دوره دهم، شماره دوم، پاییز و زمستان ۱۳۹۶

مجله فناوری اطلاعات در طراحی مهندسی
Information Technology in Engineering Design
http://ited.sinaweb.net

طبقه‌بندی داده‌های نامتعادل با ترکیب منحنی اصلی و SMOTE بهبود یافته درجه پشتیبان

ریحانه کمال پور* (۱) عادل قاضی خانی (۲)

(۱) گروه مهندسی کامپیوتر، دانشگاه بین‌المللی امام رضا (ع)، مشهد، ایران*
(۲) گروه مهندسی کامپیوتر، دانشگاه بین‌المللی امام رضا (ع)، مشهد، ایران

چکیده:

مسئله طبقه‌بندی داده‌های نامتعادل به‌عنوان یکی از چالش‌های اصلی در حوزه داده‌کاوی است. در مجموعه داده‌های نامتعادل، تعداد نمونه‌ها در کلاس‌های گوناگون اختلاف بسیاری دارند. در داده‌های نامتعادل هدف اصلی شناسایی صحیح نمونه‌های کلاس اقلیت است. به‌عنوان مثال در حوزه‌ی‌پزشکی، تعداد نمونه‌های مثبت از یک بیماری در مقابل تعداد نمونه‌های منفی بسیار کمتر است. در حالی‌که اهمیت شناسایی نمونه‌های مربوط به دسته مثبت، بسیار زیاد است. در این مقاله الگوریتمی بانام SdCurveSmote، برای این مسئله ارائه شده است. این الگوریتم شامل سه بخش کلی است بخش اول محاسبه پارامتری بنام درجه پشتیبانی برای انتخاب نمونه‌های مناسب مرزی برای عمل Smote و بخش دوم انجام عمل Smote بر روی این نمونه‌های مناسب با درجه پشتیبان بالا و تولید نمونه جدید است و در بخش آخر محاسبات منحنی اصلی وحدآستانه و بررسی نمونه‌های تولید شده می‌باشد. در این مقاله از ۵ مجموعه داده نامتعادل، برگرفته از سایت UCI و KEEL استفاده شده است که میزان عدم تعادل متفاوت دارند. الگوریتم پیشنهادی نسبت به الگوریتم‌های مشابه، با آزمون فریدمن نتایج بهتری داشت در این آزمون برای طبقه بند Adaboost نشان می‌دهد که در آن تفاوت معنادار بسیار مطلوبی بین روش پیشنهادی و روش‌های Smote و SdSmote وجود دارد.

واژه‌های کلیدی: طبقه‌بندی، داده‌های نامتعادل، نمونه افزایی، منحنی اصلی، حد آستانه عضویت

*عهده دار مکاتبات

نشانی: گروه مهندسی کامپیوتر، دانشگاه بین‌المللی امام رضا (ع)، مشهد، ایران

تلفن: ۰۹۱۵۵۱۸۱۳۲۶ **پست الکترونیکی:** r.kamalpour@imamreza.ac.ir

۱- مقدمه

تکنیک‌های یادگیری ماشین در بسیاری از حوزه‌های جهان واقعی کاربرد دارند مانند اینترنت، مطالعات تجاری و علمی، کاربردهای صنعتی و غیره [۱]. مسئله داده‌های نامتعادل زمانی رخ می‌دهد که نمونه‌های یک یا چند کلاس ذاتاً نادرند و یا به‌سختی جمع‌آوری می‌شوند [۲]؛ بنابراین مسئله کلاس‌های نامتعادل بسیار حائز اهمیت است زیرا به طور ضمنی در اکثر کاربردهای واقعی مشاهده می‌شود مانند تشخیص کلاه‌برداری، مدیریت ریسک، تحقیقات پزشکی [۳]، بازاریابی پایگاه داده [٤]، تشخیص تولد زودرس [٥]، داده‌های میکرو آرایه [٦] و غیره.

در مجموعه داده‌های نامتعادل، کلاس مثبت یا اقلیت به کلاسی گفته می‌شود که معمولاً از دیدگاه یادگیری بیشترین علاقه و توجه به آن است ودر صورتی که نادرست طبقه‌بندی شود، منجر به هزینه بیشتری می‌شود و کلاس با تعداد بیشتر داده را کلاس منفی یا اکثریت می‌نامند.

توزیع کلاس‌ها، یعنی نسبت نمونه‌های متعلق به هر کلاس در یک مجموعه داده، نقش کلیدی در طراحی طبقه‌بند ایفا می‌کند، در الگوریتم‌های استاندارد طبقه‌بندی داده‌ها، توزیع کلاس‌ها، متعادل در نظر گرفته می‌شود. ازاین‌رو در صورت استفاده از این الگوریتم‌ها در طبقه‌بندی داده‌های نامتعادل، نمی‌توان به نتایج قابل‌قبولی دست یافت؛ زیرا این الگوریتم‌های طبقه‌بند به سمت نمونه‌های آموزش کلاس بزرگ‌تر متمایل می‌شوند [۷] که این موضوع سبب افزایش تعداد خطاها در شناسایی نمونه‌های اقلیت می‌گردد. الگوریتم‌های یادگیری ماشین کلاسیک، اغلب میزان صحت بالایی برای داده‌های اکثریت به دست می‌آورند درصورتی‌که برای داده‌های اقلیت، خلاف آن است [۸].

۲- مروری بر پژوهش‌های پیشین

با توجه به اهمیت این مسئله، روش‌های زیادی در سطح الگوریتم وجود دارند که با اصلاح الگوریتم‌های یادگیری موجود، تلاش می‌کند تا فرآیند یادگیری را به سمت کلاس اقلیت سوق دهد. یکی از این روش‌ها، با نمونه‌برداری از فضای داده باعث تغییر توزیع داده‌ها می‌شود، به‌طوری‌که تغییری در الگوریتم یادگیری انجام نمی‌شود و تلاش می‌کند در مرحله پیش‌پردازش تأثیرات ناشی از عدم تعادل را برطرف کند.

