УДК 004.032.26

Вестник УрФО № 3(41) / 2021, с. 5–11

Кротова Е.Л., Уразбаева Ю.В.

DOI: 10.14529/secur210301

НЕЙРОННЫЕ СЕТИ ДЛЯ ОБНАРУЖЕНИЯ ЗЛОУМЫШЛЕННИКА В РАСПРЕДЕЛЕННОЙ АВТОМАТИЗИРОВАННОЙ СИСТЕМЕ

В работе проведено исследование расширения стандартной модели поиска злоумышленника в распределённой автоматизированной системе. Предложено дополнить систему, построенную на эвристических правилах, моделью, оценивающей отклонение поведения легального пользователя от профиля его стандартного поведения в сети. Данный метод реализован на основе перцептрона. Работа проиллюстрирована примерами реализации разработанной методики.

Ключевые слова: *компьютерная безопасность, статистическое моделирование, нейронные сети.*

Krotova E.L., Urazbaeva Yu.V.

NEURAL NETWORKS FOR DETECTING AN INTRUDER IN A DISTRIBUTED AUTOMATED SYSTEM

The paper investigates the extension of the standard model for searching for an intruder in a distributed automated system. It is proposed to supplement the system based on heuristic rules with a model that estimates the deviation of the behavior of a legal user from the profile of his standard behavior in the network. This method is implemented on the basis of a perceptron. The work is illustrated with examples of the implementation of the developed methodology.

Keywords: *computer security, statistical modeling, neural networks.*

Одна из основных направлений исследований по информационной безопасности – задача наискорейшего обнаружения злоумышленников в информационной системе и сведения к минимуму ошибок применяемых методов.

Традиционным методом обнаружения злоумышленника является использование алгоритмов на основе эвристических правил, позволяющих отличить злоумышленника от легального пользователя. Тем не менее, в настоящее время этого метода уже недостаточно для решения проблемы.

Альтернативные способы включают в себя изучение моделей активности пользователя с помощью статистических методов и их классификацию по заданному критерию. Актуальность применения статистических методов заключается в том, что, по сравнению с эвристическими методами, статистические методы позволяют в большей степени избежать ошибок при мониторинге активности пользователя, что, в свою очередь, усиливает надежность информационной системы. Анализируя реальные данные по различным действиям пользователя в информационной системе с помощью предложенного метода, мы сможем разделить пользователей на легальных и нелегальных и выявить ошибки.

Применение статистических методов для обнаружения злоумышленника в информационной системе рассматривалось такими отечественными и зарубежными учеными, как В. Столлингс, А.А. Корниенко, И.М. Слюсаренко, С.М. Доценко, Н.Н. Фимичев, Rodriguez, M.Z., Zeng L., Zhang M., Bouchachia A. [1, 2, 3]

Файлы журналов регистрации событий безопасности дают представление о состоянии информационной системы и позволяют обнаруживать аномалии в поведении пользователей и фиксировать инциденты информационной безопасности. Однако, автоматический анализ данных журналов событий безопасности затруднен, поскольку он содержит огромное количество неструктурированных данных, собранных из различных источников. Файлы журнала содержат информацию почти обо всех событиях, происходящих в информационной системе, в зависимости от уровня журнала. Для этого развернутая инфраструктура ведения журналов автоматически собирает, объединяет и хранит журналы, которые постоянно создаются большинством компонентов и устройств.

Основная проблема исследования заключается в том, что при анализе журналов инциденты обнаруживаются только задним числом. Кроме того, анализ журналов – это трудоемкая и ресурсоемкая задача, требующая знания предметной области о системе. Таким образом, обнаружение аномалий в поведении пользователей в реальном времени становится возможным благодаря постоянному мониторингу системных журналов в режиме онлайн, то есть сразу после их создания. Это позволяет своевременно реагировать на инциденты информационной безопасности и снижает вызванные ими расходы. К сожалению, эта задача вряд ли возможна для человека, поскольку данные журнала генерируются в огромных объемах и с большой скоростью. При рассмотрении крупных корпоративных систем нередко количество ежедневно создаваемых строк журнала исчисляется миллионами. Например, общедоступные журналы распределенной файловой системы Hadoop (HDFS) содержат более 4 миллионов строк журнала в день, а небольшие организации имеют дело с пиковыми значениями 22000 событий в секунду. Статистические методы позволяют анализировать большие данные журналов событий безопасности и выявлять аномалии в поведении пользователя с большей эффективностью.

