معرفی و ارزیابی میزان دقت تکنیک های داده کاوی برای کمک به پیش بینی بیماری های قلبی

محمد متو[۱] و سمیه فرهنگ ادیب[۲]
[۱] دانشجوی کارشناسی ارشد، دانشگاه قم ، Arjmand.2200@yahoo.com
[۲] کارشناسی ارشد، دانشگاه قم، Somayehadib@yahoo.com

چکیده - امروزه با افزایش حجم داده‌ها در سیستم‌های اطلاعاتی، نیاز به ابزاری است تا با کمک آن بتوان داده‌های ذخیره شده را مورد پردازش قرارداده و اطلاعات حاصل از این پردازش را در اختیار کاربران قرار داد. تکنیک‌های داده کاوی یکی از مهمترین و بهترین روشها است که به وسیله آن الگوهای مفید از دل داده‌ها خارج می شود و پس از تجزیه و تحلیل مورد استفاده قرار می گیرد. یکی از کاربردهای اصلی داده کاوی در علم پزشکی مخصوصا در شاخه بیماری های قلبی است که از مهمترین علت های مرگ و میر در جهان می باشد. در این تحقیق از چندین روش کلاسه بندی متداول نظیر روش ماشین بردار پشتیبان، درخت تصمیم، بیزین و چندین روش دیگر استفاده شده است و در آخر یک روش ترکیبی ابداعی نیز معرفی شده است. این روش‌ها بر روی مجموعه داده‌های بیماری های قلبی که از ترکیب داده‌های ۴ مرکز پزشکی مختلف جمع آوری شده است به منظور ارزیابی میزان دقت و مقایسه روش‌ها با همدیگر پیاده سازی شده است. برای محاسبه دقت این روش ها از ابزار Clementine استفاده شده است. نتایج حاصل از این روش‌ها بر روی هر ۴ مجموعه داده نشانگر این است که بهترین الگوریتم بستگی به ساختار داده‌ها، دانش اولیه و مسئله مورد نظر دارد.

dataacademy.ir

کلید واژه- تکنیک های داده کاوی، بیماری های قلبی، داده کاوی، ابزار Clementine، روش ترکیبی

۱- مقدمه

امروزه در دانش پزشکی جمع آوری داده‌های فراوان در مورد بیماری‌های مختلف از اهمیت فراوانی برخوردار است. مراکز پزشکی با مقاصد گوناگونی به جمع آوری این داده‌ها می‌پردازند. تحقیق روی این داده‌ها و به دست آوردن نتایج و الگوهای مفید در رابطه با بیماری ها ،یکی از اهداف استفاده از این داده ها است. حجم زیاد این داده ها و سردرگمی حاصل از آن مشکلی است که مانع رسیدن به نتایج قابل توجه می شود. استفاده ازتکنیک‌های داده کاوی در بیماری های قلب و عروقی با غلبه بر این مشکل و به دست آوردن روابط مفید بین عوامل خطر زا بسیارمفید بوده است. این بیماری ها با توجه به شیوع و سهمی که در مرگ و میر انسان ها داشته اند از اهمیت بالایی برخوردارند . در ابتدای قرن بیستم ۱۰ درصد کل مرگ و میرها به علت بیماری های قلبی عروقی بوده است، که در انتهای همین قرن موارد مرگ و میر ناشی از بیماری های قلبی به ۲۵ درصد افزایش یافت و پیش بینی می شود با توجه به روند کنونی تا سال ۲۰۲۵ میلادی بیشتر از ۳۵تا ۶۰ درصد موارد مرگ و میر در جهان از بیماری های قلبی عروقی ناشی شود.[۱]

استفاده از روش های آماری در بررسی داده‌ها مدتی است که رونق گرفته است، به طوری که اغلب تحلیل های پزشکی به وسیله روش های آماری نظیر SPSS انجام می شود. روش های آماری معمولا برای اثبات فرضیه مورد نظر استفاده می شود، بدین ترتیب که ابتدا نظریه ای طرح شده و سپس توسط روش های آماری، درستی آن مورد ارزیابی قرار می گیرد. همکاری متخصصان در زمینه کامپیوتر و پزشکی راه حل جدیدی را در تحلیل این داده ها و به دست آوردن الگوهای مفید و کاربردی ارائه می دهد که همان داده کاوی است[۲]. در داده کاوی بر خلاف علم آمار به دنبال پیشگویی هستند نه کشف یا اثبات. بدین معنا که با استفاده از روش های داده کاوی به دنبال تأیید آنچه از قبل وجود دارد نیستند بلکه به دنبال مشخص کردن الگوهای از قبل شناخته نشده هستند[۳].همچنین در داده کاوی بیماریهای قلبی به دنبال شناسایی افرادی که دارای بیماری قلبی

آکادمی داده
dataacademy.ir

هستند، نبوده بلکه پیدا کردن عواملی که در بروز این بیماری نقش بیشتری دارند هستیم.

