همایش ملی الکترونیکی دستاوردهای نوین در علوم مهندسی و پایه
National e-Conference on Advances in Basic Sciences and Engineering
. . . . W W W . A E B S C O N F . I R . . . .

تکنیک های تشخیص سرطان خون با استفاده از داده کاوی

خدیجه بابادی، محمد خدامرادی

مهنس نرم افزار کامپیوتر، دانشگاه جهاد اهواز، خوزستان، ایران
عضو هیات علمی دانشگاه ایذه، خوزستان، ایران
مسئول مکاتبات: خدیجه بابادی
Babadi.khadije@yahoo.com

چکیده:
استفاده از داده کاوی در پزشکی یکی از تکنیکهای پرکاربرد داده کاوی محصوب می شود که نقش حیاتی در سلامت دارد و منجر به کشف دانش جدید، سودمند و ماندگار در پایگاه داده ها می شود. امروزه بخش سلامت و پزشکی بیشترین نیاز را به داده کاوی پیدا کرده و حرکت از پزشکی سنتی به سمت پزشکی مبتنی بر شواهد از جمله مواردی است که می توان مؤکد این امر باشد زیرا هنگامی که تعداد پارامترها در تشخیص بیماری زیاد می شود ممکن است تشخیص بیماری حتی برای یک متخصص خبره پزشکی نیز به سختی امکان پذیر باشد. همین دلیل موجب شده که در چند دهه اخیر از ابزار های تشخیص کامپیوتری با هدف کمک به پزشکی با استفاده از ابزارها، احتمال بروز خطاهای احتمالی ناشی از خستگی و یا بی تجربگی فرد را کاهش دهد. سرطان خون یکی از رایج ترین سرطان ها در بین مردم است ک در این مقاله سعی شده با استفاده از تکنیک های داده کاوی مانند خوشه بندی، درخت تصمیم گیری و الگوریتم ژنتیک به تشخیص سرطان خون در کمترین زمان و با جزئیات بیشتر در اختیار پزشک قرار میدهد.
کلمات کلیدی: سرطان خون، پزشک، الگوریتم ژنتیک، بیمار، داده کاوی، درخت تصمیم

مقدمه:
داده کاوی با ایجاد پزشکی مبتنی بر شواهد نقش حیاتی در سلامت دارد و منجر به کشف دانش جدید در پایگاه های داده ای سازمان های سلامت می شود۱. چرا که برای دستیابی به پزشکی مبتنی بر شواهد باید از شناسایی شکاف و خلاء دانش در فرایندهای مراقبت سلامت کنونی شروع کرد و سپس به دنبال بهترین ادله بود، در قدم بعدی باید به بررسی صحیح بودن و معتبر بودن اقدامات شناسایی شده در بهترین ادله پرداخت و در قدم آخر این ادله را به بیماران گزارش کرد. داده کاوی دستیابی به اولین گام در این زمینه را هموار می سازد.که با استفاده از الگوریتم ژنتیک داده ها را وارد میکنیم و با تکنیک خوشه بندی داده ها دسته بندی می شود و در نهایت با استفاده از درخت تصمیم گیری یک تصمیم درست در رابطه با داده های ورودی گرفته می شود که این مراحل به پزشک مربوطه کمک میکند در کمترین زمان بادقت بالایی یک تشخیص درست در زمینه ی سرطان خون بگیرد. سرطان‌خون یا **لوسمی** یا **لوکمیا** نام دارد که انواع مختلفی مانند: لوسمی مزمن میلوئیدی ، لوسمی حاد میلوئیدی ، لوسمی مزمن لنفوئیدی و لوسمی حاد لنفوئیدی (ALL، CLL، AML، CML)دارد۲. با توجه به نوع سلول سرطانی، درمان و نوع علایم آنها با هم متفاوت هستند. باید از مغزاستخوان بیمار نمونه‌برداری و ظرف ۴۸ ساعت تصمیم‌گیری و شیمی‌درمانی شروع شود. پس با وجود این حساسیت بالا و زمان کوتاه در تصمیم گیری باید یک راه حل دقیق برای درست تصمیم گرفتن وجود داشته باشد تا از بروز خطا های احتمالی جلوگیری کرد که این تصمیم را با استفاده از الگوریتم ژنتیک و درخت تصمیم گیری برای پزشک آسان میکنیم.
با استفاده از داده کاوی و مد لسازی داده ها می توان بیماران با شرایط پرخطر را شناسایی کرد .در واقع داده کاوی با ارایه ی اطلاعات به ارایه دهندگان مراقبت، آن ها را در شناسایی بیماران پرخطر به گونه ای که بتوان کیفیت مراقبت آن ها را بهبود داد و از مشکلات آتی آ نها جلوگیری کرد، کمک می کند ۳ و با طراحی مداخله ی مناسب منجر به کاهش پذیرش های بیمارستانی می شود ۴ .به عنوان مثال تکنیک های مد لسازی پیش بینی کننده ی داده کاوی در رابطه با مدیریت بیماری دیابت منجر به ارتقای کیفیت و کاهش هزینه ی بیماران مبتلا به دیابت می شود۳ پس با این تکتیک می توانیم یک راه حل برای تشخیص بیماری سرطان خون ارائه دهیم تا بیمار برای به دست آوردن جواب آزمایش های خود به پزشکان متعدد مراجعه نکند. در این مقاله به کاهش هزینه و سرفه جویی در وقت توجه بسیار شده.

