JIPET

Journal of Intelligent Procedures in Electrical Technology
Vol. 13/ No. 51/ Summer 2022
P-ISSN: 2322-3871, E-ISSN: 2345-5594, http://jipet.iaun.ac.ir/

Publishing Health Information Without Distortion While Balancing Desired Privacy-Preserving and Utility

Abbas Karimi Rizi[1,2], *PhD Student*, Mohammad Naderi Dehkordi[1,2], *Assistant Professor*, Naser Nematbakhsh[3], *Assistant Professor*

[1]Faculty of Computer Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Iran
[2]Big Data Research Center, Najafabad Branch, Islamic Azad University, Najafabad, Iran
abbas.karimi.rizi@mau.ac.ir, naderi@iaun.ac.ir
[3]Shahid Ashrafi Esfahani University, Esfahan, Iran
nemat@eng.ui.ac.ir

Abstract:

In the age of health information analysis, the disease diagnostic code is considered as the patient's privacy. Achieving this code is the most important need for the analysts while anonymizing the code is necessary for people when publishing health information. Disease diagnostic codes, usually presented based on international classifications, are displayed in the form of a taxonomy. In real life, patients only allow the category of the disease diagnostic code to be disclosed, not the original disease diagnostic code. Conventional privacy-preserving models often distort the category of the disease diagnostic code. Preserving privacy accompanying the data utility has always been a critical issue in the dissemination of health information. In this study, a new anonymization method is presented in a way that all attributes of health information can be published without distortion to maintain the utility of the data. So, the published information protects the privacy of patients, so that the experts' expectations and the utility of analysts are desired as expected. The innovative method disseminates health information in a way that the maximum probability of disclosing the disease diagnostic code is always less than or equal to the threat threshold defined by the expert, and on the other hand, the membership analysis error is reduced. The new method is scalable under certain conditions. The results of the practical evaluation of patient data obtained from one of the hospitals in Isfahan are evidence of the effectiveness of the proposed method.

Keywords: disease diagnostic code, taxonomy, membership analysis.

Received: 18 December 2020
Revised: 21 Feb 2021
Accepted: 2 March 2021

Corresponding Author: Dr. Mohammad Naderi Dehkordi

Citation: A. Karimi-Rizi, M. Naderi-Dehkordi, N. Nematbakhsh, "Publishing health information without distortion while balancing desired privacy-preserving and utility", Journal of Intelligent Procedures in Electrical Technology, vol. 13, no. 47, pp. 31-51, September 2022 (in Persian).

نشر اطلاعات سلامت بدون تحریف با حفظ توازن میان حریم خصوصی و سودمندی مطلوب

عباس کریمی ریزی[۱،۲]، دانشجوی دکتری، محمد نادری دهکردی[۱،۲]، استادیار، ناصر نعمت‌بخش[۳]، استادیار

۱- دانشکده مهندسی کامپیوتر- واحد نجف‌آباد، دانشگاه آزاد اسلامی، نجف‌آباد، ایران
۲- مرکز تحقیقات مه داده- واحد نجف‌آباد، دانشگاه آزاد اسلامی، نجف‌آباد، ایران
abbas.karimi.rizi@mau.ac.ir, naderi@iaun.ac.ir
۳- دانشکده مهندسی- دانشگاه شهید اشرفی اصفهانی، اصفهان، ایران
nemat@eng.ui.ac.ir

چکیده: در عصر تحلیل اطلاعات سلامت، کد تشخیص بیماری، حریم خصوصی بیمار است و مهم‌ترین نیاز تحلیل‌گران، دست‌یابی به کد تشخیص بیماری و مهم‌ترین نیاز مردم، گمنام‌سازی کد تشخیص بیماری، هنگام نشر اطلاعات سلامت است. کدهای تشخیص بیماری که معمولاً بر اساس کلاس‌بندی بین‌المللی ارایه می‌شوند، قابل نمایش در قالب یک درخت تاکسونومی هستند. در زندگی واقعی، بیماران فقط اجازه می‌دهند تا رده‌ی کد تشخیص بیماری به جای کد اصلی تشخیص بیماری، منتشر و افشا گردد. مدل‌های متعارف حفظ حریم خصوصی، معمولاً باعث تحریف رده‌ی کد تشخیص بیماری می‌شوند. حفظ حریم خصوصی توأم با سودمندی داده‌ها، همواره، مسأله‌ی حایز اهمیتی در نشر اطلاعات سلامت است. در این پژوهش، روش گمنام‌سازی جدیدی ارایه می‌شود تا جهت حفظ سودمندی داده‌ها، تمام صفات اطلاعات سلامت بتواند بدون تحریف منتشر گردد؛ طوری‌که اطلاعات منتشره هم حافظ حریم خصوصی بیماران تا حد مطلوب متخصصین و هم حافظ سودمندی مطلوب تحلیلگران باشد. این روش، اطلاعات سلامت را طوری منتشر می‌کند که حداکثر احتمال افشای کد تشخیص بیماری، همواره، کوچک‌تر یا مساوی با آستانه‌ی تهدید تعریف شده توسط متخصص باشد و از طرفی میزان خطای تحلیل عضویت، کاهش یابد. روش جدید در شرایط خاص بسط‌پذیر است. نتایج ارزیابی عملی داده‌های بیماران یکی از بیمارستان‌های اصفهان گواه اثربخشی این روش جدید است.

کلمات کلیدی: تاکسونومی، تحلیل عضویت، کد تشخیص بیماری

تاریخ ارسال مقاله: ۱۳۹۹/۹/۲۸
تاریخ بازنگری مقاله: ۱۳۹۹/۱۲/۳
تاریخ پذیرش مقاله: ۱۳۹۹/۱۲/۱۲

نام نویسنده‌ی مسئول: دکتر محمد نادری دهکردی
نشانی نویسنده‌ی مسئول: نجف‌آباد- بلوار دانشگاه- دانشکده مهندسی کامپیوتر- دانشگاه آزاد اسلامی واحد نجف‌آباد

۱- مقدمه

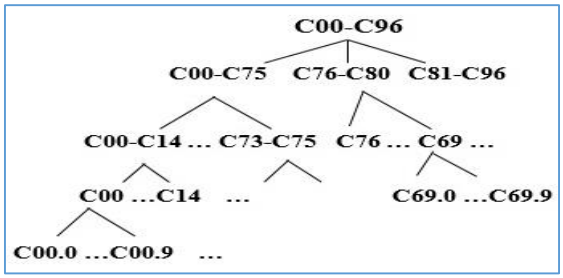
بیمارستان‌ها معمولاً جدول‌های حاوی اطلاعات سلامت بیماران مانند جدول (۱) را جهت تحلیل به مراکز تحقیقاتی معتبر تحویل می‌دهند. اگر فرض گردد که در مراکز تحقیقاتی ممکن است متخاصمی یافت شود که بخواهد به حریم خصوصی داده‌های دریافتی ورود پیدا کند، پس بیمارستان (نگهدارنده‌ی داده‌ها) موظف است، هنگام نشر داده‌ها، محافظت از حریم خصوصی را تضمین کند. نگهدارنده‌ی داده‌ها معمولاً به‌صورت دو مدل غیرقابل اعتماد و قابل اعتماد تصور می‌شود [۱].

Table (1): The original data of patients

جدول (۱): داده‌های اصلی بیماران

کد تشخیص	کدپستی، جنسیت، سن	نام	شناسه‌ی تاپل
C00.0	۱۱۰۰۰، مرد، ۲۳	حسن	۱
C00.4	۱۳۰۰۰، مرد، ۲۷	حسین	۲
C00.4	۱۹۰۰۰، مرد، ۳۵	علی	۳
C00.0	۱۲۰۰۰، مرد، ۲۹	ساسان	۴
C00.6	۵۴۰۰۰، زن، ۶۱	مریم	۵
C69.5	۲۵۰۰۰، زن، ۶۵	ناهید	۶
C69.1	۲۵۰۰۰، زن، ۶۵	سارا	۷
C69.3	۳۰۰۰۰، زن، ۷۰	سوسن	۸

با توجه به تعریف دالنیس [۲] از حفظ حریم خصوصی، صفات جدول (۱) در چهار دسته افراز می‌شوند: صفات شناسه‌ی صریح[۱](نام بیمار)، شبه‌شناسه[۲](سن، جنسیت و کد پستی)، حساس[۳](کد تشخیص بیماری) و غیرحساس[۴](مابقی صفات). مقادیر صفات شبه‌شناسه‌ی دو یا چند بیمار مانند تاپل‌های ۶ و ۷ در جدول (۱) می‌تواند یکسان باشد.



شکل (۱): تاکسونومی کد تشخیص سرطان

Figure (1): The taxonomy of cancer diagnostic code

درخت تاکسونومی مندرج در شکل (۱) که بر اساس رده‌بندی بین‌المللی معروف به ICD-O برپا گشته است حاکی از سلسه مراتب کدهای تشخیص بیماری سرطان است. گره‌های داخلی این تاکسونومی مانند گره‌ای با کد C00 که مربوط به دسته سرطان‌های مرتبط با لب است و خود به زیرکلاس‌های C00.0 (سرطان لب بالایی) تا C00.9 (جنبه‌های نامشخص لب) تقسیم می‌شود حاکی از دسته‌ی عمومی‌تری از کدهای تشخیص سرطان است. بدیهی است، بیمارستان در مرحله‌ی اول، جدولی مانند جدول (۲) که فاقد صفات شناسه‌ی صریح مانند نام بیمار است را برای انتشار فراهم می‌سازد. با توجه به جدول (۲)، هر گروه از تاپل‌هایی مانند تاپل‌های ۶ و ۷ که مقادیر شبه‌شناسه‌ی‌شان دقیقاً یکی است را اصطلاحاً گروه با شبه‌شناسه‌ی یکسان[۵]می‌نامند [۳]. مسأله‌ی قابل تأمل این است که یک متخاصم با داشتن شبه‌شناسه‌ی قربانی هدفش، (سن=۲۳، جنسیت=مرد، کدپستی=۱۱۰۰۰) می‌تواند کد تشخیص بیماری او را با احتمال صددرصد از طریق جدول (۲) با استفاده از پرسش A کشف نماید؛ چنین کشف یا پرسشی را اصطلاحاً حمله‌ی پیوند رکورد گویند.

Table (2): The original data without explicit identifier

جدول (۲): داده‌های اصلی فاقد شناسه‌ی صریح

شناسه‌ی تاپل	کدپستی،جنسیت،سن	کد تشخیص سرطان
۱	۱۱۰۰۰، مرد، ۲۳	C00.0
۲	۱۳۰۰۰، مرد، ۲۷	C00.4
۳	۱۹۰۰۰، مرد، ۳۵	C00.4
۴	۱۲۰۰۰، مرد، ۲۹	C00.0
۵	۵۴۰۰۰، زن، ۶۱	C00.6
۶	۲۵۰۰۰، زن، ۶۵	C69.5
۷	۲۵۰۰۰، زن، ۶۵	C69.1
۸	۳۰۰۰۰، زن، ۷۰	C69.3

Query A: Select Disease Code From Table 2 Where quasi-identifier=”23, Male, 11000”

اکنون اگر متخاصم بداند که قربانی هدفش یک زن ۶۵ ساله با کد پستی ۲۵۰۰۰ است، آنگاه می‌تواند کد تشخیص بیماری او را از طریق جدول (۲) با استفاده از پرسش B کشف نماید که به احتمال ۵۰ درصد برابر با C69.5 و به احتمال ۵۰ درصد C69.1 است؛ چنین کشف یا پرسشی را اصطلاحاً حمله‌ی پیوند صفت[۷] گویند.

