مقدمه‌ای بر استنتاج و یادگیری در شبکه‌های بیزی

فهیمه مرادی[۱]، علی کریم‌نژاد[۲]، سودابه شمه‌سوار[۳]

چکیده:

شبکه‌های بیزی ابزار جدیدی در مدل‌بندی پدیده‌ها و سیستم‌های ایستا و پویا هستند و در زمینه‌های مختلفی از جمله تشخیص بیماری‌ها، پیش‌بینی آب و هوا، تصمیم‌گیری و دسته‌بندی کاربرد دارند. یک شبکه بیزی یک مدل گرافی-احتمالی است که ارتباط‌های علت و معلولی بین متغیرهای تصادفی را نشان می‌دهد و از یک گراف بدون دور جهت‌دار و یک مجموعه از احتمال‌های شرطی تشکیل شده است. دو موضوع مهم در مدل‌بندی یک مجموعه داده با شبکه بیزی یادگیری ساختاری و یادگیری پارامتری شبکه است. در این مقاله یک شبکه بیزی با ساختار معلوم را در نظر می‌گیریم و با شبیه‌سازی تلاش می‌کنیم ساختار شبکه را با استفاده از دو الگوریتم متداول PC و K_2 یاد بگیریم. سپس، به یادگیری پارامترهای شبکه می‌پردازیم و برآوردهای ماکزیمم درستنمایی، ماکزیمم احتمال پسین و میانگین پسین پارامترهای مورد علاقه را به دست می‌آوریم. در ادامه، عملکرد برآوردها را با استفاده از معیار واگرایی کولبک-لایبلر مقایسه می‌کنیم و در نهایت، با استفاده از یک مجموعه داده واقعی، به یادگیری ساختاری و پارامتری شبکه می‌پردازیم تا امکان پیاده‌سازی روش‌های پیشنهادی بر روی داده‌های واقعی را نشان دهیم.

واژه‌های کلیدی: توزیع دیریکله، شبکه بیزی، یادگیری پارامتری، یادگیری ساختاری.

۱ مقدمه

شبکه بیزی به یک گراف بدون دور جهت‌دار اطلاق می‌شود که پارامترهایی در قالب احتمال شرطی این ساختار را از حالت کیفی به حالت کمی تبدیل می‌کنند. رأس‌های این گراف متغیرهای تصادفی هستند و یال‌های جهت‌دار آن بیانگر وابستگی بین رأس‌ها می‌باشند. پارامترهای موجود در شبکه میزان این وابستگی را مشخص می‌کنند. شبکه بیزی به یادگیری ارتباطات سببی کمک می‌کند و به خاطر ساختار گرافیکی‌ای که دارد به صورت شهودی قابل درک است. برای ساختن شبکه بیزی می‌توان از اطلاعات پیشین و اطلاعات افراد خبره استفاده کرد و با کمک تکنیک‌های آمار بیزی، داده‌ها و اطلاعات موجود در آن زمینه را با هم ترکیب کرده و به شبکه محتمل‌تری نسبت به شبکه‌هایی که قبلا ساخته شده است دست یافت. شبکه بیزی ترکیبی از اصول نظریه گراف، نظریه احتمال، علوم کامپیوتر و آمار می‌باشد و در بسیاری از زمینه‌ها از جمله تشخیص، پیش‌بینی، دسته‌بندی و تصمیم‌گیری کاربرد دارد. همچنین از شبکه بیزی در مدل‌بندی شبکه تنظیمی بیان ژن، مدل‌بندی ترافیک بزرگراه و تشخیص صدا نیز استفاده شده است. مدل‌بندی داده‌ها با شبکه‌ی بیزی از دو مرحله تشکیل شده است. مرحله‌ی اول تعیین ساختار شبکه است که به آن یادگیری ساختاری گفته می‌شود و مرحله‌ی دوم تعیین پارامترهای ساختار شبکه بیزی است که به آن یادگیری پارامتری گفته می‌شود. یادگیری ساختاری شبکه بیزی یک مسئله NP- سخت است و روش‌های متعددی برای یادگیری ساختاری وجود دارد. انواع روش‌هایی که برای یادگیری ساختاری شبکه بیزی وجود دارد در دو دسته‌ی کلی روش‌های محدودیت گرا و روش‌های امتیازگرا قرار می‌گیرند. هنگامی که تعداد نمونه‌ها کم باشد، روش‌های محدودیت‌گرا با خطاهای آماری زیادی مواجه هستند و هنگامی که تعداد متغیرها زیاد باشد روش‌های امتیازگرا (به علت کند شدن سرعت انجام روش‌های گشت) به مشکل برمی‌خورند. در این مقاله بر مدل‌بندی با شبکه بیزی تمرکز می‌کنیم و پس از یادگیری ساختار شبکه بیزی مورد مطالعه به یادگیری پارامتری در آن شبکه می‌پردازیم.

[۱] دانشجوی کارشناسی ارشد آمار ریاضی، گروه آمار دانشگاه تهران

[۲] دانشجوی دکتری آمار، گروه آمار دانشگاه تهران

[۳] استادیار گروه آمار دانشگاه تهران.

[۴] Tomas Bayes

ایده اصلی شبکه‌های بیزی بر پایه قانون بیز است که توسط توماس بیز[۴] (۱۷۲۰) ارائه شده و به همین دلیل آن را شبکه بیزی می‌نامند. طبق قانون بیز، محققین شبکه‌های از قبل ساخته شده برای متغیرها (که شبکه‌های پیشین[۵] نامیده می‌شوند) را در نظر می‌گیرند و آن‌ها را با مجموعه داده‌ها ترکیب می‌کنند تا شبکه پسین، که شبکه محتمل‌تری نسبت به شبکه‌های پیشین است، ساخته شود. اصطلاح شبکه بیزی اولین بار توسط پرل[۶] (۱۹۸۸) مطرح شد. امروزه با گذشت ۲۷ سال از پیدایش شبکه بیزی، این شبکه‌ها به عنوان یک ابزار قوی در علوم مختلف از جمله بیوانفورماتیک برای مدل‌بندی شبکه بیان ژن به‌کار می‌روند. در بین مطالعات انجام شده در این زمینه می‌توان به آنگ[۷] و همکاران (۲۰۰۲) و فریدمن[۸] (۲۰۰٤) اشاره کرد. آنگ و همکاران (۲۰۰۲) یک شبکه بیزی پویای زمان پیوسته برای مدل‌بندی داده‌های بیان ژن ارائه دادند. فریدمن (۲۰۰٤) داده‌های تجربی میزان بیان ژن را به صورت شبکه بیزی ایستا مدل‌بندی کرده است که در آن شبکه امکان پیش‌بینی دقیق وجود ندارد. برای مطالعه بیشتر در شبکه‌های بیزی می‌توان به جنسن[۹] و نیلسن[۱۰] (۲۰۰۷) و کوسکی[۱۱] و نوبل[۱۲] (۲۰۰۹) مراجعه کرد.

ساختار مطالب بیان شده در این مقاله به شرح زیر است: در بخش دوم تعاریف و مفاهیم مورد نیاز را بیان می‌کنیم. در بخش سوم روش‌های مختلف یادگیری ساختاری شبکه بیزی را معرفی می‌کنیم. در بخش چهارم به یادگیری پارامتری شبکه بیزی می‌پردازیم. سپس، در بخش پنجم یک مجموعه داده‌ی شبیه‌سازی شده را با شبکه بیزی مدل‌بندی می‌کنیم و در بخش ششم با استفاده از یک مجموعه داده واقعی، به یادگیری ساختار و پارامترهای شبکه می‌پردازیم. در نهایت در بخش هفتم، از مطالب بیان شده نتیجه‌گیری می‌کنیم.

