DOI:https://doi.org/10.22034/sdsb.1.4.43

توسعه پایدار و ساختمان هوشمند

https://sdsb.buqaen.ac.ir

جلد ۱، شماره ۴، زمستان ۱۴۰۴

ارائه رویکردی جدید برای نمونه‌زایی- نمونه‌زدایی در مجموعه داده نامتوازن

حسین زعفرانی[۱*]، صادق رستاد[۲]

۱- گروه مهندسی کامپیوتر- دانشکده مهندسی- دانشگاه سمنان- سمنان- ایران.

۲- گروه مهندسی کامپیوتر- دانشکده مهندسی- دانشگاه آزاد اسلامی واحد بیرجند- بیرجند- ایران.

چکیده

نامتوازن بودن مجموعه داده‌ها در دنیای واقعی، امری بسیار رایج و البته چالشی مهم در مسائل طبقه‌بندی و خوشه‌بندی آن‌ها است. مجموعه داده‌های نامتوازن، به مجموعه‌هایی اطلاق می‌شود که در آن نمونه‌های یک یا چند کلاس بر نمونه‌های کلاس‌های دیگر غلبه دارد. در این پژوهش، سعی خواهد شد تا یک مکانیزم، جهت متوازن نمودن مجموعه داده‌های نامتوازن ارائه شود که طی آن روش نمونه-زدایی و نمونه‌زایی با یکدیگر ادغام می‌شود. به طور کلّی، هدف از این طرح، ارائه یک روش ترکیبی نمونه‌زدایی- نمونه‌زایی است که بتواند به گونه‌ای کارا و مؤثر، مجموعه نامتوازن را متوازن کند. رویکرد پیشنهادی در این پژوهش به جای تمرکز بر کلاس‌های اقلیت و اکثریت، بر روی داده‌های اقلیت و اکثریت کار می‌کند. جهت شناسایی این داده‌ها، از الگوریتم خوشه‌بندی birch استفاده می-شود. از این رو، روش پیشنهادی تحت عنوان birch-resampling نام‌گذاری می‌شود. این روش از شش مرحله تشکیل شده است. جهت پیاده‌سازی این الگوریتم از نرم‌افزار «متلب» استفاده شد و برای ارزیابی نیز، از ۱۱ مجموعه دادهٔ دو و چند کلاسه با حجم و رخ عدم توازن متنوع استفاده شد، تا از همه ابعاد، عملکرد سیستم پیشنهادی مورد ارزیابی قرار گیرد. نتایج به‌دست آمده از الگوریتم پیشنهادی با ۵ روش دیگر، مورد مقایسه قرار گرفت. نتایج حاصل از این مقایسه‌ها، نشان می‌دهد که الگوریتم پیشنهادی از دقت بالایی در طبقه‌بندی داده‌ها برخوردار است.

اطلاعات مقاله

نوع مقاله: مقاله پژوهشی

تاریخ دریافت: ۱۴۰۴/۰۹/۰۸

تاریخ پذیرش: ۱۴۰۴/۱۱/۰۷

واژه‌های کلیدی:

داده نامتوازن، خوشه‌بندی Birch، طبقه‌بندی، داده کاوی.

* نویسنده مسئول، پست الکترونیکی: hzaaferani@semnan.ac.ir

Digital Object Identifier https://doi.org/10.22034/sdsb.1.4.43

۱. مقدّمه

طبقه‌بندی و خوشه‌بندی داده‌ها، یکی از تکنیک‌های مهم در یادگیری ماشین و داده‌کاوی است که این امکان را فراهم می‌آورد تا سیستم کامپیوتری بتواند داده‌ها و الگوهای مختلف را شناسایی کند و آن‌ها را از یکدیگر تشخیص دهد. روال اصلی در طبقه‌بندی و خوشه‌بندی داده‌ها، ایجاد مدل، جهت شناسایی داده‌ها از یکدیگر است. اصلی‌ترین مرحله در ساخت مدل در داده‌کاوی و یادگیری ماشین، آموزش، است [۱]. در فرایند آموزش، مجموعه‌ای از داده‌ها به سیستم داده می‌شود. سپس سیستم با شناسایی ارتباط بین آن‌ها، اقدام به ساخت مدل برای داده‌ها می‌کند. در این فرایند، سعی می‌شود تا تعداد نمونه‌های داده شده به سیستم، برای هر یک از کلاس‌ها، برابر باشد. به طور علمی‌تر، داده‌ها دارای توزیع یکسانی باشند. با این حال، بسیاری از مجموعه داده‌های موجود در دنیای واقعی، از قبیل تشخیص تقلّب، ناهنجاری، پزشکی و ... دارای توزیع یکسانی نیستند [۲]. چنین داده‌هایی، داده‌های نامتوازن نامیده می‌شوند. به عبارت بهتر، «مجموعه داده نامتوازن»، به مجموعه داده‌ای گفته می‌شود که تعداد نمونه‌های موجود در یک یا چند کلاس، از سایر کلاس‌های موجود در مجموعه دادهٔ آموزش بیشتر باشد [۲]. در این نوع داده‌ها، به کلاس‌هایی که کمترین تعداد نمونه‌ها را دارند، کلاس اقلیت و به کلاس‌هایی که تعداد نمونه‌های زیادی دارند، کلاس اکثریت، گفته می‌شود [۳].

موضوع اصلی در زمینه طبقه‌بندی داده‌های نامتوازن این است که الگوریتم‌های طبقه‌بندی استاندارد، اغلب تمایل بیشتری به کلاس اکثریت دارد و قوانینی که این نمونه‌ها را به درستی پیش‌بینی می‌کند، به درستی وزن‌دهی می‌شود، اما قوانین خاصی که نمونه‌های کلاس اقلیت را پیش‌بینی می‌کند، اغلب نادیده گرفته می‌شود و در واقع به صورت نویز با آن‌ها برخورد می‌شود، در نتیجه، نمونه‌های کلاس اقلیت، به اشتباه کلاس‌بندی خواهند شد. ازاین‌رو، نرخ طبقه‌بندی اشتباه برای داده‌های اقلیت بالا می‌رود [۴]. جهت حل این مسئله، راهکارهای مختلفی ارائه شده است که می‌توان آن‌ها را در چهار گروه تقسیم‌بندی نمود: روش سطح داده، رویکردهایی در سطح الگوریتم، روش‌های یادگیری جمعی و روش حساس به هزینه.

در این بین، روش‌های در سطح داده، دارای پیچیدگی کمتری هستند. در رویکردهای در سطح داده، توزیع کلاس نامتوازن با نمونه‌گیری مجدد در فضای داده‌ها متوازن می‌شود. این روش‌ها، با دو رویکرد نمونه‌زدایی و نمونه‌زایی، از روش‌های مؤثر در متوازن نمودن داده‌ها محسوب می‌شود [۵].

هر چند، ارائه رویکرد خاص و ویژه برای طبقه‌بندی و خوشه‌بندی داده‌های نامتوازن، به عنوان اولین راه‌حل به ذهن می‌رسد، اما راه حل آسان‌تر و مناسب‌تر، متوازن کردن مجموعه داده نامتوازن، به گونه‌ای کارآمد و مطلوب است که بتواند به افزایش کارایی الگوریتم‌ها نیز اضافه کند. در مطالعهٔ مجموعه داده نامتوازن، نیاز به یک روش با پیچیدگی کم، جهت متوازن ساختن داده‌ها است؛ به گونه‌ای که بتواند در زمان مطلوب و بدون کاستن از دقتِ طبقه‌بندیِ داده‌ها، اقدام به متوازن‌کردن داده‌ها کند. هر چند، ایجاد روش نمونه‌زایی و نمونه‌زداییِ مطلوب، خود به طور مستقل، هنوز به عنوان یک مسئله و چالش تحقیقاتی مطرح است، اما ادغام این دو روش نیز، موضوعی است که تاکنون به آن، کمتر پرداخته شده است و مطالعات اندکی در این خصوص وجود دارد و این امر می‌تواند خود به عنوان یک نوآوری برای این پژوهش مطرح گردد.

رویکرد پیشنهادی در این پژوهش به جای تمرکز بر کلاس‌های اقلیت و اکثریت، بر روی داده‌های، اقلیت و اکثریت کار می‌کند. جهت شناسایی این داده‌ها، از الگوریتم خوشه‌بندی birch استفاده می‌شود. از این رو، روش پیشنهادی تحت عنوان birch-resampling نام‌گذاری می‌شود. این الگوریتم برخلاف سایر الگوریتم‌های دیگر که به نویز حساس هستند، حساسیت کمتری نسبت به نویز دارد. علاوه براین، خود به صورت خودکار، اقدام به تولید دسته و خوشه جدید برای مجموعه داده می‌کند. از این رو، نیازی به تعیین تعداد خوشه از ابتدای راه‌اندازی الگوریتم نیست. همین امر سبب منعطف شدن این الگوریتم و توانایی کار با مجموعه داده‌های مختلف شده است.

ادامهٔ این مقاله به صورت زیر تنظیم شده است. در بخش ۲، مروری بر ادبیات مربوط به مسئله داده‌های نامتوازن و پژوهش‌های مرتبط پیشین ارائه می‌شود. سپس، در بخش ۳، جزئیات روش پیشنهادی مقاله بیان می‌شود. بخش ۴، به بیان جزئیات آزمایش‌های انجام شده و ارائه نتایج حاصل شده اختصاص یافته است، و در بخش ۵، یک جمع‌بندی از محتوای مقاله ارائه خواهد شد.

۲. مروری بر پژوهش‌های مرتبط

داده‌کاوی، فرایند کشف و استخراج الگوهای پنهان، از حجم بالایی از داده‌ها است [۶] و به کشف روابط ناشناخته و الگوهای درون‌داده‌ها می‌پردازد [۷]. هدف از فرایند داده‌کاوی، ساخت مدلی مؤثر، توصیفی یا پیشگویانه، از بین مقدار زیادی از داده است که نه تنها توانایی توصیف آن‌ها را داشته باشد، بلکه بتواند برای داده جدید نیز تعمیم‌داده شود [۸].

گام مهم و اصلی در داده‌کاوی، یادگیری ماشین است. «یادگیری» عبارت است از تغییر به نسبت پایدار در احساس، تفکّر و رفتار فرد که براساس تجربه ایجاد شده باشد. یادگیری ماشین، یک شاخه از هوش مصنوعی و یک زمینه تحقیقاتی جذاب و فعال، در حوزه علوم کامپیوتر است که به تنظیم و اکتشاف شیوه‌ها و الگوریتم‌هایی می‌پردازد که براساس آن، کامپیوترها توانایی تعلم و یادگیری پیدا می‌کند. به طور کلی، الگوریتم‌های یادگیری به چهار دسته تقسیم می‌شوند: یادگیری نظارت شده،[1] یادگیری نظارت نشده،[2] یادگیری تقویتی[3] و یادگیری نیمه نظارت شده [۹].

روش‌های نظارت شده، از پایگاه داده‌ای که حاوی داده‌های ساختارمند از کلاس‌های معیّن است، استفاده می‌کند. این تکنیک‌ها، از داده‌های گذشته یاد گرفته و آن را در آینده به کار می‌بندند [۷, ۱۰]. این نوع یادگیری، زمانی استفاده می‌شود که مجموعه داده، دارای برچسبِ کلاس است، بنابراین، داده‌ها و کلاس مربوط به هر یک، مشخص و معلوم است. از جمله مهم‌ترین روش‌های موجود در این حوزه می‌توان به روش‌های طبقه‌بندی[4] سنتی مانند k نزدیک‌ترین همسایه (KNN)، شبکه عصبی،[5] الگوریتم هیورستیک مانند ژنتیک، ماشین بردار پشتیبان (SVM) و درخت تصمیم[6] اشاره نمود [۱۱, ۱۲].

