بسم الله الرحمن الرحیم

عنوان:

مروری بر مبحث داده کاوی

استاد مربوطه: **خانم دکتر محمدی**

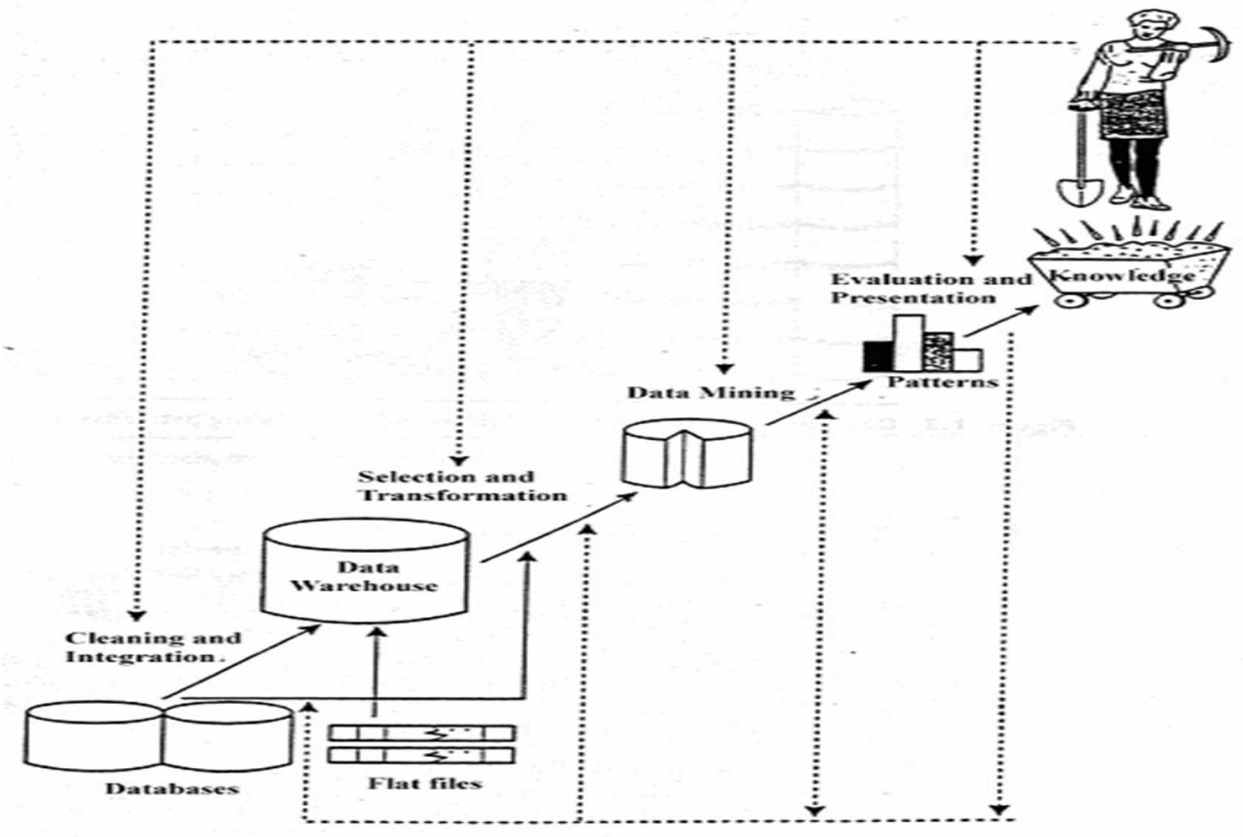
گردآورنده: **فاطمه احمدنژادیان**

پاییز و زمستان 97

بخش اول :مقدمه ای بر داده کاوی

در دو دهه قبل توانایی های فنی بشر در برای تولید و جمع آوری دادهها 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 سیستمه ای 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 و دیگر مخازن بزرگ اطلاعات، ذخیره شده است. داده کاوی بطور همزمان از چندین رشته 2 داده های عظیم، انباره داده علمی بهره می برد نظیر: تکنولوژی پایگاه داده، هوش مصنوعی، یادگیری ماشین، شبکه های عصبی، آمار، شناسایی و بازنمایی بصری داده ، محاسبات سرعت بالا ، بازیابی اطلاعات4 ، حصول دانش 3 الگو، سیستم های مبتنی بر دانش . 7داده کاوی در اواخر دهه 1980 پدیدار گشته، در دهه 1990 گامهای بلندی در این شاخه از علم برداشته شده و انتظار می رود در این قرن به رشد و پیشرفت خود ادامه دهد اغلب به صورت مترادف یکدیگر مورد استفاده قرار می گیرند. 