Journal of Information Technology Management
ISSN: 2008-5893
EISSN: 2423-5059
Spring 2018
Vol. 10; No. 1; PP. 69- 92
DOI: 10.22059/jitm.2017.233609.2043

University of Tehran
Faculty of Management

Applying High-level Agreement Ensemble Classification Voting Techniques to Distinguish Inflammatory Bowel Disease

Nayere Zaghari [1], Rahil Hosseini [2]

Abstract: Due to the complexity of medical decisions, there is a growing interest in the application of intelligence systems to support these decisions. In this paper, accordingly, the potential of several algorithms such as K Nearest Neighbor, Support Vector Machine, Random Forest, Naive Bayes, and Decision Tree was used to create an ensemble classification. Then, to obtain the voting result, high level agreement voting was used to evaluate the performance and make prediction. According to the involvement of body organs with this disease, the problem of diagnosing and differentiating various types of bowl inflammation was investigated. We should mention that higher prediction accuracy was obtained using the proposed model. The results and the comparisons of these methods showed that the proposed model indicates the highest prediction accuracy which is 98%. In the final step, the proposed model was evaluated applying the receiver operating characteristic curve model (ROC), and the area under the curve (AUC) was calculated.

Key words: ***Ensemble classification algorithms, High-level agreement voting algorithm, Inflammatory bowel disease, Noise detection, ROC curve.***

1. Ph.D. Candidate of Computer Engineering, Shahr-e-Qods Branch, Islamic Azad University, Tehran, Iran
2. Assistant Prof. of Computer engineering, Shahr-Qods Branch, Islamic Azad University, Tehran, Iran

Submitted: 19 / May / 2017
Accepted: 28 / September / 2017
Corresponding Author: Rahil Hosseini
Email: universityhosseini@gmail.com

مدیریت فناوری اطلاعات
دانشکدهٔ مدیریت دانشگاه تهران
دورهٔ ۱۰، شمارهٔ ۱
بهار ۱۳۹۷
صص. ۹۲- ۶۹

Journal of Information Technology Management
DOI: 10.22059/jitm.2017.233609.2043

به‌کارگیری تکنیک‌های دسته‌بندی ترکیبی رأی‌گیری با توافق سطح بالا برای تفکیک بیماری التهاب روده

نیره زاغری[۱]، راحیل حسینی[۲]

چکیده: به‌دلیل پیچیدگی تصمیمات پزشکی، کاربرد سیستم‌های هوشمند برای پشتیبانی از این تصمیمات افزایش یافته است. در این پژوهش از قابلیت پنج الگوریتم مختلف ماشین بردار پشتیبان، درخت تصمیم‌گیری، نایو بیزین، نزدیک‌ترین همسایه و جنگل تصادفی، یک دسته‌بند ترکیبی ساخته شده که برای به‌دست آوردن نتیجهٔ آرای این دسته‌بند، از رویکرد رأی‌گیری با توافق سطح بالا به‌منظور ارزیابی و پیش‌بینی استفاده می‌شود. با توجه به اهمیت درگیری اعضای بدن نسبت به بیماری التهاب روده، در ارزیابی و مقایسهٔ نتایج به‌دست آمده، دسته‌بند ترکیبی پیشنهاد شده برای تفکیک دو نوع بیماری التهاب روده (کرون و کولیت) به درصد صحت بالاتری دست یافته است. نتیجهٔ پژوهش و مقایسهٔ روش‌ها با توجه به آزمایش‌های انجام شده نشان داد، بالاترین صحت پیش‌بینی (۹۸ درصد) در دسته‌بندی ترکیبی پیشنهادشده (رأی‌گیری با توافق سطح بالا) به‌دست آمده است. در گام آخر، مدل ساخته شده با استفاده از نمودار مشخصه‌های عامل گیرندهٔ ROC ارزیابی شد و مساحت زیر نمودار AUC به‌دست آمد.

واژه‌های کلیدی: الگوریتم رأی‌گیری با توافق سطح بالا، الگوریتم‌های ترکیبی دسته‌بندی، بیماری التهاب روده، تشخیص نویز، نمودار مشخصه‌های عامل گیرنده.

۱. دانشجوی دکتری مهندسی کامپیوتر، واحد شهر قدس، دانشگاه آزاد اسلامی، تهران، ایران
۲. استادیار گروه مهندسی کامپیوتر، واحد شهرقدس، دانشگاه آزاد اسلامی، تهران، ایران

تاریخ دریافت مقاله: ۱۳۹۶/۰۲/۲۹
تاریخ پذیرش نهایی مقاله: ۱۳۹۶/۰۷/۰۶
نویسندهٔ مسئول مقاله: راحیل حسینی
E-mail: universityhosseini@gmail.com

مقدمه

در دنیای امروز، هزینهٔ تشخیص و درمان بیماری روز به روز افزایش می‌یابد. تشخیص بیماری التهاب روده که دو نوع کولیت و کرون دارد، بسیار پیچیده است؛ به‌گونه‌ای که سال‌ها تحت عنوان بیماری کرون داروهای این بیماری را تجویز می‌کنند، اما بعد از مدتی پی به تشخیص اشتباه خود می‌برند و داروهای بیمار را تغییر می‌دهند. این مسئله نه تنها موجب هزینهٔ زیاد خرید دارو می‌شود، بلکه بهبودی بیماری را نیز به تأخیر می‌اندازد. به‌کمک روش تشخیص نویز کلاس می‌توان نویزهای موجود درون دیتاست بیمار را تشخیص داده و آن را حذف کرد. نویز کلاس یا نویز برچسب در صورتی رخ می‌دهد که تشخیص پزشک از بیماری یک نمونه با نتیجهٔ رویکردهای مختلف رأی‌گیری متفاوت باشد. استفاده از تشخیص نویز برچسب در مقالات مختلفی برای تشخیص بیماری استفاده شده است (احمد و همکاران، ۲۰۱۷). دو بیماری کرون و کولیت اولسروز از بیماری‌های التهابی رودهٔ بزرگ هستند. کرون و کولیت اولسروز معمولاً با عفونت‌های باکتریایی یا ویروسی و گاهی بر اثر مصرف داروی خاصی شروع یا تشدید می‌شوند و رودهٔ بزرگ را تحریک کرده و موجب التهاب آن می‌شوند. در مبتلایان به بیماری‌های کرون و کولیت اولسروز، واکنش التهابی در مقابل عوامل تحریک‌کنندهٔ بیماری شدید است و درد و اسهال مزمن ایجاد می‌کند. علاوه بر این، نقطهٔ پایانی هم وجود ندارد؛ بنابراین حتی پس از حذف عوامل محرک از صحنه، التهاب همچنان ادامه دارد و در نتیجه جدار روده به‌طور دائم آسیب دیده و روند جذب و ترشح مایعات توسط روده را بر هم می‌زند. بیماری کرون هم رودهٔ بزرگ و هم رودهٔ کوچک را درگیر می‌کند، در حالی که کولیت اولسروز فقط رودهٔ بزرگ را درگیر می‌سازد. هر دو بیماری می‌توانند رودهٔ بزرگ را ملتهب کرده و در نتیجه موجب بروز اسهال مزمن شوند. اغلب این دو اختلال به‌وسیلهٔ آزمایش‌های پزشکی از یکدیگر تشخیص داده می‌شوند. از این رو، باید برای تشخیص آنها بررسی‌های کامل و دقیقی انجام شود. بنابراین مراجعه به پزشک متخصص و تشخیص درست بیماری‌های مرتبط با اسهال مزمن، اهمیت شایان توجهی دارد. علائم بیماری کرون و سندرم رودهٔ تحریک‌پذیر شبیه یکدیگرند؛ به همین دلیل می‌گویند که کرون، ماسک این بیماری را به چهره می‌زند. اما ممکن است درد مبتلایان به کرون و کولیت اولسروز شدیدتر باشد و همچنین حین مدفوع خونریزی کنند.

تشخیص تفاوت بین بیماری کرون و کولیت روده بسیار مهم است. بیماری کرون می‌تواند هر بخشی از دستگاه گوارش را تحت تأثیر قرار دهد، اما بیماری کولیت روده تنها بر روده اثر می‌گذارد. همچنین، بیماری کرون می‌تواند تمام لایه‌های دیوارهٔ روده را تحت تأثیر قرار دهد، در حالی که بیماری کولیت روده تنها بر غشای روده تأثیرگذار است. بیماری کرون بیشتر بر قسمت انتهایی رودهٔ

کوچک (ایلئوم) و قسمت ابتدایی روده اثر می‌گذارد، اما می‌تواند هر قسمتی از دستگاه گوارش، از دهان گرفته تا مقعد را تحت تأثیر قرار دهد. کولیت روده محدود به رودهٔ بزرگ می‌شود.

این مشکل ما را به این فکر واداشت که بتوانیم به‌کمک متدهای محاسبات نرم، تا حدی صحت تشخیص بیماری را بالا ببریم. در این پژوهش ابتدا به‌کمک رویکرد رأی‌گیری با آرای بالا، نمونه‌هایی که نویز کلاس دارند شناسایی شده و از دیتاست حذف می‌شوند. در گام اول نتایج آزمایش‌های بالینی و کولونوسکوپی ۱۰۰ نمونه از افرادی که بیش از ۵ تا ۱۰ سال مبتلا به بیماری التهاب روده بوده‌اند را همراه با ۱۹ ویژگی مؤثر در تشخیص این بیماری در نظر گرفتیم تا بتوانیم معین کنیم هر نمونه دچار کدام نوع بیماری (کرون یا کولیت) است.

از آنجا که قصد ما از ارائهٔ این پژوهش، مدیریت عدم قطعیت در تفکیک دو نوع بیماری التهاب روده است، ابتدا باید به‌کمک یک رویکرد دسته‌بندی ترکیبی، نمونه‌های نویزی (نمونه‌هایی که پزشک در تشخیص بیماری آنها دچار مشکل شده یا به بیان دیگر، نتیجهٔ رأی‌گیری به‌کمک دسته‌بندی ترکیبی با تشخیص پزشک برای آن نمونه متفاوت است) را پیدا کرده و حذف کنیم. بدین ترتیب می‌توانیم یک دیتاست تقریباً بدون نویز ایجاد کنیم.

