

امنیت داده ها در داده کاوی

مؤلف :

مهندس داراب زینی

کارشناسی ارشد مهندسی کامپیوتر (نرم افزار)

مدرس دانشگاه

www.ketab.ir



انتشارات راز تهران

۱۴۰۰

سرشناسه	: زینی، داراب، ۱۳۶۰-
عنوان و نام پدیدآور	: امنیت داده‌ها در داده کاوی/مؤلف داراب زینی.
مشخصات نشر	: تهران: رازنهان، ۱۴۰۰
مشخصات ظاهری	: ۴۵ص-
شابک	: 978-622-278-170-5
وضعیت فهرست نویسی	: فیپا
یادداشت	: کتابنامه: ص. ۴۳ - ۴۵.
موضوع	: داده کاوی -- تدابیر ایمنی
موضوع	: Data mining -- Security measures
موضوع	: حفاظت داده‌ها
موضوع	: Data protection
رده بندی کنگره	: QA۷۶/۹
رده بندی دیویی	: ۰۰۶/۳۱۲
شماره کتابشناسی ملی	: ۷۶۳۵۴۵۴
وضعیت رکورد	: فیپا

مشتاقان میدان انقلاب، اول جمالزاده جنوبی، کوچه دانشور، شماره ۱۲، واحد ۲، شماره تماس: ۰۲۱۶۶۹۰۱۸۴۱ - ۰۹۱۲۲۲۳۰۲۶۹
 پست الکترونیک: raznahanbook@gmail.com
 نشانی سایت: www.rznahanbook.com



که عنوان کتاب.....	امنیت داده ها در داده کاوی
که مؤلف.....	مهندس داراب زینی
که صفحه آرا.....	عاطفه نیک فر
که طراح جلد.....	عاطفه نیک فر
که ناشر.....	رازنهان
که شمارگان.....	۱۰۰۰ نسخه
که نوبت چاپ.....	اول ۱۴۰۰
که قیمت.....	۲۵۰۰۰ تومان
ISBN: 978-622-278-170-5	

فهرست

۷	 مقدمه
۹	 پیشگفتار
۱۰	 داده‌کاوی پیش‌بینی کننده یا یادگیری با نظارت
۱۰	 داده‌کاوی توصیف‌کننده یا یادگیری بدون نظارت
۱۱	 مبانی نظری
۱۱	 معرفی فرایند داده‌کاوی
۱۶	 معیارهای اندازه‌گیری سطح حریم
۱۷	 حریم خصوصی داده‌ها
۱۷	 تکنیک‌های بی‌نام کردن (K - Anonymization)
۱۸	 تکنیک‌های مبتنی بر کنترل افشاء آماری
۱۸	 تکنیک‌های مبتنی بر رمزنگاری
۱۹	 معیارهای انتخاب صفت خاصه
۱۹	 معیار به دست آوردن اطلاعات
۳۱	 ماشین بردار پشتیبان (SVM):

۲۴	 بهینه سازی حداقل بهینه
۲۵	 ماشین بردار اصلی
۲۵	 ماشین بردار کاهش
۲۵	 ماشین بردار ارتباط
۲۶	 ماشین بردار پشتیبانی دوقلو
۲۶	 الگوریتم های مبهم سازی
۲۶	 الگوریتم MSI-LNSI
۲۷	 مراحل الگوریتم MSI- LNSI
۲۸	 مثالی از الگوریتم MSI-LNSI
۳۱	 ارزیابی الگوریتم MSI-LNSI
۳۲	 مدل حفاظتی k-Anomity
۳۴	 عمومی سازی داده ها به منظور جلوگیری از استنتاج
۳۶	 رتبه بندی ویژگی ها
۳۹	 سخن پایانی
۴۱	 منابع

مقدمه

داده کاوی فرآیند استخراج الگوهای معتبر از داده‌ها است. دانش داده کاوی یکی از ده دانش در حال توسعه‌ای است که دهه آینده را با انقلاب جدیدی مواجه خواهد ساخت و بدین رو در سال‌های اخیر در دنیا گسترش فوق‌العاده سریعی داشته است. دانش داده کاوی با برخورداری از دامنه وسیع در زمینه‌های تخصصی با توصیف، تشریح، پیش‌بینی و کنترل پدیده‌های گوناگون، امروزه دارای کاربرد بسیار وسیع در حوزه‌های مختلف از جمله صنعتی، پزشکی، ارتباطات، کشاورزی، انرژی، علوم اجتماعی، فرهنگی، سیاسی، اقتصادی، بازرگانی، نظامی، آموزشی و ... است، به گونه‌ای که امروزه مرز و محدودیتی برای کاربرد این دانش در نظر گرفته نشده و زمینه‌های کاری این دانش را در تمامی عرصه‌های برخوردار از داده می‌دانند [۱].

