رویکرد فازی به حفظ حریم خصوصی در خوشه بندی با الگوریتم امید ریاضی بیشینه‌سازی

لیلا جعفر تفرشی [۱] و فرزین یغمایی [۲]

[۱] دانشجوی کارشناسی ارشد، دانشکده برق و کامپیوتر، دانشگاه سمنان
Leila.tafreshi۶۶@yahoo.com

[۲] استادیار، دانشکده برق و کامپیوتر، دانشگاه سمنان
F_yaghmaee@semnan.ac.ir

چکیده

امروزه حجم زیاد اطلاعات شخصی و محرمانه و گسترش آن‌ها در وب به منظور فعالت های تجاری، آماری و بسیاری از کاربردهای دیگر، حفظ محرمانگی داده‌ها را به چالش مهمی تبدیل کرده است. در نتیجه تکنیک‌های داده کاوی با معضل مهم محافظت از داده-های حساس مانند داده‌های بانکی، پزشکی و سایر اطلاعات محرمانه‌ی افراد مواجه شده است و شاخه‌ی جدیدی از داده کاوی به نام حفظ حریم خصوصی در داده کاوی پا به عرصه نهاده است. هدف تحقیقات در این زمینه توسعه روشی است که بتواند بدون انتشار داده‌های محرمانه، داده کاوی را انجام داده و به نتایج آن خدشه‌ای وارد نکند. در این تحقیق روشی برای حفظ حریم خصوصی در داده کاوی ارائه شده است که در آن با اعمال توابع عضویت فازی روی داده‌های حساس، علاوه بر اینکه از داده‌های محرمانه به خوبی محافظت می‌شود، نتایج معتبری از خوشه بندی با روش امید ریاضی بیشینه‌سازی حاصل می‌گردد.

کلمات کلیدی

حفظ حریم خصوصی در داده کاوی، منطق فازی، تابع عضویت فازی، خوشه بندی، الگوریتم امید ریاضی بیشینه‌سازی.

۱- مقدمه

داده کاوی فرآیند انتخاب، شناسایی و مدل کردن کمیت‌های عظیمی از داده‌ها برای کشف و استخراج قوانین و ارتباط موجود بین آنها است تا از این قوانین و ارتباط به اطلاعات و دانشی از مجموعه داده ها برسیم.
کاربردهای داده کاوی شامل موارد زیر است:

- قوانین وابستگی [۱]: الگوهایی که در آن وجود یک آیتم دلالت بر وجود آیتم دیگر دارد.
- کلاسبندی[۲]: انتساب الگوها به یک مجموعه‌ی کوچک از کلاس‌های از قبل تعریف شده به وسیله کشف بعضی روابط بین ویژگی‌ها.
- خوشه‌بندی[۳] : گروه‌بندی داده‌ها یا مجموعه الگوهایی که ویژگی‌های مشابهی دارند.
- پیشگویی [۴] :کشف الگوها برای پیشگویی منطقی درباره‌ی آینده.

داده کاوی یک تکنولوژی جدید نیست ولی کاربرد آن به طور معناداری در بخش‌های مختلف خصوصی و عمومی روبه رشد بوده و عموما صنایعی چون بانک، بیمه، پزشکی و خرده فروشی، از داده کاوی به هدف کاهش هزینه‌ها، افزایش تحقیقات و افزایش فروش استفاده می‌کنند[۱].

هر چند داده کاوی در بسیاری از برنامه‌ها تأثیرات مثبتی بر جای گذاشته است، ولی به خاطر احتمال افشای داده‌های خصوصی، باعث نگرانی‌هایی نیز شده است. هنوز هیچ تضمینی ارائه نشده است که بتوان داده‌ها را بدون تجاوز به حریم خصوصی مالک آن مورد داده کاوی قرار داد.

بعنوان مثال در شبکه‌های اجتماعی اطلاعات مختلف کاربران از جمله سن، حرفه، هویت و سایر اطلاعاتی که باعث شناسایی دقیق فرد می‌شود نیاز به محافظت دارند. در یک سیستم پزشکی، نحوه انجام داده کاوی در اطلاعات خصوصی بیماران بدون افشای اطلاعات محرمانه، یکی از مسائلی است که با آن روبه رو هستیم و یا سیستم‌های جمع‌آوری داده به صورت آنلاین، نمونه ای از ده‌ها برنامه جدیدی هستند که حریم شخصی افراد را تهدید می‌کنند.

