25 \times i\ ,\quad i: 1,\dots,10$
c_1, c_2, c_3, c_4, μ	$10^i, i: -5, \dots 5$
r	$\begin{cases} \frac{10^{-5}}{10^i} & i: -4,-3,-2,-1,0 \\ 0.1+\frac{i}{10} & i: 1,\dots,9 \end{cases}$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (۲۲)$$

$$Recall = \frac{TP}{TP+FN} \qquad (۲۳)$$

$$Precision = \frac{TP}{TP+FP} \qquad (۲۴)$$

$$F-score = \frac{2 \times Precision \times Recall}{Precisoin + Recall} \qquad (۲۵)$$

$$G-mean = \sqrt{Precision \times Recall} \qquad (۲۶)$$

جدول ۳ مقایسه‌ی بین نتایج این مدل با نتایج چند نسخه از توسعه‌های الگوریتم twin svm که به طور ویژه برای داده‌های نامتوازن طراحی شده اند را روی داده‌های سرطان و دیابت نشان می دهد. این داده‌ها در مخزن داده های keel[۱] و Kaggle[۲] در دسترس است. نتایج نشان می‌دهد در داده‌های سرطان، مدل مورد مطالعه به دقت ۱۰۰٪ دست یافته و در داده‌های دیابت رتبه‌ی دوم را دارد. در رابطه با یک بیماری همه‌گیر مانند COVID-19 و یا انواع سرطان، اگر شخص سالمی بیمار تشخیص داده شود می توان محتاطانه تدابیر لازم را اتخاذ کرد، اما اگر شخص بیماری سالم تشخیص داده شود ممکن است عواقب جبران ناپذیری داشته باشد در این

[۱] https://sci2s.ugr.es/keel/development.php

[۲] https://kaggle.com

جدول ۳: مقایسه‌ی دقت مدل مورد مطالعه در نمونه داده‌های Pima و Breast Cancer با ٤ مدل دیگر

Datasets	HFSWLSTSVM*	KWRUTSVM-CIL*	RFLSTSVM-CIL* (c_0, c_1, μ)	IFW-LSTSVM-CIL* (c_1, c_2, k, μ)	**This study** $(c, c_1, c_2, c_3, c_4, \mu, r)$
Breast Cancer	98.55	98.01	98.76 (1.5,0.1,16)	99.07 $(10^{-2}, 10^{-4}, 1, 16)$	**100** $(0.75, 2.25, 10, 10^{2}, 10^{-4}, 10^{-5}, 10^{-5}, 10^{-5})$
Pima Indian	**89.71**	71	62.77 (0.5,0.1,32)	76.77 (0.001,0.1,1,2)	79.22 $(1.25, 0.5, 10^{2}, 10^{2}, 10^{-4}, 10^{-3}, 0.1, 0.1)$

* این داده‌ها از [13]، [14] و [15] بدست آمده است.

مورد اهمیت معیار recall بسیار زیاد است، این معیار با توجه به رابطه‌ی (۱۵) به معنی دقت مدل در تشخیص درست شخص واقعا بیمار است.

جدول ۴ عملکرد مدل را با این معیار مورد ارزیابی قرار داده است. در این جدول مشاهده می شود مدل علاوه بر نتایج مطلوب معیار صحت(Accuracy)، هماهنگ با آن نتایج معیار Recall نیز بسیار مطلوب است. در مورد Lung با میزان دقت ٪۹۵.۷۰ معیار Recall نشان دهنده‌ی این است که مدل ٪۹۱.۶۷ از بیماران را به درستی تشخیص داده است. در رابطه با داده‌های heart و hepatitis مقدار recall حتی بهتر از accuracy است و در سایر موارد نیز بیشترین تفاوت بین این دو معیار در مورد آلزایمر تقریبا ٪۹ است. این مقایسه نشان می دهد روش مورد مطالعه در هر سه مورد نتایج بهتری نشان می دهد. از نظر مقدار recall در دیتاست Pima ضعیف بوده و در hepatitis با اختلاف تقریبا ٪۵ برتری دارد.

کوچکتر در بیماریهای مزمن بسیار بیشتر از گروه بزرگتر است توجه ویژه داشتیم تا بتوانیم با افزای دقت تشخیص صحیح در این بخش اعتبار مدل را بهبود ببخشیم، نتایج بررسی شده نشان داد که اعمال وزن به خطای گروه کوچکتر می تواند اهمیت این گروه را در تعیین طبقه بند صحیح منعکس کند با توجه به مطالب بررسی شده و نتایج موجود و مقایسه‌های انجام گرفته می‌توان انتظار داشت ماشین بردار پشتیبان دوقلوی با وزن گرانشی(روش مورد مطالعه) با اختصاص وزن به هر دو کلاس، در داده‌های نامتوازن عملکرد بهتری دارد. از طرفی با در نظر گرفتن مقدار نسبت عدم تعادل در گروه بزرگتر کنترل سوگیری مدل را مانند مدلهای ذکر شده حفظ می‌کند به ویژه در داده‌های بیماری‌ها که اغلب اهمیت کلاس کوچکتر از کلاس بزرگتر بیشتر است این مدل عملکرد بهتری نسبت به سایر مدل‌ها دارد.