۲ تعاریف و مفاهیم مورد نیاز

در این بخش به بیان تعاریف و مفاهیم مورد نیاز بخش‌های آتی مقاله می‌پردازیم. یک گراف بدون دور جهت‌دار را دگ[۱۳] می‌نامیم و شبکه بیزی دگی است که رأس[۱٤]های آن متغیرهای تصادفی هستند. مجموعه‌ی همه‌ی رأس‌هایی که از رأس مورد نظر یک یال[۱۵] جهت‌دار به آن‌ها وجود داشته باشد را مجموعه اولاد آن رأس و مجموعه‌ی همه‌ی رأس‌هایی که از آن‌ها یک یال جهت‌دار به رأس مورد نظر وجود داشته باشد را مجموعه اجداد آن رأس می‌نامند. شبکه‌ای که هیچ یالی در آن وجود نداشته باشد را شبکه تهی و شبکه‌ای که در آن بین هر رأس و همه‌ی رأس‌های دیگر یال وجود داشته باشد را شبکه کامل می‌نامند. یک شبکه بیزی از دو بخش B_s (ساختار بیزی[۱٦]) و B_p (احتمال بیزی[۱۷]) تشکیل شده و معمولا آن را با (B_s, B_p) نشان می‌دهند. B_s و B_p به ترتیب به ساختار و پارامترهای شبکه بیزی اشاره می‌کنند. لازم به ذکر است که معمولا عبارت‌های ساختار شبکه بیزی (B_s) و دگ (G) را به جای یکدیگر به کار می‌برند. یکی از ویژگی‌های مهم شبکه بیزی این است که هر رأس به شرط اجدادش از مجموعه رأس‌های غیراولاد مستقل است. از طرف دیگر ساختار شبکه‌های بیزی دارای دور جهت‌دار نیست و می‌توان X_iها را طوری مرتب کرد که اجداد X_i در مجموعه‌ی $\{X_1, \ldots, X_{i-1}\}$ و اولاد آن در مجموعه‌ی $\{X_{i+1}, \ldots, X_n\}$ قرار

[۵] Prior Networks
[۶] Pearl
[۷] Ong
[۸] Friedman
[۹] Jensen
[۱۰] Nielsen
[۱۱] Koski
[۱۲] Noble
[۱۳] DAG (DirectedAcyclic Graph)
[۱٤] Node
[۱۵] Edge
[۱٦] Bayesian Structure
[۱۷] Bayesian Probability

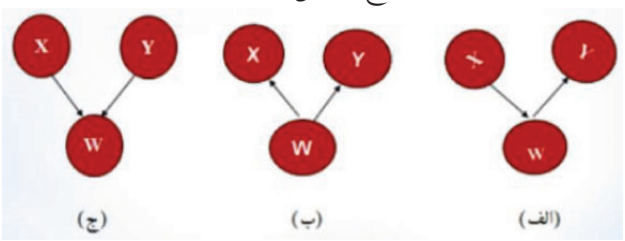
بگیرند، بنابراین طبق قانون احتمال کل، تابع احتمال توأم را می‌توان به صورت زیر نوشت:

$$P(X_1,\ldots,X_n)=\prod_{i=1}^{n}P(X_i|X_1,\ldots,X_{i-1})$$
$$=\prod_{i=1}^{n}P(X_i|X_{i\text{اجداد}}),$$

دو شبکه بیزی دارای توزیع احتمال توأم یکسان را معادل گویند. دو متغیر تصادفی X و Y را به شرط مجموعه ی $\mathcal{Z}$ مستقل می‌نامند و این مطلب را با نماد $X \perp Y|\mathcal{Z}$ نشان می‌دهند هرگاه برای هر $\mathcal{Z}$ داده شده، مستقل باشند و در این صورت داریم:

$$P(X,Y|\mathcal{Z})=P(X|\mathcal{Z})\ P(Y|\mathcal{Z}).$$

مسیر بین دو رأس X و Y توسط مجموعه‌ی $\mathcal{Z}$ بلوک شده خوانده می‌شود. هرگاه به ازای هر $W \in \mathcal{Z}$ جهت یال‌ها به صورت ساختار (الف) یا (ب) شکل ۱ باشد و یا اینکه برای هر $W \notin \mathcal{Z}$ جهت یال‌ها به صورت ساختار (ج) شکل ۱ باشد.



شکل ۱. بلوک شدن مسیر

مجموعه‌ی $\mathcal{Z}$ رأس Y را از رأس X در گراف G جدا می‌کند اگر و تنها اگر هر مسیری که از X به Y وجود دارد توسط مجموعه‌ی $\mathcal{Z}$ بلوک شود. این مطلب با نماد $d-sep_G(X,Y|\mathcal{Z})$[۱۸] نشان داده می‌شود.

۳ یادگیری ساختاری

هدف از یادگیری ساختاری یافتن بهترین ساختار برای شبکه بیزی است که با داده‌ها یا اطلاعات موجود مطابقت داشته و از لحاظ پیچیدگی بهینه باشد (دارای کمترین پیچیدگی باشد). این یادگیری از دو روش محدودیت‌گرا (مبتنی بر قید[۱۹]) و امتیازگرا[۲۰](گشت-رتبه‌بندی) انجام می‌شود که در ادامه به اختصار به معرفی هر یک می‌پردازیم.

۱.۳ روش‌های محدودیت‌گرا

در این روش‌ها ساختار شبکه بیزی با مشخص کردن رابطه‌ی استقلال شرطی بین گره‌ها به دست می‌آید و برای مشخص کردن وابستگی یا عدم وابستگی بین متغیرها از آزمون‌های استقلال استفاده می‌کنند. تعیین ساختار شبکه بیزی با روش‌های یادگیری مبتنی برقید یک فرآیند دو مرحله‌ای است. در مرحله اول، ساختار شبکه تعیین می‌شود به طوری که یال‌های ساختار بدون جهت هستند و اصطلاحا به این ساختار کالبد دگ گفته می‌شود. در این مرحله برای تعیین یال‌ها از آزمون‌های استقلال شرطی مانند آزمون استقلال χ^2، آزمون Z فیشر و آزمون استقلال نسبت درستنمایی استفاده می‌شود. سپس، در مرحله دوم یال‌های ساختاری که در مرحله اول به‌دست آمده جهت‌دار می‌شوند. در الگوریتم‌های مختلف بر اساس قیود متفاوتی یال‌ها جهت‌دار می‌شوند (جنسن و نیلسن، ۲۰۰۷).

الگوریتم‌های یادگیری ساختاری زیادی مانند الگوریتم استقلال شرطی، الگوریتم اجداد و اولاد[۲۱] (PC) (جنسن و نیلسن، ۲۰۰۷) و الگوریتم دوگانه اجداد و اولاد (ابراهیمی، ۱۳۹۰) وجود دارد. در اینجا تنها به معرفی مهم‌ترین الگوریتم یادگیری ساختاری، یعنی الگوریتم PC می‌پردازیم. برای درک بهتر این الگوریتم گزاره زیر را بیان می‌کنیم.

گزاره ۱.۳. در یک شبکه بیزی متغیرهای تصادفی X و Y به شرط مجموعه $\mathcal{Z}$ از یکدیگر مستقلند ($X \perp Y|\mathcal{Z}$) اگر و تنها اگر X و Y توسط مجموعه $\mathcal{Z}$ از یکدیگر جدا شده باشند ($d-sep_G(X,Y|\mathcal{Z})$).

بر طبق این گزاره، اگر X و Y به شرط مجموعه $\mathcal{Z}$ از یکدیگر مستقل باشند، آنگاه به ازای هر $W \in \mathcal{Z}$ جهت یال‌ها به صورت

[۱۸] "d" for "directed graph"

[۱۹] Constraint-Based

[۲۰] Scoring-Based

[۲۱] Parents and Children

یکی از دو ساختار (الف) یا (ب) شکل ۱ است.

در قسمت (الف) تابع احتمال توأم متغیرهای W، X و Y عبارت از:

$$P(W,X,Y) = P(Y|W)P(W|X)P(X)$$
$$= P(Y|W)P(W,X),$$

است و در قسمت (ب) تابع احتمال توأم متغیرهای W، X و Y عبارت از:

$$P(W,X,Y) = P(Y|W)P(X|W)P(W)$$
$$= P(Y|W)P(W,X),$$

است. از آنجایی که تابع احتمال توأم متغیرها در هر دو ساختار یکسان است این دو ساختار با یکدیگر معادلند. این دو ساختار، ساختارهای V–شکل نامیده می‌شوند. بر طبق این گزاره، اگر X و Y به شرط مجموعه $\mathcal{Z}$ از یکدیگر مستقل نباشند آنگاه به ازای هر $W \in \mathcal{Z}$ جهت یال‌ها به صورت ساختار (ج) شکل ۱ است و تابع احتمال توأم متغیرها عبارت از:

$$P(W,X,Y) = P(W|X,Y)P(Y)P(X).$$

است.

۱.۱.۳ الگوریتم یادگیری ساختاری اجداد و اولاد (PC)

این الگوریتم با آزمون استقلال χ^2، استقلال یا وابستگی متغیرها را بررسی می‌کند و برای تشخیص استقلال یا وابستگی دو رأس، به دنبال پیدا کردن مجموعه جداکننده در بین تمام همسایه‌های دو رأس است و با استفاده از مفهوم مجموعه جداکننده و دو اصل زیر به جهت‌یابی یال‌ها اقدام می‌کند (جنسن و نیلسن، ۲۰۰۷):

۱. الگوریتم از به وجود آمدن ساختار V–شکل در گراف جلوگیری می‌کند، یعنی ساختار $X \longrightarrow Y \leftrightarrow Z$ را به شکل $X \longrightarrow Y \leftarrow Z$ جهت‌دهی می‌کند.

