Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: jscit.nit.ac.ir
Volume 11, Number 1, Spring 2022, pp. 73-88
Received: 10/15/2021, Revised: 03/11/2022, Accepted: 04/27/2022

Privacy preserving Naïve Bayes classification using Bit-string encryption

Mohammad Akbari Azimian[1], Negin Daneshpour[1*] and Masoumeh Safkhani[1]
1- Computer Engineering Department, Shahid Rajaee Teacher Training University, Tehran, Iran.
[1]mohammad@sru.ac.ir, [1*]ndaneshpour@sru.ac.ir, and [1]safkhani@sru.ac.ir

Corresponding author's address: Negin Daneshpour, Faculty of Computer Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.

Abstract- The construction of classification models is widely used in data mining. There are concerns about the privacy of data owners because of the need to collect data to build models. In this paper, a Naïve Bayes classification model construction plan is presented, which performs the model construction operation with the participation of data owners and without the need to collect the original data. Instead of collecting data, the scheme uses encryption of bit strings from counting without disclosing the original data to perform the process of creating the Naïve Bayes model. This design allows the model to be built with appropriate performance without the need for trust in a third party with a minimum number of encryptions, so that in terms of time complexity, up to 87% improvement in time cost can be observed. In addition, memory consumption has not increased significantly when compared to designs that use encryption operations.

Keywords- Data Mining, Classification, Naïve Bayes, Security, Privacy Preserving, Privacy

مجله علمی پژوهشی رایانش نرم و فناوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: jscit.nit.ac.ir
جلد ۱۱، شماره ۱، بهار ۱۴۰۱ ، صفحه ۷۳-۸۸
دریافت: ۱۴۰۰/۰۷/۲۳، بازنگری: ۱۴۰۰/۱۲/۲۰، پذیرش: ۱۴۰۱/۰۲/۰۷

حفظ حریم خصوصی در کلاس‌بندی Naïve Bayes با استفاده از رمزنگاری رشته بیت‌ها

محمد اکبری عظیمیان[۱]، نگین دانشپور*[۱]، معصومه صفخانی[۱]
۱- دانشکده مهندسی کامپیوتر، دانشگاه تربیت دبیر شهید رجائی، تهران، ایران.
[1]mohammad@sru.ac.ir, [۱*]ndaneshpour@sru.ac.ir, [۱]safkhani@sru.ac.ir

* نشانی نویسنده مسئول: نگین دانشپور، تهران، لویزان، خیابان شهید شعبانلو، دانشگاه تربیت دبیر شهید رجائی، دانشکده مهندسی کامپیوتر.

چکیده- ساخت مدل‌های کلاس‌بندی به طور گسترده‌ای در داده‌کاوی مورد استفاده قرار می‌گیرد. از آنجا که برای ساخت مدل‌ها نیاز به جمع‌آوری داده است، نگرانی‌هائی در زمینه‌ی حریم خصوصی مالکین داده‌ها وجود دارد. در این مقاله یک طرح ساخت مدل کلاس‌بندی Naïve Bayes ارائه شده است که با مشارکت مالکین داده‌ها و بدون نیاز به جمع‌آوری اصل داده‌ها، عملیات ساخت مدل را انجام می‌دهد. این طرح به جای جمع‌آوری داده‌ها، با استفاده از رمزنگاری رشته بیت‌های حاصل از شمارش و بدون افشای داده‌ها، فرآیند ساخت مدل Naïve Bayes را انجام می‌دهد. این طرح بدون نیاز به اعتماد به شخص سوم[۱] با حداقل تعداد اجرای عملیات رمزنگاری، امکان ساخت مدل را با کارایی مناسب فراهم می‌کند به‌طوری‌که از نظر پیچیدگی زمانی تا ۸۷٪ بهبود در هزینه‌ی زمانی مشاهده می‌شود و حافظه‌ی مصرفی نیز افزایش چندانی نسبت به طرح‌های دارای عملیات رمزنگاری نداشته است.

واژه‌های کلیدی: داده‌کاوی، کلاس‌بندی، Naïve Bayes، امنیت، حریم خصوصی، محرمانگی

۱- مقدمه

امروزه با گسترش حجم داده‌ها و افزایش نیاز به پاسخ سریع به تحلیل‌های داده‌ای، روش‌های داده‌کاوی روز به روز در حال پیشرفت است. اما با افزایش کاربردهای تحلیل داده، نگرانی درباره‌ی امنیت و حریم خصوصی داده‌های کاربران نیز افزایش یافته است. فناوری‌های ارتباطی نوین، خدمات جدیدی را به کاربران ارائه می‌کنند و در عین حال نیاز به دسترسی و پردازش داده‌های اشخاص یا سازمان‌ها نیز افزایش یافته است. بسیاری از افراد یا سازمان‌ها یا نهادهایی که داده‌های آن‌ها برای داده‌کاوی مورد استفاده قرار می‌گیرد، دغدغه‌ی افشای اطلاعات خود یا داده‌های افرادی که به آن‌ها اعتماد کرده‌اند را دارند.

یکی از روش‌های داده‌کاوی، کلاس‌بندی[۲] است. کلاس‌بندی یک شکل از تجزیه و تحلیل داده است و مدل‌هایی استخراج می‌کند که کلاس‌های مهم داده را توصیف می‌کنند[۲۶]. یکی از دغدغه‌های ساخت مدل‌های کلاس‌بندی، حفظ محرمانگی و حریم خصوصی مالکین داده‌ها است. از آنجا که در کلاس‌بندی نیاز به جمع‌آوری داده‌ها یا مشاهدات است، حفظ امنیت اطلاعات و حریم خصوصی می‌تواند دارای اهمیت باشد. طرح‌هایی که سعی دارند فرآیند جمع‌آوری داده‌ها را بدون افشای اطلاعات انجام دهند با محدودیت‌های زیادی مواجه می‌شوند. برخی نیاز به اعتماد به شخص سوم دارند و برخی نیز منجر به کاهش دقت می‌شوند. همچنین برای بازسازی مدل نیاز به دریافت مجدد تمام داده‌های جدید خواهند داشت که معمولاً بر تردید مالکین داده‌ها برای مشارکت در ساخت مدل می‌افزاید. از آنجائی که برای ساخت بسیاری از مدل‌ها نیاز به رمزنگاری است، کارائی طرح نیز مطرح

است، بخصوص اینکه بدون رمزنگاری، ساخت مدل با حفظ حریم خصوصی و بدون کاهش دقت، غیرمحتمل یا بسیار مشکل خواهد بود.

یکی از مدل‌های پرکاربرد کلاس‌بندی، مدل کلاس‌بندی Naïve Bayes است. از آنجا که مدل Naïve Bayes براساس ریاضیات و احتمالات است، بسیار پرکاربرد بوده و ساخت و پیاده‌سازی آن آسان‌تر از بسیاری از مدل‌ها می‌باشد[۲۶].

ایده‌ی پیشنهادی این است که با ارائه‌ی طرحی، نیازمندی‌های ساخت مدل کلاس‌بندی Naïve Bayes را بدون نیاز به جمع‌آوری اصل داده‌ها و با مشارکت مالکین داده‌ها و بدون نیاز به اعتماد به شخص سوم، بدست آورد و بتوان اقدام به ساخت مدل کلاس‌بندی Naïve Bayes با حفظ محرمانگی و حریم خصوصی کرد.

۱-۱- نوآوری انجام شده

طرح پیشنهادی براساس سازوکار شمارش رمزنگاری شده کار می‌کند. نوآوری انجام شده در طرح پیشنهادی بدین شرح است:

- برای رسیدن به حفظ محرمانگی و حریم خصوصی یک ساختار داده و الگوریتم ابتکاری ارائه شده است که با استفاده از رمزنگاری هوموورفیک[۳]، شمارش رمزنگاری شده را بدون نیاز به جمع‌آوری اصل داده‌ها و با مشارکت مالکین داده انجام می‌دهد. رمزنگاری هوموورفیک نوعی از رمزنگاری است که توانایی انجام اعمال ریاضی برروی داده‌ی رمز شده را فراهم می‌کند.
- ساختار داده و الگوریتم ابتکاری پیشنهادی به گونه‌ای است که تعداد اجرای عملیات رمزنگاری مورد نیاز را تا حد امکان کم می‌کند و منجر به کارائی مناسبی برای طرح می‌شود.
- به علت انجام رمزنگاری توسط مالکین داده‌ها و پخش شدن نیاز پردازشی بین آن‌ها، هزینه‌ی پردازشی کمی به سازنده‌ی مدل تحمیل می‌شود. همچنین ساختار داده پیشنهادی به گونه‌ای است که تعداد اجرای عملیات رمزنگاری مورد نیاز توسط هر مالک داده نیز تا حد امکان کاهش یافته است.
- با استفاده از ساختار داده و الگوریتم ابتکاری، مدل Naïve Bayes به صورتی ساخته می‌شود که نیازی به اعتماد به شخص سوم و سازنده‌ی مدل و سایر مشارکت کنندگان نمی‌باشد و به دلیل اینکه ساخت مدل با مشارکت مالکین داده و بدون افشای اطلاعات است، می‌توان بدون نگرانی از افشای اطلاعات، به صورت مکرر مدل را بازسازی کرد.
- طرح برای مجموعه داده‌ی Pima Indians Diabetes Database[۲۸] و Dermatology[۲۹] و Iris[۳۰] تست شده و کارائی طرح نشان داده شده است.

۱-۲- سازماندهی

ساختار مقاله بدین شرح است که در بخش ۲ کارهای انجام شده بیان شده است و نیازمندی‌ها در بخش ۳ آمده است. در بخش ۴ ساختار کلی طرح، چالش‌ها و راه‌حل‌ها بیان شده است. در بخش ۵ ساختار طرح پیشنهادی، ارائه شده و ساختار داده و الگوریتم‌های ساخت مدل ارائه شده است. در بخش ۶ تحلیل امنیت و در بخش ۷ پیاده‌سازی و نتایج ارزیابی بیان شده است. در آخر در بخش ۸ نتیجه‌گیری ارائه شده است.

۲- کارهای مرتبط

طرح‌ها و روش‌های داده‌کاوی با حفظ حریم خصوصی، با توجه به جنبه‌های حفظ حریم خصوصی متفاوت هستند. در بعضی از طرح‌ها مدل ساخته شده است و نیاز است کاربران بتوانند با استفاده از مدل ساخته شده و بدون افشای اطلاعاتشان داده‌کاوی کنند[۱][۲]. علاوه بر این، محرمانگی مدل نیز باید حفظ شود؛ زیرا مالک مدل ممکن است به علت هزینه و زمان زیادی که صرف ساخت مدل شده است، تمایلی به افشای مدل نداشته باشد[۱۰]. یکی از کاربردهای این طرح‌ها، در سامانه‌های تشخیص پزشکی برخط است. Naïve Bayes محرمانه[۸] یکی از این طرح‌هاست که با استفاده از رمز نگاری هوموورفیک در کلاس‌بندی Naïve Bayes، حریم خصوصی را حفظ می‌کند. Bost و همکارانش[۸] یک طرح Naïve Bayes محرمانه ارائه دادند که در آن داده‌های کاربر و جدول احتمال‌های مدل، رمز شده هستند و احتمال کلاس‌ها برای داده‌ی کاربر با استفاده از خاصیت رمزنگاری هوموورفیک محاسبه می‌شود. در نهایت احتمالات به سرور داده‌کاوی ارسال می‌شود تا بیشترین احتمال محاسبه شود. ایراد این طرح این است که سرور از احتمال‌های نتیجه‌ی داده‌کاوی آگاه می‌شود. Wood و همکارانش[۵] با ایجاد یک طرح جدید در قسمت محاسبه‌ی بیشینه‌ی احتمال‌ها، این مشکل را برطرف نمودند. آن‌ها با اعمال نگاشت و ارسال تفاضل آن‌ها به جای مقادیر اصلی به سرور، مقادیر آن‌ها را پنهان کردند. همچنین با اعمال تابع تغییر دهنده برروی تفاضل‌ها، مقدار تفاضل‌ها را نیز پنهان کردند. علاوه بر این در طرح آن‌ها از سربارهای محاسباتی کاسته شده است. در زمینه‌ی تشخیص

پزشکی، Xiaoxia_Liu و همکارانش[۷] طرحی به نام PDiag[4] ارائه دادند که می‌تواند بدون افشا کردن اطلاعات پزشکی کاربر، کلاس‌بندی را انجام دهد و مدل کلاس‌بندی نیز محرمانه بماند. این طرح نیز براساس Naïve Bayes پیاده‌سازی شد و در آن به جای روش‌های رمزنگاری هوموورفیک که بسیار زمان‌بر هستند از روش تجمیع چندجمله‌ای سبک[۵] استفاده شده است. یکی دیگر از روش‌های حفظ حریم خصوصی در کلاس‌بندی Naïve Bayes طرحی است که Bost و همکارانش[۱۷] به نام BPTG-NBC ارائه کردند. این طرح مبتنی بر رمزنگاری هوموورفیک برروی لگاریتم احتمالات است که در یک عدد بزرگ ضرب شده است. این طرح نسبت به حمله‌ی STC[6] آسیب‌پذیر است. در حمله‌ی STC مهاجم مقداری از ارزش‌ها را در اجرای پروتکل جایگزین می‌کند و از ماژول مقایسه تعبیه شده در پروتکل برای یادگیری مقدار دقیق استفاده می‌کند[۱۸]. برای رفع این آسیب‌پذیری، Chong-zhi و همکارانش[۱۸] یک طرح جدید به نام PPNBC ارائه دادند و از روش پیشنهادیشان که نام آن را Double-blinding نامیدند، استفاده کردند.

