چهارمین کنفرانس انجمن رمز ایران
۲۴ الی ۲۶ مهر ماه ۱۳۸۶ دانشگاه علم و صنعت ایران
انجمن رمز ایران
Iranian Society of Cryptology

جداسازی هرز نامه‌های متنی

یک رویکرد مبتنی بر الگوریتم ژنتیک و روش دسته‌بندی SVM

سعید جلیلی
دانشگاه تربیت مدرس، گروه کامپیوتر
آزمایشگاه یادگیری ماشین
sjalili@modares.ac.ir

شیما گرانی
دانشگاه تربیت مدرس، گروه کامپیوتر
آزمایشگاه یادگیری ماشین
gerani@modares.ac.ir

چکیده: در این مقاله، یک روش ترکیبی الگوریتم ژنتیک برای انتخاب خصیصه و روش دسته‌بندی svm برای جداسازی نامه‌های هرز پیشنهاد شده‌است. روش پیشنهادی روی مجموعه نامه‌های استاندارد LingSpam ارزیابی شده‌است. نتایج حاصل از ارزیابی نشان می‌دهد که روش پیشنهادی علاوه بر حفظ و یا بالا بردن معیارهای دقت، فراخوانی و F1 ، تعداد خصیصه‌ها را حدوداً به ۰/۱ تعداد اولیه کاهش می‌دهد. همچنین، مقایسه نتایج حاصل از میانگین دقت، فراخوانی و F1 دسته‌بندی هرزنامه با روش روش پیشنهادی با روش‌هایی که از SVM به همراه یک روش فیلتری انتخاب خصیصه استفاده می‌کنند و سایر روش‌های آماری جداسازی هرزنامه‌ها نشان می‌دهد که روش پیشنهادی از نظر دقت و فراخوانی قابل مقایسه و حتی در مواردی بهتر بوده است.

واژه های کلیدی: هرزنامه، یادگیری ماشین، الگوریتم ژنتیک، SVM، انتخاب خصیصه.

۱- مقدمه

امروزه نامه الکترونیکی یکی از ابزارهای مهم و پراستفاده برای ارتباطات بین مردم است. محبوبیت روز افزون و کمی هزینه نامه‌الکترونیکی این زمینه را فراهم آورده‌است که افرادی اقدام به ارسال نامه‌های الکترونیکی نامربوط در حجم انبوه کنند. این نامه‌ها به اصطلاح هرزنامه، (Spam) یا UCE[1] نامیده می‌شوند. هرزنامه‌ها باعث اتلاف وقت، اشغال منابع، اتلاف پهنای باند و درنتیجه طولانی شدن زمان ارتباط می‌شوند. بررسی‌های انجام شده نشان داده‌است که امروزه در حدود ۷۰٪ نامه‌های کاری الکترونیکی، هرزنامه هستند[۱].

تا کنون تعدادی فیلتر ضد هرزنامه با دو رویکرد مختلف عرضه شده‌اند: در رویکرد اول که بدون بهره‌گیری از یادگیری

[1] Unsolicited Commercial Email

ماشین(non-ML) است، روش‌هایی مانند روش‌های شهودی[۲]، لیست سیاه[۳]، امضاء[۴]، بر اساس هش[۵] و آنالیز ترافیک ارایه شده‌اند. رویکرد دوم مبتنی بر یادگیری ماشین (ML[۶]) است. در این رویکرد، به یک مجموعه آموزش(حاوی نامه‌های الکترونیکی عادی و هرز) نیاز است. هریک از روش‌های یادگیری ماشین، از روی داده‌های مجموعه آموزشی داده شده، یک دسته‌بند یاد می‌گیرند. سپس دسته‌بند تولید شده، برای تعیین نوع نامه‌های الکترونیکی موجود در مجموعه آزمون(نامه‌های الکترونیکی که از قبل نوع آنها تعیین شده است وبا نامه‌های مجموعه آموزش متفاوت هستند.) به‌کار گرفته شده و کارای دسته‌بند، اندازه‌گیری می‌شود. موفقیت روش‌های یادگیری ماشین در دسته‌بندی متون[۱۴]، محققین را به سمت استفاده از این روش‌ها در فیلترسازی هرزنامه‌ها هدایت کرده‌است. در واقع فیلترسازی هرز نامه‌های هرز بر اساس محتوای متنی آنها، حالت خاصی از دسته‌بندی متون محسوب می‌شود که در آن دو کلاس هرزنامه و نامه‌های مفید مدنظر هستند[۸]. مسأله جداسازی هرزنامه‌ها به نحوی که هیچ نامه عادی به اشتباه در مجموعه نامه‌های هرز قرار داده نشود، اهمیت زیادی دارد، و از آن به عنوان یک معیار مهم در ارزیابی کارایی فیلترهای هرزنامه‌ها استفاده می‌شود. یادآوری می‌گردد هزینه به اشتباه بلاک نمودن یک نامه عادی به عنوان نامه هرز، بیشتر از هزینه آن است که یک نامه هرز از فیلتر عبور نماید. در نتیجه، این تفاوت باید در هر دو مرحله یادگیری و آزمون، در نظر گرفته شود[۵].

