فصلنامه پژوهش در علوم، مهندسی و فناوری
دوره ۳، شماره ۱، بهار ۱۳۹۶
صفحات ۱۲-۱۹
ISSN: 2476-5082
www.uctjournals.com

شناسایی اسپم در نظرات با الگوریتم های هوش گروهی و الگوریتم های یادگیری ماشین

منصور جویا [۱] و سیران خیاطی [۲]

[۱] گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد علوم تحقیقات ، ارومیه، ایران jouya.m@gmail.com
[۲] گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد علوم تحقیقات، سنندج، ایران

چکیده

در سال‌های اخیر، نظرات جعلی و اسپم مشکلی است که به شدت در حال گسترش و افزایش است. امروزه سایت‌های تجاری، هدف مناسبی برای تولیدکنندگان اسپم هستند. اغلب سایت‌های تجاری قسمتی برای نظرات کاربران دارند و کاربران می‌توانند دیدگاه خود را درباره‌ی محصول انتشار دهند که اطلاعات ارزشمندی برای مشتری‌ها و شرکت سازنده دارد. برای این‌که نظرات و تجربیات واقعی کاربران به درستی منعکس شود، شناسایی نظر جعلی بسیار مهم است. تشخیص نظرات جعلی و اسپم در نظرات به علت تکنیک‌ها و روش‌های جدید برای انجام این کار، بسیار پیچیده است. در مقاله می خواهیم با استفاده از الگوریتم های هوش گروهی به شناسایی اسپم ها با زمان کمتری بپردازیم.

واژه‌های کلیدی: اسپم، تشخیص نظر جعلی، هوش گروهی، یادگیری ماشین.

۱- مقدمه

اولین اسپم توسط گری فورک[1] برای تبلیغ شرکتش ارسال شد. ارسال اسپم تقریباً بدون هزینه است و این مسئله باعث شده شرکت‌های بازاریابی، به سمت آن حرکت کنند. بدون اغراق میتوان گفت که استفاده از نظرات در رسانه های اجتماعی در حال افزایش است. اشخاص و سازمان ها از نظررات در شبکه های اجتماعی به صورت گستردهای برای تاثیر بر تصمیمات خریداران، تصمیم گیری در انتخابات، بازاریابی و طراحی محصول، استفاده میکنند. نظرات برخط به صورت افزایندهای توسط افراد و سازمانها استفاده میشوند. بنابراین بسیار مهم است که این نظرات خالی از غرض ورزی باشند تا مراجعه کنندگان به این سایتها بتوانند به نظرات اتکا کنند. نظرات مثبت اغلب به معنی سود بیشتر و شهرت برای تجارت و اشخاص است که متاسفانه ابزاری برای بازی دادن سیستمها شده است بدین طریق که با جعل دیدگاه و یا نظر ، محصولات مشخص ، سرویسها، اشخاص و سازمانها را ارتقا داده یا بی اعتبار میکنند. این نظرات حتی بدون فاش کردن هویت اصلی شخص یا سازمانی است که شخص برای آن کار میکند، صورت میپذیرد [۱].

۲- اسپم

به سوء استفاده از ابزارهای الکترونیکی مانند ایمیل، مسنجر، گروه های خبری ایمیلی، فکس، پیام کوتاه و غیر برای ارسال پیام به تعداد زیاد و به صورت ناخواسته اسپم می‌گویند. با توجه به هزینه اندک این روش نسبت به پست سنتی که در گذشته برای ارسال پلاک به پلاک تبلیغات مورد استفاده قرار می‌گرفت و همچنین ناقص بودن قوانین بین‌المللی برای محدود کردن هرزنامه، این قبیل ایمیل‌ها در سطح وسیعی ارسال می‌شوند. به طور کلی اسپم به معنای فرستادن پیام‌های متعدد و مشابه‌است و اسپمینگ تکرار این عمل نامیده می‌شود [۲]. اسپم‌ها به گونه‌های مختلفی در اینترنت گسترش می‌یابند که تعدادی از آنها در زیر آمده‌است.

