Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: jscit.nit.ac.ir
Volume 10, Number 4, Winter 2022, pp. 29-42
Received: 04/05/2021, Revised: 07/15/2021, Accepted: 10/02/2021

Privacy Preserving Classification in Vertically Distributed Datasets

Mohammad Reza Ebrahimi Dishabi*
Department of Computer Engineering, Miyaneh Branch, Islamic Azad University, Miyaneh, Iran.
*mrebrahimy@m-iau.ac.ir

Corresponding author's address: Mohammad Reza Ebrahimi Dishabi, Miyaneh Branch, Islamic Azad University, Miyaneh, Iran.

Abstract- Nowadays, preserving the privacy of data is an important issue during data mining techniques. several algorithms have been proposed to preserve the privacy of data. The most important problems with these algorithms are their unprovable privacy, the low importance of considering the adversary's background knowledge and lack of dimensionality reduction process over the original data. In this paper, differential privacy mechanism has been used to prove the privacy of vertical distributed data to be used for classification. In differential privacy, it is no matter what background knowledge has an adversary about the published data. Also, Haar wavelet transform has been used for dimension reduction of the original data. The privacy of data has been proved mathematically and the accuracy of data measured using the K-NN algorithm. Finally, it has been mathematically proved that the proposed algorithm adds noise with less standard deviation to the data than the compared algorithm, resulting in higher classification accuracy. The result shows that our algorithm has more secure compared to previous algorithms.

Keywords- Data Mining, Classification, Haar Wavelet Transform, Differential Privacy.

مجله علمی پژوهشی رایانش نرم و فناوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: jscit.nit.ac.ir
جلد ۱۰، شماره ۴، زمستان ۱۴۰۰ ، صفحه ۲۹-۴۲
دریافت: ۱۴۰۰/۰۱/۱۶، بازنگری: ۱۴۰۰/۰۴/۲۴، پذیرش: ۱۴۰۰/۰۷/۱۰

حفظ حریم خصوصی در رده‌بندی داده‌های توزیع شده عمودی

محمدرضا ابراهیمی دیشابی*
گروه کامپیوتر، واحد میانه، دانشگاه آزاد اسلامی، میانه، ایران.
*mrebrahimy@m-iau.ac.ir

* نشانی نویسنده مسئول: محمدرضا ابراهیمی دیشابی، میانه، بلوار شهدای زینبیه (جاده ترانزیت)، ساختمان سردار شهید حاج قاسم سلیمانی (ساختمان مرکزی)، گروه کامپیوتر، واحد میانه، دانشگاه آزاد اسلامی.

چکیده- امروزه، حفظ حریم خصوصی داده‌های منتشر شده، از اهمیت زیادی برخوردار است. الگوریتم‌های مختلفی برای حفظ حریم خصوصی داده‌ها ابداع شده است. مهمترین مشکلات این الگوریتم‌ها، غیر قابل اثبات بودن حریم خصوصی آنها، اهمیت کم در نظر گرفته شده برای دانش پیش‌زمینه فرد متخاصم و عدم کاهش ابعاد داده‌های اولیه است. در این مقاله، برای اثبات حریم خصوصی داده‌ها از مفهوم حریم خصوصی تفاضلی در انتشار داده‌های توریع شده عمودی و رده‌بندی آنها استفاده شده است. در حریم خصوصی تفاضلی، دانش پیش‌زمینه فرد متخاصم در مورد داده‌های منتشر شده، اهمیت ندارد. همچنین، با استفاده از تبدیل موجک هار، ابعاد داده‌ها کاهش داده شده است. در نهایت، حریم خصوصی داده‌های منتشر شده به صورت ریاضی اثبات شده است، همچنین، دقت داده‌ها، با استفاده از الگوریتم رده‌بند «K-نزدیکترین همسایه» نیز اندازه‌گیری شده است و به صورت ریاضی اثبات شده است که الگوریتم پیشنهادی، خدشه‌هایی با انحراف معیار کمتری را نسبت به الگوریتم مورد مقایسه، به داده‌ها اضافه می‌کند و در نتیجه دقت رده‌بندی بالاتری دارد. نتایج به دست آمده نشان دهنده امن‌تر بودن الگوریتم ارائه شده نسبت به الگوریتم‌های مورد مقایسه است.

واژه‌های کلیدی: داده‌کاوی، رده‌بندی داده‌ها، تبدیل موجک هار، حریم خصوصی تفاضلی.

۱- مقدمه

مجموعه داده‌ای را در نظر بگیرید که شامل اطلاعات خصوصی و حساس افراد بوده که به نوبه‌ی خود می‌توانند ارزشمند باشند. فرض کنید این داده‌ها اطلاعات مربوط به یک سری بیماریهایی است که منطقه به منطقه مشخص شده‌اند. پژوهشگران با استفاده از این اطلاعات، نتایج مهمی را می‌توانند استخراج کنند. به عنوان مثال، با بررسی این اطلاعات، می‌توان همه‌گیر شدن یک بیماری خاص را پیش‌بینی کرد و با برنامه‌ریزی دقیق‌تر، جان انسانها را نجات داد یا می‌توان جهت توزیع بهتر تجهیزات پزشکی و دارویی در مراکز درمانی، از آنها استفاده کرد. از این‌رو این داده‌ها باید در اختیار پژوهشگران قرار داده شوند تا با استفاده از روشهای داده‌کاوی و استخراج دانش به اهداف خود برسند. از طرفی، انتشار داده‌ها، باعث فاش شدن اطلاعات خصوصی و حساس بیماران خواهد شد که بر اساس قانون، دارندگان این داده‌ها (بیمارستانها و مراکز درمانی) مجاز به این کار نیستند. پس باید به دنبال روشی جهت انتشار داده‌ها باشیم که هم حریم خصوصی افراد نقض نشود و هم اینکه پژوهشگران بتوانند از داده‌های منتشر شده استفاده شایسته به عمل بیاورند.

جهت غلبه بر مشکل حفظ حریم خصوصی در داده‌کاوی، مفهوم «داده کاوی با حفظ حریم خصوصی»[۱] (PPDM) ابداع شد و در حال حاضر، تحقیقات و پژوهشهای زیادی در این زمینه در حال انجام است. برای PPDM، راه‌ها و روش‌های گوناگونی پیشنهاد شده است. انتشار داده‌هایی که قابلیت حفظ حریم خصوصی را داشته باشند[۲]

یکی از این روش‌ها است که به اختصار PPDP نامیده می‌شود [۱, ۲]. تصادفی‌سازی داده‌ها[۳] [۳]، گمنامی-k[۴] [۴, ۵ ,۶] و گوناگونی-l[۵] [۷] از مهمترین الگوریتم‌های PPDP هستند. در PPDP، بر روی ابداع الگوریتمهایی که باعث تغییر[۶] داده‌های اصلی می‌شوند، فعالیت می‌کنند طوری‌که داده‌ها را بعد از تغییر، در اختیار کاربران قرار داده و کاربران نیز با استفاده از روش‌های مختلف داده‌کاوی، از جمله رده‌بندی، به دنبال اطلاعات مورد نظر خود می‌گردند. از این‌رو، «حفظ حریم خصوصی در رده‌بندی داده‌ها»[۷] (PPC) یکی از زیر شاخه‌های PPDP است. برای PPC، الگوریتم‌های مختلفی ارائه شده است. هر کدام از این الگوریتم‌ها، داده‌ها را طوری تغییر داده و منتشر می‌کنند که امکان نقض حریم خصوصی داده‌های منتشر شده در اثر اعمال الگوریتم‌های رده‌بندی بر روی آنها، میسر نباشد. اکثر الگوریتم‌های ارائه شده برای PPC، دو ایراد اساسی زیر را دارند:

- این الگوریتم‌ها، فرض می‌کنند که فرد مهاجم[۸]، اطلاعات پیش‌زمینه‌ی[۹] کمی درباره‌ی داده‌های منتشر شده دارند [۲]. اما، تشخیص اینکه یک فرد مهاجم چقدر اطلاعات پیش‌زمینه درباره‌ی داده‌های منتشر شده در اختیار دارد و به طبع آن، شناسایی تمام روش‌هایی که منجر به کشف اطلاعات حساس داده‌های منتشر شده می‌شوند، بسیار دشوار و شاید غیرممکن است. بنابراین اکثر الگوریتم‌های ارائه شده برای PPC در مقابل دانش پیش‌زمینه‌ی کاربران آسیب‌پذیر هستند. در [۲]، انواع حملاتی که در اثر دانش پیش‌زمینه، بر روی داده‌های منتشر شده قابل اعمال هستند را به طور کامل توضیح داده است.
- این الگوریتم‌ها، هیوریستیک‌مبنا[۱۰] بوده و حریم خصوصی قابل اثبات ریاضی، ندارند [۲, ۸].

برای غلبه بر مشکلات بالا، مفهوم «حریم خصوصی تفاضلی»[۱۱] ابداع شد. در سال‌های اخیر استفاده از حریم خصوصی تفاضلی گسترش روزافزونی پیدا کرده است. هدف اصلی از حریم خصوصی تفاضلی این است که تغییرات کوچک در داده‌های ورودی، باعث ایجاد تغییرات کوچکی در خروجی الگوریتمی که روی داده‌ها اعمال می‌شود، گردد و معنی آن این است که خروجی الگوریتم، وابسته به هیچ کدام از داده‌های ورودی نباشد. از نظر حریم خصوصی تفاضلی، اینکه یک فرد مهاجم چقدر اطلاعات کمکی در مورد یک داده خاص در اختیار دارد، مهم نیست. از طرف دیگر، قابل اثبات بودن حریم خصوصی داده‌های منتشر شده با استفاده از مفهوم حریم خصوصی تفاضلی، یکی از مهمترین مزایای این مفهوم است. بنابراین یکی از قوی‌ترین تعریف‌هایی است که تا کنون درباره‌ی حریم خصوصی داده‌ها پیشنهاد شده است و از مبنای ریاضی قوی و محکمی برخوردار است.

در این مقاله، حفظ حریم‌خصوصی داده‌های منتشر شده در هنگام رده‌بندی آنها (PPC) مورد بررسی قرار گرفته است. انگیزه اصلی ما این است که الگوریتمی بر روی داده‌های توزیع شده به منظور رده‌بندی آنها ارائه گردد که (۱) حریم خصوصی داده‌های منتشر شده حفظ گردد و با استفاده از یک مدل ریاضی، قابل اثبات باشد و (۲) ابعاد داده‌های اولیه را کاهش دهیم طوری‌که کارآمدی الگوریتم‌های رده‌بندی افزایش پیدا کند و با سرعت بالاتری داده‌ها را در رده‌های مربوطه قرار دهد.

بر این اساس، الگوریتمی بر روی داده‌های توزیعی ارائه شد. الگوریتم ارائه شده، مورد ارزیابی قرار گفت و به صورت ریاضی اثبات گردید که خاصیت حریم خصوصی تفاضلی- ϵ دارد. همچنین، با اجرای الگوریتم پیشنهادی بر روی مجموعه داده‌های مورد آزمایش، میزان کاهش ابعاد داده‌های اولیه در جدولی مشخص گردید. در نهایت، الگوریتم پیشنهادی با الگوریتمی که بر روی داده‌های متمرکز ارائه شده است و دو مورد از خاصیت‌های الگوریتم ارائه شده (یعنی، (۱) حریم خصوصی تفاضلی و (۲) کاهش ابعاد داده‌ها) را دارد (خاصیت توزیعی بودن داده‌ها را ندارد)، مورد مقایسه قرار گرفت و به صورت ریاضی نشان داده شد که الگوریتم ارائه شده، هم از نظر میزان حریم خصوصی و هم از نظر دقت رده‌بندی، بهتر از الگوریتم مورد مقایسه است.

ساختار مقاله به این صورت است. در بخش ۲، مفاهیم پایه توضیح دااده شده است. در بخش ۳، کارهای مرتبط بیان شده است. در بخش ۴، الگوریتم پیشنهادی ارائه شده است. در بخش ۵، حریم خصوصی تفاضلی الگوریتم پیشنهادی اثبات می‌گردد و مورد ارزیابی قرار می‌گیرد . در بخش ۶، نتیجه‌گیری ارائه می‌گردد.

