تاریخ دریافت مقاله: ۹۶/۰۴/۲۹
تاریخ پذیرش مقاله: ۹۶/۰۹/۱۲

روش بی‌نام‌سازی داده‌های حساس با حفظ سودمندی مبتنی بر مرتب‌سازی داده‌ها

محمد طه عسکری*
کارشناسی ارشد نرم‌افزار – دانشکده فنی و مهندسی- دانشگاه آزاد دامغان- ایران
پست الکترونیکی: taha.askaree@gmail.com

رضا مرتضوی
استادیار – دانشکده فنی و مهندسی – دانشگاه دامغان – دامغان – ایران
پست الکترونیکی: r_mortazavi@du.ac.ir

چکیده:

ریزتجمیع خانواده‌ای از روش‌های کنترل افشای آماری (SDC) از ریزداده‌ها (رکوردهای اشخاص و یا شرکت‌ها) برای پوشش ریزداده‌ها است. داده‌های منتشر شده باید ضمن حفظ سودمندی، حریم خصوصی صاحبان داده‌ها را حفظ کند. بنابراین یک توازن بین سودمندی داده‌ها و حریم خصوصی وجود دارد. در این مقاله دو الگوریتم ریزتجمیع پیشنهاد شده است که NFPN++ و ENFPN نامیده می‌شود که ابتدا داده‌ها را با توجه به رکورد آخر و رکورد ماقبل آخر و پنج رکورد آخر مرتب می‌سازد، سپس یک بخش‌بندی با کمترین اتلاف سودمندی نسبت به داده‌های مرتب شده جستجو می‌کند. نتایج تجربی نشان می‌دهد که روش پیشنهادی به کمترین اتلاف اطلاعات نسبت به روش‌های ریزتجمیع گذشته دست یافته است و توازن بهتر بین سودمندی داده‌ها و حریم خصوصی را میسر ساخته است.

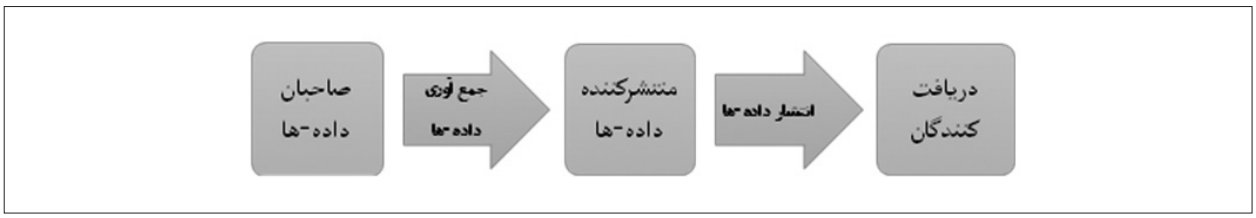
واژه‌های کلیدی: حریم خصوصی، کنترل افشای آماری، ریزتجمیع داده‌گرا، k-بی‌نامی، اتلاف اطلاعات

۱- مقدمه

پیشرفت‌های مداوم در فناوری‌های رایانه‌ای شرکت‌ها را قادر می‌سازد تا حجم عظیمی از داده‌های شخصی را جمع‌آوری نمایند، که منابع ارزشمند برای پژوهشگران و شرکت‌های دولتی می‌باشند [۱]. اسکمن حریم خصوصی را به‌عنوان «حق فرد برای تعیین این‌که چه نوع اطلاعات (شخصی) در مورد وی در حال تبادل است» یا «کنترل یک فرد روی اطلاعات شخصی‌اش» تعریف می‌کند [۲]. PPDP[1] یک ناحیه پژوهشی است که تلاش می‌کند داده‌ها را قبل از انتشار برای حفاظت اطلاعات حساس تغییر دهد. بین حریم خصوصی و سودمندی داده‌ها در PPDP توازن وجود دارد

1- Privacy Preserving Data Publishing

* نویسندهٔ مسئول



شکل ۱: فرایند انتشار داده‌ها، بر گرفته از [۱۱]

[۳]. تغییر یک پایــگاه داده با کاهش کیفیت باید به گونه‌ای باشــد که اطلاعات حساس افشا نشــوند؛ تحلیل داده‌های تغییر یافته شبیه به تحلیل داده‌های اصلی است [٤].

داده‌های منتشر شده باید بی‌نام[۲] باشند درحالی‌که طبق حریم خصوصی داده‌ها حفظ شده هستند. بنابراین قبل از انتشــار چنین داده‌هایی باید چند دستورالعمل برای ایجاد داده‌های بی‌نام اعمال شــود. چندین مدل و سازوکار برای نیازمندی‌هــای این بی‌نامی[۳] در پایــگاه داده‌ها، داده‌کاوی[٤] [٥،٦] و مقالات آماری پیشنهاد شده است [۷-۹]. این حفاظت که در ازای کاهش سودمندی داده‌های محافظت شده برای کاربران مجاز نظیر پژوهشگران به‌دست آمده است، آن‌ها را مجبور به اســتفاده از روش‌های خاص برای استخراج دانش از داده‌های منتشر شده می‌نماید [۱۰]. شکل ۱ فرایند انتشار داده‌ها را نشان می‌دهد.

از آنجا کــه حریم خصوصی یک مفهــوم چند وجهی می‌باشد، مسئلهٔ تضمین حفاظت مناسب از حریم خصوصی اشخاص آسان نیســت: اطلاعات معین (حساس) راجع به اشخاص باید پوشیده بماند، هویت اشخاص باید محافظت شده باشد، یا اعمال اشــخاص نباید قابل ردگیری[٥] باشد. عامل دیگری که بر پیچیدگی مســئله می‌افزاید وجود منابع متعدد داده‌هاست که تحلیل آن‌ها و مرتبط ساختن آن‌ها با هم، منجر به نشت نابجای اطلاعاتی می‌شود که نباید افشا شوند [۱۲].

ریزتجمیع[٦] یــک تکنیک در SDC[۷] با کاربردهایی در علم کامپیوتر نظیر انتشــار داده‌ها با حفــظ حریم خصوصی [۱۳] و ســرویس‌هایی بر اساس مکان (LBS)[۸] [١٤] است که برای خوشــه‌بندی یک مجموعه از رکوردها در گروه‌هایی با حداقــل k رکــورد می‌باشــد [۱۵]. ریزتجمیــع یکی از بیشترین روش‌هایی است که برای رضایتمندی k-بی‌نامی[۹] مورد اســتفاده قرار می‌گیرد که شــامل دو مرحله است:

۱- بخش‌بندی داده‌های داخــل گروه‌ها با حداقل k عنصر

۲- جابجایی هر رکــورد با مرکز ثقل گروهش [۳]. چندین مهــارت از k-بی‌نامــی و متغیرهایش در مقــالات امنیت پایگاه داده پیشــنهاد شده است مانند (α , k)-بی‌نامی [١٦]، (α , L)-تنوع[۱۰] [۱۷] و (p , α)-k- بی‌نامی حساس[۱۱] [۱۸].

NFPN-MHM[۱۲] [۳] یک روش ریزتجمیع داده‌گرا اســت. این مقاله توســعه‌هایی را بر روی NFPN-MHM پیشــنهاد می‌کند که NFPN++ و Enhanced-NFPN با بیشترین صحت و کمترین اتلاف سودمندی نامیده می‌شوند.

بخش‌های این مقاله مطابق ذیل سازماندهی شده است: بخش ۲ یک نظر اجمالــی روی روش‌های ریزتجمیع دارد. بخش ۳ شــامل تعریف‌هــا و مقدمه‌های مــورد نیاز برای شــرح روش‌های پیشنهادی است. بخش ٤ ارائه روش‌های پیشــنهادی و بخــش ٥ نمایش مقایســه‌ها بین روش‌های پیشنهادی و دیگر روش‌ها و نهایتاً بخش ٦ نتیجه می‌باشد.

