DOR :20.1001.1.2538161.1399.5.4.7.4

تاریخ دریافت مقاله: ۹۸/۰۸/۰۳
تاریخ پذیرش مقاله: ۹۹/۰۴/۱۶

انتخاب ویژگی با الگوریتم بهینه‌سازی حاصلخیزی زمین‌های کشاورزی برای تشخیص صفحات وب هرز

محمد سخی دل هوسین
کارشناس ارشد مهندسی کامپیوتر، گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران
پست الکترونیکی: mohammad.sakhidel@gmail.com

فرهاد سلیمانیان قره چپق*
استادیار، گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران
پست الکترونیکی: bonab.farhad@gmail.com

چکیده

در فضای اینترنت، امکان به‌کارگیری انواع سرویس‌ها و خدمات متعدد برای کاربران مهیا شده است. همزمان با رشد و گسترش استفاده از اینترنت، تعداد هرزنویسان وب افزایش یافته است. صفحات وب هرز به اشکال مختلفی چون تبلیغات تجاری و ویروس‌هایی نهان شده در صفحات وب جایگذاری می‌شود. صفحات وب هرز علاوه بر تهدید امنیت کاربران در وب، موجب هدر رفتن منابع سیستم و ایجاد ترافیک مخرب نیز می‌گردند؛ لذا ارایه راهکارهایی جهت مقابله با وب هرز ضروری به نظر می‌رسد. یکی از روش‌های شناسایی و مقابله با صفحات وب هرز، طبقه‌بندی صفحات با استفاده از الگوریتم‌های یادگیری ماشین است. در این مقاله، مدلی جدید برمبنای الگوریتم حاصلخیزی زمین‌های کشاورزی و بیز ساده با عنوان FFANB[1] برای تشخیص صفحات وب هرز پیشنهاد شده است. در مدل FFANB از الگوریتم حاصلخیزی زمین‌های کشاورزی برای انتخاب ویژگی و بیز ساده برای طبقه‌بندی نمونه‌ها استفاده شده است. هدف مدل FFANB کاهش ویژگی‌ها به منظور افزایش صحت بااستفاده از الگوریتم حاصلخیزی زمین‌های کشاورزی می‌باشد که از مجموعه داده WEBSPAM-UK2007 که از معتبرترین مجموعه داده در زمینه شناسایی صفحات وب هرز می‌باشد استفاده شده است. این مجموعه داده شامل سه دسته ویژگی با عناوین ویژگی‌های مبتنی بر محتوا (۹۶ ویژگی)، ویژگی‌های مبتنی بر پیوند (۴۱ ویژگی) و ویژگی‌های مبتنی بر پیوند تبدیل یافته (۱۳۸ ویژگی) می‌باشد که تعداد کل ویژگی‌ها برابر با ۲۷۵ ویژگی است. نتایج ارزیابی‌های صورت گرفته برروی مدل FFANB نشان دهنده درصد دقت ۰/۹۲۴۱ و صحت ۰/۹۵۸۴ می‌باشند که حاکی از برتری مدل FFANB در مقایسه با بسیاری از روش‌های پیشین می‌باشد.

واژه‌های کلیدی: صفحات وب هرز، طبقه‌بندی، الگوریتم حاصلخیزی زمین‌های کشاورزی، الگوریتم بیز ساده، انتخاب ویژگی

۱. مقدمه

هر وبگاه شامل مجموعه‌ای از صفحات وب متصل به هم است که شامل اطلاعات خاصی در مورد یک مبحث

* نویسندهٔ مسئول
1- Farmland Fertility Algorithm Naive Bayes (FFANB)

می‌باشند و معمولاً روی یک کارســـاز قرار دارند. وبگاه‌ها معمولاً انبوهی از اطلاعـــات را در قالب متن، تصویر، صدا، و فیلـــم در اختیار کاربران قـــرار می‌دهند. مهم‌ترین چالش اصلی برای موتورهای جستجوگر در جستجوی مطالب و ارائه خدمات کارآمد به کاربران، صفحات هرز می‌باشند که نقشی تخریب‌کننده در نتایج موتورهای جستجوگر از نظر دقت، سرعت و فضای ذخیره‌سازی دارند. هرز در صفحات وب چالش بزرگی اســـت که هم بـــر روی کاربران و هم بر روی تامین‌کنندگان سرویس تاثیرگذار است. کاربران زمانی که در حال انجام فعالیت‌های خود مانند جستجوی مبتنی بر وب هستند با وب هرز مواجه می‌شوند و تشخیص اطلاعات واقعـــی برای آن‌ها مشـــکل می‌گردد. در واقـــع موتورهای جســـتجوگر هرچه قدر هم که قوی و کارآمد باشند به دلیل بـــالا بودن حجم و تنوع موضوعی اطلاعات و منابع موجود در شبکه اینترنت قادر به سازماندهی تمامی منابع این شبکه نیستند و لذا تشخیص وب هرز در میان انبوهی از اطلاعات سخت است. موتورهای جستجوگر باید صفحات وب را پیدا کنند و محتوای آن‌ها را برمبنای کلمات وب هرز تجزیه کنند تا صفحات وب هرز را تشخیص دهند [۱,۲].

هدف از راه‌انـــدازی موتورهای جســـتجوگر، کمک به کاربران در پیدا کردن اطلاعـــات موجود در صفحات وب می‌باشد. در سیستم موتورهای جستجوگر با وارد کردن کلمات کلیدی توســـط کاربران، لیستی از صفحات وبگاه‌ها ظاهر می‌شـــوند که به موضوع مـــورد علاقه کاربر مرتبط می‌باشند. موتورهای جستجوگر در واقع این اطلاعات را در میان صفحات وبگاه‌ها جستجو کرده و در مدت زمان کوتاه و با بیشترین ســـرعت در اختیار کاربران قرار می‌دهند. اما نکته مهم این اســـت که همزمان با اطلاعات اســـتخراج شده یکســـری از صفحـــات وب هرز بازیابی می‌شـــوند که باید شناسایی شـــوند و صفحات آن‌ها باز نشود. برای این کار باید کلمات هرز تشخیص داده شوند و اگر یک صفحه حاوی آن کلمات باشد از بازیابی آن اجتناب شود [۳,٤].

مســـئله انتخاب ویژگی در بسیاری از کاربردها (مانند طبقه‌بنـــدی) اهمیت ویـــژه‌ای دارد، زیـــرا در این کاربردها تعـــداد زیادی ویژگی وجود دارد، که بســـیاری از آن‌ها یا بی‌اســـتفاده هستند و یا این‌که بار اطلاعاتی چندانی ندارند. حذف نکردن این ویژگی‌ها مشکلی از لحاظ اطلاعاتی ایجاد نمی‌کنند ولی بار محاسباتی را افزایش می‌دهند. با افزایش تعـــداد ویژگی‌ها، فضـــای ویژگی‌ها نیز افزایـــش می‌یابد، تجزیه و تحلیل داده‌ها و طبقه‌بندی نیز به‌طور قابل توجهی ســـخت‌تر می‌شـــود. به‌علاوه، داده‌ها به‌طور فزاینده‌ای در فضایی که اشـــغال‌کرده‌اند پراکنده می‌شـــوند که منجر به مشکلات بزرگی هم برای الگوریتم‌های بانظارت وهم بدون نظارت می‌شـــود. این پدیده به عنوان مشکل ابعاد شناخته شده است و براساس این واقعیت است که اغلب کارکردن باداده‌های با ابعاد بالا مشـــکل است. تعداد زیاد ویژگی‌ها می‌توانند اختلال در داده‌ها را افزایش دهد و درنتیجه خطای الگوریتم یادگیری نیز افزایش می‌یابد، بخصوص اگر تعداد نمونه‌ها در مقایسه با تعداد ویژگی‌ها کم باشد [٥,٦].

