ИССЛЕДОВАНИЕ ОБНАРУЖЕНИЯ АНОМАЛИЙ В СЕТЕВОМ ТРАФИКЕ С ИСПОЛЬЗОВАНИЕМ ONE-CLASS SVM

Kurbanov Sardor Nuriddinovich

LLC "Programmsoft", Republic of Uzbekistan, Tashkent

Введение:

С ростом важности компьютерных сетей в современном мире возрастает и необходимость обеспечения их безопасности. Одной из ключевых задач в этой области является обнаружение аномалий в сетевом трафике, которые могут свидетельствовать о наличии подозрительной или вредоносной активности. Для решения этой задачи широко используются методы машинного обучения, способные анализировать большие объемы данных и выявлять скрытые шаблоны, невидимые человеческому глазу.

Одним из подходов к обнаружению аномалий является применение One-Class SVM (Support Vector Machine), мощного инструмента машинного обучения, который находит свое применение в различных областях. One-Class SVM стремится разделить данные на нормальные и аномальные, создавая границу вокруг нормальных данных и идентифицируя объекты, находящиеся за её пределами как аномальные. В данной статье мы исследуем применение One-Class SVM для обнаружения аномалий в сетевом трафике.

Мы предлагаем глубокий анализ процесса обнаружения аномалий, начиная от захвата сетевого трафика и создания признаков для обучения модели, до оценки качества и адаптации модели в реальном времени. Мы рассмотрим примеры данных и выведем основные выводы о работе предложенного метода.

В данной статье мы обратим особое внимание на:

- Подготовку и преобразование данных сетевого трафика в пригодный для обучения вид.
- Обучение и применение One-Class SVM для выявления аномалий в сетевом трафике.

- Оценку качества обнаружения аномалий и подходы к адаптации модели для изменяющейся ситуации.

Через этот анализ мы попытаемся лучше понять эффективность данного подхода к обнаружению аномалий и выявим его преимущества и ограничения в реальном мире. Продвигаясь далее, давайте вглубь рассмотрим каждый этап этого метода и его реализацию на практике.

Ключевые слова: Обнаружение аномалий, Сетевой трафик, One-Class SVM, Машинное обучение, Support Vector Machine, Преобразование данных, Оценка качества, Адаптация модели, Безопасность сетей, Мониторинг ресурсов, Телекоммуникации, Data Science, Искусственный интеллект, Python, Прогнозирование классов, Статистический анализ, Преобразование признаков, Telegram бот, Информационная безопасность

Постановка задачи:

Целью данного исследования является изучение эффективности метода обнаружения аномалий в сетевом трафике на основе One-Class SVM. Мы ставим перед собой следующие задачи:

1. Захват и обработка сетевого трафика: Осуществить захват сетевого трафика и преобразовать полученные данные в структурированный формат для последующего анализа.

2. Создание признаков для обучения: Разработать подход к созданию дополнительных признаков на основе собранных данных, что поможет улучшить производительность модели.

3. Обучение One-Class SVM: Произвести обучение модели One-Class SVM на подготовленных данных для обнаружения аномалий в сетевом трафике.

4. Оценка качества обнаружения аномалий: Проанализировать результаты классификации и оценить качество обнаружения аномалий с использованием стандартных метрик.

5. Анализ результатов: Проанализировать эффективность метода, выявить сильные и слабые стороны, а также оценить возможности его дальнейшего улучшения.

Исследование будет проведено на реальных данных сетевого трафика с использованием библиотек машинного обучения и анализа данных на платформе Python. Полученные результаты помогут понять, насколько One-Class SVM подходит для обнаружения аномалий в сетевом трафике и какие аспекты стоит учитывать при его практическом применении.

1. Захват и обработка сетевого трафика:

Первый этап нашего исследования заключается в захвате сетевого трафика и его последующей обработке для дальнейшего анализа. Для этого мы используем библиотеку Scapy, позволяющую нам считывать и анализировать сетевые пакеты. В процессе захвата мы получаем информацию о различных атрибутах пакетов, таких как IP-адрес отправителя и получателя, размер пакета и другие характеристики, которые могут быть полезными при обнаружении аномалий.

Захваченные данные будут сохранены в виде списка, содержащего словари с информацией о каждом пакете. Каждый словарь будет содержать следующие поля:

- `Source_IP`: IP-адрес отправителя пакета.
- `Destination_IP`: IP-адрес получателя пакета.
- `Packet_Size`: Размер пакета в байтах.

Этот этап является ключевым для получения входных данных для обучения и оценки модели. Обработанные данные позволят нам лучше понять характеристики сетевого трафика и подготовить их для дальнейшего анализа.

2. Создание признаков для обучения:

На втором этапе исследования мы разрабатываем методику создания дополнительных признаков на основе собранных данных о сетевом трафике. Эти признаки позволяют модели лучше разделять нормальные и аномальные пакеты, так как они включают в себя дополнительную информацию о характеристиках пакетов.

