فصلنامه گزیده‌مطالب آماری
سال ۱۶، شماره ۴، زمستان ۱۳۸۴، صص ۳۴-۲۱

داده‌کاوی و کاربرد آن در آمار رسمی

علی‌اصغر حائری مهریزی*
حسین حسنی

کارشناس ارشد آمار

چکیده. امروزه داده‌کاوی به‌عنوان یکی از ابزارهای پرتوان در تحلیل داده‌ها و پایگاه‌های داده‌ای حجیم با ابعاد زیاد بین محققان در علوم مختلف از جمله محققان در حوزه آمار رسمی شناخته شده است. بر همین اساس مراکز آماری مختلف روز به روز به اهمیت داده‌کاوی پی‌برده و این مهم را در اولویت‌های برنامه‌ای خود قرار داده‌اند. هدف این مقاله معرفی داده‌کاوی و شاخه‌های مرتبط با آن، اهمیت و کاربرد آن در آمار رسمی و موضوع‌های مرتبط با آن است.

۱- مقدمه

از سال ۱۹۵۰ به بعد که رایانه در تحلیل و ذخیره‌سازی داده‌ها به کار رفت، حجم اطلاعات ذخیره شده در آن با گذشت زمان بیش‌تر شده و همچنین رو به فزونی است. بسیاری از پایگاه‌های داده‌ها چنان گسترش یافته‌اند که شامل چند صد میلیون یا چندین میلیارد رکورد ثبت شده هستند و امکان تحلیل و استخراج اطلاعات با روش‌های معمول و کلاسیک آماری از این پایگاه داده‌ها (داده‌انبارها) مستلزم داشتن دانش و ابزارهای توانمندتر است. از طرفی شدت رقابت‌ها در عرصه‌های علمی، اجتماعی، اقتصادی، سیاسی و نظامی نیز اهمیت سرعت یا زمان دسترسی به اطلاعات را دو چندان کرده است. بنا بر این نیاز به طراحی سیستم‌هایی که قادر به اکتشاف سریع اطلاعات مورد علاقه کاربران با

واژگان کلیدی: داده‌کاوی؛ فرایند کشف دانش از پایگاه داده‌ها؛ آمار رسمی.
* نویسنده عهده‌دار مکاتبات

تاکید بر حد اقل مداخله انسانی باشند از یک سو و روی آوردن به روش‌های تحلیل متناسب با حجم داده‌های زیاد از سوی دیگر به خوبی احساس می‌شود. در حال حاضر، داده‌کاوی مهم‌ترین فناوری برای بهره‌برداری موثر، صحیح و سریع از داده‌های حجیم بوده و اهمیت آن رو به فزونی است.

از طرفی سازمان‌ها و موسسه‌های آماری دارای داده انباره‌های حجیمی از اطلاعات هستند که از منابع مختلف و موضوع‌های متفاوت نشات گرفته و جمع‌آوری شده‌اند. در این خصوص داده‌کاوی به‌عنوان ابزاری توانمند نه تنها دسترسی به اطلاعات را تسهیل می‌سازد بلکه باعث می‌شود تا از دل این داده‌انبارها اطلاعات مفید و قابل اعتمادی که تا کنون نهفته بوده را به دست آورد.

