APJ. 4(3):51-59, Autumn 2023

Arman Process Journal (APJ)

Homepage: armanprocessjournal.ir

Data Mining, Applications, Requirements, Process and Tools

***P. Mohammadi*[*,1]**

[1] *Department of Computer Engineering, Science and Research Branch, Islamic Azad University, West Azerbaijan, Iran*

Received: 12 August 2023
Accepted: 23 November 2023

KEYWORDS:
Data mining,
Analysis and modeling,
distribution base,
neural network,
Data mining tools,

[1] Corresponding author
p.mohammadi.asl@gmail.com

ABSTRACT

The emergence of data mining science has made data become one of the most valuable assets of organizations and with the correct use of this trump card, software systems can produce results in a different and effective way. The process of extracting and discovering patterns and correlations from a large volume of raw data from one or more databases is called data mining. Data mining is an important and fundamental part in the analysis of distributed information of today's organizations. The data obtained from data mining can be used in business intelligence and advanced analysis. Increasing capacity, finding hidden patterns, trends and correlations in data sets is one of the main advantages of data mining tools. Due to the evolution of data storage technology and the growth of big data, the use of data mining techniques has increased dramatically in the last two decades. Using the best data mining tools helps businesses to make decisions and implement knowledge-based processes more efficiently by identifying hidden relationships and patterns in the data. Despite the technology constantly evolving to handle large-scale data, leaders still face challenges around scalability and automation. According to the importance of the topic in this article, we are going to examine the applications of data mining, requirements, process and important tools in this field. In the end, we will examine the technological perspective of data mining.

NUMBER OF REFERENCES	NUMBER OF FIGURES	NUMBER OF TABLES
23	0	0

نشریه تخصصی آرمان پردازش، دوره ۴، شماره ۳، پاییز ۱۴۰۲

فصلنامه تخصصی آرمان پردازش
(APJ)
Homepage: www.armanprocessjournal.ir

داده کاوی، کاربردها، نیازمندی ها، فرایند و ابزارها

پیام محمدی *[1]

[1] گروه مهندسی کامپیوتر، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، آذربایجان غربی، ایران

چکیده

ظهور علم داده کاوی باعث شده است که داده‌ها به یکی از سرمایه‌های بسیار ارزشمند سازمان‌ها تبدیل شوند و استفاده درست از این برگ برنده، سیستم های نرم افزاری بتوانند نتایج را به نحو متفاوت و موثری رقم بزنند. به فرایند اســـتخراج و کشـــف الگوها و همبســـتگی‌ها از میان حجم زیادی از داده‌های خام از یک یا چند بانک اطلاعاتی، داده کاوی می‌گویند. داده کاوی بخش مهم و اســـاســـی در تجزیه و تحلیل اطلاعات توزیعی ســـازمان های امروزی ا ست. داده های بد ست آمده از داده کاوی را می توان در هوش تجاری و تجزیه و تحلیل پی شرفته ا ستفاده کرد . افزایش ظرفیت، یافتن الگوها، روندها و همبستگی های پنهان در مجموعه داده ها، یکی از ا صلی ترین مزیت های ابزارهای داده کاوی اســت. با توجه به تکامل فناوری ذخیره‌ســازی داده‌ها و رشــد کلان داده‌ها، استفاده از تکنیک‌های داده‌کاوی طی دو دهه اخیر به طور چشم‌گیری افزایش یافته است. بهره گیری از ابزارهای برتر داده کاوی با مشخص کردن روابط و الگوهای پنهان در داده ها به کسب و کارها در تصمیم گیری و اجرای کارآمدتر فرایندهای دانش محور کمک می کنند. علی رغم این که این فناوری برای رســـیدگی به داده‌ها در مقیاس بزرگ به طور مداوم تکامل می‌یابد، رهبران هنوز در مورد مقیاس‌پذیری و اتوماسیون با چالش‌هایی روبرو هستند. بنا به اهمیت موضوع در این مقاله قصد داریم به بررسی کاربردهای داده کاوی، نیازمندی ها، فرایند و ابزارهای مهم این حوزه بپردازیم. در پایان نیز درخصوص دورنمای تکنولوژیکی داده کاوی بررسی می نمائیم.

