مجموعه مقالات پانزدهمین کنفرانس بین‌المللی انجمن رمز ایران
دانشکده مهندسی برق، دانشگاه تربیت دبیر شهید رجایی، تهران، ۶-۷ شهریور ۱۳۹۷

روشی ترکیبی برای حفظ حریم خصوصی در انتشار کلان داده‌ها

مهدی خشنود منصورخانی[۱]، محمد غفاریان[۲]، حمیدرضا شهریاری[۳]

[۱] دانشگاه صنعتی امیرکبیر، دانشکده مهندسی کامپیوتر و فناوری اطلاعات
m_khoshnood@aut.ac.ir
[۲] دانشگاه صنعتی امیرکبیر، دانشکده مهندسی کامپیوتر و فناوری اطلاعات
s.m.ghaffarian@aut.ac.ir
[۳] دانشگاه صنعتی امیرکبیر، دانشکده مهندسی کامپیوتر و فناوری اطلاعات
shahriari@aut.ac.ir

چکیده

امروزه، بسیاری از سازمان‌ها و شرکت‌ها اطلاعات خود را برای تجزیه‌وتحلیل در اختیار داده‌کاوان و عموم مردم قرار می‌دهند. این امر می‌تواند تهدیدی برای حریم‌خصوصی افراد باشد. ازاین‌رو حفظ حریم‌خصوصی افراد در اشتراک‌گذاری اطلاعات سازمانی مساله‌ی مهمی است که در سال‌های اخیر روش‌های بسیاری برای آن توسعه داده‌شده است. غالباً داده‌های جمع‌آوری‌شده حاوی اطلاعات حساسی هستند که نباید منتشر شوند. انباشته شدن داده‌هایی با ویژگی‌هایی مانند تنوع، حجم و سرعت، باعث بروز پدیده‌ی «کلان داده‌ها» شده‌اند. پردازش کلان داده‌ها چالش‌های جدیدی در روش‌های حفظ حریم خصوصی به وجود آورده است. از مهم‌ترین این چالش‌ها عدم کارایی روش‌های گذشته در مواجهه با حجم بالای داده‌ها در عین برقراری تعادل بین دقت داده‌ها و حفظ حریم خصوصی افراد است. در این مقاله روش پیشنهادی مبتنی بر توابع حساس به موضع و الگوریتم ژنتیک، برای تأمین شرایط k-گمنامی و l-تنوع در کلان داده ارائه می‌گردد. ارزیابی روش نشان می‌دهد این روش نسبت به روش قبلی دارای مزیت مقیاس‌پذیری، افزایش کیفیت داده‌های منتشرشده و همچنین افزایش سطح حریم‌خصوصی کاربران هنگام انتشار داده‌ها است.

کلمات کلیدی: حریم خصوصی، کلان داده‌ها، k-گمنامی، l-تنوع

۱- مقدمه

کلان داده‌ها[۱] مجموعه بزرگی از داده‌ها هستند که معمولا از چند منبع اطلاعاتی متفاوت گرفته شده‌اند. این نوع از داده‌ها از پیچیدگی زیادی برخوردار هستند که این موجب دشواری پردازش کلان داده‌ها می‌گردد به‌طوری‌که پردازش آن‌ها با استفاده از ابزارهای سنتی مدیریت پایگاه داده به سادگی امکان‌پذیر نیست. امروزه با گسترش حجم داده‌ها در کاربردهای مختلف، بخش‌های مختلف سازمانی و تجاری در جهت تولید دانش و اطلاعات مفید از کلان داده‌ها بهره-های فراوانی را می‌برند. فرایند تولید دانش و اطلاعات مفید از کلان داده‌ها معمولا با گردآوری، ذخیره، پردازش،تجزیه و تحلیل و اشتراک‌گذاری آن‌ها همراه است. لذا از جمله مسائل مهم در طی این فرایند، حفظ حریم خصوصی افراد است. حفظ حریم خصوصی با تاکید خاص بر ضمانت صحت و محفوظ ماندن داده‌های حساس هست. درواقع هدف اصلی حفظ حریم خصوصی حفاظت از اطلاعات حساس در طول فرآیند تولید دانش، پردازش، اشتراک‌گذاری و یا انتشار داده‌ها است.

مساله‌ی حفظ حریم خصوصی در انتشار داده یکی از مهم‌ترین مسائل می‌باشد که باید به آن توجه نمود. در انتشار داده کاربران نباید با اطلاعات حساسشان تعیین هویت شوند و از طرفی داده منتشرشده باید سودمند باشد.

پردازش کلان داده‌ها چالش‌های جدیدی در روش‌های حفظ حریم خصوصی به وجود آورده است. از مهم‌ترین این چالش‌ها عدم کارایی روش‌های گذشته در مواجهه با حجم داده‌ها، سطح پایین گمنامی برای انتشار و برقراری تعادل بین دقت داده‌ها و حفظ‌حریم‌خصوصی افراد است.

در روش‌های حفظ حریم خصوصی داده‌های منتشرشده، فرمت جداول به شکل زیر است:

(شناسه صریح، شبه شناسه[۲]، ویژگی‌های حساس، ویژگی‌های غیر حساس)

شناسه صریح شامل یک مجموعه از ویژگی‌ها است مثل نام که حاوی اطلاعاتی است که به‌صراحت مشخص‌کننده صاحب رکورد اطلاعاتی است.

شبه شناسه شامل یک مجموعه از اطلاعات است که صاحب رکوردهای اطلاعاتی را مشخص می‌کند. ویژگی‌های حساس شامل اطلاعات حساس فرد همانند بیماری، دستمزد و غیره است. هر ویژگی که جز سه دسته ویژگی فوق نباشند را ویژگی‌های غیر حساس می‌نامیم [۱].

