داده کاوی یک ابزار آنالیز مدیریتی

Data Mining an Analysis Implement Managemental

استاد راهنما : مهندس عمادی

ارائه دهنده :

مسعود فهرستی خمسه

یوسف مغانلو فرد

قربان مقدم زرزری

موسسه آموزش عالی روزبه زنجان

چکیده

داده کاوی فرآیندی تحلیلی برای کاوش داده های طراحی شده است،که در جستجوی الگوهای سازگار، یا روابط سیستماتیک بین متغیرها است، و سپس به تائید این یافته ها با استفاده از الگوهای تشخیص داده شده می پردازد. استخراج اطلاعات مناسب از میان انبوه داده‌ها و تبدیل آنها به دانش مورد نیاز سازمانها، بویژه در تصمیم‌گیری‌های سازمانی، نیازمند استفاده از روش‌های نوین در این حوزه است. داده کاوی یکی از این ابزار و رویکردهاست که در فضای مدیریت دانش سازمان‌ها به کشف دانش از پایگاه داده ها کمک می‌کند. این مقاله به بررسی ویژگی های منحصر به‌فرد این حوزه از فناوری و تکنیکهای استفاده از آن را نشان می دهد.

مقدمه

به مرور زمان، استخراج و کشف سریع و دقیق اطلاعات با ارزش و پنهان از پایگاه داده‌ها به‌عنوان داده‌کاوی مورد توجه قرار گرفت. به این شکل بود که فرایند داده‌کاوی به عنوان فرایند آماری و تجزیه و تحلیل درفرایند کشف دانش در پایگاه داده‌ها (KDD) پررنگ شد ،به حدی که گاه، داده‌کاوی (DM) به‌عنوان مترادف کشف دانش در پایگاه داده‌ها(KDD) مورد استفاده قرار می‌گرفت[2]. امروزه فرایند استخراج اطلاعات معتبر، از پیش ناشناخته، قابل فهم و قابل اعتماد از پایگاه داده‌های بزرگ و استفاده از آن در تصمیم‌گیری و در فعالیت‌های تجاری داده‌کاوی نامیده می‌شود[1]. در تعاریف متعدد و متنوع برای داده‌کاوی برموضوعاتی نظیر: استخراج دانش کلان، کاوش در داده‌ها، تجزیه و تحلیل داده‌ها و یافتن روابط و الگوهای مطمئن بین داده‌ها تاکید می‌شود. هدف نهایی داده‌کاوی، ایجاد سیستم‌های پشتیبانی تصمیم‌گیری سازمانی است. داده‌کاوی به استخراج اطلاعات مفید و دانش از حجم زیاد داده‌ها می‌پردازد.

داده‌کاوی، الگوهای حاوی اطلاعات را در داده‌های موجود جست‌وجو می‌کند. این الگوها و الگوریتم‌ها، می‌توانند توصیفی باشند یعنی داده‌ها را توصیف کنند و یا جنبه پیش‌بینی داشته باشند، یعنی از متغیرها برای پیش‌بینی ارزش‌های ناشناخته سایر متغیرها به‌کار روند. داده‌کاوی توصیفی، به‌دنبال یافتن اگرها در فعالیت‌ها یا اقدامات گذشته است و داده‌کاوی پیش‌بینانه با نگاه به سابقه، رفتار آینده را پیش‌بینی می‌کند[1].

حوزه فعالیت

اکتشاف در این مرحله معمولا با آماده سازی داده ها که ممکن است شامل تمیز کردن داده ها ، تبدیل داده ها ، زیر مجموعه های انتخاب آثار ضبط شده و انجام برخی از عملیات اولیه انتخاب شروع می شود . سپس بسته به ماهیت تحلیلی ، این مرحله از فرایند استخراج داده ها ممکن است شامل هر انتخاب ساده و سرراست برای یک مدل رگرسیون استادانه درست شده را به تجزیه و تحلیل اکتشافی با استفاده از طیف گسترده ای از روش های گرافیکی و آماری به منظور شناسایی متغیرهای مربوطه و تعیین پیچیدگی از طبیعت مدل ها باشد. البته ناگفته نماند که داده کاوی معمولا با نوشتن مقدار زیادی گزارش و تحقیق و استعلام در آنها اشتباه گرفته می شود. اما در واقع داده کاوی هیچ کدام از اینها را شامل نمی شود. داده کاوی توسط تجهیزات خاصی صورت می پذیرد، که عملیات کاوش را بر اساس تجزیه و تحلیل مکرر داده ها انجام می دهد. داده کاوی با آنالیز های متداول آماری نیز متفاوت است؛در زیر به بررسی نوع دیدگاه روش آماری به روش داده کاوی می پردازیم:

