دوفصلنامه مدیریت مهندسی و رایانش نرم/سال چهارم، شماره۲، پاییز و زمستان ۱۳۹۷
(DOI): 10.22091/jemsc.2018.1274 صفحات ۳۹- ۲۷
JEMSC.QOM
مدیریت مهندسی و رایانش نرم ۱۳۹۴

پیش بینی سرطان سینه با استفاده از روش خوشه‌بندی انتشار وابستگی با در نظر گرفتن وزن متغیرها*

سینا دامی[۱]
زینب حاتم چوری[۲]

چکیده

با استفاده از ابزارهای داده‌کاوی در حوزه تشخیص پزشکی محدودیت‌هایی همچون هزینه‌های بالای برخی از آزمایشات یا زمان‌بر بودن آن‌ها مرتفع می‌گردد. به علاوه، وجود خطا در برخی از آزمایشات موجب شده تا روش‌های دسته‌بندی مورد استقبال پژوهشگران قرار گیرد. در همین راستا، پژوهش جاری با تکیه بر ترکیب روش‌های خوشه‌بندی و دسته‌بندی، روش جدیدی را برای تشخیص بدخیمی سرطان سینه ارائه نموده است که در آن عمل ترکیب با استفاده از یک الگوریتم ابتکاری تکرار شونده و الگوریتم خوشه‌بندی انتشار وابستگی انجام می‌شود. این روش با استفاده از یک الگوریتم ابتکاری وزن‌هایی را برای متغیرها تولید نموده و بر اساس الگوریتم انتشار وابستگی، خوشه‌های موزون تشکیل می‌دهد. سپس شماره خوشه‌ها به عنوان یک متغیر جدید به داده‌ها افزوده شده و در مرحله بعد، الگوریتم دسته‌بند روی مجموعه داده اصلاح شده حاوی داده‌های اصلی و شماره خوشه‌ها اجرا می‌گردد. با توجه به شاخص دقت، تولید اوزان تا رسیدن به بیشترین دقت ممکن ادامه می‌یابد. طبق آزمایشات عددی انجام شده در این پژوهش، ترکیب الگوریتم خوشه‌بندی انتشار وابستگی با میانگین دقت ۹۸/۳۶ دارای بیشترین دقت بوده است. به علاوه، آزمون فرض ویلکاکسون برتری شبکه عصبی ترکیبی را نسبت به سایر روش‌ها مورد تأیید قرار داده است.

کلمات کلیدی: *آزمون فرض ویلکاکسون، خوشه‌بندی، سرطان سینه، شبکه عصبی مصنوعی، ماشین بردار پشتیبان.*

---

* تاریخ دریافت: ۹۷/۸/۲۸؛ تاریخ پذیرش: ۹۷/۱۰/۲۵.

۱. استادیار گروه مهندسی کامپیوتر، دانشکده برق و کامپیوتر، دانشگاه آزاد اسلامی واحد تهران غرب، تهران، ایران (نویسنده مسئول)
dami@wtiau.ac.ir
۲. کارشناس ارشد مهندسی کامپیوتر، دانشکده برق و کامپیوتر، دانشگاه آزاد اسلامی واحد تهران غرب، تهران، ایران
zeinab.hatam.ch@gmail.com

مقدمه

بطور کلی سرطان یا چَنگار، تقسیم نامتقارن و از کنترل خارج شده سلول‌های بدن است که منشأ آن نامعلوم بوده و از نظر محققان ممکن است به دلایل ژنتیکی و اختلال در هسته سلول باشد. سرطان سینه نوعی از این بیماری است که از بافت سینه شروع می‌شود و علائم آن ممکن است به‌صورت یک توده در سینه، تغییر شکل در سینه، گودی پوست، ترشح مایع یا پوسته شدن باشد. در بیمارانی که سلول‌های سرطانی در بدن آن‌ها چنگ‌اندازی کرده باشد علائمی از قبیل درد استخوان، غدد لنفاوی متورم، تنگی نفس یا زردی وجود دارد. تشخیص سرطان سینه با انجام نمونه‌برداری از توده مربوطه تأیید می‌شود. پس از تشخیص سرطان، آزمایش‌های بیشتری انجام می‌شوند تا مشخص شود که آیا سرطان به قسمت‌های دیگر بدن نیز سرایت کرده است یا خیر؟ و چه درمان‌هایی ممکن است نسبت به بیماری واکنش نشان دهند.

