مقاله
پژوهشی

کاهش خطای رده‌بندی تعیین بیماری تیروئید در شهرستان شوشتر با استفاده از الگوریتم بوستینگ درختی

فردوس محمدی بساتینی[1*]، بهزاد ریحانی نیا[2]

[1]مربی، دانشگاه آزاد اسلامی، واحد شوشتر، گروه ریاضی، شوشتر، ایران
[2]مربی، دانشگاه آزاد اسلامی، واحد شوشتر، گروه ریاضی، شوشتر، ایران
[*]**نویسنده مسئول:** دانشجوی دکترا آمار، مربی دانشگاه آزاد اسلامی، واحد شوشتر، گروه ریاضی، شوشتر، ایران.
پست الکترونیک: fe.mohamadi91@gmail.com

چکیده

زمینه و هدف: *غده تیروئید یکی از غدد حیاتی بدن است که می‌توان گفت به طور غیر مستقیم روی تمام ارگان‌های بدن مانند قلب، کلیه، دستگاه گوارش و غیره اثر دارد هدف این مطالعه استفاده از الگوریتم بوستینگ در کاهش خطای تشخیص غده تیروئید نرمال از غده تیروئید غیرنرمال می‌باشد. این الگوریتم یک روش قدرتمند در حوزه تشخیص و پیش‌بینی می‌باشد. الگوریتم بوستینگ به طور مکرر یک رده‌بندی کننده پایه را روی داده‌های دوباره وزن‌دار شده رشد می‌دهد و در نهایت یک ترکیب خطی از نتایج تشکیل می‌دهد و از این رو دقت را بهبود می‌بخشد.*

مواد و روش کار: *این مطالعه از نوع مقطعی است. داده‌های وضعیت غده تیروئید یک نمونه ۱۰۳تایی از مراجعه کنندگان به آزمایشگاه سلامت شهرستان شوشتر در سال ۸۹-۹۰ مورد تحلیل قرار گرفت برای تشخیص غده تیروئید نرمال از غده تیروئید غیرنرمال از درخت‌های تصمیم معمولی و درخت‌های تصمیم بوستینگ از نرم افزار R۳.۰.۱ استفاده شد. برای مقایسه نتایج از روش تحلیل رده‌بندی و سه معیار نرخ خطای رده‌بندی، حساسیت و ویژگی استفاده شد.*

یافته‌ها: *نرخ خطای رده‌بندی، حساسیت و ویژگی در مجموعه آزمون برای درخت‌های تصمیم معمولی به ترتیب ۰/۰۸۸، ۰/۹۱ و ۰/۹۲ به دست آمدند و در درخت‌های تصمیم بوستینگ سه معیار فوق به ترتیب ۰/۰۲۹، ۰/۹۵۵ و ۱ به دست آمدند.*

نتیجه‌گیری: *نتایج این مطالعه نشان داد که الگوریتم بوستینگ برای تشخیص غده تیروئید نرمال از غده تیروئید غیرنرمال بسیار موفق‌تر عمل می‌کند بنابراین استفاده از درخت‌های تصمیم بوستینگ جهت تشخیص و پیشگویی وضعیت غده تیروئید پیشنهاد می‌شود.*

واژه های کلیدی: *الگوریتم بوستینگ، نرخ خطای رده‌بندی، حساسیت، ویژگی.*

مجله دانشگاه علوم پزشکی خراسان شمالی سال ۹۴؛ دوره ۷ (۲):۳۸۱-۳۹۱
Journal of North Khorasan University 2015;7(2):381-391

وصول: ۹۳/۲/۲۸
اصلاح: ۹۳/۵/۴
پذیرش: ۹۳/۷/۲۲

مقدمه

غده تیروئید در قاعده گلو و اطراف نای قرار دارد اندازه این غده حدود ۲۰-۱۲ گرم و بسیار پر عروق و دارای قوام نرم است. غده تیروئید هورمون‌هایی را می‌سازد که به همه سلول‌های بدن می‌روند و در تنظیم سرعت تغییر غذا به نیرویی که بدن بتواند از آن استفاده کند، ضربان قلب، قدرت عضلانی، رشد، گرمای بدن و روحیه دخالت دارد. غده تیروئید دو هورمون وابسته به یکدیگر شامل تیروکسین (T4) و تری یدو تیرونین (T3) را تولید می‌کند این هورمون‌ها در طول رشد جنین نقش مهمی در تمایز سلول‌ها ایفاء می‌کنند. و توسط هورمون تحریک کننده غده تیروئید تیروتروپین (TSH) کنترل می‌شوند. متداول‌ترین مشکلات غده تیروئید کم‌کاری و پرکاری غده تیروئید می‌باشد. کم‌کاری غده تیروئید معمولاً به علت کمبود ید در بدن ایجاد می‌شود و در خانم‌ها بیشتر از آقایان می‌باشد که این می‌تواند ناشی از اثرات هورمون‌های جنسی بر روی پاسخ ایمنی بدن و یا یک عامل ژنتیکی مربوط به جنس مؤنث باشد از جمله علائم کم‌کاری تیروئید می‌توان به خستگی و ضعف، ریزش مو، اشکال در حافظه و تمرکز، یبوست، افزایش وزن همراه با کاهش اشتها و اختلالات آزمایشگاهی شامل افزایش کلسترول، افرایش تری گلیسرید و کم خونی اشاره نمود. کم‌کاری تیروئید ممکن است پس از زایمان ایجاد شود بنابراین تیروئید خانم‌هایی که پس از زایمان دچار افسردگی شدید می‌شوند باید بررسی شود [۱].

پرکاری غده تیروئید که به صورت افزایش بیش از حد عملکرد تیروئید توصیف می‌شود و شیوع آن نیز در خانم‌ها بیشتر از آقایان است، معمولاً ناشی از بیماری گریوز (Grave's) است. این بیماری زمانی اتفاق می‌افتد که سیستم ایمنی بدن تیروئید را مجبور به فعالیت بیش از اندازه می‌کند اما ممکن است بر اثر توده کوچکی که روی غده تیروئید درآمده است و هورمون‌های زیاد تولید می‌کند نیز ایجاد شود. بعضی از افرادی که دچار بیماری گریوز شده‌اند چشمان برآمده و اشک فراوان دارند دیگران ممکن است گواتر یعنی برآمدگی قابل مشاهده در جلوی گردن داشته باشند. از جمله علائم دیگر پرکاری تیروئید عدم تحمل گرما و تعریق فراوان، طپش قلب، کاهش وزن همراه با افزایش اشتهاء، لرزش اندام‌ها، ضعف عضلانی و از اختلالات آزمایشگاهی اختلال آنزیم‌های کبدی و کم خونی می‌باشد [۱].

