استفاده همزمان از همبستگی‌خطی پیرسون و ترکیب الگوریتم‌های داده‌کاوی به منظور بهبود پیش‌بینی نوع تومور در بیماران سرطانی

محسن‌غلامی[۱]. سیدجوادمیرعابدینی[۲*]

۱- دانشجوی کارشناسی ارشد دانشگاه آزاد اسلامی واحد بوشهر. پست الکترونیکی: Mohsen.gholami18@yahoo.com
۲- عضو هیات علمی دانشگاه آزاد اسلامی واحد تهران مرکزی. پست الکترونیکی: j_mirabedini@iauctb.ac.ir
تاریخ دریافت: ۹۷/۱۱/۱۷ تاریخ پذیرش:۹۸/۵/۱

چکیده

امروزه سرطان سینه از شایع‌ترین بیماری‌های سرطان در بین زنان به‌شمار می‌آید. آمارها از رشد شش درصدی این نوع سرطان در ایران حکایت می‌کند که نشان دهنده جدی بودن خطر آن می‌باشد. این در صورتی است که در صورت پیشگیری و یا تشخیص زود هنگام بیماری می‌توان تا حد زیادی از خطرات آن جلوگیری نمود. با پیشرفت علوم پزشکی، زمینه لازم جهت ایجاد سیستم‌هایی با قابلیت پیشگیری، پیش‌بینی و درمان بیماران با استفاده از فناوری‌های جدید حاصل گردیده است. داده‌کاوی پزشکی سعی در مدل‌سازی و کشف روابط بین عوامل خطرساز جهت پیش‌بینی وضعیت بیماران آینده با کمک از داده‌های دردست دارد. در این پژوهش سعی گردیده تا با مقایسه الگوریتم‌های مختلف داده‌کاوی و ترکیب این الگوریتم‌ها، روشی جدید، کارا و با دقت بالا و قابلیت پیاده‌سازی بر روی داده‌های محلی ایجاد گردد. در نهایت روش پیشنهادی که به بهبود کارایی الگوریتم بیز ساده با استفاده از الگوریتم آدابوست می‌پردازد، توانایی پیش‌بینی نوع تومور خوش‌خیم یا بدخیم با دقت۹۶.۶۷ درصد را دارا می‌باشد. داده‌های لازم جهت این فرآیند از سایت**UCI** جهت تشخیص نوع تومور با۵۶۹ رکورد و۳۲ متغیر، استخراج گردیده است.

کلمات کلیدی: ضریب همبستگی پیرسون، الگوریتم‌های دسته‌بندی، بیزساده، آدابوست

۱- مقدمه :

سرطان به رشد غیر طبیعی سلول‌ها در بدن که به سرعت گسترش می‌یابد گفته می‌شود. طبق تحقیقات صورت گرفته عواملی همچون وراثت، موادشیمیایی و مصرف دخانیات در بروز آن تاثیر گذار هستند[۱].

باتوجه به افزایش سریع روند گسترش بیماری سرطان سینه در زنان و همچنین پایین آمدن سن مبتلایان به این بیماری، لـــزوم ایجاد سیستم‌هایی جهت تشخیص و پیش‌بینی این بیماری با توجه به اینکه شناسایی زودهنگام عوامل خطرساز و پیش‌بینی زود هنگام بیماری بسیار در روند بهبود بیماران تاثیرگذار است به شدت احساس می‌گردد چرا که سرطان سینه در مقایسه با سایر سرطان‌ها یکی از عوامل عمده مرگ و میر در میان زنان به‌شمار می‌رود. [۲] از سویی امروزه افزایش هزینه‌ها جهت آزمایشات متعدد و درمان نیز با توجه به گسترش روز افزون بیماری وجود دارد[۳].

از طرف دیگر امکانات پزشکی و تعدد آزمایشات و پزشکان متخصص در همه مناطق جهت تشخیص بیماری وجود ندارد [٤] و داده کاوی پزشکی می‌تواند به کشف روابط پنهان میان متغیرها و ایجاد مدلی توانمند و کارا به سرعت بخشیدن در پیش‌بینی عوامل خطرساز و پیش‌بینی احتمال عود بیماری به صورت سریع همراه با کاهش هزینه ها از طریق کاهش آزمایشات متعدد در تشخیص و پیش‌بینی اولیه بیماری، از طریق تحلیل داده‌های محلی کمک نماید.

طبق تحقیقات صورت گرفته در کشورهای اروپایی سن ابتلا به سرطان سینه در زنان بین سنین ۵۳ الی ۵۷ سال است[۵]. این در صورتی است که میانگین سن ابتلا به این بیماری در ایران تقریباً ٤٥ سال می‌باشد. از این رو ضرورت ایجاد سیستم‌هایی جهت تشخیص عوامل خطرساز و پیش‌بینی زود هنگام بیماری در مراحل اولیه بیماری امری ضروری به نظر می رسد[٦].

ایجاد سیستم‌های ساده و موثر جهت کاهش تعداد متغیرها و افزایش دقت و کارایی الگوریتم‌ها یکی از شاخص‌های حیاتی برای تشخیص و پیش‌بینی زود هنگام و کاهش هزینه های آزمایشات متعدد، یکی از اهداف اصلی داده‌کاوی پزشکی محسوب می‌گردد[۷]. حتی این سیستم با استفاده از داده های مرتبط، می تواند در تشخیص نوع تومور مثل خوش‌خیم بودن و یا بدخیم بودن آن به پزشکان متخصص کمک کند.

در این مقاله سعی شده ابتدا الگوریتم‌های مختلف از نظر دقت و کارایی با هم مقایسه گردیده و سپس با روش‌های ترکیبی کارایی این الگوریتم‌ها بهبود یافته و در نهایت روشی کارا و ساده ایجاد می‌گردد. روش پیشنهادی علاوه بر کاهش تعداد متغیرها، باعث افزایش دقت پیش‌بینی در مقایسه با الگوریتم اولیه گردیده و عملیاتی می باشد. از این روش نیز می توان به تشخیص نوع تومور از نظر خوش‌خیمی و بدخیمی به متخصصین یاری رساند.

۲- ضرورت انجام تحقیق:

امروزه سرطان یکی از علل اصلی مرگ و میر در جهان به‌شمار می آید و تشخیص و پیش‌بینی زود هنگام ، در مرحله رشد بیماری و تشخیص نوع آن یکی از چالش‌های جدی در علوم پزشکی است. که می‌توان با پیش‌بینی و تشخیص زود هنگام و به موقع، آن را تا حد زیادی کنترل و از بروز آن جلوگیری کرد[۸].

سرطان سینه به عنوان شایع‌ترین سرطان در بین زنان می‌باشد و روند رو به رشد آن در سال‌های اخیر بسیار زیاد و نگران کننده بوده است. پیش‌بینی احتمال بروز این بیماری و یا تشخیص نوع توده و یا احتمال عود مجدد، می تواند تا حد زیادی در بهبود و یا کنترل بیماری تاثیرگذار باشد[۹].

کاهش سن مبتلایان به سرطان سینه در کشور که به میانگین سنی ٤٥ سال رسیده است در صورتی که در کشورهای غربی این نوع از سرطان بین زنان۵۳ الی۵۷ سال شایع است.

در تحقیقات صورت گرفته در خصوص علل مرگ و میر و روند تغییرات آن در سال‌های۱۳۵۸ تا ۱۳۸۰ در ایران که توسط دکتر پروین یاری و همکاران صورت پذیرفت نتایج حاصله در بازه ۲۳ ساله نشان می داد مرگ به علت بیماری‌های واگیردار در کشور کاهش یافته ولی متاسفانه بیماری‌های غیر واگیر مثل بیماری‌های قلبی و سرطان سیر صعودی داشته‌اند.

از طرفی ایجاد مدل‌های پیش‌بینی و تشخیص می‌توانند به تسهیل روند مدیریت و شناسایی بیماری کمک نمایند[۱۰].

۳- مفاهیم اولیه

۳-۱ ضریب همبستگی پیرسون

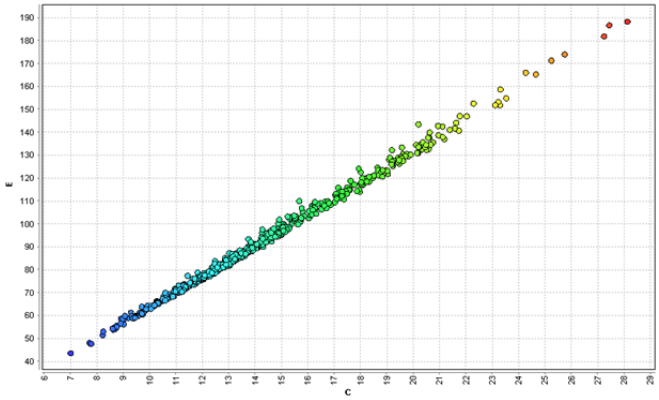
با استفاده از این تکنیک آماری می توان میزان همبستگی خطی بین دو متغیر تصادفی را شناسایی نمود. این ضریب مقداری بین ۱- و ۱ دارد. همبستگی کامل زمانی رخ می دهد که ضریب ما ۱ باشد و زمانی که ضریب ۰ باشد یعنی همبستگی بین متغیرها وجود ندارد. ضریب ۱- نیز نشان دهنده نسبت منفی کامل بین دو مقدار است[۱۱]. در واقع ما با استفاده از این تکنیک میزان ارتباط خطی بین متغیرها را شناسایی می کنیم که از فرمول زیر قابل محاسبه است:

$$\mathrm{P}X,Y = \frac{COV(X,Y)}{\sigma_X\,\sigma_Y} = \frac{E[(X-\mu X)(Y-\mu Y)]}{\sigma_X\,Q_Y}$$