اما با این مقدمه یک سؤال در ذهن تداعی می‌کند و آن، این است که چرا تا کنون مراکز و موسسه‌های آماری به این مهم کمتر پرداخته‌اند؟ پاسخ‌های متفاوتی در این خصوص مطرح شده است. شاید یکی از مستدل‌ترین آن‌ها این باشد که وظیفه اصلی مراکز و موسسه‌های آماری تولید داده است و وظیفه تحلیل داده‌ها معمولاً با مراکز و موسسه‌های دیگر است. علاوه بر آن به‌نظر می‌رسد که بینش اکتشاف و جستجو در یک پایگاه داده به‌منظور یافتن الگو یا مدل‌هایی برای آماردانانی که در حوزه آمار رسمی پژوهش می‌کنند در مراحل بعدی قرار دارد و به‌طور کلی می‌توان قسمت عمده تحلیل‌های آماری را در این سازمان‌ها به ساختن چارچوب آماری اختصاص داد. اما از آن‌جا که این مراکز و موسسه‌های آماری وظیفه تولید، جمع‌آوری، نگاهداری، حفظ و مدیریت داده انبارهای حجیم و متفاوت مربوط به جامعه، تجارت، کشاورزی، صنعت و... را به عهده دارند لذا بهترین پتانسیل را در جهت اکتشاف روابط و اطلاعات مفید به وسیله این داده‌انبارهای متفاوت خواهند داشت. با این تفکر می‌توان درک کرد که چرا امروزه به کاربرد داده‌کاوی در آمار رسمی و به‌ویژه در سازمان‌ها و موسسه‌های آماری تاکید می‌شود. بر همین اساس در حال حاضر پروژه‌های مختلفی که تمامی آن‌ها به وسیله این سازمان‌ها حمایت می‌شود، انجام شده و در حال انجام می‌باشد که به چندین مورد آن در این مقاله در حد اختصار اشاره می‌شود. در این مقاله داده‌کاوی در حد اختصار معرفی، تاریخچه و

تعاریف مختلف آن بیان شده است (برای اطلاعات بیش‌تر در این خصوص به [۳-۱] رجوع شود) سپس کاربرد داده‌کاوی در آمار رسمی مورد بررسی قرار گرفته است.

۲- تاریخچه، تعاریف و شاخه‌های مرتبط با داده‌کاوی

داده‌کاوی فرایندی است که در آغاز دهه ۹۰ میلادی پا به عرصه ظهور گذاشته و با نگرشی نو به مسئله استخراج اطلاعات از پایگاه داده‌ها می‌پردازد. در سال‌های ۱۹۸۹ و ۱۹۹۱ کارگاه‌های کشف دانش از پایگاه داده‌ها توسط پیاتتسکی و همکارانش برگزار شد. همچنین در فاصله سال‌های ۱۹۹۱ تا ۱۹۹۴ کارگاه‌های کشف دانش و معرفت از پایگاه داده‌ها نیز برگزار شد. به‌طور رسمی اصطلاح داده‌کاوی برای اولین بار توسط فایاد در اولین کنفرانس بین‌المللی « کشف دانش و داده‌کاوی » در سال ۱۹۹۵ مطرح شد. در این سال داده‌کاوی به صورت جدی وارد مباحث آمار شد [۴]. داده‌کاوی حاصل تحول تدریجی در طول تاریخ بوده و از اوایل دهه ۹۰ هم‌زمان با همه‌گیر شدن استفاده از پایگاه‌های داده‌ای، به‌عنوان یک علم مطرح شده است [۵].

موضوع داده‌کاوی شناخت مولفه‌های جدید و با ارزش، مفید، رابطه‌های منطقی و الگوهای موجود در داده‌ها است. در زمینه‌های مختلف یافتن الگوهای مفید در داده‌ها با عنوان‌های متعددی (مانند داده‌کاوی) بیان می‌شود. برای مثال از عنوان‌هایی نظیر استخراج دانش، کشف اطلاعات، برداشت اطلاعات و پردازش الگوهای داده‌ها می‌توان نام برد.

نگاهی به ترجمه لغوی داده‌کاوی ما را در درک بهتر این واژه کمک می‌کند. واژه لاتین mine به معنای استخراج از منابع نهفته و با ارزش زمین اطلاق می‌شود. ادغام این کلمه با کلمه داده‌ها (data) بر جستجویی عمیق از داده‌های قابل دسترس با حجم زیاد برای یافتن اطلاعات مفید که قبلاً نهفته بودند، تاکید دارد.

داده‌کاوی دارای تعریف‌های مختلفی است. این تعریف‌ها به مقدار زیادی به پیش زمینه‌ها و دیدگاه‌های افراد بستگی دارد. هر نویسنده، محقق و یا کاربر با توجه به نوع نگرش خود تعریف‌های مختلفی از داده‌کاوی ارائه کرده است. به‌عنوان مثال می‌توان به چند تعریف داده‌کاوی که در ادامه آمده است، اشاره کرد:

الف) داده‌کاوی فرایندی از شناخت الگوهای معتبر، جدید، مفید و قابل فهم از داده‌ها است [۴]؛

ب) داده‌کاوی به فرایند استخراج اطلاعات نهفته، قابل فهم، قابل تعقیب از پایگاه داده‌های بزرگ و استفاده از آن‌ها در تصمیم‌گیری‌های تجاری مهم اطلاق می‌شود [۶]؛

پ) داده‌کاوی، مجموعه‌ای از روش‌ها در فرایند کشف دانش است که برای تشخیص الگوها و رابطه‌های نامعلوم در داده‌ها مورد استفاده قرار می‌گیرد [۷] و [۸]؛

ت) فرایند کشف الگوهای مفید از داده‌ها را داده‌کاوی می‌گویند [۹].

