Кроме всего выше сказанного, могут быть решены проблемы безопасности портов, охраны окружающей среды. Например, уменьшение уровня шума создаваемого портом и эффективного использования энергии без выброса в атмосферу вредных веществ. Будет легче решить вопрос глобального изменения климата и повышения уровня мирового океана.

Посмотрев на прошлое и будущее ИТ в водном транспорте, можно проследить экспансию информационных технологий, благодаря которой сфера водного транспорта становится объектом всё более глубокой информатизации, и вряд ли эта экспансия сбавит темп в ближайшее столетие. Конечно разговоры о интеллектуальных судах, портах – это лишь теория, но вполне реализуемая теория, и я считаю, что скоро мы сами станем свидетелями этой индустриализации информационных технологий в водном транспорте.

ЛИТЕРАТУРА

1. Малыгин, И.Г. Индустриальные революции и водный транспорт [Электронный ресурс] / И.Г. Малыгин, В.И. Комашинский, О.А. Королев, М.Ю. Аванесов, О.А. Михалев. Режим доступа - https://www.infokosmo.ru/file/article/16554.pdf

2. Комашинский, В. И. Когнитивная метафора в развитии телекоммуникационных и индустриальных сетевых инфраструктур, или первые шаги к постинформационной эпохе / В.И. Комашинский, Д.В. Комашинский // Технологии и средства связи. – 2015. – № 1. – С. 62–67.

3. Малыгин И.Г., Шаталова Н.В., Комашинский В.И. Транспортные технологии и глобализация в период 4-й индустриальной революции (проблемы и перспективы) // Информация и Космос. 2018. № 1. С. 6–13.

4. Транспортная стратегия РФ на период до 2030 г. (утверждена распоряжением Правительства РФ от 22 ноября 2008 г. № 1734)

ИДЕНТИФИКАЦИЯ ВНУТРЕННИХ УТЕЧЕК ДАННЫХ С ИСПОЛЬЗОВАНИЕМ НЕЙРОННОЙ СЕТИ-РЕПЛИКАТОРА

Банокин П. И.
(г. Томск, Томский Политехнический Университет)
pavel805@gmail.com

IDENTIFICATION OF INTERNAL DATA LEAKS WITH USE OF REPLICATOR NEURAL NETWORK

Pavel Banokin
(Tomsk, Tomsk Polytechnic University)

Abstract. The article is devoted to the problem of internal data leaks identification. The task is solved by user's behavior analysis. Input data is collected as an array of vectors describing user's actions. Behavior analysis is performed by use of context abnormality detection functions and creation of behavior profile. A data leak is identified by changes of user behavior profile entries and detected by applying replicator neural network to behavior profile entries.

Key words: internal data leaks, behavior profile, replicator neural network.

Введение. Утечкой данных является умышленная или случайная передача конфиденциальных данных неуполномоченным лицу или группе лиц [1]. Статистические данные [2, 3] подтверждают тенденцию увеличения количества случаев утечек данных из корпоративных информационных систем. Внутренние утечки данных отличаются сложностью обнаружения, так как совершаются индивидуумами в процессе исполнения служебных обязанностей и имеющими доступ к данных корпоративной информационной системы. Поведенческий анализ является одним из способов обнаружения внутренних утечек данных и основан на предположении, что поведение сотрудника в момент совершения утечки данных отличается от модели каждодневного поведения. При использовании поведенческого анализа необходимо создание модели поведения пользователя – поведенческого профиля и последующее сравне-

ние действий пользователя с ним [4]. Найденный поведенческие отличия могут быть выражены в доступе к ранее не используемым инструментам информационной системы, нарушением порядка выполнения бизнес-процессов и др.

Полнофункциональные системы предотвращения утечек данных (*Oracle Database Vault, InfoSphere Guardium Database Activity Monitor, McAfee Network User Behavior Analysis Studio* и др.), использующие поведенческий анализ, являются программным обеспечением с закрытым исходным кодом, а используемые в них алгоритмы анализа данных о поведении пользователя скрыты от потребителя. Ввиду этого существует необходимость создания алгоритмов идентификации утечек данных и последующая их реализации в виде программного комплекса.

Данные о поведении пользователя. Единицей информации о поведении пользователя является регистрируемая запись – вектор w, содержащий числовые, текстовые и категориальные данные. Регистрируемая запись содержит информацию о действии пользователя, включая имя пользователя, идентификатор вычислительного устройства, название используемого программного приложения, название коллекции данных, название инструмента информационной системы, заголовок окна и др.

Потоком регистрируемых записей $S = \{w_1, w_2, ..., w_n\}$ является упорядоченная по времени создания их последовательность. Временным окном W является подпоследовательность записей из потока ригистрируемых S .

