دانشگاه صنعتی اصفهان

دانشکده علوم ریاضی

داده‌کاوی آماری؛
ترکیب رگرسیون لجستیک چندگانه با درخت تصمیم

پایان‌نامه کارشناسی ارشد آمار ریاضی

شکیبا خادم‌القرانی

اساتید راهنمای پایان‌نامه

دکتر علی زینل همدانی
دکتر محمد حسین سرائی

۱۳۸۵

دانشگاه صنعتی اصفهان

دانشکده علوم ریاضی

پایان نامه کارشناسی ارشد آمار ریاضی خانم شکیبا خادم‌القرانی

تحت عنوان

داده‌کاوی آماری؛
ترکیب رگرسیون لجستیک چندگانه با درخت تصمیم

در تاریخ ۸۵/۱۲/۲۲ توسط کمیته تخصصی زیر مورد بررسی و تصویب نهائی قرار گرفت.

۱– استاد راهنمای پایان‌نامه — دکتر علی زینل همدانی

۲– استاد راهنمای پایان‌نامه — دکتر محمد حسین سرائی

۳– استاد داور ۱ — دکتر محمدرضا خیام‌باشی
(گروه کامپیوتر دانشگاه اصفهان)

۴– استاد داور ۲ — دکتر سروش علیمرادی

سرپرست تحصیلات تکمیلی دانشکده — دکتر رسول نصر اصفهانی

تشکر و قدردانی

حمد و سپاس پروردگار عالمیان را، او که همواره مرا از دریای بیکران فضل الهی‌اش بهره‌مند کرد و از آغازین لحظات زندگی نعمت‌هایش را بر من ارزانی داشت و او که از همین درگاه، نعمت مادر و پدری والا را بر من عطا فرمود.

بر خود واجب می‌دانم که از اولین و بزرگ‌ترین معلمان زندگیم مادر و پدر عزیزم که مرا به جان پروردند و امید رسیدن به افق‌های روشن را در دلم شکوفا ساختند از صمیم قلب تشکر کنم.

از خواهران عزیزم که همیشه مرا مورد مهربانی‌های خود قرار داده‌اند متشکرم.

بر خود لازم می‌دانم که نهایت تشکر و سپاس صمیمانه خود را به دوست و استاد عزیزم جناب آقای دکتر علی همدانی که از رهنمودهای ایشان در مراحل مختلف زندگی‌ام بهره برده‌ام ابراز نمایم.

همچنین از جناب آقای دکتر محمد حسین سرایی استاد ارجمندم که با راهنمایی‌های ارزنده خود، در انجام تحقیقات مرا یاری نمودند صمیمانه سپاسگزارم.

از جناب آقای دکتر احمد پارسیان که برای اولین بار مرا با داده‌کاوی آشنا نمودند و مرا همواره همراه، همگام و پشتیبان بوده‌اند کمال تشکر را دارم.

از آقای دکتر محمدرضا خیام‌باشی و هم‌چنین آقای دکتر سروش علیمرادی که زحمت بازخوانی و داوری پایان‌نامه را متقبل شدند، نهایت تشکر و قدردانی را دارم.

همچنین از دیگر اساتید ارجمندم که در دانشگاه صنعتی اصفهان افتخار شاگردی ایشان را داشته و همواره بنده را تشویق و یاری نموده‌اند صمیمانه سپاسگزارم.

از تمامی دوستان و هم‌کلاسی‌های عزیزم به خصوص خانم‌ها ریحانه ریخته‌گران و مریم هاشمی که در طی دوران تحصیلی‌ام زیباترین لحظات زندگی‌ام را رقم زدند، متشکرم.

و در پایان دست همه کسانی را که در گذر زندگی چراغی فرا راهم داشته‌اند به گرمی می‌فشارم.

فهرست مطالب

چکیده:

داده‌کاوی یک شیوه نوین برای استخراج اطلاعات در فرایند تصمیم‌گیری‌های علمی است و اغلب از روش‌های آماری و یادگیری ماشین برای تجزیه و تحلیل داده‌ها استفاده می‌نماید. یک رویکرد جدید در این راستا ترکیب شیوه‌های آماری و یادگیری ماشین برای کسب اطلاعات بیشتر از استفاده جداگانه هریک می‌باشد.

در این پایان‌نامه فرایند داده‌کاوی، رگرسیون لجستیک و درخت‌های تصمیم معرفی می‌شوند و با ترکیب $CART$، یکی از الگوریتم‌های درخت‌های تصمیم، با رگرسیون لجستیک چندگانه شیوه نوینی برای تحلیل داده‌های چندرسته‌ای ارائه می‌گردد. سپس این روش ابداعی را برای داده‌های واقعی خط تولید ورق فولاد به‌کار گرفته که نتایج نشان می‌دهند تحلیل داده‌ها کاراتر و حاوی اطلاعات سازنده‌تر به خصوص برای پاسخ‌های رسته‌ای ترتیبی می‌باشد.

