سیستم‌های فازی و کاربردها
سال پنجم، شماره دوم (پائیز و زمستان ۱۴۰۱)
صص. ۷۳ تا ۹۲

شاپای چاپی: ۲۷۱۷-۴۴۰۹
شاپای الکترونیکی: ۲۷۱۷-۳۹۸۴
DOI:10.22034/JFSA.2022.354599.1141

روشی جدید برای حفظ حریم خصوصی داده ها بر اساس تجزیه نامنفی ماتریس

لعیا علی احمدی پور و عزت ولی پور

بخش علوم کامپیوتر، دانشکده ریاضی و کامپیوتر، دانشگاه شهید باهنر کرمان، کرمان، ایران
بخش ریاضی کاربردی، دانشکده ریاضی و کامپیوتر، دانشگاه شهید باهنر کرمان، کرمان، ایران

تاریخ دریافت: ۱۴۰۱/۵/۱۲ تاریخ پذیرش: ۱۴۰۱/۷/۷

نوع مقاله: علمی-پژوهشی

چکیده

نگراني شرکت‌ها از افشا و نقض حریم خصوصی کاربران در حال افزایش است و این امر منجر به تمرکز بسیاری از محققان بر روی توسعه روش‌های حفظ حریم خصوصی داده‌ها شده است. این روش‌ها روی داده‌های اصلی تاثیر می‌گذارند و با حفظ ویژگی‌ها و ارتباط بین آن‌ها، داده را بصورت دیگری منتشر مي‌کنند. در این مقاله با استفاده از روش تجزیه نامنفي ماتریس‌ها، الگوریتمي براي تولید داده‌هایی با سطوح امنیتي متفاوت، جهت انتشار براي يك مجموعه داده اصلي پیشنهاد مي‌شود. پیاده‌سازی‌ها با دو رویکرد متفاوت روی مجموعه داده‌های متنوع استاندارد نشان می‌دهند که الگوریتم پیشنهادي علاوه بر تولید داده‌هایي با سطوح امنیتي مختلف، ساختار محدودیت‌هاي داده‌ي اصلي را نیز حفظ مي‌کند. روش مطرح شده در این مقاله روی مجموعه داده‌های با ابعاد بالا پیاده‌سازی شده، در حالیکه برخی روش‌های ریزتجمیع‌کننده مبتنی بر فازی قابلیت پیاده‌سازی روی آن‌ها را ندارند. همچنین، نتایج تجربي با استفاده از روش c-میانگین فازی نشان مي‌دهد که میزان اطلاعات از دست رفته نیز بسیار ناچیز است. لذا، الگوریتم پیشنهادي مي‌تواند داده‌هایي را منتشر کند که در عین حفظ حریم خصوصی، براي پردازشگران قانوني قابل استناد باشند.

عبارات و کلمات کلیدی: حفظ حریم خصوصي داده، روش خوشه‌بندي تجزیه نامنفي ماتریس، داده ماسك شده، داده منتشر شده، اطلاعات از دست رفته.

Email(s): valipour@uk.ac.ir.

Mathematics Subject Classification:

۱ مقدمه

در سال‌های اخیر رویکردها و روش‌های متنوعی جهت حفظ حریم خصوصی[1] گسترش یافته‌اند. به طور کلي، مکانیزم‌های حافظ حریم خصوصی داده‌ها را می‌توان در دو رده‌ي روش‌های داده-محور (اهداف کلی)[2] و روش‌های محاسبات-محور (اهداف خاص)[3]قرار داد [۱۳]. در واقع روش‌هايي که براي انتشار داده مستقل از کاربرد گذشته آن استفاده مي‌شوند، را داده-محور گويند [۱۲]. از طرفي روش‌هاي گسترش يافته براي مجموعه داده‌هاي خاص را محاسبات-محور مي‌نامند [۱۵، ۱۶]. یکی از مهمترین زیرردهای روش‌هاي داده محور، روش‌های آشفته‌ساز [4] [۹] مانند ریزتجمیع‌کننده‌ها[5] [۶، ۳، ۴] و روش‌های تعویض رتبه[6] [۱۱]هستند که بر پایه آشفته کردن داده‌ها عمل مي‌کنند. روش‌های ریزتجمیع‌کننده از يك الگوريتم خوشه‌بندي[7] جهت افراز داده‌های اصلی به مجموعه‌هایی با حداقل k و حداکثر $2k$ عضو استفاده می کنند. سپس نماینده هر خوشه را تعیین و نهایتا سایر اعضا خوشه را با نماینده جانشین مي‌کنند [۱۴]. این جانشاني جزییات اطلاعات اعضا را پوشانده و حریم خصوصي را حفظ مي‌کند. علاوه بر این، یک فایل ریزتجمیع شده، به نوعي در خاصیت k-ناشناسي[8] صدق مي‌کند زیرا هر کدام از حداقل k عضو خوشه از سایرین قابل تمایز نیست. لازم به ذکر است که به مجموعه داده‌اي که پس از انجام عملیات محافظتي و حفظ حریم خصوصی به عنوان خروجي در اختیار سایرین قرار مي‌گیرد، به اصطلاح داده منتشر شده[9] مي‌گویند. مسئله ریز تجمیع کننده روی داده‌های تک متغیره در زمان چند جمله‌ای[10] قابل پیاده سازی است [۸] در حالیکه ثابت شده است این مسئله برای داد های چند متغیره NP- سخت[11] است [۱۰]. تورا [۱۴] يك روش ریزتجمیع‌کننده بر پایه خوشه‌بندي C-میانگین فازي ارایه مي‌دهد که

[1] Privacy preserving
[2] Data-driven (general purpose)
[3] Computation-driven (specific purpose)
[4] Pertubative methods
[5] Microaggreagation
[6] Swap ranking
[7] Clustering algorithm
[8] k-anonymity
[9] Published data
[10] Polynomial time
[11] NP- Hard

این روش توانایی بررسی برقراري یا عدم برقراری برخي از محدودیت‌های[۱۲] اصلي روي مجموعه داده ماسك شده را نیز دارد. هر چند روش مذکور روي مجموعه داده‌هاي با ابعاد و اندازه‌ي بزرگ قابل پیاده‌سازی نیست.