طبق گزارشات صندوق جهانی پژوهش سرطان سینه، شایع‌ترین نوع سرطان در میان زنان با تعدادی بالغ بر ۱/۷ میلیون تشخیص در سال ۲۰۱۲ می‌باشد. بطور متوسط ٪۳۷/۳ از مبتلایان این بیماری بطور کامل سلامت خود را بازیافته‌اند که بیشتر این موفقیت مرهون تشخیص زودهنگام و نیز تشخیص صحیح شدت بدخیمی توده سرطانی می‌باشد. از همین رو، تشخیص شدت بدخیمی یک توده سرطانی و در این میان سرطان سینه که مبحث اصلی این پژوهش خواهد بود بسیار حائز اهمیت است. بنابراین، توسعه مدل‌هایی با دقت قابل قبول و عملکرد مناسب همچنان موضوع اصلی بسیاری از پژوهش‌ها می‌باشد.

پیشینه پژوهش

در ادبیات موضوعی مرتبط با این حوزه تحقیقات بسیار زیادی انجام شده است. این پژوهش‌ها برخی بر روی کشف عوامل مؤثر در بیماری‌ها متمرکز بوده و برخی دیگر نیز مدل‌های پیش‌بینی را برای آن‌ها به کار گرفته‌اند. بطور کلی، داده‌کاوی به دو صورت در پزشکی مورد استفاده قرار می‌گیرد :

- رویکردهای توصیفی، که در آن‌ها با استفاده از روش‌های توصیفی در داده‌کاوی (خوشه‌بندی، قواعد انجمنی و...) اقدام به مشخص نمودن تأثیر عوامل بر روی یکدیگر نموده و متغیرهای اثرگذار در بیماری‌ها را تعیین می‌کنند. این رویکرد معمولاً در پیشگیری از بیماری‌ها مؤثر است.
- رویکردهای پیش‌گویانه، که در آن‌ها با استفاده از تکنیک‌های پیش‌بینی و دسته‌بندی اقدام به ساخت مدل‌هایی جهت تعیین نوع بیماری یا درجه بدخیمی نموده و علاوه بر آن در برخی موارد این روش‌ها را جایگزین برخی آزمایشات زمان‌بر/ هزینه‌بر می‌نمایند.

تشخیص صحیح و به موقع، از مهم‌ترین عوامل موفقیت در درمان سرطان سینه است. به عنوان مثال، یکی از روش‌های تشخیصی سرطان سینه آزمون ایمنوهستوشیمی بوده که در برخی موارد از تشخیص صحیح بیماری عاجز است. این حقیقت باعث شده است تا روش‌هایی همچون داده‌کاوی در حوزه‌های تشخیصی مذکور توسعه یابد. در حوزه استفاده از ماشین بردار پشتیبان برای تشخیص سرطان سینه، برخی از پژوهشگران با استفاده از ماشین بردار پشتیبان تصاویر مربوط به سلول‌های سرطانی را پردازش نموده و بر این اساس یک مدل دسته‌بندی برای تعیین سلول‌های سرطانی ارائه نمودند. وانگ و همکاران (۲۰۱۶) در پژوهشی دیگر، با استفاده از ماشین بردار پشتیبان داده‌های مربوط به ماموگرافی را برای تعیین توده بدخیم استفاده نموده و به دقت ٪۹۲/۹۹ دست یافتند. سامپیا و همکاران (۲۰۱۶) با ترکیب الگوریتم نزدیکترین همسایگان و ماشین بردار پشتیبان سرطان سینه را پیش‌بینی نمودند. در رویکرد مذکور، از یک تکنیک حساس به هزینه استفاده شده است که در آن نرخ خطای تشخیص سالم بودن فرد در صورتی که واقعاً مبتلا باشد مقدار بیشتری در نظر گرفته شده است. روش پیشنهادی آن‌ها نسبت به روش‌های سنتی دقت بالاتری را نشان داده است (روحی و جعفری، ۲۰۱۶). در بررسی‌های دیگر نیز یک روش مجموعه فعال برای افزایش سرعت ماشین بردار پشتیبان ارائه گردید و سپس این روش را

بر روی داده‌های مربوط به سرطان سینه پیاده‌سازی نمودند. نتایج مربوطه نشان از برتری روش پیشنهادی و افزایش سرعت در آموزش آن داشته است (سولام و همکاران ۲۰۱۰).

