دوره دوازدهم، شماره دوم، پاییز و زمستان ۱۳۹۸

مجله فناوری اطلاعات در طراحی مهندسی

Information Technology in Engineering Design

http://ited.sinaweb.net

بهبود عملکرد دسته‌بند k نزدیک‌ترین همسایه با الگوریتم بهینه‌سازی ازدحام گربه‌ها برای تشخیص ایمیل‌های هرزنامه

مهدی درستی [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 : ۱۳۹۸/۴/۱۸ تاریخ پذیرش : ۱۳۹۹/۷/۱۶)

چکیده

با گسترش اینترنت و شبکه‌های آنلاین و پیوستن کاربران به این شبکه‌ها باعث شده است که هر تبلیغی بر روی آن‌ها انجام شود و نظر کاربران را با روش‌های گوناگون جلب کنند. مهم‌ترین روشی که از آن به‌عنوان روشی برای تبلیغ استفاده می‌شود، ایمیل است. کاربران از طریق هرز نامه‌نویس‌ها با انبوهی از ایمیل هرزنامه‌ها مواجه هستند و زمان و حافظه آن‌ها با این ایمیل‌ها مختل می‌شود. یک مشکل دیگر در ایمیل هرزنامه این است که برخی از هرز نامه‌نویس‌ها از ایمیل هرزنامه به‌منظور اهداف امنیتی و دزدی هویتی استفاده می‌کنند و با این روش وارد کامپیوتر کاربر و حساب‌های شخصی کاربر می‌شوند. لذا چالش اصلی در تشخیص ایمیل هرزنامه، تفکیک ایمیل‌های هرزنامه از غیر هرزنامه است. برای حل این مشکل باید ویژگی‌های ایمیل مانند عنوان، متن و کاراکترها تشخیص داده شود. در این مقاله مدل ترکیبی برمبنای k نزدیک‌ترین همسایه و الگوریتم بهینه‌سازی اجتماع گربه برای تشخیص ایمیل هرزنامه استفاده‌شده است. از الگوریتم بهینه‌سازی اجتماع گربه برای انتخاب ویژگی و جستجو در فضای بردارهای ویژگی و از k نزدیک‌ترین همسایه برای طبقه‌بندی داده‌ها استفاده‌شده است. نتایج ارزیابی بر روی مجموعه داده *Spambase* نشان داده که دقت تشخیص مدل پیشنهادی برمبنای ۲۰ بار تکرار برابر با ۹۷.۶۱ درصد می‌باشد.

واژه‌های کلیدی: تشخیص ایمیل هرزنامه، طبقه‌بندی، k نزدیک‌ترین همسایه، الگوریتم بهینه‌سازی اجتماع گربه

* عهده‌دار مکاتبات:

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

شماره تلفن: ۰۹۱۴۱۷۶۴۴۲۷ **پست الکترونیکی:** *farhad@iaurmia.ac.ir*

۱. مقدمه

ایمیل یکی از این فنّاوری‌های برتر ارتباطی است. ایمیل درواقع یک آدرس پستی می‌باشد که در فضای مجازی و به‌صورت الکترونیکی، مسئولیت برقراری ارتباط دوسویه بین فرستنده و گیرنده پیام را دارد. و به دلیل مزیت بسیار مهم خود، به یکی از فنّاوری‌های برتر در عرصه روابط آنلاین تبدیل‌شده است. این مزیت همان برقراری ارتباط با سرعت بسیار بالا بین مخاطبان و شرکت‌ها است. این ارتباط کاملاً دوسویه می‌باشد به‌نحوی‌که مخاطب از طریق آدرس پست الکترونیکی شرکت یا سازمان به ابراز عقاید و نظرات خود در مورد آن پیام می‌پردازد. اما همگام با پیشرفت ایمیل و دنیای اینترنت، هکرها هم برای سرقت اطلاعات کاربران و تبلیغات بی‌اساس از ایمیل به‌عنوان یک‌راه سریع و کم‌هزینه استفاده می‌کند. راهکار آن‌ها برای نفوذ به داده‌های کاربران استفاده از ایمیل هرزنامه است [۱]. هرزنامه به ایمیل‌هایی گفته می‌شود که به‌قصد مزاحمت، تبلیغات یا پخش ویروس ارسال می‌شوند. هرروز با حجم انبوهی از ایمیل هرزنامه از سوی کاربران و سایت‌های ناشناس مواجه هستیم. ایمیل هرزنامه یک معضل ناخوشایند برای کاربرانی که مداوم از اینترنت و ایمیل استفاده می‌کنند شده است.

ارسال ایمیل از طریق سرویس‌های اینترنتی رایگان نظیر جیمیل و سرویس ایمیل یاهو، به یک امر رایج برای ارتباط بین کاربران معمولی تبدیل‌شده است. و حتی افراد سرشناس، سیاستمداران، کارمندان و مدیران مشاغل و کسب‌وکارهای حساس نیز از این سرویس‌ها برای ارسال پیام استفاده می‌کنند. قطعاً استفاده از این سرویس‌های رایگان، بسیار آسان است و خدمات خوبی نیز در اختیار کاربران قرار می‌گیرد، اما اطلاعات شخصی کاربران ممکن است در اختیار بنگاه‌های تبلیغاتی، هکرهای خرابکار، سارقان دیتا، سیستم‌های جاسوسی و ابرقدرت‌های پشت پرده‌ی دنیای تکنولوژی قرار بگیرد که نتیجه آن ارسال ایمیل هرزنامه به ایمیل کاربران است [۲].

امروزه ایمیل به یکی از رایج‌ترین ابزار ارتباطی در زندگی روزمره بشر تبدیل‌شده است. اما عمومیت و سادگی استفاده از ایمیل باعث شده تا مورداستفاده هرز نامه‌نویس‌ها و کلاه‌برداران اینترنتی قرار بگیرد. ازآنجایی‌که ارسال میلیونی هرزنامه باعث آزار کاربر، اتلاف زمان، هزینه، منابع شبکه و پهنای باند می‌شود. بنابراین هرزنامه به یک مشکل حیاتی تبدیل‌شده است. لذا روش‌ها و الگوریتم‌های زیادی برای جلوگیری و فیلتر هرزنامه پیشنهادشده است. هدف اصلی افزایش دقت روش تشخیص هرزنامه در ایمیل است. در این مقاله عمل تشخیص هرزنامه با استفاده از ترکیب k نزدیک‌ترین همسایه [۳] و الگوریتم بهینه‌سازی ازدحام گربه‌ها [۴] مورد ارزیابی قرار خواهد گرفت.

مدل k نزدیک‌ترین همسایه [۳] برای دسته‌بندی نمونه‌های داده‌ها استفاده می‌شود و بر اساس این اصل است که یک نمونه با k نمونه که خصوصیات مشابه بیشتری باهم دارند، دسته‌بندی می‌شود. به این صورت که k تا از نمونه‌های نزدیک بر مبنای فاصله برای نمونه جدید شناسایی می‌شوند و کلاسی که بیشترین شباهت را در میان این نمونه‌ها دارد به‌عنوان کلاس نتیجه برای نمونه جدید شخص می‌شود. بنابراین باید معیاری، برای تعیین فاصله بین نمونه‌ها مشخص شود.

الگوریتم بهینه‌سازی اجتماع گربه [۴] یکی از الگوریتم‌های فرا ابتکاری است که از جستجوی گروهی گربه‌ها برای یافتن غذا الهام گرفته‌شده است. در این الگوریتم یک جستجوی مبتنی بر جمعیت وجود دارد که در آن هر گربه باگذشت زمان موقعیت خود را برای جستجوی غذا تغییر می‌دهد. در الگوریتم بهینه‌سازی اجتماع گربه، گربه‌ها در یک فضای جستجوی دوبعدی از راه‌حل‌های ممکن مسئله، حرکت می‌کنند. در فضای جستجو یک معیار ارزیابی تعریف می‌شود و سنجش کیفیت راه‌حل‌های مسئله از طریق آن صورت می‌پذیرد. تغییر حالت هر گربه در یک گروه تحت تأثیر تجربه‌های خود و یا دانش همسایگانش بوده و رفتار جستجوی یک

گربه در گروه تحت تأثیر گربه‌های دیگر است. همین رفتار ساده باعث پیدا شدن ناحیه‌های بهینه از فضای جستجو می‌گردد.

