شناسایی ویروس ها با منطق فازی بر مبنای N-gram فایل

فاطمه شمسی [۱]، محدثه میرعمادی[۲]، فاطمه السادات هل اتایی [۳,*]

۱- دانشجوی کارشناسی،گروه علوم کامپیوتر، دانشگاه سمنان، سمنان، fatemeh.shamsi@yahoo.com

۲- دانشجوی کارشناسی،گروه علوم کامپیوتر، دانشگاه سمنان، سمنانmiremadi.mohadese@yahoo.com

۳- هیئت علمی گروه علوم کامپیوتر،گروه علوم کامپیوتر،دانشگاه سمنان، سمنان halataei@semnan.ac.ir

چکیده

یکی از چالش‌های دنیای امروز کامپیوتر پیدا کردن معیاری ساده و قوی برای شناسایی فایل های ویروسی است. در این مقاله، منطق فازی برای تشخیص ویروس پیشنهاد می‌شود. قدرت و سادگی منطق فازی در گروه بندی باعث شده است که منطق فازی نقطه عطف این مقاله باشد. علاوه بر منطق فازی، از کد باینری و N-gram نیز کمک می‌گیریم؛ به این صورت که پایگاه داده-ای از N-gram فایل‌های مختلف ایجاد کرده و سپس با استفاده از منطق فازی مشخص می‌کنیم که فایل ویروسی است و یا بی‌خطراست.

کلمات کلیدی: طبقه بندی ویروس ها ،شناسایی ویروس ، ویروس ، کد باینری ، منطق فازی ، N-gram،

۱. مقدمه

ویروس‌های کامپیوتری تهدید بزرگی برای دنیای کامپیوتری هستند. محققانی که در این زمینه کار می‌کنند برای طبقه بندی و روش‌های تشخیص این ویروس‌ها تلاش کرده و کارهای مختلفی انجام داده‌اند. نرم افزارهای آنتی ویروس تجاری فعلی یک ویروس را تنها پس از ظاهر شدن ویروس و ایجاد آسیب تشخیص می‌دهند. به همین دلیل شناسایی ویروس‌های کامپیوتری پلی‌مورف و دگرگون شونده که از روش‌های پیچیده تری برای تکامل خود استفاده می‌کند کار دشواری است. بنابراین لازم است از مدل‌های قوی برای درک تکامل و سپس تشخیص اعمال برای حذف آن‌ها استفاده کنیم. [4]

* Corresponding author
Email: halataei@semnan.ac.ir

مقاله نرم افزار محاسباتی تشخیص ویروس [1] برای شناسایی ویروس‌ها از ژنتیک الگوریتم و منطق فازی کمک گرفته شده است. ژنتیک الگوریتم برای بهینه سازی و مجموعه فازی برای طبقه بندی و ایجاد گروه‌های فازی استفاده شده است. اما پارامترهای شناسایی ویروس در این مقاله مبهم هستند. از جمله این پارامترها می‌توان به کاهش سرعت سیستم، خطای متعدد، رفتار غیر طبیعی سیستم و... اشاره کرد. واضح است که صرفا استفاده از چنین پارامترهایی برای شناسایی و طبقه بندی ویروس‌ها ، خصوصا ویروس‌های دگرگون شونده معیار مناسبی نیست .

در مقاله ترکیبی از شبکه‌های عصبی و سیستم ایمنی مصنوعی [2] از شبکه عصبی برای دسته بندی و از سیستم ایمنی مصنوعی برای آموزش به سیستم کمک گرفته شده است. پارامتر شناسایی ویروس در این مقاله تولید N-gram از کد باینری است .

از آن جایی که کد باینری همان محتوای فایل است، پس می توان نتیجه گرفت که کد باینری پارامتر مناسبی برای تشخیص ویروس است. اما استفاده از کد باینری نیازمند فضای زیادی است، ترجیح داده می شود از N-gramتولید شده از کد باینری استفاده شود تا در ذخیره فضا صرفه جویی کرد. [6]

از طرفی منطق فازی که بر اساس نظریه مجموعه‌های فازی ساخته شده است فرمتی چند ارزشی دارد یعنی برخلاف منطق کلاسیک که به هر گزاره یک ارزش ۰ (دروغ مطلق) و یا ۱ (حقیقت مطلق) می‌دهد، درجه ای از حقیقت را به هر گزاره اختصاص می‌دهد. این ویژگی علاوه بر اینکه باعث می‌شود مدل سازی به واقعیت نزدیک شود، منطق فازی را برای تصمیم گیری‌های سریع بر اساس داده‌های موجود مناسب می‌کند. [5]

ما در این مقاله توجه خود را به سمت نوع دیگری از فایل‌ها نیز برده‌ایم. فایل‌هایی موجود هستند که ممکن است رفتار آن‌ها شبیه به ویروس باشد اما خطری برای رایانه ایجاد نکنند، این نوع فایل‌ها را شبه ویروس نام گذاری کرده‌ایم.