در مقالات پیشـــین در زمینه تشـــخیص ایمیل هرزنامه از الگوریتم‌هـــای مختلف فرا ابتکاری و طبقه‌بندی اســـتفاده کردیم [۷,۸]. برای طبقه‌بندی نمونه‌ها از الگوریتم‌هایی مانند k نزدیکترین همسایه و آدابوست استفاده کردیم. دقت تشخیص مدل‌های پیشین که مبتنی بر الگوریتم‌های فرا ابتکاری بودند بســـیار بالا و دقیق بود. لـــذا در این مقاله هم از الگوریتم فرا ابتکاری به دلیل قدرت بالا در یافتن راه حل‌های بهینه جهت تشـــخیص صفحات وب هرز استفاده می‌کنیم و مدلی جدید برروی مجموعه داده وب هرز ارزیابی می‌شـــود. طبق نتایج مشخص شد که الگوریتم بهینه‌سازی حاصلخیزی زمین‌های کشـــاورزی که از دســـته الگوریتم‌های فراابتکاری است در تشخیص صفحات وب هرز کارایی خوبی دارد. در تشخیص ایمیل هرزنامـــه از مجموعه داده معتبـــر Spambase با ٥٧ ویژگی استفاده کردیم. اما در این مقاله از مجموعه داده‌های وب هرز که برمبنای ویژگی‌های وب طراحی شده‌اند استفاده می‌کنیم. در مجموعه داده‌های وب هرز، تعداد ویژگی‌ها، نوع داده‌ها و مقدار داده‌ها متفاوت از ایمیل هرزنامه است.

در این مقاله، مدلی جدید برمبنای الگوریتم بهینه‌سازی حاصلخیزی زمین‌های کشاورزی [۹] و بیز ساده [۱۰] با عنوان FFANB برای تشخیص صفحات وب هرز برروی مجموعه داده WEBSPAM-UK2007 [۱۱] پیشنهاد می‌شود. در مدل FFANB از الگوریتم بهینه‌سازی حاصلخیزی زمین‌های کشاورزی برای انتخاب ویژگی‌ها و از الگوریتم بیز ساده برای طبقه‌بندی نمونه‌ها استفاده می‌کنیم. الگوریتم بیز ساده یکی از روش‌های یادگیری ماشین برای طبقه‌بندی به شمار می‌آید. در این روش رده‌های مختلف، هر کدام به شکل یک فرضیه دارای احتمال در نظر گرفته می‌شوند. هر داده آموزشی جدید، احتمال درست بودن فرضیه‌های پیشین را افزایش و یا کاهش می‌دهد و در نهایت، فرضیاتی که دارای بالاترین احتمال باشند، به عنوان یک رده در نظر گرفته شده و برچسبی بر آن‌ها تخصیص داده می‌شود[۱۲].

۲. تحقیقات مرتبط

پژوهشگران برای تشخیص صفحات وب هرز از روش Smart-BT استفاده کرده‌اند [۱۳]. آن‌ها در این روش برای انتخاب ویژگی و کشف وابستگی بین ویژگی‌ها از آزمون کای‌دو و برای طبقه‌بندی از الگوریتم بیز ساده استفاده کرده‌اند. آزمون کای‌دو، میزان ارتباط یا وابستگی بین ویژگی‌ها را اندازه‌گیری می‌کند. آزمایش مربعی کای‌دو برپایه یک آزمایش آماری ($\chi 2$) می‌باشد. در این روش ارزش هر ویژگی با استفاده از محاسبه مقدار آماری کای‌دو و با توجه به ردهٔ مد نظر ارزیابی می‌گردد. این مقدار با استفاده از آزمون کای‌دو جزء آزمون‌های غیر پارامتری است محاسبه می‌شود. نتایج نشان داده که مدل Smart-BT در مقایسه با ماشین بردار پشتیبان، درخت تصمیم‌گیری و شبکه عصبی مصنوعی دقت بیشتری دارد. همچنین انتخاب ویژگی با استفاده از کای دو از رتبه بالاتری در مقایسه با الگوریتم‌های ژنتیک، و بهینه‌سازی اجتماع ذرات بهره‌مند است.

برای تشخیص صفحات وب هرز از روش‌های یادگیری ماشین استفاده شده است [۱۴]. ارزیابی برروی مجموعه داده‌های WEBSPAM-UK2006 و WEBSPAM-UK2007 انجام شده است. بیشترین درصد صحت متعلق به الگوریتم آدابوست است که درصد صحت آن برای مجموعه داده‌های WEBSPAM-UK2006 و WEBSPAM-UK2007 به‌ترتیب برابر با ۰/۹۳۷۰ و ۰/۸۵۲۰ است. همچنین ماشین بردار پشتیبان، شبکه عصبی مصنوعی چندلایه و بیز ساده از دقت تشخیص مناسب برخوردار هستند.

مدلی برمبنای ترکیب الگوریتم بهینه‌سازی اجتماع ذرات و انتخاب ویژگی برمبنای همبستگی برای تشخیص صفحات وب هرز پیشنهاد شده است [۱۵]. در مدل ترکیبی از انتخاب ویژگی برمبنای همبستگی برای بهبود الگوریتم بهینه‌سازی اجتماع ذرات استفاده شده است. ارزیابی برروی مجموعه داده WEBSPAM-UK2006 انجام شده است. نتایج نشان داده است که بهترین مقدار F-Measure برابر با ۸۸ درصد است.

برای تشخیص صفحات وب هرز از روش‌های بیز ساده، درخت C5.0، ماشین بردار پشتیبان، بگینگ، آدابوست و فیلتر استفاده شده است [۱۶]. ارزیابی برروی مجموعه داده WEBSPAM-UK2007 انجام شده است. نتایج نشان داده است که درصد صحت مدل فیلتر برابر با ۰/۸۱۹۰ درصد است که در مقایسه با روش‌های بیز ساده، درخت C5.0، ماشین بردار پشتیبان، بگینگ و آدابوست از کارایی بهتری برخوردار است. شبکه باور عمیق که برگرفته از شبکه‌های عصبی مصنوعی است برای تشخیص صفحات وب هرز پیشنهاد شده است [۱۷]. ایده اصلی شبکه باور عمیق، تکرار یادگیری شبکه به منظور آموزش بهینه و افزایش دقت تشخیص است. ارزیابی برروی مجموعه داده WEBSPAM-UK2007 انجام شده است. نتایج نشان داده است که درصد صحت برابر با ۰/۹۱۲۶ درصد است.