استفاده از این روش دسته‌بندی ترکیبی به‌کمک رویکرد رأی‌گیری با رأی اکثریت و رویکرد رأی‌گیری با توافق عام در مقالات بسیاری استفاده شده است، گفتنی است اسلوبان و لاوراک (۲۰۱۵) در مقاله‌ای در خصوص ضعف رویکرد رأی‌گیری با رأی اکثریت و همچنین محدودیت رویکرد توافق عام بحث کرده‌اند. این ضعف باعث می‌شود که نویزها به‌درستی تشخیص داده نشوند. همچنین از محدودیت‌های رویکرد توافق عام نیز به این نکته اشاره شده است که توانایی تشخیص خیلی از نمونه‌های نویزی را ندارد. بنابراین برای حل این مشکل، از الگوریتم رأی‌گیری با توافق بالا استفاده کردیم. فرضیهٔ ما این است که الگوریتم یادگیری با توافق سطح بالا نسبت به الگوریتم یادگیری با رأی اکثریت، صحت بیشتری دارد. گام دوم، دسته‌بندی و ارزیابی صحت خوشه‌بندی روی دیتاستی است که نویزهای آن در مرحلهٔ قبل به‌کمک رویکرد رأی‌گیری توافق سطح بالا شناسایی و حذف شده است.

در این پژوهش به‌کمک پنج الگوریتم مختلف ناهمگن به نام‌های ماشین بردار پشتیبان، درخت تصمیم‌گیری، نایو بیزین، نزدیک‌ترین همسایه و جنگل تصادفی، یک دسته‌بند ترکیبی ساخته شده و برای به‌دست آوردن نتیجهٔ آرای این دسته‌بند، از رویکرد رأی‌گیری با آرای بالا استفاده می‌شود. تفاوت رویکرد رأی‌گیری با آرای بالا و رویکرد رأی‌گیری با رأی اکثریت در این است که در طرح رأی‌گیری با رأی اکثریت باید دست کم سه الگوریتم از پنج الگوریتم موجود برای تعیین دستهٔ نوع بیماری جواب یکسانی داشته باشند تا به‌عنوان جواب رأی‌گیری شناخته

شود. با توجه به تحقیقات انجام شده در این خصوص، رویکرد رأی‌گیری با رأی اکثریت باعث می‌شود که خیلی از نمونه‌های سالم به اشتباه به‌عنوان نمونهٔ نویزی شناخته شده و در نتیجه از دیتاست حذف شوند. ولی رویکرد طرح رأی‌گیری پیشنهاد شده در این پژوهش به حذف شدن تعداد کمتری از نمونه‌های سالم منجر می‌شود. برای گرفتن نتیجهٔ آرا در رویکرد پیشنهادی ما که رأی‌گیری با آرای بالا نام دارد، باید دست کم چهار الگوریتم از پنج الگوریتم موجود یک جواب یکسان در مورد دستهٔ بیماری نمونهٔ مد نظر بدهند تا بتوان آن را نتیجهٔ رأی‌گیری این رویکرد دانست.

پیشینهٔ پژوهش

گوان، یوان، ما و لی (۲۰۱۴) از ایدهٔ رأی‌گیری چندگانه در روش‌های تعیین اختلال مبتنی بر یادگیری کلی استفاده کردند. از نظر تئوری، این مورد را می‌توان با انواع دیگری از روش‌های شناسایی اختلال نیز به‌کار برد (احمد و همکاران، ۲۰۱۷). اسلوبان و لاوراک (۲۰۱۵) رابطهٔ بین تنوع دسته‌بندی ترکیبی و عملکرد تشخیص نویز دسته را در تحلیل‌های تجربی، طرح رأی‌گیری توافق عام و طرح رأی‌گیری با رأی اکثریت بررسی کردند. اگیوز (۲۰۱۱) و بشیر، کامار، خان و ناسیم (۲۰۱۶) یک چارچوب ترکیبی را برای رأی اکثریت سلسله‌مراتبی صحت بالا در دسته‌بندی و پیش‌بینی بیماری در تمام دادگان‌های پزشکی با روش‌های دسته‌بند کننده بررسی کردند. مانند الگوریتم‌های پایه‌ای که درون دسته‌بندکنندهٔ ترکیبی چندلایهٔ رأی‌گیری با رأی اکثریت سلسله‌مراتبی به‌کار برده شده است، فقط روی یک یا دو دادگان بیماری می‌تواند صحت بالایی در دسته‌بندی و پیش‌بینی داشته باشد؛ درحالی که برای پیش‌بینی و دسته‌بندی دادگان‌های دیگر صحت پایینی دارد. از محدودیت‌های روش رأی‌گیری چند لایه این است که پیچیدگی‌های محاسباتی و زمانی بالایی دارد (اکایار، کاناتال، جاسجیت و سوری، ۲۰۰۶). گوان و همکارانش برای تشخیص نویز کلاس، الگوریتم‌های یادگیری دسته‌بندی کننده را خلق کردند که به‌عنوان فیلترهای نویز به‌کار گرفته می‌شود (کالادهار، پوتومتو، راو و وادلامدی، ۱۹۲۶). تامپسون، لاندار، ژو و شیئ (۲۰۱۴) در زمینهٔ مداخله و ارتباط سرطان رودهٔ بزرگ، روش‌های داده‌کاوی انجمنی را شناسایی کردند که به‌منظور مطالعهٔ بیشتر روی تفاوت بین گروه‌های چسبنده و غیرچسبنده استفاده می‌شوند. رگرسیون‌های منطقی چهار متغیر را پیشنهاد دادند. کاتال، الان و بالکان (۲۰۱۱) در زمینهٔ تشخیص نویز با استفاده از الگوریتم نایوبیز همراه با یک فیلتر ورود توانستند به بهبودهایی دست یابند. گوان و همکارانش (۲۰۱۴) از یادگیری ترکیبی برای شناسایی و حذف نمونه‌هایی با برچسب‌گذاری نادرست استفاده کردند. در این مقاله از ایدهٔ رأی‌گیری چندگانه در

روش‌های تعیین اختلال مبنی بر یادگیری کلی استفاده شده است. از نظر تئوری، این مورد را می‌توان با انواع دیگری از روش‌های شناسایی اختلال نیز به‌کار برد. احمد و همکارانش (۲۰۱۷) برای تشخیص بیماری کرون، ترکیب مدل فازی عصبی و طبقه‌بند فازی شبکهٔ عصبی مدل فازی را به‌عنوان فرایند کاهش بُعد برای دسته‌بندی کردن در نظر گرفتند. آگیوز به‌منظور تشخیص بیماری قلبی، سیستم فازی عصبی استنتاجی تطبیقی برای استخراج ویژگی سیگنال‌های صدای قلب، از تبدیل موجک گسسته به باندهای زیر شاخه و آنتروپی هر یک از زیرمجموعه‌های باند استفاده کرد و به‌کمک الگوریتم آنتروپی شانون آن را محاسبه نمود تا ابعاد بردارهای ویژگی را کاهش دهد (جیسون و کوپار، ۲۰۱۶).

در زمینهٔ دسته‌بندی بیماری چشم از مقایسهٔ سه استراتژی دسته‌بندی برای طبقه‌بندی کردن چهار نوع دیتاست چشم استفاده کرد. پروتکل ارائه شدهٔ در این مقاله از سه نوع مختلف از طبقه‌بند تشکیل شده است که عبارت‌اند از: شبکهٔ عصبی مصنوعی، دسته‌بند فازی و دسته‌بند فازی ـ عصبی. ویژگی‌ها از این عکس‌های خام استخراج شده و در مراحل بعد به این الگوریتم‌های دسته‌بند داده می‌شوند (تامپسون و همکاران ۲۰۱۴).

موسوتو و همکارانش در زمینهٔ بیماری التهاب روده، تحقیقاتی به‌منظور ضرورت تشخیص سریع بیماری انجام داده دادند. این مطالعه به‌کمک یادگیری ماشین (ML) برای طبقه‌بندی بیماری از داده‌های آندوسکوپی و بافت‌شناسی ۲۸۷ کودک مبتلا استفاده می‌کند. در تحقیق سه گرتچن، پورسل (۲۰۰۶) مدل تحت نظارت ML با استفاده از داده‌های اندوسکوپی تنها، بافت‌شناسی و داده‌های ترکیبی آندوسکوپی / بافت‌شناسی به دقت طبقه‌بندی ۷۱ درصد، ۷۶/۹ درصد و ۸۲/۷ درصد دست یافتند. مدل ترکیبی مطلوب روی یک گروه آماری مستقل (۴۸ بیمار مبتلا به التهاب روده از کلینیک ساتمپتون) آزمایش شد و ۸۳/۳ درصد بیماران را دقیق طبقه‌بندی کرد. این مطالعه با استفاده از مدل‌سازی ریاضی داده‌های آندوسکوپی و بافت‌شناسی به صحت تشخیص کمک می‌کند. در مجموع، این مقاله یک طرح برای استفاده از ML با داده‌های بالینی ارائه می‌دهد (گوان، یان، لی، ۲۰۱۴)

در جدول ۱، مطالعات پیشین با در نظر گرفتن درجهٔ صحت به‌دست آمده در هر روش و متدهای ارزیابی در هر مقاله دسته‌بندی شده است. در این جدول مطالعات پیشین با توجه به تعریف اهداف، نتایج ارزیابی و روش‌های متفاوت به‌کار برده شده بررسی شده‌اند.