در این مقاله با استفاده از N-gram فایل و پایگاه داده‌ای از N-gram های فایل‌های بی‌خطر ، ویروس، شبه ویروس، درصد تطبیق را محاسبه و سپس با قوانین منطق فازی تصمیم گرفته می‌شود که این فایل در کدام دسته قرار دارد.

۲. منطق فازی

فرآیندهایی که در دنیای واقعی مدل می‌شوند معمولا دقیق نیستند. در اکثر اوقات نمی‌توان واقعیتی که همراه با عدم قطعیت است را آن‌چنان که واقعا است مدل کرد و در مدل کردن محدودیت هایی داریم. نتیجه این است که چون واقعیت خوب مدل نشده ، نتیجه مدل سازی (راه حل مدل) نیز کارایی لازم را به عنوان راه حل برای دنیای واقعی ندارد. تئوری فازی اولین بار در سال ۱۹۶۵ توسط لطفی زاده یا زاده معرفی شد.زاده به ناتوانی ریاضیات کلاسیک برای پرداختن به مسائل نادقیق دنیای واقعی اشاره کرد و چارچوب جدیدی به نام تئوری فازی را پایه ریزی کرد و مبانی آن را معرفی نمود.

تئوری فازی یک چارچوب جدید است که توانایی مدل کردن واقعیت را آن چنان که هست دارد.در واقع ضعف چارچوب ریاضیات کلاسیک این بود که واژه‌هایی مثل کم، متوسط و زیاد در آن تعریف نشده بود؛ بنابراین اگر بخواهیم قواعد نادقیق (فازی) را به زبان ریاضیات کلاسیک تعریف کنیم آنچه که به دست می آید ممکن است با آن چه که مدنظر ما بوده متفاوت باشد و کارایی مدنظر ما تأمین نشود. زیرا ریاضیات کلاسیک ابزار لازم برای بیان واقعیت‌های نادقیق فوق را ندارند.[9]

چارچوب جدید فازی سعی می‌کند مدل را با واقعیت بهم نزدیک کند و فاصله بین مدل سازی و تفکر انسان را کم کند. در این چارچوب بستر مناسبی برای تعریف واژه‌های فازی مثل کم، متوسط و زیاد فراهم می‌شود که تطابق خوبی با طرز فکر و احساس انسان دارد.

قبل از آن که بخواهیم منطق فازی و قوانین آن را برای شما معرفی کنیم باید در ابتدا مجموعه‌های فازی و تفاوت آن را با مجموعه‌های کلاسیک را تعریف کنیم.

در ریاضیات کلاسیک مجموعه‌های کلاسیک را داریم که مرزهای یک مجموعه کاملا مشخص هستند و به همین علت به آنها مجموعه Crisp هم گفته می‌شود.

مجموعه‌های کلاسیک در مقابل مجموعه‌های فازی قرار دارند که در آنها مرزها نامشخص هستند.

یک مجموعه کلاسیک ۲ بخش مهم دارد :

۱-اعضای هر مجموعه از یک دیگر قابل تشخیص اند.

۲-برای هر شی داده شده، چه شی داده شده عضو مجموعه ای باشد و چه نباشد، شی موجودیت مجزایی از مجموعه دارد.

مجموعه‌های فازی با رد شرط دوم از مجموعه‌های کلاسیک متفاوت هستند. یعنی برخلاف مجموعه‌های کلاسیک نیازی به مرزهای تند و تیزی برای جدا کردن اجزای مجموعه خود از یک دیگر ندارند. عضویت هر شی در مجموعه مثل مجموعه کلاسیک تصدیق یا تکذیب نمی‌شود، اما موضوع درجه وجود دارد.

یک مجموعه فازی که روی یک مجموعه جهانی تعریف می‌شود دارای یک تابع است، که شبیه به تابع مشخصه مجموعه کلاسیک است. هر کدام از این توابع به هر شی در یک مجموعه یک درجه عضویت را اختصاص می‌دهد. یک مجموعه فازی استاندارد A روی مجموعه جهانی U تابع زیر را دارد :

$$A : U \rightarrow [0 ,1]$$

که این تابع به هر المنت x از U یک عدد A(x) از [۰،۱] را نسبت می‌دهد. و این عدد را درجه عضویت* x در A نامیده می‌شود. پس در نتیجه عدد، درجه حقیقت گزاره " x عضوی از A است " را مشخص می‌کند.