مشکل اصلی از آنجا ناشی می‌شود که چگونه می‌توان هم حریم خصوصی افراد را در نظر گرفت و هم از نتایج مفید سیستم‌های داده کاوی بهره برد. برای برطرف کردن موانع موجود در این زمینه، تحقیقات زیادی در حال انجام است [۲] .

روش‌هایی که برای حفظ محرمانگی داده‌ها در داده کاوی معرفی شده‌اند، هر کدام دارای ویژگی‌ها، مزایا و معایب خاصی نظیر از دست دادن داده‌ها هستند که ممکن است در کیفیت دانش استخراج شده از داده‌ها اثر گذارد تا جایی که داده‌ها غیر قابل استفاده شوند. در نتیجه لازم است تا از داده‌های محرمانه تا حدی که سودمندی آن ها حفظ گردد، محافظت نماییم.

در این تحقیق با کمک توابع عضویت فازی روشی را ارائه می‌کنیم تا محرمانگی داده‌ها، سودمندی آن ها و همچنین صحت نتایج داده کاوی حفظ شود.

۲- مفاهیم اولیه

۲-۱- خوشه بندی

خوشه‌بندی با یافتن یک ساختار درون یک مجموعه از داده‌های بدون برچسب درگیر است. خوشه به مجموعه‌ای از داده‌ها گفته می‌شود که به هم شباهت داشته باشند. در خوشه‌بندی سعی می‌شود تا داده‌ها به خوشه‌هایی تقسیم شوند که شباهت بین داده‌های درون هر خوشه حداکثر و شباهت بین داده‌های درون خوشه‌های متفاوت حداقل شود.

۲-۱-۱- روش های خوشه بندی

روش‌های خوشه‌بندی را می‌توان از چندین جنبه تقسیم‌بندی کرد:

۱. خوشه‌بندی انحصاری[۵] و خوشه‌بندی با هم پوشی [۶].

در روش خوشه‌بندی انحصاری پس از خوشه‌بندی هر داده دقیقاً به یک خوشه تعلق می‌گیرد. ولی در خوشه‌بندی با همپوشی پس از خوشه‌بندی به هر داده یک درجه تعلق بازاء هر خوشه نسبت داده می‌شود. به عبارتی یک داده می‌تواند با نسبتهای متفاوتی به چندین خوشه تعلق داشته باشد. نمونه‌ای از آن، خوشه‌بندی فازی است.

۲. خوشه‌بندی سلسله مراتبی[۷] و خوشه‌بندی مسطح[۸].

در روش خوشه بندی سلسله مراتبی، به خوشه‌های نهایی بر اساس میزان عمومیت آنها ساختاری سلسله مراتبی نسبت داده می‌شود. ولی در خوشه‌بندی مسطح تمامی خوشه‌های نهایی دارای یک میزان عمومیت هستند. با توجه با اینکه روش‌های خوشه‌بندی سلسله مراتبی اطلاعات بیشتر و دقیق‌تری تولید می‌کنند برای تحلیل داده‌های با جزئیات پیشنهاد می‌شوند ولی از طرفی چون پیچیدگی محاسباتی بالایی دارند برای مجموعه داده‌های بزرگ روش‌های خوشه‌بندی مسطح پیشنهاد می‌شوند. در بین الگوریتم‌های خوشه بندی انحصاری و مسطح در این تحقیق روش پیشنهادیمان را روی الگوریتم خوشه بندی پرکاربرد امید ریاضی بیشینه‌سازی[۹] بررسی کرده‌ایم.

۲-۱-۲- الگوریتم امید ریاضی بیشینه‌سازی

الگوریتم امید ریاضی بیشینه‌سازی، یک روش تکرارشونده است که به دنبال یافتن برآوردی با بیشترین درست نمایی برای پارامترهای یک توزیع پارامتری است. این الگوریتم روش متداول برای زمان‌هایی است که برخی از متغیرهای تصادفی پنهان هستند.

شرح الگوریتم:

فرض کنید که مشاهدات $x_1, x_2, \ldots, x_n$ را با d نمایش دهیم، متغیرهای پنهان $h_1, h_2, \ldots, h_n$ را با h و همه ی پارامترهای توزیع را نیز با θ.

