داده‌کاوی، راهی به‌سوی ناشناخته‌ها

سیدامیررضا نجات*، آرش علی‌اکبری**

چکیده

در عصر حاضر که به عصر اطلاعات شهرت یافته است بدون در اختیار داشتن اطلاعات کافی، جامع و به هنگام، تصمیم‌گیری برای مدیران ممکن نخواهد بود؛ از طرفی با گسترش روز افزون علم و تکنولوژی و روش‌های جمع آوری داده‌ها با حجم عظیمی از داده‌ها، مواجه شده‌ایم. استفاده بهینه از این پایگاه‌های داده در صورتی امکان‌پذیر خواهد بود که از ابزارهای کارآ و استاندارد در کنار یک برنامه اصولی و آینده‌نگر برای تبدیل داده به اطلاعات کمک بجوییم. بنابراین امروزه با فقدان یا کمبود اطلاعات مواجه نیستیم بلکه آنچه از اهمیت به سزایی برخوردار است استفاده از روش‌های مناسب و استاندارد جهت نگهداری، به روز کردن، در دسترس قرار دادن و نهایتاً کشف دانش‌های جدید از انبوه اطلاعات موجود است.

داده‌کاوی یک روش کارآمد جهت استخراج دانش و اطلاعات از حجم انبوه داده‌های موجود است. روش‌های داده‌کاوی به طور گسترده‌ای توسط متخصصین و برنامه‌ریزان بکار گرفته شده است.

در این مقاله سعی داریم ضمن معرفی داده‌کاوی به نحوه ارتباط آن با علوم مختلف و وظایف آن و در نهایت به بررسی کاربردهای آن در صنایع مختلف تولیدی، خدماتی و به خصوص کاربردهای متصور برای ناجا بپردازیم.

کلید واژه‌ها

داده‌کاوی، هوش مصنوعی، روش‌های تحلیل آماری

---

*. کارشناس ارشد آمار ریاضی ، مرکز آمار معاونت طرح و برنامه و بودجه ناجا

**. کارشناس ارشد مهندسی سیستمهای اقتصادی و اجتماعی

مقدمه

در ابتدای قرن نوزدهم برخی از کشورها وارد مرحله‌ای گردیدند که بعدها عصر صنعتی نام گرفت. در آن دوران نیروی کار و مدیریت بر همه امور تسلط داشت و تولید اطلاعات رو به افزایش نهاد به گونه‌ای که با تلاش بیشتر، تبدیل اطلاعات به دانش امکان پذیر گردید و کارکنان مجبور گردیدند علاوه بر استفاده از نیروی جسمانی خود از فکرشان نیز برای انجام کارها بهره گیرند.

وقتی بشر مفهوم تکنولوژی و صنعت را دریافت اطلاعات قابل تولید به حدی رسید که دیگر نمی‌شد آن را در ذهن یک فرد به تنهایی نگهداری کرد و این روند با شتاب بیشتری به مرحله‌ای رسید که اکنون در آن قرار گرفته‌ایم؛ دورانی که به عصر اطلاعات شهرت دارد، دورانی است که حاکمیت به کامپیوتر، تکنولوژی ارتباطات و متخصصان تعلق دارد و به جای تلاش فیزیکی محض بر لزوم استفاده از قدرت فکر تاکید می‌شود. در عصر صنعتی **"سرمایه"** منبع راهبردی به شمار می‌آمد در حالی‌که در عصر اطلاعات **"دانش[1]"** منبع راهبردی می‌باشد. در عصر اطلاعات ارزش افزوده از طریق تبدیل اطلاعات به دانش و همچنین سرعت انتقال آن حاصل می‌شود.

در عصر کنونی همزمان با پیشرفت فن‌آوری اطلاعات[2]، حجم داده‌ها نیز به طور چشمگیری افزایش یافته است و به دنبال آن تعداد پایگاه داده‌ها[3] نیز از رشد قابل توجهی برخوردار بوده است به گونه‌ای که بسیاری از پایگاه داده‌ها شامل چندصد میلیارد رکورد ثبت شده می‌باشند که تحلیل و استخراج و پردازش اطلاعات آنها با روش‌های معمول آماری مستلزم صرف زمان و هزینه بسیار بالایی است که اغلب کمترین توجیه اقتصادی برای آن یافت نمی‌شود. اگرچه وجود سیستم‌های یکپارچه اطلاعاتی و یا سیستم‌های یکپارچه بانکی از لحاظ سیستمی به سازمان‌های استفاده کننده از این اطلاعات کمک نموده تا ساختار خود را به شکل سازمان یافته‌تری اصلاح نمایند اما از طرف دیگر باعث گردیده تا به حجم

---

1. Knowledge
2. Information Technology
3. Database

داده‌ها در پایگاه داده‌های مربوط، لحظه به لحظه اضافه گردد و این موضوع سبب گردید تا تحلیل‌گران با حجم داده‌هایی در حد گیگابایت و یا ترابایت مواجه گردند.

