dataacademy.ir

مقایسه روش های مبتنی بر داده کاوی در تشخیص نوع سرطان سینه

ندا آذری[۱]، رضا احسن[۲] و حامدرحیمی[۳]
[1] دانشجوی کارشناسی نرم افزار، دانشگاه آزاد اسلامی واحد قم ، azari_kardani@yahoo.com
[۲] عضو هیئت علمی و مدیرگروه کامپیوتر ، دانشگاه آزاد اسلامی واحد قم، ahsan@qom-iau.ac.ir
[۳] عضو هیئت علمی ، موسسه آموزش عالی شهاب دانش قم ، info@hamedrahimy.ir

چکیده – داده کاوی تکنیکی جهت کشف دانش جدید از پایگاههای داده می باشد. تشخیص بیماری های مختلف در علم پزشکی یکی از زمینه های پر کاربرد داه کاوی محسوب می شود. سرطان یکی از بیماری هایی است که در دهه های اخیر بسیار گسترش پیدا کرده است . بدین منظور در این مقاله سعی شده است تا ضمن مقایسه روش های داده کاوی نظیر درخت تصمیم گیری و *svm* ، مناسب ترین روش به منظور تشخیص نوع سرطان معرفی و پیشنهاد گردد. بدین منظور از نرم افزار داده کاوی *clementine* و پایگاه داده واقع در مخزن داده دانشگاه کلیفرنیا استفاده شده است. نتایج بدست آمده از دقت مدل های ایجاد شده بر روی داده های آزمایشی نشان می دهد که الگوریتم *c5.0* دقت بالاتری نسبت به سایر الگوریتم ها دارد. با استفاده از الگوریتم *c5.0* می توان با دقت *70.59٪* نوع سرطان سینه را پیش بینی کرد.

کلید واژه- تشخیص سرطان ، درخت تصمیم گیری ، داده کاوی ، svm

1- مقدمه

بطور ساده داه کاوی به معنای استخراج دانش از مقدار زیادی داده خام است. داده کاوی گونه ای از تکنیک ها برای شناسایی و یا دانش تصمیم گیری از قطعات داده است ، به نحوی که با استخراج آنها در حوزه های تصمیم گیری ، پیش بینی ، تخمین و پیشگویی مورد استفاده قرار گیرد. داده ها اغلب حجیم و بدون ارزش هستند و به تنهایی قابل استفاده نیستند بلکه دانش نهفته در آنها قابل استفاده است . [1] از تکنیک هایی که برای داده کاوی بکار گرفته می شود می توان به خوشه بندی ، طبقه بندی کشف قوانین وابستگی و..... اشاره کرد که هر یک در پاره ای از موارد می تواند مفید باشد.

امروزه در دنیای پزشکی کشف و تشخیص سریع و به موقع بیماری بسیار حائز اهمیت است . به طوری که تشخیص زودهنگام بیماری می تواند درمان موثرتری را به همراه داشته باشد . سرطان یکی از بیماری هایی است که در دهه های اخیر بسیار گسترش پیدا کرده است . در این مقاله سعی شده است تا به تشخیص نوع سرطان سینه پرداخته شود. در بخش 2 به بررسی ادبیات موضوع پرداخته می شود. در بخش 3 توضیحی در مورد پایگاه داده سرطان سینه داده می شود. در بخش 4 الگوریتم های داده کاوی بکار گرفته شده شرح داده می شود و در بخش های پایانی خلاصه و نتایج بدست آمده در مقاله مطرح می گردد.

2- مرور ادبیات

در این بخش به معرفی برخی از مطالعاتی که در رابطه با این موضوع انجام شده است ، می پردازیم. یکی از کارهایی که در این زمینه انجام شده است، استفاده از ترکیبSvm وF-Score برای تشخیص سرطان سینه می باشد.[5] در این مقاله ابتدا از F-Score برای تشخیص ویژگیهای بهتر استفاده شده است.سپس ترکیبات مختلفی ازویژگیهای معرفی شده توسط F- Score به عنوان ورودی Svm مورد استفاده قرار می گیرند. سرانجام هر یک از ترکیبات که نتیجه بهتری را در پی داشته باشد ، برای تهیه ی مدل انتخاب می شود. ضمناً برای تشخیص پارامترهای Svm از جستجوی شبکه ای استفاده شده است. استفاده از الگوریتم LS-Svm ، یکی دیگر از روش هایی است که برای تشخیص سرطان سینه مورد استفاده قرار گرفته و منجر به کسب نتایج خوبی در رده بندی مجموعه ی داده wisconsim شده

