

تشخیص داده پرت با استفاده از ماشین یادگیری مغرط مبتنی بر فازی c-Means کوانتومی

سید امیر ادبی^۱ محبوبه هوشمند^۲ سید عابد حسینی^۳

۱- دانش آموخته کارشناسی ارشد- گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران
amir.cyber96@gmail.com

۲- استادیار- گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران
houshmand@mshdiau.ac.ir

۳- استادیار- گروه مهندسی برق، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران
hoseyni@mshdiau.ac.ir

چکیده: همواره یکی از مهم‌ترین دغدغه‌های داده‌کاوان در اختیار داشتن داده‌هایی صحیح و عاری از خطاست. داده‌هایی که اشتباهات انسانی در آن وجود نداشته باشد و رکوردهای آن تماماً پُر و حاوی داده‌هایی صحیح باشند. در این مقاله یک مدل یادگیری جدید مبتنی بر شبکه عصبی ماشین یادگیر مغرط برای تشخیص داده پرت پیشنهاد می‌شود. عملکرد شبکه‌های عصبی وابسته به پارامترهای مختلفی از قبیل ساختار آن، وزن‌های اولیه، تعداد نورون‌های لایه پنهان و نرخ یادگیری است. محاسبات کوانتومی یک روش جدید پردازش اطلاعات بر مبنای مکانیک کوانتومی است که امروزه مفاهیم آن در کاربردهای هوش مصنوعی نیز مورد استفاده قرار می‌گیرد. در روش پیشنهادی، شبکه عصبی ماشین یادگیری مغرط با استفاده از مفهوم خوشه‌بندی فازی c-Means کوانتومی، بهبود می‌یابد. این خوشه‌بندی به پیدا کردن وزن بهینه اتصالات لایه ورودی به لایه پنهان شبکه عصبی کمک می‌کند. همچنین باعث می‌شود، معماری شبکه در لایه پنهان به شکل سازنده‌ای شکل بگیرد و یادگیری بهبود یابد. عملکرد روش پیشنهادی از لحاظ صحت، نرخ صحیح مثبت و نرخ صحیح منفی نشان‌دهنده برتری روش پیشنهادی در تشخیص داده پرت نسبت به روش‌های دیگر است.

واژه‌های کلیدی: داده پرت، شبکه عصبی، فازی c-Means، محاسبات کوانتومی، ماشین یادگیری مغرط

نوع مقاله: پژوهشی

DOI: 10.52547/jiaeee.20.1.79

تاریخ ارسال مقاله: ۱۴۰۰/۰۷/۲۴

تاریخ پذیرش مشروط مقاله: ۱۴۰۱/۰۵/۰۲

تاریخ پذیرش مقاله: ۱۴۰۱/۰۸/۰۷

نام نویسنده‌ی مسئول: دکتر محبوبه هوشمند

نشانی نویسنده‌ی مسئول: ایران- مشهد- خیابان استاد یوسفی- دانشگاه آزاد اسلامی واحد مشهد- دانشکده‌ی مهندسی- گروه مهندسی کامپیوتر

۱- مقدمه

داده‌ها می‌کند. یکی از روش‌های محاسباتی نوین برای یادگیری ماشین شبکه‌های عصبی مصنوعی است. ایده اصلی این‌گونه شبکه‌ها تا حدودی الهام‌گرفته از شیوه کارکرد سیستم عصبی زیستی برای پردازش داده‌ها و اطلاعات به‌منظور یادگیری و ایجاد دانش است [۳]. عنصر کلیدی این ایده، ایجاد ساختارهایی جدید برای سامانه پردازش اطلاعات است.

ژو و همکارانش در سال ۲۰۲۰، الگوریتم نزدیک‌ترین همسایه را برای تشخیص داده پرت در داده جریانی اینترنت اشیا بکار بردند. در این مقاله، یک چارچوب جدید به نام GAAOD (کشف میانگین تقریبی مبتنی بر شبکه) را برای پشتیبانی از تشخیص داده‌های پرت مبتنی بر KNN بر روی داده‌های جریانی IoT پیشنهاد شده است [۶]. وانگ و همکارانش در سال ۲۰۱۹، یک مقاله مروری در رابطه با داده پرت ارائه کردند. تشخیص داده پرت مسئله مهمی است که در زمینه‌های مختلف تحقیقاتی و کاربردی مورد بررسی قرار گرفته است. محققان علم داده همواره به طراحی ارائه راه‌حل‌هایی محکم برای شناسایی داده پرت می‌پردازند. این مقاله پیشرفت فعلی روش‌های تشخیص دور را ارائه می‌دهد و درک بهتری از روش‌های مختلف تشخیص داده پرت را ارائه می‌دهد [۷].

کینگ و همکارانش در سال ۲۰۱۹، یک رویکرد رگرسیون مقاوم برای تشخیص داده پرت ارائه کردند. در این مقاله یک الگوریتم تشخیص داده پرت برای داده‌های سنسور دما از آزمایش موتور جت را پیشنهاد می‌شود. شناسایی مؤثر داده پرت باعث می‌شود تا مشکلات موتور به‌طور کارآمد بررسی و برطرف شود. تشخیص داده پرت در این داده‌ها چالش‌برانگیز است زیرا کنترل‌کننده انسانی در هر زمان میزان سرعت موتور را تعیین می‌کند [۸]. لیو و همکارانش در سال ۲۰۱۸ به تجزیه و تحلیل خوشه مشترک و مشکل تشخیص داده پرت پرداختند و الگوریتم خوشه‌بندی با حذف داده پرت (COR) را پیشنهاد دادند [۹]. یافتن توزیع مناسب برای داده‌های بُعد بالا بسیار دشوار است. برای حل مشکل داده‌های بُعد بالا، چند روش بدون مدل توسط جامعه‌ی مدیریت داده ارائه شده است. به عنوان مثال، از داده‌های پرت مبتنی بر فاصله و مبتنی بر چگالی. متأسفانه، اگرچه این تعاریف نیاز به فرضیه‌ای در مورد مجموعه داده ندارند هنوز هم نقص‌هایی دارند. شبکه‌های عصبی ذاتاً دارای پیچیدگی‌های زیادی می‌باشند و با معرفی شبکه‌های عصبی به‌عنوان طبقه‌بندی‌کننده غالباً مسئله مهم و حیاتی، چگونگی در نظر گرفتن ساختار شبکه، تعداد لایه‌ها و چگونگی تنظیم پارامترها، همیشه جزو چالش‌های اساسی این شبکه‌ها بوده است [۱۰]. می‌توان از روش‌های پس انتشار خطا که در واقع بر پایه گرادینان مرتبه اول هستند به‌عنوان روش‌های سنتی در شبکه‌های عصبی نام برد که این روش‌ها غالباً از مشکل ماندن در بهینه‌های محلی و همگرایی بسیار کند، رنج می‌برند. روش‌های زیادی به‌منظور رفع مشکلات این شبکه‌ها ارائه گردیده‌اند اما باز هم در هیچ‌کدام ضمانت کامل برای رسیدن به بهینه سراسری وجود ندارد. به‌منظور جلوگیری از محدودیت‌های

