بررسی روش‌های مختلف بیش نمونه‌گیری در رده‌بندی داده‌های نامتوازن

هادی مهدوی‌نیا
دانشجوی دکتری مهندسی نرم افزار دانشگاه آزاد اسلامی واحد خوراسگان
Email:hadi.mahdavinia1365@gmail.com

چکیده

در استخراج دانش از داده‌ها، از نظر توازن، به داده‌های متوازن و نامتوازن برخورد می‌شود. برای رده‌بندی داده‌های نامتوازن، اگر از روش‌های معمول رده‌بندی مانند ماشین‌های بردار و غیره استفاده شود، مشکلاتی مانند: مدل جانبدارانه، رده‌بندی اشتباه رده‌ی اقلیت، صرفه‌نظر کردن از داده‌های رده‌ی اقلیت و بیش‌پوشش به وجود خواهد آمد. برای اجتناب از مشکلات، باید از روش‌های خاصی برای رده‌بندی داده‌های نامتوازن استفاده شود. برای رده‌بندی داده‌های نامتوازن، از یکی از روش‌های سطح داده، روش‌های سطح الگوریتم، روش‌های یادگیری حساس به هزینه و روش‌های یادگیری ترکیبی استفاده می‌شود که استفاده از هر یک از روش‌ها به اهداف استخراج دانش بستگی دارد. در روش‌های سطح داده، یکی از روش‌های مرسوم، افزایش نمونه‌های رده‌ی اقلیت است که به بیش‌نمونه‌گیری معروف است. در این تحقیق انواع روش‌های بیش‌نمونه‌گیری در داده‌کاوی داده‌های نامتوازن که یکی از زیرمجموعه‌های رویکرد سطح داده است، بررسی می‌شوند.

کلمات کلیدی: داده‌های نامتوازن، رده اکثریت، رده اقلیت، بیش‌نمونه‌گیری

۱–مقدمه

یکی از مباحث معمول و رایج یادگیری ماشین[۱]، رده‌بندی[۲] مجموعه داده‌های نامتوازن[۳] است[۱]. زمانی که یک برچسب یا گروه از مؤلفه‌ی هدف در مقابل دیگر برچسب‌ها، دارای نمونه‌های کمتری باشد، مجموعه داده مورد نظر، یک مجموعه داده‌ی نامتوازن است[۲]. رده‌ای[۴] که نسبت به رده‌های دیگر مؤلفه‌ی هدف از میزان کمتری نمونه برخوردار است، رده‌ی اقلیت[۵] و مابقی نیز رده‌ی اکثریت[۶] نامیده می‌شوند[۳]. در استخراج دانش از مجموعه داده‌های نامتوازن، داده‌های رده‌یِ اقلیت، از اهمّیت ویژه‌ای برخوردار است، بنابراین سعی می‌شود نمونه‌های رده‌ی اقلیت به صورت صحیح افزایش یابد[۴].

داده‌های نامتوازن در بسیاری از زمینه‌ها، از جمله تشخیص خطا و عیب، تشخیص پزشکی، تشخیص نفوذ، تشخیص متون، تشخیص تقلب‌های بانکی، رده‌بندی جریان داده‌ها و از این قبیل کاربرد دارند[۵].

خطاهای غیر منتظره‌ای مانند صرفه‌نظر کردن از رده‌ی اقلیت، استخراج مدل جانبدارانه[۷] و محاسبه‌ی خطاهایِ رده‌هایِ مختلف برای کسب هزینه‌هایِ مختلف، از مشکلات و چالش‌های استخراج دانش از درون داده‌های نامتوازن است[۷].

---

[1] Machine Learning
[2] Classification
[3] Imbalanced datasets
[4] Class
[5] Minority Class
[6] Majority Class
[7] Bias Model

برای حل مشکلات ذکر شده، از دسته‌بندی‌های مختلف استفاده می‌شود که هر کدام مزایا و معایب خود را دارند. بارتوز کوارشکی، برای حلِ مشکلاتِ داده‌هایِ نامتوازن، سه دسته‌بندی از راه‌کارهای موجود ارائه داده است که شامل: رویکرد سطح داده[1]، یادگیری سطح الگوریتم[2] و یادگیری ترکیبی[3] است[۸]. امّا بهترین دسته‌بندی توسط ژِن و همکاران ارائه شده است که چهار دسته راه حل برای رده‌بندی داده‌های نامتوازن ارائه داده‌اند که عبارتند از رویکرد سطح داده، رویکرد سطح الگوریتم، یادگیری حساس به هزینه و یادگیری ترکیبی [۹].

