نشریه علمی- ترویجی محاسبات نرم
شمارهٔ هفتم/ بهار و تابستان ۹۴/ صفحهٔ ۷۴ـ۸۳

دریافت مقاله: ۱۳۹۴/۱/۲۲
پذیرش مقاله: ۱۳۹۴/۵/۱۵

پیش‌بینی ژن‌های بیماری با استفاده از دسته‌بند تک‌کلاسی ماشین بردار پشتیبان

اکرم وثیقی ذاکر*[۱]، سعید جلیلی[۲]

[۱] دانشجوی کارشناسی ارشد، دانشکده مهندسی برق و کامپیوتر، دانشگاه تربیت مدرس، تهران، ایران
a.vasighi@modares.ac.ir
[۲] دانشیار، دانشکده مهندسی برق و کامپیوتر، دانشگاه تربیت مدرس، تهران، ایران
sjalili@modares.ac.ir

چکیده: در بحث شناسایی و دسته‌بندی ژن‌های بیماری، متخصصان تنها به دسته‌بندی یک کلاس خاص، یعنی ژن‌های بیماری علاقه‌مند هستند و توجهی به کلاس‌های دیگر، یعنی ژن‌های غیربیماری ندارند. در مباحث یادگیری ماشین، این کاربرد تحت عنوان دسته‌بندی تک‌کلاسی مطرح می‌شود. روش‌های موجود مبتنی بر یادگیری معمولاً ژن‌های شناخته‌شدهٔ بیماری را به‌عنوان مجموعهٔ آموزشی مثبت و ژن‌های ناشناخته را به‌عنوان مجموعهٔ آموزشی منفی برای ایجاد یک دسته‌بند دودویی استفاده می‌کنند. ازآنجایی‌که در پایگاه داده‌های موجود در علم ژنتیک، مجموعه ژن غیربیماری وجود ندارد، ما در این مقاله با استفاده از روش دسته‌بند تک‌کلاسی ماشین بردار پشتیبان[۱](OCSVM)، و درنظرگرفتن تنها ژن‌های بیماری‌زا اقدام به شناسایی داده‌های بدون برچسب می‌کنیم. روش پیشنهادی نسبت به روش‌های موجود از لحاظ معیارهای دقت[۲]، بازخوانی[۳] و معیار F_1[۴] نتایج بهتری را ارائه می‌دهد.

واژه‌های کلیدی: شناسایی ژن‌های بیماری، دسته‌بندی تک‌کلاسی ماشین بردار پشتیبان.

1. One Class SVM
2. Precision
3. Recall
4. F-measure

۱. مقدمه

در پزشکی و داروشناسی، درک مکانیسم بیماری برای درمان مؤثر آن یک مسئلهٔ حیاتی است. دربارهٔ اختلالات ارثی، پیدا کردن ژن‌های عامل بیماری اولین گام مهم است. بیماری‌های ژنتیکی انسان معمولاً از اختلال ایجادشده در وظیفه یک یا چند ژن از کل ژنوم انسانی به‌وجود می‌آیند. در مراحل مختلف هر پروژهٔ تحقیقاتی، زیست‌شناسان مولکولی نیاز به انتخاب ژن یا پروتئین دارند (اغلب تا حدودی تصادفی، حتی پس از تجزیه و تحلیل داده‌های آماری دقیق) تا آن را بیشتر مورد بررسی تجربی قرار دهند و این کار به‌دلیل محدودیت منابع است. در اینجا ژن‌های شناخته‌شدهٔ عامل بیماری را به اختصار ژن‌های بیماری و بقیهٔ ژنوم را ژن‌های غیربیماری یا ژن‌های ناشناخته می‌نامیم.

در اصل، به‌دلیل دانش کم بشر تا به امروز، ژن‌های غیربیماری به‌صورت مرجعی معرفی نشده‌اند، زیرا با توجه به شرایط و موقعیت‌های خاص که ممکن است وجود داشته باشد، ژنی که امروزه به‌عنوان ژن غیربیماری شناخته می‌شود، ممکن است در آینده به‌عنوان ژن بیماری ثبت شود. در اصل، هدف از آزمایشات آزمایشگاه‌های زیستی، تشخیص ژن‌های عامل بیماری است و بعد از بررسی‌های زمان‌بر و پرهزینهٔ بسیار، به‌صورت قطعی و یا در برخی موارد به‌صورت احتمالاتی می‌توان یک ژن را برای اهداف درمانی و تولید دارو به‌عنوان ژن بیماری معرفی کرد. اما معرفی یک ژن، به‌عنوان ژن غیر بیماری در اصل در آزمایشات آن‌ها تعریف نشده است. در واقع مشخص شدن اینکه یک ژن، ژن بیماری باشد نیاز به آزمایشات زمان‌بر و هزینه‌بر زیادی دارد و ژن‌هایی را هم که روش‌های محاسباتی به‌عنوان ژن بیماری معرفی می‌کنند، با احتمالی ژن بیماری معرفی می‌شوند و برای بررسی دقیق‌تر به متخصصان زیست‌شناسی ارجاع داده می‌شوند.

به دلایل ذکرشده، کشف ژن‌های بیماری جدید هنوز به‌عنوان یک چالش مطرح است. روش‌های محاسباتی می‌توانند ژن‌ها را برای مطالعات جزئی‌تر بعدی پیش‌بینی کنند. چنین روش‌هایی می‌توانند برای سرعت بخشیدن به شناسایی ژن‌های بیماری جدید از تعداد بی‌شمار ژن‌های ناشناخته کمک کنند و این برای محققان بسیار ارزشمند است.

امروزه تعدادی از ژن‌ها ثابت شده‌اند که عامل بیماری هستند. این یک منبع ارزشمند برای توسعهٔ روش‌های یادگیری ماشین برای شناسایی ژن‌های بیماری جدید با استفاده از ژن‌های تأییدشدهٔ عامل بیماری به‌عنوان نمونه‌های آموزشی مثبت فراهم کرده است.