در بعضی طرح‌ها داده‌ها جمع‌آوری شده‌اند و مهم این است که از روی جدول‌ها و مدل داده‌کاوی نتوان به اطلاعات حساس مشارکت-کنندگان دست پیدا کرد که در واقع مدل منتشر می‌شود و حفظ حریم خصوصی مشارکت‌کنندگان در انتشار داده، دارای اهمیت است[۱][۲]. K-ناشناسی[۷] و l-تنوع[۸] از اصولی هستند که برای جلوگیری از بازشناسایی اشخاص از خروجی منتشر شده، به کار می‌روند[۱۵] و از کلیت‌بخشی[۹][۳] و فرونشانی[۱۰][۴] برای پنهان‌سازی داده استفاده می‌کنند. Radhika Kotecha و همکارانش[۱۴] طرحی را برای محرمانگی خروجی کلاس‌بندی داده‌های جریانی ارائه داده‌اند که از این دو اصل استفاده می‌کند. در این طرح از DAHOT[11][۱۶] برای کلاس‌بندی داده‌ی جریانی[۱۲] استفاده شده است.

بعضی طرح‌ها، به حفظ محرمانگی در فرآیند یادگیری می‌پردازند و تمرکز آن‌ها بر جمع‌آوری داده و انتشار مدل است[۱][۲]. داده‌هایی که برای یادگیری و ساخت مدل استفاده می‌شوند، همواره به صورت متمرکز نیستند. Tong Li و همکارانش[۱۰] طرحی ارائه داده‌اند که با استفاده از ساز و کار لاپلاس[۱۳]، حریم خصوصی افتراقی[۱۴] را برقرار می‌کند و منجر به حفظ حریم خصوصی می‌شود. Xiaoqian Liu و همکارانش[۲۵] نیز طرحی ارائه داده‌اند که حریم خصوصی افتراقی را برای درخت تصمیم[۱۵] برقرار می‌سازد.

بعضی از طرح‌ها به داده‌کاوی با حفظ حریم خصوصی در رایانش ابری[۱۶] تمرکز دارند. Ping Li و همکارانش[۱۲] طرحی به نام POCC[17] ارائه داده‌اند که از یک موجودیت مجزای چند منظوره برای کمک به ساخت کلید و رمزنگاری و همچنین مشارکت در بعضی محاسبات و مراحل، استفاده کرده است و با روش recrypt در رمز نگاری کاملاً هوموورفیک[۱۸] داده‌های رمز شده با کلید مالکین داده‌ها را به داده‌های رمزشده با کلید سازنده‌ی مدل تبدیل می‌کند. Harmanjeet Kaur و همکارانش[۱۳] طرحی به نام CLaMPP[19] ارائه داده‌اند که مدل کلاس‌بندی Naïve Bayes را برای فیلترینگ مشارکتی[۲۰] مبتنی بر کاربر با حفظ حریم خصوصی فراهم می‌کند. Yi Liu و همکارانش[۱۹] روشی برای کلاس‌بندی چند برچسبه در ابر[۲۱] ارائه کرده‌اند. طرح آن‌ها بر مبنای کلاس‌بندی ML-kNN[۲۰] می‌باشد که یک روش کلاس‌بندی چند برچسبه[۲۲] با استفاده از k-نزدیکترین همسایگی[۲۳] است. در طرح آن‌ها از رمزنگاری هوموورفیک جمعی[۲۴] برای حفظ حریم خصوصی استفاده شده است. Yiran Shen و همکارانش[۲۱] نیز طرحی را برای کلاس‌بندی در برنامه‌های موبایل مبتنی بر ابر[۲۵] ارائه داده‌اند که بر مبنای ماتریس اسپارس[۲۶] عمل می‌کند. در طرح آن‌ها یک شبکه مشابه TOR نیز برای جمع‌آوری داده‌ها و انجام کلاس‌بندی پیاده‌سازی شده است که موجب می‌شود هویت مالکین داده و کاربرانی که درخواست کلاس‌بندی نمونه‌های خود را دارند، پنهان بماند.

بعضی از طرح‌ها بر داده‌کاوی با حفظ حریم خصوصی در کلان داده‌ها تمرکز دارند. Chamikara و همکارانش[۳۲] طرحی به نام PABIDOT ارائه کردند که با تبدیل هندسی بهینه[۲۷] در مجموعه داده، آشفتگی غیر قابل بازگشت ایجاد می‌کند به طوری که داده‌های اصلی پنهان بمانند.

بعضی از طرح‌ها بر داده‌کاوی بر روی داده‌های توزیع شده تمرکز دارند. Duy-Hien Vu [۳۳] طرحی ارائه داده است که داده‌کاوی را بر روی داده‌های نیمه کاملِ توزیع شده، انجام می‌دهد و داده‌ها به سه بخش تقسیم می‌شوند. یک بخش نزد مالک داده محرمانه می‌ماند، یک بخش نزد داده‌کاو محرمانه می‌ماند و بخش سوم برای هر دو فاش می‌شود.

بعضی از طرح‌ها نیز از جنبه‌ی مقابله با حملات، به موضوع حفظ حریم خصوصی در داده‌کاوی پرداخته‌اند. Ngoc Hong Tran و همکارانش[۳۴] یک پروتکل تحلیل داده‌ی امن به نام SmartClass را برای مقابله با حملات داخلی ارائه داده‌اند. Jing Wang و همکارانش[۳۵] نیز طرحی به نام PDAM را برای جمع‌آوری داده در شبکه های هوشمند مجهز به اینترنت اشیا[۲۸] ارائه داده‌اند.

بعضی از طرح‌ها دقت مدل را پایین می‌آورند و برخی دیگر به اعتماد

به شخص سوم نیاز دارند که در واقعیت در بسیاری موارد برای کاربران فرقی بین اعتماد به شخص سوم و اعتماد به سازنده‌ی مدل وجود ندارد و به هردو بی‌اعتماد هستند. در بسیاری از طرح‌هائی که از رمزنگاری استفاده می‌کنند، مشکل کارائی وجود دارد.

تمرکز بر ساخت مدل با حفظ حریم خصوصی در طرح‌های کمی وجود دارد در حالی که اهمیت بیشتر بر دغدغه‌ی کاربران و مالکین داده‌ای که از داده‌های آن‌ها برای ساخت مدل استفاده می‌شود، بسیار ضروری است. بهتر است سعی شود طرحی ارائه شود که ساخت مدل داده‌کاوی بدون نیاز به اعتماد به شخص سوم و با مشارکت کاربران صورت گیرد. همچنین باید سعی شود طرحی با کارائی نتایج و با خروجی دقیق ارائه شود.

۳- مقدمات

در ادامه مقدمات طرح پیشنهادی که شامل کلاس‌بندی Naïve Bayes و رمزنگاری هوموورفیک است، بیان می‌شود.

۳-۱- کلاس‌بندی Naïve Bayes

کلاس‌بندی Naïve Bayes، کلاس‌بندی بر اساس احتمالات است. در این روش از ویژگی‌های داده‌های موجود و کلاس‌های آن‌ها، یک مدل احتمالی ساخته می‌شود. سپس با استفاده از این مدل می‌توان حدس زد که داده‌های جدید با چه احتمالی به کدام کلاس تعلق خواهند داشت[۲۶]. در واقع در این روش باید $P(C_L|x_1,x_2,\ldots,x_n)$ محاسبه شود که بیانگر احتمال تعلق به کلاس C_L (داده‌های با برچسب L) است در صورتی که (به شرطی که) داده‌ی مورد نظر $x_1,x_2,\ldots,x_n$ باشد که فرمول آن بر طبق قانون Bayes به صورت فرمول (۱) است[۲۶].

$$P(C_L|x_1,x_2,\ldots,x_n)=\frac{P(x_1,x_2,\ldots,x_n|C_L)P(C_L)}{P(x_1,x_2,\ldots,x_n)}=\frac{P(C_L)\prod_{i=1}^{n}P(x_i|C_L)}{P(x_1,x_2,\ldots,x_n)}$$

فرمول (۱)

C_Lای که بزرگترین $P(C_L|x_1,x_2,\ldots,x_n)$ را داشته باشد به عنوان کلاس انتخاب می‌شود. از آنجا که $P(x_1,x_2,\ldots,x_n)$ در مخرج کسر تمام احتمال‌ها یکسان است می‌توان آن را نادیده گرفت و فقط صورت کسر را حساب کرد. که فرمول (۲) بدست می‌آید[۲۶]:

$$P(C_L|x_1,x_2,\ldots,x_n)\propto P(C_L)\prod_{i=1}^{n}P(x_i|C_L)$$

فرمول (۲)

نکته: منظور از مدل Naïve Bayes، لیستی از تمام احتمال‌های $P(C_L)$ها و $P(x_i|C_L)$ها است که از روی داده‌ی یادگیری[۲۹] بدست می‌آیند.

۳-۲- رمزنگاری هوموورفیک

رمزنگاری هوموورفیک شیوه‌ای از رمزنگاری است که امکان انجام اعمال ریاضی برروی داده‌ی رمزشده را فراهم می‌سازد. به عنوان مثال می‌توان دو عدد رمزشده را با هم جمع کرد به طوری که حاصل جمع، رمزشده‌ی جمع دو عدد باشد[۱][۲۷]. رمزنگاری هوموورفیک می‌تواند به صورت هوموورفیک جزئی[۳۰] باشد [۲۷]. یعنی فقط بعضی از عملگرهای پایه را پوشش دهد؛ مانند هوموورفیک جمعی[۳۱] یا هوموورفیک ضربی[۳۲] [۵]. استفاده از این نوع رمزنگاری به دلیل سرعت بالاتر نسبت به هوموورفیک کامل، رایج است [۲۷].

یک طرح هوموورفیک جمعی نامیده می‌شود اگر [۵]:

$$[\![x]\!]+[\![y]\!]=[\![x+y]\!] \qquad \text{فرمول (۳)}$$

که x و y اعداد صحیح هستند و $[\![\ ]\!]$ بیانگر رمزنگاری شده‌ی عدد است. البته فرم دیگری از هوموورفیک جمعی به صورت فرمول (۴) است[۲۷]:

$$[\![x]\!].[\![y]\!]=[\![x+y]\!] \qquad \text{فرمول (۴)}$$

و هوموورفیک ضربی است اگر [۵]:

$$[\![x]\!].[\![y]\!]=[\![x.y]\!] \qquad \text{فرمول (۵)}$$

که x و y اعداد صحیح هستند.

۳-۳- نشانه‌ها

در طرح پیشنهادی، از علائمی استفاده شده است که تعریف آن‌ها در ادامه بیان می‌شود.

