УДК 004.89

ПОВЫШЕНИЕ КАЧЕСТВА КЛАССИФИКАЦИИ НА ОСНОВЕ СЕГМЕНТАЦИИ ИНФОРМАЦИОННЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ В УСЛОВИЯХ РЕДКИХ СОБЫТИЙ НА ПРИМЕРЕ СЕТЕВЫХ АТАК

А.А. Макеенко, И.С. Лебедев (Санкт-Петербург)

Введение

Задачи анализа временных рядов и информационных последовательностей являются важным направлением в области информационной безопасности. Особое значение они приобретают при построении систем обнаружения вторжений, где требуется не только высокая точность, но и устойчивость алгоритмов к шуму, дисбалансу классов и изменчивости сетевого трафика.

Классические методы машинного обучения, например, такие как Logistic Regression, Decision Tree и Naive Bayes, при применении напрямую к исходным потокам данных зачастую демонстрируют ограниченную эффективность. Это связано с тем, что исходные последовательности содержат большое количество избыточной информации и нерегулярности, которые снижают точность классификации. В литературе [1-3] отмечается, что выбор подходящего способа представления данных играет не меньшую роль, чем выбор самого алгоритма.

В связи с этим возникает задача формирования подпоследовательностей данных таким образом, чтобы повысить качественные показатели выборки за счет назначения отдельных моделей на определяемые сегменты.

Предлагаемый подход

В работах [4, 5] предложена концепция сегментации информационных последовательностей, позволяющая адаптировать данные к используемой модели. Отмечается, что качество классификации определяется не только самим алгоритмом a_n, но и характеристиками сегментации μ_k, что формально выражается следующим критерием:

$$Q(a_n, X, \mu_k) \to \max_{k,n} \quad (1)$$

где X – входные данные, a_n – модель, μ_k – параметры сегментации.

В настоящая работа рассматривает применение данного подхода для анализа качества идентификации сетевых атак. Основное внимание уделяется исследованию влияния параметров сегментации (размер окна, сдвиг) на итоговые метрики качества моделей в условиях дисбаланса классов и наличия редких событий. Показано, что изменение способа представления данных способно обеспечить значительный прирост точности для простых алгоритмов, не требующих больших ресурсов и времени обучения.

В работе мы рассматриваем параметры сегментации $\mu_k = (\text{window,shift})$ как гиперпараметры представления данных в критерии качества $Q(a_n, X, \mu_k)$ и показываем, что их систематический подбор заметно повышает результат даже для простых моделей при редких событиях. По сути, мы переносим акцент с «выбора модели» к «выбору представления данных» и фиксируем измеримый прирост: для Logistic Regression и Decision Tree F1 растёт на 4–5 п.п., для Naive Bayes — порядка 40 п.п. В отличие от работ, где основной упор делается на отбор признаков и балансировку классов, а также на глубокие модели [7–10], мы демонстрируем практичный и воспроизводимый рецепт настройки сегментации, не требующий тяжёлых вычислительных ресурсов.

Материалы и методы

Для проведения экспериментов использовался общедоступный датасет UNSW-NB15 [1, 2]. Датасет включает как реальные, так и синтетические сетевые атаки, сгенерированные с использованием современного генератора IXIA PerfectStorm, что обеспечивает более реалистичное отражение сетевого.

UNSW-NB15 содержит около 2,5 млн. записей, каждая из которых описывается 49 признаками, включающими:

- числовые характеристики (длительность сессии, количество байт, количество пакетов и др.);
- категориальные признаки (протокол, сервис, состояние соединения);
- метку класса, определяющую тип трафика (нормальный или одна из 9 категорий атак: Fuzzers, DoS, Exploits, Generic, Reconnaissance, Shellcode, Worms и др.).

Такое сочетание числовых и категориальных данных делает UNSW-NB15 удобным для проверки различных алгоритмов машинного обучения. В настоящей работе для целей бинарной классификации метки атак были агрегированы в один общий класс «атака», противопоставленный классу «нормального» трафика.