dataacademy.ir

است . [5] روش دیگر استفاده از یک رهیافت مبتنی بر الگوریتم ژنتیک برای استخراج قوانین تصمیم گیری است. [6] در این مقاله ، برای شناسایی ویژگی های اضافی از روش آماری بهترین زیرمجموعه استفاده شده است . ضمناً دراین روش n تا از بهترین قواعد استخراج شده با هم ترکیب می شوند وبرای تصمیم گیری مورد استفاده قرار می گیرند. روش دیگر استفاده از ترکیب متد k نزدیکترین همسایه ، متد بیز و الگوریتم ژنتیک می باشد. [3] ایده اصلی ، این است که این روش با حذف داده هایی که آموزش را مشکل می کنند ، نتایج موفقیت آمیزی در رده بندی کسب خواهدکرد.[۲] یکی دیگراز کارهایی که در این زمینه انجام گرفته ، استفاده از قوانین انجمنی و شبکه عصبی برای تشخیص سرطان سینه می باشد. دراین مقاله از قوانین انجمنی برای تشخیص وحذف ویژگی های اضافی ، و ازشبکه عصبی برای رده بندی استفاده شده است.

3- پایگاه داده ها

مجموعه داده های پزشکی زیادی درسایت مخزن داده دانشگاه کالیفرنیا به نشانی http://archive.ics.uci.edu قابل دسترس است . مجموعه داده سرطان سینه شامل 286 نمونه داده از دو کلاس خوش خیم و بدخیم می باشد. تعداد نمونه های کلاس خوش خیم 201 مورد وتعداد نمونه های کلاس بدخیم85 مورد می باشدتعداد صفات این مجموعه داده 9 صفت است علاوه برآن ها یک صفت دیگر نیز برچسب کلاس هر نمونه می باشد .
ویژگی هایی که در این مقاله برای طبقه بندی استفاده شده است در جدول شماره 1 لیست گردیده است :

جدول **1**- ویژگی های موجود در مجموعه داده

تعریف ویژگی	نام ویژگی	تعریف ویژگی	نام ویژگی
Age	سن بیمار	Deg-malig	درجه غده
Menopause	زمان یائسگی	Breast	سینه
Node-caps	روکش غده ها	Breast-Quad	چهارقسمت سینه
Inv-node	غده ها	irradiat	قابل انشار
Tumor-size	اندازه تومور	Class	نوع سرطان

4- روش های بکار گرفته شده

ما ازالگوریتم های c 5.0 وsvm برای طبقه بندی استفاده کردیم که درادامه به شرح مختصری ازاین الگوریتم ها می پردازیم .

4– 1– درخت تصمیم گیری

از متداولترین الگوریتم های دسته بندی ، درخت تصمیم می باشدکه از اجزای یادگیری مدرن محسوب می شود. هدف اصلی در درخت تصمیم گیری ، تقسیم داده ها به صورت بازگشتی به زیر مجموعه هایی است به گونه ای که هر زیر مجموعه در برگیرنده وضعیت همگنی ازمتغیرهدف باشد. این الگوریتم پیش بینی هایی را براساس روابط بین ستون ها برای پیش بینی وضعیت یک ستون که به عنوان ستون قابل پیش بینی انتخاب شده است ، استفاده می نماید. در این مقاله برای ایجاد درخت از الگوریتم Tree C5.0 استفاده شده است خروجی الگوریتم یک درخت تصمیم و یا مجموعه ای از قوانین را تولید می کند.این مدل فیلدهایی که مهمترین اطلاعات را دارا هستند ، دسته بندی می کند . هر زیرنمونه توسط اولین دسته ، از دسته اصلی ایجاد می شود (با استفاده از فیلدهای مختلف) این روال تا زمانی تکرار می شود که دیگر زیرمجموعه ها نتوانند به زیر مجموعه دیگری یا کوچکتری تقسیم شوند.سرانجام در پایین ترین قسمت درخت (جایی که برگ است) دوباره تست صورت می پذیرد و درحقیقت برگ هایی که خیلی مهم تشخیص داده نشوند هرس و چیده می شوند. [4]

4-2- ماشین بردار پشتیبان (svm)

