داده کاوی در پایگاه داده های بزرگ


بسم الله الرحمن الرحیم

داده کاوی در پایگاه داده های بزرگ

رضا خسروی

مقطع کاردانی

رشته :فناوری اطلاعات وارتباطات (IT)

تابستان ۱۳۹۲

سپاسگزاری

سپاس خدای مهربان را که اندیشه‌ام داد.

حمد و ستایش بی‌قیاس خدای را سزاست که از الطاف خود در انسان دمید و او را اشرف مخلوقات خود قرار داد. حال که به لطف او توفیق تحصیل علم و کسب دانش را پیدا نمودم ، از خداوند متعال می‌خواهم که قدم‌هایم را در راه خدمت به جامعه استوار گرداند تا بتوانم از آنچه در این سال‌ها آموخته‌ام در مسیر پیشرفت و آبادانی کشور عزیزم استفاده نمایم .

پیش گفتار

داده کاوی، فرایند مرتب سازی و طبقه بندی داده های حجیم و آشکارسازی اطلاعات مرتبط باهم می باشد. امروزه داده کاوی به عنوان یکی از ابزارهای بسیار مهم مدیران جهت شناخت وضعیت دقیق تر سازمان و همچنین کمک در اتخاذ تصمیمات مناسب کاربرد دارد. با استفاده از این تکنیک، داده های موجود در سازمان با بکارگیری ابزارهای نرم افزاری، مورد بررسی و تحلیل دقیق قرار می گیرد تا الگوهای پنهان و پیچیده ای که در آنها وجود دارد کشف و استخراج گردد. داده کاوی را می توان نسل سوم تکنولوژیهایی نامید که با داده سروکار دارند. در نسل اول یا نسل سنتی، فقط انجام پرس و جو های ساده امکان پذیر بود، مثلا تعداد فروش یک کالای خاص چقدر است؟ میزان خرید یک مشتری خاص در ماه جاری چه مبلغی است؟ در نسل دوم یا همان پردازش لحظه ای برخط (OLAP)امکان پرس و جوی همزمان چند بعدی فراهم گردید. در این روش به عنوان مثال به سوالاتی مانند: «میزان فروش محصولات به تفکیک فروشنده، خریدار و مسیر خاص چقدر است؟ » بصورت لحظه ای و با استفاده از مکعب تصمیم و گزارش ماتریسی پاسخ داده می شود. اما در نسل سوم یا همان داده کاوی فقط مساله پرس و جو و دریافت گزارش ها از داده ها نیست، بلکه از حجم انبوه داده ها، الگوهایی کشف می شود که هیچ وقت امکان کشف این الگوها در OLAP یا روش سنتی وجود نداشت. انواع اطلاعات و الگوهایی که از طریق داده کاوی بدست می آیند و کاربرد دارند عبارتند از: وابستگی، تسلسل و توالی، طبقه بندی، خوشه بندی و پیش بینی. برای استخراج این الگوها اغلب از روشهای نوینی مانند شبکه عصبی و درختهای تصمیم استفاده می شود. در عمل برای امکان انجام داده کاوی و استفاده از تکنیکهای فوق الذکر، ابتدا باید نسبت به ایجاد یک انبار داده مناسب اقدام کرد. یک انبارداده در حقیقت پایگاه داده ای است که داده های جاری و همچنین سوابق قبلی تراکنشها را در خود ذخیره کرده و با منابع خارج سازمان نیز ارتباط برقرار می کند.

اهداف کلی این مقاله عبارتند از ارایه تعریف دقیقی از انبار داده، بررسی تکنیکها و کاربردهای داده کاوی و کاربرد آن در مدیریت، معرفی شبکه عصبی به عنوان یکی از روشهای اجرای داده کاوی و بیان مفهوم درخت تصمیم و ارتباط آن با داده کاوی.

فهرست مطالب

فهرست جداول واشکال

چکیده

امروزه به دلیل وجود ابزار های مختلف برای جمع آوری داده ها و پیشرفت قابل قبول تکنولوژی پایگاه داده ، حجم انبوهی از اطلاعات در انبار داده های مختلف ذخیره شده است . این رشد انفجاری داده ها ، احتیاج به یک سری تکنیک ها و ابزار های جدید که توانایی پردازش هوشمندانه اطلاعات را دارا باشند ، نمایان میسازد.

داده کاوی با پیدا کردن مجموعه ای از الگوهای جالب از دل داده های موجود در انباره ها ، می تواند چنین نیازی را مرتفع کند.

در حال حاضر داده کاوی در پایگاه داده های بزرگ ، توسط بسیاری از محققان به عنوان یک موضوع تحقیقاتی مهم به شمار می آید.

محققان در بسیاری از رشته ها نظیر پایگاه داده ها ، یادگیری ماشین و آمار ، این موضوع را پیگیری کرده و تکنیک های مختلفی را برای داده کاوی ، تکنیک ها و روش های مختلف ارائه شده در این زمینه را معرفی کرده و آنها را طبقه بندی کند.

داده کاوی یکی از مهم ترن مراحل فرایند استخراج دانش در پایگاه داده به حساب می آید .

مراحل مختلف استخراج دانش در پایگاه داده ها به شرح ذیل است:

۱ .درک دامنه مسئله : شامل دانش های موجود و اهداف مسئله.

۲ .استخراج یک مجموعه داده : شامل انتخاب یک مجموعه داده ای و تمرکز ر روی قسمتی از داده ها.

۳ .آماده سازی و پاکسازی داده ها : شامل عملیات پایه ای نظیر حذف و تغییر داده های دارای اشکال.

٤ .یکپارچه سازی داده ها : شامل یکپارچه کردن منابع داده ای ناهمگون.

۵ .کاهش و تغییر شکل داده ها : شامل روش هایی برای تغییر شکل و کاهش ابعاد داده ها.

٦ .انتخاب نوع کاوش داده ها : شامل تعمیم و تقلیل ، طبقه بندی ، رگرسیون ، گروه بندی ، وب کاوی ، بازیابی تصویر ، کشف قوانین پیوندی و وابستگی های تابعی ، استخراج قوانین و یا ترکیبی از اینها.

۷ .انتخاب الگوریتم کاوش داده ها : شامل انتخاب متدهایی برای جست و جوی الگوها.

۸ .کاوش داده ها : شامل جست و جوی الگوهای جالب.

۹ .تفسیر : شامل تفسیر ، بازنمایی و آنالیز الگوی کشف شده.

۱۰ .استفاده از دانش کشف شده : شامل پیاده سازی دانش کشف شده در سیستم های اجرایی و اتخاذ تصمیماتی برپایه دانش مراحل مختلف کشف دانش

۱ مقدمه ای بر **داده‌کاوی**[1]

در دو دهه قبل توانایی های فنی بشر در برای تولید و جمع آوری داده ها به سرعت افزایش یافته است. عواملی نظیر استفاده گسترده از بارکد برای تولیدات تجاری، به خدمت گرفتن کامپیوتر در کسب و کار، علوم، خدمات دولتی و پیشرفت در وسائل جمع آوری داده، از اسکن کردن متون و تصاویر تا سیستمهای سنجش از دور ماهواره ای، در این تغییرات نقش مهمی دارند.

بطور کلی استفاده همگانی از وب و اینترنت به عنوان یک سیستم اطلاع رسانی جهانی ما را مواجه با حجم زیادی از داده و اطلاعات می‌کند. این رشد انفجاری در داده های ذخیره شده، نیاز مبرم وجود تکنولوژی های جدید و ابزارهای خودکاری را ایجاد کرده که به صورت هوشمند به انسان یاری رسانند تا این حجم زیاد داده را به اطلاعات و دانش تبدیل کند: داده کاوی به عنوان یک راه حل برای این مسائل مطرح می باشد. در یک تعریف غیر رسمی داده کاوی فرآیندی است، خودکار برای استخراج الگوهایی که دانش را بازنمایی می کنند، که این دانش به صورت ضمنی در پایگاه داده های عظیم، انباره داده [2] و دیگر مخازن بزرگ اطلاعات، ذخیره شده است. داده کاوی بطور همزمان از چندین رشته علمی بهره می برد نظیر: تکنولوژی

[1] Data Mining
[2] Data warehouses

پایگاه داده، هوش مصنوعی، یادگیری ماشین، شبکه های عصبی، آمار، شناسایی الگو، سیستم های مبتنی بر دانش [۳]، حصول دانش [٤]، بازیابی اطلاعات [٥]، محاسبات سرعت بالا [٦] و بازنمایی بصری داده [۷]. داده کاوی در اواخر دهه ۱۹۸۰ پدیدار گشته، در دهه ۱۹۹۰ گامهای بلندی در این شاخه از علم برداشته شده و انتظار می رود در این قرن به رشد و پیشرفت خود ادامه دهد.

واژه های «داده کاوی» و «کشف دانش در پایگاه داده» [۸] اغلب به صورت مترادف یکدیگر مورد استفاده قرار می گیرند. کشف دانش به عنوان یک فرآیند در شکل ۱-۱ نشان داده شده است.

کشف دانش در پایگاه داده فرایند شناسایی درست، ساده، مفید، و نهایتا الگوها و مدلهای قابل فهم در داده ها می باشد. داده کاوی، مرحله ای از فرایند کشف دانش می باشد و شامل الگوریتمهای مخصوص داده کاوی است، بطوریکه، تحت محدودیتهای مؤثر محاسباتی قابل قبول، الگوها و یا مدلها را در داده کشف می کند. به بیان ساده تر، داده کاوی به فرایند استخراج دانش ناشناخته، درست، و بالقوه مفید از داده اطلاق می شود. تعریف دیگر اینست که، داده کاوی گونه ای از تکنیکها برای شناسایی اطلاعات و یا دانش تصمیم گیری از قطعات داده می باشد، به نحوی که با استخراج آنها، در حوزه های تصمیم گیری، پیش بینی، پیشگویی، و تخمین مورد استفاده قرار گیرند. داده ها اغلب حجیم ، اما بدون ارزش می باشند، داده به

[3] Knowledge-based system
[4] Knowledge-acquisition
[5] Information retrieval
[6] High-performance computing
[7] Data visualization
[8] Knowledge Discovery in Database

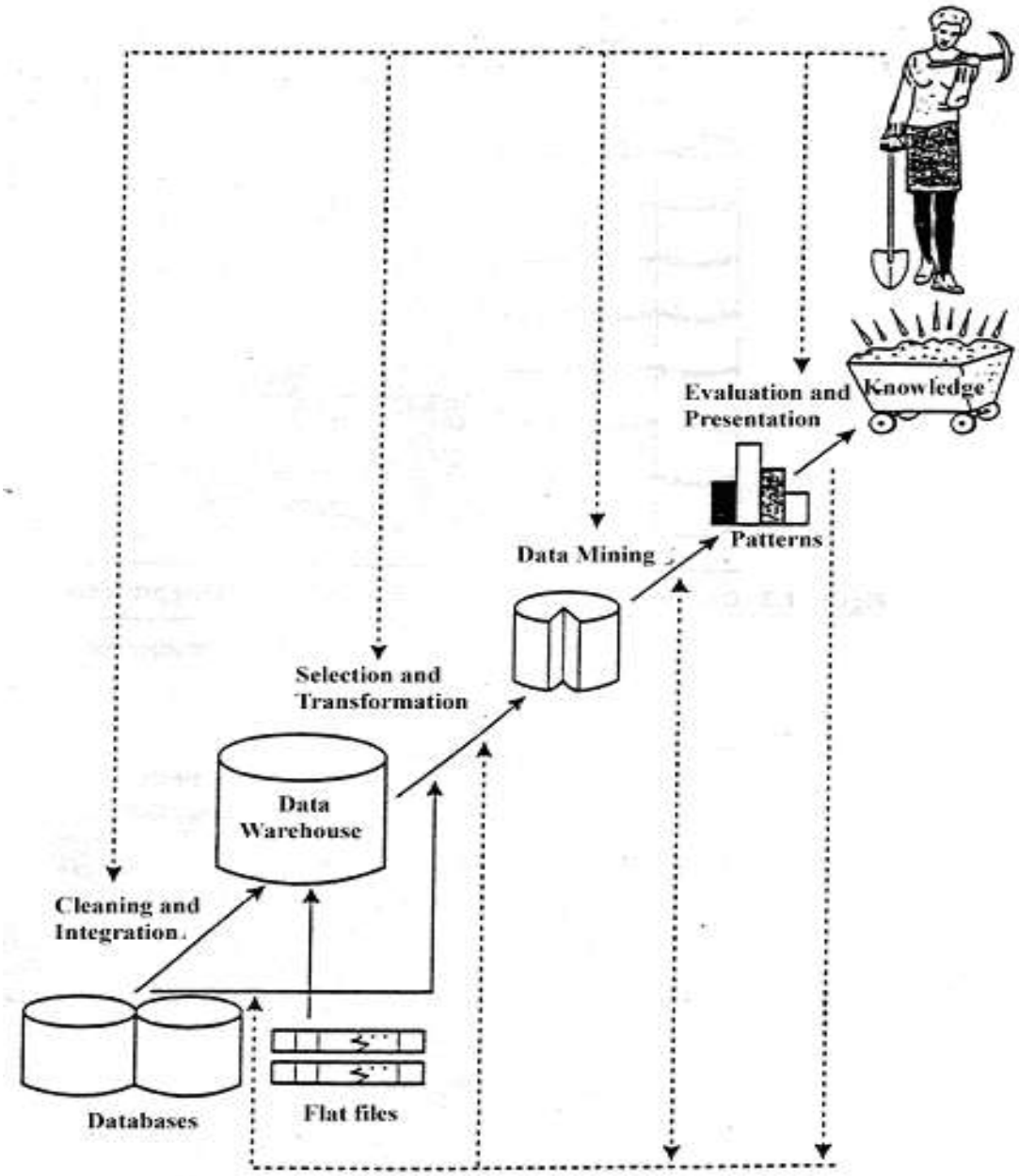
تنهایی قابل استفاده نیست، بلکه دانش نهفته در داده ها قابل استفاده می باشد. به این دلیل اغلب به داده کاوی، تحلیل داده ای ثانویه[9] گفته می شود.

۱-۱ چه چیزی سبب پیدایش داده کاوی شده است؟

اصلی ترین دلیلی که باعث شد داده کاوی کانون توجهات در صنعت اطلاعات قرار بگیرد، مساله در دسترس بودن حجم وسیعی از داده ها و نیاز شدید به اینکه از این داده ها اطلاعات و دانش سودمند استخراج کنیم. اطلاعات و دانش بدست آمده در کاربردهای وسیعی از مدیریت کسب و کار وکنترل تولید و تحلیل بازار تا طراحی مهندسی و تحقیقات علمی مورد استفاده قرار می گیرد.

داده کاوی را می توان حاصل سیر تکاملی طبیعی تکنولوژی اطلاعات دانست، که این سیر تکاملی ناشی از یک سیر تکاملی در صنعت پایگاه داده می باشد، نظیر عملیات: جمع آوری داده ها وایجاد پایگاه داده، مدیریت داده و تحلیل و فهم داده ها. در شکل ۱-۲ این روند تکاملی در پایگاه های داده نشان داده شده است.

[9] Secondary Data Analysis



شکل۱: داده کاوی به عنوان یک مرحله از فرآیند کشف دانش

تکامل تکنولوژی پایگاه داده و استفاده فراوان آن در کاربردهای مختلف سبب جمع آوری حجم فراوانی داده شده است. این داده های فراوان باعث ایجاد نیاز برای ابزارهای قدرتمند

برای تحلیل داده ها گشته، زیرا در حال حاضر به لحاظ داده ثروتمند هستیم ولی دچار کمبود اطلاعات می باشیم.

ابزارهای داده کاوی داده ها را آنالیز می کنند و الگوهای دادهای را کشف می کنند که می توان از آن در کاربردهایی نظیر: تعیین استراتژی برای کسب و کار، پایگاه دانش [10] و تحقیقات علمی و پزشکی، استفاده کرد. شکاف موجود بین داده ها و اطلاعات سبب ایجاد نیاز برای ابزارهای داده کاوی شده است تا داده های بی ارزش را به دانشی ارزشمند تبدیل کنیم.

به طور ساده داده کاوی به معنای استخراج یا «معدن کاری [11]» دانش از مقدار زیادی داده خام است. البته این نامگذاری برای این فرآیند تا حدی نامناسب است، زیرا به طور مثال عملیات معدن کاری برای استخراج طلا از صخره و ماسه را طلا کاوی می نامیم، نه ماسه کاوی یا صخره کاوی، بنابراین بهتر بود به این فرآیند نامی شبیه به «استخراج دانش از داده» می دادیم که متاسفانه بسیار طولانی است. «دانش کاوی» به عنوان یک عبارت کوتاهتر به عنوان جایگزین، نمی تواند بیانگر تاکید و اهمیت بر معدن کاری مقدار زیاد داده باشد. معدن کاری عبارتی است که بلافاصله انسان را به یاد فرآیندی می اندازد که به دنبال یافتن مجموعه کوچکی از قطعات ارزشمند از حجم بسیار زیادی از مواد خام هستیم.

[10] Knowledge base
[11] Mining

با توجه به مطالب عنوان شده، با اینکه این فرآیند تا حدی دارای نامگذاری ناقص است ولی این نامگذاری یعنی داده کاوی بسیار عمومیت پیدا کرده است. البته اسامی دیگری نیز برای این فرآیند پیشنهاد شده که بعضا بسیاری متفاوت با واژه داده کاوی است، نظیر: استخراج دانش از پایگاه داده، استخراج دانش[۱۲] ، آنالیز داده / الگو، باستان شناسی داده[۱۳]، و لایروبی داده ها[۱٤].