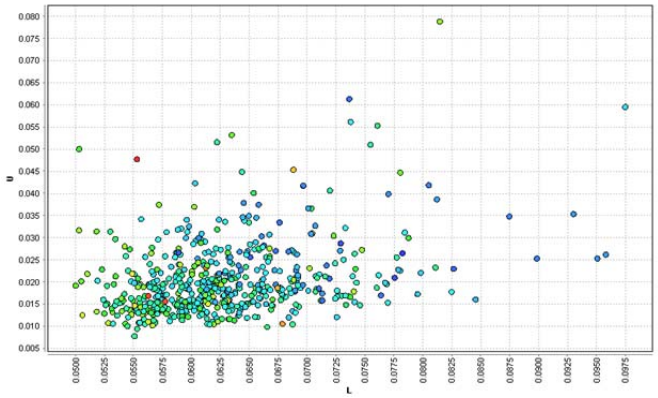
فرمول شماره (۱) ضریب همبستگی جامعه

$$r = \frac{\sum_i^n = 1\,(X_{i\_}\overline{X})\,(Y_i - \overline{Y})}{\sqrt{\sum_i^n = 1(\,X_i - \overline{X}\,)^2}\;\sqrt{\sum_i^n = 1(\,Y_i - \overline{Y}\,)^2}}$$

فرمول شماره(۲) ضریب همبستگی نمونه آماری پیرسون



شکل(۱) پراکندگی قرارگیری داده‌ها بر روی محور مختصات که نشان می‌دهد همبستگی خطی زیادی بین دو متغیر E و C وجود دارد.



شکل(۲) نشان می‌دهد با توجه به قرارگیری داده‌ها به صورت دایره‌ای همبستگی خطی کمی بین دو متغیر L و U وجود دارد(نزدیک به صفر)

۳-۲ واریانس و کواریانس:

واریانس نحوه پراکندگی داده‌ها در اطراف میانگین همان داده‌ها را مشخص می‌کند همچنین کواریانس شاخصی است برای تغییرات یک متغیر با متغیر دیگر است[۱۲].

۳-۳ الگوریتم بیز ساده

در یادگیری ماشین معمولاً در فضای فرضیه H بدنبال بهترین فرضیه‌ای هستیم که درمورد داده‌های آموزشی D صدق کند. یک راه تعیین بهترین فرضیه، این است که بدنبال محتمل ترین فرضیه‌ای باشیم که با داشتن داده‌های آموزشی D و احتمال قبلی در مورد فرضیه‌های مختلف می‌توان انتظار داشت تئوری بیز چنین راه‌حلی را ارائه می‌دهد. این روش راه حل مستقیمی است که نیازی به جستجو ندارد[۱۳].

سنگ‌بنای یادگیری بیزی را تئوری بیز تشکیل می‌دهد. این تئوری امکان محاسبه احتمال ثانویه را بر مبنای احتمالات اولیه می‌دهد[۱٤].

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

فرمول شماره (۳) قضیه بیز

همان‌طور که مشاهده می‌شود با **افزایش** P(D) مقدار P(h□D) کاهش می‌یابد. زیرا هر چه احتمال مشاهده D مستقل از h بیشتر باشد به این معنا خواهد بود که D شواهد کمتری در حمایت از h در بر دارد. **[۱۵] [۱٦]**

قضیه بیز:

P(h)= احتمال اولیه‌ای که فرضیه h قبل از مشاهده مثال آموزشی D داشته است[۱٦].

P(D)= احتمال اولیه‌ای که داده آموزشی D مشاهده خواهد شد[۱٦].

P(D|h)= احتمال مشاهده داده آموزشی D به فرض آنکه فرضیه h صادق باشد[۱٦].

۳-٤ الگوریتم آدابوست[1]

آدابوست مخفف بوستینگ تطبیقی بوده و یک الگوریتم یادگیری ماشین است که توسط یاو فروند و رابرت شاپیر ابداع شد[۱۷]. در واقع آدابوست یک متا الگوریتم است که بمظور ارتقاء عملکرد، و رفع مشکل رده‌های نامتوزان[۱۸] همراه دیگر الگوریتم‌های یادگیری استفاده می‌شود. در این الگوریتم، طبقه‌بند هر مرحله جدید به نفع نمونه‌های غلط طبقه‌بندی شده در مراحل قبل تنظیم می‌گردد[۱۹]. آدابوست نسبت به داده‌های نویزی و پرت حساس است؛ ولی نسبت به مشکل بیش‌برازش از بیشتر الگوریتم‌های یادگیری برتری

[1]- Adaboost

دارد. طبقه‌بند پایه که در اینجا استفاده می‌شود فقط کافیست از طبقه بند تصادفی(۵۰٪) بهتر باشد و به این ترتیب بهبود عملکرد الگوریتم با تکرارهای بیشتر بهبود می‌یابد[۲۰].

۳-۵ جنگل تصادفی[1]

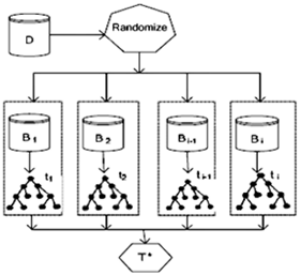
جنگل تصادفی مجموعه‌ای از درختان تصمیم است که بصورت موازی با هم در ارتباط اند و بر پایه تئوری بگینگ تشکیل شده است[۲۰] [۲۱]. این روش دسته بندی یادگیری با ناظر است که در آن درخت‌های تصمیم صفات خاصه خود را بطور تصادفی از داده‌های کاملاً مستقل انتخاب و سپس دسته‌بندی می‌کند.

جنگل تصادفی را می‌توان با دو روشForest-RI وForest-RC ایجاد کرد[۲۲]. در روش Forest-RI که بر پایه روش bagging است جنگل تصادفی بصورت تصادفی و پی در پی صفات خاصه را انتخاب و در نهایت دسته‌بندی می‌کند[۲۳]. برای ساخت و رشد درختان در جنگل تصادفی می‌توان از روش درخت تصمیم CART استفاده کرد.

در روش Forest-Rc ابتدا با استفاده از ترکیب تعدادی از صفات خاصه و تعیین ضرائب در محدوده (1 ,1-) صفات خاصه را تصادفی انتخاب و در نهایت بهترین صفت را گسترش می‌دهد. از معایب این روش، محدودیت تعداد صفات خاصه است. در صورتی‌که تعداد صفات خاصه زیاد شود باعث کاهش همبستگی میان دسته‌بندها می‌شود[۲٤].

جنگل تصادفی امکان هرس‌کردن درختان در صورت بزرگ شدن را ندارد و به همین دلیل نرخ خطا برای جنگل تعریف می شود. صحت و درستی یک جنگل تصادفی به قدرت و توانایی هر یک از دسته‌بندها و وابستگی میان آن‌ها بستگی دارد.

از مزایای جنگل تصادفی می‌توان به کارایی و دقت بالا نسبت به روش هایBagging وBoosting اشاره کرد. این الگوریتم بر روی داده‌های بزرگ با ویژگی‌های زیاد بدون حذف متغییر نتایج خوبی را ارائه می‌دهد[۲۵]. همچنین در برخورد با داده‌های پرت و نامتوازن بهتر از سایر الگوریتم‌ها جهت دسته‌بندی عمل می‌کند[۲٦].



شکل۳) نحوه عملکرد الگوریتم جنگل تصادفی[۲۷].

روش کار این الگوریتم به صورتی است که ابتدا k نمونه با استفاده از الگوریتمBagging به صورت تصادفی انتخاب می‌شود. در مرحله بعد، از k نمونه‌ها برای k درخت و k دسته‌بند استفاده می‌شود. در نهایت از بین دسته‌بندی کننده‌ها برای دسته‌بندی بهینه رای گیری صورت می‌گیرد[۲۸].

1_ Random Forest

۳-٦ درخت تصمیم[3]

الگوریتم درخت تصمیم، یک الگوریتم بازگشتی دارای ساختار درختی است که از ریشه به سمت برگ حرکت می‌کند. این الگوریتم یکی از روش‌های طبقه‌بندی داده بر پایه یادگیری با نظارت است. گره‌های برگ حاوی اطلاعات کلاس‌ها(متغیر وابسته) و گره‌های غیر برگ صفات خاصه (متغیر مستقل) می‌باشد. روش کار الگوریتم بصورت سلسله مراتبی است که بر اساس داده‌های آموزشی با انتخاب یکی از صفات خاصه در هر مرحله شروع به کار می‌کند. در ادامه با تقسیم‌بندی هر یک از صفات ادامه می‌دهد تا زمانیکه تمام داده‌ها به اطلاعات دارای برچسب واحد کلاس شوند. این روش ساده و به راحتی قابل درک و تفسیر بوده و امکان هرس کردن تصمیم هایی که قابل تعمیم نیستند را دارد. الگوریتم های درخت تصمیم به صورت چند مرحله ای عمل می‌کنند یعنی تصمیمات پیچیده به تصمیمات ساده تری تقسیم و ترکیب این تقسیمات ساده رسیدن به تصمیمات مورد انتظار است[۲۹].

