УДК 621.391 160164

А.О. НІЧЕПОРУК
Хмельницький національний університет

ВИКОРИСТАННЯ НЕЧІТКОЇ КЛАСИФІКАЦІЇ ДЛЯ ВИЯВЛЕННЯ МЕТАМОРФНИХ ВІРУСІВ В КОРПОРАТИВНІЙ МЕРЕЖІ

В роботі обґрунтовано використання нечіткої класифікації для виявлення метаморфних вірусів у корпоративній мережі. Класифікатор визначення інфікування метаморфним вірусом комп'ютерної системи реалізовано на базі системи нечіткого висновку Мамдані. Розв'язана задача формування множини правил нечіткої системи та функцій приналежності.

Ключові слова: нечітка логіка, обфускація, метаморфний вірус, опкод, вектор схожості копій метаморфних вірусів.

A.O. NICHEPORUK
Khmelnytskyi National University

USE OF FUZZY CLASSIFICATION FOR DETECTION OF METAMORPHIC VIRUS IN THE CORPORATE NETWORK

The paper proved the use of fuzzy classification for detect of metamorphic viruses in a corporate network. Classifier definition of infection metamorphic virus computer system implemented on the basis of fuzzy inference system Mamdani. Solved the problem of forming the set of rules of the fuzzy system and membership functions.

Keywords: fuzzy logic, obfuscation, metamorphic virus, opcode, feature vector of similarity for metamorphic viruses' copies.

Вступ

Однією з головних проблем в галузі комп'ютерної безпеки є розмежування між нормальною та потенційно небезпечною поведінкою програм, що притаманна вірусам, зокрема класу метаморфних вірусів. Проте, отримання результату в чіткій формі, тобто значень вірус – корисна програма, є майже неможливим, у зв'язку з маскуванням вірусних програм під довірені додатки, для приховування своєї присутності в комп'ютерній системі.

У випадку метаморфних вірусів, під поведінкою будемо розуміти не послідовність функціональних дій, що виконують програми в комп'ютерній системі, а функціонування програми на нижчому рівні абстракції, тобто поведінка команд у дизасембльованому вигляді при порівнянні з іншим таким самим дизасембльованим файлом. Це зумовлено використанням у метаморфних вірусах обфускації програмного коду, що проявляється у таких техніках заплутування програмного коду як вставка "команд-сміття", використання еквівалентних інструкцій та перестановки команд, що дає змогу створювати абсолютно ідентичні копії метаморфного вірусу у плані функціонального навантаження та різні, з точки зору структури файлу.

Постановка завдання

Задача діагностування комп'ютерних систем на наявність метаморфних вірусів передбачає вирішення задачі класифікації. Результатом класифікації є віднесення невідомого об'єкту до класу метаморфних вірусів або до класу довірених додатків, тобто $X \to y \in \{C_1, C_2\}$, де X – невідома програма, C_1 – метаморфних вірус, C_2 – довірений додаток. В якості вхідних даних для класифікації виступає вектор ознак схожості копій метаморфних вірусів [1]. Проте отримання результату класифікації у чіткому вигляді, тобто $R^c = \{0,1\}$ є складним завданням, у зв'язку з використанням деякими додатками технік обфускації коду, для захисту від копірайту [2], і є лише частковим випадком задачі нечіткої класифікації, де $R^c = [0,1]$.

Таким чином, постає задача віднесення невідомої програми, що задана вектором ознак до одного з класів з деякою функцією належності відповідному класу, тобто провести нечітку класифікацію.

Основна частина

У роботі [1] запропоновано метод виявлення метаморфних вірусів у корпоративній мережі на основі модифікованих емуляторів. Зазначений метод передбачає отримання копій метаморфних вірусів, шляхом емуляції підозрілої програми на кожному хості, та їх порівняння з використанням метрики Дамерау-Левенштейна. Використання метрики Дамерау-Левенштейна зумовлено використанням метаморфними вірусами обфускації програмного коду, що проявляється у застосуванні вставок інструкцій, які не впливають на загальний алгоритм виконання вірусної програми ("команд-сміття"), перестановку інструкцій, видалення та заміну інструкцій.

З огляду на механізми створення метаморфними вірусами власних копій, що використовують техніки вставки, видалення та переміщення власних інструкцій, для пошуку схожості зразками коду до емуляції та після використовується дистанція Дамерау – Левенштейна.

Для прийняття рішення про інфікування метаморфним вірусом системи необхідно вирішувати задачу нечіткої класифікації.