امروزه یکی از مسائل مهم در زمینهٔ داده‌کاوی و یادگیری ماشین، مسئلهٔ دسته‌بندی مجموعه داده‌های نامتوازن است. مشکل داده‌های نامتوازن، زمانی ایجاد می‌شود که تعداد نمونه‌های موجود در برخی کلاس‌ها، از سایر کلاس‌های موجود در مجموعه داده آموزشی[7] بیشتر باشند. نسبت تعداد نمونه‌های کلاس اکثریت[8] به کلاس اقلیت،[9] میزان نامتوازن بودن داده‌ها را نشان می‌دهد که در برخی از مجموعه داده‌ها، این نسبت ممکن است ۱:۱۰۰، ۱:۱۰۰۰ و یا بیشتر باشد [۱۴, ۱۵]. این مشکل، در مسائل واقعی از قبیل کشف آلودگی، مدیریت ریسک، کشف تقلب و تشخیص‌های پزشکی بسیار مشاهده شده است.

الگوریتم‌های معمول داده‌کاوی، در مواجهه با مشکل کلاس‌های نامتوازن، معمولاً عملکرد ضعیفی دارند؛ زیرا با صرف نظرکردن از نمونه‌های داده در کلاس اقلیت، اقدام به افزایش دقت کلّی می‌کنند. در صورتی که نمونه‌های موجود در کلاس اقلیت در بسیاری از کاربردها، نسبت به نمونه‌های کلاس دیگر، از اهمیت بیشتری برخوردارند. برای مثال، در مسئله تشخیص پزشکی، با وجود اینکه، نمونه‌های بیماری در مقایسه با نمونه‌های معمول در اقلیت قرار دارند [۱۶].

این روش‌های مواجهه با داده‌های متوزان را می‌توان به چهار دسته تقسیم کرد: روش‌های سطح الگوریتم، روش‌های سطح داده، روش‌های حساس به هزینه، روش‌های ترکیبی [۲۱]. روش‌های سطح داده که موسوم به روش‌های خارجی است، شامل الگوریتم‌هایی هستند که با داده‌ها سروکار دارد و روش‌های باز نمونه‌برداری[10] نیز نامیده می‌شود. این الگوریتم‌ها به صورت روش‌های پیش‌پردازش و قبل از انجام مرحله طبقه‌بندی، وارد عمل می‌شود و با متوازن کردن داده‌ها، آن‌ها را بهبود می‌دهد و آماده استفاده در الگوریتم‌های طبقه‌بندی می‌کند. هدف اصلی این پژوهش، استفاده از روش‌های در سطح داده است [۲۱].

سه راهکار برای باز نمونه‌برداری وجود دارد: نمونه‌زدایی از کلاس اکثریت، یعنی ایجاد زیرمجموعه‌ای از داده‌های اصلی با حذف نمونه‌های انتخاب شده از کلاس. نمونه‌زایی کلاس اقلیت‌ها، یعنی ایجاد یک اَبَرمجموعه از مجموعه داده‌های اصلی با تولید نمونه‌های جدید از نمونه‌های موجود یا با تکرار آن‌ها و ترکیبی که هر دو روش فوق را با هم ترکیب می‌کند.

از آنجا که هدف از این پژوهش، ارائه یک رویکرد مبتنی بر سطح داده است، لذا در اینجا نیز تحقیقات و مقالات مربوط به این حیطه مورد بحث قرار خواهد گرفت. این رویکرد از سه دسته کلّی تشکیل شده است: نمونه‌زایی، نمونه‌زدایی و ترکیبی.

یکی از روش‌های شناخته شده در زمینه نمونه‌زایی، روش نمونه‌زایی ترکیب اقلیت یا SMOTE است [۱]. در روش نمونه‌زایی ترکیب اقلیت، به ازای هر داده، تعدادی دادهٔ جدید ساخته می‌شود، ولی به همسایه‌های این داده‌ها توجهی نمی‌شود. در نتیجه، امکان ایجاد تداخل بین کلاس‌ها زیاد است [۱].

از جمله مقالاتی که به نمونه‌زدایی پرداختند، می‌توان به مقاله [۲۲] اشاره کرد که در آن دو استراتژی نمونه‌زدایی، توسط تکنیک خوشه‌بندی معرفی شده و در مرحله پیش‌پردازش اطلاعات،

1 Supervised
2 Unsupervised
3 Reinforcement Learning
4 classification
5 Neural network
6 Decision Tree
7 Training Set
8 Majority Class
9 Minority Class
10 Resampling

استفاده می‌شود. در این استراتژی‌ها، تعداد خوشه‌ها در کلاس اکثریت، برابر با تعداد نمونه‌های داده در کلاس اقلیت است. استراتژی اول از مراکز خوشه، برای نشان‌دادن کلاس اکثریت استفاده می‌کند، در حالی که استراتژی دوم، از نزدیک‌ترین همسایگان مراکز خوشه برای این کار بهره می‌گیرد. نتایج تجربی به‌دست آمده بر روی ۴۴ مجموعه‌داده کوچک و ۲ مجموعه داده بزرگ، حاکی از عملکرد مطلوب این دو استراتژی است. در مقاله [۳۶]، از نمونه‌زدایی تکاملی جهت کار با داده نامتوازن استفاده شده و برای کاهش محدودیت‌های سرعت و حافظه در این روش‌ها از رویکرد تقسیم و غلبه براساس MapReduce جهت تقسیم داده‌ها به چند زیر مجموعه بهره گرفته است، به‌گونه‌ای که در زیرمجموعه‌ها، سعی شده تا تعداد بیشتری از نمونه‌های کلاس اقلیت باقی بماند. در مقاله [۳۷]، یک مدل موازی جهت نمونه‌زدایی تکاملی به همراه استفاده از قابلیت‌های MapReduce طراحی شده است. در مقاله [۳۸]، نیز یک روش نمونه‌زدایی با استفاده از الگوریتم ژنتیک به نام GAUS برای غلبه بر محدودیت‌های موجود در روش‌های نمونه‌زدایی موجود، مانند بی-ثباتی عملکرد و از دست‌دادن داده‌ها طراحی شده است.

در مقاله [۳۹]، یک روش الگوریتم جدید پیشنهاد شده است که نمونه‌های اکثریت را از منطقه دارای هم‌پوشانی حذف می‌کند، از این رو، نمونه‌های اقلیت را بهبود می‌بخشد. در این مقاله، از روش غربالگری مبتنی بر هم‌پوشانی استفاده شده است.

در مقاله [۴۰]، یک الگوریتم انطباقی جهت نمونه‌زایی، برای غلبه بر مشکل عدم توازن کلاس‌ها، پیشنهاد شده است. در مقاله [۴۱]، یک چارچوب برای گسترش روش نمونه‌زایی توسط شبکه عصبی کانولوشن (CNN) ارائه شده است. ویژگی اصلی آن، یادگیری صریح و تحت نظارت است که داده‌های آموزشی هر نمونه، ورودی خام را با یک هدف جاسازی مصنوعی در فضای عمیق، که از زیرمجموعه خطی همسایگان در کلاس، نمونه‌برداری می‌شود، ارائه می‌دهد.

در مقاله [۴۲]، الگوریتمی تحت عنوان نمونه‌زاییِ خودسازماندهی مبتنی بر نگاشت (SOMO) ارائه شده است که یک نمایش دو بعدی از فضای ورودی را فراهم می‌کند که به تولید داده‌های ورودی مصنوعی منجر می‌شود. SOMO، شامل سه مرحله اصلی است: مرحله اول، یک نگاشت خودسازمانده، جهت نمایش دو بعدی از داده اصلی که معمولاً دارای ابعاد بالا است، تولید می‌شود. مرحله بعد، تولید نمونه‌های ترکیبی درون خوشه است و سرانجام، تولید نمونه‌های ترکیبی بین خوشه‌ها است.

در مقاله [۴۳]، یک استراتژی، برای ایجاد توازن در مجموعه داده نامتوازن ارائه شده است که نمونه‌های آن به عنوان رشته‌ها کدگذاری می‌شود. رویکرد پیشنهادی براساس تطبیق الگوریتم شناخته شده، تکنیک‌های جمع‌آوری اقلیت‌های مصنوعی (SMOTE) به فضای رشته است. به طور دقیق‌تر، تولید داده‌ها با یک رویکرد تکراری برای ایجاد رشته‌های مصنوعی درون بخش، بین دو نمونه داده شده از مجموعه آموزش به دست می‌آید.

مقاله [۴۴]، یک روش SMOTE مبتنی بر هسته وزن را پیشنهاد می‌دهد که محدودیت SMOTE را برای مشکلات غیرخطی در فضای ویژگی سیستم طبقه‌بندی‌کننده بردار پشتیبانی (SVM) برطرف می‌کند. الگوریتم پیشنهادی، یک رویکرد نمونه‌زایی و یک چارچوب سلسله‌مراتبی برای مشکلات عدم توازن چند لایه‌ای ایجاد می‌کند.

مقاله [۴۵]، یک مدل به نام CTD را با در نظر گرفتن تفاوت در هزینه انتخاب نمونه‌های کلیدی پیشنهاد می‌کند. این روش از یک الگوریتم سازنده پوشش،[1] جهت تقسیم نمونه‌های اقلیت به چند پوشش استفاده می‌کند. سپس، یک مدل تصمیم‌گیری برای انتخاب نمونه کلیدی بر اساس تراکم پوشش روی نمونه‌های اقلیت، ساخته می‌شود. در نهایت، بر اساس الگوی توزیع پوشش در نمونه-های اقلیت، پارامترهای تولید نمونه به‌دست می‌آید و پس از آن، نمونه‌های کلیدی را می‌توان برای نمونه‌زایی انتخاب کرد.

در برخی از مقالات نیز دو روش نمونه‌زایی و نمونه‌زدایی با هم مورد استفاده قرار گرفتند، که در این خصوص می‌توان به مقاله [۴۶] اشاره کرد که یک روش جدید به نام OSELM-UO برای نمونه‌زایی و نمونه‌زدایی، یادگیری ماشین آنلاین برای طبقه‌بندی نامطلوب داده‌های بزرگ ارائه کرده است که ترکیبی از ساختارهای نمونه‌زایی و نمونه‌زدایی را فراهم می‌سازد، به‌گونه‌ای که به طور همزمان، مسائل از دست رفتن بیش از حد اطلاعات را برطرف کند. در مقاله [۴۷] نیز، روش جدیدی برای نمونه‌زایی با هدف گسترش داده‌های اقلیت‌ها توسط پارامتر توزیع weibull و همچنین از روش box-whisker برای کاهش اندازه اطلاعات اکثریت براساس توزیع نرمال پیشنهاد شده است.

1 Constructive Covering Algorithm.

مقاله [۴۸] یک روش نمونه‌گیری ترکیبی به نام SCUT پیشنهاد نموده است که برای تعادل تعداد نمونه‌های آموزش در مجموعه‌های نامتوازن چند کلاسی استفاده می‌شود. رویکرد SCUT نمونه‌های کلاس‌های اقلیت را از طریق تولید نمونه‌های مصنوعی افزایش می-دهد.

در مقاله [۴۹]، یک چارچوب جمعی جدید به نام Undersampling- Boost Adaptive Ensemble برای یادگیری مجموعه داده نامتوازن پیشنهاد شده است، که ترکیبی از تکنیک EUS، Real Adaboost و اصلاح وزن حساس به هزینه و استراتژی تصمیم انطباقی است. با توجّه به نتایج تجربی و تجزیه و تحلیل‌ها، ثابت شده است که این چارچوب، یک روش امیدوارکننده در زمینهٔ طبقه‌بندی مجموعه داده‌های نامتوازن است.