روش آنالیز آماری :

یک مفسر ممکن است متوجه الگوی رفتاری شود که سبب کلاهبرداری بیمه گردد. بر اساس این فرضیه، مفسر به طرح یک سری سوال می پردازد تا این موضوع را بررسی کند. اگر نتایج حاصله مناسب نبود، مفسر فرضیه را اصلاح می کند و یا با انتخاب فرضیه دیگری مجددا شروع می کند. این روش نه تنها وقت گیر است بلکه به قدرت تجزیه و تحلیل مفسر نیز بستگی دارد.مهمتر از همه اینکه این روش هیچ وقت الگوهای کلاهبرداری دیگری را که مفسر به آنها مظنون نشده و در فرضیه جا نداده ، پیدا نمی کند.

روش داده کاوی :

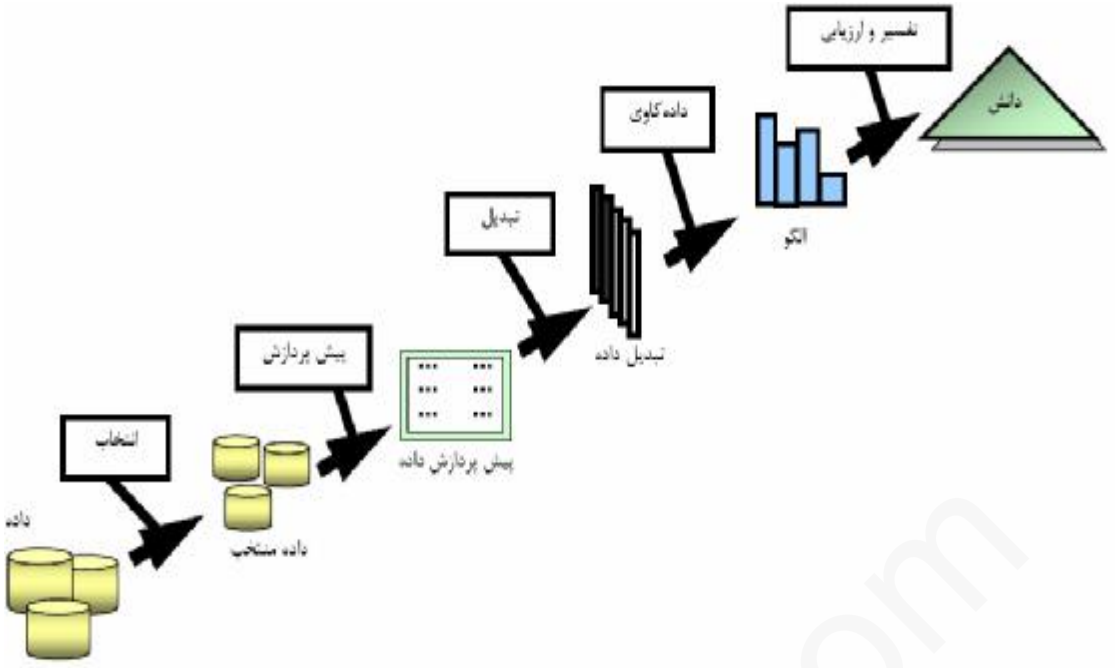
یک مفسر سیستم های داده کاوی را ساخته و پس از طی مراحلی از جمله جمع آوری داده ها، یکپارچه سازی و اخلاص داده ها به انجام عملیات داده کاوی می پردازد. داده کاوی تمام الگوهای غیرعادی را که از حالت عادی و نرمال انحراف دارند و ممکن است منجر به کلاهبرداری شوند را پیدا می کند. نتایج داده کاوی حالت های مختلفی را که مفسر باید در

مراحل بعدی تحقیق کند، نشان می دهند. در نهایت مدل های به دست آمده می توانند مشتریانی را که امکان کلاهبرداری دارند، پیش بینی نمایند.