مشاهده می‌شود که هر کس بنا به کاربرد و موارد استفاده تعریفی از داده‌کاوی ارائه کرده است.

پایه و اساس داده‌کاوی در سه شاخه قدیمی ریشه دارد که مهم‌ترین آن‌ها آمار کلاسیک است. بدون آمار، داده‌کاوی وجود نخواهد داشت. زیرا آمار زیربنای بیش‌تر فناوری‌هایی است که داده‌کاوی بر اساس آن‌ها بنا شده است.

دومین شاخه مرتبط با داده‌کاوی، هوش مصنوعی (AI) است. این شاخه بر اساس اکتشاف ساخته شده و سعی دارد پردازش‌هایی شبیه افکار انسان را در مسائل آماری به کار برد.

سومین شاخه مرتبط با داده‌کاوی، یادگیری ماشین است که تلفیقی از آمار و هوش مصنوعی است. یادگیری ماشین می‌تواند به‌عنوان هوش مصنوعی تکامل یافته مطرح شود. یادگیری ماشین سعی دارد به برنامه‌های رایانه‌ای این امکان را بدهد تا در مورد اطلاعاتی که به آن‌ها داده می‌شود، یاد بگیرند تا چنین برنامه‌هایی بتوانند متناظر با اطلاعات متفاوتی که به آن‌ها داده می‌شوند تصمیم‌گیری‌های متفاوتی انجام دهند.

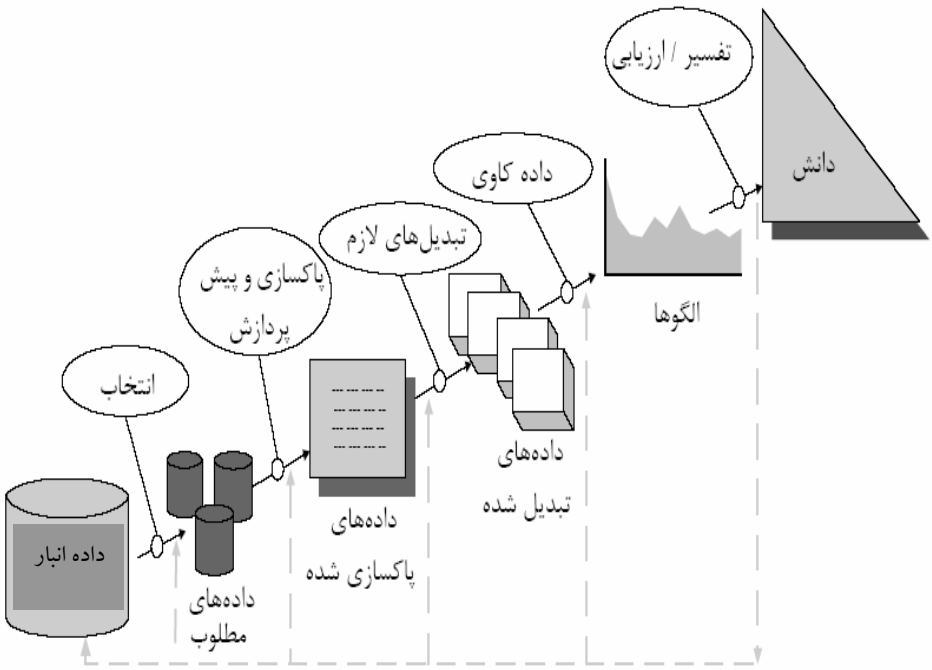
با آن‌چه گفته شد، نتیجه می‌شود که بهترین تعریف برای داده‌کاوی را می‌توان تلفیق پیشرفت‌های قدیمی و جدید علم آمار، هوش مصنوعی و یادگیری ماشین دانست. این فنون برای تحلیل داده‌ها و یافتن رابطه‌هایی که با روش‌های دیگر قابل یافتن نیستند، استفاده می‌شود [۱۰].