پس در مجموعه های فازی میزان تعلق اعضاء، نسبی است.

در مجموعه‌های کلاسیک ۰ و ۱ ارزش عددی ندارند بلکه نقش آن‌ها نمادین است اما در مجموعه‌های فازی درجه عضویت اختصاص داده شده اهمیت عددی دارد. در مجموعه کلاسیک ۰و ۱ به عنوان علامت در نظر گرفته شده اما در مجموعه فازی ۰ و ۱ و دیگر اعداد به عنوان عدد مشاهده می‌شوند.[8]

برای درک بهتر تابع عضویت فرض کنید یک جمعیت مشخصی را داریم و می‌خواهیم مشخص کنیم که این افراد جوان هستند یا خیر. به عنوان مثال وقتی می‌گوییم "حسن جوان است"، "حسن" عضو مجموعه‌ای به نام "جوان" است که عناصر آن یعنی اشخاص در سنین مختلف به اندازه‌های متفاوت عضو این مجموعه هستند. میزان عضویت افراد در مجموعه "جوان" را با عددی بین صفر و یک نشان می‌دهند که درجه عضویت نامیده می‌شود. درجه عضویت "صفر" یعنی فرد در این مجموعه هیچ عضویتی ندارد، مانند یک فرد هفتاد ساله که می‌توان میزان عضویتش را در مجموعه فازی جوان "صفر" در نظر گرفت و درجه عضویت "یک" یعنی فرد صد در صد عضو مجموعه است مانند یک فرد ۱۸ ساله. از طرفی اگر "حسن" ۳۰ ساله باشد می‌توان او را به اندازه ۰/۷ عضو مجموعه "جوان" دانست.

* Membership Digree

همان‌طور که مشاهده می‌شود، در مجموعه‌های فازی برخلاف مجموعه‌های قطعی عناصر به دو دسته عضو و غیرعضو تقسیم نمی‌شوند، بلکه بر اساس آنچه ما تعریف می‌کنیم میزان عضویت عناصر در مجموعه‌های فازی بین صفر و یک متغییر است.

هر تابع عضویت مجموعه فازی را می‌توان به صورت زیر مشخص کنیم:

$$E(x)=\begin{cases} f_E(x) & \text{when } x\in[a,b) \\ 1 & \text{when } x\in[b,c] \\ g_E(x) & \text{when } x\in(c,d] \\ 0 & \text{otherwise} \end{cases}$$

که در آن a,b,c,d اعداد حقیقی هستند که a<b<c<d است. نماد $f_E(x)$ به یک تابع پیوسته و اکیدا صعودی و نماد $g_E(x)$ به یک تابع پیوسته و اکیدا نزولی اشاره می‌کند. اگر f و g توابع عددی باشند آنگاه تابع E(x) به شکل یک ذوزنقه در می‌آید. در حالات دیگر با توجه به نوع توابع f و g تابع E(x) شکلی به خود می گیرد. مانند مثلثی، گوسی و ... .

توجه داریم که توابع f و g توسط کارشناسان تعریف می‌شوند.

علاوه بر مطالب بالا سه قانون جبری زیر را نیز معرفی می‌کنیم :

C(a)=1-a
t-norm i(a,b)=min(a,b)
t-conorm u(a,b)=max(a,b)

نکته قابل توجه تفاوت فازی و احتمال است. برای روشن کردن این تفاوت فرض کنید که در زندانی اسیر هستید و دو لیوان برای شما آورده شده است که حتماً یکی از آنها را باید بنوشید. لیوان سمت راست لیوانی است که به میزان ۱۰ درصد سمیت دارد و لیوان سمت چپ به احتمال ۱۰ درصد سم است. شما کدامیک را می‌نوشید؟!

در لیوان سمت راست مرز بین سم و غیر سم مشخص نیست (عدم قطعیت از نوع فازی). لیوان سمت چپ یا سم است یا سم نیست ولی مشخص نیست (عدم قطعیت از نوع احتمال). معمولا عدم قطعیت احتمالاتی با گذر زمان از بین می‌رود.

۳. N-gram

شاخص N-gram یک تکنیک برای پردازش پرس و جو های کلمات کلیدی است.N-gram یک دنباله N تایی از واج ها ، هجاها ، حروف ، کلمات یا جفت پایه (base pair) بر اساس نرم افزار است.در ارتباط با سایز یک N-gram، ۱ به Unigram، ۲ به Bigram و ۳ به Trigram اشاره می‌کند. البته مقدار N می تواند اعداد بزرگ تر را نیز اختیار کند.

