УДК 004.8.565.5; 004.621:681.324

Косюк Є. С., студент (Тел. +380 (96) 702 86 28. E-mail: yevgeniy.kosyuk@gmail.com)
(Київський національний університет культури і мистецтв)

ВИКОРИСТАННЯ НЕЙРОННИХ МЕРЕЖ З ПРЯМИМ РОЗПОВСЮДЖЕННЯМ СИГНАЛУ ДЛЯ РОЗПІЗНАВАННЯ СКРИПТОВОГО ШКІДЛИВОГО ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

Косюк Є.С. Використання нейронних мереж з прямим розповсюдженням сигналу для розпізнавання скриптового шкідливого програмного забезпечення. В роботі розглянуті питання підвищення ефективності систем захисту комп'ютерної інформації. Виконано аналіз способів захисту комп'ютерних систем від скриптових вірусів, досліджено методи та алгоритми розпізнавання скриптового шкідливого програмного забезпечення. Запропоновано підхід та розроблена методика оптимізації структури двохшарового персептрону, призначеного для розпізнавання скриптового шкідливого програмного забезпечення. Обґрунтовані можливі сфери використання запропонованих рішень.

Ключові слова: нейронна мережа, скриптовий вірус, захист інформації, багатошаровий персептрон

Косюк Е. С. Использование нейронних сетей с прямым распространением сигнала для распознавания скриптового вредоносного программного обеспечения. В работе рассмотрены вопросы повышения эффективности систем защиты компьютерной информации. Выполнен анализ способов защиты компьютерных систем от скриптових вирусов, исследованы методы и алгоритмы распознавания скриптового вредоносного программного обеспечения. Предложен подход и разработана методика оптимизации структуры двухслойного персептрона, предназначенного для распознавания скриптового вредоносного программного обеспечения. Обоснованные возможные сферы использования предложенных решений.

Ключевые слова: нейронная сеть, скриптовий вирус, защита информации, многослойный персептрон

1. Постановка задачі. За останні декілька років значна кількість успішних атак на ресурси комп'ютерних систем значно зросла, а більшість реалізована за допомогою вірусів, які проникають за допомогою виконання сценаріїв скриптів вбудованих у WEB-сторінки. Цей факт показує важливість вдосконалення відповідних методів та засобів захисту інформації. І хоча розробці систем захисту від вірусів присвячено багато досліджень, практичний досвід та висновки більшості літературних джерел вказують на їх відносно низьку ефективність. На данний момент, найбільш ймовірним шляхом зараження є перегляд WEB- сайтів. Для інтернет-ресурсів шкідливе програмне забезпечення, як правило, розроблюється за допомогою скриптових мов програмування. Таким чином, віруси написані на скриптових мовах програмування є одною із основних загроз безпеці комп'ютерної системи. Метою даної статті є покращення системи захисту комп'ютерної інформації за рахунок підвищення ефективності розпізнавання зашифрованих скриптових комп'ютерних вірусів за допомогою нейронних мереж з прямим розповсюдженням сигналу.

Аналіз останніх досліджень і публікацій [1...5] вказує на те, що ефективність застосування нейромережевої моделі безпосередньо залежить від того, наскільки її тип та параметри оптимізовані відповідно умов поставленої задачі. При цьому, відповідно [3], для вирішення задачі розпізнавання скриптового шкідливого програмного забезпечення оптимальним типом нейромережевої моделі є багатошаровий персептрон.

В загальному випадку багатошаровий персептрон, структура якого показана на Рис. 1, представляє собою нейронну мережу, яка складається із декількох послідовно з'єднаних між собою шарів штучних нейронів [1, 5]. Зовнішня інформація поступає у вхідний шар, основними завданнями якого є прийом та розповсюдження вхідної інформації по іншим шарам нейронної мережі. Далі знаходиться один або декілька схованих шарів нейронів, в

яких відбувається основна обробка інформації, результати якої відображаються у вихідному шарі. Зв'язки між нейронами одного шару відсутні. Інформація розповсюджується тільки у напрямку "вхід→вихід".