В качестве базового уровня (baseline) были выбраны три классических алгоритма классификации: Logistic Regression (LR), Decision Tree (DT) и Naive Bayes (NB). Несмотря на то, что эти методы являются классическими [6, 7], они по-прежнему привлекательны благодаря низкой вычислительной сложности и скорости обучения, что делает их удобным baseline для сравнения с более сложными моделями [8-10].

Logistic Regression (LR) – линейный метод, оценивающий вероятность принадлежности объекта к классу на основе логистической функции. Несмотря на простоту, LR часто используется как базовый ориентир для сравнения с более сложными моделями.

Decision Tree (DT) – деревья решений строят иерархию правил вида «если–то», позволяя эффективно работать с разнородными признаками и интерпретировать полученные решения. Однако они подвержены переобучению при большом числе признаков.

Naive Bayes (NB) – вероятностный классификатор, основанный на теореме Байеса и предположении независимости признаков. Этот метод особенно чувствителен к структуре данных, поэтому его использование в сочетании с сегментацией позволяет проверить гипотезу о том, что изменение представления входных последовательностей может компенсировать ограничения самой модели.

Во всех экспериментах указанные модели обучались на исходных данных без применения сегментации, что позволило использовать их в качестве точки отсчёта для дальнейших сравнений.

Для моделирования последовательной структуры сетевого трафика применялась оконная сегментация. Данный метод позволяет преобразовать поток пакетов в совокупность окон фиксированной длины, каждое из которых описывает локальные статистические характеристики последовательности. Такой подход снижает влияние случайного шума и позволяет выявить устойчивые закономерности, характерные для атакующего или нормального поведения сети.

В экспериментах использовались параметры:

- размер окна (window) = {300, 500, 1000, 2000} пакетов;
- сдвиг (shift) = {150, 250, 500, 1000} пакетов.

Выбор диапазонов обусловлен необходимостью охватить как малые окна (для чувствительности к локальным изменениям), так и более протяжённые (для сглаживания кратковременных флуктуаций).

Для каждого окна вычислялись агрегированные характеристики:

- для числовых признаков – среднее значение, стандартное отклонение, минимум и максимум;
- для категориальных признаков – мода и количество уникальных значений.

Формирование метки окна осуществлялось по правилу: если не менее 10% пакетов внутри окна имели метку «атака», то всё окно относилось к классу «атака». Такой порог позволяет учитывать не отдельные единичные аномалии, а именно устойчивые проявления атакующего поведения.

Для проверки гипотезы о влиянии сегментации на качество классификации был проведён полный перебор комбинаций параметров «размер окна» (window), «сдвиг» (shift) и выбранной модели классификации. Таким образом, экспериментальная схема представляла собой grid search по всем возможным вариантам (window × shift × модель).

Для каждой конфигурации выполнялось обучение и последующая оценка модели на тестовой выборке. В качестве показателей качества использовались три стандартные метрики бинарной классификации:

- precision (точность) – доля корректно распознанных атак среди всех выявленных системой;
- recall (полнота) – доля обнаруженных атак среди всех реально присутствующих;
- F1-score – гармоническое среднее precision и recall, обеспечивающее баланс между ними.

В качестве точки отсчёта рассматривались baseline-модели (Logistic Regression, Decision Tree и Naive Bayes), обученные без применения сегментации. Сравнение с ними позволяет количественно оценить вклад сегментации в повышение качества классификации.

Полученные модели и конфигурации сегментации далее были сопоставлены по метрикам качества, результаты анализа представлены в следующем разделе.

Эксперименты выполнялись в среде Google Colab с использованием Python 3.10. Для обработки данных применялись библиотеки: *pandas* версии 2.x, *numpy* 1.26.x, *scikit-learn* 1.4.x и *matplotlib* 3.8.x.

Предобработка признаков осуществлялась с помощью ColumnTransformer: числовые данные нормализовались с использованием SimpleImputer(strategy="median") и StandardScaler, категориальные - через SimpleImputer(strategy="most_frequent") и OneHotEncoder(handle_unknown="ignore", sparse_output=False). В качестве классификаторов использовались Logistic Regression (solver="lbfgs", max_iter=300, random_state=42), Decision Tree (max_depth=6, random_state=42) и Gaussian Naive Bayes.

