روشهای آماری برای داده‌کاوی

محسن محمدزاده[۱] افشین فلاح[۲]

چکیده

آمار عموماً با مسایلی در چالش است که علوم و صنعت فرا روی آن قرار می‌دهند. در گذشته این مسایل اغلب نتیجه آزمایش‌های صنعتی، کشاورزی و پزشکی بودند. که به مجموعه داده‌های محدود منجر می‌شدند. اما با پیشرفت رایانه‌ها و افزایش اطلاعات، روشهای آماری با داده‌هایی حجیم‌تر و پیچیده‌تر مواجه شده‌اند. از طرفی علومی مانند سنجش از دور[۳] و پردازش[۴] تصاویر پا به عرصه وجود گذاشته‌اند، که ذاتاً با مجموعه داده‌های بزرگ سرو کار دارند. به این ترتیب چالش‌های بوجود آمده منجر به وجود آمدن گسترده‌ای از علوم به نام داده‌کاوی شده است که بصورت گسترده از برخی روشهای آماری در جهت اهداف خود استفاده می‌کند. در این مقاله دیدگاه‌های مطرح در خصوص داده‌کاوی مورد بررسی قرار می‌گیرد و مراحل مختلف استخراج اطلاعات از مجموعه داده‌های بسیار بزرگ مطرح خواهند شد. سپس با عنایت به مسائلی که داده‌کاوی با آنها مواجه است، روش‌های آماری مناسب معرفی خواهند شد.

واژه‌های کلیدی: داده‌کاوی، پایگاه داده‌ها، انبارداده‌ها، میانگین‌گیری بیزی، شبیه‌سازی مونت کارلویی.

۱. مقدمه

پیشرفت شگفت انگیز فن‌آوری رایانه‌ای و مجهز شدن بشر به این ابزار امکان جمع آوری اطلاعات دقیق و کامل در زمینه‌های مختلف را فراهم ساخته است و منجر به پیدایش ساختارهای داده‌ی بسیار حجیم شده است. بنابراین دستیابی به اطلاعات نهفته در داده‌های حجیم که لازمه مدیریت مؤثر است، با بکاربردن سیستم‌های سنتی استفاده از پایگاه‌های داده‌ها، میسر نیست. شدت رقابتها در عرصه‌های علمی، اجتماعی، اقتصادی، سیاسی و حتی نظامی نیز اهمیّت عامل سرعت یا زمان دسترسی به اطلاعات را دو چندان کرده است. بنابراین نیاز به طراحی سیستم‌هایی که قادر به اکتشاف سریع اطلاعات مورد علاقه کاربران با تأکید بر حداقل مداخله انسانی باشند از یک طرف و روی آوردن به روش‌های آماری متناسب با حجم داده‌های زیاد از سوی دیگر به خوبی احساس می‌شود. داده‌کاوی[۵] فرآیندی است که در آغاز دهه ۹۰ پا به عرصه ظهورگذاشته و با نگرشی نو به مسئله استخراج اطلاعات از پایگاه‌های داده‌ها می‌پردازد. از سال ۱۹۹۵ داده‌کاوی بصورت جدی وارد مباحث آمار شد و در سال ۱۹۹۶ مجله اکتشاف دانش از پایگاههای داده ها اولین شماره خود را منتشر کرد. در همان سال گلایمور و همکاران[۵] موضوعات آماری مرتبط با داده‌کاوی را در مقاله‌ای مورد بررسی قرار دادند. هند [۷] نقش و اهمیت روشهای آماری در داده‌کاوی را مورد بررسی قرار داد و مادیگان [۱۰] برخی روشهای بیزی را برای کاربردهای داده‌کاوی مطرح کرد. علاوه بر مقالاتی که در این زمینه نوشته شده است، می توان به کتاب هند[۸] در زمینه مبانی داده‌کاوی اشاره کرد. در این مقاله مفاهیم پایه‌ای داده‌کاوی، دیدگاه‌های موجود در زمینه داده‌کاوی، ساختار کلی فرآیند و قسمتهای مختلف آن بطور خلاصه تشریح می‌شود و در نهایت داده‌کاوی از دیدگاه آماری و روشهای آماری مورد استفاده در داده‌کاوی مورد بحث قرار می‌گیرد.

داده‌ها نمایشی از واقعیّت‌ها، معلومات، مفاهیم، رویدادها یا پدیده‌هایی هستند که به گونه‌ای صوری و مناسب برای برقراری ارتباط، تفسیر یا پردازش توسط انسان یا ماشین مورد استفاده قرار می‌گیرند. اما اطلاع به معنی دانشی است که از طریق پردازش داده‌ها بدست می‌آید. به

