ارایهٔ یک روش جدید انتشار داده‌ها با حفظ محرمانگی با هدف بهبود دقّت طبقه‌بندی روی داده‌های گمنام

رضا ابراهیمی آتانی* و مهدی صادق‌پور
گروه مهندسی کامپیوتر، دانشکدهٔ فنی، دانشگاه گیلان، رشت، ایران

چکیده

با توسعهٔ روزافزون خدمات دولت الکترونیکی، اطلاعات شخصی افراد در قالب پایگاه‌های داده در دستگاه‌ها و ارگان‌های دولتی و خصوصی ذخیره شده است. در بسیاری از موارد برای پردازش و استخراج دانش از این منابع دادهٔ بزرگ و باارزش، نیاز به انتشار منابع داده و دراختیارگذاشتن اطلاعات به سایر نهادها و شرکت‌ها پدید می‌آید که این امر موجب ایجاد چالش‌های امنیتی در نقض حریم خصوصی افراد می‌شود. در این مقاله ضمن بررسی کامل پیشینهٔ پژوهش، حفظ محرمانگی در انتشار داده‌ها، یک روش کارآمد برای گمنام‌سازی ارائه می‌شود که هدف آن حفظ دقت طبقه‌بندی روی داده‌های گمنام است. این روش با بهره‌گیری از درخت تصمیم از انتشار اطلاعاتی که تأثیر کمی بر سودمندی داده‌های خروجی دارد و حذف آن‌ها موجب تأمین محرمانگی می‌شود، جلوگیری می‌کند. یکی از چالش‌های طرح‌هایی که از عمل‌گر گمنام‌سازی عمومی‌سازی استفاده می‌کنند، نیازمندی به ساخت درخت طبقه‌بندی برای هر شبه‌شناسه است که بیش‌تر به‌صورت خودکار صورت می‌گرفت. در طرح پیشنهادی نیازی به ساخت درخت طبقه‌بندی نیست. نتایج شبیه‌سازی و ارزیابی‌های انجام‌شده نشان می‌دهد، میان دقت الگوریتم‌های طبقه‌بندی که روی مجموعه‌داده استاندارد گمنام‌شده توسط این روش و مجموعه‌دادهٔ اولیه آموزش دیده‌اند، تفاوت اندکی وجود دارد.

واژگان کلیدی: حفظ محرمانگی، طبقه‌بندی، گمنام‌سازی، درخت تصمیم، عمل‌گر فرونشانی.

A New Privacy Preserving Data Publishing Technique Conserving Accuracy of Classification on Anonymized Data

Reza Ebrahimi Atani* & Mehdi Sadeghpour
Department of Computer Engineering, University of Guilan, Rasht, Iran

Abstract

Data collection and storage has been facilitated by the growth in electronic services, and has led to recording vast amounts of personal information in public and private organizations databases. These records often include sensitive personal information (such as income and diseases) and must be covered from others access. But in some cases, mining the data and extraction of knowledge from these valuable sources, creates the need for sharing them with other organizations. This would bring security challenges in user's privacy. The concept of privacy is described as sharing of information in a controlled way. In other words, it decides what type of personal information should be shared and which group or person can access and use it. "Privacy preserving data publishing" is a solution to ensure secrecy of sensitive information in a data set, after publishing it in a hostile environment. This process aimed to hide sensitive information and keep published data suitable for knowledge discovery techniques. Grouping data set records is a broad approach to data anonymization. This technique prevents access to sensitive attributes of a specific record by eliminating the distinction between a

* نویسندهٔ عهده‌دار مکاتبات
* Corresponding author

number of data set records. So far a large number of data publishing models and techniques have been proposed but their utility is of concern when a high privacy requirement is needed. The main goal of this paper to present a technique to improve the privacy and performance data publishing techniques. In this work first we review previous techniques of privacy preserving data publishing and then we present an efficient anonymization method which its goal is to conserve accuracy of classification on anonymized data. The attack model of this work is based on an adversary inferring a sensitive value in a published data set to as high as that of an inference based on public knowledge. Our privacy model and technique uses a decision tree to prevent publishing of information that removing them provides privacy and has little effect on utility of output data. The presented idea of this paper is an extension of the work presented in [20]. Experimental results show that classifiers trained on the transformed data set achieving similar accuracy as the ones trained on the original data set.

Keywords: Privacy preservation, Data sharing, Anonymization, Classification, Decision tree, Suppression.

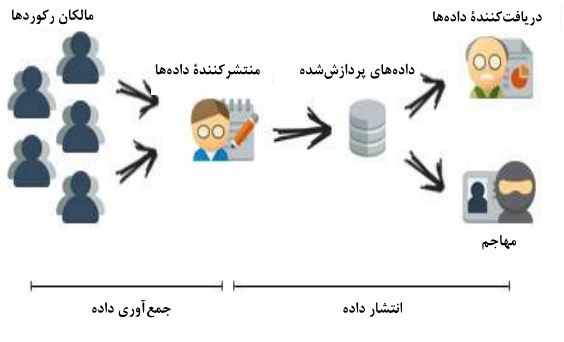
۱- مقدّمه

با توسعه و نفوذ هر چه بیشتر خدمات الکترونیکی در زندگی روزمرهٔ افراد و سازمان‌ها، سامانه‌های رایانه‌ای و پایگاه‌های داده، محل پردازش و جمع‌آوری اطلاعات بسیار مهمی از زندگی افراد شده است. نظر به حجم بسیار زیاد اطلاعات در این منابع داده، کشف الگوها و قوانین سودمند، روش‌های داده‌کاوی را به ابزاری کلیدی برای استفاده از این منابع باارزشِ داده تبدیل کرده است. مالکان این مجموعه‌داده‌ها ممکن است، به‌دلایل مختلفی تصمیم به انتشار آن‌ها بگیرند؛ اما به‌دلیل وجود اطلاعات حساس و شخصی مربوط به افراد، نمی‌توانند این داده‌ها را به‌صورت خام انتشار دهند. فرآیند انتشار با حفظ محرمانگی داده[۱] (PPDP) راه‌کاری برای حل این مشکل در محیط‌های غیرقابل اعتماد است [1,4].

در این فرآیند سعی بر آن است که ضمن مخفی نگه‌داشتن اطلاعات حساس، داده‌های منتشرشده، همچنان برای عملیات کشف دانش مفید باقی بماند [3]. در ادامه، مفاهیم پایه‌ای PPDP که در مقاله مورد استفاده قرار می‌گیرند، تعریف شده‌اند:

- **مالکان رکورد:** افرادی هستند که اطلاعات مربوط به آن‌ها در مجموعه‌داده ثبت شده است.
- **منتشرکنندهٔ داده‌ها:** سازمانی است که مجموعه‌داده را در اختیار دارد و اقدام به انتشار آن می‌کند.
- **مهاجم:** فرض بر این است که بعد از انتشار داده‌ها، مهاجم قادر خواهد بود، نسخه‌ای از آن را به‌دست آورد. وی تلاش می‌کند اطلاعاتِ حساسِ یک یا چند مالک رکورد (که قربانی نامیده می‌شوند) را از مجموعه‌داده استخراج کند. اطلاعات در دسترس مهاجم فقط به داده‌های منتشرشده محدود نیست و امکان دارد از دانش پیش‌زمینهٔ خود نیز برای انجام حملات استفاده کند.
- **دریافت‌کنندهٔ داده‌ها:** پژوهش‌گران علمی یا سازمان‌های تجاری که به‌دنبال کشف دانش از مجموعه‌دادهٔ منتشرشده هستند.
- **گمنام‌سازی داده‌ها:** فرآیندی است که داده‌های خام را به داده‌های گمنام‌شده تبدیل می‌کند. داده‌های گمنام‌شده در مقابل نوع خاصی از حملات که در یک مدل محرمانگی تعریف شده است، از افشای اطلاعات مصون می‌مانند.

در PPDP سعی بر این است که با پردازش مجموعه دادهٔ اولیه هویت مالکان رکوردها، پنهان نگه‌داشته شوند تا به حریم شخصی آنها لطمه‌ای وارد نشود. همان‌طورکه در شکل (۱) نمایش داده شده است، دو مرحلهٔ اصلی فرآیند PPDP جمع‌آوری و انتشار داده‌ها هستند [4]. در مرحله نخست منتشرکنندهٔ داده‌ها اطلاعات را از کاربران خود جمع‌آوری می‌کند. فرض بر این است که مالکان رکورد به منتشرکنندهٔ داده‌ها اطمینان دارند و حاضرند اطلاعات حساس خود را با او به اشتراک بگذارند، اما دریافت‌کنندگان داده، مورد اعتماد نیستند و اطلاعات حساس باید از آن‌ها مخفی نگاه داشته شود؛ بنابراین در مرحله بعد روش‌های حفظ محرمانگی روی داده‌ها اعمال می‌شود و خروجی در اختیار دریافت‌کنندگان داده قرار می‌گیرد.



(شکل-۱): جمع‌آوری و انتشار داده‌ها
(Figure-1): Data collection and data publishing

[۱] Privacy-Preserving Data Publishing (PPDP)

هر دریافت‌کننده ممکن است، داده‌های مورد نیاز خود را از چندین منتشرکننده جمع‌آوری کند. به‌عنوان مثال، یک شرکت بیمه برای انجام داده‌کاوی در زمینه رابطه بین شغل افراد و یک بیماری خاص، نیاز به داده‌های ذخیره‌شده در بیمارستان‌های شهر دارد. در این مثال بیماران مالکان رکورد، بیمارستان‌ها منتشرکنندگان داده و شرکت بیمه دریافت‌کننده داده‌ها است.

صفات رکوردها در یک مجموعه‌داده از لحاظ کاربرد در PPDP به چند دسته تقسیم می‌شوند. به صفاتی که مالک رکورد مایل نیست عموم از آنها اطلاع داشته باشند، صفت حساس گفته می‌شود (مانند نوع بیماری و یا میزان حقوق یک فرد). از سوی دیگر صفاتی که به‌صراحت مالک رکورد را مشخص می‌کنند، شناسهٔ صریح نامیده می‌شوند. هدف PPDP پنهان‌ساختن تناظر بین شناسه صریح و صفت حساس رکوردها است. نخستین قدم در گمنام‌سازی مجموعه داده، حذف شناسه‌های صریح از رکوردها است؛ اما همان‌طورکه در [5] نشان داده شده است، حتی درصورت حذف شناسهٔ صریح به‌کمک دسته‌ای دیگر از صفات که شبه‌شناسه[1] نامیده می‌شوند، می‌توان صفت حساس مالکان رکوردها را به‌دست آورد. کدپستی، تاریخ تولد و جنسیت را به‌عنوان مثالی از شبه‌شناسه‌ها می‌توان نام برد که هر یک به‌تنهایی برای تعیین مالک رکورد کافی نیستند؛ اما با ترکیب آنها می‌توان به هویت یک فرد خاص رسید.