در این مقاله، یک روش جداسازی نامه‌های هرز از نامه‌های عادی با رویکرد یادگیری ماشین ارایه شده‌است. این روش از الگوریتم ژنتیک برای انتخاب خصیصه و روش دسته‌بندی SVM برای یادگیری دسته‌بند استفاده می‌کند. برای سنجش کارایی، روش پیشنهادی بر روی مجموعه داده LingSpam به‌کار رفته است. به‌منظور مقایسه، از روش SVM به‌تنهایی نیز برای دسته‌بندی همان مجموعه داده، استفاده شده‌است.

ادامه این مقاله به‌این ترتیب سازماندهی شده‌است: در بخش ۲ پژوهش‌های مرتبط توضیح داده شده است. بخش ۳ اطلاعات پیش‌زمینه مورد نیاز را بیان می‌کند. در بخش ۴ روش پیشنهادی ارایه می‌شود. سپس در بخش ۵ شرایط و پارامترهای آزمایش و نتایج ارزیابی روش پیشنهادی بیان می‌شود. در نهایت در بخش ۶ نتیجه‌گیری و زمینه‌های پژوهش آینده بحث می‌شود.

۲- پژوهش‌های مرتبط

از جمله روش‌های فیلترسازی هرزنامه‌ها با رویکرد یادگیری ماشین می‌توان به روش‌های یادگیری قانون[۷][۲-۳]، درخت تصمیم[۴]، بیزین ساده[۵-۶]، SVM[8] [۷-۸] و یا ترکیب دسته‌بندهای مختلف[۱۰] اشاره نمود.

یکی از متداول‌ترین روش‌ها در زمینه دسته‌بندی متون و نامه‌های الکترونیکی، روش بیزین می‌باشد. این روش، یادگیری و دسته‌بندی سریعی دارد و به راحتی امکان یادگیری افزایشی دارد. Sahami و همکارانش[۱۱] از روش دسته‌بندی بیزین[۹] برای جداسازی نامه‌های هرز استفاده کردند. Androutsopoulos و همکارانش[۴-۶] در مجموعه مقالات خود، فیلتر بیزین ساده پیشنهاد شده توسط Sahami و همکارانش را توسعه داده و تأثیر تعداد متفاوت خصیصه و اندازه‌های متفاوت مجموعه آموزش روی کارایی فیلتر را بررسی نمودند. در عین حال، کارایی روش بیزین ساده با روش بر مبنای حافظه مقایسه شده و نتایج نشان داده‌است که هر دو روش نسبت به روش فیلتر بر مبنای کلمه‌کلیدی[۱۰] کارایی بیشتری دارند. Pantel و Lin نیزفیلتری به نام SpamCop ارایه کرده‌اند. این فیلتر نوعی فیلتر بیزین محسوب می‌شود. تنها تفاوت آن با روش بیزین ساده در تعریف نشانه‌های آماری است و همین کارایی آن را به مقدار کمی افزایش می‌دهد[۲۶].

[2] heuristics
[3] blacklisting
[4] signatures
[5] Hash based
[6] Machine Learning

[7] Rule learning
[8] Support Vector Machine
[9] Bayesian
[10] Keyword-based

حبیبی و کفائی[۲۶]، نوعی فیلتر بیزین ارایه داده‌اند که در ساختار پردازش و پیش‌پردازشی آن، تغییراتی انجام شده است. برای مثال تمامی اطلاعات نظیر کاراکترها و علایم متن اصلی، برچسب‌ها و کدهای html، سرنوشته‌ها، موضوع ، آدرس‌ها، تصاویر و غیره بررسی می‌شود. این مسأله نرخ جداسازی صحیح الگوریتم را افزایش داده است.

Drucker و همکارانش[۷] از SVM برای دسته‌بندی نامه‌ها با توجه به محتوایشان استفاده کردند. سپس کارایی روش را با روش‌های Ripper، Rocchio، و BDT[11] مقایسه نمودند. روش‌های BDT و SVM کارایی مورد قبولی از نظر صحت[12] و سرعت در آزمون‌ها نشان دادند. البته زمان یادگیری BDT بیش از حد طولانی می‌باشد. Woitaszek و همکارانش[۱۲] نیز از یک SVM ساده و یک دیکشنری شخصی برای تعیین نامه‌های الکترونیکی تجاری استفاده نمودند.

با وجود آنکه روش نزدیک‌ترین همسایه از جمله مباحث مطرح در یادگیری ماشین می‌باشد، در زمینه فیلترسازی هرزنامه‌ها، به ندرت استفاده شده‌است. Trudging و Yang [۱۳] این روش را برای دسته‌بندی نامه‌های متنی به‌کار برده و با سایر روش‌ها مقایسه کرده‌اند.

در کنار روش‌های آماری ذکر شده، روش‌های تکاملی نیز برای دسته‌بندی هرزنامه‌ها استفاده شده‌اند که در ادامه توضیح داده می‌شوند:

Stoean و همکارانش[۱۸] روشی برای فیلترسازی نامه‌های هرز با روش تکاملی خاصی به نام GC [۱۹] ارایه داده‌اند. مدل ارایه شده به منظور حل مسأله فیلترسازی نامه‌های هرز مورد استفاده قرارگرفته‌است.

