مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره اول شماره دوم. ص۱۰۳-۹۵، ۱۳۹۳

استفاده از قوانین انجمنی جهت کشف عوامل خطر در بروز سرطان معده

سید عباس محمودی[۱]*، کمال میرزائی[۲]، سید مصطفی محمودی[۳]

• دریافت مقاله: ۹۳/۱۱/۱۸ • پذیرش مقاله: ۹۳/۱۲/۱۱

مقدمه: سرطان معده دومین علت مرگ ناشی از سرطان بعد از سرطان ریه در جهان است. بروز آن در مناطق مختلف دنیا متفاوت است. 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 روش‌های علمی‌تر، مورد بررسی قرار گیرد. هدف این مقاله، بررسی این عوامل با کمک تکنیک داده‌کاوی است.

روش: داده‌های مورد نیاز برای این مطالعه، از بیماران مراجعه‌کننده به بیمارستان امام رضا(ع) شهر تبریز جمع‌آوری شده است و پس از اعمال پیش پردازش بر روی این داده‌ها، در نهایت ۴۹۰ رکورد شامل ۲۲۰ نمونه مبتلا به سرطان و ۲۷۰ نمونه سالم در یک فایل Excel جمع‌آوری شد. با استفاده از پیاده‌سازی الگوریتم Apriori در نرم‌افزار Matlab و مجموعه داده‌های نهایی، بهترین قوانین حاکم بر روی این مجموعه داده، استخراج شده است.

نتایج: در این مطالعه برای نخستین بار از مجموعه داده‌های سرطان معده و ویژگی‌های تأثیرگذار در بروز این بیماری استفاده شده است. نتایج نشان داد، افراد مبتلا به بیماری قلبی 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 رفلاکس معده بیشترین تأثیر را در ایجاد این بیماری دارد. با استفاده از الگوریتم Apriori قوانینی به دست آمد که می‌تواند به عنوان الگویی برای پیش‌بینی وضعیت بیماران و احتمال بروز این بیماری، استفاده شود.

نتیجه‌گیری: امروزه به دلیل وجود حجم انبوهی از داده‌های پزشکی، می‌توان با استفاده از رویکرد داده‌کاوی به استخراج دانش از مجموعه داده‌های پزشکی پرداخت. در این مطالعه با استفاده از الگوریتم Apriori، قوانینی استخراج شده است که می‌تواند کمک فراوانی به پزشکان در بررسی عوامل ایجاد این بیماری بکند.

کلید واژه‌ها: قوانین انجمنی، الگوریتم Apriori، سرطان معده

• **ارجاع:** محمودی سید عباس، میرزائی کمال، محمودی سید مصطفی. استفاده از قوانین انجمنی جهت کشف عوامل خطر در بروز سرطان معده. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۳؛ ۱(۲): ۱۰۳-۹۵.

۱. دانشجوی کارشناسی ارشد مهندسی نرم‌افزار، گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی، پردیس علوم تحقیقات یزد، یزد، ایران.
۲. دکترای مهندسی نرم افزار، استادیار گروه مهندسی کامپیوتر، دانشکده فنی مهندسی، دانشگاه آزاد اسلامی، واحد میبد، میبد، ایران.
۳. دکترای تخصصی پاتولوژی دهان، استادیار گروه پاتولوژی دهان، دانشکده دندانپزشکی، دانشگاه علوم پزشکی بیرجند، بیرجند، ایران.
***نویسنده مسؤول:** یزد، دانشگاه آزاد اسلامی، واحد یزد، گروه مهندسی کامپیوتر
• **شماره تماس:** ۰۳۵۱-۸۱۱۷۷۱۳
• **Email:** sa_mahmoodi_85@yahoo.com

مقدمه

امروزه در دانش پزشکی میزان داده‌های مربوط به علائم بیماران مبتلا به بیماری‌های گوناگون و روش‌های کمکی برای تشخیص این بیماری‌ها، بسیار وسیع و گسترده شده است، به طوری که معمولاً تحلیل و در نظرگرفتن کلیه عوامل دخیل توسط یک فرد، دشوار به نظر می‌رسد. بنابراین به یک سیستم مکانیزه برای کمک به کشف قوانین، شناسایی الگوهای موجود و پیش‌بینی رخدادهای آینده، کاملاً احساس می‌شود. تکنیک-های داده‌کاوی (Data Mining) به عنوان ارائه‌کننده این سیستم مکانیزه، کمک بسیاری در پیشرفت پزشکی به ویژه در زمینه تشخیص بیماری‌های گوناگون و به دست آوردن روابط مفید میان عوامل موجود در داده‌ها، کرده است [۱].