۲- مفاهیم پایه

در این بخش، مفاهیم پایه توضیح داده می‌شود.

۲-۱- تبدیل موجک هار

موجک به معنای موج کوچک است بنابراین تجزیه تحلیل موجک [۱۲] به معنای تجزیه و تحلیل یک سیگنال (با انرژی متناهی) در یک بازه‌ی کوچک است و منظور از متناهی بودن انرژی موج f این است که $f \in L^2(R)$ وجود دارد و با نماد $\int_{-\infty}^{+\infty}|f(t)|^2dt$ نمایش می‌دهند که در آن $L^2(R) = \{f(x) \mid \int_{-\infty}^{+\infty}|f(t)|^2dt < \infty\}$ است. از طریق موجک، سیگنال‌های مورد مطالعه، به شکلها و

حالتهای مفیدتری تبدیل می‌شوند که این تبدیل را تبدیل موجک می‌گویند[۱۳].

جهت محاسبه‌ی کارآمد تبدیلات موجک، «تجزیه و تحلیل چند تفکیکی»[۱۴] (MRA) توسط «مالات و مایر»[۱۵] معرفی شد. با استفاده از MRA، یک سیگنال یا یک داده، به‌صورت ترکیبی از داده‌های با آهنگ تغییر کند[۱۶] (آهسته) (داده‌هایی که به آهستگی تغییر پیدا می‌کنند) و داده‌های با آهنگ تغییر تند[۱۷] (داده‌هایی که در یک زمان کوتاه به شدت تغییر پیدا می‌کنند)، مشاهده می‌شود. MRA، پلی است که ارتباط بین تجزیه و تحلیل موجک یک سیگنال و فیلترینگ آن سیگنال را برقرار می‌کند. با استفاده از MRA و با انتخاب مناسب سطح تفکیک، تقریب خوبی برای داده به دست خواهد آمد.

در تبدیل موجک هار (غیر نرمال)، فیلترهای پایین گذر و بالا گذر به ترتیب عبارتند از $\{h_0, h_1\}$ و $\{g_0, g_1\}$ که در آن h_0=1/2، h_1=1/2، g_0=1/2 و g_1=-1/2 است. مطابق با الگوریتم «مالات و مایر» ، با اعمال فیلترهای پایین‌گذر (بالاگذر) بر روی مجموعه ضرایب تقریب (موجک) در یک سطح، مجموعه ضرایب تقریب (موجک) در سطح پایین به دست خواهد آمد. این الگوریتم تا رسیدن به ضرایب سطح صفر به صورت بازگشتی اجرا می‌شود.

۲-۲- حریم خصوصی تفاضلی

یکی از قوی‌ترین تعریف‌هایی است که توسط دی‌ورک [۹] درباره‌ی حریم خصوصی ارائه گردید. این تعریف، با در نظر گرفتن قدرت محاسباتی فرد مهاجم، حریم خصوصی داده‌ها را اندازه‌گیری می‌کند [۱۰]. از طریق حریم خصوصی تفاضلی، مطمئن خواهیم شد که حذف یا اضافه کردن یک رکورد در مجموعه‌ی داده‌ها، تاثیری در خروجی الگوریتم‌ها نخواهد داشت. به عبارت دیگر، از فاش شدن اطلاعات یک رکورد، جلوگیری خواهد کرد.

دو مجموعه داده‌ی[۱۸] T و $\acute{T}$ که فقط در یک مشخصه[۱۹] با هم تفاوت دارند، را در نظر بگیرید. اگر f تابعی روی آنها باشد به طوری که $f(T) = x$ و $f(\acute{T}) = \acute{x}$ باشند در این صورت گفته می شود که سیستم دارای خاصیت حریم خصوصی تفاضلی است اگر خروجی های x و $\acute{x}$ تقریباً با هم برابر باشند. به عبارت دیگر، کاربر متوجه تغییرات صورت گرفته در مجموعه داده‌ی T نباشد. از این‌رو به مکانیزمی مانند A نیاز داریم تا ارتباط بین خروجی واقعی و درست تابع f و کاربر را برقرار کند. مکانیزم A با دریافت خروجی تابع f، طوری آن را تغییر داده و در اختیار کاربر قرار می‌دهد تا کاربر متوجه تغییرات صورت گرفته در T نباشد. یکی از این تغییرات، اضافه کردن توزیع لاپلاس در خروجی f است. با کمی تغییر، تعریف زیر را داریم [۱۱]:

تعریف (۱) [۱۲]. مکانیزم A خاصیت حریم خصوصی تفاضلی-(ϵ, δ) دارد هرگاه برای تمام مجموعه داده‌های T و Ť (که فقط در یک مشخصه با هم تفاوت دارند) و تمام مجموعه داده‌های خروجی $\hat{D} \subseteq Range(A)$، رابطه‌ی زیر برقرار باشد:

$$Pr[A(T) \in \hat{D}] \leq e^{\epsilon}.Pr[A(\acute{T}) \in \hat{D}]+\delta$$

در حریم خصوصی تفاضلی-ϵ، مقدار $\delta = 0$ است.

در حریم خصوصی تفاضلی فرضی در مورد دانش پیش زمینه‌ی کاربر گرفته نمی شود به عبارت دیگر داشتن دانش پیش زمینه مهم نیست. حساسیت-L_1[۲۰] یکی از کلیدی‌ترین مفاهیمی است که در ساختن داده‌های با خاصیت حریم خصوصی تفاضلی از آن استفاده شده است و به‌صورت زیر تعریف می‌شود:

تعریف (۲). تابع $f: X^* \to R^k$ را در نظر بگیرید. حساسیت-L_1 تابع f عبارت است از $S(f) = \max_{T, \acute{T}} |f(T) - f(\acute{T})|$ به طوریکه مجموعه داده‌های $T, \acute{T}$ فقط در یک مشخصه با هم اختلاف داشته باشند.

برای ایجاد داده‌هایی با خاصیت حریم خصوصی تفاضلی، راه‌ها و روش‌های متفاوتی وجود دارند. مکانیزم لاپلاس یکی از آنها است. در [۹] مکانیزم لاپلاس به صورت زیر تعریف شده است:

تعریف (۳). اگر $f: X^* \to R^k$ یک تابع باشد، مکانیزم لاپلاس به‌صورت زیر تعریف می‌شود:

$$M_L(D, f(.)) = f(D) + (y_1, y_2, \dots, y_k)$$

که y_i ها، به‌صورت مستقل از توزیع لاپلاس $Lap(\frac{S(f)}{\epsilon})$ انتخاب می‌شوند.

قضیه ۱ [۱۳]. اگر F نشان دهنده‌ی مجموعه‌ی از توابع حقیقی با حساسیت $S(F)$ باشد و اگر A الگوریتمی باشد که به خروجی هرکدام از این توابع، خدشه‌ی با توزیع لاپلاس که دارای میانگین صفر و دامنه‌ی λ هستند را اضافه کند در این صورت مکانیزم A خاصیت «حریم خصوصی تفاضلی- $\left(\frac{S(F)}{\lambda}\right)$» خواهد داشت.

۱- کارهای مرتبط

دو روش کلی جهت حفظ حریم خصوصی داده‌ها در الگوریتم‌های داده‌کاوی وجود دارند که عبارتند از [۱۴] : PBA [21] و TBA [۲۲]. در PBA، یک محیط توزیعی از داده‌ها در نظر گرفته می‌شود که در آن، مالکان داده تمایل دارند تا داده‌های خود را بدون اینکه حریم خصوصی آنها نقض گردد، به منظور داده‌کاوی در اختیار دیگران قرار دهند. در این حالت، فرضی در مورد نحوه‌ی تقسیم‌بندی داده‌ها

(افقی یا عمودی) در بین افراد یا سازمانها در نظر گرفته می‌شود و معمولاً، الگوریتم‌های شناخته شده موجود را طوری تغییر می‌دهند که بتوان از آنها جهت حفظ حریم خصوصی داده‌ها، استفاده کرد. پروتکل‌های پیشنهاد شده (مانند «محاسبات چند جانبه امن»[۲۳] (SMC))، اطلاعات ارسالی بین افراد یا سازمانها را طوری کنترل می‌کنند تا از عدم انتشار اطلاعات حساس مربوط به داده‌ها مطمئن گردند. در TBA، ضمن تبدیل داده‌های اصلی به داده‌های مخدوش شده، سعی می‌کنند تا شباهت بین داده‌های اصلی و تبدیل شده حفظ گردد. به عبارت دیگر، عملگر تبدیل طوری انتخاب می‌شود تا ضمن مخدوش شدن داده‌ها، نتایج بهتری از داده‌کاوی به دست آورند. الگوریتم‌های PBA نسبت به الگوریتم‌های TBA، نتایج دقیق‌تری را در داده‌کاوی داده‌ها به دست می‌آورند. در صورتی که از الگوریتم PBA در داده‌کاوی داده‌ها استفاده گردد، امنیت داده‌ها قابل اثبات خواهد بود [۱۵]. در حالت کلی، الگوریتم‌های PBA، به دلیل سربار زیاد پیامهای ارسال شده در بین کاربران، از کارایی کمتری برخوردار هستند.

الگوریتم‌های مختلفی بر اساس روش‌های PBA و TBA ارائه شده‌اند.. اولی‌ویرا[۲۴] و زایانس[۲۵] [۱۶]، الگوریتمی بر مبنای TBA برای داده‌های توزیع شده‌ی افقی ارائه کردند. در[۱۶, ۱۷]، از «تبدیلات هندسی»[۲۶] شامل انتقال[۲۷]، مقیاس‌پذیری[۲۸] و چرخش[۲۹]، جهت مخدوش کردن داده‌های اصلی استفاده شد. یکی از مشکلات تبدیلات اقلیدسی، آسیب‌پذیر بودن آنها در مقابل حملات «ورودی-خروجی شناخته شده»[۳۰] است. در این نوع حمله، فرد مهاجم علاوه بر داشتن تعدادی از داده‌های مخدوش شده، داده‌های اصلی متناظر با آنها را نیز در اختیار دارد. [۱۶]، به منظور رفع این مشکل، از متغیرهای تصادفی یکنواخت یا نرمال استفاده کرد. به این ترتیب که این متغیرها را به ارزش مقداری مشخصه‌های حساس هر رکورد اضافه کرده سپس آنها را منتشر می‌کنند و می‌توان آن را به‌صورت $y_i = x_i + r_i$ نشان داد که x_i، نشان دهنده‌ی مقدار اصلی و r_i نشان دهنده‌ی متغیر تصادفی یکنواخت یا متغیر تصادفی نرمال است. این روش، برای داده‌های با وابستگی بالا[۳۱]، مناسب نیست و امکان آسیب‌پذیری آن از طریق روش‌هایی مانند PCA و تحلیل بیز وجود دارد. به این نوع حمله، «حمله آنالیز-ویژه»[۳۲] گفته می‌شود. [۱۸] الگوریتمی جهت حل این مشکل ارائه کرد که در آن از خدشه‌های دارای کوواریانس یکسان با داده‌های اصلی استفاده شد.

در [۱۹]، داده‌ها به مشخصه‌های مختلفی تقسیم شده سپس بین کاربران توزیع می‌گردد و در نهایت، جهت حفظ حریم خصوصی داده‌ها از پروتکل SMC که توسط یاو[۳۳] ارائه شده بود استفاده شد. در این الگوریتم، تمام محاسبات بر پایه‌ی رمزگذاری و رمزگشایی داده‌ها است. از این‌رو، دقت داده‌کاوی به دلیل مخدوش نشدن داده‌ها حفظ می‌گردد. الگوریتم ارائه شده در این روش، روی تک تک داده‌های یک مجموعه داده اعمال می‌شود از این‌رو، یکی از اشکالات این روش، زیاد بودن پیچدگی ارتباط بین کاربران است.