۱و۲- گروه آمار، دانشگاه تربیت مدرس

۳- Remote Sensing

۴- Image Processing

۵- Data Mining

بیان دیگر اطلاع، حاصل تکامل داده‌ها است. وقتی انسان تصویری را مشاهده می‌کند، این تصویر به عنوان یک داده از طریق دستگاه ورودی یعنی چشم وارد سیستم می‌شود، سپس مغز تصویر دریافت شده را با تصاویر موجود در حافظه مقایسه می‌کند، این عمل سنجش و تفسیر داده‌ها است. چنانچه تصویر دریافتی با یکی از تصاویر موجود در حافظه مطابقت داشته باشد، اطلاع حاصل می‌شود. به این ترتیب بین داده‌ها و اطلاعات یک شکاف وجود دارد که اندازه این شکاف با حجم داده‌ها ارتباط مستقیم دارد. هر چه داده‌ها حجیم‌تر باشند، این شکاف بیشتر خواهد بود و هرچه حجم داده‌ها کمتر و روش‌ها و ابزار پردازش داده‌ها کاراتر باشند، فاصله بین داده و اطلاعات کمتر می‌شود. امروزه به دلیل افزایش حجم داده‌ها علی‌رغم پیشرفت‌های چشمگیر در فن آوری رایانه‌ای و فنون پردازش داده‌ها، این شکاف بسیار گسترده شده است. مفهوم داده‌کاوی در مطالعات و مقالات متعدد برای استخراج دانش از پایگاه‌های داده مورد استفاده قرار گرفته است. این نام هر چند بی مسمّی است، امّا بنا بر مصالحی مورد استفاده قرار می‌گیرد. بی مسمّی از این جهت که استخراج طلا از صخره ها یا سنگها معمولاً صخره‌کاوی یا سنگ‌کاوی نامیده نمی‌شود، بلکه طلاکاوی نامیده می‌شود. بنابراین در واقع به داده‌کاوی باید نام مناسب تر اکتشاف دانش از پایگاه داده‌ها را نسبت داد که متأسفانه کمی طولانی است. از طرفی دانش‌کاوی فاقد تأکید لازم بر کاوش در مقادیر بزرگ داده‌ها است، لذا داده‌کاوی که بر هر دو کلمه داده و کاوش تأکید دارد یک انتخاب عمومی است. گرچه عبارات معادل زیادی در این زمینه وجود دارد که از جمله می‌توان دانش کاوی در پایگاه داده‌ها، استخراج دانش و لایروبی داده‌ها[۱] را نام برد.

در نوشتگان این زمینه علمی، دو تعبیر مختلف از داده‌کاوی وجود دارد. برخی مؤلفین مانند چتفیلد [۲] داده‌کاوی را مترادف عبارت اکتشاف دانش[۲] از پایگاه‌های داده‌ها *(KDD)* می‌دانند. در این دیدگاه تأکید صرفاً بر کارایی و حجم بسیار زیاد پایگاه‌های داده‌ها است (جیاوی، [۹]) و موضوع اصلی مدیریت آنها است (ژو، [۱۵]). دیدگاه دوّم که به داده‌کاوی به عنوان یک مرحله ضروری از فرآیند اکتشاف دانش از پایگاه‌های داده می‌نگرد، یک دیدگاه آماری از داده‌کاوی است، که توسط فایاد [۳] و هند [۸] مطرح شده است. مؤلفه‌های اصلی این دو دید گاه در بخش‌های ۲ و۳ بطور مختصر تشریح می‌شوند.

۱- Data Dredging

۲- Knowledge Discovery

۲. اکتشاف دانش از پایگاه‌های داده‌ها

به طور کلی فرآیند اکتشاف دانش از پایگاه‌های داده‌ها زیرمجموعه‌ای از فن‌آوری اطلاعات است، که در آن مراحل پاکسازی و یکپارچه‌سازی داده‌ها دو مرحله از پیش پردازش داده‌ها محسوب می‌شوند و نتیجه در انبار داده‌ها ذخیره می‌گردد.

اما در فن‌آوری اطلاعات، علاوه بر استخراج اطلاعات از داده‌ها، بسترها و ابزارهای جمع‌آوری داده‌ها قبل از تشکیل پایگاه‌های داده‌ها و نیز نحوه استفاده و بکارگیری اطلاعات حاصل از داده‌ها مورد بحث قرار می‌گیرد. شکل ۱ مراحل فرآیند اکتشاف دانش از پایگاه‌های داده‌ها را نشان می‌دهد، که در آن قسمت (الف) مرحله تشکیل پایگاههای داده‌ها، قسمت (ب) مرحله پیش پردازش داده‌ها و قسمت (ج) ساختار سیستم داده‌کاوی را نشان می‌دهد.

۲ . ۱ پیش پردازش داده‌ها

کیفیّت داده‌ها در استخراج نتایج مطلوب و اطلاعات حقیقی بسیار مؤثر است. پایگاه‌های داده‌های حجیم مستعد شمول داده‌های مزاحم، گمشده و ناپایدار هستند. از اینرو برای ارتقاء کیفیّت داده‌ها، لازم است در ابتدای کار بصور زیر پیش پردازش شوند.

۱ – **پاکسازی داده‌ها**[۳]: برای انجام یک داده‌کاوی مؤثر لازم است مقادیر گمشده جایگزین شوند، داده‌های مزاحم شناسائی و به نحوی مناسب با آنها برخورد و ناپایداری‌ها اصلاح شوند.

