

تعیین عوامل موثر در بروز سرطان معده با استفاده از رویکرد داده کاوی

سید عباس محمودی^۱، دکتر کمال میرزائی^۲، دکتر سید مصطفی محمودی^۳

چکیده

زمینه و هدف: سرطان معده دومین علت مرگ ناشی از سرطان در جهان است. با توجه به اینکه این بیماری جزو کشنده ترین بیماری ها در کشور ماست بررسی و شناخت عوامل تاثیرگذار در ایجاد این بیماری، بسیار اهمیت دارد. در این پژوهش از دو تکنیک داده کاوی یعنی الگوریتم Apriori و الگوریتم ID3 به منظور بررسی عوامل موثر در بروز سرطان معده استفاده شده است. **روش بررسی:** مجموعه داده های این پژوهش از ۴۹۰ بیمار شامل ۲۲۰ نمونه ی مبتلا به سرطان و ۲۷۰ نمونه ی سالم مراجعه کننده به بیمارستان امام رضای تبریز جمع آوری شد. با استفاده از الگوریتم Apriori و پیاده سازی آن در نرم افزار متلب، بهترین قوانین حاکم بر روی این مجموعه داده، استخراج شده است. همچنین از الگوریتم ID3 نیز جهت بررسی این عوامل استفاده شد. **یافته ها:** نتایج داده کاوی نشان می دهد که داشتن سابقه رفلاکس معده بیشترین تاثیر را در بروز این بیماری دارد. با استفاده از الگوریتم Apriori قوانینی به دست آمد که می تواند به عنوان الگویی برای پیش بینی وضعیت بیماران و احتمال بروز این بیماری و بررسی عوامل تاثیرگذار در ایجاد این بیماری استفاده شود. همچنین دقت پیش بینی به دست آمده از الگوریتم ID3 برابر ۸۵/۵۶ به دست آمد که نتیجه ی بسیار خوبی در پیش بینی سرطان معده است.

نتیجه گیری: استفاده از داده کاوی به خصوص در داده های پزشکی با توجه به حجم بالای داده ها و وجود روابط ناشناخته بین ویژگی های سیستمیک، شخصی و رفتاری بیماران بسیار مفید است. نتایج حاصل از این پژوهش می تواند به پزشکان در شناسایی عوامل موثر در بروز این بیماری و نیز پیش بینی بروز این بیماری کمک فراوانی کند.

واژه های کلیدی: داده کاوی، قوانین انجمنی، الگوریتم Apriori، سرطان معده، الگوریتم ID3

دریافت مقاله : آذر ۱۳۹۵

پذیرش مقاله : فروردین ۱۳۹۶

*نویسنده مسئول :

سید عباس محمودی؛

دانشکده فنی و مهندسی دانشگاه آزاد اسلامی

واحد یزد

Email :

sa_mahmoodi_85@yahoo.com

^۱ کارشناس ارشد مهندسی کامپیوتر، گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه آزاد اسلامی، واحد یزد، یزد، ایران

^۲ استادیار گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه آزاد اسلامی، واحد میبد، میبد، ایران

^۳ استادیار گروه پاتولوژی دهان، دانشکده دندانپزشکی، دانشگاه علوم پزشکی شهید صدوقی یزد، یزد، ایران

مقدمه

کمک زیادی کرده است، داده کاوی و تکنیک هایی است که در این علم گنجانده شده است. در دنیای پزشکی داده های مربوط به علائم بیماران مبتلا به بیماری های گوناگون و روش های کمکی برای تشخیص این بیماری ها بسیار وسیع و گسترده می باشد تا جایی که معمولاً تحلیل و در نظر گرفتن همه جانبه ی تمامی عوامل دخیل توسط یک فرد دشوار به نظر می رسد. استخراج دانش از میان حجم انبوه داده های مرتبط با سوابق بیماری و پرونده های پزشکی افراد با استفاده از فرایند داده کاوی، می تواند منجر به شناسایی قوانین حاکم بر ایجاد، رشد و تسریع بیماری ها گردیده و اطلاعات ارزشمندی را به منظور شناسایی علل رخداد بیماری ها، پیش بینی و درمان بیماری ها با توجه به عوامل محیطی حاکم در اختیار متخصصان و دست اندرکاران حوزه ی سلامت قرار دهد.

داده کاوی و کشف دانش در پایگاه داده ها، رویکردی برای یافتن روابط و الگوهای پنهان داده هاست (۲۵). این علم، فرایند استخراج الگوها از مجموعه داده های بزرگ با ترکیب روش هایی از آمار، هوش مصنوعی و مدیریت پایگاه داده است (۲۶). با استفاده از دانش داده کاوی می توان به پزشکان و بیماران در این حوزه کمک کرد. داده کاوی رشته ی علمی جدید در زمینه ی بازیابی اطلاعات از پایگاه داده ها می باشد. مطالعات معدودی در زمینه ی سرطان معده با استفاده از دانش داده کاوی انجام گرفته است. مهم ترین تحقیقی که در سال اخیر در مورد کشف عوامل موثر در بروز سرطان معده انجام شده، استفاده از الگوریتم درخت تصمیم است (۲۷). نتایج نشان داد که رفلاکس معده، مهم ترین عامل بروز سرطان معده است. همچنین فاکتورهای غذایی، مصرف مشروبات الکلی، مصرف سیگار، سن و درآمد از عوامل خطر بروز سرطان معده است. Ramachandran و همکارانش به بررسی و تجزیه و تحلیل تاثیر عوامل موثر در بروز عفونت هلیکوباکتر پیلوری، با استفاده از الگوریتم های دسته بندی پرداخته اند (۲۸). عفونت هلیکوباکتر پیلوری از عوامل اصلی در بروز سرطان معده است. در این مطالعه از سه الگوریتم داده کاوی C4.5، KNN و CN2 استفاده شد. نتایج نشان داد که بهترین الگوریتم برای دسته بندی الگوریتم C4.5 است. این الگوریتم نشان داد که خطر ابتلا به عفونت هلیکوباکتر پیلوری، عمدتاً به دلیل تغییرات استرس و سبک زندگی ناشی می شود. افرادی که در محیط های دارای ترکیبات شیمیایی خطرناک و حجم کار سنگین فعالیت دارند در معرض عفونت هلیکوباکتر پیلوری قرار دارند. همچنین از عوامل دیگر ابتلا به عفونت هلیکوباکتر پیلوری عدم فعالیت بدنی عنوان شده است.