واژگان کلیدی:
داده کاوی،
تحلیل و مدل سازی،
پایگاه توزیع،
شبکه عصبی،
ابزارهای داده کاوی،

تعداد جداول	تعداد شکل ها	تعداد مراجع
۰	۰	۲۳

مقدمه

سیر تحول زیرساخت های فناورمحور و ابزارهای دیجیتال موجب شده‌اند که در سال‌های اخیر، میزان تولید و ثبت داده‌ها به طرز چشمگیری افزایش پیدا کند. همین میسر شدن امکان ثبت و ذخیره‌سازی، افراد و کسب‌وکارها را قادر ساخته تا بتوانند اقدام به تحلیل داده‌ها و استخراج اطلاعاتی کنند که می‌تواند روند توسعه سازمان‌ها را به کلی متحول کند و به مانند چشم سومی برای مدیران، آن‌ها را در اتخاذ تصمیم‌های بهینه‌تر یاری کند. ظهور علم داده کاوی[1] باعث شده است که اکنون «داده‌ها» به یکی از سرمایه‌های بسیار ارزشمند سازمان‌ها تبدیل شوند و استفاده درست از این برگ برنده، بتواند نتایج را به نحو متفاوتی رقم بزند. به‌طوری‌که اکنون در شرکت‌های بزرگ و یا سطوح کلان اقتصادی، سیاسی و اجتماعی، بدون استناد به پژوهش‌های داده‌محور و تحلیل‌های داده‌ای از جوامع هدف، هیچ تصمیم و یا سیاستی اتخاذ نمی‌شود [۱]. به فرایند استخراج و کشف الگوها و همبستگی‌ها از میان حجم زیادی از داده‌های خام از یک یا چند بانک اطلاعاتی، دیتاماینینگ یا داده کاوی می‌گویند .به زبان ساده‌تر، داده کاوی به معنای کاوش و کشف اطلاعات مخفی یا الگوهایی است که به‌صورت طبیعی در داده‌ها وجود دارند، اما به‌راحتی توسط متخصصان اطلاعات شناسایی نمی‌شوند. عبارت داده کاوی که به عنوان KDD[2] هم شناخته می شود تا دهه ۱۹۹۰ ابداع نشده بود و پس از جمع آوری داده ها در مخازن داده، مفهوم داده کاوی به دنیا ارائه شد.دیتاماینینگ، بر پایه‌های سه رشته علمی، آمار ، هوش مصنوعی و یادگیری ماشین ساخته شده است. انبار داده فرآیند جمع آوری و مدیریت داده ها است. در این فرآیند داده ها از منابع مختلف در یک مخزن ذخیره می شود و به ویژه برای سیستم های مدیریت ارتباط با مشتری مفید هستند. این فرآیند قبل از داده کاوی اتفاق می افتد [۳-۲].

ابزارها و فناوری داده کاوی برای تحلیلBig Data ، دائما در حال تکامل هستند. پیشرفت تکنولوژی به تجزیه و تحلیل سریع‌تر و آسان‌تر داده ها کمک کرده است. امروزه هرچه مجموعه داده‌ها بزرگ‌تر و پیچیده‌تر باشند، شانس بیشتری برای یافتن موارد مرتبط وجود دارد. با شناسایی و بررسی داده های معنادار، سازمان ها می توانند از اطلاعات ارزشمند آن برای تصمیم گیری و رسیدن به اهداف جدیدی استفاده کنند. فرآیند داده کاوی ممکن است با توجه به رویکرد هر کسب و کار به چندین مرحله تقسیم شود، اما به طور کلی شامل پنج مرحله زیر است [۴].

- شناسایی الزامات کسب و کار با توجه به اهداف و سیاست های مدیران.
- شناسایی منابع داده و بررسی اینکه کدام بخش از داده باید تجزیه و تحلیل شود.
- انتخاب ابزار برتر داده کاوی و متد های مدل سازی.
- ارزیابی مدل جهت اطمینان از برآورده کردن نیازمندی های سازمان و یا کسب و کار.
- تهیه گزارشی برای ارائه نتایج داده کاوی.

کاربردها

داده کاوی بخش مهم و اساسی در تجزیه و تحلیل اطلاعات توزیعی سازمان های امروزی است. داده های بدست آمده از داده کاوی را می توان در هوش تجاری و تجزیه و تحلیل پیشرفته استفاده کرد. افزایش ظرفیت، یافتن الگوها، روندها و همبستگی های پنهان در مجموعه داده ها، یکی از اصلی ترین مزیت های بهره گیری از ابزارهای داده کاوی است. از ترکیبی از تجزیه و تحلیل داده های سنتی و تجزیه و تحلیل فرآیند داده کاوی، می توان برای تصمیم گیری و برنامه ریزی استراتژیک شرکت استفاده کرد. داده کاوی برای تجزیه و تحلیل داده ها و هوش تجاری مفید است تا به کسب و کارها کمک کند تا دانش عمیق تری نسبت به سازمان، مشتریان، رقبا و صنعت خود کسب کنند. برخی از مهم ترین کاربردهای داده کاوی عبارتند از[۷-۵]:

فروش و بازاریابی

شرکت‌ها حجم عظیمی از داده‌ها درمورد مشتریان خود را جمع‌آوری می‌کنند که می‌توانند با بررسی دموگرافیک و رفتار کاربران آنلاین، از داده‌ها برای بهینه‌سازی کمپین‌های بازاریابی خود، بهبود پیشنهادهای متقابل فروش و برنامه‌های وفاداری مشتری استفاده کنند و ROI بالاتری را در بازاریابی به دست بیاورند.

آموزش

امروزه مؤسسات آموزشی هم شروع به جمع‌آوری داده‌ها برای درک دانش‌آموزان خود و هم‌چنین بررسی محیط‌هایی که برای موفقیت آن‌ها مناسب است کرده‌اند. حوزه‌ی جدیدی به‌نام داده‌کاوی آموزشی[3] در حال ظهور است که به توسعه‌ی روش‌هایی مربوط است که دانش را از داده‌های موجود در محیط‌های آموزشی استخراج می‌کند. اهداف EDM پیش‌بینی نحوه‌ی یادگیری دانش‌آموزان در آینده، مطالعه اثرات پشتیبانی تحصیلی و ارتقای دانش علمی درمورد نحوه‌ی یادگیری است. داده‌کاوی می‌تواند در مؤسسات برای تصمیم‌گیری دقیق و هم‌چنین پیش‌بینی نتایج تحصیلی دانش‌آموزان استفاده شود. با نتایج به‌دست‌آمده، مؤسسات می‌توانند بر آنچه باید آموزش دهند و چگونگی آموزش آن تمرکز کنند [۸].

پزشکی

داده‌کاوی به پزشکان کمک می‌کند با جمع‌آوری سابقه‌ی پزشکی هر بیمار، نتایج معاینه‌ی فیزیکی، داروها و الگوهای درمانی، تشخیص‌های

[1] Data Mining
[2] Knowledge Discovery in Databases (KDD)
[3] Educational Data Mining

دقیق‌تری بدهند؛ هم‌چنین داده‌کاوی به ایجاد استراتژی‌های مدیریت منابع پزشکی مقرون‌به‌صرفه‌تر کمک بزرگی می‌کند.

تشخیص کلاه‌برداری

تا کنون میلیاردها دلار به‌دلیل کلاه‌برداری از دست رفته است. روش‌های سنتی کشف کلاه‌برداری زمان‌بر و پیچیده هستند. داده‌کاوی به ارائه الگوهای معنادار و تبدیل داده‌ها به اطلاعات کمک می‌کند. یک سیستم تشخیص کلاه‌برداری کامل باید از اطلاعات همه‌ی کاربران محافظت کند. یک روش برای ایجاد چنین سیستمی یادگیری با ناظر[1] است. این روش جمع‌آوری نمونه‌های قبلی را دربرمی‌گیرد که به دو دسته‌ی کلاه‌برداری یا غیرکلاه‌برداری طبقه‌بندی می‌شوند. در این روش الگویی با استفاده از این داده‌ها ساخته می‌شود تا تشخیص دهد نمونه کلاه‌برداری است یا خیر. این قضیه در بانک‌ها و دیگر مؤسسات مالی بسیار استفاده می‌شود و شرکت‌های مستقر در SaaS نیز برای حذف حساب‌های کاربران جعلی از مجموعه‌داده‌های خود، اقدام به اتخاذ این روش‌ها کرده‌اند. سایر کاربردهای رایج و مهم داده کاوی به شرح زیر می باشند [۱۱-۹و۵]:

- تولید محصولات صنعتی و تجاری
- یافتن جامعه هدف برای سازمان ها وکسب و کارها
- کشف الگوهای رفتاری
- پیش‌بینی فروش در فرایندهای تجاری
- دسته بندی آیتم ها بر اساس تفاوت‌های موجود
- پیش‌بینی الگوهای موردنیاز در حوزه بانکداری
- تجمیع و تمرکز بر روی داده‌های بزرگ

به‌طور کلی فرایند داده‌کاوی علاوه‌بر اینکه به سازمان کمک می‌کند داده‌های نامرتبط و بلااستفاده را از مجموعه‌ی خود حذف نماید، از طرفی اطلاعات بسیار مفید و کاربردی را در اختیار سازمان قرار می‌دهد و همچنین به فرایندهای تصمیم‌گیری سرعت می‌بخشد. در این راستا، بکارگیری صحیح نیازمند ها و فرایندهای اجرائی، چرخه حیات داده کاوی را تسهیل و بهینه سازی می نمایند. لذا در قسمت بعدی مقاله حاضر به بررسی این موارد می پردازیم.

