ارائه روش انتخاب ویژگی مبتنی بر خوشه‌بندی در مسئله تشخیص هرزنامه

وحید نصرتی[*1]
دانشجوی دکتری، مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه اراک، اراک، ایران
محسن رحمانی
دانشیار، مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه اراک، اراک، ایران

مدیریت اطلاعات
دوره ۸، شماره ۱
بهار و تابستان ۱۴۰۱

چکیده: یکی از راه‌های تشخیص هرزنامه، دسته‌بندی ایمیل‌ها به دو دسته هرزنامه و غیرهرزنامه است. کارایی بالای روش‌های یادگیری ماشین در مسائل گوناگون، باعث توسعه وسیع آنها در دسته‌بندی متون شده است. استفاده از یک سازوکار کاهش ویژگی کارآمد در الگوریتم‌های یادگیری ماشین مبتنی بر محتوا به‌منظور استخراج یک بردار ویژگی کارآمد از میان تعداد بسیار زیادی ایمیل نقش مهمی دارد. برخلاف روش‌های پیشین که فقط ویژگی‌های برتر را انتخاب کرده و باقی ویژگی‌ها را نادیده می‌گیرند، در روش پیشنهادی در این مقاله سعی شده است از ویژگی‌های انتخاب‌نشده نیز استفاده شود. روش کار به این صورت است که ابتدا یک انتخاب ویژگی اولیه اعمال شده و تعدادی ویژگی انتخاب می‌شود. سپس، ویژگی‌های انتخاب‌نشده خوشه‌بندی شده و هر خوشه به یک ویژگی جدید نگاشت می‌شود و بردار ویژگی نهایی شامل ویژگی‌های انتخاب‌شده و ویژگی‌های نگاشت‌شده از هر خوشه خواهد بود. در پژوهش حاضر، با اعمال دو روش انتخاب ویژگی اولیه و همچنین دو تابع نگاشت ویژگی‌های خوشه، در مجموع، چهار روش ارائه شد و نتایج با استفاده از دو پایگاه داده PU2 و PU3 تجزیه و تحلیل شدند. نتایج حاصل از تجزیه و تحلیل انجام‌شده نشان داد که روش مبتنی بر انتخاب ویژگی اولیه DF و تابع نگاشت پیشرفته، در بین کلیه روش‌های پیشنهادی، دارای بالاترین کارایی است. همچنین، روش‌های پیشنهادی در مقایسه با انتخاب ویژگی اولیه (بدون خوشه‌بندی) دارای کارایی بهتری هستند.

کلیدواژه‌ها: انتخاب ویژگی، ایمیل، خوشه‌بندی، دسته‌بندی، کاهش ویژگی، هرزنامه.

---

۱. نویسنده مسئول: vh_nosraty@yahoo.com

مقدمه

سرویس ایمیل به بخشی جدایی‌ناپذیر از زندگی امروزه تبدیل شده است که به‌صورت گسترده و با اهداف شخصی و شغلی استفاده می‌شود. با وجود استفاده روزافزون از سایر رسانه‌های ارتباطی مبتنی بر بستر اینترنت همچون شبکه‌های اجتماعی، ایمیل‌ها همچنان در ارتباطات اجتماعی، دانشگاهی و تجاری پیشرو بوده و همچنان به‌عنوان پیش‌نیاز سایر رسانه‌های ارتباطی و تراکنش‌های الکترونیکی استفاده می‌شوند (Ghaleb, Mohamad, Fadzli & Ghanem, 2022). در سال‌های اخیر مشاهده می‌شود که اکثریت وب‌سایت‌ها از مشتریان تقاضای ثبت شناسه ایمیل را دارند که موجب قرار گرفتن ایمیل افراد در معرض هرزنامه نویس‌هایی[1] می‌شود که می‌توانند با ارسال هرزنامه به آنها حمله کنند (Soneji, Soman, Vyas & Puthran, 2022). هرزنامه‌ها یکی از پدیده‌های آزاردهنده اینترنتی به شمار می‌روند که توانسته‌اند شرکت‌های بزرگ جهانی از جمله ای‌اوال[2]، گوگل، یاهو و ماکروسافت را نیز به چالش بکشانند (Al-Rawashdeh, Mamat & Abd Rahim, 2019). به‌موازات پژوهش و توسعه روش‌های شناسایی ایمیل‌های ناخواسته، هرزنامه‌نویس‌ها نیز از سازوکارهای جدیدی برای دور زدن فیلترها استفاده می‌کنند. این جدال همواره ادامه داشته و توقفی در توسعه روش‌های تشخیص هرزنامه وجود ندارد. بنابراین، ضرورت وجود مدل‌های کارآمد و سازگار به‌منظور تشخیص هرزنامه، همواره احساس می‌شود.

برای غلبه بر این مشکل، پژوهشگران برای دسته‌بندی ایمیل‌ها به‌عنوان مشروع یا هرزنامه، روش‌های مختلفی ارائه کرده‌اند. به‌منظور تشخیص پیام‌های مخرب در پژوهش‌های مختلف، از پنج روش مختلف استفاده شده است (Dada, Bassi, Chiroma, Adetunmbi & Ajibuwa, 2019) که از میان آنها پژوهشگران از روش‌های مبتنی بر محتوا به‌طور گسترده‌تری استفاده کرده‌اند. روش‌های مبتنی بر محتوا از تجزیه و تحلیل فراوانی[3] و توزیع[4] کلمات و عبارات در محتوای ایمیل‌ها و همچنین ایجاد قوانین فیلترینگ خودکار برای دسته‌بندی ایمیل‌های دریافتی با استفاده از روش‌های یادگیری ماشین استفاده می‌کنند. به‌طور کلی، تمامی این روش‌ها، بر اساس دو مفهوم پایه‌ای شامل یک روش یادگیری ماشین و یک روش مهندسی دانش[5] عمل می‌کنند (Gibson, Issac, Zhang & Jacob, 2020). روش‌های مهندسی دانش به فیلترینگ مبتنی بر قانون شهرت دارند که در آنها مجموعه‌ای از قوانین به‌منظور کشف هرزنامه ساخته می‌شوند و ضروری است که این قوانین به‌طور مداوم به‌روزرسانی شوند تا بتوانند با تهدیدهای جدید مقابله کنند. در مقابل، روش‌های یادگیری ماشین که به فیلترینگ مبتنی بر یادگیری شهرت دارند، در مقایسه با روش‌های مبتنی بر قانون، کارایی بهتری دارند که موجب شده تاکنون برای مسئله تشخیص هرزنامه به‌خصوص بر اساس محتوا، روش‌های یادگیری ماشین بسیاری ارائه شود. در روش‌های یادگیری ماشین مبتنی بر محتوا به‌طور کلی محتوای ایمیل‌ها به مجموعه‌ای از کلمات (ویژگی‌ها) تبدیل شده و

1. Spammers
2. AOL
3. Frequency
4. Distribution
5. Knowledge engineering

معمولاً تعداد تکرار هر ویژگی به‌عنوان مقدار آن کلمه در نظر گرفته می‌شود. در مسئله تشخیص هرزنامه با تعداد زیادی از ویژگی روبه‌رو هستیم که اعمال نشدن یک انتخاب ویژگی کارآمد می‌تواند به کاهش کیفیت دسته‌بند نهایی منجر شود. از سوی دیگر، استخراج/ انتخاب ویژگی با به‌کارگیری روش‌های هوش مصنوعی به‌منظور کاهش داده‌های اضافی و تکراری و ایجاد نتایج با دقت بالا و رضایت‌بخش روزبه‌روز اهمیت بیشتری پیدا می‌کند (Ravi Kumar, Murthuja, Anjan Babu & Nagamani, 2022).

رویکردهای استخراج/ انتخاب ویژگی بخشی از فرایند کاهش ابعاد برای انتخاب مناسب‌ترین فضای ویژگی به شمار می‌روند که به‌عنوان یک مفهوم اصلی برای افزایش اثربخشی و کارایی مدل یادگیری ماشین، پیچیدگی مدل‌ها را کاهش داده و به تسریع روند آموزش مدل کمک می‌کنند. در واقع، ویژگی‌های کاهش‌یافته به کاهش مشکل بیش‌برازش و بهبود دقت کلی مدل یادگیری ماشین کمک می‌کنند و هزینه محاسباتی و زمان آموزش را کاهش می‌دهند. با ایجاد ویژگی‌های جدید از نسخه اصلی، تعداد ویژگی‌های یک مجموعه داده کاهش می‌یابد، بنابراین، مجموعه اصلی از ویژگی‌ها تا حد زیادی با مجموعه کاهش یافته جدید ویژگی‌ها خلاصه می‌شود (Rao, Verma & Bhatia, 2021).