Задача идентификации утечек данных. Задачей идентификации утечек данных является отнесение поведения пользователя к классу безопасного поведения или классу аномального поведения. В случае отнесения к классу аномального поведения также необходимо получение характеристик действий пользователя.

Результат проверки поведения пользователя принадлежит множеству $B = \{0,1\}$, элемент «0» обозначает класс обычного поведения и элемент «1» класс аномального поведения.

Функции поиска контекстных аномалий. Для анализа потока регистрируемых записей используются функции поиска контекстных аномалий $z_i = W \to B$. Функции поиска контекстных аномалий используют в своей работе данные из отдельных измерений вектора w. Использование функций поиска контекстных аномалий обусловлено тем, что алгоритмы анализа данных сложно адаптировать для использования входных данных разных типов: текстовых, категориальных и числовых; а также разным семантическим значением отдельных измерений вектора регистрируемой записи w .

Профиль поведения пользователя. Профиль поведения $P = \left\langle V^{\langle 1 \rangle} V^{\langle 2 \rangle} ... V^{\langle n \rangle} \right\rangle$ пользователя хранит векторы числовых значений, полученные при выполнении функций поиска контекстных аномалий z_i . Элемент поведенческого профиля $V^{\langle i \rangle}$ является k-мерным вектором:

$$V^{\langle i \rangle} = \begin{pmatrix} v_1^{\langle i \rangle} \\ v_2^{\langle i \rangle} \\ \dots \\ v_k^{\langle i \rangle} \end{pmatrix},$$

где k – количество функций поиска контекстных аномалий. $v_1^{\langle i \rangle} = z^{\langle i \rangle} : W \to \mathrm{B}$.

Поведенческих профиль позволяет хранить историю поведения пользователей не в виде коллекции регистрируемых записей, а в виде векторов числовых значений, значительно сокращая требования к размеру накопителя данных.

Идентификация утечек данных. Идентификация утечек данных происходит с помощью функции оценки $ev = P \to B$. Перед запуском процесса идентификации необходимо

накопление необходимого накопление необходимого количества регистрируемых записей для создания профиля пользователя.

В качестве реализации функции итоговой оценки используется нейронная сеть-репликатор [5] – многослойный персептрон, возвращающий входные данные с ошибкой репликации $error$. При подаче входных данных, которые отсутствовали в обучающей выборке или встречались в ней с низкой вероятностью, ошибка репликации превышает наблюдаемое по завершении процесса обучения значение ошибки обучения $error$. Нейронная сеть репликатор обучается на входных данных, представленными элементами профиля поведения P . Ошибка репликации рассчитывается на основе функции квадратичной ошибки:

$error = \frac{1}{2n}\sum_{i=1}^{n}(out_i - out\_ideal_i)^2$, где

n – количество входных нейронов;

out и out_ideal – векторы выходных и входных значений соответственно.

После обучения нейронной сети при поступлении новой записи в поведенческий профиль происходит выполнение функции итоговой оценки ev. Результат итоговой оценки, равный «1» и свидетельствующий об аномальном поведении пользователя, возвращается в том случае, если ошибка репликации превышает пороговое значение $error_{max}$.

Уведомление об аномальном поведении пользователя содержит визуализацию значений ошибки репликации и записи профиля поведения (рис. 1).

