Конспект лекции

А.Е. Сулавко

Защищенное исполнение искусственного интеллекта

**Защищенное исполнение нейросетевых алгоритмов на базе линейных нейронов**

**ЗАЩИЩЕННОЕ ИСПОЛНЕНИЕ НЕЙРОСЕТЕВЫХ АЛГОРИТМОВ НА БАЗЕ ЛИНЕЙНЫХ НЕЙРОНОВ**

Как уже упоминалось в предыдущем разделе, частным случаем преобразователя образов в код является *нейросетевой преобразователь биометрия-код (НПБК).* Согласно ГОСТ Р 52633.0-2006, нейросетевой преобразователь биометрия-код представляет собой заранее обученную искусственную нейронную сеть с большим числом входов и выходов, преобразующую частично случайный вектор входных биометрических параметров "Свой" в однозначный код криптографического ключа (длинного пароля) и преобразующая любой иной случайный вектор входных данных в случайный выходной код.

Стоит отметить, что в свою очередь НПБК строится исходя из концепции построения *преобразователя биометрия-код* (ПБК). На текущий момент сложилось два основных подхода к построению ПБК: на основе нечеткого экстрактора, в рамках которого для дополнительной защиты применяется криптография (стандарты ISO/IEC 19792:2009, 24761:2009 и 24745:2011), и нейросетевой подход, поддерживаемый серией ГОСТ Р 52633, к настоящему времени не имеющих мировых аналогов [4].

Итак, говоря о НПБК, в первую очередь, целесообразно опираться на ГОСТ Р 52633.5-2011 (самостоятельное изучение). В данном стандарте отражен и подробно рассмотрен процесс автоматического обучения нейросетевых ПБК, спроектированных на базе так называемых *линейных нейронов* (т.е. классических нейронов), в НПБК представляющих собой сумматор нескольких биометрических параметров, на выходе которого подключена нелинейная пороговая функция с двумя выходными состояниями "0" и "1" [5].

Нейросетевой ПБК строится персонально для каждого субъекта, при этом формируется искусственная нейронная сеть (ИНС), количество входов которой равно числу признаков, а количество выходов — длине его личного ключа. Каждый нейрон последнего слоя генерирует один бит. Нейронная сеть обучается на биометрических образах пользователя и образах «Чужих», чтобы вырабатывать ключ субъекта при поступлении на вход его биометрического образа. Так как НПБК должен откликаться на примеры образа «Свой» одним кодом, а на другие образы «Чужой» - случайными кодами , то код «Свой» и случайные коды образов «Чужие» не совпадают между собой по состояниям части разрядов. В качестве меры близости кодов образов «Чужие» к коду «Свой» можно использовать *расстояние Хэмминга*.

Хорошо обученная нейронная сеть не нуждается в дополнительной корректировке выходов (ответа). Обучение нейросетевого ПБК должно быть абсолютно устойчивым, при этом объем обучающей выборки «Чужие» может быть сколь угодно большим. Разработчик биометрической системы может заготовить репрезентативную выборку «Чужие» заранее и использовать ее для обучения каждого ПБК. Однако число примеров образа «Свой» должно быть малым (по ГОСТ Р 52633.5-2011 достаточно 11 примеров), нельзя заставлять пользователя сотни раз вводить биометрический образ. Это обстоятельство накладывает существенные ограничения на архитектуру ИНС, используемую в основе ПБК [4].

Так, в основу стандартов ГОСТ Р 52633 легли так называемые *«широкие» нейронные сети*, представляющие собой большие сети из малого числа слоев (одного или двух). Важным отличием «широких» сетей является процедура автоматического и абсолютно устойчивого послойного обучения (без использования алгоритма градиентного спуска).

После оцифровки биометрический образ преобразуется в вектор признаков *ā = {a1, ..., aN}* фиксированной длины, таки образом образуя пространство признаков.

Первый слой классической «широкой» нейронной сети, обучаемой по алгоритму ГОСТ 52633.5, обогащает входные данные, второй — играет роль кодов, исправляющих ошибки. При этом классический нейрон такой сети базируется на функционале:

(1)

И пороговой функции активации

(2)

Модули весов нейронов первого слоя вычисляются по формуле

(3)

где *y* — отклик нейрона на образ «Свой» или «Чужой»;

*n* — количество входов нейрона;

*aj* — значение j-го признака (входа нейрона);

*f(y)* — ответ нейрона;

— порог активации нейрона;

и — математическое ожидание и среднеквадратичное отклонение значений j-го признака для образа «Свой»;

и — аналогичные показатели образов для «Чужих».

Таким образом, сеть из нейронов (1) с функцией активации (2) после обучения представляет собой нейросетевой ПБК. Ответ нейросетевого ПБК складывается из битовых значений на выходах нейронов (путем их конкатенации).

Если нейрон настроен на выход «1» при поступлении образа «Свой», то знак весового коэффициента выбирается исходя из правила: «+» при < , иначе «–». Если нейрон настраивается на «нулевой» бит, то знаки инвертируются. Параметры , , и после обучения удаляются, чтобы не компрометировать эталон. Остаются таблицы связей и весов µ, из которых нельзя непосредственно вычислить . Параметры обученных нейронов (связи и веса) называют *нейросетевым контейнером*.