برخی روش‌های مرسوم مختلف انتخاب ویژگی شامل سازوکارهای فیلتر[1]، رپر[2] و جاسازی‌شده[3] است که هر یک دارای مزایا و معایب مربوط به خود هستند. پژوهش‌های گذشته نشان داده‌اند که استفاده از تحلیل خوشه‌ای مؤثرتر از الگوریتم‌های سنتی انتخاب ویژگی است (Song, Ni & Wang, 2011). برخی پژوهش‌ها از جمله پژوهش دیلون، ملا و کومار[4] (۲۰۰۳) از خوشه‌بندی کلمات برای کاهش ابعاد داده‌های متنی استفاده کردند. اگرچه پژوهشگران در زمینه کاربرد انتخاب ویژگی مبتنی خوشه‌بندی در سایر حیطه‌ها پژوهش‌هایی انجام داده‌اند، اما تاکنون پژوهشگران اندکی کاربرد آن را در تشخیص هرزنامه بررسی کرده‌اند. برای نمونه، پژوهشگران یک سیستم فیلتر هرزنامه آنلاین بر اساس ویژگی‌های مختلف مانند لایک[5]، پخش مجدد[6]، هش‌تگ[7]، فالوورها[8] و آدرس‌های موجود در پست‌های شبکه اجتماعی فیس‌بوک ارائه دادند که از سه الگوریتم خوشه‌بندی و همچنین از Bayes ساده و درخت تصمیم برای تشخیص هرزنامه از غیرهرزنامه استفاده کرده است (Sohrabi & Karimi, 2015). در پژوهشی دیگر نیز به‌منظور انتخاب و شناسایی کلمات مهم کلیدی متون در دسته‌بندی پروفایل از خوشه‌بندی متنی استفاده شده است (Elhussein & Brahimi, 2021). از این رو، در پژوهش حاضر، چند سازوکار کاهش ویژگی مبتنی بر خوشه‌بندی ویژگی‌ها در مسئله تشخیص هرزنامه ارائه می‌شود. نوآوری پژوهش حاضر را می‌توان به‌صورت زیر برشمرد:

1. Filter
2. Wrapper
3. Embedded
4. Dhillon, Mallela & Kumar
5. Like
6. Replay
7. Hash tag
8. Followers

- روش خوشه‌بندی پیشنهادی برخلاف روش‌های قبلی به‌صورت متقارن بـه خوشـه‌بنـدی تمـام ویژگی‌ها نمی‌پردازد، بلکه ابتدا ویژگی‌ها را به دو مجموعه با ارزش بالا و با ارزش پایین تقسـیم کرده و فقط عملیات خوشه‌بندی را بین ویژگی‌های با ارزش پایین انجام می‌دهد.
- در روش‌های خوشه‌بندی پیشین به‌طور کلی در هر خوشه بـه انتخـاب ویژگـی‌هـای بااهمیـت پرداخته می‌شود و سایر ویژگی‌ها نادیده گرفته می‌شوند. روش بـه‌کاررفتـه در پـژوهش حاضـر مبتنی بر انتقال فضای بردار ویژگی اولیه به یک فضای جدید است کـه بـا بهـره‌گیـری از یـک سازوکار نگاشت، فضای بردار ویژگی هر خوشه به یک ویژگی جدید نگاشت مـی‌شـود. در روش پیشنهادی، دو سازوکار نگاشت ساده و پیشرفته ارائه و بررسی خواهد شد.
- به‌منظور تقسیم ویژگی‌های اولیه به دو مجموعه، از دو رویکرد شامل سازوکار مبتنـی بـر روش تکرار سند (DF)[1] و سازوکار مبتنی بر روش BPIL[2] اسـتفاده شـده و کـارایی آنهـا بـا یکـدیگر مقایسه می‌شود.

بنابراین، در پژوهش حاضر بر اساس دو سازوکار انتخاب ویژگی اولیه و دو سازوکار نگاشت، در مجموع چهار رویکرد ارائه شده است که با استفاده از دسته‌بند بیزین به‌عنوان یک سازوکار پرکاربرد، به‌خصوص در زمینه متن‌کاوی و روی دو پایگاه داده مهم تشخیص هرزنامه بررسی، تجزیه و تحلیل خواهند شد.

در ادامه، مقاله بر این اساس تقسیم شده است. ابتدا مقـدمات روش پیشـنهادی شـامل فراینـد کلـی دسته‌بندی مبتنی بر یادگیری ایمیل‌ها بررسی شده و بیزین پایه به‌عنوان یک دسته‌بند پایه بیان می‌شود. سپس، معیار DF و الگوریتم PBIL که در روش پیشنهادی برای انتخاب ویژگـی اولیـه بـه‌کـار مـی‌رونـد، بررسی می‌شوند. در ادامه، روش کاهش ویژگی پیشنهادی مبتنی بر خوشه‌بندی شـرح داده مـی‌شـود. در پایان، به ارزشیابی و تحلیل روش پیشنهادی پرداخته می‌شود و پس از آن بحث، نتیجه‌گیری و پیشنهادها ارائه خواهند شد.

پیشینه پژوهش

کاهش ویژگی به‌طور کلی می‌تواند به دو طریق انتخاب ویژگی و اسـتخراج ویژگـی انجـام شـود. در روش نخست، مجموعه‌ای از ویژگی‌های برتر بر اساس یـک پـارامتر خـاص انتخـاب شـده و سـایر ویژگـی‌هـای باقی‌مانده نادیده گرفته می‌شوند که ممکن است به دور ریخته‌شدن ویژگی‌های حاوی اطلاعات مفید برای تمییز بین کلاس‌ها و در عمل کاهش کارایی دسته‌بند منجر شود. از طرف دیگر، هرزنامه‌نویس می‌تواند با کشف ویژگی‌های برتر و به‌کـار نگـرفتن آنهـا در ایمیـل‌هـای خـود، فیلترینـگ را دور بزنـد (DeBarr & Wechsler, 2012). در استخراج ویژگی به‌جای انتخاب مجموعه‌ای از ویژگی‌های برتر، سـعی مـی‌شـود بـا تغییر فضای بردار اولیه به یک فضای فشرده جدید بدون کم کردن تعداد ویژگی‌ها، عملیات کاهش ویژگی انجام شود. از آنجا که در این روش، هر ویژگی نهایی به‌دسـت‌آمـده تـابعی از ویژگـی‌هـای ورودی اسـت،

1. Document Frequency
2. Population Based Incremental Learning

هرزنامه‌نویس‌ها قادر به شناسایی این ویژگی‌ها نبوده و نمی‌تواند به‌راحتی با آن مقابله کنند. این کار باعث افزایش مقاومت[1] مدل نسبت به نویز شده و شناسایی سازوکار آن توسط هرزنامه نویسان را سخت‌تر می‌کند. با این حال، گاهی اوقات توصیف داده‌ها پس از نگاشت بردار ویژگی از بین رفته و هزینه انجام این فرایند نیز می‌تواند زیاد باشد (Eesa, Abdulazeez & Orman, 2017; Aziz, Verma & Srivastava, 2017). همچنین، زمانی که تعداد ویژگی‌های غیرمرتبط زیاد باشد، این روش دارای کارایی مناسبی نیست (Zebari, Abdulazeez, Zeebaree, Zebari & Saeed, 2020).

همان‌طور که پیش‌تر بیان شد، از روش‌های مبتنی بر یادگیری ماشین به‌طور گسترده‌ای در مسئله تشخیص هرزنامه استفاده شده است. به‌دلیل تعداد زیاد کلمات که به‌عنوان ویژگی در یک مدل دسته‌بندی در نظر گرفته می‌شوند، ضرورت دارد در سیستم‌های فیلترینگ هرزنامه سازوکار انتخاب ویژگی کارآمدی وجود داشته باشد. طیف وسیعی از پژوهشگران تاکنون، مدل‌های انتخاب ویژگی مبتنی بر یادگیری ماشین را ارائه داده‌اند که یکی از مواردی که شاید کمتر به آن توجه شده روش‌های مبتنی بر خوشه‌بندی است. مائو، هو، جیانگ، وی و شن[2] (۲۰۲۰) یک مدل انتخاب ویژگی مبتنی بر خوشه‌بندی و رتبه‌بندی به نام CBFS[3] برای کاهش ویژگی پیشنهاد دادند که در آن ابتدا، فاصله بین بردارهای ویژگی محاسبه شده و سپس این بردارها را در خوشه‌های مختلف ادغام می‌کند و مرکز هر خوشه را به‌عنوان یک بردار ویژگی انتخاب می‌کند. سپس، افزایش اطلاعات و نرخ بهره ویژگی‌ها را برای ساده‌سازی بیشتر تعداد ویژگی‌ها بر اساس تولید خوشه‌بندی ادغام کرده و در نهایت، دسته‌بند را به زیرمجموعه‌ای از ویژگی‌ها اعمال می‌کند تا جریان‌های ترافیک غیرعادی را شناسایی کند.

الحاران، فتلاوی و علی[4] (۲۰۱۹) در پژوهشی به ارائه یک رویکرد انتخاب ویژگی مبتنی بر خوشه‌بندی پرداختند که دارای سه مرحله است. مرحله نخست، شامل استخراج ویژگی از طریق ماتریس هم‌زمانی[5] سیاه و سفید (GLCM)[6]، الگوی باینری محلی (LBP)[7] و فیلتر گابور است. مرحله دوم، انتخاب ویژگی با استفاده از الگوریتم خوشه‌بندی K-means و بر اساس پنج معیار ارزیابی ویژگی است. در نهایت نیز چند الگوریتم دسته‌بندی برای ارزیابی عملکرد و دقت طبقه‌بندی پیشنهادی استفاده شد که حاکی از عملکرد رضایت‌بخش آن بود. وین و حجازی[8] (۲۰۲۱) نیز یک رویکرد انتخاب ویژگی نیمه‌نظارت‌شده مبتنی بر خوشه‌بندی ویژگی و حداکثر کردن حاشیه فرضیه[9] با هدف بهبود دقت دسته‌بندی ارائه دادند که رویکرد اصلی آن، مدیریت دو جنبه اصلی انتخاب ویژگی، یعنی ارتباط و افزونگی بود. روش آنها دارای سه مرحله بود که در مرحله نخست، وزن شباهت بین ویژگی‌ها با یک گراف پراکندگی[10] نشان داده

1. Robust
2. Mao, Hu, Jiang, Wei & Shen
3. Clustering-Based Feature Selection
4. Alharan, Fatlawi & Ali
5. Co-occurrence Matrix
6. Gray Level Co-occurrence Matrix
7. Local Binary Pattern
8. Ving & Hijazi
9. Hypothesis margin
10. Sparse graph

می‌شود که در آن، هر ویژگی می‌تواند از نگاشت خطی سایر ویژگی‌ها ساخته شود. در مرحله دوم، ویژگی‌ها به‌صورت سلسله‌مراتبی دسته‌بندی می‌شوند تا خوشه‌های مشابه را شناسایی کنند و در نهایت، یک تابع هدف مبتنی بر حاشیه نیمه‌نظارت‌شده به‌منظور انتخاب ویژگی دارای بیشترین تمایز بهینه‌سازی شده که باعث حداکثر شدن ارتباط و حداقل شدن افزونگی می‌شود.

