اهمیت پردازش داده‌ها

راضیه کرمی و ملیحه‌سادات ملک‌جعفریان

مقدمه

داده‌های گردآوری‌شده قبل از هر نوع تحلیلی باید آماده‌سازی شوند. بنا بر این ویرایش، کدگذاری، ورود داده‌ها به رایانه، تعریف داده‌ها و طبقه‌بندی از جمله مراحلی هستند که باید قبل از تحلیل داده‌ها انجام گیرند. این مراحل، «پردازش داده‌ها» نامیده می‌شود. بنا بر این همواره پردازش و آماده‌سازی داده‌ها مقدم بر تحلیل آن‌ها است و هر نوع سهل‌انگاری در این مرحله می‌تواند نتایج و یافته‌های پژوهش را تحت تأثیر قرار دهد. لذا به‌دلیل اهمیت این موضوع، فرایند اساسی پردازش و تحلیل داده‌ها در شکل ۱ نشان داده شده است.

اطلاعات از داده‌ها استخراج می‌شود. داده‌های دارای سطح کیفیت پایین منجر به اطلاعات غیر مطمئن و حتی غلط می‌شود [۱]. قوت یک زنجیر توسط ضعیف‌ترین حلقه‌ی آن تعیین می‌شود. پس تعریف دقیق و صحیح داده‌ها برای کسب اطلاعات معتبر، ضروری است. در طی پردازش داده‌ها، داده‌های تحلیل‌شده به شکلی تبدیل می‌شود که نمایش داده‌ها بهتر صورت گیرد. یکی از نیازهای درک پردازش داده‌ها، تعریف دقیق مفاهیمی همچون داده‌ها، اطلاعات و پردازش است، که در قسمت بعد به تعریف هر یک از این مفاهیم می‌پردازیم. شایان ذکر است که مرحله‌ی آخر پردازش داده‌ها، نمایش داده‌ها می‌باشد که خود دارای اصول خاصی است. پردازش داده‌ها شامل همه‌ی پردازش‌ها از ثبت داده‌ها تا داده‌کاوی است.

داده‌ها

مواد اولیه‌ی مورد نیاز که پیش از اجرای عملیات پردازش و به‌عنوان ورودی یا معلوم مسئله در اختیار داریم، «داده‌ها» نامیده می‌شود. به‌عنوان مثال اگر مسئله‌ی ما تعیین معدل دانش‌آموز باشد، نمره‌های موجود، نقش داده‌ها را دارند. به عبارت دیگر، داده‌ها می‌توانند صدا، تصویر، عدد، حروف الفبا، علامت یا ترکیبی از آن‌ها باشد که به‌عنوان ورودی تلقی می‌شود [۳].

اطلاعات

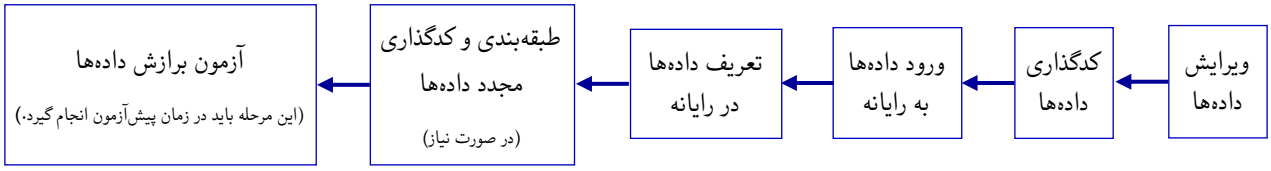
پس از هر پردازش خاص روی داده‌ها، آن‌ها به اطلاعات تبدیل می‌شوند. به‌عنوان مثال، معدل و رتبه‌ی دانش‌آموز، دیگر جزو داده‌ها نیستند، بلکه اطلاعاتی هستند که از داده‌های خام به دست آمده‌اند؛ یعنی اطلاعات، همان داده‌های خام اولیه‌اند که پردازش شده‌اند. اطلاعات نیز همانند داده‌ها در انواع و اشکال گوناگونی مانند صدا، تصویر، عدد، حروف، علامت یا ترکیبی از آن‌ها هستند [۳].

پردازش

به مجموعه‌ی عملیاتی که بر روی داده‌ها برای رسیدن به نتایج صورت می‌گیرد، پردازش گفته می‌شود. به‌عنوان مثال، مرتب‌سازی داده‌ها که به‌نوعی رتبه‌بندی داده‌ها تلقی می‌شود، جستجو در بین آن‌ها، یا محاسبات انجام‌گرفته بر روی داده‌ها، از انواع پردازش محسوب می‌شوند. به‌طور کلّی، اصطلاح پردازش داده‌ها می‌تواند هر پردازشی را که داده‌ها را از شکلی به شکلی دیگر تبدیل می‌کند در بر گیرد، اگرچه «تبدیل داده‌ها» می‌تواند اصطلاح منطقی‌تر و صحیح‌تری باشد. از این دیدگاه، پردازش داده‌ها تبدیل داده‌ها به اطلاعات خواهد بود و همچنین تبدیل مجدد اطلاعات به داده‌ها. تفاوت این‌جاست که تبدیل، نیاز به درخواست نخواهد داشت [۲].