هدف از این پژوهش، بررسی روش‌های مطرح‌شده برای مسئلهٔ پیش‌بینی ژن‌های بیماری و همچنین ارائهٔ روشی برای دسته‌بندی مناسب و صحیح ژن‌هاست؛ به‌طوری‌که از بین ژن‌های بدون برچسب، ژن‌های محتمل‌تر بیماری (به عبارت دیگر داده‌های دستهٔ مثبت) را با استفاده از یک دسته‌بند جدا کنیم. این روش با مفهوم موجود در مسئله (دسته‌بندی یک کلاس خاص، یعنی ژن‌های بیماری) سازگاری بیشتری داشته و از منظری متفاوت نسبت به کارهای پیشین ژن‌های بیماری را پیش‌بینی می‌کند.

در بخش بعد، روش‌های مختلف انجام‌شده در این زمینه را بررسی و مقایسه‌ای از عملکرد آن‌ها ارائه خواهیم داد. سپس در بخش ۳ روش پیشنهادی را مطرح و در پایان در بخش‌های ۴ و ۵ مقایسه‌ای بر روی نتایج حاصل از این روش با روش‌های پیشین و نتیجه‌گیری روش انجام‌شده را خواهیم داشت.

۲. تاریخچهٔ پژوهش

براساس اصل «ارتباط براساس شباهت[1]»، بیماری‌های مشابه یا یکسان دارای ژن‌های عاملی هستند که با یکدیگر ارتباط فیزیکی (مستقیم) یا عملکردی (غیرمستقیم) دارند، (و در شبکه‌های برهم‌کنش پروتئین-پروتئین[2] تشکیل ماژول‌های فیزیکی یا عملکردی داده‌اند)، برخی روش‌ها برای پیش‌بینی ژن‌های کاندید بیماری توسعه داده شده‌اند که در آن‌ها ژن‌ها با ویژگی‌های زیستی مختلف مانند توالی پروتئین، داده‌های حاشیه نویسی عملکردی پروتئین، داده‌های بیان ژن و شبکه‌های

1. Guilt-by-proximity
2. Protein-Protein Interaction

برهم‌کنش پروتئین-پروتئین، PPI مشخص شده‌اند. این روش‌ها در سه دسته قرار می‌گیرند: ۱. روش‌های بدون‌نظارت، ۲. روش‌های با نظارت، ۳. روش‌های نیمه‌نظارتی.

روش‌های بدون نظارت ژن‌های کاندید بیماری را براساس نزدیکی و شباهت آن‌ها به ژن‌های شناخته‌شده بیماری، با استفاده از روش‌های مختلف رتبه‌دهی اولویت‌بندی می‌کنند. اکثر این روش‌ها تنها یک لیست اولویت‌بندی‌شده از ژن‌های کاندید بیماری در اختیار قرار می‌دهند، درحالی‌که نیاز به یک حد آستانه برای تعیین اینکه آیا یک ژن متعلق به ژن‌های بیماری هست یا خیر، همچنان احساس می‌شود. یعنی تا حدی بتوان مرز بین ژن‌های بیماری و غیربیماری را مشخص‌تر کرد. بنابراین روش‌هایی برای دسته‌بندی ژن‌ها پیشنهاد شد.

Adie و همکاران [۱]، الگوریتم درخت تصمیم را روی انواع مختلفی از ویژگی‌های ژنومیکی اعمال کردند. Xu و همکاران [۲] دسته‌بند KNN (K نزدیک‌ترین همسایه) را روی $K=1$ و $K=5$ برای پیش‌بینی ژن‌های بیماری براساس ویژگی‌های توپولوژیکی شبکه‌های PPI، مانند درجهٔ پروتئین‌ها، درصد ژن‌های بیماری در همسایگی پروتئین‌ها و... به‌کار گرفتند. Smalter و همکاران [۳] دسته‌بند ماشین بردار پشتیبان (SVM) را با استفاده از ویژگی‌های توپولوژیکی PPI، ویژگی‌های مشتق‌شده از توالی و... به‌کار گرفتند. Radivojac و همکاران [۴] با استفاده از ترکیب سه دسته‌بند دودویی SVM روی سه نوع ویژگی ژن، شبکهٔ PPI، توالی پروتئین و اطلاعات عملکردی پروتئین یک دسته‌بند دودویی برای پیش‌بینی ژن‌های بیماری ایجاد کردند.

روش‌های ذکرشده دسته‌بندهای دودویی را با ژن‌های شناخته‌شدهٔ بیماری‌ها به‌عنوان مجموعه آموزشی مثبت و بقیهٔ ژن‌ها را به‌عنوان مجموعه آموزشی منفی به‌کار می‌گیرند. اما ازآنجاکه هویت واقعی داده‌های بدون برچسب (مجموعه ژن‌های ناشناخته) مشخص نیست، در نتیجهٔ درنظرگرفتن آن‌ها به‌عنوان داده‌های منفی صحیح نیست (در واقع بخشی از آن‌ها منفی و بخشی دیگر مثبت هستند که به اشتباه موارد مثبت نیز منفی در نظر گرفته می‌شود). بنابراین کارایی مطالعات انجام‌شده تاکنون با اشکال همراه است. این نوع دسته‌بندها در واقع از مجموعه نویزدار منفی که ممکن است شامل ژن‌های بیماری باشند، ساخته می‌شوند. درنتیجه، این دسته‌بندها نمی‌توانند آن‌طور که باید عملکرد خوبی داشته باشند.