مدلی برمبنای شبکه عصبی مصنوعی چندلایه و ژنتیک برای طبقه‌بندی صفحات وب هرز برروی مجموعه داده WEBSPAM-UK2007 پیشنهاد شده است. در مدل ترکیبی از شبکه عصبی مصنوعی چندلایه برای آموزش نمونه‌ها و

از ژنتیک برنامه‌نویسی شده برای طبقه‌بندی استفاده شده اســت. نتایج نشــان داده که درصد صحت برابر با ۰/۹۲ درصد است که در مقایسه با مدل‌های درصد بالاتری دارد [۱۸]. مدلی برمبنای ماشین بردار پشتیبان برای تشخیص صفحات وب هرز پیشنهاد شــده است. ابتدا ماشین بردار پشتیبان با N نقاط داده آموزش داده شده که هرکدام قبلاً به عنوان وب هرز یا وب غیرهرز طبقه‌بندی شده‌اند. در این مقاله از ترکیب ویژگی‌ها استفاده شده است. درصد نتایج با ۲۹۶ ویژگی برابر با ۰/۸ است [۱۹].

۳. مدل FFANB

هدف وب هرزنویســان این اســت کــه رتبه صفحات وب را بــه روش وب هرز به‌طور غیرقانونی افزایش دهند. لذا هدف این اســت که این صفحات هرز شناســایی شوند و ســپس حذف شــوند. اگر این محتوا شناسایی نشود و نتایج جســتجو فیلتر نشــود، صفحات وب هرز می‌توانند دقت موتورهای جســتجو را به شدت کاهش دهند. به‌طور کلی، وب هرزها قصد دارند بــه صورت غیرقانونی مرتباً رتبه صفحات هرز خود را از طریق موتورهای جستجوگر افزایش دهند که ممکن است منجر به جعل، تخریب اطلاعات و تحریف نتایج جســتجو شــود و از این طریق کل فرآیند جســتجوی اطلاعات تحت تأثیر قرار می‌گیرد. با توجه به موفقیت ابزارهای ضد وب هرز ایمیل توسط الگوریتم‌های یادگیری ماشین، ما در این مقاله برای شناسایی صفحات وب هرز از روش‌های یادگیری ماشــین استفاده می‌کنیم. به‌طور معمول، دقت تشــخیص بالا و ضریب مثبت کاذب پایین، اصلی‌ترین ویژگی‌های لازم برای شناسایی وب هرز براســاس روش‌های یادگیری ماشین می‌باشد. این امر به ویژه برای شناســایی صفحات وب هرز و اطمینان از عدم جعل وبگاه‌ها خیلی مهم و ضروری است.

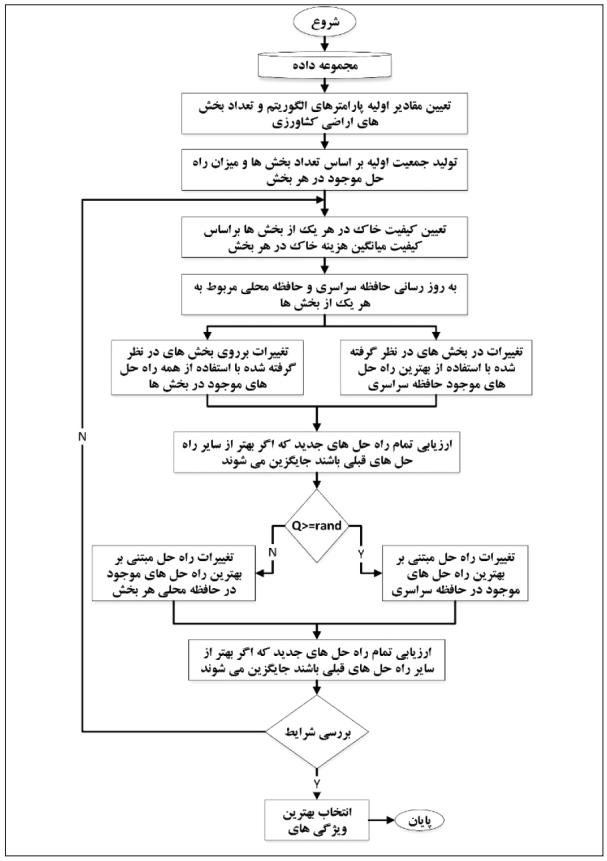
مدل FFANB، ترکیبی از الگوریتم حاصلخیزی زمین‌های کشــاورزی و بیز ساده است. در مدل FFANB از الگوریتم حاصلخیزی زمین‌های کشــاورزی برای انتخاب ویژگی و از بیز ساده برای طبقه‌بندی نمونه‌ها استفاده می‌شود. در ایــن مقاله از مجموعه داده وب هرز WEBSPAM-UK2007 استفاده شده است. این مجموعه داده توسط پژوهشگران برچسب‌گذاری شــده است و هر صفحه با عنوان «صفحه هرز» یا «صفحه غیرهرز» مشــخص شــده است. در مدل FFANB در ابتدا مجموعه داده‌ها در قالب فایل متنی خوانده می‌شــود. ســپس بااســتفاده از روش‌های پیش پردازش داده، مقدار کمینه و بیشــینه ویژگی‌ها به بازه ۰ و ۱ تبدیل می‌شود. برای توســعه مدل، در مرحله اول مقادیر هریک از ویژگی‌ها جهت طبقه‌بندی و افزایش دقت، نرمال‌ســازی می‌شــوند. جهت نرمال‌ســازی به عنوان یکی از گام‌های پیش‌پردازش داده از روش کمینه-بیشــینه طبق معادله(۱) استفاده می‌شود [۲۰].

$$x_n = \frac{(x_r - x_{min})}{(x_{max} - x_{min})} \qquad (۱)$$

در معادلــه(۱) xr ،xn ،xmax ،xmin و بــه ترتیــب نشان‌دهنده مقادیر واقعی، استاندارد شده، حداکثر و حداقل داده‌های تحت بررسی هستند. قالب بردارها برمبنای مقدار ویژگی‌ها ایجاد می‌شود و توســط الگوریتم حاصلخیزی زمین‌های کشــاورزی بهترین ویژگی‌ها در میان بردارها انتخاب می‌شوند.

۳-۱ انتخاب ویژگی

در الگوریتم حاصلخیزی زمین‌های کشاورزی در ابتدا فضا را به چهار بخش تقسیم می‌کنیم. مجموعه داده‌ها را در میان چهار بخش تقسیم می‌کنیم. یکی از بخش‌ها را انتخاب می‌کنیم و ویژگی‌ها مربوط به آن بخش را برمبنای میانگین فاصله ویژگی‌های یک رکورد ارزیابی می‌کنیم. ویژگی‌های انتخاب شــده از یک رکورد برمبنای معیــار فاصله انجام می‌شــود. اگر میانگین فاصله در یک رکورد کمتر باشــد به منزله این اســت که برازندگی ویژگی‌های انتخاب شده دقت بیشــتری خواهند داشــت. هر کدام از بخش‌ها شامل یک حافظه محلی و یک حافظه سراســری هســتند که در واقع حافظه سراسری همان بهترین میانگین به‌دست آمده



شکل۱: روندنمای مدل FFANB به منظور انتخاب ویژگی

می‌باشـــد و حافظه محلی دربرگیرنده ارزیابی رکوردها به منظور نزدیک شده به حافظه سراسری می‌باشد.