При сравнении подходов нейронных сетей и других статистических методов можно увидеть, что статистические методы используют формулы, а нейронные сети – графическую интерпретацию. При использовании нейронных сетей основное время занимает обучение сетей, тогда как в статистике основное время посвящается анализу задачи, для которого требуются предшествующие знания. Нейросетевой подход, в свою очередь, в большинстве случаев может без таких знаний обойтись.

Вопрос о том, какие методы лучше использовать для обнаружения злоумышленника в информационной системе, остается открытым. Выбор метода зависит от ситуации, которая в основном определяется наличием априорной информации о данных, по которым можно выявить злоумышленника.

Мы остановимся на исследовании методов, основанных на нейронных сетях, так как это более молодая, в данный момент развивающаяся, область, доступная для применения в различных сферах деятельности. С помощью нейронных сетей можно анализировать большее количество данных и получить возможность обнаруживать злоумышленни-

ка в информационной системе за более короткое время за счет обучения [4, 5 с. 1-7].

Искусственная нейронная сеть (ИНС) – математическая модель, а также её программное или аппаратное воплощение, построенная по принципу организации и функционирования биологических нейронных сетей – сетей нервных клеток живого организма. Это понятие возникло при изучении процессов, протекающих в мозге, и при попытке смоделировать эти процессы.

ИНС представляют собой систему соединённых и взаимодействующих между собой простых процессоров (искусственных нейронов). Такие процессоры обычно довольно просты (особенно в сравнении с процессорами, используемыми в персональных компьютерах). Каждый процессор подобной сети имеет дело только с сигналами, которые он периодически получает, и сигналами, которые он периодически посылает другим процессорам. И, тем не менее, будучи соединёнными в достаточно большую сеть с управляемым взаимодействием, такие локально простые процессоры вместе способны выполнять довольно сложные задачи.

После того как нейронная сеть обучена множеством последовательных команд защищаемой системы или одной из ее подсистем, сеть представляет собой «образ» нормального поведения. Процесс обнаружения аномалий представляет собой определение показателя неправильно предсказанных команд, то есть фактически обнаруживается отличие в поведении объекта.

Преимущества:

– успех данного подхода не зависит от природы исходных данных;

– нейронные сети легко справляются с зашумленными данными;

– автоматически учитываются связи между различными измерениями, которые, несомненно, влияют на результат оценки.

При обнаружении злоумышленника в информационной системе предполагается, что его поведение отличается от поведения легального пользователя и эти различия можно оценить количественно. Невозможно будет увидеть совершенно разную работу в информационной системе нелегального пользователя по сравнению с легальным, тем не менее можно отследить в их поведении общие черты и рассчитать вероятность ошибки.

Основная задача исследования заключается в том, чтобы проанализировать большой объем данных по действиям пользователя в информационной системе и обучить нейронную сеть анализировать новые данные, что позволит определять, является пользователь легальным или нелегальным.

Входными параметрами модели является вектор, представляющий собой множество бинарных данных, характеризующих действия пользователя в информационной системе.

Ядром математической модели является нейронная сеть, обученная анализировать входные данные и выявлять в них аномалии, что будет интерпретироваться как аномальное поведение пользователя.

Выходные параметры:

– 0;

– 1,

	A	B	C	D	E	F	G	H	I	J	K	L
1	1	0	0	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0	0	0	0	0
5	1	0	0	0	0	0	0	0	0	0	0	0
6	1	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0
8	1	0	0	0	0	0	0	0	0	0	0	0
9	1	0	0	0	0	0	0	0	0	0	0	0
10	1	0	0	0	0	0	0	0	0	0	0	0
11	1	0	0	0	0	0	0	0	0	0	0	0
12	1	0	0	0	0	0	0	0	0	0	0	0
13	1	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0
15	1	0	0	0	0	0	0	0	0	0	0	0
16	1	0	0	0	0	0	0	0	0	0	0	0
17	1	0	0	0	0	0	0	0	0	0	0	0
18	1	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0
20	1	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0
22	1	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0
25	1	0	0	0	0	0	0	0	0	0	0	0
26	1	0	0	0	0	0	0	0	0	0	0	0

Рис. 1. Массив данных в MS Excel

где 0 – пользователь легальный, 1 – пользователь нелегальный.