امروزه دیگر با مشکل کمبود داده و اطلاعات روبه‌رو نیستیم و به جای آن با مسئله داده‌ها و ارتباط درست، از میان حجم انبوهی از داده‌ها روبه‌رو هستیم. از آنجاکه درک روابط و ساختار حاکم بر داده‌ها و اطلاعات می‌تواند اطلاعات مهمی را ارائه نماید، نیازمند مکانیسمی هستیم تا به بهترین وجه عصاره‌ی این انبوه داده‌ها تهیه‌شده و استفاده شود [۱]. از این‌رو استفاده گسترده از داده‌کاوی بیش‌ازپیش در مرکز توجه است. از آنجایی‌که دستیابی به نتایج واقعی و مؤثر بدون برخورداری از ورودی‌های صحیح، قابل‌اعتماد و مؤثر ممکن نیست؛ پیش از هر تحلیلی بایستی از صحت و تناسب داده‌ها و اطلاعات موجود اطمینان داشته باشیم. سنگ‌بنای اولیه داده‌کاوی، داده است و از این سو یک عملیات داده‌کاوی مفید نیاز به داده‌های خوب و مناسب دارد. داده‌های مناسب جهت عملیات داده‌کاوی می‌تواند حاصل پیش‌پردازش درست و صحیح داده باشد، در واقع برای کشف دانش به کمک داده‌کاوی باید مقدماتی صورت گیرد که مجموعه‌ای از این مقدمات را آماده‌سازی داده می‌گویند. از اصلی‌ترین عملیات آماده‌سازی داده‌ها پاک‌سازی داده‌ها است که با هدف از بین بردن عدم قطعیت در داده‌کاوی مورد استفاده قرار گرفته و شامل پُر کردن داده‌های گم‌شده، هموار کردن نویزها، شناخت و حذف داده‌های پرت [۲] و رفع ناسازگاری‌هاست [۳].

در این مقاله ضمن معرفی داده‌های پرت و بررسی انواع روش‌های تشخیص داده پرت سعی داریم روشی بهبودیافته جهت تشخیص این داده‌ها و چگونگی از بین بردن آن‌ها را ارائه دهیم. شناسایی داده‌های پرت مسئله‌ی مهمی در کاوش داده است و توسط محققین زیادی بررسی شده است. داده پرت شی‌ای در مجموعه داده است که با دیگر اشیا در آن مجموعه داده فرق دارد. روش‌های کاوش داده‌های پرت را می‌توان در رشته‌های مختلف مانند شناسایی کلاه‌برداری از کارت اعتباری، ورود غیرمجاز به شبکه، نظارت بر محیط و غیره استفاده کرد [۴].

دو کار باید برای شناسایی داده پرت انجام شود. (۱) باید تعریف داده پرت در مجموعه مورد نظر مشخص شود. (۲) باید روشی مناسب برای شناسایی آن‌ها طراحی شود. مسئله‌ی داده پرت ابتدا در جامعه‌ی آماری بررسی شد. آن‌ها فرض می‌کنند که دیتاست دارای توزیع خاصی است و یک داده در صورتی داده پرت است که از این توزیع متفاوت باشد. به‌رحال، یافتن توزیع مناسب برای داده‌های بُعد بالا سخت است. یکی از رویکردهایی که امروزه بسیار مورد استفاده قرار می‌گیرد و می‌تواند نقش بسزایی در تشخیص داده پرت داشته باشد رویکردهای یادگیری ماشین است [۵]. هدف یادگیری ماشین این است که رایانه‌ها و سامانه‌ها بتوانند به تدریج و با افزایش داده‌ها کارایی بهتری در انجام وظیفه مورد نظر پیدا کنند. یادگیری ماشین کمک فراوانی به صرفه‌جویی در هزینه‌های عملیاتی و بهبود سرعت عمل تجزیه و تحلیل

معماری شبکه با اضافه کردن نورون‌ها در لایه پنهان به شکل سازنده‌ای شکل بگیرد و یادگیری بهبود یابد. در ادامه‌ی مقاله در بخش ۲، خلاصه‌ای از ماشین یادگیری مغرط و خوشه‌بند فازی c-Means تعاریف اولیه بررسی می‌شود و سپس در بخش ۳ روش پیشنهادی بررسی خواهد شد و در بخش ۴ نتایج آزمایش و در بخش ۵ نتیجه‌گیری نهایی آورده شده است.

۲- تعاریف اولیه

۲-۱- ماشین یادگیری مغرط (ELM)

ماشین یادگیری مغرط، به طور گسترده در الگوهای طبقه‌بندی در سال‌های اخیر استفاده شده است. این ابزار تعداد نورون‌های لایه مخفی شبکه را تنظیم می‌نماید و وزن‌های ورودی و بایاس‌های نورون‌های لایه پنهان را به صورت تصادفی انتخاب می‌نماید. وزن‌های لایه خروجی در این شبکه به صورت تحلیلی با روش حداقل مربعات محاسبه می‌گردد. کل این فرآیند بدون هیچ تکراری انجام شده و قابلیت تعمیم‌پذیری شبکه را ارتقا می‌بخشد [۱۶]. ELM یک شبکه تعمیم‌پذیر تک لایه پنهان روبه‌جلو (SLFNs) است. اساس ELM لایه‌های پنهان SLFNs است که نیاز به تنظیم نخواهد داشت [۱۷].