۲- مرور ادبیات

۲-۱- بیماری قلبی و نشانه های اولیه آن

اصطلاح بیماری قلبی به تعدادی از بیماری‌هایی که سیستم گردش خون (شامل قلب ورگ های خونی) وشیوهای که در آن خون پمپاژ شده ودر سراسر بدن گردش می کند و آن را تحت تأثیر قرار می دهد اشاره می‌کند و همچنین می توان از آن به عنوان حمله قلبی نیز یاد کرد. کاردیومیوپاتی وبیماری های قلب و عروق برخی ازبیماری‌های قلبی هستند. بیماری قلبی عروقی منجر به بیماری شدید،از کار افتادگی و مرگ می شود[۴]و[۵]. باریک شدن شریان کرونری منجر به کاهش خون و اکسیژن رسانی به قلب و در نتیجه منجر به بیماری کرونری قلب می شود. آنفارکتوس میوکارد، عموما به عنوان حملات قلبی ودرد قلبی، و یا درد قفسه سینه شناخته می شوند، که توسط بیماری کرونری در بر گرفته می شود. انسداد ناگهانی شریان کرونری، عموما ناشی از لخته شدن خون در یک حمله قلبی است. درد قفسه سینه وقتی که خون دریافت شده توسط عضلات قلب ناکافی باشد بوجود می آید.

فشار خون بالا، بیماری عروق کرونر، بیماری دریچه قلب، سکته مغزی، یا تب روماتیسمی/ بیماری روماتیسمی قلب، اشکال مختلف بیماریهای قلبی وعروقی هستند [۶]. علائم بیماری از مهم ترین ابزارهای تشخیص در تعیین نوع بیماری هر فرد می باشد و پزشک با تکیه به شرح حال دقیق از روش های تشخیصی مکمل استفاده می کند.

۲-۲- تشخیص بیماری قلبی

با توجه به نشانه‌های اولیه ای که برای بیماری‌های قلبی وجود دارد پزشکان براساس روش‌های تشخیص که در ادامه معرفی خواهیم نمود بیماری های قلبی را تشخیص می دهند. از فاکتورهای خطر برای بیماریهای قلبی می توان به مواردی نظیر فشار خون بالا، دیابت، کشیدن سیگار، کلسترول بالا، سابقه خانوادگی، چاقی، استفاده ازکوکائین ومحرک‌های مشابه نام برد. برخی از روش‌های تشخیص بیماری‌های قلبی را در زیر مشاهده می کنید[۷] :

الکتروکاردیوگرام یا نوار قلب: در این روش با استفاده از الکترودهایی که روی قفسه سینه و اندام‌ها قرار می‌گیرد، گزارشی از فعالیت الکتریکی قلب ثبت می‌شود.

اکوکاردیوگرافی: در واقع سونوگرافی مخصوص قلب می‌باشد که با استفاده از امواج ماوراء صوت به بررسی ساختمان و کار قلب می‌پردازد. با اکوکاردیوگرافی به سایر حفره‌های قلب، ناهنجاری های مادرزادی قلب، میزان انقباض و بیماری‌های دریچه ای قلب پی می‌بریم. بعد از نوار قلب،اکوکاردیوگرافی معمول‌ترین اقدام تشخیص در بیماری‌های قلب و عروق بوده و کامل کننده اطلاعاتی است که توسط معاینه به دست می‌آید.

تست ورزش: در این روش درحالی‌که بیمار روی دستگاه دویدن ثابت (تردمیل) حرکت می‌کند از بیمار نوار قلب گرفته می‌شود و درصورت وجود تنگی در عروق قلب، همزمان با افزایش ضربان قلب با ورزش، در نوار قلب بیمار تغییراتی حاصل می‌شود که پزشک را به تشخیص گرفتگی عروق کرونر هدایت می‌کند. درصورت مثبت بودن تست ورزش، بیمار با تشخیص پزشک نیاز به انجام آنژیوگرافی عروق کرونر دارد.از فاکتورهای خطر برای بیماریهای قلبی می توان به مواردی نظیر فشار خون بالا، دیابت، کشیدن سیگار، کلسترول بالا، سابقه خانوادگی، چاقی، استفاده ازکوکائین ومحرک‌های مشابه نام برد.

۲-۳- داده کاوی

داده کاوی استخراج اطلاعات مفهومی، ناشناخته و به صورت بالقوه مفید از پایگاه داده می باشد. داده کاوی علم استخراج اطلاعات مفید از پایگاه‌های داده یا مجموعه داده ای می باشد[۸]. داده کاوی پل ارتباطی میان علم آمار، علم کامپیوتر، هوش مصنوعی، الگوشناسی، فراگیری ماشین و بازنمایی بصری داده می باشد وفرآیندی پیچیده جهت شناسایی الگوها و مدلهای صحیح، جدید و به صورت بالقوه مفید، درحجم وسیعی از داده می باشد، به طریقی که این الگو ها و مدل‌ها برای انسان‌ها قابل درک باشند.داده‌ها اغلب حجیم می باشند و به تنهایی قابل استفاده نیستند، بلکه دانش نهفته در داده‌ها قابل استفاده می باشد[۹]. بنابراین بهره‌گیری از قدرت فرآیند داده کاوی جهت شناسایی الگوها و مدل‌ها و نیز ارتباط عناصر مختلف در پایگاه داده جهت کشف دانش نهفته در داده‌ها و نهایتا تبدیل داده به اطلاعات، روز به روز ضروری‌تر می‌شود. تشخیص بیماری‌های مختلف در علم پزشکی یکی از زمینه‌های پرکاربرد داده کاوی

آکادمی داده
dataacademy.ir

محسوب می‌شود. استخراج قواعد طبقه بندی، نوعی داده کاوی است که در آن دانشی به شکل چندین قانون ساده و فهم پذیر از داده کشف شده و در آینده برای تصمیم گیری و پیشگویی به کاربرده می شود[۱۰]. با به کارگیری الگوریتم های داده کاوی می توان سیستم‌های هوشمندی ابداع کرد که به شکل خودکار و بدون نیاز به نظارت پزشک قادر به فهم و تفسیر ویژگی‌های پزشکی افراد باشند و یا اطلاعات مفیدی را اکتشاف کنند که متخصصان را در قضاوت صحیح یاری رساند[۱۱].