الف – مقادیر گمشده : داده‌هایی هستند که به هر دلیلی در هنگام تحلیل داده‌ها در اختیار تحلیلگر قرار ندارند. وجود چنین داده‌هایی می‌تواند تحلیل داده‌ها را بسیار دشوار سازد. در صورت وجود مقادیر گمشده در داده‌ها باید به گونه‌ای مناسب در مورد آنها تصمیم‌گیری شود، به عنوان مثال ممکن است رکورد مربوطه حذف یا بجای آن یک مقدار ثابت، میانگین مقادیر صفت مورد نظر، میانگین نمونه‌های مشابه یا محتمل‌ترین مقدار جایگزین گردد.

ب – داده‌های مزاحم: انحرافی تصادفی یا غیرتصادفی در یک متغیر اندازه‌گیری شده هستند، که به عنوان مثال می‌توانند نتیجه خطای اندازه‌گیری یا یک اثر پنهان باشند و باید علّت وجود یک داده مزاحم به خوبی بررسی و در مورد آنها تصمیم‌گیری شود.

ج – داده‌های ناپایدار: اینگونه داده‌ها شامل تغییراتی بی‌قاعده هستند که تحلیل آنها را دچار مشکل می‌سازد. برخی از انواع آنها را می‌توان با تبدیل مناسب اصلاح کرد یا برای تحلیل آنها روش‌های خاصی را بکار گرفت.

۳- Data Cleaning

۲ – تلفیق داده‌ها[1] : از آنجا که داده‌ها از منابع مختلف جمع‌آوری می‌شوند، ممکن است دارای ناسازگاری‌هایی مانند تفاوت در مقیاس باشند یا صفت‌های مختلف به‌گونه‌ای با یکدیگر مرتبط باشند که برخی از آنها بر حسب تعدادی دیگر قابل حصول باشند. در اینگونه موارد لازم است داده‌ها بگونه‌ای یکپارچه شوند که حتی الامکان دارای کمترین تفاوت باشند و از ورود صفات مشابه یا تکراری در تحلیل داده‌ها پرهیز شود.

۳ – تبدیل داده‌ها: گاهی برای خلاصه‌سازی یا بکارگیری روش‌های تحلیل داده‌ها، لازم است به یکی از روش‌های زیر داده‌ها به شکلی مناسب تبدیل شوند.

- **هموارسازی:** برای حذف افت و خیز شدید در داده‌ها از تکنیک‌های هموارسازی مانند خوشه‌بندی و رگرسیون استفاده می‌شود.
- **انبوهش**[2]: نوعی خلاصه‌سازی است که با عملیات جبری روی برخی مقادیر و بدست آوردن مقادیر کلی‌تر اجرا می‌شود. به عنوان مثال مقادیر فروش روزانه می‌توانند بصورت مقادیر ماهانه یا سالانه انباشته شوند.
- **تعمیم داده‌ها:** با استفاده از سلسله مراتب مفاهیم، داده‌های ابتدایی یا سطح پایین مانند سن بوسیله مفاهیم سطح بالاتر مانند جوان، میانسال، بزرگسال جایگزین می‌شوند.
- **نرمالسازی داده‌ها:** داده‌ها به نحوی مقیاس بندی می‌شوند که در داخل یک محدوده مشخص و کوچکتر قرار گیرند.

۴- فروکاهی داده‌ها: از تکنیک‌هایی مانند فروکاهی[3] بعد، فشرده‌سازی و مجزاسازی داده‌ها برای فروکاهی داده‌ها استفاده می‌شود.

الف- فروکاهی بعد: مجموعه داده‌ها ممکن است شامل صفت‌های متعددی باشد که تنها تعدادی از آنها در طول کاوش مد نظر باشند. فروکاهی بعد، اندازه مجموعه داده‌ها را با حذف این قبیل صفت‌ها یا ابعاد فروکاهی می‌دهد. عمدتاً روش‌هایی برای انتخاب زیر مجموعه‌ای از صفت‌ها بکار می‌روند که توزیع احتمال داده‌های منتخب تا حد ممکن به توزیع احتمال داده‌های اصلی نزدیک باشد.

ب– فشرده سازی: استفاده از تبدیل‌ها یا انجام مراحلی برای فروکاهی داده‌ها است. اگر داده‌های اصلی را بتوان طی مراحلی از داده‌های فشرده، بدون از دست دادن اطلاعی بدست آورد، روش فشرده‌سازی بی‌زیان نامیده می‌شود و اگر تنها قادر به بدست آوردن تقریبی از داده‌های اصلی باشیم، روش همراه با زیان نامیده می‌شود. روش‌های بی‌زیان را تنها برای محدوده خاصی از داده‌ها می‌توان به کار برد. دو روش همراه با زیان عبارتند از :

– **تبدیلهای موجکی**[4]: وقتی این تبدیل برای بردار داده‌های V به کار می‌رود، آنرا به یک بردار 'V از ضرائب موجک تبدیل می‌کند که هر دو دارای طول یکسان هستند. مفید بودن این روش در فشرده‌سازی داده‌ها به این دلیل است که داده‌های تبدیل شده (ضرائب موجک) می توانند از جایی به بعد بریده شوند، به این معنی که می‌توان با ذخیره سازی همه ضرائب موجک بزرگتر از یک آستانه و حذف بقیه، تقریبی از داده‌ها را بدست آورد (جیاوی، [۹]).