مقدمه

اندازه‌گیری یک ویژگی یا صفت از جامعه که معمولاً یک متغیر است، مجموعه‌ای از اعداد را شامل می‌شود که به طور متداول آن را داده می‌نامند. داده‌ها بر دو قسمند:
داده پیوسته که یک عدد حقیقی است و به صورت اعشاری بیان می‌شود.
داده گسسته که آن را داده جدا از هم نیز می‌نامند.
امروزه پیشرفت شگفت‌انگیز فن‌آوری رایانه‌ای و مجهزشدن بشر به این ابزار، موجب پیشرفت فوق‌العاده در کسب و ذخیره‌سازی داده‌ها و همچنین به وجود آمدن پایگاه داده‌های بزرگ در زمینه‌های مختلف شده است.

داده‌کاوی فرایندی است که با نگرشی نو به مسئله استخراج اطلاعات نهفته پرداخته و در حال حاضر مهم‌ترین فن‌آوری جهت بهره‌برداری مؤثر از داده‌های حجیم می‌باشد و اهمیت آن به عنوان یک علم رو به فزونی است. روش‌های آماری و یادگیری ماشین متداول‌ترین تکنیک‌های داده‌کاوی می‌باشند.
با توجه به متداول بودن داده‌های رسته‌ای چندگانه و اهمیت تحلیل رده‌بندی در بررسی نتایج و فرایند پیش‌گویی در اکثر زمینه‌های مطالعاتی و خصوصاً داده‌کاوی، روش‌هایی که تاکنون ارائه شده‌اند از لحاظ نحوه رده‌بندی، میزان ارائه اطلاعات، نحوه نمایش نتایج و ... مورد بررسی قرار گرفتند.
متداول‌ترین روش آماری در تحلیل و مدل‌سازی داده‌های رسته‌ای دودویی و چندگانه، روش مدل‌سازی رگرسیون لجستیک است. در یادگیری ماشین، درخت تصمیم یک روش معمول در رده‌بندی داده‌ها می‌باشد و الگوریتم $CART$ نیز از مشهورترین الگوریتم‌های ساخت درخت به شمار می‌آید.
در این پایان‌نامه، علاوه بر معرفی دو روش فوق، روش مدل‌سازی لجیت و درخت تصمیم در تحلیل

داده‌های رسته‌ای، چند کاربرد از روش‌های آماری و یادگیری ماشین در فرایند داده‌کاوی نیز معرفی می‌شوند. در این کاربردها، محققان رویکرد نوینی را ابداع نموده‌اند که در آن به ترکیب این دو روش مستقل پرداخته و این امر موجب ارائه اطلاعاتی با جزئیات بیشتر می‌گردد[۱۱،۲۸،۳۸،۴۰].

با تحقیق و تفحص در تحلیل داده‌های رسته‌ای چندگانه ترتیبی، پیرو این رویکرد نوین، روش جدیدی برای تحلیل و مدل‌سازی ابداع گردید. در این روش با ترکیب روش رگرسیون لجستیک چندگانه و الگوریتم $CART$ در روش درخت تصمیم، شیوه جدیدی برای رده‌بندی داده‌های رسته‌ای ارائه می‌شود. در شیوه مدل‌سازی لجیت تجمعی، متغیرهای معنی‌دار برای رده‌های مختلف متغیر هدف شناسایی می‌شوند و تأثیرگذاری متغیرهای پیش‌گو یا توضیح‌دهنده بر اساس آزمون‌های والد، تابع امتیاز و ... بررسی می‌گردد و میزان اثرگذاری هر متغیر دقیقاً محاسبه می‌شود. ولیکن نتایج این تحلیل به صورت گرافیکی نبوده و چنانچه هدف تحلیل، رده‌بندی متغیر مربوطه باشد، به دلیل این‌که از احتمالات تجمعی اثرات استفاده می‌شود، به سادگی با این روش امکان‌پذیر نمی‌باشد و حتی در مواردی که از دیگر مدل‌های لجیت برای تحلیل داده‌های دودویی یا چندگانه استفاده شود، نتایج قابل مقایسه با نتایج تحلیل درخت تصمیم نمی‌باشد، به عبارت دیگر درخت تصمیم نمایش منحصر به فرد، جذاب، ساده و سریعی را برای رده‌بندی ارائه می‌دهد که ترکیب این دو روش در تحلیل داده‌ها اطلاعاتی با جزئیات بیشتر، جذاب‌تر و حتی معتبرتر را فراهم می‌سازد و به عنوان کاربردی از این پژوهش، این روش ابداعی بر روی داده‌هایی از صنعت فولاد به کار گرفته می‌شود.

حاصل بررسی‌ها و مطالعات به عمل آمده در شش فصل به شرح زیر دسته‌بندی شده است:

فصل اول شامل پیشینه داده‌کاوی، معرفی داده‌کاوی، سیستم داده‌کاوی، ارتباط آن با کشف دانش از پایگاه داده‌ها و آمار، و معرفی برخی نرم‌افزارهای مفید و متداول در این زمینه، می‌باشد.