در این مقاله روش جدیدی برای حفظ حریم خصوصی داده‌ها معرفی می‌شود که از تجزیه نامنفي ماتریس‌ها به عنوان روش خوشه‌بندی بهره می‌برد. ایده استفاده از خوشه‌بندی در الگوریتم پیشنهادی مشابه با روش‌های ریزتجمیع‌کننده است. هرچند خوشه‌بندی مورد استفاده در اینجا بر عکس ریزتجمیع‌کننده‌ها نامقید است. علاوه بر این هدف ما تولید چندین مجموعه داده ماسک شده[۱۳] با سطوح امنیتی متفاوت است که همگی بتوانند کاندید داده منتشر شده باشند. در این صورت تصمیم گیرنده از آزادی بیشتری برای انتخاب داده منتشر شده برخوردار است. همچنین امروزه، بسیاري از تصمیمات کاربران، مبتني بر پردازش داده‌هاي منتشرشده هستند. از طرفی مکانیزم‌هاي حفظ حریم خصوصي از کیفیت داده‌ها می‌کاهند، لذا معیارهایي براي سنجش کیفیت داده‌های منتشر شده پیشنهاد شده‌اند. یکي از مهمترین آن‌ها، معیار اطلاعات از دست رفته (IL)[۱۴] است [۱۲]. بنابراین در پیاده‌سازی الگوریتم پیشنهادی روی مجموعه داده‌های واقعی علاوه بر تولید کاندیداهای متفاوت برای انتشار، به بررسی برقراری ساختار محدودیت‌های[۱۵] اصلی روی داده‌های ماسک شده نیز پرداخته شده است. همچنین با استفاده از خوشه‌بندی c-میانگین فازی مقدار IL را نیز برای تمام داده‌های تولید شده محاسبه شده است. نهایتا، این امکان را برای تصمیم گیرنده فراهم شده که با در نظر گرفتن تعامل بین حفظ حریم خصوصی و اطلاعات از دست رفته، یکی از داده‌های ماسک شده را برای انتشار انتخاب کند.

ادامه مقاله به صورت زیر ارایه شده است. در بخش ۲ مفاهیم و پیش‌نیازهای اولیه‌ی مورد استفاده در مقاله مطرح شده‌اند. بخش ۳ به جزییات روش پیشنهادی و الگوریتم آن اختصاص یافته است. در بخش ۴ نتایج تجربی پیاده‌سازی الگوریتم و در پایان نتیجه‌گیری و پیشنهادهایی برای تحقیقات آتی مطرح شده‌اند.

[12] Constraints
[13] Masked Data
[14] Information loss
[15] Constraints

۲ مفاهیم اولیه و پیش‌نیازها

در این بخش مفاهیم خوشه‌بندی تجزیه نامنفی ماتریس و اطلاعات از دست رفته به اختصار معرفی می‌شوند.

۱.۲ روش خوشه‌بندی تجزیه نامنفی ماتریس

در این مقاله مجموعه‌ي همه ماتريس‌هاي نامنفي $m \times n$ را با نماد $\mathbb{R}_{\geq 0}^{m \times n}$ نمایش داده شده‌اند.

تعریف ۱.۲. ([۱۷]) فرض کنید $D \in \mathbb{R}_{\geq 0}^{m \times n}$ و $k \leqslant rank(D)$ داده شده باشند. تجزیه D به ماتریس‌های پایه[۱۶]

$W \in \mathbb{R}_{\geq 0}^{m \times k}$ و ضرایب[۱۷] $H \in \mathbb{R}_{\geq 0}^{k \times n}$ بطوریکه $D \approx WH$ را NMF گویند.

يك ديدگاه متداول براي يافتن ماتريس‌هاي W و H حل مساله بهينه سازي زير است [۷]:

$$\min_{W \geq 0, H \geq 0} \|D - WH\|_F, \qquad (۱)$$

که در آن $\| \ \|_F$ محاسبه نرم فروبینوس[۱۸] هست. بریا و همکاران [۲] برخی از الگوريتم‌هاي حل مساله (۱) را مرور کرده‌اند و به تعدادي از کاربردهاي آن پرداخته‌اند.

NMF بطور گسترده‌اي در حوزه‌های مختلف علوم داده استفاده مي‌شود. به عنوان مثال در خوشه‌بندي، يک مجموعه از بردارهای داده $d_1, \ldots, d_n \in \mathbb{R}^m$ وجود دارند که آن‌ها را به عنوان ستون‌هاي ماتريس D قرار مي‌دهيم:

$$D = [d_1, \ldots, d_n].$$

[16] Basis matrix

[17] Coefficient matrix

[18] Frobenius norm

منظور از یک تجزیه $k-$رتبه از ماتریس D یافتن تجزیه‌ی NMF -ی است که در آن ماتریس‌های W و H دارای رتبه k هستند که مقدار k از قبل داده شده است. از منظر علوم داده این تجزیه معادل خوشه‌بندي مشاهدات d_i در k خوشه است. در این کاربرد از NMF، $W_{m\times k}$ به عنوان ماتریس مراکز خوشه تعبیر می‌شود که ستون w_k بیانگر مرکز خوشه k-ام است. همچنین $H_{k\times n}$ ماتریس درجات عضویت خوشه‌ها است بطوریکه درایه h_{ij} نشان دهنده درجه عضویت d_j در خوشه i-ام است [۷].

۲.۲ اطلاعات از دست رفته

هر چند داده‌های ماسک شده برای حفظ حریم خصوصی و امنیت کاربران ضروری هستند، اما از طرف دیگر بخش مهمی از بقا، کارایی و توسعه شرکت‌ها به این داده‌ها و نتایج حاصل از پردازش آن‌ها وابسته است. درصورتی که داده ناصحیح یا مخدوش در اختیار آن‌ها قرار گیرد، می‌تواند روی تصمیمات کلیدی برای توسعه، تاثیر نامطلوب بگذارد. این تاثیر منفی بنا به کاربرد داده اثر مستقیمی روی جامعه هدف دارد؛ به عنوان مثال در مصارف پزشکی این امر ممکن است سلامت افراد را به خطر اندازد. همچنین در بسیاری از موارد تصمیم برپایه اطلاعات نادرست براحتی منجر به کاهش درآمدهای مالی شرکت و یا حتی ورشکستگی آن می‌گردد. بنابراین بررسی و علم به اینکه ماسک کردن داده برای امنیت چه میزان منجر به تحریف و از دست رفتن اطلاعات حساس داده‌ها می‌شود، اهمیت زیادی دارد. دانشمندان علوم داده معیارهایی[۱۹] را براي اندازه‌گیري میزان اطلاعات از دست رفته گسترش داده‌اند.