نیازمندی ها و فرایند

یکی از اهداف بنیادین داده کاوی تولید دانش جدیدی است که کاربران بتوانند بر اساس آن به اهداف دانش محور دست یابند. این کار بوسیله ساختن مدلی از جهان واقعی بر پایه داده هایی که از منابع گوناگون بدست می آید صورت می گیرد. نتیجه مدل سازی یک سری توضیحات در مورد الگوها و ارتباطات داده ای که می تواند براساس فرایندی به صورت مطمئنی جهت پیش بینی آینده مورد استفاده قرار گیرد. بمنظور پرهیز از اغتشاش در مراحل مختلف داده کاوی ایجاد فرایند و تصویری از سلسله مراتبی از انتخابات و تصمیم ها که نیاز مند آن می باشیم ضروری می باشد. فرآیند داده کاوی، مجموعه‌ای از گام‌ها و اقداماتی است که لازم است از زمان آغاز کار و جمع آوری داده تا استخراج اطلاعات و دانش کاربردی از آن انجام شود. عموما فرآیند داده کاوی بر روی حجم عظیمی از داده‌ها اعمال می‌شود و چون این کار از توان انسان خارج است، از فناوری‌های خاصی برای این کار استفاده می‌شود. اساسا دو نکته برای موفق بودن یک فرایند داده کاوی وجود دارد. اول اینکه یک فرموله سازی دقیق از مساله ای است که باید حل شود. دومین نکته استفاده از داده صحیح است. پس از انتخاب داده ای که در دسترس است یا شاید خرید داده خارجی ممکن است نیازمند بکارگیری روشهای انتقال داده یا دسته بندی باشیم. مهم ترین گامها و الزامات درجهت ایجاد چنین فرایند جامعی به شرح زیر می باشد [۱۴-۱۲و۹] :

سیاست گذاری استراتژیک

در گام آغازین لازم است هدف نهایی از جستجوی داده را بصورت شفاف مشخص نمائیم. 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.

تحلیل 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 یابد.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. 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. برخی از الگوریتمها یک پایگاه داده از قوانین، فاکتورهای ایمن و فراهم آوردن امکان جستجو را ایجاد می نمایند.

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

[1] Supervised Learning

فهم داده موجود و هم برای پیش بینی اینکه هر نمونه جدید چگونه کار می کند استفاده شود. داده کاوی مدلهای طبقه بندی را بوسیله امتحان کردن داده طبقه بندی شده(موارد) و نهایتا 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.

انتخاب نوع پیش بینی و حدس بازگشتی

گام بعدی تصمیم در مورد انتخاب نوعی پیش بینی که از همه مناسب تر است می باشد. دراین خصوص راهکارهای اصلی طبقه بندی و حدس زدن و سیاست گذاری می باشند. 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 و هم برای رده بندی و 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 [۱۵].

گزینش نوع مدل انتخابی

ام بعدی گزینش نوع مدل می باشد که عبارت است از ایجاد یک رویکرد مانند سری های زمانی یا شبکه های عصبی برای پیاده سازی 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 دارند. 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.

شبکه های عصبی به طور خاصی مورد استفاده اند چرا که آنها ابزاری موثر برای مدلسازی مسائل بزرگ و پیچیده که ممکن است در آنها صدها متغیر پیش بینی کننده که فعل و انفعالات زیادی دارند وجود داشته باشد. شبکه های عصبی زیستی بطور غیر قابل مقایسه ای پیچیده تر هستند. شبکه های عصبی می توانند در مسائل طبقه بندی یا حدسهای بازگشتی (که در آنها متغیر خروجی پیوسته است) استفاده شوند. یک شبکه عصبی با یک لایه داخلی شروع می شود که در آن هر گره به یک متغیر پیشگو منسوب می گردد. این گره های ورودی به یک تعداد از گره ها در لایه پنهان متصل می شوند. گره ها در لایه پنهان می توانند به گره هایی در یک لایه پنهان دیگر یا به یک لایه خروجی متصل شود. لایه خروجی خود شامل یک یا بیشتر متغیرهای جواب می باشد. درخت های انتخاب نیز راهی برای نمایش یک سری از قوانین که به یک کلاس یا مقدار منجر می شود می باشند. 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. استنتاج قانون نیز 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 وضعیت ها و حالتهای ممکن را پوشش ندهند [۱۶].

