هجدهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران
دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف، تهران، ۲۲-۲۴ اسفندماه ۱۳۹۱

تشخیص بازدیدکنندگان مخرب و غیر مخرب وب سایت‌ها مبتنی بر شبکه های باور بیزی

امیر حسین زارعی[۱]، سوده لایقی[۲]، مجید وفایی جهان[۳] ، مهرداد جلالی[۴]

[۱] گروه نرم افزار کامپیوتر، دانشگاه آزاد اسلامی، مشهد، ایران
Zareie@ymail.com

[۲] گروه نرم افزار کامپیوتر، دانشگاه آزاد اسلامی، مشهد، ایران
So_layeghi@yahoo.com

[۳] گروه نرم افزار کامپیوتر، دانشگاه آزاد اسلامی، مشهد، ایران
Vafaeijahan@mshdiau.ac.ir

[۴] گروه نرم افزار کامپیوتر، دانشگاه آزاد اسلامی، مشهد، ایران
Jalali@mshdiau.ac.ir

چکیده

در این مقاله بازدیدکنندگان وب سایت‌ها به چهار گروه انسان، روبات وب با رفتار خوب، روبات وب با رفتار مخرب و بازدیدکنندگان ناشناخته تقسیم بندی شده است. روبات‌های وب یا خزنده های وب برنامه های نرم افزاری هستند که دائماً به صورت خودکار ساختار لینک‌های وب سایت‌ها را مورد پیمایش قرار می‌دهند. هدف روبات‌های وب کشف و بازیابی محتوا و دانش از وب می‌باشد. این روبات‌ها هم به منظور اعمال مفیدی مانند کشف لینک‌های خراب و هم اعمال مخربی مانند حمله توزیع شده مختل کننده سرویس طراحی شده‌اند. تشخیص روبات‌های وبی که تلاش در تقلید رفتار انسان دارند به عنوان مهم‌ترین چالش دسته بندی است. در این مقاله با استفاده از شبکه باور بیزی، به عنوان رهیافت کاربردی جدید به منظور تشخیص بازدیدکنندگان وب سایت‌ها پرداخته شده است. روش پیشنهادی با روش‌های شبکه عصبی، ماشین بردار پشتیبان و *C4.5* استیوانوویچ و بیزین ساده استسپولو مقایسه شده است، که استفاده از روش پیشنهادی و استخراج سه ویژگی جدید باعث گردیده است که دقت روش پیشنهادی نتایج بهتری نسبت به سایر روش‌ها در تشخیص روبات‌های وب داشته باشد؛ و همچنین نشان داده شده است هر چه تعداد مجموعه داده آموزش بیشتر باشد دقت دسته بندی بالاتر خواهد بود.

کلمات کلیدی

روبات‌های وب، شبکه های باور بیزی، فایل ثبت وقایع، کاربرد کاوی وب

۱- مقدمه

با گسترش شبکه جهانی اینترنت، بسیاری از زوایای زندگی انسان نیز تحت تأثیر این پدیده قرار گرفته است. به طوری که در کشورهای صنعتی، بسیاری از امور روزمره، از خریدهای روزانه گرفته تا آموزش و تجارت، همگی از طریق اینترنت صورت می‌گیرد. با پیشرفت تکنولوژی‌های مرتبط با کامپیوتر و افزایش قدرت برنامه‌نویسان هر روزه برنامه های سودمندی بر روی اینترنت به دنیا عرضه می‌شود. در مقابل، قدرت هکرها نیز افزایش یافته و برنامه های مخرب قدرتمندتری تولید شده است. پس نیاز است تا در دنیای مجازی بتوان تنها با یک برنامه خودکار نرم افزاری تفاوت میان یک کاربر انسان و یک برنامه نرم افزاری را تشخیص داد.

تا کنون روش‌های مختلفی برای تشخیص روبات‌های وب پیشنهاد شده است. بومهارت و همکارانش روبات‌های وب را به چهار دسته: روش‌های ساده، تله، ارزیابی رفتار حرکتی روبات‌ها، مدل سازی الگوی رفتاری روبات‌های وب خلاصه کرده‌اند[1]. دسته بندی دوران و همکارانش چهار دسته: تحلیل نحوی ثبت وقایع، الگوی ترافیک، تکنیک‌های یادگیری تحلیلی، سیستم تست تورینگ می‌باشد [2]. طبق دسته بندی دوران روش‌های تحلیل نحوی ثبت وقایع شامل:

بررسی رشته های عامل کاربر [4] تکنیک تحلیل چند گامی ثبت وقایع[5]می‌باشد و روش‌های تحلیل الگوی ترافیک شامل: تشخیص روبات‌های وب از طریق تحلیل نحوی و تحلیل الگو [6]، تشخیص روبات‌های وب بر اساس الگوهای منبع درخواست [7]، تشخیص بر اساس الگو های نرخ درخواست [8]، تشخیص با استفاده از معیار ترافیک [9] می‌باشد.