برای جلوگیری از این مشکل می‌توان با استفاده از عمل‌گرهای گمنام‌سازی، موجب شد تا شبه‌شناسه‌های هر رکورد در چندین رکورد دیگر نیز تکرار و مهاجم در شناسایی مالک رکورد دچار ابهام شود. از میان عمل‌گرهای گمنام‌سازی پرکاربرد، می‌توان به عمل‌گرهای عمومی‌سازی[2] و فرونشانی[3] اشاره کرد [2]. در عمل‌گر عمومی‌سازی، مقدار یک صفت به‌کمک درخت طبقه‌بندی[4] به بازه‌ای بزرگ‌تر تعمیم داده می‌شود؛ برای مثال در صفت نشانی، مقدار ایران با خاورمیانه جایگزین می‌شود و در مرحله بعد می‌توان خاورمیانه را با آسیا جایگزین کرد.

مشکل این روش، نیاز به تشکیل درخت طبقه‌بندی برای هر شبه‌شناسه است که در بیش‌تر موارد این کار به‌صورت دستی و توسط یک فرد خبره انجام می‌گیرد [3]. علاوه‌براین ممکن است، بین درخت‌های طبقه‌بندی تشکیل‌شده توسط افراد مختلف، تفاوت وجود داشته باشد و موجب به‌دست‌آمدن نتایج گوناگون شود. همچنین زمانی که از عمل‌گر عمومی‌سازی سلولی[5] در یک مجموعه‌داده استفاده شود، مجموعهٔ یادشده دیگر برای عمل طبقه‌بندی[6] مناسب نیست و منجر به مشکل پدیدهٔ اکتشاف داده[7] می‌شود [4]. در مقابل، عمل‌گر فرونشانی، مقدار شبه‌شناسه موردنظر را با مقدار گمشده[8] جایگزین می‌کند. مقدار گمشده که به‌طورمعمول با علامت "?" نمایش داده می‌شود، بیان‌گر ناشناخته‌بودن مقدار صفت موردنظر است. استفاده از عمل‌گر فرونشانی باید با دقت انجام گیرد؛ زیرا تعداد زیاد شبه‌شناسه‌هایی که مقدار گمشده دارند، موجب کاهش کیفیت مجموعه داده می‌شود [2].

در این مقاله الگوریتم جدیدی برای حفظ محرمانگی طبق چهارچوب معرفی‌شده در [20] معرفی می‌شود. در روش پیشنهادی از عمل‌گر فرونشانی استفاده می‌شود، درنتیجه مشکلات یادشده را برای عمومی‌سازی به همراه نخواهد داشت. این روش از درخت تصمیمی که از روی داده‌های خام ایجاد می‌شود، به‌منظور انتخاب شبه‌شناسه‌ها برای فرونشانی استفاده می‌کند و مجموعه‌داده گمنام‌شده توسط آن برای عمل طبقه‌بندی مناسب است. نتایج به‌دست‌آمده از آزمایش‌ها نشان می‌دهد که دقت طبقه‌بندی در این حالت، تفاوت اندکی با طبقه‌بندی روی مجموعه داده اولیه دارد.

در ادامه ساختار مقاله بر این اساس ارائه می‌شود: در بخش دوم به‌اختصار پیشینهٔ پژوهش درخصوص انتشار داده با حفظ محرمانگی ارائه می‌شود. در بخش سوم مدل‌های حفظ محرمانگی و روابط ریاضی مورد نیاز برای معرفی الگوریتم آورده شده است. در بخش چهارم مقاله، الگوریتم پیشنهادی تشریح می‌شود. بخش پنجم نتایج پیاده‌سازی و ارزیابی طرح گمنام‌سازی پیشنهادی ارائه و مورد نقد و بررسی واقع و سرانجام در بخش ششم، مقاله جمع‌بندی و نتیجه‌گیری می‌شود.

۲- پیشینهٔ پژوهش

دسته‌ای از پژوهش‌های انجام‌شده در زمینهٔ PPDP به‌دنبال حفظ کیفیت داده‌های گمنام‌شده برای انجام یک عمل داده‌کاوی معین هستند. به بیان دیگر، تلاش می‌شود پردازش رکوردها به‌گونه‌ای انجام گیرد که مجموعه‌دادهٔ گمنام‌شده

[1] Quasi Identifier
[2] Generalization
[3] Suppression
[4] Taxonomy Tree
[5] Cell Generalization
[6] Classification
[7] Data Exploration
[8] Missing Value

علاوه‌بر این‌که یک مدل حفظ محرمانگی را ارضا می‌کند، برای استفاده در عمل داده‌کاوی خاصی بهینه شده باشد.

طبقه‌بندی یکی از عملیات اصلی داده‌کاوی محسوب می‌شود. به همین دلیل پژوهش‌های بسیاری برای کاهش اختلاف نتیجه طبقه‌بندی‌کردن روی داده‌های خام و گمنام، انجام شده است. پژوهش ارایه‌شده در [21] را می‌توان از نخستین طرح‌های کاربردی در این زمینه به‌شمار آورد. این روش برای هر عمل عمومی‌سازی G، میزان محرمانگی که بر اثر G به‌دست می‌آید (P) و مقدار اطلاعاتی را که از دست می‌رود (I) محاسبه می‌کند. تا زمانی که مدل k-anonymity در کل مجموعه‌داده برقرار نشده است، عملی که معیار $\frac{I(G)}{P(G)}$ را کمینه کند، روی داده‌ها اعمال می‌شود.

روش معرفی‌شده در [22] روند گمنام‌سازی را به‌صورت بالابه‌پایین انجام می‌دهد؛ به‌عبارت دیگر کار را از حالتی که تمام شبه‌شناسه‌ها در عمومی‌ترین حالت خود قرار دارند، آغاز می‌کند و با استفاده از عمل‌گر خاص‌سازی در هر مرحله، مقدار یک شبه‌شناسه را یک گره در درخت طبقه‌بندی پایین می‌آورد (عکس عمل عمومی‌سازی). در این روش نیز انتخاب شبه‌شناسه‌ها برای خاص‌سازی، با توجه به میزان گمنامی حاصل‌شده و اطلاعات ازدست‌رفته انجام می‌شود. اجرای الگوریتم تا زمانی که خاص‌سازی بیشتر موجب نقض مدل محرمانگی نشود، ادامه می‌یابد. این روش نیازی به درخت طبقه‌بندی برای شبه‌شناسه‌های عددی ندارد و درخت مورد نیاز این شبه‌شناسه‌ها طی اجرای الگوریتم ساخته می‌شود. این ایده در [10] بهبود بیشتری پیدا کرده است و مطابق آن اگر درخت طبقه‌بندی برای شبه‌شناسه‌ای موجود نباشد از عمل‌گر فرونشانی استفاده می‌شود. این روش از فرونشانی سراسری استفاده می‌کند؛ یعنی وقتی که فرونشانی بر شبه‌شناسه q با مقدار x اعمال می‌شود، q در تمام رکوردهایی که در آن‌ها مقدار x دارد، فرونشانده می‌شود.

در روش kACTUS [23] درخت طبقه‌بندی از فرآیند گمنام‌سازی حذف شده است و فقط از عمل‌گر فرونشانی استفاده می‌شود. این روش از فرونشانی محلی استفاده می‌کند که باعث می‌شود، فرونشاندن شبه‌شناسه‌ای از یک رکورد، تأثیری بر سایر رکوردها نداشته باشد و درنتیجه کیفیت مجموعه‌داده کمتر کاهش می‌یابد. شکل (۲) نحوه کارکرد این روش را برای برقراری k-anonymity با پارامتر k=100 نشان می‌دهد. در قسمت الف، صفت D مورد پردازش قرار می‌گیرد. از میان فرزندان این گره، فقط یکی بیش از صد رکورد دارد و دو گره دیگر درمجموع ۶۵ رکورد دارند؛ بنابراین ۳۵ رکورد از طبقهٔ A برای ایجاد دستهٔ صدتایی نگه داشته می‌شود و ۱۴۵ رکورد دیگر با شبه‌شناسه‌های {A='a3', B='b1', C='c1', D='d2'} به مجموعه رکوردهای گمنام‌شده اضافه می‌شوند. در قسمت ب، صفت E مورد بررسی قرار می‌گیرد. به‌دلیل این‌که هیچ‌کدام از فرزندان آن بیش از صد رکورد ندارند، پس از فرونشانی E با هم ادغام می‌شوند (قسمت ج) و هیچ رکوردی به مجموعه رکوردهای گمنام‌شده اضافه نمی‌شود. این روند تا پیمایش کل درخت ادامه می‌یابد.

هرچند دقت طبقه‌بندی روی داده‌های گمنام‌شده توسط این روش در حدی مطلوب است، اما به‌دلیل این‌که مجموعه‌داده را طبق مدل محرمانگی k-anonymity گمنام می‌کند، رکوردهای منتشرشده در مقابل حملاتی از قبیل حمله هم‌نوعی و حملهٔ دانش پیش‌زمینه آسیب‌پذیر هستند [23] و [24].

به‌طوراصولی دست‌یابی به محرمانگی مطلق غیر ممکن است؛ یعنی نمی‌توان کاری کرد که پس از انتشار مجموعه‌داده، مهاجم هیچ‌گونه اطلاعاتی از مالکان رکورد به‌دست نیاورد. به همین دلیل با توجه به کاربرد، مدل‌های مختلفی برای سطوح متفاوت از حفظ محرمانگی پیشنهاد شده است که هرکدام در مقابله با حملات معینی کارآمد هستند. در جدول (۱) مقایسه‌ای بین عملگرهای گمنامی انجام گرفته است.

اگر در زمان گمنام‌سازی از کاربردی که داده‌ها در آن استفاده خواهند شد، آگاه باشیم، می‌توانیم داده‌هایی متناسب با آن کاربرد ایجاد کنیم. طبقه‌بندی یکی از عملیات مهم داده‌کاوی است. به همین دلیل انتشار داده‌هایی که کیفیت مناسب برای طبقه‌بندی داشته باشند، بسیار بااهمیت است. در این شرایط ملاک باکیفیت‌بودن داده‌ها، تشابه نتایج طبقه‌بندی روی داده‌های خام و گمنام است.

در این مقاله ضمن استفاده از ایدهٔ فرونشانی شبه‌شناسه‌های با بهره اطلاعاتی کمتر، به جای گمنام‌سازی طبق k-anonymity از مدل احتمالاتی ارائه‌شده در [4] استفاده می‌شود که در مقایسه با مدل شناخته‌شده Differential privacy [7] به عملکرد خوبی در حفظ سودمندی و محرمانگی داده‌ها دست یافته است.