فصل دوم به مفهوم مدل‌سازی در آمار و متغیر رسته‌ای اشاره می‌کند، سپس به بررسی مدل‌سازی لجیت برای داده‌هایی با پاسخ رسته‌ای دوگانه و چندگانه می‌پردازد.

فصل سوم به معرفی درخت تصمیم و عوامل مؤثر در ساخت درخت اختصاص دارد.

فصل چهارم برخی کاربردهای ترکیب روش‌های آماری و درخت تصمیم را شرح می‌دهد.

فصل پنجم به معرفی روش ابداعی ترکیب رگرسیون لجستیک و درخت تصمیم می‌پردازد و کاربرد این شیوه را بر روی داده‌های صنعت فولاد ارائه می‌نماید.

فصل ششم خلاصه نتایج و پیشنهادات این مطالعه را مطرح می‌کند.

فصل ۱

داده‌کاوی

پیشرفت شگفت انگیز فن‌اوری رایانه‌ای و مجهز شدن بشر به این ابزار، سبب پیشرفت فوق‌العاده در کسب و ذخیره‌سازی داده‌ها و همچنین به‌وجود آمدن پایگاه داده‌های بزرگ در زمینه‌های مختلف شده است، لذا مشکل دنیای امروز نبود داده کافی برای تصمیم‌گیری‌های علمی نیست و محققان در بیشتر زمینه‌های مطالعاتی از جمله کشاورزی، پزشکی، تجاری، ترافیک، اینترنت و . . . با سیلانی از داده‌های خام مواجه هستند.

امروزه اهمیت دستیابی به اطلاعات نهفته در داده‌های حجیم که لازمه مدیریت مؤثر است، روبه فزونی است که با به‌کار بردن سیستم‌های سنتی میسر نمی‌باشد، از این رو محققان برای ارائه تحلیل‌های مفید و کارامد نیازمند ترکیب روش‌های مختلف می‌باشند.

داده‌کاوی فرایندی است که به عنوان یک ابزار پرتوان به تجزیه و تحلیل حجم بالای داده‌ها و پایگاه‌های داده‌های با ابعاد زیاد می‌پردازد.

در این فصل ابتدا به پیشینه‌ای از داده‌کاوی اشاره شده و سپس به معرفی آن پرداخته می‌شود، سیستم داده‌کاوی و ارتباط آن با کشف دانش از پایگاه داده‌ها و آمار مطرح شده و با معرفی برخی نرم‌افزارهای مفید و متداول در این زمینه، فصل خاتمه می‌یابد.

۱—۱ پیشینه داده‌کاوی

داده‌کاوی[۱] و کشف دانش از پایگاه داده‌ها[۲] (KDD) از جمله موضوع‌هایی هستند که همزمان با ایجاد و استفاده از پایگاه داده‌ها در اوایل دههٔ ۸۰ برای جستجوی دانش در داده‌ها شکل گرفته‌اند. شاید بتوان لوول (۱۹۸۳) را اولین شخصی که گزارشی در مورد داده‌کاوی تحت عنوان شبیه سازی فعالیت‌های داده‌کاوی ارائه نموده است، معرفی نمود. همزمان با وی پژوهشگران و متخصصان تحلیل داده‌ها، علوم رایانه، آمار، هوش مصنوعی، یادگیری ماشین و . . . نیز به پژوهش در این زمینه و زمینه‌های مرتبط پرداخته‌اند.
به طور جدی موضوع داده‌کاوی از اوایل دههٔ ۹۰ مطرح شد. پژوهش‌ها و مطالعه‌های زیادی تاکنون صورت گرفته و همچنین سمینارها، دوره‌های آموزشی و کنفرانس‌های متعددی برگزار شده‌اند[۴۲].
سال ۱۹۹۱ پیاتتسکی — شاپیرو استقلال آماری قاعده‌ها در داده‌کاوی را بررسی نمود. سال ۱۹۹۵ هافمن و نش استفاده از داده‌کاوی و داده‌انبار توسط بانک‌های آمریکا را مطرح نموده و بیان کردند که چگونه این سیستم‌ها برای بانک‌های آمریکا رقابت بیشتری ایجاد می‌کنند. چت‌فیلد مشکلات ایجاد شده توسط داده‌کاوی را بررسی نمود[۹] و همچنین مقاله‌ای تحت عنوان مدل‌های خطی غیر دقیق داده‌کاوی و استنباط آماری ارائه کرد. هندری نیز دیدگاه اقتصاد سنج‌ها روی داده‌کاوی را تهیه نمود. در این سال انجمن داده‌کاوی در اولین کنفرانس بین‌المللی کشف دانش و داده‌کاوی شروع به‌کار نمود. سال ۱۹۹۶ ایمیلنسکی و منیلا دیدگاهی از داده‌کاوی به عنوان پرس و جو کننده از پایگاه داده‌های استنتاجی[۳] را پیشنهاد کردند. فایاد، پیاتتسکی — شاپیرو، اودوراسامی پیشرفت‌های کشف دانش و داده‌کاوی را عنوان کردند[۱۶]. سال ۱۹۹۷ منیلا خلاصهٔ مطالعه‌ای روی اساس داده‌کاوی را ارائه نمود. فریدمن مقاله‌ای در ارتباط با مفهوم آمار و داده‌کاوی [۱۹] و در سال ۱۹۹۸ هند مقاله‌ای تحت عنوان داده‌کاوی: آمار و بیشتر؟ [۲۲] ارائه نمودند. سال ۲۰۰۱ هند و اسمیت بحث‌های مقایسه‌ای بین آمار و داده‌کاوی را مطرح کردند[۲۴].
کهنرت و همکارانش در سال ۲۰۰۰ دیدگاه جدیدی را ابداع کردند. برای کسب دانش و اطلاعات بیشتر از پایگاه داده‌ها به ترکیب مدل‌های ناپارامتری با رگرسیون لجستیک پرداختند و کاربرد این تکنیک را روی آسیب‌های حوادث وسایل نقلیه موتوری ارائه نمودند[۲۸]. پس از آن پیرو این شیوه، در سال ۲۰۰۲ چانگ روش $CART$[۴] و مدل‌سازی لگاریتم خطی را ترکیب و کاربرد آن را روی داده‌ها تولد بررسی نمود[۱۱]. اندرو ورث و مارک کرونین نیز در سال ۲۰۰۳ کاربرد روش های تحلیل ممیزی، رگرسیون