Для дальнейшего анализа в качестве ядра модели выбран перцептрон, который является простейшей нейронной сетью, способной к обучению и решению поставленной задачи.

Для разработки тестовых заданий использован массив данных, представленных в бинарном виде, которые характеризуют действия пользователя в информационной системе.

Первый столбец представляет собой реальное положение пользователя:

0 – легальный;

1 – нелегальный.

Все последующие столбцы – данные, которые соответствуют поведению пользователя.

Часть массива представлена на рисунке 1.

Так как массив содержит более 700 видов оцениваемых данных по более 1500 пользователей, очень сложно провести его анализ.

Для того, чтобы сократить массив и составить таблицу из наиболее показательных примеров, проведена корреляция. Рассчитаны коэффициенты корреляции для каждого столбца относительно первого.

Выбраны 30 столбцов, модуль коэффициента корреляции которых наиболее приближен к 1. Фрагмент полученной таблицы представлен на рисунке 2.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
690	0	1	1	0	0	0	0	0	0	1	1	1	1	0	0	0
691	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
692	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
693	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
694	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
695	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
696	0	1	0	0	1	0	0	0	1	1	1	0	0	0	0	0
697	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
698	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
699	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
700	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
701	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
702	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
703	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
704	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
705	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
706	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
707	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
708	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
709	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
710	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
711	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
712	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
713	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0
714	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
715	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
716	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0
717	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
718	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
719	КОРРЕЛ	-0,08141	0,07409	-0,07723	-0,08987	0,074714	-0,07723	-0,07723	0,09866	-0,07396	-0,09839	-0,08406	-0,10922	0,085359	-0,07723	0,091926
720	МОДУЛЬ	0,08141	0,07409	0,077232	0,089874	0,074714	0,077232	0,077232	0,09866	0,073959	0,098393	0,084059	0,109218	0,085359	0,077232	0,091926

Рис. 2. Сокращенная таблица в MS Excel

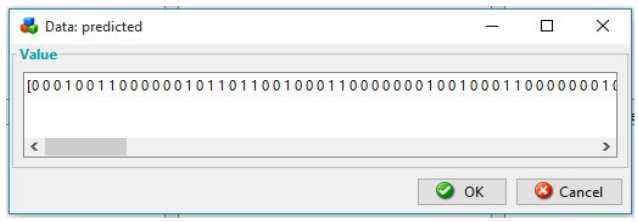


Рис. 3. Полученные выходные данные нейронной сети на персептроне

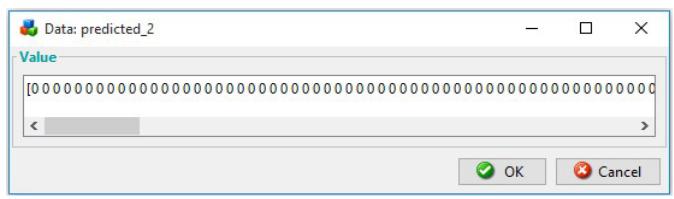


Рис. 4. Полученные выходные данные линейной нейронной сети

Полученная таблица далее использована для реализации метода в пакете MATLAB.

Полученную таблицу разделим пополам по строкам. Первую часть будем использовать для построения и обучения нейронной сети. Вторую – для реализации метода и выявления ошибок относительно реальных данных.

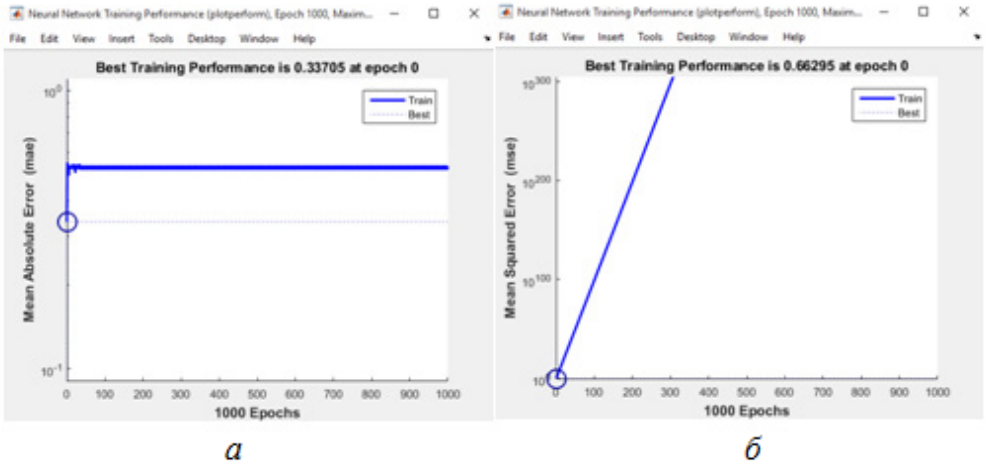
В пакете MATLAB проведено обучение нейронной сети на персептроне. На обучен-

ной нейронной сети смоделированы выходные данные для второй части таблицы с реальными данными.