Query B: Select Disease Code From Table 2 Where quasi-identifier=”65, Female, 25000”

۱-۱- مسأله‌ی اصلی و ضرورت تحقیق

تکنیک‌های متعارف نشر داده‌ها با حفظ حریم خصوصی معمولاً موجب کاهش سودمندی داده‌ها برای تحلیلگران می‌شوند. مسأله‌ی اصلی، تعدیل میزان حفظ حریم خصوصی و میزان سودمندی داده‌های منتشره است. چون در زندگی واقعی، بین صفات بیماران معمولاً رابطه‌های متعددی وجود دارد -برای مثال، در اطلاعات سلامت بیماران، صفتی مانند کد تشخیص بیماری با صفاتی مانند پزشک، علائم و درمان مرتبط است و تحریف صفت کد تشخیص بیماری ممکن است منجر به شکست تحلیل شود- لذا ضرورت دیده شد تا روشی کنکاش گردد که اطلاعات سلامت بیماران بدون تحریف به‌نحوی منتشر گردند که علاوه‌بر حفظ حریم خصوصی بیماران تا یک آستانه‌ی مطلوب متخصصین، سودمندی مطلوبی نیز برای تحلیلگران حوزه‌ی سلامت تأمین گردد.

۱-۲- عامل اصلی حفظ حریم خصوصی

روش‌های گمنام‌سازی[۸] معمولاً بر اساس دو تکنیک متعارف اقدام به گمنام‌سازی داده‌های حساس جهت حفظ حریم خصوصی می‌کنند: افزونگی تکنیکی و تضعیف ارتباط صفات. افزونگی عامل اصلی برای دفع حملات پیوندی در راستای حفظ حریم خصوصی است. برای مثال جدول (۳) که تعمیم‌یافته‌ی مقادیر شبه‌شناسه‌ی (سن و کدپستی) جدول (۲) است، حایز افزونگی تکنیکی است. عمل تعمیم مقادیر شبه‌شناسه، باعث می‌شود که جدول داده‌های منتشره دارای خاصیت گمنامی سطح k[۹] گردد. جدول داده‌هایی که کاردینالیتی هر گروه با شبه‌شناسه‌ی یکسانش، حداقل k باشد، دارای خاصیت گمنامی سطح k است [۴]. برای مثال، جدول (۳) حایز گمنامی سطح ۴ است؛ زیرا کاردینالیتی هر گروه با شبه‌شناسه‌ی یکسانش، حداقل ۴ است. خاصیت گمنامی سطح k، معمولاً حمله‌ی پیوند رکورد را دفع می‌سازد [۴]، چرا که هر تاپلی در یک گروه با شبه‌شناسه‌ی یکسان، بین حداقل k-1 تاپل دیگر در آن گروه، گمنام می‌شود.

برای مثال، حاصل پرسش C متناظر با پرسش A روی جدول (۳)، دو کد تشخیص سرطان (C00.0 و C00.4) است و همین باعث حفظ حریم خصوصی می‌شود. عیب اصلی تعمیم، کاهش سودمندی داده‌هاست.

Query C: Select Disease Code From Table 3 Where Age>=20 and Age<=40 and Sex=” Male” and Zip Code>=10000 and Zip Code<=20000”

Table (3): The generalized Quasi-identifiers

جدول (۳): شبه‌شناسه‌های تعمیم‌یافته

	کدپستی، جنسیت، سن	کد تشخیص سرطان
گروه با شبه‌شناسه‌ی یکسان اول	۱۰۰۰۰-۲۰۰۰۰، مرد، ۲۰-۴۰	C00.0
	۱۰۰۰۰-۲۰۰۰۰، مرد، ۲۰-۴۰	C00.4
	۱۰۰۰۰-۲۰۰۰۰، مرد، ۲۰-۴۰	C00.4
	۱۰۰۰۰-۲۰۰۰۰، مرد، ۲۰-۴۰	C00.0
گروه با شبه‌شناسه‌ی یکسان دوم	۲۱۰۰۰-۶۰۰۰۰، زن، ۴۱-۸۰	C00.6
	۲۱۰۰۰-۶۰۰۰۰، زن، ۴۱-۸۰	C69.5
	۲۱۰۰۰-۶۰۰۰۰، زن، ۴۱-۸۰	C69.1
	۲۱۰۰۰-۶۰۰۰۰، زن، ۴۱-۸۰	C69.3

عامل دیگر دفع حملات پیوندی، تضعیف ارتباط صفات است. جدولی که هر گروه از تاپل‌هایش حاوی حداقل l مقدار مجزا برای یک صفت حساس در آن گروه باشد، حایز خاصیت تنوع سطح یک l‌است [۵]. جدول (۳) دارای خاصیت تنوع سطح دو است. زیرا هر گروهش حداقل دارای دو نوع کد تشخیص بیماری متمایز است. خاصیت تنوع سطح یک، حمله‌ی پیوند صفت را دفع می‌سازد [۵]. برخی روش‌ها، مانند آناتومی [۶]، از تکنیک تضعیف ارتباط صفات استفاده می‌نمایند. برای مثال، آناتومی، دو جدول (۴) و (۵) را با دریافت جدول (۲) جهت انتشار تولید می‌نماید.

Table (4): The quasi-identifier attributes

جدول (۴): صفات شبه‌شناسه

شناسه‌ی گروه	کدپستی، جنسیت، سن
۱	۱۱۰۰۰، مرد، ۲۳
۱	۱۳۰۰۰، مرد، ۲۷
۲	۱۹۰۰۰، مرد، ۳۵
۲	۱۲۰۰۰، مرد، ۲۹
۲	۵۴۰۰۰، زن، ۶۱
۲	۲۵۰۰۰، زن، ۶۵
۱	۲۵۰۰۰، زن، ۶۵
۱	۳۰۰۰۰، زن، ۷۰

صفت ابداعی شناسه‌ی گروه در دو جدول (۴) و (۵)، جهت بازسازی ناقص ارتباط صفات شبه‌شناسه و حساس استفاده می‌شود و کاربرد دیگری ندارد. حاصل پیوند طبیعی دو جدول (۴) و (۵)، جدولی است با ۳۲ تاپل که ۲۴ تاپل آن به‌عنوان تاپل زاید یا افزونه منظور می‌شوند و همین ۲۴ تاپل زاید باعث گمنامی ۸ تاپل واقعی می‌گردند. تشخیص تاپل‌های زاید، بدون در دست داشتن جدول اصلی، کاری تقریباً غیرممکن است. آناتومی ارتباط صفات را به شدت تضعیف می‌سازد [۶]. برای مثال، حاصل پرسش D متناظر با پرسش A روی دو جدول (۴) و (۵)، چهار کد تشخیص بیماری (C00.0, C00.4, C69.1 و C69.3) است و همین امر باعث حفظ حریم خصوصی بالا و کاهش شدید سودمندی می‌شود.

Query D: Select Disease Code From Table 4, Table 5 Where Group ID of Table 4=Group ID of Table 5 and quasi-identifier="23, Male, 11000"

Table (5): The sensitive attributes

جدول (۵): صفات حساس

شناسه‌ی گروه	کد تشخیص
۱	C00.0
۱	C00.4
۲	C00.4
۲	C00.0
۲	C00.6
۲	C69.5
۱	C69.1
۱	C69.3

۱-۳- تحلیل عضویت

منظور از تحلیل عضویت،[۱] یافتن موارد یا تاپل‌هایی از جدول‌های منتشره است که مطابق با یک درخت تاکسونومی، در یک دسته‌ی مشخص قرار می‌گیرند. برای مثال، حاصل پرسش E به‌عنوان یک تحلیل عضویت که می‌خواهد مشخصات بیمارانی را بیابد که مطابق شکل (۱) کد تشخیص بیماری‌شان در رده‌ی کد تشخیص C69 (هر نوع سرطان چشم) هستند، جدول (۶) است. پرسش E مثال بارزی برای نشان دادن ارتباط بین صفات شبه‌شناسه و حساس (کد تشخیص بیماری) است.

Query E: Select quasi-identifier From Table 2 Where Disease Code in {C69}

Table (6): The query E

جدول (۶): پرسش E

کدپستی، جنسیت، سن	شناسه‌ی تاپل
۲۵۰۰۰، زن، ۶۵	۶
۲۵۰۰۰، زن، ۶۵	۷
۳۰۰۰۰، زن، ۷۰	۸

منظور از تحلیل تجمیع،[۲] اجرای توابع تجمعی مانند Count بر روی صفاتی از یک جدول است [۶]. برای مثال، حاصل پرسش F به‌عنوان یک تحلیل تجمیع، برابر با مقدار ۳ است. برای این کار می‌توان ابتدا با اجرای پرسش E تاپل‌های واجد رده‌ی کد تشخیص C69 را استخراج نمود، یعنی جدول (۶)، سپس تعداد تاپل‌هایش را شمارش کرد.

Query F: Select Count (quasi-identifier) From Table 2 Where Disease Code in {C69}

جهت ارزیابی سودمندی جدول‌های منتشره، در این مقاله، تنها بر تحلیل عضویت تمرکز می‌شود و میزان خطای تحلیل عضویت، به‌عنون میزان سودمندی داده‌های منتشره ارزیابی می‌گردد. فرض کنید، با دو جدول (۴) و (۵) آناتومی بخواهید پرسش G را اجرا نمایید. نتیجه‌ی پرسش G، جدول (۷) است. تاپل‌های حاصل از پرسش G، در دو کلاس مجزا افراز می‌شوند: معتبر و نامعتبر.

Query G: Select quasi-identifier From Table 4, Table 5 Where Group-ID of Table 4= Group-ID of Table 5
And Disease Code in {C69.1, C69.3, C69.5}

تعریف ۱ (تاپل معتبر). تاپلی در جدول حاصل از یک پرسش (مانند پرسش G) روی جدول‌های منتشره که همانندش در جدول حاصل از اجرای پرسشی متناظر با آن پرسش (مانند پرسش E) روی جدول اصلی یافت شود، یک تاپل معتبر تلقی می‌گردد.

برای مثال، تاپل‌های ۱ تا ۵ در جدول (۷) نامعتبر و تاپل‌های ۶ تا ۸ در جدول (۷) معتبرند. با توجه به تعریف صحت عضویت[۷] می‌توان نرخ صحت عضویت و نرخ خطای عضویت را چنین تعریف نمود.