[۱] *Data Mining*
[۲] *Knowledge Discovery of Database*
[۳] *Inductive Database*
[۴] *Classification and Regression Trees*

لجستیک و تحلیل طبقه بندی درختی روی گسترش مدل‌های طبقه‌بندی را شرح داده‌اند[۴۰].

۱—۲ داده‌کاوی

نگاهی به ترجمه لغوی داده‌کاوی، در درک بهتر واژه مؤثر می‌باشد.
کلمه $'Mine'$ به معنای استخراج از منابع نهفته و با ارزش زمین می‌باشد، ادغام آن با کلمه داده $'Data'$ به جستجویی عمیق از داده‌های قابل دسترس با حجم زیاد برای کسب اطلاعات مفید که قبلاً نهفته بودند، تأکید دارد.

۱—۲—۱ داده‌کاوی چیست؟

این اصطلاح را آماردانان، تحلیل‌گران داده‌ها و انجمن سیستم‌های اطلاعات مدیریت به‌کار برده‌اند. برای داده‌کاوی تعاریف متنوعی ارائه شده که به نوع نگرش و دیدگاه محققان در علوم مختلف اشاره دارد. برخی از این تعاریف عبارتند از:

۱. داده‌کاوی فرایند شناخت الگوهای معتبر، جدید، مفید و قابل فهم از داده‌ها می‌باشد.

۲. داده‌کاوی به فرایند استخراج اطلاعات نهفته، قابل فهم و قابل تعقیب از پایگاه داده‌های بزرگ و استفاده از آن در تصمیم‌گیری‌های تجاری مهم، اطلاق می‌شود.

۳. داده‌کاوی، مجموعه‌ای از روش‌ها در فرایند کشف دانش می‌باشد که برای تشخیص الگوها و روابط نامعلوم در داده‌ها مورد استفاده قرار می‌گیرد.

۴. فرایند کشف دانش الگوهای مفید از داده‌ها را داده‌کاوی می‌گویند.

۵. فرایند انتخاب، کاوش و مدل‌سازی داده‌های حجیم، جهت کشف روابط نهفته با هدف کسب نتایج واضح و مفید، را داده‌کاوی می‌گویند.

که به طور کلی همه تعاریف داده‌کاوی به فرایند استخراج دانش از پایگاه داده‌ها اشاره دارند[۱۹].

۱—۲—۲ ابزارهای داده‌کاوی

برای استخراج اطلاعات از پایگاه داده‌های بزرگ، چنانچه اشاره شد سیستم‌های سنتی به تنهایی کارا نمی‌باشند و لازم است در این راستا از روش‌ها، فنون و الگوریتم‌های به‌کار رفته در سایر رشته‌ها از قبیل آمار، هوش مصنوعی[۵]، یادگیری ماشین[۶] و . . . استفاده شود.

[۵] *Artificial Inteligence*
[۶] *Machine Learning*

به طور کلی داده‌کاوی، فرایند استخراج دانش از پایگاه داده‌های بزرگ، در سه شاخه قدیمی ریشه دارد[۴۱] 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 توزیع استاندارد، انحراف معیار، واریانس، بازه‌های اطمینان، تحلیل رگرسیون، تحلیل ممیزی، تحلیل خوشه‌ای و . . . استفاده می‌کند تا جزئیات و روابط بین داده‌ها به طور دقیق مورد بررسی قرار گیرند، لذا یقیناً در قلب ابزار و روش‌های داده‌کاوی، امروزه تحلیل‌های آمار کلاسیک نقش مهمی را ایفا می‌کنند.