۱-۲ مراحل کشف دانش

کشف دانش دارای مراحل تکراری زیر است:

۱- پاکسازی داده ها[۱۵] (از بین بردن نویز و ناسازگاری داده ها).

۲- یکپارچه سازی داده ها[۱٦] (چندین منبع داده ترکیب می شوند).

۳- انتخاب داده ها[۱۷] (داده های مرتبط با آنالیزازپایگاه داده بازیابی می شوند).

٤- تبدیل کردن داده ها[۱۸] (تبدیل داده ها به فرمی که مناسب برای داده کاوی باشد مثل خلاصه سازی[۱۹] و همسان سازی[۲۰]

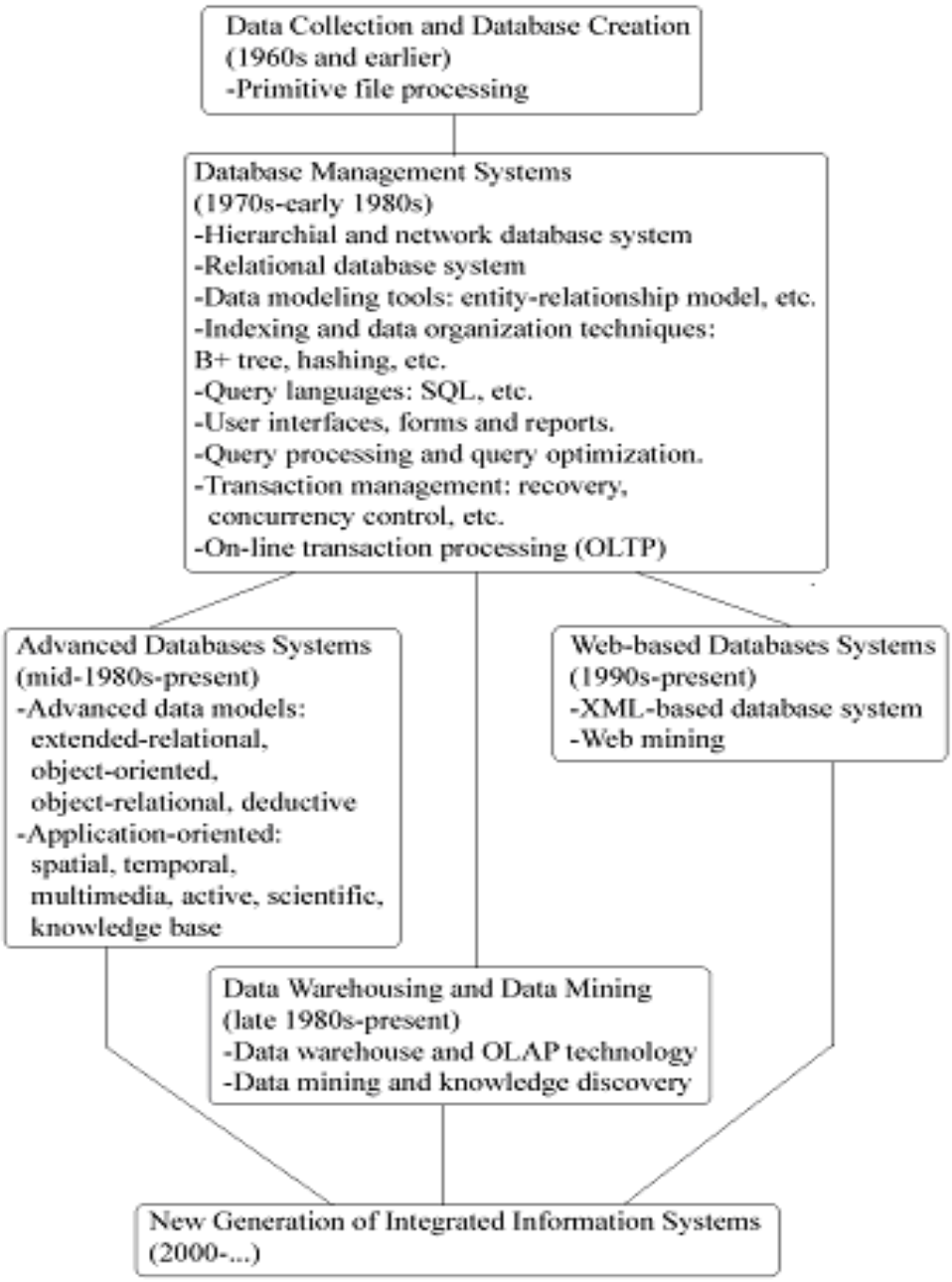
[12] Knowledge extraction
[13] Data archaeology
[14] Data dredging
[15] Data cleaning
[16] Data integration
[17] Data selection
[18] Data transformation
[19] Summary
[20] Aggregation

۵–داده کاوی (فرایند اصلی که روالهای هوشمند برای استخراج الگوها از داده ها به کار گرفته می شوند.)

٦–ارزیابی الگو[۲۱] (برای مشخص کردن الگوهای صحیح و مورد نظربه وسیله معیارهای اندازه گیری)

۷–ارائه دانش[۲۲] (یعنی نمایش بصری، تکنیکهای بازنمایی دانش برای ارائه دانش کشف شده به کاربر استفاده می شود).

[21] Pattern evaluation
[22] Knowledge presentation



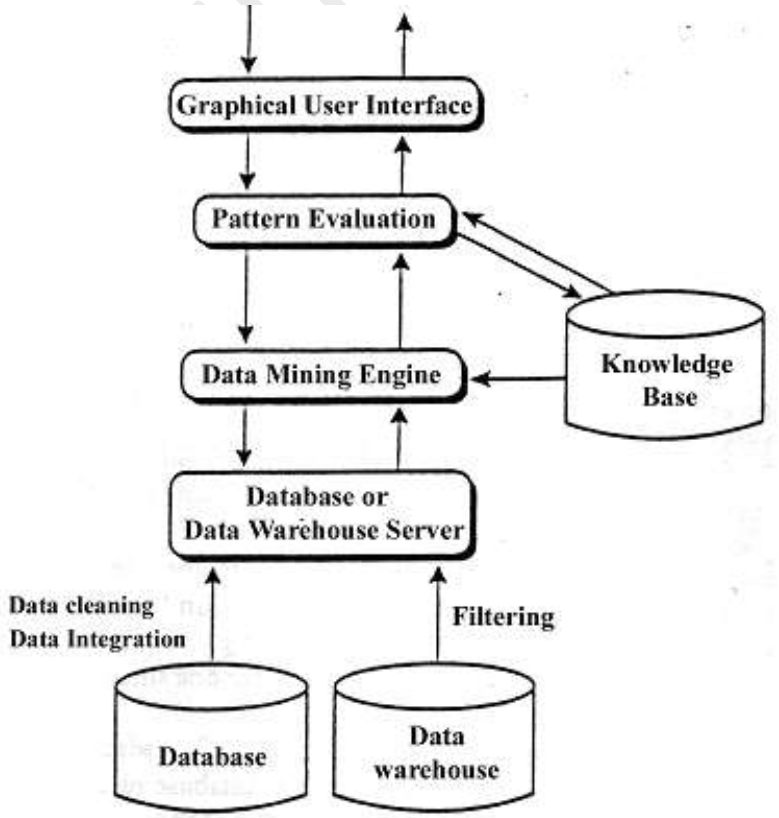
شکل۲: سیر تکاملی صنعت پایگاه داده

هر مرحله داده کاوی باید با کاربر یا پایگاه دانش تعامل داشته باشد. الگوهای کشف شده به کاربر ارائه می شوند و در صورت خواست او به عنوان دانش به پایگاه دانش اضافه می شوند.

توجه شود که بر طبق این دیدگاه داده کاوی تنها یک مرحله از کل فرآیند است، البته به عنوان یک مرحله اساسی که الگوهای مخفی را آشکار می سازد. با توجه به مطالب عنوان شده، دراینجا تعریفی از داده کاوی ارائه می دهیم:

"داده کاوی عبارتست از فرآیند یافتن دانش از مقادیر عظیم داده های ذخیره شده در پایگاه داده، انباره داده ویا دیگر مخازن اطلاعات" .

بر اساس این دیدگاه یک سیستم داده کاوی به طور نمونه دارای اجزاء اصلی زیر است که شکل۱-۳ بیانگر معماری سیستم است.



شکل۳: معماری یک نمونه سیستم داده کاو

۱- پایگاه داده، انباره داده یا دیگر مخازن اطلاعات: که از مجموعه ای از پایگاه داده ها، انباره داده، صفحه گسترده[۲۳]، یا دیگر انواع مخازن اطلاعات. پاکسازی داده ها و تکنیکهای یکپارچه سازی روی این داده ها انجام می شود.

۲- سرویس دهنده پایگاه داده یا انباره داده: که مسئول بازیابی داده های مرتبط بر اساس نوع درخواست داده کاوی کاربر می باشد.

۳- پایگاه دانش: این پایگاه از دانش زمینه [۲٤] تشکیل شده تا به جستجو کمک کند، یا برای ارزیابی الگوهای یافته شده از آن استفاده می شود.

٤- موتور داده کاوی [۲۵] : این موتور جزء اصلی از سیستم داده کاوی است و به طور ایدآل شامل مجموعه ای از پیمانه[۲٦] هایی نظیر توصیف[۲۷]، تداعی[۲۸]، کلاسبندی[۲۹]، آنالیزخوشه ها[۳۰]، و آنالیز تکامل وانحراف[۳۱]، است.

[23] Spread sheets
[24] Domain knowledge
[25] Data mining engine
[26] Module
[27] Characterization
[28] Association
[29] Classification
[30] Cluster analysis
[31] Evolution and deviation analysis

۵- پیمانه ارزیابی الگو [۳۲] : این جزء معیارهای جذابیت [۳۳] را به کار می بندد و با پیمانهء داده کاوی تعامل می کند بدینصورت که تمرکز آن بر جستجو بین الگوهای جذاب می باشد، و از یک حد آستانه جذابیت استفاده می کند تا الگوهای کشف شده را ارزیابی کند.

٦- واسط کاربرگرافیکی[۳٤] : این پیمانه بین کاربر و سیستم داده کاوی ارتباط برقرار می کند، به کاربر اجازه می دهد تا با سیستم داده کاوی از طریق پرس وجو [۳۵] ارتباط برقرار کند، این جزء به کاربر اجازه می دهد تا شمای پایگاه داده یا انباره داده را مرور کرده، الگوهای یافته شده را ارزیابی کرده و الگوها را در فرمهای بصری گوناگون بازنمایی کند.

با انجام فرآیند داده کاوی، دانش، ارتباط یا اطلاعات سطح بالا از پایگاه داده استخراج می شود و قابل مرور از دیدگاههای مختلف خواهد بود. دانش کشف شده در سیستم های تصمیم یار، کنترل فرآیند، مدیریت اطلاعات و پردازش پرس وجو[۳٦] قابل استفاده خواهد بود.

بنابراین داده کاوی به عنوان یکی از شاخه های پیشرو در صنعت اطلاعات مورد توجه قرار گرفته و به عنوان یکی از نوید بخش ترین زمینه های توسعه بین رشته ای در صنعت اطلاعات است.

[32] Pattern evaluation module
[33] Interesting measures
[34] Graphical User Interface (GUI)
[35] Query
[36] Query processing

۱-۳ جایگاه داده کاوی در میان علوم مختلف

ریشه های داده کاوی در میان سه خانواده از علوم، قابل پیگیری می باشد. مهمترین این خانواده ها، آمار کلاسیک [۳۷] می باشد. بدون آمار، هیچ داده کاوی وجود نخواهد داشت، بطوریکه آمار، اساس اغلب تکنولوژی هایی می باشد که داده کاوی بر روی آنها بنا می شود. آمار کلاسیک مفاهیمی مانند تحلیل رگرسیون، توزیع استاندارد، انحراف استاندارد، واریانس، تحلیل خوشه، و فاصله های اطمینان را که همه این موارد برای مطالعه داده و ارتباط بین داده ها می باشد، را در بر می گیرد. مطمئنا تحلیل آماری کلاسیک نقش اساسی در تکنیکهای داده کاوی ایفا می کند.

دومین خانواده ای که داده کاوی به آن تعلق دارد هوش مصنوعی [۳۸] می باشد. هوش مصنوعی که بر پایه روشهای ابتکاری می باشد و با آمار ضدیت دارد، تلاش دارد تا فرایندی مانند فکر انسان، را برای حل مسائل آماری بکار بندد. چون این رویکرد نیاز به توان محاسباتی بالایی دارد، تا اوایل دهه ۱۹۸۰ عملی نشد. هوش مصنوعی کاربردهای کمی را در حوزه های علمی و حکومتی پیدا کرد، اما نیاز به استفاده از کامپیوترهای بزرگ با عث شد همه افراد نتوانند از تکنیکهای ارائه شده استفاده کنند.

[37] Classic Statistics
[38] Artificial Intelligence

سومین خانواده داده کاوی، یادگیری ماشین [۳۹] می باشد، که به مفهوم دقیقتر، اجتماع آمار و هوش مصنوعی می باشد. درحالیکه هوش مصنوعی نتوانست موفقیت تجاری کسب کند، یادگیری ماشین در بسیاری از موارد جایگزین آن گردید. از یادگیری ماشین به عنوان تحول هوش مصنوعی یاد شد، چون مخلوطی از روشهای ابتکاری هوش مصنوعی به همراه تحلیل آماری پیشرفته می باشد. یادگیری ماشین اجازه می دهد تا برنامه های کامپیوتری در مورد داده ای که آنها مطالعه می کنند، مانند برنامه هایی که تصمیمهای متفاوتی بر مبنای کیفیت داده مطالعه شده می گیرند، یادگیری داشته باشند و برای مفاهیم پایه ای آن از آمار استفاده می کنند و از الگوریتمها و روشهای ابتکاری هوش مصنوعی را برای رسیدن به هدف بهره می گیرند.

داده کاوی در بسیاری از جهات، سازگاری تکنیکهای یادگیری ماشین با کاربردهای تجاری است. بهترین توصیف از داده کاوی بوسیله اجتماع آمار، هوش مصنوعی و یادگیری ماشین بدست می آید. این تکنیکها سپس با کمک یکدیگر، برای مطالعه داده و پیدا کردن الگوهای نهفته در آنها استفاده می شوند. بعضی از کاربردهای داده کاوی به شرح زیر است:

[39] Machine Learning

- **کاربردهای معمول تجاری:** از قبیل تحلیل و مدیریت بازار، تحلیل سبد بازار، بازاریابی هدف، فهم رفتار مشتری، تحلیل و مدیریت ریسک؛
- **مدیریت و کشف فریب:** کشف فریب تلفنی، کشف فریبهای بیمه ای و اتومبیل، کشف حقه های کارت اعتباری، کشف تراکنشهای مشکوک مالی (پولشویی)؛
- **متن کاوی**[40]: پالایش متن (نامه های الکترونیکی، گروههای خبری و غیره)؛
- **پزشکی:** کشف ارتباط علامت و بیماری، تحلیل آرایه های DNA ، تصاویر پزشکی؛
- **ورزش:** آمارهای ورزشی؛
- **وب کاوی**[41]: پیشنهاد صفحات مرتبط، بهبود ماشینهای جستجوگر یا شخصی سازی حرکت در وب سایت؛

۱-٤ داده کاوی چه کارهایی نمی تواند انجام دهد؟

داده کاوی فقط یک ابزار است و نه یک عصای جادویی. داده کاوی به این معنی نیست که شما راحت به کناری بنشینید و ابزارهای داده کاوی همه کار را انجام دهد.

[40] Text Mining
[41] Web Mining

داده کاوی نیاز به شناخت داده ها و ابزارهای تحلیل و افراد خبره در این زمینه ها را از بین نمی برد.

داده کاوی فقط به تحلیلگران برای پیدا کردن الگوها و روابط بین داده ها کمک می کند و در این مورد نیز روابطی که یافته می شود باید به وسیله داده های واقعی دوباره بررسی و تست گردد.

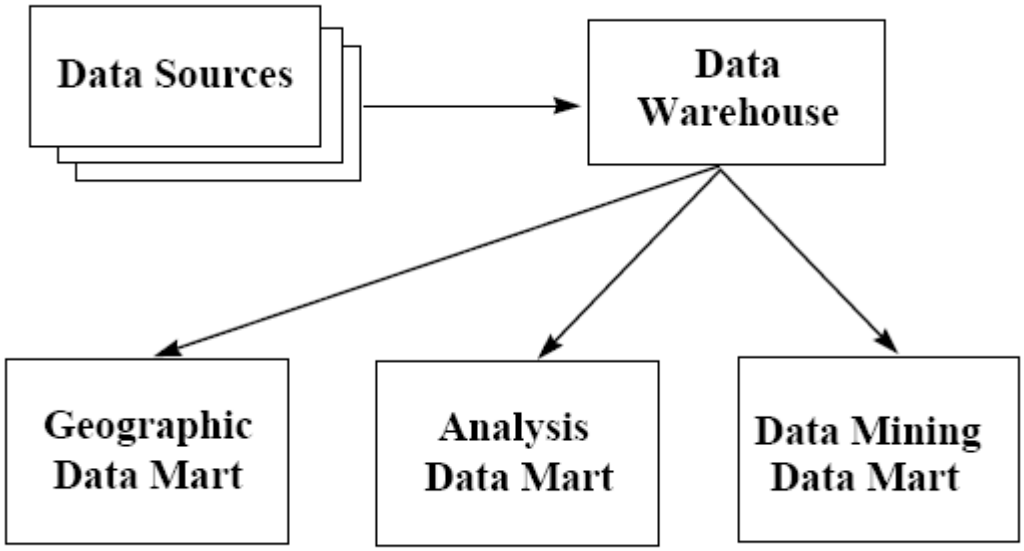
۱-۵ داده کاوی و انبار داده ها [۴۲]

معمولا داده هایی که در داده کاوی مورد استفاده قرار می گیرند از یک انبار داده استخراج می گردند و در یک پایگاه داده[۴۳] یا مرکز داده [۴۴]ای ویژه برای داده کاوی قرار می گیرند.