نوع دیگری از مخدوش‌سازی اطلاعات، مخدوش‌سازی ضربی[۳۴] است. این نوع مخدوش‌سازی بر پایه لم جانسون-لیندن‌اشتراس[۳۵] [۲۰] بنا نهاده شده است. اگر داده‌های با بعد بالاتر را به داده‌های با بعد پایین‌تر نگاشت کنیم، فاصله بین داده‌ها، تقریباً حفظ شده و به دلیل از دست رفتن برخی اطلاعات در هنگام نگاشت از بعد بالاتر به بعد پایین‌تر، بازیابی اطلاعات اصلی با استفاده از داده‌های با بعد پایین‌تر، امکان‌پذیر نخواهد شد. PCA، یکی از بهترین روش‌ها جهت استفاده در کاهش ابعاد داده‌ها است. با استفاده از این روش، مشخصه‌های با واریانس بالاتر انتخاب (نسبت به سایر مشخصه‌های یک مجموعه داده) و سپس، داده‌ها را به این مشخصه‌ها نگاشت می‌کنند. پیچیدگی زمانی این روش $o(nd^2) + o(d^3)$ است که n، بعد داده‌ی اصلی و d، بعد داده‌ی به‌دست آمده از نگاشت است. به دلیل بالا بودن پیچدگی زمانی الگوریتم فوق برای داده‌های با ابعاد بالا، در [۲۱, ۲۲] روشی به نام «افکنش تصادفی»[۳۶] (RP) ارائه شد. RP نسبت به PCA، از دقت پایین‌تر برخوردار است ولی پیچیدگی زمانی کمتری نسبت به آن دارد. RP، را می‌توان به‌صورت $Y = XM$، که در آن، X مجموعه داده‌ی اصلی با بعد $n \times d$ و M، ماتریس تصادفی با بعد $d \times e$ است، نشان داد. درایه‌های ماتریس M، دارای «توزیع مستقل و یکسان»[۳۷] هستند وY نیز، ماتریس حاصل از نگاشت داده‌های با بعد n به بعد e است. انتخاب ماتریس تصادفی M، یکی از مهمترین مسایلی است که در این الگوریتم مورد بررسی قرار می‌گیرد [۲۳].

الگوریتم‌های فوق، نه تنها هیوریستیک‌مبنا بوده و حریم خصوصی قابل اثباتی ندارند بلکه در مقابل دانش پیش‌زمینه کاربران نیز آسیب پذیر هستند. بنابراین، امکان استفاده از حملات «پیوند رکوردی»، «پیوند مشخصه‌ای»، «پیوند جدولی» و «حملات احتمالی» به منظور آسیب رساندن به حریم خصوصی داده‌ها وجود خواهد داشت. برای این منظور، مفهوم «حریم خصوصی تفاضلی[۳۸]» [۹] برای غلبه بر مشکل دانش پیش‌زمینه کاربران ابداع شد. در این مفهوم، حذف یا اضافه کردن یک مشخصه جدید در داده‌ها، تاثیری در خروجی الگوریتمهای داده‌کاوی نخواهد داشت. الگوریتم‌های مختلفی در محیط‌های تفاضلی ارائه شده‌اند. در [۲۴] و [۲۵]، الگوریتم‌هایی به منظور محاسبه‌ی نقاط مرکزی الگوریتم k-means ارائه شده‌اند. پایه و اساس الگوریتم ارائه شده در [۲۴]، ساختاری به نام SULQ [39] است. در [۲۶] از فرایندی به نام GSB [۴۰] به منظور انتشار امن نقاط

مرکزی الگوریتم k-means استفاده شده است. از طریق ساختار GSB امکان انتشار امن تابعی مانند $f(x)$ بر روی مجموعه داده‌ی x وجود خواهد داشت. در [۲۷] ساختاری به نام PINQ [۴۱] (که در واقع یک ساختار توسعه پذیر برای تجزیه و تحلیل دادها است) به منظور حفظ بی قید و شرط حریم خصوصی داده‌های مورد نظر ارائه شد. در [۲۷] با استفاده از ساختار PINQ ویرایش ساده‌ای از الگوریتم k-means (که در مرجع [۲۴] معرفی شده است) پیاده‌سازی شده است. در [۲۸] الگوریتمی به منظور ایجاد ساختارهای هندسی مبتنی بر حریم خصوصی تفاضلی به نام Core-Sets معرفی شد به این صورت که ابتدا، الگوریتم فوق را بر روی مجموعه داده اولیه T اعمال کرده تا Core-Setsی به نام C را ایجاد نماید. سپس تمام الگوریتم‌های خوشه‌بندی مبتنی بر k-median به جای اجرا روی T بر روی C اجرا خواهند شد.

در [۲۹]، الگوریتمی برای حفظ حریم خصوصی در «کاوش دنباله‌های متناوب[۴۲]» مبتنی بر حریم خصوصی تفاضلی ارائه شده است. این الگوریتم از دو مرحله تشکیل شده است. در مرحله اول، پیش پردازشی روی داده‌ها انجام می‌دهد و در مرحله دوم، عمل کاوش روی داده‌ها را اجرا می‌کند. در مرحله پیش‌پردازش از تقسیم‌بندی جدیدی برای تغییر در پایگاه دادها استفاده کرده است تا از این طریق، محرمانگی داده‌ها را ارتقاء دهد. در [۳۰]، یک الگوریتم مبتنی بر حریم خصوصی تفاضلی برای حل مشکل عدم یکنواختی داده های چند رسانه ای[۴۳] دو بعدی ارائه داده است. الگوریتم های معمولی حفظ حریم خصوصی ، فضای داده های مکانی[۴۴] را به شبکه تقسیم می کنند و سپس به هر شبکه در همان مقیاس نویز می دهند. به منظور حل مشکل افزایش خطای نسبی و کاهش دقت ، یک الگوریتم تخصیص پویای خدشه‌های حریم خصوصی تفاضلی پیشنهاد شده است. در [۳۱]، با استفاده از خدشه‌های لاپلاس، الگوریتمی برای حفظ حریم خصوصی داده‌ها ارائه شده است. در [۳۲]، انتشار داده‌ها در شبکه‌های اجتماعی مورد بررسی قرار گرفته است و الگوریتمی مبتنی بر حریم خصوصی تفاضلی ارائه شده است که از طریق آن می‌توان داده‌های حساس را در اختیار دیگران قرار داد. در [۳۳] با استفاده از تبدیلات فوریه، حریم خصوصی تفاضلی را در رده‌بندی داده‌های یک شبکه هوشمند[۴۵] ایجاد کرده است. در [۳۴]، از طریق اضافه کردن خدشه به داده‌ها، مجموعه داده‌های تفاضلی را در یک محیط رایانش لبه‌ای[۴۶]، ایجاده کرده و با سایر افراد به اشتراک می‌گذارند. در [۳۵]، الگوریتمی برای رده‌بندی داده‌ها معرفی شده است که پایه و اساس آن، اضافه کردن خدشه‌های لاپلاس (یکی از روش‌های ایجاد داده‌های مبتنی بر حریم خصوصی تفاضلی) به داده‌ها می‌باشد. یکی از معایب اصلی الگوریتم ارائه شده در [۳۵] این است که به منظور افزایش کارایی داده‌های ایجاد شده، الگوریتم K-NN (K-Nearest Neighbor) را تغییر داده است. در [۳۶] الگوریتمی برای حفظ حریم خصوصی در رده‌بندی داده‌ها در محیط رایانش ابری ارائه شده است و به منظور رسیدن به حریم خصوصی تفاضلی داده‌ها از خدشه‌های لاپلاس استفاده شده است. در [۳۷, ۳۸]، الگوریتم‌هایی برای حفظ حریم خصوصی رده‌بندی داده‌ها ارائه شده است یکی از مهمترین مشکلات این الگوریتم‌ها این است که یا حریم خصوصی قابل اثباتی ندارند یا اینکه (در صورت استفاده از مفهوم حریم خصوصی تفاضلی) همزمان، قادر به کاهش ابعاد داده‌های اولیه نیستند. در مرجع [۳۹]، الگوریتمی تفاضلی برای حفظ حریم خصوصی داده‌ها در خوشه‌بندی ارائه شده است. ایراد این الگوریتم این است که سطح توقف تجزیه داده‌ها به صورت خودکار مشخص نمی‌گردد. اکثر این الگوریتم‌ها در مقابل دانش پیش‌زمینه کاربران آسیب پذیر هستند.

با توجه به مطالب گفته شده، حفظ حریم خصوصی داده‌ها در داده‌کاوی، یکی از مهمترین مسائلی است که پژوهشگران روی آن کار می‌کنند. یکی از روش‌های حفظ حریم خصوصی، روش TBA است. در این روش، داده‌های اصلی را طوری مخدوش می‌کنند که شباهت بین داده‌های اصلی و مخدوش شده حفظ گردد. الگوریتم‌ها مختلفی مبتنی بر TBA ارائه شده است. اکثر این الگوریتم‌ها، داده‌هایی را تولید می‌کنند که حریم خصوصی آنها با استفاده از مدل ریاضی قابل اثبات نیست. برای رفع این مشکل، مفهوم حریم خصوصی تفاضلی ابداع شد. در حریم خصوصی تفاضلی، داده‌های اصلی را با اضافه کردن خدشه‌هایی مخدوش می‌کنند ولی تفاوت آن با الگوریتم‌های قبلی این است که حریم خصوصی این داده‌ها از طریق مدل ریاضی قابل اثبات است. یکی دیگر از مشکلات در داده‌کاوی، ابعاد بالای داده‌ها است. هر قدر ابعاد داده‌ها کاهش پیدا کند به همان میزان دقت داده‌کاوی کاهش پیدا می‌کند. از این‌رو، کاهش ابعاد داده‌ها باید طوری باشد که تعادلی بین کاهش ابعاد و دقت داده‌کاوی برقرار گردد. تا کنون، با جستجوهای صورت گرفته، الگوریتمی که بتواند چهار مشخصه (۱) قابل اثبات بودن حریم خصوصی داده‌ها با مدل ریاضی، (۲) کاهش ابعاد داده‌ها، (۳) اعمال روی داده‌های توزیع شده و (۴) رده‌بندی را همزمان مورد بررسی قرار دهد، پیدا نشد. از این‌رو در این مقاله، الگوریتمی با استفاده از مفاهیم «تبدیل موجک» و «حریم خصوصی تفاضلی» برای ارائه یک راه حل قابل قبول برای این سه مورد پیشنهاد گردید.

۴- الگوریتم پیشنهادی

در این بخش الگوریتم پیشنهادی بر روی داده‌های توزیعی عمودی ارائه شده است.

۴-۱- فرضیات مساله

فرض‌های در نظر گرفته شده در این مقاله، عبارتند از (۱) مجموعه داده‌هایی که الگوریتم‌های ارائه شده بر روی آنها اعمال خواهند شد را به‌صورت یک ماتریس دو بعدی $T_{m\times n}$ در نظر می‌گیریم. به عبارت دیگر، این ماتریس دارای m سطر و n ستون است که هر سطر اطلاعات مربوط به یک موجودیت خاص با n مشخصه را نشان می‌دهد، (۲) مقادیر مربوط به مشخصه‌های هر رکورد را به‌صورت یک دنباله‌ی عددی در نظر می‌گیریم و (۳) مقدار مثبت T_{Max} چنان وجود دارد که قدر مطلق تمام درایه‌های ماتریس $T_{m\times n}$، کوچکتر یا مساوی T_{Max} است.

