ISSN: 2745-3673
http://www.Rcsj.ir

مجلـه‌ی علمی پژوهش در علوم رایانه (سال سوم)
شماره ۱۲ / زمستان ۱۳۹۷ / ص ۷۳-۶۴

مروری بر الگوریتم‌های حفظ حریم‌خصوصی در داده‌کاوی

توحید ساکت
گروه کامپیوتر، واحد میانه، دانشگاه آزاد اسلامی، میانه، ایران.

نام نویسنده مسئول:
توحید ساکت

چکیده
شرکت‌ها و سازمان‌های بسیاری با هدف پیشرفت در زمینه‌های گوناگون، نیاز به استخراج دانش مطلوب و مفید از اطلاعات خود دارند. تکنیک‌های داده‌کاوی متعددی جهت استخراج دانش پنهان از پایگاه داده‌های بزرگ معرفی شده است که قوانین انجمنی یکی از پرکاربردترین آنها می‌باشد. اشتراک‌گذاری مکرر داده‌ها جهت اعمال تکنیک‌های داده‌کاوی و همچنین امکان استخراج اطلاعات حساس و محرمانه از این داده و نگرانی مالکان داده از افشاء اطلاعات خصوصی، منجر به معرفی حوزه جدیدی تحت عنوان حفظ حریم خصوصی در داده‌کاوی گردید. بهره‌گیری از روش‌های حفظ حریم‌خصوصی و ایمن‌سازی داده، علاوه بر کشف الگوریتم‌های مناسب، احتمال کشف اطلاعات حساس و محرمانه را به حداقل می‌رساند. در این مقاله به بررسی الگوریتم‌های مربوط به حفظ حریم‌خصوصی در انتشار داده‌ها خواهیم پرداخت. هدف اصلی این است که مزایا و معایب الگوریتم‌های موجود را مشخص کرده و مقایسه‌ای از آنها را ارائه دهیم تا مالکان داده منبعی به منظور انتخاب الگوریتم‌های مناسب برای حوزه کاری خود در اختیار داشته باشند.

واژگان کلیدی: حریم خصوصی، داده‌کاوی، الگوریتم Rock، الگوریتم max-mix.

مجلـه‌ی علمی پژوهش در علوم رایانه ، شماره ۱۲ ، زمستان ۱۳۹۷، ص ۷۳- ۶۴

ISSN: 2745-3673
http://www.Rcsj.ir

مقدمه

حریم خصوصی در فضای مجازی از چالش‌هایی است که عدم پایبندی عده‌ای از کاربرانِ عمدتا جوان به برخی اصول، و یا خوشبینی گروه‌ها و افرادی از مخاطبان شبکه‌های اجتماعی در آن باعث بروز فاجعه‌ای هولناک می‌شود که حیثیت و آبروی افراد را بازیچه قرار می‌دهد. در طول سالیان اخیر، فضای مجازی و دنیای ارتباطات با چنان رشد شتابنده و فزاینده‌ای همراه شده که امروزه کمتر کسی را می‌بینیم که به نوعی پیوندی با چنین فناوری نداشته باشد. در سایه رقابت سنگین و پایاپای نرم افزارها و فناوری‌های ارتباطی در شبکه‌های اجتماعی جمع کثیری از افراد و گروه‌های جامعه از اقصی نقاط کشور با فرهنگ‌ها، بینش‌ها، روش‌ها و منش‌های متنوع و متکثر، جذب فضاهای مذکور شده‌اند. در کنار ثمرات و مزایای فناوری مورد اشاره همچون تسهیل ارتباطات، جابجایی و گردش اطلاعات، معایب و عوارض ناخوشایندی نیز با چنین فناوری‌هایی همراه است که هر از چندگاهی مسبب چالش‌های بسیاری می‌شود. سئوال مطرح در این فضا و بستر (دنیای مجازی) معطوف به این موضوع می‌شود که براستی چه انگیزه و علتی در عقبه حسن کنجکاوی شومی که آبرو و حیثیت افراد را بر باد می‌دهد وجود دارد؟ دیگرانی که مانند بسیاری از افراد جامعه، حق زندگی دارند و همچون ما و تمامی افراد و آحاد جامعه حریم خصوصی دارند و حفظ و صیانت از آن حق طبیعی آنها به شمار می‌رود. براستی بر اساس کدام فرآیند و تغییر رفتاری و اجتماعی، اینچنین حریم خصوصی دیگران برای برخی از طیف‌ها و گروه‌های اجتماعی جذاب شده و حاضریم این جذابیت نامیمون را با دیگران نیز سهیم شویم [۱].

حفظ حریم خصوصی در داده کاوی به عنوان یک پیش نیاز مطلق برای تبادل اطلاعات محرمانه از نظر تحلیل داده‌ها، اعتبار سنجی، و انتشار پدید آمده است. احساسات و مجادلات مشکوک، عدم تمایل ارائه دهندگان مختلف اطلاعات را به سمت حفاظت از قابلیت اطمینان داده‌ها از افشا شدن می‌کشاند که اغلب منجر به رد مطلق در اشتراک گذاری داده‌ها و یا به اشتراک گذاری اطلاعات نادرست می‌شود. تکنیک‌های فعلی حفظ حریم خصوصی در داده کاوی، بر اساس، قانون انجمنی (اجتماع)، قانون انجمنی پنهان، طبقه بندی، خوشه بندی، طبقه بندی انجمنی، داده کاوی برون سپاری، توزیع شده، و k-anonymity طبقه بندی می‌شوند. روش‌های پیشرفته حفظ حریم خصوصی در داده کاوی همیشه در حال تقاضا برای تبادل اطلاعات امن و قابل اعتماد از طریق اینترنت هستند. افزایش چشمگیر ذخیره سازی اطلاعات شخصی مشتریان، منجر به پیچیدگی فزاینده‌ی الگوریتم‌های داده کاوی شده که تاثیر قابل توجهی در به اشتراک گذاری اطلاعات گذاشته است. در میان چندین الگوریتم موجود، حفظ حریم خصوصی در داده کاوی (PPDM[۱]) نتایج عالی در ارتباط با ادراک درونی حفظ حریم خصوصی و داده کاوی ارائه می‌دهد. حریم خصوصی باید از سه جنبه داده کاوی شامل قوانین انجمنی، طبقه‌بندی و خوشه‌بندی (sachan و همکارانش، ۲۰۱۳) محافظت کند. مشکلات پیش رو در داده کاوی به طور گسترده‌ای در بسیاری از جوامع مانند پایگاه داده، کنترل افشای آماری و جامعه رمزنگاری (نایاک و دوی، ۲۰۱۱) سنجیده می‌شوند. ظهور تکنولوژی جدید محاسبات ابری، به همکاران کسب و کار اجازه می‌دهد تا داده‌ها را به اشتراک بگذارند و اطلاعات را برای منافع متقابل عرضه کنند. همه این‌ها مربوط به قابلیت تجمعی برای ذخیره تک تک داده‌های کاربران همراه با پیچیدگی رو به رشد الگوریتم‌های داده کاوی است که بر روی تبادل اطلاعات تاثیر می‌گذارد [۳].

