تاریخ دریافت مقاله: ۹۶/۱۰/۰۳
تاریخ پذیرش مقاله: ۹۷/۰۳/۱۹

بهبود الگوریتم بهینه‌سازی علف‌های هرز مهاجم با K نزدیک‌ترین همسایه در طبقه‌بندی ایمیل هرزنامه

فرهاد سلیمانیان قره چپق*
استادیار، گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران
پست الکترونیکی: farhad@iaurmia.ac.ir

مهدی وفادار
کارشناس ارشد مهندسی کامپیوتر، گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد ارومیه، ارومیه، ایران
پست الکترونیکی: v.mehdi1364@gmail.com

محسن موتمن فر
کارشناس ارشد مهندسی کامپیوتر، گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد ارومیه، ارومیه، ایران
پست الکترونیکی: motaman.mohsen@yahoo.com

چکیده

امروزه ایمیل که یک وسیله ارتباطی سریع و کم‌هزینه می‌باشد به یکی از مهم‌ترین ابزارهای مدرن برای ارتباط در محیط‌های برخط تبدیل شده است. اما مشکلی اصلی این است که ارسال هرزنامه باعث آزار کاربران، اتلاف زمان، هزینه، منابع شبکه و پهنای باند می‌شود. بنابراین ایمیل هرزنامه به یک مشکل جدی برای کاربران و سازمان‌هایی که مدت زمان زیادی از وقت خود را با ایمیل و کامپیوتر کار می‌کنند تبدیل شده است. در ایمیل هرزنامه بیشتر اوقات، کلاهبرداران با فرستادن هرزنامه قصد نفوذ به کامپیوتر یا کلاهبرداری را دارند که این گونه هرزنامه‌ها اغلب حاوی نرم‌افزارها و پیوندهای مخرب هستند که در درون خود انواع بدافزارها و ویروس‌ها را دارند. در این مقاله یک مدل بهبود یافته مبتنی بر الگوریتم بهینه‌سازی علف‌های هرز مهاجم و K نزدیک‌ترین همسایه برای تشخیص ایمیل هرزنامه پیشنهاد شده است. در مدل پیشنهادی با استفاده از الگوریتم بهینه‌سازی علف‌های هرز مهاجم، ویژگی‌های مهم برای تشخیص ایمیل هرزنامه را انتخاب می‌کنیم و سپس توسط K نزدیک‌ترین همسایه، نمونه‌ها را طبقه‌بندی می‌کنیم. انتخاب ویژگی باعث می‌شود که درصد دقت تشخیص بیشتر باشد. ارزیابی مدل پیشنهادی برروی مجموعه داده Spambase انجام شده است. نتایج نشان داده که مقدار صحت در مدل پیشنهادی برابر ۹۱,۱۱٪است که در مقایسه با مدل‌های دیگر دقت تشخیص بهتری دارد.

واژه‌های کلیدی: تشخیص ایمیل هرزنامه، الگوریتم بهینه‌سازی علف‌های هرز مهاجم، K نزدیک‌ترین همسایه

۱. مقدمه

پرکاربردترین و مهم‌ترین ابزار ارتباطی درجهان امروز که برای همه کاربران ساده و دردسترس می‌باشد ایمیل

* نویسندهٔ مسئول

است. پیشرفت قابل توجه در عرصه فناوری‌های ارتباطات جهانی و همچنین ایمیل این امکان را به کاربران می‌دهد که در تمام نقاط جهان و با دسترس بودن اینترنت به راحتی می‌توانید ایمیل‌های خود را به سرتاسر جهان ارسال نمایید و بدون آن‌که زمان زیادی را تلف کنید فقط با چند ثانیه ایمیل خود را به گیرنده ارسال نمایید [۱]. از طرفی دیگر، دسترسی آسان، سریع و بی‌شمار به ایمیل توسط کاربران موجب شده است تا حجم وسیعی از محتوا و اطلاعات باساختار و بدون ساختار به ایمیل کاربران ارسال گردد. از آنجایی که امکان تحقیق و نیز فرصت کافی برای بررسی اعتبار آن‌ها در زمان کوتاه وجود ندارد، لذا به یک تهدید امنیتی تبدیل می‌شوند و در بیشتر موارد به شکل ایمیل هرزنامه برای کاربران ارسال می‌شوند [۲]. هرزنامه برای اهداف مختلف بدون اجازه و تمایل دریافت کننده، ارسال می‌شود به طوری که کاربر پس از دریافت آن ایمیل (که می‌تواند با اهداف خرابکارانه یا با هدف تبلیغات ارسال شود) نمی‌تواند جلوی دریافت مجدد آن را بگیرد [۳]. درنتیجه، تهدیدهای امنیتی در محیط برخط اینترنت استفاده از ایمیل را آسیب‌پذیر کرده‌اند که یکی از مهم‌ترین تهدیدها هرزنامه است [٤]. امروزه هرزنامه مانند سیلی عظیم، و با کپی‌های فراوان به ایمیل‌های کاربران ارسال می‌شود.

در بیشتر مواقع هرزنامه‌ها فقط باعث مزاحمت کاربرانی که آن را دریافت می‌کند می‌شوند اما گاهی قضیه فراتر از این است و نفوذگرها با فرستادن هرزنامه قصد نفوذ به کامپیوتر کاربران یا کلاهبرداری را دارند که این گونه هرزنامه‌ها اغلب حاوی نرم‌افزارها و پیوندهای مخربی هستند که در درون خود انواع بدافزارها، ویروس‌ها و باج افزارها را دارند [٥]. تاکنون روش‌های مختلفی برای مبارزه با هرزنامه ایجاد شده که یکی از راهکارها استفاده از نرم‌افزارهای فیلترینگ است که در این برنامه‌ها با استفاده از کلید واژه‌های خاصی، موضوع پیام و ایمیل‌ها بررسی می‌شود و در صورت شناسایی آن‌ها ایمیل دریافت شده حذف می‌شود [٦]. کاربرانی که در ایمیل خود هرزنامه دریافت می‌کنند، می‌توانند نشانی ایمیل مورد نظر را در قسمت ایمیل‌های مسدود شده قرار دهند تا از دریافت مجدد ایمیل از آن نشانی جلوگیری شود.

استفاده از فناوری اطلاعات و ارتباطات در حوزه‌های مختلف در حال رشد و گسترش بوده و اطلاعات کاربران برخط با ایمیل هرزنامه در خطر حمله نفوذگرها قرار دارد. به همین دلیل شناسایی و تشخیص ایمیل هرزنامه یک راهکار مفید برای جلوگیری از نفوذ افراد ناشناس به حساب کاربری کاربران می‌باشد [۷]. در این مقاله با بررسی ویژگی‌ها و مقادیر مربوط به مجموعه داده Spambase [۸] با استفاده از ترکیب الگوریتم بهینه‌سازی علف‌های هرز مهاجم[1] [۹] و K نزدیک‌ترین همسایه[2] [۱۰] روشی موثر برای دقت تشخیص ایمیل هرزنامه پیشنهاد می‌کنیم. در مدل پیشنهادی از الگوریتم بهینه‌سازی علف‌های هرز مهاجم برای انتخاب ویژگی و از k نزدیک‌ترین همسایه برای طبقه‌بندی نمونه‌ها استفاده می‌شود. الگوریتم بهینه‌سازی علف‌های هرز مهاجم یکی از الگوریتم‌های بهینه‌سازی جدید و توانمند است که بر اساس تقلید از قابلیت تطابق‌پذیری و تولید مثل، روابط بین مقادیر را پیدا می‌کند. این الگوریتم با الهام از تکثیر و رشد علف‌های هرز ابداع شده است. طبق تعریف، علف‌های هرز به گیاهانی ناخواسته اطلاق می‌شود که دارای رفتار تهاجمی برای رشد بوده و به عنوان تهدیدی برای دیگر گیاهان زراعی مفید می‌باشند و جلوی رشد آن‌ها را می‌گیرند. این الگوریتم در عین سادگی، در یافتن نقاط بهینه بسیار مؤثر و سریع می‌باشد. در واقع این الگوریتم براساس ویژگی‌های طبیعی علف‌های هرز مانند تولید بذر، رشد و تنازع برای بقا عمل می‌کند.

