# استنتاج فازی در تشخیص نفوذ روبات‌های وب به شبکه‌های کامپیوتری

# Fuzzy Inference for Intrusion Detection of Web Robots in Computer Networks

مهدیه ذبیحی*، دانشجوی کارشناسی ارشد مهندسی کامپیوتر، دانشگاه بین‌المللی امام رضا، مشهد،
m.zabihi@imamreza.ac.ir

مجید وفایی جهان، عضو هیئت علمی گروه مهندسی کامپیوتر، دانشگاه آزاد، مشهد، vafaeiJahan@mshdiau.ac.ir

جواد حمیدزاده، عضو هیئت علمی گروه مهندسی کامپیوتر، موسسه آموزش عالی سجاد، مشهد،
j_hamidzadeh@sadjad.ac.ir

**چکیده:** تمایز انسان و روبات از حیث تامین امنیت شبکه‌های کامپیوتری، باعث طرح مساله تشخیص روبات وب شده‌است؛ که حل دقیق آن، سایت‌ها را از دید روبات‌های مخرب مصون داشته و کارایی سرورها را با کاهش اولویت در پاسخ‌دهی به روبات‌ها افزایش می‌دهد. هدف این مقاله، ارائه روشی جدید مبتنی بر سیستم استنتاج فازی در تشخیص روبات‌های وب است. در روش پیشنهادی، علاوه بر اعمال آنالیز همبستگی جهت کاهش تعداد ویژگی‌های توصیفی بازدیدکنندگان وب، از درخت تصمیم استفاده شده‌است که ضمن فازی‌سازی این ویژگی‌ها، موجب فائق آمدن بر مشکل نفرین ابعاد شده و تعداد قوانین لازم برای سیستم استنتاج فازی را کاهش می‌دهد و بدین‌ترتیب موجب سهولت طراحی آن می‌گردد. نتایج آزمایش‌ها نشان می‌دهد؛ الگوریتم پیشنهادی با نرخ هشدار غلط ۰,۱۳ انسان را از روبات وب متمایز ساخته و با دقت ۹۷٪ قادر به تشخیص روبات‌ها است. نتایج مقایسه‌ها نشان‌دهنده کاهش نرخ هشدار غلط و افزایش دقت روش پیشنهادی نسبت به روش‌های مرز دانش می‌باشد.

**کلمات کلیدی:** روبات‌های وب، سیستم استنتاج فازی، درخت تصمیم، فایل ثبت وقایع، آنالیز همبستگی.

## مقدمه

در این مقاله با دیدگاهی جدید نسبت به بازدیدکنندگان وب و فازی‌سازی ویژگی‌های توصیفی آن‌ها، راهی دقیق برای تشخیص نفوذ در شبکه‌های کامپیوتری یافته و با بهره‌گیری از درخت تصمیم[1]، تعداد قوانین لازم برای سیستم استنتاج فازی[2] را کاهش و طراحی این سیستم را آسان می‌سازیم. با توجه به اهمیت ویژگی‌های مرتبط، از آنالیز همبستگی[3] بین برچسب داده‌ها و هر یک از ویژگی‌های موردنظر استفاده کرده و برای تبدیل هر ویژگی به یک متغیر فازی، از درخت تصمیم استفاده می‌کنیم. اگرچه در [۱] نیز از سیستم استنتاج

[1] Decision Trees
[2] Fuzzy Inference Systems
[3] Correlation Analysis

فازی برای تشخیص روبات‌های وب استفاده شده-است؛ در این‌جا تعداد نمونه‌های مجموعه آموزشی و آزمایشی به مراتب بیشتر بوده و استفاده از راه‌کارهایی برای ساده‌سازی طراحی سیستم استنتاج فازی اهمیت بسزایی می‌یابد.

ادامه این مقاله به‌صورت زیر سازمان‌دهی شده-است: بخش ۲ به معرفی کارهای پیشین پرداخته و بخش ۳ نیز الگوریتم پیشنهادی را معرفی می‌کند. آزمایش‌های انجام شده در بخش ۴ و مقایسه‌ها و نتیجه‌گیری نهایی نیز در بخش ۵ ارائه می‌شود.

