وزارت علوم، تحقیقات و فناوری
دانشگاه تحصیلات تکمیلی علوم پایه
گاوازنگ - زنجان

طبقه‌بندی داده‌های نامتوازن با استفاده از ترکیب طبقه‌بندها و توصیف‌گرهای بردار پشتیبان

پایان‌نامهٔ کارشناسی ارشد

عباس پیرمحمدی

استاد راهنما: **دکتر علی امیری**

استاد مشاور: **دکتر محسن افشارچی**

۲۴ اردیبهشت ۱۳۹۳

تقدیم به آنهایی که می خواهند بیشتر بدانند.

چکیده

مسئله طبقه‌بندی داده‌های نامتوازن به‌عنوان یکی از چالش‌های اصلی در حوزه‌ی داده‌کاوی، مورد توجه بسیاري از محققان و پژوهش‌گران قرار گرفته است. در سال‌های اخیر تحقیقات ارزشمند زیادی برای حل مسئله طبقه‌بندی داده‌های نامتوازن انجام شده است. در بین این تحقیقات، رهیافت‌های مبتنی بر ترکیب طبقه‌بندها از موفقیّت قابل توجهی برخوردار بوده‌اند.

علیرغم کارهای بسیار مؤثر انجام شده در ترکیب طبقه‌بندها هنوز برخی چالش‌ها از قبیل؛ عدم توجه به اهمیت نمونه‌ها در متوازن سازی، تعیین تعداد مناسب طبقه‌بندها و عدم بهینه‌سازی وزن طبقه‌بندها در ترکیب طبقه‌بندها به‌صورت حل نشده باقی مانده است. بنابراین در این پایان‌نامه تلاش شده است رهیافتی برای چالش‌های مطرح شده ارائه شود. در این راستا، ترکیبی از الگوریتم‌ها جهت ایجاد ساختار پیشنهادی ارائه شده است. سیستم پیشنهادی شامل ۳ بخش کلی است: در بخش اول توصیف‌گر بردار پشتیبان داده جهت نمونه‌برداری از دو کلاس اقلیت و اکثریت استفاده شده است. این روش برخلاف روش‌های بیش نمونه‌برداری و زیر نمونه‌برداری از داده‌هایی که توصیف بهتری از کل داده‌ها دارد، نمونه‌برداری می‌کند. در بخش دوم تعداد بهینه طبقه‌بندهای پایه در ترکیب طبقه‌بندها برای نمونه‌های منتخب تعیین می‌شود و در بخش پایانی با استفاده از الگوریتم ژنتیک عمل بهینه‌سازی وزن‌ها در ترکیب طبقه‌بند انجام می‌شود. روش پیشنهادی با تعدادی از الگوریتم‌های موجود در این حوزه مقایسه شده و نتایج نشان دهنده مطلوبیت ساختار پیشنهادی در مقایسه با برخی از الگوریتم‌ها از لحاظ دقت است.

واژه‌های کلیدی: *شبیه‌سازی تبرید تدریجی، توصیف‌گر بردار پشتیبان داده، ترکیب طبقه‌بندها، الگوریتم ژنتیک*

فهرست

فهرست تصاویر

فصل اول

بیان مسئله

۱.۱ مقدمه

امروزه پردازش داده‌های با توزیع احتمال نامتوازن[۱]، به دلیل گستردگي آنها در بسیاری از مسائل دنیای واقعی، از قبیل تشخیص تقلّب [۱، ۲]، تشخیص ناهنجاری [۳]، تشخیص پزشکی [۴] و تشخیص نشت نفت [۵] ، مورد توجه بسیاري از محققان و پژوهش‌گران قرار گرفته است. در داده‌های نامتوازن، معمولاً تعداد نمونه‌های یکی از کلاس‌ها خیلی بیشتر از نمونه‌های کلاس دیگر است. کلاس با تعداد داده‌های بیشتر را کلاس اکثریت[۲] و کلاس با داده‌های کمتر را کلاس اقلیت[۳] می‌گوییم. در این مسائل معمولاً نسبت داده‌های کلاس اقلیت به اکثریت، اغلب ۱ به ۱۰۰ و یا بیشتر است [۶]. در داده‌های نامتوازن چالش اصلی شناسایی صحیح نمونه‌های کلاس اقلیت است. به‌عنوان مثال در حوزه پزشکی، تعداد نمونه‌های مثبت از یک بیماری در مقابل تعداد نمونه‌های منفی بسیار کمتر است. در حالی‌که

