Подход к отбору информативных признаков сетевого трафика для оперативного обнаружения компьютерных атак

Кузьмичев Владимир Андреевич
кандидат технических наук, начальник лаборатории (научно-исследовательской), Военно-космическая академия имени А.Ф. Можайского, vka@mil.ru,

Ильина Марина Анатольевна
младший научный сотрудник, Военно-космическая академия имени А.Ф. Можайского, vka@mil.ru

В статье рассматривается задача оперативного обнаружения компьютерных атак на основе анализа статистических характеристик сетевого трафика. Описаны проблема избыточности признакового описания сетевых сессий и негативное влияние взаимной зависимости параметров на скорость и устойчивость работы алгоритмов классификации. Предложен двухэтапный подход к отбору признаков, сочетающий устранение линейно зависимых параметров на основе коэффициента корреляции Пирсона и последующее ранжирование оставшихся характеристик по критерию информативности с помощью ансамбля решающих деревьев. Приведены результаты экспериментальной проверки предложенного подхода на размеченных данных сетевого трафика, подтвердившие возможность сокращения числа анализируемых признаков и ускорения классификации более чем в 3 раза при сохранении точности и полноты обнаружения атак на уровне не ниже 98%.
Ключевые слова: обнаружение компьютерных атак, сетевой трафик, отбор информативных признаков, решающие деревья, классификация трафика.

Введение

В настоящее время обеспечение безопасности современных информационно-телекоммуникационных сетей существенно осложняется широким распространением протоколов сквозного шифрования, вследствие чего традиционные средства глубокого анализа пакетов (DPI) теряют свою практическую эффективность. В этих условиях основным подходом к своевременному выявлению компьютерных атак становится несигнатурный анализ агрегированных характеристик сетевых сессий: объёмов передаваемых данных, межпакетных интервалов и служебных флагов заголовков.

Однако современные средства агрегации сетевого трафика формируют избыточные описания соединений, включающие от 80 и более первичных характеристик (признаков сетевых сессий). Обработка полного набора признаков такой размерности в высокоскоростных каналах связи требует значительных аппаратных ресурсов и приводит к задержкам при классификации трафика в реальном времени. Кроме того, многие сетевые характеристики функционально дублируют друг друга, создавая эффект мультиколлинеарности, а включение в анализ адресной информации и временных меток (IP-адресов и портов) ведёт к переобучению моделей под конкретную сетевую топологию вместо выявления общих закономерностей компьютерных атак.

В этой связи актуальной научно-технической задачей представляется разработка подходов к обоснованному снижению размерности признакового пространства, позволяющих исключить топологически зависимые и избыточные характеристики сетевых сессий без снижения качества распознавания угроз.

Таким образом, целью исследования является разработка и экспериментальное обоснование двухэтапного подхода к отбору информативных признаков сетевого трафика, обеспечивающего ускорение классификации при сохранении точности обнаружения компьютерных атак на уровне не ниже 98%.

1 Анализ существующих подходов к отбору признаков сетевого трафика

Вопросам применения методов машинного обучения для анализа сетевого трафика и обнаружения компьютерных атак посвящено значительное число исследований [1, 2]. Для описания трафика авторы традиционно используют статистические агрегаторы — CICFlowMeter [7], Argus и другие [6], — генерирующие от нескольких десятков до более чем сотни параметров на каждую сессию.

Для снижения размерности признакового пространства в литературе выделяют три группы методов [4, 5]:

Методы фильтрации основаны на статистических критериях — взаимной информации, дисперсии, коэффициентах корреляции Пирсона или Спирмена. Они вычислительно просты, но учитывают только линейные и изолированные связи, оставляя без внимания сложные нелинейные зависимости, характерные для сетевых атак.

Обёрточные методы оценивают информативность подмножества признаков путём многократного обучения классификатора на различных их комбинациях. Такой подход обеспечивает высокую точность отбора, однако его вычислительная сложность резко растёт с увеличением объёма данных.