اگر داده های انتخابی جزیی از انبار داده ها باشند بسیار مفید است چون بسیاری از اعمالی که برای ساختن انباره داده ها انجام می گیرد با اعمال مقدماتی داده کاوی مشترک است و در نتیجه نیاز به انجام مجدد این اعمال وجود ندارد ، از جمله این اعمال پاکسازی داده ها می باشد.

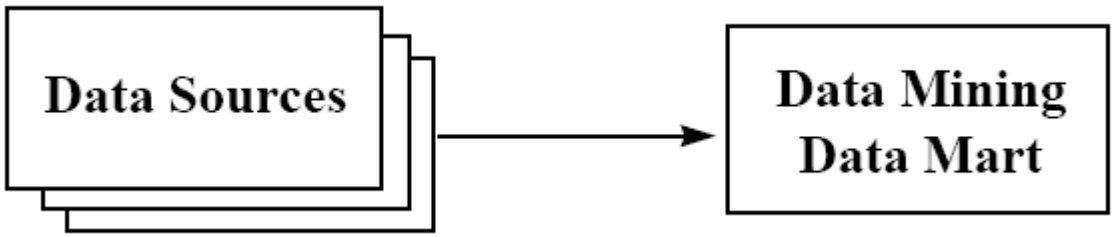
پایگاه داده مربوط به داده کاوی می تواند جزیی از سیستم انبار داده ها باشد و یا می تواند یک پایگاه داده جدا باشد.

[42] Data warehouse
[43] Database
[44] Data mart



شکل ٤ . داده ها از انباره داه ها استخراج می گردند

ولی با این حال وجود انباره داده ها برای انجام داده کاوی شرط لازم نیست و بدون آن هم اگر داده ها دریک یا چندین پایگاه داده باشند می توان داده کاوی را انجام دهیم و بدین منظور فقط کافیست داده ها را در یک پایگاه داده جمع آوری کنیم و اعمال جامعیت داده ها و پاکسازی داده ها را روی آن انجام دهیم. این پایگاه داده جدید مثل یک مرکز داده ای عمل می کند.



شکل ۵ .داده ها از چند پایگاه داده استخراج شده اند

۱-٦ داده کاوی و OLAP

بسیاری فکر می کنند که داده کاوی و OLAP دو چیز مشابه هستند در این بخش سعی می کنیم این مسئله را بررسی کنیم و همانطور که خواهیم دید این دو ابزار های کاملا متفاوت می باشند که می توانند همدیگر را تکمیل کنند.

OLAP جزیی از ابزارهای تصمیم گیری [45]می باشد.سیستم های سنتی گزارش گیری و پایگاه داده ای **آنچه** را که در پایگاه داده بود توضیح می دادند حال آنکه در OLAP هدف بررسی **دلیل** صحت یک فرضیه است.

بدین معنی که کاربر فرضیه ای در مورد داده ها و روابط بین آنها ارائه می کند و سپس به وسیله ابزار OLAP با انجام چند Query صحت آن فرضیه را بررسی می کند.

اما این روش برای هنگامی که داده ها بسیار حجیم بوده و تعداد پارامترها زیاد باشد نمیتواند مفید باشد چون حدس روابط بین داده ها کار سخت و بررسی صحت آن بسیار زمانبر خواهد بود.

تفاوت داده کاوی با OLAP در این است که داده کاوی برخلاف OLAP برای بررسی صحت یک الگوی فرضی استفاده نمی شود بلکه خود سعی می کند این الگوها را کشف کند.

درنتیجه داده کاوی و OLAP می توانند همدیگر را تکمیل کنند و تحلیل گر می تواند به وسیله ابزار OLAP یک سری اطلاعات کسب کند که در مرحله داده کاوی می تواند مفید باشد و همچنین الگوها و روابط کشف شده در مرحله داده کاوی می تواند درست نباشد که با اعمال تغییرات در آنها می توان به وسیله OLAP بیشتر بررسی شوند.

[45] Decision Support Tools

۱-۷ کاربرد یادگیری ماشین و آمار در داده کاوی

داده کاوی از پیشرفت هایی که در زمینه هوش مصنوعی و آمار رخ می دهد بهره می گیرد . هر دو این زمینه ها در مسائل شناسایی الگو و طبقه بندی داده ها کار می کنند و بالتبع در داده کاوی استفاده مستقیم خواهند داشت. و هر دو گروه در شناخت و استفاده از شبکه های عصبی و درخت های تصمیم گیری فعال می باشند.

داده کاوی جانشین تکنیک های آماری سابق نمی باشد بلکه وارث آنها بوده و در واقع تغییر و گسترش تکنیک های سابق برای متناسب ساز ی آنها با حجم داده ها و مسائل امروزی می باشد. تکنیک های کلاسیک برای داده های محدود و مسائل ساده مناسب بوده اند حال آنکه با پیچیده شدن مسائل و رشد روزافزون داده ها نیاز به تغییر آنها کاملا طبیعی است.به عبارت دیگر داده کاوی ترکیب تکنیک های کلاسیک با الگوریتم های جدید مثل شبکه های عصبی و درخت تصمیم گیری می باشد.

مهمترین نکته این است که داده کاوی راهکاری است برای مسائل تجاری امروز به کمک تکنیک های آماری و هوش مصنوعی برای افراد حرفه ای که قصد دارند یک مدل پیش بینی ایجاد نمایند.

۲- توصیف داده ها در داده کاوی

۲-۱ خلاصه سازی و به تصویر در آوردن داده ها

قبل از اینکه بتوان روی مجموعه ای از داده ها ،داده کاوی انجام بدهیم و یک مدل پیش بینی مناسب ابجاد کنیم ، باید بتوان داده ها را به خوبی شناخت که برا ی شروع این کار می توان از پارامترهایی مثل میانگین , انحراف معیار و.... استفاده کنیم.

ابزارهای تصویرسازی داده ها و گراف سازی برای شناخت داده ها بسیار مفید می باشند و نقش آنها در آماده سازی داده ها بسیار مفید و غیر قابل انکار است ، مثلا با استفاده از این ابزار می توان توزیع مقادیر مختلف داده ها را در یک نمودار مشاهده کرد و میزان داده های دارای خطا را به طور تقریبی حدس زد.

مهمترین مشکل این ابزار این است که معمولا تحلیل ها دارای تعداد زیادی پارامتر هستند که به هم مربوطند و باید رابطه این پارامترها را که چند بعدی می باشد در دو بعد نمایش دهند که این کار اگر هم عملی باشد برای استفاده از آنها نیاز به افراد خبره می باشد.

۲-۲ خوشه بندی [46]

هدف از خوشه بندی این است که داده های موجود را به چند گروه تقسیم کنند و در این تقسیم بندی داده های گروه های مختلف باید حداکثر تفاوت ممکن را به هم داشته باشند و داده های موجود در یک گروه باید بسیار به هم شبیه باشند .

برخلاف کلاس بندی (که در ادامه خواهیم دید) در خوشه بندی ، گروه ها از قبل مشخص نمی باشند و همچنین معلوم نیست که بر حسب کدام خصوصیات گروه بندی صورت می گیرد. درنتیجه پس از انجام خوشه بندی باید یک فرد خبره خوشه های ایجاد شده را تفسیر کند و در بعضی مواقع لازم است که پس از بررسی خوشه ها بعضی از پارامترهایی که در خوشه بندی در نظر گرفته شده اند ولی بی ربط بوده یا اهمیت چندانی ندارند حذف شده و جریان خوشه بندی از اول صورت گیرد.

پس از اینکه داده ها به چند گروه منطقی و توجیه پذیر تقسیم شدند از این تقسیم بندی می توان برای کسب اطلاعات در مورد داده ها یا تقسیم داده ها جدید استفاده کنیم.

[46] Clustering

از مهمترین الگوریتم هایی که برای خوشه بندی استفاده می شوند می توان Kohnen و الگوریتم K-means را نام برد.

۲-۳ تحلیل لینک[47]

تحلیل داده ها یکی از روش های توصیف داده هاست که به کمک آن داده ها را بررسی کرده و روابط بین مقادیر موجود در بانک اطلاعاتی را کشف می کنیم.از مهمترین راههای تحلیل لینک **کشف وابستگی** [48]و **کشف ترتیب** [49]می باشد.

منظور از کشف وابستگی یافتن قوانینی در مورد مورادی است که با هم اتفاق می افتند مثلا اجناسی که در یک فروشگاه احتمال خرید همزمان آنها زیاد است.

کشف ترتیب نیر بسیار مشابه می باشد ولی پارامتر زمان نیز در آن دخیل می باشد.

وابستگی ها به صورت A->B نمایش داده می شوند که به A مقدم و به B موخر یا نتیجه گفته می شود.مثلا اگر یک قانون به صورت زیر داشته باشیم :

" اگر افراد چکش بخرند آنگاه آنها میخ خواهند خرید "

در این قانون مقدم **خرید چکش** و نتیجه **خرید میخ** می باشد.

[47] Link Analysis
[48] Association discovery
[49] Sequence discovery

۳- مدل های پیش بینی داده ها

۳-۱ Classification

در مسائل classification هدف شناسایی ویژگیهایی است که گروهی را که هر مورد به آن تعلق دارد را نشان دهند. از این الگو میتوان هم برای فهم دادههای موجود و هم پیشبینی نحوه رفتار مواد جدید استفاده کرد.

دادهکاوی مدلهایclassification را با بررسی دادههای دستهبندی شده قبلی ایجاد میکند و یک الگوی پیشبینی کننده را بصورت استقرایی مییابند. این موارد موجود ممکن است از یک پایگاه داده تاریخی آمده باشند.]**Error! Reference source not found.**[

۳-۲ Regression

Regression از مقادیر موجود برای پیشبینی مقادیر دیگر استفاده میکند. در سادهترین فرم، regression از تکنیکهای آماری استاندارد مانند linear regression استفاده می-کند. متاسفانه، بسیاری مسائل دنیای واقع تصویرخطی سادهای از مقادیر قبلی نیستند. بناراین

تکنیکهای پیچیدهتری(logistic regression، درختهای تصمیم، یا شبکههای عصبی) ممکن است برای پیشبینی مورد نیاز باشند.

انواع مدل یکسانی را میتوان هم برای regression و هم برای classification استفاده کرد. برای مثال الگوریتم درخت تصمیم CART را میتوان هم برای ساخت درختهای classification و هم درختهای regression استفاده کرد. شبکههای عصبی را نیز می - توان برای هر دو مورد استفاده کرد .

۳-۳ Time series

پیشبینی های Time series مقادیر ناشناخته آینده را براساس یک سری از پیشبینی گرهای متغیر با زمان پیشبینی میکنند. و مانند regression، از نتایج دانسته شده برای راهنمایی پیشبینی خود استفاده میکنند. مدلها باید خصوصیات متمایز زمان را در نظر گیرند و بویژه سلسلهمراتب دورهها را.

٤ مدل ها و الگوریتم های داده کاوی

در این بخش قصد داریم مهمترین الگوریتم ها و مدل های داده کاوی را بررسی کنیم. بسیاری از محصولات تجاری داده کاوی از مجموعه از این الگوریتم ها استفاده می کنند و معمولا هر کدام آنها در یک بخش خاص قدرت دارند وبرای استفاده از یکی از آنها باید بررسی های لازم در جهت انتخاب متناسب ترین محصول توسط گروه متخصص در نظر گرفته شود.

نکته مهم دیگر این است که در بین این الگوریتم ها و مدل ها ، **بهترین** وجود ندارد و با توجه به داده ها و کارایی مورد نظر باید مدل انتخاب گردد.

٤-١ شبکه های عصبی [50]

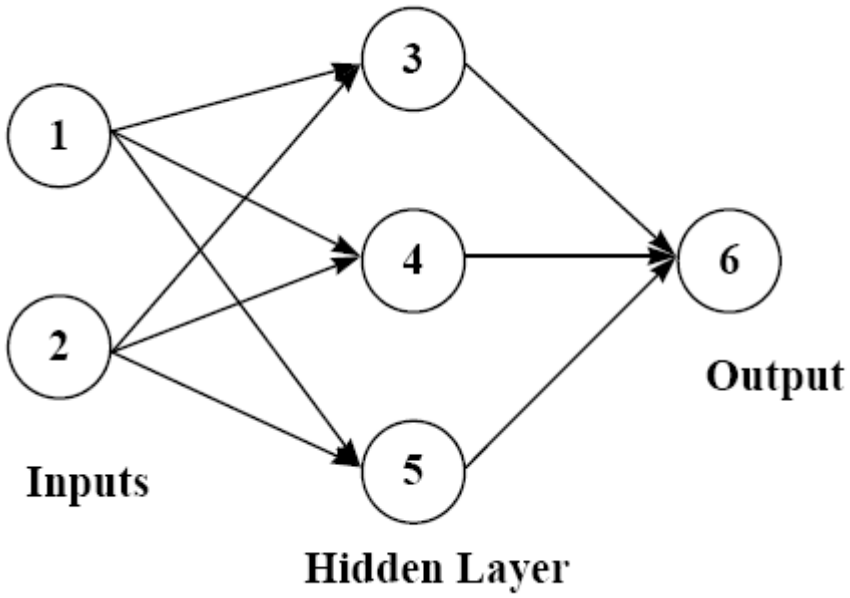
شبکه های عصبی از پرکاربردترین و عملی ترین روش های مدل سازی مسائل پیچیده و بزرگ که شامل صدها متغیر هستند می باشد. شبکه های عصبی می توانند برای مسائل کلاس بندی (که خروجی یک کلاس است) یا مسائل رگرسیون (که خروجی یک مقدار عددی است) استفاده شوند.

[50] Neural Networks

هر شبکه عصبی شامل یک لایه ورودی [۵۱]می باشد که هر گره در این لایه معادل یکی از متغیرهای پیش بینی می باشد. گره های موجود در لایه میانی وصل می شوند به تعدادی گره در لایه نهان[۵۲] . هر گره ورودی به همه گره های لایه نهان وصل می شود.

گره های موجود در لایه نهان می توانند به گره های یک لایه نهان دیگر وصل شوند یا می توانند به لایه خروجی [۵۳]وصل شوند.

لایه خروجی شامل یک یا چند متغیر خروجی می باشد.



شکل شماره٦

[51] Input Layer
[52] Hidden Layer
[53] Output Layer

شبکه عصبی با یک لایه نهان

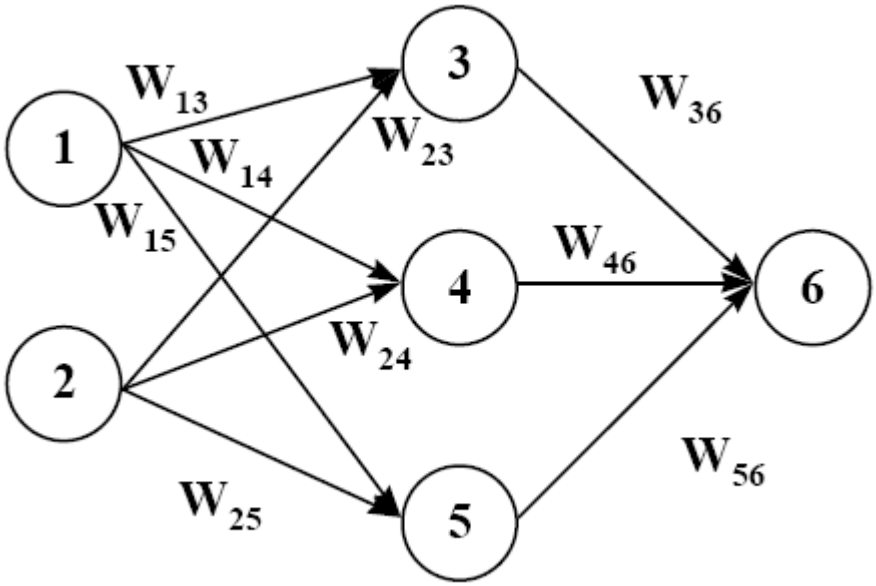
هر یال که بین نود های X,Y می باشد دارای یک وزن است که با Wx,y نمایش داده می شود. این وزن ها در محاسبات لایه های میانی استفاده می شوند و طرز استفاده آنها به این صورت است که هر نود در لایه های میانی (لایه های غیر از لایه اول) دارای چند ورودی از چند یال مختف می باشدکه همانطور که گفته شد هر کدام یک وزن خاص دارند.

هر نود لایه میانی میزان هر ورودی را در وزن یال مربوطه آن ضرب می کند و حاصل این ضرب ها را با هم جمع می کند و سپس یک تابع از پیش تعیین شده (تابع فعال سازی) روی این حاصل اعمال می کند و نتیجه را به عنوان خروجی به نودهای لایه بعد می دهد.

وزن یال ها پارامترهای ناشناخته ای هستند که توسط تابع آموزش [54]و داده های آموزشی که به سیستم داده می شود تعیین می گردند.

