تاریخ دریافت مقاله: ۹۶/۰۵/۳۰
تاریخ پذیرش مقاله: ۹۶/۰۹/۰۸

بهبود رده‌بندی داده‌های نامتوازن در الگوریتم جنگل تصادفی با استفاده از شبکه‌های عصبی–فازی

هادی مهدوی‌نیا
دانشگاه آزاد اسلامی واحد دولت‌آباد، گروه کامپیوتر، اصفهان، ایران
پست الکترونیکی: hadi.mahdavinia1365@gmail.com

اعظم ربیعی*
استادیار گروه کامپیوتر، دانشگاه آزاد اسلامی واحد دولت آباد، اصفهان، ایران
پست الکترونیکی: a.rabiee@iauda.ac.ir

خلاصه

امروزه پژوهشگران برای داده‌کاوی[1]، با دو نوع داده، مواجه هستند: ۱- داده‌های متوازن ۲- داده‌های نامتوازن[2]. چالش اصلی، در داده‌کاوی داده‌های نامتوازن است. از طرفی، یکی از روش‌های داده‌کاوی، رده‌بندی[3] است. برای رده‌بندی داده‌های نامتوازن با چالش‌هایی مانند: استخراج مدل جانبدارانهٔ متمایل به داده‌های آموزشی، رده‌بندی اشتباه رده اقلیت[4]، صرفنظر کردن از داده‌های مهم ردهٔ اقلیت و بیش‌پوشش[5] مواجه هستیم؛ از این رو از روش‌های مرسوم و معمول نمی‌توان برای رده‌بندی این نوع داده‌ها استفاده کرد. در این تحقیق، سعی شده است رده‌بندی داده‌های نامتوازن با الگوریتم جنگل تصادفی[6] را با استفاده از روش شبکه‌های عصبی–فازی انفیس[7]، بهبود بخشیده و معیارهای مختلف برای ارزیابی این روش، سنجیده شود.

واژه‌های کلیدی: رده‌بندی، داده‌های نامتوازن، بیش‌پوشش، رده اکثریت[8]، رده اقلیت.

۱. مقدمه

یکی از فنون[9] مهم یادگیری از داده‌ها، رده‌بندی است. در رده‌بندی کردن داده‌ها، دو نوع از مجموعه داده‌ها، از نظر توازن[10] مشاهده می‌شود: داده‌های متوازن و داده‌های نامتوازن[۱]. یادگیری رده‌بندی‌کننده‌ها از داده‌های نامتوازن یا اُریب[11]، یک مسئلهٔ بسیار مهم است[۲]. در واقع منظور از داده‌های نامتوازن این است که تعداد نمونه‌های یک یا چند رده[12]، بیشتر از نمونه‌های یک یا تعدادی از رده‌های دیگر باشد[۳]. در مجموعه داده‌های نامتوازن و رده‌های دودوئی[13]، رده اقلیت، مورد توّجه است[٤][٥].

از چالش‌های بزرگ و اساسی در داده‌های نامتوازن، می‌توان به: امکان نادیده‌گرفتن ردهٔ اقلیت، تمایل نمونه‌ها

* نویسندهٔ مسئول

1- data mining
2- imbalanced dataset
3- classification
4- minority class
5- overfitting
6- random forest
7- aNFIS
8- majority class
9- technique
10- balance
11- skewed
12- class
13- binary Class

به جمعیتی که جزء مشکلات عدم توازن نیستند و به‌دست آوردن خطاهـــای رده‌های مختلف برای کســـب هزینه‌های مختلف، اشـــاره کرد[۳]. در رده‌بنـــدی داده‌های نامتوازن، ممکن است رده اقلیت از درستیِ[۱۴] کمی برخوردار باشد[٦]. این رده‌های اقلیت، خود نیز می‌توانند باعث کاهش دقّت[۱۵]و افزایش خطاهای الگوهای کشف شده شوند[۷]. همچنین رده اقلیت، باعث می‌شود تا ســـاخت مجموعه داده آموزشی[۱۶] نیز بســـیار سخت شـــود[۸]. به عبارتی، داده‌ها و رده‌های نامتوازن، می‌توانند باعث ایجـــاد اختلال در الگوریتم‌های یادگیری شوند[۹].

برای حل این قبیل چالش‌ها، معمولاً از چهار دسته‌بندی استفاده می‌شود. دسته‌بندی مفصل‌تری توسط ژونگلِیانگ ژِن و همکاران ارائه شده اســـت که راه‌حل‌های موجود را برای برخورد با داده‌های نامتوازن، به چهار دستهٔ رویکرد سطح داده، رویکرد ســـطح الگوریتم، یادگیری حساس به هزینه و یادگیری گروهی تقسیم کرده است[۸].

یکی دیگر از راه‌حل‌های موجود در مواجهه با داده‌های نامتوازن، رویکـــرد الگوریتمی و اســـتفاده از درخت‌های تصمیم اســـت؛ البته این نـــوع از الگوریتم‌هـــا دارای انواع مختلفی از جمله درخت‌های غیرحســـاس به اندازهٔ رده[۱۷]، درخت‌هـــای گروهی و از این قبیل اســـت و یکـــی از انواع الگوریتم‌های درخت‌های گروهی، جنگل تصادفی اســـت[۳] [۱۰]. در الگوریتم جنگل تصادفی، برای پیدا کردن بهترین رده‌بندی‌کنندهٔ داده‌های نامتوازن، از رأی‌گیری[۱۸] بین نتایج چندین درخت استفاده می‌شود. در واقع، مسئله اصلی، پیدا کردن بهترین رده‌بنـــدی برای مجموعه داده‌های نامتوازن اســـت که این کار، بسیار سخت است زیرا با افزایش تعداد درختان، که یک متغیر وابسته به تعداد نمونه‌های مجموعه داده اســـت، انتخاب بهترین مدل نیز سخت می‌شود. برای بهبود انتخاب بهترین جواب حاصل که همان مدل رده‌بندی اســـت، بایـــد بهبود مناســـبی در عملکـــرد الگوریتم جنگل تصادفی انجام شـــود. همچنین مدت زمان اجرای الگوریتم جنـــگل تصادفی بـــر روی داده‌های بزرگ[۱۹]، بســـیار زیاد اســـت که این خود یک نقطه ضعف بسیار بزرگ محسوب می‌شـــود. در این تحقیق، هدف اصلی، بهبود روند انتخاب بهترین مـــدل رده‌بندی برای داده‌های نامتوازن اســـت که این کار، توسط شـــبکه‌های عصبی–فازی در روند ساخت درخت و رأی‌گیری، صورت می‌گیرد.