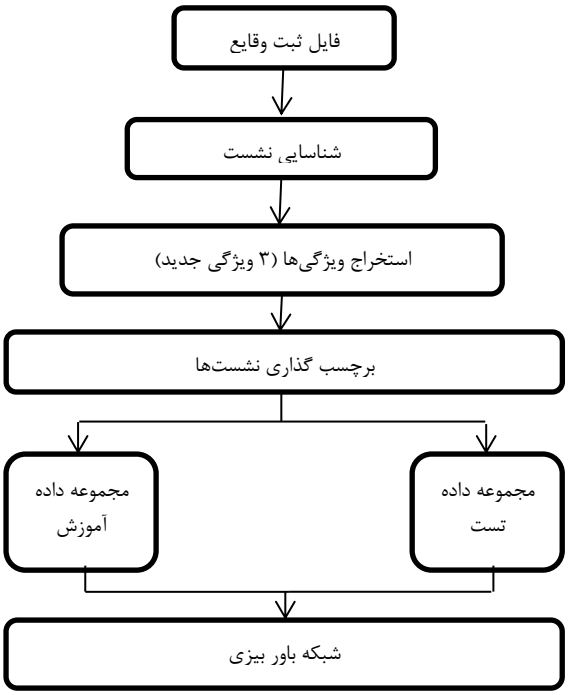
تکنیک‌های یادگیری تحلیلی شامل: تشخیص با استفاده از درخت تصمیم [10]، تشخیص با استفاده از شبکه عصبی [11]، تشخیص بر اساس شبکه بیزین ساده[12,13]، تشخیص با استفاده از مدل مخفی مارکوف [14]می‌باشد و تکنیک‌های سیستم تست تورینگ شامل: تشخیص بر اساس تست کپچا [15,16]، تشخیص با رفتار مروری انسان [17] می‌باشد. استیوانوویچ و همکارانش بازدیدکنندگان وب سایت‌ها را چهار دسته: انسان، روبات‌ها با رفتار خوب، روبات‌ها با رفتار مخرب، بازدیدکننده ناشناخته در نظر گرفته‌اند و با روش‌های دسته بندی، مشخص کرده‌اند هر نشست مربوط به کدام دسته می‌باشد [3].

در این مقاله هدف اصلی، کشف دانش از فایل‌های ثبت وقایع به منظور دسته بندی و تشخیص بازدیدکنندگان وب سایت‌ها به کمک روش شبکه باور بیزی می‌باشد. این فرایند شامل سه فاز اصلی به شرح زیر می‌باشد. ۱- پیش پردازش : که در این مرحله ورودی، فایل ثبت وقایع و خروجی، نشست کاربران می‌باشد. فاز پیش پردازش شامل یکسری ریز مرحله می‌باشد : (پاک‌سازی اطلاعات، شناسایی کاربران، شناسایی نشست). ۲-کشف الگو: که در این فاز ورودی، نشست کاربران می‌باشد و برای دسته بندی از روش شبکه باور بیزی استفاده شده است. ۳- تحلیل الگوی کشف شده: در این فاز با استفاده از معیار دقت دسته بندی به تحلیل الگوی کشف شده پرداخته شده است.

ادامه این مقاله به صورت زیر سازماندهی شده است: در بخش دوم آماده سازی مجموعه داده، بخش سوم روش شبکه باور بیزی، بخش چهارم نتایج آزمایشات و در بخش پنجم نتیجه گیری مقاله ارائه شده است.

۲- آماده سازی مجموعه داده

کلیه مراحل آماده سازی مجموعه دادها که در شکل (۱) مشاهده می‌شود به این شرح است: ۱- ورودی فایل ثبت وقایع ۲- شناسایی نشست ۳- استخراج ویژگی برای هر نشست (استفاده از ویژگی‌های روش‌های قبلی و استخراج سه ویژگی جدید) ۴- بر چسب گذاری هر نشست ۵-مجموعه داده را به دو دسته مجموعه داده آموزش و تست تقسیم می‌کنیم ۶- استفاده از روش شبکه باور بیزی برای دسته بندی. با فرض صحت فرایند برچسب زدن، هدف اصلی بررسی دقت طبقه بندی است [3]. هر ورودی فایل ثبت وقایع به ترتیب حاوی اطلاعات زیر می‌باشد: تاریخ، ساعت، متد (GET,HEAD,...)، فایل درخواست شده، آدرس IP کلاینت، رشته عامل کاربر، کوکی، رشته ارجاع، کد پاسخ، تعداد بایتی که از کلاینت به سرور ارسال می‌شود.



شکل (۱) : مراحل آماده سازی سیستم پیشنهادی

۲-۱- شناسایی نشست

ابتدا تمام درخواست‌های HTTP بر اساس IP و عامل کاربر[1] یکسان گروه بندی می‌شوند سپس از یک رویکرد وقفه برای شکستن این گروه‌ها به زیر گروه های دیگر استفاده می‌شود (اگر زمان وقفه بین دو درخواست متوالی از یک زیرگروه IP بیش از یک حد آستانه باشد این طور فرض شود که آن کاربر، یک نشست جدید را شروع کرده است). معمولاً حد آستانه را ۳۰ دقیقه در نظر می‌گیرند. بدون شک یک عدم قطعیت در این رویکرد وجود دارد [3].

۲-۲- استخراج ویژگی برای هر نشست

پایه انتخاب ویژگی‌ها را بر اساس مطالعاتی که از رفتار روبات‌های وب در [1,3,10,18,19] داشتیم در نظر می‌گیریم. ویژگی‌هایی که از هر نشست استخراج می‌شود به شرح زیر است: ۱-حداکثر نرخ کلیک، ۲-مدت نشست، ۳-درصد درخواست تصویر، ۴-درصد درخواست صفحات HTML، ۵-درصد پاسخ خطای 4xx، ۶-درخواست فایل Robots.txt، ۷-درصد درخواست‌های با ارجاع خالی، ۸-درصدی از درخواست HTTP از نوع HEAD ، ۹-نرخ درخواست دنباله متوالی، ۱۰-عمق درخواست صفحه، ۱۱-تعداد بایتی که از کلاینت به سرو ارسال می‌شود. ویژگی‌های یک تا یازده قبلاً برای تشخیص روبات‌های وب استفاده شده است بعلاوه سه ویژگی جدید که به شرح زیر آورده شده است، ویژگی‌های هستند که از هر نشست استخراج می‌شود.