هدف در انتخاب ویژگی در واقع پیدا کردن کوچک‌ترین زیرمجموعه از ویژگی‌های ورودی با بیشترین خاصیت پیشگویانه یا بیشترین اطلاعات جداکننده در کل مجموعه داده‌ها است. با انتخاب ویژگی‌های مناسب از میان

1- Invasive Weed Optimization (IWO)
2- K-Nearest Neighbor (KNN)

ویژگی‌های فراوان استخراج شده، می‌توان علاوه بر دست یافتن به نرخ دقت دسته‌بندی بالا، هزینه‌های محاسباتی را کاهش داده و از استخراج ویژگی‌های غیر ضروری خودداری کرد.

ساختار کلی مقاله را به‌شرح زیر سازماندهی کردیم: در بخش دوم، کارهای قبلی را توضیح می‌دهیم. در بخش سوم، مدل پیشنهادی و توضیحات لازم در مورد الگوریتم بهینه‌سازی علف‌های هرز مهاجم و K نزدیک‌ترین همسایه را توضیح خواهیم داد. در بخش چهارم، ارزیابی و نتایج مدل پیشنهادی و مقایسه‌ها را توضیح می‌دهیم و نهایتاً در بخش پنجم به نتیجه‌گیری و کارهای آینده خواهیم پرداخت.

۲. تحقیقات انجام شده

در دودهه اخیر، با رشد چشمگیر استفاده از ایمیل و از طرفی حجم ایمیل‌های ناخواسته و مزاحم تحت عنوان هرزنامه باعث انگیزه محققان برای طراحی و پیاده‌سازی سیستم‌هایی جهت فیلترسازی ایمیل‌های هرزنامه با الهام از تکنیک‌های طبقه‌بندی شد.

مدل شبکه عصبی مصنوعی بر پایهٔ داده‌گردانی با روش گروهی[3] [۱۱] برای طبقه‌بندی ایمیل هرزنامه پیشنهاد شده است. این شبکه عصبی مصنوعی یکی از پرکاربردترین شبکه‌های عصبی مصنوعی است که از توانایی بالایی در مدل‌سازی داده‌های پیچیده برخوردار است. این روش یک روش مدل‌سازی آماری ردهٔیک نیست؛ بلکه فرآیندی منظم برای غلبه بر ضعف‌های آماری و شبکه‌های عصبی مصنوعی است. این نوع شبکه عصبی، حاوی مجموعه‌ای از نرون‌ها است که از پیوند جفت‌های مختلف از طریق یک چندجمله‌ای درجه دوم به وجود می‌آیند. شبکه با ترکیب چند جمله‌ای‌های درجه دوم حاصل از تمامی نرون‌ها برای یک مجموعه از ورودی‌ها با کمترین خطای خروجی تعریف می‌شود. ارزیابی شبکه عصبی مصنوعی بر پایهٔ داده‌گردانی با روش گروهی برروی مجموعه داده جهانی ومعتبر Spambase انجام شده است. مجموعه داده Spambase شامل ۵۷ ویژگی و ٤٦٠١ نمونه است. نمونه‌ها در دو ردهٔ هرزنامه و غیر هرزنامه طبقه‌بندی شده‌اند. نتایج با انتخاب ویژگی‌های مختلف نشان داده که مقدار معیار صحت در شبکه عصبی مصنوعی بر پایهٔ داده‌گردانی با روش گروهی در مقایسه با مدل‌های پرسپترون چندلایه و بیز ساده بیشتر است. بیشترین مقدار صحت در مدل شبکه عصبی مصنوعی بر پایهٔ داده‌گردانی با روش گروهی برابر ۹۲,٤ و در مدلهای پرسپترون چندلایه و بیز ساده[4] به‌ترتیب برابر ۹۱,۷ و ۷۵,٤ است.

مدل ترکیبی بهینه‌سازی اجتماع ذرات-الگوریتم انتخاب منفی[5] به منظور طبقه‌بندی ایمیل هرزنامه پیشنهاد شده است [۱۲]. ارزیابی برروی مجموعه داده Spambase انجام شده است. در مدل ترکیبی بهینه‌سازی اجتماع ذرات-الگوریتم انتخاب منفی از مدل بهینه‌سازی اجتماع ذرات برای جستجوی مقدار ویژگی‌ها در فضای مسئله و از مدل الگوریتم انتخاب منفی برای انتخاب ویژگی‌ها استفاده شده است. روش پیشنهادی دارای دو فاز آموزش و آزمایش است. در فاز آموزش ابتدا با مدل الگوریتم انتخاب منفی و از طریق فیلترینگ نمونه‌ها تعدادی ویژگی انتخاب می‌شوند. در فاز آزمایش، براساس مدل بهینه‌سازی اجتماع ذرات-الگوریتم انتخاب منفی تصمیم مناسبی در راستای هرزنامه یا غیر هرزنامه بودن یک نمونه انجام می‌شود. نتایج نشان داده که مدل بهینه‌سازی اجتماع ذرات-الگوریتم انتخاب منفی در مقایسه با مدل‌های بیز ساده، انتخاب ویژگی متمایز-ماشین بردار پشتیبان[6] و الگوریتم انتخاب منفی دقت تشخیص بالاتر و در مقایسه با مدل ماشین بردار پشتیبان دقت تشخیص کمتری دارد. مقدار معیار صحت در مدل بهینه‌سازی اجتماع ذرات-الگوریتم انتخاب منفی برابر ۸۳,۲۰ و در مدل‌های بیز ساده، انتخاب ویژگی متمایز-ماشین بردار پشتیبان و الگوریتم انتخاب منفی به‌ترتیب برابر ۷۸,۸، ۷۱ و ٦٨,٨٦ است.