ساختار ELM در شکل ۱ نشان داده شده است. با توجه به N نمونه مجزای دلخواه $(x_j, t_j), j = 1, 2, \dots, N$ ،

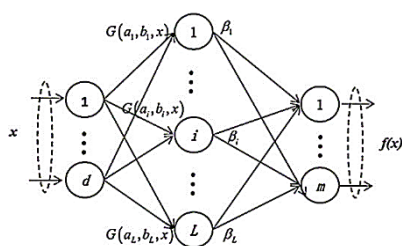
$$t_j = [t_{j1}, t_{j2}, \dots, t_{jm}]^T \text{ و } x_j = [x_{j1}, x_{j2}, \dots, x_{jd}]^T \in R^d$$

خروجی تابع SLFN با L گره پنهان در لایه خروجی را می‌توان توسط رابطه (۱) بیان کرد:

$$f(x_j) = \sum_{i=0}^L \beta_i G(a_i, b_i, x_j) = o_j \quad (1)$$

$$a_i \in R^d, b_i \in R, \beta_i \in R^m$$

که $a_i = [x_{i1}, x_{i2}, \dots, x_{id}]^T$ بردار وزن اتصال نورون پنهان i ام و نورون‌ها ورودی است، و b_i بایاس نورون پنهان i ام است، $\beta_i = [\beta_{i1}, \beta_{i2}, \dots, \beta_{im}]^T$ بردار وزن اتصال نورون پنهان i ام و نورون خروجی است. برای ماشین یادگیری مغرط، تمام معادلات بالا را می‌توان به صورت فشرده $H\beta = T$ نوشت. به طور کلی، ELM



شکل (۱): شبکه عصبی تک لایه پنهان رو به جلو

[۱۹]

ذکر شده که در بالا اشاره گردید برای اولین بار شخصی به نام هووانگ الگوریتم جدیدی را با نام ماشین یادگیری مغرط معرفی نمود. ماشین یادگیری مغرط یک مدل بسیار مناسب و ساده برای آموزش شبکه‌های عصبی تک لایه روبه‌جلو بوده که دارای سرعتی هزاران بار بیشتر از روش‌های مبتنی بر گرادیان سنتی می‌باشد و به طور گسترده در مسائل طبقه‌بندی و رگرسیون کاربرد دارد [۱۰]. اما با وجود همه مزیت‌هایی که در این شبکه کارآمد وجود دارد، اما این شبکه با چالش‌هایی هم روبه‌رو است. در ماشین یادگیری مغرط به دلیل تعیین تصادفی پارامترهای گره‌های پنهان ممکن است بعضی از پارامترهای غیر بهینه تولید شوند که در عملکرد و ثبات شبکه تأثیر گذراند. ماشین یادگیری مغرط برای به حداقل رساندن خطای آموزش از کل داده‌های آموزشی استفاده می‌کند که گاهی این موضوع مشکل بیش‌برازش را به همراه خواهد داشت [۱۱]. در ماشین یادگیری مغرط نورون‌های لایه پنهان به صورت تصادفی تعیین شده و به صورت ثابت در نظر گرفته می‌شوند که در مدل‌هایی با تعداد زیادی از نورون‌های پنهان ممکن است منجر به پیچیدگی بالای الگوریتم شود.

هدف در این مقاله ارائه رویکردی برای بهبود این شبکه عصبی است. الگوریتم خوشه‌بندی فازی c-Means (FCM) [۱۲] یک روش یادگیری بدون نظارت است. الگوریتم FCM می‌تواند توزیع نمونه‌های اصلی را یاد بگیرد و یک مجموعه نمونه جدید را با یک مقیاس کوچک‌تر اما با عملکرد بهتر شکل دهد به صورتی که توزیع نمونه‌های اصلی را تغییر نمی‌دهد [۱۳]. FCM همچنین می‌تواند تداخل ایجاد شده در داده‌های اصلی توسط داده‌های نویزی را به اندازه بیشتری کاهش دهد و در نتیجه طبقه‌بندی داده‌ها را بهبود می‌بخشد. با این حال الگوریتم خوشه‌بندی FCM اغلب به دام بهینه محلی گرفتار شده و از مشکل مقداردهی اولیه رنج می‌برد. بدین منظور از رویکرد کوانتوم برای بهبود این الگوریتم خوشه‌بندی بهره خواهیم برد تا علاوه بر بهبود این الگوریتم از آن برای مقداردهی اولیه ورودی شبکه عصبی ماشین یادگیری مغرط استفاده نماییم و به هدف نهایی که همان شناسایی داده‌های پرت و بدست آوردن یک طبقه‌بندی با دقت مناسب و سطح خوب از تفسیر است دست یابیم [۱۴].

این مقاله بر مسئله‌ی تشخیص داده پرت با استفاده از ماشین یادگیری مغرط مبتنی بر فازی c-Means کوانتومی تمرکز دارد. اهمیت این مطالعه عبارت است از: شناسایی داده‌های پرت مبتنی بر شبکه‌های عصبی، که مزیت‌های بالایی نسبت به تعاریف موجود در مجموعه داده‌های پیچیده دارد. در روش پیشنهادی، شبکه عصبی ماشین یادگیری مغرط با استفاده از مفهوم خوشه‌بندی فازی بهبودیافته با کوانتوم، که در آن پارامتر درجه عضویت با استفاده از مفهوم محاسبات کوانتومی تکامل یافته بهبود می‌یابد. مفهوم محاسبات کوانتومی یک فضای جستجوی بزرگ برای یک انتخاب درجه عضویت ایجاد می‌کند، که به پیدا کردن وزن بهینه کمک می‌کند و همچنین معماری شبکه را بهینه می‌کند [۱۵]. این مهم باعث می‌شود که

تنظیم شده برای حل مسائل یادگیری است [۱۸].

$$\min \beta \frac{1}{2} \|\beta\|^2 + \frac{C}{2} \|T - H\beta\|^2 \quad (2)$$

که در آن C ضریب جریمه خطاهای آموزش است. فرم بسته رابطه (۳) به صورت زیر است:

$$\beta = \left(H^T H + \frac{I}{C} \right)^{-1} H^T H \quad (3)$$

یا

$$\beta = H^T \left(H H^T + \frac{I}{C} \right)^{-1} T \quad (4)$$

که در آن I ماتریس واحد است. می توان از مزایای کلیدی الگوریتم ماشین یادگیری مفرد به موارد زیر اشاره نمود [۲۰]:

- سرعت بسیار بالا برای آموزش مدل
- پیچیدگی محاسباتی بسیار کم
- سادگی در ساختار و شبیه سازی راحت
- قدرت تعمیم دهی مناسب و بالا نسبت به روش های مبتنی بر گرادیان
- نداشتن مشکل ماندن در بهینه های محلی بر خلاف روش های مبتنی بر گرادیان
- قابلیت کار با انواع توابع فعال سازی

۲-۲- خوشه بندی فازی c-Means (FCM)