[۱] Imbalanced

[۲] Majority

[۳] Minority

اهمیت شناسایی نمونه‌های مربوط به دسته مثبت، بسیار زیاد است.

توزیع کلاس، نسبت نمونه‌های متعلق به هر کلاس در یک مجموعه داده، نقش کلیدی در طراحی طبقه‌بند ایفا می‌کند، در الگوریتم‌های استاندارد طبقه‌بندی داده‌ها، توزیع کلاس‌ها متوازن در نظر گرفته می‌شود. از این‌رو در صورت استفاده از این الگوریتم‌ها در طبقه‌بندی داده‌های نامتوازن، نمی‌توان به نتایج قابل قبولی دست یافت؛ زیرا در این الگوریتم‌ها طبقه‌بند به سمت نمونه‌هاي آموزشی کلاس بزرگ‌تر متمایل می‌شود که این موضوع سبب افزایش تعداد خطاها در شناسایی نمونه‌هاي مثبت می‌شود.

بنابراین در سال‌های اخیر مسئله‌ی طبقه‌بندی داده‌های نامتوازن به عنوان یکی از چالش‌های اصلی در حوزه‌ی داده‌کاوی، توجه بسیاری از محققّین را به خود جلب نموده است [۷]. در این پایان‌نامه، مسئله طبقه‌بندی داده‌های نامتوازن مورد توجه قرار گرفته است. بیان دقیق مسئله و چالش‌های اصلی پایان‌نامه، طرح کلی رهیافت پیشنهادی و ساختار کلی فصل‌های آتی در ادامه این فصل آمده است.

۲.۱ بیان مسئله و چالش‌های پیش‌رو

مسئله اصلی این پایان‌نامه طراحی یک طبقه‌بند برای داده‌های با توزیع نامتوازن است. همان‌طور که در بخش قبل اشاره شد، استفاده از طبقه‌بندهای استاندارد نمی‌تواند برای حل این مسئله مؤثر باشد. در سال‌های اخیر تحقیقات ارزشمند زیادی برای حل این مسئله ارائه شده است. در بین روش‌های ارائه شده، رهیافت‌های مبتنی بر ترکیب طبقه‌بندها از موفقیّت قابل توجهی برخوردار بوده‌اند. از جمله این الگوریتم‌ها می توان $AdaBoost$ [۸]، $SMOTEBoost$ [۹]، $RUSBoost$ [۱۰] و غیره اشاره کرد. علیرغم کارهای بسیار موثر انجام شده در ترکیب طبقه‌بندها هنوز چالش‌های زیر به‌صورت حل نشده باقی مانده است:

۱. الگوریتم‌های ترکیب طبقه‌بندها عمدتاً از روش‌های زیر نمونه‌برداری[۱] و بیش نمونه‌برداری[۲] برای متوازن نمودن داده‌ها استفاده می‌کنند. این کار با افزایش تصادفی تعداد نمونه‌های اقلیت یا کاهش تعداد نمونه‌های اکثریت صورت می‌گیرد. بدیهی است افزایش بی‌رویه و کورکورانه نمونه‌ها می‌تواند منجر به تولید و گسترش برخی نمونه‌ها که در توصیف نقش موثری را ندارند، گردد. بدیهی است داده‌هایی که در مرز داده‌ها قرار می‌گیرد نسبت به داده‌هایی که در مرکز داده‌ها قرار دارند از اهمیت بالایی برخوردار هستند. عدم توجه به اهمیت نمونه‌ها در متوازن‌سازی داده‌ها یک نقطه ضعف در کارهای گذشته بوده و از این‌رو به‌عنوان یکی از چالش‌های اصلی در کار ما قرار گرفته است.