۲. پیشینه کاری

برای اولین بـــار به صورت رســـمی، رده‌بندی داده‌ها به وســـیله الگوریتم جنگل تصادفی، توسط لئو بریمن در سال ۲۰۰۱ مطرح شد[۱۱]؛ از آن پس بود که این الگوریتم بـــرای رده‌بندی داده‌ها مورد اســـتفاده قـــرار گرفت. یکی از معضـــلات رده‌بنـــدی داده‌ها، عدم توازن اســـت. ژولین توماس و همکاران در سال ۲۰۰٦ در دانشگاه لومیر لیون فرانســـه به بهینه‌ســـازی و ارزیابی جنگل تصادفی برای داده‌های نامتوازن، با اســـتفاده از روشگان[۲۰] دو مرحله‌ای پرداختند[۱۲]. دِ ریچارد کاتلر و همکاران در ســـال ۲۰۰۷ حتّی از الگوریتم جنگل تصادفـــی برای رده‌بندی داده‌های بوم‌شناســـی[۲۱] اســـتفاده کردند و نتایج حاصله را با نتایج حاصـــل از ٤ روش دیگـــر رده‌بندی، مقایســـه کردند[۱۳]. خوش‌گفتار و همکاران در ســـال ۲۰۰۷ با اســـتفاده از نرم افزار وِکا[۲۲]، یک مطالعه تجربی بر روی رده‌بندی داده‌های نامتوازن با استفاده از الگوریتم جنگل تصادفی داشتند[۱۰]. پیِرو بونیســـونه و همکاران در ســـال ۲۰۱۰، با استفاده از منطـــق فازی، ســـعی در بهبـــود عملکـــرد الگوریتم جنگل تصادفی داشـــتند و ســـعی کردند در مرحله رأی‌گیری از منطق فازی[۲۳] اســـتفاده کنند و بدین ترتیب الگوریتم جنگل تصادفی مرحلهٔی را ارائه دادند[۱٤]. در سال ۲۰۱۱ خلیلیا و همکاران در دانشگاه میسوری از الگوریتم جنگل تصادفی برای پیش‌بینی میزان ریسک تشخیص بیماری از داده‌های

14- accuracy
15- precision
16- training dataset
17- decision tree algorithms insensitive to the class Sizes
18- voting
19- big data
20- methodology
21- ecology
22- Weka
23- fuzzy logic

بسیار نامتوازن اســتفاده کردند[۱۵]. همچنین وریکاس و همکاران در سال ۲۰۱۱، طی یک تحقیق مروری، سازگاری و اصــل کلیت رتبه‌بندی در جنگل تصادفی را با اســتفاده از آزمایش‌های جدید، بررسی کردند[۱۶]. در سال ۲۰۱۲، وُراشــالی وای کولکارنی و پرادیپ کِی سینها، از دانشگاه پونا هند، هرس[۲۴] جنگل تصادفی را بررســی کردند و طی یک مقاله مروری موضوع هــرس را در درخت‌های تولید شده در جنگل تصادفی به چالش کشیدند[۱۷]. سارا دِل ریو و همکاران از دانشگاه گرانادا اسپانیا در سال ۲۰۱٤، مبحث داده‌های نامتوازن در MapReduce را با اســتفاده از جنگل تصادفی، مطرح کردند[۱۸]. ژینگ یائو وُنگ و همکاران در سال ۲۰۱٤، حتی از الگوریتم جنگل تصادفی برای رده‌بندی متون نامتوازن اســتفاده کرده‌اند که نام الگوریتم خود را FORESTEXTER گذاشــتند[۱۹]. در ســال ۲۰۱٦، هی ژائو و همکاران، متد جدیدی به نام SOB[۲۵] مطرح کرده‌اند که ترکیبی از تکنیک نمونه‌برداری افزایش ردهٔ اقلیت و Bagging است که به صورت رده رده[۲٦] عمل می‌کند[۲۰].

۳. ضرورت تحقیق

داده‌های نامتــوازن در بســیاری از زمینه‌ها، از جمله پردازش تصاویر مغناطیسی، داده‌های بازاریابی، تشخیص نشتی مواد نفتی، تشخیص تغییرات سطح زمین با استفاده از تصــاویر سنجش از راه دور، رده‌بندی متون، رده‌بندی صحنه‌ها[۵]، تشــخیص خطا در کارت‌هــای اعتباری[۲۱]، زیســت داده‌ورزی[۳]و پزشکی[۲۲] مشــاهده می‌شوند. برای مثال در زمینه پزشکی، در تشخیص بیماری سرطان، تعداد افراد مبتلا به سرطان بسیار کمتر از تعداد افراد سالم است؛ حال اگر فرد مبتلا به بیماری سرطان، سالم تشخیص داده شود باعث تلف شــدن زمان درمان و همچنین بروز فاجعه برای زندگی آن فرد می‌شــود[۲۲]. به عنوان مثالی دیگــر، روزانه مطالب و متون زیادی در اینترنت، انتشــار پیــدا می‌کند که هر کدام از درجه اهمیّت خاصی برخوردار هستند؛ حال اگر هدف کشــف ارتباط مفاهیم بسیار داغ و مهم با مطالب نشر شده در اینترنت باشد، مطالب داغ و مهم نیز، خود داده‌های اقلیت محسوب می‌شوند[۷].

اغلب مجموعه داده‌های دنیای واقعی، عمدتاً ترکیبی از نمونه‌های طبیعــی و درصد کمی نمونه‌های غیر طبیعی یا جالب اســت. در این موارد، هزینه رده‌بندی اشــتباه نمونه غیرطبیعی[۲۷] یا جذاب به عنوان نمونه طبیعی بسیار بیشتر از هزینه خطای معکوس است[۲۳].

۴. روش‌های رده‌بندی داده‌های نامتوازن

در برخــورد بــا داده‌هــای نامتوازن جهــت رده‌بندی آن‌ها، اکثراً روش‌های چهار دسته‌ای ارائه می‌شوند که در برخورد با داده‌های نامتوازن، الزاماً استفاده از رویکردهای زیر پیشنهاد نمی‌شوند.

٤-۱. رویکردهای تعبیه شده

این رویکرد در درون یادگیرنده یا در مرحلهٔ یادگیری پیاده‌سازی می‌شود و یا ممکن است به عنوان یک رویکرد داخلــی در نظر گرفته شــده باشــد. این رویکرد شــامل روش‌هــای اعمــال و تنظیم و تعیین وزن و یا شــامل یک تابــع نادیده گرفتن بــرای نمونه‌های مختلــف و همچنین یادگیرنده‌هایــی که ذاتــاً برای برخی روش‌هــای انتخاب زیرنمونه عمل می‌کنند، باشــد. یادگیــری فعال، روش‌های مبتنی بر هسته و رویکردهای انتساب هزینه در این رویکرد قرار می‌گیرند[۲٤].