В данной работе мы предлагаем использовать два дополнительных признака: `Feature1` и `Feature2`. Для их генерации мы будем использовать случайные

значения из нормального распределения. Это позволит внести случайность в данные, что может помочь модели лучше выявлять скрытые аномалии.

Формула для генерации случайных признаков:

$$Feature1 = N(\mu_1, \sigma_1)$$

$$Feature2 = N(\mu_2, \sigma_2)$$

где $N(\mu, \sigma)$ представляет собой нормальное распределение с параметрами среднего μ и стандартного отклонения σ.

Эти признаки добавляют дополнительный уровень информации о пакетах, который может быть полезен для обучения модели One-Class SVM. Следующие этапы обработки будут включать стандартизацию данных и обучение модели с учетом созданных признаков.

3. Обучение One-Class SVM:

На этом этапе мы переходим к обучению модели One-Class SVM на подготовленных данных. One-Class SVM представляет собой метод машинного обучения, который находит гиперплоскость, максимально удаленную от нормальных данных, и определяет объекты, находящиеся за её пределами, как аномалии.

Мы будем использовать ядро радиального базисной функции (RBF) для One-Class SVM. Формула для вычисления расстояния между пакетами x_1 и x_2 в пространстве признаков выглядит следующим образом:

$$K(x, x_i) = \exp(-\gamma||x - x_i||^2)$$

где $||x - x_i||^2$ - это Евклидово расстояние между векторами признаков x и x_i, а γ - это параметр ядра.

Задачей обучения One-Class SVM является поиск гиперплоскости f(x), которая максимально охватывает нормальные данные:

$$F(x) = sgn(\sum_{i=1}^{n} a_i K(x, x_i) - \rho)$$

где a_i - множители Лагранжа, ρ - пороговое значение.

После обучения модели мы можем использовать её для классификации новых пакетов на нормальные и аномальные на основе полученного значения функции (f(x).

Мы проведем обучение One-Class SVM на наших подготовленных данных, включая дополнительные признаки, и будем использовать эту модель для дальнейшего анализа и оценки качества обнаружения аномалий.

4. Оценка качества обнаружения аномалий:

Для оценки эффективности метода обнаружения аномалий на основе One-Class SVM, мы применяем стандартные метрики оценки классификации. Для начала, мы создаем метки истинности аномалий на основе порогового значения размера пакета. Пакеты, размер которых превышает 99-й перцентиль, помечаются как аномалии (-1), остальные - как нормальные (1).

Затем, используя обученную модель One-Class SVM, мы получаем предсказанные классы для всех пакетов. Сравнивая предсказанные классы с истинными метками аномалий, мы можем вычислить следующие метрики:

1. Точность (Precision): Доля действительно аномальных пакетов среди всех предсказанных аномальных пакетов.

$$\text{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

2. Полнота (Recall): Доля действительно аномальных пакетов среди всех истинно аномальных пакетов.

$$\text{Recall} = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

3. F-мера (F1-score): Среднее гармоническое между точностью и полнотой. Она учитывает оба аспекта классификации.

$$\text{F1} - \text{score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Оценка качества позволяет нам понять, насколько хорошо модель способна выявлять аномалии в сетевом трафике. Результаты оценки также могут служить основой для сравнения с другими методами обнаружения аномалий и для принятия решения о дальнейшей оптимизации модели.

5. Анализ результатов:

Анализ результатов является заключительным этапом в рамках постановки задачи. На этом этапе мы осуществляем всестороннюю оценку работы модели, оцениваем её эффективность в контексте обнаружения аномалий в сетевом трафике. Давайте подробнее рассмотрим этот этап и как он соотносится с предыдущими шагами:

1. Оценка метрик:

Точность, полнота, и F1-мера:* Мы применяем ранее рассмотренные метрики для оценки качества модели. Точность дает нам представление о том, насколько часто наша модель правильно выявляет аномалии. Полнота оценивает, насколько успешно модель обнаруживает все реальные аномалии. F1-мера предоставляет сбалансированную меру между точностью и полнотой. Ложные срабатывания и пропуски:* Анализ ложных срабатываний и пропусков является важным шагом. Если модель имеет высокую точность, но низкую полноту, это может свидетельствовать о том, что модель неспособна выявлять все реальные аномалии.

2. Адаптация и оптимизация:

Автоматическая адаптация модели: Мы оцениваем, насколько успешно модель адаптируется к изменениям в сетевом трафике. Автоматическая адаптация позволяет модели оставаться эффективной даже при изменяющихся условиях. Если мы выявляем необходимость в частых корректировках, это может указывать на необходимость оптимизации модели или выборе более устойчивого метода.