نمونه‌برداری به دو شکل، نمونه زدایی از کلاس اکثریت و نمونه افزایی از کلاس اقلیت و یا ترکیبی از هر دو روش است [۹].

ساده‌ترین روش نمونه‌زدایی، نمونه‌زدایی به‌طور تصادفی است که در آن نمونه‌هایی از کلاس اکثریت به‌صورت تصادفی حذف می‌شود تا زمانی که کلاس اقلیت درصدی از کلاس اکثریت شود، به این ترتیب تعادل در مجموعه آموزشی برقرار می‌شود. از معایب این روش، از دست دادن برخی اطلاعات باارزش است و موجب زیر برازش داده‌های کلاس اکثریت می‌شود [۱۰].

روش دیگر خوشه‌بندی کاهشی است که نوعی روش نمونه زدایی است که در آن با استفاده از روش خوشه‌بندی کاهشی، نمونه‌های اکثریت را به تعدادی خوشه تقسیم و سپس با رتبه‌بندی نمونه‌های هر خوشه، تعدادی نمونه انتخاب کرده و این نمونه‌های اکثریت انتخابی را به همراه نمونه‌های اقلیت به‌عنوان مجموعه داده متعادل شده در نظر می‌گیرد [۱۱].

نمونه‌افزایی یکی دیگر از روش‌ها است که در آن برای به تعادل رساندن توزیع کلاس، از روش جایگزینی نمونه‌های کلاس اقلیت استفاده شده که نیازی به اطلاعات اضافی ندارد و از داده‌های موجود دوباره استفاده می‌کند تا تعادل در مجموعه آموزشی برقرار شود. از معایب این روش، افزایش اندازه مجموعه آموزشی است که این امر منجر به افزایش زمان یادگیری طبقه‌بند می‌شود.

روش Smote یکی دیگر از روش‌های نمونه افزایی است که داده‌های مصنوعی را بر اساس شباهت ویژگی‌های بین نمونه‌های کلاس اقلیت ایجاد می‌کند. در این روش برای زیرمجموعه $S_{min} \in S$، بر اساس فاصله اقلیدسی در فضای nبعدی K تا از نزدیک‌ترین همسایه نمونه $x_i \in S_{min}$ را انتخاب می‌کند. برای ایجاد نمونه مصنوعی، یکی از k همسایه‌های نزدیک نمونه به طور تصادفی انتخاب و سپس

با ضرب اختلاف این دو نمونه در یک عدد تصادفی بین [۱، ۰] و اضافه کردن نتیجه حاصل شده به نمونه x_i، نمونه جدید از رابطه (۱) به دست می‌آید که نمونه جدید بر روی خط متصل‌کننده بین دو نمونه قرار دارد [۱۲].

$$X_{new} = x_i + (x_j - x_i) * \delta \qquad (۱)$$

روش MSMoteنیز نوعی روش نمونه‌افزایی بهبود یافته است که با محاسبه مرکز نمونه‌های اقلیت و تعیین فاصله اقلیدسی هر یک از نمونه‌ها نسبت به مرکز نمونه‌ها، آن را در سه گروه امن، مرزی و نمونه‌های نویزدار مخفی طبقه‌بندی می‌کند. در ایجاد نمونه مصنوعی از داده‌های امن، یک نمونه از بین همسایه‌های نزدیک آن به طور تصادفی انتخاب می‌کند و برای نمونه‌های مرزی نیز نزدیک‌ترین همسایه را انتخاب می‌کند و برای داده‌های نویزی هیچ اقدامی انجام نمی‌پذیرد [۱۳].

روش MwMote روشی جدید بر پایه Smote است که ابتدا نمونه‌هایی از اقلیت را که در میان همسایگان اکثریت می‌باشد و هیچ همسایه اقلیتی ندارد را به عنوان نویز شناسایی و حذف می‌کند تا در ایجاد داده‌های مصنوعی نقشی نداشته باشند. پس از آن برای هر یک از نمونه‌های اقلیت، مجموعه‌ای از همسایگان اکثریت ایجاد می‌کند و برای نمونه‌های اکثریت نیز مجموعه‌ای از همسایگان اقلیت ایجاد می‌کند و برای هر یک از نمونه‌های موجود در این مجموعه‌ها یک وزن انتخاب و آن را به احتمال تبدیل می‌کند و از آن برای تولید بهتر نمونه‌های مصنوعی استفاده می‌کند [۱٤].

روش SdSmoteنیز روش جدید دیگری از نمونه‌افزایی است، از آنجایی که برخی از نمونه‌های اقلیت بسیار آسان برای یادگیری هستند درنتیجه تمامی نمونه‌های اقلیت برای تولید نمونه مصنوعی جدید مناسب نیستند، پس لازم است نمونه‌هایی انتخاب شوند که به آسانی آموزش داده نمی‌شوند. این نمونه‌ها معمولاً نزدیک مرز تصمیم‌گیری قرار دارند که بیشتر الگوریتم‌ها به دنبال این نمونه‌ها برای ایجاد داده مصنوعی هستند. در این الگوریتم، با تعریف پارامتری بنام درجه پشتیبان، سعی در انتخاب نمونه‌های مرزی برای ایجاد نمونه‌های مصنوعی دارد که برای محاسبه این پارامتر از مفهوم فاصله بین نمونه‌ها استفاده می‌نماید.

