





دانشگاه آزاد اسلامی

واحد دولت آباد

دانشکده فنی مهندسی

پایان نامه برای دریافت درجه کارشناسی ارشد در رشته کامپیوتر

گرایش نرم افزار

عنوان

طبقه بندی داده های پزشکی نامتوازن چند کلاسه بر اساس الگو

کاوی عمیق



معاونت پژوهش و فن آوری

به نام خدا

مشور اخلاق پژوهش

بیاباری از خداوند بجهان و اعتماد به اینکه عالم محضر خداست و همواره ناظر بر اعمال انسان و به منظور پادداشت مقام بلند دانش و پژوهش و نظریه اهمیت جایگاه دانشگاه در
اعتلای فرهنگ و تمدن بشری و دانشجویان و اعضاء هیئت علمی و احدی دانشگاه آزاد اسلامی متعهد می گردیم اصول زیر را در انجام فعالیت های پژوهشی مد نظر قرار داده و
از آن تخلفی نکنیم:

- ۱- اصل حقیقت جویی: تلاش در راستای پی جویی حقیقت و وفاداری به آن و دوری از هرگونه پنهان سازی حقیقت
- ۲- اصل رعایت حقوق: التزام به رعایت کامل حقوق پژوهشگران و پژوهشگران (انسان، حیوان و نبات) و سایر صاحبان حق
- ۳- اصل مالکیت مادی و معنوی: تعهد به رعایت کامل حقوق مادی و معنوی دانشگاه و کلیه بکاران پژوهش
- ۴- اصل منافع ملی: تعهد به رعایت مصالح ملی و در نظر داشتن پیشبرد و توسعه کشور در کلیه مراحل پژوهش
- ۵- اصل رعایت انصاف و امانت: تعهد به اجتناب از هرگونه جانب داری غیر علمی و حفاظت از اموال، تجهیزات و منابع در اختیار
- ۶- اصل رازداری: تعهد به صیانت از اسرار و اطلاعات محرمانه افراد سازمان ها و کشور و کلیه افراد و ستادهای مرتبط با تحقیق
- ۷- اصل احترام: تعهد به رعایت حریم ها و حرمت ها در انجام تحقیقات و رعایت جانب تقد و خودداری از هرگونه حرمت شکنی
- ۸- اصل ترویج: تعهد به رواج دانش و اشاعه نتایج تحقیقات و انتقال آن به بکاران علمی و دانشجویان به غیر از مواردی که منع قانونی دارد
- ۹- اصل برانست: التزام به برانست جویی از هرگونه رفتار غیر حرفه ای و اعلام موضع نسبت به کسانی که حوزه علم و پژوهش را به مثابه ای غیر علمی می آلائند

تعهدنامه اصالت رساله یا پایان نامه

اینجانب زهرا فروزان فر دانش آموخته مقطع کارشناسی ارشد در رشته مهندسی کامپیوتر

که در تاریخ..... از پایان نامه / رساله خود با عنوان طبقه‌بندی داده‌های پزشکی نامتوازن چند کلاسه بر اساس الگو کاوی عمیق با کسب نمره و درجه دفاع نموده‌ام بدینوسیله متعهد می‌شوم:

(۱) این پایان نامه / رساله حاصل تحقیق و پژوهش انجام شده توسط اینجانب بوده و در مواردی که از دستاوردهای علمی و پژوهشی دیگران (اعم از پایان نامه، کتاب، مقاله و ...) استفاده نموده‌ام، مطابق ضوابط و رویه موجود، نام منبع مورد استفاده و سایر مشخصات آن را در فهرست مربوطه ذکر و درج کرده‌ام.

(۲) این پایان نامه / رساله قبلاً " برای دریافت هیچ مدرک تحصیلی (هم سطح، پائین‌تر یا بالاتر) در سایر دانشگاه‌ها و مؤسسات آموزش عالی ارائه نشده است .

(۳) چنانچه بعد از فراغت از تحصیل، قصد استفاده و هرگونه بهره‌برداری اعم از چاپ کتاب، ثبت اختراع و ... از این پایان نامه داشته باشم، از حوزه معاونت پژوهشی واحد مجوزهای مربوطه را اخذ نمایم.

(۴) چنانچه در هر مقطع زمانی خلاف موارد فوق ثابت شود، عواقب ناشی از آن را می‌پذیرم و واحد دانشگاهی مجاز است با اینجانب مطابق ضوابط و مقررات رفتار نموده و در صورت ابطال مدرک تحصیلی ام هیچ‌گونه ادعایی نخواهم داشت.