جدول ۱. مقایسهٔ مطالعات پیشین

پدیدآورنده	اهداف	روش	نتایج و پارامترهای ارزیابی
آگیوز (۲۰۱۱)	تشخیص بیماری قلبی	سیستم حمایت از تصمیم‌گیری پزشکی برای دسته‌بندی سیگنال‌های سیستم فازی	روش توسعه‌یافته به صحت دسته‌بندی ۹۸/۳۳ دست یافت
اچاریا و همکاران (۲۰۰۶)	دسته‌بندی بیماری چشم	شبکهٔ عصبی و فازی و فازی ـ عصبی.	طبقه‌بندهای پیشنهاد شده در این مقاله می‌توانند با صحت بیشتر از ۸۵درصد تصویر ناشناخته را طبقه‌بندی کنند
گوان و همکاران (۲۰۱۴)	یادگیری ترکیبی برای شناسایی و حذف نمونه‌هایی با برچسب‌گذاری نادرست	رأی‌گیری چندگانه برای شناسایی نویز	نویز را دقیق‌تر از سیستم رأی دهی یک باره شناسایی می‌کند Mf & cf =38 & 24 4 &89
اسلوبان و لاوراک (۲۰۱۵)	رابطهٔ بین تنوع دسته‌بندی ترکیبی و عملکرد تشخیص نویز دسته	طرح رأی‌گیری توافق عام و طرح رأی‌گیری با رأی اکثریت	الگوریتم رأی‌گیری توافق عام در پیدا کردن نویز دقت بسیار زیادی دارد که باعث می‌شود خیلی از نویزها را نتواند تشخیص دهد؛ یعنی میزان تشخیص صحیح نویز پایین است (۵۶ درصد) precision & recall =56%
بشیر، قمر، حسنخان و نسیم (۲۰۱۶)	سیستم‌های حمایت از تصمیم‌گیری پزشکی	رأی اکثریت سلسله‌مراتبی	رأی اکثریت سلسله‌مراتبی صحت بالایی در دسته‌بندی و پیش‌بینی بیماری در تمام دادگان‌های پزشکی به ما می‌دهد
گوان و همکاران (۲۰۱۳)	تشخیص نویز کلاس	فیلترینگ رأی اکثریت، فیلترینگ توافق عام و فیلترینگ رأی اکثریت	وی توانست دادهٔ الف و دادهٔ ب را به درستی به‌عنوان داده‌هایی که برچسب‌گذاری اشتباه شده‌اند، تشخیص دهد
کالادهار، پوتوموتو و همکاران (۲۰۱۵)	عناصر یادگیری آماری در دادگان بیماری التهاب روده	جنگل تصادفی، درخت AD، لجستیک	الگوریتم‌های Navie Bayes و Part با میزان صحت ۹۷/۲۲ و الگوریتم‌های Simple Cart و ZeroR با میزان صحت حداقل ۵۰ درصد توانسته‌اند نمونه‌های آزمایشی را دسته‌بندی کنند.

ادامهٔ جدول ۱

پدیدآورنده	اهداف	روش	نتایج و پارامترهای ارزیابی
کوپر و پورسل (۲۰۱۳)	رابطه‌ای که بین رژیم و بیماری کرون است	داده‌کاوی	تجزیه و تحلیل رگرسیون سری زمانی نشان داد که ۳ نفر از ۷ نفر شرکت‌کننده فعل و انفعالات مواد غذایی غیرپیچیده و کلاسیک را نشان دادند. این فعل و انفعالات ساده یا اثر متقابل ساده، شامل ارتباط غذایی شناخته شده با غذاهایی که دیر هضم هستند.
تامسون و همکاران (۲۰۰۶)	ارتباط سرطان رودهٔ بزرگ با امریکایی‌های قدیمی	داده‌کاوی انجمنی رگرسیون لجستیک	باید به بیماران توصیه کنند که نبود علائم یا کولیت ارزش غربالگری را زیر سؤال نمی‌برد.
کاتال و الان (۲۰۱۴)	حضور نویز کلاس و نویز ویژگی در دادگان‌های اندازه‌گیری نرم‌افزار و تأثیر منفی که بر عملکرد دسته‌بند کننده‌های مبنی بر یادگیری ماشین دارند	از الگوریتم نایوبیز همراه با یک فیلتر ورود تعداد	چنانچه برچسب‌های کلاس واحدهای نویزی شناخته شده درست باشند، کارکرد پیش‌بینی‌کننده‌های خطا با استفاده از الگوریتم نایوبیز همراه با یک فیلتر ورود تعداد، بهبود پیدا می‌کنند.
علی مردانی و همکاران (۱۳۹۴)	ارائهٔ روش نظارتی برای نظرکاوی در زبان فارسی با استفاده از لغت‌نامه و الگوریتم SVM	استفاده از طبقه‌بند ماشین بردار پشتیبان	در این مقاله روشی پژوهشی برای نظرکاوی در زبان فارسی ارائه شده است که از لغت نامه sentiwordnet و الگوریتم SVM استفاده می‌کند.
اولیایی و همکاران (۲۰۱۶)	ارائهٔ یک ساختار پروتئین بر پایهٔ شبکه‌های پیچیده و آنالیز کاهشی	استفاده از شبکهٔ عصبی و ماتریس‌های پراکندگی	در این مقاله یک ساختار جدید با ترکیبی از شبکه‌های پیچیده عصبی و ماتریس‌های پراکندگی برای هر چه بهتر تشخیص ساختار پروتئین‌ها ارائه شده است

روش‌شناسی پژوهش

در دنیای امروز علی‌رغم پیشرفت‌های صورت‌گرفته در امر تشخیص پزشکی بیماری‌ها، هنوز بیماری التهاب روده که به دو بیماری کولیت و کرون دسته‌بندی می‌شود، نه‌تنها در مراحل اولیه قابل تشخیص نیست، بلکه تا چندین سال بعد نیز تشخیص داده نمی‌شود. این مشکل ما را به این فکر واداشت که بتوانیم به کمک متدهای محاسبات نرم، تا حدی صحت تشخیص بیماری را

بالا ببریم. در گام اول نتایج آزمایش‌های بالینی و کولونوسکوپی ۱۰۰ نمونه از افرادی که بیش از ۵ تا ۱۰ سال مبتلا به بیماری التهاب روده بوده‌اند را همراه با ۱۹ ویژگی مؤثر در تشخیص این بیماری در نظر گرفتیم تا معین کنیم هر نمونه دچار بیماری کرون است یا کولیت. از آنجا که هدف از ارائهٔ این پژوهش، مدیریت عدم قطعیت در تفکیک دو نوع بیماری التهاب روده است، باید ابتدا به‌کمک یک رویکرد دسته‌بندی ترکیبی، نمونه‌های نویزی (نمونه‌هایی که پزشک در تشخیص بیماری آنها با مشکل مواجه شده است) یا به بیان دیگر، نتیجه رأی‌گیری به کمک دسته‌بندی ترکیبی با تشخیص پزشک برای آن نمونه متفاوت است را پیدا کرده و حذف کنیم. بدین ترتیب یک دیتاست تقریباً بدون نویز ایجاد می‌شود. استفاده از این روش دسته‌بندی ترکیبی به‌کمک رویکرد رأی‌گیری با رأی اکثریت و رویکرد رأی‌گیری با توافق عام را در مقالات مختلف بارها مشاهده شده است، اما اسلوبان و لاوراک (۲۰۱۵) در تحقیق خود، به ضعف رویکرد رأی‌گیری با رأی اکثریت و همچنین محدودیت رویکرد توافق عام اشاره کرده‌اند. همین ضعف باعث می‌شود که نویزها به درستی تشخیص داده نشوند. در این پژوهش، برای تشخیص نویز از پنج الگوریتم و برای دسته‌بندی از الگوریتم جنگل تصادفی استفاده شده است. بعد از حذف نویز، دیتاست بدون نویز را دسته‌بندی کردیم. دسته‌بندی ترکیبی روی تشخیص نویز اعمال شده است. سپس نمونه‌های سالم را دسته‌بندی کرده و صحت نمونه‌های سالم توسط الگوریتم جنگل تصادفی صحت دسته‌بند اندازه‌گیری شده است.

دربارهٔ محدودیت‌های رویکرد توافق عام به این نکته اشاره شده است که قادر به تشخیص خیلی از نمونه‌های نویزی نیست. بنابراین برای رفع این مشکل از الگوریتم رأی‌گیری با توافق بالا استفاده کردیم. فرضیهٔ ما این است که الگوریتم یادگیری با توافق سطح بالا صحت بالاتری به نسبت الگوریتم یادگیری با رأی اکثریت دارد. گام دوم، دسته‌بندی و ارزیابی صحت خوشه‌بندی روی دیتاستی است که در مرحلهٔ قبل به‌کمک رویکرد رأی‌گیری توافق سطح بالا نویزهایش تشخیص داده شده و از دیتاست حذف شده‌اند. به‌کمک پنج الگوریتم مختلف یک دسته‌بند ترکیبی می‌سازیم و برای به‌دست آوردن نتیجهٔ آرای این دسته‌بند از رویکرد رأی‌گیری با آرای بالا استفاده می‌کنیم. تفاوت رویکرد رأی‌گیری با آرای بالا با رویکرد رأی‌گیری با رأی اکثریت در این است که در طرح رأی‌گیری با رأی اکثریت باید از پنج الگوریتم موجود، دست کم سه الگوریتم جواب یکسانی برای تعیین دستهٔ نوع بیماری داشته باشند تا به‌عنوان جواب رأی‌گیری شناخته شود. با توجه به تحقیقات انجام شده این خصوصیت رویکرد رأی‌گیری با رأی اکثریت باعث می‌شود که خیلی از نمونه‌های سالم به اشتباه به‌عنوان نمونهٔ نویزی شناخته شده و در نتیجه از دیتاست حذف شوند. ولی رویکرد طرح رأی‌گیری پیشنهادی ما به حذف شدن تعداد

کمتری از نمونه‌های سالم منجر می‌شود. برای گرفتن نتیجهٔ آرا در رویکرد پیشنهادی ما که رأی‌گیری با آرای بالا نام دارد، باید دست کم چهار الگوریتم از پنج الگوریتم موجود جواب یکسانی برای دستهٔ بیماری نمونهٔ مد نظر داشته باشند تا بتوان آن را نتیجهٔ رأی‌گیری این رویکرد دانست.