دومین شاخه مرتبط با داده‌کاوی، هوش مصنوعی است که برای تلخیص داده‌ها، تحلیل کاوشگرانهٔ داده‌ها، ساختن مدل و . . . به‌کار می‌رود. این شاخه بر اساس اکتشاف، ساخته شده و سعی دارد پردازش‌هایی شبیه افکار انسان را در مسایل آماری به‌کار برد. این شاخه مستلزم قدرت پردازش رایانه‌ای بالایی بود که تا اوایل دهه ۸۰، زمانی که رایانه‌ها با قدرت بالا و قیمت معقول عرضه شدند، امکان پذیر نبود. هوش مصنوعی تنها کاربردهای اندکی در پژوهش‌های پیشرفته، سازمان‌های دولتی ویژه و بازارهایی خاص داشت ولی در آن دوران ابررایانه‌های لازم برای بهره‌برداری از این فناوری به حدی محدود بود که کاربردی برای سایر افراد و سازمان‌های جامعه نداشت.

سومین شاخه مرتبط با داده‌کاوی، یادگیری ماشین است که تلفیقی از آمار و هوش مصنوعی است. یادگیری ماشین می‌تواند به عنوان هوش مصنوعی تکامل یافته مطرح شود زیرا اکتشاف‌های هوش مصنوعی را با تحلیل‌های آماری پیشرفته ادغام می‌کند. این شاخه 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 کیفیت‌های 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. مهم‌ترین تکنیک‌های این شاخه شبکه‌های عصبی، درخت تصمیم و الگوریتم ژنتیک می‌باشند.

۱—۲—۳ روش‌های داده‌کاوی

اساساً دو روش برای داده‌کاوی عنوان می‌شود که از لحاظ ایجاد، طراحی مدل و یافتن الگوها با یکدیگر تفاوت دارند. اولین روش، مربوط به ساخت مدل است، جدا از مشکلاتی که ذاتاً در مجموعه داده‌های بزرگ وجود دارد مشابه روش‌های کاوشگرانه آماری مرسوم، می‌باشد. در این روش، هدف ارائه خلاصه‌ای

کلی از مجموعه‌ای از داده‌ها برای شناخت و شرح خصوصیت‌های اصلی شکل توزیع است(هند،۱۹۹۸). تحلیل خوشه‌ای بخشی از مجموعه داده‌ها، مدل رگرسیونی برای پیش‌گویی و قاعده رده‌بندی با ساختار درختی مثال‌هایی از این روش می‌باشند.

در طراحی مدل، بعضی مواقع تفاوت‌هایی بین مدل‌های تجربی و ماشینی تشخیص داده می‌شود (باکس و هانتر، ۱۹٦۵، کاکس و هند، ۱۹۹۵). مدل تجربی که بعضی مواقع مدل عملیاتی هم نامیده می‌شود، سعی می‌کند روابط بین داده‌ها را بدون در نظر گرفتن تئوری‌های زیربنایی، پایه‌ریزی و مدل‌بندی کند.

مدل ماشینی که گاهی مدل بنیادی یا مدل پدیده شناختی هم نامیده می‌شود، بر اساس برخی تئوری‌ها یا مکانیسم‌هایی که برای فرایند تولید داده‌های زیربنایی به‌کار می‌رود پایه‌ریزی شده‌اند. داده‌کاوی بنا به تعریف، بیشتر با مدل عملیاتی سروکار دارد.

روش دوم داده‌کاوی، رویکرد تشخیص الگو است. این رویکرد سعی دارد انحراف‌هایی هر چند کوچک از حد مطلوب را تشخیص دهد، تا الگوها و روندهای غیر معمول نمایان شود. مثال‌هایی نظیر الگوهای نامعمول (برای تشخیص کلاهبرداری) در استفاده از کارت‌های اعتباری برای خرج کردن و موضوع‌هایی که الگوهایی با ویژگی‌های نامشابه با دیگران دارند و . . ..

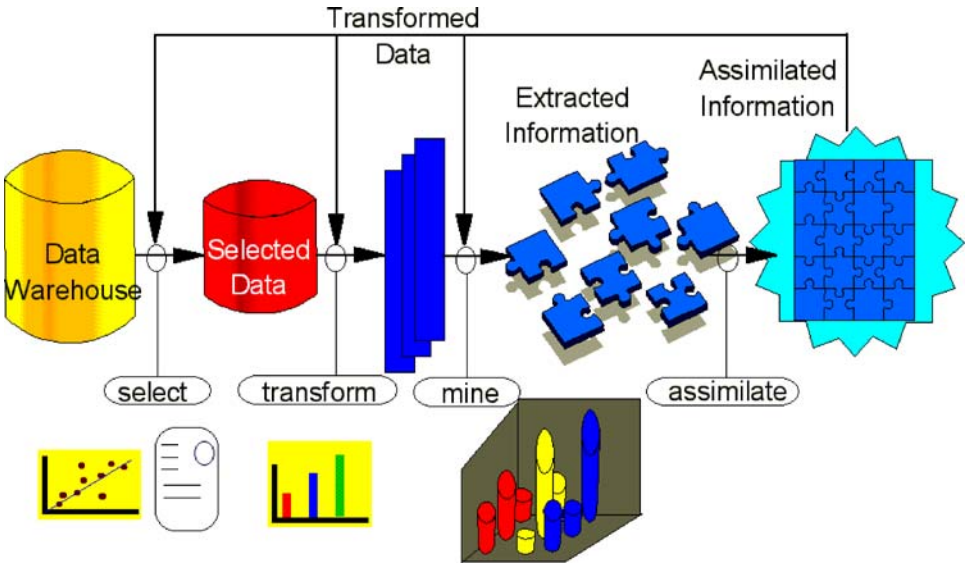
این رده از راهبردها است که موجب گردیده داده‌کاوی به عنوان علم جستجوی اطلاعات باارزش از بین توده عظیمی از داده‌ها به حساب آید.

