مقاله مروری

فصلنامهٔ بیماری‌های پستان ایران، سال ششم، شماره اول، بهار ۱۳۹۲

مروری بر ۷ الگوریتم برتر داده کاوی در پیش‌بینی بقا، تشخیص و عود بیماران مبتلا به سرطان پستان

لیلا قاسم احمد*: کارشناس ارشد مدیریت فناوری اطلاعات، دانشگاه آزاد اسلامی، واحد علوم و تحقیقات تهران

چکیده

مقدمه: پیش بینی تشخیص، بقا و عود بیماران مبتلا به سرطان پستان، همواره از چالش‌های مهم برای محققین و پزشکان بوده است. امروزه به مدد علوم بیوانفورماتیک، امکان رفع این چالش‌ها با بهره‌گیری از اطلاعات قبلی ثبت شده از بیماران تا حدود زیادی محقق گردیده است. با تکنولوژی‌های کم هزینه سخت افزاری و نرم افزاری، داده‌ها با کیفیت بهتر و در حجم‌های بالاتر به صورت خودکار ذخیره می‌گردند و به کمک تجزیه و تحلیل بهتر آنها، این حجم عظیم از داده‌ها به صورتی کارآمد و موثرتر پردازش می‌شوند. هدف اصلی این مقاله معرفی تعدادی از الگوریتم‌های پرکاربرد و شناخته شده داده‌کاوی در سرطان پستان است.

روش بررسی: الگوریتم‌های داده‌کاوی، مدل‌های بهینه‌ای هستند که در پیش بینی تشخیص، بقا و عود سرطان پستان به کار رفته و دقت قابل توجهی از خود نشان داده‌اند. نتایج حاصل از این الگوریتم‌ها، نه تنها به پزشکان در تصمیم گیری بهتر کمک می‌کند بلکه باعث آشکار شدن برخی از الگوهای پنهان و ناشناخته می‌شود که شاید توجه خاصی به آنها معطوف نبوده است. این الگوریتم‌ها شامل: شبکه‌های عصبی مصنوعی (Artificial Neural Networks/ANNs)، درختان تصمیم گیری (Decision Trees)، شبکه‌های بیزی (Bayes Nets)، بیزی ساده (Naive Bayes)، رگرسیون لجستیک (Logistic Regression)، بردار پشتیبان ماشین (Support Vector Machine) و روش‌های ترکیبی درختان تصمیم و شبکه‌های بیزی (Decision Tree with Naive Bayes) هستند. از این الگوریتم‌ها، برای دسته بندی، خوشه بندی، یادگیری آماری که مهم‌ترین روش‌های داده کاوی هستند، استفاده می‌شود.

یافته‌ها: در این مقاله، ۷ الگوریتم برتر داده‌کاوی در پیش بینی بقا، تشخیص و عود بیماران مبتلا به سرطان پستان معرفی می‌گردند و با معرفی هر الگوریتم، پیشینه‌ای از تحقیقات صورت گرفته در سرطان پستان به کمک الگوریتم مورد نظر، نتایج حاصل از آن و همچنین تحقیقات فعلی در حال انجام در این خصوص ارائه می‌شود. درختان تصمیم و ماشین بردار پشتیبان، در تحقیقات مختلف انجام شده، معمولا نتایج بهتر و دقیق‌تری در زمینه دقت، حساسیت و ویژگی ارایه کرده‌اند.

نتیجه‌گیری: موفقیت این الگوریتم‌ها، به فاکتورهای متعددی چون وجود متغیرهای مورد نیاز، بزرگتر بودن پایگاه داده، کم بودن تعداد داده‌های مفقوده و دسترسی به داده‌های صحیح و درست بستگی دارد.

واژه‌های کلیدی: سرطان پستان، الگوریتم‌های داده‌کاوی، پیش بینی، شبکه‌های عصبی مصنوعی، درختان تصمیم‌گیری، شبکه‌های بیزی، رگرسیون لجستیک، بردار پشتیبان ماشین.

* نشانی نویسنده پاسخگو: دانشگاه آزاد اسلامی، واحد علوم و تحقیقات تهران، لیلا قاسم احمد.
نشانی الکترونیک: lga_77@yahoo.com

[DOR: 20.1001.1.17359406.1392.6.1.6.7]

مقدمه

علم داده‌کاوی، به کشف الگوهای پنهان و ناشناخته در میان حجم عظیمی از داده‌ها می‌پردازد که گاه از دید متخصصین علوم پزشکی، پنهان می‌مانند. در این میان روش‌های مختلفی در زمینه پیش بینی، بقا و عود بیماران مبتلا به سرطان پستان، به کار رفته‌اند که گاهی نتایج حاصل از آنها به طور باورنکردنی پشتیبان تصمیمات پزشکان بوده‌اند. قرار نیست که این روش‌ها جایگزین تصمیمات متخصصین و محققین گردند، اما با بهره‌گیری از الگوهای خاص و تکرار شونده، می‌توانند مدد رسان آنها را در شرایط حساس باشند(۱).

مواد و روش‌ها

الگوریتم‌های داده‌کاوی:

الگوریتم‌های مختلفی در علم داده‌کاوی به کار می‌روند که، طبق تحقیقات متعدد به عمل آمده در سال‌های اخیر برخی از آنها نتایج بسیار قابل قبولی در دنیای واقعی، از خود نشان داده‌اند. الگوریتم‌های داده‌کاوی به کمک نرم افزارهای متداول در داده‌کاوی مثل نرم افزار «کلمنتاین» یا «وکا» پیاده سازی می‌شوند. «دقت[۱]»، «حساسیت[۲]» و «ویژگی[۳]» سه مورد حایز اهمیت جهت ارزیابی نتایج حاصله هستند(۳).

