دوره دوازدهم، شماره دوم، پاییز و زمستان ۱۳۹۸

مجله فناوری اطلاعات در طراحی مهندسی

**Information Technology in Engineering Design**

**http://ited.sinaweb.net**

# تشخیص ایمیل‌های هرزنامه با استفاده از ترکیب الگوریتم‌های کلونی مورچه و آدابوست

مهدی قربانی وند [۱]  فرهاد سلیمانیان قره چپق * [۲]

(۱) گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران.

(۲) گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران. *

(تاریخ دریافت : ۱۳۹۸/۱۱/۲۱  تاریخ پذیرش : ۱۳۹۹/۳/۱۲)

**چکیده**

**حمله‌های اینترنتی و گسترش فعالیت هکرها برای به سرقت بردن اطلاعات کاربران مختلف روزبه‌روز در حال افزایش است. تهدیدهای اینترنتی هوشمندتر شده‌اند، اهدافشان متنوع‌تر و محدوده عملیاتی‌شان نیز گسترده‌تر شده است. اما بااین‌وجود، زمینه‌های اصلی فعالیت‌های آن‌ها تغییر چندانی نکرده است. بیشتر نفوذها از طریق سرویس‌های ایمیل انجام می‌گیرد. عمومیت و سادگی استفاده از ایمیل باعث شد تا مشکلاتی برای کاربران ایجاد گردد و مفهومی به نام ایمیل هرزنامه وارد دنیای فن‌آوری اطلاعات گردید. لذا باید نوع ایمیل‌ها ازنظر هرزنامه و غیر هرزنامه تشخیص داده شوند. در این مقاله، به‌منظور تشخیص ایمیل هرزنامه از مدل ترکیبی بهینه‌سازی کلونی مورچه و آدابوست استفاده‌شده است. در الگوریتم ترکیبی از الگوریتم بهینه‌سازی کلونی مورچه برای انتخاب ویژگی و از آدابوست برای طبقه‌بندی نمونه‌ها استفاده‌شده است. در الگوریتم آدابوست برمبنای اختصاص وزن به نمونه‌ها، نوع نمونه‌های ضعیف و قوی تشخیص داده می‌شوند. ارزیابی نتایج بر روی مجموعه داده *Spambase* با ٤٦٠١ نمونه نشان داده که درصد صحت مدل پیشنهادی برابر ۹٦/۲۷ درصد است و در مقایسه با الگوریتم‌های بهینه‌سازی اجتماع ذرات، ژنتیک، بهینه‌سازی کلونی مورچه، شبکه‌های عصبی پرسپترون چندلایه، شبکه‌های عصبی شعاعی و درخت تصمیم *C4.5* دقت بیشتری دارد.**

**واژه‌های کلیدی: تشخیص ایمیل هرزنامه ، الگوریتم بهینه‌سازی کلونی مورچه، الگوریتم آدابوست، بهینه‌سازی**

* عهده‌دار مکاتبات:

نشانی: گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران.

شماره تلفن: ۰۹۱٤۱۷٦٤٤۲۷  پست الکترونیکی: *farhad@gmail.com*

## ۱. مقدمه

اینترنت همواره از جهت‌های گوناگون موردنقد و ارزیابی قرار می‌گیرد. اما واقعیت این است که این شبکه عظیم مانند هر اجتماع عادی انسانی دیگر، برای افراد تهدیدها و خطراتی دارد. از نفوذ داده‌های مخرب گرفته تا تخریب داده‌های سالم و برهم زدن نظم شبکه همه و همه تنها به یک مورد بستگی دارد و آن بحث نظارت بر تبادل اطلاعات در محیط اینترنت است. ایمیل هرزنامه، شامل مجموعه‌ی گسترده‌ای از ایمیل‌های تبلیغاتی آلوده و مخرب است که بدون اجازه و با اهداف متفاوتی موجب خسارت، از بین رفتن داده‌ها و سرقت اطلاعات شخصی می‌شود. هرزنامه‌های تبلیغاتی با باز کردن صفحات خاص یا ظاهر کردن برگه‌های متعددی، کاربران را به خرید محصولات و یا استفاده از سرویس‌های خود توصیه می‌کنند. این ایمیل‌ها می‌توانند داده‌های افراد را موردحمله امنیتی قرار داده و در برخی موارد تنظیمات کامپیوتری را تغییر دهند. لذا باید نوع و محتوای آن‌ها تشخیص داده شود و از نفوذ آن‌ها جلوگیری شود.

در دنیای امروز، ایمیل بسیاری از جنبه‌های زندگی بشر را تحت تأثیر قرار داده است به‌گونه‌ای که حتی فضای دولت‌ها، شهرها، بانک‌ها، تجارت و غیره به فضایی دیجیتالی تبدیل‌شده است. امروزه ایمیل به یکی از مهم‌ترین و مبرم‌ترین نیازهای بشر و به‌عنوان ابزار ارتباطی در زندگی روزمره تبدیل‌شده است [۱]. این ابزار نوین، زندگی انسان را تحت تأثیر قرارداد و باگذشت زمان تکامل‌یافته و ساده‌تر، عام‌تر و کاربردی‌تر گردیده و جزء لاینفکی از دنیای مجازی و اینترنت شده است. متأسفانه همین عمومیت و سادگی استفاده از ایمیل باعث شد تا مشکلاتی برای کاربران ایجاد گردد و با توسعه آن مفهومی به نام ایمیل هرزنامه وارد دنیای کامپیوتر گردید [۲].

تعاریف متعددی از اینکه ایمیل هرزنامه چیست و اینکه چه فرقی با ایمیل‌های معتبر دارد، وجود دارد. کوتاه‌ترین تعریف متداول از بین تعاریف موجود در مورد ایمیل هرزنامه ، این‌گونه بیان‌شده که «یک ایمیل الکترونیکی ناخواسته» می‌باشد [۳][٤]. در تعریفی دیگر؛ ایمیل هرزنامه یک ایمیل «ناخواسته» است که به‌طور نامشخص، به‌طور مستقیم و یا غیرمستقیم توسط فردی که نسبتی با گیرنده ایمیل ندارد، فرستاده‌شده است [۵]. یک تعریف دیگر که به‌طور گسترده‌ای موردقبول واقع‌شده است بیان می‌کند که ایمیل هرزنامه اینترنتی یک یا چند پیغام ناخواسته است که به‌عنوان بخشی از مجموعه بزرگ‌تر از پیغام‌ها فرستاده می‌شود به‌طوری‌که همه این پیغام‌ها دارای یک محتوای یکسان هستند.

متون علمی زیادی تاکنون ارائه‌شده‌اند که به مشخصات پدیده ایمیل هرزنامه پرداخته‌اند، به‌طورکلی ایمیل هرزنامه با اهداف زیر ارسال می‌گردد: تبلیغ برای یک کالای خاص، سرویس خاص، یک ایده خاص، فریب کاربران برای استفاده از اطلاعات محرمانه آن‌ها، انتقال یک نرم‌افزار خرابکار به کامپیوتر کاربر و یا ایجاد یک خرابی به‌صورت موقتی در سرویس‌دهنده‌ی ایمیل [٦].

ایمیل هرزنامه که روزی کم‌اهمیت جلوه می‌نمود، امروزه به معضلی جدی برای میلیون‌ها کاربر ایمیل مبدل شده و مشکلات زیادی برای آن‌ها ایجاد نموده است و بدین ترتیب نقطه آغازین برای توسعه راهکارهای نوین مدیریت خودکار ایمیل‌های هرزنامه ایجاد گردید و استفاده از ابزار و متدهایی برای شناسایی و فیلتر ایمیل هرزنامه‌ها ضرورتی غیرقابل‌انکار شد. تاکنون فیلترهای ضد ایمیل هرزنامه مختلفی عرضه‌شده‌اند که اکثر آن‌ها بر اساس تطبیق قوانین ثابت عمل می‌کنند. قوانین این سامانه‌ها به‌صورت دستی توسط کاربر تعیین می‌شوند و شامل ویژگی‌ها و مشخصات ثابت ایمیل‌های نامعتبر یا

ایمیل هرزنامه هستند و عمل حذف ایمیل هرزنامه بر اساس آن‌ها انجام می‌شود، مثلاً فیلترهای استفاده‌شده در یاهو که کاربر نشانی‌هایی را به‌عنوان فرستنده ایمیل هرزنامه معرفی کرده و سرویس‌دهنده ایمیل بر اساس نشانی‌های معرفی‌شده، هر ایمیل دریافتی از آن‌ها را جداگانه ذخیره یا حذف می‌کند.

روش‌های مبارزه با ایمیل هرزنامه را می‌توان به دودسته تقسیم کرد [۷]:

۱) تغییر پروتکل (احراز هویت همه‌ی فرستندگان، برای هر ایمیل، متدی برای کپسوله‌سازی آدرس ایمیل):

یکی از راه‌های پایان دادن به ایمیل هرزنامه، بهبود و یا تعویض استانداردهای موجود انتقال ایمیل با جایگزین‌های جدید و ضد ایمیل هرزنامه است، چون این روش نیاز به مقدار زیادی ارتقا یا جایگزینی پروتکل‌های کنونی دارند، از آن‌ها استقبال چندانی نشد.

۲) فیلترها:

فیلترهای ایمیل هرزنامه با توجه به کار آیی بالا امروزه بسیار مورداستفاده قرار می‌گیرند. در یک دسته‌بندی کلی، فیلترها و تشخیص‌دهنده‌های ایمیل هرزنامه را می‌توان به پنج بخش دسته‌بندی کرد:

فیلترهای مبتنی بر محتوای ایمیل (کلمات و تصاویر)، فیلترهای مبتنی بر فهرست یا سرآیند، فیلترهای مبتنی بر عملیات آغازین، فیلترهای مبتنی بر تشخیص هویت فرستنده و فیلترهای مبتنی بر روش‌های شبکه‌های اجتماعی.

تقسیم‌بندی دیگری نیز برای فیلترهای ایمیل هرزنامه ذکرشده که عبارت است از:

الف) فیلترهای مبتنی بر روش غیر یادگیری که بر اساس اعتبارسنجی و الگوسازی کار می‌کنند.

ب) فیلترهای مبتنی بر یادگیری ماشین مانند فیلترهای بیزین، شبکه عصبی، ماشین بردار پشتیبان و... که این دسته کاربرد زیادی دارد [۸،۹].

در این مقاله، تشخیص ایمیل هرزنامه با استفاده از مدل ترکیبی الگوریتم بهینه‌سازی کلونی مورچه [۱۰] و آدابوست [۱۱] که هردو از الگوریتم‌های یادگیری ماشین می‌باشند، انجام می‌شود. از الگوریتم بهینه‌سازی کلونی مورچه برای انتخاب ویژگی و از آدابوست برای طبقه‌بندی نمونه‌ها استفاده می‌شود.

## ۲. کارهای قبلی

ایمیل هرزنامه ازآنجایی‌که شامل فایل‌های پیوستی مانند ویروس و عوامل نرم‌افزاری جاسوسی می‌باشد می‌تواند برای یک سیستم و دریافت‌کنندگان آن خطرناک باشد و باعث از بین رفتن اطلاعات شود. بنابراین نیاز به ابزارهایی جهت تشخیص هرزنامه بسیار حیاتی است. بسیاری از فن‌های تشخیص هرزنامه‌ها بر اساس روش‌های یادگیری ماشین پیشنهادشده‌اند و نتایج مقبولی داشته‌اند.

الگوریتم ماشین بردار پشتیبان میانگین [۱۲] که ترکیبی از ماشین بردار پشتیبان و کا-میانگین است برای طبقه‌بندی ایمیل هرزنامه پیشنهادشده است. در مدل ماشین بردار پشتیبان میانگین از کا-میانگین برای خوشه‌بندی نمونه‌ها و از ماشین بردار پشتیبان برای طبقه‌بندی استفاده‌شده است. در مدل کا-میانگین از معیار فاصله اقلیدسی برای تشخیص شباهت بین نمونه‌ها استفاده‌شده است. ارزیابی بر روی مجموعه داده Spambase با ٤٦٠١ نمونه انجام‌شده است. نتایج نشان داده که دقت مدل ماشین بردار پشتیبان و ماشین بردار پشتیبان میانگین به ترتیب برابر ۹٦،۳۰ و ۹۸،۰۱ می‌باشد.