۴-۲- مدل تهدید

در الگوریتم‌های توزیعی، یکی از دو مدل ارتباطی زیر را در بین سیستم‌ها در نظر می‌گیرند [۴۰]: ۱) مدل بدخواه[۴۷] ۲) مدل نیمه-صادق[۴۸]. در مدل بدخواه، طرف بدخواه مجبور نیست تا قواعد در نظر گرفته شده در پروتکل را رعایت کند و امکان دارد با استفاده از داده‌هایی که از سایرین دریافت می‌کند، شروع به سازمان دادن حمله‌هایی بر ضد افراد شرکت کننده نماید. ولی در مدل نیمه-صادق، فرض بر این است که شرکت کنندگان قابل اعتماد و کنجکاو هستند. در این مدل، افراد شرکت کننده، قواعد در نظر گرفته شده را رعایت کرده ولی ممکن است بخواهند اطلاعاتی بیشتر از آنچه برای یک فرد در نظر گرفته شده است را به دست آورند. در اکثر الگوریتم‌های ارائه شده برای حفظ حریم خصوص داده‌ها، از مدل نیمه-صادق برای ارائه الگوریتم خود استفاده می‌کنند. در این مقاله فرض می‌شود که افراد[۴۹] شرکت کننده در الگوریتم، نیمه-صادق هستند. همچنین از یک «طرف سوم»[۵۰] (TTP)، به منظور کنترل ارتباط بین افراد شرکت کننده در الگوریتم استفاده شده است. TTP می‌تواند مستقل یا یکی از افراد شرکت کننده در مساله باشد. TTP را هم به عنوان یک فرد نیمه-صادق در نظر می‌گیریم. همچنین فرض می‌کنیم که سایت‌ها با استفاده از یک سیستم رمزگذاری عمومی، اطلاعات را به یکدیگر ارسال یا از یکدیگر دریافت می‌کنند. به عبارت دیگر، هر سایت دارای یک کلید عمومی و یک کلید خصوصی است. همچنین، TTP، کلید عمومی تمام سایت‌ها و تمام سایت‌ها نیز کلید عمومی TTP را می‌دانند.

۴-۳- جزییات الگوریتم پیشنهادی

به منظور طراحی الگوریتم برای داده‌های توزیع شده به صورت عمودی، از خاصیت خطی تبدیلات موجک [۴۱] استفاده می‌کنیم. این خاصیت بیان می‌کند که اگر مجموعه‌ی داده‌ی T را به دو مجموعه داده‌ی جدا از هم عمودی T_1 و T_2 تقسیم کنیم (یعنی $T = (T_1, T_2)$) در این صورت رابطه‌ی زیر را داریم: (HWT^{51} مخفف تبدیل موجک هار است)

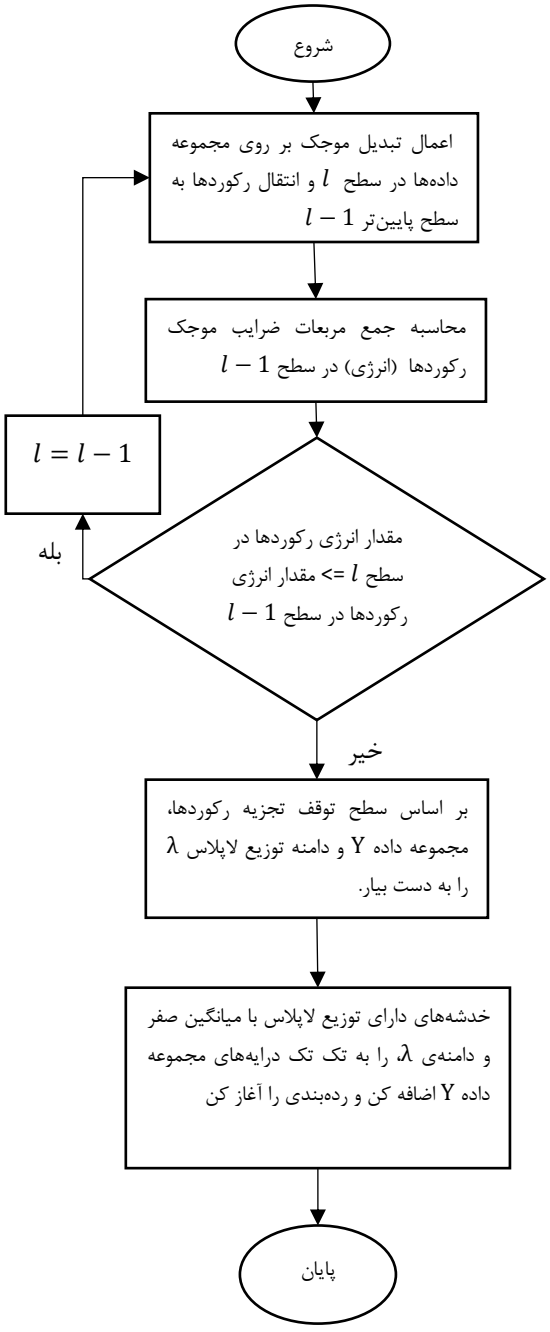
$$HWT(T) = HWT(T_1) + HWT(T_2) \qquad (۱)$$

رابطه‌ی (۱) بیان می‌کند که اگر تبدیل موجک هار را بر روی تک تک مجموعه‌های T_1 و T_2 اعمال کنیم و سپس آنها را با هم جمع کنیم، نتیجه‌ی به‌دست آمده همان نتیجه‌ی حاصل از اعمال تبدیل موجک هار بر روی مجموعه داده‌ی T است. لازم به ذکر است که اندیس مجموعه داده‌های T_1 و T_2 در رابطه‌ی $T = (T_1, T_2)$، به ترتیب برابر با ۱ و ۲ است. در تبدیل موجک هار، اندازه هر مجموعه داده باید توانی از دو باشد. از این‌رو اگر اندازه هر مجموعه داده توانی از دو نباشد باید به اندازه کافی عدد صفر به انتهای آن اضافه شود تا به توان دو برسد. به منظور افزایش دقت رده‌بندی و جلوگیری از همپوشانی ضرایب تولید شده، از روش زیر برای اضافه کردن عدد صفر به انتهای مجموعه داده‌ها استفاده می‌شود:

فرض کنید که تعداد سایت‌ها برابر با G بوده و مجموعه داده‌ی T_g با ابعاد $m \times n_g$ متعلق به سایت g باشد. ماتریس T_g را به‌صورت $T_g = \begin{pmatrix} t_{11}^g & \cdots & t_{1n_g}^g \\ \vdots & \ddots & \vdots \\ t_{m1}^g & \cdots & t_{mn_g}^g \end{pmatrix}$ در نظر بگیرید. واضح است که $1 \le g \le G$ است. همچنین فرض کنید که نماد $|.|$ نشان دهنده‌ی تعداد مشخصه‌های (اندازه‌ی) یک مجموعه داده باشد. ترکیب کلی عمودی $T_{m\times n} = (T_1, \dots, T_g, \dots, T_G)_{m\times n}$ را در نظر بگیرید که در آن رابطه $n = \sum_{g=1}^{G} n_g$ برقرار است. واضح است که اندیس (موقعیت) T_g در T برابر با g است. کوچکترین عدد صحیح $\acute{n}$ را طوری انتخاب می‌کنیم که بزرگتر یا مساوی n بوده و توانی از ۲ باشد. هر سایت g، مجموعه داده‌ی R_g با ابعاد $m \times (\acute{n} \times G)$ را طوری می‌سازد که درایه‌های آن برابر با صفر باشد. سپس، درایه‌ی t_{ij}^g را از ماتریس T_g برداشته و آن را در سطر i ام و ستون $((g-1) \times \acute{n} + j)$ ام ماتریس R_g کپی می‌کنیم. واضح است که فاصله‌ی اقلیدسی دو رکورد از ماتریس T_g با فاصله‌ی اقلیدسی رکوردهای متناظر با آنها در ماتریس R_g برابر است. بنابراین، بدون از دست دادن اطلاعات، امکان استفاده از R_g به جای T_g وجود خواهد داشت. از این‌رو، اگر رکوردهای ماتریس R_g را در سطح اولیه

$\log(\acute{n})$ در نظر گرفته و تبدیل موجک هار را تا سطح ۰ بر روی آنها اعمال کنیم، به هیچ عنوان همپوشانی ضرایب تقریب در سطوح مختلف اتفاق نخواهد افتاد و این باعث افزایش دقت رده‌بندی خواهد شد. برای مثال، مجموعه داده‌ی کلی T با یک رکورد و ۶ مشخصه را به‌صورت $(4, 2, 1, 3, 5, 1)_{1\times6}$ در نظر بگیرید. همچنین فرض کنید که T به دو قسمت عمودی $T_1=(4, 2, 1)_{1\times3}$ و $T_2 = (3, 5, 1)_{1\times3}$ تقسیم شده و در دو سایت مختلف پخش شده باشند. بدون اینکه خللی در بیان مساله پیش بیاید فرض می‌کنیم T_1 در سایت ۱ و T_2 در سایت ۲ قرار داده شده‌اند. به منظور اعمال تبدیل موجک هار بر روی مجموعه داده‌های T_1 و T_2، طول رکوردهای آنها باید توانی از ۲ باشد. در غیر این‌صورت، به تعداد کافی صفر به تک تک رکوردهای این مجموعه‌ها اضافه می‌شود تا اندازه‌ی هر رکورد توانی از ۲ باشد. در مثال فوق مقادیر $G = 2$، $n = 6$، $n_2 = 3$، $n_1 = 3$، $m = 1$ و $\acute{n} = 8$ است. جدول ۱، تجزیه مجموعه داده‌های متعلق به سایت‌های ۱ و ۲ (مثال فوق) را با استفاده از روش گفته شده برای اضافه کردن صفر به مجموعه داده‌ها را نشان می‌دهد. همان‌طوری که ملاحظه می‌شود، اگر از رابطه‌ی ۱، برای محاسبه‌ی ضرایب تقریب و ضرایب موجک مجموعه داده‌ی ترکیبی کلی T استفاده شود، در این صورت هیچ گونه هم‌پوشانی در بین ضرایب تقریب و ضرایب موجک مجموعه داده‌های متعلق به سایت‌ها اتفاق نخواهد افتاد.

شکل ۱، فلوچارت الگوریتم پیشنهادی را نشان می‌دهد. در محله اول، تبدیلات موجک روی مجموعه داده‌ها اعمال می‌شود. در مرحله بعد، انرژی رکوردها در سطح پایین‌تر با انرژی رکوردهای سطح بالاتر مقایسه شده و بر اساس آن، سطح توقف تجزیه مشخص می‌گردد. در نهایت با توجه به سطح توقف، میزان خدشه‌ها مشخص شده و به مجموعه داده‌ها اضافه می‌گردد. همچنین، شکل ۲، شبه کد الگوریتم ارائه شده بر روی داده‌های توزیع شده به‌صورت عمودی را نشان می‌دهد. خط ۱ و ۲ الگوریتم، پیش پردازشی را بر روی مجموعه داده اولیه به منظور اعمال روش گفته شده، انجام می‌دهد. در بین خطوط ۳ تا ۱۰، تبدیل موجک هار بر روی مجموعه داده اعمال می‌شود تا آن را تجزیه نماید. در مرجع [۴۲]، تبدیل موجک هار را بر روی مجموعه داده‌های اعمال کرده سپس مجموع انرژی مربوط به ضرایب موجک تمام رکوردهای سطح جاری را با هم جمع کرده و آن را با مقدار حاصل از جمع انرژی سطح قبلی (که قبلاً محاسبه کرده بود) مقایسه می‌نماید. اگر جمع انرژی سطح قبلی کمتر از جمع انرژی سطح جاری باشد در این صورت پیغام «توقف تجزیه» را به تمام سایت‌ها ارسال کرده و سطح قبلی را به عنوان سطح توقف تجزیه مشخص می‌کند. در غیر این صورت، تبدیل موجک بر روی داده‌ها تا رسیدن به شرط گفته شده در بالا ادامه پیدا می‌کند. در [۴۲] نشان داده شده است که ضرایب تقریب به دست آمده یک مجموعه داده که شرط انرژی گفته شده در آن برقرار باشد، جایگزین خوبی برای مجموعه داده‌های اولیه خواهد بود.