معیارهای مختلفی جهت انتخاب صفات خاصه در این الگوریتم وجود دارد که می‌توان به موارد زیر اشاره نمود:

Information Gain : این معیار مشخص کننده این است که یک ویژگی تا چه مقدار می‌تواند مثال‌های آموزشی را بر اساس دسته بندی تفکیک کند[٢٤] .این معیار در ID3 برای انتخاب صفات خاصه استفاده می‌شود .

$$IG(A)=Entropy\ (D)\ -\ Entropy\ A(D)$$

فرمول٤) محاسبه بهره اطلاعاتی

$$Entropy(D)=-\sum_{i=1}^{c} pi \times \log2(pi)$$

فرمول٥)محاسبه آنتروپی

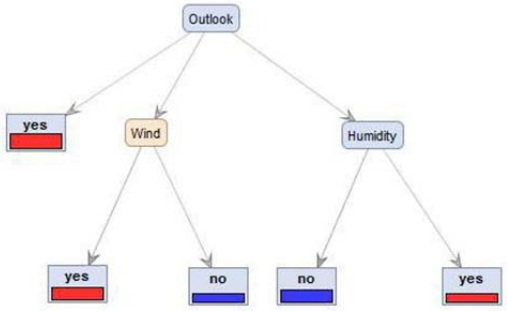
این فرمول‌ها برای محاسبه مفهوم آنتروپی که در واقع میزان خلوص (بی نظمی یا عدم خالص بودن) مجموعه‌ای از مثال‌ها را مشخص می‌کند[۲۰] بکار برده می‌شود. در توضیح فرمول شماره ۱ و ۲ بهتر است بگوییم:

- C معرف تعداد برچسب‌های کلاس موجود در داده‌ها.
- Pi معرف احتمال اینکه نمونه‌ای از داده‌ها متعلف به کلاس i ام باشد.
- Vتعداد اعضای دامنه صفت خاصه A .
- Dj قسمتی از داده اولیه که مقدار صفت خاصه آن‌ها Vj باشد.

Gain Ratio: این معیار در واقع برای نرمال‌سازی Information Gain استفاده می‌شود. این روش صفات خاصه دارای دامنه زیاد را به صفات خاصه با دامنه کم ترجیح می‌دهد که دارای دقت بالا و همچین کاهش پیچیدگی نسبت به عملکرد Information Gain است.

Gini Index : این معیار تمام صفات خاصه را بصورت دودویی دسته بندی می‌کند سپس مقدارGini را محاسبه کرده و کمترین مقدار صفت خاصه را انتخاب می‌کند. این معیار برای انتخاب گره در الگوریتم CART مورد استفاده قرار می‌گیرد[۳۰].

3- Decision Tree



شکل٤) نمونه ای از رشد درخت از ریشه به برگ

در توضیح شکل ۱ ، Outlook گره ریشه و wind و Humidity گره‌ها و مقادیر yes و no برگ‌ها هستند. به طور کلی از مزایای درخت تصمیم می‌توان به امکان استفاده برای انواع داده‌های پیوسته و گسسته ، استخراج قوانین قابل فهم و آسان ، آماده‌سازی آسان داده‌ها ، سازگارکردن داده های فاقد مقدار، شناسایی تفاوت میان زیرگروه‌ها، دسته‌بندی ویژگی‌ها با تاثیرات زیاد و محاسبات کم برای دسته‌بندی داده‌ها و تفسیر آسان درخت اشاره کرد.

معایب این الگوریتم، مصرف بالای حافظه ، سخت و زمانبر بودن پیاده‌سازی برای مجموعه داده‌های بزرگ ، هزینه بالا برای هرس کردن ، زیاد بودن تعداد گره های پایانی در صورت همپوشانی، انباشته شدن خطای لایه‌ها بر روی یکدیگر در صورت بزرگ شدن درخت است.

۳-۷ K نزدیکترین همسایه[1]:

این الگوریتم برای اولین بار در سال ۱۹۵۰ معرفی شد. اما به دلیل سرعت پایین محاسبات کامپیوترها و نیاز به حافظه بالا در آن زمان، تا سال ۱۹۶۰مورد توجه قرار نگرفت[۳۰]. یادگیری خود را بر اساس میزان تشابه و فاصله انجام می‌دهد. از این الگوریتم در تشخیص الگو و پیش‌بینی بصورت گسترده استفاده می‌شود.

در این روش دسته‌بندی داده‌ها بصورت بردار یا نقاطی در فضای چند بعدی ویژگی‌ها به گونه‌ای که مفهومی به نام فاصله بوجود می‌آید صورت می‌پذیرد. بعد از آن هریک از داده‌ها نرمال‌سازی شده چراکه به داده‌های نویز حساس و در صورت وجود بر روی صحت پاسخگویی اثر می‌گذارد . در این الگوریتم می‌توان از فاصله اقلیدوسی یا فاصله منهتن برای سنجش فاصله میان صفات استفاده کرد. اگر بطور کلی هر یک از مجموعه داده‌ها را در دو کلاس مثبت و منفی برچسب بزنیم ، این الگوریتم با در نظر گرفتن k تعداد دلخواه همسایه را برای تعیین و انتخاب کلاس‌ها برای دسته‌بندی انتخاب می‌نماید که پیشتر توسط کاربر انتخاب می‌شود. پیچیدگی ذخیره‌سازی و محاسباتی این الگوریتم $O(n)$ است [۱۱].

این الگوریتم دارای دقت خوبی است . سرعت انجام محاسبه آن پایین است. یکی از مشکلات این روش این است که قابلیت تعمیم ندارد (برخلاف روش رگرسیون) و با اضافه کردن داده جدید می بایست دوباره مدل سازی صورت گیرد . این الگوریتم را Lazy یا تنبل نیز می‌نامند چرا که تا داده آزمایشی وارد نشود مدل سازی انجام نمی‌گردد.

برای محاسبه فاصله می‌توان از معیارهای متفاوتی استفاده کرد که یکی از این معیارها فاصله اقلیدسی است. فاصله دو نقطه p و q اندازه پاره‌خطی است که آنها را به هم متصل می‌کند که می توان به صورت زیر تعریف کرد.

1-K Nearest Neighbor (KNN)

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\cdots+(p_n-q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i-q_i)}$$

فرمول٦) محاسبه فاصله اقلیدسی

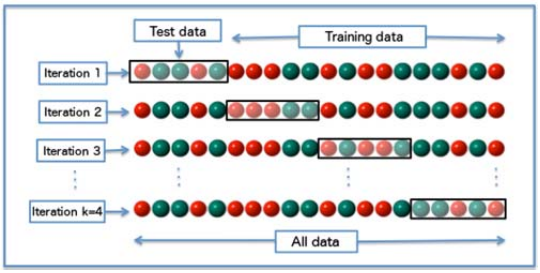
۳-۸ الگوریتم RIPPER

از این الگوریتم برای کشف قوانین به صورت مستقیم با استفاده از معیار بهره اطلاعاتی استفاده می شود. این الگوریتم شامل دو مرحله رشد و هرس کردن است. هرس کردن تا زمانیکه میزان خطا بیشتر از ۵۰ درصد باشد ادامه می یابد. به عبارت ساده تر قوانینی که تاثیری بر روی دقت مدل ندارد ضعیف شناخته شده و حذف می شود.

۳-۹ روش ارزیابی مدل با استفاده از K-Fold

در این پژوهش از روش K-Fold جهت ارزیابی مدل استفاده شده است. در این روش نوع اعتبارسنجی داده‌ها به K زیرمجموعه افراز می‌شوند. از این K زیرمجموعه، هر بار یکی برای اعتبارسنجی و K-1 تای دیگر برای آموزش بکار می‌روند. این روال K بار تکرار می‌شود و همه داده‌ها دقیقاً یکبار برای آموزش و یکبار برای اعتبارسنجی بکار می‌روند[۳۱]. در نهایت میانگین نتیجه این K بار اعتبارسنجی به عنوان یک تخمین نهایی برگزیده می‌شود. البته می‌توان از روش‌های دیگر برای ترکیب نتایج استفاده کرد. بطور معمول از ۱۰-Fold استفاده می‌شود[۳۲].

در روش K-Fold طبقه‌ای سعی می‌شود نسبت داده های هر کلاس در هر زیرمجموعه و در مجموعه اصلی یکسان باشد.



شکل۵) نحوه انتخاب مجموعه داده تست جهت ارزیابی در روش K-Fold[۳۳].

٤- داده‌ها

در این مقاله از مجموعه داده استاندارد سرطان سینه ویسکانسین از مخزن یادگیری UCI استفاده می‌گردد که حاوی ٥٦٩ پرونده و دارای ۳۲ متغیر است که جهت تشخیص نوع تومور استفاده می‌شود و متغیر هدف آن دارای دو نوع خوش خیم و بدخیم می‌باشد. لازم به توضیح است مقادیر ویژگی ها از تصاویر عکس‌برداری شده استخراج شده‌اند[٣٤].