مدل بهینه‌سازی اجتماع ذرات-شبکه عصبی مصنوعی تابع شعاعی پایه [۱۳] برمبنای شبکه عصبی مصنوعی و الگوریتم بهینه‌سازی اجتماع ذرات برای تشخیص ایمیل هرزنامه بر روی مجموعه داده Spambase اجراشده است. در مدل بهینه‌سازی اجتماع ذرات-شبکه عصبی مصنوعی تابع شعاعی پایه از شبکه‌های عصبی تابع پایه

شعاعی برای آموزش و تست نمونه‌ها و از بهینه‌سازی اجتماع ذرات برای بهینه‌سازی مقدار وزن نرون‌ها استفاده‌شده است. نتایج نشان داده که دقت الگوریتم ترکیبی مدل بهینه‌سازی اجتماع ذرات-تابع شعاعی پایه برابر ۹۳/۱ درصد است که در مقایسه با مدل‌های پرسپترون چندلایه و تابع شعاعی بیشتر است.

الگوریتم سیستم ایمنی مصنوعی برای تشخیص ایمیل هرزنامه برمبنای انتخاب ویژگی‌های مهم پیشنهادشده است [۱٤]. روال مدل این‌گونه است که در ابتدا عملیات پیش‌پردازش انجام می‌شود و کلمات مهم در متن ایمیل شناسایی و تکرار آن‌ها شمارش می‌شود و همچنین نمادها و کلمات مشکوک شناسایی و برای طبقه‌بندی از آن‌ها استفاده می‌شود. ارزیابی بر روی مجموعه داده TREC07 با ٤٥٤۱۹ ایمیل(۲۵۲۲۰ غیر هرزنامه و ۵۰۱۹۹ هرزنامه) نشان داده که دقت تشخیص برابر ۸٦/۲۰ ٪ می‌باشد.

مدل نیوی بیز [۱۵] برمبنای آموزش و تست نمونه‌ها بر روی مجموعه داده Ling-spam با ۹٦۰ نمونه اجراشده است. مدل نیوی بیز شامل مراحل پیش‌پردازش و استخراج ویژگی می‌باشد. برای تشخیص ویژگی‌های مهم از مراحل تکرار و تست استفاده‌شده است و ویژگی‌هایی انتخاب می‌شوند که دقت بیشتری دارند. نتایج نشان داده که مدل نیوی بیز در مقایسه با ماشین بردار پشتیبان مقدار خطای کمتری دارد.

مدل ترکیبی الگوریتم انتخاب منفی-شبکه عصبی مصنوعی [۱٦] به‌منظور انتخاب ویژگی و طبقه‌بندی ایمیل هرزنامه پیشنهادشده است. در مدل الگوریتم انتخاب منفی-شبکه عصبی مصنوعی از الگوریتم انتخاب منفی برای انتخاب ویژگی و از الگوریتم شبکه عصبی مصنوعی برای طبقه‌بندی نمونه‌ها استفاده‌شده است. الگوریتم انتخاب منفی یکی از الگوریتم‌های انتخاب ویژگی است که از بردارهای اولیه و وزن دهی برای انتخاب ویژگی‌ها استفاده می‌کند. الگوریتم شبکه عصبی مصنوعی که مهم‌ترین نوع آن پرسپترون چندلایه است برای طبقه‌بندی و کاهش خطا دقت بالایی دارد. نتایج بر روی ٤٦۰۱ نمونه ایمیل هرزنامه از مجموعه Spambase نشان داده که دقت مدل الگوریتم انتخاب منفی-ماشین بردار پشتیبان برابر ۹٤/۳۰٪ است که در مقایسه با ماشین بردار پشتیبان دقت بیشتری دارد.

مهم‌ترین مدل‌های داده‌کاوی برای تشخیص ایمیل هرزنامه بر روی مجموعه داده Spambase اجرا و تست‌شده‌اند [۱۷]. این مدل‌ها برمبنای تصمیم‌گیری، آموزش و تست نمونه‌ها عملیات طبقه‌بندی را انجام می‌دهند. نتایج نشان داده که دقت مدل‌های K نزدیک‌ترین همسایه و جنگل تصادفی در مقایسه با مدل‌های دیگر بیشتر است.

الگوریتم‌های ماشین بردار پشتیبان، درخت تصمیم J48، شبکه‌های بیز و روش یادگیری مبتنی بر نمونه ( Lazy IBK) که از فن‌های داده‌کاوی هستند بر روی ٤۳۰۷ نمونه ایمیل (٦۳۵ ایمیل هرزنامه) تست و اجراشده‌اند [۱۸]. ارزیابی در محیط وکا انجام‌شده است. نتایج نشان داده که درصد صحت در J48 برابر ۹۳/۳۱ درصد است که در مقایسه با مدل‌های دیگر بیشتر است.

مدل ترکیبی [19 NB-K-Means] برای تشخیص ایمیل هرزنامه بر روی مجموعه داده‌های Ling-spam، Assassin و Spambase تست و اجراشده است. از مدل نیوی بیز به‌منظور فاصله تشابه و از کا-میانگین برای مشابه بودن نمونه‌های داخل هر خوشه استفاده‌شده است. نتایج نشان داده که دقت مدل ترکیبی در مقایسه با نیوی بیز بیشتر است.

مدل ترکیبی بهینه‌سازی اجتماع ذرات-شبکه عصبی مصنوعی چندلایه به‌منظور تشخیص ایمیل هرزنامه پیشنهادشده است [۲۰]. از الگوریتم بهینه‌سازی اجتماع ذرات برای انتخاب ویژگی‌ها و از مدل پرسپترون

چندلایه برای آموزش و تست داده‌ها و طبقه‌بندی استفاده‌شده است. در مدل بهینه‌سازی اجتماع ذرات- شبکه عصبی مصنوعی چندلایه از شبکه عصبی پرسپترون با تابع فعال‌سازی سیگموئید برای لایه پنهان و ۸۰ درصد داده‌ها برای آموزش و ۲۰ درصد برای تست استفاده‌شده است. تعداد لایه‌های مخفی در شبکه عصبی مصنوعی پرسپترون چندلایه بین ۳-۱۵ لحاظ شده و تکرار الگوریتم بهینه‌سازی اجتماع ذرات برای انتخاب ویژگی برابر ۲۰۰ است. ارزیابی بر روی مجموعه داده Ling-Spam با ۲۵۸۹ ایمیل(٤٨١ ایمیل هرزنامه و ۲۱۷۱ ایمیل غیر هرزنامه) و مجموعه داده Spam-Assassin با ٦۰۰۰ نمونه ایمیل انجام‌شده است. ارزیابی بر روی مجموعه داده Spam-Assassin و Ling-Spam نشان داده که درصد صحت در مدل بهینه‌سازی اجتماع ذرات- شبکه عصبی مصنوعی چندلایه به ترتیب برابر ۹۹/۹۸ درصد و ۹۹/۷۹ درصد است. مقایسه‌ها نشان داده که مدل بهینه‌سازی اجتماع ذرات-شبکه عصبی مصنوعی چندلایه در مقایسه با مدل‌های ماشین بردار پشتیبان با تابع کرنل، ماشین بردار پشتیبان با تابع شعاعی و شبکه‌های عصبی پس انتشار دقت تشخیص بهتری دارد.

مدل‌های درخت تصمیم، ماشین بردار پشتیبان و شبکه‌های عصبی پس انتشار و ترکیب آن‌ها بر روی دو مجموعه داده با ١٤ ویژگی تست و اجر شده‌اند [۲۱]. مجموعه داده اولی شامل ٥٠٤ ایمیل (۳۳٦ غیر هرزنامه و ١٦٨ هرزنامه) و مجموعه داده دومی شامل ٦٥٧ ایمیل(۳۸۷ غیر هرزنامه و ۲۷۰ هرزنامه) است. در الگوریتم درخت تصمیم از آنتروپی، الگوریتم ماشین بردار پشتیبان از تابع کرنل و شبکه عصبی پس انتشار از خطای میانگین استفاده‌شده است. نتایج بر روی مجموعه داده اولی نشان داده که درصد صحت در مدل ترکیبی برابر ۹۱،۰۷ و در مدل‌های درخت تصمیم، ماشین بردار پشتیبان و شبکه عصبی پس انتشار به ترتیب برابر ۸۹/۸۸، ۸۸/٦۹ و ۸۹/۸۸ است و بر روی مجموعه داده دومی درصد صحت در مدل ترکیبی برابر ۹۱/۷۸ و در مدل‌های درخت تصمیم، ماشین بردار پشتیبان و شبکه عصبی پس انتشار به ترتیب برابر ۹۰/۸۷، ۹۰/۸۷ و ۸۹/۰٤ است.

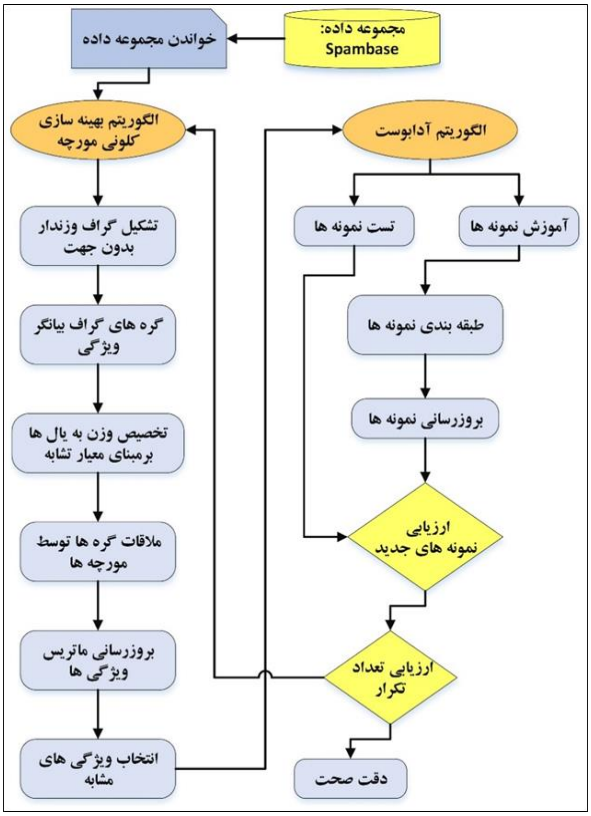
مدل ماشین بردار پشتیبان برمبنای تابع تبدیل شعاعی برای تشخیص ایمیل هرزنامه پیشنهادشده است [۲۲]. مدل تابع شعاعی از منحنی‌های نرمال در اطراف نقاط داده استفاده می‌کند و آن‌ها را طوری باهم جمع می‌کند که مرز تصمیم را بتوان به‌نوعی تعریف کرد. با استفاده از کرنل تابع شعاعی، فضای ویژگی از طریق یک تبدیل غیرخطی به دست می‌آید. برای ارزیابی مدل ماشین بردار پشتیبان- تابع شعاعی پایه از مجموعه داده‌های Spambase با ٤٦۰١ نمونه و SMS Spam با ٥٥٧٤ نمونه استفاده‌شده است. نتایج نشان داده که مدل ماشین بردار پشتیبان- تابع شعاعی پایه نمونه‌های هرزنامه را با ۹۳/۹۲ درصد تشخیص داده است.

برای طبقه‌بندی ایمیل هرزنامه از الگوریتم‌های مختلفی مانند ماشین بردار پشتیبان، کا-میانگین، شبکه‌های عصبی مصنوعی، الگوریتم بهینه‌سازی اجتماع ذرات، الگوریتم نیوی بیز، k نزدیک‌ترین همسایه و درختان تصمیم‌گیری استفاده‌شده است. هر مدل مزایا و معایب خاصی برای طبقه‌بندی دارد. بیشتر مدل‌های پیشنهادشده در بخش آموزش و تست مشکل‌دارند و باید تقسیم‌بندی این دو فاکتور به‌گونه‌ای انجام بگیرد که دقت تشخیص زیاد باشد. برای شبکه‌های عصبی مصنوعی باید وزن نرون‌ها دقیق باشد تا خطای آموزش کمتر باشد. درختان تصمیم‌گیری به زمان زیادی برای کشف بهترین ویژگی‌ها نیاز دارند و لذا باید مقایسه‌های زیادی انجام گیرد. در مدل‌های فرا ابتکاری این احتمال وجود دارد الگوریتم در یک‌راه‌حل گیر کند و با چندین تکرار به جواب‌های یکسان برسد. هدف از ارائه مدل پیشنهادی در این مقاله

این است که مشکلات مدل‌های موجود از قبیل انتخاب ویژگی و طبقه‌بندی را تا حدی برطرف کنیم.