سرطان یکی از علت های منجر به مرگ در سرتاسر جهان است. بعد از بیماری های قلبی و عروقی، سرطان دومین علت مرگ و میر در بیشتر کشورهاست (۱). در بین سرطان ها، سرطان معده یکی از علل عمده ی مرگ ناشی از سرطان می باشد (۲). از نظر میزان بروز، سرطان معده چهارمین سرطان شایع و دومین سرطان منجر به مرگ در جهان است (۳ و ۴). بروز این سرطان در دنیا (خصوصاً در کشورهای توسعه یافته) روندی کاهشی دارد (۵ و ۶). به طوری که در امریکا میزان بروز این سرطان در دهه های اخیر رو به کاهش می باشد (۷ و ۸). این وضعیت در کانادا نیز مشاهده شده است و از سال ۱۹۸۴ تا ۲۰۱۳ میزان بروز این سرطان از ۱۸/۴ به ۹/۵ در هر ۱۰۰ هزار نفر کاهش یافته است (۹). در سایر کشورهای اروپایی نیز این سرطان شایع نبوده حال آنکه در کشورهای آسیایی و در حال توسعه بروز این سرطان در حال افزایش می باشد (۱۰ و ۱۱). در ایران برخلاف کشورهای پیشرفته میزان بروز سرطان معده در حال افزایش است. این افزایش خصوصاً در غرب ایران قابل توجه و به عنوان یک معضل مطرح می باشد (۱۲ و ۱۳). این افزایش در سایر مناطق کشور نیز مشاهده گردیده است (۱۴). در ایران نیز این سرطان دارای اهمیت بالایی است و از مشکلات عمده ی بهداشتی و سلامت است (۱۲ و ۱۳). براساس آخرین تحقیقات انجام شده در ایران، در سال ۸۸ سرطان معده با ۹/۳٪ سومین سرطان شایع در کشور در مجموع زنان و مردان است (۱۵). با توجه به میزان شیوع این بیماری و میزان مرگ و میر بالای سرطان معده در کشور لازم است با بررسی علل و عوامل تأثیرگذار در بروز این بیماری، این بیماری مورد بررسی بیشتری قرار گیرد.

مطالعات مختلفی در زمینه ی بررسی عوامل مختلف بروز این بیماری انجام شده است. تحقیقات نشان می دهد عفونت هلیکوباکتر پیلوری عامل بسیار مهم و تأثیرگذاری در ایجاد خطر ابتلا به سرطان معده بوده است (۱۶). به طور کلی عفونت اولیه هلیکوباکتر پیلوری باعث ایجاد یک گاستریت خفیف می گردد. در بعضی افراد این التهاب، به زخم معده منجر خواهد شد. در صورت ادامه ی روند بیماری زایی و عدم درمان زخم معده، التهاب آتروفیک معده ایجاد خواهد شد. افرادی که دچار این نوع التهاب هستند در خطر ایجاد بدخیمی و سرطان قرار دارند (۱۷). در برخی از مطالعات عوامل ژنتیکی را به عنوان پایه مطرح کرده اند (۱۸ و ۱۹). برخی دیگر فاکتورهای محیطی (۲۰ و ۲۱) و تعدادی نیز رژیم غذایی (۲۲ و ۲۳) و مصرف سیگار را ذکر کرده اند (۲۴). یکی از دانش هایی که در کنار علم پزشکی به بیماران و پزشکان

مراجعه کننده به بیمارستان امام رضای تبریز بین سال های ۹۲-۹۰ است. جامعه ی مورد مطالعه بیماران دارا و فاقد سرطان معده بودند که به بیمارستان مراجعه کرده بودند. نمونه گیری به شیوه ی تصادفی و حجم نمونه با استفاده از فرمول (۱)، فرمول حجم نمونه برای برآورد یک نسبت، ۴۷۴ به صورت زیر محاسبه شد ($p=0/5$, $Z=1/96$, $d=0/045$, $q=0/5$).

$$(1) n = \frac{Z^2 p (1-q)}{d^2}$$

در نهایت، جهت اطمینان بیشتر تعداد ۴۹۰ نمونه انتخاب شد. پایگاه داده اولیه مورد نظر با فرمت اکسل می باشد که این اطلاعات از طریق پرسش نامه جمع آوری شده است. مجموعه داده های اولیه شامل ویژگی (مثل نام، سن، شغل، سابقه ی سرطان در فامیل، سابقه ی سرطان معده در فامیل، مصرف نمک و غیره) قبل از پیش پردازش است. پس از دریافت داده ها و شناخت مفاهیم مربوط به داده ها فاز پیش پردازش داده ها شروع می شود. در این مرحله برحسب نیاز عملیات های مربوط به پیش پردازش بر روی داده ها اعمال شد که در ادامه بیان می شود.