– **کوفتن داده‌ها**[5]: یک مجموعه داده حجیم را به مجموعه‌ای بسیار کم حجم‌تر تبدیل می‌کند، به نحوی که نتایج حاصل از تحلیل آماری براساس مجموعه داده کوفته شده، شبیه نتایج حاصل از داده‌های اصلی است (مادیگان و همکاران، [۱۱]).

ج– مجزاسازی: از تکنیک‌های مجزاسازی برای فروکاهی گستردگی مقادیر یک صفت استفاده می‌شود. این کار با تقسیم محدوده صفت‌ها به فواصل مشخص و نسبت دادن یک برچسب به هر محدوده صورت پذیرد. به عنوان مثال سلسله مراتب مفاهیم برای یک صفت داده شده منجر به مجزا شدن سطوح مختلف آن صفت می‌شود.

۲.۲ ساختار یک سیستم داده‌کاوی

پس از پیش پردازی داده‌ها، نتایج در پایگاهی جدید یا انبار داده‌ها ذخیره می‌شوند. قسمت ب از شکل ۱، ساختار یک سیستم داده‌کاوی را نشان می‌دهد که ورودی آن مخازن بزرگ داده‌ها و خروجی آن اطلاعات مورد نیاز کاربر است و پایان کار اکتشاف دانش از پایگاه‌های داده‌ها تلقی می‌شود. در حالت کلی یک سیستم داده‌کاوی می‌تواند شامل مؤلفه‌های زیر باشد:

الف- پایگاه داده‌ها: برای ذخیره‌سازی داده‌های حجیم به کار می‌روند.

۱- Data Integration
۲- Aggregation
۳- Data Reduction
۴- Wavewlete
۵- Data Squashing

ب- خادم: مسئول بازیافت داده‌های مربوط به کار کاوش براساس اهداف داده‌کاوی است.

ج- پایگاه دانش: محدوده‌ای از دانش است که جهت هدایت فرآیند کاوش یا ارزیابی جذابیت الگوهای حاصل به کار می‌رود. این اطلاعات می‌تواند شامل دانستنی‌هایی مانند عقاید کاربر یا مفاهیم طبقه‌بندی شده جهت سازماندهی صفت‌ها، آستانه‌های تعیین جذابیّت و غیره باشد.

د- موتور داده‌کاوی: که بصورت گسترده از روش‌های آماری برای توصیف، پیوند دادن، رده‌بندی، تحلیل خوشه‌ای، پیشگویی و غیره استفاده می‌کند، هسته اصلی سیستم داده‌کاوی را تشکیل می‌دهد.

و- واحد ارزیابی الگو: با استفاده از معیارهای جذّابیّت الگو و آستانه‌های جذّابیّت و همچنین ارتباط متقابل با واحدهای دیگر داده‌کاوی، کاوش را به سمت تمرکز بر استخراج الگوهای جذّاب هدایت می‌کند.

ه- رابط گرافیکی کاربر: بین کاربران و سیستم داده‌کاوی یک رابطه گرافیکی برقرار می‌کند.

با توجّه به ساختار سیستم داده‌کاوی که در شکل ۲ نمایش داده شده است، برای انجام داده‌کاوی لازم است بخشهای مورد علاقه کاربر از پایگاه‌های داده، پس زمینه اطلاعاتی مفید برای ارزیابی الگوهای حاصل و نحوه ارائه اطلاعات استخراج شده، مشخص شوند.

(جای شکل)

تعیین این موارد شرایطی را فراهم می سازد تا کاربر در خلال کاوش با سیستم داده‌کاوی رابطه برقرار کند. برای تلفیق همه موارد فوق می‌توان یک زبان داده‌کاوی[1] (DMQ) طراحی کرد، تا سیستم انعطاف‌پذیری بیشتری داشته باشد.