### ۳. مدل پیشنهادی

مدل پیشنهادی، ترکیبی از الگوریتم بهینه‌سازی کلونی مورچه و آدابوست است. در مدل پیشنهادی از الگوریتم بهینه‌سازی کلونی مورچه به‌منظور انتخاب ویژگی و از آدابوست برای طبقه‌بندی نمونه‌ها استفاده می‌شود. در شکل (۱)، فلوچارت مدل پیشنهادی نشان داده‌شده است.




شکل ۱: فلوچارت مدل پیشنهادی

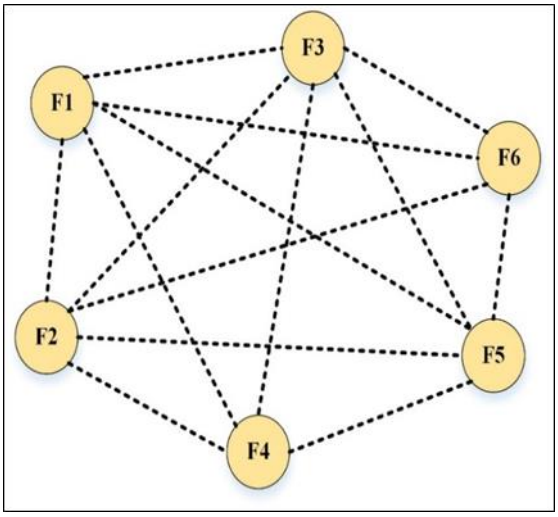
در ابتدا عمل نرمالیزه‌سازی را بر روی ویژگی‌ها انجام می‌دهیم. این عمل به‌منظور همگام‌سازی داده‌ها و افزایش صحت درستی انجام می‌گیرد. زیرا اگر داده‌ها در بازه بیشتر یا کمتر باشند باعث بی‌نظمی در داده‌ها و کاهش دقت خواهد شد. در معادله(۱)، v مقدار اصلی، minA و maxA حداقل و حداکثر مقدار ویژگی هستند [۲۳].

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\max_A - new\min_A) + new\min_A \qquad (۱)$$

قبل از فرایند انتخاب ویژگی باید فضای جستجو را برای الگوریتم بهینه‌سازی کلونی مورچه تشکیل دهیم. لذا، در ابتدا فضای جستجو به‌عنوان یک گراف G=(F,E) وزن‌دار بدون جهت نمایش داده می‌شود. که شامل مجموعه ویژگی‌ها است که هر ویژگی در گراف با یک گره نمایش داده می‌شود. و بیانگر لبه‌های گراف است. وزن یال‌های گراف برمبنای معیار تشابه بین دو گره Fi و Fj تعیین می‌شود. در معادله(۲)، پارامترهای Fi و Fj دو ویژگی در فضای pبعدی می‌باشند [۲۴].

$$sim(F_i, F_j) = \left|\frac{F_i . F_j}{\|F_i\|\|F_j\|}\right| = \left|\frac{\sum_{k=1}^{p} F_{ik} F_{jk}}{\left(\sqrt{\sum_{k=}^{p} F_{ik}^{\ 2}}\right)\left(\sqrt{\sum_{k=1}^{p} F_{jk}^{\ 2}}\right)}\right| \qquad (۲)$$

مدل پیشنهادی شامل مراحل انتخاب ویژگی و تشخیص کلاس است. در مجموعه داده هر نمونه یک ویژگی محسوب می‌شود و F یک بردار ویژگی از نمونه‌های موجود در مجموعه داده است. در گراف الگوریتم بهینه‌سازی کلونی مورچه، گره‌ها ویژگی‌ها را نشان می‌دهند و یال‌های بین آن‌ها نشان‌دهنده ویژگی بعدی برای انتخاب می‌باشند. هر گره معرف یک ویژگی است. به عبارتی ماتریس اولیه توسط گره‌ها تشکیل می‌شود. طبق معادله(۲)، مقادیر بین ویژگی‌ها برابر ۰ و ۱ است. اگر مقدار برابر ۱ باشد بدین معنی است که بین دو ویژگی تشابه وجود دارد و اگر برابر با ۰ باشد بدین معنی است که دو ویژگی مشابه هم نیستند. در شکل(۲)، گراف ویژگی‌ها به‌منظور انتخاب ویژگی نشان داده‌شده است. رئوس گراف شامل ویژگی‌ها و یال‌ها شامل وزن بین دو ویژگی می‌باشد. نوآوری مدل پیشنهادی در تخصیص وزن به یال‌ها است. یال‌هایی که توسط مورچه‌های بیشتری ملاقات می‌شوند به‌منزله این است که یک یال مناسب هستند و لذا گره‌های متصل به آن یال به‌عنوان ویژگی، انتخاب می‌شوند.




**شکل ۲: نمایش انتخاب ویژگی‌ها به‌عنوان یک گراف**

**مورچه kام بر روی ویژگی iام قرار می‌گیرد و ویژگی jام می‌تواند برمبنای احتمال انتخاب شود. برای انتخاب ویژگی بعدی از معادله(۳)، استفاده می‌شود [۲٤].**

$$j = \arg\max_{u \in J_i^k} \{[\tau_u][\eta(F_i, F_u)]^\beta\} \qquad (۳)$$

**در معادله(۳)، پارامتر $J_i^k$ مجموعه ویژگی‌های ملاقات نشده و $\tau_u$ مقدار فرومون داده‌شده به یال‌ها u است. و $\eta(F_i,F_u)=1/sim(F_i,F_u)$ مقدار عکس تشابه بین ویژگی‌های i و u است و مقدار فرومون را کنترل می‌کند. همچنین عمل به‌روزرسانی ماتریس ویژگی‌ها باعث می‌شود که ویژگی‌های مهم انتخاب شوند و مدل در زمان کمتری به جواب بهینه دست پیدا کند. در مدل پیشنهادی دقت طبقه‌بندی با استفاده از آدابوست برمبنای معادله (٤) که فاکتور اصلی برازندگی است محاسبه می‌شود [۲۵].**

$$Accuracy(i) = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \qquad (٤)$$

**در معادله(٤)، تفسیر پارامترها به شرح زیر می‌باشد. پارامتر TP (درست مثبت)؛ در این حالت تست نمونه‌ها باید مثبت باشد و به‌عنوان مثبت تعریف می‌شوند. پارامتر TN (درست منفی)؛ در این حالت تست نمونه‌ها باید منفی باشد و به‌عنوان منفی تعریف می‌شوند. پارامتر FP (نادرست مثبت)؛ در این حالت تست نمونه‌ها باید منفی باشد و به‌عنوان مثبت تعریف می‌شوند. پارامتر FN (نادرست منفی)؛ در این حالت تست نمونه‌ها باید مثبت باشد و به‌عنوان منفی تعریف می‌شوند. تابع هدف در مدل پیشنهادی برمبنای معادله(۵)، محاسبه می‌شود [۲٦].**

$$F(i) = \frac{Accuracy(i)}{1 + \lambda.n(i)} \qquad (۵)$$

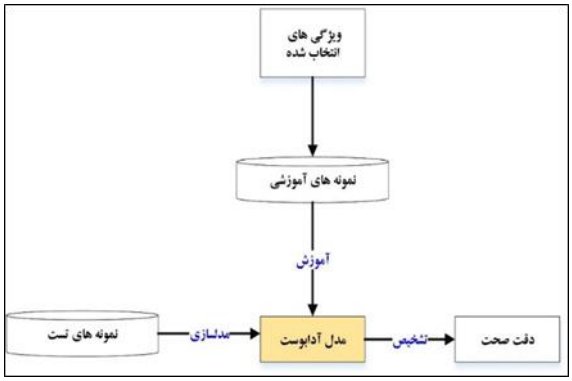
**در معادله(۵)، F(i) مقدار برازندگی مورچه iام، n(i) تعداد ویژگی انتخاب‌شده و Accuracy دقت طبقه‌بندی می‌باشد. همچنین مقدار پارامتر =λ۰,۰۱ می‌باشد.**

**۱-۳- طبقه‌بندی ویژگی‌ها**

**الگوریتم آدابوست یکی از روش‌های طبقه‌بندی است که برمبنای تکرار و خطا سعی می‌کند که نمونه‌های مشابه را دریک دسته قرار دهد. آموزش بر روی هر ویژگی انجام می‌گیرد و لذا برای حداقل کردن خطا از معادله(٦) استفاده می‌شود. در معادله(٦)، w وزن ویژگی‌ها و x و y مقدار ویژگی‌ها است [۱۱].**

$$\varepsilon_j = \sum_{i=1}^{n} w_i \left| h_j(x_i) - y_i \right| \qquad (٦)$$

**الگوریتم آدابوست به‌منظور کاهش خطا مراحل آموزش را در چندین مرحله تکرار می‌کند و در هر مرحله نمونه‌ها مشابه به هم در یک دسته قرار می‌گیرند. در مرحله آموزش الگوریتم آدابوست، برمبنای نرخ صحت به هر دسته‌بند ضعیف، وزنی اختصاص می‌یابد و به هر نمونه آموزش نیز وزنی منصوب می‌شود؛ که نشان‌دهنده درستی فرآیند دسته‌بندی است. درروند توالی اضافه شدن یک دسته‌بند ضعیف، اگر هر یک از نمونه‌های آموزش به‌درستی دسته‌بندی شود، وزنش کاهش می‌یابد. در غیر این صورت، وزن آن افزایش خواهد یافت؛ بنابراین دسته‌بند در تکرار بعد می‌تواند بر روی نمونه‌های اشتباه دسته‌بندی‌شده تمرکز کند.**




**شکل ۳: طبقه‌بندی ویژگی‌ها برمبنای مدل آدابوست**

هنگامی‌که که ویژگی‌ها توسط الگوریتم بهینه‌سازی کلونی مورچه انتخاب شدند باید برای طبقه‌بندی وارد الگوریتم آدابوست شوند. الگوریتم آدابوست برمبنای نمونه‌های آموزشی، مدل طبقه‌بندی را ایجاد می‌کند و سپس نمونه‌های تست برمبنای مدل‌سازی می‌توانند طبقه‌بندی شوند و درنهایت دقت تشخیص مدل پیشنهادی ارزیابی می‌گردد.

## ٤. ارزیابی و نتایج

ارزیابی و مقایسه مدل پیشنهادی در محیط سی‌شارپ ۲۰۱۷ انجام‌شده است. در این مقاله مجموعه داده‌های Spambase [۲۸] که یکی از اصلی‌ترین مجموعه داده‌ها است را مورد ارزیابی قرار می‌دهیم. مجموعه داده spambase مجموعه‌ای از ایمیل‌ها می‌باشد که حاوی ٤٦۰۱ پیام و ۵۷ ویژگی (یک ویژگی کلاس) بوده که توسط هاپکینز و همکارانش گردآوری‌شده است [۲۸]. این مجموعه داده پیش‌پردازش شده می‌باشد. مجموعه داده Spambase شامل دو کلاس ایمیل هرزنامه با ۱۸۱۳ نمونه (۳۹/٤ ٪) و ایمیل غیر هرزنامه با ۲۷۸۸ نمونه (٦۰/٦ ٪) است. ٤۸ ویژگی اول از مجموعه داده Spambase از تکرار بعضی از کلمات خاص گرفته‌شده است. ٦ ویژگی بعدی درصد وقوع یک کاراکتر خاص مانند ";"، "("، "["، "!"، "$" و "#" هستند. ۳ ویژگی بعدی نشان‌دهنده معیارهای مختلف تکرار حروف در متن پیام هستند. درنهایت آخرین ویژگی برچسب کلاس بوده که نشان‌دهنده این است که آیا یک نمونه ایمیل هرزنامه بوده و یا اینکه آن نمونه ایمیل غیر هرزنامه می‌باشد.