یکی از مشکلات شایعی که در مجموعه داده های اولیه است، پایین بودن کیفیت آن است. به عملیاتی که به برطرف شدن مشکل کیفیت داده ها می انجامد، پاکسازی داده گفته می شود. یکی از مهم ترین مواردی که باعث کاهش کیفیت داده می شود، وجود مقادیر از دست رفته (Missing Values) است. به دلایلی ممکن است بعضی از مقادیر مربوط به برخی ویژگی ها Null باشند. به این گونه مقادیر، مقادیر از دست رفته می گویند. برای مدیریت مقادیر از دست رفته در این مرحله رکوردهایی که حداقل یکی از ویژگی های آن ها Null بود، حذف شد. مجموعه ویژگی های مورد استفاده در این پژوهش به همراه بازه مقادیر آن ها در جدول ۱ نشان داده شده است.

جدول ۱: ویژگی های شصتی و رفتاری

ردیف	نوع ویژگی	نام ویژگی	مقدار
۱	ویژگی های شخصی و رفتاری	جنس	مرد، زن
۲		گروه خونی	A, B, AB, O
۳		مصرف سیگار	بله، خیر
۴		مصرف الکل	بله، خیر
۵		در معرض مواد شیمیایی	بله، خیر
۶		وزن	$BMI < 18/5$, $BMI > 24/9$, $18/5 < BMI < 29/18$, $BMI > 30$, $BMI > 25$
۷		میزان تحرک	سبک، متوسط، زیاد
۸		سن	زیر ۴۰ سال، بین ۴۱ تا ۶۰ سال، ۶۱ به بالا
۹		مصرف نمک	نمی خورد، زیاد، کم

Kirshners و همکارانش در مطالعه ی خود یک روش چند لایه با ترکیب تکنیک های خوشه بندی و درخت تصمیم یک سیستم تشخیص و پیش بینی خطر ابتلا به سرطان های ریه، پستان، دهان، گردن رحم، خون و سرطان معده ارایه کرده اند (۲۹). نتایج حاصل از درخت تصمیم در این مطالعه نشان داد که جنسیت افراد، خطرات شغلی، کاهش وزن، سابقه ی خانوادگی، مصرف الکل، درد شکم همراه با خون در مدفوع از عوامل و علایم سرطان معده است.

در مطالعه ی Silvera در سال ۲۰۱۴ که به بررسی نقش عوامل خطر تغذیه ای بر ابتلا به سرطان مری و معده با استفاده از مدل رده بندی درختی انجام شده است، رفلکس مری- معده ای به عنوان یکی از مهم ترین عوامل خطر ابتلا به آدنوکارسینومای کاردیا و غیر کاردیاست. همچنین مصرف گوشت قرمز، میوه هایی غیر از مرکبات، چای سیاه و سبزیجات خام به عنوان عوامل خطر ابتلا به سرطان مری در این مدل مطرح بوده است (۳۰). با توجه به میزان شیوع این بیماری و میزان مرگ و میر بالای سرطان معده در کشور، لازم است علل و عوامل تأثیر گذار در بروز این بیماری، با دقت بیشتر و روش های علمی تر، بررسی شود. هدف این مقاله استخراج قوانین و الگوهای پنهان از داده های سرطان معده است که به کمک آن می توان بدون نیاز به روش های تشخیصی، احتمال ابتلا به این بیماری را تشخیص داد و نیز عوامل موثر در این بیماری را شناسایی کرد.

روش بررسی

مطالعه ی حاضر از نوع توصیفی- مقطعی بوده و در برگرنده ی دو مرحله ی اصلی است. مرحله اول جمع آوری و آماده سازی داده و مرحله دوم طراحی و پیاده سازی الگوریتم های داده کاوی است. مجموعه داده های مورد استفاده، داده های جمع آوری شده از بیماران



روزانه، بین ۱ تا ۳ بار در هفته، بین ۱ تا ۳ بار در ماه	مصرف سبزی	۱۰
نمی خورد، بین ۱ تا ۳ بار در هفته، بین ۱ تا ۳ بار در ماه	مصرف غذاهای دودی شده	۱۱
دارد، ندارد	مصرف شیر	۱۲
نمی خورد، بین ۱ تا ۳ بار در هفته، بین ۱ تا ۳ بار در ماه	مصرف فست فود	۱۳
نمی خورد، بین ۱ تا ۳ بار در هفته، بین ۱ تا ۳ بار در ماه	مصرف غذاهای سرخ شده	۱۴
روزانه، بین ۱ تا ۳ بار در هفته، بین ۱ تا ۳ بار در ماه	مصرف میوه	۱۵
بله، خیر	سابقه ی آلرژی	۱۶ ویژگی های سیستمیک
بله، خیر	سابقه ی سرطان در فامیل	۱۷
بله، خیر	سابقه ی سرطان معده در فامیل	۱۸
بله، خیر	سابقه ی بیماری های قلبی عروقی	۱۹
بله، خیر	سرطان معده	۲۰
خوب، متوسط، ضعیف	وضعیت عمومی سرطان	۲۱
بله، خیر	سابقه ی رفلاکس معده	۲۲
بله، خیر	سابقه ی جراحی معده	۲۳
بله، خیر	سابقه ی التهاب معده	۲۴
بله، خیر	سابقه ی عفونت معده	۲۵
نرمال، ورم، سرخ و قرمز، زخم	وضعیت مخاط	۲۶
کاردیا، غیرکاردیا	محل سرطان	۲۷

که تمام ویژگی ها به ویژگی های دودویی تبدیل شدند. به عنوان مثال در جدول ۱، ۳ نوع ویژگی (سن، گروه خونی، مصرف سیگار) که به ترتیب ویژگی های پیوسته، دسته ای، دودویی هستند مشاهده می شود. ویژگی سن به ۳ دسته (زیر ۴۰ سال، ۴۱ تا ۶۰ سال، ۶۱ به بالا) گسسته سازی شده است. ویژگی دسته ای گروه خونی به ۴ دسته A، B، AB، O و ویژگی دودویی مصرف سیگار به دو مقدار (بله، خیر) دسته بندی شده است. حال یک مجموعه داده جدید ایجاد شد که شامل (زیر ۴۰ سال، ۴۱ تا ۶۰ سال، ۶۱ به بالا)، A، B، AB، O، مصرف سیگار دارد، مصرف سیگار ندارد است. هر ویژگی، دودویی و مفهوم مشخصی دارد.