در زمینه انتخاب وکاهش خصیصه به منظور دسته‌بندی متون، روش‌های آماری فراوانی ارایه شده است[۲۵و ۲۰]. Koprinska و همکارانش در[۲۰]، روشی برای انتخاب خصیصه به نام TFV ارایه کرده‌اند. این روش نسبت به IG که از قوی‌ترین و متداول‌ترین روش‌های انتخاب خصیصه به‌شمار می‌رود، به همراه SVM کارایی بهتری در تفکیک نامه‌ها نشان داده‌است.

به‌منظور انتخاب خصیصه متون با روش‌های تکاملی، Castillo و Serrano، در [۱۶] یک روش روکشی بر اساس GA ، داده‌اند. در این روش از چندین معیار آماری برای اختصاص رتبه به خصیصه‌ها (کلمات) استفاده شده است.

در زمینه انتخاب خصیصه برای دسته‌بندی نامه‌های هرز با استفاده از روش‌های تکاملی کاری مشاهده نشده‌است.

۳- پیش زمینه

۳-۱- پیش‌پردازش

برای استفاده از نامه‌ها جهت یادگیری، باید آن‌ها را به شکل مناسبی نمایش داد. به همین دلیل، ابتدا مراحل پیش پردازش و کاهش خصیصه‌ها، روی نامه‌ها انجام گرفته و سپس دسته‌بند ساخته می‌شود.

در مرحله پیش‌پردازش معمولاً سه عمل حذف کلمات زاید[۱۳]، حذف برچسب‌ها و ریشه‌یابی روی کلمات اسناد صورت می‌گیرد. کلمات زاید کلمات معمولی هستند که حاوی اطلاعات چندانی نمی‌باشند، به علاوه در تمامی متون به تعداد زیاد وجود داشته و تأثیری در متمایز ساختن متن نسبت به سایر متون ندارند، مانند حروف ربط و حروف اضافه. برچسب ها نیز مانند برچسب های HTML و XML می باشند، که همه آنها از محتوای نامه حذف می‌گردند. البته در نظر گرفتن برچسب‌ها در دسته‌بندی و جداسازی هرز نامه‌ها می‌تواند بسیار مفید باشد. زیرا اکثر نامه‌های هرز شامل برچسب‌های مشابهی هستند. اما در آزمایشات انجام شده در این مقاله، این اطلاعات در نظر گرفته نشده‌اند. در ریشه‌یابی نیز، به جای مشتقات مختلف یک کلمه، تقریباً ریشه کلمه آورده می‌شود که بدین ترتیب تعداد کلمات سند تا حدودی کاهش می‌یابد. البته انجام ریشه‌یابی کلمات ضروری نمی‌باشد. الگوریتم‌های متعددی برای این عمل وجود دارد، که از این میان می‌توان به ریشه‌گیر porter و lemmatizer اشاره نمود.

در مرحله شاخص‌گذاری، در واقع شیوه نمایش نامه تعیین می‌گردد. معمول‌ترین شیوه، استفاده از مدل فضای برداری است که یک روش غیرمعنایی می‌باشد. یعنی هر نامه را می‌توان به-

[11] Boosting Decision Trees
[12] Accuracy
[13] StopWord

صورت یک بردار نشان داد که هر عنصر آن، حضور یا عدم حضور یک کلمه یا تعداد وزن آن کلمه در نامه می‌باشد.

روش‌های متعددی برای وزن‌دهی به کلمات یک سند بر اساس تکرار آن‌ها وجود دارد. که عبارتند از: وزن‌دهی دودویی، وزن‌دهی با تکرار کلمات، وزن‌دهی TFIDF، وزن‌دهی TFC، وزن‌دهی LTC و وزن‌دهی آنتروپی[۱۴،۲۱،۲۲].

مرحله بعدی، کاهش ابعاد بردار متناظر با یک نامه است. چون در دسته‌بندی نامه‌ها، تعداد خصیصه‌ها، به عبارت دیگر تعداد کلمات نامه فراوان است، ابعاد مسأله بسیار بزرگ خواهد شد و کارکردن روی آن زمان‌بر و پرهزینه است. به همین دلیل برای کاهش ابعاد بردارهای نامه، می‌توان از روش‌های متعدد انتخاب خصیصه متون استفاده کرد[۲۵]. در روش‌های انتخاب خصیصه سعی بر این است که بهترین خصیصه‌ها انتخاب شوند طوری که با حذف سایر خصیصه‌ها، دقت دسته‌بندی تغییر چندانی نداشته باشد. روش‌های انتخاب خصیصه به دو دسته کلی روش‌های روکشی و روش‌های تصفیه تقسیم می‌شوند. در روش‌های روکشی، زیر مجموعه خصیصه‌ها با توجه به خطای دسته‌بندی برای دسته‌بند موجود، بدست می‌آید. در حالی‌که در روش‌های تصفیه، خصیصه‌ها را بدون استفاده از دسته‌بند، انتخاب می‌کنند. روش‌های تصفیه ساده و کم هزینه هستند. به همین علت بیشتر از این روش‌ها استفاده می‌شود. انواع روش‌های تصفیه عبارتند از: روش آستانه‌یابی تکرار سند(DF)، روش بهره اطلاعاتی(IG)، روش اطلاعات متقابل(MI) ، روش CHI، روش ضریب همبستگی و روش SCHI [۱۴].