بی تردید بسیاری از این داده‌ها بی‌کیفیت، پرخطا و یا متناقص می‌باشند و بدیهی است که استفاده از داده‌های بی‌کیفیت همیشه منجر به نتایج بی‌کیفیت‌تر می‌گردد. از سوی دیگر مدیران برای تصمیم‌گیری‌ها، سیاست گذاری‌ها و پیش‌بینی‌های[1] خود نیاز به تحلیل داده‌های موجود دارند و بر مبنای آنچه که تشریح گردید نیاز به الگویی سودمند جهت استفاده بهینه از اطلاعات، همواره یکی از دغدغه‌های مدیران عصر حاضر می‌باشد.

در این مقاله، ابتدا به بررسی پیشینه و مفهوم داده‌کاوی و مراحل آن اشاره شده است. سپس به ساختار کلی داده‌کاوی و نحوه ارتباط آن با علوم مختلف از جمله آمار و هوش مصنوعی و مقایسه آنها پرداخته شده، در ادامه وظایف آن به عنوان یک ابزار مورد تشریح قرار گرفته و در نهایت به معرفی کاربردهای آن در صنایع مختلف تولیدی، خدماتی و به خصوص کاربردهای متصور در ناجا پرداخته شده است.

داده‌کاوی[2]

پیشینه داده کاوی

داده کاوی یکی از روش‌های کارآمد در کشف اطلاعات مفید از میان انبوه عظیمی اطلاعات تعریف شده است، که با استخراج الگوها و روابط بین داده‌ها ارزش‌های پنهانی آنها را آشکار می‌نماید. با کشف این ارزش‌ها می‌توان ارزش متغیرهای دیگر را پیش بینی نمود و در امر تصمیم‌گیری از آنها بهره جست. *(Hand,1998, P.10)*

داده‌کاوی را بایستی به عنوان علمی جدید و نوین نگریست. در مورد سابقه تاریخی داده‌کاوی شاید بتوان اوول (۱۹۸۳) را اولین شخصی دانست که مطالعه‌ای تحت عنوان *"شبیه‌سازی فعالیت‌های داده‌کاوی"* ارائه نمود. پژوهش‌های جدی‌تر روی موضوع داده‌کاوی از اوایل دهه ۹۰ آغاز گردید به گونه‌ای که آهنگ علاقه پژوهشگران به این موضوع در اواخر دهه ۹۰ با اوایل همان دهه قابل مقایسه نبود. داده‌کاوی تا امروز به عنوان موضوعی

---

1. Forecast
2. Data mining

جذاب، با شاخه‌های گوناگون برای محقیقین و علاقـه منـدان در حـوزه‌هـای مختلـف علـوم همانند آمار[1]، هوش مصنوعی[2]، نرم افزار، صـنایع و ... شـناخته مـی‌شـود. (سـعیدی، ۱۳۸۴، ص۲)

مفهوم داده کاوی

عبارت داده‌کاوی مترادف بـا عبـارات کـاوش در داده‌هـا، اسـتخراج[3] دانـش، برداشـت اطلاعات، لایروبی کردن اطلاعات و کشف دانش در پایگاه داده‌ها می‌باشد.

کشف دانش در پایگاه داده فرایند شناسایی درست، ساده و مفید و نهایتاً کشف الگوهـا و مدل‌های قابل فهم در داده‌ها می‌باشد به عبارت دیگر داده‌کاوی به فرایند استخراج اطلاعات ناشناخته، درست و مفید از انبوهی از داده‌ها اطلاق می‌گردد. در تمامی مـوارد حجـم داده‌هـا بسیار زیاد می‌باشد و کشف دانش نهفته در میان انبوهی از داده‌هـا هـدف غـایی داده‌کـاوی می‌باشد. به بیان ساده‌تر داده‌کاوی به معنای استخراج یا معدن کاری دانـش از میـان مقـدار زیادی داده خام است .یک نگاه کلی که نهایتاً به کشف ساختارهای جالب توجه، غیر منتظـره و با ارزش از داخل این مجموعه وسیع از داده‌ها می‌انجامد که شناسایی این روابـط، الگوهـا و روندهای جدید و معنی دار به دست آمده از داده‌کاوی به فرایند تصمیم‌گیری و طـرح یـک مدل به منظور پیش بینی و تصمیم‌گیری کمک می‌نماید.

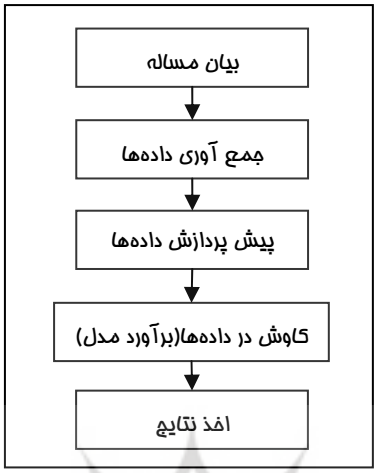
در حالت کلی می‌تـوان گفـت داده‌کـاوی عبـارت اسـت از " فراینـد جسـتجو و کشـف مدل‌های گوناگون، مختصرسازی[4] و اخذ مقـادیر مفیـد و بـا ارزش از مجموعـه‌ای وسـیع از داده‌ها."