در مقاله [۲۲]، یک روش ترکیبی جدید، براساس نمونه‌گیری مبتنی بر مجموعه (HUSBoost) برای مدیریت مجموعه داده‌های نامتوازن ارائه شده است که شامل سه مرحلهٔ پاک‌سازی داده‌ها، متوازن‌کردن داده‌ها و طبقه‌بندی داده‌ها است. ابتدا داده‌های نویزدار حذف می-شود. سپس چند زیر مجموعه متوازن با استفاده از روش نمونه‌زادیی تصادفی ایجاد شده و سپس در هر زیر مجموعه متعادل، جنگل‌های تصادفی (RF)، آدابوست با درخت تصمیم و آدابوست با ماشین بردار پشتیبانی (SVM) به صورت موازی، اجرا و بعد با استفاده از روش رأی‌گیری، نتایج هر یک با هم ترکیب می‌شود.

همچنین در مقاله [۳۰] نیز، یک رویکرد جدید و الگوریتمی به نام AdaOUBoost پیشنهاد شده که تطبیقی از نمونه‌زایی و نمونه‌زدایی است. در AdaOUBoost، ابتدا، نمونه‌زدایی مبتنی بر خوشه‌بندی، برای نمونه‌های اکثریت انجام می‌شود. سپس از الگوریتم border- SMOTE برای نمونه‌زایی داده‌ها، اقلیت استفاده شده و در نهایت، عمل طبقه‌بندی انجام خواهد شد.

در مقاله[۲]، یک رویکرد سطح داده، سطح الگوریتم، جهت رفع مشکل طبقه‌بندی در داده نامتوازن ارائه شده است. در این روش، ابتدا مجموعه داده‌های نامتوازن به چندین مجموعه داده متوازن تبدیل می‌شود. در واقع، داده‌های کلاس اکثریت در مجموعه داده به چند زیر مجموعه تقسیم می‌شود، درحالی که، داده کلاس اقلیت در همه زیرمجموعه‌ها تکرار می‌شود. سپس یک الگوریتم طبقه‌بندی خاص بر روی هر یک از این زیرمجموعه‌ها ایجاد شده و در نهایت، نتایج طبقه‌بندی این طبقه‌بندها برای داده‌های جدید با یک قاعده گروهی معین، ترکیب می‌گردد.

همان‌طور که در این بخش مشاهده شده، نمونه‌زایی و نمونه‌زدایی، به علت سادگی در پیاده‌سازی، قابل توجّه بسیاری از افراد بوده است. با این حال، هنوز چالش‌هایی وجود دارد که باید به آن‌ها رسیدگی شود. از جمله اینکه، هرچند روش‌های مختلفی جهت متوازن کردن مجموعه داده ارائه شده است، با این حال، این روش-ها در اغلب موارد، منجر به از دست دادن اطلاعات سودمند، وقوع اشتباهات غیر منتظره و یا تغییر توزیع داده اصلی می‌شود [۲]. به عبارت بهتر، در نمونه‌زدایی، انتخاب داده مناسب جهت حذف، امری بسیار دشوار است. حال اینکه هدف از متوازن ساختن مجموعه داده، برابر کردن نمونه‌های هر کلاس است. از سوی دیگری، در روش‌های نمونه‌زایی نیز، باید به طور دقیق و مناسب، اقدام به تولید نمونه‌های اقلیت شود. در غیراین صورت، این احتمال وجود دارد تا نمونه‌هـایی ایجـاد گردد که مشابه با داده‌های اکثریت است و می‌تواند منجر به نـامطلوب‌شـدن دقت طبقه‌بندی شود و یا نمونه‌هایی کاملاً تکراری شود که در این صورت تأثیری در طبقه‌بندی نداشته باشد.

۳. الگوریتم خوشه‌بندی Birch

الگوریتم خوشه‌بندی Birch یکی از الگوریتم‌های رایج در خوشه-بندی سلسله‌مراتبی است، که برای کار بر روی داده‌های عددی با حجم بالا طراحی شده و توانایی و قابلیت بسیار بالایی در تشخیص داده‌های نویز و مزاحم دارد [۲۸].

این الگوریتم از ویژگی خوشه[1]یا به اختصار CF جهت انجام خوشه-بندی استفاده می‌کند و همین امر سبب شده تا الگوریتم از سرعت و قابلیت مقیاس‌پذیری بالایی برخوردار باشد[۲۸]. ویژگی خوشه، یک روش خلاصه‌سازی اطلاعات است که نقاط داده برای هر زیرخوشه را نگهداری می‌کند. به طور رسمی، ویژگی خوشه را می‌توان به صورت زیر تعریف نمود [۳۱]:

برای N نقطه داده d بعدی در یک خوشه، $\{\vec{X}_i\}\ i = 1,2,\dots,N$، درآیه ویژگی خوشه به عنوان یک سه‌تایی به صورت $CF = (N, \overrightarrow{LS}, SS)$ تعریف شده که N تعداد نقطه داده در خوشه، $\overrightarrow{LS}$ مجموع خطی از N نقطه داده $(\sum_{i=1}^{N} \vec{X}_i)$ و SS مجموع مربعات N نقطه داده $(\sum_{i=1}^{N} \vec{X}_i{}^2)$ است.

اگر دو خوشه $CF_1 = (N_1, \overrightarrow{LS}_1, SS_1)$ و $CF_2 =$ $(N_2, \overrightarrow{LS}_2, SS_2)$، درآیه‌های دو زیر خوشه جدا از هم باشند، پس

1.Clustering Feature

درآیه CF زیر خوشه‌ای که توسط ادغام این دو زیرخوشه حاصل می‌شود، عبارتند از [۳۱]:

$$CF_1 + CF_2 = (N_1 + N_2, LS_1 + LS_2, SS_1 + SS_2) \qquad (۲-۲)$$

در الگوریتم Birch از درخت متوازن استفاده می‌شود. درخت متوازن CFها را جهت اجرای الگوریتم، ذخیره می‌کند. هر گره از این درخت که حاوی گره فرزند باشد، در خود، مجموع CFهای فرزندانش را نگهداری می‌کند، بدین‌ترتیب خلاصه‌ای از اطلاعات فرزندان خود را خواهد داشت. درخت، دارای دو پارامتر به نام‌های فاکتور شاخه‌بندی B و حد آستانه T است. حداکثر تعداد فرزندان گره‌های غیر پایانی با فاکتور شاخه‌بندی B مشخص می‌شوند و حد آستانه T حداکثر فاصله‌ای که میان دو نمونه از خوشه‌های برگ قرار دارد، اطلاق می‌شود [۲۸].

پس از درج و یا حذف یک CF در درخت و یا در صورت افزایش مقادیر B و T از مقدار تعیین شده توسط کاربر، گره‌های درخت متوازن تقسیم و یا ادغام می‌شود. واضح است که مقادیر B و T در اندازه درخت تولید شده نهایی، نقش بسزایی ایفا می‌کند. برای مثال، اگر حافظه کافی جهت ذخیره درخت CF وجود نداشته باشد، می‌توانیم با مقدار جدید T درخت را بازسازی و درخت کوچک‌تری تولید کنیم؛ بنابراین، می‌توان ادعا کرد، الگوریتم با منابع در دسترس، بهترین خوشه‌بندی را انجام می‌دهد. برای بازسازی درخت نیز لازم نیست تمام نمونه‌ها بازخوانی شوند، فقط کافی است با کمک گره‌های پایانی درخت قدیمی، درخت جدید را تولید کرد.

خوشه‌بندی Birch از نقاط داده چند بعدی استفاده می‌کند تا خوشه‌های با کیفیت با محدودیت زمانی و حافظه موجود تولید کند. این تکنیک خوشه‌بندی از ویژگی خوشه برای فشرده‌سازی اطلاعات در مورد زیر خوشه‌های نقاط استفاده می‌کند. ویژگی خوشه یا CF در درخت متوازن به نام CF-tree سازمان‌دهی می‌شود [۳۲].

درخت CF-tree یک درخت متوازن[1] با دو پارامتر است: فاکتور شاخه (B برای نودهای غیر برگ و L برای نودهای برگ) و آستانه T. هر نود غیر برگ، شامل حداکثر B درآیه به شکل $\{CF_i, child_i\}$ که i=1,2,…,B است که $child_i$ یک اشاره‌گر به iامین نود فرزند است و CF_i درآیه CF از زیرشاخه حاوی این فرزند است، بنابراین یک نود غیر برگ، یک زیرخوشه از همه زیر خوشه‌های ارائه شده توسط درآیه‌های اش ارائه می‌دهد.

یک نود برگ، حاوی حداکثر L درآیه است و هر درآیه یک CF است. علاوه بر این، هر نود برگ، دو اشاره‌گر دارد به نام‌های prev و next که برای اتصال همه نودهای برگ، برای بررسی کارآمد استفاده می‌کند. نود برگ همچنین، یک زیر خوشه از همه زیرخوشه‌های ارائه شده توسط درآیه‌هایش ارائه می‌دهد، اما همه درآیه‌ها در نود برگ باید حد آستانه لازم را برآورده کند. اندازه درخت، تابعی از T است. T بزرگ‌تر، منجر به ایجادِ درخت کوچک‌تر می‌شود.

درخت CF هنگامی‌که داده جدید درج می‌شود، به‌صورت دینامیک ساخته خواهد شد. این مسئله به منظور درج جدید در زیرخوشه صحیح برای خوشه‌بندی استفاده می‌شود. درخت CF، مجموعه داده به صورت فشرده ارائه می‌دهد، زیرا هر درآیه در نود برگ، یک داده مجزا نیست، بلکه یک زیرخوشه شامل نقاط داده بسیار با قطر یا شعاع کمتر از حد آستانه T است [۳۱].

۴. رویکرد پیشنهادی

رویکرد پیشنهادی در این مقاله، علاوه بر کلاس اقلیت و اکثریت، به داده‌ها و تراکم نمونه‌ها نیز توجّه می‌کند. به عبارتی بهتر، در مجموعه داده نامتوازن، همه نمونه‌ها دارای یک وزن و ارزش یکسان نیستند. برخی از نمونه‌ها، به لحاظ موقعیت و قرارگیری در فضای نمونه‌ها، دارای موقعیت استراتژیک و حسّاسی هستند. حذف چنین نمونه‌ای در عملیات نمونه‌زدایی شاید بتواند تا حدی نتیجه طبقه‌بندی را بهبود دهد و طبقه‌بندی را متعادل کند، اما موجب از دست رفتن داده‌ای ارزشمند نیز شده است، بنابراین در فرایند نمونه‌زایی و نمونه‌زدایی، قبل از اعمال هر گونه تغییر باید به نمونه‌ها و موقعیت آن‌ها نیز توجّه کرد.

در این پژوهش با استفاده از الگوریتم خوشه‌بندی birch، اقدام به تعیین تراکم بین داده‌ها خواهد شد. این الگوریتم، می‌تواند داده‌ها را در یک درخت سازمان دهد، طوری‌که داده‌های نزدیک به یکدیگر و مشابه، در یک شاخه از درخت قرار گیرند. در ادامه به تشریح الگوریتم پیشنهادی می‌پردازیم.

فاز اول: تقسیم داده‌ها

در این مرحله، مجموعه داده به دو بخش تقسیم می‌شود، شامل داده آموزش و داده آزمایش. داده آموزش، جهت آموزش‌دادن سیستم طبقه‌بندی مورد استفاده قرار خواهد گرفت. داده آزمایش نیز، جهت تأیید مدلِ طبقه‌بندی و آزمایش طبقه‌بند استفاده خواهد شد. سپس، در مجموعه داده‌های آزمایش، نمونه‌ها با توجّه به برچسب کلاس، از یکدیگر تفکیک می‌شوند. بعد از جداسازی نمونه‌ها، با توجّه به تعداد داده‌های هر کلاس، کلاس‌ها به عنوان اقلیت و اکثریت، برچسب دریافت می‌کنند. اگر مجموعه داده، چند کلاسه باشد، داده‌ای که بیشترین مقدار را دارد، کلاس اکثریت و باقی به عنوان اقلیت مقداردهی می‌شوند.

