مجله تخصصی اپیدمیولوژی ایران؛ تابستان ۱۳۹۷، دوره ۱۴، شماره ۲: صفحات ۱۷۶-۱۶۶.
مقاله پژوهشی

بررسی رابطه علت و معلولی متغیرهای مرتبط با سرطان پستان با استفاده از شبکه‌های بیزی

ساغر حیدری[۱]، امیر کاوسی[۲]، وحید رضائی تبار[۳]

[۱] دانشجوی کارشناسی ارشد آمار زیستی، دانشکده پیراپزشکی، دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران

[۲] دانشیار، گروه علوم پایه، دانشکده سلامت، ایمنی و محیط زیست، دانشگاه علوم پزشکی شهید بهشتی، نهران، ایران

[۳] استادیار، گروه آمار، دانشکده علوم ریاضی، دانشگاه علامه طباطبایی، تهران، ایران

نویسنده رابط: امیر کاوسی، گروه علوم پایه، دانشکده سلامت، ایمنی و محیط زیست، دانشگاه علوم پزشکی شهید بهشتی، تهران، حکیمیه تهرانپارس، جنب پارک ساحل، تلفن تماس: ۰۹۱۲۲۴۰۴۵۱۴، پست الکترونیک: kavousi @sbmu.ac.ir

تاریخ دریافت: ۱۳۹۶/۰۶/۴؛ پذیرش: ۹۶/۱۲/۱۲

مقدمه و اهداف: سرطان پستان شایع‌ترین سرطان در ایران است. سرطان پستان قابل درمان و پیشگیری بوده و اساس کار هم تشخیص سریع این بیماری است. برای این منظور لازم است تا عوامل مؤثر بر سرطان و روابط علیتی بین آن‌ها مورد بررسی قرار گیرد. شبکه‌های بیزی نوعی از ابزار داده‌کاوی هستند که می‌توانند روابط علی بین متغیرهای مؤثر در تشکیل سلول‌های سرطانی را با ترسیم گراف به خوبی نمایش دهند. در این مقاله از یادگیری ساختاری شبکه بیزی با استفاده از الگوریتم ژنتیک برای یافتن روابط علیتی بین متغیرهای سرطان پستان استفاده شده است.

روش کار: این مطالعه از نوع کاربردی است و داده‌های آن مربوط به ۹۰۰ بیمار مبتلا به سرطان سینه در استان کرمان برای سال‌های ۸۶-۱۳۷۸ است. برای تجزیه و تحلیل داده‌ها از یک گراف که نشان دهنده روابط علت و معلولی است، استفاده شده است.

یافته‌ها: جراحی اصلی‌ترین روش درمانی سرطان پستان است و با توجه به احتمال‌های شرطی برآورد شده در شبکه‌ی بیزی، اغلب زنانی که تحت عمل جراحی قرار می‌گیرند، با امید بیش‌تری می‌توانند به زندگی خود ادامه دهند. هم‌چنین ۸۱ درصد از بیمارانی که تحت عمل جراحی قرار نگرفته‌اند و درمان آن‌ها صرفاً به شیوه شیمی‌درمانی یا رادیوتراپی صورت گرفته است، احتمال ادامه زندگی آن‌ها کم‌تر بوده است.

نتیجه‌گیری: افراد در رده سنی ۶۵-۴۰ سال، بیش‌تر درگیر سرطان هستند. هم‌چنین می‌توان نتیجه گرفت که متغیرهای سن، جراحی، **شیمی**درمانی و رادیوتراپی اثر مستقیم روی وضعیت بیمار دارند و یال از سمت آن‌ها به سمت متغیر وضعیت بیمار است.

واژگان کلیدی: سرطان پستان، شبکه بیزی، الگوریتم ژنتیک

مقدمه

کشنده‌ترین سرطان در ایران سرطان معده است که بیش‌ترین فراوانی آن نیز در مردان دیده می‌شود (۱). بعد از آن سرطان پستان در زنان شیوع زیادی دارد. سرطان پستان با تغییرات سلولی ایجاد می‌شود که تحت تأثیر دو تئوری است: تئوری اول مربوط به سلول بنیادی است. این سلول که از بدو تولد با نوزاد به دنیا می‌آید و در سینه جای گرفته است، در یک دوره‌ای تحت تأثیر عوامل محرک محیطی تبدیل به سلول سرطانی می‌شود که این دوره می‌تواند از سنین کم تا بالا باشد. در این تئوری می‌توان تصور کرد که اگر همه زنان عالم برای همیشه زنده بمانند بالاخره روزی به این سرطان مبتلا می‌شوند. تئوری دوم، تئوری جهش ژنی است که در حقیقت همان سلول‌های طبیعی هر عضوی از جمله سینه، تحت تأثیر عوامل محیطی خود تغییر ماهیت داده و سرطانی می‌شوند. به هر حال هر کدام از این دو تئوری سرطان سینه را یک بیماری مرتبط با ژن می‌شناسد که با اختلال‌هایی که در ژن سلولی ایجاد می‌شود سرطانی می‌شود، در حالی که این بیماری اصلاً ژنتیک نیست و حداکثر تا ۷ درصد موارد سرطان سینه بیماری ژنتیک تلقی می‌شود به این مفهوم که والدین دارای ژن اختصاصی سرطان هستند و آن را منتقل می‌کنند.

سرطان پستان وقتی رخ می‌دهد که سلول‌ها در سینه شروع به رشد خارج از کنترل نموده و می‌توانند در نزدیکی بافت‌ها گسترش یافته و در سراسر بدن پخش شوند. مجموعه‌های بزرگی از این بافت خارج از کنترل، تومور نامیده می شود. برخی از تومورها در واقع سرطان نیستند زیرا آن‌ها نمی‌توانند منتشر و پخش شوند و یا زندگی سلول‌های دیگر را تهدید نمایند این تومورها، تومورهای خوش‌خیم نامیده می‌شوند. تومورهایی که می‌توانند در سراسر بدن یا نزدیک بافت‌ها منتشر و پخش شوند؛ به‌عنوان سرطان در

نظر گرفته و تومورهای بدخیم نامیده می‌شوند. بیش‌تر اوقات تومورها را در پستان موقعی تشخیص می‌دهند که قبلاً بیماری ایجاد شده است (۲).