- اسپم ایمیل
- اسپم پیامدهی آنی یا لحظه‌ای
- موتورهای جستجوی با اهداف اسپم
- اسپم وبلاگی
- اسپم انجمنی

۳- هوش گروهی

هوش گروهی نوعی روش هوش مصنوعی است که استوار بر رفتارهای گروهی در سامانه‌های نامتمرکز و خود سامانده بنیان شده است. این سامانه‌ها معمولاً از جمعیتی از کنشگران ساده تشکیل شده است که بطور محلی با یکدیگر و با پیرامون خود در همکنشی هستند. با وجود اینکه معمولاً هیچ کنترل تمرکز یافته‌ای، چگونگی رفتار کنش‌گران را به آنها تحمیل نمی‌کند، همکنشی های محلی آنها به پیدایش رفتاری عمومی می‌انجامد. الگوریتم های هوش گروهی شامل الگوریتم مورچه‌ها، بهینه‌سازی گروهی ذره، سامانه ایمنی ساختگی، الگوریتمهای بر پایه زنبور و الگوریتم چکه آب‌های هوشمند می باشد [۳].

۳-۱- الگوریتم مورچه‌ها

الگوریتم مورچه ها یکی از زیر شاخه‌های هوش گروهی است که در آن از رفتار مورچه‌های واقعی برای یافتن کوتاه‌ترین گذرگاه میان لانه و چشمه خوراکی الگوبرداری شده است. هر مورچه برای یافتن خوراک در گرداگرد لانه به گونه تصادفی

[1] Gary Fork

حرکت و در طی راه با بهره‌گیری از ماده شیمیایی به نام فرومن، از خود ردی بر جای می‌گذارد. هر چه شمار مورچه‌های گذر کرده از یک گذرگاه بیشتر باشد، میزان فرومن انباشته روی آن گذرگاه نیز افزایش می‌یابد. مورچه‌های دیگر نیز برای گزینش گذرگاه، به میزان فرومن آن می‌نگرند و به احتمال زیاد گذرگاهی را که دارای بیشترین فرومن است بر می‌گزینند [۴].

۳-۱-۱- الگوریتم کلونی مورچگان در فضای پیوسته

الگوریتم کلونی مورچگان یا ACO ابتدا برای حل مسایل بهینه سازی در فضای گسسته مطرح شد و با توجه به کارایی مناسب و قابلیت های آن نسخه های مختلفی از این الگوریتم برای حل مسایل بهینه سازی در فضای پیوسته نیز ارائه شد. از پرکاربرد ترین این الگوریتمها میتوان به الگوریتم ACOR که توسط سوشا و دریگو مطرح شده است اشاره کرد. در الگوریتم ACO در فضای گسسته، در هر مرحله یک مورچه بر اساس یک رابطه احتمالاتی مسیر خود را میسازد. در این رابطه احتمالاتی مسیر با فرومون بیشتر احتمال انتخاب شدن بیشتری دارد و ساختار احتمالاتی بر پایه توزیع احتمال گسسته بنا شده است. ایده اصلی الگوریتم ACOR تبدیل این توزیع احتمال گسسته به یک تابع چگالی احتمال پیوسته میباشد. در الگوریتم ACOR ابتدا یک جمعیت اولیه از راه حل ها به صورت تصادفی انتخاب میشوند. سپس یک آرشیو از راه حل ها ایجاد شده و تعداد k راه حل از راه حل های بهتر موجود در جمعیت اولیه بر حسب توابع برازندگی شان مطابق شکل ۱ در آرشیو جای میگیرند. برای تعیین راه حل های جدید ابتدا یکی از راه حل های موجود در الگوریتم با یک رابطه احتمالاتی به عنوان مورچه راهنما انتخاب میشود [۵]. جهت این انتخاب احتمالاتی به هر سطر j بر حسب تابع برازندگی آن ، یک ضریب وزنی ω نسبت میدهیم. (رابطه ۱)

رابطه ۱

$$\omega_j = \frac{1}{qk\sqrt{2\pi}} e^{-\frac{(j-1)^2}{2q^2k^2}}$$

احتمال انتخاب شدن سطر j نیز از رابطه ۲ بدست میآید :