به این دلیل، روش‌های دیگری مجموعه ژن‌های ناشناخته را به‌عنوان مجموعهٔ بدون برچسب (به‌جای یک مجموعه منفی) در نظر گرفتند و تکنیک‌های یادگیری PU[1] را روی مجموعه مثبت و بدون برچسب اعمال کردند و بنابراین به نتایج بهتری رسیدند.

برای نمونه، Mordelet و همکاران [۵] یک روش یکپارچه[2] به نام ProDiGe ارائه دادند که به‌طور تصادفی، زیرمجموعه‌هایی از مجموعه بدون برچسب انتخاب می‌کند و سپس چندین دسته‌بند با استفاده از BSVM (Bias SVM) برای جداکردن نمونه‌های مثبت از هرکدام از زیرمجموعه‌ها آموزش می‌دهد و این کار را تکرار می‌کند. دسته‌بند نهایی از تجمیع دسته‌بندهای جداگانه به‌دست می‌آید. روش ProDiGe با تمام نمونه‌های موجود در زیرمجموعه‌های تصادفی رفتار یکسانی دارد.

Yang و همکاران [۶] یک الگوریتم یادگیری PU چندسطحی را با محاسبهٔ شباهت بین نمونه‌ها در مجموعه بدون برچسب و نمونه‌های مثبت در مجموعهٔ مثبت‌ها پیشنهاد دادند. آن‌ها احتمال بودن نمونه‌های مثبت یا منفی در مجموعهٔ بدون برچسب را با استفاده از الگوریتم RWR[3] تخمین زدند و سپس یک دسته‌بند چندسطحی وزن‌دار SVM (WSVM) با استفاده از این مجموعه‌ها به علاوهٔ مجموعه مثبت ایجاد کردند و به نتایج بهتری از روش‌های دیگر رسیدند.

۳. روش پیشنهادی

داده‌های بدون برچسب ما آن بخشی از ژنوم انسانی است که جزء لیست ژن‌های شناخته‌شدهٔ بیماری نیستند، اما جزء لیست ژن‌هایی هستند که در شبکهٔ PPI در ارتباط با ژن‌های بیماری هستند. ما این ژن‌ها را طبق اصل «ارتباط براساس شباهت»،

1. Positive-Unlabeled
2. Bagging
3. Random Walk with Restart

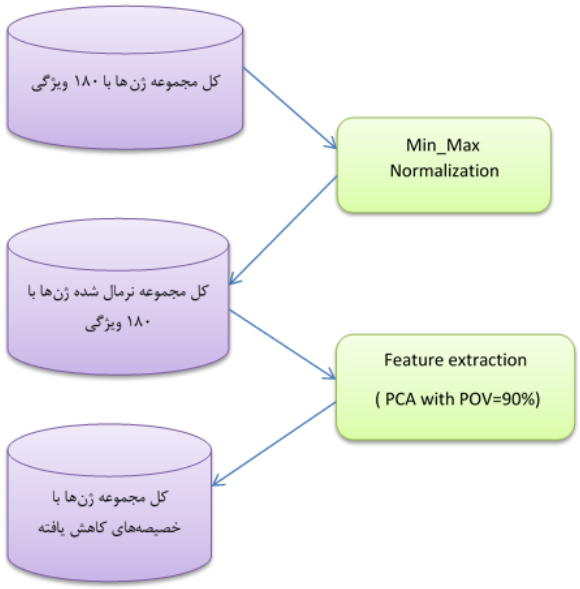
ژن‌های محتمل‌تری برای کاندید ژن بیماری بودن در نظر می‌گیریم، زیرا ایجاد اختلال در مسیرهای مختلفی که یک ژن درگیر است، باعث به‌وجودآمدن بیماری‌هایی با مکانیسم مشابه، مانند انواع بیماری‌های عصبی یا سرطان می‌شود. در مسئلهٔ ما، ژن‌های شناخته‌شدهٔ بیماری به‌عنوان داده‌های مثبت در اختیار هستند، اما متأسفانه داده‌های منفی در اختیار نیست.

همان‌طور که در بخش کارهای پیشین ارائه شد، روش‌های موجود در دستهٔ نظارتی، داده‌های بدون برچسب را به‌عنوان دادهٔ منفی در نظر گرفته و یک دسته‌بند دوکلاسی یادگیری می‌کنند. اما ازآنجاکه هویت واقعی داده‌های بدون برچسب مشخص نیست، درنتیجه درنظرگرفتن آن‌ها به‌عنوان داده‌های منفی صحیح نیست. بنابراین دسته‌بندهای یادگرفته‌شده نمی‌توانند کارایی خوبی داشته باشند.

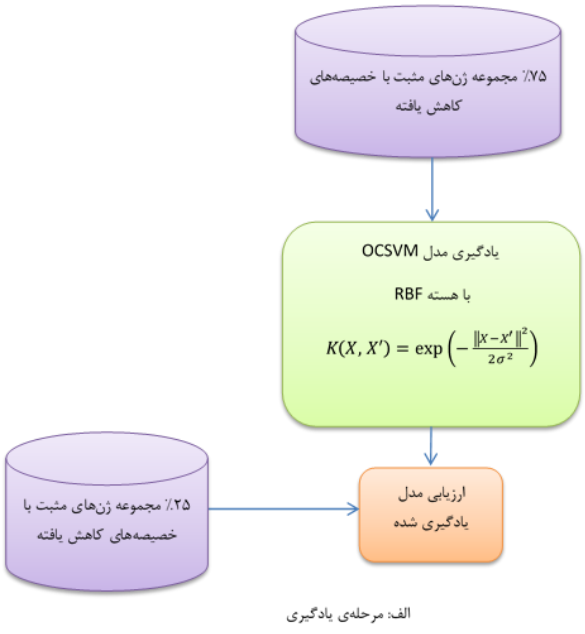
این نکته قابل توجه است که در دسته‌بندی تک‌کلاسی، داده‌های منفی یا ارائه نشده است یا به درستی نمونه‌برداری نشده است. ازآنجاکه تنها داده‌های ژن‌های بیماری (داده‌های مثبت) به درستی برچسب خورده‌اند، مسئلهٔ تشخیص ژن‌های جدید کاندید بیماری می‌تواند به‌عنوان یک مسئلهٔ دسته‌بندی تک‌کلاسی در نظر گرفته شود.