تعریف ۲.۲. ([۱۲]) فرض کنید D مجموعه داده اصلي روي دامنه X و D' مجموعه داده ماسك شده آن روی دامنه X' باشند. براي يك تحليل داده[۲۰] مفروض (يعني $f : X \to X'$) ميزان اطلاعات از دست رفته براي مجموعه داده‌هاي D و D' به صورت

[19] Measures

[20] Data analysis

زیر تعریف می‌شود:

$$IL_f(D, D') = divergence(f(D), f(D')), \qquad (۲)$$

که دیورژانس راهی برای مقایسه دو عضو X' می‌باشد. همچنین براي هر دو عضو $D_۱, D_۲ \in X'$ روابط زیر باید برقرار باشند:

- $divergence(D_۱, D_۱) = ۰$,
- $divergence(D_۱, D_۲) \geqslant ۰$,
- $divergence(D_۱, D_۲) = divergence(D_۲, D_۱)$.

ملاحظه ۳.۲. روشن است که هر تابع نرم یا شبه‌نرم روی X' کاندیدایی برای محاسبه دیورژانس است. همچنین با توجه به نوع و کاربري داده‌ها، معیارهاي متنوعي براي محاسبه دیورژانس ارائه می‌شوند (براي مشاهده مثال‌ها و توضیحات بیشتر به [۱۲] و منابع آن رجوع کنید).

۳ روش پیشنهادی

در این مقاله جهت حفظ امنیت و حریم خصوصي داده‌هاي مورد مطالعه، از دیدگاه خوشه‌بندي NMF استفاده می‌کنیم. در دنیای واقعی لازم است که داده‌ها برای استفاده در کاربردهای مختلف سطوح امنیتی متفاوتی داشته باشند. لذا، ما در این مقاله داده‌های ماسک شده متنوعی از نظر میزان شباهت به داده‌های اصلی تولید می‌کنیم تا قدرت اختیار تصمیم گیرنده را برای استفاده از این داده‌ها بالاتر ببریم. علاوه بر این به مساله حفظ محدودیت‌هاي لازم روی ویژگی‌های داده‌ها نیز می‌پردازیم و برقراري یا عدم برقراري آن‌ها را روي داده‌هاي ماسک شده بررسي می‌کنیم.

الگوریتم ۳ فرایند اجرای روش پیشنهادی را نشان می‌دهد.

[!h]

ورودی:

- ماتریس داده اصلی $D = [v_1, \ldots, v_N] \in \mathbb{R}^{F \times N}$.
- محدودیت‌های روی ویژگی‌ها ($\mathcal{C}_i$) برای $i = 1, \ldots, I$.

خروجی:

- یک خانواده از ماتریس‌های داده ماسک شده $\mathcal{A} = \{A^k | k = 3, \ldots, rank(D)\}$ است.
- خطاهای $error^k$ و e_i^k.

گام ۱. رتبه ماتریس D را بیابید و قرار دهید $r = rank(D)$.

گام ۲. برای رتبه های $k = 3, \ldots, r$

خوشه بندی $NMF(D, k)$ را محاسبه کنید.

ماتریس های W^k و H^k را بیابید.

برای $j = 1, \ldots, N$

بردارهای داده ماسک شده

$$a_j^k \approx \sum_{l=1}^{k} H_{lj}^k W_l^k \in \mathbb{R}^F \qquad (3)$$

را محاسبه کنید.

قرار دهید $A^k = [a_j^k]_{j=1,\ldots,N}$.

میزان خطاي نسبي را حساب کنید:

$$error^k = \frac{1}{FN} \left(\sum_{j=1}^{N} \| v_j - a_j^k \|_1 \right). \qquad (4)$$

گام ۳. برای ماتریس‌های ماسک شده A^k، $k = 3, \ldots, r$

برای $i = 1, \ldots, I$

اگر محدودیت $\mathcal{C}_i$ روی A^k برقرار است قرار دهید $e_i^k = 0$،

در غیر اینصورت با توجه به نوع محدودیت میزان خطای نقض آن، e_i^k، را محاسبه کنید.

توضیح الگوریتم روش پیشنهادی. الگوریتم ۳ تعداد N داده را به عنوان ورودی می‌پذیرد که دارای F ویژگی کمی هستند. همچنین با توجه به تاثیرات بین داده‌ای در دنیای واقعی ممکن است محدودیت‌هایی روی آنها وجود داشته باشد. به عنوان مثال فرض کنید که یک کارخانه قصد تولید N محصول را داشته باشد و ویژگی‌های اول، دوم و سوم به ترتیب نشان‌دهنده هزینه نیروی انسانی، هزینه مواد اولیه و هزینه کل برای تولید هر واحد از محصولات باشند. بنابراین ویژگی سوم (سطر ۳-ام ماتریس D) ترکیب خطی از دو ویژگی دیگر است. لذا داده‌های ماسک شده نیز باید در این محدودیت صدق کنند.

گام ۱ جهت ساخت داده‌های ماسک شده با سطوح امنیتی متفاوت (A^k) رتبه ماتریس داده (D) را محاسبه می‌کند.

گام ۲ تجزیه‌های NMF با رتبه‌هاي $k = ۳\ldots, rank(D)$ را بدست مي‌آورد. لازم به ذکر است تجزیه NMF براي رتبه‌هاي $k = ۱, ۲$ نتایج قابل استنادی تولید نمي‌کند، بنابراین این رتبه‌ها بررسي نشده‌اند. همچنین در این گام براي هر ماتریس ماسك شده A^k میزان خطای نسبی $error^k$ یعني میزان تفاوت نسبي A^k با داده اصلي D محاسبه مي‌شود.

گام ۳ خطاهای حاصل از نقض محدودیت e_i^k را براي هر مجموعه داده‌ي ماسك شده A^k و هر محدودیت C_i محاسبه می‌کند.