داده‌کاوی و کشف دانش در پایگاه داده‌ها، رویکردی برای یافتن روابط و الگوهای پنهان داده‌ها است [۱]. این علم، فرآیند استخراج الگوها از مجموعه داده‌های بزرگ با ترکیب روش‌هایی از آمار، هوش مصنوعی و مدیریت پایگاه داده است [۲]. امروزه داده‌کاوی در بسیاری از زمینه‌ها مانند بازاریابی، تشخیص تقلب، کشف علم و پزشکی استفاده می‌شود. قوانین انجمنی (Association Rules) یکی از تکنیک‌های اصلی داده‌کاوی است و تقریباً مهم‌ترین شکل از کشف و استخراج الگوها در سیستم‌های یادگیری غیر هدایت شده می‌باشد. این روش تمام الگوهای جالب و تکرارپذیر در پایگاه داده‌ها را بازیابی می‌کند [۳].

برای داده‌کاوی پزشکی در زمینه سرطان معده، کارهای زیادی انجام نگرفته است. به طوری که بیشتر پژوهش‌های انجام شده بر روی مجموعه داده‌های سایر سرطان‌ها است که در ادامه به صورت مختصر بیان می‌شود. یکی از مطالعات انجام شده، در زمینه شناسایی عوامل پیشگیری ابتلا به سرطان سینه است. در این مطالعه، از الگوریتم Apriori، برای استخراج قوانین انجمنی از مجموعه داده‌های پیشگیری از سرطان سینه، استفاده شده است. مطابق نتایج به دست آمده از این تحقیق، فاکتورهایی نظیر فعالیت بدنی، نداشتن سابقه خانوادگی، اجتناب از درمان با جایگزینی هورمون، چک کردن سینه‌ها به صورت ماهانه و استفاده از رژیم غذایی کم چرب، به عنوان عوامل پیشگیری از سرطان سینه شناسایی شدند [۴].

مطالعه دیگری نیز روی مجموعه داده‌های SEER (Surveillance Epidemiology and End Results) برای استخراج قوانین انجمنی با استفاده از الگوریتم Apriori انجام شده است. مجموعه داده‌های این مطالعه، وضعیت افراد سرطانی را پس از بهبود پنج ساله مشخص می‌کند. در حقیقت این مجموعه داده نشان می‌دهد که پس از ۵ سال، بهبودی حاصل شده یا این که رخداد مجدد بیماری اتفاق افتاده است [۵].

در تحقیق دیگری در زمینه کشف روابط پنهان از مجموعه داده‌های سرطان دهان، به استخراج الگوهای پنهان و تشخیص سرطان دهان پرداخته شده است. در این تحقیق نیز از الگوریتم Apriori جهت کشف روابط پنهان بین ویژگی‌های موجود در پایگاه داده استفاده شد. پایگاه داده مورد استفاده در این تحقیق شامل ۳۳ ویژگی و ۱۰۲۰ رکورد می‌باشد که از پرونده بیماران استخراج شده است. نتایج حاصل از تجزیه و تحلیل قوانین استخراجی در این تحقیق نشان می دهد که اگر فرد علائم بالینی به صورت زخم، سابقه مصرف مواد مخدر (توتون، تنباکو، سیگار و یا الکل) و فشارخون خون بالا داشته باشد، آنگاه مشکوک به سرطان است و باید از طریق نمونه‌برداری و دیگر روش‌های تشخیصی تأیید شود [۶].

از جمله تحقیقات مهمی که در دو سال اخیر در کشف روابط و الگوهای پنهان از داده‌های پزشکی در ایران انجام شده می‌توان به استفاده از داده‌کاوی برای کشف الگوهای پنهان در داده‌های سرطان سینه اشاره کرد. در این پژوهش مهم نیز از الگوریتم Apriori استفاده کرده است [۷].

سرطان معده دومین علت مرگ ناشی از سرطان بعد از سرطان ریه در جهان است. بروز آن در مناطق مختلف دنیا متفاوت است؛ اما به طور کلی یک مشکل اساسی در کشورهای در حال توسعه محسوب می‌شود [۸]. 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 [۹]. 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 روش‌های علمی‌تر، مورد بررسی قرار گیرد. هدف این مقاله، بررسی این عوامل با کمک تکنیک داده‌کاوی است که تاکنون در هیچ پژوهشی انجام نشده است.