با توجه به اینکه مدل‌های گرافیکی، روابط علیتی بین متغیرها را به خوبی در قالب یک گراف نشان می‌دهند، در این پژوهش از مدل گرافیکی برای بررسی روابط بین متغیرهای مربوط به سرطان پستان استفاده می‌شود. برای این منظور از یک مدل گرافیکی احتمالی به نام شبکه بیزی روی متغیرهایی مانند سن، سطح‌بندی تومورها (گرید)، ریخت‌شناسی غده و روش‌های مختلف درمان مانند (جراحی، شیمی درمانی، رادیوتراپی) استفاده کرده و یک گراف به‌دست می‌آید. در نهایت با استفاده از این گراف، روابط علت و معلولی بین متغیرها و هم‌چنین وضعیت بیمار مشخص شده است.

یک شبکه بیزی، گرافی متشکل از رأس‌ها (گره‌ها) و یال‌های جهت‌دار میان آن‌هاست که رأس‌ها بیانگر متغیرهای تصادفی هستند (۳). دو نوع یادگیری در شبکه‌های بیزی مطرح است، یادگیری ساختاری و یادگیری پارامتری. در یادگیری ساختاری، گراف حاصل از روابط علت و معلولی بین متغیرها ارائه می‌شود و در یادگیری پارامتری، توزیع‌های شرطی بین متغیرها برآورد می‌شود (۴). یادگیری ساختاری شبکه بیزی یک مسأله چندجمله‌ای سخت است، زیرا تعداد ساختارهای ممکن شبکه، با افزایش تعداد گره‌های آن به‌صورت نمایی افزایش یافته و تولید همه ساختارهای ممکن و انتخاب بهترین ساختار از میان آن‌ها کاری بسیار پیچیده و زمان‌بری است (۵).

در این مقاله یادگیری ساختاری شبکه بیزی بر اساس الگوریتم ژنتیک انجام می‌پذیرد. الگوریتم ژنتیک با یک جواب اولیه که به طور ابتکاری ایجاد می‌شود شروع به کار می‌کنند. سپس یک جواب همسایگی که بهبود در تابع هدف ایجاد نماید، انتخاب می‌شود و تا زمانی که دیگر بهبودی در تابع هدف ایجاد نشود، تکرار می‌شوند. الگوریتم ژنتیک به طور گسترده و قوی در جست‌وجوهای تصادفی و تکنیک‌های بهینه‌سازی کاربرد دارد، شاید امروزه الگوریتم ژنتیک شناخته شده‌ترین نوع الگوریتم‌های تکاملی باشد. عموماً الگوریتم ژنتیک از ۵ جزء اصلی زیر تشکیل شده است (۶).

۱. نمایش ژنتیکی جواب‌های مساله؛
۲. یک راه برای ایجاد جمعیت اولیه از راه‌حل‌ها؛
۳. یک تابع ارزیابی رتبه جواب‌ها از نظر تناسب آن‌ها؛
۴. مقادیری برای پارامترهای الگوریتم ژنتیک؛
۵. اپراتورهای ژنتیکی که در طول تولید مثل ترکیب ژنتیکی فرزندان را تغیر می‌دهد.

برای مسأله یادگیری ساختار شبکه بیزی با الگوریتم ژنتیک، اغلب فرض می‌کنند ترتیبی بین متغیرهای مساله وجود دارد. بدیهی است با در نظر گرفتن فرض ترتیب بین متغیرها، فضای جست‌وجو کاهش یافته و در زمان اجرای الگوریتم و حافظه اشغال شده صرفه‌جویی می‌شود. برای انتخاب ترتیب بین متغیرها، از روش معرفی شده توسط کو و همکاران (۲۰۱۴) استفاده می‌شود. کو و همکاران با استفاده از توزیع دریخله ترتیب بین دو متغیر را تعیین کرده و آن را به چند متغیر تعمیم داده‌اند (۷). هم‌چنین در ارتباط با کارهای انجام شده روی سرطان پستان، کروز و همکاران (۲۰۰۷) برای نخستین بار شبکه بیزی مربوط به سرطان سینه را ترسیم کرده و پیش‌بینی لازم برای آن را انجام داده‌اند (۸). لازم به ذکر است که روش آن‌ها برای یادگیری ساختاری شبکه بیزی، از الگوریتم‌های موجود در نرم‌افزار R بوده است. هم‌چنین سیمئوز و همکاران (۲۰۱۵)، متغیرهای بیش‌تری در مقایسه با روش کروز و همکاران در نظر گرفته و گراف مربوط به شبکه بیزی را به‌دست آورند (۹). مشکل اصلی روش‌های ذکر شده عدم استفاده از روش جدید در یادگیری ساختاری است و فقط روی تغییر متغیرها تمرکز کرده‌اند. در صورتی که در این مقاله روش جدید برای یادگیری ساختار شبکه بیزی ارائه شده است. شبکه‌های بیزی برای متغیرهای مرتبط با سرطان سینه در ایران انجام نشده است. هدف از انجام این کار بیان رابطه علی بین متغیرهای مرتبط با سرطان سینه برای بررسی وضعیت بیمار است. نتایج نشان می‌دهد که سن، شیمی‌درمانی، رادیوتراپی و جراحی تاثیر مستقیم روی وضعیت بیمار دارند. میزان تاثیرگذاری هرمتغیر با محاسبه احتمال‌های شرطی و حاشیه‌ایی قابل محاسبه

بخش‌های این مقاله به شرح زیر است: دربخش ۲ به تشریح روش کار پرداخته و شبکه بیزی، الگوریتم ژنتیک و روش کو و همکاران (۲۰۱۴) توضیح داده می‌شود. در بخش ۳ به تحلیل یافته‌ها پرداخته می‌شود و در بخش ۴ بحث و نتیجه گیری ارائه می‌شود.

روش کار

جامعه مورد مطالعه تمام بیماران مبتلا به سرطان پستان در استان کرمان که زمان تشخیص بیماری آن‌ها از سال ۱۳۷۸ تا ۱۳۸۶ بوده است و این اطلاعات در مرکز ثبت سرطان استان کرمان موجود است. در این مطالعه نمونه‌گیری وجود ندارد. روش جمع‌آوری داده‌ها، از طریق چک لیست بوده، تعداد کل بیماران

مبتلا به سرطان پستان برابر ۹۰۰ نفر که از این میان ۸۰۳ نفر زنده و ۹۷ نفر فوت کرده بودند. پیگیری وضعیت بیماران در یک بازه‌ی زمانی ۵ ساله مورد بررسی قرار گرفته است. این داده‌ها نخستین بار توسط مددی‌زاده و بهرام‌پور (۱۳۹۳) برای پیش‌بینی مرگ‌ومیر بیماران سرطان پستان با استفاده از مدل مارکوف پنهان به‌کارگرفته شده است (۱).