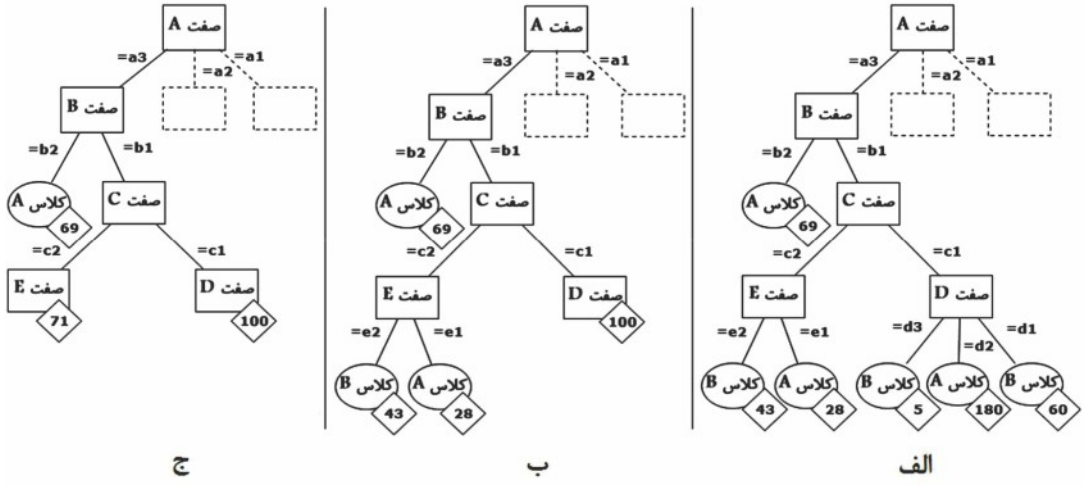
۳- مدل‌های حفظ محرمانگی در انتشار داده

پیش از تلاش برای حفظ محرمانگی باید مشخص کنیم چه زمانی این هدف برآورده می‌شود. تعریف ایده‌آل از حفظ

محرمانگی هنگام انتشار داده‌ها، در [12] ارائه شده است. بر طبق این تعریف، بعد از آن‌که داده‌ها در اختیار مهاجم قرار می‌گیرد، نسبت به زمانی که داده‌ها را در اختیار نداشته است، نباید هیچ دانشی درمورد مالکان رکورد به او اضافه شود. با درنظرگرفتن اطلاعاتی که ممکن است از قبل در اختیار مهاجم باشد، این سطح از محرمانگی قابل دست‌یابی نیست [13]. به همین دلیل در بیش‌تر کارهایی که در حوزهٔ محرمانگی داده انجام می‌گیرد از مدل‌های تعدیل‌یافته‌تری استفاده می‌شود که در ادامه به شرح آن‌ها پرداخته می‌شود.

۱ -۳- مدل پیونددهی رکورد

در حملات پیونددهی رکورد، مهاجم شبه‌شناسه‌های قربانی را با رکوردهای انتشاریافته تطبیق می‌دهد؛ بعد از انجام این کار دسته‌ای کوچک از رکوردها باقی می‌ماند، که ممکن است، متعلق به قربانی باشد.



(شکل-۲): روند اجرای الگوریتم kACTUS

(Figure-2): kACTUS Algorithm operation flow

(جدول-۱): مقایسه عملگرهای گمنام‌سازی

(Table-1): A comparison on anonymity operators

نام روش	نحوه عملکرد	رکوردها معرف یک نمونه واقعی هستند؟	قابلیت استفاده از روش‌های استاندارد داده‌کاوی	میزان تغییر مقدار صفات
عمومی سازی	تعمیم مقدار صفت به بازه‌ای بزرگتر	بله	بله	نسبتا کم
فرونشانی	حذف مقدار صفت	بله	بله	نسبتا زیاد
تجزیه	جدا کردن صفات حساس از شبه‌شناسه‌ها	تا حدودی	خیر	بدون تغییر
برش دهی	جدا کردن صفات مرتبط به هم و تعویض مقدار آن‌ها در رکوردهای مختلف	خیر	بله	بدون تغییر
اضافه‌کردن نوفه	تولید داده‌های جدید به وسیله افزودن یک مقدار تصادفی به داده‌های اولیه	خیر	بله	نسبتا کم
تولید داده‌های مصنوعی	تولید داده‌های جدید با توجه به خصوصیات آماری داده‌های اولیه	خیر	بله	به‌طور کامل عوض می‌شود

پیونددهی رکورد، زمانی رخ می‌دهد که مهاجم موفق شود رکورد مربوط به قربانی را به‌طور یکتا شناسایی کند [1]. به‌عنوان مثال در جدول (۳) صفات تحصیلات، جنسیت و سن به‌عنوان شبه‌شناسه‌ها و حقوق به‌عنوان صفت حساس درنظر گرفته شده است. پس از انتشار این جدول اگر مهاجم به دنبال رکوردی با شبه‌شناسه‌های <کارشناسی، مرد، ۲۹> باشد، به‌راحتی می‌تواند آن را پیدا کند؛ زیرا تنها یک رکورد متناظر وجود دارد و درنتیجه حمله پیونددهی رکورد با موفقیت انجام می‌شود.

روش k-anonymity [5] برای جلوگیری از چنین حملاتی، مجموعه‌داده را به دسته‌های هم‌ارزی با اندازه k، تقسیم می‌کند. در این حالت به‌ازای هر مقدار ممکن از

شبه‌شناسه‌ها، یا هیچ رکوردی در مجموعه‌داده وجود ندارد و یا k رکورد (و بیشتر) موجود است، درنتیجه مهاجم بیش‌تر با درجه اطمینان 1/k می‌تواند رکورد قربانی را شناسایی کند. این روش در قسمت «ب» از جدول (۳) برای گمنام‌سازی مورد استفاده قرار گرفته است. همان‌طور که دیده می‌شود، هر یک از رکوردهای این جدول دستِ‌کم با یک رکورد دیگر شبه‌شناسه‌های یکسانی دارد. به چنین جدولی 2-anonymous می‌گویند.

در برخی پایگاه داده‌ها مانند فهرست بیماران یک درمانگاه، ممکن است n رکورد متعلق به یک فرد باشد. در چنین مواردی پس از اعمال k-anonymity این رکوردها در یک دستهٔ هم‌ارزی قرار می‌گیرند. در این حالت مهاجم با درجهٔ اطمینان n/k رکورد قربانی را به‌دست می‌آورد که با نیاز تعریف‌شده در k-anonymity مطابقت ندارد. برای حل این مشکل می‌توان از مفهومی عمومی‌تر به نام (X,Y)-anonymity [14] استفاده کرد. در این روش X و Y دو زیرمجموعه از صفات مجموعه‌داده هستند که اشتراکی با هم ندارند. برای برآورده‌شدن محرمانگی در (X,Y)-anonymity، هر مقدار در X باید دستِ‌کم با k مقدار متمایز از Y ارتباط داشته باشد. اگر Y را شناسه و X را مجموعه شبه‌شناسه‌ها درنظر بگیریم، با انتخاب مقدار مناسب برای k می‌توان بیشینه تعداد رکوردهای متعلق به یک فرد در یک دسته هم‌ارزی را کنترل کرد.

(جدول-۲): مثالی ازگمنام سازی k-anonymity با K=۲
(Table-2): A example of k-anonymity with k=2

الف) مجموعه‌داده خام

تحصیلات	جنسیت	سن	حقوق
کارشناسی ارشد	مرد	۲۶	۱۰۰۰
کارشناسی	مرد	۳۰	۱۰۵۰
کاردانی	مرد	۳۴	۹۰۰
دکترا	مرد	۲۸	۱۱۰۰
کارشناسی	زن	۲۵	۹۵۰
کارشناسی	زن	۲۷	۹۵۰
کاردانی	زن	۲۶	۹۵۰

ب) جدول 2-anonymous

تحصیلات	جنسیت	سن	حقوق
ارشد و بالاتر	مرد	[۲۵-۳۰)	۱۰۰۰
پایین‌تر از ارشد	مرد	[۳۰-۳۵)	۱۰۵۰
پایین‌تر از ارشد	مرد	[۳۰-۳۵)	۹۰۰
ارشد و بالاتر	مرد	[۲۵-۳۰)	۱۱۰۰
پایین‌تر از ارشد	زن	[۲۵-۳۰)	۹۵۰
پایین‌تر از ارشد	زن	[۲۵-۳۰)	۹۵۰
پایین‌تر از ارشد	زن	[۲۵-۳۰)	۹۵۰

۲-۳- مدل پیونددهی صفت

همان‌طورکه گفته شد، در حمله پیونددهی رکورد مهاجم در صدد تشخیص رکورد مربوط به قربانی است؛ اما در بیش‌تر موارد به‌دست‌آوردن صفت حساس قربانی، مهاجم را به هدف خود می‌رساند و نیازی به شناسایی رکورد نیست [1]. برای مثال در جدول (۲) مقدار صفت حساس در سه رکورد آخر برابر است و مهاجم با دراختیارداشتن شبه‌شناسه‌های <کاردانی، زن، ۲۶> بدون آن‌که رکورد قربانی را شناسایی کند، از میزان حقوق وی مطلع می‌شود. این مشکل ناشی از عدم تنوع در مقدار صفت حساس در یک دستهٔ هم‌ارزی است. l-diversity [15] یکی از روش‌های مورد استفاده برای مقابله با حمله پیونددهی صفت است. برای تأمین l-diversity در یک جدول، در هر دستهٔ هم‌ارزی باید دستِ‌کم l مقدار مختلف برای صفت حساس موجود باشد. این خاصیت سبب می‌شود تا مهاجم بعد از پیداکردن دسته‌ای که قربانی به آن تعلق دارد، با مقادیر یکسانی از صفت حساس مواجه نشود. اگر در یک دسته هم‌ارزی l مقدار متفاوت ولی نزدیک به هم وجود داشته باشد، باز هم مشکل عدم تنوع بروز پیدا می‌کند.

برای مثال قسمت «ب» از جدول (۳) مدل 3-diversity را ارضا می‌کند؛ اما مقادیر صفت حساس در نخستین گروه هم‌ارزی بسیار نزدیک به هم هستند. برای رفع این نقص، در [16] روشی به نام t-Closeness ارائه شد که بر طبق آن تفاوت توزیع صفت حساس در هر دسته هم‌ارزی نسبت به توزیع آن در کل جدول باید از حد آستانه t کمتر باشد. این روش از فاصله Earth Mover [17] برای سنجش تفاوت بین توزیع‌ها استفاده می‌کند. در جدول (۴) نتیجه گمنام‌سازی یک مجموعه‌داده به روش‌های l-diversity و t-Closeness مقایسه شده است.

برقرارکردن t-Closeness در یک جدول، نیازمند استفاده زیاد از عملگرهای گمنام‌سازی است که باعث کاهش قابل توجه در کیفیت داده‌ها می‌شود. برای برطرف‌کردن این ضعف تکنیکی انعطاف‌پذیرتر به نام(n,t)-Closeness ارائه شده است [18]. اگر به‌ازای هر گروه هم ارزی E1 مجموعه‌ای از رکوردها با نام E2 موجود باشد که:

- بیش از n رکورد داشته باشد.
- E1 زیرمجموعه آن باشد.
- فاصله توزیع صفت حساس در E1 و E2 کمتر از t باشد.

آن‌گاه آن مجموعه‌داده (n,t)-Closeness را ارضا می‌کند. در این روش هرچه n کوچک‌تر درنظر گرفته شود، کیفیت داده‌ها کمتر لطمه می‌بیند و در حالتی که n برابر با تعداد کل رکوردها باشد، معادل روش t-Closeness است.