کاربر انسان برای مشاهده صفحات وب نیاز به مرورگرهای وب دارد در حالی که روبات‌های وب نیازی به استفاده از مرورگرهای وب ندارد. ۱۲-درصدی از درخواست فایل CSS: مرورگرهای وب به صورت خودکار یک درخواست برای فایل CSS ارسال می‌کنند در حالی که روبات‌های وب نیازی به مشاهده فایل CSS ندارد، پس اگر در یک نشست تمام درخواست‌ها از یک نوع منبع باشد و فایل CSS مشاهده نشده باشد آنگاه آن نشست مشکوک به روبات وب می‌باشد. ۱۳-درصدی از درخواست فایل‌های دیگر: مرورگرهای وب به صورت خودکار کلیه منابع جاسازی شده در یک صفحه وب را مشاهده می‌کنند در یک نشست اگر تنها یک نوع از منابع درخواست شده باشد آن نشست مشکوک به روبات وب می‌باشد. بنابراین اگر یک صفحه وب دیده شده اما نه همه منابع جاسازی شده در آن، آنگاه آن نشست می‌توان مشکوک به روبات وب باشد. ۱۴-درصدی از کوکی‌ها: کوکی‌ها اطلاعاتی هستند که سرور HTTP می‌تواند به همراه منبع درخواست شده به ماشین کاربر ارسال کند. مرورگر کاربر ممکن است این اطلاعات را ذخیره کند و متعاقباً هنگام ارسال درخواست‌های بعدی اطلاعات آن را به سرور HTTP پس بفرستد. اگر درصد کوکی‌ها در یک نشست صفر باشد، آنگاه آن نشست می‌توان مشکوک به روبات باشد.

۳-۲- بر چسب گذاری هر نشست

برچسب گذاری نشست‌ها بر اساس چهار کلاس به شرح زیر است [20-24] : ۱- مقایسه عامل کاربر با لیست به روز شده عامل‌های کاربر مرورگرهای شناخته شده: در صورتی که رشته عامل کاربر با لیست به روز شده عامل‌های کاربر مرورگرهای شناخته شده مطابقت داشته باشد و به فایل Robots.txt دسترسی پیدا نکرده باشد به عنوان کاربر انسان برچسب گذاری می‌شود. ۲- مقایسه عامل کاربر با لیست به روز شده عامل‌های کاربر روبات‌های وب با رفتار خوب شناخته شده: در صورتی که رشته عامل کاربر با لیست به روز شده عامل‌های کاربر روبات‌های وب با رفتار خوب شناخته شده مطابقت داشته باشد به عنوان روبات وب با رفتار خوب برچسب گذاری می‌شود. ۳- مقایسه عامل کاربر با لیست به روز شده عامل‌های کاربر روبات‌های وب مخرب شناخته شده: در صورتی که رشته عامل کاربر با لیست به روز شده عامل‌های کاربر روبات‌های وب مخرب شناخته شده مطابقت داشته باشد به عنوان روبات وب مخرب برچسب گذاری می‌شود. ۴- بقیه نشست‌ها به عنوان کاربر ناشناخته برچسب گذاری می‌شود.

۳- شبکه باور بیزی

یک شبکه بیزی توزیع احتمالی توأم مجموعه ای از v متغیر $v = (x_1,\ldots,x_v)$ را به صورت یک گراف جهت دار بدون دور و مجموعه ای از جداول احتمال شرطی برای هر متغیر، نشان می‌دهد [25-27]. مجموعه والد های x_i که با π_i نشان داده می‌شود، مجموعه نودهایی هستند که در شبکه یک کمان از آن‌ها به نود x_i وجود دارد. ساختار شبکه بیانگر این مسئله است که هر نود با دانستن اطلاعات والدین بلافصل اش از متغیر های غیر فرزند خود مستقل است. بنابراین احتمال یک واقعه دلخواه $x = (x_1,\ldots,x_v)$ می‌تواند به صورت $P(X)=\prod_{i=1}^{v} P(x_i \mid \pi_i)$ محاسبه شود. به طور کلی نمایش توزیع احتمال توأم مجموعه ای از v متغیر گسسته به فضایی از مرتبه نمایی نسبت به v نیاز دارد. شبکه های بیزی این فضا را به مرتبه نمایی نسبت به $\max_{i\in\{1,\ldots,v\}}|\pi_i|$ کاهش می‌دهند [34].

۳-۱- آموزش شبکه های باور بیزی

برای توصیف یک شبکه بیزی باید دو مورد را فراهم آورد: توپولوژی (ساختار) گراف، و پارامتر های جداول احتمال شرطی مربوط به تمام متغیرها. مسئله ای که معمولاً برای ساخت شبکه های بیزی مطرح است، یادگیری هر دوی این موارد بر اساس مجموعه مثال‌های آموزشی است [29-32]. پارامتر های یک شبکه بیزی در واقع مقادیر موجود در جداول احتمال شرطی هر متغیر می‌باشند. ساده‌ترین حالت آموزش پارامتر های شبکه وقتی است که ساختار شبکه مشخص بوده و کلیه متغیرها کاملاً مشاهده پذیر باشند، که در این تحقیق کلیه متغیرها کاملاً مشاهده پذیر می‌باشد. در این حالت فرض می‌شود که هدف یادگیری پیدا کردن برآورد امکان بیشینه[۲] پارامتر های جداول احتمال شرطی هر یک از متغیرها می‌باشد [34].