تعداد گره ها و تعداد لایه های نهان و نحوه وصل شدن گره ها به یکدیگر معماری (توپولوژی) شبکه عصبی را مشخص می کند. کاربر یا نرم افزاری که شبکه عصبی را طراحی می کند باید تعداد نودها ، تعداد لایه های نهان ، تابع فعال سازی و محدودیت های مربوط به وزن یال ها را مشخص کند.

[54] Training method



شکل شماره (7)

Wx,y وزن یال بین X و Y است.

از مهمترین انواع شبکه های عصبی Feed-Forward Backpropagation می باشد که در اینجا به اختصار آنرا توضیح می دهیم.

Feed-Forward به معنی این است که مقدار پارامتر خروجی براساس پارامترهای ورودی و یک سری وزن های اولیه تعیین می گردد. مقادیر ورودی با هم ترکیب شده و در لایه های

نهان استفاده می شوند و مقادیر این لایه های نهان نیز برای محاسبه مقادیر خروجی ترکیب می شوند.

Backpropagation : خطای خروجی با مقایسه مقدار خروجی با مقدار مد نظر در داده های آزمایشی محاسبه می گردد و این مقدار برای تصحیح شبکه و تغییر وزن یال ها استفاده می گردد و از گره خروجی شروع شده و به عقب محاسبات ادامه می یابد.

این عمل برای هر رکورد موجود در بانک اطلاعاتی تکرار می گردد.

به هر بار اجرای این الگوریتم برای تمام داده های موجود در بانک یک **دوره** [55]گفته می شود. این دوره ها آنقدر ادامه می یابد که دیگر مقدار خطا تغییر نکند.

از آنجایی که تعداد پارامترها در شبکه های عصبی زیاد می باشد محاسبات این شبکه ها می تواند وقت گیر باشد. ولی اگر این شبکه ها به مدت کافی اجرا گردند معمولا موفقیت آمیز خواهند بود. مشکل دیگری که ممکن است به وجود بیاید Overfitting می باشد و آن بدین صورت است که که شبکه فقط روی داده ها آموزشی خوب کار می کند و برای سایر مجموعه داده ها مناسب نمی باشد. برای رفع این مشکل ما باید بدانیم چه زمانی آموزش شبکه را متوقف کنیم.یکی از راه ها این است که شبکه را علاوه بر داده های آزمایشی روی داده های تست نیز مرتبا اجرا کنیم و جریان تغییر خطا را در آنها بررسی کنیم.اگر در این داده ها به

[55] Epoch

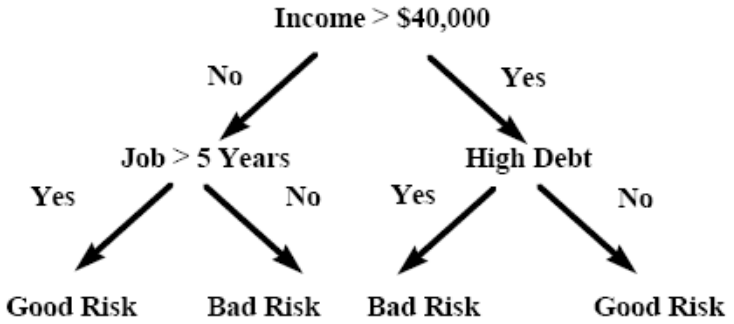
جایی رسیدیم که میزان خطا رو به افزایش بود حتی اگر خطا در داده های آزمایشی همچنان رو به کاهش باشد آموزش را متوقف کنیم.

از آنجایی که پارامترهای شبکه های عصبی زیاد است یک خروجی خاص می تواند با مجموعه های مختلفی از مقادیر پارامترها ایجاد گردد درنتیجه این پارامترها مثل وزن یالها قابل تفسیر نبوده و معنی خاصی نمی دهند .

یکی از مهمترین فواید شبکه های عصبی قابلیت اجرای آنها روی کامپیوترهای موازی می باشد.

٤-٢ Decision trees

درختهای تصمیم روشی برای نمایش یک سری از قوانین هستند که منتهی به یک رده یا مقدار میشوند. برای مثال، میخواهیم متقاضیان وام را به دارندگان ریسک اعتبار خوب و بد تقسیم کنیم. شکل یک درخت تصمیم را که این مسئله را حل میکد نشان میدهد و همه مؤلفههای اساسی یک یک درخت تصمیم در آن نشان داده شده است : نود تصمیم، شاخهها و برگها.



شکل شماره (۸) درخت تصمیم گیری

براساس الگوریتم، ممکن است دو یا تعداد بیشتری شاخه داشته باشد. برای مثال، CART درختانی با تنها دو شاخه در هر نود ایجاد میکند. هر شاخه منجر به نود تصمیم دیگر یا یک نود برگ میشود. با پیمایش یک درخت تصمیم از ریشه به پایین به یک مورد یک رده یا مقدار نسبت میدهیم. هر نود از دادههای یک مورد برای تصمیمگیری درباره آن انشعاب استفاده میکند.

درختهای تصمیم از طریق جداسازی متوالی دادهها به گروههای مجزا ساخته میشوند و هدف در این فرآیند افزایش فاصله بین گروهها در هر جداسازی است.

یکی از تفاوتها بین متدهای ساخت درخت تصمیم اینستکه این فاصله چگونه اندازهگیری میشود. درختهای تصمیمی که برای پیشبینی متغیرهای دستهای استفاده میشوند، درخت - های classification نامیده میشوند زیرا نمونهها را در دستهها یا ردهها قرار میدهند.

درختهای تصمیمی که برای پیشبینی متغیرهای پیوسته استفاده میشوند درختهای regression نامیده میشوند.

هر مسیر در درخت تصمیم تا یک برگ معمولا قابل فهم است. از این لحاظ یک درخت تصمیم میتواند پیشبینیهای خود را توضیح دهد، که یک مزیت مهم است. با این حال این وضوح ممکن است گمراهکننده باشد. برای مثال، جداسازی های سخت در درختهای تصمیم دقتی را نشان میدهند که کمتر در واقعیت نمود دارند. (چرا باید کسی که حقوق او ٤٠٠٠٠١ است از نظر ریسک اعتبار خوب باشد درحالیکه کسی که حقوقش ٤٠٠٠٠ است بد باشد. بعلاوه، از آنجاکه چندین درخت میتوانند دادههای مشابهای را با دقت مشابه نشان دهند، چه تفسیری ممکن است از قوانین شود؟

درختهای تصمیم تعداد دفعات کمی از دادهها گذر میکنند(برای هر سطح درخت حداکثر یک مرتبه) و با متغیرهای پیشبینیکننده زیاد بخوبی کار میکنند. درنتیجه، مدلها بسرعت ساخته میشوند، که آنها را برای مجموعهداده های بسیار مناسب میسازد. اگر به درخت اجازه دهیم بدون محدودیت رشد کند زمان ساخت بیشتری صرف میشود که غیرهوشمندانه است، اما مسئله مهمتر اینستکه با دادهها overfit میشوند. اندازه درختها را میتوان از طریق قوانین توقف کنترل کرد. یک قانون معمول توقف محدود کردن عمق رشد درخت است.

راه دیگر برای توقف هرس کردن درخت است. درخت میتواند تا اندازه نهایی گسترش یابد، سپس با استفاده از روشهای اکتشافی توکار یا با مداخله کاربر، درخت به کوچکترین اندازهای که دقت در آن از دست نرود کاهش مییابد.

یک اشکال معمول درختهای تصمیم اینستکه آنها تقسیمکردن را براساس یک الگوریتم حریصانه انجام میدهند که در آن تصمیمگیری اینکه براساس کدام متغیر تقسیم انجام شود، اثرات این تقسیم در تقسیمهای آینده را درنظر نمیگیرد.

بعلاوه الگوریتمهایی که برای تقسیم استفاده میشوند، معمولا تکمتغیری هستند: یعنی تنها یک متغیر را در هر زمان در نظر میگیرند. درحالیکه این یکی از دلایل ساخت سری مدل است، تشخیص رابطه بین متغیرهای پیشبینی کننده را سختتر میکند.

٣-٤ Multivariate Adaptive Regression Splines(MARS)

در میانههای دهه ٨٠ یکی از مخترعین CART، Jerome H. Friedman، متدی را برای برطرفکردن این کاستیها توسعه داد.

کاستیهای اساسی که او قصد برطرف کردن آنها را داشت عبارتند از :

- پیشبینی های غیرپیوسته(تقسیم سخت)
- وابستگی همه تقسیمها به تقسیمهای قبلی

به این دلیل او الگوریتم MARS را توسعه داد. ایده اصلی MARS نسبتا ساده است، درحالیکه خود الگوریتم نسبتا پیچیده است. بسیار ساده ایده عبارت است از :

- جایگزینی انشعابهای غیرپیوسته با گذر های پیوسته که توسط یک جفت از خطهای مستقیم مدل میشوند. در انتهای فرآیند ساخت مدل، خطوط مستقیم در هر نود با یک تابع بسیار هموار که spline نامیده میشود جایگزین میشوند.
- عدم نیاز به اینکه تقسیمهای جدید وابسته به تقسیمهای قدیمی باشند.

متأسفانه این به معنی اینست که MARS ساختار درختی CART را ندارد و نمیتواند قوانینی را ایجاد کند. از طرف دیگر، MARS به صورت خودکار مهمترین متغیرهای پیشبینی کننده و همچنین تعامل میان آنها را مییابد. MARS همچنین وابستگی میان پاسخ و هر پیشبینی کننده را معین میکند. نتیجه ابزار رگرسیون اتوماتیک، خودکار و step-wise است.

MARS، مانند بیشتر الگوریتمهای شبکههای عصبی و درخت تصمیم، تمایل به overfit شدن برای دادههای آموزشدهنده دارد. که میتوان آنرا به دو طریق درست کرد. اول اینکه،

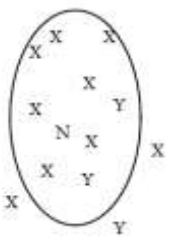
cross validation بصورت دستی انجام شود و الگوریتم برای تولید پیشبینی خوب روی مجوعه تست تنظیم شود. دوم اینکه، پارامترهای تنظیم متفاوتی در خود الگوریتم وجود دارد که cross validation درونی را هدایت میکند.

٤-٤ Rule induction

استنتاج قوانین متدی برای تولید مجموعهای از قوانین است که موارد را دستهبندی میکند. اگرچه درختهای تصمیم میتوانند مجموعهای از قوانین را ایجاد کند، متدهای استنتاج قوانین مجموعهای از قوانین مستقل را ایجاد میکند. که لزوما یک درخت را ایجاد نمیکنند. از آنجا که استنتاجگر قوانین اجباری به تقسیم در هر سطح ندارد، و میتواند به آینده بنگرد، قادر است الگوهای متفاوت و گاها بهتری برای ردهبندی بیابد. برخلاف درختان، قوانین ایجاد شده ممکن است همه موارد ممکن را نپوشاند. همچنی« برخلاف درختان، قوانین ممکن است در پیشبینی متعارض باشند، که در هر مورد باید قانونی را برای دنبال کردن انتخاب کرد. یک روش برای حل این تعارضات انتصاب یک میزان اطمینان به هر قانون است و استفاده از قانونی است که میزان اطمینان بالاتری دارد.

٤-٥ K-nearest neibour and memory-based reansoning(MBR)

هنگام تلاش برای حل مسائل جدید، افراد معمولا به راهحل های مسائل مشابه که قبلا حل شدهاند مراجعه میکنند. K-nearest neighbor(k-NN) یک تکنیک دستهبندی است که از نسخهای از این متد استفاده میکند. در این روش تصمیمگیری اینکه یک مورد جدید در کدام دسته قرار گیرد با بررسی تعدادی(k) از شبیهترین موارد یا همسایهها انجام میشود. تعداد موارد برای هر کلاس شمرده میشوند، و مورد جدید به دستهای که تعداد بیشتری از همسایهها به آن تعلق دارند نسبت داده میشود.



محدوده همسایگی (بیستر همسایه ها در دسته X قرار گرفته اند)

اولین مورد برای بکاربردن k-NN یافتن معیاری برای فاصله بین صفات در دادهها و محاسبه آن است. در حالیکه این عمل برای دادههای عددی آسان است، متغیرهای دستهای نیاز به برخورد خاصی دارند. هنگامیکه فاصله بین مواد مختلف را توانستیم اندازه گیریم، میتوانیم از

مجموعه مواردی که قبلا دستهبندی شدهاند را بعنوان پایه دستهبندی موارد جدید استفاده کنیم، فاصله همسایگی را تعیین کنیم، و تعیین کینم که خود همسایهها را چگونه بشماریم.

K-NN بار محاسباتی زیادی را روی کامپیوتر قرار میدهد زیرا زمان محاسبه بصورت فاکتوریلی از تمام نقاط افزایش مییابد. درحالیکه بکابردن درخت تصمیم یا شبکه عصبی برای یک مورد جدید فرایند سریعی است، K-NN نیاز به محاسبه جدیدی برای هر مورد جدید دارد. برای افزایش سرعت K-NN معمولا تمام دادهها در حافظه نگهداری میشوند.

فهم مدلهای K-NN هنگامیکه تعداد متغیرهای پیشبینی کننده کم است بسیار ساده است. آنها همچنین برای ساخت مدلهای شامل انواع داده غیر استاندارد هستند، مانند متن بسیار مفیدند. تنها نیاز برای انواع داده جدید وجود معیار مناسب است.

٤-٦ رگرسیون منطقی[56]

رگرسیون منطقی یک حالت عمومی تر از رگرسیون خطی می باشد.قبلا این روش برای پیش بینی مقادیر باینری یا متغیرهای دارای چند مقدار گسسته (کلاس) استفاده می شد. از آنجایی که مقادیر مورد نظر برای پیش بینی مقادیر گسسته می باشند نمی توان آنرا به روش رگرسیون خطی مدلسازی کرد برای این منظور این متغیرهای گسسته را به روشی تبدیل به متغیر عددی

[56] Logistic regression

و پیوسته می کنیم وبرای این منظور مقدار لگاریتم احتمال متغیر مربوطه را در نظر می گیریم و برای این منظور احتمال پیشامد را بدین صورت در نظر می گیریم :

احتمال اتفاق نیفتادن پیشامد/ احتمال اتفاق افتادن پیشامد

و تفسیر این نسبت مانند تفسیری است که در بسیاری از مکالمات روزمره در مورد مسابقات یا شرط بندی ها یه موارد مشابه به کار می رود .مثلا وقتی می گوییم شانس بردن یک تیم در مسابقه ۳ به ۱ است در واقع از همین نسبت استفاده کرده و معنی آن این است که احتمال برد آن تیم ۷۵٪ است.

وقتی که ما موفق شدیم لگاریتم احتمال مورد نظر را بدست آوریم با اعمال لگاریتم معکوس می توان نسبت مورد نظر و از روی آن کلاس مورد نظر را مشخص نمود.

٤-۷ تحلیل تفکیکی ۵۷

این روش از قدیمی ترین روش های ریاضی وار گروه بندی داده ها می باشد که برای اولین بار در سال ۱۹۳٦ توسط فیشر استفاده گردید. روش کار بدین صورت است که داده ها را مانند داده های چند بعدی بررسی کرده و بین داده ها مرزهایی ایجاد می کنند (برای داده ها دو بعدی خط جدا کننده، برای داده های سه بعدی سطح جدا کننده و ..) که این مرزها مشخص

[57] Discriminant analysis

کننده کلاس های مختلف می باشند و بعد برای مشخص کردن کلاس مربوط به داده های جدید فقط باید محل قرارگیری آن را مشخص کنیم.

این روش از ساده ترین و قابل رشدترین روش های کلاس بندی می باشد که در گذشته بسیار استفاده می شد.

این روش به سه دلیل محبوبیت خود را از دست داد :اول اینکه این روش فرض می کند همه متغیرهای پیش بینی به صورت نرمال توزیع شده اند که در بسیاری از موارد صحت ندارد . دوم اینکه داده هایی که به صورت عددی نمی باشند مثل رنگها در این روش قابل استفاده نمی باشند. سوم اینکه در این روش فرض می شود که مرزهای جدا کننده داده ها به صورت اشکال هندسی خطی مثل خط یا سطح می باشند حال اینکه این فرض همیشه صحت ندارد.

نسخه های اخیر تحلیل تفکیکی بعضی از این مشکلات را رفع کرده اند به این طریق اجازه می دهند مرزهای جدا کننده بیشتر از درجه ۲ نیز باشند که باعث بهبود کارایی و حساسیت در بسیاری از موارد می گردد.

٤-٨ مدل افزودنی کلی (GAM)[58]

این روش ها در واقع بسطی بر روش های رگرسیون خطی و رگرسیون منطقی می باشند. به این دلیل به این روش افزودنی می گویند که فرض می کنیم می توانیم مدل را به صورت مجموع چند تابع غیر خطی (هر تابع برای یک متغیر پیش بینی کننده) بنویسیم. GAM می تواند هم به منظور رگرسیون و هم به منظور کلاس بندی داده ها استفاده گردد. این ویژگی غیر خطی بودن توابع باعث می شود که این روش نسبت به روشهای رگرسیون خطی بهتر باشد .

٤-٩ Boosting

در این روش ها مبنی کار این است که الگوریتم پیش بینی را چندین بار و هر بار با داده های آموزشی متفاوت (که با توجه به اجرای قبلی انتخاب می شوند) اجرا کنیم و در نهایت آن جوابی که بیشتر تکرار شده را انتخاب کنیم. این روش اگر چه وقت گیر است ولی جواب های آن مطمئن تر خواهند بود. این روش اولین بار در سال ١٩٩٦ استفاده شد و در این روزها با توجه به افزایش قدرت محاسباتی کامپیوترها بر مقبولیت آن افزوده گشته است.