اصول الگریتم های ژنتیک

الگوریتم ژنتیک (GA) هر نقطه در یک پارامتر با فضای راه حل را به یک رشته بیتی به نام کروموزم رمز گذاری می کنند. این نقاط در یک فضای n بعدی، نشان نمی دهند. در حالی که نمونه ها در دیگر روش ها و متولوژی های کاوش داده های مجموعه داده هایی هستند که بیشتر به منظور آموزش و آزمون مورد استفاده قرار می گیرند ولی در عین حال مجموعه ای از نقاط n بعدی در الگوریتم ژنتیکی بخشی از یک الگوریتم ژنتیکی (GA) بوده و آنها به صورت مداوم و مکرر در فرایند بهینه سازی، تولید می شوند. هر نقطه یا رشته دودویی یک راه حل بالقوه برای مساله ای که باید حل شود را نشان می دهد و در الگوریتم های ژنتیک (GA)، متغیرهای تصمیم گیری یک مسئله بهینه سازی به وسیله ساختاری از یک یا چند رشته ای کد گذاری می شود، به نحوی که این مفهوم مشابه کروموزم ها در سیستم ژنتیک طبیعی می باشد. رشته ها کد گذاری در حقیقت از جنبه ها و خصیصه هایی که دقیقا مشابه ژن ها هستند، تشکیل شده است. این خصیصه ها در موقعیت های مختلف این رشته قرار دارند، جایی که هر خصیصه موقعیت (مکان) خود و یک مقدار آلل را که با روش کدگذاری مورد نظر سازگار است، دارا می باشد. ساختار های رشته در کروموزوم ها مشابه فرایند تحولات طبیعی از طریق عملکردهای مختلف برای تولید راه حل های جایگزین بهتر ارائه می گردد. کیفیت کرومزوم های جدید بر اساس مقدار "برازش" یا "انطباق" تخمین و برآورد می شود که این امر می تواند به عنوان تابع هدف برای مساله بهینه سازی در نظر گرفته شود.۵

جدول۱- مفاهیم اساسی در الگوریتم های نتیک

مفهوم در تحولات طبیعی	مفهوم در الگوریتم های ژنتیک
کروموزوم	رشته
والد اول	موقعیت در رشته
آلل	مقدار مکانی(معمولا صفر یا ۱)
ژنوتیپ	ساختار رشته
فنوتیپ	مجموعه خصوصیات(ویژگی ها)
ژن	خصیصه ها و ویژگی ها در رشته