۲. در الگوریتم‌های ترکیب طبقه‌بند ارائه شده در کارهای قبل، تعداد طبقه‌بندهای پایه و اهمیت آنها مورد توجه قرار نگرفته است. معمولاً بیشتر کارهای انجام شده با تعداد طبقه‌بند پایه ثابت به حل مسئله پرداخته‌اند و هیچ توجهی به انتخاب بهینه تعداد آنها نشده است.

۳. در بیشتر کارهای گذشته، رأی‌گیری طبقه‌بندها معمولاً به‌صورت وزن‌دار و بر اساس قدرت تصمیم آن روی کل داده‌ها تعیین می‌شد. تنظیم این وزن‌ها جهت افزایش قدرت تصمیم کل طبقه‌بند مرکب به‌دست آمده تاکنون مورد توجه قرار نگرفته است.

۳.۱ رهیافت پیشنهادی

در این پایان‌نامه تلاش شده است رهیافتی برای چالش‌های مطرح شده در بخش قبل ارائه شود. در این راستا، ترکیبی از الگوریتم‌ها جهت ایجاد ساختار پیشنهادی ارائه شده است.

سیستم پیشنهادی شامل ۳ بخش کلی است: در بخش اول جهت نمونه‌برداری از هر دو کلاس اقلیت و

---

[۱] Under sampling

[۲] Over sampling

اکثریت از روش SVDD استفاده شده است. این روش بر خلاف سایر روش‌های نمونه‌برداری که از کل مجموعه داده‌ها نمونه‌برداری می‌کنند، تنها از نمونه‌هایی که در مرز داده‌ها قرار دارند نمونه‌برداری می‌کند. در بخش دوم روش پیشنهاد شده توسط هرناندز-لباتو[۱] و همکارانش روی نمونه‌های منتخب اعمال می‌شود که منجر به تعیین حداقل تعداد طبقه‌بندهای پایه در ترکیب طبقه‌بندها می‌شود. در بخش پایانی با استفاده از الگوریتم ژنتیک عمل بهینه‌سازی وزن‌ها در ترکیب طبقه‌بند انجام می‌شود.

۴.۱ ساختار پایان‌نامه

این پایان‌نامه شامل شش فصل است. در فصل اول، مسئله اصلی و چالش‌های پیش‌رو بیان شد. در فصل دوم، داده‌های نامتوازن و الگوریتم‌های طبقه‌بندی داده‌های نامتوازن مطالعه شده است. مفهوم توصیف‌گر بردار پشتیبان داده و مفهوم شبیه‌سازی تبرید تدریجی[۲] و سایر مفاهیم مورد نیاز در فصل سوم آمده است. در فصل چهارم رهیافت پیشنهادی برای طبقه‌بندی داده‌های نامتوازن بیان شده است. در فصل پنجم نتایج آزمایشات آمده است و فصل ششم شامل نتیجه‌گیری و کارهای آتی است.

---

[۱] Hernandez-Lobato

[۲] Simulated Annealing (SA)

فصل دوم

داده‌های نامتوازن

۱.۲ مقدمه

در سال‌های اخیر مشکل نامتوازن بودن کلاس‌ها مورد توجه محققّان در زمینه داده‌کاوی قرار گرفته است. در موارد متعددی کلاسی که از نقطه نظر دامنه کاربردی، اهمیت زیادی دارد (کلاس اصلی) شامل تعداد نمونه‌های کمتری نسبت به کلاس اکثریت است.