تعاریف اولیه

نویز برچسب یا همان نویز کلاس

نویز یک ارتباط است که نمی‌توان آن را به درستی طبقه‌بندی کرد. نویز با قوانینی که برنامه برای تعیین و دسته‌بندی نوع بیماری در یک زبان خاص استفاده می‌کند، همخوانی ندارد. داده‌های نویزی به‌طور خلاصه داده‌هایی هستند که هنگام ثبت یا تغییر آنها اشتباهی رخ داده و مقدار آنها نامعتبر است. نویز کلاس بدین معناست که تشخیص پزشک برای نوع بیماری نمونه با جواب به‌دست آمده از رأی‌گیری با آرای بالا برای همان نمونه، متفاوت باشد. این کار نه تنها برای نشان دادن بهبود در پیش پردازش، پاکسازی و درک داده صورت می‌گیرد، بلکه به‌منظور بهبود دسته‌بندی کردن داده‌ها برای پیش‌بینی با کیفیت بالا نیز انجام می‌شود. زمانی که داده‌هایی با ویژگی‌های مشخص که متعلق به کلاس «الف» است، به‌دلیل تشخیص اشتباه الگوریتم یا همان تشخیص اشتباه روش دسته‌بندی مانند دسته‌بندی ترکیبی، داده را متعلق به یک کلاس دیگر مانند «ج» دانستند، داده را درون کلاس «ج» دسته‌بندی کردند. یکی از دلایل تشخیص اشتباه کلاس توسط روش‌های دسته‌بندی این است که مقادیر صفت‌های داده‌ای که در اصل متعلق به کلاس «الف» است، به مقادیر صفت‌های داده‌های متعلق به کلاس «ج» بسیار نزدیک است. در این صورت احتمال کلاس‌بندی اشتباه داده زیاد می‌شود که به این نوع اشتباه در دسته‌بندی، نویز برچسب گفته می‌شود. در این پژوهش از روش‌هایی مانند مدل سیستم فازی ـ عصبی، تِیوری راف و شبکهٔ عصبی راف و دسته‌بندی‌هایی از قبیل K نزدیک‌ترین همسایه[1]، ماشین بردار پشتیبان[2]، جنگل تصادفی[3]، نایو بیزین[4] و درخت تصمیم‌گیری[5] استفاده شده است که در ادامه هر یک از این روش‌های دسته‌بندی را جداگانه شرح می‌دهیم.

---

1. K Nearest Neighbor
2. Support Vector Machine
3. Random Forest
4. Naive Bayes
5. Decision Tree

روش دسته‌بندی نایو بیز

در این جا برای بررسی چگونگی انجام دسته‌بندی بیـزی، از تئـوری اولیـهٔ بیـز آغـاز مـی‌کنـیم. یادگیری احتمالی می‌تواند معادل محاسبهٔ $p(C = c|d)$ باشد، نمونهٔ x را مقادیر ویژگی مشاهده شدهٔ a_1 تا a_k در نظر بگیرید. این مقادیر برای پیش‌بینی یک کلاس گسستهٔ C به‌کار می‌روند. هدف ما پیش‌بینی و انتخاب دسته‌ای است که در آن احتمال P ماکزیمم شود. فرضی که در بیـز ساده وجود دارد این است که ویژگی‌ها به‌طور شرطی از هم مستقل‌اند. فرض می‌کنیم کـه بـرای یک دستهٔ C همهٔ ویژگی‌ها به‌طور شرطی از هم مستقل هستند. از مزایای بیز ساده می‌تـوان بـه اجرای راحت و نتایج خوب آن برای بسیاری از کاربردها اشاره کرد. معایب بیز ساده این است که استقلال شرطی دسته‌ها فرضی است، اما در مواردی که این فرض برقرار نیست، دقت مدل پایین می‌آید. درعمل وابستگی وجود دارد و فرض استقلال همواره برقرار نیست.

روش K نزدیک‌ترین همسایه

این الگوریتم از سه گام تشکیل می‌شود. گـام نخسـت، محاسـبهٔ فاصـلهٔ نمونـهٔ ورودی بـا تمـام نمونه‌های آموزشی است. گام دوم، نمونه‌های آموزشی را براساس فاصله و انتخاب چنـد همسـایهٔ نزدیک‌تر مرتب می‌کند. در گام سوم، از دسته‌ای کـه اکثریـت را در همسـایه‌هـای نزدیـک دارد، به‌عنوان تخمین برای دستهٔ نمونهٔ ورودی استفاده می‌شود. از مزایای روش نزدیک‌ترین همسایه، ساده بودن و در عین حال مؤثر بودن آن است و از معایب این روش می‌توان به سرعت پایین آن اشاره کرد.

روش درخت تصمیم گیری

ساختار کلی درخت تصمیم به این صورت است که یک گـرهٔ ریشـه در بـالای آن و بـرگ‌هـا در پایین آن هستند. یک نود جدید در گرهٔ ریشه وارد مـی‌شـود. در ایـن گـره یـک آزمـون صـورت می‌گیرد تا معلوم شود که این نود به کدام یک از گره‌های فرزند (شاخهٔ پایین‌تر) تعلـق دارد. ایـن فرایند آنقدر ادامه پیدا می‌کند تا نود جدید به گرهٔ برگ برسد. تمام نودهایی که بـه یـک بـرگ از درخت می‌رسند، در یک دسته قرار می‌گیرند. از مزایای درخت تصمیم‌گیری این است که نیاز بـه محاسبات پیچیده‌ای برای دسته‌بندی داده‌ها ندارد. ایـن روش نشـان مـی‌دهـد کـدام فیلـدها یـا متغیرها، تأثیر مهمی در پیش‌بینی و دسته‌بندی دارند. هرچه متغیـر بـه ریشـه نزدیـک‌تـر باشـد، اهمیت آن بیشتر است. در بیان معایب این روش باید گفت که بعضی از روش‌های درخت تصمیم تنها می‌توانند روی متغیرهای هدف دودویی مانند بله یا خیر دسته‌بندی یـا پـیش‌بینـی کننـد. در برخی روش‌ها هنگامی که تعداد مثال‌ها یا رکوردهای هر دسته کم باشند، درصد خطا بالا می‌رود.

همچنین در برخی الگوریتم‌های دیگر نیز برای پیدا کردن بهترین ویژگی، وضعیت هر ویژگی نگهداری می‌شود که نیاز به حافظهٔ زیادی دارد.

روش ماشین بردار پشتیبان

این الگوریتم، اساساً یک جداکنندهٔ دودویی است. یعنی برای دسته‌بندی دو کلاس از آن استفاده می‌کنیم. یک تشخیص الگوی چند کلاسه می‌تواند به وسیلهٔ ترکیب ماشین‌های بردار پشتیبان دو کلاسه حاصل شود. برای حل مسائل چند کلاسه، رهیافت کلی، کاهش مسئلهٔ چند کلاسه به چندین مسئلهٔ دودویی است. هر یک از مسائل با یک جداکنندهٔ دودویی حل می‌شود. سپس خروجی جداکننده‌های دودویی ماشین بردار پشتیبان با هم ترکیب شده و به این ترتیب مسئلهٔ چند کلاسه حل می‌شود. بردارهای پشتیبان به زبان ساده، مجموعه‌ای از نقاط در فضای چند بعدی داده‌ها هستند که مرز دسته‌ها را مشخص می‌کنند و مرزبندی و دسته‌بندی داده‌ها براساس آنها انجام می‌شود و با جابه‌جایی یکی از آنها، خروجی دسته‌بندی ممکن است تغییر کند. از معایب این الگوریتم می‌توان به این نکات اشاره کرد که داده‌های گسسته و غیر عددی با این روش سازگار نیستند و باید تبدیل شوند. همچنین ماشین‌های مبتنی بر بردار پشتیبان به محاسبات پیچیده و زمان‌بر نیاز دارند و به‌دلیل پیچیدگی محاسباتی، حافظهٔ زیادی مصرف می‌کنند. از جمله مزیت‌های روش ماشین بردار پشتیبان این است که برای دسته‌بندی و تفکیک داده‌ها، الگوریتم‌های بسیار قدرتمندی هستند، به‌خصوص زمانی که با سایر روش‌های یادگیری ماشین مانند روش جنگل تصادفی تلفیق شوند. این روش برای مسائل با دقت بسیار بالا که به ماشین داده‌ها نیاز داریم، به شرط اینکه توابع نگاشت را به درستی انتخاب کنیم، بسیار خوب عمل می‌کند.

جنگل‌های تصادفی

جنگل تصادفی درخت تصمیم‌های زیادی را تولید می‌کند. برای طبقه‌بندی یک شیء جدید از بردار، ورودی را در انتهای هر یک از درختان جنگل تصادفی قرار می‌دهد. هر درخت به ما یک طبقه‌بندی می‌دهد و می‌گوییم این درخت به آن کلاس «رأی» داده است. جنگل، طبقه‌بندی کننده‌ای که بیشترین رأی را (بین همه درخت‌های جنگل) داشته باشد، انتخاب می‌کند.

دسته‌بندی ترکیبی و انواع آن

دسته‌بندی ترکیبی به معنای استفاده از چند الگوریتم دسته‌بندی (در قسمت قبل توصیف شده‌اند) برای تصمیم‌گیری در تعیین دسته برای داده است. این روش از تعیین کلاس دقت، صحت و

فراخوانی بیشتری نسبت به استفاده از یک الگوریتم منفرد دارد. برای مثال، در این مقاله از دو طرح رأی‌گیری در دسته‌بندی ترکیبی[1] استفاده شده است که در ادامه به شرح مختصری از هر دو طرح پرداخته می‌شود.