رده‌بندی به معنی جدا کردن مجموعه‌های متمایز نمونه‌ها و تخصیص نمونه‌های جدید به کلاس‌های تعریف شده قبلی می‌باشد. و هدف از کاربرد آن در تحقیقات کلینیکی تشخیص و پیشگویی بیماری می‌باشد برای این منظور با انتخاب صفت‌های مناسبی از کلاس‌ها و اندازه‌گیری آنها در نمونه‌های بدست آمده از کلاس‌های مختلف راهکار مناسب برای رده‌بندی یک نمونه جدید به یکی از این کلاس‌های تعریف شده قبلی ارائه می‌گردد [۲].

درخت‌های تصمیم‌گیری یکی از رایج‌ترین روش‌های رده‌بندی هستند که به ویژه استفاده از این درخت‌ها برای رده‌بندی بیماران بر اساس وجود یا عدم وجود نشانه‌های بیماری یا سلامتی رو به افزایش است. درخت‌های تصمیم‌گیری به دلیل ساختار ساده و گرافیکی امروزه از پرکاربردترین روش‌ها در تشخیص‌های پزشکی می‌باشند زیرا این روش‌ها معمولاً مطابق با فکر کردن پزشک تصمیم‌گیری می‌کنند درخت‌های تصمیم‌گیری یکی از معروف‌ترین روش‌های تحلیل داده‌ها هستند زیرا نسبتاً سریع ساخته می‌شوند و مدل‌های تفسیرپذیر فراهم می‌کنند، آمیخته‌ای از متغیرهای عددی، طبقه‌ای و مقادیر گم شده را در می‌توانند در ترکیب مدل قرار می‌دهند و تحت تبدیلات متغیرها پایدار هستند بنابراین مقیاس‌بندی و یا تبدیلات مشکلی ایجاد نمی‌کند. نسبت به اثر داده‌های پرت ایمن هستند، انتخاب متغیرها را خود انجام می‌دهند، همچنین لازم نیست شکل رابطه پارامتری بین متغیرهای پیشگو و متغیر پاسخ معلوم باشد [۳]، این ویژگی‌ها عمدتاً دلیل استفاده از درخت‌های تصمیم به عنوان رایج‌ترین روش در داده کاوی هستند. درخت تصمیم در مسائلی کاربرد دارد که بتوان آن را به صورتی مطرح نمود که پاسخ واحدی به صورت نام یک کلاس ارائه دهد. برای مثال می‌توان درخت تصمیمی ساخت که به این سؤال پاسخ دهد که بیماری مریض کدام است؟ و یا درختی ساخت که به این سؤال پاسخ دهد آیا مریض به هپاتیت مبتلاست؟ و برای مسائلی مناسب است که نمونه‌های آموزش (training) به صورت زوج (کلاس- ویژگی)

مشخص شده باشند و تابع هدف دارای خروجی با مقادیر گسسته باشد مثلاً هر نمونه با بله و خیر تعیین شود [۴]. در هر درخت تصمیم ریشه بالاترین گره است و نمونه‌ها از بالا به پایین درخت عبور می‌کنند در هر گره داخلی یک تصمیم گرفته می‌شود تا اینکه به گره نهایی که برگ نامیده می‌شوند برسند، هر گره داخلی شامل یک سؤال است که بر اساس آن یک تقسیم‌بندی صورت می‌گیرد و هر برگ این درخت یک کلاس در رده‌بندی را تعیین می‌کند درخت‌های تصمیم از روش تقسیم‌بندی مکرر دوتایی برای تقسیم نمونه‌ها به دو زیر مجموعه مجزا استفاده می‌کنند و در هر زیر مجموعه نمونه‌ها به کلاسی رده‌بندی می‌شوند که اکثر نمونه‌های زیر مجموعه در آن کلاس قرار دارند [۶،۵].

توضیحات بالا متمرکز بر الگوریتم CART که رایج‌ترین روش ساختن درخت است می‌باشد روش دیگر الگوریتم ID3، و نسخه‌های بعدی آن یعنی C4/5 و C5/0 می‌باشند [۷]. الگوریتم ID3 محدود به متغیرهای پیشگو طبقه‌ای بود این الگوریتم از مفهوم آنتروپی و نفع اطلاعات برای تقسیم‌بندی استفاده می‌کرد. الگوریتمC4/5 تکمیل شده الگوریتم ID3 است این الگوریتم قابل کاربرد برای داده‌های پیوسته است. آخرین نسخه از درخت تصمیمC4/5 الگوریتم C5/0 نامیده می‌شود. این الگوریتم بسیار شبیه به الگوریتم CART است. تنها خاصیت برجسته C5/0 یک طرح برای به دست آوردن قواعد تقسیم مجموعه‌ها است. به این ترتیب که بعد از رشد دادن درخت قواعد تقسیم که گره‌های نهایی را می‌سازند می‌توانند بعضی مواقع ساده‌تر شوند یعنی یک یا چندین شرط می‌توانند حذف شوند بدون اینکه تغییری در زیر مجموعه‌های متعلق به آن گره ایجاد شود بنابراین با یک مجموعه ساده شده از قواعد که هر گره نهایی را تعریف می‌کنند به پایان می‌رسیم اما این دیگر از روش ساختن درخت تبعیت نمی‌کند [۴]؛ با توجه به توضیحات بالا در این مطالعه از الگوریتم CART برای ساختن درخت استفاده می‌شود. از آنجا که استفاده از درخت‌های تصمیم در تحقیقات کلینیکی رو به افزایش است نگرانی‌ها درباره دقت روش‌های درخت ساختار در رگرسیون و رده‌بندی افزایش یافته است [۳]. به عبارت دیگر یک مشکل در ارتباط با درخت‌ها واریانس زیاد آنها است اغلب یک تغییر کوچک می‌تواند تبدیل به یک سری کاملاً متفاوت از تقسیم‌ها شود و بنابراین تفسیر آنها را تا حدودی ناپایدار می‌سازد. دلیل عمده این ناپایداری طبیعت سلسله مراتبی فرآیند است یک خطا در تقسیم بالا به سمت پایین در همه تقسیم‌های زیرین پخش می‌شود [۶]. بنابراین یک محدودیت درباره‌ی درخت‌های تصمیم قدرت پیشگویی آن‌ها است این درخت‌ها معمولاً از دقت بالایی برخوردار نیستند به همین دلیل همواره سعی شده است که با اصلاحات جدیدتر بر روی درخت‌های تصمیم کارایی آن‌ها در رده‌بندی افزایش یابد. از آنجا که در حوزه پزشکی هدف از رده‌بندی معمولاً تشخیص و یا پیشگویی بیماری می‌باشد بنابراین خطا در این حوزه می‌تواند صدمات جبران‌ناپذیری در پی داشته باشد

