МАШИННОЕ ОБУЧЕНИЕ В БАНКИНГЕ: ПЛЮСЫ И МИНУСЫ

ДАНИЛ ЗАКОЛДАЕВ, К. Т. Н.
АЛИСА ВОРОБЬЕВА, К. Т. Н.

В банковской сфере все чаще разрабатываются и внедряются в эксплуатацию системы биометрической идентификации человека с применением инструментария машинного обучения, что способствует в том числе появлению определенных рисков. В статье рассмотрен ландшафт таких рисков, на ряде примеров показаны уязвимости подобных систем, а также способы их эксплуатации. Судя по приведенным доводам, в арсенал компетенций современного специалиста по защите информации в России необходимо добавить навыки, связанные с машинным и глубоким обучением, когнитивными технологиями.

В 2018 г. объем безналичных операций практически вдвое превысил число операций по снятию наличных средств с банковских карт. Это косвенно подтверждает то, о чем говорят уже давно, — деньги становятся «электронными». Соответственно, и преступления переходят в киберпространство: компьютерное мошенничество, фишинг, социальная инженерия и самое простое — преступления, связанные с кражей платежной информации. Раньше оценку параметров — какой клиент пришел в банк, не мошенник ли он, совершается ли операция под давлением третьего лица, находится ли человек в состоянии стресса или опьянения, склонен ли к совершению необдуманных операций — осуществлял сотрудник банка. В этом ему помогали личная внимательность и опыт. Сегодня же все банковские операции переходят в цифровую среду, в том числе мобильные банки, и на первый рубеж защиты вместо естественного интеллекта выходит интеллект искусственный. При этом некоторые решения, принятые специально обученной программой, объяснить не представляется возможным. Также при активном внедрении во многие сферы жизни человека искусственный интеллект сегодня демонстрирует уязвимость к большому числу атак, и исследования в этой области только начинаются. Однако при корректном и аккуратном использовании современных методов машинного и глубокого обучения в решении названных проблем можно добиться хороших результатов.

В декабре Сбербанк поддержал инициативу Центрального банка по удаленной идентификации клиентов и начал сбор биометрических данных для Единой биометрической системы (ЕБС) Единой системы идентификации и аутентификации (ЕСИА). В качестве ключа в этой системе будут использоваться голос и лицо человека. Помимо «классической» проблемы защиты персональных и биометрических данных — обеспечения их конфиденциальности, целостности и доступности, появились и достаточно новые и интересные проблемы. Первая — это так называемая liveness detection — способность детектировать подделку вместо живого человека. А вторая — выявление аномалий в поведении пользователей, например для обнаружения мошенников во время проведения удаленных операций или для определения того, что какая-либо операция совершается без согласия клиента либо под давлением.

БИОМЕТРИЧЕСКИЕ ИДЕНТИФИКАЦИЯ И АУТЕНТИФИКАЦИЯ

Рассмотрим существующие подходы к проведению атак, приводящих к возможности «подделки клиента» банка, т. е. спуфинг-атак на системы биометрии.

Современные биометрические системы идентификации/аутентификации обладают достаточно высокой точностью, процент верного распознавания приближается к значению 0,92 и выше (тем не менее есть проблемы, например с распознаванием близнецов). Сегодня эти системы работают без участия человека, что делает их уязвимыми к так называемым спуфинг-атакам. Простейший пример такой атаки — предъявление фотографии вместо реального человека. Эта фотография может быть даже найдена злоумышленником на просторах Интернета. Способов реализации таких атак множество — от простой фотографии и специального макияжа до специально созданных 3D-масок. Стоит отметить, что этот специальный макияж создается с использованием генетических алгоритмов, а нейронные сети (сети 3D-GAN) способны проектировать трехмерную модель лица по фотографии. Та же ситуация и с системами голосовой биометрии — синтез речи осуществляется по короткому доступному фрагменту. Еще одним интересным примером могут быть вредоносные заплатки, которые, например, при нанесении на одежду вовсе не позволяют распознать наличие человека в кадре.

«Фокусы» такого рода сегодня довольно распространены, и часто появляются новые публикации, связанные с обманом систем биометрической идентификации и аутентификации (по сгенерированным отпечаткам пальцев, по подложке фотографий вместо реального пользователя). Такие атаки используются достаточно давно и хорошо известны специалистам по безопасности, это те же «брут форс» или «маскарад», но то, как и где они реализуются, — совершенно новая и сложная область.