۳. روشهای آماری کاوش

هند [۸] داده‌کاوی را بصورت خلاصه‌سازی و تحلیل داده‌های غالباً حجیم برای یافتن روابط مفید و قابل اعتماد تعریف کرده است. در این دیدگاه داده‌کاوی رابط بین آمار وعلم کامپیوتر است و از پیشرفتهای هر دو رشته در جهت استخراج اطلاعات از پایگاههای داده‌ها استفاده می‌کند. از طرف دیگر داده‌کاوی همانند آمار، اما با تأکیدهای متفاوت به آموختن از داده ها و یا تبدیل داده‌ها به اطلاعات، مربوط می‌شود. در این راستا با توجه به اینکه ماهیت داده‌ها در آمار با داده‌کاوی متفاوت است، داده‌کاوی به برخی از روش‌های آماری که دارای ویژگیهای خاصی می‌باشند توجه بیشتری نشان می‌دهد. پارزن [۱۲] برای آن دسته از روشهای آماری که مورد توجه داده‌کاوان قرار دارند، نام روش‌های کاوش آماری[2] را پیشنهاد نموده است. یکی از ویژگی‌های مورد توجه داده‌کاوی برای روش‌های آماری، سادگی تعبیر آنها است. از اینرو به استفاده از مدل‌های نسبتاً ساده و قابل تعبیر مانند نمودارها گرایش زیادی وجود دارد. در داده‌کاوی مواردی که در آنها با تعداد بسیار زیادی متغیر، مدل و یا فرضیه مواجه هستیم، فراوان است. از طرفی داده‌کاوی یک فرآیند اکتشافی و تکراری است به این معنی که در خلال تحلیل داده‌ها اطلاعات جدید کشف می‌شوند و فرضیّه های قبلی اصلاح و فرضیّه‌های جدید فرمول بندی می‌شوند و این کار ممکن است با داده‌های زیادی، بارها تکرار شود. لذا از دیدگاه آماری روش‌هایی با کارایی محاسباتی بالا، تحلیل های تقریبی و تحلیل‌های محاسباتی تقریبی، مورد توجه خاص داده‌کاوی هستند(گلایمور و همکاران [۵]).
از طرفی در دیدگاه آماری تأکید بردقت و صحّت یا بصورت کلی اعتبار روشها و نتایج استخراج شده از پایگاه‌های داده‌های بسیار بزرگ است (هند، ۲۰۰۱)، از اینرو مشخص کردن میزان عدم حتمیّت نتایج حاصل از روش‌های آماری بسیار مهم تلقی می‌شود. لذا آن دسته از روش‌های آماری که عدم حتمیّت را در نتایج خود لحاظ می‌کنند، جایگاه خاصی در داده‌کاوی دارند و روش‌های مختلف بیزی به این دلیل که علاوه بر عدم حتمیّت نمونه‌گیری، عدم حتمیّت حاکم بر پارامترها را نیز در نتایج خود منعکس می‌کنند، کاربرد زیادی دارند. با توجه به موارد مطروحه، مدل‌ها و الگوها، توابع امتیاز، روش‌های بهینه سازی و راهکارهای مدیریّت داده‌ها چهار مؤلفه اصلی الگوریتم‌های داده‌کاوی را تشکیل می‌دهند (هند، [۸]).
تأکید بیشتر داده‌کاوی بر بعضی روش‌های آماری، به معنی عدم استفاده از سایر روش‌های آماری نیست و در عمل از طیف گسترده‌ای از روش‌های آماری برای تحلیل داده‌ها استفاده می‌شود، که در اینجا به دو مسئله مهم داده‌کاوی و راه حل‌های آماری برای آنها اشاره می‌شود.

۲.۱ محاسبات دشوار

چون در داده‌کاوی حجم داده‌ها زیاد و مدلها غالباً پیچیده هستند، یکی از خصوصیّات داده‌کاوی، دشواری محاسبات آن است. به عنوان مثال از دیدگاه بیزی خروجی تحلیل های آماری اغلب بصورت امیدهایی مانند میانگین، واریانس یا کواریانس شرطی کمیّت مورد علاقه مانند

۱- Data Mining Query

۲- Statistical Methods Mining

$$E\left[h(\theta|x_1,\cdots,x_n)d\theta\right] = \int h(\theta).\pi(\theta|x_1,\cdots,x_n)d\theta \quad (۱)$$

هستند، که در آن $\pi(\theta|x_1,\cdots,x_n)$ توزیع پسین پارامترها به شرط مشاهدات است. اغلب محاسبه اینگونه امیدها و بدست آوردن یک شکل بسته برای آنها امکان پذیر نیست. بگونه‌ای که تحلیل گران را وادار می‌سازد از تحلیل دقیق مدل‌های بیزی سلسله مراتبی و محاسبات توابع درستنمایی پیچیده دوری کنند (مادیگان،[۱۰]). پیشرفت‌های چشمگیر اخیر در روشهای مونت کارلو تحلیل‌گران را از این محدودیّت‌های رها ساخته است. در این روش‌ها ابتدا نمونه $\theta_1,\cdots,\theta_M$ از توزیع پسین $\pi(\theta|x_1,\cdots,x_n)$ تولید می‌شود، سپس با استفاده از قانون اعداد بزرگ انتگرال (۱) بصورت

$$\int h(\theta).\pi(\theta|x_1,\cdots,x_n)d\theta = \lim_{M\to\infty}\frac{1}{M}\sum_{i=1}^{M}h(\theta_i) \quad (۲)$$

تقریب زده می‌شود. اما بدلیل پیچیدگی فرم توزیع پسین اغلب بدست آوردن نمونه‌ای از آن نیز بسیار دشوار است. دشواری مضاعف داده‌های حجیم نیز موجب می‌شود که روش انتگرال‌گیری مونت کارلویی در هر تکرار کندتر شود.