مرحله ی دوم پژوهش، پیاده سازی الگوریتم های داده کاوی با استفاده از نرم افزار MATLAB است. در این پژوهش از دو الگوریتم داده کاوی کشف قوانین انجمنی و درخت تصمیم ID3 استفاده شد که در ادامه، مختصر بیان می شود.

کشف قوانین انجمنی (Association Rules Mining)

در این روش، هدف، شناسایی الگوهای مفید و جالب در پایگاه داده است. قوانین انجمنی یک جمله ی شرطی به صورت $A \rightarrow B$ بوده که A و B مجموعه ویژگی ها هستند. به طور کلی کشف قوانین انجمنی در پایگاه داده های بزرگ، جهت یافتن رابطه یا ارتباط بین ویژگی های موجود در پایگاه داده است. شکل ۱ الگوریتم Apriori را نشان می دهد.

در مرحله ی کاهش داده، ویژگی های اضافی مثل نام و نام خانوادگی، شماره پرونده، آدرس، تلفن و وضعیت تأهل که متغیرهای تاثیرگذاری نیستند، حذف شد. همان طور که در جدول ۱ مشاهده می شود، کلیه ویژگی های پیوسته در این مجموعه داده به ویژگی های گسسته تبدیل شد. به عنوان مثال ویژگی سن در این مجموعه داده که یک ویژگی پیوسته است به ۳ بازه (زیر ۴۰ سال، ۴۱-۶۰ و ۶۱ به بالا) دسته بندی شد. در مرحله ایجاد ویژگی، با ترکیب کردن ویژگی ها، ویژگی های جدیدی ایجاد می شوند که بار اطلاعاتی بیشتری دارند. به عنوان مثال در مجموعه ویژگی ها دو ویژگی قد و وزن بود. ویژگی معروف (Body Mass Index) BMI یا شاخص توده بدن با استفاده از این دو ویژگی ساخته شد که از طریق فرمول (۲) محاسبه می شود.

$$BMI = \frac{\text{وزن (kg)}}{\text{قد (m)}^2} \quad (2)$$

همان طور که در جدول ۱ مشاهده می شود، این ویژگی به مجموعه ویژگی های قبلی اضافه شد. در مرحله ی آخر تمام ویژگی ها به ویژگی های دودویی تبدیل شدند. به این صورت که هر ویژگی اسمی یا ترتیبی به چندین ویژگی دودویی شکسته شد. در واقع بعد از گسسته سازی، ویژگی های پیوسته به چندین ویژگی دسته ای شکسته می شود. سپس هر ویژگی چندین مقادیر دسته ای و هر مقدار دسته ای باید یک ویژگی دودویی داشته باشد. سرانجام یک پایگاه داده جدید ایجاد شد

```

1:  $K=1$ .
2:  $F_K = \{i \mid i \in I \wedge \frac{\sigma(\{i\})}{N} \geq \text{minsup}\}$ . {Find all frequent 1- itemsets}
3: repeat
4:  $K = K + 1$ .
5:  $C_K = \text{apriori-gen}(F_{K-1})$ . {Generate candidate itemsets}
6: for each transaction  $t \in T$  do
7:  $C_t = \text{subset}(C_K, t)$ . {Identify all candidates that belong to  $t$ }
8: for each candidate itemset  $c \in C_t$  do
9:  $\sigma(c) = \sigma(c) + 1$ . {Increment support count}
10: end for
11: end for
12:  $F_K = \{c \mid c \in C_K \wedge \frac{\sigma(c)}{N} \geq \text{minsup}\}$  {Extract the frequent  $K$ - itemsets}
13: Until  $F_K = \emptyset$ 
14: Answer =  $\cup F_K$ .

```

شکل ۱: شبه‌کد الگوریتم Apriori (۲۵)

باشد که همگی یک کلاس داشته باشند ولی هر برگ برای قرار گرفتن در دسته مورد نظر، علت متفاوتی دارد. یکی از الگوریتم‌های پایه‌ای درخت تصمیم، الگوریتم ID3 است (۲۶). در این الگوریتم داده‌ها باید دارای مقادیر گسسته باشند. انتخاب اساسی در الگوریتم ID3، انتخاب یک صفت برای آزمایش در هر گره در درخت است. این صفت باید به دسته‌بندی مثال‌ها کمک کند. برای معرفی یک معیار کمی یا عددی خوب برای بیان ارزش یک صفت، یک ویژگی آماری به نام بهره‌ی اطلاعات (Information Gain) تعریف می‌شود که وظیفه‌ی آن، اندازه‌گیری کیفیت تقسیم‌کنندگی مثال‌های آموزشی توسط یک صفت می‌باشد. استراتژی جستجوی ID3:

(۱) درختان کوتاه‌تر را به بلندتر ترجیح می‌دهد.

(۲) درختانی که صفاتی با بالاترین بهره‌ی اطلاعاتی را نزدیک‌تر به ریشه جای می‌دهند، برمی‌گزیند.

حساسیت، ویژگی و دقت، سه معیار مهم در تحقیقات پزشکی هستند که به طور معمول استفاده می‌شود. دقت مدل یا نرخ دسته‌بندی حساسیت و ویژگی با توجه به جدول ۲ محاسبه می‌شود.

الگوریتم Apriori یکی از محبوب‌ترین الگوریتم‌ها در کشف قوانین انجمنی است (۲۵). در حقیقت این الگوریتم مبنای دیگر الگوریتم‌های کشف قوانین انجمنی است.