رمز گذاری الگو و مقداردهی اولیه

یک الگوریتم ژنتیک با طراحی نمایشی برای یک راه حل مساله مشخص، شروع می شود.در اینجا منظور از راه حل، در نظر گرفتن هر مقداری است که برای یک راه حل صحیح می تواند، ارزیابی شود. برای مثال، فرض کنید خواسته باشیم تابع $y = 5 - (x-1)^2$ را بیشینه کنیم. سپس x=2 یک راه حل بوده و x=2.5 نیز راه حل دیگری است، همچنین، x=3 راه حل صحیح مسئله است که y را بیشینه می کند. نمایش هر راه حل برای الگوریتم ژنتیک بستگی به طراح آن داشته و در حقیقت آن بستگی دارد به آنچه که هر راه حل به نظرمی رسد و اینکه چه مشکلی از راه حل برای اعمال یک الگوریتم ژنتیک مفید و مناسب است. متداول ترین نمایش یک را حل، رشته ای از کاراکترها می باشد، یعنی رشته ای که از کدها برای نمایش خصوصیات به نحوی که کاراکترها متعلق به یک الفبای ثابت می باشد. هر چه مجموعه الفبای یاد شده بزرگتر باشد، اطلاعات بیشتری می توانید توسط هر کاراکتر در رشته ارائه شود. بنابراین، عناصر کمتری در یک رشته برای رمزگذاری میزان مشخصی از اطلاعات ضروری می باشد. هرچند، در اکثر کاربردهای دنیای واقعی، الگوریتم های ژنتیک (GA) از یک الگو یا طرح دودویی – کدگذاری استفاده می کنند.

فرایند رمزگذاری، نقاط در یک فضای خصوصیات را به نمایش رشته بیتی تبدیل می کند. برای مثال، یک نقطه (11, 6, 9) در یک فضای خصوصیت سه بعدی، با بازه [0, 15] برای هر بعد، می تواند به عنوان یک رشته دودویی الحاق شده در نظر گرفته شود:

(11, 6, 9) = (101101101001)

به نحوی که هر مقدار اعشاری مربوط به خصوصیت به عنوان یک ژن ترکیب شده از ۴ بیت رمزگذاری می شود(با استفاده از یک کدگذاری دودویی).در اینجا باید توجه داشت که مجموعه ای از تمام مقادیر خصوصیات های رمزگذاری شده در یک رشته بیتی نشان دهنده یک کروموزوم است. در الگوریتم های ژنتیک، ما نه تنها بر روی یک تک کروموزوم، بلکه بر روی مجموعه ای از کروموزوم ها به نام جمعیت عمل می کنیم. برای مقدار دهی اولیه جمعیت، می توانیم به صورت ساده تعدادی از کروموزوم ها را به صورت تصادفی در نظر بگیریم. اندازه این جمعیت همچنین یکی از با اهمیت ترین انتخاب هایی است که هر کاربر الگوریتم ژنتیک با آن مواجه می باشد و ممکن است در خیلی از کاربردها این مسئله، بحرانی باشد. در اینجا این سئوال مطرح می شود که آیا به طور کلی به راه حل تقریبی خواهیم رسید و اگر پاسخ این سئوال مثبت باشد، این امر با چه سرعتی انجام می گیرد؟ اگر اداره جمعیت خیلی کوچک باشد، الگوریتم ژنتیک ممکن است خیلی سریع همگرا شده و ممکن است به یک راه حل منجر شود که تنها به صورت داخلی بهینه باشد. اگر اندازه جمعیت خیلی بزرگ باشد، الگوریتم ژنتیک ممکن است منابع محاسباتی را از دست داده و زمان انتظار برای یک بهبود ممکن است، خیلی طولانی گردد.6

مروری بر الگریتم خوشه بندی مبنی بر الگوریتم ژنتیک

کریشنا و مورتی : یک الگوریتم ژنتیک ترکیبی را ابداع کردند که جواب بهینه کلی را برای مساله خوشه بندی با تعداد خوشه های مشخص بدست می دهد. در این شکل ابداعی از الگوریتم ژنتیک از الگوریتم کلاسیک شیب نزولی در فرایند خوشه بندی استفاده شده است. در الگوریتم K-means مبنی بر الگوریتم ژنتیک (GKA)، عملگر K-means به عنون

عملگر جستجو تعریف می شود و به جای عملگر نقاطع مورد استفاده قرار میگیرد. در روش GKA عملگر جهش خاصی متناسب با مساله خوشه بندی تعریف شده است که جهش مبتنی بر فاصله نامیده می شود. با استفاده از تئوری زنجیره مارکوف اثبات می شود که روش GKA به جواب بهینه کلی همگرا می شود. یکی از مهم ترین مشکلات در روش های خوشه بندی از نوع افراز داده ها، یافتن افرازی از داده هاست که با داشتن تعداد مشخصی خوشه، مجموع تغیرات درون خوشه ای (TWCV) را کمینه کند. کمینه سازی مقدار TWCV در روش GKA انجام می گیرد.

