ПРИМЕНЕНИЕ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ ВЫЯВЛЕНИЯ АНОМАЛЬНОГО ПОВЕДЕНИЯ ПОЛЬЗОВАТЕЛЕЙ ЦЕНТРОВ ОБРАБОТКИ ДАННЫХ

Саенко И.Б.[1], Котенко И.В.[2], Аль-Барри М.Х.[3]

Цель статьи: *разработка метода выявления аномального поведения пользователей центров обработки данных, основанного на использовании искусственных нейронных сетей.*

Метод исследования: *теоретический и системный анализ открытых источников данных по обнаружению SQL-запросов и построению искусственных нейронных сетей, разработка и программная реализация метода выявления аномального поведения пользователей центров обработки данных с использованием искусственных нейронных сетей, экспериментальная оценка разработанного метода.*

Полученный результат: *предложен подход к выявлению аномального поведения пользователей центров обработки данных, основанный на внедрении в систему защиты аналитического блока, содержащего модуль искусственных нейронных сетей. Предложена структура искусственной нейронной сети в виде семи последовательно соединенных нейронных слоев фиксированной размерности, имеющих различные функции активации. Описан порядок формирования набора данных для обучения нейронной сети исходя из набора записей регистрационного журнала базы данных. Приведены примеры реализации и экспериментальной оценки предложенного метода, подтверждающие его результативность и высокую эффективность.*

Область применения предложенного подхода – компоненты обнаружения аномалий и кибератак, предназначенные для повышения эффективности систем мониторинга и управления информационной безопасностью.

Ключевые слова: *кибербезопасность, центр обработки данных, обнаружение аномалий, искусственная нейронная сеть, аналитический блок.*

DOI:10.21681/2311-3456-2022-2-87-97

1. Введение

Вопросы обнаружения аномального поведения пользователей в автоматизированных системах управления являются актуальными на протяжении последних десятилетий. По настоящее время они не имеют однозначных ответов, так как появляются новые технологии обработки информации.

Одной из таких технологий, которая получает широкое распространение во многих сферах, являются облачные технологии, связанные с использованием центров обработки данных (ЦОД). ЦОДы играют важную роль в системах управления различного назначения. Они составляют информационно-техническую основу облачной инфраструктуры, поскольку поддерживают хранилище разнородной информации, используемой пользователями в своих интересах [1]. По этой причине ЦОДы являются объектами, на которые в первую очередь нацелены нарушители безопасности с целью получения информации или нарушения работы центров. Однако эти нарушители могут быть как внутренними, так и внешними [2].

Проблема обнаружения аномалий в поведении пользователей ЦОД не в последнюю очередь связана с постоянным увеличением их распространения и усложнением топологии, а также обновлением аппаратного и программного обеспечения. Это, в свою очередь, вызывает определенные проблемы в управлении и значительно повышает требования к квалификации персонала.

1 Саенко Игорь Борисович, доктор технических наук, ведущий научный сотрудник, ФГБУН «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН), г. Санкт-Петербург, Россия. E-mail: ibsaen@comsec.spb.ru

2 Котенко Игорь Витальевич, доктор технических наук, профессор, главный научный сотрудник и заведующий лабораторией проблем компьютерной безопасности, ФГБУН «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН), г. Санкт-Петербург, Россия. E-mail: ivkote@comsec.spb.ru

3 Аль-Барри Мазен Хамед, адъюнкт, Военная академия связи имени Маршала Советского Союза С.М. Буденного, г. Санкт-Петербург, Россия. E-mail: mazenb51@gmail.com

Для обеспечения информационной безопасности ЦОД используются межсетевые экраны, антивирусные средства, системы обнаружения вторжений, системы мониторинга [3] и системы контроля доступа [4,5]. При этом при построении систем защиты информации могут использоваться различные методы поиска аномалий, обладающие той или иной степенью эффективности. Однако эти аномалии обычно обнаруживаются в сетевом трафике. Но аномалии сетевого трафика не отражают неправильное, ненормальное поведение пользователей при работе с базами данных (БД). Эти действия можно обнаружить только путем анализа регистрационных журналов баз данных. Такой анализ, направленный на выявление аномального поведения пользователей ЦОД, в настоящее время либо не проводится, либо проводится не в полном объеме. Во многом это связано с особенностями наборов данных, формирующих записи регистрационных журналов БД, и сложностью использования наиболее известных классификаторов и методов машинного обучения для их анализа (SVM, Байесовские сети, методы k-средних, случайного леса, логистической регрессии и т.д.) [6]. Также до сих пор не были должным образом рассмотрены вопросы использования искусственных нейронных сетей в контексте анализа действий пользователей и предотвращения различных атак или злоупотреблений, спровоцированных ими.

Предлагаемый в статье подход ориентирован на обнаружение аномалий в регистрационных журналах БД и основан на использовании аналитических блоков, содержащих искусственные нейронные сети. При этом предполагается использовать комбинированные нейронные сети прямого распространения (многослойные персептроны), в которых отдельные слои имеют различные функции активации. Такие сети просты в реализации и при этом имеют достаточно высокую производительность, благодаря чему они хорошо функционируют на больших наборах данных. С другой стороны, комбинирование слоев с различными функциями активации позволяет устранить многие недостатки, присущие обычным многослойным персептронам.

