УДК 519.6

В.М.ПОШТАРЕНКО, канд.техн.наук, НТУ «ХПІ»;
А.Ю.ВАРЛИГІНА, НТУ «ХПІ»;
В.С.КИТАЙНИК, НТУ «ХПІ»

АНАЛІЗ ДОЦІЛЬНОСТІ ВИКОРИСТАННЯ ШТУЧНИХ ІМУННИХ СИСТЕМ У ЗАДАЧАХ ВИЯВЛЕННЯ АНОМАЛІЙ

Представлено імітаційну модель виявлення аномалії в комп'ютерних мережах на основі штучних імунних систем.

Presented the simulation model of the anomaly detection in the computer networks based on artificial immune systems.

Постановка проблеми. В останні роки все більше технологій, що використовуються людиною для створення систем підтримки прийняття рішень, робототехніці, інтелектуальному аналізі даних тощо, були запозичені у природи. Генетичні алгоритми [1] та нейронній мережі [1, 2] вже набули сьогодні великої популярності і фігурують не тільки у теоретичних дослідженнях, а і практично запроваджуються для вирішення широкого кола питань. Штучні ж імунні системи вивчені недостатньо, хоча їх структура і алгоритми функціонування також можуть «стати у пригоді» для вирішення проблем виявлення аномалій, розпізнаванні образів, побудові систем захисту тощо[3, 4]. Штучні імунні системи – це перспективний напрямок досліджень, адже особливості організації і функціонування цієї системи можуть бути використані у областях, де генетичні алгоритми і нейронні мережі можуть програвати за деякими характеристиками.

Аналіз літератури. На сьогоднішній день існує ряд публікацій [3, 4 тощо], в яких досить детально аналізуються штучні імунні системи, їх властивості та області застосування, розглядаються різні підходи до реалізації штучних імунних систем, виконується порівняння штучних імунних систем з іншими техніками, що були запозичені у природи. Зокрема у [3] виконано порівняння імунних та нервових систем людини, [6, 8] розглядають схожі та відмінні властивості штучних імунних систем та генетичних алгоритмів, але ці порівняння не є досить повними, тому у рамках даної статті вони будуть розширені. У [7] детально розглядаються алгоритми моделювання роботи імунних систем для вирішення задач виявлення аномалій. Джерела [1, 2] надають вичерпну інформацію щодо генетичних алгоритмів та нейронних мереж відповідно, але жодне з них навіть не згадує про штучні імунні системи, їх властивості і переваги.

Метою статті є розробка імітаційної моделі виявлення аномалій на основі імунних мереж.

Обґрунтування доцільності застосування штучних імунних систем у задачах виявлення аномалій.

Головним принципом дії людської імунної системи є порівняння певних «шаблонів» з тілами, що перебувають усередині організму, і виявлення, таким чином, сторонніх предметів, які отриману назву антигенів.

Роль згаданих шаблонів виконують лімфоцити, які постійно генеруються спинним мозком і тимусом з урахуванням інформації, що втримується в ДНК (така інформація увесь час накопичується, процес цей називається еволюцією генної бібліотеки), і поширюються організмом через лімфатичні вузли, причому кожний тип лімфоцита відповідає за виявлення якогось обмеженого числа антигенів. При генеруванні лімфоцитів є одна дуже важлива стадія, названа негативною селекцією, на якій відбувається своєрідний тест на відповідність рідним клітинам організму. Іншими словами, завдяки негативній селекції створюються «шаблони», що містять ту інформацію, що усередині організму відсутня, і якщо якесь тіло підходить під даний шаблон, виходить, воно чуже. У випадку виявлення лімфоцитами антигену на підставі відповідного шаблона виробляються антитіла, які й знищують його. Потім активується ще один процес – клональна селекція [3, 7], під час якої відбувається своєрідний природний відбір антитіл: виживають лише ті, що максимально підходять під знайдений антиген. При цьому відомості про створені антитіла «заносяться» у згадувану вище генну бібліотеку.