برای ارزیابی از پنج معیار دقت ، بازخوانی ، F-Measure، صحت و نرخ خطا استفاده‌شده است. نتایج ارزیابی برمبنای تکرار و تعداد ویژگی انجام‌شده است. تعداد تکرار و تعداد جمعیت اولیه در الگوریتم بهینه‌سازی کلونی مورچه به ترتیب برابر ۱۰۰۰ و ۱۰۰ هستند. همچنین در الگوریتم آدابوست یک نمونه با تمامی همسایه مقایسه می‌شود تا هنگامی‌که بهترین همسایه برای آن پیدا شود. در بیشتر موارد برای ارزیابی طبقه‌بندی و پیش‌بینی نمونه‌ها از ۸۰درصد داده‌ها برای آموزش و از ۲۰درصد داده‌ها برای تست استفاده می‌شود. لذا در این مقاله، تقسیم‌بندی داده‌ها طبق شیوه ذکرشده انجام‌گرفته است.

نتایج مدل پیشنهادی، باید در مرحله ارزیابی توسط معیارهای مهم ارزیابی [۲۷] مورد تحلیل قرار گیرد تا بتوان کار آیی آن را تعیین نمود. این معیارها را می‌توان هم برای مجموعه داده‌های آموزشی در مرحله یادگیری و هم برای مجموعه رکوردهای آزمایشی در مرحله ارزیابی محاسبه نمود.

| | |
|---|---|
| دقت $\Pr ecision = \frac{TP}{TP + FP}$ | (۷) |
| بازخوانی $\mathrm{Re}call = \frac{TP}{TP + FN}$ | (۸) |
| $F - Measure = \frac{2 * \Pr ecision * \mathrm{Re}call}{(\Pr ecision + \mathrm{Re}call)}$ | (۹) |
| درصد صحت $Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$ | (۱۰) |
| نرخ خطا $\mathrm{False\,Discovery\,Rate} = \frac{FP}{FP + TP}$ | (۱۱) |

معیار دقت شامل پارامترهای TP و FP است. در این معیار نمونه‌های نادرست مثبت هم به‌عنوان مثبت درست در دقت تشخیص تأثیرگذار هستند. معیار بازخوانی شامل پارامترهای TP و FN است. در این معیار

نمونه‌های نادرست منفی هم در دقت تأثیرگذار هستند. معیار F-Measure به‌منظور ارزیابی کلی معیارهای دقت و بازخوانی استفاده می‌شود.

در جدول(۱)، نتایج مقایسه مدل پیشنهادی با مدل‌های بهینه‌سازی کلونی مورچه و آدابوست برمبنای تکرار نشان داده‌شده است. جدول(۱)، نشان می‌دهد که مدل پیشنهادی با ۱۰۰۰ بار تکرار، بهترین درصد صحت را دارد و مقدار آن برابر با ۹۶/۲۷ درصد است. همچنین درصد صحت مدل پیشنهادی در تکرار ۵۰۰ برابر با ۹۲/۳۶ درصد است و درصد صحت مدل بهینه‌سازی کلونی مورچه و آدابوست در تکرار ۱۰۰۰ام به ترتیب برابر ۸۵/۱۳ و ۸۶/۲۷ درصد می‌باشند.

در مدل پیشنهادی تعداد ویژگی‌ها توسط الگوریتم بهینه‌سازی کلونی مورچه انتخاب می‌شوند. هدف انتخاب کردن ویژگی‌هایی است که دقت طبقه‌بندی با استفاده از آن‌ها حداکثر باشد. برای مثال برای ۳۲ ویژگی انتخابی در اجرای اول و دوم دقت تشخیص متفاوت است. و هر اجرا برای ۳۲ ویژگی، مقدار درصد صحت متفاوتی دارند. در ابتدا مدل پیشنهادی برمبنای تعداد ویژگی‌های کمتر از ۵۷ ارزیابی‌شده است و سپس به‌منظور اینکه کار آیی کلی مدل سنجیده شود از تعداد کل ویژگی‌ها برای ارزیابی استفاده‌شده است.

همان‌طور که در جدول(۱)، مشاهده می‌کنید زمان اجرای مدل پیشنهادی در مقایسه با الگوریتم بهینه‌سازی کلونی مورچه و آدابوست بیشتر است. به دلیل اینکه، مدل پیشنهادی از دو مرحله مهم برای طبقه‌بندی استفاده می‌کند. در هر تکرار تعداد ویژگی‌هایی که توسط الگوریتم بهینه‌سازی کلونی مورچه کشف می‌شوند با استفاده از آدابوست طبقه‌بندی می‌شوند و درصد صحت نتایج به‌دست‌آمده مقایسه می‌گردند و در بین آن‌ها، بهترین راه‌حل کشف می‌شود. جدول (۱) نشان می‌دهد که با حداکثر تکرار، درصد صحت بیشتر است. درصد صحت یکی از معیارهای اصلی برای سنجش و دقت تشخیص می‌باشد. در بیشتر تحقیقات بررسی‌شده از این معیار برای ارزیابی استفاده‌شده است. همان‌طور که در جدول (۱) مشاهده می‌شود دلیل انتخاب ۱۰۰۰ تکرار این است که درصد صحت بیشتر از بقیه تکرارها بوده است. خروجی مدل‌های الگوریتم بهینه‌سازی کلونی مورچه، آدابوست و مدل پیشنهادی را برای تعداد ویژگی‌های ۵، ۱۰، ۱۲، ۱۶، ۲۰، ۲۵، ۳۰، ۳۵، ۴۰، ۴۵، ۵۰، ۵۴ و ۵۷ با یکدیگر مقایسه نمودیم که نتایج به شرح جدول(۱) می‌باشد. در جدول(۱)، نتایج مدل پیشنهادی برمبنای ۱۰۰۰ تکرار و انتخاب ویژگی نشان داده‌شده است. جدول(۱)، نشان می‌دهد که اگر تعداد ویژگی‌ها کمتر باشد درصد صحت بیشتر است. زیرا تعداد همسایه‌های مشابه برای طبقه‌بندی بهتر تشخیص داده می‌شوند. درصد صحت مدل پیشنهادی با تمامی ویژگی‌ها(۵۷ ویژگی) برابر با ۹۶/۲۷ درصد است. جدول(۱)، نشان می‌دهد که با افزایش تعداد ویژگی‌ها، زمان اجرا نیز افزایش‌یافته است. البته در مدل پیشنهادی زمان اجرا در مقایسه با بهینه‌سازی کلونی مورچه و آدابوست بیشتر است.

در جدول(۱)، نتایج الگوریتم بهینه‌سازی کلونی مورچه به‌تنهایی به شرح زیر به‌دست‌آمده است. در ابتدا، انتخاب ویژگی توسط الگوریتم بهینه‌سازی کلونی مورچه انجام می‌گیرد و سپس با استفاده از نزدیک‌ترین همسایه عمل طبقه‌بندی انجام‌شده است. همچنین نتایج الگوریتم آدابوست به‌تنهایی به شرح زیر انجام‌شده است. در ابتدا، ویژگی‌ها به‌صورت تصادفی انتخاب می‌شوند و سپس طبقه‌بندی برمبنای الگوریتم آدابوست انجام می‌گیرد.

جدول ۱: نتایج مدل‌های بهینه‌سازی کلونی مورچه، آدابوست و مدل پیشنهادی برمبنای تکرار و انتخاب ویژگی برای ۵ تا ۳۰ ویژگی

| زمان اجرا | نرخ خطا | درصد صحت | F-Measure | بازخوانی | دقت | الگوریتم | تعداد ویژگی | تعداد تکرار |
|---|---|---|---|---|---|---|---|---|
| ۰.۹۵۲۱ | ۷.۸۲ | ۹۲.۰۶ | ۹۲.۵۱ | ۹۲.٦٨ | ۹۲.۳۵ | بهینه‌سازی کلونی مورچه | ۵ | ۱۰۰۰ |
| ۰.۹٦٤٨ | ٦.۲۵ | ۹٤.۱۳ | ۹٤.۳۲ | ۹٤.۳۹ | ۹٤.۲٦ | آدابوست | | |
| ۱.۳۹٤۷ | ۱.۲۰ | ۹۸.۸۳ | ۹۹.۰۰ | ۹۹.٤۹ | ۹۸.۵۳ | مدل پیشنهادی | | |
| ۰.۹۸۵٤ | ۹.۱٦ | ۹۱.۲٤ | ۹۱.۹۷ | ۹۲.۰۷ | ۹۱.۸۹ | بهینه‌سازی کلونی مورچه | ۱۰ | |
| ۰.۹۸٦۳ | ۷.۲۹ | ۹۳.۱۷ | ۹۳.۸۹ | ۹٤.۰۹ | ۹۳.۷۱ | آدابوست | | |
| ۱.٤۲۵۲ | ۱.٤۹ | ۹۸.۵۱ | ۹۸.۷٦ | ۹۹.۰۷ | ۹۸.٤٦ | مدل پیشنهادی | | |
| ۰.۹۹۲۱ | ۹.۲۳ | ۹۲.۵۳ | ۹۱.٦۳ | ۹۱.٦۸ | ۹۱.۵۸ | بهینه‌سازی کلونی مورچه | ۱۲ | |
| ۰.۹۹۵۳ | ۸.۱۳ | ۹۳.٤۲ | ۹۳.۵۰ | ۹۳.۷۳ | ۹۳.۲۸ | آدابوست | | |
| ۱.٤۳۱۰ | ۱.۸۱ | ۹۸.۱۹ | ۹۸.۵۳ | ۹۸.۹٦ | ۹۸.۱۲ | مدل پیشنهادی | | |
| ۰.۹۹۸۹ | ۹.۱۵ | ۹۱.٦۸ | ۹۱.٦٤ | ۹۲.۰۳ | ۹۱.۳٦ | بهینه‌سازی کلونی مورچه | ۱٦ | |
| ۰.۹۹٦۲ | ۸.٤۹ | ۹۳.۱۷ | ۹۳.٤۲ | ۹۳.۷۹ | ۹۳.۰٦ | آدابوست | | |
| ۱.٤۵٦۲ | ۲.۷۰ | ۹۷.۳۲ | ۹۸.۱۸ | ۹۸.۵۳ | ۹۷.۸٤ | مدل پیشنهادی | | |
| ۰.۹۹۷٦ | ۹.۲٦ | ۹۲.۰۸ | ۹۰.۵۱ | ۹۰.٦٤ | ۹۰.۳۹ | بهینه‌سازی کلونی مورچه | ۲۰ | |
| ۰.۹۹۹۲ | ۸.۰۳ | ۹۳.۱٤ | ۹۲.۹۳ | ۹۳.۰۱ | ۹۲.۸٦ | آدابوست | | |
| ۱.٤۵۸۱ | ۲.۹۳ | ۹۷.۰۸ | ۹۷.۸۹ | ۹۸.۱۷ | ۹۷.٦۲ | مدل پیشنهادی | | |
| ۱.۰۱۲٦ | ۹.۳٦ | ۹۰.۱٦ | ۹۰.۸۱ | ۹۰.۸۱ | ۹۰.۳۲ | بهینه‌سازی کلونی مورچه | ۲۵ | |
| ۱.۰۱۵٤ | ۷.٤۹ | ۹۲.۸۱ | ۹۲.٦۷ | ۹۲.۸۹ | ۹۲.٤۷ | آدابوست | | |
| ۱.٤۹۸۱ | ۲.۵۳ | ۹۷.٤۷ | ۹۷.٤٤ | ۹۷.٦٤ | ۹۷.٤۲ | مدل پیشنهادی | | |
| ۱.۰۲۳٦ | ۱۰.۱٦ | ۸۹.۳۱ | ۸۹.۷۸ | ۸۹.۸٦ | ۸۹.۷۲ | بهینه‌سازی کلونی مورچه | ۳۰ | |
| ۱.۰۳۰۵ | ۹.٦۹ | ۹۰.۱٦ | ۹۰.۹۹ | ۹۱.۱۰ | ۹۰.۸۹ | آدابوست | | |
| ۱.٤۹۸٤ | ۲.۰۸ | ۹۷.۹۲ | ۹۷.۱٤ | ۹۷.۳۷ | ۹٦.۹۲ | مدل پیشنهادی | | |