З кожного хоста в корпоративній мережі надходить вектор ознак схожості копій метаморфних вірусів на сервер для здійснення його класифікації. Визначимо вектор ознак наступним чином:

$$\overline{S} = \langle dL, T, D, I, R, M \rangle, \tag{1}$$

де dL – відстань Дамерау – Левенштейна для функціонального блоку між програмами F_p та F_s;

T – кількість необхідних операцій обміну опкодів для перетворення блоку програми F_p у F_s ($F_p = F_s$);

D – кількість необхідних операцій видалення опкоду;

I – кількість необхідних операцій вставки опкоду;

R – кількість необхідних операцій заміни відповідних опкодів;

M – кількість співпадінь між опкодами в функціональному блоці програми F_p та F_s.

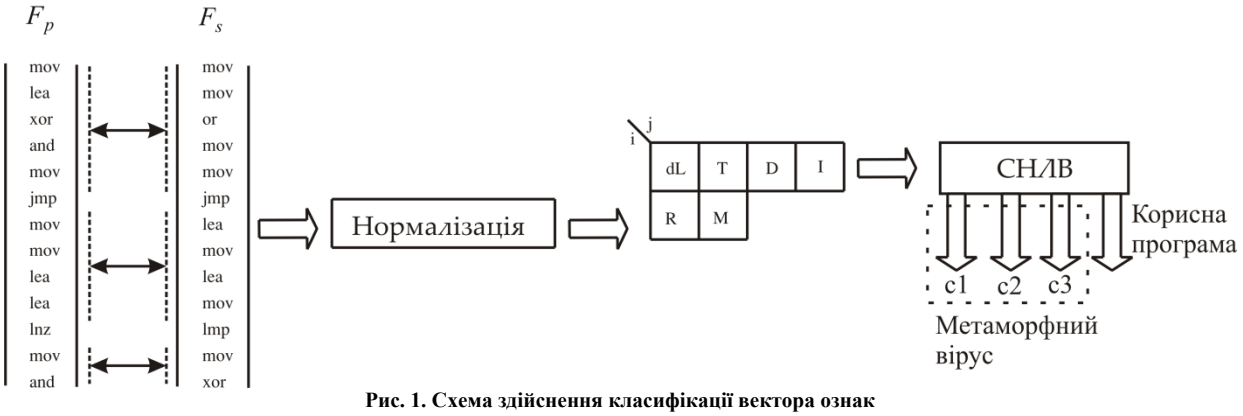

Рис. 1. Схема здійснення класифікації вектора ознак

Нечітка класифікація

В якості вхідних даних для нечіткої класифікації виступає вектор ознак схожості копій метаморфних вірусів, тобто $X = (x_{dL}, x_T, x_D, x_I, x_R, x_M)$. Вихідними даними для системи є класи рішень, що представляють собою множину класів метаморфних вірусів, $Y \in \{C_{NGVCK}, C_{VCL32}, C_{G2}, C_{BP}\}$, де $C_{NGVCK}, C_{VCL32}, C_{G2}$ – класи метаморфних вірусів, що відрізняються між собою інтенсивністю та складністю операцій вставки команд-сміття, переміщення інструкцій та використанням еквівалентних команд; C_{BP} – клас довірених додатків. Тоді нечітким класифікатором буде відображення виду:

$$X = (x_{dL}, x_T, x_D, x_I, x_R, x_M) \to Y \in \{C_{NGVCK}, C_{VCL32}, C_{G2}, C_{BP}\} \tag{2}$$

Згідно з [4] база правил для нечіткої класифікації задається:

$$R_j = if\,(t_{1j}\ and\ t_{2j}\ and\ ...\ and\ t_{nj}\ з\ вагою\ w_j) \to Y = d_j, \quad j = \overline{1,k}, n = 6 \tag{3}$$

де k – кількість правил у базі;

$d_j \in \{C_{NGVCK}, C_{VCL32}, C_{G2}, C_{BP}\}$ – результат j-го правила (тільки одне значення з множини);

$w_j \in [0,1]$ – ваговий коефіцієнт, що визначає вагу j-го правила $j = \overline{1,k}$;

t_{ij} – нечіткий терм з множини термів, що описує i-у ознаку вектора схожості копій метаморфних вірусів в j-му правилі, $j = \overline{1,k}$, $i = \overline{1,6}$

Для здійснення нечіткої класифікації обчислимо ступінь виконання j-го правила:

$$\mu_j(X) = w_j \cdot (\mu_j(x_{dL}) \wedge \mu_j(x_T) \wedge \mu_j(x_D) \wedge \mu_j(x_I) \wedge \mu_j(x_R) \wedge \mu_j(x_M)), \quad j = \overline{1,k} \tag{4}$$

