

ارائه درس داده کاوری

فصل دوم آماده سازی داده ها :

آماده سازی داده‌ها کتابخانه‌های دیجیتال آرشیوی از تصاویر اطلاعات پزشکی بیماران داده‌های مربوط به تجارت و خرید و فروش و همچنین داده‌های علمی نمونه‌های بارزی از داده‌ها هستند دانش از آنها بدون شک مهم است. وقتی مقیاس داده‌ها و کاربر روی آنها بالاتر از قابلیت‌های انسانی قرار می‌گیرند نیاز به تکنولوژی‌های محاسباتی به جای تحلیل دستی و سنتی بیشتر احساس می‌شود نکته حائز اهمیت در این میان آماده سازی داده‌ها برای یک تحلیل هوشمند است .

1-1-2) انواع داده‌ها و خصوصیات آنها

صرف نظر از اینکه داده‌ها ممکن است 1-ساخت یافته 2-نیمه ساخت یافته و یا 3-غیر ساخت یافته باشند ممکن است دو گونه از داده‌ها داشته باشیم داده‌های کمی و داده‌های کیفی

1-2-1) متغیرهای کمی: هرگاه صفت خاصه و داده مورد نظر را بتوان شمارش و یا اندازه‌گیری کرد و سپس آن را به صورت عدد بیان نمود یک متغیر کمی خواهیم داشت که به آن متغیر عددی نیز می‌گویند. که به دو ورت گسسته و پیوسته بیان می‌شوند.

1. پیوسته: متغیرهای پیوسته متغیرهایی هستند که می‌توانند کلیه مقادیر حقیقی بین محدوده‌ای را داشته باشند وزن و قد نمونه‌هایی از این نوع هستند
2. گسسته متغیرهایی مثل تعداد فرزندان یک خانواده و یا تعداد دروس انتخاب شده توسط یک دانشجو که در اثر شمارش به دست می‌آیند از تغییر متغیرهای گسسته یا جدا هستند

2-1-2) متغیرهای کیفی

متغیرهایی مانند جنسیت اشخاص محل تولد آدرس و رنگ چشم را متغیرهای کیفی می‌نامیم

2-2) ابعاد بالای داده ها: یکی از مشکلات فرایند داده کاوی از آنجا ناشی میشود که ما با مقدار خیلی زیاد و متنوعی از داده ها روبرو هستیم، که هریک دارای رفتار متفاوتی هستند. جمع آوری صفات خاصه در انبار داده ها بطور طبیعی افزایش ابعاد داده‌ها را به دنبال دارد و این یکی از مشکلات اساسی شناخته شده در داده کاوی است.

3-2) آمار: این بخش به صورت خلاصه به بیان برخی از شاخص‌های مرکزی و شاخص‌های پراکندگی در آمار توصیفی می‌پردازیم

1-2-3) شاخص‌های مرکزی: این شاخص‌ها مقادیری هستند که با استفاده از آنها مجموعه‌ای از داده‌ها در یک اندازه یا عدد که نماینده آن مجموعه است بیان می‌شود مهم‌ترین و همچنین پرکاربردترین شاخص‌های مرکزی می‌توان به 1- میانگین 2- میانه 3- نما یا مد اشاره کرد

میانگین: یکی از معروف‌ترین شاخص‌های مرکزی میانگین است که انواع آن عبارت است از میانگین حسابی میانگین هندسی و میانگین هارمونیک

میانه: عددی است که توضیح داده‌ها را به دو قسمت مساوی تقسیم می‌کند به نحوی که نیمی از داده‌ها بزرگتر و نیمی دیگر کوچکتر از آن هستند نابراین در مجموعه اعداد مرتب شده داده میانی به عنوان میانه لحاظ می‌گردد

نما یا مد: داده‌ای که فراوانی آن از سایر داده‌ها بیشتر باشد را نما یا مد می‌نامیم اگر دو داده مجاور دارای بیشترین فراوانی باشند نصف مجموع آن به عنوان نما لحاظ می‌گردد

2-3-2) **شاخص‌های پراکندگی:** یکی از موضوعات مهم در مورد داده‌ها میزان تغییرات و پراکندگی آنها است بدین معنی که اندازه‌گیری تا چه اندازه از داده به داده دیگر تغییر می‌کند.