جدول ۱ معرفی ویژگی‌های مجموعه داده[۳۵]

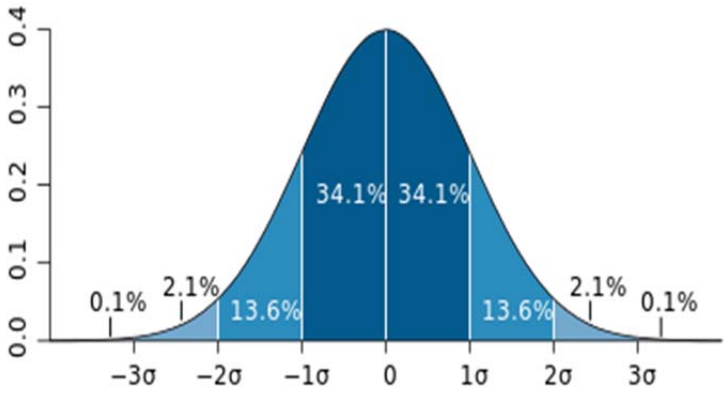
ردیف	نام متغیر	توضیحات
۱	ID	شماره پرونده
۲	diagnosis	متغیر هدف که شامل دو مقدار خوش خیم و بد خیم است
۳	radius	شعاع
٤	texture	بافت
٥	perimeter	محیط
٦	area	مساحت
۷	smoothness	همواری
۸	compactness	میزان غلظت
۹	concavity	میزان فرورفتگی
۱۰	concave_points	نقاط توخالی (تعداد مقادیر توخالی حد فاصل)
۱۱	symmetry	تقارن
۱۲	Fractal_dimension	ابعاد فراکتاکال
۱۳	radius_SE	شعاع SE
۱٤	texture_SE	بافت SE
۱٥	perimeter_SE	محیط SE
۱٦	area_SE	مساحت SE
۱۷	smoothness_SE	همواری SE
۱۸	compactness_SE	فشردگیSE
۱۹	concavity_SE	فرورفتگی SE
۲۰	concave_points_SE	نقاط مقعر SE
۲۱	symmetry_SE	تقارن SE
۲۲	fractal_dimension_SE	ابعاد فراکتاکال SE
۲۳	radius_Worst	بدترین شعاع
۲٤	texture_Worst	بدترین بافت
۲٥	perimeter_Worst	بدترین محیط
۲٦	area_Worst	بدترین مساحت
۲۷	smoothness_Worst	بدترین همواری
۲۸	compactness_Worst	بدترین فشردگی
۲۹	concavity_Worst	بدترین تورفتگی
۳۰	concave_points_Worst	نقاط تو رفتگی ارزشمند
۳۱	symmetry_Worst	بدترین تقارن
۳۲	fractal_dimension_Worst	بدترین بعد فراکتال

٤-١ پیش پردازش داده ها:

٤-١-١ بررسی مقادیر خارج از محدوده و گم شده

انحراف معیار اندازه‌گیری چگونگی پراکندگی یک ویژگی است که با کمک آن می‌توان داده‌های ناسازگار(با توجه به میزان فاصله) را مشخص نمود. انحراف به معنی‌دوری از مقدار نرمال است. انحراف معیار، عددی برای نشان دادن میزان گسترش اعداد است. علامت آن حرف یونانی سیگما σ است. با استفاده از انحراف معیار، ما یک راه استاندارد برای یافتن مقدار نرمال، مقدار بیش از نرمال و مقدار کمتر از نرمال در دست داریم.

یک قانون سرانگشتی خوب این است که معمولاً داده‌های بین $\bar{x} - 2\sigma$ و $\bar{x} + 2\sigma$ داده‌های باارزش‌تری محسوب شده و داده‌های خارج از این فاصله در نظر گرفته نمی‌شود که اصطلاحاً به داده‌های کم ارزش داده‌های خارج از محدوده[1] نیز می‌گویند[۳۲]. می‌توان با استفاده از نمودار ۳-۱ این قضیه را بهتر درک کرد که به آن اصطلاحاً نمودار توزیع نرمال داده‌های تصادفی یا نمودار محدوده‌های قابل نرمال گویند.



نمودار ۱: نمودار توزیع طبیعی[۱۱]

در توضیح نمودار ۱ باید بگوییم که قسمت آبی تیره در فاصله یک برابر انحراف معیار از میانگین توزیع قرار دارد و قسمت آبی روشن و آبی تیره به طور توام، در فاصله دو برابر انحراف معیار از میانگین توزیع قرار دارند. در توزیع طبیعی، اولی برابر با ٦٨٪ سطح زیر نمودار و دومی برابر با ۹۵٪ سطح زیر نمودار است.

معمولاً با افزایش تعداد داده‌ها توزیع آن‌ها به منحنی توزیع نرمال میل پیدا می‌کند[۱۱]. در توزیع نرمال، ٦٨،٢٪ داده‌ها در فاصله کمتر از یک انحراف معیار نسبت به میانگین قرار دارند. این مقدار برای فاصله‌های دو و سه انحراف معیار، به ترتیب ۹۵،٤٪ و ۹۹،۷٪ است. به بیان دیگر، احتمال آن که اختلاف یک داده با میانگین، بیش از سه انحراف معیار باشد، تنها ۰،۳٪ (تقریباً معادل ۱ در ۳۰۰) است[۱۱] [۳٦].

برای مثال، صفت age که میانگین آن ۳٦.۷۳۱ است در حالی که انحراف معیار آن ۱۰.٦٤۷ است. حال مطابق با قانون سر انگشتی معرفی شده می‌بایست فاصله دو برابر انحراف معیار از میانگین را محاسبه کنیم که به صورت زیر عمل می‌کنیم:

[1]- outlier

(۳۶.۷۳۱+(۲*۱۰.۶۴۷))= ۵۸.۰۲۵

(۳۶.۷۳۱-(۲*۱۰.۶۴۷))= ۱۵.۴۳۷

فرمول شماره (۷): محاسبه فاصله دو برابر انحراف معیار از میانگین

صفت age دارای یک بازه از ۱۷ تا ۵۷ است که نشان می‌دهد همه مشاهدات ما در فاصله کمتر از دو برابر انحراف معیار از میانگین که ۵۸.۰۲۵ و ۱۵.۴۳۷ است قرار دارد و در محدوده می‌باشد.

در بررسی‌های صورت گرفته مشخص گردید مقادیر مجموعه داده ما فاقد مقادیر خارج از محدوده است و گم شده است.

۵- روش کار مدل پیشنهادی

در این روش ابتدا متغیرهای دارای ضریب همبستگی بیشتر با استفاده از ضریب همبستگی پیرسون[1] شناسایی شده و در ادامه با انتخاب متغیرهای قوی‌تر (بالاتر از +۰.۴ یا پایین تر از -۰.۴) مدل‌سازی صورت می‌گیرد. با مقایسه نتایج حاصل از مدل‌سازی این روش انتخاب متغیر کارایی بهتری نسبت به سایر روش‌های انتخاب صفات از خود نشان می‌دهد و بیست متغیر جهت مدل سازی انتخاب می‌شود.

در ادامه از الگوریتم بیز ساده جهت مدل سازی استفاده شده و با استفاده از الگوریتم آدابوست تقویت شده و توانایی لازم جهت پیش-بینی نوع تومور را از خود نشان می‌دهد.در این پژوهش آدابوست وظیفه دارد تا هفت دسته بند متفاوت برای بیز ساده ایجاد کند که بیشتر روی تاپل های سخت تر تمرکز دارد منظور از تاپل های سخت تر مقادیری است که در دست بندهای قبلی غلط پیش‌بینی شده اند. آدابوست با انتخاب زیر مجموعه هایی از مجموعه داده اصلی به مدل سازی پرداخته و در نهایت هفت دسته بند ایجاد شده به وسیله رای گیری و در نهایت نیز برای پیش بینی از همین هفت دسته بند استفاده می‌شود.

با توجه به هدف این تحقیق، که بهبود کارایی و دقت پیش‌بینی نوع تومور بیماران سرطانی است، ابتدا چندین الگوریم دسته‌بندی مقایسه شده و پس از آن با استفاده از ضریب همبستگی پیرسون، متغیرهایی که دارای میانگین همبستگی و ارتباط بیشتری نسبت به سایر متغیرها هستند انتخاب و عمل آموزش و ایجاد مدل جدید بر روی این متغیرها صورت می‌گیرد. با حذف متغیرهای ضعیف که دارای همبستگی صفر یا کمی هستند علاوه بر افزایش کارایی پارامترهایی همچون Accuracy، Recall ، Precision، Kappa **تقویت شده و** میزان بازه احتمالی خطا در هر بار آموزش نیز کاهش پیدا کرده است و در ادامه با استفاده از الگوریتم آدابوست، فرآیند آموزش بهبود پیدا می‌کند. وجود بازه احتمالی خطا به دلیل این است که در این پژوهش از روش **Cross Validation** جهت ارزیابی مدل استفاده گردیده است. دقت مدل پیشنهادی پس از ارزیابی ۹۶.۶۷ درصد می‌باشد.

