فصلنامه گزیده‌مطالب آماری
سال ۱۷، شماره ۲، تابستان ۱۳۸۵، صص ۷۴-۴۷

داده‌کاوی و کاربرد آن در کیفیت داده‌ها

عبدالحمید حقیقی*، شکوفه قصوری، آزاده برفی‌پور، علیرضا شماخی
علی‌اصغر حائری مهریزی، حسین حسنی، محمدرضا یگانگی

مرکز آمار ایران

چکیده. کیفیت داده‌ها یکی از مفاهیم پیچیده و ساختار نیافته است و حل مسائل کیفیت داده‌ها به اطلاعات وابسته زیاد در حوزه مشخص نیاز دارد که این اطلاعات حاصل تجربه کارشناسی است. از طرفی از دهه ۹۰ به بعد نیز که پایگاه‌های داده‌ای حجیم ایجاد شد امکان این که فقط تجربه کارشناسی به کیفیت داده‌ها کمک کند وجود ندارد. بنا بر این بهترین راه برای افزایش کیفیت داده‌ها نظارت و اعتبارسنجی مداوم داده‌ها و فراداده‌ها بر اساس نظریات کارشناسی و با استفاده از نرم‌افزارها و الگوریتم‌های آماری می‌باشد. برای بهبود کیفیت، باید به‌طور خاص روی فراداده‌ها تمرکز کرد و در این بین داده‌کاوی و جستجوی داده‌ها می‌تواند دید وسیع‌تری را برای ما ایجاد کند و شکاف‌های داده‌ها را تا حدودی پر کند. اما از طرفی مسئله اتوماسیون کامل بسیار پیچیده و سخت است و الگوریتم‌های موجود، بخش کوچکی از مسئله را حل می‌کنند. این مقاله به کیفیت داده‌ها پرداخته و برای بهبود کیفیت داده‌ها استفاده از روش‌های داده‌کاوی را مطرح می‌نماید. در انتهای مقاله نیز به استفاده از برخی روش‌های داده‌کاوی در طرح هزینه و درامد خانوارهای شهری و روستایی مرکز آمار ایران پرداخته می‌شود و پیش‌نیازها، مسائل و مشکلات و پیشنهادها در به‌کارگیری روش‌های داده‌کاوی در این طرح ارائه می‌شود.

واژگان کلیدی: کیفیت داده‌ها؛ داده‌کاوی.

* نویسنده عهده‌دار مکاتبات

۱- مقدمه

از سال ۱۹۵۰ به بعد که رایانه، در تحلیل و ذخیره‌سازی داده‌ها به کار رفت، حجم اطلاعات ذخیره شده در آن پس از حدود ۲۰ سال دو برابر شد و همزمان با پیشرفت فناوری اطلاعات، حجم داده‌ها در پایگاه داده‌ها هر دو سال یک بار، دو برابر شد و همچنان با سرعت بیش‌تری نسبت به گذشته حجم اطلاعات ذخیره شده بیش‌تر و بیش‌تر می‌شود. با وجود شبکه جهانی وب، سیستم‌های یکپارچه اطلاعاتی، سیستم‌های یکپارچه بانکی، تجارت الکترونیکی و... لحظه به لحظه به حجم داده‌ها در پایگاه داده‌ها اضافه شده و باعث به‌وجود آمدن انبارهای (توده‌های) عظیمی از داده‌ها شده است، به‌طوری که ضرورت کشف و استخراج سریع و دقیق دانش از این پایگاه داده‌ها را بیش از پیش نمایان کرده است.

شدت رقابت‌ها در عرصه‌های علمی، اجتماعی، اقتصادی، سیاسی و نظامی نیز اهمیت سرعت یا زمان دسترسی به اطلاعات را دو چندان کرده است. بنا بر این نیاز به طراحی سیستم‌هایی که قادر به اکتشاف سریع اطلاعات مورد علاقه کاربران با تأکید بر حداقل مداخله انسانی باشند از یک سو و روی آوردن به روش‌های تحلیل متناسب با حجم داده‌های حجیم از سوی دیگر، به‌خوبی احساس می‌شود. در حال حاضر، داده‌کاوی مهم‌ترین فناوری برای بهره‌برداری مؤثر، صحیح و سریع از داده‌های حجیم است و اهمیت آن رو به فزونی است.

با توجه به‌وجود اطلاعات ارزشمند در پایگاه‌های داده‌ای در اواخر دهه ۸۰ میلادی، تلاش برای استخراج و استفاده از اطلاعات پایگاه‌های داده‌ای شروع شد. داده‌کاوی فرایندی است که در آغاز دهه ۹۰ پا به عرصه ظهور گذاشته و با نگرشی نو، به مسئله استخراج اطلاعات از پایگاه داده‌ها می‌پردازد. در سال ۱۹۸۹ و ۱۹۹۱ کارگاه‌های کشف دانش از پایگاه داده‌ها توسط پیاتتسکی و همکارانش و در فاصله سال‌های ۱۹۹۱ تا ۱۹۹۴کارگاه‌های فوق، توسط فایاد و پیاتتسکی و دیگران برگزار شد. به‌طور رسمی اصطلاح داده‌کاوی برای اولین بار توسط « فیاض » در اولین کنفرانس بین‌المللی « کشف دانش و داده‌کاوی » در سال ۱۹۹۵ مطرح شد. از سال ۱۹۹۵ داده‌کاوی به صورت جدی وارد مباحث آمار شد [۸] و در سال ۱۹۹۶، اولین شماره مجله کشف دانش از پایگاه داده‌ها

منتشر شد. امروزه کنفرانس‌های مختلفی در این زمینه در سراسر دنیا برگزار می‌شود. داده‌کاوی حاصل تحول تدریجی در طول تاریخ بوده و از اوایل دهه ۹۰ همزمان با همه‌گیر شدن استفاده از پایگاه‌های داده‌ای به‌عنوان یک علم مطرح شده است [۹].

موضوع داده‌کاوی شناخت چیزهای جدید و با ارزش، بالقوه مفید، رابطه‌های منطقی و الگوهای موجود در داده‌ها است (چانگ و گری، ۱۹۹۹). در جوامع مختلف یافتن الگوهای مفید در داده‌ها با عناوین متعددی (مانند داده‌کاوی) بیان می‌شود. برای مثال از عنوان‌هایی نظیر استخراج دانش، کشف اطلاعات، برداشت اطلاعات، پردازش الگوهای داده‌ها (فایاد و همکاران، ۱۹۹۶) می‌توان نام برد.

عبارت « داده‌کاوی » توسط آمارشناسان، محققان پایگاه‌های داده‌ها و سیستم‌های اطلاعات مدیریتی و جوامع بازرگانی به کار برده می‌شود. عبارت کشف دانش در پایگاه داده‌ها عموماً برای اشاره به فرایند کلی کشف دانش مفید از داده‌هایی که داده‌کاوی گام مهمی در این فرایند است، مورد استفاده قرار می‌گیرد [۸] گام‌های دیگری در فرایند کشف دانش در پایگاه داده‌ها نظیر آماده کردن داده‌ها، انتخاب داده‌ها، تمیز کردن داده‌ها و درک درست از فرایند داده‌کاوی موجب می‌شود تا اطلاعاتی که برای ما مفید هستند از داده‌ها استخراج شوند. داده‌کاوی از تحلیل‌های سنتی داده‌ها و رویکردهای آماری نشأت گرفته است به‌طوری که شامل فنون تحلیلی‌ای است که از شاخه‌های دیگری تشکیل شده است، مانند:

- تحلیل‌های عددی؛
- الگوهای سازگار و سطوحی از هوش مصنوعی مانند یادگیری ماشین؛
- شبکه‌های عصبی و الگوریتم‌های ژنتیک؛
- ....

با وجود این بسیاری از داده‌کاوی‌ها بر روش‌های سنتی و رویکردهای تحلیل داده‌های مبتنی بر فرضیه تکیه دارد. اساساً دو رویکرد برای داده‌کاوی وجود دارد که از لحاظ ایجاد و طراحی مدل و یافتن الگوها با هم فرق دارند. اولین رویکرد که مربوط به ساخت مدل است (جدا از مشکلاتی که ذاتاً در مجموعه داده‌های بزرگ وجود دارد) مشابه

روش‌های کاوشگرانه آماری مرسوم است. در این حالت هدف این است تا خلاصه‌ای کلی از مجموعه‌ای از داده‌ها برای شناخت و توضیح خصوصیت‌های اصلی شکل توزیع به دست آوریم [۹]. مثال‌هایی از این قبیل مدل‌ها شامل تحلیل خوشه‌ای بخشی از مجموعه داده‌ها، مدل رگرسیونی برای پیشگویی و قاعده رده‌بندی با ساختار درختی است.

نوع دوم رویکرد داده‌کاوی، رویکرد تشخیص الگو است. این رویکرد سعی بر آن دارد تا انحراف‌هایی هر چند کوچک (از حد مطلوب) را تشخیص دهد (که در هر صورت حائز اهمیت هستند)، تا الگوها و روندهای غیر معمول نمایان شود. مثال‌هایی نظیر الگوهای نامعمول (برای تشخیص کلاهبرداری) در استفاده از کارت‌های اعتباری و موضوع‌هایی که الگوهایی با ویژگی‌های نامشابه با سایر الگوها دارند از این نوع کاربرد است. این دسته از راهبردهاست که موجب می‌شود تا داده‌کاوی به‌عنوان علم جستجوی اطلاعات با ارزش از بین توده عظیمی از داده‌ها به حساب آید. به‌طور کلی در پایگاه‌های داده‌ای کسب و کار (تجاری) ضعف درک الگوها به خاطر پیچیدگی زیاد آن‌هاست. این پیچیدگی‌ها در اثر ناپیوسته بودن، نامفهوم بودن و کامل نبودن به‌وجود می‌آیند [۸]. هر چند اکثر الگوریتم‌های داده‌کاوی می‌توانند اثر این گونه خصوصیت‌های نامربوط را در تشخیص الگوی اصلی تمییز دهند، ولی قدرت پیش‌گویی الگوریتم‌های داده‌کاوی با افزایش این انحراف‌ها کاهش می‌یابد (برای راهنمایی بیش‌تر به [۲،۱۲،۱۴،۱۵،۱۶،۱۸،۱۷] مراجعه کنید).