ایمیل‌های ناشناس، تبلیغاتی، مخرب و حتی ایمیل‌های جعلی برای سرقت اطلاعات کاربران ازجمله مواردی هستند که به آن‌ها هرزنامه گفته می‌شود [۵]. هرزنامه‌نویسان از طریق ارسال ایمیل می‌توانند به اهداف خود نظیر دسترسی به اطلاعات شخصی کاربران، انتشار ویروس و... دست یابند [۶]. در این مقاله از ترکیب k نزدیک‌ترین همسایه و الگوریتم بهینه‌سازی ازدحام گربه برای تشخیص ایمیل هرزنامه استفاده می‌کنیم. در مدل ترکیبی از الگوریتم بهینه‌سازی ازدحام گربه برای انتخاب ویژگی‌ها [۷][۸] و از k نزدیک‌ترین همسایه برای طبقه‌بندی ویژگی‌ها استفاده می‌شود. به دلیل اینکه برای تشخیص ایمیل هرزنامه با ویژگی‌های زیادی مواجه هستم از الگوریتم بهینه‌سازی ازدحام گربه استفاده برای انتخاب ویژگی استفاده می‌کنیم.

۲- کارهای قبلی

هرزنامه‌ها صرفاً فقط به‌عنوان زباله نیستند و ازآنجایی‌که شامل فایل‌های پیوستی مانند ویروس و عوامل نرم‌افزاری جاسوسی هستند می‌توانند برای یک سیستم و دریافت‌کنندگان آن خطرناک باشند و باعث از بین رفتن اطلاعات باشند. بنابراین ما نیاز به ابزارهای جهت تشخیص هرزنامه داریم. بسیاری از تکنیک‌های تشخیص هرزنامه‌ها بر اساس روش‌های یادگیری ماشین پیشنهادشده است.

مدل‌های ماشین بردار پشتیبان، درخت J48، شبکه بیزین و Lazy IBK که از تکنیک‌های داده‌کاوی هستند بر روی ۴۳۰۷ ایمیل (۶۳۵ ایمیل هرزنامه) تست و اجراشده‌اند [۹]. ارزیابی در محیط وکا انجام‌شده است. نتایج نشان داده که درصد صحت در J48 برابر ۹۳.۱۳ است که در مقایسه با مدل‌های دیگر بیشتر است. مدل ترکیبی NB-K-Means 10]] 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 K-Means 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 [۱۱]. از الگوریتم بهینه‌سازی اجتماع ذرات برای انتخاب ویژگی‌ها و از شبکه عصبی مصنوعی 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 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(۴۸۱ ایمیل هرزنامه و ۲۱۷۱ ایمیل غیر هرزنامه) و مجموعه داده Spam-Assassin با ۶۰۰۰ 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 و ۹۹.۷۹ 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 پایه و شبکه عصبی پس انتشار دقت تشخیص بهتری دارد.

مدل‌های درخت تصمیم‌گیری، ماشین بردار پشتیبان و شبکه عصبی پس انتشار و ترکیب آن‌ها بر روی دو مجموعه داده با ۱۴ ویژگی تست و اجر شده‌اند [۱۲]. مجموعه داده اولی شامل ۵۰۴ ایمیل (۳۳۶ غیر هرزنامه و ۱۶۸ هرزنامه) و مجموعه داده دومی شامل ۶۵۷ ایمیل(۳۸۷ غیر هرزنامه و ۲۷۰ هرزنامه) است. در مدل درخت تصمیم‌گیری از آنتروپی، مدل 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 تصمیم‌گیری، ماشین بردار پشتیبان و شبکه عصبی پس انتشار به ترتیب برابر ۸۹.۸۸، ۸۸.۶۹ و ۸۹.۸۸ بوده است و بر روی مجموعه داده دومی درصد صحت در مدل ترکیبی برابر ۹۱.۷۸ و در مدل‌های درخت تصمیم‌گیری، ماشین بردار پشتیبان و شبکه عصبی پس انتشار به ترتیب برابر ۹۰.۸۷، ۹۰.۸۷ و ۸۹.۰۴ بوده است.

ماشین بردار پشتیبان برمبنای تابع شعاعی پایه برای تشخیص ایمیل هرزنامه پیشنهادشده است [۱۳]. مدل شبکه تابع شعاعی پایه 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 پایه فضای ویژگی از طریق یک تبدیل غیرخطی به دست می‌آید. برای ارزیابی مدل ماشین بردار پشتیبان برمبنای تابع شعاعی پایه از مجموعه داده‌های Spambase با ۴۶۰۱ نمونه و SMS Spam با ۵۵۷۴ نمونه استفاده‌شده است. نتایج نشان داده که مدل ماشین بردار پشتیبان برمبنای تابع شعاعی پایه نمونه‌های هرزنامه را با دقت ۹۳.۹۲ تشخیص داده است.

الگوریتم سیستم ایمنی مصنوعی که از الگوریتم‌های مبتنی بر جمعیت است برای تشخیص ایمیل هرزنامه بر روی مجموعه داده TREC07 که شامل ۷۵۴۱۹ ایمیل است تست و اجراشده است [۱۴]. از الگوریتم سیستم ایمنی مصنوعی برای انتخاب ویژگی برمبنای تکرار و تست استفاده‌شده است. نتایج نشان داده که دقت مدل بالا ۸۰ درصد بوده است. مدل [15 RAN-LSH] که ترکیبی از شبکه تابع شعاعی پایه و ماشین بردار پشتیبان است برای تشخیص ایمیل هرزنامه پیشنهادشده است. ارزیابی بر روی دو مجموعه داده که در سال ۲۰۱۳ گردآوری‌شده‌اند و هرکدام شامل ۸۹۴۸ (۱۳۱۱ ایمیل هرزنامه و ۷۶۳۷ ایمیل غیر هرزنامه) و ۲۰۹۰۵ (۸۸۶۳ ایمیل و ۱۲۰۴ ایمیل غیر هرزنامه) نمونه ایمیل هرزنامه هستند انجام‌شده است. نتایج نشان داده که مدل RAN-LSH در پیش‌بینی دقت مناسبی دارد. به دلیل اینکه مجموعه دادها شامل ویژگی‌های زیادی هستند از انتخاب ویژگی استفاده‌شده است. در مدل RAN-LSH برای انتخاب ویژگی‌ها از ماشین بردار پشتیبان و برای آموزش و تست از شبکه تابع شعاعی پایه استفاده‌شده است. همچنین برای انتخاب ویژگی‌های برازنده از لایه وسطی مدل RAN-LSH استفاده می‌شود.

روش درخت رگرسیون تجمیعی بیزین [۱۶] که برمبنای رگرسیون کار می‌کند بر روی Ling-Spam با ۲۸۹۳ نمونه، PU1 با ۱۰۹۹ نمونه و Spambase با ۴۶۰۱ نمونه تست و اجراشده است. همچنین از مدل‌های طبقه‌بندی به‌منظور تخمین دقت استفاده‌شده است. دقت مدل درخت رگرسیون تجمیعی بیزین برای سه مجموعه داده به ترتیب برابر ۱۰۰٪، ۱۰۰٪ و ۹۷.۶۰٪ بوده است.

