Оглавление

[Введение 3](#_Toc101219612)

[1. Исследование модели поведения пользователя в веб-среде и поиск подходящих характеристик 4](#_Toc101219613)

[2. Анализ существующих средств защиты информации, исследующих шаблоны поведения человека в веб-среде 5](#_Toc101219614)

[3. Формирование исходных данных на основе проведения исследования для обучения алгоритмов искусственного интеллекта, их анализ и предобработка 6](#_Toc101219615)

[4. Обоснование и выбор алгоритмов для построения моделей машинного обучения и их разработка 7](#_Toc101219616)

[5. Тестирование работы реализованной программы по выявлению НСД и оценка полученных результатов с помощью метрик 8](#_Toc101219617)

[Заключение 9](#_Toc101219618)

# Введение

# 1. Исследование модели поведения пользователя в веб-среде и поиск подходящих характеристик

Qwe

Для разбиения потока событий на временные окна можно использовать следующие методики:

1) Разбиение потока событий на временные окна по количеству событий в окне (данную методику предпочтительнее использовать при умеренной работе пользователя с клавиатурой [47, 62]);

2) Разбиение потока событий на временные окна по длительности работы пользователя в рамках данного окна (данную методику предпочтительнее использовать, если пользователь активно взаимодействует с клавиатурой на протяжении всего времени своей работы за компьютером). Существующие исследования показывают, что в данном случае оптимальный размер временного окна – 2 минуты [37, 47]. Как в первом, так и во втором случае, можно строить окна фиксированной (по количеству событий либо времени) либо плавающей длины (задавая интервалы для возможных размеров окон). Также возможно использовать принудительное разбиение на временные окна при возникновении следующих ситуаций [37, 45, 47]:

• Длительная пауза в работе пользователя (в данном случае, события, пришедшие до и после паузы, будут принадлежать разным временным окнам);

• Смена пользователем рабочего процесса либо просматриваемой webстраницы (в данном случае, события из разных процессов или web-страниц будут принадлежать разным временным окнам).

# 2. Анализ существующих средств защиты информации, исследующих шаблоны поведения человека в веб-среде

Обзор современных индустриальных решений по динамической аутентификации пользователей на основе клавиатурного почерка. На сегодняшний день количество существующих на рынке индустриальных решений по аутентификации пользователей на основе динамики их работы с клавиатурой персонального компьютера постоянно увеличивается. И если раньше данные системы ограничивались только анализом ввода пары логин / пароль (рассматривалась исключительно статическая аутентификация), то сейчас активно развиваются системы, способные анализировать поведение пользователя за компьютером непрерывно.

BehavioWeb (BehavioSec)

Одним из наиболее известных коммерческих решений в области непрерывной фоновой аутентификации пользователей по клавиатурному почерку является продукт BehavioWeb компании BehavioSec [8, 11, 21]. Для анализа поведения пользователя в нем используются ритм и скорость набора текста, а также сила нажатия на клавиши. Программное обеспечение встраивается в web-сайт или web-приложение. Для этого используются поставляемая JavaScript-библиотека, а также J2EE-модуль, встраиваемый в web-сервер для осуществления процедуры аутентификации. Разработчики обращают внимание на то, что их решение анализирует изменение характеристик ввода пользователя с течением времени и периодически обновляет модель пользователя. Однако, используемые для этого алгоритмы не называются.

KeyTrac

Не менее популярным решением является продукт KeyTrac [8, 11, 22], позволяющий осуществлять в фоновом режиме как аутентификацию, так и идентификацию пользователей компьютера, основываясь на динамике их клавиатурного ввода. Данные пользователей (продолжительности нажатия на клавиши клавиатуры, а также продолжительности перескока между клавишами) записываются с помощью компонента KeyTrac Recorder и отправляются на сервер компании, где происходит их сравнение с построенной ранее моделью. При этом построение модели пользователя способно осуществляться на любом произвольно вводимом тексте, а не только при многократном вводе одних и тех же фраз. Для передачи данных используется предоставляемый KeyTrac API. Далее сервер возвращает свой вердикт в виде булевой величины true / false – соответствуют ли присланные тестовые данные рассматриваемому легитимному профилю либо нет. Для встраивания данного решения в web-сайт также предлагается использовать предоставляемую JavaScript-библиотеку. Разработчики системы утверждают, что их решение нечувствительно к смене используемого языка ввода, а также к смене используемого оборудования. Используемые для этого алгоритмы, а также методы построения модели и дальнейшей классификации не называются. Поскольку все данные динамики работы пользователей с клавиатурой анонимизируются, разработчики рекомендуют использовать свой продукт в том числе и в приложениях электронной коммерции.