در شـــکل (۱) روندنمای مدل FFANB برمبنای انتخاب ویژگی توسط الگوریتم حاصلخیزی زمین‌های کشاورزی نشان داده شده است.

بـــرای ارزیابی مدل FFANB، ابتدا لازم اســـت مجموعه داده‌هـــا به دو بخش آموزش و آزمایش تقســـیم شـــوند. داده‌های بخش آموزش مـــدل را تولید می‌کنند و داده‌های بخش آزمایـــش با کمک تعدادی رکورد، مدل تولید شـــده را آزمایش و برچســـب مربوط بـــه رکوردهای مذکور را تعییـــن می‌نمایند. نرخ صحت یک روش طبقه‌بندی بر روی مجموعه داده‌های آموزشی، درصد مشاهداتی از مجموعه آموزش است که به درســـتی توسط روش مورد استفاده طبقه‌بندی شـــده است. برای محاســـبه این معیار داده‌های آزمون استفاده می‌شوند. همچنین می‌توان نرخ خطا را بر اساس معیار صحت محاسبه کرد.

در مدل FFANB ابتدا در مرحلـــهٔ اول مجموعه داده با تمام‌های ویژگی‌ها اســـتخراج شده و جهت انتخاب بهترین ویژگی‌ها از الگوریتم حاصلخیزی زمین‌های کشـــاورزی اســـتفاده می‌شـــود. هر موقعیت معادل یک آرایه به تعداد تمام ویژگی‌ها است. ابعاد فضای جستجو در موقعیت هر بخش از اراضی برابر با تعداد ویژگی‌ها است. هر خانه از آرایه می‌تواند عدد صفر و یا یک داشته باشد که به ترتیب به معنی آن است که یک ویژگی انتخاب نشده و یا انتخاب شـــده اســـت. به‌طور مثال اگر ۱۵ ویژگی برای بررسی به الگوریتم حاصلخیزی زمین‌های کشاورزی داده می‌شود، فضای جســـتجو برای موقعیت هر بخـــش از اراضی ۱۵ بعدی می‌شود. هر بعد متناظر با یک ویژگی است. موقعیت هر بخش از اراضی با توجه به تابع سیگموئید به محدوده ۰ و ۱ نگاشت داده می‌شود. پس مقدار هر بعد از موقعیت با تابع ســـیگموئید به مقدار ۰ و ۱ تبدیل می‌شود که مقدار ۰ به معنی حـــذف ویژگی متناظر و مقدار ۱ به معنی حفظ ویژگی متناظر با آن بعد است.

$$S(x) = \frac{1}{1 + e^{-x}} \qquad (۲)$$

$$x_i = \begin{cases} 1, & if\ S(x) \geq rand \\ 0, & if\ S(x) < rand \end{cases} \qquad (۳)$$

برای تبدیل اعداد پیوسته به دودویی از تابع سیگموئید طبق معادله(۲)، استفاده می‌شود. خروجی تابع سیگموئید در یـــک محدوده عددی خاصـــی (عموماً بین صفر و یک) قـــرار می‌گیرد. در این تابع جواب ۰ یا ۱ نخواهد بود بلکه مجموعـــه اعدادی بین صفر و یک اســـت. لذا در الگوریتم حاصلخیـــزی زمین‌های کشـــاورزی بـــرای تبدیل حالت پیوسته به گسســـته موقعیت هر بخش از اراضی xi طبق معادلـــه(۳)، تعریف می‌شـــود. اگر موقعیـــت هر بخش از اراضی به صورت معادله(٤) تعریف شـــود آنگاه موقعیت جدیـــد هر بخش از اراضـــی بعد از اعمال ســـیگموئید به صورت معادله(۵) خواهد بود. ســـپس ویژگی‌های متناظر با ستون‌های غیرصفر، به طبقه‌بند اعمال می‌شوند.

$$Pos1 = [3.10, 3.56, -3.12, 2.15, 3.10, -2.26, 0.52, -1.84, 2.80, -3.15] \quad (۴)$$

$$Pos1 = [1,1,0,1,1,0,1,0,1,0] \quad (۵)$$

بعد از انتخـــاب ویژگی‌ها، تابع برازندگی برای ارزیابی کارایی ویژگی‌های انتخاب شـــده اســـتفاده می‌شود. تابع برازندگی در مـــدل FFANB طبق معادله(٦) تعریف شـــده است [۲۱].

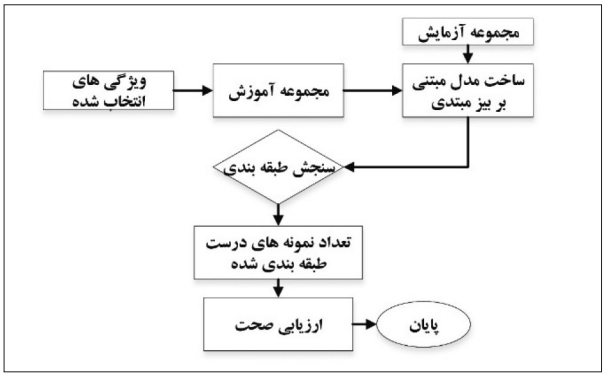
$$f(x_i) = \sigma \times E_{xi} + (1-\sigma) \times \left(\frac{|x_i|}{d}\right) \quad (۶)$$

در معادلـــه(٦) Exi مقدار خطـــای طبقه‌بندی، |xi| تعداد ویژگی‌های انتخاب شـــده، d تعـــداد کل ویژگی‌ها. پارامتر σ یـــک مقـــدار تصادفی در بازه ۰ و ۱ اســـت که به منظور توازن بین دقت طبقه‌بندی و تعداد ویژگی‌های انتخاب شده استفاده می‌شود.

۳-۲ طبقه‌بندی

وجود رده‌هـــای نامتـــوازن در طبقه‌بندی بـــه یکی از چالش‌های بزرگ در این زمینه تبدیل شده است. مجموعه داده نامتوازن براساس تعریف عبارت است از یک مجموعه داده‌ای که تعداد نمونه‌های متعلق به یک رده در آن با تعداد نمونه‌های ردهٔ دیگر به‌طور مساوی توزیع نشده باشد. رده با تعداد داده‌های بیشـــتر را ردهٔ اکثریت و رده با داده‌های کمتـــر را ردهٔ اقلیت می‌نامنـــد. در الگوریتم‌هـــای طبقه‌بند اســـتاندارد، توزیع رده‌ها متوازن در نظر گرفته می‌شود و این دســـته از الگوریتم‌ها در مواجهه با مجموعه داده‌های نامتوازن عملکرد مناسبی را از خود ارایه نمی‌دهند؛ چرا که الگوریتم‌های معمول طبقه‌بند به سمت نمونه‌های آموزشی رده بزرگ‌تر متمایل می‌شوند که این موضوع باعث افزایش خطـــا در شناســـایی نمونه‌های اقلیت می‌شـــود. به منظور ارزیابی به مســـائل مربوط به مجموعه داده‌های نامتوازن تکنیک‌های متعددی معرفی شـــده‌اند که در سه دسته زیر طبقه‌بندی می‌شوند:

روش ســـطح الگوریتم: ایـــن روش بـــه الگوریتم‌های



شکل۲: روندنمای مدل FFANB به منظور طبقه‌بندی

یادگیری طبقه‌بند کمک می‌کند تا فرآیند یادگیری را به سمت ردهٔ اقلیت سوق دهد.