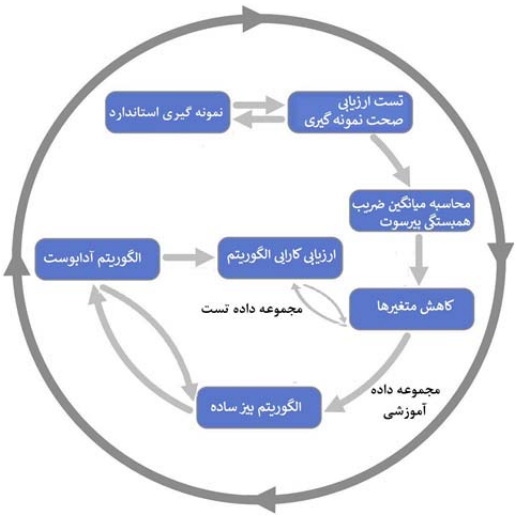
۵-۱ مراحل اجرایی

مراحل اجرایی روش پیشنهادی به شرح زیر است:

۱. نمونه‌گیری استاندارد
۲. تست ارزیابی صحت نمونه‌گیری

[1] Pearson correlation

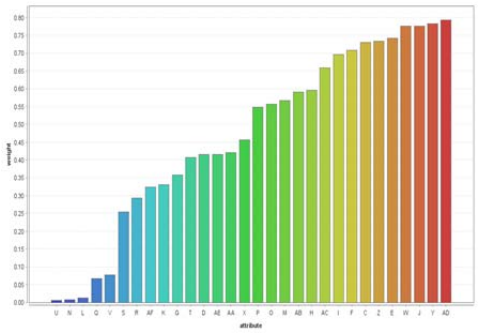
۳. محاسبه میانگین ضریب همبستگی با روش پیرسون و انتخاب متغیرهای قوی‌تر
٤. کاهش ابعاد و حذف متغیرهای با همبستگی کمتر
٥. فرآیند آموزش با الگوریتم بیزساده
٦. ترکیب الگوریتم بیز ساده[1] با الگوریتم آدابوست
۷. تست مدل ترکیبی ضریب همبستگی با پیرسون و بیزساده و آدابوست



شکل ٦) فلوچارت روش پیشنهادی

٦- کاهش تعداد متغیرها

جهت بهبود پارامترهای مختلف با استفاده از میانگین ضریب همبستگی پیرسون برای متغیرها کاهش ابعاد صورت می گیـرد و بیسـت متغیر برتر انتخاب می‌شود و عملیات آموزش و تست صورت می‌گیرد.



نمودار ۲ :تست و مرتب‌سازی میزان همبستگی متغیرها با روش پیرسون

[1] Naive Bayes

جدول ۲: مقدار وزن تعیین شده به همراه نام متغیر

ردیف	نام متغیر	وزن
۱	AD	۰.۷۹۳۵٦٦
۲	Y	۰.۷۸۲۹۱٤
۳	J	۰.۷۷٦٦۱٤
٤	W	۰.۷۷٦٤٥٤
٥	E	۰.۷٤۲٦۳٦
٦	Z	۰.۷۳۳۸۲٥
۷	C	۰.۷۳۰۰۲۹
۸	F	۰.۷۰۸۹۸٤
۹	I	۰.٦۹٦۳٦
۱۰	AC	۰.٦٥۹٦۱
۱۱	H	۰.٥۹٦٥۳٤
۱۲	AB	۰.٥۹۰۹۹۸
۱۳	M	۰.٥٦۷۱۳٤
۱٤	O	۰.٥٥٦۱٤۱
۱٥	P	۰.٥٤۸۲۳٦
۱٦	X	۰.٤٥٦۹۰۳
۱۷	AA	۰.٤۲۱٤٦٥
۱۸	AE	۰.٤۱٦۲۹٤
۱۹	D	۰.٤۱٥۱۸٥
۲۰	T	۰.٤۰۸۰٤۲
۲۱	G	۰.۳٥۸٥٦
۲۲	K	۰.۳۳۰٤۹۹
۲۳	AF	۰.۳۲۳۸۷۲
۲٤	R	۰.۲۹۲۹۹۹
۲٥	S	۰.۲٥۳۷۳
۲٦	V	۰.۰۷۷۹۷۲
۲۷	Q	۰.۰٦۷۰۱٦
۲۹	L	۰.۰۱۲۸۳۸
۲۹	N	۰.۰۰۸۳۰۳
۳۰	U	۰.۰۰٦٥۲۲

نمودار ۳ :نمودار Hinton ضریب همبستگی متغیرها

در توضیح نمودار۳ هر مربع نشان دهنده یک متغیر است و هر چه مربع قرمز رنگ بزرگ تر باشد، نشان دهنده ضریب همبستگی بیشتر آن است.

۷- شبه کد روش پیشنهادی

Calculate Pearson correlation coefficient

$$(1)\Gamma = \frac{n(\sum xy)-(\sum x)(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2]}}$$

Selecting variables with more correlation coefficient

Algorithm: AdaBoost. A boosting algorithm-Create an ensemble of classifiers. Each one gives a weighted vote.

Input:

- D, a set of d class- labled training tuples;
- K, the number of round (one classifier is generated per round);
- a classification learning scheme.

Output: A composite model.

Method:

(1) initialize the weight of each tuple in D to 1/d;
(2) for i=1 to k do // for each round:
(3) sample D with replacement according to the tuple weights to obtain D_i;
(4) Use training set D_i to derive a model, M_i;
(5) Compute error(M_i), the error rate of M_i (Eq. 8.34)
(6) If error(M_i) > 0.5 then
(7) go back to step 3 and try again;
(8) Endif
(9) For each tuple in D_i that was correctly classified do
(10) Multiply the weight of the tuple by error(M_i)/(1-error(m_i)); //update weights
(11) Normalize the weight of each tuple;

(12) Endfor

To use the ensemble to classify tupe, X:
(1) Initialize weight of each class to 0;
(2) For i=1 k do //for each classifier:
(3) $W_i = log \frac{1-error(M_i)}{error(M_i)}$; // weight of the classifiers vote
(4) $C=M_i(X)$; // get class prediction for X from M_i
(5) Add w_i to weight for class c
(6) Endfor
(7) Return the class with the largest weight;

۷-۱ تشریح کدهای به کار رفته

استفاده از فرمول پیرسون ضریب همبستگی میان متغیرها شناسایی شده و متغیرهای دارای ضریب بیشتر از +۰.٤ و کمتر از -۰.٤ انتخاب می‌شود.

$$\Gamma = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

فرمول شماره (۸): محاسبه ضریب همبستگی خطی پیرسون

در این فرمول x و y دو متغیر و n تعداد تاپل‌ها است. به عنوان مثال در مجموعه داده D با دو متغیر xو yدر صورتی که پرونده ٦ بیمار وجود داشته باشد مقدار n عدد ٦ خواهد بود. همچنین $\sum xy$ ضرب مقادیر هر ردیف متغیر x در y خواهد بود که در نهایت با هم جمع می‌شود. هدف از این کار انتخاب متغیرهای قوی‌تر و حذف متغیرهای ضعیف‌تر است، که در نهایت بیست متغیر جهت مدل‌سازی مرحله بعد انتخاب می‌شود.

در ادامه مجموعه داده که دارای کاهش تعداد متغیرها است با روش آدابوست تقویت می‌شود. آدابوست الگوریتمی از شاخه بوستینگ است.

در روش boosting به هر یک از تاپل‌های آموزشی وزنی تخصیص داده می‌شود[20] . پس از ساخته دسته‌بند m_i وزن‌ها تغییر خواهند کرد تا دسته‌بند mi+1 که پس از mi تولید می‌شود، توجه بیشتری را بر روی تاپل‌هایی که به درستی توسط mi دسته‌بندی نشده‌اند، داشته باشند.

دسته‌بندی نهاییm^* رای نهایی را با ترکیب رای‌های هر یک از دسته‌بندهای پایه محاسبه می‌کند. جایی که وزن رای هر یک از این دسته‌بندها تابعی از صحت آن است.

از میان روش‌های boosting الگوریتم Adaboost یکی از الگوریتم‌های رایج محسوب می‌شود که در این پژوهش نیز از این روش، برای افزایش یادگیری مدل استفاده می‌شود. مجموعه داده‌های D از تعداد d تاپل که هر یک به شکل (x_i, y_i) هستند تشکیل شده است و در آن بر چسب کلاس تاپل x_i با y_i نشان داده می‌شود.

در ابتدا الگوریتم Adaboost مقدار وزن یکسان $1/d$ را برای هر تاپل آموزشی در نظر می‌گیرد که در خط شماره ۱ شبه کد نشان داده شده است. d تعداد تاپل‌ها است. در این روش تلفیقی برای تولیدk دسته‌بند بهk مرحله نیاز است که ما تعدادk را عدد ۷ در نظر گفته‌ایم. در این پژوهش از الگوریتم بیز ساده به عنوان دسته‌بند پایه استفاده می‌شود.

در مرحله i ام با کمک نمونه‌گیری با جایگزینی مجموعه آموزشی Di شکل می‌گیرد.

در این روش ممکن است تاپل‌های یکسانی بیش از یک بار انتخاب شود. شانس انتخاب m تاپل بر اساس وزن تعیین می‌شود.

با استفاده از مجموعه داده‌های آموزشی Di مدل mi ساخته می‌شود و با کمک همین مجموعه، خطای مدل نیز محاسبه می‌گردد. پس از آن وزن‌های تاپل‌های آموزشی بر اساس اینکه چگونه دسته‌بندی شده‌اند، تنظیم می‌شود.