- **شبکه بیزی**

شبکه‌های بیزی، یکی از ابزارهای داده‌کاوی و گروهی از مدل‌های گرافیکی احتمالی هستند که عمدتاً نشان‌دهنده روابط علیتی و معلولی بین متغیرها هستند. شبکه بیزی یک گراف جهت‌دار بدون دور است. این گراف با مجموعه رئوس V و مجموعه یال‌های E که زیرمجموعه‌ای از $V \times V$ است، به صورت $G = (V, E)$ نشان داده می‌شود. فضای E شامل حاصل‌ضرب همه سطوح گره‌های ممکن است. هر گره گراف مربوط به یک متغیر تصادفی در دامنه بوده و این گراف، یک خانواده از توزیع‌های احتمال را روی متغیرهای V نشان می‌دهد. یال‌های جهت‌دار E بیانگر وابستگی بین رأس‌های مربوط به آن یال است. یک یال $Y \rightarrow X$ در گراف رابطه بین والدY و فرزند X را شرح می‌دهد. به علاوه هر گره یک جدول احتمال شرطی دارد که احتمال هر ترکیب ممکن یک گره و والدین خود را نشان می‌دهد. اگر یک گره والدی نداشته باشد، احتمال‌های شرطی همان احتمال‌های حاشیه‌ای آن گره خواهد بود (۳). به عبارت دیگر، ایده‌ی اصلی در طراحی مدل گرافیکی، استفاده از ساختار علت و معلولی است و از نظریهٔ گراف برای ایجاد یک رابط مناسب استفاده می‌شود. این ساختارها به‌علت ظاهر گرافی که دارند نمایش مناسبی برای داده‌ها هستند که به‌صورت شهودی قابل درک است.

یادگیری در شبکه‌های بیزی به دو روش ساختاری و پارامتری صورت می‌پذیرد. در یادگیری ساختاری باید به دنبال یافتن ساختاری بهینه برای شبکه بیزی بود که با داده‌های موجود، بیش‌ترین تطابق را داشته باشد. در یادگیری پارامتری نیز احتمال‌های شرطی برآورد می‌شود. یادگیری ساختاری، خود دارای دو رویکرد اصلی است: رویکرد مبتنی بر قید که بر اساس تحلیل وابستگی موجود در داده‌ها، ساختار بهینه را می‌یابد و رویکرد مبتنی بر رتبه‌بندی- امتیازدهی که از یک تابع امتیاز برای رتبه‌بندی ساختارهای ممکن استفاده کرده و سپس با بهره‌گیری از یک الگوریتم جست‌وجو، به دنبال کشف ساختاری با بیش‌ترین امتیاز است. این روش نسبت به روش مبتنی بر قید دقیق‌تر بوده و خروجی آن ساختاری منحصر به‌فرد خواهد بود. در این پژوهش از روش مربوط به الگوریتم ژنتیک که زیر مجموعه روش رتبه‌بندی امتیازدهی است، برای یادگیری ساختاری شبکه بیزی استفاده شده است (۱۰).

الگوریتم ژنتیک در یادگیری ساختاری شبکه ب

- الگوریتم ژنتیک نوع خاصی از الگوریتم‌های تکامل است که از تکنیک‌های زیست‌شناسی مانند وراثت و جهش استفاده می‌کند. این الگوریتم برای نخستین بار توسط جان هنری هالند[1] (۱۹۷۵) معرفی شده است (۱۱).

ساختار کلی یک الگوریتم ژنتیک را می‌توان چنین تصویر کرد که پیش از هر چیز باید مکانیسمی برای تبدیل جواب هر مسأله به یک کروموزوم تعریف شود. پس از آن مجموعه‌ای از کروموزوم‌ها که در حقیقت مجموعه‌ای از جواب‌های مسأله هستند، به عنوان جمعیت اولیه در نظر گرفته می‌شوند. پس از تعریف جواب اولیه باید با به کارگیری عملیات ژنتیک[2] اقدام به ایجاد کروموزوم‌های جدید موسوم به فرزند[3] نمود. این عمل با استفاده از اپراتورهای اصلی تقاطع و جهش انجام می‌شود. پس از ایجاد جمعیت فرزندان باید با استفاده از عمل ارزیابی[4] نسبت به محاسبه برازندگی هر عضو جمعیت اقدام نمود. فرایند انتخاب[5] بر اساس مقدار برازندگی هر کروموزم انجام می‌شود، بنابراین فرایند ارزیابی مهم‌ترین بحث در فرایند انتخاب می‌باشد. بر این اساس پس از تکرار چند نسل، جمعیتی که بهترین جواب را در خود دارد، ایجاد خواهد شد. ساختار شبکه بیزی برای استفاده از الگوریتم ژنتیک به صورت ماتریس با درایه‌های صفر و یک و با ابعاد $n \times n$ به صورت زیر در نظر گرفته می‌شوند.

$$c_{ij} = \begin{cases} 1 & \text{اگر } i \text{ والد j باشد} \\ 0 & \text{در غیر این صورت} \end{cases} \qquad (1)$$

که در آن c_{ij} عنصر مربوط به سطر iام و ستون jام است. این ماتریس، ماتریس مجاورت نامیده می‌شود. اعضای جمعیت اولیه با کنار هم قرار دادن ستونی درایه‌های ایجاد می‌شوند (۱۲). این جمعیت اولیه به عنوان ورودی الگوریتم ژنتیک که در واقع یک مدل گرافیکی است، در نظر گرفته می‌شود. برای مساله یادگیری ساختار شبکه بیزی، اغلب فرض می‌کنند ترتیبی بین متغیرهای مساله وجود دارد، بدین‌ترتیب ماتریس اتصال شبکه به یک ماتریس بالا مثلثی تبدیل می‌شود. با در نظر گرفتن این فرض

[1] Holland
[1] Genetic Operation
[2] Offspring
[3] Evaluation Operation
[4] Selection

اپراتورهای تقاطع و جهش الگوریتم ژنتیک بسته هستند و فضای جستجوی کم‌تری را می‌گیرند. در این مورد طول رشته مربوط به ساختار شبکه بیزی با n گره، به جای n^2 حالت به $\binom{n}{2}$ تبدیل می‌شود، بدیهی است با در نظر گرفتن فرض ترتیب فضای جست‌وجو کاهش یافته و در زمان اجرای الگوریتم و حافظه اشغال شده صرفه‌جویی می‌شود. در بخش بعد در مورد تعیین ترتیب بین متغیرها توضیحات لازم ارایه می‌شود.

