آشنایی با مفهوم DATA MINING و بررسی کاربرد های آن

سارا شهسواری

گروه مهندسی نرم افزار کامپیوتر،واحد یادگارامام خمینی(ره) شهرری ،دانشگاه آزاد اسلامی، تهران، ایران

چکیده:

علم داده کاوی (Data Mining) شناختهی knowledge Discovery in Databases است، که به عنوان یک فرایند کشف دانش، شامل: پاکسازی داده ها، انتخاب داده ها، یکپارچه سازی داده ها، تبدیل داده ها، کشف الگو، ارزیابی الگو و ارائه دانش است. یکی از علت هایی که این علم مورد توجه قرار گرفته است، ضریب اطمینان بالای تصمیمات اتخاذ شده بر اساس تحلیل های داده ای و نتایجی است که ایجاد می شود. داده کاوی در زمینه های مختلفی می تواند کاربرد داشته باشد از جمله: سلامت عمومی، پزشکی، تحقیقات جرم شناسی و غیره. جالب است بدانید؛ آنالیز داده ها باعث صرفه جویی در هزینه ها و بهبود کسب وکار می شود و می تواند مشتریان جدید و جریان های درامدی را شناسایی کند. هدف از این پژوهش این است که به این موضوع پی ببریم؛ چرا داده کاوی تقاضای زیادی دارد و چگونه بخشی از تکامل طبیعی فناوری اطلاعات است. در گذشته برای تجزیه و تحلیل مجموعه داده های بزرگ مدت زمان زیادی صرف می شد، اما هم اکنون یکی از اصلی ترین مزایای Data mining سرعت است.

واژگان کلیدی : داده کاوی، کشف دانش، آنالیز داده ها، پاکسازی داده ها، تبدیل داده ها

مقدمه

سیر تحول زیر ساخت های فناور محور و ابزار های دیجیتال؛ موجب شده اند که در سال های اخیر، میزان تولید و ثبت داده ها به طرز چشمگیری افزایش پیدا کند. همین میسر شدن امکان ثبت و ذخیره سازی، افراد و کسب و کار هارا قادر ساخته تا بتوانند اقدام به تحلیل داده ها و استخراج اطلاعاتی کنند که می تواند روند توسعه سازمانها را به کل متحول کند Gupta). (and Chandra, 2020 داده کاوی فرایندی است که به جستجوی الگو های معنادار توسط الگوریتم ها و روش ها پرداخته و به عنوان یک ابزار تحلیلی با قابلیت جستجوی پیچیده داده ای؛ دانشی را که در انبار های داده مدفون شده اند و علم آمار ناتوان از تحلیل آنهاست، یافته و استخراج می کند. می توان گفت پل ارتباطی میان علم آمار، علم کامپیوتر، هوش مصنوعی، الگوشناسی، فراگیری ماشین و بازنمایی بصری داده است(جماعت و عسگری، ۱۳۸۹). ظهور علم داده کاوی باعث شده است که هم اکنون "داده ها" به یکی از سرمایه های بسیار ارزشمند سازمانها تبدیل شوند و استفاده درست از این برگ برنده می تواند نتایج را به نحو متفاوتی رقم بزند. به طوری که اکنون در شرکت های بزرگ و یا سطوح کلان اقتصادی، سیاسی و اجتماعی، بدون استنادیه پژوهش های داده محور و تحلیل های داده ای از جوامع هدف، هیچ تصمیم و یا سیاستی اتخاذ نمی شود (Gupta MK Chandra, 2020). عبارت داده کاوی که به عنوان knowledge Discovery in Databases (KDD)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؛ در اصل انبارداده فرایند جمع آوری و مدیریت داده ها است. در این فرایند؛ داده ها از منابع مختلف در یک مخزن ذخیره می شوند و به ویژه برای سیستم های مدیریت ارتباط با مشتری مفید هستند؛ که این فرایند قبل از داده کاوی اتفاق می افتد (Oweis et al, 2015)(Roy et al, 2014). در اصل این فناوری موجب انقلابی در عینیت بخشیدن به مفاهیم مدیریت در کسب و کارهای بزرگ می شود و مهم ترین کاربرد آن در تلاش هایی است که برای استنتاج قواعد وابستگی از داده های تراکنشی انجام می شود (Hong et al, 2014).

