



بررسی داده های نامتوازن و انتخاب گروه مناسب طبقه بندی با استفاده از روش های جمعی برای داده های مربوط به سرطان سینه

ابراهیم بهشتی نیا^۱، دکتر هاله همایونی^۲، دکتر منصور امینی لاری^۳

^۱ دانشجوی کارشناسی ارشد مهندسی کامپیوتر، موسسه آموزش عالی آپادانا شیراز، ایران، S.beheshti66@gmail.com

^۲ عضو هیات علمی، موسسه غیرانتفاعی آپادانا شیراز، ایران، Haleh.homayouni@gmail.com

^۳ استاد یار و عضو هیئت علمی دانشگاه آزاد اسلامی واحد مرودشت، Aminilari@gmail.com

چکیده

یکی از مسائل مهم در زمینه داده کاوی، مسأله دسته بندی مجموعه داده های نامتوازن است. اصطلاح عموماً به مجموعه داده ای گفته می شود که تعداد نمونه ها در کلاس های «مجموعه داده نامتوازن» گوناگون، اختلاف بسیاری داشته باشند. در این نوع داده ها، به کلاس هایی که کم ترین تعداد نمونه ها را دارند، کلاس اقلیت گفته می شود. به دلیل این که اکثر الگوریتم های یادگیری، یک دسته بند را با فرض برابر بودن تعداد نمونه های آموزشی هر کلاس آموزش می دهند، بنابراین زمانی که این الگوریتم ها را به مجموعه داده های نامتوازن اعمال می کنیم، دسته بند آموزش داده شده، غالباً براساس نمونه های کلاس اکثریت آموزش می بیند. این موضوع به پیش بینی بسیار ضعیف نمونه های کلاس های اقلیت منجر می شود، زیرا آموزش کلاس اقلیت به درستی انجام نشده است. در این پایان نامه هدف مورد نظراین است که روش هایی برای دسته بندی ارائه شوند که علاوه بر بهبود دسته بندی در داده های کلاس های اقلیت، قابلیت دسته بندی داده های کلاس های اکثریت نیز، در سطح قابل قبولی حفظ شود. از آنجایی که در کارهای صورت گرفته پیشین عموماً از داده های مصنوعی برای کار دسته بندی استفاده شده است لذا در این پایان نامه داده های مربوط به سرطان سینه توسط نگارنده جمع آوری شده. روشهای پیشنهادی که بیشتر در قالب متدهای جمعی هستند بر روی این داده ها (دیتاست) با استفاده از نرم افزار راپیدماینر به مرحله اجرا، خروجی و در نهایت مقایسه بین روشها خواهیم پرداخت.

واژه های کلیدی

داده های نامتوازن، الگوریتم های طبقه بندی، روشهای جمعی، سرطان سینه

مقدمه

امروزه پردازش داده های با توزیع احتمال نامتوازن^۲، به دلیل گستردگی آنها در بسیاری از مسائل دنیای واقعی، از قبیل تشخیص تقلب تشخیص ناهنجاری، تشخیص پزشکی و تشخیص نشت نفت مورد توجه بسیاری از محققان و پژوهشگران قرار گرفته است. در دادههای نامتوازن، معمولاً تعداد نمونه های یکی از کلاس ها خیلی بیشتر از نمونه های کلاس دیگر است. کلاس با تعداد داده های

بیشتر را کلاس اکثریت^۳ و کلاس باداده های کمتر را کلاس اقلیت^۴ می گوئیم. در این مسائل معمولاً نسبت داده های کلاس اقلیت به اکثریت، اغلب ۱ به ۱۰۰ یا بیشتر است. در داده های نامتوازن چالش اصلی شناسایی صحیح نمونه های کلاس اقلیت است. به عنوان مثال در حوزه پزشکی، تعداد نمونه های مثبت از یک بیماری در مقابل تعداد نمونه های منفی بسیار کمتر است. در حالی که اهمیت شناسایی نمونه های مربوط به دسته مثبت، بسیار زیاد است. توزیع کلاسی، نسبت نمونه های متعلق به هر کلاس در یک مجموعه داده، نقش کلیدی در طراحی طبقه بند ایفا می کند، در الگوریتم های استاندارد طبقه بندی داده ها، توزیع کلاسها متوازن در نظر گرفته میشود. از این رو در صورت استفاده از این الگوریتم ها در طبقه بندی داده های نامتوازن، نمی توان به نتایج قابل قبولی دست یافت؛ زیرا در این الگوریتمها طبقه بند به سمت نمونه های آموزشی کلاس بزرگتر متمایل میشود که این موضوع سبب افزایش تعداد خطاها در شناسایی نمونه های مثبت می شود. بنابراین در سالهای اخیر مسئله ی طبقه بندی داده های نامتوازن به عنوان یکی از چالشی های اصلی در حوزه ی داده کاوی، توجه بسیاری از محققین را به خود جلب نموده است. در این پایان نامه، مسئله طبقه بندی داده های نامتوازن با استفاده از روش پیشنهادی (روش جمعی) مورد توجه قرار گرفته است.

اهمیت موضوع

کشف یک توده در پستان میتواند از مهمترین اتفاقات زندگی یک زن باشد که باعث اضطراب میشود. برخلاف تومورهای خوشخیم پستان، توده های بدخیم در صورت عدم تشخیص و درمان به موقع، بیوقفه رشد میکنند، به بافت های اطراف و حتی به نقاط دوردست گسترش می یابند و میتوانند در صورت عدم درمان منجر به مرگ شوند. امروزه دانشمندان علوم بهداشتی به واژه جدیدی به نام غربالگری دست یافته اند که هدف آن تشخیص زودرس، کاهش مرگ و میر، کاهش ناتوانیهای ناشی از بیماری و عوارض درمان است. اساس غربالگری در سرطان پستان، تشخیص این بیماری در زمانی است که توده کوچک بوده و هنوز به قسمت های دیگر بدن انتشار پیدا نکرده است. در مراحل اولیه، ممکن است بیمار کاملاً بدون علامت باشد و پس از مراجعه به منظور کنترل، شک به وجود سرطان