Сегментация выполнялась при значениях параметров окна $window \in \{300,500,1000,2000\}$ и сдвига $shift \in \{150,250,500,1000\}$. Метка окна формировалась по правилу: если не менее 10% пакетов в нём относились к классу «атака», то окно считалось атакующим ($\tau = 0.10$). Для каждой комбинации параметров и модели рассчитывались метрики precision, recall и F1. Все скрипты и результаты экспериментов сохранены в виде Jupyter-ноутбуков и изображений, что обеспечивает воспроизводимость исследования.

Результаты

В разделе представлены результаты экспериментов по оценке влияния параметров сегментации на качество классификации сетевых атак. Сначала приведено сравнение базовых моделей (без сегментации) с их модификациями после применения оконной сегментации. Далее проанализирована зависимость метрик качества от размеров окна и величины сдвига, как в целом по выборке, так и отдельно для каждой модели. Для наглядности результаты представлены в виде графиков и таблиц, что

позволяет проследить закономерности и выявить оптимальные конфигурации параметров сегментации.

Сравнение baseline и сегментации

На рис. 1 приведено сравнение значений F1 для трёх базовых моделей до и после применения сегментации. Видно, что во всех случаях сегментация приводит к улучшению качества классификации. Для Logistic Regression и Decision Tree прирост составляет 4–5 п.п., что подтверждает устойчивое положительное влияние выбранного метода. Особенно заметен рост качества для Naive Bayes: показатель F1 увеличился более чем на 40 п.п., что объясняется тем, что агрегирование признаков сглаживает шум и делает распределения данных ближе к предположениям этой модели.

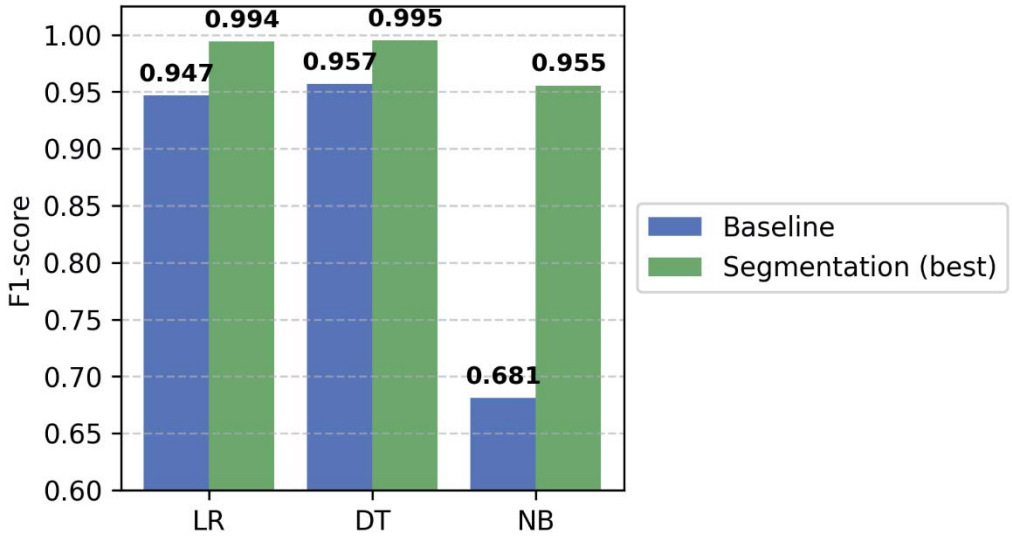

Рис. 1. Сравнение моделей по F1: baseline и сегментация

Зависимость от параметров окна и сдвига

Для анализа влияния параметров сегментации была построена тепловая карта (рис.2), отображающая значения метрики F1 при различных комбинациях размеров окна и величины сдвига. По оси абсцисс откладывается значение сдвига, по оси ординат – размер окна, а цветовая шкала отражает достигнутое качество классификации. Видно, что лучшие результаты наблюдаются при использовании больших окон (1000–2000) и умеренных значений сдвига (250–1000), что подтверждает гипотезу о том, что агрегирование большего объёма информации улучшает устойчивость модели к шуму.

