УДК 004.6
doi:10.31799/1684-8853-2025-6-15-27
EDN: ERTCQY

Научные статьи
Articles

Метод фильтрации признаков по критериям стабильности и значимости

О. С. Исаева[a]***,*** *доктор техн. наук, ведущий научный сотрудник, orcid.org/0000-0002-5061-6765, isaeva@icm.krasn.ru*
[a]*Институт вычислительного моделирования СО РАН — обособленное подразделение ФИЦ КНЦ СО РАН, Академгородок, 50/44, Красноярск, 660036, РФ*

Введение: *анализ сетевого трафика интернета вещей осложнен высокой размерностью, избыточностью и нестабильностью признаков. Наблюдается сильная корреляция, мультиколлинеарность и шум, что снижает качество кластеризации и затрудняет интерпретацию. Кроме того, легитимный и аномальный трафик часто перекрываются, что осложняет формализацию границ между классами. В этой связи требуется метод отбора признаков, обеспечивающий устойчивость, компактность и семантическую интерпретируемость.* ***Цель:*** *разработать и экспериментально оценить новый метод для построения устойчивого и интерпретируемого признакового пространства в задачах кластеризации сетевого трафика — Progressive Feature Filtering with Stability and Significance (PFF-SS, PF^{2}S).* ***Методы:*** *описан пошаговый алгоритм PF^{2}S, сочетающий анализ линейных (корреляция, VIF) и нелинейных (взаимная информация) зависимостей с оценкой стабильности и информативности. На каждом этапе исключаются избыточные, слабо значимые или нестабильные признаки.* ***Результаты:*** *применение PF^{2}S к датасету сетевого трафика интернета вещей позволило сократить число признаков с более чем 300 до 17, сохранив высокую информативность. Сравнение с пространствами, редуцированными методом главных компонент и методом рекурсивного исключения признаков, показало, что PF^{2}S обеспечивает более высокие метрики стабильности, интерпретируемости и качестве кластеризации. Метод не преобразует признаки, как метод главных компонент, а сохраняет их исходную семантику. По сравнению с методом рекурсивного исключения признаков PF^{2}S обеспечил отсутствие мультиколлинеарности, более низкую сложность модели и на 17,6 % более высокий силуэтный коэффициент. Кластеры, построенные на основе PF^{2}S-пространства, оказались устойчивыми (высокий скорректированный индекс Рэнда) и семантически интерпретируемыми.* ***Практическая значимость:*** *PF^{2}S формирует компактное и устойчивое признаковое пространство, пригодное для систем обнаружения аномалий в сетевом трафике интернета вещей.* ***Обсуждение:*** *перспективным направлением является адаптация PF^{2}S для потоковой обработки данных и интеграция с сигнатурными методами выявления аномалий и онтологиями сетевого трафика.*

Ключевые слова *— интернет вещей, устойчивость признаков, информативность признаков, кластеризация K-средних, агломеративная кластеризация, спектральная кластеризация, модель гауссовых смесей, метод главных компонент, метод рекурсивного исключения признаков, анализ сетевого трафика, обнаружение аномалий.*

Для цитирования: Исаева О. С. Метод фильтрации признаков по критериям стабильности и значимости. *Информационно-управляющие системы*, 2025, № 6, с. 15–27. doi:10.31799/1684-8853-2025-6-15-27, EDN: ERTCQY
For citation: Isaeva O. S**.** Feature filtering method based on stability and significance criteria. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2025, no. 6, pp. 15–27 (In Russian). doi:10.31799/1684-8853-2025-6-15-27, EDN: ERTCQY

Введение

Широкое применение систем искусственного интеллекта сопровождается ростом числа задач, в которых признаковое пространство характеризуется высокой размерностью, разнотипностью и избыточностью. Это особенно актуально в предметных областях, где данные генерируются в реальном времени и содержат сотни, а иногда и тысячи признаков, описывающих поведение устройств, пользователей и правила их взаимодействия [1]. В таких условиях эффективность моделей машинного обучения снижается из-за увеличения вычислительной сложности, риска переобучения и потери интерпретируемости. Сокращение признакового пространства становится ключевым условием построения устойчивых, интерпретируемых и масштабируемых моделей.

К таким задачам относится анализ сетевой активности в системах интернета вещей (Internet of Things, IoT), где применение традиционных протоколов безопасности затруднено требованиями к облегченности и энергоэффективности решений из-за ограничений в вычислительной мощности и времени автономной работы устройств [2]. Архитектура IoT объединяет физические и виртуальные объекты: датчики, исполнительные механизмы, облачные сервисы, специальные сетевые протоколы, транспортные средства коммуникации и пользователей [3]. Основными целями безопасности IoT являются обеспечение конфиденциальности, целостности и доступности предлагаемых услуг [4]. Аномалии в таких сетях могут быть вызваны как техническими сбоями (например, выходом из строя IoT-устройства), так и целенаправленными кибератаками (например, DoS — «отказ в обслуживании», MITM — «атака посредника», spoofing — «подмена доверенного лица»). Эффективное выявление таких угроз требует

построения моделей, способных различать нормальное поведение и деструктивные воздействия на основе анализа многомерных данных [5]. IoT-сети характеризуются динамичностью процессов, гетерогенностью устройств и постоянным изменением признаков нормального поведения, что затрудняет формализацию границ между «нормой» и «аномалией» [6].