برای محاسبه درجه پشتیبانی، با فرض اینکه تعداد داده‌های کلاس مثبت m و تعداد داده‌های کلاس منفی n باشند، هریک از نمونه‌های کلاس مثبت را مانند x_i درنظرگرفته و مطابق رابطه (۲)، مجموع فاصله این نمونه را از تمام نمونه‌های کلاس منفی، محاسبه‌می‌کند و پس‌ازآن مجموع S_i های مرحله قبل را محاسبه می‌کند.

$$s_i = \sum_{j=1}^{n} \sqrt{\| x_i - x_j \|^2} \qquad (۲)$$

$$s = \sum_{i=0}^{m} s_i$$

پس از محاسبه مجموع فاصله تمامی داده‌های مثبت تا تمامی داده‌های منفی، میانگین فاصله‌ها را از رابطه (۳) محاسبه می‌کند.

$$s_{ave} = \frac{s}{m * n} \qquad (۳)$$

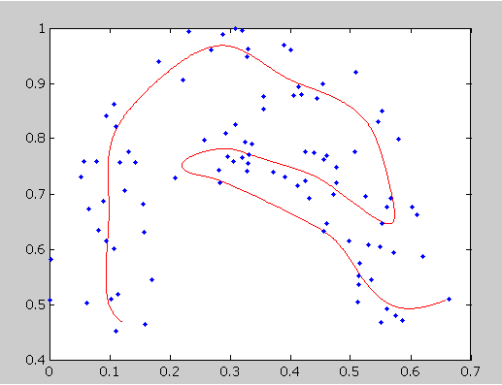
در این مرحله S_{ave} را به عنوان یک پارامتر فاصله در نظر گرفته و به ازای تمامی نمونه‌های مثبت، تک‌تک داده‌های مثبت را به عنوان مرکز دایره در نظر گرفته و دایره‌ای فرضی به شعاع S_{ave} رسم می‌کند و تعداد نمونه‌های کلاس منفی که در این دایره قرار دارند را شمرده و آن را به عنوان درجه پشتیبان نمونه مثبت منتخب در نظر می‌گیرد. این پارامتر محاسبه شده را به عنوان میزان احتمال انتخاب نمونه مثبت برای انجام عمل نمونه افزایی در نظر می‌گیرد [۱۵].

این پارامتر باعث شناخت نمونه‌های مرزی شده و همان‌طور که روش Smote مشکل over fitting نمونه‌افزایی را حل می‌کند، این روش نیز تا حدی مشکل blindness روش نمونه‌افزایی را کاهش می‌دهد.

علاوه بر روش‌های ذکر شده، مفهومی نیز به نام منحنی اصلی در سال ۱۹۸۳ برای اولین بار توسط هاستی بیان گردید [۱٦]. پس از آن، این تئوری در حل برخی مسائل مانند تجسم

داده‌ها و تحلیل اکولوژی مورد استفاده قرار گرفت. همچنین این منحنی در اصلاح مکان آهنربا در برخورد دهنده خطی استنفورد، پردازش تصویر و مدل‌سازی خطوط یخ، آنالیز دست‌نوشته‌ها و دستگاه‌های پاور، تشخیص لهجه، کنترل فرآیند و...مورد استفاده قرار گرفته است. به‌طورکلی منحنی اصلی یک منحنی هموار است که از وسط یک مجموعه داده p بعدی عبور می‌کند و یک توصیف غیرخطی از داده‌هاست. این منحنی غیر پارامتریک است و شکل آن توسط داده‌ها ایجاد می‌گردد. تفاوت میان منحنی اصلی و رگرسیون در این است که منحنی اصلی یک روش غیر پارامتریک برای توضیح مسیر داده‌ها، بدون توجه به رابطه میان آن‌ها است. با گذشت زمان انواع مختلفی از این نوع منحنی‌ها، توسط محققین مختلف پایه‌گذاری گردید که می‌توان منحنی‌های اصلی HS، TBR، و K رانام برد [۱۷] در تمامی این منحنی‌ها، منحنی مانند یک اسکلت در میان داده‌ها قرارداد و داده‌ها مانند یک ابر اطراف آن را پوشانده‌اند. در این میان منحنی اصلی K برای یک مجموعه داده X که جز R^d باشد، یک منحنی مانند f^* با طول L است، اگر f^* بتواند Δf را برای همه منحنی‌هایی با طول L و یا کمتر مینیمم کند. به طوری که f یک تابع پیوسته از I به R^d باشد و Δf نیز فاصله مربع مورد نظر بین X و f باشد که از رابطه (٤) به دست می‌آید. نمونه‌ای از این منحنی در شکل ۱ آمده است.

$$\Delta(f) = E[\Delta(x,f)] = E[\inf_{\lambda} \| x - f(\lambda) \|^2] = E\left[\| x - f(\lambda_f(x)) \|^2\right] \quad (٤)$$
$$\lambda_f(x) = \sup\{\lambda := \| x - f(\lambda) \| = \inf_{\tau} \| x - f(\tau) \|\}$$



شکل ۱- نمایی از منحنی اصلی

در روشی دیگر از این منحنی استفاده شده است که در آن از پارامتری به نام حد آستانه عضویت اقلیت‌ها استفاده شده است که برای محاسبه این پارامتر، با توجه به رابطه (۵)، ابتدا فاصله پروجکشن تمامی نمونه‌های اقلیت تا منحنی اصلی محاسبه می‌گردد پس‌ازآن به وسیله تقسیم این فواصل بر ماکزیمم فاصله، نرمال‌سازی انجام و میانگین این فاصله‌ها را به عنوان حد آستانه عضویت داده‌های اقلیت در نظر می‌گیرد. پس‌ازآن از این پارامتر برای بررسی داده‌های مصنوعی جدید استفاده می‌کند [۱۸].

$$d_i = \| w - f(\lambda_f(w)) \|_2^2 \quad (٥)$$
$$b_i = d_i$$
$$B = \max(b_1, b_2, \dots, b_N)$$
$$\mu_i = \frac{b_i}{B}$$
$$\sigma_x = \text{mean}(\mu_i)$$

گاهی استفاده ترکیبی از دو روش نمونه‌افزایی و نمونه‌زدایی برای متعادل نمودن نمونه‌ها [۱۹] نیز رایج است و تحقیقات نشان داده است که ترکیب روش‌ها با هم و یا ترکیب روش‌های نمونه‌برداری با روش‌های فازی و یا الگوریتم‌های تکاملی در مقابل روش‌های تک مانند نمونه‌افزایی ونمونه‌زدایی نتایج بهتری دارند [۲۰]. البته در مجموعه داده‌های متفاوت استفاده از هر یک از روش‌ها به تنهایی و یا ترکیب آن‌ها باعث ایجاد بهبود و یا عدم بهبود در عملکرد طبقه‌بندی می‌شوند [۲۱].

