ции рисков инвестиционных проектов в сфере автоматизации промышленных предприятий// Инновационный вестник Регион. -2013.-No4.2.С. 55-60.
18. Наношкин А.Г., Макашов П.Л. Управление качеством ИТ-проекта // Современные научные исследования и инновации. 2015. № 10 [Электронный ресурс]. URL: http://web.snauka.ru/issues/2015/10/57965
19. Макашова В.Н., Трейбач Е.Л., Чусавитина Г.Н. Методика оценки ИТ-стартапа//Теплотехника и информатика в образовании, науке и производстве: сб. докладов IV Всероссийской научно-практической конференции студентов, аспирантов и молодых учёных (ТИМ'2015) с международным участием, посвящённой 95-летию основания кафедры и университета (Екатеринбург, 26–27 марта 2015 г.). – Екатеринбург: УрФУ, 2015. –С.319-323
20. Бикчурина А.И., Макашова В.Н. Расчет экономической эффективности проекта по разработке Веб-приложения//Информационные технологии в науке, управлении, социальной сфере и медицине: сборник научных трудов II Международной конференции «Информационные технологии в науке, управлении, социальной сфере и медицине»/Часть III/под. ред. О.Г. Берестневой, О.Г. Гергет; Национальный исследовательский Томский политехнический университет. - Томск: Изд-во Томского политехнического университета, 2015. -356 с. -26-28 с.
21. Чусавитина Г.Н., Макашова В.Н., Колобова О.Л. Управление ИТ-проектами [Текст] : учебно-методическое пособие / Г. Н. Чусавитина, В. Н. Макашова, О. Л. Колобова ; М-во образования и науки Российской Федерации, Магнитогорский гос. технический ун-т им. Г. И. Носова. - Магнитогорск : МГТУ, 2015. - 140 с

ИДЕНТИФИКАЦИЯ УТЕЧЕК ДАННЫХ НА ОСНОВЕ КЛАССИФИКАТОРА ПОВЕДЕНИЯ ПОЛЬЗОВАТЕЛЕЙ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

П.И. Банокин
(г. Томск, Томский политехнический университет)
e-mail: pavel805@gmail.com

DATA BREACH DETECTION BASED ON SOFTWARE USERS' BEHAVIOR CLASSIFIER

P.I. Banokin
(Tomsk, Tomsk Polytechnik University)

Annotation. The article considers the process of internal data breach detection. The proposed process is based on a set of functions which detect contextual abnormalties and a neural network performing final evaluation. The approach of users' sessions analysis is presented. Analysis methods for different data types are described.

Keywords: data breach protection, neural network, data mining

Введение. Утечки данных являются одной из главных причин финансовых убытков предприятия [1]. Для предотвращения утечек данных обеспечивается физическая безопасность носителей информации, проводятся образовательные семинары среди сотрудников предприятия, проверяются и разграничиваются права доступа и используются аппаратные и программные средств [2]. Идея предотвращения утечек данных с помощью программного анализа поведения пользователей основана на том, что действия пользователя при хищении данных отличаются от его каждодневного поведения при исполнении служебных обязанностей [3]. Поэтому можно выявить возможные случаи утечек данных, сравнивая поведение пользователей за разные промежутки времени или сравнивая поведение пользователей одной

группы между собой. Поведение пользователя программных приложений характеризуется большим числом параметров, среди которых могут быть названия программного приложения, используемой коллекции данных, тип совершаемого действия, название клиентского устройства, дата и время совершения действия и др. Так как поведенческие данные многомерны и каждая размерность может быть численной или категориальной, проводить анализ таких данных целесообразно по отдельным подпространствам. Анализ поведения в каждом подпространстве может использовать метрические алгоритмы классификации, методы математической статистики и теории вероятности, нечеткую логику [4] и др. В настоящей статье для анализа предлагается использовать следующий набор классификаторов и поведенческих данных, приведенных в таблице 1.

Таблица 1. Поведенческие данные и классификаторы

Источник данных	**Название**	**Тип данных**	**Возможные методы анализа**
Операционная система	Данные буфера обмена	Строковые данных произвольной длины и содержания	Сигнатурный анализ, методы обработки естественного языка.
Операционная система	Время совершения операций аутентификации и деаутентификации	Дата и время (преобразуются в численный формат)	Правило трех-сигм, сравнение выборок данных.
Операционная система	Название исполняемого файла с текущим фокусом управления, заголовок активного окна	Строковые категориальные данные	Построение наборов частых множеств признаков с помощью алгоритма Apriori и последующее сравнение созданных наборов.
Веб-браузер	Перечень посещенных URL-адресов	Строковые категориальные данные	

Процесс идентификации данных. В рассматриваемом процессе идентификации утечек данных на основе анализа поведения пользователя принимается во внимание сложность и разнородность входных данных (рис. 1).