۲- تعاریف مختلف داده‌کاوی

نگاهی به ترجمه لغوی داده‌کاوی، به ما در درک بهتر این واژه کمک می‌کند. واژه لاتین Mine به معنای استخراج از منابع نهفته و با ارزش زمین اطلاق می‌شود. ادغام این کلمه با کلمه Data به معنی داده بر جستجویی عمیق از داده‌های قابل دسترس با حجم زیاد برای یافتن اطلاعات مفید که قبلاً نهفته بودند، تأکید دارد.

داده‌کاوی دارای تعریف‌های مختلفی است. این تعریف‌ها به مقدار زیادی به پیش‌زمینه‌ها و نقطه‌نظرهای افراد بستگی دارد. هر نویسنده، محقق و کاربر با توجه به دیدگاه و نوع نگرش خود تعریف‌های مختلفی از داده‌کاوی ارائه کرده‌اند. به‌عنوان مثال

می‌توان به چند تعریف داده‌کاوی که در ادامه آمده است، اشاره کرد:

الف) داده‌کاوی فرایندی از شناخت الگوهای معتبر، جدید، بالقوه مفید و قابل فهم از داده‌هاست [۸]؛

ب) داده‌کاوی به فرایند استخراج اطلاعات نهفته، قابل فهم، قابل تعقیب از پایگاه داده‌های بزرگ و استفاده از آن‌ها در تصمیم‌گیری‌های تجاری مهم اطلاق می‌شود [۲۳]؛

پ) داده‌کاوی، مجموعه‌ای از روش‌ها در فرایند کشف دانش است که برای تشخیص الگوها و رابطه‌های نامعلوم در داده‌ها مورد استفاده قرار می‌گیرد [۱۶] و [۱۱]؛

ت) فرایند کشف الگوهای مفید از داده‌ها را داده‌کاوی می‌گویند [۲۱].

لذا مشاهده می‌شود که هر کس بنا به کاربرد و موارد استفاده، تعریفی از داده‌کاوی ارائه کرده است. البته از سالیان پیش آمارشناسان با نام‌های مختلف مانند صید داده‌ها، لایروبی داده‌ها و بررسی داده‌ها به نوعی از داده‌کاوی استفاده کرده‌اند. با وجود این که داده‌کاوی یک رشته جدید علمی است، ولی امروزه کاربردهای متنوع و گسترده‌ای در رشته‌هایی مانند بازرگانی، پزشکی، مهندسی، علوم رایانه، صنعت، کنترل کیفیت، ارتباطات و کشاورزی پیدا کرده است.

۳- شاخه‌های مرتبط با داده‌کاوی

پایه و اساس داده‌کاوی در سه شاخه قدیمی ریشه دارد که مهم‌ترین آن‌ها آمار کلاسیک است. بدون آمار داده‌کاوی وجود نخواهد داشت، زیرا آمار زیربنای بیش‌تر فناوری‌هایی است که داده‌کاوی بر اساس آن‌ها بنا شده است. آمار کلاسیک از ابزاری همچون انحراف معیار، واریانس، بازه‌های اطمینان، تحلیل رگرسیون، تحلیل تشخیصی، تحلیل خوشه‌ای و... استفاده می‌کند تا جزئیات و رابطه‌ها بین داده‌ها را به‌طور دقیق مورد بررسی قرار دهد. تمامی این موارد در داده‌کاوی کاربرد دارند. لذا در قلب ابزار و روش‌های داده‌کاوی تحلیل‌های مربوط به آمار کلاسیک نقش مهمی را ایفا می‌کند.

دومین شاخه مرتبط با داده‌کاوی هوش مصنوعی AI (Artificial Intelligence)

است. این شاخه بر اساس اکتشاف ساخته شده و سعی دارد پردازش‌هایی شبیه افکار انسان را در مسائل آماری به کار برد. این شاخه مستلزم قدرت پردازش رایانه‌ای بالایی است که تا اوایل دهه ۱۹۸۰، زمانی که رایانه‌ها قدرت و سرعت پردازش کافی را نداشتند، امکان‌پذیر نبود. در آن زمان هوش مصنوعی تنها کاربردهای اندکی در پژوهش‌های پیش‌رفته، سازمان‌های دولتی ویژه و بازارهایی خاص داشت ولی در آن دوران ابر رایانه‌های لازم برای بهره‌وری از این فناوری به حدی محدود بود که کاربردی برای سایر افراد و ارگان‌های جامعه نداشت (برای راهنمایی بیش‌تر به [۴–۶] مراجعه کنید).

سومین شاخه مرتبط با داده‌کاوی یادگیری ماشین است که تلفیقی از آمار و هوش مصنوعی است. یادگیری ماشین می‌تواند به‌عنوان هوش مصنوعی تکامل‌یافته مطرح شود. زیرا در این روش اکتشاف‌های هوش مصنوعی با تحلیل‌های آماری پیشرفته ادغام می‌شود. یادگیری ماشین سعی دارد به برنامه‌های رایانه‌ای این امکان را بدهد تا در مورد اطلاعاتی که به آن‌ها داده می‌شود، یاد بگیرند تا چنین برنامه‌هایی بتوانند متناظر با اطلاعات متفاوتی که به آن‌ها داده می‌شوند تصمیم‌گیری‌های متفاوتی انجام دهند. این تصمیم‌گیری‌ها بر اساس اصول پایه‌ای آمار انجام می‌گیرد. علاوه بر آمار، الگوریتم‌های هوش مصنوعی و هوش مصنوعی اکتشافی پیشرفته، ابزاری برای رسیدن به این هدف می‌باشند (برای راهنمایی بیش‌تر به [۲۰–۱۹] مراجعه کنید).

بهترین تعریف برای داده‌کاوی را می‌توان تلفیق پیشرفت‌های قدیمی و جدید آمار، هوش مصنوعی و یادگیری ماشین دانست. این فنون برای تحلیل داده‌ها و پیدا کردن رابطه‌هایی که با روش‌های دیگر قابل یافتن نیستند، استفاده می‌شود. در جامعه امروزی که با اطلاعات و داده‌های زیادی سر و کار داریم یافتن رابطه‌های بین داده‌ها و بهینه‌سازی زمان تحلیل داده‌ها کاری دشوار است، اما داده‌کاوی پاسخی به این مشکل است.

داده‌کاوی یک رشته نسبتاً جدید علمی است که از انجام پژوهش‌ها، حداقل در رشته‌های مختلف آمار، یادگیری ماشین، علوم رایانه به‌خصوص مدیریت پایگاه داده‌ها شکل گرفته است. البته مرزهای این رشته‌ها در داده‌کاوی مبهم و بعضی وقت‌ها دارای اشتراک‌های فراوانی هستند.

٤- کاربرد داده‌کاوی در آمار رسمی

اطلاعات یکی از نیازهای مبرم سیاست مداران در تصمیم‌گیری‌های کلان هر کشور است و بدون داشتن اطلاعات کافی، جامع و بهنگام تصمیم‌گیری ممکن نیست یا با نقایص زیادی مواجه است. امروزه با گسترش روزافزون علم آمار و استفاده از روش‌های مختلف جمع‌آوری داده‌ها شاهد حجم انبوهی از داده‌ها در مراکز تولید آمار هستیم که تنها با داشتن یک برنامه اصولی و ابزارهای کارامد و متناسب با اهداف از پیش تعیین شده می‌توان نیازهای کاربران از پایگاه داده‌ها را پاسخ گفت. پایگاه داده‌ها به‌عنوان بزرگ‌ترین منبع اطلاعاتی، حاوی حجم وسیعی از داده‌های جمع‌آوری شده است که روز به روز به حجم این پایگاه داده‌ها نیز افزوده می‌شود و متأسفانه به‌دلیل نبود آینده‌نگری کافی در زمان تشکیل و رشد آن و نبود یک استاندارد دقیق و همه‌جانبه در این مورد استفاده از آن را دچار مشکل کرده است، به‌طوری که برای یافتن اطلاعات مورد نیاز باید هزینه‌های زمانی بسیاری سپری شود. در عصر حاضر نیاز مبرم و ضروری پس از جمع‌آوری داده، چگونگی انجام محاسبات بر روی داده‌هایی با حجم عظیم و به دست آوردن و استخراج اطلاعات (دانش) مبتنی بر خواسته‌ها و نیازهای بشر است. آن‌چه امروزه اهمیت بسیار زیادی پیدا کرده است، کمبود یا نبود اطلاعات مورد نیاز نیست بلکه کمبود یا نبود روش‌هایی مناسب و استاندارد به‌منظور نگهداری، به روز کردن، در دسترس قرار دادن و در حالت آرمانی‌تر کشف دانش جدید از اطلاعات موجود است. چگونگی انجام محاسبات و تئوری‌های مربوط به آن و استخراج دانش و اطلاعات از حجم داده‌های انبوه موضوع اصلی کشف دانش در پایگاه اطلاعات است [۱۰].