انبار داده به مجموعه ای از داده‌ها گفته می شود که از منابع مختلف اطلاعاتی سازمان جمع آوری، دسته بندی و ذخیره می شود. در واقع یک انبار داده مخزن اصلی کلیه داده های حال و گذشته یک سازمان می باشد که برای همیشه جهت انجام عملیات گزارش گیری و آنالیز در دسترس مدیران می‌باشد. انبار داده بیماری‌های قلبی شامل غربالگری از اطلاعات مربوط به بیماران قلبی است. در ابتدا،انبار داده‌ها پیش پردازش می شود و در مراحل بعد با استفاده از ابزارهای داده کاوی نظیر کلمنتاین عملیات نهایی بر روی داده‌های انبار داده انجام می شود[۱۲].

۲-۴- Clementine

کلمنتاین یکی از پرکاربردترین ابزارهای داده کاوی است. محقق ایرانی دانشگاه هاروارد به همراه گروهی از محققان درMIT موفق به ابداع این ابزار داده کاوی شده اند که می تواند الگوهای معنا داری را در میان مجموعه ای گسترده از داده‌ها ردیابی کند.با استفاده از این ابزار می‌توان الگوهای چندگانه پنهان در انواع مجموعه داده‌ها از قبیل داده‌های بهداشتی یا آمار مسابقات بیسبال کشف کرد.ابزار کلمنتاین یک محیط کاری برای انجام فرآیندهای داده‌کاوی است که به ما کمک می کند مدلهای پیشگو را با استفاده از دانش فنی مورد نیاز ساخته و آنها را به منظور تصمیم‌گیری بهتر در فرآیندهای تجاری به کار بگیریم. این ابزار که براساس مدل استاندارد CRISP-DM طراحی شده است، از تمامی مراحل فرآیند داده کاوی از زمان ورود داده خام تا حاصل شدن نتایج مورد انتظار پشتیبانی می نماید.

محققان این ابزار را بر روی چندین مجموعه بزرگ داده‌ها از جمله، مجموعه داده‌هایی در رابطه با تریلیونها ریزجاندارانی که درون روده زندگی می کنند، آزمایش کردند. این ابزار توانست در حدود ۲۲ میلیون مقایسه را در میان داده‌های دریافتی انجام داده و الگوهای پنهان در آنها را به چند صد الگوی جالب توجه که پیش از این مورد مشاهده قرار نگرفته بودند، محدود سازد.به گفته "مایکل میتزن‌ماخر" استاد علوم رایانه ای دانشگاه هاروارد هدف از استفاده از چنین ابزاری بررسی بیشترین ابعاد ممکن در داده‌ها و بیشترین ارتباطات ممکن در میان آنها است تا در نهایت برترین داده‌ها انتخاب شوند.این ابزار نوعی ابزار اکتشافی است که می تواند ارتباطات میان داده‌ها را کشف کرده و براساس اهمیت رتبه بندی کند.براساس گزارش تی جی دیلی، امکان جستجو برای یافتن یک الگو در شرایطی که محقق هنوز نمی داند دقیقا در جستجوی چه چیزی است، به وجود خواهد آمد و می تواند ایده‌های جدید و ارتباطاتی ایجاد کند که تاکنون کسی به فکر جستجو برای یافتن آنها نیفتاده است. به صورت ویژه این ابزار برای اکتشاف در میان مجموعه داده‌هایی مناسب است که می توانند حاوی بیش از یک الگوی مهم باشند. Clementine انواع گوناگونی از روش‌های مدل سازی مانند روش‌های مبتنی بر یادگیری ماشین، هوش مصنوعی و آمار را ارائه می دهد. از طریق این روش‌ها می توان اطلاعات جدیدی را از داده ورودی استخراج نموده و براساس آن مدل های پیشگویی مورد نظر را ایجاد کرد. هر روش نقاط قوت خاص خود را دارد و برای حل دسته خاصی از مسائل سودمند است.گره های مدل سازی در قالب ماژول‌های پایه Classification، Clustering و Association تقسیم می شود [۱۳] و [۱۴]. ما دراین مقاله بر روی روش کلاسه بندی (Classification) متمرکز شده که اساس کلاسه بندی (Classification) مبتنی بر الگوریتم های یادگیری بانظارت است. در الگوریتم‌های یادگیری با نظارت ، هدف از کاوش داده‌ها مشخص بوده و می دانیم که به دنبال چه نوع دانشی می گردیم ولی در الگوریتم های یادگیری بدون نظارت مانند خوشه بندی (Clustering) ، هدف کاملا تعریف شده نیست.

۳- مجموعه داده ها

مجموعه داده‌های بیماری قلبی از ۴ موقعیت زیر جمع آوری شده اند:

- ✓ کلینیک کلیولند (Cleveland)
- ✓ موسسه مجارستانی قلب و عروق (Hungarian)
- ✓ مرکز پزشکی، لانگ بیچ (long-beach-va)
- ✓ دانشگاه علوم پزشکی زوریخ (Switzerland)

داده‌های تمام این پایگاه داده ها فرمت یکسانی دارد.این پایگاه داده‌ها ۷۶ سطر Attribute داشته در حالیکه فقط

dataacademy.ir

آکادمی داده
dataacademy.ir

۱۴سطر از آنها استفاده می شوند،همه Attribute ها مقدار عددی دارند و مجموعه داده‌های استفاده شده برای پژوهش ترکیبی از ۴ مجموعه داده بالاست که تعداد آنها ۹۲۰ رکورد می باشد.