تعریف ۲ (نرخ صحت عضویت).[۱] اگر NV تعداد تاپل‌های معتبر و NI تعداد تاپل‌های نامعتبر در جدول حاصل از اجرای یک پرسش بر روی جدول‌های منتشره و نرخ صحت عضویت با نماد MA نشان داده شود، آنگاه MA برابر است با:

Table (7): The query G

جدول (۷): پرسش G

	کدپستی، جنسیت، سن	شناسه‌ی تاپل
تاپل‌های نامعتبر	۱۱۰۰۰، مرد، ۲۳	۱
	۱۳۰۰۰، مرد، ۲۷	۲
	۱۹۰۰۰، مرد، ۳۵	۳
	۱۲۰۰۰، مرد، ۲۹	۴
	۵۴۰۰۰، زن، ۶۱	۵
تاپل‌های معتبر	۲۵۰۰۰، زن، ۶۵	۶
	۲۵۰۰۰، زن، ۶۵	۷
	۳۰۰۰۰، زن، ۷۰	۸

$$MA = \frac{NV}{NV+NI} \quad (۱)$$

تعریف ۳ (نرخ خطای عضویت)[۱۰]: اگر نرخ خطای عضویت با نماد ME نشان داده شود، آنگاه معیار ME که درواقع، کاملاً مخالف معیار MA است، برابر است با:

$$ME = \frac{NI}{NV+NI} = 1 - MA \quad (۲)$$

برای مثال، مطابق با دو معادله‌ی (۱) و (۲) نرخ صحت عضویت و نرخ خطای عضویت برای پرسش G برابر است با:

$$MA = \frac{3}{8} = 0.375, \quad ME = \frac{5}{8} = 0.625$$

تکنیک آناتومی علی‌رغم این‌که مقادیر تمام صفات را بدون تغییر منتشر می‌سازد، اما دارای نرخ خطای عضویت (ME) بالایی است. جهت ارزیابی میزان سودمندی داده‌های منتشره، می‌توان برای تمام پرسش‌های Q_i اجرا شده توسط تحلیلگر روی جدول‌های منتشره از معیار خطای تحلیل عضویت[۱۱] (MAE) و یا مقدار خطا (Err) یا فاصله‌ی L2 استفاده نمود [۷]:

$$Err_{Q_i} = (ME_{Q_i})^2 \quad (۳)$$

$$MAE = \sum_{\forall Q_i} Err_{Q_i} \quad (۴)$$

منظور از ME_{Q_i}، نرخ خطای عضویت [معادله‌ی (۲)] و Err_{Q_i} مقدار خطا برای پرسش Q_i است. بدیهی است، با محاسبه‌ی Err_{Q_i} برای تمام پرسش‌های Q_i می‌توان گفت که یک تکنیک گمنام‌سازی ایده‌آل، تکنیکی است که مقدار MAE را بسیار نزدیک به صفر سازد.

۱-۴- انگیزه

وقتی در حوزه‌ی نشر داده‌ها با حفظ حریم خصوصی یا پی‌پی‌دی‌پی[۱۲] (PPDP) فرض بر این است که متخاصم، مقادیر شبه‌شناسه‌ی قربانی هدف را می‌داند و همچنین می‌داند که ایشان در بیمارستان بستری بوده است، چه بسا می‌تواند با وجود ابزارهای اطلاعاتی نوین (مانند شبکه‌های اجتماعی) و با در دست داشتن جدول‌های منتشره مدعی بر حفظ حریم خصوصی، به‌راحتی متوجه شود که قربانی هدف در چه بخشی از بیمارستان بستری بوده است و از رده‌ی کد تشخیص بیماری ایشان مطلع گردد. همین امر موجب شد تا روشی تحقیق گردد که بتوان اطلاعات بیماران را به گونه‌ای منتشر نمود که رده‌های کد تشخیص تا حد مطلوب متخصص همراه با مقادیر اصلی تمامی صفات، بدون تحریف، منتشر گردند؛ طوری‌که از طرفی کد اصلی تشخیص بیماران همچنان محرمانه باقی بمانند و از طرف دیگر سودمندی داده‌های منتشره برای تحلیل‌های عضویت، تأمین شود.

۱-۵- مشارکت

در این مقاله یک روش جدید گمنام‌سازی جهت نشر اطلاعات سلامت با حفظ حریم خصوصی مطلوب متخصصین و تأمین سودمندی مطلوب تحلیلگران حوزه‌ی سلامت ارائه شده است. اول، پیشینه‌ای برای روش‌های گمنام‌سازی فراهم خواهیم کرد. دوم، چارچوبی جهت روش جدید بر

اساس آستانه‌ی تهدید و تاکسونومی موردنظر متخصصین تبیین می‌کنیم. سوم، نشان می‌دهیم که روش جدید حافظ حریم خصوصی مشخص و تحت کنترل متخصص است. چهارم، نشان می‌دهیم که روش پیشنهادی به‌طور قابل توجهی باعث افزایش سودمندی داده‌ها به‌واسطه‌ی کاهش خطای تحلیل عضویت می‌شود. پنجم، روش جدید را با روش هم‌تایش مقایسه و ارزیابی می‌کنیم. ششم، نشان می‌دهیم که روش جدید در شرایط خاص به‌راحتی برای هر تعداد صفت حساس، دارای خاصیت بسط‌پذیری است. هفتم، یک الگوریتم جهت کاربردی نمودن روش جدید، ارایه می‌دهیم و آن را تحلیل می‌کنیم. هشتم، به واسطه‌ی آزمایش‌های گسترده اثبات می‌کنیم که روش پیشنهادی از نظر سودمندی داده‌ها به‌طور قابل توجه نسبت به هم‌تایش بهتر عمل می‌کند. سرانجام، مقاله را با بخش نتیجه و راهنمایی برای کارهای آینده به پایان می‌رسانیم.

۲- پیشینه تحقیق

روش‌های گمنام‌سازی معمولاً از حیث تعداد صفت حساس به دو دسته تقسیم می‌شوند: تک صفت حساس و چند صفت حساس. روش‌های متعارف تک صفت حساس: در سال ۱۹۹۸، سوبینی و همکارش، تکنیک تعمیم [۳] را ارایه دادند که منجر به تحریف و تغییر داده‌های اصلی می‌گردد. در سال ۲۰۰۶، تائو و ژیائو، روش حفظ حریم خصوصی شخصی [۸] و وانگ و همکاران گمنامی سطح (α,k) را مطرح نمودند [۹] که چون این دو روش مبتنی‌بر تکنیک تعمیم مقادیر حساس‌اند، باعث کاهش شدید سودمندی داده‌ها می‌شوند. در سال ۲۰۰۷، ژانگ و همکاران گمنامی سطح (k, e) را جهت تنها صفات حساس عددی[۷] مانند حقوق و دستمزد [۱۰] و وانگ و همکاران، روش آستانه‌ی درصد اطمینان را جهت دفع حمله‌ی پیوند صفت [۱۱] ارایه دادند. این دو روش نیز مبتنی‌بر تکنیک تعمیم مقادیر شبه‌شناسه هستند. در سال ۲۰۰۷، لی و همکاران روش نزدیک بودن t را پیشنهاد نمودند [۱۲] که منجر به طرح حمله‌ی جدیدی با نام حمله‌ی چولگی[۸] شد. در سال ۲۰۰۷، راستوگی و همکاران با تعدیل سطح حریم خصوصی و سودمندی داده‌ها، در شرایط خاص، یک روش حریم خصوصی سطح (d, γ) را پیشنهاد نمودند [۱۳]. در سال ۲۰۰۸، بلوم و همکاران یک مدل حریم خصوصی به نام حریم خصوصی توزیعی را برای یک مدل پرسش غیرتعاملی ارایه دادند [۱۴]. در سال ۲۰۰۱۳، دیورک روش حریم خصوصی تفاضلی را مطرح نمود [۱۵]. حریم خصوصی تفاضلی معمولاً برای بانک‌های اطلاعاتی آماری[۹] و پرسش‌های محدود و مشخص استفاده می‌شود.

روش‌های متعارف چند صفت حساس: در سال ۲۰۱۳، هان و همکارانش، روش برش را جهت نشر داده‌های دارای چند صفت حساس ارایه دادند [۱۶] که باعث تحریف داده‌ها می‌گردد. در سال ۲۰۱۵، کینگای لو و همکارانش، روشی برای حداکثر ۲ تا ۳ صفت حساس عددی مطرح نمودند [۱۷]. در سال ۲۰۱۶، کریستوفر و سوزان، روش جدیدی تحت نام آناتومی همراه با برش را جهت حفظ محرمانگی داده‌های با ابعاد بالا و دارای چند صفت حساس ارایه دادند [۱۸] که این روش نیز سودمندی داده‌ها را کاهش می‌دهد. در سال ۲۰۱۸، حسن و همکاران حملات ترکیبی را مورد مطالعه قرار دادند [۱۹]. در روش‌هایی که در سال ۲۰۱۹، توسط کانوال و همکارانش [۲۰] و فاروک و همکارانش [۲۱] مطرح گردید، فرض بر این است که یک شخص ممکن است چندین رکورد دارای چند صفت حساس داشته باشد. در سال ۲۰۲۰، تائو و همکاران با تمرکز روی داده‌های منتشره در دو جدول جداگانه که اولی حاوی صفات همبسته‌ی حساس است و دومی حاوی تنها صفات شبه‌شناسه‌ی تعمیم‌یافته است، روش جدیدی را ارایه دادند[۲۲] که مبتنی‌بر تعمیم است.

ضعف تمام روش‌های مزبور، تحریف یا تضعیف ارتباط برخی صفات است. در جدول (۸)، سه تکنیک متعارف تعمیم، آناتومی و برش که معمولاً اساس روش‌های گمنام‌سازی‌اند، همراه با تکنیک جدید اس‌ان‌آی که اواخر ۲۰۱۹ جهت شبکه‌های اجتماعی [۷] مطرح گردید، مقایسه گردیده‌اند. با توجه به جدول (۸)، ملاحظه می‌گردد که اس‌ان‌آی نسبت به سه تکنیک دیگر از منظر سودمندی داده‌ها و حفظ حریم خصوصی متعادل و قابل کنترل است. چراکه با ابداع صفت جدیدی با نام صفت حساس جزئی هم باعث حفظ ارتباط صفات می‌گردد و هم باعث حل داده‌های پرت (دور افتاده) می‌شود اما کاربردش در گراف‌های متناظر با شبکه‌های اجتماعی است. در این مقاله قرار است با استفاده از نقاط قوت روش اس‌ان‌آی روش جدیدی برای پایگاه داده‌های جدولی ارایه گردد و با روش هم‌تایش آناتومی مقایسه گردد.

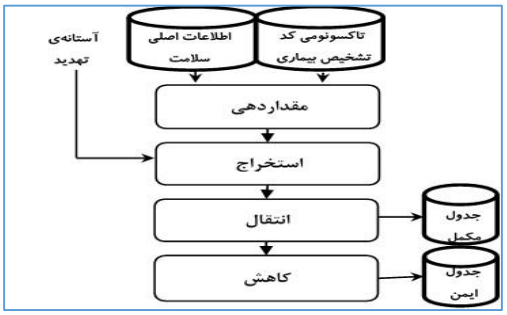
Table (8): The comparison of four basic techniques

جدول (۸): مقایسه‌ی چهار تکنیک پایه

کاربرد اصلی	حل داده‌های پرت	ابداع صفت جدید	تحریف داده	تضعیف شدید ارتباط صفات	حریم خصوصی کنترل شده	تکنیک
جدول	-	-	✓	✓	-	تعمیم
جدول	-	✓	-	✓	-	آناتومی
جدول	-	-	✓	-	-	برش
شبکه‌ی اجتماعی	✓	✓	-	-	✓	اس‌ان‌آی

۳- روش جدید گمنام‌سازی

شکل (۲)، حاکی از چارچوبی برای[۲] روش پیشنهادی جدید جهت نشر اطلاعات سلامت با حفظ حریم خصوصی و سودمندی مطلوب تحلیل‌های عضویت است. در این روش، فرض می‌شود که ۱- صفت کد تشخیص بیماری، دارای یک درخت تاکسونومی و یک آستانه‌ی تهدید پیشنهادی از طرف متخصص است؛ ۲- ناشر یا نگهدارنده‌ی داده‌ها مجاز است که حساسیت مقادیر صفت کد تشخیص بیماری را تا حد مجاز مشخص شده توسط متخصص (آستانه‌ی تهدید) کاهش دهد و چنین مقادیر با حساسیت کم‌تر را تحت نام رده‌ی کد تشخیص بیماری به جای کد اصلی تشخیص بیماری منتشر سازد؛ ۳- نگهدارنده‌ی داده‌ها هیچ دانشی در رابطه با تحلیل داده‌ها ندارد.