یکی از راهکارهای پیشنهادی برای حصول به این هدف استفاده از سیستم‌های داده‌کاوی است. سیستم‌های داده‌کاوی این امکان را به کاربر می‌دهد که بتواند انبوه داده‌های جمع‌آوری شده را تفسیر و الگوها و دانش (اطلاعات) نهفته در آن را استخراج نماید. سازمان‌ها و مراکز ملی آماری نیز از این فن بی‌بهره نبوده‌اند به‌طوری که می‌توان به کاربرد داده‌کاوی در آمارهای رسمی اشاره کرد. از آن‌جا که مراکز جمع‌آوری داده‌های آمارهای رسمی با حجم انبوهی از داده‌ها مواجه هستند، لذا نیاز به استفاده از این فن مبرم و ضروری به‌نظر می‌رسد. در ادامه به چند کاربرد داده‌کاوی در افزایش بهبود کیفیت داده

اشاره شده است:

- داده‌های دورافتاده (پرت): یکی از موارد استفاده داده‌کاوی کشف داده‌های دورافتاده با استفاده فنون و الگوریتم‌های مختلف است؛
- داده‌های گمشده: فنون داده‌کاوی در براورد مشاهده‌های گمشده نیز کاربرد دارد. در این بین استفاده از روش‌های خوشه‌بندی می‌تواند راهگشا باشد [۲۲].

اخیراً پژوهش‌های مختلف و وسیعی در زمینه کاربرد داده‌کاوی در آمارهای رسمی صورت گرفته و چندین کارگاه تحت عنوان کاویدن داده‌های رسمی (اولین کارگاه در سال ۲۰۰۲ در فنلاند و دومین در سال ۲۰۰۴ در ایتالیا) برگزار شده است که نشان‌دهنده آن است این موضوع مورد توجه سازمان‌های ملی آماری، محققان آماری و سایر کاربران قرار گرفته است و بر این نکته نیز تأکید دارند که کاربرد داده‌کاوی در آمارهای رسمی موضوع جدید و نوپایی است و در مراحل تکامل قرار دارند و باید سازمان‌های ملی آماری بر استفاده از این ابزار پرتوان تأکید کنند. امروزه با وجود فناوری اطلاعات و سیستم‌های اطلاع‌رسانی شاخه‌های جدیدی به نام وب‌کاوی و متن‌کاوی نیز به‌وجود آمده که یکی از مباحث اصلی و جدید در فناوری اطلاعات است و همگی نشان‌دهنده الزام به استفاده از این ابزار است. در واقع به جرأت می‌توان گفت انتشار اطلاعات در دنیای حاضر با وجود حجم بسیار بالای اطلاعات بدون ابزار داده‌کاوی میسر نیست و اگر مراکز آماری به این مهم مجهز نباشند با مسائل و مشکلات فراوانی در نگهداری، بهنگام کردن، افزایش کیفیت، تفسیر و ارائه اطلاعات مواجه خواهند بود [۱۰].

۵- کیفیت داده‌ها

کیفیت داده‌ها یکی از مفاهیم پیچیده و ساختار نیافته است. مسئله مهم در کیفیت داده‌ها این است که حل مسائل کیفیت داده‌ها نیاز به اطلاعات وابسته زیاد در حوزه مشخص دارد که این اطلاعات حاصل تجربه کارشناسی است. تنها کارشناسان می‌توانند قوانین و جریان داده‌ها را به درستی تعیین کنند و تهیه چنین قواعدی یک مرحله اساسی در اعتبارسنجی داده‌هاست [۷].

کیفیت داده‌ها فعالیتی پیوسته است که از شروع جمع‌آوری داده‌ها تا آخرین مرحله تحلیل آن‌ها ادامه دارد. نیاز به روز کردن تعاریف متداول و معیارهای کیفیت داده‌ها روز به روز بیش‌تر احساس می‌شود و به این دلیل است که فرایند کیفیت داده‌ها و شاخص‌های مورد نیاز برای اندازه‌گیری مؤثر و نظارت بر کیفیت داده‌ها به‌طور مستمر مورد بررسی قرار می‌گیرد [۷].

بر اساس مطالعه انجام شده توسط مؤسسه داده‌انبار مسائل کیفیت داده‌های جاری، سالانه میلیون‌ها دلار هزینه در بردارد. همچنین بر اساس پژوهش انجام‌یافته از طریق مصاحبه با متخصصان داده‌کاوی براورد شده است که بین ۳۰ تا ۸۰ درصد تحلیل‌های داده‌ای صرف پاکسازی و فهم داده‌ها می‌شود. به این ترتیب اهمیت کیفیت داده‌ها به مرور وضوح بیش‌تری می‌یابد، و دلیل صحت این گفته ایجاد نرم‌افزارها، ابزارها، مشاوره‌های شرکت‌ها و سمینارهایی در مورد مسئله کیفیت داده‌ها است.

به‌طور عمده روش‌های به کار گرفته شده برای افزایش کیفیت داده را می‌توان در مورد سه رده اساسی تقسیم کرد که عبارتند از:

الف) داده‌های گمشده
ب) داده‌های ناقص
پ) داده‌های دورافتاده [۷] و [۱۳]

داده‌های گمشده یک ویژگی ثابت داده‌های حجیم است، به‌طوری که خانه‌های منفرد، ستون‌ها، ردیف‌ها و یا کل یک بخش می‌تواند گمشده باشد. روش‌هایی از ساده تا پیچیده برای جانهی مقادیر گمشده وجود دارد. گاهی اوقات، داده‌ها گم نشده‌اند بلکه ناقصند. برای مثال، ممکن است بدانیم که فرایند حداقل ۱۰ روز اجرا شده است. اما دقیقاً نمی‌دانیم چه زمانی متوقف شده است. بنا بر این، همه آن‌چه می‌دانیم این است که سیستم برای حداقل ۱۰ روز کار کرده است. بعضی مواقع دیگر، نقاط داده‌ای خارج از انتظار ما تلقی می‌شوند. چنین نقاط داده‌ای دورافتاده (پرت) نامیده می‌شوند. این‌ها نقاطی هستند که می‌توانند به‌طور بالقوه موجود باشند. روش‌های مختلف برای تشخیص و اندازه‌گیری نقاط دورافتاده در تحلیل‌های مختلف وجود دارد. در ادامه هر یک از این

عوامل تا حد بسیار اندکی توصیف شده‌اند و برای اطلاعات بیش‌تر می‌توانید به منابع و کتب مختلف در این خصوص مراجعه کنید.

٦- مقادیر گمشده

دلایل بسیاری مبنی بر وجود ایراد و اشکال در مجموعه داده‌ها وجود دارد. داده‌های ائتلافی اولین گزینه است، زیرا با یکپارچه‌سازی مجموعه‌های متفاوتی که ممکن است دارای نقاط مشترک باشند و مخصوص یک مجموعه داده خاص باشند، به‌وجود آمده‌اند. در برخی موارد، هنگامی که داده‌های دقیق یکسان از منابع مختلف جمع‌آوری می‌شود (به‌طور مثال، خرید از شعبه‌های مختلف یک مغازه خرده‌فروشی)، یک منبع خاص ممکن است اطلاعاتش را به موقع برای همگردانی در بانک اطلاعاتی یکپارچه ارسال نکند. به‌طور مثال تقاضای فرد به‌خصوصی ثبت نشده باشد و یا برخی مولفه‌های داده‌ای (به‌طور مثال شماره تلفن مشتری) به‌وسیله افراد متقاضی وارد نشده باشد. دیگر موضوعات کیفیت داده‌ای زمانی پیش می‌آیند که مقادیر یکسانی برای نمایش پیش‌فرض‌ها و مقادیر گمشده به کار می‌روند [۷].

۷- چرا باید مراقب باشیم؟

چرا باید در مورد داده‌های گمشده نگران باشیم؟ اول این که اگر همه داده‌ها با زمینه داده گمشده را کنار بگذاریم، ممکن است حجم عظیمی از داده‌ها (مثلاً ممکن است چیزی بین ۳۰ تا ۷۰ درصد) از دست بدهیم. به‌علاوه، مطالعه الگوها و روندها در داده‌های گمشده می‌تواند به پیدا کردن و تشریح دلایل و علل به‌وجود آمدن آن‌ها کمک کنند و دیگر موضوعات کیفیت داده را آشکار سازند. نهایتاً، داده‌های گمشده می‌توانند اریبی‌های جدی در تحلیل را که به ندرت در یک روند تصادفی مخفی می‌شوند معرفی کنند. از طرفی گزارش‌هایی که بر پایه داده‌هایی با حجم زیاد داده‌های گمشده حاصل شده باشند، فاقد اطمینان مطلوب هستند.

۸- جانهی مقادیر گمشده

فرایند تخمین مقادیر داده‌های گمشده، جانهی مقادیر گمشده نامیده می‌شود. این تکنیک باید با دقت زیادی به کار گرفته شود. شایان ذکر است مقادیر جانهی را می‌توان برای تحلیل انبوه مناسب نیز به کار برد ولی هیچ مقدار جانهی شده‌ای به تنهایی قابل اطمینان نیست (چون یک براورد است).