ساخت مدل براساس الگوریتم مناسب

در ادامه الگوریتمهای زیادی برای ساخت مدلها در دسترس هستند. 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.

ساخت پایگاه داده داده کاوی

این گام به همراه دو گام بعدی هسته آماده سازی و ذخیره و بازیابی داده را تشکیل می دهند. 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. امروزه کاربران 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 [۱۷]:

- جمع آوری داده ها
- توضیح داده ها
- انتخاب داده ها
- تعیین کیفیت داده ها و پاک کردن آن
- تثبیت و یکپارچگی
- ساختن فوق داده (داده هایی که خود بیانگر توضیحی در مورد داده های موجود می باشند.)
- بارکردن پایگاه داده مربوط به داده کاوی
- نگهداری پایگاه داده مربوط به داده کاوی
- فعالیت های فوق ممکن است لزوما به ترتیب ذکر شده انجام نشوند.
- درگام بعد باید از تکنیکهای جستجو در جهت کشف و بازیابی صحیح داده ها از پایگاه های داده استفاده نمود.

مدل سازی داده کاوی

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

- انتخاب متغیرها
- انتخاب سطرها
- ساختن متغیرهای جدید
- تغییر شکل متغیرها

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. 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.

ارزیابی و تفسیر

بعد از ساخت یک مدل شما باید نتایج آن را ارزیابی نموده و همچنین اهمیت آن را نیز توضیح دهید. 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 در این قسمت توضیح داده شد برای انجام هر فرآیند داده کاوی لازم و ضروری به نظر می رسند. بهره گیری از ابزارهای مناسب در فرایند داده کاوی جریان های کاری را تسهیل و بهینه سازی می نماید. در قسمت بعدی از مقاله حاضر به بررسی ابزارهای محبوب و رایج حوزه داده کاوی می پردازیم.

ابزارهای داده کاوی

ابزارهای برتر داده کاوی شامل تکنیک هایی مکانیزه برای فرآیند یافتن بهینه الگوها و روابط در مقادیر زیاد می باشند. این ابزارها به کسب‌وکارها کمک می‌کنند تا درباره نیازهای مشتریان، افزایش درآمد، کاهش هزینه‌ها، بهبود روابط با مشتری و موارد دیگر اطلاعات بیشتری کسب کنند، به همین علت، انتخاب این ابزار ها از اهمیت زیادی برخوردار است. با وجود تعداد زیاد ابزارهای رایگان، یکی از سخت ترین کارها در کل فرآیند داده کاوی، انتخاب ابزار مناسب است. ابزارهای منبع باز، گزینه های خوبی برای شروع هستند،چون دائماً به روز می شوند. مهمترین ویژگی هایی که باید در هنگام انتخاب ابزار های داده کاوی به آن توجه نمود عبارتند از [۱۹-۱۸]:

متن باز بودن یا نبودن

بیشتر ابزار برتر داده کاوی متن باز، هستند اما گاهی اوقات تفاوت های کمی با هم دارند .

امکان یکپارچه سازی داده ها

برخی از ابزارهای داده کاوی با مجموعه داده های بزرگ بهتر کار می کنند، در حالی که برخی دیگر داده های کوچکتر بهتر کار می کنند. وقتی گزینه های ابزار های داده کاوی را بررسی می کنید، انواع داده هایی که بیشتر با آنها سر و کار دارید را در نظر بگیرید.

کاربردی بودن و قابلیت استفاده

هر ابزار داده کاوی، یک رابط کاربری دارد که تعامل با محیط کار و تعامل با داده ها را آسان تر می کند. بعضی از ابزارهای داده کاوی، ماهیت آموزشی دارند در حالی که برخی دیگر، بر اساس نیازهای شرکت ها طراحی شده اند.

زبان برنامه نویسی

اکثر ابزارهای متن باز داده کاوی، به زبان جاوا توسعه یافته اند؛ ولی بسیاری از آنها از اسکریپت های R و Python هم، پشتیبانی می کنند.