- v بیانگر هر رکورد داده است بطوریکه $v=x_1,x_2,\ldots,x_n|L$ که هر x_i یک خصیصه از رکورد v را نشان می‌دهد و L بیان کننده‌ی برچسب رکورد است.
- C_L بیانگر کلاس با برچسب L است.
- b_v بیانگر فرمت بیتی خاصی از رکورد v است که جزئیات و نحوه‌ی ساخت آن در طرح بیان خواهد شد.
- A تعداد کل رکوردهای مشارکت‌کنندگان است.

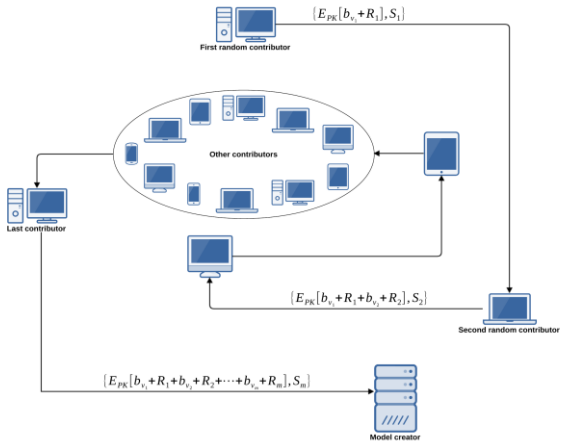
- M بیانگر تعداد مشارکت‌کنندگان است. بدیهی است اگر هر مشارکت‌کننده دارای یک رکورد باشد، M با A برابر خواهد بود.
- هر مقدار m بیانگر اندیس یک مشارکت‌کننده است.
- R_i بیانگر یک عدد تصادفی است که برای مشارکت کننده‌ی iم در نظر گرفته شده است.
- PK کلید عمومی سازنده‌ی مدل است.
- SK کلید خصوصی سازنده‌ی مدل است.
- $E_{PK}[]$ بیانگر رمزنگاری توسط کلید PK است.
- $D_{SK}[]$ بیانگر رمزگشایی توسط کلید SK است.
- S_i امضای مشارکت‌کننده‌ی iم است.

۴- بیان مسئله

فرض کنید تعدادی مالک داده و یک سازنده‌ی مدل وجود دارند و قرار است طی یک فرآیند، سازنده‌ی مدل با استفاده از داده‌های مالکین داده‌ها، کلاس‌بند Naïve Bayes بسازد. مالکین داده تمایلی ندارند داده‌های خود را در اختیار سازنده‌ی مدل یا شخص ثالثی قرار دهند. اما حاضر به مشارکت در ساخت کلاس‌بند بدون افشای داده‌هایشان هستند. این مشارکت می‌تواند به علت دریافت خدمات از یک سرویس‌دهنده در ازای مشارکت یا هر دلیل دیگری که مالک داده را مشتاق به مشارکت کند، باشد. از این رو برای ساخت کلاس‌بند Naïve Bayes، باید فرآیندی طراحی شود که سازنده‌ی مدل به نیازمندی‌های ساخت کلاس‌بند Naïve Bayes که همان شمارش در رکوردهای داده با خصوصیات و برچسب مورد نظرش است، دست یابد و از طرفی رکورد یا رکوردهای مالکین داده‌ها فاش نشود. در واقع طرح پیشنهادی، بر حریم خصوصی در مجموعه داده‌ی یادگیری[۳۳] تمرکز دارد و باید مراحل ساخت مدل و به دست آوردن احتمال‌های مختلف در مدل Naïve Bayes را در بر بگیرد. هیچکدام از مالکین داده به یکدیگر یا سازنده‌ی مدل و یا شخص ثالث اعتماد ندارند و صرفاً در فرآیند شمارش که لازمه‌ی ساخت کلاس‌بند Naïve Bayes است، مشارکت می‌کنند.

۴-۱- معماری

معماری کلی طرح پیشنهادی، در شکل (۱) آورده شده است.



شکل ۱: معماری کلی طرح پیشنهادی

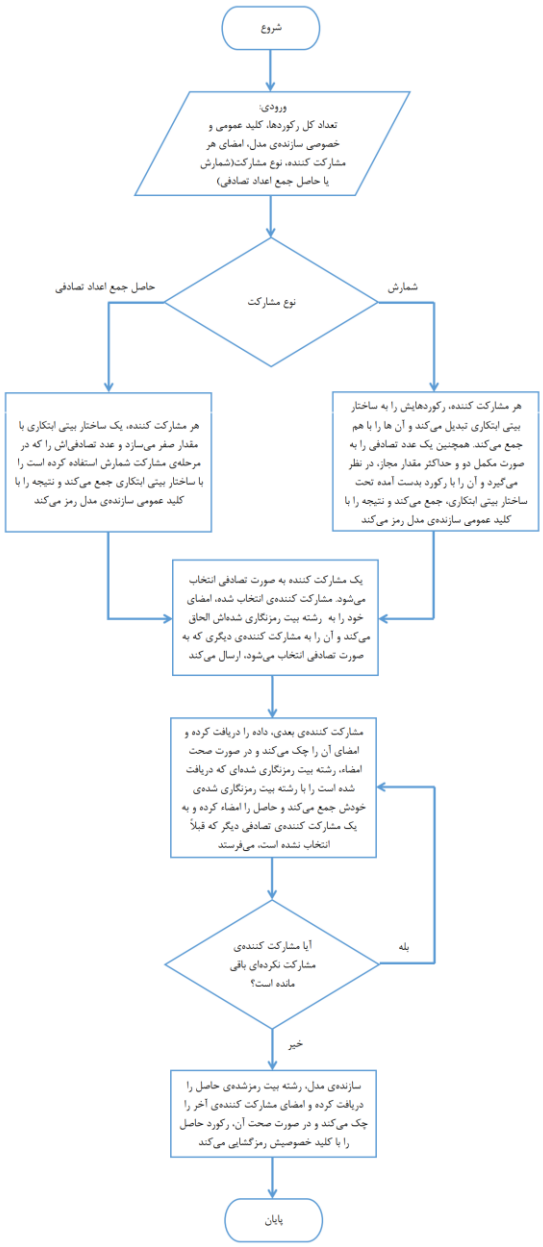
مطابق شکل (۱) موجودیت‌هایی[۳۴] که در طرح وجود دارند، عبارتند از:

- مشارکت‌کننده: مالک داده‌ای است که فرآیند یادگیری از رکورد یا رکوردهای متعلق به آن انجام می‌شود و در فرآیند ساخت مدل مشارکت می‌کند.
- سازنده‌ی مدل: مدل Naïve Bayes را با مقادیر شمارش شده، طی فرآیند یادگیری با طرح پیشنهادی، می‌سازد.

شکل (۱) معماری مراحل یک بار مشارکت همه‌ی مالکین داده‌ها را مطابق طرح پیشنهادی، نشان می‌دهد. ابتدا تمام مشارکت کننده‌ها شمارش‌های مورد نیاز را بر روی رکورد یا رکوردهایشان انجام می‌دهند و حاصل آن‌ها را به ساختار بیتی ابتکاری که در بخش ۵-۲ شرح داده می‌شود، تبدیل می‌کنند. همچنین داده بیتی حاصل را با یک عدد تصادفی متعلق به خودشان نیز جمع می‌کنند تا در صورت دسترسی به داده‌های میانی موجودیت‌ها، داده‌ی بیتی آن‌ها برای سازنده‌ی مدل فاش نشود. سپس با انجام مراحلی که در بخش ۵-۳ بیان می‌شود، شمارش با حفظ حریم خصوصی انجام می‌شود.

سازنده‌ی مدل، برای ساخت مدل Naïve Bayes به دو نوع شمارش نیاز خواهد داشت. یک نوع شمارش، شمارش برای محاسبه‌ی احتمال‌های برچسب‌ها و نوع دیگر شمارش، شمارش برای محاسبه‌ی احتمال‌های شرطی است. در نهایت سازنده‌ی مدل از ساختار بیتی ابتکاری بدست آمده و استخراج نتیجه‌ی شمارش‌ها از ساختار بیتی ابتکاری، می‌تواند احتمال‌ها را محاسبه کند.

در ادامه عملکرد کلی طرح پیشهادی در فلوچارت ۱ نشان داده شده است. البته این فرآیند یک بار برای محاسبه‌ی احتمال‌های برچسب‌ها و سپس به تعداد برچسب‌ها برای محاسبه‌ی احتمال‌های شرطی اجرا می‌شود.



فلوچارت ۱: عملکرد کلی طرح پیشهادی

۴-۲- مدل تهدید

در طرح پیشنهادی، موجودیت‌ها معتمد ولی کنجکاو[۳۵] در نظر گرفته شده‌اند و نمی‌توانند خلاف فرآیند رفتار کنند یا خرابی ایجاد کنند. طرح پیشنهادی به گونه‌ای طراحی شده است تا مانع کنجکاوی یا دسترسی موجودیت‌ها به داده‌های سایر موجودیت‌ها شود. حتی سازنده‌ی مدل نیز نباید بتواند به داده‌های مالکین داده‌ها دسترسی داشته باشد. همچنین توانائی دسترسی به داده‌ها توسط مهاجم‌هایی که در میان موجودیت‌ها نیستند، وجود نخواهد داشت.

۴-۳- چالش‌ها

طرح باید به گونه‌ای باشد که با انجام شمارش، مقادیر رکوردها برای سازنده‌ی مدل و برای سایر مالکین داده و برای مهاجم‌ها، فاش نشود. راه‌های مختلفی برای افشای اطلاعات وجود دارد که باید جلوی آن‌ها گرفته شود:

- ❖ داده‌ی هر مشارکت‌کننده باید از مشارکت‌کنندگان دیگر پنهان بماند.
- ❖ داده‌ی هر مشارکت‌کننده باید از سازنده‌ی مدل پنهان بماند.
- ❖ داده‌ی هر مشارکت‌کننده باید نسبت به دسترسی و خرابکاری افراد و موجودیت‌های خارج از فرآیند پنهان بماند.
- ❖ طول رشته بیت قابل رمزنگاری بنا بر چند بیتی بودن رمزنگاری محدودیت دارد و در طرح پیشنهادی از رشته بیت‌های بلند استفاده می‌شود.

در طرح پیشنهادی، با استفاده از رمزنگاری هوموفیک، داده‌ی هر مشارکت‌کننده از مشارکت‌کنندگان دیگر و افراد و موجودیت‌های خارج از فرآیند پنهان می‌ماند. همچنین از امضاءها برای جلوگیری از اضافه شدن داده توسط مهاجم استفاده می‌شود.

از آنجا که سازنده‌ی مدل فقط به حاصل جمع شمارش‌ها دسترسی پیدا می‌کند، به داده‌ی هر مالک دسترسی ندارد. اما برای اطمینان از اینکه نمی‌تواند به داده‌های میانی فرآیند مشارکت دست پیدا کند، هر مالک داده، داده‌ی بیتی خود را با یک عدد تصادفی جمع می‌کند. سپس در یک مشارکت جداگانه توسط همه‌ی مشارکت‌کنندگان، حاصل جمع این اعداد تصادفی نیز بدست می‌آید و از داده‌ی بیتی نهائی حذف می‌شود.

برای رفع محدودیت طول رشته بیت قابل رمزنگاری، ساختار بیتی ابتکاری بیان شده در بخش ۵-۲، در بخش ۵-۴ به گونه‌ای به تکه‌های کوچکتر تقسیم شده است که این محدودیت برطرف شود و قابلیت شمارش نیز حفظ شود.

۵- طرح پیشنهادی

در ادامه کلاس‌بندی Naïve Bayes با حفظ حریم خصوصی مشارکت‌کنندگان طراحی می‌شود.

۵-۱- بررسی اجمالی

در طرح پیشنهادی، هر مشارکت‌کننده دارای یک یا چند رکورد داده است. هر مشارکت‌کننده، داده‌ی خود را طوری پنهان می‌کند که مشارکت‌کنندگان دیگر و سازنده‌ی مدل نتوانند از محتوای آن باخبر شوند ولی بتوانند محاسبات را انجام دهند. برای انجام محاسبات از رمزنگاری هوموورفیک استفاده شده است که قابلیت اعمال عملگرهای ریاضی برروی داده‌ی رمزنگاری شده را بدون اینکه محتوای داده‌ی رمز شده فاش شود، فراهم می‌کند. البته برای طرح پیشنهادی، پشتیبانی از عملگر جمع کفایت می‌کند. هر مشارکت-کننده به وسیله‌ی امضای مخصوص به خود شناخته می‌شود تا از دستکاری توسط موجودیت‌های خارجی جلوگیری شود. هدف این است که با استفاده از ساختار بیتی ابتکاری و الگوریتم شمارش با حفظ حریم خصوصی، یک مدل کلاس‌بندی Naïve Bayes ساخته شود.