Методы защиты (liveness detection) от таких атак также разнообразны — это и применение видеофрагментов вместо фотографий, и просьбы к пользователю совершить какое-либо случайное действие (например, моргнуть левым глазом), и использование дополнительных факторов (термограммы лица, текстуры и рельефа кожи, 3D-модели лица). Арсенал средств обеспечения безопасности

зависит только от того, сколько разработчики готовы потратить на усовершенствование системы. Но все же отметим, что здесь тоже действует классическое правило: система должна быть основана на секретности ключа, а не алгоритма. Алгоритм рано или поздно злоумышленники раскроют, особенно учитывая современные возможности.

СОСТЯЗАТЕЛЬНЫЕ АТАКИ НА БИОМЕТРИЧЕСКИЕ СИСТЕМЫ

Известно, что большое число биометрических систем основано на нейронных сетях, а нейронные сети уязвимы, например, к состязательным атакам (следует учитывать, что это всего лишь один из видов уязвимостей). Состязательная атака (adversarial attack) — это способ обмануть нейронную сеть с целью изменения «ответа» системы на необходимый злоумышленнику. Состязательный пример (adversarial samples) — некий пример тестовых данных, в который внесены искажения, приводящие к некорректному распознаванию. Такими искажениями, в частности, могут служить добавление шума или изменение нескольких пикселей (а иногда даже одного) на исходной фотографии, вследствие чего человек на ней будет распознан некорректно. Важно, что такие искажения незаметны человеческому глазу, а «видит» их только нейронная сеть.

Эти атаки возможны из-за разных механизмов распознавания образов. Человек видит некую картину, а нейронная сеть «видит» только пиксели, расположенные в определенном порядке. Соответственно, для нас отличие одного пользователя от другого состоит в разрезе глаз, форме лица и оттенке кожи. Для нейронной сети — это только особенности в яркости определенных пикселей или расположении опорных точек, которые четко ассоциированы с определенным человеком. Те особенности и корреляции, которые она обнаружила в обучающих данных и которые, однако, могут быть ошибочными. Возможным решением этой проблемы сегодня считают отбор идентификационных признаков «в ручном режиме», что позволяет убрать ложные корреляции и использовать только действительные.

При реализации атак на системы биометрической идентификации, основанной на методах машинного обучения, злоумышленник может преследовать различные цели. От простого сбора информации о модели, на которой основана система, до модификации системы путем подмены обучающих данных.

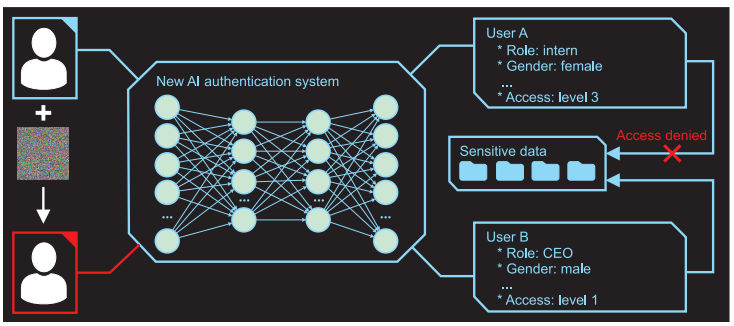

РИС. 1. ◄
Пример простой атаки в обход

Приведем основные сценарии действий злоумышленников.

Начнем с наименее опасных атак — разведывательных. Для злоумышленника система идентификации является «черным ящиком», поэтому, для того чтобы осуществлять дальнейшие злонамеренные действия, ему необходима информация о системе, об алгоритме, лежащем в ее основе, о структуре обучающих данных.

Наиболее распространенный тип атаки — атака в обход (evasion attack). Злоумышленнику известны «ответы» системы. У него есть возможность использовать сгенерированные состязательные примеры. Например, добавляя незначительный шум к изображению лица и анализируя ответы системы, он может подобрать такие искажения, которые для нейронной сети сделают пользователя А неотличимым от пользователя Б. Схематичное изображение этого вида атаки приведено на рис. 1.

Наконец, самый опасный тип атак — отравляющая атака (poisoning attack), под ней подразумевается «загрязнение» обучающих данных специально сгенерированными образцами, что приводит к неверному построению модели.

Можно выделить два ключевых типа атак на системы, основанные на методах машинного обучения:

- Атаки во время обучения. Злоумышленник пытается влиять на модель путем непосредственного изменения обучающего набора данных. Такие атаки схематически изображены на рис. 2.
- Атаки на этапе тестирования. Злоумышленник не вмешивается в целевую модель, а заставляет ее выдавать необходимый ему результат путем искажения тестовых данных.