۳-۲- الگوریتم ژنتیک

در الگوریتم ژنتیک، هر یک از راه حل‌های احتمالی مسأله، به صورت یک کروموزوم نمایش داده می‌شود. سپس، برای یافتن راه حل بهینه، طی چندین نسل و بر اساس قوانین انتخاب طبیعی[14] و ابقاء برازنده‌ترین[15]، جمعیت کروموزوم‌ها توسط عملگرهای ژنتیک، نظیر انتخاب، آمیزش و جهش تکامل می‌یابند و راه حل‌های احتمالی، توسط تابع برازندگی ارزیابی می‌شوند. از جمله مزایای الگوریتم ژنتیک، آن است که برخلاف بیشتر الگوریتم‌های یادگیری سنتی، جستجوی سراسری انجام می‌دهند. در حالی‌که معمولاً از نظر محاسباتی، پیچیدگی بالایی دارند.

۴- روش GA+SVM:

روش پیشنهادی شامل دو بخش انتخاب خصیصه و آموزش دسته‌بند می‌باشد. این روش برای انتخاب خصیصه، از روش روکشی و الگوریتم ژنتیک استفاده می‌کند. برای آموزش دسته‌بند نیز از روش SVM استفاده می‌کند.

در الگوریتم ژنتیک ارایه شده در مرحله انتخاب خصیصه، طول کروموزوم‌ها برابر با تعداد کلمات متفاوت مجموعه آموزش در نظر گرفته می‌شود. ژن‌ها از نوع دودویی بوده و True یا False بودن آنها نشان‌دهنده انتخاب یا عدم انتخاب خصیصه مربوطه است. در نتیجه هر کروموزوم، می‌تواند نشان‌دهنده یک مجموعه خصیصه منتخب است. برای ارزیابی هر کروموزوم، ابتدا نامه‌های مجموعه آموزش و آزمون با توجه به خصیصه‌های انتخاب شده توسط آن کروموزوم و با روش وزن‌دهی TFIDF نمایش داده می‌شوند. در مرحله بعد، دسته‌بند SVM از روی این داده‌ها آموزش داده می‌شود. سپس دسته‌بند آموزش دیده شده، برای دسته‌بندی مجموعه آزمون به‌کار می‌رود. نتایج حاصل از دسته‌بندی مجموعه آزمون، مطابق الگوریتم ارایه شده در شکل ۱، در محاسبه برازندگی کروموزوم، به‌کار گرفته می‌شود:

```
If ((SR > R0) && (SP ==1)) {
  Fitness = (SR + (F#unSelected / F#all)) / 2;
} else if (SR > R0 && SP >P0) {
Fitness=(SR-(2*(1 - SP))+( F#unSelected / F#all))/4;
} else if (SP ==1) {
  Fitness = (SR + (F#unSelected / F#all))/2;
} else {
  Fitness = 0 ;}
```

در عبارات بالا، F#all نشان‌دهنده تعداد کل خصیصه‌ها؛ یعنی تعداد کلمات غیر تکراری موجود در مجموعه نامه‌های آموزش است. F#unSelected نیز بیانگر تعداد خصیصه‌های انتخاب نشده است که در حقیقت زیر مجموعه‌ای از F#all هستند. SP و SR مطابق رابطه‌های (۱) و (۲) بدست می‌آیند. Ro و Po به

[14] Natural selection
[15] Survival of the fittest

ترتیب نشان‌دهنده دقت و فراخوانی دسته‌بند SVM به تنهایی، یعنی بدون استفاده از GA و با F#all خصیصه هستند.

در ساخت دسته‌بند هرزنامه، توجه به این مطلب ضروری است که هزینه دسته‌بندی اشتباه یک نامه عادی به عنوان هرز خیلی بیشتر از هزینه دسته‌بندی اشتباه یک نامه هرز به عنوان عادی است. در نتیجه می‌توان گفت بالاتر بودن مقدار دقت(SP)، از اهمیت بیشتری نسبت به مقدار فراخوانی برخوردار است. در واقع چنانچه مقدار SP یک باشد، به این معنی است که هیچ نامه عادیی در دسته نامه‌های هرز قرار نگرفته است.

در دسته‌بند پیشنهادی، هدف بالا بردن دقت تا اندازه ۱، افزایش فراخوانی تا حد ممکن، و در عین حال کاهش خصیصه است. در شرایطی که افزایش دقت و فراخوانی ممکن نباشد، حفظ دقت و فراخوانی اولیه و کم کردن تعداد خصیصه‌ها نیز مناسب است. تابع برازندگی برای ارزیابی هر کرومزوم ۴ حالت را در نظر می‌گیرد:

حالت اول(مطلوب‌ترین): زمانی که دقت دسته‌بند برابر یک بوده و فراخوانی از فراخوانی اولیه(با کل خصیصه‌ها) بالاتر است. در این صورت، برازندگی کرومزوم مربوطه برابر با مجموع فراخوانی فعلی و نسبت تعداد خصیصه‌های انتخاب نشده به تعداد کل خصیصه‌های در نظر گرفته می‌شود. برای نرمال سازی، این مجموع بر ۲ تقسیم می‌شود. زیرا فراخوانی و نسبت خصیصه‌های انتخاب نشده به کل خصیصه‌ها، هر دو بین ۰ تا ۱ تغییر می‌کنند. در اینجا هدف ماکزیمم نمودن برازندگی کرومزوم‌هاست. همچنین واضح است که کاهش تعداد خصیصه‌های انتخاب شده به معنی افزایش تعداد خصیصه‌های انتخاب نشده است. در نتیجه در تابع برازندگی تعداد خصیصه‌های انتخاب نشده در نظر گرفته شده است.