طبقه‌بندی بیزین [۱۷] برای تشخیص ایمیل هرزنامه بر روی سه مجموعه داده که شامل ۱۰۰۰، ۱۵۰۰ و ۲۰۰۰ ایمیل هستند اجرا و تست‌شده است. در طبقه‌بندی بیزین از روابط احتمالاتی استفاده می‌شود. نتایج نشان داده دقت تشخیص برای سه مدل بیشتر از ۹۳ درصد بوده است. برای ارزیابی از معیارهای صحت، دقت و بازخوانی استفاده‌شده است. درصد صحت به ترتیب برای سه مجموعه داده برابر ۹۳.۹۸، ۹۴.۸۵ و ۹۶.۴۶ بوده است. الگوریتم سیستم ایمنی مصنوعی به‌منظور تشخیص ایمیل هرزنامه بر روی ۴۶۰۱ نمونه تست و اجراشده است [۱۸]. برای محاسبه تشابه بین ویژگی‌ها از فاصله اقلیدسی استفاده‌شده است. نتایج نشان داده که دقت تشخیص برابر ۹۴.۲۸ درصد بوده است.

مدل k نزدیک‌ترین همسایه که یکی از مدل‌های طبقه‌بندی است بر روی مجموعه داده TREC2005 با ۱۰۰۰ نمونه ایمیل تست و اجراشده است [۱۹]. برای تست از k با مقادیر ۱، ۳، ۵، ۷ و ۹ استفاده‌شده است. نتایج نشان داده که دقت تشخیص در مدل k نزدیک‌ترین همسایه برابر ۹۲ درصد بوده است. برای تشخیص ایمیل هرزنامه از مدل فازی برای تست و آموزش ۱۸۹۵ نمونه ایمیل استفاده‌شده است [۲۰]. در ابتدا داده‌ها پیش‌پردازش و سپس فراوانی

آن‌ها مشخص‌شده است. در مدل فازی به هر ویژگی یک مقدار عضویت به‌منظور رده‌بندی نسبت داده می‌شود. نتایج نشان داده که دقت تشخیص برابر ۸۶.۹ درصد بوده است. در جدول(۱) مقایسه مدل‌های پیشنهادشده برای تشخیص ایمیل هرزنامه برمبنای روش‌های مختلف نشان داده‌شده است.

جدول ۱ مقایسه مدل‌های پیشنهادشده برای تشخیص ایمیل هرزنامه برمبنای روش‌های مختلف

دقت تشخیص	انتخاب ویژگی	تعداد	مجموعه داده	مدل‌ها	رفرنس
۹۳.۳۱	X	۴۳۰۷	مجموعه داده هرزنامه	ماشین بردار پشتیبان	[۹]
۸۸.۳۹				J48	
۹۳.۰۸				شبکه بیزین	
۸۹.۲۳				Lazy IBK	
۸۸.۳۳	√	۲۵۸۹	Ling-Spam	NB-K-Means	[۱۰]
۹۱.۱۲		۶۰۰۰	Assassin		
۹۶.۵		۴۶۰۱	Spambase		
۹۹.۷۹	√	۲۵۸۹	Ling-Spam	بهینه‌سازی اجتماع ذرات – شبکه عصبی مصنوعی چندلایه	[۱۱]
۹۹.۹۸		۶۰۰۰	Assassin		
۸۹.۸۸	X	۵۰۴	مجموعه داده هرزنامه	درخت تصمیم‌گیری	[۱۲]
۸۸.۶۹				ماشین بردار پشتیبان	
۸۹.۸۸				شبکه عصبی پس انتشار	
۹۰.۸۷				درخت تصمیم‌گیری	
۹۰.۸۷				ماشین بردار پشتیبان	
۸۹.۰۴				شبکه عصبی پس انتشار	
				شبکه عصبی پس انتشار با ماشین بردار پشتیبان	
۹۳.۹۲	√	۴۶۰۱	Spambase	ماشین بردار پشتیبان	[۱۳]
		۵۵۷۴	SMS Spam		
۸۰.۰۰	√	۷۵۴۱۹	TREC07	الگوریتم سیستم ایمنی مصنوعی	[۱۴]
۹۰.۲۱	√	۸۹۴۷	Dataset1	RAN-LSH	[۱۵]
۸۹.۵۴		۲۰۹۰۵	Dataset2		
۱۰۰.۰۰	√	۲۸۹۳	Ling-Spam	درخت رگرسیون تجمیعی بیزین	[۱۶]
۱۰۰.۰۰		۱۰۹۹	PU1		
۹۷.۶۰		۴۶۰۱	Spambase		
۹۳.۹۸	X	۱۰۰۰	Dataset1	بیزین	[۱۷]
۹۴.۸۵		۱۵۰۰	Dataset2		
۹۶.۴۶		۲۰۰۰	Dataset3		
۹۴.۲۸	√	۴۶۰۱	Spambase	الگوریتم	[۱۸]

					سیستم ایمنی مصنوعی
۹۲.۰۰	X	۱۰۰۰	TREC2005	K نزدیک‌ترین همسایه	[۱۹]
۸۶.۰۹	X	۱۸۹۵	مجموعه داده هرزنامه	منطق فازی	[۲۰]

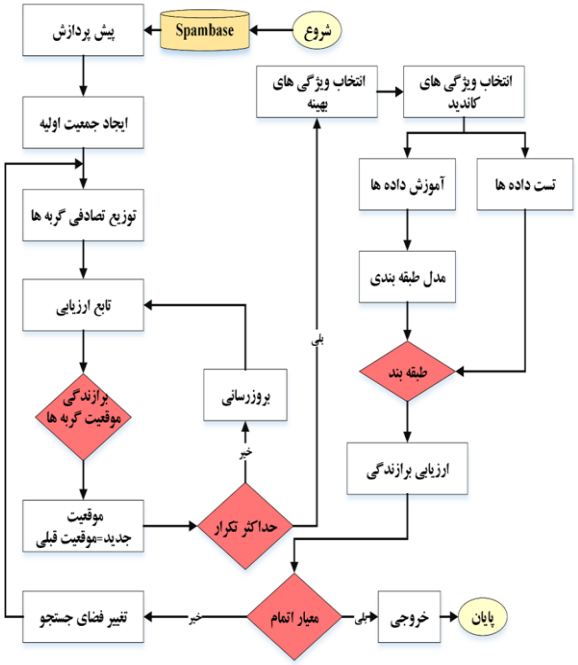
۳- مدل پیشنهادی

در این بخش، مدل پیشنهادی که بهبود k نزدیک‌ترین همسایه با الگوریتم بهینه‌سازی ازدحام گربه‌ها است توضیح داده می‌شود. در مدل پیشنهادی در ابتدا مجموعه داده Spambase [۲۱] که شامل ۵۷ ویژگی است خوانده می‌شود و سپس عمل نرمال‌سازی و تصحیح داده‌های حذف‌شده انجام می‌گیرد. در مرحله پیش پردازش ویژگی‌هایی که مقدارشان مشخص نیست با توجه به حداقل و حداکثر مقدار آن‌ها تعیین می‌شود. تعیین مقدار ویژگی‌ها طبق معادله(۱) انجام می‌شود [۲۲]. در معادله(۱)، پارامترهای x_{max} و x_{min} مربوط به بیشترین و کمترین مقدار ویژگی‌ها هستند.

$$x' = (x_{\max} - x_{\min}) \times \frac{(x_i - x_{\min})}{(x_{\max} - x_{\min})} + x_{\min} \qquad (۱)$$

در مدل پیشنهادی عمل انتخاب ویژگی برمبنای الگوریتم بهینه‌سازی اجتماع گربه انجام می‌شود. برای تشکیل جمعیت اولیه از مقدار ویژگی‌های مجموعه داده Spambase استفاده می‌شود و هر بردار برمبنای مقادیر آن مقداردهی می‌شود. بدین‌صورت که بردارهای اولیه که شامل مقادیر ویژگی‌ها هستند با استفاده از عملگرهای جستجو و به‌روزرسانی به هم نزدیک می‌شوند و هر ناحیه جستجو شامل بردارهایی با مقدار مشابه می‌باشد. برای یافتن ناحیه‌ها از برازندگی گربه‌ها استفاده می‌شود. هر گربه با جستجو در ناحیه شامل یک مقدار است که برای یافتن موقعیت جدید به آن کمک می‌کند. درصورتی‌که یافتن ناحیه‌ها به بهینه محلی ختم شود عمل به‌روزرسانی انجام می‌شود و گربه‌ها یک موقعیت جدید را پیدا می‌کنند. در شکل(۱)، فلوچارت مدل پیشنهادی نشان داده‌شده است.