بیان مسئله

به طور کلی؛ استفاده همگانی از وب و اینترنت، به عنوان یک سیستم اطلاع رسانی جهانی مارا مواجه با حجم زیادی از داده و اطلاعات می کند. این رشد انفجاری در داده های ذخیره شده، نیاز مبرم وجود تکنولوژی های جدید و ابزار های خودکاری را ایجاد کرده که به صورت هوشمند به انسان یاری رسانند؛ تا این حجم زیاد داده را به اطلاعات و دانش تبدیل کند (Han and Kamber,2001). 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 الگو ها، روند ها و همبستگی های پنهان در مجموعه داده ها، یکی از اصلی ترین مزیت های بهره گیری از ابزار های داده کاوی است. داده کاوی به کسب و کار ها کمک میکند تا دانش عمیق تری نسبت به؛ سازمان، مشتریان، رقبا و صنعت خود کسب کنند (Bartschat et al, 2019)(Padhy et al, 2012).

تفاوت داده و اطلاعات

بسیاری از مردم به اشتباه می پندارند که داده با اطلاعات تفاوتی ندارد و عموما؛ این دو واژه را به جای هم استفاده می کنند. داده (Data)؛ می تواند هر نوع از کاراکتر شامل متن، عدد، کلمه، صدا و تصویر باشد و در صورتی که توسط انسان مشاهده شود لزوما معنای خاصی هم در بر نخواهد داشت. داده ها عموما خام ، دسته بندی و طبقه بندی نشده هستند و در صورتی که

بخواهیم از آنها به صورت مستقیم استفاده کنیم عموما بی فایده خواهد بود. حال آنکه پس از طبقه بندی، دسته بندی و ساختاردهی به داده ها، اطلاعات (Information) به وجود می آید. می توان از داده ها برای تصمیم گیری و یا ایجاد دانش درمورد یک مقوله استفاده کرد. اطلاعات عموما برای کاربر مفهوم دارد و قابل استفاده است. در ادامه مثال هایی از داده و اطلاعات را مشاهده میکنید (احمدی، 1397):

تاریخ دمای سراسر جهان در صد سال گذشته داده است، حال آنکه روند افزایش یا کاهش دما در این سال ها یک اطلاعات است.

نمرات دانشجویان یک کلاس داده است، اما ترتیب نمرات، میانگین و ارتباط نمره این درس با معدل دانشجو یا یک درس دیگر یک نمونه از اطلاعات است.

مدیریت ذخیره سازی و دستیابی به اطلاعات

به طور کلی؛ فرایند داده کاوی علاوه بر اینکه به سازمان کمک می کند داده های نامرتبط و بلا استفاده را از مجموعه ی خود حذف نماید، از طرفی اطلاعات بسیار مفید و کاربردی را در اختیار سازمان قرار می دهد و همچنین به فرایند های تصمیم گیری سرعت می بخشد. در این راستا، بکارگیری صحیح نیازمند ها و فرایند های اجرائی، چرخه حیات داده کاوی را تسهیل و بهینه سازی می نمایند (Halkidi et al, 2011) (Goebel and Gruenwald, 1999) (Padhy et al, 2012). داده های اطلاعاتی به عنوان یکی از منابع حیاتی سازمان شناخته می شود و بسیاری از سازمان ها با اطلاعات و دانش سازمانی خود مانند سایر دارایی های ارزشمندشان برخورد می کنند.

داده (Data) به اطلاعات خام سازمان اتلاق می شود و اطلاعات (Information) به داده های پردازش شده. همچنین داده های پردازش شده پس از طبقه بندی و آنالیز به دانش سازمان (Knowledge) تبدیل می گردند؛ حال تصور نمائید، دسترسی به اطلاعات در شرایطی که داده ها به روش نامناسبی نگهداری شوند و یا روش ضابطه مندی جهت دستیابی به آنها وجود نداشته باشد تا چه حد مشکل است. برای رسیدن به یک سیستم اطلاعاتی مناسب، داده ها می بایست به صورتی منطقی طبقه بندی و ذخیره شوند تا استفاده از آنها ساده تر بوده، با کارایی بیشتری تحلیل شوند و سریعتر مورد استفاده قرار گیرند و در نتیجه مدیریت بهتری بر آنها اعمال شود (احمدی، 1397).