- **روش کو و همکاران (۲۰۱۴) در تعیین ترتیب بین متغیرها**

اکنون برای ترتیب بین متغیرها از روش جدیدی که توسط کو و همکاران (۲۰۱۴) معرفی شده، استفاده می‌شود. اساس کار روش کو و همکاران (۲۰۱۴) مبتنی بر توزیع دیریکله به‌صورت

$$Dir\left(p(c_1),\dots,p(c_{r_i});f(c_1),\dots,f(c_{r_i})\right) = \frac{\Gamma(\sum_{k=1}^{r_i} f(c_k))}{\prod_{k=1}^{r_i}\Gamma(f(c_k))}\prod_{k=1}^{r_i} p(c_k)^{f(c_k)} \qquad (2)$$

است ن اصل استوار است که هر اندازه فاصله توابع احتمال دیریکله متغیری بر روی تغییرات والدش بیشتر باشد، متغیر موجود به‌درستی به عنوان والد انتخاب گردیده است. منظور از والد متغیری است که از آن به سمت متغیر دیگر که فرزند نامیده می‌شود، جهت یا یال وجود داشته باشد. در این الگوریتم ابتدا جدول توافقی بین هر زوج متغیر تشکیل شده و والد احتمالی هر متغیر تعیین می‌گردد. برای این منظور، فرض می‌شود که متغیر فرزند با سطوح مختلف آن در سطر و متغیر والد با سطوح مختلف آن در ستون جدول توافقی باشند. در این صورت به منظور اینکه آیا متغیر ستون، می‌تواند والد احتمالی متغیر سطر باشد به صورت زیر عمل می‌شوکه در آن r_i حالات ممکن متغیر تصادفی، c_k -k امین سطح یک متغیر، $p(c_k)$ احتمال c_k و $f(c_k)$ فراوانی c_k است.

ایده آنها بر این اصل استوار است که هر اندازه فاصله توابع احتمال دیریکله متغیری برروی تغییرات والدش بیشتر باشد، متغیر موجود به‌درستی به‌عنوان والد انتخاب گردیده است. در این الگوریتم ابتدا جدول توافقی بین هر زوج متغیر تشکیل شده و والد احتمالی هر متغیر تعیین می‌گردد. برای این منظور، فرض می‌شود که متغیر فرزند با سطوح مختلف آن در سطر و متغیر والد با سطوح مختلف آن در ستون جدول توافقی باشند. روش کار بصورت زیر است:

- مقدار احتمال توزیع دریخله را برای ستون اول با در نظر گرفتن‌فراوانی و احتمال‌های مربوط به آن محاسبه می‌کنی مقدار احتمال توزیع دریکله را برای ستون اول با در نظر گرفتن فراوانی و احتمال‌های مربوط به آن محاسبه می‌کنیم

سپس مقدار احتمال توزیع دریکله را برای سایر ستون‌ها، با ثابت در نظر گرفتن احتمال‌های مربوط به فراوانی ستون اول محاسبه می‌شود. به عبارت دیگر در رابطه ۲ ، احتمال‌های $p(c_k)$ همان احتمالات ستون اول بوده و فقط فراوانی‌ها از ستون‌های دیگر برای محاسبه احتمال دریکله، انتخاب می‌شود. هدف از انجام این کار، محاسبه احتمال دریکله سایر ستون‌ها با در نظر گرفتن احتمال‌های ستون اول است.

- مقدار مینیمم و ماکسیمم قدم‌های قبل به عنوان U_1 و R_1 در نظر گرفته می‌شوند.
- این عملیات برای تمامی سطوح متغیر والد تکرار می‌شود. بنابراین به تعدد سطوح متغیر والد، مقادیر R و U خواهیم داشت.

در نهایت با استفاده از رابطه ۳، میزان وابستگی متغیر فرزند X_i به متغیر والد X_j محاسبه می‌شود. در واقع رابطه ۳، فاصله بین سطوح متغیر والد احتمالی و فرزند را محاسبه می‌کند که بیش‌تر بودن این فاصله نشان دهنده انتخاب صحیح متغیر والد است. هرچه میزان این فاصله بیشتر باشد نشان از تحت تأثیر بودن متغیر فرزند تحت متغیر والد است و بیانگر انتخاب درست متغیر والد احتمالی است.

$$score(X_j, X_i) = \sum_{k=1}^{r_j} (\log R_k - \log U_k) \qquad (3)$$

گام‌های فوق برای هر جفت از متغیرها به‌کار می‌رود تا والدین احتمالی یا همان ترتیب بین متغیرها بر اساس مقایسه بین امتیازها تعیین شود.

- **روش پیشنهادی**

برای اینکه بتوانیم از میان شبکه‌های ایجاد شده توسط الگوریتم ژنتیک با درنظرگرفتن ترتیب بین متغیرها، بهترین و منطبق‌ترین شبکه با داده را انتخاب کنیم باید از تابع امتیاز کمک بگیریم. تابع امتیاز معادل تابع برازندگی در الگوریتم ژنتیک است. شبکه‌های بیزی ایجاد شده به وسیله الگوریتم ژنتیک با استفاده از متر K2 ارزیابی می‌شوند. متر K2 به صورت زیر تعریف می‌شود (۱۳):

$$K_2(G,D) = \log(P(G)) + \sum_{i=1}^{n}\sum_{j=1}^{q_i}\left(\log\left(\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\right) + \sum_{k=1}^{r_i}\log(N_{ijk}!)\right)$$

که در آن،

D : پایگاه داده،

G: ساختار گراف جهت‌دار شبکه بیزی

r_i:حالات ممکن یک متغیر تصادفی (رأس یا گره) است

q_i: تعداد حالت‌های ممکن که والدین رأس X_i می تواند بگیرند.

N_{ijk}:تعداد نمونه‌هایی که در آن‌ها متغیر X_i ، k امین مقدارش را بگیرد و والدین آن j امین حالت خود را اختیار نمایند.

N_{ij}: تعداد نمونه‌هایی که والدین X_i، j امین حالت خود را بگیرند.

عبارت Log(P(G)) در متر K2، تاثیری در مقدار تابع هدف ندارد. بنابراین با در نظر گرفتن مقدار P(G) معادل یک، می‌توان از این عبارت صرف‌نظر کرد.

با توجه به تابع امتیاز تعریف شده، الگوریتم پیشنهادی به صورت زیر ارائه می‌شود:

- ابتدا ترتیب بین متغیرها بر اساس روش کو و همکاران تعین می‌شود.
- سپس مجموعه‌ای از گراف‌های تصادفی در نظر گرفته شده و ماتریس مجاورت آن‌ها به‌دست می‌آید. لازم به ذکر است با توجه به ترتیب بین متغیرها، ماتریس مجاورت بالا مثلثی خواهد بود.
- جمعیت اولیه بر اساس ماتریس مجاورت به عنوان ورودی الگوریتم ژنتیک در نظر گرفته می‌شود.