هر رکورد داده به صورت $<x_1,x_2,...,x_n|L>$ است که $x_1,x_2,...,x_n$ خصیصه‌ها و L بیانگر برچسب می‌باشد.

در ادامه مراحل مشارکت مالکین داده‌ها در طرح پیشنهادی بیان می‌شود:

- مرحله‌ی مقداردهی اولیه[۳۶]: در این مرحله، مقدارهای اولیه طرح مقداردهی می‌شوند. کلیدهای رمزنگاری (PK,SK) ایجاد می‌شوند و امضای هر مشارکت‌کننده $(S_1,S_2,...,S_m)$ ایجاد می‌شود که امضاءها و کلیدهای عمومی از طریق کانال امن در اختیار سازنده مدل و مشارکت‌کنندگان قرار داده می‌شوند.
- مرحله‌ی آماده‌سازی[۳۷]: در این مرحله، هر مشارکت‌کننده رکورد یا رکوردهایش را به ساختار بیتی ابتکاری که در بخش ۵-۲ بیان می‌شود، تبدیل می‌کند و همچنین عدد تصادفی مربوط به خود را می‌سازد و رمزنگاری و محاسبات مورد نیاز برای مشارکتش در ساخت مدل را انجام می‌دهد.
- مرحله‌ی شمارش: در این مرحله، سازنده‌ی مدل با مشارکت مالکین داده‌ها طی یک فرآیند امن، شمارش‌های مورد نیاز برای ساخت مدل Naïve Bayes را انجام می‌دهد و مدل را می‌سازد.

مراحل آماده‌سازی و شمارش به تعداد مورد نیاز تکرار می‌شوند.

۵-۲- ساختار بیتی ابتکاری

در طرح پیشنهادی، هر مشارکت‌کننده باید رکورد یا رکوردهایش را به شکلی تغییر دهد که بتوان خصیصه‌ها و برچسب‌ها را در یک فرآیند ریاضی شمارش کرد. در حالت عادی، رکورد $<x_1,x_2,...,x_n|L>$ را می‌توان در یک آرایه، به صورت شکل (۲) نشان داد.

برچسب	خصیصه n	.....	خصیصه ۲	خصیصه ۱
L	x_n	.....	x_2	x_1

شکل ۲: شمای کلی هر رکورد

در شکل (۲)، x_1 تا x_n نشان‌دهنده‌ی خصیصه‌ها و L نشان‌دهنده‌ی برچسب رکورد است.

اگر یک رکورد رمز شود، دیگر نمی‌توان بدون رمزگشایی مقدار خصیصه‌ها و برچسب‌ها را شمارش کرد. در طرح پیشنهادی برای این منظور، رکوردها به ساختار بیتی ابتکاری خاصی تبدیل می‌شوند که این هدف محقق شود.

هر رکورد دارای خصیصه‌هایی است که مقادیر خاص خود را می‌پذیرند (مانند جنسیت، سن، ...). ابتدا باید تعداد مقادیری که هر خصیصه‌ی x_i می‌پذیرد را در نظر گرفت. به عنوان مثال اگر x_i جنسیت است، مقادیر مرد و زن را می‌پذیرد. پس تعداد مقادیر قابل پذیرش برای خصیصه‌ی x_i عدد ۲ می‌شود. حال باید برای هر مقدار قابل پذیرش، یک رشته بیت با طول ثابت در نظر گرفت و با در کنار هم قرار دادن آن‌ها رشته بیت مربوط به آن خصیصه را ساخت. به عنوان مثال برای جنسیت، می‌توان دو رشته بیت را در نظر گرفت و با در کنار هم قرار دادن آن‌ها رشته بیت مربوط به خصیصه‌ی جنسیت را ساخت. در رشته بیت متناظر با مقداری که در خصیصه‌ی x_i وجود دارد، عدد ۱ قرار داده می‌شود و در بقیه‌ی رشته بیت‌ها مقدار صفر در نظر گرفته می‌شود. در شکل (۳) یک نمونه رکورد که به ساختار بیتی ابتکاری تبدیل شده است، نشان داده شده است.

برچسب	خصیصه n	.....	خصیصه ۲	خصیصه ۱
0000001 0000000	0000000 0000001 0000000	.....	0000000 0000000 0000000 0000001	0000001 0000000

شکل ۳: یک رکورد با ساختار بیتی ابتکاری در طرح پیشنهادی

همانطور که در شکل (۳) مشاهده می‌شود، مقدار هر خصیصه با رشته بیتی که بیت کم ارزش آن ۱ است مشخص می‌شود. به عنوان مثال اگر خصیصه‌ی اول، معادل جنسیت در نظر شود و طول رشته بیت برای هر یک از مقادیر قابل پذیرش در خصیصه، برابر ۷ باشد، کل خصیصه دارای رشته بیتی به طول ۱۴ خواهد بود که از قراردادن رشته بیت‌های مربوط به همه‌ی مقادیر قابل پذیرش در خصیصه در کنار هم بدست می‌آید. به طوری که رشته بیت "00000010000000" معادل زن بودن و رشته بیت

"00000000000001" معادل مرد بودن، خواهد بود.

حال اگر هر رکورد، در مجموع به صورت رشته بیت‌های پشت سر هم در نظر گرفته شود که از کنار هم قراردادن رشته بیت‌های هر خصیصه بدست آمده است، با جمع بیتی دو رکورد می‌توان تعداد مقادیر موجود برای هر مقدار قابل پذیرش در هر خصیصه را نیز جمع کرد و هر رشته بیت متناظر با مقدار قابل پذیرش در خصیصه، بعد از عمل جمع، تعداد وجود آن مقدار برای آن خصیصه را مشخص خواهد کرد. به عنوان مثال اگر در حاصل جمع چندین رکورد، رشته بیت دوم خصیصه‌ی اول حاوی عدد k بود این معنی را می‌دهد که k رکورد وجود داشته است که مقدار خصیصه‌ی اول آن‌ها برابر با مقدار متناظر با رشته بیت دوم بوده است. به عنوان مثال اگر رشته بیت دوم خصیصه‌ی اول بیانگر مرد بودن باشد، k رکورد وجود داشته است که مقدار خصیصه‌ی اول(جنسیت) آن‌ها مرد بوده است.

نکته‌ای که در اینجا باید به آن توجه شود طول رشته بیت‌ها است. اگر تعداد کل رکوردها، A در نظر گرفته شود و قرار باشد در فرآیند جمع رشته بیت‌ها سرریز اتفاق نیفتد، طول هر رشته بیت که برای هر مقدار مورد پذیرش در خصیصه، در نظر گرفته می‌شود، باید حداقل برابر با $\lfloor log_2 A \rfloor + 1$ باشد. در این صورت می‌توان A تعداد عدد ۱ را در این رشته بیت جمع کرد بدون اینکه سرریز رخ دهد.

به دلیل اینکه در طرح پیشنهادی هر مشارکت‌کننده رکورد یا رکوردهایش را با عدد تصادفی خودش جمع می‌کند، جمع عدد تصادفی هر مشارکت‌کننده با رکورد یا رکوردهایش نباید دارای رقم نقلی[۳۸] باشد تا باعث سرریز از ساختار بیتی نشود. برای اینکه در جمع رکورد با عدد تصادفی، سرریز اتفاق نیفتد باید برای مقدار عدد تصادفی محدودیت در نظر گرفت. برای این منظور باید بزرگترین عددی که یک رکورد می‌تواند داشته باشد را ساخت که برای ساخت این عدد باید برای تمامی خصیصه‌های رکورد از بین رشته بیت‌های متناظر با مقادیر قابل پذیرش در آن خصیصه، در سمت چپ‌ترین رشته بیت مقدار ۱ قرار داده شود. در شکل (۴) یک نمونه از چنین رکوردی نمایش داده شده است.

برچسب	خصیصه n	.....	خصیصه ۲	خصیصه ۱
0000001 0000000	0000001 0000000 0000000	.....	0000001 0000000 0000000 0000000	0000001 0000000

شکل ۴: بزرگترین مقدار عددی که ممکن است در یک رکورد ذخیره شود

در شکل (۴) با قرار گرفتن عدد ۱ در سمت چپ‌ترین رشته بیت در هر خصیصه، بزرگترین عدد ممکن برای یک رکورد بدست آمده است. حال با کم‌کردن بزرگترین عدد ممکن برای یک رکورد از بزرگترین عددی که می‌توان با مجموعه‌ی تمام رشته بیت‌های یک رکورد ساخت، می‌توان به حداکثر مقدار ممکن برای عدد تصادفی رسید که اگر با هر رکوردی جمع شود منجر به سرریز نخواهد شد. البته می‌توان به جای این محدودیت $\lfloor log_2 A \rfloor + 1$ بیت به سمت چپ رکورد اضافه کرد تا بتواند بیت‌های حاصل از جمع حداکثر تعداد A رکورد با تعداد A عدد تصادفی را در خود نگه دارد.

علاوه بر بیت‌های رقم نقلی احتمالی در جمع رکورد متعلق به یک مشارکت‌کننده با عدد تصادفی‌اش، جمع رکوردهای مشارکت‌کنندگان با هم نیز به علت وجود اعداد تصادفی می‌تواند باعث تولید بیت‌های رقم نقلی شود. بنابراین باید $\lfloor log_2 A \rfloor + 1$ بیت نیز برای نگهداری بیت‌های رقم نقلی احتمالی حاصل از جمع تعداد A عدد تصادفی با هم، به سمت چپ رکورد اضافه شود. شکل (۵) یک نمونه رکورد با بیت‌های اضافه را نشان می‌دهد.

برچسب	خصیصه n	.....	خصیصه ۲	خصیصه ۱	
0000001 0000000	0000000 0000001 0000000	.....	0000000 0000000 0000000 0000001	0000001 0000000	0000000

شکل ۵: یک رکورد با ساختار بیتی ابتکاری با بیت‌های نگهداری رقم نقلی

۵-۳- شمارش با حفظ حریم خصوصی

با استفاده از ساختار بیتی ابتکاری، می‌توان شمارش با حفظ حریم خصوصی را با اجرای چند مرحله انجام داد:

۱) هر مشارکت‌کننده، رکورد یا رکوردهایش را به ساختار بیتی ابتکاری توضیح داده شده در بخش ۵-۲ تبدیل می‌کند. اگر مشارکت‌کننده دارای بیش از یک رکورد باشد ساختارهای بیتی‌ای که برای رکوردهایش ساخته است را با هم جمع می‌کند. همچنین یک عدد تصادفی را به صورت مکمل دو و حداکثر مقدار مجاز که در بخش ۵-۲ بیان شده است، در نظر می‌گیرد و آن را با رکورد بدست آمده تحت ساختار بیتی ابتکاری، جمع می‌کند.