روش‌های نام‌برده انتخاب ویژگی به‌طور کلی در هر خوشه به انتخاب ویژگی‌های برتر پرداخته و سایر ویژگی‌ها را نادیده می‌گیرند. دهقان و منصوری (۲۰۱۸) یک روش انتخاب ویژگی با استفاده خوشه‌بندی سلسه‌مراتبی[1] پایین به بالا[2] ارائه کردند که پس از خوشه‌بندی ویژگی‌ها، ویژگی‌های نهایی از بین خوشه‌های برتر انتخاب شده و سایر خوشه ها نادیده گرفته می‌شوند. هوانگ، ژانگ، ونگ و ژانگ[3] (۲۰۱۸) نیز یک روش انتخاب ویژگی SVM-RFE[4] مبتنی بر خوشه‌بندی ارائه دادند که ابتدا ویژگی‌ها را بر اساس ارزشمندی آنها مرتب کرده، سپس عملیات خوشه‌بندی را روی آنها انجام می‌دهد و در پایان در هر خوشه ویژگی‌های برتر انتخاب می‌شوند.

با توجه به مطالب گفته‌شده می‌توان بیان داشت که خوشه‌بندی می‌تواند به‌طور مؤثری برای بهبود فرایند انتخاب ویژگی به‌کار گرفته شود. در روش پیشنهادی تلاش می‌شود با بهره‌گیری از ویژگی‌های انتخاب‌نشده مرحله نخست، از انتخاب ویژگی و خوشه‌بندی آنها به‌منظور افزایش قدرت افتراق بردار ویژگی نهایی استفاده شود. به همین دلیل، برخلاف روش‌های مبتنی بر خوشه‌بندی بیان‌شده، تمام ویژگی‌های یک خوشه به یک ویژگی جدید نگاشت و به بردار ویژگی مرحله نخست اضافه می‌شوند. در واقع، در هر خوشه عملیات استخراج ویژگی وجود دارد و هیچ ویژگی‌ای حذف نخواهد شد.

مقدمات روش پیشنهادی

در این بخش، برخی مقدمات روش پیشنهادی شامل دسته‌بند بیزین، معیار DF و الگوریتم PBIL معرفی می‌شود.

دسته‌بند بیزین

روش‌های مبتنی بر یادگیری ماشین که عملیات دسته‌بندی ایمیل‌ها را بر اساس محتوا انجام می‌دهند، دارای سه مرحله کلی هستند. مرحله نخست، عملیات پیش‌پردازش[5] است که کلمات بی‌تأثیر حذف شده، لغات ریشه‌یابی شده و هر ایمیل به شکل مجموعه‌ای از کلمات (ویژگی) نمایش داده می‌شود. برای نمایش ویژگی روش‌های زیادی وجود دارد که در این پژوهش، از روش تکرار کلمه (TF)[6] برای نمایش داده استفاده شده که مقدار آن معادل «تعداد تکرار وقوع یک کلمه در یک پیام» است (Li, Xia, Zong &

1. Hierarchical
2. Bottom-up
3. Huang, Zhang, Wang & Zhang
4. Support Vector Machine Recursive Feature Elimination
5. Pre-processing
6. Term Frequency

Huang, 2009). پس از اینکه داده‌های ورودی به‌شکل مطلوب نمایش داده شدند، در مرحله بعد عملیات انتخاب/ کاهش ویژگی برای بهینه کردن ویژگی‌های انتخابی به‌کار می‌رود و در پایان عملیات، دسته‌بندی بر اساس این ویژگی‌ها انجام می‌شود.

تاکنون، روش‌های یادگیری ماشین بسیاری از جمله ماشین بردار پشتیبان (SVM)[1]، شبکه عصبی مصنوعی (ANN)[2] و k نزدیک‌ترین همسایه (KNN)[3] برای دسته‌بندی ایمیل‌ها به‌کار رفته است، اما بیزین به‌طور خاص در کاربردهای تجاری و متن‌باز[4] بسیار استفاده می‌شود (Metsis, Androutsopoulos & Paliouras, 2006) که دلیل آن می‌تواند سادگی آن باشد که موجب می‌شود بتوان به‌راحتی آن را پیاده‌سازی کرد. پیچیدگی اجرایی الگوریتم بیزین خطی بوده و دارای دقت به‌نسبت بالایی است که باعث رقابت‌پذیری آن با سایر الگوریتم‌های یادگیری ماشین می‌شود (Androutsopoulos, Koutsias, Chandrinos & Spyropoulos, 2004). به همین دلیل در این مقاله، از بیزین پایه به‌عنوان دسته‌بند استفاده می‌شود.

در تئوری بیزین احتمال اینکه یک پیام با بردار ویژگی $\vec{x} = \langle x_1, \ldots, x_m \rangle$ متعلق به کلاس c باشد به‌صورت زیر محاسبه می‌شود (Metsis et al, 2006):

$$p(c|\vec{x}) = \frac{p(c).p(\vec{x}|c)}{p(\vec{x})}$$ رابطهٔ ۱)

که $p(c)$ برابر است با احتمال پیشین[5] کلاس c، $p(\vec{x}|c)$ برابر است با احتمال اینکه $\vec{x}$ در کلاس c ظاهر شود و $p(\vec{x})$ هم به‌صورت زیر محاسبه می‌شود:

$$p(c_s).p(\vec{x}|c_s) + p(c_h).p(\vec{x}|c_h)$$ رابطهٔ ۲)

که در آن، c_s کلاس هرزنامه و c_h کلاس غیرهرزنامه محسوب می‌شود. دسته‌بند بیزین دارای پیاده‌سازی‌های مختلفی است که در این پژوهش از نسخه‌ای از بیزین با عنوان روش بیزین چندجمله‌ای با TF[6] (Nosrati & Rahmani, 2021) استفاده می‌شود.

معیار تکرار سند DF

معیار DF، یکی از معیارهای معروف و ساده انتخاب ویژگی در داده‌های متنی به شمار می‌رود. مقدار این معیار برای هر کلمه/ ویژگی i عبارت است از تعداد اسنادی که ویژگی i در آن وجود دارد. در این معیار فرض می‌شود ویژگی‌هایی که در اسنادی بیشتری تکرار شده باشند، اهمیت بیشتری دارند و آنها را برای عملیات انتخاب ویژگی در نظر می‌گیرد.

1. Support Vector Machine
2. Artificial Neural Network
3. K-nearest neighbor
4. Open-source
5. Priori probabilities
6. Multinomial NB, TF attributes

الگوریتم تخمین توزیع PBIL

الگوریتم PBIL که نخستین بار Baluja آن را معرفی کرد، یک الگوریتم از خانواده الگوریتم‌های تخمین توزیع (EDA)[1] است که در آن فرض می‌شود هیچ‌گونه وابستگی بین متغیرها وجود ندارد (Baluja & Caruana, 1995) و با نمایش باینری افراد کار می‌کند. این الگوریتم همانند الگوریتم ژنتیک، با ساخت یک نسل اولیه شروع به کار می‌کند و در هر دور از اجرای الگوریتم، تعدادی از نسل کنونی برای ساخت نسل بعدی انتخاب می‌شوند، اما برخلاف الگوریتم ژنتیک فاقد عملگرهای جهش و ادغام است. بعد از انتخاب افراد برای ساخت نسل بعد، عملیات تخمین توزیع روی آن محاسبه شده و بر اساس آن به ساخت فرزندان اقدام می‌شود. عملیات تخمین و توزیع در واقع جایگزین عملگرهای جهش و انتخاب در الگوریتم ژنتیک شده است. این موضوع باعث می‌شود که نسل والد در تولید نسل فرزند تأثیر مستقیم نداشته باشد، بلکه فقط توزیع آنها به ساخت فرزندان منجر می‌شود. شبه کد مربوط به الگوریتم PBIL در شکل ۱ نشان داده شده است.

```
******Initialize Probability Vector******
for i :=1 to LENGTH do P[i] = 0.5;
while (NOT termination condition)
                    *****Generate Samples*****
for i :=1 to NUMBER_SAMPLES do
solution_vectors[i] := generate_sample_vector_according_to_probabilities (P);
evaluations[i] :=Evaluate_Solution (solution_vectors[i]);
solution_vectors = sort_vectors_from_best_to_worst_according_to_evaluations():
                    ****Update Probability Vector towards best solutions****
for j :=1 to NUMBER_OF_VECTORS_TO_UPDATE_FROM
for i :=1 to LENGTH do P[i] := P[i] * (1.0 - LR) + solution_vectors[j][i]* (LR);
PBIL CONSTANTS:
NUMBER_SAMPLES: the number of vectors generated before update of the probability vector (200)
LR: the learning rate, how fast to exploit the search performed (0.005).
NUMER_OF_VECTORS_TO_UPDATE_FROM: the number of vectors in the current population which
are used to update the probability vector (2)
LENGTH: number of bits in the solution (determined by the problem encoding).
```

شکل ۱. شبه کد مربوط به الگوریتم PBIL (Baluja and Caruana, 1995)

از جمله کاربردهایی که در آن الگوریتم PBIL دارای کارایی مناسب است، می‌توان به بهینه‌سازی ترکیبی[2]، خوشه‌بندی افزایشی[3] و انتخاب ویژگی اشاره کرد (Hong, Kwong, Chang & Ren, 2008). در مقاله حاضر نیز از الگوریتم PBIL به‌منظور انتخاب ویژگی اولیه استفاده می‌شود.

1. Estimation of Distribution Algorithms
2. Combinatory optimization
3. Additive clustering

روشی پیشنهادی مبتنی بر خوشه‌بندی ویژگی‌ها

خوشه‌بندی روشی است که برای بهینه‌سازی در بسیاری از شاخه‌های علوم استفاده می‌شود. اساس کار آن به این صورت است که داده‌هایی که قرار است خوشه‌بندی شوند را به گروه‌های مختلف با عنوان خوشه تقسیم‌بندی می‌کند، به‌نحوی که داده‌های مشابه در خوشه‌های یکسان و داده‌های متفاوت تا حد امکان در خوشه‌های مختلف قرار می‌گیرند. عملیات خوشه‌بندی یک پایگاه داده را می‌توان در سطوح مختلف از جمله در سطح ویژگی‌ها (عمودی) یا نمونه‌های آموزشی (افقی) اجرا کرد (Vega-Pons & Ruiz-Shulcloper, 2011) که در پژوهش حاضر از خوشه‌بندی عمودی روی ویژگی استفاده می‌شود.

