۱.۱ مقدمه

امروزه داده کاوی به عنوان یک روش جدید برای نظم دادن به پایگاه‌های بزرگ و در حال افزایش داده‌ها ظهور پیدا کرده است. در دنیای امروزی حتی در کارهای ساده‌ای مانند تلفن‌زدن، استفاده از کارت‌های اعتباری یا خریدهای روزانه، جای پای تکنولوژی‌های مدرن دیده می‌شود. افزایش آزمایشات در تمام زمینه‌های علمی و ثبت نتایج آن باعث ذخیره انبوه داده‌ها به حجم چندین پتابایت (هزار ترابایت) شده است. دستگاه‌های جمع‌آوری اتوماتیکی داده که در کسب و کار امروزی مورد استفاده قرار می‌گیرند توانایی تولید ترابایت (هزار گیگابایت) داده در ساعت را دارند. بزرگترین انبار داده[۱] جهان، سیستم والمرت[۲]، ۵۰۰ ترابایت داده را شامل می‌شود که بارگذاری آن بر روی کامپیوترهای بسیار بزرگ است. داده کاوی بدنبال نیاز به دست‌کاری انبار داده بوجود آمد تا الگوهایی منطقی از داده‌ها بدست آورد که ممکن است برای مدیریت سازمان یا شرکت تولید کننده داده مفید باشد. این الگو می‌تواند یک خلاصه‌سازی ساده از داده، طبقه‌بندی داده و یا مدل مربوط به آن باشد. داده کاوی به عنوان کشف دانش [۳] به دنبال کشف دانش مدیریتی از داده‌های خام می‌باشد [۲۲].

۲.۱ داده کاوی چیست؟

هدف داده کاوی در یک حیطه معین علمی، عبارت است از ایجاد درک از حجم زیادی از داده‌هایی که در اغلب موارد به صورت هدایت نشده[۴] جمع‌آوری شده‌اند. این عبارت، یک تعریف مستقیم بوده و اهداف داده کاوی را نیز شرح می‌دهد و فهم آن ساده است. بیشتر افرادی که داده کاوی

[۱] Warehouse

[۲] Walmart System

[۳] Knowledge Discovery

[۴] Unsupervised

را مورد استفاده قرار می‌دهند، اشخاص متخصص در یک زمینه خاص علمی بوده و نه تنها به داده‌ها دسترسی دارند، بلکه خود نیز به جمع‌آوری آن می‌پردازند. افرادی که در فعالیت‌های تجاری مشارکت دارند، بزرگترین گروه استفاده کننده از تکنیک‌ها و مفاهیم داده کاوی را تشکیل می‌دهند. زیرا بطور مستقیم و همیشگی حجم زیادی از داده‌ها را جمع‌آوری کرده و تمایل زیادی دارند که به درک درست و کاملی از داده‌ها برسند. هدف این افراد آن است که شرکت‌های تجاریشان سودآورتر شده و نیز در مقایسه با سایر شرکت‌ها، توان رقابت بیش‌تر و عملکرد موفق‌تر داشته باشند. صاحبان داده نه تنها ترجیح می‌دهند که داده خود را بهتر بشناسند، بلکه مایل هستند تا دانش جدیدی (که در درون داده‌ها پنهان است) را در رابطه با زمینه فعالیت خویش کسب نمایند.