ملاحظه ۱.۳. الگوریتم ۳ متناظر با هر ماتریس ماسک شده A^k، $k = ۳, \ldots, r$ یک دنباله از خطاهای $error^k$ و e_i^k، $i = ۱, \ldots, I$ نیز تولید می‌کند. تصمیم گیرنده می‌تواند بر اساس خطاها، شرایط موجود و بویژه نوع داده مورد مطالعه؛ داده مناسب را جهت انتشار انتخاب نماید. به عنوان نمونه در مسائل داده‌کاوی بهتر است A^*-یی انتخاب شود که برای آن

$$error^* = \min_{k=۳,\ldots,r} error^k.$$

هزینه ۱۶٪	هزینه ۷٪	کل هزینه	ساعات پرداخت شده	نرخ دستمزد	دستمزد	مجموع ساعت کل
$f_۱$	$f_۲$	$f_۳$	$f_۴$	$f_۵$	$f_۶$	$f_۷$
۱۵	۲۳	۴۲٫۰۱	۲۳	۵۰	۱۱۵۰	۳۷
۱۲	۴۳	۵۹٫۹۳	۲۸	۷۰	۱۹۶۰	۳۷
۶۴	۲۲۹	۳۱۹٫۲۷	۱۲	۸۴	۱۰۰۸	۲۵
۱۲	۴۵	۶۲٫۰۷	۲۹	۷۳	۲۱۱۷	۳۰
۲۸	۳۹	۷۴٫۲۱	۹	۳۰	۲۷۰	۴۰
۷۱	۱۰۲	۱۹۱٫۵	۱۰	۶۳	۶۳۰	۲۰
۲۳	۶۴	۹۵٫۱۶	۹	۷۴	۶۶۶	۱۰
۲۵	۱۰۲	۱۳۸٫۱۴	۷۲	۳۰	۲۱۶۰	۸۰
۴۸	۲۳۰	۳۰۱٫۷۸	۲۶	۳۰	۷۸۰	۳۵
۳۲	۵۰	۹۰٫۶۲	۶	۴۵	۲۷۰	۱۵
۹۰	۲۰۰	۳۱۸٫۴	۸	۴۵	۳۶۰	۱۵
۱۶	۱۰۰	۱۲۵٫۵۶	۳۴	۵۵	۱۸۷۰	۴۵

جدول ۱: *مجموعه داده‌هاي اصلي*

از طرفي در مسائل آماري كه ارتباطات ميان داده‌اي از اهميت زيادي برخوردارند، بهتر است A^*-يي جهت انتشار انتخاب شود كه در آن ميزان خطاهاي نقض محدوديتها، e_i^k، به نسبت كمتر باشد.

تورا [۱۳، ۱۴] مثال كوچكي از يك مجموعه داده همراه با تعدادي محدوديت ارايه داده است. در اينجا جهت روشنتر شدن روش پيشنهادي، الگوريتم ۳ را روي اين مثال پيادەسازي مي‌كنيم. لازم به ذكر است كه اين مجموعه داده در حالت كلي و گسترش يافته آن [۵] در بخش ۴ مطالعه خواهد شد.

مثال ۲.۳. جدول ۱ شامل داده‌هاي دوازده نفر است. اين جدول اطلاعات افراد روي هفت ويژگي هزينه ۱۶٪، هزينه ۷٪، هزينه كل، ساعات پرداخت شده، نرخ دستمزد، مجموع دستمزد و ساعت كل را شامل مي‌شود. همانطور كه در جدول ملاحظه مي‌كنيد مقادير اين ويژگي‌ها با f_j، $j = ۱, \ldots, ۷$ نشان داده شده‌اند. همچنين داده‌ها در سه محدوديت زير صدق مي‌كنند:

$$\begin{aligned} \mathcal{C}_۱ &: \quad f_۳ = \alpha_۱ f_۱ + \alpha_۲ f_۲, \\ \mathcal{C}_۲ &: \quad f_۶ = f_۴ * f_۵, \\ \mathcal{C}_۳ &: \quad f_۴ \leq f_۷. \end{aligned} \tag{۵}$$

ورودی: روشن است که D برابر با ترانهاده ماتریس داده در جدول ۱ است. همچنین $N = ۱۲$ و $F = ۷$.

گام ۱. $r = rank(D) = ۶$.

گام ۲. برای تولید خانواده داده‌های ماسک شده $\mathcal{A} = \{A^۳, \ldots, A^۶\}$ ، گام ۲ چهار بار تکرار میشود. همچنین در هر تکرار میزان خطای نسبی $(error^k)$ از (۴) محاسبه شده و که این مقادیر در جدول ۲ ارائه شده‌اند.

گام ۳. این گام به بررسي برقراري محدوديت‌ها روي ماتريس‌هاي ماسك شده اختصاص دارد.

بررسی محدودیت $\mathcal{C}_۱$

ابتدا روی ماتریس D با حل مساله کمترین مربعات[۲۱]:

$$\min_{(\alpha_۱,\alpha_۲)} \parallel f_۳ - (\alpha_۱ f_۱ + \alpha_۲ f_۲) \parallel_۲ \qquad (۶)$$

مقدار $(\alpha_۱^*, \alpha_۲^*) = (۱/۱۶, ۱/۰۷)$ را محاسبه مي‌كنيم. لازم به ذكر است كه جهت حل مساله (۶) از روش تجزیه QR [۷] استفاده مي‌كنيم. سپس مساله (۶) را به طور مشابه برای هر کدام از A^kها بازنویسی و $(\alpha_۱^k, \alpha_۲^k)$ را محاسبه مي‌كنيم. در اينجا برای محاسبه مقدار خطای نقض محدودیت $\mathcal{C}_۱$، فرمول زیر را پیشنهاد می‌کنیم:

$$e_۱^k = |\alpha_۱^k - \alpha_۱^*| + |\alpha_۲^k - \alpha_۲^*| \qquad (۷)$$

بررسی محدودیت $\mathcal{C}_۲$

$f_۶ = f_۴ * f_۵$ يك محدوديت غيرخطي است. ما فرمول زير را جهت خطاي حاصل از نقض محدوديت $\mathcal{C}_۲$ براي هر A^k، تعريف مي‌كنيم:

[21] Least square

$$e_{۲}^{k} = \frac{۱}{۱۲}\left(\sum_{j=۱}^{۱۲}\left|A_{۶j}^{k} - \left(A_{۴j}^{k} * A_{۵j}^{k}\right)\right|\right). \qquad (۸)$$

بررسی محدودیت $\mathcal{C}_{۳}$

در این حالت صدق یا عدم صدق نامساوي $f_{۴} \leq f_{۷}$ باید تحقیق شود. لذا براي هر A_k،

$$e_{۳}^{k} = \begin{cases} \frac{۱}{۱۲}\sum_{A_{۴j}^{k} > A_{۷j}^{k}}\left(A_{۴j}^{k} - A_{۷j}^{k}\right) & \text{if} \exists j \in \{۱, \ldots, ۱۲\} : A_{۴j}^{k} > A_{۷j}^{k} \\ ۰. & \text{O.W.} \end{cases} \qquad (۹)$$