Встроенные методы, основанные на оценке значимости признаков в ансамблях деревьев, используют индекс Джини или прирост информации (Information Gain), в частности в алгоритме случайного леса. Они быстро сходятся и выявляют нелинейные закономерности, но склонны завышать значимость непрерывных числовых признаков с высокой вариативностью.

Анализ ряда опубликованных работ [1, 2, 4, 5] позволяет выделить ряд существенных ограничений.

Во-первых, при использовании встроенных методов зачастую не учитывается мультиколлинеарность: ранжирование по индексу Джини применяется без предварительной фильтрации, из-за чего в итоговый набор попадают сильно скоррелированные признаки (например, средняя длина пакета и средний размер сегмента, $r > 0{,}9$), что создаёт избыточность вычислений и снижает устойчивость модели.

Во-вторых, требованиям оперативности уделяется недостаточно внимания: критерием оптимизации почти всегда выступает точность классификации, тогда как время вычисления признаков и задержки при обработке потока в реальном времени остаются вне рассмотрения.

В-третьих, в ряде работ в состав анализируемых параметров ошибочно включаются топологически зависимые данные — IP-адреса, служебные порты, абсолютные временные метки, — что приводит к ложной информативности и снижает способность модели обобщать результаты при изменении архитектуры сети.

Таким образом, актуальной задачей представляется разработка комплексного подхода, объединяющего предварительную фильтрацию линейно зависимых признаков с последующей оценкой их нелинейного вклада ансамблевыми алгоритмами, исключающего топологические артефакты и обеспечивающего минимальное время обработки сетевых сессий.

2 Двухэтапный подход к отбору информативных признаков сетевого трафика

Предлагаемый подход направлен на устранение двух ключевых проблем: статистической избыточности параметров сетевых сессий и неоправданных вычислительных затрат при их обработке. Общая схема включает предварительную очистку данных и два последовательных аналитических этапа (рис. 1).



Рис. 1. Схема двухэтапного подхода к формированию сокращённого признакового пространства

2.1 Предварительная обработка и нормализация данных

На подготовительном этапе из исходного набора исключаются параметры, прямо указывающие на сетевую топологию: идентификатор потока (Flow ID), сетевые адреса источника и приёмника (Src IP, Dst IP), порты (Src Port, Dst Port) и временные метки (Timestamp). Использование таких признаков искусственно завышает показатели классификации на обучающей выборке за счёт запоминания адресов конкретного стенда, но лишает модель устойчивости при переносе на другую сетевую конфигурацию.

Дополнительно удаляются признаки с нулевой дисперсией (поля с постоянными или нулевыми значениями на всём интервале наблюдения) и сессии с ошибочными записями. Оставшиеся числовые характеристики приводятся к единому масштабу с помощью Z-преобразования:

$$z = \frac{x - \mu}{\sigma},$$

где x – исходное значение признака, μ – математическое ожидание, σ – среднеквадратическое отклонение.

2.2 Первый этап: фильтрация мультиколлинеарных признаков

Многие агрегированные характеристики сетевого потока функционально связаны между собой. Наличие пар признаков с сильной линейной зависимостью (мультиколлинеарностью) дестабилизирует решающие правила классификатора и увеличивает затраты на вычисление избыточных параметров.

Для выявления таких связей строится корреляционная матрица на основе коэффициента Пирсона:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

где x_i, y_i – значения исследуемой пары признаков для i-й сессии; $\bar{x}, \bar{y}$ – их выборочные средние.

Пороговое значение коэффициента корреляции для отсечения признаков выбрано равным $|r_{th}| = 0{,}85 \ldots 0{,}90$. Из каждой коррелированной пары исключается признак, требующий больших вычислительных затрат либо обладающий меньшей вариативностью. Примеры исключённых пар приведены в таблице 1.

В результате применения корреляционного фильтра размерность пространства сокращается с 65 до 38–40 взаимно слабокоррелированных признаков.