شکل (۲): چارچوب روش جدید

Figure (2): The new method framework

با توجه به شکل (۲)، مسؤلیت جمع‌آوری اطلاعات اصلی سلامت و تهیه‌ی درخت‌های تاکسونومی مبتنی‌بر کدهای تشخیص بیماری، با نگهدارنده‌ی داده‌هاست که می‌تواند از فرآیند ای‌تی‌ال[۱] (ETL) [۲۳] استفاده نماید. اطلاعات اصلی سلامت می‌توانند در یک رایانه‌ی محلی، یک سرویس‌دهنده، انبار داده[۲] یا حتی محیط رایانش ابری[۳] مستقر باشند. با فرآیند متعارف ای‌تی‌ال (ETL) ابتدا داده‌های مورد نیاز نشر، از منابع اطلاعاتی مختلف گزینش می‌شوند، سپس در قالب جدول‌های فاقد صفات شناسه‌ی صریح و درخت‌های تاکسونومی و همچنین آستانه‌های تهدید درمی‌آیند و نهایتاً تحت عنوان منابع آماده شده، برای نشر بارگذاری می‌گردند. برای مثال، چنانچه یک نگهدارنده‌ی داده‌ها بخواهد جدول (۱) را به‌گونه‌ای منتشر سازد که حداکثر احتمال افشای کد تشخیص بیماری، بیشتر از ۰/۴ نباشد و مقادیر هیچ صفتی نیز تغییر نکند، باید از طریق جدول (۱)، ابتدا جدول (۲) و درخت تاکسونومی مانند شکل (۱) و همچنین آستانه‌ی تهدید برابر با ۰/۴ را آماده‌سازی و بارگذاری نماید.

چارچوب روش جدید، جهت حفظ حریم خصوصی و سودمندی داده‌ها، با دریافت منابع آماده شده، با اجرای چهار فعالیت دو جدول ایمن و مکمل را تولید می‌کند. برای مثال، چنانچه، جدول (۲) حاوی اطلاعات سلامت بیماران به همراه درخت تاکسونومی کدهای تشخیص بیماری

مندرج در شکل (۱) با آستانه‌ی تهدید برابر با ۰/۴ به‌عنوان ورودی‌های چارچوب شکل (۲)، آماده‌سازی و بارگذاری گردد؛ با استفاده از روش جدید، دو جدول (۹) و (۱۰) جهت نشر، تولید می‌گردند.

Table (9): The Immune data

جدول (۹): داده‌های ایمن

ردەی کد تشخیص	کدپستی، جنسیت، سن	شناسه‌ی تاپل
C00	۱۱۰۰۰، مرد، ۲۳	۱
C00	۱۳۰۰۰، مرد، ۲۷	۲
C00	۱۹۰۰۰، مرد، ۳۵	۳
C00	۱۲۰۰۰، مرد، ۲۹	۴
C00	۵۴۰۰۰، زن، ۶۱	۵
C69	۲۵۰۰۰، زن، ۶۵	۶
C69	۲۵۰۰۰، زن، ۶۵	۷
C69	۳۰۰۰۰، زن، ۷۰	۸

جدول (۹) به‌عنوان جدول ایمن حاوی فقط صفات شبه‌شناسه (سن، جنسیت و کدپستی) و یک صفت جدید با نام صفت ردەی کد تشخیص است. صفت جدید درواقع حاوی مقادیری با حساسیت کم‌تر، عمومی‌تر و قابل نشر از دید متخصص است. جدول (۱۰) که گویای یک جدول مکمل است، درواقع، فاقد صفات شبه‌شناسه و حاوی مقادیر کد تشخیص به همراه فراوانی‌شان در جدول اصلی و همچنین مقادیر ردەی کد تشخیص است. صفت ردەی کد تشخیص عامل پیوند دو جدول ایمن و مکمل است.

سطح حفظ حریم خصوصی: با پیوند طبیعی جدول‌های (۹) و (۱۰) نمی‌توان، جدول اصلی را بازسازی نمود؛ زیرا حاصل چنین پیوندی، جدولی است با ۲۳ تاپل نامعتبر و ۸ تاپل معتبر. برای مثال، چون تاپل ۱ از جدول (۹) دارای ردەی کد تشخیص C00 است، لذا با عمل پیوند دو جدول (۹ و ۱۰) باعث می‌گردد تا جدول حاصل از چنین پیوندی، حاوی یک گروه با شبه‌شناسه‌ی یکسان دارای کاردینالیتی ۵ گردد (۱ تاپل با کد C00.6، ۲ تاپل با کد C00.0 و ۲ تاپل با کد C00.4)؛ همچنین چون تاپل ۶ از جدول (۹) دارای ردەی کد تشخیص C69 است، لذا جدول حاصل از پیوند جدول‌های (۹) و (۱۰)، حاوی یک گروه با شبه‌شناسه‌ی یکسان دارای کاردینالیتی ۳ می‌گردد (۱ تاپل با کد C69.1، ۱ تاپل با کد C69.3 و ۱ تاپل با کد C69.5).

Table (10): The complementary data

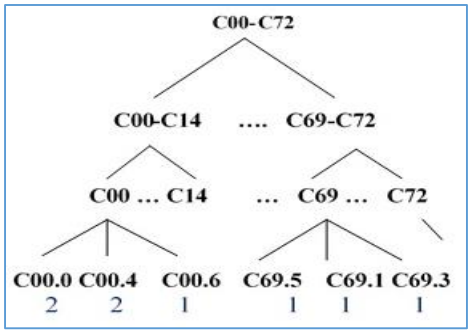
جدول (۱۰): داده‌های مکمل

فراوانی	ردەی کد تشخیص	کد تشخیص
۲	C00	C00.0
۲	C00	C00.4
۱	C00	C00.6
۱	C69	C69.1
۱	C69	C69.3
۱	C69	C69.5

همان‌طور که بیان شد، تشخیص تاپل‌های معتبر از زاید و نامعتبر، بدون در دست داشتن جدول اصلی، کاری تقریباً غیرممکن است. جدول‌های (۹) و (۱۰) هر کدام به تنهایی علاوه‌بر این‌که حافظ حریم خصوصی بیماران سرطانی‌اند قابل استفاده‌ی مؤثر و مفید توسط تحلیلگرند که در قسمت بعد تشریح می‌گردد.

۳-۱- عملکرد روش جدید

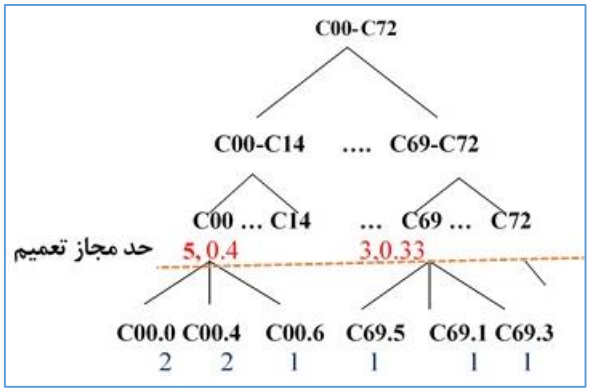
اولین فعالیت چارچوب روش جدید در شکل (۲)، با نام مقداردهی اقدام به مقداردهی فراوانی برگ‌های درخت تاکسونومی از طریق جدول اصلی می‌نماید. برای مثال، فعالیت مقداردهی با دریافت جدول (۲) و شکل (۱) که حاکی از درخت تاکسونومی است، درخت تاکسونومی مندرج در شکل (۳) را تولید می‌کند. برای نمونه چون در جدول (۲) تعداد تاپل‌های حاوی کد C00.0 برابر با ۲ است، لذا در شکل (۳) مقدار فیلد فراوانی برگ C00.0 برابر با ۲ لحاظ می‌شود.



شکل (۳): تاکسونومی حاوی فراوانی

Figure (3): The taxonomy having frequency

دومین فعالیت با نام استخراج، ابتدا اقدام به محاسبه‌ی دو فیلد فراوانی و ماکزیمم فراوانی نسبی گره‌های داخلی درخت تاکسونومی می‌کند تا بتواند یک حد مجاز تعمیم استخراج نماید. برای مثال، فعالیت استخراج با دریافت شکل (۳)، درخت تاکسونومی مندرج در شکل (۴) را تولید می‌کند که حاوی حد مجاز تعمیم است. برای نمونه چون در شکل (۴) مجموع فراوانی برگ‌های C00.0، C00.4 و C00.6 واقع در زیردرخت C00 برابر با ۵ است و بزرگ‌ترین فراونی نسبی برگ‌های C00.0، C00.4 و C00.6 در زیردرخت C00 برابر با 2/5=0.4 است، لذا در شکل (۴) مقدار فیلد فراوانی و ماکزیمم فراوانی نسبی گره‌ی داخلی C00 برابر با ۵ و ۰/۴ لحاظ می‌شود. دقت نمایید که جهت محاسبه‌ی فیلد ماکزیمم فراوانی نسبی C00، ابتدا ماکزیمم فراوانی برگ‌های C00.0، C00.4 و C00.6 واقع در زیردرخت C00 را که برابر با ۲ است به‌دست آوردیم و سپس بر فیلد فراوانی C00 که برابر ۵ است، تقسیم نمودیم.



شکل (۴): تاکسونومی حاوی حد تعمیم

Figure (4): The taxonomy having generalization limit

با توجه به شکل (۴)، فعالیت استخراج بعد از محاسبه‌ی ماکزیمم فراوانی نسبی گره‌های داخلی، درخت تاکسونومی را از سمت برگ‌ها به سمت ریشه می‌پیماید تا اولین گره‌های داخلی را که مقدار فیلد فراوانی نسبی‌شان از مقدار آستانه‌ی تهدید تعریف شده توسط متخصص کوچک‌تر یا

مساوی است پیدا کند و به‌عنوان حد مجاز تعمیم نشانه‌گذاری کند. برای نمونه در شکل (۴) چون فیلد ماکزیمم فراوانی نسبی گره‌های داخلی C00 و C69 از ۰/۴ (آستانه‌ی تهدید) کوچک‌تر یا مساوی‌اند، لذا به‌عنوان حد مجاز تعمیم نشانه‌گذاری می‌گردند که با خط‌چین نشان داده شده‌اند.

تذکر ۱: ماکزیمم فراوانی نسبی هر گره‌ی داخلی، درواقع، مبین حداکثر احتمال افشای مقادیر واقع در برگ‌های زیردرخت آن گره‌ی داخلی است. برای نمونه، مقدار ۰/۴ برای گره‌ی داخلی C00 در شکل (۴)، گویای این است که اگر ناشر به جای کدهای C00.0، C00.4 و C00.6 در جدول ایمن یعنی جدول (۹)، مقدار C00 را قید و منتشر سازد، آنگاه متخاصم مقدار هر یک از این سه کد را حداکثر با احتمال ۴۰ درصد می‌تواند، استنتاج و افشا نماید.

اصل ۱: فعالیت استخراج به‌عنوان مهم‌ترین فعالیت، با پیمایش درخت تاکسونومی از پایین به بالا، مجموعه‌ای از نزدیک‌ترین گره‌های داخلی به برگ‌ها را تحت عنوان حد مجاز تعمیم استخراج می‌نماید تا چنین گرهایی حایز سه شرط ذیل باشند:

- ماکزیمم فراوانی نسبی هر یک از گره‌های داخلی واقع در حد مجاز تعمیم، کوچک‌تر یا مساوی با آستانه‌ی تهدید باشد.
- وجه اشتراک برگ‌های زیردرخت‌های هر یک از گره‌های داخلی واقع در حد مجاز تعمیم برابر با تهی باشد.
- اجتماع برگ‌های زیردرخت‌های تمام گره‌های داخلی واقع در حد مجاز تعمیم برابر با کل برگ‌های درخت تاکسونومی باشد.