ابزار های داده کاوی

ابزار های برتر داده کاوی شامل تکنیک هایی مکانیزه برای فرایند یافتن بهینه الگو ها و روابط در مقادیر زیاد می باشند. این ابزار ها به کسب و کار ها کمک می کنند تا درباره نیاز های مشتریان، افزایش درآمد، کاهش هزینه ها، 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 از ابزار های رایگان، یکی از سخت ترین کار ها در کل فرایند داده کاوی، انتخاب ابزار مناسب است. ابزار های منبع باز، گزینه های خوبی برای شروع هستند، چون دائما بروز می شوند. مهم ترین ویژگی هایی که باید در هنگام انتخاب ابزار های داده کاوی به آن توجه نمود عبارتند از (Pereira et al, 2022)(Odan and Daraiseh, 2015):

متن باز بودن یا نبودن

بیشتر ابزار های برتر داده کاوی متن باز، هستند اما اغلب تفاوت های کمی باهم دارند.

امکان یکپارچه سازی داده ها

برخی از ابزار های داده کاوی با مجموعه داده های بزرگ، بهتر کار می کنند، در حالی که برخی دیگر با داده های کوچکتر بهتر کار می کنند. و به این نکته باید توجه داشت که وقتی گزینه های ابزار های داده کاوی را بررسی می کنید، انواع داده هایی که بیشتر با آنها سروکار 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 ابزار های داده کاوی، ماهیت آموزشی دارند در حالی که برخی دیگر، بر اساس نیاز های شرکت ها طراحی شده اند.

زبان برنامه نویسی

اکثر زبان های متن باز داده کاوی، به زبان جاوا توسعه یافته اند؛ اما بسیاری از آنها از اسکریپت های R و Python هم، پشتیبانی می کنند. ابزار های داده کاوی به سازمان ها و کسب و کار ها کمک می کنند تا درباره نیازمندی ها، اهداف و گام های پیاده سازی داده کاوی از اهمیت ویژه ای در فرایند داده کاوی برخوردار است. در ادامه برخی از ابزار های محبوب داده کاوی و کارکرد های آن ها را معرفی و مقایسه می نمائیم (Malkawi et al, 2020)(Mikut and Reischl , 2011):

Rapid Miner

Rapid Miner یک پلت فرم رایگان و متن باز داده کاوی است؛ که توسط شرکت Rapid Miner توسعه یافته است. رپید ماینر دارای صد ها 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 است.

Oracle Data Mining

Oracle Data Mining یکی 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؛ شامل چندین الگوریتم برای کارهایی مانند طبقه بندی، رگرسیون، تشخیص ناهنجاری، پیش بینی و غیره است.

IBM SPSS Modeler

این ابزار یکی از محبوب ترین و قدرتمند ترین ابزار های داده کاوی و تحلیل پیشرفته داده ها است. که توسط شرکت IBM توسعه داده شده است و به تحلیل و پیش بینی داده ها در حوزه های مختلف کمک می کند. 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 تحلیل های پیشرفته بر روی داده ها می کند. از مزایای این نرم افزار می توان به موارد زیر اشاره نمود (Wendler and Grottrup, 2016):

- ✓ 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 کار های تحلیلی

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

ماژول های موجود در این نرم افزار عبارتند از:

PASW Association

PASW Classification

PASW Segmentation

PASW Modeler Solution Publisher

یکی از ویژگی های این نرم افزار، این است که؛ 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

Weka یکی از نرم افزار های منبع باز و قدرتمند برای داده کاوی و یادگیری ماشین است. این ابزار توسط دانشگاه Waikato در نیوزلند توسعه داده شده و به افراد با سطوح تخصصی مختلف اجازه می دهد تا تحلیل داده های خود را انجام دهند. ابزارWeka از رابط کاربری گرافیکی قدرتمندی برخوردار است که به کاربران امکان انجام تحلیل های پیشرفته بر روی داده ها را بدون نیاز به دانش برنامه نویسی می دهد. این محیط شامل؛ روش هایی برای همه مسائل استاندارد داده کاوی مانند: رگرسیون، رده بندی، خوشه بندی، کاوش قواعد انجمنی و انتخاب ویژگی می باشد. با در نظر گرفتن اینکه، داده ها بخش مکمل کار هستند، بسیاری از ابزار های پیش پردازش داده ها و مصور سازی آنها فراهم گشته است. همه الگوریتم ها، ورودی های خود را به صورت یک جدول رابطه ای به فرمت ARFF دریافت می کنند. این فرمت داده ها می تواند از یک فایل خوانده شده یا به وسیله یک درخواست از پایگاه داده ای تولید گردد.