نمی‌توان به این سوال که کدام الگوریتم بهترین الگوریتم به کار رفته در مورد سرطان پستان است، پاسخ دقیقی داد. علت این موضوع، به هدف از تحقیق، شرایط تحقیق، تعداد داده‌ها و متغیرهای موجود و همچنین بسیاری از فاکتورهای دیگر برمی‌گردد. همچنین، اهداف اصلی تحقیق از قبیل پیش‌بینی کردن طول عمر بیماران، تشخیص بیماری، عود بیماری نیز می‌تواند الگوریتم مناسب را تغییر دهد. در خصوص شواهد مورد نیاز جهت پرکاربرد بودن این الگوریتم‌ها می‌توان گفت که، الگوریتم‌های زیادی برای داده کاوی تعریف شده است که تعدادی از آنها نیز حاصل ترکیبی از چند الگوریتم هستند. از میان مدل‌های شناخته شده، برخی مدل‌ها طی سالیان اخیر در داده‌کاوی پزشکی و سرطان (پستان)، کاربرد بیشتری داشته است. همچنین نمونه‌های کلی از کاربرد داده‌کاوی در پزشکی به کمک الگوریتم‌های آن نیز در منابع متعددی ذکر شده است. مقاله مروری اخیر با استناد به این تحقیق‌ها گردآوری شده است و مثال‌های ذکر شده برای هر مدل که در انتهای هر مدل و الگوریتم آمده است، با ذکر منبع و نمونه تحقیق انجام شده در خصوص آن مدل است(۴ الی۸).

شبکه عصبی مصنوعی[۴] :

یک شبکه عصبی مصنوعی ایده‌ای است برای پردازش اطلاعات که از سیستم عصبی زیستی الهام گرفته شده و مانند مغز به پردازش اطلاعات می‌پردازد. عنصر کلیدی این ایده، ساختار جدید سیستم پردازش اطلاعات است. این سیستم از شمار زیادی عناصر پردازشی فوق‌العاده بهم پیوسته تشکیل شده که برای حل یک مسأله با هم هماهنگ عمل می‌کند.ANNها، نظیر انسان‌ها، با مثال یاد می‌گیرند. یک ANN برای انجام وظیفه‌ای مشخص، مانند شناسایی الگوها و دسته بندی اطلاعات، در طول یک پروسه یادگیری، تنظیم می‌شود. در سیستم‌های زیستی یادگیری با تنظیماتی در اتصالات سیناپسی که بین اعصاب قرار دارد همراه است. این روش ANN ها هم می‌باشد(۹).

شبکه‌های عصبی مصنوعی و به زبان ساده‌تر شبکه‌های عصبی، سیستم‌ها و روش‌های محاسباتی نوینی هستند برای یادگیری ماشینی، نمایش دانش و در انتها اعمال دانش به دست آمده در جهت بیش‌بینی پاسخ‌های خروجی از سامانه‌های پیچیده. ایده اصلی این گونه شبکه‌ها (تا حدودی) الهام‌گرفته از شیوه کارکرد سیستم عصبی زیستی، برای پردازش داده‌ها، و اطلاعات به منظور یادگیری و ایجاد دانش قرار دارد. عنصر کلیدی این ایده، ایجاد ساختارهایی جدید برای سامانه پردازش اطلاعات است. این سیستم از شمار زیادی عناصر پردازشی فوق‌العاده بهم پیوسته با نام نورون تشکیل شده که برای حل یک مسأله با هم هماهنگ عمل می‌کند.

[۱] دقت: عبارت است از تعداد نمونه‌هایی که به درستی تشخیص داده می شوند، نسبت به کل نمونه‌ها.

[۲] حساسیت: احتمال پیش‌بینی درست وضعیت مورد نظر توسط الگوریتم‌ها (مثبت واقعی تقسیم بر منفی کاذب+ مثبت واقعی).

[۳] ویژگی: احتمال پیش‌بینی درست عدم وجود وضعیت مورد نظر توسط الگوریتم‌ها (منفی واقعی تقسیم بر مثبت کاذب+ منفی واقعی).

[4] Artificial Neural Network -ANN

DOR: 20.1001.1.17359406.1392.6.1.6.7]

اجزای یک شبکه عصبی:

یک شبکه عصبی شامل اجزای سازنده لایه‌ها و وزن‌ها است. رفتار شبکه نیز وابسته به ارتباط بین اعضا است. در حالت کلی در شبکه‌های عصبی سه نوع لایه نورونی وجود دارد:

لایه ورودی: دریافت اطلاعات خامی که به شبکه تغذیه شده‌است.

لایه‌های پنهان: عملکرد این لایه‌ها به وسیله ورودی‌ها و وزن ارتباط بین آنها و لایه‌های پنهان تعیین می‌شود. وزن‌های بین واحدهای ورودی و پنهان تعیین می‌کند که چه وقت یک واحد پنهان باید فعال شود.

لایه خروجی: عملکرد واحد خروجی به فعالیت واحد پنهان و وزن ارتباط بین واحد پنهان و خروجی بستگی دارد.

شبکه‌های تک لایه و چند لایه‌ای نیز وجود دارند که سازماندهی تک لایه که در آن تمام واحدها به یک لایه اتصال دارند بیشترین مورد استفاده را دارد و پتانسیل محاسباتی بیشتری نسبت به سازماندهی‌های چند لایه دارد. در شبکه‌های چند لایه واحدها به وسیله لایه‌ها شماره‌گذاری می‌شوند (به جای دنبال کردن شماره گذاری سراسری). هر دو لایه از یک شبکه به وسیله وزن‌ها و در واقع اتصالات با هم ارتباط می‌یابند. در شبکه‌های عصبی چند نوع اتصال و یا پیوند وزنی وجود دارد.

پیشرو: بیشترین پیوندها از این نوع است که در آن سیگنال‌ها تنها در یک جهت حرکت می‌کنند. از ورودی به خروجی هیچ بازخوردی (حلقه) وجود ندارد. خروجی هر لایه بر همان لایه تاثیری ندارد.