۲. الگوریتم از به وجود آمدن دور در شبکه بیزی جلوگیری می‌کند. از آن جایی که بنابر تعریف، یک شبکه بیزی گرافی بدون‌دور جهت‌دار است در جهت‌دهی به یال‌ها این نکته همواره مدنظر است و هر زمان که جهت یالی باعث تشکیل یک دور جهت‌دار شود به حالت مخالف جهت‌دهی خواهد شد.

نهایتا ممکن است تعدادی از یال‌ها بدون جهت باقی بمانند که در این صورت با یک گراف بدون دور جزئا جهت‌دار روبه‌رو هستیم که متناظر با آن یک خانواده از گراف‌های بدون دور جهت‌دار وجود دارد و باید یکی از ساختارها را به تصادف انتخاب کنیم.

الگوریتم اجداد و اولاد برای ساختن یک دگ دو مرحله زیر را طی می‌کند:

۱. **ساختن کالبد دگ**: الگوریتم ابتدا یک گراف کامل روی تمام متغیرها در نظر می‌گیرد. سپس برای هر دو متغیر، مانند X و Y، مشخص می‌کند که دارای وابستگی هستند یا خیر. این کار با استفاده از آزمون استقلال χ^2 به این صورت انجام می‌شود که ابتدا مجموعه $\mathcal{Z}$ را تهی در نظر گرفته و درست بودن رابطه‌ی $X \perp Y|\emptyset$ را بررسی می‌کند. اگر با p– مقدار در نظر گرفته شده برای آزمون χ^2، فرض صفر رد نشد یعنی دو رأس از هم مستقل بودند یال بین آن‌ها حذف می‌شود، در غیر این صورت یکی از همسایه های X را به مجموعه $\mathcal{Z}$ اضافه کرده و نتیجه را بررسی می‌کند. به همین ترتیب برای همه‌ی زیرمجموعه‌های تک عضوی از مجموعه‌ی رأس‌هایی که با X همسایه هستند، این کار را انجام می‌دهد. در هر مرحله که استقلال دو رأس تأیید شد الگوریتم متوقف و یال بین X و Y حذف می‌شود، اما اگر یال باقی بماند الگوریتم از آزمون‌های استقلال مرتبه دو استفاده می‌کند، یعنی با انتخاب $\mathcal{Z}$ از زیرمجموعه‌های دو عضوی از همسایه‌های X، شرط $X \perp Y \,|\mathcal{Z}$ را بررسی می‌کند. الگوریتم این کار را ادامه می‌دهد تا جایی که تعداد عناصر مجموعه‌ی $\mathcal{Z}$ برابر تعداد همسایه‌های X شود (در نظر داشته باشید که $\mathcal{Z}$ زیرمجموعه‌ای از همسایه‌های X است). اگر در مرحله‌ای رابطه‌ی $X \perp Y \,|\mathcal{Z}$ تأیید شود یال بین X و Y حذف خواهد شد در غیر این صورت یال مربوطه باقی می‌ماند. از آن‌جایی که در الگوریتم ترتیب در نظر گرفتن X و Y مهم است، فقط زمانی یال بین X و Y باقی می‌ماند که نه در بین همسایه‌های X و نه در بین همسایه‌های Y، مجموعه $\mathcal{Z}$ که X و Y را از هم جدا کند

وجود نداشته باشد. به این ترتیب تعدادی از یال‌ها حذف خواهد شد و کالبد دگ مشخص می‌شود (جنسن و نیلسن، ۲۰۰۷).

۲. **جهت‌دار کردن یال‌های ساختاری که در مرحله‌ی اول مشخص شده**: در این مرحله پس از مشخص شدن ساختار گراف، الگوریتم سه تایی W، X و Y را پیدا می‌کند با این ویژگی که در ساختاری که در مرحله‌ی قبل به دست آمده، دو رأس X و W همسایه باشند و دو رأس Y و W نیز همسایه باشند اما دو رأس X و Y همسایه نباشند. در این صورت اگر رأس W متعلق به مجموعه‌ی جداکننده $separator_{X,Y}$ نباشد جهت یال XW از X به W و جهت یال YW از Y به W تعیین می‌شود (دلیل این امر با توجه به این که مجموعه‌ی $separator_{X,Y}$ همه‌ی مسیرهای بین X و Y را بلوک می‌کند واضح است). زمانی که تمام ساختارهایی که به این شکل هستند جهت‌یابی شدند، الگوریتم یال‌های باقی‌مانده در ساختار دگ را با استفاده از دو اصل مهم ذکرشده جهت‌دهی می‌کند (جنسن و نیلسن، ۲۰۰۷).

۲.۱.۳ محدودیت‌های روش‌های محدودیت‌گرا

اگر تعداد نمونه کم باشد یا داده گم شده داشته باشیم، روش‌های محدودیت‌گرا با خطاهای آماری مواجه‌اند. این روش‌ها، در جهت‌دهی به بعضی از یال‌ها ناتوان هستند و روش‌های ناپایداری هستند به طوری که یک اشتباه کوچک اولیه می‌تواند کار را به جایی سوق دهد که نتیجه با گراف اصلی بسیار متفاوت باشد.

۲.۳ روش‌های امتیازگرا

در این روش‌ها با استفاده از یک متر، میزان تطابق شبکه‌ها را با اطلاعات موجود محاسبه می‌کنند و به دنبال شبکه‌ای هستند که بیشترین تطابق را با داده‌ها داشته باشد. ساختن دگ با استفاده از روش‌های امتیازگرا شامل دو مرحله می‌شود. در مرحله اول، یک روش گشت مشخص می‌شود که تمام ساختارهای دگ ممکن را بررسی کند. در مرحله دوم، یک متر مناسب در نظر گرفته می‌شود که میزان تطابق هر ساختار را با داده‌ها یا مجموعه اطلاعات ارزیابی کند. این دو مرحله را تا جایی ادامه می‌دهند که هیچ ساختار ممکنی دارای تطابق بیشتری نباشد (ابراهیمی، ۱۳۹۰).

۱.۲.۳ انواع مترها

مترها به دو دسته زیر تقسیم‌بندی می‌شوند.

مترهای وابسته به توزیع پیشین (متر بیزی): این مترها با در نظر گرفتن توزیع داده‌ها $P(D|G)$ و توزیع پیشین شبکه $P(G)$، توزیع احتمال پسین شبکه $P(G|D)$ را محاسبه می‌کنند و شبکه‌ای که مقدار احتمال پسین آن ماکسیمم باشد را به عنوان بهترین شبکه انتخاب می کنند. هدف پیدا کردن شبکه G است به طوری که $P(G|D) = \frac{P(G,D)}{P(D)}$ را ماکسیمم کند. چون $P(D)$ در همه‌ی شبکه‌ها یکسان است کافی است شبکه‌ای را پیدا کنیم که $P(G,D)$ را ماکسیمم کند. چون کار کردن با لگاریتم آسان‌تر است معمولا در مترها به جای $P(G,D)$، $\log(P(G,D))$ را محاسبه می‌کنیم. از جمله مترهای بیزی می‌توان به متر بیزی دیریکله (BD) و حالت‌های خاص آن، مترهای K_2، BD_e و BD_{eu} اشاره کرد (هکرمن[۲۲] و همکاران، ۱۹۹۵؛ هکرمن، ۱۹۹۶).

مترهای وابسته به مفاهیم نظریه اطلاعات (متر غیربیزی): این مترها میزان تطابق دگ با اطلاعات مجموعه داده‌ها را اندازه‌گیری می‌کنند و ساختاری که تطابق بیشتری با مجموعه داده‌ها داشته باشد به عنوان بهترین ساختار انتخاب می‌کنند. از جمله مترهای غیربیزی می‌توان به متر کوتاهترین توصیف[۲۳](MDL) که شامل مترهای معیار اطلاع بیزی[۲۴](BIC)، معیار اطلاع آکائیک[۲۵](AIC) و لگاریتم درستنمایی[۲۶](LL) است، اشاره کرد (ابراهیمی، ۱۳۹۰). توجه کنید که یک متر باید حداقل دارای این دو ویژگی باشد: تعادلی بین دقت ساختار و پیچیدگی آن برقرار کند و از نظر محاسباتی قابل حل باشد. یکی از مترهایی که دو ویژگی بیان

[۲۲] Heckerman

[۲۳] Minimum Description length

[۲۴] Bayesian Information Criterion

[۲۵] Akaike Information Criterion

[۲۶] Log-Likelihood

شده را دارد متر BIC است که شامل عبارتی برای اندازه‌گیری میزان برازندگی مدل به داده‌ها و عبارت دیگری برای اندازه‌گیری پیچیدگی مدل است.