В рамках исследований по обеспечению безопасности сети IoT в Красноярском научном центре СО РАН создана и внедрена инфраструктура для сбора данных и имитации угроз, позволяющая генерировать реалистичные сценарии сетевой активности [7]. Схема IoT-сети построена по шаблону «Издатель — Подписчик» с использованием протокола MQTT. Различные сценарии атак для этого протокола рассмотрены в работе [8]. Обязательным элементом архитектуры сети является брокер, отвечающий за прием и маршрутизацию сообщений. Исследования проводятся для брокеров, развернутых в нескольких популярных платформах (Eclipse Mosquitto, EMQX, NanoMQ, VerneMQ) с различными конфигурациями политик безопасности. Настройки брокеров осуществляются адаптивно [9], но вопросы безопасности для такой разнообразной сети остаются. С помощью программных агентов фиксируется весь сетевой трафик, поступающий на стандартные и шифрованные порты как во внутренней сети, так и извне [10]. В настоящее время собраны датасеты, описывающие временные, статистические, протокольные и поведенческие характеристики трафика.

Полученные датасеты характеризуются высокой размерностью как по числу объектов (пакетов, накопленных за длительный период), так и по числу признаков (которых более 300). При этом легитимные и аномальные сетевые сессии могут быть как долгосрочными, так и кратковременными, что затрудняет их исследование на основе временных характеристик. Наблюдается высокий уровень шума, многие признаки являются избыточными, сильно коррелированными и чувствительными к вариативности данных. Особую сложность представляет перекрытие классов на подмножествах признаков. В работе [11] показано, что на качество классификации значительно влияет степень такого пересечения. В этих условиях возникает необходимость в применении надежных и интерпретируемых методов снижения размерности, способных выделить устойчивое ядро информативных признаков.

Существующие методы снижения размерности можно разделить на фильтрующие, встраиваемые, обертывающие и методы преобразования [12]. Несмотря на их различия, большинство подходов не обеспечивают высокой стабильности отбора признаков. В работе [13] показано, что в реальных задачах эффективность сокращения пространства признаков целесообразно оценивать по критерию его устойчивости (стабильности), т. е. способности метода воспроизводить схожие наборы признаков при вариации обучающих выборок. Это свойство особенно важно в условиях IoT, где нормальное поведение динамично, а данные подвержены изменчивости из-за обновлений устройств, сбоев или изменений в сетевой нагрузке.

Фильтрующие методы (например, на основе корреляции, взаимной информации или условной энтропии) просты и вычислительно эффективны, что делает их популярными для предварительного анализа в задачах с большим числом признаков [14]. Однако фильтрующие методы не учитывают взаимодействия между признаками и не зависят от целевой модели. Это может привести к отбору избыточных признаков (например, нескольких коррелирующих переменных) или пропуску информативных комбинаций.

Встраиваемые методы интегрируют отбор признаков в процесс обучения, что позволяет учитывать скрытую структуру данных. К таким методам относится LASSO (Least Absolute Shrinkage and Selection Operator), который применяет L_1-регуляризацию для обнуления коэффициентов при слабых признаках, деревья решений и градиентный бустинг (XGBoost, LightGBM), где признаки ранжируются по важности [15]. Эти методы эффективно учитывают нелинейности и взаимодействия, но в условиях шумных и разреженных данных, типичных для сетевых сессий интернета вещей, даже небольшие изменения выборки приводят к существенно разным результатам, что снижает доверие к модели.

Методы преобразования признаков, такие как метод анализа главных компонент (Principal Component Analysis, PCA), t-SNE, UMAP и autoencoders [16], подразделяются на глобальные и локальные в зависимости от того, какие структурные свойства данных они стремятся сохранить. Глобальные методы ориентированы на сохранение расстояний между всеми парами точек в пространстве, в то время как локальные методы фокусируются на сохранении структуры локальных окрестностей. В результате локальные подходы могут лучше передавать внутреннюю структуру кластеров, но при этом искажать глобальную топологию данных; напротив, глобальные методы воспроизводят общую структуру распределения, но могут терять детали локальной кластеризации. Современные методы стремятся сбалансировать эти аспекты. Например, t-SNE моделирует распределение попарных сходств в исходном многомерном пространстве и воспроизводит его в низкоразмерном представ-

лении, что позволяет сохранить как локальные, так и частично глобальные структуры. Тем не менее такие методы остаются чувствительными к выбору гиперпараметров и теряют интерпретируемость, поскольку представляют данные в виде линейных или нелинейных комбинаций исходных признаков, а не в терминах самих признаков [17]. Это делает их неприменимыми для задач, где важно понимать, какие именно исходные признаки вносят вклад в модель.