روش

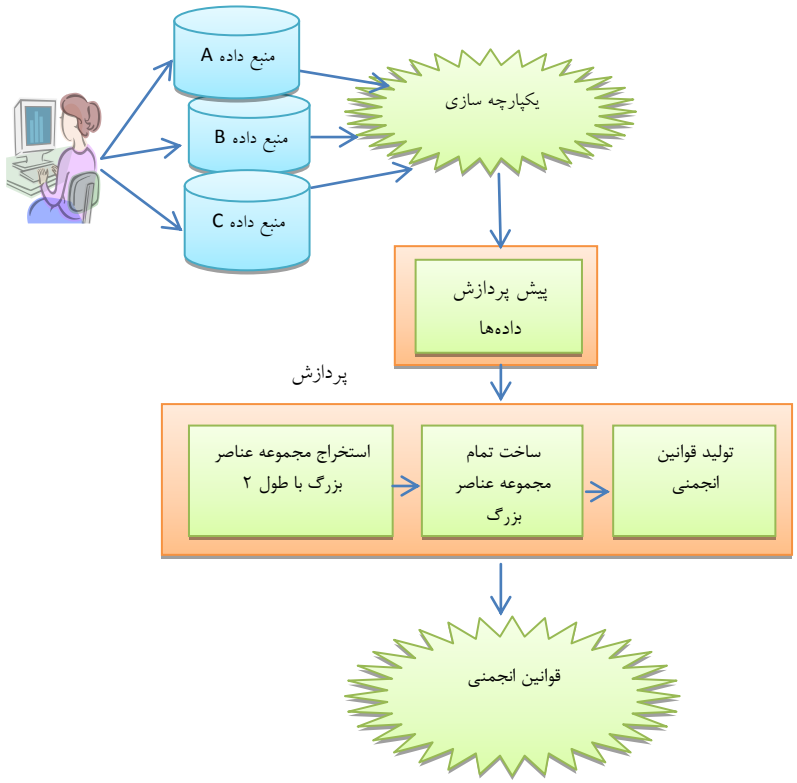
مجموعه داده مورد استفاده: مجموعه داده‌های مورد استفاده در این مطالعه، داده‌های جمع‌آوری شده بیماران مراجعه کننده به بیمارستان امام رضا (ع) شهر تبریز در سال‌های ۹۰-۱۳۸۶ است. پایگاه داده اولیه مورد نظر با فرمت Excel شامل ۵۶۰

رکورد می‌باشد که این اطلاعات از طریق پرسشنامه جمع‌آوری شده است. در ادامه عملیات پیش پردازش و پاکسازی بر روی داده‌ها صورت گرفت. مرحله پیش پردازش به منظور بهبود داده‌ها انجام می‌شود [۱۰] که شامل انجام فرآیندهایی از قبیل تصحیح و یا حذف داده‌های بدون مقدار، تعیین محدوده مجاز و تصحیح مقادیر غیرمجاز، انجام محاسبات مجدد برای برخی از ویژگی‌ها و تبدیل آن‌ها به ویژگی‌های دیگر، گسسته‌سازی ویژگی‌های پیوسته و در نهایت تبدیل مقادیر ویژگی است [۱۱,۱۲]. در این مرحله تمام ویژگی‌ها به ویژگی‌های دودویی تبدیل شدند. به این صورت که هر ویژگی اسمی یا ترتیبی به چندین ویژگی دودویی شکسته شد. در واقع بعد از گسسته‌سازی، ویژگی‌های پیوسته به چندین ویژگی دسته‌ای، شکسته می‌شوند. سپس هر ویژگی چندین مقادیر دسته‌ای و هر مقدار دسته‌ای باید یک ویژگی دودویی داشته باشد. سرانجام یک پایگاه داده جدید ایجاد شد. که تمام ویژگی‌ها، به ویژگی‌های دودویی تبدیل شدند. به عنوان مثال در مجموعه داده‌های سرطان، ۳ نوع ویژگی"سن، گروه خونی و مصرف سیگار" که به ترتیب ویژگی‌های پیوسته، دسته‌ای و دودویی هستند. ویژگی سن به ۳ دسته "زیر ۴۰ سال، ۴۱ تا ۶۰ سال و ۶۱ به بالا" گسسته‌سازی شده است. ویژگی دسته‌ای گروه خونی به ۴ دسته"A، B، AB و O" و ویژگی دودویی مصرف سیگار به دو مقدار "بله، خیر"دسته‌بندی 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 است و مفهوم مشخصی دارد. مجموعه داده نهایی در این تحقیق بعد از پیش‌پردازش شامل ۲۲۰ نمونه بیمار مبتلا به سرطان و ۲۷۰ نمونه سالم است. جدول ۱ ویژگی‌های مجموعه داده نهایی پس از پیش‌پردازش نهایی را نشان می‌دهد.