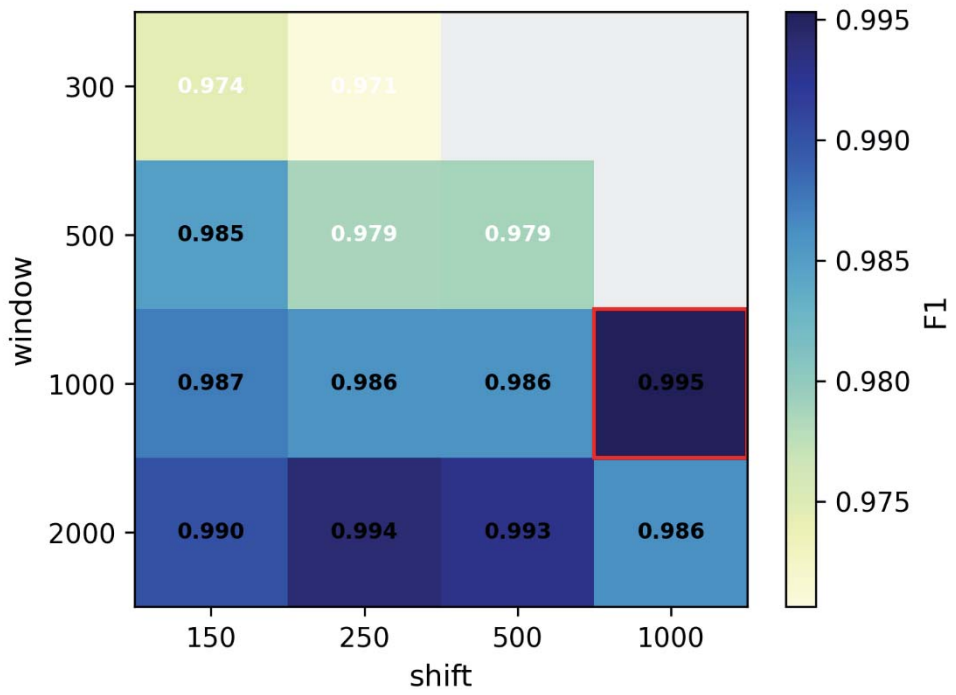

Рис. 2. Тепловая карта значений F1 (максимум по моделям) в зависимости от параметров сегментации

На рис. 3 показано, как у модели Logistic Regression метрика F1 меняется при увеличении размера окна для разных значений сдвига.

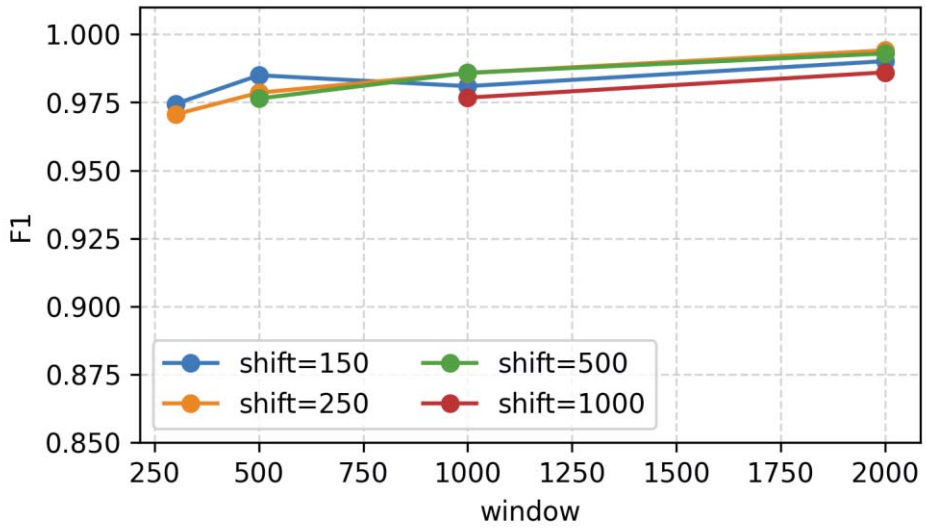

Рис. 3. Logistic Regression: зависимость F1 от размера окна при разных значениях сдвига

На рис. 4 представлена зависимость метрики F1 у Logistic Regression от величины сдвига при различных размерах окна. Видно, что модель сохраняет стабильное качество во всём диапазоне параметров, при этом наибольшие значения достигаются при больших окнах.