[58] Generalized Additive Models

5- سلسله مراتب انتخابها

هدف دادهکاوی تولید دانش جدیدی است که کاربر بتواند از آن استفاده کند. این هدف با ساخت مدلی از دنیای واقع براساس دادههای جمعآوری شده از منابع متفاوت بدست میآید. نتیجه ساخت این مدل توصیفی از الگوها و روابط دادههاست که میتوان آنرا برای پیشبینی استفاده کرد. سلسه انتخابهایی که قبل از آغاز باید انجام شود به این شرح است :

- هدف تجاری
- نوع پیشبینی
- نوع مدل
- الگوریتم
- محصول

در بالاترین سطح **هدف تجاری** قرار دارد: ه د ف نهایی از کاوش دادهها چیست؟ برای مثال، جستجوی الگوها در دادهها ممکن است برای حفظ مشتریهای خوب باشد، که ممکن است مدلی برای سودبخشی مشتریها و مدل دومی برای شناسایی مشتریهایی که ممکن از دست دهیم میسازیم. اطلاع از اهداف و نیازهای سازمان ما را در فرموله کردن هدف سازمان یاری میرساند.

مرحله بعدی تصمیمگیری درباره **نوع پیشبینی** مناسب است:

(۱) classification : پیشبینی اینکه یک مورد در کدام گروه یا رده قرار میگیرد. یا (۲) regression : پیشبینی اینکه یک متغیر عددی چه مقداری خواهد داشت.

مرحله بعدی انتخاب **نوع مدل** است: یک شبکه عصبی برای انجام regression ،و یک درخت تصمیم برای classification. همچنین روشهای مرسوم آماری برای مانند logistic regression، discriminant analysis، و یا مدلهای خطی عمومی وجود دارد.

الگوریتمهای بسیاری برای ساخت مدلها وجود دارد. میتوان یک شبکه عصبی را با backpropagation، یا توابع radial bias ساخت. برای درخت تصمیم، میتوان از میان CART، C5.0، Quest، و یا CHAID انتخاب کرد.

هنگام انتخاب یک **محصول** دادهکاوی، باید آگاه بود که معمولا پیادهسازیهای متفاوتی از یک الگوریتم دارند. این تفاوتهای پیادهسازی میتواند بر ویژگیهای عملیاتی مانند استفاده از حافظه و ذخیره داده و همچنین ویژگیهای کارایی مانند سرعت و دقت اثر گذارند.

در مدلهای پیشگویانه، مقادیر یا ردههایی که ما پیشبینی میکنیم متغیرهای پاسخ، وابسته، یا هدف نامیده میشوند. مقادیری که برای پیشبینی استفاده میشوند متغیرهای مستقل یا پیش - بینیکننده نامیده میشوند.

مدلهای پیشگویانه با استفاده از دادههایی که مقادیر متغیرهای پاسخ برای آنها از قبل دانسته شده است ساخته یا آموزش داده میشوند. این نحوه آموزش supervised learning نامیده میشود، زیرا که مقادیر محاسبه شده یا تخمینزده شده با نتایج معلومی مقایسه می شوند.(در مقابل، تکنیکهای توصیفی مانند clustering، unsupervised learning نامیده میشوند زیرا که هیچ نتیجه از پیش معلومی برای راهنمایی الگوریتم وجود ندارد.)

6- مراحل فرایند کشف دانش از پایگاه داده ها

فرایند کشف دانش از پایگاه داده ها شامل پنج مرحله است که عبارتند از :

۱. انبارش داده ها[۵۹]
۲. انتخاب داده ها
۳. تبدیل داده ها
٤. کاوش در داده ها
۵. تفسیر نتیجه

همانگونه که مشاهده می شود داده کاوی یکی از مراحل این فرایند است که به عنوان بخش چهارم آن نقش مهمی در کشف دانش از داده ها ایفا می کند .

٦-۱انبارش داده ها

وجود اطلاعات صحیح و منسجم یکی از ملزوماتی است که در داده کاوی به آن نیازمندیم . اشتباه و عدم وجود اطلاعات صحیح باعث نتیجه گیری غلط و در نتیجه اخذ تصمیمات ناصحیح در سازمانها می گردد و منتج به نتایج خطرناکی خواهد گردید که نمونه های آن کم نیستند .

[۵۹] Data Warehousing

اکثر سازمانها دچار یک خلا اطلاعاتی [60] هستند . در اینگونه سازمانها معمولا سیستم های اطلاعاتی در طول زمان و با معماری و مدیریت های گوناگون ساخته شده اند ، به طوری که سازمان اطلاعاتی یکپارچه و مشخصی مشاهده نمی گردد . علاوه بر این برای فرایند داده کاوی به اطلاعات خلاصه و مهم در زمینه تصمیم گیریهای حیاتی نیازمندیم .

هدف از فرایند انبارش داده ها فراهم کردن یک محیط یکپارچه جهت پردازش اطلاعات است . در این فرایند ، اطلاعات تحلیلی و موجز در دوره های مناسب زمانی سازماندهی و ذخیره می شود تا بتوان از آنها در فرایند های تصمیم گیری که از ملزومات آن داده کاوی است ، استفاده شود . به طور کلی تعریف زیر برای انبار داده ها ارائه می گردد :

انبار داده ها ، مجموعه ای است موضوعی [61] ، مجتمع [62] ، متغیر در زمان [63] و پایدار [64] از داده ها که به منظور پشتیبانی از فرایند مدیریت تصمیم گیری مورد استفاده قرار می گیرد .

انبارش داده ها خود موضوع مفصلی است که مقاله ها و رساله ها ی گوناگونی در مورد آن نگاشته شده اند . در این فصل به منظور آشنایی با این فرایند به آن اشاره ای شد .

---

[60] Information Gap
[61] Subject Oriented
[62] Integrated
[63] Time Variant
[64] NonVolatile

٦-٢انتخاب داده ها

انبار داده ها شامل انواع مختلف و گوناگونی از داده ها است که همه آنها در داده کاوی مورد نیاز نیستند . برای فرایند داده کاوی باید داده ها ی مورد نیاز انتخاب شوند . به عنوان مثال در یک پایگاه داده های مربوط به سیستم فروشگاهی ، اطلاعاتی در مورد خرید مشتریان ، خصوصیات آماری آنها ، تامین کنندگان ، خرید ، حسابداری و ... وجود دارند . برای تعیین نحوه چیدن قفسه ها تنها به داده ها یی در مورد خرید مشتریان و خصوصیات آماری آنها نیاز است . حتی در مواردی نیاز به کاوش در تمام محتویات پایگاه نیست بلکه ممکن است به منظور کاهش هزینه عملیات ، نمونه هایی از عناصر انتخاب و کاوش شوند .

٦-٣تبدیل داده ها

هنگامی که داده های مورد نیاز انتخاب شدند و داده ها ی مورد کاوش مشخص گردیدند ، معملا به تبدیلات خاصی روی داده ها نیاز است . نوع تبدیل به عملیات و تکنیک داده کاوی مورد استفاده بستگی دارد : تبدیلاتی ساده همچون تبدیل نوع داده ای به نوع دیگر تا تبدیلات پیچیده تر همچون تعریف صفات جدید با انجام عملیاتهای ریاضی و منطقی روی صفات موجود .

٦-٤کاوش در داده ها

داده های تبدیل شده با استفاده از تکنیکها و عملیاتهای داده کاوی مورد کاوش قرار می گیرند تا الگوهای مورد نظر کشف شوند .

٦-٥تفسیر نتیجه

اطلاعات استخراج شده با توجه به هدف کاربر تجزیه و تحلیل و بهترین نتایج معین می گردند . هدف از این مرحله تنها ارائه نتیجه (بصورت منطقی و یا نموداری) نیست ، بلکه پالایش اطلاعات ارایه شده به کاربر نیز از اهداف مهم این مرحله است .

7-عملیاتهای داده کاوی

در داده کاوی ، چهار عمل اصلی انجام می شود که عبارتند از :

١. مدلسازی پیشگویی کننده
٢. تقطیع پایگاه داده ها
٣. تحلیل پیوند
٤. تشخیص انحراف

از عملیاتهای اصلی مذکور ، یک یا بیش از یکی از آنها در پیاده سازی کاربرد های گوناگون داده کاوی استفاده می شوند . به عنوان مثال برای کاربرد های خرده فروشی معمولا از عملیات تقطیع و تحلیل پیوند استفاده می شود در حالی که برای تشخیص کلاهبرداری ، می توان از هر

یک از چهار عملیات مذکور استفاده نمود . علاوه برا ین می توان از دنباله ای از عملیاتها برای یک منظور خاص استفاده کرد . مثلا برای شناسایی مشتریان ، ابتدا پایگاه تقطیع می شود و سپس مدلسازی پیشگویی کننده در قطعات ایجاد شده اعمال می گردد .

تکنیکها ، روشها و الگوریتمهای داده کاوی ، راههای پیاده سازی عملیاتهای داده کاوی هستند . اگر چه هر عملیات نقاط ضعف و قوت خود را دارد ، ابزارهای گوناگون داده کاوی عملیاتها را بر اساس معیارهای خاصی ، انتخاب می کنند . این معیارها عبارتند از :

- تناسب با نوع داده های ورودی
- شفافیت خروجی داده کاوی
- مقاومت در مقابل اشتباه در مقادیر داده ها
- میزان صحت خروجی
- توانایی کار کردن با حجم بالای داده ها

در جدول زیرتکنیک های وابسته به هر یک از عملیاتهای چهار گانه مشخص شده اند:

نام عملیات	تکنیک های داده کاوی
مدلسازی پیشگویی کننده	رده بندی ، پیشگویی مقدار
تقطیع پایگاه داده ها	خوشه بندی آماری ، خوشه بندی
تحلیل پیوند	کشف بستگی ، کشف الگوهای متوالی ، کشف دنباله های زمانی مشابه
تشخیص انحراف	آمار ، تجسم مدل

جدول شماره (۱)

عملیاتها و تکنیکهای داده کاوی

۷-۱مدلسازی پیشگویی کننده

مدلسازی پیشگویی کننده ، شبیه تجربه یادگیری انسان در به کار بردن مشاهدات برای ایجاد یک مدل از خصوصیات مهم پدیده ها است . در این روش از تعمیم دنیای واقعی و تعمیم دنیای واقعی و قابلیت تطبیق داده های جدید با یک قالب کلی ، استفاده می شود .

در این مدل ، می توان با تحلیل یک پایگاه داده های موجود ، خصوصیات مجموعه های داده را تعیین کرد . این مدل با استفاده از روش یادگیری نظارت شده، شامل دو فاز آموزش و

آزمایش ایجاد شده است . در فاز آموزش با استفاده از نمونه های عظیمی از داده های سابقه ای ، مدلی ساخته می شود که که به آن مجموعه آموزشی گو یند . در فاز آزمایش این مدل روی داده هایی که در مجموعه آموزشی قرار ندارند ، اعمال می شود تا صحت و خصو صیات آن تایید گردد .

از کاربردهای عمده این مدل می توان به مدیریت مشتریان ، تصویب اعتبار ، بازاریابی مستقیم در خرده فروشی و ... اشاره کرد .

۷-۲تقطیع پایگاه داده ها

هدف از تقطیع پایگاه داده ها ، تقسیم آن به تعداد نامعینی از قطعات یا خوشه هایی [۶۵] از رکوردهای مشابه است ، یعنی رکوردهایی که خصوصیاتی مشابه دارند و می توان آنها را همگن فرض کرد . پیوستگی داخلی این قطعات بسیار زیاد است در حالی که همبستگی خارجی میان آنها کم می باشد .

در این مدل بر خلاف مدل قبل ، از یادگیری نظارت نشده برای تعیین زیرشاخه های ممکن از جمعیت داده ای استفاده می شود . دقت تقطیع پایگاه داده ها از روشهای دیگر کمتر است ، بنابراین در مقابل خصوصیات نامربوط و افزونگی ، حساسیت کمتری از خود نشان می دهد .

---

[۶۵] Clusters

از کاربردهای این روش می توان به شناسایی مشتریان ، بازاریابی مستقیم و ... اشاره کرد . در شکل ٤ مثالی از تقطیع پایگاه داده ها دیده می شود .

در این مثال ، پایگاه داده ها شامل ۲۰۰ مشاهده است که در آن ۱۰۰ اسکناس تقلبی و ۱۰۰ اسکناس واقعی هستند . داده ها دارای شش بعد می باشند که هر بعد مربوط به یک معیار از اندازه اسکناس ها است . با استفاده از تقطیع پایگاه داده ها می توان خوشه های متناظر با اسکناسهای معتبر و تقلبی را تشخیص داد . دو خوشه از اسکناسهای تقلبی وجود دارند و این بدان معنی است که حداقل دو گروه مبادرت به تولید و چاپ اسکناسهای تقلبی می کنند .

تقطیع پایگاه داده ها با آمارگیری مرتبط است که در ان از فاصله میان رکوردها و درصد قرار گرفتن داده های ورودی در خوشه ها ، جهت تجزیه و تحلیل استفاده می شود .

۷-۳تحلیل پیوند

در این روش پیوند هایی مرسوم به بستگی [۶۶] میان رکوردها و یا مجموعه ای از رکوردها بازشناسی می شوند . سه رده ویژه از تحلیل پیوند وجود دارند که عبارتند از :

۱. کشف بستگی[۶۷]
۲. کشف الگوهای متوالی[۶۸]
۳. کشف دنباله های زمانی مشابه[۶۹]

---

[۶۶] Association
[۶۷] Association Discovery
[۶۸] Sequential Pattern Discovery
[۶۹] Similar time Sequences

برای قوانین وابستگی دو پارامتر معرفی می گردند :

۱. درجه پشتیبانی[۷۰] : کسری از جمعیت است که در یک قاعده ، هم مقدم و هم تالی را دارند . در واقع درصدی از تراکنشها که شامل همه اقلام ظاهر شده در مقدم و تالی باشند . فرض کنیم که تنها در ۰۰۰۱/ . ٪ از تراکنشهای خرید ، شیر و پیچ گوشتی با هم باشند ، بنابراین درجه پشتیبانی برای قانون " پیچ گوشتی ← شیر " بسیار پایین است . این مساله نشان می دهد که مدرکی برای اثبات رابطه میان " شیر " و " پیچ گوشتی " وجود ندارد .

۲. درجه اطمینان[۷۱] : در یک جمعیت مورد بررسی ، کسری از موارد است که وقتی مقدم قاعده در آنها ظاهر شده است ، تالی نیز در آنها وجود دارد . به عنوان مثال در قانون " پنیر ← نان " اگر درجه اطمینان برابر ۸۰٪ تراکنشهای خرید ، اگر نان وجود داشته باشد ، پنیر نیز وجود دارد . باید توجه داشت که مقدار درجه اطمینان با تعویض مقدم و تالی در قاعده ، ممکن است به شدت تغییر کند .

دامنه اندازه پایگاه های داده امروزه به ترا بایت رسیده است این پایگاه داده به همراه اطلاعات فراوانی که به صورت ناشناخته در آن تعبیه گردیده می بایشد مساله این است که چگونه می توان از میان این جنگل عظیم اطلاعاتی به همراه درختهای پیچیده آن اطلاعاتی را استنتاج

---

[۷۰] Support
[۷۱] Confidence

نمود؟با استفاده از داده کاوی می توان این هزینه را کم نمود و در عوض بازدهی بیشتری بدست آورد.در حال حاضر شرکتهای بی شماری سعی دارند با استفاده از این روش به مشتریان خود پیشنهادات بهتری برای خرید ارائه دهند تا فروش آنها بالاتر رفته و در عوض ضرر و زیان موجود از این طریق کمینه گردد.

داده کاوی فرآیندی است که طی آن با استفاده از انواع مختلف ابزار تحلیل داده به دنبال کشف الگوها و ارتباطات میان داده های موجود که ممکن است منجر به استخراج اطلاعات جدیدی از پایگاه داده گردند می باشد.

اولین وساده ترین گام تحلیل داده در داده کاوی توضیح و شرح مشخص داده (از جمله معنی داده وانحراف استاندارد کلمه)می باشد که این کار می تواند به وسیله نمدارها و گراف هاییوهمچنین کلماتی که با این کلمه ارتباط معنایی نزدیکی دارند انجام گردد در نیجه جمع آوری جستجو و انتخاب داده درست در این بخش بسیار مهم و حیاتی می باشد.

اما این کار به تنهایی کار خاصی انجام نمی دهد شما باید یک مدل پیش بینی کننده بر اساس الگهایی که از نتایج دانش به دست آورده شده بسازید سپس آزمایش کنید که آیا ان مدل با نمونه اصلی سازگار است یک مدل خوب نباید با جهان واقع تفاوت چندانی داشته باشد.