خلاصه نتایج را در جدول ۲ ملاحظه می‌کنید. علامت ✓ نشان‌دهنده برقراری محدودیت در ماتریس داده ماسک شده است. البته درصورت عدم برقراری محدودیت، مقدار خطای نقض نیز ارائه شده است. در پایان تصمیم گیرنده آزادی زیادی جهت انتخاب یک داده ماسک شده برای انتشار دارد. به عنوان نمونه $A^{۳}$ و $A^{۴}$ هر دو تقریبا در تمام محدودیتها صدق می‌کنند هر چند میزان خطای نسبی متفاوتی دارند. در شرایطی که ترجیح می‌دهیم داده با شباهت بیشتر به داده اصلی (درعین وجود تفاوت) را منتشر کنیم، بهتر است $A^{۳}$ را به عنوان داده منتشر شده انتخاب کنیم. اما اگر خطر افشا خیلي زیاد باشد، بهتر است داده $A^{۴}$ را منتشر کنیم.

A^k	$A^{۳}$	$A^{۴}$	$A^{۵}$	$A^{۶}$
$error^k$	۰/۱۱۱	۰/۱۸۷	۰/۱۵۳	۰/۱۵۴
$(\alpha_{۱}^k, \alpha_{۲}^k)$	(۱/۱۶, ۱/۰۷)	(۱/۱۶, ۱/۰۷)	(۱/۱۵, ۱/۰۵۹)	(۱/۱۲, ۱/۰۵)
$\mathcal{C}_{۱}^k$	✓	✓	$e_{۱}^{۵} = ۰/۰۱۷$	$e_{۱}^{۵} = ۰/۰۵۳$
$\mathcal{C}_{۲}^k$	$e_{۲}^{۵} = ۰/۰۰۲$	$e_{۲}^{۵} = ۰/۰۰۳$	$e_{۲}^{۵} = ۰/۰۰۱$	$e_{۲}^{۵} = ۰/۰۰۲$
$\mathcal{C}_{۳}^k$	✓	✓	✓	✓

جدول ۲: *نتایج اجرا الگوریتم ۳ روی مثال ۲.۳*

۴ نتایج تجربی

در این بخش عملکرد روش پیشنهادی از دو دیدگاه مورد مطالعه قرار مي‌گيرد. در دیدگاه اول، الگوریتم ۳ روي مجموعه داده استاندارد $Census\ data\ set$ [۵] كه يك محك [۲۲] در روش‌هاي ريزتجميع كننده و حافظ حريم خصوصي است، پياده سازي مي‌شود. رویکرد دوم به بررسي كارايي و نحوه عملكرد داده‌هاي ماسك شده در کاربردهای عملی به عنوان نمونه: داده‌كاوي، مي‌پردازد.

پيـاده‌سـازي‌ها در نـرم افـراز مـتـلـب ۲۰۲۱(a) در يـك سـيسـتـم بـا پـردازنـده مـركـزي Core ۷i بـا حـافـظـه اصـلـي Gb۸ انـجـام شـده‌انـد. از آنـجـا كـه NMF روشـی اسـت كـه در اجـراهای مـتـفـاوت نـتـايـج يـكـسـان تـولـيـد نـمـی‌كـنـد و الـگـوريـتـم ۳ از NMF بـهره مـی‌بـرد، لـذا نـتـابـج اعـلام شـده مـقـدار مـيـانـگـيـن ده بار اجرا هستند.

۱.۴ رویکرد اول پیاده‌سازی

در رویکرد اول شکل اصلی مجموعه داده مطرح شده در مثال ۲.۳ یعنی داده استاندارد $Census\ data\ set$ [۵] بررسی شده است. در این مجموعه، اطلاعات هزاروهشتاد نفر $(N = ۱۰۸۰)$ روی سیزده ویژگی مختلف $(F = ۱۳)$ گردآوری شده که سه محدودیت (۵) نیز در آن برقرار است. خلاصه نتایج پیاده‌سازی الگوریتم ۳ روی مجموعه داده $Census\ data\ set$ در جدول ۳ آمده است. توجه کنید که در این مثال $rank(D) = ۱۲$ محاسبه شده است.

در توضیح جدول ۳ موارد زیر قابل ذکر هستند:

- با توجه به اینکه $rank(D) = ۱۲$ به کمك الگوريتم ۳، نه مجموعه داده ماسك شده $(A^{۳} - A^{۱۲})$ توليد مي‌شود.

[22] Benchmarck

A^k	$A^۳$	$A^۴$	$A^۵$	$A^۶$	$A^۷$	$A^۸$	$A^۹$	$A^{۱۰}$	$A^{۱۱}$	$A^{۱۲}$
$error^k$	۰٫۰۰۰۵	۰٫۰۰۰۵	۰٫۰۰۱۱	۰٫۰۰۱	۰٫۰۰۰۱	۰٫۰۰۰۱	۰٫۰۰۱۲	۰٫۰۰۰۳	۰	۰
$\begin{pmatrix}\alpha_۱\\ \alpha_۲\end{pmatrix}$	$\begin{pmatrix}۰٫۰۰۳۶\\ ۰٫۰۴۱۰\end{pmatrix}$	$\begin{pmatrix}۰٫۰۰۳۶\\ ۰٫۰۴۱۱\end{pmatrix}$	$\begin{pmatrix}۰٫۰۰۳۷۵\\ ۰٫۰۴۰۸\end{pmatrix}$	$\begin{pmatrix}۰٫۰۰۳۶\\ ۰٫۰۴۳۶\end{pmatrix}$	$\begin{pmatrix}۰٫۰۱۷۶\\ ۰٫۰۰۶۴۹\end{pmatrix}$	$\begin{pmatrix}۰٫۰۰۳۲\\ ۰٫۰۳۵۷\end{pmatrix}$	$\begin{pmatrix}۰٫۰۰۳۳\\ ۰٫۰۴۲۸\end{pmatrix}$	$\begin{pmatrix}۰٫۰۰۲۳\\ ۰٫۰۴۳۴\end{pmatrix}$	$\begin{pmatrix}۰٫۰۰۸۲\\ ۰٫۰۰۳۳\end{pmatrix}$	$\begin{pmatrix}۰٫۰۰۴۵\\ ۰٫۰۴۹۵\end{pmatrix}$
$e_۱^k$	۰٫۰۰۷۸	۰٫۰۰۸۷	۰٫۰۰۴۳	۰٫۰۰۳۵	۰٫۰۱۶۲	۰٫۰۰۳۰	۰٫۲۰۴	۰٫۱۱۱	۰٫۰۰۸۰	۰٫۰۱۲۷
$e_۲^k$	۷۲×۱۰^{-۷}	۸۲×۱۰^{-۷}	۲×۱۰^{-۷}	۲×۱۰^{-۷}	۵×۱۰^{-۷}	۵×۱۰^{-۷}	۱×۱۰^{-۷}	۲×۱۰^{-۷}	۳۸×۱۰^{-۶}	۷×۱۰^{-۷}
$e_۳^k$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