٤-۲. رویکردهای سطح داده

ایده اصلی این رویکرد مربوط به مرحله پیش‌پردازش داده‌ها است. در واقع در این رویکرد، برای حل عدم تعادل، توزیع رده‌ها نیز دســتکاری شده تا تعادل در کل رده‌های مجموعه داده‌ها حاصل شود. این رویکرد، بیشتر در ترکیب بــا رویکرد ترکیبی که در ادامه توضیح داده خواهد شــد،

24- pruning
25- Stratified Oversampling Bagging
26- stratified
27- abnormal

جدول ۱ - انواع روش‌های رده‌بندی داده‌های نامتوازن و مزایا و معایب آن‌ها

روش	مزایا	معایب
رویکردهای تعبیه شده	انطباق با داده‌های حجیم، عدم نیاز به پیش‌پردازش داده‌ها	محاسبات سنگین برای اختصاص وزن
رویکردهای سطح داده	عدم نیاز به فرمول و محاسبه وزن، استفاده کردن از رویکردهای استاندارد	حذف و نادیده گرفته شدن برخی از داده‌های پراهمیّت و مهم، پیش‌پردازش سخت در داده‌های بزرگ
رویکردهای حساس به هزینه	دقّت و صحّت بالا، جلوگیری از رده‌بندی اشتباه ردهٔ اقلیت، کاهش خطای وزنی نسبت به روش‌های دیگر	پیدا کردن بهترین فرمول لازم برای وزن‌دهی ردهٔ اقلیت، محاسبات سنگین
رویکردهای ترکیبی	دقّت تشخیص بالا، استفاده از رده‌بندی کننده‌های استاندارد	انجام پیش‌پردازش روی داده‌ها، انطباق روش‌ها با یکدیگر

استفاده می‌شود[۳]. از زیرروش‌های این رویکرد که بسیار استفاده می‌شوند، می‌توان به روش افزایش نمونه‌های رده یا رده‌های اقلیت، کاهش نمونه‌های رده یا رده‌های اکثریت، RUS[۲۸]، ROS[۲۹] و SMOTE[۳۰] اشاره کرد[۸].

٤-۳. رویکرد حساس به هزینه

در این رویکرد، برای افزایش دقّت و صحّت، به نمونه‌های هر رده، وزن اختصاص داده می‌شود. از زیرروش‌های این رویکرد می‌توان به درخت‌های تصمیم حساس به هزینه و شبکه‌های عصبی حساس به هزینه اشاره کرد[۳].

در واقع در این رویکرد، برای رده‌های اقلیتی که به صورت اشتباهی رده‌بندی می‌شوند نیز وزن بیشتری نسبت به ردهٔ اکثریت در نظر گرفته می‌شود؛ بنابراین در فرآیند یادگیری، هدف، کاهش خطاهای وزنی، به جای افزایش نرخ صحّت است[۸].

٤-٤. رویکرد ترکیبی

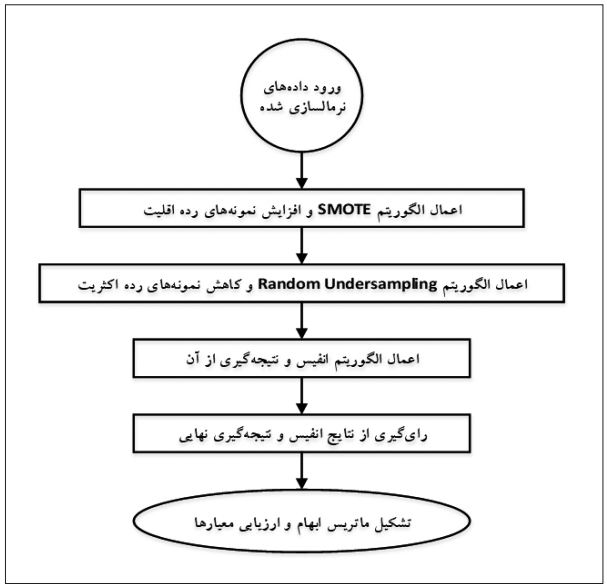
این رویکرد، اغلب برای روش‌های رده‌بندی استاندارد که برای رده‌بندی داده‌های نامتوازن، نامناسب هستند، به کار می‌رود. رویکردهای ترکیبی بر مبنای سطح داده، یکی از زیررویکردهای این دسته است. در این رویکرد، هر رده‌بندی با یک مجموعه دادهٔ پیش‌پردازش شده، آموزش می‌بیند. در زیررویکرد سطح داده، رده‌بندی کننده‌های مختلف با مجموعه داده‌های پیش‌پردازش شده استفاده می‌شود. در واقع رده‌بندی کننده‌های استاندارد با مجموعه داده‌هایی که پیش‌پردازش شده‌اند نیز مورد استفاده قرار می‌گیرند[۳].

در رویکرد ترکیبی، زمانی که از رده‌بندی کننده‌های مستقل استفاده می‌شود، صحّت، پیش‌بینی بالاتر و بهتری نسبت به تک رده‌بندی کننده‌های قوی دارد. بنابراین می‌توان گفت که رده‌بندی کننده‌های ترکیبی داده‌های نامتوازن، بیشتر برای احقاق دقّت تشخیص بالاتر، استفاده می‌شوند[۲۵]. یکی از انواع زیررویکردهای ترکیبی، زیررویکرد SMOTEBAGGING است که شامل Bagging و بخش متفاوتی از SMOTE و Oversampling در هر تکرار است[۳].

هر یک از روش‌های رده‌بندی داده‌های متوازن نیز دارای نقاط ضعف و قوت هستند که به صورت اجمالی در جدول ۱ بیان شده‌اند.

با توّجه به جدول ۱، روش سطح داده، برای رده‌بندی داده‌های نامتوازن که دارای حساسیت زیادی هستند، مناسب نمی‌باشند زیرا ممکن است از داده‌های مهم و اثرگذار در نتیجه، صرفنظر شود؛ همچنین در رویکرد سطح داده، حتی با ترکیب روش‌های آن، باز کلیهٔ فعالیت‌ها در مرحلهٔ پیش‌پردازش انجام می‌شود امّا در نقطه مقابل،

28- Random Under-Sampling
29- Random Over-Sampling
30- Synthetic Minority Over-sampling Technique



شکل ۱ : مراحل روش پیشنهادی

رویکردهای ترکیبی بسیار مؤثرتر و بهتر هستند، زیرا هم می‌تـوان از رویکردهای سـطح داده و هم از رویکردهای الگوریتمی جهت رده‌بندی داده‌های نامتوازن استفاده کرد.