مجموعه داده‌ها شامل Attribute های زیر می باشد :

- Id :شماره شناسایی بیمار
- Age :سن
- Sex :جنسیت (1 = مرد ،0 = زن)
- Painloc : محل درد قفسه سینه (1 = substernal ، 0 = otherwise
- Painexer :(1= provoked by exertion; otherwise=0)
- Relrest : (1= relieved after rest; otherwise = 0)
- Cp : نوع درد قفسه سینه
- Trestbps : فشار خون ایستا
- Chol :کلسترول
- Famhist : سابقه خانوادگی (1 = yes; 0 = no)
- Restecg : نتایج کاردیو گرافی
- ECG: month of exercise reading
- Thaldur : مدت زمان تست ورزش
- Thalach : حداکثر ضربان قلب
- Thalrest : ضربان قلب در زمان استراحت
- Num : تشخیص بیماری قلبی(وضعیت آنژیو گرافی)
  - Value 0 : احتمال باریک شدن قطر رگ کمتر از ۵۰٪ است.(سالم)
  - Value 1 : احتمال باریک شدن قطر رگ بیشتر از ۵۰٪ است.(بیمار)

از ۱۶صفت ذکر شده صفت۱ مشخصه شناسایی و صفت ۱۶ به عنوان کلاس است ،که در کل از ۱۴ صفت برای پیش بینی استفاده شده است. با توجه به مجموعه داده های بیماری قلب در نمودار۱ می توان کلاس بیماری قلبی را مشاهده کرد که به دو قسمت سالم(Healty) و بیمار (Risky)تقسیم می شود.

Value	Proportion	%	Count
healty		44.67	411
risky		55.33	509

نمودار۱- کلاس بیماری قلب در دو قسمت سالم وبیمار

۴- الگوریتم های مورد استفاده

۴-۱- ماشین بردار پشتیبان

یکی از روش‌های یادگیری، یادگیری بانظارت است که از آن برای رده بندی و رگرسیون استفاده می‌کنند.این روش از جمله روش‌های نسبتا جدیدی است که در سال‌های اخیرکارایی خوبی نسبت به روش‌های قدیمی‌تر برای طبقه‌بندی از جمله شبکه‌های عصبی پرسپترون نشان داده است. مبنای کاری دسته بندی کنندهٔ ماشین بردار پشتیبان یا SVM دسته‌بندی خطی داده‌ها است که در تقسیم خطی داده‌ها سعی می‌کنیم خطی را انتخاب کنیم که حاشیه اطمینان بیشتری داشته باشد[۱۵]. 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 تقسیمِ خطی برای اینکه ماشین بتواند داده‌های با پیچیدگی بالا را دسته‌بندی کند داده‌ها را به وسیله ی تابع PHI به فضای با ابعاد خیلی بالاترمی‌بریم.

۴-۲- درخت تصمیم

dataacademy.ir

درخت تصمیم یک ابزار برای پشتیبانی از تصمیم است که از درخت برای مدل سازی استفاده می‌کند. به طور معمول درخت تصمیم در تحقیق درعملیات استفاده شده و به طور خاص در آنالیز تصمیم و نیز برای تشخیص استراتژی که با بیشترین احتمال به هدف برسد به کاربرده می‌شود. استفاده دیگر درخت تصمیم، توصیف محاسبات احتمال شرطی است. در میان ابزارهای پشتیبانی تصمیم، درخت تصمیم و دیاگرام تصمیم دارای مزایایی هستندازجمله فهم ساده ، کارکردن با داده‌های بزرگ و پیچیده ، استفاده مجدد آسان،در صورتی که درخت تصمیم برای یک مسئله ساخته شد، نمونه‌های مختلف از آن مسئله را می‌توان با آن درخت تصمیم محاسبه کرد و همچنین قابلیت ترکیب با روش‌های دیگر اشاره کرد.

چهار الگوریتم برای ساخت درخت تصمیم موجود هستند. این الگوریتم ها اساسا یک عمل شبیه به هم را پیاده سازی می کنند. آنها تمامی صفات پایگاه داده را بررسی می کنند تا به صفاتی برسند که بهترین رده بندی و پیش بینی را با تقسیم داده به زیر گروه ها انجام می دهد. این فعالیت به صورت بازگشتی تکرار می شود تا باز هم زیر گروه‌ها به زیر گروه‌های دیگری شکسته شوند. صفات هدف یا ورودی می توانند از نوع عددی و یا طبقه ای بر حسب الگوریتم مورد استفاده باشند اگر

آکادمی داده
dataacademy.ir

یک بازه مورد استفاده قرار گیرد نتیجه کار یک درخت رگرسیون خواهد بود، اما اگر ورودی ها بصورت رده ای باشند نتیجه کار یک درخت رده بندی خواهد بود[۱۶].

- **C& R Tree**

این درخت یک درخت تصمیم تولید می کند که سعی در پیش بینی و رده بندی مشاهدات آینده دارد.این روش سعی در کم کردن ناخالصی در هر رده دارد.یک گره وقتی کاملا عاری از ناخالصی است که تمامی عناصر یک زیر گروه آن متعلق به یک رده از فیلد هدف باشند. صفت پیش بینی کننده و فیلد هدف می توانند از دو نوع بازه و رده ای باشند.تمامی تقسیم بندی ها دودویی خواهند بود،به این معنی که فقط دو زیر گروه از هر گره منشعب خواهند شد.