جدول ۳: *نتايج پيادهسازي الگوريتم ۳ روي مجموعه داده Census data set*

- در سطر دوم برابر بودن میزان خطاهای نسبی برای دو نوع داده ماسک شده به معنای یکی بودن آن‌ها نیست. به عنوان نمونه اگر چه $error^۲ = error^۳$ اما مسلما $A^۲ \neq A^۳$. در واقع این‌ها دو مجموعه داده ماسک شده متفاوت هستند که از نظر تصمیم گیرنده با توجه به شرایط دیگر ممکن است یکی از آن‌ها بر دیگری ارجح باشد.

- سطرهاي سوم و چهارم به بررسي محدوديت اول (۵) می‌پردازند. با حل مساله (۶) روي مجموعه داده اصلي $Census\ data\ set$، $\alpha_۱ = ۰٫۰۰۳۵$ و $\alpha_۲ = ۰٫۰۴۱۷$ محاسبه شده‌اند. سطر سوم مقادير $\alpha_۱$ و $\alpha_۲$ حاصل از حل مساله (۶) براي داده‌هاي ماسك شده متفاوت و سطر چهارم ميزان خطاي نقض محدوديت اول (۷) را نمايش مي‌دهند.

- در سطر پنجم، ميزان خطاي نقض محدوديت دوم (۵) محاسبه شده از رابطه (۹) آمده است. با توجه به مقادير حاصل، از نظر يك تصميم گيرنده خوشبين محدوديت $\mathcal{C}_۲$ در مجموعه‌هاي ماسك شده نيز برقرار است. هر چند در اين جدول ما براي تصميم گيرنده بدبين خطاها را اعلام كرده‌ايم.

- نماد ✓ در سطر آخر نشانگر برقراري محدوديت سوم (۵) در تمام مجموعه‌هاي ماسك شده است.

در مقایسه با کارهای تورا [۱۴، ۱۳] با توجه به NP- سخت بودن روش‌های ریزتجمیع کننده، روش فازی پیشنهادی خود را روی زیر مجموعه بسیار کوچکی از مجموعه داده [۵] (به ترتیب با ۱۲ داده و سه ویژگی؛ ۱۲ داده و هفت ویژگی) پیاده‌سازی کرده است. روش

های ایشان بر پایه یافتن مراکز خوشه متفاوت فقط با اعمال نویز با توزیع های نرمال متفاوت است و همچنین در [۱۴] تنها برقراری محدودیت اول و در [۱۳] سه محدودیت (۵) را بررسی کرده است.

۲.۴ رویکرد دوم پیاده‌سازی

در این بخش میزان اطلاعات از دست رفته را در کاربرد عملی (خوشه‌بندي) زمانیکه داده‌ها با روش پیشنهادی جهت تامین امنیت ماسک شده‌اند، می‌سنجیم. بدین منظور از شش مجموعه داده استاندارد مورد استفاده در روش‌هاي داده‌کاوي استفاده مي‌کنیم. ابتدا براي هر مجموعه داده با روش c- میانگین فازي (FCM)[۲۳] [۱] مجموعه داده اصلي را خوشه‌بندي مي‌کنیم، سپس عملکرد روش FCM را با استفاده از معیار دقت نرخ خوشه‌بندي[۲۴]:

$$CR_{FCM} = \frac{\text{تعداد نمونه‌هایی که بدرستی خوشه‌بندی شده‌اند}}{\text{تعداد کل نمونه‌ها}}, \qquad (۱۰)$$

ارزیابي مي‌کنیم. در ادامه، روش FCM را روي مجموعه داده‌هاي ماسك شده پیاده‌سازي و CR_{FCM} را براي آن‌ها محاسبه می‌کنیم. در پایان، براي بررسي میزان اطلاعات از دست رفته از تعریف ۲ استفاده مي‌کنیم. در این راستا، تابع f مساله خوشه‌بندي است و دیورژانس را مقایسه CR_{FCM}ها در نظر مي‌گیریم و معیار زیر را پیشنهاد مي‌دهیم:

$$IL(D, A^k) = |CR_{FCM}(D) - CR_{FCM}(A^k)|. \qquad (۱۱)$$

جزییات مجموعه‌هاي داده و نتایج پیاده‌سازي در ادامه توضیح داده شده است.

- مجموعه داده Haberman's Survival شامل سیصد و شش داده ($N = ۳۰۶$) روي سه ویژگي ($F = ۳$) است که در دو کلاس دسته‌بندي مي‌شوند [۲۱]. نرخ

[23] Fuzzy c-means (FCM)

[24] Clustering rate (CR)

خوشه‌بندی مجموعه داده اصلی برابر با $CR_{FCM} = ۰/۶۵۰۳$ و $rank(D) = ۳$ هستند. خلاصه نتایج پیاده‌سازی را در جدول ۴ مشاهده می‌کنید.

	$A^۳$
$error^k$	$۴/۰۹ \times ۱۰^{۱۴}$
CR_{FCM}	۰/۵۲۶۱
IL	۰/۱۲۴۲

جدول ۴: مجموعه داده *Haberman's Survival*

- مجموعه داده Iris شامل صد و پنجاه داده ($N = ۱۵۰$) روي چهار ويژگي ($F = ۴$) است که در سه کلاس دسته‌بندي مي‌شوند [۲۰]. نرخ خوشه‌بندی مجموعه داده اصلی برابر با $CR_{FCM} = ۰/۸۹۳۳$ و $rank(D) = ۴$ هستند. خلاصه نتایج پیاده‌سازی را در جدول ۵ مشاهده می‌کنید.