داده‌کاوی به عنوان یک فن آوری قابل توجه برای به دست آوردن دانش از مقادیر عظیمی از داده‌های استخراج شده است. با این حال، نگرانی استفاده از این تکنولوژی باعث نقض حریم خصوصی افراد شده است. این امر منجر به واکنش شدید در برابر تکنولوژی شده است. مثلا "قانون توقف داده کاوی" که توسط وزارت دفاع ایالات متحده معرفی شد که می‌توانست تمام برنامه‌های داده کاوی (از جمله تحقیق و توسعه) را ممنوع کند. در حالی که شاید بیش از حد شدید به عنوان یک مثال فرضی، خرابی تجهیزات و عدم بهبود برنامه‌های حفظ حریم خصوصی یک نگرانی واقعی است. این نگرانی به طور کلی بر حریم خصوصی اطلاعات، با استانداردهای همراه وجود دارد [۴].

با رشد سریع در اندازه و تعداد پایگاه داده‌ها، کاوش دانش، قوائد یا اطلاعات سطح بالا از داده ها به منظور پشتیبانی از تصمیم گیری‌ها و پیش بینی رفتارهای آتی ضروی به نظر می رسد. کاوش قوانین انجمنی یکی از وظایف مهم در داده کاوی، روند یافتن روابطی مابین خصیصه‌ها یا ما بین مقادیر آن ها در یک پایگاه داده بزرگ است که در جهت امر تصمیم گیری کمک ساز باشند. یافتن چنین روابطی داخل یک مجموعه وسیعی از داده ها به علت ماهیت نمایی آن کار ساده‌ای نیست [۶-۵]. در سال‌های اخیر، داده کاوی به دلیل گسترش داده‌های الکترونیکی که توسط شرکت‌های بزرگ تولید می‌شود به عنوان یک تهدید برای حریم خصوصی شده است. این امر منجر به افزایش نگرانی در مورد حفظ حریم خصوصی از داده‌های اساسی منجر شد. در سال های اخیر، تعدادی از تکنیک‌ها برای اصلاح و یا تبدیل داده‌ها برای حفظ حریم خصوصی ارائه شده است. مدل‌ها و الگوریتم‌ها برای بررسی برخی از تکنیک‌های مورد استفاده برای حفظ حریم خصوصی و حفظ داده‌ها ممکن است حفظ حریم خصوصی در داده کاوی را در بر داشته باشند [۷].

---

[۱] Privacy Preserving data mining

ISSN: 2745-3673
http://www.Rcsj.ir

سابقه

در سال ۲۰۱۳ Belwal و همکارانش اساس پشتیبانی و اعتماد قواعد حساس را بدون تغییر مستقیم پایگاه داده‌ی معین، کاهش دادند. با این حال، تغییر به طور غیر مستقیم می‌تواند از طریق پارامترهای در حال ترکیب مرتبط با معاملات پایگاه داده و قوانین انجمنی انجام شود. اضافات جدید شامل M پشتیبانی (پشتیبانی اصلاح شده)، M اعتماد (اعتماد اصلاح شده) و شمارنده پنهان کننده است. الگوریتم، از تعریف پشتیبانی و اعتماد استفاده کرد. بنابراین، قانون انجمنی حساس مورد نیاز را بدون هیچ گونه عوارض جانبی مخفی کرد [۸]. در سال ۲۰۰۸ Aggarwal و Yu بر دو عامل مهم مربوط به کاوش قانون انجمنی مانند اعتماد و پشتیبانی تاکید کردند. برای یک قانون انجمنی X => Y، پشتیبانی عبارتست از درصد معاملات در مجموعه داده که شامل X U Y است. اعتماد (قدرت نیز نامیده می‌شود) یک قانون انجمنی X => Y عبارتست از نسبت تعداد معاملات توسط X [۹]. در سال ۲۰۱۱ مشکلات پیش رو در داده کاوی به طور گسترده‌ای در بسیاری از جوامع مانند پایگاه داده، کنترل افشای آماری و جامعه رمزنگاری توسط Nayak و Devi سنجیده شدند [۱۰].

در سال ۲۰۱۱ Jain و همکاران الگوریتم جدیدی به منظور افزایش و کاهش پشتیبانی از آیتم قانون LHS[۲] و RHS[۲]برای مخفی کردن یا حفظ قوانین انجمنی ایجاد کردند. الگوریتم پیشنهادی، از این لحاظ سودمند است که حداقل اصلاح را بر روی اطلاعات ثبت شده ایجاد می‌کند تا مجموعه‌ای از قوانین را با زمان CPU کمتری نسبت به کار قبلی مخفی کند. این الگوریتم، تنها محدود به قانون انجمنی است [۱۱]. در سال ۲۰۱۱ Islam و Brankovic یک معماری حدید شامل تکنیک‌های مختلف ارائه کردند که تمام ویژگی‌ها در پایگاه داده را تحت تاثیر قرار داد. یافته‌های تجربی نشان داد که معماری ارائه شده در حفظ الگوهای اصلی در یک مجموعه داده مختل و آشفته، بسیار مؤثر است [۱۲].