مراحل داده‌کاوی

در یک نگاه کلی و اجمالی به فرایند داده کاوی، می‌توان نحوه انجام آن را از ابتدا تا انتهـا به صورت نمودار شماره (۱) در نظر گرفت:

---

1. Statistics
2. Intelligent
3. Mining
4. Reduction



نمودار شماره (۱): فرایند انجام داده کاوی

در اکثر موارد، الگوهای داده‌کاوی بایستی در امر تصمیم‌گیری موثر باشند و چنین الگوهایی برای مفید بودن، لازم است قابل تفسیر و قابل فهم باشند؛ لذا بایستی دقت داشت تا حد امکان الگوها ساده باشند زیرا آنها تنها ابزاری برای تفکر هستند. الگوها هیچ‌گاه جای تفکر را نمی‌گیرند بلکه کمک می‌کنند تا تفکر دقیق تر، عمیق‌تر و قوی‌تر صورت پذیرد. بایستی به این نکته اذعان داشت که "سیستم‌های پیچیده لزوماً نیازی به الگوهای پیچیده ندارند." مراحل فوق را در یک تقسیم بندی جزئی‌تر می‌توان به صورت زیر بیان نمود:

۱- شناسایی هدف:

در این مرحله مشخص می‌گردد کاربر چه چیزی را می‌خواهد و تا چه سطحی از اطلاعات را در نظر دارد که از پایگاه داده‌ها اخذ نماید.

۲-انتخاب داده‌ها[1]:

در این مرحله بایستی داده‌ها بر مبنای معیارهای مشخص انتخاب گردند.

۳-آماده سازی داده‌ها:

فرمت قابل استفاده داده‌ها و شناسایی متغیرهای زائد از اهداف این مرحله می‌باشد.

---

1. Data Selection

۴-ارزیابی داده‌ها:

معیارهایی از قبیل نوع توزیع داده‌ها، ویژگی و ساختار پایگاه داده‌ها، شرایط کلی داده‌ها و... چارچوب این بخش را تعیین می‌کنند.

۵-قالب بندی پاسخ:

ارائه فرمت پاسخ به شکل تصویر، نمودار، شبکه عصبی و ... خروجی این بخش است.

۶- انتخاب ابزار:

در این مرحله ابزارهای مناسب برای داده‌کاوی انتخاب می‌گردد و انطباق آن با کامپیوتر بررسی می‌شود.

۷-الگوسازی:

فرایند داده‌کاوی به صورت اصلی از این مرحله آغاز می‌گردد. این بخش شامل جستجوی الگوها در یک مجموعه داده‌ها، طبقه‌بندی، ارزشیابی داده‌ها و ... است. مواردی از قبیل صحت الگو، خطاهای الگو، توسعه الگو و ... در این مرحله صورت می‌پذیرد.

۸- اعتبار سازی یافته‌ها:

این مرحله شامل آزمون کردن الگوها می‌باشد. در تجزیه و تحلیل داده‌کاوی بایستی در مورد نتایج تجزیه و تحلیل با مدیر، مجری طرح و یا تحلیل‌گر بحث گردد.

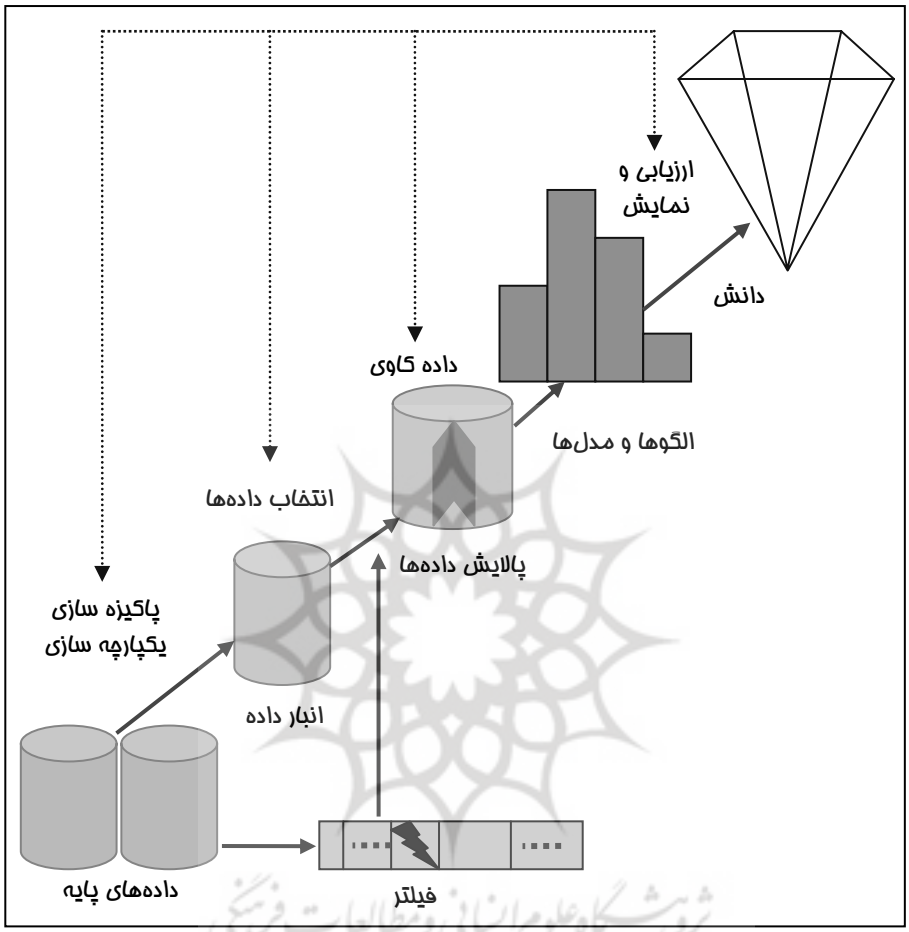
۹- ارائه نتایج:

در این بخش گزارش نهایی برای کاربر تهیه می‌شود، این گزارش بایستی با استناد به کل فرایند داده‌کاوی باشد.