Таким чином можна дійти висновку, що дана схема може бути корисною для забезпечення безпеки комп'ютерних систем, адже у даному випадку розподілена, гнучка, самоорганізована штучна імунна система обумовлює її максимальну ефективність у системах виявлення вторгнень [4].

Система виявлення вторгнень для одного сегмента мережі, побудована на принципах штучної імунної системи, підрозділяється на основну й набір вторинних мереж. Основна є аналогом спинного мозку, а вторинні – аналогами лімфатичних вузлів. В основній системі виявлення вторгнень на базі штучної імунної системи імітуються два процеси [3, 7] – еволюція генної бібліотеки й негативна селекція.

На етапі еволюції генної бібліотеки відбувається нагромадження інформації про характер аномалій мережевого трафіку. Генна бібліотека штучної імунної системи повинна містити «гени» (це можуть бути, наприклад, дані про характерну кількість пакетів, їхній довжині, структурі, типових помилках і т.д.), на підставі яких будуть генеруватися особливі програмні агенти-детектори, що є аналогами лімфоцитів [8]. Початкові дані для формування генної бібліотеки вибираються, виходячи з особливостей застосовуваних мережних протоколів, зокрема їх слабких з погляду захисту місць. Надалі, при

виявленні детекторами аномальної активності в мережі до бібліотеки будуть додаватися відповідним цим проявам нові «гени» [3]. Через обмеженість розміру генної бібліотеки, у ній зберігаються тільки «гени», що проявляють себе найбільшу кількість разів.

На другому етапі шляхом довільного комбінування «генів» відбувається генерування так званих пре-детекторів (аналоги «зародкових» лімфоцитів), які потім за допомогою механізму тої самої негативної селекції перевіряються на, на несумісність з нормальним мережним трафіком. При цьому використовуються дані про характер такого трафіку (профілі), формовані так званим автоматичним профайлером, що постійно аналізує потік даних, який надходить від маршрутизатора, що стоїть на вході в мережний сегмент [8]. Кінцевою метою в цьому випадку є створення обмеженого набору детекторів, за допомогою якого можна було б виявити максимальну кількість мережевих аномалій. Цей набір розсилається по всіх вузлах мережі, створюючи вторинну систему виявлення вторгнень [6]. Варто відзначити, що розроблені на сьогоднішній день алгоритми негативної селекції оперують імовірнісними характеристиками – замість точної відповідності використовується часткова, ступінь якої може досить варіюватися. Її зміна в остаточному підсумку повинна призвести до зменшення або збільшення частоти «помилкових спрацьовувань».

При виявленні аномалії відбувається клональна селекція [3, 7] – відповідний їй детектор «розмножується» і розсилається на всі вузли. Остаточне ж рішення про те, відбувається вторгнення в чи мережу ні, приймається на підставі даних від декількох вузлів. Кожний вузол, а також основна система виявлення вторгнень постачені ще одним компонентом – комунікатором, що, зокрема, оперує таким параметром, як рівень ризику. У випадку, якщо на якімсь вузлі помічена підозріла активність, комунікатор підіймає свій рівень ризику й відсилає відповідне повідомлення комунікаторам інших вузлів і основній системі виявлення вторгнень, і ті також піднімають свої рівні ризику. З появою аномалій відразу на декількох вузлах протягом короткого проміжку часу цей рівень дуже швидко росте, і якщо буде досягнутий заданий поріг, адміністратор мережі одержить сигнал тривоги.

Імітаційна модель імунного алгоритму виявлення аномалій.

Нормальна поведінка системи часто характеризується дискретними тимчасовими рядами спостережень. В цьому випадку проблему виявлення аномалій можна сформулювати як задачу знаходження неприпустимих відхилень у характеристиках системи. Для виявлення аномалій і помилок в мережах відомі різноманітні методи, такі як контрольні карти, методи моделювання, використання експертних систем, розпізнання образів й кластеризацію, марковські моделі і нейронні мережі [9]. Проте для більшості методів вимагається наявність апріорної інформації про різноманітні умови виник-

нення аномалій або точна теоретична модель моніторінгуєвих системи. У роботі розглядається можливість виявлення аномалій шляхом імітаційного моделювання.