(جدول-۳): مقایسه روش‌های l-diversity و t-Closeness

(Table-3): A comparison between l-diversity and t-closeness techniques

الف) مجموعه‌داده خام

	کدپستی	سن	حقوق
۱	۴۷۶۷۷	۲۹	۳۰۰۰
۲	۴۷۶۰۲	۲۲	۴۰۰۰
۳	۴۷۶۷۸	۲۷	۵۰۰۰
۴	۴۷۹۰۵	۴۳	۶۰۰۰
۵	۴۷۹۰۹	۵۲	۱۱۰۰۰
۶	۴۷۹۰۶	۴۷	۸۰۰۰
۷	۴۷۶۰۵	۳۰	۷۰۰۰
۸	۴۷۶۷۳	۳۶	۹۰۰۰
۹	۴۷۶۰۷	۳۲	۱۰۰۰۰

ب) 3-diverse

	کدپستی	سن	حقوق
۱	۴۷۶**	۲*	۳۰۰۰
۲	۴۷۶**	۲*	۴۰۰۰
۳	۴۷۶**	۲*	۵۰۰۰
۴	۴۷۹**	≥ ۴۰	۶۰۰۰
۵	۴۷۹**	≥ ۴۰	۱۱۰۰۰
۶	۴۷۹**	≥ ۴۰	۸۰۰۰
۷	۴۷۶**	۳*	۷۰۰۰
۸	۴۷۶**	۳*	۹۰۰۰
۹	۴۷۶**	۳*	۱۰۰۰۰

پ) 0.167–closeness

	کدپستی	سن	حقوق
۱	۴۷۶۷*	≤ ۴۰	۳۰۰۰
۳	۴۷۶۷*	≤ ۴۰	۵۰۰۰
۸	۴۷۶۷*	≤ ۴۰	۹۰۰۰
۴	۴۷۹۰*	≥ ۴۰	۶۰۰۰
۵	۴۷۹۰*	≥ ۴۰	۱۱۰۰۰
۶	۴۷۹۰*	≥ ۴۰	۸۰۰۰
۲	۴۷۶۰*	≤ ۴۰	۴۰۰۰
۷	۴۷۶۰*	≤ ۴۰	۷۰۰۰
۹	۴۷۶۰*	≤ ۴۰	۱۰۰۰۰

(جدول-۴): روش δ-Presence

(Table-4): δ-Presence technique

P

	نام	کدپستی	سن	ملیت
a	Alice	۴۷۹۰۶	۳۵	ایالات‌متحده
b	Bob	۴۷۹۰۳	۵۹	کانادا
c	Chris	۴۷۹۰۶	۴۲	ایالات‌متحده
d	Dirk	۴۷۶۳۰	۱۸	برزیل
e	Eunice	۴۷۶۳۰	۲۲	برزیل
f	Frank	۴۷۶۳۳	۶۳	پرو
g	Gail	۴۸۹۷۳	۳۳	اسپانیا
h	Harry	۴۸۹۷۲	۴۷	بلغارستان
i	Iris	۴۸۹۷۰	۵۲	فرانسه

T^*

	کدپستی	سن	ملیت
b	۴۷*	*	امریکا
c	۴۷*	*	امریکا
f	۴۷*	*	امریکا
h	۴۸*	*	اروپا
i	۴۸*	*	اروپا

۱-۲-۳- مدل پیونددهی جداول

در برخی شرایط، صرف تشخیص این‌که رکورد قربانی در مجموعه‌داده انتشار یافته وجود دارد یا خیر، می‌تواند نقض محرمانگی تلقی شود. به‌عنوان نمونه زمانی که یک بیمارستان داده‌های مربوط به اشخاصی را که دچار بیماری خاصی هستند، منتشر می‌کند، پی‌بردن به هویت هر یک از افراد، نوع بیماری وی را آشکار می‌سازد. در زمینه مدل پیونددهی جداول نسبت به مدل پیونددهی رکورد و صفت، کارهای کمتری انجام شده که معروف‌ترین آن‌ها δ-Presence است [19].

طبق این مدل، احتمال شناسایی رکورد قربانی در مجموعه‌داده انتشاریافته باید به سقف δ٪ محدود شود. این مفهوم با مثالی که در جدول (۴) آورده شده است، بیشتر شرح داده می‌شود. فرض می‌شود جدول P از قبل در دسترس عموم قرار دارد. جدول T زیرمجموعه‌ای از رکوردهای P است که مقدار صفت حساس در آن‌ها یکسان است. T^* نسخه گمنام‌سازی شدهٔ T است که پس از انتشار آن مهاجم نباید با احتمال بیشتر از δ قادر باشد که بگوید یکی از رکوردهای P در T^* وجود دارد. احتمال این‌که هریک از رکوردهای a-f در T^* حضور داشته باشند، برابر با $\frac{|\{b,c,f\}|}{|\{a,b,c,d,e,f\}|}=\frac{3}{6}$ است. این احتمال برای رکوردهای g-i برابر با $\frac{|\{h,i\}|}{|\{g,h,i\}|}=\frac{2}{3}$ است. بنابراین مهاجم حداکثر با احتمال $\frac{2}{3}$ می‌تواند از وجود یک رکورد دلخواه از P در جدول T^* اطمینان داشته باشد.

همان‌طور که گفته شد، در روش δ-Presence مهاجم حداکثر δ٪ از وجود رکورد قربانی در جدول منتشرشده اطمینان دارد؛ بنابراین احتمال این‌که بتواند با موفقیت حمله پیونددهی رکورد یا صفت را انجام دهد، نیز حداکثر δ٪ است؛ لذا این روش برای مقابله با حملات پیونددهی رکورد و صفت نیز کارایی دارد. عیب عمده روش δ-Presence این است که در محاسبهٔ احتمالات، فرض می‌شود منتشرکنندهٔ داده و مهاجم به یک جدول خارجی مشترک (مجموعه‌داده P) دسترسی دارند. درحالی‌که چنین فرضی در دنیای واقعی زیاد محتمل نیست.

۲-۲-۳- مدل احتمالاتی

در مدل‌های احتمالاتی هدف جلوگیری از پیونددهی رکورد قربانی به یک رکورد، صفت و یا جدول خاص نیست؛ بلکه بیش‌تر این مدل‌ها تلاش می‌کنند تغییر دانش مهاجم، قبل و بعد از مشاهدهٔ داده منتشر شده، به مقدار کوچکی محدود شود.

طرح ارائه‌شده در [13] یکی از سخت‌گیرانه‌ترین مدل‌های حفظ محرمانگی است. بر طبق این مدل، انتشار یا عدم انتشار رکورد قربانی نباید در نتیجه بررسی‌هایی که مهاجم روی مجموعه‌داده انجام می‌دهد، تأثیر قابل توجهی بگذارد. به بیان ریاضی، اگر بررسی‌های مهاجم توسط تابع تصادفی F مدل شود، به‌ازای هر مجموعه‌داده D و D' که در وجود یک رکورد اختلاف دارند، باید رابطه زیر برقرار باشد:

$$\forall S \in Range(F)\ (\Pr[F(D) = S] \le e^{\varepsilon} \times \Pr[F(D') = S])$$

که S صفت حساس و ε پارامتری است که هرچه کوچک‌تر انتخاب شود، میزان تأثیر یک رکورد بر نتیجه بررسی‌ها کمتر می‌شود. Differential Privacy تعریف محکمی از محرمانگی ارائه می‌دهد که تفاوتی میان صفات حساس و سایر صفات قائل نیست. دست‌یابی به چنین هدفی مستلزم اضافه‌کردن نوفه فراوان به داده‌ها است که موجب کاهش کیفیت مجموعه‌داده می‌شود. طرح ارایه‌شده در [20] با درنظرداشتن این مشکل مدلی تعدیل‌یافته‌تر معرفی کرده است: اطمینانی که مهاجم پس از مشاهدهٔ مجموعه‌دادهٔ گمنام‌شده نسبت به صفت حساس قربانی به‌دست می‌آورد، نباید بیشتر از یک حدس تصادفی باشد. ازآنجاکه طرح ارائه‌شده در این مقاله بر پایه همین مدل بنا شده است، در ادامه با جزییات کامل‌تری تشریح می‌شود.

مدل حفظ محرمانگی معرفی‌شده در [20] (AGF) بر اساس دو مفهوم اطمینان مورد انتظار (EC) و اطمینان مشاهده‌شده (OC) شکل گرفته است. اطمینان مشاهده‌شده یک رکورد، بیان‌گر این است که مهاجم با چه احتمالی می‌تواند صفت حساس آن را به قربانی نسبت دهد. اگر رکورد قربانی (v) عضو دسته هم‌ارزی به‌اندازه |E(v)| باشد و تعداد تکرار صفت حساس v در این دسته هم‌ارزی را #S(v) بنامیم، از دید مهاجم احتمال آن‌که صفت حساس v متعلق به قربانی باشد، برابر $\frac{\#S(v)}{|E(v)|}$ است. به‌دلیل این‌که در این روش در ابتدا از رکوردها با نرخ β نمونه‌برداری می‌شود، باید احتمال گزینش رکورد قربانی در مرحله نمونه‌برداری نیز مدنظر قرار بگیرد. همان‌طورکه رابطه (۱) نشان می‌دهد از ضرب این دو احتمال، اطمینان مشاهده‌شده رکورد r به‌دست می‌آید.

$$ObservedConfidence(r) = \beta \times \frac{\#S(r)}{|E(r)|} \qquad (۱)$$

هرچه اطمینان مشاهده‌شده رکوردها کمتر باشد، مهاجم با اطمینان کمتری می‌تواند بین صفات حساس و شناسه‌های صریح رابطه برقرار کند؛ بنابراین باید مقدار آن را به سقف معینی محدود کرد. فرض می‌شود مجموعه‌داده شامل n رکورد باشد و D_0 نیز مجموعه‌داده‌ای با اندازه n است که به‌صورت تصادفی از فضای نمونه رکوردها ایجاد شده است. برطبق این مدل، اطمینان مشاهده‌شدهٔ یک رکورد در مجموعه‌داده گمنام‌شده، نباید از احتمال دیده‌شدن آن رکورد در D_0 بیشتر باشد. این احتمال اطمینان مورد انتظار نامیده می‌شود. با فرض این‌که مجموعه‌داده شامل n رکورد باشد، احتمال وجود رکورد r در D_0 برابر با احتمال مکمل "۰" پیروزی در n آزمون با احتمال پیروزی Pr(r) است و با استفاده از توزیع دوجمله‌ای طبق رابطه (۲) محاسبه می‌شود:

$$ExpectedConfidence(r) = 1 - f(0; n, \Pr(r)) = 1 - (1 - \Pr(r)^2) \qquad (۲)$$

که Pr(r) با فرض مستقل‌بودن صفات از یکدیگر به‌وسیله ضرب احتمال شبه‌شناسه‌ها در احتمال صفت حساس به‌دست می‌آید:

$$\Pr(r) = (\prod_i \Pr(q_i)) \times \Pr(s) \qquad (۳)$$

با توجه به رابطه (۱) و (۲)، برای تأمین محرمانگی طبق قاعده ذکرشده، به‌ازای هر رکورد موجود در مجموعه‌داده، باید رابطه (۴) برقرار باشد:

$$\text{ObservedConfidence(r)} \le \text{ExpectedConfidence(r)} \qquad (۴)$$

وقتی رابطه (۴) برای رکوردی برقرار نبود، یکی از شبه‌شناسه‌ها باید برای عمومی‌سازی انتخاب شود. این کار بر اساس معیاری به نام Left Domain Size(LDS) انجام می‌گیرد. LDS شبه‌شناسه‌ای با مقدار v، تعداد گره‌هایی است که در درخت طبقه‌بندی در سطح برابر یا بالاتر از v قرار دارد. شبه‌شناسه‌ای که مقدارش کمترین LDS را دارد برای عمومی‌سازی برگزیده خواهد شد.