**جدول (۲)، نتایج مدل‌های بهینه‌سازی کلونی مورچه، آدابوست و مدل پیشنهادی برمبنای ۱۰۰۰ بار تکرار و تعداد ویژگی‌های متفاوت نشان می‌دهد. اگر تعداد ویژگی برابر با ۳۵ باشد آنگاه درصد صحت برابر با ۹٦/۷٦ درصد است. اگر تعداد ویژگی برابر با ٤۰ باشد آنگاه درصد صحت برابر با ۹٦/۷٤ درصد است. اگر تعداد ویژگی برابر با ٤۵ باشد آنگاه درصد صحت برابر با ۹٦/۹۳ درصد است. اگر تعداد ویژگی برابر با ۵۰ باشد آنگاه درصد صحت برابر با ۹٦/۷۸ درصد است. اگر تعداد ویژگی برابر با ۵٤ باشد آنگاه درصد صحت برابر با ۹٦/۳۰ درصد است. اگر تعداد ویژگی برابر با ۵۷ باشد آنگاه درصد صحت برابر با ۹٦/۲۷ درصد است.**

**جدول ۲: نتایج مدل‌های بهینه‌سازی کلونی مورچه، آدابوست و مدل پیشنهادی برمبنای تکرار و انتخاب ویژگی برای ۳۵ تا ۵۷ ویژگی**

| زمان اجرا | نرخ خطا | درصد صحت | F-Measure | بازخوانی | دقت | الگوریتم | تعداد ویژگی | تعداد تکرار |
|---|---|---|---|---|---|---|---|---|
| ۱.۰۳۲۸ | ۱۰.۷۹ | ۸۹.٤۵ | ۸۹.٦۱ | ۸۹.۷۵ | ۸۹.٤۸ | بهینه‌سازی کلونی مورچه | ۳۵ | ۱۰۰۰ |
| ۱.۰٤۱۳ | ۹.٤۸ | ۹۰.۳٤ | ۹۰.٤۳ | ۹۰.٦۳ | ۹۰.۲۵ | آدابوست | | |
| ۱.۵۰۲۳ | ۳.۲٤ | ۹٦.۷٦ | ۹٦.۹٤ | ۹۷.۱۵ | ۹٦.۷٤ | مدل پیشنهادی | | |
| ۱.۰٤۱۵ | ۱۱.٦۳ | ۸۸.۵٦ | ۸۸.٦۹ | ۸۸.۹۱ | ۸۸.٤۹ | بهینه‌سازی کلونی مورچه | ٤۰ | |
| ۱.۰٤۳۲ | ۹.۲۰ | ۹۰.۳۲ | ۸۹.۹۲ | ۹۰.۰۵ | ۸۹.۹٤ | آدابوست | | |
| ۱.۵۹۱۳ | ۳.۲٦ | ۹٦.۷٤ | ۹۷.۰۷ | ۹۷.۸٦ | ۹٦.۳۱ | مدل پیشنهادی | | |
| ۱.۰٤۹۲ | ۱۳.۷۹ | ۸۸.۱٦ | ۸۷.۳٦ | ۸۷.۵۹ | ۸۷.۱۵ | بهینه‌سازی کلونی مورچه | ٤۵ | |
| ۱.۰۵۳۲ | ۱۲.۰٤ | ۸۹.٤۹ | ۸۹.٦۳ | ۸۹.۹۲ | ۸۹.۳۵ | آدابوست | | |
| ۱.٦۲۱۵ | ۳.۲۵ | ۹٦.۹۳ | ۹۵.۸۲ | ۹٦.۰۵ | ۹۵.٦۱ | مدل پیشنهادی | | |
| ۱.۰۵۹٦ | ۱٤.۹۲ | ۸۷.۲۳ | ۸٦.٤۱ | ۸٦.۷۳ | ۸٦.۱۱ | بهینه‌سازی کلونی مورچه | ۵۰ | |
| ۱.۰٦۰۲ | ۱۳.۰۲ | ۸۸.۹۲ | ۸۸.۸۲ | ۸۹.۰٦ | ۸۸.۵۹ | آدابوست | | |
| ۱.٦۲٦۹ | ۳.٦۰ | ۹٦.۷۸ | ۹۵.۸٦ | ۹٦.۲۸ | ۹۵.٤۵ | مدل پیشنهادی | | |
| ۱.۰٦٤۸ | ۱٤.٦۲ | ۸٦.۷۹ | ۸۵.۳۸ | ۸۵.۷۲ | ۸۵.۰٦ | بهینه‌سازی کلونی مورچه | ۵٤ | |
| ۱.۰٦۵۵ | ۱۳.٤۸ | ۸۷.۱٤ | ۸۷.۰٤ | ۸۷.۱۷ | ۸٦.۹۲ | آدابوست | | |
| ۱.٦۳۲۵ | ۳.۹۰ | ۹٦.۳۰ | ۹۵.۵۱ | ۹۵.۸۱ | ۹۵.۲۲ | مدل پیشنهادی | | |
| ۱.۰٦۳۸ | ۱۵.۲٦ | ۸۵.۱۳ | ۸٤.۲۷ | ۸٤.۵۲ | ۸٤.۰۳ | بهینه‌سازی کلونی مورچه | ۵۷ | |
| ۱.۰٦٤۰ | ۱٤.۸٤ | ۸٦.۲٦ | ۸۵.۷۱ | ۸۵.۸۲ | ۸۵.٦۲ | آدابوست | | |
| ۱.۷۲۳۱ | ۳.۹۰ | ۹٦.۲۷ | ۹۵.۰۳ | ۹۵.۰۹ | ۹٤.۹۸ | مدل پیشنهادی | | |

### ٤-۱- مقایسه و ارزیابی

**برای مقایسه و ارزیابی مدل پیشنهادی با مدل‌های دیگر از مجموعه داده Spambase [28] استفاده‌شده است. مقاله‌های مقایسه‌ای شامل عنوان دقیق تشخیص ایمیل هرزنامه نیستند اما به دلیل اینکه برمبنای نوع کاربرد از مجموعه داده Spambase برای انتخاب ویژگی و طبقه‌بندی استفاده کرده‌اند. بدین منظور، در این بخش برای مقایسه فقط از مجموعه داده Spambase استفاده‌شده است.**

**در جدول(۳)، مقایسه مدل پیشنهادی با مدل‌های مختلف برمبنای درصد صحت نشان داده‌شده است. درصد صحت مدل پیشنهادی (۹٦/۲۷ درصد) در مقایسه با مدل‌های ارائه‌شده که در آن‌ها از الگوریتم‌هایی مانند درخت تصمیم‌گیری C4.5 (۸۵/۳۳ درصد)، نیوی بیز (۹۲/۸۹ درصد)، K نزدیک‌ترین همسایه (۸۰/۸۳ درصد) و ماشین بردار پشتیبان (۸۵/۸۵) به‌عنوان طبقه‌بندی کننده و الگوریتم‌های مطرح انتخاب ویژگی استفاده‌شده، بیشتر است.**

به‌منظور معادل‌سازی مدل پیشنهادی با مدل‌های ارائه‌شده [۲۹] از معیار انتخاب ویژگی به‌عنوان ملاک درصد صحت استفاده‌شده است. تعداد ویژگی‌ها در الگوریتم انتخاب ویژگی مبتنی بر همبستگی [۲۹] به‌طور میانگین برابر با ۱۸ می‌باشد. تعداد ویژگی‌ها در الگوریتم انتخاب ویژگی مبتنی بر تقابل به‌طور میانگین برابر با ۲۰ می‌باشد. تعداد ویژگی‌ها در الگوریتم انتخاب ویژگی مبتنی بر فیلتر سازگاری به‌طور میانگین برابر با ۱۶ می‌باشد. تعداد ویژگی‌ها در الگوریتم انتخاب ویژگی مبتنی بر نرخ اطلاعات به‌طور میانگین برابر با ۱۸ می‌باشد. تعداد ویژگی‌ها در الگوریتم انتخاب ویژگی مبتنی بر الگوریتم ReliefF به‌طور میانگین برابر با ۱۷ می‌باشد.

در [۲۹]، الگوریتم‌های C4.5، نیوی بیز، k نزدیک‌ترین همسایه و ماشین بردار پشتیبان با استفاده از روش انتخاب ویژگی مبتنی بر همبستگی بهبودیافته، انتخاب ویژگی مبتنی بر تقابل بهبودیافته، انتخاب ویژگی مبتنی بر سازگاری بهبودیافته، انتخاب ویژگی مبتنی بر نرخ اطلاعات بهبودیافته و انتخاب ویژگی مبتنی بر الگوریتم ReliefF بهبودیافته دارای درصد صحت بیشتری هستند.

جدول ۳: مقایسه مدل پیشنهادی با مدل‌های مختلف برمبنای درصد صحت

| الگوریتم‌های انتخاب ویژگی | الگوریتم‌های طبقه‌بندی/درصد صحت | | | | منابع |
|---|---|---|---|---|---|
| | ماشین بردار پشتیبان | K نزدیکترین همسایه | نیوی بیز | C4.5 | |
| انتخاب ویژگی مبتنی بر همبستگی | ۸۵.۸۵ | ۷۹.۱٤ | ۵۷.٦۹ | ۸۱.۱٦ | [۲۹] |
| انتخاب ویژگی مبتنی بر همبستگی کلاس‌بندی | ۸۲.۲۷ | ۷۷.۳۱ | ۵۷.۲٤ | ۷۹.۲۷ | |
| انتخاب ویژگی مبتنی بر همبستگی بهبودیافته | ۸۵.٤٦ | ۷۹.۹۲ | ۵۸.۸۷ | ۷۹.۷۳ | |
| انتخاب ویژگی مبتنی بر تقابل | ۸۰.۳۱ | ۷۹.۷۳ | ۵۷.۹۵ | ۷۸.۱٦ | [۲۹] |
| انتخاب ویژگی مبتنی بر تقابل کلاس‌بندی | ۸۱.٤۹ | ۷٦.۸٦ | ۷٤.۷۷ | ۸۰.۸۳ | |
| انتخاب ویژگی مبتنی بر تقابل بهبودیافته | ۸۱.۱۰ | ۷٦.۹۲ | ۷۸.٤۲ | ۸۰.٤٤ | |
| انتخاب ویژگی مبتنی بر سازگاری | ۸۱.۸۸ | ۸۰.۸۳ | ۹۱.۰۰ | ۸٤.٦۲ | [۲۹] |
| انتخاب ویژگی مبتنی بر سازگاری کلاس‌بندی | ۸۱.۹٤ | ۷۷.۳۸ | ۹۲.۸۹ | ۷۹.۳٤ | |
| انتخاب ویژگی مبتنی بر سازگاری بهبودیافته | ۸۱.۱٦ | ۷٦.۷۹ | ۹۲.۰۵ | ۸۵.۲۷ | |
| انتخاب ویژگی مبتنی بر نرخ اطلاعات | ۸۳.۸۳ | ۷۸.٦۲ | ۷٦.۵۳ | ۸۳.۸۳ | [۲۹] |
| انتخاب ویژگی مبتنی بر نرخ اطلاعات کلاس‌بندی | ۸۳.۳۸ | ۷۸.٤۲ | ۸۹.۷۰ | ۸۵.۳۳ | |
| انتخاب ویژگی مبتنی بر نرخ اطلاعات بهبودیافته | ۸۳.۳۸ | ۷۷.۷۱ | ٦٦.۹۵ | ۸۱.٦۸ | |
| انتخاب ویژگی مبتنی بر الگوریتم ReliefF | ۸۱.۹٤ | ۷٦.۹۹ | ٤۱.۸۵ | ۷۸.۸۱ | [۲۹] |
| انتخاب ویژگی مبتنی بر الگوریتم ReliefF کلاس‌بندی | ۸۳.۵۷ | ۸۰.۷۰ | ۹۱.۷۹ | ۸٤.۷۵ | |
| انتخاب ویژگی مبتنی بر الگوریتم ReliefF بهبودیافته | ۸۳.۷۷ | ۸۰.۱۸ | ۹۲.۰۵ | ۸٤.۸۸ | |
| مدل پیشنهادی | ۹٦.۲۷ | | | | |

طبق جدول(٤)، در [۱۲] از دو تکنیک ماشین بردار پشتیبان برای تشخیص ایمیل هرزنامه استفاده‌شده است. در ماشین بردار پشتیبان درصد صحت برابر با ۹٦/۳۰ درصد و در ماشین بردار پشتیبان میانگین برابر با ۹۸/۰۱ است. در ماشین بردار پشتیبان برای بهینه‌سازی طبقه‌بندی از الگوریتم کا-میانگین استفاده‌شده است. درصد صحت ماشین بردار پشتیبان میانگین در مقایسه با مدل پیشنهادی بیشتر است. در [۳۰]، درصد صحت درخت تصمیم‌گیری در مقایسه با آنالیز تشخیص خطی بیشتر است. در [۳٦]، درصد صحت ماشین بردار پشتیبان خطی، ماشین بردار پشتیبان چندجمله‌ای و ماشین بردار پشتیبان کرنل به ترتیب برابر با ۹۳/۲٤، ۹۳/۳۷ و ۹۳/۸۷ می‌باشد.