دسته‌بندهای تک‌کلاسی مانند دسته‌بند تک‌کلاسی ماشین بردار پشتیبان (OCSVM) به داده‌های با برچسب منفی در مجموعه آموزش نیازی ندارد و کاربرد آن در دسته‌بندی و شناسایی ژن‌های کاندید بیماری مورد مطالعه قرار نگرفته است. بنابراین بسیار مناسب است که یک مدل دسته‌بند تک‌کلاسی برای جداسازی مجموعه مثبت به‌عنوان ژن‌های بیماری از ژن‌های دیگر ایجاد کنیم که نرخ خطای آموزش را کاهش می‌دهد و این می‌تواند بهتر از یک دسته‌بند دوکلاسی با مجموعه آموزشی منفی غیرقابل اعتماد عمل کند.

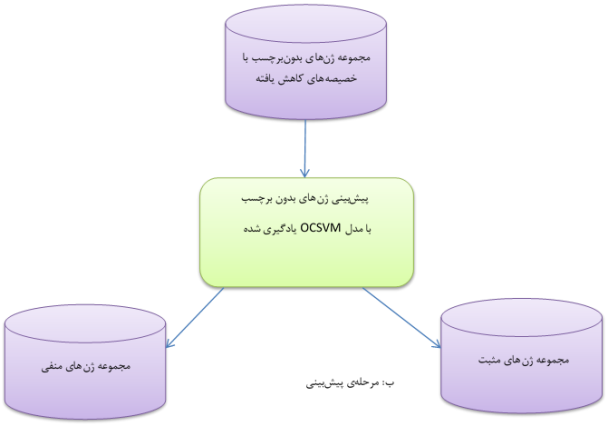
ابتدا کل مجموعهٔ داده‌ها را با روش Min-Max نرمال‌سازی کرده و سپس با استفاده از روش PCA تعداد ویژگی‌ها را کاهش دادیم. سپس پارامترهای دسته‌بند OCSVM را تنظیم کرده و سپس این دسته‌بند را با پارامترهای بهینهٔ به‌دست‌آمده یادگیری کرده و ژن‌های بدون برچسب را با استفاده از این دسته‌بند برچسب‌دهی کردیم. نمودار کلی استخراج ویژگی در شکل (۱) نشان داده شده است. همچنین نمودار کلی روش پیشنهادی در شکل‌های (۲) و (۳) ارائه شده است.



شکل (۱): نمودار کلی استخراج ویژگی‌ها در روش پیشنهادی



شکل (۲): نمودار کلی روش پیشنهادی (مرحلهٔ یادگیری)



شکل (۳): نمودار کلی روش پیشنهادی اول (مرحلهٔ پیش‌بینی)

ابتدا دسته‌بندی تک‌کلاسی را توضیح می‌دهیم و سپس روش پیشنهادی خود را ارزیابی می‌کنیم و در نهایت روش خود را با روش‌های معرفی‌شدهٔ قبلی مقایسه می‌کنیم. نتایج ما نشان می‌دهد که به‌کارگیری روش OCSVM برای تشخیص ژن‌های بیماری بهتر از روش‌های موجود نتیجه می‌دهد.

۱.۳. دسته‌بندی تک‌کلاسی

در مسئلهٔ یادگیری تک‌کلاسی یکی از کلاس‌ها (کلاس مثبت یا هدف) توسط نمونه‌های مجموعهٔ آموزش به‌خوبی مشخص شده و به‌درستی برچسب خورده است. در کلاس‌های دیگر (غیر هدف) یا هیچ‌گونه نمونه‌ای وجود ندارد، یا تعداد آن‌ها بسیار بسیار کم است یا اینکه مفهوم واحدی از نمونهٔ منفی در دست نیست و بنابراین به‌درستی به‌عنوان نمونهٔ منفی نمونه‌برداری نشده‌اند. برای مثال، این مسئله در دسته‌بندی متون یا صفحات وب و همین‌طور کشف ناهنجاری دیده می‌شود. اصطلاح دسته‌بند تک‌کلاسی ابتدا در کار تحقیقاتی Moya و همکاران در سال ۱۹۹۳ به‌کار گرفته شد [۷]. محققان دیگر، اصطلاحات دیگری مانند تشخیص دادهٔ پرت[1]، تشخیص دادهٔ جدید[2] و یادگیری مفهوم[3] را در تحقیقات خود به‌کار برده‌اند. این اصطلاحات از کاربردهای مختلف یادگیری تک‌کلاسی نشئت می‌گیرند.

در یک دسته‌بند تک‌کلاسی کوشش می‌شود نمونه‌های متعلق به کلاس هدف تشخیص داده شود و بقیهٔ نمونه‌ها رد شود. روش OCSVM به‌عنوان یک روش یادگیری تک‌کلاسی توسط Schölkopf و همکاران [۸] معرفی شد. کلاس موردنظر، کلاس هدف نامیده می‌شود. ایدهٔ پنهان در OCSVM، توصیف کلاس هدف توسط یک تابع است که بخش عمده‌ای از آن را به ناحیه‌ای که تابع در آن صفر نباشد، نگاشت می‌کند. برای این منظور، مبدأ به‌عنوان تنها عضو کلاس غیرهدف (به‌عنوان داده پرت) در نظر گرفته می‌شود و سپس مسئله با یافتن یک فوق صفحه (تابع تصمیم) با حداکثر فاصله از مبدأ که سطح ناحیهٔ شامل داده را از ناحیهٔ بدون داده جدا می‌کند، حل می‌شود. این موضوع در شکل (۴) سمت چپ نشان داده شده است. عبارت اولیهٔ OCSVM به‌صورت زیر تعریف می‌شود:

$$\min_{w,\rho,\xi} \frac{1}{2} w^t w - \rho + \frac{1}{vl} \sum_{i=1}^{l} \xi\, i \qquad (۱)$$