ساده‌ترین شیوه جانهی، بر پایه برخورد با هر ویژگی به تنهایی و صرف نظر از هر گونه رابطه درونی با ویژگی‌های دیگر است. براوردهای نقطه‌ای از قبیل میانگین یا میانه می‌تواند برای مقادیر گمشده جایگزین شود. برای مثال {۳ ،... ،... ،... ،۱،۲،۳،۱،۳،۱} دارای سه مقادیر گمشده است که می‌تواند: با ۲، یعنی میانه مقادیر گم نشده جایگزین شود یا می‌توان یک توزیع با استفاده از مقادیر گم نشده شبیه‌سازی کرد و هر بار که با یک داده گمشده مواجه می‌شویم از آن توزیع برای تولید آن استفاده کنیم. بنا بر این در این مثال ساده، توزیع شبیه‌سازی شده به صورت زیر است:

$$P(۱) = ۳/۷ \qquad P(۲) = ۱/۷ \qquad P(۳) = ۳/۷$$

به‌طوری که توزیع مقادیر گمشده عیناً از توزیع کلی تبعیت می‌کند. به‌وضوح، فرضیه در نظر گرفته شده در این‌جا آن است که مقادیر گمشده از توزیع یکسان مانند مقادیر گم نشده پیروی می‌کنند. به‌وضوح روش براورد نقطه‌ای و شیوه شبیه‌سازی ساده و بر پایه فرضیات است خود داده‌هاست، اجرای آن بسیار ساده، کم‌هزینه، درک و تفسیر آن آسان است.

البته برای جانهی مقادیر می‌توان از روش‌های پیچیده‌تر جانهی استفاده کرد که ارتباطات درونی بین ویژگی‌ها و مقادیر چندگانه را به‌جای مقادیر منفرد به کار می‌گیرد. در جانهی چندگانه، از تمامی مقادیر ممکن مرتبط با یک مقدار گمشده استفاده می‌شود. جانهی مقادیر چندگانه به مجموع داده‌های چندگانه منجر می‌شود. یک روش جانهی چندگانه عمومی، روش رگرسیون است. در این روش مدل رگرسیون برای هر صفت با استفاده از صفات قبلی برازش می‌شود. برای روشن شدن مطلب به مثال زیر که در مورد قد، وزن، سن، شاخص قد است و در آن تعدادی از مشاهدات گم شده است.

شاخص قد	قد	سن	وزن
۵	۱۰	۲	۲۰
۳	۹	۵	۱۵
.	۱۰	۵	۲۵
.	.	۴	۲۰
.	.	۱	۱۰

برای این منظور از مدل‌های مختلفی می‌توان استفاده اما هر دفعه سعی می‌شود از متغیری استفاده کنیم که کمترین مشاهده گمشده را داشته باشد. لذا مدل زیر حاصل می‌شود:

$$\text{قد} = \alpha + \beta_1 \text{ سن } + \beta_2 \text{ وزن}$$

بعد از اجرای هر مدل رگرسیون، مقادیر گمشده با مقادیر پیش‌بینی شده از مدل همراه با یک جمله خطا (یک انحراف مقیاس‌دار) جایگزین می‌شود و سپس می‌توان مدل‌های رگرسیون دیگری به صورت زیر اجرا کرد:

$$\text{شاخص قد} = \alpha + \beta_1 \text{ سن } + \beta_2 \text{ وزن} + \beta_3 \text{ قد}$$

و به همین ترتیب، تا این که مقادیر گمشده با مقادیر رگرسیونی جایگزین شوند. از آن‌جایی که جمله خطا به صورت تصادفی تغییر می‌کند، می‌توانیم مجموعه داده‌های چندگانه با چرخشی از طریق فرایند جانهی تولید کنیم. هر مجموعه داده به تنهایی تحلیل یا تحلیل‌ها از مجموعه داده‌های چندگانه را با هم ترکیب می‌کند تا یک مجموعه قابل اعتماد واحد از نتایج به دست آید.

روش دیگر گروه‌بندی داده‌ها با استفاده از صفت گمشدگی است. مجدداً با این فرض که در این روش یک متغیر نشانگر (با مقادیر ۰ و ۱) برای نشان دادن این که آیا ویژگی y_j گمشده است یا خیر ساخته می‌شود. یک رگرسیون لجستیک بر پایه گم شدن برای هر مشاهده ساخته می‌شود. داده‌ها با احتمال این که صفت y_j گمشده است یا خیر گروه‌بندی می‌شوند. مقادیر y_j برای مقادیر گمشده از مقادیر معلوم y_j درون هر گروه با

استفاده از ساز و کاری که تقریب بیزی بوت استرپ نامیده می‌شود، تولید می‌شود.

برای الگوهای مقادیر گمشده می‌توان از روش MCMC (مونت کارلو زنجیره مارکف) برای شبیه‌سازی داده‌ها استفاده کرد. فرض می‌شود داده‌ها دارای یک توزیع چند متغیره نرمال‌اند. سپس مقادیر گمشده با (a) ساختن توزیع شرطی مقادیر گمشده به شرط مقادیر مشاهده شده و (b) محاسبه پارامترهای توزیع نرمال چند متغیره با استفاده از نمونه، براورد می‌شوند. گام‌های a و b تکرار می‌شوند. تا زمانی که براوردها پایدار شوند، فرایند تکرار می‌شود. زنجیره مارکف به دوتایی زیر اشاره می‌کند: $(\tilde{y}\,;\,\tilde{\theta})$ که $\tilde{y}$ مجموعه براوردهای مقادیر گمشده و $\tilde{\theta}$ مجموعه پارامترهای براورد شده در بازگشت iام می‌باشد. مقادیر iامین بازگشت تنها به مقادیر بازگشت $(i-1)$ بستگی دارد. از این رو زنجیره مارکف نام گرفته است. همان‌طور که انتظار می‌رود این روش پر هزینه است و برای مجموعه داده‌های بزرگ با توجه به افزایش حجم پیچیدگی جانهی غیر ممکن می‌شود. اما، بهترین استفاده از روش MCMC شاید شبیه‌سازی مقادیر باشد. روش‌های مختلفی در این زمینه وجود دارد ولی به این سه روش اشاره شد زیرا اجرای نرم‌افزاری قابل اطمینان این تکنیک‌ها در SAS موجود است. داده‌های گمشده و رفتار آن‌ها یک موضوع مهم کیفیت داده‌ها برای داده‌کاوها است. اما باید توجه داشت یک راه حل مناسب بستگی به منابع محاسباتی و تحلیل خطاها در تقریب زدن مقادیر گمشده دارد [۷].

۹- داده‌های ناقص

وضعیت‌هایی وجود دارد که داده‌ها موجودند اما تغییر کرده‌اند و یا تا زمان تحویل به کاربرها دستکاری شده‌اند. دو کلاس (رده) از این داده‌ها وجود دارد که از سوی آمارشناسان داده‌های ناقص نامیده شده‌اند. یک مجموعه، دارای داده‌های بریده شده است. برای مثال خانوارهایی که کمتر از ۲۰ هزار تومان در سال هزینه می‌کنند ممکن است در بانک اطلاعاتی جزو خانوارها به حساب نیایند. زمانی که چنین داده‌هایی از مجموعه داده‌ها کاسته شود، بر حجم نمونه تأثیر می‌گذارند. فراداده و حوزه مهارت در حل و فصل داده‌های بریده شده بسیار حیاتی هستند، در غیر این صورت آشکارسازی چنین اریبی‌هایی مشکل خواهد بود. نوع دیگر داده‌ها، داده‌های سانسور شده نامیده می‌شود. چنین داده‌هایی

معمولاً در طول تحلیل تا هنگامی که زمان تا اتفاق نوع خاصی از پیشامد (زمان سرآمدن قبل از زلزله بعدی، زمان تا هنگامی که یک ماشین کار می‌کند قبل از این که از کار بیفتد، زمان تا قبل از نمایان شدن نشانه‌های بیماری، زمان تا انجام موفقیت‌آمیز) به وقوع بپیوند، مورد مطالعه قرار می‌گیرند. در این مورد تأکید بر روی درک توزیع احتمال فواصل زمانی است که قادر باشد دوره‌های با شانس بالا در زمانی که پیشامد مورد نظر می‌تواند اتفاق افتد را جدا سازد. لذا در مورد چنین داده‌هایی معنی گمشده یا ناقص بودن در فواصل اتفاق می‌افتد. برای مثال، یک بیمار ممکن است در طول مطالعه علایمی از خود نشان ندهد و در نتیجه داده‌ها بریده نباشند. برای مثال، می‌دانیم که نشانه‌ها برای حداقل دو سال اتفاق نمی‌افتند، اما ممکن است بعد از دو سال و یک هفته یا بعد از ۱۰ سال صعود کند. همان‌گونه که ملاحظه می‌شود قادر به مشاهده تفاوت بین این دو نیستیم. بریدگی و سانسور داده‌ها به‌شیوه‌های قابل انتظار و غیر قابل انتظار اتفاق می‌افتد. یک مثال هشدار دهنده سانسور زمانی است که قالب‌های زمانی به صورت پیش‌فرض در نظر گرفته می‌شوند. چنین مثال‌هایی از سانسورهای مستند نشده می‌تواند با کمک هسیتوگرام و توزیع‌های فروانی آشکار شود. وجود چنین داده‌هایی نتایج را بسیار غیر قابل پیش‌بینی می‌کنند. به‌عنوان مثال ممکن است حذف افرادی که در سال کمتر از ۲۰ هزار تومان هزینه می‌کند عملی باشد. اما، اگر تحلیل‌گر داده‌ها از بریدگی داده‌ها آگاه نباشد و یا تعداد رکوردهای بریده شده در این روش را نداند، تحلیل‌های اشتباهی را ارائه خواهد کرد. با توجه به ناپیوستگی بین جمع‌آوری داده‌ها و داده‌کاوی، فراداده در زمینه داده‌های ناقص بسیار با اهمیت است [۷].