8 واژه های «داده کاوی» و «کشف دانش در پایگاه داده «کشف دانش به عنوان یک فرآیند در شکل1-1 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 تنهایی قابل استفاده نیست، بلکه دانش نهفته در داده ها قابل استفاده می باشد. به این دلیل گفته می شود. اغلب به داده کاوی، تحلیل داده ای ثانویه 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 شکل1-2 این روند تکاملی در پایگاه های داده نشان داده شده است.



شکل 1 مراحل مختلف کشف دانش

تکامل تکنولوژی پایگاه داده و استفاده فراوان آن در کاربردهای مختلف سبب جمع آوری حجم فراوانی داده شده است

ین 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 و تحقیقات علمی و پزشکی، استفاده کرد. شکاف موجود بیندادە ها و اطلاعات سبب ایجاد نیاز برای ابزارهای داده کاوی شده است تا داده های بی ارزش را به دانشی ارزشمند تبدیل کنیم

مراحل کشف دانش

کشف دانش دارای مراحل تکراری زیر است:

1 پاکسازی داده ها (از بین بردن نویز و ناسازگاری داده ها)

-2 یکپارچه سازی داده ها(چندین منبع داده ترکیب می شوند.)

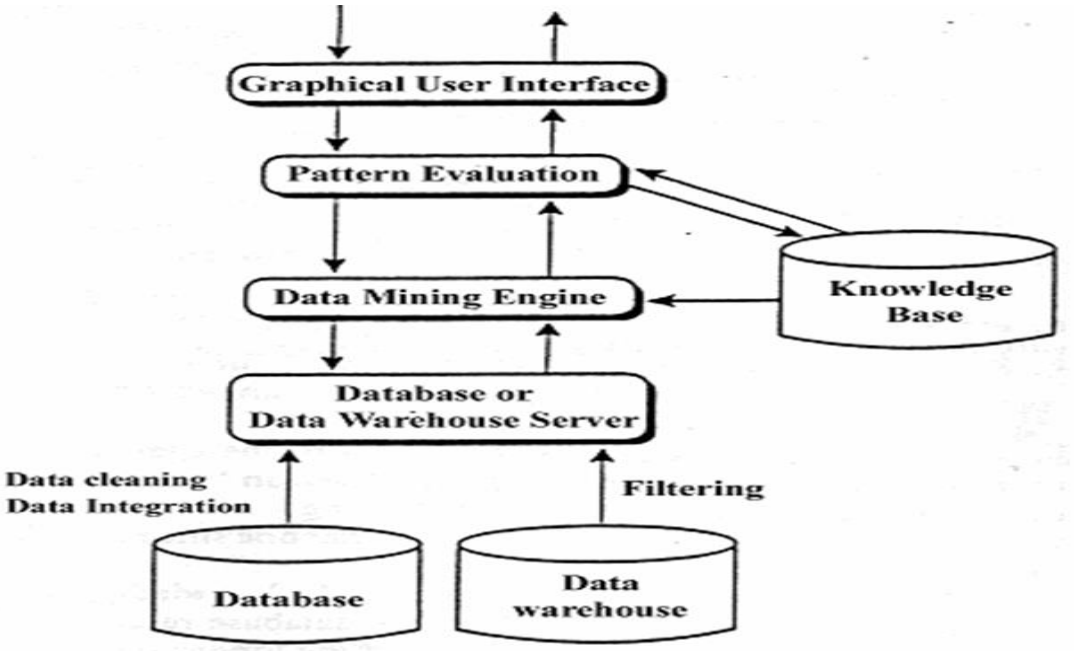
-3 انتخاب داده ها(داده های مرتبط با آنالیزازپایگاه داده بازیابی می شوند.)