جدول ۱: مجموعه داده نهایی پس از پیش‌پردازش نهایی

ردیف	ویژگی	مقادیر	ردیف	ویژگی	مقادیر
۱	بیمار بودن	بله، خیر	۱۶	وضعیت مخاط	نرمال، ورم، سرخ و قرمز، زخم
۲	محل سرطان	کاردیا، غیرکاردیا	۱۷	وضعیت عمومی سرطان	خوب، متوسط، ضعیف
۳	جنس	مرد، زن	۱۸	BMI	کمتر از ۱۸,۵، ۱۸,۵-۲۴,۹، ۲۵-۲۹,۵، بیشتر از ۳۰
۴	گروه خونی	A,B,AB,O	۱۹	میزان تحرک	سبک، متوسط، زیاد
۵	در معرض مواد شیمیایی	بله، خیر	۲۰	میزان مصرف نمک	نمی‌خورد، زیاد، کم
۶	مصرف سیگار	بله،خیر	۲۱	میزان مصرف سبزیجات	روزانه، بین ۱تا ۳ بار در هفته،۱تا ۳ بار درماه
۷	مصرف الکل	بله، خیر	۲۲	میزان مصرف مواد غذایی دودی شده	نمی‌خورد، بین ۱ تا ۳ بار درهفته، بین ۱ تا ۳ بار در ماه
۸	سابقه سرطان در فامیل	بله، خیر	۲۳	میزان مصرف میوه	روزانه، بین ۱ تا ۳ بار در هفته، بین ۱ تا ۳ بار در ماه
۹	سابقه سرطان معده در فامیل	بله، خیر	۲۴	میزان مصرف فست فود	نمی‌خورد، ۱ تا ۳ بار در هفته، ۱ تا ۳ بار در ماه
۱۰	سابقه بیماری قلبی عروقی	بله، خیر	۲۵	سن	زیر ۴۰ سال، بین ۴۱ تا ۶۰ سال، ۶۱ به بالا
۱۱	سابقه آلرژی	بله، خیر	۲۶	مصرف شیر	بله، خیر
۱۲	سابقه عفونت معده	بله، خیر	۲۷	ظروف پخت غذا	آلومینیم، تفلن، مس، لعابی
۱۳	سابقه التهاب معده	بله، خیر	۲۸	ظروف ذخیره سازی غذا	ملامین، آلومینیم، پلاستیک، چینی، استیل، مس
۱۴	سابقه رفلاکس معده	بله، خیر	۲۹	میزان مصرف غذای های سرخ شده	روزانه، بین ۱ تا ۳ بار در هفته، بین ۱ تا ۳ بار در ماه، نمی‌خورد

کشف قوانین انجمنی: قوانین انجمنی، یکی از تکنیک‌های اصلی داده‌کاوی است و تقریباً مهمترین شکل کشف و استخراج الگوها در سیستم‌های یادگیری است [۱۳]. قوانین انجمنی ارتباطات جذاب بین مجموعه بزرگی از داده‌ها را کشف می‌کنند که این ارتباطات می‌تواند به تصمیم‌گیرندگان کمک‌کند [۳]. در واقع قوانین انجمنی شرایطی را نشان می‌دهند که در یک مجموعه‌ی داده، مکرراً با هم اتفاق می‌افتند. قوانین استخراج شده در حقیقت حضور برخی ویژگی‌ها را براساس سایر ویژگی‌ها شرح می‌دهند [۱۴]. چارچوب کلی فرآیند کشف قوانین انجمنی در این مطالعه در شکل (۱) نشان داده است.



شکل ۱ : جریان کاری کلی فرآیند کشف قوانین انجمنی بر روی مجموعه داده‌های سرطان معده

در این پژوهش از الگوریتم Apriori برای استخراج قوانین انجمنی استفاده شده است. Apriori یکی از مهمترین یافته‌ها در تاریخ استخراج قوانین انجمنی است که تاکنون معرفی شده است [۹]. قبل از معرفی الگوریتم مربوط به کاوش قوانین انجمنی، نیاز به معرفی یک سری مفاهم پایه است.

۱- مجموعه آیتم‌های موجود در یک پایگاه اطلاعاتی با Itemset=$\{X_1,X_2,\dots\}$ نمایش داده می‌شوند.

۲- برای هر قانون که به شکل $X\rightarrow Y$ است، دو مقدار Support و Confidence مشخص می‌شود [۲].

۲-۱- Support(S) احتمال وجود همزمان X و Y به صورت توأم در تراکنش است.

۲-۲- Confidence(C) احتمال شرطی است برای آن که تراکنش دارای X، دارای Y نیز باشد.

بنابراین قانون $X\rightarrow Y$ با (S=50%,C=66.7%) بدین معنی است که X و Y به صورت توأم در ۵۰ درصد از کل تراکنش‌ها وجود دارند و در ۶۶/۷ درصد از تراکنش‌ها، هرجا X در تراکنش حضور داشته، Y نیز حضور داشته است.