Дальнейшая структура статьи является следующей. В разделе 2 анализируется состояние исследований в области анализа и поиска аномалий. В разделе 3 обсуждаются вопросы получения наборов данных для обучения и тестирования нейронных сетей. Раздел 4 описывает структуру предлагаемого аналитического блока, содержащего нейросетевой модуль, и порядок его функционирования. Экспериментальные результаты по обнаружению аномалий представлены и обсуждены в разделе 5. Заключение содержит общие выводы и направления дальнейших исследований.

2. Существующие методы анализа и поиска аномалий

Популярные методы обнаружения аномалий в действиях пользователей ЦОД можно свести в несколько укрупненных групп.

Первая группа содержит методы, осуществляющие тестирование на основе модели данных. Результатом испытаний является выявление точек, имеющих существенное отклонение от построенной модели. Эти точки являются аномалиями [7–9].

Вторая группа основана на расчете метрик. В качестве примера можно привести расчет расстояния до k-го соседа, определяя тем самым отклонение от нормы [10, 11].

Третья группа методов производит определение аномалий с помощью статистических тестов [12, 13]. Обычно эти методы используются для выявления отдельных аномалий.

К четвертой группе можно отнести методы поиска аномалий, основанные на индукции и кластеризации задачи на более мелкие фрагменты [14, 15].

В пятую группу можно включить методы машинного обучения, такие как подбор эллипсоидальных данных [16] или метод опорных векторов (SVM) [17].

Шестую группу формируют методы обнаружения аномалий путем вычисления среднего балла по результатам выполнения нескольких алгоритмов обнаружения [18,19].

Кроме того, отдельно можно выделить методы, основанные на шаблонах сетевых атак [20], контекстном поиске [21] и анализе состояния сигнатур [22].

Анализ известных методов обнаружения аномалий позволяет сделать вывод, что аномалии поведения пользователей ЦОД можно классифицировать следующим образом:

- выделение отдельного экземпляра данных, аномального по отношению к другим экземплярам данных;
- выделение отдельного экземпляра данных, аномального в определенном контексте;
- выявление набора данных, аномального в отношении других данных, при этом каждый отдельный экземпляр данных этого набора не является аномальным.

Эти свойства аномалий будут использоваться далее при построении искусственных нейронных сетей.

3. Получение исходных данных для искусственной нейронной сети

В качестве исходных данных для обучения и последующей работы искусственных нейронных сетей целесообразно использовать результаты анализа регистрационных журналов (журналов транзакций) баз данных, используемых в ЦОД, преобразованные в числовую форму.

Как правило, журнал транзакций хранится в виде одного или нескольких постоянно обновляемых файлов. В журнале, в текстовом виде содержится информация о событиях базы данных и о пользовательских запросах. По умолчанию сюда включена следующая информация:

- дата запроса;
- время запроса;
- источник запроса;
- информация о событии (запросе).

В работе в качестве источника для формирования набора исходных данных (датасета) использовались журналы транзакций базы данных образовательного портала образовательного учреждения, включающей около 4000 таблиц, размеры которых варьировались от десяти до нескольких сотен записей. Одновременно к базе данных могли обращаться до нескольких сотен пользователей, и их права доступа различались в зависимости от категории учетной записи.

Исходные данные для работы нейронной сети формируются следующим образом.

Вначале происходит накопление информации о периодичности обращений пользователя к базе данных и к конкретным таблицам, типах запросов, их частоте.

Выделяется типовой диапазон времени суток, в котором данный пользователь работает с базой данных.

Сохраняется адресная информация об источнике запроса (получателе выборки данных).

По завершении накопления определенного объема данных, выполняется их статистическая обработка. Результаты обработки используются для генерации обучающей (контрольной) выборки чисел, используемой для обучения (работы) искусственной нейронной сети.

Примеры записей из журнала транзакций базы данных представлены на рисунке 1. До ключевого слова `statement` каждая запись содержит информацию о дате и времени запроса, а также пользователе, совершившего запрос. После этого ключевого слова содержится текст SQL-запроса, с которым пользователь обращался к базе данных.

Всего в представленном фрагменте содержится пять SQL-запросов. Эти запросы сформированы двумя пользователями с идентификаторами 2536820 и 2536821. Запросы обращаются к трем таблицам: “p_group”, “512_2021” и “p_caf_num_code”. Каждый запрос является простым, так как содержит один оператор SELECT.

Преобразование записей журнала транзакций в числовую форму осуществляется следующим образом.

Вначале при анализе конкретного запроса выполняется разложение записи на отдельные составляющие.

Формируется выборка вида [0, 0, 0,..., 0]. Размерность выборки определяет размер входного слоя нейронной сети. Выполняется инициализация выборки нулями.

Затем производится первичный анализ времени запроса. Если обращение пользователя попадает в сохраненный ранее диапазон обращений, определенному значению в выборке присваивается значение 1. В противном случае высчитывается степень отклонения в диапазоне значений [0, 1].