تحلیلهای داده‌کاوی به دو روش با ناظر و بدون ناظر و از طریق الگوریتمهایی چون شبکه های عصبی(NN)، طبقه بندی و درخت تصمیم (C&RT)، ژنتیک، تحلیل سبد خرید، شبکه کوهونن قابل اجراست. علاوه بر این الگوریتمهای رایج، همچنان الگوریتمهای جدیدی برای اهداف تحقیقات علمی یا تجاری از طریق طرحهای پژوهشی دانشگاهی، تولید می‌شود. ویژگیهای منحصر بفرد داده کاوی را می توان به صورت زیر برشمرد[3]:

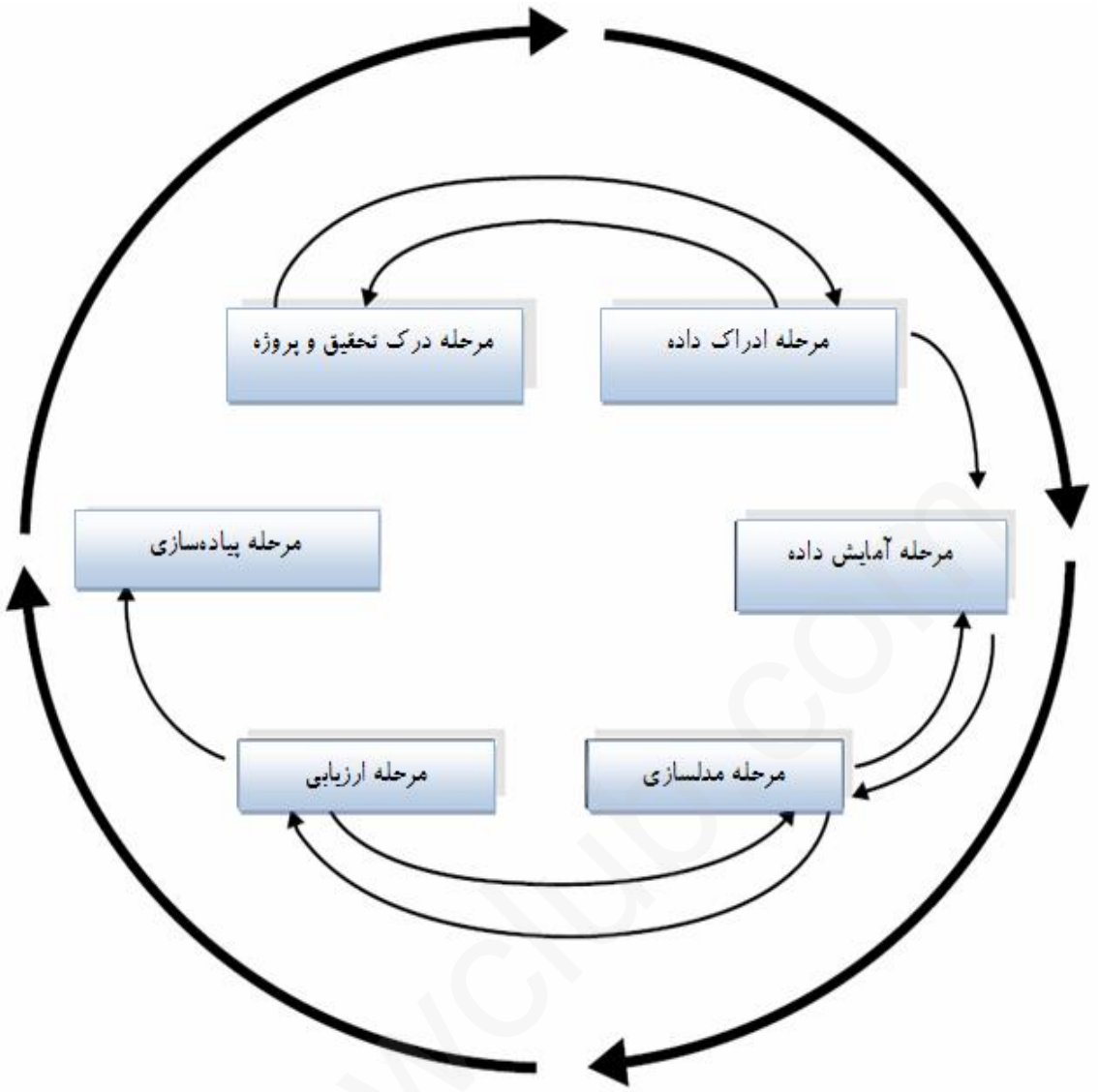
- نه تنها بر فاز تحلیل، بلکه بر طراحی مطالعه و جمع آوری داده نیز تاثیر می گذارند
- امکان جستجوی پاسخ سؤالات دقیق و با پیچیدگی بالا را در دادههای جمع‌آوری شده فراهم می کنند.
- قادرند که به سؤالات بطور واضح و مشخص پاسخ دهند. مزیت اصلی و تفاوت آنها با سایر تکنیکها نیز در همین است که بجای ارائه صرف استراتژی کلان، پاسخهای دقیق در اختیار محقق قرار می دهند.
- امکان سنجش اثر متغیرهای مختلف بر روی متغیرهای وابسته را فراهم می کنند.
- به مدیران کمک می‌کنند که تأثیر سناریوهای آتی را مورد ارزیابی قرار دهند و با مدلسازی گزینه های متعدد و کمک به تصمیم گیری در شرایط عدم قطعیت به انتخاب مسیر حرکت بپردازند.