در این صورت لگاریتم درست نمایی کل داده‌ها (پنهان و نمایان : مشاهدات) برابر خواهد بود با:

$$\pounds(\theta) = \log p(d;\theta) = \log \sum_{h} p(d,h;\theta) \qquad (۱)$$

از آنجا که لگاریتم تابع اکیداً صعودی است، می‌توان لگاریتم درست نمایی کل داده‌ها را نسبت به θ بیشینه کرد. البته آرگومان لگاریتم یک مجموع است و نمی‌توان به سادگی پاسخ تحلیلی برای θ یافت. از این رو، الگوریتم ترفندی را برای بیشینه کردن حد پایین لگاریتم درست نمایی بکار می‌برد. این حد پایین از نابرابری جنسن بدست می‌آید. بر اساس نابرابری جنسن که از کوژ بودن تابع لگاریتم استفاده می‌کند برای هر دسته k تایی از t_i ها و w_i ها اگر $t_i > 0$ و $\sum w_i = 1$ ، خواهیم داشت:

$$\sum_{i=1}^{k} w_i \log t_i \le \log \sum_{i=1}^{k} w_i t_i \qquad (۲)$$

اکنون $\pounds$ را به صورت زیر باز می‌نویسیم

$$\pounds(\theta) = \log \sum_{h} q(h) \frac{p(d,h;\theta)}{q(h)} \ge \sum_{h} q(h) \log \frac{p(d,h;\theta)}{q(h)} = \tau(q,\theta) \qquad (۳)$$

با گزینش $q(h) = p(h;d,\theta)$ نابرابری بالا تنگ می‌شود. این به معنای آن است که نابرابری به برابری تبدیل می‌شود. این گام الگوریتم همانند بیشینه کردن حدپایین درست نمایی (τ) نسبت به q است. در نتیجه روش کار الگوریتم امید ریاضی بیشینه کردن به صورت زیر است:

۱. پارامترها را با مقدار آغازین $\theta^{(0)}$ می‌دهیم.

۲. تا زمان همگرایی به بیشینه محلی ادامه می‌دهیم:

۱. امید ریاضی

$$q^t = \arg\max_{q} \tau(q,\theta^{(t)}) \qquad (۴)$$

۲. بیشینه کردن

$$\theta^{(t+1)} = \arg\max_{\theta} \tau(q^{(t)}, \theta) \qquad (۵)$$

۳. مقادیر نهایی θ و q را باز گردان.

بدین ترتیب در هر گام الگوریتم، حد پایین درست نمایی کل داده‌ها افزایش می‌یابد تا آنجا که در یک بیشینه محلی همگرا شود. برای رهایی از بیشینه-های محلی، این الگوریتم را معمولاً چندین بار با شرایط آغازین متفاوت اجرا می‌کنند.

۲-۲- منطق فازی

اولین بار در پی تنظیم نظریه‌ی مجموعه‌های فازی به وسیله‌ی پروفسور لطفی زاده [۳] در صحنه ی محاسبات نو ظاهر شد. منطق فازی از منطق ارزش‌های "صفر و یک" نرم‌افزارهای کلاسیک فراتر رفته و درگاهی جدید برای دنیای علوم نرم‌افزاری می‌گشاید، زیرا فضای شناور و نامحدود بین اعداد صفر و یک را نیز در منطق و استدلال‌های خود به کار می‌گیرد.

۲-۲-۱- تابع عضویت فازی

یک مجموعه فازی با تابع عضویت مشخص می‌شود. تابع عضویت، مجموعه مقادیر جهانی X را به عددی در بازه $[۰,۱]$ می‌نگارد.

در این تحقیق از سه تابع عضویت پی ، S شکل و Z شکل استفاده کرده ایم که در ادامه روابط آن ها را مشاهده می‌کنیم.

تابع عضویت پی:

$$F(x;a,b,c,d) = \begin{cases} 0, x \le a \\ a \le x \le \frac{a+b}{2} \to 2\left(\frac{x-b}{b-a}\right)^2 \\ \frac{a+b}{2} \le x \le b \to 1-2\left(\frac{x-b}{b-a}\right)^2 \\ b \le x \le c \to 1 \\ c \le x \le \frac{c+d}{2} \to 1-2\left(\frac{x-c}{d-c}\right)^2 \\ \frac{c+d}{2} \le x \le d \to 2\left(\frac{x-d}{d-c}\right)^2 \\ x \ge d \to 0 \end{cases} \qquad (۶)$$

تابع عضویت Z شکل:

$$F(x;a,b) = \begin{cases} x \le a \to 1 \\ a \le x \le \frac{a+b}{2} \to 1-2\left(\frac{x-a}{b-a}\right)^2 \\ \frac{a+b}{2} \le x \le b \to 2\left(\frac{x-b}{b-a}\right)^2 \\ x \ge b \to 0 \end{cases} \qquad (۷)$$

تابع عضویت S شکل:

$$F(x;a,b) = \begin{cases} x \le a \to 0 \\ a \le x \le \frac{a+b}{2} \to 2\left(\frac{x-b}{b-a}\right)^2 \\ \frac{a+b}{2} \le x \le b \to 1-2\left(\frac{x-b}{b-a}\right)^2 \\ x \ge b \to 1 \end{cases} \qquad (۸)$$

۳- مروری بر کارهای انجام شده

داده کاوی با حفظ حریم خصوصی، اولین بار در سال ۲۰۰۰ با انتشار دو مقاله با این عنوان در [۴, ۵] معرفی شد .این دو مقاله به دو مسئله اصلی زیر پرداختند:

۱. حفظ محرمانگی داده‌های جمع آوری شده.