پسرو: داده‌ها از گره‌های لایه بالا به گره‌های لایه پایین بازخورانده می‌شوند.

جانبی: خروجی گره‌های هر لایه به عنوان ورودی گره‌های همان لایه استفاده می‌شوند. در شکل زیر ساختار یک نرون طبیعی و یک نرون مصنوعی نشان داده شده است (۱۰).

یکی از مثال‌های کاربرد شبکه عصبی در پیش بینی بقا بیماران مبتلا به سرطان پستان بوده است. در تحقیقی که در سال ۲۰۰۷ توسط گروهی از محققین دانشگاه توکیو انجام گردید، از شبکه‌های عصبی مصنوعی در این خصوص استفاده شد. از ۳۷ هزار و ۲۵۶ بیماری که اطلاعات آنها در یک دوره پنج ساله از پایگاه داده استخراج شده بود. ۸۱ متغیر از قبیل سن، جنسیت، مرحله بیماری، وضعیت ازدواج و .... انتخاب گردیدند. نتایج مطالعات میزان دقت این مدل را در پیش بینی بقا بیماران ۸۴.۵٪ را نشان داد. یعنی مدل داده‌کاوی به کار رفته بر مبنای شبکه‌های عصبی به میزان ۸۴.۵٪ توانست میزان بقای بیماران مبتلا به سرطان پستان را به درستی پیش بینی کند(۱۱).

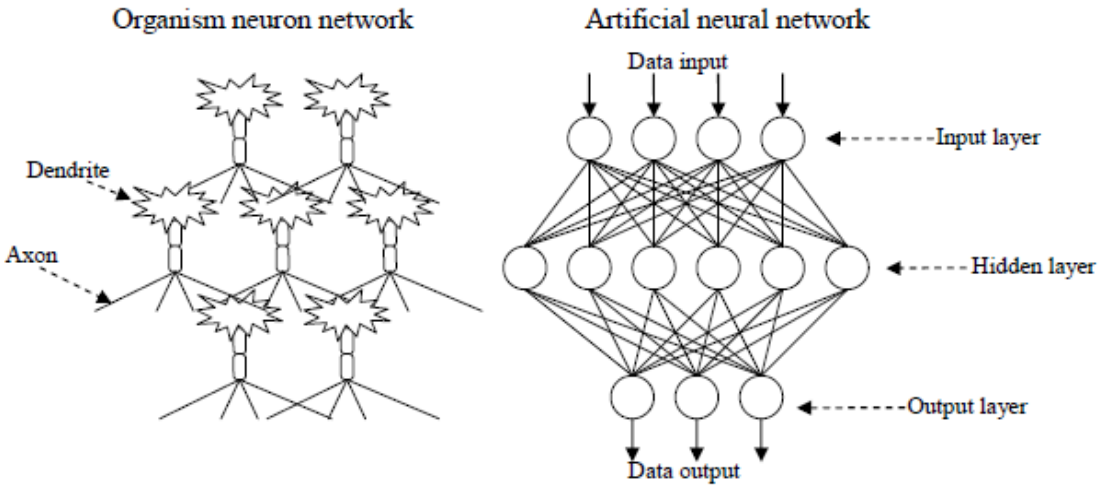
درخت تصمیم[5] :

درخت تصمیم درختی است که در آن نمونه‌ها را به نحوی دسته‌بندی می‌کند که از ریشه به سمت پائین رشد می‌کنند و در نهایت به گره‌های برگ می‌رسد:

- هر گره داخلی یا غیر برگ[6] با یک ویژگی[7] مشخص می‌شود. این ویژگی سوالی را در رابطه با مثال ورودی مطرح می‌کند.
- درهر گره داخلی به تعداد جواب‌های ممکن با این سوال شاخه[8] وجود دارد که هر یک با مقدار آن جواب مشخص می‌شوند.
- برگ‌های این درخت با یک کلاس و یا یک دسته از جواب‌ها مشخص می‌شوند.

یک درخت معمولا تشکیل شده از ریشه، شاخه‌ها، گره‌ها (جایی که شاخه‌ها منشعب می‌شوند) و برگ‌ها. درخت‌های تصمیم هم به صورت مشابه تشکیل شده‌اند از گره‌ها که با دایره نشان داده می‌شوند و شاخه‌ها که با پاره خط‌های اتصال بین گره‌ها نشان داده می‌شوند. درخت تصمیم را به منظور سادگی در رسم، معمولا از چپ براست یا از بالا به پایین رسم می‌کنند به طوری که ریشه در بالا قرار بگیرد. گره اول را ریشه می‌گویند. انتهای یک زنجیره «ریشه- شاخه- گره-گره» را یک «برگ» می‌نامند. از هر یک از گره‌های داخلی (یعنی هر گره‌ای که برگ نباشد)، دو یا چند شاخه دیگر می‌توانند منشعب شوند. هر گره مربوط به یک خصوصیت معین است و شاخه‌ها به معنای بازه‌ای از مقادیر هستند. این بازه‌های مقادیر باید بخش‌های مختلف مجموعه مقادیر معلوم برای خصوصیت‌ها را به‌دست دهند. هنگامی که دقیقا دو شاخه از یک گره داخلی منشعب شوند (چنین درختی را درخت دو