۵- نتیجه گیری

در این مقاله در رابطه با داده‌های نامتوازن و اهمیت آن‌ها در الگوریتم‌های طبقه‌بندی برپایه‌ی twin svm بحث کردیم. به این نکته که اهمیت گروه

جدول ٤: عملکرد مدل و معیارهای ارزیابی

Dataset $(c, c_1, c_2, c_3, c_4, \mu, r)$	IR	Kernel	Accuracy	Recall	Precision	f-score	g-mean
Lung $(1,1.25,1,10^{3},10^{-5}, 10^{-3}, 10^{-4}, 10^{-5})$	7	rbf	**0.9569**	**0.9167**	0.7857	0.8461	0.8487
Breast Cancer $(0.75,2.25,10,10^{2}, 10^{-4}, 10^{-5}, 10^{-5}, 10^{-5})$	1.63	rbf	**1**	**1**	1	1	1
Pima $(1.25,0.5, 10^{2}, 10^{2}, 10^{-4}, 10^{-3}, 1,1)$	1.98	linear	0.7705	0.7159	0.6923	0.7039	0.7040
Ckd $(1,1.75, 10^{4}, 10^{5}, 10^{3}, 10^{-2}, 10^{-2}, 0.2)$	1.74	linear	**1**	**1**	1	1	1
Hepatitis $(0.5,1.25, 10^{2}, 10^{2}, 10^{5}, 10, 10^{-5}, 10^{-1}, 10^{-4})$	3.32	linear	**0.9149**	**1**	0.6364	0.7778	0.7977
Heart $(0.25,2,1,10^{2}, 10^{2}, 10^{-4}, 10^{-3}, 0.3)$	1.23	linear	**0.8406**	**0.8689**	0.7910	0.8290	0.7516
Alzheimer $(1.75,2, 10^{-2}, 10^{-1}, 1,0.1,0.1,0.1)$	2.33	rbf	0.7972	0.7025	0.6967	0.6996	0.6996

مراجع

[1] V. Vapnik ",*Support Vector Networks*", Nature of Statistical Learning Theory (ser. Statistics for Engineering and Information Science). New York, NY, USA: Springer, 2000.

[2] JA Nasiri, AM Mir," *An Enhanced KNN-based twin support vector machine with stable learning rules*", Neural computing and applications 32 (16), 12949-12969,2020.

[3] K Mozafari, JA Nasiri, NM Charkari, S Jalili, "*Informatics and Computational Intelligence (ICI), 2011.*

[4] Jayadeva, R. Khemchandani, and S. Chandra, "*Twin support vector machines for pattern classification*", IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 5, pp. 905–910, May 2007.

[5] Xiong Si, Lu Jing, "*Mass Detection in Digital Mammograms Using Twin Support Vector Machine-based CAD System*", WASE International Conference on Information Engineering, 2009.

[6] DUAN Hua1, FENG Tong1, LIU Songning1, ZHANG Yulin1 , and SU Jionglong, "*Tumor Classification of Gene Expression Data by Fuzzy Hybrid Twin SVM*", Chinese Journal of Electronics Vol.31, No.1, Jan. 2022.

[7] Qiaolin Yea, Chunxia Zhaoa, Shangbing Gaoa, Hao Zheng," *Weighted Twin Support Vector Machines with Local Information and its application*", Neural Networks, 31-39, 2012.

[8] XiaohanYuan, Shuyu Chen, Chuan Sun & LuYuwen, "*A novel early diagnostic framework for chronic diseases with class imbalance*", Scientifc Reports, 12:8614, 2022.

[9] M. A. Ganaie, M. Tanveer, Senior Member, IEEE, and Chin-Teng Lin, *"Large-Scale Fuzzy Least Squares Twin SVMs for Class Imbalance Learning",*IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 30, NO. 11, NOVEMBER 2022.

[10] B. Richhariya and M. Tanveer, "*A robust fuzzy least squares twin support vector machine for class imbalance learning*", Appl. Soft Comput., vol. 71 ,pp. 418–432, 2018.

[11] Mokhtar S. Bazaraa John J. Jarvis Hanif D ",*Linear Programming and Network Flows",* New York: Wiley,1977.

[12] Avrim Blum, John Hopcroft, and Ravindran Kannan, "*Foundations of Data Science*", Thursday 4th January, 2018.

[13] Divya Tomar and Sonali Agarwal, "*Hybrid Feature Selection Based Weighted Least Squares Twin Support Vector Machine Approach for Diagnosing Breast Cancer, Hepatitis, and Diabetes*", Hindawi Publishing Corporation Advances in Artificial Neural Systems Volume 2015.

[14] M.A. Ganaie, M. Tanveer, "*KNN weighted reduced universum twin SVM for class imbalance learning*",

[15] M. Tanveer, Senior Member, IEEE, M. A. Ganaie, A. Bhattacharjee, and C. T. Lin, "*Intuitionistic Fuzzy Weighted Least Squares Twin SVMs*", IEEE.