۳- داده‌کاوی، کشف دانش از پایگاه داده‌ها (KDD) و مراحل داده‌کاوی

در متون مربوط به داده‌کاوی دو تعبیر مختلف از داده‌کاوی وجود دارد. برخی مؤلفان مانند چتفیلد، داده‌کاوی را مترادف عبارت کشف دانش از پایگاه داده‌ها می‌دانند [۱۱]. گروه دوم از جمله فایاد، داده‌کاوی را به‌عنوان یک مرحله ضروری از فرایند بزرگتر کشف دانش از پایگاه داده‌ها (KDD) در نظر می‌گیرند [۴].

بسیاری از نویسندگان فرایند KDD را شامل ۹ مرحله دانسته‌اند و شکل زیر را مرجعی برای توصیف آن می‌دانند [۱۲] و [۱۳]. همان‌گونه که در شکل پیداست این فرایند دارای مراحل مختلفی است که داده‌کاوی یکی از این مراحل است.

داده‌کاوی نیز مانند KDD دارای مراحل مختلفی است. به‌طور کلی می‌توان مراحل



مراحل فرایند کشف دانش از پایگاه داده‌ها

مختلف داده‌کاوی را به صورت ده بند زیر بیان کرد (برای اطلاعات بیش‌تر بـه [۲] رجـوع شود).

الف) تعیین موضوع؛
ب) انتخاب داده‌ها؛
پ) آماده کردن داده‌ها؛
ت) بازرسی داده‌ها؛
ث) انتخاب ابزارها؛
ج) قالب پاسخ؛
چ) طرح‌ریزی مدل؛
ح) ارزیابی یافته‌ها؛
خ) ارائه یافته‌ها؛
د) هماهنگ کردن پاسخ‌ها.

۶- کاربرد داده‌کاوی در آمار رسمی

آن‌چه امروزه اهمیت بسیاری پیدا کرده کمبود یا نبود داده‌های مورد نیاز نیست بلکه کمبود یا نبود روش‌هایی مناسب و استاندارد به‌منظور نگهداری، بـه روزکـردن، در دسـترس قـرار دادن و در حالت آرمانی‌تر، کشف دانش جدید از داده‌های موجود است. یکی از راهکارهای پیشنهادی برای حصول به این هدف، استفاده از سیستم‌های داده‌کاوی است. سیستم‌هـای داده‌کاوی این امکان را به کاربر می‌دهند که بتواند انبوه داده‌های جمع‌آوری شده را تفسیر و الگوها و دانش (اطلاعات) نهفته در آن را استخراج نماید.

از آن‌جا که مراکز و موسسه‌های آماری با داده‌انبارهای حجیم و با ابعاد زیادی مواجـه هستند لذا نیاز بـه اسـتفاده از ایـن فـن ضـروری بـه‌نظـر مـی‌رسـد. در سـال‌هـای اخیـر پژوهش‌های مختلف و وسیعی در زمینه کاربرد داده‌کاوی در آمارهای رسمی صورت گرفته و چندین کارگاه با عنوان کاوش داده‌های رسمی (اولین کارگاه در سال ۲۰۰۲ در فنلانـد و دومین در سال ۲۰۰۴ در ایتالیا) برگزار شده است [۱۴] و [۱۵] و نشان‌دهنده آن است کـه

این موضوع مورد توجه سازمان‌های ملی آمار، محققان آماری و سایر کاربران قرار گرفته است و بر این نکته نیز تاکید دارند که کاربرد داده‌کاوی در آمارهای رسمی موضوع جدید و نوپایی است که در مرحله تکامل قرار دارد و باید سازمان‌های ملی آمار بر استفاده از این ابزار پرتوان تاکید کنند. به‌عنوان مثال کیفیت که امروزه یکی از معیارهای مهم در ارزیابی انتشارات سازمان‌های ملی و بین‌المللی آماری است مورد توجه داده‌کاوان و محققان در این زمینه‌ها قرار گرفته است [۱۶] و [۱۷].