مدل QUANT [۱۳] که ترکیبی از شبکه عصبی مصنوعی

3- Group Method of Data Handling

4- Naive Bayes

5- Particle Swarm Optimization-Negative Selection Algorithm

6- Distinguishing Feature Selection and Support Vector Machine

و درخت تصمیم‌گیری است برای تشخیص ایمیل هرزنامه پیشنهاد شده است. در این مدل داده‌ها با استفاده از شبکه عصبی مصنوعی آموزش داده و آزمایش می‌شوند و با استفاده از درخت C4.5 طبقه‌بندی می‌شوند. از الگوریتم C4.5 جهت تحلیل ویژگی‌های اصلی موثر بر ایمیل هرزنامه و تشخیص استفاده شده است. در درخت C4.5 هر مسیر از ریشه به سمت یک گره، نمایانگر یک قانون طبقه‌بندی می‌باشد. ارزیابی برروی دو مجموعه داده SpamAssassin و Corpus 2006 انجام شده است. نتایج برروی مجموعه داده SpamAssassin نشان داده که مقدار صحت در مدل QUANT برابر ۸۹,۱۵ و در مدل‌های بیز ساده، بهینه‌سازی حداقل متوالی[۷] و C4.5 به‌ترتیب برابر ۸۱,۰۸، ۸۸,۶۲ و ۷۳,۰۸ است. و همچنین برروی مجموعه داده Corpus ۲۰۰۶ مقدار صحت در مدل QUANT برابر ۹۰,۸۷ و در مدل‌های بیز ساده، بهینه‌سازی حداقل متوالی و C٤,٥ به‌ترتیب برابر ۸۸,۱۵، ۸۹,۷۹ و ۸۸,۱۵ است.

مدل ترکیبی شبکه عصبی مصنوعی-بهینه‌سازی اجتماع ذرات به منظور تشخیص ایمیل هرزنامه پیشنهاد شده است [۱٤]. از الگوریتم بهینه‌سازی اجتماع ذرات برای انتخاب ویژگی‌ها و از مدل پرسپترون چندلایه برای آموزش و آزمایش داده‌ها و طبقه‌بندی استفاده شده است. در مدل شبکه عصبی مصنوعی-بهینه‌سازی اجتماع ذرات از شبکه عصبی پرسپترون با تابع فعال‌سازی سیگموئید برای لایه پنهان و ۸۰ درصد داده‌ها برای آموزش و ۲۰درصد برای آزمایش استفاده شده است. تعداد لایه‌های مخفی در مدل پرسپترون چندلایه بین ۱۵-۳ لحاظ شده و تکرار الگوریتم بهینه‌سازی اجتماع ذرات برای انتخاب ویژگی برابر ۲۰۰ است. ارزیابی برروی مجموعه داده LingSpam با ٤۸۱ ایمیل هرزنامه و ۲۱۷۱ ایمیل غیر هرزنامه و مجموعه داده SpamAssassin با ٦۰۰۰ نمونه ایمیل انجام شده است. ارزیابی برروی مجموعه داده SpamAssassin و LingSpam نشان داده که مقدار معیار صحت در مدل شبکه عصبی مصنوعی-بهینه‌سازی اجتماع ذرات به‌ترتیب برابر ۹۹,۹۸ و ۹۹,۷۹ است. مقایسه‌ها نشان داده که مدل شبکه عصبی مصنوعی-بهینه‌سازی اجتماع ذرات در مقایسه با مدل‌های ماشین بردار پشتیبان با تابع هسته، ماشین بردار پشتیبان با تابع شعاعی پایه[۸] و شبکه عصبی مصنوعی تابع شعاعی پایه دقت تشخیص بهتری دارد.

مدل‌های درخت تصمیم‌گیری، ماشین بردار پشتیبان و شبکه عصبی مصنوعی تابع شعاعی پایه و ترکیب آن‌ها بر روی دو مجموعه داده با ١٤ ویژگی آزمایش و اجرا شده‌اند [۱۵]. مجموعه داده اولی شامل ٥۰٤ ایمیل (۳۳٦ ایمیل و ۱٦۸ ایمیل هرزنامه) و مجموعه داده دومی شامل ٦٥۷ ایمیل(۳۸۷ ایمیل و ۲۷۰ ایمیل هرزنامه) است. در مدل درخت تصمیم‌گیری از آنتروپی، مدل ماشین بردار پشتیبان از تابع کرنل و شبکه عصبی مصنوعی تابع شعاعی پایه از خطای میانگین استفاده شده است. نتایج بر روی مجموعه داده اولی نشان داده که مقدار صحت در مدل ترکیبی برابر ۹۱,۰۷ و در مدل‌های درخت تصمیم‌گیری، ماشین بردار پشتیبان و شبکه عصبی مصنوعی تابع شعاعی پایه به‌ترتیب برابر ۸۹,۸۸، ۸۸,٦۹ و ۸۹,۸۸ است. و برروی مجموعه داده دومی مقدار صحت در مدل ترکیبی برابر ۹۱,۷۸ و در مدل‌های درخت تصمیم‌گیری، ماشین بردار پشتیبان و شبکه عصبی مصنوعی تابع شعاعی پایه به‌ترتیب برابر ۹۰,۸۷، ۹۰,۸۷ و ۸۹,۰٤ است.

ادریس و همکارانش [۱٦] مدل ترکیبی تکامل تفاضلی-الگوریتم انتخاب منفی[۹] را برای تشخیص ایمیل هرزنامه پیشنهاد کرده‌اند. در این مدل از الگوریتم تکاملی تفاضلی برای جستجوی ویژگی‌ها در فضای مسئله استفاده شده است. الگوریتم تکاملی تفاضلی یکی از جدیدترین روش‌های جستجو است. الگوریتم تکاملی تفاضلی به عنوان روشی قدرتمند و سریع برای مسائل بهینه‌سازی در فضاهای پیوسته معرفی شده است. ارزیابی مدل تکامل تفاضلی-الگوریتم انتخاب منفی برروی مجموعه داده Corpus با ۱۸۱۳ ایمیل هرزنامه و ۲۷۸۸ ایمیل غیرهرزنامه طبق

7- Sequential Minimal Optimization

8- Radial Basis Function

9- Differential Evolution-Negative Selection Algorithm

معادله (۱) انجام شده اســت. معادله(۱) به منظور ضریب همبستگی[10] نمونه‌ها استفاده شده است.

$$CC = \frac{[(TP)(TN) - (FP)(FN)]}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \qquad (۱)$$

پارامترهای درست مثبت(TP)، درست منفی(TN)، کاذب مثبت(FP)، کاذب منفــی(FN) از پارامترهــای اصلی برای معیار ضریب همبستگی هستند. خروجی دقت معیار ضریب همبســتگی در معادلــه() برمبنای پارامترهای ذکر شــده می‌باشد. پارامتر درســت مثبت، بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها مثبت بوده و الگوریتم دسته‌بندی نیز دســته آن‌ها را به‌درستی مثبت تشــخیص داده است. پارامتر درســت منفی، بیانگر تعداد رکوردهایی اســت که دســته واقعی آن‌ها منفی بوده و الگوریتم دســته‌بندی نیز دسته آن‌ها را به‌درستی منفی تشخیص داده است. پارامتر کاذب مثبت، بیانگر تعداد رکوردهایی است که دسته واقعی آن‌هــا منفی بوده و الگوریتم دســته‌بندی، دســته آن‌ها را به اشتباه مثبت تشــخیص داده است. پارامتر کاذب منفی، بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها مثبت بوده و الگوریتم دســته‌بندی دسته آن‌ها را به‌اشتباه منفی تشخیص داده است.