---

[۱] Big Data

[۲] Quasi Identifier

گمنامی یکی از روش‌های حفظ حریم خصوصی داده‌های منتشرشده است که از شناسایی اطلاعات حساس جلوگیری می‌کند. در این فرآیند داده‌ها به صورتی تغییر می‌کنند که هنگام انتشار یا استفاده از آن‌ها اطلاعات کلیدی قابل‌شناسایی نباشند. دو روش اصلی برای گمنام کردن داده‌ها وجود دارد که در محافظت از حریم خصوصی کلان داده‌ها استفاده می‌شوند. این روش‌ها k–گمنامی[۳] l–تنوع[۴] هستند.

k–گمنامی: این روش راهکارهایی را برای پنهان نمودن هویت کاربران بیان می‌کند، از مهم‌ترین این راهکارها، تغییر مقادیر اصلی داده‌ها، عوض کردن مقادیر داده با هم یا تعمیم مقادیر داده، است. رویکرد محافظتی k–گمنامی زمانی استفاده می‌شود که محتویات اطلاعات هر فرد حداقل با k-1 فرد دیگر یکسان باشد. k–گمنامی با تعمیم صفت‌های کاربران محقق می‌شود[۲].

l–تنوع: هدف تنوع L برقراری تنوع برای ویژگی حساس در هر گروه از شبه شناسه‌ها است. روش l–تنوع درواقع به دنبال این است که در هر گروه از شبه شناسه حداقل L مقدار متفاوت برای ویژگی‌های حساس وجود داشته باشد[۳].

ساختار کلی این مقاله به این صورت ا ست که در بخش ۲ در مورد کارهای صورت گرفته در این حوزه صحبت می‌کنیم. در بخش ۳ مفاهیم پایه را معرفی می‌کنیم. در بخش ۴ روش ارائه‌شـــده را بررســـی می‌کنیم. در بخش ۵ نتایج ارزیابی را نشـــان می‌دهیم و در بخش ۶ نیز جمع‌بندی و کارهای آتی را ارائه می‌دهیم.

۲- کارهای گذشته

در [۴] روشی برای برقراری l–تنوع و k–گمنامی ارائه‌شده است. در این مدل ارائه‌شده از روش خوشه‌بندی k-member برای خوشه‌بندی داده‌ها استفاده‌شده است. درواقع ایده اصلی این بوده که با استفاده از الگوریتم خوشه‌بندی K-member مسئله گمنام‌سازی را به مسئله خوشه‌بندی تبدیل کند و مجموعه‌ای از کلاس‌های هم‌ارز را پیدا کند که در آن داده‌ها به توجه به خوشه‌های نهایی تعمیم داده شوند.

در [۵] روشی برای حفظ حریم خصوصی در جریان داده‌ها ارائه‌شده است. در این روش بافری که شامل مجموعه‌ای از خوشه‌ها است برای عمومی‌سازی داده‌ها وجود دارد. زمانی که یک داده جدید وارد می‌شود میزان شباهتش با خوشه‌های موجود سنجیده می‌شود. اگر با یکی از خوشه‌های موجود شباهتی داشت به آن خوشه تعلق می‌گیرد در غیر این صورت این داده باید یک مدت‌زمان مشخصی را انتظار بکشد. اگر پس از زمان تعیین‌شده داده‌های شبیه به آن وارد نشد، در آن صورت دسته‌ای جدید با K نزدیک‌ترین همسایه برای داده موردنظر ساخته می‌شود.

در [۶] یک تکنیک جدید به اسم برش[۵] ارائه‌شده که داده‌ها را به‌صورت عمودی و افقی افراز می‌کند که از افشای عضویت محافظت می‌کند. در این روش ابتدا یک افراز از صفات بر اساس محاسبه همبستگی بین صفات و خوشه‌بندی صفات ایجاد می‌کنند. بعد از ایجاد یک افراز از صفات، افراز بر اساس رکوردها ایجاد می‌کنیم. درواقع به این صورت که یک برش به‌صورت عمودی و افقی بر روی دادها ایجاد می‌کند. بعد از ایجاد این تقسیم‌بندی بر روی‌داده‌ها نظم داده‌های موجود در هر کلاس هم ارزی را با تغییر مکان‌های داده‌ها از بین می‌برد.

در [۷] روشی برای برقراری l–تنوع در انتشار داده‌ها ارائه‌شده است. در این رویکرد در ابتدا برای ویژگی‌های حساس یک درخت طبقه‌بندی ایجاد می‌کند و سپس بر اساس این درخت طبقه‌بندی و ویژگی‌های حساس داده‌ها را تقسیم-بندی می‌کند. در مرحله بعد برای رسیدن به‌شرط حداقل میزان تنوع l در هر مرحله یک داده از این دسته‌ها انتخاب می‌کند و با توجه به کمترین فاصله که این داده با داده‌های دسته‌های دیگر دارد در دسته نهایی قرار می‌دهد.

در [۸] روش به نام LSH-RC برای حفظ حریم خصوصی کاربران بر مبنای نگاشت کاهش ارائه دادند که با کمک توابع حساس به موضع و خوشه‌بندی k-member است. این روش به این صورت عمل می‌کند که به‌جای اینکه همه رکوردها را برای محاسبه میزان شباهت به هم مقایسه کند می‌آید آن‌هایی را که کاندید به شباهت هستند را باهم مقایسه می‌کند یعنی مقدار درهم‌سازی داده‌ها را حساب می‌کند و در دسته‌هایی آن‌ها را قرار می‌دهد. بعدازاینکه داده‌ها را به دسته‌هایی تقسیم کرد در مرحله بعد تعداد عناصر هر دسته را برای بررسی برقرار بودن شرط k گمنامی بررسی می‌کند. که برای انجام این کار از الگوریتم خوشه‌بندی k-member استفاده می‌کند.

از مهم‌ترین مشکلات بیان شده برای روش‌های پیشین عدم کارایی این رویکردها زمانی که حجم داده‌ها افزایش پیدا می‌کند، کیفیت پایین داده‌های انتشاریافته و همچنین سطح پایین حریم خصوصی افراد است. بنابراین در این مقاله روشی ارائه می‌شود که مشکلات بیان‌شده را برطرف و یا بهبود دهد.

۳- مفاهیم پایه

در این بخش به معرفی و تعریف مفاهیم پایه‌ای می‌پردازیم که با کمک آن‌ها روش پیشنهادی را ارائه دادیم.

۳-۱- توابع حساس به موضع

این نوع توابع باعث کاهش ابعاد داده با ابعاد بالا می‌شود. از این نوع توابع برای پیدا کردن شباهت بین رکوردهای داده استفاده می‌شود. این توابع داده‌های ورودی را به‌گونه‌ای به یکسری دسته‌ها تقسیم می‌کند که رکوردهایی که کاندید شباهت هستند با احتمال بالا در یک دسته قرار می‌گیرند[۹].