قابلیت و توانایی استفاده از داده‌کاوی امروزه به موضوع مهمی در بین محققان و پژوهشگران در حوزه آمار رسمی تبدیل شده است. تعداد تحلیل‌ها در مورد استفاده از فنون داده‌کاوی که در آن‌ها مدل یا الگو در مجموعه داده‌های آمار رسمی با استفاده از این فن کشف شده باشد تا چندی پیش اندک بودند. البته شایان ذکر است از گذشته نیز موسسه‌ها و مراکز آماری بدون در نظر گرفتن نام داده‌کاوی به نوعی از این فن یا قسمت‌هایی از آن نظیر تحلیل اکتشافی داده‌ها یا الگوریتم‌های انتخاب مدل استفاده کرده‌اند اما نه با وسعت کاربرد داده‌کاوی‌ای که امروزه شاهد آن هستیم.

برای انجام داده‌کاوی صحیح و موفق باید موارد مهمی را در نظر گرفت که به‌طور کلی پیش‌تر به ده مورد آن اشاره شد. اما برای انجام کاربردهای موفق داده‌کاوی در زمینه آمار رسمی باید به موارد زیر نیز توجه خاصی داشت:

- داده‌های انبوهه شده

موسسه‌های آماری در خصوص گردآوری داده‌های حاصل از عملیات میدانی هزینه‌های مالی و زمانی زیادی را مصرف می‌کنند اما آن‌ها تنها سازمانی نیستند که داده‌ها را تحلیل یا استفاده می‌کنند بلکه همان‌طور که پیش‌تر اشاره شد تحلیل داده‌ها معمولاً توسط موسسه‌ها یا سازمان‌های دیگری نیز انجام می‌شود یا این‌که حد اقل سازمان‌ها، اشخاص مختلف و پژوهش‌گران از این اطلاعات استفاده می‌کنند. اما بر اساس قانون، موسسه‌ها یا سازمان‌های آماری مجاز به انتشار و در اختیار قرار دادن پاسخ افراد به سازمان‌ها، ارگان‌ها، موسسه‌های تجاری و هر مصرف‌کننده دیگری نیستند. لذا آنان این داده‌ها را پیش از انتشار به‌منظور حفظ محرمانگی به صورت انبوهه ارائه می‌کنند. بنا بر

این تحلیلگران داده‌ها با داده‌های کلان مواجه هستند نه با داده‌های خرد (که معمولاً در آمار کلاسیک از آن‌ها استفاده می‌شود). تحلیل داده‌های کلان و توسعه روش‌های کلاسیک آماری برای تحلیل چنین داده‌هایی موضوع جدیدی را تحت عنوان تحلیل داده‌های نمادین به وجود آورده است [۱۸] و [۱۹].

- کیفیت داده

مقوله اهمیت کیفیت در آمار رسمی و بحث در مورد کاربرد داده‌کاوی در کیفیت داده به نوشتار دیگری احتیاج دارد و فقط هدف از مطرح کردن آن در این قسمت، یادآوری موضوع آن است. مفاهیم مختلفی در کیفیت داده مطرح است که در این‌جا به آن‌ها نمی‌پردازیم. به‌طور کلی می‌توان کاربرد داده‌کاوی را در کیفیت داده به دو قسمت عمده اندازه‌گیری و بهبود یا ارتقای کیفیت داده تقسیم کرد. در واقع هدف اصلی از به‌کارگیری داده‌کاوی کشف، تعیین کمی، تشریح و تصحیح موارد و خطاها در پایگاه‌های داده‌ای حجیم است. در این خصوص روش‌های زیادی مطرح است از جمله تعیین و شناسایی نقاط پرت، گمشده، موثر و چگونگی رفتار با آن‌ها، روش‌های خوشه‌بندی، تحلیل‌های وابستگی، شبکه‌های عصبی، هوش مصنوعی و بسیاری از روش‌های دیگر که به ذکر این چند مورد بسنده می‌کنیم [۱۶] و [۲۰].