واضح است وقتی که ساختار شبکه نیز معلوم نباشد باید ابتدا ساختار و سپس پارامتر های آن را آموزش داد. مسئله آموزش ساختار راه حل کم هزینه و اولیه ای نداشته و مستلزم استفاده از روش‌های جستجو است. فرایند یادگیری ساختار به دو دسته محدودیت گرا [33] و امتیازگرا[26] تقسیم می‌شود. دسته دوم بسیار محبوب‌تر بوده و در این تحقیق نیز از دسته دوم، الگوریتم اتوماتای یادگیر استفاده می‌شود. مسئله یافتن بهترین شبکه، یک مسئله NP-Hard می‌باشد [27]. بنابراین برای جستجوی شبکه نزدیک به بهینه از روش‌های جستجوی اکتشافی استفاده می‌گردد.

۳-۲- آموزش ساختار توسط اتوماتاهای یادگیر

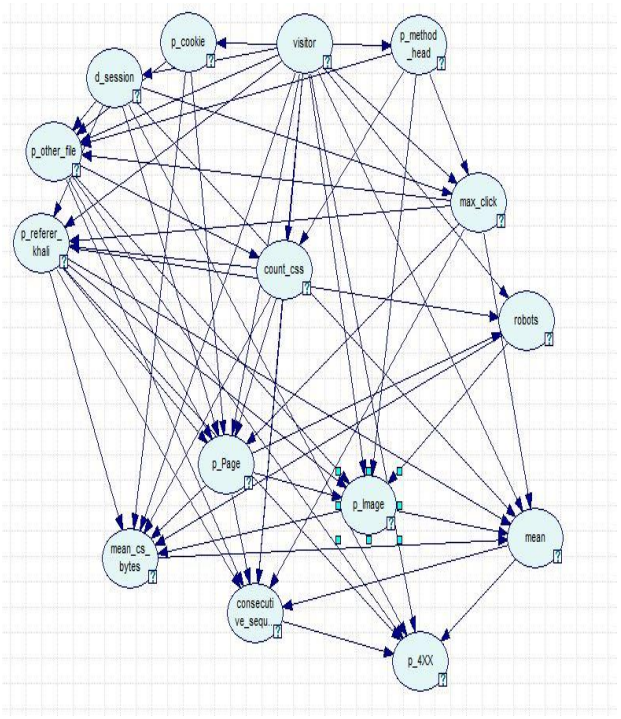
اتوماتاهای یادگیر به دو دسته غیر قطعی با ساختار ثابت و غیر قطعی با ساختار متغیر تقسیم می‌شوند. در این تحقیق یک روش مبتنی بر اتوماتاهای یادگیر با ساختار متغیر برای آموزش ساختار شبکه های بیزی استفاده شده است. این روش از چندین اتوماتای یادگیر برای جستجوی ساختار نزدیک به بهینه استفاده می‌کند. این ساختار نزدیک به بهینه، باید ساختاری باشد که مقدار بیشینه تابع ارزیابی شبکه را داشته باشد [37-39]. این روش از مجموعه ای از اتوماتاها برای جستجو استفاده می‌کند. در آن به ازای هر یک از اتصالات ممکن در

گراف شبکه یک اتوماتا در نظر گرفته می‌شود. این اتصالات می‌توانند هر یک از کمان‌های یک گراف کامل با n گره باشند.

کلیه اتوماتاها به صورت موازی در هر مرحله جستجو، یک عمل را بر مبنای بردار احتمالات خود انتخاب می‌کنند. عمل انتخاب شده هر اتوماتا به شبکه اعمال شده و باعث تولید یک گراف می‌گردد. سپس پارامتر های جداول احتمال شرطی هر یک از متغیرها توسط روش برآورد امکان بیشینه محاسبه می‌شود و گراف جهت دار تولید شده مورد بررسی قرار می‌گیرد تا یک گراف قابل قبول باشد. برای این منظور باید اولاً متغیر کلاس هیچ والدی نداشته باشد و ثانیاً هیچ دوری در گراف بدست آمده وجود نداشته باشد. اگر شبکه حاصل شده قابل قبول نباشد کلیه اتوماتاها جریمه شده و اگر شبکه قابل قبول باشد بر اساس مقایسه امتیاز مرحله فعلی و قبلی مورد پاداش و یا جریمه قرار می‌گیرد. جستجو تا جایی ادامه می‌یابد که امتیاز مرحله فعلی از امتیاز مرحله قبلی بهتر نشود. هر اتوماتا مستقلاً عملی را از مجموعه اعمالش انتخاب کرده و به محیط اعمال می‌کند سپس همه اتوماتاها از محیط پاسخ یکسانی دریافت می‌کنند. روش آموزش ساختار پیشنهادی در [38] با الگوریتم‌های رقیب مقایسه شده است و نتایج نشان داد که این روش جستجوی علاوه بر داشتن هزینه محاسباتی کمتر، می‌تواند ساختار با امتیازی نزدیک‌تر به امتیاز بهینه را بیابد.

توپولوژی شبکه باور بیزی حاصل از مجموعه ای از اتوماتاها برای تشخیص روبات‌های وب در شکل (۲) نشان داده شده است. با در نظر گرفتن ۱۴ ویژگی و یک نود کلاس، با چهار حالت بازدید کننده ناشناخته، کاربر انسان، روبات وب خوب و روبات وب مخرب، کران بالای ۷ برای حداکثر تعداد والد های هر گره، تابع $(LL(B|D))$ تابع امتیازدهی [34]، رکورد های استخراج شده از فایل ثبت وقایع به عنوان ورودی، خروجی این الگوریتم با مفروضات ذکر شده ماتریسی است با ۱۵ سطر و ستون، معادل ۱۵ گره شبکه که یک‌های ماتریس نشان دهنده رابطه علی یا یک کمان در گراف تولیدی است.