۲. کاوش مجموعه‌ای از داده‌هایی که بین چند کاربر به اشتراک گذاشته شده است.

گروه اول یک پروتکل رمزنگاری ساخت درخت‌های تصمیم برای داده‌هایی که به اشتراک گذاشته شده‌اند، ابداع کردند و گروه دوم یک الگوریتم رندم سازی ارائه داده‌اند که به کاربران زیادی اجازه می‌دهد داده‌های خود را برای داده کاوی به اشتراک گذارند درحالی که داده‌های محرمانه شان محفوظ مانده و افشا نشود.

از آن سال تاکنون، روش ها و فنون داده کاوی با حفظ حریم خصوصی به اهداف مختلفی همچون پنهان سازی داده‌ها، آشفته‌سازی داده‌ها[۶]، پنهان سازی دانش، داده کاوی با حفظ حریم خصوصی توزیع شده و اشتراک دانش آگاه از حریم خصوصی برای وظایف داده کاوی مختلف ارائه شده‌اند.

دو دسته روش اصلی داده کاوی با حفظ حریم خصوصی عبارتند از پریشانی تصادفی[۱۰] [۴,۷] و محاسبات چند جانبه امن[۱۱] [۵].

روش پریشانی تصادفی مبنی بر افزایش یا ضرب خدشه تصادفی در مقادیر داده‌ها است. این روش جهت پنهان‌سازی داده‌ها مفید است. اگرچه در این روش‌ها می‌توان توزیع داده‌های اصلی را به خوبی از روی داده‌های تغییر یافته بازسازی کرد؛ اما در طی این تبدیل فواصل بین رکوردهای داده‌های اصلی به خوبی حفظ نمی‌شود. این مسئله منجر به کاهش صحت اجرای الگوریتم‌های داده کاوی بر اساس فاصله بر روی این داده‌ها می‌شود. علاوه بر این، این روش‌ها باعث کاهش داده نیز نمی‌شوند.

روش محاسبات چند جانبه امن، تاکنون برای محدوده گسترده‌ای از الگوریتم-های داده کاوی در محیط‌های توزیع شده، جایی که داده‌ها به صورت افقی یا عمودی بین چند طرف ارتباطی توزیع شده‌اند، به کار رفته است [۸,۹]. اما اغلب این روش‌ها بر روی فنونی تمرکز دارند که تنها برای الگوریتم‌های داده کاوی خیلی خاص مفید بوده و روند آنها به گونه‌ای است که استفاده از آنها نیازمند تغییر الگوریتم‌های داده کاوی است و اغلب امکان تعمیم آنها برای سایر الگوریتم‌های داده کاوی وجود ندارد. به علاوه این روشها دارای هزینه محاسباتی و ارتباطاتی بالایی بوده و تنها برای محیط‌های توزیع شده مناسب هستند.

علاوه بر دو دسته قبلی، فنونی نیز جهت حفظ حریم خصوصی برای الگوریتم-های داده کاوی از جمله خوشه بندی و طبقه بندی ارائه شده است.

یکی از روش‌های اصلی ارائه شده برای حفظ حریم خصوصی الگوریتم‌های داده کاوی بر اساس فاصله اقلیدسی در [۱۴] ارائه شده است. این روش داده

را بر اساس روش‌های مختلف تبدیل داده هندسی[۱۲] ، تغییر می‌دهد. روش‌های مختلف تبدیل داده هندسی شامل انتقال، مقیاس پذیری و چرخش فاصله اقلیدسی بین نقاط داده را به خوبی حفظ می‌کند؛ اما از آنجایی که تبدیل برای همه رکوردها به صورت یکسان انجام می‌شود، اگر طرف سوم بتواند مقادیر اصلی تنها یک سطر را بازیابی کند، کلیه مقادیر داده‌های اصلی بازسازی خواهد شد. در ضمن این روش نیز باعث کاهش ابعاد نمی‌شود.