- بهنگامی

این موضوع را می‌توان یکی از مولفه‌های اصلی کیفیت داده در نظر گرفت. در حال حاضر موسسه‌های دولتی و خصوصی در تلاشند تا زمان گردآوری داده‌ها و زمانی که تصمیم‌ها بر اساس نماگرهای آماری منتشره از این داده‌ها اتخاذ می‌شود را کاهش دهند. یک مثال از این مورد اندازه‌گیری نرخ بهره محاسبه شده توسط اداره آمار اروپا و تصمیم اتخاذ شده توسط بانک مرکزی اروپا برای تعیین نرخ مالیات است. داشتن برنامه، الگوریتم یا فنونی که فاصله این دو زمان را کاهش دهد همواره مد نظر بوده است. به‌طور قطع برنامه‌های خودکار با قابلیت هوشمند بودن و به روز بودن که بتواند این فاصله زمانی را کاهش دهد

نیز مد نظر است، به‌طوری که داده‌کاوی یکی از پاسخ‌های موجود به این مشکل مهم است [۱۴] و [۱۵].

- محرمانگی

به‌نظر می‌رسد که داده‌کاوی با حفظ محرمانگی آمارهای رسمی متناقض باشد، اما این‌گونه نیست. داده‌کاوی در پی یافتن روابط آماری و الگوها در داده‌ها است نه در بین اطلاعات فردی اشخاص و جزئیات اطلاعات آن‌ها. بنا بر این باید متذکر شد که داده‌کاوی تشخیص الگو است و نه تشخیص اشخاص. اصل محرمانگی یکی از اصول آمار رسمی است و این موضوع در تمامی آمارگیری‌ها برای پاسخگویان بیان شده و به طرق مختلف اطمینان پاسخگو را در محرمانه نگاه داشتن اطلاعات فردی وی تضمین می‌کند. بر همین اساس موضوع محرمانگی و روش‌های افشای اطلاعات به‌منظور استفاده بهینه اطلاعات با در نظر گرفتن اصول اولیه مورد نظر محققان بوده است و امروزه چگونگی استفاده از داده‌کاوی در این خصوص نقش مهمی را ایفا می‌کند [۲۱].

- فراداده

امروزه دیگر برای بیان یک نتیجه نمی‌توان تنها به یک مجموعه داده‌ها اشاره کرد به خصوص در ابعاد زیاد و مفاهیم کلان. کاوش داده‌های رسمی از منابع اطلاعاتی مختلف یا منابع انتشاراتی متفاوت صورت می‌گیرد و همین امر باعث شده تا علاوه بر اطلاعات موجود در پایگاه‌های داده‌ای مختلف به اطلاعات بیش‌تری در این خصوص که مرتبط با آن باشد و در بیان، نتیجه‌گیری و دقت اطلاعات کمک کند، نیاز باشد. در این باره باید گفت که وجود فراداده به بالا بردن کیفیت نتیجه داده‌کاوی کمک کند [۲۲].

موضوع کاربرد داده‌کاوی در آمار رسمی چنان مورد توجه مراکز و موسسه‌های آماری مختلف بوده که چندین پروژه مهم، پرهزینه و با مدت زمان نسبتاً طولانی در این زمینه انجام شده و در حال انجام است. بسیاری از این

پروژه‌ها توسط اتحادیه آمار اروپا انجام شده و هدف آن‌ها ایجاد نرم‌افزاری ویـژه و کارامد با استفاده از ابزار داده‌کاوی با قابلیت استفاده در آمار رسمی اسـت. در ایـن خصوص می‌توان به پروژه‌های زیر اشاره کرد:

- ASSO ،SODAS

ASSO یا به عبارتی تحلیل سیستم داده‌های رسمی در ژانویه سال ۲۰۰۱ برای مدت ۳ سال به اجرا در آمـد. ایـن پـروژه روش‌هـا، روش شـناختی و ابزارهـای نرم‌افزاری را برای تحلیل داده‌های مختلط چند بعـدی (عـددی و غیـر عـددی) حاصل از پایگاه‌های داده‌ای مختلف ارائه می‌کند. نتیجه این پروژه، نـرم‌افـزاری است که SODAS۲ نامیده شده است. در واقع این نرم‌افزار توسعه‌یافته نرم‌افزار SODAS است [۲۳].