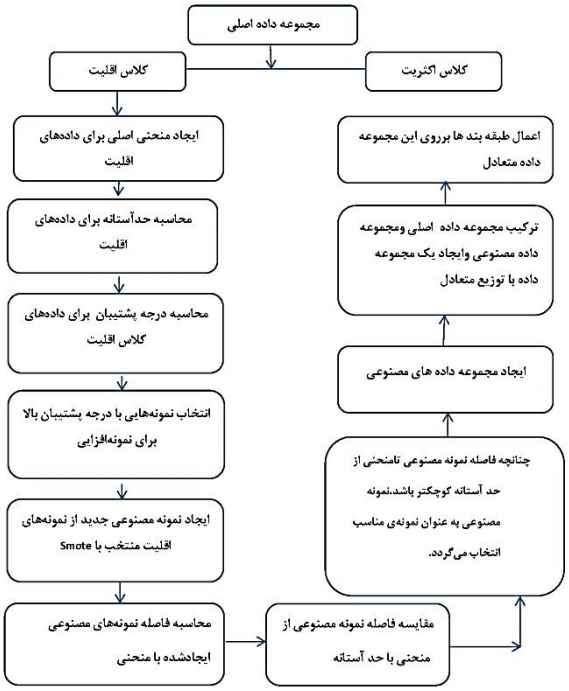
پس از اعمال هر یک از روش‌های ذکر شده بر روی داده‌های نامتعادل و ایجاد یک مجموعه داده نسبتاً متعادل و یا کاملاً متعادل، لازم است که این داده‌ها توسط طبقه‌بندهای مناسب، طبقه‌بندی شده تا بتوانند مورد تحلیل قرار گیرند.

به دلیل اهمیت بالای داده‌های کلاس اقلیت و در جهت طبقه‌بندی صحیح آن‌ها، با وجود انجام تحقیقات بسیار، هنوز مواردی مانند عدم توجه به اهمیت نمونه‌ها در متعادل‌سازی و عدم توجه به نمونه‌های مصنوعی جدید، تعیین تعداد مناسب طبقه‌بندها، عدم بهینه‌سازی وزن طبقه‌بندها در ترکیب طبقه بندها و...در این زمینه وجود دارد و در این مقاله تلاش

شده است رهیافتی برای چالش مطرح شده‌ی اول ارائه شود. در این مقاله یک الگوریتم ترکیبی برای طبقه‌بندی داده‌های نامتعادل ارائه شده که در آن با محاسبه پارامتر درجه پشتیبان و نمونه‌افزایی نمونه‌های مناسب، به بررسی نمونه‌های مصنوعی تولید شده پرداخته و پس از متعادل نمودن مجموعه داده، آن را مورد طبقه‌بندی قرار داده است. در ادامه، الگوریتم پیشنهادی و ارزیابی کارایی آن و در نهایت نتیجه‌گیری آمده است.

۳- الگوریتم پیشنهادی

در الگوریتم پیشنهادی، برای طبقه‌بندی داده‌های نامتعادل، ابتدا سعی بر انتخاب نمونه‌های مناسب جهت نمونه‌افزایی نموده و پس از انجام عمل نمونه‌افزایی به بررسی نمونه‌های جدید تولید شده توسط مفهومی به نام منحنی اصلی می‌پردازد و در این بررسی، نمونه‌های جدید حفظ و یا حذف می‌شوند و پس از ایجاد یک مجموعه داده متعادل، بدین ترتیب طبقه‌بندی با صحت بیشتر و خطای کمتر انجام می‌شود. الگوریتم پیشنهادی که در ادامه بانام SdCurveSmote از آن یاد شده است، ترکیبی از روش Smote بهبود یافته درجه پشتیبان با عنوان SdSmote [۱۵]، با منحنی اصلی است [۱۸]. در روش Smote بهبود یافته درجه پشتیبان، بر اساس درجه پشتیبان که برای هر داده اقلیت بدست می‌آید، نمونه‌های مناسب جهت عمل Smote انتخاب و نمونه‌های مصنوعی ایجاد می‌گردد. در روش پیشنهادی پس از ایجاد داده‌های مصنوعی، این داده جدید نیز توسط مفهوم منحنی اصلی مورد بررسی قرار می‌گیرد و با ترکیب این دو روش بهبود عملکرد طبقه‌بندی مشاهده می‌گردد. در شکل ۲ فلوچارت الگوریتم آمده است.



شکل ۲- چهارچوب روش پیشنهادی

الگوریتم پیشنهادی شامل ۱۰ مرحله اصلی است که به ترتیب در ذیل آمده است:

مرحله اول: داده‌های کلاس مثبت و کلاس منفی از میان کل داده‌ها انتخاب و به دو مجموعه داده مجزا تقسیم می‌شوند. در این مرحله با توجه به کلاس نمونه‌ها، دو مجموعه داده مثبت و منفی ایجاد می‌گردند که مجموعه داده مثبت شامل داده‌های اقلیت و مجموعه داده منفی نیز شامل داده‌های اکثریت می‌باشند.

مرحله دوم: منحنی اصلی مربوط به داده‌های مثبت ایجاد می‌گردد که این منحنی نشان‌دهنده توزیع داده‌های کلاس مثبت است و از آن برای بررسی داده‌های تولید شده جدید استفاده می‌گردد.

مرحله سوم: پس از ایجاد منحنی اصلی در جهت پیش برد کار لازم است که در این مرحله، فاصله داده‌های مجموعه داده مثبت تا منحنی محاسبه و پس از نرمال‌سازی فواصل و به کمک میانگین آن‌ها بر اساس رابطه (۵)، پارامتر حد آستانه عضویت داده‌های مثبت محاسبه گردد.