در رویکـرد سـطح داده، زیررویکردهـای متنوع‌تری وجود دارند که این باعث می‌شـود بتـوان برای رده‌بندی داده‌های نامتوازن از زیررویکردهای متناسب‌تری استفاده کرد.

از رویکرد حسـاس به هزینه زمانی استفاده می‌شود که اهمّیـت مقادیر صفت خاصه زیاد اسـت امّا زمانی که تعـداد نمونه‌های رده‌های مختلف زیاد و از اهمّیت کمتری برخوردار باشـند می‌توان از روش و رویکرد سطح داده استفاده کرد.

در واقـع بهترین روش، رویکرد ترکیبی و سـاده‌ترین روش، رویکرد سطح داده است که از دقّت و صحّت کمتری برخوردار است.

۵. روش پیشنهادی

روش پیشـنهادی ارائه شده در این تحقیق، سعی دارد با در نظرگیـری معیارهای مهمی همانند دقّت، مسـاحت زیر منحنـی ROC، زمان، صحّت، نرخ خطـا[31]، F-Measure G-Means، حساسـیت[32] و ویژگی، نشـان دهـد که دارای شایستگی رده‌بندی داده‌های نامتوازن است. قبل از ورود به الگوریتم یا روش پیشـنهادی، داده‌ها باید بین محدودهٔ صفـر و یک، نرمال‌سـازی شـده باشـند. همچنین برای انتخاب سـتون‌های تصادفی، از بین کل مجموع داده‌های نرمال‌سازی شـده، باید میزان همبسـتگی[33] کلیهٔ صفات خاصه با ستون هدف نیز محاسبه شده باشند.

کلیهٔ مراحل روش پیشـنهادی در شـکل ۱ نشان داده شده اسـت و برخی از مراحل می‌تواند به صورت موازی نیز اجرا شود.

در مرحلهٔ اول، داده‌های نرمال‌سازی شده، وارد برنامه شده و به دو دسته تقسیم می‌شوند: ۱- دسته رده اکثریت که ستون هدف آن‌ها مقدار صفر دارد. ۲- دستهٔ ردهٔ اقلیت که مقدار ستون هدف آن‌ها، مقدار یک دارد. سپس الگوریتم افزایـش نمونه‌هـای ردهٔ اقلیت که SMOTE اسـت بر روی دستهٔ دوم، اجرا می‌شود و تعداد نمونه‌های تصادفی که از قبل، درصد آن‌ها نیز مشخص شده است، تولید می‌شوند. همچنین، الگوریتم کاهش تصادفی نمونه‌ها، بر روی دستهٔ اول انجام می‌شـود و با کاهش نمونه‌های ردهٔ اکثریت، به درصد مورد نظر ما بین ردهٔ اقلیت و ردهٔ اکثریت می‌رسد. شـایان ذکر است دو مرحلهٔ کاهش نمونه‌های ردهٔ اکثریت و افزایش نمونه‌های ردهٔ اقلیت، می‌تواند به صورت موازی هم اجرا شود.

پس از ایجاد توازن تعیین شـده، حـال باید وارد حلقهٔ اصلی تکرار شـده و هسـتهٔ اصلی روش پیشنهادی اجرا شـود. در این حلقه، هر بار به صـورت تصادفی، تعدادی ستون از بین ستون‌های مجموعه دادهٔ آماده شده در مرحلهٔ قبل، انتخاب می‌شـود. به همین منظور برای هر ستون یک وزن که همان نسـبت همبسـتگی صفات خاصه با ستون هدف اسـت، در نظر گرفته می‌شود. سـپس، از بین همان مجموعه با سـتون‌های تصادفی انتخاب شده، به میزان ۸۰ درصد از داده‌ها، برای مجموعه داده‌های آموزشـی و ۲۰

31- error rate

32- sensitivity
33- correlation

درصد از داده‌ها بـــرای مجموعه داده‌های آزمون، انتخاب می‌شـــود. در مرحله بعد، آموزش هسته و ساختار انفیس صورت می‌گیرد و بعد از آن با استفاده از هر دو مجموعه دادهٔ آموزشـــی و مجموعه دادهٔ آزمون، خروجی سیستم انفیس نیز تولید می‌شود.

بعـــد از این که کلیـــهٔ نتایج حاصل از تکرار مشـــخص انفیـــس، به‌دســـت آمـــد از رابطـــه ۱ برای انجـــام عملیات رأی‌گیری استفاده می‌شود و نتایج رأی‌گیری نیز به همراه نتایج ستون هدف، ماتریس ابهام را تشکیل می‌دهند.

$$V_{Final\ voting} = step(\sum_{i=1}^{n}(o_i - T)) \quad (۱)$$

در رابطه ۱، $V_{Final\ voting}$، بیانگر نتیجه رأی‌گیری کل نتایج حاصل از انفیس اســـت. منظـــور از step، همان تابع پله‌ای اســـت که نتیجهٔ حاصل از عملیات داخـــل پرانتز، به صفر یا یک منجر می‌شـــود. i نیز بیانگر شـــمارندهٔ نتایج حاصل از تکرار انفیس اســـت؛ به عبارت بهتر، شـــمارندهٔ مربوط به ســـتون‌های جدول نتایج انفیس اســـت که تا n ادامه پیدا می‌کند. n نیز تعداد کل خروجی‌های حاصل از تکرار انفیس اســـت. o_i، خروجی مربوط به هر تکرار انفیس است؛ برای مثال خروجی تکرار دوم انفیس، عدد ۰/۲ است که $o_i = 0.2$، بیانگر همین مســـئله است. T یک عدد ثابت است که در این تحقیق ۰/۵ است.

در مرحله پایانی، کلیهٔ معیارها، با اســـتفاده از ماتریس ابهام نیز ارزیابی می‌شوند.