شکل ۱ فلوچارت مدل پیشنهادی



در مدل پیشنهادی در ابتدا بردار v که شامل ویژگی‌ها هستند ایجاد می‌شود. هر بردار v شامل یک فضا است که هر گربه برای شروع از آن استفاده می‌کند. هنگامی‌که گربه‌ها در فضای محیط پخش شدند باید آن‌ها به نقاط مشابه هدایت شوند و ویژگی‌ها را برمبنای هرزنامه و غیر هرزنامه تشخیص دهند. برای تشخیص از فاصله موقعیتی هر گربه استفاده می‌شود. اگر گربه‌ها به سمت نقاط غیرمشابه بروند باید به‌روزرسانی در محیط انجام شود. تشخیص نقاط بهینه برمبنای مقدار قبلی هر گربه انجام می‌شود. به‌روزرسانی سرعت گربه‌ها در فضای d بعدی، طبق معادله(۲) محاسبه می‌شود.

$$v_{k,d} = v_{k,d} + r_1 \times c_1 \times (x_{best,d} - x_{k,d}), \qquad (۲)$$
$$where...d = 1,2,...,M$$

در معادله(۲)، پارامتر x_(best,d) موقعیت بهترین گربه، x_(k,d) موقعیت گربه kام و c1 و r1 مقادیر تصادفی بین [۰,۱] هستند. همچنین به‌روزرسانی موقعیت گربه kام طبق معادله(۳) محاسبه‌شده است.

$$x_{k,d} = x_{k,d} + v_{k,d} \qquad (۳)$$

به‌روزرسانی موقعیت به دلیل یافتن نقاط بهینه و رسیدن به موقعیت بهترین گربه انجام می‌شوند. موقعیت هر گربه با توجه به موقعیت بقیه گربه‌ها، به‌روزرسانی می‌شود. در شروع الگوریتم به اطلاعاتی مانند بهترین موقعیت هر گربه، سرعت و بهترین موقعیت مجاور نیاز است.

تابع ارزیابی در الگوریتم بهینه‌سازی اجتماع گربه برمبنای بردارهای ایجادشده توسط گربه‌ها محاسبه می‌شود. در الگوریتم بهینه‌سازی اجتماع گربه، هر گربه یک‌راه‌حل کاندید در قالب یک بردار است که طول هر بردار برابر با ۵۷ می‌باشد. به دلیل اینکه تعداد ویژگی‌ها در مجموعه داده Spambase برابر با ۵۷ ویژگی است. پس از ایجاد جمعیت اولیه برای ارزیابی برازندگی هر راه‌حل از تابع ارزیابی طبق معادله(۴) استفاده می‌شود.

$$X_i = (1/Sum_{x(i)}/b_{(i)}) \quad (۴)$$

در معادله(۴)، پارامتر $sum_{x(i)}$ مجموع بردار iام و $b_{(i)}$ طول بردار iام است و X_i مقدار برازندگی گربه iام را نشان می‌دهد. اگر مقدار Xi بیشتر باشد، در این صورت گربه iام توانسته گربه‌های بیشتری را در همسایگی خود داشته باشد و لذا نقاط بهینه‌ی بیشتری را در فضای مسئله کشف کرده است.

۱-۳- انتخاب ویژگی

در الگوریتم بهینه‌سازی اجتماع گربه برای تبدیل روش پیوسته به گسسته، ابتدا موقعیت هر گربه در هر بُعد را برمبنای روش پیوسته به دست می‌آوریم. سپس، این مقدار با استفاده از تابع سیگموید به مقداری بین صفر و یک تبدیل می‌شود. تابع سیگموئید، به‌منظور تعیین مقادیر موقعیت گربه‌ها به مقدار صفر یا یک برای به‌روزسانی موقعیت‌های جدید استفاده می‌شود. در اکثر موارد برای تبدیل فضای پیوسته به گسسته از تابع سیگموئید استفاده می‌شود. لذا تغییرات الگوریتم بهینه‌سازی اجتماع گربه طبق معادله(۵)، انجام می‌گیرد. تابع rand اعداد تصادفی در بازه [۰،۱] تولید می‌کند.

$$S(X_{ij}) = \frac{1}{1+e^{-x_{ij}}} \quad (۵)$$

$$if\ rand < S(X_{ij})\ then\ x_{ij} = 1\ else\ x_{ij} = 0 \quad (۶)$$

انتخاب ویژگی برمبنای بردارهای بهینه انجام می‌شود که هر بردار برای تشابه ویژگی از فاصله استفاده می‌کند. در بردار ویژگی، نقاطی انتخاب می‌شوند که ایندکس آن‌ها برابر 1 باشد و نقاط دیگر مانند 0 انتخاب نمی‌شود. به دست آوردن مقدار صفر و یک برمبنای دودویی کردن اعداد و این عمل توسط تابع سیگموئید انجام می‌گیرد. محاسبه بردارهای راه‌حل هر گربه توسط تابع ارزیابی برمبنای فاصله محاسبه می‌شود. بدین معنی که هر عنصر از بردار با مقدار همسایه خود محاسبه می‌شود. انتخاب بردار بهینه در الگوریتم بهینه‌سازی ازدحام گربه‌ها برمبنای بیشینه‌سازی بردارها انجام می‌شود. یک بردار بهینه باید مقدار بیشتری(X_i) برمبنای تابع برازندگی در مقایسه با بردارهای دیگر داشته باشد. اگر مقدار میانگین بردار بیشتر باشد نشان‌دهنده این است که فاصله بین ویژگی‌های کمتر است.

۲-۳- مدل K نزدیک‌ترین همسایه

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 به‌صورت $<a_1(x), a_2(x), ..., a_n(x)>$ داشته باشیم از فاصله اقلیدسی طبق معادله(۷) برای به دست آوردن فاصله بین دو ویژگی x_i و x_j استفاده می‌کنیم.

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2} \quad (۷)$$

در مدل k نزدیک‌ترین همسایه یک دسته باید شامل k نمونه از مجموعه نمونه‌های آموزشی باشد بطوریکه نزدیک‌ترین نمونه‌ها شامل نمونه آزمایشی باشند و بر اساس برتری دسته یا برچسب مربوط به آن‌ها در مورد نمونه آزمایشی جدید تصمیم‌گیری می‌شود. به عبارتی دسته‌ای باید انتخاب شود که در همسایگی خود بیشترین

تعداد نمونه منتسب به آن دسته وجود داشته باشد. درنتیجه، پس‌ازآنکه 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.

۳-۳- معیارهای دقت و خطا

در این بخش مهم‌ترین معیارها انتخاب‌شده است و برای ارزیابی مدل پیشنهادی از آن‌ها استفاده می‌شود [۲۴،۲۳]. به‌منظور نشان دادن دقت بر روی کلاس‌ها به‌منظور نشان دادن برازندگی از صحت استفاده می‌شود.