Далее определяется источник запроса и вычисляется степень соответствия адреса источника той подсети, из которой ранее приходили запросы пользователя. Задается значение соответствия адресной

```
2022-03-04 07:34:06.238 UTC [2536820] postgres@2122 LOG:  statement:
SELECT DISTINCT facult FROM "p_group" ORDER BY facult;
2022-03-04 07:34:12.604 UTC [2536821] postgres@2122 LOG:  statement:
SELECT "29" FROM "512_2021" ORDER BY count;
2022-03-04 07:34:12.609 UTC [2536821] postgres@2122 LOG:  statement:
SELECT caf_num FROM p_caf_num_code ORDER BY id;
2022-03-04 07:34:12.610 UTC [2536821] postgres@2122 LOG:  statement:
SELECT "potok_num" FROM "p_group" WHERE groups='512'
2022-03-04 07:34:12.612 UTC [2536821] postgres@2122 LOG:  statement:
SELECT "groups" FROM "p_group" WHERE potok_num='5101'
```

Рис. 1. Примеры записей из журнала транзакций базы данных

информации, сохраненной ранее в диапазоне [0, 1]. Полученный результат также попадает в выборку исходных данных (в набор данных).

Затем анализируется содержание запроса к базе данных. На основе накопленных статистических данных вычисляется вероятность обращения пользователя к обозначенной в запросе таблице, которая задается в диапазоне [0, 1].

Аналогичным образом задаются вероятности использования запросов определенного типа (SELECT, UPDATE, DELETE, INSERT и т.д.), а также вероятности обращения в запросе к определенным полям в таблице. Полученные данные также включаются в выборку исходных данных для нейронной сети.

Соответственно, для каждой выборки исходных данных формируется выборка выходных значений нейронных данных, определяющая тип поведения пользователя.

Для имитации аномального поведения пользователей были внесены определенные изменения в журналы транзакций, которые в дальнейшем использовались для формирования обучающих и контрольных наборов данных, в сумме составляющих не более 2% от общего количества записей.

Суть изменений заключалась в следующем:

- для пользователей, у которых нет прав на запись и изменение данных, добавлены записи UPDATE, UPDATE, DELETE, INSERT, CREATE TABLE, DROP TABLE;
- добавлены обращения пользователей к определенным записям в служебных таблицах базы данных, а также к таблицам, к которым они ранее никогда не обращались;
- добавлены запросы, характерные для отдельных пользователей, но в необычное для работы время, например, в нерабочее время, или с сетевых адресов, с которых они ранее не обращались к базе данных.

Помимо вышеперечисленного, больше никаких изменений в журналы транзакций и в полученные на их основе наборы данных не вносилось.

4. Структура аналитического блока обнаружения аномалий

Для обнаружения аномалий в действиях пользователей ЦОД целесообразно использование реализованных программно аналитических блоков. Аналитический блок имеет в своем составе (рисунок 2):

- модуль преобразования исходных данных;
- модуль, содержащий искусственные нейронные сети;
- модуль для интерпретации полученных результатов.

Модуль преобразования исходных данных выполняет функцию считывания необходимых записей из журнала транзакций баз данных и преобразует их содержимое в числовую форму с последующим формированием выборок, выступающих в качестве исходных данных для работы нейронной сети.

Рис. 2. Обобщенная схема аналитического блока

При необходимости выполняется нормализация исходных данных в пределах [0, 1]. Для этой цели используется, например, помощи сигмоидальная логистическая функция, имеющая следующий вид:

$$\tilde{x}_{ik} = \frac{1}{e^{-a(x_{ik}-x_{ci})}+1}, \quad (1)$$

где x_{ik} и $\tilde{x}_{ik}$ – начальное и нормированное значения *k*-го элемента данных *i*-го интервала соответственно; x_{ci} – центр нормированного *i*-го интервала, *a* – параметр наклона функции.

При формировании обучающего набора данных необходима обработка полученных выборок человеком-экспертом, в целях формирования пар входных/выходных наборов данных, необходимых для обучения нейронной сети.

Модуль искусственных нейронных сетей (модуль ИНС) включает в себя одну или несколько нейронных сетей, выполняющих, после соответствующего обучения, функции преобразования числовых выборок, полученных в модуле преобразования исходных данных. Преобразование осуществляется к виду, приближенному к одному из выходных наборов данных, на которых ранее эти нейронные сети обучались.

Обобщенная структура искусственной нейронной сети представлена на рисунке 3. Она включает в себя три модуля.

Каждый из модулей включает:

— некоторое количество слоев нейронов, имеющих линейную (ступенчатую) или нелинейную (гиперболический тангенс) функции активации;
— слои с фильтрующей (Relu) функцией активации;
— отдельный модуль отсева (Dropout), выполняющий функции удаления избыточных связей между слоями нейронов.

Идея использования слоев с разными функциями активации в модуле ИНС основана на следующих предположениях. Каждая функция активации имеет свои преимущества и недостатки. Так, для линейной функции, одной из самых простых в реализации, недостатком является отсутствие смысла в построении многослойных структур, так как несколько линейных слоев всегда можно с успехом заменить одним линейным слоем. Нелинейная активация обеспечивает высокую точность в граничных областях, а также при больших весах. Совместное использование слоев с этими двумя типами функции активации приводит к тому, что недостатки, присущие каждому из слоев, устраняются, а достоинства сохраняются. В результате совместное расположение этих двух слоев позволяет уменьшить количество нейронов в обоих слоях по сравнению с использованием только слоя с линейной или нелинейной функцией активации. В результате время обучения сети сокращается.