طبقه‌بندی داده‌های نامتوازن یک مسئله رایج در بسیاری از دامنه‌ها، از قبیل تشخیص معاملات تقلبی [۲]، تشخیص پزشکی [۴] و تشخیص نشت نفت [۵] است. در همه سناریوها وقتی کلاس اکثریت شامل ۹۹-۹۸ درصد از کل نمونه‌ها باشد، یک طبقه‌بند ساده در مرحله یادگیری نمونه‌های کلاس اقلیت را نادیده می‌گیرد و در مرحله آزمایش هر چیزی را از کلاس اکثریت برچسب‌دهی می‌کند، بنابراین این طبقه‌بند می‌تواند به بالاترین دقت ممکن برسد. معیارهایی از قبیل؛ آنالیز منحنی مشخصه عملکرد

دریافت کننده[۱] ، دقت [۲] ، فراخوانی[۳] ، اندازه‌گیری-F [۴] و میانگین-G [۵] برای درک کارایی الگوریتم یادگیری روی مجموعه داده‌های نامتوازن استفاده می‌شود. با توجه به اهمیت این نوع داده‌ها، امروزه الگوریتم‌های متعددی برای طبقه‌بندی داده‌های نامتوازن بوجود آمده است. بدین منظور در ادامه فصل، نگاهی بر داده‌های نامتوازن خواهیم داشت. سپس چگونگی مدیریت این نوع داده‌ها و روش‌های مبتنی بر ترکیب طبقه‌بندها را بررسی خواهیم نمود و در آخر معیارهای اندازه‌گیری کارایی در دامنه نامتوازن را بیان خواهیم کرد.

۲.۲ مجموعه داده‌های نامتوازن

مشکل توزیع نامتوازن داده در کلاس‌ها در مواقعی مطرح می‌شود که نمونه‌های موجود در برخی کلاس‌ها از سایر کلاس‌ها بیشتر باشد، این مسئله بخصوص در کاربردهای دوکلاسی مطرح است که یك کلاس از نمونه‌های زیادی نسبت به کلاس دیگر برخوردار است. در تعداد زیادی از حوزه‌های داده‌کاوی، مشکل عدم توازن کلاس‌ها یکی از مسائل مهم می‌باشد.

الگوریتم‌های معمول داده‌کاوی در مواجهه با مشکل کلاس‌های نامتوازن معمولاً عمل‌کرد ضعیفی دارند. زیرا با صرف‌نظر کردن از نمونه‌های داده در کلاس اقلیت، اقدام به افزایش دقت کلی 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 یك مسئله تشخیص پزشکی با وجود اینکه نمونه‌های بیماری در مقایسه با نمونه‌های معمول در اقلّیت قرار دارند؛ ولی هدف از طبقه‌بندی در واقع یافتن

---

۱ Receiver Operating Characteristic)(ROC)

۲ Precision

۳ Recall

۴ F-measure

۵ G-mean

نمونه‌های بیماری است. در مجموعه داده‌ای با نرخ عدم توازن ۱:۱۰۰ (یعنی با فرض وجود یك نمونه کلاس اقلیت به ازای ۱۰۰ نمونه از کلاس اکثریت) روش‌های طبقه‌بندی استاندارد با نادیده گرفتن تمامی نمونه‌های اقلیت به یك طبقه‌بندی با دقت بالای ٪۹۹ دست می‌یابد، در صورتی که در طبقه‌بندی داده‌های نامتوازن با توجه به اهمیت بالای شناخت نمونه‌های کلاس اقلیت، طبقه‌بندی صحیح آنها خیلی مهم است [۶، ۹، ۱۱] .

در اغلب الگوریتم‌های طبقه‌بندی، تمایل در جهت کلاسی است که بیشترین تعداد نمونه‌ها را دارند، از این‌رو توانایی کمی در پیش‌گویی صحیح داده‌های کلاس اقلیت از خود نشان می‌دهند. در چنین الگوریتم‌هایی، نمونه‌های کلاس اقلیت نسبت به نمونه‌های کلاس اکثریت نادرست طبقه‌بندی می‌شوند [۱۲، ۱۳]. مسائل مرتبط در شکل‌گیری داده‌های نامتوازن عبارتند از [۶] :

۱. کم بودن تعداد نمونه‌ها : به‌طور کلی در مجموعه داده‌های نامتوازن، تعداد نمونه‌های اقلیت به اندازه کافی نیست.