نام و نام خانوادگی:

تاریخ و امضاء:

فهرست مطالب

عنوان.....	صفحه.....
چکیده	۱
فصل اول کلیات تحقیق	
۱-۱- مقدمه	۲
۱-۲- بیان مسئله	۳
۱-۳- ضرورت تحقیق	۴
۱-۴- اهداف پایان نامه	۵
۱-۵- نوآوری پایان نامه	۶
۱-۶- ساختار پایان نامه	۶
فصل دوم ادبیات و پیشینه تحقیق	
۲-۱- مقدمه	۷
۲-۲- داده‌های نامتوازن	۷
۲-۳- الگوریتم‌های طبقه‌بندی مجموعه داده‌های نامتوازن	۹
۲-۴- یادگیری ماشین	۱۰
۲-۵- انتخاب ویژگی	۱۲
۲-۵-۱- گام‌های حل مسئله انتخاب ویژگی	۱۴
۲-۵-۲- دسته‌بندی روش‌های انتخاب ویژگی	۱۵
۲-۵-۲-۱- الگوریتم جستجو	۱۶
۲-۵-۲-۱-۱- جهت جستجو	۱۶
۲-۵-۲-۱-۲- تکنیک جستجو	۱۷
۲-۵-۲-۱-۳- معیار ارزیابی	۱۸
۲-۶- شبکه عصبی	۱۹
۲-۶-۱- ساختار شبکه عصبی	۲۱

- ۲-۶-۲- پرسپترون و الگوریتم پس انتشار خطا..... ۲۱
- ۲-۷- الگوریتم فاخته..... ۲۲
- ۲-۸- روش‌های مبتنی بر اتوماتای یادگیر..... ۲۴
- ۲-۸-۱- اتوماتای یادگیر..... ۲۶
- ۲-۹- کارهای پیشین..... ۲۷
- ۲-۱۰- جمع‌بندی..... ۳۰

فصل سوم روش پیشنهادی

- ۳-۱- مقدمه..... ۳۱
- ۳-۲- روش پیشنهادی..... ۳۱
- ۳-۳- پایگاه داده..... ۳۱
- ۳-۴- پیش‌پردازش..... ۳۳
- ۳-۵- انتخاب ویژگی در روش پیشنهادی..... ۳۳
- ۳-۵-۱- نمایش محل سکونت فاخته..... ۳۵
- ۳-۵-۲- یافتن راه حل‌های جدید و پرواز لوی..... ۳۶
- ۳-۵-۳- تابع هزینه..... ۳۶
- ۳-۵-۴- الگوریتم فاخته و پارامترهای آن..... ۳۷
- ۳-۶- طبقه‌بندی در روش پیشنهادی..... ۳۹

فصل چهارم شبیه‌سازی و ارزیابی نتایج

- ۴-۱- مقدمه..... ۴۲
- ۴-۲- معیارهای ارزیابی..... ۴۲
- ۴-۳- مشخصات پایگاه داده..... ۴۵
- ۴-۴- انتخاب ویژگی..... ۴۵
- ۴-۵- طبقه‌بندی با استفاده از شبکه عصبی و اتوماتای یادگیر..... ۴۶
- ۴-۶- مقایسه نتایج..... ۴۷

۴-۷- جمع بندی ۵۱

فصل پنجم نتیجه گیری و پیشنهادات آتی

۵-۱- نتیجه گیری ۵۲

۵-۲- پیشنهادات آتی ۵۳

منابع ۵۴

فهرست شکل‌ها

عنوان.....	صفحه.....
شکل (۲-۱): روال کلی الگوریتم‌های انتخاب ویژگی.....	۱۴
شکل (۲-۲): دسته‌بندی روش‌های انتخاب ویژگی.....	۱۶
شکل (۲-۳): شبکه عصبی مصنوعی.....	۲۰
شکل (۲-۴): شبه کد الگوریتم جستجوی فاخته.....	۲۳
شکل (۲-۵): مثالی از مسیر پرواز لوی.....	۲۴
شکل (۳-۱): فرآیند روش پیشنهادی.....	۳۲
شکل (۳-۲): جریان انتخاب ویژگی مبتنی بر رپر.....	۳۴
شکل (۳-۳): فلوچارت انتخاب ویژگی.....	۳۵
شکل (۳-۴): نمونه‌ای از نمایش ویژگی‌ها در روش پیشنهادی.....	۳۶
شکل (۳-۵): اتصال شبکه عصبی اتوماتا.....	۴۰
شکل (۴-۱): الگوریتم طبقه‌بند به ازای ۲۹ تکرار.....	۴۶
شکل (۴-۲): الگوریتم طبقه‌بند به ازای ۵۰ تکرار.....	۴۷
شکل (۳-۴): مقایسه روش پیشنهادی با استفاده و بدون استفاده از انتخاب ویژگی.....	۴۸
شکل (۴-۴): مقایسه روش پیشنهادی با استفاده از الگوریتم‌های انتخاب ویژگی.....	۴۹
شکل (۴-۵): مقایسه عملکرد شبکه عصبی و شبکه عصبی بهینه شده با اتوماتای یادگیر.....	۴۹
شکل (۴-۶): مقایسه عملکرد روش پیشنهادی با کارهای پیشین.....	۵۰
شکل (۴-۷): نمودار ROC مقایسه الگوریتم‌ها.....	۵۱