Knime

پلت فرم KNIME، یک ابزار قدرتمند و منبع باز برای داده کاوی، تجزیه و تحلیل داده های پیشرفته است. این ابزار توسط تیم KNIME توسعه داده شده و به کاربران اجازه می دهد تا به کمک یک رابط کاربری گرافیکی، فرایند های پیچیده تحلیلی را بدون نیاز به مهارت های برنامه نویسی انجام دهند.

H20

این ابزار ؛ یک پلت فرم یادگیری ماشین متن باز است که هدف آن دسترسی همه ی افراد به فناوری هوش مصنوعی است و از متداول ترین الگوریتم های ML پشتیبانی می کند و همچنین، به کاربران کمک کند تا مدل های یادگیری ماشین را به روشی سریع و ساده بسازند، حتی اگر متخصص نباشند.

Orange

ابزار داده کاوی ORANGE، یک نرم افزار متن باز و قدرتمند برای تجزیه و تحلیل داده و ایجاد مدل های یادگیری ماشینی است. این ابزار توسط دانشگاه Ljubljana در اسلوونی توسعه داده شده است و به کاربران اجازه می دهد، با استفاده از رابط کاربری گرافیکی ساده و آسان، تحلیل داده های خود را انجام دهند.

Apache Mahout

Apache Mahout؛ یک 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

این ابزار، روش ها و رویه های مختلفی را برای اجرای قابلیت های تحلیلی مختلف ارائه می کند که خواسته ها و اهداف سازمان را ارزیابی می کند. این نرم افزار شامل مدل سازی پیش بینی کننده و مدل سازی تجویزی است. ابزار داده کاوی SAS به دلیل طراحی و پردازش حافظه توزیع شده، بسیار مقیاس پذیر است.

روش طبقه بندی داده کاوی

یکی از روش های اکتشاف دانش که عموما در داده کاوی به کار می رود، روش طبقه بندی و استفاده از الگوریتم های درخت تصمیم می باشد درخت های تصمیم می توانند قواعد قابل فهمی را تولید کنند و حتی در یک درخت بزرگ یا پیچیده هم، یک مسیر را به راحتی می توان طی کرد و این باعث می شود که تفسیر دسته بندی ها یا پیش بینی ها نسبتا آسان باشد. الگوریتم های مختلفی جهت ساخت درخت های تصمیم معرفی شده است، که از روش های معروف آن می توان به: C5, C4.5., ID35, OC, CART, CHAID, SERCH, AID روش های QUEST, Algorithms SAS اشاره کرد (مشکانی و ناظمی، 1388).

ساخت پایگاه داده؛ داده کاوی

این گام به همراه دو گام بعدی هسته آماده سازی و ذخیره و بازیابی داده را تشکیل می دهند. 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 کاربرا ن به طور روز افزونی در حال انتخاب پایگاه داده های خاص منظوره ای هستند که این نیاز های داده کاوی را به نحو مناسبی حمایت کند. در هر صورت اگر داده موجود در انبار داده شما اجازه می دهد که مراکز منطقی داده ای ایجاد کنید 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 کاوی به شکل زیر می باشد (Turner et al , 2012):

- 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.)
- باز کردن پایگاه داده مربوط به داده کاوی
- نگهداری پایگاه داده مربوط به داده کاوی

باید به این نکته توجه داشت؛ که فعالیت های فوق ممکن است لزوما به ترتیب ذکر شده انجام نشوند.

مدل سازی داده کاوی

مهم ترین مسئله برای یاداوری در مورد ساخت مدل آن است که این کار یک فرایند تکراری است. 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 مسئله 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 تقسیم شود: یکی برای آماده کردن مدل و دیگری جهت تست مدل مربوطه.