یکی از روش‌های عمومی مونت کارلو روش نمونه‌گیری از نقاط مهم[۱] است. تقریب مونت کارلویی (۲)، به توانائی نمونه‌گیری از توزیع $\pi(\theta|x_1,\cdots,x_n)$ بستگی دارد. اگر نمونه‌گیری از این توزیع بسادگی مقدور نباشد ولی برای توزیع دیگری مانند $g(\circ)$ این امکان فراهم باشد، می‌توان تکنیک نمونه‌گیری از نقاط مهم را بکار برد. رابطه (۱) را می توان بصورت

(۳)

$$E\left[h(\theta|x_1,\cdots,x_n)d\theta\right] = \int h(\theta)\frac{\pi(\theta|x_1,\cdots,x_n)}{g(\theta)}g(\theta)d\theta = \lim_{M\to\infty}\frac{1}{M}\sum_{i=1}^{M}\omega_i h(\theta_i)$$

نوشت، که در آن θ_i یک استخراج از $g(\theta)$ و $\omega_i = \dfrac{\pi(\theta_i,\cdots,\theta_n)}{g(\theta_i)}$ است. طبیعتاً برای مفید بودن این روش باید نمونه‌گیری از توزیع $g(\theta)$ ساده باشد.

گرچه نمونه‌گیری از نقاط مهم روش مفیدی برای محاسبات است. اما برای مدل‌های پیچیده که در داده‌کاوی زیاد مطرح هستند، یافتن توزیع نمونه‌گیری معقول $g(\theta)$ می‌تواند بسیار دشوار باشد (گیلکس و همکاران، [۴]). در عمل تا اواخر دهه ۱۹۸۰ هیچ روشی برای اینگونه مسائل وجود نداشته، اما از آن به بعد روش‌های دیگری به نام روش‌های مونت کارلوی زنجیر مارکوفی[۲] (MCMC) کاربرد آمار بیزی را متحول ساختند. در این روش برخلاف نمونه‌گیری از نقاط مهم که نمونه‌هایی مستقل تولید می‌شوند، براساس یک زنجیر مارکوف نمونه وابسته $\theta_1,\cdots,\theta_M$ از توزیع مانای $\pi(\theta|x_1,\cdots,x_n)$ تولید می‌شود. ساختن چنین زنجیر مارکوفی اغلب کار دشواری نیست و همین امر یکی از دلایل برتری آن بر روش نمونه‌گیری از نقاط مهم است. الگوریتم متروپولیس- هاستینگ، الگوریتمی عمومی برای روش MCMC است، که ریجوی و همکاران [۱۴] با ارائه یک روش دنباله‌ای استفاده از برخی صافی‌ها کارایی آنرا برای مسائل داده‌کاوی افزایش داده‌اند.

از سوی دیگر معمولاً نظریه آمار برای توصیف پسین‌ها در نمونه های بزرگ، نتایج مجانبی ارائه می‌دهد که در کاربردها بسیار راه گشا هستند، اما حتی در پایگاه‌های داده‌های بسیار بزرگ ممکن است تعداد حالات مربوط به یک سئوال خاص بسیار کم باشد و نظریه آمار نتواند راه حلی را ارائه نماید. در اینگونه موارد خانواده‌ای از روشهای شبیه‌سازی به نام نمونه‌گیری گیبز که حالت خاص از الگوریتم متروپولیس– هاستینگ است، با مکانیسم‌های آماری سازگار شده‌اند و امکان محاسبه تقریبی بسیاری از توزیع‌ها را فراهم ساخته‌اند. به طور کلی با وجود اینکه روش‌های MCMC بسیاری از مشکلات محاسباتی را مرتفع می‌سازند، اما همواره دارای کارایی خوبی نیستند و ممکن است هنگام کار با مجموعه داده‌های حجیم به کندی عمل کنند. در این حالات می‌توان از تحلیل‌های تقریبی که اخیرا توسط مادیگان و همکاران [۱۱] توسعه زیادی یافته‌اند، استفاده کرد.

۲.۳ مدلسازی

یکی از مسائل مهم در داده‌کاوی مدل‌سازی است که معمولا در عمل با دنباله‌ای از مدلها مواجه هستیم. با فرض اینکه مجموعه مدلهای $\{M_k; k\in I\}$ بر اساس مشاهدات $D=\{x_1,\cdots,x_n\}$ حاصل شده باشد، روش‌های مختلفی برای انتخاب یک مدل از بین مدل‌های موجود وجود دارد که هر کدام داری نقاط ضعف و قوتی هستند. به عنوان مثال اگر هدف انتخاب مدل صحیح از بین مدل‌های موجود باشد، تصمیم بهینه می تواند انتخاب مدل با احتمال پسین ماکزیمم باشد (برناردو و اسمیت، [۱]). احتمال پسین را می توان بصورت