۲. تداخل یا کلاس تفکیک‌پذیر[۱] : این حالت زمانی رخ می‌دهد که قوانین متمایز کردن کلاس‌ها سخت باشد. در نتیجه، قوانین کلی‌تر سبب می‌شوند داده‌های کلاس اقلیت نادرست طبقه‌بندی شوند [۱۱]. اگر در هم تداخلی بین کلاس‌ها نباشد، آنگاه هر طبقه‌بند ساده با توجه به توزیع کلاس می‌تواند به‌طور مناسب یاد بگیرد. شکل ۱.۲ بیان‌گر این حالت می‌باشد.

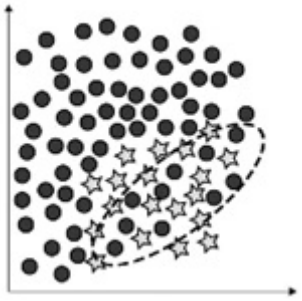
۳. جدا کننده‌های کوچک[۲] : جدا کننده‌های کوچک، جدا کننده‌هایی هستند که در طبقه‌بندی یادگیرنده تعدادی از نمونه‌های آموزشی را پوشش می‌دهند [۱۴]. وجود جدا کننده‌های کوچک در یک مجموعه داده، زمانی روی می‌دهد که مفهوم ارائه شده توسط کلاس اقلیت متشکل از زیرمفهوم‌ها [۳] باشد [۱۵] . بنابراین، در اغلب مسائل شاهد حضور این زیر مفهوم‌ها هستیم.

---

[۱] Overlapping or class separability

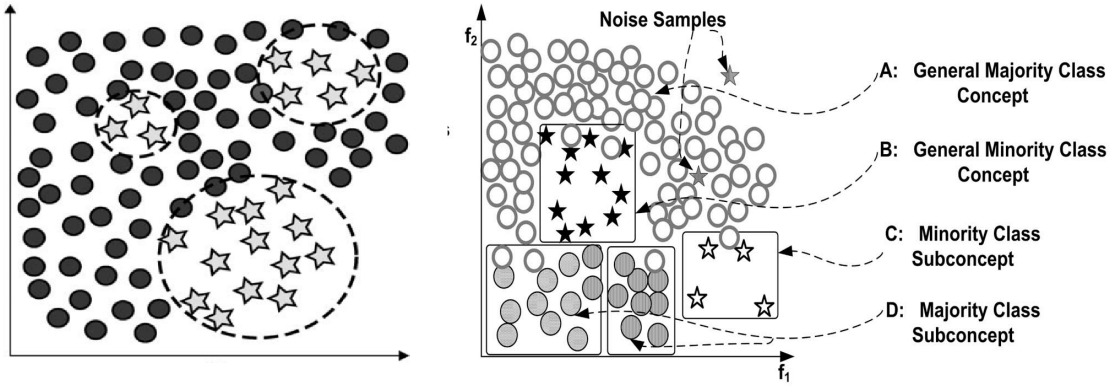
[۲] Small disjuncts

[۳] Subconcepts



شکل ۱.۲: در هم تداخل کلاسی [۶]

وجود زیر مفهوم‌ها، پیچدگی مسئله را افزایش می‌دهد، زیرا تعداد نمونه‌ها معمولاً میان آنها به‌طور متعادل نمی‌باشد. نمودار سمت چپ شکل ۲.۲ ازمتشکل سه زیر مفهوم از کلاس اقلیت است.



شکل ۲.۲: نمودار جدا کننده‌های کوچک [۶،۱۳]

۳.۲ روش‌های مواجهه با مسئله داده‌های نامتوازن

با توجه به اهمیت مسئله مجموعه داده‌های نامتوازن در برنامه‌های کاربردی، تعداد زیادی از روش‌ها در برخورد با این مسئله توسعه داده شده‌اند. در این ادبیات، تکنیک‌های متنوعی برای حل مسئله در ارتباط

با کلاس نامتوازن پیشنهاد شده است، که در سه گروه: روش سطح داده[۱] ، روش سطح الگوریتم[۲] و روش حساس به هزینه[۳] تقسیم می شوند [۶]. در ادامه به توصیف هر کدام از این روش‌ها خواهیم پرداخت.