۲- پیشینه

این بخش چند روش ریزتجمیع را به طور خلاصه شرح می‌دهد که بر روی ریزداده‌ها اعمال می‌شوند. پیشنهاد همه روش‌ها پیدا کردن یک گروه‌بندی از داده‌ها با کمترین اتلاف

2- anonymous
3- anonymity
4- data mining
5- traceable
6- microaggregation
7- Statistical Disclosure Control

8- Location-Based Services
9- k-anonymity
10- diversity
11- sensitive k-anonymity
12- Nearest Far Point Next MHM

اطلاعات و خطر افشـــا است که سودمندی بیشتر داده‌ها و حریم خصوصی را به دنبال دارد. این مســـئله می‌تواند در زمان چندجمله‌ای بـــرای مجموعه داده‌های تک‌متغیره حل شود، اما یک مسئله NP-Hard برای مجموعه داده‌های چند متغیره می‌باشد [۱۹].

یـــک مجموعه (فایل) ریـــزداده، شـــامل مجموعه‌ای از رکوردهاســـت که هر یک حاوی اطلاعـــات یک موجودیت مشخص (مانند فرد یا سازمان) هستند، و معمولاً به صورت یک ماتریس با n ســـطر (n رکورد) و v ســـتون (مقادیر v خصیصـــه برای هر رکورد) در نظر گرفته می‌شـــود [۲۰]. در ریزداده شناســـه‌ها[۱۳] مجموعه‌ای از خصیصه‌ها هستند که حاوی اطلاعاتی اســـت که صریحاً موجودیت متعلق به رکـــورد (صاحب رکورد) را مشـــخص می‌کنند [۲۱]؛ مانند نام یا شـــماره حساب بانکی، و یا کد پرسنلی کارمندان یک سازمان. و همچنین نیمه‌شناسه‌ها[۱۴] مجموعه خصیصه‌هایی هستند که به صورت بالقوه صاحب رکورد را می‌شناسانند [۲۱]؛ بر خلاف خصیصه‌های شناسه، به تنهایی نمی‌توانند یک موجودیت را شناســـایی کنند، اما ممکن است بتوان با کنار هم قرار دادن چند خصیصهٔ نیمه‌شناسه، دقیقاً به یک موجودیت مشخص رسید [۲۰]. به‌عنوان مثال، ممکن است ســـن، کد پستی و درآمد ماهیانه افراد در کنار هم، آن‌ها را از یکدیگر به خوبی متمایز نمایند.

روش‌های ریزتجمیع به دو دستهٔ اندازه‌ثابت و داده‌گرا تقســـیم می‌شـــوند. روش‌های اندازه‌ثابت داده‌ها را درون گروه‌هایی با اندازه k تفکیک می‌کنند.

یکی از معروف‌تریـــن روش‌های ریزتجمیع اندازه‌ثابت MDAV[۱۵] [۲۲] اســـت. این روش مکرراً خوشـــه‌ها را تولید می‌کنـــد. در هـــر تکـــرار، ابتدا مرکـــز ثقـــل از رکوردهای خوشه‌بندی نشـــده را دوباره محاسبه و دو خوشه جدید تشـــکیل می‌دهد. اولین خوشه شـــامل r، دورترین رکورد از مرکز ثقل سراســـری و k-1 نزدیک‌ترین همســـایه‌اش و دومین آن شامل s، دورترین رکورد از r و k-1 نزدیک‌ترین همسایه‌ها از s می‌باشد. نهایتاً عضوهای خوشه‌های جدید از لیســـت را حذف می‌کند. این تکرارها تـــا کمتر از m<2k رکوردهای خوشـــه‌بندی نشـــده باقیمانده ادامه دارد. اگر m>k-1 باشد همه رکوردهای باقیمانده یک خوشه جدید را تشکیل می‌دهند یا در غیر این صورت هر رکورد باقیمانده به نزدیک‌ترین خوشه‌اش متصل می‌شود.

روش دیگر اندازه‌ثابت شناخته شده CBFS[۱۶] [۲۳] است. CBFS شبیه MDAV اســـت اما یک‌بار مرکز ثقل را محاسبه می‌کند و یک خوشـــه جدید در هر تکرار تشـــکیل می‌دهد. این روش ســـریع‌تر از MDAV اســـت چون مرکز ثقل فقط یک‌بار محاسبه می‌شود. ساختمان داده‌های درخت KD در KD-CBFS [۲۴] بـــا شـــتاب دادن بـــه رویـــه جســـتجوی k نزدیک‌ترین همســـایه‌ها مورد استفاده است. این پیچیدگی CBFS را از $O(n^2)$ به $O(n.\log n)$ بهبود می‌دهد و قابلیت اجرا برای حجم داده‌های بزرگ را ایجاد می‌کند.

یک نســـخه از MDAV پیشنهاد شده اســـت که MDAV-generic [۲۵] نامیده می‌شـــود. آســـتانه پایانی تکرارهای MDAV عمومی با 3k تنظیم شده است.

در مقابل، روش‌های داده‌گرا خوشه‌های با اندازهٔ متغیر را تولید می‌کنند. نســـخهٔ داده‌گرای MDAV، V-MDAV [۱۵] است. بعد از ساخت هر خوشه با اندازه k، V- MDAV اندازهٔ خوشـــه را تا 2k-1 رکورد گسترش می‌دهد. e_{min} یک رکورد خوشـــه‌بندی نشده را مشخص می‌کند که حداقل فاصله را از نزدیک‌ترین عضو خوشه‌بندی شده‌اش دارد. این فاصله، d_{in} و فاصلـــه بین e_{min} و نزدیک‌ترین رکورد خوشـــه‌بندی نشـــده‌اش، d_{out} می‌باشـــد. رکورد e_{min} به خوشـــه اضافه می‌شود، اگر $din<\gamma d_{out}$ باشد که γ یک آستانه تعریف شده توسط کاربر است.

روش دیگر GSMS[۱۷] [۲۶] اســـت. ایـــن روش دارای دو مرحله می‌باشد، در مرحله اول $\left\lfloor \frac{N}{k} \right\rfloor$ تکرار انجام می‌شود که در هر تکرار خوشه نامزدی (خوشه‌ای شامل رکورد نامزد برای مرکز خوشه بودن و k-1 نزدیک‌ترین همسایه‌اش) که

13- identifiers
14- quasi-identifiers
15- Maximum Distance to Average Vector

16- Centroid-Based Fixed-Size
17- Group Selection Based on Sequential Minimization of SSE

باعث کمینه شدن SSE[18]جاری از رکوردهای باقیمانده شود را رد می‌کند. در پایان این مرحله، $\left\lfloor \frac{N}{k} \right\rfloor$ خوشـــه k عضوی تشـــکیل می‌گردد. در مرحله دوم تعـــداد اندک رکوردهای باقیمانده را به نزدیک‌ترین خوشه‌ها منتسب می‌نماید.

MHM[19] [۲۷] یک راه حل بهینه را در زمان چند جمله‌ای برای مجموعه داده‌های تک‌متغیره میســـر می‌ســـازد. این روش همه داده‌ها را در یک ترتیب صعودی مرتب می‌کند، سپس خوشه‌های با مقدار SSE کمینه برای داده‌های مرتب شده را جســـتجو می‌کند. جزئیات بیشتر این الگوریتم در بخش ٤ شرح داده شده اســـت. MHM فقط برای مجموعه داده‌های تک‌متغیره بهینه است، اگرچه روش‌های اکتشافی گوناگونی وجود دارند که مرتب‌سازی داده‌های چندمتغیره را پیشنهاد می‌کنند که با اعمال MHM یک قسمت‌بندی بهینه با رابطه مرتب‌سازی داده شـــده ارائه می‌نماید (لزوماً این معادل با بخش‌بندی بهینه مجموعه داده‌ها نیست). به‌عنوان مثال ژینیتا و همکاران [۲۸] یک روش بر اساس منحنی‌های فضا–پرکن[20] هیلبرت پیشـــنهاد کرده‌اند تا همه رکوردهای داده‌ها در یک توالی مرتب شـــود. این روش خیلی ســـریع است اما مجموعه داده‌های محافظت شده با اتلاف اطلاعات بالا را تولید می‌کند.

IMHM[21] [۱] یک پیاده‌ســـازی کارا از MHM اســـت که وزن‌های لبه را به صورت تکراری محاسبه می‌کند، بنابراین O(nk) عملیات حسابی نیاز دارد.