بر همین اساس جایگزین‌ها و تعمیم‌هایی برای روش‌های رده‌بندی کلاسیک ارائه شده است یکی از مهم‌ترین این تعمیم‌ها الگوریتم بوستینگ است بوستینگ در لغت به معنای تقویت کردن است زیرا انگیزه اولیه دستیابی به روش بوستینگ ترکیب خروجی رده‌بندی کننده‌های ضعیف برای تولید یک رده‌بندی کننده چندگانه مفید جهت تقویت عملکرد آن‌ها است [۸،۹]. بوستینگ در ترکیب با درخت‌های تصمیم به عنوان رده‌بندی‌کننده ضعیف یکی از بهترین رده‌بندی‌کننده‌ها را نتیجه می‌دهد [۱۰]. اما به نظر می‌رسد بوستینگ در متون پزشکی ناشناخته است. در این مطالعه ما روی رایج‌ترین الگوریتم بوستینگ که آدابوست M1 نامیده می‌شود تمرکز می‌کنیم [۱۱]. الگوریتم آدابوست به طور مداوم یک مجموعه از رده‌بندی کننده‌های ضعیف وزن‌دار شده را تولید می‌کند که با هم ترکیب می‌شوند تا یک رده‌بندی کننده قدرتمند کلی بسازند. که با احتمال زیاد در کاهش نرخ خطای رده‌بندی نادرست نسبت به هر یک از رده‌بندی کننده‌های منفرد بسیار موفق‌تر است [۱۲].

از جمله کاربردهای درخت‌های تصمیم بوستینگ می‌توان به مطالعه‌ای اشاره کرد که از الگوریتم بوستینگ در ترکیب با درخت‌های تصمیم جهت رده بندی تومور بر اساس داده های نمایش دهنده ژن در میان ۷۲ بیمار پرداخته شد [۱۳]، و یا مقایسه درخت‌های رده‌بندی

بوستینگ با درخت‌های تصمیم معمولی بر روی اطلاعات بدست آمده از بیمارانی که به مدت ۳۰ روز در بیمارستان بستری شده بودند [۳]، در مطالعه ای برای جداسازی حرکات غیرنرمال چشم در میان ۸۸ بیمار شیزفرنی از یک نمونه ۸۸ تایی کنترل از درخت‌های تصمیم بوستینگ و مدل‌های شبکه عصبی استفاده شد [۱۴]، در مطالعه‌ای دیگر برای شناسایی ۵ نشانه مفید برای تشخیص تومورهای بدخیم سرطان پستان در یک نمونه ۱۸تایی از سگ‌ها از درخت‌های تصمیم بوستینگ استفاده شد [۱۵].

همانطور که گفته شد تعیین صحیح وضعیت نرمال یا غیر نرمال بودن غده تیروئید امری مهم در حوزه سلامت انسان است، از این رو، در این مطالعه استفاده از الگوریتم بوستینگ جهت کاهش خطای رده‌بندی کننده درختی در تعیین وضعیت غده تیروئید مراجعه کنندگان به یکی از آزمایشگاه‌های شهرستان شوشتر مورد نظر است.

به عبارتی مطالعه حاضر با هدف مقایسه دقت پیشگویی درخت‌های تصمیم بوستینگ با درخت‌های تصمیم معمولی در پیش‌بینی وضعیت غده تیروئید و در عین حال شناسایی متغیرهای مؤثرتر در پیشگویی وضعیت غده تیروئید انجام گردید.

روش کار

مطالعه حاضر یک مطالعه‌ی مقطعی بود که در آن از داده‌های مراجعه کنندگان به آزمایشگاه سلامت شوشتر در بازه زمانی آذر سال ۸۹ تا آذر سال ۹۰ استفاده شد با استفاده از روش نمونه‌گیری تصادفی ساده نمونه ۱۰۳ تایی از برگه جواب آزمایش بایگانی شده‌ی افراد مراجعه کننده انتخاب شد و با با رعایت موارد اخلاقی و بدون ذکر نام آن‌ها برگ آزمایشگاهی آنها مورد مطالعه قرار گرفت. داده‌های این مطالعه مربوط به وضعیت غده تیروئید افراد می‌باشند وضعیت غده تیروئید هر فرد در دو وضعیت نرمال و غیر نرمال در نظر گرفته شد. از آنجا که برای تعیین وضعیت غده تیروئید هر فرد درآزمایشاتی که از او به عمل می‌آید سه فاکتور تری یدو تیرونین، تیروکسین و تیروتروپین اندازه گیری می‌شود و با توجه به مقدار این هورمون‌ها وضعیت غده تیروئید افراد مشخص می‌شود در این مطالعه نیز سه فاکتور فوق به صورت زیر به عنوان متغیرهای پیشگو در نظر گرفته شدند؛ X_1: تری یدو تیرونین (بر حسب نانو گرم در هر میلی لیتر) X_2: تیروکسین (بر حسب میکروگرم در صد میلی لیتر) X_3: تیروتروپین (بر حسب میکروگرم در صد میلی لیتر).

برای رده‌بندی وضعیت غده تیروئید، y، به دو کلاس «غده تیروئید نرمال» و «غده تیروئید غیر نرمال» از سه متغیر پیشگو X_1، X_2، X_3 و متغیر پاسخ $y \in \{-1,1\}$ استفاده شد. برای مقایسه روش‌های رده‌بندی درخت‌های تصمیم معمولی و درخت‌های تصمیم بوستینگ دو سوم از داده‌های نمونه به روش تصادفی ساده به عنوان مجموعه آموزشی جهت برآورد مدل انتخاب شد، بقیه نمونه‌ها به عنوان مجموعه آزمون برای بررسی و مقایسه روش‌های رده‌بندی به کار رفتند.