در سال ۲۰۱۶ پژوهشی در زمینه تشخیص توده از روی آزمایشات ماموگرافی انجام گرفته است که در نهایت به دقت ٪۸۱ دست یافته است (آرولا و همکاران، ۲۰۱۶). در همان سال پژوهشی با موضوع دسته‌بندی و تشخیص عوامل مؤثر به منظور انتخاب مشخصه انجام داده است و در این پژوهش ۹ عامل از بین ۱۷ متغیر موجود به عنوان عوامل مؤثر انتخاب شده و دسته‌بندی با استفاده از آن‌ها تا دقت ٪۹۶/۷ گزارش شده است (جیا و وانگ، ۲۰۱۶). در پژوهشی دیگر با استفاده از شبکه عصبی پیش‌خور در تحلیل عوامل ژنتیکی مؤثر در سرطان سینه به دقت ٪۹۴/۲ و ایجاد ٪۱۱/۳ بهبود نسبت به روش‌های پیشین دست یافته است (نوشاد و همکاران، ۲۰۱۶).

برخی از محققین با استفاده از ماشین بردار پشتیبان و الگوریتم K-means یک روش خوشه‌بندی- دسته‌بندی ارائه نموده‌اند. در این تحقیق داده‌ها در ابتدا خوشه‌بندی شده و سپس برای هر خوشه یک دسته‌بند مورد استفاده قرار گرفته است. روش مذکور با افزایش دقت و سرعت همگرایی همراه بوده است (ژانگ و همکاران، ۲۰۱۴).

عده‌ای از پژوهشگران با استفاده از ماشین بردار پشتیبان و الگوریتم کلونی مورچگان توانستند به بهبود دقت نسبت به حالت سنتی دست یابند. این پژوهشگران علت استفاده از الگوریتم کلونی را تنظیم پارامترهای ماشین بردار پشتیبان از قبیل پارامتر جریمه و تابع کرنل اعلام نموده‌اند (میشرا و همکاران، ۲۰۱۵).

برخی از پژوهشگران به‌صورت هم‌زمان از الگوریتم درخت تصمیم و ماشین بردار پشتیبان در پژوهش خود استفاده نموده‌اند (سیواکامی، ۲۰۱۵).

برخی از پژوهشگران با استفاده از نقشه‌های خودسازمانده برای تعیین نقاط مشکوک و انجام عمل دسته‌بندی از طریق دادن مراکز خوشه به ماشین بردار پشتیبان به دقت بالاتری دست یافته‌اند (قیومی، ۱۳۹۲).

در همین راستا، این تحقیق با موزون در نظر گرفتن داده‌ها و همچنین استفاده از ترکیب روش خوشه‌بندی انتشار وابستگی و الگوریتم‌های دسته‌بندی موجود در ادبیات موضوع قصد دارد روش جدیدی را برای پیش‌بینی سرطان سینه مورد استفاده قرار دهد. همین مسئله، ایده اصلی این تحقیق خواهد بود. در این رابطه، برخی از متغیرهای موجود در مجموعه داده که در مدلسازی به کار گرفته می‌شوند عبارتند از: ضخامت توده، یکنواختی اندازه سلول‌های سرطانی، یکنواختی شکل سلول‌های سرطانی، چسبندگی حاشیه‌ای، عادی بودن هستک، میتوز و... خواهد بود. رویکرد مورد بحث در بهبود شاخص‌های عملکرد دسته‌بندی دو روش مذکور را ترکیب خواهد نمود.

