НЕЙРОСЕТЕВАЯ МОДЕЛЬ ДЛЯ ОБНАРУЖЕНИЯ АНОМАЛИЙ

ШуршевФ.В., д.т.н. профессор,
ТагировР.Р., магистрант,
АГТУ, г. Астрахань, Россия

Аннотация. Рассмотрена гибридная нейросетевая архитектура, которая может бы автоматически обнаруживать отклонения в поведении сетевого трафика на основе его структуры. Предложена комбинированная оценка аномальности, учитывающая ошибку реконструкции и реакцию на изменение внутренних признаков. Для повышения точности введена комбинированная функция потерь, регулируемая параметром, позволяющим настраивать вклад каждой компоненты.

Ключевые слова: нейронные сети, обработка информации, обнаружение аномалий, информационная безопасность.

Одной из наиболее острых проблем в области информационной безопасности является выявление аномалий в сетевом трафике. Основная причина заключается в том, что классические системыIDS ориентированы на заранее известные векторы атак и не способны адекватно реагировать на ранее неизвестные виды угроз. В связи с этим актуальной становится задача разработки методов, способных обнаруживать аномалии на основе поведения сетевого трафика. Перспективным направлением в решении данной задачи является использование нейронных сетей, в частности автокодировщиков. Эти модели способны выявлять скрытые закономерности, характерные для нетипичного поведения, что позволяет повысить точность обнаружения угроз.

Модель нейронной сети

В состав предлагаемой нейронной сети входят три основные компонента: энкодер, декодер и дискриминатор. Каждый из них реализован в

виде многослойного перцептрона с использованием полносвязных слоёв с функцией активации LeakyReLU с исключением равным 20%, а также – гиперболического тангенса и сигмоиды в выходных слоях. Кодировщик состоит из трёх последовательных слоёв. Первый слой содержит 64 нейрона и имеет функцию активации LeakyReLU с параметром отрицательного наклона равным 0.2. Для предотвращения переобучения применяется исключение с шансом 20%. Второй и третий слои имеют по 32 нейрона и также используют функцию активации LeakyReLU с тем же параметром отрицательного наклона и исключением 20%. Декодер симметрично повторяет структуру энкодера и имеет выходной слой, который состоит из 80 нейронов с функцией активации tanh. Это позволяет ограничить значения выходных признаков в диапазоне -1 до 1.

Дискриминатор – отдельный модуль, задача которого отличать настоящий сетевой поток от сгенерированного автокодировщиком. Первые два слоя содержат по 64 нейрона и используют функцию активации LeakyReLU с более высоким параметром отрицательного наклона 0.3. Для обоих слоёв применяется исключение с шансом 20%. Третий слой имеет 32 нейрона, ту же функцию активации и то же исключение. Завершающий слой дискриминатора – это одиночный нейрон с функцией активации сигмоида, который выдаёт значения от 0 до 1, который интерпретируется как вероятность того, что вектор является настоящим, а не созданным генератором. В качестве оптимизатора в модели используется Adam [1].

Функция потерь

В генеративно-состязательных сетях генератор обучается максимизировать вероятность того, что дискриминатор не распознает генерированные данные. Эта цель [2] описывается следующим образом:

$$\max_G E_{z\sim p(z)}[logD(G(z))] \quad (1)$$

Однако данный подход зачастую приводит к переобучению: генератор воспроизводит ограниченный набор шаблонов, успешно обманывающих

дискриминатор и теряет способность к обобщению. Это значительно снижает пригодность такой модели для задачи обнаружения аномалий, где важна чувствительность к отклонениям.

Для решения данной проблемы в исследовании применяется подход сопоставления признаков [3]. Вместо использования итогового скалярного ответа дискриминатора, сравниваются внутренние признаки. Это позволяет добиться более стабильного обучения генератора.

Функция потерь дискриминатора при использовании подхода сопоставления признаков записывается так:

$$L_D = \Sigma||f(x) - f\big(G(z)\big)|| \quad (2)$$

где f(x) – признаки реального образца, f(G(z)) – признаки сгенерированного образца.

В отличие от стандартного подхода, такая функция потерь заставляет генератор воспроизводить внутреннюю структуру реальных данных. Общая функция потерь модели обнаружения аномалий определяется как сумма ошибки реконструкции и ошибки сопоставления признаков:

$$L = L_R + L_D \quad (3)$$

где $L_R = \Sigma|x - G(x)|$ – ошибка реконструкции [4], представляющая разницу между сжатыми и восстановленными из представления данными. Использование совмещённой функции потерь может [5] повысить чувствительность модели.

Оценка аномальности

Процесс обнаружения аномалий на этапе эксплуатации основывается на вычислении комбинированной функции аномальности, отражающей степень отклонения входных признаков от нормального поведения. В основе данного подхода лежит модифицированная функция потерь, рассмотренная ранее. Формально функция аномальности потока X определяется следующим образом:

$$A(x) = \lambda L_d(x) + (1 - \lambda) L_g(x) \quad (4)$$

где: L_d – функция потерь дискриминатора, L_g – ошибка восстановления, $\lambda \in [0,1]$ – вес, задающий вклад ошибки восстановления в итоговую оценку.

При обработке нового потока низкое значение функции аномальности указывает на то, что он хорошо реконструируется и его внутренние признаки соответствуют внутренним признакам нормальных данных. Напротив, высокое значение функции аномальности свидетельствует о наличии нетипичных признаков, что может указывать на потенциальное нарушение или атаку. Дополнительно, компонент L_d может быть использован для интерпретации результатов [6]. Анализ распределения ошибки реконструкции по признакам позволяет локализовать те характеристики потока, которые внесли наибольший вклад в итоговую оценку. Это означает возможность не только обнаружить аномалию, но и обнаружить его источник.

Заключение

В работе предложена нейросетевая модель для обнаружения аномалий в сетевом трафике на основе автокодировщика и генеративно-состязательной сети.Разработана комбинированная функция аномальности и функцию потерь, учитывающие как ошибку реконструкции, так и отклонения во внутренних признаках.Эксперименты показали стабильную работу модели на реальных данных и позволили сделать вывод о возможности её практического применения.

Литература

1. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)
2. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S.,Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM 63(11), 139–144 (2020)

3. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: Advances in Neural Information Processing Systems. (2016) 2226–2234
4. Кингма Д. П., Веллинг М. An Introduction to Variational Autoencoders // Foundations and Trends in Machine Learning. — 2019. — Т. 12, № 4. — С. 307–392.
5. ШлеглЬТ., СебёкП., ВальдштейнС. М., Шмидт-ЭрфуртУ., ЛангсГ. Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery // Proceedings of the International Conference on Information Processing in Medical Imaging (IPMI). 2017.
6. Yeh, R., Chen, C., Lim, T.Y., Hasegawa-Johnson, M., Do, M.N.: Semantic image inpainting with perceptual and contextual losses. arXiv:1607.07539 (2016)