محاسبات مربوطه با برنامه‌نویسی در نرم افزار R نسخه ۳.۰.۱ انجام شده است البته برای انجام رده‌بندی درخت‌های تصمیم معمولی بسته نرم افزاری tree و برای انجام درخت‌های تصمیم بوستینگ بسته‌های نرم‌افزاری rpart و ada به R اضافه شدند.

در اجرای درخت‌های تصمیم معمولی برای تعیین بهینه تعداد گره‌های پایانی از روش اعتبار سنجی متقابل بر روی داده‌های آموزشی استفاده شد به این ترتیب که که در این روش ۱۰درصد از داده‌های آموزشی کنار گذاشته می‌شود و با ۹۰ درصد باقیمانده مدل برآورد می‌شود و در ادامه دقت مدل روی ۱۰ درصد بقیه داده‌های آموزشی بررسی می‌شود. هر مدلی که دارای خطای مینیمم باشد انتخاب می‌شود. در این مطالعه پس از اجرا روش اعتبار سنجی متقابل بر روی مجموعه آموزشی یک درخت با ۳ گره پایانی دارای خطای مینیمم بود. در پایان برای بررسی قدرت پیشگویی، مدل انتخاب شده از طریق روش اعتبار سنجی بر روی داده‌های مجموعه آزمون اجرا می‌شود.

در اجرای درخت‌های تصمیم بوستینگ از الگوریتم آدابوستM1 با تابع زیان نمایی استفاده شد همچنین در تعداد تکرارهای الگوریتم بوستینگ با ۵۰ تکرار بهترین نتیجه به دست می‌آمد.

در منابع مختلفی که مورد مطالعه قرار گرفتند مشاهده شد که برای مقایسه دقت درخت‌های تصمیم معمولی با درخت‌های تصمیم بوستینگ از نرخ خطای رده‌بندی (misclassification) که به صورت $\frac{\sum_{i\neq j} n_{ij}}{n}$

محاسبه می‌شود استفاده می‌شود، n_{ij} ،$i \neq j$ تعداد نمونه‌هایی از کلاس i است که به اشتباه به کلاس j رده‌بندی شده‌اند و n تعداد کل نمونه‌ها است؛ نرخ خطای رده‌بندی یک استاندارد برای ارزیابی رده‌بندی کننده‌های دوتایی می‌باشد در این معیار فرض می‌شود که هزینه رده‌بندی نادرست در دو کلاس یکسان است [۱۶].

همچنین دو معیار حساسیت (sensitivity) و ویژگی (specificity) که معمولاً در رده‌بندی پزشکی برای قضاوت درباره یک قاعده رده‌بندی محاسبه می‌شوند نیز استفاده شده است که هر چقدر مقدار این دو معیار نزدیک به ۱ باشد دلالت بر خوبی روش رده‌بندی دارند [۶]. تعریف این دو معیار برای داده‌های این تحقیق عبارت است از:

حساسیت: احتمال پیش‌بینی اینکه دارای غده تیروئید غیر نرمال هستند در حالی که واقعاً غده تیروئید غیر نرمال دارند.

ویژگی: احتمال پیش‌بینی اینکه دارای غده تیروئید نرمال هستند در حالی که واقعاً غده تیروئید نرمال دارند.

برای آشنایی بیشتر با روش تحلیل، فرض کنید متغیر پاسخ y که نشان دهنده شماره کلاس است دو مقدار گسسته بگیرد یعنی $y \in \{1و-1\}$ و فرض کنید $\mathbf{x}$ نشان دهنده یک بردار از متغیرهای پیشگو باشد و رده‌بندی کننده $g(\mathbf{x})$ پیش‌بینی ارائه می‌دهد که یکی از دو مقدار $\{1و-1\}$ را می‌گیرد که وقتی با بوستینگ استفاده می‌شود رده‌بندی‌کننده پایه نامیده می‌شود. در مرحله اول الگوریتم وزن همه مشاهدات یکسان است (وزن‌ها=$\frac{1}{n}$) و رده‌بندی‌کننده به داده‌ها برازش داده می‌شود و یک کلاس برای هر نمونه به دست می‌آید نسبت نمونه‌های نادرست رده‌بندی شده محاسبه است در نهایت وزن‌های نمونه اصلاح می‌شود برای نمونه‌های درست رده‌بندی شده وزن‌ها تغییر نمی‌کند اما وزن نمونه‌های نادرست رده‌بندی شده در تکرار جدید افزایش می‌یابد سپس این فرآیند چندین بار (M بار) تکرار می‌شود سپس رده‌بندی کننده نهایی یک ترکیب وزنی از M رده‌بندی کننده است رده‌بندی کننده‌هایی که نرخ رده‌بندی درست بالاتری دارند در ترکیب نهایی وزن بیشتری می‌گیرند [۶]. الگوریتم آدابوست M1 با هر رده‌بندی کننده‌ای بکار می‌رود اما در اکثر موارد با درخت‌های تصمیم به عنوان رده‌بندی کننده پایه بکار می‌رود چنانکه حتی استفاده از کنده درخت (کنده درخت یک رده‌بندی کننده با دقیقاً یک تقسیم دوتایی و دقیقاً دو گره نهایی یا برگ است) به عنوان رده‌بندی کننده ضعیف همراه با بوستینگ نشان داده شده که یک بهبود اساسی در خطای پیشگویی آن در مقایسه با درخت‌های تصمیم معمولی حاصل می‌شود [۳].