طرح رأی گیری رأی اکثریت

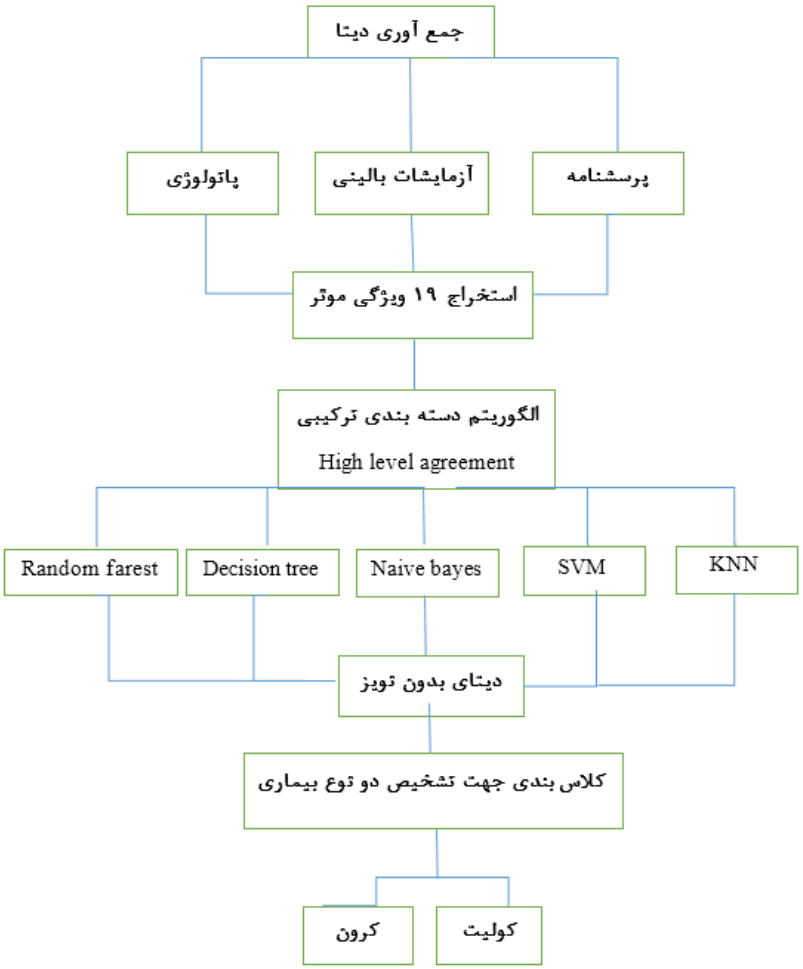
در روش طرح رأی‌گیری رأی اکثریت[2] باید بیشتر از نصف تعداد الگوریتم‌ها به‌صورت جداگانه، مقدار یکسانی از برچسب را به‌عنوان جواب خروجی بدهند. جواب دسته‌بند ترکیبی همان جواب خروجی تعداد اکثریت الگوریتم‌ها می‌شود. از معایب این روش می‌توان گفت در کلاس‌بند ترکیبی خیلی از داده‌های سالم، به‌عنوان نمونهٔ نویزی شناخته می‌شوند؛ زیرا اگر تعداد اکثریت الگوریتم‌ها یعنی بیشتر از نصف آنها، آن داده را غیر نویزی تشخیص دهند، داده دچار فرایند تغییر برچسب یا حذف شدن نمی‌شود و با اینکه تعدادی از الگوریتم‌ها که شمار آنها کمتر از نصف تعداد کل الگوریتم‌هاست، آن داده را نویزی تشخیص دادند با درصد خطای زیادی در تشخیص نویز مواجه می‌شویم (اسلوبان و لاوراک، ۲۰۱۵).

طرح رأی گیری با توافق بالا

در طرح رأی‌گیری از چند نفر برای دسته‌بندی ترکیبی با توافق بالا استفاده می‌شود. داده‌هایی که توسط سه تا پنج الگوریتم، نویزی شناخته شوند، از دسته‌بندی حذف می‌شوند؛ زیرا با اطمینان نسبتاً زیادی آن داده نویزی است. اگر برچسب داده به‌کمک نصف تعداد الگوریتم‌ها درست پیش‌بینی شود، باید آن داده دوباره برچسب‌گذاری شود. در این صورت با اطمینان بهتری می‌توانیم نویز را تشخیص دهیم و داده را در دستهٔ مناسب، کلاس‌بندی کنیم (علی مردانی و آقایی، ۱۳۹۴). بعد از تشخیص نویز، دیتای بدون نویز با استفاده از الگوریتم‌های محاسبات نرم به‌منظور مدیریت عدم قطعیت برای صحت تشخیص بیماری کرون و کولیت، کلاس‌بندی می‌شود. در این مرحله چون خبرگان از ترجیحات غیر قطعی استفاده می‌کنند، از سه روش سیستم فازی عصبی، تئوری مجموعهٔ راف و الگوریتم بهینه شدهٔ شبکهٔ عصبی راف (بهینه شدن مقدار α و β توسط الگوریتم بهینه‌ساز رقابت استعماری) برای مدیریت عدم قطعیت و مقایسه استفاده می‌کنیم.

---

1. Classifying ensemble voting scheme
2. Majority voting scheme



شکل ۱. مدل بررسی الگوریتم رأی‌گیری با توافق سطح بالا

روش کار تحقیق

ابتدا به‌کمک پنج الگوریتمی که در قسمت تعاریف اولیه بیان شدند، یک دسته‌بند ترکیبی می‌سازیم. سپس برای به‌دست آوردن نتیجهٔ آرای این دسته‌بند از رویکرد رأی‌گیری با آرای بالا استفاده می‌کنیم. تفاوت رویکرد رأی‌گیری با آرای بالا و رویکرد رأی‌گیری با رأی اکثریت در این است که در طرح رأی‌گیری با رأی اکثریت باید دست کم سه الگوریتم از پنج الگوریتم موجود

جواب یکسانی برای تعیین دستهٔ نوع بیماری داشته باشند تا به‌عنوان جواب رأی‌گیری شناخته شود؛ در حالی که در طرح رأی‌گیری با آرای بالا، باید چهار یا پنج الگوریتم جواب مشترکی در تعیین دسته‌بندی نمونه داشته باشند تا بتوان آن نمونه را نتیجهٔ رأی‌گیری این رویکرد دانست. با توجه به تحقیقات اسلوبان و لاوراک (۲۰۱۵) این خصوصیت رفتاری رویکرد رأی‌گیری با رأی اکثریت موجب می‌شود که خیلی از نمونه‌های سالم به اشتباه به‌عنوان نمونهٔ نویزی شناخته شده و در نتیجه از دیتاست حذف شوند. اما رویکرد طرح رأی‌گیری پیشنهاد شدهٔ ما به حذف شدن تعداد کمتری از نمونه‌های سالم منجر می‌شود. شواهد تجربی نشان می‌دهد صحت دسته‌بندی دیتاست کلون که نویزهای کلاس درون آن به‌کمک رویکرد رأی‌گیری با توافق سطح بالا تشخیص داده شده و حذف شده‌اند، بالاتر از دیتاستی است که نویزهای کلاس درون آن توسط رویکرد رأی‌گیری با رأی اکثریت تشخیص داده شده و حذف شده است.

در گام اول برای تهیهٔ دیتاست، از نظرسنجی ۱۰۰ بیمار و آزمایش‌ها و نتایج پاتولوژی به‌عمل آمده در مرحلهٔ خاموشی و شعله‌وری آنان کمک گرفتیم. سپس به بررسی یک ویژگی مؤثر هایپرپلازی و درگیری رودهٔ کوچک و تشخیص نهایی پاتولوژی بیماران مبنی بر کرون یا کولیت پرداختیم. در جدولی که تشکیل دادیم، برای برچسب تشخیص و تفکیک نمونه، عدد ۱ را برای بیماری کرون و عدد ۲ را برای تشخیص بیماری کولیت در نظر گرفتیم و مقدار نتایج ویژگی‌های مؤثر در تشخیص بیماری را از آزمایش‌ها استخراج کرده و در ستون‌های جداول به‌عنوان معیار مقایسات تفکیک و تشخیص بیماری بررسی کردیم.

این ویژگی‌ها را به فایل اکسل وارد می‌کنیم، سپس برای هر نمونه نوع بیماری کولیت یا کرون را در ستون آخر که ستون برچسب است، قرار می‌دهیم. گام دوم تشخیص نویز کلاس دیتاست به‌دست آمده است. نویز کلاس به این معناست که جوابی که دکتر به‌عنوان تشخیص بیماری نمونه داده است با جواب به‌دست آمده از رأی‌گیری با آرای بالا برای همان نمونه، متفاوت باشد. این کار نه تنها برای نشان دادن بهبود در پیش پردازش، پاکسازی و درک داده صورت می‌گیرد، بلکه به‌منظور بهبود در دسته‌بندی کردن داده‌ها به گونه‌ای که به پیش‌بینی با کیفیت بالا منجر شود نیز، انجام می‌شود. برای اثبات برتری عملکرد طرح رأی‌گیری با توافق بالا به نسب طرح رأی‌گیری با اکثریت آرا در تشخیص نمونه‌های نویزی، باید به‌صورت مصنوعی ۱۰ درصد نویز کلاس به دیتاست اضافه کنیم. پس از تشخیص نویز به کمک رأی‌گیری با آرای بالا، نوبت به حذف نمونه‌های نویزی می‌رسد. پس از حذف نمونه‌های نویزی دیتاستی که به‌دست می‌آوریم، تعداد سطرهای کمتری دارد، اما اطمینان بالاتری به سالم بودن نمونه‌های دیتاست بعد از حذف نویز به نسبت نمونه‌های دیتاست قبل از حذف نویز داریم.