شکل ۱: فلوچارت الگوریتم پیشنهادی

به منظور افزایش دقت رده‌بندی، احتمال اینکه در سطوح پایین‌تر بتوانیم به یک مجموعه داده‌ای برسیم که کارایی خوبی داشته باشد، وجود دارد. از این‌رو در خط ۱۱، برا اساس دقت به دست آمده برای مجموعه داده، سطح مورد نظر برای توقف را انتخاب می‌کنیم. فرض کنید شرط توقف تجزیه برای مقدار S برابر با ۲ برقرار باشد. مطابق

با خط ۱۲ الگوریتم ***PrivacyCL***، مقادیر sa_1، ea_1، sa_2 و ea_2 به ترتیب برابر با ۱، ۲، ۵ و ۶ محاسبه خواهند شد. از این‌رو مقادیر d_1 و d_2 برابر با ۲ خواهند بود. سپس، ماتریس‌های Y_1*=(3, 0.5)* و Y_2*=(4, 0.5)* بر اساس خط ۱۳ محاسبه می‌شوند. فرض کنید مقدار متغیر ***hasnegativevalues*** برابر با ***false*** باشد. در این صورت مقدار ϑ برابر با ۱ خواهد شد (خط ۱۴). با فرض $\epsilon = 1$، ماتریس‌های تصادفی Δ_1*=(0.03, -0.05)* و Δ_2*=(0.39, -0.80)* را ساخته و سپس $P_1=Y_1+\Delta_1$*=(3.03,0.045)* و $P_2=Y_2+\Delta_2$*=(4.39,-0.30)* را محاسبه کرده و آنها را منتشر می‌کنیم. در خط ۲، به منظور غلبه بر مشکل هم‌پوشانی ضرایب تقریب در توزیع عمودی داده‌ها (که باعث کاهش کارایی داده‌های تولید شده می‌شود)، از روشی که برای اضافه کردن مقادیر صفر به مجموعه داده گفته شد، برای اضافه کردن صفر به مجموعه داده‌ی R_g استفاده شده است. بنابراین، صفرهایی را به سمت چپ و راست مجموعه داده‌ی R_g اضافه می‌کنیم. از این‌رو، بعد از توقف تجزیه در سطح S، ضرایب واقعی سایت g، در بین اندیس‌های sa_g و ea_g قرار خواهند گرفت. بنابراین، فقط ضرایب قرار گرفته در بین اندیس‌های sa_g و ea_g در مجموعه داده‌ی Y_g قرار داده می‌شود.

۵- ارزیابی الگوریتم ارائه شده

در این بخش، الگوریتم ارائه شده را بر اساس حریم خصوصی و دقت رده‌بندی ارزیابی می‌کنیم.

۵-۱- ارزیابی حریم خصوصی

ابتدا حریم خصوصی تفاضلی الگوریتم ارائه شده، بررسی می‌گردد. برای این منظور قضیه زیر را دارم.

قضیه(۲): الگوریتم پیشنهادی ***PrivacyCL*** خاصیت حریم خصوصی تفاضلی-ϵ دارند.

اثبات. فرض کنید مجموعه داده‌ی $\mathrm{Y_g}$ با ابعاد $m \times d_g$ در اثر اجرای خطوط ۱ تا ۱۳ الگوریتم ***PrivacyCL*** (شکل ۲) بر روی مجموعه داده‌ی اصلی $\mathrm{T_g}$ به ابعاد $m \times \mathrm{n_g}$ به دست آمده باشد. همچنین فرض کنید که فقط یکی از مشخصه‌های مجموعه داده‌ی اصلی $\mathrm{T_g}$ را به اندازه‌ی η تغییر بدهیم. بدون اینکه خللی در اثبات قضیه وارد شود، فرض کنید مقدار درایه‌ی $(t_g)_{\alpha\beta}$ $(1 \le \alpha \le m, 1 \le \beta \le n_g)$ به مقدار $(\mathrm{t_g})_{\alpha\beta} + \eta$ تغییر پیدا کرده باشد. مجموعه داده‌ی به دست آمده را $\acute{\mathrm{T}}_\mathrm{g}$ می‌نامیم. واضح است که با اجرای خطوط ۱ تا ۱۳ الگوریتم ***PrivacyCL*** بر مجموعه داده $\acute{\mathrm{T}}_\mathrm{g}$، فقط درایه $(\mathrm{y_g})_{\alpha\theta}$ مربوط به مجموعه داده‌ی $\mathrm{Y_g}$ تغییر پیدا می‌کند که در آن مقدار θ برابر با $1 + \left\lfloor \frac{2^S \times \beta}{\acute{n}} \right\rfloor$ است چرا که در تبدیل موجک هار، طول فیلترهای بالاگذر و پایین‌گذر برابر با ۲ است و اگر تغییری در یک مشخصه ایجاد شود در این صورت تغییر اعمال شده فقط به یکی از ضرایب به دست آمده درهر سطح پایین منتقل می‌گردد. لازم به ذکر است که در ارائه‌ی الگوریتم ***PrivacyCL*** از حالت غیر نرمال تبدیل موجک هار استفاده شده است. برای اندازه‌گیری مقدار تغییر اتفاق افتاده در درایه $(\mathrm{y_g})_{\alpha\theta}$، مراحل زیر را طی می‌کنیم:

در خط ۲ الگوریتم ***PrivacyCL*** ، درایه‌های $\acute{\mathrm{T}}_\mathrm{g}$ بر عدد $\mathrm{T_{Max}}$ تقسیم می‌شوند. بنابراین اندازه‌ی تغییر اتفاق افتاده در درایه‌ای با سطر αام و ستون βام مربوط به مجموعه داده‌ی $\acute{\mathrm{T}}_\mathrm{g}$ نسبت به درایه‌ی متناظر آن در مجموعه داده‌ی $\mathrm{T_g}$ برابر با $\vartheta = \frac{\eta}{\mathrm{T_{Max}}}$ خواهد بود. در خطوط ۳ تا ۱۱، تبدیل موجک هار تا سطح توقف S، و به‌صورت مستقل بر روی رکوردهای $\acute{\mathrm{T}}_\mathrm{g}$ اعمال می‌شوند. بنابراین تمام ضرایب تقریب به دست آمده برای هر رکورد در سطح S، حاصل جمع دو به دوی تقسیم شده‌ی درایه‌های همان رکورد درسطح اولیه بر $\frac{\acute{n}}{2^S}$ است. از این‌رو مقدار ϑ در سطح S برابر با $\frac{2^S \times \vartheta}{\acute{n}} = \frac{\vartheta}{(\frac{\acute{n}}{2^S})}$ است. لازم به ذکر است که تبدیل موجک هار به‌صورت مستقل بر روی رکوردها اعمال می‌شود. بنابراین، به غیر از درایه‌ی $(\mathrm{y_g})_{\alpha\theta} \in \mathrm{Y_g}$ $(\theta = \left\lfloor \frac{2^S \times \beta}{\acute{n}} \right\rfloor + 1)$، هیچکدام از درایه‌های مجموعه داده‌ی $\mathrm{Y_g}$ تغییر پیدا نخواهند کرد (درایه‌های ماتریس $\mathrm{Y_g}$، همان ضرایب تقریب رکوردها در سطح توقف تجزیه (یا S) هستند).

طبق مطالب گفته شده در بالا، اندازه‌ی تغییر درایه‌ی $(\mathrm{y_g})_{\alpha\theta}$ (نسبت به درایه‌ی متناظر آن که از $\mathrm{T_g}$ به دست آمده است) برابر با $\frac{2^S \times \vartheta}{\acute{n}}$ خواهد شد.

جدول ۱: تجزیه مجموعه داده‌های متعلق به سایت‌ها با استفاده از روش جدید اضافه کردن صفر

سطح	مجموعه داده‌ی متعلق به سایت ۱	مجموعه داده‌ی متعلق به سایت ۲
۳	$\mathrm{R_1}$=(4, 2, 1,0,0,0,0,0,0,0,0,0,0,0,0,0)$_{1\times16}$	$\mathrm{R_2}$=(0,0,0,0,0,0,0,0,3, 5, 1,0,0,0,0,0)$_{1\times16}$
۲	$\mathrm{R_{1_{ac^2}}}$=(3, 0.5, 0, 0,0,0,0,0) ,$\mathrm{R_{1_{wc^2}}}$=(1, 0.5, 0, 0,0,0,0,0)	$R_{2_{ac^2}}$ = (0,0,0,0,4,0.5,0, 0) , $R_{2_{wc^2}}$ = (0, 0, 0, 0,-1,0.5,0,0)
۱	$\mathrm{R_{1_{ac^1}}}$=(1.75, 0, 0, 0) , $R_{1_{wc^1}}$ = (1.25, 0, 0, 0)	$R_{2_{ac^1}}$ = (0,0,2.25,0) , $R_{2_{wc^1}}$ = (0, 0, 1.75, 0)
۰	$R_{1_{ac^0}}$=(0.875, 0) , $R_{1_{wc^0}}$ = (0.875, 0)	$R_{2_{ac^0}}$ = (0,1.125) , $R_{2_{wc^0}}$ = (0,1.125)

الگوریتم پیشنهادی *PrivacyCL* برای داده‌های توزیعی عمودی

- ورودی:

- ماتریس از تقسیمات عمودی $T_{m\times n} = (T_1, T_2, ..., T_G)$: به طوریکه ماتریس T_g $(1 \le g \le G)$ با ابعاد $m \times n_g$، متعلق به سایت g است. همچنین رابطه‌ی $\sum_{g=1}^{G} n_g = n$ برقرار است.
- T_{Max} : قدر مطلق تمام درایه‌های مجموعه داده‌ی کلی $T_{m\times n}$، کوچکتر یا مساوی T_{Max} هستند. مقدار T_{Max} با همکاری تمام سایت‌ها و TTP مشخص می‌شود.
- ϵ : پارامتر مربوط به میزان حریم خصوصی است.
- a : دقت رده‌بندی مورد نظر.
- ***hasnegativevalues*** : که مشخص می‌کند آیا درایه‌های مجموعه داده‌ی $T_{m\times n}$ می‌توانند مقادیر منفی داشته باشند یا خیر؟ (این مقدار با همکاری تمام سایت‌ها و TTP مشخص می‌شود).

- خروجی: ماتریس مخدوش شده‌ی $P_{m\times d} = (P_1, P_2, ..., P_G)$ **:** به طوری‌که ماتریس P_g $(1 \le g \le G)$ با ابعاد $m \times d_g$، مخدوش شده‌ی ماتریس T_g است. همچنین رابطه‌ی $\sum_{g=1}^{G} d_g = d$ برقرار است.

هر سایت g، خطوط ۱ تا ۱۱ را اجرا می‌کند:

۱) کوچکترین عدد صحیح بزرگتر یا مساوی n که توانی از ۲ باشد را پیدا کرده و آن را $\acute{n}$ بنامید. (با همکاری تمام سایت‌ها و TTP مشخص می‌شود).

۲) تمام درایه‌های ماتریس T_g را بر T_{Max} تقسیم کن و مجموعه داده‌ی دو بعدی R_g با ابعاد $m \times (\acute{n} \times G)$ را طوری بساز که درایه‌های آن برابر با صفر باشد. سپس، درایه‌ی t_{ij}^{g} را از ماتریس T_g برداشته و آن را در سطر i ام و ستون $((g-1) \times \acute{n} + j)$ ام ماتریس R_g کپی کن.

۳) قرار بده $l = \log(\acute{n})$ و $EH = +\infty$. (تمام رکوردهای اولیه، در سطح $\log(\acute{n})$ در نظر گرفته می‌شوند).

۴) مراحل ۵ تا ۱۰ را تا خارج شدن از حلقه ادامه بده

۵) با استفاده از تبدیل موجک هار، m رکورد مجموعه داده‌ی R_g را به طور مستقل تجزیه کن تا در سطح $l-1$ قرار بگیرند.

۶) WC_i^{l-1} (ضرایب موجک رکورد i ام در سطح $l-1$ ام) و AC_i^{l-1} (ضرایب تقریب رکورد i ام در سطح $l-1$ ام) را برای هر رکورد به‌دست بیاور.

۷) «جمع مربعات ضرایب موجک» یعنی $EL_g = \sum_m (WC_m^{l-1})^2$ را محاسبه کن و به *TTP* ارسال کن.

۸) *TTP*، جمع انرژیهای ارسال شده توسط همه سایت‌ها را را محاسبه می‌کند تا سطح مورد نظر برای توقف تجزیه داده‌ها مشخص گردد. یعنی مقدار $EL = \sum_{g=1}^{G} EL_g$ توسط *TTP* محاسبه می‌شود.

۹) اگر $EL \le EH$ باشد آنگاه:

۹-۱) قرار بده $EH = EL$ و $l = l - 1$ و برو به مرحله ۵.