الگوریتم سریع K-means مبتنی بر ژنتیک (FGKA) از الگوریتم GKA الهام گرفته شده است ، اما در بسیاری از جنبه ها نسبت به GKA بهبود داده شده است. آزمایشات نشان می دهد که اگر چه امکان همگرایی روش K-means به یک جواب بهینه محلی وجود دارد، روش های GKA و FGKA همواره به جواب کلی همگرا میشوند، اما روش FGKA بسیار سریع تر از GKA عمل میکند.7

لین و دیگران روشی را ارائه کردند که در آن مرکز خوشه ها مستقیما از میان مجموع داده امتخاب می شود. در این روش ابتدا یک جدول جستجو ایجاد می شود و فاصله بین زوج از نقاط محاسبه شدهو در این جدول قرار داده میشود. در نتیجه برای محاسبه تابع برازندگی نیاز به انجام محاسبات تکراری نیست و مقادیر مورد نیاز از جدول استخراج می شود. اتخاذ این رویه سبب سرعت بیشتر در محاسبه تابع براندازگی و در نتیجه سریع تر شدن کل الگوریتم می گردد. در این روش، برای نمایش جوابها در کروموزوم ها از نمایش دوتایی به جای نمایش حقیقی استفاده شده است و همچنین شاخص دیویس – بولدین برای سنجش اعتبار خوشه ها مورد استفاده قرار گرفته است.8

کاتاری و دیگران ، روشی را برای خوشه بندی داده ها ارائه نمودند که در آن از الگوریتم ژنتیک بهبود یافته استفاده شده و عملگرهای تقاطع و جهش به شکل کاراتری تعریف شده اند. به علاوه روش جستجو سیمپلکس نلدر – مید (NM) و ورش K-means نیز در الگوریت ارائه شده مورد استفاده قرار گرفت است تا الگوریتم ترکیبی از مزایا و پتانسیلهای هر دئ روش برخوردار باشند.9

تبیین بخش های مختلف الگوریتم ژنتیک

نحو نمایش رشته ها

کروموزوم ها از اعداد حقیقی تشکیل شده اند و مقادیر و مختصات مربوط به مراکز خوشه ها را در خود جای داده اند. طول رشته ها ثابت و برای مقدار K_{max} است. مقدار k یعنی تعداد خوشه ها به صورت تصادفی از بازه $[k_{min}, k_{max}]$ انتخاب می شود که مقادیر k_{max}و k_{min} جزء ورودی های مساله بوده که می بایست توسط کاربر معین شوند. پس از مشخص شدن مقدار k ، تعداد k ژن مرکز خوشه ها را در خود جای می دهند و مابقی ژن ها یک عدد خاص (در روش پیشنهادی ما عدد ۱۰۰۰) تخصیص داده می شود تا مشخص شود که ژن مربوطه خالی است و مرکز خوشه ای در آن قرار نگرفته است.10

مقدار دهی اولیه جمعیت

به ازای هر رشته (یا کروموزوم) i در جمعیت (i=1, ……,p و P برابر با اندازه جمعیت است)، یک مقدار تصادفی k_iدر بازه تعریف شده تولید می شود.سپس k_iنقطه به صورت تصادفی از میان داده ها انتخاب می شود و به صورت تصادفی در میان خانه ها رشته قرار داده می شود. در نهایت به ژن های خالی رشته مقدار ۱۰۰۰ تخصیص داده می شود.10

پیش نیاز اصلی در درخت تصمیم گیری

برای اعمال بعضی از روش ها که بر پایه ی یادگیری استقرار هستند، چندین پیش نیاز اصلی باید برآورده شود:

۱. توصیف مقدار ویژگی یا صفت: داده ها برای تحلیل، باید به شکل یک فایل مسطح ارائه گردند. تمام اطلاعات در مورد یک موضوع یا مثال باید بر حسب مجموعه ی ثابتی از ویژگی ها یا صفات بیان شوند. هر ویژگی و صفت ممکن است، مقادیر عددی داشته باشد، اما ویژگی های به کار رفته برای توصیف نمونه ها نباید از یک مورد به مورد دیگر متفاوت باشد. این محدودیت منجر به ارائه می گردد که در آن نمونه ها دارای یک ساختار متغیر ذاتی هستند.