پس از آموزش ساختار و پارامتر شبکه، باید بتوان آن را مورد پرسش قرار داد. منظور از پرسش پیش بینی مقدار تعدادی متغیر مجهول با دانستن مقدار تعدادی متغیر معلوم می‌باشد. در واقع شبکه باید بتواند مقدار یک احتمال شرطی خواسته شده را محاسبه نماید [28]. اگر مقدار کلیه متغیرها در مثال‌ها معلوم باشد، مسئله استنتاج چندان پیچیده نخواهد بود. در حالتی که در مثال‌ها مقادیر نامعلوم وجود داشته باشد باید از روش‌های استنتاج تقریبی استفاده کرد. در این مقاله موتورهای استنتاج انتخاب شده، دارای CPD[3] گسسته می‌باشند که سازگار با متغیرهای این مطالعه باشند؛ و به دلیل اینکه همه مقادیر متغیرها در مثال‌ها معلوم است از استنتاج دقیق استفاده کرده‌ایم، در این شرایط رایج‌ترین راه حل تبدیل شبکه باور بیزی به یک درخت با خوشه بندی نودها به هم و ساخت یک درخت اتصال می‌باشد.



شکل (۲) : ساختار شبکه باور بیزی بر روی مجموعه داده پارس وب سایت[4] با استفاده از الگوریتم اتوماتای یادگیر برای تشخیص روبات‌های وب

۴- نتایج و شبیه سازی‌ها

نشست‌های استخراج شده از فایل ثبت وقایع شامل انواع روبات‌های وب متنی و غیر متنی از قبیل موتورهای جستجوی متفاوت، جمع کنندگان تصاویر و ... می‌باشد. در جدول (۱) مشخصات دو مجموعه داده پارس هاستینگ[5] و پارس وب سایت نشان داده شده است [35,36]. همان طور که در این جدول مشاهده می‌کنید تعداد نشست‌های استخراج شده در مجموعه داده پارس وب سایت بیشتر از مجموعه داده پارس هاستینگ است و نتایج بدست آمده از مجموعه داده پارس وب سایت از دقت کلی بالاتری برخوردار می‌باشد. ۸۰ درصد از مجموعه داده به عنوان مجموعه داده های آموزشی و ۲۰ درصد باقی مانده به عنوان مجموعه داده های تست در نظر گرفته شده است که مشخصات این دو مجموعه داده در جدول (۲) و جدول (۳) قابل مشاهده است.

جدول (۱) : مشخصات مجموعه داده‌ها

مشخصات / مجموعه داده	تعداد کل نشست‌ها	تعداد بازدید کننده ناشناخته	تعداد کاربر انسان	تعداد روبات وب غیر مخرب	تعداد روبات وب مخرب
PH	۲۴۲۲	۱۰۳۹	۱۰۹۲	۱۲۷	۱۶۴
PW	۱۳۷۳۴	۳۵۶۶	۸۵۰۷	۱۴۵۹	۲۰۲

جدول (۲) : مشخصات مجموعه داده‌ها آموزش

مشخصات / مجموعه داده	تعداد کل نشست‌ها	تعداد بازدید کننده ناشناخته	تعداد کاربر انسان	تعداد روبات وب غیر مخرب	تعداد روبات وب مخرب
PH	۱۹۳۸	۹۵۲	۷۷۱	۱۰۲	۱۱۳
PW	۱۰۹۸۷	۳۲۴۶	۶۴۳۶	۱۱۴۴	۱۶۱

جدول (۳) مشخصات مجموعه داده‌ها تست

مشخصات / مجموعه داده	تعداد کل نشست‌ها	تعداد بازدید کننده ناشناخته	تعداد کاربر انسان	تعداد روبات وب غیر مخرب	تعداد روبات وب مخرب
PH	۴۸۴	۸۷	۳۲۱	۲۵	۵۱
PW	۲۷۴۷	۳۲۰	۲۰۷۱	۳۱۵	۴۱

یک روش ساده برای تخمین خطا بین دسته‌ها، استفاده از ماتریس تداخل[۶] است. ماتریس تداخل، نحوه‌ی توزیع خطا روی دسته های مختلف را مشخص می‌کند.که در جدول (۴) نمونه ای از این ماتریس مشاهده می‌شود. با توجه به این ماتریس نحوه محاسبه صحت و فراخوانی برای کلاس A به ترتیب در فرمول‌های (۱) و (۲) آورده شده است [40].

جدول (۴) : ماتریس تداخل

		کلاس واقعی			
		A	B	C	D
نتیجه پیش بینی شده	A	TP_A	E_{AB}	E_{AC}	E_{AD}
	B	E_{BA}	TP_B	E_{BC}	E_{BD}
	C	E_{CA}	E_{CB}	TP_C	E_{CD}
	D	E_{DA}	E_{DB}	E_{DC}	TP_D

$$Precision_A = TP_A / (TP_A + E_{AB} + E_{AC} + E_{AD}) \quad (۱)$$
$$Recall_A = TP_A / (TP_A + E_{BA} + E_{CA} + E_{DA}) \quad (۲)$$

برای مجموعه داده پارس هاستینگ با استفاده از روش پیشنهادی، دقت کلی[۷]، میانگین فراخوانی[۸]، میانگین صحت[۹] و F_1 به ترتیب برابر ۰,۹۵۲۴، ۰,۹۰۵۲، ۰,۸۸۴۳ و ۰,۸۹۴۷ بدست آمده است. ماتریس تداخل برای مجموعه داده پارس هاستینگ در جدول (۵) و دقت دسته بندی در جدول (۶) قابل مشاهده است.