۱- Importance Sampling

۲- Marcov Chain Monte Carlo

$$P(M_i|D) \propto P(D|M_i).P(M_i) \qquad (۴)$$

نوشت که در آن محاسبه جمله درستنمایی حاشیه‌ای[۱] $P(D|M_i)$ عموماً حتی در حالات عادی نیز کار دشواری است و اغلب هیچ اطمینانی مبنی بر صحت یکی از مدل‌های خاص وجود ندارد. علاوه بر آن ارائه تعریفی از بهترین مدل به سادگی مقدور نمی‌باشد. بعلاوه انتخاب یک مدل به معنی بی ارزش بودن سایر مدل‌ها نیست و ممکن است رقبای بسیار خوبی برای مدل انتخاب شده وجود داشته باشد. در داده‌کاوی که معمولاً تعداد مدل‌های مطرح برای یک مجموعه داده بسیار زیاد است، مشکلات مطروحه بیش از پیش خود را نشان می‌دهند. در داده‌کاوی عموماً تعدادی مدل رقیب برای داده‌ها وجود دارد که ممکن است همه آنها معقول به نظر برسند، اما دارای نتایج متفاوتی باشند. راهکاری که در اینگونه موارد برای مدلسازی مطرح می‌باشد، میانگین گیری بیزی از مدلها است. این روش بدون اینکه مدل خاص را به عنوان بهترین مدل انتخاب کند با عنایت به عدم حتمیّت هر یک از مدل‌ها مدل جدیدی را بر اساس تمام مدل‌های موجود ارائه می‌کند. به عنوان مثال اگر Δ کمیّتی مورد علاقه باشد، این روش با استفاده از مقدار

$$P(\Delta|D) = \sum_k P(\Delta|D, M_k).P(M_k|D) \qquad (۵)$$

که در آن $P(\Delta|D, M_k)$ توزیع پیشگو[۲]است، به استنباط در مورد Δ می‌پردازد. با استفاده از معیارهای مختلف می‌توان نشان داد نتایج حاصل از این روش از تمام مدل‌های مطرح بهتر است (رفتری و همکاران، [۱۳]).

هر یک از روش‌های مطروحه نیازمند معیاری برای بررسی میزان برازش مدل نهایی به داده‌هاست. معمولاً مانده‌های $r_i = y_i - E(Y_i|D)$ که در آن $\{y_1,\dots,y_M\}$ نمونه‌ای مستقل از D است، نقطه شروع بسیار خوبی برای سنجش برازش هستند. علاوه بر آن عرض خم[۳] پیشگوی شرطی $P(Y_i|y(i))$ که در آن $y(i)$ نشان دهنده بردار مشاهدات بدون $y(i)$ است، نیز معیار مفیدی در این زمینه می‌باشد و برحسب مورد می‌توان یکی از این دو ابزار را برای ارزیابی میزان برازش مدل به داده‌ها بکار برد.

از دو روش مطروحه برای مدلسازی، روش اول که یکی از مدل‌ها را بعنوان مدل مطلوب برمی‌گزیند، روش سنتی موسوم به انتخاب مدل است که به برخی از اشکالات و نارسایی‌های آن اشاره شد. اما روش میانگین‌گیری از مدل‌ها روش جدیدی است که در سال‌های اخیر به دلیل کارایی بالای آن و رفع نارسایی‌های روش انتخاب مدل، بصورت گسترده مورد توجه و بررسی قرار گرفته است (رفتری و همکاران [۱۳]).

٤. بحث و نتیجه‌گیری

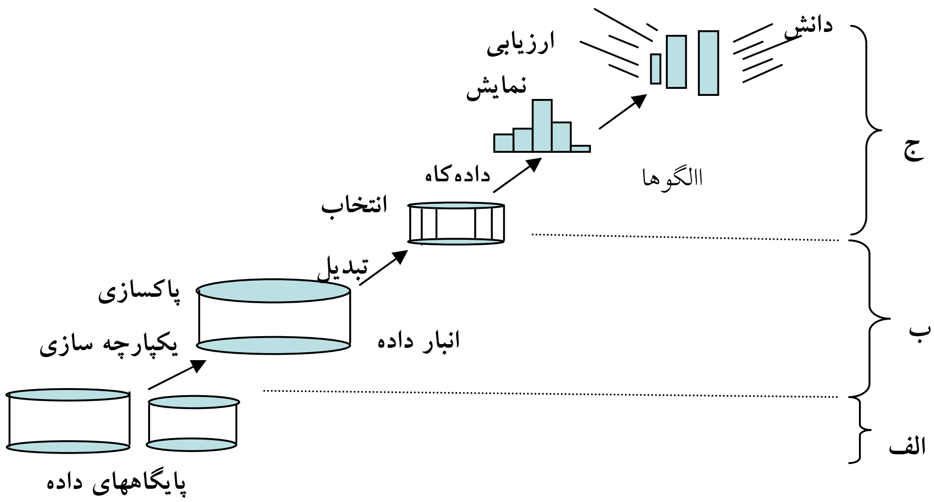
فرآیند اکتشاف دانش از پایگاه‌های داده‌ها و داده‌کاوی به چگونگی اکتشاف دانش از پایگاه‌های داده‌ها می‌پردازند. در این زمینه دو دیدگاه مطرح وجود دارد. دیدگاه پایگاه داده‌ها، داده‌کاوی را معادل فرآیند اکتشاف دانش از پایگاه‌های داده‌ها می‌داند و در آن تأکید بر نحوه مدیریت پایگاه‌های داده‌های بسیار حجیم است. اما در دیدگاه آماری داده‌کاوی جزئی از فرآیند اکتشاف دانش از پایگاه‌های داده‌های است و همانند آمار به بحث در مورد نحوه استخراج اطلاعات از داده‌ها می‌پردازد. در این دیدگاه با توجه به نوع داده‌ها و کاربرد و اهداف مورد نظر، برخی از روشهای آماری از جمله روش‌های دارای تعبیر ساده مانند نمودارها، روش‌های مختلف مدل‌سازی مانند انتخاب مدل و میانگین‌گیری از مدلها و روشهای تحلیل داده‌های حجیم مانند فنون مختلف شبیه‌سازی، نسبت به سایر موضوعات بیشتر مورد استفاده قرار می‌گیرند. اما دایره روش‌های آماری مورد استفاده در داده‌کاوی بسیار گسترده‌تر از این بوده و هر آنچه را که مربوط به آموختن از داده‌های با تبدیل داده‌ها به اطلاعات می‌شود.