## ۲- کارهای مرتبط

براساس دسته‌بندی Doran و Gokhale [۲]، یکی از اولین کارهای انجام شده در حوزه تشخیص روبات-های وب، مقاله Tan و Kumar در سال ۲۰۰۲ است که با ارائه روشی جدید جهت استخراج نشست‌ها از فایل ثبت وقایع، تعریف ۲۵ ویژگی و استفاده از درخت تصمیم C4.5، به حل مساله تشخیص روبات وب می‌پردازد [۳]. بعد از آن Bomhardt و همکارانش در سال ۲۰۰۵ با استفاده از شبکه عصبی، درخت تصمیم و تعریف یک ابزار پیش‌پردازش فایل ثبت وقایع[4]، نمونه داده‌های دو سایت را طبقه‌بندی می‌کنند [۴]. در سال ۲۰۰۹ نیز Stassopoulou و Dikaiakos با استفاده از ۶ ویژگی و به‌کارگیری شبکه بیزین به طبقه‌بندی نمونه‌ها می‌پردازند [۵].

همانطور که در بخش قبل نیز اشاره شد؛ کار ارائه شده در [۱] نیز مبتنی بر سیستم استنتاج فازی است که البته، تنها بر روی تعداد محدودی نمونه آزمایش شده و حالت جامع‌تر آن در این مقاله ارائه می‌گردد.

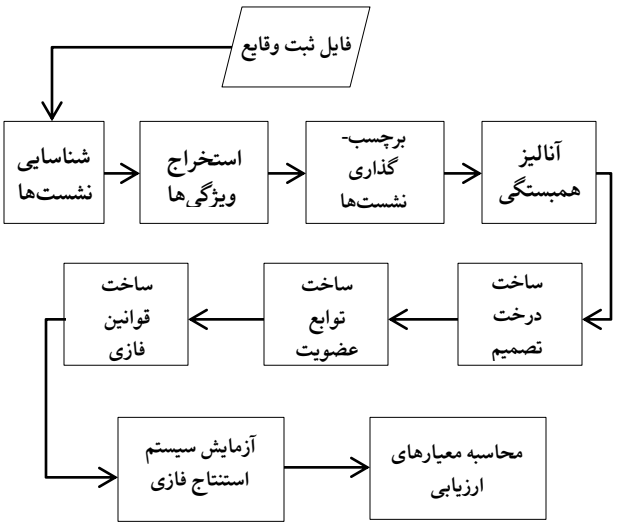
## ۳- الگوریتم پیشنهادی

در این قسمت، به ارائه الگوریتم پیشنهادی این مقاله که در شکل (۱) نمایش داده شده‌است؛ می‌پردازیم.

در فاز استخراج نشست‌ها، به روش‌های مرسوم، نشست‌های موجود در فایل ثبت وقایع یک وب سرور استخراج شده و در فاز بعد، برای هر نشست، ۱۴ ویژگی توصیفی بازدیدکنندگان وب تعیین می‌گردد.

بعد از برچسب‌گذاری نشست‌ها به دو نوع انسان و روبات در مرحله ۳، نوبت به اعمال آنالیز همبستگی بین این برچسب‌ها و هر یک از ۱۴ ویژگی استفاده‌شده می‌رسد تا تعداد این ویژگی‌ها کاهش یابد. زیرا، اگرچه داشتن ویژگی‌های بیشتر معادل داشتن اطلاعات بیشتر در مورد داده‌هاست؛ با افزایش تعداد ویژگی‌ها، تعداد قوانین سیستم استنتاج فازی زیاد شده و همین امر موجب سخت و وقت‌گیر شدن طراحی سیستم استنتاج فازی می‌گردد. جدول (۱) مقادیر آنالیز همبستگی بین برچسب نمونه‌ها و هریک از ویژگی‌ها را نشان می‌دهد.

می‌دانیم مقدار صفر یا نزدیک به صفر برای آنالیز همبستگی، گواه عدم کفایت ویژگی موردنظر در ایجاد تمایز بین انسان و روبات است [۳]. پس در روش پیشنهادی برای مقادیر آنالیز همبستگی مثبت، بازه $(1,\infty]$ و برای مقادیر همبستگی منفی بازه $[-1,-\infty)$ در نظر گرفته می‌شود. مطابق توضیحات، ده ویژگی از مجموع چهارده ویژگی به‌عنوان ویژگی نهایی انتخاب می‌گردد.