با توجه به تابع امتیاز معرفی شده، عملیات تقاطع و جهش در طول الگوریتم ژنتیک برای دست‌یابی به گراف بهتر در نظر گرفته می‌شود تا بهترین ساختار با بالاترین امتیاز به‌دست آید.

یافته‌ها

متغیرهای پیش بینی کننده در مدل عبارت‌اند از: جنس، سن بیمار در زمان تشخیص (در ۵ سطح ۳۰-۱۵، ۴۵-۳۰، ۶۰-۴۵، ۷۵-۶۰ و بالای ۷۵ سال)، سطح‌بندی تومور (grade) (با سه سطح، سلول‌های کاملاً متفاوت (درجه۱)، سلول‌های نسبتاً متفاوت (درجه۲) و سلول‌های مشابه (درجه۳)، متغیر مورفولوژی (ریخت‌شناسی توده) در ۵ سطح نئوپلاسم، کارسینوم، کارسینوم مجرایی،کارسینوم مدولی، کارسینوم لوبولی و همچنین متغیرهای شیمی درمانی، جراحی و رادیوتراپی در دو سطح بلی و خیر و متغیر وضعیت بیمار در دو سطح زنده ماندن و یا فوت شدن. کارسینوم مجرایی و کارسینوم مدولی و کارسینوم لوبولی می‌باشند.

در قسمت تحلیل نتایج ابتدا به معرفی روشی برای ارزیابی شبکه ساخته شده توسط الگوریتم ژنتیک پرداخته می‌شود. برای این منظور از شبکه آسیا که جزء معروف‌ترین شبکه‌های بیزی در حوزه‌ی یادگیری ماشین است، استفاده شده است.

شبکه آسیا شامل ۸ متغیر دو دویی (درست و نادرست) است (۱۴). شکل شماره ۱، گراف مربوط به شبکه آسیا را نشان می‌دهد. این شبکه مربوط به بیماری ریوی در بازدید از آسیا است. متغیرهای این شبکه شامل تنگی نفس، بیماری سل، سرطان ریه، برونشیت، بازدید از آسیا، سیگارکشیدن، اشعه‌ی ایکس قفسه‌ی سینه، سل در مقابل سرطان ریه برونشیت هستند. از این شبکه می‌توان داده با حجم نمونه‌های مختلف تولید کرد. این داده‌ها برای ارزیابی شبکه‌های ساخته شده با الگوریتم‌های مختلف مورد استفاده قرار می‌گیرد. در واقع به لحاظ آماری یک اعتبار سنجی انجام می‌شود. برای این منظور، الگوریتم پیشنهادی روی داده‌ها پیاده شده و شبکه بیزی ساخته می‌شود. حال سؤال این است که این شبکه بیزی ساخته شده چقدر معتبر است. برای جواب به این سؤال از امتیاز مربوط به یال‌ها استفاده می‌شود. امتیاز مربوط به یال‌ها شامل تعداد یال‌های انطباقی، تعداد یال‌های معکوس، اضافه و گمشده می‌باشد. یال‌های انطباقی در واقع همان یال‌هایی هستند که در شبکه اصلی با همان جهت وجود دارد. یال‌های معکوس نیز یال‌هایی هستند که فقط جهت عکس در مقایسه با گراف اصلی را دارند. به عبارت دیگر هدف اصلی در بحث یادگیری ساختاری شبکه‌های بیزی، ایجاد ساختاری با بیشترین میزان تطبیق با شبکه‌ی اصلی است. برای ارزیابی، نمونه‌های با اندازه‌های ۱۰۰۰، ۲۰۰۰، ۵۰۰۰ و ۱۰۰۰۰ از شبکه‌ی اصلی آسیا تولید شده و الگوریتم پیشنهادی روی این مجموعه داده پیاده می‌شود. امتیاز یال‌ها در جدول شماره ۱ ارایه شده است. به‌دلیل بالا بودن تعداد یال‌های انطباقی در مقابل خطاها، شبکه‌ی بیزی ساخته شده از اعتبار بالایی برخوردار است.

پس از مشخص شدن ترتیب بین متغیرها، الگوریتم ژنتیک را با در نظر گرفتن پارامترهای موجود در جدول شماره ۲ اجرا می‌شود. در شکل شماره ۲، نتیجه حاصل از الگوریتم ژنتیک ارایه شده است.

هم‌چنین گراف مربوط به شرایط مختلفی که بیمار می‌تواند زنده بماند و یا فوت کند نیز در شکل‌های شماره ۳ و ۴ ارائه

شده است.

با درنظرگرفتن این نکته که بسیاری از زنان مبتلا به سرطان سینه تمایل به انتخاب نوع درمان خود دارند، شبکه بیزی ترسیم شده می‌تواند اطلاعات بسیار خوبی را در اختیار آنان قرار دهد. این گراف و احتمال‌های مربوط به آن راهنمای خوبی برای انتخاب نوع درمان است.باتوجه به نتایج دو شکل ۳ و ۴، افراد در رده سنی ۶۵-۴۰ سال، بیشتر درگیر این بیماری هستند. بنابراین با توجه به یال موجود بین سن و وضعیت زنده ماندن یا نماندن، می‌توان نتیجه گرفت که سن تاثیر مستقیم روی وضعیت زندگی دارد. همچنین با توجه به مقایسه شکل‌های ۳ و ۴، جراحی اصلی‌ترین روش درمانی سرطان سینه است و اغلب زنانی که تحت عمل جراحی قرار می‌گیرند، می‌توانند به زندگی خود ادامه دهند. زیرا همان‌طور که در شکل ۴ مشخص است، ۸۱٪ از بیمارانی که تحت عمل جراحی قرار نگرفته‌اند و درمان آن‌ها صرفاً به شیوه شیمی درمانی یا رادیوتراپی صورت گرفته است، امکان ادامه زندگی آن‌ها کمتر است. از سویی دیگر، بیمارانی که تحت عمل جراحی قرار گرفته‌اند و کمتر در معرض پرتودرمانی یا شیمی درمانی بوده‌اند، زنده می‌مانند. در حالت کلی می‌توان نتایج زیر را برای وضعیت زنده ماندن و یا فوت بیماران استخراج کرد.