۱—۲—۴ مراحل داده‌کاوی

به طور خلاصه داده‌کاوی شامل مراحل زیر است:

تعیین موضوع: در ابتدا باید روشن کرد که با تجزیه و تحلیل، انتظار داریم به چه چیزی دست یابیم. بنابراین از قبل باید هدف از داده‌کاوی را تعیین و ثابت کنیم که این هدف قابل اندازه‌گیری است یا نه؟

انتخاب داده‌ها: گام بعدی، انتخاب داده‌ها برای رسیدن به این هدف است. این امر می‌تواند زیر مجموعه‌ای از پایگاه داده‌ای باشد که در اختیار داریم که برای این کار باید تا حد امکان داده‌ها را بخش‌بندی نمود.



نمودار ۱—۱ مراحل داده‌کاوی

آماده کردن داده‌ها: در مرحله قبل داده‌ها گرداوری شد. حال باید تصمیم گرفت کدام‌یک از ویژگی‌های موجود قابل استفاده هستند.

- اتخاذ تدابیری برای بررسی داده‌های گمشده، خطاهای خارجی و داده‌های پرت.
- تعیین متغیرهای اضافی در مجموعه داده‌ها و تصمیم‌گیری درباره اینکه کدام فیلد باید خارج شود.
- در صورت لزوم تصمیم‌گیری درباره انواع تبدیل‌ها: لگاریتمی، نمایی، توان دوم، . . .
- بازرسی شهودی مجموعه داده‌ها برای درک پایگاه داده‌ها.

بازرسی داده‌ها: ارزیابی ساختار داده‌ها به منظور تعیین ابزارهای مناسب.

انتخاب ابزارها: اهداف کاری و ساختار داده‌ها دو رهنمون مهم برای انتخاب ابزار داده‌کاوی مناسب هستند. هر دو باید تحلیل‌گر را به ابزار یکسان هدایت کنند. به هنگام ارزیابی مجموعه‌ای از ابزارهای بالقوه معمولاً پرسش‌های زیر مطرح می‌شوند:

- آیا مجموعه داده‌ها صرفاً رسته‌ای است؟
- چه برنامه‌ای برای حمایت از ابزارهای منتخب داریم؟
- چه قالبی از داده‌ها می‌تواند برای ابزارها مفهوم داشته باشد؟

قالب پاسخ : قالب پاسخ را بازرسی داده‌ها، هدف کاری و انتخاب ابزار، تعیین می‌کنند. سؤال‌های کلیدی در این مرحله عبارتند از:

– قالب بهینه حل درخت تصمیم چیست؟

– قالب مناسب گزینه چیست؟

– هدف از حل چیست؟

– کاربر به چه چیزی احتیاج دارد: نمودارها، گزارش‌ها، برنامه‌ها،. . .؟

طرح‌ریزی مدل : در این مرحله است که فرایند داده‌کاوی آغاز می‌شود. معمولاً اولین گام، استفاده از چند رده تصادفی برای جدا کردن داده‌ها در مجموعه تولیدی، مجموعه آزمون، ایجاد و ارزیابی یک مدل است. تولید قاعده‌های رده‌بندی، درخت‌های تصمیم، خوشه‌بندی زیر گروه‌ها، برنامه‌ها، وزن‌ها و ارزیابی داده‌ها یا نسبت خطاها در این مرحله قرار دارند.

ارزیابی یافته‌ها: تهیه یک گزارش نهایی برای کاربران، که این گزارش باید مستند به فرایند داده‌کاوی بی‌عیب و نقص، شامل آماده‌سازی داده‌ها، ابزارهای به‌کار رفته، نتایج آزمون، کد منابع، قاعدها و . . . باشد.

برخی از این نتایج عبارتند از:

– آیا داده‌های اضافی تجزیه و تحلیل را بهبود خواهند بخشید؟

– از تجزیه و تحلیل داده‌کاوی چه نتایجی می‌تواند پیشنهاد شود؟

– آیا یافته‌ها مناسب هدف کاری هستند؟

هماهنگ کردن پاسخ ها : یافته‌ها با کلیه علایق کاربران در واحدهای کاری مناسب تقسیم می‌شود. می‌توان به‌طور کامل نتایج تجزیه و تحلیل در روش‌های کار را به‌هم پیوند داد. اگر چه ابزارهای داده‌کاوی، پایگاه داده‌ها را به طور خودکار تجزیه و تحلیل می‌کنند، ولی اگر مراقب نباشیم ممکن است به یافته‌های نادرست و نتایج غلط منجر شود. در نظر داشته باشید که داده‌کاوی فرایندی کاری با هدف مشخص برای استخراج اطلاعات مفید از داده‌های موجود در پایگاه داده‌ها است.