یافته‌های پژوهش

مجموعهٔ داده مورد استفاده در ارزیابی‌ها

مجموعهٔ داده با سه روش پرسشنامه و نتایج پاتولوژی و آزمایش‌های بالینی از ۲۰۰ نمونه بیمار مبتلا به التهاب روده که بین ۵ تا ۱۰ سال درگیر بیماری کرون و کولیت بودند، جمع‌آوری شده است. از بین این ویژگی‌ها، ۱۹ ویژگی مؤثر در تشخیص بیماری توسط پزشک در نظر گرفته شده است. برای مجموعهٔ دیتاست نیز در مراحل پیش پردازش از ۲۰۰ نمونهٔ جمع‌آوری شده، ۱۰۰ نمونه که دارای ویژگی برتر بودند، انتخاب شدند. این دیتاست توسط نظرسنجی از ۱۰۰ بیمار نمونه و تهیهٔ آزمایش‌های به‌عمل آمده در مرحلهٔ خاموشی و شعله‌وری آنها و بررسی یک ویژگی مؤثر هایپرپلازی (درگیری رودهٔ کوچک و تشخیص نهایی پاتولوژی بیماران مبنی بر کرون یا کولیت) ثبت شد و در جدولی به‌عنوان برچسب تشخیص و تفکیک نمونه (عدد ۱ برای بیماری کرون و عدد ۲ برای بیماری کولیت) در نظر گرفته شد. مقدار ویژگی‌های مؤثر در تشخیص بیماری از آزمایش‌ها به‌دست آمدند و در ستون هر جدول به‌عنوان معیار مقایسات تفکیک و تشخیص بیماری بررسی شدند. ویژگی‌های مؤثر با توجه به نظر پزشک متخصص از آزمایش‌های بالینی و پاتولوژی و پرسشنامه استخراج شده که در جدول ۲ با توجه به ویژگی مؤثر در تشخیص دو نوع بیماری کولیت و کرون به تفسیر ویژگی‌ها پرداخته شده است. برای نمونه در بیماری کرون، بیماران به‌ندرت خونریزی گوارشی یا درگیری مفاصل دارند و بیشتر از مشکل کبد رنج می‌برند، در صورتی که بیماران کولیتی با خونریزی‌های شدید و درد مفاصل درگیرند و کمتر با مشکل کبد مواجه می‌شوند.

جدول ۲. معرفی ویژگی‌های مؤثر درتشخیص بیماری کولیت و کرون

Features	Range	
Eosiphilin	[۰-۱]	ائوزینوفیل‌ها به افزایش سطح التهاب که نقش مفیدی در جداسازی و کنترل ناحیهٔ ملتهب برعهده دارد، کمک می‌کنند. ولی گاهی اوقات سطح التهاب بیش از میزان لازم است و این به بروز علائم مشکل‌زا و حتی آسیب‌های بافتی منجر می‌شود. برای مثال، ائوزینوفیل‌ها در بروز علائم آسم و آلرژی‌هایی چون تب یونجه نقش کلیدی دارند. گذشته از این، سایر اختلالات سیستم ایمنی نیز می‌تواند به مزمن شدن التهاب منجر شود.
MCH	[۲۶-۳۲]	مخفف و مختصر شده Mean Corposcular of Hemoglobin است و میزان متوسط هموگلوبین در هر سلول را نشان می‌دهد.
WBC	[۴۰۰۰-۱۰۰۰۰]	مخفف WHITE BLOOD CELL است و اشاره به تعداد و شمارش سلول‌های سفید خون دارد که شامل نوتروفیل، منوسیت، لنفوسیت، بازوفیل و ائوزینوفیل می‌شود. اندازه‌گیری مقدار گلبول‌های سفید خون یکی از روش‌های اصلی تعیین وجود عفونت در بدن است چون این سلول‌ها که جزء سیستم دفاعی بدن هستند، در شرایط بیماری‌های عفونی و غیرعفونی واکنش‌های مختلفی از خود نشان می‌دهند.

ادامهٔ جدول ۲

Features	Range	
RBC	[۴/۳۰–۵/۸۰]	مخفف RED BLOOD CELL است و به تعداد گلبول‌های قرمز خون اشاره می‌کند. خونریزی‌های گوارشی یا خون‌ریزی‌های واضح از محل زخم، سوءتغذیه و فقر آهن یا کمبود ویتامین B۱۲ ، شکستن سلول‌های خونی یا همولیز در اثر بعضی بیماری‌های خاص مثل فاوسیم، بعضی مشکلات ژنتیکی مثل گلبول‌های قرمز سلول داسی‌شکل و مشکلات مغز استخوان، موجب پایین آمدن مقدار گلبول‌های قرمز می‌شوند.
Hemoglobin,	[۱۲–۱۷]	میزان رنگدانهٔ خون انسان بوده و مقدار آن در خون نشان‌دهندهٔ کم خونی یا پرخونی یا طبیعی بودن خون فرد است.
ALT-SGPT	[۰–۳۸]	قسمت عمدهٔ ALT-SGPT برعکس AST به‌طور طبیعی در کبد یافت می‌شود. نمی‌توان گفت که این آنزیم منحصراً در کبد قرار دارد، اما کبد جایی است که دربرگیرندهٔ بیشترین غلظت این آنزیم است. این آنزیم در نتیجهٔ آسیب کبدی وارد خون می‌شود، بنابراین از این آنزیم نسبتاً به‌عنوان شناسایی‌کنندهٔ ویژهٔ موقعیت کبدی استفاده می‌شود.
Esr2th,	[۰–۲۰]	erythrocyte sedimentation rate.
Esr1th	[۰–۳۰]	وقتی التهابی در بدن جود دارد، پروتئین‌های خاصی باعث چسبیدن گلبول‌های قرمز به هم شده و در نتیجه موجب رسوب آنها بیشتر از حالت طبیعی می‌شوند. این پروتئین‌ها توسط کبد و سیستم ایمنی در شرایط غیر طبیعی مثل عفونت، بیماری خود ایمنی یا سرطان ساخته می‌شوند.
crp	[۰–۵۰]	اندازه‌گیری CRP قابل اعتمادترین وسیلهٔ تشخیص و کنترل التهاب‌های حاد باکتریایی و همچنین عفونت‌های پنهان است. در بیماری‌های اوتوایمیون نیز با اندازه‌گیری مقدار CRP می‌توان از شدت و پیشرفت بیماری آگاهی یافت؛ زیرا به ندرت اتفاق می‌افتد که بدون التهاب بافتی از بدن، مقدار آن همیشه بالا باشد.
calprotectine,	[۰–۲۵]	برای تشخیص بیماران مبتلا به کرون، کولیت اولسراتیو و سرطان‌های کلورکتال کالپروتکتین پروتئینی آنتی باکتریال با قابلیت اتصال به کلسیم و روی است و به‌دنبال فعال شدن نوتروفیل‌ها یا اتصال مونوسیت‌ها به آندوتلیال، کالپروتکتین موجود در غشا، این سلول‌ها آزاد شده و میزان آن در سرم یا مایعات بدن افزایش می‌یابد.
alk	[۱۸۰–۱۲۰۰]	آنزیمی است که در بافت‌های مختلف بدن وجود دارد. برای مثال در کبد، گلبول‌های سرخ و.... بالا بودن این آنزیم در مقادیر غیر طبیعی نشان‌دهندهٔ حضور بیماری در فرد است، مانند آسیب‌های کبدی ناشی از هپاتیت یا کبد چرب.
JointInvolve	[۱]	درگیری مفاصل
Age	[۱–۷۰]	سن
City	[۱–۲]	محل زندگی (نواحی خشک و نواحی ساحلی)
Bleeding,	[۱–۲]	خونریزی

ادامهٔ جدول ۲

Features	Range	
Platete	[۱۵۰-۴۵۰]	بیماری آرتریت روماتویید، کم‌خونی فقر آهن، مشکلات بعد از برداشتن طحال، بعضی سرطان‌ها و بیماری‌های ژنتیکی خاص باعث افزایش مقدار پلاکت می‌شوند.
Hematocrit	[۳۶/۰-۴۵/۰]	نسبت حجم سلولی خون و بخش مایع خون را نشان می‌دهد. سوختگی، اسهال شدید، بیماری‌های انسدادی ریوی، از دست دادن زیاد آب، تولید بیش از حد گلبول قرمز، عوامل افزایش HCT هستند. بیماری‌هایی که باعث به‌وجود آمدن شکل‌های غیرطبیعی گلبول قرمز می‌شوند (مثل بیماری گلبول قرمز داسی‌شکل) مقدار HCT را تغییر می‌دهند. وقتی مقدار گلبول سفید به شدت بالا باشد بر مقدار HCT مؤثر است. در صورت طبیعی بودن اندازه‌های گلبول قرمز، مقدار هماتوکریت ۳ برابر هموگلوبین است. هماتوکریت را نباید بلافاصله بعد از خون‌ریزی شدید اندازه‌گیری کرد.
MCV	[۷۶-۹۶]	بی‌نظمی در این بخش می‌تواند به علت کم خونی یا سندرم خستگی مزمن باشد.

همچنین چهار ویژگی مؤثر سن، محل زندگی (برای ناحیهٔ آب و هوایی خشک عدد ۲ و نواحی مرطوب عدد ۱)، خونریزی و درگیری مفاصل را توسط پرسشنامه از بیماران تهیه کرده و در ستون جدول داده‌ها به‌عنوان ویژگی مؤثر درج کردیم. ستون آخر که به پنل‌بندی بیماری اختصاص داده شده، از نتایج کلونوسکوپی و پاتولوژی بیماران با مقدار ۱ برای بیماری کرون و ۲ برای بیماری کولیت به‌دست آمده است.