مرحله چهارم: برای انتخاب داده‌های مناسب جهت انجام عمل نمونه‌افزایی Smote، نیاز است که پارامتری به نام درجه

پشتیبان برای تک‌تک داده‌های مثبت محاسبه گردد که بر اساس روابط (۲) و (۳) در ابتدای کار فاصله میان هر یک ازنمونه‌های مثبت با تمامی نمونه‌های منفی محاسبه و پس از آن مجموع این فواصل و میانگین آن‌ها حساب می‌گردد. از این میانگین به‌عنوان شعاع دایره فرضی به مرکز هر یک از داده‌های مثبت استفاده نموده و تعداد داده‌های منفی موجود در این دایره فرضی را شمارش کرده و آن را به عنوان درجه پشتیبان نمونه مثبت در نظر می‌گیرد و هر نمونه مثبتی که درجه پشتیبان بالاتری داشته باشد، احتمال انتخاب آن نمونه برای عمل Smote بیشتر است بدین ترتیب داده‌های مناسب و مرزی برای عمل Smote مشخص می‌شوند.

مرحله پنجم: با توجه به میزان عدم تعادل، داده‌های مثبت با درجه پشتیبان بالا جهت انجام عمل نمونه‌افزایی انتخاب می‌شوند.

مرحله ششم: با اعمال Smote بر روی‌داده‌های منتخب، نمونه جدید را برای به تعادل رساندن مجموعه داده، ایجاد می‌کنیم.

مرحله هفتم: بااینکه داده‌های مناسب را برای نمونه‌افزایی انتخاب شده‌اند و عمل Smote نیز داده‌ای بین داده مثبت اصلی و همسایگانش ایجاد می‌کند اما لازم است که این داده جدید باز هم مورد بررسی قرار گیرد تا ازنظر توزیع داده‌ای در محدوده داده‌های اصلی باشد.

برای این کار لازم است تا فاصله تصویر نمونه جدید نسبت به منحنی اصلی محاسبه گردد تا موقعیت نمونه جدید نسبت به منحنی مشخص گردد.

مرحله هشتم: فاصله محاسبه‌شده در مرحله قبل را با حد آستانه عضویت داده‌های اقلیت مقایسه کرده، چنانچه فاصله نمونه‌ی جدید با منحنی از این حد آستانه کمتر باشد، نشان‌دهنده این است که نمونه جدید در محدوده منحنی است و یک داده مناسب برای متعادل نمودن مجموعه داده است در غیر این صورت حذف می‌گردد.

مرحله نهم: با توجه به نرخ عدم تعادل و تعداد داده‌های لازم برای تولید، سه مرحله ٦ و ٧ و ٨ تکرار می‌شوند. پس از تولید تعداد کافی نمونه‌های جدید، این داده‌ها را به مجموعه داده اصلی اضافه نموده و به یک مجموعه داده متعادل برای انجام عمل طبقه‌بندی دست می‌یابد.

مرحله دهم: پس از ایجاد مجموعه داده متعادل، لازم است با اعمال طبقه بندهای مناسب، به بررسی میزان عملکرد آن‌ها و تحلیل روش پیشنهادی بپردازد.

a) ارزیابی کارایی روش پیشنهادی

۱-۱-۱- مجموعه داده‌ها

در این مقاله، از ۵ مجموعه داده از مخزن مجموعه داده KEEL و UCI[۲۳ و ۲۲] با نرخ عدم توازن متفاوت استفاده شده است که جدول ۱، ویژگی‌های مجموعه داده‌های نامتعادل استفاده شده را نمایش می‌دهد.

جدول ۱- مشخصات مجموعه داده‌ها

ردیف	نام مجموعه داده	تعداد نمونه‌ها	تعداد ویژگی‌ها	تعداد داده‌های اقلیت	تعداد داده‌های اکثریت	درصد عدم توازن
1	Abalone19	4174	8	32	4142	129.44
2	Blood	748	5	178	570	3.2
3	JM1	10885	22	2106	8779	4.17
4	Winequality-red-4	1599	11	53	1546	29.17
5	Yeast1	1484	8	429	1055	2.46

در جدول ۱، برای هر مجموعه داده، تعداد نمونه‌ها، تعداد ویژگی‌ها و نسبت تعداد نمونه‌های کلاس اکثریت به تعداد نمونه‌های کلاس اقلیت (نرخ عدم تعادل) نشان داده شده است. برای بررسی کارایی یک طبقه‌بند لازم است که بر اساس معیاری آن را مورد ارزیابی قرار داده تا به میزان عملکرد آن دست پیدا کنیم.

۱-۱-۲- ارزیابی عملکرد

در این بخش‌بر اساس روابط (٦) و (۷) و (۸) و (۹) از سه معیار، اندازه‌گیری اف و میانگین هندسی و سطح زیر منحنی، به ارزیابی روش پیشنهادی ومقایسه آن با دو روش دیگر می‌پردازد.

$$F-measure=\frac{2*precision*recall}{precision+recall} \quad (٦)$$

$$precision=\frac{TP}{(TP+FP)} \quad (۷)$$

$$recall=\frac{TP}{TP+FN}$$

$$G-mean=\sqrt{precision*recall} \quad (۸)$$

$$AUC=\frac{1+Tprate-Fprate}{2} \quad (۹)$$

برای انجام این مقایسه، ابتدا مجموعه داده‌های نامتعادل، هر کدام ۱۰ بار در محیط متلب ۲۰۱٤، توسط سه الگوریتم Smote SdCurveSmote [۱۲],SdSmote [۱۵] متعادل شده و پس‌ازآن در محیط وکا ۳,۷,۹ به کمک سه‌طبقه بند C4.5,Bagging,Adaboost مورد طبقه‌بندی قرارگرفته و بدین ترتیب میانگین نتایج حاصله از معیارهای ارزیابی با هم مقایسه گردیدند.

نتایج در جدول ۲ معیارهای اندازه‌گیری اف، میانگین هندسی و سطح زیر منحنی را نشان می‌دهند.