روش دیگری که MHM اســـتفاده می‌کند مرتب‌ســـازی داده‌ها با روش نزدیک‌ترین نقطه بعدی (NPN)[22] [۲۹] است. NPN یک توالی از رکوردها، در شروع با انتخاب دورترین رکـــورد از مرکز ثقل را تولید می‌کنـــد. در هر مرحله اگر r انتهای توالی باشـــد، نزدیک‌ترین رکورد به r به آن اضافه می‌شود، تا همه رکوردها به توالی اضافه شوند. ساختمان توالی بـــا پیـــروی از نزدیک‌ترین رکوردهـــا می‌تواند یک الگوریتم NPN ســـردرگم را تولید کند و یک رکورد دور را به انتهای توالی اضافه می‌کند.

NFPN-MHM [۳] این مســـئله را نشانی‌دهی می‌کند. این روش با پیچیدگی زمانی یکسان با NPN اجرا می‌شود.

این مقاله گســـترش‌هایی را روی NFPN-MHM پیشنهاد می‌کند. روش‌های پیشنهادی جدید معمولاً سودمندی بیشتر و مجموعه داده‌های محافظت شده را تولید می‌نمایند.

۳- تعریف‌ها

یک مجموعه داده عددی T نرمال شـــده شامل n رکورد موجود اســـت. هر رکـــورد d خصوصیت عـــددی دارد و می‌تواند به‌عنوان یک نقطه در فضای d- بعدی ارائه شـــود. برای هر رکـــورد داده x_i، $1<i<n$، x_{ij} jامین خصوصیت از x_i را مشـــخص می‌کند. یک عدد صحیح $k>1$ در اولین قدم از الگوریتم ریزتجمیع داده شـــده است. این روش، داده‌ها را درون g گـــروه مجـــزا با حداقل k عضو تقســـیم می‌کند به طـــوری که $\sum_{i=1}^{g} n_i = n$ که n_i اندازه گروه iام اســـت. x_{ij}^{p} jامین خصوصیت از iامین رکورد از p امین گروه می‌باشد، c^p مرکز ثقل p امین گروه و میانگین برداری از همه اعضای گروه است. c^p مانند رابطه ۱ محاسبه می‌شود:

$$c_j^p = \frac{1}{n_p} \sum_{i=1}^{n_p} x_{ij}^p \qquad (۱)$$

که c_j^p jامین خصوصیت از c^p و n_p اندازه گروه اســـت. مرکز ثقل سراســـری که به وسیله c مشخص شده، مرکز ثقل همه رکوردها در T است و می‌تواند با یک شیوه مشابه محاسبه شود. در انتشار داده‌ها با حفظ حریم خصوصی یک تـــوازن بین ســـودمندی داده‌هـــا و محرمانگی وجود دارد. IL[23] یک معیار محبوب برای تعیین کیفیت ســـودمندی داده‌های بی‌نام است و مانند رابطه ۲ محاسبه میشود:

$$IL = 100 \times \frac{SSE}{SST} \qquad (۲)$$

کـــه SSE (رابطه ۳) جمع مربـــع خطا برای هر رکورد از مرکز ثقل خوشه‌اش است و SST (رابطه ٤) جمع مربع خطا برای هر رکورد از مرکز ثقل سراســـری و رابطه ٥ فاصله

18- Sum of Square Error
19- Multivariate Hansen Mukherjee algorithm
20- space-filling
21- Iterative MHM
22- Nearest Point Next MHM
23- Information Loss

بین دو رکورد است :

$$SSE = \sum_{p=1}^{g}\sum_{i=1}^{n_p} distance\left(x_i^p, c^p\right) \quad (۳)$$

$$SST = \sum_{i=1}^{n} distance\left(x_i, c\right) \quad (۴)$$

$$distance\left(x_i, x_j\right) = \sum_{l=1}^{d}\left(x_{il} - x_{jl}\right)^2 \quad (۵)$$

وقتی IL = 0 است یعنی این‌که هر رکورد با مرکز ثقل خوشه‌اش معادل است، بنابراین اطلاعاتی از دست رفته نیست و داده‌ای تغییر پیدا نکرده است.

از طرف دیگر توازن به محرمانگی اشاره دارد. خطر افشا با معیار خطر افشای اطلاعات حساس بیان می‌شود. DLD[۲۴] (رابطه ٦) یک روش محبوب برای تعیین کیفیت خطر افشا است:

$$DLD = 100 \times \frac{m}{n} \quad (۶)$$

که m تعداد کلی رکوردهای منطبق است. فرض کنید T′ نسخه بی‌نام شده از T است و رکورد $x_i \in T$ با یک رکورد بی‌نام شده $x_i' \in T'$ تعویض شده است. رکورد x_i منطبق است اگر و فقط اگر رابطه ۷ برقرار باشد:

$$x_i = \underset{x_j}{arg\,min}\left(distance\left(x_j, x_i'\right)\right), \quad 0 \le j \le n \quad (۷)$$

معمولاً افزایش اتلاف اطلاعات، کاهش خطر افشا و بالعکس را نتیجه می‌دهد. بنابراین چند معیار وجود دارد که توازن بین IL و DLD سنجیده می‌شود. SI[۲۵] یک معیار است که به وسیله رابطه ۸ محاسبه می‌شود.

$$SI = \alpha \,.\, DLD + (1 - \alpha) \,.\, IL \quad (۸)$$

که $0 \le \alpha \le 1$ اهمیت DLD در SI را کنترل می‌کند.

۴- روش‌های پیشنهادی

در الگوریتم NFPN-MHM [۳] تشکیل توالی بر اساس آخرین نقطه صورت می‌گرفت. در روش اول به رکورد آخر و ماقبل آخر و در روش دوم به پنج رکورد آخر نیز توجه شده که نتایج بهتری را به دنبال دارد. NFPN++ و Enhanced-NFPN الگوریتم MHM را روی داده‌های مرتب شده همانند NFPN-MHM استفاده می‌کنند، اما تفاوت در شیوه مرتب‌سازی داده‌ها است. این دو مرحله از این دو الگوریتم عبارتند از: ۱- مرتب‌سازی داده‌ها و ۲- خوشه‌بندی با استفاده از MHM. این دو مرحله مطابق ذیل توصیف می‌شوند: مراحل‌نهایی تجمیع و غیرنرمال‌سازی مجموعه داده‌های محافظت شده همانند کارهای قبلی در مقالات هستند و در اینجا بررسی نمی‌شود.

٤-۱- گام مرتب‌سازی داده‌ها

اولین قدم مرتب کردن داده‌ها در یک توالی از رکوردها است، به گونه‌ای که مجاورت رکوردهای نزدیک در فضای دامنه اصلی به قدر ممکن در توالی به‌دست آمده حفظ شوند. به این ترتیب داده‌های محافظت شده سودمندتر تولید می‌شود زیرا در گام بعدی، به طور پی‌درپی فقط رکوردهای مجاور با یکدیگر خوشه‌بندی می‌شوند، بنابراین خوشه‌های متراکم‌تر تولید می‌شوند. شکل ۲ رویه مرتب‌سازی الگوریتم NFPN++ و شکل ۳ الگوریتم ENFPN را نشان می‌دهد.

دورترین نقطه از مرکز ثقل سراسری سر توالی است. نقاط بعدی به انتهای توالی بر اساس رابطه ۹ اضافه می‌شوند. x_i را به‌عنوان انتهای توالی در نظر بگیرید، xj به x_i اضافه خواهد شد اگر و فقط اگر فاصلهٔ رابطه ۹ برای xj کمینه باشد:

$$D_{ij} = \begin{cases} d_{ij} & if\ (d_{ij} < 1) \\ distance\left(x_j, x_i\right) & otherwise \end{cases} \quad (۹)$$

که d_{ij} طبق رابطه ۱۰ تعریف می‌شود:

$$d_{ij} = \frac{distance\left(x_j, x_i\right)}{distance\left(x_j, c\right)} \quad (۱۰)$$

فاصلهٔ بین x_i و x_j مانند d_{ij} محاسبه می‌شود فقط وقتی که x_j نزدیک‌تر به x_i از مرکز ثقل است (خط ٦ و ۷ از شکل ٤).