قبل از عملیات خوشه‌بندی بایستی مسائلی همچون معیار خوشه‌بندی، اندازه هر خوشه و تعداد ویژگی‌هایی که در هر خوشه قرار می‌گیرند، مشخص شوند. منظور از معیار خوشه‌بندی این است که آیا ویژگی‌هایی که در یک خوشه قرار می‌گیرند، ارتباطی با یکدیگر دارند و اگر دارند، بر اساس چه معیاری است. برای مثال، تشابه داشتن یا تشابه نداشتن در یک پارامتر خاص می‌توانند معیار خاصی برای خوشه‌بندی باشد. در روش پیشنهادی پس از خوشه‌بندی ویژگی‌ها، هر خوشه به یک ویژگی نگاشت می‌شود و مسئله مهمی که باید مشخص شود، چگونگی نگاشت ویژگی‌های یک خوشه برای به دست آوردن ویژگی معادل آن خوشه است. بر اساس شکل ۲، فرایند خوشه‌بندی پیشنهادی را می‌توان به‌طور کلی به سه مرحله تقسیم کرد که عبارت‌اند از:

۱. **مرحله نخست:** تقسیم ویژگی‌ها به دو مجموعه شامل آنهایی که قرار است خوشه چندتایی (ارزش پایین) باشند و آنهایی که قرار است خوشه‌بندی نشوند (ارزش بالا). نقطه‌ای که در آن بردار ویژگی شکسته می‌شود را نقطه شکست می‌نامیم.
۲. **مرحله دوم:** انجام عملیات خوشه‌بندی در بخشی که بایستی خوشه‌بندی چندتایی انجام شود.
۳. **مرحله سوم:** نگاشت ویژگی‌های هر خوشه برای به دست آوردن ویژگی معادل آن خوشه.

در ادامه، هر یک از این مراحل به تفصیل تشریح خواهند شد.

مرحله نخست: اعمال انتخاب ویژگی اولیه و تقسیم ویژگی‌ها

نخستین گام در خوشه‌بندی ویژگی‌ها بر اساس شکل ۲ تقسیم ویژگی‌ها است. همان‌طور که گفته شد، در این گام ویژگی‌ها به دو مجموعه با ارزش بالا و ویژگی‌های با ارزش پایین تقسیم می‌شوند که به‌ترتیب با رنگ تیره و رنگ روشن در شکل ۲ نشان داده شده‌اند. در مجموعه ویژگی‌های با ارزش بالا، هیچ‌گونه خوشه‌بندی انجام نمی‌گیرد و در واقع، عملیات خوشه‌بندی فقط در مجموعه ویژگی‌های با ارزش پایین انجام می‌شود. درصد ویژگی‌های اختصاص‌یافته به هر یک از این دو مجموعه باید قبل از اجرای الگوریتم مشخص باشد. در این پژوهش، برای تقسیم ویژگی دو استراتژی در نظر گرفته شده است که عبارت‌اند از:

۱. تقسیم مبتنی بر معیار DF: در این روش، درصد ویژگی‌های اختصاص‌یافته به هر یک از دو مجموعه به‌عنوان ورودی در نظر گرفته می‌شود که بایستی توسط کاربر تعیین شود. سپس، فرایند اختصاص ویژگی‌ها به هر یک از دو مجموعه آغاز می‌شود. در این روش، ابتدا ویژگی‌ها بر اساس معیار DF مرتب

می‌شوند و با توجه به درصد واردشده، مجموعه ویژگی‌ها به دو بخش تقسیم می‌شوند و بخش بالایی لیست مرتب‌شده با عنوان ویژگی‌های با ارزش بالا و بخش پایین به‌عنوان ویژگی‌های با ارزش پایین در نظر گرفته می‌شوند. روند انجام این کار به‌صورت زیر خواهد بود:

$$\text{کل ویژگی‌ها} = \begin{cases} \{\text{ویژگی‌های ابتدایی لیست مرتب شده بر اساس DF}\} = \text{ویژگی‌های با ارزش بالا} \\ \{\text{ویژگی‌های انتهایی لیست مرتب شده بر اساس DF}\} = \text{ویژگی‌های با ارزش پایین} \end{cases}$$

۲. تقسیم مبتنی بر معیار PBIL: در این روش برخلاف روش قبلی، درصد ویژگی‌های اختصاص‌یافته به هر یک از دو مجموعه توسط کاربر تعیین نمی‌شود، بلکه این مقدار توسط الگوریتم PBIL تعیین می‌شود. یعنی ابتدا انتخاب ویژگی توسط الگوریتم PBIL انجام می‌گیرد و ویژگی‌های انتخاب شده توسط آن به مجموعه ویژگی‌های با ارزش بالا و ویژگی‌های انتخاب‌نشده به مجموعه ویژگی‌های با ارزش پایین اختصاص می‌یابند. در این روش به‌منظور تقسیم ویژگی‌ها به‌صورت زیر عمل می‌شود:

$$\text{کل ویژگی‌ها} = \begin{cases} \{\text{ویژگی‌های انتخاب شده توسط PBIL}\} = \text{مجموعه ویژگی‌های با ارزش بالا} \\ \{\text{ویژگی‌های انتخاب نشده توسط PBIL}\} = \text{مجموعه ویژگی‌های با ارزش پایین} \end{cases}$$

مرحله دوم: عملیات خوشه‌بندی

همان‌طور که در شکل ۲ نشان داده شده است، عملیات خوشه‌بندی فقط روی ویژگی با ارزش پایین انجام می‌شود. تعداد خوشه‌ها در این روش به‌عنوان پارامتر ورودی در نظر گرفته شده است و اندازه هر خوشه نیز از تقسیم تعداد کل ویژگی‌ها بر تعداد خوشه‌ها به دست می‌آید. در پژوهش حاضر اندازه خوشه‌ها برابر ۲ در نظر گرفته شده است، به این معنا که هر خوشه شامل دو ویژگی خواهد بود. عملیات خوشه‌بندی نیز بر اساس معیار DF انجام می‌شود و ویژگی‌هایی که مقادیر تکرار سند در آنها به یکدیگر نزدیک‌تر باشد، در یک خوشه قرار می‌گیرند.

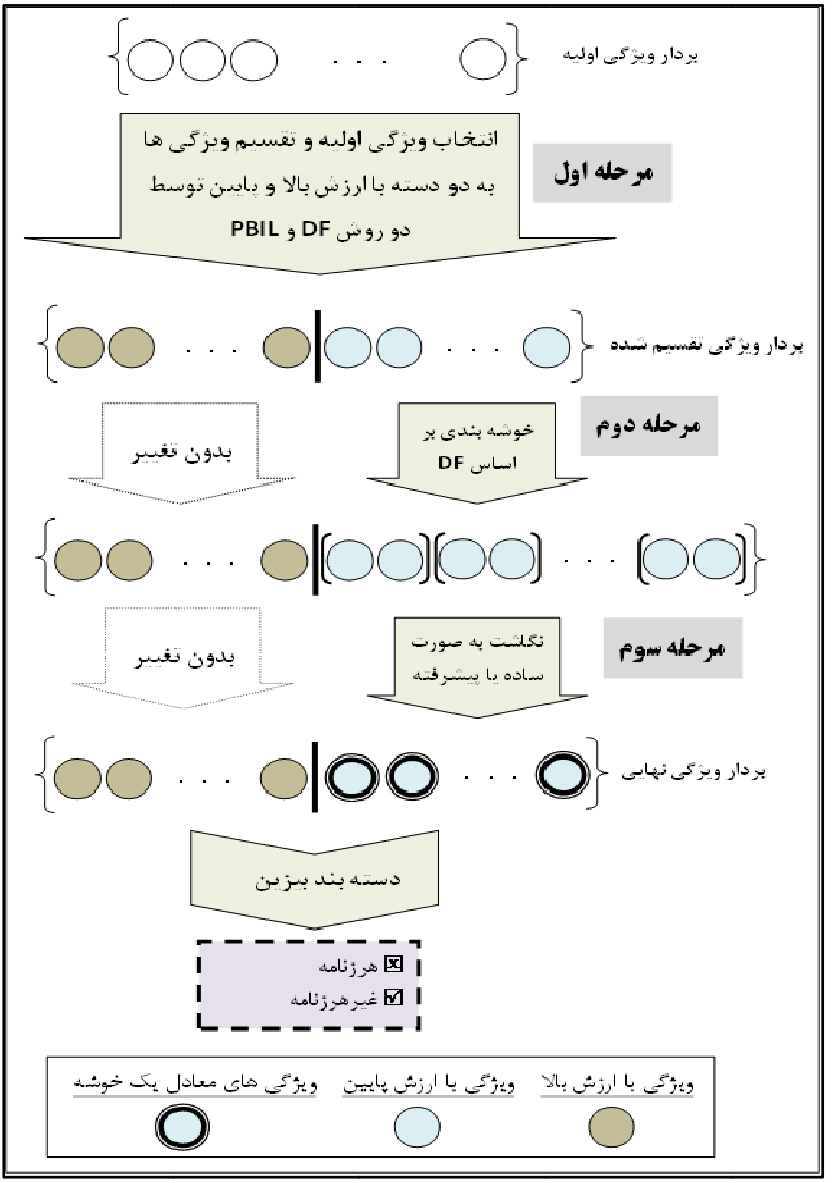
مرحله سوم: نگاشت ویژگی‌های هر خوشه

پس از خوشه‌بندی ویژگی‌ها، در مرحله بعد بایستی هر خوشه که دربرگیرنده مجموعه‌ای از ویژگی‌ها است به یک ویژگی نگاشت شود. در این پژوهش دو سازوکار نگاشت ساده و پیشرفته به‌کار رفته است. در ادامه، سازوکار هر یک از این دو روش بررسی می‌شود.