جدول ۲- نتایج معیارها

F-measure	Blood	Jm1	Yeast1	Abalone19	Winequality-red-4
Smote+C4.5	71.37	82.59	77.06	97.75	93.73
SDSmote+C4.5	78.2	85	77.59	98.82	95.41
SDCurveSmote+C4.5	79.76	85.24	78.17	98.82	95.41
Smote+Bagging	75.21	84.64	80.89	98	95.22
SDSmote+Bagging	80.39	86.07	81.06	98.97	96.25
SDCurveSmote+ Bagging	81.66	86.16	81.02	98.96	96
Smote+Adaboost	71.78	66.49	72.94	84.56	73.68
SDSmote+Adaboost	73.99	71.58	69.94	92.5	80.59
SDCurveSmote+ Adaboost	74.56	82.28	70.51	92.51	84.56

G-mean	Blood	Jm1	Yeast1	Abalone 19	Winequality-red-4
Smote+C4.5	72.42	82.23	77.01	97.75	93.67
SDSmote+C4.5	79.26	84.94	76.26	98.82	95.38
SDCurveSmote+C4.5	80.46	85.17	77	98.82	95.39
Smote+Bagging	75.54	84.53	80.57	98	95.18
SDSmote+Bagging	81.09	86.2	80.49	98.7	96.23
SDCurveSmote+ Bagging	82.3	86.31	80.45	98.6	95.99
Smote+Adaboost	69.9	67.23	73.18	84.1	73.05
SDSmote+Adaboost	73.55	69.44	71.14	92.71	79.65
SDCurveSmote+ Adaboost	74.98	69.91	71.75	92.73	84.02

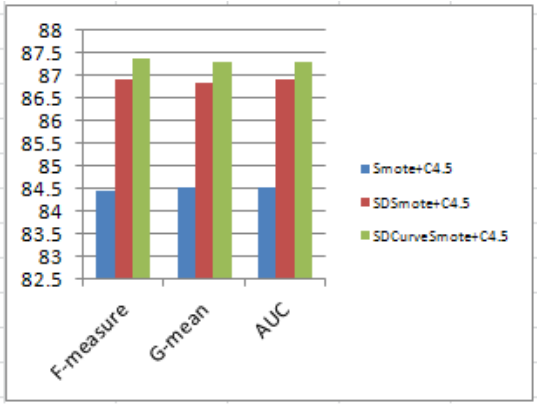
AUC	Blood	Jm1	Yeast1	Abalone19	Winequality -red-4
Smote+C4.5	72.44	82.25	77.01	97.76	93.69
SDSmote+C4.5	79.3	84.94	76.44	98.82	95.38
SDCurveSmote+C4.5	80.49	85.17	77.15	98.82	95.39
Smote+Bagging	75.54	84.53	80.58	98.01	95.19
SDSmote+Bagging	81.11	86.2	80.52	98.97	96.23
SDCurveSmote+Bagging	82.32	86.31	80.45	98.96	96
Smote+Adaboost	69.99	67.23	73.18	84.38	73.07
SDSmote+Adaboost	73.56	69.64	71.16	92.95	79.81
SDCurveSmote+Adaboost	74.99	70.09	71.77	92.97	84.79

مقایسه‌ها نشان می‌دهد که روش پیشنهادی برای اکثر داده‌ها نتایج مطلوب‌تری نسبت به دو روش دیگر دارد. میانگین نتایج معیارها نیز در جدول ۳ آمده است که بهبود کلی طبقه‌بندی را با روش پیشنهادی نشان می‌دهد.

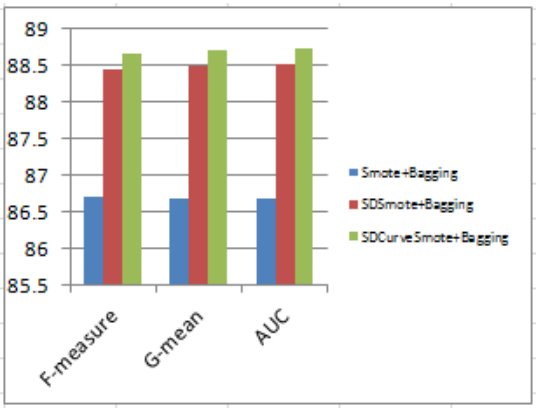
جدول ۳- میانگین معیارها

	F-measure	G-mean	AUC
Smote+C4.5	84.43	84.52	84.54
SDSmote+C4.5	86.93	86.85	86.9
SDCurveSmote+C4.5	87.39	87.28	87.31
Smote+Bagging	86.71	86.68	86.68
SDSmote+Bagging	88.45	88.5	88.51
SDCurveSmote+Bagging	88.66	88.71	88.72
Smote+Adaboost	73.6	73.25	73.33
SDSmote+Adaboost	77.48	77	77.12
SDCurveSmote+Adaboost	78.63	78.34	78.47

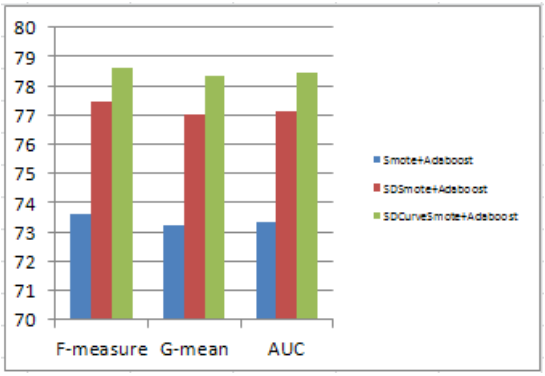
همان‌طور که در نمودارهای موجود در شکل‌های ۳ و ٤ و ٥ نیز مشاهده می‌گردد میانگین نتایج هر یک از معیارها، بهبود روش پیشنهادی را نشان می‌دهد.



شکل ۳- مقایسه روش‌ها با طبقه بند c4.5



شکل ٤- مقایسه روش‌ها با طبقه بند Bagging



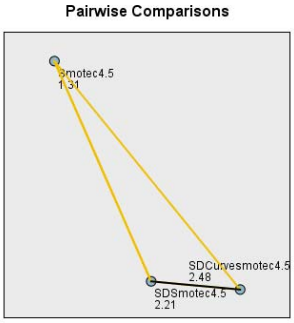
شکل ٥- مقایسه روش‌ها با طبقه بند Adaboost

برای بررسی دقیق‌تر نتایج نیز از آزمون آماری فریدمن استفاده شده است.