الگوریتم درخت تصمیم ID3: درخت تصمیم‌گیری یکی از ابزارهای قوی و متداول برای دسته‌بندی و پیش‌بینی است. درخت تصمیم به تولید قانون می‌پردازد. درخت تصمیم به این صورت است که یک گره ریشه در بالای آن کشیده شده و برگ‌های آن در پایین می‌باشند. یک رکورد در گره ریشه وارد می‌شود و در این گره یک آزمون صورت می‌گیرد تا معلوم شود که این رکورد به کدامیک از گره‌های فرزند (شاخه پایین‌تر) می‌رود. معمولاً روش‌های مختلفی برای انتخاب این آزمون اولیه وجود دارد، ولی هدف همه آن‌ها یکی است: انتخاب روشی که بهترین جداسازی را در کلاس‌های هدف انجام دهد. این فرایند آنقدر ادامه می‌یابد تا رکورد به گره برگ برسد. تمام رکوردهایی که به یک برگ از درخت می‌رسند در یک کلاس قرار می‌گیرند. همچنین برای رسیدن از ریشه به یک برگ تنها، یک راه وجود دارد و آن راه در واقع بیان قانونی است که برای دسته‌بندی رکوردها ایجاد شده است. ممکن است تعداد زیادی برگ وجود داشته

جدول ۲: ماتریس درهم‌ریختگی

نوع دسته	دسته تشخیص داده شده	
	مثبت	منفی
مثبت	TP	FN
دسته واقعی منفی	FP	TN

نشان دهنده ی درصد موفقیت روش دسته بندی کننده در تشخیص نمونه های مربوط به هرکدام از دسته هاست. نرخ فراخوانی یا ویژگی که همانند معیار قبل برای هر کدام از دسته های موجود محاسبه می گردد، درصد قابلیت اعتماد به خروجی روش دسته بندی کننده را نشان می دهد(۳).

$$(۴) \text{ Precision} = \frac{TP}{TP+FP}$$

$$(۵) \text{ Recall} = \frac{TP}{TP+FN}$$

یافته ها

نتایج این تحقیق از جنبه های مختلفی حایز اهمیت است. این که: این تحقیق یکی از مسایل واقعی را پوشش می دهد، به این معنی که داده های مورد بررسی واقعی بوده، اهداف و مطالب متناسب با یک مساله واقعی تدوین شده و نتایج آن نیز مورد تأیید کارشناسان این حوزه قرار گرفته است. در این تحقیق برای رسیدن به یک نتیجه ی قابل قبول یک فرایند کشف دانش از داده های واقعی طراحی و اجرا شد. این فرایند شامل پیش پردازش داده ها، آماده سازی، یکپارچه سازی داده ها، کشف الگوهای مکرر و تفسیر قوانین به دست آمده بود. هر یک از این مراحل، مستلزم صرف وقت و دقت بسیار است. همان طور که قبلاً هم بیان شد الگوریتم Apriori یکی از مهم ترین الگوریتم های داده کاوی در حوزه ی کشف قوانین انجمنی است. جهت بررسی و ارزیابی این تحقیق، مقدار درجه پشتیبانی و اطمینان مینیم به ترتیب ۰/۲ و ۰/۹ در نظر گرفته شد. تعدادی از بهترین قوانین به دست آمده در این مطالعه در جدول ۳ مشاهده می شود.

دسته ها در مسأله ی تشخیص سرطان با چهار دسته ی مثبت و منفی و چهار عدد TP, FN, FP, TP با توجه به نوع دسته مثبت و منفی محاسبه می گردند.

TP، شامل نمونه هایی است که جزو نمونه های دسته ی مثبت است و الگوریتم، آن را به درستی در دسته ی مثبت تشخیص داده است.

FP، شامل نمونه هایی است که جزو نمونه های دسته ی منفی است و الگوریتم آن را به صورت نادرستی در دسته ی مثبت تشخیص داده است.

FN، شامل نمونه هایی است که جزو نمونه های دسته ی مثبت است و الگوریتم آن را به صورت نادرستی در دسته ی منفی تشخیص داده است.

TN، شامل نمونه هایی است که جزو نمونه های دسته ی منفی است و الگوریتم آن را به درستی در دسته ی منفی تشخیص داده است. نرخ دسته بندی بیانگر دقت الگوریتم پیاده سازی شده در دسته بندی دسته های مختلف موجود در مسأله ی تشخیص سرطان است. این معیار در واقع درصد دسته بندی درست الگوریتم می باشد. به عبارتی نرخ دسته بندی، تعداد نمونه هایی است که به درستی دسته بندی شده است و نسبت تعداد نمونه هایی است که به درستی تشخیص داده می شوند به کل نمونه ها:

$$(۳) \text{ Classification Rate} = \frac{(TP+TN)}{(TP+TN+FN+FP)}$$

نرخ صحت یا میزان حساسیت که برای هر کدام از دسته های موجود قابل محاسبه می باشد، جهت تعیین دقت دسته بندی برای هر کدام از دسته ها در نظر گرفته شده است. در واقع این معیار