۱۰- داده‌های دورافتاده

اکثر پایگاه داده‌های دنیای واقعی شامل تعدادی مقادیر استثنائی هستند که معمولاً به‌عنوان داده‌های دورافتاده نام‌گذاری شده‌اند. دورافتادگی نقاط دورافتاده از دو جهت، یکی افزایش کیفیت داده‌های اصلی و دیگری کاهش تأثیر مقادیر دورافتاده در فرایند کشف دانش در پایگاه داده‌ها دارای اهمیت می‌باشد. اکثر روش‌های موجود کشف داده‌های دورافتاده بر اساس بررسی دستی داده‌های ارائه شده پایه‌ریزی شده‌اند [۶].

تعریف آماری یک داده دورافتاده به توزیع متغیر در مسئله بستگی دارد. مندل‌هال از عبارت « داده‌های دورافتاده » برای مقادیری که خیلی از میانه توزیع در هر جهت دورافتاده‌اند استفاده می‌کند. این تعریف اساساً به متغیرهایی با مقادیر پیوسته با یک تابع چگالی احتمالی همواره محدود می‌شود. به هر حال فاصله عددی تنها عامل در کشف نقاط دورافتاده پیوسته نیست. اهمیت فراوانی نقاط دورافتاده در یک تعریف کاملاً متفاوت توسط پایل تأکید شده است. « یک داده دورافتاده، تک یا دارای فراوانی خیلی کم است. وقوع مقدار یک متغیر دورتر از بخش عمده مقادیر متغیر است » فراوانی وقوع، یک معیار مهم برای کشف داده‌های دورافتاده در داده‌های رسته‌ای (اسمی) است و یک معیار مشترک مناسب در پایگاه داده‌های دنیای واقعی است. یک تعریف کلی‌تر از یک داده دورافتاده عبارت است از:

یک مشاهده (یا زیرمجموعه‌ای از مشاهدات) که به‌نظر می‌رسد با بقیه مجموعه داده‌ها سازگاری ندارد.

دلیل واقعی وقوع نقطه دورافتاده معمولاً ناآگاهی کاربران داده‌ها یا تحلیل‌گران است. گاهی اوقات یک مقدار ناقص در نتیجه کیفیت پایین یک مجموعه داده‌ها یعنی خطای ورود داده‌ها یا تبدیل داده‌ها می‌باشد. اندازه‌گیری‌های فیزیکی به‌خصوص وقتی که با استفاده از تجهیزات اجرا شود، ممکن است تعداد مشخصی مقادیر نادرست تولید کند. در این موارد هیچ اطلاع سودمندی از طریق نقاط دورافتاده منتقل نمی‌شود. به هر حال ممکن است یک نقطه دورافتاده مقدار درستی باشد. به‌عنوان مثال، اگر خوشه‌های نقاط دورافتاده از نوسانات در رفتار فرایند کنترل نتیجه شود، مقادیر آن‌ها برای کنترل فرایند اهمیت دارند [۶].

۱۱- چرا نقاط دورافتاده باید مجزا شوند؟

دلیل اصلی جداسازی نقاط دورافتاده همبسته بودن آن با تضمین کیفیت داده‌ها است. مقادیر استثنایی با احتمال زیادی نادرست هستند. مطابق با تعریفی که توسط واند و وانگ ارائه شده است داده‌های غیر معتبر یک ناهماهنگی را میان وضعیت پایگاه داده‌ها و

وضعیت دنیای واقعی ارائه می‌کند. بنا بر این برداشتن و جایگزینی داده‌های دورافتاده می‌تواند کیفیت داده‌های ذخیره شده را بالا ببرد. علاوه بر آن جدا کردن داده‌های دورافتاده ممکن است تأثیر مثبتی بر نتایج تحلیل‌های داده‌ها و داده‌کاوی داشته باشد. براوردهای آماری ساده مانند میانگین نمونه و انحراف استاندارد ممکن است به‌طور معنی‌داری با داده‌های دورافتاده تکی که خیلی از میانه توزیع دور هستند، اریب شوند. در مدل‌های رگرسیونی، داده‌های دورافتاده ممکن است روی ضریب همبستگی براورد شده اثر بگذارند. وجود نقاط دورافتاده در آزمون‌ها و تست‌ها برای روش‌های یادگیری درخت تصمیم ممکن است مشکلاتی به‌همراه داشته باشد که توسط میتچل شرح داده شده است. به‌عنوان مثال، به‌کارگیری یک مقدار دورافتاده پیش‌بینی‌کننده صفت اسمی ممکن است لزوماً تعداد شاخه‌های درخت تصمیم مرتبط با این صفت را افزایش ندهد. در عوض منجر به محاسبه نادرست معیار انتخاب صفت شود. در نتیجه ممکن است درستی پیش‌بینی‌کننده نتایج درخت تصمیم کاهش یابد. جداسازی نقاط دورافتاده یک مرحله مهم در آماده‌سازی مجموعه داده‌ها برای هر نوع تحلیل برروی داده‌ها است [۶].

یک روش عینی– کمی برای کشف نقاط دورافتاده عددی بدون بررسی، روش گرافیکی بر اساس نمودار جعبه‌ای است که میانه همه مشاهدات و چارک اول و چارک سوم را ارائه می‌دهد، پایه‌ریزی شده است. انتظار می‌رود که بیش‌تر مقادیر در دامنه میان چارکی (H) جای گرفته باشند. مقادیری که در بازه $\pm ۱/۵H$ قرار می‌گیرند اصطلاحاً « نقاط دورافتاده خفیف » و مقادیری که خارج از کرانه‌های $\pm ۳H$ قرار می‌گیرند، اصطلاحاً « نقاط دورافتاده کرانگین » نامیده می‌شوند. از آن‌جایی که این روش یک شق عملی بازرسی دستی هر نمودار جعبه‌ای را ارائه می‌دهد، می‌تواند تنها با متغیرهای پیوسته‌ای با توزیع‌های احتمال یک نمایی سر و کار داشته باشد. طبقه‌بندی به سرعت با حرکت یک مقدار در طول یکی از $۱/۵H$ یا کرانه‌های $۳H$ تغییر می‌کند [۶].

یک روش برای کشف داده‌های نادرست استفاده از یکی از فنون داده‌کاوی (مثل شبکه عصبی یا درخت تصمیم) برای ایجاد یک مدل پیش‌بینی است. اکثر « الگوهای شگفت‌انگیز » (با کمترین احتمال برای پیش‌بینی درست توسط مدل) قابل اعتماد نیستند و با آن‌ها به‌عنوان داده‌های دورافتاده باید برخورد شود. به هر حال این روش، این حقیقت

را که انطباق داده‌ها ممکن است وابسته به توزیع اصلی صفت‌های پایگاه داده‌ها و بعضی از فاکتورهای ذهنی و وابسته به کاربر باشد، نادیده می‌گیرد [۶].

یک راه بدیهی این است که رکوردهای دورافتاده را اگر حداقل یک مشاهده دورافتاده دارند از تحلیل بیرون بگذاریم. این، مشابه صرف نظر کردن از رکوردهای دارای مقادیر گمشده توسط برخی از روش‌های داده‌کاوی می‌باشد روش دیگر، تصحیح خطاهای فوق با استفاده از مقادیر دیگر می‌باشد.

یکی از روش‌های مورد استفاده در داده‌کاوی خوشه‌بندی است. اغلب اطلاعات و داده‌های موجود در پایگاه‌های داده‌ها توزیع‌های ناشناخته یا پیچیده‌ای دارند که به راحتی نمی‌توان آن توزیع‌ها را شناسایی نمود و مورد استفاده قرار داد. بنا بر این برای تحلیل داده‌ها و اطلاعات موجود در پایگاه‌های داده‌ها استفاده از روش‌هایی که نیاز به دانستن توزیع متغیرها ندارد از اهمیت خاصی برخوردار است. خوشه‌بندی یکی از روش‌هایی است که با توزیع داده‌های موجود سر و کار نداشته و اغلب با استفاده از معیارهای تشابه و عدم تشابه به خوشه‌بندی داده‌ها می‌پردازد [۱] و [۳].

دلایل زیادی را می‌توان برای ارزشمند بودن تحلیل خوشه‌ای ارائه کرد:

الف) تحلیل خوشه‌ای می‌تواند در یافتن گروه‌های واقعی در پایگاه‌های داده‌ای موجود به کاربران کمک کند؛

ب) تحلیل خوشه‌ای می‌تواند برای کاهش داده‌ها به کار برده شود؛

پ) تحلیل خوشه‌ای می‌تواند در شناسایی نقاط دورافتاده مورد استفاده قرار گیرد؛

ت) تحلیل خوشه‌ای با توزیع داده‌ها سر و کار ندارد.

به‌عنوان یک وظیفه داده‌کاوی، تحلیل خوشه‌ای می‌تواند مانند یک ابزار که به تنهایی عهده‌دار بینش در توزیع داده‌ها است، برای مشاهده مشخصه‌های هر خوشه و تمرکز روی یک مجموعه ویژه از خوشه‌ها و تحلیل بیش‌تر به کار رود. به‌عنوان راهی دیگر، تحلیل خوشه‌ای ممکن است به‌عنوان یک مرحله پیش‌پردازش برای الگوریتم‌های دیگر به کار رود به‌طوری که مشخصه‌ها و رده‌بندی‌ها ممکن است روی خوشه‌های کشف شده اعمال شوند.

تمرکز اصلی تحلیل خوشه‌ای عمدتاً بر اساس فاصله است. ابزارهای تحلیل خوشه‌ای بر اساس k میانگین، k مدوئید و همچنین چند روش دیگر، در سیستم‌ها یا بسته‌های نرم‌افزار تحلیل آماری موجود است. S-Plus, SAS, SPSS چند نمونه از این نرم‌افزارها می‌باشند [۱].