شکل ۱- فلوچارت الگوریتم پیشنهادی

[4] RDT

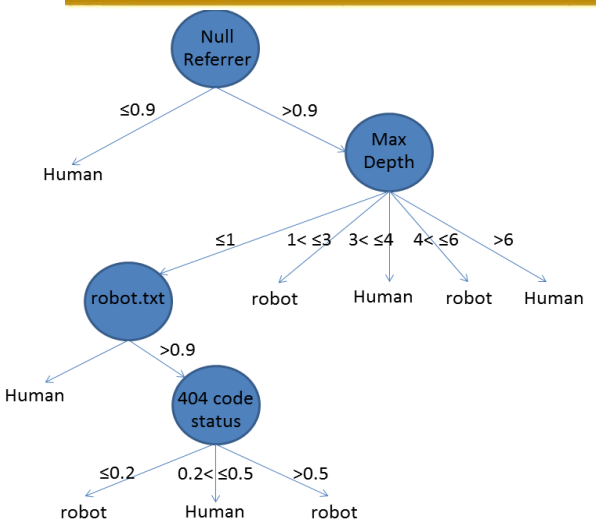
جدول ۱- مقادیر آنالیز همبستگی برای هریک از ویژگی‌ها

| نام ویژگی | نتیجه آنالیز |
|---|---|
| حجم اطلاعات درخواستی | -۰,۰۱۱ |
| نرخ درخواست دنباله متوالی | -۰,۱۲ |
| انحراف معیار عمق درخواست | -۰,۲۸ |
| درصد پاسخ با کد وضعیت ۴۰۴ | ۰,۱۴ |
| تعداد تقاضاها | -۰,۰۲ |
| درخواست فایل‌های pdf/pss | ۰,۰۷ |
| درصد درخواست در شب | ۰,۳۳ |
| درخواست فایل "robots.txt" | ۰,۴۶ |
| طول نشست | ۰,۱۱ |
| درصد درخواست فایل css | -۰,۲۴ |
| نرخ تغییرات نوع فایل درخواستی | -۰,۱۲ |
| درصد درخواست با ارجاعات خالی | ۰,۴۶ |
| حداکثر عمق درخواست | -۰,۲۹ |
| نسبت درخواست html به تصویر | ۰,۰۴ |

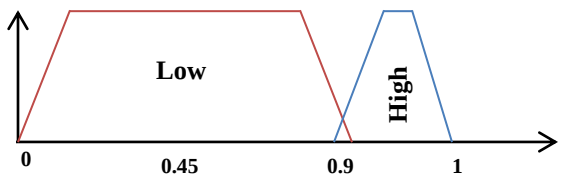
نهایتا در فاز ۵، نوبت به ساخت درخت تصمیم بر روی مجموعه داده‌های آموزشی می‌رسد.

هدف از این کار علاوه بر فازی‌سازی[5] ویژگی‌ها، حل مشکل نفرین ابعاد[6] نیز می‌باشد؛ چرا که در ساخت درخت تصمیم، در هر مرحله یک ویژگی با بیشترین بهره اطلاعات، که می‌تواند متمایزکننده بهتری به حساب آید؛ انتخاب می‌شود [۶][۷]. بنابراین با رسم این درخت تا حدی می‌توان از تعداد ویژگی‌های نهایی کاست. شکل (۲)، درخت تصمیم بدست آمده بر روی داده‌های آزمایشی این مقاله را نشان می‌دهد. همان‌طور که از این شکل پیداست؛ از مجموع ده ویژگی انتخابی در فاز ۴، تنها چهار ویژگی در رسم درخت و نهایتا استخراج قوانین فازی استفاده می‌شوند.

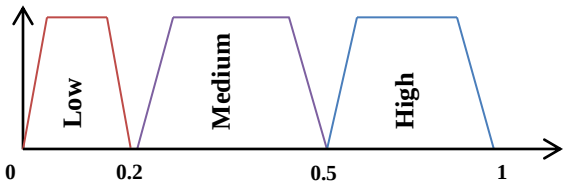
در فاز ۶، کلیه توابع عضویت چهار متغیر استفاده-شده در ساخت درخت تصمیم، براساس وزن‌های هر شاخه از درخت مشخص می‌گردند (شکل ۳).



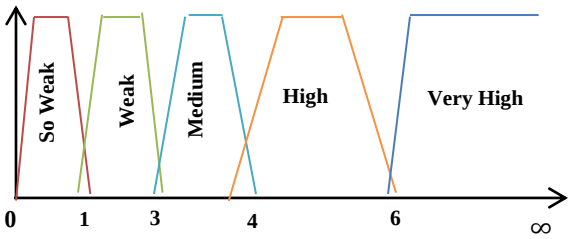

شکل ۲- درخت تصمیم C4.5 براساس ده ویژگی انتخابی



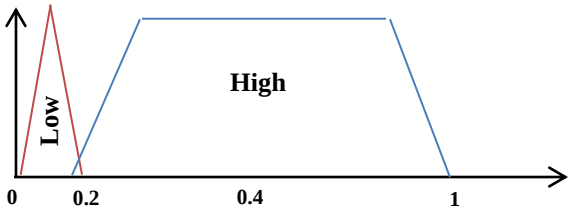

۳-الف- تابع عضویت ویژگی "درصد درخواست‌های بدون ارجاع"