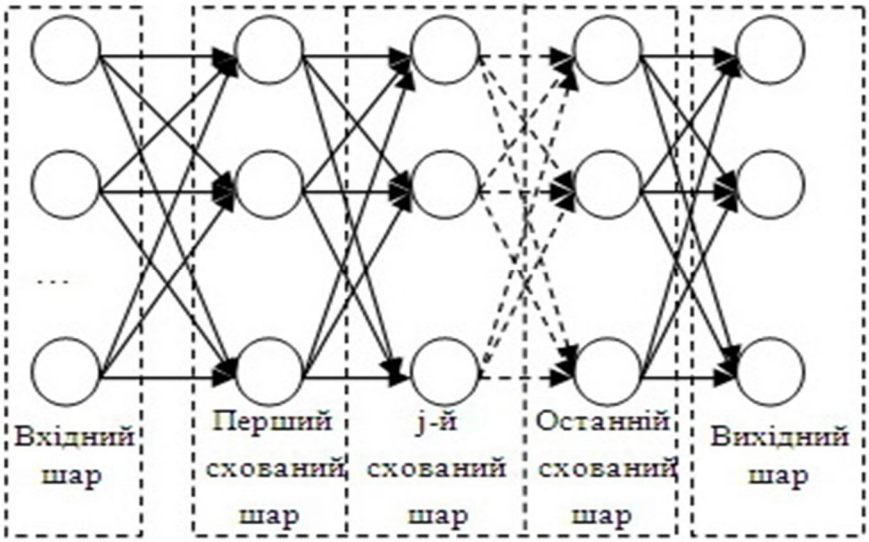

Рис. 1 Структура багатошарового персептрону

Для вхідних нейронів найчастіше використовується лінійна функції активації. Для схованих нейронів як правило використовують сигмоїдальну функцію активації або гіперболічний тангенс

$$F(NET) = \frac{1}{1+e^{-a\times NET}}, \tag{1}$$

$$F(NET) = \frac{e^{a\times NET} - e^{-a\times NET}}{e^{a\times NET} + e^{-a\times NET}}, \tag{2}$$

де *NET* – сумарний зважений вхідний сигна; *F(NET)* – вихідний сигнал нейрону.

В більшості випадків вихідні нейрони виконують тільки розрахунок власних вхідних сигналів, тому функція активації для них не використовується. Розрахунок параметрів *j*-го нейрону в *l*-му схованому шарі здійснюється так

$$NET_j^l = \sum_{i=1}^{K_j^l} w_{ij}^l x_{ij}^l, \tag{3}$$

$$OUT_j^l = F(NET_j^l - \theta_j^l), \tag{4}$$

$$x_{ij}^{l+1} = OUT_j^l, \tag{5}$$

де *i* – номер входу; *j* – номер нейрону в шарі; *l* – номер схованого шару;

K_j^l – кількість вхідних зв'язків *j*-го нейрону в *l*-му шарі;

w_{ij}^l – ваговий коефіцієнт *i*-го входу *j*-го нейрону в *l*-му шарі;

θ_j^l – пороговий рівень активації *j*-го нейрону в *l*-му шарі;

x_{ij}^l – *i*-й вхідний сигнал нейрону в *l*-му шарі;

F – функція активації; OUT_j^l – вихідний сигнал;

NET_j^l – сумарний вхідний сигнал j-го нейрону в l-му схованому шарі.

В [2, 5] наведено вирази для оцінки оптимальної кількості синаптичних зв'язків та кількості схованих нейронів в двохшаровому персептроні з сигмоїдальними функціями активації:

$$\frac{N_0 P}{1+log_2 P} \le L_w \le N_1\left(\frac{P}{N_1}+1\right)(N_1+N_0+1)+N_1, \tag{6}$$

$$\frac{P}{10}-N_1-N_0 \le L_w \le \frac{P}{2}-N_1-N_0, \tag{7}$$

$$L_w < P \times \varepsilon_{max}, \tag{8}$$

де L_w – кількість синаптичних зв'язків; ε_{max} – максимальна допустима помилка узагальнення; N_0 – кількість вхідних нейронів; N_1 – кількість схованих нейронів.

Також наведено формулу для визначення максимальної кількості образів (P), яку може запам'ятати двохшаровий персептрон з пороговою функцією активації:

$$\frac{L_w}{N_0} < P < \frac{L_w}{N_0} log_2\left(\frac{L_w}{N_0}\right), \tag{9}$$

Зазначимо, що вирази (6)...(9) наведені в [2, 7] без належного теоретичного обґрунтування, методика їх визначення описана недостатньо, а практичний досвід свідчить про низьку точність. Перспективним шляхом виправлення вказаних недоліків може стати створення відповідної методики оптимізації.