KeystrokeID (ID Control)

Другим известным решением, осуществляющим непрерывный анализ динамики работы пользователей с клавиатурой компьютера, является продукт KeystrokeID компании ID Control [8, 23]. Для дальнейшего анализа здесь используются такие характеристики 19 набора, как промежутки времени между нажатием и отпусканием одной клавиши и промежутки времени между двумя последовательными нажатиями разных клавиш. Сбор данных, характеризующих динамику работы пользователей с клавиатурой, реализуется при помощи Java-апплета. Вначале модель обучается при вводе пользователем логина и пароля, далее происходит дообучение на данных фоновой работы пользователя. Используемые при этом алгоритмы не раскрываются. Данный продукт прост в использовании и используется для защиты компьютеров пользователей от сетевых атак. Scout Analytics Интересным решением является продукт компании Scout Analytics [8, 24], используемый для корректной аутентификации пользователей системы хранения и просмотра электронных публикаций. Он позволяет предотвратить написание отзывов о публикациях разными людьми с одного аккаунта, тем самым помогая составить более честную и объективную оценку публикаций. Алгоритмы, используемые в данном решении, запатентованы [25]. Для анализа используются код нажатой клавиши, тип события (нажатие / отжатие), а также время произошедшего события. Далее, в качестве рассчитываемых признаков выступают продолжительности нажатий на отдельные клавиши клавиатуры, продолжительности перескоков при нажатии двух или трех клавиш, а также различные статистики от данных величин (среднее значение, медиана, среднеквадратичное отклонение, максимум, минимум и т.д.). В качестве классификатора используется нейронная сеть. Модель поведения пользователя строится на основании его поведения за последние 60 дней (тем самым, каждый день происходит обновление модели). Также, для дополнительного статистического анализа используется информация об устройстве, с которого был осуществлен вход на сайт, и используемый при этом IPадрес.

TypingDNA

Одним из последних появившихся на рынке решений в области непрерывной аутентификации пользователей на основе клавиатурного почерка является продукт TypingDNA [26]. Разработчики предлагают использовать данное решение в учебном процессе: в течение семестра студенты работают за компьютерами во время занятий, в то время как в фоновом режиме происходит сбор данных динамики их работы с клавиатурой компьютера и обучение персональных моделей на этих данных. Во время контрольных мероприятий также производятся сбор пришедших от клавиатуры событий и сопоставление этих данных с построенными ранее моделями. Таким образом, удается выявить недобросовестных студентов, выполняющих контрольные работы несамостоятельно. Вторая область применения данного продукта – защита клиентов онлайн банковских систем от несанкционированного доступа посторонних лиц к их учетным записям. В качестве характеристических признаков выступают продолжительности нажатий, а также продолжительности перескоков для 44 наиболее часто используемых пользователем клавиш. Разработчики гарантируют высокое качество работы системы при условии, что для построения модели пользователем было напечатано на клавиатуре без продолжительных пауз не менее 100 символов. Сбор данных осуществляется при помощи Java-апплета. Разработчики предоставляют TypingDNA API и заявляют, что данное решение совместимо с большинством современных языков программирования. Используемые в данном продукте алгоритмы машинного обучения не называются.

KeystrokeDNA

Также набирает популярность коммерческое решение KeystrokeDNA [27], разработчики которого позволяют встраивать его в любое web-приложение и советуют использовать данный продукт для защиты компьютеров пользователей от сетевых вторжений. При этом, для анализа поведения пользователей используются ритм и скорость набора текста. Разработчики данного решения заявляют о стабильности его работы как при смене языка ввода, так и при смене используемой клавиатуры. Для установки и встраивания KeystrokeDNA пользователям предоставляется KeystrokeDNA API, а также собственная JavaScript-библиотека и подробная инструкция по их использованию. Более никакой информации о данном продукте в свободном доступе не предоставляется. Отметим, что помимо решений, основанных только на непрерывной аутентификации пользователей по динамике их работы с клавиатурой персонального компьютера (ноутбука), на рынке также распространены коммерческие продукты, осуществляющие многофакторную динамическую аутентификацию пользователей. Данные решения помимо клавиатурного почерка также анализируют динамику работы пользователей с мышью, с файловой системой и другие различные биометрические показатели. Разработчики данных продуктов (как анализирующих только клавиатурный почерк, так и решений, основанных на многофакторной аутентификации) заявляют, что точность аутентификации их решений достигает порядка 90%. Рассмотрим наиболее известные решения по непрерывной многофакторной аутентификации пользователей, использующие в качестве одной из своих анализируемых характеристик клавиатурный почерк пользователей.