де $\mu_j(x_i)$ – ступінь приналежності значення x_i нечіткому терму t_{ij};

Наступним етапом є розрахунок ступеня приналежності вхідного вектора X до кожного класу $C_{NGVCK}, C_{VCL32}, C_{G2}, C_{BP}$:

$$\mu_{\hat{N}_x}(Y) = \underset{\forall j: d_j = Cq}{agg}(\mu_j(X)), q = \overline{1,4} \tag{5}$$

де $\mu_{C_x}(Y)$ – степінь приналежності вхідного вектора класам, agg – агрегування результату нечіткого висновку по кожному правилу бази знань.

Результатом логічного $\hat{Y}$ – нечітка множина, що визначається:

$$\hat{Y} = (\frac{\mu_{C_{NGVCK}}}{C_{NGVCK}}, \frac{\mu_{C_{VCL32}}}{C_{VCL32}}, \frac{\mu_{C_{G2}}}{C_{G2}}, \frac{\mu_{C_{BP}}}{C_{BP}}) \tag{6}$$

Результатом класифікації – рішення з максимальною степеню приналежності в нечіткій множині *Y:*

$$Y = \arg\max((\mu_{C_{NGVCK}}(X), \mu_{C_{VCL32}}(X), \mu_{C_{G2}}(X), \mu_{C_{BP}}(X)) \quad (7)$$

У випадку приналежності невідомого об'єкту одразу до декількох класів, тобто коли у (5) результатом буде $Y = \max(\mu_{C_{NGVCK}}(X)) \vee \max(\mu_{C_{VCL32}}(X)) \vee \max(\mu_{C_{G2}}(X)) \vee \max(\mu_{C_{BP}}(X))$ вибирається один із результатів за медіаною.

Нормалізація вхідних даних

З метою уникнення впливу на відстань між двома найвіддаленішими точками класифікації використовується нормалізація даних [5]. Оскільки, даними для вхідного вектора схожості копій метаморфних вірусів є операції вставки, переміщення, видалення та заміни інструкцій, що можуть сильно варіюватися у метаморфному вірусі від покоління до покоління, для здійснення нечіткої класифікації необхідно виконати нормалізацію вхідних даних. Наприклад, для метаморфного вірусу NGVCK кількість операцій співпадіння інструкцій буде на порядок меншою за кількість операцій вставки (генерація "команд-сміття"). Тому, для кожної ознаки вектора схожості копій метаморфного вірусу $\overline{S}$ використаємо лінійну нормалізацію:

$$\underset{\forall s_i \in \overline{S},\ i=\overline{1,6}}{s(x)} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (8)$$

Нечітка класифікація копій метаморфних вірусів

Побудуємо нечіткий класифікатор для віднесення невідомої програми до одного з класів. Кожна невідома програма може бути віднесена до одного із чотирьох класів – трьох класів метаморфних вірусів та одного класу довірених додатків. Класи метаморфних вірусів представляються метаморфними генераторами NGVCK, VCL32 та G2, що отримані з VX Heavens [6]. Клас довірених додатків представляє собою корисні виконуванні файли, до яких застосована техніка обфускації.

У якості вхідних лінгвістичних змінних приймемо: «ступінь схожості підозрілої програми з її копією за дистанцією Левенштейна» (dL), «ступінь схожості підозрілої програми з її копією за кількістю операцій вставки» (I), «ступінь схожості підозрілої програми з її копією за кількістю операцій видалення» (D), «ступінь схожості підозрілої програми з її копією за кількістю операцій заміни» (R), «ступінь схожості підозрілої програми з її копією за кількістю операцій перестановки» (T), та «ступінь схожості підозрілої програми з її копією за кількістю операцій співпадіння» (M). Для кожної лінвістичної змінної задано терм-множину less low, low, medium, high та more high.