۱۰-استفاده از نتایج:

هدف غایی داده‌کاوی استفاده از نتایج کشف شده جهت تصمیم گیری، سیاست گذاری و پیش بینی به منظور ایجاد یک موقعیت جدید و بهتر می‌باشد.

نمودار شماره (۲) مراحل مختلفی را که با کمک روش‌های داده کاوی، داده‌های پایه طی فرایندی به دانش تبدیل می‌شود به تصویر می‌کشد:



نمودار شماره (۲): مراحل کشف دانش با کمک داده کاوی

ساختمان داده‌کاوی

در بسیاری از موارد پایه و اساس داده‌کاوی به دو مقوله آمار و هوش مصنوعی تقسیم شده است که روش‌های مصنوعی به عنوان روش‌های یادگیری ماشین در نظر گرفته می‌شوند. فرق اساسی میان روش‌های آماری و روش‌های یادگیری ماشین بر اساس فرض‌ها و یا طبیعت داده‌هایی که پردازش می‌شوند می‌باشد. معمولا فرض بر این است که در

تکنیک‌های آماری توزیع داده‌ها مشخص است و در اکثر موارد توزیع نرمال است و در نهایت درستی و یا نادرستی نتایج نهایی به درست بودن فرض اولیه وابسته است حال آنکه در مقابل، روش‌های یادگیری ماشین از هیچ فرضی در مورد داده‌ها استفاده نمی‌کنند. روش‌های آماری در داده‌کاوی بیشتر زمانی استفاده می‌گردند که تعداد داده‌ها خیلی زیاد نباشد و اطلاعات بیشتری راجع به داده‌ها بتوان به دست آورد. این روش‌ها ابزاری برای کشف ارتباطات میان داده‌ها به کار می‌روند و بر خلاف روش‌های شبکه عصبی که فرایند نسبتاً مبهمی دارند، این روش‌ها کاملاً شفاف و واضح می‌باشند به علاوه دقت نتیجه‌گیری‌ها و تعبیر خروجی‌ها در این روش‌ها بهتر می‌باشد. به طور کلی روش‌های آماری زمانی که تفسیر داده‌ها توسط روش‌های دیگر مشکل به نظر می‌رسد، ابزاری کاملاً مفید می‌باشند. جدول شماره (۱) تفاوت روش‌های آماری را با روش‌های هوش مصنوعی نشان می‌دهد:

جدول شماره (۱): مقایسه روش‌های آماری با سایر روش‌های داده کاوی

روش‌های آماری	سایر روش‌های داده‌کاوی
داشتن فرضیه اولیه	بدون فرضیه اولیه
مورد استفاده در محدوده کوچکی از داده‌ها	کاربرد در محدوده وسیع‌تری از داده‌ها
استفاده بیشتر برای داده‌های کمی	کاربرد در انواع مختلفی از داده‌ها
استفاده از روابط ریاضی	استفاده از روش‌های یادگیری و هوش مصنوعی

آمار و داده‌کاوی

آمار شاخه‌ای از ریاضیات است که به جمع‌آوری، توضیح و تفسیر داده‌ها می‌پردازد. آمار در زندگی روزمره کاربردهای فراوانی دارد و در مقایسه با داده‌کاوی از قدمت بسیار بالایی برخوردار است لیکن می‌توان از تکنیک‌های آماری به عنوان روش‌های کلاسیکی از داده‌کاوی نام برد. از مهمترین تحلیل‌های آماری که به شدت در داده‌کاوی مورد استفاده قرار می‌گیرد می‌توان به موارد زیر اشاره نمود:

رگرسیون[1]:

یکی از هدف‌های اصلی بسیاری از پژوهش‌های آماری ایجاد وابستگی‌هایی است تا پیش‌بینی یک یا چند متغیر را بر حسب سایرین میسر گرداند. یک ابزار مفید در اینجا

1. Regression

استفاده از رگرسیون می‌باشد. یک معادله رگرسیون خطی n متغیره به شکل زیر بیان می‌گردد:

$$Y_i = \alpha + \beta_1 X_{1i} + ... + \beta_n X_{ni} + \varepsilon_i$$

که در آن Y_i متغیر وابسته و $X_1,...,X_n$ متغیرهای مستقل می‌باشند. پس از برآورد پارامترهای $\alpha,\beta_1,...,\beta_n$ می‌توان از معادله فوق جهت پیش‌بینی متغیر وابسته بر اساس مقادیر دلخواه متغیر مستقل اقدام نمود.