در شکل ۲، r رکورد آخر و rLast رکورد ماقبل آخر می‌باشد. همچنین ۲ وزن یا اهمیت رکورد آخر و (۲-1) وزن رکورد ماقبل آخر می‌باشد. همچنین می‌توان با توجه به

24- Distance base Linkage Disclosure
25- Score Index

اهمیـــت DLD، IL و SI مقادیر ۲های گوناگون قرار داد که در بخش نتایج شرح داده خواهد شد.

در شکل ۳، خط ۸ الگوریتم، اگر اندازهٔ دنبالهٔ o بیشتر از پنج باشد میانگین پنج رکورد انتهای توالی به‌عنوان معیار تصمیم‌گیری برای پارامتر d1 در شکل ٤ می‌باشد. در غیر ایـــن صورت میانگین کل رکوردهـــای دنبالهٔ o که کمتر از ٥ می‌باشـــد در نظر گرفته می‌شود. با انجام آزمایش‌ها بر روی میانگین چند رکورد آخر (به‌عنوان مثال از میانگین ۲ رکورد تا ۱۰ رکورد آخـــر) بهترین نتیجه برای میانگین ٥ رکورد آخر به‌دســـت آمد.شکل ٥ مقایسه بین مرتب‌سازی NFPN و NFPN++ و شکل ٦ مقایسه بین مرتب‌سازی NFPN و ENFPN برای مجموعه دادهٔ نمونه یکسان را نشان می‌دهد.

٤-۲- خوشه‌بندی داده‌های مرتب شده

دومیـــن قـــدم از ENFPN و NFPN++ اعمـــال الگوریتـــم ریزتجمیع بهینه بـــرای داده‌های تک‌متغیره روی مجموعه دادهٔ T چندمتغیره با رابطه ترتیب تهیه شـــده از داده‌ها در گام قبل اســـت. الگوریتـــم MHM [۲۷] مجموعه دادهٔ T با n رکـــورد را می‌پذیرد، یک k صحیـــح و ترتیبی از رکوردها به‌عنـــوان ورودی و بخش‌بندی بهینه با نســـبت به صفر، رضایت‌مندی k–بی‌نامی با حداقل مقدار اتلاف اطلاعات را بر می‌گرداند. MHM رکوردهای مرتب شـــده در یک گراف جهت‌دار بدون دایـــره G را در نظر می‌گیرد که هر رکورد به‌عنوان یک گره در نظر گرفته می‌شود.

همان‌طور که در شـــکل ۷ نشان داده شده، MHM ابتدا رکـــورد موقت x_0 را قبل از گره x_1 اضافه می‌کند و ســـپس لبه‌های شـــروع k از هـــر x_i، $0\le i\le n-k$، بـــه x_{i+j}، $k\le j\le 2k-1$ و $i+j\le n$ را اضافه می‌کند. هر لبه e_{ij} به‌عنوان یک خوشه شامل رکوردهـــا $\{x_{0[i]}, \dots, x_{0[j-1]}\}$ فـــرض می‌شـــود، و وزن لبه با مجموع مربعات خطا از این خوشـــه معادل اســـت. بعد از اضافه شـــدن لبه‌ها به G، مسیرهای چندگانه از x_0 به x_n در G وجود دارند. هر مسیر یک k–بخش‌بندی معتبر از داده‌ها را نشـــان می‌دهد و طول آن SSE نهایی از این خوشه‌بندی را تعریف می‌کند. کوتاه‌ترین مسیر از x_0 به x_n خوشه‌بندی

```
Input: dataset T
Output: ordered sequence o on records of T
1  l ← T
2  o ← [ ]
3  c ← global_centroid
4  rNew ← furthst record from c
5  add rNew to the end of o
6  remove rNew from l
7  while l is not empty
8      n ← NearestFarSearch ( rNew , c , l )
9      add n to the end of o
10     rLast ← rNew
11     r ← n
12     γ ← weight
13     rNew ←Mean (r * γ  ,  rLast * (1-γ) )
14 return o
```

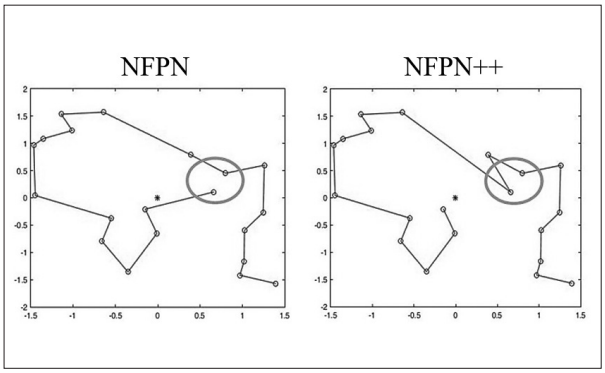
شکل ۲: الگوریتم مرتب‌سازی NFPN++

```
Input: dataset T
Output: ordered sequence o on records of T
1  l ← T
2  o ← [ ]
3  c ← global_centroid
4  rNew ← furthest record from c
5  add rNew to the end of o
6  remove rNew from l
7  while l is not empty
8      if size(o)>=5
9          rNew ←Mean (last five records in o)
10         n ← NearestFarSearch ( rNew , c , l )
11         add n to the end of o
12     else
13         rNew ←Mean (all record in o)
14         n ← NearestFarSearch ( rNew , c , l )
15         add n to the end of o
16 return o
```

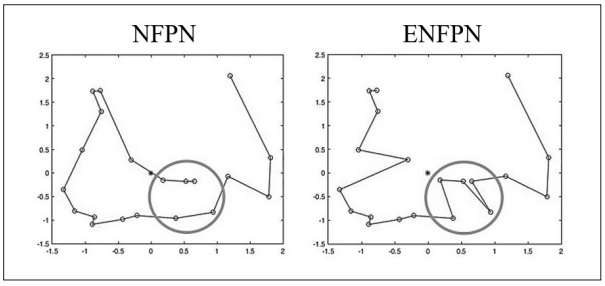
شکل ۳: الگوریتم مرتب‌سازیENFPN

```
Input: record r, record c, list l
Output: record n in l, furthest record from c between nears to r
1  mind ← +∞
2  n ← NULL
3  For each record x in l
4      d1 ← EuclideanDistance ( x , r )
5      d2 ← EuclideanDistance ( x , c )
6      if ( d2 > d1 )
7          d ← d1 / d2
8      else
9          d ← d1
10     if ( d < mind )
11         mind ← d
12         n ← x
13 return n
```

شکل ۴: الگوریتم جستجوی نزدیک‌ترین دور



شکل ۵: مقایسه بین مرتب‌سازی‌های NFPN و NFPN++ روی یک نمونه داده. هر محور یک صفت را نمایش می‌دهد.



شکل ۶: مقایسه بین مرتب‌سازی‌های NFPN و ENFPN روی یک نمونه داده. هر محور یک صفت را نمایش می‌دهد.

بهتر با رابطه ترتیب داده شده از داده‌ها را مشخص می‌کند که حداقل اتلاف اطلاعات را به وسیله تقسیم SSE بر SST را متحمل می‌شود.

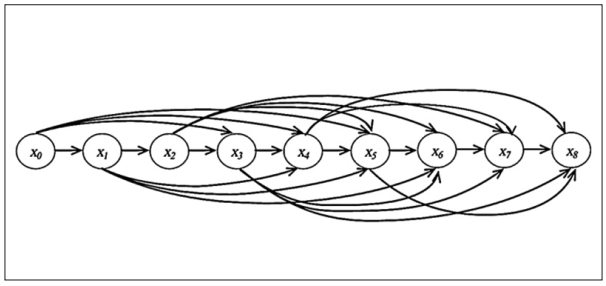
۵- نتایج تجربی

نتایـــج NFPN++ با توجه به اهمیت ۳ معیار DLD، IL و SI با مقادیر مختلف ۲ ارائه شده است. تعیین مقدار دقیق ۲ با توجه به جداول ۱ و ۲ و ۳ می‌باشد. این مقادیر با آزمایش ۲های مختلف از ۰,۱ تا ۰,۹ به‌دســت آمده و بهترین آن‌ها طبق این جدول‌ها می‌باشــد. جداول ٤ و ٥ و شــکل‌های ۸ تا ۱۰ نتایج آزمایش‌ها را در مقایســه با ســه روش، NPN NFPN و MDAV نشان می‌دهند. روش پیشنهاد شده مقادیر IL سراســری کمینه را در ۱۸ از ۲٤ آزمایش نشان می‌دهد. همچنین مقادیر DLD کمینه را برای تمام آزمایش‌ها نشــان می‌دهد.