در نهایت هر مشارکت‌کننده ساختار بیتی ابتکاری نهائی خود را که با عدد تصادفی جمع شده است، با کلید عمومی سازنده‌ی مدل رمز می‌کند.

$$E_{PK}[b_{v_i} + R_i]$$

۲) یک مشارکت‌کننده به صورت تصادفی انتخاب می‌شود. مشارکت‌کننده‌ی انتخاب شده، امضای خود را به رشته بیت رمزنگاری شده‌اش الحاق می‌کند و آن را به مشارکت‌کننده‌ی دیگری که به صورت تصادفی انتخاب می‌شود، ارسال می‌کند.

$$\{E_{PK}[b_{v_1} + R_1], S_1\}$$

۳) مشارکت‌کننده‌ی بعدی، داده را دریافت کرده و امضای آن را چک می‌کند و در صورت صحت امضاء، رشته بیت رمزنگاری شده‌ای که دریافت شده است را با رشته بیت رمزنگاری شده‌ی خودش جمع می‌کند و حاصل را امضاء کرده و به یک مشارکت کننده‌ی تصادفی دیگر که قبلاً انتخاب نشده است، می‌فرستد.

$$E_{PK}[b_{v_1} + R_1] + E_{PK}[b_{v_2} + R_2] = E_{PK}[b_{v_1} + R_1 + b_{v_2} + R_2] \rightarrow \{E_{PK}[b_{v_1} + R_1 + b_{v_2} + R_2], S_2\}$$

۴) مشارکت‌کنندگان دیگر نیز مرحله‌ی(۳) را تکرار می‌کنند تا اینکه به آخرین مشارکت‌کننده برسد و آخرین مشارکت‌کننده داده‌ی حاصل را به سازنده‌ی مدل ارسال کند.

$$E_{PK}[b_{v_1} + R_1 + b_{v_2} + R_2 + \cdots + b_{v_{i-1}} + R_{i-1}] + E_{PK}[b_{v_i} + R_i] = E_{PK}[b_{v_1} + R_1 + b_{v_2} + R_2 + \cdots + b_{v_i} + R_i] \rightarrow \{E_{PK}[b_{v_1} + R_1 + b_{v_2} + R_2 + \cdots + b_{v_i} + R_i], S_i\}$$

۵) سازنده‌ی مدل، رشته بیت رمزشده‌ی حاصل را دریافت کرده و امضای مشارکت‌کننده‌ی آخر را چک می‌کند و در صورت صحت آن، رکورد حاصل را با کلید خصوصیش رمزگشایی می‌کند.

$$D_{SK}\left[E_{PK}[b_{v_1} + R_1 + b_{v_2} + R_2 + \cdots + b_{v_m} + R_m]\right] = b_{v_1} + R_1 + b_{v_2} + R_2 + \cdots + b_{v_m} + R_m = b_{v_1} + b_{v_2} + \cdots + b_{v_m} + R_1 + R_2 + \cdots + R_m$$

۶) مراحل ۱ تا ۴ مجددا تکرار می‌شود با این تفاوت که هر مشارکت-کننده به جای مجموع رکورد یا رکوردهایش با عدد تصادفی خود، فقط عدد تصادفی‌اش را در رکورد با ساختار بیتی ابتکاری قرار می‌دهد و محاسبات را انجام داده و به مشارکت کننده‌ی بعدی ارسال می‌کند. همچنین مشارکت‌کنندگان تصادفی آغازین قبلی نمی‌توانند مجدداً مشارکت‌کننده‌ی آغازین باشند که این شرط همان‌گونه که در بخش ۴-۳ بیان شد، باعث جلوگیری از افشای اطلاعات با بررسی داده‌های میانی می‌شود. در نهایت آخرین مشارکت‌کننده، داده‌ی حاصل را به سازنده‌ی مدل ارسال می‌کند.

$$\{E_{PK}[R_1 + R_2 + \cdots + R_m], S_m\}$$

۷) سازنده‌ی مدل، رشته بیت رمزشده‌ی حاصل را دریافت کرده و امضای مشارکت‌کننده‌ی آخر را چک می‌کند و در صورت صحت آن، رکورد حاصل را با کلید خصوصیش رمزگشایی می‌کند و به مجموع اعداد تصادفی می‌رسد.

$$D_{SK}[E_{PK}[R_1 + R_2 + \cdots + R_m]] = R_1 + R_2 + \cdots + R_m$$

۸) در نهایت سازنده‌ی مدل با کم کردن حاصل جمع اعداد تصادفی از ساختار بیتی بدست آمده، به ساختار بیتی نهائی می‌رسد که حاصل جمع همه‌ی ساختارهای بیتی رکوردهای مشارکت‌کنندگان است.

$$b_{v_1} + b_{v_2} + \cdots + b_{v_m} + R_1 + R_2 + \cdots + R_m - (R_1 + R_2 + \cdots + R_m) = b_{v_1} + b_{v_2} + \cdots + b_{v_m}$$

۹) در ساختار بیتی نهائی بدست آمده، رشته بیت‌های متناظر با هر مقدار قابل پذیرش در یک خصیصه، بیانگر تعداد رکوردهایی است که مقدار آن خصیصه‌شان آن مقدار متناظر بوده است و سازنده‌ی مدل می‌تواند حاصل شمارش‌ها را بدست آورد.

در الگوریتم(۱) مراحل یک بار مشارکت همه‌ی مالکین داده‌ها در شمارش با حفظ حریم خصوصی بیان شده است.

Algorithm(1): Contribution
// Contribute in counting with privacy preserving
Input:
A, total number of vectors
(*PK*, *SK*), public and private key of the model creator
(S_1,S_2, …,S_m), signatures of contributors
isForSumRandNums, specifies that is it for collecting random numbers?
Output:
1: **for** each contributor i **do**
2: **if** isForSumRandNums
3: Create $b_{v_i}$with *zero* value;
4: **else**
5: Create b_v for each vector owned by the contributor and add them together to get the b_{v_i};
6: **end if**
3: Make a random number R_i which does not cause an overflow;
4: $b_{v_i} + R_i \rightarrow E_{PK}[b_{v_i} + R_i]$;
5: **end for**
6: Choose first random contributor;
7: $E_{PK}[b_{v_1} + R_1] \rightarrow \{E_{PK}[b_{v_1} + R_1], S_1\}$;
8: **while** there is a contributor who did not contributed **do**
9: Choose a random contributor *i* who did not contributed;
10: Obtain $\{E_{PK}[b_{v_1} + R_1 + \cdots + b_{v_{i-1}} + R_{i-1}], S_{i-1}\}$;
11: Check that S_{i-1} is correct;
12: $E_{PK}[b_{v_1} + R_1 + \cdots + b_{v_{i-1}} + R_{i-1}] + E_{PK}[b_{v_i} + R_i] = E_{PK}[b_{v_1} + R_1 + \cdots + b_{v_i} + R_i]$;
13: $E_{PK}[b_{v_1} + R_1 + \cdots + b_{v_i} + R_i] \rightarrow \{E_{PK}[b_{v_1} + R_1 + \cdots + b_{v_i} + R_i], S_i\}$;
14: **end while**
15: Contributor obtains $\{E_{PK}[b_{v_1} + R_1 + \cdots + b_{v_m} + R_m], S_m\}$;
16: Check that S_m is correct;
17: $D_{SK}\left[E_{PK}[b_{v_1} + R_1 + \cdots + b_{v_m} + R_m]\right] = b_{v_1} + R_1 + \cdots + b_{v_m} + R_m = b_{v_1} + \cdots + b_{v_m} + R_1 + \cdots + R_m$;

الگوریتم ۱: مشارکت همه‌ی مالکین داده‌ها در شمارش با حفظ حریم خصوصی

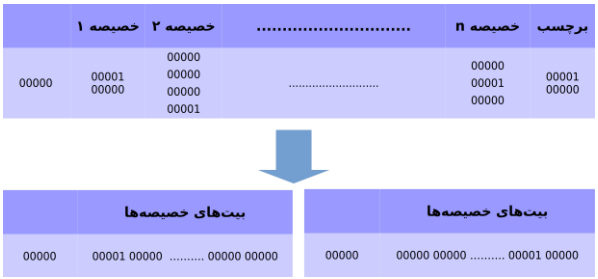
در الگوریتم(۲) مراحل یک بار شمارش با حفظ حریم خصوصی که با دو بار مشارکت همه‌ی مالکین داده‌ها همراه است، بیان شده است.

Algorithm(2): Counting
// Counting with privacy preserving
Input:
A, total number of vectors
(*PK*, *SK*), public and private key of the model creator
(S_1,S_2,…,S_m), signatures of contributors
Output:
1: $b_{v_1} + b_{v_2} + \cdots + b_{v_m} + R_1 + R_2 + \cdots + R_m \leftarrow$ Contribution(*PK*, *SK*, (S_1,S_2,…,S_M));
2: $R_1 + R_2 + \cdots + R_m \leftarrow$ Contribution(*PK*, *SK*, (S_1,S_2,…,*Sm*), *isForSumRandNums*);
3: $b_{v_1} + b_{v_2} + \cdots + b_{v_m} + R_1 + R_2 + \cdots + R_m - (R_1 + R_2 + \cdots + R_m) = b_{v_1} + b_{v_2} + \cdots + b_{v_m}$

الگوریتم ۲: یک بار شمارش با حفظ حریم خصوصی

۵-۴- رفع محدودیت طول رشته بیت قابل رمزنگاری

طول رشته بیت قابل رمزنگاری بنا بر چند بیتی بودن رمزنگاری محدودیت دارد. به عنوان مثال در رمزنگاری ۲۰۴۸ بیتی، رشته بیتی با طول کمتر از ۲۰۴۸ را می‌توان رمزنگاری کرد. از آنجا که طول رشته بیت در ساختار بیتی ابتکاری می‌تواند بسیار طولانی باشد، ممکن است نیاز باشد ساختار بیتی ابتکاری را به قسمت‌های کوچکتر تقسیم کرد. برای این منظور می‌توان ساختار بیتی ابتکاری را به رشته بیت‌هائی کوچکتر که ضریبی از $\lfloor log_2 A \rfloor + 1$ باشد، تقسیم‌بندی کرد و برای هر تکه نیز عدد تصادفی جداگانه‌ای در نظر گرفت. همچنین باید به اندازه‌ی $\lfloor log_2 A \rfloor + 1$ بیت نیز برای نگهداری بیت‌های رقم نقلی احتمالی حاصل از جمع تعداد A عدد تصادفی با هم، در سمت چپ رشته بیت‌ها در نظر گرفت. شکل (۶) یک نمونه از تقسیم‌بندی را نشان می‌دهد که یک ساختار بیتی ابتکاری به دو رشته بیت با ضریب $\lfloor log_2 A \rfloor + 1$ تقسیم شده است.



شکل ۶: تقسیم یک رکورد با ساختار بیتی ابتکاری به دو رشته بیت

همانطور که در شکل (۶) مشاهده می‌شود $\lfloor log_2 A \rfloor + 1$ بیت در سمت چپ هر رشته برای نگهداری بیت‌های رقم نقلی احتمالی حاصل از جمع تعداد A عدد تصادفی با هم، در نظر گرفته شده است.

اگر تقسیم‌بندی به گونه‌ای انجام شود که هر رشته بیت، بلندترین ضریبی از $\lfloor log_2 A \rfloor + 1$ باشد که می‌توان رمزنگاری کرد، حالت بهینه خواهد بود. حال کافی است هر مشارکت‌کننده به جای ساختار بیتی ابتکاری، هر رشته بیت را با یک عدد تصادفی جمع کرده و به صورت جداگانه رمزنگاری کند.

$$b_{v_i} \rightarrow b1_{v_i}, b2_{v_i}, \cdots, bn_{v_i}$$
$$\rightarrow \{E_{PK}[b1_{v_i} + R_{i_1}], E_{PK}[b2_{v_i} + R_{i_2}], \cdots, E_{PK}[bn_{v_i} + R_{i_n}]\}$$

بنابراین اگر در الگوریتم مشارکت مالکین داده‌ها، تک تک رشته بیت‌های متناظر به صورت مجزا جمع شوند، تغییری در نتیجه ایجاد نخواهد شد و الگوریتم می‌تواند با ساختارهای بیتی ابتکاری بلند نیز کار کند.

$$\{E_{PK}[b1_{v_i} + R_{i_1}], E_{PK}[b2_{v_i} + R_{i_2}], \cdots, E_{PK}[bn_{v_i} + R_{i_n}]\}$$
$$+\{E_{PK}[b1_{v_j} + R_{j_1}], E_{PK}[b2_{v_j} + R_{j_2}], \cdots, E_{PK}[bn_{v_j} + R_{j_n}]\}$$
$$= \{E_{PK}[b1_{v_i} + R_{i_1} + b1_{v_j} + R_{j_1}], E_{PK}[b2_{v_i} + R_{i_2} + b2_{v_j} + R_{j_2}], \cdots, E_{PK}[bn_{v_i} + R_{i_n} + bn_{v_j} + R_{j_n}]\}$$

۵-۵- ساخت مدل کلاس‌بندی Naïve Bayes با حفظ حریم خصوصی

از آنجا که مدل کلاس‌بندی Naïve Bayes نیاز به شمارش دارد، با استفاده از ساختار بیتی ابتکاری و الگوریتم شمارش با حفظ حریم خصوصی، می‌توان مدل کلاس‌بندی Naïve Bayes را ساخت.