انجام گام‌های آزمایش

در گام اول برای جمع‌آوری دیتاست از سه روش پرسشنامه، آزمایش‌های بالینی و نتیجهٔ پاتولوژی ۲۰۰ بیمار استفاده کردیم و بر اساس نظرسنجی آنها، ۱۰۰ بیمار به‌عنوان نمونهٔ مؤثر توسط فرد خبره انتخاب شدند و آزمایش‌هایی در مرحلهٔ خاموشی و شعله‌وری آنها انجام گرفت. سپس به بررسی یک ویژگی مؤثر هایپرپلازی و درگیری رودهٔ کوچک و تشخیص نهایی پاتولوژی بیماران مبنی بر کرون یا کولیت پرداختیم. در فایل جمع‌آوری دیتاست برچسب ۱ و ۲ به‌عنوان برچسب تشخیص و تفکیک نمونه (۱ برای بیماری کرون و ۲ برای تشخیص بیماری کولیت) در نظر گرفته شده است. ۱۹ ویژگی مؤثر در تشخیص بیماری را از آزمایش‌ها استخراج کرده و در ستون‌های جدول ۲ به‌عنوان معیار ارزیابی تفکیک و تشخیص بیماری بررسی کردیم. در مرحلهٔ بعد با استفاده از الگوریتم رأی‌گیری با توافق بالا تشخیص نویز را انجام دادیم. گفتنی است پس از تشخیص نویز به‌کمک رأی‌گیری آرای بالا، دیتاست تعداد سطرهای کمتری دارد، اما به سالم بودن دیتاست بعد از حذف نویز در مقایسه با دیتاست قبل از حذف نویز، اطمینان بیشتری داریم. سپس برای ساختن، آموزش و ارزیابی طبقه‌بند مد نظر روش cross validation را انتخاب کرده

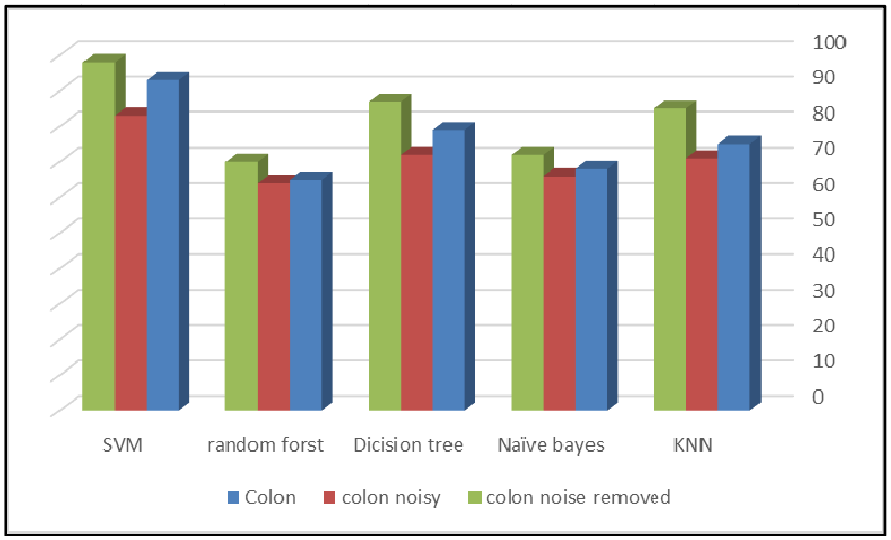
و تعداد foldها را ۱۰ در نظر گرفتیم. از مدل‌های ایجاد شده در هر روش می‌توان برای شناسایی سایر آنها که در دیتاست آموزش (train) و آزمایش ما موجود نبود، استفاده کرد، البته این مدل‌ها قدرت خود را با روش cross-validation نشان داده‌اند، در همین روش نیز ابتدا دیتاست به بخش‌های مساوی به تعداد foldها تقسیم شده، سپس به‌صورت متوالی، بخشی از دیتاست برای آزمایش و سایر بخش‌ها برای آموزش استفاده می‌شود تا تمام بخش‌ها برای مدل یک بار در حالت آزمایش استفاده شوند. پس در هر بار بعد از آموزش، سیستم با گروه جدیدی مواجه می‌شود که قبلاً آنها را ندیده و باید نوع آنها را تشخیص دهد. البته مراحل آموزش و آزمایش سیستم به‌صورت کاملاً مجزا و با دیتاست‌های مجزا نیز انجام شده است. برای این منظور دیتاست اصلی به‌صورت random با درصد ۷۰–۳۰ جدا می‌شود (ابتدا دیتاست را randomize نموده، سپس ۷۰ درصد آن را برای یادگیری سیستم و ۳۰ درصد باقی را برای آزمایش سیستم جدا می‌کنیم). در این حالت ابتدا سیستم را با بخش جدا شده برای یادگیری، آموزش می‌دهیم؛ سپس با استفاده از داده‌های آزمایش دقت آن را در تشخیص ملورهای جدید ارزیابی می‌کنیم، نتایج به‌دست آمده (درصد میزان درستی تشخیص سیستم) برای داده‌های آزمایشی در هر روش نشان می‌دهد مدل (سیستم)های به‌دست آمده برای تشخیص مدل‌های جدید قابل اعتماد خواهند بود.

نتایج ارزیابی الگوریتم تشخیص نویز رأی گیری با توافق بالا

صحت انتخاب پنج الگوریتم دسته‌بند در جدول ۳ مشاهده می‌شود. نتیجهٔ صحت دسته‌بندی به‌کمک رویکرد رأی‌گیری با توافق سطح بالاتر بهتر از دیتاستی است که نویزهای کلاس درون آن، توسط رویکرد رأی‌گیری با رأی اکثریت تشخیص داده شده و حذف شدند. در شکل ۲ دیاگرام مقایسهٔ الگوریتم‌های دسته‌بند ترکیبی مشاهده می‌شود که صحت الگوریتم‌های تشخیص نویز مورد ارزیابی قرار گرفته است. در بهترین حالت، صحت ۹۸ درصد با الگوریتم پیشنهادی رأی‌گیری با توافق سطح بالا به‌دست آمده است.

جدول ۳. نتایج صحت ارزیابی پنج دسته‌بند الگوریتم رأی‌گیری توافق با سطع بالا

SVM	Random forest	Decision tree	Naïve Bayes	KNN	name
۵۹۳	٪۶۵	٪۷۹	٪۶۸	٪۷۵	Colon
٪۸۳	٪۶۴	٪۷۲	٪۶۶	٪۷۱	Colon noisy
٪۹۸	٪۷۰	٪۸۷	٪۷۲	٪۸۵	Colon noise removed



شکل ۲. ارزیابی الگوریتم‌های دسته‌بند

در مرحلهٔ ارزیابی عملکرد دو رویکرد رأی‌گیری با توافق بالا و رأی‌گیری با رأی اکثریت، از دو معیار صحت دسته‌بندی و انحراف معیار استفاده می‌کنیم. بدین صورت که نمونه‌های باقی مانده از دیتاست کلون بعد از تشخیص نمونه‌های نویزی آن توسط رویکرد رأی‌گیری با توافق بالا به‌عنوان یک دیتاست ذخیره می‌شود. سپس صحت و انحراف معیار دسته‌بندی دیتاست به‌دست آمده از طریق رویکرد رأی‌گیری با توافق بالا را به‌کمک الگوریتم پایهٔ درخت تصمیم‌گیری اندازه‌گیری می‌کنیم. در مرحلهٔ آخر، نمونه‌های باقی‌مانده بعد از حذف نمونه‌های نویزی به‌کمک روش رأی‌گیری با رأی اکثریت در دیک دیتاست جداگانه ذخیره می‌شوند. مقدار صحت و انحراف معیار برای این دیتاست نیز به‌کمک الگوریتم پایهٔ درخت تصمیم‌گیری اندازه‌گیری می‌شود. با توجه به جدول ۴، نتایج تجربی نشان داده است که صحت به‌دست آمده برای دیتاستی که به‌کمک طرح رأی‌گیری با توافق بالا به‌دست آمده است بیشتر از صحت دیتاستی است که به‌کمک طرح رأی‌گیری با رأی اکثریت به‌دست آمده است. انحراف از معیار دیتاستی که به‌کمک رویکرد رأی‌گیری با توافق بالا به‌دست آمده نیز، پایین‌تر از انحراف معیار دیتاستی است که به‌کمک طرح رأی‌گیری با رأی اکثریت تولید شده است.

شایان ذکر است که هرچه مقدار انحراف از معیار پایین‌تر باشد، نمونه‌های نویزی آن دیتاست بهتر دسته‌بندی شده است.

جدول ٤. صحت ارزیابی الگوریتم با رأی‌گیری با توافق بالا

روش	صحت	انحراف معیار
دیتاستی که به‌کمک طرح رأی‌گیری با توافق بالا به‌دست می‌آید.	٪۸۵	۰/۱۲
دیتاستی که به‌کمک طرح رأی‌گیری با رأی اکثریت به‌دست می‌آید.	٪۸۱	۰/۱۶

ارزیابی نتیجهٔ پیش‌بینی تفکیک دو نوع بیماری التهاب روده (کرون و کولیت)

در کاربردهای تشخیص مانند کاربردهای پزشکی، وقتی از طبقه‌بندهای آماری استفاده می‌شود، مشکل نامتوازن شدن کلاس‌ها به‌وجود می‌آید. این مشکل از احتمالات اولیهٔ زیادی که بین کلاس‌ها وجود دارد نشئت می‌گیرد و موجب کارایی ضعیف طبقه‌بندی می‌شود (اولایی، یعقوبی و یعقوبی، ۲۰۱۶). به همین دلیل از آزمون‌های خاصی که بر اساس ماتریس تلفیق[1] هستند، استفاده می‌شود. یکی از انواع این آزمون‌های خاص، آزمون ارزیابی نمودار مشخصه‌های عامل گیرنده است که برای بیان میزان درستی تشخیص از معیارهای خاص استفاده می‌کند. برای اندازه‌گیری میزان طبقه‌بندی از پارامتر AUC که همان سطح زیرمنحنیROC است استفاده می‌شود. این پارامتر یک عدد اسکالر برای مقایسهٔ طبقه‌بندی‌کننده‌های مختلف ارائه می‌دهد. از آنجا که AUC قسمتی از یک مربع واحد است، مقدار آن عددی بین صفر تا ۱ در نظر گرفته می‌شود. زمانی که طبقه‌بندی به‌صورت تصادفی انجام شود، مقدار AUC آن کمتر از ۰/۵ به‌دست می‌آید. مقدار AUC هر چه به ۱ نزدیک‌تر باشد، عملکرد بهتر طبقه‌بندی را نشان می‌دهد. در این پژوهش بیماری کولیت با کد ۱ و بیماری کرون با کد ۰ نمایش داده می‌شود. به‌منظور ارزیابی تئوری دسته‌بندی ترکیبی با رأی‌گیری توافق سطح بالا در تحلیل تفکیک دو نوع بیماری التهاب روده، از منحنی ROC استفاده شده است و روش از نظر صحت و اعتبار نتایج به‌دست آمده از طریق نمودار ROC مشخص می‌شود. نتایج را می‌توان به‌صورت زیر بیان کرد:

TP: تشخیص درست نوع التهاب روده (کولیت یا کرون) در بیماران با استفاده از نتایج آزمون.