- افراد در رده سنی ۴۵ تا ۶۰ سال، بیشتر درگیر تومور درجه دو هستند و افراد در این رده سنی بیشتر از افراد دیگر زندگی خود را از دست داده‌اند.
- ۸۱ درصد از بیمارانی که تحت عمل جراحی قرار نگرفته‌اند و درمان آن‌ها صرفاً به شیوه شیمی درمانی یا رادیوتراپی صورت گرفته است، زندگی خود را از دست داده‌اند.
- ۹۵ درصد بیماران با گرید تومور دو، تحت شیمی درمانی و رادیو تراپی قرار گرفته‌اند و زندگی خود را از دست داده‌اند.
- ۶۵ درصد از کسانی که رادیوتراپی انجام داده‌اند، عمل جراحی نکرده و این افراد زندگی خود را از دست داده‌اند.
- ۹۱درصد بیماران تحت شیمی درمانی زندگی خود را از دست دادند.
- ۹۰ درصد بیماران با گرید تومور دو، که تحت رادیوتراپی قرار نگرفته‌اند و عمل جراحی انجام داده‌اند، زنده مانده‌اند.
- رادیوتراپی قرار نگرفته‌اند و عمل جراحی انجام داده‌اند، زنده مانده‌اند ۵۲ درصد از بیمارانی که تحت عمل جراحی قرار گرفته‌اند و کمتر در معرض پرتودرمانی یا شیمی درمانی بوده‌اند، زنده مانده‌اند.

البته واضح است شیوه درمان با توجه به نوع سلول درگیر و سطح پیشرفت بیماری، متفاوت است.که با تعریف سناریوهای مختلف روی گراف‌های فوق، می‌توان نتایج را بر این اساس گزارش کرد. جراحی به اشکال مختلف زیر انجام می‌شود:

- رادیکال ماستکتومی[۱]: برداشتن کامل پستان با عضلات پکتورال زیرین و غدد لنفاوی آگزیلاری است.
- رادیکال ماستکتومی وسیع[۲]: شامل رادیکال ماستکتومی و دیسکسیون کامل غدد لنفاوی داخلی است.
- رادیکال ماستکتومی تعدیل شده[۳]: در این روش رادیکال ماستکتومی بدون برداشت عضله پکتورالیس ماژور انجام می‌شود و پوست و غدد زیر بغل نیز با آن وسعت برداشته نمی‌شوند و لذا نتایج بهتری از لحاظ عملکرد و زیبایی برای بیمار به همراه دارد.
- ماستکتومی توتال[۴]: این عمل شامل برداشتن تمام نسج پستان که به طرف زیر بازو کشیده می‌شود و پوست اطراف آن است.
- ماستکتومی کاملا وسیع[۵]: علاوه بر ماستکتومی بخش نسبتاً وسیع از غدد لنفاوی برداشته می‌شود.
- ماستکتومی پارشیال یا سگمنتال یا لامپکتومی[۶]: این روش جراحی محافظتی نیز نامیده می‌شود و فقط شامل برداشتن توده‌ها و بافت‌های سرطانی است.

مددی‌زاده و همکاران (۱۳۹۳) با استفاده از مدل مارکوف پنهان روی داده‌های سرطان پستان، به پیش‌بینی وضعیت بیمار پرداخته و دقت مدل را برابر ۹۳ درصد گزارش کرده‌اند. برای این منظور تعدا حالات پنهان در مدل مارکوف برابر ۵ به‌دست آمده است. در این مقاله دقت پیش‌بینی وضع بیمار با استفاده از احتمالاهای شرطی به میزان ۹۷ درصد به‌دست آمده که در مقایسه با کار قبل، دقت پیش‌بینی وضع بیمار افزایش یافته است. برای محاسبه میزان دقت پیش‌بینی در شبکه‌های بیزی، بعد از

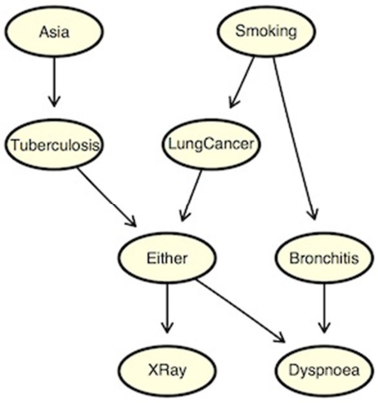
---

[۱] Radical mastectomy
[۲] Extended Radical mastectomy
[۳] Modified Radical mastectomy
[۴] Total mastectomy
[۵] Extended mastectomy
[۶] Surgery, lumpectomy, segmental mastectomy

به‌دست آمدن ساختار شبکه بیزی، توابع احتمال‌های شرطی والد و فرزندان محاسبه شده و در نهایت احتمال وضعیت زنده ماندن یا فوت شدن بیمار با استفاده از تابع احتمال‌های حاشیه‌ای محاسبه می‌شود. با انطباق دادن وضعیت پیش‌بینی شده با وضعیت واقعی بیماران، میزان دقت پیش‌بینی به‌دست می‌آید. لازم به ذکر است که تابع احتمال حاشیه‌ای وضع بیمار از رابطه زیر به‌دست می‌آید:

$$P(S) = \sum P(G).P(A).P(M|G).P(G|A).P(R|M,G)P(C|G,M).\ P(Su|M).\ P(S|C,R,A,Su)$$

که در آن S وضع بیمار، G جنس بیمار، A سن بیمار در زمان تشخیص، M مورفولوژی، R رادیوتراپی، C شیمی درمانی و Su جراحی است.



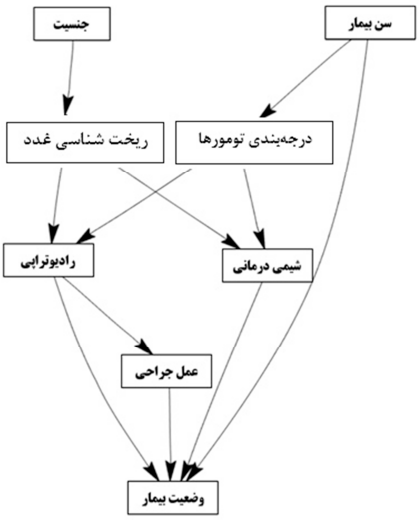
شکل شماره **۱**- شبکه آسیا

جدول شماره **۱**- امتیاز یال‌ها

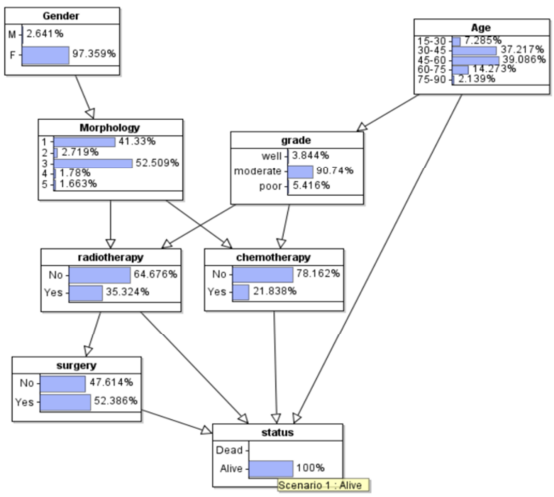
داده	ملاک (تعداد یال‌ها)	حجم نمونه			
		۱۰۰۰	۲۰۰۰	۵۰۰۰	۱۰۰۰۰
Asia	انطباقی با جهت درست	۵	۵	۶	۷
	گمشده	۰	۰	۰	۰
	انطباقی با جهت معکوس	۴	۳	۳	۳
	اضافی	۰	۰	۱	۱
	خطا (جمع گمشده، معکوس و اضافی)	۴	۳	۴	۴