۱۰) اگر $EL > EH$ باشد آنگاه:

۱۰-۱)قرار بده $ML = l$ و برو به مرحله ۱۱.

۱۱) از بین سطح های ۰ تا ML سطحی را پیدا کن که اگر الگوریتم *K-NN* (برای مقدار $K=5$) را بر روی ضرایب تقریب در آن سطح، به منظور رده‌بندی، اعمال کنیم در این صورت، دقت رده‌بندی بزرگتر یا مساوی a باشد.

۱۱-۱) در صورت پیدا شدن سطح مورد نظر با شرایط گفته شده:

۱۱-۱-۱) مقدار سطح پیدا شده را در متغیر S قرار بده

در غیر این صورت:

۱۱-۱-۲) مقدار ML را در متغیر S قرار بده.

۱۲) قرار بده $sa_g = (g-1).2^s + 1$، $ea_g = g.2^s - \left\lfloor \frac{2^s \times (\acute{n} - n_g)}{\acute{n}} \right\rfloor$ و $d_g = ea_g - sa_g + 1 = 2^s - \left\lfloor \frac{2^s \times (\acute{n} - n_g)}{\acute{n}} \right\rfloor$.

۱۳) ماتریس Y_g با ابعاد $m \times d_g$ را بساز، سپس تمام ضرایب تقریب سطح S ام رکورد i ام مجموعه داده‌ی R_g که اندیس آنها بین sa_g و ea_g است را به ترتیب از اندیس پایین تا اندیس بالا انتخاب کرده و آن را در رکورد i ام ماتریس Y_g قرار بده.

۱۴) اگر $hasnegativevalues = true$ باشد قرار بده $\vartheta = 2$ وگرنه قرار بده $\vartheta = 1$.

۱۵) قرار بده $\lambda = \frac{2^S \times \vartheta}{n \acute{\times} \epsilon}$.

۱۶) ماتریس تصادفی Δ_g با ابعاد $m \times d_g$ را که درایه‌های آن، متغیرهای تصادفی دارای توزیع لاپلاس با میانگین صفر و دامنه‌ی λ هستند را بساز.

۱۷) قرار بده $P_g = Y_g + \Delta_g$ که در آن ماتریس P_g از بعد $m \times d_g$ می‌باشد.

۱۸) $P = (P_1, P_2, ..., P_G)$ را به خروجی ارسال کن.

شکل ۲: الگوریتم PrivacyCL بر روی داده‌های توزیعی عمودی

مطابق با خط ۱۴، اگر فرض کنیم که درایه‌های ماتریس T_g هم مثبت و هم منفی می‌توانند باشند، در این صورت بیشینه قدر مطلق مقدار ϑ برابر با ۲ خواهد بود زیرا، در بدترین حالت اگر $(t_g)_{\alpha\beta} = T_{Max}(-T_{Max})$ باشد در این صورت طبق فرض در نظر گرفته شده (که قدر مطلق تمام درایه‌های ماتریس T_g، کمتر یا مساوی T_{Max} هستند)، بیشینه مقدار η می‌تواند $2 \times T_{Max}$ $(2 \times T_{Max}) - 2$ باشد (تا رابطه‌ی $|(t_g)_{\alpha\beta} + \eta| \le T_{Max}$ برقرار گردد). از این‌رو، $|\eta| \le 2 \times T_{Max} \Rightarrow \left|\vartheta = \frac{\eta}{T_{max}}\right| \le 2$ خواهد بود (طبق فرض در نظر گرفته شده $|t_{\alpha\beta}| \le T_{Max}$ است). اما اگر فرض کنیم که درایه‌های ماتریس T_g فقط مقادیر مثبت می‌توانند اختیار کنند، در این صورت مطابق با توضیح بالا، بیشینه مقدار قدر مطلق ϑ برابر با

۱ خواهد بود. بنابراین، اندازه‌ی تغییر درایه‌ی $(y_g)_{\alpha\vartheta}$ برابر با $\frac{2^s\times\vartheta}{\acute{n}}$ است که در آن مقدار ϑ برابر با ۲ یا ۱ خواهد بود.

هر کدام از درایه‌های $(y_g)_{ij}$ $(1\le i\le m, 1\le j\le d_g)$ در ماتریس Y_g را می‌توان به عنوان نگاشتی از یک تابع حقیقی f_{ij} در نظر گرفت طوریکه f_{ij} مجموعه داده‌ی اولیه T_g را به $(y_g)_{ij}$ نگاشت کند. به عبارت دیگر $f_{ij}(T_g)=(y_g)_{ij}$ $((1\le i\le m, 1\le j\le d_g))$ است. از این‌رو، مجموعه‌ای شامل توابع حقیقی، به‌صورت $F=\{f_{ij}|1\le i\le m, 1\le j\le d_g\}$ را در نظر می‌گیریم طوریکه درایه‌های T_g را به درایه‌های Y_g نگاشت کند (در واقع، تابع حقیقی f_{ij}، خطوط ۱ تا ۱۱ الگوریتم ***PrivacyCL***، را بر روی درایه‌ی $(y_g)_{ij}$ اجرا می‌کند). فرض کنید که فقط یکی از درایه‌های ماتریس T_g، یعنی $(t_g)_{\alpha\beta}$، را به اندازه‌ی η تغییر داده و مجموعه داده‌ی جدید را $\acute{T}_g$ بنامیم. اگر مجموعه توابع F را به طورمستقل بر روی T_g و $\acute{T}_g$ اعمال کنیم، در این صورت، ماتریس‌های Y_g به دست آمده برای T_g و $\acute{T}_g$، در تمام درایه‌ها به غیر از درایه‌ی $\theta=\left\lfloor\frac{2^s\times\beta}{\acute{n}}\right\rfloor+1$ مقادیر یکسانی را خواهند داشت. مطابق با توضیحات بالا، مقدار تغییر به دست آمده برابر با $\frac{2^s\times\vartheta}{\acute{n}}$ خواهد بود که در آن مقدار ϑ برابر با ۱ یا ۲ است. پس رابطه (۲) را خواهیم داشت.

$$\left|\begin{matrix}f_{ij}(T_g)\\-f_{ij}(\acute{T}_g)\end{matrix}\right| = \begin{cases}0 & where\\ & i\in\{1,2,...,m\}-\{\alpha\}\\ & and\ j\in\{1,2,...,d_g\}\\ & -\{\left\lfloor\frac{2^s\times\beta}{\acute{n}}\right\rfloor+1\}\\ \frac{2^s\times\vartheta}{\acute{n}} & where\ i=\alpha,\\ & j=\left\lfloor\frac{2^s\times\beta}{\acute{n}}\right\rfloor+1\\ & and\ (\vartheta=2\ or\ \vartheta=1)\end{cases} \qquad (۲)$$

بنابراین با استفاده از رابطه‌ی ۲ رابطه ۳ را خواهیم داشت داریم:

$$\sum_{f_{ij}\in F}\left|f_{ij}(T_g)-f_{ij}(\acute{T}_g)\right|=\frac{2^s\times\vartheta}{\acute{n}}\ , \qquad (۳)$$
$$1\le i\le m\ and\ 1\le j\le d_g$$

از این‌رو، با استفاده از تعریف ۲، حساسیت-L_1 مجموعه توابع F برابر با $S(F)=\frac{2^s\times\vartheta}{\acute{n}}$ خواهد شد. پس، با استفاده ازقضیه ۱، اگر متغیرهای تصادفی لاپلاس با میانگین صفر و دامنه‌ی λ را به خروجی‌های هر تابع حقیقی $f\in F$ اضافه کنیم، در این صورت، الگوریتم ***PrivacyCL*** خاصیت حریم خصوصی تفاضلی-ϵ خواهد داشت که در آن، مقدار λ از رابطه‌ی زیر محاسبه می‌شود:

$$\epsilon=\frac{S(F)}{\lambda}\Longrightarrow\lambda=\frac{S(F)}{\epsilon}=\frac{2^s\times\vartheta}{\acute{n}\times\epsilon} \qquad (۴)$$

بنابراین خواهیم داشت:

$$\lambda=\frac{2^s\times\vartheta}{\acute{n}\times\epsilon} \qquad (۵)$$

در خط ۱۶ الگوریتم ***PrivacyCL***، ماتریس تصادفی Δ به ابعاد $m\times d_g$ را طوری می‌سازیم که درایه‌های آن دارای توزیع لاپلاس با میانگین صفر و دامنه‌ی $\lambda=\frac{2^s\times\vartheta}{\acute{n}\times\epsilon}$ باشد. سپس، در خط ۱۷، ماتریس Δ را به ماتریس Y_g اضافه کرده و نتیجه‌ی به دست آمده را منتشر می‌کنیم. بنابراین، الگوریتم پیشنهادی ***PrivacyCL*** ، حریم خصوصی تفاضلی-ϵ دارد■.

۵-۲- ارزیابی دقت رده‌بندی

برای پیاده‌سازی الگوریتم ارائه شده، از زبان برنامه نویسی C#.Net و سیستم مدیریت پایگاه داده‌ی SQL SERVER استفاده شده است. پایه و اساس الگوریتم ارائه شده، متغیرهای تصادفی هستند. از این‌رو در هر بار اجرای الگوریتم ارائه شده‌ی مورد نظر، نتیجه‌ی متفاوتی را نسبت به نتیجه‌ی حاصل از اجرای قبلی به دست خواهیم آورد. بنابراین به منظور افزایش دقت نتایج، الگوریتم ارائه شده را ۱۰۰ بار بر روی یک مجموعه داده‌ی مورد نظر اجرا می‌کنیم و در هر بار، نتایج به دست آمده را در پایگاه داده‌ی SQL SERVER ذخیره می‌کنیم. در نهایت با استفاده از دستورات SQL، نتایج ذخیره شده را به دست آورده و آنها را به‌صورت جدول بیان می‌کنیم. الگوریتم پیشنهادی را بر روی هشت مجموعه داده‌ی واقعی و یک مجموعه داده‌ی تصنعی[۵۲] (SCCTS) اجرا نمودیم. این مجموعه داده‌ها از مرجع [۴۳] به دست آمده‌اند. جدول ۲، مشخصات این مجموعه داده‌ها را بیان می‌کند.

به منظور رده‌بندی داده‌ها، الگوریتم‌های مختلفی ارائه شده‌اند. یکی از مهمترین و شناخته شده‌ترین آنها، الگوریتم *K-NN* است. این الگوریتم، در اکثر پژوهش‌های ارائه شده در حوزه‌ی رده‌بندی، به عنوان استانداردی برای مقایسه بین الگوریتم ارائه شده و الگوریتم‌های موجود مورد استفاده قرار گرفته است. در این مقاله، از الگوریتم *K-NN* به منظور رده‌بندی داده‌ها استفاده شده است. الگوریتم ارائه شده، خدشه‌هایی را جهت ایجاد شرایط حریم خصوصی تفاضلی به داده‌های نهایی اضافه می‌کند. از این‌رو، بعد از انتخاب ده درصد داده‌ها به عنوان «داده‌های آزمون»، هر آزمایش را ۱۰۰ بار تکرار کرده (۱۰۰ بار الگوریتم *K-NN* را بر روی داده‌های نهایی اجرا می‌کنیم) و بیشینه «دقت رده‌بندی» حاصل از اجرای الگوریتم را محاسبه می‌کنیم. همچنین مقدار ϵ، یک مقدار عمومی[۵۳]

در نظر گرفته شده است ولی به نظر می‌رسد که مقادیر انتخاب شده برای ϵ، بهتر است بیشتر از عدد ۱ نباشد. در این مقاله، مقدار ϵ را عدد یک در نظر گرفته‌ایم. همچنین مقدار پارامتر a برابر با ۰.۸۵ در نظر گرفته شده است.