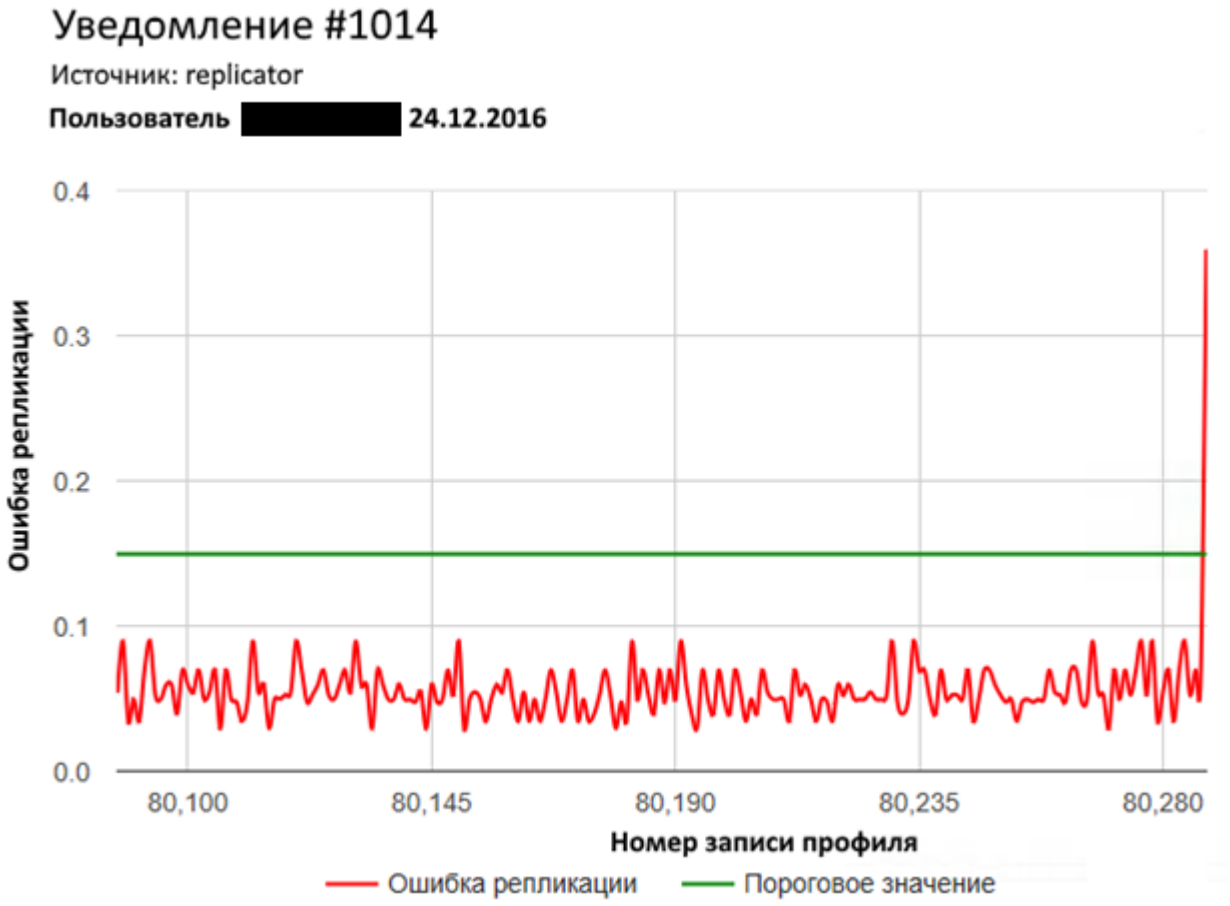

Рис. 1. Уведоление о возможном случае утечки данных

Для установления конфигурации внутренних слоев нейронной сети-репликатора использованы данные из источника [5] и функции активации *tanh, sigmoid* и ступенчатая (*stepwise*). Конфигурации нейронной сети проверены на поведенческом профиле из 300 записей и приведены в таблице 1. Размерность каждой записи профиля поведения составила 12 измерений. Оптимальной конфигурацией является конфигурация №2.

Таблица 1. Сравнение архитектур нейронной сети-репликатора

№	Архитектура сети, количество нейронов в слое (название функции активации)	Ошибка обучения	Максимальная ошибка репликации для элементов профиля	Минимальное значение ошибки репликации для постороннего элемента	Минимальная ошибка репликации для редких элементов профиля (P<1%)
1	13(sigmoid) 13(sigmoid) 13(sigmoid)	0,04	0,09	0,11	0,11
2	13(sigmoid) 14(sigmoid) 17(sigmoid) 14(sigmoid) 13(sigmoid)	0,07	0,10	0,17	0,12
3	13(sigmoid) 11(tanh) 10(stepwise) 11(tanh) 13(sigmoid)	0,82	0,69	0,59	0,40-0,94

Заключение. Нейронная сеть-репликатор в сочетании с предложенными способами обработки данных о поведении пользователей позволяет идентифицировать поведенческие аномалии и возможные случаи утечек данных. Способ обработки поведенческих данных с использованием функций поиска контекстных аномалий отличается гибкостью и возможностью адаптации к разным условиям эксплуатации корпоративных информационных систем.

1. Shabtai A., Elovici Y., Rokach L. A survey of data leakage detection and prevention Solutions. - Berlin, Germany: Springer, 2012.
2. Data leakage worldwide: The high cost of insider threats // Cisco. URL: http://www.cisco.com/en/US/solutions/ collateral/ns170/ns896/ns895/white_paper_c11-506224.pdf
3. Data breach investigations report 2012 // Verizon Enterprise Solutions Worldwide Site. URL: http://www.verizonbusiness.com/resources/reports/rp_data-breach-investigations-report-2012_en_xg.pdf
4. Denning D. E. An Intrusion Detection Model // IEEE transactions on softw are engineering. - 1987. - №SE-13, 2. - С. 222-232.
5. Hawkins S., He H., Williams G., Baxter R. Outlier Detection Using Replicator Neural Networks // International Conference on Data Warehousing and Knowledge Discovery. - NY, USA: Springer, 2002. - С. 170-180.