روش فازی c-Means یکی از معروف ترین روش های خوشه بندی فازی است [۲۱]. در این روش میزان عضویت یا تعلق هر شیء داده به هر خوشه در ماتریس عضویت $U = (u_1, u_2, \dots, u_n) = [u_{ij}]_{c \times c}$ که c تعداد خوشه ها و n تعداد اشیا داده است، مشخص می شود. در این روش دو محدودیت اصلی وضع شده است؛ اول اینکه هیچ خوشه ای نباید تهی باشد $(0 < \sum_{i=1}^n u_{ij} < n \forall i \in \{1, \dots, c\})$ و محدودیت دوم که محدودیت نرمال سازی نامیده می شود، بیان می کند که مجموع درجه عضویت هر داده به همه خوشه ها باید برابر یک باشد $(\sum_{i=1}^n u_{ij} = 1 \forall j \in \{1, \dots, n\})$.

$$M_{FCM} = \left\{ U \in [0, 1]^{n \times n} \mid \sum_{i=1}^c u_{ij} = 1, j = 1, \dots, n, \sum_{j=1}^n u_{ij} > 0, i = 1, \dots, c \right\} \quad (5)$$

روش FCM سعی می کند تا برای یک مجموعه نقاط داده، افزایشی $(c$ عدد خوشه به صورت فازی) بیابد که مجموع پراکندگی های تمام متغیرهای j در هر خوشه i را کمینه نماید. بر این اساس، تابع هدف FCM به صورت زیر نوشته می شود [۲۲]:

$$J_{FCM}(U, V, X) = \sum_{i=1}^c \sum_{j=1}^n d_{ij}^m (x_i - v_i)^2 \quad (6)$$

که در آن $d_{ij} = (x_j - v_i)$ فاصله بین داده x_j و مرکز خوشه i است و m میزان فازی بودن است (معمولاً $m=2$). پس اگر m به سمت یک $(m \rightarrow 1)$ میل کند، خوشه بندی قطعی یا سخت تر می شود و برعکس اگر m به سمت بی نهایت $(m \rightarrow \infty)$ میل کند، خوشه بندی فازی تر خواهد شد. البته تابع J_{FCM} را نمی توان به طور مستقیم کمینه نمود، بنابراین باید از الگوریتم های تکراری استفاده کرد. برای حل این مشکل از جایگزینی بهینه به صورت زیر استفاده می شود [۲۳]:

الف- یک عدد مثبت کوچک برای داده و مقادیر مناسبی برای m ، c را انتخاب (با استفاده از توابع روایی) و ماتریس V (مرکز یا میانه خوشه ها) را به صورت تصادفی تشکیل و مقدار t را نیز صفر قرار می دهیم.

ب- ماتریس عضویت را در $t = 0$ محاسبه یا در $t > 0$ بروز رسانی می نماییم. یعنی درجه عضویت برای پارامترهای ثابتی از خوشه ها به صورت زیر بهینه سازی می شوند.

(۷)

$$U_{ij}^{(t+1)} = \frac{\left(\frac{1}{|x_j - v_i|^2} \right)^{\frac{1}{m-1}}}{\sum_{l=1}^c \left(\frac{1}{|x_j - v_l|^2} \right)^{\frac{1}{m-1}}} = \frac{1}{\sum_{l=1}^c \left(\frac{|x_j - v_l|^2}{|x_j - v_i|^2} \right)^{\frac{2}{m-1}}} \quad l = 1, \dots, N$$

در این رابطه مشاهده می شود که درجه عضویت علاوه بر فاصله همان داده تا خوشه به فاصله بین این داده و خوشه های دیگر نیز بستگی خواهد داشت.

ج- در گام آخر، ماتریس مرکز خوشه ها را با درجه عضویت های بهینه شده، بروز رسانی می نماییم.

$$v_i^{t+1} = \frac{\sum_{j=1}^n (u_{ij}^{(t+1)})^m x_j}{\sum_{j=1}^n (u_{ij}^{(t+1)})^m} \quad (8)$$

فرمول بروز رسانی ماتریس مرکز خوشه ها به پارامترهایی نظیر مکان، شکل و اندازه خوشه ها وابسته است. علاوه بر پارامترهای مذکور، نحوه اندازه گیری فاصله بسیار مؤثر خواهد بود.

د- تکرار گام‌های ب و ج تا زمانی که $|U^{(t+1)} - U^t| < \varepsilon$ یا $|v^{(t+1)} - v^t| < \varepsilon$ برقرار باشند.

۳- روش پیشنهادی

۳-۱- الگوریتم خوشه‌بندی فازی c-Means کوانتومی

محاسبات کوانتومی یک روش جدید پردازش اطلاعات بر مبنای مکانیک کوانتومی است که امروزه مفاهیم آن در کاربردهای هوش مصنوعی نیز مورد استفاده قرار می‌گیرد [۲۴]. یک رجیستر کوانتومی (Q) از چندین کیوبیت تشکیل شده است.

$$Q = (q_1 | q_2 | \dots | q_l) \quad (9)$$

یک بیت در حالت کلاسیک در هر زمان فقط یک حالت ۰ یا ۱ را دارد درحالی‌که یک کیوبیت (کوچک‌ترین واحد اطلاعات در کامپیوترهای کوانتومی) قادر است به‌صورت برهم‌نهی خطی از ۰ و ۱ نمایش داده شود.

$$q_i = \alpha_i |0\rangle + \beta_i |1\rangle = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \quad (10)$$

که اعداد مختلط α و β به ترتیب نمایانگر احتمال قرار داشتن یک کیوبیت در حالت $|0\rangle$ و $|1\rangle$ هستند به‌نحوی که:

$$|\alpha_i|^2 + |\beta_i|^2 = 1 \quad (11)$$

گیت چرخشی کوانتومی به‌عنوان یک عملگر تبدیل در نظر گرفته می‌شود.

در معماری ماشین یادگیری مفراط یادگیری در لایه پنهان با استفاده از مفهوم الگوریتم فازی c-Means آغاز می‌شود. استفاده از FCM در یادگیری لایه پنهان و تعداد نورون‌ها، به یافتن مناسب تعداد خوشه‌ها (c) کمک می‌کند. بنابراین، مقادیر اولیه وزن لایه پنهان (W_h) به‌عنوان مراکز اولیه خوشه (V) در FCM برای یادگیری لایه پنهان استفاده می‌شوند. مراکز خوشه (V) با تغییر پارامتر فازی‌ساز (m) با استفاده از مفهوم محاسبات کوانتومی به‌صورت تکراری در FCM به‌روزرسانی می‌شوند. مراکز خوشه (V) بروز شده وزن‌های به‌روزرسانی لایه پنهان را نشان خواهد داد. پارامتر فازی‌ساز (m) نقش حیاتی در FCM را ایفا می‌کند، به‌طوری‌که درجه فازی بین خوشه‌ها (یا نورون‌ها) شکل گرفته را کنترل می‌کند. بنابراین برای مقداردهی اولیه پارامتر فازی‌ساز (m)، ما از مفهوم محاسبات کوانتومی استفاده می‌کنیم. این نشان‌دهنده فازی بودن پارامتر از نظر یک رجیستر