Третья функция активации Relu сочетает в себе свойства линейности и нелинейности. Во всех точках, кроме нуля, она линейная. При нуле она становится нелинейной. Из-за своей линейной природы слой с функцией Relu можно использовать в качестве хорошего аппроксиматора. Таким образом, ИНС становится более устойчивой к шуму входного набора данных. Кроме того, тот факт, что функция Relu равна нулю в отрицательной области, способствует разрядке нейронной функции в ИНС, т.е. многие нейроны в этом слое приобретают нулевые значения. А это также обеспечивает сокращение времени обучения в случае очень большой размерности входного вектора.

Выходной линейный слой необходим для уменьшения размерности выходных данных до необходимого уровня. Он использует линейную функцию активации как самую простую.

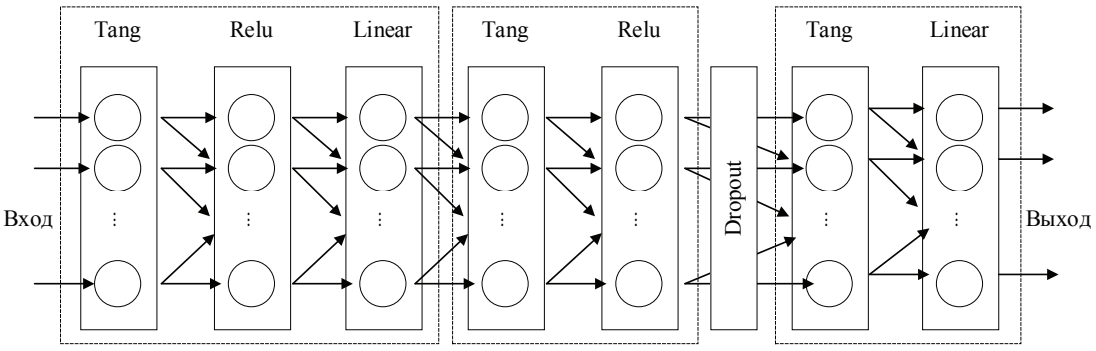

Рис. 3. Обобщенная структура искусственной нейронной сети

Для нейронных сетей с предложенной структурой в качестве метода обучения целесообразно использовать метод адаптивного шага обучения (Adadelta) [23]. Этот метод позволяет получать качественные результаты за приемлемый период времени.

Метод Adadelta применяет экспоненциальное скользящее среднее для оценки второго момента градиента g_t. Обновление параметров происходит следующим образом:

$$g_{t+1} = \gamma g_i + (1-\gamma)\nabla f_i(\theta_t)^2, \quad (2)$$

$$v_{t+1} = -\frac{\sqrt{x_t + \varepsilon}\nabla f_i(\theta_t)}{\sqrt{g_{t+1} + \varepsilon}}, \quad (3)$$

$$x_{t+1} = \gamma x_t + (1-\gamma)v_{t+1}^2, \quad (4)$$

$$\theta_{t+1} = \theta_t + v_{t+1}, \quad (5)$$

где f_i – функция, рассчитанная на *i*-й части данных, *t* – шаг итерации, x_t – скользящее среднее, γ– гиперпараметр.

Модуль для интерпретации полученных результатов выступает в качестве фильтра, определяющего близость результата, полученного на выходе искусственной нейронной сети, к эталонным значениям, заданным ранее человеком-экспертом для обучающего набора данных.

5. Выявление аномалий в действиях пользователей

Для реализации предложенного подхода к выявлению аномалий в действиях пользователей ЦОД целесообразно размещение аналитического блока на узле компьютерной сети ЦОД с возможностью получения данных из журнала транзакций БД в режиме времени, близком к реальному. Место аналитического блока в структуре ЦОД показано на рисунке 4.

Для тестирования использовались три разновидности аналитических блоков, включающие в свой состав одну, две и три искусственные нейронные сети, соответственно. Обучение каждой нейронной сети проводилось с использованием разных наборов данных. Объем наборов данных изменялся от 15 до 30 тысяч записей. Приложение, реализующее аналитический блок, было написано на языке Python (версия Django). Для реализации искусственных нейронных сетей использовалась библиотека PyTorch. Эксперименты проводились на компьютере со следующими характеристиками: процессор Intel Core i5, оперативная память DDR4 16Gb, видеокарта GeForce RTX 3060 и SSD на 128 ГБ.

Тестовый набор данных был сформирован путем добавления в нормальный набор данных аномальных записей. Аномальные записи формировались случайным образом путем дублирования имеющихся записей и изменений имен таблиц данных в SQL-запросах.