۱–۳ داده‌کاوی و آمار

داده کاوی و آمار هر دو با روش‌های تحلیل و مدل‌بندی داده‌ها مرتبط می‌باشند، لذا اشتراک زیادی بین این دو رشته وجود دارد. به عنوان یک لطیفه یکی از نویسندگان در پاسخ به سؤال **داده‌کاوی چیست** بیان می‌دارد که " **همان آمار است اما با یک نام جذاب‌تر**".

داده‌کاوی هم‌پوشانی زیادی با آمار کلاسیک دارد، به طوری که حتی برخی از محققان داده‌کاوی را زیر مجموعه‌ای از آمار تلقی می‌کنند. از طرف دیگر داده‌کاوی از ابزار و روش‌های دیگری از جمله یادگیری ماشین، نظریه و فناوری پایگاه داده‌ها و . . . استفاده می‌نماید و ضرورتی ندارد که به برخی از سطوحی که مد نظر آمار است مربوط شود.

داده‌کاوی اساساً یک شیوه تحلیل اطلاعات است که در آن سعی می‌شود با ایجاد ارتباط بین آمار و علوم رایانه‌ای، روش‌های کارا و سریعی برای استخراج اطلاعات نهفتهٔ مجموعه‌ای بزرگ از داده‌ها ارائه گردند. روابط در داده‌کاوی غالباً به صورت الگوها و مدل‌هایی از جمله معادلات رگرسیونی، سری‌های زمانی، خوشه‌بندی، رده‌بندی، گراف و غیره بیان می‌شوند.

داده‌کاوی و آمار هر دو به طرح استنباط از داده‌ها مربوط می‌شوند. هدف از استنباط ممکن است درک الگوهای همبستگی و پیوندهای سببی بین مقادیر داده‌ها، یا ساخت پیش‌گویی از مقادیر داده‌های قبلی باشد. آمار کلاسیک دربر گیرنده مدلی برای توزیع احتمال داده‌ها و ساخت استنباط به صورت بیان احتمال است. روش‌های داده‌کاوی در بسیاری از موارد مسائلی را دربر می‌گیرد که به سادگی در چارچوب آمار کلاسیک برازش داده نمی شوند و از آمار فاصله گرفته‌اند.

در آمار با جامعه سروکار داریم در حالی که داده‌کاوی با پایگاه داده‌ها سروکار دارد.

داده‌کاوی با نمونه‌گیری مانند آمار سروکار ندارد، یعنی دوباره می‌توان از داده‌ها استفاده نمود.

داده‌کاوی از میان مجموعه مدل‌های برازش داده شده، بهترین مدل را انتخاب می‌کند ولی در آمار تنها یک مدل برازش داده می‌شود.

آمار مدل برازش داده شده را برای تعمیم به جامعه به‌کار می‌برد ولی در داده‌کاوی مدل برازش داده شده انتخابی برای پیش‌گویی داده‌های جدید است.

مجموعه داده‌ها در داده‌کاوی می‌تواند خیلی وسیع‌تر و گسترده‌تر از مجموعه داده‌های معمول در آمار، شامل چند صد میلیون یا هزار میلیارد رکورد باشد.

نوع داده‌ها در آمار کمی و کیفی است ولی داده‌کاوی داده‌های کمی، کیفی و متنی را شامل می‌شود.

نوع متغیرهای ورودی در آمار عددی و در داده‌کاوی از نوع عددی، طبقه‌ای و متنی می‌باشند.

نوع تمرکز در آمار روی مدل و در داده‌کاوی روی الگو است.

زیر بخش‌های اصلی آمار براورد، توزیع‌های احتمالی، آزمون فرضیه‌ها امتیازبندی مدل و پیش‌گویی است و زیر بخش‌های داده‌کاوی مدل‌بندی پیش‌گویی‌ها، بخش‌بندی پایگاه داده‌ها و ساخت فرضیه‌ها است.
هدف اصلی از به‌کارگیری روش‌ها در آمار استفاده از براوردگرها و توزیع‌ها برای ادغام اطلاعات است و در داده‌کاوی هدف اصلی کشف دانش مورد علاقه است.
شکل نتایج در آمار به صورت مدل‌های کلی براورد می‌شود ولی در داده‌کاوی مدل‌های موضعی محاسبه می‌شوند.
ارزش اطلاعاتی نتایج در آمار معلوم و محدود است در حالی که در داده‌کاوی نامعلوم و نامحدود می‌باشد.
در آمار جستجو برای دستیابی به نتایج محدود به جستجوهای جهت‌دار بوده و با نتایج نیز آشنا هستیم ولی در داده‌کاوی جستجوها اکثراً از طریق کاربر تعیین شده و ممکن است جهت‌دار باشند، اما اصولاً روش خودکار بوده و نوع نتایج نامعلوم است.
در مقایسه با آمار، داده‌کاوی توجه کمتری به ویژگی‌های مجانبی استنباط‌های بزرگ نمونه‌ای دارد و فلسفه کلی یادگیری، بیشتر شامل ملاحظه پیچیدگی مدل‌ها و محاسباتی که آنها نیاز دارند، می‌باشد.