۴- ارزیابی و نتایج

در این بخش به ارزیابی نتایج مدل پیشنهادی بر روی مجموعه داده Spambase [21] با ۴۶۰۱ نمونه و ۵۷ ویژگی در محیط ویژوال سی‌شارپ دات‌نت ۲۰۱۷ می‌پردازیم. ارزیابی برمبنای معیارهای مختلفی مانند تعداد k در مدل k نزدیک‌ترین همسایه ، تعداد تکرار و تعداد نسل در الگوریتم بهینه‌سازی ازدحام گربه‌ها و درصد آموزش و تست داده‌ها انجام‌شده است. برای اجرای اولیه برنامه مقادیر پارامترهای k، تعداد جمعیت اولیه، تعداد تکرار و تعداد نسل به ترتیب برابر ۳، ۵۰، ۱۰۰ و ۲۰

$$P = \frac{TP}{TP+FP} \times 100 \quad \text{دقت} \quad (۸)$$

$$R = \frac{TP}{TP+FN} \times 100 \quad \text{بازخوانی} \quad (۹)$$

$$F - Measure = \frac{2 \times P \times R}{(P+R)} \quad \text{اندازه‌گیری F} \quad (۱۰)$$

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad \text{صحت} \quad (۱۱)$$

$$ErrorRate = 1 - Accuracy \quad \text{نرخ خطا} \quad (۱۲)$$

پارامتر TN بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها منفی بوده و الگوریتم دسته‌بندی نیز دسته آن‌ها را به‌درستی منفی تشخیص داده است.TP بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها مثبت بوده و الگوریتم دسته‌بندی نیز دسته آن‌ها را به‌درستی مثبت تشخیص داده است.FP بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها منفی بوده و الگوریتم دسته‌بندی دسته آن‌ها را به‌اشتباه مثبت تشخیص داده است.FN بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها مثبت بوده و الگوریتم دسته‌بندی دسته آن‌ها را به‌اشتباه منفی تشخیص داده است.

لحاظ شده‌اند و مقادیر آموزش و تست داده‌ها به ترتیب برابر ۸۰ و ۲۰ درصد است. مقادیر اولیه پارامترها در مراحل اجرا بسیار تأثیرگذار هستند لذا برمبنای اجراهای متعدد و تست‌های مختلف مقادیر ذکرشده برای مدل پیشنهادی انتخاب‌شده‌اند.

۴-۱- مدل K نزدیک‌ترین همسایه

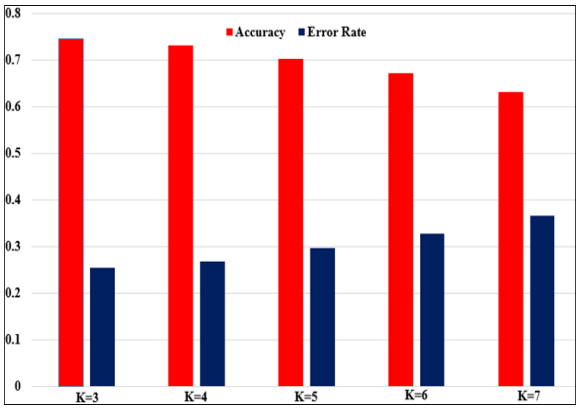
در جدول(۲)، نتایج ارزیابی مدل k نزدیک‌ترین همسایه برمبنای مقادیر مختلف k بر روی مجموعه داده Spambase نشان داده‌شده است.در جدول(۲)، k=3 و k=7 به ترتیب بیشترین و کمترین درصد صحت را دارند. به دلیل اینکه در k=3 همسایه‌های مشابه و نزدیک ارزیابی می‌شوند و لذا دقت تشخیص بیشتر است. اما در k=7 همسایه‌های دورتر ممکن است متعلق به یک دسته نباشد و به اشتباه به‌عنوان داده درست تشخیص داده شوند.

جدول ۲ ارزیابی مدل K نزدیک‌ترین همسایه برمبنای مقادیر مختلف K

معیارها	K=3	K=4	K=5	K=6	K=7
دقت	۷۲.۳۴	۷۰.۲۱	۶۸.۲۷	۶۷.۹۰	۶۵.۲۱
بازخوانی	۷۳.۶۹	۷۱.۶۲	۶۸.۸۷	۶۸.۲۷	۶۵.۶۴
اندازه‌گیری F	۷۳.۰۱	۷۰.۹۱	۶۸.۵۷	۶۸.۰۸	۶۵.۴۲
صحت	۷۴.۶۱	۷۳.۲۵	۷۰.۳۱	۶۷.۱۹	۶۳.۲۷
نرخ خطا	۰.۲۵۳۹	۰.۲۶۷۵	۰.۲۹۶۹	۰.۳۲۸۱	۰.۳۶۷۳

در شکل(۲) نمودار مدل k نزدیک‌ترین همسایه برمبنای معیارهای صحت و نرخ خطا نشان داده‌شده است. طبق شکل(۲) اگر تعداد k بیشتر باشد آنگاه درصد صحت کمتر و نرخ خطا افزایش می‌یابد.

شکل ۲ نمودار مدل K نزدیک‌ترین همسایه برمبنای معیارهای صحت و نرخ خطا



۲-۴- ارزیابی مدل پیشنهادی برمبنای تکرار

در جدول(۳)، مدل پیشنهادی با ۱۰۰ بار تکرار و برمبنای مقادیر مختلف k ارزیابی شده است. درصد صحت با k=3 برابر با ۸۹.۳۴ و با k=7 برابر با ۷۹.۸۱ درصد می‌باشد.

جدول ۳ ارزیابی مدل پیشنهادی با ۱۰۰بار تکرار

معیارها	K=3	K=4	K=5	K=6	K=7
دقت	۸۱.۳۲	۷۸.۶۵	۷۶.۹۴	۷۵.۲۴	۷۳.۱۴
بازخوانی	۸۰.۱۸	۷۹.۰۶	۷۷.۱۹	۷۶.۳۴	۷۴.۶۲
اندازه‌گیری F	۸۰.۷۵	۷۸.۸۵	۷۷.۰۶	۷۵.۷۹	۷۳.۸۷
صحت	۸۹.۳۴	۸۶.۳۲	۸۱.۶۴	۷۶.۲۴	۷۹.۸۱
نرخ خطا	۰.۱۴۶۶	۰.۱۵۶۸	۰.۱۸۳۶	۰.۲۳۷۶	۰.۲۰۱۹

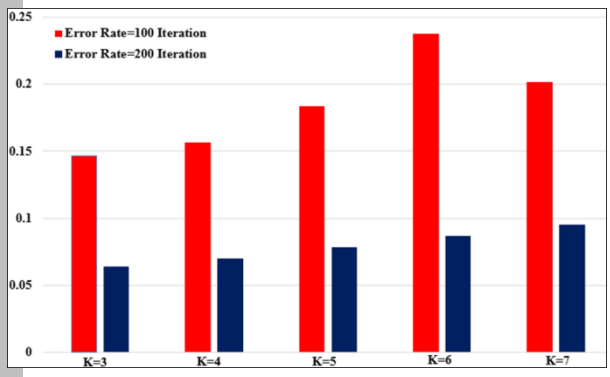
در جدول(۴)، مدل پیشنهادی با ۲۰۰ بار تکرار برمبنای مقادیر مختلف k ارزیابی شده است. با افزایش تعداد تکرار دقت تشخیص مدل پیشنهادی بیشتر شده است و درصد صحت برای ۲۰۰ بار تکرار برابر ۹۷.۶۱ است. درصد صحت با k=3 برابر با ۹۷.۶۱ و با k=7 برابر با ۹۰.۴۸ درصد می‌باشد.

جدول ۴ ارزیابی مدل پیشنهادی با ۲۰۰بار تکرار

معیارها	K=3	K=4	K=5	K=6	K=7
دقت	۸۶.۲۳	۸۵.۳۶	۸۴.۱۲	۸۳.۶۹	۸۰.۱۵
بازخوانی	۸۸.۱۱	۸۶.۲۸	۸۵.۹۴	۸۵.۵۴	۸۲.۶۴
اندازه‌گیری F	۸۷.۱۶	۸۵.۸۲	۸۵.۰۲	۸۴.۶۰	۸۱.۳۸
صحت	**۹۷.۶۱**	۹۳.۰۱	۹۲.۱۸	۹۱.۳۴	۹۰.۴۸
نرخ خطا	۰.۰۶۳۹	۰.۰۶۹۹	۰.۰۷۸۲	۰.۰۸۶۶	۰.۰۹۵۲

در شکل(۳)، نمودار مدل پیشنهادی برمبنای معیار نرخ خطا با ۱۰۰ و ۲۰۰ بار تکرار نشان داده‌شده است. همانطور که در شکل(۳) مشاهده می‌کنید دقت خطا برای مقادیر مختلف k در حالت ۲۰۰ بار تکرار کمتر است.