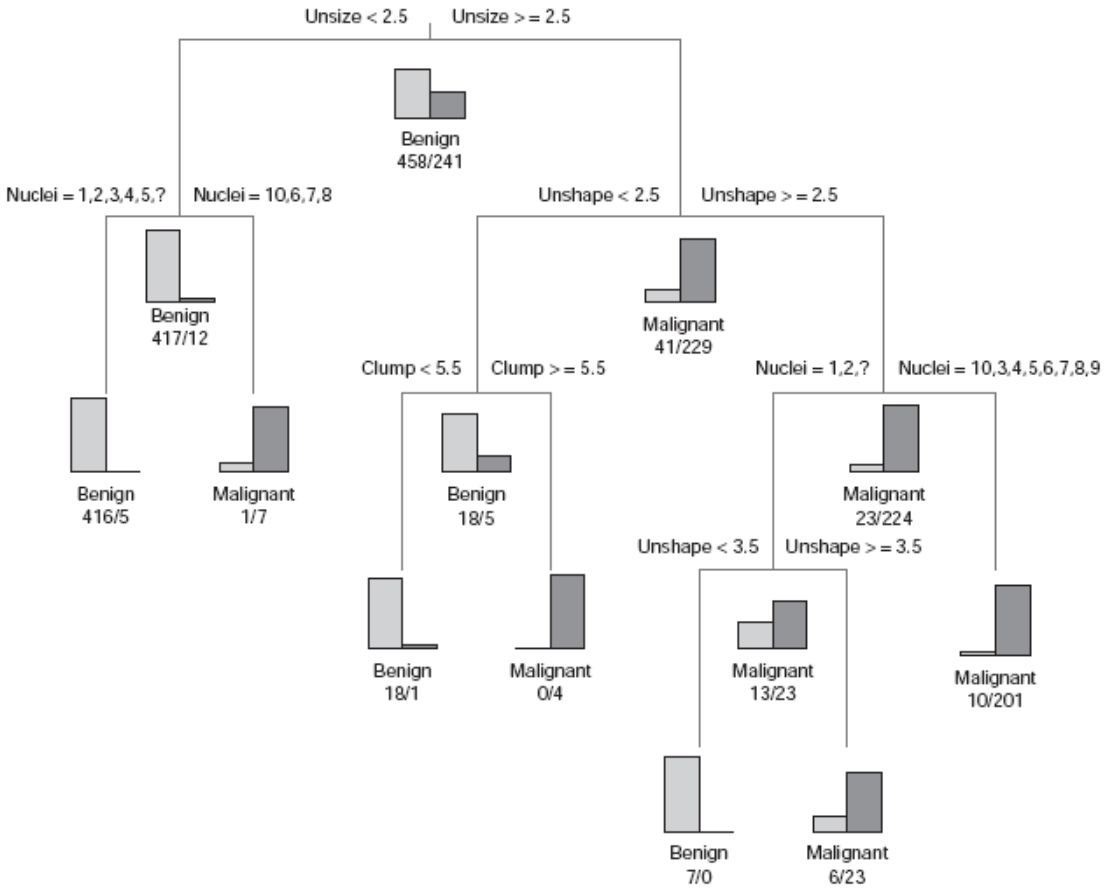
[5] Decision Tree
[6] none leaf
[7] attribute
[8] branch



شکل ۱: ساختار یک نرون مصنوعی و یک نرون طبیعی

حالته می‌گویند)، هر یک از این دو شاخه می‌تواند نماینده یک عبارت درست یا غلط برحسب خصوصیات معلوم باشد (۱). شکل دو یک درخت تصمیم به کار رفته در پیش بینی سرطان پستان را نشان می‌دهد.

در این مثال، داده‌های ۶۹۹ بیمار که دارای سرطان خوش خیم یا بدخیم بودند بررسی شد. و از ۹ متغیر جمع آوری شده از بیوپسی استفاده گردید. سپس دسته‌بندی مورد نظر مطابق شکل ایجاد شده و درخت‌ها شکل گرفتند. در



شکل ۲: نمونه‌ای از به‌کارگیری درخت تصمیم در سرطان پستان

DOR: 20.1001.1.17359406.1392.6.1.6.7]

اینجا هدف پیش بینی بیماری، بر اساس وضعیت ۹ متغیر جمع آوری شده از بیوپسی است. شاخه‌های جدید، مطابق درخت تصمیم شکل دو ایجاد شده و به نتیجه نهایی می‌رسند(۱۳و۱۲).

کاربرد درخت تصمیم:

- درخت تصمیم در مسایلی کاربرد دارد که بتوان آنها را بصورتی مطرح نمود که پاسخ واحدی به صورت نام یک دسته یا کلاس ارائه دهند.
- برای مثال می‌توان درخت تصمیمی ساخت که به این سوال پاسخ دهد: بیماری مریض کدام است؟ و یا درختی ساخت که به این سوال پاسخ دهد: آیا مریض به سرطان پستان مبتلاست؟
- برای مسائلی مناسب است که مثال‌های آموزشی به صورت زوج (مقدار- ویژگی) مشخص شده باشند.
- تابع هدف دارای خروجی با مقادیر گسسته باشد. مثلا هر مثال با بله و خیر تعیین شود.
- نیاز به توصیف‌گر فصلی باشد.

این موارد، از کاربردهای مهم درخت تصمیم هستند(۱).

ویژگی‌های درخت تصمیم:

- برای تقریب توابع گسسته به کار می‌رود.
- نسبت به نویز داده‌های ورودی مقاوم است.
- برای داده‌های با حجم بالا کاراست از این رو در داده-کاوی استفاده می‌شود.
- می‌توان درخت را به صورت قوانین if-then نمایش داد که برای استفاده قابل فهم است.
- امکان ترکیب عطفی(AND) و فصلی(OR) فرضیه‌ها را می‌دهد.
- در مواردی که مثال‌های آموزشی که فاقد همه ویژگی‌ها هستند نیز قابل استفاده است.