برای مقایسه و اثربخشی روش‌ها، استفاده از روش آنالیز واریانس توصیه شده است. بدین منظور با ورود داده‌ها در نرم‌افزار SPSS و با اســتفاده از جــداول آنالیز واریانس به بررســی و مقایســه روش‌ها با یکدیگر و همچنین مقایسهٔ مؤلفه‌های دیگر از قبیل مجموعه داده‌ها و پارامتر k پرداخته شد.

شکل‌های ۸ و ۹ به ترتیب نمودارهای حاصل از جداول ٤ و ٥ را نشان می‌دهد.

نتایج به‌دست آمده از آنالیز واریانس، اختلاف معنی‌دار آماری در استفاده از روش‌ها را نشان داد ($P<0.05$ ، مقدار احتمال باید کمتر از ۰,۰۵ باشد تا اختلاف معنی‌دار باشد).



شکل ۷: گراف ایجاد شده به وسیله MHM برای یک مجموعه دادهٔ تک‌متغیره با ۸ رکورد، k=3

جدول ۱: تعیین مقدار ۷ برای IL کمینه

	EIA	Census	Tarra gona
k = 3	0.1	0.2	0.3
k = 4	0.3	0.5	0.6
k = 5	0.5	0.5	0.6
k = 6	0.5	0.5	0.6
k = 7	0.5	0.5	0.6
k = 8	0.5	0.5	0.6
k = 9	0.5	0.5	0.6
k = 10	0.4	0.5	0.6

ســپس چون آزمون آماری معنی‌دار اســت با استفاده از آزمون‌های تعقیبی آماری (آزمون LSD) به مقایســه روش پیشنهادی با دیگر روش‌ها پرداخته شد. نتایج ضمن تایید برتری روش پیشــنهادی نســبت به روش‌های پیشین، به برتری نسبی روش NFPN++ اشاره دارد. همچنین اختلاف به شدت معنی‌دار روش پیشنهادی با روش NPN مشاهده شد.

مقادیر خاکستری در جداول ٤ و ٥ عدم موفقیت روش پیشنهادی را نشان می‌دهد. اما نسبت به دو روش دیگر که موفق نبوده‌اند، موفق بوده است.

نتایــج تایید می‌کند کــه هیچ یک از IL و یــا DLD برای مقایســه به صورت منحصر به فرد کافی نیســتند، به این خاطر در کاربردهای متفــاوت، اهمیت متفاوتی به هر یک از آن‌ها می‌دهند. SI برای در نظر گرفتن هر دوی IL و DLD با وزن‌های مختلف استفاده می‌شود. α سطح اهمیت خطر افشا است، بنابراین اهمیت اتلاف اطلاعات به وسیله (1-α)

جدول ۲: تعیین مقدار γ برای DLD کمینه

	EIA	Census	Tarra gona
k = 3	0.9	0.9	0.9
k = 4	0.9	0.9	0.9
k = 5	0.9	0.9	0.9
k = 6	0.9	0.9	0.9
k = 7	0.9	0.9	0.9
k = 8	0.9	0.9	0.9
k = 9	0.9	0.9	0.9
k = 10	0.9	0.9	0.9

جدول ۳: تعیین مقدار γ برای SI کمینه

EIA	α = 0.3	γ = 0.7
	α = 0.5	γ = 0.7
	α = 0.7	γ = 0.7 , 0.9
Census	α = 0.3	γ = 0.3 , 0.5
	α = 0.5	γ = 0.7
	α = 0.7	γ = 0.8 , 0.9
Tarra gona	α = 0.3	γ = 0.6 , 0.9
	α = 0.5	γ = 0.9
	α = 0.7	γ = 0.9

جدول ۴: نتایج مقایسه‌های IL برای الگوریتم NFPN++

	Method	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	k = 10
EIA	MDAV	0.48	0.67	1.67	1.31	2.17	2.87	3.18	3.83
	NPN	0.5	0.69	1	1.31	1.92	2.14	2.36	2.49
	NFPN	**0.41**	0.61	**0.86**	**1.1**	1.83	2.08	2.29	2.45
	NFPN++	**0.43**	**0.6**	**0.86**	**1.11**	**1.74**	**1.92**	**2.11**	**2.18**
Census	MDAV	5.69	7.49	9.09	10.38	**11.58**	**12.39**	**13.29**	**14.16**
	NPN	6.21	8.96	11.05	13.28	15.31	17.36	18.65	20.23
	NFPN	5.63	7.6	9.3	11.03	12.74	14.35	15.43	16.68
	NFPN++	**5.47**	**7.35**	**8.92**	**10.28**	**11.76**	**12.95**	**14.13**	**15.42**
Tarragona	MDAV	16.93	19.55	22.46	26.33	27.52	29.69	31.21	33.19
	NPN	17.53	21.69	28.17	30.63	32.21	34.01	35.96	38.72
	NFPN	15.47	18.79	21.81	25.87	28.81	30.28	31.07	33.23
	NFPN++	**15.23**	**18.21**	**21.55**	**24.8**	**27.32**	**28.65**	**30.37**	**32.4**

مشخص می‌شـــود. مقادیر پایین‌تر اتلاف اطلاعات صحت بیشتری را نشان می‌دهد که معمولاً هزینهٔ بالاتر خطر افشا را به دنبال دارد. اگرچه یک کاهش در SI بهبود سراســـری در هر دوی IL و DLD را نشـــان می‌دهد و یک روش ساده برای مقایســـه کردن روش‌های بی‌نام‌ســـازی متفاوت را میســـر می‌سازد. شکل ۱۰ نتایج مقایســـه‌های SI با مقادیر مختلف α با توجه به γهای جدول ۳ برای الگوریتم NFPN++ را نشان می دهد.

جداول ٦ و ۷ و شـــکل‌های ۱۱ تا ۱۳ نتایج آزمایش‌های ENFPN را در مقایســـه بـــا دو روش NPN و NFPN نشـــان می‌دهند. روش پیشنهاد شده مقادیر IL سراسری کمینه را در ۱۸ از ۲۷ آزمایش نشـــان می‌دهد. همچنین مقادیر DLD کمینه را در ۱۹ از ۳۳ آزمایش نشان می‌دهد.

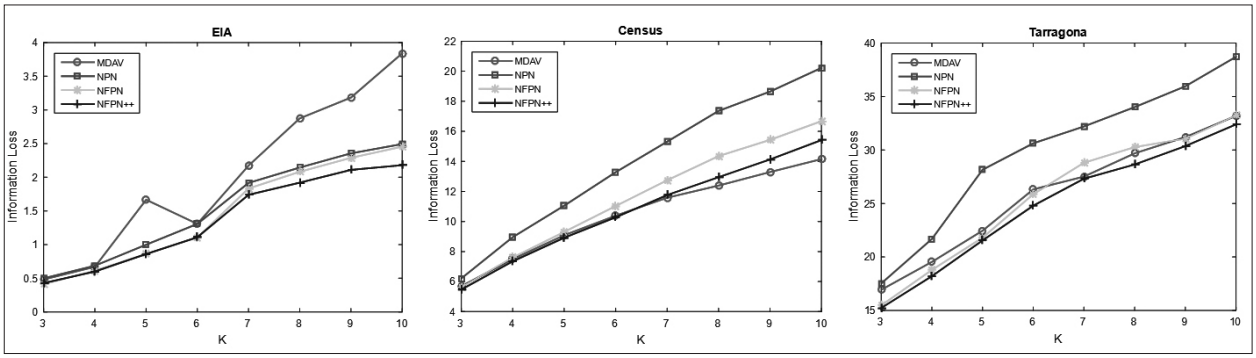
شـــکل‌های ۱۱ و ۱۲ به ترتیـــب نمودارهای حاصل از جداول ٦ و ۷ را نشان می‌دهد.