شکل ۳ نمودار مدل پیشنهادی برمبنای معیار نرخ خطا با ۱۰۰ و ۲۰۰ بار تکرار



۳-۴- ارزیابی مدل پیشنهادی برمبنای تعداد نسل

در جدول(۵)، مدل پیشنهادی با ۱۰۰ بار تکرار و ۱۰۰ نسل و برمبنای مقادیر مختلف k ارزیابی شده است. همانطور که در جدول(۵)، مشاهده می‌شود تعداد نسل در دقت تشخیص تأثیرگذار بوده و درصد صحت برای ۱۰۰ نسل برابر ۹۱.۳۸ درصد است.

جدول ۵ ارزیابی مدل پیشنهادی با ۱۰۰ نسل

معیارها	K=3	K=4	K=5	K=6	K=7
دقت	۸۹.۳۱	۸۹.۰۲	۸۸.۳۶	۸۷.۲۴	۸۶.۳۲
بازخوانی	۹۰.۳۲	۹۱.۳۴	۸۹.۶۴	۸۸.۴۷	۸۹.۳۲
اندازه‌گیری F	۸۹.۸۱	۹۰.۱۷	۸۹.۰۰	۸۷.۸۵	۸۷.۷۹
صحت	۹۱.۳۸	۹۲.۰۰	۹۰.۵۸	۸۹.۳۴	۸۸.۵۶
نرخ خطا	۰.۰۸۶۲	۰.۰۸۰۰	۰.۰۹۴۲	۰.۱۰۶۶	۰.۱۱۴۴

در جدول(۶)، مدل پیشنهادی با ۱۰۰ بار تکرار و ۱۵۰ نسل و برمبنای مقادیر مختلف k ارزیابی شده است. درصد صحت برای ۱۵۰ نسل برابر ۹۶.۳۲ است. رابطه افزایش دقت با افزایش نسل در مدل پیشنهادی این است که اگر تعداد نسل‌ها بیشتر باشد احتمال جستجوی سراسری و یافتن نقاط مشابه بیشتر است.

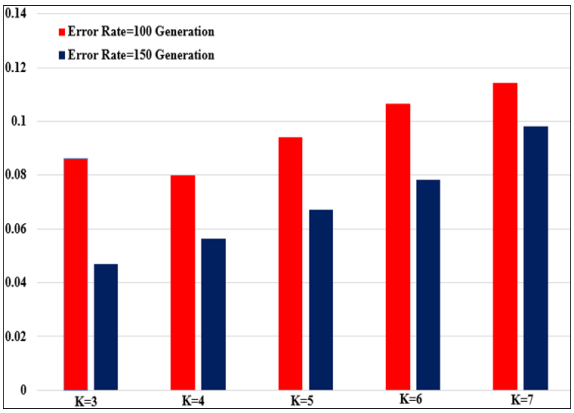
جدول ۶ ارزیابی مدل پیشنهادی با ۱۵۰ نسل

معیارها	K=3	K=4	K=5	K=6	K=7
دقت	۹۲.۶۸	۹۱.۳۶	۹۱.۷۸	۹۰.۳۱	۸۹.۲۱
بازخوانی	۹۳.۱۴	۹۲.۸۴	۹۱.۹۹	۹۰.۵۶	۸۹.۶۶

اندازه‌گیری F	۹۲.۹۱	۹۲.۰۹	۹۱.۸۸	۹۰.۴۳	۸۹.۳۴
صحت	**۹۶.۳۲**	۹۴.۳۷	۹۳.۲۸	۹۲.۱۸	۹۰.۱۸
نرخ خطا	۰.۰۴۶۸	۰.۰۵۶۳	۰.۰۶۷۲	۰.۰۷۸۲	۰.۰۹۸۲

در شکل(۴)، نمودار مدل پیشنهادی برمبنای معیار نرخ خطا با ۱۰۰ و ۱۵۰ نسل نشان داده‌شده است. همانطور که در شکل(۴) مشاهده می‌کنید دقت خطا برای مقادیر مختلف k در حالت ۱۵۰ نسل کمتر است.

شکل ۴ نمودار مدل پیشنهادی برمبنای معیار نرخ خطا با ۱۰۰ و ۱۵۰ نسل



۴-۴- ارزیابی مدل پیشنهادی برمبنای آموزش و تست

در جدول(۷)، مدل پیشنهادی با ۱۰۰ بار تکرار برمبنای مقادیر مختلف k ارزیابی شده است. درصد داده‌های آموزش و تست در روند افزایش دقت خیلی تأثیرگذار هستند. درصورتی‌که مقدار آن‌ها به درستی تعیین نشود احتمال کاهش دقت وجود دارد. زیرا برای اجرای مدل باید داده‌های مناسب برای آموزش وجود داشته باشند و مدل در مرحله تست با مشکل عدم تشخیص مواجه نشود.

جدول ۷ ارزیابی مدل پیشنهادی با ۵۰ و ۵۰ درصد آموزش و تست

معیارها	K=3	K=4	K=5	K=6	K=7
دقت	۸۹.۴۷	۸۷.۲۵	۸۶.۳۲	۸۵.۲۴	۸۴.۹۷
بازخوانی	۹۰.۲۴	۸۸.۱۶	۸۹.۳۱	۸۸.۰۹	۸۵.۱۶
اندازه‌گیری F	۸۹.۸۵	۸۷.۷۰	۸۷.۷۹	۸۶.۶۴	۸۵.۰۶
صحت	۸۹.۳۷	۸۶.۳۷	۸۵.۱۲	۸۴.۰۵	۸۵.۲۸
نرخ خطا	۰.۱۰۶۳	۰.۱۳۶۳	۰.۱۴۸۸	۰.۱۵۹۵	۰.۱۴۷۲

در جدول(۸)، مدل پیشنهادی با ۱۰۰ بار تکرار و برمبنای مقادیر مختلف k ارزیابی شده است. اگردرصد آموزش و تست به ترتیب برابر ۶۰ و ۴۰ باشند دقت تشخیص برابر ۹۳.۳۷ است.

جدول ۸ ارزیابی مدل پیشنهادی با ۶۰ و ۴۰ درصد آموزش و تست

معیارها	K=3	K=4	K=5	K=6	K=7
دقت	۸۶.۰۲	۸۷.۲۴	۸۵.۲۴	۸۳.۲۱	۸۲.۹۴
بازخوانی	۸۹.۳۴	۸۸.۲۰	۸۶.۹۰	۸۴.۶۱	۸۳.۵۵
اندازه‌گیری F	۸۷.۶۵	۸۷.۷۲	۸۶.۰۶	۸۳.۹۰	۸۳.۲۴
صحت	۹۳.۳۷	۹۲.۱۰	۸۹.۲۷	۸۸.۵۱	۸۶.۷۳
نرخ خطا	۰.۰۹۶۳	۰.۰۹۹۰	۰.۱۰۷۳	۰.۱۱۴۹	۰.۱۳۲۷

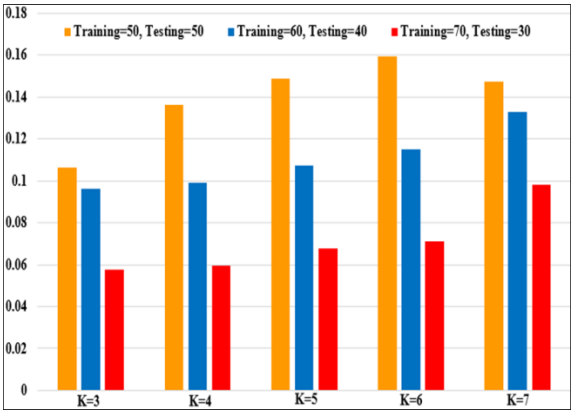
در جدول(۹)، مدل پیشنهادی با ۱۰۰ بار تکرار و برمبنای مقادیر مختلف k ارزیابی شده است. اگر درصد آموزش و تست به ترتیب برابر ۷۰ و ۳۰ باشند دقت تشخیص برابر ۹۵.۲۵ است. همچنین مقدار آموزش و تست در مقدار k تأثیرگذار هستند.