درخت‌های «و، یا» که نام دیگر آنها درختان تصمیم است، نمونه‌ها را با مرتب کردن آنها در درخت از گره ریشه به سمت گره‌های برگ دسته بندی می‌کنند. درخت تصمیم درختی است که در آن نمونه‌ها را به نحوی دسته بندی می‌کند که از ریشه به سمت پائین رشد می‌کنند و در نهایت به گره‌های برگ می‌رسد. هر گره داخلی یا غیر برگ با یک ویژگی مشخص می‌شود. این ویژگی سوالی را در رابطه با مثال ورودی مطرح می‌کند. درهر گره داخلی به تعداد جواب‌های ممکن با این سوال شاخه وجود دارد که هر یک با مقدار آن جواب مشخص می‌شوند. برگ‌های این درخت با یک کلاس و یا یک دسته از جواب‌ها مشخص می‌شوند. علت نامگذاری آن با درخت تصمیم این است که این درخت فرایند تصمیم گیری برای تعیین دسته یک مثال ورودی را نشان می‌دهد. درخت تصمیم در مسائلی کاربرد دارد که بتوان آنها را به صورتی مطرح نمود که پاسخ واحدی به صورت نام یک دسته یا کلاس ارائه دهند. همچنین برای مسائلی مناسب است که مثال‌های آموزشی به صورت زوج (مقدار- ویژگی) مشخص شده باشند. تابع هدف دارای خروجی با مقادیر گسسته باشد. مثلاً هر مثال با بله و خیر تعیین شود و یا نیاز به توصیف‌گر فصلی باشد (۱ و ۱۴).

ماشین‌های بردار پشتیبان[9] :

ماشین بردار پشتیبان یکی از روش‌های یادگیری با نظارت است که از آن برای طبقه‌بندی استفاده می‌کنند. در حالی که روش‌هایی مانند درخت تصمیم‌گیری را نمی‌توان به راحتی در مسائل مختلف به کار برد. این روش از جمله روش‌های نسبتاً جدیدی است که در سال‌های اخیر کارایی خوبی نسبت به روش‌های قدیمی‌تر برای طبقه‌بندی از جمله شبکه‌های عصبی پرسپترون نشان داده است. در شکل زیر نمونه‌ای از مدل SVM در تشخیص سرطان پستان معرفی شده است. حل معادلهٔ پیدا کردن خط بهینه برای داده‌ها به وسیله روش‌های QP که روش‌های شناخته شده‌ای در حل مسائل محدودیت‌دار هستند صورت می‌گیرد. قبل از تقسیم خطی برای اینکه ماشین بتواند داده‌های با پیچیدگی بالا را دسته‌بندی کند داده‌ها را به وسیله تابع phi به فضای با ابعاد خیلی بالاتر می‌بریم. برای اینکه بتوانیم مساله ابعاد خیلی بالا را با استفاده از این روش‌ها حل کنیم از قضیه دوگانی لاگرانژ برای تبدیل مساله مینیمم‌سازی مورد نظر به فرم دوگانی آن از تابع ساده‌تری به نام تابع هسته استفاده می‌کنیم. از توابع هسته مختلفی از جمله هسته‌های نمایی، چند جمله‌ای و سیگموید می‌توان استفاده نمود.

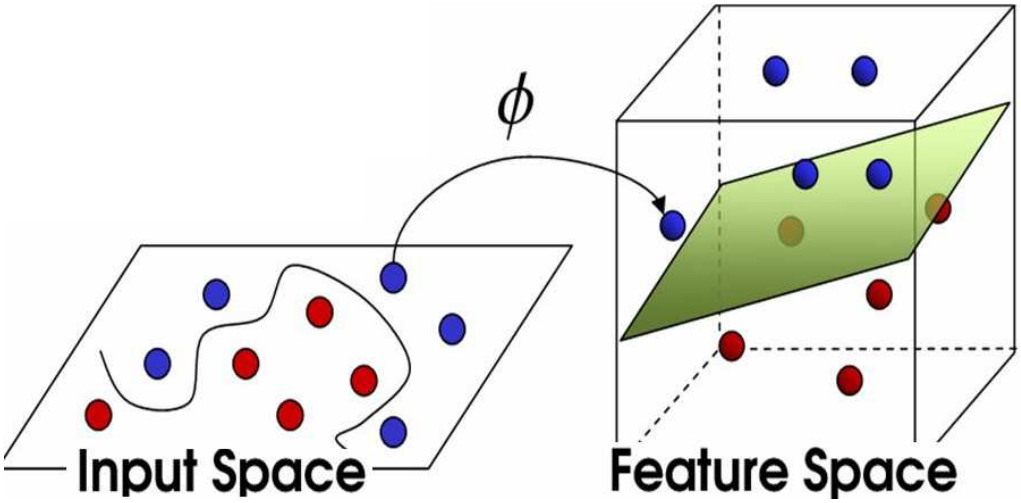
مبنای کاری دسته‌بندی کننده SVM دسته‌بندی خطی داده‌ها است و در تقسیم خطی داده‌ها سعی می‌شود خطی

[9] Support Vector Machine- SVMs

[DOR: 20.1001.1.17359406.1392.6.1.6.7]

انتخاب شود که حاشیه اطمینان بیشتری داشته باشد (۱ و ۱۵).

این روش زمان مورد نیاز جهت تشخیص را به میزان حیرت آوری کاهش می‌دهد(۱۶).



شکل ۳: نمونه‌ای از مدل SVM در تشخیص سرطان پستان

در تحقیقی که در سال ۲۰۰۵ در کشور تایوان انجام گردید، اثبات پاتولوژیک بودن تعدادی از تومورهای جامد[10] در سونوگرافی، توسط روش ماشین بردار پشتیبان انجام شد. در این مطالعه دو پایگاه داده مشتمل بر تصاویر اولتراسونیک، ارزیابی گردیدند. از مدل داده‌کاوی SVM به روش طبقه‌بندی استفاده شد و بافت‌های توموری به دو دسته خوش خیم و بدخیم تقسیم شدند. پایگاه داده اول شامل۱۴۰ مورد تصویر (۵۲ مورد بدخیم و ۸۸ مورد خوش خیم) بود و پایگاه دوم نیز شامل ۲۵۰ مورد تصاویر از ندول‌های جامد (۵۲ مورد بدخیم و ۸۸ مورد خوش خیم). ناحیه مورد نظر[11] سونوگرافی و ویژگی‌های بافتی[12] برای ایجاد دو کلاس یا طبقه مورد استفاده قرار گرفتند. نتایج نشان داد که تشخیص بیماری در این روش، بسیار سریع‌تر از شبکه‌های عصبی پرسپترون بود که این موضوع خود برای پزشکان، حایز اهمیت زیادی بود.