۲.۲.۳ روش‌های گشت

به منظور انجام یادگیری ساختاری با روش‌های امتیازگرا، باید پس از در نظر گرفتن یک تابع رتبه‌بندی (متر)، ساختار شبکه بیزی دارای بیشترین امتیاز (رتبه) را بین مجموعه‌ی همه‌ی ساختارهای شبکه‌ای ممکن جستجو کنیم. یعنی، مسئله‌ی یادگیری ساختاری به یک مسئله‌ی گشت محدود می‌شود. چالش این مسئله آن است که تعداد کل ساختارهای ممکن با تعداد رأس‌ها رابطه‌ی نمایی دارد. اگر n تعداد رأس‌های یک شبکه بیزی باشد آنگاه تعداد کل ساختارهای ممکن از رابطه‌ی زیر به دست می‌آید:

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} \frac{n!}{i!(n-i)!} 2^{i(n-i)} f(n-1),$$

که در آن $f(0) = 1$ است. برای nهای بزرگ، در نظر گرفتن همه‌ی ساختارها غیرممکن است. بنابراین، محققین روش‌های گشت اکتشافی را در نظر می‌گیرند و مکررا با ایجاد تغییرات کوچک روی ساختار کنونی، مناسب‌ترین ساختار را در فضای ساختارهای ممکن جستجو می‌کنند. به دگ‌هایی که با یک تغییر در دگ کنونی به وجود می‌آیند همسایه‌های آن دگ می‌گویند (جنسن و نیلسن، ۲۰۰۷).

عملگرهایی که یک تغییر در دگ کنونی ایجاد می‌کنند به شرح زیر است:

۱. عملگر افزاینده‌ی یال: یک یال بین دو رأس ایجاد می‌کند.

۲. عملگر حذف‌کننده‌ی یال: یال بین دو رأس را حذف می‌کند.

۳. عملگر معکوس‌کننده‌ی یال: جهت یال بین دو رأس را برعکس می‌کند.

یک خصوصیت مهم این عملگرها این است که آن‌ها تنها یک تغییر موضعی در ساختار موجود می‌دهند. به عنوان مثال، اگر یک یال از رأس X_i به رأس X_j وارد شود یا یال بین آن‌ها حذف شود آنگاه تنها خانواده (مجموعه اجداد) رأس X_j تغییر می‌کند و اگر یال بین آن‌ها معکوس شود آنگاه خانواده‌های هر دو رأس X_i و X_j تغییر می‌کند. از این خصوصیت در مترهای تجزیه‌پذیر استفاده می‌شود. متر BIC مثالی از یک متر تجزیه‌پذیر برای مجموعه داده‌های کامل است. با استفاده از مترهای تجزیه‌پذیر می‌توان افزایش یا کاهش رتبه ساختاری که با تغییر دادن یک یال به دست آمده را نسبت به ساختار اولیه اندازه‌گیری کرد. به عنوان مثال، اگر یک یال از رأس X_i به رأس X_j وارد کنیم آنگاه تنها رتبه‌ی رأس X_j تغییر خواهد کرد یعنی برای اندازه‌گیری تغییر رتبه ساختار جدید نسبت به ساختار اولیه کافی است اختلاف مترهای رأس X_j را از رابطه زیر محاسبه کنیم:

$$\triangle(X_i \to X_j) = score(X_j, Pa(X_j) \cup \{X_i\}, D) - score(X_j, Pa(X_j), D)$$

اگر $\triangle(X_i \to X_j) > 0$ باشد آنگاه ساختار جدید رتبه بیشتری نسبت به ساختار اولیه دارد و در نتیجه بهتر به داده‌ها برازش داده می‌شود. به طور کلی، همه‌ی روش‌های گشت دو مرحله دارند:

مرحله شروع: در این مرحله یک گراف اولیه (گراف تهی، شبکه پیشین و ...) در نظر گرفته می‌شود.

مرحله جستجو: بر اساس متر انتخاب شده ، میزان تغییر تطابق ساختارهایی که به عنوان مثال با اضافه شدن، حذف شدن یا برعکس شدن یک یال به دست می‌آیند، محاسبه می‌شود. این فرایند تا جایی ادامه پیدا می‌کند که هیچ ساختار ممکنی دارای تطابق بیشتری نباشد (جنسن و نیلسن، ۲۰۰۷).

۳.۲.۳ الگوریتم گشت K_2:

یکی از روش‌های پایه‌ای و پرکاربرد گشت است. ورودی‌های این الگوریتم متغیرهای رتبه‌بندی شده و ماکسیمم تعداد اجدادی است که می‌خواهیم رأس‌ها داشته باشند. در این الگوریتم ابتدا یک ساختار تهی برای شبکه در نظر می‌گیریم سپس گره‌ها را طوری که گره ولد بیشترین امتیاز را داشته باشد و اجداد X_i در مجموعه‌ی $\{X_1, ..., X_{i-1}\}$ قرار بگیرند، وارد شبکه می‌کنیم. رتبه بندی می‌تواند بر اساس مترهای مختلف مانند K_2، BIC و $BDeu$ انجام شود. این الگوریتم به دنبال پیدا کردن مجموعه اجداد هر رأس به صورتی است که در ترتیب ایجاد شده برای هر رأس X_i، رأسی که بیشتر از بقیه رتبه شبکه را افزایش می‌دهد از مجموعه $\{X_1, ..., X_{i-1}\}$ به مجموعه اجداد X_i اضافه می‌کنیم. این کار را

تا جایی ادامه می‌دهیم که تعداد اعضای مجموعه اجداد X_i از عدد از پیش تعیین شده بیشتر نشود (جنسن و نیلسن، ۲۰۰۷).

٤ یادگیری پارامتری

یادگیری پارامتری به معنی برآورد پارامتر است. هدف از یادگیری پارامتری شبکه بیزی برآورد احتمال شرطی هر رأس شبکه به شرط اجدادش می‌باشد. پس از ساخت یک شبکه بیزی، لازم است که یک سری از مقادیر احتمال از مدل طراحی شده استخراج شود که به این فرآیند، استنتاج می‌گویند. انواع استنتاج عبارتند از استنتاج دقیق و استنتاج تقریبی. در استنتاج دقیق مجموعه داده کامل است و برآورد میزان احتمال شرطی هر رأس شبکه از روش‌های متداول برآوردیابی مانند روش ماکسیمم درستنمایی و در برخی از موارد روش بیزی دقیقا محاسبه می‌شود. در روش ماکسیمم درستنمایی برآورد پارامترها با ماکسیمم کردن چگالی احتمال توأم متغیرها یعنی بر اساس مجموعه داده به دست می‌آید در حالی که در روش بیزی برآورد پارامترها با ماکسیمم کردن چگالی احتمال توأم و چگالی احتمال پیشین پارامترها یعنی با ترکیب کردن اطلاع نتیجه شده از داده‌ها با دانش پیشین در مورد پارامترها به دست می‌آید. معیارهای متداول برای برآورد نقطه‌ای پارامترها در روش ML [۲۷] برآورد ML و در روش بیزی برآورد ماکسیمم احتمال پسین[۲۸](MAP) و میانگین احتمال پسین[۲۹](PM) هستند.

در استنتاج تقریبی مجموعه داده‌ها کامل نیست و برای برآورد پارامترها از روش‌های تقریبی مانند الگوریتم امید ریاضی- ماکسیمم سازی[۳۰] (EM) استفاده می‌شود. الگوریتم EM یکی از تکنیک‌های پر کاربرد در به دست آوردن برآورد نقطه‌ای[۳۱] پارامترهای یک توزیع از یک مجموعه داده‌ی ناکامل(دارای مقادیر گمشده) است. در هر تکرار الگوریتم EM دو گام وجود دارد. ابتدا گام E که مرحله امیدگیری از تابع درستنمایی است. سپس گام M که مرحله ماکسیمم‌سازی امید تابع درستنمایی است (فریدمن، ۱۹۹۷).

۱.٤ روش‌های یادگیری پارامتری

اگر مقادیر همه‌ی متغیرها از پیش معلوم باشد (داده گمشده نداشته باشیم) انجام چنین استنتاجی ساده است ولی معمولا فقط بخشی از متغیرها مشاهده می‌شود. بنابراین، فرآیند یادگیری پارامتری در شبکه‌های بیزی با مشخص بودن یا نبودن ساختار شبکه و همچنین قابل مشاهده بودن یا نبودن متغیرها روندهای متفاوتی را طی می‌کند. این روندها در جدول ۱ نشان داده شده‌اند (مورفی، ۲۰۰۱).