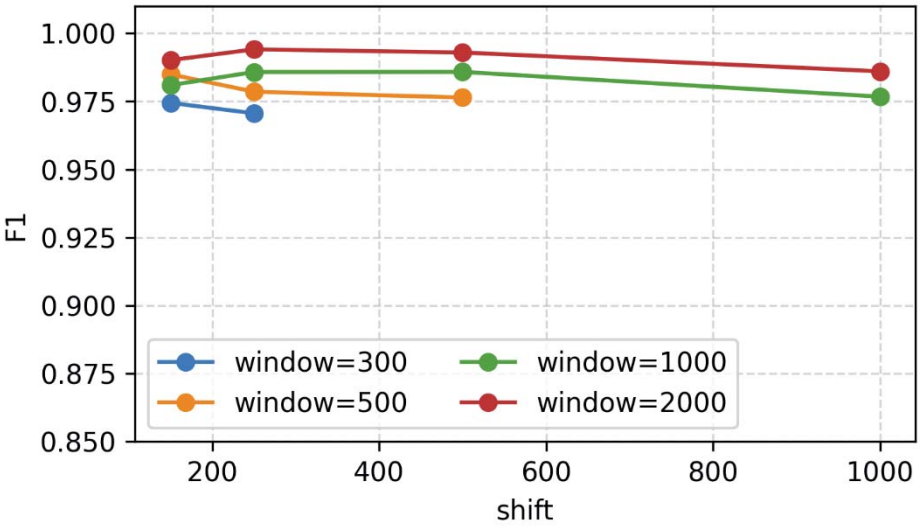

Рис. 4. Logistic Regression: зависимость F1
от сдвига при разных размерах окна

На рис. 5 показана зависимость F1 для модели Decision Tree от размера окна при различных значениях сдвига. Видно, что качество классификации остаётся высоким, а наибольший рост наблюдается при переходе к большим окнам.

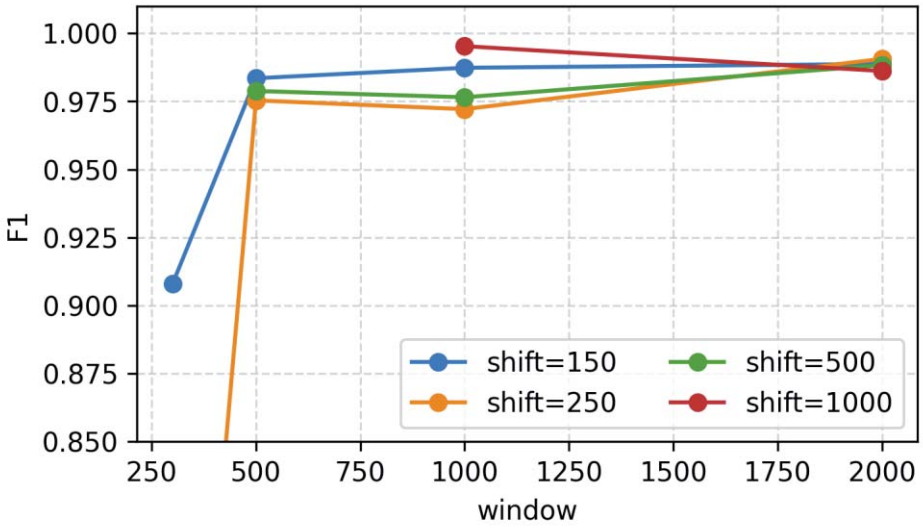

Рис. 5. Decision Tree: зависимость F1 от размера окна
при разных значениях сдвига

На рис. 6 представлена зависимость F1 у Decision Tree от величины сдвига при разных размерах окна. Модель демонстрирует устойчивые результаты, а максимальные значения метрики достигаются при больших окнах и умеренных сдвигах.