- نگاشت ساده

در این روش، ویژگی حاصل از یک خوشه از حاصل جمع تمامی مقادیر آن خوشه به دست می‌آید. فرض کنید یک خوشه شامل k ویژگی به‌صورت $(x_1, x_2, \ldots, x_k)$ باشد. تابع نگاشتی که برای این کار انتخاب شده به‌صورت رابطه ۳ است:

رابطه ۳)
$$S = \sum_{i=1}^{k} x_i$$



شکل ۲. فرایند انتخاب ویژگی پیشنهادی

همان‌طور که در رابطه ۳ مشاهده می‌شود، مقدار ویژگی حاصل از یک خوشه برابر با حاصل جمع مقادیر تمام ویژگی‌هایی است که در آن خوشه قرار دارند. با انجام این کار تمام ویژگی‌هایی که در یک خوشه قرار دارند از نظر دسته‌بند دارای ارزش یکسانی برای تمایز بین دسته‌ها هستند. حال این مسئله مطرح می‌شود که نگاشت ویژگی‌ها در یک خوشه و در نظر گرفتن آنها به‌عنوان یک ویژگی باعث کاهش قدرت تمایز کنندگی ویژگی‌ها شده و در نتیجه، دقت دسته‌بند کاهش یابد. در مقابل، قوت‌هایی وجود دارند که تأثیر آنها بیشتر از این مسئله است. این قوت‌ها به این صورت هستند که در بردار ویژگی که در فاز آموزش و بدون خوشه‌بندی به دست می‌آید، معمولاً ویژگی‌هایی از تمام پیام‌های نمونه آموزشی وجود دارد و زمانی که هم‌پوشانی این ویژگی‌ها در نمونه‌های آموزشی کم باشد، باعث می‌شود که تعداد کمی از ویژگی‌ها در هر پیام در فرایند ساخت بردار ویژگی آن پیام دخیل باشند. در نتیجه، زمانی که در فاز تست بردار ویژگی برای هر نمونه آموزش ساخته می‌شود، ممکن است مقدار بسیاری از ویژگی‌ها (به‌خصوص برای پیام‌های کوتاه) صفر باشد. اما در فرایند خوشه‌بندی امکان صفر بودن تمام مقادیر خوشه بسیار کم است. این کار باعث کاهش تعداد ویژگی‌هایی می‌شود که دسته‌بند استفاده می‌شود و هزینه محاسباتی و سربار زمانی نیز کاهش می‌یابد.

- نگاشت پیشرفته

برخلاف نگاشت ساده که در آن ویژگی‌ها با استفاده از یک معادله از پیش تعریف‌شده با یکدیگر ادغام می‌شوند و تمام ویژگی‌های یک خوشه از دیدگاه دسته‌بند نقش یکسانی دارند، در نگاشت پیشرفته برای ترکیب ویژگی‌ها از معادله‌ای استفاده می‌شود که بتوان تا حدی وابستگی بین ویژگی‌ها را در نظر گرفت، به‌طوری که بتوان الگوی بود یا نبود ویژگی‌های یک خوشه را در یک کلاس خاص اقتباس کرد. این الگوها می‌توانند شامل دو یا تعداد بیشتری ویژگی باشند که مقدر رخداد آنها در یک پیام می‌تواند به تعیین کلاس آن پیام کمک کند. در این بخش، الگوهای ما در ساده‌ترین حالت خود شامل دو ویژگی هستند که بر اساس رابطه ۴ با یکدیگر ترکیب می‌شوند (Nosrati, Rahmani, Jolfaei & Seifollahi, 2022):

$$\min\left(\sqrt[2]{|x^2 - y^2|}, \min(x, y)\right)$$ رابطه ۴)

که در این معادله، x مقدار ویژگی اول و y مقدار ویژگی دوم (هر خوشه شامل دو ویژگی است) هر خوشه است و Min نیز تابع مینیمم است که کوچک‌ترین عضو را برمی‌گرداند.

ارزیابی روش پیشنهادی

پایگاه داده استفاده‌شده

پایگاه داده‌های استفاده‌شده برای ارزیابی روش‌های پیشنهادی PU2 و PU3 هستند[1] که شامل مجموعه‌ای از ایمیل‌های شخصی است و برای اینکه محتوای آنها شخصی بماند، تمام کلمات با یک شناسه عددی

1. The PU corpora is available at http://www.csmining.org/index.php/cdmc-2013.html

جایگزین شده‌اند. معیارهای ارزیابی با رویه k-fold cross validation با مقدار k برابر ۱۰ انجام و گزارش شده است که در آن مجموعه داده به ۱۰ بخش مجزا تقسیم می‌شود و به‌ازای هر بخش الگوریتم یک بار اجرا می‌شود. در هر بار اجرا، یک بخش به‌عنوان مجموعه تست و ۹ بخش باقی‌مانده به‌عنوان مجموعه آموزشی استفاده می‌شوند. معیارهای ارزیابی در مجموعه تست در ۱۰ دور اجرا و سپس میانگین‌گیری شده و به‌عنوان معیار کل مجموعه داده برگردانده می‌شود. PU2 شامل ۷۱۰ پیام است که ۲۰ درصد آنها هرزنامه هستند و PU3 شامل ۴۱۳۰ پیام است که ۴۴ درصد آنها هرزنامه هستند.

معیارهای ارزیابی

معمولاً هر الگوریتم با هدف اجرای یک یا چند وظیفه مطرح می‌شود و در پایان، عملکرد آن با توجه به اینکه تا چه حد توانسته این وظایف را بهینه انجام دهد، سنجیده می‌شود. در واقع، پارامتر کارایی کمیتی برای سنجش این است که هدف مدنظر تا چه اندازه بهینه انجام شده است. برای ارزیابی دسته‌بندهای هرزنامه نیز از چند پارامتر ارزیابی استفاده می‌شود (Mohammad & Zitar, 2011) که در پژوهش حاضر از سه پارامتر صحت[1]، دقت[2] و بازخوانی[3] استفاده می‌شود. هرزنامه‌ها به‌عنوان کلاس مثبت و غیرهرزنامه‌ها به‌عنوان کلاس منفی در نظر گرفته شده‌اند.

معیار صحت

$$Accuracy\ (\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100$$ رابطه ۵)

معیار دقت

$$Precision\ (\%) = \frac{TP}{TP + FP} \times 100$$ رابطه ۶)

معیار بازخوانی

$$Recall(\%) = \frac{TP}{TP + FN} \times 100$$ رابطه ۷)

که در آن داریم:

TN (منفی درست): برابر است با تعداد پیام‌های منفی که دسته‌بند آنها را به‌درستی منفی تشخیص داده است.

TP (مثبت درست): برابر است با تعداد پیام‌های مثبت که دسته‌بند آنها را به‌درستی مثبت تشخیص داده است.

1. Accuracy
2. Precision
3. Recall

FP (مثبت اشتباه): برابر است با تعداد پیام‌های منفی که دسته‌بند آنها را به اشتباه مثبت تشخیص داده است.

FN (منفی اشتباه): برابر است با تعداد پیام‌های مثبت که دسته‌بند آنها را به اشتباه منفی تشخیص داده است.

ارزیابی

بر اساس آنچه گفته شد، سازوکارهای انتخاب ویژگی پیشنهادی را بر اساس روش تقسیم ویژگی‌ها و نگاشت ویژگی‌های هر خوشه می‌توان همانند جدول ۱ به چهار روش تقسیم‌بندی کرد که در این بخش به تجزیه و تحلیل هر یک از این چهار روش و مقایسه عملکرد آنها با یکدیگر پرداخته می‌شود و همچنین نتایج آنها با روش‌های پایه انتخاب ویژگی مقایسه می‌شود.

جدول ۱. نام اختصاری چهار سازوکار پیشنهادی

نام روش	سازوکار اجرایی
Df_Simple	تقسیم مبتنی بر معیار DF با استفاده از نگاشت ساده
Df_Advance	تقسیم مبتنی بر معیار DF با استفاده از نگاشت پیشرفته
PBIL _Simple	تقسیم مبتنی بر معیار PBIL با استفاده از نگاشت ساده
BPIL _ Advance	تقسیم مبتنی بر معیار PBIL با استفاده از نگاشت پیشرفته

از آنجا که هدف روش پیشنهادی بهبود روش‌های انتخاب ویژگی پایه با خوشه‌بندی ویژگی‌های انتخاب‌نشده است، به‌منظور نشان دادن برتری روش‌های پیشنهادی، عملکرد آن با دو روش دیگر مقایسه می‌شوند که عبارت‌اند از: روش مبتنی بر ویژگی‌های برتر و روش مبتنی بر کل ویژگی‌ها. در روش مبتنی بر ویژگی‌های برتر، عمل دسته‌بندی فقط بر اساس ویژگی‌های با ارزش بالای حاصل از انتخاب ویژگی اولیه انجام می‌گیرد و ویژگی‌های با ارزش پایین حذف می‌شوند. در روش مبتنی بر کل ویژگی‌ها، عملیات دسته‌بندی بر اساس کل ویژگی‌ها یعنی بر اساس کل ویژگی‌های با ارزش بالا و ارزش پایین انجام می‌شود.

در روش‌هایی که در آنها تقسیم ویژگی بر اساس معیار DF انجام می‌شود، ۵۰ درصد ویژگی‌ها به‌عنوان ویژگی برتر انتخاب شده و باقی به‌عنوان ویژگی با ارزش پایین در نظر گرفته می‌شوند. در روش‌های مبتنی بر PBIL نیز انتخاب این مجموعه از اختیار ما خارج است و خود الگوریتم به تعیین آن اقدام می‌کند. همچنین، در روش‌های مبتنی بر PBIL تعداد جمعیت اولیه برابر ۱۰ بوده و الگوریتم در ۱۰ دوره اجرا می‌شود.