با هدف:

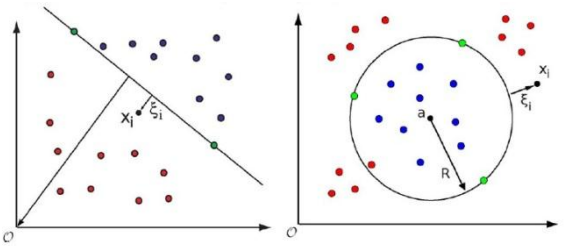
$$w^t \phi\left(xi\right) \geq \rho - \xi i$$
$$\xi i \geq 0, i = 1, \ldots, l$$

در این فرمول، پارامترهای ρ و w در راه حل این عبارت، تابع تصمیم خطی برای l دادهٔ آموزشی را تشکیل می‌دهد. پارامتر ξ ضرر حاصل از آموزش همراه با تحمل مقداری جریمه است و مقدار عبارت $\sum_{i=1}^{l} \xi$ خطای آموزش را بیان می‌کند. پارامتر جریمه یا نرخ عدم پذیرش دسته‌بند $v \in (۰,۱)$ برای کنترل مصالحه بین پیچیدگی مدل $(\frac{۱}{۲} w^t w - \rho)$ و خطای آموزش به‌کار می‌رود. تابع $\phi\left(xi\right)$، نیز تابع نگاشت است.

فرمول دیگری از دسته‌بند تک‌کلاسی SVM با نام SVDD وجود دارد که توسط Tax و Duin [۹] معرفی شد که به‌جای یافتن یک فوق صفحه که نمونه‌های کلاس هدف را با حاشیهٔ حداکثر از مبدأ جدا کند، یک فوق کره با حداقل شعاع می‌یابد که شامل تنها نمونه‌های کلاس هدف باشد و نمونه‌هایی که بیرون کره قرار می‌گیرند، به عنوان داده‌های پرت شناخته می‌شوند. این موضوع در شکل (۴) سمت راست نشان داده شده است.

1. Outlier Detection
2. Novelty Detection
3. Concept Learning

اثبات شده است که وقتی با کرنل‌های دارای خواص مشابه مانند RBF یا Gaussian و داده‌های نرمال‌شده کار می‌کنیم، روش‌های OCSVM و SVDD در بیشتر موارد، هر دو نتیجهٔ یکسانی به‌دست می‌آورند [۸]. همچنین Tax نشان داد که کرنل Gaussian برای بیشتر مجموعهٔ داده‌ها بهتر کار می‌کند. به‌علاوه به‌دلیل یک نقص تکنیکی SVDD که اغلب به مجموعهٔ داده بزرگی نیاز دارد، مخصوصاً در فضای ویژگی با ابعاد بالا بسیار ناکارآمد عمل می‌کند. از طرف دیگر، روش OCSVM می‌تواند الگوهای پیچیده و همچنین داده‌های پرت را شناسایی کند [۹].



شکل (۴): (سمت چپ) OCSVM: فوق صفحه با حداکثر حاشیه تمام داده‌های هدف (نقاط آبی) را از مبدأ با نگاشت در سمت بالای فوق صفحه و داده‌های پرت (نقاط قرمز) را در سمت پایین جدا می‌کند. (سمت راست) SVDD: فوق کره با شعاع R و مرکز a تمام نقاط هدف را محاط کرده است.

با توجه به مطالب بالا، ما از دسته‌بند تک‌کلاسی ماشین بردار پشتیبان (OCSVM) را برای شناسایی ژن‌های بیماری استفاده می‌کنیم. به‌منظور آموزش دسته‌بند تک‌کلاسی OCSVM ما از کرنل RBF استفاده کرده‌ایم که در رابطهٔ (۲) ارائه شده است:

$$K(X,X') = \exp\left(-\frac{\|X-X'\|^2}{2\sigma^2}\right) \qquad (۲)$$

در این رابطه، X بردار ویژگی نمایش‌دهندهٔ داده‌ها و σ پارامتر شباهت است که اندازه‌گیری دقیق شباهت بین بردارهای ویژگی داده به این پارامتر بستگی دارد.

۱.۳. چالش‌های دسته‌بندی تک‌کلاسی

ازآنجایی‌که داده‌های آموزشی ما شامل هیچ دادهٔ منفی نیست، مدل آموزش داده شده می‌تواند تنها نرخ مثبت صحیح را گزارش کند. تضمین گزارش صحت بالا در زمانی که مدل به یک مجموعه اعتبارسنجی مجزا اعمال می‌شود که شامل نمونه‌های مثبت و بدون برچسب است، کار بسیار سختی است. بنابراین یک چالش اصلی در مسئله دسته‌بندی تک‌کلاسی، ارزیابی دسته‌بند یادگیری شده است [۱۰].