Завдання виявлення аномалій може бути зведено до задачі виявлення змін у паттерні нормальної активності. При підготовці даних беремо до уваги зміни форми подання даних при збереженні їхнього інформаційного змісту. Будь-які зміни, що перевершують допустимі варіації паттернів даних, повинні бути повністю представлені в новій формі. Це може викликати труднощі, якщо необхідно виявляти дуже малі зміни в потоці реальних даних [3]. Для цього аналогова величина спочатку нормується по відношенню до певної фіксованої варіації, що дозволяє визначити інтервал, до якого вона відноситься, і після цього належність інтервалу кодується у бінарній формі. Проте якщо величина виявляється за межами інтервалу (*MIN, MAX*), то вона повинна кодуватися всіма нулями або всіма одиницями, залежно від границі інтервалу, за яку вона вийшла. Тоді, якщо кожний набір даних кодується m бінарними числами (величина m обирається залежно від необхідної точності), поміж максимальним значенням *MAX* й мінімальним *MIN* існує $2^m - 2$ інтервалів що розрізнюються. Відповідно, розмір інтервалу d рівний $(MAX - MIN)/(2^m - 2)$. Отже, для аналогової величини x $MIN \le x \le MAX,$ де $MAX = MIN + (2^m - 2)\, d$, і вона може бути відведена до певного інтервалу (з абсолютною помилкою по амплітуді d) та закодована бінарним числом по номеру цього інтервалу. Наприклад, якщо амплітуда x така, що $MIN + n_a d \le x \le MIN + (n_a + 1)\, d$, тоді вона кодується бінарним рядком, відповідній номеру інтервалу n_a (де n_a може мінятися в межах від 1 до $2^m - 2$).

Перерахуємо основні параметри процесу підготовки даних.

m – параметр, визначаючий точність, із якої дані представляються в бінарній формі. Наприклад, 5-бітне кодування дозволяє розподілити дані по 30 інтервалам у діапазоні [*MIN, MAX*].

w – число елементів, що кодується у кожному паттерні (свого рядка).

SHIFT – число елементів, на яке даний паттерн зміщений по відношенню до попереднього. Наприклад, якщо *SHIFT* = 1, а розмір вікна рівний w, то паттерни будуть мати наступний вигляд: $\{X_1, X_2, ..., X_w\}$, $\{X_2, X_2, ..., X_{(w+1)}\}$ й т.п.

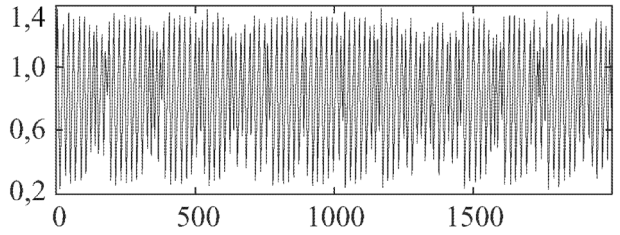
В ході цих експериментів використовувались послідовності, що генеруються рівнянням Макея-Гласса, що моделювали результати вимірів.

В першому тимчасові ряді одержувалися за допомогою послідовностей Макея-Гласса і включали експерименти щодо спостереження за динамікою процесу в різноманітних умовах. Рішення цього рівняння використовувались в багатьох галузях науки. Рівняння має наступний вигляд:

$$\frac{dx}{dt} = \frac{ax(t-m)}{[1+x^c(t-m)]} - bx(t).$$

Використовувались наступні значення параметрів: $a = 0{,}2$; $b = 0{,}1$;

$c = 10$, величина параметру запізнювання m визначає складність вигляду рішення рівняння. Рішення цього рівняння використовуються як завдання для короткострокового прогнозування по результатах наявних вимірів. Окрім цього, може бути поставлена задача виявлення змін величини параметру m. Для отримання нормальних послідовностей брались значення $m = 30$, для невеликих змін – $m = 27$, для одержання значних аномалій використовується $m=17$. Послідовності дістаються чисельним вирішенням рівняння методом Рунге-Кутта 4-го порядку. Інтервал вибірки на кожному кроку інтегрування обмежений величиною. Для усунення впливу початкових умов, пропускаються перші 1000 вимірів від початкового значення. В моделі використовували 2000 вимірів тимчасового ряду Макея-Гласса (див. рисунок).