در حین عملیات گمنام‌سازی، ممکن است، عمل‌گر عمومی‌سازی روی تعداد زیادی از شبه‌شناسه‌های یک رکورد اعمال شود. چنین رکوردی کیفیت داده‌های خروجی را کاهش می‌دهد. به همین منظور از معیار تخریب برای سنجش میزان کیفیت رکوردها استفاده می‌شود و رکوردهایی که میزان تخریب آن‌ها از حدِّ از پیش تعریف‌شده δ′ بیشتر باشد از مجموعه خروجی حذف می‌شوند. میزان تخریب رکورد r با

شبه‌شناسه‌های q_0 تا q_d از رابطه (۵) به‌دست می‌آید.

$$\delta = \frac{1}{d}\sum_{i=1}^{d}\left(1 - \frac{Currentlevel\ of\ q'_i}{Maximal\ level\ in\ q_i}\right) \qquad (۵)$$

اگر تخریب "۰" باشد به معنی این است که تمام شبه‌شناسه‌های رکورد، مقدار اصلی خود را حفظ کرده‌اند و تخریب "۱" به معنی عمومی‌سازی تمام شبه‌شناسه‌ها تا گره ریشه است.

در جدول (۵) مدل‌های مطرح محرمانگی بر اساس ویژگی‌ها و نقاط ضعف هرکدام مقایسه شده‌اند.

(جدول-۵): مقایسه مدل‌های حفظ محرمانگی

(Table-5): A Comparison between Privacy models

نام مدل	رابطه‌ای که در مجموعه‌داده باید برقرار شود	ویژگی ها	نقطه ضعف
k-anonymity	گروه‌بندی رکوردها در دسته‌های هم ارزی به اندازه k	محافظت در برابر پیونددهی رکورد، تخریب کم	چند رکورد از یک فرد در یک دسته هم‌ارزی قرار می‌گیرد
(X,Y)-anonymity	هر مقدار در زیرمجموعه X حداقل با k مقدار متمایز از Y ارتباط داشته باشد	رفع ضعف ذکر شده برای k-anonymity	مقدار صفت حساس در اعضای یک دسته هم ارزی می‌تواند یکسان باشد
l-diversity	وجود l مقدار متنوع از صفت حساس در هر دسته هم‌ارزی	محافظت در برابر پیونددهی صفت، تخریب کم	l مقدار متنوع می‌توانند نزدیک به هم باشند
t-closeness	تفاوت توزیع صفت حساس در هر دسته هم‌ارزی نسبت به توزیع آن در کل جدول از t کمتر باشد	رفع ضعف ذکر شده برای l-diversity	تخریب زیاد
(n,t)-closeness	هر گروه هم ارزی ابرمجموعه‌ای با بیش از n رکورد داشته باشد که فاصله توزیع صفت حساس در آن دو کمتر از t است.	رفع ضعف ذکر شده برای t-closeness	امکان وجود چند صفت حساس در مجموعه‌داده درنظر گرفته نشده است
δ-presence	احتمال وجود رکورد قربانی در مجموعه‌داده انتشار یافته به δ٪ محدود شود	محافظت در برابر پیونددهی رکورد، صفت، جدول	فرض دسترسی منتشرکننده داده و مهاجم به یک جدول خارجی مشترک (P)
differential privacy	انتشار یا عدم انتشار یک رکورد، در نتیجه بررسی‌هایی که روی مجموعه‌داده انجام می‌گیرد تاثیر قابل توجهی نداشته باشد	نزدیک‌ترین مدل به تعریف ایده‌آل از محرمانگی	تخریب زیاد
AGF	اطمینان مشاهده شده رکوردها کوچکتر از اطمینان مورد انتظار باشد	رفع ضعف ذکر شده برای differential privacy	فرض مستقل بودن مقدار صفات

۴- طرح الگوریتم پیشنهادی

در این بخش از مقاله، الگوریتم جدیدی برای حفظ محرمانگی در انتشار داده‌ها ارایه می‌شود که طبق چهارچوب معرفی‌شده در [20] طراحی و خروجی آن برای استفاده در طبقه‌بندی، بهینه شده است. یکی از مشکلات موجود در روش [20]، نیاز به حضور فردی خبره به‌منظور تشکیل درخت طبقه‌بندی برای هر شبه‌شناسه است. ممکن است افراد مختلف، درخت‌های طبقه‌بندی مختلفی برای صفات در نظر بگیرند که این امر موجب به‌دست‌آمدن نتایج گوناگون می‌شود. علاوه‌براین مجموعه‌داده حاصل از اجرای این روش عام‌منظوره است، یعنی در کاربردهای مختلف داده‌کاوی می‌توان از آن استفاده کرد؛ ولی برای استفاده در کاربرد خاصی بهینه نیست. هنگامی‌که داده‌ها با هدف استفاده در طبقه‌بندی منتشر می‌شوند، می‌توان با تغییر در نحوه گمنام‌سازی، به نتایج بهتری دست یافت.

به‌دلیل این‌که [20] از عمل‌گر عمومی‌سازی سلولی برای گمنام‌سازی استفاده می‌کند، در نحوه طبقه‌بندی رکوردهای خروجی ابهام به‌وجود می‌آید [1]. به‌عنوان نمونه اگر مقدار صفت «ملیت» در برخی رکوردها از ایران به آسیا تبدیل شود، الگوریتم طبقه‌بندی که به‌وسیله داده‌های گمنام‌شده آموزش دیده و قرار است رکورد جدیدی را طبقه‌بندی کند، باید از رابطه آسیا و ایران اطلاع داشته باشد؛ زیرا در داده‌های واقعی صفت ملیت هیچ‌گاه مقدار آسیا ندارد. همچنین ممکن است، در بعضی موارد قوانین طبقه‌بندی متناقضی به وجود بیاید. به‌عنوان‌مثال اگر صفت «ملیت» مقدار ایران داشت، آن رکورد متعلق به طبقهٔ یک، و اگر مقدار آسیا داشت، متعلق به طبقهٔ دو است.

در روش پیشنهادی از عمل‌گر فرونشانی استفاده می‌شود؛ در نتیجه مشکلات ذکرشده برای عمومی‌سازی را به همراه نخواهد داشت. در این روش، انتخاب شبه‌شناسه‌ها برای فرونشانی به‌وسیله درخت تصمیمی که از روی داده‌های خام ایجاد می‌شود، صورت می‌پذیرد.

در روش پیشنهادی، انتخاب شبه‌شناسه‌ها برای فرونشانی، به‌وسیله درخت تصمیم انجام می‌شود و در این کار اولویت با شبه‌شناسه‌هایی است که در پایین درخت قرار دارند. بدین منظور ابتدا درخت تصمیم از روی داده‌ها ساخته می‌شود؛ سپس اطمینان مشاهده‌شده و اطمینان مورد انتظار برای هر یک از رکوردها محاسبه شده و برقراربودن رابطه (۴) بررسی می‌شود. درصورتی‌که این رابطه برقرار نباشد، مسیر رکورد در درخت تصمیم مشخص شده و شبه‌شناسه‌ای که در انتهای مسیر قرار دارد، فرونشانده می‌شود. با فرونشانی شبه‌شناسه‌ها، نقاط تمایز رکوردها کمتر می‌شود و اندازه دسته‌های هم‌ارزی افزایش می‌یابد. با توجه به رابطه (۱) این امر منجر به کاهش اطمینان مشاهده‌شده رکوردها می‌شود و به‌تدریج قاعده محرمانگی برای تمام رکوردهای مجموعه‌داده برقرار می‌شود.

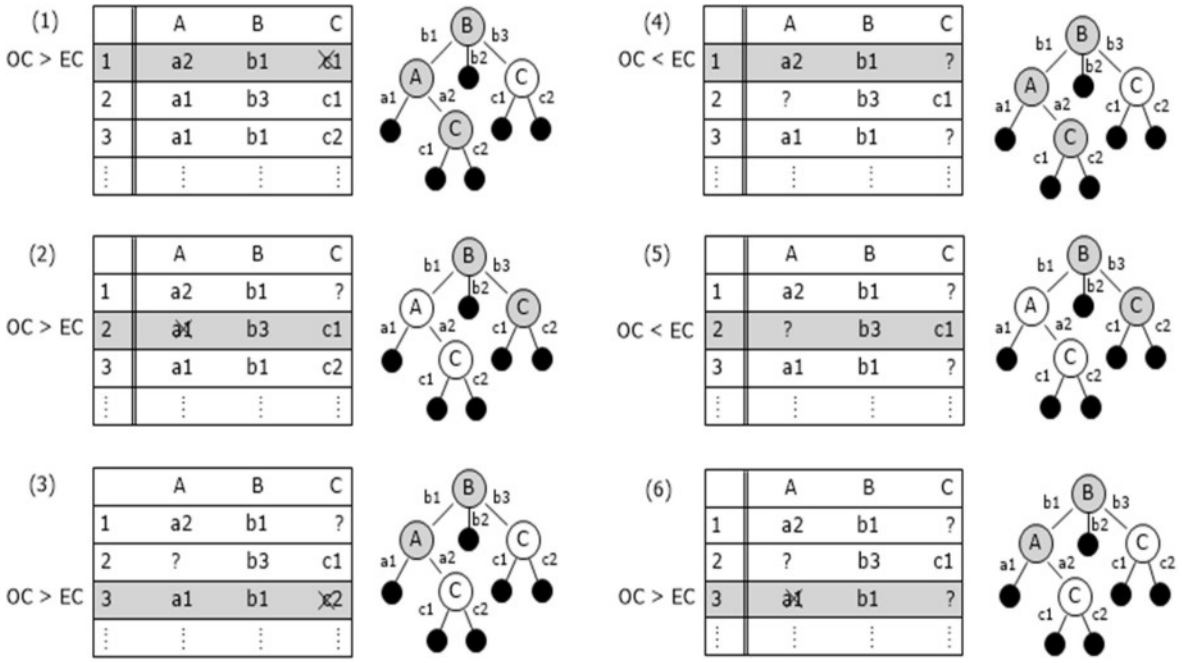
در شکل (۳) مثالی ساده از نحوه کارکرد روش پیشنهادی نمایش داده شده است. برای ساده‌سازی فقط سه رکورد نخست آورده شده است. مجموعه‌داده شامل شبه‌شناسه‌های A، B و C است. کار از پردازش رکورد نخست در قسمت (۱) آغاز می‌شود؛ فرض می‌شود که این رکورد رابطه (۴) را ارضا نمی‌کند؛ همان‌طور که در درخت متناظر دیده می‌شود، گره C آخرین گرهی است که در مسیر رکورد نخست، روی آن تصمیم‌گیری می‌شود؛ لذا مقدار این شبه‌شناسه باید حذف شود. رابطه OC<EC برای رکورد دوم هم برقرار نیست.

شبه‌شناسه A در تعیین طبقهٔ این رکورد نقشی ندارد، بنابراین فرونشانده و به‌دلیل مشابه در قسمت (۳) نیز مقدار شبه‌شناسه C حذف می‌شود.