محققینی که تنها روابط دو به دو را در نظر می گیرند و از داده کاوی استفاده نمی کنند، ابزار قدرتمندی را از دست می دهند که می تواند اطلاعات سودمندی را در اختیار آنان قرار دهد. در مسائل واقعی چندین متغیر به طور همزمان بر روی پاسخ تاثیر می گذارند، از این رو آنالیزهای چندمتغیره جواب های دقیقتر و نزدیک به واقع تری را فراهم می کند. در شکل (۱) فرایند کسب دانش از پایگاه دادهها به صورت شماتیک بیان شده است[4] همانطور که ملاحظه می شود یکی از گام های این فرایند، داده کاوی می باشد. موفقیت در این مرحله کاملا متاثر از سه گام قبل است بگونه ای که اگر هر کدام از مراحل قبلی به درستی انجام نپذیرد، نتایج حاصل از داده کاوی نه تنها مفید نبوده ممکن است گمراه کننده نیز باشد.



شکل (۱): فرآیند تبدیل دادها به دانش

تکنیکهای داده کاوی از جمله تکنیکهای نوین علمی هستند که در توصیف، تشریح، پیش بینی و کنترل پدیدهها به کار می روند[3]. این تکنیکها به اندازه‌گیری، تشریح و پیش‌بینی درجه وابستگی میان متغیرها میپردازند. روشهای داده‌کاوی نه تنها بر جنبه های تحلیلی مطالعات، بلکه در طراحی و ابزارهای جمع آوری داده برای تصمیم‌گیری و حل مسائل نیز تأثیر می‌گذارند. موفق‌ترین پروژههای داده‌کاوی، در چارچوب فرآیند استانداردی اجرا می شودکه توسط یک تیم کاری در شرکت SPSS در قالب پروژهای به نام CRISP-DM ارائه شده است[۵]. برطبق CRISP-DM یک پروژه داده‌کاوی معین شامل چرخه حیاط شش مرحله‌ای است که توالی مراحل را نشان می دهد شکل (۲). هر مرحله از ترتیب مراحل اغلب نتیجه وابستگی مراحل قبلی را نیز دربر دارد. مهمترین وابستگی بین مراحل نمایش پیکانها می باشد. خاصیت تکراری CRISP حاکی از چرخه بیرونی است که اغلب منجر به راه



شکل(۲) CRISP–DM در فرایند تکرار و سازگاری مراحل

حلی برای مسئله تحقیقی یا تجاری با سوالات اضافی جالب توجه می شود. در زیر مراحل کاری در داده‌کاوی را توضیح می‌دهیم:

مرحله درک پروژه و فهم حوزه کاربرد: اولین مرحله پردازش استاندارد CRISP–DM می باشد که به صورت آشکار اهداف و نیازمندیها آن مشخص می شود. ترجمه اهداف و محدودیت آن در قاعده‌سازی، تعریف مسئله داده‌کاوی و مهیا کردن استراتژی اولیه برای نائل شدن به اهداف تعریف می‌شود.