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 و راه دوم؛ به کار بردن مدل ها در مجموعه داده های مختلف است. این مدل می تواند، برای مشخص نمودن رکورد ها بر اساس گروه بندی شان و یا مقدار دهی یک امتیاز مثلا احتمال انجام یک عمل استفاده گردد. به طور مثال: در یک سیستم تشخیص فرآیند؛ الگوهای موجود فرآیند می توانند با الگو های کشف شده تلفیق شوند. هنگامی که مواد مفروض این فرآیند برای ارزیابی به بررسی کنندگان فرستاده می شوند، بررسی کنندگان ممکن است نیاز داشته باشند که به رکورد هایی در پایگاه داده که مربوط به قسمت های ادعا شده توسط یک سازنده است دسترسی پیدا کنند. به طور کلی مراحلی که در این قسمت توضیح داده شد، برای انجام هر فرایند داده کاوی ضروری به نظر می رسند.

کاربرد های داده کاوی

فروش و بازاریابی

شرکت ها حجم عظیمی از داده ها در مورد مشتریان خود را جمع آوری می کنند که می توانند با بررسی دموگرافیک و رفتار کاربران آنلاین، از داده ها برای بهینه سازی کمپین های بازاریابی خود بهبود پیشنهاد های متقابل فروش و برنامه های وفاداری مشتری استفاده کنند و ROI بالاتری را در بازاریابی بدست بیاورند (Padhy et al, 2012) (Bartschat et al, 2019).

آموزش الکترونیکی

داده کاوی؛ یک رشته ی علمی و فرایندی است که باید به صورت یک پروژه پیاده سازی شود؛ یکی از شاخه های جالب آن داده کاوی آموزشی (EDM) می باشد (مقصودی وهمکاران، 1391). در سال های اخیر، تکنیک های داده کاوی و استفاده از آنها در امر آموزش مورد توجه قرار گرفته است که این زمینه ی تحقیقاتی جدید به امر توسعه روش های کشف دانش از داده های محیط آموزشی، خصوصا دانشجویان می پردازد (Romero and Ventura, 2007)(مقصودی و همکاران، 1391). بنا به تحقیقات موجود، روش های داده کاوی آموزشی اکثرا با روش های عادی داده کاوی متفاوت است؛ زیرا این حوزه نیازمند تشریح سلسله مراتب چند سطحی از داده های آموزشی و وابستگی ضمنی میان آنها می باشد (مقصودی و همکاران، 1391).

با اعمال داده کاوی بر داده های آموزش الکترونیکی می توان به تحلیل روش های آموزشی، خصوصیات محتوا و میزان فراگیری پرداخت (Mustapasa et al, 2010)(رضا پور و همکاران، 1392). نکته قابل توجه این است، که با استفاده از داده کاوی آموزشی می توان الگو و قوانین موجود در سیستم را کشف کرد که این قوانین به طور واضح در سیستم مشخص نیستند. استفاده از این الگو های کشف شده می تواند چاره ساز مشکلاتی باشد که امروزه دانشجویان در سیستم های آموزش الکترونیکی با آن مواجه هستند (مقصودی و همکاران، 1391).

پزشکی

داده کاوی به پزشکان کمک می کند که؛ با جمع آوری سابقه ی پزشکی هر بیمار، نتایج معاینه ی فیزیکی، دارو ها و الگو های در مانی، تشخیص های دقیق تری بدهند؛ همچنین داده کاوی به ایجاد استراتژی های مدیریت منابع پزشکی مقرون به صرفه تر کمک بزرگی می کند (محمدی، 1402).