جدول (۵) : ماتریس شبکه بیزی برای مجموعه داده پارس هاستینگ

Precision	کلاس واقعی					
	روبات وب مخرب	روبات وب غیر مخرب	کاربر انسان	بازدید کننده ناشناخته		نتیجه پیش بینی شده
۰,۹۰۸۰	۰	۰	۸	۷۹	بازدید کننده ناشناخته	
۰,۹۸۷۵	۰	۰	۳۱۷	۴	کاربر انسان	
۰,۷۲	۶	۱۸	۰	۱	روبات وب غیر مخرب	
۰,۹۲۱۶	۴۷	۴	۰	۰	روبات وب مخرب	
	۰,۸۸۶۸	۰,۸۱۸۱	۰,۹۷۵۴	۰,۹۴۰۵	Recall	

جدول (۶) : دقت دسته بندی بر روی مجموعه داده پارس هاستینگ

جمع	روبات وب مخرب	روبات وب غیر مخرب	کاربر انسان	بازدید کننده ناشناخته	
۴۶۱	۴۷	۱۸	۳۱۷	۷۹	مقادیر صحیح دسته بندی شده
۴۸۴	۵۱	۲۵	۳۲۱	۸۷	تعداد نشست‌های واقعی
۰,۹۵۲۴	۰,۹۲۱۵	۰,۷۲	۰,۹۸۷۵	۰,۹۰۸۰	دقت

برای مجموعه داده پارس وب سایت با استفاده از روش پیشنهادی، دقت کلی، میانگین فراخوانی، میانگین صحت و F_1 به ترتیب برابر ۰,۹۵۸۸، ۰,۷۸۷۰، ۰,۷۴۱۲ و ۰,۷۶۳۴ بدست آمده است. ماتریس تداخل برای مجموعه داده پارس وب سایت در جدول (۷) و دقت دسته بندی در جدول (۸) قابل مشاهده است.

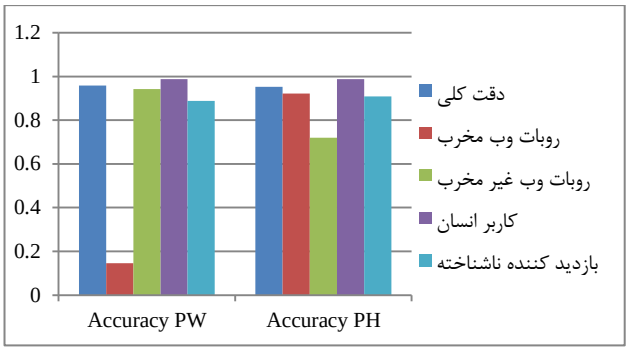
جدول (۷) : ماتریس شبکه بیزی برای مجموعه داده پارس وب سایت

Precision	کلاس واقعی					
	روبات وب مخرب	روبات وب غیر مخرب	کاربر انسان	بازدید کننده ناشناخته		نتیجه پیش بینی شده
۰,۸۸۷۵	۲	۵	۲۹	۲۸۴	بازدید کننده ناشناخته	
۰,۹۸۸۴	۴	۵	۲۰۴۷	۱۵	کاربر انسان	
۰,۹۴۲۸	۸	۲۹۷	۲	۸	روبات وب غیر مخرب	
۰,۱۴۶۳	۶	۴	۳۰	۱	روبات وب مخرب	
	۰,۳	۰,۹۵۵۰	۰,۹۷۱۱	۰,۹۲۲۱	Recall	

جدول (۸) : دقت دسته بندی بر روی مجموعه داده پارس وب سایت

جمع	روبات وب مخرب	روبات وب غیر مخرب	کاربر انسان	بازدید کننده ناشناخته	
۲۶۳۴	۶	۲۹۷	۲۰۴۷	۲۸۴	مقادیر صحیح دسته بندی شده
۲۷۴۷	۴۱	۳۱۵	۲۰۷۱	۳۲۰	تعداد نشست‌های واقعی
۰,۹۵۸۸	۰,۱۴۶۳	۰,۹۴۲۸	۰,۹۸۸۴	۰,۸۸۷۵	دقت

نمودار میله ای مقایسه دقت دسته بندی دو مجموعه داده پارس هاستینگ و پارس وب سایت در شکل (۳) قابل مشاهده است مجموعه داده آموزش پارس وب سایت بیشتر از مجموعه داده آموزش پارس هاستینگ است و قابل مشاهده است که دقت کلی دسته بندی بر روی مجموعه داده پارس وب سایت بیشتر از پارس هاستینگ می‌باشد.