فهرست جداول

عنوان.....	صفحه.....
جدول (۳-۱): ویژگی های مجموعه داده.....	۳۳.....
جدول (۳-۲) : پارامترهای انتخاب ویژگی.....	۳۸.....
جدول (۴-۱): ماتریس درهم ریختگی.....	۴۳.....
جدول (۴-۲): مشخصات مجموعه داده.....	۴۵.....
جدول (۴-۳): نتایج روش پیشنهادی در محیط متلب.....	۴۷.....
جدول (۴-۴): مقایسه روش های انتخاب ویژگی.....	۴۸.....
جدول (۴-۵): مقایسه روش پیشنهادی با الگوریتم های متداول یادگیری ماشین.....	۵۱.....

چکیده

طبقه‌بندی داده‌ها در دنیای واقعی همواره با مشکل نامتوازن بودن مجموعه داده‌ها روبرو است. این مشکل هنگامی رخ می‌دهد که تعداد نمونه‌ها در کلاس‌های گوناگون اختلاف بسیاری داشته باشند. نحوه برخورد با داده‌های نامتوازن به عنوان یک مسئله چالش‌برانگیز در داده‌کاوی شناخته می‌شود، به نحوی که شناسایی داده‌های رده کوچک‌تر با اشتباهات بسیاری روبرو است. بنابراین انتخاب ویژگی‌های منحصربفرد و مناسب جهت طبقه‌بندی داده‌های با رده کوچک‌تر، چالش اساسی این پژوهش می‌باشد. امروزه پردازش داده‌های نامتوازن پزشکی، به دلیل گستردگی آن‌ها در بسیاری از مسائل دنیای واقعی، مورد توجه محققان قرار گرفته است. در این پژوهش نیز هدف، ارائه‌ی روشی جهت طبقه‌بندی داده‌های نامتوازن پزشکی می‌باشد. در این پایان نامه از مجموعه داده کم کاری تیروئید از مخزن UCI استفاده شده است. به منظور دستیابی به یک زیرمجموعه بهینه از ویژگی‌ها در مرحله انتخاب ویژگی از الگوریتم ماشین بردار پشتیبان به عنوان تابع هزینه و الگوریتم فاخته به عنوان استراتژی جستجو استفاده شده است. روش پیشنهادی روشی با دقت بالا است که با استفاده از بهینه‌سازی شبکه عصبی توسط اتوماتای یادگیر به دقت ۹۹,۶ درصد برای طبقه‌بندی داده‌ها دست یافته است.

کلمات کلیدی: طبقه‌بندی، داده‌های نامتوازن، شبکه عصبی، اتوماتای یادگیر

فصل اول

کلیات تحقیق

۱-۱-مقدمه

پیشرفت‌های بوجود آمده در جمع‌آوری داده‌ها و قابلیت‌های ذخیره‌سازی در طی دهه‌های اخیر باعث شده در بسیاری از علوم با حجم بزرگی از اطلاعات روبه‌رو شویم. داده‌کاوی کوششی برای بدست آوردن اطلاعات مفید از میان این داده‌هاست. رشد بی‌رویه‌ی داده‌ها در سطح جهان اهمیت داده‌کاوی را دو چندان کرده است. طبقه‌بندی یکی از قابلیت‌های داده‌کاوی است که امروزه دارای کاربرد بسیار وسیعی در حوزه‌های مختلف است به گونه‌ای که حد و مرزی برای کاربرد این دانش در نظر گرفته نشده است. زمینه‌های کاری این دانش را از ذرات کف اقیانوس‌ها تا اعماق فضا می‌دانند. داده‌کاوی در حوزه‌ی پزشکی یکی از ارزش‌ترین و در عین حال پرچالش‌ترین کاربردها در زمینه‌ی کشف دانش محسوب می‌شود [۱].