رابطه ۲

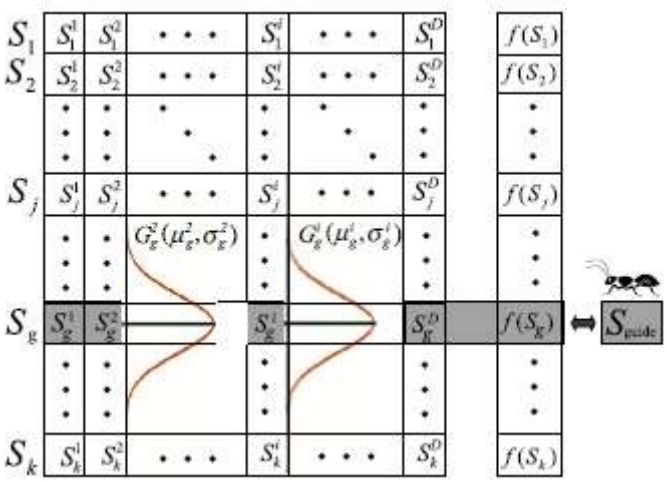
$$P_j = \frac{\omega_j}{\sum_{r=1}^{k} \omega_r}$$

سپس با استفاده از یک چرخ رولت میتوان یکی از مورچه ها را انتخاب کرد. چنانچه سطر g به عنوان مورچه راهنما انتخاب شود، سعی میکنیم جوابی در در همسایگی نزدیک این جواب تولید کنیم. در عین حال برای ایجاد تنوع نباید احتمال انتخاب نقاط دورتر را هم صفر کنیم. بدین منظور از یک تابع توزیع احتمال گاوسی حول متغیرهای راه حل g استفاده کرده و با نمونه برداری از این تابع راه حل های جدید تولید میکنیم. میانگین تابع گاوسی را برابر مقدار متغییر i ام سطر g S_g^i در نظر گرفته و انحراف معیار با بکار بردن ربطه ۳ به صورت تطبیقی محاسبه می شود.

رابطه ۳

$$\sigma_g^i = \varepsilon \sum_{e=1}^{k} \frac{|S_e^i - S_g^i|}{k-1}$$

ε یک پارامتر مثبت بوده و برای کنترل پراکندگی استفاده میشود. سپس از بین راه حل های تولید شده و راه حل های موجود در آرشیو، تعداد k راه حل بهتر را انتخاب و اطلاعات آرشیو را به روزرسانی میکنیم. پس از تعداد مناسبی تکرار انتظار داریم سطرهای آرشیو به مقدار بهینه همگرا شود.



شکل ۱: آرشیو راه حل ها در الگوریتم ACOR

۳-۲- بهینه‌سازی گروهی ذره

یک الگوریتم جستجوی اجتماعی است که از روی رفتار اجتماعی دسته‌های پرندگان مدل شده است. در ابتدا این الگوریتم به منظور کشف الگوهای حاکم بر پرواز همزمان پرندگان و تغییر ناگهانی مسیر آنها و تغییر شکل بهینه دسته به کار گرفته شد [۶].

۳-۳- سامانه ایمنی ساختگی[۲]

از سامانه ایمنی پرهامی (طبیعی) برگرفته شده و با مدل کردن برخی یاخته‌های این سامانه می‌کوشد تا از اینگونه سامانه‌ها در راستای بهینه‌سازی و یا یادگیری دستگاه [۳]بهره بگیرد [۷].

۳-۴- الگوریتمهای بر پایه زنبور

چندین الگوریتم در زمینه الگوبرداری از رفتار گروهی زنبورها در کندو و یا بیرون آن به ویژه رفتار آنها در خوراکجویی [۴]برای بهینه‌سازی پیوسته[۵] پیشنهاد شده‌اند [۸].

۳-۵- الگوریتم چکه آب‌های هوشمند

الگوریتم چکه آبهای هوشمند[۶] بر پایه بررسی رفتار گروهی چکه‌های آب در رودخانه‌ها برای بهینه‌سازی پیشنهاد شده است. این الگوریتم با گسیل کردن چکه‌های هوشمند آب بر روی گراف مسئله (پرسمان)، می‌کوشد راه‌های بهینه تر را برای گذر چکه‌های یاد شده هموارتر گرداند [۹].