تذکر ۲: احراز شرط اول در اصل ۱، عامل تضمین کننده‌ی حفظ حریم خصوصی تا حد مطلوب متخصص (آستانه‌ی تهدید) است. درواقع، میزان حفظ حریم خصوصی با روش جدید به واسطه‌ی همین احراز شرط اول در اصل ۱ باعث می‌شود که کاملاً مشخص و قابل کنترل توسط ناشر گردد.

سومین فعالیت با نام انتقال، اقدام به تبدیل و انتقال یک درخت تاکسونومی حاوی حد مجاز تعمیم به یک جدول تحت نام جدول مکمل می‌نماید. برای مثال، فعالیت انتقال با دریافت درخت تاکسونومی مندرج در شکل (۴)، جدول (۱۰) را تحت نام جدول مکمل تولید می‌کند. با توجه به جدول (۱۰)، ملاحظه می‌گردد که هر جدول مکمل دارای سه صفت است. صفت رده‌ی کد تشخیص که عامل پیوند دهنده‌ی دو جدول ایمن و مکمل است، درواقع، بر مبنای حد مجاز تعمیم تعریف می‌گردد؛ چنین صفتی مقادیرش فقط می‌تواند از مقادیر موجود در گره‌های داخلی مندرج در حد مجاز تعمیم بر روی یک درخت تاکسونومی، برگرفته شود.

چهارمین فعالیت با نام فعالیت کاهش، جدول داده‌های اصلی و جدول مکمل را دریافت و با کاهش درجه حساسیت صفت حساس کد تشخیص بیماری، یک جدول ایمن تولید می‌کند. برای مثال، اگر جدول (۲) و جدول (۱۰) ورودی فعالیت کاهش لحاظ شوند، آنگاه خروجی این فعالیت، جدول (۹) است که دارای مقادیری با حساسیت قابل نشر است؛ چراکه فاقد کدهای اصلی تشخیص بیماری است.

روش جدید مطابق با شکل (۲) مدعی است که در شرایط خاص (یعنی، وجود یک آستانه‌ی تهدید و یک درخت تاکسونومی برای صفت کد تشخیص بیماری)، تمام داده‌های جدول اصلی را بدون تغییر یا تعمیم با حفظ ارتباط نسبی صفات، در قالب یک جدول ایمن و یک جدول مکمل، طوری منتشر می‌سازد که همواره حداکثر احتمال افشای هر مقدار کد تشخیص بیماری، برابر با آستانه‌ی تهدید باشد و سودمندی مطلوبی برای بیشتر تحلیل‌های عضویت فراهم گردد. جهت ارزیابی این ادعا، روش جدید از دو حیث ارزیابی می‌شود: میزان حفظ حریم خصوصی و میزان سودمندی داده‌ها.

۳-۲- میزان حفظ حریم خصوصی

چون جدول مکمل (۱۰) فاقد صفات شبه‌شناسه و جدول ایمن (۹) فاقد صفت حساس کد تشخیص بیماری هستند، لذا هر یک از این دو جدول به‌تنهایی حافظ حریم خصوصی‌اند. مسأله‌ی اصلی این است که میزان حفظ حریم خصوصی در جدول حاصل از پیوند دو جدول مکمل و ایمن چه ارزیابی می‌گردد؟ برای چنین ارزیابی، فرض کنید متخاصمی با داشتن دو جدول (۹) و (۱۰) بخواهد کد تشخیص بیماری یک زن ۷۰ ساله با کد پستی ۳۰۰۰۰ را بیابد. بدیهی است که راحت‌ترین کار اجرای پرسش H است؛ پرسشی که گویای یک پیوند طبیعی بین جدول‌های (۹) و (۱۰) است:

Query H: Select Disease code From Table 9, Table 10 Where Class Code of Table 9= Class Code of Table 10 And quasi-identifier in {"۷۰، زن، ۳۰۰۰۰"}

حاصل اجرای پرسش H، جدول (۱۱) است. بدیهی است با توجه به جدول (۱۱) بیمار موردنظر با احتمال ۰/۳۳دارای یکی از سه کد C69.1، C69.3 و C69.5 است. چراکه احتمال افشای هر یک برابر است با:

$$\Pr\{\text{کد تشخیص} = \text{"C69.1"}\} = \Pr\{\text{کد تشخیص} = \text{"C69.3"}\} = \Pr\{\text{کد تشخیص} = \text{"C69.5"}\} = \frac{1}{3} = 0.33 \leq 0.4$$

Table (11): The query H

جدول (۱۱): پرسش H

فراوانی	کد تشخیص	...، جنسیت، سن	شناسه‌ی تاپل
۱	C69.5	...، زن، ۷۰	۸
۱	C69.3	...، زن، ۷۰	۸
۱	C69.1	...، زن، ۷۰	۸

ملاحظه می‌گردد که احتمال افشای هر کد تشخیص بیماری همواره کوچک‌تر از ۰/۴ (آستانه‌ی تهدید) است؛ چراکه میزان حفظ حریم خصوصی با تکنیک جدید، به واسطه‌ی رعایت اصل ۱ در روش جدید و وجود مقدار آستانه‌ی تهدید کاملاً مشخص و قابل کنترل است. در فعالیت استخراج، طبق اصل ۱، مقرر گردید که مقادیر صفت ابداعی رده‌ی کد تشخیص که عامل پیوند دو جدول ایمن (۹) و مکمل (۱۰) و به تبع آن تضمین حفظ حریم خصوصی مطلوب متخصصین هستند، تنها مقادیر گره‌های واقع در حد مجاز تعمیم باشند.

۳-۳- میزان سودمندی داده‌ها

با روش جدید، جدول‌های ایمن و مکمل، هر یک به تنهایی می‌توانند توسط تحلیلگران مورد استفاده واقع شوند؛ چراکه جدول ایمن حاوی رده‌ی کد تشخیص و جدول مکمل حاوی خود مقادیر کد تشخیص همراه با فراوانی‌شان است. جدول (۹) به‌عنوان جدول ایمن، چون حاوی صفت رده‌ی کد تشخیص بیماری است که از تعمیم صفت کد تشخیص بیماری حاصل می‌گردد لذا در خیلی مواقع می‌تواند پاسخگوی تحلیل‌های عضویت باشد و یا حتی نقش مؤثری در پاک‌سازی داده‌های پرت داشته باشد. جهت ارزیابی روش جدید از حیث تحلیل عضویت، تمام پرسش‌های ممکن که در جدول (۱۲) ذکر شده است، برای سه حالت بررسی می‌گردند: ۱- ارزیابی سودمندی جدول ایمن به تنهایی؛ ۲- ارزیابی جدول ایمن به همراه جدول مکمل؛ ۳- ارزیابی جدول مکمل به تنهایی. دقت نمایید که پرسش‌های مندرج در جدول (۱۲) بر اساس جدول اصلی است، لذا چنانچه بخواهیم این پرسش‌ها را روی جدول‌های منتشره اعمال نماییم باید بازسازی گردند.

Table (12): All questions according to Figure (4)

جدول (۱۲): تمام پرسش‌های مطابق با شکل (۴)

Query	Select	From	Where
Q1	Quasi-identifier	Original Table	Disease in {C69.1}
Q2	Quasi-identifier	Original Table	Disease in {C69.3}
Q3	Quasi-identifier	Original Table	Disease in {C69.5}
Q4	Quasi-identifier	Original Table	Disease in {C00.6 }
Q5	Quasi-identifier	Original Table	Disease in {C00.4}
Q6	Quasi-identifier	Original Table	Disease in {C00.0}
Q7	Quasi-identifier	Original Table	Disease in {C00}
Q8	Quasi-identifier	Original Table	Disease in {C69}

ارزیابی سودمندی جدول ایمن به تنهایی: برای برخی از پرسش‌ها مانند Q_7 و Q_8 در جدول (۱۲)، جدول ایمن به تنهایی پاسخ‌گوست. برای مثال، پرسش Q_8 جهت یافتن مشخصات بیمارانی که دچار یکی از بیماری‌هایی با کد C69.1، C69.3 و C69.5 هستند، لحاظ شده است. بدیهی است، با استفاده از جدول اصلی (۲)، پرسش Q_8 به‌راحتی قابل اجراست. کافی است پرسش I اجرا گردد. حاصل اجرای پرسش I ، تاپل‌های ۶ تا ۸ از جدول (۲) است.

Query I: Select quasi-identifier From Table 2 Where Disease Code of Table 2 in {"C69.1", "C69.3"," C69.5"}

اما، اگر دو جدول (۹) و (۱۰) در دست باشد، کافی است پرسش J تنها روی جدول (۹) اجرا گردد. حاصل اجرای J تاپل‌های ۶ تا ۸ از جدول (۹) است که هر سه تاپل معتبر تلقی می‌شوند و درواقع حاصل پرسش J فاقد تاپل نامعتبر است.

Query J: Select quasi-identifier From Table 9 Where Disease Code of Table 9 in {"C69"}

با این حساب، برای پرسش J طبق معادلات (۱) و (۲) داریم:

$$MA = \frac{3}{3} = 1 \ , \ ME = \frac{0}{3} = 0$$

و چون خطای تحلیل عضویت برابر با صفر است، پس سودمندی داده‌ها برای چنین تحلیل عضویتی به‌صورت ۱۰۰٪ تأمین می‌گردد.

ارزیابی سودمندی جدول ایمن و جدول مکمل: برای برخی از پرسش‌ها مانند Q_1 تا Q_6 ترکیب جدول ایمن و جدول مکمل پاسخ‌گوست. برای مثال، فرض کنید، تحلیلگری بخواهد با پرسش Q_1 مندرج در جدول (۱۲)، مشخصات بیمارانی که دچار سرطان قرنیه چشم (C69.1) هستند را بیابد. اگر ایشان جدول (۲) را داشته باشد، به راحتی درمی‌یابد که تنها تاپل ۷ از جدول (۲) حایز چنین بیماری است. اما، اگر دو جدول (۹) و (۱۰) را داشته باشد کافی است پرسش K را روی دو جدول (۹) و (۱۰) اجرا نماید. حاصل اجرای پرسش K، تاپل‌های ۶ تا ۸ است که فقط تاپل ۷ معتبر است. یعنی حاصل K تنها حاوی ۱ تاپل معتبر و ۲ تاپل نامعتبر است.

Query K: Select quasi-identifier From Table 9, Table 10 Where Class Code of Table 9= Class Code of Table 10 And Disease code in {"C69.1"}

بنابراین طبق معادلات (۱) و (۲) داریم:

$$MA = \frac{1}{3} = 0.33, \ ME = \frac{2}{3} = 0.67$$

و از این‌رو، چون خطای تحلیل عضویت برابر با ۰/۶۷ است، لذا سودمندی داده‌ها برای چنین پرسش‌هایی زیاد تأمین نمی‌گردد؛ چراکه جدول ایمن و جدول مکمل باید حافظ حریم خصوصی کدهای تشخیص اصلی باشند و نباید اجازه دهند احتمال افشای مقادیر اصلی کد تشخیص بیماری ا ز آستانه‌ی تهدید بیشتر شوند.