در این تحقیق، در فصل دوم، خلاصه‌ای از انواع روش‌های رده‌بندی داده‌های نامتوازن تشریح می‌گردد؛ در فصل سوم انواع روش‌های بیش‌نمونه‌گیری معرفی می‌گردد و در فصل چهارم نتیجه‌گیری و نقاط ضعف و قوت هر یک از روش‌ها به صورت اجمالی بیان می‌گردد.

۲- دسته‌بندی روش‌های رده‌بندی داده‌های نامتوازن

برای رده‌بندی داده‌های نامتوازن، می‌توان از روش‌های مختلفی استفاده کرد که هر کدام مزایا و معایب خود را دارند. در رده‌بندی داده‌های نامتوازن، روش‌های اکتشافی در چهار دسته قرار می‌گیرند. حتی می‌توان هر روش در یک رویکرد را با روشی دیگر در رویکرد دیگر ترکیب کرد تا به کارائی مناسب و هدف مدنظر در رده‌بندی مجموعه داده‌ی نامتوازن رسید.

۲-۱-روش‌های سطح الگوریتم

در این روش‌ها، ساختار پایه‌ای الگوریتم‌ها تغییر پیدا می‌کند تا وابستگی و جانبداری مدل استخراجی از داده‌های نامتوازن، کاهش پیدا کند. یکی از تکنیک‌های مرسوم در این روش‌ها استفاده و اختصاص وزن به رده‌های پراهمیّت است[۱۰].

۲-۲- روش‌های سطح داده

با دیدگاه پیش‌پردازش داده‌ها، این روش، سعی در ایجاد توازن در رده‌های مؤلفه‌ی هدف دارد که این روش‌ها با تغییر در تعداد نمونه‌های رده‌های مختلف به این هدف می‌رسد[۶]. به عبارتی توزیع داده‌ها در مرحله یادگیری، باعث حل مشکل استخراج دانش و به خصوص رده‌بندی داده‌های نامتوازن می‌شود[۱۱].

از زیرروش‌هایِ این رویکرد که بسیار استفاده می‌شوند، می‌توان به روش بیش‌نمونه‌گیری و کم‌نمونه‌گیری اشاره کرد[۱۲].

۲-۳- روش‌های حساس به هزینه

در این رویکرد، هزینه‌های خطاهای هر رده یکسان نیست؛ بنابراین برای افزایش دقّت و صحت، به نمونه‌هایِ هر رده، وزن اختصاص داده می‌شود. در واقع در این رویکرد، برای رده‌هایِ اقلیتی که به صورت اشتباهی رده‌بندی می‌شوند، وزن بیشتری نسبت به رده‌یِ اکثریت در نظر گرفته می‌شود؛ بنابراین در فرآیند یادگیری، هدف، کاهشِ خطاهایِ وزنی، به جای افزایش نرخ صحت است. به عبارتی هدف اصلی این رویکرد کاهش هزینه رده‌های اشتباه رده‌بندی شده به جای کاهش خطای رده‌های اشتباه رده‌بندی شده است[۹][۶].

از روش‌های این رویکرد می‌توان به درخت‌هایِ تصمیمِ حساس به هزینه، شبکه‌هایِ عصبیِ حساس به هزینه، BEE-Miner و MEPAR اشاره کرد[۶].

۲-۴- روش‌های ترکیبی

[1] Data level approach
[2] Algorithm level approach
[3] Ensemble learning

در این روش، تکنیک‌های مختلف از سه دسته بندی فوق، با یکدیگر ترکیب شده و روش ترکیبی را برای ارائه یک مدل در رده‌بندی داده‌های نامتوازن مورد استفاده قرار می دهد. در این نوع از روش‌های داده‌کاوی داده‌های نامتوازن، دقّت بالایی از مدل ارائه می‌شود[۶][۱۴].
از روش‌های مرسوم این رویکرد می‌توان به AdaBoost و Boosting اشاره کرد[۱۵].