با رشد و گسترش دادن پایگاه داده، می‌توان تصاویر اولتراسونیک جدیدی را جمع آوری و به عنوان موارد منبع قابل ارجاع در حین انجام تشخیص بیماری استفاده نمود.

شبکه‌های بیزی[13] و بیزی ساده[14] :

هدف کلاس‌بندی این است که یک مورد را بر پایه مقادیر متغیرهای صفات گوناگون به یک کلاس نسبت دهد. بسیاری از روش‌های کلاس بندی تلاش می‌کنند تابع روشنی از مجموعه وابسته به مقادیر صفات به یک برچسب کلاس بسازند. یادگیری در کلاس‌بندی Bayesian یعنی تخمین زدن توزیع احتمالات وابسته. پس از اینکه چنین تخمینی را ساختیم، مقادیر را کلاس‌بندی می‌کنیم و کلاسی را که احتمال بیشتری دارد معین می‌نماییم.

یک روش یادگیری بسیار عملی روش Naive Bayes learner است. در کاربردهایی نظیر دسته بندی متن و تشخیص پزشکی این روش کارائی قابل مقایسه‌ای با شبکه‌های عصبی و درخت تصمیم دارد. این روش در مسایلی کاربرد دارد که:

- نمونه x توسط ترکیب عطفی ویژگی‌ها قابل توصیف بوده و این ویژگی‌ها به صورت شرطی مستقل از یکدیگر باشند.

[10] solid
[11] ROI (Region of interest)
[12] textual features
[13] Bayes Nets
[14] Naive Bayes

DOR: 20.1001.1.17359406.1392.6.1.6.7]

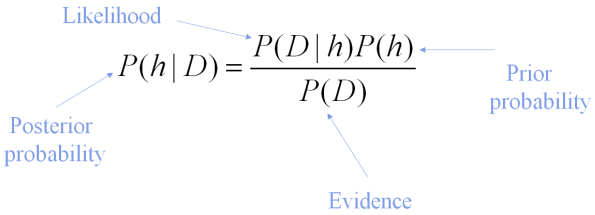
- تابع هدف f(x) بتواند هر مقداری را از مجموعه محدود v داشته باشد.
- مجموعه مثال‌های آموزشی نسبتاً زیادی در دست باشد.

سنگ بنای یادگیری بیزی را تئوری بیز تشکیل می‌دهد. این تئوری امکان محاسبه احتمال ثانویه را بر مبنای احتمالات اولیه می‌دهد:

P(h)= احتمال اولیه‌ای که فرضیه h قبل از مشاهده مثال آموزشی D داشته است(prior probability). اگر چنین احتمالی موجود نباشد می‌توان به تمامی فرضیه‌ها احتمال یکسانی نسبت داد.

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

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



در روش یادگیری Naive Bayes Classifier مقادیر مختلف (P(vj و P(ai| vj با استفاده از دفعات تکرار آنها تخمین زده می‌شود. مجموعه این تخمین‌ها فرضیه‌ای را تشکیل می‌دهد که با استفاده از رابطه زیر برای دسته-بندی داده جدید به کار می‌رود(۱۶و۱).

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_{i=1}^{n} P(a_i \mid v_j)$$

کلاس‌بندی بیزی به روش‌های مختلفی انجام می‌شود. در نمونه‌ای از مدل‌سازی ایجاد شده به کمک شبکه‌های بیزی، پیش بینی شدت وخامت توده‌های پستانی به کمک روش بیزی بررسی گردید. پایگاه داده مملو از داده‌های مفقوده و گمشده بود. در این مطالعه از دو روش مختلف کلاس‌بندی توسط مجموعه‌های بیزی استفاده شد. هدف از این تحقیق، پیشنهاد الگوریتمی برای کمک به پزشکان در تصمیم‌گیری‌هایشان بود. به این منظور که آیا در ضایعات مشکوک دیده شده در تصویر ماموگرام برای بیمار بیوپسی انجام شود یا اینکه به جای بیوپسی پیگیری و فالوآپ کوتاه مدت. علی‌رغم اینکه ماموگرافی موثرترین و در دسترس‌ترین ابزار برای غربالگری سرطان پستان است، ماموگرام‌ها قادر به کشف همه موارد سرطان پستان نیستند. دقت پیش بینی مجموعه‌های بیزی در تحقیق مذکور معادل ۹۱.۸۳٪ بود. نتایج تجربی نشان داد که شبکه‌های بیزی، می‌توانند تکنیک‌های قابل رقابت با شبکه‌های عصبی پرسپترون نیز باشند و همچنین به عنوان یک روش آلترناتیو و جایگزین سایر تکنیک‌ها در سرطان پستان به کار روند(۱۸).

رگرسیون لجستیک[15]:

رگرسیون لجستیک یک مدل آماری رگرسیون برای متغیرهای وابسته دودویی است. این مدل را می‌توان به عنوان مدل خطی تعمیم یافته‌ای که از تابع لوجیت به عنوان تابع پیوند استفاده می‌کند و خطایش از توزیع چند جمله‌ای پیروی می‌کند، به حساب آورد. در حالی که رگرسیون لجستیک یک مدل بسیار قدرتمند محسوب می‌شود، بسته به تجربه مدل‌ساز در زمینه در ارتباط با داده‌ها و آنالیز داده‌ها دارد و مدل‌ساز باید براساس تجربه پاسخ صحیح را در رابطه با متغیرها مشخص کند(۱۹).