2.3 Второй этап: ранжирование признаков по критерию информативности

Оставшиеся после фильтрации признаки слабо коррелируют между собой, однако их вклад в разделение классов трафика (легитимный трафик и типы атак) существенно различается. Для нелинейной оценки значимости используется ансамбль решающих деревьев (Random Forest) [3].

Таблица 1

Примеры функционально зависимых пар признаков сетевого трафика

Оставляемый признак	Исключаемый дублирующий признак	Коэффициент корреляции r_{xy}	Физический смысл зависимости
Средний размер пакета (Average Packet Size)	Средняя длина пакета (Packet Length Mean)	0,95	Совпадение статистических моментов распределения длин пакетов
Длительность соединения (Flow Duration)	Суммарный интервал в прямом направлении (Fwd IAT Total)	0,94	Длительность сессии определяется преимущественно временем передачи пакетов в прямом направлении
Интенсивность пакетов потока (Flow Packets/s)	Интенсивность пакетов в прямом направлении (Fwd Packets/s)	0,90	При асимметричном обмене на трафик прямого направления приходится до 90% общего числа пакетов
Суммарный объём прямого потока (Total Length of Fwd Pkts)	Объём данных в прямом подпотоке (Subflow Fwd Bytes)	0,91	Понятие подпотока в агрегаторе дублирует суммарный объём сессии

Значимость каждого параметра определяется через суммарное взвешенное уменьшение критерия неоднородности Джини во всех узлах расщепления T построенных деревьев:

$$I(f) = \sum_{t \in T, v(t)=f} p(t) \cdot \Delta Gini(t),$$

где $v(t) = f$ означает, что в узле t расщепление проводится по признаку f; $p(t)$ – доля обучающих наблюдений, попавших в узел t; $\Delta Gini(t)$ – уменьшение неоднородности при разбиении.

Для выбора оптимальной границы отсечения параметров вводится критерий эффективности K_{eff}, отражающий баланс между качеством классификации и временем обработки:

$$K_{eff}(m) = \frac{F_1(m)}{t_{proc}(m)},$$

где $F_1(m)$ – F-мера классификации на подмножестве из m наиболее информативных признаков, $t_{proc}(m)$ – время формирования признаков и принятия решения.

Насыщение метрики F_1 наступает при включении первых 16–18 признаков: дальнейшее расширение набора не повышает точность, но приводит к линейному росту времени классификации. В результате формируется компактный рабочий набор признаков (минимальные размеры сегментов, среднеквадратические отклонения межпакетных интервалов, ключевые счётчики флагов), пригодный для работы классификатора в режиме реального времени.

Таблица 2

Сравнительная оценка эффективности способов формирования признакового пространства

Способ формирования признакового пространства	Число признаков	Точность (Precision)	Полнота (Recall)	F1-мера	Время обработки выборки, с	Снижение объема данных в ОЗУ, раз
Исходный набор (без отбора признаков)	65	0,94	0,94	0,94	12,90	1,0
Отбор по коэффициенту Пирсона относительно метки класса	33	0,97	0,96	0,97	7,21	1,5
Одномерный отбор по значимости Джини (без фильтрации корреляции)	25	0,97	0,97	0,97	6,85	1,9
Предлагаемый двухэтапный подход (фильтр Пирсона + отбор по Джини)	17	0,98	0,98	0,98	3,42	2,7

3 Экспериментальные исследования и анализ результатов

Для экспериментальной проверки предложенного двухэтапного подхода использовался массив данных реального сетевого трафика [7], включающий профиль легитимного взаимодействия (BENIGN) и сессии типовых компьютерных атак: распределённый отказ в обслуживании (DDoS, DoS), сканирование портов (PortScan) и подбор аутентификационных данных (Brute Force) [2].