2 Artificial Immune System
3 Machine Learning
4 foraging
5 continuous optimization

۴- الگوریتم های یادگیری ماشین

به عنوان یکی از شاخه‌های وسیع و پرکاربرد هوش مصنوعی، یادگیری ماشین به تنظیم و اکتشاف شیوه‌ها و الگوریتم‌هایی می‌پردازد که بر اساس آنها رایانه‌ها و سامانه‌ها توانایی تعلم و یادگیری پیدا می‌کنند. یادگیری ماشین عبارت است از اینکه چگونه می توان برنامه ای نوشت که از طریق تجربه یادگیری کرده و عملکرد خود را بهتر کند. یادگیری ممکن است باعث تغییر در ساختار برنامه و داده ها شود. یادگیری ماشین زمینه تحقیقاتی نسبتا جدیدی از هوش مصنوعی است که در حال حاضر دوران رشد و تکامل خود را می گذراند و زمینه ای بسیار فعال در علوم کامپیوتر می باشد [۱۰].

۴-۱- الگوریتم ماشین یادگیری افراطی[۷]

یادگیری ماشین افراطی بر اساس نظریه به حداقل رساندن خطر تجربی بنا نهاده شده است و از شبکه تک لایه پیشخور برای آموزش شبکه های عصبی تک لایه پنهان پیشخور[۸]، استفاده می کند [۱۱]. فرایند یادگیری تنها به یک تکرارکننده نیاز دارد و از تکرار متعدد و به حداقل رساندن محلی جلوگیری می کند. در مقایسه با الگوریتم های شبکه عصبی معمولی، الگوریتم ماشین یادگیری افراطی قادر به دستیابی به سرعت های آموزش سریع تر هستند و می توانند بر مشکل بیش‌برازش[۹] غلبه کنند. فرض کنید مجموعه نمونه های مختلف $D = (x_i.o_i).i = 1 \dots P$ را در اختیار داشته باشیم، که در آن $\{x_i\}and\{o_i\}$ باشد بنابراین هدف از این رابطه پیدا کردن $\{\mathrm{xi}\}\epsilon^{\mathrm{m}}.\mathrm{and}\{\mathrm{oi}\}\epsilon^{\mathrm{n}}$ است. شبکه های پیشخور تک لایه پنهان استاندارد با گره های $\mathbf{N}$ می توانند از لحاظ ریاضی توسط رابطه ۴ مدل سازی شوند:

رابطه ۴

$$y_j = \sum_{k=1}^{N} h_k f(w_k . x_j)$$

که در آن $1 \leq \mathrm{j} \leq \mathrm{P}$ است، w_k نمایانگر پارامترهای عناصر K در لایه های پنهان است، h_k اشاره به وزنی دارد که عنصر پنهان k را با لایه خروجی متصل می کند و f نشان دهنده تابعی است که خروجی لایه پنهان را ایجاد می کند. رابطه ۵ می تواند در نمایش ماتریسی بعنوان $y = G.h$ بیان شود که در آن h بردار وزنی لایه خروجی است و $\mathbf{G}$ برابر است با:

رابطه ۵

$$G = \begin{matrix} f(w_1 . x_1) \dots\dots f(w_n . x_1) \\ \mathrm{f}(w_1 . \mathrm{x}_p) \dots\dots f(w_N . \mathrm{x}_p) \end{matrix}$$

که در آن N تعداد گره های پنهان است. الگوریتم ماشین یادگیری افراطی مقدار دهی اولیه تصادفی پارامترها را در لایه پنهان w^k پیشنهاد می کند. که وزن های لایه خروجی به دست آمده توسط معکوس تعمیم یافته با توجه به $h = G^+ . 0$ بیان می شوند که در آن $G^+ = (G^T . G)^{-1}$ است. G^T ماتریس شبه معکوس است.