۵-۵-۱- ایده‌ی اصلی

همان‌طور که گفته شد در کلاس‌بندی Naïve Bayes برقرار است:

$$P(C_L|x_1, x_2, \ldots, x_n) \propto P(C_L) \prod_{i=1}^{n} P(x_i|C_L)$$

پس باید $P(C_L)$ها و $P(x_i|C_L)$ها محاسبه شوند. برای محاسبه‌ی $P(C_L)$ باید تعداد رکوردهای با برچسب L بر تعداد کل رکوردها تقسیم شود و برای محاسبه‌ی $P(x_i|C_L)$ باید تعداد رکوردهایی که پارامتر iام آن‌ها برابر با x_i و برچسب آن‌ها برابر L است، بر تعداد رکوردهایی که برچسب آن‌ها L است تقسیم شود. برای بدست آوردن تعداد رکوردها با شرایط تعیین شده از ساختار بیتی ابتکاری و الگوریتم شمارش با حفظ حریم خصوصی، استفاده می‌شود.

۵-۵-۲- مقداردهی اولیه

برای مقداردهی اولیه، کلید عمومی و خصوصی سازنده‌ی مدل بر

اساس رمزنگاری هومومورفیک ساخته می‌شود (PK, SK) و کلید عمومی در اختیار همه‌ی مشارکت‌کنندگان قرار می‌گیرد. پشتیبانی از جمع برای نوع رمزنگاری کفایت می‌کند. اگر مشارکت‌کنندگان بتوانند بیش از یک رکورد داشته باشند، باید تعداد کل رکوردهای مشارکت‌کنندگان (A) نیز مشخص شود؛ در غیر این صورت تعداد کل رکوردهای مشارکت‌کنندگان برابر با تعداد مشارکت‌کنندگان خواهد بود. همچنین هر مشارکت‌کننده امضای مخصوص به خود (S_i) را برای چک‌کردن صحت تعلق داده به خود می‌سازد و در اختیار بقیه‌ی موجودیت‌ها قرار می‌دهد.

۵-۵-۳- محاسبه‌ی احتمال‌های برچسب‌ها

برای محاسبه‌ی احتمال برچسب‌ها کافی است یک بار فرآیند شمارش اجرا شود. در این صورت سازنده‌ی مدل در رشته بیت‌های متناظر با هر مقدار ممکن برچسب L، می‌تواند تعداد وجود آن برچسب در رکوردها را بدست آورده و با تقسیم آن‌ها بر کل تعداد رکوردها، احتمال برچسب $P(C_L)$ را محاسبه کند. همچنین برای بهبود کارائی می‌توان به جای کل بیت‌های رکوردها فقط بیت‌های بخش برچسب را در ساختار بیتی ابتکاری قرار داد. در این صورت یک ساختار بیتی ابتکاری کوچکتر بدست می‌آید که سرعت محاسبات و مصرف حافظه را کاهش می‌دهد. یک نمونه از چنین ساختار بیتی ابتکاری‌ای با دو برچسب قابل پذیرش در شکل (۷) نشان داده شده است که پس از انجام یک بار مشارکت همه‌ی مالکین داده‌ها، تعداد وجود هر برچسب در رشته بیت‌های متناظر با آن برچسب قرار گرفته است.

	برچسب
0000000	L_1 count = 0010011101 L_2 count = 0100110110

شکل ۷: ساختار بیتی ابتکاری مخصوص شمارش تعداد برچسب‌ها که پس از پایان یک بار مشارکت همه‌ی مالکین داده‌ها بدست آمده است

۵-۵-۴- محاسبه‌ی احتمال‌های شرطی

برای محاسبه‌ی احتمال‌های شرطی باید شمارش به شرط داشتن برچسب مورد نظر انجام شود. پس باید به تعداد برچسب‌های ممکن، فرآیند شمارش اجرا شود؛ به صورتی که در هر بار اجرای فرآیند شمارش، اگر مقدار برچسب موجود در رکورد هر مشارکت‌کننده با مقدار برچسب مورد نظر L در آن اجرا برابر بود، خود رکورد و در غیر این صورت رکوردی با بیت‌های کاملا صفر را در ساختار بیتی ابتکاری‌اش قرار می‌دهد. در این صورت سازنده‌ی مدل در رشته بیت‌های متناظر با هر مقدار ممکن در یک خصیصه، می‌تواند تعداد وجود آن مقدار خصیصه در رکوردها را به شرط اینکه برچسب موجود در رکورد، همان برچسب مورد نظر باشد، بدست آورد و با تقسیم آن بر کل تعداد رکوردهای دارای برچسب مورد نظر (که در شمارش محاسبه‌ی احتمال‌های برچسب‌ها بدست آمده است)، مقدار احتمال شرطی $P(x_i|C_L)$ را محاسبه کند. همچنین از آنجا که شمارش به شرط یک برچسب خاص انجام می‌شود می‌توان بیت‌های مربوط به برچسب را نیز از ساختار بیتی ابتکاری حذف کرد. شکل (۸) یک نمونه ساختار بیتی ابتکاری را که پس از اجرای فرآیند شمارش به شرط وجود یک برچسب خاص، بدست آمده است نشان می‌دهد.

	خصیصه ۱	خصیصه ۲	.....	خصیصه n
0000000	0101101 1011010	0101001 1001110 0101011 0010011	.....	0010100 0011001 0101100

شکل ۸: ساختار بیتی ابتکاری پس از پایان شمارش به شرط وجود یک برچسب خاص

فرآیند شمارش به ازای هر مقدار ممکن برچسب L، اجرا می‌شود و سازنده‌ی مدل با استفاده از رشته بیت‌های متناظر با مقادیر ممکن در خصیصه‌ها در ساختار بیتی ابتکاری، تعداد هر مقدار ممکن x_i برای یک خصیصه را به شرط وجود برچسب L، بدست می‌آورد و با تقسیم آن بر تعداد رکوردهای دارای برچسب L که قبلاً در محاسبه‌ی احتمال‌های برچسب‌ها بدست آمده است، می‌تواند احتمال‌های شرطی $P(x_i|C_L)$ را حساب کند.

۵-۵-۵- الگوریتم ساخت مدل کلاس‌بندی Naïve Bayes با حفظ حریم خصوصی

در الگوریتم(۳) نحوه‌ی ساخت مدل کلاس‌بندی Naïve Bayes با حفظ حریم خصوصی در طرح پیشنهادی نشان داده شده است.

Algorithm(3): Naïve Bayes model with privacy preserving
// Counting with privacy preserving
Input:
A, total number of vectors
Output:
// Generate public key and private key of model creator
PK, SK = generate_keypair();
// Generate sign key of each contributor
for each contributor *i* **do**
S_i = generate_signkey();
end for
$b_{v_1} + b_{v_2} + \cdots + b_{v_m} \leftarrow$ Run counting algorithm with innovative bit structure for labels;
for each label *L* **do**
Extract the number of label *L* from $b_{v_1} + b_{v_2} + \cdots + b_{v_m}$;
Calculate the $P(C_L)$ for each label *L*;
end for
for each label *L* **do**
$b_{v_1} + b_{v_2} + \cdots + b_{v_m} \leftarrow$ Run counting algorithm for conditional probabilities with the condition labeled *L*;
for each attribute x_i **do**
for each possible value of x_i **do**
Extract the number of vectors with labeled *L* and possible value of x_i, from $b_{v_1} + b_{v_2} + \cdots + b_{v_m}$;
Calculate the $P(x_i|C_L)$ for possible value of x_i with the condition labeled *L*;
end for
end for
end for

الگوریتم ۳: ساخت مدل کلاس‌بندی Naïve Bayes با حفظ حریم خصوصی

۶- تجزیه و تحلیل امنیتی

در ادامه طرح پیشنهادی از دو منظر حریم خصوصی مالکیت[۳۹] و حریم خصوصی آماری[۴۰] مورد بررسی قرار می‌گیرد. حریم خصوصی مالکیت، به عدم افشای داده‌های مالکین داده‌ها توجه دارد، در صورتی که حریم خصوصی آماری، بر محرمانگی مدل و اطلاعات آماری آن توجه دارد.

۶-۱- حریم خصوصی مالکیت

در طرح پیشنهادی، هر مشارکت‌کننده داده‌ای رمزنگاری شده دریافت می‌کند که نمی‌تواند محتویات آن را رمزگشائی کند. مهاجم‌های احتمالی نیز در صورت بدست آوردن داده‌های رمزنگاری شده، نمی‌توانند به محتویات آن‌ها دسترسی پیدا کنند. همچنین به علت وجود امضاءهای مشارکت‌کنندگان، امکان اضافه‌کردن داده‌ی کاذب توسط مهاجم‌ها وجود ندارد. به علت استفاده از اعداد تصادفی، داده‌ی مشارکت‌کنندگان در مراحل میانی نیز از سازنده‌ی مدل پنهان می‌ماند و با هر مشارکت و انجام هر عمل جمع، بر مخفی بودن داده‌ها افزوده می‌شود. اگر سازنده‌ی مدل قصد بررسی داده‌های میانی مشارکت‌کنندگان را داشته باشد فقط به داده‌های میانی جمع شده با یکدیگر که با اعداد تصادفی نیز جمع شده‌اند دست می‌یابد. همچنین تصادفی بودن انتخاب مشارکت‌کننده‌ی بعدی مانع از کنجکاوی در محتوای داده‌ها می‌شود. به دلیل تصادفی بودن انتخاب مشارکت‌کننده‌ی بعدی، این کنجکاوی فقط زمانی می‌تواند خطرساز باشد که در هر دو بار مشارکت اجرا شده در فرآیند یک شمارش، اولین مشارکت‌کننده یکسان باشد و اختلاف داده‌ی آن با عدد تصادفی‌اش بررسی شود. هرچند احتمال وقوع این حالت بسیار کم است ولی با منع یکسان بودن مشارکت‌کننده‌ی آغازین در دو مشارکت اجرا شده در فرآیند یک شمارش، از احتمال نشت اطلاعات جلوگیری می‌شود. از طرفی برای اطمینان از کنجکاوی‌های بیشتر، می‌توان محدودیت تعداد تکراری نبودن مشارکت‌کنندگان آغازین را بیشتر کرد؛ زیرا با هر عمل جمع، اعداد بیشتر با هم ترکیب می‌شوند و تجزیه و تحلیل آن‌ها به نیت سوء مشکل‌تر می‌شود.

۶-۲- حریم خصوصی آماری

در طرح پیشنهادی، محاسبات نهائی برای بدست آوردن داده‌های آماری همگی توسط سازنده‌ی مدل انجام می‌شود و راهی برای بدست آوردن آن‌ها وجود ندارد. داده‌های میانی نیز توسط رمزنگاری و اعداد تصادفی پنهان‌سازی شده‌اند. همچنین به علت وجود امضاءهای مشارکت‌کنندگان، امکان اضافه‌کردن داده‌ی کاذب و بررسی نتیجه‌ی آن توسط مهاجم‌ها وجود ندارد.