Базовая структура искусственной нейронной сети на начальном этапе экспериментов включает 7 слоев фиксированной размерности. Первый (входной) слой имеет 12 нейронов с нелинейной функцией активации (гиперболический тангенс). Второй и третий слои также включают по 12 нейронов с нелинейной (Relu) и линейной (ступенчатой) функциями активации. Четвертый, пятый и шестой слои состоят из 18

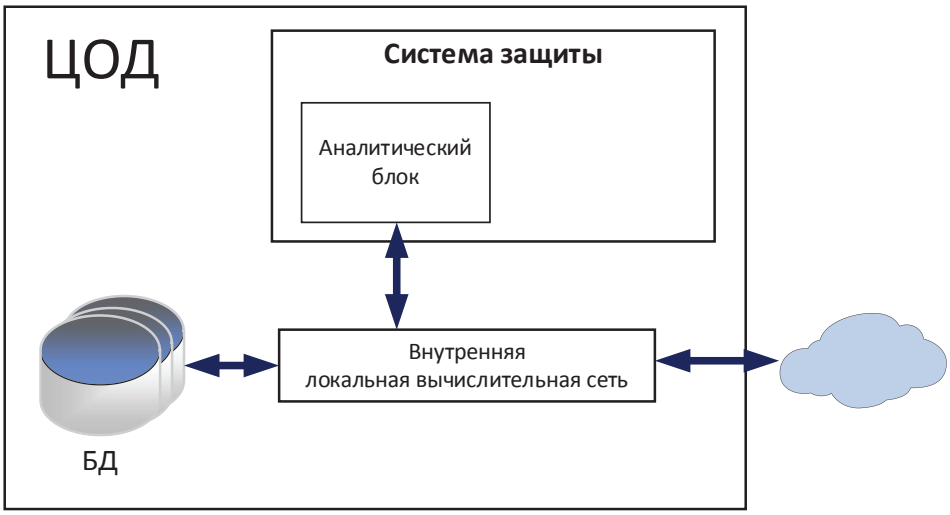

Рис. 4. Место аналитического блока в структуре ЦОД

нейронов каждый. Для этих слоев используются нелинейные функции активации: для четвертого и шестого слоев – гиперболический тангенс, для пятого – Relu. К связям между нейронами пятого и шестого слоев применяется функция Dropout. Седьмой слой является выходным и состоит из одного нейрона с линейной функцией активации.

Тестовый набор данных был создан путем случайного добавления аномальных записей в журналы транзакций. Затем на основе дополненного журнала транзакций формировался соответствующий набор данных.

Тесты проводились на наборах данных разного размера, содержащих 15000, 25000 или 30000 записей.

В таблице 1 представлены результаты тестирования аналитического блока, включающего одну, две или три искусственные нейронные сети, обученные на наборах данных разного объема, содержащих 15000, 25000 или 30000 записей.

Ложная интерпретация результата определяется как сумма частоты ложноположительных результатов (*FPR*) и частоты ложноотрицательных результатов (*FNR*). *FPR* и *FNR* определяются следующим образом:

$$FPR = FP / N, \quad (6)$$

$$FNR = FN / N, \quad (7)$$

где N – объем тестируемой выборки, $N = TP + TN + FP + FN$; TP – количество правильных положительных решений; TN – количество правильных отрицательных решений; FP – количество ошибочных положительных решений; FN – количество ошибочных отрицательных решений.

В таблице 2 приведены затраты времени на обучение нейронных сетей в зависимости от их количества и размера обучающей выборки. Значения среднего времени обучения рассчитаны по результатам трех тестов. Программная реализация каждого аналитического блока позволила обучать искусственные нейронные сети одновременно, в параллельных потоках.

Анализ полученных экспериментальных результатов позволяет сделать следующие выводы. Наилучший результат по точности выявления аномалий в поведении пользователей ЦОД показал аналитический блок, в состав которого входят три искусственные нейронные сети, обученные на наборе данных из 30000 записей. Ошибка обучения уменьшается по мере увеличения количества итераций. Точность обучения можно увеличить вдвое, если в аналитическом блоке использовать не одну, а три искусственные нейронные сети.

Анализируя данные о времени обучения, можно сделать следующий вывод. Время обучения аналитического блока увеличивается с увеличением размера набора данных по линейному закону. Эту закономерность легко увидеть, если сравнить время обучения друг с другом для размеров наборов данных, равных 15000 и 30000, и разного количества нейронных сетей. Так, с одной нейронной сетью эти времена равны, соответственно, 550 и 1100 секунды, с двумя нейронными сетями – 990 и 1980 секунд, с тремя ней-

Таблица 1

Результаты тестирования аналитического блока

Объем набора данных, 10^3	Ложная интерпретация результата, %		
	1 нейронная сеть	2 нейронные сети	3 нейронные сети
15	41	37	20
25	25	19	12
30	19	14	9

Таблица 2

Время обучения для различных наборов данных

Объем набора данных, 10^3	Время обучения, сек.		
	1 нейронная сеть	2 нейронные сети	3 нейронные сети
15	550	990	1990
25	920	1650	3310
30	1100	1980	3980

ронными сетями – 1990 и 3980 секунд. Получается, что двукратное увеличение размера набора данных также удваивает время обучения.