یک تابع درهم‌سازی بر اساس تعریف ریاضی به‌صورت زیر است:

تعریف ۱: فرض کنید دو فاصله d_1 و d_2 که بر اساس یک معیار فاصله‌ای به‌دست‌آمده است رابطه $d_1 < d_2$ بین آن‌ها برقرار باشد. تابع درهم‌سازی که برای آن‌ها تعریف می‌شود به‌صورت (d_1,d_2,p_1,p_2)–حساس هست که به‌صورت زیر تعریف می‌شوند[۱۰]:

- اگر $d(x,y)<d_1$ باشـــد با حداقل احتمال p_1، رکوردهای x و y کاندیدای شباهت هستند.
- اگر $d(x,y)>d_2$ باشـــد با حداکثر احتمال p_2، رکوردهای x و y کاندیدای شباهت هستند.

اگر h تابع درهم‌ساز باشد زمانی که $h(x)=h(y)$ به این معنی است که x و y توسط تابع درهم‌ساز در یک دسته یکسان قرارگرفته‌اند. درواقع برای استفاده از توابع حساس به موضع ما از تکنیک banding استفاده می‌کنیم. با توجه به خانواده توابع در هم‌ساز دو متغیر α و λ در نظر گرفته می‌شود. در این تکنیک

---

[۳] K-Anonymity

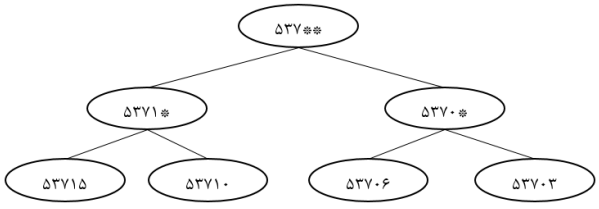
[۴] L-Diversity

[۵] slicing

مقدارهای توابع درهم‌ساز را به λ باند که هرکدام شامل α سطر است تقسیم می‌کند. α وλ دو متغیر صحیح است که توسط کاربر انتخاب می‌شود. با توجه به تعریف تکنیک banding دو رکورد در یک دسته یکسان قرار می‌گیرند اگر مقدار درهم‌سازی همه α در حداقل یک باند یکسان باشد. اگر S را میزان شباهت بین دو رکورد را مشخص کند بنابراین احتمال اینکه این دو رکورد در یک دسته قرار گیرد برابر $1-(1-s^{\alpha})^{\lambda}$ است.

زمانی که بخواهیم از توابع حساس به موضع استفاده کنیم باید معیار شباهتی برای این توابع معرفی کنیم که بر اساس آن معیار، توابع درهم‌ساز را تولید کنیم و میزان شباهت بین داده‌ها برای تولید کاندیدهای شباهت را بسنجیم. حال با توجه به این‌که داده‌های دارای ویژگی‌های مستقل از هم هست و اینکه مقدار هر ویژگی یا ابعاد در جایگاه خودش معنا دارد و همچنین هر رکورد از داده‌ها تبدیل به یک بردار خلوت[۶]می‌شود بنابراین با یک نرمال‌سازی داده‌ها ما از توابع درهم‌ساز تصادفی مبتنی بر شباهت کسینوسی استفاده می‌کنیم که تخمین خوبی از شباهت‌های بین داده‌ها ارائه می‌دهد. برای استفاده از این معیار ابتدا شبه شناسه هرکدام از رکوردها را با توجه به درخت طبقه‌بندی[۷]به برداری از ویژگی‌ها تبدیل می‌کنیم.

بردار ویژگی بر اساس شبه شناسه و درخت طبقه‌بندی ساخته می‌شود و آن را با r* نشان می‌دهیم. در اینجا با توجه به مقدار گره شبه شناسه و گره‌های اجداد به‌جز گره ریشه، در بردار ویژگی مقدار یک و برای بقیه مقادیر صفر در نظر گرفته می‌شود.برای مثال درخت طبقه‌بندی برای ویژگی مثل کد پستی در شکل ۱ آمده است.



شکل (۱) : درخت طبقه‌بندی کد پستی

فرض کنید برای در نظر گرفتن جنسیت نیز از یک بیت که صفر نشان‌دهنده مرد بودن و یک نشان‌دهنده زن بودن هست استفاده می‌کنیم. همچنین برای تبدیل مقدار کد پستی ۵۳۷۱۵ به برداری از صفر و یک از شکل ۲ که بر اساس درخت طبقه‌بندی به‌دست‌آمده است استفاده می‌کنیم.

۵۳۷۱۵							
۵۳۷۱۵						۵۳۷۱*	
۵۳۷۱۵	۵۳۷۱۰	۵۳۷۱۱	۵۳۷۰۶	۵۳۷۰۷	۵۳۷۰۳	۵۳۷۱*	۵۳۷۰*
۱	۰	۰	۰	۰	۰	۱	۰

شکل (۲) : بردار ویژگی کد پستی

بنابراین بردار نرمال سازی شده برای ویژگی جنسیت و کد پستی برای فردی که جنسیت مرد دارد و کد پستی آن ۵۳۷۱۵ باشد به‌صورت r*= < ۰،۱،۰،۰،۰،۰،۰،۱،۰> است که در اینجا اولین بیت را برای جنسیت در نظر گرفتیم. البته ذکر این نکته ضروری هست که ویژگی‌های کمی مثل سن را به‌گونه‌ای نرمال‌سازی می‌کنیم که مقداری بین صفر و یک بگیرند و بعد آنها را در بردار ویژگی قرار می‌دهیم.

۲-۳-فاصله

خوشه‌بندی[۸]در آمار و یادگیری‌ماشینی، یکی از شاخه‌های یادگیری بی‌نظارت[۹] می‌باشد و فرآیندی است که در طی آن، نمونه‌ها در دسته‌هایی که اعضای آن مشابه یکدیگر می‌باشند تقسیم می‌شوند که به این دسته‌ها خوشه گفته می‌شود. برای خوشه‌بندی باید یک معیار برای ارزیابی میزان شباهت داده‌ها استفاده کنیم که این معیار درواقع همان فاصله‌ای است که رکوردهای داده از همدیگر دارند. این توابع فاصله توسط انواع صفات تعیین می‌شوند. انواع صفات را می‌توان به صفات عددی (کمی) و صفات کیفی تقسیم کرد و در اغلب موارد مجموعه داده‌ها شامل صفات عددی و نیز کیفی می‌باشند. توابع فاصله‌ای برای این دو نوع صفت متفاوت هستند بنابراین فاصله بین مقادیر عددی را می‌توان به‌صورت رابطه ۱ تعریف کرد:

$$D(x_i,y_i)=\frac{|x-y|}{R} \qquad (۱)$$

R در اینجا دامنه تغییرات برای ویژگی i است.