۵- روش الگوریتم پیشنهای

برای شناسایی اسپم با استفاده از الگوریتم کلونی مورچگان در فضای پیوسته و ادغام آن با الگوریتم ماشین یادگیری افراطی نتایج زمای بهتری نسبت به الگوریتم ماشین یادگیری افراطی به تنهای بدست آمده است. روش الگوریتم به این

---

6 Intelligent Water Drops
7 Extreme Learning Machine
8 single-hidden layer feedforward
9 Overfitting

صورت می باشد که در ابتدا داده ها موجود در دیتاست را استخراج کرده و الگوریتم کلونی مورچگان در فضای پیوسته را روی داده های خوانده شده اعمال می کنیم نتایج بدست آماده به دو دسته کاربران عادی و کاربرات غیر عادی تقسیم میشود. این عمل تقسیم کاربران باعث می شود زمان کمتری جهت بررسی اسپم بر روی کاربران غیر عادی بجای بررسی کل کاربران سپری شود. در ادامه با توجه به اطلاعات بدست آمده از کاربران غیر عادی، داده ها بدست آمده را با الگوریتم ماشین یادگیری افراطی مورد بررسی جهت تشخصی کاربران اسپم قرار میدهیم.

۶- نتایج شبیه‌سازی

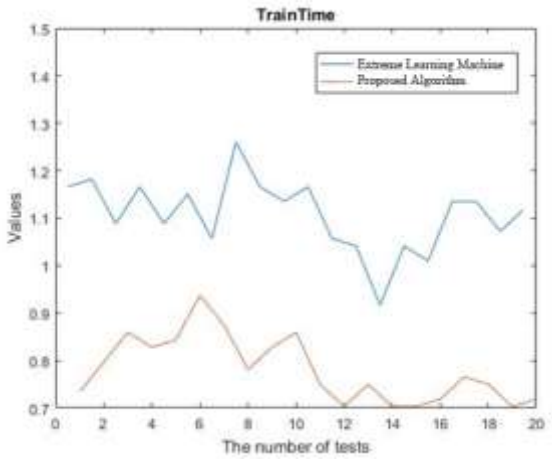
در الگوریتم پیشنهادی شناسایی اسپم در نظرات کاربران با یکی از الگوریتم های هوش گروهی بنام الگوریتم کلونی مورچگان در فضای پیوسته و ادغام آن با یکی از الگوریتم های یادگیری ماشین بنام الگوریتم های یادگیری ماشین افراطی استفاده شده و با الگوریتم ماشین یادگیری افراطی به تنهای مورد بررسی و مقایسه قرار داده ایم که از نظر زمان برای آموزش و زمان تست که دو پارامتر در تشخیص کاربران اسپم هستند، نتیجه مطلوب تر و بهینه تری بدست آمده است.

۶-۱- نتایج زمان آموزش

همانطورکه در شکل ۲ نمایش داده شده است نتایج زمان آموزش روش پیشنهادی با مقایسه با الگوریتم ماشین یادگیری افراطی به تنهای بهینه‌تر و کاربردی‌تر می‌باشد. مشخصات بررسی در جدول ۱ نمایش داده شده است.

جدول ۱: پارامترهای بررسی ویژگی زمان آموزش

تعداد آزمایش	Dataset
۲۰	۵۰۰۰ نفر



شکل ۲: بررسی ویژگی زمان آموزش

میانگین بدست آمده از بررسی با مشخصات جدول ۱ را در جدول ۲ می‌توان مشاهده کرد.

جدول ۲: بررسی میانگین ویژگی زمان آموزش

الگوریتم پیشنهادی	الگوریتم ماشین یادگیری افراطی به تنهای

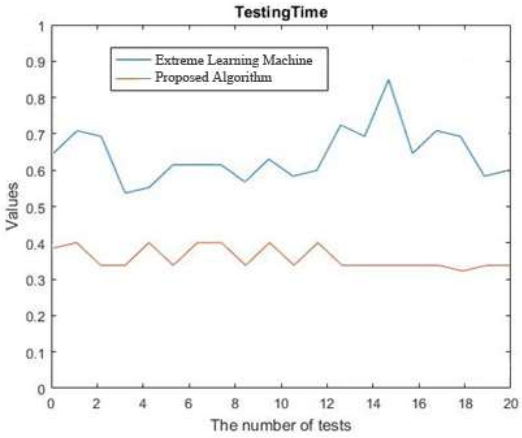
۰,۹۹۵۰	۱,۴۵۲۰

۶-۲- بررسی زمان تست

همانطورکه در شکل ۳ نمایش داده شده است نتایج بررسی زمان تست روش پیشنهادی با مقایسه با الگوریتم ماشین یادگیری افراطی به تنهای بهینه‌تر و کاربردی‌تر می‌باشد. مشخصات بررسی در جدول ۳ نمایش داده شده است.