جدول ۱: روش‌های یادگیری پارامتری

ساختار	مشاهده‌پذیری	روش
مشخص	کامل	ماکسیمم درستنمایی و بیزی
مشخص	ناقص	الگوریتم EM
نامشخص	کامل	گشت در فضای ساختارهای ممکن + ماکسیمم درستنمایی و بیزی
نامشخص	ناقص	گشت در فضای ساختارهای ممکن + الگوریتم EM

- اگر ساختار شبکه معلوم باشد و تمامی متغیرها قابل مشاهده باشند می‌توان برآورد ML ، MAP و یا PM احتمال‌های شرطی هر رأس به شرط اجدادش را از روی داده‌های آموزشی و دانش پیشین در مورد پارامترها به دست آورد.

- درصورتی که ساختار شبکه از قبل معلوم باشد ولی فقط برخی از مقادیر متغیرها قابل مشاهده باشند، یادگیری مشکل‌تر خواهد بود. در این صورت از الگوریتم EM برای برآورد پارامترها استفاده می‌کنیم.

اگر ساختار شبکه معلوم نباشد یادگیری مشکل بوده و بسته به اینکه تمامی متغیرها و یا بخشی از آن‌ها قابل مشاهده باشند به صورت زیر عمل می‌کنیم.

- اگر ساختار شبکه معلوم نباشد اما تمامی متغیرها قابل مشاهده باشند، ابتدا با روش‌های گشت (مانند روش گشت

[۲۷] Maximum Likelihood

[۲۸] Maximum A Posteriori

[۲۹] Posterior Mean

[۳۰] Expectation-Maximization

[۳۱] Point Estimation

K_2) بهترین ساختار را می‌یابیم و سپس برآورد ML، MAP و یا PM پارامترها را به دست می‌آوریم.

- درصورتی که ساختار شبکه معلوم نباشد و فقط برخی از مقادیر متغیرها قابل مشاهده باشند، ابتدا با روش‌های گشت (مانند روش گشت K_2) بهترین ساختار را می‌یابیم و سپس از الگوریتم EM برای برآورد پارامترها استفاده می‌کنیم.

۲.٤ یادگیری پارامتری شبکه بیزی با روش ماکسیمم درستنمایی و روش بیزی

دگ $G=(V,E)$ را در نظر بگیرید که در آن $V=\{X_1,...,X_n\}$ مجموعه‌ای از n متغیر تصادفی و E مجموعه‌ای از یال‌های جهت‌دار درون فضای $V\times V$ است. فرض کنید متغیر X_i برای $i=1,...,n$ مقادیرش را از مجموعه‌ی $\mathcal{X}_i=\{x_i^{(1)},\dots,x_i^{(r_i)}\}$ بگیرد و مجموعه‌ی همه‌ی برآوردهای ممکن آزمایش به صورت زیر باشد:

$$\begin{aligned}\mathcal{X}&=\mathcal{X}_1\times\ldots\times\mathcal{X}_n\\&=\left\{(x_1^{(i_1)},\dots,x_n^{(i_n)})|i_j=1,\dots,r_j,\,j=1,\dots,n\right\}.\end{aligned}$$

مجموعه داده‌ها را به صورت یک نمونه‌ی mتایی از متغیرهای تصادفی $X_1,...,X_n$ به صورت زیر در نظر بگیرید:

$$x=\begin{pmatrix}\underline{x}_{(1)}\\ \vdots\\ \underline{x}_{(m)}\end{pmatrix},$$

که در آن $\underline{x}_{(l)}=(x_{l,1},\dots,x_{l,n})$ بیانگر l–امین نمونه است. فرض کنید مجموعه‌ی Π_i دارای q_i ترکیب از اجداد متغیر X_i به صورت $\Pi_i=\{\pi_i^{(q_1)},\dots,\pi_i^{(q_i)}\}$ باشد که در آن به این نکته $\pi_i^{(j)}$ اشاره می‌کند که j–امین ترکیب از Π_i مشاهده شده است.

متغیر برنولی n_{ijkl} را به صورت زیر تعریف می‌کنیم:

$$n_{ijkl}=\begin{cases}1, & \text{if } (x_i^{(k)},\pi_i^{(j)}) \text{ is found in } \underline{x}_{(l)}\\ 0, & otherwise\end{cases}$$

که در آن $(x_i^{(k)},\pi_i^{(j)})$ یک ترکیب از خانواده‌ی (X_i,Π_i) است. فرض کنید $\underline{\theta}$ مجموعه پارامترهای تعریف شده برای $i=1,\dots,n$، $k=1,\dots,r_i$ و $j=1,\dots,q_i$ به صورت زیر باشد:

$$\theta_{ijk}=P\left(X_i=x_i^{(k)}|\Pi_i=\pi_i^{(j)}\right),$$

که در آن $\sum_{k=1}^{r_i}\theta_{ijk}=1$ است.

تابع احتمال برای یک نمونه‌ی $\underline{x}_{(l)}$ از دگ $G=(V,E)$ به صورت زیر است:

$$p_{\underline{X}|\underline{\Theta}}(\underline{x}_{(l)}|\underline{\theta},E)=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\theta_{ijk}^{n_{ijkl}}.$$

بنابراین تابع احتمال توأم برای نمونه‌های مستقل $\underline{x}_{(1)},\dots,\underline{x}_{(m)}$ به صورت زیر خواهد بود:

$$p_{X|\underline{\Theta}}(x|\underline{\theta},G)=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\prod_{l=1}^{m}\theta_{ijk}^{n_{ijkl}}=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\theta_{ijk}^{n_{ijk}}$$

که در آن $n_{ijk}=\sum_{l=1}^{m}n_{ijkl}$ است. بنابراین برآورد ML پارامتر θ_{ijk} از رابطه‌ی زیر به دست می‌آید:

$$\delta_{ijk}^{ML}=\frac{n_{ijk}}{n_{ij}},\qquad(۱)$$

که در آن $n_{ij}=\sum_{k=1}^{r_i}n_{ijk}$ است (برای مشاهده‌ی جزئیات بیشتر کوسکی و نوبل، ۲۰۰۹ را ببینید).

به منظور به دست آوردن برآورد MAP و PM پارامتر θ_{ijk} برای $k=1,\dots,r_i$، توزیع پیشین مزدوج

$$(\theta_{ij1},\dots,\theta_{ijr_i})\sim Dir(\alpha_{ij1},\dots,\alpha_{ijr_i})$$

را با تابع چگالی احتمال زیر را در نظر بگیرید:

$$\pi(\theta_{ij1},\dots,\theta_{ijr_i})=\frac{\Gamma(\alpha_{ij})}{\prod_{k=1}^{r_i}\Gamma(\alpha_{ijk})}\prod_{k=1}^{r_i}\theta_{ijk}^{\alpha_{ijk}-1},$$

که در آن $0<\theta_{ijk}<1$، $\sum_{k=1}^{r_i}\theta_{ijk}=1$ و $\alpha_{ij}=\sum_{k=1}^{r_i}\alpha_{ijk}$، $\alpha_{ijk}>0$.

تابع احتمال پسین برای مجموعه داده‌ی $\underline{x}_{(1)},\dots,\underline{x}_{(m)}$ از رابطه‌ی زیر به دست می‌آید:

$$\begin{aligned}\pi(\theta_{ij1},\dots,\theta_{ijr_i}|x)&\propto\pi(\theta_{ij1},\dots,\theta_{ijr_i})p(x|\theta_{ij1},\dots,\theta_{ijr_i})\\&\propto\prod_{k=1}^{r_i}\theta_{ijk}^{n_{ijk}+\alpha_{ijk}-1},\end{aligned}$$

یعنی

$$(\theta_{ij1},...,\theta_{ijr_i})|x\sim Dir(n_{ij1}+\alpha_{ij1},...,n_{ijr_i}+\alpha_{ijr_i}).$$

به وضوح توزیع پسین حاشیه‌ای به صورت $\theta_{ijk}|x\sim Beta(n_{ijk}+\alpha_{ijk},n_{ij}+\alpha_{ij}-n_{ijk}-\alpha_{ijk})$ است.