برای صفات کیفی باید از درخت‌های طبقه‌بندی کمک بگیریم. بنابراین برای هرکدام از صفات کیفی ما یک درخت طبقه‌بندی ایجاد می‌کنیم. فاصله بین دو صفت به طول مسیر آن‌ها در درخت طبقه‌بندی‌شان برمی‌گردد. بنابراین فاصله در صفات کیفی به‌صورت رابطه ۲ تعریف می‌شود:

$$D(xi,yi)=\frac{L(X,Y)}{H} \qquad (۲)$$

حال برای محاسبه فاصله کلی بین رکوردها از ترکیب دو معیار اندازه‌گیری فاصله کمی و کیفی استفاده می‌کنیم. فرض کنید که A_1 تا A_n ابعاد داده‌ها یا همان صفات داده‌ها هست که به‌عنوان شبه شناسه (QI) در نظر گرفتیم. مجموعه QI را به دو مجموعه QI_L (صفات کیفی) و QI_L(صفات کمی) افراز می‌کنیم که برای آن‌ها روابط ۳ و ۴ برقرار است. فرض کنید i و j نیز دو رکورد از داده‌ها هست که می‌خواهیم برای آن‌ها فاصله را محاسبه کنیم.

$$QI = QI_N \cup QI_L \qquad (۳)$$

$$QI_L \cap QI_N = \emptyset \qquad (۴)$$

$$F(Y)=d(i,j)=\sum_{A_p \in QI_N} W_p \frac{|A_{i,p}-A_{j,p}|}{R_p} + \sum_{p \in QI_L} W_p \frac{L(A_{i,p},A_{j,p})}{H_p} \qquad (۵)$$

تابع F(Y) برای نشان دادن فاصله بین دو رکورد است. همان‌طور که پیش‌تر گفته شد داده‌های ما دارای صفات کیفی و کمی باهم هست بنابراین محاسبه فاصله بین دو رکورد در رابطه ۵ آمده است که این رابطه بر اساس روابط ۱ و ۲ که برای محاسبه فاصله بین رکوردهای کمی و کیفی است محاسبه می‌شود.

۴- روش پیشنهادی

روشی که ارائه می‌شود یک روش دومرحله‌ای است. در مرحله اول مجموعه داده ا صلی تو سط توابع ح ساس به مو ضع به د سته‌هایی کوچک‌تر

---

[۶] Sparse
[۷] Taxonomy
[۸] Clustering
[۹] Unsupervised

تقسیم می‌شوند و در مرحله بعدی به کمک الگوریتم ژنتیک با توجه به قیودی که بر روی آن تعریف شده است سعی می‌کنیم به گمنام‌سازی داده‌های دست پیدا کنیم.

در مرحله اول ابتدا داده‌ها را به یک بردار با توجه با آنچه قبلاً گفته‌شده تبدیل می‌کنیم. سپس توابع درهم‌ساز را برای تقسیم‌بندی و تولید کاندیدهای شباهت بر روی‌داده‌ها اجرا می‌کنیم. در مرحله آخر هم بر روی هرکدام از دسته‌هایی که از مرحله قبل به‌دست‌آمده است الگوریتم ژنتیک را اجرا می‌کنیم تا به خوشه‌هایی با شرایط مدنظر دست پیدا کنیم.

```
Input = Dataset D, privacy parameter k
and L
Output = Anonymous Dataset D*
01: Convert data record r to vector r*
02: Run Hash based LSH to obtain a set
of buckets
03: For each bucket run the genetic
algorithm with constraint to obtain K-
anonymity With L-diversith clusters D*
04: return D*
```

شکل (۳) : شمای کلی الگوریتم گمنامی ارائه شده

۴-۱- گام اول

در مرحله اول داده‌ها را با کمک توابع حساس به موضع دسته‌بندی می‌کنیم. توابع حساس به موضع در افراز داده‌ها برای به دست آوردن مقیاس‌پذیری و قابلیت اطمینان استفاده می‌شود و با احتمال بالا داده‌های مشابه در یک دسته قرار می‌گیرند.

افراز بر پایه توابع حساس به موضع با نگاشت-کاهش

فرایند افراز با استفاده از توابع حساس به موضع شامل دو مرحله است، یکی دسته‌بندی داده‌ها بر اساس مقدار درهم‌سازی تولیدی برای داده‌ها و دیگری ادغام کردن افرازها است. اصل این فرایند افراز تولید شماره دسته برای هر رکورد داده است. درواقع هر تابع نگاشت در اینجا یک رکورد از شبه شناسه‌ها از ورودی دریافت می‌کند و یک مقدار درهم‌سازی که همان شماره دسته می‌باشد را تولید می‌کند؛ یعنی در این مرحله ما به کمک تابع نگاشت یک کلید[۱۰] و ارزش[۱۱] برای داده‌ها تولید می‌کنیم که کلید برای داده‌ها همان شماره دسته هست که از روی درهم‌سازی داده‌ها تولیدشده و ارزش نیز در اینجا یک رکورد از شبه شناسه‌ها است.

```
Input = Data record r, hash function H
Output = Key-value pair(bucketID,r)
  Initialize bucketID with an empty string
  For i,1≤i≤a do  //a=number of hash function
            h ← H(i) // Compute hash value
       bucketID ← concat(bucketId,h)
  return (bucketID,r).
```

شکل (۴) : شبه کد تابع نگاشت برای افراز داده‌ها

۴-۲- گام دوم

در این مرحله هدف به دست آوردن مجموعه خوشه‌های نهایی است که برای آن‌ها شروط k–گمنامی و l–تنوع برقرار باشد. برای رسیدن به این هدف از الگوریتم ژنتیک استفاده می‌شود تا شرایط k–گمنامی و l–تنوع را برای خوشه-های نهایی برقرار کند.

