УДК 539.3
Каганюк О.К. к.т.н. доц., Бортник К.Я.к.т.н. доц., Свиридюк В.В.
Луцький національний технічний університет

АНАЛІЗ АНОМАЛЬНИХ СТАНІВ ТРАФІКА КОМП'ЮТЕРНОЇ МЕРЕЖІ НА БАЗІ НЕЙРОМЕРЕЖ

Каганюк О.К., Бортник К.Я., Свиридюк В.В. Аналіз аномальних станів трафіка комп'ютерної мережі на базі нейромереж. У статті запропонований метод обробки даних перед навчанням нейронної мережі для ідентифікації аномальних станів КС.Задача дослідження аномалій трафіку є однією з найуспішніших застосувань нейронних мереж для рішення задач класифікації.
Ключові слова: метод, аномальні стани, нейромережа, трафік, ідентифікація.

Каганюк О.К., Бортник К.Я., Свиридюк В.В. Анализ аномальных состояний трафика компьютерной сети на базе нейронных сетей. В статье предложен метод обработки данных перед обучением нейронной сети для идентификации аномальных состояний КС. Задача исследования аномалий трафика является одной из самых успешных применений нейронных сетей для решения задач классификации.
Ключевые слова: метод, аномальные состояния, нейросеть, трафик, идентификация.

Kahanyuk O., Bortnyk K., Sviridyuk V. Analysis of abnormal states of computer network traffic based on neural networks. The paper proposed a method of processing data before training a neural network to identify abnormal states KS.Zadacha traffic anomalies research is one of the most successful applications of neural networks for solving classification problems.
Keywords: method, abnormal conditions, neural network, traffic identification.

Постановка наукової проблеми. Виявлення мережевих атак є в даний момент є однією з найбільш гострих проблем мережевих технологій. Однією з актуальних наукових завдань в даний час є аналіз (і подальше прогнозування) самоподібної структури трафіку в сучасних мультисервісних мережах. Для вирішення цього завдання необхідний збір і подальший аналіз різноманітної статистики в діючих мережах. [2,4].

Метою та завданням даної роботи є ідентифікація аномальних станів комп'ютерних систем на основі паралельної обробки трафіка КС, використовуючи колектив нейромереж.

Об'єктом дослідження є інтелектуальні технології, базовані на нейромережевих принципах, які орієнтовані на розв'язок прикладних задач.

Предметом дослідження є методи та моделі ідентифікації аномальних станів трафіка КМ комітетом нейромереж.

Виклад основного матеріалу й обґрунтування отриманих результатів дослідження. Більшість сучасних методів виявлення атак використовують деяку форму аналізу контрольованого простору на основі правил або статистичного підходу. В якості контрольованого простору можуть виступати журнали реєстрації або мережевий трафік. Цей аналіз спирається на набір заздалегідь визначених правил, які створюються адміністратором або самою системою виявлення атак.

Будь-який поділ атаки або в часі, або серед кількох зловмисників є важким для виявлення за допомогою експертних систем. За великої різноманітності атак і хакерів навіть спеціальні постійні оновлення бази даних експертної системи ніколи не дадуть гарантій точної ідентифікації всього діапазону атак.

Використання нейронних мереж є одним із способів подолання зазначених проблем експертних систем. На відміну від експертних систем, які можуть дати користувачеві певну відповідь, відповідають чи ні аналізовані характеристики закладеним в базу даних правилам, нейронна мережа проводить аналіз інформації та надає можливість оцінити, чи узгоджуються дані з характеристиками, які вона навчена розпізнавати. У той час як ступінь відповідності нейромережевого подання може досягати 100 %, достовірність вибору повністю залежить від якості системи в аналізі прикладів наданої задачі.

Спочатку нейромережу навчають правильної ідентифікації та попередньо підібраною вибіркою прикладів предметної області. Реакція нейромережі аналізується, і система налаштовується таким чином, щоб досягти задовільних результатів. На додаток до початкового періоду навчання нейромережа набирається також досвіду з протягом часу, у міру того, як вона проводить аналіз даних, пов'язаних з предметною областю.