مهمترین شاخص‌های پراکندگی عبارتند از: 1- دامنه تغییرات و میانگین انحرافات 2- انحراف چارکی 3- واریانس و انحراف استاندارد.

2-3-3) **کواریانس و ضریب همبستگی:** در نظریه احتمالات کواریانس اندازه تغییرات هماهنگ ۲ متغیر تصادفی است اگر دو متغیر یکی باشند کواریانس برابر با واریانس خواهد بود.

2-3-4) **توزیع نرمال:** مهم ترین توضیح پیوسته در آمار و احتمال توزیع نرمال است که اغلب متغیرهای تصادفی پیوسته در طبیعت از آن پیروی می‌کنند

2-4) **لزوم آماده سازی داده‌ها:** پایگاه داده‌های امروزی به دلیل حجم بسیار بالا مستعد داده‌های نادرست و ناسازگار هستند. به طور حتم داده‌های ناقص و افزونه می‌تواند در نتایج عملیاتی مثل داده کاوی خلل ایجاد کند اصطلاح داده‌های ناقص به وضعیتی اشاره می‌کند که داده‌ها موجود حاوی اطلاعات کافی جهت استخراج دانش جدید نیستند. به طور کلی و خلاصه می‌توان دو وظیفه زیر را برای مرحله آماده سازی داده‌ها در نظر گرفت: 1- سازماندهی داده‌ها در یک شکل استاندارد تا برای پردازش در عمل داده کاوی مناسب باشند . 2- آماده سازی مجموعه داده‌ها تا الگوریتم‌های داده کاوی بتوانند با کارایی بالایی اجرا شوند . به تکنیک آماده‌سازی داده‌ها تکنیک‌های پیش پردازش داده‌ها نیز گفته می‌شود .

2-5) **تکنیک‌های آماده‌سازی داده‌ها:** آماده سازی داده‌ها به مراحل قبل از داده کاوی اطلاق می‌گردد هرچند از این تکنیک‌ها می‌توان در حین اجرای الگوریتم‌های داده کاوی نیز استفاده نمود آماده سازی داده‌ها یکی از مهم‌ترین گام‌ها در فرایند توسعه مدل‌ها به شمار می‌رود کیفیت داده‌های ورودی در ساده‌ترین تحلیل تا ساخت مدل‌های پیچیده یکی از کلیدهای موفقیت انجام یک پروژه به حساب می‌آید

در این بخش به بررسی تکنیک‌های آماده‌سازی داده‌ها در قالب چهار نوع عملیات 1- پالایش داده‌ها 2- جمع آوری داده‌ها 3- تغییر شکل داده‌ها 4- کاهش حجم داده‌ها می‌پردازیم.

1-5-2) **پالایش داده‌ها:** حتی در برنامه‌های کاربردی واقعی مقدار بسیار زیاد داده‌ها می‌توان نمونه‌هایی را یافت که مقداری برای صفحات خاص آنها وجود ندارد اگر تکنیک‌ها از الگوریتم قدرتمندی استفاده کنند شاید این مقادیر ناقص در نتیجه نهایی فرایند بی تاثیر یا حداقل کم اثر باشد اما تکنیک‌های داده کاوی کم و بیش به این مقادیر حساس هستند. ساده‌ترین راه حل برای این مشکل صرف نظر از نمونه‌های ناقص است و این در صورتی امکان‌پذیر است که داده‌های در دسترس به اندازه کافی وجود داشته باشد و داده‌های ناقص کسر کوچکی از داده‌ها را تشکیل دهند.

در مواردی که تعداد عضوهای دامنه ی یک متغیر کم یا تعداد داده های ناقص کم باشند، میتوان به صورت دستی نمونه های ناقص را کامل کرد. همانطور که مشخص است، این راهکار برای پایگاه داده ی بسیار بزرگ زمانگیر است و از همه مهمتر تعیین مقدار برای داده‌های ناقص حتی به صورت دستی مشکل بزرگی است. استفاده از هر مقداری برای تکمیل داده های ناقص ممکن است در تحلیل نهایی موثر باشد.