روش ســـطح داده: این روش با بازنمونه‌گیری از فضای داده باعث تغییر توزیع داده‌ها می‌شود به‌طوری که تغییری در الگوریتم یادگیری ایجاد نمی‌شود و تلاش می‌کند در مرحله پیش‌پردازش تأثیرات ناشی از ناتوازنی را برطرف کند.

روش یادگیری حساس به هزینه: این روش مابین روش الگوریتمـــی و داده‌ای قرار دارد. به‌طوری‌که هم در ســـطح داده و هـــم در ســـطح الگوریتم تغییر ایجـــاد خواهد کرد. مهم‌ترین نقطه ضعف این رویکرد تعریف هزینهٔ طبقه‌بندی نادرست می‌باشد که عموما در مجموعه داده وجود ندارند.

در قســـمت طبقه‌بنـــدی از الگوریتـــم بیز ســـاده برای طبقه‌بندی نمونه‌های صفحات وب هرز اســـتفاده می‌شود. در شکل(۲) روندنمای مدل FFANB برمبنای انتخاب ویژگی توسط الگوریتم بیز ساده نشان داده شده است.

یک مجموعه داده از نمونه‌های آموزشـــی با برچسب رده(نوع رده مشخص است) و یک مجموعه داده نمونه‌های آزمایـــش E با n مقدار ویژگی $(a_1, a_2, \dots, a_n|c)$ در نظر گرفته می‌شـــود و طبقه‌بنـــدی بیز برای دســـته‌بندی E به صورت معادله(۷) تعریف می‌شود.

$$c(E) = a \quad c \in C, \max P(c). \; P(a_1, a_2, \dots, a_n|c) \quad (۷)$$

فرض پایه الگوریتم بیز ســـاده این است که در هر رده مقادیر ویژگی‌ها از یکدیگر مستقل باشند. بنابراین با استفاده از قانون احتمالی استقلال رابطه(۸) تعریف می‌شود.

$$p(a_1, a_2, \dots, a_n|c) = p(a_1|c)p(a_2|c) \dots p(a_n|c)$$
$$p(a_1, a_2, \dots, a_n|c) = \prod_{i=1}^{n} p(a_i|c) \quad (۸)$$

با جایگذاری معادله (۷) در (۸)، نتیجه به صورت معادله(۹) نشان داده می‌شود که طبقه‌بند بیز ساده نامیده می‌شود. در این رابطه، E طبقه‌بند بیز ساده روی نمونه آزمایش E است. با استفاده از معادله(۹) تمامی احتمالات می‌توانند مستقیما از روی داده‌های آموزشی تعیین شوند.

$$c_N(E) = a \ \ c \in C, \max P(c) \prod_{i=1}^{n} p(a_i|c) \quad (۹)$$

۳-۳ معیارهای ارزیابی

در این مقاله، جهت ارزیابی کارایی طبقه‌بندی از چهار معیار دقت[۲]، بازخوانی[۳]، F-Measure، و صحت[۴] استفاده شده است. دقت نشان دهنده تعداد نمونه‌های درست به تعداد کل نمونه‌ها است. معیار بازخوانی نشان می‌دهد که اگر نمونه‌ای، نوع A تشخیص داده شد با چه احتمالی نوع A می‌باشد. هرچه میزان بازخوانی بالاتر باشد بیانگر این است که قابلیت شناسایی درست رده‌ها بیشتر است. در مقابل هرچه میزان دقت بیشتر باشد، بیانگر این است که الگوریتم اشتباه‌های کمتری در شناسایی رده‌ها دارد. درصد صحت یک روش طبقه‌بندی بر روی مجموعه داده‌های آموزشی، درصد مشاهداتی از مجموعه آموزش است که به درستی توسط روش مورد استفاده طبقه‌بندی شده است[۲۲].

$$\text{دقت} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (۱۰)$$

$$\text{بازخوانی} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (۱۱)$$

$$\text{F} - \text{Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} * \text{Recall}} \quad (۱۲)$$

$$\text{صحت} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (۱۳)$$

۴. ارزیابی و نتایج

در این بخش مدل FFANB که ترکیبی از الگوریتم حاصلخیزی زمین‌های کشاورزی و بیز ساده است به منظور طبقه‌بندی برروی مجموعه داده‌های WEBSPAM-UK2007 ارزیابی می‌گردد. این مجموعه داده شامل ۳۷۷۶ نمونهٔ وب غیرهرز و ۲۲۲ نمونهٔ وب هرز در بخش آموزش و شامل ۱۹۳۳ نمونهٔ وب غیرهرز و ۱۲۲ نمونهٔ وب هرز در بخش آزمایش می‌باشد. به‌منظور انجام آزمایش‌ها، از نمونه‌های با برچسب نامشخص صرفنظر شده است. همچنین به‌صورت تصادفی ۷۵ درصد از داده‌ها به‌عنوان داده‌های آموزش برای ساخت مدل و ۲۵ درصد از آن به‌عنوان داده‌های آزمون در نظر گرفته شده است.

در فرآیند اجرای مدل FFANB تعداد تکرار یا نسل می‌تواند ثابت یا براساس معیار توقف همچون خطا باشد. در این مقاله تعداد تکرار مدل FFANB در راستای طبقه‌بندی داده‌ها متغیر و در بازه ۱۰۰ تا ۵۰۰ مرتبه در نظر گرفته شده است. در این تحقیق جهت مقایسه عملکرد مدل FFANB با سایر مقالات، تعداد جمعیت برابر و یکسان در نظر گرفته شده است.

الگوریتم‌های فراابتکاری از ماهیت و رفتاری تصادفی در فرآیند جستجو و کشف بهینه سراسری برخوردار می‌باشند، لذا براساس رفتار تصادفی در فرایند ارائه تأثیر ماهیت تصادفی نشان داده می‌شود. در راستای ارزیابی ثبات و پایداری و عملکرد الگوریتم‌های فراابتکاری، این الگوریتم‌ها چندین مرتبه به‌طور مستقل اجرا می‌شوند و سپس بر اساس تحلیل آماری عملکرد آن‌ها مورد ارزیابی و سنجش قرار می‌گیرد. در این مقاله، در راستای ارزیابی و سنجش عملکرد و کارایی مدل FFANB در راستای طبقه‌بندی داده‌ها برای هر مجموعه داده به‌طور مجزا ۱۵ مرتبه فرآیند طبقه‌بندی مورد آزمایش و شبیه‌سازی قرار گرفته است.