بنابراین برآورد MAP برای θ_{ijk} عبارت از:

$$\delta_{ijk}^{\pi,MAP}(x)=\frac{n_{ijk}+\alpha_{ijk}-1}{n_{ij}+\alpha_{ij}-2}.\qquad(۲)$$

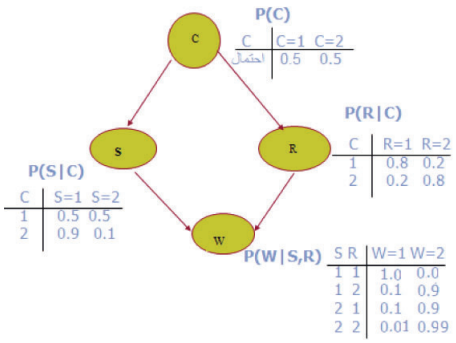
است و برآورد PM برای θ_{ijk} عبارت از:

$$\delta_{ijk}^{\pi,PM}(x)=\frac{n_{ijk}+\alpha_{ijk}}{n_{ij}+\alpha_{ij}}. \qquad (۳)$$

است.

۵ مدل‌بندی یک مجموعه داده شبیه‌سازی شده با شبکه بیزی

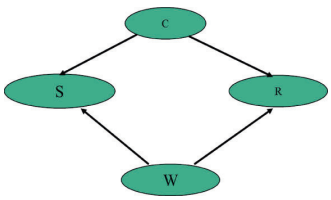
در این بخش یک مجموعه داده‌ی شبیه‌سازی شده را با شبکه بیزی مدل‌بندی می‌کنیم. برای این منظور ابتدا ساختار شبکه بیزی را با روش‌های محدودیت‌گرا و امتیازگرا به ترتیب با الگوریتم‌های یادگیری ساختاری PC و K_2 تعیین می‌کنیم و سپس به مسئله‌ی یادگیری پارامتری شبکه بیزی حاصل با روش ماکسیمم درستنمایی و روش بیزی می‌پردازیم. شبکه بیزی متشکل از ٤ رأس $C = Cloudy$، $S = Sprinkler$، $R = Rain$ و $W = Wetgrass$ شکل ۲ را در نظر بگیرید.



شکل ۲: شبکه بیزی از ٤ متغیر دو وضعیتی $C = Cloudy$، $S = Sprinkler$، $R = Rain$ و $W = Wetgrass$.

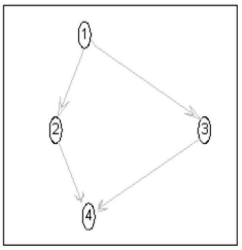
۱.۵ یادگیری ساختاری

ابتدا الگوریتم PC روش‌های محدودیت‌گرا را بر روی ۱۰۰ داده‌ی شبیه‌سازی شده از شبکه شکل ۲ اجرا می‌کنیم. خروجی الگوریتم به صورت شکل ۳ است.



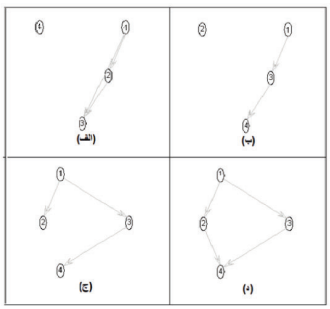
شکل ۳: دگ حاصل از اجرای الگوریتم PC روی داده‌های شبیه‌سازی شده به حجم ۱۰۰ از شبکه شکل ۲

از مقایسه‌ی این شبکه با شبکه واقعی (شکل ۲) ملاحظه می‌کنیم که این شبکه با شبکه واقعی اختلاف زیادی دارد. این نتیجه دور از انتظار نبود زیرا روش‌های محدودیت‌گرا در مواقعی که تعداد نمونه کم باشد خطاهای آماری دارند. از آن جایی که روش‌های امتیازگرا نسبت به روش‌های مبتنی بر قید جواب‌های دقیق‌تری دارند و مواقعی که تعداد نمونه‌ها کم باشد قابلیت اجرا دارند، به تعیین ساختار شبکه بیزی این مجموعه داده با الگوریتم K_2، با متر بیزی K_2 و متر غیر بیزی BIC از روش‌های امتیازگرا می‌پردازیم. به منظور اجرای الگوریتم K_2، متغیرها را (به دلخواه) به صورت $C = 1$، $S = 2$، $R = 3$ و $W = 4$ شماره‌گذاری کرده و داده‌ها را به صورت ۱۰ مجموعه داده به حجم‌های ۵، ۱۰، ۱۵و ... و٤٥ و ۵۰ در نظر می‌گیریم. سپس الگوریتم $K2$ را با در نظر گرفتن عدد ۲ به عنوان ماکسیمم تعداد اجداد هر رأس اجرا می‌کنیم. به منظور مقایسه شبکه واقعی (شکل ۲) با شبکه‌های بازسازی شده توسط الگوریتم K_2، متغیرها را به صورت $C = 1$، $S = 2$، $R = 3$ و $W = 4$ شماره‌گذاری کرده و ساختار شبکه شکل ۲ را با کدهای نرم افزار متلب، در شکل ٤ رسم کرده‌ایم.



شکل ٤: دگ متشکل از ٤ متغیر دو وضعیتی $W = 1$، $S = 2$، $R = 3$ و $C = 4$.

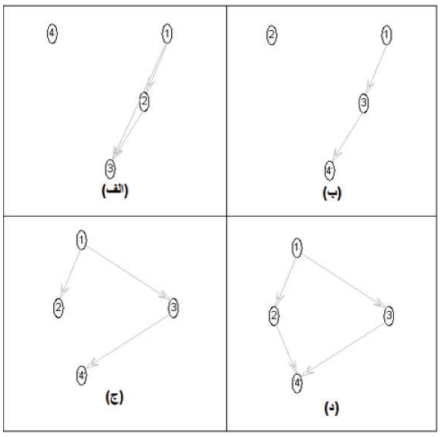
ابتدا الگوریتم K_2 را بر اساس متر بیزی K_2 اجرا می‌کنیم و مشخص می‌کنیم که شبکه حاصل از کدام حجم نمونه با شبکه اولیه یکسان است. خروجی الگوریتم به صورت شکل ۵ است.



شکل ۵: DAG حاصل از اجرای الگوریتم K_2 با متر بیزی K_2

روی داده‌های شبیه‌سازی شده از شبکه شکل ۲، به (الف) حجم ۵، (ب) حجم ۱۰، (ج) حجم‌های ۱۵، ۲۰ و ۲۵، (د) حجم‌های ۳۰، ۳۵، ۴۰، ۴۵ و ۵۰.

از مقایسه این شبکه‌ها با شبکه واقعی ملاحظه می‌کنیم که شبکه‌های حاصل از نمونه‌های به حجم‌های ۳۰، ۳۵، ۴۰، ۴۵ و ۵۰ با شبکه واقعی یکسانند. پس برای ساختن یک شبکه بیزی با الگوریتم K_2 با متر بیزی K_2، از چهار متغیر دو وضعیتی، کافی است یک نمونه به حجم ۳۰ از متغیرها داشته باشیم. از آن جایی که متر بیزی K_2 نا آگاهی‌بخش است یک بار دیگر الگوریتم K_2 را بر اساس متر غیر بیزی BIC اجرا می‌کنیم و مشخص می‌کنیم که شبکه حاصل از کدام حجم نمونه با شبکه اولیه یکسان است. خروجی الگوریتم به صورت شکل ٦ است.



شکل ٦: دگ حاصل از اجرای الگوریتم K_2 با متر غیر بیزی BIC روی داده‌های شبیه‌سازی شده از شبکه شکل ۲، به (الف) حجم ۵، (ب) حجم ۱۰، (ج) حجم‌های ۱۵، ۲۰ ، ۲۵ و ۳۵ (د) حجم‌های ۳۰، ۴۰، ۴۵ و ۵۰.

از مقایسه این شبکه‌ها با شبکه واقعی ملاحظه می‌کنیم که شبکه‌های حاصل از نمونه‌های به حجم‌های ۳۰، ۴۰، ۴۵ و ۵۰ با شبکه واقعی یکسانند. پس برای ساختن یک شبکه بیزی با الگوریتم K_2 با استفاده از متر غیر بیزی BIC، از چهار متغیر دو وضعیتی، کافی است یک نمونه به حجم ۴۰ از متغیرها داشته باشیم.