روش دیگر ارائه شده حفظ حریم خصوصی الگوریتم‌های داده کاوی بر اساس فاصله اقلیدسی، افکنش تصادفی[۱۳] است [۱۵]. بر خلاف روش‌های قبلی، در روش افکنش تصادفی، تعداد ابعاد از m به k کاهش می‌یابد. آزمایش ها نشان داده فاصله اقلیدسی، زمانی که k کوچک باشد، تا حد زیادی خراب می‌شود. علاوه بر این برای داده‌های با تعداد رکورد زیاد، زمان اجرای این روش طولانی است.

روش دیگری برای حفظ حریم خصوصی الگوریتم‌های داده کاوی بر اساس فاصله اقلیدسی ارائه شده که بر اساس استفاده از تبدیل فوریه است [۱۶].

یکی دیگر از روش‌ها تبدیل بر پایه ی دوران[۱۴] است[۱۰]. این روش یک روش تبدیل مکانی برای حفظ حریم خصوصی داده‌ها در خوشه بندی است. با استفاده از این تبدیل علاوه بر این که حریم خصوصی داده‌ها حفظ می‌شود، نتایج معتبری از خوشه بندی دریافت خواهیم کرد. در این روش داده‌ها بعد از تبدیل شدن نیازی به نرمال سازی ندارند.

در سال ۲۰۰۳ یکسری روش‌های تبدیل هندسی معرفی شدند[۱۱]. این روش‌ها تضمین می‌دادند که کاوش داده‌ها محرمانگی داده‌ها را تا حد مشخصی به خطر نمی‌اندازد. این متدها برای حفظ حریم خصوصی در خوشه بندی خصوصا روش‌های خوشه بندی سلسله مراتبی ارائه شده بود. در این روش داده‌های محرمانه‌ی عددی دستخوش تغییراتی می‌شدند ولی ویژگی-های عمومی داده برای خوشه بندی محفوظ می‌ماند.

در سال ۲۰۰۹ روش پاسخ رندم و تبدیل داده هندسی داده‌ها را ترکیب کرده و روشی به نام پاسخ رندم تبدیل داده هندسی[۱۲] ابداع کردند. این روش می-تواند حریم خصوصی داده‌های عددی را حفظ کند.

در سال ۲۰۱۱ از توابع فازی برای حفظ حریم خصوصی در داده کاوی استفاده شد که از داده‌های محرمانه محافظت می‌کند[۱۳]. ولی صحت این روش فقط روی الگوریتم خوشه بندی k-means اثبات شده و محدود به داده‌های عددی است.

در این تحقیق ما پس از بررسی توابع فازی، یکی از سه تابع عضویت پی، S شکل و Z شکل را روی داده‌های محرمانه اعمال نموده و نشان داده‌ایم که نتایج خوشه بندی با روش امید ریاضی بیشینه‌سازی معتبر خواهد بود.

٤- فن پیشنهادی

در این تحقیق روشی برای حفظ حریم خصوصی در داده کاوی ارائه می‌دهیم. در این روش علاوه بر این که حریم خصوصی داده‌ها حفظ می‌شود، نتایج معتبری از خوشه بندی دریافت خواهیم کرد. در ادامه روند کار را شرح می‌دهیم:

مرحله ۱. مجموعه داده‌ها را با روش امید ریاضی بیشینه‌سازی خوشه بندی می‌کنیم.

مرحله۲. داده‌های محرمانه غیر عددی را به مقادیر عددی می‌نگاریم.

مرحله۳. توابع فازی از جمله s-shaped ، z-shaped یا p-shaped را روی داده‌های محرمانه اعمال کرده تا این داده‌ها به طور دقیق قابل شناسایی نبوده و حریم خصوصی آنها حفظ شود.

مرحله ۳. مجموعه داده‌های تغییر داده شده را با همان روشی که در مرحله اول ذکر شد خوشه بندی می‌کنیم.

مرحله ۴. نتایج حاصل از خوشه بندی داده‌های تغییر داده شده و داده‌های اصلی را مقایسه می‌کنیم.

٥- ارزیابی فن پیشنهادی

در این بخش پس از معرفی معیارهای ارزیابی و مجموعه داده‌های استفاده شده، به بیان نتایج اجرای این فن بر روی چند مجموعه داده خواهیم پرداخت. برای ارزیابی روش ارائه شده، نتایج خوشه بندی داده‌های تغییر داده شده با روش پیشنهادی را با نتایج خوشه بندی داده‌های اصلی و نتایج خوشه بندی داده‌هایی که بخش محرمانه آنها را حذف کرده‌ایم، مقایسه می‌کنیم.