دسته‌بندی داده‌ها: داده‌های کمی را می‌توان دسته‌بندی کرد و این تکنیک هم می‌تواند به منظور تشخیص و حذف داده‌های نویز و مزاحم استفاده شود هم اینکه برای کاهش حجم داده‌ها نیز روش مفیدی است .

خوشه بندی: فرایند گروه بندی مجموعه‌ای از داده‌ها و قرار دادن آنها در طبقاتی از نمونه‌های مشابه خوشه‌بندی نام دارد. یک خوشه مجموعه‌ای از داده‌ها است که نسبت به دیگر داده‌های همان خوشه شبیه بوده ولی متفاوت از نمونه‌های دیگر خوشه‌ها هستند. تحلیل خوشه‌ها کاربردهای بسیار مختلفی از جمله تشخیص الگو تحلیل داده‌ها پردازش تصاویر و تحلیل‌های تجاری دارد.

رگرسیون: این تکنیک آماری قدرتمند می‌تواند در کلیه مراحل فرایند داده کاوی داده‌ها استفاده شود . رگرسیون خطی ، رگرسیون چندگانه ، رگرسیون غیر خطی از انواع آن به شمار می‌روند . در رگرسیون خطی دو نمونه مختلف از داده‌ها به صورت یک خط راست مدل سازی می‌شوند. رگرسیون چندگانه حالت توسعه یافته‌ای از رگرسیون خطی است با این تفاوت که در آن بیش از دو متغیر وجود دارد و مقدار یکی با توجه به دیگر متغیرها تخمین زده می‌شود. با تعمیم مدل‌های خطی

می‌توانیم برای متغیرهای طبقه بندی شده نیز تخمینی داشته باشیم رگرسیون لجستیکو رگرسیون پویسون واع دیگری از رگرسیون‌ها هستند که این کار را انجام می‌دهند

2-5-2) جمع آوری داده‌ها: جمع آوری و ترکیب داده‌ها از چندین منبع به درون یک محل منسجم به نام انبارداده‌ها یکی از وظایف اولیه در فرایند داده کاوی محسوب میشود. این منابع می‌تواند شامل چند پایگاه داده ناهمگن باشد افزونگی و در نتیجه ناسازگاری داده‌ها کی از مسائل مهمی هستند که در جمع آوری داده‌ها باید به آن توجه کرد جمع آوری و ترکیب داده‌ها از چندین منبع به درون یک محل منسجم به نام انبار داده‌ها یکی از وظایف اولیه در فرایند داده کاوی محسوب می‌شود صفحات خاصه یکسانی که از منابع مختلف جمع آوری می‌شوند را می‌توان توسط داده‌های متا که معمولاً هر پایگاه داده یا انبار داده شامل آن است تشخیص داد و بدین ترتیب از افزونگی جلوگیری نمود.

3-5-2) تغییر شکل داده‌ها: شکل مناسب داده‌ها به عنوان ورودی الگوریتم‌های داده کاوی نقش بسزایی در این فرایند بازی می‌کند و در مرحله آماده سازی داده‌ها این نقش پررنگ است استفاده از توابع تجمعی، نرمال سازی، خوشه‌بندی و رگرسیون روش‌های مرسوم هستند که برای تغییر و تبدیل شکل داده‌ها می‌توان از آنها استفاده نمود در جمع آوری داده‌ها و ساخت انبار داده‌ها از توابع تجمعی نظیر ماکزیمم مینیوم میانگین و مجموع استفاده می‌شود ن توابع علاوه بر تغییر شکل داده‌ها می‌توانند از حجم آن نیز بکاهند.