جدول شماره **۲**- مقادیر مشخصه‌های الگوریتم ژنتیک

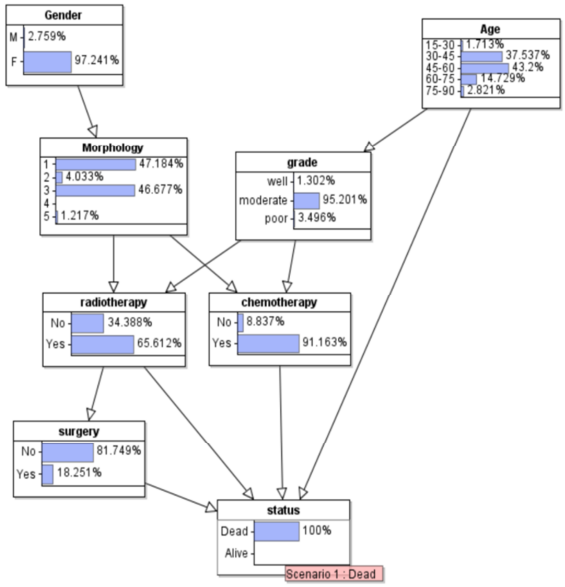
تعداد دفعات تکرار الگوریتم	**MaxIt**	**۵۰**
تعداد اعضای جمعیت اولیه	nPop	۲۰
درصد تقاطع	Pc	٪۵۰
درصد جهش	Pm	٪۳۰
نرخ جهش	Mu	۰/۰۵



شکل شماره ۲- مدل گرافیکی حاصل از به‌کارگیری الگوریتم ژنتیک



شکل شماره **۳**- گراف احتمال شرطی متغیرهای مختلف وقتی بیمار زنده می‌ماند



شکل شماره ٤- گراف احتمال شرطی متغیرهای مختلف وقتی بیمار فوت می کند

بحث

با وجودی که در دو دهه‌ی اخیر پیشرفت‌های مهمی در زمینه‌ی تشخیص زودرس و درمان به موقع و کاهش مرگ و میر برای سرطان پستان در زنان ایجاد شده است، اما هنوز این سرطان جزء شایع‌ترین بیماری‌های بدخیم زنان است. این سرطان که نخستین سرطان در بین زنان ایرانی است منجر به از دست رفتن ۰/۰۱ سال به ازای هر یک هزار نفر جمعیت می‌شود و بار ناشی از سرطان پستان، ۶ درصد سال‌های عمر از دست رفته به خاطر سرطان‌ها را در کشور به خود اختصاص می‌دهد (۲، ۱)

تشخیص در مراحل اولیه سرطان به عنوان یک عامل حیاتی در درمان موفق به شمار می‌رود. بنابراین تشخیص روابط علیتی بین متغیرهای مرتبط با سرطان پستان از مباحث بسیار مهم در تشخیص بهترین نوع درمان است. تاکنون شبکه‌های بیزی برای تشخیص بهترین نوع درمان در ایران برای سرطان پستان به‌کار گرفته نشده است که در این تحقیق این مهم انجام پذیرفته است. سرطان پستان می‌تواند هم به صورت یک بیماری غیر تهاجمی بوده و به محل آغاز بیماری محدود شود و یا تهاجمی بوده و به قسمت‌های مختلف گسترش یابد. سرطان‌های غیر تهاجمی شامل کارسینومای داکتال اینسایتو[۱] و کارسینومای لوبولار اینسایتو[۲] هستند. سرطان پستان تهاجمی هنگامی رخ می‌دهد که سلول‌ها از سطح غشای اولیه که بافت همبند پایه را پوشش می‌دهند گسترش می‌یابند. این بافت پر از رگ‌های خونی و کانال‌های لنفاتیک است که قادر به حمل سلول‌های سرطانی به خارج از بافت پستان هستند. سرطان‌های پستان تهاجمی، کارسینومای داکتال تهاجمی[۳] و کارسینومای لوبولار تهاجمی[۴] را شامل می‌شود در مقاصد درمانی، سرطان پستان را به ۵ مرحله[۵] بالینی تقسیم می‌کنند:

۱- مرحله صفر؛ سلول‌های سرطانی را توصیف می‌کند که غیر تهاجمی بوده، اما ریسک بلند مدتی از تهاجمی شدن مطرح می‌کنند. در این مرحله اندازه تومور اولیه بسیار کوچک و کم‌تر از یک سانتی‌متر است.

۲- مرحله اول؛ تومورهایی را توصیف می‌کند که بیش‌تر از ۲ سانتی‌متر عرض نداشته و به خارج از بافت پستان و گره‌های لنفاوی زیر بغل گسترش نیافته‌اند.

۳- مرحله دوم؛ تومور در حدود ۲ سانتی‌متر بوده و به گره‌های لنفاوی زیر بازو گسترش یافته و یا اندازه تومور اولیه حدود ۵ سانتی‌متر بوده اما به گره‌های لنفاوی گسترش نیافته است.

۴- مرحله سوم؛ تومور بیش از ۵ سانتی‌متر عرض داشته و به گره‌های لنفاوی یا سایر بافت‌های اطراف توسعه یافته است.

۵- مرحله چهارم؛ سرطان‌های متاستاز داده نامیده می‌شوند و به سایر قسمت‌های بدن نیز گسترش یافته‌اند. خطر بیماری‌ها با بالا رفتن هر مرحله افزایش یافته و شانس زنده ماندن کاهش می‌یابد.

با توجه به موارد فوق و تحلیل نتایج، افراد در رده سنی ۶۵-۴۰ سال، بیش‌تر درگیر سرطان پستان هستند. بنابراین با توجه به یال موجود بین سن و وضع زنده ماندن یا نماندن، می‌توان نتیجه گرفت که سن تأثیر مستقیم روی وضعیت زنده ماندن دارد.