در داده‌کاوی تلاش‌های پژوهشگران برای یافتن روش‌های تحلیل خوشه‌ای کارامد و مؤثر در پایگاه‌های داده‌ای بزرگ متمرکز است. خوشه‌بندی یک شاخه پژوهشی با چالش‌های فراوان است که کاربردهای بالقوه آن ملزومات ویژه خودش را مطرح می‌کند. ملزومات کلی خوشه‌بندی در داده‌کاوی عبارتند از:

مقیاس‌پذیری، توانایی کار با ویژگی‌های مختلف، کشف خوشه‌ها با شکل دلخواه، مینیمم ملزومات برای این که قلمرو دانش بتواند پارامترهای ورودی را تعیین کند، توانایی کار با داده‌های نوفه‌ای، عدم حساسیت به‌ترتیب وارد شدن داده‌های ورودی، بعد زیاد، خوشه‌بندی مبتنی بر محدودیت و قابلیت تفسیر و قابلیت کاربرد.

یکی از مهم‌ترین ملزومات خوشه‌بندی قابلیت تفسیر و قابلیت کاربرد روش خوشه‌بندی می‌باشد، در حقیقت بعد از خوشه‌بندی کاربر توقع دارد که نتایج خوشه‌بندی قابل تفسیر، قابل فهم و قابل استفاده باشد. یعنی ممکن است خوشه‌بندی به تفسیرهای معنایی مشخص و کاربردهای مرتبط با آن نیاز داشته باشد. با در نظر گرفتن این نیازها، مطالعه تحلیل خوشه‌بندی به صورت زیر نتیجه می‌شود [۱].

۱۲- انجام داده‌کاوی بر روی داده‌های طرح هزینه و درامد خانوار

۱-۱۲- داده‌کاوی داده‌های طرح هزینه و درامد خانوار با استفاده از روش پردازش تحلیل آنی (OLAP)

برای بررسی داده‌های حجیم یکی از روش‌های مناسب استفاده از جدول‌ها و اطلاعات تولید شده با روش پردازش تحلیلی آنی (OLAP) است، این روش قابلیت استخراج جدول‌های زیادی را دارد به‌طوری که می‌توان با تغییر دادن وضعیت متغیرهای گروه‌بندی به‌سرعت اطلاعات سایر متغیرها را در سطوح مختلف متغیرهای گروه‌بندی نمایش داد.

OLAP استخراج‌ها و نمایش‌های دلخواه از داده‌ها را از دیدگاه‌های مختلف مد نظر قرار می‌دهد. این دیدگاه‌ها عموماً به‌عنوان بعد معروف هستند. هر بعد معمولاً می‌تواند چند مرحله از انباشتگی را دارا باشد. به‌عنوان مثال می‌توان آمارهای خلاصه مربوط به هر یک از بخش‌های طرح هزینه و درامد خانوار یا هر یک از اقلام مصرفی خانوار را بر اساس کدهای ۶ رقمی، ۵ رقمی و... استخراج نمود. OLAP ابزاری برای خلاصه‌سازی و فشرده کردن داده‌ها است و از این روش می‌توان برای تشخیص ماهیت داده‌ها با استفاده از معیارهای تمرکز، معیارهای پراکندگی و... استفاده نمود. به‌عنوان مثال برای پاسخ دادن یا بررسی این که میانگین، دامنه تغییرات و... هر یک از اقلام چقدر است و چرا این اتفاق افتاده است؟ استفاده از این روش خیلی ساده‌تر و جامع‌تر از روش‌هایی است که در حال حاضر برای بررسی پایگاه داده‌های طرح هزینه و درامد خانوار و مشخص نمودن داده‌های دورافتاده و... استفاده می‌شود.

استفاده از OLAP در بررسی دو پایگاه داده‌ای شهری و روستایی طرح هزینه و درامد خانوار مشخص نمود که در برخی از اقلام هزینه برخی از خانوارها دارای مقادیر خیلی بزرگ‌تری نسبت به سایر اعضای نمونه هستند و این امر سبب شده تا میانگین هزینه اقلام به صورت چشم‌گیری افزایش یابد. استفاده از این روش می‌تواند به سرعت داده‌هایی که ممکن است دورافتاده باشند را معرفی نماید (برای راهنمایی بیش‌تر به گزارش طرح مطالعاتی داده‌کاوی و کاربرد آن در کیفیت داده‌ها مراجعه کنید).

۲-۱۲- داده‌کاوی داده‌های طرح هزینه و درامد خانوار با استفاده از ابزار ساده آماری

با استفاده از ابزار آماری ساده‌ای چون دستور Explore در نرم‌افزار SPSS آماره‌های خلاصه‌ای استخراج شد که نشان داد داده‌ها در بخش‌های پرسشنامه از توزیع نرمال پیروی نکرده و دارای چولگی هستند. همچنین ابزار ساده فوق بزرگ‌ترین و کوچک‌ترین ۵ مقدار کرانگین هزینه‌ها و درامدها را نشان می‌دهد (برای راهنمایی بیش‌تر به گزارش طرح مطالعاتی داده‌کاوی و کاربرد آن در کیفیت داده‌ها مراجعه کنید).

۳-۱۲- داده‌کاوی داده‌های طرح هزینه و درامد خانوار با استفاده از روش‌های خوشه‌بندی

با توجه به یافته‌های تحلیل، اغلب اطلاعات و داده‌های موجود در طرح هزینه و درامد خانوار توزیع‌های ناشناخته یا پیچیده‌ای دارند که به راحتی نمی‌توان آن توزیع‌ها را شناسایی نمود و مورد استفاده قرار داد. بنا بر این برای تحلیل داده‌ها و اطلاعات فوق، استفاده از روش‌هایی که نیاز به دانستن توزیع متغیرها ندارد از اهمیت خاصی برخوردار است. خوشه‌بندی یکی از روش‌هایی است که به توزیع داده‌های موجود وابسته نمی‌باشد.

اغلب در پایگاه‌های داده‌ها با حجم بالا تعداد متغیرها، حجم داده‌ها یا هر دو خیلی زیاد است که طرح هزینه و درامد خانوار نمونه بارزی از آن است، برای خوشه‌بندی متغیرها و داده‌ها در این پایگاه‌ها یک روش بسیار مفید و ارزشمند تحلیل خوشه‌ای است.

از آن‌جا که تعداد متغیرها و همچنین تعداد خانوار نمونه در طرح هزینه و درامد خانوار زیاد می‌باشد، استفاده از روش‌های مرسوم خوشه‌بندی پیشنهاد نمی‌شود. لذا برای انجام تحلیل خوشه‌ای از الگوریتم کلارا که یکی از روش‌های خوشه‌بندی پم (بخش‌بندی اطراف مدوئید) و از جمله روش‌های خوشه‌بندی در داده‌کاوی است استفاده شد. این روش‌ها در مقایسه با روش‌های خوشه‌بندی سلسله‌مراتبی نیاز به صرف وقت خیلی کمتر و امکانات سخت‌افزاری معمولی دارد. از آن‌جا که در روش خوشه‌بندی تعداد خوشه‌ها را نمی‌توان با استفاده از استدلال علمی از قبل تعیین کرد، لذا از حداقل انتخاب تعداد خوشه شروع به خوشه‌بندی نموده و این کار را تا آن‌جا که نمودار سایه‌نما مناسب به دست آید (ماکسیمم مقدار میانگین سایه‌نما) ادامه دادیم.

برای خوشه‌بندی داده‌ها از آن‌جا که مقدار داده‌ها برای اقلام مصرفی دارای تغییرپذیری زیادی است و این عامل می‌تواند نقش مهمی را در خوشه‌بندی داده‌ها داشته باشد لذا برای این که تأثیر تغییرپذیری داده‌ها را از خوشه‌بندی حذف کنیم، قبل از خوشه‌بندی ابتدا داده‌ها را استاندارد کردیم.

شایان ذکر است که روش‌های خوشه‌بندی‌ای هم وجود دارند که تعداد خوشه بهینه را به دست می‌دهند، ولی این روش‌ها به فرض‌هایی نیاز دارند که در نظر گرفتن آن فرض‌ها روی داده‌های طرح هزینه و درامد خانوار و خیلی از داده‌های دیگر امکان‌پذیر

نیست، به‌عنوان مثال در خوشه‌بندی دومرحله‌ای که تعداد خوشه‌های بهینه را ارائه می‌کند یکی از فرض‌های مورد نیاز، نرمال بودن داده‌ها (نرمال یک یا چند متغیره) می‌باشد که ما نمی‌توانیم در طرح هزینه و درامد خانوار چنین فرضی را در نظر بگیریم چرا که با توجه به نمودار بافت‌نگار داده‌ها ملاحظه می‌شود که داده‌ها دارای چولگی شدیدی از نرمال بودن هستند، بنا بر این اصلی‌ترین فرض مورد نیاز برای داده‌ها برقرار نیست. با بخش‌بندی داده‌ها و محاسبه میانگین، میانه، مینیمم و ماکسیمم و... می‌توان به اکتشاف، دیداری کردن، انتشار داده‌ها، برازش مدل اکتشافی و پاک‌سازی داده‌ها پرداخت که در حقیقت این امر باعث کاهش داده‌ها با استفاده از آماره‌های خلاصه می‌باشد. همچنین با بخش‌بندی داده‌ها به خوشه‌هایی دست می‌یابیم که از اندازه (حجم) و تغییرپذیری کمتری برخوردار هستند (برای راهنمایی بیش‌تر به گزارش طرح مطالعاتی داده‌کاوی و کاربرد آن در کیفیت داده‌ها مراجعه کنید).