فرض کنید که عبارت re*ve را داریم؛ یعنی می‌خواهیم تمام کلماتی که با re شروع و با ve خاتمه می‌یابد را پیدا کنیم. پس ما دو N-gram، re و ve را داریم، پروسه جست و جو در لیست‌ها انجام می‌شود و لیستی از تطبیق به ما ارایه می‌شود.

نکته حائز اهمیت در N-gram این است که این تکنیک در عین سادگی، جست‌وجوی آن می‌تواند گران باشد. به همین دلیل محققان در حال ایجاد روش های جدید و کم هزینه در جست‌وجو هستند یکی از این نمونه کار ها مقاله Wesley Jin and Charles Hines [3] است.

اما اکنون این سوال مطرح می‌شود که N-gram معرفی شده که بیشتر شبیه به یک ابزار در جست‌وجوی متون است چه کمکی به ما در شناسایی ویروس ها می‌کند ؟ برای پاسخ به این سوال مراحل ایجاد یک شاخص N-gram و طرز جست‌وجو در آن را توضیح می‌دهیم. [7]

۱)استخراج N-gram

همان‌طور که قبلا گفته شد N-gram ها دنباله‌هایی از بایت‌ها به طول N هستند. برای تولید N-gram ها از تکنیک کشیدن یا از ته زدن(shingling) استفاده می‌کنیم، به این صورت که یک بافر به طول بزرگ تر یا مساوی N داریم. یک پنجره کشویی به طول N را فرض می‌کنیم و در هر واحد زمان پنجره کشویی را یک بایت به جلو می‌بریم. محتویات پنجره در هر بار تکرار یک N-gram را تشکیل می‌دهد.

برای مثال رشته ABCDEFG را در نظر بگیرید. N-gram های با N=4 به صورت زیر است :

ABCD,BCDE,CDEF,DEFG

۲)لیست سازی :

ایجاد لیستی است از فایل هایی که در آن داده‌ها می‌توانند یافت شوند. این لیست شامل دو بخش است؛ اول فهرست N-gram ها و دوم برای هر N-gram نوشته‌هایی به صورت زیر داریم :

<FileID1, offset11, offset12, …, frequency1>

که در آن FileIDx یک شناسه منحصر به فرد از فایل‌هایی است که شامل این N-gram می‌شود.offsetxy موقعیت N-gram در فایل x را مشخص می‌کند و frequency نشان دهنده تعداد دفعاتی است که N-gram در فایل تکرار شده است.

لیست N-gram ها اجازه می‌دهد تا لیست خاص برای N-gram ورودی، به سرعت پیدا شود.

۳)جست و جو :

جست و جو در شاخص‌ها شامل ۴ مرحله است :

- تجزیه رشته ورودی به N-gram های موجود در رشته
- اصلاح ورودی‌های ارسالی برای هر N-gram
- پیدا کردن FileID رایج در بین ورودی
- اطمینان پیدا کردن از اینکه offset ها به طور متوالی منظم هستند.

برای مثال یک شاخص n=4 که شامل ۳ فایل زیر است را در نظر بگیرید :

File 1: AAADEADBBB
File 2: ADEADBEEFC
File 3: DEADBEECBEEF

و فرض کنید رشته‌ای را که می خواهیم جست و جو کنیم DEADBEEF است.4-gram های آن به صورت زیر است :

DEAD, EADB, ADBE, DBEE, BEEF

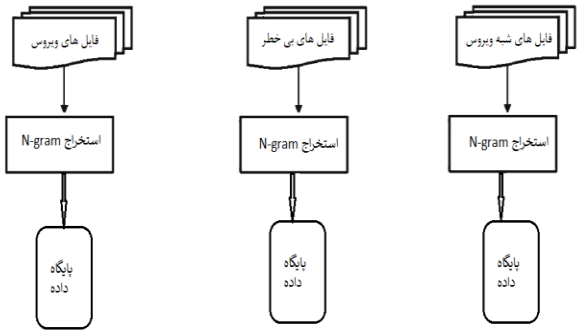
فایل ۱ شامل سه تا از پنج تا 4-gram است.

فایل ۲ شامل تمام 4-gram ها است و تمام 4-gram ها به ترتیب در امتداد بعدی آمده است.

فایل ۳ نیز شامل تمام 4-gram ها است اما با این حال BEEF در موقعیت ۸ است که DBEE را در موقعیت ۳ دنبال نمی‌کند. بنابراین این یک تطابق نیست.