فرآیند کاری الگوریتم بدین صورت است که ابتدا تمام مجموعه عناصر تک عضوی مکرر را پیدا می‌کند، سپس براساس آن، مجموعه عناصر دو عضوی مکرر پیدا می‌شوند، در ادامه براساس این مجموعه عناصر دو عضوی، مجموعه عناصر سه عضوی ساخته می‌شوند و این فرآیند به همین ترتیب ادامه می‌یابد تا هیچ مجموعه عنصر مکرر بزرگ‌تری پیدا نشود. این الگوریتم در دو مرحله الحاق و هرس، کار می‌کند [۴]. در ادامه این دو مرحله توضیح داده می‌شوند.

مرحله الحاق: ابتدا باید مطمئن شویم که عناصر برمبنای ترتیب حروف الفبا مرتب شده باشند. L_{K-1} را با l_i و اعضای آن‌ها را به صورت $l_i[j]$ نشان داده می‌شود. در اینجا i بیانگر شماره مجموعه و j بیانگر شماره عنصر در مجموعه می‌باشد. اگر K-2 عنصر اول دو مجموعه با یکدیگر برابر باشند، آنگاه دو مجموعه L_{K-1} با یکدیگر قابل پیوست هستند، یعنی:

$$(l1[1] = l2[1]) \;\&\; (l1[2] = l2[2]) \;\&\; \dots\; (l1[k-2] = l2[k-2]) \;\&\; (l1[k-1] < l2[k-1])$$

توجه شود که دو عنصر آخر به ترتیب مرتب شده‌اند و از وجود عناصر تکراری جلوگیری می‌کنند. اجتماع دو مجموعه قابل

پیوست، ترکیب را به وجود می‌آورد (البته مرتب از نظر الفبایی). با این روش ترکیب مجموعه حاصل K عضو آخر از نظر ترتیبی از مجموعه دوم خواهد بود.

P in L_{K-1} = (1 2 3)
Q in L_{K-1} = (1 2 4)
Join: Result in C_K = (1 2 3 4)

مرحله هرس: C_K (مجموعه کاندید) مجموعه‌ای ازL_K ها است که هر عنصر آن یا مکرر است یا نیست؛ اما تمام عناصر مکرر در آن قرار دارند. حاصل تمام عناصر این مجموعه باید بررسی شوند تا مشخص شود که آیا مکرر هستند یا خیر؟ اما چون ممکن است تعداد آن‌ها زیاد باشد، لذا برای کاهش حجم محاسبات، از اصل Apriori استفاده می‌شود. به این صورت که اگر یکی از زیر مجموعه‌ها مکرر نباشد، آن مجموعه نیز مکرر نخواهد بود. حال برای پیدا کردن مجموعه‌های مکرر، کافی ست مجموعه‌های غیر مکرر را از آن‌ها جدا کنیم، به این صورت که اگر عضوی از C_K دارای زیر مجموعه‌های K-1 عضوی باشد که در L_{K-1} عضو نباشد، آن عضو C_K مکرر نخواهد بود. این الگوریتم C_Kها را با اتصال اقلام پرتکرار بزرگ حاصل از فاز قبلی و حذف آن‌هایی که در فاز قبلی بوده‌اند، به طور مجزا تولید می‌کند. بدین ترتیب تعداد C_K های اضافی به طور چشمگیری کاهش می‌یابند. C_Kها در این الگوریتم مطابق شبه کد شکل (۲) محاسبه می‌گردد.

Apriori-gen (Lk-1)

11. Join step
12. insert into C_k
13. select p.item1, p.item2 ,… , $p.item_{k-1}$, $q.item_{k-1}$
14. from L_{k-1} p, L_{k-1} q
15. where p.item1=q.item1, …,$p.item_{k-2}$ = $q.item_{k-2}$, $p.item_{k-1}$< $q.item_{k-1}$
16. Prune step
17. forall itemsets c $C_k$do
18. forall (k-1)-subsets s of c do
19. if (s Lk-1) then
20. delete c from C_K

شکل ۲: شبه کدتولید مجموعه کاندید(C_k)

پس از محاسبه C_k ها، میزان پشتیبان هر یک از اعضای آن‌ها را محاسبه می‌کنیم و فقط آن‌هایی را که از حداقل میزان پشتیبانی برخوردارند را در L_k قرار می‌دهیم. شکل ۳ شبه کد الگوریتم را به صورت کلی نشان می‌دهد.

The Apriori Algorithm

1. L1 = {large 1–itemsets}
2. for (k = 2; $L_{k-1} \neq \emptyset$; k++) do begin
3. Ck = apriori – gen (L_{k-1});
4. forall transactions t $\in$D do begin
5. Ct = subset (C_k ,t)
6. for all candidates c $\in$Ct do
7. c.count++;
8. end
9. end
10. Lk = {c $\in C_k$ | c.count $\geq$ minsup}
11. end
12. Answer = U_k L_k

شکل۳: شبه کد الگوریتم Apriori

پس از آنکه مجموعه عناصر بزرگ با تکرار قابل قبول استخراج شد، نوبت به استخراج قوانین انجمنی می‌رسد. برای هرمجموعه مکرر L، تمام زیر مجموعه‌های غیر تهی نوشته می‌شود. برای هر زیر مجموعه حاصل s نیز قوانین زیر نوشته می‌شود.