۲. گروه ها و کلاس های از پیش تعریف شده: کلاس ها و گروه هایی که نمونه ها به آنها منسوب شده اند، باید زودتر ایجاد شوند. در اصطلاح یادگیری ماشین این یادگیری نظارتی است.

۳. کلاس ها و گروه های گسسته: کلاس ها باید به سرعت ترسیم شوند. یک مورد یا به یک گروه و کلاس خاص تعلق دارند یا ندارد. انتظار می رود که نمونه ها بسیار زیادتری نسبت به کلاس ها وجود داشته باشد.

۴. داده های کافی: تعمیم استقراری ارائه شده به شکل درخت تصمیم با شناسایی الگو در داده ها محقق می شود. این روش اگر تعداد کافی از الگوهای قوی از انطباق تصادفی قابل تمایط باشند، معتبر خواهد بود. همان گونه که این تفاوت معمولا وابسته به آزمون های آماری است، تعداد کافی از این نمونه ها برای اینکه این آزمون ها موثر باشند، باید وجود داشته باشند.

۵. مدل های طبقه بندی منطقی: این روش ها تنها آن دسته از رده بندها را ایجاد می کنند که بتوانند به عنوان درختان تصمیم یا قوانین تصمیم بیان شوند. این شکل ها اساسا توصیف یک کلاس به یک عبارت منطقی که پایه های آن عبارتی درباره ی مقادیر و ویژگی های خاص هستند، محدود می کند. بعضی کاربردها، صفات و ویژگی های ارزشی (وزنی) یا ترککیبات حسابی آنها برای یک توصیف قابل اطمینان از کلاس ها نیاز است.11

