УДК 681.3.06

І.А. Терейковський

Київський національний університет будівництва і архітектури, Київ

РОЗПІЗНАВАННЯ СКРИПТОВИХ ВІРУСІВ ЗА ДОПОМОГОЮ НЕЙРОННОЇ МЕРЕЖІ З РАДІАЛЬНИМИ БАЗИСНИМИ ФУНКЦІЯМИ

Розглянуто можливість використання нейронної мережі з радіальними базисними функціями в системах розпізнавання скриптових вірусів. Розроблено модель такої нейронної мережі. Доведено доцільність її застосування.

Ключові слова: антивірус, скриптовий вірус, захист інформації, нейронна мережа з радіальними базисними функціями

Постановка проблеми

На сьогоднішній день захист від комп'ютерних вірусів є однією із найбільш важливих та актуальних проблем в галузі захисту інформації. Необхідність посилення антивірусного захисту підтверджується великою кількістю відомих прикладів зараження комп'ютерних систем, яке призводить не тільки до втрати функціональних можливостей, але й до несанкціонованого використання заражених систем. Наприклад, досить часто заражені вірусом комп'ютерні системи несанкціоновано розсилають спам-повідомлення або/та приймають участь у розподілених DoS-атаках. Разом з тим вважається, що основні труднощі створення ефективного захисту пов'язані з розпізнаванням вірусів. При цьому велика увага звертається на розпізнавання вірусів, які розмножуються за допомогою мережі Internet. Засобами розповсюдження таких вірусів є заражені електронні листи та веб-сайти. Небезпека посилюється популяризацією веб-орієнтованих соціальних мереж, використання яких потребує встановлення на комп'ютер клієнта потенційно небезпечного спеціалізованого програмного забезпечення, розташованого на веб-сервері. Практичний досвід та статистичні дані, що опубліковані на веб-сайтах "Лабораторії Касперського" та "Dr. Web", свідчать, що листи електронної пошти та веб-сайти заражаються за допомогою вірусів написаних на скриптових мовах програмування. Також можна зробити висновок про необхідність вдосконалення сучасних систем розпізнавання скриптових вірусів, що і є основною проблемою даної статті. Проблема безпосередньо пов'язана з таким важливим науково-практичним завданням, як забезпечення надійності функціонування розподілених комп'ютерних систем та мереж.

Аналіз останніх досліджень і публікацій, на які спирається автор

Відповідно до вітчизняної нормативної документації вважатимемо, що комп'ютерний вірус - це програма, що розмножується та поширюється самочинно. Разом з тим комп'ютерний вірус може порушувати цілісність інформації, програмне забезпечення та режим роботи обчислювальної техніки. При цьому вірус, написаний на скриптовій мові програмування, будемо називати скриптовим вірусом. Скриптовий вірус, який використовує можливості макромов, вбудованих в системи обробки даних, називають макровірусом [3, 4]. Незважаючи на середовище розповсюдження, скриптовий вірус представляє собою скрипт або макрос, що виконується внаслідок реалізації певної події. На сьогодні найбільш поширені макровіруси, що пристосовані для функціонування в середовищі MS Office, написані на мові програмування Visual Basic for Applications. Це пояснюється популярністю як самого пакету MS Office, так і мови VBA. Крім того, в сучасних версіях операційної системи Windows вбудовано скриптовий інтерпретатор Windows Scripting Host, який дозволяє виконувати скрипти (макроси), що написані на мовах програмування VBScript та JScript. Запуск макроса може бути здійснено користувачем при відкритті файлу. Ця особливість використовується для активізації вірусів, що розповсюджуються за допомогою файлів, прикріплених до листів електронної пошти та вірусів, які маскуються під спеціалізоване програмне забезпечення, необхідне для перегляду ресурсів веб-орієнтованих соціальних мереж. Результати досліджень [3, 4, 5] вказують на те, що більшість таких вірусів функціонує в середовищі інтерпретатора Wscript і написані на мові програмування VBScript. При цьому базою VBScript і Visual Basic for Applications є мова Microsoft Visual Basic. Тому, за своєю суттю та за технологією

створення описані скриптові віруси не відрізняються від макровірусів та макротроянів MS Office.