"s $\rightarrow$ (L - s) "

سپس اطمینان از رابطه (۱) حساب می‌شود و اگر بیشتر از حد قابل قبول بود، پذیرفته می‌گردد.

$$confidence(A \rightarrow B) = P(B|A) = \frac{Support\_count(A \cup B)}{Support\_Count(A)} \quad (1)$$

نتایج

این تحقیق از جنبه‌های مختلفی حائز اهمیت است. نخست این که این تحقیق یکی از مسائل واقعی را پوشش می‌دهد. به این معنی که داده‌های مورد بررسی واقعی بوده، اهداف و مطالب متناسب با یک مسأله واقعی تدوین شده و نتایج آن نیز مورد تأیید کارشناسان این حوزه قرار گرفته است. در این تحقیق برای رسیدن به یک نتیجه قابل قبول، یک فرآیند کشف دانش از داده‌های واقعی طراحی و اجرا شد. این فرآیند شامل پیش پردازش داده‌ها، کشف الگوهای مکرر و تفسیر قوانین به دست آمده بود. هر یک از این مراحل مستلزم صرف وقت و دقت بسیار است. همان‌طور که بیان‌شد، الگوریتم Apriori یکی از مهم‌ترین الگوریتم‌های داده‌کاوی در حوزه 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 می‌شود. 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.

جدول ۳ : بهترین قوانین به دست آمده از الگوریتم Apriori

ردیف	قانون	نتیجه	درجه پشتیبان
۱	اگر محل سرطان کاردیا، سابقه رفلاکس معده دارد و سابقه عفونت معده ندارد	بیمار بودن	۰/۲۰۶۱۲
۲	اگر محل سرطان کاردیا، سابقه بیماری قلبی عروقی ندارد و سابقه سرطان در فامیل ندارد	بیمار بودن	۰/۲۱۰۲۰
۳	اگر محل سرطان غیر کاردیا، سابقه رفلاکس معده دارد و سابقه عفونت معده ندارد	بیمار بودن	۰/۱۷۷۵۵
۴	اگر محل سرطان غیر کاردیا، سابقه بیماری قلبی عروقی _ندارد، سابقه سرطان در فامیل ندارد	بیمار بودن	۰/۱۸۷۷۵
۵	اگر محل سرطان کاردیا، سابقه بیماری قلبی عروقی ندارد، سابقه سرطان در فامیل ندارد و سابقه سرطان معده در فامیل ندارد	بیمار بودن	۰/۱۷۷۵۵
۶	اگر محل سرطان غیر کاردیا، سابقه بیماری قلبی عروقی ندارد، سابقه سرطان در فامیل ندارد وسابقه سرطان معده در فامیل ندارد	بیمار بودن	۰/۱۸۷۷۵
۷	اگر محل سرطان کاردیا و سابقه رفلاکس معده دارد	بیمار بودن	۰/۲۲۲۴
۸	اگر محل سرطان کاردیا و سابقه بیماری قلبی عروقی ندارد	بیمار بودن	۰/۲۰۲۰
۹	اگر محل سرطان کاردیا و سابقه عفونت معده ندارد	بیمار بودن	۰/۲۰۶۱۲
۱۰	اگر محل سرطان غیرکاردیا و سابقه رفلاکس معده دارد	بیمار بودن	۰/۲۲۴۴
۱۱	اگر محل سرطان غیرکاردیا و سابقه بیماری‌های قلبی عروقی ندارد	بیمار بودن	۰/۲۱۴۲
۱۲	اگر محل سرطان غیرکاردیا و سابقه عفونت معده ندارد	بیمار بودن	۰/۲۱۲۲۴
۱۳	اگر مصرف نمک ندارد	سابقه رفلاکس معده دارد	۰/۱۷۷۵۵
۱۴	اگر میزان مصرف نمک زیاد دارد	سابقه رفلاکس معده دارد	۰/۲۰۶۱۲
۱۵	اگر مصرف شیر ندارد	سابقه رفلاکس معده دارد	۰/۲۲۴۴
۱۶	اگر جنس مرد، وضعیت مخاط نرمال، در معرض مواد شیمیایی نیست، سابقه التهاب معده ندارد، مصرف الکل ندارد و ۲۵<BMI>۲۹/۵	سابقه رفلاکس معده دارد	۰/۱۷۳۴۶
۱۷	اگر وضعیت مخاط نرمال، سابقه عفونت معده ندارد، مصرف الکل ندارد، میزان تحرک متوسط و سن۶۱ به بالا	سابقه رفلاکس معده دارد	۰/۱۸۳۶۷