نتایج مــدل ترکیبی تکامــل تفاضلی–الگوریتم انتخاب منفــی نشــان داده که مقــدار ضریب همبســتگی در مدل الگوریتــم انتخاب منفی و تکامل تفاضلی–الگوریتم انتخاب منفی به‌ترتیب برابر ٦٨,٨٦ و ٨٠,٦٦ است. و مقدار ضریب همبستگی در مدل الگوریتم انتخاب منفی و تکامل تفاضلی–الگوریتم انتخاب منفی به‌ترتیب برابر ٣٦,٠٦ و ٦٠,٠٨ است.

مدل بهبود داده شده بهینه‌سازی اجتماع ذرات–الگوریتم انتخاب منفی [۱۷] برای تشخیص طبقه‌بندی ایمیل هرزنامه پیشنهاد شده است. ارزیابی مدل بهینه‌سازی اجتماع ذرات–الگوریتم انتخاب منفی بــرروی مجموعه داده Corpus با ۱۸۱۳ ایمیل هرزنامه و ۲۷۸۸ ایمیل غیرهرزنامه انجام شده اســت. در مدل بهینه‌سازی اجتماع ذرات–الگوریتم انتخاب منفی برای تغییر موقعیت ذرات و سرعت ذرات از یک مدل ریاضی جدید طبق معادله(۲) استفاده شده است.

(۲)

$$dv_i = \left\{ -\propto \sum_{j \neq i} \left[\left[\frac{r}{x_{ij}}\right]^p - \left[\frac{r}{x_{ij}}\right]^q \right] x_{ij} - \beta \sum_{j \neq i} \left[\left[\frac{r}{x_{ij}}\right]^p - \left[\frac{r}{x_{ij}}\right]^q \right] V_{ij} + f_i(t, x_i, v_i) \right\} dt$$

در معادلــه(۲)، x موقعیــت ذرات، v ســرعت ذرات، r پارامتر یادگیری، t تعداد تکرار، مقدار p و q در بازه [۰,۱] می‌باشــد. هدف معادله(۲)، همگرایی به جواب بهینه است. همگرایــی یعنی این‌که الگوریتم مقادیر مناســب و بهینه را پیدا کند.

نتایج نشــان داده که مدل بهینه‌ســازی اجتماع ذرات–الگوریتم انتخاب منفی در مقایســه با مدل‌های بهینه‌سازی اجتماع ذرات، الگوریتم انتخاب منفی، بیز ســاده، ماشــین بردار پشــتیبان، انتخــاب ویژگی متمایز–ماشــین بردار پشتیبان[11]، شبکه عصبی مصنوعی و فازی دقت تشخیص بهتری دارد.

در جدول(۱)، مزایا و معایب هر مدل نشــان داده شده اســت. مدل‌های مختلفی برای تشــخیص ایمیــل هرزنامه پیشــنهاد شده اســت. اما هر مدل می‌تواند مزایا و معایبی داشته باشد.

۳. مدل پیشنهادی

در این بخش، مدل پیشــنهادی که ترکیبی از الگوریتم بهینه‌سازی علف‌های هرز مهاجم و K نزدیک‌ترین همسایه است را توضیح می‌دهیم. از الگوریتم بهینه‌سازی علف‌های هرز مهاجم برای انتخاب ویژگی و از K نزدیک‌ترین همسایه بــرای طبقه‌بندی داده‌ها اســتفاده می‌کنیم. در شــکل(۱)، روندنمای کلی مدل پیشنهادی نشان داده شده است.

در ابتــدا، خوانــدن مجموعه داده‌ها که شــامل ٤٦٠١ نمونه می‌باشد انجام می‌گیرد. نمونه‌ها سطر به سطر وارد یــک ماتریس 57x4601 می‌شــوند. در ابتدا ۵۷ ویژگی از مجموعه داده Spambase به عنوان ورودی برای الگوریتم بهینه‌سازی علف‌های هرز مهاجم در نظر گرفته می‌شوند.

10- Correlation Coefficient

11- Distinguishing Feature Selection and Support Vector Machine

جدول ۱: مزایا و معایب مدل‌های پیشنهاد شده برای تشخیص ایمیل هرزنامه

مراجع	مدل‌ها	مزایا	معایب
[۱۱]	مدل شبکه عصبی مصنوعی بر پایهٔ داده‌گردانی با روش گروهی	*مدل شبکه عصبی مصنوعی بر پایهٔ داده‌گردانی با روش گروهی، شبکه‌ای خودسازمانده و یک سویه است که از چندین لایه و هر لایه از چندین نرون تشکیل یافته است و لذا آموزش تا کمترین خطا انجام می‌گیرد.	*سرعت همگرایی آن خیلی کند است و به نوفه‌های موجود در مجموعه داده‌های ورودی و خروجی که جهت آموزش شبکه عصبی به کار می‌روند بسیار حساس است.
[۱۲]	بهینه‌سازی اجتماع ذرات- الگوریتم انتخاب منفی	*انتخاب ویژگی برمبنای نزدیک‌ترین فاصله *ماتریسی جدید ویژگی‌ها که تولید شده با ماتریس اولیه مقایسه می‌شود و اگر ماتریس جدید دارای دقت بیشتری باشد جایگزین ماتریس اولیه می شود.	*گیر افتادن در راه حل غیربهینه
[۱۳]	ترکیبی از شبکه عصبی مصنوعی و درخت تصمیم‌گیری	*استفاده از قوانین درخت تصمیم گیری *کاهش حجم ویژگی‌ها با استفاده از درخت تصمیم‌گیری *آموزش و آزمایش نمونه‌ها *کاهش خطا در هر مرحله از آموزش	*افزایش زمان محاسباتی
[۱۴]	شبکه عصبی مصنوعی- بهینه‌سازی اجتماع ذرات	*آموزش و آزمایش نمونه‌ها برمبنای مقایسه *جستجو در فضا برای انتخاب بهترین ویژگی‌ها	*گیر افتادن در راه حل غیربهینه
[۱۵]	درخت تصمیم‌گیری	*استفاده از قوانین تصمیم‌گیری *پیدا کردن مرز بهینه بین رده‌های هرزنامه و غیرهرزنامه	*افزایش زمان محاسباتی
	ماشین بردار پشتیبان		
	شبکه عصبی مصنوعی تابع شعاعی پایه		
[۱۶]	تکامل تفاضلی-الگوریتم انتخاب منفی	*کشف ویژگی‌های جدید بعد از هر دور مقایسه *احتمال تشخیص درست ویژگی‌ها به دلیل روش ادغام	*افزایش زمان محاسباتی
[۱۷]	بهینه‌سازی اجتماع ذرات- الگوریتم انتخاب منفی	* انتخاب ویژگی برمبنای نزدیک‌ترین فاصله	*گیر افتادن در راه حل غیربهینه

جمعیت اولیه برمبنای نمونه‌های مجموعه داده Spambase ایجاد می‌شوند. مراحل الگوریتم بهینه‌سازی علف‌های هرز برای انتخاب ویژگی را می‌توان به طور خلاصه به صورت زیر بیان کرد:

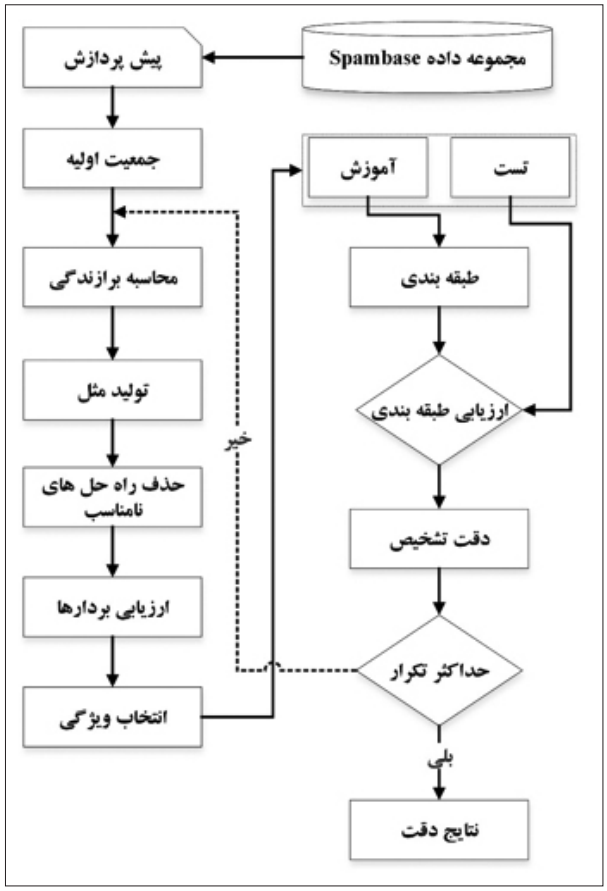
الف) تولید جمعیت اولیه (برمبنای مجموعه داده) و ارزیابی تابع هدف آن‌ها. یک جمعیت اولیه در فضای حل مسئله برمبنای نمونه‌های اولیه پراکنده و سپس ارزیابی می‌شوند. از محاسبه فاصله بین ویژگی‌ها برای ارزیابی استفاده می‌گردد. ویژگی‌هایی که فاصله کمتری دارند انتخاب می‌شوند.

ب) تولید مثل بر مبنای شایستگی و بروزرسانی انحراف معیار. در مراحل بعدی ویژگی‌هایی انتخاب می‌شوند که به ویژگی‌های انتخاب شده در مراحل اولیه نزدیک‌تر باشند و همچنین مقدار برازندگی آن‌ها کمتر باشد. با این روش از تولید مثل به عنوان راهکاری برای نزدیک شدن به ویژگی‌ها بهینه استفاده می‌شود.

لذا، فضای مسئله را براساس ویژگی‌های مجموعه داده ایجاد می‌کنیم. سپس عملیات پیش پردازش برروی نمونه‌های ماتریس انجام می‌گیرد. هدف پیش پردازش این است که نمونه‌هایی با مقدار تهی و مقدار نامعتبر شناسایی شوند و یک مقدار مناسب با توجه به بازه ویژگی‌ها برای آن‌ها تعیین شود. به عبارتی نرمال‌سازی برروی مقدار نمونه‌ها انجام گیرد.

۳-۱ انتخاب ویژگی

مجموعه داده Spambase دارای ویژگی‌های گمراه کننده، نامرتبط، زائد و نوفه‌دار می‌باشد که در صحت عملکرد و زمان محاسباتی تاثیر منفی دارند. برای حل



شکل۱: روندنمای مدل پیشنهادی

این مشـــکل می‌توان از انتخـــاب زیرمجموعه ویژگی بهینه اســـتفاده نمود. انتخاب ویژگی مدل را ساده‌تر می‌کند، این کار باعث می‌شود که هزینه محاسبات کاهش یابد. با حذف ویژگی‌هـــای غیرمفید، مدل شـــفاف‌تر و جامع‌تر می‌گردد. همچنین باعث تســـریع در فرآیند یادگیری، کاهش فضای ذخیره‌ســـازی و بهبود کارایی مانند صحت می‌شـــود. در نتیجـــه وجود الگوریتم انتخاب ویژگـــی برای کاهش ابعاد داده‌ها در داده‌های با ابعاد بالا ضروری است.

الگوریتم بهینه‌ســـازی علف‌های هـــرز مهاجم یکی از الگوریتم‌های بهینه‌سازی جدید و توانمند است که از قابلیت تطابق‌پذیری و تصادفی بودن کولونی علف‌های هرز، الهام گرفته شـــده است. هر عضو از جمعیت بر طبق توانایی‌اش می‌تواند بین دو مقدار تعیین شده حداکثر($S_{\max}$) و حداقل ($S_{\min}$) تولید دانه کند. تعداد دانه‌هایی که هر گیاه می‌تواند تولیـــد کند به طور خطـــی از کمترین تعـــداد دانه ممکن تا بیشـــترین تعداد تغییر می‌کند. تعداد دانه‌هایی که هر علف هرز می‌تواند تولید کند طبق معادله (۳) محاســـبه می‌شود. در معادله(۱)، پارامتر $f_{\min}^{k}$ حداقـــل تابع هزینه کلونی در تکرار k است و پارامتر $f_{\max}^{k}$ حداکثر تابع هزینه کلونی در تکرار k است.

(۳)

$$seed_i^k = \left\lfloor S_{\max} - (S_{\max} - S_{\min}) \frac{f_i^k - f_{\min}^k}{f_{\max}^k - f_{\min}^k} \right\rfloor$$

طبق معادله (۳)، برای هر کدام از ویژگی‌های مجموعه داده Spambase برمبنای فضای مسئله، فاصله بهینه پیدا می‌شود. در ابتدا یک ویژگی انتخاب می‌شود و سپس مقدار آن ویژگی برمبنای معادلـــه(۳)، با مقدار ویژگی‌های دیگر مقایســـه می‌گـــردد. در انتها در بین مقادیر به‌دســـت آمده از محاســـبه فاصله بیـــن ویژگی‌ها، یک ویژگـــی به عنوان نزدیک‌تریـــن ویژگی از لحاظ مقدار، انتخاب می‌شـــود. در شـــکل(۲)، بردار تولید دانه‌ها و انتخـــاب دانه‌های بهینه از بردار دانه‌ها نشان داده شده است.

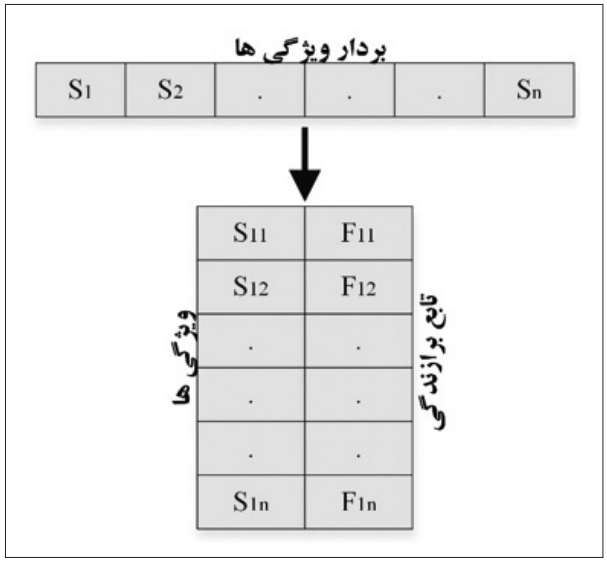
از الگوریتم بهینه‌ســـازی علف‌های هـــرز مهاجم برای انتخاب بهترین ویژگی‌ها به منظور بالابردن دقت تشخیص و کاهش زمان محاســـباتی اســـتفاده شده است. به منظور ارزیابی دانه‌هـــا در الگوریتم بهینه‌ســـازی علف‌های هرز مهاجم از معیار برازندگی اســـتفاده می‌کنیم. معیار صحت اصلی‌ترین معیار ارزیابی مربوط به طبقه‌بندی و تشخیص نمونه‌ها در فاز آموزش اســـت. سازماندهی نمونه‌های هر رده به درســـتی و یا نادرستی دسته‌بند بستگی دارد که به صورت درست مثبت، درســـت منفی، کاذب منفی و کاذب مثبت تعریف می‌شود.