В основному для захисту від скриптових вірусів, як і для вірусів в цілому, використовуються антивірусні сканери. Принцип їх роботи базується на постійній або періодичній перевірці файлів, секторів дисків та системної пам'яті на предмет виявлення в них відомих та невідомих вірусів. Найчастіше для виявлення вірусів використовується метод пошуку сигнатур. Сигнатура вірусу представляє собою характерну для цього вірусу послідовність команд програмного коду. Для виявлення відомих вірусів аналізуємий програмний код співвідноситься з базою даних відомих сигнатур вірусів. Особливістю пошуку сигнатур скриптових вірусів є те, що сканер може аналізувати програмний код в текстовому вигляді. В тому випадку, коли фрагмент аналізуємого програмного коду відповідає певній сигнатурі, це свідчить про зараження файлу певним вірусом. З цієї причини метод сигнатур дозволяє розпізнавати тільки відомі віруси та відкриває шлях для обходу антивірусного захисту поліморфним та стелс-вірусам.

Ще одним важливим недоліком методу сигнатур є необхідність постійного оновлення антивірусної бази даних користувачами та постійного функціонування розгалуженої служби підтримки, яка виявляє нові віруси та оновлює базу даних. В деяких антивірусних сканерах для розпізнавання невідомих вірусів разом з методом сигнатур використовується так званий евристичний аналіз. При цьому в різних антивірусних системах застосовуються різні евристичні методи, реалізація яких практично не документується. Проте аналіз джерел [4, 5] вказує на те, що в більшості випадків базою цих методів є статистичний аналіз послідовності виконання програмного коду об'єкта, що перевіряється. Відзначимо, що сучасні евристичні методи дозволяють виявити тільки близько 50% вірусів, сигнатура яких не представлена в антивірусній базі даних. В [3,4] пропонується підвищити достовірність розпізнавання за допомогою використання нейронних мереж. Також [4] розроблена модель нейронної мережі типу «багатошаровий персептрон» призначена для розпізнавання скриптових вірусів. Однак, висока обчислювальна складність навчання багатошарового персептрону ускладнює його практичне застосування та зумовлює необхідність розвідувального аналізу поставленої задачі.

В теорії нейронних мереж [1, 2] рекомендується проводити розвідувальний аналіз за допомогою нейронної мережі з радіальними базисними функціями (РБФ). Базова модель РБФ складається із трьох шарів: вхідного, схованого та вихідного. В задачу вхідного шару входить розподіл вхідних даних за нейронами схованого шару з Гаусівською функцією активації:

$$\varphi_j(net) = exp\left(-\frac{1}{2\sigma}\sum_{i=1}^{N}(c_i - x_i)^2\right), \qquad (1)$$

де $\varphi_j(net)$ – функція активації j-го нейрону проміжного шару; net – сумарний вхідний сигнал; x – вхідний вектор; c – центр функції Гауса; σ – радіус функції Гауса; N – кількість вхідних нейронів.

Кожен із схованих нейронів призначений для зберігання окремого еталонного образу, який відповідає окремому класу. Як правило, кількість нейронів у схованому шарі більша кількості вхідних нейронів. Після нелінійного перетворення сигнали від нейронів схованого шару потрапляють до вихідного шару нейронів, які мають лінійні функції активації. Сукупність значень активностей всіх схованих нейронів визначає вектор, на який відображається вхідний вектор.

Мережа РБФ потребує навчання, яке реалізується поетапно методом «з вчителем». На першому етапі розраховуються кількість нейронів в схованому шарі та коефіцієнти (центр і радіус функції Гауса) для функцій активації нейронів схованого шару. Для розрахунку центру функції Гауса рекомендується використовувати метод "К-середніх". Наступним етапом навчання є розрахунок радіусів функцій Гауса. Після розрахунку параметрів функції Гауса, які за своєю суттю представляють вагові коефіцієнти нейронів схованого шару необхідно визначити вагові коефіцієнти нейронів вихідного шару.

Формулювання мети статті

Розробити модель нейронної мережі РБФ призначеної для використання в антивірусному сканері для розпізнавання скриптових вірусів написаних на мові програмування Microsoft Visual Basic.