رگرسیون لجستیک[1] (آماد):

هنگامی که متغیر وابسته یک متغیر دودویی باشد از مدل رگرسیون لجستیک (آماد) استفاده می‌گردد. در حقیقت رگرسیون آماد احتمال وقوع متغیر وابسته را به ازای مقادیر مشخصی از مقادیر متغیر مستقل پیش‌بینی می‌نماید. این مدل مبتنی بر لگاریتم طبیعی نسبت فردها می‌باشد. در حالت کلی بیان یک معادله رگرسیون آماد به صورت زیر می‌باشد:

$$Ln(odds\ ratio) = \alpha + \beta_1 X_{1i} + ... + \beta_n X_{ni} + \varepsilon_i$$

پس از پردازش بر روی داده‌ها، نسبت فردهای برآورد شده به دست می‌آید و بعد هم احتمال وقوع متغیر وابسته به شکل زیر محاسبه می‌گردد:

$$\text{مقدار متغیر وابسته= برآورد احتمال موفقیت} = \frac{\text{نسبت فردهای برآورد شده}}{\text{نسبت فردهای برآورد شده+۱}}$$

آنالیز واریانس[2]:

با استفاده از آنالیز واریانس می‌توان به سوالات کلیدی از این دست پاسخ داد که آیا می‌توان تفاوت‌های مشاهده شده میان میانگین‌های دسته‌های نمونه‌ای مشابه را معلول تصادف دانست؟ یا این‌که میان میانگین‌های جامعه‌های مورد نمونه‌گیری تفاوت‌های واقعی وجود دارند؟

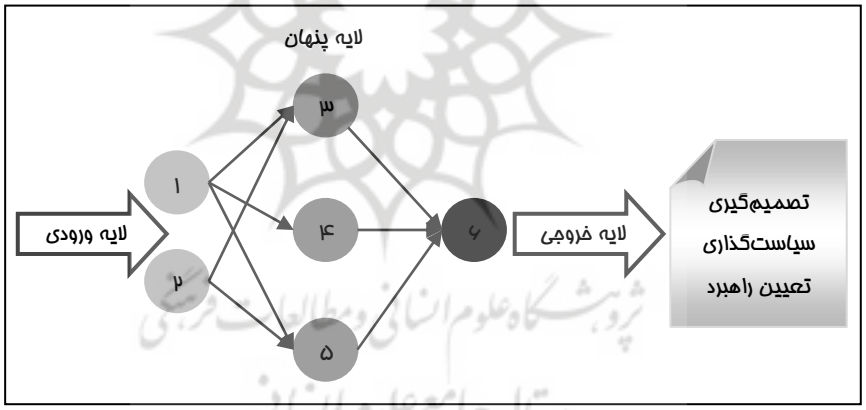
---

1. Logistic Regression
2. Analysis of Variance (ANOVA)

به طور مثال ممکن است علاقه‌مند باشیم تا بدانیم آیا میان موثر بودن سه شیوه برخورد متفاوت با مجرمین اختلاف معنی‌داری وجود دارد یا نه؟ چون اختلاف‌هایی که مشاهده می‌شوند می‌توانند همواره معلول عواملی به جز عوامل مشخصی باشند در مورد مثال فوق ممکن است اختلاف در وضعیت مجرمین پس از اعمال شیوه‌های مربوط ناشی از تفاوت در میزان تحصیلات، وضعیت اقتصادی خانواده و ... باشد. مباحث کلیدی‌تر در این زمینه را می‌توان در بحث طراحی آزمایش‌ها[1] مورد بررسی قرار داد.

هوش مصنوعی[2] و داده‌کاوی

شبکه‌های عصبی از پرکاربردترین و عملی‌ترین روش‌های الگوسازی مسائل پیچیده و بزرگ که شامل صدها متغیر هستند می‌باشد. شبکه‌های عصبی می‌توانند برای مسایل کلاس‌بندی و یا مسایل رگرسیون استفاده گردند. هر شبکه عصبی شامل یک لایه ورودی، یک لایه پنهان و یک لایه خروجی است.



نمودار شماره (۳): ساختار شبکه عصبی

تعداد گره‌ها و تعداد لایه‌های پنهان و نحوه وصل شدن گره‌ها به یکدیگر معماری شبکه عصبی را مشخص می‌کند. هر یال که دو گره را به یکدیگر متصل می‌نماید دارای یک وزن می‌باشد که اوزان معمولاً ناشناخته هستند که توسط تابع آموزش و داده‌های آموزشی که به

1. Design of Experiment
2. Artificial Intelligence

سیستم داده می‌شود تعیین می‌گردند. کاربر یا نرم افزاری که شبکه عصبی را طراحی می‌کند باید تعداد گره‌ها، تعداد لایه‌های پنهان، تابع فعال سازی و محدودیت‌های مربوط به وزن یال‌ها را مشخص نماید.