Смоделированные нейронной сетью на персептроне выходные представлены на рисунке 3.

Для сравнения аналогичные действия проведены с линейной нейронной сетью. В данном случае в структуре результирующих данных преобладают нули, что говорит о том, что сеть определяет пользователя как легального в большинстве случаев. Выходные данные линейной нейронной сети представлены на рисунке 4.

Таким образом, сделан вывод, что персептрон наиболее подходит для решения поставленной задачи, так как для других сетей задача решается хуже.



а *б*

Рис. 5. Графики среднеквадратического отклонения нейронной сети на персептроне (а) и линейной нейронной сети (б)

На рисунке 5 представлены графики среднеквадратического отклонения для реализованных нейронных сетей. Среднеквадратическое отклонение для сети на персептроне меньше, чем для линейной нейронной сети, что так же свидетельствует о том, что сеть на персептроне способна точнее определить злоумышленника в информационной системе.

Далее проведено исследование уровней ошибок I-го и II-го рода для существующих методов и предложенного метода [6].

Ошибки I-го и II-го рода используются для проверки статистических гипотез и принятия решения на основе критерия, который может давать ложный результат.

H_0 – нулевая гипотеза, которая в данном случае соответствует нормальному поведению пользователя в информационной системе.

H_1 – альтернативная гипотеза, соответствующая аномальному поведению пользователя в информационной системе.

Возможны четыре варианта принятия решения:

– реальное поведение пользователя соответствует гипотезе H0, гипотеза H0 (пользователь легален) верно принята;

– реальное поведение пользователя соответствует гипотезе H0, гипотеза H0 (пользователь легален) неверно отвергнута, что представляет собой ошибку I-го рода;

– реальное поведение пользователя соответствует гипотезе H1, гипотеза H0 (пользователь легален) верно отвергнута;

– реальное поведение пользователя соответствует гипотезе H1, гипотеза H0 (пользователь легален) неверно принята, что представляет собой ошибку II-го рода.

Ошибка I-го рода возникает, когда легального пользователя принимают за злоумышленника, а ошибка II-го рода – когда злоумышленник определяется в информационной системе как легальный пользователь.

Для определения количества ошибок I-го и II-го рода было проведено сравнение реальных данных и данных, смоделированных нейронной сетью на персептроне в пакете MATLAB.

В таблице Excel было посчитано:

– количество легальных и нелегальных пользователей;

– количество ошибок I-го и II-го рода.

Ошибка I-го рода возникает, когда легального пользователя принимают за нелегального:

	A	B	C	D	E	F	G	H	I	J	K
1	Реальные данные:	1	0	1	1	0	1	1	1	0	1
2	Результаты нейронной сети:	0	0	0	1	0	0	1	1	0	0
3	Ошибки I-го рода:		0			0				0	
4	Ошибки II-го рода:	1		1	0		1	0	0		1
5											
6	Количество легальных пользователей:	108									
7	Количество нелегальных пользователей:	251									
8	Количество ошибок I-го рода	23									
9	Количество ошибок II-го рода	63									

Рис. 6. Определение количества ошибок I-го и II-го рода

– реальные данные – 0;

– результаты нейронной сети – 1.

Ошибка II-го рода возникает, когда нелегального пользователя принимают за легального:

– реальные данные – 1;

– результаты нейронной сети – 0.

В строках «Ошибки I-го рода» и «Ошибки II-го рода» использована функция ЕСЛИ соответственно для нуля или единицы в строке «Реальные данные». Если результат нейронной сети равен реальным данным в ячейку заполняется 0 – нет ошибки, если не равен – 1 – ошибка есть.

Таким образом количество ошибок I-го рода равно 23 (вероятность ошибки I-го рода – 21,3%), а количество ошибок II-го рода – 63 (вероятность ошибки II-го рода – 25,1%). Расчет ошибок представлен на рисунке 6.