۳- روش‌های بیش‌نمونه‌گیری[1]

در فصل ۲، انواع روش‌های استخراج دانش از داده‌های نامتوازن در رده‌بندی داده‌های نامتوازن بیان گردید. یکی از روش‌های ذکر شده، روش سطح داده است که دارای زیر رویکردهایی همچون: کم‌نمونه‌گیری[2]، بیش‌نمونه‌گیری و روش‌های تلفیقی نمونه‌گیری است.
از آنجائی که رده‌ی اقلیت یک رده‌ی مهّم است و داده‌های موجود در آن نیز از اهمیّت ویژه ای برخوردار است، روش‌های ارائه شده در بیش‌نمونه‌گیری از حساسیت بیشتری برخوردار است. بنابراین در این فصل، انواع روش‌های موجود در زیررویکرد بیش‌نمونه‌گیری بررسی می‌گردد.

۳-۱- روش بیش‌نمونه‌گیری اقلیت مصنوعی[3]

معروف‌ترین روش افزایش رده‌ی اقلیت، این روش است. SMOTE جایگزین روش تولید رده‌ی اقلیت مصنوعی با جایگزینی است. اولین بار این روش از تشخیص موفقیت‌آمیز دست نوشته‌ها الهام گرفت. روش افزایش مصنوعی رده‌ی اقلیت به جای عمل در فضای داده‌ای در فضای مؤلفه‌ای یا صفت خاصه عمل می‌کند که هر سطر به عنوان نقطه مرکزیت و سپس تعداد k سطر به عنوان نزدیکترین همسایه نقطه مرکزیت به عنوان نمونه مصنوعی رده‌ی اقلیت تولید می‌شود[۱۶].
یکی از ورودی‌های الگوریتم، عدد T است که بیانگر تعداد نمونه‌های رده‌ی اقلیت است؛ عدد N نیز مقدار درصد تولید نمونه تصادفی از رده‌ی اقلیت است و عدد k نیز تعداد نزدیک‌ترین همسایه به نقطه مرکزیت است. آرایه‌ی Synthetic نیز تمامی نمونه‌های مصنوعی تولید شده را در خود نگه می‌دارد و مقدار بازگشتی الگوریتم SMOTE آرایه‌ای از نمونه‌های مصنوعی تولید شده از رده‌ی اقلیت است[۱۶].
در سال‌های اخیر تلاش‌هایی برای بهبود این روش صورت گرفته است و روش‌هایی مانند Safe-Level-SMOTE و همچنین Borderline-SMOTE ارائه شده است[۱۷][۱۸].

۳-۲- بیش‌نمونه‌گیریِ خوشه‌بندی[4]

در این روش، الگوریتم مورد نظر، ابتدا تعداد سرخوشه‌های مشخصی را از بین داده‌های رده‌ی اقلیت پیدا می‌کند و سپس k نزدیک‌ترین همسایه به هر کدام از سرخوشه‌ها را محاسبه کرده و ارائه می‌دهد. بدین ترتیب داده‌های مصنوعی نزدیک به رده‌ی اقلیت داده‌ها را پیدا کرده و به مجموعه داده‌ی رده‌ی اقلیت اضافه می‌کند. قابل ذکر است روش k-means یک روش خوشه‌بندی است که با ارزش محاسباتی کم برای داده‌هایی با بُعد بالا می‌تواند مورد استفاده قرار گیرد[۱۹].
این روش می‌تواند با انواع روش‌های خوشه‌بندی مانند k-Mediods یا DBSCAN و CFDP[5] و روش‌های دیگر خوشه‌بندی صورت گیرد.

---
[1] Oversampling
[2] Undersampling
[3] Synthetic Minority Oversampling Technique(SMOTE)
[4] Cluster-based oversampling
[5] Clustering by Fast search and find of Density Peaks

۳-۳- روش CTD [1]

در رده‌ی اقلیت، داده‌های مرزی بسیار حیاتی‌تر نسبت به مراکز رده‌های اقلیت هستند. بنابراین روش‌های موجود، بیشتر، نمونه‌های مرزی رده‌ی اقلیت را در نظر می‌گیرند. در زمان انتخاب نمونه‌های رده اقلیت برای افزایش رده اقلیت با استفاده تولید نمونه تصادفی و مصنوعی، نمونه‌ها می‌توانند در سه دسته قرار بگیرند[۲۰]:

۱- نمونه هایی با تأثیر تجمعی آشکار در رده‌ی اقلیت، که می توانند تأثیر کمی در کارائی مدل داشته باشند.

۲- نمونه های دارای نویز یا نمونه‌های جدا شده از رده اقلیت که از نمونه‌های رده‌ی اقلیت دور هستند و تأثیر تجمعی آشکار دارند. این نمونه‌ها نه تنها تأثیری مثبتی در کارائی مدل ندارند بلکه تأثیر منفی بر روی کارائی مدل نیز هم دارد.