در حقیقت شبکه عصبی یک الگوی محاسباتی انتزاعی از مغز انسان است که همانند آن حاوی نورون‌های مصنوعی (یا واحدهای پردازش) و روابط و اتصالات می‌باشد. استفاده از شبکه‌های عصبی به دلایل زیر مفید و سودمند می‌باشد:

- **غیر خطی بودن:** یک نورون به عنوان یک واحد پایه می‌تواند یک عنصر پردازش خطی یا غیر خطی باشد، اما کل شبکه عصبی کاملاً غیر خطی می‌باشد. این ویژگی برای الگوهای هوش مصنوعی که ذاتاً مکانیسم‌های دنیای واقعی غیر خطی مسئول تولید داده یادگیری برای آن هستند، بسیار مهم می‌باشد.
- **یادگیری از مثال‌ها:** یک شبکه عصبی معیارهای درون ارتباطی‌اش را با به‌کارگیری مجموعه‌ای از نمونه‌های آموزشی و یادگیری اصلاح می‌نماید و اثرات نهایی یک فرایند یادگیری، پارامترهای یک شبکه را تنظیم می‌کند.
- **انطباق:** شبکه‌های مصنوعی یک قابلیت درون ساختی برای انطباق معیارهای درون ارتباطی جهت تغییر در محیط احاطه شده دارند. به بیان ساده‌تر با تغییر شرایط محیطی یک شبکه مصنوعی قابلیت تطبیق خود با شرایط جدید را پیدا می‌نماید.
- **پاسخ‌های مستند:** یک شبکه عصبی علاوه بر تهیه اطلاعات درباره گروهی خاص می‌تواند به منظور تصمیم‌گیری و تصمیم سازی طراحی گردد.
- **تحمل خرابی:** یک شبکه عصبی ذاتاً توان مقاومت در برابر خرابی را دارا بوده و عملکرد آن تحت شرایط ویژه از قبیل عدم ارتباط نورون‌ها و اختلال و داده‌های نامعلوم کاهش نمی‌یابد.
- **یکنواختی تجزیه تحلیل و طراحی:** اصولاً از شبکه‌های عصبی به عنوان پردازشگرهای اطلاعات استفاده می‌کنند. شبکه‌های عصبی این توان را دارا هستند تا اطلاعات را تجزیه و تحلیل نموده و به عنوان خروجی و در جهت پیش‌بینی از آنها استفاده نمایند.

وظایف داده‌کاوی

داده‌کاوی به طور معمول وظایف زیر را بر عهده دارد:

الف) توضیح و تفسیر[1]

خروجی‌داده‌کاوی به هر زبانی که بیان گردد بایستی قابل تفسیر باشد تا مدیران بتوانند از آن جهت تصمیم‌گیری‌های سازمان متبوعشان استفاده نمایند. فارغ از این‌که از تحلیل‌های آماری استفاده شده است یا از تکنیک‌های هوش‌مصنوعی، داده‌کاوی بایستی توسط تحلیلگران، به زبانی قابل فهم برای مدیران ترجمه گردد. در داده‌کاوی این امکان وجود دارد تا این عمل به بهترین شکل انجام پذیرد لذا می‌توان توضیح و تفسیر نتایج را از مهمترین وظایف داده‌کاوی به‌شمار آورد.

ب) تخمین[2]

در تخمین به دنبال تعیین مقدار یک مشخصه مجهولی می‌باشیم. در روش‌های آماری استفاده در این مورد به طور کلی شامل تخمین نقطه و یا فواصل اطمینان می‌باشد. دلیل استفاده از تخمین، وجود پارامترهای مجهول در جامعه می‌باشد در این موارد از یک آماره جهت برآورد کردن پارامترهای مجهول استفاده می‌گردد.

ج) پیش‌بینی[3]

در پیش‌بینی مقدار یک متغیر بر اساس متغیرهای دیگر برآورد می‌گردد. ممکن است عوامل متعددی در وقوع یک حادثه نقش داشته باشند در اینجا وظیفه داده‌کاوی تعیین سهم هر یک از این متغیرها و پیش‌بینی متغیر پاسخ به ازای مقادیر مختلف متغیرهای مستقل می‌باشد. رگرسیون یک ابزار سودمند و کارآ در این زمینه است.

---

1. Description
2. Estimation
3. Forecasting

د) وابسته سازی و ایجاد رابطه[1]

از وظایف داده‌کاوی کشف روابط پنهان میان انبوه داده‌ها است. با توجه به حجم عظیم داده‌های تولید شده، کشف دانش جدید از طریق آشکار سازی ارتباطات نهفته که از طریق روش‌های معمول آماری امکان‌پذیر نمی‌باشد از کلیدی‌ترین کارکردهای داده‌کاوی به شمار می‌رود.

و) کلاس بندی[2] (طبقه‌بندی)

توسط روش‌های کلاس‌بندی داده‌ها بر اساس ویژگی‌های خاص خود طبقه‌بندی می‌گردند. از این الگو می‌توان برای فهم داده‌های موجود و پیش‌بینی نحوه رفتار آنها استفاده نمود. گاهی الگوی پیش‌بینی کننده این روش‌ها به صورت استقرایی می‌باشد. در کلاس بندی، دسته‌ها یا کلاس‌ها به صورت "از پیش تعریف شده" می‌باشد در این حالت ابتدا کلاس‌ها تعریف می‌شوند سپس داده‌ها بر مبنای انطباق خصوصیاتشان با کلاس‌های تعریف شده به دسته‌ها تخصیص می‌یابند. در روش‌های کلاس‌بندی داده‌های درون هر کلاس دارای خصوصیات مشابهی می‌باشند و اغلب در میان خود کلاس‌ها نیز خصوصیات مشترک یافت می‌گردد.