جدول ۳: بهترین قوانین به دست آمده از الگوریتم Apriori

درجه پشتیبان	نتیجه	قانون
۰/۲۰۶۱۲	بیمار بودن	سابقه ی رفلاکس معده دارد، سابقه ی عفونت معده ندارد
۰/۲۱۰۲۰	بیمار بودن	سابقه ی بیماری قلبی عروقی ندارد، سابقه ی سرطان در فامیل ندارد
۰/۱۷۷۵۵	بیمار بودن	سابقه ی رفلاکس معده دارد، سابقه ی عفونت معده ندارد
۰/۱۸۷۷۵	بیمار بودن	سابقه ی بیماری قلبی عروقی ندارد، سابقه ی سرطان در فامیل ندارد
۰/۱۷۷۵۵	بیمار بودن	سابقه ی بیماری قلبی عروقی ندارد، سابقه ی سرطان در فامیل ندارد
۰/۱۸۷۷۵	بیمار بودن	سابقه ی بیماری قلبی عروقی ندارد، سابقه ی سرطان در فامیل ندارد
۰/۲۲۲۴	بیمار بودن	سابقه ی رفلاکس معده دارد
۰/۲۰۲۰	بیمار بودن	سابقه ی بیماری قلبی عروقی ندارد
۰/۲۰۶۱۲	بیمار بودن	سابقه ی عفونت معده ندارد
۰/۲۲۴۴	بیمار بودن	سابقه ی رفلاکس معده دارد
۰/۲۱۴۲	بیمار بودن	سابقه ی بیماری های قلبی عروقی ندارد

زمینه‌ی ریسک فکتورهای سرطان معده انجام شد (۳۱)، نتایج نشان داد که ارتباط معناداری بین میزان مصرف نمک در رژیم غذایی و سرطان معده وجود دارد. نتایج این تحقیق هم با نتایج پژوهش ما منطبق است. در هر دو تحقیق انجام شده ارتباط معناداری بین میزان مصرف نمک و سرطان معده وجود دارد.

در مطالعات اخیر (۳۲) محققان به این نکته دست یافته اند که بیماران قلبی عروقی به علت مصرف یک سری داروها در معرض خطر کمتری برای ابتلا به سرطان معده هستند. همان طور که مشاهده شد نتایج تحقیق ما نیز نشان داد که افراد مبتلا به بیماری قلبی و عروقی کمتر در معرض خطر ابتلا به سرطان معده هستند.

از نکات جذاب این پژوهش این است که علاوه بر نتایج به دست آمده که بیان شد، با استفاده از قوانین ایجاد شده می توان برای یک نمونه‌ی جدید با ویژگی‌های مشخص می تواند پیش بینی کرد که این فرد احتمالاً در آینده دچار این بیماری خواهد شد که با کنترل عوامل تاثیرگذار در بروز این بیماری می توان امیدوار بود که از بروز این بیماری تا حدی اجتناب کرد. یکی از محدودیت‌هایی که در این پژوهش وجود داشت داده‌های ناقص بود. همچنین یکی دیگر از محدودیت‌هایی که در این مطالعه با آن مواجه بودیم تعداد کم بیماران مورد بررسی بود که باید در مطالعات بعدی با تعداد بیشتری از بیماران توسعه یابد. همچنین نقطه قوت این مطالعه، استفاده از یک مجموعه داده‌ی واقعی بیماران است که معمولاً کمتر در مطالعات مشابه صورت می‌گیرد. زیرا معمولاً در کشور ما، پژوهشگرانی که به مطالعاتی از این دست می‌پردازند بیش از آنکه جنبه‌ی پزشکی آن را در نظر بگیرند، به دنبال تست مدلها و متدهای آن در حوزه‌ی مهندسی هستند و بیشتر از مجموعه داده‌هایی بهره می‌برند که ماهیت واقعی آنها کمتر بومی و یا غیربومی هستند.

نتیجه گیری

با استفاده از داده‌های کاوی می توان کمک فراوانی به پزشکان در بررسی علل و عوامل موثر در بروز بیماری‌ها و همچنین روش‌های پیشگیری از بیماری‌ها کرد. هدف از این مقاله بررسی عوامل تاثیرگذار در ایجاد بیماری سرطان معده با استفاده از تکنیک‌های داده‌کاوی بود. در این پژوهش با استفاده از الگوریتم Aprior بهترین قوانین حاکم بر مجموعه داده‌های بیماران مراجعه کننده به بیمارستان امام رضای تبریز کشف و علل و عوامل تاثیرگذار در ایجاد بیماری سرطان معده، بررسی شد. همچنین نتایج حاصل از الگوریتم ID3 نشان از توان بالای این مدل در پیش‌بینی احتمال بروز این بیماری در فرد با

در ضمن مقدار اطمینان هر یک از قوانین ۱۰۰ درصد است. قابل توجه است که انتخاب این قوانین از بین تعداد قوانین زیادی که در این آزمایش تولید شده، زیر نظر کارشناسان پاتولوژی انجام شده است. در بررسی عوامل موثر بر ابتلا به سرطان معده با استفاده از الگوریتم ID3، نتایج نشان داد که سابقه‌ی جراحی معده، گروه خونی، سابقه‌ی رفلاکس معده و مصرف سیگار از عوامل مهم در بروز سرطان معده است. در بررسی کارایی مدل با استفاده از مساحت زیر منحنی Roc Area، مقدار مساحت ۸۵/۳ درصد تعیین شد که نشان دهنده‌ی توان بالای الگوریتم ID3 در تعیین عوامل موثر در ابتلا به سرطان معده است. همچنین در بخش ارزیابی این روش، مقادیر حساسیت، ویژگی و دقت به ترتیب ۸۳/۸، ۷۹/۶ و ۸۵/۵۶ به دست آمد که نشان از توان بالای این مدل در پیش‌بینی احتمال بروز این بیماری در فرد با توجه به ویژگی‌های بیان شده در این مطالعه است. با کنترل عوامل تاثیرگذار در بروز این بیماری می توان امیدوار بود که از بروز بیماری تا حدی اجتناب کرد یا آن را به تعویق انداخت.