- **QUEST**

این درخت عملکردی همانند درخت C&R دارد،اما سعی دارد زمان لازم برای ساخت درخت تصمیم توسط C&R را کاهش دهد. در ضمن صفت پیش بینی کننده آن مثل C&R می تواند هم طبقه ای و هم بازه باشد، اما فیلد هدف آن باید حتما رده ای باشد. درخت تولید شده توسط آن یک درخت دودویی است .

- **CHAID**

برخلاف درخت C&R و QUEST، CHAID می تواند درختی تولید کند که در برخی موارد به صورت غیر دودویی عمل کند، یعنی یک گروه آن به سه زیر گروه و یا بیشتر شکسته شود. صفات پیش بینی کننده و هدف می تواند هم از نوع بازه و هم از نوع رده ای باشد.

- **C5**

برای ساخت یک درخت تصمیم و یا مجموعه قوانین استفاده می گردد. این گره بر حسب نیاز می تواند برای برخی گره ها بیش از دو زیر گروه ایجاد کند. صفت هدف آن حتما باید از نوع رده ای باشد.

در بین ۴ الگوریتم درخت تصمیم ما در ادامه این مقاله از الگوریتم های C5، C&R Tree و CHAID بهره خواهیم برد.

۴-۳- بیزین

Bayes ارائه کننده مدلی پیش بینی کننده در رابطه با احتمال خروجی/ نتایج خاصی است. الگوریتم Naive Bayes یا بیزین الگوها و یا ارتباط بین داده‌ها را با شمارش تعداد وقوع مشاهده آنها اندازه گیری می‌نماید. این الگوریتم سپس مدلی را ایجاد می‌کند که نشان‌دهنده الگوها و ارتباط آنها با یکدیگر است. پس از ایجاد مدل، می توان از آن به عنوان الگوئی پیش بینی کننده استفاده نمود. این الگوریتم کمک خواهد کرد تا بتوان مدل‌هایی را برای طبقه بندی و پیش بینی اهداف متعددی انجام داد مانند،کدام مشتری‌ها علاقه مند هستند تا محصول خاصی را خریداری نمایند، کدام مشتری‌ها قادر هستند بیش از ۳۰۰۰ دلار خریداری نمایند، شناسائی مشتری‌هائی که امکان دارد محصولات شرکت را ترک کرده و از محصولات رقبا خریداری نمایند، پیش بینی محصولاتی که میل به شکست داشته و میزان احتمال آن.

الگوریتم Naïve Bayes این پیش بینی‌ها را توسط تئوری Bayes (که فرض می‌کند مقادیر داده مستقل از یکدیگر می باشند) انجام می‌دهد. این الگوریتم ایجاد مدل را در سریع‌ترین زمان، ممکن می سازد، که برای مسائلی همچون طبقه بندی با ۲ کلاس یا بیشتر قابل انجام خواهد بود [۱۷] .

dataacademy.ir

۴-۴- لیست تصمیم

با در دسترس داشتن رکوردهای مجموعه یادگیری،این الگوریتم قادر به تولید دو نوع معروف از کلاس بندهای جدول تصمیم است که کار هردو بر اساس یک جدول جستجو می باشد.هر جدول تصمیم دارای ساختاری متشکل از دو قسمت است ، الگو ؛ شامل لیستی از صفات می باشد و بدنه ؛ مجموعه از رکوردهای برچسب دار است.هر رکورد دارای مقادیری برای صفات الگو و یک مقدار برای برچسب می باشد .مجموعه ی رکوردها با مقادیر یکسان برای الگو را ،یک سلول می نامیم.

اولین کلاس بند DTMaj نام دارد که در صورت خالی بودن سلولی که با رکورد جدید برابری می کند ،کلاس اکثریت داده های یادگیری را بر می گرداند. دومین کلاس بند DTLoc نام دارد که در صورت خالی بودن سلولی که با رکورد جدید برابری می کند، به دنبال نظیری با تعداد صفات کم تر برای آن می گردد. این روش پاسخ را از میان همسایگان محلی بر می گرداند و برای مجموعه داده‌های یکنواخت کارا می باشد زیرا تغییرات کوچک در صفات ،باعث تغییر در برچسب نمی شود. فرآیند

آکادمی داده
dataacademy.ir

ساخت جدول تصمیم ،شامل پیدا کردن یک لیست بهینه از صفات برای الگو می باشد. این لیست به گونه ای است که کمترین میزان خطا را روی جمعیت رکوردها دارد. برای انتخاب صفات بهینه ، از معیار خستگی استفاده می شود. می توان از معیارهای دیگری نظیر بهره ی اطلاعات، نرخ بهره و اندیس جینی نیز استفاده نمود[۱۸].

۴-۵- روش ترکیبی

یکی از مشهورترین روش های ترکیبی روش رای اکثریت می باشد که در بین خروجی‌های کلاسه بندها ،کلاسی به عنوان نتیجه نهایی انتخاب می شود که توسط تعداد کلاسه بندی بیشتری به عنوان خروجی پیشنهاد داده باشد. این روش علیرغم سادگی،کارایی تمام کلاسه بندها را بدون توجه به ویژگی‌های هر یک،یکسان در نظر می گیرد [۱۹]. در این مقاله ما الگوریتم های C5 ، SVM وDecision List را به روش رای اکثریت با هم ترکیب کرده ایم تا الگوریتم ترکیبی(Tarkib) بدست آید.