درهنگام تلاش برای کشف الگوها و مدل های طبقه بندی ، یادگیری ماشین می تواند یک ابزار قوی به شمار رود. بیشترین استفاده از تکنیک های داده کاوی رده بندی مسائل و مثال های جدید درون کلاس های خاص است. ماشین بردارپشتیبان در واقع یک طبقه بندی کننده دودویی است که دو کلاس را با استفاده از یک مرز خطی از هم جدا می کند . در این روش با استفاده از تمامی باندها و یک الگوریتم بهینه سازی ، نمونه هایی که مرزهای کلاس ها را تشکیل می دهند به دست می آورنداین نمونه ها را بردارهای پشتیبان گویند . تعدادی از نقاط آموزشی که کمترین فاصله تا مرز تصمیم گیری را دارند می توانند به عنوان زیر مجموعه ای برای تعریف مرزهای تصمیم گیری و به عنوان بردار پشتیبان در نظر گرفته شوند. این تکنیک به صورت

dataacademy.ir

موفق برای مشکلاتی که برای طبقه بندی الگوهای مختلف وجود دارد ، بـــه کارگرفتـــه مـــی شـــود. ایـــن روش نیازمنـــد حل مشکل بهینه سازی درجه دوم و نیاز زمان آموزش است.

در این روش مرز خطی بین دو کلاس به گونـه ای محاسـبه مـی شودکه :

1- تمام نمونه های کلاس **1+** در یک طرف مرز و تمام نمونه های کلاس **1-** در طرف دیگر مرز واقع شوند.

2- مرز تصمیم گیری به گونه ای باشد که فاصـله نزدیـک تـرین نمونه های آموزشی هر دو کلاس از یکدیگر در راستای عمـود بـر مرز تصمیم گیری تا جایی که ممکن است.

توابع هسته متعددی همچون تابع پایـه شـعاعی ، چندجملـه ای، سیگموئید و .. وجود دارد که روابط مربوطه به ترتیب در معادلات (1,2,3) نشان داده شده است. [7]

در این مقاله توابع پایه شعاعی و چندجملـه ای بکارگرفتـه شـده است.

$$K(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \quad (1)$$

$$K(x,y) = (x.y+1)^d \quad (2)$$

$$K(x,y) = \tanh(v(x,y)+c) \quad (3)$$

5- ارزیابی

در این بخش ابـزاری کـه بـرای بکـارگیری داده کـاوی اسـتفاده نمودیم معرفی می کنیم و عملیات پـیش پردازشـی کـه قبـل از اجرای الگوریتم های داده کاوی بر روی مجموعـه داده هـا انجـام دادیم توضیح خواهیم داد.در پایان به تحلیل نتایج الگـوریتم هـا می پردازیم.

5-1- محیط عملیاتی

ما برای بکارگیری تکنیکهای داده کاوی از ابزارکلمنتاین اسـتفاده نمودیم. این ابزار الگوریتم های مختلف داده کـاوی ماننـد طبقـه بندی خوشه بندی ، قواعد انجمنی و عملیات پیش پـردازش داده ها و..... را شامل می شود. تفاوت بین این نرم افزار و نرم افزارهای دیگر در پردازش داده ها از طریق بکارگیری تعـدادی گـره اسـت که درقالب یک رویه با یکدیگر ارتباط دارند. علاوه بـرآن ، بعـد از اتمام مراحل نتایج بصورت گرافیکی می توانـد بـه کـاربر نمـایش داده شود. [2]

5-2- پیش پردازش داده ها

دراین مرحله ، رکوردهـای موجـود در مجموعـه داده هـا ، آمـاده فرآیند داده کاوی می شود. ابتدا مشکلات موجود در داده ها کـه بر کیفیت داده اثـر منفـی مـی گـذارد ، بـا برطـرف کـردن ایـن مشکلات ممکن است ویژگی هایی که تا قبل اثر مثبتـی بـر روی متغیر هدف نداشته باشـند بـا برطـرف کـردن مشـکلات موجـود اثرمثبتی بر روی هدف مسئله داشـته باشـند قبـل از توضـیح در مورد مشکلات موجود و نحوه ی برطرف کردن این مشکلات لازم است ذکر شود این مرحله با مرحله انتخاب ویژگی ها در ارتباط است مشکلاتی که اثر منفی بر کیفیت داده ها می گذارند عبارت است از :

1- ناسازگاری موجود در داده ها
2- مقادیر مفقود شده بعضی از ویژگی ها در برخی از رکوردها
3- مقادیر پرت موجود در داده ها

5-2-1- مقادیر مفقود شده

مجموعه داده های پزشکی با تعداد زیـادی از ویژگـی هـا مواجـه است که مقادیر مفقود شده دارند. در مجموعه داده ای که در این پژوهش مورد استفاده قرار گرفته است ، تعداد ویژگی های کامل در مجموعه داده برابر 97٪ است . تعـداد ویژگـی هـایی کـه بـا مشکل مقادیرمفقود شده مواجه هستند، برابر با یک ویژگی اسـت که درجدول شماره 2 به همـراه درصـد مقادیرمفقودشـده آورده شده است.