درخت‌های تصمیم معمولی دارای توصیف ساده هستند و به شکل گرافیکی نمایش داده می‌شوند سهم هر متغیر پیشگو در رده‌بندی به آسانی ارزیابی و بررسی می‌شود با این وجود در درخت‌های تصمیم بوستینگ ارزیابی سهم هر متغیر پیشگو سخت‌تر است و باید به روش دیگری تفسیر شوند. در کاربردهای داده‌کاوی به ندرت پیش می‌آید که همه متغیرهای پیشگو دارای اثرات یکسان باشند و اغلب فقط تعداد کمی از آن‌ها دارای اثر اساسی روی پاسخ هستند. بنابراین اغلب مفید است که اهمیت نسبی و یا سهم هریک از متغیرهای پشگو در پیش‌بینی پاسخ را بدانیم. برای یک درخت تصمیم معمولی معیار

$$I_h^2(T) = \sum_{i=1}^{J-1} \hat{i}_t^2 \, I(v(t) = h)$$

برای اندازه گیری اهمیت متغیر پیشگو x_h شده است. مجموع بالا روی J-1 گره داخلی درخت است در گره t یکی از متغیرهای پیشگو $x_{v(t)}$ برای تقسیم ناحیه مربوط به آن گره به دو زیر ناحیه استفاده می‌شود و آن متغیری انتخاب می‌شود که بیشترین بهبود در $\hat{i}_t^2$ یعنی مربع ریسک خطا بدهد. مربع اهمیت نسبی x_h مجموع مربع چنین بهبودهایی روی تمام گره‌های داخلی که این متغیر به عنوان متغیر تقسیم در آن‌ها استفاده شده می‌باشد. معیار فوق برای اندازه‌گیری اهمیت متغیر در درخت‌های تصمیم بوستینگ به صورت زیر اصلاح می‌شود.

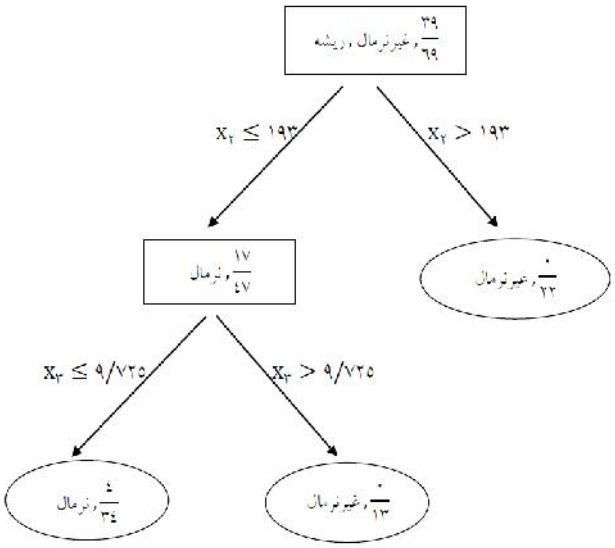
$$I_h^2 = \frac{1}{M} \sum_{m=1}^{M} I_h^2(T_m)$$

بنابراین اهمیت یک متغیر پیشگو برابر میانگین اهمیت آن متغیر در طول M درخت تصمیم با مجموعه داده‌های مکرراً وزن‌دار شده می‌باشد[۶].

یافته ها

در نمونه جمع‌آوری شده در ۴۲ نفر غده تیروئید در وضعت نرمال و برای ۶۱ نفر غده تیروئید غیر نرمال تشخیص داده شده بود، مجموعه آموزشی شامل ۶۹ نمونه بود که ۳۰ نمونه دارای غده تیروئید نرمال و ۳۹ نمونه دارای غده تیروئید غیرنرمال بودند و مجموعه آزمون شامل ۳۴ نمونه شد که ۱۲ نمونه دارای غده تیروئید نرمال و ۲۲ نمونه دارای غده تیروئید غیر نرمال بودند. شکل ۱ درخت تصمیم معمولی ساخته شده برای داده‌های غده تیروئید در مجموعه آموزشی را نشان می‌دهد، مشاهده می‌شود که یک درخت تصمیم با سه گره پایانی است. در هر گره کلاس تخصیص داده شده به مشاهدات و همچنین نسبت نقاط اشتباه رده‌بندی شده آورده شده است. بنابراین در گره شماره ۱ که همان ریشه است کل نمونه‌ها یعنی ۶۹ نمونه قرار دارند و چون تعداد نمونه‌های غیرنرمال بیشتر است نمونه‌های این کلاس به کلاس غیرنرمال رده‌بندی شده‌اند (همان‌گونه که قبلاً گفته شد در هر گره نمونه‌ها به کلاس اکثریت در آن گره رده‌بندی می‌شوند)، مقدار زیان کل در این گره ۹۴/۴۷ می‌باشد سپس نمونه‌هایی که برای آن‌ها۱۹۳ $X_2 \leq$ گره شماره ۲ را تشکیل می‌دهند که ۴۷ نمونه در این گره قرار می‌گیرد و زیان رده‌بندی در این گره ۶۱/۵۱ می‌باشد، نمونه‌هایی که برای آن‌ها۱۹۳ $X_2 >$ است گره شماره ۳ را تشکیل می‌دهند که ۲۲ نمونه در این گره قرار می‌گیرد زیان در این گره صفر می‌باشد زیرا تمام نمونه‌های موجود در این گره واقعاً متعلق به کلاس غیرنرمال می‌باشند و بنابراین در این گره تمام نمونه‌ها به درستی رده‌بندی شده‌اند از این رو این گره یک گره نهایی یا برگ می‌باشد . سپس نمونه‌هایی که برای آن‌ها ۹/۷۲۵ $X_3 \leq$ می‌باشد گره شمار ۴ را تشکیل می‌دهند و زیان در این گره ۲۴/۶۳ می‌باشد، این گره یک گره نهایی می‌باشد. نمونه‌هایی که برای آن‌ها۹/۷۲۵ $X_3 >$ است گره شماره ۵ را تشکیل می‌دهند و زیان در این گره صفر می‌باشد زیرا همه نمونه‌های این گره در واقع متعلق به کلاس غیر نرمال بوده‌اند و بنابراین همگی به درستی رده‌بندی شده‌اند و این گره هم یک گره نهایی می‌باشد، نتایج رده‌بندی در جدول ۱ آمده است.