۶. شبیه‌سازی

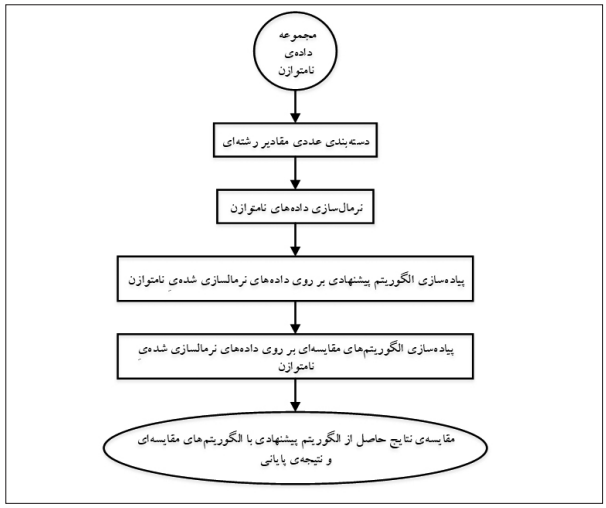
در ایـــن تحقیق از داده‌هـــای creadit_card که از پایگاه دادهٔ kaggle است، اســـتفاده شده است. این مجموعه داده، تراکنش‌های کارت‌های اعتباری ناشـــناس در سال ۲۰۱۳ را بـــا دو برچســـب اصلی یا تقلبی مشـــخص می‌کند. کلیهٔ نمونه‌هـــای این مجموعه داده، مربـــوط به تراکنش‌های دو روز کارت‌های اعتباری است. این مجموعه داده که نظارت شده[۳۴] است دارای ۲۸۴۸۰۷ نمونه[۳۵] است.

این مجموعه داده دارای ۳۰ صفت خاصه است که یک ستون هدف آن نیز دو رده‌ای است. در ستون هدف، تعداد نمونه‌های ردهٔ اقلیت که با عدد یک مشـــخص شـــده است، ۴۹۲ نمونه است و این بدین معنی است که ۴۹۲ نمونه تقلبی وجـــود دارد و تعداد نمونه‌هـــای ردهٔ اکثریت نیز ۲۸۴۳۱۵ نمونه است.

در شکل ۲، کلیهٔ مراحل شبیه‌سازی در این تحقیق، نشان داده شده است. همان‌طور که در شکل ۲ مشاهده می‌شود، در ابتدا باید مقادیر رشـــته‌ای مجموعـــه دادهٔ نامتوازن به مقادیر عددی تبدیل شـــود که به اصطلاح، مقادیر رشته‌ای به دســـته‌بندی عددی تبدیل می‌شـــوند. البته هیچ اجباری در اســـتفاده از اعداد صحیح و یا اعشـــاری نیست ولی در بســـیاری از موارد از اعداد صحیح اســـتفاده می‌شود. در مرحلهٔ بعـــد، کلیهٔ مقادیر مجموعه داده‌هـــای نامتوازن که مقادیر عددی هستند، بین محدودهٔ عددی صفر و یک قرار می‌گیرند که به اصطلاح، نرمال‌سازی می‌شوند. در مرحلهٔ سوم، الگوریتم و روش پیشنهادی بر روی داده‌های نرمال شده، پیاده‌ســـازی می‌شود. در مرحلهٔ بعد، همان داده‌های نرمال شـــده، توســـط روش‌های دیگری برای مقایســـه با روش پیشنهادی، استفاده می‌شود. این روش‌ها عبارتند از: شـــبکه‌های عصبی، جنگل تصادفی، درخت تصمیم C4.5 و انفیـــس. در مرحله پایانی، تمام نتایج حاصل از ۵ روش، با یکدیگر مقایسه شده و نتیجه نهایی نیز بیان می‌گردد.

در این تحقیق، ابتدا مجموعه دادهٔ نرمال‌ســـازی شده، توسط روش پیشـــنهادی، استفاده شده و نتایج حاصل از آن نیز طبق جداولی به‌دســـت آمد. برای به‌دســـت آوردن نتایـــج، از طریق الگوریتم و روش پیشـــنهادی به این‌گونه عمل می‌شـــود که ابتدا باید مواردی از قبیل میزان درصد توازن ردهٔ اقلیت و اکثریت، تعداد نزدیک‌ترین همسایه برای محاسبه افزایش تعداد نمونه‌های ردهٔ اکثریت یا همان روش SMOTE، میزان درصد افزایش ردهٔ اقلیت، تعداد سرخوشه‌ها و هـــم چنین تعـــداد همســـایه‌های سرخوشـــه‌ها، در نظر گرفته می‌شود.

34- supervised
35- sample



شکل ۲ : مراحل کلی انجام شبیه‌سازی

۷. نتایج شبیه‌سازی

نتایج حاصل از شبیه‌سازی در جدول ۲، به صورت کامل مشاهده می‌گردد. کلیهٔ معیارها به درصد بیان شده است و مدت زمان اجرای الگوریتم‌ها نیز، به ثانیه بیان شده است؛ معیارها در میزان تکرار مختلف ارزیابی شده‌اند.

معیار دقّت نیز در روش پیشنهادی، با توازن ۵۰٪-۵۰٪ (ردهٔ اقلیت و ردهٔ اکثریت، هر دو مساوی باشند)، بالای ۹۰ درصد است؛ همچنین مساحت زیر منحنی ROC، در روش پیشنهادی بالای ۹۶ درصد است. میانگین دقّت ۹۲/۸۹ و همچنین میانگین نمودار زیر منحنی ROC برای ردهٔ اقلیت با عدد ۹۶/۵۷، برای روش پیشنهادی قابل مشاهده است که روند رو به رشد و خوبی را نشان می‌دهد. مدّت زمان اجرای روش پیشنهادی در توازن نسبتاً مساوی برای هر دو رده، در هر تکرار، نسبت به همهٔ روش‌های دیگر موجود در این تحقیق، پایین‌تر و به بیان دیگر، بهتر است.

البته، در روش شبکه‌های عصبی-مصنوعی، معیارها، رشد خوبی دارند و مساحت زیر منحنی ROC در روش شبکه‌های عصبی-مصنوعی، میزان بالایی با مقدار ۹۷/۰۴ درصد را دارد امّا میزان دقّت الگوریتم شبکه‌های عصبی-مصنوعی بر روی داده‌های نامتوازن، نسبت به روش پیشنهادی، کمتر است.

در روش انفیس نیز، به دلیل مدّت زمان اجرای طولانی، ادامه نتیجه‌گیری الگوریتم انفیس بر روی داده‌های نامتوازن، به صرفه نبوده است. بنابراین، ادامه رویه و اجرای الگوریتم انفیس بر روی داده‌های نامتوازن، بسیار زمانگیر است. الگوریتم درخت تصمیم C4.5 نیز در حالت تساوی توازن بین رده‌های اقلیت و اکثریت و با پیش‌پردازش داده‌ها، بر روی داده‌های نامتوازن انجام شده است. در این روش، میانگین کلیهٔ پارامترها، سیر صعودی دارد و این سیر صعودی در معیارهای دقّت و مساحت زیر منحنی ROC مشهود است، امّا مدّت زمان اجرای الگوریتم درخت تصمیم C4.5 بر روی داده‌های نامتوازن، نیز بیشتر از الگوریتم و روش پیشنهادی است.