-4 تبدیل کردن داده ها(تبدیل داده ها به فرمی که مناسب برای داده کاوی باشد مثل خلاصه سازی و همسان‌سازی)

-5داده کاوی (فرایند اصلی که روالهای هوشمند برای استخراج الگوها از داده ها به کار گرفته می شوند)

-6ارزیابی الگو(برای مشخص کردن الگوهای صحیح و مورد نظربه وسیله معیارهای اندازه گیری)

-7ارائه دانش (یعنی نمایش بصری، تکنیکهای بازنمایی دانش برای ارائه دانش کشف شده به کاربر استفاده میشوند)



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:

1- پایگاه داده، انباره داده یا دیگر مخازن اطلاعات: که از مجموعه ای از پایگاه داده ها، انباره داده، صفحه گسترده

دیگر انواع مخازن اطلاعات. پاکسازی داده ها و تکنیکهای یکپارچه سازی روی این داده ها انجام می شود.

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

-3 پایگاه دانش: این پایگاه از دانش زمینه‌تشکیل شده تا به جستجو کمک کند، یا برای ارزیابی . الگوهای یافته شده از آن استفاده می شود.

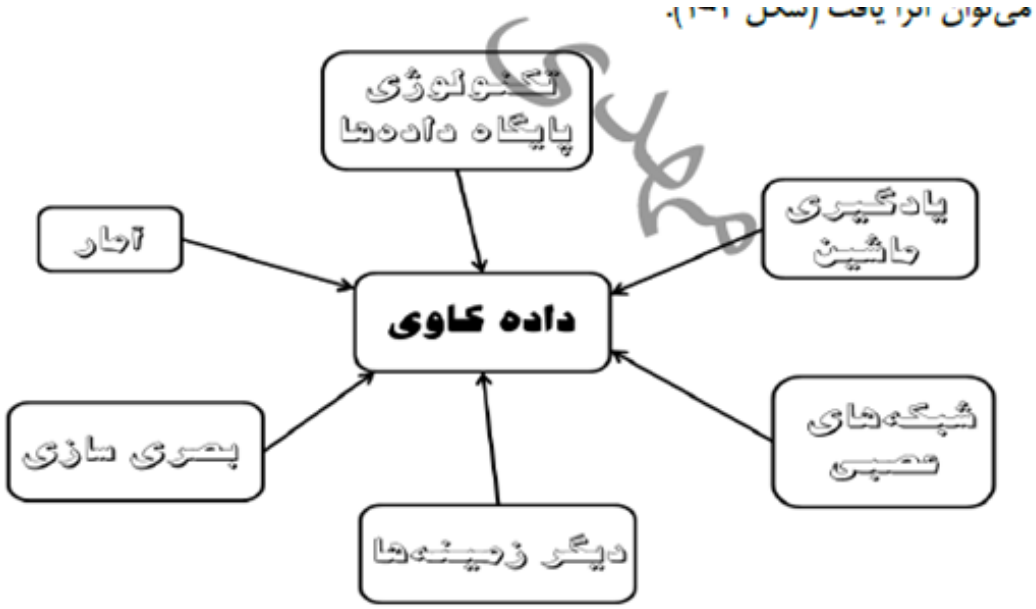
-4 موتور داده کاوی : این موتور جزء اصلی از سیستم داده کاوی است و به طور ایدآل شامل مجموعه ای از پیمانه‌هایی نظیر توصیفcharacterization، تداعیAssociation، کلاسبندیِ Classification، آنالیزخوشه ها Cluster analysis، و آنالیز تکامل وانحرافٍ Evolution and division analysis، است.