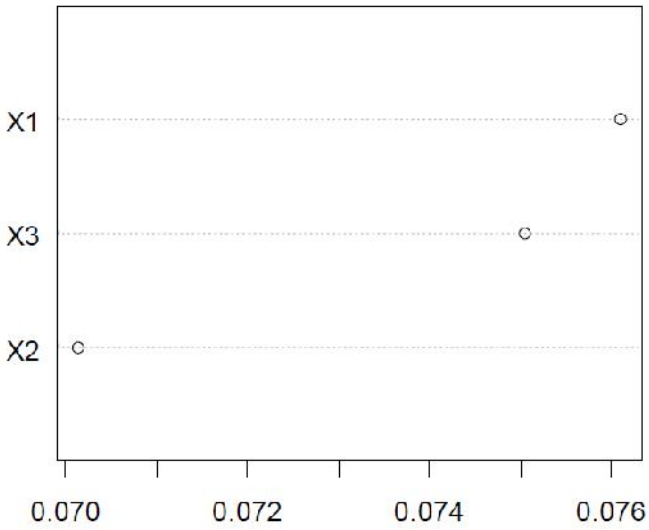
با توجه نتایج بالا مشاهده می‌شود که مقدار نرخ خطای رده‌بندی، حساسیت و ویژگی در درخت تصمیم بوستینگ برای مجموعه آموزشی و مجموعه آزمون به طور معناداری از مقادیر متناظرشان در درخت تصمیم معمولی بهتر می‌باشند. در درخت‌های تصمیم بوستینگ می‌توان نمودار اهمیت متغیرها را در رده‌بندی داده‌ها رسم کرد این نمودار در شکل ۲ رسم شده است مشاهده می‌شود که متغیر X_1 بیشترین اهمیت سپس متغیر X_3 در رتبه دوم



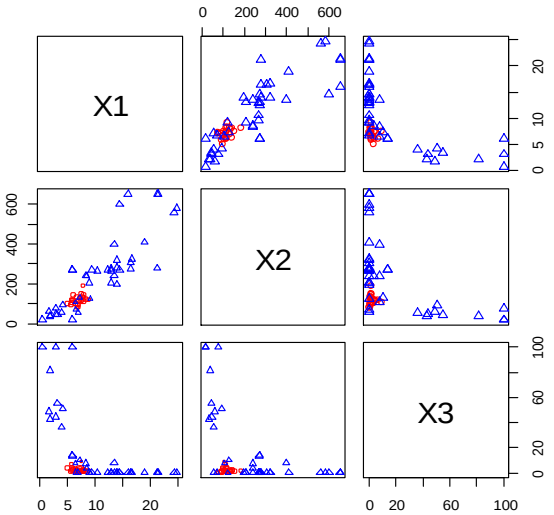
شکل۱: درخت تصمیم معمولی ساخته شده برای تعیین وضعیت غده تیروئید

جدول ۱: نتیجه رده‌بندی تعیین وضعیت غده تیروئید در درخت تصمیم معمولی و بوستینگ

نوع درخت تصمیم	مجموعه	ویژگی	حساسیت	نرخ خطای رده‌بندی
معمولی	آموزشی	۱	۰/۸۹۷	۰/۰۵۸
	آزمون	۰/۹۲	۰/۹۱	۰/۰۸۸
بوستینگ	آموزشی	۱	۱	۰
	آزمون	۱	۰/۹۵۵	۰/۰۲۹



شکل ۲: نمودار اهمیت متغیرهای پیشگو در تعیین وضعیت غده تیروئید



شکل ۳: نمودار جفت متغیرهای پیشگو در تعیین وضعیت غده تیروئید

اهمیت قرار دارد و در نهایت X_2 دارای کمترین اهمیت می‌باشد. همچنین در درخت‌های تصمیم بوستینگ می‌توان نمودار جفت‌ها را که یک تصویر از روابط جفت متغیرها در میان مجموعه داده‌ها به دست می‌دهد رسم نمود این نمودار در شکل ۳ نشان داده شده است. نمودارهای بالایی کلاس واقعی را برای هر رابطه جفت متغیرها نشان می‌دهد در حالی که نمودار پایین کلاس پیش‌بینی شده را برای هر مشاهده نشان می‌دهد. همچنین مشاهدات در نمودارهای پایین با برآورد احتمال کلاس مقیاس‌بندی شده‌اند. بنابراین اندازه نقطه نماینگر برآورد احتمال است بنابراین این نمودار کمک می‌کند نقاطی که برای الگوریتم بوستینگ رده‌بندی آن‌ها سخت است شناخته شوند.

بحث

معمولا هر تحقیق تلاشی است در جهت پیشرفت در یک حوزه از علم، هر تحقیق باید دنباله آخرین پژوهش‌های قبل از خود در آن زمینه باشد و همچنین نقطه شروعی باشد برای ارتقاء فعالیت‌های بعدی در این زمینه. در این بخش به تحلیل یافته‌های این تحقیق پرداخته می‌شود.

درخت‌های تصمیم با وجود کاربرد فراوان به دلیل ساختار ساده و قابلیت تفسیر آسان ابزارهای سودمندی در تحلیل داده‌ها به ویژه در علم پزشکی هستند اما این درخت‌ها مدل‌های دقیقی نیستند و اغلب با واریانس زیاد همراه می‌شوند [۴]. به طور کلی علاقه زیادی به استفاده از روش‌های رده‌بندی در تحقیقات کلینیکی وجود دارد. یک قاعده رده‌بندی دوتایی رده‌بندی نمونه‌ها به یکی از دو کلاس متقابلاً منحصربفرد بر اساس مشخصه‌های مشاهده شده از نمونه‌ها تخصیص می‌دهد. طبقات دوتایی معمول در تحقیقات پزشکی شامل مرگ/بقا، بیمار/غیربیمار، برگشت بیماری/بهبود بیماری و بستری شدن در بیمارستان/بستری نشدن در بیمارستان می‌باشد. در تحقیقات پزشکی درخت‌های رده‌بندی روش‌های رده‌بندی دوتایی رایج می‌باشند. در زمینه داده کاوی پیشرفت‌هایی در درخت‌های تصمیم کلاسیک انجام شده است با این وجود این پیشرفت‌ها در تحقیقات پزشکی شناخته شده نیستند و این روش‌های توسعه یافته مانند درخت‌های تصمیم بوستینگ به ندرت در تحقیقات پزشکی استفاده می‌شوند [۳]. هدف اصلی این مطالعه استفاده از الگوریتم بوستینگ جهت کاهش خطای درخت‌های تصمیم در تعیین وضعیت غده تیروئید به دو حالت غده تیروئید نرمال و غده تیروئید غیر نرمال بود برای مقایسه نتایج از نرخ خطای رده‌بندی و دو معیار حساسیت و ویژگی استفاده شد که در نهایت معلوم شد درخت‌های تصمیم بوستینگ بسیار بهتر از درخت‌های تصمیم معمولی عمل می‌کنند. از آن‌جا که ملاک انتخاب مدل در این تحقیق نتایج مجموعه آزمون است، در نمونه داده‌های غده تیروئید نرخ خطای رده‌بندی، حساسیت و ویژگی برای مجموعه آزمون در درخت‌های تصمیم معمولی به ترتیب ۰/۰۸۸، ۰/۹۱ و ۰/۹۲ به دست آمدند و در درخت‌های تصمیم بوستینگ سه معیار فوق به ترتیب ۰۲۹/، ۰/۹۵۵ و ۱ به دست آمدند. بنابراین درخت‌های تصمیم بوستینگ عملکرد بهتری از خود نشان دادند.