۴.شناسایی ویروس ها

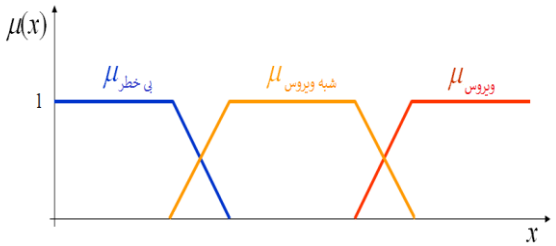
در این روش در ابتدا باید به سیستم آموزش دهیم تا بتواند ویروس ها را تشخیص دهد. برای این کار در ابتدا باید سه پایگاه داده ایجاد کنیم.برای سه نوع فایل ویروس، شبه ویروس، بی خطر N-gram های آنان را تولید کرده و پایگاه داده‌های جدا ایجاد می‌کنیم.



اکنون فایل ناشناخته جدیدی را به سیستم ارائه می‌دهیم. در ابتدا سیستم N-gram آن را تولید می‌کند. سپس N-gram تولید شده را با رشته N-gram های موجود در پایگاه داده‌ها مقایسه می‌کنیم.

اکنون باید با کمک قوانین و منطق فازی تشخیص دهیم آیا ویروس است یا شبه ویروس و یا بی خطر است؟

نمودار توابع عضویت مجموعه فازی {بی خطر ،شبه ویروس ،ویروس } به صورت زیر می‌باشد.



برای نوشتن قوانین منطق فازی از t-conorm استفاده می‌کنیم. برای مثال فرض کنیم فایل x را داریم، تطابق N-gram فایل x با پایگاه داده بی‌خطر ۵٪ ، با پایگاه داده شبه ویروس ۷۰٪و با پایگاه داده ویروس ۳۵٪ است. آنگاه با توجه به t-conorm اظهار می‌کنیم که فایل x یک فایل شبه ویروس است.

۵. نتیجه‌گیری

ایده مطرح شده اگرچه بسیار ساده می‌باشد، اما باید توجه کنید که در بسیاری از مباحث قدرت بالای حل مسائل با استفاده از فازی ثابت شده است. البته که این مساله باید به صورت عملی بررسی شود.

اما یکی از نکات ساده ولی مهم ایجاد پایگاه داده های ذکر شده است. مسلما هر چه پایگاه داده بزرگ تر باشد احتمال خطا پایین تر می‌آید .اما مبحث نگه داری و جست‌وجو در این پایگاه داده مساله ای است که نمی‌توان از آن چشم پوشی کرد.

۶. مراجع

1. Obi J.C and Okpor D.M .(2013),” SOFT-COMPUTING VIRUS IDENTIFICATION SYSTEM” , International Journal of Fuzzy Logic Systems (IJFLS) Vol.3, No2,PP 63-72

2. Mai Trong Khang and Vu Thanh Nguyen and Tuan Dinh Le (2015),” A Combination of Artificial Neural Network and Artificial Immune System for Virus Detection”, REV Journal on Electronics and Communications, Vol. 5, No. 3–4,PP 52-57

3. Wesley Jin and Charles Hines and Cory Cohen and Priya Narasimhan (2012),” A Scalable Search Index for Binary Files”, 7th International Conference on Malicious and Unwanted Software , PP 94-103

4. Ankur Singh Bist ,(2014),” Fuzzy Logic for Computer Virus Detection”, International Journal of Engineering Sciences & Research Technology,PP 771-773

5. Pratik C. Patel and Upasna Singh ,(2013),”Detection of Data Theft using Fuzzy Inference System”, 3rd IEEE International Advance Computing Conference (IACC),PP 702-707

6. He Huang,(2015),” BINARY CODE REUSE DETECTION FOR REVERSE ENGINEERING AND MALWARE ANALYSIS”, Master of Applied Science in Information Systems Security, Montr´eal, Qu´ebec, Canada.

7. D Krishna Sandeep Reddy and Arun K Pujari,(2006),” *N*-gram analysis for computer virus detection”, J Comput Virol,vol.2,PP 231-239

8. Radim Belohlavek ,2011,” Concepts and fuzzy logic “, Graphic Composition, London, England

9. Lotfi A. Zadeh ,(2008), “Is there a need for fuzzy logic?”, Information Sciences,vol. ,pp. 2751–2779

10.https://www.mitre.org/capabilities/cybersecurity/overview/cybersecurity-blog/snugglefish-provides-quick-pattern-matching