جدول**2**– ویژگی ها با درصد مقادیر مفقود شده

ویژگی	درصد مقادیر مفقود شده
Node-caps	0.3

چهار سیاست برای رفع مقادیر مفقود شده اتخـاذ شـده اسـت در زیر درمورد هرکدام از سیاست ها توضیح داده شده است.

سیاست اول : اگر یک ویژگی در بیش از 50٪ رکوردها با مقـادیر مفقود شده مواجه باشد ، در این صورت این ویژگـی نمـی توانـد ویژگی موثری در تحلیل هـا باشـد . در نتیجـه چنـین ویژگـی از مجموعه ویژگی ها حذف می شود.

سیاست دوم :اگر یک ویژگی در کمتر از یک درصـد رکوردهـا بـا مقادیر مفقود شده مواجه باشد ، در آن صورت اگر ویژگی از نـوع عددی باشد میـانگین مقـادیر موجـود در آن ویژگـی جـایگیزین مقادیر مفقود شده میشود و اگر ویژگی ازنـوع اسـمی یـا ترتیبـی باشد در این صورت مد مقـادیر موجـود در آن ویژگـی جـایگزین مقادیر مفقود شده ویژگی می شود.

با توجه به جدول بالا ویژگی نام برده شده در کمتر از یک درصـد

dataacademy.ir

رکوردها با مقادیر مفقود شده مواجه است.این ویژگی از نوع اسمی بوده و مقدار مد جایگزین مقادیر مفقود شده می شود. مقدار مد جایگزین شده در جدول شماره 3 آمده است

جدول **3**– ویژگی ها و مقادیر جایگزین شده به جای مقادیر مفقود شده

ویژگی	مقدارجایگزین شده، مقادیر مفقود شده
Node-caps	No

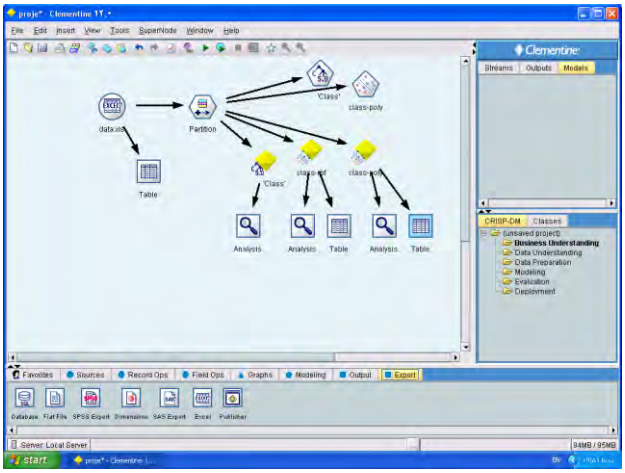
سیاست سوم : اگر ویژگی درکمتر از 10٪ رکوردها با مقادیر مفقود شده مواجه باشد ، در آن صورت با توجه به نوع ویژگی مقدار میانگین یا مد را در مقادیر موجود آن ویژگی در هر کلاس محاسبه میشود و با توجه به کلاس رکوردی که در ویژگی موجود با مقدار مفقود شده مواجه است ، مقدار میانگین یا مد مربوط به همان کلاس جایگزین می شود.

سیاست چهارم: اگر ویژگی در بیش از 10٪ رکوردها با مقادیر مفقود شده مواجه باشد ، در آن صورت از الگوریتم های موجود در طبقه بندی برای برآورد مقادیر مفقود شده در آن ویژگی استفاده می شود.

6- ایجاد مدل

همانگونه که بیان گردید داده ها بوسیله نرم افزار کلمنتاین محاسبه و تحلیل گردیده و دو مدل C 5.0 و SVM با یکدیگر مقایسه گردیده اند.

مدلسازی تحلیل در شکل شماره 1 نمایش داده شده است :



شکل 1: نمای گرافیکی مدل در نرم افزار

با استفاده از node partition آزمون داده ها در مدل ها انجام شده است.

داده ها به دو بخش داده های test و train بصورت تصادفی و با نسبت80 درصد آموزش و20 درصد تست تقسیم شده اند. از الگوریتم های مختلفی برای پیش بینی نوع سرطان سینه استفاده شده است. دقت حاصل از الگوریتمC5.0 در جدول 4 آورده شده است. درجدول 5 و 6 دقت حاصل از اجرای تابع های الگوریتم svm آمده است.