В свою очередь, принято считать, что есть три основные стратегии атак во время обучения модели, основанные на состязательных возможностях:

- Внедрение данных. Злоумышленник не имеет никакого доступа к обучающим данным, а также к самому алгоритму. Он может скомпрометировать целевую модель, вставляя состязательные примеры в обучающий набор данных.
- Модификация данных. Злоумышленник не имеет доступа к алгоритму обучения, но имеет полный

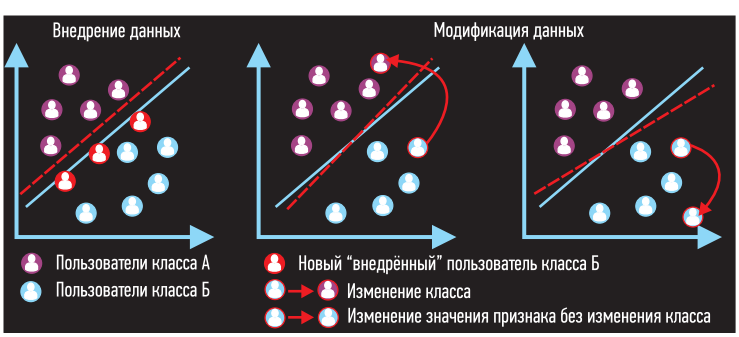

РИС. 2. ◄
Атаки во время обучения

доступ к обучающим данным. Он изменяет целевую модель, модифицируя данные до того, как они будут использованы для обучения.

- Логическое искажение. У злоумышленника есть возможность вмешиваться в алгоритм обучения.

Очевидно, что весьма трудно разработать стратегию борьбы со злоумышленниками, которые могут изменить логику обучения, тем самым контролируя саму модель, — на главную позицию выходит внутренняя корпоративная безопасность, особенно среды разработки и обучения.

Не менее серьезной задачей является разработка методологий модификации обучающей выборки на этапе ее передачи внешним разработчикам с целью проведения машинного обучения — таким образом, чтобы чувствительные данные стали недоступны, а внутренние связи при этом не нарушились.

Состязательные атаки на этапе тестирования не вмешиваются ни в данные, ни в целевую модель, а заставляют ее выдавать некорректный результат. Эффективность реализации таких атак зависит от количества информации о модели, доступной для злоумышленника. Атаки на фазу тестирования могут быть классифицированы как атаки «белого» и «черного ящика».

При атаке «белого ящика» злоумышленник обладает полной информацией о модели. Например, он знает тип используемой нейронной сети и количество слоев, алгоритм, применяемый в процессе обучения (в частности, оптимизатор градиентного спуска). Он также владеет информацией о параметрах обученной модели. Злоумышленник использует доступные сведения для определения пространства признаков, в котором модель может быть уязвима, то есть участков, где модель имеет высокую частоту появления ошибок. Затем модель атакуется с применением методов генерации состязательных примеров. Доступ к внутренним весам модели при атаке «белого ящика» соответствует очень сильной состязательной атаке.

При атаке «черного ящика», напротив, не предполагается, что у атакующего есть какие-либо знания о модели. Суть этих атак сводится к тому, что злоумышленник пытается натренировать собственную модель, имитирующую цель его атаки. Далее он создает состязательные примеры на своей модели, используя стратегию атаки «белого ящика», и применяет их на целевой модели. То есть фактически он заставляет целевую модель классифицировать данные неверно, тем самым нарушая результаты ее работы.

Виды атак «черного ящика» различаются по тому, что именно доступно злоумышленнику, как он получает обучающие данные.

В неадаптивных атаках черного ящика (non-adaptive black-box attack) злоумышленник имеет полный доступ к обучающим данным. В случае двух других типов атак он фактически пытается восстановить обучающие данные. При адаптивной атаке «черного ящика» злоумышленник не имеет доступа к обучающим данным, но может получить доступ к «ответам» системы. Имея некий неразмеченный набор данных (найденные им в открытом доступе фотографии сотрудников компании или клиентов банка) и делая запросы к модели, он получает возможность произвести разметку этих данных. Иногда при атаке «черного ящика» злоумышленник не имеет ни доступа к обучающим данным, ни возможности вводить собственные тестовые примеры. Однако он может собирать пары «ввод-вывод» из целевого классификатора — это так называемая строгая атака «черного ящика» (strict black-box attack). Такая стратегия довольно успешно работает на больших наборах пар входов-выходов. После долгого наблюдения злоумышленник также получает размеченную выборку.

БОРЬБА С МОШЕННИКАМИ В ВИРТУАЛЬНОМ ФИНАНСОВОМ ПРОСТРАНСТВЕ

Помимо биометрических механизмов определения личности клиента банка, интересной задачей сегодня является выявление аномалий в поведении пользователей. Эта задача относится к проблеме нахождения паттернов данных, не соответствующих ожидаемому поведению. Обнаружение аномалий широко используется в информационной безопасности, а в банковской сфере может применяться для обнаружения вторжений и несанкционированного доступа, мошенничества при проведении банковских транзакций и др.