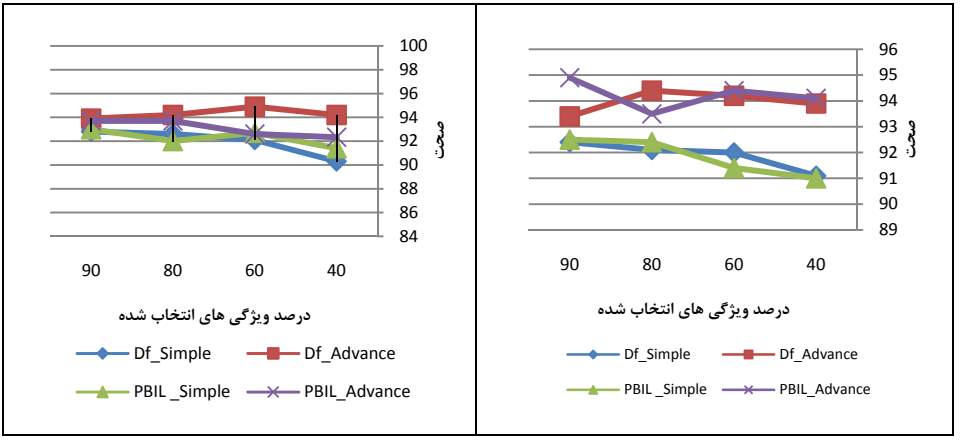
نتایج هر یک از چهار روش پیشنهادی

در وهله نخست به مقایسه نتایج چهار روش پیشنهادی بیان‌شده در جدول ۱ پرداخته می‌شود که نتایج آن بر اساس سه معیار صحت، دقت و بازخوانی در شکل‌های ۳ تا ۵ نشان داده شده است. همان‌طور که در این شکل‌ها مشاهده می‌شود، در اکثریت موارد روش‌های Df_Advance و PBIL_Advance به‌ترتیب دارای بهترین نتایج هستند و از دو روش دیگر عملکرد بهتری داشته‌اند. از آنجا که این دو روش مبتنی بر روش نگاشت پیشرفته بر اساس رابطه ۴ هستند، می‌توان بیان داشت که روش‌های مبتنی بر نگاشت پیشرفته، در مقایسه با روش نگاشت ساده، دارای نتایج بهتری بوده است. دو روش Df_Simple و PBIL_Simple نیز در اکثریت موارد کمترین نتایج را داشته‌اند.

برای داشتن درک بهتر، میانگین نتایج به‌دست‌آمده به‌ازای هر یک از روش‌های نگاشت و روش‌های انتخاب ویژگی اولیه در جدول ۲ نیز نشان داده شده است. همان‌طور که در جدول مشاهده می‌شود، روش نگاشت پیشرفته در هر سه معیار ارزیابی، در مقایسه با روش ساده، دارای خروجی بهتری بوده است. همچنین، روش انتخاب ویژگی مبتنی بر DF توانسته است در مقایسه با روش مبتنی بر PBIL، نتایج بهتری داشته باشد.

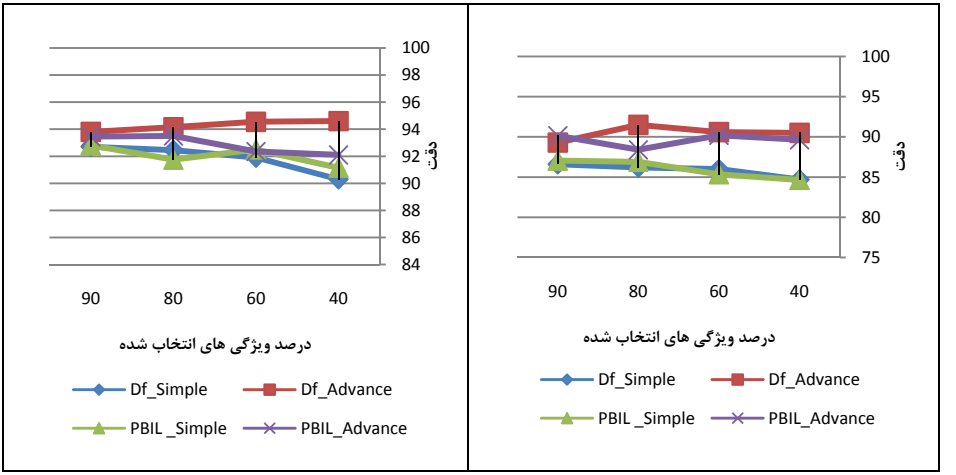
جدول ۲. مقایسه میانگین نتایج روش‌های نگاشت و انتخاب ویژگی اولیه

		صحت	دقت	بازخوانی
روش نگاشت	ساده	۹۲/۷	۹۰/۰۸	۹۲/۳۱
	پیشرفته	۹۳/۲۸	۹۰/۸۷	۹۲/۶۷
روش انتخاب ویژگی اولیه	مبتنی بر DF	۹۳/۰۳	۹۰/۶۲	۹۲/۵۲
	مبتنی بر PBIL	۹۲/۸۵	۹۰/۱۲	۹۲/۴۰



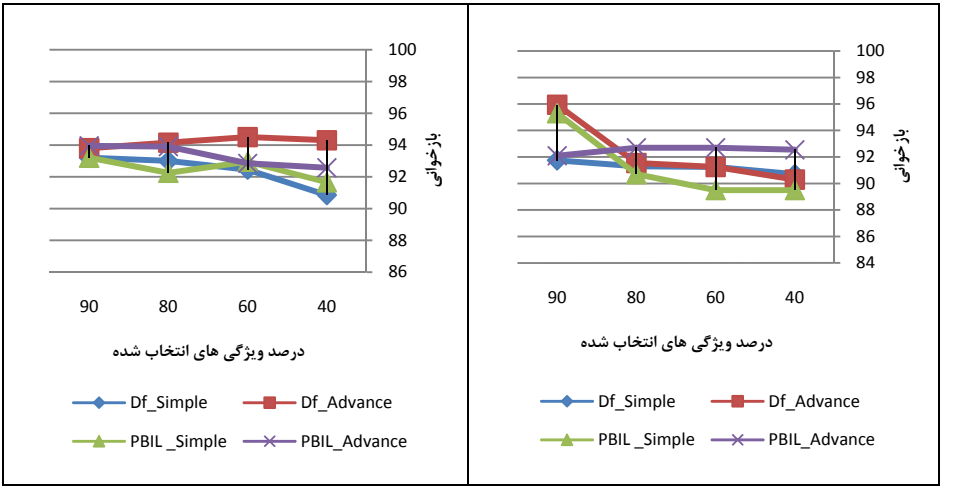
پایگاه داده PU3 پایگاه داده PU2

شکل ۳. نتایج پارامتر صحت (٪) روش‌های پیشنهادی در دو پایگاه داده PU2 و PU3



پایگاه داده PU3 پایگاه داده PU2

شکل ۴. نتایج پارامتر دقت (٪) روش‌های پیشنهادی در دو پایگاه داده PU2 و PU3



پایگاه داده PU3 پایگاه داده PU2

شکل ۵. نتایج پارامتر بازخوانی (٪) روش‌های پیشنهادی در دو پایگاه داده PU2 و PU3

مقایسه عملکرد روش‌های پیشنهادی با روش‌های انتخاب ویژگی پایه

در ادامه، به مقایسه عملکرد روش‌های پیشنهادی با دو رویکرد انتخاب ویژگی پایه DF و PBIL پرداخته می‌شود تا نشان داده شود که روش پیشنهادی قادر بوده است عملکرد این دو روش را بهبود بخشد. همچنین، نتایج به‌دست‌آمده در فقدان هرگونه روش انتخاب ویژگی گزارش خواهد شد تا مشخص شود کاربرد انتخاب ویژگی چه تأثیری در عملکرد مدل داشته است.

همان‌طور که در بخش‌های قبلی نیز بیان شد، بر اساس روش‌های خوشه‌بندی و همچنین بر اساس چگونگی نگاشت ویژگی‌های یک خوشه در این پژوهش، چهار رویکرد کلی مطرح شد که نتایج مقایسه هر یک در دو پایگاه داده PU2 و PU3 به‌طور جداگانه در جدول‌های ۳ تا ۶ آورده شده است. همان‌طور که در جدول‌ها نیز مشاهده می‌شود، روش‌های پیشنهادی Df_Simple، Df_Advance و PBIL_Advance توانسته‌اند در مقایسه با دو روش دیگر عملکرد بهتری داشته باشند و فقط روش PBIL_Simple نتوانسته است در مقایسه با روش مبتنی بر ویژگی‌های برتر عملکرد بهتری داشته باشد. همچنین، مشاهده می‌شود که روش مبتنی بر کل ویژگی‌ها در اکثریت موارد دارای ضعیف‌ترین نتایج بوده است.

روش پیشنهادی Df_Simple در هر سه معیار صحت، دقت و بازخوانی توانسته است در اکثریت موارد در مقایسه با دو روش دیگر عملکرد بهتری داشته باشد. روش پیشنهادی Df_Advance در دو معیار صحت و دقت توانسته است در اکثریت موارد نتایج بهتری داشته باشد، اما در معیار بازخوانی روش مبتنی بر ویژگی‌های برتر توانسته است عملکرد بهتری داشته باشد. روش پیشنهادی PBIL_Simple نیز در پایگاه داده PU2 نتوانسته بهترین نتایج را داشته باشد، اما در پایگاه داده PU3 تقریباً عملکرد قابل رقابتی با روش مبتنی بر ویژگی‌های برتر داشته است. روش پیشنهادی PBIL_Advance نیز بر اساس هر سه معیار صحت، دقت و بازخوانی توانسته است در اکثریت موارد در مقایسه با دو روش دیگر نتایج بهتری کسب کند. بنابراین با توجه به نتایج به‌دست‌آمده می‌توان گفت که روش پیشنهادی توانسته است به‌نحو قابل قبولی در مقایسه با روش مبتنی بر ویژگی‌ها برتر بهتر عمل کند.