Метою даного дослідження є підвищення ефективності систем захисту комп'ютерної інформації, аналіз існуючих рішень по захисту комп'ютера від скриптових вірусів, розробка методів та алгоритмів деобфускації зашифрованих скриптів та аналізу на предмет вірусної активності. Для досягнення поставленої задачі розробляється модель нейронної мережі призначеної для використання в антивірусному сканері для розпізнавання скриптових вірусів написаних на мові програмування Microsoft Visual Basic.

2. Виклад основного матеріалу дослідження. Одним із найбільш важливих критеріїв ефективності функціонування багатошарового персептрону є помилка узагальнення [1, 2, 5]. Виходячи з цих міркувань в якості критерію оптимізації кількості синаптичних зв'язків оберемо помилку узагальнення двохшарового персептрону. Запишемо відповідну математичну модель

$$\varepsilon(L_w) \to min, L_w \in \{1,\infty,1\}, \tag{10}$$

де ε – помилка узагальнення; L_w – кількість синаптичних зв'язків.

В [1] доведено, що в загальному випадку помилка узагальнення, складається із помилки апроксимації (ε_a) та помилки опису моделі (ε_o)

$$\varepsilon = \varepsilon_a + \varepsilon_o. \tag{11}$$

Також показано, що помилку апроксимації багатошарового персептрону можливо оцінити так

$$\varepsilon_a \sim \frac{N_1}{L_w}, \tag{12}$$

де N_1 – кількість компонент вхідного вектора (розмірність вхідного вектору).

В першому наближенні помилку опису моделі багатошарового персептрону можливо оцінити так:

$$\varepsilon_o \sim \frac{L_w}{P}, \tag{13}$$

де P – кількість навчальних прикладів.

Зазначимо, що в (12, 13) знак «~» означає пропорційність. Перепишемо пропорції (12, 13) наступним чином:

$$\varepsilon_a = k_a \times \frac{N_1}{L_w}, \tag{14}$$

$$\varepsilon_o = k_o \times \frac{L_w}{P}, \tag{15}$$

де k_a, k_o – фіксовані коефіцієнти, що належать масиву натуральних чисел.

В першому наближені можна прийняти, що

$$k_a = k_o = k, \tag{16}$$

де k – деяке натуральне число.

Підстановка (14)...(16) в (11) дозволяє отримати наступний вираз для оцінки помилки узагальнення:

$$\varepsilon = \left(\frac{k \times N_1}{L_w} + \frac{k \times L_w}{P} \right). \tag{17}$$

Для знаходження точки мінімуму функції (10) з урахуванням (17 продиференцюємо функцію $\varepsilon(L_w)$:

$$\frac{d\varepsilon}{dL_w} = \frac{d\left(\frac{k \times N_1}{L_w} + \frac{k \times L_w}{P} \right)}{dL_w}. \tag{18}$$

Проведемо спрощення виразу (18):

$$\frac{d\varepsilon}{dL_w} = \frac{d\left(k \times N_1 \times L_w^{-1}\right)}{dL_w} + \frac{d\left(k \times P^{-1} \times L_w\right)}{dL_w}. \tag{19}$$

Остаточно диференціал описується виразом:

$$\frac{d\varepsilon}{dL_w} = \frac{-k \times N_1}{L_w^2} + k \times P^{-1}. \tag{20}$$

Перейдемо до знаходження критичних точок функції $\varepsilon(L_w)$. Для цього, враховуючи (20), слід розв'язати наступне рівняння

$$\frac{d\varepsilon}{dL_w} = 0. \tag{21}$$

Після нижченаведених перетворень (22)...(24) отримаємо вираз (25) для розрахунку критичної точки функції $\varepsilon(L_w)$.

$$\frac{-k \times N_1}{L_w^2} + k \times P^{-1} = 0; \tag{22}$$

$$\frac{k}{P} = \frac{k \times N_1}{L_w^2} \ ; \tag{23}$$

$$L_w^2 = N_1 \times P \, . \tag{24}$$

$$L_w^0 = \sqrt{N_1 \times P} \, , \tag{25}$$

де L_w^0 – критична точка.