با توجه به اینکه الگوریتم ژنتیک برای یادگیری ساختاری شبکه بیزی به شدت تحت تأثیر چگونگی ترتیب بین متغیرها است، بنابراین انتخاب روش مناسب برای ترتیب بین متغیرها از اهمیت بالایی برخوردار است. بنابراین مقایسه روش‌های مختلف تعیین ترتیب از کارهای آتی در این زمینه به شمار می‌آید. هم‌چنین استفاده از سایر الگوریتم‌های فراابتکاری نیز پیشنهادهای آتی برای پژوهش است.

نتیجه‌گیری

در این پژوهش نشان داده شد که، جراحی اصلی‌ترین روش درمانی سرطان پستان است و اغلب زنانی که تحت عمل جراحی قرار می‌گیرند، می‌توانند با امید بیش‌تری به زندگی خود ادامه دهند. زیرا همان‌طور که در شکل شماره ۳ مشخص است، ۸۱ درصد از بیمارانی که تحت عمل جراحی قرار نگرفته‌اند و درمان آن‌ها صرفاً به شیوه شیمی درمانی یا پرتودرمانی صورت گرفته است، احتمال ادامه زندگی آن‌ها کم‌تر است. از سویی دیگر، بیمارانی که تحت عمل جراحی قرار گرفته‌اند و کم‌تر در معرض پرتودرمانی یا شیمی درمانی بوده‌اند، زنده مانده‌اند. بنابراین در یک نتیجه‌گیری کلی می‌توان گفت که عمل جراحی نقش تعیین کننده‌ای در زندگی افرادی دارد که مبتلا به سرطان سینه هستند.

---

[۱] Ductal Carcinoma Insitu (DCIS)
[۲] Lobular Carcinoma Insitu (LCIS)
[۳] Intra Ductal Carcinoma (IDC)
[۴] Invasive Lobular Carcinoma (ILC)
[۵] Stage

تشکر و قدردانی

نویسندگان از جناب آقای دکتر عباس بهرام‌پور عضو هیأت علمی دانشگاه علوم پزشکی کرمان و آقای فرزان مددی‌زاده برای در اختیار گذاشتن داده‌های سرطان پستان در استان کرمان، تقدیر و تشکر به عمل می‌آورند.

منابع

1. Madadizadeh F, Bahrampour A. Mortality prediction of breast cancer patients using Hidden Markov model. Master of Scince Thesis, Kerman University of Medical Sciences. 1393.
2. Tavakolpour V, Tavakolpour S. Evaluation of genes involved in hereditary breast cancer, First National Congress of Biology and Natural Sciences of Iran, Tehran. 1393; 1-7.
3. Pearl J. Probabilistic Reasoning in Expert Systems: Networks of Plausible Reasoning. San Mateo, CA: Morgan Kaufmann (Pubs.); 1988.
4. Heckerman D. Bayesian networks for data mining. Data mining and knowledge discovery. 1997; 1: 79-119.
5. Larrañaga P, Poza M, Yurramendi Y, Murga RH, Kuijpers CMH. Structure learning of Bayesian networks by genetic algorithms: A performance analysis of control parameters. IEEE transactions on pattern analysis and machine intelligence. 1996; 18: 912-26.
6. Michalewicz Z, Schoenauer M. Evolutionary algorithms for constrained parameter optimization problems. Evolutionary computation. 1996; 4: 1-32.
7. Ko S, Kim D-W. An efficient node ordering method using the conditional frequency for the K2 algorithm Pattern Recognition Letters. 2014; 40: 80-7.
8. Cruz-Ramírez, N., Acosta-Mesa, H. G., Carrillo-Calvet, H., Nava-Fernández, L. A., & Barrientos-Martínez, R. E. Diagnosis of breast cancer using Bayesian networks: A case study. Computers in Biology and Medicine. 2007; 37(11), 1553-1564.
9. Simoes PW, Silva GD, Moretti GP, Simon CS, Winnikow EP, Nassar SM, Medeiros LR, Rosa MI. Meta analysis of the use of Bayesian networks in breast cancer diagnosis. *Cadernos De* SaúDe PúBlica. 2015; 31: 26-38.
10. Eskandari F, Rezaei Tabar V, Naghizadeh S. Risk assessment and decision analysis in Bayesian networks. Allameh Tabataba'i University Press. 1396.
11. Holland JH. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT press; 1992.
12. Nielsen TD, Jensen FV. Bayesian networks and decision graphs: Springer Science & Business Media; 2009.
13. Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Machine learning. 1992; 9: 309-47.
14. Lauritzen SL, Spiegelhalter DJ. Local computations with probabilities on graphical structures and their application to expert systems. Journal of the Royal Statistical Society Series B (Methodological). 1988:157-224.

Iranian Journal of Epidemiology Summer 2018; 14 (2): 166-176.

Original Article

Causal Relationship between Variables Related to Breast Cancer using Bayesian Networks

Heidari S[1], Kavousi A[2], Rezaei V[3]

1- Msc Student, Department of Biostatistics, SBMU School of Para-Medical Sciences, Shahid Beheshti Medical University, Tehran, Iran
2- Associate Professor, School of Health, Safety and Environment, Shahid Beheshti Medical University, Tehran, Iran
3- Assistant Professor, Department of Statistics, Faculty of Mathematics and Computer Sciences, Allameh Tabataba'I University, Tehran, Iran
Corresponding author: Kavousi A, kavousi @sbmu.ac.ir
(Received 26 August 2017; **Accepted** 3 March 2018)

Background and Objectives: Breast cancer is the most common cancer in Iran. It can be prevented by rapid diagnosis of the disease. Thus, it is necessary to determine the causal relationships between variables related to breast cancer. Bayesian network is a data mining tool that shows the causal relationship between different variables. In this paper, a Bayesian network was applied to find causal relationships between breast cancer variables using a genetic algorithm in a graphical model.

Methods: in this applied study, data were collected from 900 breast cancer patients in Kerman Province from 1999 to 2008. For data analysis, we used a probabilistic graphical model representing the causal relationship between variables.

Results: The results showed that surgery was the most important treatment for breast cancer. Based on the conditional and marginal probabilities, the women who underwent surgery had higher hopes of living longer. Moreover, 81% of the patients who did not undergo surgery only received chemotherapy or radiotherapy were less likely to have long lives.

Conclusion: People aged 40-65 years are more likely to have breast cancer. Moreover, the variables of age, surgery, chemotherapy, and radiotherapy had a direct effect on the status of the patients and there were direct edges from these variables to the status of the patients.

Keywords: Breast cancer, Bayesian network, Genetic algorithm