۱-۱-۳- آزمون فریدمن

با توجه به آنالیز انجام گرفته در نرم‌افزار SPSS، معیارهای به‌دست‌آمده دارای توزیع نرمال نبوده و بهترین آزمون آماری برای تحلیل این معیارها روش فریدمن است. از میان معیارها، AUC برای آزمون فریدمن انتخاب گردید.

همان‌طور که نتایج آزمون در شکل ٦ نشان می‌دهد، در طبقه بند C4.5، روش Smote با روش SdSmote تفاوت معنادار بسیار مطلوبی دارد، همچنین روش Smote با SdCurveSmote نیز تفاوت معناداری دارد.



Each node shows the sample average rank.

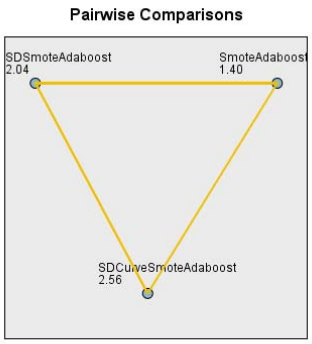
Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
Smotec4.5-SDSmotec4.5	-.904	.196	-4.609	.000	.000
Smotec4.5-SDCurvesmotec4.5	-1.173	.196	-5.982	.000	.000
SDSmotec4.5-SDCurvesmotec4.5	-.269	.196	-1.373	.170	.509

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

شکل ٦- نتایج آزمون فریدمن- معیار AUC- طبقه بند C4.5

در شکل ۷ نیز آزمون فریدمن را برای طبقه بند Adaboost نشان می‌دهد که در آن تفاوت معنادار بسیار مطلوبی بین روش‌های Smote و SdSmote همچنین بین Smote و SdCurveSmote وجود دارد.

علاوه بر این دو تفاوت معنادار، تفاوت معناداری نیز بین دو روش SdSmote و SdCurveSmote وجود دارد.



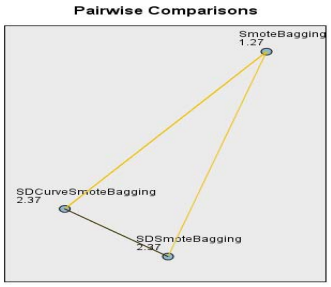
Each node shows the sample average rank.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
SmoteAdaboost-SDSmoteAdaboost	-.635	.196	-3.236	.001	.004
SmoteAdaboost-SDCurveSmoteAdaboost	-1.154	.196	-5.883	.000	.000
SDSmoteAdaboost-SDCurveSmoteAdaboost	-.519	.196	-2.648	.008	.024

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

شکل ۷- نتایج آزمون فریدمن- معیار AUC- طبقه بند Adaboost

در تست بعدی با آزمون فریدمن، شکل ۸ معیار AUC مربوط به طبقه بند Bagging آمده است که در آن مانند طبقه بند C4.5 تفاوت معناداری بین Smote و SdSmote، همچنین بین Smote و SdCurveSmote وجود دارد.



Each node shows the sample average rank.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
SmoteBagging-SDCurveSmoteBagging	-1.096	.196	-5.589	.000	.000
SmoteBagging-SDSmoteBagging	-1.096	.196	-5.589	.000	.000
SDCurveSmoteBagging-SDSmoteBagging	.000	.196	.000	1.000	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

شکل ۸- نتایج آزمون فریدمن-معیار AUC-طبقه بند C4.5

٤- نتیجه‌گیری و کارهای آتی

در این مقاله مسئله طبقه‌بندی داده‌های نامتعادل موردبررسی قرارگرفته و ایده اصلی کاهش عدم تعادل در مجموعه داده و بهبود طبقه‌بندی است. الگوریتم پیشنهادی بانام sdcurvesmote از طریق پارامتر درجه پشتیبان، داده‌هایی مناسب که احتمال مرزی بودن آن‌ها بیشتر است را انتخاب نموده و به کمک عمل Smote، داده‌های مصنوعی را ایجاد کرده و در مرحله بعدی، بررسی نمونه‌های جدید ایجادشده، توسط مفهوم منحنی اصلی است تا بتوان کنترلی نیز بر روی داده‌های مصنوعی داشته باشد به طوری که این داده‌ها از محدوده داده‌های اقلیت اصلی خارج نشوند وچنانچه با چنین داده‌ای مواجه گردد آن را حذف در غیر این صورت آن را به مجموعه داده اصلی اضافه می‌کند. پس از ایجاد یک مجموعه داده متعادل آن را با سه روش طبقه‌بندی C4.5,Bagging,Adaboost و روش ارزیابی 10-fold-cross-validation کلاس‌بندی نموده و به کمک سه معیار اندازه‌گیری اف، میانگین هندسی و سطح زیر منحنی مورد ارزیابی قرار داده است.

پس از اجرای طبقه‌بندها، سه معیار مورد بررسی قرار گرفته و میانگین نتایج نشان‌دهنده بهبود عملکرد الگوریتم پیشنهادی است. همچنین توسط تست آماری فریدمن نتایج الگوریتم‌ها مورد بررسی قرار گرفتند و تفاوت معناداری بین الگوریتم پیشنهادی و روش‌های دیگر مشاهده گردید.

در مطالعات آتی می‌توان کنار انجام عمل Smote جهت نمونه افزایی، با تعریف پارامتری مناسب برای داده‌های اکثریت از روش‌های نمونه‌زدایی نیز استفاده کرد. همچنین می‌توان بر روی مجموعه داده‌های نامتعادل با ابعاد ویژگی بالا [۲٤] نیز تحقیقاتی انجام داد و از آنجایی که برخی صفات تأثیر خاصی بر روی طبقه‌بندی ندارند علاوه بر متعادل‌سازی این داده‌ها به بررسی ویژگی‌های آن نیز پرداخت. در مورد داده‌های بزرگ نیز که مجموعه‌ای از داده‌های بزرگ و پیچیده هستند که بارگذاری آن‌ها بر روی کامپیوتر مشکل می‌باشد، چالش اصلی، تجزیه و تحلیل آن‌ها زمانی است که به صورت نامتعادل می‌باشند [۲۵]، علاوه براین موارد داده‌هایی با چگالی بالا [۲٦] نیز می‌توانند مواردی برای توسعه تحقیق باشند.