جدول ۲ : مشخصات مجموعه داده‌های آزمایش

عنوان مجموعه داده	تعداد مشخصه	تعداد رکورد	تعداد کلاس
breast cancer Wisconsin (B.C.W)	۳۱	۵۶۹	۲
chess	۳۶	۳۱۹۶	۲
glass	۱۰	۲۱۴	۷
habermans survival (H.S)	۳	۳۰۶	۲
Ionosphere	۳۵	۳۵۱	۲
iris	۴	۱۵۰	۳
SCCTS	۶۰	۶۰۰	۶
SPECTF	۴۵	۲۶۷	۲

جدول ۳، نتایج حاصل از اجرای الگوریتم ***PrivacyCL*** را بر روی ۲ سایت نشان می‌دهد. در این جدول، اعداد نوشته شده در ستون‌های Par.1 و Par.2 به ترتیب نشان دهنده‌ی تعداد مشخصه‌های پخش شده بر روی سایت ۱ و سایت ۲ است. Max-Ac بیشینه مقدار «دقت رده‌بندی» در ۱۰۰ بار آزمایش است. d_r d_o ، D.P.# و T_{Max} نیز به ترتیب، «تعداد مشخصه مجموعه داده‌ی اصلی»، «تعداد مشخصه‌ی مجموعه داده‌ی منتشر شده» و «تعداد مراحل تجزیه توسط موجک هار» و «بیشینه مقدار موجود در مجموعه داده» است. همان‌طوریکه ملاحظه می‌شود، الگوریتم‌های ارائه شده دقت رده‌بندی مناسبی را علاوه بر داشتن خاصیت حریم خصوصی تفاضلی، دارد. بر اساس جستجوهایی که انجام شد، مشخص شد که تا کنون، الگوریتمی در محیط‌های توزیعی برای تبدیل داده‌های اصلی به داده‌های مبتنی بر حریم خصوصی تفاضلی (با هدف رده‌بندی داده‌ها) همراه با کاهش ابعاد داده‌های منتشر شده ارائه نشده است. اکثر الگوریتم‌های ارائه شده در محیط‌های توزیعی برای PPC هیوریستیک‌مبنا بوده و در مقابل دانش پیش‌زمینه کاربران حریم خصوصی داده‌ها را نقض می‌کنند. با این حال، در مرجع [۱۲]، الگوریتم متمرکزی به نام ***PrivateProjection*** ارائه شده است که هم خاصیت حریم خصوصی تفاضلی دارد و هم ابعاد داده‌های اولیه را کاهش می‌دهد. پایه و اساس الگوریتم ***PrivateProjection*** تبدیلات جانسون-لیندن‌اشتراس [۲۰] است. الگوریتم ***PrivateProjection*** به این صورت اجرا می‌شود: مجموعه داده‌ی اولیه $T_{m\times n}$ که درایه‌های آن مقادیر بولی *true* و *false* (به جای مقادیر *true* و *false*، می‌توان از مقادیری که در بازه‌ی [0, 1] قرار دارند استفاده کرد) هستند را در نظر بگیرید. ابتدا، ماتریس افکنش[۵۴] تصادفی R به ابعاد $n \times d$ را ایجاد کرده سپس آن را در مجموعه داده‌ی T ضرب می‌کنیم تا ماتریس $Y_{m\times d}$ که ابعاد آن به مراتب کمتر از ابعاد ماتریس T است، ایجاد گردد. در قدم بعدی، ماتریس تصادفی $\Delta_{m\times d}$ که درایه‌های آن دارای توزیع نرمال هستند را ایجاد کرده و آن را با ماتریس $Y_{m\times d}$ جمع می‌کنیم تا مجموعه داده‌ی $P_{m\times d}$ به دست آید.

در نهایت، ماتریس $P_{m\times d}$ را منتشر می‌کنیم تا کاربران از آن استفاده نمایند.

در [۱۲]، ثابت شده است که الگوریتم ***PrivateProjection*** خاصیت«حریم خصوصی تفاضلی- (ϵ, δ)[55]» دارد. در این قسمت ثابت می‌کنیم که الگوریتم پیشنهای، هم حریم خصوصی بالاتر و هم دقت رده‌بندی بهتری نسبت به الگوریتم ***PrivateProjection*** دارد. لم زیر در مورد الگوریتم ***PrivateProjection*** در [۱۲] اثبات شده است:

لم ۱: [۱۲]. افکنش تصادفی $R_{n\times d}$ که درایه‌های آن دارای توزیع نرمال با میانگین صفر و واریانس $1/d$ هستند را در نظر بگیرید. اگر در الگوریتم ***PrivateProjection***، درایه‌های ماتریس تصادفی $\Delta_{m\times d}$ داری توزیع نرمال $N(0, \sigma^2)$ باشند در این صورت ***PrivateProjection*** خاصیت حریم خصوصی تفاضلی-(ϵ, δ) دارد هرگاه روابط زیر برقرار باشند:

$$\sigma \geq \frac{4}{\epsilon}\sqrt{\ln\left(1/\delta\right)},\ d > 2(\ln(n) + \ln\left(2/\delta\right),$$

$$\epsilon < \ln\left(1/\delta\right) \blacksquare.$$

برای مقایسه الگوریتم ارائه شده با الگوریتم ***PrivateProjection***، لم زیر را اثبات می‌کنیم.

لم ۲. فرض کنید ماتریس $P_g = Y_g + \Delta_g$ در اثر اعمال الگوریتم ***PrivacyCL*** بر روی مجموعه داده‌ی T_g به دست آمده باشد که در آن، درایه های ماتریس تصادفی Δ_g دارای توزیع لاپلاس با میانگین صفر و دامنه $\lambda = \frac{2^s\times\vartheta}{\acute{n}\times\epsilon}$ است. در این صورت، انحراف معیار خدشه‌های اضافه‌شده به ماتریس Y_g از رابطه‌ی زیر پیروی می‌کنند:

$$\sigma \leq \frac{\sqrt{2}}{\epsilon} \blacksquare.$$

جدول ۳: نتایج حاصل از اجرای *PrivacyCL* و *K-NN* بر روی دو سایت ($\epsilon = 1$)

مجموعه داده‌ها	تعداد رکوردها در هر سایت Par.1	Par. 2	d_r در هر سایت	$\acute{n}$	T_{Max}	D.P#	بیشینه دقت رده‌بندی در ۱۰۰ آزمایش Max-AC
Breast Cancer Wisconsin (d_o=31, s=0, $\lambda = 0.03$, ϑ =1)	15	16	1	32	911320502	5	0.91
Chess (d_o=36, s=5, $\lambda = 0.5$, ϑ =1)	18	18	18	64	73	1	0.77
Glass (d_o=10, s=1, $\lambda = 0.125$, ϑ =1)	5	5	2	16	214	3	1.0
Habermans survival (H.S) (d_o=3, s=0, $\lambda = 0.25$, ϑ =1)	1	2	1	4	83	2	1.0
Ionosphere (d_o=35, s=0, $\lambda = 0.031$, ϑ =2)	17	18	14	64	1	6	0.97
Iris (d_o=4, s=0, $\lambda = 0.25$, ϑ =1)	2	2	1	4	7.9	2	1.0
SCCTS (d_o=60, s=0, $\lambda = 0.031$, ϑ =2)	30	30	1	64	63.8281	6	0.86
SPECTF (d_o=45, s=0, $\lambda = 0.015$, ϑ =1)	22	23	1	64	89	6	1.0

اثبات: همان‌طوری که قبلاً گفته شد، به منظور اعمال تبدیل موجک هار، داده‌های اصلی در سطح اولیه قرار داده می‌شوند. اما، در سطح اولیه، هنوز تجزیه‌ای صورت نگرفته است از این‌رو، در این سطح، ضرایب تقریب را به جای ضرایب موجک مورد استفاده قرار خواهیم داد. بنابراین، جمع انرژی ضرایب موجک تمام رکوردها در سطح پایین‌تر به مراتب کمتر از جمع انرژی ضرایب تقریب در سطح اولیه (که همان داده‌های اصلی هستند) خواهد بود. پس حداقل یک بار عمل تجزیه‌ی داده‌ی اصلی با استفاده از تبدیل موجک هار، اجرا شده و در بدترین حالت، ابعاد داده‌ی تبدیل شده نصف ابعاد داده‌ی اصلی خواهد بود. به عبارت دیگر $S \leq log(\acute{n}) - 1$ است (S نشان دهنده‌ی سطح توقف تجزیه است). بنابراین نامعادله‌ی $\frac{2^s}{\acute{n}} \leq \frac{1}{2}$ را خواهیم داشت. همچنین بیشینه مقدار متغیر ϑ برابر با ۲ است. بنابراین $\lambda = \frac{2^s \times \vartheta}{\acute{n} \times \epsilon} \leq \frac{1}{\epsilon}$ خواهد شد. از طرفی، واریانس توزیع لاپلاس با میانگین صفر و دامنه‌ی λ برابر با $\sigma^2 = 2\lambda^2$ است. پس، رابطه‌ی $\sigma = \lambda\sqrt{2} \leq \frac{\sqrt{2}}{\epsilon}$ را به دست خواهیم آورد.■

اکنون، دو الگوریتم ***PrivacyCL*** و ***PrivateProjection*** را بر اساس مقدار خدشه‌ی اضافه شده به داده‌های منتشر شده مورد مقایسه قرار می‌دهیم. در الگوریتم ***PrivacyCL*** ، مقدار δ برابر با صفر است و برای مقدار ثابتی از ϵ، حداکثر مقدار σ برابر با $\frac{\sqrt{2}}{\epsilon}$ خواهد شد. در صورتی که در الگوریتم ***PrivateProjection***، مقدار σ وابسته به مقدار δ است. مطابق با لم ۱، نامعادله‌ی $\sigma \geq \frac{4}{\epsilon}\sqrt{\ln\left(1/\delta\right)}$ برای الگوریتم ***PrivateProjection*** برقرار است. بنابراین، اگر برای مقدار ثابتی از ϵ، مقدار δ به سمت صفر میل کند($\delta \to 0$)، در این صورت $\sigma \to \infty$ میل خواهد کرد. به عنوان مثال، فرض کنید که مقدار ϵ و δ به ترتیب برابر با ۱ و ۰.۱ در نظر گرفته شوند. در این‌صورت، مقدار σ در الگوریتم ***PrivacyCL*** حداکثر برابر با $\sqrt{2}$ خواهد شد در صورتی که درالگوریتم ***PrivateProjection*** این مقدارحداقل برابر با ۶.۰۶ خواهد شد. همچنین نامعادله‌ی چیبیشف[۵۶] [۴۴] بیان می‌کند که ۰.۷۵ درصد از متغیرهای تصادفی ایجاد شده از یک توزیع احتمالی، در فاصله‌ی دو برابر «انحراف معیار» از مقدار میانگین قرار می‌گیرند. بنابراین، الگوریتم ***PrivacyCL*** به مراتب خدشه‌ی کمتری را نسبت به الگوریتم ***PrivateProjection*** به داده‌های منتشر شده اضافه خواهد کرد. از طرفی، خدشه‌ی کمتر باعث تغییر کمتری در مقدار فاصله‌ی اقلیدسی خواهد شد. از این‌رو، دقت رده‌بندی بهتری را ایجاد خواهد کرد. پس، رده‌بندی الگوریتم ***PrivacyCL*** به مراتب بهتر از دقت رده‌بندی الگوریتم ***PrivateProjection*** خواهد شد.

حال دو الگوریتم ***PrivacyCL*** و ***PrivateProjection*** را بر اساس درجه حریم خصوصی مقایسه می‌کنیم. بر اساس شکل ۲، در الگوریتم ***PrivacyCL***، از متغیرهای لاپلاس با میانگین صفر و دامنه‌ی λ، به منظور ایجاد داده‌های تفاضلی استفاده می‌گردد ولی در الگوریتم ***PrivateProjection***، از متغیرهای تصادفی نرمال برای ایجاد داده‌های تفاضلی استفاده می‌شود. اما، متغیرهای لاپلاس، حریم خصوصی داده‌های منتشر شده را در بدترین فرضیات در نظر گرفته شده برای نقض حریم خصوصی، حفظ می‌کند در صورتی که متغیرهای تصادفی نرمال، حالت ملایم‌تری از فرضیات را در نظر می‌گیرند. ثابت شد که الگوریتم ***PrivacyCL*** خاصیت حریم

خصوصی تفاضلی- ϵ دارد. از طرف دیگر، مطابق با تعریف ۱، حریم خصوصی تفاضلی-$(\epsilon, 0)$ همان حریم خصوصی تفاضلی- ϵ است. به عبارت دیگر، مقدار پارامتر δ در الگوریتم ***PrivacyCL*** برابر با صفر است. اما مطابق با لم ۱، در صورتی که مقدار δ در الگوریتم ***PrivateProjection*** به سمت صفر میل کند $(\delta \to 0)$ در این صورت مقدار d نیز به سمت بی‌نهایت میل خواهد کرد (بر اساس شرط $d > 2(\ln(n) + \ln(2/\delta))$ در لم ۱، که این نادرست است چرا که مقدار d محدود می‌باشد. به عبارت دیگر، مقدار δ در الگوریتم ***PrivateProjection*** به هیچ عنوان صفر نخواهد شد. پس در یک مقدار ثابت در نظر گرفته شده برای پارامتر ϵ، درجه حریم خصوصی الگوریتم ***PrivacyCL*** به مراتب بهتر از درجه‌ی حریم خصوصی الگوریتم ***PrivateProjection*** است.