در مسئله‌ی تعیین معدل یک دانش‌آموز، عملیات محاسبه‌ی مجموع نمره‌ها و تقسیم عدد حاصل بر تعداد درس‌ها، پردازش مسئله خواهد بود. در بحث پردازش داده‌ها نباید از فرایند پیش‌پردازش غافل شد. پیش‌پردازش داده‌ها شامل همه‌ی تبدیلاتی است که بر روی داده‌های خام

راضیه کرمی کارشناس آمار شرکت آب و فاضلاب شهری استان همدان است.
ملیحه‌سادات ملک‌جعفریان کارشناس آمار شرکت آب و فاضلاب شهری استان سمنان است.



شکل ۱- فرایند پردازش داده‌ها

صورت می‌گیرد و آن‌ها را به‌صورتی در می‌آورد که برای پردازش‌های بعدی نظیر استفاده در دسته‌بندی، ساده‌تر و مؤثرتر باشد. فرایند پردازش در سیستم رایانه را می‌توان با شکل ۲ بیان کرد.

◉ پیش‌پردازش داده‌ها

ابزارها و روش‌های مختلفی برای پیش‌پردازش وجود دارد مانند بهنجار کردن، که داده‌ها را به داده‌هایی جدید با بازه‌ی تغییرات یا توزیع مناسب تبدیل می‌کند، سفید کردن، که در ناهمبسته کردن داده‌ها استفاده می‌شود، کاهش ابعاد، که برای حذف داده‌های تکراری، اضافی یا نامربوط برای دسته‌بندی استفاده می‌شود.

◉ بهنجار کردن

در بسیاری از کاربردها بازه‌ی تغییرات ویژگی‌ها یکسان نیست. برای رفع این مشکل از دو روش زیر استفاده می‌شود: روش اول با یک تبدیل خطی یا ناخطی، داده‌ها را در بازه‌ای قرار می‌دهد که توسط کاربر انتخاب می‌شود. این بازه در کاربردهای شبکه‌ی عصبی، معمولاً به‌صورت (۱ , ۱-) یا (۱ , ۰) انتخاب می‌شود. اگر مقادیر کمینه و بیشینه‌ی داده‌ها را فرض کنیم، از رابطه‌ی (۱) برای بهنجار کردن استفاده می‌شود:

$$y_{ij} = y_{i,\min} + (y_{i,\max} - y_{i,\min}) \times \frac{(x_{ij} - x_{i,\min})}{(x_{i,\max} - x_{i,\min})} \quad (۱)$$

این تبدیل، خطی است و مقیاس داده‌ها را تغییر می‌دهد. در این حالت، تابع توزیع تغییر نمی‌کند.

روش دوم با استفاده از ویژگی‌های آماری است. برای این کار از تخمین میانگین و واریانس داده‌ها استفاده می‌شود:

$$y_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i} \quad (۲)$$

در این حالت، داده‌های تبدیل‌شده در هر بعد و ویژگی، دارای میانگین صفر و واریانس یک می‌باشند. این تبدیل، یک تبدیل خطی است.

در رابطه‌های (۱) و (۲) منظور از متغیرهای تعریف‌شده، داده‌های ما می‌باشد. منظور از iها دسته‌های داده، و منظور از jها داده‌های موجود در دسته‌ها است. به‌عنوان مثال، $x_{۱۲}$، داده‌ی دوم در دسته‌ی اول است. برای هر دسته، میانگین و انحراف معیار و کمینه و بیشینه قابل تعریف است.

◉ سفید کردن

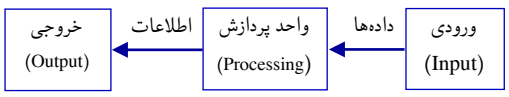
سفید کردن برای ناهمبسته کردن رشته‌ای از داده‌ها مورد استفاده قرار می‌گیرد. رشته‌ی سفید شده دارای میانگین صفر، واریانس یک و همبستگی صفر است. سفید کردن رشته‌ای از داده‌های تصادفی، بسیار شبیه عمود کردن تعدادی بردار است.

◉ کاهش ابعاد

دلیل کاهش ابعاد را می‌توان راحت‌تر شدن تحلیل‌های بعدی، افزایش عملکرد جداکننده بر اساس نمایش بهتر (پایدارتر)، حذف اطلاعات تکراری یا نامربوط، یا تلاشی برای کشف ساختار با به دست آوردن نمایش گرافیکی از داده‌ها دانست.