ه) خوشه بندی[3]

گاهی اطلاعات و داده‌های موجود در پایگاه داده‌ها توزیع‌های ناشناخته و یا پیچیده‌ای دارند که به راحتی نمی‌توان آن توزیع‌ها را شناسایی نمود و مورد استفاده قرار داد. لذا برای تحلیل داده‌ها و اطلاعات موجود در پایگاه داده‌ها استفاده از روش‌هایی که نیاز به دانستن توزیع متغیرها ندارد از اهمیت خاصی برخوردار است. خوشه بندی یکی از روش‌هایی است که با توزیع داده‌های موجود سر و کار نداشته و اغلب با استفاده از معیارهای تشابه و عدم تشابه به خوشه بندی داده‌ها می‌پردازد. اغلب داده‌های درون یک خوشه دارای بیشترین شباهت می‌باشند در حالی‌که میان خود خوشه‌ها تفاوت‌های معنی‌داری وجود دارد. در

---

1. Association
2. Classification
3. Clustering

خوشه‌بندی بر خلاف کلاس بندی پس از تجزیه و تحلیل داده‌ها خوشه‌ها شکل می‌گیرند و عملیات تخصیص صورت می‌پذیرد.

موارد زیر می‌تواند دلایل استفاده از خوشه بندی را توجیه نماید:

۱- تحلیل خوشه‌ای می‌تواند حجم داده‌ها را کاهش دهد؛

۲- تحلیل خوشه‌ای می‌تواند در شناسایی داده‌های دور افتاده مورد استفاده قرار گیرد؛

۳- تحلیل خوشه‌ای با توزیع داده‌ها ارتباط ندارد؛

۴- تحلیل خوشه‌ای می‌تواند به کشف گروه‌های واقعی در پایگاه داده‌ها به کاربران کمک نماید.

به عنوان یک وظیفه داده کاوی، تحلیل خوشه‌ای می‌تواند همانند یک ابزار که به تنهایی عهده‌دار بینش در توزیع داده‌ها باشد و برای مشاهده مشخصه‌های هر خوشه و تمرکز روی یک مجموعه ویژه از خوشه‌ها و تحلیل بیشتر به کار رود.

کاربردهای داده کاوی:

دستیابی به عملکردی موفق در زمینه داده‌کاوی بیش از همه به خلاقیت، تلاش و دانش طراحان و فعالان این بخش بستگی دارد.داده کاوی شباهت بسیاری به حل یک معما یا پازل دارد.قطعات یک پازل به تنهایی بسیار ساده و جذاب هستند اما هنگامی که با تعداد بسیاری از این قطعات ساده روبرو می‌شویم در ابتدا دچار سردرگمی و احساس نگرانی از چگونگی شکل‌دهی به این ساختار پیچیده می‌شویم در ادامه هنگامی که تعدادی از قطعات در محل خود قرار گرفتند و روش انجام کار مشخص گردید به ادامه کار علاقه‌مند می‌شویم. در مقام مقایسه داده‌کاوی و پازل طراحان فرایند ابتدا اطلاعات کافی از ارتباط حجم عظیمی از پایگاه‌های داده‌ها ندارند در حالی‌که ظاهر داده‌ها به تنهایی ساده و قابل درک و توضیح هستند زمانی که طراح با خلاقیت و دانش خود پی به ارتباط و ساختار این پایگاه‌های داده می‌برد جنبه‌های متفاوتی از پایگاه‌های داده معلوم می‌شود.

کاربرد داده‌کاوی اغلب در جایی که انبار جامع داده‌ها[1]، مراکز اختصاصی و سیستم پشتیبانی تصمیم[2] استقرار دارند نمود پیدا می‌کند.

---

1. Data Warehouse
2. Dicision Support Systems (DSS)

بنابراین صنایع خرده فروشی، تولید، مخابرات، ارتباطات، بهداشت عمومی، بیمه، حمل و نقل و ...جامعه هدف در داده‌کاوی است.داده کاوی کاربردهایی نیز در تشخیص هزینه‌های غیر مجاز سیستم حسابداری، رقابت در بازار سرمایه، حمایت و رضایت مشتریان، ارائه روش‌های جدید خرید، پیش بینی خرید مشتریان و مواردی از این دست دارد.

در زمینه کشف جرم و تخلف، بسیاری از مجریان قانون و واحدهای بازرسی به شناسایی فعالیت کلاهبرداران و کشف روش‌های نوین ارتکاب جرم می‌پردازند. همچنین متخصصین را در تشخیص الگوهای بحران رفتاری در رابطه با مواد مخدر، معاملات و فعالیت‌های پول‌شویی، اقدامات گروه‌های آدم ربا، شناسایی قاچاق‌چیان در نقاط مرزی یاری می‌رساند.

سایت آمازون[1] بزرگترین فروشگاه الکترونیکی کتاب با استفاده از بانک اطلاعات دقیق، قوی و تخصصی خود ارتباط معناداری بین فعالیت‌های تک تک کاربران و سایت برقرار نموده است و بر اساس اطلاعات خود اطلاعات کاربرآنرا طبقه بندی می‌کند و بین اطلاعات کاربر جدید و اطلاعات مربوط به کالا و خدمات سایت ارتباط برقرار می‌کند.