Создание шаблонов типичного поведения клиентов при выполнении банковских операций, взаимодействии с банковскими приложениями или веб-сайтом позволяет выявлять аномалии и отклонения от штатных действий. Применение искусственного интеллекта при этом основано на прецедентах, но не на предыдущих случаях атак, а на типичном поведении пользователя. Представим, что злоумышленник получил доступ в онлайн-банк клиента. После этого он начинает совершать некие нетипичные действия, которые существенно отличаются от обычного поведения пользователя. Следовательно, в ходе выполнения этих действий злоумышленник скомпрометирует себя перед системой, основанной на методах машинного обучения.

На этапе борьбы с банковским мошенничеством помогают методы, созданные на стыке совершенно разных областей — машинного обучения, лингвистики, информационной безопасности. Например, выявление мошеннических чат-ботов. Эти боты действуют с взломанных аккаунтов в мессенджерах и социальных сетях, имитируя реального человека. Они ведут диалоги с пользователями, уговаривая и вынуждая «помочь другу» и совершить банковский перевод, — это современный вариант реализации методов социальной инженерии.

Также всем давно известно такое явление, как фишинг (это «рыбалка», когда злоумышленник забрасывает крючок с привлекательной наживкой). Целью такой рыбалки является совершение клиентом банковского перевода на счета злоумышленников или ввод платежной информации. Единственным вариантом защиты раньше было повышение грамотности в области безопасности. Банки пытались научить своих клиентов не попадаться на крючок через специально созданные тесты, опросы и даже игры. Сегодня выявлять такие письма могут методы, основанные на машинном обучении.

Приведенные примеры атак на механизмы биометрической аутентификации выявлены в результате анализа ландшафта рисков, порождаемых парадигмой машинного обучения. Эти риски необходимо учитывать при массовом внедрении биометрии на основе машинного обучения, сочетая их с традиционными методами аутентификации при малейшем подозрении на ошибки первого рода

(злоумышленник успешно выдает себя за авторизованного пользователя). При абсолютном доверии к таким методам возможно возникновение ущербов (финансовых, имиджевых и др.), значительно превышающих положительный эффект от их реализации.

КАКИЕ СПЕЦИАЛИСТЫ НУЖНЫ ОТРАСЛИ

Плавное сокращение операций с использованием наличных, рост числа пользователей банковских онлайн-сервисов, упрощение процедур совершения операций для пользователя — все это требует повышенного качества защиты и безопасности. Которое, однако, должно обеспечиваться в фоновом, абсолютно незаметном для пользователя и клиента режиме. Более того, необходимый уровень защиты должен быть создан с использованием хоть и новейших технологий, но абсолютно прозрачных и досконально изученных специалистами по информационной или кибербезопасности. Сегодня требуются «эргономичные» методы обеспечения безопасности, ориентированные в равной мере на качество защиты и на удобство пользователя. Создание дополнительных сложных или нестандартных для пользователя процедур ведет к потере клиентов, что является абсолютно недопустимым. В связи с этим современный специалист в области кибербезопасности, особенно в банковских и финансовых организациях, должен не только уметь настраивать и использовать современные инструменты и средства защиты, но и досконально знать и понимать основы используемых технологий, природу угроз. Строго необходимыми являются знания в области современных методов машинного обучения и анализа данных, поскольку именно они сегодня могут служить базой для создания новых эффективных средств защиты. Вместе с тем специалист должен использовать эти технологии не только как набор инструментария, он должен быть способен сам создавать подобные методы и алгоритмы. Необдуманное применение готовых технологий и решений, без глубокого понимания процессов, часто приводит к реализации угроз. Технологии меняются стремительно, и поэтому сегодня основная задача подготовки специалистов по информационной безопасности — дать им способность к непрерывному профессиональному развитию, заложив основы в виде глубокого понимания современных технологий.

Ни для кого не секрет, что ключевым аспектом в подготовке специалистов является тесная связь с будущим работодателем. Формирование запроса на требующиеся компетенции, экспертиза образовательного процесса, проведение занятий специалистами организации и даже выполнение студенческих исследований в областях, с которыми будут связаны их профессиональные задачи, — это необходимый, хотя в то же время идеальный и часто труднодостижимый минимум.

Одним из примеров решения проблемы, связанной с обучением нужных специалистов, является корпоративная магистерская программа «Кибербезопасность в банковской сфере», которая реализуется в Университете ИТМО, на факультете безопасности информационных технологий (БИТ), совместно с ПАО «СберБанк». Отличительная особенность программы заключается в том, что подготовка магистров основана на решении реальных «боевых» задач совместно с действующими специалистами по информационной безопасности банка.