۷- پیاده‌سازی و ارزیابی

کارائی، یک معیار مهم در مدل‌های داده‌کاوی است. طرح پیشنهادی در شرایط مختلفی پیاده‌سازی و ارزیابی شده است تا کارائی آن سنجیده شود. پیاده‌سازی و ارزیابی بر روی چند مجموعه داده[۴۱] انجام شده و با طرح‌های مشابه نیز مقایسه شده است. بسیاری از طرح‌ها در حیطه‌ی کلاس‌بندی با حفظ حریم خصوصی، به حفظ حریم خصوصی رکوردی که برای آن کلاس‌بندی انجام می‌شود، پرداخته‌اند و مقایسه‌ی آن‌ها با طرح پیشنهادی درست نیست. برای مقایسه نیاز است که طرح پیشنهادی با طرح‌های مشابه در حیطه‌ی ساخت مدل مقایسه شود؛ اما مشکلاتی در این نوع مقایسه وجود دارد. طرح‌های ارائه شده در این حیطه کم است و تعداد معدود ارائه شده نیز بسیار ناهمگن هستند. برخی از طرح‌ها برروی سبک خاصی

از کلاس‌بندی تمرکز دارند که با کلاس‌بندی ما متفاوت است. به عنوان مثال طرح Yi Liu و همکارانش[۱۹] برکلاس‌بندی چند برچسبه تمرکز دارد یا طرح ClaMPP[۱۳] برای فیلترینگ مشارکتی مبتنی بر کاربر طراحی شده است. برخی از طرح‌ها نیز صرفاً به صورت تئوری به این موضوع پرداخته‌اند به عنوان مثال POCC[۱۲] صرفاً به جمع‌آوری داده‌ها به صورت رمزنگاری شده پرداخته است و یک راهکار مشخص برای انجام کلاس‌بندی برروی داده‌ی رمز شده بدون رمزگشائی آن ارائه نکرده است. برای رفع این مشکل، از آنجا که بیشتر مقایسه‌های مربوط به کارائی تحت‌تأثیر رمزنگاری است و تاثیر عوامل دیگر در مقایسه با رمزنگاری بسیار ناچیز است، ساخت مدل کلاس‌بندی Naïve Bayes ساده با شمارش رمزنگاری شده نیز انجام شده است که در واقع این پیاده‌سازی، یک پیاده‌سازی با حفظ حریم خصوصی نیست بلکه در این پیاده‌سازی سعی شده است حداقل میزان کارائی مورد نیاز برای یک طرح که از رمزنگاری شمارش‌ها استفاده می‌کند، بدست آید. در واقع هر طرحی که از ساختار داده‌ی متفاوت استفاده نمی‌کند و تنها با اتکا به شمارش مدل را می‌سازد، حداقل به تعداد اجرای عملیات رمزنگاری که در این پیاده‌سازی وجود دارد، نیاز به اجرای عملیات رمزنگاری خواهد داشت. بنابراین می‌توان به مقایسه‌ی خوبی برای کارائی طرح نسبت به طرح‌های مشابه که وجود دارند یا در آینده ارائه خواهند شد، رسید. یادگیری Naïve Bayes خصوصی افتراقی[۴۲] [۱۰] نیز با تغییراتی قابلیت مقایسه با طرح پیشنهادی را خواهد داشت و مقایسه‌ها با این طرح نیز انجام شده است. در پیاده‌سازی‌ها از رمزنگاری پایلیر[۴۳] که یک رمزنگاری هوموفیک جمعی مطرح است، استفاده شده است. البته برای طرح پیشنهادی می‌توان از انواع رمزنگاری‌های هوموفیک دیگر که از جمع پشتیبانی می‌کنند، نیز استفاده کرد ولی طرح یادگیری Naïve Bayes خصوصی افتراقی با رمزنگاری پایلیر ارائه شده است.

۷-۱- جزئیات پیاده‌سازی و محیط اجرا

برای پیاده‌سازی از زبان پایتون استفاده شده است. همچنین برای پیاده‌سازی رمزنگاری پایلیر از کتابخانه‌ی Python-Paillier[۳۱] استفاده شده است. اجراها با رمزنگاری‌های ۲۰۴۸ بیتی و ۲۵۶ بیتی انجام شده است.

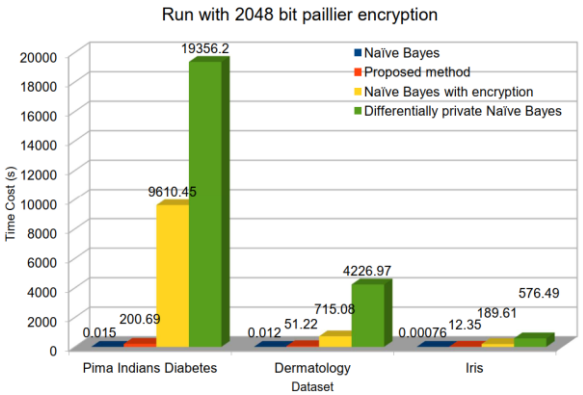
در پیاده‌سازی طرح یادگیری Naïve Bayes خصوصی افتراقی و همکارانش، از آنجا که در این طرح خود رمزنگاری نیز برای اضافه-کردن نویز دستکاری می‌شود و از آنجا که این امر منجر به یکسان نبودن رمزنگاری و کاهش اعتبار مقایسه می‌شود، به جای نویزها اعداد با جمع ۱ قرار داده شده است و به جای نویزهای افتراقی رمزنگاری شده، از رمزنگاری شده‌ی عدد صفر استفاده شده است که تفاوتی در کارائی ایجاد نمی‌کند و هزینه‌ی زمانی اجرا در طرح پیاده‌سازی شده کوچکتر یا مساوی طرح اصلی خواهد بود.

برای اجراها، کامپیوتری دارای 8GiB رم و پردازنده‌ی Intel Core i5-4460 که دارای قدرت پردازش فرکانسی 3.2 گیگاهرتز است، استفاده شده است. همچنین اجراها در محیط سیستم عامل ۶۴ بیتی دبیان نسخه‌ی ۳.۱۰ و با مفسر پایتون نسخه‌ی ۴.۷.۳ انجام شده است.

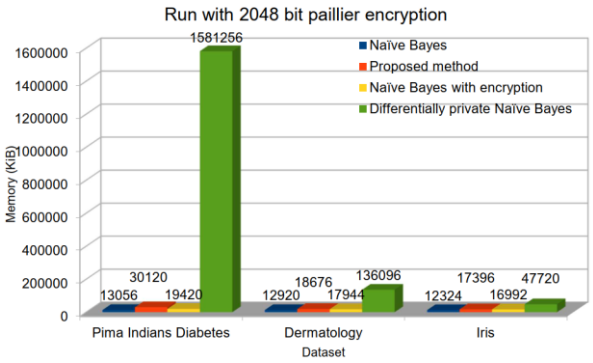
برای تست اجرا از مجموعه داده‌های Pima Indians Diabetes Database[۲۸] و Dermatology[۲۹] و Iris[۳۰] استفاده شده است. مجموعه داده‌ی Pima Indians Diabetes Database دارای ۷۶۸ رکورد و ۸ نوع خصیصه است و در مجموع، خصیصه‌های آن ۱۲۵۴ مقدار ممکن را می‌پذیرند. از آنجا که مقادیر قابل پذیرش برای خصیصه‌های این مجموعه داده زیاد است می‌تواند برای تست طرح پیشنهادی با ساختار داده‌ی ابتکاری که رشته بیت‌های آن بلند است، مورد استفاده قرار گیرد. مجموعه داده‌ی Dermatology دارای ۳۶۶ رکورد و ۳۳ نوع خصیصه است و در مجموع، خصیصه‌های آن ۱۸۹ مقدار ممکن را می‌پذیرند که تعداد خصیصه‌های زیاد ولی مقادیر قابل پذیرش آن‌ها نرمال است. مجموعه داده‌ی Iris دارای ۱۵۰ رکورد و ۴ نوع خصیصه است و در مجموع، خصیصه‌های آن ۱۲۳ مقدار ممکن را می‌پذیرند که یک مجموعه داده‌ی کوچک است.

۷-۱-۱- ارزیابی

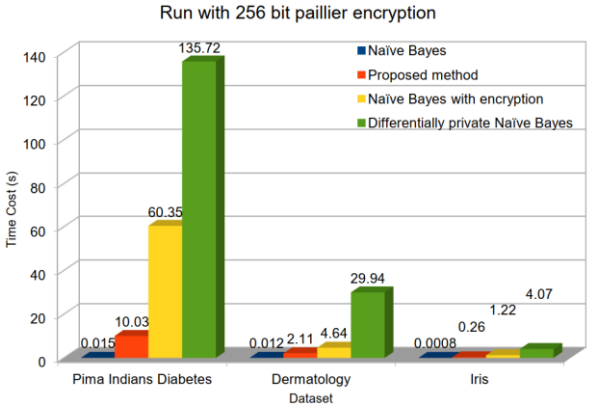
هر رکورد برای یک مشارکت‌کننده در نظر گرفته شده است، پس به تعداد رکوردها مشارکت‌کننده وجود دارد. نتیجه‌ی اجرا با رمزنگاری‌های ۲۰۴۸ بیتی و ۲۵۶ بیتی با نمودارهای زمان پردازش و حافظه‌ی مصرفی در شکل‌های (۹) و (۱۰) و (۱۱) و (۱۲) نشان داده شده است.



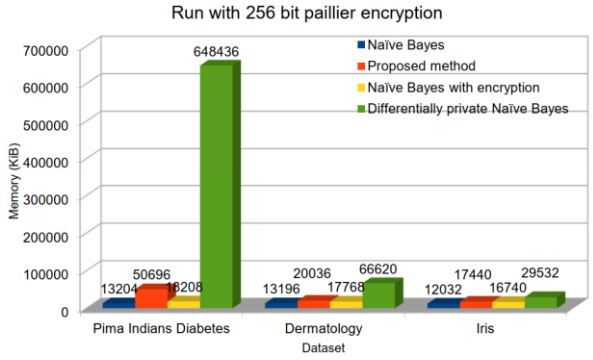
شکل ۹: هزینه‌ی زمانی با رمزنگاری ۲۰۴۸ بیتی



شکل ۱۰: مصرف حافظه با رمزنگاری ۲۰۴۸ بیتی



شکل ۱۱: هزینه‌ی زمانی با رمزنگاری ۲۵۶ بیتی



شکل ۱۲: مصرف حافظه با رمزنگاری ۲۵۶ بیتی

مطابق شکل‌های (۹) و (۱۰) و (۱۱) و (۱۲)، یادگیری Naïve Bayes خصوصی افتراقی دارای پیچیدگی زمانی و حافظه‌ی بسیار زیادی است. در واقع در طرح یادگیری Naïve Bayes خصوصی افتراقی به علت وجود شخص سوم که باعث رمزنگاری چندباره‌ی اعداد می‌شود و استفاده از نویز، تعداد اجرای عملیات رمزنگاری افزایش پیدا می‌کند و از آنجا که طول بیت داده‌ای که رمزنگاری شده است، فارق از مقدار عددی آن یکسان است، افزایش تعداد اجرای عملیات رمزنگاری منجر به افزایش حافظه‌ی مصرفی می‌شود. همچنین افزایش مقادیر ممکن برای خصیصه‌ها بر این طرح اثر نامطلوب می‌گذارد.

طول بیت داده‌ای که رمزنگاری شده است، جدا از مقدار عددی آن یکسان است. بنابراین هرچه عدد بزرگتری در داده‌ی رمزنگاری شده قرار داده شود، تعداد اجرای عملیات رمزنگاری کاهش پیدا می‌کند. طرح پیشنهادی به علت قرار دادن بزرگترین رشته بیت ممکن در هر رمزنگاری، تعداد اجرای عملیات رمزنگاری را کاهش می‌دهد و از طرفی از مصرف حافظه‌ی زیاد برای اعداد رمزنگاری شده‌ی کوچک جلوگیری می‌کند. برتری کارائی طرح پیشنهادی از نظر هزینه‌ی زمانی و حافظه مصرفی در شکل‌های (۹) و (۱۰) و (۱۱) و (۱۲) مشاهده می‌شود. با مقایسه‌ی شکل‌های (۱۰) و (۱۲) مشاهده می‌شود که با کاهش طول بیت رمزنگاری به ۲۵۶ بیتی و همچنین افزایش مقادیر ممکن برای خصیصه‌ها، حافظه‌ی مصرفی طرح پیشنهادی تا حدود ۵۰٪ درصد افزایش پیدا می‌کند. این افزایش به علت نیاز به تقسیم‌بندی بیشتر ساختارهای بیتی ابتکاری است که با کاهش طول بیت قابل رمزنگاری اتفاق می‌افتد. از این رو استفاده از طرح پیشنهادی برای مواقعی که نیاز به رمزنگاری‌های با طول بیت زیاد است نیز مفید است.