یک جنبه‌ی مهم روی الگوریتم‌ها و کاربردهای داده کاوی حفظ حریم خصوصی و ابزارهایی برای توسعه و ارزش یابی به منظور انتخاب معیار ارزش یابی مناسب می‌باشد، اما واقعیت این نیست که الگوریتم‌های داده کاوی حفظ حریم خصوصی که تحت شاخص‌های متنوعی قرار دارند، بهتر از دیگر الگوریتم‌ها باشند، معمولا، یک الگوریتم می‌تواند در زمینه کارایی عملی باشد و یا کمی بهتر از دیگر الگوریتم‌ها باشد. بسیار مهم است که برای کاربران مجموعه‌ای از معیارها فراهم شود تا آنها را قادر سازد تا مناسب ترین الگوریتم را برای حفظ حریم خصوصی داده ها انتخاب کنند.

قبل از ارائه معیارهای مختلف برای تعیین سطح حریم خصوصی باید دو جنبه را در مورد اطلاعات شخصی یاد آور شد.

(الف) اطلاعات شخصی و محرمانه‌ای که در مجموعه داده‌های ورودی قرار دارند.

(ب) اطلاعات محرمانه و حساسی که در نتایج داده کاوی کشف می‌شوند.

در این تحقیق به مورد اول(داده‌های محرمانه) پرداخته‌ایم.

برای ارزیابی کیفیت خوشه بندی از نرم افزار وکا[۱۵] استفاده کرده‌ایم. این نرم افزار مجموعه‌ای از الگوریتم‌های روز یادگیری ماشین و ابزارهای پیش پردازش داده‌ها می‌باشد.

٥-١- مجموعه داده‌ها

آزمایش‌ها بر روی چهار مجموعه داده واقعی که از UCI Machine Learning Repository گرفته شده، انجام شده است. نام‌ها و ویژگی هایشان در جدول(۱) مشاهده می‌شود.

جدول (۱) : ویژگی‌های مجموعه داده‌ها

مجموعه داده‌ها	تعداد رکوردها	تعداد ویژگی‌ها
ADULT	۴۸۸۴۲	۱۴
Bank Marketing	۴۵۲۱۱	۱۷
Breast Cancer	۵۶۹	۳۱
German Credit	۱۰۰۰	۲۰

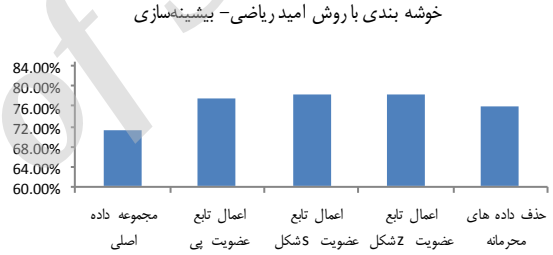
۵-۲- معیارهای ارزیابی

(۱) کیفیت خوشه بندی
(۲) محرمانگی داده‌ها
(۳) سودمندی داده‌ها

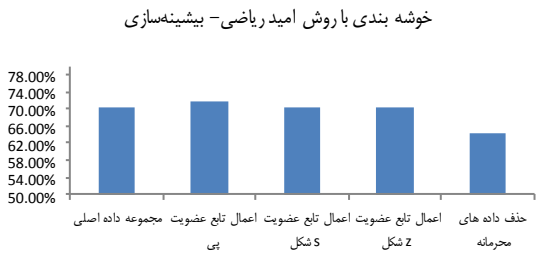
۵-۲-۱- کیفیت خوشه بندی

زمانی که از روش‌های خوشه بندی با حفظ حریم خصوصی استفاده می‌کنیم، خوشه بندی حاصل از مجموعه داده‌های تبدیل شده بایستی تا حد امکان با خوشه‌های حاصل از داده‌های اصلی یکسان باشند. برای ارزیابی کیفیت خوشه بندی از معیار صحت خوشه بندی استفاده کرده‌ایم. این معیار مشخص می‌کند چگونه و تا چه حد خوشه بندی داده‌های اصلی و تبدیل شده نزدیک و شبیه هستند.