البته محدودیت‌هایی در استفاده از درخت‌های تصمیم بوستینگ وجود دارد. تفسیر درخت‌های رده‌بندی و رگرسیون معمولی ساده و شفاف است و با یک گرافیک دوبعدی قابل نمایش است با این وجود ترکیبات خطی از درخت‌ها این خاصیت مهم را از بین می‌برد و بنابراین باید به روش متفاوت دیگری تفسیر شوند [۱۷]. بسیاری از پزشکان به دلیل سادگی تفسیر درخت نهایی درخت‌های رگرسیون معمولی و درخت‌های تصمیم معمولی را جذاب می‌دانند همچنین پیش‌بینی‌ها و رده‌بندی‌هایی برای بیماران جدید به سادگی می‌تواند توسط پزشک انجام شود. اما درخت‌های تصمیم بوستینگ این توانایی را به پزشک نمی‌دهند که بتواند ساده و سریع رده‌بندی‌هایی را برای بیماران جدید انجام دهد. در انتخاب بین درخت‌های تصمیم معمولی و درخت‌های تصمیم بوستینگ محقق باید یک تعامل بین به دست آوردن دقت بیشتر در قدرت پیش‌بینی و از دست دادن قابل تفسیر بودن و افزایش دشواری در کابرد درخت بوستینگ نهایی برقرار کند [۳].

رده‌بندی نقش مهمی در پزشکی بازی می‌کند، وقتی پاسخ وجود یا عدم وجود یک بیماری خاص است رده‌بندی نمونه‌ها بر اساس حالت‌های بیماری به پزشکان اجازه می‌دهد که وقتی بیماری برای یک گروه از نمونه‌ها رد می‌شود از بررسی‌ها و درمان‌های بیشتر که به نظر موجه

می‌آمدند خودداری کند و وقتی وجود بیماری رد نمی‌شود بررسی‌ها و درمان‌های بعدی قابل توجیه می‌شوند و با جدیت بیشتری دنبال می‌شوند. رده‌بندی دقیق حالت‌های بیماری اجازه می‌دهد که از منابع محدود مراقبت‌های پزشکی بهتر استفاده شود. وقتی پاسخ، پیشامدی مانند مرگ و میر باشد آنگاه بسته به اینکه پاسخ تأیید و یا رد شود معالجات بعدی لازم الاجرا می‌شوند مثلاً یک مراقبت آرام‌بخش و مسکن ممکن است برای بیماری که احتمالاً به زودی می‌میرد تجویز شود همچنین ممکن است یک درمان تهاجمی برای بیمارانی که احتمالاً به زودی می‌میرند برای جلوگیری از مسیر طبیعی بیماری تجویز شود. بنابراین ممکن است پزشکان بیشتر از گذشته پذیرا درخت‌های تصمیم بوستینگ شوند که می‌تواند یک پاسخ دوتایی (مثلاً مرگ یا بقا) تولید کند به جای اینکه یک عدم قطعیت پنهان تولید شود برای وقتی که خروجی‌ها به شکل احتمال ارائه می‌شوند [۳]. نتیجه به دست آمده در این مطالعه همسو با بسیاری از مطالعات مشابه دیگر است برای مثال در مطالعه‌ای که تأثیر الگوریتم بوستینگ بر روی درخت‌های تصمیم معمولی و مدل‌های شبکه عصبی در رده‌بندی داده‌های شیمیایی را بررسی نموده است و از سه گروه داه شیمیایی مختلف جهت مقایسه نتایج استفاده شد الگوریتم بوستینگ در کاهش خطای رده‌بندی بسیار موفق عمل کرد [۱۷]، همچنین در مطالعه دیگری که بر روی بیماران بستری شده به مدت ۳۰ روز در بیمارستان انجام شده بود الگوریتم بوستینگ یک بهبود جزیی تا متوسط بر روی درختان تصمیم معمولی ایجاد نمود [۳]. برای پیش‌بینی میزان مرگ و میر در آسیب مغزی مطالعه‌ای درخت‌های تصمیم بوستینگ با چندین روش دیگر از جمله درخت‌های تصمیم معمولی بر روی یک نمونه شامل ۱۶۰۳ مورد آسیب مغزی را مورد بررسی قرار داده است نتایج نشان داد که درخت‌های بوستینگ تفاوت معنی‌داری در دقت پیش‌بینی نسبت به روش‌های دیگر دارند [۱۸].

نتیجه‌گیری

با توجه به یافته‌های این مطالعه می‌توان نتیجه‌گیری کرد که الگوریتم بوستینگ می‌تواند برای روش‌های رده‌بندی به کار رود تا عدم دقت درخت‌های تصمیم معمولی را تعدیل کند. در تعیین وضعیت غده تیروئید وقتی از معیار نرخ خطا رده‌بندی و یا حساسیت استفاده شد درخت‌های تصمیم بوستینگ بسیار بهتر از درخت‌های تصمیم معمولی عمل کردند اما وقتی از معیار ویژگی استفاده می‌شود درخت‌های تصمیم بوستینگ یک بهبود نسبتاً متوسط نسبت به درخت‌های تصمیم معمولی ایجاد می‌کنند. همچنین در درخت‌های تصمیم بوستینگ متغیر X_1 مهمترین متغیر در رده‌بندی داده‌ها تشخیص داده شد که این متغیر میزان هورمون تیروکسین می‌باشد در حالی که در درخت‌های تصمیم معمولی در بیشتر موارد فقط از دو متغیر X_2 (تری یدو تیرونین) و متغیر X_3 (تیروتروپین) برای تقسیم فضای متغیرها استفاده شد. بنابراین بطور کلی استفاده از الگوریتم بوستینگ جهت کاهش خطای رده‌بندی تعین وضعیت غده تیروئید پیشنهاد می‌شود. همچنین پیشنهاد می‌شود که استفاده از الگوریتم بوستینگ در درخت‌های رگرسیونی در زمینه تحلیل داده‌های پزشکی نیز بررسی شود، در این تحقیق از درخت‌های تصمیم بوستینگ برای تعیین وضعیت غده تیروئید نرمال از غیر نرمال استفاده شد می‌توان جهت ادامه کار این مطالعه از همین روش جهت تعیین غده تیروئید پرکار از غده تیروئید کم‌کار در نمونه‌های که برای آنها غده تیروئید غیرنرمال تشخیص داده شده است استفاده نمود. در تحلیل داده‌های غده تیروئید درخت تصمیم بوستینگ هورمون تیروکسن را مهمترین متغیر در رده‌بندی وضعیت غده تیروئید به دو حالت نرمال و غیر نرمال تشخیص داد که لازم است پزشکان در این زمینه بررسی بیشتر انجام دهند شاید از انجام آزمایشات اضافه در این زمینه جلوگیری شود.