ابزارهای داده کاوی به سازمان ها و کسب‌وکارها کمک می‌کنند تا درباره نیازمندی ها، اهداف و گامهای پیاده سازی داده کاویموثرتر عمل نمایند. به همین علت، انتخاب ابزار مناسب داده کاوی از اهمیت ویژه ای در فرایند داده کاوی برخوردار است. در ادامه برخی از ابزارهای محبوب داده کاوی و کارکردهای آن ها را معرفی و مقایسه می نمائیم [۲۲-۲۰]:

RapidMiner

RapidMinerیک پلت فرم رایگان و متن باز داده کاوی است که توسط شرکت RapidMiner توسعه یافته است. رپیدماینر دارای صدها الگوریتم برای آماده سازی داده ها، یادگیری ماشین، یادگیری عمیق، متن کاوی و تجزیه و تحلیل پیش بینی است. این ابزار برتر داده کاوی، با استفاده از زبان برنامه نویسی جاوا توسعه یافته.

Oracle Data Mining

Oracle Data Miningیکی از اجزای Oracle Advanced Analytics است که به تحلیلگران داده این امکان را می دهد که مدل های مورد نظر خود را بسازند. این ابزار برتر داده کاوی، شامل چندین الگوریتم برای کارهایی مانند طبقه بندی، رگرسیون، تشخیص ناهنجاری، پیش بینی و غیره است

.

IBM SPSS Modeler

ابزار IBM SPSS Modelerیکی از محبوب‌ترین و قدرتمندترین ابزارهای داده‌کاوی و تحلیل پیشرفته داده‌ها است. این ابزار توسط شرکت IBM توسعه داده شده و به تحلیل و پیش‌بینی داده‌ها در حوزه‌های مختلف کمک می‌کند. SPSS Modeler به کمک رابط کاربری گرافیکی جذاب و بدون نیاز به دانش تخصصی برنامه نویسی، افراد مختلف مانند تحلیلگران داده، مهندسین و محققین را قادر به انجام فرایند داده کاوی و تحلیل‌های پیشرفته بر روی داده‌ها می کند. از مزایای این نرم افزار می توان به موارد زیر اشاره نمود [۲۳]:

- داشتن روش های بسیار متنوع برای تحلیل داده ها
- سرعت بسیار بالا در انجام محاسبات و استفاده از اطلاعات پایگاه داده ها
- داشتن محیط گرافیکی به منظور راحتی بیشتر کاربر برای انجام کارهای تحلیلی

در نسخه جدید امکان پاک سازی و آماده سازی داده ها به صورت کاملاً اتوماتیک انجام می شود. این نرم افزار تمامی نرم افزارهای پایگاه داده معروف مانند Microsoft Office و SQL و ... را پشتیبانی می کند .

ماژول های موجود در این نرم افزار عبارتند از:

- PASW Association
- PASW Classification
- PASW Segmentation
- PASW Modeler Solution Publisher

این نرم افزار هم بر روی کامپیوتر شخصی و هم بر روی سرور قابل نصب است و از Windows های ۳۲ و ۶۴ بیتی نیز پشتیبانی می کند.

Weka

Wekaیکی از نرم‌افزارهای منبع باز و قدرتمند برای داده‌کاوی و یادگیری ماشین است. این ابزار توسط دانشگاه وایکاتو در نیوزیلند توسعه داده شده و به افراد با سطوح تخصصی مختلف اجازه می‌دهد تا تحلیل داده‌های خود را انجام دهند. ابزار Weka از رابط کاربری گرافیکی قدرتمندی برخوردار است که به کاربران امکان انجام تحلیل‌های پیشرفته بر روی داده‌ها را بدون نیاز به دانش برنامه‌نویسی می‌دهد. این محیط، شامل روش‌هایی برای همه مسایل استاندارد داده کاوی مانند رگرسیون، رده‌بندی، خوشه‌بندی، کاوش قواعد انجمنی و انتخاب ویژگی می‌باشد. با در نظر گرفتن اینکه، داده‌ها بخش مکمل کار هستند، بسیاری از ابزارهای پیشپردازش داده‌ها و مصورسازی آنها فراهم گشته است. همه الگوریتم‌ها، ورودی‌های خودرا به صورت یک جدول رابطه-ای به فرمت ARFF دریافت می‌کنند. این فرمت داده‌ها،می‌تواند از یک فایل خوانده شده یا به وسیله یک درخواست از پایگاه داده‌ای تولیدگردد.

Knime

پلتفرم KNIME، یک ابزار قدرتمند و منبع باز برای داده‌کاوی، تجزیه و تحلیل داده های پیشرفته است. این ابزار توسط تیم KNIME توسعه داده شده و به کاربران اجازه می‌دهد تا به کمک یک رابط کاربری گرافیکی، فرآیندهای پیچیده تحلیلی را بدون نیاز به مهارت‌های برنامه‌نویسی انجام دهند.