۵- مراجع

1-H. He, and E. A. Garcia, “Learning from imbalanced data, IEEE Transactions on Knowledge and Data Engineering”, vol. 21, 1263-1284, 2009.

2- A. Fernández, M. J. del Jesus, and F. Herrera, “On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced data-sets”, Expert Systems with Applications, vol. 36, 9805-9812, 2009.

3- Y.H. Lee, P.J.H. Hu, T.H. Cheng, T.C. Huang, and W.Y. Chuang. “A preclustering-based ensemble learning technique for acute appendicitis diagnoses”. Artificial Intelligence in Medicine, 58(2):115–124, 2013.

4- E. Duman, Y. Ekinci, and A. Tanrıverdi. “Comparing alternative classifiers for database marketing: The case of imbalanced datasets”, Expert Systems with Applications, 39(1):48–53, 2012.

5- D. P. Williams, V. Myers, and M. S. Silvious. “Mine classification with imbalanced data”, IEEE on Geoscience and Remote Sensing Letters, vol. 6, 528-532, 2009.

6- H. Yu, J. Ni, and J. Zhao. ACOSampling: “An ant colony optimizationbased undersampling method for classifying imbalanced dna microarray data”, Neurocomputing, 101:309–318, 2013.

7- R.Dubay, J. Zhou,Y.Wang, P.Thompson, J.ye,”Analysis of sampling techniques for imbalanced data: An n = 648 ADNI study”, Elsevier,vol 87,220-241,2014.

8- Enislay Ramentol, Sarah Vluymans, Nele Verbiest, Yail´e Caballero, Rafael Bello, Chris Cornelis, and Francisco Herrera, “IFROWANN: Imbalanced Fuzzy-Rough Ordered Weighted Average Nearest Neighbor Classification”, IEEE Transactions on Fuzzy Systems, 1063-6706 (c) 2013

9- M. Alibeigi, S. Hashemi, A. Hamzeh, DBFS: “an effective density based feature selection scheme for small sample size and high dimensional imbalanced data sets”, Data Knowl.Eng.81–82 67–103, 2016.

10- S-J Yen and Y-S Lee, “chuster-based undersampling Elsevier for imbalanced data distribution”, Expert Systems with Applications.Vol 36,5718-5727, 2009.

۱۱- محبوبه مهدی زاده، مهدی افتخاری،"ارائه روش جدید نمونه‌گیری مبتنی بر رتبه‌بندی نمونه‌ها و خوشه‌بندی کاهشی"، کنفرانس مهندسی کامپیوتر و توسعه پایدار، ایران، ۱۳۹۲

12- Chawla NV, Bowyer K, Hall L, Kegelmeyer W, "SMOTE: synthetic minority over-sampling technique", J Artif Intell Res, 16:341-378, 2002.

13- Hu, Shengguo, Liang, Yanfeng, Ma, Lintao, and He, Ying, “Msmote: improving classification performance when training data is imbalanced”, in Computer Science and

Engineering, 2009. WCSE'09. Second International Workshop on,vol. 2, pp. 13–17. IEEE, 2009.
14- Sukarna Barua, Md. Monirul Islam, Xin Yao, Fellow, and Kazuyuki Murase, "MWMOTE—Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning", IEEE Transactions on Knowledge and Data Engineering, VOL. 26, NO. 2, 2014,
15- Kewen Li, Wenrong Zhang, Qinghua Lu, Xianghua Fang, "An Improved SMOTE Imbalanced Data Classification Method Based on Support Degree", International Conference on Identification, Information and Knowledge in the Internet of Things, 2014.
16- J.Zhang, l. Wang, "An overview of principal curves", Chinese Journal of Computers, 26(2):129-146, 2003.
17- Dewang Chen,Jiateng Yin, Shiying Yang,Lingxi Li, Peter Pudney, "Constraint local principal curve: Concept,algorithms and applications", Elsevier,p223, 2016.
18- J. Wang, W. Mao, "Online Sequential Classification of Imbalanced Data by Combining Extreme Learning Machine and improved SMOTE Algorithm" IEEE, International Conference on Neural Networks (IJCNN), 2015.
19- Wu QingQiang, School of Software, Xiamen University, "Mixed-sampling Approach to Unbalanced Data Distributions": A Case Study involving Leukemia's Document Profiling, Vol 8, 2011.
20- Ginny Y. Wong_, Frank H.F. Leung_, and Sai-Ho Ling, "A Novel Evolutionary Preprocessing Method Based on Over-sampling and Under-sampling for Imbalanced Datasets", IEEE, 2013
21- Katarzyna Borowska1, Magdalena Topczewska2, "Data preprocessing in the classification of the imbalanced data", Advances in Computer Science Research, vol. 11, pp. 31-46, 2014.
22- UCI, Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets.html, 2016
23-KEEL, Data set Respority, http://sci2s.ugr.es/keel/imbalanced.php, 2016
24- Hualong Yu and Jun Ni, "An Improved Ensemble Learning Method for Classifying High-Dimensional and Imbalanced Biomedicine Data", IEEE Transactions on Computational Biology and Bioinformatics, Vol. 11, NO. 4, 2014
25-Mehrdad jalali, S. Fallah Mehneh, J. Gazalan Toosi, "An optimized approach for unbalanced big data categorizing using fuzzy clustering",IEEE International Congress on (ICTCK), 2014
26- Jia Pengfei, Zhang Chunkai, He Zhenyu, "A new sampling approach for classification of imbalanced data sets with high density", IEEE, International Conference on (BIGCOMP), 2014