جدول(٤)، نشان می‌دهد که در [۳۱]، درصد صحت بهبود بهینه‌سازی اجتماع ذرات-نزدیک‌ترین همسایه، بهبود بهینه‌سازی اجتماع ذرات-k نزدیک‌ترین همسایه و بهبود بهینه‌سازی اجتماع ذرات-k نزدیک‌ترین همسایه وزنی به ترتیب برابر با ۸٤/۲۸، ۸۳/۷۹ و ۸۳/۱۸ می‌باشد و درصد صحت بهبود بهینه‌سازی کلونی مورچه-نزدیک‌ترین همسایه، بهبود بهینه‌سازی کلونی مورچه-k نزدیک‌ترین همسایه و بهبود بهینه‌سازی کلونی مورچه-k نزدیک‌ترین همسایه وزنی به ترتیب برابر با ۸۳/۰۱، ۸۳/٤٤ و ۸۲/۱۱ می‌باشد. به‌منظور معادل‌سازی مدل پیشنهادی با مدل‌های مبتنی بر بهینه‌سازی اجتماع ذرات [۳۱]، تکرار و تعداد نسل به ترتیب برابر با ۲۰۰ و ۵۰ لحاظ شده‌اند.

جدول ٤: مقایسه مدل پیشنهادی با مدل‌های مبتنی بر بهینه‌سازی گروهی برمبنای درصد صحت

| درصد صحت | مدل‌ها | منابع |
|---|---|---|
| ۹۶.۳۰ | ماشین بردار پشتیبان | [۱۲] |
| ۹۸.۰۱ | ماشین بردار پشتیبان میانگین | |
| ۵۶.۰۰ | بیزین | [۳۰] |
| ۹۱.۰۰ | درخت تصمیم‌گیری | |
| ۸۸.۰۰ | آنالیز تشخیص خطی | |
| ۸۶.۵۷ | بهینه‌سازی اجتماع ذرات با گسترش توپولوژی همسایگی-نزدیکترین همسایه | [۳۱] |
| ۸۷.۲۱ | بهینه‌سازی اجتماع ذرات با گسترش توپولوژی همسایگی-k نزدیکترین همسایه | |
| ۸۷.۴۵ | بهینه‌سازی اجتماع ذرات با گسترش توپولوژی همسایگی-k نزدیکترین همسایه وزنی | |
| ۸۴.۳۲ | بهینه‌سازی اجتماع ذرات با گسترش توپولوژی همسایگی محلی-نزدیکترین همسایه | |
| ۸۵.۲۱ | بهینه‌سازی اجتماع ذرات با گسترش توپولوژی همسایگی محلی-k نزدیکترین همسایه | |
| ۹۱.۷۵ | الگوریتم جستجوی گرانشی باینری-اطلاعات متقابل | [۳۲] |
| ۹۲.۱۳ | الگوریتم جستجوی گرانشی باینری | |
| ۸۷.۱۹ | بهینه‌سازی اجتماع ذرات-بهینه‌سازی کلونی مورچه | [۳۳] |
| ۸۷.۵۴ | جستجوی حریصانه تصادفی وفقی -بهینه‌سازی جفت‌گیری زنبور عسل | |
| ۷۴.۰۰ | انتخاب ویژگی برمبنای جنگل تصادفی | [۳۴] |
| ۶۴.۰۰ | الگوریتم ژنتیک برمبنای مارکوف | |
| ۹۴.۰۰ | الگوریتم آدابوست-انتخاب ویژگی تصادفی-تفکیک فیشر خطی | [۳۵] |
| ۹۲.۳۰ | الگوریتم آدابوست- تفکیک فیشر خطی | |
| ۹۳.۲۴ | ماشین بردار پشتیبان خطی | [۳۶] |
| ۹۳.۳۷ | ماشین بردار پشتیبان چندجمله‌ای | |

به‌منظور معادل‌سازی مدل پیشنهادی با مدل‌های فرا ابتکاری و داده‌کاوی [۳۷] ذکرشده در جدول (۵)، تعداد تکرار، تعداد جمعیت اولیه و تعداد نسل به ترتیب برابر با ۱۰۰، ۵۰۰ و ۲۰ لحاظ شده‌اند. همچنین تعداد ویژگی‌ها در الگوریتم بهینه‌سازی اجتماع ذرات برابر با ۵۱ ویژگی، الگوریتم بهینه‌سازی کلونی مورچه برابر با ۵۶ ویژگی، الگوریتم ژنتیک برابر با ۵۶ ویژگی و الگوریتم جستجوی ممنوعه برابر با ۳۴ ویژگی می‌باشد. تعداد تکرار در مدل‌های انتخاب ویژگی مبتنی بر الگوریتم ژنتیک [۳۹]، انتخاب ویژگی مبتنی بر یادگیری افزایشی [۳۹]، انتخاب ویژگی مبتنی بر کواریانس تغیر ناپذیر [۳۹] و انتخاب ویژگی مبتنی بر درخت [۳۹] برابر با ۲۰۰ بار تکرار می‌باشد.

جدول ۵: مقایسه مدل پیشنهادی با مدل‌های فرا ابتکاری و داده‌کاوی برمبنای درصد صحت

| درصد صحت | مدل‌ها | منابع |
|---|---|---|
| ۸۷,۱۳ | بهینه‌سازی اجتماع ذرات | [۳۷] |
| ۸۶,۷۸ | بهینه‌سازی کلونی مورچه | |
| ۸۵,۵۹ | الگوریتم ژنتیک | |
| ۸۲,۸۰ | الگوریتم جستجوی ممنوعه | |
| ۸۷,۳۹ | نزدیک‌ترین همسایه بازگشتی کاهشی | [۳۸] |
| ۸۵,۶۶ | نزدیک‌ترین همسایه بازگشتی کاهشی | |
| ۸۶,۶۵ | نزدیک‌ترین همسایه بازگشتی کاهشی | |
| ۸۸,۹۲ | انتخاب ویژگی مبتنی بر الگوریتم ژنتیک | [۳۹] |
| ۸۸,۶۳ | انتخاب ویژگی مبتنی بر یادگیری افزایشی | |
| ۸۹,۸۰ | انتخاب ویژگی مبتنی بر کواریانس تغیر ناپذیر | |
| ۸۹,۶۰ | انتخاب ویژگی مبتنی بر درخت | |
| ۸۹,۵۸ | شبکه عصبی مصنوعی خودسازمان‌ده | [۴۰] |
| ۵۷,۶۹ | انتخاب ویژگی مبتنی بر همبستگی-نیوی بیز | |
| ۵۷,۹۵ | انتخاب ویژگی مبتنی بر تقابل- نیوی بیز | |
| ۹۱,۰۰ | انتخاب ویژگی مبتنی بر سازگاری- نیوی بیز | |
| ۷۶,۵۳ | نرخ اطلاعات با نیوی بیز | |
| ۷۹,۱۴ | انتخاب ویژگی مبتنی بر همبستگی-ماشین بردار پشتیبان | |
| ۷۹,۷۳ | انتخاب ویژگی مبتنی بر تقابل-ماشین بردار پشتیبان | |
| ۸۰,۸۳ | انتخاب ویژگی مبتنی بر سازگاری-k نزدیک‌ترین همسایه | |
| ۷۸,۶۲ | K نزدیک‌ترین همسایه-نرخ اطلاعات | |
| ۹۶,۲۷ | مدل پیشنهادی | |

در جدول(٦)، مقایسه مدل پیشنهادی با مدل ماشین بردار پشتیبان و الگوریتم ژنتیک بر اساس متغیرهایی همچون نوع تابع کرنل و تعداد بردار پشتیبان، برمبنای آموزش و تست نشان داده‌شده است. درصد داده‌های آموزش و تست در مدل پیشنهادی به ترتیب برابر با ۸۰٪ و ۲۰٪ هستند. به دلیل اینکه داده‌های آموزش بیشتر هستند لذا درصد صحت در بخش آموزش بیشتر است.

و از طرفی چون درصد داده‌های تست کمتر هستند لذا درصد صحت کمتر است. جدول(٦)، نشان می‌دهد که درصد صحت در ماشین بردار پشتیبان–الگوریتم ژنتیک در بیشتر موارد بالای ۹۵٪ است. بیشترین مقدار ماشین بردار پشتیبان–الگوریتم ژنتیک در حالت آموزش برابر با ۹۵،۸۱ می‌باشد.

جدول ٦: مقایسه مدل پیشنهادی با مدل ماشین بردار پشتیبان و الگوریتم ژنتیک برمبنای آموزش و تست

| مدل پیشنهادی | | دقت | تعداد بردار پشتیبان | تعداد ویژگی | نوع تابع کرنل | الگوریتم | منابع |
|---|---|---|---|---|---|---|---|
| دقت تست | دقت آموزش | تست | | | | | |
| ۸۹.۳۲ | ۹۶.۲۷ | ۸۶.۲۰ | ۹۹۰ | ۵۷ | چند جمله‌ای | ماشین بردار پشتیبان | [۴۲] |
| ۸۷.۱۱ | ۹۷.۴۶ | ۸۶.۲۰ | ۹۷۲ | ۵۲ | | ماشین بردار پشتیبان-الگوریتم ژنتیک | |
| ۸۹.۱۶ | ۹۶.۲۷ | ۸۷.۲۰ | ۷۷۹ | ۵۷ | | ماشین بردار پشتیبان | |
| ۸۸.۷۲ | ۹۷.۰۶ | ۸۸.۰۶ | ۷۳۶ | ۵۱ | | ماشین بردار پشتیبان-الگوریتم ژنتیک | |
| ۸۷.۵۲ | ۹۶.۴۳ | ۸۶.۸۰ | ۹۹۴ | ۵۷ | | ماشین بردار پشتیبان | |
| ۸۷.۴۶ | ۹۵.۷۹ | ۸۶.۰۰ | ۹۳۸۸ | ۴۹ | | ماشین بردار پشتیبان-الگوریتم ژنتیک | |
| ۸۷.۱۵ | ۹۶.۰۷ | ۸۷.۰۰ | ۹۳۲ | ۵۷ | پایه شعاعی | ماشین بردار پشتیبان | |
| ۸۶.۲۲ | ۹۶.۸۹ | ۸۵.۶۰ | ۸۷۰ | ۵۲ | | ماشین بردار پشتیبان-الگوریتم ژنتیک | |
| ۸۷.۳۴ | ۹۶.۲۱ | ۸۶.۸۰ | ۹۴۶ | ۵۷ | | ماشین بردار پشتیبان | |
| ۸۷.۹۴ | ۹۷.۰۸ | ۸۵.۸۰ | ۸۷۵ | ۵۳ | | ماشین بردار پشتیبان-الگوریتم ژنتیک | |
| ۸۶.۹۸ | ۹۵.۴۶ | ۸۶.۶۰ | ۹۷۹ | ۵۷ | | ماشین بردار پشتیبان | |
| ۸۷.۳۸ | ۹۶.۹۰ | ۸۶.۰۰ | ۹۰۹ | ۴۹ | | ماشین بردار پشتیبان-الگوریتم ژنتیک | |

در جدول(۷)، مقایسه مدل پیشنهادی با مدل‌های مختلف برمبنای درصد صحت نشان داده‌شده است. مدل پیشنهادی در مقایسه با مدل‌های درخت تصمیم C4.5، نیوی بیز، یادگیری برپایه نمونه IB1 و ماشین بردار پشتیبان به‌عنوان طبقه‌بندی و برخی الگوریتم‌های مطرح انتخاب ویژگی، دقت صحت بیشتری دارد. الگوریتم نیوی بیز در مقایسه با فن‌های دیگر، درصد صحت کمتری دارد. اما برمبنای الگوریتم انتخاب ویژگی مبتنی بر سازگاری (معیار فاصله) از برازش بهتری برخوردار است. درصد صحت برمبنای نرخ اطلاعات در درخت تصمیم C4.5 در مقایسه با ماشین بردار پشتیبان بیشتر است.