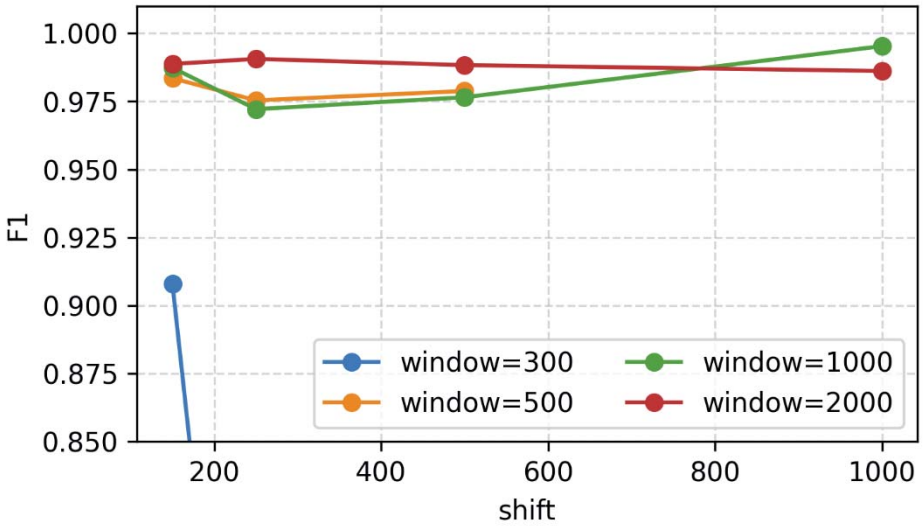

Рис. 6. Decision Tree: зависимость F1 от сдвига
при разных размерах окна

На рис. 7 показана зависимость F1 для модели Naive Bayes от размера окна при различных значениях сдвига. Видно, что с увеличением окна метрика существенно улучшается, особенно при средних значениях сдвига, что подтверждает эффективность сегментации именно для этой модели.

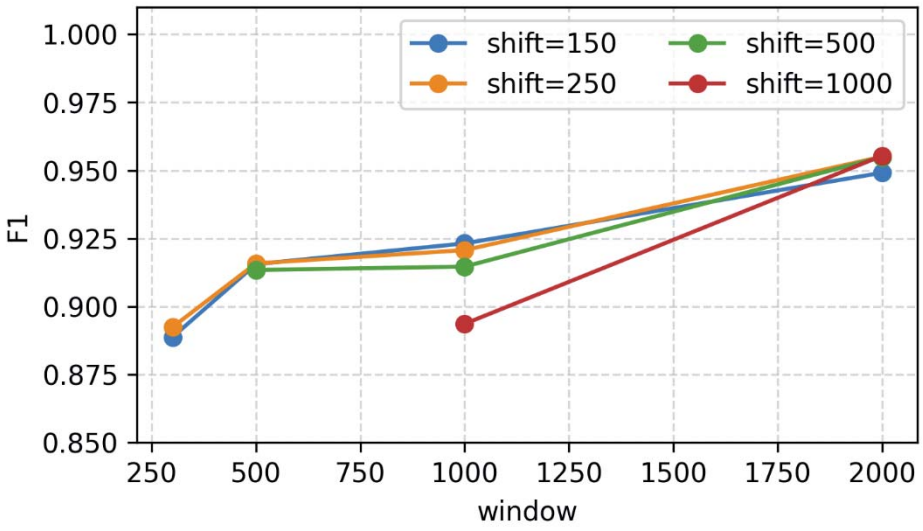

Рис. 7. Naive Bayes: зависимость F1 от размера окна
при разных значениях **сдвига**

На рис. 8 представлена зависимость F1 для Naive Bayes от величины сдвига при разных размерах окна. Наилучшие результаты достигаются при больших окнах (2000), где метрика сохраняется на уровне выше 0.95 независимо от сдвига.