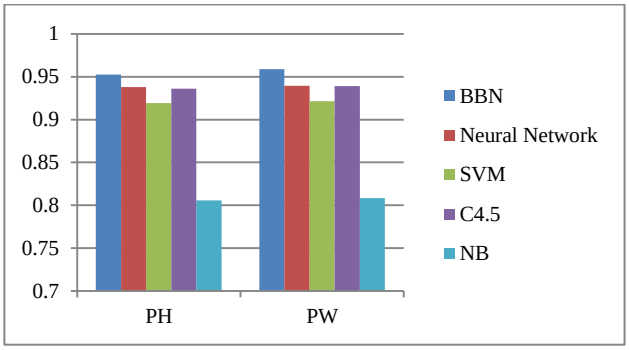


شکل (۳) : نمودار میله ای مقایسه دقت دسته بندی دو مجموعه داده

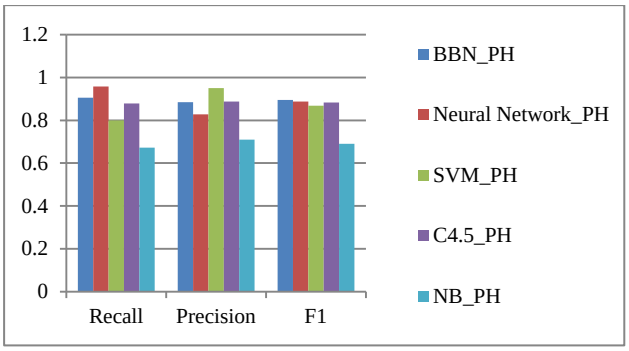
برای ارزیابی دقیق‌تر کارایی روش پیشنهادی آن را با روش استسوپلو، مبتنی بر مدل دسته بندی بیزین ساده(NB) [13] و با روش‌های شبکه عصبی، SVM و C4.5 استیوانوویچ [3] که دو ویژگی جدید نسبت به روش استسوپلو ارائه کرده است مقایسه شده است. مقایسه الگوریتم پیشنهادی با کارهای انجام شده در جدول (۹)، شکل (۴) و شکل (۵) قابل مشاهده است. نتایج بدست آمده از ارزیابی‌های مختلف نشان داد که دقت کلی دسته بندی در مدل پیشنهادی با استفاده از مجموعه داده پارس هاستینگ برابر ۰,۹۵۲۴ و با استفاده از مجموعه داده پارس وب سایت برابر ۰,۹۵۸۸ است و بهتر از سایر روش‌های انجام شده، است و همچنین برروی مدل پیشنهادی و کارهای انجام شده، F_1 با استفاده از مجموعه داده پارس هاستینگ و پارس وب سایت، محاسبه شده است نتایج نشان داد که F_1 برروی مدل پیشنهادی در مقایسه با کارهای انجام شده بهتر است.

جدول (۹) : مقایسه دقت کلی دسته بندی الگوریتم پیشنهادی با کارهای انجام شده

الگوریتم / مجموعه داده	BBN	Neural Network	SVM	C4.5	NB
PH	۰,۹۵۲۴	۰,۹۳۸۰	۰,۹۱۹۴	۰,۹۳۵۹	۰,۸۰۵۷
PW	۰,۹۵۸۸	۰,۹۳۹۵۷	۰,۹۲۱۳۶	۰,۹۳۹۲	۰,۸۰۸۵



شکل (۴) : نمودار میله ای مقایسه دقت کلی دسته بندی الگوریتم پیشنهادی با کارهای انجام شده



شکل (۵) : نمودار میله ای مقایسه فراخوانی، صحت و F_1 الگوریتم پیشنهادی بر روی مجموعه داده پارس هاستینگ با کارهای انجام شده

۵- نتیجه گیری

در این مقاله سه ویژگی جدید و روشی مبتنی بر شبکه باور بیزی به عنوان رهیافت کاربردی جدید، برای تشخیص بازدیدکنندگان وب سایت‌ها با استفاده از فایل ثبت وقایع مربوط به سرورهای پارس هاستینگ و پارس وب سایت ارائه شده است. یکی از مهم‌ترین کاربرد های شبکه های باور بیزی، استفاده از آن‌ها به عنوان دسته بندی کننده است. شبکه های باور بیزی یکی از معدود متدولوژی های است که از عهده پیچیدگی روابط، تعداد متغیرهای زیاد و عدم قطعیت روابط بر می‌آید و همچنین می‌تواند ارتباط بین متغیرها را بدست آورد.

تشخیص و دسته بندی روبات‌های وبی که تلاش در تقلید رفتار انسان دارند به عنوان مهم‌ترین چالش دسته بندی است. با به‌کارگیری روش پیشنهادی تا حدودی این چالش کاهش داده شده است. در این مقاله با استفاده از پارامتر دقت و F_1 دسته بندی سیستم پیشنهادی مورد ارزیابی قرار گرفته است روش پیشنهادی با روش استسوپلو مبتنی بر مدل دسته بندی بیزین ساده(NB) و با روش‌های شبکه عصبی، ماشین بردار پشتیبان و C4.5 استیوانوویچ که دو ویژگی جدید نسبت به روش استسوپلو ارائه کرده است مقایسه شده است. نتایج بدست آمده از ارزیابی‌های مختلف نشان داد که دقت کلی دسته بندی در مدل پیشنهادی و با استفاده از سه ویژگی جدید، بهتر از سایر روش‌های انجام شده، است و همچنین دقت دسته بندی الگوریتم پیشنهاد شده بر روی دو مجموعه داده پارس وب سایت و پارس هاستینگ مورد بررسی قرار گرفت. تعداد رکوردهای آموزش مجموعه داده پارس وب سایت بیشتر از مجموعه داده پارس هاستینگ است و مشاهده شد دقت کلی دسته بندی بر روی مجموعه داده پارس وب سایت بالاتر از پارس هاستینگ است و در پایان نشان داده شد که F_1 بر روی روش پیشنهادی و مجموعه داده پارس هاستینگ در مقایسه با کارهای انجام شده بهتر است.

مراجع

[1] C. Bomhardt, W. Gaul, L. Schmidt-Thieme, "Web robot detection techniques: overview and limitations", *Data Mining and Knowledge Discovery*, (2011), Vol.22 183-210.

[2] D. Stevanovic, A. An, and N. Vlajic, "Feature evaluation for web crawler detection with data mining techniques", *Expert Systems with Applications: An International Journal*, (2012), Vol.39 8707-8717.

[3] T. Kabe, M. Miyazaki, "Determining WWW user-agents from server access log", *In: Proceedings of seventh international conference on parallel and distributed systems*, (2000), 173–178.

[4] P. Huntington, D. Nicholas, and H.R.J. M., "Web robot detection in the scholarly information environment", *Journal of Information Science*, (2008), Vol.34 726-741.

[5] N. Geens, J. Huysmans, and J. Vanthienen, "Evaluation of Web Robot Discovery Techniques: A Benchmarking Study", *in Proc. Industrial Conference on Data Mining*, (2006), 121-130.

[6] W. Guo, S. Ju, and Y. Gu, "Web robot detection techniques based on statistics of their requested URL resources", *in Proc. CSCWD (1)*, (2005), 302-306.