به‌طور کلی در سامانه‌های یادگیری ماشین معیارهای متعددی برای ارزیابی و بررسی عملکرد الگوریتم‌ها و دسته‌بندها وجود دارد. یکی از روش‌هایی که برای ارزیابی عملکرد الگوریتم‌های یادگیری ماشین و دسته‌بندها مورد استفاده قرار می‌گیرد، استفاده از ماتریس تداخل است. این ماتریس یک ماتریس مربعی است که ابعاد آن تعداد دسته‌های مسئلهٔ دسته‌بندی است. در ساده‌ترین حالت، زمانی که دو دسته برای داده‌ها وجود داشته باشد، یک ماتریس ۲*۲ خواهیم داشت. برای سادگی فرض کنید که هر نمونه به یکی از دو دستهٔ مثبت و منفی تعلق داشته باشد. در این حالت، ماتریس تداخل به‌صورت جدول (۱) تعریف می‌شود. توضیح هریک از درایه‌های این ماتریس به شرح زیر است:

- TP: تعداد نمونه‌های مثبتی که به‌درستی مثبت تشخیص داده شده‌اند.
- TN: تعداد نمونه‌های منفی که به‌درستی منفی تشخیص داده شده‌اند.
- FP: تعداد نمونه‌های منفی که به اشتباه مثبت تشخیص داده شده‌اند.
- FN: تعداد نمونه های مثبتی که به اشتباه منفی تشخیص داده شده‌اند.

جدول (۱): ماتریس تداخل برای مسائل دو کلاسی

برچسب واقعی	برچسب پیش‌بینی‌شده	
	منفی	مثبت
مثبت	منفی کاذب (FN)	مثبت صحیح (TP)
منفی	منفی صحیح (TN)	مثبت کاذب (FP)

در حالت کلی، زمانی که تعداد دسته‌های مسئله دسته‌بندی n باشد، ماتریس تداخل یک ماتریس $n \times n$ به‌صورت $C = [c_{ij}]$ خواهد شد که درایه c_{ij} نشان‌دهندهٔ تعداد نمونه‌های دستهٔ iام است که جزء دستهٔ jام پیش‌بینی (دسته‌بندی) شده است. با توجه به اینکه مسائل دسته‌بندی به‌صورت ذاتی مسائل

دوکلاسی هستند، درنتیجه معیارهای دیگری از روی ماتریس تداخل قابل محاسبه است که در جدول (۲) معرفی شده است.

عدم حضور نمونه‌های منفی، کار تخمین معیارهای ارزیابی کارایی کلاسیک مثل دقت را مشکل می‌کند. در مواقعی که هیچ نمونهٔ منفی موجود نباشد، تنها TP و FN می‌تواند محاسبه شود و بنابراین طبق تعاریف p ، r از بین معیارهای کارایی، تنها محاسبهٔ معیار بازخوانی امکان‌پذیر است و محاسبهٔ دقت نیاز به FP دارد که موجود نیست.

جدول (۲): معیارهای ارزیابی

معیار	نحوهٔ محاسبه	توصیف
دقت[1]	$p = \frac{TP}{TP+FP} \times 100$	درصد پیش‌بینی‌هایی که به‌درستی مثبت شناسایی شده‌اند.
بازخوانی[2]	$r = \frac{TP}{TP+FN} \times 100$	درصد پیش‌بینی‌هایی که به‌درستی مثبت بازخوانی شده‌اند.
معیار F[3]	$F_1 - measure = (2 \times p \times r)/(p+r)$	میانگین همساز r و p

به‌عنوان جایگزین، فرمول‌های (۳)، (۴) و (۵) می‌توانند به‌عنوان معیارهای ارزیابی مدل یادگیری‌شده استفاده شوند [۱۱]. در اینجا X را به‌عنوان متغیر تصادفی نمایانگر بردار ورودی و Y را بردار برچسب‌های واقعی در نظر می‌گیریم:

$$p = precision = P[Y = 1 | f(X) = 1] \quad (۳)$$

$$r = recall = \mathrm{P}[f(X) = 1 | Y = 1] \quad (۴)$$

$$F\_measure = recall^2 / P[f(X) = 1] \quad (۵)$$

معیار بازخوانی می‌تواند به‌عنوان نسبتی از نمونه‌های مثبت موجود در داده‌های مثبت مجموعهٔ اعتبارسنجی که به‌طور صحیح پیش‌بینی شده‌اند، تخمین زده شود. مقدار احتمال $\mathrm{P}[f(X) = 1]$ هم به‌عنوان نسبتی از داده‌های مثبت پیش‌بینی‌شدهٔ موجود در کل مجموعه اعتبارسنجی تخمین زده شود [۱۰] و [۱۱]. بنابراین از دیدگاه احتمالاتی، معیار بازخوانی، احتمال نمونه‌های مثبت، $Y = 1$ را که توسط تابع دسته‌بند $f(x)$ به‌طور صحیح به‌عنوان مثبت دسته‌بندی شده‌اند، نشان می‌دهد و معیار دقت احتمال اینکه یک نمونه که توسط $f(x)$ به‌عنوان مثبت دسته‌بندی شده است، واقعاً مثبت بوده باشد، $Y = 1$ را نشان می‌دهد.

۴. ارزیابی روش پیشنهادی

۱.۴. مجموعهٔ داده

ژن‌های مورد بررسی در این مقاله، از پژوهش Yang و همکارانش [۶] استخراج شده است. داده‌های ژن‌های بیماری در این پژوهش، از آخرین نسخهٔ GENECARD [۱۲] و OMIM [۱۳] دانلود شده است. در مجموع تعداد ۲۱۳۶ ژن بیماری را که ۱۸۰ ویژگی دارند، به‌عنوان مجموعهٔ برچسب‌دار مثبت انتخاب کردیم. تعداد ۳۱۷۴ ژن ناشناخته را از ژن‌های دانلود شده از Ensembl [۱۴] انتخاب کرده و به‌عنوان مجموعهٔ دادهٔ بدون برچسب در نظر گرفتیم.

۲.۴. محیط و ابزار پیاده‌سازی

برای پیاده‌سازی OCSVM از کتابخانهٔ LIBSVM [۱۵] در نرم‌افزار MATLAB استفاده کرده‌ایم که یک ابزار نرم‌افزاری مجتمع برای دسته‌بند، رگرسیون و تخمین توزیع SVM است و می‌تواند دسته‌بند تک‌کلاسی OCSVM را هم اجرا کند.