کوانتومی است که از چندین کیوبیت ساخته شده است. همان‌طور که قبلاً بحث شد، نمایش کیوبیت از نظر تنوع جمعیت نسبت به دیگر نمایش‌ها به دلیل برهم‌نهی خطی حالات در یک فضای جستجو به‌طور احتمالی ویژگی بهتری دارد. الگوریتم FCM برای چندین نسل با تحول پارامتر فازی در هر نسل (g) با استفاده از گیت چرخشی کوانتومی اجرا می‌شود. همچنین بهره‌برداری مناسب برای انتخاب مقدار بهینه m را فراهم می‌کند. پارامتر فازی‌ساز (m)، از نظر یک مکانیک کوانتومی برای هر نسل (g) به‌صورت M^g ، نشان داده می‌شود، که به‌صورت زیر تعریف می‌شود:

$$M^g = Q^g \quad (12)$$

که در آن Q^g از دو کیوبیت تشکیل شده است که به شرح زیر است:

$$Qg = (q_1^g | q_2^g) \quad (13)$$

در ماشین یادگیری مفراط کوانتومی، دو کیوبیت به اندازه کافی برای نشان دادن پارامتر فازی‌ساز (m) برای یافتن ارزش خود در دامنه $[2/5, 1/5]$ کافی هستند. این دو کیوبیت فضای جستجو را به چهار زیر بخش تقسیم می‌کنند و مقدار m را از این زیر بخش‌ها به‌طور احتمالی و نه به‌صورت پی‌درپی پیدا می‌کنند.

۳-۲- ماشین یادگیری مفراط (ELM)

پس از به دست آمدن مقادیر اولیه ورودی توسط الگوریتم خوشه‌بندی فازی c-Means کوانتومی فرآیند طبقه‌بندی ماشین یادگیری مفراط آغاز می‌شود، فرض کنید که n نمونه به همراه خروجی شان وجود دارد که به‌صورت (۱۴) نشان داده می‌شوند.

$$L = \{(x_i, y_i) : i = 1, 2, \dots, n\} \quad (14)$$

که هر داده x_i طبق رابطه (۱۵) تعریف می‌شود:

$$x_i = (x_i^1, x_i^2, \dots, x_i^D)^T \in R^D, (i = 1, 2, \dots, n) \quad (15)$$

که i شماره نمونه آموزشی و D بعد نمونه آموزشی می‌باشد و y_i خروجی داده x_i است که طبق رابطه (۱۶) تعریف می‌شود:

$$y_i = (y_i^1, y_i^2, \dots, y_i^C)^T \in R^C, (i = 1, 2, \dots, n) \quad (16)$$

که برای مسئله طبقه‌بندی خروجی طبق رابطه (۱۷) تعریف می‌شود:

و ماتریس وزن اتصالات بین لایه مخفی و خروجی به صورت (۲۴) تعریف می‌گردد:

$$\beta = \begin{bmatrix} \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \beta_{11} & \dots & \beta_{1c} \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ \beta_{11} & \dots & \beta_{1c} \end{bmatrix} \quad (24)$$

بنابراین خروجی نورون‌های لایه خروجی برای داده x طبق رابطه (۲۵) محاسبه می‌شوند:

$$f_i(x) = \sum_{j=1}^L \beta_{j,i} G(a_j, x), i = 1, \dots, C \quad (25)$$

با داشتن n نمونه و خروجی‌های متناظرشان، می‌توان رابطه (۲۶) را به فرم ماتریسی بازنویسی نمود:

$$\hat{Y} = \Phi \beta \quad (26)$$

که وزن‌های لایه مخفی و بایاس‌های نورون‌های لایه مخفی به صورت تصادفی توسط توزیع پیوسته‌ای تولید می‌شوند. با انجام این کار، وزن‌های بین لایه مخفی و خروجی توسط بهینه‌سازی خطای آموزش تعریف‌شده توسط تابع هدف رابطه (۲۷) به دست می‌آیند.

$$\min_{\beta} \|\Phi \beta - Y\| \quad (27)$$

راه‌حل تحلیلی رابطه (۲۷) به صورت (۲۸) است:

$$\beta = \left(\Phi^T \Phi + \frac{I}{C} \right)^{-1} \Phi^T Y \quad (28)$$

رابطه (۲۸) تنها زمانی مؤثر است که شرط $n \geq L$ برقرار باشد. هنگامی که $n < L$ فرمول هم‌ارز برای حل β با رابطه (۲۹) به دست می‌آید:

$$\beta = \Phi^T \left(\Phi \Phi^T + \frac{I}{C} \right)^{-1} Y \quad (29)$$

فلوچارت روش پیشنهادی در شکل ۲ آمده است.

$$y_i^j \in \{0,1\} (j = 1,2,\dots,C), \sum_{j=1}^C y_i^j = 1 \quad (17)$$

بنابراین، ماتریس خروجی نمونه‌ها به صورت (۱۸) تعریف می‌شود:

$$Y = \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} y_1^1 & \dots & y_1^c \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ y_n^1 & \dots & y_n^c \end{bmatrix} \quad (18)$$

خروجی لایه مخفی ELM به صورت (۱۹) است:

$$\Phi(x) = (\phi_1(x), \phi_2(x), \dots, \phi_l(x)) \quad (19)$$

به طوری که:

$$\phi_i(x) = G(a_i, b_i, x), (i = 1,2,\dots,L) \quad (20)$$

که L تعداد نورون‌های لایه مخفی را نشان می‌دهد و $\phi_i(x)$ خروجی i امین گره لایه مخفی می‌باشد و a_i بردار ستونی D بُعدی هستند که وزن‌های اتصال‌دهنده لایه ورودی به نورون مخفی i شبکه است و b_i یک عدد است که بایاس نورون i می‌باشد. G تابع فعال سازی است که می‌تواند به فرم‌های گوناگون نظیر تابع سیگموئید یا گوسین و ... باشد، تابع فعال‌سازی سیگموئید به صورت (۲۱) تعریف می‌شود:

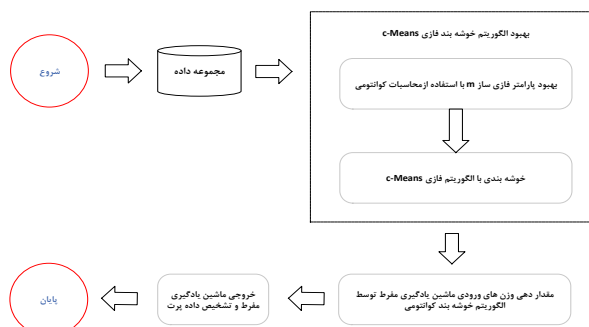
$$G(a, b, x) = \frac{1}{1 + \exp(-(a^T x + b))} \quad (21)$$