ترکیب الگوریتم خوشه‌بندی انتشار وابستگی و روش‌های دسته‌بندی، انعطاف‌پذیری مدل ساخته شده را بالا برده و می‌تواند منجر به افزایش دقت شود. به علاوه، به نظر می‌رسد از بین عواملی همچون سن، نوع رژیم غذایی، عوامل ارثی و... یک یا چند متغیر بیشترین تأثیر را روی بدخیمی توده خواهند داشت. این پژوهش ترکیبی از تحقیق توصیفی و تجربی خواهد بود. به طوری که در بخش‌های اولیه تحقیق یک مطالعه پیمایشی روی تحقیقات گذشته صورت گرفته و پس از آن با استفاده از تحلیل نتایج حاصل از حل مدل، در مورد عملکرد آن نتیجه‌گیری خواهد شد. در این راستا، در ابتدا جمع‌آوری و پیش‌پردازش داده‌ها صورت گرفته و سپس مدل ترکیبی پیشنهادی بر روی آن پیاده‌سازی می‌شود و در نهایت نیز دقت مدل پیشگو با استفاده از شاخص‌های مناسب اندازه‌گیری خواهد شد.

روش‌شناسی پژوهش

روش پیشنهادی، با استفاده از یک الگوریتم ابتکاری، وزن‌هایی را برای متغیرها در نظر گرفته و عمل دسته‌بندی را به صورت موزون انجام می‌دهد. وزن‌دهی به داده‌ها می‌تواند ضمن بهبود دقت نهایی مدل، در تشخیص عوامل مؤثر در ابتلا به سرطان سینه نیز مؤثر باشد. در این راستا، گام‌های مربوط به متدولوژی تحقیق که در زیربخش‌های بعدی به صورت دقیق‌تر مورد بحث قرار می‌گیرد مطابق با مراحل زیر خواهد بود:

۱. جمع‌آوری داده‌ها
۲. پیش‌پردازش داده‌ها
۳. پیش‌بینی مقادیر مفقوده
۴. تشخیص داده‌های پرت
۵. الگوریتم ترکیبی پیشنهادی
۶. شاخص‌های عملکرد

جمع‌آوری داده‌ها

داده‌های مورد استفاده در این تحقیق متعلق به پایگاه داده UCI Repository در دانشگاه کالیفرنیا به آدرس https://archive.ics.uci.edu/ml/datasets می‌باشد که مربوط به بیماری سرطان سینه بوده و شامل ۶۹۹ مشاهده است که هر یک دارای ۱۰ مشخصه (صفت) می‌باشند. خروجی مربوط به این مجموعه داده به صورت صفر و یک است که مقدار صفر به معنای عدم وجود بیماری یا خوش‌خیم بودن توده مشاهده شده در عکسبرداری‌های پزشکی است و مقدار ۱ نیز به معنای بدخیم بودن توده می‌باشد. این مجموعه توسط گروه آنکولوژی دانشگاه علوم پزشکی الجوبلیانا در کشور یوگوسلاوی جمع‌آوری شده و در دسترس پژوهشگران قرار گرفته است.

پیش‌پردازش داده‌ها

پیش‌پردازش داده‌ها در دو بخش انجام خواهد شد. بخش اول شامل تکمیل مقادیر مفقوده و بخش دوم نیز دربرگیرنده تشخیص مقادیر پرت و حذف آن‌ها است که می‌تواند روی نتایج نهایی اثرگذار باشد.

پیش‌بینی مقادیر مفقوده

بطورکلی روش‌های پرکردن مقادیر مفقوده عبارتند از :

- حذف داده

- درنظرگرفتن یک مقدار سراسری برای همه مقادیر مفقوده
- استفاده از مقدار میانگین متغیر برای همه مقادیر مفقوده
- استفاده از مقدار میانه داده‌ها برای مقادیر مفقوده
- تعیین مقادیر مفقوده با استفاده از روش‌های یادگیری: در این روش‌ها متغیری که مقدار آن مفقود شده است به عنوان متغیر وابسته انتخاب شده و سایر متغیرها نیز نقش متغیر مستقل را خواهند داشت. سپس با استفاده از یک روش یادگیری مثل رگرسیون خطی (که خطر بیش برازش آن را تهدید نمی‌کند) مقدار مفقوده پیش‌بینی می‌شود.