Рішення рівняння Макея-Гласса.

Результати застосування моделі імунної системи для виявлення невеликих змін у послідовностях Макея-Гласса. Результати аналізу наведені у таблиці.

Параметри кодування	R	Nr	Виявлення аномалій	
			Середнє	Надійність
Win size = 4 Win shift = 4 Self length, l = 20 Self size, Ns = 500	9	30	11,34 (3,52)	100 %
		40	14,96 (3,78)	100 %
		50	16,99 (3,20)	100 %
	10	30	4,78 (2,77)	98 %
		50	7,82 (3,16)	100 %
		70	11,48 (3,51)	100 %
Win size = 4 Win shift = 4 Self length, l = 20 Self size, Ns = 500	10	20	8,08 (3,30)	100 %
		30	12,12 (4,08)	100 %
		40	14,58 (4,82)	100 %
	12	30	5,80 (2,70)	96 %
		50	9,38 (2,99)	100 %
		70	14,80 (4,82)	100 %

Тут r — граничне значення критерію подібності, Nr — число детекто-

рів. В колонках 4 і 5 наведені середня за 50 прогонів кількість виявлень в успішних прогонах (середнє і стандартне відхилення) і надійності виявлення відповідно.

Відзначимо, що незначні зміни виявляються майже так ж ефективно, як і значні, бо в закодованому вигляді число змінених рядків в обидвох випадках приблизно однаково (в інтервалі від 1000 до 1500). Виявлено, що ефективність алгоритму змінюється залежно від величини порога подібності r для даної довжини рядка l і параметрів кодування. При збільшенні r детектори стають дошкульними до будь-яких змін в даних, так що для отримання необхідного рівня надійності необхідно використовувати більше число детекторів. З іншого боку, якщо r занадто маленьке, то створення достатнього набору детекторів по даному може не вийти, бо при обраному значенні r не існує рядків , що відрізняються.

Висновки

Шляхом імітаційного моделювання встановлено, що для оптимізації роботи алгоритму виявлення аномалій потрібен вибір схожої величини параметру r. Зокрема, величина параметру r може використовуватися при настройці співвідношення надійності виявлення та ризику помилкової позитивної відповіді.

Список літератури: 1. *Munakata T.* Fundamentals of the New Artificial Intelligence. – London: Springer, 2008. – 256 p. **2.** *Хайкин С.* Нейронные сети: полный курс. – М.: Издательский дом «Вильямс», 2006. – 1104с. **3.** *Дасгупта Д.* Искусственные иммунные системы и их применение. – М.: ФИЗМАТЛИТ, 2006. – 344 с. **4.** *Гвозденко А.* Искусственные иммунные системы как средство сетевой самозащиты. – http://itc.ua/node/4270. **5.** *Castro L., Timmis J.* Artifical Immune Systems: A New Computational Intelligence Approach. – London: Springer, 2002. – 364 p. **6.** *Burke K.Edmund, Kendall Graham.* Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques. – New York: Springer, 2006. – 626 p. **7.** *Gonzalez F.* A study of artificial immune systems applied to anomaly detection. – Memphis: Memphis University Edition, 2003. – 184 p. **8.** *Aickelin U.* Artificial Immune Systems – A New Paradigm for Heuristic Decision Making – Nottingham: Nottingham University Edition, 2004. – 200 p. **9.** *Hunt J.E., Cooke D.E.* An adaptive, distributed learning system, based on the immune system**. –** Wales**:** Wales University Edition, 1995. – 2540 p.

Надійшла до редколегії 02.04.2009.