BioCatch

Одним из решений, осуществляющих динамическую многофакторную аутентификацию пользователей, является широко известный продукт Biocatch [8, 28], анализирующий в режиме онлайн динамику работы пользователей с клавиатурой, мышью и web-ресурсами. При анализе клавиатурного почерка, помимо прочего данное решение позволяет определять, правшой или левшой является рассматриваемый пользователь, вычисляет размер его руки и также учитывает эти характеристики для определения легитимности пользователя. Всего данное решение вычисляет порядка 2000 характеристических признаков, из которых алгоритмами машинного обучения отбираются 20 наиболее значимых. Но какие именно это признаки и что за алгоритмы используются для их отбора – разработчики умалчивают. Данный продукт предлагается использовать для защиты web-приложений от несанкционированного доступа, а также для защиты пользователей онлайн банковских систем от доступа злоумышленников к их счетам. Аналитикам предоставляется широкий инструмент мониторинга, отображающий на различных графиках возможные риски и аномалии, а также формирующий высоко информативные отчеты о работе системы.

BioTracker (Plurilock)

Одним из наиболее известных решений по непрерывной многофакторной аутентификации пользователей является продукт BioTracker компании Plurilock [8, 11, 29]. Для того, чтобы с высокой точностью аутентифицировать пользователя, приложению необходимо обучиться на порядка 30 минутах его непрерывной работы с клавиатурой персонального компьютера. Дополнительно для улучшения качества построенной модели также анализируется работа пользователя с мышью. Необходимо заметить, что BioTracker периодически перестраивает модель пользователя, что позволяет учитывать изменение динамики его работы с клавиатурой и мышью в течение длительного периода времени. Разработчики рекомендуют использовать данный продукт как для защиты персональных компьютеров от сетевых атак, так и в приложениях электронной коммерции и государственных структурах – BioTracker может работать и как самостоятельное решение, так и встраиваться в готовые приложения пользователей. Продукт поддерживается всеми популярными сегодня операционными системами (Windows, Linux, Mac OS). Алгоритмы, используемые в данном решении, не называются.

CVMetrics, Tickstream (Intensity Analytics)

Продукты CVMetrics и Tickstream компании Intensity Analytics [8, 30] также широко используются для непрерывной аутентификации пользователей по динамике их 22 работы с клавиатурой и мышью. Данные решения нашли свое применение в сфере электронной коммерции, энергетике, судебно-медицинской экспертизе, государственных структурах – местах, где защита компьютеров от посторонних вторжений является одной из важнейших задач. Помимо анализа взаимодействия пользователей с мышью и клавиатурой, продукт Tickstream также дополнительно анализирует текстовые документы, с которыми работает пользователь. Для анализа клавиатурного почерка, и CVMetrics, и Tickstream для каждой нажатой клавиши записывают ее номер, соответствующее событие (нажатие / отжатие), а также временную метку, когда данное событие произошло. Разработчики замечают, что основным в их методике является именно анализ динамики работы пользователей с клавиатурой, потому что, в отличие от динамики работы с мышью, клавиатурный почерк является более стабильной характеристикой. Механизм аутентификации основан на вычислении различных статистик, определяющих степень похожести тестируемых данных с данными легитимного профиля. Однако, какие именно статистики вычисляются для этого в данных продуктах – разработчики умалчивают. Symantec VIP Высокое качество работы также демонстрирует продукт Symantec VIP [31], предлагающий свои решения как для защиты домашних компьютеров пользователей от несанкционированного доступа, так и для защиты систем электронной коммерции от сетевых вторжений. Динамическая многофакторная аутентификация пользователя осуществляется на основе анализа динамики его работы с клавиатурой и мышью. Процедура аутентификации производится спустя каждые 125 клавиатурных нажатий, при этом характеристики взаимодействия пользователя с мышью используются для уточнения построенной модели. Для принятия решения о легитимности рассматриваемого пользователя используется статистический подход. В случае обнаружения злоумышленника происходит блокировка системы, также система может сфотографировать злоумышленника. Symantec VIP поддерживается операционными системами Windows и Mac OS. В случае встраивания Symantec VIP в web-сайт, он будет корректно работать во всех популярных сегодня браузерах: Internet Explorer, Firefox, Chrome, Safari. Конкретные алгоритмы, используемые в данном решении, разработчиками не называются.