5 پیمانه ارزیابی الگوPattern evaluation module: این جزء معیارهای جذابیت Interesting measures 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 تاالگوهای کشف شده را ارزیابی کند.

6- واسط کاربرگرافیکی(GUI) hical User Interface: 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 وجوارتباط 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

جایگاه داده کاوی در میان علوم مختلف



شکل ۲-۱: داده‌کاوی و تجمعی از زمینه‌های مختلف

ریشه های داده کاوی در میان سه خانواده از علوم، قابل پیگیری می باشد مهمترین این خانواده ها، آمارکلاسیک 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 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 1980 عملی نشد. هوش مصنوعی کاربردهای کمی را ابتکاری می باشد و با آمار ضدیت دارد، تلاش دارد تا فرایندی مانند فکر انسان، را برای حل مسائل آماری بکار بندد.

سومین خانواده داده کاوی، یادگیری ماشین 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

•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

•پزشکی: کشف ارتباط علامت و بیماری، تحلیل آرایه های DNA ، تصاویر پزشکی؛

•ورزش: آمارهای ورزشی؛

وب کاوی

:پیشنهاد صفحات مرتبط، بهبود ماشینهای جستجوگر یا شخصی سازی حرکت در وب سایت؛

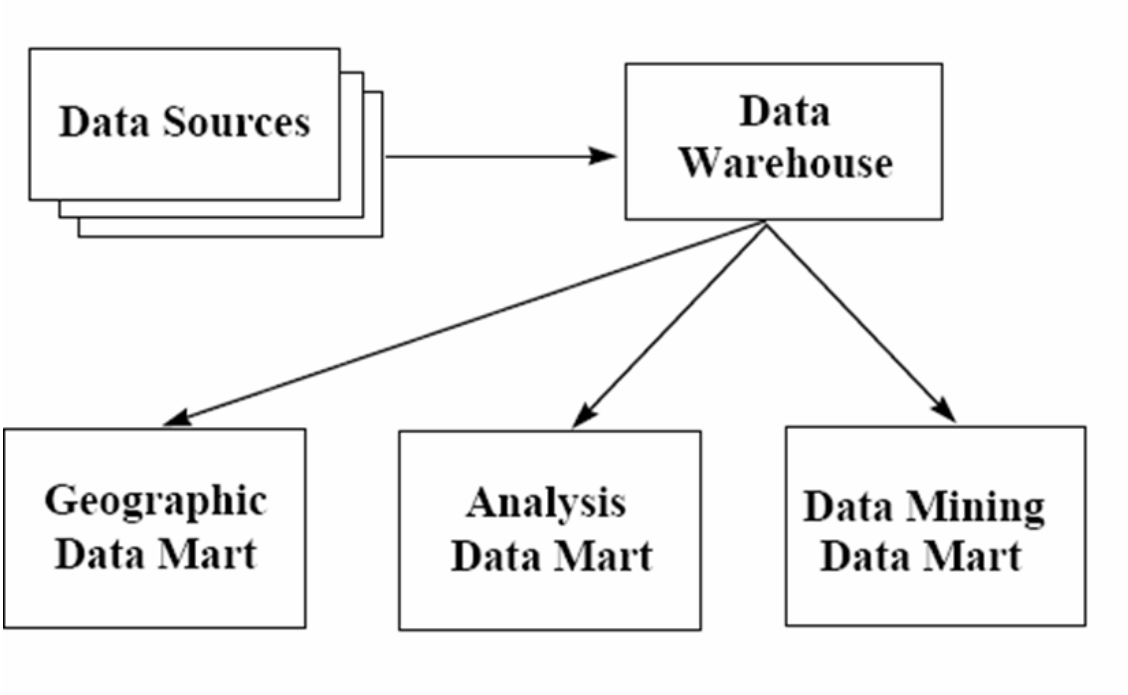
داده کاوی و انبار داده ها

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

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 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

.