روش پنجم از چهار روش دیگردقیق‌تراست. زیرا داده‌های پرت، نامتقارن بودن توزیع داده‌ها و... که به شدت بر روی روش‌های ۱ تا ۳ اثرگذار است، در روش پنجم تأثیر کمتری را نشان می‌دهد. از سوی دیگر، استفاده از روش پنجم خطر بیش‌برازش را نیز خواهد داشت. زیرا استفاده از یک روش پیچیده (مثلاً شبکه عصبی با تعداد زیادی لایه پنهان) ممکن است روی داده‌های آموزش کاملاً منطبق شود ولی در فاز تست مدل دقت و صحت بسیار پایینی داشته باشد. معمولاً برای غلبه بر خطر بیش‌برازش از مدل‌های پیچیده غیرخطی استفاده نمی‌شود. به همین دلیل به منظور تخمین مقادیر مفقوده از رگرسیون خطی چندگانه استفاده خواهد شد.

تشخیص داده‌های پرت

داده‌های پرت معمولاً به داده‌هایی اطلاق می‌گردد که در یک مجموعه نسبت به سایر اعضا بسیار کوچک‌تر یا بزرگ‌تر باشند. وجود این داده‌ها می‌تواند نتایج را تحت تأثیر قرار داده و در واقع صحت نتایج حاصل را خدشه‌دار کند. بنابراین لازم است چنین مواردی مورد تحلیل قرار گرفته و در صورت وجود از میان داده‌ها حذف شوند.

برخی از روش‌های موجود در حوزه تشخیص داده‌های پرت به صورت زیر است:

- آزمون گرابز
- روش بونفرونی

- روش خوشه‌بندی
- شبکه عصبی
- ماشین بردار پشتیبان تک کلاسه. بطور کلی این روش برای حل مسئله‌ای است که در آن یک دسته به عنوان دسته هدف موجود بوده و بقیه دسته‌ها به عنوان داده‌های پرت شناخته می‌شوند. مقصود اصلی این روش عبارت است از یافتن یک مرز در اطراف داده‌های مربوط به دسته هدف که با استفاده از آن داده‌های پرت شناسایی شوند.

الگوریتم ترکیبی پیشنهادی

مدل ترکیبی خوشه‌بندی-دسته‌بندی پیشنهادی شامل استفاده از الگوریتم خوشه‌بندی انتشار وابستگی، الگوریتم ابتکاری وزن‌دهی و ترکیب آن‌ها با روش‌های دسته‌بندی همچون ماشین بردار پشتیبان و شبکه عصبی می‌باشد. به همین منظور، گام‌های زیر پس از پیش‌پردازش داده‌ها برای ساخت مدل طی خواهد شد:

۱. الگوریتم ابتکاری پیشنهادی جمعیتی از اوزان را تولید می‌کند.
۲. مجموعه داده مربوط به بیماری سرطان سینه با ضرب اوزان تولید شده، موزون گردیده و برای انجام عمل خوشه‌بندی در حافظه قرار می‌گیرد. الگوریتم خوشه‌بندی انتشار وابستگی بر روی مجموعه داده موزون اجرا شده و خوشه‌های به دست آمده برای هر یک از داده‌ها ذخیره‌سازی می‌شود. الگوریتم خوشه‌بندی انتشار وابستگی، نوعی از روش‌های یادگیری بدون نظارت است که بر پایه عبور پیام میان داده‌ها توسعه یافته است. برخلاف سایر روش‌های خوشه‌بندی موجود در ادبیات موضوع، الگوریتم انتشار وابستگی نیاز به تعیین تعداد خوشه‌ها ندارد و در تکرارهای متوالی تعداد خوشه‌ها را به صورت خودکار تعیین می‌کند. به عبارت بهتر، الگوریتم مذکور مشابه الگوریتم k-medoids داده‌های نمونه را می‌یابد که در واقع نماینده خوشه‌ها هستند.