۲.۵ یادگیری پارامتری

اکنون که ساختار شبکه بیزی مشخص شد، به مسئله‌ی یادگیری پارامتری شبکه بیزی حاصل می‌پردازیم. رأس‌های شبکه را به صورت متغیرهای $C = X_1$، $S = X_2$، $R = X_3$ و $W = X_4$ در نظر بگیرید. همه‌ی متغیرها دو وضعیتی هستند و مقادیر ۱ و ۲ را می‌گیرند. بنابراین طبق نمادهای ذکر شده در بخش قبل، $n = 4$ و $r_1 = r_2 = r_3 = r_4 = 2$ است. با توجه به شکل ۲، $q_4 = 4$، $q_2 = q_3 = 2$ و $q_1 = 0$ است یعنی مجموعه‌ی جد برای متغیرها به صورت زیر است:

$$\Pi_1 = \{\}, \quad \Pi_2 = \{\pi_2^{(1)}, \pi_2^{(2)}\}$$
$$\Pi_3 = \{\pi_3^{(1)}, \pi_3^{(2)}\}, \quad \Pi_4 = \{\pi_4^{(1)}, \ldots, \pi_4^{(4)}\},$$

که در آن:

$$\pi_2^{(1)} = 1,\ \pi_2^{(2)} = 2, \pi_3^{(1)} = 1,\ \pi_3^{(2)} = 2,\ \pi_4^{(1)} = (1,1),$$
$$\pi_4^{(2)} = (1,2),\ \pi_4^{(3)} = (2,1),\ \pi_4^{(4)} = (2,2).$$

در این‌جا با در نظر گرفتن مشاهدات متغیر X_3 از $m = 100$ نمونه شبیه‌سازی شده، برآورد ML، MAP و PM پارامترهای θ_{311} و θ_{321} را از رابطه‌های (۱)، (۲) و (۳) به دست می‌آوریم. این برآوردها در جدول ۲ نشان داده شده‌اند.

جدول ۲: مقدار واقعی و برآورد ML، MAP و PM پارامترهای θ_{311} و θ_{321}.

پارامتر	مقدار واقعی	برآورد ML	برآورد MAP	برآورد PM
θ_{311}	۰/۸۰۰	۰/۸۱۱	۰/۸۱۸	۰/۸۰۷
θ_{312}	۰/۲۰۰	۰/۲۸٦	۰/۲۵۹	۰/۲٦۵

لازم به ذکر است که به منظور به دست آوردن برآوردهای MAP و PM این پارامترها، توزیع پیشین مزدوج

$$(\theta_{311}, \theta_{321}, \theta_{312}, \theta_{322}) \sim Beta(16, 4, 4, 16)$$

را بر اساس دانش پیشین خود طوری در نظر گرفته‌ایم که میانگین توزیع حاشیه‌ای هر پارامتر با مقدار واقعی پارامتر یکسان باشد. به منظور کاهش محاسبات، برآورد ML، MAP و PM پارامترهای θ_{312} و θ_{322} را به جای محاسبه‌ی مستقیم از رابطه‌های (۱)، (۲) و (۳)، به کمک رابطه‌های $\theta_{311} + \theta_{312} = 1$ و $\theta_{321} + \theta_{322} = 1$ محاسبه می‌کنیم. برآورد این پارامترها در جدول ۳ آمده است. بقیه‌ی پارامترهای شبکه نیز به طور مشابه برآورد می‌شوند.

جدول ۳: مقدار واقعی و برآورد ML، MAP و PM پارامترهای θ_{312} و θ_{322}.

پارامتر	مقدار واقعی	برآورد ML	برآورد MAP	برآورد PM
θ_{321}	۰/۲۰۰	۰/۱۸۹	۰/۱۸۲	۰/۱۹۳
θ_{322}	۰/۸۰۰	۰/۷۱٤	۰/۷٤۱	۰/۷۳۵

اکنون می‌خواهیم عملکرد برآوردهای ML، MAP و PM پارامتر θ_{311} را با آماره‌های میانگین (Mean) و انحراف استاندارد[۳۲](STD) برآوردها و معیار واگرایی کولبک-لایبلر (KL) با یکدیگر مقایسه کنیم. برای این منظور، مراحل زیر را در نظر می‌گیریم:

گام ۱. متغیرهای $(x_1,\ldots,x_4)$ را در $m = 25, 50, 100$ نمونه شبیه‌سازی می‌کنیم.

گام ۲. $d[1,k] = \delta_{31k}^{ML}$، $d[2,k] = \delta_{31k}^{\pi,MAP}$ و $d[3,k] = \delta_{31k}^{\pi,PM}$ را برای هر $k = 1, 2$، با در نظر گرفتن توزیع پیشین مزدوج $Beta(16, 4)$ برای پارامترهای $(\theta_{311}, \theta_{312})$ محاسبه می‌کنیم.

گام ۳. گام‌های ۱ و ۲ را $N = 10000$ بار تکرار می‌کنیم. سپس بر اساس داده‌های تولید شده، آماره‌های کمی میانگین و انحراف استاندارد و واگرایی کولبک-لایبلر هر برآورد را از رابطه‌های زیر محاسبه می‌کنیم:

$$\text{Mean } d[i,k] = \frac{1}{N}\sum_{r=1}^{N} d[i,k,r],$$

$$\text{STD } d[i,k] = \left(\frac{1}{N-1}\sum_{r=1}^{N}\left(d[i,k,r] - \text{Mean } d[i,1]\right)^2\right)^{\frac{1}{2}},$$

$$\text{KLD } d[i] = \frac{1}{N}\sum_{r=1}^{N}\left(\theta_{311}\log_2(\theta_{311}/d[i,1,r])\right.$$
$$\left. + \theta_{312}\log_2(\theta_{312}/d[i,2,r])\right)$$

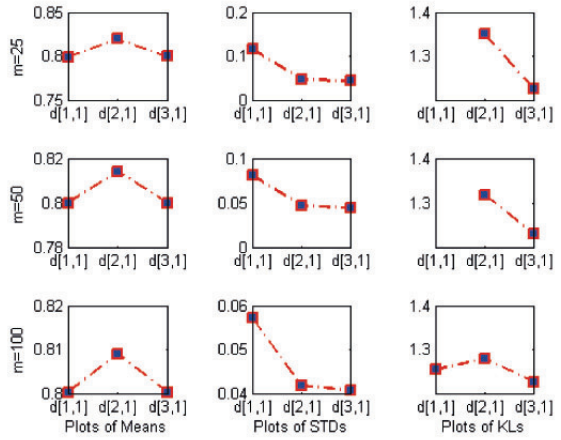
که در آن $i = 1,\ldots,3$، $k = 1, 2$ و $d[i,k,r]$ برای $r = 1,\ldots,N$، برآورد $d[i,k]$ در r-امین تکرار است.

آماره‌های کمی میانگین و انحراف استاندارد و معیار واگرایی KL برآوردهای پارامتر θ_{311} برای مقادیر متفاوت m، محاسبه شده و در جدول ٤ آورده شده است. علامت $*$ در این جدول بیانگر این است که معیار واگرایی KL قابل محاسبه نیست.

جدول ٤: آماره‌های کمی میانگین و انحراف استاندارد و معیار واگرایی KL برآوردهای پارامتر θ_{311} برای مقادیر متفاوت m.

	m	$d[1,1]$	$d[2,1]$	$d[3,1]$
	۲۵			
Mean		۰/۷۹۹۱	۰/۸۱۹۵	۰/۷۹۹۷
STD		۰/۱۱٦۳	۰/۰٤٦٤	۰/۰٤۳۵
KL		*	۱/۳۵۰۷	۱/۲۲٦۱
	۵۰			
Mean		۰/۸۰۰۰	۰/۸۱٤۱	۰/۸۰۰۰
STD		۰/۰۸۱۲	۰/۰٤٦۵	۰/۰٤٤٤
KL		*	۱/۳۱۷۱	۱/۲۲۹۵
	۱۰۰			
Mean		۰/۸۰۰۰	۰/۸۰۹۰	۰/۸۰۰۱
STD		۰/۰۵۷۱	۰/۰٤۱۸	۰/۰٤۰٦
KL		۱/۲۵٤۲	۱/۲۷۹۳	۱/۲۲۵۸

نمودارهای شکل ۷ نشان می‌دهند که عملکرد برآوردها با تغییر حجم نمونه به مقدار چشمگیری تغییر نمی‌کند. به عبارت دیگر، دقت برآوردها تحت تأثیر حجم نمونه نیست و حجم نمونه‌ی $m = 25$ نیز برای به دست آوردن برآورد دقیق پارامتر θ_{311} کافی است. به منظور انجام یک استنتاج استقرایی، نمودارهای میانگین $N = 10000$ تکرار برآوردهای پارامتر θ_{311} برای مقادیر متفاوت m، در شکل ۷ به تصویر کشیده شده است.



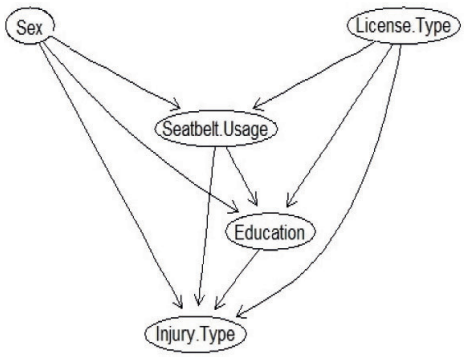
شکل ۷: نمودارهای میانگین برآوردهای ML، MAP و PM و معیار واگرایی KL پارامتر θ_{311} برای مقادیر متفاوت m.