(۵)

$$accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

۳-۲ طبقه‌بندی: K نزدیک‌ترین همسایه

بعـــد از انتخاب ویژگی‌ها، نمونه‌هـــا را به دو مجموعه آموزش و آزمایش تقســـیم می‌کنیـــم. برازندگی دانه‌ها با اســـتفاده از داده‌هـــای آزمایش ارزیابی می‌شـــود. در طی



شکل۲ : انتخاب دانه بهینه از بردار دانه‌ها

فرآیند تکامل الگوریتم، دانه‌ها به سمت نقاط بهینه همگرا می‌شوند. با استفاده از این روش تعداد ویژگی کاهش می‌یابد و بهینه‌ترین ویژگی‌ها را انتخاب می‌کنیم. در آخرین تکرار بهترین دانه از نظر معیار نزدیکی به عنوان دانه نهایی که نشان دهنده بهترین دانه است انتخاب بهینه شود. مدل K نزدیک‌ترین همسایه در اغلب موارد برای طبقه‌بندی به کار می‌رود، هرچند که می‌توان از آن برای تخمین و پیش‌بینی نیز استفاده نمود. در مدل K نزدیک‌ترین همسایه یک نمونه طبقه‌بندی نشده ممکن است به‌سادگی با مقایسه آن با شبیه‌ترین نمونه‌ها در مجموعه آموزشی یافت شود. بنابراین لازم است معیاری را برای تعیین فاصله بین نمونه‌ها مشخص نماییم. اگر یک بردار ویژگی به صورت $< a_1(x), a_2(x),..., a_n(x) >$ داشته باشیم از فاصله اقلیدسی طبق معادله (٤) برای به‌دست آوردن فاصله بین دو ویژگی x_i و x_j استفاده می‌کنیم.

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2} \qquad (٤)$$

در روش K نزدیک‌ترین همسایه، یک دسته باید شامل K نمونه از مجموعه نمونه‌های آموزشی باشد، به‌طوری که مشابه‌ترین نمونه‌ها در یک دسته باشند و براساس برتری دسته یا برچسب مربوط به آن‌ها در مورد نمونه آزمایشی جدید تصمیم‌گیری می‌شود. درنتیجه، پس از آن‌که فاصله اقلیدسی بین نقاط محاسبه شد، با مرتب‌سازی عناصر برحسب فاصله اقلیدسی، از میان k همسایه، برچسبی که از میان k همسایه دارای اکثریت است به نمونه ناشناخته داده می‌شود.

نمونه‌های آزمایشی، نمونه‌هایی هستند که از قبل به مدل داده نمی‌شوند، بلکه هدف نمونه‌های آزمایشی، ارزیابی نمونه‌های آموزشی است. برای مثال طبق جدول(۲)، فرض کنید که ۱۰ نمونه آموزشی و یک نمونه آزمایشی داریم. هدف این است که برای نمونه آزمایش یک رده برمبنای فاصله اقلیدسی و با استفاده از الگوریتم K نزدیک‌ترین همسایه مشخص شود. در مرحله اول باید فاصله بین نمونه‌ها محاسبه شود و سپس عمل رتبه‌بندی برمبنای مقدار فاصله انجام گیرد. سپس نمونه آزمایش برمبنای تعداد k به ردهٔ Y یا N اختصاص داده می‌شود. اگر k=3 باشد در این حالت [Y=۱]، [Y=۲] و [Y=۳] است. لذا ردهٔ نمونه آزمایشی برابر Y است. زیرا تعداد Y بیشتر از N است.

۴. ارزیابی و نتایج

در این بخش ارزیابی و نتایج مدل پیشنهادی برروی مجموعه داده Spambase با ٤٦٠١ نمونه و ۵۷ ویژگی که شامل ۳۹,٤٪ ایمیل هرزنامه و ٦۰,۰٦٪ ایمیل غیرهرزنامه است در محیط برنامه‌نویسی ویژوال سی‌شارپ ۲۰۱۷ انجام شده است. مجموعه داده Spambase شامل ۵۷ ویژگی عددی است. بعضی از این ویژگی‌ها مقدار نزدیک به همدیگر دارند و اگر ویژگی‌های مشابه به الگوریتم طبقه‌بندی داده شوند، دقت تشخیص بیشتر خواهد بود. زیرا دسته‌ها، شامل نمونه‌های دقیق‌تری خواهند بود. منظور این است که فاکتورهای درست مثبت، درست منفی، کاذب مثبت و کاذب منفی در دسته‌های متعلق به خودشان قرار می‌گیرند. ویژگی‌ها توسط الگوریتم بهینه‌سازی علف‌های هرز مهاجم انتخاب می‌شوند و ویژگی‌هایی که در تکرارهای برنامه بیشترین دقت صحت را داشته باشند

جدول۲: نمایش نمونه‌های آموزش و آزمایش

داده‌ها	ردیف	a	b	ردهٔ	فاصله	رتبه‌بندی
داده‌های آموزشی	۱	۱۱	۵	Y	SQRT(12-11)^2+(4-5)^2=1.41	۱
	۲	۱۲	۶	Y	SQRT(12-12)^2+(4-6)^2=1.41	۲
	۳	۱۳	۷	N	SQRT(12-13)^2+(4-7)^2=3.16	۷
	۴	۱۵	۴	N	SQRT(12-15)^2+(4-4)^2=3	۵
	۵	۱۴	۵	N	SQRT(12-14)^2+(4-5)^2=2.23	۴
	۶	۱۲	۸	Y	SQRT(12-12)^2+(4-8)^2=4	۸
	۷	۱۳	۵	Y	SQRT(12-13)^2+(4-5)^2=1.41	۳
	۸	۱۶	۶	N	SQRT(12-16)^2+(4-6)^2=4.47	۱۰
	۹	۱۲	۷	N	SQRT(12-12)^2+(4-7)^2=3	۶
	۱۰	۱۳	۸	Y	SQRT(12-13)^2+(4-8)^2=4.12	۹
داده آزمایش	۱	۱۲	۴	Y	-	