۶- نتیجه‌گیری

حفظ حریم خصوصی داده‌ها در داده‌کاوی یکی از مهمرین بحث‌ها است. روش‌ها و الگوریتم‌های مختلفی برای حفظ حریم خصوصی داده‌ها ابداع شده است. اکثر این الگوریتم‌ها دارای معایب (۱) حریم خصوصی قابل اثباتی ندارند (۲) فرض می‌کنند که افراد مهاجم، دانش پیش‌زمینه کمی نسبت به داده‌ها دارند و (۳) اغلب روی داده‌های با ابعاد بالا، کارایی کمی دارند. در این مقاله برای رفع این مشکل‌ها از مفهوم حریم خصوصی تفاضلی همراه با تبدیل موجک گسسته هار استفاده گردید. علاوه برا اثبات حریم خصوصی تفاضلی الگوریتم ارائه شده، نشان داده شد که الگوریتم ارائه شده ابعاد داده‌ها را کاهش داده و دقت خوبی در رده‌بندی داده‌ها دارد.

مراجع

[1] C.C. Aggarwal and P.S Yu, "Privacy-Preserving Data Mining: Models and Algorithms" Springer-Verlag, New York, 2008.

[2] B.C.M. Fung, K. Wang, R. Chen and P.S. Yu, "Privacy-Preserving Data Publishing: A Survey on Recent Developments," ACM Computing Surveys 42(4), Article 14, 2010.

[3] R. Agawam, R. Srikant, "Privacy-Preserving Data Mining," Proceedings of ACM SIGMOD Conference, pp. 439-450, 2000.

[4] L. Sweeney "k-Anonymity: A Model for Protecting Privacy," Int. J. Unc. Fuzz. Knowl. Based Syst., 10 (5), pp. 557-570, 2002.

[5] P. Samarati, L. Sweeney, "Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppression," IEEE Symp. on Security and Privacy, pp. 188-199, 1998.

[6] R.J. Bayardo, R. Agrawal, "Data Privacy through Optimal K-Anonymization," Proceedings of ICDE Conference, pp. 217-228, 2005.

[7] A. Machanavajjhala, J. Gehrke, and D. Kifer, "L-diversity: Privacy beyond k-anonymity," IEEE ICDE Conference, pp. 24, 2006.

[8] X. Gu, M. Li, L. Xiong and Y. Cao, "Providing Input-Discriminative Protection for Local Differential Privacy," 2020 IEEE 36th International Conference on Data Engineering (ICDE), pp. 505-516, 2020.

[9] C. Dwork, "Differential Privacy," Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP'06), pp. 1–12, 2006.

[10] B.C.M. Fung, K. Wang, A.W.C Fu and P.S. Yu, "Introduction to Privacy-Preserving Data Publishing Concepts and Techniques," Taylor & Francis Group, 2011.

[11] C. Dwork, F. McSherry, K. Nissim and A. Smith, "Calibrating noise to sensitivity in private data analysis," in TCC, pp. 265–284 2006.

[12] K. Kenthapadi, A. Korolova, I. Mironov and N. Mishra,"Privacy via the Johnson-Lindenstrauss Transform," CoRR, abs/1204.2606, 2012.

[13] X. Xiaokui, W. Guozhang, G.Johannes, "Differential Privacy via Wavelet Transforms," IEEE Transactions on Knowledge and Data Engineering 23(8), pp. 1200-1214, 2011.

[14] A.G. Divanis, V.S. Verykios," Association Rule Hiding for Data Mining," Springer, 2010.

[15] Y. Cui, W.K. Wong, and D.W. Cheung, "Privacy-Preserving Clustering with High Accuracy and Low Time Complexity," In Proceedings of the 14th International Conference on Database Systems for Advanced Applications, DASFAA 2009, Brisbane, Australia, 2009.

[16] S.R.M. Oliveira, and O.R. Za¨iane, " Privacy preserving clustering by data transformation," Proc. of the 18th Brazilian Symposium on Databases, Manaus, Amazonas, Brazil, pp. 304–318, 2003.

[17] W. Du and Z. Zhan, "Using randomized response techniques for privacy-preserving data mining," In Proc. of the 9th International Conf. on Knowledge Discovery and Data Mining (KDD'03), pp. 505–510, 2003.

[18] Z. Huang, W. Du, and B. Chen, "Deriving Private Information from Randomized Data," In Proceedings of the ACM SIGMOD Conference on Management of Data, Baltimore, Maryland, USA, pp. 37-48, 2005.

[19] J. Vaidya and C. Clifton, "Privacy-Preserving K-means Clustering over Vertically Partitioned Data," In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 206 – 215, 2003.

[20] W.B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz Maps into a Hilbert Space," Contemporary Mathematics 26, 189–206, 1984.

[21] E. Bingham and H. Mannila, "Random Projection In Dimensionality Reduction: Applications To Image And Text Data," In Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 245–250, 2001.

[22] X.Z. Fern, and C.E. Brodley, "Random projection for high dimensional data clustering: a cluster ensemble approach," In Proc. of the 20th international conf. on machine learning, Washington DC, USA, pp. 102-110, 2003.

[23] S.R.M. Oliveira and O.R. Za¨ıane, "A Privacy-Preserving Clustering Approach Toward Secure and Effective Data Analysis for Business Collaboration," Computers & Security 26(1), 81-93, 2007.

[24] A. Blum, C. Dwork, F. McSherry, and K. Nissim, "Practical Privacy: The SuLQ Framework," In: Proceedings of the 2005 SIGMOD/PODS Conference, pp. 128–138, 2005.

[25] C. Dwork, and N. Nissim, "Privacy-Preserving Data mining on Vertically Partitioned Databases," In: Proceedings of the 24th Annual International Cryptology Conference Advances in Cryptology (CRYPTO'04), Santa Barbara, California, USA, pp. 528-544, 2004.

[26] K. Nissim, S. Raskhodnikova, and A. Smith, "Smooth Sensitivity and Sampling in Private Data Analysis," In: Proceedings of the 39th ACM Symposium on the Theory of Computing (STOC'07), pp. 75-84, 2007.

[37] Z. Li, M. Sun, "Privacy-Preserving Classification of Personal Data with Fully Homomorphic Encryption: An Application to High-Quality Ionospheric Data Prediction," International Conference on Machine Learning for Cyber Security, pp. 437-446, 2020.

[38] P. Li, J. Li, Z. Huang, CZ. Gao, W.B. Chen, K. Chen, "Privacy-preserving outsourced classification in cloud computing," Cluster Computing, Vol. 21, pp. 277–286, 2018.

[39] M.R.E. Dishabi, M.A. Azgomi, "Differential privacy preserving clustering in distributed datasets using Haar wavelet transform," Intelligent Data Analysis, vol. 19, no. 6, pp. 1323-1353, 2015.

[40] R. Canetti, U. Feige, O. Goldreich, and M. Naor, "Adaptively Secure Multi-Party Computation," Proceedings of the 28th Annual Symposium on Theory of Computing (STOC'96), ACM Press, pp. 639–648, 1996.

[41] C.S. Burrus, R.A. Gopinath, and H. Guo, "Introduction to Wavelets and Wavelet Transforms," A Primer Prentice Hall, 1997.

[42] H. Zhang, T.B. Ho, Y. Zhang and M.S. Lin, "Unsupervised Feature Extraction for Time Series Clustering Using Orthogonal Wavelet Transform," Journal Informatica 30(3), pp. 305-319, 2006.

[43] "Machine Learning Repository," URL: http://archive.ics.uci.edu/ml, Visited: 2020-05-20.

[44] D.F. Vysochanskij, and Y.I. Petunin, "Justification of the Three Sigma Rule for Unimodal Distributions," Theory of Probability and Mathematical Statistics 21, 25–36, 1980.

[27] F. McSherry, "Privacy Integrated Queries: An Extensible Platform for Privacy-Preserving Data Analysis," In: Proceedings of the 2009 SIGMOD/PODS Conference, pp. 19-30, 2009.

[28] D. Feldman, A. Fiat, H. Kaplan, and K. Nissim, "Private Coresets." In: Proceedings of the 41st Symposium on Theory of Computing (STOC), pp. 361–370, 2009.

[29] S. Xu, X. Cheng, S. Su, et al., Differentially private frequent sequence mining, IEEE Trans. Knowl. Data Eng. 28 (11), 2016.

[30] G. Zhou, S. Qin, H. Zhou, et al., A differential privacy noise dynamic allocation algorithm for big multimedia data, Multimedia Tools Appl. (8), pp. 1-19, 2018.

[31] H. Shen, Z. Lu, A new lower bound of privacy budget for distributed differential privacy, in: proceedings of International Conference on IEEE Parallel and Distributed Computing, Applications and Technologies. pp. 25–32, 2018.

[32] Q. Wang, Y. Zhang, X. X. Lu, Z. Wang, Real-time and spatiotemporal crowd-sourced social network data publishing with differential privacy, IEEE Trans. Dependable Secure Comput. 15 (4), 2018.

[33] L. Ou, Z. Qin, S. Liao, T. Li, and D. Zhang, "Singular spectrum analysis for local differential privacy of classifications in the smart grid," IEEE Internet of Things Journal, 2020.

[34] Sun, X., Xu, R., Wu, L. et al. A differentially private distributed data mining scheme with high efficiency for edge computing. J Cloud Comp 10, 7, 2021.

[35] S. Mukherjee, et al, "A privacy preserving technique for distance-based classification with worst case privacy guarantees," Data & Knowledge Engineering, 66, pp. 264–288, 2008.

[36] W. Fan, J. He, M. Guo, P. Li, Z. Han, R. Wang, "Privacy preserving classification on local differential privacy in data centers," Journal of Parallel and Distributed Computing, Vol. 135, pp. 70-82, 2020.

[1] Privacy Preserving Data Mining (PPDM)
[2] Privacy Preserving Data Publishing
[3] Randomization
[4] K-Anonymity
[5] L-Diversity
[6] Transformation Methods
[7] Privacy Preserving Classification
[8] adversary
[9] background knowledge
[10] heuristic
[11] differential privacy
[12] wavelet analysis
[13] wavelet transform
[14] multi-resolution analysis
[15] Mallat & Meyer
[16] smooth background
[17] fluctuation
[18] data set
[19] attribute
[20] L_1-sensitivity
[21] protocol-based approach
[22] transformation-based approach
[23] secure multi-party computation
[24] Oliveira
[25] Zaiance
[26] geometric transformation
[27] shift
[28] scaling
[29] rotation
[30] known input-output
[31] high correlation
[32] Eigen-analysis attack
[33] Yao
[34] multiplicative perturbation
[35] Johnson-Lindenstrauss
[36] random projection
[37] independent and identically distributed
[38] differential privacy
[39] sub-linear query
[40] generic sampling-based procedure
[41] privacy integrated queries
[42] frequent sequence mining
[43] multi-media
[44] spatial data space
[45] smart grid
[46] Edge computing
[47] malicious model
[48] semi-honest
[49] party
[50] third party
[51] Haar wavelet transform
[52] synthetic
[53] public
[54] projection matrix
[55] (ϵ,δ)-differential privacy
[56] Chebychev's inequality