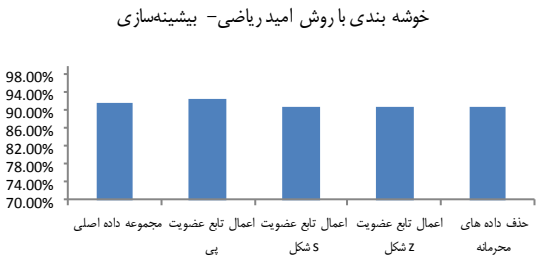
پس از اجرای فن پیشنهادی روی مجموعه داده‌های جدول (۱) به اجرای الگوریتم خوشه بندی امید ریاضی بیشینه‌سازی روی داده‌های اصلی، تبدیل شده با فن پیشنهادی و مجموعه‌ای از داده‌ها که بخش محرمانه آن را حذف کرده‌ایم، پرداختیم. اشکال ۱ تا ۴ نتایج بدست آمده را نشان می‌دهند.



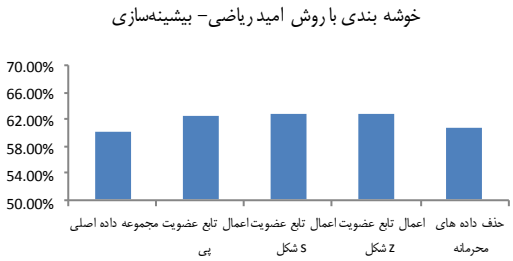
شکل (۱) : صحت خوشه بندی مجموعه داده‌ی Adult



شکل (۲) : صحت خوشه بندی مجموعه داده‌ی Bank Marketing



شکل (۳) : صحت خوشه بندی مجموعه داده‌ی Breast Cancer



شکل (٤) : صحت خوشه بندی مجموعه داده‌ی German Credit

۵-۲-۲- محرمانگی داده‌ها

به طور کلی، معیار کمی مورد استفاده برای اندازه گیری حفظ حریم خصوصی داده‌ها درجه ای از شک و عدم قطعیت است، که بر اساس آن می‌توان داده های خصوصی اصلی را حدس زد. در الگوریتم‌های مختلف با درجه‌های متفاوتی از عدم قطعیت روبرو هستیم.

در روش ارائه شده از آنجایی که بخش محرمانه‌ی داده‌ها تحت توابع فازی تغییر می‌یابند و بر اساس ماهیت توابع فازی می‌دانیم که حدس دقیق داده های محرمانه‌ی اصلی از روی داده‌های تغییر داده شده غیر ممکن است. در نتیجه محرمانگی داده‌ها در این روش به خوبی حفظ می‌شود.

۵-۲-۳- سودمندی داده‌ها

در نتیجه فرآیند حفظ محرمانگی داده‌ها باید سودمندی استفاده از آن‌ها نیز در نظر گرفته شود زیرا در روند پنهان سازی یا تحریف داده‌های محرمانه و شخصی مسلماً پایگاه داده‌ها دچار تغییراتی می‌شوند[۳].

به منظور پنهان سازی اطلاعات مهم، اطلاعات نادرست باید به پایگاه داده وارد شود یا مقادیر داده‌های محرمانه مسدود شود. اگرچه تکنیک‌های نمونه، اطلاعاتی که در پایگاه داده‌ها ذخیره شده‌اند را تغییر نمی‌دهند، اما چون این اطلاعات ناقص هستند، باز هم استفاده از داده‌ها را کاهش می‌دهند. هر چه تغییرات پایگاه داده‌ها بیشتر شود استفاده پایگاه داده‌ها کمتر می‌شود. در روش ارائه شده از آنجایی که داده‌های محرمانه مسدود نشده و اطلاعات نادرست به آن‌ها اضافه نمی‌شود، در نتیجه سودمندی داده‌ها به خوبی حفظ می‌شود.

٦- نتیجه گیری

پروسه استخراج دانش از داده‌ها در فرایند داده کاوی و انتشار داده‌ها، یا دانش حاصل از آنها، بدون افشاسازی بخش محرمانه، مسائل پیچیده و مهمی هستند. بنابر این نیاز به فنی است که علاوه بر حفظ حریم خصوصی داده‌ها، نتایج داده کاوی معتبری داشته باشد. در این تحقیق با بررسی انواع توابع فازی، از توابع S-shape، P-shape و Z-shape برای تغییر بخش محرمانه داده‌ها استفاده شد و همانطور که مشاهده شد این توابع داده‌ها را به گونه‌ای تغییر می‌دهند که صحت خوشه بندی مجموعه داده‌ها با روش امید ریاضی بیشینه‌سازی کاهش نیابد. و علاوه بر حفظ حریم خصوصی داده‌ها سودمندی آن ها نیز حفظ می‌شود. در کارهای آینده می‌توان راهکاری برای حفظ حریم خصوصی در سایر روش‌های داده کاوی ارائه داد.