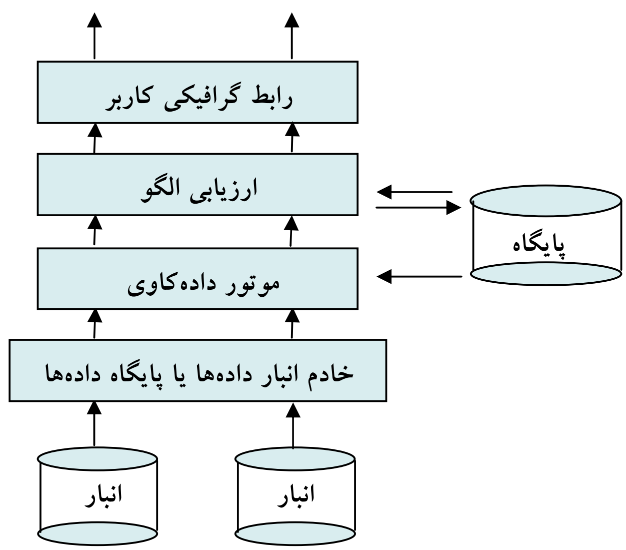
۱- Marginal Likelihood

۲- Predictive Distribution

۳- Ordinate

شکل ۱ : فرآیند اکتشاف دانش از پایگاه‌های داده‌ها





شکل ۲ : یک سیستم داده‌کاوی

مراجع

[1] Bernardo, J. M. and Smith, A. F. M., 1994, *Bayesian Theory*, John Wiley , Chichester.

[2] Chatfield, G., 1995, Model Uncertainty, Data Mining and Statistical Inference, *Journal of the Royal Statistical Society,* A, 58, 3, 419-466.

[3] Fayyad, U., Piatesky-Shapiro, G., Smyth, P., 1996, From Data Mining to Knowlredge Discovery in Data Bases, *American Association for Artifical Inteligence*, Fall, 37-53.

[4] Gilks, D., Richardson, S. and Spiegelhalter, D., 1996, *Marckov Chain Monte Carlo in Practice*, Chapman and Hall, London.

[5] Glaymour, C., Madigan, D., Fregtbon, D. and Smayth, P., 1996, Statistical Themes and Lessons for Data Mining, *Data Mining and Knowledge Discovery*, 1, 25-42.

[6] Hall, P., 1998, *Solving Data Mining Problems through Pattern Recognition*, New Jersey.

[7] Hand, D. J., 1998, Data Mining Statistics and More, *The American Statistician*, 52, 2.

[8] Hand, D. J., 2001, Principels of Data Mining, MIT Press, Cambridge.

[9] Jiawei, H., 2001, Data Mining: Concepts and Techniques, Taylor & Francis, New York,.

[10] Madigan, D., 2000, Statistical Methods for Bayesian Data Mining, http://www.stat. rutgers.edu/madigan/papers.

[11] Madigan, D., Rachavan, N., Nason, M., and Rideway, G., 2002, Likelihood-Based Data Squashing: A Modeling Approch to Instance Construction, *Data Mining and Knowledge Discolvery*, 6, 173-190.

[12] Parzen, E., 2001, Data Mining, Statistical Methods Mining and History of Statistics. http://stat.tamu.edu/ftp/pub/eparzen/future.pdf.

[13] Raftery, A., Madigan, D. and Hoeting, J., 1997, Bayesian Model Averaging for Linear Regression Models, *Journal of American Statistical Association*, 92, 179-191.

[14] Ridgewaye, G., and Madigan, D., 2002, A Sequential Monte Carlo Method for Bayesian Analysis of Massive Datasets, *Data Mining and Knowledge Discovery*, 6, 130-139.

[15] Zhou, Z. H., 2003, Three Perspectives of Data Mining, *Artificial Intelligence*, 143, 139-146.