در جدول(۱) نتایج مدل FFANB برمبنای تکرارهای مختلف نشان داده شده است. درصد صحت در تکرار ۱۰۰ و ۵۰۰ به‌ترتیب برابر با ۰/۸۱۲۳ و ۰/۹۵۸۴ است. طبق جدول(۱) می‌توان اظهار کرد که درصد صحت با افزایش

2- Precision
3- Recall
4- Accuracy

جدول۱: نتایج مدل FFANB برمبنای تکرارهای مختلف

صحت	F-Measure	بازخوانی	دقت	تعداد تکرار
۰/۸۱۲۳	۰/۸۱۳۹	۰/۸۲۶۵	۰/۸۰۱۶	۱۰۰
۰/۸۵۴۹	۰/۸۴۸۹	۰/۸۵۲۰	۰/۸۴۵۹	۲۰۰
۰/۸۹۷۲	۰/۸۸۷۹	۰/۸۹۴۶	۰/۸۸۱۳	۳۰۰
۰/۹۲۳۵	۰/۹۱۸۸	۰/۹۲۰۸	۰/۹۱۶۸	۴۰۰
۰/۹۵۸۴	۰/۹۲۷۶	۰/۹۳۱۱	۰/۹۲۴۱	۵۰۰

تکرارها، بیشتر شـــده و همچنین درصد معیارهای دقت و بازخوانی و F-Measure هم افزایش یافته است.

در جدول(۲) نتایج مدل FFANB برمبنای تعداد ویژگی‌ها برروی دسته مبتنی بر محتوا نشان داده شده است. تعداد ویژگی‌های دســـته مبتنی بـــر محتوا برابر بـــا ۹۶ ویژگی می‌باشـــد. اگر تعداد ویژگی‌ها کمتر باشـــد درصد صحت بیشـــتر خواهد بود. درصد صحت بـــا ۱۰ ویژگی برابر با ۰/۹۸۱۳، بـــا ٤۰ ویژگی برابر بـــا ۰/۹٦۳۹ و با ۹٦ ویژگی برابر با ۰/۹۳٥٦ اســـت. درصد دقت با ۱۰ ویژگی برابر با ۰/۹۷۹۱، بـــا ٤۰ ویژگی برابر بـــا ۰/۹٦۰٥ و با ۹٦ ویژگی برابر با ۰/۹۳۰۸ است. درصد بازخوانی با ۱۰ ویژگی برابر بـــا ۰/۹۸۰٥، با ٤۰ ویژگی برابر با ۰/۹٦٤۷ و با ۹٦ ویژگی برابر با ۰/۹۳۷۸ است.

در جدول(۳) نتایج مدل FFANB برمبنای تعداد ویژگی‌ها برروی دسته مبتنی بر پیوند نشـــان داده شده است. تعداد ویژگی‌هـــای دســـته مبتنی بر پیونـــد برابر بـــا ٤۱ ویژگی می‌باشد. درصد صحت با ۸ ویژگی برابر با ۰/۹۷٦۳، با ۳۰ ویژگی برابر با ۰/۹٦٤۹ و با ٤۱ ویژگی برابر با ۹٥ اســـت. درصد دقت با ۸ ویژگی برابر با ۰/۹٦۲٥، با ۳۰ ویژگی برابر بـــا ۰/۹٤۷۲ و با ٤۱ ویژگی برابر با ۰/۹۳۹۱ اســـت. درصد بازخوانی با ۸ ویژگی برابر با ۰/۹٦۳۲، با ۳۰ ویژگی برابر با ۰/۹٤۸۱ و با ٤۱ ویژگی برابر با ۰/۹٤۱۳ است.

در جدول(٤) نتایج مدل FFANB برمبنای تعداد ویژگی‌ها برروی دســـته مبتنی بر پیوند تبدیل یافته نشان داده شده اســـت. تعداد ویژگی‌های دسته مبتنی بر پیوند تبدیل یافته برابر با ۱۳۸ ویژگی می‌باشد. درصد صحت با ۱۰ ویژگی

جدول۲: نتایج مدل FFANB برمبنای تعداد ویژگی‌ها برروی دسته مبتنی بر محتوا

صحت	F-Measure	بازخوانی	دقت	تعداد ویژگی	
۰/۹۸۱۳	۰/۹۷۹۸	۰/۹۸۰۵	۰/۹۷۹۱	۱۰	مبتنی بر محتوا
۰/۹۸۰۰	۰/۹۸۰۳	۰/۹۸۲۱	۰/۹۷۸۶	۱۵	
۰/۹۷۸۱	۰/۹۷۷۲	۰/۹۷۸۹	۰/۹۷۵۶	۲۰	
۰/۹۷۵۵	۰/۹۷۷۳	۰/۹۸۰۲	۰/۹۷۴۵	۲۵	
۰/۹۶۷۳	۰/۹۶۳۹	۰/۹۶۶۵	۰/۹۶۱۳	۳۰	
۰/۹۶۳۹	۰/۹۶۲۶	۰/۹۶۴۷	۰/۹۶۰۵	۴۰	
۰/۹۶۱۶	۰/۹۶۴۵	۰/۹۶۸۲	۰/۹۶۰۹	۴۵	
۰/۹۵۴۹	۰/۹۵۴۴	۰/۹۵۷۳	۰/۹۵۱۶	۵۵	
۰/۹۵۱۸	۰/۹۵۲۶	۰/۹۵۴۶	۰/۹۵۰۶	۶۰	
۰/۹۴۷۵	۰/۹۴۷۵	۰/۹۵۰۶	۰/۹۴۴۴	۷۰	
۰/۹۴۳۶	۰/۹۴۸۲	۰/۹۴۹۳	۰/۹۴۷۱	۷۵	
۰/۹۴۰۵	۰/۹۴۴۶	۰/۹۴۶۹	۰/۹۴۲۳	۸۰	
۰/۹۳۸۲	۰/۹۳۴۷	۰/۹۳۴۱	۰/۹۳۵۴	۸۵	
۰/۹۳۵۶	۰/۹۳۴۳	۰/۹۳۷۸	۰/۹۳۰۸	۹۶	

برابـــر بـــا ۰/۷۹۰۰، با ۳۰ ویژگی برابر بـــا ۰/۷۷۳٤، با ٦۰ ویژگـــی برابر با ۰/۷٥٤۹، با ۹۰ ویژگـــی برابر با ۰/۷٤۲۳ و با ۱۳۸ ویژگی برابر با ۰/٦۹٥۲ اســـت. درصد دقت با ۱۰ ویژگی برابر با ۰/۷۹۱۱، با ۳۰ ویژگی برابر با ۰/۷۷٤٥، با ٦۰ ویژگی برابر با ۰/۷٥۱۲، با ۹۰ ویژگی برابر با ۰/۷٤٦٤ و با ۱۳۸ ویژگی برابر با ۰/٦۹٥۷ است.