[1] Height-balanced tree

فاز دوم: محاسبه فاصله داده‌ها

در این گام، فاصله بین داده‌ها محاسبه می‌شود. همچنین، حداقل فاصله بین داده‌های کلاس‌های موجود در مجموعه داده، مشخص می‌شود. در واقع، این گام مشخص می‌کند که بین دو داده، در دو کلاسِ متفاوت، حداقل چه میزان فاصله وجود دارد؟ جهت تعیین فاصله، از فاصله اقلیدوسی استفاده می‌شود.

$$\text{dist} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ \dots \ + (x_{in} - x_{jn}}$$

فاز دوم: راه‌اندازی الگوریتم Birch

در این فاز، الگوریتم خوشه‌بندی راه‌اندازی و دو پارامتر اصلی این الگوریتم، مقداردهی می‌شود. در این راستا، حداقل فاصله دو کلاسِ مختلف، به عنوان حد آستانه در نظر گرفته می‌شود.

که C تعداد کلاس‌ها و $samples$ تعداد نمونه‌ها را نشان می‌دهد. $dist_{i,j}$ نیز حداقل فاصله نمونه‌های دو کلاس مختلف را از یکدیگر نشان می‌دهد. این حد آستانه، تضمین می‌کند که هیچ یک از داده‌های اقلیت و اکثریت، با هم در یک خوشه قرار نخواهند گرفت. اندازه درخت نیز که به منزله تعداد انشعابات درخت است، با توجّه به تعداد نمونه‌های کلاس اقلیت و اکثریت مشخص می‌گردد و این مقدار برابر است با:

$$\text{B} = 2 \times \left\lceil \frac{Num_{majority}}{Num_{miniority}} \right\rceil$$

که $Num_{majority}$ تعداد نمونه‌های اکثریت، و $Num_{miniority}$ تعداد نمونه‌های اقلیت است. اگر مجموعه داده‌ای چند کلاسه باشد، میانگین کلاس‌های اقلیت در نظر گرفته می‌شود. B نشان‌دهنده فاکتور شاخه (تعداد انشعابات) درخت است. بدین ترتیب، مشخص می‌شود که هر گره برگ چه میزان نود را می‌تواند در خود جای دهد.

فاز سوم: ساخت درخت

در این مرحله، درخت متوازن، با توجّه به حد آستانه و فاکتور انشعاب یا شاخه، ساخته می‌شود. در این درخت، دو نوع عنصر وجود دارد، شامل عنصر والد و عنصر برگ. هر عنصر والد در این درخت، حاوی CF یا ویژگی خوشه و هر عنصر برگ، حاوی نمونه‌ها است. در اینجا CF حاوی سه بخش است: بخش اول، تعداد عناصر موجود در خوشه را نمایش می‌دهد، بخش دوم، میانگین مقادیر نمونه‌های موجود در خوشه را تعیین می‌کند و بخش سوم، برچسب کلاس را نشان می‌دهد. این برچسب با توجه به نمونه‌های درج شده در گره، تعیین می‌گردد. از مقدار موجود در بخش دوم، می‌توان به عنوان سرخوشه نیز استفاده کرد.

جهت ساخت درخت، مراحل زیر صورت می‌پذیرد:

به عنوان گام اول، ابتدا یک ریشه و یک برگ تولید می‌شود.

سپس نمونه‌ها برای قرارگیری در درخت، فراخوانی می‌شوند.

اگر برگ و درنتیجه، نود، ریشه خالی بود، داده به برگ اضافه می‌شود، اطلاعات CF برای برگ و ریشه نیز، به‌روز می‌شود.

در غیر این صورت، الگوریتم، مناسب‌ترین برگ برای درج داده در درخت پیدا کرده و داده را در آن درج می‌کند. فرایند یافتن برگ مناسب، بدین صورت است، که ابتدا توسط ریشه و اطلاعات CF موجود در آن مناسب‌ترین و نزدیک‌ترین فرزند ریشه، انتخاب می‌شود. به عبارت بهتر، به جای مقایسه نمونهٔ جدید با تمام نمونه‌های

$$T = \min_{\substack{i,j \in C \\ i \neq j}} (dist_{i,j})$$

موجود در درخت، این نمونه تنها با CFهای موجود در درخت مقایسه می‌شود. هر برگی که فاصله CF آن تا داده جدید، کمتر از حد آستانه بود، به عنوان برگ کاندید انتخاب شده و نمونهٔ جدید در این برگ، استقرار می‌یابد. اگر تعداد نمونه‌های موجود در برگ، مساوی با فاکتور انشعاب بود، این برگ، موجودیت خود را به والد تغییر داده و دو برگ برای آن تولید می‌شود. همه داده‌های موجود در برگ، به لحاظ فاصله، مورد مقایسه قرار می‌گیرند، نمونه‌ای که میانگینِ فاصله بیشتری نسبت به سایر نودها داشته باشد، در یک برگ و باقی نمونه‌ها، در برگِ دیگر قرار می‌گیرد.

داده‌های موجود در این برگی که نمونه در آن جای گرفته، استخراج شده و CF آن برگ، با احتساب نمونه جدید، به‌روز می‌شود.

اطلاعات CF برای همه والدهای برگ، به‌روز می‌شود.

عمل فراخوانی نمونه‌ها و تولید درخت، تا زمان اتمام نمونه‌ها، ادامه پیدا می‌کند.

بدین ترتیب یک درخت متوازن تولید می‌شود. سپس با توجّه به نمونه‌های موجود در:

فاز چهارم: بررسی درخت و انجام نمونه‌زدایی

در این فاز، خوشه‌های ساخته شده، مورد بررسی قرار می‌گیرد. هر برگ به عنوان یک خوشه عمل می‌کند. در اینجا تنها گره‌های برگ مورد ارزیابی و بررسی قرار می‌گیرد.

ابتدا یک گره برگ، فراخوانی شده، سپس داده‌های موجود در آن به لحاظ برچسب کلاس و همچنین تعداد نمونه، مورد بررسی قرار می-گیرد.

اگر گره برگ باشد، و برچسب کلاس، کلاس اکثریت را نشان دهد، تعداد نمونه‌های آن بررسی می‌شود. اگر تعداد نمونه‌های کلاس از مقدار MajorityNumber بیشتر باشد، باید عمل نمونه‌زدایی صورت پذیرد، یعنی یک یا چند نمونه از این برگ حذف شود. جهت تعیین MajorityNumber به صورت زیر عمل می‌شود:

$$\text{MajorityNumber} = \left[\frac{Num_{majority}}{Num_{miniority}}\right]$$

عمل نمونه‌زدایی تا زمانی انجام می‌شود که تعداد نمونه‌های موجود در برگ برابر با MajorityNumber شود.

اگر گره برگ باشد، و برچسب کلاس، کلاس اکثریت را نشان دهد و تعداد نمونه‌های آن کمتر از مقدار MajorityNumber باشد، به معنی این است که داده‌های موجود در این برگ از موقعیت استراتژیک برخوردارند و به عبارت بهتر، در اقلیت هستند؛ بنابراین، باید عمل نمونه‌زایی صورت پذیرد. یعنی یک یا چند نمونه به این برگ اضافه شود.

اگر گره برگ باشد، و برچسب کلاس، کلاس اقلیت را نشان دهد، تعداد نمونه‌های آن بررسی می‌شود. اگر تعداد نمونه‌های موجود در برگ از مقدار MajorityNumber کمتر باشد، باید عمل نمونه‌زدایی صورت پذیرد.

در غیر این صورت، داده‌های موجود در گره برگ دست نمی‌خورد و عمل نمونه‌زایی یا عمل نمونه‌زدایی، صورت می‌گیرد.

فاز پنجم: نمونه‌زایی– نمونه زدایی

برای تمام گره‌هایی که برای نمونه‌زایی انتخاب شدند (چه از کلاس اقلیت باشند و چه از کلاس اکثریت،) عمل زیر انجام می‌شود:

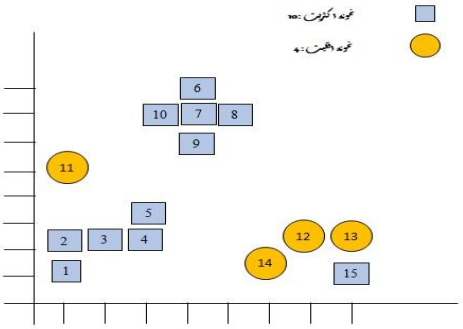
$$X_{new} = x_i + r$$

که x_i، نمونهٔ موجود در کلاس، $rand$ یک عدد تصادفی است که با توجّه به دامنه تغییرات هر ویژگی در مجموعه داده، مشخص می-شود. برای مثال، اگر برای یک ویژگی، داده‌ها بین مقدار ۰ و ۱ باشد، این مقدار تصادفی نیز بین ۰ تا ۱ تعیین می‌شود و اگر بین ۰ تا ۲۰۰ باشد، $rand$ بین ۰ تا ۲۰۰ است.

اگر مقدار فاصله X_{new} تا نمونه x_i از حد آستانه کمتر بود، نمونه X_{new}، به عنوان نمونه جدید به مجموعه داده اضافه می‌شود. در غیر این صورت، عمل تولید نمونه تکرار می‌شود، تا زمانی که به اندازه مشخص شده، نمونه تولید شود. برای گره‌های برگِ اقلیت، این مقدار برابر با Num_Sample_{leaf} و برای گره‌های اکثریت، برابر با num_{sample} است.

جهت نمونه‌زدایی نیز، ابتدا فاصله داده‌های موجود در برگ، محاسبه می‌شود. سپس برای هر نمونه، میانگین فواصل، محاسبه می‌شود. در نهایت نمونه یا نمونه‌هایی که کمترین میانگین فاصله را دارند، انتخاب و از مجموعه داده و گره برگ حذف می‌شوند.

یک مثال بارز و روشن از عمل نمونه‌زایی و نمونه‌زدایی:



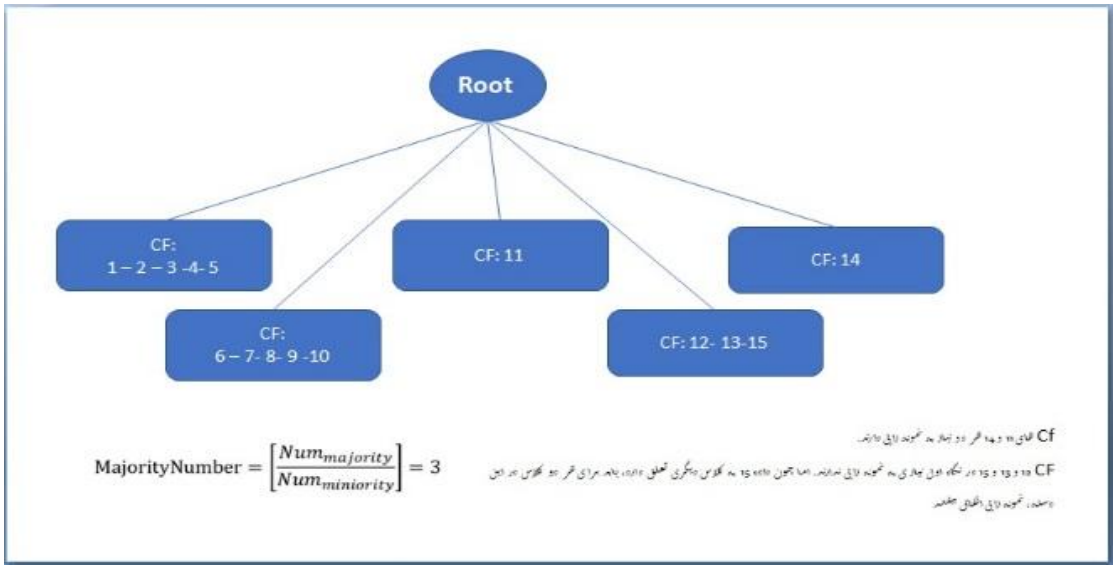
ID	X	Y	Class
1	1	1	Square
2	1	2	Square
3	2	2	Square
4	3	2	Square
5	3	3	Square
6	4	8	Square
7	4	7	Square
8	5	7	Square
9	4	6	Square
10	3	7	Square
11	1	5	Circle
12	7	2	Circle
13	8	2	Circle
14	6	1	Circle
15	8	1	Square