آخرین گام نیز تشخیص صحت وسقم عملکرد مدل بصورت تجربی می باشدم.برای مثال از یک بانک مربوط به مشتریان وپاسخ هایی که به یک پیشنهاد خاص داده اند یک مدل می

سازید که بر اساس آن مشخص می شود که کدام حدس وانتظار بیشترین نزدیکی را با یک پیشنهاد مانند پیشنهاد قبلی دارد و اینکه آیا شما می توانید بر این حدس اعتماد کنید یا نه؟

8-قابلیتهای DataMining:

باید توجه داشته باشید که داده کاوی یک ابزار جادویی نیست که بتواند در پایگاه داده شما به دنبال الگوهای جالب بگردد و اگر به الگویی جدیدی برخورد کرد آن را به شما اعلام کند بلکه صرفا الگوها و روابط بین داده ها را به شما اعلام می کند بدون توجه به ارزش آنها. بنابراین الگوهایی که به این وسیله کشف می شوند باید با جهان واقع تطابق داشته باشند. به عنوان مثال داده کاوی می تواند با تعیین نرخ در آمدهایی که بطور مثال بین $50/000 و $65/000 است که برای خرید روزنامه خاصی در میان فروشندگان است تعیین کند که اکثر کالاهای مورد نیاز مردم چه رنجی از قیمت بوده وکدام ها هستند؟

به این ترتیب شما می توانید از هدف خرید مردم بدون اینکه فاکتورهایی برای خرید کالاهای خود در نظر بگیرید مطلع شوید؟

برای تضمین بدست آمدن نتایج با معنی لازم است که شما بتوانید داده های خود را تحلیل کنید کیفیت خروجی شما به اطلاعات خارج از پایگاه داده (به عنوان مثال داده ای باارزشی که

متفاوت از داده های نوعی در پایگاه داده شماست) ستونهای ظاهرا بی ارتباط یا با ارتباط نزدیک به بقیه پایگاه داده(مانند تاریخ تولید یا انقضای کالا) بستگی نزدیکی دارند .الگوریتم بر اساس حساسیتشان به داده ها روشهای متفاوتی دارند. اما غیر عاقلانه است که به محصول داده کاوی صرفا به برای تمام تصمیم گیری هایمان تکیه کنیم.

داده کاوی بطور اتوماتیک و بدون رهنمایی قادر به کشف راه حل ها نیست. شما ترجیحا به جای بیان یک هدف مبهم مانند "کمک به ارتقای پاسخ دهی به در خواست ها mail من " شما باید از داده کاوی برای یافتن خصیصه های افرادی که

(۱): به درخواست های شما پاسخ می دهند

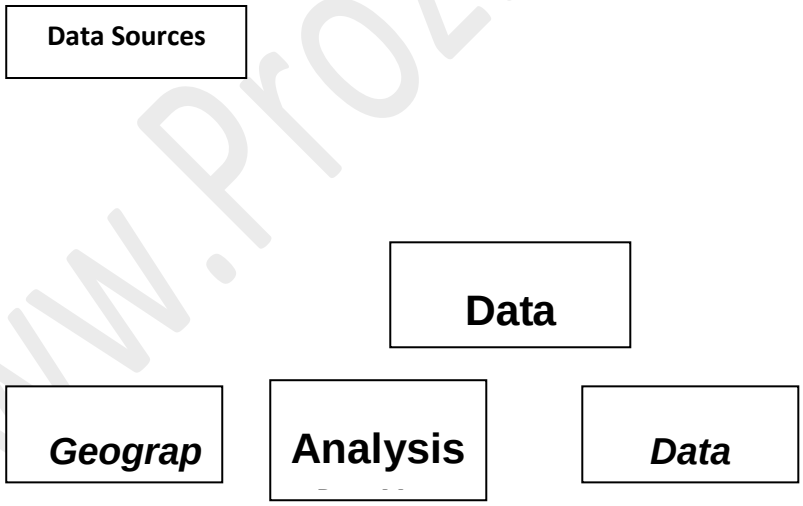
(۲): به درخواست های شما پاسخ داده و خرید زیادی می کنند

استفاده کنید. الگو هایی که داده کاوی برای یافتن به این دو هدف استفاده می کنند متفاوت است.

اگر چه یک ابزار خوب برای داده کاوی شما را از پیچیدگی های تکنیکهای آماری راحت می سازد اما به شما برای فهمیدن کار های ابزاری که انتخاب کرده اید و همچنین الگوریتمهایی که بر پایه آن کار می کند نیازمند است. انتخابی که شما برای ابزار مورد نیاز انجام می دهید و بهینه سازی هایی را که شما انجام می دهید در دقت و سرعت کار بسیار تاثیر دارد.

۸-۱داده کاوی و انبار داده ها :

اغلب داده ای که مورد کاوش قرار می گیرد ابتدا از یک انبار داده آماده شده به داخل یک پایگاه داده کاوی سرازیر می شود. این کار مزایای زیادی دارد. پایگاه داده کاوی می تواند به جای یک انبار فیزیکی داده یک انبار منطقی از داده ها باشد. به شرط آنکه انبار داده DBMS بتواند دامنه های منابع اضافی از داده کاوی را نیز پوشش دهد. روند شرح داده شده در شکل زیر آمده است:



شکل شماره(۹)

۸-۲داده کاوی , آمار و یادگیری ماشین

داده کاوی فوایدی از پیشرفتهای رشته هوش مصنوعی را در خود جای داده است که هم شامل قواعدی برای مسائل تشخیص الگو و طبقه بندی می باشد وهم ارتباطاتی که از طریق کاربرد شبکه های عصبی و درختهای تصمیم گیری برای فهم مسائل صورت می گیرد می باشد.

داده کاوی در این زمینه دارای الگوریتم های نسبتا جدیدی مانند شبکه عصبی و درخت تصمیم ورهیافت های جدیدی برای الگوریتم های قدیمیتر مانند الگوریتم های تفکیک کننده دارد.

نکته مهم آنکه داده کاوی کاربرد این تکنیکها را برای مسائل تجاری مشابه بالا به طریقی که این تکنیکها را برای کاربر خبره دانش و آمارگیر متخصص قابل دسترس سازد استفاده می شود.

۸-۳کاربردهای داده کاوی

داده کاوی به سرعت در حال محبوبیت است به خاطر کمک های اساسی آن.

سازمانهای زیادی در حال استفاده از داده کاوی برای کمک به مدیریت تمام فازهای ارتباط با مشتری شامل به دست آوردن مشتریان جدید, افزایش سود از طریق مشتریان موجود و حفظ کردن مشتریان خوب هستند.با تعیین مشخصات یک مشتری خوب یک شرکت می تواند با

همان مشخصات اهداف آینده خویش را پیش بینی کند. با پرونده سازی برای مشتری که یک محصول خاص را خردی می نماید این شرکت می تواند توجه خود را به مشتریان مشابهی که از این محصول خرید نکرده اند معطوف دارد با پرونده سازی برای مشتریانی که این سازمان را ترک کرده اند یک شرکت می تواند مشتریانی را که خطر رفتن آنها نیز وجود دارد را نگه دارد چرا که نگهداری یک مشتری موجود بسیار کم هزینه تر از بدست آوردن یک مشتری جدید هزینه می برد. داده کاوی ارزشهایی را از طریق بررسی یک طیف وسیعی از کارخانه ها پیشنهاد می کند.شرکتهای ارتباطات از راه دور و کارت های اعتباری دو شاخه بزرگ در استفاده از داده کاوی برای تشخیص استفاده کلاه بردارانه از خدمات آنها می باشند. شرکتهای بیمه و درآمد هم علاقمند به استفاده از این تکنولوژی برای کاهش کلاه برداری می باشند. کاربردهای دارویی نواحی مفید دیگری هستند که داده کاوی در آنها دست دارد داده کاوی می تواند برای تشخیص تاثیر اعمال جراحی, آزمایش های دارویی ودرمان استفاده گردد.

شرکتهایی که در خرید و فروشهای مالی فعالیت می کنند از داده کاوی برای تعیین شاخصه های بازار و صنعت برای تشخیص کارایی درآمد استفاده می کنند. خرده فروشها از داده کاوی برای تصمیم در مورد اینکه کدام محصول در فروشگاه ها در آمد زاست به منظور دسترسی به ارتقای کیفیت کار خود استفاده بیشتری می نمایند. شرکتهای دارویی در حال کاوش پایگاههای داده بزرگی از ترکیبات شیمیایی و مواد ژنتیکی برای کشف مواد که می توانند گزینه خوبی برای ساخت به عنوان دارو باشند.

۸-٤داده کاوی موفق:

دو نکته برای موفق بودن یک داده کاوی وجود دارد. اول اینکه یک فرموله سازی دقیق از مساله ای است که شما باید حل کنید. دومین نکته استفاده از داده صحیح است. پس از انتخاب داده ای که در دسترس شماست یا شاید خرید داده خارجی شما ممکن است نیازمند شوید آن را به روشهایی انتقال داده یا دسته بندی کنید.

۸-۵تحلیل ارتباطات:

تحلیل ارتباط یک رهیافت توصیفی برای اکتشاف داده است که می تواند به مشخص سازی ارتباطات میان مقادیر در پایگاه داده کمک نماید.دو رهیافت عام برای رسیدن به تحلیل ارتباطی اکتشاف ارتباطی و اکتشاف توالی می باشد.اکتشاف ارتباطات قوانینی را در مورد مواردی را که باید با هم در یک رویداد ظاهرشوند مانند تراکنش خرید را می ابد.تحلیل سبد عرضه یک نمونه شناخته شده از کشف ارتباط می باشد.کشف توالی بسیبار شبیه کشف ارتباط است با توجه به این نکته که در اینجا توالی یک ارتباط است که در طول یک بازه زمانی صورت می گیرد.

ارتباطات به صورت A=>B نوشته می شود که به A مقدم یا طرف سمت چپ و به B تالی یا طرف سمت راست می گویند.برای مثال در قانون ارتباطی "اگر مردم یک چکش بخرند آنگاه می توانند میخ بخرند" جمله مقدم "خرید چکش" و جمله تالی "خرید میخ" می باشد.

براحتی میتوان نسبت تراکنشهایی را که شامل مورد یا لیستی ازموارد خاص می باشد با شمردن آنها تعیین کرد (که در اطنجا موارد میخ ها و چکش هامی باشد) را تعیین کرد.تعداد موجود از یک نوع ارتباط خاص که در یک پایگاه داده به نظر می رسد را موجودی یا شیوع آن مورد می گویند.اگر برای مثال گفته شود که از هر ۱۰۰۰ تراکنش ۱۵ تای آن شامل "میخ و چکش" می باشد موجودی این ارتباط ۱.۵٪خواهد بود.یک موجودی کم(مثلا یک در میلیون) می تواند بیانگر این باشد که ان ارتباط خاص در پایگاه داده چندان مهم نیست.

برای کشف قوانین معنا دار ما باید به فراوانی متناسب دفعات اتفاق موارد و ترکیباتشان نیز بنگریم.باداشتن تعداد دفعات اتفاق مورد A مورد B چند بار اتفاق می افتد؟به عبارت دیگر سوال این است که ببینیم "هنگامی که مردم یک چکش می خرند چه تعداد از این افراد میخ هم می خرند؟ عبارت دیگر برای این پیش بینی شرطی اطمینان نام دارد.

فرض کنید پایگاه داده فرضی مان رابه صورت زیر و با جزئیات بیشتر برای بیان این مفاهیم در نظر بگیریم:

تمام تراکنشهای سخت افزار :۱۰۰۰

تعداد تراکنشهایی که شامل "چکش " می باشد:۵۰

تعداد تراکنشهایی که شامل "میخ" می باشد:۸۰

تعداد تراکنشهایی که شامل "تخته " می باشد:۲۰

تعداد تراکنشهایی که شامل " میخ و چکش"می باشد:۱۵

تعداد تراکنشهایی که شامل " میخ و تخته " می باشد:۱۰

تعداد تراکنشهایی که شامل " چکش و تخته" می باشد: ۱۰

تعداد تراکنشهایی که شامل " چکش و تخته و میخ " می باشد:۵

حال قادر به محاسبه ایم:

موجودی "میخ و چکش"=۱.۵٪

موجودی " میخ و چکش وتخته"=۰.۵٪

درصد اطمینان "چکش=>میخ" = ۳۰٪

درصد اطمینان " میخ=> چکش" = ۱۹٪

درصد اطمینان " چکش و میخ=>تخته" = ۳۳٪

درصد اطمینان " تخته=> چکش و میخ " =۲۵٪

بنابراین ما می بینیم که احتمال اینکه یک خرنده چکش میخ هم بخرد(۳۰٪) بیشتر از احتمال آن است که فردی که میخ می خرد چکش هم بخرد(۱۹٪).ارتباط چکش و میخ به اندازه ای بزرگ است که یک قانون با معنی باشد.

Lift(نسبتا پیشرفت) یکی از معیارهای اندازه گیری قدرت یک ارتباط است.هر چه lift بزرگتر باشد تاثیر اتفاقات A بر احتمال اینکه B اتفاق بیفتد بیشتر است.lift بصورت نسبت

(اطمینان A=>B) تقسیم بر فراوانی B محاسبه می شود:

برای مثال ما:

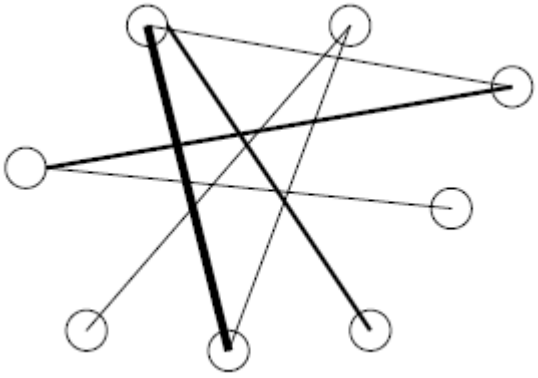
Lift "چکش=>میخ" :۳.۷۵

Lift " چکش و میخ =>تخته ":۱۶.۵

الگوریتمهای ارتباط این قوانین را با معادل مرتب سازی داده هنگام شمارش دفعاتی که می توانند درصد اطمینان و موجودی را محاسبه کنند می یابد. اثراتی که هر یک از این قوانین می توانند داشته باشند یکی از معیارهای تفاوت این الگوریتم هاست. این معیار مهم است زیرا که نتایج ترکیبی بسیار زیادی از تعداد بی شماری از قوانین بدست می آید حتی برای سبد های خرید. برخی از الگوریتمها یک پایگاه داده از قوانین, فاکتورهای ایمن, و فراهم آوردن امکان جستجو(برای مثال تمام ارتباطاتی که در آن کلمه بستنی در قوانین به عنوان نتیجه آمده و فاکتوری برابر ۸۰٪را دارند نشان بده)را ایجاد می نمایند.

اغلب تصمیم گیری در مورد کار با قوانینی که شما کشف کرده اید دشوار است.به عنوان مثال در یک نقشه خرید برای مشتریان در یک فروشگاه قراردادن تمام اجناس مرتبط منطقی به صورت فیزیکی در کنار یکدیگر ممکن است ارزش کامل سبد خرید را کاهش دهد – مشتریان ممکن است در مجموع ارزش کمتری خرید کنند چون آنها بر خلاف نقشه خرید مورد نظر شما در حین راه رفتن در مغازه اجناس مورد دلخواه خود را خرید می کنند. در چنین حالتی تقریب و تحلیل ارتباطات معمولا برای بدست آوردن هر گونه سودی از قوانین مرتبط با هم مورد نیاز خواهد بود.

روشهای گرافیکی می توانند در نمایش ساختار ارتباطات نقش داشته باشند. در شکل زیر هر یک از دوایر یک مقدار یا یک رویداد را نمایش می دهد. خطوط ارتباطی میان این دایره ها یک ارتباط را نشان می دهند. خطوط کلفت تر ارتباطات قوی تر و فراوان تری را نمایش می دهند.



شکل شماره (۱۰)-سلسله مراتبی از انتخاب ها

هدف داده کاوی تولید دانش جدیدی است که کاربر بتواند بر اساس آن کار خود را جلو برد. این کار بوسیله ساختن مدلی از جهان واقعی بر پایه داده هایی که از منابع گوناگون بدست می آید صورت گیرد که این منابع می تواند شامل تراکنشهای هماهنگ, تاریخ مربوط به هر مشتری, اطلاعات نمایش گرافیکی, داده کنترل فرآیند و پایگاه داده های مرتبط خارجی مانند اطلاعات اعتبار اداری و ... باشد. نتیجه مدل سازی یک سری توضیحات در مورد الگوها و ارتباطات داده ای که می تواند به صورت مطمئنی جهت پیش بینی آینده مورد استفاده قرار گیرد.

برای جلوگیری از سرگردانی در مراحل مختلف داده کاوی ایجاد تصویری از سلسله مراتبی از انتخابات و تصمیم ها که نیاز مند آن هستید در ذهن قبل از شروع کار به شما کمک خواهد کرد:

- هدف کار

- نوع پیش بینی

- نوع مدل انتخابی

- الگوریتم

- محصول

اولین گام مشخص نمودن هدف کار می باشد :

هدف نهایی از جستجوی این داده چیست؟ برای مثال جهت یافتن الگوهای مفیدی در داده خود برای این که به شما کمک کند مشتریان خود را حفظ کنید شما باید یک مدل برای پیش بینی سودبخشی به مشتری و مدل دیگری برای شناسایی مشتریانی که آنجا را ترک کرده اند طراحی کنید. دانش شما از احتیاجات و اهداف سازمانتان شما را به سمت فرموله کردن اهداف مدلهایتان راهنمایی خواهد کرد.

گام بعدی تصمیم در مورد انتخاب نوعی پیش بینی که از همه مناسب تر است می باشد:

(۱) طبقه بندی: تعیین این که این مورد خاص در کدام کلاس یا دسته قرار می گیرد.