یکی از نمونه مدل‌های ایجاد شده در به کارگیری رگرسیون لجستیک، آنالیز سرطان پستان با استفاده از رگرسیون لجستیک بوده است. در این مطالعه، تشخیص سرطان پستان از طریق ماموگرام‌ها با استفاده از این روش انجام شد. رادیولوژیست‌ها می‌توانند برای قضاوت مناسب در مورد وجود سرطان پستان، از این روش استفاده کنند. داده‌ها از پرسشنامه‌هایی که توسط رادیولوژیست‌ها گرداوری شده و حاصل مشاهدات آنها بیماران بود. نتایج به دست آمده به کمک روش رگرسیون لجستیک ارزش فاکتورهای مهم و اثرگذار را در سرطان پستان نشان داد. دسته بندی به دست آمده از ۱۷۶ نمونه و با استفاده از متغیرهای به دست آمده از تاریخچه بیمار شامل وجود توده، سابقه بیماری در بستگان درجه یک، سن یائسگی، حجم توده و همچنین از ماموگرام یعنی، ضایعات موجود در پستان[16]، کلسیفیکاسیون و ... انجام شد. متغیرهای مستقل و وابسته به کمک SPSS آنالیز شدند و

[15] Logistic Regression
[16] architectural distortion

متغیرهای مرتبط انتخاب شدند و پس از اطمینان از مناسب بودن رگرسیون لجستیک برای آنالیز، مدل قابل اطمینان آن ساخته شد. جدول یک طبقه بندی نتایج به دست آمده از آنالیز تاریخچه بیماران و ماموگرام به کمک روش رگرسیون لجستیک را نشان می‌دهد.

جدول ۱: طبقه بندی نتایج به دست آمده از آنالیز تاریخچه بیماران و ماموگرام به کمک روش رگرسیون لجستیک

	Groups	Percentage Correct
History Analysis	130 samples	91.7%
History Mammogram	46 Samples	67.4%

نتایج نشان داد که بیماری که در غربالگری ماموگرام وی یک توده کشف شده بود، پنج برابر بیشتر احتمال داشت که دچار سرطان پستان شود. بیمارانی با ضایعات موجود در پستان و کلسیفیکاسیون هیجده بار بیشتر احتمال ابتلا داشتند. بنابراین بیمارای که یک یا ترکیبی از این شرایط را داشت، بیشتر احتمال ابتلا داشت(۲۰ و ۲۱). این مطالعه به رادیولوژیست‌ها کمک می‌کند تا به درستی و با استفاده از تاریخچه بیمار و ماموگرام، ابتلا وی به سرطان پستان را تشخیص دهند.

روش‌های ترکیبی درختان تصمیم و شبکه‌های بیزی[17]:

گاهی جهت افزایش میزان دقت و حساسیت مدل‌های نهائی به دست آمده، ترکیبی از دو مدل در روند تحقیق به کار می‌رود. این امر به ویژه در تحقیقات پزشکی حایز اهمیت است، که صحت نتایج حاصله از حساسیت بسیار بالایی برخوردار است. در سال ۲۰۰۷ در دانشگاه توکیو تحقیقی به کمک بهره گیری از مدل ترکیبی درخت تصمیم و شبکه‌های بیزی ساده در بیماران مبتلا به سرطان پستان صورت گرفت و میزان طول عمر و مرگ بیماران مشخص شد. مدل‌های ترکیبی گاهی نتایج دقیق‌تری به دست می‌دهند(۲۲ و ۲۳).

در مطالعه‌ای که به منظور مقایسه الگوریتم‌های مختلف داده‌کاوی در پیش بینی بقا بیماران مبتلا به سرطان پستان انجام شده بود، یکی از روش‌های منتخب، ترکیبی از دو روش درختان تصمیم و شبکه‌های بیزی بود. هدف از انتخاب یک مدل ترکیبی در کنار سایر مدل‌ها، توسعه مدل مناسب برای پیش بینی دقیق‌تر میزان بقا بیماران بود. در نتایج به دست آمده، میزان دقت این مدل کمتر از مدل درخت تصمیم به تنهایی بود.

جدول ۲: نتایج حاصل از بکارگیری الگوریتم‌های مختلف در پیش بینی بقا بیماران مبتلا به سرطان پستان

Algorithms	Accuracy (%)
Logistics Regression	85.8
Decision Tree	85.6
DT with Naïve Bayes	84.2
Artificial Neural Network	84.5
Naïve Bayes	83.9
Bayes Net	83.9
Decision Trees (ID3)	82.3

در عین حال، موضوع مهم این بود، که به کارگیری شبکه‌های بیزی میزان دقت را افزایش می‌داد و حاصل آن، مدل مناسب و مطمئن‌تری بود و در واقع مدلی مابین درخت تصمیم و شبکه‌های بیزی ارایه می‌کرد. جدول دو نتایج حاصل از به‌کارگیری الگوریتم‌های مختلف در پیش بینی بقا بیماران مبتلا به سرطان پستان را نشان می‌دهد(۱۱).