حالت دوم: زمانی که دقت دسته‌بند یک نباشد، اما هم دقت و هم فراخوانی، از حد اولیه خود بالاتر باشند. در این صورت، برازندگی کرومزوم مربوطه برابر با مجموع فراخوانی فعلی و نسبت تعداد خصیصه‌های انتخاب نشده به تعداد کل خصیصه‌ها در نظر گرفته می‌شود. در این رابطه، به اندازه دو برابر فاصله دقت از مقدار یک، به عنوان جریمه، از فراخوانی کسر می‌شود. برای نرمال سازی، این مجموع بر ۴ تقسیم می‌شود.

حالت سوم: زمانی که دقت دسته‌بند یک است اما فراخوانی از مقدار اولیه کمتر شده است. با وجود آنکه فراخوانی از مقدار اولیه کمتر شده و این مطلوب نمی‌باشد، اما به علت یک بودن دقت، در این حالت برازندگی صفر داده نمی‌شود. به این امید که کرومزوم بتواند در نسل‌های بعد، فراخوانی را نیز افزایش دهد. در این حالت، برازندگی برابر با مجموع فراخوانی فعلی و نسبت تعداد خصیصه‌های انتخاب نشده به تعداد کل خصیصه‌ها در نظر گرفته می‌شود. در این وضعیت نیز برای نرمال‌سازی مجموع بر ۲ تقسیم می‌شود. با وجود یکسان بودن رابطه اول و سوم، از آنجا که در این حالت فراخوانی کمتر از فراخوانی اولیه است و در حالت اول فراخوانی بیش از فراخوانی اولیه است، کرومزوم‌های با حالت سوم، برازندگی کمتری نسبت به کرومزوم‌های حالت اول خواهند داشت.

حالت چهارم: زمانی که دقت و فراخوانی، هر دو از مقدار اولیه کمتر شده‌اند. در این حالت برازندگی صفر در نظر گرفته می‌شود.

۵- ارزیابی روش پیشنهادی

۵-۱- مجموعه هرزنامه

مجموعه هرزنامه در واقع مجموعه‌ای از نامه‌های الکترونیکی می‌باشد که شامل نامه‌های عادی و هرزنامه است. چندین مجموعه استاندارد برای آزمایشات روی نامه‌های الکترونیکی موجود است که پرکاربردترین آن‌ها، LingSpam نام دارد[۲۳]. این مجموعه شامل ۲۸۹۳ نامه است که به صورت زیر تقسیم می‌شوند:

- ۲۴۱۲ نامه عادی که به‌طور تصادفی از آرشیو نامه‌های الکترونیکی گرفته شده‌اند.
- ۴۸۱ هرزنامه که هرکدام توسط یک نفر دریافت شده‌اند.

از معروف‌ترین مجموعه‌های دیگر می‌توان به PU1 [۲۳] و Spambase اشاره‌کرد. PU1 حاوی ۱۰۹۹ نامه است که شامل ۶۱۸ نامه مفید و ۴۸۱ هرزنامه می‌باشد. تعداد اندک نامه‌ها در این مجموعه دقت تشخیص هرزنامه‌ها را پایین می‌آورد و در نتیجه کاربرد محدودی دارد.

۵-۲- روش ارزیابی K-fold cross validation

در این روش، مجموعه نامه‌ها به K بخش تقسیم می‌شود. سپس فرایند یادگیری دسته‌بند و آزمون آن K بار تکرار شده و هر بار K-1 بخش برای یادگیری و یک بخش برای آزمون استفاده می‌شود. نتیجه کلی ارزیابی کارایی، میانگین K تکرار خواهد بود[۱۴].

۵-۳- معیارهای ارزیابی کارایی

وقتی کارایی فیلترهای هرزنامه را بررسی می‌کنیم، چهار کمیت برای هر دسته مورد توجه قرار می‌گیرد:

$n_{S\to S}$: تعداد نامه‌های هرز که هرز شناخته شده‌اند.

$n_{S\to L}$: تعداد نامه‌های هرز که عادی شناخته شده‌اند.

$n_{L\to L}$: تعداد نامه‌های عادی که عادی شناخته شده‌اند.

$n_{L\to S}$: تعداد نامه‌های عادی که هرز شناخته شده‌اند.