گمنام‌سازی داده‌ها

پس از دسته‌بندی داده در مرحله قبل توسط توابع حساس به موضع، اکنون در هر دسته، به‌گونه‌ای داده‌ها را خوشه‌بندی می‌کنیم که هر گروه هم‌زمان هر دو شرط k–گمنامی و l–تنوع را داشته باشد. بنابراین در هر خوشه حداقل تعداد رکورد با شبه شناسه یکسان باید k و حداقل تنوع صفات حساس رکوردها باید l باشد، درعین‌حال کیفیت داده‌ها حداکثر ممکن را داشته باشد. یافتن خوشه‌های بهینه طبق [۱] یک مسأله np-hard است. یکی از ابزارهای حل مسائل بهینه‌سازی با چند قید، الگوریتم‌های تکاملی است. در اینجا برای حل این مسأله از الگوریتم ژنتیک استفاده می‌کنیم. در مدل‌سازی الگوریتم ژنتیک از تکامل ژنتیکی به‌عنوان یک الگوی حل مسئله استفاده می‌شود. مسئله‌ای که باید حل شود دارای ورودی‌هایی می‌باشد که طی یک فرایند الگوبرداری شده از تکامل ژنتیکی به راه‌حل‌ها تبدیل می‌شود سپس راه‌حل‌ها به‌عنوان کاندیداها توسط تابع ارزیاب[۱۲] مورد ارزیابی قرار می‌گیرند و چنانچه شرط خروج مسئله فراهم‌شده باشد الگوریتم خاتمه می‌یابد[۱۱].

پارامتر k: پارامتر k–گمنامی هست که توسط کاربر تعریف می‌شود و اندازه خوشه‌ها با توجه به این متغیر تعیین می‌شود.

در این مرحله هر ژن نشان‌دهنده یک مرکز خوشه است و طول کروموزوم‌ها را برابر با بیشترین مقدار ممکن یعنی تعداد دادهای دسته تقسیم‌بر K که پارامتری هست که توسط کاربر تعریف می‌شود در نظر می‌گیریم. اگر در پایان تعداد تکرارهای الگوریتم ژنتیک، خوشه‌ای به حداقل تعداد اعضا نرسد با نزدیک‌ترین خوشه که شرایط آن را نقض نکند ادغام می‌شود.

اگر c_j مرکز خوشه jام باشد که $1 \leq j \leq p$، آنگاه تابع هزینه به‌صورت رابطه ۶ تعریف می‌شود.

$$g(c_1,c_2,\ldots,c_k)=\sum_{j=1}^{p}\sum_{x\in cj} F(Y) \quad (۶)$$

$$k \leq |c| \leq 2k-1 \quad (۷)$$

$$\sum_{b=1}^{c}\prod_{m=1}^{c} sig|X_{bn}-X_{(b-m)n}|>\beta \quad (۸)$$

---

[۱۰] key

[۱۱] value

[۱۲] Fitness Function

تابع F(Y) تابع هزینه هست، در این مسأله برابر فاصله رکوردها تا مرکز خوشه‌ای که متعلق به آن است که این تابع در قسمت قبل تعریف شد و هدف کمینه کردن این تابع است. رابطه‌های۷ و ۸ قیود مسئله هست. قبل از اینکه به قیود مسئله بپردازیم ابتدا تعریفی برای کلاس هم ارزی ارائه می‌دهیم.

کلاس هم ارزی: به رکوردهایی که درنهایت دارای شبه شناسه یکسانی دارند را یک کلاس هم ارزی می‌گوییم و با E آن را نشان می‌دهیم.

رابطه ۷ قید اول مساله هست که هدف آن برقراری شرط k–گمنامی هست. در k–گمنامی هر دسته از خوشه‌ها یا کلاس هم ارزی می‌تواند بین k تا 2k-1 رکورد در آن خوشه یا کلاس هم ارزی قرار گیرد.

رابطه ۸ قید دوم برای برقراری شرط تنوع در هر کلاس هم ارزی هست که مقدار n نشان‌دهنده ستون صفت حساس هست که باید در هر کلاس حداقل L–تنوع را در این صفت از رکوردها داشته باشیم.

Sig: یک تابع پله‌ای هست که زمانی دو تا رکورد در ویژگی حساس باهم تفاوت داشته باشند مقدار یک و زمانی که یکسان باشد مقدار صفر را برمی‌گرداند.

$$\text{Sig} = \begin{cases} 0 & if\ x_{in} \in E \\ 1 & if\ x_{in} \notin E \end{cases} \qquad (۹)$$

$g(c_1,c_2,\ldots,c_k)$: نشان‌دهنده مراکز دسته‌ها و یا همان کلاس‌های هم ارزی

β: متغیری برای برقراری میزان تنوع در هر دسته و یا همان کلاس هم ارزی و $\beta \in N$ (N مجموعه اعداد طبیعی)

K: متغیری که درجه گمنام سازی را مشخص می‌کند که مقدار آن $k \geq 2$ و $k \in N$

شبه کد الگوریتم ژنتیک در شکل ۵ آمده است:

```
Input = bucket I, privacy parameter k and L
Output = Anonymous Dataset D*
While not termination do
 Parent1,parent2 = getTwoInstance(population)
 Ch1,ch2 = crossover(parent1,parent2)
 Mutation(ch1)
 Mutation(ch2)
 Ch1_fitness = AssignDataAndGetFitness(ch1)
 Ch2_fitness = AssignDataAndGetFitness(ch2)
 Population=
applygeneticoperators(population,ch1,ch2)
```

شکل (۵) : شبه کد الگوریتم جست‌وجوی با استفاده از محاسبات ژنتیک

۵–ارزیابی

از مهم‌ترین معیارها برای ارزیابی روش ارائه‌شده اثبات مقیاس‌پذیری و میزان از دست دادن اطلاعات[3]هست. در کارهای گمنام‌سازی چون هدف اصلی انتشار داده است به‌طوری‌که هم حریم خصوصی افراد حفظ شود و هم داده‌های انتشاریافته کیفیت لازم برای داده‌کاوی را داشته باشد(درواقع هدف ایجاد تعادل بین گمنام سازی و کیفیت داده است) بنابراین مقدار اطلاعاتی که از دست می‌رود مهم هست. در اینجا میزان از دست دادن اطلاعات IL به‌صورت زیر محاسبه می‌شود:

$$\text{IL} = |e|.\left(\sum_{A_k \in QI_N} \frac{(MAX_{A_k} - MIN_{A_k})}{|QI_N|} + \sum_{A_k \in QI_L} \frac{L(A_k)}{H_k}\right) \qquad (۱۰)$$

$$\text{Total\_IL} = \sum_{e \in D} IL(e) \qquad (۱۱)$$

در اینجا e نشان‌دهنده تعداد داده‌هایی است که در یکی از کلاس‌های هم‌ارزی قرار گرفته است.