در سال ۲۰۱۰ Naeem و همکارانش یک معماری ارائه دادند که قوانین انجمنی محدود را با حذف کامل عوارض جانبی معلوم مانند تولید ناخواسته قوانین انجمنی غیراصل به نمایش گذاشت در حالی که هیچگونه شکستی در آن وجود ندارد. در این معماری، مقیاس‌های آماری استاندارد به جای چارچوب متعارف پشتیبانی و اعتماد مورد استفاده قرار می‌گیرند تا قوانین انجمنی، به خصوص روش توزین بر اساس گرایش مرکزی ایجاد کنند [۱۳]. در سال ۲۰۱۰ Kamakshi و Babu سه مدل شامل کلاینت‌ها، مراکز داده، و پایگاه داده در هر سایت معرفی کردند. مرکز داده کاملا مجهول است، به طوری که نقش کلاینت‌ها و پایگاه داده سایت قابل تعویض است [۱۴].

در سال ۲۰۱۱ Mukkamala و Ashok مجموعه‌ای از روش‌های نگاشت مبتنی بر فازی را در زمینه ویژگی‌های حفظ حریم خصوصی و توانایی حفظ همان ارتباط با سایر زمینه‌ها مقایسه کردند. این مقایسه یعنی چه: (۱) چهار اصلاح جلودار مربوط به تعریف تابع فازی، (۲) معرفی هفت راه برای پیوستن مقادیر کاربردی مختلف از یک داده خاص به یک مقدار واحد، (۳) استفاده از چندین معیار شباهت برای مقایسه داده‌های اصلی و داده‌های نگاشت شده، و (۴) ارزیابی تاثیر نگاشت بر روی قانون انجمنی بدست آمده، قرار دارد [۱۵]. ارتباط تکنیک حفظ حریم خصوصی در داده کاوی به طور کامل توسط Matwin در سال ۲۰۱۳ مورد آنالیز و بحث قرار گرفته است. استفاده از روش‌های خاص، توانایی آنها را برای جلوگیری از استفاده تبعیض آمیز از داده‌کاوی نشان داد. برخی از روش‌ها پیشنهاد کردند که هر گروه نشاندار نباید بیشتر در تعمیم داده‌ها نسبت به جمعیت عمومی، هدف قرار داده شود [۱۶].

در سال ۲۰۱۳ Sachan و همکارانش راه حل‌های فعلی حفظ حریم خصوصی را برای خدمات ابری مورد آنالیز قرار دادند، که در آنها راه حل بر اساس اجزای رمزنگاری پیشرفته تعیین می‌شود. این راه حل، دسترسی ناشناس، توانایی جدا کردن و حفظ محرمانه بودن داده‌های منتقل شده را ارائه داد. در نهایت، این راه حل پیاده‌سازی می‌شود، نتایج تجربی به دست می‌آیند و عملکرد مقایسه می‌گردد [۱۷]. در سال ۲۰۱۳ Vatsalan و همکارانش تکنیکی به نام (PPRL[۳]) به معنی «ارتباط ضبط حریم خصوصی» را بررسی کردند، که با حفاظت از حریم خصوصی، ارتباط پایگاه داده‌ها به سازمان‌ها را مجاز کرد. بنابراین، یک طبقه بندی مبتنی بر روش‌های PPRL برای آنالیز آنها در ۱۵ بُعد پیشنهاد شده است [۱۸]. در سال ۲۰۱۰ Vijayarani و همکارانش در مورد تکنیک‌هایی از جامعه آماری کنترل افشا، جامعه پایگاه داده، و جامعه رمزنگاری توضیح دادند. ابزار کمتری از داده‌ها، نیاز به هزینه بالا دارد [۱۹].

در سال ۲۰۰۹ Li و Liu یک الگوریتم استخراج قانون انجمنی برای حفظ حریم خصوصی معرفی کردند. الگوریتم پیشنهادی بر اساس محدودیت پرس و جو و اختلال داده است. داده‌های اصلی را می‌توان با استفاده از الگوریتم پنهان یا مختل کرد تا راندمان حریم خصوصی بهبود یابد. این یک روش مؤثر برای تولید آیتم‌های مکرر از داده‌های تبدیل یافته است. نتایج تجربی نشان داد که روش پیشنهادی برای تولید مقادیر قابل قبول از تعادل حریم خصوصی با انتخاب مناسب پارامترهای تصادفی، کارآمد است [۲۰]. در سال ۲۰۱۲ Qi و

[2] Limite Hidden single
[2] Rule Hidden single
[3] Privacy Preserving Record Linkage

ISSN: 2745-3673
http://www.Rcsj.ir

Zong چند تکنیک موجود از داده‌کاوی را برای حفاظت از حریم خصوصی بسته به توزیع داده‌ها، اعوجاج، الگوریتم‌های استخراج، و پنهان کردن داده‌ها و قوانین مرور کردند. با توجه به توزیع داده‌ها، در حال حاضر تنها چند الگوریتم برای حفاظت از حریم خصوصی در داده کاوی بر اساس داده‌های متمرکز و پراکنده استفاده می‌شود [۲۱].

الگوریتم PSO

الگوریتم PSO که توسط Kennedy و Eberhart پیشنهاد شده است یک تکنیک بهینه سازی تصادفی بر مبنای جمعیت می باشد و از رفتارهای اجتماعی دسته پرندگان و ماهی‌ها الهام می‌گیرد. در الگوریتم PSO ابتدا سیستم با یک جمعیتی از جواب های تصادفی مقدار دهی اولیه می‌شود و سپس با بروز رسانی نسل‌ها جواب بهینه جستجو می‌شود. بر خلاف روش مشابه الگوریتم ژنتیک این الگوریتم هیچ عملگر تکاملی مانند ادغام و جهش ندارد و از سرعت همگرایی بالاتری نسبت به آن برخوردار است. در PSO جواب‌های بالقوه که ذره نامیده می‌شوند در کل فضای مسئله با دنبال کردن بهینه ترین ذره کنونی به حرکت در می‌آیند.