تشکر و قدردانی

مقاله حاضر برگرفته از طرح پژوهشی با شماره ۸۸۶ است که در دانشگاه آزاد اسلامی واحد شوشتر انجام شده است و کلیه اعتبار طرح پژوهشی مذکور توسط دانشگاه آزاد اسلامی واحد شوشتر تأمین شده است. که بدین ترتیب نویسندگان مراتب تقدیر و تشکر خود را اعلام می‌دارند، همچنین نویسندگان مقاله از آزمایشگاه سلامت شهرستان شوشتر که داده‌های مطالعه را در اختیار آنها قرار داد مراتب سپاس خود را اعلام می‌دارند.

References

1. Fauci A, Wald B, Kasper D,“ et al”, Harrisson’s Principle of internal medicine, translated by Sajjadian KH, 2th printing, Arjmand press, Tehran, 2008, 188-335.
2. Basatini FM, Multilayer logistic discrimination using neural networks, (Dissertation) Computer and mathematical science college of Shahid Chamran University, June 2006.
3. Austin PC, Lee DS, Boosted classification trees result in minor to modest improvement in the accuracy in classifying cardiovascular outcomes compared to conventional classification trees, Am J cardiovasc Dis. 2011;1(1), 1-15.
4. Ripley BD, Pattern recognition and neural networks, 7th printing, Cambridge University Press, 2004, 213-235.
5. Austin PC, A comparison of classification and regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality, Stat. med 2007; 26, 2937-2957.
6. Hastie T, Tibshirani R, Freidman JH, The elementary of statistical learning: Prediction, Inference and Data-Mining, Springer-verla, 2001, 266-323.
7. Quinlan JR, C4/5: programs for machine learning, San Mateo, CA: Morgan Kaufman,1993.
8. Buhlman P and Yu B, Boosting with L2-loss: regression and classification, JASA 2003; 98, 324-349.
9. Friedman JH, Greedy function approximation: A gradient boosting machine, Annals of Statistics 2001; 29, 1189-1232.
10. Breiman L, Friedman JH, Olsen RA and Ston CJ, Classification and regression trees, Monterey, CA: Wadsworth and Books/Cole,1984.
11. Freund Y and Schapire R, A decision-theoretic generalization of online learning and an application to boosting, JCSS 1997; 55, 119-139.
12. Culp M, Johnson K and Michailidis G, ada: An R package for stochastic boosting, J STAT SOFTW 2006; 17(2):1-27.
13. Dettling M and Buhlman P, Boosting for tumor classification with gene expression data, Bioinformatics 2003;19(9): 1061-1069.
14. Benson PJ, Beedie SA, Shephard E, Giegling I, Rujescu D Clair D St, Simple viewing tests can detect eye movement abnormalities that distinguish schizophrenia cases from controls with exceptional accuracy, Biol Psychiatry, 2012; 72: 716-724.
15. Pawlowski KM, Maciejewski H, Majchrzak K, Dolka I, Mol J A, Motyl T, Five markers useful for the distinction of canine mammary malignancy, BMC veterinary research, 2013;9:138:1-9.
16. Mease D, Abraham JW, Andreas B, Boosted classification trees and class Probability /quantile estimation, JMLR,2007; 8: 409-439.
17. He P, Xu CJ, Liang YZ , Fang KT, Improving the classification accuracy in chemistry via boosting technique, CILS 2004; 70, 39-46.
18. Sut S, Simsek O, Comparison of regression tree data mining methods for prediction of mortality in head injury, ESA 2011; 38, 15534-15539.

Decreasing in misclassification of determination thyroid disease in Shoushtar town using tree boosting algorithm

*mohammadi basatini F[1] *, reyhani niya B[2]*

[1]Instructor,Department of mathematic, Shoushtar Branch,Islamic Azad University,Shoushtar,Iran
[2]Instructor,Department of mathematic, Shoushtar Branch,Islamic Azad University,Shoushtar,Iran
***Corresponding Author:** Islamic Azad University,Shoushtar,Iran
Email: fe.mohamadi91@gmail.com

Abstract

Background & Objectives: *Thyroid is a vital gland, which affect all of the body oragans such as heart, digestive system, kidney and so on. The intention of this research is to decreas in wrong determination of normal thyroid gland from abnormal using boosting algorithm. This algorithm is a powerful method in diagnosis and prognosis. It iteratively grows base classifer on a sequence of reweighted datasets then takes a linear combination of consequencs and we hope improves accuracy at final.*

Material & Methods: *A total of 103 patients' data corrolated to November 2010 until November 2011 from Shoushtar salamat laboratory were analyzed for detemination thyroid gland state. Conventional decision trees and boosting decision trees were made for diagnosis normal thyroid gland from abnormal thyroid gland using R softwere vedersion 3.0.1.*

Results: *Our findings revealed that for conventional decision trees misclassification rate , sensitivity and specificity with test set were 0.088 , 0.91 and 0.92 respectively .However these figures considered by boosting desion trees were 0.029 , 0.955 and 1 crrespondingly.*

Conclution: *The boosting decision trees had possibily superior sucsses in diagnosis normal tiroid gland ftom unnormal . So using boosting decisin trees propose in determination thyroid gland state.*

Keyword: *boosting algorithm, misclassification rate, sensitivity, specificity.*

Original Article

Journal of North Khorasan University 2015;7(2): 381-391

Recivied: 18 May 2014
Revised:26 Jul 2014
Accepted:14 Oct 2014