۵- نتایج

dataacademy.ir

۵-۱- پیش پردازش داده ها

پیش پردازش وآماده سازی داده‌ها مهمترین و زمان برترین مرحله در پروژه های داده‌کاوی است. تقریبا ۶۰ تا ۹۰ درصد زمان انجام یک پروژه داده‌کاوی صرف این مرحله می شود و ۷۵ تا ۹۰ درصد موفقیت پروژه‌ها داده کاوی به آن بستگی دارد[۲۰].فرآیندهایی که در پیش پردازش انجام می شود عبارت است از، تجمیع ، نمونه گیری ، کاهش بعد ، انتخاب زیر مجموعه ویژگی ها، ایجاد ویژگی و تبدیلات داده و بر اساس نوع کاربردی که عمل داده کاوی باید بر روی آن انجام شود، تکنیک‌های مختلفی برای هر یک از این اعمال مورد استفاده قرار می گیرد[۲۰].

برای پیش پردازش بر روی مجموعه داده بیماری قلبی باید مقادیر از دست رفته یا Missing Value را توسط یکی از روش هایی نظیر روش های استفاده از بیشترین تکرار، استفاده از میانگین گیری ، استفاده از روش نزدیکترین همسایه، حذف کردن کل سطری که داده از دست رفته دارد و یا تکرار فیلدی که داده از دست رفته دارد به ازای مقادیر ممکن، مدیریت کنیم.

با توجه به مجموعه داده موجود می توان از روش های مختلفی استفاده کرد ما در این پژوهش برای مدیریت مقادیر از دست رفته از روش استفاده از بیشترین تکرار استفاده کرده ایم بدین صورت که به جای مقادیر از دست رفته اگر داده گسسته باشد از بین مقادیر موجود بیشترین تکرار را به داده از دست رفته و اگر داده پیوسته باشد میانگین مقادیر داده های موجود برای آن فیلد را به عنوان مقادیر از دست رفته اختصاص می دهیم .

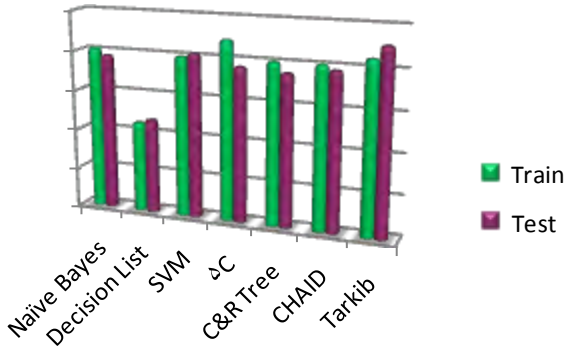
۵-۲- نتایج آزمایش

در این پژوهش ۷ الگوریتم روش Classification برای پیشگویی صحت(Accuracy) بیماری قلبی مورد استفاده قرار می گیرند که عبارتند از SVM، C5 ، CHAID ،C&R Tree، Naive Bayes، Decision List و روش ترکیبی (Tarkib).

بعد از پیش پردازش داده‌ها، داده ها را به دو مجموعه داده Train و Test تقسیم می کنیم. با استفاده از داده های مجموعه Train مدل مورد نظر ما ساخته می شود و سپس داده های Test خود را بر روی مدل ساخته شده توسط مجموعه داده Train آزمایش می کنیم،در این پژوهش با توجه به Attribute ها مجموعه داده را به دو بخش که ۷۰ درصد از داده ها برای Training و ۳۰ درصد برای Testing می باشد تقسیم می شوند. ۷ الگوریتم مطرح شده را بر روی داده ها اجرا می کنیم ، هدف اصلی بدست آوردن Accuracy به وسیله الگوریتم های طبقه بندی در دو مجموعه داده Train و Test می باشد. مجموعه داده ها شامل ۹۲۰ نمونه با ۱۴Attribute مختلف است. این نمونه ها در مجموعه داده نشانگر نتیجه انواع مختلف تست هایی است که برای پیشگویی صحت بیماری قلبی صورت گرفته است می باشد. عملکرد کلاسی‌فایر ها مورد ارزیابی قرار می گیرد و سپس نتایج آن تجزیه و تحلیل می شود و عملکرد را می توان براساس نرخ خطا مشخص کرد. با مقایسه میان این الگوریتم‌ها به این نتیجه می رسیم که با اینکه سعی شده است روش ترکیبی را به صورتی اعمال شود که بالاترین Accuracy را در میان دیگر الگوریتم ها داشته باشد ولی بالاترین دقت در بین الگوریتم ها برای مجموعه Train مربوط به الگوریتم C5 و برای مجموعه Test مربوط به الگوریتم ترکیبی است.با توجه به نمودار ۲ الگوریتم های C5 و Tarkib که ترکیبی از الگوریتم های C5، SVM و Decision List می باشد بهتر از الگوریتم های دیگر کار می کند زیرا نسبت به بقیه صحت بالاتری داشته و الگوریتم Decision List نسبت به

آکادمی داده
dataacademy.ir

بقیه الگوریتم ها از دقت پایین‌تری برخوردار می باشد. نمودار ۲ ، مقایسه مقادیر پارامتر میزان درستی مربوط به الگوریتم های اعمال شده روی مجموعه داده بیماری های قلبی را نمایش می دهد. لازم به تأکید است که، از آنجایی که الگوریتم رگرسیون روی این نوع داده ها جواب قابل قبولی نخواهد داد پس این الگوریتم را درنظر نگرفته ایم.