زیرنویس‌ها

[۱] Association Rule
[۲] Classify
[۳] Cluster
[٤] Prediction
[٥] Exclusive or Hard Clustering
[٦] Overlapping or Soft Clustering
[۷] Hierarchical
[۸] Flat
[۹] Expectation Maximization
[۱۰] Random Perturbation
[۱۱] Secure Multiparty Computation
[۱۲] Geometric Data Transformation
[۱۳] Random Perturbation
[۱٤] Rotation-Based Transformation
[۱٥] Weka

مراجع

[۱] "حفظ حریم خصوصی در داده کاوی" صدیقه پیردومویی، انسیه ناظمی. همایش ملی مهندسی رایانه و مدیریت فناوری اطلاعات. ۱۳۹۳.

[۲] "پیش نمایشی از داده کاوی حفظ حریم خصوصی و بررسی چالش‌های آن". احمد قلیچی و احمد فراهی. اولین کنفرانس ملی نوآوری در مهندسی کامپیوتر و فناوری اطلاعات. ۱۳۹۲.

[۳] Zadeh L.A., ۱۹٦٥, *Fuzzy sets*, Information and Control.

[٤] Agrawal, R. and Srikant, R., *Privacy-Preserving Data Mining*, In Proc. Of the ۲۰۰۰ ACM SIGMOD Conf. on Management of Data, Dallas, pp. ۱۰۰-۱۰٥, ۲۰۰۲.

[٥] Lindell, Y. and Pinkas, *B., Privacy preserving data mining*, Proc. ۱۲th Ann. Int'I Conf. Advances in Cryptology, Springer-Verlag pp. ۳٦-٥٤ ۲۰۰۰.

[٦] Mohnish Patel, Prashant Richariya, Anurag Shrivastava , *A Review On Privacy Preserving Data Mining*, India, Scholars Journal of Engineering and Technology (SJET) Kim, J. J. Winkler, W. E., Multiplicative noise for masking continuous data, Statistical Research Division U.S., pp.۱۱۰-۱۱۸, ۲۰۰۳.

[۷] Kim, J. J. Winkler, W. E., "Multiplicative noise for masking continuous data" Statistical Research Division, U.S., pp.۱۱۰-۱۱۸, ۲۰۰۳.

[۸] Lin, X., Clifton, C. and Zhu, M., *Privacy preserving clustering with distributed EM mixture modeling*, Knowledge and Information Systems, Vol. ۸, pp.٦۸–۸۱, ۲۰۰٥.

[۹] Merugu, S. and Ghosh, J., *Privacy-preserving distributed clustering using generative models*, Proc. of the ۳rd International Conf. on Data Mining (ICDM'۰۳), pp.

[۱۰] S. R. M. Oliveira, and O. R.Zaiyane, *Achieving Privacy Preservation When Sharing Data for Clustering*, In Proceedings of the Workshop on Secure Data Management in a Connected World, in conjunction with VLDB"۲۰۰٤.Toronto, Ontario, Canada, pp. ٦۷–۸۲, ۲۰۰٤a.Object Management Group.

[۱۱] S.R. M. Oliveira, O.R. Zaiyane (۲۰۰۳), *Privacy Preserving Clustering by Data Transformation,* in proceedings of ۱۸ th Brazilian Conference on Databases.

[۱۲] Jie Liu,Yifeng XU,Harbin, *privacy preserving clustering by Random Response Method of Geometric Transformation*, In proceedings of Fourth international conference on internet computing for science and engineering ۲۰۰۹.

[۱۳] B. Karthikeyan, G. Manikandan, V.Vaithiyanathan, *A Fuzzy based approach for Privacy Preserving Clustering*, Journal of Theoretical and Applied Information Technology, ۳۱st October ۲۰۱۱. Vol. ۳۲ No.۲.

[۱٤] Oliveira, S. R. M. and Za¨iane, O. R., *Privacy preserving clustering by data transformation*, Proc. of the ۱۸th Brazilian Symposium on Databases, Manaus, Amazonas, Brazil., pp. ۳۰٤–۳۱۸, ۲۰۰۳.

[۱٥] Oliveira, S. and Zaiane, O.R., Privacy-preserving clustering by object similarity-based representation and dimensional reduction transformation, Proc. of Workshop on privacy and security aspects of data mining(PSDM'۰۴), pp. ۲۱–۳۰, ۲۰۰۴.

[۱٦] Mukherjee, S., Chen, Z. and Gangopadhyay, A., *A Preserving Technique for Euclidean Distance-Based Mining Algorithms Using Fourier-Related Transforms*, The VLDB Journal, Vol. ۱٥, pp. ۲۹۳–۳۱٥, ۲۰۰٦.