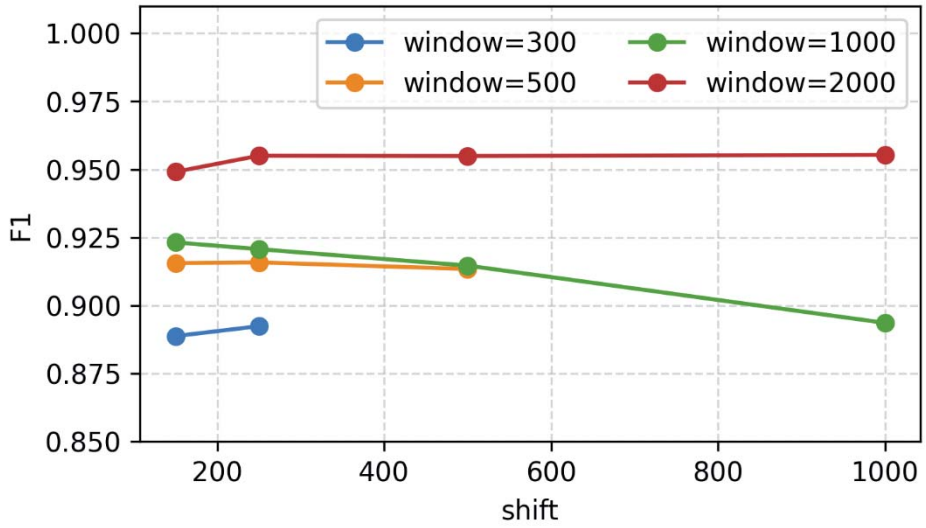

Рис. 8. Naive Bayes: зависимость F1 от сдвига
при разных размерах окна

Таблицы результатов

Для количественной оценки проведено сравнение значений F1 для базовых моделей и их лучших конфигураций после сегментации. Результаты сведены в табл. 1.

Таблица 1. Сравнение baseline и лучшей конфигурации сегментации по F1

Модель	F1 (без сегм.)	F1 (с сегм.)	ΔF1, %	Окно	Сдвиг
NB	0,681	0,955	+40,3	2000	1000
LR	0,947	0,994	+5,0	2000	250
DT	0,957	0,995	+4,0	1000	1000

Как видно, наибольший прирост наблюдается у модели Naive Bayes (ΔF1 = +40,3 п.п.), в то время как Logistic Regression и Decision Tree также показывают заметное улучшение (+5 и +4 п.п. соответственно).

Более детализированный анализ метрик precision, recall и F1 представлен в табл. 2. Это позволяет оценить, какие именно аспекты качества классификации улучшаются за счёт сегментации.

Таблица 2. Подробное сравнение precision, recall и F1 (расширенный вариант)

Модель	P (без)	P (с)	ΔP, %	R (без)	R (с)	ΔR, %	F1 (без)	F1 (с)	ΔF1, %	Окно	Сдвиг
NB	0,519	0,93	79,3	0,99	0,982	-0,8	0,681	0,955	+40.3	2000	1000
LR	0,929	1	7,6	0,965	0,988	2,4	0,947	0,994	+5,0	2000	250
DT	0,981	1	1,9	0,934	0,991	6,1	0,957	0,995	+4,0	1000	1000

Из таблицы видно, что для Naive Bayes особенно значителен рост precision (+79,3%), что связано с уменьшением числа ложных срабатываний.

Для Logistic Regression и Decision Tree также фиксируется положительная динамика как по precision, так и по recall, что подтверждает общую эффективность метода.

Таким образом, эксперимент показал закономерное улучшение качества классификации при использовании сегментации. Полученные результаты требуют обсуждения в контексте существующих методов и ограничений.

Обсуждение

Полученные результаты показывают, что сегментация информационных последовательностей выступает в роли универсального приёма, позволяющего существенно повысить эффективность классификации. Особенно выраженный эффект наблюдается у модели Naive Bayes, для которой прирост метрики F1 превысил 40 п.п. Такой результат объясняется тем, что агрегирование признаков снижает вариативность и шум, приближая распределения данных к предположениям метода о независимости признаков. В то же время Logistic Regression и Decision Tree демонстрируют более умеренный прирост (3–5 п.п.), что подтверждает, что сегментация является важным гиперпараметрическим фактором даже для устойчивых алгоритмов.

Важно отметить, что традиционные подходы к повышению качества в задачах анализа сетевого трафика опираются преимущественно на отбор признаков или методы балансировки классов [4, 6]. В данной работе показано, что сегментация может рассматриваться как самостоятельный инструмент повышения качества, дополняющий или даже замещающий эти методы [6, 7, 10].