مرحله انتخاب دادها : این مرحله شامل جمع آوری دادها برای استفاده از تحلیل اکتشافی و مشخص کردن اطلاعات اولیه برای ارزیابی داده‌های با کیفیت و انتخاب دادهای مفید و مورد نیاز می باشد.

مرحله آماده سازی داده‌ها: آماده کردن داده‌های اولیه خام به داده‌های نهایی، این دادها در کلیه مراحل بعدی استفاده می شود و از این نظر این مرحله تحلیل و تلاش بیشتری را می طلبد. انتخاب عناصر و شناسه‌های تحلیل شده را برای کاوش داده‌ها اختصاص می دهیم. و با تمیز کردن دادهای خام آن را برای ابزارهای مدلسازی آماده می‌کنیم.

مرحله مدلسازی: با انتخاب و به‌کار بستن تکنیکهای مدلسازی مناسب و روش داده‌کاوی معین نتایج مدلسازی را بهینه می‌کنیم که در صورت نیاز می توانیم با برگشت به عقب تحلیل مدلسازی را بهینه تر نماییم.

مرحله ارزیابی: مشخص کردن اینکه آیا مدل انتخابی، ما را به اهدافمان که در اولین مرحله تعیین کردیم می‌رساند. اتخاذ تصمیم راجع به استفاده از نتایج داده‌کاوی برای اعتبارسنجی نیز در این مرحله انجام می‌شود.

مرحله تحکیم و گسترش :استفاده کردن از مدل ایجاد شده، برای مثال می تواند تولید یک گزارش ساده از خروجیها را نام برد، و برای یک مثال پیچیده تکمیل کردن پردازش داده‌کاوی موازی در سایر حوزه‌ها می باشد که این الگوها به یک دانش مفید و قابل استفاده تبدیل می شوند و پس از بهبود آنها، الگوهایی که کارا محسوب می‌شوند در یک سیستم اجرایی به کار گرفته خواهند شد.

نتیجه گیری

بررسی اجمالی پژوهشهای صورت گرفته در حوزه دانش ابزراهای داده‌کاوی نشان می دهد که تحقیقات عمیق و اساسی در این باره خصوصاً در ایران اندک شمار است. از سوی دیگر با افزایش سرعت تحول در علوم، ضرورت استفاده از دانشهای نوین بیش از پیش محرز شده است . داده کاوی به عنوان یک رشته علمی نوین در زمینه بازیابی و استخراج اطلاعات می تواند نقش مهمی در جهت دستیابی به این اهداف داشته باشد. امروزه اکثر نرم افرار های پایگاه داده ای مثل ORACLE و SQL Server نیز شامل ابزارهایی داده کاوی شده اند ولی نرم افزار های تخصصی داده‌کاوی همچون Knowledge ,Intelligent Miner , Darwin , Mine Set Studio, Data Mind از مهمترین ابزار های داده کاوی به شمار می روند. در این مقاله قابلیتهای داده کاوی و مراحل کاری آن معرفی شد که در گامهای بعد می توان تأثیر آن را در عمل آزمود.

مراجع

[1] B. Fernandez / Et. Al., "Knowledge Management"/ Cho. 12, 2004.

[2] N.Balac/ "Introduction To Data Mining" , 2006

[3] Hair ،Joseph F., *"Multivariate Data Analysis"*, Prentice Hall, 2005.

[4] Daniel T. Larose, *"Discovering Knowledge in Data: An Introduction to Data Mining"* , 2004 .
[5] www.spss.com/ CRISP DM/ Downloads
[6] Pang-Ning Tan, Steinbach, *"Introduction to Data Mining"*, 2005 .