[7] O. Duskin, DG. Feitelson, "Distinguishing humans from robots in web search logs: preliminary results using query rates and intervals", *In: Proceedings of 2009 workshop on Web Search Click Data*, (2009), 15-19.

[8] X. Lin, L. Quan, and H. Wu, "An Automatic Scheme to Categorize User Sessions in Modern HTTP Traffic", *in Proc. GLOBECOM*, (2008), 1485-1490.

[9] P. Tan and V. Kumar, "Discovery of Web Robot Sessions Based on their Navigational Patterns", *Data Mining and Knowledge Discovery*, (2002), Vol.6 9-35.

[10] D. Stevanovic, N. Vlajic, A. An, "Detection of malicious and non-malicious website visitors using unsupervised neural network learning", *Applied Soft Computing*, (2013), Vol.13 698-708.

[11] A. Stassopoulou and M.D. Dikaiakos, "A Probabilistic Reasoning Approach for Discovering Web Crawler Sessions", *in Proc. APWeb/WAIM*, (2007), 265-272.

[12] A. Stassopoulou and M.D. Dikaiakos, "Web robot detection: A probabilistic reasoning approach", *Computer Networks: The International Journal of Computer and Telecommunications Networking*, (2009), Vol.53 265-278.

[13] WZ. Lu, SZ. Yu, "Web robot detection based on hidden Markov model", *In: Proceedings of international conference on communications, circuits and systems*, (2006), 1806–1810.

[14] L.V. Ahn, M. Blum, N.J. Hopper, and J. Langford, "CAPTCHA: Using Hard AI Problems for Security", *Proceedings of the 22nd international conference on Theory and applications of cryptographic techniques*, (2003), 294-311.

[15] M. Motoyama, K. Levchenko, C. Kanich, D. McCoy, G.M. Voelker, and S. Savage, "Re: CAPTCHAs-Understanding CAPTCHA-Solving Services in an Economic Context", *Proceedings of the 19th USENIX conference on Security*, (2010), 435-462.

[16] K. Park, V.S. Pai, K. Lee, and S.B. Calo, "Securing Web Service by Automatic Robot Detection", *In Proceedings of USENIX Annual Technical Conference, General Track.*, (2006), 255-260.

[17] M.D. Dikaiakos, A. Stassopoulou, and L. Papageorgiou, "Characterizing Crawler Behavior from Web Server Access Logs", *in Proc. EC-Web*, (2003), 369-378.

[18] M.D. Dikaiakos, A. Stassopoulou, L. Papageorgiou, "An investigation of WWW crawler behavior: characterization and metrics", *Computer Communications*, (2005), Vol.28 880–897.

[19] http://www.user-agents.org/, August 2011.

[20] http://www.botsvsbrowsers.com, August 2011.

[21] http://www.useragentstring.com, August 2011.

[22] www.robotstxt.org, 2007.

[23] http://user-agent-string.info/list-of-ua/bots-ip, 2012.

[24] T. M. Mitchell, "Machine Learning", *presented at McGraw Hill series in computer science*, (1997), 1-414.

[25] K. P. Murphy, "An Introduction to Graphical Models", *Technical Report, Intel Research Technical Report*, (2001).

[26] W. Lam and F. Bacchus, "Learning Bayesian Belief Networks: An Approach Based on the MDL Principle", *Elsevier, Computational Intelligence*, (1994), Vol. 10 269-293.

[27] C. Haung, A. Darviche, "Inference in Belief Networks: A Procedural Guide", *Elsevier Science, International Journal of Approximate Reasoning*, (1994), Vol.15 225-263.

[28] J. Cheng and R. Greiner, "Learning Bayesian Belief Network Classifiers: Algorithms and System", *Proceedings of 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence*, (1998), Vol. 2056 141-151.

[29] J. Cheng, D. A. Bell, W. Liu, "An Algorithm for Bayesian Belief Network Construction from Data", *In Proceedings of AI & STAT'97*, (1997), 83-90.

[30] W.L. Buntine, "A Guide to the Literature on Learning Probabilistic Networks from Data", *Knowledge and Data Engineering (TKDE), IEEE Transactions on*, (1996), Vol.8 195-210.

[31] D. Margaritis, "Distribution-Free Learning of Bayesian Network Structure in Continuous Domain", *20th National Conference on Artificial Intelligence, AAAI*, (2005), Vol.2 825-830.

[32] P. Spirtes, C. Glymour, R. Scheines, "Causation, Prediction, and Search", *Springer, Lecture Notes in Statistics*, (1993).

[33] D. Grossman and P. Domingos, "Learning Bayesian Network Classifiers by Maximizing Conditional Likelihood", *ACM, 21st International Conference of Machine Learning (ICML)*, (2004).

[34] http://parshosting.com, November 2012.

[35] http://www.parswebsite.com, November 2012.

[36] M. A. L. Thathachar, P. S. Sastry, "Varieties of Learning Automata: An Overview", *IEEE, Transaction on Systems, Man, and Cybernetics*, (2002), Vol. 32 711- 722.

[37] N.A. Rezvani, AND M.R. Meybodi, "A Learning Automata-Based Technique for Training Bayesian Networks", *Proceedings of the 2nd International Conference on Advanced Computer Theory and Engineering (ICACTE)*, (2009), Vol.1.

[38] S. H. Zahiri, "Learning Automata Based Classifier", *Elsevier Science, Pattern Recognition Letters*, (2008), Vol. 29 40–48.

[39] http://www.ife.ee.ethz.ch/education/WS1_HS2012_02, December 2012.

زیرنویس‌ها

[1] User Agent
[2] Maximum likelihood estimate (MLE)
[3] Conditional Probability Distribution
[4] Pars Web Site (PW)
[5] Pars Hosting (PH)
[6] Confusion Matrix
[7] Accuracy
[8] Recall
[9] Precision