با استفاده از این کمیت‌ها، معیارهای دقت(SP[16])، فراخوانی (SR[17])، و SF1[18] تعریف می‌شوند:

$$SP = \frac{n_{S\to S}}{n_{S\to S} + n_{L\to S}} \qquad (۱)$$

$$SR = \frac{n_{S\to S}}{n_{S\to S} + n_{S\to L}} \qquad (۲)$$

$$SF1 = \frac{2\times SP\times SR}{SP + SR} \qquad (۳)$$

۵-۴- ارزیابی کارایی روش پیشنهادی

برای ارزیابی، روش پیشنهادی روی مجموعه نامه‌های الکترونیکی LingSpam آزمون شده است. این مجموعه در چهار مدل، با توجه به اینکه ریشه‌یابی و یا حذف StopWords انجام شده‌است یا نه، موجود است. در این مقاله، نتایج بر روی مدل lemm(فقط ریشه‌گیری شده) انجام شده‌است، زیرا بطور معمول از این مدل استفاده می‌شود. همچنین برای وزن‌دهی خصیصه‌ها از روش TFIDF استفاده شده‌است.

در الگوریتم ژنتیک پیشنهادی، تعداد ۵۰ کرومزوم به عنوان جمعیت اولیه در نظر گرفته شده است. همچنین، از الگوریتم انتخاب مسابقه‌ای۱۹ استفاده شده است. تعداد عمل آمیزش در هر بار تکرار ۲۵ مرتبه در نظر گرفته شده‌است. احتمال جهش نیز ۰/۰۱ در نظر گرفته شده است. شرط پایان الگوریتم، یکسان بودن برازندگی بهترین کرومزوم در طی ۵۰ نسل متوالی است. همچنین برای سنجش دقیقتر کارایی روش پیشنهادی، از روش 10-fold cross validation استفاده شده‌است. روش پیشنهادی (GA+SVM) بر روی هر بخش از ۱۰ بخش نامه‌های LingSpam، ۳ بار اجرا شده و بهترین نتیجه برای هر بخش به تفکیک در جدول (۱) آورده شده است. نتایج نشان می‌دهد که روش پیشنهادی در تمامی ۱۰ بخش، ضمن کاهش خصیصه‌ها به حدود ۰/۱ تعداد اولیه، دقت، فراخوانی و F1 را نسبت به حالت اولیه افزایش داده است و در شرایطی که امکان افزایش نبوده ، مقدار اولیه حفظ شده است. شکل (۱) مقایسه ارزیابی روش پیشنهادی(GA+SVM)و روش SVM را روی کل مجموعه نامه نشان می‌دهد.

نتایج مقایسه میانگین ۱۰ بخش در روش پیشنهادی و دو روش TFV+ SVM [۲۰] و IG+SVM [۲۰]، در جدول (۲) نشان داده شده‌است. نتایج نشان می‌دهد که روش پیشنهادی، نسبت به روش IG+SVM، بهتر بوده و همانند روش TFV+ SVM دقت را روش مقدار یک نگه می‌دارد. یعنی هیچیک از نامه‌های عادی به اشتباه به عنوان هرزنامه از فهرست نامه‌های عادی حذف نشده است. شکل (۲) مقایسه نتایج ارزیابی روش پیشنهادی(GA+SVM)، TFV+SVM و IG+SVM را نشان می‌دهد.

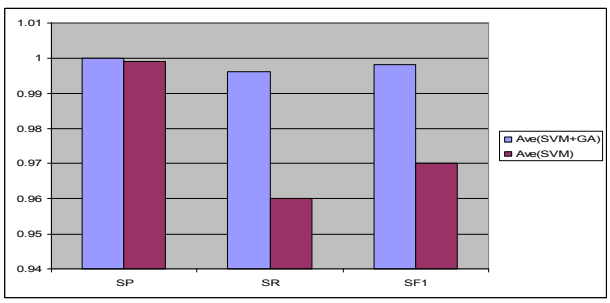
جدول (۳) نیز نتایج مقایسه روش پیشنهادی را با روش‌های بیزین ساده، spamcop و روش حبیبی و همکارش نشان می‌دهد. این مقایسه نیز نشان می‌دهد که روش پیشنهادی از نظر دقت از تمامی روش‌ها بهتر است. از نظر فراخوانی نیز نسبت به روش‌های بیزین و SpamCop، بیشتر از ۰/۲ بالاتر است و با روش حبیبی و کفائی برابر می‌باشد. یادآوری می‌گردد در کل روش پیشنهادی نسبت به روش حبیبی و همکارش بهتر است

[16] Spam Precision
[17] Spam Recall
[18] SpamF1
[19] Tournament Selection

زیرا هیچ‌یک از نامه‌های عادی را به اشتباه به عنوان هرزنامه از فهرست نامه‌های عادی حذف نمی‌کند. در حالی که در روش حبیبی و همکارش این اتفاق می‌افتد. شکل(۳) مقایسه نتایج ارزیابی روش پیشنهادی(GA+SVM)، بیزین ساده، SpamCop و حبیبی را نشان می‌دهد.