الگوریتم درخت تصادفی بدون پیش‌پردازش داده‌ها بر روی داده‌های نامتوازن با همان درصد توازن اولیه، اجرا شده است که همانطور که مشاهده می‌گردد، نسبت به روش پیشنهادی دارای درصد بیشتری در مساحت زیر منحنی ROC است. امّا معیار دقّت آن از درصد کمتری نسبت به روش پیشنهادی برخوردار است. مدّت زمان اجرای الگوریتم جنگل تصادفی بر روی ۲۸۴۸۰۷ نمونه، زیاد است. در نتیجه، ادامه شبیه‌سازی با ۱۵ و ۲۰ تکرار یا همان تعداد جنگل‌های تصادفی، به صرفه نبوده و مدّت زمان زیادی اتلاف می‌شود.

۸. نتیجه‌گیری

در این تحقیق، سعی شد، معیارهای مختلفی برای روش پیشنهادی، ارزیابی شوند که از تمامی معیارهای بیان شده، معیارهای دقّت، مساحت زیر منحنی ROC و زمان اجرای روش پیشنهادی، مهم‌تر بودند. با توجه به نتایج حاصل از اجرای الگوریتم و روش پیشنهادی بر روی مجموعه دادهٔ creadit_card، نتایج زیر حاصل می‌شود:

- مدّت زمان اجرای روش پیشنهادی نسبت به روش‌های دیگری که در این تحقیق بیان شده‌اند، بسیار بهتر است و این نشان می‌دهد که روش پیشنهادی از نظر پیچیدگی زمانی، در وضعیت مناسبی قرار دارد. حتی روش پیشنهادی، در

جدول ۲ : نتایج حاصل از روش پیشنهادی، انفیس، شبکه‌های عصبی-مصنوعی، درخت تصمیم C4.5 و جنگل تصادفی بر روی داده‌های نامتوازن creadit_card از دادگان kaggle