جدول **4**- دقت مدل برای پیش بینی سرطان سینه با الگوریتم C5.0

دقت	آموزش		آزمایش	
صحیح	169	٪77.52	48	٪70.59
غلط	49	٪22.48	20	٪29.41
کل	218	٪100	68	٪100

ساختار شبکه svm دارای 2 تابع Poly (چند جمله ای) و Rbf (پایه شعاعی) می باشد. داده ها بمانند پایین تجزیه شده اند.

جدول **5**- دقت مدل برای پیش بینی سرطان با الگوریتم SVM (تابع rbf)

دقت	آموزش		آزمایش	
صحیح	212	٪97.25	44	٪64.71
غلط	6	٪2.75	24	٪35.29
کل	218	٪100	68	٪100

جدول **6**- دقت مدل برای پیش بینی سرطان با الگوریتم SVM (تابع poly)

دقت	آموزش		آزمایش	
صحیح	216	٪99.08	41	٪60.29
غلط	2	٪0.92	27	٪39.71
کل	218	٪100	68	٪100

7- تحلیل نتایج

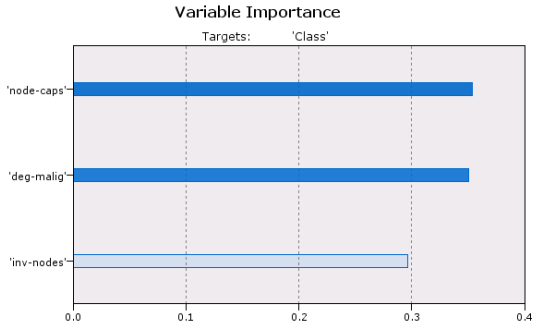
همانطور که در شکل های بالا مشاهده می کنید ، بالاترین دقت برای پیش بینی نوع سرطان سینه ، مربوط به الگوریتم C5.0 است. بعبارت دیگربا استفاده از مدل ایجاد شده توسط الگوریتم C5.0 می توان نوع سرطان سینه بیماران را با دقت ٪70.59 پیش بینی کرد.

8- نتیجه گیری

ازتکنیک های داده کاوی می توان درکشف قوانین و استخراج اطلاعات مفید در داده های پزشکی بهره گرفت . در این مقاله ما

dataacademy.ir

با بکارگیری داده کاوی برروی داده های سرطان سینه برآن بودیم تا اطلاعاتی راکشف نماییم تا بتواند درتشخیص نوع سرطان (خوش خیم یا بدخیم) مورد استفاده پزشکان قراربگیرد. بر اساس داده هایی که در اختیار ما قرارداشت نتایج نشان دهنده آن بودکه مهمترین فیلدهای تاثیر گذار درتشخیص نوع سرطان مطابق شکل شماره 2عبارتند از: روکش غده ها ، درجه غده ، غده ها



شکل 2: فیلدهای تاثیر گذار در تشخیص نوع سرطان

با توجه به اینکه هر چه میزان رکورد های در اختیار بیشتر باشد نتایج بهتری حاصل می گردد استفاده از مجموعه داده کاملتر در تحقیقات آینده سودمند می باشد. همچنین چنانچه داده های جمع آوری شده به صورت پیگیر در فاصله زمانی دراختیار داده کاوی باشد بدین معنا که داده های مربوط به افراد را در فواصل زمانی چندین بار جمع آوری کرده باشیم با دنبال کردن این اطلاعات و واردکردن نقش زمان در آن ها می توان در آینده کاربردهای دیگر را مد نظر قرار داد.

مراجع

[1] D.Hand; h.Mannila and P.Smyth (2001). Principles of Data Mining. MIT Press.
[2] SPSS Inc. "Clementine 12.0 Algorithm Guide". http://www.spss.com 105, 1999.
[3] Aci, Mwhmet; Inan Cigdem; Avci, Mutlu. "Ahybrid classification method of k nearest neighbor, Bayesian methods and genetic algorithm". Expert Systems with Applications, 37(7), pp.5061-5067, July2010.
[4] J.P.Marques de Sa , "Applied statistics: using SPSS, STATISTICA, and MATLAB",Springer, 2003
[5] Akay, Mehmet Fathi."SUPPORT vector Machines combined with feature selection for breast diagnosis". Expert Systems with Applications,36(2),pp.3240-3247,March2009
[6] Chen, Ta-Cheng; Hsu, Tung-Chou." A Gas based approach for mining breast cancer pattern".Expert Systems with Applications,30(4),pp.674-681,May2006
[7] K.Rpbert, S.Mika; "An Introduction of Kernel Based Learning Algorithms", IEEE Transactions on neural Networks, 12(2):pp.181-202, 2001.