Тестирование проводилось методом пятикратной перекрёстной проверки (5-fold cross-validation) на независимой контрольной выборке объёмом 30% от общего массива данных. В качестве базового классификатора использовался ансамбль решающих деревьев (Random Forest).

Сравнение эффективности предложенного подхода с альтернативными способами отбора признаков представлено в таблице 2.

Анализ полученных результатов показывает, что изолированный одномерный отбор по значимости Джини оставляет в модели скрытую мультиколлинеарность: схожие по физическому смыслу параметры конкурируют между собой, из-за чего для стабилизации точности требуется не менее 25 признаков.

Предложенный двухэтапный подход позволил:

сократить размерность входного вектора сессии с 65 до 17 параметров (почти в 4 раза);

ускорить классификацию в 3,77 раза (с 12,90 до 3,42 с), что важно при обработке потоковых данных в реальном времени;

снизить затраты оперативной памяти на хранение векторов признаков в 2,7 раза;

повысить F1-меру обнаружения атак с 0,94 до 0,98 за счёт устранения зашумляющих и избыточных параметров.

Зависимость точности классификации и времени обработки от числа используемых признаков представлена на рисунке 2.

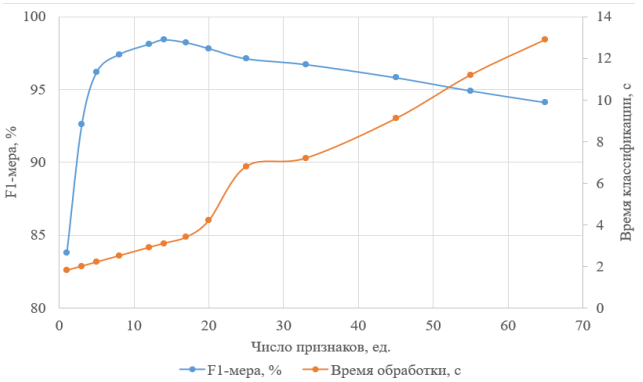


Рис. 2. Зависимость качества классификации и времени обработки от числа признаков

Как видно из графика, представленного на рис. 2, увеличение признакового пространства свыше 17–20 параметров практически не даёт прироста достоверности классификации, но приводит к линейному росту времени вычислений.

Заключение

В ходе исследования была решена задача оптимизации описания сетевых сессий для систем оперативного обнаружения компьютерных атак:

Показано, что использование полного набора из 65 признаков нецелесообразно из-за их избыточности, взаимной зависимости и наличия топологически зависимых параметров, что увеличивает время классификации и снижает обобщающую способность модели.

Разработан двухэтапный подход к отбору признаков, включающий исключение адресных данных, фильтрацию линейно зависимых параметров по коэффициенту Пирсона и ранжирование оставшихся признаков по критерию Джини.

Экспериментально подтверждено, что сокращённый набор из 17 признаков позволил ускорить классификацию в 3,77 раза, уменьшить расход оперативной памяти в 2,7 раза и повысить F1-меру с 0,94 до 0,98.

Полученные результаты относятся к одному набору данных и одному классификатору (Random Forest), поэтому в дальнейших исследованиях представляется целесообразным проверить устойчивость отобранных признаков на других наборах данных и типах атак, а также сопоставить предложенный подход с иными классификаторами и методами снижения размерности. Кроме того, перспективным направлением дальнейших исследований видится оценка предложенного подхода на потоковых данных в условиях дисбаланса классов и дрейфа характеристик трафика, что потребует разработки процедуры адаптивного пересмотра набора признаков и её реализации в программном прототипе системы обнаружения атак.

Литература

1. Гетьман А. И., Иконникова М. К. Обзор методов классификации сетевого трафика с использованием машинного обучения // Труды Института системного программирования РАН. 2020. Т. 32, вып. 6. С. 137–154. DOI: 10.15514/ISPRAS-2020-32(6)-11.