T = Minimum distance = 1

$$B = 2 \times \left[\frac{Num_{majority}}{Num_{miniority}}\right] = 5$$

شکل ۱

	data1	data2	data3	data4	data5	data6	data7	data8	data9	data10	data11	data12	data13	data14	data15
data1	0.00	1.00	1.41	2.24	2.83	7.62	6.71	7.21	5.83	6.32	4.00	6.08	7.07	5.00	7.00
data2	1.00	0.00	1.00	2.00	2.24	6.71	5.83	6.40	5.00	5.39	3.00	6.00	7.00	5.10	7.07
data3	1.41	1.00	0.00	1.00	1.41	6.32	5.39	5.83	4.47	5.10	3.16	5.00	6.00	4.12	6.08
data4	2.24	2.00	1.00	0.00	1.00	6.08	5.10	5.39	4.12	5.00	3.61	4.00	5.00	3.16	5.10
data5	2.83	2.24	1.41	1.00	0.00	5.10	4.12	4.47	3.16	4.00	2.83	4.12	5.10	3.61	5.39
data6	7.62	6.71	6.32	6.08	5.10	0.00	1.00	1.41	2.00	1.41	4.24	6.71	7.21	7.28	8.06
data7	6.71	5.83	5.39	5.10	4.12	1.00	0.00	1.00	1.00	1.00	3.61	5.83	6.40	6.32	7.21
data8	7.21	6.40	5.83	5.39	4.47	1.41	1.00	0.00	1.41	2.00	4.47	5.39	5.83	6.08	6.71
data9	5.83	5.00	4.47	4.12	3.16	2.00	1.00	1.41	0.00	1.41	3.16	5.00	5.66	5.39	6.40
data10	6.32	5.39	5.10	5.00	4.00	1.41	1.00	2.00	1.41	0.00	2.83	6.40	7.07	6.71	7.81
data11	4.00	3.00	3.16	3.61	2.83	4.24	3.61	4.47	3.16	2.83	0.00	6.71	7.62	6.40	8.06
data12	6.08	6.00	5.00	4.00	4.12	6.71	5.83	5.39	5.00	6.40	6.71	0.00	1.00	1.41	1.41
data13	7.07	7.00	6.00	5.00	5.10	7.21	6.40	5.83	5.66	7.07	7.62	1.00	0.00	2.24	1.00
data14	5.00	5.10	4.12	3.16	3.61	7.28	6.32	6.08	5.39	6.71	6.40	1.41	2.24	0.00	2.00
data15	7.00	7.07	6.08	5.10	5.39	8.06	7.21	6.71	6.40	7.81	8.06	1.41	1.00	2.00	0.00

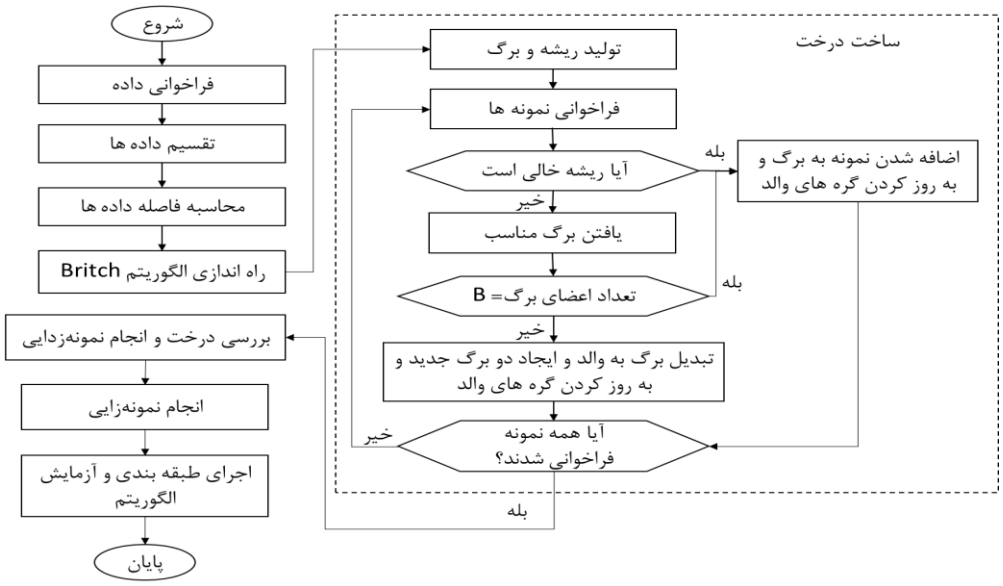


شکل ۲

فاز ششم: طبقه‌بندی

در این فاز با استفاده از یکی از روش‌های طبقه‌بندی همچون k نزدیک‌ترین همسایه، ماشین بردار پشتیبان (SVM) و ... مجموعه داده جدید، مورد آزمایش و بررسی قرار می‌گیرد.

در ادامه در شکل (۳) فلوچارت الگوریتم پیشنهادی آمده است.



شکل ۳: الگوریتم روش پیشنهادی

۵. نتایج و ارزیابی

جهت پیاده‌سازی الگوریتم پیشنهادی از نرم‌افزار متلب۴نسخه ۲۰۱۸ استفاده شده است. مجموعه داده‌های مورد استفاده در این پژوهش در جدول ۱ نشان داده شده است.

جدول ۱: نمایش مجموعه داده‌های نامتوازن

نام مجموعه داده	تعداد ویژگی‌ها	تعداد نمونه‌ها	تعداد کلاس	حداقل تعداد نمونه اقلیت	تعداد نمونه اکثریت	نرخ عدم توازن
Pima	8	768	2	268	500	0.466
Breast	9	699	2	241	458	0.474
German	10	1000	2	300	700	0.572
Haber-man	3	306	2	81	235	0.655
Blood Transfusin	5	748	2	178	570	0.788
ionosphere	34	241	2	126	225	0.44
Yeast3	8	1484	2	163	1341	0.979
Glass	9	214	6	17	197	0.914
lymphography	18	148	4	2	81	0.976
Ecoli	7	336	8	2	143	0.986
Letter	16	20000	26	447	813	0.45

جهت ارزیابی نیز، روش نمونه‌زایی و نمونه‌زدایی پیشنهادی، با روش-ها و الگوریتم‌های دیگر در این حیطه مورد مقایسه قرار می‌گیرد. این روش‌ها و الگوریتم‌ها در جدول (۲) آمده است.

جدول ۲: معرفی روش‌ها و الگوریتم‌های موجود برای ارزیابی

شماره مرجع	نام روش	سال
[۱]	SMOTE	۲۰۰۲
[۵۱]	Bor-SMOT	۲۰۰۵

14 Matlab 2018b.

[۵۲]	Bagging	۲۰۱۵
[۵۳]	AdaBoost	۲۰۱۳
[۵۴]	AdaBoost2	۲۰۱۸

در جدول‌های ۳ تا ۸، نتایج به‌دست آمده براساس معیارهای دقت، recall،FNR ، تشخیص، صحت و F آورده شده است.

جدول ۳: ارزیابی الگوریتم پیشنهادی به لحاظ معیار دقت.

نام مجموعه داده	Main-data		SMOTE		Bor-SMOT		Bagging	AdaBoost	AdaBoost2	birch-resample	
	Knn	Svm	Knn	Svm	Knn	Svm		-		Knn	svm
Pima	۰/۶۷۵۳	۰/۷۶۷	۰/۶۷۵۳	۰/۷۶۷	۰/۶۷۵۳	۰/۷۸۷	۰/۷۵۸۴	۰/۷۲۹۹	۰/۷۶۱	۰/۷۶۷۹	۰/۷۸۷
Breast	۰/۹۶۸۶	۰/۹۷۷۱	۰/۹۶۸۶	۰/۹۷۴۳	۰/۹۶۸۶	۰/۹۷۴۳	۰/۹۸	۰/۹۷۴۳	۰/۹۷۴۳	۰/۹۶۵۷	۰/۹۸۵۷
German	۰/۵۱۲۸	۰/۵۵۶۷	۰/۶۱۲۸	۰/۵۴۲۹	۰/۶۱۲۸	۰/۴۳۵۱	۰/۵۷۴۵	۰/۵۴۴۵	۰/۵۶۶۵	۰/۶۱۲۸	۰/۶۸۶۶
Haber-man	۰/۶۲۹۹	۰/۶۲۹۹	۰/۶۲۹۹	۰/۶۲۱۸	۰/۶۲۹۹	۰/۶۲۶۲	۰/۶۲۱۸	۰/۶۱۸۸	۰/۶۲۱۸	۰/۶۲۳۴	۰/۷۵۳۲
Blood Transfusion	۰/۶۱۲۷	۰/۵۱۴۳	۰/۷۶	۰/۷۵۷۳	۰/۷۶	۰/۷۷۳۳	۰/۷۶۲۷	۰/۷۶۵۳	۰/۷۷۰۷	۰/۷۶	۰/۷۷۳۳
ionosphere	۰/۵۶۳۶	۰/۵۰۳۴	۰/۸۶۳۶	۰/۵۰۳۴	۰/۸۶۳۶	۰/۸۲۵	۰/۸۳۷۵	۰/۸۳۷۵	۰/۸۵۴۵	۰/۸۶۹۳	۰/۸۸۰۷
Yeast3	۰/۸۱۹۲	۰/۸۹۷۷	۰/۸۱۹۲	۰/۸۹۷۷	۰/۹۱۹۲	۰/۸۹۵	۰/۸۴۴۸	۰/۸۳۹۴	۰/۸۴۳۵	۰/۹۱۹۲	۰/۹۳۸۱
Glass	۰/۸۲۸۸	۰/۷۷۳۹	۰/۸۳۳۹	۰/۸۱۰۲	۰/۸۳۳۹	۰/۸۱۰۲	۰/۸۴۳۳	۰/۸۷۰۳	۰/۸۴۶	۰/۸۷۰۹	۰/۸۷۵۸
Lympho-graphy	۰/۸۲۲۴	۰/۸۱۸۹	۰/۸۲۲۴	۰/۸۱۸۹	۰/۸۱۸۹	۰/۸۲۲۴	۰/۸۲۴۷	۰/۸۲۲۴	۰/۸۲۵۹	۰/۸۵۶۳	۰/۸۵
Ecoli	۰/۹۴۰۷	۰/۷۵۸	۰/۸۴۰۷	۰/۸۷۵	۰/۹۴۰۷	۰/۸۴۰۷	۰/۸۷۲۶	۰/۸۴۶۷	۰/۷۶۵۵	۰/۹۴۹۴	۰/۹۵۵۲
Letter	۰/۹۷۹۶	۰/۹۷۹۵	۰/۹۹۹۶	۰/۹۹۹۵	۰/۹۹۹۶	۰/۹۹۹۵	۰/۹۹۹۶	۰/۹۹۹۷	۰/۹۹۹۶	۰/۹۸۵۸	۰/۹۸۶۶