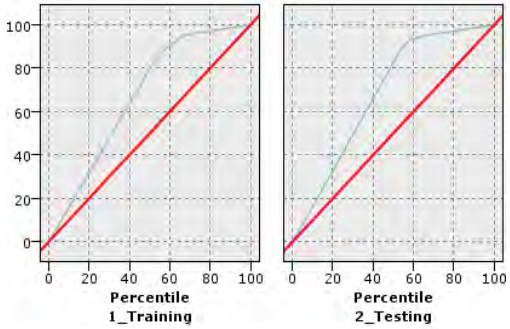


نمودار ۲- مقایسه دقت الگوریتم ها

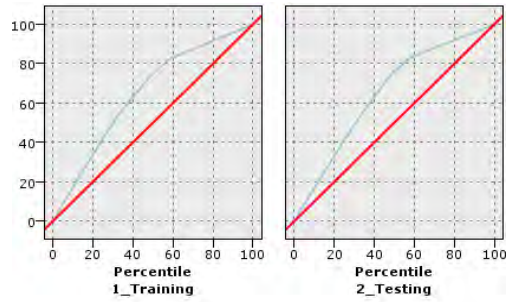
۵-۳- منحنی یادگیری

منحنی یادگیری یا Learning curve ، نشان دهنده میزان پیشرفت در یادگیری به شکل نمودار از روی امتحان‌های موفقیت‌آمیز یادگیرنده است . به‌عبارت دیگر منحنی یادگیری جنبه‌ها و قسمت‌های معینی از پیشرفت در یادگیری است .منحنی یادگیری یکی از راه های ارزیابی الگوریتم های Classification می باشد، هر چه سطح زیر نمودار سنگین تر باشد قدرت یادگیری بیشتر می شود. هر چه شیب منحنی تندتر باشد و شطح زیر منحنی به یک نزدیکتر شود یادگیری در حد بالایی می باشد و هر چه منحنی به خط قرمز رنگ نزدیکتر شود تصمیم گیری شانسی و بر حسب تصادف خواهد بود. با توجه به نتایج مرحله قبل الگوریتم ترکیبی بالاترین دقت در پیش بینی و الگوریتم Decision List پایین ترین دقت را در بین سایر الگوریتم ها دارا می باشد. از این رو در زیر منحنی های مربوط به هر کدام از الگوریتم ها را برای دو مجموعه داده Train و Test نمایش داده می شوند. در نمودار مربوط به الگوریتم ترکیبی نمودار ۳ مشاهده می شود که هرچه به سمت بالای نمودار می رویم قدرت یادگیری به ۱۰۰ نزدیکتر می شود. نمودار ۴ مربوط به منحنی الگوریتم Decision List می باشد.

dataacademy.ir



نمودار ۳- منحنی های مربوط به الگوریتم ترکیبی



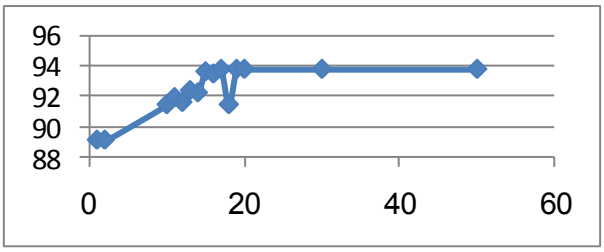
نمودار ۴- منحنی های مربوط به الگوریتم Decision List

۵-۴- افزایش دقت

همانطور که در بخش قبل مشاهده شد الگوریتم C5 از دقت بالاتری نسبت به دیگر الگوریتم ها بر روی مجموعه داده Train برخوردار می باشد و در این بخش به دنبال بالا بردن دقت این الگوریتم با استفاده از تکنیک Boosting هستیم. یکی از تکنیک هایی که می‌توان با استفاده از آن میزان دقت را بهبود داد، تکنیک Boosting در درخت تصمیم C5 می باشد،عملکرد این تکنیک بدین صورت است که چندین مدل را در یک دنباله تولید کرده، اولین مدل را با روش معمول C5 می سازد و سپس مدل دوم با تمرکز بر روی رکوردهای غلط رده بندی شده در مدل اول ساخته می شود و مدل سوم با تمرکز بر روی رکوردهای غلط رده بندی شده در مدل دوم ساخته می شود و به همین منوال تا آخر ادامه می یابد.

با توجه به نمودار ۲ الگوریتم C5 دارای دقت ۸۹.۱۳ ٪ برای داده های Train و دقت ۷۶.۸۱ ٪ برای داده های Test می باشد که با اعمال تکنیک Boosting با مرتبه تکرار ۱۷ بر روی الگوریتم C5 ، می توان دقت الگوریتم C5 را به ترتیب به ۹۳.۷۹ ٪ و ۸۲.۶۱ ٪ افزایش داد و اگر مرتبه تکرار از ۱۷ بالاتر رود از دقت این الگوریتم همانطور که در نمودار ۵ قابل مشاهده است ادامه می یابد.