جدول۳: ارزیابی مدل پیشنهادی با انتخاب ویژگی‌های مختلف

مقدار صحت	ویژگی‌های انتخاب شده	تعداد ویژگی
۸۹،۰۳	۹، ۳۵، ۷، ۵۲، ۵۳، ۲۱، ۲۳، ۲۵، ۳۷، ۵۷، ۸، ۲۰، ۵۵، ۶، ۱۸، ۳۰، ۲، ۱۶، ۱۹، ۵۶، ۵، ۲۴، ۲۷، ۳، ۱۱، ۲۶، ۱۷، ۱۳	۲۸
۹۰،۲۳	۱۶، ۱۹، ۵۶، ۵، ۲۴، ۲۷، ۳، ۲۳، ۲۵، ۱۷، ۳۷، ۵۷، ۸، ۲۰، ۵۵، ۶، ۱۱، ۲۶، ۷، ۵۲، ۵۳، ۲۱، ۱۸، ۳۰، ۲، ۹، ۳۵، ۱۲، ۱۵، ۲۸،۴، ۱۰، ۴۸، ۵۰، ۲۲	۳۵
۹۰،۰۶	۵۷، ۸، ۲۰، ۵۵، ۶، ۱۸، ۳۰، ۲، ۹، ۳۵، ۱۲، ۱۵، ۲۸، ۱، ۲۹، ۴۶، ۱۰، ۱۳، ۳۱، ۴۸، ۳۲، ۴۵، ۱۴، ۳۳، ۲۲،۷، ۵۲، ۵۳، ۲۱، ۲۳، ۲۵، ۱۶، ۱۹، ۵۶، ۵، ۲۴، ۲۷، ۳، ۱۱، ۲۶، ۱۷، ۳۷	۴۲
۹۲،۸۷	۲۶، ۱۷، ۳۷، ۵۷، ۸، ۲۰، ۵۵، ۶، ۱۸، ۳۰، ۲، ۹، ۵۴، ۴، ۱۴، ۲۲، ۳۶، ۳۹، ۷، ۵۲، ۳۵، ۱۲، ۱۵، ۲۸، ۱، ۲۹، ۴۶، ۱۰، ۱۳، ۳۱، ۳۲، ۴۰، ۴۲، ۳۳، ۴۵، ۵۳، ۲۱، ۲۳، ۲۵، ۲۶، ۱۹، ۵۶، ۵، ۲۴، ۲۷، ۳، ۱۱	۴۷
۹۱،۳۷	۱۶، ۱۹، ۵۶،۲۸، ۱، ۲۹، ۴۶، ۱۰، ۱۳، ۳۱، ۳۲، ۴۰، ۴۲، ۳۳، ۴۵، ۵۴، ۴۱، ۴۳، ۴۴، ۴۸، ۴۹، ۷، ۵۲، ۵۳، ۲۱، ۲۳، ۲۵، ۴، ۱۴، ۲۲، ۳۶، ۳۹، ۵، ۲۴، ۲۷، ۳، ۱۱، ۲۶، ۱۷، ۳۷، ۵۷، ۸، ۲۰، ۵۵، ۶، ۱۸، ۳۰، ۲، ۹، ۳۵، ۱۲، ۱۵	۵۲

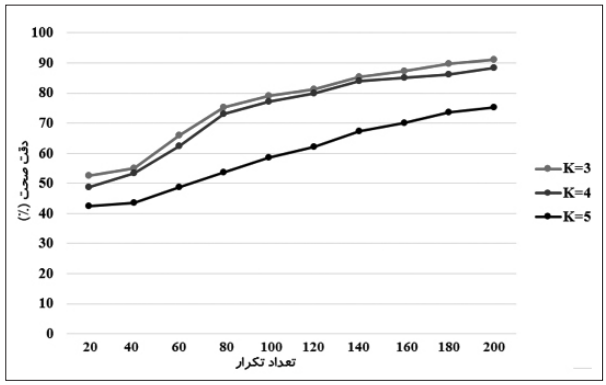
انتخاب می‌شود.

حداکثر تعداد تکرار در الگوریتم بهینه‌سازی علف‌های هرز مهاجم برابر ۲۰۰ است و مقدار K در K نزدیک‌ترین همسایه با سه مقدار مختلف تست شده است. در ابتدا مدل پیشنهادی را با انتخاب ویژگی‌های مختلف ارزیابی کردیم و با هر تعداد ویژگی نتایج متفاوتی حاصل شده است. در جدول(۳)، ارزیابی مدل پیشنهادی با انتخاب ویژگی‌ها مختلف برمبنای معیار صحت نشان داده شده است.

در جدول (٤)، مدل پیشنهادی با تعداد تکرار الگوریتم بهینه‌سازی علف‌های هرز مهاجم و مقدار K در مدل K نزدیک‌ترین همسایه برای ۵۷ ویژگی ارزیابی شده است.

الگوریتم الگوریتم بهینه‌سازی علف‌های هرز مهاجم به تعداد ۲۰۰ تکرار اجرا می‌گردد. همان‌طور که مشاهده می‌کنید تعداد تکرار الگوریتم بهینه‌سازی علف‌های هرز مهاجم در دقت تشخیص خیلی موثر است و با افزایش تعداد تکرار مقدار معیار صحت بیشتر شده است. دلیل این‌که با تعداد تکرار، مقدار صحت افزایش می‌یابد این است که الگوریتم بهینهسازی علف‌های هرز مهاجم در یافتن ویژگی‌ها در فضای جستجو دقت بیشتری حاصل می‌کند و از گیرافتادن در بهینگی محلی فرار می‌کند.

در شکل(۳)، نمودار مقایسه تعداد تکرار در الگوریتم بهینه‌سازی علف‌های هرز مهاجم و مقدار K در مدل K نزدیک‌ترین همسایه نشان داده شده است. در شکل(۳)،



شکل ۳: نمودار مقایسه تعداد تکرار در الگوریتم بهینه‌سازی علف‌های هرز مهاجم و مقدار K در مدل K نزدیک‌ترین همسایه

نمایان است که مقدار k=۳ در مقایسه با k=٤ و k=٥ بیشتر است. دلیل این‌که درصد معیار صحت در k=۳ بیشتر است برای این‌که فاصله بین ویژگی‌ها برای طبقه‌بندی بهتر شناسایی می‌شود.

در جدول(٥)، مقایسه مدل پیشنهادی با مدل‌های دیگر نشان داده شده است. همان‌طور که مشاهده می‌کنید مدل پیشنهادی در مقایسه با مدل‌های بهینه‌سازی اجتماع ذرات، الگوریتم انتخاب منفی، بیز ساده، ماشین بردار پشتیبان، انتخاب ویژگی متمایز-ماشین بردار پشتیبان، شبکه عصبی مصنوعی و فازی مقدار صحت بیشتری دارد. در میان مدل‌های جدول(٥)، مقدار صحت در مدل ماشین بردار پشتیبان به مدل پیشنهادی نزدیک‌تر و صحت در مدل بهینه‌سازی اجتماع ذرات-الگوریتم انتخاب منفی با اندکی جزئی بیشتر از مدل پیشنهادی است.

در جدول(٤) نشان دادیم که مقدار صحت با مقادیر مختلف K نتایج متفاوتی دارد. طبق نتایج این مقاله، به این نتیجه دست یافتیم که تعداد ویژگی‌ها، مهم‌ترین فاکتور در افزایش دقت صحت می‌باشند. انتخاب ویژگی‌های موثر به الگوریتم K نزدیک‌ترین همسایه کمک می‌کنند تا برمبنای فاصله بتواند دسته‌بندی دقیقی انجام دهد.