جدول۱ مقایسه ارزیابی روش پیشنهادی(GA+SVM)و روش SVM به تفکیک هر بخش از داده‌ها

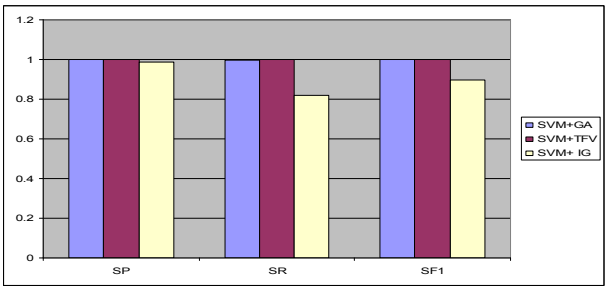
بخش	SVM				روش پیشنهادی			
	SP_o	SR_o	$SF1_o$	$F\#_o$	SP	SR	SF1	F#
P1	۱/۰	۰/۹۸	۰/۹۹	۴۶۶۵۳	۱/۰	۱/۰	۱/۰	۵۳۸۶
P2	۱/۰	۰/۹۴	۰/۹۷	۴۶۲۲۱	۱/۰	۱/۰	۱/۰	۵۳۱۴
P3	۱/۰	۰/۹۸	۰/۹۹	۴۶۱۵۸	۱/۰	۱/۰	۱/۰	۵۶۳۳
P4	۱/۰	۰/۹۸	۰/۹۹	۴۵۳۴۵	۱/۰	۱/۰	۱/۰	۵۲۱۷
P5	۱/۰	۰/۹۸	۰/۹۹	۴۵۹۳۰	۱/۰	۱/۰	۱/۰	۵۱۹۲
P6	۱/۰	۰/۹۸	۰/۹۹	۴۶۴۶۶	۱/۰	۱/۰	۱/۰	۵۰۹۰
P7	۱/۰	۰/۹۸	۰/۹۹	۴۶۹۷۵	۱/۰	۱/۰	۱/۰	۵۳۸۲
P8	۱/۰	۰/۹۴	۰/۹۷	۴۶۴۰۰	۱/۰	۱/۰	۱/۰	۹۱۵۹
P9	۰/۹۸	۰/۹۸	۰/۹۸	۴۶۴۲۷	۱/۰	۰/۹۸	۰/۹۹	۴۷۰۰
P10	۱/۰	۰/۹۲	۰/۹۶	۴۶۰۵۲	۱/۰	۱/۰	۱/۰	۸۰۲۵
Ave	۰/۹۹۹	۰/۹۶	۰/۹۷	۴۶۳۶۳	۱/۰	۰/۹۹۶	۰/۹۹۸	۵۹۰۹



شکل۱: مقایسه ارزیابی روش پیشنهادی(GA+SVM)و روش SVM

جدول۲ مقایسه کارایی سه روش دسته‌بندی هرزنامه

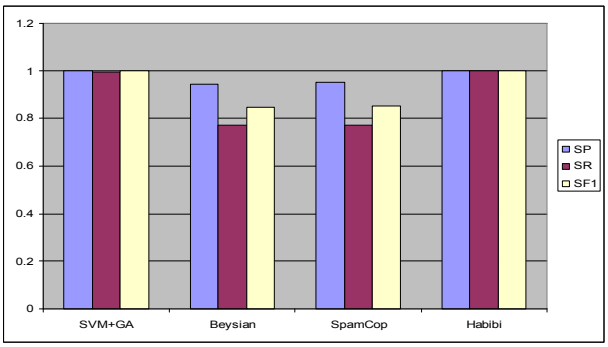
نوع فیلتر	SP	SR	SF1
IG +SVM [۲۰]	۰/۹۹۰	۰/۸۱۹	۰/۸۹۶
TFV+SVM [۲۰]	۱/۰	۱/۰	۱/۰
GA+SVM	۱/۰	۰/۹۹۸	۰/۹۹۹



شکل۲: مقایسه نتایج ارزیابی روش پیشنهادی(GA+SVM)، IG+SVM و TFV+SVM

جدول۳ مقایسه کارایی SVM+GA با فیلترهای غیر SVM

نوع فیلتر	SP	SR	SF1
بیزین ساده[۲۶]	۰/۹۴۴	۰/۷۶۹	۰/۸۴۷
SpamCop [۲۶]	۰/۹۵۱	۰/۷۷۱	۰/۸۵۱
روش حبیبی و کفائی [۲۶]	۰/۹۹۹	۰/۹۹۸	۰/۹۹۹
GA+SVM	۱/۰	۰/۹۹۸	۰/۹۹۹



شکل۳: مقایسه نتایج ارزیابی روش پیشنهادی(GA+SVM)، بیزین ساده، SpamCop و حبیبی

۶- نتیجه‌گیری و پژوهش‌های آینده

در این مقاله روش پیشنهادی(GA+SVM) به عنوان یک روش جداسازی هرزنامه‌ها معرفی شد. این روش روی مجموعه استاندارد LingSpam اعمال شد. نتایج نشان داد که روش پیشنهادی علاوه بر جداسازی هرزنامه‌ها با دقت و فراخوانی بالا، امکان کاهش خصایص تا حدود ۰/۱ تعداد اولیه را نیز دارا

است. مقایسه روش پیشنهادی با سایر روش‌های جداسازی هرزنامه‌ها، نشان داد که روش پیشنهادی به خوبی قابل مقایسه با بهترین روش‌های آماری موجود است. روش پیشنهادی از نظر دقت بهتر از تمامی روش‌ها و از نظر مقدار فراخوانی و F1 بهتر از روش‌هایی مانند IG+SVM، SpamCop و بیزین ساده عمل می‌کند. در ادامه قصد داریم با مطالعه و بررسی بیشتر در زمینه انتخاب تابع برازندگی، و بهبود عملگرها، میزان کارایی روش را افزایش داده و تعداد خصایص را به میزان بیشتری کاهش دهیم. همچنین بررسی در زمینه امکان به‌کارگیری PSO[20] در زمینه انتخاب خصیصه و جداسازی هرزنامه‌ها، از دیگر زمینه‌های تحقیقاتی آینده می‌باشد.