بعد از این‌که تمام رکوردها یک دور پردازش شدند، در قسمت (۴) دوباره نوبت رکورد نخست فرا می‌رسد. به‌دلیل فرونشانی‌های انجام‌شده در دور نخست، اطمینان مشاهده‌شده (OC) رکوردهای نخست و دوم کاهش یافته و قاعده محرمانگی برای این دو رکورد برقرار شده است؛ بنابراین در قسمت‌های (۴) و (۵) هیچ‌گونه تغییری در مقدار شبه‌شناسه‌ها صورت نمی‌پذیرد. در قسمت (۶)، رابطه OC<EC برای رکورد سوم برقرار نیست. با توجه به این‌که مقدار C در دور قبل حذف شده بود، در این دور پایین‌ترین گره، یعنی A برای فرونشانی انتخاب می‌شود. این روال تا زمانی که رابطه (۴) برای تمام رکوردها برقرار شود، ادامه می‌یابد.

به‌دلیل این‌که روش پیشنهادی از فرونشانی استفاده می‌کند، تنها دو حالت برای شبه‌شناسه q_i از رکورد r متصور است: مقدار اصلی خود را حفظ کرده و یا مقدار گمشده گرفته است. میزان تخریب رکورد r با شبه‌شناسه‌های q_0 تا q_d از ۰ به‌دست می‌آید.

$$\delta = \frac{1}{d}\sum_{i=1}^{d} g(r,q_i)\,,\quad g(r,q_i) = \begin{cases} 0: q_i\ isn't\ missed\ from\ r \\ 1: q_i\ is\ missed\ from\ r \end{cases} \qquad (۶)$$



(شکل-۳): نحوه عملکرد روش پیشنهادی

(Figure-3): Operation of the proposed technique

الگوریتم ارایه‌شده در شکل (۴) روند گمنام‌سازی داده‌ها را نمایش می‌دهد. مجموعه‌داده (D)، مجموعه شبه‌شناسه‌ها (Q)، احتمال مقادیر مختلف برای هر شبه‌شناسه (P)، نرخ نمونه‌گیری (β) و حد تخریب مورد قبول (δ′) به‌عنوان ورودی به الگوریتم داده می‌شود و با اجرای الگوریتم مجموعه‌داده گمنام شده (D*) ایجاد می‌شود. مرحله یک شبه‌شناسه‌های عددی را بازه‌بندی و مقادیر این صفات را با بازه‌ای که در آن قرار دارند، تعویض می‌کند. طول بازه‌ها می‌تواند از حد پایین و بالای اعداد در هر صفت به‌دست آید. این کار به‌منظور کمترساختن پراکندگی مقادیر در صفات عددی، انجام می‌شود. مرحلهٔ دو درخت تصمیم را از روی مجموعه‌داده تشکیل می‌دهد. این درخت در مراحل بعد برای انتخاب صفتی که کمترین بهره اطلاعاتی را دارد، مورد استفاده قرار می‌گیرد. مرحلهٔ سه با نرخ β و به‌صورت تصادفی از مجموعه‌داده نمونه‌برداری می‌کند.

در مراحل ۵-۴ بررسی می‌شود که اگر اصل محرمانگی تعریف‌شده در رابطه (۴) برای تمامی رکوردها صادق باشد، اجرای الگوریتم خاتمه یابد؛ در غیر این صورت رکوردها برای پردازش بیشتر به مرحلهٔ شش فرستاده می‌شوند. مراحل ۷-۶ اطمینان مورد انتظار و اطمینان مشاهده‌شده رکورد r را از رابطه (۱) و رابطه (۲) محاسبه می‌کنند.

```
Anonymize (D, Q, P, β, δ')

Output: D*

 1: D* ← DiscretizeNumericQIs(D, Q)
 2: DT ← GenerateDecisionTree(D*)
 3: D* ← RandomSample(D*, β)
 4: WHILE all records in D* does not satisfy Eq. 3-4 DO
 5:     FOR each r ∈ D* DO
 6:  ec ← ExpectedConfidence(r, D*, Q, P)
 7:  oc ← ObservedConfidence(r, D*, Q)
 8:  IF ec < oc THEN
 9:      q ← QIWithLowestInfoGain(r, Q, DT)
10:            r ← Suppress (r, q)
11:         ELSE
12:            δ = Distortion(r, Q)
13:            IF δ > δ' THEN
14:                D* ← D* - r
15:            END IF
16:         END IF
17:    END FOR
18: END WHILE
19: RETURN D*
```

(شکل-۴): روند گمنام‌سازی داده‌ها
(Figure-4): Annonymizaton technique

اگر رکورد r هنوز رابطه (۴) را ارضا نمی‌کند، باید عمل‌گر فرونشانی روی یکی از شبه‌شناسه‌های آن اعمال شود. مرحلهٔ نه با استفاده از درخت تصمیم DT از میان شبه‌شناسه‌ها، صفت q که کمترین بهره اطلاعاتی را دارد، انتخاب می‌کند. نحوه این کار در الگوریتم ارایه‌شده در شکل (۵) شرح داده می‌شود. مرحلهٔ ده فرونشانی را روی شبه‌شناسه q از رکورد r اعمال می‌کند.

مرحلهٔ دوازده میزان تخریب رکورد را از طریق ∘صفر محاسبه می‌کند و در صورتی که از حد آستانه δ′ بیشتر باشد، رکورد مربوطه از مجموعه D* حذف می‌شود. مرحلهٔ نوزده مجموعه D* را که شامل رکوردهای گمنام شده است، در خروجی قرار می‌دهد.

نحوه انتخاب صفتی که باید فرونشانده شود در شکل (۵) بیان شده است. برای این‌که اعمال فرونشانی کمترین تأثیر را در نتیجهٔ طبقه‌بندی بگذارد، در هر مرحله شبه‌شناسه‌هایی که دورتر از ریشه درخت طبقه‌بندی هستند، فرونشانده می‌شوند. ورودی‌های الگوریتم رکورد r، مجموعه شبه‌شناسه‌ها Q و درخت تصمیم DT هستند و پس از اتمام کار شبه‌شناسه qi برگردانده می‌شود.

```
QIWithLowestInfoGain(r, Q, DT)

Output: qi

 1: qi ← Q[0]
 2: CQ ← ∅
 3: FOR each a ∈ Q DO
 4:     IF a is not missed from r THEN
 5:  CQ ← CQ ∪ { a }
 6:     END IF
 7: END FOR
 8: n ← Root(DT)
 9: WHILE n ∉ Leaves(DT) DO
10:     IF n ∈ CQ THEN
11:         qi ← n
12:         CQ ← CQ - { n }
13:     END IF
14:     n ← NextNode(DT, n, r)
15: END WHILE
16: IF | CQ | > 0 THEN
17:     qi ← RandomSelect(CQ)
18: END IF
19: RETURN qi
```

(شکل-۵): انتخاب یک شبه‌شناسه برای فرونشانی
(Figure-5): Selection of a quasi identifier for Suppression

مراحل ۵-۲ مجموعه‌ای از شبه‌شناسه‌ها ایجاد می‌کند که مقدارشان در رکورد r هنوز فرونشانده نشده است. شبه‌شناسه‌ای که در انتها بازگردانده می‌شود از این مجموعه (CQ) انتخاب خواهد شد.

مراحل ۹-۸، بررسی صفت‌ها را از ریشهٔ درخت شروع می‌کند و تا زمانی که گره فعلی یکی از برگ‌های درخت نیست، به این کار ادامه می‌دهد. مراحل ۱۲-۱۰ درصورتی‌که گره فعلی عضوی از مجموعه شبه‌شناسه‌های CQ باشد، آن را به‌عنوان کم‌اهمیت‌ترین شبه‌شناسه‌ای که تاکنون بررسی شده قرار می‌دهد. همچنین حذف این شبه‌شناسه از CQ موجب می‌شود شبه‌شناسه‌ای که بیش از یک‌بار در مسیر ریشه تا برگ ظاهر می‌شود، فقط در نخستین دفعه مورد بررسی قرار گیرد.

مراحل ۱۷-۱۶ بررسی می‌کنند تمام شبه‌شناسه‌ها در مسیر ریشه تا درخت ملاقات شده‌اند یا خیر. درصورتی‌که هنوز چند شبه‌شناسه در مجموعه CQ وجود داشته باشد، یکی از آن‌ها به‌صورت تصادفی به‌عنوان شبه‌شناسه خروجی انتخاب می‌شود. مرحله نوزده صفت qi را که انتظار می‌رود فرونشاندن آن تأثیر کمی درنتیجه طبقه‌بندی داشته باشد در خروجی قرار می‌دهد.

۵- پیاده‌سازی طرح پیشنهادی و نتایج شبیه‌سازی و ارزیابی

پیاده‌سازی الگوریتم با استفاده از زبان جاوا و به‌کمک کتابخانه وکا [25] انجام گرفته است. وکا یکی از مشهورترین سکوهای یادگیری ماشین است که به زبان جاوا و توسط دانشگاه وایکاتو نیوزلند توسعه پیدا کرده است. کاربر می‌تواند از طریق واسط گرافیکی و واسط خط فرمان از وکا استفاده کند و یا آن را به‌عنوان یک کتابخانه به پروژه برنامه‌نویسی خود اضافه کند. با توجه به امکانات فراوانی که API وکا برای برنامه‌نویس فراهم می‌کند، نیازی به کدنویسی برای عملیات سطح پایین نیست و بسیاری از قسمت‌های لازم برای پیاده‌سازی و ارزیابی الگوریتم مانند طبقه‌ها و توابع مفید برای کار با مجموعه‌داده‌ها و الگوریتم‌های طبقه‌بندی، توسط این کتابخانه انجام می‌گیرد.

مجموعه‌داده‌های ورودی باید با فرمت csv ذخیره شده باشند. نخستین خط از فایل ورودی شامل نام صفات مجموعه‌داده است. طبق قرارداد نام شبه‌شناسه‌ها با #QI خاتمه می‌یابد و آخرین صفت به‌عنوان صفت حساس درنظر گرفته می‌شود. لازم به ذکر است که خواندن داده‌ها از پایگاه‌داده و فهرست‌گذاری روی شبه‌شناسه‌ها می‌تواند باعث بهبود سرعت اجرای الگوریتم شود.

فایل گمنام‌سازی‌شده به‌صورت arff. {نام الگوریتم گمنام‌سازی}. {نام مجموعه‌داده} در پوشه out ذخیره می‌شود. فرمت arff علاوه‌بر داده‌ها شامل سرآیندی است که فهرستی از نام و نوع صفات در آن نگهداری می‌شود.

نتایج آماری حاصل از اجرای برنامه (زمان اجرا و دقت طبقه‌بندی توسط الگوریتم‌های مختلف) در فایل متنی با نام results.txt ذخیره می‌شود.

۵-۱- ارزیابی و نتایج به‌دست‌آمده

در این بخش به بررسی روش پیشنهادی (PM) و مقایسه عملکرد آن با روش ارائه‌شده در [20] پرداخته می‌شود. همان‌طورکه گفته شد، نیاز روش‌های مبتنی بر عمومی‌سازی به درخت طبقه‌بندی، به‌عنوان نقطه ضعف آن‌ها تلقی می‌شود. برای بررسی این موضوع عملکرد روش AGF درحالی‌که درخت طبقه‌بندی در ورودی به آن داده نشده نیز مورد ارزیابی قرار گرفته است (در ادامه AGF1L نامیده می‌شود). در این حالت درختی به عمق یک که تمام مقادیر دامنه صفت در برگ‌ها و مقدار ANY در ریشه قرار دارد به‌صورت خودکار ایجاد می‌شود.