MAX_{A_k}, MIN_{A_k} : نشان‌دهنده بیشترین و کمترین مقداری است که برای ویژگی A_k در دسته نهایی قرارگرفته است.

QI_N : نشان‌دهنده دامنه تغییرات ویژگی‌های عددی است.

$L(A_k)$: نشان‌دهنده ارتفاع پدر تمام مقدارهای ویژگی A_k هست که در دسته نهایی قرارگرفته‌اند.

H_k : نشان‌دهنده ارتفاع درخت طبقه‌بندی ویژگی A_k است.

D: مجموعه‌ای از تمام کلاس‌های هم‌ارزی است.

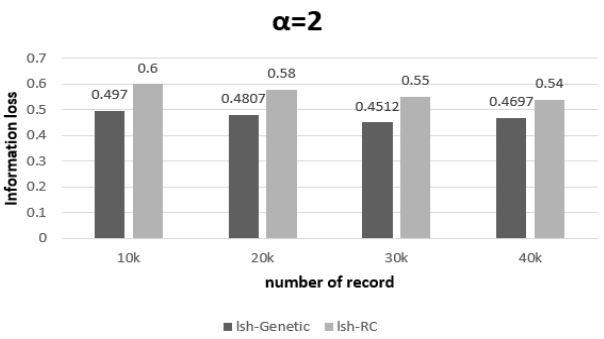
در ادامه به اندازه‌گیری هر یک از این معیارها پرداخته و روش خود را در برابر آن‌ها محک می‌زنیم.

۵–۱ از دست رفتن اطلاعات

به‌منظور ارزیابی روش پیشنهادی از مجموعه داده بزرگسالان[4]از مجموعه داده‌های یادگیری ماشینUCI استفاده می‌کنیم که یک مجموعه داده عمومی است که عموماً به‌عنوان مجموعه داده‌های برای آزمایش الگوریتم‌های گمنام-سازی سازی استفاده می‌شوند. نتایج ارزیابی را روش مقاله [۸] که با عنوان LSH-RC شناخته می‌شود مقایسه می‌کنیم.

آزمایش اول: تأثیر مقدار α در توابع حساس به موضع بر میزان از دست رفتن اطلاعات

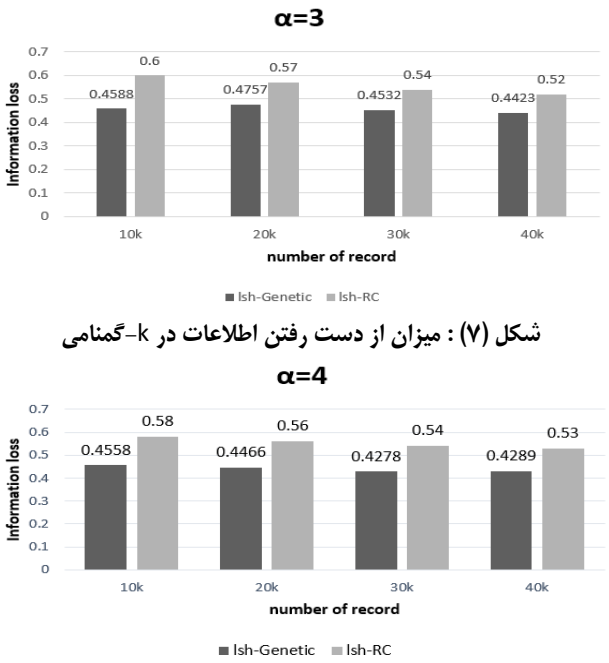
در آزمایش اول به بررسی تأثیر پارامترα در رابطه‌ی $1-(1-s^{\alpha})^{\lambda}$ برای تقسیم‌بندی اولیه داده‌ها می‌پردازیم. در این ارزیابی برای کاهش مقدار افزونگی مقدار پارامتر λ=۱ در نظر گرفته‌شده که مشخص‌کننده تعداد باندها می‌باشد. همچنین پارامتر α که مشخص‌کننده تعداد توابع درهم‌سازی در هر باند است که در آزمایش اول برابر ۲ در نظر گرفته‌شده است. در این آزمایش به بررسی میزان از دست رفتن اطلاعات با تعداد رکوردهای مختلف برای k–گمنامی زمانی که مقدار k=10 است می‌پردازیم. شکل۶، ۷ و ۸ میزان از دست رفتن اطلاعات را در روش ارائه‌شده با روش LSH-RC نشان می‌دهد.



شکل (۶) : میزان از دست رفتن اطلاعات در k–گمنامی

[1] Adult Data Set [4]

[1] information loss [3]



شکل (۷) : میزان از دست رفتن اطلاعات در k–گمنامی

شکل (۸) : میزان از دست رفتن اطلاعات در k–گمنامی

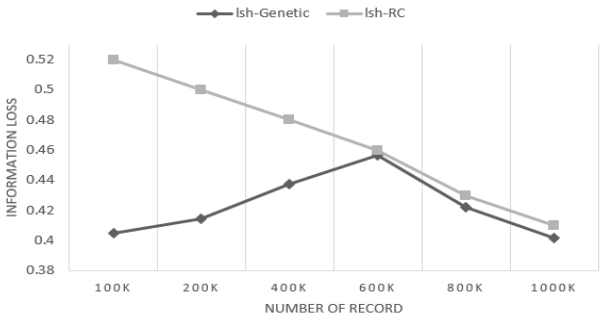
همان‌طور که در نمودارها مشخص‌شده در تمام آزمایش‌ها مقدار از دست رفتن اطلاعات در روش پیشنهادی نسبت به روش LSH-RC بهتر بوده است. بهترین نتایج زمانی حاصل‌شده است که تعداد توابع درهم‌سازی ۴ بوده است. همان‌طور که مشاهده می‌شود در هر دو روش با افزایش تعداد توابع درهم‌ساز تعداد دسته‌ها بیشتر و درنتیجه سایز دسته‌ها کوچک‌تر و احتمال هم دسته شدن داده‌های نامشابه کمتر می‌شود.