۱۳- جانهی اقلام بی‌پاسخ طرح هزینه و درامد خانوارهای شهری و روستایی

از روی فایل‌های مربوط به داده‌های طرح هزینه و درامد خانوار مربوط به فصل ۱ سال ۱۳۸۴ که بی‌پاسخی‌ها در آن با @ مشخص شده بودند یک فایل جدید ساخته شد که تنها شامل طرز تهیه یک (کد خرید) برای اقلام مصرفی بود. سپس @ها به بی‌پاسخی تبدیل شد. این‌کار فقط برای داده‌های سال ۱۳۸۴ قابل استفاده است چرا که روش تکمیل پرسشنامه‌های سال‌های قبل امکان بررسی بی پاسخی قلم را نمی‌داد. بررسی‌های صورت گرفته نشان داد که بسیاری از موارد بی‌پاسخ در صورتی که طبق دستورالعمل طرح عمل می‌شد دارای پاسخ بودند. به‌طور مثال بیمه تحصیلی دانش‌آموزان در سال ۱۳۸۴ برای هر دانش‌آموز ۷۰۰۰ ریال بوده است، در صورتی که تعداد زیادی از پرسشنامه‌های تکمیل شده در استان تهران برای قلم مذکور بی‌پاسخ بوده‌اند. با این حال پس از جانهی موارد بی‌پاسخی اقلام مشخص شد که جانهی موارد فوق، به‌علت این که در اکثر حالات موارد بی‌پاسخی مربوط به اقلامی است که سهم کمی در براوردهای هزینه دارند، تأثیر بسیار اندکی در براورد میانگین خواهد داشت.

برای جانهی و براورد هر یک از بی‌پاسخی‌ها (@ها)، ابتدا خانوارها بـر اسـاس سـایر اقلام در سطح کد ۶ رقمی که مرتبط با متغیر مربوط بودند خوشه‌بنـدی شـدند. بـه‌عنـوان مثال اگر در زیرمجموعه اقلام مربـوط بـه کـدهای ۰۱۱۱۲ و ۰۱۱۱۱ قـسمت غـلات در بخش هزینه‌های خوراکی، یکی از خانوارها به قلمی پاسخ نداده بود خانوارها را بـر اسـاس سایر متغیرهای آن زیرمجموعه خوشه‌بندی کردیم سـپس بـر اسـاس اطـلاع مربـوط بـه عضویت آن خانوار به خوشه مربوط، آمار خلاصه آن قلم به‌تفکیک خوشه‌ها محاسـبه شـد سپس جانهی بر اساس این که خانوار به کدام خوشه تعلق دارد انجام شد و در صورتی کـه این کار امکان‌پذیر نبود از میانگین قلم در خانوارهای نمونـه اسـتان در فـصل مـورد نظـر استفاده شد.

یکی از دلایل براورد بی‌پاسخی و جانهی اقلام این اسـت کـه اگـر خـانواری اطـلاع مربوط به قلمی را پاسخ نداده باشد الگوریتم‌های خوشـه‌بنـدی آن خـانوار را در محاسـبات خود حذف خواهند کرد. بنا بر این، برای این که اطلاعات مربوط به خانوار از تحلیل حذف نشود نیاز به براورد بی‌پاسخی قبل از خوشه‌بندی نهایی ضروری بـه‌نظـر مـی‌رسـد (بـرای راهنمایی بیش‌تر به گزارش طـرح مطالعـاتی داده‌کـاوی و کـاربرد آن در کیفیـت داده‌هـا مراجعه کنید).

۱۴- شناسـایی و جـانهی (اصـلاح) هزینـه بخـش‌هـای دارای مـشاهده دورافتاده با استفاده از خوشه‌بندی

پس از جانهی بی‌پاسخی‌ها، سـرجمع‌هـای هـر بخـش پرسـشنامه را محاسـبه کـرده، بـه خوشه‌بندی هزینه‌هـای مربـوط بـه هـر بخـش پرسـشنامه پرداختـه شـد، بـدین منظـور خوشه‌بندی با استفاده از متغیرهایی چون بعد خانوار، فصل آمارگیری و هزینه هر بخـش و نیز هر ۱۳ بخش به‌طور کلی در سطح مناطق روستایی کل کـشور و اسـتان تهـران بـرای تمامی خانوارهای نمونه فصل ۱ سال ۱۳۸۴ به‌عنوان نمونه انجام شد. بـرای ایـن کـار بـا استفاده از الگوریتم خوشه‌بندی کلارا، خانوارها در بخش‌های هزینه خوشه‌بندی شدند کـه بر اساس ضریب سایه‌نما که تعلق داده‌ها به خوشـه‌هـا را نمـایش مـی‌دهـد، بـزرگ‌تـرین میانگین ضریب سایه‌نما که بیانگر بهترین تعداد خوشه‌ها می‌باشد انتخاب شد. سپس داخل

خوشه‌ها برای هر بخش از روش‌های ذکر شـده اسـتفاده و داده‌هـای دورافتـاده شناسـایی شدند.

پس از انتخاب تعداد خوشه بهینه، آماره‌های خلاصه مربوط بـه هـر بخـش از قبیـل تعداد خانوارهای متعلق به هر خوشه، میانگین، میانه، چارک‌ها، دهک‌ها، انحراف اسـتاندارد و بافت‌نگار استخراج شد که با استفاده از آن‌ها می‌توان مشخص نمود آیا خوشه‌هـا شـامل مشاهده دورافتاده هستند یا خیر.

در خوشه‌بندی نیز ممکن است برخی از خوشه‌ها نسبت به سـایر خوشـه‌هـا اعـضای کمتری داشته باشند، این خوشه‌ها ممکن است حاوی مشاهده‌های دورافتاده باشـند. بـرای شناسایی و تصمیم‌گیری در رابطه با آن‌ها می‌توان داده‌هـای هـر بخـش کـه در خـارج از فاصله ۱/۵ برابر دامنه میان‌چارکی قرار دارند را به‌عنوان مشاهدات دورافتاده بالقوه معرفـی نمود. سپس با استفاده از سایر فراداده‌های موجود و مـرتبط بـا خـانوار در فایـل داده‌هـای مربوط به خانوار نظر قطعی در مورد دورافتاده بودن یا نبودن اطلاعات خانوار داد.

شایان ذکر است پس از بررسی‌های انجام شده مشخص شد که اکثر محاسبات تحت تأثیر متغیر فصل و انتخاب مناطق شهری و روستایی استان و یا کل کشور به‌عنوان سطح جغرافیایی هستند.

خانوارهای شناسایی شده بـه‌عنـوان مـشاهده دورافتـاده، در دو روش قبـل و بعـد از خوشه‌بندی در مواردی با هم متفاوت بودند ضمن این که اطلاعاتی که برای اصلاح آن‌ها از داده‌ها نتیجه می‌شود نیز با هم تفاوت داشتند. می‌تـوان گفـت کـه خوشـه‌بنـدی باعـث شناسایی داده‌هایی شده است که بدون استفاده از خوشه‌بندی به‌نظر مـی‌رسـید داده‌هـایی باشند که شک‌برانگیز نیستند. با استفاده از برخی داده‌های بنـد ۸ قـسمت دوم پرسـشنامه (که از سال ۱۳۸۴ به پرسشنامه هزینه و درامد اضافه شده‌اند) و نیـز رجـوع بـه داده‌هـایی چون درامد (هزینه کل) و هزینه غیر خوراک خـانوار مـی‌تـوان تـصمیم‌گیـری نمـود کـه داده‌های مذکور به درستی دورافتاده هستند یا خیر. در صورتی که پـس از بررسـی داده‌هـا کارشناس موضوعی تشخیص دهد که داده دورافتاده است، مـی‌تـوان بـرای حـل مـشکل بسته به مورد، اقدام به رفع مشکل از طریق:

الف) استفاده از نظر کارشناس اجرایی استان پس از رجوع به پرسشنامه خانوار؛

ب) اصلاح (جانهی) مقدار دورافتاده با استفاده از اطلاعات خانوار (یا خانوارهای) مشابه در خوشه مورد نظر؛
پ) حذف مقدار دورافتاده برای خانوار مربوط؛
ت) حذف اطلاعات کل خانوار مربوط از تحلیل کرد.

برای بررسی نقاط دورافتاده درامد نیز پس از شناسایی آن‌ها با استفاده از خوشه‌بندی در هر فصل و به‌تفکیک مناطق شهری و روستایی هر استان، با رجوع به سایر اطلاعات افراد شاغل و دارای درامد خانوار و نیز با مقایسه با هزینه کل خانوار می‌توان تصمیم‌گیری نمود که داده‌های فوق واقعاً دورافتاده می‌باشند یا خیر. در صورت تشخیص داده‌های فوق به‌عنوان داده‌های واقعاً دورافتاده می‌توان از همان راهکارهایی که برای مقابله با داده‌های دورافتاده هزینه ذکر شد، استفاده نمود.

۱۵- مشکلات و پیشنهادها

۱-۱۵- مشکلات در فراخوانی فایل داده‌های طرح هزینه و درامد شهری و روستایی در نرم‌افزارهای آماری

برای داده‌کاوی در داده‌های طرح آمارگیری از هزینه و درامد خانوار می‌باید این داده‌ها به قالبی تبدیل شود که برای استفاده در نرم‌افزارهای آماری مانند SAS و SPSS قابل استفاده باشد زیرا تحلیل اطلاعات رابطه‌ای در نرم‌افزارهای آماری به سهولت امکان‌پذیر نیست و اصطلاحاً اطلاعات باید به قالب تخت و در یک رکورد چیده شود. از آن‌جایی که فایل داده‌های هزینه و درامد خانوار در قالب Access تهیه شده است در تبدیل فایل‌های موجود هزینه و درامد خانوار برای استفاده در نرم‌افزارهای آماری، به‌علت حجم انبوه داده‌ها مشکلات زیادی وجود داشت که زمان و نیروی زیادی را صرف نمود.