آکادمی داده
dataacademy.ir



نمودار۵- افزایش دقت الگوریتم C5 با تکنیک Boosting

۶- نتیجه گیری

بررسی‌ها و مقایسات انجام شده روی مجموعه داده بیماری های قلبی و یافتن بهترین و کاراترین و همچنین بدترین الگوریتم‌ها در تشخیص بیماری ما را به این نتیجه می رسانند که هرگز نمی توان الگوریتمی را به عنوان الگوریتم بهینه معرفی کرد، بر طبق مقاله ای که در سال ۲۰۱۰ در زمینه تشخیص بیماری های قلبی انجام شد [۷] بهترین الگوریتم برای محاسبه بالاترین دقت پیش بینی بیماری های قلبی ،الگوریتم Naïve Bayse شناخته شده است در صورتیکه در این پژوهش الگوریتم Naïve Bayes به نسبت دیگر الگوریتم ها دقت پیش بینی پایینی دارد.بنابراین در این بین الگوریتم های ترکیبی جواب قابل قبولتری برای ما ارائه می دهند.

داده کاوی در مدیریت بهداشت و درمان برخلاف زمینه‌های دیگر مدیون این حقیقت است که داده‌ها در حال حاضر ناهمگن هستند و اینکه محدودیتهای خاص اخلاقی، حقوقی و اجتماعی درمورد اطلاعات خصوصی پزشکی وجود دارد. داده‌های مربوط به مراقبت بهداشت و درمان به طور طبیعی حجیم هستند و آنها از منابع گوناگون بدست می آیند و همه آنها به طور کامل در ساختار یا کیفیت مناسب نیستند. امروزه، بهره برداری ازدانش و تجربه متخصصان متعدد و داده‌های غربالگری بالینی بیماران جمع آوری شده در یک پایگاه داده درطی روند تشخیص، به طورگسترده ای استفاده می شود.

مراجع

[۱] ماهنامه تخصصی پزشکی شماره ۱۳۱ ،اسفند ماه ۱۳۹۰

[2] M. Chen, J. Han, and P. S. Yu. Data Mining: An Overview from Database Perspective. IEEE Trans. Knowl. Dat. Eng., vol: 8, no:6, pp: 866-883, 1996.

[3] Carlos Ordonez, "Improving Heart Disease Prediction Using Constrained Association Rules," Seminar Presentation atUniversity of Tokyo, 2004

[4] Heart Disease from http://chineseschool. netfirms.com/heart-disease-causes.html

[5] "Hospitalization for Heart Attack, Stroke, or Congestive Heart Failure among Persons with Diabetes", Special report: 2001 – 2003, New Mexico.

[6] Heart disease from http://en.wikipedia.org/wiki/Heart_disease

[7] Diagnosis Of Heart Disease Using Datamining Algorithm, Asha Rajkumar , Mrs. G.Sophia Reena, P a g e 38 Vol. 10 Issue 10 Ver. 1.0 Sepetember 2010.

[8] D. Hand,H. Mannila,P. Smyth(2001).Principles of Data Mining.MITPress,Cambridge.

[9] Huang T.-M., Kecman V., Kopriva I. (2006), Kernel Based Algorithms for Mining Huge Data Sets, Supervised, Semi-supervised, and Unsupervised Learning, Springer-Verlag, Berlin, Heidelberg, 260 pp. 96 illus., Hardcover, ISBN 3-540-31681-7 .

[10] Franck Le Duff , CristianMunteanb, Marc Cuggiaa, Philippe Mabob, "Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method", Studies in health technology and informatics ,107(Pt 2):1256-9, 2004.

[11] Boleslaw Szymanski, Long Han, Mark Embrechts, Alexander Ross, KarstenSternickel, Lijuan Zhu, "Using Efficient Supanova Kernel For Heart Disease Diagnosis", proc. ANNIE 06, intelligent engineering systems through artificial neural networks, vol. 16, pp:305-310, 2006.

[12] W.Frawley and G. Piatetsky. Knowledge Discovery I DataBases.ISSN 0738-4602.

[13] Kiyong Noh, HeonGyu Lee, Ho-Sun Shon, Bum JuLee,andKeun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", springer,Vol:345 , pp721-727, 2006.

[14] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases", In Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile, August 29-September 1994.

[15] Vojislav Kecman: "Learning and Soft Computing Support Vector Machines, Neural Networks, Fuzzy Logic Systems", The MIT Press, Cambridge, MA, 2001.

[۱۶] مینایی بهروز ،نصیری مهدی ،حسنی دانیال ،شناسا ابراهیم ؛آموزش گام به گام داده کاوی با Clementine ؛انتشارات گروه مهندسی – پژوهشی ساحر ،انتشارات پارسه ،چاپ اول ،پاییز ۱۳۹۰.

[۱۷] جام سحر خشایار ،کتاب رهیافتی بر هوشمندی کسب و کار.

[۱۸] محمد رضا کیوان پور و لیلا خلعتبری ،"مقایسه الگوریتم های کلاس بندی در تشخیص دیابت و نارسایی قلبی "مجموعه مقالات سومین کنفرانس داده کاوی ایران ،دانشگاه علم و صنعت ،آذر ۱۳۸۸ .

[۱۹] حمیدرضا طهماسبی و حسن احمدی ،"افزایش دقت کلاسه بندی در داده کاوی با استفاده از ترکیب کلاسه بندها ".مجموعه مقالات سومین کنفرانس داده کاوی ایران ،دانشگاه علم و صنعت ،آذر ۱۳۸۸ .

[۲۰] علی محمد احمدوند ،بهروز مینایی بیدگلی والهام آخوند زاده ،"تحلیل رضایت مندی شهروندان با استفاده از تکنیک های داده کاوی ".مجموعه مقالات سومین کنفرانس داده کاوی ایران ،دانشگاه علم و صنعت ،آذر ۱۳۸۸ .

dataacademy.ir

آکادمی داده
dataacademy.ir