۳-ب- تابع عضویت ویژگی "درصد درخواست‌ها با کد ٤٠٤"




۳-ج - تابع عضویت "حداکثر عمق درخواست‌ها"




۳-د - تابع عضویت ویژگی "درخواست فایل robots.txt"

[5] Fuzzification
[6] Curse Dimensionality

نهایتا در فاز ۷، قوانین سیستم استنتاج فازی بر طبق توابع عضویت استخراج شده، با پیمایش از ریشه درخت به سمت برگ‌های آن، تعیین می‌گردند:

جدول ۲- قوانین تعیین شده برای سیستم استنتاج فازی

| |
|---|
| If (**null Referrer is High**) & (**Max Depth is So Weak**) & (**robot.txt is High**) & (**E404 Status is Low**) then (**output is Robot**) |
| If (**null Referrer is High**) & (**Max Depth is So Weak**) & (**robot.txt is High**) & (**E404 Status is Medium**) then (**output is Human**) |
| If (**null Referrer is High**) & (**Max Depth is So Weak**) & (**robot.txt is High**) & (**E404 Status is High**) then (**output is Robot**) |
| If (**null Referrer is High**) & (**Max Depth is So Weak**) & (**robot.txt is Low**) then (**output is Human**) |
| If (**null Referrer is High**) & (**Max Depth is Weak**) then (**output is Robot**) |
| If (**null Referrer is High**) & (**Max Depth is Medium**) then (**output is Human**) |
| If (**null Referrer is High**) & (**Max Depth is High**) then (**output is Robot**) |
| If (**null Referrer is High**) & (**Max Depth is Very High**) then (**output is Human**) |
| If (**null Referrer is Low**) then (**output is Human**) |

بعد از اتمام فرآیند آموزش مدل استنتاج فازی، نوبت به آزمودن این مدل و ارزیابی عملکرد آن می‌رسد.

## ۴- آزمایش‌ها

در این مقاله از فایل ثبت وقایع وب سرور دانشگاه امام رضا[7]، با ۱۱۷۰ نمونه روبات و ۱۶۷۹۹ نمونه انسان استفاده شده‌است.

دو معیار حائز اهمیت در مساله تشخیص نفوذ به شبکه‌های کامپیوتری مطابق روابط زیر هستند:

$$TPR = \frac{TP}{FN + TP} = 0.97$$ رابطه (۱)

$$FPR = \frac{FP}{TN + FP} = 0.13$$ رابطه (۲)

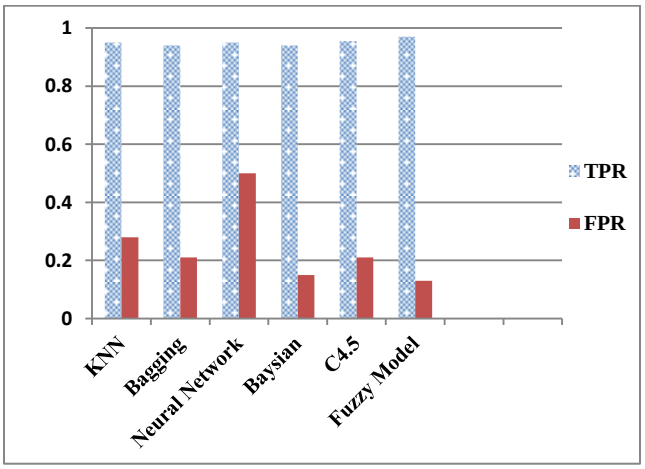
[7] www.imamreza.ac.ir

معیار TPR[8] دقت الگوریتم پیشنهادی در تشخیص روبات‌ها و معیار FPR[9] نیز نرخ هشدار غلط یا تعداد انسان‌هایی که به اشتباه روبات شناسایی شده‌اند؛ را نشان می‌دهد [۶][۷].

## ۵- مقایسه‌ها و نتیجه‌گیری

در این بخش، دو معیار نام‌برده الگوریتم پیشنهادی، در مقایسه با تعدادی از طبقه‌بندهای مرز دانش مقایسه می‌گردد. مطابق شکل (۴)، نرخ هشدار غلط الگوریتم پیشنهادی در مقایسه با سایرین پایین‌تر بوده و با توجه به معیار TPR، دقت آن نیز بیشتر است.