В якості функцій приналежності для входів було обрано трапецевидну, для виходу – трикутну. Наприклад, функції приналежності для лінгвістичної змінної dL можна описати рівняннями:

$$\mu_{low}(dL) = \begin{cases} 0, & 72 < x \\ \dfrac{72 - x}{64}, & 8 < x \le 72 \\ 1, & 0 \le x \le 8 \end{cases}$$

$$\mu_{medium}(dL) = \begin{cases} 0, & x < 16 \\ \dfrac{x - 16}{49}, & 16 \le x < 65 \\ 1, & 65 \le x < 96 \\ \dfrac{145 - x}{49}, & 96 \le x < 145 \\ 0, & 145 \le x \end{cases}$$

$$\mu_{high}(dL) = \begin{cases} 0, & 96 < x \\ \dfrac{x - 96}{38}, & 96 \le x < 134 \\ 1, & 134 \le x \le 161 \end{cases}$$

На рис. 2 представлено графік функції приналежності для лінгвістичної змінної дистанція Левенштейна. З рисунка видно, що для ознаки дистанція Левенштейна значення 119 (операцій вставки, видалення, перестановки інструкцій) степінь приналежності до нечіткого терму low становить 0,47, до medium – 0,28, до всіх решти нечітких термів – 0.

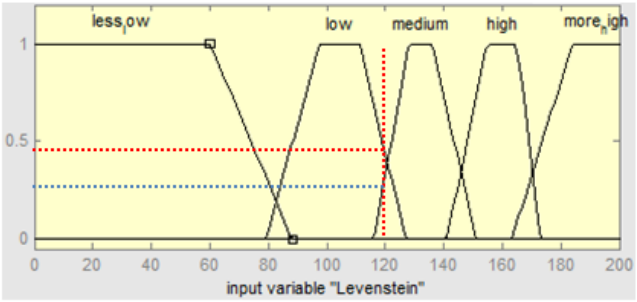

Рис. 2. Графік функції приналежності для ознаки – "дистанція Левенштейна"

Для здійснення нечіткого логічного висновку в системі задіяно 38 правил, приклад яких подано нижче:

1. if (dL is Low) and (T is Low) and (D is Medium) and (I is Hight) and (R is Low) and (M is Medium) then class1
2. if (dL is Low) and (T is Medium) and (D is Medium) and (I is Hight) and (R is Low) and (M is Medium) then class1

…

38 if (dL is Hight) and (T is Low) and (D is Hight) and (I is Hight) and (R is Medium) and (M is Low) then class3

Для навчання системи, використовуючи методику [1], експериментальним шляхом було отримано 300 векторів схожості копій метаморфних вірусів. Для отримання кожного вектора ознак було проаналізовано 200 рядків дизасембльованого коду копії метаморфного вірусу. Мінімальне та максимальне значення ознак для вектора схожості копій метаморфних вірусів для кожного класу представлено у таблиці 1.

Дані з таблиці 1 свідчать, що незважаючи на відмінність двох копій метаморфного вірусу одна від одної, що зумовлено використанням техніки обфускації програмного коду вірусу, загальна кількість операцій вставки "команд-сміття", перестановки інструкцій та використання еквівалентних команд залишається відносно сталою в певному діапазоні. Це пояснюється використанням всередині метаморфних генераторів алгоритмів для генерації нової копії вірусу, у яких закладені ті чи інші закони розподілу величин [7].

Таблиця 1

Тестові значення ознак вектора схожості копій метаморфних вірусів

Класи метаморфних вірусів	Ознаки вектора схожості копій метаморфних вірусів					
	Дистанція Дамерау-Левенштейна	Перестановка інструкцій	Видалення інструкцій	Вставка інструкцій	Заміна інструкцій	Співпадіння інструкцій
C_{NGVCK}	138-161	26-31	42-49	41-51	24-41	39-62
C_{VCL32}	120-149	17-22	37-41	29-39	37-47	51-80
C_{G2}	98-127	32-41	25-32	20-28	21-26	73-102

Таблиця 2

Значення нечітких терм множин для вхідних параметрів нечіткого класифікатора

№ п/п	Назва вхідного параметру	Позначення	Нечіткі терм-множини				
			less low	low	medium	high	more high
1	Дистанція Дамерау-Левенштейна	dL	0-88	88-126	127-138	139-161	162-200
2	Перестановка інструкцій	T	0-16	17-22	26-31	32-41	42-200
3	Видалення інструкцій	D	0-24	25-32	37-41	42-49	50-200
4	Вставка інструкцій	I	0-19	20-28	29-39	41-51	52-200
5	Заміна інструкцій	R	0-20	21-25	26-36	37-41	42-200
6	Співпадіння інструкцій	M	0-38	39-62	63-73	74-102	103-200

Для тестування нечіткого класифікатора задіяно 80 векторів ознак схожості копій метаморфних вірусів, які отримані з 60 метаморфних вірусів та 20 корисних програм.