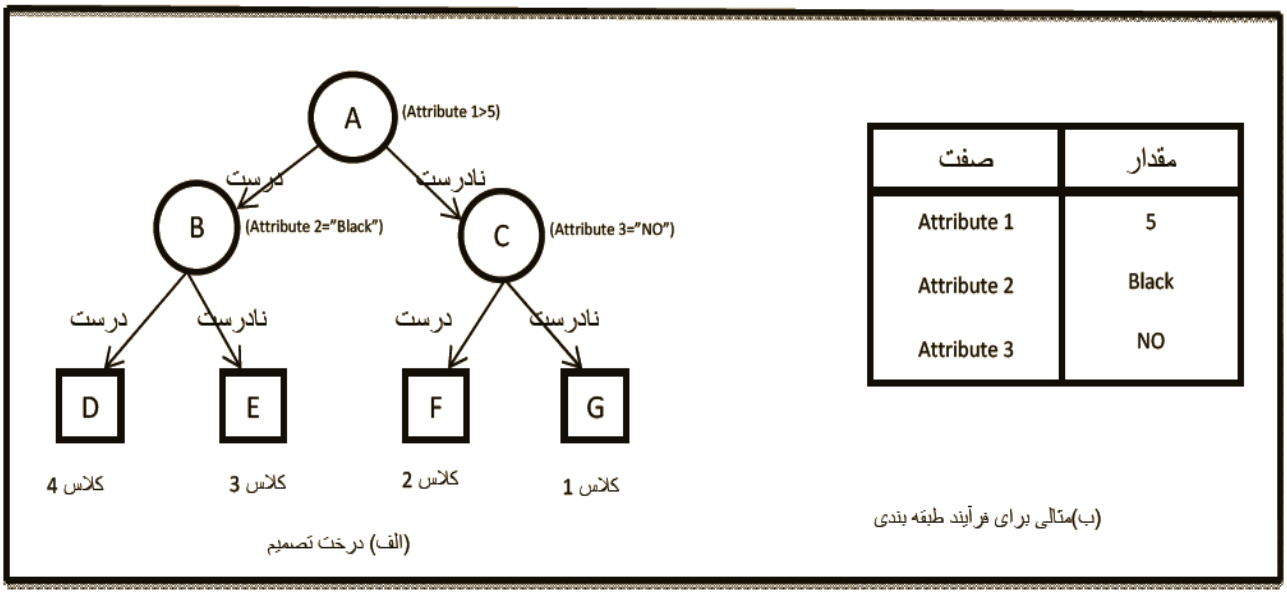
الگوریتم C4.5: تولید درخت تصمیم

مهمترین بخش الگوریتم C4.5، فرایند ایجاد درخت اولیه ی تصمیم از مجموعه نمونه های آزمایشی است. به عنوان یک نتیجه، این الگوریتم یک رده بند به شکل یک درخت تصمیم ایجاد می کند. یک ساختار با دو نوع گره: یک گره برگ که یک کلاس را نشان می دهد یا یک گره تصمیم که برخی آزمون ها را برای اجرا بر روی مقدار یک صفت و ویژگی با یک شاخه و زیر درخت برای هر نتیجه ی ممکن از آزمون تعیین می کند. یک درخت تصمیم می تواند برای طبقه بندی یک نمونه ی جدید با شروع از ریشه درخت و حرکت در آن تا رسیدن به یک گره برگ مورد استفاده قرار گیرد. در هر گره تصمیم غیر برگ، نتیجه ی مشخصه ها برای آزمایش در گره تعیین می شود و توجه به ریشه ی زیر درخت انتخاب شده منتقل می شود. برای مثال، اگر مدل طبقه بندی مسئله به همراه درخت تصمیم در شکل الف و نمونه برای طبقه بندی در شکل ب ارائه شوند، سچس الگوریتم، مسیری را در میان گروه هایA ، C وF (برگ) ایجاد می کند تا تصمیم نهایی طبقه بندی را ارائه کند.

چهارچوب الگوریتم C4.5 بر پایه ی روش CLS هانت برای ایجاد و ساخت یک درخت تصمیم از یک مجموعه ی T از نمونه ها ی آموزشی می باشد. اجازه دهید گروه ها و کلاس ها به صورت $\{C_1, C_2, \ldots, C_K\}$ نمایش داده شوند. در اینجا سه امکان برای محتوای مجموعه ی T وجود دارد:

۱. T شامل یک نمونه یا بیشتر می باشد که همه متعلق به یک تک گروه (کلاس)C_j می باشد، درخت تصمیم برای T، یک برگ است که گروه C_jرا شناسایی کی کند.

جدول ۲- طبقه بندی یک نمونه جدید مبتنی بر مدل درخت تصمیم



۲. در اینجا حاوی هیچ نمونه ای نیست و درخت تصمیم دوباره یک برگ است، اما گروهی که با برگ متناظر می باشند، باید از اطلاعات دیگر T از قبیل گروه اکثریت کلی در T معین شود.

۳. T شامل نمونه هایی است که به ترکیبی از گروه ها تعلق دارد. در این شرایط، ایده اصلی پالایش T در زیرمجموعه از نمونه هایی است که به سمت یک مجموعه ی تک گروهی از نمونه ها پیشروی می کند بر پایه ی تک ویژگی، یک آزمون مناسب که نتیجه یا نتایج اختصاصی متقابلی دارد$\{O_1, O_2, \ldots, O_n\}$ انتخاب شوند. T به زیر درخت های $T_1, T_2, \ldots, T_n$ تقسیم می شود به نحوی که T_1 شامل تمام نمونه ها در T است که دارای نتایج O_i از آزمون انتخابی می باشد.

روند ساخت درخت مشابه برای هر زیرمجموعه از نمونه های آزمایشی به صورت بازگشتی اعمال شده است. بنابراین، شاخه iام منتج به درخت تصمیم ایجاد شده از زیر مجموعه ی T_i از نمونه های آموزشی می شود. تصمیم متوالی مجموعه ی نمونه های آموزشی تا تمام زیرمجموعه ها شامل نمونه های متعلق به یک نک گروه پیش می رود. ۱۱

مقادیر ویژگی ناشناخته

بخش قبلی الگوریتم C4.5 بر پایه ی این فرضیه بود که تمام مقادیر برای همه ی ویژگی ها تعیین می شد. اما در یک مجموعه ی داده ای بعضی مقادیر ویژگی برای برخی نمونه ها ناشناخته هستند. با چنین اشکالات و نقصان هایی در کاربرد های دنیای واقعی وجود دارد. این مسئله ممکن است اتفاق بیافتد. چرا که مقدار مربوط به یک نمونه خاص نیست با مقدار

ثبت نمی شود. وقتی که داده ها جمع آوری می شوند یا یک خطا توسط فردی که داده ها را در یک پایگاه داده ای وارد می کند، این ایجاد می شود. برای حل مشکل مقادیر ناشناخته دو انتخاب وجود دارد:

۱. حذف تمام نمونه ها در یک پایگاه داده که دارای داده های ناشناخته هستند یا
۲. تعین الگوریتم جدید یا تغییر یک الگوریتم موجود که با داده های ناشناخته کار می کند.