برای بهبود بهره‌وری یک فروشگاه، مدیر فروشگاه با کمک داده‌کاوی از داده‌های انبار داده، الگوهایی را ارائه می‌کند که مشخص می‌نماید چه مشتریانی چه محصولات یا خدماتی را در چه زمانی و به چه میزان و از چه طریقی خریداری می‌نمایند. بنابراین از طریق داده‌کاوی ارزش مشتریان تعیین و رفتار آینده آنها پیش‌بینی می‌شود و از این طریق می‌توان تصمیمات آگاهانه‌ای را اتخاذ نمود.

در صنعت بیمه و پیش بینی میزان استقبال از بیمه نامه‌های جدید، تشخیص کلاهبرداری‌ها و مشخص کردن رفتارهای نامتناسب و تشخیص نیاز مشتریان و خواسته‌های آنها از مواردی هستند که با استفاده از داده‌کاوی قابل پیش‌بینی می‌باشد.

با استفاده از گردآوری و تحلیل داده‌های مربوط به آمار شمارش مراجعین به کتابخانه، میز امانت، امانت بین کتابخانه‌ای، گسترش مجموعه، تهیه مواد، کاربرد منابع الکترونیکی و روند استفاده از وب، از داده‌کاوی برای تصمیم بهینه در کتابخانه استفاده می‌شود.[2]

---

1. www.amazon.com
2. www.ketabdar.org

در مدیریت پرواز فدرال از داده‌کاوی درمرور و بازبینی داده‌ها و اطلاعات پیرامون سقوط هواپیماها استفاده می‌کنند و موارد نقص و ایرادات عمده حوادث را تشخیص داده و اقـدامات پیشگیرانه لازم و احتیاطی را لحاظ می‌کنند.

در بخش مدیریت امنیت شهری بعضی از تحلیل‌گران پیشنهاد داده‌اند که بـا اسـتفاده از این ابزار می‌توان به شناسایی فعالیت‌های تروریستی مانند ارتباطات، نقل و انتقـالات مـالی و شناسایی و ردیابی تروریست‌ها از طریق ثبت داده‌های مرتبط به سـفر و جابجـایی‌هـای آنهـا کمک گرفت.

نتیجه گیری

با عنایت به موارد ذکر شده و اهمیت تجزیـه و تحلیـل داده‌هـا و همچنـین حجـم انبـوه اطلاعات در نیروی انتظامی‌جمهوری اسلامی ایران و با توجه به قابلیت‌هـای روش‌هـای داده کاوی، در موارد زیر می‌توان این تکنیک را به کار برد:

- کشف ارتباط میان متغیرهایی از قبیل سطح تحصیلات، جمعیت بیکار، جمعیت معتاد، پشتوانه مذهبی و سطح اقتصادی خانواده‌هـا، نـرخ جـرم و جنایـت، میـزان شـاخص آسیب‌های اجتماعی و دیگر متغیرهای کمّی یا کیفی مرتبط بـا شـاخص‌هـای امنیتـی درحوزه‌های استحفاظی کلانتری‌ها یا منطقه‌ای وسیع‌تر؛
- خوشه‌بندی مناطق یا استان‌ها بر اساس میزان امنیت ؛
- کشف رابطه میان انواع جرایم؛
- کشف رابطه میان شاخص‌های امنیتی با متغیرهای موجود؛
- کلاس‌بندی کلانتری‌ها بر اساس نحوه عملکردشان ؛
- خوشه‌بندی مجرمین ؛
- پیش‌بینی وقوع انواع جرم ؛
- دسته‌بندی جرایم؛
- امکان شناسایی، تحلیل و بررسی فرصت‌ها و تهدیدها؛
- ارائه راه‌کارهایی جهت بهبود عملکرد در حوزه‌های ماموریتی و پشتیبانی؛
- .....

منابع :

- حقیقی، عبدالحمید و دیگران (۱۳۸۵)، "***داده‌کاوی و کاربرد آن در کیفیت داده‌ها***"، فصلنامه گزیده مطالب آماری شماره ۲.
- دیوید، هند (۱۳۸۶)، "***داده‌کاوی چیست؟***"، ترجمه: حمید معظمی گودرزی، گزیده مطالب آماری، شماره ۵۲.
- سعیدی، احمد (۱۳۸۴)، "***داده‌کاوی، مفهوم و کاربرد آن در آموزش عالی***"، ماهنامه آموزش عالی شماره ۱۸.
- کانتاردزیک، محمد (۱۳۸۵)، "***داده‌کاوی***"، ترجمه: علی خانزاده، تهران نشر علوم رایانه.
- حائری مهریزی، علی‌اصغر "***داده‌کاوی: مفاهیم و روش‌ها و کاربردها***"، پایان‌نامه کارشناسی ارشد، دانشگاه علامه‌طباطبایی.

- *Hand. D. J (1998)" **Review of Data mining**", The American statistician, 52, 112-118*
- *Pidd. M. (2004), "**Tools for thinking, modeling in management science**", Lancaster university .*
- *Jeffery W. Seifert,(2004), "**Analyst in information science and technology policy**", data mining : an overview.*
- *www.ketabdar.org*