مقایسه‌ی نتایج ارزیابی با Naïve Bayes رمزنگاری شده، نشان می‌دهد که طرح پیشنهادی می‌تواند از بسیاری از طرح‌های مشابه کارائی بهتری از نظر پیچیدگی زمانی و حافظه‌ی مصرفی داشته باشد، زیرا با حداقل تعداد اجرای عملیات رمزنگاری در Naïve Bayes رمزنگاری شده، کماکان طرح پیشنهادی برتری کارائی ۸۷٪ را از نظر هزینه‌ی زمانی نشان می‌دهد و از نظر حافظه‌ی مصرفی نیز هزینه‌ی اضافی زیادی تحمیل نمی‌کند و برای رمزنگاری ۲۰۴۸ بیتی افزایشی در حدود ۶٪ الی ۳۰٪ دارد. در صورتی که بسیاری از طرح‌ها حافظه‌ی مصرفی را بسیار افزایش می‌دهند، مانند یادگیری Naïve Bayes خصوصی افتراقی که حافظه‌ی مصرفی را در حدود ۲۰۰٪ الی ۸۵۰۰٪ برابر افزایش می‌دهند.

۸- نتیجه‌گیری

در تحقیق و پژوهش صورت گرفته، طرحی جدید برای ساخت مدل کلاس‌بندی Naïve Bayes با حفظ محرمانگی و حریم خصوصی بدست آمد. این طرح براساس سازوکار شمارش رمزنگاری شده، کار می‌کند و تلاش می‌شود با ایجاد یک بستر امن بر اعتماد مالکین داده‌ها افزوده شود. طرح پیشنهادی می‌تواند مدل را بدون از دست رفتن دقت و بدون نیاز به اعتماد به شخص سوم بسازد. با استفاده از طرح پیشنهادی می‌توان مدل را با مشارکت مالکین داده‌ها به تعداد دفعات زیاد بازسازی کرد. از طرفی به دلیل اینکه پردازش، بین مشارکت‌کنندگان تقسیم می‌شود، هزینه‌ی محاسباتی سازنده‌ی مدل نیز به شدت کاهش می‌یابد. کارائی یک معیار مهم در مدل‌های

داده‌کاوی است. حفظ محرمانگی و حریم خصوصی در داده‌کاوی معمولا با از دست رفتن یکی از معیارهای کارائی و دقت انجام می‌شود. نتایج نشان می‌دهد طرح پیشنهادی، علاوه بر حفظ دقت، کارائی قابل قبولی را نیز از نظر پیچیدگی زمانی و حافظه‌ی مصرفی، ارائه می‌کند. به‌طوری‌که از نظر پیچیدگی زمانی تا ۸۷٪ بهبود در هزینه‌ی زمانی مشاهده می‌شود و حافظه‌ی مصرفی نیز افزایش چندانی نسبت به طرح‌های دارای عملیات رمزنگاری نداشته است.

امروزه با افزایش توان سامانه‌های محاسباتی، امنیت رمزنگاری‌های با طول بیت کوتاه، به مرور کاسته می‌شود. از این جهت طرح پیشنهادی می‌تواند به عنوان ایده‌ای که با افزایش طول بیت رمزنگاری، کمتر دچار چالش می‌شود، مورد توجه قرار گیرد.

مراجع

[1] Ricardo Mendes, João P. Vilela, “Privacy-Preserving Data Mining: Methods, Metrics, and Applications”, IEEE Access, Volume 5, Pages 10562 - 10582, June 2017.

[2] Yousra Abdul Alsahib S.Aldeen, Mazleena Salleh and Mohammad Abdur Razzaque, “A comprehensive review on privacy preserving data mining”, SpringerPlus, Volume 4, Pages 1-36, November 2015.

[3] Ke Wang, Philip S. Yu, Sourav Chakraborty, “Bottom-up generalization: a data mining solution to privacy protection”, Fourth IEEE International Conference on Data Mining (ICDM'04), 2004.

[4] Benjamin C. M. Fung, Ke Wang, Philip S. Yu, “Top-down specialization for information and privacy preservation”, 21st IEEE International Conference on Data Engineering (ICDE'05), 2005.

[5] Alexander Wood, Vladimir Shpilrain, Kayvan Najarian, Delaram Kahrobaei, “Private naive bayes classification of personal biomedical data: application in cancer data analysis”, Computers in Biology and Medicine, Volume 105, Pages 144-150, February 2019.

[6] Kai Xing, Chunqiang Hu, Jiguo Yu, Xiuzhen Cheng, Fengjuan Zhang, “Mutual Privacy Preserving k-Means Clustering in Social Participatory Sensing IEEE Transactions on Industrial Informatics, Volume 13, Pages 2066-2076, August 2017.

[7] Xiaoxia Liu, Hui Zhu, Rongxing Lu, Hui Li, “Efficient privacy-preserving online medical primary diagnosis scheme on naive bayesian classification”, Peer-to-Peer Networking and Applications, Volume 11, Pages 334-347, March 2018.

[8] R Bost, Raluca Ada Popa, Stephen Tu, Shafi Goldwasser, “Machine learning classification over encrypted data”, Network and Distributed System Security Symposium (NDSS), 2015.

[9] Alexey Gribov, Delaram Kahrobaei, Vladimir Shpilrain, “Private-keyfully, homomorphic encryption in rings”, Groups Complexity Cryptology, Volume 10, Pages 17-27, March 2018.

[10] Tong Li, Jin Li, Zheli Liu, Ping Li , Chunfu Jia, “Differentially private Naive Bayes learning over multiple data sources”, Information Sciences, Volume 444, Pages 89-104, May 2018.

[11] Cynthia Dwork, Aaron Roth, “The algorithmic foundations of differential privacy”. Foundations and Trends® in Theoretical Computer Science, Volume 9, Pages 211-407, August 2014.

[12] Ping Li, Jin Li, Zhengan Huang, Chong-Zhi Gao, Wen-Bin Chen, Kai Chen, “Privacy-preserving outsourced classification in cloud computing”, Cluster Computing, Volume 21, Pages 277-286, March 2018.

[13] Harmanjeet Kaur, Neeraj Kumar, Shalini Batra, “ClaMPP: a cloud-based multi-party privacy preserving classification scheme for distributed applications”, The Journal of Supercomputing, Volume 75, Pages 3046-3075, June 2019.

[14] Radhika Kotecha, Sanjay Garg, “Preserving output-privacy in data stream classification”, Progress in Artificial Intelligence, Volume 6, Pages 87-104, June 2017.

[15] Pierangela Samarati, Latanya Sweeney, “Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression”, IEEE Symposium on Research in Security and Privacy, Pages 188–206, 1998.

[16] Pedro Domingos, Geoff Hulten, “Mining high-speed data streams”, Proceedings of 6th ACM International Conference on Knowledge Discovery and Data Mining, August 2000.

[17] Raphael Bost, Raluca Ada Popa, Stephen Tu, Shafi Goldwasser, “Machine learning classification over encrypted data”, 22nd Annual Network and Distributed System Security Symposium (NDSS), 2015.

[18] Chong-zhi Gao, Qiong Cheng, Pei He, Willy Susilo, Jin Li, “Privacy-preserving Naive Bayes classifiers secure against the substitution-then-comparison attack”, Information Sciences, Volume 444, Pages 72-88, May 2018.

[19] Yi Liu, Yu Luo, Youwen Zhu, Yang Liu, Xingxin Li, “Secure multi-label data classification in cloud by additionally homomorphic encryption”, Information Sciences, Volume 468, Pages 89-102, November 2018.

[20] Min-Ling Zhang, Zhi-Hua Zhou, “A k-nearest neighbor based algorithm for multi-label classification”, IEEE International Conference on Granular Computing, 2005.

[21] Yiran Shen, Chengwen Luo, Dan Yin, Hongkai Wen, Rus Daniela, Wen Hu, “Privacy-preserving sparse representation classification in cloud-enabled mobile applications”, Computer Networks, Volume 133, Pages 59-72, March 2018.

[22] Kai Schramm, Gregor Leander, Patrick Felke, Christof Paar, “A collision-attack on aes”, Cryptographic Hardware and Embedded Systems-CHES, Springer, Volume 3156, Pages 163–175, 2004.

[23] Kai Xing, Chunqiang Hu, Jiguo Yu, Xiuzhen Cheng, Fengjuan Zhang, “Mutual privacy preserving k-means clustering in social participatory sensing”, IEEE Transactions on Industrial Informatics, Volume 13, Pages 2066-2076, Aug 2017.

[24] Qingchen Zhang, Hua Zhong, Laurence T. Yang, Zhikui Chen, Fanyu Bu, “PPHOCFS: privacy preserving high-order CFS algorithm on the cloud for clustering multimedia data”, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), Volume 12, Pages 66-80, November 2016.

[25] Xiaoqian Liu, Qianmu Li, Tao Li, Dong Chen, “Differentially private classification with decision tree ensemble”, Applied Soft Computing, Volume 62, Pages 807-816, January 2018.

[26] Jiawei Han, Micheline Kamber, Jian Pei, “Data Mining: Concepts and Techniques, Third Edition”, Elsevier, June 2011.

[27] Wikipedia, the free encyclopedia, “Homomorphic encryption”, September 2021, url https://en.wikipedia.org/wiki/Homomorphic_encryption.

[28] National Institute of Diabetes and Digestive and Kidney Diseases, “ PIMA Indians diabetes dataset”, License: CC: Public Domain, 1990, url https://www.kaggle.com/uciml/pima-indians-diabetes-database

[29] Nilsel Ilter, H Altay Guvenir, “Dermatology dataset”, License: Open Data Commons, 1998, url https://archive.ics.uci.edu/ml/datasets/dermatology

[30] R.A. Fisher, Michael Marshall, “Iris dataset”, License: Open Data Commons, 1988, url https://archive.ics.uci.edu/ml/datasets/Iris

[31] Python-paillier, "A Python 3 library for Partially Homomorphic Encryption using the Paillier crypto system", License: GPL v3, 2021, url https://python-paillier.readthedocs.io

[32] M.A.P. Chamikara, Peter Bertok, Dongxi Liu, Seyit Camtepe, Ibrahim Khalil, "Efficient privacy preservation of big data for accurate data mining", Information Sciences, Volume 527, Pages 420-443, July 2020.

[33] Duy-Hien Vu, "Privacy-Preserving Naive Bayes Classification in Semi-Fully Distributed Data Model", Computers & Security, Volume 115, Article 102630, April 2022.

[34] Ngoc Hong Tran, Nhien-AnLe-Khac, M-Tahar Kechadi, "Lightweight privacy-Preserving data classification", Computers & Security, Volume 97, Article 101835, April 2020.

[35] Jing Wang, Libing Wu, Sherali Zeadally, Muhammad Khurram Khan, Debiao He, "Privacy-preserving Data Aggregation against Malicious Data Mining Attack for IoT-enabled Smart Grid", ACM Transactions on Sensor Networks, Volume 17, Pages 313-338, August 2021.

پاورقی‌ها:

[1] Third Party
[2] Classification
[3] Homomorphic Encryption
[4] Privacy-preserving Medical Primary Diagnosis
[5] Lightweight Polynomial Aggregation Technique
[6] Substitution-then-Comparison
[7] K-anonymity
[8] l-diversity
[9] Generalization
[10] Specialization
[11] HOeffding Tree
[12] Stream
[13] Laplace Mechanism
[14] Differential Privacy
[15] Decision Tree
[16] Cloud Computing
[17] Privacy-preserving Outsourced Classification in Cloud Computing
[18] Fully Homomorphic
[19] Cloud-based Multi-party Privacy Preserving Classification Scheme for Distributed Applications
[20] Collaborative Filtering
[21] Cloud
[22] Multi Label
[23] k-nearest neighbors
[24] Additively Homomorphic
[25] Cloud-enabled Mobile Applications
[26] Sparse Representation
[27] Optimal Geometric Transformations
[28] IoT-enabled smart grids
[29] Train
[30] Partially Homomorphic
[31] Additively Homomorphic
[32] Multiplicatively Homomorphic
[33] Training Set
[34] Entities
[35] Honest-but-Curious
[36] Initialization Phase
[37] Aggregation Phase
[38] Carry
[39] Ownership Privacy
[40] Statistics Privacy
[41] Dataset
[42] Differential
[43] Paillier