جدول ۳. نتایج مقایسه‌ای روش پیشنهادی Df_Simple

		PU2				PU3			
		۴۰	۶۰	۸۰	۹۰	۴۰	۶۰	۸۰	۹۰
صحت	روش پیشنهادی	۹۱/۱	۹۲	۹۲/۱	۹۲/۴	۹۰/۳	۹۲/۱	۹۲/۶	۹۲/۸
	ویژگی‌های برتر	۸۹/۳	۹۰/۳	۹۰/۶	۹۰/۴	۸۹/۴	۹۰/۳	۹۰/۹	۹۱/۳
	کل ویژگی‌ها	۹۰/۶	۹۱/۴	۹۱/۵	۹۱/۷	۹۰/۹	۹۱/۹	۹۲/۱	۹۲/۳
دقت	روش پیشنهادی	۸۴/۷	۸۶	۸۶/۲	۸۶/۶	۹۰/۳	۹۱/۹	۹۲/۴۵	۹۲/۷
	ویژگی‌های برتر	۸۲/۱۵	۸۳/۴۵	۸۳/۸۵	۸۳/۶۵	۸۹/۴۵	۹۰/۳	۹۰/۸۵	۹۱/۱۵
	کل ویژگی‌ها	۸۳/۸۵	۸۵/۱	۸۵/۳	۸۵/۵	۹۰/۸۵	۹۱/۷۵	۹۲	۹۲/۱
بازخوانی	روش پیشنهادی	۹۰/۷	۹۱/۲۵	۹۱/۳	۹۱/۷۵	۹۰/۸۵	۹۲/۴۵	۹۳	۹۳/۲
	ویژگی‌های برتر	۸۹/۵۵	۹۰/۴۵	۹۰/۶	۹۰/۵۵	۸۹/۹۵	۹۰/۸	۹۱/۴	۹۱/۷۵
	کل ویژگی‌ها	۹۰/۶	۹۱/۱۵	۹۱/۲۵	۹۱/۳	۹۱/۴	۹۲/۳	۹۲/۵۵	۹۲/۷

جدول ۴. نتایج مقایسه‌ای روش پیشنهادی Df_Advance

		PU2				PU3			
		۴۰	۶۰	۸۰	۹۰	۴۰	۶۰	۸۰	۹۰
صحت	روش پیشنهادی	۹۳/۹	۹۴/۲	۹۴/۴	۹۳/۴	۹۴/۲	۹۴/۹	۹۴/۲	۹۳/۹
	ویژگی‌های برتر	۹۳/۵	۹۳/۸	۹۳/۵	۹۲/۵	۹۴/۳	۹۴/۳	۹۴	۹۳/۳
	کل ویژگی‌ها	۹۰/۶	۹۱/۴	۹۱/۵	۹۱/۷	۹۰/۹	۹۱/۹	۹۲/۱	۹۲/۳
دقت	روش پیشنهادی	۹۰/۵	۹۰/۶	۹۱/۵	۸۹/۳	۹۴/۶	۹۴/۵۵	۹۴/۱۵	۹۳/۸
	ویژگی‌های برتر	۸۸/۹۵	۸۹/۱۵	۸۸/۷	۸۷/۳۵	۹۴/۵۵	۹۴/۴	۹۳/۸۵	۹۳/۱۵
	کل ویژگی‌ها	۸۳/۸۵	۸۵/۱	۸۵/۳	۸۵/۵	۹۰/۸۵	۹۱/۷۵	۹۲	۹۲/۱
بازخوانی	روش پیشنهادی	۹۰/۳	۹۱/۲۵	۹۱/۵۵	۹۵/۹۵	۹۴/۳	۹۴/۵	۹۴/۱۵	۹۳/۸
	ویژگی‌های برتر	۹۱/۴	۹۲/۳۵	۹۱/۹۵	۹۵/۳	۹۴/۴	۹۴/۶۵	۹۴/۱	۹۳/۳۵
	کل ویژگی‌ها	۹۰/۶	۹۱/۱۵	۹۱/۲۵	۹۱/۳	۹۱/۴	۹۲/۳	۹۲/۵۵	۹۲/۷

جدول ۵. نتایج مقایسه‌ای روش پیشنهادی PBIL _Simple

		PU2				PU3			
		۴۰	۶۰	۸۰	۹۰	۴۰	۶۰	۸۰	۹۰
صحت	روش پیشنهادی	۹۱	۹۱/۴	۹۲/۴	۹۲/۵	۹۱/۴	۹۲/۷	۹۲	۹۳
	ویژگی برتر	۹۳/۴	۹۳/۴	۹۳/۵	۹۳/۷	۹۲/۲	۹۲/۲	۹۲/۱	۹۲/۶
	کل ویژگی‌ها	۹۰/۶	۹۱/۴	۹۱/۵	۹۱/۷	۹۰/۹	۹۱/۹	۹۱/۹	۹۲/۱
دقت	روش پیشنهادی	۸۴/۶۵	۸۵/۳۵	۸۶/۹	۸۷/۰۵	۹۱/۱۵	۹۲/۵	۹۱/۷۵	۹۲/۸
	ویژگی برتر	۸۸	۸۸/۳۵	۸۸/۶	۸۸/۵۵	۹۲/۰۵	۹۲	۹۱/۹	۹۲/۴
	کل ویژگی‌ها	۸۳/۸۵	۸۵/۱	۸۵/۳	۸۵/۵	۹۰/۸۵	۹۱/۷۵	۹۱/۷۵	۹۲
بازخوانی	روش پیشنهادی	۸۹/۵	۸۹/۵	۹۰/۷	۹۵/۳	۹۱/۶۵	۹۲/۹۵	۹۲/۲۵	۹۳/۲
	ویژگی برتر	۹۳/۲	۹۲/۱	۹۲/۲	۹۳/۰۵	۹۲/۵۵	۹۲/۵۵	۹۲/۴	۹۲/۸۵
	کل ویژگی‌ها	۹۰/۶	۹۱/۱۵	۹۱/۲۵	۹۱/۳	۹۱/۴	۹۲/۳	۹۲/۳	۹۲/۵۵

جدول ۶. نتایج مقایسه‌ای روش پیشنهادی PBIL_Advance

		PU2				PU3			
		۴۰	۶۰	۸۰	۹۰	۴۰	۶۰	۸۰	۹۰
صحت	روش پیشنهادی	۹۴/۱	۹۴/۴	۹۳/۵	۹۴/۹	۹۲/۳۳	۹۲/۶۰	۹۳/۷۰	۹۳/۷۰
	ویژگی‌های برتر	۹۴/۴	۹۳/۹	۹۳/۱	۹۳/۷	۹۲/۱۰	۹۲/۶۰	۹۳/۳۰	۹۳/۴۰
	کل ویژگی‌ها	۹۰/۶	۹۱/۴	۹۱/۵	۹۱/۷	۹۱/۱۰	۹۱/۹۰	۹۲/۱۰	۹۲/۳۰
دقت	روش پیشنهادی	۸۹/۶۵	۹۰/۲	۸۸/۴	۹۰/۱۵	۹۲/۱۰	۹۲/۳۵	۹۳/۵۰	۹۳/۴۵
	ویژگی‌های برتر	۹۰/۱۵	۸۹/۵	۸۷/۷	۸۸/۷۵	۹۱/۸۸	۹۲/۳۵	۹۳/۱۰	۹۳/۲۰
	کل ویژگی‌ها	۸۳/۸۵	۸۵/۱	۸۵/۳	۸۵/۵	۹۰/۹۵	۹۱/۷۵	۹۲/۰۰	۹۲/۱۰
بازخوانی	روش پیشنهادی	۹۲/۵۵	۹۲/۷	۹۲/۷	۹۲/۱	۹۲/۵۷	۹۲/۸۵	۹۳/۹۰	۹۳/۹۵
	ویژگی‌های برتر	۹۵/۵۵	۹۲/۲	۹۲/۴۵	۹۲/۵۵	۹۲/۴۰	۹۲/۹۰	۹۳/۵۵	۹۳/۷۵
	کل ویژگی‌ها	۹۰/۶	۹۱/۱۵	۹۱/۲۵	۹۱/۳	۹۱/۵۲	۹۲/۳۰	۹۲/۵۵	۹۲/۷۰

بحث

در این مقاله، یک سازوکار جدید کاهش ویژگی ارائه شد که در آن، برخلاف اکثریت روش‌های کنونی، بعد از عملیات انتخاب ویژگی اولیه و انتخاب زیرمجموعه‌ای از ویژگی‌های برتر، باقی ویژگی‌های انتخاب‌نشده نادیده گرفته نمی‌شوند، بلکه عملیات خوشه‌بندی و نگاشت روی آنها انجام می‌گیرد و هر خوشه به یک ویژگی جدید نگاشت می‌شود و بردار ویژگی نهایی از مجموعه ویژگی اولیه و ویژگی‌های نگاشت شده به دست می‌آید. این موضوع باعث می‌شود که برخلاف روش‌های مرسوم انتخاب ویژگی، هیچ اطلاعاتی دور ریخته نشود و در عین حال که فضای بردار ویژگی کاهش می‌یابد، هم‌زمان از اطلاعات کلیه ویژگی‌ها برای افزایش عملکرد مدل دسته‌بندی استفاده شود. مزیت دیگر روش پیشنهادی این است که به‌دلیل بهره‌گیری از تمامی ویژگی‌ها، هرزنامه‌نویس‌ها قادر به شناسایی ویژگی‌ها و تغییر آنها نیستند. در واقع، در روش پیشنهادی تلاش شده است تا از مزیت‌های هر دو روش انتخاب ویژگی و کاهش ویژگی مبتنی بر نگاشت ویژگی‌ها استفاده شود.

به‌منظور بررسی عملکرد روش پیشنهادی، عملکرد آن با نتایج انتخاب ویژگی اولیه بدون خوشه‌بندی مقایسه شد که حاکی از بهبود آن بود. در واقع، می‌توان این‌گونه استدلال کرد که افزودن ویژگی‌های نادیده‌گرفته از مرحله انتخاب ویژگی باعث افزایش اطلاعات و بهبود عملکرد دسته‌بند شده است. از طرف دیگر، روش پیشنهادی با روش مبتنی بر کل ویژگی‌ها نیز مقایسه شد که در اکثریت موارد نتایج بهتری به دست آمد. در واقع، در حالتی که هیچ‌گونه عملیات انتخاب ویژگی روی پیام‌ها انجام نمی‌شد و از تمامی ویژگی‌ها به‌منظور فرایند دسته‌بندی استفاده می‌شود، نتایج همواره بدتر از زمانی است که فقط روی مجموعه ویژگی‌های برتر اعمال می‌شود. این موضوع مؤید اهمیت و تأثیر مثبت فرایند انتخاب ویژگی بر مدل پیش‌بینی است.