Перевірка (26), дозволяє твердити, що точка L_w^0 є точкою мінімуму функції $\varepsilon(L_w)$.

$$\begin{cases} \text{якщо } L_w < L_w^0 \quad \text{то} \quad \dfrac{d\varepsilon}{dL_w} > 0, \\ \text{якщо } L_w > L_w^0 \quad \text{то} \quad \dfrac{d\varepsilon}{dL_w} < 0. \end{cases} \tag{26}$$

Враховуючи (14) та (25) запишемо рівняння для приблизної оцінки оптимальної кількості синаптичних зв'язків багатошарового персептрону, що відповідає мінімуму ε_0

$$L_w^{opt} = \sqrt{N_1 \times P} \, . \tag{27}$$

Як видно, із показаної на Рис. 2, типової структури двохшарового персептрону призначеного для розпізнавання скриптового шкідливого програмного забезпечення, кількість його синаптичних зв'язків (L) розраховується так:

$$L_w = N_1 \times L + N_0 \times L \, , \tag{28}$$

де L – кількість нейронів в схованому шарі; N_1 – кількість вхідних нейронів; N_0 – кількість вихідних нейронів.

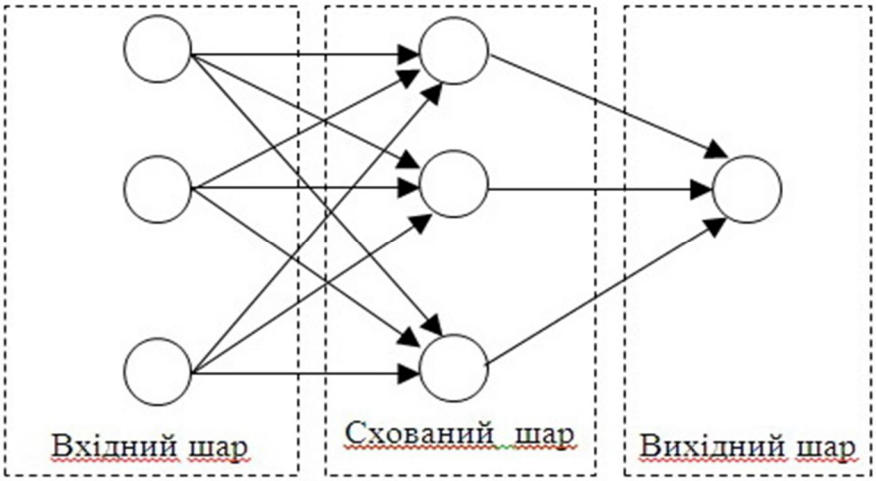

Рис. 2. Структура двохшарового персептрону

Відповідно (28), з врахуванням того, що N_0=1, кількість нейронів в схованому шарі двохшарового персептрону визначається так:

$$L = \frac{L_w}{N_1 + 1} \, . \tag{29}$$

Для розрахунку оптимальної кількості схованих нейронів слід в (29) підставити (27):

$$L^{opt} = \frac{\sqrt{N_1 \times P}}{N_1 + 1}, \tag{30}$$

де L^{opt} – оптимальна кількість схованих нейронів в двохшаровому персептроні.

При достатньо великих значення N_1 можна вважати, що

$$N_1 + 1 \approx N_1 \ . \tag{31}$$

Підстановка (31) в (30) дозволяє записати остаточний вираз для розрахунку оптимальної кількості схованих нейронів в двохшаровому персептроні:

$$L^{opt} \cong \sqrt{\frac{P}{N_1}} \ . \tag{32}$$

Зазначимо, що по відношенню до (6, 7) вираз (32) дозволяє більш точно визначити діапазон оптимальної кількості схованих нейронів в двохшаровому персептроні.

3. Висновки. Запропоновано підхід та розроблена методика оптимізації структури двохшарового персептрону, призначеного для розпізнавання скриптового шкідливого програмного забезпечення. Результатом методики є розрахункові вирази, які дозволяють на основі обсягів навчальної вибірки та вхідних параметрів, визначити оптимальну кількість синаптичних зв'язків та оптимальну кількість схованих нейронів.

Основні перспективи подальших розробок у даному напрямку полягають у розробці методів оптимізації структури синаптичних зв'язків двохшарового персептрону при визначеній оптимальній кількості схованих нейронів.

Література

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

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

3. Терейковський І. Нейронні мережі в засобах захисту комп'ютерної інформації / І. Терейковський. – Київ : ПоліграфКонсалтинг. – 2007. – 209 с.

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

5. Хайкин С. Нейронные сети : полный курс / Хайкин С. ; пер. с англ. Н. Н. Куссуль. – 2-е изд., испр. – Москва : Вильямс, 2006. – 1104 с.

Дата надходження в редакцію: 26.03.2015 р. Рецензент: д.т.н., проф. Г. М. Розорінов