تخلفات رانندگان

آمار بالای تخلفات ترافیکی و حوادث ناشی از آن در کشور، موجب بروز صدمات جانی و مالی جبران ناپذیری می گردد که عوامل انسانی در این میان، اصلی ترین سبب بروز تخلفات می باشند. نیروی انتظامی با توجه به سامانه های اطلاعاتی متعددی که از سال ها پیش در حوزه های مختلف ماموریتی ایجاد نموده است؛ حجم عظیمی از اطلاعات را در اختیار دارد که تجزیه و تحلیل این اطلاعات به مدد رویکرد داده کاوی، امکان کشف روابط، روند ها و الگو های مخفی بین داده ها و دستیابی به دانش نوین در زمینه چالش های آشکار و نهان ناجا را میسر خواهد ساخت. در نیروی انتظامی اطلاعات مرتبط با گواهینامه های صادر شده برای رانندگان و آموزشگاه های رانندگی که متولی آموزش متقاضیان دریافت گواهینامه می باشند و همچنین، اطلاعات تخلفات رانندگی در بانک های اطلاعاتی مختلف، در اختیار پلیس راهور ناجا است و با توجه به حجم زیاد این اطلاعات، امکان استفاده از رویکرد داده کاوی برای تشخیص ارتباط تخلفات رانندگان با مشخصات فردی رانندگان و آموزشگاه های رانندگی وجود دارد: برخی از پژوهش های انجام شده در زمینه ی تحلیل راهبردی تخلفات رانندگان با استفاده از روش های داده کاوی به شرح زیر است:

روش داده کاوی با هدف دستیابی به قوانین تصمیم گیری و الگوی شناسایی رانندگان پرخطر در تخلفات مورد استفاده واقع شده است. در اجرای روش دسته بندی در داده کاوی، دو روش شبکه های عصبی و درخت تصمیم در اطلاعات تخلفات به کار گرفته شده است و در انتها برای بهبود نتایج، از رویکرد ترکیبی شامل روش خوشه بندی و درخت تصمیم استفاده شده است (جعفری و صمدیان ، 1391).

هوشیار و شریفی، طی تحقیقی در سال ۱۳۹۵؛ تصادفات درون شهری، شهر ارومیه در سال ۱۳۹۲ را مورد تحلیل فضایی قرار دادند و نتایج تحقیق آنها نشان داد که بیشترین تعداد تصادفات در مناطق ۱و۳ و اطراف بخش مرکزی شهر ارومیه رخ داده است و توزیع نقاط تصادفی برای شهر ارومیه به صورت خوشه ای بوده است.

شاه محمدی و رجبی، در سال ۱۳۹۷در مقاله خود، به تحلیل جرایم اخلاقی فضای سایبر با رویکرد داده کاوی پرداختند. جامعه آماری پژوهش، پایگاه داده جرایم سایبری پلیس فتا است و حجم نمونه، بخشی از پایگاه داده جرایم سایبری است که مربوط به جرایم اخلاقی فضای سایبر در سال های ۱۳۸۹ تا ۱۳۹۴ می باشد. نتایج نشان داد که بهترین الگو برای تحلیل و استخراج قوانین حاکم بر داده های جرایم اخلاقی فضای سایبر، الگوی درخت تصمیم C5 است؛ همچنین، مشخصه های تاثیر گذار بر وقوع جرایم اخلاقی فضای سایبر به ترتیب شغل، تحصیلات، جنسیت، سن، تاهل و نقش فرد است، در این پژوهش، قوانین حاکم بر جرم اخلاقی فضای سایبر احصا گردید که با تحلیل این قوانین می توان راهکارهای پیشگیرانه ای را ارائه نمود.

تشخیص کلاهبرداری

تا کنون میلیارد ها دلار به دلیل کلاهبرداری از دست رفته است. روش های سنتی کشف کلاهبرداری زمان بر و پیچیده هستند. داده کاوی به ارائه الگو های معنا دار و تبدیل داده ها به اطلاعات کمک می کند. یک سیستم تشخیص کلاهبرداری کامل باید از اطلاعات همه ی کاربران محافظت کند. یک روش برای ایجاد چنین سیستمی یادگیری با ناظر است. این روش جمع آوری نمونه های قبلی را در بر می گیرد که به دو دسته ی کلاه برداری یا غیر کلاه برداری طبقه بندی می شوند. در این روش الگویی با استفاده از این داده ها ساخته می شود تا تشخیص دهد نمونه کلاهبرداری است یا خیر. این قضیه در بانک ها و دیگر موسسات مالی بسیار استفاده می شود و شرکت های مستقر در SAAS نیز برای حذف حساب های کاربران جعلی از مجموعه داده های خود، اقدام به اتخاذ این روش ها کرده اند (Padhy et al, 2012)(Goebel et al, 1999)(Halkidi et al, 2011).