Проактивность в выявлении новых угроз: Рассматривается, насколько успешно модель выявляет новые, ранее неизвестные угрозы. Проактивное обнаружение новых аномалий является важным аспектом, особенно в условиях постоянно меняющейся киберугрозы.

3. Сравнение с предыдущими оценками:

Сравнение с результатами предыдущих итераций: Если данная оценка результатов является частью повторяющегося процесса, то проводится

сравнение с предыдущими результатами. Это помогает выявить тенденции в улучшении или ухудшении производительности модели, что может подсказать о необходимости корректировок в методологии или параметрах.

Сравнение с другими методами: Если были использованы или рассмотрены другие методы обнаружения аномалий, результаты текущей модели сравниваются с ними. Это может помочь в определении того, насколько эффективна выбранная модель по сравнению с альтернативными подходами.

4. Дальнейшие действия:

Оптимизация параметров: Если результаты не удовлетворяют ожидания, проводится анализ параметров модели. Возможно, необходимо провести оптимизацию параметров, таких как настройка ядра One-Class SVM или изменение параметров адаптации.

Дополнительное обучение: Результаты анализа могут подсказать о необходимости дополнительного обучения модели на новых данных. Это может включать в себя расширение обучающего набора или изменение стратегии обучения.

Развитие стратегий безопасности: На основе анализа результатов можно выработать новые стратегии безопасности или улучшить существующие. Это может включать в себя внесение изменений в процессы мониторинга безопасности или дополнительные шаги по обеспечению безопасности в сети.

Анализ результатов является циклическим процессом, который позволяет постоянно совершенствовать методику обнаружения аномалий и повышать уровень безопасности информационной системы.

Заключение: Обеспечение Информационной Безопасности через Обнаружение Аномалий в Сетевом Трафике

В ходе нашего исследования по обеспечению информационной безопасности через обнаружение аномалий в сетевом трафике мы успешно пройдем через ряд важных этапов, охватывающих постановку задачи, захват и обработку сетевого трафика, создание признаков для обучения, обучение модели One-Class SVM, оценку качества обнаружения аномалий и анализ результатов.

1. Захват и обработка сетевого трафика:

Мы начали с определения задачи захвата сетевого трафика, что является фундаментальным этапом в обеспечении безопасности сети. Применяя инструменты, такие как Scapy, мы реализовали механизм захвата и обработки пакетов, что является неотъемлемой частью процесса обнаружения аномалий.

2. Создание признаков для обучения:

Для эффективного обучения модели мы внедрили создание дополнительных признаков на основе данных сетевого трафика. Этот этап позволяет модели учесть разнообразные аспекты пакетов и повысить её способность обнаруживать аномалии в сетевой активности.

3. Обучение One-Class SVM:

Применяя метод One-Class SVM, мы обеспечили модели способность выявлять аномалии на основе обучающего набора данных. Этот этап предоставляет ключевой инструмент для обнаружения нешаблонных и подозрительных сценариев в сетевом трафике.

4. Оценка качества обнаружения аномалий:

С использованием метрик, таких как точность, полнота и F1-мера, мы осуществили оценку качества работы модели. Этот этап позволяет нам не только количественно измерить эффективность, но и проанализировать ложные срабатывания и пропуски, что существенно для оптимизации процесса обнаружения аномалий.

5. Анализ результатов:

Анализ результатов стал заключительным этапом, на котором мы провели всестороннюю оценку эффективности модели. Результаты оценки, сравнение с предыдущими итерациями и альтернативными методами обеспечивают основу для дальнейших шагов в обеспечении безопасности.

Завершая наше исследование, мы видим, что обнаружение аномалий в сетевом трафике является неотъемлемым компонентом обеспечения информационной безопасности. Наш подход, основанный на тщательной постановке задачи, эффективной обработке данных и анализе результатов, предоставляет

практический инструмент для выявления и реагирования на потенциальные угрозы. Применение таких инновационных методов обеспечивает надежную защиту современных информационных систем от всевозможных атак и нешаблонных сценариев, что оставляет нас на передовой борьбы за безопасность в цифровой эпохе.

Список литературы:

1. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
2. Scapy: The Python Packet Crafting for Network Security (2021). Retrieved from https://scapy.net/
3. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
4. Psutil - Cross-platform library for retrieving information on running processes and system utilization (2021). Retrieved from https://psutil.readthedocs.io/en/latest/
5. Rousseeuw, P. J., & Leroy, A. M. (1987). Robust regression and outlier detection. John Wiley & Sons.
6. Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data (pp. 93-104).
7. Wu, K., Li, Z., Zhao, T., & Leung, V. C. (2019). A survey on intrusion detection with machine learning: A forward looking perspective. IEEE Transactions on Emerging Topics in Computational Intelligence, 3(4), 362-371.
8. Han, J., & Kamber, M. (2006). Data mining: concepts and techniques. Elsevier.