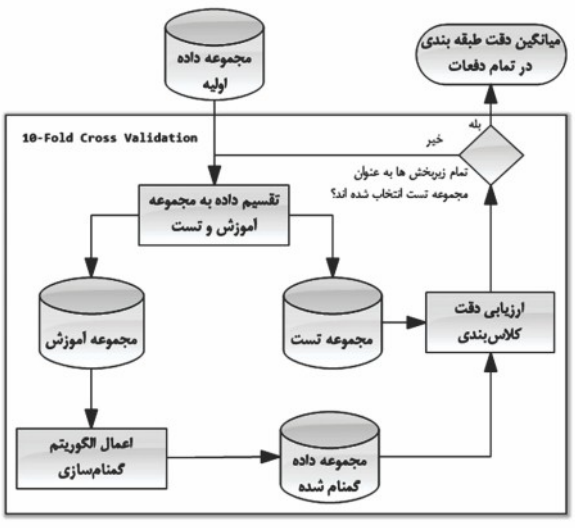
درخت تصمیم مورد نیاز روش پیشنهادی توسط الگوریتم J48 ایجاد و هنگام بازه‌بندی شبه‌شناسه‌های عددی، طول بازه‌ها ۵٪ دامنه مقادیر صفت درنظر گرفته شده است. در آزمایش‌های انجام‌شده، برای همه روش‌ها نرخ نمونه‌گیری برابر با ۹۰٪ و حد تخریب مورد قبول ۰٫۶ است. درخت‌های طبقه‌بندی مورد استفاده روش AGF مطابق [26] ایجاد شده‌اند.

روند ارزیابی دقت طبقه‌بندی، در شکل (۶) نمایش داده شده است. در ابتدا مجموعه‌داده به ده قسمت مساوی تقسیم می‌شود. یک قسمت به‌عنوان مجموعه آزمون و نُه قسمت دیگر به‌عنوان مجموعه آموزش انتخاب می‌شود؛ سپس الگوریتم گمنام‌سازی بر روی مجموعه آموزش اعمال و الگوریتم طبقه‌بندی با استفاده از آن آموزش داده می‌شود. در ادامه رکوردهای مجموعه آزمون برای اندازه‌گیری دقت طبقه‌بند ساخته شده، مورد استفاده قرار می‌گیرد. این فرآیند تا زمانی که تمام ده قسمت به‌عنوان مجموعه آزمون انتخاب شوند، ادامه می‌یابد و در انتها میانگین دقت طبقه‌بندی در تمام دورها محاسبه می‌شود.

مجموعه‌داده مورد استفاده Adult [27] نام دارد که در کارهای زیادی از آن برای ارزیابی طبقه‌بندی استفاده شده است [10]، [22] و [20]. این مجموعه‌داده شامل ۴۸۸۴۲ رکورد و پانزده صفت است.

صفت طبقه، میزان درآمد یک فرد در طول سال (Income) است که دو مقدار 50k≤ و یا 50k> می‌تواند داشته باشد. همین صفت به‌عنوان صفت حساس نیز درنظر گرفته شده است. مشخصات سایر صفت‌ها در جدول (۶) آورده شده است.

برای انجام آزمایش‌های مختلف ۶، ۹ و ۱۲ صفت از این مجموعه‌داده به‌عنوان شبه‌شناسه درنظر گرفته شده‌اند که به‌ترتیب Adult.6، Adult.9 و Adult.12 نامیده می‌شوند. صفت‌هایی که در هر یک از این حالات به‌عنوان شبه‌شناسه انتخاب شده‌اند در جدول (۷) قابل مشاهده است.



(شکل-۶): روند ارزیابی الگوریتم‌های گمنام‌سازی
(Figure-6): Evaluation of anonymization algorithms

(جدول-۶): صفت‌های مجموعه‌داده Adult
(Table-6): Atributes of Adult Data set

صفت	نوع	تعداد/دامنه مقادیر	صفت	نوع	تعداد/دامنه مقادیر
Age	پیوسته	17 - 90	Occupation	گسسته	14
Capital-gain	پیوسته	0 - 99999	Work-class	گسسته	8
Capital-loss	پیوسته	0 - 4356	Martial-status	گسسته	7
Education-num	پیوسته	1 - 16	Relationship	گسسته	6
Final-weight	پیوسته	13492 - 1490400	Sex	گسسته	2
Hours-per-week	پیوسته	1 - 99	Native-country	گسسته	40
Education	گسسته	16	Race	گسسته	5

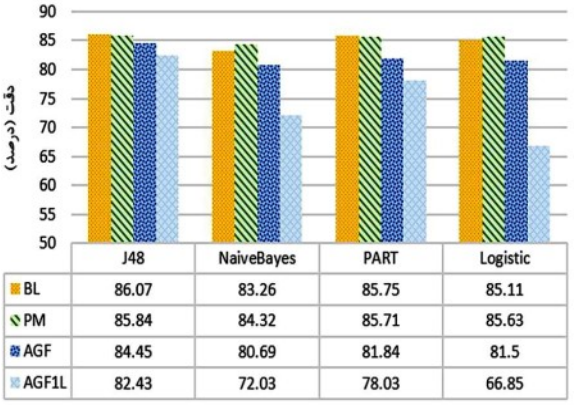
(جدول-۷): صفت‌های انتخاب شده به‌عنوان شبه‌شناسه
(Table-7): Selected attributes for quasi identifier

عنوان	شبه‌شناسه‌ها
Adult.6	age, education, sex, race, occupation, native-country
Adult.9	age, education, sex, race, occupation, native-country, workclass, marital-status, capital-loss
Adult.12	age, education, sex, race, occupation, native-country, workclass, marital-status, capital-loss, relationship, hours-per-week, capital-gain

۲-۵- دقت طبقه‌بندی

برای ارزیابی تأثیر گمنام‌سازی بر دقت طبقه‌بندی از الگوریتم‌های J48، NaiveBayes، PART و Logistic استفاده شده است. در نخستین آزمایش نّه صفت از مجموعه‌داده به‌عنوان شبه‌شناسه انتخاب شده‌اند. همان‌طور که در شکل (۷) نمایش داده شده است نتایج به‌دست‌آمده نشان می‌دهد دقت طبقه‌بندهایی که با مجموعه‌داده گمنام‌شده توسط روش پیشنهادی آموزش دیده‌اند، بیشتر است. برای مقایسه بهتر، دقت طبقه‌بندی بر روی داده‌های گمنام‌نشده نیز اندازه‌گیری شده است. نتایج به‌دست‌آمده برای این حالت، در نمودارها با

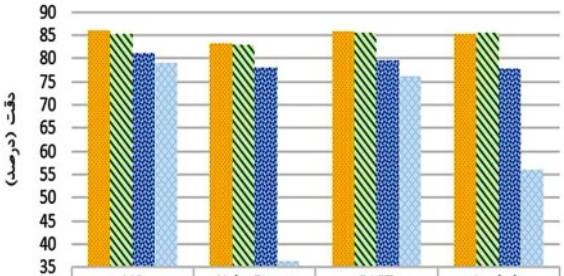
نام BL نمایش داده شده است. به‌منظور بررسی تأثیر تعداد شبه‌شناسه‌ها بر کیفیت خروجی هر الگوریتم، در آزمایش دوم از دوازده شبه‌شناسه استفاده شده است. با توجه به نمودار شکل (۸) روش پیشنهادی در این حالت توانسته است، دقت طبقه‌بند را نزدیک به حالت قبل نگاه دارد؛ درحالی‌که روش‌های AGF و AGF1L با افت دقت بیشتری مواجه شده‌اند.



	J48	NaiveBayes	PART	Logistic
BL	86.07	83.26	85.75	85.11
PM	85.84	84.32	85.71	85.63
AGF	84.45	80.69	81.84	81.5
AGF1L	82.43	72.03	78.03	66.85

(شکل-۷): دقت طبقه‌بندی – مجموعه داده Adult.9

(Figure-7): Classification accuracy in Adult.9 Data set



	J48	NaiveBayes	PART	Logistic
BL	86.07	83.26	85.75	85.11
PM	85.18	82.98	85.39	85.43
AGF	81.03	77.88	79.55	77.65
AGF1L	79.09	36.34	76.1	55.9

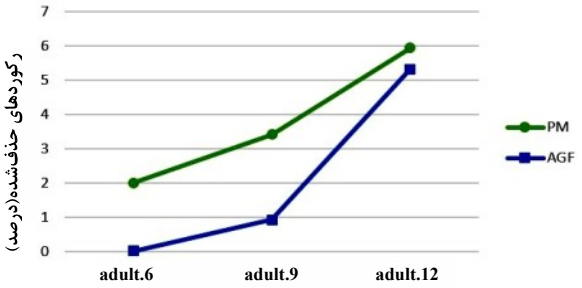
(شکل-۸): دقت طبقه‌بندی – مجموعه‌داده Adult.12

(Figure-8): Classification accuracy in Adult.12 Data set

۳-۵- رکوردهای غیرقابل انتشار

گفته‌شد که در طول اجرای الگوریتم عملگرهای دسته‌سازی مقدار تعدادی از صفات هر رکورد را تغییر می‌دهند و به‌وسیله معیار تخریب مشخص می‌شود که رکورد موردنظر کیفیت لازم را برای انتشار دارد یا خیر. شکل (۹) نسبت رکوردهای غیرقابل انتشار به کل رکوردهای مجموعه‌داده را نشان می‌دهد.

همان‌طور که انتظار می‌رفت، به‌دلیل استفاده از فرونشانی در روش پیشنهادی تعداد رکوردهای غیرقابل انتشار بیشتر است. با این حال وقتی تعداد زیادی از صفت‌ها به‌عنوان شبه‌شناسه در نظر گرفته شوند، در روش AGF عمومی‌سازی تا سطح بالایی از درخت طبقه‌بندی انجام خواهد شد و نتیجه مشابه استفاده از عملگر فرونشانی خواهد بود.



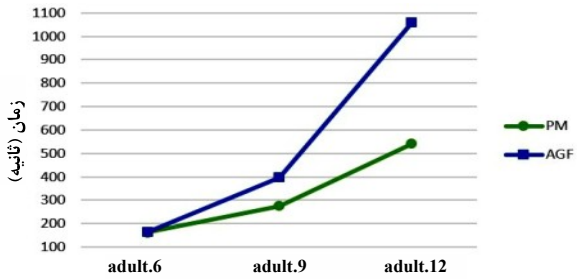
(شکل-۹): درصد رکوردهای غیرقابل انتشار

(Figure-9): A comparison on Unpublishable records

۴-۵- زمان اجرا

تعداد شبه‌شناسه‌ها و اندازه مجموعه‌داده عواملی هستند که بر زمان اجرای الگوریتم تأثیر دارند. برای آزمودن تأثیر این دو عامل، آزمایش‌های جداگانه‌ای انجام شده است. زمان اجرای کدها، بر روی سیستمی با پردازنده AMD Phenom II 1090T (3.2 GHz) و 4GB حافظه اصلی اندازه‌گیری شده است.