NuData Security (Mastercard)

Интересным решением является продукт NuData Security всемирно известной платежной корпорации Mastercard [32]. Решение NuData Security используется для обеспечения безопасности электронных переводов посредством данной платежной 23 системы. Аутентификация ведется на основании непрерывного анализа динамики работы пользователя с клавиатурой и мышью. Также при построении модели пользователя учитываются характеристики устройства, с которого осуществляется вход на сайт платежной компании, и время, проведенное пользователем на данном сайте. В случае обнаружения аномалии, все операции по счету блокируются. Разработчики отмечают, что получить доступ к их системе, копируя поведение легитимного пользователя, практически невозможно. Для обучения модели требуется достаточно небольшой промежуток времени – порядка одного посещения пользователем web-сайта компании с осуществлением платежной операции. Данное решение поддерживается операционными системами Windows и Mac OS. Разработчики NuData Security отмечают высокое качество работы их системы и низкий процент ложных срабатываний. При этом, используемые в данном решении алгоритмы машинного обучения не раскрываются.

NoPassword

Также набирает популярность недавно появившееся на мировом рынке коммерческое решение NoPassword [33], осуществляющее непрерывную многофакторную аутентификацию пользователей по динамике их работы с клавиатурой и мышью, а также с файловой системой компьютера или планшета. При этом, и для аутентификации на компьютере, и для аутентификации на планшете используются единые алгоритмы, которые демонстрируют высокое качество работы системы в том числе и при использовании сессий удаленного рабочего стола. Для анализа клавиатурного почерка данное решение вычисляет продолжительности нажатий и продолжительности перескоков при нажатии клавиш клавиатуры. Разработчики предлагают использовать данный продукт для защиты как домашних, так и служебных компьютеров пользователей от несанкционированного доступа злоумышленников. Для встраивания данного решения пользователям предоставляются удобные в использовании NoPassword API и NoPassword SDK. Также администраторам системы предлагаются обладающий широкими возможностями инструмент мониторинга активности пользователей в режиме онлайн и возможность гибкой настройки политик доступа. Данное решение поддерживается всеми наиболее популярными сегодня браузерами и операционными системами. Используемые в данном продукте алгоритмы машинного обучения не называются.

# 3. Формирование исходных данных на основе проведения исследования для обучения алгоритмов искусственного интеллекта, их анализ и предобработка

Таблиц соответствия кодов клавиш и букв английского алфавита

Постобработка векторов признаков Большинство современных методов машинного обучения сильно чувствительны к шкалированию входных данных. Для того, чтобы решить данную проблему, в существующих работах вводится шаг нормализации [12, 36, 66, 71, 72] (или ее подвида – стандартизации [37, 38]) данных. Нормализация данных выполняет отображение исходного пространства признаков таким образом, что их значения после этого преобразования принадлежат отрезку [0, 1]. После же процедуры стандартизации каждый признак имеет среднее значение, равное нулю, и дисперсию, равную единице. Благодаря этому, ни один из используемых признаков не будет доминировать в строящейся модели, и классификатор корректно учтет каждый из них. Суть стандартизации заключается в осуществлении следующего преобразования: 𝑥 → 𝑥−𝐸𝑥 √𝐷𝑥 , (1) где 𝑥 – значение рассматриваемого признака, а 𝐸𝑥 и 𝐷𝑥 – математическое ожидание и дисперсия данного признака соответственно, полученные на обучающем наборе.

# 4. Обоснование и выбор алгоритмов для построения моделей машинного обучения и их разработка

Список доступных URI (таблица)

# 5. Тестирование работы реализованной программы по выявлению НСД и оценка полученных результатов с помощью метрик

# Заключение