(۲) حدس زدن اینکه یک متغیر چه مقدار عددی خواهد داشت(اگر متغیری باشد که با زمان تغییر کند این کار حدس سریهای زمانی نامیده می شود).در مثال بالا شما می توانید از این

حدس برای پیش بینی مقدار سوددهی و طبقه بندی برای پیش بینی اینکه کدام مشتریان ممکن است خرید شما را ترک کنند استفاده کنید.

حالا نوبت به نوع مدل می رسد:

که عبارت است از یک شبکه عصبی برای انجام حدس فوق الذکر و یک درخت تصمیم برای طبقه بندی. مدلهای آماری سنتی نیز برای انتخاب از مدلهای معمولی خطی , تحلیل تفکیکی و حدس منطقی وجود دارد. مهمترین نوع این مدلها برای داده کاوی در بخش بعد (الگوریتمها و مدلهای داده کاوی)توضیح داده می شود.

الگوریتمهای زیادی برای ساخت مدلهایتان در دسترس هستند. شما می توانید با استفاده از توابع شعاعی یا انتشاری شبکه عصبی بسازید. برای درخت تصمیم شما می توانید از میان طرق cart, c5.0, CHAID,Quest یکی را انتخاب کنید. برخی از این الگوریتم ها در مدلها و الگوریتمهای داده کاوی توضیح داده شده است.

هنگام انتخاب یک محصول داده کاوی باید توجه داشت که این محصولات پیاده سازیهای مختلفی از یک الگوریتم خاص دارند حتی اگر این الگوریتم برای همه آنها نام یکسانی داشته باشد. این تفاوتها در پیاده سازی می تواند بر روی مشخصه های قابل استفاده مانند استفاده از حافظه و ذخیره داده و همچنین بر روی مشخصه های کارایی مانند سرعت و دقت تاثیر بگذارند.

بسیاری از اهداف تجاری به بهترین شکل به وسیله ساخت انواع مختلفی از مدلها با استفاده از الگوریتمهای مختلف به دست می آیند. شما ممکن است تا زمانی که راه های مختلفی را امتحان نکنید قادر نباشید تعیین کنید کدام نوع مدل بهترین است.

9- طبقه بندی

مسائل طبقه بندی به شناسایی خصوصیاتی منجر می شوند که مشخص می نمایند هر مورد به کدام گروه تعلق دارد.این الگو هم می تواند برای فهم داده موجود و هم برای پیش بینی اینکه هر نمونه جدید چگونه کار می کند استفاده شود. برای مثال شما ممکن است بخواهید پیش بینی کنید که آیا اشخاص برای پاسخگویی به درخواست یک میل مستقیم که ممکن است به یک دستگاه تلفن با مسافت زیاد آسیب برساند می توانند گروه بندی شوند یا برای یک عمل جراحی باید گروه بندی شوند.

داده کاوی مدلهای طبقه بندی را بوس یله امتحان کردن داده طبقه بندی شده(موارد) و نها یتا یافتن یک الگوی پیش گو ایجاد می کند. این موارد موجود می تواند از یک پایگاه داده تاریخی ناشی شود مانند اطلاعات افرادی که تحت معالجه دارویی خاصی هستند و یا به سمت یک خدمت با مسافت دور جذب شده اند.یا اینکه از تجربه هایی که طی آن یک نمونه از تمام پایگاه داده در جهان واقعی تست شده باشد و نتایج آن برای ایجاد یک گروه بند استفاده شده

باشند منتج شود. برای مثال یک نمونه از لیستی از پیامها به عنوان پیشنهاد فرستاده خواهد شد و نتایج پیام رسانی برای ساخت یک مدل طبقه بندی جهت بکار گرفته شدن در تمام پایگاه داده استفاده خواهد شد.

۹-۱حدس بازگشتی

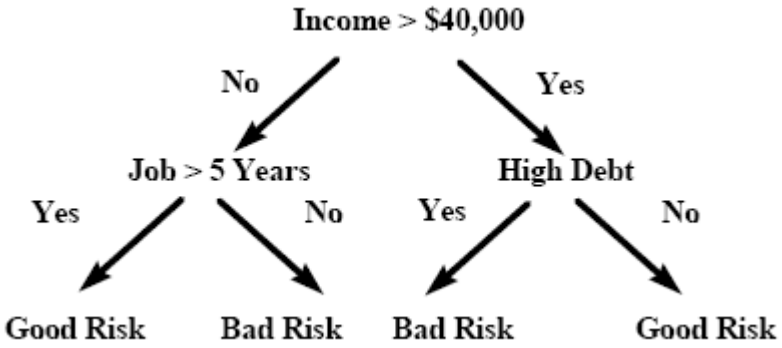
حدس بازگشتی از داده های موجود برای پیش بینی این که مقادیر داده های دیگر چه خواهد بود استفاده می کند. در ساده ترین حالت حدس مذکور از تکنیکهای آماری مانند حدس خطی استفاده می کند. متاسفانه بسیاری از مسائل جهان واقع تصویری خطی از مقادیر قبلی نیستند. برای نمونه مقادیر فروش, ارزش فروش, ارزش سهام و نرخ ورشکستگی محصول برای پیش بینی سخت می باشد زیرا آنها ممکن است بر فعل و انفعالات پیچیده حاصل از چندین متغیر پیش بینی کننده متکی باشند. بنابراین تکنیکهای پیچیده تری ممکن است برای پیش بینی متغیرهای آینده ضروری باشند. انواع مدل یکسان اغلب می توانند هم برای حدس بازگشتی وهم برای طبقه بندی استفاده شوند. برای مثال الگوریتم درخت تصمیم CART (درختهای حدس وطبقه بندی)هم برای ساخت درختهای حدس و هم برای ساخت درختهای طبقه بندی به کار می رود. شبکه های عصبی هم می توانند هر دو نوع مدل نام برده شده را ایجاد نمایند.

۹-۲سری های زمانی

سری های زمانی پیش بینی کننده مقادیری را که هنوز مقدارشان مشخص نیست بر اساس یک سری از پیشگوهای متغیر با زمان پیش بینی می کنند. مانند حدس بازگشتی این روش هم از نتایج معلوم قبلی برای اعمال پیشگویی های بعدی اش بهره می برد. مدلها باید خواص منحصر بفرد زمان علی الخصوص سلسله مراتب دوره های زمانی مانند دوره های فصلی تاثیرات تقویمی مانند تعطیلات محاسبات تاریخی و ملاحظات خاص مانند تطبیق گذشته با حال را ذخیره نمایند.

۹-۳درخت های انتخاب

درخت های انتخاب راهی برای نمایش یک سری از قوانین که به یک کلاس یا مقدار منجر می شود می باشند. برای مثال شما ممکن است بخواهید درخواستهای وام را برحسب ریسک اعتبار خوب یا بد طبقه بندی کنید. شکل بعد یک مدل ساده از یک درخت انتخاب به همراه توضیح در مورد تمام بسته های پایه آن یعنی گره انتخاب, شاخه ها و برگهای آن که این مساله را حل می کند نشان می دهد.



شکل شماره (۱۱)

اولین بسته گره بالایی تصمیم یا ریشه می باشد که یک بررسی جهت برقراری شرط خاصی می نماید. گره ریشه در این مثال "Income>$40,000" می باشد. نتایج این بررسی منجر می شود که درخت به دوشاخه تقسیم گرددکه هر یک نشان دهنده جوابهای ممکن است.در این مورد بررسی شرط مذکور می تواند دارای جواب خیر یا بله باشد در نتیجه دو شاخه داریم.

براساس نوع الگوریتم هر گره می تواند دو یا تعداد بیشتری شاخه داشته باشد. برای مثال CART درختهایی با تنها دوشاخه در هر گره تولید می کند.چنین درختی یک درخت دودویی می باشد.

مدلهای مختلف درخت تصمیم بطور عمومی در داده کاوی برای کاوش داده و برای استنتاج درخت و قوانین آن که برای پیش بینی مورد استفاده قرار می گیرد استفاده می شوند. یک تعداد از الگوریتمهای مختلف می توانند برای ساخت درختهای تصمیم شامل CHAID, Quest,CART و C5.0 بکار روند.

اندازه درخت می تواند از طریق قوانین متوقف شونده که رشد درخت را محدود می کنند کنترل شود.

●تئوری بیز .

تئوری بیز یکی از روش های آماری برای طبقه بندی به شمار می آید . در این روش کلاس های مختلف ، هر کدام به شکل یک فرضیه دارای احتمال در نظر گرفته می شوند.

هر رکورد آموزشی جدید ، احتمال درست بودن فرضیه های پیشین را افزایش و یا کاهش می دهد و در نهایت ، فرضیاتی که دارای بالاترین احتمال شوند ، به عنوان یک کلاس در نظر گرفته شده و برچسبی بر آنها زده می شود . این تکنیک با ترکیب تئوری بیز و رابطه سببی بین داده ها ، به طبقه بندی می پردازد.

●رگرسیون .

رگرسیون نیز یکی از روش های آماری برای طبقه بندی به شمار می آید . هدف از رگرسیون ، پیش بینی مقدار یک متغیر پیوسته بر اساس مقادیر متغیر های دیگر است . رگرسیون به دو دسته خطی و غیر خطی تقسیم می شود.

برای مثال می توان پیش بینی میزان فروش یک محصول جدید را بر اساس میزان تبلیغات صورت گرفته بر روی آن ، از روش رگرسیون انجام داد.

به جز روش های ذکر شده ، روش های دیگری نیز برای طبقه بندی موجود است که می توان به K_ Nearest Neighborhood ، Case_ Based Reasoningو الگوریتم ژنتیک

اشاره کرد.

●گروه بندی داده ها .

به فرایند دسته بندی اشیای فیزیکی یا انتزاعی به کلاس هایی از اشیاء متشابه ، گروه بندی (طبقه بندی بدون ناظر) می گویند.

گروه بندی جزو روش های تشریح کننده به حساب می آید . این روش با تفکر تقسیم و حل ، به دسته بندی داده های موجود در یک سیستم بزرگ پرداخته و آنها را به مولفه های کوچک تر تقسیم می کند.

یک گروه بندی را زمانی مناسب گویند که اشیای داده ای درون هر گروه بسیار به یکدیگر شبیه بوده و با اشیای گروه های دیگر تفاوت بسیار داشته باشند . معیار شباهت و تفاوت بین اشیای داده ای توسط یک تابع فاصله مشخص می شود . بسته به نوع داده ، توابع فاصله متفاوتی موجود است که از آن جمله می توان به تابع فاصله Minkowski ، تابع فاصله اقلیدسی ضریب Jaccark اشاره کرد . در ادامه به روش های مختلف برای گروه بندی داده ها پرداخته می شود.

●بخش بندی

در این تکنیک یک بخش بندی از پایگاه داده D با n شیء به k گروه انجام می گیرد . این کار توسط معیاری که برای گروه بندی در نظر گرفته شده ، انجام می شود . روش های مختلفی از جمله K_ means ، K_ medoids ، PAM ، CLARAو CLARANSبرای دسته بندی

موجود است.

●سلسله مراتبی .

این تکنیک از فاصله ماتریسی به عنوان شرط گروه گروه بندی استفاده می کند . این روش به جای مشخص کردن تعداد گروه ها در ابتدای کار ، احتیاج به یک شرط خاتمه برای پایان دادن به عملیات گروه بندی دارد.

روش های مختلفی نیز برای این تکنیک مطرح شده است که از آن جمله می توان به روش AGNES، DLANA، BLRCH، CUREو CHAMELEON اشاره کرد.

●گروه بندی بر اساس تراکم .

در این تکنیک ، گروه بندی بر اساس میزان تراکم نقاط به هم پیوسته مشخص می شود . دو پارامتر Eps و MinPts در این تکنیک در نظر گرفته می شود که Eps مشخص کننده ماکزیمم شعاع همسایگی و MinPts مشخص کننده مینمم تعداد نقاط درون همسایگی Eps است.

روش های مختلفی نظیر DBSCAN ، OPTLCS، DENCLUEو CLIQUE نیز در این تکنیک مورد مطالعه قرار گرفته است.● کاوش قوانین پیوندی .

یکی دیگر از تکنیک های داده کاوی ، کاوش قوانین پیوندی است.

از جمله کارهایی که کاوش قوانین پیوندی برای ما انجام می دهد می توان به پیدا کردن وابستگی ها و همبستگی های موجود در بین داده ها ، یافتن الگوهایی که غالبا در بین داده ها وجود دارند و همچنین پیدا کردن یک سری ساختار سببی در بین آیتم ها و

اشیای موجود در پایگاه داده های تعاملی و رابطه ای اشاره کرد . قبل ازمعرفی الگوریتم های مربوط به کاوش قوانین پیوندی ، نیاز به معرفی یک سری مفاهیم پایه است .

۱ .مجموعه آیتم های موجود در یک پایگاه اطلاعاتی با ...} ،۲X ،۱X ItemSet = { نمایش داده می شوند.

۲ .برای هر قانون به شکل X -> Y است ، دو مقدار Support و Confidence مشخص می شود.

۳ Support .احتمال وجود همزمان X و Y به صورت توام در تراکنش است.

٤ Confidence .احتمال شرطی است برای آنکه تراکنش دارای X ، دارای Y نیز باشد.

بنابراین قانون X -> Y با=S (۵۰٪ و=C ٦٦.۷٪ (بدین معنی است که X و Y به صورت توام در ۵۰ درصداز کل تراکنش ها وجود دارند و در ۷/٦٦ درصد از تراکنش ها ، هر جا X در تراکنش حضور داشته ، Yنیز حضور داشته است . کاوش قوانین پیوندی در پایگاه داده ها شامل دو مرحله زیر است:

۱ .کشف بزرگ ترین مجموعه آیتم ها (مجموعه آیتم هایی که دارای مقدار Support بالاتر از یک مقدار خاص باشند. (

۲ .استفاده از مجموعه آیتم های کشف شده در مرحله قبل و ساخت قوانین پیوندی.

به طور کلی بیشتر کارها برای بهینه کردن اجرای مرحله اول یعنی کشف بزرگ ترین مجموعه آیتم انجام می گیرد ، زیرا با داشتن بزرگ ترین مجموعه آیتم ، پیدا کردن قوانین به صورت

مستقیم ، ممکن می شود . در ادامه به معرفی الگوریتم های مختلف ارائه شده برای کشف بزرگ ترین مجموعه آیتم می پردازیم.

- Apriori

هدف در این الگوریتم ، پیدا کردن بزرگ ترین مجموعه آیتم است که حداقل Supportو Confidenceرا رعایت کند . دو فرض زیر در این الگوریتم در نظر گرفته می شود:

۱ .هر زیر مجموعه از یک مجموعه آیتم تکرار شونده ، تکرار شونده است) . یعنی اگر فرضاً مجموعه { a ،b ،c } تکرار شونده باشد ، آنگاه مجموعه { a ،b } ، { a } نیز تکرار شونده است(.

۲ .هر فوق مجموعه از یک مجموعه آیتم تکرار نشونده ، است . (یعنی اگر فرضاً مجموعه { a ،b } تکرار شونده نباشد ، آنگاه مجموعه { a ،b ،c } ، { a ،b } نیز تکرار شونده نیست(.

الگوریتم Apriori به این صورت است که در هر بار ، یک سری مجموعه آیتم بزرگ با طول K+۱ را از روی مجموعه آیتم های کاندید با طول K می سازد و این کار را تا رسیدن به یک مجموعه آیتم با بیشترین طول انجام می دهد . مجموعه آیتم های کاندید در هر دفعه با ضرب مجموعه کاندید در خودش به دست می آید از مشکلات این روش می توان به حجم بسیار بالای تراکنش های موجود در پایگاه داده ، طولانی بودن زمان جست و جوی آنها در هر بار و تعداد زیاد کاندیدها در هر مرحله اشاره کرد . ایده های مطرح شده برای بهینه سازی الگوریتم Aprioriعبارتند از:

۱ .کاهش تعداد دفعات جست و جو در پایگاه داده تراکنشی.

۲ .کاهش تعداد کاندیدها.

۳ .ساده کردن شمارش برای. Support

- DHP

این روش مشابه با الگوریتم Apriori بوده و تنها تفاوت آن در ایجاد مجموعه کاندید در هر مرحله است . در روش Apriori مجموعه کاندید ، با ضرب مجموعه آیتم بزرگ به دست آمده تا این مرحله در خودش ، به وجود آمد . اما در روش DHPبرای ساخت مجموعه کاندید در هر مرحله ، از یک جدول hash استفاده می شود و تنها یک سری از مجموعه آیتم های موجود در حاصلضرب به عنوان مجموعه کاندید پذیرفته می شود.

)مجموعه آیتم هایی که دارای Support بالاتری هستند(.

الگوریتم DHP با استفاده از کاهش تعداد کاندیدها ، الگوریتم Apriori را بهبود می بخشد . روش های دیگری نیز برای بهبود الگوریتم Apriori مطرح شده اند ؛ روش DlC تعداد جست و جو ها را کاهش می دهد ، روش Partition پایگاه داده را به دو قسمت تقسیم کرده و درهر کدام به دنبال بزرگ ترین مجموعه آیتم محلی می گردد و در نهایت بر اساس آنها بزرگ ترین مجموعه آیتم کلی را پیدا می کند .

روش Sampling هر دفعه یک مجموعه آیتم را به عنوان نمونه انتخاب کرده و بزرگ ترین مجموعه آیتم را پیدا می کند و سپس مرزهای مجموعه را بررسی کرده و پایگاه را برای مجموعه آیتم های بزرگ تر ، جست و جو می کند.

- خلاصه سازی و کلی نگری داده ها در سطوح مختلف .