و تابع فعال‌سازی گوسین به صورت (۲۲) تعریف می‌شود:

$$G(a, b, x) = \exp(-b \|a - x\|^2) \quad (22)$$

ماتریس خروجی لایه پنهان، به ازای داده‌های بر چسب‌دار آموزشی به صورت (۲۳) تعریف می‌شود:

$$\Phi = \begin{bmatrix} \phi(x_1) \\ \cdot \\ \cdot \\ \cdot \\ \phi(x_n) \end{bmatrix} = \begin{bmatrix} \phi_1(x_1) & \dots & \phi_l(x_1) \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ \phi_1(x_n) & \dots & \phi_l(x_n) \end{bmatrix} \quad (23)$$



شکل (۲): فلوجارت روش پیشنهادی

۴-۲- معیار ارزیابی

به منظور ارزیابی روش پیشنهادی و مقایسه با سایر روش‌ها از سه معیار متداول استفاده خواهد شد که هر کدام به شرح زیر است:

- صحت، عبارت از میزان نزدیکی مقدار اندازه‌گیری شده به مقدار صحیح است؛ به عبارت دیگر از همخوانی یک مقدار آزمایشی یا میانگین چند سنجش با یک مقدار نظری (حقیقی) یک کمیت، صحت به دست می‌آید. دو مقدار TP و TN مهم‌ترین مقادیری هستند که در یک مسئله دودسته‌ای باید بیشینه شوند.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (30)$$

- معیار ویژگی به عنوان (نرخ صحیح منفی) در نظر گرفته می‌شود. معیار ویژگی توانایی شناسایی نمونه‌های طبیعی توسط مدل را نشان می‌دهد. زمانی که این معیار همانند حساسیت افزایش می‌یابد و اختلاف آن‌ها خیلی کمتر از (0.1) باشد آنگاه طبقه‌بندی کننده مناسب‌ترین خواهد بود. معادله این معیار به شرح زیر خواهد بود:

$$Specificity = \frac{TN}{TN + FP} \quad (31)$$

- معیار حساسیت به عنوان (نرخ صحیح مثبت) در نظر گرفته می‌شود. معیار حساسیت ارزیابی می‌کند که چقدر مدل توانایی شناسایی سوابق غیرطبیعی را دارا می‌باشد. حساسیت بالاتر یک طبقه‌بندی بهتر را نشان می‌دهد. این عبارت معمولاً از نسبت نمونه‌های صحیح طبقه‌بندی شده به کل نمونه‌ها حاصل می‌شود. میزان صحیح مثبت بیشتر یعنی میزان حساسیت بیشتر که منجر به دقت بیشتر هم خواهد شد. معادله حساسیت به شرح زیر خواهد بود:

$$Sensitivity = \frac{TP}{TP + FN} \quad (32)$$

۴-۳- ارزیابی نتایج

در این بخش به ارزیابی روش پیشنهادی به منظور به دست آوردن بالاترین میزان دقت و نرخ صحیح مثبت و منفی، برای مجموعه دادگان مورد نظر، خواهیم پرداخت. روش پیشنهادی توسط نرم‌افزار متلب ۲۰۱۴ بر روی سیستمی با پردازنده Intel Core i7-8550 4.0 GHz با ۱۶ گیگابایت رم و سیستم عامل ویندوز ۱۰ شبیه‌سازی شده و بر روی مجموعه دادگان توسط معیارهای ارزیابی مورد ارزیابی قرار گرفته است. نتایج به دست آمده از روش پیشنهادی بر روی مجموعه دادگان توسط معیارهای ارزیابی و مقایسه با سایر روش‌ها به شرح جدول ۲ و نمودارهای ۱ تا ۳ است.

۴- نتایج آزمایش‌ها

آنچه در این بخش بررسی خواهد شد شامل جزئیات مربوط به پیاده‌سازی روش پیشنهادی می‌باشد. دادگان استفاده شده در این بخش و نحوه جمع‌آوری آن‌ها ذکر خواهد شد و برای شفاف شدن موضوع داده‌های استفاده شده، در ادامه درج می‌شود.

۴-۱- مجموعه داده

برای ارزیابی عملکرد روش پیشنهادی از مجموعه داده‌های زیر (جدول ۱) در این پژوهش استفاده شده است که به شرح زیر است:

- مجموعه داده سرطان سینه از بیمارستان Srinagarind در تایلند به دست آمده است. اطلاعات این مجموعه داده از سال ۱۹۹۰ تا سال ۲۰۰۱ جمع‌آوری شده است. این مجموعه داده شامل ۲۴۶۲ نمونه و ۲۶ ویژگی است. در این مجموعه داده برخی صفات بیش از ۳۰٪ داده از دست رفته دارند، این در حالی است که بعضی صفات تنها یک مقدار دارند. با توجه به این موارد گفته شده مجموعه داده نهایی شامل ۷۳۶ نمونه و ۱۱ ویژگی و کلاس دودویی است.
- مجموعه داده shuttle که شامل ۹ ویژگی و ۵۸۰۰۰ نمونه است که همه آن‌ها عددی هستند. تقریباً ۸۰٪ از داده‌ها متعلق به کلاس ۱ است.
- مجموعه داده census که دارای ۴۰ ویژگی و ۲۹۹۲۸۵ نمونه است. این مجموعه داده شامل داده‌های سرشماری وزنی استخراج شده از بررسی‌های جمعیتی فعلی ۱۹۹۴ و ۱۹۹۵ که توسط اداره آمار ایالات متحده انجام شده است. داده‌ها حاوی ۴۰ متغیر وابسته به جمعیت‌شناسی و اشتغال هستند. وزن نمونه نشان‌دهنده تعداد افراد در جمعیت است که هر رکورد به دلیل نمونه‌برداری طبقه‌ای نشان داده می‌شود.