H2O

H2Oیک پلت فرم یادگیری ماشین متن باز است که هدف آن دسترسی همه ی افراد به فناوری هوش مصنوعی است. این ابزار برتر داده کاوی، از متداول‌ترین الگوریتم‌های ML پشتیبانی می‌کند و به کاربران کمک کند تا مدل‌های یادگیری ماشین را به روشی سریع و ساده بسازند، حتی اگر متخصص نباشند.

Orange

ابزار داده‌کاوی اورنج، یک نرم‌افزار متن باز و قدرتمند برای تجزیه و تحلیل داده و ایجاد مدل‌های یادگیری ماشینی است. این ابزار توسط دانشگاه لیوبلیانا در اسلوونی توسعه داده شده است و به کاربران اجازه می‌دهد با استفاده از رابط کاربری گرافیکی ساده و آسان، تحلیل داده‌های خود را انجام دهند.

Apache Mahout

Apache Mahoutیک پلت فرم متن باز و یک ابزار برتر داده کاوی برای ایجاد برنامه های کاربردی مقیاس پذیر با استفاده از یادگیری ماشین است .

SAS Enterprise Miner

از سیستم داده کاوی SAS برای بهینه سازی و داده کاوی استفاده می شود. روش‌ها و رویه‌های مختلفی را برای اجرای قابلیت‌های تحلیلی مختلف ارائه می‌کند که خواسته‌ها و اهداف سازمان را ارزیابی می‌کند. این نرم افزار شامل مدل‌سازی توصیفی، مدل‌سازی پیش‌بینی‌کننده و مدل‌سازی تجویزی[1] است. ابزار داده کاوی SAS به دلیل طراحی و پردازش حافظه توزیع شده، بسیار مقیاس پذیر است.

نتیجه‌گیری و راهکارهای آتی

امروزه متخصصین علم داده برای سازمان های سراسر جهان حیاتی شده اند، زیرا سازمان ها بیش از هر زمان دیگری به دنبال دستیابی به اهداف بزرگ تر با علم داده هستند. با توجه به تکامل فناوری ذخیره‌سازی داده‌ها و رشد کلان داده‌ها، استفاده از تکنیک‌های داده‌کاوی طی دو دهه اخیر به طور چشم‌گیری افزایش یافته است. هدف داده کاوی تبدیل داده‌های خام سازمان‌ها به دانش مفید است. علی رغم این که این فناوری برای رسیدگی به داده‌ها در مقیاس بزرگ به طور مداوم تکامل می‌یابد، رهبران هنوز در مورد مقیاس‌پذیری و اتوماسیون با چالش‌هایی روبرو هستند. فرایند داده کاوی درجهت رفع چالش های پیاده سازی داده کاوی شامل چندین مرحله از جمع آوری داده‌ها تا مصورسازی اطلاعات ارزشمند از مجموعه‌ی کلان داده‌ها با استفاده از مدلسازی فرایند است. همچنین ابزارهای برتر داده کاوی با مشخص کردن روابط و الگوهای پنهان در داده ها به کسب و کارها در تصمیم گیری بهتر کمک می کنند. بنا به اهمیت موضوع در این مقاله به بررسی کاربردهای داده کاوی، نیازمندی ها، فرایند و ابزارهای مهم این حوزه پرداختیم. در پایان لازم به ذکر است، آینده برای داده کاوی و علم داده بسیار روشن است. بزودی مدیران و کاربران سازمان‌ها در جهان داده های دیجیتال با انباشتی از داده ها رو به رو خواهند شد که مدیریت این داده‌ها برای سازمان بسیار حیاتی است که توسط علم داده امکان‌پذیر خواهد بود. به دلیل پیشرفت در فناوری، فناوری‌های استخراج اطلاعات ارزشمند از داده ها بسیار پیشرفت خواهد کرد. تا چند دهه قبل فقط سازمان‌هایی مانند ناسا می‌توانستند از ابر رایانه های خود برای تجزیه و تحلیل داده ها استفاده کنند هزینه ذخیره سازی و محاسبه داده ها بسیار زیاد بوده و در توان کمپانی‌های کوچک نبود اما اکنون، شرکت‌ها انواع کارهای مختلف را با یادگیری ماشین، هوش مصنوعی و یادگیری عمیق با هزینه های مناسب تر و حجم انبوهی از داده‌ها در کوتاه‌ترین زمان ممکن انجام می‌دهند و این خود کمک بزرگی به تمامی کسب و کارهای کوچک و بزرگ در راستای رسیدن به موفقیت خواهد بود.