◉ نمایش اطلاعات بعد از پردازش داده‌ها

نمایش صحیح اطلاعات برای فهم درست کاربر ضروری است. پردازش اطلاعات باید این توانایی را به کاربر بدهد که اطلاعات مورد نظر خود را به راحت‌ترین و واضح‌ترین وجه استخراج کند. به‌عنوان مثال، داده‌های



شکل ۲- فرایند پردازش در سیستم رایانه

جدول ۱- نمایش اطلاعات به‌شکل جدول

ماه	مصرف آب (لیتر)
فروردین	۲,۴۹۶,۳۶۷
اردیبهشت	۵,۰۱۷,۱۸۸
خرداد	۲,۷۳۱,۱۳۰
تیر	۳,۷۹۷,۸۳۴
مرداد	۳,۳۵۷,۰۶۴
شهریور	۵,۷۳۵,۳۴۳
مهر	۳,۴۵۷,۵۸۶
آبان	۲,۷۴۲,۲۳۹
آذر	۳۴۲,۸۷۴
دی	۳,۳۲۰,۰۸۶
بهمن	۴,۰۷۷,۲۲۷
اسفند	۳,۳۴۰,۳۳۰

شکل ۳- نمایش اطلاعات به‌شکل نمودار

مربوط به تولید آب به‌تفکیک ماه‌های سال که از سامانه‌ی مدیریت اطلاعات (Web MIS) شرکت مهندسی آب و فاضلاب کشور در سال ۱۳۹۲ استخراج شده است، به دو صورت جدول ۱ و شکل ۳ نمایش داده شده است.

◉ علم داده‌کاوی

از داده‌کاوی به‌عنوان مرحله‌ای از فرایند کشف دانش یاد می‌شود که الگوها یا مدل‌ها را در میان انبوهی از داده‌ها پیدا می‌کند. علاوه بر این، داده‌کاوی علمی است که از تلفیق علوم متفاوت همچون آمار، یادگیری ماشینی، پایگاه‌های اطلاعاتی و مانند آن شکل می‌گیرد و مواد اولیه‌ی به کار رفته در آن، داده‌ها هستند. از این رو سنگ بنای عملیات داده‌کاوی خوب، به‌کارگیری و دسترسی به داده‌های اولیه‌ی خوب و مناسب است. در واقع برای کشف دانش به‌کمک داده‌کاوی باید مقدماتی صورت گیرد که مجموعه‌ی این اقدامات را آماده‌سازی داده‌ها گویند.

اهمیت آماده‌سازی داده‌ها در این واقعیت است که «فقدان داده‌های باکیفیت، برابر با فقدان کیفیت در نتایج کاوش است». در جدول ۲ مقایسه‌ای بین اهمیت آماده‌سازی داده‌ها نسبت به سایر گام‌های کشف دانش به‌کمک داده‌کاوی صورت گرفته است. با این حال، متأسفانه بسیاری اهمیت آماده‌سازی داده‌ها را فراموش می‌کنند یا آن را کم‌اهمیت می‌انگارند. اهمیت این موضوع سبب شده تا بسیاری، نتایج حاصل از داده‌کاوی را صرفاً در صورتی قابل اعتنا بدانند که از پیش‌پردازش مناسبی برخوردار باشد.

جدول ۲- مقایسه‌ی اهمیت گام آماده‌سازی داده‌ها با سایر گام‌های داده‌کاوی

گام‌های داده‌کاوی	درصد زمان صرف‌شده از کل کار	درصد اهمیت در موفقیت نهایی کار
آماده‌سازی داده‌ها	۷۵	۷۵
بررسی داده‌ها	۲۰	۱۵
مدل‌سازی داده‌ها	۵	۱۰

◉ نتیجه‌گیری

داده‌های خام معمولاً دچار مشکلاتی مانند نوفه، اریبی، تغییرات شدید در بازه‌ی دینامیکی و نمونه‌برداری هستند و استفاده از آن‌ها به همان شکل، موجب تضعیف طراحی‌ها و تحلیل‌های بعدی خواهد شد. با اجرای پردازش صحیح بر روی داده‌ها، می‌توان تحلیل بهینه و راحت‌تری را انجام داد. علم داده‌کاوی که به‌دنبال یافتن الگو و مدل در بین انبوهی از داده‌ها است، با پیمودن صحیح روند پیش‌پردازش و پردازش داده‌ها، ما را به این مهم نزدیک و مراحل داده‌کاوی را میسر و بهینه خواهد کرد. همچنین، نوع نمایش داده‌ها در کنار یکدیگر، یا به بیان دیگر، نمایش متغیرهای گوناگون پس از ویرایش و پردازش‌های لازم نیز نقش مهمی را در تحلیل و حتی در نوع برداشت کاربر از اطلاعات به‌دست‌آمده دارد.

مرجع‌ها

[۱] محمدی تاکانی، سید محسن (۱۳۸۴). معماری کامپیوتر.

[۲] کلانتری، خلیل (۱۳۸۶). پردازش و تحلیل داده‌ها در تحقیقات اجتماعی-اقتصادی.

[3] Oliver and Champons (2004). Data Processing and Information Technology.

[4] Garys. Popkin (2008). Introduction to Data Processing.