	$A^۳$	$A^۴$
$error^k$	۰	$۲/۰۳۷ \times ۱۰^{۱۵}$
CR_{FCM}	۰/۹۰۰۰	۰/۸۹۳۳
IL	۰/۰۱	۰

جدول ۵: مجموعه داده *Iris*

- مجموعه داده Glass شامل دويست و چهارده داده ($N = ۲۱۴$) روي نه ويژگي ($F = ۹$) است که در شش کلاس دسته‌بندي مي‌شوند [۲۲]. نرخ خوشه‌بندی مجموعه داده اصلی برابر با $CR_{FCM} = ۰/۵۳۷۴$ و $rank(D) = ۹$ هستند. خلاصه نتایج پیاده‌سازی را در جدول ۶ مشاهده می‌کنید.

	$A^۳$	$A^۴$	$A^۵$	$A^۶$	$A^۷$	$A^۸$	$A^۹$
$error^k$	۴/۰۲	۳/۷۴	۲/۲۴	۳/۲۳	۴/۱۶	۲/۷۱	$۶/۲۴ \times ۱۰^{۱۲}$
CR_{FCM}	۰/۵۷	۰/۵۷	۰/۴۰	۰/۵۶	۰/۵۷	۰/۵۶	۰/۵۳
IL	۰/۰۴	۰/۰۳	۰/۰۱	۰/۰۲	۰/۰۵	۰/۱	۰

جدول ۶: مجموعه داده *Glass*

- مجموعه داده Breast cancer شامل ششصد و نود و نه داده ($N = ۶۹۹$) روي ده ويژگي ($F = ۱۰$) است که در دو کلاس دسته‌بندي مي‌شوند [۱۹]. نرخ خوشه‌بندی روي مجموعه داده‌ي اصلي برابر با $CR_{FCM} = ۰/۹۸۳۶$ و $rank(D) = ۹$ هستند. خلاصه نتایج پیاده‌سازی را در جدول ۷ مشاهده می‌کنید.

	$A^{۳}$	$A^{۴}$	$A^{۵}$	$A^{۶}$	$A^{۷}$	$A^{۸}$	$A^{۹}$
$error^k$	۱/۲۷	۱/۴۸	۱/۸۳	۲/۳۵	۳/۹۱	۱۰/۲۵	۱×۱۰^{۱۵}
CR_{FCM}	۰/۹۵۵۷	۰/۹۶۵۷	۰/۹۶۱۴	۰/۹۵۵۷	۰/۹۵۷۱	۰/۹۵۵۷	۰/۹۵۴۲
IL	۰/۰۲۷	۰/۰۱۷	۰/۰۲۲	۰/۰۲۷	۰/۰۲۶	۰/۰۲۷	۰/۰۲۹

جدول ۷: مجموعه داده *Breast Cancer*

- مجموعه داده Wine شامل صد و هفتاد و هشت داده ($N = ۱۷۸$) روي سيزده ويژگي ($F = ۱۳$) است که در سه کلاس دسته‌بندي مي‌شوند [۱۸]. نرخ خوشه‌بندي روش c- ميانگين فازي روي داده‌هاي اصلي برابر با $CR_{FCM} = ۰/۶۸۵۴$ است. از آنجا که $rank(D) = ۱۳$ پس الگوریتم ۳ یازده مجموعه داده متفاوت ($A^{۳} - A^{۱۳}$) ایجاد می‌کند. خلاصه نتایج پیاده‌سازی الگوریتم ۳، روش FCM و اطلاعات ازدست رفته در جدول ۸ نشان داده شده‌اند.

	$A^{۳}$	$A^{۴}$	$A^{۵}$	$A^{۶}$	$A^{۷}$	$A^{۸}$	$A^{۹}$	$A^{۱۰}$	$A^{۱۱}$	$A^{۱۲}$	$A^{۱۳}$
$error^k$	۰/۲۴	۱/۲۰	۰/۰۷	۱/۲۴	۱/۲۳	۱/۲۴	۲/۲۵	۱/۲۵	۱/۹۸	۱/۹۹۸	$۸/۲۲۵ \times ۱۰^{۱۰}$
CR_{FCM}	۰/۶۸۵۴	۰/۶۸۵۴	۰/۷۴۷۲	۰/۶۸۵۴	۰/۶۸۵۴	۰/۶۸۵۴	۰/۶۸۵۴	۰/۶۸۵۴	۰/۶۸۵۴	۰/۶۸۵۴	۰/۶۸۵۴
IL	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰

جدول ۸: مجموعه داده *Wine*

- مجموعه داده Heart SPECT شامل دویست و شصت و هفت داده ($N = ۲۶۷$) روي بيست و دو ويژگي ($F = ۲۲$) است که در دو کلاس دسته‌بندي مي‌شوند [۲۳]. نرخ خوشه‌بندی روي مجموعه داده‌ي اصلي برابر با $CR_{FCM} = ۰/۶۲۹۲$ و $rank(D) = ۲۲$ هستند. خلاصه نتایج پیاده‌سازی را در جدول‌های ۹ و ۱۰ مشاهده می‌کنید.

	$A^{۳}$	$A^{۴}$	$A^{۵}$	$A^{۶}$	$A^{۷}$	$A^{۸}$	$A^{۹}$	$A^{۱۰}$	$A^{۱۱}$
$error^k$	۴٫۱۵	۴٫۹۱	۵٫۴۱	۵٫۷۳	۶٫۴۲	۶٫۸۲	۶٫۹۸	۷٫۶۶	۷٫۸۳
CR_{FCM}	۰٫۵۹	۰٫۶۰	۰٫۶۱	۰٫۶۱	۰٫۶۱	۰٫۶۱	۰٫۶۲	۰٫۶۲	۰٫۶۲
IL	۰٫۰۳	۰٫۰۲۶	۰٫۰۱	۰٫۰۱	۰٫۰۱	۰٫۰۱	۰٫۰۰۷	۰٫۰۰۳	۰٫۰۰۷

جدول ۹: مجموعه داده *SPECT Heart* (قسمت اول)

	$A^{۱۲}$	$A^{۱۳}$	$A^{۱۴}$	$A^{۱۵}$	$A^{۱۶}$	$A^{۱۷}$	$A^{۱۸}$	$A^{۱۹}$	$A^{۲۰}$	$A^{۲۱}$	$A^{۲۲}$
$error^k$	۸٫۸۷	۹٫۸۹۱	۰٫۰۲	۱۰٫۸۹	۱۳٫۴۰	۱۴٫۷۴	۱۸٫۲۷	۱۸٫۲۰	۲۱٫۳۳	۳۶٫۴۲	۴٫۶۷[۱۵]
CR_{FCM}	۰٫۶۱	۰٫۶۲	۰٫۶۳	۰٫۶۱	۰٫۶۲	۰٫۶۱	۰٫۶۱	۰٫۶۲	۰٫۶۱	۰٫۶۱	۰٫۶۰
IL	۰٫۰۱	۰٫۰۰۷	۰٫۰۰۳	۰٫۰۱	۰	٫۰۱	۰٫۰۱	۰٫۰۰۳	۰٫۰۱	۰٫۰۱	۰٫۰۲