امروزه در عصر اطلاعات با افزایش استفاده از فناوری‌های نوین حجم بالایی از داده‌های شخصی افراد نظیر اطلاعات اعتباری، عادات خرید، تاریخچه پزشکی و ... جمع‌آوری شده و باگذشت زمان نیز میزان آن‌ها روز به روز در حال افزایش هست. از این رو استفاده از راهکارهای مناسب برای طبقه‌بندی و تولید اطلاعات از میان انبوهی از داده‌ها امری ضروری و حیاتی است. در این وضعیت استفاده از فناوری‌هایی نظیر داده کاوی، به عنوان مکانیزمی برای استخراج دانش پنهان از حجم زیاد داده‌ها در قالب الگوها و مدل‌های مختلف، بیش از پیش ضروری به نظر می‌رسد. اما تحلیل داده‌ها با استفاده از تکنیک‌های داده کاوی زمانی به عنوان یک تهدید واقعی محسوب می‌گردد که به قابلیت استخراج دانش از داده‌ها توسط این تکنیک‌ها که صاحب داده از آن بی‌اطلاع است، توجه شود. در چنین وضعیتی به منظور مواجه با این مسئله صاحبان داده یا از اشتراک داده‌ها اجتناب می‌ورزند، یا بخش‌هایی را به اشتراک می‌گذارند که استخراج دانش را با مشکل روبرو می‌نماید [۱].

علی‌رغم مزیت‌های زیاد داده کاوی، این فرآیند با مسائل اخلاقی من جمله حریم خصوصی، حقوق مالکیت معنوی، امنیت داده و ... روبرو هست. امنیت به عنوان حق افراد در تعیین زمان، چگونگی و میزان استفاده از اطلاعات خود توسط دیگران، تعریف می‌شود. حریم خصوصی، یافتن مصالحه بین حق اجتماعی برای دانستن و حق انفرادی برای زندگی خصوصی تعریف می‌گردد.

امنیت داده و حریم خصوصی، اگرچه اغلب به جای یکدیگر استفاده می‌شوند، اما نشان‌دهنده دو جنبه متفاوت از حفاظت داده هستند و برای آن‌ها فنون مختلفی توسعه یافته است. در اینجا هدف یافتن یک مصالحه بین دو هدف در تعارض است: سودمندی حاصل از داده کاوی و حفظ حریم خصوصی. تلاش‌ها در جهت یافتن راه‌حلی برای این دو هدف منجر به ایجاد شاخه جدیدی در داده کاوی بانام داده کاوی با حفظ حریم خصوصی گردید. هدف اصلی داده کاوی با حفظ حریم خصوصی، توسعه الگوریتم‌هایی برای تغییر داده است، به گونه‌ای که داده‌های

خصوصی قبل و بعد از انجام داده کاوی و حتی دانش خصوصی پس از داده کاوی خصوصی بمانند. در این وضعیت ارائه روش هایی که حریم خصوصی افراد را در حین فرآیند داده کاوی تضمین می نمایند، ضروری به نظر می رسد [۱].

داده کاوی فرآیند انتخاب، شناسایی و مدل کردن کمیت های عظیمی از داده ها برای کشف و استخراج قوانین و ارتباط موجود بین آن ها است، تا از این قوانین و ارتباط به اطلاعات و دانشی از مجموعه داده ها برسیم. کاربردهای داده کاوی شامل موارد زیر است:

قوانین وابستگی: الگوهایی که در آن وجود یک آیتم دلالت بر وجود آیتم دیگر دارد.

کلاس بندی: انتساب الگوها به یک مجموعه کوچک از کلاس های از قبل تعریف شده به وسیله کشف بعضی روابط بین ویژگی ها.

خوشه بندی: گروه بندی داده ها یا مجموعه الگوهایی که ویژگی های مشابهی دارند.

پیشگویی: کشف الگوها برای پیشگویی منطقی درباره آینده [۲].

کاربرد داده کاوی در بخش های مختلف خصوصی و عمومی رو به رشد است و عموماً صناعی چون بانک، بیمه، پزشکی و خرده فروشی، از داده کاوی به هدف کاهش هزینه ها، افزایش تحقیقات و افزایش فروش استفاده می کنند.

در کنار تأثیرات مثبت داده کاوی، احتمال افشای اطلاعات خصوصی نیز وجود دارد. هنوز هیچ تضمینی ارائه نشده است که بتوان داده ها را بدون تجاوز به حریم خصوصی مالک آن، مورد داده کاوی قرار داد. به عنوان مثال در شبکه های اجتماعی اطلاعات مختلف کاربران از جمله سن، حرفه، هویت و سایر اطلاعاتی که باعث شناسایی دقیق فرد می شود نیاز به محافظت دارند. در یک سیستم پزشکی، نحوه انجام داده کاوی در اطلاعات خصوصی بیماران بدون افشای اطلاعات محرمانه، یکی از مسائلی است که با آن روبه رو هستیم و یا دستگاه های جمع آوری داده به صورت آنلاین، نمونه ای از ده ها برنامه جدیدی هستند که حریم شخصی افراد را تهدید می کنند. مشکل اصلی از آنجا ناشی می شود که چگونه می توان هم حریم خصوصی افراد را در نظر گرفت و هم از نتایج مفید دستگاه های داده کاوی بهره برد.

برای برطرف کردن موانع موجود در این زمینه، تحقیقات زیادی در حال انجام است. روش هایی که برای حفظ محرمانگی داده ها در داده کاوی معرفی شده اند، هر کدام دارای ویژگی ها، مزایا و معایب خاصی نظیر از دست دادن داده ها هستند که ممکن است در کیفیت دانش استخراج شده از داده ها اثر گذارد تا جایی که داده ها غیر قابل استفاده شوند. در نتیجه لازم است تا از داده های محرمانه تا حدی که سودمندی آن ها حفظ گردد، محافظت نماییم [۳].