Важливою перевагою нейронних мереж при виявленні зловживань є їх здатність «вивчати » характеристики умисних атак та ідентифікувати елементи, які не схожі на ті, що спостерігалися в мережі раніше.

Мережі з прямим зв'язком є універсальним засобом апроксимації функцій, що дозволяє їх використовувати у вирішенні задач класифікації. Як правило, нейронні мережі виявляються найбільш ефективним способом класифікації, тому що генерують фактично велике число регресійних моделей (які використовуються у вирішенні задач класифікації статистичними методами).

Багаті можливості відображення особливо важливі в тих випадках, коли на основі кількох оцінок будується високорівнева процедура прийняття рішень. Відомо багато додатків нейронних мереж з прямим зв'язком до завдань класифікації. Як правило, вони виявляються ефективніше інших методів, тому що нейронна мережа генерує нескінченне число нелінійних регресійних моделей.

На жаль, у застосуванні нейронних мереж у практичних завданнях виникає ряд проблем. По-перше, заздалегідь не відомо, якої складності (розміру) може знадобитися мережа для досить точної реалізації відображення. Ця складність може виявитися надмірно високою, що потребує складної архітектури мереж. Так доведено, що найпростіші одношарові нейронні мережі здатні вирішувати тільки лінійно роздільні завдання. Це обмеження можна подолати при використанні багатошарових нейронних мереж. У загальному вигляді можна сказати, що в мережі з одним прихованим шаром, вектор, відповідний вхідному зразку, перетворюється в прихований шар в деякий новий простір, який може мати іншу розмірність, а потім гіперплощини, відповідні нейронам вихідного шару, поділяють його на класи. Таким чином мережа розпізнає не тільки характеристики вихідних даних, але і характеристики характеристик, сформовані прихованим шаром.

Все це підкреслює важливість етапу попередньої обробки даних. Чим більш компактно представлені характеристики зразків, тим менше залежність від настроюваних параметрів мережі (0 або 1).

Підвищення якості навчання нейронної мережі можливе при використанні ефективних методів попередньої обробки даних. Для обробки даних перед навчанням нейронної мережі пропонується використовувати метод, заснований на понятті профілю компактності і комбінаторних формулах для ефективного обчислення функціонала ковзкого контролю. Метод застосовується для підготовки даних в задачах класифікації [3]. Щоб звести своє завдання попередньої обробки даних у моделюванні до задачі попередньої обробки даних при класифікації даних (яка має рішення, що використовує профіль компактності), пропонується провести кластерний аналіз на вихідних параметрах нейронної мережі. Таким чином, будемо мати задачу класифікації, для якої відоме рішення попередньої обробки даних.

Метод, починаючи з повної вибірки, послідовно виключає об'єкти. На кожному кроці вибирається той об'єкт, виключення якого мінімізує функціонал. Виявляється, що процес відсіву об'єктів розбивається на дві стадії. Спочатку виключаються шумові, потім виключаються неінформативні периферійні об'єкти.

Процес зупиняється, коли залишаються об'єкти, виключення яких помітно збільшує функціонал, тоді в масиві даних залишаються опорні об'єкти.

Основним результатом застосування комбінаторної формули для оцінки функціоналу повного ковзкого контролю є те, що вона однаково добре підходить як для виключення шумових об'єктів, так і для скорочення множини прецедентів, будучи при цьому ефективно обчислюваним, точним значенням функціоналу.

Метод спирається на припущення, яке називається гіпотезою компактності: схожі об'єкти набагато частіше лежать в одному класі, ніж в різних. У цьому випадку межа між класами має досить просту форму, а класи утворюють компактно локалізовані області в просторі об'єктів (у математичному аналізі компактними називаються обмежені замкнуті множини, гіпотеза компактності не має нічого спільного з цим поняттям).