در دو دهه گذشته، رشد سریع داده‌ها در پزشکی از چالش‌های اساسی می‌باشد. با توجه به پیشرفت سریع روش‌های بیوتکنولوژی و روش‌های تجزیه و تحلیل داده‌های زیستی، روش‌های داده‌کاوی در حال رشد می‌باشد. امروزه برای تجزیه و تحلیل و پیش‌بینی داده‌های بزرگ پزشکی از الگوریتم‌های داده‌کاوی و یادگیری ماشین استفاده می‌شود. هدف اصلی از تجزیه و تحلیل این داده‌ها، توسعه مدل‌های مختلف پیش‌بینی در حوزه پزشکی می‌باشد. تشخیص بیماری یک موضوع مهم در تحقیقات فعلی است. ویژگی اصلی مدل تشخیص

بیماری، تصمیم‌گیری سریع پزشکان و کاهش اشتباهات در تشخیص می‌باشد و داده‌های پزشکی را خیلی سریع بررسی می‌کند. همچنین از آنجایی که داده‌های پزشکی شامل تعداد زیادی ویژگی است و محققان تنها زیرمجموعه‌ای از این ویژگی‌ها را در نظر می‌گیرند و این ویژگی‌ها برای تشخیص بیماری حیاتی است، بنابراین بررسی این ویژگی‌های مهم باید بسیار دقیق باشد [۲]. همچنین باید در برخورد با نمونه‌های اقلیت دقت شود. به همین منظور در این پایان‌نامه از الگوریتم ترکیبی انتخاب ویژگی و همچنین الگوریتم بهینه شده شبکه عصبی برای طبقه‌بندی نمونه‌ها استفاده شده است.

۲-۱- بیان مسئله

طبقه‌بندی^۱ داده‌های نامتوازن^۲ به دلیل اهمیت آن‌ها در بسیاری از کاربردهای دنیای واقعی مورد توجه است. یک مجموعه داده نامتوازن گفته می‌شود اگر تعداد نمونه‌هایی که معرف یک کلاس هستند توسط نمونه‌های سایر کلاس‌ها مغلوب شود. امروزه پردازش داده‌های نامتوازن، به دلیل گستردگی آن‌ها در بسیاری از مسائل دنیای واقعی، مورد توجه بسیاری از محققان قرار گرفته است. مسائل طبقه‌بندی نامتوازن باینری، مسائلی هستند که شامل دو کلاس هستند که تعداد نمونه‌های یک کلاس بسیار کمتر از کلاس دیگر است. طبقه‌ی مثبت یا اقلیت به طبقه‌ای گفته می‌شود که تعداد نمونه‌های کمتری دارد؛ درحالی که طبقه‌ی منفی یا اکثریت دارای تعداد نمونه‌های بیشتری است. الگوریتم‌های یادگیری کلاسیک بر روی مجموعه داده‌های نامتوازن، نتایج نسبتاً بهینه‌ای را تولید می‌کند، زیرا هدف نهایی این الگوریتم‌ها، بیشینه کردن دقت می‌باشد. به همین علت، این دسته از الگوریتم‌ها هنگام برخورد داده‌های نامتوازن در کل دقت بالایی دارند، اما بیشتر مواقع کارایی پایینی بر روی کلاس اقلیت که اغلب مهم‌ترین کلاس در داده‌های نامتوازن است، خواهند داشت و این در حالی است که اهمیت نمونه‌های اقلیت به مراتب بیشتر از نمونه‌های اکثریت است. به همین علت این الگوریتم‌ها برای پردازش داده‌های نامتوازن مناسب نمی‌باشند [۲]. تاکنون روش‌هایی برای کار روی این داده‌ها ارائه شده است. روش‌های

¹ Classification

² Imbalanced data

ارائه شده به سه دسته‌ی کلی نمونه‌گیری از داده‌ها^۱، اصلاح الگوریتم‌های موجود^۲ و یادگیری حساس به هزینه^۳ تقسیم می‌شوند.