بحث و نتیجه‌گیری

هدف اصلی به کارگیری الگوریتم‌های داده‌کاوی در علوم پزشکی، بهره گیری بهتر از پایگاه داده و کشف دانش پنهان، به منظور کمک به پزشکان جهت تصمیم‌گیری بهتر است. در سال‌های اخیر تحقیقات متعددی در خصوص داده‌کاوی در سرطان پستان انجام شده و بسیاری از الگوریتم‌ها نیز در این زمینه موفقیت آمیز بوده‌اند. اما نکته قابل توجه این است که میزان موفقیت این الگوریتم‌ها، به فاکتورهای متعددی بستگی دارد و نمی‌توان یک روش را به عنوان بهترین روش برگزید. عواملی چون تعداد متغیرها، بزرگتر بودن پایگاه داده، کم بودن تعداد داده‌های مفقوده و دسترسی به داده‌های مناسب و درست، شانس موفقیت در داده‌کاوی را افزایش می‌دهند و نتایج الگوریتم‌ها را به موفقیت نزدیک‌تر می‌کنند. الگوریتم‌های

[17] Decision Tree with Naive Bayes

DOR: 20.1001.1.17359406.1392.6.1.6.7]

بکار رفته در سرطان پستان تاکنون، در مورد دقت، حساسیت و ویژگی، نتایج متفاوتی ارایه نموده‌اند. مثلاً تحقیقی در خصوص پیش بینی بقا بیماران مبتلا به سرطان پستان به کمک سه الگوریتم درخت تصمیم، شبکه‌های عصبی و رگرسیون لجستیک انجام شد که در نتایج حاصل، بیشترین میزان دقت مربوط به درخت تصمیم بود و پس از آن شبکه‌های عصبی، رگرسیون لجستیک، کمترین میزان دقت را داشت. درختان تصمیم و ماشین بردار پشتیبان، در تحقیقات مختلف انجام شده، معمولا نتایج بهتر و دقیق‌تری ارایه کرده‌اند. ولی همین تحقیق وقتی با پایگاه‌های داده مختلف در کشورهای مختلف انجام شده، نتایج متفاوتی را نشان داده است. دلیل اصلی، حساسیت بالای داده‌کاوی نسبت به نوع پایگاه داده است. پس جهت کسب نتایج بهتر، در ابتدا باید از صحت و درستی اطلاعات موجود در پایگاه داده، وجود متغیرهای مورد نیاز و مناسب و تعداد کافی رکوردها اطمینان حاصل کرد. پس از آن می‌توان از الگوریتم‌های مناسب بهره جسته و به روش مناسب‌تری در داده کاوی دست یافت.

References

۱.غضنفری مهدی، علیزاده سمیه، تیمورپور بابک. داده کاوی و کشف دانش. انتشارات دانشگاه علم و صنعت ایران، ۱۳۸۷.

2. American Cancer Society. Breast Cancer Facts & Figures 2005-2006. Atlanta: American Cancer Society, Inc. (http://www.cancer.org/).

3. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. in Proceedings of the 14th International Conference on AI (IJCAI), San Mateo, CA: Morgan Kaufmann 1995; 1137–45.

4. Aftarczuk K, Kozierkiewicz A. Evaluation of selected data mining algorithms implemented in Medical Decision Support Systems. Report of Institute of Information Science & Engineering, University of Technology. Wroclaw 2009; (1).

5. Witten IH, Frank E. Data Mining, Practical Machine Learning Tools and Techniques. 2nd Elsevier 2008.

6. http://msdn.microsoft.com/en-us/ library/ ms175595.aspx

7. Xindong Wu, Vipin Kumar, J. Ross Quinlan, Top 10 Algorithms in Data Mining, Knowledge and Information Systems, 14(2008), 1: 1-37.

8. Plamena Andreeva, Maya Dimibova and Petra Radeve, “Data mining Learning models and Algorithms for medical applications”, page no.44, 2004.

9. Lundin M, Lundin J, Burke HB, Toikkanen S, Pylkkanen L, Joensuu H. Artificial neural networks applied to survival prediction in breast cancer. Oncology 1999; 57:281-6.

10. Jerez-Aragone´s JM, Gomez-Ruiz JA, Ramos-Jimenez G, Munoz-Perez J, Alba-Conejo E. A combined neural network and decision trees model for prognosis of breast cancer relapse. Artif Intell Med 2003; 27:45-63.

11. Endo A, Shibata T, Tanaka H. Comparison of Seven Algorithms to Predict Breast Cancer Survival. Biomedical Soft Computing and Human Sciences 2008;13, (2):11-6.

12. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Belmont CA: Wadsworth 1984.

13. http://thehealthscience.com/wiki/Decision-Tree.

14. Friedman JH, Kohavi R, Yun Y. Lazy decision trees. In: Proceedings of the thirteenth national conference on artificial intelligence, San Francisco, CA. AAAI Press/MIT Press 1996; 717–24.

15. Thongkam J, Xu G, Zhang Y, Huang F. Support vector machines for outlier detection in cancers survivability prediction. In International workshop on

[DOR: 20.1001.1.17359406.1392.6.1.6.7]

health data management, APWeb'08 2008; 99-109.

16. Chen D, Chang RF, Huang YL. Breast cancer diagnosis using selforganizing map for sonography. Ultrasound Med Biol 2000; 26(3): 405- 11.

17. Ridgeway G, Madigan D, Richardson T. Interpretable boosted naive Bayes classification. In:Agrawal R, Stolorz P, Piatetsky-Shapiro G (eds) Proceedings of the fourth international conference on knowledge discovery and data mining. AAAI Press, Menlo Park 1998; 101-4.

18. Alaa M. Elsayad- Predicting the severity of breast masses with ensemble of Bayesian classifiers‖ journal of computer science 2010; 6(5): 576-84.

19. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Monterey, CA: Wadsworth & Brooks/ Cole Advanced Books & Software; 1984.

20. http://www.arpapress.com/Volumes/Vol10Issue1/IJRRAS_10_1_02.pdf

21. http://www.rsu.ac.th/rjas/article/abstract_120201_20120810_1740.pdf

22. Kohavi R. Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision Tree Hybrid. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining 1996; 202-7.

23. Delen D, Walker G, Kadam A. Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods. Artificial Intelligent Medical 2005; 34(2):p 113-27.

DOR: 20.1001.1.17359406.1392.6.1.6.7]