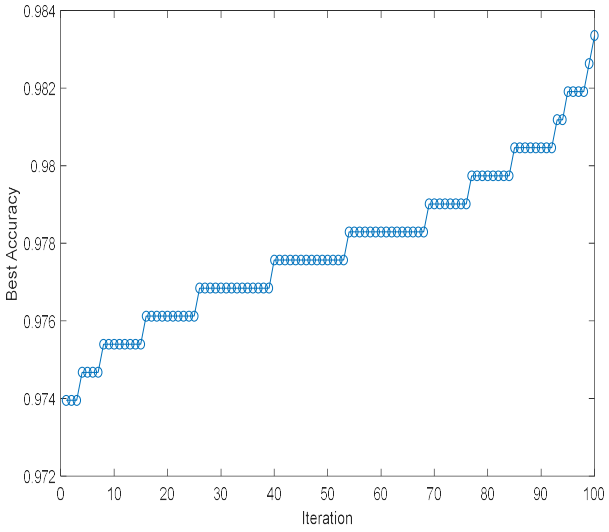
۳. خوشه‌های حاصل شده به عنوان یک متغیر جدید به مجموعه داده افزوده می‌شود. پس از انجام این کار، تعداد مشخصه‌های موجود برابر با ۱۱ خواهد بود. با انجام این عمل، یک مجموعه داده اصلاح شده ایجاد گردیده و در حافظه ذخیره‌سازی خواهد شد.
۴. الگوریتم‌های ماشین بردار پشتیبان و شبکه عصبی به صورت جداگانه روی مجموعه داده اصلاح شده اجرا می‌شود. در این مرحله با استفاده از روش اعتبارسنجی متقاطع ۱۰-لایه مجموعه‌های آموزش و تست انتخاب شده‌اند که ۹۰٪ داده‌ها را برای آموزش و ۱۰٪ را برای تست به کار می‌گیرند.
۵. شاخص دقت برای سنجش عملکرد مدل محاسبه شده و سپس در حافظه ذخیره می‌شود. در صورتی‌که شاخص دقت نسبت به تکرارهای پیشین برتر باشد، وزن تولید شده جایگزین وزن پیشین می‌شود.
۶. در صورتی که شرط توقف برقرار باشد، الگوریتم متوقف می‌شود. در غیر این صورت، به گام (۱) باز می‌گردیم.

در واقع، مدل ترکیبی پیشنهادی در تکرارهای متوالی سعی می‌کند بهترین اوزان را که منجر به بهبود دقت عملکرد می‌شود تولید نموده و از این طریق دقت نهایی مدل دسته‌بند را بهبود دهد.

در این رابطه، اندازه جمعیت در الگوریتم ابتکاری وزن‌دهی برابر با ۱۰ در نظر گرفته شده و پس از ۱۰۰ تکرار، بهترین اوزان که منجر به ایجاد بهترین دقت شده‌اند ذخیره‌سازی گردیده و مقدار دقت نهایی نیز ثبت شده است. شایان ذکر است که در هر تکرار از الگوریتم، دقت ماشین بردار پشتیبان و شبکه عصبی محاسبه شده و میانگین آن به عنوان دقت نهایی الگوریتم مدنظر قرار گرفته است که در واقع، میانگین دقت ماشین بردار پشتیبان و شبکه عصبی توسط الگوریتم وزن‌دهی حداکثرسازی می‌شود. بنابراین در هر اجرای الگوریتم، تعداد ۱۰۰۰ وزن تولید شده و عمل خوشه‌بندی بر روی داده‌های موزون بر اساس آن‌ها انجام می‌شود. در مرحله بعد، داده‌های اصلی به همراه شماره خوشه‌ها وارد

مدل دسته‌بندی شده و میانگین دقت محاسبه می‌گردد. الگوریتم در هر تکرار به گونه‌ای سعی در بهبود اوزان تولید شده می‌نماید که میانگین دقت در هر تکرار بهبود یابد.

در هر تکرار از الگوریتم، بهترین دقت به دست آمده از ۱۰ وزن تولید شده ذخیره‌سازی شده و این عمل برای ۱۰۰ تکرار انجام می‌شود. در نهایت، پس از پایان الگوریتم، بهترین دقت به دست آمده به عنوان دقت نهایی گزارش می‌شود. بهبود دقت مدل بر حسب تکرارهای متوالی الگوریتم در شکل (۱) نشان داده شده است.



شکل ۱. میانگین دقت مدل پیشنهادی در تکرارهای متوالی الگوریتم

یافته‌های پژوهش

پس از اجرای مدل به تعداد ۱۰۰ مرتبه، نتایج حاصل بر پایه شاخص دقت به صورت جدول (۱) خلاصه‌سازی شده است.