مسائل بانکداری

داده کاوی بخشی از فرایند استخراج معرفت است که در آن الگو های مفید و ضمنی در پایگاه داده ها جستجو می شوند. در این میان با افزایش کاربرد سیستم های اطلاعات جغرافیایی، پایگاه های بزرگی از داده های متنوع جغرافیایی در دسترس قرار گرفته اند که؛ کمک شایانی به انجام تحلیل های کامل تر و دقیق تر می نمایند. برای مثال فرض کنید؛ به دنبال بررسی و اجرای یک روش داده کاوی پیشرفته روی داده های فضایی موجود در بانک ملت ایران می باشید که با داده های مختلف بانکی از قبیل مکان شعب، شاخص های بانکی مانند: درآمد، سود، هزینه، تعداد کارکنان، میزان مراجعه و مانند آن تلفیق خواهد شد. بدین معنی که بعد از انجام مراحل لازم جهت آماده سازی داده ها، با ملاحظات لازم به دلیل فضایی بودن آنها برای عملیات داده کاوی؛ شامل پردازش و پاکسازی داده ها و ساخت انبار داده ها و همچنین در نظر گرفتن روش های دسترسی به داده های فضایی، الگوریتمی برای استخراج قوانین وابستگی توسعه و پیاده سازی خواهد شد و از آن برای کشف روابط موجود ما بین مقادیر مختلف و جغرافیایی مانند: ترکیب جمعیتی، کاربری های منطقه، وضعیت سنی، درآمد، تحصیلات، موقعیت رقبا، شبکه معابر و مانند آن از یک طرف و شاخص های بانکی شعب مانند سود، هزینه، درآمد، کارایی و مانند آن از طرف دیگر استفاده خواهد شد. دانش استخراج شده از این فرایند، در تصمیم گیری های مختلف مدیران در حوزه مدیریت شعب، مانند: مکان یابی، توسعه، تلفیق و تنظیم شعب، کاربرد و اهمیت بالایی خواهد داشت (احمدی، 1397).

سایر کاربرد های رایج و مهم داده کاوی به شرح زیر می باشند (Padhy et al, 2012) (Goebel et al, 1999)(Halkidi et al, 2011):

- تولید محصولات صنعتی و تجاری
- یافتن جامعه هدف برای سازمان ها و کسب و کار ها
- کشف الگو های رفتاری
- پیش بینی فروش در فرایند های تجاری
- دسته بندی آیتم ها بر اساس تفاوت های موجود
- تجمیع و تمرکز بر روی داده های بزرگ

بحث و نتیجه گیری

باتوجه به این که امروزه سازمان ها؛ به دنبال دستیابی به اهداف بزرگ تر با بهره گیری از علم داده هستند. در این مقاله به بررسی داده کاوی و کاربرد های آن پرداختیم. با توجه به مطالعاتی که در این زمینه داشتیم، به این مسئله پی بردیم که تا چند دهه قبل فقط سازمان هایی مانند ناسا می توانستند از ابر رایانه های خود برای تجزیه و تحلیل داده ها استفاده کنند؛ به این علت که، هزینه ذخیره سازی و محاسبه کردن داده ها بسیار بالا بوده و در واقع در توان شرکت های کوچک نبوده است، اما هم اکنون؛ شرکت ها انواع فعالیت ها را با استفاده از؛ یادگیری ماشین، هوش مصنوعی و یادگیری عمیق با هزینه های مناسب تر و با وجود حجم انبوهی از داده ها در کوتاه ترین زمان ممکن انجام می دهند که این موضوع باعث موفقیت حجم انبوهی از کسب و کارها شده است. در پایان باید به این نکته توجه داشت که؛ آینده ای بسیار روشن در انتظار داده کاوی و علم داده است. زیرا در آینده مدیریت حجم گسترده ای از داده ها برای سازمان ها مشکلاتی به وجود خواهد آورد و این کار توسط علم داده امکان پذیر خواهد بود.