۱.۳.۲ روش‌های سطح داده

در این روش با اضافه کردن مرحله پیش‌پردازش[۴] قبل از طبقه‌بندی، موجب متوازن شدن مجموعه داده‌های نامتوازن می‌شود. در روش پیش پردازش داده از روش نمونه‌برداری استفاده می‌شود. نمونه برداری به دو شکل است، زیر نمونه‌برداری از کلاس اکثریت، بیش نمونه‌برداری از کلاس اقلیت، یا ترکیبی از هر دو روش است.

۱. زیر نمونه‌برداری: یکی از معروفترین روش‌های نمونه‌برداری به‌طور تصادفی است. در زیر نمونه‌برداری به‌طور تصادفی نمونه‌هایی از کلاس اکثریت حذف می شود، تا زمانی که کلاس اقلیت درصدی از کلاس اکثریت شود؛ به این ترتیب توازن در مجموعه آموزشی برقرار می‌شود. شکل ۳.۲ بیانگر این روش می‌باشد . از معایب این روش، از دست دادن اطلاعات با ارزش است و موجب زیر برازش[۵] داده‌های کلاس اکثریت می‌شود.

۲. بیش نمونه‌برداری: این روش برای به‌تعادل رساندن توزیع کلاس، از روش جایگزینی نمونه‌های کلاس اقلّیت استفاده می‌کند، که نیازی به اطلاعات اضافی ندارد و از داده‌های موجود دوباره استفاده می‌کند تا توازن در مجموعه آموزشی برقرار شود. شکل ۴.۲ نمایش روش بیش نمونه‌برداری

---

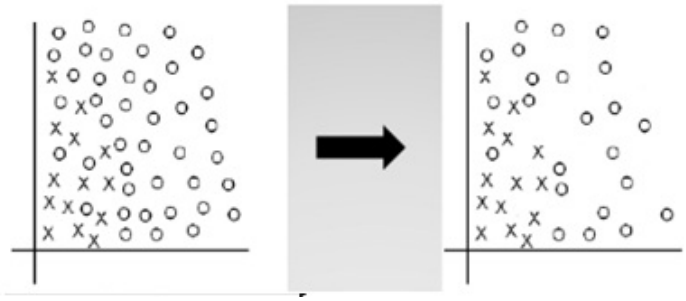
۱ Data level

۲ Algorithm level

۳ Cost-sensitive

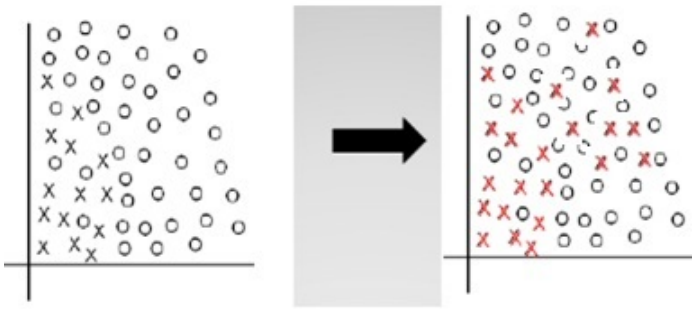
۴ Preprocessing

۵ Under fitting



شکل ۳.۲: حذف تصادفی نمونه‌های کلاس اکثریت [۱۶]

است. از معایب این روش باعث بیش برازش[۱] مدل آموزشی می‌شود و همچنین روش بیش نمونه‌برداری اندازه مجموعه آموزشی را افزایش می‌دهد که این امر منجر به افزایش زمان یادگیری طبقه‌بند می‌شود.