۳- نمونه هایی که به دو دسته‌ی بالا مرتبط نیستند و نقش بسزایی در بهبود تشخیص رده‌های اقلیت دارد. در واقع این نمونه‌ها تأثیر زیادی در بهبود رده‌بندی‌های پایین رده‌ای دارد. شکل (۱) نمای شماتیک از سه دسته داده را نشان می‌دهد.

این روش هر سه نمونه را در تولید نمونه‌های رده اقلیت استفاده می‌کند. از داده‌ها با دامنه‌ی مثبت (pos)، نمونه‌ها با دامنه‌ی منفی (neg) و از نمونه‌ها با دامنه مرزی (bnd) استفاده می‌شود. به زبانی ساده، استفاده از روش‌هایی برای تشخیص این سه دسته و استفاده از آنها برای تولید رده‌های اقلیت، از عملیات‌های روش CTD است. در این روش همچنین برای نرخ نمونه‌گیری از روش SMOTE استفاده می‌شود[۲۰].

۴-۳- روش FRO [2]

این روش به طور خاص به مشکل رده‌بندی داده‌های نامتوازن با خصوصیات عددی می‌پردازد. برای تعیین وزن قوانین در این روش از رابطه (۱) استفاده می‌شود.

رابطه (۱)
$$rw_j = \frac{\sum_{xi \,\epsilon\, classCj} f_{AJ}(X_I)}{\sum_{i=1}^{m} f_{A_j}(xi)}$$

در این رابطه C_j یک برچسب از یک رده است؛ m تعداد قوانین فازی است؛ rw_j وزن محاسبه شده قانون j است و A_j یک ارزش زبانی است که توسط یک تابع عضویت به نام f تعریف شده است[۲۱].

از این رابطه می توان نتیجه گرفت که هر چه وزن قوانین بیشتر باشد، مساحت پوشش داده شده توسط قوانین امن‌تر خواهد بود. این الگوریتم در دو نوع رده‌ی دودوئی و رده‌ی چند متغیره استفاده می‌شود. همچنین در این روش مقادیر گمشده[3] نیز پوشش داده می‌شوند[۲۱].

۵-۳- رویکرد نمونه‌گیری مصنوعی سازگار [4]

یکی دیگر از روش‌های افزایش رده‌ی اقلیت، استفاده از روش ADASYN است. تحقیقات زیادی بر روی این روش و تفاوت آن با روش SMOTE شده است. ایده‌ی اصلی این روش استفاده از یک توزیع وزن‌دار برای نمونه‌های متفاوت رده‌ی اقلیت مطابق با سطوح مختلف یادگیری است که این داده‌ها، نسبت به داده‌های واقعی رده اقلیت، یادگیری سخت‌تری دارند. این روش یادگیری از داده‌های نامتوازن، سعی در بهبود دو جنبه از داده‌های نامتوازن را دارد[۲۲]:

۱-کاهش معرفی یک مدل جانبدارانه

[1] CCA based three-way decision model
[2] Fuzzy Rule-based Oversampling
[3] Missing values
[4] ADASYN: Adaptive Synthetic Sampling Approach

۲-تغییر مرز رده‌بندی در مورد داده‌های سخت به طور سازگار

ورودی ADASYN یک مجموعه داده‌ی آموزشی به تعداد m نمونه است. رویه‌ی اصلی ADASYN به صورت مراحل ذیل است[۲۲]:

۱- ابتدا درجه‌ی عدم توازن محاسبه می شود. در رابطه (۲) تعداد نمونه‌های رده‌ی اقلیت با m_S و تعداد نمونه‌های رده‌ی اکثریت با m_l نشان داده شده است.

$$d = m_s/m_l \ , \ d \in (0, 1] \qquad \text{رابطه (۲)}$$

۲- اگر $d < d_{th}$ باشد:

I. محاسبه تعداد نمونه های مصنوعی مورد نیاز برای رده‌ی اقلیت از رابطه (۳) حاصل می‌شود.

$$G = (ml - ms) \times \beta \ , \ \beta \in [0, 1] \qquad \text{رابطه (۳)}$$

β مشخص کننده‌ی سطح مطلوبِ تولید داده‌های مصنوعی بعد از تولید داده‌های مصنوعی است که $\beta = 1$ به معنای تولید یک مجموعه داده‌ی متعادل کامل بعد از فرآیند تولید نمونه‌ی مصنوعی از رده‌ی اقلیت است. همچنین G کل تعداد نمونه‌های داده‌ی مصنوعی مورد نیاز برای رده‌ی اقلیت است.