بحث

قوانین به دست آمده، نشان می‌دهد که داشتن سابقه‌ی رفلاکس معده بیشترین تأثیر را در بروز این بیماری دارد. افراد مبتلا به بیماری قلبی و عروقی کمتر در معرض خطر ابتلا به سرطان معده هستند. در ضمن رفلاکس معده با مصرف نکردن نمک، مصرف زیاد نمک و مصرف نکردن شیر ارتباط دارد. همچنین از الگوریتم ID3 نیز جهت بررسی این عوامل استفاده شد. ID3 نشان داد که که سابقه‌ی جراحی معده، گروه خونی، سابقه‌ی رفلاکس معده، مصرف سیگار از عوامل مهم در بروز سرطان معده است. همچنین در بخش ارزیابی این روش، مقادیر حساسیت، ویژگی و دقت به ترتیب ۸۳/۸، ۷۹/۶ و ۸۵/۵۶ به دست آمد که نشان از توان بالای این مدل در پیش‌بینی احتمال بروز این بیماری در فرد با توجه به ویژگی‌های بیان شده در این مطالعه است. مطالعه‌ی در زمینه‌ی استفاده از قوانین انجمنی تاکنون در حوزه‌ی سرطان معده در کشور انجام نشده است. حتی مطالعات محدود انجام شده در نقاط دیگر دنیا با الگوریتم‌های متفاوت داده‌کاوی انجام شده است. Navarro Silvera و همکارانش در مطالعه‌ی در سال ۲۰۱۴ از الگوریتم درخت تصمیم استفاده کردند (۲۷). نتایج نشان داد که رفلاکس معده، مهم‌ترین عامل بروز سرطان معده است. همان طور که مشاهده می‌شود نتایج به دست آمده از این تحقیق کاملاً منطبق با نتیجه‌ی پژوهش ما می‌باشد. در مطالعه‌ی که در سال ۲۰۱۲ توسط D'Elia و همکارانش در



تشکر و قدردانی

این مقاله برگرفته از پایان نامه ی کارشناسی ارشد در سال ۹۳ دانشگاه آزاد اسلامی واحد یزد به شماره ثبت ۶۲۴۴۱۰۰۶۹۱۱۰۳۰ می باشد. پژوهشگران از زحمات کلیه عزیزانی که در اجرای این طرح همکاری نمودند به ویژه دکتر مرتضی قوجازاده عضو محترم هیات علمی دانشگاه علوم پزشکی تبریز تشکر و قدردانی می نمایند.

توجه به ویژگی های بیان شده در این مطالعه است. پیشنهاد می شود جهت یافتن نتایج بهتر، از پایگاه داده های بزرگتر و همچنین استفاده از ویژگی های دیگری مانند نتایج آزمایش ها استفاده شود. همچنین استفاده از روش های انتخاب ویژگی و ترکیب آن با روش ارایه شده در این تحقیق جهت کاهش تعداد ویژگی ها و انتخاب ویژگی های مهم و تاثیرگذار در ایجاد این بیماری، یکی دیگر از راه های ارتقای این پژوهش است.

منابع

1. Shils ME, Shike M, Ross AC, Caballero B & Cousins RJ. Modern nutrition in health and disease. 10th ed. Philadelphia: Lippincott Williams & Wilkins; 2006: 1290-1.
2. Ushijima T & Sasako M. Focus on gastric cancer. *Cancer Cell* 2004; 5(2): 121-5.
3. Parkin DM, Bray F, Ferlay J & Pisani P. Global cancer statistics, 2002. *A Cancer Journal for Clinicians* 2005; 55(2): 74-108.
4. Murphy G, Pfeiffer R, Camargo MC & Rabkin CS. Meta-analysis shows that prevalence of epstein-barr virus-positive gastric cancer differs based on sex and anatomic location. *Gastroenterology* 2009; 137(3): 824-33.
5. Azizi F, Hatami H & Janghorbani M. Epidemiology and control of common diseases in Iran. 3th ed. Tehran: Khosravi Publication; 2010: 45-7[Book in Persian].
6. Bertuccio P, Chatenoud L, Levi F, Praud D, Ferlay J, Negri E, et al. Recent patterns in gastric cancer: A global overview. *Int J Cancer* 2009; 125(3): 666-73.
7. Baranovsky A & Myers MH. Cancer incidence and survival in patients 65 years of age and older. *A Cancer Journal for Clinicians* 1986; 36(1): 26-41.
8. Fuchs CS & Mayer RJ. Gastric carcinoma. *The New England Journal of Medicine* 1995; 333(1): 32-41.
9. Canadian Cancer Society. Canadian cancer statistics 2015. Available at: <http://www.cancer.ca/~media/cancer.ca/CW/cancer%20information/cancer%20101/Canadian%20cancer%20statistics/Canadian-Cancer-Statistics-2015-EN.pdf?la=en>. 2015.
10. Murray CJ & Lopez AD. Alternative projections of mortality and disability by cause 1990-2020: Global burden of disease study. *The Lancet* 1997; 349(9064): 1498-504.
11. Hartgrink HH, Jansen EP, Van Grieken NC & Van de Velde CJ. Gastric cancer. *The Lancet* 2009; 374(9688): 477-90.
12. Sadjadi A, Malekzadeh R, Derakhshan MH, Sepehr A, Nouraie M, Sotoudeh M, et al. Cancer occurrence in Ardabil: Results of a population-based cancer registry from Iran. *Int J Cancer* 2003; 107(1): 113-8.
13. Rahimi F & Heidari M. Time trend analysis of stomach cancer incidence in the west of Iran. *Journal of Health & Development* 2012; 1(2): 100-11[Article in Persian].
14. Hasanzadeh J, Hosseini Nezhad Z, Molavie Vardanjani H & Farahmand M. Gender differences in esophagus, stomach, colon and rectum cancers in Fars, Iran, during 2009-2010: An epidemiological population based study. *Journal of Rafsanjan University of Medical Sciences* 2013; 12(5): 333-42[Article in Persian].
15. Korosh E, Goya MM & Ramazani R. Cancer registries report 2009. Tehran: Ministry of Health and Medical Education; 2012: 154[Book in Persian].
16. Jafar S, Jalil A, Soheila N & Sirous S. Prevalence of helicobacter pylori infection in children, a population-based cross-sectional study in west Iran. *Iran Journal Pediatrics* 2013; 23(1): 13-8.
17. Mccoll KE & El-Omar E. How does H. pylori infection cause gastric cancer? *The Keio Journal of Medicine* 2002; 51(2): 53-6.