ایجاد درک در تعریف داده کاوی عبارت است از دانش یا مدلی که بتوان آن را به کمک عبارت ساده (به عنوان مثال، از طریق قوانین) توصیف نمود. عبارت کلیدی که در تعریف داده‌کاوی به آن اشاره شده است، حجم زیادی از داده‌ها می‌باشد. هدف از داده‌کاوی تحلیل مجموعه‌های کوچک از داده‌ها نیست، زیرا، می‌توان آن‌ها را به کمک بسیاری از تکنیک‌های استاندارد تحلیل کرد یا آنکه آن‌ها را به صورت دستی انجام داد. با چند مثال منظور خود را از حجم مناسب داده بیان می‌کنیم. شرکت مخابراتی ATT روزانه بیش از سیصد میلیون تماس تلفنی را دریافت و به حدود صد میلیون نفر از مشتریان خود خدمت‌دهی می‌کند. این شرکت، اطلاعات خود را در یک پایگاه داده چندین ترابایتی ذخیره می‌سازد. $Wal-Mart$ روزانه در حدود بیست و یک میلیون تراکنش را در تمامی شعب خود مدیریت کرده و اطلاعات بدست آمده را در یک پایگاه داده چند ده ترابایتی ذخیره می‌سازد.

روشن است که هیچ کدام از پایگاه‌های داده‌ای را که در بالا به آن اشاره شد نمی‌توان به کمک نیروی انسانی یا حتی بهترین الگوریتم‌ها تحلیل کرد. به منظور کاهش مقدار و بعد این حجم عظیم داده‌ها از اطلاعات به تکنیک‌های داده کاوی نیاز است.

جمع‌آوری داده به صورت هدایت نشده بسیار ارزان‌تر و سریع‌تر از جمع‌آوری داده‌ها به صورت هدایت شده است. زیرا در جمع‌آوری داده به صورت هدایت شده به ورودی‌های معلوم با خروجی‌های متناظر نیاز است. بنابراین، اگر جمع‌آوری داده‌ها صرفاً به صورت هدایت نشده انجام شود چگونه می‌توان از آن استفاده کرد؟ به منظور حل این مشکل که یکی از دشوارترین مسائل در داده کاوی می‌باشد، به الگوریتم‌هایی نیاز است که توانایی یافتن گروه‌بندی‌ها، خوشه‌ها[۵] ، رابطه‌ها و وابستگی‌های طبیعی و موجود در داده‌ها را داشته باشد.

ساده‌ترین وضعیتی که در داده کاوی با آن مواجه هستیم، زمانی است که تمامی نقاط داده‌ای بطور کامل هدایت‌شده باشند. تعداد زیادی از تکنیک‌های داده کاوی موجود، به منظور کار با چنین داده‌هایی کاملاً مناسب هستند؛ اما ممکن است که از نظر مقیاس‌پذیری[۶] با یکدیگر متفاوت باشند. یک الگوریتم داده‌کاوی که هم با داده‌های کوچک و هم با داده‌های بزرگ بخوبی کار کند، مقیاس‌پذیر نامیده می‌شود. متاسفانه تعداد کمی از الگوریتم‌ها دارای چنین ویژگی هستند.

کسب موفقیت در یک پروژه داده کاوی به در دسترس بودن دانش موجود در حیطه علمی مورد نظر، بسیار وابسته است و بنابراین برای داده‌کاوها ضروری است تا با افراد متخصص در حیطه علمی مورد نظر و صاحبان داده همکاری بسیار نزدیکی داشته باشند. استخراج دانش جدید فرآیندی است که نیاز به تعامل (با متخصصین در حیطه علمی مورد نظر) و تکرار (در فرایند استخراج دانش) بسیار زیاد دارد. نمی‌توان یک سیستم داده‌کاوی که با موفقیت در یک حیطه علمی معین پیاده‌سازی شده است را به سادگی و در ارتباط با یک حیطه علمی دیگر بکار گرفته و انتظار کسب نتایج خوب را داشت [۲۹].

[۵] Clusters

[۶] Scalability

۳.۱ تفاوت روش‌های آماری و داده کاوی

آمار و تحلیل‌های آماری ابزار تبدیل داده به اطلاعات برای استفاده در علوم مختلف است. امروزه مفهوم داده کاوی [۷] همگام با روش‌های آماری مورد استفاده محققان قرار می‌گیرد. روش‌های کلاسیک داده کاوی از قبیل شبکه‌های عصبی، روش‌های قوی‌تری برای داده‌های واقعی به ما می‌دهند و همچنین استفاده از آنها برای کاربرانی که تجربه کمتری دارند راحت‌تر است و بهتر می‌توانند از آن استفاده کنند. اما معمولاً داده‌ها اطلاعات زیادی در اختیار ما نمی گذارند، این روش‌ها با اطلاعات کمتر بهتر می توانند کار کنند وهمچنین اینکه برای داده‌های وسیع کاربرد دارند.