ارزیابی سودمندی جدول مکمل به تنهایی: برای برخی تحلیل‌های عضویت به ویژه تحلیل‌های تجمیع، جدول مکمل به تنهایی پاسخ‌گوست. برای مثال، اگر تحلیلگری بخواهد با پرسش F تعداد بیماران واجد رده‌ی کد C69 را بیابد به‌راحتی می‌تواند پاسخ خود را مستقیماً از جدول (۱۰) به‌دست آورد. جدول مکملی که فاقد مقادیر شبه‌شناسه است و همین باعث حفظ حریم خصوصی بیماران می‌گردد. جدول مکمل حتی پاسخ‌گوی تحلیل‌های تجمیع روی مقادیر اصلی کد تشخیص بیماری است. برای مثال، تحلیلگری به‌راحتی می‌تواند تعداد بیماران با کد C69.1 را از طریق جدول (۱۰) بیابد، بدون این که بتواند به مشخصات بیماران دست پیدا کند.

۴- روش جدید در مقیاسه با آناتومی

آناتومی در مقایسه با روش جدید، یکی از مهم‌ترین روش‌هایی است که مدعی است تمام مقادیر صفات جدول اصلی را عیناً بدون تحریف منتشر می‌کند. آیا آناتومی می‌تواند جایگزین روش پیشنهادی جدید گردد؟ شکل (۵) گویای نتایج اجرای تمام پرسش‌های مندرج در جدول (۱۲) با به‌کارگیری دو روش آناتومی و روش جدید به‌طور جداگانه است.

شکل (۵): نرخ خطای عضویت برای پرسش‌های Q1 تا Q8

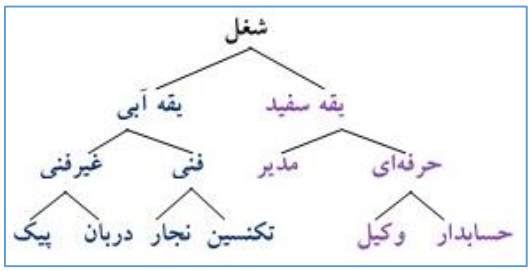
Figure (5): The membership error rate for queries Q1-Q8

همان‌طور که در شکل (۵) دیده می‌شود، با روش جدید پرسش‌هایی که شروط رده‌ی کد تشخیص بیماری‌شان از سمت برگ‌ها دور می‌شوند و به سمت ریشه‌ی درخت تاکسونومی نزدیک می‌شوند، باعث کسب خطای تحلیل عضویت نزدیک یا حتی برابر با صفر می‌شوند. اما با آناتومی معمولاً هیچ پرسشی دارای خطای تحلیل عضویت برابر با صفر نمی‌شود؛ چراکه آناتومی بدون توجه به درخت تاکسونومی، جدول اصلی را جهت تأمین نیاز تنوع سطح l، به‌طور تصادفی به گروه یا باکت‌های مجزا تقسیم و افراز می‌کند. درصورتی‌که روش جدید طبق اصل ۱ با توجه به درخت تاکسونومی و هدفمند، چنین افرازهایی را انجام می‌دهد تا حداکثر احتمال افشای هر مقدار کد تشخیص بیماری کم‌تر از آستانه‌ی تهدید موردنظر متخصص درآید. روش جدید برخلاف آناتومی جهت جدول حاوی چند صفت حساس، قابل اجراست، با این محدودیت که هر صفت حساس باید دارای یک درخت تاکسونومی و یک آستانه‌ی تهدید باشد.

۴-۱-بسط‌پذیری روش جدید

روش جدید برخلاف آناتومی هر تعداد صفت حساس دارای یک درخت تاکسونومی را به‌طور مستقل پشتیبانی می‌کند. برای مثال، با توجه به شکل (۶)، روش جدید جهت نشر جدول (۱۳) که حاوی دو صفت حساس کد تشخیص با آستانه تهدید برابر با ۰/۴ و عنوان شغلی با آستانه تهدید برابر با ۰/۵ و درخت تاکسونومی شکل (۶) است، جدول‌های (۱۴)، (۱۰) و (۱۵) را منتشر می‌سازد.

میزان حفظ حریم خصوصی: چون روش جدید برای هر صفت حساس جداگانه عمل می‌کند و یک جدول مکمل مستقل تولید می‌نماید؛ لذا طبق اصل ۱ و قضیه‌ی پیشامدهای مستقل، حافظ حریم خصوصی چندین صفت حساس است.



شکل (۶): تاکسونومی عنوان شغلی
Figure (6): The job taxonomy

Table (13): The original data having two sensitive attributes

جدول (۱۳): داده‌های اصلی حاوی دو صفت حساس

عنوان شغلی	کد تشخیص	کدپستی، جنسیت، سن
دربان	C00.0	۱۱۰۰۰، مرد، ۲۳
پیک	C00.4	۱۳۰۰۰، مرد، ۲۷
مدیر	C00.4	۱۹۰۰۰، مرد، ۳۵
وکیل	C00.0	۱۲۰۰۰، مرد، ۲۹
حسابدار	C00.6	۵۴۰۰۰، زن، ۶۱
وکیل	C69.5	۲۵۰۰۰، زن، ۶۵
تکنسین	C69.1	۲۵۰۰۰، زن، ۶۵
تکنسین	C69.3	۳۰۰۰۰، زن، ۷۰

میزان سودمندی داده‌ها: فرض کنید، تحلیلگری با پرسشی بخواهد مقادیر شبه‌شناسه‌ی بیمارانی را بیابد که در رده‌ی شغلی یقه‌آبی و رده‌ی کد تشخیص C00 قرار دارند. با جدول (۲) نتیجه‌ی چنین پرسشی دو تاپل ۱ و ۲ از جدول (۲) است. اما، اگر جدول (۱۴) در دسترس باشد، راحت‌ترین

راه، اجرای پرسش L است که نتیجه‌اش همان دو تاپل ۱ و ۲ از جدول (۱۴) با نرخ خطای عضویت صفر است؛ زیرا جدول (۱۴) حاوی اطلاعات کامل موردنیاز پرسش L است.

پرسش L مثال بارزی برای ارزیابی میزان ارتباط صفات کد تشخیص، عنوان شغلی و شبه‌شناسه است. در روش جدید، هرچه شروط پرسش‌ها از سمت برگ‌ها به سمت ریشه‌ی درخت‌های تاکسونومی نزدیک‌تر شوند، نرخ خطای عضویت نیز به سمت صفر نزدیک‌تر می‌شود.

Query L: Select quasi-identifier From Table 14 Where job in ("یقه آبی") And disease code in ("C00")

Table (14): The immune data

جدول (۱۴): داده‌های ایمن

رده‌ی شغلی	رده‌ی کد تشخیص	کدپستی، جنسیت، سن
یقه آبی	C00	۱۱۰۰۰، مرد، ۲۳
یقه آبی	C00	۱۳۰۰۰، مرد، ۲۷
یقه سفید	C00	۱۹۰۰۰، مرد، ۳۵
یقه سفید	C00	۱۲۰۰۰، مرد، ۲۹
یقه سفید	C00	۵۴۰۰۰، زن، ۶۱
یقه سفید	C69	۲۵۰۰۰، زن، ۶۵
یقه آبی	C69	۲۵۰۰۰، زن، ۶۵
یقه آبی	C69	۳۰۰۰۰، زن، ۷۰

Table (15): The complementary data

جدول (۱۵): داده‌های مکمل

عنوان شغلی	رده‌ی شغلی	فراوانی
دربان	یقه آبی	۱
پیک	یقه آبی	۱
مدیر	یقه سفید	۱
وکیل	یقه سفید	۲
حسابدار	یقه سفید	۱
تکنسین	یقه آبی	۲

۵- الگوریتم روش جدید

الگوریتم روش جدید مندرج در شکل (۷)، جدول داده‌های اصلی حاوی اطلاعات بیماران مانند جدول (۲)، درخت تاکسونومی مانند شکل (۱) و آستانه‌ی تهدید مانند ۰/۴ را به‌عنوان ورودی دریافت و یک جدول ایمن مانند جدول (۹) و یک جدول مکمل مانند جدول (۱۰) قابل نشر را به‌عنوان خروجی تولید می‌کند. الگوریتم شامل چهار فعالیت است که در ادامه، تشریح می‌گردد.

فعالیت اول الگوریتم (خطوط ۱ تا ۳)، مسؤل مقداردهی فیلد فراوانی (F) برگ‌های درخت تاکسونومی است. از لحاظ پیاده‌سازی، هر برگ از درخت تاکسونومی، دارای دو فیلد S و F است: فیلد S برای ثبت یک مقدار کد تشخیص بیماری و فیلد F برای ثبت فراوانی S در جدول داده‌های اصلی است. با اجرای فعالیت اول، درواقع، فراوانی هر کد تشخیص بیماری در جدول اصلی به یک برگ در درخت تاکسونومی منتقل و نگاشت می‌گردد.

Algorithm:
Input: *An original table, a taxonomic tree, and an attack threshold*
Output: *An immune table and a supplementary table*
Begin:
/ Initialization activity (lines 1-3) updates the leaves in the taxonomic tree */*
1. *For each leaf* $x \in$ *Taxonomicc tree do*
2. $x[Frequency\ ] \leftarrow$ *frequency of* $x[$*disease code*$]$ *in Original Table;*
3. *End for*

/ Extraction activity (lines 4-10) extracts the permitted generalization limit in the taxonomic tree*/*
*/*the lines 4-7 calculate the internal nodes */*
4. *For each internal node* $y \in$ *Taxonomic tree do*
5. $y[Frequency] \leftarrow$ *summation of Frequency of leaves in subtree* y;
6. $y[Relative\ frequency\ ] \leftarrow \frac{Maximum\ Frequency\ of\ leaves\ in\ subtree\ y}{y[Frequency]}$
7. *End for*

*/*the lines 8-10 extract the permitted generalization limit*/*
8. *For Taxonomic tree do*
9. *Traverse Taxonomic tree by bottom-up and find* m *internal nodes* $y_k \in$ *Taxonomic tree as a permitted generalization limit where*
$y[Relative\ frequency] \leq$ *attack threshold and* $\cup_{k=1}^{m}\{leaves\ in\ subtree\ of\ y_k\} = \{all\ leaves\ of\ Taxonomic\ tree\}$ *and* $\{leaves\ in\ subtree\ of\ y_k\} \cap$ *leaves in subtree of* $y_j\} = \emptyset$.
10. *End for*

/ Transformation activity (lines 11-17) transfers the taxonomic tree to the supplementary table (ST) */*
11. *Create ST with three attributes (S, P, F).*
12. *For each* $x \in$ *Taxonomicc tree do*
13. *Insert a new record t' in ST.*
14. $t'[S] \leftarrow x[S]$.
15. $t'[F] \leftarrow x[F]$.
16. $t'[P] \leftarrow$*ancestor of* x *that is in the permitted generalization limit*.
17. *End for*

/ Reduction activity (lines 17-20) reduces sensitivity of the sensitive attribute of Original Table and extracts an immune table */*
18. *For each the tuple t of Original Table do // tuple* $t \in$ *Original Table*
19. *find an internal node* $t' \in$ *Supplementary Taxonomic Tree where* $t[Disease\ Code]=t'[Disease\ Code]$ *then* $t[Disease\ Code] \leftarrow t'[Disease\ Code]$.
20. *End for*
21: *Return the Original table and the Taxonomic tree.*
End

شکل (۷): الگوریتم روش جدید
Figure (7): The new method algorithm

فعالیت دوم الگوریتم (خطوط ۴ تا ۱۰) مسؤل محاسبه‌ی دو فیلد فراوانی و ماکزیمم فراوانی نسبی گره‌های داخلی درخت تاکسونومی (خطوط ۴ تا ۷) و به تبع آن استخراج حد مجاز تعمیم (خطوط ۸ تا ۱۰) است. از لحاظ پیاده‌سازی، هر گره‌ی داخلی، دارای سه فیلد F, P و R جهت ثبت رده، فراوانی و ماکزیمم فراوانی نسبی برگ‌هایش است. فیلد P هر گره‌ی داخلی، گویای یک کد رده برای تمام مقادیر واقع در برگ‌های زیردرخت آن گره‌ی داخلی است؛ فیلد F هر گره‌ی داخلی، گویای مجموع فراوانی تمام برگ‌های واقع در زیردرخت آن گره‌ی داخلی است؛ اما فیلد R هر گره‌ی داخلی، گویای ماکزیمم فراوانی نسبی تمام برگ‌های واقع در زیردرخت آن گره‌ی داخلی است. فیلد ماکزیمم فراوانی نسبی هر گره‌ی داخلی از طریق مجموع فراوانی تمام برگ‌های واقع در زیردرخت آن گره‌ی داخلی (خط ۵) و فیلد ماکزیمم فراوانی نسبی‌اش از طریق تقسیم بزرگ‌ترین فراوانی برگ‌های واقع در زیردرختش بر فیلد فراوانی خود گره‌ی داخلی (خط ۶) مقداردهی می‌گردد.