При этом зависимость времени обучения от количества нейронных сетей в аналитическом блоке носит нелинейный характер. Точнее, подчиняется степенной зависимости. Итак, если рассматривать экспериментальные данные, полученные для набора данных объемом 15000 записей, то легко увидеть, что переход от одной нейронной сети в аналитическом блоке к двум нейронным сетям увеличивает время обучения примерно в два раза. При этом переход с двух нейронных сетей на три также удваивает время обучения. Получается, что переход от одной нейронной сети к трем приводит к увеличению времени обучения уже в четыре раза. Эту зависимость можно увидеть для двух других наборов данных с объемами 25000 и 30000 записей.

Таким образом, перекрестная проверка выходных данных разных нейронных сетей и их дальнейшая обобщенная интерпретация позволяют значительно улучшить конечный результат работы аналитического блока. В этом случае не требуется усложнение структуры искусственной нейронной сети и, соответственно, не будет увеличения трудозатрат на обучение искусственных нейронных сетей.

6. Заключение

В статье рассмотрен подход к выявлению аномального поведения пользователей ЦОД, основанный на применении искусственных нейронных сетей, входящих в состав специализированного программно-реализованного аналитического блока, расположенного на одном из узлов локальной сети ЦОД.

По результатам анализа полученных результатов подтвержден вывод о возможности использования аналитического блока в существующем виде в качестве вспомогательного инструмента для выявления разного рода аномалий, в том числе аномалий в регистрационных журналах баз данных ЦОД.

Дальнейшие исследования планируется проводиться в следующих направлениях. Во-первых, с целью снижения трудозатрат, которые необходимо затрачивать на обучение и дальнейшее функционирование нейронной сети без потери качества результатов ее работы, планируется провести исследования по применению ИНС других архитектур, например, сетей обратного распространения, рекуррентных, LSTM, сверточных и т.д. Другое направление связывается с усложнением или изменением структуры модулей преобразования исходных данных и интерпретации полученных результатов, что также может привести к повышению точности определения аномалий.

Рецензент: *Молдовян Николай Андреевич, доктор технических наук, профессор, главный научный сотрудник лаборатории кибербезопасности и постквантовых криптосистем СПб ФИЦ РАН, Санкт-Петербург, Россия. E-mail: nmold@mail.ru*

Работа выполнена при частичной финансовой поддержке бюджетной темы FFSU-2019-0002.

Литература

1. Кожанков В.Н., Иванов И.И., Бондаренко Е.Ю., Моисеев А.С. Анализ нормативной правовой базы в сфере создания и эксплуатации центров обработки данных // Информационная безопасность - актуальная проблема современности. Совершенствование образовательных технологий подготовки специалистов в области информационной безопасности. 2021. Т. 1. № 1(14). С. 122-125.
2. Касенова Д.А. Необходимость обеспечения информационной безопасности центра обработки данных // Modern Science. 2021. № 10-1. С. 436-439.
3. Законодательно-правовое и организационно-техническое обеспечение информационной безопасности автоматизированных систем и информационно-вычислительных сетей. Котенко И.В., Котухов М.М., Марков А.С. и др. Под редакцией И.В. Котенко / Санкт-Петербург, 2000. 190 с.
4. Котенко И. В., Полубелова О.В., Саенко И.Б., Чечулин А.А. Применение онтологий и логического вывода для управления информацией и событиями безопасности // Системы высокой доступности, Т.8, № 2, 2012. С.100-108.
5. Kotenko I., Stepashkin M. Network Security Evaluation based on Simulation of Malefactor's Behavior // Proceedings. International Conference on Security and Cryptography, SECRYPT 2006. Polytechnic Institute of Setubal. Setubal, 2006. P. 339-344.
6. Гайдышев И.П. Оценка качества бинарных классификаторов // Вестник Омского университета. 2016. № 1(79). С. 14-17.
7. Kurt M.N., Yilmaz Y., Wang X. Sequential Model-Free Anomaly Detection for Big Data Streams // 2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2019, pp. 421- 425. DOI: 10.1109/ALLERTON.2019.8919759.