۵. نتیجه‌گیری و کارهای آینده

در دنیای امروزی با پیشرفت علوم کامپیوتری باید برای موارد ناشناخته در محیط‌های برخط و هجوم ایمیل

جدول ۴: ارزیابی مدل پیشنهادی با فاکتورهای تعداد تکرار و مقدار K

تعداد تکرار	مقدار K		
	K=۵	K=۴	K=۳
۲۰	۴۲,۳۱	۴۸,۶۷	۵۲,۶۴
۴۰	۴۳,۴۰	۵۳,۴۱	۵۵,۰۲
۶۰	۴۸,۶۲	۶۲,۴۰	۶۵,۹۴
۸۰	۵۳,۶۱	۷۳,۰۱	۷۵,۱۳
۱۰۰	۵۸,۵۰	۷۷,۲۵	۷۹,۰۶
۱۲۰	۶۲,۰۴	۷۹,۸۴	۸۱,۳۵
۱۴۰	۶۷,۲۸	۸۳,۹۷	۸۵,۳۰
۱۶۰	۷۰,۰۰	۸۵,۰۹	۸۷,۱۸
۱۸۰	۷۳,۴۹	۸۶,۰۴	۸۹,۶۲
۲۰۰	۷۵,۳۲	۸۸,۳۲	۹۱,۱۱

جدول ۵: مقایسه مدل پیشنهادی با مدل‌های دیگر برمبنای صحت

مدل‌ها	مقدار صحت
بهینه‌سازی اجتماع ذرات-الگوریتم انتخاب منفی [۱۷]	۹۱,۲۲
بهینه‌سازی اجتماع ذرات [۱۷]	۸۱,۳۲
الگوریتم انتخاب منفی [۱۷]	۶۸,۸۶
بیز ساده [۱۸]	۷۹,۳
ماشین بردار پشتیبان [۱۹]	۹۰
انتخاب ویژگی متمایز-ماشین بردار پشتیبان [۲۰]	۷۱
شبکه عصبی مصنوعی [۲۱]	۸۶
مجموعه فازی [۲۲]	۸۶,۹
مدل پیشنهادی	۹۱,۱۱

هرزنامه، مدل‌هایی طراحی و پیاده‌سازی کرد که توانایی تشخیص مفید و نامناسب بودن اطلاعات را داشته باشند. معمولاً شاهد بوده‌ایم که الگوریتم‌های یادگیری ماشین در تشخیص و طبقه‌بندی کارایی و دقت بالایی داشته‌اند. در این مقاله برای طبقه‌بندی ایمیل هرزنامه از مدل ترکیبی الگوریتم بهینه‌سازی علف‌های هرز مهاجم و K نزدیک‌ترین همسایه استفاده شد. ما نشان دادیم که در مدل پیشنهادی با تعداد ویژگی‌های مختلف، مقدار صحت نتایج متفاوتی دارد. برای مدل پیشنهادی مقدار صحت برابر ۹۱,۱۱٪ به‌دست آمد و در مقایسه با مدل‌هایی مانند ماشین بردار پشتیبان

و بیز ســـاده دقت صحت بیشتری داشت. مهم‌ترین مشکل در تشـــخیص ایمیل هرزنامه، انتخاب ویژگی‌های مناسب از میان انبوهی ویژگی‌ها اســـت که در آینده امیدواریم این مشکل را با الگوریتم‌های متنوع فراابتکاری بهینه‌تر کنیم و به دقت ۱۰۰٪ دست یابیم.

مراجع

Q. Fu, B. Feng, D. Guo, Q. Li, Combating the evolving spammers in online social networks, Computers & Security, Vol. 72, pp. 60-73, 2018.

Y. Cohen, D. Hendler, A. Rubin, Detection of malicious webmail attachments based on propagation patterns, Knowledge-Based Systems, in press, corrected proof, Available online 11 November 2017.

S. Rathore, V. Loia, J.H. Park, SpamSpotter: An efficient spammer detection framework based on intelligent decision support system on Facebook, Applied Soft Computing, In press, corrected proof, Available online 22 September 2017.

L. Nguyen; A.Q. Tran; L.T. Bui, DMEA-II and its application on spam email detection problems, Seventh IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), IEEE, pp. 1-6, 2014.

S. Salehi, A. Selamat; M. Bostanian, Enhanced genetic algorithm for spam detection in email, IEEE 2nd International Conference on Software Engineering and Service Science, pp. 594-597, 2011

M. Takesue, Cascaded Simple Filters for Accurate and Lightweight Email-Spam Detection, Fourth International Conference on Emerging Security Information, Systems and Technologies, pp. 160-165, 2010.

W. Ma, D. Tran; D. Sharma, A Novel Spam Email Detection System Based on Negative Selection, Fourth International Conference on Computer Sciences and Convergence Information Technology, IEEE, pp. 987-992, 2009

https://archive.ics.uci.edu/ml/datasets/Spambase

A.R. Mehrabian, C. Lucas, A novel numerical optimization algorithm inspired from weed colonization, Ecol. Inform. 1(4): 355-366, 2006

Martin, Instance-Based Learning: Nearest Neighbor with Generalization, Doctoral dissertation, University of Waikato, 1995

El-Sayed M. El-Alfy, R.E. Abdel-Aal, Using GMDH-based networks for improved spam detection and email feature analysis, Applied Soft Computing, Vol. 11, Issue 1, pp. 477-488, 2011.

I. Idris, A. Selamat, N.T. Nguyen, S. Omatu, O. Krejcar, K. Kuca, M. Penhaker, A combined negative selection algorithm–particle swarm optimization for an email spam detection system, Engineering Applications of Artificial Intelligence, Vol. 39, pp. 33-44, 2015

Mu-Chun Su, Hsu-Hsun Lo, Fu-Hau Hsu, A neural tree and its application to spam e-mail detection, Expert Systems with Applications, Vol. 37, 7976-7985, 2010

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.

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.

I. Idris, A. Selamat, S. Omatu, Hybrid email spam detection model with negative selection algorithm and differential evolution, Engineering Applications of Artificial Intelligence, Vol. 28, 97-110, 2014

I. Idris, A. Selamat, Improved email spam detection model with negative selection algorithm and particle swarm optimization, Applied Soft Computing, Vol. 22, pp. 11-27, 2014

Y. Zhang, H.Y. Li, M. Niranjan, P. Rockett, Applying cost-sensitive multi objective genetic programming to feature extraction for spam e-mail filtering, Lecture Notes in Computer Science, Genetic Programming, Berlin/Heidelberg, Springer, Vol. 4971, pp. 325-336, 2008

T. Fagbola, S. Olabiyisi, A. Adigun, Hybrid GA-SVM for efficient feature selection in e-mail classification, Comput. Eng. Intell. Syst, 3 (3), 2012.

A.K. Uysal, S. Gunal, A novel probabilistic feature selection method for text classification, Knowl.-Based Syst, Vol. 36, pp. 226-235, 2012.

L. Ozgur, T. Gungor, F. Gurgen, Spam mail detection using artificial neural network and Bayesian filter, in: Z. Yang, H. Yin, R. Everson (Eds.), Intelligent Data Engineering, and Automated Learning- IDEAL 2004, Springer, Berlin/Heidelberg, 2004, pp. 505-510, 2004.

R. Ariaeinejad, A. Sadeghian, Spam detection system: A new approach based on interval type-2 fuzzy sets, in: 24th Canadian Conference on Electrical and Computer Engineering (CCECE, 2011), 2011.