با مجموعه دستورات واقع در خطوط ۸ تا ۱۰، درخت تاکسونومی به‌طور جداگانه از پایین به بالا پیمایش می‌گردد تا مجموعه‌ای از گره‌های داخلی‌اش تحت نام حد مجاز تعمیم، استخراج و نشانه‌گذاری گردد. با توجه به شروط واقع در خط ۹، گره‌های داخلی می‌توانند در مجموعه‌ی حد مجاز تعمیم قرار گیرند که اولاً مقدار فیلد ماکزیمم فراوانی نسبی‌شان کوچک‌تر یا برابر با آستانه‌ی تهدید باشد، ثانیاً اشتراک برگ‌های گره‌های واقع در مجموعه‌ی حد مجاز تعمیم، تهی باشد و ثالثاً اجتماع برگ‌های گره‌های واقع در مجموعه‌ی حد مجاز تعمیم، برابر با کلیه‌ی برگ‌های درخت تاکسونومی باشد- مانند C00 و C69 در شکل (۴).

فعالیت سوم الگوریتم (خطوط ۱۱ تا ۱۷) مسؤل تولید یک جدول مکمل از طریق درخت تاکسونومی و حد مجاز تعمیم است- مانند جدول (۱۰). در این بخش، درواقع، به‌ازای هر برگ درخت تاکسونومی یک تاپل جدید در جدول مکمل ایجاد و مقادیر هر برگ (خط ۱۴) به همراه فراوانی (خط ۱۵) و جد آن برگ که در حد مجاز تعمیم واقع است (خط ۱۶) به تاپل جدیداً ایجاد شده، منتقل می‌گردد.

بخش چهارم الگوریتم (خطوط ۱۸ تا ۲۰)، مسؤل تولید یک جدول ایمن قابل نشر دارای سودمندی مطلوب و حافظ حریم خصوصی است- مانند جدول (۹). با این فعالیت، درواقع، با جایگزینی مقادیر کد تشخیص بیماری در جدول اصلی با مقادیر رده‌ی کد تشخیص، حساسیت مقادیر کد تشخیص هر تاپل از جدول اصلی تا حد مجاز تعمیم مشخص شده، کاهش می‌یابد (خط ۱۹) و سودمندی مطلوب تحلیل عضویت حاصل می‌گردد.

۵-۱- تحلیل الگوریتم

الگوریتم شکل (۷) نیازمند $O(n/b)$ عمل ورودی و خروجی است؛ جایی‌که n کاردینالیتی جدول داده‌های اصلی و b سایز صفحات دیسک است. پیچدگی زمانی الگوریتم نیز $O(m^2)$ است؛ جایی‌که m حاکی از تعداد برگ‌های درخت تاکسونومی است. چون تعداد گره‌های درخت تاکسونومی محدود، مشخص و ثابت است، لذا جهت پیاده‌سازی درخت تاکسونومی برپا شده برای کد تشخیص بیماری می‌توان از یک آرایه‌ی یک بعدی استفاده کرد که هر عنصرش حاوی یک رکورد دارای ۷ فیلد شماره رکورد، نوع گره (داخلی یا برگ)، کد تشخیص بیماری گره، کد تشخیص بیماری گره والد، شماره رکورد والد (جهت برقراری یال هر گره با گره والد)، فراوانی و فراوانی نسبی باشد. از فیلد شماره رکورد والد جهت پیمایش درخت از سَمت برگ‌ها به سمت ریشه استفاده می‌شود. اگر مجموع تعداد برگ و گره‌های داخلی درخت تاکسونومی برابر با λ لحاظ گردد، پس الگوریتم نیازمند $O(\lambda)$ حافظه است. با فعالیت اول الگوریتم (خطوط ۱ تا ۳) یک آرایه‌ی یک بعدی دارای λ عنصر در حافظه پیاده‌سازی و حفظ می‌گردد و در فعالیت دوم که مهم‌ترین و زمان‌برترین فعالیت الگوریتم است، تمام محاسبات بر روی این آرایه صورت می‌گیرد.

چون هر درخت تاکسونومی، معمولاً در عمل دارای تعداد برگ و گره‌ی داخلی کاملاً محدود، مشخص و تعریف شده توسط یک متخصص است و از طرفی هر جدول داده‌های حاوی اطلاعات سلامت بیماران با هر کاردینالیتی روی برگ‌های درخت تاکسونومی دارای گره‌های محدود نگاشت می‌گردد، لذا الگوریتم از قابلیت اطمینان بالایی برخوردار است. الگوریتم درواقع تمام عملیات و محاسبات لازم را مستقل از دیسک یا هر انباره‌ی ذخیره‌سازی دیگری، فقط روی آرایه‌ی برپا شده‌ی متناظر با درخت تاکسونومی انجام می‌دهد.

۶- آزمایش

برای آزمایش الگوریتم جدید، اطلاعات ۶۷۱۷ مورد سرطانی در یکی از بیمارستان‌های اصفهان، کد تشخیص سرطان به‌عنوان صفت حساس و آستانه‌ی تهدید برابر با ۰/۳۴ درنظر گرفته می‌شود. شکل (۱) نمایانگر بخشی از تاکسونومی صفت کد تشخیص سرطان با ۳۱ گره‌ی داخلی و ۴۵۴ برگ است که در آزمایش استفاده می‌گردد. در شکل (۱)، درخت تاکسونومی براساس رده‌بندی بین‌المللی ICD-O [۲۴] کدهای تشخیص سرطان برپا می‌گردد. جدول (۱۶) مبین برخی رده‌های کد تشخیص سرطان نزدیک به ریشه‌ی درخت تاکسونومی مندرج در شکل (۱) است.

گره‌های داخلی، درواقع، حاکی از دسته‌ی کلی‌تری از کدهای سرطان است. برای مثال، گره‌ی داخلی با کد C00 مربوط به دسته سرطان‌های مرتبط با لب است که خود به زیرکلاس‌ها یا گره‌های خارجی C00.0 تا C00.9 تقسیم می‌شود. کد C00.0 مربوط به جنبه‌های خارجی لب بالایی، کد C00.1 مربوط به جنبه‌های خارجی لب پایینی و نهایتاً کد C00.9 مربوط به جنبه‌های نامشخص لب است. گره‌ی داخلی با کد C00-C14 که خود به ۱۵ گره‌ی داخلی دیگر تقسیم می‌شود (C00 تا C14)، مربوط به دسته‌ی عمومی‌تری است؛ یعنی دسته سرطان‌های مرتبط با لب، دهان و گلو است. جدول (۱۷) تعدادی پرسش برای آزمایش دو روش جدید و آناتومی به‌طور جداگانه را شامل می‌شود.

شکل (۸) نمایانگر نرخ خطای تحلیل عضویت آن پرسش‌ها با آستانه‌ی تهدید برابر با ۰/۳۴ برای دو روش جدید و آناتومی است. با توجه به شکل (۸) ملاحظه می‌شود که نرخ خطای عضویت پرسش 'M برای روش جدید برابر با صفر و برای آناتومی بزرگ‌تر از ۰/۵ است. با توجه به جدول (۱۷) چون شرط پرسش 'M برابر با کد C00-C75 است و با توجه به جدول (۱۶) این کد مربوط به دسته کدهای واقع در بازه‌ی C00 تا C75 (سرطان‌های لب، دهان، گلو، اندام گوارشی، تیروئید و غیره) است، پس صفر بودن نرخ خطای عضویت پرسش 'M حاکی از این است که تحلیلگر، به‌راحتی می‌تواند بیماران دارای رده‌ی کد تشخیص C00-C75 (بیماران حاوی کد C00.0 تا C75.9) را از جدول ایمن استخراج نماید.

Table (16): The classification of the cancer codes

جدول (۱۶): کلاس‌بندی کدهای سرطان [۲۴]

بازه‌ی کدها	کلاس توده‌های بدخیم	بازه‌ی کدها	کلاس توده‌های بدخیم
C00-C14	لب، دهان و گلو	C51-C58	اندام تناسلی زن
C15-C26	اندام گوارشی	C60-C63	دستگاه تناسلی مردان
C30-C39	اندام‌های تنفسی و داخل قفسه سینه	C64-C68	دستگاه ادراری
C40-C41	استخوان و غضروف مفصلی	C69-C72	چشم، مغز و دیگر بخش‌های سیستم عصبی مرکزی
C43-C44	ملانوم و دیگر توده‌های بدخیم پوست	C73-C75	تیروئید و دیگر غدد درون ریز
C45-C49	مزوتلیال و بافت نرم	C76-C80	محل‌های بد تعریف شده، ثانوی و نامشخص
C50-C50	پستان	C81-C96	لنفاوی، خون‌ریزی و بافت مرتبط

Table (17): The main Queries with the taxonomy in Figure (1)

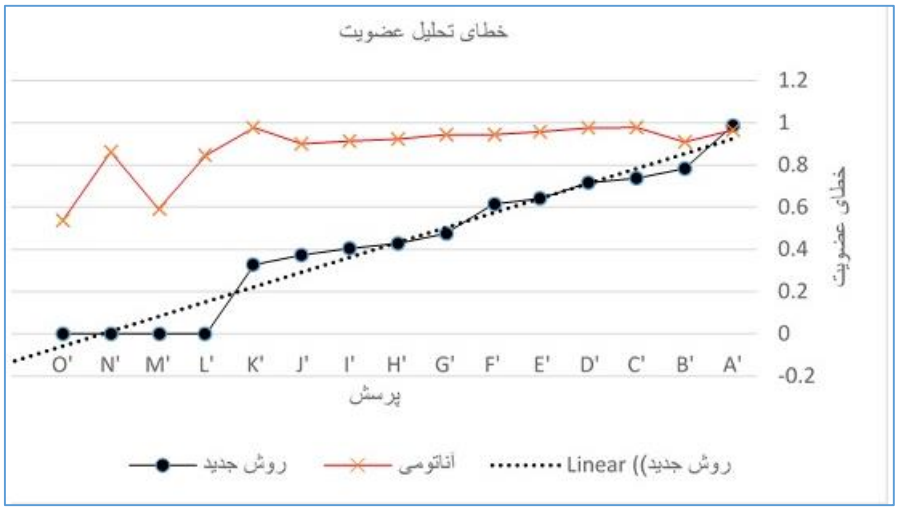
جدول (۱۷): پرسش‌های اصلی با تاکسونومی شکل (۱)