بحث و نتیجه‌گیری

در این پژوهش با استفاده از تکنیک‌های داده‌کاوی، به بررسی عوامل تأثیرگذار در ایجاد بیماری سرطان معده پرداختیم. سرطان معده از جمله شایع‌ترین سرطان‌ها محسوب می‌شود، به طوری که در کشور ما در رأس سرطان‌های شایع قرار گرفته است [۱۵]. با توجه به این موضوع، بر آن شدیم تا پژوهشی پیرامون بررسی عوامل مؤثر در ایجاد این بیماری انجام دهیم.

سرطان از جمله بیماری‌هایی هست که عوامل بسیار زیادی در ابتلا به آن مؤثر می‌باشد. این ویژگی بیانگر مؤثر بودن استفاده از تکنیک‌های داده‌کاوی است [۱۶]. تکنیک داده‌کاوی مورد استفاده در این پژوهش، الگوریتم Apriori است که یکی از مهمترین الگوریتم‌های داده کاوی در حوزه کشف قوانین انجمنی می‌باشد. قوانین به دست آمده نشان می‌دهد که داشتن سابقه رفلاکس معده بیشترین تأثیر را در بروز این بیماری دارد که این نتیجه با اکثر مطالعات در این زمینه مطابقت دارد [۱۷]. در تحقیقات اخیر [۱۸] نیز نشان داده شده است که بیماران قلبی عروقی به علت مصرف یک سری داروها، در معرض خطر کمتری برای ابتلا به سرطان معده هستند. بنا بر این افراد مبتلا به بیماری قلبی و عروقی کمتر در معرض خطر ابتلا به سرطان معده هستند. همچنین رفلاکس معده با مصرف نکردن نمک، مصرف زیاد نمک و مصرف نکردن شیر نیز ارتباط دارد. از نکات جالب این پژوهش این است که علاوه بر نتایج به دست آمده که بیان شد، با استفاده از قوانین ایجاد شده می‌توان برای یک نمونه جدید با ویژگی‌های مشخص پیش‌بینی کرد که این فرد احتمالاً در آینده دچار این بیماری خواهد شد و با کنترل عوامل تأثیرگذار در بروز این بیماری، می‌توان امیدوار بود که از بروز این بیماری تا حدی اجتناب کرد.

امروزه با استفاده از داده‌کاوی می‌توان کمک فراوانی به پزشکان در بررسی علل و عوامل تأثیر گذار در بروز بیماری‌ها و همچنین روش‌های پیشگیری از بیماری‌ها کرد. در زمینه استفاده از داده-های سرطان معده در زمینه کشف قوانین انجمنی، تاکنون هیچ پژوهش مستندی انجام نشده است. در این پژوهش، با استفاده از الگوریتم Apriori، بهترین قوانین حاکم بر مجموعه داده‌های بیماران مراجعه کننده به بیمارستان امام رضا (ع) شهر تبریز و عوامل تأثیرگذار در ایجاد این بیماری کشف و بررسی شد.

References

1. Chou SM, Lee TS, Shao YE, Fei Chen IF. Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines. Expert Systems with Applications. 2004;27(1):133-42.
2. Tan P, Steinbach M, Kumar V. Introduction to data mining. 2th ed. Boston: Addison-Wesley; 2006.
3. Ramageri M. Data mining techniques and applications. Indian Journal of Computer Science and Engineering. 2011;4(1): 301-5.
4. Nahar J, Tickle KS, Shawkat Ali A B, Chen YP. Significant cancer prevention factor extraction: an association rule discovery approach. J Med Syst. 2009:35(3): 353-67.
5. Fan Q, Zhu C, Xiao J, Wang B, Yin L, Xu X, et al. An application of apriori algorithm in SEER breast cancer data. International Conference on Artificial Intelligence and Computational Intelligence (AICI); 23 - 24 Oct 2010; Sanya, China. 2010. p. 114-116.
6. Sharma N, Om H. Extracting Significant patterns for oral cancer detection using apriori algorithm. Intelligent Information Management. 2014; 6(2): 30-37.
7. Tabatabyi F, Minayi B. Using data mining to discover hidden patterns in breast cancer data [Dissertation]. Qom: University of Qom; 2012. Persian.
8. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin. 2011;61(2):69-90.
9. Etemad K, Goya M, Ramazami R, Modiran M, Partoipoor E, Salavati F, et al. Cancer Registration report in 2009. Ministry of Health and Medical Education; 2012. p. 120-45. Persian.
10. Ghazanfari M, Alizadeh S, Teymourpour B. editors. Data mining and knowledge discovery. 2th ed. Tehran: Publication of University of Science and Technology; 2011. Persian
11. Dzeroski S, Hristovski D, Peterlin B. Using data mining and OLAP to discover patterns in a database of patients with Y-chromosome deletions. Proc AMIA Symp. 2000:215-9.
12. Hoseini M. Developing a predictive model based on the Sarem hospital infertility data. [Dissertation]. Tehran: K.N.Toosi University of Technology; 2012. Persian.
13. Han J, Kamber M. Data Mining: concepts and techniques. 3th ed.UK: Morgan Kaufmann; 2006.
14. Hu R. Medical data mining based on association rules. Computer and Information Science. 2010; 3(4): 104-8.
15. Keyhanian S, Farhadifar N, Fotoukian Z, Pouya M, Saravi M. Epidemiologic and malignancy indices of gastric cancer in patients referred to oncology clinic at ramsar emam sajjad hospital during 2002-2009. J Shahid Sadoughi Univ Med Sci. 2012; 20(1):110-18. Persian.