Виклад основного матеріалу дослідження

Як і для більшості нейронних мереж, розробка моделі РБФ починається з визначення вхідних та вихідних параметрів. При цьому номенклатура вхідних параметрів повинна відображати здатність скриптових вірусів до саморозповсюдження та характерні спільні ознаки скриптових вірусів та троянів [3, 4]. В даній статті розглядаються скриптові віруси написані на мові програмування Microsoft Visual Basic, яка дозволяє працювати з файловою системою, встановлювати мережеві з'єднання, маніпулювати процесами і потоками, здійснювати виклик функцій API ОС та запускати

зовнішні програми. Шляхи розповсюдження поштових скриптових вірусів та макровірусів не обмежені програмним середовищем зараженого документу. Наприклад, макровірус MS Word може заражати документи AutoCAD або командні файли. Засобами розповсюдження макровірусу в цьому випадку можуть бути об'єкти відповідних бібліотек, функції API ОС та програмні додатки. Крім того, вхідні параметрів нейронної мережі повинні враховувати характерні ознаки скриптових вірусів та троянів, які можливо розділити на групи: автоматизації запуску, ігнорування помилок, маскування та деструктивних функцій.

Приблизний перелік ознак представлений в табл. 1. Безпосередньо вхідними параметрами будуть фрагменти коду (назви функцій, параметрів, об'єктів, бібліотек, методів та властивостей об'єктів), що відповідають ознакам скриптових вірусів та макровірусів. Наприклад, однією із характерних ознак макровірусу є використання автомакросів. Відповідними параметрами будуть AutoOpen, AutoClose та інші аналогічні назви макросів. Вхідні параметри можуть мати два значення: «1», якщо ознака присутня, та «-1» – в протилежному випадку. Кількість вхідних параметрів N буде дорівнювати кількості відповідних фрагментів.

Таблиця 1

Характерні спільні ознаки скриптових вірусів та троянів

Назва групи	Перелік ознак
Автоматизація запуску	Використання автомакросів
Ігнорування помилок	Використання операторів ігнорування помилок та переходу на певний рядок програмного коду після виникнення помилок
Маскування вірусу	Шифрування/дешифрування макросу, захист макросу паролем, відключення захисту від макровірусів, блокування та перевизначення кодів клавіш, зміна шрифту макросів, запис/зчитування інформації в буфер обміну, відключення редактора VBA, знищення панелі інструментів для роботи з макросами та шаблонами, ігнорування повідомлень програмного середовища, поліморфізм макровірусу, використання функцій, що порушують функціонування антивіруса, знищення або перейменування файлу та модулю з вірусом, знищення процедури з вірусом
Деструктивні функції	Форматування жорстких дисків, модифікація та знищення файлів, встановлення паролів на файли, встановлення мережевих з'єднань, доступ до поштових клієнтів

Закінчення таблиці 1

Відповідно висновків [1;2] встановлено, що вихідний шар РБФ буде складатися тільки із одного нейрону. При цьому з метою забезпечення гнучкості системи розпізнавання визначено, що вихід РБФ повинен містити в собі числову оцінку класифікації скрипта. Це дозволяє проводити класифікацію всіх скриптів на три класи: безпечні, віруси та підозрілі. До підозрілих слід віднести скрипти, в яких знайдено тільки окремі ознаки вірусів, наприклад, зашифрований програмний код. Реалізувати такий вихід мережі РБФ можливо за рахунок одного вихідного елементу. Встановлено, що вихід РБФ для навчальних прикладів, які відповідають вірусам, дорівнює 1, а вихід для навчальних прикладів, що відповідають безпечним макросам, дорівнює – 1. На практиці можливим результатом функціонування РБФ будуть відмінні від -1 та 1 величини вихідних сигналів. Так, мережа буде сигналізувати про оцінку належності скрипта до одного із класів. Остаточну класифікацію можливо провести відповідно до методики, яка представлена в табл. 2.