شکل ۴.۲: جایگزینی تصادفی نمونه‌های کلاس اقلّیت (نمودار سمت راستی اصلاح شده است.) [۱۷]

۳. روش بیش نمونه‌برداری مصنوعی کلاس اقلّیت (SMOTE)[۲] الگوریتم SMOTE داده‌های مصنوعی را بر اساس شباهت فضای ویژگی بین نمونه‌های کلاس اقلیت ایجاد می‌کند. برای زیر مجموعه $S_{min} \in S$ ، K تا از نزدیکترین همسایه نمونه $x_i \in S_{min}$ را انتخاب می‌کنیم که k یک عدد صحیح می‌باشد. نحوه انتخاب k تا از همسایه‌های نزدیک نمونه مورد نظر براساس

---

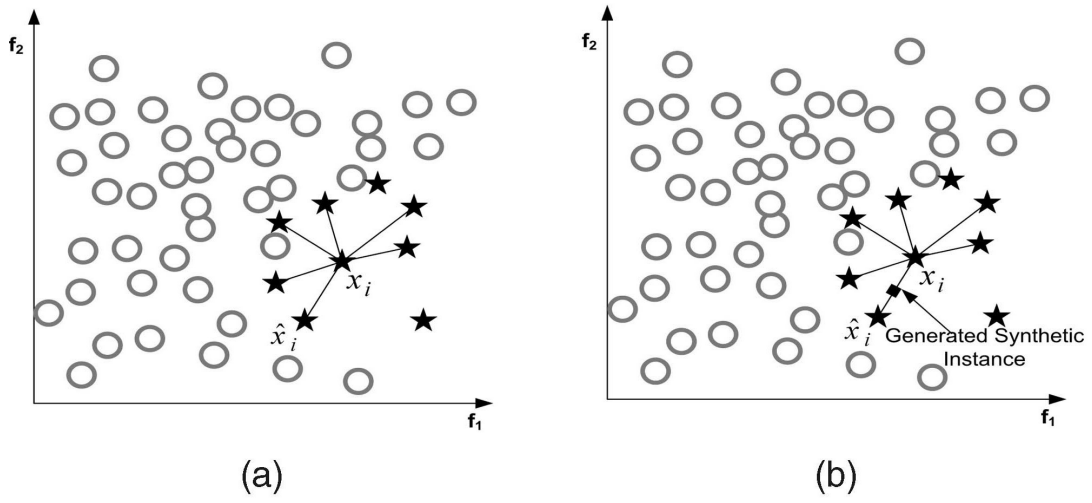
[۱] Overfitting

[۲] Synthetic Minority Oversampling Technique (SMOTE)

فاصله اقلیدسی در فضای n بعدی می‌باشد. برای ایجاد نمونه مصنوعی، یکی از k همسایه‌های نزدیک نمونه x_i به‌طور تصادفی انتخاب می‌کنیم، سپس با ضرب اختلاف این دو نمونه در یک عدد تصادفی بین [۰،۱] و اضافه کردن نتیجه حاصل شده به نمونه x_i ، نمونه جدید به‌دست می‌آید.

$$x_{new} = x_i + (\hat{x}_i - x_i) \times \delta$$

که $\hat{x}_i$ یکی از k همسایه نزدیک نمونه x_i است و $\delta \in [0, 1]$ یک عدد تصادفی است. نمونه جدید در بخشی از خط متصل کننده بین دو نمونه قرار می‌گیرد [۱۷].



شکل ۵.۲: نمودار a) مثالی از ۶ همسایه نزدیک نمونه x_i و نمودار b) ایجاد داده براساس فاصله اقلیدسی [۱۳]

۴. اصلاح شده روش بیش نمونه‌برداری مصنوعی کلاس اقلیت (MSMOTE)[۱] نسخه اصلاح شده روش SMOTE است. این الگوریتم با محاسبه مرکز نمونه‌های کلاس اقلّیت و تعیین فاصله اقلیدسی هر یک از نمونه‌ها نسبت به مرکز نمونه‌ها، آن‌ها را در سه گروه، امن، مرزی و نمونه‌های نویزدار نهان طبقه‌بندی می‌کند. روش‌های تولید نمونه‌های مصنوعی به‌وسیله MSMOTE

[۱] Modified Synthetic Minority Oversampling Technique (MSMOTE)