اگر یکی از ذرات مسیر خوبی را بیابد، سایر ذرات به دنبال آن ذره حرکت می کنند هرچند که از آن خیلی دور باشند. رفتار جمعی با استفاده از ذرات داخل فضای چند بعدی که دارای دو مشخصه مکان و سرعت هستند مدل می‌شود. این ذرات در کل این فضا حرکت می‌کنند و بهترین مکانی را که تاکنون ملاقات نکرده اند را به خاطر می‌سپارند. آنها این موقعیت‌های خوب را به اطلاع یکدیگر رسانده و موقعیت و سرعت حرکت خود را بر اساس این موقعیت‌های خوب تنظیم می‌کنند. اگر بخواهیم این روند را به صورت دقیق تر بیان کنیم مراحل زیر را خواهیم داشت: ذرات در ابتدا با جمعیتی تصادفی از جواب ها مقدار دهی اولیه می شوند. این جمعیت اولیه به صورت تکراری در کل فضای جستجوی بعدی حرکت می‌کنند و به دنبال جواب های جدید می گردند. برای هر ذره تابع برازندگی f به منظور اندازه گیری کیفیت جواب محاسبه می‌شود تا بهترین ذره مشخص گردد. هر ذره دارای یک مکان و یک سرعت است که به ترتیب توسط بردارهای مکان Xi (که اندیس ذره می‌باشند) و سرعت Vi نشان داده می‌شوند. هر ذره بهترین مکان خود تا لحظه کنونی را در بردار نگهداری می‌کند [۲۲].

الگوریتم Apriori

این الگوریتم از اولین الگوریتم‌هایی است که جهت یافتن مجموعه اقالم مکرر از آن استفاده می‌شود. نام آن برگرفته از شیوه‌هایی است که از آن استفاده می‌کند، یعنی استفاده از دانش مرحله قبل که در ادامه آن را شرح می‌دهیم. الگوریتم Apriori یک الگوریتم جستجوی سطحی است، که با پایان کاوش در سطح مرحله‌ی k ام به مرحله‌ی بعدی یعنی K+1 می‌رود. این عمل تا محقق شدن شرط یا شروط پایانی تکرار می‌شود. در مرحله‌ی k ام مجموعه اقالم k تایی تولید خواهند شد. پس از محاسبه‌ی مقدار پشتیبان برای هرکدام و مقایسه‌ی آن با مقدار minsup الگوهای مکرر k تایی شناسایی می‌شوند. در مرحله‌ی بعدی الگوریتم با کمک الگوهای مکرر k تایی، مجموعه اقالم k+1 تایی کاندید که بالقوه می‌توانند مکرر باشند را تولید می‌کند. به همین ترتیب با توجه به مقدار minsup برخی حذف شده و مجموعه اقالم مکرر k+1 تایی تشکیل خواهند شد. این عمل تا یافتن آخرین مجموعه قلم مکرر ادامه پیدا می‌کند .این الگوریتم در حین اجرا از قاعده های موسوم به قاعده Apriori استفاده می‌کند که بدین صورت بیان می‌شود: "اگر یک الگوی مکرر داشته باشیم، کلیه‌ی زیرمجموعه‌های آن نیز مکرر هستند." به عبارت دیگر اگر مجموعه اقالم مکرر نباشد، هر مجموعه که شامل است نیز نمی‌تواند مکرر باشد [۲].

الگوریتم ROCK

در بیشتر الگوریتم‌های خوشه بندی تشابه میان نمونه‌ها معیار ارزیابی است و در هر مرحله نمونه‌های مشابه در یک خوشه قرار می‌گیرند اما با این شرط الگوریتم مستعد خطاست. برای مثال دو خوشه‌ای را در نظر بگیرید که فقط چند نمونه از هر یک به هم نزدیک و مشابه باشند. با معیار مزبور دو خوشه می‌توانند ادغام شوند در حالیکه این عمل مناسب نیست. الگوریتم ROCK که برای خوشه بندی بر روی داده‌هایی با صفات خاصی طبقه‌بندی شده و دودویی طراحی شده است، با کمک محاسبه‌ی تعداد همسایه‌های مشترک میان نمونه‌ها راهکار مناسبی ارائه می‌کند. بطور معمول الگوریتم‌های خوشه‌بندی از معیارهایی که ارائه شدند، جهت ارزیابی تشابه یا فاصله میان نمونه‌ها استفاده می‌کنند. نتایج تجربی نشان می‌دهند این معیارها نمی‌توانند با نوع ویژگی طبقه‌بندی شده خوشه‌های مناسبی را در خروجی تولید کنند.

تعداد پیوندهای میان Oi و Oj به عنوان تعداد همسایه‌های مشترک میان این دو نمونه تلقی می‌شود. اگر این تعداد زیاد باشد، امکان یکی بودن خوشه‌های این دو نمونه بیشتر است. با در نظر گرفتن تعداد همسایه‌های مشترک می‌توان گفت الگوریتم ROCK در مقابل

ISSN: 2745-3673
http://www.Rcsj.ir

داده‌های نویز و خارج از محدوده مقاومتر از الگوریتم‌های خوشه‌بندی است که فقط بر روی تشابه میان نمونه‌ها متمرکز می‌شوند. بر اساس توضیح مختصر در می‌یابیم که الگوریتم ROCK با کمک از تابع ارزیابی تشابه، حد آستانه‌ی آن و ماتریس تشابه در مرحله‌ی اول به ساخت یک گراف پراکنده می‌پردازد. در گراف مفهوم همسایه‌های مشترک نیز لحاظ شده است. مرحله بعدی اجرای یک الگوریتم خوشه‌بندی سلسله‌مراتبی بر روی این گراف است که استفاده از یک معیار ارزیابی مناسب جهت تولید خوشه‌های خروجی می‌تواند آن را بهبود بخشد. برای مجموعه داده‌هایی با حجم بالا می‌توان از روش‌های نمونه‌گیری نیز استفاده نمود [۲].

الگوریتم DIANA

تکنیک های خوشه بندی سلسله مراتبی در دو دسته پایین به بالا و بالا به پایین گروه بندی شده‌اند. الگوریتم‌هایی که تاکنون به برسی آنها پرداخته ایم، در دسته اول گنجانده می‌شوند. این بدین معنی است که با قرار دادن هر نمونه در خوشه‌ای مجزا شروع و پس از معرفی معیاری جهت ارزیابی تشابه، آنها را ادغام می‌کنیم تا جایی که کلیه نمونه ها در یک خوشه قرار می‌گیرند.