هدف از ارائه‌ی روش‌های فوق افزایش دقت روی کلاس اقلیت است. به این دلیل که نمونه‌های کلاس اقلیت دارای اهمیت بیشتری هستند و لذا طبقه‌بندی درست آن‌ها بسیار مهم می‌باشد. هر یک از این روش‌ها سیاستی را در جهت برخورد با نامتوازن بودن داده‌ها دنبال می‌کنند. در نمونه‌گیری از داده‌ها، هدف ایجاد توازن در توزیع داده‌ها است. در این روش پس از متوازن کردن مجموعه داده‌های آموزشی می‌توان طبقه‌بندی استاندارد را برای طبقه‌بندی داده‌ها به کار برد. در دسته‌ی دوم هدف این است که با بهبود الگوریتم‌های موجود و یا با ارائه‌ی روش‌هایی جدید، مسائل مربوط به طبقه‌بندی داده‌های نامتوازن را حل کنند [۳]. در این پایان‌نامه بر روی دسته‌ی دوم تمرکز شده است و هدف این است که با کمک الگوکاو و با ایجاد بهبودی در الگوریتم طبقه‌بندی شبکه عصبی، یک روش مناسب برای طبقه‌بندی داده‌های نامتوازن بر مبنای این الگوریتم ارائه داده شود. برای ارزیابی روش پیشنهادی از مجموعه داده در دسترس عموم مخزن UCI، بیماری کم کاری تیروئید استفاده شده است. ابتدا با استفاده از ترکیب الگوریتم فاخته و ماشین بردار پشتیبان زیرمجموعه‌ای بهینه از ویژگی‌ها بدست آمده و سپس با استفاده از الگوریتم شبکه عصبی بهینه شده به طبقه‌بندی داده‌ها پرداخته می‌شود.

۳-۱- ضرورت تحقیق

طبقه‌بندی داده‌های نامتوازن یک مسأله مورد رقابت در زمینه‌ی داده کاوی است که در سال‌های اخیر مورد توجه زیادی قرار گرفته است. دلیل اهمیت این مسأله کاربرد عمده و برجسته آن در بسیاری از مسائل دنیای واقعی از جمله تشخیص بیماری‌ها، بیوانفورماتیک، مدیریت ریسک و ... است. در مجموعه داده‌های نامتوازن تعداد نمونه‌های یک یا چند کلاس بسیار کمتر از سایر کلاس‌ها است. مسائل طبقه‌بندی نامتوازن باینری، مسائلی

¹ Datasampling

² Algorithm modification

³ Cost sensitive learning

هستند که شامل دو کلاس هستند که تعداد نمونه‌های یک کلاس بسیار کمتر از کلاس دیگر است. طبقه‌ی مثبت یا اقلیت به طبقه‌ای گفته می‌شود که تعداد نمونه‌های کمتری دارد؛ درحالی که طبقه‌ی منفی یا اکثریت دارای تعداد نمونه‌های بیشتری است. بیشتر الگوریتم‌های یادگیری روی کلاس اکثریت دقت بالایی را به دست می‌آورند اما در مورد کلاس اقلیت ضعیف عمل می‌کنند و اغلب داده‌های تستی که متعلق به کلاس کوچکتر هستند اشتباه طبقه‌بندی می‌شوند و این در حالی است که طبقه‌ی مثبت یا اقلیت مهمتر از کلاس اکثریت است و در صورتی که نادرست طبقه‌بندی شود، منجر به هزینه‌ی بیشتری می‌شود [۴].

۴-۱- اهداف پایان نامه

در این پایان نامه، هدف این است که با استفاده از الگوریتمی کارا مبتنی بر رویکردهایی در سطح الگوریتم، به گونه‌ای مجموعه داده‌های نامتوازن طبقه‌بندی شوند تا تعداد نمونه‌های کلاس‌های اقلیتی که به درستی طبقه‌بندی می‌شوند، افزایش یابند. همچنین توانایی طبقه‌بند برای تشخیص نمونه‌های کلاس اکثریت و کمینه نمودن خطا کلی، در سطح قابل قبولی حفظ شود. با توجه به اهمیت مسئله طبقه‌بندی داده‌های نامتوازن پزشکی در این پایان‌نامه برای رسیدن به پاسخ این مسئله یک راه‌حل با اهداف زیر ارائه شده است.