اگر تاپلی به درستی دسته‌بندی نشده باشند، وزن آن افزایش می‌یابد و وزن تاپل‌هایی که به درستی دسته‌بندی شده‌اند کاهش داده می‌شود.

در واقع وزن یک تاپل، دشواری دسته‌بندی آن تاپل را منعکس می‌کند، به صورتی که وزن بالاتر نشان می‌دهد اغلب این تاپل‌ها به درستی دسته‌بندی نشده است.

وزن‌های جدید تاپل‌ها برای تولید مجموعه آموزشی دسته‌بند بعدی در مرحله‌ی بعد استفاده می‌شود. ایده اصلی کار از جایی سرچشمه می‌گیرد که ما هنگام ساخت دسته‌بند مایلیم که مدل بیشتر بر روی دسته‌بندی تاپل‌هایی تمرکز کند که در مرحله قبلی به درستی دسته‌بندی نشده‌اند.

ممکن است برخی از دست‌بندها در دسته‌بندی تاپل‌های دشوار بهتر از بعضی دیگر عمل کنند، بدین طریق یک سری از دسته‌بندهایی ساخته شده‌اند که همدیگر را تکمیل می‌کنند.

مجموع وزن‌های تاپل‌هایی در D_J که مدل M_i آنها را به درستی دسته بندی نکرده است، به عنوان نسبت خطای مدل درنظر گرفته می‌شود، بنابراین داریم:

$$error(Mi) = \sum_{j=1}^{d} w_j \times err(X_j)$$

فرمول شماره (۹): محاسبه خطای دسته بند

که در آن تابع $err(X_j)$ هرگاه تاپل X_j به درستی دسته‌بندی نشده باشد، برابر با ۱ و در غیر این صورت برابر با صفر خواهد بود. چنانچه خطای حاصل از دسته‌بند Mi از مقدر ۰/۵ تجاوز کند، کارایی مدل ضعیف تشخیص داده می‌شود و از مدل صرف نظر خواهد شد، در عوض با تولید یک مجموعه آموزشی جدید Di سعی در ساختن مدل جدید Mi خواهیم داشت. نسبت خطای Mi بر روی چگونگی بهنگام‌سازی وزن‌های تاپل‌های آموزشی تاثیر می‌گذارد. چنانچه تاپلی در مرحله i ام به درستی دسته‌بندی شود، وزن آن در مقدار $error(Mi)/ (1-error(Mi))$ ضرب می‌شود(خط ۱۰ شبه کد). پس از بهنگام‌سازی وزن کلیه تاپل‌هایی که به درستی دسته‌بندی شده‌اند، وزن تمام تاپل‌ها(حتی آن‌هایی که به درستی دسته‌بندی نشده‌اند) به نحوی نرمال‌سازی می‌شود که مجموع وزن‌ها همانند قبل باقی بماند. برای نرمال‌سازی یک وزن کافی است آن را در مجموع وزن‌های قدیم ضرب و بر مجموع وزن‌های جدید تقسیم کنیم.

در نتیجه همانطور که قبل از این نیز توضیح داده شد، با این کار وزن تاپل‌هایی که به درستی دسته‌بندی نشده‌اند افزایش و وزن تاپل-هایی که به درستی دسته‌بندی شده‌اند کاهش می‌یابد. پس از پایان الگوریتم، از روش تلفیقی دسته‌بندها برای پیشگویی برچسب کلاس تاپلی مانند x استفاده می‌شود. بر خلاف روش بگینگ که در آن تمام دسته‌بندها دارای رای یکسانی هستند، در این روش بر اساس عملکرد هر دسته‌بند وزنی برای رای آن در نظر گرفته می‌شود. در واقع رای‌های آن‌ها دارای وزن‌های یکسانی نیست. هر چه نسبت خطای مدل کمتر باشد، در نتیجه صحت آن بیشتر خواهد بود و بنابراین وزن بالاتری را برای رای آن لحاظ می‌کنیم. وزن رای دسته بند =Mi برابر است با:

$$\log \frac{1 - \text{error}(M_i)}{\text{error}(M_i)}$$

فرمول شماره (۱۰): محاسبه وزن رای دسته بند

مجموع وزن‌های دسته‌بندی‌هایی که کلاس c را برای x پیشگویی می‌کنند محاسبه می‌کنیم. این کار برای تمام برچسب کلاس‌ها انجام می‌شود و بیشترین مقدار، تعیین کننده کلاس تاپل x خواهد بود.

۷-۲ محاسبه وزن رای دسته بندها:

همانطور که در تشریح مدل ذکر گردید در الگوریتم آدابوست در نهایت با استفاده از رای گیری دسته‌بندها پیش‌بینی انجام می‌گردد. البته این رای‌گیری به این صورت است که با توجه به وزن هر دسته‌بند با توجه به صحت آن مشخص می‌شود. این باعث می‌شود که تاثیر رای دسته‌بندها با هم یکسان نباشد.(هر چه وزن دسته‌بند بیشتر باشد، تاثیر رای بیشتری دارد.)

برای محاسبه وزن دسته بند از فرمول زیر استفاده می‌کنیم.

$$\log \frac{1 - \text{error}(M_i)}{\text{error}(M_i)}$$

فرمول شماره (۱۱): محاسبه نسبت خطای دسته بند

$\text{Error}(M_i)$ همانطور که قبلاً نیز تشریح داده شد نسبت خطای هر دسته‌بند است.

در مدل پیشنهادی از ۷ دسته‌بند استفاده شده که وزن هر دسته‌بند به صورت زیر می‌باشد. در تشریح این مقادیر بهتر است بگوییم Model 1 نسبت خطای کمتر و در نتیجه صحت بیشتری نسبت به دسته‌بندهای دیگر داشته است در نتیجه رای آن تاثیر بیشتری نسبت به سایر دسته‌بندها دارد.

Adaboost
Model 1 [w= 2.820] (Naive Bayes)
Model 2 [w= 1.333] (Naive Bayes)
Model 3 [w= 1.495] (Naive Bayes)
Model 4 [w= 1.231] (Naive Bayes)

Model 5 [w= 0.273] (Naive Bayes)
Model 6 [w= 0.587] (Naive Bayes)
Model 7 [w= 0.975] (Naive Bayes)

۷-۳ روش ارزیابی مدل و تاثیر افزایش حجم داده‌ها بر روی‌مدل

در این پژوهش از روش Cross validation (k-fold) جهت ارزیابی مدل استفاده گردیده است. این تکنیک بطور ویژه در کاربردهای پیش‌بینی مورد استفاده قرار می‌گیرد تا مشخص شود مدل موردنظر تا چه اندازه در عمل مفید خواهد بود. این روش مناسب برای ارزیابی مجموعه داده‌هایی با تعداد کم و برای دوری از Overfitting پیشنهاد می‌شود. در این مقاله از مقدار پیش‌فرض و متعارف تعداد تکرار مدل و مجموعه‌داده که ده بار است، استفاده شده است. نتایج حاصل از بررسی‌ها با استفاده از این روش نشان می‌دهد مدل پیشنهادی قابلیت تعمیم در دنیای واقعی را داشته و توانایی لازم در صورت افزایش تعداد بیماران و حجم داده‌ها را با سرعت مناسب دارد.

لازم به توضیح است الگوریتم‌ها با لپ‌تاپLenovo G50 با CPU: Core i3 اینتل اجرا و مدل‌سازی شده‌اند.

همچنین برای محاسبه کارایی روش پیشنهادی از نظر سرعت و مواجه با تعداد بالای بیماران ابتدا با ۵۶۹ پرونده سپس با ۱۱۳۸ پرونده و بعد از آن با ۱۷۰۷ پرونده و در نهایت با ۲۲۷۶ پرونده مدل سازی گردید که روش پیشنهادی در بیشترین تعداد بیماران قادر به پاسخگویی در کمتر از ٤ ثانیه بود که نسبت به سایر مدل ها بسیار سریع‌تر است.

۸- نتایج مدل سازی روش پیشنهادی

با استفاده از روش ترکیبی کاهش ابعاد با استفاده از اندازه‌گیری میزان همبستگی توسط تکنیک پیرسون و حذف متغیرهایی که دارای همبستگی کمتری هستند و استفاده از الگوریتم بیز ساده و ترکیب آن با الگوریتم آدابوست این الگوریتم بهینه‌سازی شده و کارایی و دقت آن به نحو قابل توجهی افزایش پیدا می‌کند.

جدول ۳: ماتریس نتایج مدل سازی روش پیشنهادی

	true M	true B	class precision
pred. M	۲۰۳	۱۰	٪۹۵.۳۱
pred. B	۹	۳٤۷	٪۹۷.٤۷
class recall	٪۹۵.۷۵	٪۹۷.۲۰	
Kappa Kohen	۰.۹۲۸		

در توضیح جدول ۳ قطر اصلی نشان دهنده تعداد پیش‌بینی صحیح مدل و قطر فرعی نشان دهنده تعداد خطای مدل است. به عنوان مثال مدل پیشنهادی ۳٤۷ مورد را B (Benign) و ۲۰۳ مورد را M (Malignant) به درستی پیش‌بینی کرده است. از این ماتریس برای محاسبه پارامترهای مختلف مدل استفاده می گردد.