در روش پیشنهادی پس از دسته‌بندی با استفاده از تابع حساس به موضع، هر دسته را با استفاده از الگوریتم تکاملی ژنتیک به دسته‌های مطلوب‌تری تقسیم می‌کند. در این الگوریتم سعی می‌شود تا مراکز دسته بهینه‌ای پیدا شوند تا علاوه بر اینکه میزان از دست رفتن اطلاعات کم شود، تعداد عناصر هر دسته نیز در محدوده‌ای تعریف‌شده حفظ شود تا حریم خصوصی نیز نقض نشود

آزمایش دوم: محاسبه میزان از دست رفتن اطلاعات در k–گمنام سازی

در این ارزیابی به بررسی میزان از دست رفتن اطلاعات برای داده‌ها با حجم بالا می‌پردازیم. در این ارزیابی برای افزایش تعداد نمونه‌ها از مجموع داده‌های آزمون و آموزشی استفاده می‌کنیم. حال به دلیل اینکه ارزیابی را برای کلان داده‌ها محاسبه می‌کنیم برای افزایش تعداد رکوردهای مجموعه داده از روش موجود در مقاله [۱۲] استفاده می‌کنیم. در این روش برای افزایش سایز مجموعه داده در ابتدا هر رکورد از داده‌ها را در نظر می‌گیریم و از روی آن β-1 رکورد از آن ایجاد می‌کنیم. این رکورد از داده‌ها به این صورت ایجاد می‌شود که برای هر رکورد q بعد را به‌صورت یک توزیع یکنواخت انتخاب می‌کند. در مرحله بعد با توجه به این q بعد انتخابی، مقدارشان به‌صورت تصادفی با یک مقدار در دامنه آن صفت جایگزین می‌شود.

هدف از این ارزیابی محاسبه میزان از دست رفتن اطلاعات در روش ارائه‌شده و مقایسه آن با روش LSH-RC برای k–گمنامی زمانی که k=10 است می‌پردازیم .

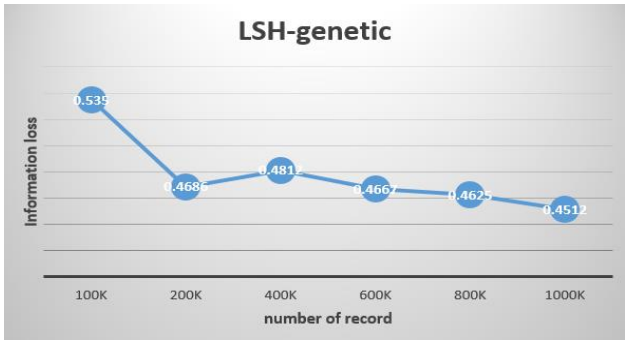


شکل (۹) : میزان از دست رفتن اطلاعات در k–گمنامی

نتایج آزمایش در شکل ۹ نشان می‌دهد که میزان از دست رفتن اطلاعات در تمام حجم داده‌ها از روش LSH-RC بهتر است. این بهبود به چند دلیل صورت گرفته است. یکی اینکه در روش ارائه‌شده مسئله را به یک مسئله بهینه‌سازی تبدیل کرده است و خوشه‌بندی ما یک خوشه‌بندی مقید بهینه هست درصورتی‌که در روش LSH-RC بعدازاینکه داده‌ها را با توابع حساس به موضع به دسته‌ها تقسیم کرد برای رسیدن به k–گمنام‌سازی از یک خوشه بندی ساده برای برقراری تعداد اعضای هر خوشه استفاده کرده است. دلیل بعدی این است که در روش ارائه‌شده سعی بر این بوده تا جای ممکن خوشه‌هایی با اندازه نزدیک به k با توجه به بازه‌های مجاز برای تعداد اعضای هر خوشه تولید کند درصورتی‌که در روشLSH-RC هدف ایجاد خوشه‌هایی با تعداد اعضای بین k تا 2k-1 شده است. همان‌طور که مشخص هست تعمیم برای داده‌ها زمانی که داده‌های بیشتری در یک خوشه هست نسبت به زمانی تعداد کمتری در خوشه هست امکان اینکه بیشتر صورت گیرد زیاد است. همچنین از روند رشد نمودارها می‌توان دریافت که با وجود افزایش تعداد رکوردهای پایگاه داده، کماکان این برتری پایدار می‌ماند.

آزمایش سوم: میزان از دست رفتن اطلاعات درروش L–تنوع و k–گمنامی به‌صورت هم‌زمان

در این ارزیابی به بررسی میزان از دست رفتن اطلاعات حین برقراری حفظ حریم خصوصی افراد زمان انتشار داده برای روش‌های k–گمنامی و l–تنوع به‌صورت هم‌زمان می‌پردازیم. برای افزایش تعداد رکوردها از رویکردی که در آزمایش قبل گفته شد استفاده می‌کنیم. برای انجام این آزمایش مقدار k=10 و مقدار l=4 در نظر می‌گیریم.

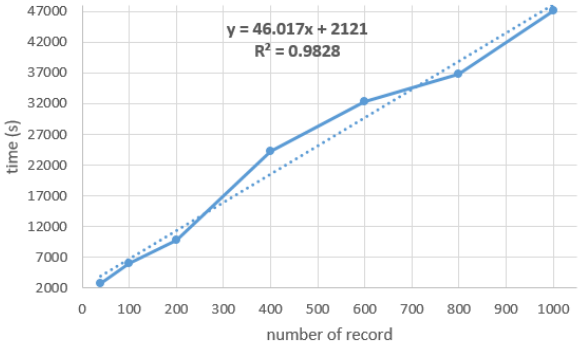


شکل (۱۰) : میزان از دست رفتن اطلاعات در k–گمنامی و l–تنوع

نتایج آزمایش‌ها در شکل ۱۰ نشان می‌دهد که میزان از دست رفتن اطلاعات در روش l–تنوع و k–گمنامی به‌طور هم‌زمان نسبت به روش ارائه‌شده برای k–گمنامی بیشتر هست. دلیل این موضوع این است که ما در k–گمنامی سعی در برقراری شرط تعداد اعضای هر خوشه است ولی در l–تنوع تلاش این هست که علاوه بر برقراری تعداد اعضای هر خوشه شرط تنوع را نیز برای خوشه‌ها رعایت کنیم. درواقع برقراری شرط تنوع گاهی اوقات باعث می‌شود داده‌هایی که فاصله زیادی تا مراکز خوشه دارد برای برقراری تنوع در هر خوشه قرار بگیرد که این باعث می‌شود تعمیم دامنه‌های آن خوشه بیشتر شود و داده‌های پراکنده‌ای در بازه‌ها قرار گیرد. این موضوع باعث می‌شود میزان از دست رفتن اطلاعات نسبت به روش k–گمنامی بیشتر شود.