یکی دیگر از روش های داده کاوی ، خلاصه سازی و کلی نگری داده ها در سطوح مختلف

است . به طور کلی اغلب داده های موجود در پایگاه داده دارای جزئیات فراوانی هستند . برای مثال در پایگاه داده فروش ، رابطه کالا شامل فیلدهای اطلاعاتی نظیر شماره کالا ، نام کالا ، سال ساخت ، قیمت و غیره است . جزئیات زیاد سبب پایین آمدن سطح ادراکی می شود و برای تصمیم گیری بر اساس اطلاعات قبلی ، نیاز به سطوح ادراکی بالاتری است.

داده کاوی با انجام خلاصه سازی و کلی نگری در داده ها در سطوح مختلف به کمک شتافته و سطوح ادراکی بالاتری را ایجاد می کند . در ادامه رهیافت های مختلف ارائه شده برای این کار ، مورد بررسی قرار گرفته است.

- رهیافت هرم داده ها .

ایده اصلی در این رهیافت ، جمع آوری نتایج محاسبات پرهزینه ای است که اغلب مورد درخواست بوده و نگهداری از آنها در یک ساختار چند بعدی به نام هرم داده ها صورت می گیرد . این محاسبات معمولا شامل توابعی نظیر مجموع ، میانگین ، ماکزیمم و غیره بر روی مجموعه صفات خاصه است . هرم داده ها معمولا بر روی پایگاه داده تحلیلی ایجاد می گردد . برای کلی نگری و ویژه نگری داده ها به ترتیب می توان Roll_Up و Drill_Down را بر روی هرم داده ها انجام داد . در بیشتر مواقع هرم داده ها دارای سه بعد بوده که دو بعد آن معمولاً زمان و مکن و بعد دیگر یک آیتم اطلاعاتی است.

برای مثال میزان فروش آیتم X در مهر ماه سال قبل در منطقه شمال تهران ، می توان نمایش دهنده یک درخواست که غالباً مورداستفاده قرار می گیرد ، باشد . از مشکلات این روش می

توان به محدود بودن محاسبات فقط بر روی انواع داده ای ساده و همچنین استفاده از توابع محاسبات ساده اشاره کرد.

- رهیافت استنتاج بر اساس صفت خاصه .

در رهیافت هرم داده ها ، محاسبات بر روی پایگاه داده تحلیلی به صورت offline انجام می گیرد . برای حل این مشکل ، رهیافت استنتاج بر اساس صفت خاصه ابتدا یک درخواست داده کاوی را به صورت DMQL مشخص می کند . سپس یک پرس و جو از روی DMQL داده شده می سازد و از پایگاه داده به صورت ONLINE درخواست می کند . سپس بر روی داده های به دست آمده ، تکنیک های کلی نگری داده ها نظیر حذف صفت خاصه ، بالا رفتن از درخت ادراک و غیره را اعمال می کند.

۹-٤استنتاج قانون

استنتاج قانون روشی برای بدست آوردن یک سری از قوانین برای طبقه بندی موارد می باشد. اگرچه درختهای تصمیم می توانند یک سری قوانین تولید کنند روشهای استنتاج قانون یک مجموعه از قوانین وابسته که ضرورتا درختی تشکیل نمی دهند را تولید می نماید. چون استنتاج کننده قوانین لزوما انشعابی در هر سطح قرار نمی دهد و می تواند گام بعدی را تشخیص دهد گاهی اوقات می تواند الگوهای مختلف و بهتری را برای طبقه بندی بیابد. برخلاف درختان قوانین تولیدی ممکن است تمام حالتهای ممکن را پوشش ندهند.

۹-۵الگوریتمهای ژنتیک

الگوریتمهای ژنتیک برای یافت الگوها استفاده نمی شود بلکه بیشتر به منظور راهنمایی در مورد فرآیند یادگیری الگوریتمهای داده کاوی مانند شبکه های عصبی مورد استفاده قرار می گیرد. الگوریتمهای ژنتیک به عنوان یک متد جهت انجام یک جستجوی هدایت شده برای مدلهای خوب در فضای حل مساله عمل می کند.

این الگوریتمها, الگوریتمهای ژنتیک نامیده می شوند چون بطور بی قاعده ای الگوی تکامل زیستی که در آن اعضای یک نسل بر سر انتقال خصوصیات خود به نسل بعد رقابت می کنند تا نهایتا بهترین مدل یافت شود را دنبال می کنند. اطلاعاتی که باید انتقال داده شود در قالب کروموزمها که شامل پارامترهایی برای ساختن مدل می باشد قرار می گیرد.

10فرآیند داده کاوی

مدلهای فرآیند

با توجه به اینکه یک فرآیند سیستماتیک برای داده کاوی موفق ضروری است بسیاری از فروشندگان و همفکران مشاور آنها یک مدل فرآیند برای راهنمایی کاربر خود که از طریق یک سری مراحل مشخص او را به نتایج خوبی هدایت خواهد کرد طراحی کردند. برای مثال SPSS از مراحل پنجگانه تشخیص دسترسی تحلیل عمل و اتوماسیون و SAS از مراحل نمونه گیری, جستجو, تغییر و بهبود, مدل سازی و تعیین استفاده می نماید.

اخیرا ائتلاف فروشندگان وکاربران شامل سیستمهای مهندسی NCR کپنهاک, راه حلهای جامعSPSS و بانک OHRA در حال ساختن یک فرآیند خاص که به فرآیند استاندارد صنعتی داده کاوی (CRISP-DM) موسوم است می باشند. این فرآیند برای پردازش مدلهای شرکتهای دیگر که یک کاره یا دو کاره هستند یکسان می باشد. این فرآیند شروع خوبی برای کمک به مردم جهت فهم مراحل ضروری در داده کاوی موفق می باشد.

۱۰-۱مدل فرآیند دو سویه

مدل فرآیند دو سویه که در زیر توضیح داده شده است برخی از موارد پیش بینی را از مدل CRISP-DM به ارث می برد.

گامهای اصلی داده کاوی جهت کشف دانش عبارتند از:

۱ - تعریف مساله

۲ - ساختن پایگاه داده مربوط به داده کاوی

۳ - جستجوی داده

٤ - آماده ساختن داده برای مدل سازی

٥ - ساختن مدل

٦ - ارزیابی مدل

۷ - ساخت مدل ونتایج

به سراغ این گامها می رویم تا فرآیند کشف دانش را بهتر متوجه شویم.

۱ - تعریف مساله

در ابتدای امر پیش زمینه کشف دانش فهم درست داده و مساله می باشد. بدون این فهم درست هیچ الگوریتمی صرف نظر از خبره بودن آن نمی تواند نتیجه مطمئنی برای شما حاصل نماید و همچنین شما قادر نخواهید بود که مسائلی را که سعی در حل آن دارید تعریف کرده و همچنین داده را جهت کاوش آماده نموده و یا نتایج را به طور صحیح تفسیر نمائید. برای استفاده بهتر از داده کاوی شما باید یک بیان واضح از هدف خود داشته باشید.

11-ساختن یک پایگاه داده داده کاوی

این گام به همراه دو گام بعدی هسته آماده سازی داده را تشکیل می دهند. در مجموع گامهای گفته شده وقت و کار بیشتری از سایر گامها می برند. ممکن است شما گامهای تکراری در آماده سازی داده و ساختن مدل داشته باشید چرا که در هر مرحله ممکن است به نکته ای برسید که شما را بر آن دارد داده خود را بهبود بخشید. این گامهای آماده سازی داده می تواند ۵۰٪ تا ۹۰٪ وقت و کار از تمام فرآیند کشف دانش را به خود اختصاص دهد.

داده ای که می خواهد کاوش شود باید در یک پایگاه داده ذخیره شود. بر اساس مقدار داده, پیچیدگی داده و استفاده هایی که قرار است از آن شود یک فایل معمولی و یا یک SpreadSheet برای این کار کافی است.

به احتمال زیاد شما می خواهید داده موجود در انباره داده را تغییر دهید. به علاوه شما ممکن است بخواهید فیلدهای جدیدی که از فیلدهای موجود محاسبه شده است را به انبار داده خود بیافزایید.این یکی از دلایل استفاده از یک پایگاه داده جداگانه است.

دلیل دیگر برای این کار آن است که انبار داده های یکی شده ممکن است به آسانی انواع جستجوهایی را که شما برای فهم داده به آنها نیاز دارید انجام ندهد. مانند پرس و جوهایی که داده را خلاصه می کند, گزارشات چند بعدی و بسیاری از انواع دیگر از گرافها یا مصورات.

و دلیل آخر اینکه شما ممکن است بخواهید این داده را در یک سیستم مدیریت پایگاه داده به همراه یک طراحی فیزیکی متفاوت از انبار داده خود ذخیره کنید. مردم به طور روز افزونی در حال انتخاب پایگاه داده های خاص منظوره ای هستند که این نیازهای داده کاوی را به نحو مناسبی حمایت کند. به هرحال اگر داده موجود در انبار داده شما اجازه می دهد که مراکز منطقی داده ای ایجادکنید و اگر شما می توانید تقاضای داده کاوی را ارضا نمایید پایگاه داده شما به خوبی وظیفه خود را انجام می دهد.

مراحل لازم برای ساخت یک پایگاه داده داده کاوی به شکل زیر می باشد:

۱ - جمع آوری داده ها

۲ - توضیح داده ها

۳ - انتخاب داده ها

٤ - تعیین کیفیت داده ها و پاک کردن آن

٥ - تثبیت و یکپارچگی

٦ - ساختن فوق داده (داده هایی که خود بیانگر توضیحی در مورد داده های موجود می باشند.)

۷ - بارکردن پایگاه داده مربوط به داده کاوی

۸ - نگهداری پایگاه داده مربوط به داده کاوی

این کارها ممکن است لزوما به همین ترتیب گفته شده انجام نگردند.

۱۱-۱جستجوی داده

به بخش توضیح داده برای داده کاوی که توضیح مختصری راجع به اشکال, تجزیه و تحلیل ارتباط و دیگر وسایل جستجوی داده می باشد نگاهی بیاندازید.

هدف شناسایی مهمترین فیلدها در پیش بینی نتیجه و تعیین اینکه کدام یک از داده های بدست آمده مفید می باشد است.

در یک مجموعه داده ای با صدها یا حتی هزاران ستون جستجوی داده می تواند کار و زمان بر باشد. یک واسط مناسب و جواب کامپیوتر سریع در این فاز مهم و حیاتی می باشند زیرا هنگامی که شما برای دریافت پاسخ برخی گراف ها مجبور باشید ۲۰ دقیقه صبر کنید ماهیت جستجوی شما به کلی تغییر خواهد کرد.

۱۱-۲آماده سازی داده برای مدل سازی

این آخرین گام آماده سازی داده قبل از ساخت مدلهاست. چهار قسمت مهم در این مرحله وجود دارد:

۱ - انتخاب متغیرها

۲ - انتخاب سطرها

۳ - ساختن متغیرهای جدید

٤ - تغییر شکل متغیرها

۱۱-۳ساختن مدل داده کاوی

مهمترین مساله برای یادآوری در مورد ساخت مدل آن است که این کار یک فرآیند تکراری است. شما برای جستجو به مدلهای جایگزین جهت یافتن سودمندترین آنها جهت حل مسائلتان نیاز دارید. آنچه که شما در جستجوی یک مدل مناسب یاد می گیرید می تواند شما را به بازگشتن به عقب و انجام برخی تغییرات در داده مورد استفاده خود و حتی بهبود بیان ساله راهنمایی کند.

هنگامی که شما در مورد نوع پیش بینی که می خواهید انجام دهید تصمیم گرفتید باید یک نوع مدل برای ساخت تصمیم خود انتخاب کنید.

آماده سازی و آزمایش مدل داده کاوی احتیاج به این دارد که داده به حداقل دو گروه شکسته شود: یکی برای آماده کردن مدل و دیگری جهت تست مدل مربوطه. اگر شما از آماده سازی و تست متفاوتی استفاده ننمائید دقت مدل خواهد بود.

۱۱-٤تائید اعتبارساده

پایه ای ترین روش تست داده تایید اعتبار ساده می باشد. برای انجام این کار چون درصدی از پایگاه داده را به عنوان یک تست پایگاه داده کنار بگذارید و به هر صورت از آن در برآورد و ساخت مدل استفاده ننمائید. این درصد معمولا بین ۵ تا ۳۳ می باشد.

۱۱-۵ارزیابی و تفسیر

بعد از ساخت یک مدل شما باید نتایج آن را ارزیابی نموده و همچنین اهمیت آن را نیز توضیح دهید.

12ماتریسهای پیچیدگی

برای مسائل طبقه بندی یک ماتریس پیچیدگی ابزار مفیدی برای فهم نتایج می باشد. یک ماتریس پیچیدگی تعداد مقادیر کلاس (گروه)مجازی را در مقایسه با تعداد مقادیر کلاس(گروه) پیش بینی شده نشان می دهد. نه تنها چگونگی پیش بینی مدل توسط این ماتریس نشان داده می شود بلکه نشان دهنده جزئیاتی است که برای نشان دادن موارد اشتباه ضروری است. ستونها کلاسهای مجازی و سطرها کلاسهای پیش بینی شده را نشان می دهند. بنابراین قطرهای این ماتریس بیانگر تمام پیش بینی های درست می باشند. در ماتریس پیچیدگی می بینید که مدل ما ۳۸ تا از ٤٦ تا کلاس **B** را به درستی پیش بینی کرده است اما ۸ تا از آنها اشتباها کلاس بندی شده اند. ۲ تا به عنوان کلاس **A** و٦ تا به عنوان کلاس **C** می باشند.

	Actual		
Prediction	**Class A**	**Class B**	**Class C**
Class A	45	2	3
Class B	10	38	2
Class C	4	6	40

در حالات خاص اگر قیمت های گوناگون با اشتباهات مختلفی در ارتباط باشند یک مدل با دقت کمتر ممکن است بر یک مدل با دقت بیشتر و در ضمن قیمت بیشتر به خاطر انواع اشتباهاتی که ایجاد می کند ترجیح داده شود. برای مثال فرض کنید در ماتریس بالا هر جواب درست قیمتی معادل ۱۰ دلار و هر جواب نادرست برای کلاس A ۵ دلار , برای کلاس B ۱۰ دلار و برای کلاس C ۲۰ دلار داشته باشد. بنابراین هزینه شبکه ای ماتریس معادل:

$$(123 * \$10) - (5 * \$5) - (12 * \$10) - (10 * \$20) = \$885.$$

خواهد داشت.

اما ماتریس شکل بعد را در نظر بگیرید. دقت تا ۷۹٪ کاهش پیدا کرده است هنگامی که همان قیمتهای قبلی را بر روی این ماتریس اعمال کنیم هزینه کل برابر:

$$(118 * \$10) - (22 * \$5) - (7 * \$10) - (3 * \$20) = \$940$$

Prediction	*Actual* Class A	Class B	Class C
Class A	40	12	10
Class B	6	38	1
Class C	2	1	40

بنابراین اگر بخواهید مقدار ارزشی مدل را بیشینه کنید بهتر است که مدلی با دقت کمتر ولی در عوض با ارزش شبکه ای بیشتر انتخاب نمائید.

۱۲-۱ایجاد معماری مدل و نتایج

هنگامی که یک مدل ساخته و تایید اعتبار می شود می تواند در دو راه اصلی مورد استفاده قرار گیرد. راه اول برای تحلیل گر است که اعمالی را بر اساس دید ساده از مدل و نتایج آن معرفی می کند. راه دوم بکاربردن مدلها در مجموعه داده ای مختلف است. این مدل می تواند برای مشخص نمودن رکوردها بر اساس گروه بندیشان و یا مقدار دهی یک امتیاز مثلا احتمال انجام یک عمل استفاده گردد.

هنگام به دست آوردن یک کاربرد پیچیده داده کاوی اغلب اگر چه بخش بحرانی اما کوچک پروژه نهایی به حساب می آید. برای مثال دانشی که از داده کاوی کشف می شود می تواند با دانش متخصصان داده و تراکنشهای ورودی ترکیب شود. در یک سیستم تشخیص فرآیند الگوهای موجود فرآیند می توانند با الگوهای کشف شده تلفیق شوند. هنگامی که موارد مفروض این فرآیند برای ارزیابی به بررسی کنندگان فرستاده می شوند بررسی کنندگان ممکن است نیاز داشته باشند که به رکوردهایی در پایگاه داده که مربوط به قسمتهای ادعا شده توسط یک سازنده است دسترسی پیدا کنند.

به طور کلی مراحلی که توضیح داده شد برای انجام هر فرآیند داده کاوی لازم به نظر می رسد.

نتیجه گیری .

داده کاوی یک گرایش تحقیقاتی است که به سرعت در حال پیشرفت بوده و پژوهش های بسیاری بر روی آن انجام شده است . محققان و برنامه نویسان رشته های مختلف ، در هنر داده کاوی با هم شریک هستند ، از این رو مهیا کردن یک بررسی کلی نسبت به متدهای داده کائی ، کار دشواری است.

در هر صورت این مقاله سعی کرد ، یک بررسی اجمالی بر روی تکنیک های مختلف داده کاوی درپایگاه داده های بزرگ انجام داده و روش های مختلف ارائه شده را مورد معرفی و مقایسه قرار دهد.

منابع و مراجع:

۱- سایت گروه داده کاوی ایران، http://irdatamining.com/articles/visualization/introduction.htm

۲-سایت ویکی پدیا داده کاوی http:// fa.wikipedia.org/wiki

3-سایت http://www.irandatamining.com/IRDM/Fa/Portal

4-سایت http://www.prozhe.com