تعارض منافع

«هیچ‌گونه تعارض منافع توسط نویسندگان بیان نشده است»

منابع و مآخذ

[1] Gupta MK, Chandra P. A comprehensive survey of data mining. International Journal of Information Technology. 2020 Dec;12(4):1243-57.

[2] Oweis NE, Owais SS, George W, Suliman MG, Snášel V. A survey on big data, mining:(tools, techniques, applications and notable uses). InIntelligent Data Analysis and Applications: Proceedings of the Second Euro-China Conference on Intelligent Data Analysis and Applications, ECC 2015 2015 (pp. 109-119). Springer International Publishing.

[3] Roy U, Zhu B, Li Y, Zhang H, Yaman O. Mining big data in manufacturing: requirement analysis, tools and techniques. InASME International Mechanical Engineering Congress and Exposition 2014 Nov 14 (Vol. 46606, p. V011T14A047). American Society of Mechanical Engineers.

[4] Mughal MJ. Data mining: Web data mining techniques, tools and algorithms: An overview. International Journal of Advanced Computer Science and Applications. 2018;9(6).

[5] Padhy N, Mishra DP, Panigrahi R. The survey of data mining applications and feature scope. arXiv preprint arXiv:1211.5723. 2012 Nov 25.

[6] Mariscal G, Marban O, Fernandez C. A survey of data mining and knowledge discovery process models and methodologies. The Knowledge Engineering Review. 2010 Jun;25(2):137-66.

[7] Bartschat A, Reischl M, Mikut R. Data mining tools. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2019 Jul;9(4):e1309.

[1] Prescriptive

[8] Vikram K, Siddipet MD, Upadhayaya N. Data mining tools and techniques: a review. Logistics management. 2011;2(8).

[9] Goebel M, Gruenwald L. A survey of data mining and knowledge discovery software tools. ACM SIGKDD explorations newsletter. 1999 Jun 1;1(1):20-33.

[10] Prakash BA, Ashoka DV, Aradhya VM. Application of data mining techniques for software reuse process. Procedia Technology. 2012 Jan 1;4:384-9.

[11] Halkidi M, Spinellis D, Tsatsaronis G, Vazirgiannis M. Data mining in software engineering. Intelligent Data Analysis. 2011 Jan 1;15(3):413-41.

[12] Rohanizadeh SS, BAMENI MM. A proposed data mining methodology and its application to industrial procedures.

[13] Gonzalez R, Kamrani A. A survey of methodologies and techniques for data mining and intelligent data discovery. InData Mining for Design and Manufacturing: Methods and Applications 2001 Oct 31 (pp. 41-59). Boston, MA: Springer US.

[14] Jackson J. Data mining; a conceptual overview. Communications of the Association for Information Systems. 2002;8(1):19.

[15] Madni HA, Anwar Z, Shah MA. Data mining techniques and applications—A decade review. In2017 23rd international conference on automation and computing (ICAC) 2017 Sep 7 (pp. 1-7). IEEE.

[16] PhridviRaj MS, GuruRao CV. Data mining–past, present and future–a typical survey on data streams. Procedia Technology. 2014 Jan 1;12:255-63.

[17] Turner CJ, Tiwari A, Olaiya R, Xu Y. Process mining: from theory to practice. Business Process Management Journal. 2012 Jun 1;18(3):493-512.

[18] Santos-Pereira J, Gruenwald L, Bernardino J. Top data mining tools for the healthcare industry. Journal of King Saud University-Computer and Information Sciences. 2022 Sep 1;34(8):4968-82.

[19] Al-Odan HA, Al-Daraiseh AA. Open source data mining tools. In2015 International Conference on Electrical and Information Technologies (ICEIT) 2015 Mar 25 (pp. 369-374). IEEE.

[20] Malkawi R, Saifan AA, Alhendawi N, BaniIsmaeel A. Data mining tools evaluation based on their quality attributes. International Journal of Advanced Science and Technology. 2020 Mar;29(3):13867-90.

[21] Kadaru BB, UmaMaheswararao M. An overview of general data mining tools. Int Res J Eng Technol. 2017 Sep;4(9):930-6.

[22] Mikut R, Reischl M. Data mining tools. Wiley interdisciplinary reviews: data mining and knowledge discovery. 2011 Sep;1(5):431-43.

[23] Wendler T, Gröttrup S. Data mining with SPSS modeler: theory, exercises and solutions. Springer; 2016 Jun 6.

[24]