Порог нейрона обычно настраивается исходя из откликов нейрона на обучающие примеры «Чужой» по правилу

(4)

где — единый для всех нейронов эмпирически подбираемый коэффициент, влияющий на баланс между FRR и FAR. В теории такой способ должен дать вероятность ошибки «ложного принятия» нейроном образа «Чужой», приближенно равную 0,5. Если допустить, что выходы нейронов независимы, каждый добавляемый нейрон, который настраивается по формуле (4), должен снижать FAR примерно в 2 раза. *Информационная энтропия (или просто энтропия)* ответов нейросетевого ПБК на образы «Чужих» в этом случае должна быть близка к длине ключа. Но в действительности, чем больше входов у нейронов, тем выше корреляция между их выходами и FAR, но чем больше информации поступает на вход каждому нейрону, тем ниже FRR. Действует и обратная логика: чем больше нейронов, тем ниже FAR, но выше FRR.

Процедура обучения «широкой» ИНС является однонаправленной и не подразумевает обратной разработки. Однако восстановление биометрического образа и личного ключа пользователя из таблиц нейросетевых функционалов все же возможно.

Контролируя допустимое число ошибочных бит в ответе «широкой» сети, можно балансировать FRR и FAR (например, применяя корректирующие коды или второй слой нейронов для исправления нескольких неверных бит). Однако эта возможность одновременно является уязвимостью. Хакер может собрать большую базу примеров произвольных паролей, воспроизведенных различными подписантами («Чужими»), и оценить среднюю стабильность ответов ПБК для каждого подписанта («Чужого») по формуле

(5)

где k — номер подписанта («Чужого»); L — количество нейронов; l — номер нейрона; Pl(1) — вероятность (или относительная частота) появления «единицы» (можно заменить на Pl(0)) в l-м разряде ответа ПБК (в выходе l-го нейрона) на примеры образа k-го «Чужого». Оценка относительных частот Pl(1) и Pl(0) может проводиться на основании нескольких образов подписанта.

Данная атака снижает количество вариантов перебора на несколько порядков. Даже если злоумышленник не обладает примерами подделок и какой-либо информацией о пароле пользователя-жертвы, данная атака вполне осуществима (в этом случае нарушителю потребуется гораздо больше времени).

Также, важно отметить, что образы «Чужие» обладают так называемой «симметрией» стабильности ответов относительно образа «Свой» (свойством «симметрии»). Это означает, что стабильность ответов ПБК при предъявлении образов «Чужих» возрастает, но не только если ответы близки (в метрике Хэмминга) к ключу пользователя, но и если они близки к инверсии ключа (инверсный код возникает, если все биты ответа ПБК являются ошибочными). Инверсный код можно обратить и получить ключ пользователя. Из этого следует, что у каждого образа «Свой» в нейросетевом логическом базисе существует его инверсия. Если на вход классического нейросетевого ПБК (обученного по ГОСТ 52633.5) подать инверсию образа «Свой», то на выходе ПБК появится инверсный ключ, который можно обратить. Данное свойство позволяет ускорить процедуру направленного перебора биометрических образов в 2 раза (осуществляя одновременно поиск наиболее близкого и наиболее дальнего образа «Чужого» относительно образа «Свой»).

В связи со всеми вышеперечисленными факторами, имеет смысл осуществлять защиту нейросетевых контейнеров, в том числе от атак извлечения знаний. Для защиты таблиц нейросетевых функционалов нужно применять *механизм защищенного нейросетевого контейнера (ЗНК)*. Сначала необходимо выстроить нейроны в цепочку. После обучения ПБК таблицы каждого нейрона шифруются наложением гаммы, представляющей собой контрольную сумму выходов всех предыдущих нейронов в цепочке:

6)

где *tablesl* — таблицы параметров соответствующего нейрона; *hash()* — криптографическая хеш-функция (например, md5); *pass* — пароль, который является опциональным и служит для дополнительной (двухфакторной) защиты; *bit*l — выход, на который настраивается l-й нейрон в цепочке.

При обработке биометрического образа нейросетевым ПБК в режиме ЗНК происходит процесс «распаковки» нейронов — параметры каждого следующего нейрона в цепочке дешифруются по той же формуле (6). Для получения на выходе ПБК верного ключа пользователя требуется, чтобы все нейроны «проголосовали» правильно. Если хотя бы один нейрон в цепочке выдаст ошибочный бит, это повлечет неверную дешифровку параметров всех последующих нейронов. В свою очередь последующие нейроны будут давать случайные выходы, и возникнет эффект хеширования биометрического образа «Чужого». В итоге ответы нейросетевого ПБК становятся случайными, их энтропия возрастает. При этом важен тот факт, что FRR и FAR в режиме ЗНК не меняются при пороге принятия, равном нулю. Стабильность ответов ПБК при поступлении на вход образов «Чужих» становится низкой и перестает возрастать, если образ «Чужого» близок к образу «Своего». Однако режим ЗНК все же накладывает ограничения: балансировать FRR и FAR, корректируя несколько ошибочных ответов ПБК, становится невозможно.

**Список источников**