٤-۲ مقایسه و ارزیابی

در جدول(٥) مقایســـه مدل FFANB بـــا مدل‌های دیگر نشان داده شده است. درصد صحت الگوریتم ژنتیک، شبکه عصبی مصنوعی و ماشـــین بردار پشتیبان به‌ترتیب برابر بـــا ۰/۹۲٤۱، ۰/۹۳٦٦ و ۰/۹٤ اســـت. درصد دقت الگوریتم بهینه‌ســـازی کلونی مورچه برابر با ۰/۸٤٤۰ است. درصد دقـــت و F-Measure در ماشـــین بردار پشـــتیبان برابر با ۰/۹۱ و ۰/۹٥ است. درصد دقت، بازخوانی و F-Measure در شبکه باور عمیق به‌ترتیب برابر با ۰/۹٦٥۹، ۰/۹٦۷٤ و ۰/۹٦٦٦ اســـت. درصد دقت، بازخوانی و F-Measure در شبکه باور عمیق با آدابوســـت به‌ترتیب برابر با ۰/۹۲۷٦، ۰/۹۳۲۷ و ۰/۹۳۰۱ اســـت. درصـــد دقـــت، بازخوانـــی و

جدول۳: نتایج مدل FFANB برمبنای تعداد ویژگی‌ها برروی دسته مبتنی بر پیوند

صحت	F-Measure	بازخوانی	دقت	تعداد ویژگی	
۰/۹۷۶۳	۰/۹۶۲۸	۰/۹۶۳۲	۰/۹۶۲۵	۸	مبتنی بر پیوند
۰/۹۷۲۹	۰/۹۶۱۸	۰/۹۶۲۶	۰/۹۶۱۱	۱۰	
۰/۹۷۰۳	۰/۹۵۸۳	۰/۹۵۸۰	۰/۹۵۸۷	۱۵	
۰/۹۶۸۲	۰/۹۵۷۶	۰/۹۵۹۶	۰/۹۵۵۶	۲۲	
۰/۹۶۴۹	۰/۹۴۷۶	۰/۹۴۸۱	۰/۹۴۷۲	۳۰	
۰/۹۶۲۱	۰/۹۴۴۶	۰/۹۴۵۵	۰/۹۴۳۸	۳۵	
۰/۹۵۵۵	۰/۹۴۲۸	۰/۹۴۴۲	۰/۹۴۱۴	۳۹	
۰/۹۵۰۰	۰/۹۴۰۲	۰/۹۴۱۳	۰/۹۳۹۱	۴۱	

جدول۴: نتایج مدل FFANB برمبنای تعداد ویژگی‌ها برروی دسته مبتنی بر پیوند تبدیل یافته

صحت	F-Measure	بازخوانی	دقت	تعداد ویژگی	
۰/۷۹۰۰	۰/۷۹۲۱	۰/۷۹۳۲	۰/۷۹۱۱	۱۰	مبتنی بر پیوند تبدیل یافته
۰/۷۸۵۲	۰/۷۸۴۲	۰/۷۸۵۸	۰/۷۸۲۶	۲۰	
۰/۷۷۳۴	۰/۷۷۶۷	۰/۷۷۸۹	۰/۷۷۴۵	۳۰	
۰/۷۶۰۰	۰/۷۶۵۳	۰/۷۶۹۱	۰/۷۶۱۵	۴۰	
۰/۷۵۸۲	۰/۷۵۸۰	۰/۷۵۹۸	۰/۷۵۶۳	۵۰	
۰/۷۵۴۹	۰/۷۵۵۶	۰/۷۶۰۰	۰/۷۵۱۲	۶۰	
۰/۷۵۰۷	۰/۷۵۶۹	۰/۷۵۹۰	۰/۷۵۴۹	۷۰	
۰/۷۴۶۹	۰/۷۴۲۶	۰/۷۴۳۵	۰/۷۴۱۸	۸۰	
۰/۷۴۲۳	۰/۷۴۸۰	۰/۷۴۹۷	۰/۷۴۶۴	۹۰	
۰/۷۲۱۶	۰/۷۲۷۲	۰/۷۲۹۳	۰/۷۲۵۱	۱۰۰	
۰/۷۱۸۹	۰/۷۲۱۸	۰/۷۲۲۵	۰/۷۲۰۹	۱۱۰	
۰/۷۰۳۵	۰/۷۰۲۹	۰/۷۰۴۷	۰/۷۰۱۱	۱۲۰	
۰/۷۰۱۳	۰/۷۰۷۴	۰/۷۰۹۲	۰/۷۰۵۶	۱۳۰	
۰/۶۹۵۲	۰/۶۹۷۵	۰/۶۹۹۴	۰/۶۹۵۷	۱۳۸	

F-Measure در ماشین بردار پشتیبان با آدابوست به‌ترتیب برابر با ۰/۹۱۴۵، ۰/۹۲۴٦ و ۰/۹۲۳۱ است.

درصد دقت در ماشین بردار پشتیبان [۱۹] برابر با ۰/۹۱ و معیار F-Measure برابر با ۰/۹۵ است. در میان مدل‌های مقایسه‌ای، شبکه باور عمیق از درصد دقت بیشتری برخوردار است. یادگیری عمیق برگرفته از شبکه عصبی مصنوعی است که شامل مجموعه‌ای از الگوریتم‌ها است که در تلاشند تا مفاهیم انتزاعی و پیچیده را در سطوح و لایه‌های مختلف مدل کنند. ایدهٔ اصلی شبکه‌های

جدول۵: مقایسه مدل FFANB با مدل‌های دیگر

صحت	F-Measure	بازخوانی	دقت	مدل‌ها	منابع
۰/۹۲۴۱	–	–	–	الگوریتم ژنتیک	[۲۳]
۰/۹۳۶۶	–	–	–	شبکه عصبی مصنوعی	
۰/۹۴۰۰	–	–	–	ماشین بردار پشتیبان	
			۰/۸۴۴۰	الگوریتم بهینه‌سازی کلونی مورچه	[۲۴]
–	۰/۹۶۶۶	۰/۹۶۷۴	۰/۹۶۵۹	شبکه باور عمیق	[۱۷]
–	۰/۹۳۰۱	۰/۹۳۲۷	۰/۹۲۷۶	شبکه باور عمیق+آدابوست	
–	۰/۹۲۳۱	۰/۹۲۴۶	۰/۹۱۴۵	ماشین بردار پشتیبان + آدابوست	
–	۰/۹۵۰۰	–	۰/۹۱۰۰	ماشین بردار پشتیبان	[۱۹]
۰/۹۵۸۴	۰/۹۲۷۶	۰/۹۳۱۱	۰/۹۲۴۱	مدل FFANB	–

مبتنی بر یادگیری عمیق، نزدیک شدن به بهترین راه حل‌ها برمبنای آموزش و آزمایش می‌باشد. برتری درصد صحت مدل FFANB در مقایسه با مدل‌های دیگر به دلیل این است که الگوریتم حاصلخیزی زمین‌های کشاورزی در چندین منطقه اقدام به کشف جواب‌های بهینه می‌نماید و سپس در بین آن‌ها، مقدار سراسری برای جواب مسئله انتخاب می‌شود.