جدول ۳: پارامترهای بررسی ویژگی زمان تست

تعداد آزمایش	Dataset
۲۰	۵۰۰۰ نفر



شکل ۳: بررسی ویژگی زمان تست

میانگین بدست آمده از بررسی با مشخصات جدول ۳ را در جدول ۴ می‌توان مشاهده کرد.

جدول ۴: بررسی میانگین ویژگی زمان تست

الگوریتم پیشنهادی	الگوریتم ماشین یادگیری افراطی به تنهای
۰,۶۲۵۰	۰,۸۴۵۲

۷- جمع‌بندی و نتیجه گیری

در این مقاله به انجام شبیه سازی و بیان گزارش‌ها، نتایج و تحلیل‌های مربوط به آن‌ها، برای ارزیابی روش پیشنهادی مطرح شده و مقایسه آن پرداختیم. نتایج حاصل بیانگر این نکته بود که، روش پیشنهادی ترکیبی ما، موجب حصول نتایج بهتری در زمان آموزش و زمان تست در مقایسه با سایر روش‌های موجود بوده است. یعنی در مقایسه روش پیشنهادی ما که ترکیب الگوریتم‌های الگوریتم کلونی مورچگان در فضای پیوسته و الگوریتم ماشین یادگیری افراطی با روش‌های مطرح شده دیگر در این زمینه، عملکرد بهتر روش پیشنهادی ما نسبت به روش های دیگر نشان داده شده است. نتایج شبیه‌سازی‌های انجام شده موفقیت روش پیشنهادی را نشان می‌دهد. این نتایج بیانگر عملکرد خوب این روش می‌باشد که تشخیص در زمان

کمتری را به همراه خواهد داشت. استفاده از حالت ترکیبی موجب بهبود قابل توجه عملکرد روش پیشنهادی ما در این مقاله شده است.

منابع

1. Sedhai, Surendra, and Aixin Sun. "Semi-Supervised Spam Detection in Twitter Stream." *arXiv preprint arXiv:1702.01032* (2017).
2. Ren, Yafeng, and Donghong Ji. "Neural networks for deceptive opinion spam detection: An empirical study." *Information Sciences* 385 (2017): 213-224.
3. Dorigo, Marco. "Editorial: Ten years of swarm intelligence." *Swarm Intelligence* 10.4 (2016): 245-246.
4. Rajappa, Gautham Puttur, Joseph H. Wilck, and John E. Bell. "An Ant Colony Optimization and Hybrid Metaheuristics Algorithm to Solve the Split Delivery Vehicle Routing Problem." *International Journal of Applied Industrial Engineering (IJAIE)* 3.1 (2016): 55-73.
5. Desell, Travis, et al. "Evolving Deep Recurrent Neural Networks Using Ant Colony Optimization." *European Conference on Evolutionary Computation in Combinatorial Optimization*. Springer International Publishing, 2015.
6. Esmin, Ahmed AA, Rodrigo A. Coelho, and Stan Matwin. "A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data." *Artificial Intelligence Review* 44.1 (2015): 23-45.
7. Alonso, F. R., D. Q. Oliveira, and AC Zambroni de Souza. "Artificial immune systems optimization approach for multiobjective distribution system reconfiguration." *IEEE Transactions on Power Systems* 30.2 (2015): 840-847.
8. Aydın, Doğan. "Composite artificial bee colony algorithms: From component-based analysis to high-performing algorithms." *Applied Soft Computing* 32 (2015): 266-285.
9. Mokhtari, Hadi. "A nature inspired intelligent water drops evolutionary algorithm for parallel processor scheduling with rejection." *Applied Soft Computing* 26 (2015): 166-179.
10. Jordan, Michael I., and Tom M. Mitchell. "Machine learning: Trends, perspectives, and prospects." *Science* 349.6245 (2015): 255-260.
11. Tang, Jiexiong, Chenwei Deng, and Guang-Bin Huang. "Extreme learning machine for multilayer perceptron." *IEEE transactions on neural networks and learning systems* 27.4 (2016): 809-821.