При сравнении с количеством ошибок I-го и II-го рода других статистических методов был сделан вывод, что при равной вероятности ошибки I-го рода предложенный метод имеет меньшую вероятность ошибки II-го рода. Например, при заданной вероятности ошибки I-го рода равной вероятности ошибки I-го рода предложенного метода (≈20%) вероятности ошибки II-го рода для критериев Пирсона и Колмогорова равны ≈50% и ≈30% соответственно, что свидетельствует о том, что предложенный метод будет более надежным при обнаружении злоумышленника в информационной системе.

Исследование статистических методов обнаружения злоумышленника в информационной системе показало, что метод, основанный на нейронных сетях, наиболее подходит для реализации на предприятии, так как не требует большого объема памяти, обладает хорошим быстродействием, требует меньше времени на реализацию и позволяет анализировать больший объем данных. Использование метода на основе нейронных сетей позволяет более эффективно обеспечить обнаружение злоумышленника в информационной системе.

Литература

1. У. Столлингс. Современные компьютерные сети. 2–е издание. «Питер» 2003, 784 с.

2. Фергюсон, Нильс, Шнайер, Брюс. Практическая криптография.: Пер. с англ. М.: Издательский дом "Вильямс", 2005.

3. https://habr.com/ru/company/nix/blog/478286/ Выявление мошенничества с помощью алгоритмов случайного леса, нейронного автокодировщика и изолирующего леса.

4. https://www.securitylab.ru/blog/company/pt/345640.php Обнаружение веб–атак с помощью рекуррентных нейронных сетей.

5. Мустафаев А.Г. Нейросетевая система обнаружения компьютерных атак на основе анализа сетевого трафика // Вопросы безопасности. – 2016. – № 2. – С. 1 – 7. DOI: 10.7256/2409–7543.2016.2.18834 URL: https://nbpublish.com/library_read_article.php?id=18834

6. Ивченко Г.И., Медведев Ю.И. Введение в математическую статистику: Учебник. М.: Издательство ЛКИ, 2010.

References

1. U. Stollings. Sovremennyye komp'yuternyye seti. 2–ye izdaniye. «Piter» 2003, 784 s.

2. Fergyuson, Nil's, Shnayyer, Bryus. Prakticheskaya kriptografiya.: Per. s angl. M.: Izdatel'skiy dom "Vil'yams", 2005.

3. https://habr.com/ru/company/nix/blog/478286/ Vyyavleniye moshennichestva s pomoshch'yu algoritmov sluchaynogo lesa, neyronnogo avtokodirovshchika i izoliruyushchego lesa.

4. https://www.securitylab.ru/blog/company/pt/345640.php Obnaruzheniye veb–atak s pomoshch'yu rekurrentnykh neyronnykh setey.

5. Mustafayev A.G. Neyrosetevaya sistema obnaruzheniya komp'yuternykh atak na osnove analiza setevogo trafika // Voprosy bezopasnosti. – 2016. – № 2. – S. 1 – 7. DOI: 10.7256/2409–7543.2016.2.18834 URL: https://nbpublish.com/library_read_article.php?id=18834

6. Ivchenko G.I., Medvedev YU.I. Vvedeniye v matematicheskuyu statistiku: Uchebnik. M.: Izdatel'stvo LKI, 2010.

КРОТОВА Елена Львовна, кандидат физико-математических наук, доцент кафедры «Высшая математика», Пермский национальный исследовательский политехнический университет. 614990, Пермский край, г. Пермь, Комсомольский проспект, д. 29. lenkakrotova@yandex.ru

KROTOVA Elena Lvovna, Candidate of Physical and Mathematical Sciences, Associate Professor of the Department of Higher Mathematics, Perm National Research Polytechnic University. 614990, Perm Territory, Perm, Komsomolsky prospect, 29. lenkakrotova@yandex.ru

УРАЗБАЕВА Юлия Владимировна, учебный мастер кафедры «Высшая математика», Пермский национальный исследовательский политехнический университет. 614990, Пермский край, г. Пермь, Комсомольский проспект, д. 29. yulyia.urazbaeva@mail.ru

URAZBAEVA Yulia Vladimirovna, training master of the Department of Higher Mathematics, Perm National Research Polytechnic University. 614990, Perm Territory, Perm, Komsomolsky prospect, 29. yulyia.urazbaeva@mail.ru