داده‌های جمع آوری شده گاهی خیلی از فرض‌های قدیمی آماری را در نظر نمی گیرند، از قبیل اینکه مشخصه ها باید مستقل باشند، تعیین توزیع داده‌ها، داشتن کمترین همپوشانی در فضا و زمان و تخلف کردن از هر کدام از فرض‌ها می تواند مشکلات بزرگی ایجاد کند زمانی که یک کاربر سعی می کند که نتیجه‌ای را بدست آورد.

پایه و اساس داده کاوی به دو مقوله آمار و هوش مصنوعی تقسیم شده است که هوش مصنوعی به عنوان روش‌های یادگیری ماشین [۸] در نظر گرفته می شوند. فرق اساسی بین روش‌های آماری و روش‌های یادگیری ماشین بر اساس فرض‌ها و یا طبیعت داده‌هایی است که پردازش می شوند. بعنوان یک قانون کلی فرض‌های تکنیک‌های آماری بر این اساس است که توزیع داده‌ها مشخص است که بیشتر موارد فرض بر این است که توزیع نرمال است و در نهایت درستی یا نادرستی نتایج نهایی به درست بودن فرض اولیه وابسته است. در مقابل روش‌های یادگیری ماشین از هیچ فرضی در مورد داده‌ها استفاده نمی کند و همین مورد باعث تفاوت‌هایی بین این دو روش می شود. به هر حال ذکر این نکته ضروری به نظر می‌رسد که بسیاری از روش‌های یادگیری

[۷] Data mining

[۸] Learning machine

ماشین برای ساخت از حداقل چند استنتاج آماری استفاده می کنند که این مسأله بطور خاص در مدل شبکه‌های عصبی دیده می‌شود.

بطور کلی روش‌های آماری روش‌های قدیمی‌تری هستند که به حالت‌های احتمالی مربوط می‌شوند. داده کاوی جایگاه جدیدتری دارد که به هوش مصنوعی و روش‌های یادگیری ماشین مربوط هستند.

روش‌های آماری بیشتر زمانی که تعداد داده‌ها کمتر است و اطلاعات بیشتری در مورد داده‌ها می توان بدست آورد استفاده می‌شوند به عبارت دیگر این روش‌ها با مجموعه داده‌های کوچکتر سر و کار دارند. همچنین به کاربران ابزارهای بیشتری برای امتحان کردن داده‌ها با دقت بیشتر فهمیدن ارتباطات بین داده‌ها را می‌دهد. برخلاف روش‌هایی از قبیل شبکه عصبی که فرآیند مبهمی دارد. پس به طور کلی این روش در محدوده مشخصی از داده‌های ورودی بکار می‌رود.

روش‌های آماری چون پایه ریاضی دارند نتایج دقیق تری نسبت به روش‌های داده کاوی به دست دهند ولی استفاده از روابط ریاضی موجود در آن‌ها نیازمند داشتن اطلاعات بیشتری در مورد داده‌ها است. مزیت دیگر روش‌های آماری در تعبیر و تفسیر داده‌ها است. هر چند روش‌های آماری به خاطر داشتن ساختار ریاضی تفسیر سخت تری دارند ولی دقت نتیجه گیری و تعبیر خروجی‌ها در این روش بهتر است بطور کلی روش‌های آماری زمانی که تفسیر داده‌ها توسط روش‌های دیگر مشکل است بسیار مفید هستند.

تفاوت‌های کلی روش‌های آماری و روش‌های داده کاوی در جدول ۱.۱ ارائه شده است.