روش	میزان توازن	معیارها	تکرار										میانگین
			۲		۵		۱۰		۱۵		۲۰		
			Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	
proposed method	۵۰-۵۰	Accuracy	۹۷,۰۲	۹۶,۴۵	۹۵,۶۴	۹۶,۴۸	۹۶,۴۴	۹۶,۴۶	۹۶,۴۹	۹۶,۶۴	۹۵,۸۹	۹۵,۹۹	۹۶,۳۵
		Error rate	۲,۹۷	۳,۵۴	۴,۳۵	۳,۵۱	۳,۵۵	۳,۵۳	۳,۵۰	۳,۳۵	۴,۱۰	۴,۰۰	۳,۶۴
		F-Measure	۹۶,۹۸	۹۶,۳۲	۹۵,۴۷	۹۶,۳۶	۹۶,۳۴	۹۶,۳۳	۹۶,۳۷	۹۶,۵۲	۹۵,۷۲	۹۵,۸۲	۹۶,۲۲
		G-Mean	۹۷,۱۰	۹۶,۶۱	۹۵,۸۲	۹۶,۶۳	۹۶,۵۸	۹۶,۶۳	۹۶,۶۶	۹۶,۸۰	۹۶,۱۱	۹۶,۲۲	۹۶,۵۱
		Precision	۹۴,۲۴	۹۲,۹۹	۹۲,۰۵	۹۳,۲۵	۹۳,۲۷	۹۲,۹۵	۹۲,۹۹	۹۳,۲۸	۹۱,۹۱	۹۱,۹۹	۹۲,۸۹
		Sensitivity	۹۹,۸۸	۹۹,۹۰	۹۹,۱۶	۹۹,۶۹	۹۹,۶۰	۹۹,۹۵	۱۰۰	۱۰۰	۹۹,۸۶	۱۰۰	۹۹,۸۰
		Specificity	۹۴,۴۰	۹۳,۴۳	۹۲,۵۹	۹۳,۶۵	۹۳,۶۵	۹۳,۴۲	۹۳,۴۴	۹۳,۷۱	۹۲,۵۱	۹۲,۵۸	۹۳,۳۳
		AUC	۹۷,۱۴	۹۶,۶۷	۹۵,۸۷	۹۶,۶۷	۹۶,۶۳	۹۶,۶۹	۹۶,۷۲	۹۶,۸۵	۹۶,۱۸	۹۶,۲۹	۹۶,۵۷
		Time	۴۹,۴		۵۹,۴		۷۷,۸		۸۳,۲		۹۰,۹		۷۲,۱۴
ANFIS	۰,۱۷-۹۹,۸۳	Accuracy	۹۹,۹۰	۹۹,۹۰	---	---	---	---	---	---	---	---	۹۹,۹۰
		Error rate	۰,۰۹	۰,۰۹	---	---	---	---	---	---	---	---	۰,۰۹
		F-Measure	۶۹,۴۲	۶۷,۵۰	---	---	---	---	---	---	---	---	۶۸,۴۶
		G-Mean	۹۲,۲۴	۹۰,۹۱	---	---	---	---	---	---	---	---	۹۱,۵۷
		Precision	۵۸,۶۰	۵۷,۰۲	---	---	---	---	---	---	---	---	۵۷,۸۱
		Sensitivity	۸۵,۱۵	۸۲,۷۱	---	---	---	---	---	---	---	---	۸۳,۹۳
		Specificity	۹۹,۹۲	۹۹,۹۲	---	---	---	---	---	---	---	---	۹۹,۹۲
		AUC	۹۲,۵۴	۹۱,۳۲	---	---	---	---	---	---	---	---	۹۱,۹۳
		Time	۳۴۸۲۵,۶		---		---		---		---		۳۴۸۲۵,۶
ANN	۰,۱۷-۹۹,۸۳	Accuracy	۹۹,۹۵	۹۹,۹۵	۹۹,۹۴	۹۹,۹۶	۹۹,۹۵	۹۹,۹۶	۹۹,۹۵	۹۹,۹۶	۹۹,۹۵	۹۹,۹۶	۹۹,۹۵
		Error rate	۰,۰۴	۰,۰۴	۰,۰۵	۰,۰۳	۰,۰۴	۰,۰۳	۰,۰۴	۰,۰۳	۰,۰۴	۰,۰۳	۰,۰۳
		F-Measure	۸۵,۳۵	۸۶,۸۷	۸۲,۹۵	۸۸,۹۱	۸۵,۲۱	۸۹,۲۰	۸۶,۶۸	۸۹,۳۷	۸۵,۶۱	۸۹,۳۷	۸۶,۹۵
		G-Mean	۹۵,۲۳	۹۶,۴۰	۹۵,۵۴	۹۸,۳۰	۹۶,۲۶	۹۸,۰۸	۹۶,۸۵	۹۸,۴۲	۹۶,۴۰	۹۸,۴۲	۹۶,۹۹
		Precision	۸۰,۵۸	۸۱,۵۲	۷۵,۹۷	۸۲,۳۱	۷۸,۸۵	۸۳,۱۳	۸۰,۵۴	۸۲,۹۲	۷۹,۳۳	۸۲,۹۲	۸۰,۸۰
		Sensitivity	۹۰,۷۲	۹۲,۹۷	۹۱,۳۳	۹۶,۶۵	۹۲,۷۰	۹۶,۲۳	۹۳,۸۴	۹۶,۹۱	۹۲,۹۷	۹۶,۹۱	۹۴,۱۲
		Specificity	۹۹,۹۶	۹۹,۹۶	۹۹,۹۵	۹۹,۹۶	۹۹,۹۶	۹۹,۹۷	۹۹,۹۶	۹۹,۹۷	۹۹,۹۶	۹۹,۹۷	۹۹,۹۶
		AUC	۹۵,۳۴	۹۶,۴۷	۹۵,۶۴	۹۸,۳۱	۹۶,۳۳	۹۸,۱۰	۹۶,۹۰	۹۸,۴۴	۹۶,۴۷	۹۸,۴۴	۹۷,۰۴
		Time	۱۱۴,۱		۲۴۷,۶		۴۷۹,۰		۸۸۶,۵		۱۰۹۰,۶		۵۶۳,۵
C4.5	۵۰-۵۰	Accuracy	۹۶,۹۸	۹۹,۱۹	۹۷,۲۸	۹۸,۷۵	۹۶,۹۲	۹۸,۴۵	۹۶,۶۸	۹۹,۳۰	۹۷,۳۵	۹۹,۲۲	۹۸,۰۱
		Error rate	۳,۰۱۳	۰,۸۰۶	۲,۷۱۲	۱,۲۴۱	۳,۰۷۹	۱,۵۴۵	۳,۳۱۴	۰,۶۹۱	۲,۶۴۳	۰,۷۷۲	۱,۹۸
		F-Measure	۹۷,۰۳	۹۹,۱۸	۹۷,۲۳	۹۸,۷۴	۹۶,۸۸	۹۸,۴۲	۹۶,۶۵	۹۹,۳۰	۹۷,۳۴	۹۹,۲۲	۹۷,۹۹
		G-Mean	۹۶,۹۷	۹۹,۲۰	۹۷,۳۱	۹۸,۷۸	۹۶,۹۴	۹۸,۴۹	۹۶,۶۹	۹۹,۳۱	۹۷,۳۶	۹۹,۲۳	۹۸,۰۲
		Precision	۹۶,۳۳	۹۸,۳۷	۹۶,۰۱	۹۷,۵۱	۹۵,۵۸	۹۶,۹۰	۹۶,۲۶	۹۸,۶۱	۹۶,۸۰	۹۸,۴۵	۹۷,۰۸
		Sensitivity	۹۷,۷۴	۱۰۰	۹۸,۴۹	۱۰۰	۹۸,۲۲	۱۰۰	۹۷,۰۴	۱۰۰	۹۷,۸۸	۱۰۰	۹۸,۹۳
		Specificity	۹۶,۲۱	۹۸,۴۱	۹۶,۱۵	۹۷,۵۸	۹۵,۶۸	۹۷,۰۰	۹۶,۳۳	۹۸,۶۳	۹۶,۸۳	۹۸,۴۷	۹۷,۱۲
		AUC	۹۶,۹۸	۹۹,۲۰	۹۷,۳۲	۹۸,۷۹	۹۶,۹۵	۹۸,۵۰	۹۶,۶۹	۹۹,۳۱	۹۷,۳۶	۹۹,۲۳	۹۸,۰۳
		Time	۱۲۲,۲		۲۴۹,۴		۴۹۳,۸		۶۸۹,۲		۹۱۰,۱		۴۹۲,۹
RF	۰,۱۷-۹۹,۸۳	Accuracy	۹۹,۹۴	۹۹,۹۶	۹۹,۹۵	۹۹,۹۶	۹۹,۹۴	۹۹,۹۶	---	---	---	---	۹۹,۹۵
		Error rate	۰,۰۵	۰,۰۳	۰,۰۴	۰,۰۳	۰,۰۵۲	۰,۰۳۸	---	---	---	---	۰,۰۴
		F-Measure	۸۰,۳۹	۸۸,۶۰	۸۴,۵۰	۸۷,۷۳	۸۳,۴۷	۸۷,۵۶	---	---	---	---	۸۵,۳۷
		G-Mean	۹۷,۲۴	۹۸,۷۰	۹۷,۷۳	۹۹,۰۹	۹۷,۵۴	۹۸,۷۲	---	---	---	---	۹۸,۱۷
		Precision	۶۹,۸۸	۸۱,۲۲	۷۵,۷۳	۷۹,۲۶	۷۴,۳۳	۷۹,۴۷	---	---	---	---	۷۶,۶۴
		Sensitivity	۹۴,۶۱	۹۷,۴۶	۹۵,۵۷	۹۸,۲۳	۹۵,۱۸	۹۷,۵۰	---	---	---	---	۹۶,۴۲
		Specificity	۹۹,۹۴	۹۹,۹۶	۹۹,۹۵	۹۹,۹۶	۹۹,۹۵	۹۹,۹۶	---	---	---	---	۹۹,۹۵
		AUC	۹۷,۲۸	۹۸,۷۱	۹۷,۷۶	۹۹,۱۰	۹۷,۵۶	۹۸,۷۳	---	---	---	---	۹۸,۱۹
		Time	۸۳۵,۳		۲۰۵۸,۶		۴۱۳۵,۷		---		---		۲۳۴۳,۲

پارامتر زمان، میانگین بهتری نسبت به شبکه‌های عصبی- مصنوعی دارد.

• روش پیشنهادی در معیار دقّت نیز نسبت به روش‌های جنگل تصادفی، انفیس و بخصوص شبکه‌های عصبی مصنوعی، که حتّی از نظر مدّت زمان اجرا، میانگین به نسبت نزدیک به روش پیشنهادی دارد، نتایج خوبی را ارائه داده است که میانگین دقّت بالای ۹۲ درصد برای روش پیشنهادی بیانگر این نتیجه‌گیری است امّا نسبت به روش ارائه شدهٔ مشابه که C4.5 نام دارد، این دقّت کم است. البته توجیه‌پذیر است که هستهٔ الگوریتم پیشنهادی که الگوریتم جنگل تصادفی است با هستهٔ روش C4.5، متفاوت است.