داده کاوی وOLAP

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.

.سیستم های سنتی گزارش گیری و پایگاه داده ای آنچه را که OLAP 45 جزیی از می باشد ابزارهای تصمیم گیری 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

بخش دوم:توصیف داده ها در داده کاوی

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 ایجاد 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 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

خوشه بندی

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 اطلاعاتدر مورد داده ها یا تقسیم داده ها جدید استفاده کنیم.

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

تحلیل لینک

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 و کشف ترتیب 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 B->A نمایش داده می شوند که به A مقدم و به B موخر یا نتیجه گفته می شود.مثلا اگریک 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.

بخش سوم : مدل های پیش بینی داده ها

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

داده کاوی مدلهای classification 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.

Regression

Regressionاز مقادیر موجود برای پیشبینی مقادیر دیگر استفاده میکند. در سادهترین فرم، regressionاز تکنیکهای آماری استاندارد مانند regression linear استفاده میکند. متاسفانه، بسیاری مسائل دنیای واقع تصویرخطی سادهای از مقادیر قبلی نیستند. بناراین تکنیکهای پیچیده تری(regression logistic ،درختهای تصمیم، یا شبکههای عصبی) ممکن است برای پیشبینی مورد نیاز باشند.

Time series

پیشبینی های series Time مقادیر ناشناخته آینده را براساس یک سری از پیشبینی گرهای متغیر با زمان پیش بینی 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 سلسله مراتب دورهها را.

بخش چهارم: 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

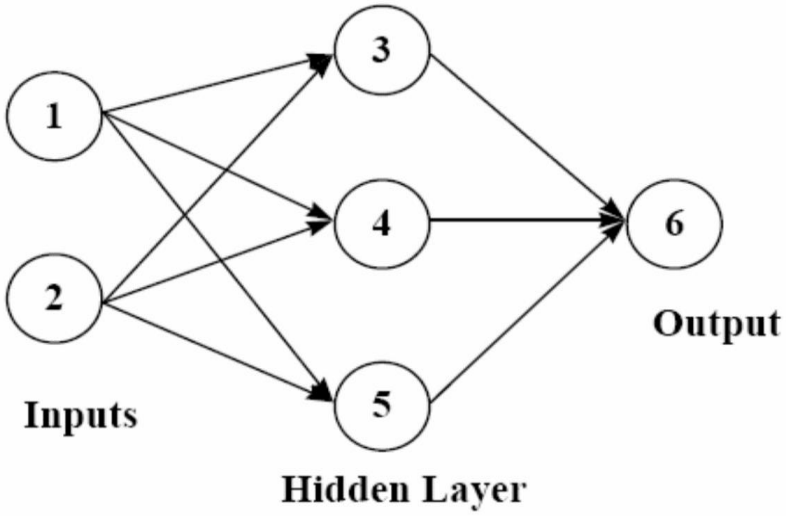
شبکه های عصبی

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

هر شبکه عصبی شامل یک لایه ورودی می باشد که هر گره در این لایه معادل یکی از متغیرهای پیش بینی می . هر گره ورودی به همه گره های 52 باشد. گره های موجود در لایه میانی وصل می شوند به تعدادی گره در 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 وصل شوند

لایه خروجی شامل یک یا چند متغیر خروجی می باشد.