در داده کاوی داده‌ها اغلب بر اساس همپوشانی نمونه هاست، نسبت به اینکه بر اساس احتمال داده‌ها باشد همپوشانی نمونه ها برای آشنایی همه انواع پایه ها برای تخمین پارامترها مشهور است. همچنین اغلب استنتاج‌های آماری نتایج ممکن است مشارکتی باشد تا اینکه سببی باشند. تکنیک‌های ماشین را به سادگی می توان تفسیر کرد. مثلاً روش شبکه عصبی بر اساس یک مدل ساده بر اساس مغز انسان استوار است. یعنی همان ساختار مغز انسان را اجرا

جدول ۱.۱: تفاوت روش‌های آماری و روش‌های داده کاوی

روش‌های داده کاوی	روش‌های آماری
بدون فرض اولیه	داشتن فرض اولیه
در انواع مختلفی از داده‌ها کاربرد دارند نه فقط داده‌های عددی	تنها برای داده‌های عددی کاربرد دارند
در محدوده وسیع تری از داده‌ها کاربرد دارند	در محدوده کوچکی از داده‌ها کاربرد دارند
به داده‌های درست بستگی دارد	توانایی حذف noise و فیلتر کردن داده‌های نامشخص را دارد
استفاده از شبکه عصبی	روش‌های رگرسیون و استفاده از معادلات
استفاده از روش‌های یادگیری ماشین و هوش مصنوعی	استفاده از روابط ریاضی
در یادگیری غیر نظارتی کاربرد بیشتری دارد	در آمار توصیفی و تحلیل خوشه‌ای کاربرد بیشتری دارد

می‌کنند ولی خروجی‌های بسیاری از روش‌های آماری ساختار ریاضی دارند، مثلاً یک معادله است که تعبیر و تفسیر آن مشکل‌تر است. در مورد روش‌های آماری باید این مطلب را گفت بدون توجه به اینکه مدل کاربردی، مدل آماری است یا خیر، تست‌های آماری می‌تواند برای تحلیل نتایج مفید باشد [۲۴].

۴.۱ وظایف داده کاوی

داده کاوی معمولا وظایف یا به عبارت بهتر استراتژی‌های زیر را در داده‌ها بکار می برد:

- توضیح و تفسیر
- تخمین
- پیش‌بینی
- کلاس بندی

- خوشه سازی
- وابسته سازی و ایجاد رابطه

البته باید گفت که روش‌های داده کاوی تنها به یك استراتژی خاص محدود نمی شوند و گاهی نتایج یك همپوشانی را بین روش‌ها نشان می‌دهد. برای مثال درخت تصمیم ممکن است که در کلاس بندی تخمین و پیش‌بینی کاربرد داشته باشد. روش‌های آماری در مباحث تخمین و پیش‌بینی کاربرد دارند. در تحلیل آماری تخمین و پیش‌بینی عناصری از استنباط آماری هستند. استنباط آماری شامل روش‌هایی برای تخمین و تست فرضیات درباره جمعیتی از ویژگی‌ها براساس اطلاعات حاصل از نمونه هستند. یك جامعه شامل مجموعه ای از عناصر از قبیل افراد، آیتم یا داده‌هایی است که در یك مطالعه خاص آمده است. بنابراین در اینجا به طور خلاصه به توضیح این دو استراتژی می‌پردازیم.

۱.۴.۱ تخمین

در تخمین به دنبال این هستیم که مقدار یک مشخصه خروجی مجهول را تعیین کنیم، مشخصه خروجی در مسایل تخمین بیشتر عددی هستند تا قیاسی، بنا بر این مواردی که بصورت قیاسی هستند باید به حالت عددی تبدیل شوند. مثلا موارد بله، خیر به ۰ و ۱ تبدیل می شود.

تکنیك‌های نظارتی تحلیل داده قادرند یکی از دو نوع مسائل کلاس‌بندی یا تخمین را حل کنند، نه اینکه هر دو را. یعنی اینکه تکنیکی که کار تخمین را انجام می‌دهد، کلاس‌بندی نمی‌کند.