شاخص‌های عملکرد

برای ارزیابی عملکرد مدل پیشنهادی و همچنین مقایسه با روش‌های سنتی از شاخص‌های متفاوتی می‌توان استفاده نمود. برای تشریح بیشتر نمادگذاری زیر را در نظر بگیرید:

T: تعداد کل داده‌ها.

TP: تعداد افراد بیماری که توسط سیستم به درستی بیمار معرفی شده‌اند.

FP: تعداد افراد سالمی که توسط سیستم به اشتباه بیمار تشخیص داده شده‌اند.

TN: تعداد افراد سالمی که توسط سیستم به درستی سالم معرفی شده‌اند.

FN: تعداد افراد بیماری که توسط سیستم به اشتباه سالم تشخیص داده شده‌اند.

در این صورت شاخص دقت برای ارزیابی دسته‌بندی به صورت زیر تعریف می‌شوند:

دقت: نشان‌دهنده تعداد کل داده‌هایی به درستی دسته‌بندی شده‌اند به کل داده‌های موجود.

رابطه ۱) $$Accuracy = \frac{TP+TN}{T}$$

جدول ۱. دقت حاصل از حل مدل پیشنهادی و مدل مرسوم (٪)

ردیف	مدل پیشنهادی		مدل مرسوم (بدون خوشه‌بندی)	
	ماشین بردار پشتیبان	شبکه عصبی	ماشین بردار پشتیبان	شبکه عصبی
۱	۹۸/۱۱	۹۸/۵۵	۸۸/۴۲	۹۶/۲۳
۲	۹۷/۹۹	۹۸/۴۹	۸۸/۴۲	۹۶/۰۹
۳	۹۸/۰۹	۹۸/۱۳	۸۷/۹۹	۹۴/۴۹
۴	۹۸/۰۹	۹۸/۴۹	۸۸/۷۱	۹۶/۰۹
۵	۹۸/۱۱	۹۸/۴۹	۸۸/۲۷	۹۵/۹۵
۶	۹۸/۲۷	۹۷/۹۵	۸۸/۴۲	۹۶/۶۷
۷	۹۸/۸۹	۹۸/۴۴	۸۸/۵۶	۹۶/۰۹
۸	۹۸/۱۱	۹۸/۱۱	۸۸/۲۷	۹۵/۵۰
۹	۹۸/۲۲	۹۸/۴۹	۸۸/۴۲	۹۶/۵۲
۱۰	۹۸/۱۱	۹۸/۴۹	۸۷/۹۹	۹۵/۲۲
میانگین	۹۸/۲۰	۹۸/۳۶	۸۸/۳۵	۹۵/۸۹

نتیجه‌گیری و پیشنهادها

به منظور مقایسه دو روش مدل پیشنهادی و مرسوم، از آزمون رتبه علامت ویلکاکسون استفاده می‌شود. در نهایت، برای رتبه‌بندی روش‌های مورد استفاده از شبکه غلبه[1] استفاده شده است که در آن، برای هر زوج روش یک یال جهت‌دار به سمت روش برتر ترسیم می‌شود. در نهایت، روشی که بیشترین تعداد یال‌های ورودی را داشته باشد به عنوان روش برتر شناسایی می‌شود.

طبق آزمایشات عددی انجام شده در این پژوهش، ترکیب الگوریتم خوشه‌بندی انتشار وابستگی با میانگین دقت ۹۸/۳۶ دارای بیشترین دقت بوده است. به علاوه، آزمون فرض ویلکاکسون برتری شبکه عصبی ترکیبی را نسبت به سایر روش‌ها مورد تأیید قرار داده است.

در راستای مدل توسعه یافته در این پژوهش، یک مورد به عنوان پیشنهاد آتی مطرح می‌گردد:

یکی از چالش‌های موجود در حوزه دسته‌بندی، ساخت روش‌هایی است که بتوانند ساختارهای غیرخطی را تفکیک نموده و از این طریق عمل دسته‌بندی را با دقت بالاتری انجام دهند. معمولاً برای افزایش انعطاف دسته‌بندها و بهبود دقت از نگاشت‌های غیرخطی استفاده می‌شود. در همین مورد استفاده از محور مختصات قطبی برای بهبود عملکرد مدل‌های دسته‌بند پیشنهاد می‌گردد. به این صورت که داده‌ها به محور مختصات قطبی نگاشت شوند و سپس عمل دسته‌بندی بر روی آن‌ها صورت پذیرد.