شکل ۱۳ نتایج مقایسه‌های SI با مقادیر مختلف α برای الگوریتم ENFPN را نشان می دهد.

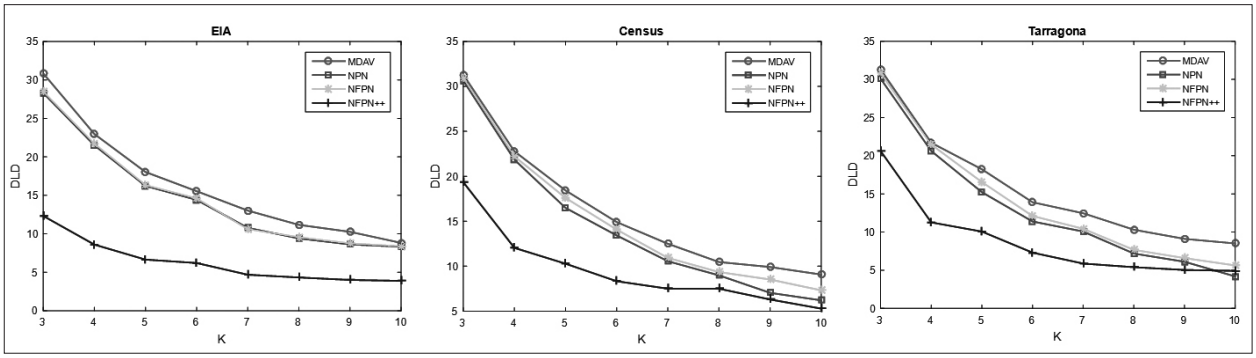
بـــا توجه بـــه جـــدول ٦ نتایج نشـــان می‌دهـــد روش پیشـــنهادی برای مجموعه دادهٔ EIA موفق نبوده اســـت اما

جدول ۵: نتایج مقایسه‌های DLD برای الگوریتم NFPN++

	Method	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	k = 10
EIA	MDAV	31.23	23.39	18.35	15.88	13.39	11.39	10.51	9.07
	NPN	28.67	21.85	16.54	14.71	11.05	9.68	8.9	8.6
	NFPN	28.86	22.07	16.64	14.88	10.90	9.82	9.02	8.65
	NFPN++	**12.32**	**8.60**	**6.65**	**6.21**	**4.69**	**4.33**	**4.01**	**3.89**
Census	MDAV	31.3	22.78	18.43	14.91	12.5	10.46	9.91	9.07
	NPN	30.65	21.85	16.48	13.43	10.56	8.98	7.04	6.2
	NFPN	30.93	22.22	17.59	14.07	10.93	9.35	8.52	7.31
	NFPN++	**19.35**	**12.04**	**10.28**	**8.33**	**7.50**	**7.50**	**6.30**	**5.28**
Tarragona	MDAV	31.41	21.82	18.47	14.03	12.47	10.31	9.11	8.51
	NPN	30.34	20.86	15.23	11.39	10.19	7.19	6.12	4.2
	NFPN	31.06	21.58	16.67	12.23	10.43	7.79	6.71	5.64
	NFPN++	**20.62**	**11.27**	**10.07**	**7.31**	**5.88**	**5.40**	**5.04**	**4.92**



شکل ۸ : نمودارهای حاصل از جدول ۴

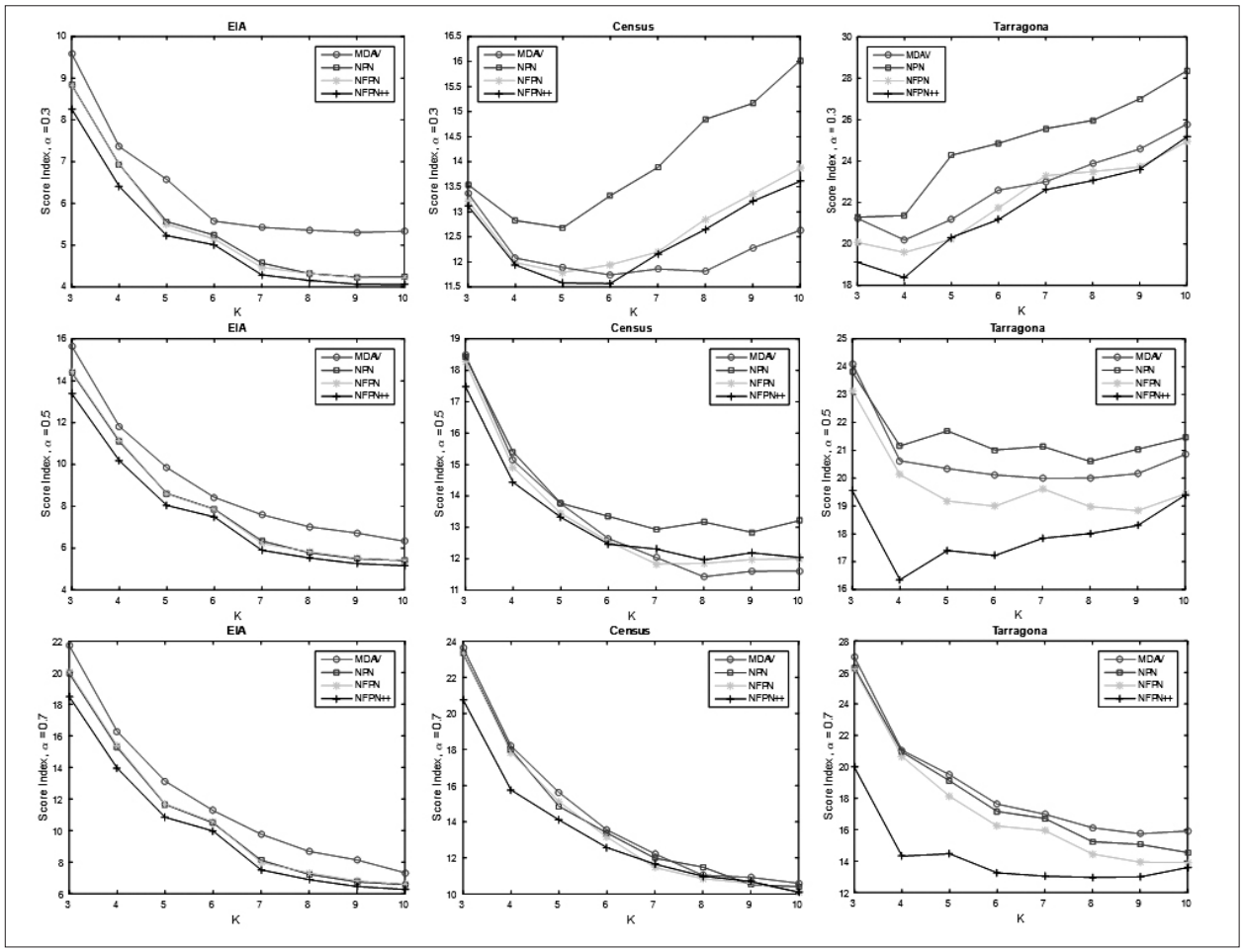


شکل ۹ : نمودارهای حاصل از جدول ۵

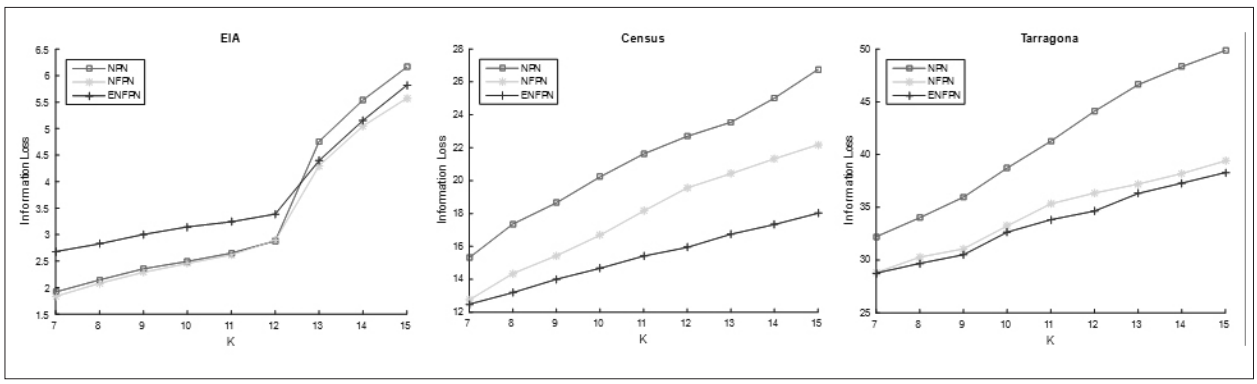
از طرفی طبق جدول ۷ برای تمام آزمایش‌ها موفق عمل کرده است. از طرفی دیگر با آزمایش‌های انجام گرفته مشاهده گردید که معیار IL برای k>6 نتایج بهتری را به دنبال دارد.