با افزایش تعداد شبه‌شناسه‌ها نقاط تمایز رکوردها بیشتر شده و انتظار می‌رود پردازش بیشتری برای برقرارکردن رابطه (۴) نیاز باشد. به‌منظور بررسی تأثیر تعداد شبه‌شناسه‌ها بر زمان مورد نیاز برای گمنام‌سازی، از مجموعه‌داده Adult با ۶، ۹ و ۱۲ شبه‌شناسه استفاده شد. همان‌طور که در شکل (۱۰) نمایش داده شده است، زمان مورد نیاز برای گمنام‌سازی توسط الگوریتم پیشنهادشده در مقایسه با AGF پایین‌تر است و همچنین تأثیرپذیری کمتری از تعداد شبه‌شناسه‌ها دارد. علت بهبود سرعت اجرا این است که AGF مقدار شبه‌شناسه‌ها را سطح به سطح در درخت طبقه‌بندی بالا می‌برد؛ اما در الگوریتم پیشنهادی مقدار شبه‌شناسه‌ها در یک مرحله حذف می‌شود.



(شکل-۱۰): تأثیر تعداد شبه‌شناسه‌ها بر زمان اجرا

(Figure-10): The impact of Quasi Identifiers on simulation run time

به‌منظور بررسی میزان مقیاس‌پذیری، در آزمایش‌های مختلف ۴۰، ۶۰، ۸۰ و ۱۰۰ هزار رکورد به ورودی الگوریتم داده شدند. همان‌طور که گفته شد، مجموعه‌داده Adult شامل ۴۸۸۴۲ رکورد است. برای افزایش تعداد رکوردها، برای هر صفت، احتمال داشتن مقادیر مختلف حساب شد؛ سپس بر طبق احتمالات به‌دست آمده، مقادیر مختلف صفات در کنار هم قرار داده شد و رکوردهای مورد نیاز تولید شدند. در این آزمایش شش صفت به‌عنوان شبه‌شناسه انتخاب شدند. نتایج به‌دست‌آمده در شکل (۱۱) نمایش داده شده است.

(شکل-۱۱): تأثیر تعداد رکوردها بر زمان اجرا

(Figure-11): The impact of records on simulation run time

(جدول-۸): مقایسه روش پیشنهادی و AGF

(Table-8): Comparison of proposed technique with AGF

روش	کاربرد	نیاز به درخت طبقه‌بندی	زمان	رکوردهای غیرقابل انتشار	دقت طبقه‌بندی
PM	طبقه‌بندی	خیر*	کمتر*	بیشتر	بیشتر*
AGF	عام‌منظوره*	بله	بیشتر	کمتر*	کمتر

۶- جمع‌بندی و نتیجه‌گیری

در این مقاله استفاده از عمل‌گر فرونشانی برای دست‌یابی به چهارچوب محرمانگی تعریف‌شده در [20] مورد بررسی قرار گرفت و به جای به کار گیری عمومی‌سازی سلولی از فرونشاندن صفاتی که بهرهٔ اطلاعاتی کمتری دارند، استفاده شد. اگرچه عمل‌گر فرونشانی نسبت به عمومی‌سازی، مقدار صفات را کم‌مفهوم‌تر می‌کند، ولی همان‌طورکه نتایج به‌دست‌آمده نشان می‌دهد، استفادهٔ درست از این عمل‌گر، علاوه بر این‌که می‌تواند نتیجه طبقه‌بندی را نزدیک به طبقه‌بندی روی داده‌های خام نگه دارد، مشکلات ساخت درخت طبقه‌بندی برای هر شبه‌شناسه را به همراه ندارد. در جدول (۸) روش پیشنهادی و روش ارائه‌شده در [20] با یکدیگر مقایسه شده‌اند. در هر ویژگی، روشی که عملکرد بهتری داشته با علامت ستاره(*) مشخص شده است. با توجه به این جدول، در شرایطی که داده‌ها با هدفی غیر از استفاده در طبقه‌بندی گمنام می‌شوند و یا در حالتی که رکوردهای مجموعه‌داده کم تعداد است، استفاده از روش پیشنهادی مناسب نخواهد بود.

۷- مراجع 7- References

[1] B. C. M. Fung, K. Wang, A. Wai-Chee Fu and P. S. Yu, (2010), Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques, Chapman and Hall/CRC.

[2] J. Bennett and S. Lanning, (2007), "The Netflix Prize", Proceedings of the KDD Cup Workshop, pp. 3-6.

[3] D. Nettleton, (2014), "Data Privacy and Privacy-Preserving Data Publishing," in Commercial Data Mining: Processing, Analysis and Modeling for Predictive Analytics Projects, Morgan Kaufmann, pp. 266-277.

[4] B. Fung, K. Wang and P. Yu, (2010), "Privacy-Preserving Data Publishing: A Survey of Recent Developments", ACM Computing Surveys, vol. 42, no. 4,

[5] L. Sweeney, (2002), "k-Anonymity: A Model for Protecting Privacy", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557-570.

[6] K. S. Babu, (2013), Utility-Based Privacy Preserving Data Publishing, PhD thesis, National Institute of Technology Rourkela.

[7] N. Mohammed, B. C. M. Fung, P. C. K. Hung and C. K. Lee, (2009), "Healthcare Data: A Case Study on the Blood Transfusion Service", Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1285-1294.

[8] D. Dou and S. Coulondre, (2012), "Detecting Privacy Violations in Multiple Views Publishing," in Database and Expert Systems Applications, Springer-Verlag Berlin Heidelberg, 506–513.

[9] A. Anjum and G. Raschia, (2013), "Anonymizing Sequential Releases under Arbitrary Updates", Proceedings of the Joint EDBT/ICDT 2013 Workshops, pp. 145-154.

[10] B. Fung, K. Wang and P. Yu, (2007), "Anonymizing Classification Data for Privacy Preservation", IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 5, pp. 711-725.

[11] V. S. Susan and T. Christopher, (2014), "A Survey on Privacy Preservation in Data Publishing", International Journal of Computer Science and Mobile Computing, vol. 3, no. 3, pp. 188-193.

[12] T. Dalenius, (1977), "Towards a Methodology for Statistical Disclosure Control", Statistik Tidskrift, vol. 15, 429–222.

[13] C. Dwork, (2006), "Differential Privacy," in Automata, Languages and Programming, Springer Berlin Heidelberg, pp. 1-12.

[14] K. Wang and B. C. M. Fung, (2006), "Anonymizing sequential releases", Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 414-423.

[15] A. Machanavajjhala, D. Kifer, J. Gehrke and M. Venkitasubramaniam, (2007), "l-diversity: Privacy beyond k-anonymity", ACM Transactions on Knowledge Discovery from Data, vol. 1, no. 1, article 3.

[16] N. Li, T. Li and S. Venkatasubramanian, (2007), "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity", IEEE 23rd International Conference on Data Engineering, pp. 106 - 115.

[17] Y. Rubner, C. Tomasi and L. J. Guibas, (2000), "The Earth Mover's Distance as a Metric for Image Retrieval", International Journal of Computer Vision, vol. 40, no. 2, pp. 99 - 121.

[18] N. Li, T. Li and S. Venkatasubramanian, (2010), "Closeness A New Privacy Measure for Data Publishing", IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 7, pp. 943-956.

[19] M. E. Nergiz, M. Atzori and C. Clifton, (2007), "Hiding the Presence of Individuals from Shared Databases", InProc. of ACM International Conference on Management of Data, pp. 665-676.

[20] A. S. Sattar, J. Li, X. Ding, J. Liu and M. Vincent, (2013), "A general framework for privacy preserving data publishing", Knowledge-Based Systems, vol. 54, 276–287.

[21] K. Wang, P. Yu and S. Chakraborty, (2004), "Bottom-Up Generalization: A Data Mining Solution to Privacy Protection", Fourth IEEE International Conference on Data Mining, pp. 249 - 256.

[22] B. Fung, K. Wang and Y. P.S, (2005), "Top-Down Specialization for Information and Privacy Preservation", Proc. 21st International Conference on Data Engineering, pp. 205-216.

[23] S. Kisilevich, L. Rokach, Y. Elovici and B. Shapira, (2010), "Efficient Multidimensional Suppression for K-Anonymity", IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 3, pp. 334 - 347.

[24] A. Hussien, N. Hamza and A. Hefny, (2013), "Attacks on Anonymization-Based Privacy-Preserving: A Survey for Data Mining and Data Publishing", Journal of Information Security, vol. 4, no. 2, pp. 101-112.

[25] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, (2009), "The WEKA Data Mining Software: An Update", ACM SIGKDD Explorations Newsletter, vol. 11, no. 1, pp. 10-18.

[26] "Taxonomy trees of the Adult data set", [Online]. Available: http://ddm.cs.sfu.ca/dmsoft/Privacy/products/adultHierarchy.txt. [Accessed 8 May 2016].

[27] "UCI Machine Learning Repository: Adult Data Set", [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Adult. [Accessed 8 May 2016].

[28] M. Nergiz, C. Clifton and A. Nergiz, (2009), "Multirelational k-Anonymity", IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 8, pp. 1104-1117.

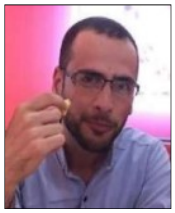
[۲۹] مهدی صادق پور، " حفظ محرمانگی در انتشار داده‌ها به وسیله گمنام‌سازی دسته‌ای"، پایان نامه کارشناسی ارشد مهندسی کامپیوتر – نرم‌افزار، دانشگاه گیلان، ۱۳۹۴.

[29] Mehdi Sadeghpour, "Privacy Preserving Data Publishing using Group Based Anonymization", MSc thesis in Software engineering, University of Guilan, 2015.

رضا ابراهیمی آتانی مدرک کارشناسی خود را از دانشگاه گیلان در سال ۱۳۸۱ دریافت و همچنین درجهٔ کارشناسی ارشد و دکترای خود را از دانشگاه علم و صنعت در سال ۱۳۸۳ و ۱۳۸۹ اخذ کرده است. در حال حاضر، ایشان عضو هیئت علمی و دانشیار در دانشکده فنی و مهندسی دانشگاه گیلان است. زمینه‌های پژوهشی مورد علاقهٔ ایشان عبارتند از: طراحی و تحلیل الگوریتم‌ها و پروتکل‌های رمزنگاری و امنیتی جهت کاربرد در شبکه‌های رایانه‌ای و بی‌سیم.

نشانی رایانامهٔ ایشان عبارت است از:

rebrahimi@guilan.ac.ir



مهدی صادق پور مدرک کارشناسی و کارشناسی ارشد خود را در سال‌های ۱۳۹۲ و ۱۳۹۴ از گروه مهندسی کامپیوتر دانشگاه گیلان اخذ کرد. موضوعات مورد علاقه ایشان مهندسی نرم افزار و طراحی و توسعه سامانه‌های امن و با حفظ حریم خصوصی سازمانی است.

نشانی رایانامهٔ ایشان عبارت است از:

mehdi.sadeghpour@live.com