ره حل اول ساده است، اما وقتی مقادیر ناشناخته زیادی در یک مجموعه از نمونه های موجود باشد، این راه حل غیر قابل است. برای در نظر گرفتن راه حل دوم، چندین سوال باید پاسخ داده شود:

۱. چگونه دو نمونه با تعداد متفاوتی از مقادیر ناشناخته مقایسه میشوند؟
۲. نمونه های آموزشی با مقادیر ناشناخته نمی توانند با یک مقدار خاص آزمون همراه شوند بنابراین آنها را نمی توان به هیچ زیرمجموعه ای از نمونه ها نسبت داده. چگونه با این نمونه ها در تقصیم بندی باید برخورد نمود؟

تمام این سئوالات و سئوالات دیگر با هر تلاشی برای یافتن یک راه حل برای داده های نامعلوم به وجود آمده اند. چندین الگوریتم طبقه بندی با داده های نامعلوم سرو کار دارند، معمولا بر چاپه تکمیل یک مقدار نامعلوم با محتمل ترین مقدار است یا بر پایه ی بررسی توزیع احتمالی تمام مقادیر برای ویژگی معین میباشد. هیچ یک از این روش ها به طور یکسان دارای برتری خاصی نیستند.

در C4.5، اصلی پذیرفته شده وجود دارد که نمونه های با مقادیر نامعلوم، احتمالا طبق تکرار نسبی مقادیر معلوم توزیع می شود. Info(T) و $Info_x(T)$ مانند قبل محاسبه شوند به استثنای اینکه تنها نمونه های با مقادیر معلوم مشخصه ها به حساب می آیند. پس پارامتر سودمند می تواند به طور معقولی با یک فکتور F تحصیح شود که در این صورت احتمال اینکه یک مشخصه معین، معلوم باشد را نشان می دهد.۱۱

هرس کردن درخت تصمیم

حذف یک زیر درخت یا بیشتر و جایگزینی آنها با برگ ها، یک درخت تصمیم را ساده می کنند و این وظیفه اصلی در هرس کردن درخت تصمیم است. در جایگزین زیر درخت با یک درخت الگوریتم مربوطه انتظار می رود که **میزان خطاهای پیشگویی شده** را پایین ببرد و کیفیت یک مدل طبقه بندی را افزایش دهد. یک میزان خطا تنها بر پایه ی یک مجموعه داده های آموزشی، ارزیابی متناسبی را فراهم نمی کند. یک احتمال برای ارزیابی میزان خطای پیشگویی شده استفاده از مجموعه جدید دیگری از نمونه های آزمایشی است اگر آنها در دسترس باشند.این تکنیک ها اساسا نمونه های موجود در بلوک های هم اندازه را تقسیم می کنند و برای هر بلوک، درخت از تمام نمونه ها ایجاد شود، این بلوک را مستثنی می کند و با یک بلوک معین از نمونه ها می آزماید. با نمونه های آموزشی و آزمایشی، ایده ی اصلی، هرس کردن درخت تصمیم،جابه جایی قسمت های درخت (زیر درخت ها) می باشد، به نحوی که در صحت طبقه بندی نمونه های آموزشی مشاهده نشده، ایجاد درختی با پیچیدگی کمتر و بنابراین درخت قابل درک تر تاثیری ندارد. دو روش وجود دارد که می تواند روش تقسیم کردن بازگشتی را اصلاح نماید:

۱. تصمیم بر عدم تقصیم بیشتر یک مجموعه از نمونه ها تحت برخی شرایط: شاخص توقف معمولا بر پایه ی بعضی آزمون های آماری است از قبیل آزمون x^2. اگر تفاوت های مهمی در دقت طبقه بندی قبل و بعد از تقسیم وجود نداشته باشد، آنگاه یک گره جاری به عنوان یک برگ را ارائه می دهد. تصمیم فوق، قبل از فرآیند تقسیم اتخاذ می شود و بنابراین، این روش، **پیش هرس** کردن نامیده می شود.
۲. جابه جایی با دقت بعضی از ساختارهای درختی با استفاده از معیار انتخاب شده ی دقیق: بعد از ساخت و ایجاد درخت، فرآیند **پس هرس کردن** انجام می گیرد.۱۱