Як правило, об'єкти навчання не є рівноцінними. Серед них можуть знаходитися типові представники класів - еталони. Якщо класифікується об'єкт близький до ідеалу, то, швидше за все, він належить тому ж класу. Ще одна категорія об'єктів - неінформативні, або периферійні. Вони щільно оточені іншими об'єктами того ж класу. Якщо їх видалити з вибірки, це практично не позначиться на якості навчання. Нарешті, у вибірку може потрапити деяка кількість шумових викидів - об'єктів, що знаходяться в чужому класі. Зазвичай їх видалення тільки покращує якість класифікації.

Виключення з вибірки шумових і неінформативних об'єктів дає кілька переваг одночасно: підвищується якість класифікації, скорочується обсяг збережених даних і зменшується час класифікації, що витрачається на пошук найближчих еталонів [5].

Перейдемо до розгляду функціоналу вибірки, що мінімізується. Нехай є множина об'єктів *X* і множина імен класів *Y*. Задана навчальна вибірка пар «об'єкт-відповідь»:

$$x^{im} = \{(x_1, y_1),...,(x_m, x_m)\} \epsilon\ x \times y$$

Нехай на множині об'єктів задана функція відстані p(*x, x '*). Ця функція повинна бути досить адекватною моделлю подібності об'єктів. Чим менше значення цієї функції, тим більше схожі об'єкти *x, x '*.

Для довільного об'єкта u розташуємо об'єкти навчальної вибірки x_i в порядку зростання відстаней до *u*:

$$p(u, x_{1u}) \le p(u, x_{2u}) \le ... \le p(u, x_{mu}),$$

де через x_{iu} позначається елемент навчальної вибірки, який є *i*-м сусідом об'єкта *u*. Аналогічне позначення введемо і для відповіді на *i*-му сусіді – y_{iu}.

Кожен об'єкт $u \in X$ породжує свою перенумерацію вибірки.

Розглядається метод найближчого сусіда, який відносить об'єкт *u*, що класифікується, до того класу, якому належить найближчий до *u* об'єкт навчальної вибірки: $a(u, x^m) = y_{1u}$.

Профіль компактності вибірки X^m є функція:

$$R(j, x^m) = \frac{1}{m} \sum_{i=1}^{m} [y_i \ne y_{ix}].$$

Іншими словами, профіль компактності *R (j)* - це частка об'єктів вибірки, для яких *j*-й сусід лежить в іншому класі.

Профіль компактності є формальним виразом гіпотези компактності - припущення про те, що схожі об'єкти набагато частіше лежать в одному класі, ніж в різних. Вибірка X^l розбивається всілякими способами на дві непересічні підвибірки:

$$x^l = x_n^m \text{ Y } x_n^k,$$

де x_n^m - навчальна підвибірка довжини m;

x_n^k - контрольна підвибірка довжини k;

k = L - m, n = 1,..., N - номер розбиття.

Для кожного розбиття *n* будується алгоритм $a_n(u, x)_n^m$. Функціонал повного ковзкого контролю (*complete cross-validation, CCV*) визначається як середня (по всіх розбиттям) помилка на контролі:

$$CCV(x^L) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{k} \sum_{x_z \in x_n^z} [a_n(x_i, x_n^m) \ne y_i]$$

Функціонал повного ковзкого контролю характеризує узагальнюючу здатність методу найближчого сусіда.

Справедлива формула для ефективного обчислення *CCV* через профіль компактності:

$$CCV(x^L) = \sum_{j=1}^{k} R(j, x^L)\, \Gamma(j),$$

де $\Gamma(j) = \frac{C_{L-1-j}^{m-1}}{C_{L-1}^{m}}$

Комбінаторний множник *Г(j)* швидко убуває із зростанням *j*. Для мінімізації функціоналу *CCV* достатньо, щоб при малих *j* профіль $R(j, X^L)$ брав значення, близькі до нуля. Це означає, що близькі об'єкти повинні лежати переважно в одному класі. Таким чином, профіль дійсно є формальним виразом гіпотези компактності [6].