جدول ۹ ارزیابی مدل پیشنهادی با ۷۰ و ۳۰ درصد آموزش و تست

معیارها	K=3	K=4	K=5	K=6	K=7
دقت	۹۲.۳۶	۹۱.۵۷	۹۰.۱۵	۸۹.۶۴	۸۸.۲۹
بازخوانی	۹۳.۲۷	۹۴.۳۹	۹۱.۵۷	۸۹.۸۹	۸۹.۰۴
اندازه‌گیری F	۹۲.۸۱	۹۲.۹۶	۹۰.۸۵	۸۹.۷۶	۸۸.۶۶
صحت	**۹۵.۲۵**	۹۴.۰۶	۹۳.۲۴	۹۲.۸۷	۹۰.۲۰
نرخ خطا	۰.۰۵۷۵	۰.۰۵۹۴	۰.۰۶۷۶	۰.۰۷۱۳	۰.۰۹۸۰

در شکل(۵)، نمودار مدل پیشنهادی برمبنای معیار نرخ خطا با مراحل آموزش و تست نشان داده‌شده است. همانطور که در شکل(۵) مشاهده می‌کنید دقت خطا در مرحله آموزش و تست به ترتیب با ۷۰ و ۳۰ درصد در مقایسه با مراحل دیگر کمتر است.

شکل ۵ نمودار مدل پیشنهادی برمبنای معیار نرخ خطا با مراحل آموزش و تست



۵-۴- ارزیابی مدل پیشنهادی برمبنای تعداد ویژگی

در جدول(۱۰)، مدل پیشنهادی با ۱۰۰ بار تکرار و با ویژگی‌های مختلف ارزیابی شده است. اگر تعداد ویژگی‌ها کمتر باشد درصد دقت تشخیص افزایش یافته و نرخ خطا کاهش می‌یابد. در جدول(۱۰)، مشاهده می‌کنید برای ۲۰ ویژگی درصد صحت برابر ۹۸.۳۴ است.

جدول ۱۰ ارزیابی مدل پیشنهادی برمبنای تعداد ویژگی و K

معیارها	انتخاب ویژگی، k=3						
	۲۰	۲۵	۳۰	۳۵	۴۵	۵۰	۵۲
دقت	۹۲.۳۶	۹۴.۵۸	۹۶.۴۸	۹۵.۷۸	۹۴.۷۸	۹۳.۲۱	۹۴.۵۰
بازخوانی	۹۴.۱۳	۹۵.۶۸	۹۷.۱۸	۹۶.۳۱	۹۶.۲۰	۹۴.۴۹	۹۵.۶۳
اندازه گیری F	۹۳.۲۴	۹۵.۱۳	۹۶.۸۳	۹۶.۰۴	۹۵.۴۸	۹۳.۸۵	۹۵.۰۶
صحت	۹۸.۳۴	۹۵.۲۸	۹۴.۸۹	۹۵.۴۸	۹۳.۱۱	۹۶.۳۷	۹۴.۸۷
نرخ خطا	۰.۰۳۶۶	۰.۰۴۷۲	۰.۰۵۱۱	۰.۰۴۵۲	۰.۰۶۸۹	۰.۰۳۶۳	۰.۰۵۱۳

در جدول(۱۱)، مدل پیشنهادی با ۱۰۰ بار تکرار و با ویژگی‌های مختلف ارزیابی شده است. در جدول(۱۱)، مشاهده می‌کنید برای ۳۰ ویژگی درصد صحت برابر ۹۵.۸۴ است.

جدول ۱۱ ارزیابی مدل پیشنهادی برمبنای تعداد ویژگی و K

معیارها	انتخاب ویژگی، k=5						
	۲۰	۲۵	۳۰	۳۵	۴۵	۵۰	۵۲
دقت	۹۲.۳۶	۹۳۸۴	۹۴.۵۲	۹۳.۱۸	۹۱.۴۸	۹۲.۱۷	۹۰.۶۴
بازخوانی	۹۳.۲۵	۹۴.۰۸	۹۵.۸۱	۹۴.۲۶	۹۲.۵۱	۹۳.۰۱	۹۱.۰۶
اندازه گیری F	۹۲.۸۰	۹۳.۹۶	۹۵.۱۶	۹۳.۷۲	۹۱.۹۹	۹۲.۵۹	۹۰.۸۵
صحت	۹۴.۸۴	۹۳.۴۰	۹۵.۸۴	۹۱.۵۶	۹۲.۳۷	۹۱.۰۵	۹۰.۷۷
نرخ خطا	۰.۰۵۱۶	۰.۰۰۶۶	۰.۰۴۱۶	۰.۰۰۸۴	۰.۰۷۶۳	۰.۰۸۹۵	۰.۰۹۲۳

در جدول(۱۲)، مدل پیشنهادی با ۱۰۰ بار تکرار و با ویژگی‌های مختلف ارزیابی شده است. در جدول(۱۲)، مشاهده می‌کنید برای ۲۰ ویژگی درصد صحت برابر ۹۳.۲۷ است.

جدول ۱۲ ارزیابی مدل پیشنهادی برمبنای تعداد ویژگی و K

معیارها	انتخاب ویژگی، k=7						
	۲۰	۲۵	۳۰	۳۵	۴۵	۵۰	۵۲
دقت	۹۰.۲۳	۹۱.۲۴	۹۱.۸۴	۹۲.۳۱	۹۰.۳۶	۹۱.۲۴	۹۰.۸۴
بازخوانی	۹۱.۳۴	۹۱.۶۶	۹۱.۹۹	۹۲.۸۵	۹۱.۳۲	۹۱.۴۵	۹۱.۳۲
اندازه گیری F	۹۰.۷۸	۹۱.۴۵	۹۱.۹۱	۹۲.۵۸	۹۰.۸۴	۹۱.۳۴	۹۱.۰۸
صحت	**۹۳.۲۷**	۸۹.۳۱	۹۲.۷۰	۹۰.۸۹	۹۱.۶۳	۸۹.۴۱	۹۰.۴۳
نرخ خطا	۰.۰۸۷۳	۰.۱۰۶۹	۰.۰۷۳۰	۰.۰۹۱۱	۰.۰۸۳۷	۰.۱۰۵۹	۰.۰۹۵۷

۶-۴- مقایسه و ارزیابی

در این بخش در جدول(۱۳) دقت تشخیص مدل پیشنهادی با مدل‌های دیگر مقایسه شده است. مدل پیشنهادی در مقایسه با اکثر مدل‌ها دقت تشخیص بیشتری دارد و در فاکتورهای تعداد نسل و مرحله آموزش و تست در مقایسه با مدل درخت رگرسیون تجمیعی بیزین دقت کمتری دارد.

مدل پیشنهادی در مرحله انتخاب ویژگی و تعداد تکرار بیشترین دقت را در میان تمامی مدل‌ها دارد.