روش پیشنهادی یک چارچوب کلی کاهش ابعاد به شمار می‌رود که می‌توان آن را روی هر یک از روش‌های انتخاب ویژگی موجود اعمال کرد. اگرچه در پژوهش حاضر فقط دو روش انتخاب ویژگی DF و PBIL در چارچوب پیشنهادی بررسی، تجزیه و تحلیل شدند، اما روش پیشنهادی محدود به این دو روش نیست. در واقع، عملکرد هر یک از روش‌های انتخاب ویژگی کنونی را می‌توان با اعمال روش پیشنهادی روی آنها و استفاده مجدد از ویژگی‌هایی که توسط هر یک از روش‌ها نادیده گرفته می‌شود، بهبود بخشید. یکی از موارد شایان ذکر روش پیشنهادی این است که در روش پیشنهادی از یک روش خوشه‌بندی ساده استفاده شده است که ویژگی‌ها را در خوشه‌های دوتایی تقسیم‌بندی می‌کند. اگرچه ممکن است هنگامی که تعداد ویژگی زیاد باشد با خوشه‌های زیادی سر و کار داشته باشیم، اما روش خوشه‌بندی پیشنهادی برخلاف روش‌هایی همچون K-mean دارای بار محاسباتی بالایی نیست و به‌طور ساده ویژگی‌ها را بر اساس معیار DF مرتب و ویژگی‌ها را به ترتیبی که در لیست مرتب‌شده قرار گرفته‌اند در خوشه‌های مجزا قرار می‌دهد.

نتیجه‌گیری و پیشنهادها

در این مقاله، چند روش کاهش ویژگی مبتنی بر خوشه‌بندی بیان و بررسی شد. روش پیشنهادی به این شکل بود که کل ویژگی‌های اولیه هر پایگاه داده با اعمال یک انتخاب ویژگی به دو مجموعه با ارزش بالا و مجموعه با ارزش پایین تقسیم می‌شدند، سپس در ویژگی‌های با ارزش پایین خوشه‌بندی دوتایی انجام شد و ویژگی‌های هر خوشه به یک ویژگی نهایی نگاشت شده و با ویژگی‌های با ارزش بالا ترکیب شده و بردار ویژگی نهایی را می‌سازند. تقسیم ویژگی‌ها بر اساس دو معیار DF و PBIL و ترکیب ویژگی‌ها نیز به دو صورت ساده و پیشرفته انجام شد، بنابراین با توجه به چگونگی تقسیم ویژگی‌ها و چگونگی نگاشت ویژگی‌ها چهار سازوکار خوشه‌بندی در مقاله حاضر ارائه و پیاده‌سازی شدند. نتایج نشان داد که به‌طور کلی روش مبتنی بر نگاشت پیشرفته دارای کارایی بیشتری است. نتایج حاصل از پیاده‌سازی روش‌های پیشنهادی نشان داد که روش Df_Advance یعنی روش مبتنی بر انتخاب ویژگی DF و نگاشت پیشرفته ویژگی‌های خوشه توانسته است بهترین نتایج را داشته باشد. همچنین، مقایسه نتایج به‌دست‌آمده از روش‌های پیشنهادی با روش‌های انتخاب ویژگی DF و BPIL حاکی از برتری روش‌های پیشنهادی در مقایسه با روش‌های انتخاب ویژگی پایه است. پیشنهادها برای ادامه کار می‌تواند بررسی سایر روش‌های انتخاب ویژگی، همچون روش‌های مبتنی بر یادگیری عمیق برای تقسیم ویژگی‌ها باشد. همچنین، بررسی سایر توابع خطی و غیرخطی نگاشت خوشه‌ها و سایر روش‌های خوشه‌بندی و اندازه‌های مختلف خوشه‌بندی نیز می‌تواند به‌عنوان ادامه این پژوهش مطرح شود.

فهرست منابع

Alharan, A. F., Fatlawi, H. K., & Ali, N. S. (2019). A cluster-based feature selection method for image texture classification. *Indonesian Journal of Electrical Engineering and Computer Science*, 14(3), 1433-1442.

Al-Rawashdeh, G., Mamat, R., & Abd Rahim, N. H. B. (2019). Hybrid water cycle optimization algorithm with simulated annealing for spam e-mail detection. *IEEE Access*, 7, 143721-143734.

Androutsopoulos, I., Koutsias, J., Chandrinos, K. V., & Spyropoulos, C. D. (2000, July). An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages. *In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 160-167).

Aziz, R., Verma, C. K., & Srivastava, N. (2017). Dimension reduction methods for microarray data: a review. *AIMS Bioengineering*, 4(2), 179-197.

Baluja, S., & Caruana, R. (1995). Removing the genetics from the standard genetic algorithm. *In Machine Learning Proceedings 1995* (pp. 38-46). Morgan Kaufmann.

Dada, E. G., Bassi, J. S., Chiroma, H., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*, 5(6), e01802.

DeBarr, D., & Wechsler, H. (2012). Spam detection using random boost. *Pattern Recognition Letters*, 33(10), 1237-1244.

Dehghan, Z., & Mansoori, E. G. (2018). A new feature subset selection using bottom-up clustering. *Pattern Analysis and Applications*, 21(1), 57-66.

Dhillon, I. S., Mallela, S., & Kumar, R. (2003). A divisive information theoretic feature clustering algorithm for text classification. *The Journal of machine learning research*, 3, 1265-1287.

Eesa, A. S., Abdulazeez, A. M., & Orman, Z. (2017). A DIDS Based on The Combination of Cuttlefish Algorithm and Decision Tree. *Science Journal of University of Zakho*, 5(4), 313-318.

Elhussein, M., & Brahimi, S. (2021). Clustering as feature selection method in spam classification: uncovering sick-leave sellers. *Applied Computing and Informatics*. https://doi.org/10.1108/ACI-09-2021-0248

Ghaleb, S. A., Mohamad, M., Fadzli, S. A., & Ghanem, W. A. H. (2021). Training Neural Networks by Enhance Grasshopper Optimization Algorithm for Spam Detection System. *IEEE Access*, 9, 116768-116813.

Gibson, S., Issac, B., Zhang, L., & Jacob, S. M. (2020). Detecting spam email with machine learning optimized with bio-inspired metaheuristic algorithms. *IEEE Access*, 8, 187914-187932.

Hong, Y., Kwong, S., Chang, Y., & Ren, Q. (2008). Unsupervised feature selection using clustering ensembles and population based incremental learning algorithm. *Pattern Recognition*, 41(9), 2742-2756.

Huang, X., Zhang, L., Wang, B., Li, F., & Zhang, Z. (2018). Feature clustering based support vector machine recursive feature elimination for gene selection. *Applied Intelligence*, 48(3), 594-607.

Li, S., Xia, R., Zong, C., & Huang, C. R. (2009, August). A framework of feature selection methods for text categorization. *In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (pp. 692-700).

Mao, J., Hu, Y., Jiang, D., Wei, T., & Shen, F. (2020). CBFS: a clustering-based feature selection mechanism for network anomaly detection. *IEEE Access*, 8, 116216-116225.

Metsis, V., Androutsopoulos, I., & Paliouras, G. (2006, July). Spam filtering with naive bayes-which naive bayes? *In CEAS*, 17, 28-69.

Mohammad, A. H., & Zitar, R. A. (2011). Application of genetic optimized artificial immune system and neural networks in spam detection. *Applied Soft Computing*, 11(4), 3827-3845.

Nosrati, V., Rahmani, M. (2021) Ensemble Bayesian Classification Using Genetic Algorithm Wrapper Feature Selection in Spam Detection, *Iranian Journal of Information Management*, 6 (2), 250-277.

Nosrati, V., Rahmani, M., Jolfaei, A., & Seifollahi, S. (2022). A Weak-Region Enhanced Bayesian Classification for Spam Content-Based Filtering. *Transactions on Asian and Low-Resource Language Information Processing.*

Rao, S., Verma, A. K., & Bhatia, T. (2021). A review on social spam detection: Challenges, open issues, and future directions. *Expert Systems with Applications*, 186, 115742.

Ravi Kumar, G., Murthuja, P., Anjan Babu, G., & Nagamani, K. (2022). An Efficient Email Spam Detection Utilizing Machine Learning Approaches. *In Innovative Data Communication Technologies and Application* (pp. 141-151). Springer, Singapore.

Sohrabi, M. K., & Karimi, F. (2015). A clustering based feature selection approach to detect spam in social networks. *International Journal of Information and Communication Technology Research*, 7(4), 27-33.

Soneji, H. N., Soman, A. S., Vyas, A., & Puthran, S. (2022). A Comprehensive Review of Fraudulent Email Detection Models. *In Proceedings of the Seventh International Conference on Mathematics and Computing* (pp. 109-127). Springer, Singapore.

Song, Q., Ni, J., & Wang, G. (2011). A fast clustering-based feature subset selection algorithm for high-dimensional data. *IEEE transactions on knowledge and data engineering*, 25(1), 1-14.

Vega-Pons, S., & Ruiz-Shulcloper, J. (2011). A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(03), 337-372.

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *Journal of Applied Science and Technology Trends*, 1(2), 56-70.

A Clustering Based Feature Selection Method in Spam Detection

Vahid Nosrati [*1]
Ph.D. Candidate, Computer Engineering, Faculty of Engineering, Arak University, Arak, Iran

Mohsen Rahmani
Associate Prof., Computer Engineering, Faculty of Engineering, Arak University, Arak, Iran

Abstract
One of the ways to detect spam is classifying emails into two categories: spam and non-spam. The high efficiency of machine learning methods in various fields has developed them in text clasification problems. The mechanism of machine learning-based classifiers that classify emails according to their content is based on a set of features, where due to the high volume of emails, using an efficient feature reduction algorithm plays an important role. Unlike the previous methods which select only the superior features and ignore the rest of the unselected features, in the proposed method of this article we try to use unselected features as well. The method is that after applying an initial feature selection, the unselected features are clustered and then each cluster is mapped to a new feature and the final feature vector forms from the selected ones and those mapped from the clusters. In this study, by applying two methods of selecting the initial feature and also two mapping functions, four methods were presented and analyzed using two datasets PU2 and PU3. The results of the analysis showed that the method based on feature selection DF and the advanced mapping function has the highest efficiency among all the proposed methods. Also, the proposed methods are more efficient than base feature selection methods (without clustering).

Keywords: Classification, Clustering, Email, Feature Reduction, Feature Selection, Spam.

1. Corresponding Author: vh_nosraty@yahoo.com