II. پیدا کردن k نزدیکترین همسایه برای هر نمونه‌ای که عضو مجموعه داده‌ی رده‌ی اقلیت است. در روش ADASYN از فاصله اقلیدسی در فضای مؤلفه‌ای(صفات خاصه) برای پیدا کردن k همسایه نزدیک استفاده می‌شود. نام مجموعه k نزدیک‌ترین همسایه، r_i است که با رابطه‌ی (۴) بدست می آید که Δ_i همان تعداد نمونه‌ها در k نزدیکترین همسایه‌ی x_i که وابسته به رده‌ی اکثریت است. بنابراین $r_i \in [0, 1]$ است.

$$r_i = \Delta_i/K \ , \ i = 1, \dots, m_s \qquad \text{رابطه (۴)}$$

III. با استفاده از رابطه (۵)، نرمال‌سازی r_i با توزیع چگالی $(\sum_i \hat{r_i} = 1)$ صورت می‌گیرد.

$$\hat{r_i} = r_i / \sum_{i=1}^{m_s} r_i \qquad \text{رابطه (۵)}$$

IV. محاسبه تعداد نمونه‌های رده‌ی اقلیت مورد نیاز که عضو رده‌یِ اقلیت است. این کار به وسیله‌یِ رابطه‌ی (۶) صورت میگیرد.

$$g_i = \hat{r_i} \times G \qquad \text{رابطه (۶)}$$

V. برای هر نمونه x_i رده‌یِ اقلیت تعداد g_i نمونه‌ی مصنوعی تولید می‌شود که این تولید نمونه با انتخاب تصادفی یک نمونه‌ی رده‌ی اقلیت در یک حلقه با رابطه (۷) صورت می‌گیرد که $(x_{zi} - x_i)$ همان تفاوت برداری صفات خاصه یا همان مؤلفه‌های رده‌ی اقلیت است و یک عدد تصادفی بین صفر و یک است.

$$s_i = x_i + (x_{zi} - x_i) \times \lambda \ , \ \lambda \in [0, 1]$$ رابطه (۷)

۴- نتایج

هر یک از روش‌های بیش‌نمونه‌گیری با توّجه به نوع داده و یادگیرنده انتخاب می‌شـوند. یکی از کابردهای مهّم SMOTE در بازیابی اطلاعات در سـاخت کوله‌ی کلمات است. SMOTE و دیگر انشعابات آن مانند SMOTE و Borderline-SMOTE با یکدیگر تفاوت دارد که این تفاوت باعث تغییر در به کارگیری و همچنین نوع هدف دارد.

هر یک از روش‌های مطرح شـــده در بیش‌نمونه‌گیری برای مؤلفه‌ی هدف دودوئی و چند رده‌ای می‌توانند مورد اســـتفاده قرار بگیرند.

اســـتفاده انفرادی از روش‌های بیش‌نمونه‌گیری بازدهی کامل در بســـیاری از روش‌های داده‌کاوی و به خصـــوص رده‌بندی داده‌های نامتوازن ندارد و ســـعی می‌شـــود برای افزایش کارائی ســـیســـتم، روش‌های بیش‌نمونه‌گیری به همراه روش‌های کم‌نمونه‌گیری مورد استفاده قرار می‌گیرند.

در روش ADASYN برای نمونه‌های رده‌ی اقلیت وزن در نظر گرفته می‌شود و براساس وزن هر کدام از نمونه‌های رده‌ی اقلیت، نمونه‌ی مصـنوعی تولید می‌شـود و این در حالی‌ست که در روش SMOTE، تولید نمونه‌های مصنوعی براساس K نزدیکترین همسایه صورت می‌گیرد که کلیه‌ی نمونه‌های رده‌ی اقلیت براساس یک شانس برابر، برای تولید نمونه‌ی تصادفی از نزدیکترین همسـایه اسـتفاده می‌کنند. در روش ADASYN صـحّت و دقّت مدل ایجاد شـده و یا دانش استخراج شده، بالاتر است زیرا نمونه‌های تولید شده از رده‌ی اقلیت، بیشترین شباهت را به اصل نمونه‌های وزن‌دار رده‌ی اقلیت دارند.

در روش FRO، حتی اگر مؤلفه‌ها بین صفر و یک نباشند، خروجی تولید شده بین صفر و یک تولید می‌شوند، بنابراین نیاز به نرمال‌سازی نیست.