پس می‌توان گفت؛ داشتن دیدی فازی نسبت به ویژگی‌های توصیفی بازدیدکنندگان وب، می‌تواند از سایر روش‌های رقیب دقیق‌تر و کاراتر باشد. البته با توجه به زیاد بودن تعداد این ویژگی‌ها، استفاده از راهکارهایی مثل درخت تصمیم C4.5 ، علاوه بر فازی‌سازی ویژگی‌های توصیفی، مشکل نفرین ابعاد را حل کرده و در کاهش تعداد قوانین سیستم استنتاج فازی و سهولت طراحی آن موثر است.




شکل ٤- مقایسه معیارهای TPR و FPR الگوریتم پیشنهادی و تعدادی طبقه‌بند رقیب

[8] True Positive Rate
[9] False Positive Rate

## مراجع

[1] J. Rajabnia, M. Zabihi, and M. Vafaei Jahan, "Web Robot Detection With Fuzzy Inference System Based on decision trees", in *The Seventh Iran Data Mining Conference*, Tehran, 2013.

[2] D. Doran and S. S. Gokhale, "Web robot detection techniques: overview and limitations", *Data Mining and Knowledge Discovery*, vol. 22, pp. 183-210, 2010.

[3] P.-N. Tan and V. Kumar, "Discovery of web robot sessions based on their navigational patterns", *Data Mining and Knowledge Discovery*, vol. 6, pp. 9-35, 2002.

[4] C. Bomhardt, W. Gaul, and L. Schmidt-Thieme, "Web Robot detection pre-processing web logfiles for Robot Detection", *New Developments in Classifiation and Data Analysis*, pp. 113-124, 2005.

[5] W. Siler and J. J. Buckley, *Fuzzy Expert Systems and Fuzzy Reasoning*. Hoboken, New Jersey: John Wiley & Sons, 2005.

[6] P. Hayati, V. Potdar, K. Chai, and A. Talevski, "Web Spambot Detection Based on Web Navigation Behaviour", in *24th IEEE International Conference on Advanced Information Networking and Applications (AINA)*, Perth, Western Australia, 2010, pp. 797-803.

[7] J. Han and M. Kamber , *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2011.

[8] D. Doran and S. S. Gokhale, "A classification framework for web robots", *Journal of the American Society for Information Science and Technology*, vol. 63, no. 12, pp. 2549-2554, 2012.

[9] G. K. Kanji, *100 Statistical Tests*, 3rd ed. SAGE Publication, 2006.

[10] D. Petrilis and C. Halatsis, "Two-level Clustering of Web Sites Using Self-Organizing Maps", *Neural Processing Letters*, vol. 27, no. 1, pp. 85-95, 2008.

[11] S. M. Ross, *Introduction to Probability and Statistics for Engineers and Scientists*. California: Academic Press, 2009.

[12] A. Stassopoulou and M. D. Dikaiakos, "web robot detection:a probabilistic reasoning approach", *Computer Network*, vol. 53, pp. 265-278, 2009.

[13] D. Stevanovic, A. An, and N. Vlajic, "Feature evaluation for web crawler detection with data mining techniques", *Expert System with Application*, vol. 39, pp. 8707-8717, 2012.

[14] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis*. Whiley Press, 2009.

[15] D. J. Hand, H. Mannila, and P. Smyth, *Principles of data mining*. MIT Press, 2001.

[16] S. Known, M. Oh, D. Kim, J. Lee, Y. Kim, and S. Cha, "Web Robot Detection Based on Monotonous Behaviour", *Proceedings of the Information Science and Industrial Applications, 4*, 2012.

[17] Losawar, M. Joshi, "Data Pre-processing in Web Usage Mining", *International Conference on Artificial Intelligence and Embedded Systems (ICAIES)*, Singapore, 2012.

[18] X. Lin, L. Quan, H. Wu, "An Automatic Scheme to Categorize User Sessions in Modern HTTP Traffic", in *Global Telecommunications Conference*, pp.1485-1490, 2008.

[20] V. Sumalatha, K. Ramani, K. Lakshmi, "Fuzzy Inference System to Control PC Power Failures", *International Journal of Computer Applications*, vol. 28, no. 4, pp. 10-17, 2011.

[21] (2014) user-agent-string.info. [Online]. http://user-agent-string.info.

[22] (2014) Bots vs. Browsers. [Online]. http://www.botsvsbrowsers.com.