FP: تشخیص غلط نوع التهاب روده (کولیت یا کرون) در بیماران با استفاده از نتایج آزمون.

TN: تشخیص درست فقدان بیماری التهاب روده (کولیت یا کرون) در بیماران با استفاده از نتایج آزمون.

FN: تشخیص‌های اشتباه فقدان التهاب روده (کولیت یا کرون) در بیماران با استفاده از نتایج آزمون.

N: تعداد کل بیماران

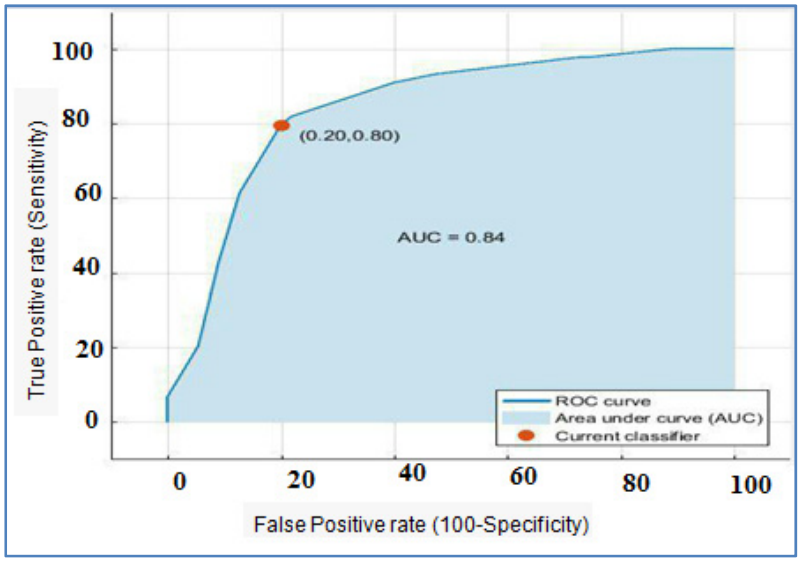
1. Confujen

ما می توانیم نتایج را به‌صورت کسر نیز ارائه کنیم. در این صورت داریم:

$$FNF + TPF = 1$$ رابطهٔ ۱)

به بیان دیگر، FNF کسر FN (تشخیص‌های غلط منفی) را به ما می دهد، همچنین می‌توانیم TPF را به‌صورت کسر TP (تشخیص‌های درست مثبت) حساب کنیم. به‌طور مشابه کسر FP (تشخیص‌های غلط مثبت) وکسر TN (تشخیص‌های درست منفی) به موارد بالا اضافه می‌شوند. بیمارانی که واقعاً کولیت یا کرون ندارند (در مثال ما) یا باید TN یا FP تشخیص داده شوند.

شکل ۳ نمودار مد نظر را که در محیط MATLAB با استفاده از کدنویسی انجام گرفته است، نشان می‌دهد. در نهایت سطح زیرمنحنی AUC که نشان‌دهندهٔ دقت مدل است، به‌دست خواهد آمد. هرچه سطح زیرمنحنی بیشتر باشد، دقت مدل به‌دست آمده بیشتر است. در این پژوهش نمودار مد نظر برای روش دسته‌بند ترکیبی رأی‌گیری با توافق سطح بالا با صحت ۹۸ درصد محاسبه شده است.



شکل ۳. تحلیل نمودار ROC

مفاهیمی که در محورهای نمودار وجود دارد، عبارت‌اند از:

Sensitivity یا حساسیت عبارت است از این که چقدر آزمایش مد نظر ما توانسته است بیماران (کولیت یا کرون) را انتخاب کند؛ یعنی TP/P است. به بیان دیگر حساسیت تعداد بیماران درست انتخاب شده توسط آزمون به نسبت کل بیمارانی که واقعاً مبتلا به کولیت یا کرون هستند را نشان می‌دهد.

Specificity: عبارتست از توانایی آزمایش برای انتخاب درست بیمارانی که کولیت یا کرون دارند.

در شکل ۳ برای ارزیابی الگوریتم پیشنهادی از این پارامتر استفاده شده است. همان‌طور که مشاهده می‌شود مقدار AUC = ۰/۸۴ است. نتایج به‌دست آمده نشان از عملکرد بهتر الگوریتم رأی‌گیری با توافق بالا است.

نتیجه‌گیری

به‌دلیل تأثیر پارامترهای مختلف در تشخیص بیماری‌ها، در این پژوهش از قابلیت سیستم رأی‌گیری با توافق سطح بالا برای پیش‌بینی با استفاده از داده‌های بدون نویز به‌دست آمده از روش‌های داده‌کاوی به‌منظور ارزیابی و تشخیص نویز استفاده شد. در مدل رأی‌گیری با توافق سطح بالا با یک پروسهٔ آموزش عمل تخمین روی داده‌های جمع‌آوری شده از ۱۰۰ نمونه بیمار مبتلا به التهاب روده انجام گرفته است. تشخیص و حذف نویز کلاس به‌کمک رویکرد رأی‌گیری با آرای بالا، نتایج بهتری با توجه به خطای به‌دست آمدهٔ قبل از تشخیص نویز و بعد از مرحلهٔ دوم داده است و در نهایت مدل پیشنهادی (رأی‌گیری با توافق سطح بالا) بهترین نتایج را در سه روش با توجه به داده‌های بدون نویز با الگوریتم‌های ذکر شده به‌دست آمده است و طبق ریسک بالای خطا در تشخیص این دو بیماری الگوریتم‌های مورد استفاده در این پژوهش کمک زیادی به تشخیص و تفکیک دو نوع بیماری کرده‌اند.

پیشنهادها

در ادامهٔ این پژوهش و برای عملکرد بهتر روش، مدل سیستم فازی نوع دوم یا سیستم فازی مدل ایشوبوشی برای ارزیابی و پیش‌بینی کلاس‌بندی با در نظر گرفتن درجهٔ اطمینان کلاس‌ها در روش آنلاین برای مدیریت دقیق‌تر عدم قطعیت و ابهام در تشخیص دو نوع بیماری التهاب روده (کولیت و کرون) پیشنهاد می‌شود. همچنین استفاده از طرح رویکرد رأی‌گیری با آرای بالای ناهمگون برای تشخیص نویز کلاس و حذف نمونه‌های نویزی برای خلق دیتاست سالم

که در مقاله‌ای توسط اسلوبان و لاوراک بررسی شده است را به‌عنوان کارهای آینده پیشنهاد می‌کنیم.

فهرست منابع

علی مردانی، س.؛ آقایی، ع. (۱۳۹۴). ارائهٔ روش نظارتی برای نظرکاوی در زبان فارسی با استفاده از لغت‌نامه و الگوریتم SVM. *مدیریت فناوری اطلاعات*، ۷(۲)، ۳۶۲– ۳۴۵.

References

Ahmed, S. S., Dey, N., Ashour, A. S., Sifaki-Pistolla, D., Bălas-Timar, D., Balas, V. E., & Tavares, J. M. R. (2017). Effect of fuzzy partitioning in Crohn's disease classification: a neuro-fuzzy-based approach. *Medical & biological engineering & computing,* 55(1), 101-115.

Alimardani, S., Aghaie, A. (2015). Opinion Mining in Persian Languageusing svm algorithm *Journal of Information Technology Management*, 7(2), 345-362. (*in Persian*)

Bashir, S., Qamar, U., Khan, F. H., & Naseem, L. (2016). HMV: A medical decision support framework using multi-layer classifiers for disease prediction. *Journal of Computational Science*, 13, 10-25.

Borut Sluban, A. & NadaLavrač, N. (2015). Relating ensemble diversity and performance: A study in class noise detection. *Neurocomputing*, 160, 120–131.

Catal, C., Alan, O. & Balkan, K. (2011). Class noise detection based on software metrics and ROC curves. *Information Sciences*, 181(21), 4867-4877.

Cooper, J.G., Purcell, G.P. (2006). Data Mining for Correlations between Diet and Crohn's Disease Activity. *AMIA Symposium Proceedings*, Page – 897.

Guan, D., Yuan, W., & Shen, L. (2013, July). Class noise detection by multiple voting. IEEE. In Natural Computation (ICNC). *Ninth International Conference on.* pp. 906-911.

Guan, D., Yuan, W., Ma, T., & Lee, S. (2014). Detecting potential labeling errors for bioinformatics by multiple voting. *Knowledge-Based Systems*, 66, 28-35.

Kaladhar, D. S. V. G. K., Pottumuthu, B. K., Rao, P. V. N., Vadlamudi, V., Chaitanya, A. K., & Reddy, R. H. (1926). The Elements of Statistical Learning in Colon Cancer Datasets: Data Mining, Inference and Prediction. *Algorithms Research*, 2(1), 8-17.

Mossotto, E., Ashton, J.J., Coelho, T., Beattie, R.M., MacArthur, B.D., Ennis, S. (2017). *Classification of Paediatric Inflammatory Bowel Disease using Machine Learning*, 2017 May 25. doi: 10.1038/s41598-017-02606-2.

Olyaee, M. H., Yaghoubi, A., & Yaghoobi, M. (2016). Predicting protein structural classes based on complex networks and recurrence analysis. *Journal of Theoretical Biology,* 404, 375-382.

Sluban, B., & Lavrač, N. (2015). Relating ensemble diversity and performance: a study in class noise detection. *Neurocomputing*, 160, 120-131.

Thompson, V. L. S., Lander, S., Xu, S., & Shyu, C. R. (2014). Identifying key variables in African American adherence to colorectal cancer screening: the application of data mining. *BMC public health*, 14(1), 1173.

Uğuz, H. (2011). Adaptive neuro-fuzzy inference system for diagnosis of the heart valve diseases using wavelet transform with entropy. *Neural Computing and Applications*, 21 (7), 1617-1628.