**وب سایت تخصصی برق و الکترونیک ECA**

**عنوان :**

# مقدمه‌ای بر داده‌کاوی

**نگارش :**

**تیم تالیف مقاله**

**زمستان ۱۳۸۸**

## ۱ - مقدمه ای بر داده‌کاوی[1]

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 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 داده‌های 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 پایگاه داده، هوش مصنوعی، یادگیری ماشین، شبکه های عصبی، آمار، شناسایی الگو، سیستم های مبتنی بر دانش[3]، حصول دانش[4]، بازیابی اطلاعات[5]، محاسبات سرعت بالا[6] و بازنمایی بصری داده[7] .

---

[1] Data Mining
[2] Data warehouses
[3] Knowledge-based system
[4] Knowledge-acquisition
[5] Information retrieval
[6] High-performance computing
[7] Data visualization

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

اصلی ترین دلیلی که باعث شد داده کاوی کانون توجهات در صنعت اطلاعات قرار گیرد، 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.

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[8] 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.

**۱-۱ مراحل کشف دانش**

کشف دانش دارای مراحل تکراری زیر است :

[8] Knowledge base

۱- پاکسازی داده ها[9] (از بین بردن نویز و ناسازگاری داده ها).

۲- یکپارچه سازی داده ها[10] (چندین منبع داده ترکیب می شوند).

۳- انتخاب داده ها[11] (داده های مرتبط با آنالیزازپایگاه داده بازیابی می شوند).

٤- تبدیل کردن داده ها[12] (تبدیل داده ها به فرمی که مناسب برای داده کاوی باشد مثل خلاصه سازی[13] و همسان سازی[14]

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

٦-ارزیابی الگو[15] (برای مشخص کردن الگوهای صحیح و مورد نظربه وسیله معیارهای اندازه گیری)

۷-ارائه دانش[16] (یعنی نمایش بصری، تکنیکهای بازنمایی دانش برای ارائه دانش کشف شده به کاربر استفاده می شود).

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" .

---

[9] Data cleaning
[10] Data integration
[11] Data selection
[12] Data transformation
[13] Summary
[14] Aggregation
[15] Pattern evaluation
[16] Knowledge presentation

۱- پایگاه داده، انباره داده یا دیگر مخازن اطلاعات: که از مجموعه ای از پایگاه داده ها، انباره داده، صفحه گسترده[۱۷]، 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[۱۸] تشکیل شده تا به جستجو کمک کند، یا برای ارزیابی الگوهای یافته شده از آن استفاده می شود.

٤- موتور داده کاوی[۱۹] : این موتور جزء اصلی از سیستم داده کاوی است و به طور ایدآل شامل مجموعه ای از پیمانه[۲۰] هایی نظیر توصیف[۲۱]، تداعی[۲۲]، کلاسبندی[۲۳]، آنالیزخوشه ها[۲٤]، و آنالیز تکامل وانحراف[۲۵]، است.

۵- پیمانه ارزیابی الگو[۲٦] : این جزء معیارهای جذابیت[۲۷] 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[۲۸] : 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[۲۹] ارتباط برقرار کند، این جزء به کاربر اجازه می دهد

---

[17] Spread sheets
[18] Domain knowledge
[19] Data mining engine
[20] Module
[21] Characterization
[22] Association
[23] Classification
[24] Cluster analysis
[25] Evolution and deviation analysis
[26] Pattern evaluation module
[27] Interesting measures
[28] Graphical User Interface (GUI)
[29] Query

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[۳۰] 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.

**۱-۲ 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[۳۱] 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[۳۲] 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

[30] Query processing
[31] Classic Statistics
[32] Artificial Intelligence

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.

سومین خانواده داده کاوی، یادگیری ماشین[۳۳] 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)؛

[33] Machine Learning

- **متن کاوی**[۳۴]: 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**[۳۵]: پیشنهاد صفحات مرتبط، بهبود ماشینهای جستجوگر یا شخصی سازی حرکت در وب سایت؛

**۱-۳ 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.

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.

**۱-٤ داده کاوی و انبار داده ها** [۳۶]

معمولا داده هایی که در داده کاوی مورد استفاده قرار می گیرند از یک انبار داده استخراج می گردند و در یک پایگاه داده[۳۷] یا مرکز داده [۳۸]ای ویژه برای داده کاوی قرار می گیرند.

---

[34] Text Mining
[35] Web Mining
[36] Data warehouse
[37] Database
[38] Data mart

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.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.

**۱-۵ داده کاوی و OLAP**

بسیاری فکر می کنند که داده کاوی و OLAP دو چیز مشابه هستند این دو ابزار های کاملا متفاوت می باشند که می توانند همدیگر را تکمیل کنند.