• از روش پیشنهادی می‌توان برای داده‌های بسیار بزرگ استفاده کرد. روش پیشنهادی، به راحتی مرحله پیش‌پردازش داده‌ها را انجام داده و سپس با دقّت خوبی، رده‌بندی داده‌های نامتوازن با میزان توازن مختلف را انجام می‌دهد.

۹. چالش‌ها و پیشنهادها

در هر تحقیق علمی، باید نقاط و مسیرهای آینده مشخص باز باشد تا آن تحقیق، پیشرفت کرده و رشد نماید. در این تحقیق نیز مسیرها و نقاط حساسی برای بررسی و پیشرفت بیشتر وجود دارد که به شرح ذیل بیان می‌گردند.

• در این تحقیق از دستور تشکیل ساختار انفیس براساس خوشه‌یابی شعاعی استفاده شده است که مقدار ثابت ۰/۵ برای پارامتر شعاع در نظر گرفته شده است؛ امّا با توجه به این که برای هر مجموعه داده می‌تواند این شعاع تغییر کند، نیاز به بررسی بیشتر بر روی این پارامتر است.

به دلیل این که در روش پیشنهادی، ساخت نمونه‌های تصادفی برای ردهٔ اقلیت صورت می‌گیرد، می‌توان برای بهبود این روش، از روش پیش‌پردازش داده‌ها استفاده کرد که نمونه‌های تصادفی دقیق‌تر و مشابه‌تری نسبت به ردهٔ اقلیت، تولید کنند.

• در روش پیشنهادی، در زمان انتخاب یک زیر مجموعه از مجموعهٔ اصلی داده‌ها، برای تشکیل یک انفیس، بعضی از صفات خاصه انتخاب نمی‌شوند؛ می‌توان بر روی این موضوع که چگونه می‌توان بدون افزایش پیچیدگی زمانی و مدّت زمان اجرای الگوریتم و روش پیشنهادی، کلیهٔ صفات خاصه در تشکیل ساختار انفیس و نتایج آن سهیم باشند، تحقیق کرد.

مراجع

[1]H. Sain and S. W. Purnami, “Combine Sampling Support Vector Machine for Imbalanced Data Classification,” Procedia Comput. Sci., vol. 72, pp. 59–66, 2015.

[2]S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, “Handling imbalanced datasets: A review,” GESTS Int. Trans. Comput. Sci. Eng., vol. 30, no. 1, pp. 25–36, 2006.

[3]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.

[4]J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano, “Experimental perspectives on learning from imbalanced data,” in Proceedings of the 24th international conference on Machine learning, 2007, pp. 935–942.

[5]A. D’Addabbo and R. Maglietta, “Parallel selective sampling method for imbalanced and large data classification,” Pattern Recognit. Lett., vol. 62, pp. 61–67, 2015.

[6]C. Jian, J. Gao, and Y. Ao, “A new sampling method for classifying imbalanced data based on support vector machine ensemble,” Neurocomputing, vol. 193, pp. 115–122, 2016.

[7]J. Fan, Z. Niu, Y. Liang, and Z. Zhao, “Probability Model Selection and Parameter Evolutionary Estimation for Clustering Imbalanced Data without Sampling,” Neurocomputing, 2016.

[8]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.

[9]S. Vluymans, I. Triguero, C. Cornelis, and Y. Saeys, “EPRENNID: An evolutionary prototype reduction based ensemble for nearest neighbor classification of imbalanced data,” Neurocomputing, vol. 216, pp. 596–610, 2016.

[10]T. M. Khoshgoftaar, M. Golawala, and J. Van Hulse, “An empirical study of learning from imbalanced data using random forest,” in 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007), 2007, vol. 2, pp. 310–317.

[11]L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[12]J. Thomas, P.-E. Jouve, and N. Nicoloyannis, “Optimisation and evaluation of random forests for imbalanced datasets,”

2014.
[20] H. Zhao, X. Chen, T. Nguyen, J. Z. Huang, G. Williams, and H. Chen, “Stratified Over-Sampling Bagging Method for Random Forests on Imbalanced Data,” in Pacific-Asia Workshop on Intelligence and Security Informatics, 2016, pp. 63–72.
[21] S. Datta and S. Das, “Near-Bayesian Support Vector Machines for imbalanced data classification with equal or unequal misclassification costs,” Neural Networks, vol. 70, pp. 39–52, 2015.
[22] G. Haixiang, L. Yijing, L. Yanan, L. Xiao, and L. Jinling, “BPSO-Adaboost-KNN ensemble learning algorithm for multiclass imbalanced data classification,” Eng. Appl. Artif. Intell., vol. 49, pp. 176–193, 2016.
[23] 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.
[24] W. A. Rivera and P. Xanthopoulos, “A priori synthetic over-sampling methods for increasing classification sensitivity in imbalanced data sets,” Expert Syst. Appl., vol. 66, pp. 124–135, 2016.
[25] 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.
in International Symposium on Methodologies for Intelligent Systems, 2006, pp. 622–631.
[13]D. R. Cutler et al., “Random forests for classification in ecology,” Ecology, vol. 88, no. 11, pp. 2783–2792, 2007.
[14]P. Bonissone, J. M. Cadenas, M. C. Garrido, and R. A. Díaz-Valladares, “A fuzzy random forest,” Int. J. Approx. Reason., vol. 51, no. 7, pp. 729–747, 2010.
[15]M. Khalilia, S. Chakraborty, and M. Popescu, “Predicting disease risks from highly imbalanced data using random forest,” BMC Med. Inform. Decis. Mak., vol. 11, no. 1, p. 1, 2011.
[16]A. Verikas, A. Gelzinis, and M. Bacauskiene, “Mining data with random forests: A survey and results of new tests,” Pattern Recognit., vol. 44, no. 2, pp. 330–349, 2011.
[17]V. Y. Kulkarni and P. K. Sinha, “Pruning of random forest classifiers: A survey and future directions,” in Data Science & Engineering (ICDSE), 2012 International Conference on, 2012, pp. 64–68.
[18] S. del Río, V. López, J. M. Benítez, and F. Herrera, “On the use of MapReduce for imbalanced big data using Random Forest,” Inf. Sci. (Ny)., vol. 285, pp. 112–137, 2014.
[19] Q. Wu, Y. Ye, H. Zhang, M. K. Ng, and S.-S. Ho, “ForesTexter: an efficient random forest algorithm for imbalanced text categorization,” Knowledge-Based Syst., vol. 67, pp. 105–116,