منابع

جماعت، علی و عسگری، فرید، 1389، مدیریت ریسک اعتباری در سیستم بانکی با رویکرد داده کاوی، نشریه مطالعات کمی در مدیریت، 1(3)، 115-126

احمدی، اکرم، 1397، داده کاوی و کاربرد های آن در سنجش رضایت مشتری، دانشگاه تهران

مشکانی، علی، ناظمی، عبدالرضا، 1388، داده کاوی کاربردی، دانشگاه آزاد اسلامی واحد نیشابور

مقصودی، بهروز و سلیمانی، صادق و امیری، علی و افشارچی، محسن، 1391، ارتقای کیفیت آموزش در رسانه های آموزش الکترونیکی با استفاده از داده کاوی، نشریه علمی پژوهشی فناوری آموزش، سال ششم، جلد ششم، شماره چهارم

رضا پور، محمد و سپهری، محمد مهدی و رضا پور، حسن، 1392، تعیین روش های برتر سنجش فراگیران در دوره های یاد گیری الکترونیکی رویکرد داده کاوی، پژوهش مدیریت در ایران، دوره هفدهم، شماره چهارم

محمدی، پیام، 1402، داده کاوی (کاربرد هاو نیازمندی ها و فرایند و ابزارها)، دانشگاه آزاد اسلامی، واحد علوم تحقیقات،آذربایجان

جعفری، ادریس، صمدیان، منیره السادات، 1391، کاربرد داده کاوی در بررسی رفتار رانندگان مختلف در کلان شهر ها، فصلنامه علمی ترویجی مطالعات راهور، سال نهم، شماره هفدهم

Gupta, MK and Chandra P.A. (2020). comprehensive survey of data mining. International Journal of Information Technology. Dec. 12 (4).1234-57.

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

Roy, U. And Zhu, B. And Li, Y. And Zhang, H. And Yaman, O. (2014). 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.

Hong, TP. And Lee, YC. And Wu, MT. (2014). An effective parallel approach for genetic fuzzy data mining. Expert Systems with Applications 41(2) 111-111.

A J, Han 1. And B M, Kamber 2. (2001). Data Mining: Concepts and Techniques San Diego Academic Press (2001).

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

Bartschat, A. And ReischI, M. And Mikut, R. (2019). Data mining tools. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. (2019) Jul.9(4). e1309.

Goebel, M. And Gruenwald, LA. (1999) survey of data mining and knowledge discovery software tools. ACM SIGKDD explorations newsletter. (1999) Jun 1 1(1).20-33.

Halkidi, M. And Spinellis, D. And Tsatsaronis, G. And Vazirgiannis, M. (2011) Data Mining in software engineering. Intelligent Data Analysis. (2011) Jan 1.15(3).413-41.

Santos-Pereira, J. And Gruenwald, L. And Bernardino, J. (2022). Top data mining Tools for the healthcare industry. Journal of king Saud University Computer and Information Sciences. (2022) sep 1.34(8). 4968-82.

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

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

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

Wendler, T. And Grottrup, S. (2016). Data mining with SPSS modeler: theory exercises and solutions. Springer (2016) Jun 6.

Turner, CJ. And Tiwari, A. And Olaiya, R Xu Y. (2012). PD rocess mining: from theory to practice. Business Process Management Journal. (2012) Jun 1.18(3).493-512.

Romero, C. And Ventura, S. (2007). Educational data mining: A survey from (1995) to (2005). Expert Systems with Applications. pp.135-146

Mustapasa, O. And et al. (2010). Implementation of semantic web mining on E-learning. Science Direct. Istanbul-Turkey Department of Software Engineering University of Bahcesehir January. (2010).

Familiarity with the concept of data mining and examining its applications

Sara Shahsavari

Department of computer software Engineering, Yadegar-e- Imam Khomeini (RAH) Shahre-Rey Branch, Islamic Azad University, Tehran, Iran

Abstract

The science of data mining is known as knowledge discovery in databases. Which as a knowledge discovery process, includes: data cleaning, data selection, data integration, data transformation, pattern discovery, pattern evaluation and knowledge presentation. One of the reasons what this science has been paying attention to is the high confidence factor of the decisions made based on the data analysis and the results that are created Data mining can be used in various fields, such as: public health, medicine, criminology research, etc. It is interesting to know that data analysis saves costs and improves business and can identify new customers and revenue streams. The purpose of this research is to find out why data mining is in high demand and how it is part of the natural evolution of information technology. In the past, it took a lot of time to analyze large data sets, but now one of the main advantages of data mining is speed.

Keywords: data mining, knowledge discovery, data analysis, data cleaning, data transformation, information technology