از جدول ٤ و شکل ۷ مشاهده می‌کنیم که برآوردهای PM دقیق‌تر از MAP پارامتر θ_{311} را برآورد می‌کنند. این مطلب در مقایسه مقادیر انحراف استاندارد نیز به وضوح قابل مشاهده است. به علاوه معیار واگرایی KL برآوردهای PM خیلی کمتر از MAP است.

[۳۲] Standard Deviation

٦ پیاده‌سازی بر روی داده‌های واقعی

در این بخش به پیاده‌سازی روش‌های یادگیری ساختاری و پارامتری بر روی یک مجموعه داده واقعی تصادف می‌پردازیم. داده‌های تصادف شامل اطلاعات عمومی رانندگانی است که در شبکه جاده‌ای کشور دچار سانحه تصادف رانندگی شده‌اند. این داده‌ها مربوط به یک دوره زمانی ٥٦ ماهه از فروردین ۱۳۸۸ تا آذر ۱۳۹۲ است. در این دوره زمانی حدود ٦٥٥۰۰ تصادف جاده‌ای به ثبت رسیده است (کریم‌نژاد، ۱۳۹۳). متغیرهای مورد استفاده عبارت از جنسیت[۳۳] (مرد، زن)، میزان تحصیلات[۳٤] (بی‌سواد، زیردیپلم، دیپلم، فوق دیپلم، کارشناسی، کارشناسی ارشد و بالاتر)، استفاده یا عدم استفاده از کمربند ایمنی[۳٥]، نوع آسیب دیدگی[۳٦] (آسیب ندیده، جرحی، فوتی) و نوع گواهینامه[۳۷] (بدون گواهینامه، گواهینامه پایه دوم، گواهینامه پایه اول، گواهینامه مشروط، گواهینامه ویژه) هستند. با توجه به اینکه حجم داده‌ها در این مطالعه زیاد است، برای یادگیری ساختار شبکه تصادف از الگوریتم PC بهره می‌گیریم (همان‌گونه که قبلا بیان شد، وقتی حجم داده‌ها کم باشد، روش‌های محدودیت‌گرا با خطاهای آماری مواجه هستند اما اگر حجم داده‌ها به اندازه کافی بزرگ باشد، می‌توان از روش‌های محدودیت‌گرا در یادگیری ساختار شبکه بهره گرفت). شکل ۸ ساختار نهایی حاصل از اجرای الگوریتم PC را بر روی داده‌های تصادف نشان می‌دهد.



شکل ۸: ساختار به دست آمده با استفاده از الگوریتم PC

همان‌گونه که در شکل ۸ مشاهده می‌کنیم، گره‌های جنسیت، میزان تحصیلات، نوع گواهینامه و استفاده یا عدم استفاده از کمربند ایمنی، همگی جدهای گره نوع آسیب‌دیدگی هستند. حال فرض کنید یادگیری پارامتری در این شبکه مورد علاقه باشد و بخواهیم بدانیم که به طور مثال، احتمال عدم آسیب‌دیدگی راننده مردی با تحصیلات دیپلم و گواهینامه پایه دوم رانندگی در حالی که از کمربند ایمنی استفاده کرده است چقدر است. با توجه به ساختار شبکه تصادف به دست آمده در شکل ۸ و با استفاده از داده‌های واقعی در دسترس، برآورد ماکسیمم درستنمایی این میزان احتمال را محاسبه نموده‌ایم که مقدار آن ۰/۹۸۲۹۹ به دست آمده است. بدیهی است می‌توان برآورد ماکسیمم درستنمایی سایر پارامترهای این شبکه را نیز به دست آورد. به همین ترتیب در صورت وجود اطلاعات قابل استناد و صحیح، می‌توان توزیع‌های پیشین مناسبی را (مشابه آنچه در زیربخش ۲.۵ بیان شد) تعیین نمود و برآوردهای MAP و PM را نیز گزارش نمود.

۷ نتیجه‌گیری

در این مقاله یک مجموعه داده‌ی شبیه‌سازی شده را با شبکه بیزی مدل‌بندی کردیم. ابتدا ساختار شبکه را با الگوریتم یادگیری ساختاری PC از روش محدودیت‌گرا مشخص کردیم. از مقایسه‌ی شبکه حاصل از خروجی الگوریتم با شبکه واقعی ملاحظه کردیم که شبکه حاصل با شبکه واقعی اختلاف زیادی دارد. این نتیجه دور از انتظار نبود زیرا روش‌های محدودیت‌گرا در مواقعی که تعداد نمونه کم باشد با خطاهای آماری مواجه‌اند. از آن جایی که روش‌های امتیازگرا نسبت به روش‌های محدودیت‌گرا جواب‌های دقیق‌تری دارند و مواقعی که تعداد نمونه کم باشد قابلیت اجرا دارند، به تعیین ساختار شبکه بیزی این مجموعه داده‌ها با الگوریتم K_2، با متر بیزی K_2 و متر غیر بیزی BIC از روش‌های امتیازگرا پرداختیم. خروجی الگوریتم‌ها بیانگر این هستند که برای ساختن یک شبکه

[۳۳] Sex
[۳٤] Education
[۳٥] Seatbelt Usage
[۳٦] Injury Type
[۳۷] License Type

بیزی با الگوریتم K_2 از چهار متغیر دو وضعیتی، به ترتیب با متر بیزی K_2 و متر غیر بیزی BIC، کافی است نمونه‌های به حجم‌های ۳۰ و ۴۰ از متغیرها داشته باشیم. پس از مشخص شدن ساختار، به مسئله‌ی یادگیری پارامتری شبکه بیزی حاصل با روش ماکسیمم درستنمایی و روش بیزی پرداختیم. مشاهده کردیم که عملکرد برآوردها با تغییر حجم نمونه به مقدار چشمگیری تغییر نمی‌کند. به عبارت دیگر، دقت برآوردها تحت تأثیر حجم نمونه نیست و حجم نمونه‌ی $m = 25$ نیز برای به دست آوردن برآورد دقیق پارامترها کافی است. در آخر نشان دادیم که برآوردهای PM خیلی دقیق‌تر از MAP پارامترها را برآورد می‌کند و معیار واگرایی KL برآوردهای PM خیلی کمتر از MAP است. مقایسه‌ی عملکرد برآوردهای ML با MAP و PM به دلیل تفاوت آن‌ها در وابستگی‌شان به دانش پیشین، نادرست است. سپس به اجرای روش‌های یادگیری ساختاری و پارامتری بر روی داده‌های واقعی متشکل از اطلاعات عمومی رانندگانی که در شبکه جاده‌ای کشور دچار سانحه تصادف رانندگی شدند، پرداختیم.

مراجع

[۱] ابراهیمی، علی. تنظیم شبکه بیان ژن بر مبنای :شبکه بیزی؛ پایان‌نامه کارشناسی ارشد، دانشگاه شهید بهشتی، ۱۳۹۰.

[۲] کریم‌نژاد، علی. تحلیل داده‌های تصادف جاده‌ای با استفاده از مدل‌های آماری؛ طرح پژوهشی دفتر تحقیقات کاربردی پلیس راهور، تهران، ۱۳۹۳.

[3] Friedman, N, (1997), *Learning Belief Networks in the Presence of Missing Values and Hidden Variables*; ICML.

[4] Fridman, N, (2004), *Inferring cellular networks using probabilistic graphical models*, Science, **303**:799-805.

[5] Heckerman, D, (1996), *A Tutorial on Learning with Bayesian Networks*, Technical Report, Microsoft Research.

[6] Heckerman, D, Geiger, D and Chickering, D. M, (1995), *Learning BNs: the combination of knowledge and statistical data*, Machine Learning, Vol. 20, pp. 197–243.

[7] Jensen, F. V and Nielsen, T. D, (2007), *Bayesian Networks and Decision Graphs*; Second Edition, Springer Science +Business Media, LLC.

[8] Koski, T and Noble, J.M, (2009), *Bayesian Networks - An Introduction*; Wiley.

[9] Murphy, K and Saira, M, (2001), *Modelling Gene Expression Data Using Dynamic Bayesian Networks*; Computer Science Division, University of California, Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720.

[10] Ong, I. M, Glasner, J. D and Page, D, (2002), *Modelling regulatory pathways in E. Coli from time series expression profiles*, Bioinformatics, Vol. 18, pp. S241–S248.

[11] Pearl, J, (1988), *Probabilistic Reasoning in Intelligent Systems*, Pacific Symposium, On Biocomputing.