- هدف اصلی

طبقه‌بندی داده‌های نامتوازن پزشکی چندکلاسه بر اساس الگوکاوی عمیق

- اهداف فرعی

یافتن ویژگی‌های بهینه با استفاده از الگوریتم فاخته و ماشین بردار پشتیبان

طبقه‌بندی داده‌های نامتوازن با استفاده از شبکه عصبی بهینه شده با اتوماتای یادگیر

۵-۱- نوآوری پایان‌نامه

الگوریتم‌های موجود و سنتی اکثراً تمایل به سمت نمونه‌های کلاس اکثریت دارند و نمونه‌های کلاس اقلیت را معمولاً به عنوان نویز یا داده پرت در نظر می‌گیرند. این مشکل یعنی برخورد با مشکل عدم توازن کلاس با

استفاده از الگوریتم‌های بهبودیافته از این جهت مهم می‌باشد که الگوریتم‌های طراحی شده در این زمینه تمایل به طرف نمونه‌های کلاس اقلیت داشته و آن‌ها را بر خلاف الگوریتم‌های سنتی نادیده نمی‌گیرند بلکه به هدف اصلی که رسیدن به نمونه‌های اقلیت است نزدیک می‌شوند. در همین راستا، در مرحله‌ی طبقه‌بندی از شبکه عصبی به‌مراه اتوماتای یادگیر بهره خواهیم برد. شبکه عصبی مصنوعی یک سامانه پردازشی داده‌ها است. در این شبکه‌ها به کمک دانش برنامه‌نویسی، ساختار داده‌ای طراحی می‌شود که می‌تواند همانند نرون عمل کند. که به این ساختار داده نرون گفته می‌شود. بعد با ایجاد شبکه‌ای بین این نرون‌ها و اعمال یک الگوریتم آموزشی به آن، شبکه را آموزش می‌دهند، تنظیمات شبکه عصبی مانند تعداد لایه‌ها، تعداد نرون‌ها عموماً با سعی و خطا انجام می‌شود لذا اگر سیستمی طراحی گردید که تعداد لایه‌ها و تعداد نرون‌ها شبکه را بتواند مشخص نماید، شبکه عصبی کارایی بهتری دارد. در این پژوهش قصد داریم به کمک شبکه عصبی داده‌های نامتوازن را طبقه‌بندی کنیم. سپس به کمک یادگیری تقویتی پارامترهای شبکه عصبی تنظیم می‌گردد تا داده‌های نامتوازن با دقت بهتر طبقه‌بندی شوند. الگوریتم مورد استفاده در یادگیری تقویتی الگوریتم اتوماتای یادگیر است که وظیفه آن تنظیم پارامتر شبکه عصبی مانند تعداد لایه‌ها، تعداد نرون، نوع نرون‌ها و غیره است.

۶-۱- ساختار پایان نامه

این پایان‌نامه در پنج فصل تنظیم شده است. در ادامه، فصل دوم به مرور راهکارهای ارائه شده‌ی قبلی در خصوص مقابله با چالش عدم توازن پرداخته است. فصل سوم به تشریح روش پیشنهادی پرداخته می‌شود که در آن با استفاده از شبکه عصبی و اتوماتای یادگیر، عملیات طبقه‌بندی صورت می‌گیرد. فصل چهارم به بررسی و ارزیابی نتایج حاصل از روش پیشنهادی بر روی مجموعه داده استاندارد مخزن UCI کم‌کاری تیروئید پرداخته می‌شود. در نهایت در فصل پنجم نتیجه‌گیری و کارهای آینده را خواهیم داشت.

فصل دوم

ادبیات و پیشینه تحقیق

۱-۲- مقدمه

تاکنون روش‌های متعددی جهت طبقه‌بندی مجموعه داده‌های نامتوازن، معرفی شده است. در این فصل، برخی از روش‌های مهم و برجسته در طبقه‌بندی مجموعه داده‌های نامتوازن، به اختصار بیان می‌شوند. ابتدا به بیان مبانی تحقیق پرداخته شده و سپس کارهای مرتبط مورد بررسی قرار گرفته‌اند.

۲-۲- داده‌های نامتوازن

مشکل داده‌های نامتوازن زمانی ایجاد می‌شود که تعداد نمونه‌های یک کلاس از کلاس دیگر به طور قابل توجهی بیشتر هستند. کلاس اقلیت در داده‌های آموزشی کمتر است و این منجر به این می‌شود که خطای این کلاس مقدار کمی داشته باشد. اغلب الگوریتم‌ها تلاش می‌کنند که مدل‌های خوبی برای تعیین کلاس اکثریت پیدا کنند و یا در شرایط بهتر میان کلاس اقلیت و اکثریت توازن ایجاد کنند. نتیجه این خواهد شد که الگوریتم‌هایی که به کلاس اکثریت اولویت می‌دهند دارای دقت بالاتری هستند و در نگاه اول به نظر بهتر از الگوریتم‌هایی هستند که میان دو کلاس توازن برقرار می‌کنند [۵].