۷- مراجع

[1] Aladdin Knowledge Systems, Anti-Spam white paper, http://www.eAladdin.com.

[2] X. Carreras, L. Ma´rquez, "Boosting trees for anti-spam email filtering," *Proc. of fourth Int'l Conf. on Recent Advances in NaturalLanguage Processing*, 2001, pp. 58–64.

[3] R. Segal, et. al , "Spamguru: an enterprise anti-spam filtering system", *Proc. of First Conf. on Email and Anti-Spam*, 2004.

[4] I. Androutsopoulos et. al. ," An evaluation of naive Bayesian anti-spam filtering", *Proc. of the Workshop on Machine Learning in the New Information Age: 11th European Conference on Machine Learning*, 2000, pp. 9–17.

[5] I. Androutsopoulos, et. al. , " Learning to filter spam e-mail: a comparison of a Naı¨ve Bayesian and a memory-based approach", *Proc. of the workshop: Machine Learning and Textual Information Access*, 2000, pp. 1–13.

[6] I. Androutsopoulos et. Al. , "An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with encrypted personal e-mail messages, " *Proc. of the 23rd Annual International ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2000, pp. 160–167.

[7] H. Drucker, D. Wu, and V.N. Vapnik, "Support vector machines for spam categorization," *IEEE Trans. Neural Netw.* Vol. 10 (No. 5) (1999) 1048–1054.

[8] A. Kołcz, J. Alspector, "SVM-based filtering of e-mail spam with content-specific misclassification costs," *Proc. of TextDM'01 Workshop on Text Mining*, 2001.

[9] W.W. Cohen," Learning rules that classify e-mail, " *Proc. of AAAI Spring Symposium on Machine Learning in Information Access*,1996, pp. 18–25.

[10] G. Sakkis, et. al. , "Stacking classifiers for anti-spam filtering of e-mail," *Proc. of the 6th Conf. on Empirical Methods in Natural Language Processing*, 2001, pp. 44–50.

[11] M. Sahami, et. al," A Bayesian approach to filtering junk e-mail," *Learning for Text Categorization–Papers from the AAAI Workshop*, 1998, pp. 55–62.

[12] M. Woitaszek, M. Shaaban, and R. Czernikowski, "Identifying junk electronic mail in microsoft outlook with a support vector machine," *Proc. of the 2003 Symposium on Applications and the Internet*, 2003, pp. 166–169.

[13] D.C. Trudgian, Z.R. Yang, "Spam classification using nearest neighbour techniques," *Proc. of Fifth International Conf. on Intelligent Data Engineering and Automated Learning*, 2004, pp. 578–585.

[14] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, Vol. 34, No. 1, 2000.

[18] C. Stoean, et al., "Evolutionary Detection of Rules for Text Categorization. Application to Spam Filtering", *Advances in Intelligent Systems and Technologies, Proc. ECIT2004 –3 rd Europ. Conf. Intell. Syst. and Tech.*, pp. 21-23, 2004.

[19] D. Dumitrescu, "*Genetic Chromodynamics*," Studia Universities Babes Bolyai, Ser. Informatica, pp. 39-50, 2000.

[20] I. Koprinska, J. Poon, I. Clark, and J. Chan, "Learning to Classify e-mails", *INFORMATION SCIENCES*, Elsevier, No. 177, p. 2167-2187, 2007.

[21] K. Aas, L. Eikvil , "Text Categorization: A Survey" , *International Conference on machine learning*, pp.128-156, 1999.

[22] F. Debole and F. Sebastiani , " Supervised term weighting for automated text categorization", proceedings of SAC-03, 18th ACM Symposium on Applied Computing, Melbourne, US, pp. 784-788, 2003.

[23] LingSpam and PU1 dataset, <http://www.aueb.gr/users/ion/publications.html>.

[۲۵] جلیلی سعید، بیطرفان مهدی ، " انتخاب خصیصه به روش ترکیبی فیلتری- روکشی در دسته بندی متون "، هشتمین کنفرانس سالانه انجمن کامپیوتر ایران، دانشگاه فردوسی مشهد، ۶-۸ اسفند ۱۳۸۱.

[۲۶] حبیبی جعفر ، کفائی مهران، "ارایه الگوریتم آماری هوشمند ضد هرزنامه‌های متنی"، دهمین کنفرانس سالانه انجمن کامپیوتر ایران، مرکز تحقیقات مخابرات ایران، ۲۷-۲۹ اسفند، بهمن ۱۳۸۳.

[20] Particle Swarm Optimization