در روش CTD نمونه‌های مرزی بسیار پر اهمیّت هستند، امّا در روش بیش‌نمونه‌گیریِ خوشه‌بندی، آن نمونه‌هایی از اهمیّت بیشتری برخوردار که در سرخوشه باشند.

در روش‌های بیش‌نمونه‌گیری نباید داده‌های مصـــنوعی تولید شـــده از رده‌ی اقلیت، باعث یادگیری ســـخت شـــوند و یا مدل غیرمناسب و ناکارا ارائه دهند.

۵- منابع

[1] *Q. Zou, S. Xie, Z. Lin, M. Wu, and Y. Ju, "Finding the best classification threshold in imbalanced classification," Big Data Res., vol. 5, pp. 2–8, 2016.*

[2] *D. Devi and B. Purkayastha, "Redundancy-driven modified Tomek-link based undersampling: a solution to class imbalance," Pattern Recognit. Lett., vol. 93, pp. 3–12, 2017.*

[3] *S. Abdellatif, M. A. Ben Hassine, S. Ben Yahia, and A. Bouzeghoub, "ARCID: A new approach to deal with imbalanced datasets classification," in International Conference on Current Trends in Theory and Practice of Informatics, 2018, pp. 569–580.*

[4] *M. A. H. Farquad and I. Bose, "Preprocessing unbalanced data using support vector machine," Decis. Support Syst., vol. 53, no. 1, pp. 226–233, 2012.*

[5] *Y. Qian, Y. Liang, M. Li, G. Feng, and X. Shi, "A resampling ensemble algorithm for classification of imbalance problems," Neurocomputing, vol. 143, pp. 57–67, 2014.*

[6] *J. F. Díez-Pastor, J. J. Rodríguez, C. García-Osorio, and L. I. Kuncheva, "Random Balance: Ensembles of variable priors classifiers for imbalanced data," Knowledge-Based Syst., vol. 85, pp. 96–111, 2015.*

[7] *S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. J. Kennedy, "Training deep neural networks on imbalanced data sets," in Neural Networks (IJCNN), 2016 International Joint Conference on, 2016, pp. 4368–4374.*

[8] *B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," Prog. Artif. Intell., vol. 5, no. 4, pp. 221–232, 2016.*

[9] *Z. Zhang, B. Krawczyk, S. Garcìa, A. Rosales-Pérez, and F. Herrera, "Empowering one-vs-one decomposition with ensemble learning for multi-class imbalanced data," Knowledge-Based Syst., 2016.*

[10] *J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," J. Big Data, vol. 6, no. 1, p. 27, 2019.*

[11] *I. Nekooeimehr and S. K. Lai-Yuen, "Cluster-based weighted oversampling for ordinal regression (CWOS-Ord)," Neurocomputing, vol. 218, pp. 51–60, 2016.*

[12] *T. Elhassan and M. Aljurf, "Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method.","2016.*

[13] *L. Zhao et al., "A cost-sensitive meta-learning classifier: SPFCNN-Miner," Futur. Gener. Comput. Syst., 2019.*

[14] *R. Barandela, R. M. Valdovinos, and J. S. Sánchez, "New applications of ensembles of classifiers," Pattern Anal. Appl., vol. 6, no. 3, pp. 245–256, 2003.*

[15] *A. Zhukov, N. Tomin, V. Kurbatsky, D. Sidorov, D. Panasetsky, and A. Foley, "Ensemble methods of classification for power systems security assessment," Appl. Comput. informatics, 2017.*

[16] *N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.*

[17] *C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem," in Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2009, pp. 475–482.*

[18] *H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," in International Conference on Intelligent Computing, 2005, pp. 878–887.*

[19] *M. S. Santos, P. H. Abreu, P. J. García-Laencina, A. Simão, and A. Carvalho, "A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients," J. Biomed. Inform., vol. 58, pp. 49–59, 2015.*

[20] *Y. T. Yan, Z. B. Wu, X. Q. Du, J. Chen, S. Zhao, and Y. P. Zhang, "A three-way decision ensemble method for imbalanced data oversampling," Int. J. Approx. Reason., vol. 107, pp. 1–16, 2019.*

[21] *G. Liu, Y. Yang, and B. Li, "Fuzzy rule-based oversampling technique for imbalanced and incomplete data learning," Knowledge-Based Syst., vol. 158, pp. 154–174, 2018.*

[22] *H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008, pp. 1322–1328.*