16. Kiyani B, Atashi AR. A prognostic model based on data mining Techniques to predict breast cancer Recurrence. Journal of Health and Biomedical Informatics. 2014;1(1):26-31. Persian.
17. Malekzadeh R, Derakhshan MH, Malekzadeh Z. Gastric cancer in Iran: epidemiology and risk factors. Arch Iran Med. 2009;12(6):576-83.
18. Kumar V, Abbas Ak, Aster J. Robbins basic pathology. 9th ed. Philadelphia:Saunders; 2012.
19. Bashshur RL, Shannon GW, Krupinski EA, Grigsby J, Kvedar JC, Weinstein RS, et al. National telemedicine initiatives: essential to healthcare reform. Telemed J E Health. 2009;15(6):600-10.
20. Cwiek MA, Rafiq A, Qamar A, Tobey C, Merrell RC. Telemedicine licensure in the United States: the need for a cooperative regional approach. Telemed J E Health. 2007;13(2):141-7.

Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2015; 1(2): 95-103

Using Association Rules for the Detection of Risk Factors in Gastric Cancer

Seyed Abbas Mahmoodi[1*], Kamal Mirzaee[2], Seyed Mostafa Mahmoodi[3]

• **Received:** 7 Feb, 2015 • **Accepted:** 2 Mar, 2015

Introduction: Gastric cancer is the second cause of death from cancer after lung cancer in the world It`s incidence is varied in different regions of the words. Due to the prevalence rate of the disease and high mortality rates for gastric cancer in the country, it is necessary to examine the influential factors in the incidence of this disease by more accurately and scientific methods. The purpose of this study is to examine this factor with data mining techniques.

Method: The required data for this study was collected from patients referring to Imam Reza Hospital, Tabriz. After applying data pre-processing, totally 490 records were collected in an Excel file samples including 220 cancer cases and 270 normal specimens. The best rules based on the datasets were extracted by using Apriori algorithm implemented in MATLAB software and final data set.

Results: In this study, gastric cancer datasets and features affecting the incidence of this disease have been used for the first time. The results showed that risk for gastric cancer in people with cardiovascular disease are less. In addition, gastric reflux is associated with not using salt and milk, and high salt intake. Gastric reflux has also the most influence on creating this disease. Some rules were obtained by using Apriori algorithm that can be used as a model to predict the status of patients and the incidence of this disease.

Conclusion: Nowadays Due to massive amounts of medical data, knowledge can be extracted from datasets by using data mining approach. In this study, some rules were extracted by using Apriori algorithm that can provide physicians with great help to examine the causes of this disease.

Key words: Association rules, Apriori algorithms, Gastric cancer

•**Citation:** Mahmoodi SA, Kamal Mirzaee K, Mahmoodi SM.Using Association Rules for the Detection of Risk Factors in Gastric Cancer. Journal of Health and Biomedical Informatics 2015; 1(2): 95-103.

1. M.Sc. in Software Engineering, Computer Engineering Dept., Islamic Azad University, Yazd Science and Research Branch ,Yazd, Iran.
2. Ph.D. in Software Engineering, Assistant Professor of Computer Engineering Dept., School of Engineering Technical, Islamic Azad University, Maybod Branch, Maybod, Iran.
3. Ph.D. in Oral Pathology, Assistant Professor of Oral Pathology Dept., School of Dentistry, Birjand University of Medical Sciences, Birjand, Iran.

***Correspondence:** Computer Engineering Dept., Islamic Azad University, Yazd Branch, Yazd, Iran.
•**Tel:** 0351-8117713 •**Email:** sa_mahmoodi_85@yahoo.com