جدول (۱): ویژگی‌های دادگان

مجموعه داده	تعداد ویژگی	تعداد نمونه
سرطان سینه	۱۱	۷۳۶
shuttle	۹	۵۸۰۰۰
census	۴۰	۲۹۹۲۸۵

جدول (۲): مقایسه نتایج روش پیشنهادی با سایر روش‌ها بر روی دادگان

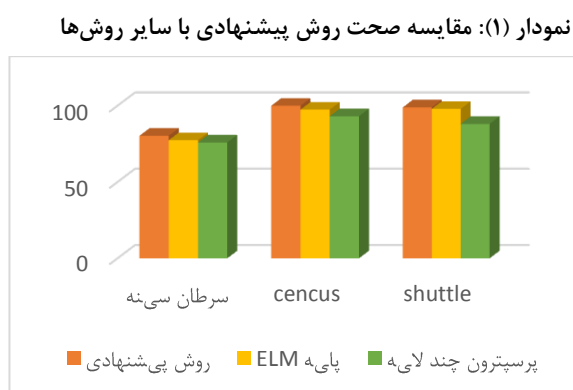
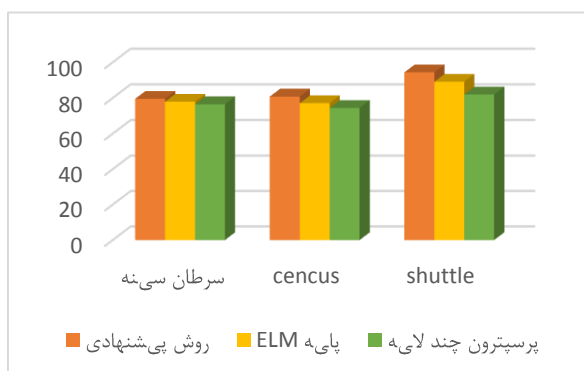
مجموعه داده	روش‌ها	صحت	نرخ صحیح مثبت	نرخ صحیح منفی
سرطان سینه	روش پیشنهادی	۷۹/۵۴	۸۰/۳۱	۷۷/۹۳
	ماشین یادگیری مفرط پایه [۲۵]	۷۷/۹۲	۷۷/۴۰	۷۶/۳۰
	شبکه پرسپترون چند لایه	۷۶/۴۷	۷۵/۹	۷۵/۱
census	روش پیشنهادی	۸۰/۷۹	۹۹/۹۶	۸۶/۲۹
	ماشین یادگیری مفرط پایه [۲۵]	۷۷/۰۶	۹۷/۴۲	۷۸/۱۲
	شبکه پرسپترون چند لایه	۷۴/۴	۹۳/۱	۷۳/۴
shuttle	روش پیشنهادی	۹۴/۳۶	۹۸/۹۰	۹۴/۴۳
	ماشین یادگیری مفرط پایه [۲۵]	۸۹/۰۹	۹۷/۸۷	۸۸/۸۴
	شبکه پرسپترون چند لایه	۸۱/۹۳	۸۸/۰۶	۸۰/۷۶

روش پیشنهادی در تشخیص داده پرت بر روی یکایک دادگان نویدبخش تحقق اصلی‌ترین عملیات آماده‌سازی داده‌ها یعنی پاک‌سازی داده‌هاست که با هدف از بین بردن عدم قطعیت در داده‌کاوای مورد استفاده قرار گرفته و شامل پُر کردن داده‌های گم‌شده، هموار کردن نویزها، شناخت و حذف داده‌های پرت و برطرف کردن ناسازگاری هاست. علت برتری روش پیشنهادی را می‌توان در موارد زیر یافت:

- مقداردهی مناسب وزن و پارامترهای شبکه به دلیل ساختار روش پیشنهادی
- استفاده از محاسبات کوانتومی جهت تشخیص داده‌های پرت

۵- نتیجه‌گیری

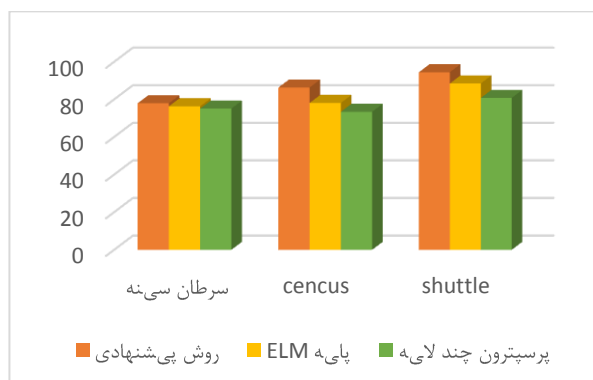
در این مقاله مسئله‌ی شناسایی داده پرت را بر روی مجموعه‌ای از دادگان بررسی کردیم و تعریف جدیدی از داده پرت را ارائه دادیم که مزیت‌های بیشتری از تعاریف موجود برای مجموعه داده‌های پیچیده دارد. برای شناسایی داده‌های پرت راهکاری با استفاده از یک شبکه عصبی کوانتومی مبتنی بر منطق فازی ارائه شد. بدین‌صورت که



الگوریتم خوشه‌بندی فازی c-Means با استفاده از محاسبات کوانتومی بهبود یافته و از خروجی آن برای مقداردهی اولیه پارامترهای ماشین یادگیری مفرط استفاده شد. در نهایت روش پیشنهادی بر روی سه مجموعه داده مورد ارزیابی قرار گرفت. نتایج آزمایش‌ها برتری روش پیشنهادی را از نظر صحت، نرخ صحیح مثبت، نرخ صحیح منفی و کیفیت راه‌حل‌های به‌دست‌آمده تأیید می‌نماید. در این پژوهش تلاش شد تا گامی در جهت بهبود نتایج حاصل در تحقیقات داده‌کاوای در حوزه یادگیری ماشین، برداشته شود. در این حوزه یک تشخیص درست و صحیح توسط این روش‌ها می‌تواند تأثیر حیاتی در روند تشخیص و حذف داده پرت، توسط داده کاوان داشته باشد. با وجود نتایج دلگرم‌کننده، برخی از محدودیت‌ها باید در کارهای آینده رسیدگی شود.