۲-۵ مقیاس‌پذیری و کارایی

در این بخش به بررسی مقیاس‌پذیری روش پیشنهادی از منظر زمان اجرای الگوریتم، برای حجم مختلفی از داده‌ها می‌پردازیم. برای اثبات مقیاس‌پذیری و کارایی روش ارائه‌شده زمان اجرا را برای تعداد رکوردهای مختلفی از داده‌ها محاسبه می‌کنیم. نمودار شکل ۱۱ نتایج محاسبات را نشان می‌دهد.



شکل (۱۱) : زمان اجرا برای تعداد رکوردهای مختلف

همان‌طور که در نمودار مشخص است زمان اجرا با توجه به افزایش تعداد رکوردها افزایش‌یافته است. شیب نمودار زمان اجرا همان‌طور که از نمودار پیدا است با سرعت پایین در حال افزایش است. حال اگر خط تقریبی که با این زمان اجرا متناسب است را به دست بیاوریم، می‌بینیم خطی با معادله $y = 46.017x + 2121$ است. این خط تناسب داده‌شده، یک خط با معادله درجه اول هست. پارامتر R^2 معیار آماری نزدیکی داده‌ها، به خط رگرسیون برازش شده می‌باشد. به R^2، ضریب تعیین یا ضریب تشخیص نیز گفته می‌شود. ضریب تعیین نشان می‌دهد چه مقدار از تغییرات متغیر وابسته تحت تأثیر متغیر مستقل مربوطه بوده و سایر تغییرات متغیر وابسته مربوط به سایر عوامل می‌باشد. مقدار ضریب تعیین در نمودار برابر ۰,۹۸ است که این مقدار اثباتی بر رشد زمان اجرا به‌صورت خطی است. ازاین‌رو می‌توان نشان داد روش پیشنهادی در مواجه با کلان داده‌ها رفتار شبه خطی خود را حفظ می‌کند و می‌تواند در زمان قابل‌قبول نتایج را ارائه دهد.

۶–جمع‌بندی و کارهای آتی

روش k–گمنامی به‌طور موفقیت‌آمیز نمی‌تواند از افشای ویژگی حساس جلوگیری کند زیرا یک کلاس هم ارزی که همه مقادیر آن یکسان است می‌تواند باعث نقض حریم‌خصوصی شود. روش l–تنوع بهتر از روش k–گمنامی در حفاظت از اطلاعات خصوصی کاربران عمل می‌کند و مشکلات روش k–گمنامی را برطرف می‌کند. بنابراین روشی که در این مقاله ارائه شد برای برقراری k–گمنامی و l–تنوع به‌صورت هم‌زمان است. این رویکرد ارائه‌شده یک روش دو مرحله‌ای است که در مرحله اول با استفاده از توابع درهم‌ساز داده‌ها تقسیم‌بندی می‌شوند و کاندیدهای شباهت برای داده‌ها به دست می‌آید. در مرحله دوم نیز بر روی هرکدام از دسته‌هایی که از مرحله قبل به‌دست‌آمده است الگوریتم ژنتیک اجرا می‌شود تا به خوشه‌هایی با شرایط مدنظر دست پیدا کنیم. در ادامه تحقیقات می‌توان به بررسی دیگر روش‌های حفظ حریم‌خصوصی مانند t-closeness پرداخت به‌گونه‌ای که با ویژگی کلان داده‌ها سازگار باشد و یا ارائه روشی برای زمانی که تعداد ویژگی‌های حساس بیش از یکی باشد.

مراجع

[1] Fung, B., et al., *Privacy-preserving data publishing: A survey of recent developments.* ACM Computing Surveys (CSUR), 2010. **42**(4): p. 14.

[2] Sweeney, L., *k-anonymity: A model for protecting privacy.* International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002. **10**(05): p. 557-570.

[3] Machanavajjhala, A., et al., *L-diversity: Privacy beyond k-anonymity.* ACM Transactions on Knowledge Discovery from Data (TKDD), 2007. **1**(1): p.۳ .

[4] Yin, C., et al., *An improved anonymity model for big data security based on clustering algorithm.* Concurrency and Computation: Practice and Experience, 2017. **29**.(۷)

[5] Cao, J., et al., *Castle: Continuously anonymizing data streams.* IEEE Transactions on Dependable and Secure Computing, 2011. **8**(3): p. 337-352.

[6] Li, T., et al., *Slicing: A new approach for privacy preserving data publishing.* IEEE transactions on knowledge and data engineering, 2012. **24**(3): p. 561-574.

[7] Wang, H ,.et al., *(l, e)-diversity--a privacy preserving model to resist semantic similarity attack.* Journal of Computers, 2014. **9**(1): p. 59-65

[8] Zhang, X., et al. Scalable local-recoding anonymization using locality sensitive hashing for big data privacy preservation. in Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. 2016. ACM

[9] Slaney, M. and M. Casey, *Locality-sensitive hashing for finding nearest neighbors [lecture notes].* IEEE Signal Processing Magazine, 2008. **25**(2): p. 128-131.

[10] Leskovec, J., A. Rajaraman, and J.D. Ullman, *Mining of massive datasets*. 2014: Cambridge university press.

[11] Maulik, Ujjwal, and Sanghamitra Bandyopadhyay. "Genetic algorithm-based clustering technique." *Pattern recognition* 33.9 (2000): 1455-1465.

[12] Mohammed, N., et al., *Centralized and distributed anonymization for high-dimensional healthcare data.* ACM Transactions on Knowledge Discovery from Data (TKDD), 2010 .p. 18.