نتیجه

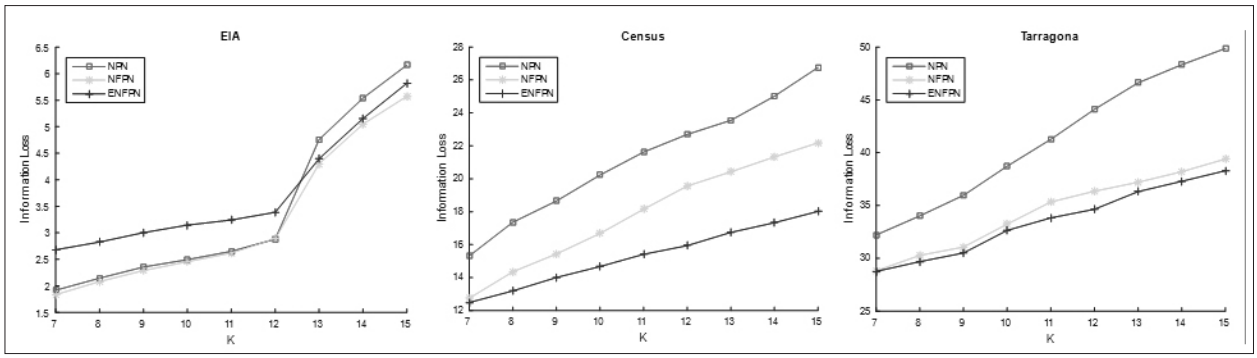
یک توازن بین حریم خصوصی و سودمندی داده‌ها در انتشار ریزداده‌ها وجود دارد. ریزتجمیع یکی از رویکردهای مورد استفاده در بسیاری از موارد برای حل



شکل ۱۰ : نتایج مقایسه‌های SI با مقادیر مختلف α با توجه به γ های جدول ۳ برای الگوریتم NFPN++



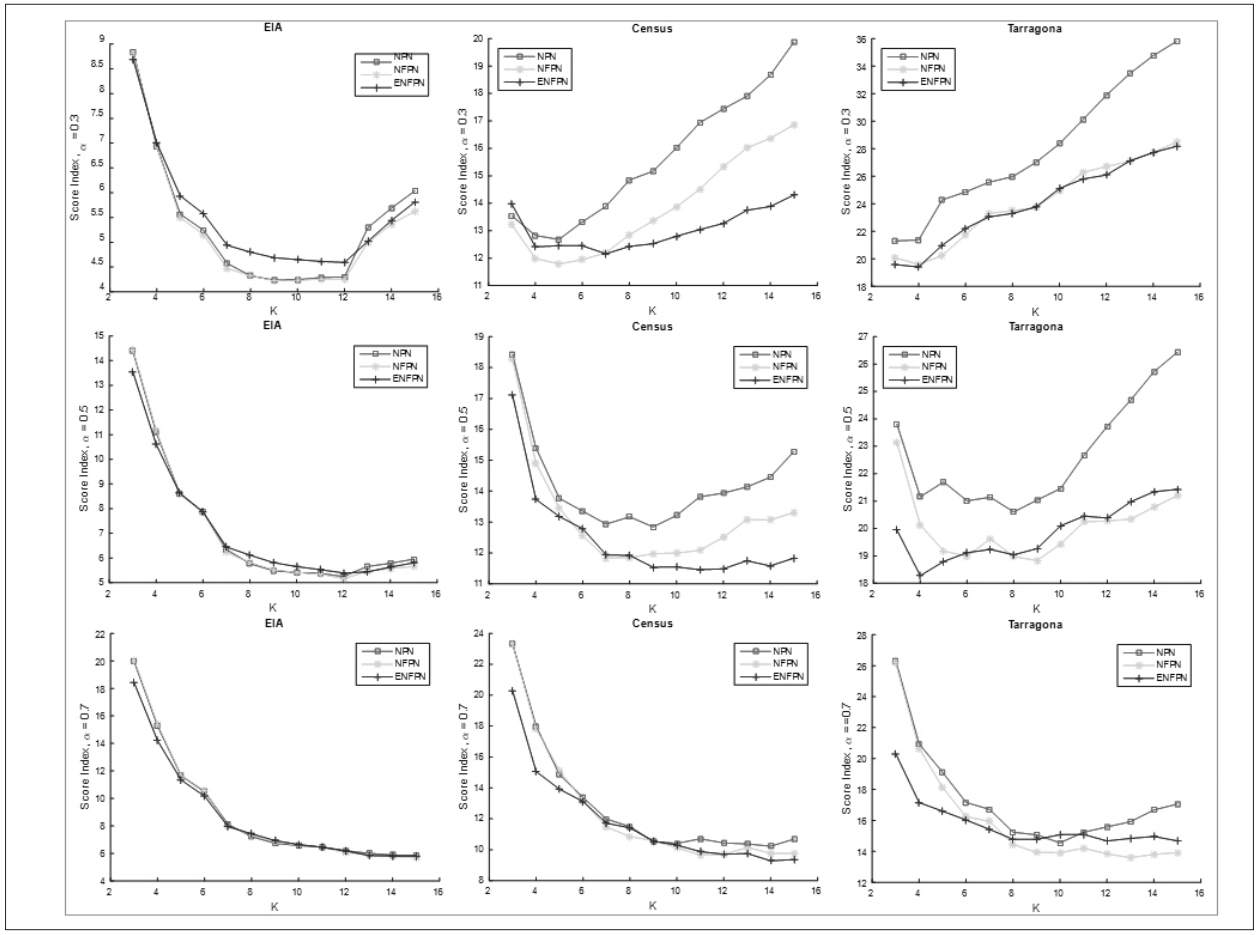
شکل ۱۱ : نمودارهای حاصل از جدول ۶



شکل ۱۲ : نمودارهای حاصل از جدول ۷

جدول ۶: نتایج مقایسه‌های IL برای الگوریتم ENFPN

	Method	k = 7	k = 8	k = 9	k = 10	k = 11	k = 12	k = 13	k = 14	k = 15
EIA	NPN	1.92	2.14	2.36	2.49	2.65	**2.88**	4.76	5.54	6.16
	NFPN	**1.83**	**2.08**	**2.29**	**2.45**	**2.62**	2.89	**4.30**	**5.05**	**5.57**
	ENFPN	2.68	2.83	3.00	3.14	3.24	3.39	4.40	5.16	5.82
Census	NPN	15.31	17.36	18.65	20.23	21.62	22.69	23.55	25.01	26.76
	NFPN	12.74	14.35	15.43	16.68	18.15	19.56	20.42	21.32	22.18
	ENFPN	**12.49**	**13.19**	**14.01**	**14.67**	**15.42**	**15.94**	**16.74**	**17.33**	**18.02**
Tarragona	NPN	32.21	34.01	35.96	38.72	41.27	44.11	46.62	48.33	49.87
	NFPN	28.81	30.28	31.07	33.23	35.33	36.35	37.21	38.20	39.40
	ENFPN	**28.76**	**29.69**	**30.51**	**32.64**	**33.83**	**34.65**	**36.32**	**37.29**	**38.30**