علیرغم محبوبیت و عمومیت داشتن این دسته، برخی از الگوریتم‌ها نظیر DIANA و DISMEA عملکردی برعکس روش های فوق را دنبال می‌کنند. به این صورت که در ابتدا کلیه نمونه ها در یک خوشه قرار می‌گیرند و سپس در هر مرحله خوشه ها به خوشه های کوچکتر تقسیم می‌شوند، تا جایی که هر نمونه در یک خوشه قرار می‌گیرد. در هر مرحله از الگوریتم DIANA بزرگترین خوشه تقسیم می‌شود تا در مرحله n-1 ام در یک خوشه مجزا قرار بگیرد. فرض کنید خوشه‌ی C خوشه‌ی مورد نظری است که باید به دو خوشه‌ی دیگر یعنی A و B تقسیم شود. تعداد نمونه‌های موجود در خوشه‌ی C بیشتر از دو نمونه است و پس از عمل تقسیم، خوشه‌های A و B دارای نمونه‌ی مشترکی نیستند. الگوریتم DIANA در ابتدا کلیه‌ی نمونه‌های خوشه‌ی C را در خوشه‌ی A قرار می‌دهد که خوشه B در ابتدا خالی است. در مرحله اول یک نمونه از خوشه‌ی A به خوشه‌ی B منتقل می‌شود. در مراحل بعدی الگوریتم جستجو برای انتقال دیگر نمونه‌ها از A به خوشه B ادامه می‌یابد [۲].

الگوریتم K-means

داده به صورت عمودی یک بار تقسیم می‌شود که این کار معایب خاص خود را دارد. اول، ما باید تصمیم بگیریم که مدل چگونه به اشتراک گذاشته است. گفتن این که داده‌های گروه ما مراکز خوشه‌ها هستند آسان است. با این حال، جدا از عضویت خوشه، چه اطلاعات دیگری توسط فرد به اشتراک گذاشته شده است؟ آیا همه گروه‌ها تمام اطلاعات مربوط به هر خوشه و یا انجام آنها را که تنها اطلاعات محدود به ویژگی های خود هستند را می‌دانند. کدام یک از این گزینه‌ها فضای دنیای واقعی را بهتر درک می‌کند؟ مسائل خوشه‌بندی داده‌ها به صورت تقسیم عمودی پیشنهاد دادند. یک پروتکل حفظ حریم خصوصی، خوشه بندی K-means را انجام می‌دهد. هر چند همه گروه‌ها انتساب نهایی از مراکز داده‌ها به خوشه را می‌دانند، مطمئناً آنها فقط اطلاعات جزئی برای هر خوشه را حفظ می‌کنند. فرض می‌شود اطلاعات مراکز این خوشه‌ها کامل نیست، به عنوان مثال، هر یک از سایت‌ها می‌توانند تنها اجزایی از آن را که مطابق با ویژگی‌های آن است یاد بگیرند. بنابراین، تمام اطلاعات در مورد ویژگی‌های یک سایت (و نه فقط ارزش‌های فردی) به صورت شخصی نگهداری شوند. اگر به اشتراک گذاری مورد نظر امن باشد می‌توان یک ارزیابی از نگرانی‌های حفظ حریم خصوصی را پس از شناخت ویژگی‌ها انجام داد [۴].

ورودی این الگوریتم n نمونه داده و مقدار k که تعداد خوشه‌های خروجی را مشخص می‌کند، می‌باشد. در ابتدا تعداد k نمونه به صورت اتفاقی از میان کل نمونه‌ها انتخاب می‌شوند. این نمونه‌ها به عنوان نماینده k خوشه شناخته خواهند شد. گاهی به آنها مرکز خوشه نیز اطلاق می‌شود. هر یک از نمونه‌های باقیمانده عضوی از خوشه‌ای خواهند بود که یکی از این نماینده ها (k نمونه) متعلق به آن است. به عبارت دیگر با کمک معیارهایی همچون فاصله اقلیدسی تشابه هر یک از نمونه‌های باقیمانده را با k نماینده محاسبه می‌کنیم و نمونه مورد نظر به هر یک نزدیکتر بود، به عضویت آن خوشه در می‌آید. پس از آن برای هر خوشه، با محاسبه میانگین میان اعضای خوشه نماینده جدیدی انتخاب می‌گردد. این فرآیند تا پوشش معیاری جهت خاتمه کار تکرار می‌شود. برای مثال این فرآیند می‌تواند تا هنگامی که دیگر هیچ یک از نمونه‌ها خوشه‌های خود را تغییر ندهند ادامه پیدا کند. معمولا به حداقل رساندن درصد خطا معیار خوبی برای، شرط پایانی است.

پیچیدگی زمانی این الگوریتم برابر است با O(nkt)، که در آن n معرف تعداد نمونه ها، k تعداد خوشه‌ها و t تعداد تکرارهایی است که الگوریتم پس از آن خاتمه می یابد. روشن است که تعداد خوشه‌ها و تعداد تکرارها بسیار کوچکتر ار تعداد نمونه ها هستند. لذا می‌توان ادعا نمود که این الگوریتم در کار با حجم بالای داده ها نسبتا مقیاس پذیر و کارا است. ضرورت مشخص نمودن تعداد خوشه‌ها در ابتدای الگوریتم هم عیب و هم حسن آن به شمار می‌رود. در برخی از کاربردهای عملی باید تعداد خوشه‌های نهایی مشخص باشند، ولی در اغلب کاربر دها دید روشنی از تعداد خوشه‌ها ندارد و پسند کاربر این است که الگوریتم خود به صورت خودکار این تعداد را بدست آورد.