8. Ramapatruni S., Narayanan S.N., Mittal S., Joshi A., Joshi K. Anomaly Detection Models for Smart Home Security // 2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS). 2019. Pp. 19-24. DOI: 10.1109/BigDataSecurity-HPSC-IDS.2019.00015.
9. Wang E., Song Y., Xu S., Guo J., Qu P., Pang T. A detection model for anomaly on ADS-B data // 2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA). 2020. Pp. 990-994. DOI: 10.1109/ICIEA48937.2020.9248249.
10. Трясучкин В.А., Синцева М.М. Исследование оптимизации гиперпараметров алгоритма k-ближайших соседей // Вестник Пензенского государственного университета. 2019. № 2 (26). С. 63-68.
11. Харитонов С.П. Метод “ближайшего соседа” для математической оценки распределения биологических объектов на плоскости и на линии // Вестник Нижегородского университета им. Н.И. Лобачевского. Серия: Биология. 2005. № 1. С. 213-221.
12. Богданов А.И. Статистические тесты стабильности математических моделей прогнозирования // Вестник Санкт-Петербургского государственного университета технологии и дизайна. Серия 1: Естественные и технические науки. 2019. № 4. С. 9-13.
13. Баранов В.А., Конышев М.Ю., Привалов А.А., Шестаков А.В. Верификация криптографических алгоритмов на основе использования метода симуляции двоичных случайных последовательностей с заданными статистическими свойствами // Наукоемкие технологии в космических исследованиях Земли. 2019. Т. 11. № 6. С. 45-52. DOI: 10.24411/2409-5419-2018-10294.
14. Пекунов В.В. Индукция правил трансформации естественно-языковой постановки задачи в смысловую модель порождения решающей программы // Программные системы и вычислительные методы. 2020. № 3. С. 29-39. DOI: 10.7256/2454-0714.2020.3.33789.
15. Бурькова Е.В., Извекова Л.А. Применение метода кластеризации данных для решения задачи оценки рисков информационной безопасности // Национальная безопасность и стратегическое планирование. 2019. № 2 (26). С. 81-86.
16. Астапов В.Н. Оценивание с помощью эллипсоидов параметров линейной регрессии при ограничениях на вектор входных функций // Международный журнал прикладных и фундаментальных исследований. 2018. № 10. С. 9-15.
17. Куликов А.Л., Бездушный Д.И., Шарыгин М.В., Осокин В.Ю. Анализ применения метода опорных векторов в многомерной релейной защите // Известия Российской академии наук. Энергетика. 2020. № 2. С. 123-132.
18. Yin A., Zhang C. BOFE: Anomaly Detection in Linear Time Based on Feature Estimation // 2018 IEEE International Conference on Data Mining Workshops (ICDMW). 2018. Pp. 1128-1133. DOI: 10.1109/ICDMW.2018.00162.
19. Pratap U., Canudas-de-Wit C., Garin F. Average state estimation in presence of outliers // 2020 59th IEEE Conference on Decision and Control (CDC). 2020. Pp. 6058-6063. DOI: 10.1109/CDC42340.2020.9303809.
20. Смирнова Е.В., Абачараева Э.Р. Современные угрозы вирусных атак на компьютерные сети и критерии их оценивания // Технологии инженерных и информационных систем. 2020. № 3. С. 3-12.
21. Тельнов В.П. Контекстный поиск как технология извлечения знаний в сети интернет // Программная инженерия. 2017. Т. 8. № 1. С. 26-37. DOI: 10.17587/prin.8.26-37.
22. Стадник А.Н., Алпеев Е.В., Скрыль С.В. Методика формирования базы классификаций компьютерных атак на основе применения интеллектуального анализа сигнатур компьютерных атак // Вопросы оборонной техники. Серия 16: Технические средства противодействия терроризму. 2021. № 3-4 (153-154). С. 108-116.
23. Li D., Qiao Z., Song T., Jin Q. Adaptive Natural Policy Gradient in Reinforcement Learning // 2018 IEEE 7th Data Driven Control and Learning Systems Conference (DDCLS). 2018. Pp. 605-610. DOI: 10.1109/DDCLS.2018.8515994.

APPLICATION OF ARTIFICIAL NEURAL NETWORKS TO REVEAL ABNORMAL BEHAVIOR OF DATA CENTER USERS

Saenko I.B.[4], Kotenko I.V.[5], All-Barri M.H.[6]

The purpose of the article: *development of the method for detecting anomalous behavior of users of data centers based on the use of artificial neural networks.*

Research method: *theoretical and system analysis of open data sources for detecting SQL queries and creating artificial neural networks, development and software implementation of a method for detecting anomalous behavior of data center users using artificial neural networks, experimental evaluation of the developed method.*

The result obtained: *an approach to detecting anomalous behavior of users of data centers is proposed, based on the introduction of an analytical block containing a module of artificial neural networks into the protection system. The structure of an artificial neural network is proposed in the form of seven sequentially connected neural layers of a fixed dimension with different activation functions. The procedure for generating a data set for training a neural network based on a set of database log records is described. Examples of the implementation and experimental evaluation of the proposed method are given, confirming its effectiveness and high efficiency.*

The area of use of the proposed approach is anomaly and cyberattack detection components designed to improve the efficiency of information security monitoring and management systems.

Keywords: *cyber security, data center, anomaly detection, artificial neural network, analytical unit.*