Результати тестування подамо у наступних відношеннях:

$$TPR = \frac{TP}{TP + FN} = \frac{57}{57 + 3} = 0{,}95$$ вірно ідентифіковані копії метаморфних вірусів

$FPR = \frac{FP}{FP+TN} = \frac{2}{2+18} = 0{,}1$ корисна програма ідентифікована як метаморфний вірус

$AR = \frac{TP+TN}{TP+TN+FP+FN} \cdot 100\% = \frac{57+18}{57+18+2+3} \cdot 100\% = 0{,}9375\%$ загальний показник точності класифікації (accuracy rate)

$ER = 1 - \frac{TP+TN}{TP+TN+FP+FN} \cdot 100\% = 1 - 0{,}9375 = 0{,}0625\%$ загальний показник помилки класифікації

Таким чином, нечітким класифікатором вірно розпізнано 57 метаморфних копій та 18 корисних програм, 2 корисних програми віднесено до класу метаморфних вірусів та 3 метаморфних копії віднесено до класу довірених додатків.

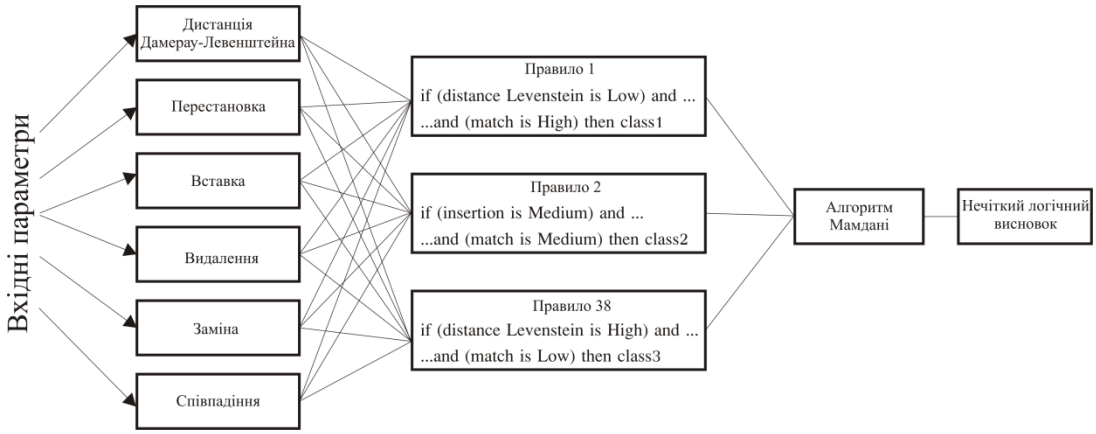

Рис. 3. Схема системи нечіткого логічного висновку для визначення ступеня приналежності кожної копії вірусу до одного з класу метаморфних вірусів

Висновки

Запропонований нечіткий класифікатор вектора схожості копій метаморфних вірусів на основі дистанції Дамерау-Левенштейна, дозволяє здійснювати виявлення метаморфних копії з точністю 94%. Закладені механізми для формування вектора ознак дозволяють здійснювати виявлення метаморфних вірусів, що створенні за допомогою генератора NGVCK, VCL32, G2. Дана система може бути масштабована шляхом доповнення нечіткої бази правилами для виявлення нового типу метаморфних вірусів.

Література

1. O. Pomorova, O. Savenko, S. Lysenko, A. Nicheporuk. Metamorphic virus detection technique based on the modified emulators: In proc. of the 12th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer, ICTERI. 2016, pp. 375–383.

2. S. Garg, C. Gentry, S. Halevi, M. Raykova, A. Sahai and B. Waters. Candidate indistinguishability obfuscation and functional encryption for all circuits. FOCS. 2013.

3. Штовба С.Д. Проектирование нечетких систем средствами MATLAB / Штовба С.Д. – М. : Горячая линия – Телеком, 2007. – 288 с.

4. A. Rotshtein, S. Shtovba. Identification of a nonlinear dependence by a fuzzy knowledgebase in the case of a fuzzy training set. Cybernetics and Systems Analysis. 2006. Vol. 42, № 2. pp.176– 182.

5. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP / [А. А. Берсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод]. – 2-е изд., перераб. и доп. – СПб : БХВ-Петербург, 2007. – 384 с.

6. VX Heavens Computer virus collection. URL: http://vx.netlux.org

7. B.B. Rad, M. Masrom. Metamorphic Virus Variants Classification Using Opcode Frequency Histogram. In 14th WSEAS international conference of computers, WSEAS. 2010, pp. 147–155.

Рецензія/Peer review : 22.8.2016 р. Надрукована/Printed : 25.8.2016 р.
Рецензент: д.т.н. Мартинюк В. В.