ISSN: 2745-3673
http://www.Rcsj.ir

از آنجا که در هر مرحله با محاسبه میانگین نمونه‌ها سروکار داریم، بنابراین برای ویژگی هایی که محاسبه میانگین برای آنها دارای مفهومی نیست، این الگوریتم مناسب نیست. به علاوه از آنجا که داده‌های نویز و خارج از محدوده مانند داده‌هایی با مقدار خیلی کم و یا زیاد بر روی مقدار میانگین اثرگذار است، این الگوریتم در مواجهه با این نو داده‌ها عملکرد مناسبی از خود نشان نمی‌دهد و به عبارتی دیگر مقاوم نیست. این روش برای کشف خوشه‌هایی که دارای شکل‌های غیرکروی و در اندازه‌هایی بسیار متفاوت هستند نیز مناسب نیست. پیچیدگی فضای مورد نیاز آن $O(k+n)$ می باشد و چنانچه ذخیره سازی تمام نمونه ها در حافظه اصلی امکان پذیر باشد زمان دسترسی به تمام نمونه ها خیلی سریع بوده و الگوریتم بسیار کارآمد است [۲].

الگوریتم k-Medoids

الگوریتم k-Medoids عملکردی بسیار شبیه به الگوریتم k-Means دارد، با این تفاوت که در الگوریتم k-Medoids به جای استفاده از میانگین، از خود نمونه‌ها برای مرکز ثقل و نمایندگی خوشه‌ها استفاده می‌شود. با انتخاب نمونه‌های واقعی جهت نمایش یک خوشه، حساسیت روش نسبت به نمونه‌های نویز و خارج از محدوده کاهش می‌یابد. فراموش نکنید که الگوریتم k-Means به دلیل اینکه حتی تعداد کمی از این داده‌ها می‌تواند در مقدار میانگین تأثیر بگذارد، الگوریتم به این گونه از داده‌ها بسیار حساس است. بنابراین روش-k Medoidsبرخلاف k-Means به جای اینکه مقادیر میانگین نمونه‌ها را دریافت کند، از مرکزی‌ترین نمونه موجود در خوشه به عنوان نمایش و نماینده خوشه استفاده می‌کند. به همین دلیل این الگوریتم حساسیت کمی نسبت به داده‌های خارج از محدوده از خود نشان می‌دهد.

در این الگوریتم همانندk-Means در ابتدا باید مقدار k را مشخص کنید. پس از آن تعداد k نمونه به عنوان نماینده‌های اولیه k خوشه به صورت اتفاقی انتخاب می‌شوند. پس از تشکیل ماتریس تشابه، هر یک از نمونه‌های باقیمانده باید در یکی از این k خوشه قرار گیرند. در این الگوریتم می‌توانیم به جای تشکیل ماتریس تشابه، فاصله هر یک از نمونه‌های باقیمانده را با k نمونه اولیه محاسبه کنیم. هر نمونه به نزدیکترین نماینده تعلق دارد. تا این مرحله از الگوریتم تفاوتی میان k-Means و k-Medoids مشاهده نمی‌شود. پس از این با جایگزینی یک نمونه از داده ها با یکی از k نمونه نماینده، کیفیت و مناسب بودن خوشه‌های بدست آمده از این جایگزینی بررسی می‌شوند. در صورت بهبود در نتایج، مجاز به جایگزینی نماینده مزبور خواهیم بود. یکی از اولین الگوریتم‌های k-Medoids با نام PAM معرفی شد.

در این الگوریتم پس از انتخاب k نماینده به صورت اتفاقی از میان نمونه‌ها، با تکرار سعی می‌شود تا نماینده‌گان بهتری برای خوشه‌ها انتخاب شوند. جایگزینی نمونه‌ها و نماینده‌ها درصورتی انجام می‌شود که بیشترین کاهش را در مقدار خطا داشته باشیم. در این الگوریتم کلیه ترکیبات دوتایی از نمونه‌ها ارزیابی می‌شوند، به صورتی که یکی از نمونه ها نماینده یا مرکز ثقل باشد. مجموعه بهترین نمونه‌های هر خوشه در یک تکرار، نماینده‌گان خوشه‌ها برای تکرار بعدی را شکل می‌دهند [۲].

الگوریتم BBA

Sun و Yu در سال ۲۰۰۵ نخستین روش مبتنی بر مرز را برای مخفی‌سازی قوانین انجمنی پیشنهاد دادند. الگوریتم پیشنهادی آنها به صورت حریصانه با در نظر گرفتن میزان تأثیر تغییر دادن قلم کاندیدا بر روی مرز مثبت اصلاح شده، تغییرات را انجام می‌دهد. مـجموعه-اقلامِ مرزی بـــه صورت ضمنـی وضعیت دیـگر مجموعه-اقلام را از نظر مکرر یا غیر مکرر بودن مشخص می‌کنند. در نتیجه کیفیت مرزها به صورت مستقیم بر کیفیت پایگاه داده پاک‌سازی شده تأثیر می‌گذارد. در این الگوریتم برای اندازه‌گیری میزان آسیبی که حذف یک قلم به کیفیت پایگاه‌داده پاک‌سازی شده وارد می‌کند، به هر کدام از مجموعه-اقلامِ مرز مثبت اصلاح شده، وزنی اختصاص داده شده است. الگوریتم برای حذف یک مجموعه-اقلام حساس میزان تأثیر حذف هر کدام از اقلام با توجه به وزن آن‌ها را بر روی مرز مثبت اصلاح شـده محاسبـه می‌کنـد و سپـس قلم کاندیدایی که کم‌ترین تأثیر را ایجاد می‌کند انتخاب می‌شود و در برخی از تراکنش‌ها که به دقت انتخاب شده‌اند این قلم حذف می‌گردد [۲۴-۲۳].