۱—۴ داده‌کاوی و کشف دانش از پایگاه داده‌ها

در متون مربوط به داده‌کاوی دو تعبیر مختلف از آن وجود دارد. برخی مؤلفان مانند چت‌فیلد (۱۹۹۵) داده‌کاوی را مترادف کشف دانش از پایگاه داده‌ها می‌دانند[۹]. گروه دوم از جمله فایاد (۱۹۹۶) داده‌کاوی را به عنوان یک مرحله ضروری از فرایند بزرگتر کشف دانش از پایگاه داده‌ها(KDD) در نظر می‌گیرند[۱۷].
عبارت داده‌کاوی توسط آماردانان، محققین پایگاه داده‌ها و سیستم‌های اطلاعات مدیریتی و جوامع بازرگانی به‌کار برده می‌شود. عبارت کشف دانش از پایگاه داده‌ها(KDD) عموماً برای اشاره به فرایند کلی کشف دانش مفید از داده‌هایی که داده‌کاوی گام مهمی در این فرایند است، مورد استفاده قرار می‌گیرد.
به طور خلاصه KDD شامل مراحل زیر است:

۱. **پاکسازی داده‌ها**[۷] : حذف داده‌های ناپایدار و مزاحم.
۲. **یکپارچه سازی داده‌ها**[۸] : ترکیب منابع متعدد، پراکنده و احتمالاً ناهمگن داده‌ها.
۳. **انتخاب داده‌ها**[۹] : بازیابی داده‌های مربوط به عمل کاوش از پایگاه داده‌ها.

[۷] *Data Cleaning*
[۸] *Data Integration*
[۹] *Data Selection*

۴. **تبدیل داده‌ها**[۱۰] : تبدیل داده‌ها به اشکالی مناسب برای به‌کار بردن روش‌های مختلف.

۵. **داده‌کاوی** : مرحله‌ای ضروری از فرایند KDD است که در آن از روش‌های مختلف آماری و یادگیری ماشین برای استخراج الگوها استفاده می‌شود، که خود شامل مراحل زیر است:

– انتخاب عملیات داده‌کاوی

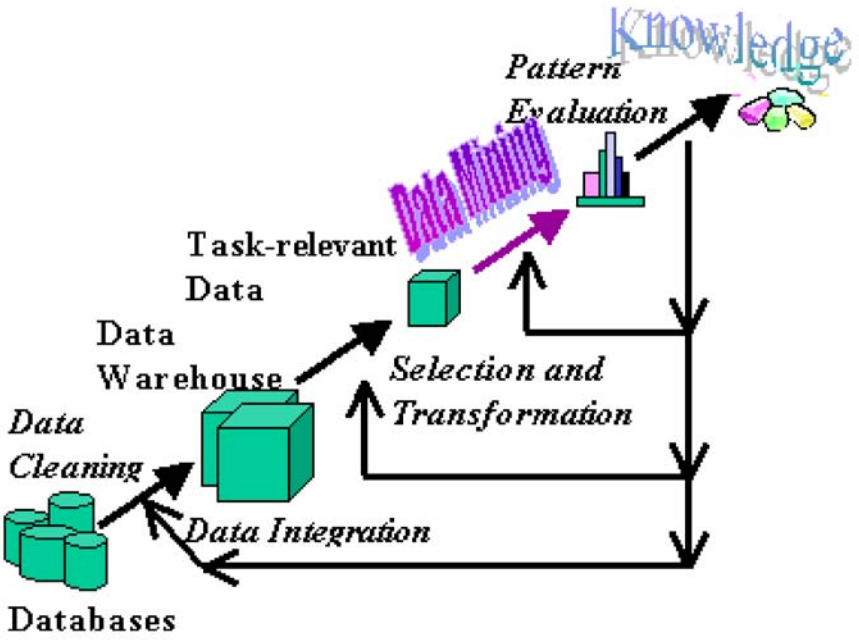
– انتخاب روش داده‌کاوی

– جستجو برای یافتن الگوی مناسب

۶. **ارزیابی الگوها**[۱۱] : شناسایی الگوهای جذاب ارائه دانش، بر اساس معیارهای جذابیت.

۷. **ارائه دانش** [۱۲] : ارائه دانش استخراج شده با استفاده از تکنیک‌های نمایش اطلاعات.

شکل ۱–۲ مراحل فرایند کشف دانش از پایگاه داده‌ها را نشان می‌دهد.



نمودار ۱–۲ داده‌کاوی هسته اصلی کشف دانش از پایگاه داده‌ها

[۱۰] *Data Transformation*
[۱۱] *Patttern Evaluation*
[۱۲] *Knowledge Presentation*