جدول ٤: مدل میزان کارایی روش پیشنهادی

RD	Algorithm	Accuracy	Recall	Precision	Kappa Kohen
1	Proposed method	%٩٦.٦٧ +/-٢.٧٧	%٩٦.٤٦ +/-٣.٦١	%٩٦.٥٣ +/-2.٤١	٠.٩٢٨

در توضیح جدول ٤ معیار Kappa هر چه مقدارش نزدیک به عدد یک باشد عملکرد بهتری دارد. ارزیابی معیارهـای مختلـف نشـان دهنده کارایی بالای روش پیشنهادی است. مقادیر بدست آمده در این جدول حاصل از ماتریس ارزیابی مدل جدول شماره ۳ است به صورتی که پس از محاسبه مقدار دقت مدل پیشنهادی 96.67 درصد تخمین زده شده است که برای محاسبه آن از فرمول زیر استفاده می کنیم.

$$Accuracy= (203 + 347) /569 = 96.67$$

فرمول شماره (۱۲): محاسبه مقدار پارامتر Accuracy

مقادیر فوق از جدول شماره ۳ که حاصل از ارزیابی مدل است استخراج گردیده است. قطر اصلی این ماتریس که شامل اعداد ۲۰۳ و ۳٤۷ است تعداد بیمارانی است که سیستم به درستی پیش‌بینی کرده و به آن برچسب درست داده است. تعداد کـل مجموعـه داده نیـز ٥٦۹ مورد است. از این رو دقت مدل ۹٦.٦۷ درصد تخمین زده شده است که بالاتر از سایر روش ها است.

۸-۱ مقایسه روش پیشنهادی با سایر روش ها:

جدول ٥: مقایسه روش پیشنهادی با سایر روش ها

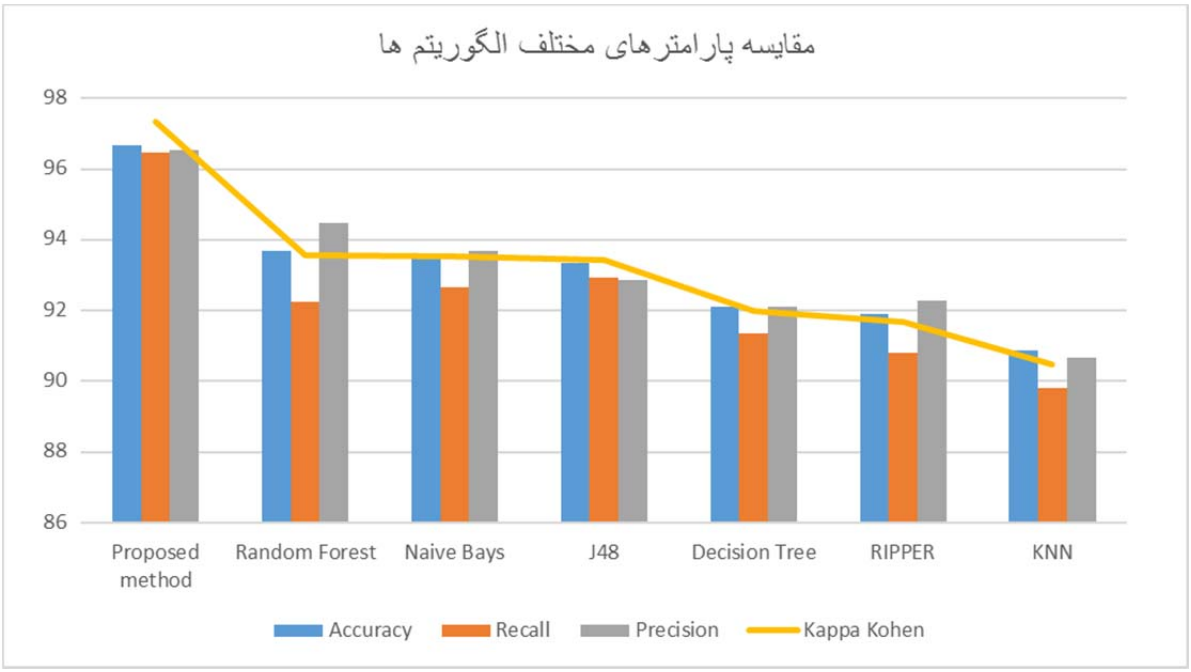
RD	Algorithm	Accuracy	Recall	Precision	Kappa Kohen
۱	**Proposed method**	**%٩٦.٦٧ +/-٢.٧٧**	**%٩٦.٤٦ +/-٣.٦١**	**%٩٦.٥٣ +/-٢.٤١**	**٠.٩٢٨**
۲	Random Forest	%٩٣.٦٨ +/-٥.٢٧	%٩٢.٢٤ +/-٦.٧٩	%٩٤.٤٨ +/-٤.٣١	٠.٨٥٩
۳	Naive Bays	%٩٣.٥١ +/-٥.٠٢	%٩٢.٦٧ +/-٦.٢٤	%٩٣.٦٨ +/-٤.٥٠	٠.٨٥٨
٤	J48	%٩٣.٣٣ +/-٤.٠٦	%٩٢.٩٣ +/-٤.٩٠	%٩٢.٨٧ +/-٤.٠٥	٠.٨٥٦
٥	Decision Tree	%٩٢.١٠ +/-٣.٩٤	%٩١.٣٧ +/-٥.٢٣	%٩٢.١١ +/-٣.٣٩	٠.٨٣٠
٦	RIPPER	%٩١.٩٢ +/-٤.٣١	%٩٠.٧٩ +/-٥.٠٥	%٩٢.٢٩ +/-٣.٩٩	٠.٨٢٤
۷	KNN	%٩٠.٨٦ +/-٣.٢١	%٨٩.٨٢ +/-٤.١٦	%٩٠.٦٦ +/-٢.٩٩	٠.٨٠٢

نتایج جدول شماره ٥ نشان می‌دهد، مدل پیشنهادی ترکیبی از پیرسون، بیـز و آدابوسـت در همـه پارامترهـای ارزیـابی شـده از سـایر الگوریتم‌ها بهتر عمل می‌کند. توانایی این روش ترکیبی وقتی بهتر مشخص می‌گردد که در کنار پارامترها با توجه بـه اسـتفاده از روش CrossValidation میانگین خطای مدل در ده بار ارزیابی به ۲.۷۷ درصد کاهش پیدا کرده است در صورتی‌که مدل اولیه بیـز سـاده تلورانس ٥.۰۲ درصد داشت. مدل پیشنهادی نیز دارای دقت ۹٦.٦۷ درصد است. پس از مدل پیشنهادی الگوریتم درخت تصـادفی بـا

دقت ۹۳.٦۸ درصد قرار می‌گیرد این الگوریتم دارای خطای میانگین ۵.۲۷ درصدی در ده بار مدل سازی است که نسبت به روش ترکیبی بسیار بیشتر است.

۸-۲ نقاط قوت الگوریتم پیشنهادی نسبت به سایر الگوریتم‌ها

دلیل تقویت مدل پیشنهادی نسبت به سایر الگوریتم‌ها این است که تکنیک آدابوست بر خلاف سایر روش‌ها بر روی مقادیری که پیش‌بینی آن‌ها سخت‌تر است تمرکز دارد. به این‌صورت که پس از مدل‌سازی اولیه سیستم با ترکیب مجموعه داده‌های مختلف در کنار یکدیگر سعی دارد تا قوانینی را کشف و هویدا سازد که بتواند داده‌های سخت را به درستی برچسب‌دهی کند از این رو با وزن دار کردن آن‌ها میزان تمرکز خود را بر روی داده‌ها افزایش یا کاهش می‌دهد. طبیعی است که داده‌های سخت‌تر بیشتر در مدل‌سازی شرکت کرده تا در نهایت سیستم توانایی پیش‌بینی وضعیت آن‌ها را نیز داشته باشد در صورتی که این تکنیک در سایر الگوریتم‌های بکار گرفته شده کاربرد ندارد.



نمودار٤: مقایسه کارایی سایر الگوریتم ها با روش پیشنهادی

۹- نتیجه‌گیری

با توجه به روند رو به رشد سرطان سینه در بین زنان و پایین آمدن سن مبتلایان به این بیماری در ایران، لزوم تشخیص نوع تومور به صورت خوش‌خیم و یا بدخیم، با دقت بالا به شدت احساس می‌شود. در این پژوهش با استفاده از ضریب همبستگی پیرسون، متغیرهایی که دارای میانگین ضریب همبستگی بیشتری هستند، انتخاب شده و در ادامه متغیرهای با همبستگی کمتر از مجموعه داده حذف گردید و کاهش ابعاد صورت گرفت. مقایسه عملکرد الگوریتم‌ها قبل و بعد از کاهش ابعاد، نشان می‌دهد دو الگوریتم بیزساده

و شبکه‌های عصبی، به‌نحو قابل توجهی در پارامترهای مختلف بهبود یافته‌اند. از این‌رو در ادامه برای بهینه‌سازی بیشتر الگوریتم بیز ساده با الگوریتم آدابوست نیز ترکیب شد. حاصل ترکیب استفاده از ضریب همبستگی پیرسون و بیز ساده با آدابوست، ایجاد مدلی ترکیبی است که توانایی پیش‌بینی و شناسایی نوع تومور سرطانی با دقت ۹۶.۶۷ درصد را دارا است. از مزایای روش پیشنهادی سهولت پیاده‌سازی و سرعت و دقت بیشتر می‌توان اشاره نمود. استفاده از همبستگی پیرسون و روش بکار رفته در مقایسه با روش‌های پیشین راحت‌تر با آموزش ساده‌تر و با دقت بالاتر می‌باشد.