۵. نتیجه‌گیری

صفحـــات وب در فضای اینترنت بـــه فضایی محبوب برای تبادل اطلاعات، ثبت و مشـــاهده اطلاعات تبدیل شده است. این میزان محبوبیت و جامعیت، هدف مناسبی جهت فعالیت‌های مخرب و وب هرزنویسان شده است. امنیت و نظـــارت بر صفحات وب یکی از مهم‌ترین ملزومات فضای اینترنـــت به دلیل جلوگیـــری از نفوذ بدافزارها می‌باشـــد. به‌طـــور کلی بدافزارها پس از ورود به سیســـتم می‌توانند اقداماتی نظیر ســـرقت اطلاعات، تخریب اطلاعات و افشای اطلاعات را انجام دهند. بنابراین ایجاد روشی که بتواند به صورت کارا به شناسایی و جلوگیری از وب هرز بپردازد، همواره مورد نیـــاز خواهد بود. پس از بیان کارهای انجام شده در مورد تشخیص صفحات وب هرز، یک مدل مبتنی بر یادگیری ماشـــین برای تشخیص وب هرز پیشنهاد شد. مـــدل FFANB ترکیبی از الگوریتـــم حاصلخیزی زمین‌های کشـــاورزی و بیز ســـاده بود. به منظـــور افزایش صحت الگوریتم طبقه‌بندی از انتخاب ویژگی استفاده شد. انتخاب ویژگی در مجموعه داده‌های اصلی در ابتدا با تعداد ویژگی کم شـــروع شـــد و به مرور تعداد ویژگی‌ها افزایش یافت. اگر تعداد ویژگی‌ها کمتر باشـــد آنگاه درصد صحت، دقت، بازخوانی و F-Measure بیشتر خواهد شد. نتایج ارزیابی نشان داد که مدل FFANB در مقایسه با الگوریتم‌های دیگر از درصد صحت بالاتری بهره‌مند است.

منابع

1. M. Yu, J. Zhang, J. Wang, J. Gao, T. Xu, R. Yu, The Research of Spam Web Page Detection Method Based on Web Page Differentiation and Concrete Cluster Centers, International Conference on Wireless Algorithms, Systems, and Applications, WASA 2018: Wireless Algorithms, Systems, and Applications, pp. 820-826, 2018.
2. X.Y. Lu, M.S. Chen, J.L. Wu, P.C. Chang, M.H. Chen, A novel ensemble decision tree based on under-sampling and clonal selection for web spam detection, Pattern Analysis and Applications, Vol. 21, Issue 3, pp. 741-754, 2018.
3. K. Hans, L. Ahuja, S.K. Muttoo, Performance Evaluation of Neural Network Training Algorithms in Redirection Spam Detection, Nature Inspired Computing pp 177-183, 2018.
4. R. Narayan, J.K. Rout, S.K. Jena, Review Spam Detection Using Semi-supervised Technique, Progress in Intelligent Computing Techniques: Theory, Practice, and Applications, pp. 281-286, 2018.
5. A. Sharaff, Spam Detection in SMS Based on Feature Selection Techniques, Emerging Technologies in Data Mining and Information Security, pp. 555-563, 2018.
6. M.K. Sohrabi, F. Karimi, A Feature Selection Approach to Detect Spam in the Facebook Social Network, Arabian Journal for Science and Engineering, Vol. 43, Issue 2, pp. 949-958, 2018.
7. N. Karimpour, F.S. Gharehchopogh, A New Approach to Detect Spam Emails Using the Hybrid Model of Ant Colony and Firefly Algorithms, Computing Science Journal (CSJ), Vol. 14, pp. 1-17, 2019
8. F.S. Gharehchopogh, M. Vafadar, M. Motamanfar, Improving the Invasive Weed Optimization algorithm with the nearest neighbor in the classification of spam emails, Computing Science Journal (CSJ), 10: 5464, 2018
9. H. Shayanfar, F.S. Gharehchopogh, Farmland fertility: A new metaheuristic algorithm for solving continuous optimization problems, Vol. 71, pp. 728-746, 2018.
10. A. McCallum, K. Nigam, A Comparison of Event Models for Naive Bayes Text Classification, In AAAI-98 workshop on learning for text categorization, Vol. 752, pp. 41-48, 1998.
11. Webpage Spam data set, http://chato.cl/webspam/datasets/uk2007/, [last available 2019.09.09]
12. J. Wu, A generalized tree augmented naive Bayes link prediction model, Journal of Computational Science, Vol. 27, pp. 206-217, 2018.
13. F. Asdaghi, A. Soleimani, An effective feature selection method for web spam detection, Knowledge-Based Systems, Vol. 166, 15 pp. 198-206, 2019.
14. K.L. Goh, A.K. Singh, Comprehensive Literature Review on Machine Learning Structures for Web Spam Classification, Procedia Computer Science, Vol. 70, pp. 434-441, 2015.
15. S. Singh, A.K. Singh, Web-Spam Features Selection Using CFS-PSO, Procedia Computer Science, Vol. 125, pp. 568-575, 2018.
16. J.F. Glez, D. Ruano-Ordas, J.R. MEndez, F. Fdez-Riverola, R. PavOn, A dynamic model for integrating simple web spam classification techniques, Expert Systems with Applications, Vol. 42, Issue 21, pp. 7969-7978, 2015.
17. Y. Li, X. Nie, R. Huang, Web spam classification method based on deep belief networks, Expert Systems with Applications, Vol. 96, pp. 261-270, 2018
18. A.H. Keyhanipour, B. Moshiri, Designing a web spam classifier based on feature fusion in the Layered Multi-population Genetic Programming framework, Proceedings of the 16th International Conference on Information Fusion, IEEE, pp. 53-60, 2013
19. R.C. Patil and D.R. Patil, Web spam detection using SVM classifier, 2015 IEEE 9th International Conference on Intelligent Systems and Control (ISCO), Coimbatore, IEEE, pp. 1-4, 2015.
20. S. Jain, S. Shukla, R. Wadhvani, Dynamic selection of normalization techniques using data complexity measures, Expert Systems with Applications, Vol. 106, pp. 252-262, 2018
21. M.A. Elaziz, A.A. Ewees, R.A. Ibrahim, S. Lu, Opposition-based moth-flame optimization improved by differential evolution for feature selection, Mathematics and Computers in Simulation, In press, corrected proof, Available online 2 August 2019
22. P. Galdi, R. Tagliaferri, Data Mining: Accuracy and Error Measures for Classification and Prediction, Encyclopedia of Bioinformatics and Computational Biology, Vol. 1, pp. 431-436, 2019
23. S.H.R. Mohammadi, M.A.Z. Chahooki, Web Spam Detection Using Multiple Kernels in Twin Support Vector Machine, Information Retrieval, pp. 1-10, 2016
24. B. Manaskasemsak, J. Jiarpakdee, A. Rungsawang, Adaptive Learning Ant Colony Optimization for Web Spam Detection, International Conference on Computational Science and Its Applications, ICCSA 2014: Computational Science and Its Applications – ICCSA 2014, Vol. 8584, pp. 642-653, 2014