[1] Dominance Network

منابع

- Arevalo J, González FA, Ramos-Pollán R, Oliveira JL, Lopez MA. (2016). Representation learning for mammography mass lesion classification with convolutional neural networks. Computer methods and programs in biomedicine. 127, 248-57.
- De Sampaio WB, Silva AC, de Paiva AC, Gattass M. (2015). Detection of masses in mammograms with adaption to breast density using genetic algorithm, phylogenetic trees, LBP and SVM. Expert Systems with Applications. 42(22), 8911-28.
- Ghayomi Zade. A. (2013). Clustering and Diagnosis of Breast Cancer via Thermal Images Using a Combination of SVM and SOM Neural Network. ijbd. 2013; 5 (4), 13-22
- Hassanien AE, Moftah HM, Azar AT, Shoman M. (2014). MRI breast cancer diagnosis hybrid approach using adaptive ant-based segmentation and multilayer perceptron neural networks classifier. Applied Soft Computing. 14, 62-71.
- He, X., Wang, Z., Jin, C., Zheng, Y., Xue, X. (2012). A simplified multi-class support vector machine with reduced dual optimization, Pattern Recognition Letters, 33, 71-82.
- Jiao Z, Gao X, Wang Y, Li J. (2016). A deep feature based framework for breast masses classification. Neurocomputing. 197, 221-31
- Mishra G, Ananth V, Shelke K, Sehgal D, Valadi J. (2015). Hybrid ACO Chaos-Assisted Support Vector Machines for Classification of Medical Datasets. InProceedings of Fourth International Conference on Soft Computing for Problem Solving 2015. Springer India. 91-101
- Naush J, González FA, Ramos-Pollán R, Oliveira JL, Lopez MA. (2016). Representation learning for mammography mass lesion classification with convolutional neural networks. Computer methods and programs in biomedicine. 127, 248-57.
- Naushad SM, Ramaiah MJ, Pavithrakumari M, Jayapriya J, Hussain T, Alrokayan SA, Gottumukkala SR, Digumarti R, Kutala VK. (2016). Artificial neural network-based exploration of gene-nutrient interactions in folate and xenobiotic metabolic pathways that modulate susceptibility to breast cancer. Gene. 580(2), 159-68.
- Rouhi R, Jafari M. (2016). Classification of benign and malignant breast tumors based on hybrid level set segmentation. Expert Systems with Applications. 46, 45-59.
- Sivakami K. (2015). Mining Big Data: Breast Cancer Prediction using DT-SVM Hybrid Model.
- Sweilam NH, Tharwat AA, Moniem NA. (2010). Support vector machine for diagnosis cancer disease: A comparative study. Egyptian Informatics Journal. 11(2), 81-92.
- Wang P, Hu X, Li Y, Liu Q, Zhu X. (2016). Automatic cell nuclei segmentation and classification of breast cancer histopathology images. Signal Processing. 122, 1-3.
- World Health Organization. (2014) "Cancer Fact sheet N°297".
- Zheng B, Yoon SW, Lam SS. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications. 41(4), 1476-82.
- Zheng-Feng LI, Guang-Jin XU, Jia-Jun WA, Guo-Rong DU, Wen-Sheng CA, Xue-Guang SH. (2016). Outlier Detection for Multivariate Calibration in Near Infrared Spectroscopic Analysis by Model Diagnostics. Chinese Journal of Analytical Chemistry. 44(2), 305-9.

استناد به این مقاله: شناسه دیجیتال (DOI): 10.22091/jemsc.2018.1274

دامی، سینا، حاتم چوری، زینب. (۱۳۹۷). «پیش بینی سرطان سینه با استفاده از روش خوشه‌بندی انتشار وابستگی با در نظر گرفتن وزن متغیرها». *مدیریت مهندسی و رایانش نرم*، ۴ (۲)، ۳۹-۲۷