Таблиця 2

Класифікація скриптів відповідно до величини виходу мережі РБФ

Величина вихідного сигналу Y	Назва класу
$0,5<Y$	Вірус
$Y<-0,5$	Безпечний макрос
$-0,5\leq Y\leq 0,5$	Підозрілий макрос

Як статистичний матеріал були використані сигнатури скриптових вірусів, що входять до складу бази даних антивірусних пакетів. Вхідні параметри РБФ були отримані шляхом аналізу цих сигнатур на предмет наявності потенційно небезпечних операторів. Кількість вхідних параметрів, а значить, і кількість вхідних елементів РБФ дорівнює 105. Відповідно наявного статистичного матеріалу та рекомендацій [1; 2], сформовано навчальну вибірку з 560 прикладів. Через обмежену кількість сигнатур вірусів частина прикладів повторювалась

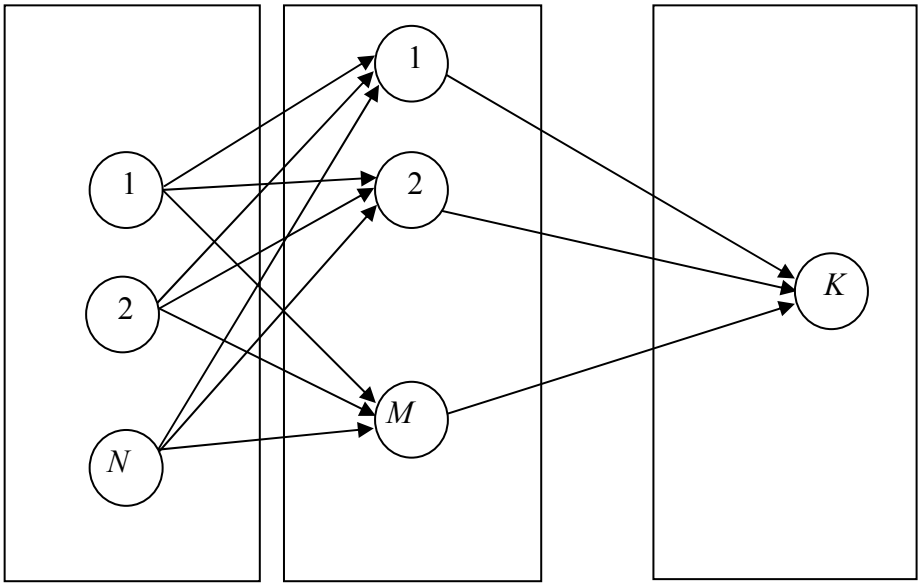

Рис.1. Структура мережі РБФ для розпізнавання скриптових вірусів

Використано приклад, в якому наявні всі параметри, характерні для вірусу, та приклад, в якому відсутні всі параметри, характерні для вірусу. Половина прикладів відповідала скриптовим вірусам, а інша половина – безпечним скриптам.

Визначення множини вхідних та вихідних параметрів та результатів [2] дозволяє перейти до розрахунку параметрів моделі РБФ, структура якої показана на рис. 1, а математичне забезпечення складають вирази (1-6). Зазначимо, що на рис. 1 символами *N*, *M* та *K* позначено кількість вхідних, схованих та вихідних нейронів. Сумарний вхідний сигнал (*net*) для *j*-го нейрону в схованому шарі від вхідного вектора (*x*):

$$net_j = \sqrt{\sum_{i=1}^{N} (x_i - w_{ij})^2} \qquad (2)$$

де x_i – *i*-а компонента вхідного вектору *x*, w_{ij} – ваговий коефіцієнт *j*-го схованого нейрону з *i*-м вхідним нейроном, *N* – кількість вхідних нейронів.

Розрахунок сумарного вхідного сигналу для будь-якого нейрону вихідного шару проводиться відповідно (1).

Розрахунок загальної кількості синаптичних зв'язків (Z_Σ) мережі РБФ реалізується так

$$Z_\Sigma = Z_1 + Z_2, \qquad (3)$$

$$Z_1 = N \times M, \qquad (4)$$

$$Z_2 = M \times K, \qquad (5)$$

де Z_1 – кількість синаптичних зв'язків схованих нейронів, *N* – кількість вхідних нейронів, *M* – кількість схованих нейронів, Z_2 – кількість синаптичних зв'язків вихідних нейронів, *K* – кількість вихідних нейронів.

У процесі навчання корекція вагових коефіцієнтів *j*-го нейрону вихідного шару відбувається так:

$$\Delta w_j = \frac{\eta \times net \times \delta}{M}, \qquad (6)$$

де *M* – кількість схованих нейронів, η – норма навчання, $\delta = (w^f - w^j)$ – помилка вихідного сигналу для *j*-го нейрону, w_f – фактичний вихід, w_o – очікуваний вихід мережі, *net* – сумарний вхідний сигнал вихідного нейрону.