نتیجه گیری

در این مقاله یک تکنیک برای تشخیص سرطان خون ارائه شده است. ما در روش پیشنهادی خود، از الگوریتم ژنتیک، الگوریتم خوشه بندی و درخت تصمیم گیری استفاده کردیم. که یک بیمار وقتی جواب آزمایشات خود را به پزشک نشان می دهد و پزشک مربوطه بدون ترس از خطا های احتمالی ناشی از خستگی و کم تجربگی می تواند جوابب به بیمار خود بدهد.

در این مقاله در ابتدا از الگوریتم ژنتیک استفاده شده که پزشک داده ها را در الگوریتم وارد می کند هر داده با یک نام و یک مفهوم در الگوریتم ژنتیک بررسی می شود مانند یک رشته که به عنوان کروموزوم یا خصیصه ها و ویژگی ها در رشته به عنوان ژن و .... نام گذاری می شود و بعد در الگوریتم خوشه بندی آنها را ترکیب و دسته بندی کنیم در نهایت یک درخت تصمیم می توان برای طبقه بندی یک نمونه ی جدید با شروع از ریشه درخت و حرکت آن تا رسیدن به یک گره برگ مورد استفاده قرار گیرد. در هر گره تصمیم غیر برگ، نتیجه ی مشخصه ها برای آزمایش در گره تعیین می شود و در آخر یک سری عملیات ریز در درخت تصمیم بر روی داده ها انجام می شود که یک تصمیم نهایی در مورد نتیجه آزمایش به دست می آید. در این مقاله به سرفه جویی در وقت و هزینه توجه بسیار شده.

منابع

(۱) Canlas RD. Data Mining in Healthcare: Current Applications and Issues [Online]. 2009 [cited 2009 Aug 9]; Available from: URL:
(۲) http://fa.wikipedia.org/wiki

(۳) Obenshain MK. Application of data mining techniques to healthcare data. Infect Control Hosp Epidemiol 2004; 25(8): 690-5.
(۴) Cios KJ. Medical data mining and knowledge discovery. IEEE Eng Med Biol Mag 2000; 19(4): 15-6.
(۵) داده کاوی (Data mining) / تالیف: مهمد کانتاریک ؛ ترجمه امیر غلیخانزاده – بابل: نشر علوم رایانه ، ۱۳۸۵ / الگوریتم های ژنتیک صفحه ۲۵۶
(6) داده کاوی (Data mining) / تالیف: مهمد کانتاریک ؛ ترجمه امیر غلیخانزاده – بابل: نشر علوم رایانه ، ۱۳۸۵ / الگوریتم های ژنتیک صفحه ۲۶۰
(7) K. Krishna and M. N . Murty, "Genetic K-Means Algorithm", IEEE Transaction On Systems, Man, And cybernetics—Part B:CYBERNETICS, Vol. 29,
No. 3, June 1999
(8) H.J. Lin, F.W. Yang and Y.T. Kao," An Efficient GAbased Clustering Technique", Tamkang Journal of Science and Engineering, Vol. ۸, No, ۲pp ۱۱۳ _122 , 2005.
(9) V. Katari, S. C. Satapathy, JVR Murthy,P. Reddy ,"A Hybridized Improved Genetic Algorithm with Variable Length Chromosome for Image Clustering", International Journal of Computer Science and Network Security, VOL. ۷ No.۱۱, November ۲۰۰۷
(10) مقاله ارائه روشی مبنی بر الگوریتم ژنتیک جهت خوشه بندی خودکار داده های مختلط عددی و دسته ای / نویسنده : مسعود یقینی، مهدی ورد
(11) داده کاوی (Data mining) / تالیف: مهمد کانتاریک ؛ ترجمه امیر غلیخانزاده – بابل: نشر علوم رایانه ، ۱۳۸۵ / درختان تصمیم و قوانین تصمیم صفحه ۱۶۲