روش‌های آماری مورد استفاده در این مورد بطورکلی شامل تخمین نقطه‌ای و فاصله اطمینان می‌باشد. تحلیل های آماری تخمین و تحلیل‌های یك متغیره و ... از این جمله می‌باشند.

در توضیح اینکه چرا به سراغ تخمین می رویم باید گفت که مقدار واقعی پارامترها برای ما ناشناخته است. مثلاً مقدار واقعی میانگین یك جامعه مشخص نیست. در خیلی از موارد تعیین میانگین مجموعه‌ای از داده‌ها برای ما مهم است. مثلاً میانگین نمرات درسی یك کلاس،

میانگین تعداد نفراتی که در یك روز به بانك مراجعه می‌کنند، متوسط مقدار پولی که افراد در یك شعبه خاص از بانك واریز می‌کنند و موارد این‌چنینی.

زمانی که مقدار یک آماره را برای برآورد کردن پارامتر یک جامعه به کار ببریم، آن پارامتر را تخمین زده‌ایم و به مقدار این آماره، برآورد نقطه‌ای پارامتر می‌گوییم. در واقع از کلمه نقطه‌ای برای تمایز بین برآورد کننده‌های نقطه‌ای و فاصله‌ای استفاده می‌کنیم. مهمترین تخمین زننده‌ها برآورد واریانس و میانگین جامعه هستند. خود برآورد کننده‌ها دارای خاصیت هایی چون نااریبی، کارایی، ناسازگاری، بسندگی و... هستند که هر یک به بیان ویژگی خاصی از آنها می‌پردازند و میزان توانایی آنها را در تخمین درست و دقیق یک پارامتر تعیین می‌کنند. در تخمین نیازمند داشتن اندازه نمونه هستیم.

در مواردی نیز تخمین فاصله‌ای برای ما اهمیت دارد. فاصله اطمینان شامل فاصله‌ای است که با درصدی از اطمینان می‌توانیم بگوییم که مقدار یك پارامتر درون این فاصله قرار می گیرد. به عبارت دیگر اگر چه برآورد نقطه‌ای طریقه متداول توصیف برآوردها است اما درباره آن، جا برای پرسش‌های زیادی باقی است. مثلاً برآورد نقطه ای به ما نمی گوید که برآورد بر چه مقداری از اطلاعات مبتنی است و چیزی درباره خطا بیان نمی‌کند. بنابراین می‌توانیم که برآورد پارامتر را با استفاده از اندازه نمونه و مقدار واریانس یا اطلاعات دیگری درباره توزیع نمونه گیری کامل کنیم. این کار همچنین ما را قادر می‌سازد که اندازه ممکن خطا را برآورد کنیم. پس در خیلی از موارد تعیین نقطه دقیق یك پارامتر ممکن نیست ولی فاصله اطمینان، اطمینان ما را از قرار گرفتن مقدار پارامتر در یك بازه تضمین می‌کند.

۲.۴.۱ پیش‌بینی

هدف از انجام پیش‌بینی تعیین ترکیب خروجی با استفاده از رفتار موجود می‌باشد. یعنی در واقع رسیدن به یک نتیجه بوسیله اطلاعات موجود از داده‌ها. مشخصه‌های خروجی در این روش هم

می‌توانند عددی باشند و هم قیاسی. این استراتژی در بین استراتژی های داده کاوی از اهمیت خاصی بر خوردار است و مفهوم کلی‌تری را نسبت به موارد دیگر دارد. خیلی از تکنیک‌های نظارتی که برای کلاس بندی و تخمین مناسب هستند در واقع کار پیش‌بینی انجام می‌دهند. آنچه از کتاب‌های آماری و داده کاوی تحت عنوان پیش‌بینی برمی آید رگرسیون و مباحث مربوط به آن است. در واقع در اکثر این کتاب‌ها هدف اصلی از انجام تحلیل‌های آماری برای داده کاوی، رگرسیون داده‌هاست و این به عنوان وظیفه اصلی روش‌های آماری معرفی می‌شود [۲۴].