Для визначення оптимальних параметрів, з урахуванням [1; 2] було побудовано чотири різних моделі РБФ. Спільною рисою всіх моделей була кількість вхідних параметрів (вхідних нейронів) *N*=105 та вихідних параметрів (вихідних нейронів) *K*=1. Відмінності в моделях полягали в кількості схованих нейронів (*M*). Для першої мережі *M*=72, для другої *M*=105, для третьої *M*=210, для четвертої *M*=315. Відповідно в першій РБФ кількість схованих нейронів в два рази менша кількості вхідних параметрів, в другій – дорівнює цій кількості, в третій – в два рази більша, в четвертій – в три рази більша. Таким чином, відповідно до теоретичних рекомендацій [2], була здійснена спроба адаптації структури РБФ до умов поставленої задачі. Прийнято, що радіус функції Гауса σ=0,5, норма навчання η=0,1. Навчання мережі здійснювалось

відповідно до виразів (2-3). Перевірена якість навчання кожної із мереж при 1, 10, 100, 500 та 1000 навчальних ітерацій.

Результати перевірки показали, що для перших трьох типів мереж середня відносна похибка вихідного сигналу знаходилась в межах 30-50%. Тільки для четвертої мережі РБФ з кількістю схованих нейронів M=315 при 1000 навчальних ітераціях вказана похибка зменшилась приблизно до 15%. На тестових прикладах, що не входили до навчальної вибірки, середня відносна помилка вихідного сигналу збільшилась в 1,5-2 рази. На погляд автора, що збігається з висновками [2], причиною низької якості РБФ в першу чергу є емпіричність визначення важливих параметрів мережі: кількості схованих нейронів, радіусу функції Гауса та норми навчання. Тому для якісного використання РБФ потрібно провести додаткові теоретичні дослідження, спрямовані на розробку чіткої методики формування параметрів мережі, які адекватно відповідають умовам задачі розпізнавання скриптових вірусів. Разом з тим, проведені експерименти вказують на перспективність вдосконалення антивірусних засобів за рахунок використання нейронних мереж. Також очевидно, що на практиці слід використовувати або більш потужні типи класичних моделей нейронних мереж, або розробити нову модель призначену для розпізнавання вірусів.

Висновки

Проведені дослідження вказують на доцільність використання нейронної мережі РБФ з метою розвідувального аналізу даних в задачах розпізнавання скрипкових вірусів.

Для підвищення ефективності застосування мережі РБФ в антивірусних засобах слід вдосконалити методику розрахунку таких параметрів мережі, як кількість схованих нейронів, радіус функції Гауса та норма навчання.

Основні перспективи подальших розвідок у даному напряму полягають у розробці моделей нейронних мереж, адаптованих для розпізнавання комп'ютерних вірусів.

Список літератури

1. Ежов А. А. Нейрокомпьютинг и его применения в экономике и бизнесе / А. А. Ежов, С. А. Шумский. – М. : МИФИ, 1998. – 224 с.

2. Каллан Р. Основные концепции нейронных сетей / Каллан Р. ; пер. с англ. А. Г. Сивака. – М. : Вильямс, 2003. – 288 с.

3. Терейковський І.А. Використання нейронних мереж при розпізнаванні макровірусів / Правове, нормативне та метрологічне забезпечення системи захисту інформації в Україні Випуск 2 (13) 2006 р., с.176-183.

4. Терейковський І. А. Розпізнавання скриптових вірусів за допомогою багатошарового перспетрону / І. А. Терейковський // Защита информации: сб. науч. трудов НАУ. – 2007. – Выпуск 14. – С. 206–212.

5. Огарок А. Виртуальные войны. Искусственный интеллект на защите от вирусов и программных закладок / А. Огарок, Д. Комашинский, Д. Школьников // Конфидент. – 2003. – №2 (50). – С. 64–69, 97.

Стаття надійшла до редколегії 7.12.2010

Рецензент: д-р техн. наук, проф. С.В. Цюцюра, Київський національний університет будівництва і архітектури, Київ.