18. Palli D, Galli M, Caporaso NE, Cipriani F, Decarli A, Saieva C, et al. Family history and risk of stomach cancer in Italy. *Cancer Epidemiology Biomarkers Prev* 1994; 3(1): 15-8.
19. Inoue M, Tajima K, Yamamura Y, Hamajima N, Hirose K, Kodera Y, et al. Family history and subsite of gastric cancer: Data from a case-referent study in Japan. *International Journal Cancer* 1998; 76(6): 801-5.
20. Haenszel W, Kurihara M, Segi M & Lee RK. Stomach cancer among Japanese in Hawaii. *Journal of the National Cancer Institute* 1972; 49(4): 969-88.
21. Kamineni A, Williams MA, Schwartz SM, Cook LS & Weiss NS. The incidence of gastric carcinoma in Asian migrants to the United States and their descendants. *Cancer Causes Control* 1999; 10(1): 77-83.
22. Tsugane S & Sasazuki S. Diet and the risk of gastric cancer: Review of epidemiological evidence. *Gastric Cancer* 2007; 10(2): 75-83.
23. Rocco A & Nardone G. Diet H. pylori infection and gastric cancer: Evidence and controversies. *World Journal Gastroenterol* 2007; 13(21): 2901-12.
24. Vineis P, Alavanja M, Buffler P, Fontham E, Franceschi S, Gao YT, et al. Tobacco and cancer: Recent epidemiological evidence. *J Natl Cancer Inst* 2004; 96(2): 99-106.
25. Hand DJ. Data mining: Statistics and more? *The American Statistical Association* 2002; 52(2): 112-8.
26. Han J, Kamber M & Pei J. *Data mining: Concepts and techniques*. UK: Morgan Kaufmann; 2000: 150-2.
27. Navarro Silvera SA, Mayne ST, Gammon MD, Vaughan TL, Chow WH, Dubin JA, et al. Diet and lifestyle factors and risk of subtypes of esophageal and gastric cancers: Classification tree analysis. *Annals of Epidemiology* 2014; 24(1): 50-7.
28. Ramachandran P, Giniya N & Bhuvanewari T. Early detection and prevention of cancer using data mining techniques. *International Journal of Computer Applications* 2014; 97(13): 48-53.
29. Kirshners A, Polaka I & Aleksejeva L. Gastric cancer risk analysis in unhealthy habits data with classification algorithms. *Information Technology and Management Science* 2015; 18(1): 97-102.
30. Silvera SAN, Mayne ST, Gammon MD, Vaghuan TL, Chow WH, Dubin JA, et al. Diet and lifestyle factors and risk of subtypes of esophageal and gastric cancers: classification tree analysis. *Ann Epidemiol* 2014; 24(1): 50-7.
31. D'Elia L, Rossi G, Ippolito R, Cappuccio FP & Strazzullo P. Habitual salt intake and risk of gastric cancer: A meta-analysis of prospective studies. *Clinical Nutrition* 2012; 31(4): 489-98.
32. Kumar V, Abbas A & Fausto N. *Robbins basic pathology*. 9th ed. Phila Delphia: Saunders; 2013: 231-2.

Determining the Effective Factors in the Incidence of Gastric Cancer by Using Data Mining Approach

Mahmoodi Seyed Abbas¹ (M.S.) - Mirzaie Kamal² (Ph.D.) - Mahmoodi Seyed Mostafa³ (D.D.)

1 Master of Science in Computer Engineering, Computer Engineering Department, School of Engineering, Islamic Azad University, Yazd Branch, Yazd, Iran

2 Assistant Professor, Computer Engineering Department, School of Engineering, Islamic Azad University, Maybod Branch, Maybod, Iran

3 Assistant Professor, Oral Pathology Department, School of Dentistry, Shahid Sadoughi University of Medical Sciences, Yazd, Iran

Abstract

Received: Nov 2016

Accepted: Mar 2017

Background and Aim: Gastric cancer is the second leading cause of cancer death in the world. Due to the prevalence of the disease and the high mortality rate of gastric cancer in Iran, the factors affecting the development of this disease should be taken into account. In this research, two data mining techniques such as Apriori and ID3 algorithm were used in order to investigate the effective factors in gastric cancer.

Materials and Methods: Data sets in this study were collected among 490 patients including 220 patients with gastric cancer and 270 healthy samples referred to Imam Reza hospital in Tabriz. The best rules related to this data set were extracted through Apriori algorithm and implementing it in MATLAB. ID3 algorithm was also used to investigate these factors.

Results: The results showed that having a history of gastro esophageal reflux has the greatest impact on the incidence of this disease. Some rules extracted through Apriori algorithm can be a model to predict patient status and the incidence of the disease and investigate factors affecting the disease. The prediction accuracy achieved through ID3 algorithm is 85.56 which was a very good result in the prediction of gastric cancer.

Conclusion: Using data mining, especially in medical data, is very useful due to the large volume of data and unknown relationships between systemic, personal, and Behavioral Features of patients. The results of this study could help physicians to identify the contributing factors in incidence of the disease and predict the incidence of the disease.

Keywords: Data Mining, Association Rules, Apriori Algorithm, Gastric Cancer, ID3 Algorithm

* Corresponding Author:

Mahmoodi SA;

Email:

sa_mahmoodi_85@yahoo.com