Пропонується використовувати кластерний аналіз для розділення значень виходів мережі на групи, щоб звести задачу попередньої обробки даних при моделюванні за допомогою нейронної мережі до задачі попередньої обробки даних при класифікації, в якій використовується теорія профілю компактності.

В якості характеристики близькості вихідних значень нейронної мережі взято евклідову відстань між точками. Для довільного вектора *v* з числом елементів *n* евклідова норма знаходиться наступним чином:

$$\|v\| = \sqrt{\sum_{L=1}^{m} |v_l|^2} .$$

Евклідова відстань є найпопулярнішою метрикою в кластерному аналізі: вона відповідає інтуїтивним уявленням про близькість і, крім того, дуже вдало вписується своєю квадратичною формою у традиційно статистичні конструкції. Геометрично вона найкраще об'єднує об'єкти в кулястих скупченнях, які дуже типові для слабо корельованих сукупностей.

На першому кроці кластерного аналізу кожен об'єкт вважається окремим кластером. На наступному кроці об'єднуються два найближчих об'єкта, які утворюють новий клас, визначаються відстані від цього класу до всіх інших об'єктів, і розмірність матриці відстаней скорочується на одиницю. Процедура повторюється на поточній матриці відстаней, поки не буде досягнуто деяке число кластерів.

Висновки та перспективи подальшого дослідження. Таким чином, пропонований метод попередньої обробки даних дає більш якісне навчання нейронної мережі. Попередня обробка полягає у видаленні з масиву суперечливих прикладів. Пошук таких прикладів заснований на теорії профілю компактності в задачі класифікації. Щоб використовувати відомі рішення (теорію профілю компактності) в задачі моделювання, необхідно за допомогою кластерного аналізу виділити групи над значеннями вихідних параметрів нейронної мережі [11]. Чим краще буде зроблена попередня обробка, тим легше буде вирішена задача по виявленню аномальних станів трафіка комп'ютерної мережі.

1. Тимощук П.В., Лобур М.В.. Основи теорії проектування нейронних мереж // Навчальний посібник, Львів, 2007, с. 328
2. Ажмухамедов И.М., Марьенков А.Н. Обеспечение информационной безопасности компьютерных сетей на основе анализа сетевого трафика //№ 1 / 2011
3. Астахов А., Актуальные вопросы выявления сетевых атак, http://www.jetinfo.ru/2002/3/1/article1.3.2002.html
4. Беляев А., Петренко С. Системы обнаружения аномалий: новые идеи в защите информации // Экспресс-Электроника №2,2004 г. - С. 86 - 96.
5. Введение в сетевые атаки, http://www.tshram.com/hacker/net_attacks.shtml#1
6. Воронцов К.В. Комбинаторные оценки качества обучения по прецедентам // Докл. РАН. – 2004. – Т. 394, №2. – С. 175–178
7. Потемкин В.Г. Справочник по MATLAB, http://matlab.exponenta.ru/ml/book2/index.php
8. Дюк В., Самойленко А. Data Mining: учебный курс, СПб: Питер, 2001.
9. Емельянова Ю. Г., Талалаев А. А., Тищенко И. П., Фраленко В. П. Нейросетевая технология обнаружения сетевых атак на информационные ресурсы // Программные системы: теория и приложения : электрон. научн. журн. 2011. № 3(7), С. 3–15.
10. Жульков Е. Поиск уязвимостей в современных системах IDS// Открытые системы. СУБД -2003. - N 7/8. С. 37 - 42.
11. Качановский Ю.П., Коротков Е.А. ПРЕДОБРАБОТКА ДАННЫХ ДЛЯ ОБУЧЕНИЯ НЕЙРОННОЙ СЕТИ // Фундаментальные исследования. – 2011. – № 12 (часть 1). – стр. 117-120.