اول از همه، با وجود همه مزیت‌هایی که در ماشین یادگیری مفرط وجود دارد این شبکه با چالش‌هایی نیز روبه‌رو است از این‌رو بهبود هرچه بیشتر این روش نیز می‌تواند در نتایج به‌دست‌آمده از روش پیشنهادی تأثیر بسزایی داشته باشد. دوم اینکه، مدل ماشین یادگیری مفرط از بیش برآزش رنج می‌برد. بیش‌برآزش به پدیده نامطلوبی در

نمودار (۳): مقایسه نرخ صحیح منفی روش پیشنهادی با سایر روش‌ها



در نتایج به‌دست‌آمده از محیط شبیه‌سازی، روش پیشنهادی توانسته است با اختلاف قابل‌ملاحظه‌ای بهبود داشته باشد. نتایج برتر

- sensor networks: A survey and research direction analysis." *Computer Networks* 180: 107376, 2020.
- [14] Shukur, Hanan, Subhi RM Zeebaree, Abdulaheem Jamil Ahmed, Rizgar R. Zebari, Omar Ahmed, Barea Shams Aldeen Tahir, and Mohammed AM Sadeeq. "A state of art survey for concurrent computation and clustering of parallel computing for distributed systems." *Journal of Applied Science and Technology Trends* 1, no. 4, 148-154, 2020.
- [15] Alghushairy, O., Alsini, R., Soule, T., & Ma, X. A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams. *Big Data and Cognitive Computing*, 5(1), 1, 2021.
- [16] MCA, D. Devakumari. "Outlier Detection using Clustering Algorithm a Survey." Vol.6, No.10, 2017.
- [17] Chowdhury, K. P. "Supervised Machine Learning and Heuristic Algorithms for Outlier Detection in Irregular Spatiotemporal Datasets." *Journal of Environmental Informatics* 33, no. 1, 2019.
- [18] Wang, Xite, Derong Shen, Mei Bai, Tiezheng Nie, Yue Kou, and Ge Yu. "Cluster-based outlier detection using unsupervised extreme learning machines." In *Proceedings of ELM-2015 Volume 1*, pp. 135-146. Springer, Cham, 2016.
- [19] Bai, Mei, Xite Wang, Junchang Xin, and Guoren Wang. "An efficient algorithm for distributed density-based outlier detection on big data." *Neurocomputing* 181: 19-28, 2016.
- [20] Laskov, Pavel, Patrick Düssel, Christin Schäfer, and Konrad Rieck. "Learning intrusion detection: supervised or unsupervised?" In *International Conference on Image Analysis and Processing*, pp. 50-57. Springer, Berlin, Heidelberg, 2005.
- [21] Wu, Chengmao, and Yan Chen. "Adaptive entropy weighted picture fuzzy clustering algorithm with spatial information for image segmentation." *Applied Soft Computing* 86: 105888, 2020.
- [22] Tang, Yiming, Fuji Ren, and Witold Pedrycz. "Fuzzy C-means clustering through SSIM and patch for image segmentation." *Applied Soft Computing* 87: 105928, 2020.
- [23] Borlea, Ioan-Daniel, Radu-Emil Precup, Alexandra-Bianca Borlea, and Daniel Iercan. "A unified form of fuzzy C-means and K-means algorithms and its partitional implementation." *Knowledge-Based Systems* 214: 106731, 2021.
- [24] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information*, Cambridge University Press, 2010.
- [25] Wang, Kuaini, Jinde Cao, and Huimin Pei. "Robust extreme learning machine in the presence of outliers by iterative reweighted algorithm." *Applied Mathematics and Computation* 377: 125186, 2020.
- آمار گفته می‌شود که در آن درجه آزادی مدل بسیار بیشتر از درجه آزادی واقعی انتخاب شده و در نتیجه اگرچه مدل روی داده استفاده شده برای یادگیری بسیار خوب نتیجه می‌دهد، اما بر روی داده جدید دارای خطای زیاد است. استفاده از ضریب رگولاریزیشن مناسب می‌تواند یکی از راه‌های مقابله با این مشکل باشد. دوم اینکه، اعمال روش پیشنهادی بر روی مسائل علمی-مهندسی و واقعی برای بررسی هر چه بیشتر روش ارائه شده و همچنین بررسی تعمیم‌پذیری هر چه بیشتر روش ارائه شده در حل سایر مسائل می‌تواند چالش‌های پیش رو را نمایان و برطرف سازد.

مراجع

- [1] [1] Chua, S. J., Wrigley, S., Hair, C., & Sahathevan, R. Prediction of delirium using data mining: A systematic review. *Journal of Clinical Neuroscience*, 91, 288-298, 2021.
- [۲] احمدی، علیرضا، مجیدیان، سینا و کهالی، محمدحسین، تکمیل ماتریس گراف در حضور داده پرت، نشریه مهندسی برق و الکترونیک ایران، جلد ۲۰، شماره ۱، ۱۴۰۱.
- [3] Xu, F., Li, Z., Yue, J., & Qu, S. A Systematic Review of Educational Data Mining. *Intelligent Computing*, 764-780, 2021.
- [4] Smiti, A. A critical overview of outlier detection methods. *Computer Science Review*, 38, 100306, 2020.
- [5] Yuan, Z., Chen, H., Li, T., Liu, J., & Wang, S. Fuzzy information entropy-based adaptive approach for hybrid feature outlier detection. *Fuzzy Sets and Systems*, 2020.
- [6] Zhu, Rui, Xiaoling Ji, Danyang Yu, Zhiyuan Tan, Liang Zhao, Jiajia Li, and Xiufeng Xia. "Knn-based approximate outlier detection algorithm over iot streaming data." *IEEE Access* 8: 42749-42759, 2020.
- [7] Wang, Hongzhi, Mohamed Jaward Bah, and Mohamed Hammad. "Progress in outlier detection techniques: A survey." *IEEE Access* 7: 107964-108000, 2019.
- [8] Hullait, Harjit, David S. Leslie, Nicos G. Pavlidis, Steve King. "Robust Functional Regression for Outlier Detection." In *International Workshop on Advanced Analysis and Learning on Temporal Data*, pp. 3-13. Springer, Cham, 2019.
- [9] Liu, Hongfu, Jun Li, Yue Wu, and Yun Fu. "Clustering with Outlier Removal." *arXiv preprint arXiv:1801.01899*, 2018.
- [10] Alowayr, A. D., Alsalooli, L. A., Alshahrani, A. M., & Akaichi, J. A Review of Trajectory Data Mining Applications. In *2021 International Conference of Women in Data Science at Taif University (WiDSTaif)* (pp. 1-6). IEEE, 2021, March.
- [11]. Khalaf, Alaa, Jasim Mohammed Dahr, Ihab Ahmed Najim, Mohammed BM Kamel, Ali Salah Hashim, Wid Aqeel Awadh, and Aqeel Majeed Humadi. "Supervised Learning Algorithms in Educational Data Mining: A Systematic Review." *Southeast Europe Journal of Soft Computing* 10, no. 1, 55-70, 2021.
- [۱۲] پارسا، پویا، صفابخش، رضا، روش جدید تقطیع تصویر بر مبنای خوشه‌بندی فازی مبتنی بر تکامل تفاضلی چندهدفه، نشریه مهندسی برق و الکترونیک ایران، جلد ۱۳، شماره ۲، ۱۳۹۵.
- [13] Shahraki, Amin, Amir Taherkordi, Øystein Haugen, and Frank Eliassen. "Clustering objectives in wireless