Практическая значимость заключается в том, что даже простые классификаторы при использовании сегментации достигают качества, сопоставимого с более сложными моделями, что открывает возможность построения лёгких и ресурсоэффективных систем обнаружения вторжений. В дальнейшем перспективным направлением является исследование адаптивной сегментации, при которой параметры окна изменяются динамически в зависимости от свойств потока, а также интеграция данного подхода с ансамблевыми и глубокими моделями.

Заключение

В работе показано, что сегментация информационных последовательностей является эффективным инструментом повышения качества классификации. Экспериментальные результаты на датасете UNSW-NB15 подтвердили, что даже простые модели, такие как Naive Bayes и Decision Tree, после применения сегментации достигают уровня качества, сопоставимого с более сложными алгоритмами машинного обучения. Наибольший прирост метрик наблюдается у модели Naive Bayes, что объясняется снижением шума и приближением распределений признаков к предположениям метода.

Практическая ценность исследования состоит в возможности разработки простых и ресурсоэффективных систем обнаружения атак и анализа сетевого трафика, что особенно важно для реальных условий эксплуатации, где ограничены вычислительные ресурсы. Предложенный подход универсален и применим к широкому спектру задач анализа временных рядов и потоков информации, включая телекоммуникационные системы, финансовый мониторинг и медицинскую диагностику.

Следует отметить, что исследование ограничивалось использованием фиксированных параметров окна и сдвига, а также применением базовых алгоритмов классификации. В дальнейшем перспективным направлением является разработка методов адаптивной сегментации, интеграция предложенного подхода с ансамблевыми моделями и глубоким обучением, а также исследование его применимости к другим типам данных.

Литература

1. **Moustafa, Nour, and Jill Slay.** UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). Military Communications and Information Systems Conference (MilCIS), 2015. IEEE, 2015.
2. **Moustafa, Nour, and Jill Slay.** The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set. Information Security Journal: A Global Perspective (2016): 1-14.
3. **Adugna T.D., Ramu A., Haldorai A.** A Review of Pattern Recognition and Machine Learning // Journal of Machine and Computing. 2024. Vol. 4, № 1. 210-220 p. https://doi.org/10.53759/7669/jmc202404020.
4. **Lebedev I. S.** Sequential information processing using adaptive pattern analysis in assessing the state of systems. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2025, no. 3, pp. 25-36 (In Russian). doi:10.31799/1684-8853-2025-3-25-36, EDN: SSGKZU.
5. **Lebedev, I. S., & Sukhoparov, M. E.** (2024). Improving the Quality Indicators of Multilevel Data Sampling Processing Models Based on Unsupervised Clustering. Emerging Science Journal, 8(1), 355-371. doi:10.28991/ESJ-2024-08-01-025.
6. **García-Teodoro P. et al.** Anomaly-based network intrusion detection: Techniques, systems and challenges // Comput. Secur. 2009. Vol. 28, № 1-2. P. 18-28. https://doi.org/10.1016/j.cose.2008.08.003.
7. **Hussain Bhuyan M., Bhattacharyya D.K., Kalita J.K.** Survey on Incremental Approaches for Network Anomaly Detection // Int. J. Commun. Networks Inf. Secur. 2012. Vol. 3, № 3.
8. **Chinnasamy R. et al.** Deep learning-driven methods for network-based intrusion detection systems: A systematic review // ICT Express. Elsevier, 2025. Vol. 11, № 1. P. 181–215. https://doi.org/10.1016/J.ICTE.2025.01.005.
9. **Zhang Y., Muniyandi R.C., Qamar F.** A Review of Deep Learning Applications in Intrusion Detection Systems: Overcoming Challenges in Spatiotemporal Feature Extraction and Data Imbalance // Appl. Sci. 2025, Vol. 15, Page 1552. Multidisciplinary Digital Publishing Institute, 2025. Vol. 15, № 3. P. 1552. https://doi.org/10.3390/APP15031552.
10. **Ring M. et al.** A survey of network-based intrusion detection data sets // Comput. Secur. Elsevier Advanced Technology, 2019. Vol. 86. P. 147-167. https://doi.org/10.1016/J.COSE.2019.06.005.