جدول ۷: مقایسه مدل پیشنهادی با مدل‌های طبقه‌بندی برمبنای درصد صحت

| مدل‌ها/درصد صحت | | | | الگوریتم انتخاب ویژگی | منابع |
|---|---|---|---|---|---|
| ماشین بردار پشتیبان | IB1 | نیوی بیز | درخت C4.5 | | |
| ۹۰.۳۰ | | | | مجموعه راف و الگوریتم جستجوی پراکنده | [۴۳] |
| ۸۵.۸۵ | ۷۹.۱۴ | ۵۷.۶۹ | ۸۱.۱۶ | انتخاب ویژگی مبتنی بر همبستگی | [۴۴] |
| ۸۲.۲۷ | ۷۷.۳۱ | ۵۷.۲۴ | ۷۹.۲۷ | انتخاب ویژگی مبتنی بر همبستگی | |
| ۸۰.۳۱ | ۷۹.۷۳ | ۵۷.۹۵ | ۷۸.۱۶ | انتخاب ویژگی مبتنی بر تقابل | |
| ۸۱.۴۹ | ۷۶.۸۶ | ۷۴.۷۷ | ۸۰.۸۳ | انتخاب ویژگی مبتنی بر تقابل | |
| ۸۱.۸۸ | ۸۰.۸۳ | ۹۱.۵۵ | ۸۴.۶۲ | انتخاب ویژگی مبتنی بر سازگاری | |
| ۸۱.۹۴ | ۷۷.۳۸ | ۹۲.۸۹ | ۷۹.۳۴ | انتخاب ویژگی مبتنی بر سازگاری | |
| ۸۳.۸۳ | ۷۸.۶۲ | ۷۶.۵۳ | ۸۳.۸۳ | انتخاب ویژگی مبتنی بر نرخ اطلاعات | |
| ۸۳.۳۸ | ۷۸.۴۲ | ۸۹.۷۰ | ۸۵.۳۳ | انتخاب ویژگی مبتنی بر نرخ اطلاعات | |
| ۸۳.۱۸ | | | | بهینه‌سازی اجتماع ذرات | [۴۵] |
| ۹۶.۲۷ | | | | مدل پیشنهادی | |

در جدول(۸)، مقایسه مدل پیشنهادی با مدل‌های مختلف برمبنای درصد صحت نشان داده‌شده است. در مدل‌های مورد مقایسه از الگوریتم‌های نیوی بیز، ماشین بردار پشتیبان، k نزدیک‌ترین همسایه، بوستینگ، شبکه عصبی مصنوعی چندلایه، بهینه‌سازی چندبعدی و درخت تصمیم‌گیری به‌عنوان طبقه‌بندی استفاده‌شده و از برخی الگوریتم‌های مطرح انتخاب ویژگی مانند بهینه‌سازی اجتماع ذرات و الگوریتم ژنتیک برای انتخاب ویژگی استفاده گردیده است.

بوستینگ روشی است که برمبنای دستیابی به یک قانون بسیار دقیق جهت پیش‌بینی از طریق ترکیب کردن تعداد زیادی قوانین ضعیف و نادقیق استوار است. در این روش تعدادی دسته‌بند وجود دارند که هرکدام با استفاده از یک مجموعه داده که به‌صورت تصادفی از مجموعه داده اصلی انتخاب می‌شوند آموزش می‌بینند و با رأی‌گیری اکثریت با یکدیگر ترکیب می‌شوند. درواقع در این روش اصولاً توزیع تصادفی که برای یک دسته‌بند استفاده می‌شود به‌گونه‌ای است که در آن احتمال انتخاب داده‌هایی که توسط دسته‌بندی‌های قبلی نادرست برچسب‌گذاری شده‌اند، بیشتر خواهد بود. الگوریتم k نزدیک‌ترین همسایه برمبنای یادگیری توسط نمونه‌های آموزش استوار است. هر نمونه، نماینده یک نقطه در فضای n بعدی می‌باشد. همه نمونه‌های آموزش در یک الگوی n بعدی فضایی ذخیره می‌شوند. زمانی

که یک نمونه ناشناخته داده می‌شود، طبقه‌بندی k نزدیک‌ترین همسایه، به دنبال k نمونه آموزش که به نمونه ناشناخته نزدیک‌ترین هستند الگوی فضایی را جستجو می‌کند. نزدیکی بر اساس فاصله اقلیدسی تعریف می‌شود. پس از یافتن k داده مشابه با نمونه آزمایشی، با رأی اکثریت برچسب نمونه ناشناخته انتخاب می‌شود.

نتیجه حاصله از مقایسه‌های به‌عمل‌آمده مدل‌های ذکرشده با مدل پیشنهادی، نشانگر این نکته است که مدل پیشنهادی درصد صحت بیشتری دارد. طبق نتایج جدول(۸)، الگوریتم بوستینگ و درخت تصمیم‌گیری در مقایسه با الگوریتم‌های دیگر درصد صحت بیشتری دارند. مدل پیشنهادی در مقایسه با الگوریتم‌های دیگر درصد صحت بیشتری دارد. زیرا مدل پیشنهادی از فاکتور وزن دهی به یال‌ها برای انتخاب ویژگی استفاده می‌کند. در مدل پیشنهادی ماتریس وزن دهی به یال‌ها در هر تکرار به‌روزرسانی می‌شود و مورچه‌هایی که یک یال را بیشتر تکرار کنند در هر مرحله مقدار آن یال به‌روزرسانی می‌گردد و این مزیت باعث می‌شود که دقیق‌ترین ویژگی‌ها برای طبقه‌بندی انتخاب شوند.

جدول ۸: مقایسه مدل پیشنهادی با مدل‌های تصمیم‌گیری برمبنای درصد صحت

| منابع | الگوریتم‌های انتخاب ویژگی | الگوریتم‌های طبقه‌بندی | | | | | |
|---|---|---|---|---|---|---|---|
| | | درخت تصمیم گیری | بهینه سازی چندبعدی | پرسپترون چندلایه | بوستینگ | K نزدیکترین همسایه | ماشین بردار پشتیبان | نیوی بیز |
| [٤٦] | انتخاب ویژگی برمبنای تشابه ویژگی | ٧٧.٠٠ | ٨٠.٠٠ | ٧٩.٥٠ | ٦٦.٨٥ | ٨٠.٨١ | ٧٩.٠٠ | ٦٦.٧٠ |
| | انتخاب ویژگی برمبنای امتیاز لاپلاس | ٦٩.١٠ | ٧٩.٣٠ | ٨٠.٦٠ | ٦٩.٢٨ | ٨٢.٦٨ | ٨٣.٨٠ | ٦٩.٣٠ |
| | انتخاب ویژگی برمبنای چندخوشه‌ای | ٧٢.٦٠ | ٧٣.٣٠ | ٦٩.٣٠ | ٦٥.٢٥ | ٨٢.٢٧ | ٨٠.٠٠ | ٦٥.٣٠ |
| | انتخاب ویژگی مبتنی بر گراف | ٦٩.٩٠ | ٧١.٦٠ | ٧٠.٢٠ | ٧٥.٧١ | ٨٤.٣١ | ٨٦.٧٠ | ٧٥.٦٠ |
| | انتخاب ویژگی مبتنی بر همبستگی | ٧٠.١٠ | ٧٢.٦٠ | ٧١.٢٠ | ٧٠.٠٠ | ٧٨.٥٩ | ٧٩.١٠ | ٧٦.٣٠ |
| | انتخاب ویژگی مبتنی بر سازگاری | ٦٥.١٠ | ٦٢.٠٠ | ٦١.٠٠ | ٦٨.٩٥ | ٦٩.٠٣ | ٧٠.٠٠ | ٧٠.٠٠ |
| | بهینه‌سازی اجتماع ذرات | ٧٦.٠٠ | ٧٣.٦٠ | ٧١.٠٠ | ٧٢.٣٥ | ٨١.٠٠ | ٧٩.١٠ | ٧٣.٥٠ |
| | الگوریتم ژنتیک | ٧٠.١٠ | ٧٠.٣٠ | ٧٢.٥٠ | ٦٩.٩٩ | ٦٣.٣٩ | ٦٢.١٠ | ٧٠.٢٠ |
| | انتخاب ویژگی مبتنی بر الگوریتم ژنتیک بهبودیافته | ٩٢.٦٠ | ٨٨.٠٠ | ٩٢.٠٠ | ٩٠.١٨٨ | ٩٢.٢٢ | ٩١.٥٠ | ٨٠.٩٠ |
| | مدل پیشنهادی | ٩٦.٢٧ | | | | | | |

نتایج مدل پیشنهادی نشان داد که تعداد تکرار و تعداد ویژگی‌ها در افزایش درصد صحت مؤثر هستند. با افزایش تعداد تکرار، مدل پیشنهادی می‌تواند مقایسه بین ویژگی را تا حد ممکن انجام دهد و هر نمونه به‌درستی به دسته متعلق به خودش داده شود.

روش پیشنهادی باوجوداینکه زمان اجرای بیشتری دارد اما در مقابل درصد صحت بیشتری دارد. برتری روش پیشنهادی در مقایسه با مدل‌های دیگر این است که الگوریتم آدابوست یک متاالگوریتم هوشمند است؛ که به‌منظور ارتقاء عملکرد و رفع مشکل طبقاتی از قبیل تخصیص نمونه‌ها به دسته‌های اشتباه استفاده می‌شود. در این الگوریتم، دسته‌بندی هر مرحله جدید به نفع نمونه‌های غلط دسته‌بندی‌شده در مراحل قبل تنظیم می‌گردد. با استفاده از الگوریتم بهینه‌سازی کلونی مورچه از میان ویژگی‌های موردنظر، تعدادی از ویژگی‌های مناسب و مرتبط باهدف انتخاب‌شده و از آن‌ها برای آموزش الگوریتم آدابوست استفاده می‌شود. بعد از حصول اطمینان از دقت آموزش الگوریتم آدابوست، ویژگی‌ها دسته‌بندی می‌شوند.

الگوریتم‌های فرا ابتکاری راهکار مناسبی برای انتخاب ویژگی هستند. این الگوریتم‌ها ویژگی‌های مناسب را انتخاب می‌کنند و باعث می‌شوند که دقت طبقه‌بندی افزایش یابد [٤٧،٤٨].