۳.۴. پیاده‌سازی

کل مجموعه داده‌های مابعد از نرمال شدن با روش Min-Max و کاهش ویژگی‌ها با استفاده از روش PCA با مقدار POV برابر با ۹۰٪ از ۱۸۰ ویژگی به ۶۶ ویژگی کاهش یافتند. در بررسی مقادیر مختلف POV به‌ازای مقادیر ۸۵٪ تا ۹۹٪ با مقدار افزایشی ۱ واحد مشاهده شد که تقریباً به یک میزان کاهش تعداد ویژگی‌ها را داریم.

یک مشکل روش OCSVM اینست که نتایج آن به پارامترهای آزاد بسیار حساس است و تنظیم آن‌ها بسیار مشکل است [۸]. دو پارامتر مهم که در الگوریتم OCSVM نیاز به تنظیم دارد، عرض کرنل RBF، γ، و نرخ عدم پذیرش، ν، دسته‌بند OCSVM است. پارامتر عرض کرنل RBF میزان

1. Precision
2. Recall
3. F-measure

پیچیدگی مدل را مشخص می‌کند و هرچه مقدار بیشتری داشته باشد، مدل پیچیده‌تر خواهد بود. پارامتر نرخ عدم پذیرش نیز مشخص می‌کند که چقدر داده‌های هدف احتمالاً به‌عنوان داده‌های پرت (به‌عبارت دیگر ژن سالم) دسته‌بندی می‌شوند.

برای دستیابی به بهترین کارایی، OCSVM را با استفاده از کل مجموعه داده ژن‌های بیماری، آموزش دادیم و پارامترهای آن را با استفاده از ارزیابی متقابل ۱۰ تایی[1] تنظیم کردیم. در این مقاله، عرض کرنل RBF، γ، به ۴ و نرخ عدم پذیرش دسته‌بند، ν، به ۰.۱ تنظیم شده است.

همچنین ما برای یادگیری دسته‌بند به‌طور تصادفی ۷۵٪ از نمونه‌های مثبت را به‌عنوان مجموعهٔ آموزشی و بقیه را به‌عنوان مجموعهٔ اعتبارسنجی انتخاب کردیم و سپس در مرحلهٔ آزمون، ۳۱۷۴ ژن ناشناخته را به دسته‌بند یادگیری شده داده تا ژن‌های بدون برچسب را برچسب بزند. همچنین ما از روابط ارزیابی کارایی که در بخش ۲.۳ معرفی کردیم، استفاده کرده و به دقتِ ۷۹.۷٪، بازخوانیِ ۷۸.۲ و معیارِ F برابر با ۷۸.۹ رسیدیم.

۵. مقایسه با دیگران

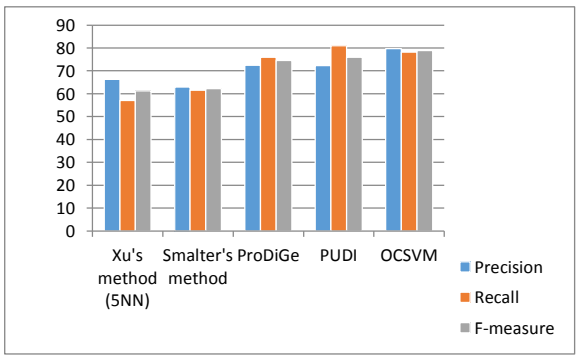
ما در این بخش روش پیشنهادی را با چهار روش تشخیص ژن‌های بیماری که عبارتند از روش PUDI، روش ProDiGe، روش Smalter و روش Xu با استفاده از معیارهای کارایی معرفی شده در روابط ۳ و۴ و ۵ مقایسه می‌کنیم. یاداوری می‌گردد تمامی روش‌های فوق روی مجموعه داده ژن‌های بیماری مستخرج از پایگاه داده OMIM مورد ارزیابی قرار گرفته‌اند.

نتایج مقایسه شده را در جدول (۳) و شکل (۵) ارائه کرده ایم. طبق این نتایج، مشاهده می‌کنیم که روش ما نسبت به روش PUDI که نسبت به سایر روش‌ها بهتر عمل کرده است، حدود ۷.۴٪ در معیار دقت بهتر و در معیار بازخوانی حدود ۲.۸٪ بدتر عمل کرده است. در کل در معیار F حدود ۲.۹٪ بهتر از روش PUDI عمل کرده و همچنین نسبت به سایر روش‌ها بهترین نتیجه را در هر سه معیار ارزیابی داده است.

این مقایسه با سایر روش‌ها به‌صورت کمّی انجام شده است اما بایستی توجه داشت که نتایج روش‌های دیگر با مقداری خطا مواجه است که فقط براساس روش‌های آزمایشگاهی می‌توان در طول زمان به مقدار خطای آن‌ها پی برد.

جدول (۳): مقایسهٔ روش پیشنهادی با چهار روش قبلی که از پایگاه داده OMIM استفاده کرده‌اند

	Precision p (%)	Recall r (%)	F_measure f (%)
OCSVM (Our Method)	**79.7**	78.2	**78.9**
PUDI [6]	72.3	**81**	76
ProDiGe [5]	72.4	75.9	74.5
Smalter's method [3]	62.9	61.5	62.2
Xu's method (1NN) [2]	65.0	55.6	59.9
Xu's method (5NN) [2]	66.3	57.1	61.3



شکل (۵): مقایسهٔ روش پیشنهادی با سایر روش‌ها در پیش‌بینی ژن‌های بیماری روی پایگاه دادهٔ OMIM

۶. نتیجه‌گیری

براساس این اصل که ژن‌های بیماری‌های مشابه احتمالاً ویژگی‌های مشابه دارند، تا کنون برخی روش‌های یادگیری ماشین برای پیش‌بینی ژن‌های بیماری جدید از ژن‌های شناخته‌شدهٔ بیماری به‌کار گرفته شده است. برخی از روش‌های پیشین یک مدل دسته‌بند دودویی با استفاده از ژن‌های شناخته‌شدهٔ بیماری، به‌عنوان مجموعه آموزشی مثبت، و ژن‌های ناشناخته، به‌عنوان مجموعه آموزشی منفی ایجاد می‌کردند. اما مجموعه منفی که

1. 10-fold cross validation

آن‌ها استفاده می‌کردند، دارای نویز است؛ زیرا ژن‌های ناشناخته می‌تواند شامل ژن‌های سالم، یعنی مجموعه مثبت هم باشد. بنابراین نتایج ارائه‌شده توسط این گونه روش‌ها قابل اعتماد نیست.