OLAP جزیی از ابزارهای تصمیم گیری [۳۹]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.

[39] Decision Support Tools

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 .

### ۱-٦ 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.

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

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.

### 2-۲ خوشه بندی ٤٠

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 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.

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

[40] Clustering

**۲-۳ تحلیل لینک[41]**

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** [42]و **کشف ترتیب** [43]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 می‌شوند که به A مقدم و به B موخر یا نتیجه گفته می شود.مثلا اگر یک قانون به صورت زیر داشته باشیم.

[41] Link Analysis
[42] Association discovery
[43] 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 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.

### 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 می‌کند. متاسفانه، بسیاری مسائل دنیای واقع تصویرخطی ساده‌ای از مقادیر قبلی نیستند. بنابراین 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 درخت-های regression استفاده کرد. شبکه‌های عصبی را نیز می‌توان برای هر دو مورد استفاده کرد.

**Time series ۳-۳**

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 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 ٤٤**

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 [٤٥]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[٤٦] . 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 [٤٧]وصل شوند.

لایه خروجی شامل یک یا چند متغیر خروجی می باشد.

هر یال که بین نود های X,Y می باشد دارای یک وزن است که با Wxy 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.

---

[44] Neural Networks
[45] Input Layer
[46] Hidden Layer
[47] Output Layer

وزن یال ها پارامترهای ناشناخته ای هستند که توسط تابع آموزش [48]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 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** [49]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 Over fitting می باشد و آن بدین صورت است که که شبکه فقط روی داده

[48] Training method
[49] Epoch

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 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.

## Decision trees ۲-٤

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، 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، 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.

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 over fit 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 اینست که 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.

**Multivariate Adaptive Regression Splines(MARS) ۳-٤**

در میانه‌های دهه ۸۰ یکی از مخترعین CART، Jerome H. Friedman_، متدی را برای برطرف-کردن 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 over fit 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.

**Rule induction ٤-٤**

استنتاج قوانین متدی برای تولید مجموعه‌ای از قوانین است که موارد را دسته‌بندی می‌کند. اگرچه درخت-های 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 رده-بندی 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.

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[50]

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 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.

[50] Logistic regression

### ٤-٧ تحلیل تفکیکی ٥١

این روش از قدیمی ترین روش های ریاضی وار گروه بندی داده ها می باشد که  برای اولین بار در سال ١٩٣٦ 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 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.

اما به دلایلی محبوبیت خود را از دست داده است : ١-این روش فرض می کند همه متغیرهای پیش بینی به صورت نرمال توزیع شده اند که در بسیاری از موارد صحت ندارد . ٢- داده هایی که به صورت عددی نمی باشند مثل رنگها در این روش قابل استفاده نمی باشند. ٣- 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 ٢ نیز باشند که باعث بهبود کارایی و حساسیت در بسیاری از موارد می گردد.

[51] Discriminant analysis

### ٤-٨ مدل افزودنی کلی (GAM)[52]

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 .

[52] Generalized Additive Models

## ۵ - 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: هدف 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 مانند logistic regression، discriminant analysis، و یا مدلهای خطی عمومی وجود دارد.

**الگوریتم**های 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 پیاده‌سازی های 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 پیش‌بینی‌کننده 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 می‌شوند.( در مقابل، تکنیک‌های توصیفی مانند clustering، unsupervised learning نامیده می‌شوند زیرا که هیچ نتیجه از پیش معلومی برای راهنمایی الگوریتم وجود ندارد.)

**سپاسگزاری**:

با سپاس فراوان از تمامی عزیزانی که در گرد آوری این مقاله کمک و یاری کرده اند.

## مراجع :


- **Daniel T. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*. WILEY 2004.**
- **David Hand, Heikki Mannila , Padhraic Smyth. *Principles of Data Mining.* The MIT Press . 2001.**
- **J.Han, and M.Kamber, *"Data Mining: Concepts and Techniques"*, San Diego Academic Press, 2001.**
- **Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth.*From Data Mining to Knowledge Discovery in Databases.*1996.**
- **Berry. Michael, *Survey of text Mining*, *Clustering, Classification and Retrieval.* 2004.**
- **Daniel T. Larose , *Data Mining Methods and Models,* February 2006, Wiley-IEEE Press**
- ***http://en.wikipedia.org/wiki/Data_mining***
- ***http://dataminingarticles.com/***
- ***http://databases.about.com/od/datamining/a/datamining.htm***
- ***http://databases.about.com/od/datamining/a/datamining.htm***