جدول ۱۰: مجموعه داده *SPECT Heart* (قسمت دوم)

در جدول‌های ۱۰-۴ میزان خطای نسبی $error^k$ برای $k = rank(D)$ از سایر kها بیشتر و عدد خیلی بزرگی است. از طرف دیگر میزان اطلاعات از دست رفته برای تمام kها تقریبا یکسان و مقدار ناچیزی است. لذا انتخاب $A^{rank(D)}$ به عنوان داده منتشر شده، ضمن حفظ حریم خصوصی به طور موثر، برای کاربران قانونی که قصد پردازش داده‌ها را دارند؛ نیز بسیار مناسب است.

۵ نتیجه‌گیری و پیشنهاد برای کارهای آتی

این مقاله به مساله چالش برانگیز انتشار داده‌ها با توجه به حفظ حریم خصوصي، حفظ ويژگي‌ها و محدوديت‌هاي داده‌هاي اصلي و كاهش اطلاعات از دست رفته، پرداخته است. در اين راستا يك الگوريتم بر اساس تجزيه نامنفي ماتريس‌ها طراحي شده كه داده‌هايي با سطوح امنيت متفاوت را براي انتشار توليد مي‌كند. رفتار الگوريتم روي يك مجموعه داده استاندارد پيچيده در مسائل امنيت كه همراه با محدوديت‌هاي ساختاري است، بررسي شده است. نتايج تجربی نشان مي‌دهند كه روش پيشنهادی در مقايسه با برخی از روش‌های فازی علاوه بر حفظ امنيت، ساختار محدوديت‌هاي داده‌های اصلی با اندازه و ابعاد بزرگ را همچنان روي داده‌هاي منتشر شده نيز حفظ می‌كند. از طرف ديگر نتايج پياده‌سازي الگوريتم روي سه مجموعه داده متنوع نشان دهنده كارايي الگوريتم پيشنهادي

از منظر افزايش امنيت و كاهش اطلاعات از دست رفته در مساله خوشه‌بندي، روي اين مجموعه‌ها است. در پايان، با توجه به كارايي روش مطرح شده در اين مقاله، توسعه روش تجزيه نامنفي ماتريس‌ها بصورت فازی به عنوان يك ريزتجميع‌كننده را براي تحقيقات آتي پيشنهاد مي‌كنيم.

مراجع

[1] Bezdek, J.C., Ehrlich, R, Full, W. (1984) FCM: the fuzzy c-means clustering algorithm, Computers and Geo-sciences , 10 , 191-203.

[2] Berrya, M.W, Browne, M, Langville, A.N, Pauca, V.P and R.J. Plemmons. (2007) Algorithms and applications for approximate nonnegative matrix factorization, Computational Statistics and Data Analysis, 52, 155–173.

[3] Castro, O., Gentile, C., Spagnolo-Arrizabalaga, E. (2022) An algorithm for the microaggregation problem using column generation, Computers and Operations Research,68, 105817.

[4] César Fadel, A., Satoru Ochi, L., André de Moura Brito, J., Silva Semaan, G. (2021)Microaggregation heuristic applied to statistical disclosure control, Information Sciences, 548, 37-55.

[5] Domingo-Ferrer, J., and Mateo-Sanz J.M. (2002) Practical data-oriented microaggregation for statistical disclosure control, IEEE Transactions on Knowledge and Data Engineering, 14, 189-201.

[6] Edgar, B. Antoni,M. Agusti, A. (2022) Privacy-preserving process mining: A microaggregation-based approach, Journal of Information Security and Applications, 68, 103235.

[7] Elden, L. (2007) *Matrix Methods in Data Mining and Pattern Recognition*, Society for Industrial and Applied Mathematics 106-110.

[8]]Hansen, S.L, Mukherjee, S. (2003) A polynomial algorithm for optimal univariate microaggregation, in IEEE Transactions on Knowledge and Data Engineering, 4, 1043-1044.

[9] Kiran, A., and Shirisha, N. (2022) K-Anonymization approach for privacy preservation using data perturbation techniques in data mining. Materials Today: Proceedings.

[10] Oganian, A, Domingo-Ferrer, J. (2000) On the Complexity of Optimal Microaggregation for Statistical Disclosure Control, Statistical Journal of the United Nations Economic Commission for Europe, 4, 345–354.

[11] Rodriguez-Garcia, M., Batet, M., Sánchez, D. (2019) Utility-preserving privacy protection of nominal data sets via semantic rank swapping, Information Fusion, 45, 282-295.

[12] Torra, V. (2017) Masking methods. In: Torra, V. (ed.) *Data Privacy: Foundations, New Developments and the Big Data Challenge*. Studies in Big Data, 28, 191–238.

[13] Torra, V. (2008) Constrained microaggregation: adding constraints for data editing, Transactions on data privacy, 1, 86–104.

[14] Torra, V. (2020) Fuzzy Clustering-based Microaggregation to Achieve Probabilistic K-anonymity for Data with Constraints, Journal of Intelligent and Fuzzy Systems, 39, 5999–6008.

[15] Vaidya, J, Zhu, Y. and C. Clifton, *Privacy Preserving Data Mining*, in Advances in Information Security, Springer, **19** 2006, 1-121.

[16] Yao, A.C. (1982) Protocols for secure computations, 23rd Annual Symposium on Foundations of Computer Science, 160-164.

[17] Wang, Y. X, and Zhang, Y. J. (2013) Nonnegative Matrix Factorization: A Comprehensive Review, in IEEE Transactions on Knowledge and Data Engineering, 6, 1336-1353.

[18] https://archive.ics.uci.edu/ml/datasets/wine.

[19] https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original).

[20] https://archive.ics.uci.edu/ml/datasets/iris.

[21] https://archive.ics.uci.edu/ml/datasets/haberman’s+survival.

[22] https://archive.ics.uci.edu/ml/datasets/glass+identification.

[23] https://archive.ics.uci.edu/ml/datasets/spect+heart.