2. Горюнов М. Н., Мацкевич А. Г., Рыболовлев Д. А. Синтез модели машинного обучения для обнаружения компьютерных атак на основе набора данных CICIDS2017 // Труды Института системного программирования РАН. 2020. Т. 32, вып. 5. С. 81–94. DOI: 10.15514/ISPRAS-2020-32(5)-6.

3. Шелухин О. И., Ванюшина А. В., Габисова М. Е. Фильтрация нежелательных приложений интернет-трафика с использованием алгоритма классификации Random Forest // Вопросы кибербезопасности. 2018. № 2 (26). С. 44–51.

4. Talabis M., McPherson R., Miyamoto I., Martin J. Information Security Analytics. Elsevier, 2015. 166 p.

5. Domingos P. A Few Useful Things to Know about Machine Learning // Communications of the ACM. 2012. Vol. 55, № 10. Pp. 78–87.

6. Ring M., Wunderlich S., Scheuring D., Landes D., Hotho A. A survey of network-based intrusion detection data sets // Computers & Security. 2019. Vol. 86. Pp. 147–167. DOI: 10.1016/j.cose.2019.06.005.

7. Sharafaldin I., Lashkari A. H., Ghorbani A. A. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization // Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP 2018). 2018. Pp. 108–116. DOI: 10.5220/0006639801080116.

An approach to selecting informative network traffic features for operational computer attack detection

Kuzmichev V.A., Ilyina M.A.
A.F. Mozhaisky Military Space Academy

JEL classification: C10, C50, C60, C61, C80, C87, C90

The article considers the problem of operational computer attack detection based on the analysis of statistical characteristics of network traffic. The problem of redundancy in feature descriptions of network sessions and the negative impact of parameter interdependencies on the speed and stability of classification algorithms are described. A two-stage approach to feature selection is proposed, combining the elimination of linearly dependent parameters based on the Pearson correlation coefficient and the subsequent ranking of the remaining characteristics by informativeness using an ensemble of decision trees. The results of an experimental evaluation of the proposed approach on labeled network traffic data are presented, confirming the possibility of reducing the number of analyzed features and speeding up classification by more than three times while maintaining attack detection precision and recall at a level of at least 98%.

Keywords: computer attack detection, network traffic, informative feature selection, decision trees, traffic classification.

References

1. Getman A.I., Ikonnikova M.K. A Survey of Network Traffic Classification Methods Using Machine Learning. Trudy ISP RAN/Proc. ISP RAS, 2020, vol. 32, issue 6, pp. 137–154. DOI: 10.15514/ISPRAS-2020-32(6)-11. (in Russian)

2. Goryunov M.N., Matskevich A.G., Rybolovlev D.A. Synthesis of a Machine Learning Model for Detecting Computer Attacks Based on the CICIDS2017 Dataset. Trudy ISP RAN/Proc. ISP RAS, 2020, vol. 32, issue 5, pp. 81–94. DOI: 10.15514/ISPRAS-2020-32(5)-6. (in Russian)

3. Shelukhin O.I., Vanyushina A.V., Gabisova M.E. Filtering of Unwanted Internet Traffic Applications Using the Random Forest Classification Algorithm. Voprosy kiberbezopasnosti [Cybersecurity Issues], 2018, no. 2 (26), pp. 44–51. (in Russian)

4. Talabis M., McPherson R., Miyamoto I., Martin J. Information Security Analytics. Elsevier, 2015, 166 p.

5. Domingos P. A Few Useful Things to Know about Machine Learning. Communications of the ACM, 2012, vol. 55, no. 10, pp. 78–87.

6. Ring M., Wunderlich S., Scheuring D., Landes D., Hotho A. A Survey of Network-Based Intrusion Detection Data Sets. Computers & Security, 2019, vol. 86, pp. 147–167. DOI: 10.1016/j.cose.2019.06.005.

7. Sharafaldin I., Lashkari A.H., Ghorbani A.A. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP 2018), 2018, pp. 108–116. DOI: 10.5220/0006639801080116.