جدول ۴: ارزیابی الگوریتم پیشنهادی به لحاظ معیار recall

نام مجموعه داده	Main-data		SMOTE		Bor-SMOT		Bagging	AdaBoost	AdaBoost2	birch-resample	
	Knn	Svm	Knn	Svm	Knn	Svm		-		Knn	svm
Pima	۰/۷۶۸۲	۰/۷۸۵۷	۰/۷۶۸۲	۰/۷۸۳۸	۰/۷۶۸۲	۰/۷۸۱۹	۰/۷۷۵۴	۰/۷۸۲۹	۰/۷۷۲۴	۰/۷۷۱۶	۰/۸۶۰۹
Breast	۰/۹۵۸۷	۰/۹۵۴۶	۰/۹۶۵۸	۰/۹۵۸۷	۰/۹۵۵۸	۰/۹۵۸۷	۰/۹۶۲۹	۰/۹۷۴۴	۰/۹۶۳۲	۰/۹۶۵۵	۰/۹۷۵۲
German	۰/۷۱۶۷	۰/۷۳۹۲	۰/۷۱۶۷	۰/۷۴۹	۰/۷۱۶۷	۰/۶۸۹۳	۰/۷۵۶۸	۰/۷۵۶۸	۰/۷۵۶۲	۰/۷۱۶۷	۰/۷۶۲۹
Haber-man	۰/۸۴۸۱	۰/۷۴۳۱	۰/۸۴۸۱	۰/۷۴۳۱	۰/۸۴۸۱	۰	۰/۸۲۴۷	۰/۷۹۶۱	۰/۸۰۸۱	۰/۸۵۵۳	۰/۸۱۳۶
Blood Transfusin	۰/۷۸۲۶	۰/۷۸۵۳	۰/۷۸۲۶	۰/۷۸۵۳	۰/۷۸۲۶	۰/۷۷۹۳	۰/۷۴۳۹	۰/۷۴۳۹	۰/۷۲۵۱	۰/۷۸۳۴	۰/۷۲۵۷
ionosphere	۰/۸۲۳۵	۰/۸۷۴	۰/۸۲۳۵	۰/۸۷۴	۰/۸۲۳۵	۰/۸۰۱۸	۰/۸۳۱۶	۰/۸۴۶۹	۰/۸۴۰۷	۰/۸۲۹۶	۰/۸۸۸۹
Yeast3	۰/۸۵۱۸	۰/۸۹۷۸	۰/۸۵۱۸	۰/۸۹۷۸	۰/۸۵۱۸	۰/۹۱۳۴	۰/۹۱۷۲	۰/۹۱۶۹	۰/۹۴۱۷	۰/۹۵۷۳	۰/۹۵۱۵
Glass	۰/۶۲۱۶	۰/۴۸۶۵	۱	۱	۱	۱	۱	۰	۱	۰/۶۱۲۶	۰/۴۷۷۵
Lympho graphy	۱	۱	۱	۱	۰/۸۹۷۸	۱	۱	۱	۱	۰/۴۲۵	۰/۴
Ecoli	۰/۷۷۷۸	۰/۵۳۴۹	۰/۷۷۷۸	۰/۵۳۴۹	۰/۷۷۷۸	۰/۷۷۷۸	۰/۹۲	۰/۷۸۵۷	۰/۸۸۴۶	۰/۷۹۷۷	۰/۸۲۰۸
Letter	۰/۵	۰/۲۵	۰/۵	۰/۲۵	۰/۵	۰/۲۵	۰/۳۳۳۳	۰	۰/۳۳۳۳	۰/۸۱۵۳	۰/۸۲۵۴

جدول ۵: ارزیابی الگوریتم پیشنهادی به لحاظ معیار FNR

نام مجموعه داده	Main-data		SMOTE		Bor-SMOT		Bagging	AdaBoost	AdaBoost2	birch-resample	
	Knn	Svm	Knn	Svm	Knn	Svm		-		Knn	svm
Pima	۰/۴۶۷۱	۰/۲۰۸۸	۰/۴۶۷۱	۰/۲۰۲۲	۰/۴۶۷۱	۰/۱۹۵۴	۰/۲۹	۰/۳۷۸۰	۰/۲۷۳۷	۰/۴۶۴۱	۰/۳۳۵۵
Breast	۰/۰۲۱۶	۰/۰۳۰۰	۰/۰۳۰۰	۰/۰۱۷۵	۰/۰۳	۰/۰۱۷۵	۰/۰۲۱۵	۰/۰۲۵۸	۰/۰۱۳۰	۰/۰۳۴۲	۰/۰۰۸۷
German	۰/۶	۰/۵	۰/۶۵۲۵	۰/۶۵۱۲	۰/۶۵۲۵	۰/۷۰۳۷	۰/۱۵۴۷	۰/۴۰۹۴	۰/۳۱۹۱	۰/۶۵۲۵	۰/۵۲۵۴
Haber-man	۰/۰۷۵۷	۰/۱۰۲۷	۰/۶	۰/۵۰۰۰	۰/۶	۰/۷۳۲۰	۰/۵۶۱۴	۰/۵۸۸۲	۰/۵۸۱۸	۰/۶۰۲۶	۰/۴۴۴۴
Blood Transfusion	۰/۵	۰/۵۱۴۳	۰/۵	۰/۵۱۴۳	۰/۵	۰/۳۵۲۹	۰/۴۸۳۹	۰/۴۶۴۳	۰/۴۲۳۱	۰/۵	۰/۷۷۴۸
ionosphere	۰	۰/۰۲۰۴	۰	۰/۰۲۰۴	۰	۰/۱۷۱۹	۰/۰۵۰۸	۰/۰۷۹۴	۰/۰۱۷۲	۰	۰/۱۳۵۶
Yeast3	۰/۳۵۴۴	۰/۱۱۱۱	۰/۳۵۴۴	۰/۱۱۱۱	۰/۳۵۴۴	۰/۴۸۲۳	۰/۱۸۱۸	۰/۲۲۸۶	۰/۱۹۴۰	۰/۳۶۷۸	۰/۲۰۶۳
Glass	۰/۰۷۵۷	۰/۱۰۲۷	۰/۰۷۵۷	۰/۱۰۲۷	۰/۰۷۵۷	۰/۱۰۲۷	۰/۰۶۴۹	۰/۱۲۹۷	۰/۰۶۱۳	۰/۰۷۷۵	۰/۱۰۴۵
Lympho graphy	۰/۰۷۸۶	۰/۹۲۹۳	۰/۰۷۸۶	۰/۰۸۲۱	۰/۸۲۱	۰/۰۷۸۶	۰/۰۷۶۸	۰/۰۷۸۶	۰/۰۷۵۰	۰/۰۸۲۱	۰/۰۸۷۵
Ecoli	۰/۰۳۳۹	۰/۰۴۰۵	۰/۰۳۳۹	۰/۰۴۰۵	۰/۰۳۳۹	۰/۰۳۳۹	۰/۰۱۹۸	۰/۰۲۷۳	۰/۰۲۲۳	۰/۰۲۸۹	۰/۰۲۵۶
Letter	۰/۰۰۴	۰/۰۰۵	۰/۰۰۲	۰/۰۰۰۲	۰/۰۰۲	۰/۰۰۰۲	۰/۰۰۴	۰/۰۰۳	۰/۰۰۰۴	۰/۰۰۷۴	۰/۰۰۷۰

جدول۶ : ارزیابی الگوریتم پیشنهادی به لحاظ معیار تشخیص.

نام مجموعه داده	Main-data		SMOTE		Bor-SMOT		Bagging	AdaBoost	AdaBoost2	birch-resample	
	Knn	Svm	Knn	Svm	Knn	Svm		-		Knn	svm
Pima	۰/۵۳۲۹	۰/۷۹۱۲	۰/۵۳۲۹	۰/۷۹۷۸	۰/۵۳۲۹	۰/۸۰۴۶	۰/۷۱	۰/۶۲۲	۰/۷۲۶۳	۰/۵۳۵۹	۰/۶۶۴۵
Breast	۰/۹۷	۰/۹۷۸۴	۰/۹۷	۰/۹۸۲۵	۰/۹۷	۰/۹۸۲۵	۰/۹۷۸۵	۰/۹۷۴۲	۰/۹۸۷	۰/۹۶۵۸	۰/۹۹۱۳
German	۰/۳۴۷۵	۰/۴۱۸۶	۰/۳۴۷۵	۰/۳۴۸۸	۰/۳۴۷۵	۰/۲۹۶۳	۰/۶۹۳۹	۰/۵۹۰۶	۰/۶۸۰۹	۰/۳۴۷۵	۰/۴۷۷۶
Haber-man	۰/۴	۰/۵	۰/۴	۰/۵	۰/۴	۰/۲۶۸	۰/۴۳۸۶	۰/۴۱۱۸	۰/۴۱۸۲	۰/۳۹۷۴	۰/۵۵۵۶
Blood Transfusio	۰/۵	۰/۴۸۵۷	۰/۵	۰/۴۸۵۷	۰/۵	۰/۶۴۷۱	۰/۵۱۶۱	۰/۵۳۵۷	۰/۵۷۶۹	۰/۵	۰/۲۲۵۲
ionosphere	۰/۸۲	۰/۸۷۹۶	۰/۸۵۹۷	۰/۹۲۹۶	۰/۸۵۹۷	۰/۸۲۸۱	۰/۹۴۹۲	۰/۹۲۰۶	۰/۹۲۲۸	۱	۰/۹۶۴۴
Yeast3	۰/۶۴۵۶	۰/۶۸۸۹	۰/۶۴۵۶	۰/۶۸۸۹	۰/۶۴۵۶	۰/۵۱۷۷	۰/۶۱۸۲	۰/۷۷۱۴	۰/۶۰۶	۰/۶۳۲۲	۰/۷۹۳۷
Glass	۰/۸۲۴۳	۰/۸۹۷۳	۰/۸۲۴۳	۰/۸۹۷۳	۰/۸۲۴۳	۰/۸۲۷۳	۰/۸۳۵۱	۰/۷۸۰۳	۰/۸۳۸۷	۰/۹۲۲۵	۰/۸۹۵۵
Lympho graphy	۰/۸۲۱۴	۰/۸۱۷۹	۰/۸۶۲۷	۰/۸۱۷۹	۰/۸۱۷۹	۰/۸۶۲۷	۰/۸۲۳۲	۰/۸۲۱۴	۰/۸۲۵	۰/۹۱۷۹	۰/۹۱۴۳
Ecoli	۰/۸۶۶۱	۰/۸۵۹۵	۰/۹۶۶۱	۰/۹۵۹۵	۰/۹۶۶۱	۰/۹۶۶۱	۰/۹۳۰۲	۰/۹۱۲۷	۰/۹۲۷۷	۰/۹۷۱۱	۰/۹۷۴۴
Letter	۰/۲۹۹۸	۰/۲۵	۰/۹۸۹۸	۰/۹۸۹۸	۰/۹۸۹۸	۰/۹۸۹۸	۰/۹۸۹۸	۰/۹۸۶۷	۰/۹۸۹۸	۰/۹۹۲۶	۰/۹۹۳

جدول ۷: ارزیابی الگوریتم پیشنهادی به لحاظ معیار صحت.