- KESO

KESO یا به عبارتی استخراج دانش برای مراکز آماری پروژه‌ای تحت حمایـت اداره آمار اروپا بود که در ژانویه سال ۱۹۹۹ به مدت ۳ سال انجام شد. هدف این پروژه ایجاد سیستم داده‌کاوی کارا است که احتیاج‌های مراکز آماری را در تهیـه و استفاده پایگاه‌های داده‌ای مراکز آماری بر طرف سازد [۲۴].

- SPIN

SPIN یا سیستم داده‌کاوی فضایی در ژانویـه سـال ۲۰۰۰ شـروع شـد. هـدف اصلی این پروژه آن بود که مراکز آماری را در ارائه بهنگام اطلاعات آماری با در نظر گرفتن کمترین هزینه یاری نماید. همچنین امکانات جدیدی را برای تحلیل داده‌های GIS فراهم نمود [۲۵].

موارد مختلف و متفاوتی را می‌توان در مورد داده‌کـاوی و کـاربرد آن در آمـار رسـمی بیان کرد. در ادامه به چندین مورد از کاربردهای داده‌کـاوی کـه در مقالـه‌هـا، کتـاب‌هـا و همایش‌های مختلف در مورد آن پژوهش صورت گرفته و با آمار رسمی مرتبط است، اشاره شده است:

- کاوش داده‌های سرشماری

- کاوش داده‌های انباشته شده
- کاربرد داده‌کاوی به‌منظور کنترل کیفیت در جمع‌آوری و انتقال داده‌ها
- الگوریتم داده‌کاوی بهنگام برای بهنگام کردن انتقال آمارهای رسمی
- فراداده‌ها و داده‌کاوی
- فنون داده‌کاوی به‌منظور تشخیص داده‌های پرت
- حفظ محرمانگی در داده‌کاوی
- فنون داده‌کاوی برای مقایسه کیفی آمار
- توصیف منابع تولیدکننده داده‌های رسمی و مسائل داده‌کاوی مرتبط
- داده‌کاوی فضایی آمارهای رسمی
- داده‌کاوی و تاثیر آن بر کیفیت داده‌ها
- فنون داده‌کاوی در براورد مشاهده‌های گم شده

۷- نتیجه‌گیری

داده‌کاوی یک رشته نسبتاً جدید علمی است که از انجام پژوهش‌ها، حد اقل در رشته‌های مختلف آمار، یادگیری ماشین، علوم رایانه به‌خصوص مدیریت پایگاه داده‌ها شکل گرفته است. فرایند خودکار کشف دانش موضوعی است که در مراکز آماری روز به روز اهمیت ویژه‌ای پیدا کرده است به‌طوری که بیان می‌شود داشتن نرم‌افزارهای کارا با تجربیات انسانی و مداخله آن‌ها در امر تحلیل مطلوب نیست و باید رفته رفته به نرم‌افزارهای خودکار و هوشمند با حد اقل مداخله انسانی روی آورد. اکنون داده‌کاوی شاخه‌ای در آمار رسمی دارد یا شاید بهتر است گفته شود آمار رسمی از داده‌کاوی به‌عنوان ابزاری پرتوان در راستای اهداف خود استفاده می‌کند و جزو جدایی‌ناپذیر آن محسوب می‌شود. این موضوع به‌عنوان یکی از مباحث مهم در آمار رسمی تبدیل شده و اهمیت آن روز به روز بیش‌تر شده و خواهد شد. برگزاری سمینارهای اختصاصی در زمینه آمار رسمی با دیدگاه داده‌کاوی و پروژه‌های چندساله و پرهزینه و سرمایه‌گذاری‌های مراکز و موسسه‌های معتبر آماری تاکیدی بر این ادعا است. در واقع به جرأت می‌توان گفت با وجود داده‌انبارهای حجیم که روز به روز بر تعداد و حجم آن‌ها افزوده می‌شود و روش‌های نوین جمع‌آوری

داده‌ها، اگر مراکز آماری به این مهم مجهز نباشند با مسائل و مشکلات فراوانی در نگهداری، بهنگام کردن اطلاعات، بهبود کیفیت، تفسیر و انتشار اطلاعات رو به رو خواهند شد.