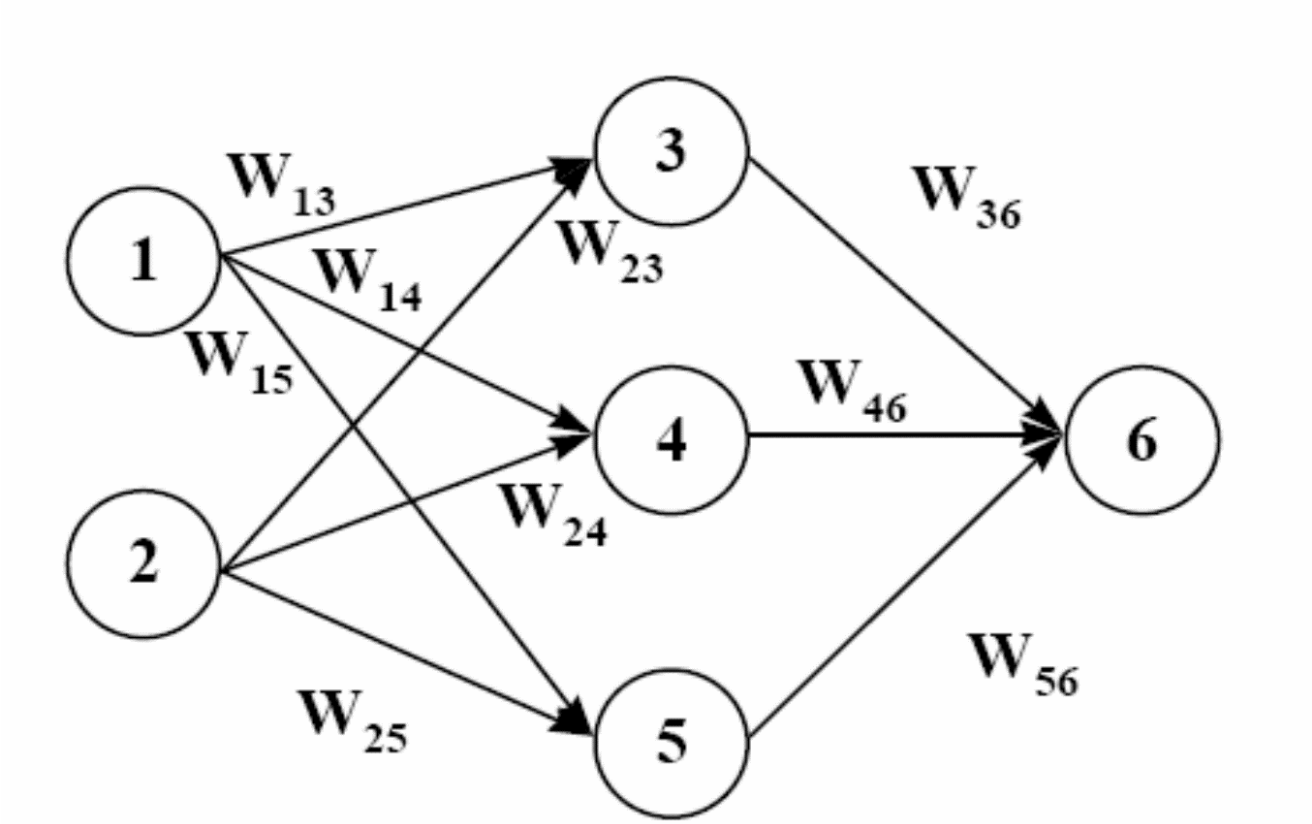


هر یال که بین نود های Y,X می باشد دارای یک وزن است که با y,Wx نمایش داده می شود. این وزن ها در محاسبات لایه های میانی استفاده می شوند و طرز استفاده آنها به این صورت است که هر نود در لایه های میانی (لایه های غیر از لایه اول) دارای چند ورودی از چند یال مختف می باشدکه همانطور که گفته شد هر کدام یک وزن خاص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

.