Рис. 1

Обычно работа пользователя с компьютером или мобильным устройством представлена в виде серий последовательно совершаемых действий. Серии действий разделены временным интервалом или операциями аутентификации и деаутентификации. В процессе предварительной обработки выявляются серии действий пользователя (рис. 2).

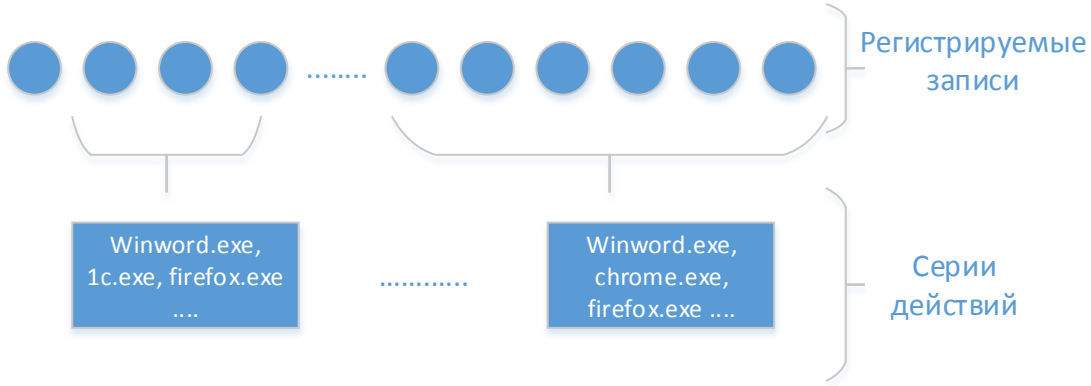

Рис. 2

После выявления серий действий пользователя происходит их анализ набором классификаторов. Итоговая оценка действия пользователя вычисляется по следующей формуле: $f = a_1 f_1 + a_2 f_2 + ... + a_n f_n$, где a_i - весовые коэффициенты, а f_i – функции выявления контекстных аномалий (классификаторы). Каждая из функций является бинарным классификатором и возвращает значения 0 (безопасное действие) или 1 (возможный случай утечки данных). Сумма всех весовых коэффициентов равна 1. Вычисленная итоговая оценка находится в диапазоне [0; 1]. В зависимости от порогового значения формируется уведомление о возможном случае утечки данных (рис. 3). Если процесс идентификации функционирует в режиме обучения, администратор указывает правильный результат проверки для проверяемого набора данных и вычисленных значений классификаторов.

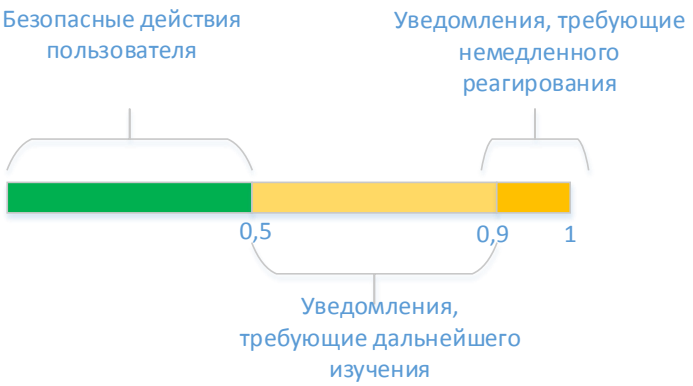

Рис. 3

Значения весовых коэффициентов, при которых обеспечивается минимальное количество ложных срабатываний, зависят от специфики работы сотрудников предприятия, включая продолжительность рабочего дня, возможность удаленной работы, служебные обязанности и др. Поэтому весовые коэффициенты классификаторов должны быть настроены вручную или с помощью механизма обучения с учителем. Обучающая пара состоит из значений классификаторов и результата проверки.

Способ вычисления итоговой оценки с помощью, представленной выше формулой имеет сходство персептроном без скрытого слоя – простейшей нейронной сетью, состоящей из входных нейронов и сумматора.

Заключение. В описанном в статье процессе идентификации утечек данных требуется проверка эффективности различных типов нейронных сетей. Для визуальной оценки такой проверки возможно использование метода анализа эффективности классификатора с помощью ROC-кривых.

ЛИТЕРАТУРА

1. 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 (Дата обращения: 15.02.2016)
2. Data Leakage Worldwide: The Insider Threat and the Cost of Data Loss // Cisco. URL: http://www.cisco.com/en/US/solutions/collateral/ns170/ns896/ns895/Cisco_STL_Data_Leakage_2008.pdf (Дата обращения: 15.02.2016)
3. Dorothy Elizabeth Robling Denning. 1982. Cryptography and Data Security. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
4. Ефремов А.А. Вычисление нечеткой вероятности безотказной работы систем с нечеткими параметрами моделей надежностие.–Доклады Томского государственного университета систем управления и радиоэлектроники. 2015. № 2 (36). С. 136-140.