البته امکان تبدیل فایل‌های با قالب بانک داده، در نرم‌افزارهای آماری وجود دارد. اما حجیم بودن داده‌های این طرح و نیاز به سخت‌افزار قوی باعث شد که مجبور شویم فایل‌های بانک داده‌ای هزینه و درامد خانوار را در دو فایل داده‌ها و کدهای اقلام با قالب متنی تهیه نموده و سپس از طریق درون بردکردن فایل داده‌ها و فایل اقلام در

نرم‌افزارهای آماری آن‌ها را ادغام نماییم. مشکل دیگر در خواندن فایل حاوی اقلام بدون پاسخ (که با علامت @ مشخص شده‌اند) در نرم‌افزارهای آماری است. نرم‌افزارهای آماری پس از خواندن فایل دارای @ آن‌ها را تبدیل به صفر می‌کنند. با توجه به مشکلات مربوط به کدهای حرفی در تبدیل فایل‌ها به یکدیگر، همچنین توصیه‌های مربوط به استانداردهای آماری پیشنهاد می‌شود از کدهای عددی مانند ۱- و... به‌جای کدهای حرفی استفاده شود.

با وجود این که در این طرح مشکل تبدیل نرم‌افزار داده‌ها پس از صرف مشکلات و زمان زیاد به‌نحوی حل شد اما به‌نظر می‌رسد اگر حل مسئله به راحتی مقدور نیست باید برای حل آن فکری اساسی نمود. به عبارت دیگر، در صورت نبود راهکار ساده، باید اولویت نرم‌افزارهای داده‌آمایی برای داده‌های طرحی مانند هزینه و درامد خانوار که نیاز به پردازش‌های زیادی دارد با نرم‌افزارهای آماری باشد. چرا که در انجام تغییرات قالب فایل‌ها، ممکن است تغییراتی در داده‌ها از جمله حذف برخی از رکوردها و یا جابه‌جایی در فیلدهای آن‌ها صورت گیرد.

۲-۱۵- وجود قواعد و استثناهای زیاد و نیاز به فراداده‌های کنترلی بیش‌تر برای تحلیل

یک چالش دیگر در داده‌کاوی داده‌های طرح هزینه و درامد خانوار پیچیدگی‌های ناشی از استثناهای زیاد قواعد مربوط به کنترل داده‌های این طرح است که برخی از آن‌ها به‌علت ماهیت طرح فوق است و تا حدود زیادی وابستگی به عامل انسانی را طلب می‌کند. همچنین به‌نظر می‌رسد که برای انجام داده‌کاوی مؤثرتر ایجاد ارتباط بیش‌تری بین اقلام هزینه نیاز می‌باشد و بدین منظور نیاز است که فراداده‌های کنترلی بیش‌تری از خانوار پرسیده و به کار گرفته شود. هر چند که قدم‌های مثبتی در این راه برداشته شده است (اضافه شدن بند ۸ قسمت دوم پرسشنامه از سال۱۳۸۴).

۳-۱۵- مشکلات مربوط به خطاهای داده‌آمایی در طرح هزینه و درامد خانوارهای شهری و روستایی

یک منبع عمده در خطاهای کیفیت داده‌ها وارد کردن دستی داده‌ها و مداخله دستی است. با توجه به این که دقت و سرعت انتشار نتایج از ابعاد کیفیت بوده و نتایج داده‌کاوی و حل مشکلات داده‌های طرح هزینه و درامد خانوارهای شهری و روستایی نشان می‌دهد که در بسیاری از موارد، کارشناسان اجرایی استان‌ها اشتباه بودن مقادیر را مربوط به خطاهای داده‌آمایی می‌دانند و از آن‌جایی که همواره پیشگیری، آسان‌تر و مفیدتر از درمان بوده و انجام عملیات داده‌آمایی، وریف و بازبینی داده‌ها در استان و حل سایر مشکلات ادیتی طرح در مرکز آمار ایران زمان زیادی را می‌طلبد. لذا پیشنهاد می‌شود آمارگیری این طرح با استفاده از رایانه و نرم‌افزارهای داده‌آمایی و بازبینی همزمان همراه شود تا علاوه بر دقت، سرعت بیش‌تری در نتایج طرح حاصل شود و انجام مکاتبات استانی و زمان انتظار برای بررسی و پاسخ به سوالات به حداقل برسد. همچنین برای نرم‌افزار داده‌آمایی در استان تمهیداتی در نظر گرفته شود که کارهای بسیار ساده آماری همچون مشخص کردن چند مقدار بالا از هر قلم در هر فصل آمارگیری و قابلیت مرتب کردن ارزش (برای اقلام خوراکی و غیر خوراکی) و مقدار (برای اقلام خوراکی) به صورت صعودی را داشته باشد تا کارشناس اجرایی طرح در استان بررسی مجددی روی آن‌ها انجام داده و صحت داده‌ها را مورد تأیید قرار دهد. این عمل باعث می‌شود تا کنترل داده‌ها توسط کارشناسان موضوعی مرکز آمار ایران به حداقل برسد و از هدر رفتن زمان و نیروی کارشناسی زیادی جلوگیری شده و دقت داده‌ها و سرعت انتشار افزون شود.

٤-۱۵- اتوماسیون کامل

کیفیت داده‌ها نمی‌تواند به‌طور کامل با مجموعه‌ای از اعداد و قواعد تسخیر شود. این قواعد باید به‌طور مداوم مورد بررسی دقیق قرار گرفته، به روز و مستند شوند و لذا باید سیستم‌هایی پویا برای نشان دادن، مجزا کردن و تصحیح داده‌هایی که در مجموعه قواعد قرار نمی‌گیرند طراحی شود. بنا بر این، ایجاد یک راه حل کلی با استفاده از مجموعه‌ای از

محدودیت‌ها برای کیفیت داده‌ها برای همیشه که در آن به‌طور کاملاً خودکار و بدون دخالت عامل انسانی، داده‌ها بتوانند بررسی دقیق شده و رکوردهایی که نمی‌توانند این فرایند را براورده کنند مجزا شوند، در این طرح تا حدودی به دور از واقعیت است.

مرجع‌ها

[۱] حائری مهریزی، علی‌اصغر. داده‌کاوی: مفاهیم و روش‌ها و کاربردها، پایان‌نامه کارشناسی ارشد، دانشگاه علامه طباطبایی، تهران. ۱۳۸۲. به راهنمایی دکتر علی زندی‌نیا.

[۲] حائری مهریزی، علی‌اصغر و نواب‌پور، حمیدرضا (۱۳۸۱) ششمین کنفرانس بین‌المللی آمار ایران، دانشگاه تربیت مدرس.

[۳] ناظمی، عبدالرضا. رده‌بندی و داده‌کاوی، پایان‌نامه کارشناسی ارشد، دانشگاه فردوسی مشهد، مشهد، ۱۳۸۳. به راهنمایی دکتر علی مشکانی.

[۴] رهنمون، رامین. هماوندی، آناهیتا. هوش مصنوعی « رهیافتی نوین »، ترجمه، چاپ چهارم. انتشارات ناقوس، تهران، ۱۳۸۳.

[۵] هوش مصنوعی رهیافتی نوین، استوارت راسل/ پیتر نورویگ، ترجمه رامین رهنمون/ آناهیتا هماوندی، انتشارات ناقوس، تهران. ۱۳۸۳.

[6] Mark last, Abraham Kandel, Automated detection of outliers in real-world data.

[7] Walter A.Shewhart and Samuel S.Wilks, Exploratory data mining and data cleaning, Wiley Series in probability and statistic.

[8] Fayyad, U.M,Piatetsky Shapiro, Smyth P and Uthurusamy R.(eds.)(1996). Advances in Knowledge Discovery and Data Mining. Menlo Park, California: AAAI Press.

[9] Han, J.,and Kamber.M.(2001). Data Mining: Concept and Techniques. Morgan Kaufmann.

[10] David J. Hand, Heikki Mannila and Padhraic Smyth, Principle of Data Mining, MIT Press, 2001.G. Saporta (2000), Data Mining and Official Statistics, Roma.

[11] Hand, D. J. Why data mining is more than statistics writ large. Statistical aspects of data mining and knowledge discovery in databases.

[12] Jiawei Han and Micheline Kamber, Data Mining: Concept and Techniques, Morgan Kaufmann, 2001.

[13] J. Hipp, U. Guntzer, U. Grimmer (2002), Data Quality Mining.

[14] Mehmed Kantardiz, Data Mining: Concepts, Models, Methods, and Algorithms, John Wiley & Sons, 2003.

[15] Michael Berry and Gordon Lindoff, Mastering Data Mining, John Wiley & Sons, 1997.

[16] Hand, D., Mannila, M., and Padhraic, S. Principle of Data Mining. MIT Press.
[17] M. Hernandez and S. Stolfo. Real-world data is dirty: Data cleansing and the merge/purge problem. Data Mining and Knowledge Discovery, 2(1):9–37, 1998.
[18] T. Dasu, and T. Johnson. Exploratory Data Mining and Data Cleaning. 2003.
[19] Machine Learning, Tom Mitchell, McGraw Hill, 1997.
[20] Introduction machine learning\nils J. Nilsson stanford university\Stanford 1997.
[21] George, H.J. Enhancements to the Data Mining Process. Ph.D.Thesis, Department of Computer Science, Stanford University.
[22] R. J. A. Little and D. B. Rubin. Statistical Analysis with Missing Data. Wiley, NewYork, 1987.
[23] Jerome, H. Data Mining and Statistics: What's the Connection? URL:http://stat.stanford.edu/~jhf/dm-stat.ps.Z.