از مهمترین انواع شبکه های عصبی Backpropagation Forward-Feed می باشد

Forward-Feedبه 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

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

براساس الگوریتم، ممکن است دو یا تعداد بیشتری شاخه داشته باشد. برای مثال، CARTدرختانی 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، درختهای 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 400001 است از نظر ریسک اعتبار خوب باشد درحالیکه کسی که حقوقش 40000 است بد باشد. بعلاوه، ازآنجاکه 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 متغیرهاپیشبینی کننده 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 کهغیرهوشمندانه است، اما مسئله مهمتر اینستکه با دادهها overfit میشوند. اندازه درختها را میتوان از طریق قوانینتوقف 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 روش های 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 نمی گیرد

نتیجه گیری

در دنیای امروز و در اقتصاد دیجیتالی و به خصوص در حوزه های خدمات دولت الکترونیکی، اطلاعات زیادی در فرمت متن وجود دارند که میتوان به راحتی آنها را در کلاسهای از پیش تعریف شده طبقه بندی و رده بندی کرد که البته حدود 80 درصد از اطلاعات در دسترس به عنوان اسناد متنی در دسترس است. این اطلاعات اغلب در بیشتر داده های توصیفی مانند گزارشها، اطلاعات به دست آمده از مشتریان، ساخت مستندات کیفیت، تحقیقات میدانی و تجزیه و یادداشتها و غیره هستند. برای بهبود عملکرد و ارائۀ خدمات تحلیلهای تئوری زمینهای باکیفیتتر در آینده و ارائۀ راه حل، باید اطلاعات موجود را به فرمتهای قابل استفاده تبدیل کرد.

تصمیمگیرندگان و کارکنان دانشی سازمان و بهخصوص مدیران دانشی، تصمیمات کسب وکار خویش را از طریق کشف الگوهای دانش به کار میگیرند که سبب کاهش هزینه های سربار از خدمات، بهبود کیفیت و مدیریت بهتر میشود. همزمان با رشد فزایندۀ تحولات اقتصادی اجتماعی، تأثیر دانش و مدیریت تجربه های سازمانی به ویژه سازمانها و ارگانهای دولتی

به شدت احساس میشود مدیریت دانش، توانایی سازمانها برای یادگیری از محیط خود و مشارکت دادن دانش در فرایندهای کسب وکار و تصمیمگیری را افزایش میدهد

روشهای داده کاوی مزایایی دارد که سبب مدیریت بهتر منابع دانش و فعالیتهای مدیریت دانش میشود. داده کاوی در کشف دانش مفید برای کمک به پردازش اطلاعات و بهبود بهره

وری کارکنان دانشی سازمان استفاده مـیشـود. نتیجـۀداده کاوی، افـزایش ارزش افـزودۀ کسب وکار به منظور تسهیل فرایند تصمیم گیری و کاهش هزینه، نسـبت بـه سـایر تکنیـکهـای پردازش متن است. در اصل برای بهدست آوردن مزایـای رقـابتی تـر و بهـره بـرداری از اطلاعـات چندگانه، روشهای کشف دانش در نظر گرفته میشود..

منابع :

رستمی, مجتبی و محسن حاجی زین العابدینی، ۱۳۹۴، داده کاوی و کاربرد آن در مدیریت دانش، هشتمین کنفرانس ملی و دومین کنفرانس بین المللی مدیریت دانش، تهران، موسسه اطلاع رسانی نفت، گاز و پتروشیمی

غنیمت, الناز و محمد هادی صدرالدینی، ۱۳۹۵، بررسی ضرورت بهره گیری از تکنیک های داده کاوی در مدیریت پروژه های صنایع دریایی، هجدهمین همایش صنایع دریایی، جزیره کیش، انجمن مهندسی دریایی ایران، https://www.civilica.com/Paper-NSMI18-NSMI18_036.html

پیرحیاتی, آزاده، ۱۳۹۷، روش های داده کاوی برای یافتن داده های تکراری در پایگاه داده های بزرگ، دومین کنفرانس ملی دانش و فناوری علوم مهندسی ایران، تهران، موسسه برگزار کننده همایش های توسعه محور دانش و فناوری سام ایرانیان، https://www.civilica.com/Paper-MGCONF02-

زنجانی، احسان . مقدمه ای بر داده کاوی ، پایان نامه کارشناسی 92