## ۵. نتیجه‌گیری و کارهای آینده

ایمیل هرزنامه، که به‌عنوان ایمیل ناشناس شناخته می‌شود، می‌تواند به شبکه‌های کاربران آسیب برساند، سرورهای ایمیل را مسدود، و صندوق ورودی ایمیل کاربران را با پیام‌های مخرب و ناخواسته پر نماید. در این مقاله مدلی برمبنای الگوریتم بهینه‌سازی کلونی مورچه و آدابوست برای تشخیص ایمیل هرزنامه پیشنهاد شد. نتایج مدل پیشنهادی نشان داد که استفاده از انتخاب ویژگی باعث افزایش دقت تشخیص می‌شود. همچنین مقایسه‌ها نشان داد که مدل پیشنهادی در مقایسه با اکثر مدل‌ها درصد صحت بیشتری دارد. درصد صحت مدل

پیشنهادی با انتخاب همه ویژگی‌ها برابر با ۹۶/۲۷ درصد بود. مدل پیشنهادی در مقایسه با مدل‌هایی مانند ماشین بردار پشتیبان، k نزدیک‌ترین همسایه، درخت تصمیم‌گیری و نیوی بیز، درصد صحت بیشتری دارد. در مقابل، در مقایسه با ماشین بردار پشتیبان میانگین از درصد صحت کمتری بهره‌مند است. همچنین نتایج نشان داد که انتخاب ویژگی با استفاده از الگوریتم بهینه‌سازی کلونی مورچه در مقایسه با بهینه‌سازی اجتماع ذرات، الگوریتم ژنتیک، الگوریتم جستجوی گرانشی و الگوریتم جستجوی ممنوعه از درصد صحت بالاتری برخوردار بوده است. برای کارهای آینده در نظر داریم که از مدل‌های فازی و داده‌کاوی برای تشخیص ایمیل هرزنامه استفاده کنیم. مدل فازی برمبنای قوانین می‌تواند بهترین ویژگی‌ها را انتخاب کند و مدل‌های داده‌کاوی برمبنای تصمیم، نمونه‌ها را طبقه‌بندی کنند.

## مراجع


[1] F.A. Hamzeh, F.S. Gharehchopogh, Feature Selection Based on Harmony Search Algorithm with Naive Bayes Algorithm to Spam Email Detection, Information Technology in Engineering Design, 11(2):31-46, 2019. (Persian)

[2] S. Amjad, F.S. Gharehchopogh, A Novel Hybrid Approach for Email Spam Detection based on Scatter Search Algorithm and K-Nearest Neighbors, Journal of Advances in Computer Engineering and Technology, 5(4): 181-194, 2019

[3] N. Saidani, K. Adi, M.S. Allili, A Semantic-Based Classification Approach for an Enhanced Spam Detection, Computers & SecurityIn press, journal pre-proofAvailable online 9 January 2020.

[4] M.K. Chae; A. Alsadoon, P.W.C. Prasad, A. Elchouemi, Spam filtering email classification (SFECM) using gain and graph mining algorithm, IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC), pp. 1-7, 2017.

[5] P. Cortez, A. Correia, P. Sousa, M. Rocha, M. Rio, Spam Email Filtering Using Network-Level Properties, Industrial Conference on Data Mining, ICDM 2010: Advances in Data Mining. Applications and Theoretical Aspects, pp. 476-489, 2010.

[6] W. Li, N. Zhong, C. Liu, Evaluating the Error Risk of Email Filters Based on ROC Curve Analysis, Communications, and Discoveries from Multidisciplinary Data, Studies in Computational Intelligence, Vol. 23, pp 299-314, 2008.

[7] M.E. Boujnouni, M. Jedra, N. Zahid, Email Spam Filtering Using the Combination of Two Improved Versions of Support Vector Domain Description, International Joint Conference, pp. 99-109, 2015.

[8] V. Cheng, C. Li, Combining Supervised and Semi-supervised Classifier for Personalized Spam Filtering, Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2007: Advances in Knowledge Discovery and Data Mining, pp. 449-456, 2007.

[9] X.H. Pham, N.H. Lee, J.J. Jung, A.S. Niaraki, Collaborative spam filtering based on incremental ontology learning, Telecommunication Systems, Vol. 52, Issue 2, pp. 693-700, 2013.

[10] M.Dorigo, L.M. Gambardella, The colony system: A cooperative learning approach to the traveling salesman problem, IEEE Transactions on Evolutionary Computation, Vol. 1, No.1, pp. 53-66, April 1997.

[11] Y. Freund and R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences, 5(1), pp. 119-139, 1997.

[12] N.O.F. Elssied, O. Ibrahim, W. Abu-Ulbeh, an improved of spam E-mail classification mechanism using K-means clustering, Journal of Theoretical and Applied Information Technology, Vol. 60, No.3, pp. 568-580, 2014.

[13] M. Awad, and M. Foqaha, Email Spam Classification Using Hybrid Approach of RBF Neural Network and Particle Swarm Optimization, International Journal of Network Security & Its Applications (IJNSA) Vol.8, No.4, pp. 17-28, July 2016.

[14] W. Ma, D. Tran, and D. Sharma, A Novel Spam Email Detection System Based on Negative Selection, 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology, IEEE, 2009.

[15] P. Sao, K. Prashanthi, E-mail Spam Classification Using Naïve Bayesian Classifier, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Vol. 4 Issue 6, pp. 2792-2796, June 2015.
[16] I. Idris, E-mail Spam Classification with Artificial Neural Network and Negative Selection Algorithm, International Journal of Computer Science & Communication Networks, Vol.1, No. 3, pp. 227-231, 2010.
[17] J. Shapoorjee, K. Nagar1, C. Bhadane, Comparative Study of Data Mining Techniques for Classifying Spam Email, International Journal of Emerging Technology and Advanced Engineering, Vol. 5, Issue 10, pp. 132-135, 2015.
[18] A. Sharaff, N.K. Nagwani, A. Dhadse, Comparative Study of Classification Algorithms for Spam Email Detection, Emerging Research in Computing, Information, Communication and Applications, pp. 237-244, 2015.
[19] N.O.F. Elssied and O. Ibrahim, K-Means Clustering Scheme for Enhanced Spam Detection, Research Journal of Applied Sciences, Engineering and Technology 7(10): 1940-1952, 2014
[20] A.R. Behjat, A. Mustapha, H. Nezamabadi-pour, Md. Nasir Sulaiman, and N. Mustapha, A PSO-Based Feature Subset Selection for Application of Spam /Non-spam Detection, Springer-Verlag Berlin Heidelberg, M-CAIT 2013, CCIS 378, pp. 183-193, 2013
[21] Kuo-Ching Ying, Shih-Wei Lin, Zne-Jung Lee, Yen-Tim Lin, An ensemble approach applied to classify spam e-mails, Expert Systems with Applications, Vol. 37, Issue 3, pp. 2197-2201, 2010.
[22] H. He, A. Tiwari, J. Mehnen, T. Watson, C. Maple, Y. Jin, B. Gabrys, Incremental information gain analysis of input attribute impact on RBF-kernel SVM spam detection, IEEE Congress on Evolutionary Computation (CEC), pp. 1022-1029, 2016.
[23] N.K. Sreeja, A. Sankar, Pattern Matching based Classification using Ant Colony Optimization based Feature Selection, Applied Soft Computing, Vol. 31, pp. 91-102, 2015.
[24] A. Huang, Similarity measures for text document clustering. In: Proceedings of the Sixth New Zealand Computer Science Research Student Conference, pp. 49-56, 2008.
[25] Y. Wan, M. Wang, Z. Ye, X. Lai, A feature selection method based on modified binary coded ant colony optimization algorithm, Applied Soft Computing, Vol. 49, pp. 248-258, 2016.
[26] Z. Ye, W. Liu, H. Chen, E. Zhao, A Novel Feature Selection Approach Based on Swarm Intelligence, International Workshop on Intelligent Systems and Applications, IEEE, 2009.
[27] R.S. Michalski, I. Bratko, M. Kubat, Machine Learning, and Data Mining: Methods and Applications, New York: Wiley, 1998.
[28] Spambase Data Set, https://archive.ics.uci.edu/ml/datasets/spambase , [Last Available: Oct 2019]
[29] V.B. Canedo, N.S. Marono, and A.A. Betanzos, A Distributed Feature Selection Approach Based on a Complexity Measure, International Work-Conference on Artificial Neural Networks, IWANN 2015: Advances in Computational Intelligence, pp. 15-28, 2015.
[30] S. Cateni, V. Colla, M. Vannucci, A Fuzzy System for Combining Filter Features Selection Methods, International Journal of Fuzzy Systems, Vol. 19, Issue 4, pp. 1168-1180, 2017.
[31] Y. Marinakis and M. Marinaki, A Hybridized Particle Swarm Optimization with Expanding Neighborhood Topology for the Feature Selection Problem, International Workshop on Hybrid Metaheuristics, HM 2013: Hybrid Metaheuristics, pp. 37-51, 2013.
[32] H. Bostani, M. Sheikhan, Hybrid of binary gravitational search algorithm and mutual information for feature selection in intrusion detection systems, Soft Computing, Vol. 21, Issue 9, pp 2307-2324, 2017.
[33] Y. Marinakis, M. Marinaki, and N. Matsatsinis, A Hybrid Clustering Algorithm Based on Multi-swarm Constriction PSO and GRASP, International Conference on Data Warehousing and Knowledge Discovery, DaWaK 2008: Data Warehousing and Knowledge Discovery pp. 186-195, 2008.
[34] A. Noura, H. Shili, and L.B. Romdhane, Reliable Attribute Selection Based on Random Forest (RASER), International Conference on Intelligent Systems Design and Applications, ISDA 2016: Intelligent Systems Design and Applications, pp. 11-24, 2016.
[35] T. Arodz, Boosting the Fisher Linear Discriminant with Random Feature Subsets, Computer Recognition Systems, pp. 79-86, 2005.
[36] S. Maldonado and G. L'Huillier, SVM-Based Feature Selection and Classification for Email Filtering, Pattern Recognition, Applications, and Methods pp. 135-148, 2013.
[37] Y. Marinakis, M. Marinaki, and N. Matsatsinis, A Hybrid Particle Swarm Optimization Algorithm for Clustering Analysis, International Conference on Data Warehousing and Knowledge Discovery, DaWaK 2007: Data Warehousing and Knowledge Discovery, pp. 241-250, 2007.

[38] B.R. Dai and S.M. Hsu, An Instance Selection Algorithm Based on Reverse Nearest Neighbor, Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2011: Advances in Knowledge Discovery and Data Mining pp. 1-12, 2011.
[39] L. Inza, P. Larraiiaga, B. Sierra, Feature Subset Selection by Estimation of Distribution Algorithms, Estimation of Distribution Algorithms, pp. 269-293, 2002.
[40] N. Grozavu, Y. Bennani, M. Lebbah, Cluster-Dependent Feature Selection through a Weighted Learning Paradigm, Advances in Knowledge Discovery and Management, pp. 133-147, 2010.
[41] L.M. Fernandez, V.B. Canedo, A.A. Betanzos, Centralized vs. distributed feature selection methods based on data complexity measures, Knowledge-Based Systems, Vol. 117, pp. 27-45, 2017.
[42] Y. Ying, W. Xiaolong, L. Bingquan, Combining feature scaling estimation with SVM classifier design using GA approach, Journal of Electronics, Vol. 22, Issue 5, pp. 550-557, 2005.
[43] M. Mohamad and A. Selamat, A New Hybrid Rough Set and Soft Set Parameter Reduction Method for Spam E-Mail Classification Task, Pacific Rim Knowledge Acquisition Workshop, PKAW 2016: Knowledge Management and Acquisition for Intelligent Systems, pp. 18-30, 2016.
[44] V.B. Canedo, N.S. Marono, and J.C. Rabunal, Scaling Up Feature Selection: A Distributed Filter Approach, Conference of the Spanish Association for Artificial Intelligence, CAEPIA 2013: Advances in Artificial Intelligence, pp. 121-130, 2013.
[45] M. Marinaki, Y. Marinakis, A hybridization of clonal selection algorithm with iterated local search and variable neighborhood search for the feature selection problem, Memetic Computing, Vol. 7, Issue 3, pp. 181-201, 2015.
[46] B.S. Pardo, I.P. Diaz, V.B. Canedo, A.A. Betanzos, Ensemble feature selection: Homogeneous and heterogeneous approaches, Knowledge-Based Systems, Vol. 118, pp. 124-139, 2017.
[47] F.S. Gharehchopogh, H. Gholizadeh, A comprehensive survey: Whale Optimization Algorithm and its applications, Swarm and Evolutionary Computation, 47: 1-24, 2019
[48] F.S. Gharehchopogh, H. Shayanfar, H. Gholizadeh, A comprehensive survey on symbiotic organisms search algorithms, Artificial Intelligence Review, Springer Netherlands, pp. 1-48, 2019.