شکل ۱۳ : نتایج مقایسه‌های SI با مقادیر مختلف α برای الگوریتم ENFPN

جدول ۷: نتایج مقایسه‌های DLD برای الگوریتم ENFPN

	Method	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$	$k=10$	$k=11$	$k=12$	$k=13$
EIA	NPN	28.67	21.85	16.54	14.71	11.05	9.68	8.9	8.6	8.11	7.62	6.55
	NFPN	28.86	22.07	16.64	14.88	10.90	9.82	9.02	8.65	8.09	7.43	6.62
	ENFPN	**25.73**	**19.60**	**15.40**	**13.61**	**10.22**	**9.41**	**8.63**	**8.16**	**7.82**	**7.40**	**6.48**
Census	NPN	30.65	21.85	16.48	**13.43**	**10.56**	**8.98**	**7.04**	**6.2**	**6.02**	**5.19**	**4.72**
	NFPN	30.93	22.22	17.59	14.07	10.93	9.35	8.52	7.31	6.02	5.46	5.74
	ENFPN	**25.00**	**17.04**	**15.00**	**13.61**	11.39	10.65	9.07	8.43	7.50	7.04	6.76
Tarragona	NPN	30.34	20.86	15.23	11.39	10.19	**7.19**	**6.12**	**4.2**	**4.08**	**3.36**	**2.76**
	NFPN	31.06	21.58	16.67	12.23	10.43	7.79	6.71	5.64	5.16	4.20	3.48
	ENFPN	**20.86**	**15.47**	**13.31**	**11.39**	**9.71**	8.39	8.03	7.55	7.07	6.12	5.64

این مسئله است. روش‌های ریزتجمیع گوناگون در مقالات پیشـــنهاد شده اســـت. این مقاله روش ریزتجمیع NFPN++ و ENFPN، را به‌عنـــوان توســـعه روش پیش‌تر ارائه شـــدهٔ NFPN-MHM با یک صحت بهبود یافته پیشـــنهاد کرد. این روش‌هـــا معمولاً توازن بهتر بین اتـــلاف اطلاعات و خطر افشا نســـبت به روش‌های دیگر در سایر مقالات را نشان می‌دهند. همچنین انعطاف‌پذیری لازم را برای اهمیت دادن به هر یک از اتلاف اطلاعات، کنترل افشا و شاخص امتیاز را دارند.

سپاس‌گزاری

در اینجا لازم هســـت از آقای دکتر مصطفی ســـالاری بـــرای به اشـــتراک‌گذاری کـــد NFPN و کمک‌های ایشـــان سپاس‌گزاری نماییم.

مراجع

[1] Mortazavi, R., Jalili, S., & Gohargazi, H., Multivariate microaggregation by iterative optimization. Applied intelligence, 2013. 39(3): p. 529-544.

[2] Schoeman, F.D., Philosophical dimensions of privacy: An anthology. 1984: Cambridge University Press.

[3] Salari, M., Jalili, S., & Mortazavi, R., A utility preserving data-oriented anonymization method based on data ordering. in Telecommunications (IST), 2014 7th International Symposium on. 2014. IEEE.

[4] Torra, V., Navarro-Arribas, G., & Stokes, K., An Overview of the Use of Clustering for Data Privacy, in Unsupervised Learning Algorithms. 2016, Springer. p. 237-251.

[5] Hong, T.-P., Lin, C.-W., Yang, K.-T., & Wang, S.-L., Using TF-IDF to hide sensitive itemsets. Applied Intelligence, 2013. 38(4): p. 502-510.

[6] Yin, Y., Kaku, I., Tang, J., & Zhu, J., Privacy-preserving data mining. Data mining. Springer, Berlin, 2011.

[7] Domingo-Ferrer, J., & Torra, V., Disclosure control methods and information loss for microdata. Confidentiality, disclosure, and data access: theory and practical applications for statistical agencies, 2001: p. 91-110.

[8] Oganian, A., & Karr, A. F., Masking methods that preserve positivity constraints in microdata. Journal of Statistical Planning and Inference, 2011. 141(1): p. 31-41.

[9] Domingo -Ferrer ,J .,Solanas ,A .,& Martinez -Balleste, A., Privacy in Statistical Databases: k-Anonymity Through Microaggregation. in Granular Computing (GrC). 2006.

[10] Xu, C., Wang, Y., Gu, Y., Lin, S., & Yu, G., Efficient fuzzy ranking queries in uncertain databases. Applied Intelligence, 2012. 37(1): p. 47-59.

[11] Kiran, P., & Kavya, N., A Survey on Methods, Attacks and Metric for Privacy Preserving Data Publishing. International Journal of Computer Applications, 2012. 53(18): p. 20-28

[12] De Capitani Di Vimercati, S., Foresti, S., Livraga, G., & Samarati, P., Data privacy: definitions and techniques. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2012. 20(06): p. 793-817.

[13] Domingo-Ferrer, J., & Sramka, M., Disclosure Control by Computer Scientists: An Overview and an Application of Microaggregation to Mobility Data Anonymization. 2011: p. 1082-1086.

[14] Rebollo-Monedero, D., Forné, J., & Soriano, M., An algorithm for k-anonymous microaggregation and clustering inspired by the design of distortion-optimized quantizers. Data & Knowledge Engineering, 2011. 70(10): p. 892-921.

[15] Solanas, A., Martinez-Balleste, A., & Domingo-Ferrer, J., V-MDAV: a multivariate microaggregation with variable group size. in 17th COMPSTAT Symposium of the IASC, Rome. 2006.

[16] Wong, R. C.-W., Li, J., Fu, A. W.-C., & Wang, K., (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data

Engineering, IEEE Transactions on, 2005. 17(7): p. 902-911.
[23] Solé, M., Muntés-Mulero, V., & Nin, J., Efficient microaggregation techniques for large numerical data volumes. International Journal of Information Security, 2012. 11(4): p. 253-267.
[24] Domingo-Ferrer, J., & Torra, V., Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Mining and Knowledge Discovery, 2005. 11(2): p. 195-212.
[25] Panagiotakis, C., & Tziritas, G., Successive group selection for microaggregation. Knowledge and Data Engineering, IEEE Transactions on, 2013. 25(5): p. 1191-1195
[26] Hansen, S. L., & Mukherjee, S., A polynomial algorithm for optimal univariate microaggregation. IEEE Transactions on Knowledge & Data Engineering, 2003(4): p. 1043-1044.
[27] Ghinita, G., Karras, P., Kalnis, P., & Mamoulis, N., Fast data anonymization with low information loss. in Proceedings of the 33rd international conference on Very large data bases. 2007. VLDB Endowment.
[28] Domingo-Ferrer, J., Martínez-Ballesté, A., Mateo-Sanz, J. M., & Sebé, F., Efficient multivariate data-oriented microaggregation. The VLDB Journal, 2006. 15(4): p. 355-369.
mining. 2006. ACM.
[17] Sun, X., Li, M., & Wang, H., A family of enhanced-diversity models for privacy preserving data publishing. Future Gener ComputSyst, 2011. 27(3): p.348–356.
[18] Sun, X., Sun, L., & Wang, H., Extended k-anonymity models against sensitive attribute disclosure. Computer Communications, 2011. 34(4): p. 526-535.
[19] Oganian, A., & Domingo-Ferrer, J., On the complexity of optimal microaggregation for statistical disclosure control. Statistical Journal of the United Nations Economic Commission for Europe, 2001. 18(4): p. 345-353.
[20] Navarro-Arribas, G., & Torra, V., Information fusion in data privacy: A survey. Information Fusion, 2012. 13(4): p. 235-244.
Fung, B., Wang, K., Chen, R., & Yu, P. S., Privacy-preserving data publishing: A survey of recent developments. ACM Computing Surveys (CSUR), 2010. 42(4): p. 14.
[21] Domingo-Ferrer, J., & Mateo-Sanz, J. M., Practical data-oriented microaggregation for statistical disclosure control. Knowledge and Data Engineering, IEEE Transactions on, 2002. 14(1): p. 189-201.
[22] Laszlo, M., & Mukherjee, S., Minimum spanning tree partitioning algorithm for microaggregation. Knowledge and Data