Query	Select	From	Where
A'	Quasi-identifier	Original Table	Code in {C00-C14}
B'	Quasi-identifier	Original Table	Code in {C15-C26}
C'	Quasi-identifier	Original Table	Code in {C30-C39}
D'	Quasi-identifier	Original Table	Code in {C40-C41}
E'	Quasi-identifier	Original Table	Code in {C43-C44}
F'	Quasi-identifier	Original Table	Code in {C45-C49}
G'	Quasi-identifier	Original Table	Code in {C50-C50}
H'	Quasi-identifier	Original Table	Code in {C51-C58}
I'	Quasi-identifier	Original Table	Code in {C60-C63}
J'	Quasi-identifier	Original Table	Code in {C64-C68}
K'	Quasi-identifier	Original Table	Code in {C69-C72}
L'	Quasi-identifier	Original Table	Code in {C73-C75}
M'	Quasi-identifier	Original Table	Code in {C00-C75}
N'	Quasi-identifier	Original Table	Code in {C76-C80}
O'	Quasi-identifier	Original Table	Code in {C81-C96}

این حالت یعنی جدول ایمن برای پرسش 'M صددرصد سودمند واقع می‌شود. لازم به تذکر است که این می‌تواند بیماران با رده‌ی کد تشخیص C00-C75 را از جدول‌های شبه‌شناسه و حساس، با احتمالی کم‌تر از ۵۰ درصد استخراج نماید. متخاصم فقط قادر به استخراج رده‌ی کد تشخیص بیماری قربانی هدف یا به عبارتی بازه‌ای از کدهاست؛ نه این که بتواند به کد اصلی او دست پیدا کند. سودمندی جدول‌های حاصل از آناتومی برای پرسش 'M کم‌تر از ۵۰ درصد است. با توجه به شکل (۸) ملاحظه می‌شود که برای هر دو روش جدید و آناتومی نرخ خطای عضویت پرسش 'A تقریباً برابر با ۱ است.

نزدیک بودن نرخ خطای عضویت پرسش 'L به مقدار ۱، حاکی از این است که تحلیلگر، به‌راحتی نمی‌تواند مشخصات بیماران با کد تشخیص واقع در بازه‌ی C00 تا C14 را از جدول‌های منتشره توسط روش جدید یا آناتومی استخراج نماید. وقتی تحلیلگر با پرسشی مانند 'A شروع به یافتن اطلاعات جزئی‌تر می‌نماید، یعنی به سمت برگ‌های درخت تاکسونومی حرکت می‌کند تا به اطلاعات جزئی‌تر قربانی هدف دست یابد، خطای تحلیل نیز به ۱ نزدیک‌تر می‌شود تا حافظ حریم خصوصی بیماران باشد.

مطابق با شکل (۸)، خطای تحلیل عضویت (MAE) طبق معادله‌ی ۴ برای تکنیک جدید برابر با ۴/۲۵ و برای آناتومی برابر با ۱۱/۹۹ محاسبه شده است و همین گویای برتری روش جدید نسبت به تنها روش هم‌تایش یعنی آناتومی است. با تغییر مقدار آستانه‌ی تهدید، یعنی ۰/۳۴، ملاحظه گردید که هر چه این مقدار، بیشتر شود، سودمندی نیز بیشتر می‌شود و برعکس.



شکل (۸): نرخ خطای عضویت برای پرسش‌های 'A تا 'O

Figure (8): The membership error rate for queries A'-O'

۸- نتیجه‌گیری

روش جدید، تمام مقادیر جدول اصلی را بدون تغییر در قالب جدول‌های ایمن و مکمل، منتشر می‌سازد؛ طوری که چنین جدول‌هایی، هر کدام به‌تنهایی یا با ترکیب دیگر جدول‌ها، حافظ سطح حریم خصوصی مورد نظر متخصص و سودمندی مطلوب برای بیشتر تحلیل‌های عضویت است. روش جدید، با ایجاد و نشر رده‌ی مقادیر حساس در جدول ایمن باعث تشدید افزونگی تکنیکی که عامل اصلی تضمین حفظ حریم خصوصی است و تقویت همبستگی نسبی صفات که عامل اصلی افزایش سودمندی داده‌هاست، می‌گردد. روش جدید، جدول داده‌های دارای چند صفت حساس همبسته را نیز به نحو مطلوب برای بیشتر تحلیل‌های عضویت پشتیبانی می‌کند. تنها محدودیتش این است که هر صفت حساس باید دارای یک درخت تاکسونومی و یک آستانه‌ی تهدید موردنظر متخصص باشد. روش جدید، راه را برای محققان باز می‌گذارد تا بتوانند سودمندی داده‌ها را بیشتر افزایش دهند. چراکه این روش، نه تنها به فراوانی مقادیر حساس معطوف نمی‌شود، بلکه به مقدار آستانه‌ی تهدید و ساختار درخت تاکسونومی نیز بستگی دارد.

References

مراجع

[1] T. Dalenius, "Finding a needle in a haystack or identifying anonymous census records", Journal of Official Statistics, vol. 2, no.3, pp. 315-328, 1986.

[2] L. H. Cox, "Suppression methodology and statistical disclosure control", Journal of the American Statistical Assocciation, vol. 75, pp. 377-385, 1980.

[3] P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information", PODS '98: Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, May. 1998 (doi.org/10.1145/275487.275508).

[4] L. Sweeney, "k-anonymity: A model for protecting privacy", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 05, pp. 557-570, 2002 (doi: 10.1142/S0218488502001648).

[5] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, "L–diversity: Privacy beyond k-anonymity", ACM Transactions on Knowledge Discovery from Data, vol. 1, no. 1, pp. 1-16, March. 2007 (doi: 10.1145/1217299.1217302).

[6] X. Xiao and Y. Tao, "Anatomy: Simple and effective privacy preservation", Proceedings of the VLDB, pp. 139-150, Sept. 2006 (doi: 10.5555/1182635.1164141).

[7] A. Karimi Rizi, M. Naderi Dehkordi, N. Nemat bakhsh, "SNI: Supervised anonymization technique to publish social networks having multiple sensitive labels", Security and Communication Networks, vol. 2019, Article ID 8171263, pp. 1-23, 2019 (doi.org/10.1155/2019/8171263).

[8] X. Xiao, Y. Tao, "Personalized privacy preservation", Proceedings of the SIGMOD, pp. 229-240, June 2006 (doi: 10.1145/1142473.1142500).
[9] R. C. W. Wong, J. Li, A. W. C. Fu, K. Wang, "(α, k)-anonymity: An enhanced k-anonymity model for privacy-preserving data publishing", Proceedings of the KDD, pp. 754-759, Aug. 2006 (doi: 10.1145/1150402.1150-499).
[10] Q. Zhang, N. Koudas, D. Srivastava, T. Yu, "Aggregate query answering on anonymized tables", Proceeding of the IEEE/ICDE, Istanbul, Turkey, pp. 116-125, 2007 (doi: 10.1109/ICDE.2007.367857).
[11] K. Wang, B. C. M. Fung, P. S. Yu, "Handicapping attacker's confidence: An alternative to k- Anonymization", Knowledge and Information Systems, vol. 11, pp. 345-368, 2007 (doi: 10.1007/s10115-006-0035-5).
[12] L. Ninghui, L. Tiancheng, S. Venkatasubramanian, "t-Closeness: Privacy beyond k-anonymity and l-diversity", Proceeding of the IEEE/ICDE, Istanbul, Turkey, pp. 106-115, 2007 (doi: 10.1109/ICDE.2007.3-67856).
[13] V. Rastogi, D. Suciu, S. Hong, "The boundary between privacy and utility in data publishing", VLDB '07: Proceedings of the 33rd international conference on Very large data bases, pp. 531-542, September 2007.
[14] A. Blum, K. Liqett, A. Roth, "A learning theory approach to non-interactive database privacy", Proceedings of the ACM, pp. 609-618, 2008 (doi.org/10.1145/1374376.1374464).
[15] C. Dwork, A. Roth, "The algorithmic foundations of differential privacy", Foundations and Trends in Theoretical Computer Sience, vol. 19, no. 3-4, pp 211-407, 2014 (doi: 10.1561/0400000042).
[16] J. Han, F. Luo, J. Lu, H. Peng, "SLOMS: A privacy preserving data publishing method for multiple sensitive attributes microdata", Journal of Software, vol. 8, no. 12, pp. 3096-3104, 2013 (doi: 10.4304/jsw.8.12.3096-3104).
[17] Q. Liu, H. Shen, Y. Sang, "A privacy-preserving data publishing method for multiple numerical sensitive attributes via clustering and multi-sensitive bucketization", Proceeding of the PAAP, Beijing, China, pp. 220-223, 2014 (doi: 10.1109/PAAP.2014.56).
[18] V. S. Susan, T. Christopher, "Anatomisation with slicing: a new privacy preservation approach for multiple sensitive attributes", SpringerPlus, vol. 5, no. 964, pp. 1-18 2016 (doi.org/10.1186/s40064-016-2490-0).
[19] A. Hasan, Q. Jiang, H. Chen, and S. Wang, "A new approach to privacy-preserving multiple independent data publishing", Applied Sciences, vol. 8, no. 5, pp. 783, 2018 (doi.org/10.3390/app8050783).
[20] T. Kanwal, S.A.A. Shaukat, A. Anjum, S.R. Malik, K.K.R Choo, A Khan, N Ahmad, M. Ahmad, S.U. Khan, "Privacy-preserving model and generalization correlation attacks for 1:M data with multiple sensitive attributes", Information Sciences, vol. 488, pp. 238-256, 2019 (doi: 10.1016/j.ins.2019.03.004).
[21] A. Anjum, N. Farooq, S. U. R. Malik, A. Khan, M. Ahmed, M. Gohar, "An effective privacy preserving mechanism for 1: M microdata with high utility", Sustainable Cities and Society, vol. 45, pp. 213, Feb. 2019 (doi.org/10.1016/j.scs.2018.11.037).
[22] R. Khan, X. Tao, A. Anjum, H. Sajjad, S.R. Malik, A. Khan, F. Amiri, "Privacy preserving for multiple sensitive attributes against fingerprint correlation attack satisfying c-diversity", Wireless Communications and Mobile Computing, vol. 2020, Article ID 8416823, pp. 1-18, 2020 (doi.org/10.1155/2020/8416823).
[23] S. K. Bansal, "Towards a semantic extract-transform-load (ETL) framework for big data integration", in Proceeding of the IEEE/ICBD, pp. 522-529, Anchorage, AK, USA, June/July 2014 (doi: 10.1109/BigDat-a.Congress.2014.82).
[24] K. Fearon, F. Strasser, S. D. Anker, Bosaeus, E. Bruera, R. L. Fainsinger, "Definition and classification of cancer cachexia: An international consensus", The Lancet Oncology, 2011 (doi.org/10.1016/S1470-2045(10)70218-7).

زیرنویس‌ها:

[1] Explicit identifier attributes
[2] Quasi-identifier attributes
[3] Sensitive attributes
[4] Non-sensitive attributes
[5] QI-group
[6] Record linkage attack
[7] Attribute linkage attack
[8] Anonymization
[۹] K-anonymity
[۱۰] l-diversity

۱۷Membership analysis
۱۸Aggregate analysis
۱۹Membership accuracy rate
۲۰Membership error rate
۲۱Membership analysis error
[1] PPDP: privacy-preserving data publishing
۲۲Numerical
۲۳Skewness attack
۲۴Statistical database
۲۵Framework
۲۶ETL: Extraction, transfer, load
۲۷Data warehouse
۲۸Cloud computing