الگوریتم max-min

Moustakides و Verykios دو الگوریتم مبتنی بر مرز را براساس معیار max-min[4] ارائه نمودند. هر دو الگوریتم مرز مثبت اصلاح شده‌ی مجموعه-اقلامِ متکرر را برای مشاهده میزان تأثیر تغییر آزمایشی یک قلم، به کار بردند. سپس آن‌ها تغییری را اعمال می‌کنند

[4] Max-Min Criterion

ISSN: 2745-3673
http://www.Rcsj.ir

که به صورت مؤثر همه دانش حساس را مخفی می‌کند و همچنین کم‌ترین تأثیر را بر روی مجموعه‌اقلامِ مرز مثبت اصلاح شده ایجاد می‌کند. آزمایش‌های ارائه شده در این مقاله‌ها نشان می‌دهد که این الگوریتم‌ها نتایج بهتری نسبت به الگوریتم BBA تولید می‌کنند. الگوریتم‌هایی که در بخش‌های قبلی مطرح شدند اگر چه الگوریتم‌ها بسیار سریع و کارآمدی هستند اما ممکن است که در نقاط کمینه محلی به دام بیافتند و نتوانند جواب بهینه کلی را بیابند. دسته دیگری از الگوریتم‌ها وجود دارند که مسأله مخفی‌سازی قوانین انجمنی را به یک مسأله بهینه‌سازی تبدیل می‌کنند (برای مسأله بدست آمده راه حل کارآمد ریاضی وجود ندارد) و با ساده‌سازی (معیار یا معیارهای کیفیت داده‌ها)، تقریب بسیار مناسبی از مسأله بدست می‌آورند که توسط روش‌های ریاضی قابل حل است و با حل این مسأله ساده شده جواب بهینه را بدست می‌آورند. این دسته از الگوریتم‌ها را الگوریتم‌های دقیق می‌خوانند. الگوریتم‌هایMenon، Inline، Two-Phase Iterative و Hybrid در این دسته قرار می‌گیرند. در ادامه به بررسی الگوریتم Menon می‌پردازیم [۲۵].

الگوریتم Menon

Menon و همکارانش الگوریتمی برای مخفی‌سازی قوانین انجمنی پیشنهاد دادند که از دو بخش دقیق و مکاشفه‌ای تشکیل شده بود. بخش دقیق این الگوریتم با استفاده از پایگاه داده اولیه یک مسأله ارضای محدودیت[5](CSP) در جهان مجموعه-اقلام حساس ایجاد می‌کند. هدف این است که تراکنش‌هایی که باید پاک‌سازی شوند طوری تعیین شود که تعداد تراکنش‌های پاک‌سازی شده کمینه شود و در عین حال همه دانش حساس مخفی گردد [۲۶]. فرآیند بهینه سازی در حل این CSP با استفاده از تابع معیاری[6] که از معیار دقت نشأت می‌گیرد، انجام می‌شود. هدف قیـود که به فرم برنامه‌ریزی عدد صحیح[7] ظاهر می‌شوند ایـن است کـه تراکنش‌هایی از پایگاه‌داده را که نیازمنـد پاک‌سازی هستند را استخراج کنند. با استفاده از حل کننده‌های برنامه‌ریزی عدد صحیح جواب بهینه برای این CSP بدست می‌آید و تراکنش‌هایی که باید پاک‌سازی شوند مشخص می‌گردند. در مرحله بعد با استفاده از یکی از روش‌های مکاشفه‌ای این تراکنش‌ها پاک‌سازی می‌شوند. هدف بخش دقیق الگوریتم Menon این است که تعداد تراکنش‌های لازم برای پاک‌سازی پایگاه داده اولیه را کمینه کند. واضح است که تعداد کمینه تراکنش‌هایی که باید برای یک مجموعه-اقلام حساس که باید پاک‌سازی شود با پشتیبان کنونی آن در پایگاه داده منهای آستانه حداقل پشتیبان بعلاوه یک برابر است. هنگامی که این تعداد از تراکنش‌ها طوری تغییر داده شوند که دیـگر مجموعه-اقلام جاری را پشتیبانی نکنند، مقدار پشتیبان آن به زیر آستانه حداقل پشتیبان خواهد رسید و در نتیجه مجموعه‌اقلام مخفی خواهد شد [۲۷-۲۸].

[5] Constraint Satisfaction Problem
[6] Criterion Function
[7] Integer Programming

ISSN: 2745-3673
http://www.Rcsj.ir

نتیجه گیری

مشکلی که الگوریتم های حفظ حریم خصوصی در تولید قوانین انجمنی دارند این است که در تولید قوانین ترتیب آیتم‌ها نقش مهمی در آنها بازی می‌کند و این باعث می‌شود که تعداد قوانین تولید شده در آنها زیاد شود، در نتیجه انتخاب یک قانون بهتر از مجموعه قوانین ممکن سخت خواهد بود. عیب الگوریتم PSO این است که با جواب های تصادفی مقدار دهی اولیه می‌شود ولی در مقابل سرعت همگرایی بالایی نسبت به الگوریتم‌های دیگر دارد که این باعث می شود جواب بهینه زودتر بدست آید.

الگوریتم Apriori از دانش مرحله قبل استفاده می‌کند و جستجوی آن سطحی است ولی مزیتی که این الگوریتم دارد این است که فضای جستجوی کمتری دارد. الگوریتم K-Means درصد خطای کمتری نسبت به الگوریتم های دیگر دارد اما عیبی که دارد این است که فقط برای محاسبه میانگین مناسب است. الگوریتم K-Medoids عیب الگوریتم K-Means را که فقط برای محاسبه میانگین مناسب است را ندارد، عیب این الگوریتم پیچیدگی زمانی بالای آن است. الگوریتم Rock در مقابل داده‌های نویز دار و خارج از محدوده، مقاوم تر از الگوریتم‌های خوشه بندی است. الگوریتم DIANA عملکردی برعکس الگوریتم‌های فوق دارد، این الگوریتم به نمونه های خارج از محدوه تعیین شده عکس العمل مناسبی نشان نمی‌دهد.

الگوریتم BBA میزان آسیبی که حذف یک قلم بر کیفیت پایگاه داده وارد می‌کند را به خوبی محاسبه می‌کند بنابراین می‌توان گفت این الگوریتم دقت خوبی دارد. الگوریتم Max-Min برای جلوگیری از افشای اطلاعات همه دانش حساس را مخفی می‌کند و همچنین دقت این الگوریتم نسبت به الگوریتم BBA بیشتر است. الگوریتم Menon برخلاف الگوریتم های دیگر از دو بخش دقت و کشف تشکیل شده است، این الگوریتم جواب بهینه را بدست آورده و همچنین مجموعه-اقلام حساس را به خوبی مخفی می‌کند در نتیجه دقت این الگوریتم بیشتر است. با بررسی مزایا و معایب الگوریتم های مربوط به حفظ حریم خصوصی در داده کاوی مشخص شد که الگوریتم های بهینه و کاراتری در این رابطه وجود داد، از جمله الگوریتم‌های مناسب و داری معایب کم می توان به الگوریتم های Max-Min و الگوریتم Menon اشاره کرد. این دو الگوریتم همه دانش حساس را مخفی می‌کنند بنابراین امکان سوء استفاده اینترنتی از اطلاعات منتشر شده افراد با استفاده از آنها کمتر خواهد بود، در نتیجه این دو الگوریتم از دقت و امنیت بالاتری برخوردار هستند.