References

1. Kozhankov V.N., Ivanov I.I., Bondarenko E.Yu., Moiseev A.S. Analysis of the regulatory legal framework in the field of creation and operation of data processing centers // Information security is an urgent problem of our time. Improving educational technologies for training specialists in the field of information security. 2021. Vol. 1. No. 1(14). P. 122-125 (in Russian).
2. Kasenova D.A. The need to ensure information security of the data center // Modern Science. 2021. No. 10-1. P. 436-439 (in Russian).
3. Legislative, legal, organizational and technical support of information security of automated systems and information and computer networks. Kotenko I.V., Kotukhov M.M., Markov A.S., et al. Edited by I.V. Kotenko / St. Petersburg, 2000. 190 p. (in Russian).
4. Kotenko I.V., Polubelova O.V., Saenko I.B., Chechulin A.A. Application of ontologies and inference for managing security information and events // High availability systems, Vol.8, No. 2, 2012. P. 100-108 (in Russian).
5. Kotenko I., Stepashkin M. Network Security Evaluation based on Simulation of Malefactor's Behavior // Proceedings. International Conference on Security and Cryptography, SECRYPT 2006. Polytechnic Institute of Setubal. Setubal, 2006. P. 339-344.
6. Gaydyshev I.P. Assessment of the quality of binary classifiers // Bulletin of the Omsk University. 2016. No. 1(79). P. 14-17 (in Russian).
7. Kurt M.N., Yilmaz Y., Wang X. Sequential Model-Free Anomaly Detection for Big Data Streams // 2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2019, pp. 421-425, doi: 10.1109/ALLERTON.2019.8919759.
8. Ramapatruni S., Narayanan S.N., Mittal S., Joshi A., Joshi K. Anomaly Detection Models for Smart Home Security // 2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS). 2019. Pp. 19-24. DOI: 10.1109/BigDataSecurity-HPSC-IDS.2019.00015.
9. Wang E., Song Y., Xu S., Guo J., Qu P., Pang T. A detection model for anomaly on ADS-B data // 2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA). 2020. Pp. 990-994. DOI: 10.1109/ICIEA48937.2020.9248249.
10. Tryasuchkin V.A., Sintseva M.M. Investigation of optimization of hyperparameters of the k-nearest neighbors' algorithm // Bulletin of the Penza State University. 2019. No. 2 (26). P. 63-68 (in Russian).

4 Igor B. Saenko, Dr.Sc. (in Tech.), Leading researcher at St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russia. E-mail: ibsaen@comsec.spb.ru

5 Igor V. Kotenko, Dr.Sc., Professor, Head of Laboratory of Computer Security Problems at St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russia. E-mail: ivkote@comsec.spb.ru

6 Mazen Al-Barri, Adjunct at Military Academy of Communications named after Marshal of the Soviet Union S.M. Budyonny, St. Petersburg, Russia. E-mail: mazenb51@gmail.com

11. Kharitonov S.P. The “nearest neighbor” method for the mathematical evaluation of the distribution of biological objects on a plane and on a line // Bulletin of the Nizhny Novgorod University named N.I. Lobachevsky. Series: Biology. 2005. No. 1. P. 213-221 (in Russian).
12. Bogdanov A.I. Statistical tests of the stability of mathematical forecasting models // Bulletin of the St. Petersburg State University of Technology and Design. Series 1: Natural and technical sciences. 2019. No. 4. P. 9-13 (in Russian).
13. Baranov V.A., Konyshev M.Yu., Privalov A.A., Shestakov A.V. Verification of cryptographic algorithms based on the use of the method of simulation of binary random sequences with given statistical properties // Science-intensive technologies in space research of the Earth. 2019. Vol. 11. No. 6. P. 45-52. DOI: 10.24411/2409-5419-2018-10294 (in Russian).
14. Pekunov V.V. Induction of rules for transforming a natural language problem statement into a semantic model for generating a solver // Program Systems and Computational Methods. 2020. No. 3. P. 29-39. DOI: 10.7256/2454-0714.2020.3.33789 (in Russian).
15. Burkova E.V., Izvekova L.A. Application of the data clustering method for solving the problem of information security risk assessment // National Security and Strategic Planning. 2019. No. 2 (26). P. 81-86 (in Russian).
16. Astapov V.N. Ellipsoid Estimation of Linear Regression Parameters under Constraints on the Vector of Input Functions // International Journal of Applied and Fundamental Research. 2018. No. 10. P. 9-15 (in Russian).
17. Kulikov A.L., Bezdushny D.I., Sharygin M.V., Osokin V.Yu. Analysis of the application of the support vector machine in multidimensional relay protection. Proceedings of the Russian Academy of Sciences. Energy. 2020. No. 2. P. 123-132 (in Russian).
18. Yin A., Zhang C. BOFE: Anomaly Detection in Linear Time Based on Feature Estimation // 2018 IEEE International Conference on Data Mining Workshops (ICDMW). 2018. Pp. 1128-1133. DOI: 10.1109/ICDMW.2018.00162.
19. Pratap U., Canudas-de-Wit C., Garin F. Average state estimation in presence of outliers // 2020 59th IEEE Conference on Decision and Control (CDC). 2020. Pp. 6058-6063. DOI: 10.1109/CDC42340.2020.9303809.
20. Smirnova E.V., Abacharaeva E.R. Modern Threats of Virus Attacks on Computer Networks and Criteria for Their Evaluation // Engineering and Information Systems Technologies. 2020. No. 3. P. 3-12 (in Russian).
21. Telnov V.P. Contextual search as a technology for extracting knowledge on the Internet // Program engineering. 2017. Vol. 8. No. 1. 3. 26-37. DOI: 10.17587/prin.8.26-37 (in Russian).
22. Stadnik A.N., Alpeev E.V., Skryl S.V. Methodology for the formation of a classification base for computer attacks based on the use of intelligent analysis of computer attack signatures. Questions of defense technology. Series 16: Technical means of countering terrorism. 2021. No. 3-4 (153-154). P. 108-116 (in Russian).
23. Li D., Qiao Z., Song T., Jin Q. Adaptive Natural Policy Gradient in Reinforcement Learning // 2018 IEEE 7th Data Driven Control and Learning Systems Conference (DDCLS). 2018. Pp. 605-610. DOI: 10.1109/DDCLS.2018.8515994.