عدم حضور داده‌های منفی،یعنی ژن‌های غیر بیماری، یک موضوع چالش‌برانگیز برای پیش‌بینی ژن‌های بیماری محسوب می‌شود. انتخاب یک روش محاسباتی مناسب که بتواند با این محدودیت ذاتی مسئلهٔ ما مقابله کند، یک راه حل جایگزین با حداکثر کارایی دسته‌بند دودویی است. در این مقاله، پس از نمایش داده‌های مسئله به‌صورت مناسب، با این ایده که ما تنها به جداسازی ژن‌های بیماری علاقه‌مند هستیم و درضمن، اطلاعاتی از ژن‌های غیربیماری در دست نداریم، روش دسته‌بندی تک‌کلاسی را برای ساخت مدلی مناسب برای جداسازی ژن‌های بیماری از سایر ژن‌ها پیشنهاد دادیم. در مقایسه با چهار روش پیشین که سعی در جداسازی ژن‌های بیماری از سایر ژن‌ها داشته‌اند، نشان دادیم که استفاده از روش دسته‌بند تک‌کلاسی به‌طور مؤثرتری عمل می‌کند و به کارایی بهتری نسبت به روش‌هایی پیشین می‌رسد. ما در آینده روش‌های پیچیده‌تر یادگیری ماشین را برای حل مسئلهٔ تشخیص ژن‌های بیماری بررسی خواهیم کرد.

مراجع

[1] E. a Adie, R. R. Adams, K. L. Evans, D. J. Porteous, and B. S. Pickard, "*Speeding disease gene discovery by sequence based candidate prioritization.*," BMC Bioinformatics, vol. 6, p. 55, Jan. 2005.

[2] J. Xu and Y. Li, "*Discovering disease-genes by topological features in human protein-protein interaction network*," Bioinformatics, vol. 22, no. 22, pp. 2800–2805, 2006.

[3] A. Smalter, S. F. Lei, X. W. Chen, F. L. Seak, and X. W. Chen, "*Human Disease-Gene Classification with Integrative Sequence-Based and Topological Features of Protein-Protein Interaction Networks*," Bioinformatics and Biomedicine, (BIBM 2007). IEEE International Conference on 2007 , pp. 209–216, Nov. 2007.

[4] P. Radivojac, K. Peng, W. T. Clark, B. J. Peters, A. Mohan, S. M. Boyle, and S. D. Mooney, "*An integrated approach to inferring gene-disease associations in humans*," Proteins, vol. 72, no. 3, pp. 1030–1037, 2008.

[5] F. Mordelet, J.-P. P. Vert, "*{ProDiGe}: Prioritization Of Disease Genes with multitask machine learning from positive and unlabeled examples*," {BMC} Bioinformatics., vol. 12, no. 1, p. 389, Jan. 2011.

[6] P. Yang, X.-L. L. Li, J.-P. P. Mei, C.-K. K. Kwoh, S.-K. K. Ng, "*Positive-unlabeled learning for disease gene identification*.," Bioinformatics, vol. 28, no. 20, pp. 2640–7, Oct. 2012.

[7] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, R. C. Williamson, "*Estimating the Support of a High-Dimensional Distribution*," Neural Computation., vol. 13, no. 7, pp. 1443–1471, 2001.

[8] D. M. J. Tax, R. P. W. Duin, "*Support vector domain description*," Pattern Recognit. Letters., vol. 20, no. 11–13, pp. 1191–1199, 1999.

[9] C.-C. Chang, C.-J. Lin, "*{LIBSVM}: A library for support vector machines*," {ACM} Transactions on Intelligent Systems and Technology (TIST)., vol. 2, no. 3, pp. 1–27, 2011.

[10] Liu, B., Dai, Y., Li, X., Lee, W.S. and Yu, P.S., "*Building text classifiers using positive and unlabeled examples*." Third IEEE International Conference on Data Mining (ICDM), pp. 179-186, 2003.

[11] Lee WS, Liu B. "*Learning with positive and unlabeled examples using weighted logistic regression*.", Twentieth International Confrance of Machine Learning (ICML), vol. 3, pp. 448-455. 2003.

[12] Khan SS, Madden MG. "*A Survey of Recent Trends in One Class Classification*." Artificial Intelligence and Cognitive Science, pp. 188-197. Springer Berlin Heidelberg, 2010.

[13] Safran M, Dalah I, Alexander J, Rosen N, Stein TI, Shmoish M, Nativ N, Bahir I, Doniger T, Krug H, Sirota-Madi A. "*GeneCards Version 3: the human gene integrator*," Database: the journal of biological databases and Curation, vol. 2010, pp. 20, 2010.

[14] V. A. McKusick, "*Mendelian Inheritance in Man and its online version, {OMIM}*," The

American Journal of Human Genetics., vol. 80, no. 4, p. 588, 2007.

[15] P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, and S. Fitzgerald, "*Ensembl 2012*," Nucleic Acids Research., p. gkr991, 2011.