ISSN: 2745-3673
http://www.Rcsj.ir

منابع و مراجع

[۱] باشگاه خبرنگاران جوان، "وقتی حریم خصوصی افراد در فضار مجازی قربانی می شود"، کد خبر: ۵۸۹۹۲۱۸، گروه: اجتماعی(دنیای ارتباطات)، تاریخ انتشار: ۲۵ آذر ۱۳۹۵.

[۲] مهدی اسماعیلی، " مفاهیم و تکنیک های داده کاوی"، data mining. concepts and techniques.pdf. تیرماه ۱۳۹۱.

[3] YOUSRA ABDUL ALSAHIB S.ALDEEN et al, "A comprehensive review on privacy preserving data mining", DOI: 10.1186/s40064-015-1481-x, Article in SpringerPlus · November 2015.

[4] Sushil Jajodia et al, " PRIVACY PRESERVING DATA MINING", Consulting Editor, Center for Secure Information Systems, George Mason University, Fairfax, VA 22030-4444.

[5] Tan, P. N., Steinbach, M., Kumar, V., (2005). Introduction to Data Mining.Addison Wesley.

[6] Adamo, J. M., (2001). Data Mining for Association Rules and Sequential Patterns, Springer-Verlag, New York.

[7] Ahmed K. Elmagarmid and Amit P. Sheth, "Privacy-Preserving Data Mining-Models and Algorithms", edite by: charu. C, aggarwal and Philip s. yu, Purdue University, West Lafayette, IN 47907; ISBN: 978-0-387-28759-1.

[8] Belwal et al, "Hiding sensitive association rules efficiently by introducing new variable hiding counter". In: IEEE international conference on service operations and logistics, and informatics, 2008, IEEE/SOLI 2008, vol 1, pp 130–134. Doi:10.1109/SOLI.2008.4686377.2013.

[9] C. Aggarwal and S. Yu, "A general survey of privacy-preserving data mining models and algorithms". In: Privacy preserving data mining, Chap 2. Springer, New York, pp 11–52. Doi: 10.1007/978-0-387-48533, 2008.

[10] G. Nayak and S. Devi, "A survey on privacy preserving data mining: approaches and techniques". Int J Eng Sci Tech 3(3): 2117–2133. 2011.

[11] Jain et al, "An efficient association rule hiding algorithm for privacy preserving data mining". Int J Comp Sci Eng 3(7):2792–2798, 2011.

[12] m. Islam and L. Brankovic," Privacy preserving data mining: a noise addition framework using a novel clustering technique". Knowl Based Syst 24(8):1214–1223, 2011.

[13] Naeem et al," Hiding sensitive association rules using central tendency". In: 6th international conference on advanced information management and service (IMS), pp 478–484. http://ieeexplore.ieee.org/xpls/abs_all. Jsp? Arnumber=5713497, 2010.

[14] P. Kamakshi and A. Babu,"Preserving privacy and sharing the data in distributed environment using cryptographic technique on perturbed data" 2(4) , 2010.

[15] R. Mukkamala and V. G. Ashok, "Fuzzy-based methods for privacy-preserving data mining". In: IEEE eighth international conference on information technology: new generations (ITNG), 2011.

[16] S. Matwin, "Privacy-preserving data mining techniques: survey and challenges". In: Discrimination and privacy in the information society. Springer, Berlin, Heidelberg, pp 209–221. 2013.

[17] Sachan et al, "An analysis of privacy preservation techniques in data mining". In: Advances in computing and information technology, vol 3. Springer, pp 119–128, 2013.

[18] Vatsalan et al, "A taxonomy of privacy-preserving record linkage techniques". INF Syst 38(6):946–969, 2013.

ISSN: 2745-3673
http://www.Rcsj.ir

[19] Vijayarani et al, "a survey Privacy preserving data mining based on association rule". In: IEEE international conference on communication and computational intelligence (INCOCCI), 2010.

[20] W. Li and J. Liu, "Privacy preserving association rules mining based on data disturbance and inquiry limitation". In: 2009 Fourth International Conference on Internet Computer Science Engineering. pp 24–29, 2009.

[21] X. Qi and M. Zong, "An overview of privacy preserving data mining". Procedia Environ Sci 12(Icese 2011):1341–1347, 2012.

[22] Kennedy, J., Eberhart, R. C., (1995). Particle Swarm Optimization, Proc. of IEEE in Int. Conf. on Neural Networks, pp. 1942-1948, Piscataway, NJ.

[23] Sun, X. and Yu, P.S. A border-based approach for hiding sensitive frequent itemsets. Data Mining, Fifth IEEE International Conference on, (2005), 8.

[24] Sun, X. and Yu, P.S. Hiding Sensitive Frequent Itemsets by a Border-Based Approach. Journal of Computing Science and Engineering 1, 1 (2007), 74–94.

[25] Moustakides, G.V. and Verykios, V.S. A MaxMin approach for hiding frequent itemsets. Data & Knowledge Engineering 65, 1 (2008), 75-89.

[26] Menon, S., Sarkar, S., and Mukherjee, S. Maximizing accuracy of shared databases when concealing sensitive patterns. Information Systems Research 16, 3 (2005), 256.

[27] Lee, G., Chang, C.-Y., and Chen, A.L.P. Hiding Sensitive Patterns in Association Rules Mining. Computer Software and Applications Conference, Annual International, IEEE Computer Society (2004), 424-429.

[28] Reddy, M. and Wang, R. Estimating data accuracy in a federated database environment. Information Systems and Data Management, (1995), 115–134.