نام مجموعه داده	Main-data		SMOTE		Bor-SMOT		Bagging	AdaBoost	AdaBoost2	birch-resample	
	Knn	Svm	Knn	Svm	Knn	Svm		-		Knn	svm
Pima	۰/۲۸۴	۰/۰۷۶	۰/۲۸۴	۰/۰۷۲	۰/۲۸۴	۰/۰۶۸	۰/۱۱۶	۰/۱۹۲	۰/۱۰۴	۰/۲۸۴	۰/۳۰۸
Breast	۰/۰۵۸۳	۰/۰۴۱۷	۰/۰۵۸۳	۰/۰۳۳۳	۰/۰۵۸۳	۰/۰۳۳۳	۰/۰۴۱۷	۰/۰۵	۰/۰۲۵	۰/۰۶۶۷	۰/۱۱۶۷
German	۰/۲۶۲۹	۰/۲۱۴۳	۰/۲۶۲۹	۰/۴۸	۰/۲۶۲۹	۰/۶۵۱۴	۰/۰۸۵۷	۰/۱۴۸۶	۰/۰۸۵۷	۰/۲۶۲۹	۰/۳
Haber-man	۰/۰۴۱۸	۰/۰۴۴۶	۰/۴۰۱۸	۰/۰۴۴۶	۰/۴۰۱۸	۱	۰/۲۸۵۷	۰/۲۶۷۹	۰/۲۸۵۷	۰/۴۱۹۶	۰/۴۴۲۹
Blood Transfusion	۰/۰۵۲۶	۰/۰۶۳۲	۰/۰۵۲۶	۰/۰۶۳۲	۰/۰۵۲۶	۰/۰۲۱۱	۰/۰۵۲۶	۰/۰۴۵۶	۰/۰۳۸۶	۰/۵۰۶۱	۰/۷۱۲۳
ionosphere	۰	۰/۰۰۸۹	۰	۰/۰۰۸۹	۰	۰/۰۹۸۲	۰/۰۲۶۸	۰/۰۴۴۶	۰/۰۰۸۹	۰/۰۱۲۴	۰/۰۷۱۴
Yeast3	۰/۰۴۲۴	۰/۰۰۱۵	۰/۰۴۲۴	۰/۰۰۱۵	۰/۰۴۲۴	۰/۱۰۳	۰/۰۱۸۲	۰/۰۲۴۲	۰/۰۱۹۷	۰/۰۴۸۵	۰/۰۴۹۷
Glass	۰/۳۷۸۴	۰/۵۱۳۵	۰/۳۴۴۳	۰/۴۱۶۱	۰/۳۴۴۳	۰/۴۱۶۱	۰/۳۱۰۳	۰/۴۹۹۶	۰/۴۹۹۶	۰/۳۸۷۴	۰/۵۲۲۵
Lympho graphy	۰/۸۶۲۷	۰/۸۶۷۹	۰/۸۶۲۷	۰/۸۶۷۹	۰/۸۶۲۷	۰/۸۶۲۷	۰/۸۶	۰/۸۶۲۷	۰/۸۵۷۱	۰/۵۷۵	۰/۶
Ecoli	۰/۲۱۸۱	۰/۲۳۳۳	۰/۲۱۸۱	۰/۲۳۳۳	۰/۲۱۸۱	۰/۲۱۸۱	۰/۱۲۹۷	۰/۱۷۶۵	۰/۱۴۳۶	۰/۲۰۲۳	۰/۲۷۹۲

جدول ۸: ارزیابی الگوریتم پیشنهادی به لحاظ معیار F

نام مجموعه داده	Main-data		SMOTE		Bor-SMOT		Bagging	AdaBoost	AdaBoost2	birch-resample	
	Knn	Svm	Knn	Svm	Knn	Svm		-		Knn	svm
Pima	۰/۴۱۴۷	۰/۱۳۸۶	۰/۴۱۴۷	۰/۱۳۱۹	۰/۴۱۴۷	۰/۱۲۵۱	۰/۲۰۱۸	۰/۳۰۸۴	۰/۱۸۳۳	۰/۴۱۵۲	۰/۴۵۳۷
Breast	۰/۱۰۹۹	۰/۰۷۹۹	۰/۱۱۰۰	۰/۰۶۴۴	۰/۱۰۹۹	۰/۰۶۴۴	۰/۰۷۹۹	۰/۰۹۵۱	۰/۰۴۸۷	۰/۱۲۴۸	۰/۲۰۸۵
German	۰/۳۸۴۷	۰/۳۳۲۳	۰/۳۸۴۷	۰/۵۸۵۱	۰/۳۸۴۷	۰/۶۶۹۸	۰/۱۵۴۰	۰/۲۴۸۴	۰/۱۵۴۰	۰/۳۸۴۷	۰/۴۳۰۷
Haber-man	۰/۵۴۵۳	۰/۰۸۴۱	۰/۵۴۵۳	۰/۰۸۴۱	۰/۵۴۵۳	۰/۰۰۰۰	۰/۴۲۴۴	۰/۴۰۰۹	۰/۴۲۲۲	۰/۵۶۳۰	۰/۵۷۳۶
Blood Transfusion	۰/۰۹۸۶	۰/۱۱۷۰	۰/۰۹۸۶	۰/۱۱۷۰	۰/۰۹۸۶	۰/۰۴۱۱	۰/۰۹۸۳	۰/۰۸۵۹	۰/۰۷۳۳	۰/۶۱۴۹	۰/۷۱۸۹
ionosphere	۰/۰۰۰۰	۰/۰۱۷۶	۰/۰۰۰۰	۰/۰۱۷۶	۰/۰۰۰۰	۰/۱۷۵۰	۰/۰۵۱۹	۰/۰۸۴۷	۰/۰۱۷۶	۰/۰۲۴۴	۰/۱۳۲۲
Yeast3	۰/۰۸۰۸	۰/۰۰۳۰	۰/۰۸۰۸	۰/۰۰۳۰	۰/۰۸۰۸	۰/۱۸۵۱	۰/۰۳۵۷	۰/۰۴۷۲	۰/۰۳۸۶	۰/۰۹۲۳	۰/۰۹۴۵
Glass	۰/۴۷۰۴	۰/۴۹۹۶	۰/۵۱۲۲	۰/۵۸۷۷	۰/۵۱۲۲	۰/۵۸۷۷	۰/۴۷۳۶	۰/۰۰۰۰	۰/۶۶۶۳	۰/۴۷۴۶	۰/۴۹۹۰
Lympho graphy	۰/۹۲۶۳	۰/۹۲۹۳	۰/۹۲۶۳	۰/۹۲۹۳	۰/۸۷۹۹	۰/۹۲۶۳	۰/۹۲۴۷	۰/۹۲۶۳	۰/۹۲۳۱	۰/۴۸۸۸	۰/۴۸۰۰
Ecoli	۰/۳۴۰۷	۰/۳۲۴۹	۰/۳۴۰۷	۰/۳۲۴۹	۰/۳۴۰۷	۰/۳۴۰۷	۰/۲۲۷۳	۰/۲۸۸۲	۰/۲۴۷۱	۰/۳۲۲۷	۰/۴۱۶۷
Letter	۰/۵۲۱۸	۰/۳۶۳۶	۰/۵۲۱۸	۰/۳۶۳۶	۰/۵۲۱۸	۰/۳۶۳۶	۰/۴۴۶۳	۰/۰۰۰۰	۰/۴۴۸۱	۰/۳۰۱۲	۰/۲۸۸۲

با تجمیع نتایج فوق، می‌توان گفت: الگوریتم پیشنهادی در مواجهه با داده‌های دو کلاسه و چند کلاسه که نرخ عدم توازن کمی دارند.

۶. نتیجه‌گیری و پیشنهادات

این پژوهش، رویکردی جهت متوازن نمودن مجموعه داده‌های نامتوازن ارائه می‌کند. هدف اصلی این پژوهش، ارائه یک روش کارآمد برای نمونه‌زدایی در مجموعه داده اکثریت در داده نامتوازن، و ارائه یک روش کارآمد برای نمونه‌زایی در مجموعه داده اقلیت در داده نامتوازن، بوده است، لذا، در این پژوهش سعی شد تا دو روش نمونه‌زدایی و نمونه‌زایی با یکدیگر ادغام شوند و به گونه‌ای کارا و مؤثر، به متوازن‌کردن مجموعه داده‌های نامتوازن بپردازند.

روش اصلی این پژوهش، متکی بر الگوریتم خوشه‌بندی birch است. این الگوریتم، برخلاف سایر الگوریتم‌های دیگر که به نویز حساس هستند، حساسیت کمتری نسبت به نویز داشته. علاوه براین، خود به صورت خودکار اقدام به تولید دسته و خوشه جدید برای مجموعه داده می‌کند و از این رو، نیازی به تعیین تعداد خوشه از ابتدای راه‌اندازی الگوریتم نیست. همین امر، سبب منعطف شدن این الگوریتم و توانایی کار با مجموعه داده‌های مختلف شده است.

علاوه براین، نکته دیگری که می‌توان آن را به عنوان نوآوری این تحقیق دانست، این است که رویکرد پیشنهادی به جای تمرکز بر کلاس‌های اقلیت و اکثریت، بر روی داده‌های اقلیت و اکثریت کار می‌کند. به عبارتی بهتر، در مجموعه داده نامتوازن، همه نمونه‌ها دارای یک وزن و ارزش یکسان نیستند. برخی از نمونه‌ها، به لحاظ موقعیت و قرارگیری در فضای نمونه‌ها، دارای موقعیت استراتژیک می‌تواند به خوبی عمل کند و داده‌های این نوع مجموعه داده‌ها را به خوبی شناسایی کند. اما با افزایش نرخ عدم توازن، یا کاهش تعداد ویژگی‌ها، عملکرد الگوریتم نیز کاهش می‌یابد.

و حساسی هستند. حذف چنین نمونه‌ای در عملیات نمونه‌زدایی، شاید بتواند تا حدی نتیجه طبقه‌بندی را بهبود دهد و طبقه‌بندی را متعادل کند، اما موجب از دست رفتن داده‌ای ارزشمند نیز شده است، بنابراین، الگوریتم پیشنهادی، ابتدا موقعیت و وضعیت هر گره را مشخص کرده و سپس در صورت نیاز آن را مورد عمل نمونه‌زایی و یا نمونه‌زدایی قرار می‌دهد.

جهت پیاده‌سازی روش پیشنهادی، از نرم‌افزار متلب استفاده شد. جهت اجرای الگوریتم و بررسی کارایی آن نیز، از ۱۱ مجموعه داده شامل Pima، Breast، German، Haber-man، Blood Transfusion، ionosphere، Yeast3، Glass، lymphography، Ecoli، Letter استفاده گردید. این مجموعه داده‌ها، دوکلاسه و چند کلاسه، با نرخ عدم توازن مختلف بودند.

جهت مقایسه عملکرد الگوریتم پیشنهادی نیز، این الگوریتم با پنج روش شامل، SMOTE، Bor-SMOT، Bagging، AdaBoost، AdaBoost2 مورد ارزیابی و مقایسه قرار گرفت. جهت انجام این مقایسه، پس از اجرای الگوریتم‌ها، ماتریس در هم ریختگی، تشکیل شد و با استفاده از آن مقادیرِ معیارهای مثبت درست TP، منفی نادرست FN، منفی درست TN، مثبت نادرستFP مشخص شد. با استفاده از این مقادیر نیز معیارهای اصلی برای ارزیابی، محاسبه شدند. معیارهای اصلی برای ارزیابی عملکرد الگوریتم طبقه‌بندی عبارت است از: دقت، معیار recall، نرخ منفی نادرست FNR، معیار تشخیص، معیار صحت و معیار F بودند.

با توجه به نتایج کسب شده، می‌توان گفت، الگوریتم پیشنهادی در مواجهه با داده‌های دو کلاسه و چند کلاسه که نرخ عدم توازن کمی دارند، می‌تواند به خوبی عمل کند و داده‌های این نوع مجموعه داده‌ها را به خوبی شناسایی کند. اما با افزایش نرخ عدم توازن، یا کاهش تعداد ویژگی‌ها، عملکرد الگوریتم نیز کاهش می-یابد.

هر چند الگوریتم پیشنهادی، نتایج قابل قبولی ارائه کرده است، اما هنوز چالش‌هایی وجود دارد که می‌تواند به عنوان مسیر باز تحقیق مطرح شود؛ الگوریتم پیشنهادی از الگوریتم طبقه‌بندی knn و SVM جهت طبقه‌بندی نمونه‌های تولید شده استفاده کرده است. حال، بررسی عملکرد این الگوریتم، با سایر روش‌های طبقه‌بندی موضوعی است که می‌تواند به عنوان مسیر آتی تحقیق عنوان شود.

مطابق با نتایج کسب شده، با افزایش نرخ عدم توازن، کارایی الگوریتم پیشنهادی نیز، کاهش می‌یابد، بنابراین، موضوع دیگر که می‌تواند به عنوان مسیر آتی تحقیق مطرح شود، بهبود الگوریتم برای کار با داده‌های دارای نرخ عدم توازن بالا است.