۵.۱ اهداف تحلیل رگرسیون

با انجام رگرسیون اهداف زیر دنبال می‌شود:

۱- به دست آوردن رفتار متغیر وابسته با استفاده از متغیرهای پیش بین. مثلا در نمونه ای این رفتار خطی است یا اینکه شکل منحنی خواهد داشت.

۲- پیش‌بینی بر اساس داده‌ها برای نمونه‌های آینده، که هدف اصلی در داده کاوی از طریق متدهای آماری است. مثلا از روی اطلاعاتی مثل داشتن کارت اعتباری یك فرد جدید،نوع جنسیت او، سن فرد و میزان درآمد سالیانه او بتوان حدس زد که این فرد از بیمه عمر استفاده می‌کند یا خیر. یا اینکه با داشتن اطلاعات در مورد داشتن یا نداشتن کارت اعتباری و بیمه عمر و سن فرد بتوان جنسیت فرد را تعیین کرد.

۳- استنباط استنتاجی یا تحلیل حساسیت. تعیین اینکه اگر متغیر پیش بین تا اندازه‌ای تغییر کند، متغیر وابسته چقدر تغییر می‌کند یعنی می‌خواهیم بدانیم تغییرات y چگونه تابعی از تغییرات x است.

روش‌های مختلف رگرسیون برای داده کاوی وجود دارد. رگرسیون خطی بیشترین کاربرد را دارد و همچنین مشتقات آن حائز اهمیت است. یك نمونه از آن مشتقات، رگرسیون خطی سلسله مراتبی یا رگرسیون چند سطحی است. این روش یکی از ابزارهای تحلیل داده‌های پیچیده

از قبیل افزایش فرکانس در تحقیقات مقداری را شامل می‌شود. مدل‌های رگرسیون چند سطحی برای حالت‌هایی که همپوشانی در سطوح مختلف وجود دارد مفید است. برای مثال اطلاعات آموزشی ممکن است اطلاعاتی از قبیل اطلاعات فردی دانش آموزان (نام، نام خانوادگی و در کل پیش زمینه خانوادگی)، اطلاعات سطح کلاس از قبیل ویژگی های معلم و همچینین اطلاعات درباره مدرسه همانند سیاست آموزشی و غیره باشد. حالت دیگر مدلهای چند سطحی، تحلیل داده‌های بدست آمده از نمونه‌های خوشه‌بندی شده است. یك خانواده از مدلهای رگرسیون، به عنوان متغیرهای شاخص برای رتبه بندی یا خوشه بندی به کار می‌رود علاوه بر اینکه همپوشانی را اندازه می گیرد. با نمونه خوشه بندی شده مدلسازی چند سطحی برای توسعه نمونه هایی که داخل خوشه نیستند، لازم است.

در روش رگرسیون چند سطحی یا سلسله مراتبی محدودیتی برای تعداد سطوح تغییر که می‌تواند انجام شود، وجود ندارد. روش‌های بیزی در تخمین پارامترهای مجهول کمك می‌کند، هرچند که محاسبات پیچیده‌ای دارد. ساده‌ترین توسعه از رگرسیون همپوشانی مجموعه‌ای از متغیرهای شاخص برای کلاس بندی نمونه های آموزشی یا رتبه بندی و خوشه بندی در نمونه‌های داده شده است [۲۴].

مدل رگرسیونی مورد نظر ما که در داده کاوی به وفور مورد استفاده قرار می‌گیرد، رگرسیون لجستیک است. در فصل دو به معرفی این مدل رگرسیون می‌پردازیم و روش برازش این مدل به داده‌ها و استفاده از آن برای پیش‌بینی مقدار متغیر پاسخ را شرح می‌دهیم.