جدول ۱۳ مقایسه مدل پیشنهادی با مدل‌های دیگر

درصد صحت	مدل‌ها	رفرنس‌ها
۹۶.۵	NB-K-Means	[۱۰]
۹۷.۶۰	درخت رگرسیون تجمیعی بیزین	[۱۶]
۹۴.۲۸	الگوریتم سیستم ایمنی مصنوعی	[۱۸]
۹۱.۷	شبکه عصبی مصنوعی گروهی	[۲۵]
۸۳.۲۰	الگوریتم انتخاب منفی-بهینه‌سازی اجتماع ذرات	[۲۶]
۶۸.۸۶	الگوریتم انتخاب منفی	[۲۷]
۸۰.۶۶	الگوریتم انتخاب منفی-تکامل تفاضلی	
۹۱.۲۲	الگوریتم انتخاب منفی-بهینه‌سازی اجتماع ذرات	[۲۸]
۸۱.۳۲	بهینه‌سازی اجتماع ذرات	
۶۸.۸۶	الگوریتم انتخاب منفی	
۸۸.۵۶	شبکه بیزین	[۲۹]
۸۹.۷	بوستینگ	
۹۱.۵۴	درخت تصادفی	
۹۲.۳۲	JRIP	
۹۲.۳۴	درخت J48	
۹۳.۲۸	شبکه عصبی مصنوعی چندلایه	
۹۳.۵۶	تحلیل کاپا	
۹۳.۸۹	جنگل تصادفی	
۷۹.۲۸	نیوی بیز	[۳۰]
۸۹.۸۰	شبکه بیزین	
۹۰.۴۱	ماشین بردار پشتیبان	
۹۳.۳۴	درخت تابعی	
۹۲.۹۷	درخت J48	
۹۴.۸۲	جنگل تصادفی	
۹۰.۹۳	درخت تصادفی	
۹۲.۴۳	درخت تصمیم‌گیری CART	
۹۷.۶۱	۲۰۰= تعداد تکرار	مدل پیشنهادی
۹۶.۳۲	۱۵۰= تعداد نسل	
۹۵.۲۵	۷۰= آموزش ۳۰= تست	
۹۸.۳۴	۲۰= انتخاب ویژگی	

۵. نتیجه‌گیری و کارهای آینده

همزمان با گسترش استفاده از کامپیوترها و مطرح شدن شبکه‌های کامپیوتری و به دنبال آن اینترنت، حیات کامپیوترها و کاربران آنان دستخوش تغییرات اساسی شده است. به دلیل گسترش روز افزون ایمیل هرزنامه و همچنین خسارات فراوانی که وارد می‌کنند، لازم است به دنبال روش‌هایی در جهت مقابله با آن‌ها باشیم. در این مقاله مدلی برای تشخیص ایمیل هرزنامه برمبنای k نزدیک‌ترین همسایه و الگوریتم بهینه‌سازی اجتماع گربه پیشنهادشده است. در مدل پیشنهادی از k نزدیک‌ترین همسایه برای آموزش و تست داده‌ها در مرحله طبقه‌بندی و از الگوریتم بهینه‌سازی اجتماع گربه در مرحله جستجو، به‌روزرسانی موقعیت‌ها و انتخاب ویژگی‌های برازنده استفاده‌شده است. نتایج برمبنای فاکتورهایی مانند تعداد تکرار، تعداد نسل، آموزش و تست و انتخاب ویژگی انجام‌شده است. بیشترین دقت تشخیص در مدل پیشنهادی مربوط به انتخاب ویژگی است. همچنین مقایسه‌ها نشان داده که مدل پیشنهادی در مقایسه با مدل‌های داده کاوی همانند شبکه بیزین، بوستینگ، درخت J48 و شبکه عصبی مصنوعی چندلایه دقت تشخیص بهتری دارد. برای کارهای اینده در نظر داریم مدلی را برمبنای ترکیب الگوریتم‌های فرا ابتکاری ارائه دهیم که در تشخیص و انتخاب ویژگی‌های مهم کارایی و سرعت بالایی داشته باشد.

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] M. Ghorbanivand, F.S. Gharehchopogh, Spam Email Detection using Hybrid of Ant Colony Optimization and Adaboost, Information Technology in Engineering Design, in press, 12(2), 2019. (Persian)
[3] B. Martin, Instance-Based Learning: Nearest Neighbor with Generalization, Doctoral dissertation, University of Waikato, 1995
[4] Chu. Shu-Chuan, Computational intelligence based on the behavior of cats, International Journal of Innovative Computing, Information and Control, Vol. 3, 2007
[5] F.S. Gharehchopogh, N. Karimpour, A New Approach to Detect Spam Emails Using the Hybrid Model of Ant Colony and Firefly Algorithms, Computing Science Journal (CSJ), in press, Nov 2019. (Persian)
[6] F.S. Gharehchopogh, M. Vafadar, M. Motaman, Improve of Invasive Weed Optimization algorithm with K nearest neighbor for email spam classification, Computing Science Journal (CSJ), Vol. 10, pp. 54-64, 2018. (Persian)
[7] H. Majidpour, F.S. Gharehchopogh, An Improved Flower Pollination Algorithm with AdaBoost Algorithm for Feature Selection in Text Documents Classificationm, Journal of Advances in Computer Research, 9(1): 29-40, 2018
[8] A. Allahverdipour, F.S. Gharehchopogh, An Improved K-Nearest Neighbor with Crow Search Algorithm for Feature Selection in Text Documents Classification, Journal of Advances in Computer Research, 9(2): 37-48, 2018.
[9] 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
[10] 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
[11] 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
[12] 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.
[13] 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
[14] 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, pp. 987-992, 2009
[15] Siti-Hajar-Aminah Ali, S. Ozawa, J. Nakazato, T. Ban, J. Shimamura, An autonomous online malicious spam email detection system using extended RBF network, International Joint Conference on Neural Networks (IJCNN), pp. 1-7, 2015
[16] S. Abu-Nimeh, D. Nappa, X. Wang, S. Nair, Bayesian Additive Regression Trees-Based Spam Detection for Enhanced Email Privacy, Third International Conference on Availability, Reliability and Security, pp. 1044-1051, 2008
[17] S.B. Rathod, T.M. Pattewar, Content based spam detection in email using Bayesian classifier, International Conference on Communications and Signal Processing (ICCSP), pp. 1257-1261, 2015
[18] I. Idris, A. Selamat, Negative selection algorithm in artificial immune system for spam detection, Malaysian Conference in Software Engineering, pp. 379-382, 2011.
[19] L. Firte, C. Lemnaru, R. Potolea, Spam detection filter using KNN algorithm and resampling, Proceedings of the 2010 IEEE 6th International Conference on Intelligent Computer Communication and Processing, pp. 27-33, 2010.
[20] R. Ariaeinejad, A. Sadeghian, Spam detection system: A new approach based on interval type-2 fuzzy sets, 24th Canadian Conference on Electrical and Computer Engineering(CCECE), pp. 379-384, 2011.
[21] Data set Spambase, https://archive.ics.uci.edu/ml/datasets/Spambase, [last available 2017.10.10].

[22] B.K. Singh, K. Verma, A.S. Thoke, Investigations on Impact of Feature Normalization Techniques on Classifier's Performance in Breast Tumor Classification, International Journal of Computer Applications, Vol. 116, No. 19, pp. 11-15, 2015.
[23] R.S. Michalski, I. Bratko, M. Kubat, Machine Learning, and Data Mining: Methods and Applications, New York: Wiley, 1998.
[24] V. Garcia, R.A. Mollineda, and J.S. Sanchez, Index of Balanced Accuracy: A Performance Measure for Skewed Class Distributions, Iberian Conference on Pattern Recognition and Image Analysis, IbPRIA 2009: Pattern Recognition and Image Analysis, pp. 441-448, 2009
[25] El-Sayed M. El-Alfy, R.E. Abdel-Aal, Using GMDH-based networks for improved spam detectionand email feature analysis, Applied Soft Computing, Vol. 11, Issue 1, pp. 477-488, 2011.
[26] 11. 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
[27] 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.
[28] 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
[29] S. Sharma, A. Arora, Adaptive Approach for Spam Detection, IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 4, No. 1, pp. 23-26, July 2013
[30] 42. M. Rathi, V. Pareek, Spam Mail Detection through Data Mining-A Comparative Performance Analysis, I.J. Modern Education and Computer Science, 12, pp. 31-39, 2013