منابع

[1] Thangaraju, P., Mehala, R., (2015); "Novel Classification based approaches over Cancer Diseases ", International Journal of Advanced Research in Computer and Communication Engineering, Vol 4 , Issue 4, 294-297

[2] Karim Khani Zand Hamid, (2015); "A COMPARITIVE SURVEY ON DATA MINING TECHNIQUES FOR BREAST CANCER DIAGNOSIS AND PREDICTION", Indian Journal of Fundamental and Applied Life Sciences, Vol ٥, ٤٣٣٠-٤٣٣٩

[3] Majidi Zolbanin Hamed, Delen Dorsan, Hassan Zadeh Amir, (201٥); "Predicting overall survivability in comorbidity of cancers: A data mining approach ", Elsevier Decision Support Systems 74,150-161

[4]Venkatalakshmi B, Shivsankar M.V, (2014);" Heart Disease Diagnosis Using Predictive Data mining", International Journal of Innovative Research in Science, Engineering and Technology, Volume 3, Issue 3,1873-1877

[5] Vikas Chaurasia, Saurabh Pal, (2014); "A Novel Approach for Breast Cancer Detection using Data Mining Techniques", International Journal of Innovative Research in Computer and Communication Engineering, Vol 2, Issue 1,2456-2465

[6]غلامی، م. برومندنیا، ع. (۱۳۹۵)، ارائه روشی با استفاده از ترکیب بهره اطلاعاتی، K نزدیکترین همسایه و شبکه های عصبی جهت پیش بینی وضعیت جنین ، یازدهمین کنفرانس علوم و تکنولوژی، مشهد، ایران، بهمن ۹۵.

[7] میرعابدینی سیدجواد، غلامی، م. (۱۳۹۵)؛ «ارائه روشی جهت افزایش دقت تشخیص بیماری‌های قلبی با ترکیب الگوریتم‌های داده‌کاوی(MAWB)»، سومین همایش ملی مهندسی رایانه و مدیریت فناوری اطلاعات، سوم، تهران، دانشگاه شهید بهشتی

[8] P.Ramachandran, Dr.N.Girija, Dr.T.Bhuvaneswari, (2013); "Cancer Spread Pattern – an Analysis using Classification and Prediction Techniques ", International Journal of Advanced Research in Computer and Communication Engineering, Vol 2, Issue 6,2363-2367

[9] Kharya Shweta, (2012); "USING DATA MINING TECHNIQUES FOR DIAGNOSIS AND PROGNOSIS OF CANCER DISEASE ", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol 2, Issue 2,55-66

[10] Safdari Reza, Ghazisaeedi Marjan, et al (2013); "A Model for Predicting Myocardial Infarction Using Data Mining Techniques", Iranian Journal of Medical Informatics, Vol 2, Issue 4,1-6

[11] غلامی، م. (۱۳۹٦)، داده‌کاوی برای همه، تهران، انتشارات ناقوس، چاپ اول.

[12] North, M., (2012), Data mining for the masses. Amazon, First Edition

[13] Purusothaman, G., Krishnakumari, P., (2015), A Survey of Data Mining Techniques on Risk Prediction: Heart Disease, Indian Journal of Science and Technology, Vol 8(12), 2–5.

[14] غلامی، م. نجفی، ن، (۱۳۹۵)، بررسی و مقایسه اثربخشی الگوریتم های داده کاوی جهت پیشبینی بیماری پارکینسون، اولین کنفرانس بین المللی چشم اندازهای نو در مهندسی برق و کامپیوتر، تهران، ایران، دانشگاه علم و صنعت، بهمن ۹۵.

[15]Durgalakshmi, B., Vijayakumar, V., (2015), Progonosis and Modelling of Breast Cancer and its Growth Novel Naive Bayes, Procedia Computer Science, Vol 50, 551, 553.

[16]چوبینه، پ. غلامی، م. (۱۳۹۵)، مقایسه شش الگوریتم برتر حوزه داده کاوی، یازدهمین کنفرانس علوم و تکنولوژی، مشهد، ایران، بهمن ۹۵.

[17] Guru Rao, C.V., Sreenivasa Rao, M., (2016), Cluster Analysis of Medical Research Data using R, Global Journal of Computer Science and Technology: C Software & Data Engineering, Vol 16, 17–22.

[18] Karim Khani Zand, H., (2015), A COMPARITIVE SURVEY ON DATA MINING TECHNIQUES FOR BREAST CANCER DIAGNOSIS AND PREDICTION, Indian Journal of Fundamental and Applied Life Sciences, Vol 5,4330-4339.

[19] Boughorbel, S., Al-Ali, R., Elkum, N., (2016), Model Comparison for Breast Cancer Prognosis Based on Clinical Data, PLOS ONE, 15-1.

[20] Han, J., Jian, P., (2011), Data mining: concepts and techniques, Elsevier.

[21] 5. L. Breiman, J. H. Friedman. (1984),"Classification and regression trees," Monterey.

[22] حقیقی. مهری،"داده کاوی و یادگیری ماشین: مروری بر دسته بندی کننده ها"، کنفرانس بین المللی یافته های نوین، تهران، ایران، شهریور ۹٤.

[23] مرتضی‌پور.رضا، مطلق‌زاده،مهسان، استفاده از شبکه بیزین در طبقه‌بندی، چهارمین کنفرانس مهندسی برق و الکترونیک، دانشگاه آزاد گناباد، شهریور،۱۳۹۱.

[24] نظری. احمد،"ارائه الگوریتم ترکیبی بهینه بردار ماشین پشتیبان و جنگل تصادفی در تشخیص به موقع بیماری های قلبی"، کنفرانس بین المللی پژوهش در علوم و تکنولوژی، مالزی، اذز ۹٤.

[25] Y. Liu and P. Zhao and et al. (2015), "A Boosting Algorithm for Item Recommendation with Implicit Feedback," Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

[26] Krishnaiah, V., Narsimha,G., Subhash Chandra, N., (2016), Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review, International Journal of Computer Applications, Vol 136, No 2, 43–51.

[27] سعادت. محمد، زمانی بروجنی. فرساد،"مروری بر روش های بهبود کارایی تکنیک جنگل تصادفی"، اولین همایش ملی فناوری اطلاعات، ارتباطات و محاسبات نرم، دانشگاه آزاد خوراسگان، اصفهان، ایران. اردیبهشت ۱۳۹۵.

[28] Han, Y. Liu, X. Sun. (2013), "A scalable random forest algorithm based on MapReduce," Software Engineering and Service Science (ICSESS), 2013 4th IEEE International Conference on, pp.849-852.

[29] Parijaee Moghaddam. A and Mousavi. S , Learning Decision Tree Using Neural Network for Stability and Flexibilit, Iranian Journal of Medical Informatics, vol. 1, no. 3, pp. 39-44, 2013.

[30] T. M. Khoshgoftar, M. Golawala, and J. Van Hulse. (2007), "An Empirical Study of Learning from Imbalanced Data Using Random Forest.," presented at the 19th IEEE Conference on Tools with Artificial Intelligence.

[31]Zheng, G. (2017), Logistic Regression, Model Selection, and Cross Validation, personal.umich.edu, March. 25.

[32] Soundarya, M., Balakrishnan, R., (2014), Survey on Classification Techniques in Data mining, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, 7550-7552.

[33] Zheng, G. Logistic Regression, Model Selection, and Cross Validation, personal.umich.edu, March. 25, 2017.

[34] Street, W., Wolberg, W., Mangasarian, O., (1992), Nuclear Feature Extraction For Breast Tumor Diagnosis. International Symposium on Electronic ImagingScience and Technology, VOL, 1905, 861-870.

[35] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[36] Christian, S., Winston,W., Zappa, Ch., (2011), Data Analysis and Decision Making, Forth edition, 14-16.

use Pearson’s Linear Correlation and the combination of Data Mining Algorithms simultaneously to improve prognosis of a kind of tumor in cancer patients

mohsen gholami[1], Seyed Javad Mirabedini[2*]

Islamic Azad University Bushehr, Iran - mohsen.gholami18@outlook.com
Department of Computer, Central Tehran Branch, Islamic Azad University, Tehran, Iran j_mirabedini@iauctb.ac.ir

Abstract:
Nowadays, breast cancer is the most common cancer disease among women. Statistics shows a six percent increase in Iran which indicates it as a serious danger. However, its danger can be prevented increasingly by early diagnosis or prediction. By medical science progress, the way for developing of a system with the capability of prevention, prognosis and cure by using the new technologies is paved. Medical data mining tries to design a model and find relationships among risky factors to predict the condition of future patients with the aid of current data. We try to compare different data mining algorithms and combination of these algorithms to develop a new, efficient method with high accuracy and capability to perform on local data. Finally, proposed method which improves efficiency of Naive Bayes with Adaboost algorithm can predict the kind of benign or malign tumor with the 96/67% accuracies. Required data for this procedure is extracted from UCI site to diagnose the kind of tumor with 569 records and 32 variables.

Keywords: Pearson’s correlation coefficient, classification algorithms, Naive Bayes, Adaboost