4-5-2) کاهش داده‌ها:

پالایش داده‌ها و تغییر شکل آنها و همچنین تکنیک‌هایی که در مرحله جمع آوری داده‌ها استفاده می‌شوند برای داده‌هایی با حجم پایین و حتی متوسط مناسب هستند استخراج دانش از داده‌هایی حجم بسیار بالا مستلزم زمان زیادی است شاید مقدار زیاد داده‌ها نتایج بهتری را بتوانیم به دست بیاوریم ولی نمی‌توان به جرات گفت داده‌های کم دارای بار اطلاعاتی کمی هستند تکنیک‌های کاهش داده‌ها می‌توانند از دست دادن درستی داده‌ها و بدون به مخاطره انداختن نتایج نهایی داده کاوی وارد عمل شود اهداف کلی روش‌های داده کاوی: تسریع در روند داده کاوی، افزایش دقت و کارایی و ساده سازی داده‌ها می‌باشد. در تکنیک‌های کاهش داده‌ها عمل کاهش می‌تواند در سه سطح انجام شود کاهش صفات خاصه یا یا ستون، کاهش نمونه‌ها یا سطر و کاهش در مقادیر یک صفت خاصه

1-4-5-2) کاهش صفات خاصه: تکنیک‌های کاهش صفات خاصه و یا به عبارت دیگر انتخاب صفات خاصه مناسب بر اساس دانش برنامه کاربردی و تعریف یک معیار می‌تواند زیر مجموعه مناسب از صفحات خاصه اولیه را انتخاب کنند. چنانچه داده‌های اولیه حاوی تعداد صفحات خاصه کمی باشند کلیه زیر مجموعه صفحات خاصه را جهت یافتن صفات خواسته بهینه بررسی پارامترهای ارزیابی بر اساس سازگاری نمونه‌ها، محتویات آنها، فاصله بین نمونه‌ها و یا وابستگی‌های آماری تعریف می‌شود. به هر حال انتخاب صفات خاصه مفید می‌تواند مشکل و زمانبر باشد به خصوص هنگامی که رفتار داده‌ها خوب شناخته شده نیست.

روش‌های کاهش ابعاد داده به دو دسته روش‌های مبتنی بر استخراج ویژگی و همچنین روش‌های مبتنی بر انتخاب صفات خاص تقسیم می‌شوند. در روش اول یک فضای چند بعدی به یک فضایی با ابعاد کمتر نگاشت می‌شود در واقع با ترکیب صفات خاصه موجود تعداد کمتری ویژگی به وجود می‌آورند به طوری که این ویژگی‌ها دارای تمام یا بخش اعظمی از اطلاعات موجود در صفحات خاصه اولیه باشند این روش‌ها به دو دسته خطی و غیر خطی تقسیم می‌شوند و تکنیک pca بهترین روش برای کاهش ابعاد داده‌ها به صورت خطی است در روش‌های مبتنی بر انتخاب صفات خاصه تلاش می‌شود با انتخاب زیر مجموعه‌ای از ویژگی‌های اولیه، ابعاد داده‌ها کاهش داده شود.

بررسی صفات خاص به صورت مجزا برای کاهش ابعاد: یکی از تکنیک‌های موجود برای انتخاب صفات خاصه مناسب، استفاده از میانگین و واریانس داده‌ها

محاسبه آنتروپی برای رتبه‌بندی صفات خاصه: ایده اصلی در این روش بدین صورت است که ما صفات خاصی را حذف می‌کنیم که حذف آن لطمه‌ای به مشخصات کلی داده‌ها وارد نمی‌سازد به عبارتی دیگر بود با نبود آن باز هم می‌توانیم بین نمونه‌ها تمایز قائل شویم و در درجه نظم داده‌ها تغییرات محسوسی حاصل نمی‌شود

تحلیل مؤلفه‌های اصلی یا pca: عمومی‌ترین روش‌های آماری به منظور کاهش ابعاد داده‌ها روش تحلیل مؤلفه‌های اصلی است در این روش واریانس کل صفات خاصه موجود تحلیل می‌شود. مؤلفه‌ها طوری برآورد می‌گردند تا واریانس صفات خاص را در کمترین ابعاد نشان

2-4-5-2) کاهش نمونه‌ها:

بدون شک تعداد نمونه‌ها بیشترین مقدار داده‌ای هستند که ما در حجم بسیار داده‌ها با آنها روبرو هستیم و باید با استفاده از تکنیک‌های کاهش داده‌ها مقدار آن را در حد مطلوبی قرار دهیم کاهش نمونه‌ها یکی از پیچیده‌ترین وظایف در روش‌های کاهش داده‌ها محسوب می‌شود اما با این حال راه‌هایی وجود دارند که می‌توانیم با کمک آنها حجم نمونه‌ها را کم کنیم یکی از ساده‌ترین و معمولی‌ترین آنها روش‌های نمونه‌گیری است

تکنیک‌های نمونه‌گیری: از عمده‌ترین موضوع نمونه‌گیری انتخاب واقعی نمونه است. جلوگیری از جهت‌گیری در فرایند انتخاب و افزایش دقت در تمام مراحل ما را به انتخاب واقعی نزدیک می‌کند. اندازه نمونه یعنی تعداد نمونه‌های انتخابی یکی از پارامترهای مهم در روش‌های نمونه‌گیری است که متاسفانه راهکار قاطعی برای یافتن آن وجود ندارد

روش‌های نمونه‌گیری عبارتند از:

نمونه‌گیری تصادفی: از ساده‌ترین تکنیک‌های نمونه‌گیری به شمار می‌رود که در دو صورت با جایگزینی و بدون جایگزینی انجام می‌شود. در این گونه از روش‌ها هر یک از نمونه‌ها دارای شانس مساوی و مستقل برای انتخاب هستند. در نوع بدون جایگزینی هر نمونه فقط یک بار شانس انتخاب شدن را دارد در صورتی که در نمونه‌گیری تصادفی با جایگزینی هر نمونه بیش از یک بار می‌تواند انتخاب شود اغلب روش‌های آماری از نوع دوم یعنی با جایگزینی استفاده می‌کنند.

نمونه‌گیری منظم: نوع دیگری از نمونه‌گیری است که در آن تمام نمونه‌ها بدون هرگونه ترتیبی قرار گرفته و سپس با یک نظم معین نمونه مورد نظر انتخاب می‌شود.

نمونه‌گیری طبقه‌ای: از این روش هنگامی استفاده می‌شود ه مایلیم نمونه‌هایی از هر طبقه در نتیجه نهایی وجود داشته باشد.

نمونه‌گیری خوشه‌ای: نوع دیگری از روش‌های نمونه‌گیری است که در آن به جای نمونه گروه‌ها یا خوشه‌ها به صورت تصادفی انتخاب می‌شوند این روش هنگامی مناسب است که انتخاب نمونه مشکل یا غیر ممکن باشد. اگر عمل نمونه‌گیری با استفاده از خوشه فقط یک بار انجام شود به آن نمونه‌گیری خوشه تک مرحله‌ای گفته می‌شود چنانچه این عمل بیش از یک بار صورت پذیرد به نمونه‌گیری خوشه چند مرحله موسوم است

3-4-5) کاهش مقادیر یک صفت خاصه: مقادیر موجود برای یک صفت خاصه محدوده‌ای را تشکیل می‌دهند. بدون شک این مقادیر زیر مجموعه‌ای از دامنه صفت خاصه مزبور است.

تکنیک‌های مجزاسازی یا گسسته‌سازی می‌توانند برای کاهش مقادیر در پایگاه داده استفاده شود این تکنیک‌ها قرار دادن داده‌های پیوسته در تعدادی فاصله است. این تکنیک‌ها عموماً در توصیف ساده‌تر داده‌ها و فهم راحت‌تر و در نهایت نتایج پایانی فرایند داده کاوی تاثیر بسزایی دارند

گسسته‌سازی مبتنی بر آنتروپی: آنتروپی یکی از معیارهای رایج جهت گسسته‌سازی به شمار می‌رود گسسته‌سازی مبتنی بر آنتروپی یک تکنیک جداسازی بالا به پایین است بر روی داده‌های با برچسب کلاس انجام می‌شود این بدین معنی است که این روش با کمک اطلاعات توزیع کلاس نقاط انفصال را برای داده‌ها محاسبه می‌کند. نقاط انفصال آنهایی هستند که مقادیر داده‌ها بر مبنای آنها افزاز می‌شود

تهیه کننده: زهرا سعیدی، نجمه فاریابی، اسما نژاد سالاری