در این الگوریتم از روشی مشابه با روش SMOTE جهت تولید داده‌ها و فرایند نمونه‌زایی استفاده شد. در حالی که در زمان تولید نمونه، باید به ویژگی‌ها، دامنه تغییر آن‌ها، نوع آن‌ها (کمی- کیفی- گسسته – پیوسته بودن متغیر) توجّه داشت و بر اساس آن‌ها، اقدام به تولید نمونه جدید کرد. ارائه یک معیار جدید، جهت تولید نمونه، موضوعی دیگر است که باید به آن پرداخته شود.

منابع

[1]. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique", Journal of artificial intelligence research, vol. 16, pp. 321-357, 2002.

[2]. Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu, and Y. Zhou, "A novel ensemble method for classifying imbalanced data", Pattern Recognition, vol. 48, pp. 1623-1637, 2015.

[۳] ر. بهشتی‌نیا، ه. همایونی، ا. امینی لاری، ه. منصور، "بررسی داده‌های نامتوازن و انتخاب گروه مناسب طبقه‌بندی با استفاده از روش‌های جمعی برای داده‌های مربوط به سرطان سینه"، آخرین پیشرفت‌ها در مهندسی برق، الکترونیک، کامپیوتر. ۱۳۹۶.

[4]. M. A. Tahir, J. Kittler, and F. Yan, "Inverse random under sampling for class imbalance problem and its application to multi-label classification", Pattern Recognition, vol. 45, pp. 3738-3750, 2012.

[5]. M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 42, pp. 463-484, 2012.

[۶]. ه. بنایی، ح. خ. نیت، "نقش و کاربرد هوش عملیاتی و داده کاوی در کشف تقلب برخط"، ششمین کنفرانس ملی تجارت و اقتصاد الکترونیکی، ۱۳۹۰.

[۷]. م. ا. خرازیان، ش. بختیاری، "کشف رفتارهای مشکوک در بانک‌داری الکترونیکی"، فصلنامه روند، بی‌تا.

[8]. F. Chen, P. Deng, J. Wan, D. Zhang, A. V. Vasilakos, and X. Rong, "Data mining for the internet of things: literature review and challenges", International Journal of Distributed Sensor Networks, 2015.

[9]. M. S. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi, and A. P. Sheth, "Machine learning for Internet of Things data analysis: A survey", Digital Communications and Networks, 2017.

[10]. C. A. Paasch, "Credit card fraud detection using artificial neural networks tuned by genetic algorithms: Hong Kong University of Science and Technology (Hong Kong)"|, 2008.

[11]. D. Olszewski, "Fraud detection using self-organizing map visualizing the user profiles", Knowledge-Based Systems, vol. 70, pp. 324-334, 2014.

[12]. H. Joudaki, A. Rashidian, B. Minaei-Bidgoli, M. Mahmoodi, B. Geraili, M. Nasiri, et al., "Using data mining to detect health care fraud and abuse: a review of literature", Global journal of health science ,vol. 7, p. 194, 2015.

[13]. P. Kulkarni, "Introduction to reinforcement and systemic machine learning", 2012.

[14]. S. Wang and X. Yao, "Diversity analysis on imbalanced data sets by using ensemble models", in 2009 IEEE Symposium on Computational Intelligence and Data Mining, pp. 324-331, 2009.

[15]. A. FernáNdez, V. LóPez, M. Galar, M. J. Del Jesus, and F. Herrera, "Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches", Knowledge-based systems, vol. 42, pp. 97-110, 2013.

[16]. T. Fawcett, "Learning from imbalanced classes", Fawcett, Tom: Author, Retrieved from Silicon Valley Data Science website. https://svds. com/learningimbalanced-classes, 2016.

[17]. G. M. Weiss and F. Provost, "Learning when training data are costly: The effect of class distribution on tree induction",

[30]. B. S. Everitt, S. Landau, M. Leese, and D. Stahl, "Hierarchical Clustering", in Cluster Analysis, ed: John Wiley & Sons, Ltd, pp. 71-110, 2011.

[31]. T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: an efficient data clustering method for very large databases1", in ACM Sigmod Record, pp. 103-114. 1996.

[32]. A. Garg, A. Mangla, N. Gupta, and V. Bhatnagar, "PBIRCH: A scalable parallel clustering algorithm for incremental data", in Database Engineering and Applications Symposium, 2006. IDEAS'06. 10th International, pp. 315-316, 2006.

[33]. M. A. Tahir, J. Kittler, K. Mikolajczyk, and F. Yan, "A multiple expert approach to the class imbalance problem using inverse random under sampling", in International Workshop on Multiple Classifier Systems, pp. 82-91, 2009.

[34]. G. Haixiang, L. Yijing ,J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications", Expert Systems with Applications, vol. 73, pp. 220-239, 2017, 2017.

[35]. L. Yijing, G. Haixiang, L. Xiao, L. Yanan, and L. Jinling, "Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data", Knowledge-Based Systems, vol. 94, pp. 88-104, 2016.

[36]. I. Triguero, M. Galar, D. Merino, J. Maillo, H. Bustince , and F. Herrera, "Evolutionary undersampling for extremely imbalanced big data classification under apache spark", in 2016 IEEE Congress on Evolutionary Computation (CEC), pp. 640-647, 2016.

[37]. I. Triguero, M. Galar, S. Vluymans, C. Cornelis, H. Bustince ,F. Herrera, et al., "Evolutionary undersampling for imbalanced big data classification", in 2015 IEEE Congress on Evolutionary Computation (CEC), pp. 715-722, 2015.

[38]. J. Ha and J.-S. Lee, "A new under-sampling method using genetic algorithm for imbalanced data classification", in Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication, pp. 95, 2016.

[39]. P. Vuttipittayamongkol, E. Elyan, A. Petrovski, and C. Jayne, "Overlap-Based Undersampling for Improving Imbalanced Data Classification", in International Conference on Intelligent Data Engineering and Automated Learning, pp. 689-697, 2018.

[40]. F. Ren, P. Cao, W. Li, D. Zhao, and O. Zaiane, "Ensemble based adaptive over-sampling method for imbalanced data learning in computer aided detection of microaneurysm", Computerized Medical Imaging and Graphics, vol. 55, pp. 54-67, 2017.

[41]. S. Ando and C. Y. Huang, "Deep over-sampling framework for classifying imbalanced data", in Joint European Journal of artificial intelligence research, vol. 19, pp. 315-354, 2003.

[18]. N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study", Intelligent data analysis, vol. 6, pp. 429-449, 2002.

[19]. R. C. Prati, G. E. Batista, and M. C. Monard, "Class imbalances versus class overlapping: an analysis of a learning system behavior", in Mexican international conference on artificial intelligence, pp. 31.۲-۳۲۱, 2004.

[20]. M. M. S. Shelke, P. R. Deshmukh, and V. K. Shandilya, "A Review on Imbalanced Data Handling Using Undersampling and Oversampling Technique", 2017.

[21]. M. Bach, A. Werner, J. Żywiec, and W. Pluskiewicz, "The study of under-and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis", Information Sciences, vol. 384, pp. 174-190, 2017.

[22]. W.-C. Lin, C.-F. Tsai, Y.-H. Hu, and J.-S. Jhang, "Clustering-based undersampling in class-imbalanced data", Information Sciences, vol. 409, pp. 17-26, 2017.

[23]. L. Yu, R. Zhou, L. Tang, and R. Chen, "A DBN-based resampling SVM ensemble learning paradigm for credit classification with imbalanced data", Applied Soft Computing, vol. 69, pp. 192-202, 2018.

[24]. D. Devi and B. Purkayastha, "Redundancy-driven modified Tomek-link based undersampling: a solution to class imbalance", Pattern Recognition Letters, vol. 93, pp. 3-12, 2017.

[۲۵]. م. امیرآبادی، ک. ن. فرجام، م. ر. ناصح، "یادگیری داده‌های نامتوازن توسط ترکیبی از الگوریتم‌های طبقه‌بندی و خوشه‌بندی مبتنی بر سیستم فازی قانون محور"، کنگره بین‌المللی علوم و مهندسی, ۱۳۹۶.

[26]. D. S. Vijayarani and M. P. Jothi, "Hierarchical and Partitioning Clustering Algorithms for Detecting Outliers in Data Streams", International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, pp. 6205-6207, 2014.

[27]. P. Berkhin, "A survey of clustering data mining techniques", in Grouping multidimensional data, ed: Springer, pp. 25-71, 2006.

[۲۸]. م. اسماعیلی، "داده کاوی و مفاهیم آن"، ۱۳۹۴.

[29]. C. A. Ralanamahatana, J. Lin, D. Gunopulos, E. Keogh, M. Vlachos, and G. Das, "Mining time series data", in Data mining and knowledge discovery handbook, ed: Springer, pp. 1069-1103, 2005.

Conference on Machine Learning and Knowledge Discovery in Databases, pp. 770-785, 2017.

[42]. G. Douzas and F. Bacao, "Self-Organizing Map Oversampling (SOMO) for imbalanced data set learning", Expert systems with Applications, vol. 82, pp. 40-52, 2017.

[43]. F. J. Castellanos, J. J. Valero-Mas, J. Calvo-Zaragoza, and J. R. Rico-Juan, "Oversampling imbalanced data in the string space", Pattern Recognition Letters, vol. 103, pp. 32-38, 2018.

[44]. J. Mathew, C. K. Pang, M. Luo, and W. H. Leong, "Classification of imbalanced data by oversampling in kernel space of support vector machines", IEEE transactions on neural networks and learning systems, vol. 29, pp. 4065-4076, 2018.

[45]. 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", International Journal of Approximate Reasoning, vol. 107, pp. 1-16, 2019.

[46]. J. Du, C.-M. Vong, Y. Chang, and Y. Jiao, "Online Sequential Extreme Learning Machine with Under-Sampling and Over-Sampling for Imbalanced Big Data Classification", in Proceedings of ELM-2016, ed: Springer, pp. 229-239, 2018.

[47]. C.-W. Yeh, D.-C. Li, L.-S. Lin, and T.-I. Tsai, "A Learning Approach with Under-and Over-Sampling for Imbalanced Data Sets", in 2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), pp. 725-729, 2016.

[48]. A. Agrawal, H. L. Viktor, and E. Paquet, "SCUT: Multi-class imbalanced data classification using SMOTE and cluster-based undersampling", in 2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), pp. 226-234, 2015.

[49]. W. Lu, Z. Li, and J. Chu, "Adaptive ensemble undersampling-boost: a novel learning framework for imbalanced data", Journal of Systems and Software, vol. 132, pp. 272-282, 2017.

Presenting a Novel Approach for Oversampling-Undersampling in Imbalanced Datasets

Hossein Zaaferani[1*] | Sadegh Rastad[2]

1- Department of Computer Engineering - Faculty of Engineering - Semnan University- Semnan- Iran.
2- Department of Computer Engineering - Faculty of Engineering - Islamic Azad University, Birjand Branch - Birjand - Iran.

Abstract

Imbalanced datasets are very common in the real world and pose a significant challenge in classification and clustering tasks. Imbalanced datasets are defined as those where the instances of one or more classes significantly outnumber the instances of other classes. This research aims to present a mechanism for balancing imbalanced datasets by integrating undersampling and oversampling techniques. Overall, the goal of this study is to propose a hybrid undersampling-oversampling method that can efficiently and effectively balance imbalanced datasets. The proposed approach focuses on minority and majority data points rather than entire minority and majority classes. To identify these data points, the Birch clustering algorithm is employed. Consequently, the proposed method is named Birch-Resampling. This method consists of six stages. The MATLAB software was used for the implementation of this algorithm, and for evaluation, 11 two-class and multi-class datasets with varying sizes and imbalance ratios were utilized to comprehensively assess the performance of the proposed system. The results obtained from the proposed algorithm were compared with five other methods. The outcomes of these comparisons demonstrate that the proposed algorithm exhibits high accuracy in data classification

Keywords: Imbalanced Data, Birch Clustering, Classification, Data Mining.