مرجع‌ها

[۱] حائری مهریزی، علی‌اصغر و نواب‌پور، حمیدرضا. (۱۳۸۱). آشنایی با داده‌کاوی. ششمین کنفرانس بین‌المللی آمار ایران. دانشگاه تربیت مدرس.

[۲] حائری مهریزی، علی‌اصغر. (۱۳۸۲). داده‌کاوی: مفاهیم و روش‌ها و کاربردها. پایان‌نامه کارشناسی ارشد، دانشگاه علامه طباطبایی.

[۳] ناظمی، عبدالرضا. (۱۳۸۳). رده‌بندی و داده‌کاوی. پایان‌نامه کارشناسی ارشد، دانشگاه فردوسی مشهد.

[٤] Fayyad, U.M., Piatetsky-Shapiro , Smyth P., and Uthurusamy R. (eds.) (۱۹۹٦). Advances in Knowledge Discovery and Data Mining. Menlo Park, California: AAAI Press.

[۵] Han, J., and Kamber. M. (۲۰۰۱). Data Mining: Concept and Techniques. Morgan Kaufmann.

[٦] Jerome, H. Data Mining and Statistics: What's the Connection? URL: http://stat.stanford.edu/~jhf/dm-stat.ps.Z.

[۷] Hand, D., Mannila, M., and Padhraic, S. (۲۰۰۱). Principle of Data Mining. MIT Press.

[۸] Hand, D. J. (۲۰۰۰). Why data mining is more than statistics writ large. Statistical aspects of data mining and knowledge discovery in databases. Topic ۳٦. ٤۳۳-٤۳٦.

[۹] George, H. J. (۱۹۹۷). Enhancements to the Data Mining Process. Ph.D. Thesis, Department of Computer Science, Stanford University.

[۱۰] Hastie, T., Tibshirani, R., and Friedman, J. (۲۰۰۱). The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer-Verlag,.

[۱۱] Chatfield, C. (۱۹۹۵). Model uncertainty, data mining and statistical inference. J. R. Statist. Soc. A. ۱۵۸(۳). ٤۱۹-٤٦٦.

[۱۲] Berry, M., and Lindoff, G. (۱۹۹۷). Mastering Data Mining. John Wiley & Sons.

[۱۳] Kantardiz, M. (۲۰۰۳). Data Mining: Concepts, Models, Methods, and Algorithms, John Wiley & Sons.

[۱۴] Workshop۱: Helsinki. (۲۰۰۲). Mining Official data. ٦th European Conference on Principal and Practice of Knowledge Discovery in database, Finland.

[۱۵] Workshop۲: Mining Official Data. ۳- ٤, June ۲۰۰٤. University of Bari, Italy.

[۱۶] Hassani, H., and Anari, M. (۲۰۰۵). Using Data Mining for Data Quality Improvement. ISI Conferences. Proceeding.

[۱۷] Saporta. G. (۲۰۰۰). Data Mining and Official Statistics. Quinta Conferenza Natuionale de Statistica, ISTAT,Roma.

[۱۸] Bock, H.H., and Diday. E (eds). (۲۰۰۰). Analysis of Symbolic Data. Exploration Methods for Extracting Statistical Information for Complex data. Springer-Verlag.

[۱۹] Klosgen, W., and May, M. (۲۰۰۲). Census Data Mining: An Application. Working Paper.

[۲۰] Hipp, J., Guntzer, U., and Grimmer. U. (۲۰۰۲). Data Quality Mining

[۲۱] Nanopoulos, Ph., and King, J. (۲۰۰۲). Important Issues on Statistical confidentiality. In proceeding of Mining official Data Workshop. Helsinki University.

[۲۲] D'Angiolini, G. (۲۰۰۲). Developing a Metadata Infrastructure for Official data: the ISTAT experience. In proceeding of Mining Official Data workshop. Helsinki University.

[۲۳] ASSO Project Analysis System of Symbolic Official data, http://www.info.fundp.ac.be/asso/

[۲٤] KESO (Knowledge Extraction for Statistical Offices)

[۲۵] http://db.cwi.nl/projecten/project.php٤?prjnr=۷۷
[۲٦] SPIN! (Spatial Mining for Data of Public Interest:
[۲۷] http://www.ais.fraunhofer.de/KD/SPIN/index.html.