Обертывающие методы, такие как метод рекурсивного исключения признаков (Recursive Feature Elimination, RFE) и метод исключения предсказуемых признаков (Predictable Feature Elimination, PFE), обеспечивают высокую точность, но требуют значительных вычислительных ресурсов. RFE итеративно исключает наименее важные признаки на основе обученной модели, но требует многократного переобучения, что делает его неэффективным для больших данных [18]. PFE оценивает предсказуемость каждого признака по остальным с помощью вспомогательной модели машинного обучения и последовательно исключает признаки, которые могут быть восстановлены через остальные [19]. В отличие от PCA, PFE сохраняет интерпретируемость, так как работает с исходными признаками, но не учитывает целевую переменную и не оценивает стабильность отбора. Сравнение эффективности методов для разных типов задач приводится, например, в [20, 21].

На практике для выбора адекватного метода требуется компромисс между статистической значимостью признаков и воспроизводимостью их отбора. Многие подходы фокусируются только на одном из этих аспектов, что снижает надежность итогового решения, особенно в условиях шумных, разреженных или несбалансированных данных.

Целью данной работы является разработка нового метода отбора признаков, сочетающего пошаговое сокращение признакового пространства с одновременной оценкой значимости и стабильности признаков. Результатом работы стал оригинальный метод прогрессивной фильтрации признаков на основе стабильности и значимости (Progressive Feature Filtering with Stability and Significance, PFF-SS, или PF^{2}S), основанный на итеративной фильтрации. Метод позволяет на каждом шаге выявлять зависимые признаки (коррелирующие, мультикоррелирующие или информационно-связанные), которые ранжируются по комбинированному критерию значимости и стабильности (последняя оценивается через бутстрэп-выборки), выполнять оценку сложности модели при исключении признака и в результате исключать наименее устойчивые и слабые признаки, не увеличивающие обобщающую способность модели. Метод предназначен для данных, собранных в инфраструктуре интернета вещей, развернутой в рамках корпоративной сети научного центра.

Постановка задачи исследования данных IoT

Набор данных, содержащий сетевой трафик интернета вещей, охватывает шесть ключевых категорий: временные характеристики, флаги протоколов TCP и MQTT, параметры скорости соединений, статистические данные по заголовкам пакетов, свойства полезной нагрузки и объемные характеристики при массовой передаче данных [22]. Требуется построить устойчивые классы сетевой активности, которые можно использовать для разметки данных и последующего анализа новых наблюдений в целях выявления сетевых аномалий. Эта задача сводится к задаче кластеризации — автоматического разбиения объектов на группы на основе схожести их признаковых описаний.

Пусть $\mathbf{X} \in \mathbb{R}^{n \times m}$ — матрица наблюдений, где каждая строка соответствует объекту, а каждый столбец — признаку. Обозначим $\mathbf{I} = \{1, 2, \ldots, n\}$ — множество индексов объектов, $\mathbf{J} = \{1, 2, \ldots, m\}$ — множество индексов признаков, x_{ij} — значение j-го признака для i-го объекта, где $i \in \mathbf{I}$, $j \in \mathbf{J}$. Каждый объект $i \in \mathbf{I}$ описывается вектором $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im}) \in \mathbb{R}^m$. Каждый признак $j \in \mathbf{J}$ описывается вектором $\mathbf{p}_i = (x_{1j}, x_{2j}, \ldots, x_{mj}) \in \mathbb{R}^n$. Матрица наблюдений может быть представлена через множество объектов или множество признаков:

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix} = \begin{bmatrix} p_1 & p_2 & \cdots & p_m \end{bmatrix}. \quad (1)$$

Обозначим множество объектов как $\chi = (x_i | i \in \mathbf{I})$, множество признаков как $\mathbf{P} = \{p_j \mid j \in \mathbf{J}\}$. Задача кластеризации заключается в разбиении множества объектов χ на K непересекающихся подмножеств:

$$\chi = \bigcup_{k=1}^{K} \chi_k, \quad (2)$$

где $\chi_k \subseteq \chi$, $\chi_k \neq \varnothing$, $\chi_{k1} \cap \chi_{k2} = \varnothing$ для $\forall\, k1 \neq k2$. χ_k — множество объектов, отнесенных к k-му кластеру, $k = [1, K]$.

Такое разбиение выполняет соответствующее разбиение множества индексов I на подмножества

$$\mathbf{C}_k = \{i \in \mathbf{I} \mid x_i \in \chi_k\} \quad (3)$$

такие, что

$$\mathbf{I} = \bigcup_{k=1}^{K} \mathbf{C}_k, \quad (4)$$

где $\mathbf{C}_k \neq \varnothing$, $\mathbf{C}_{k1} \cap \mathbf{C}_{k2} = \varnothing$ для $\forall\, k1 \neq k2$.

Для построения такого разбиения было рассмотрено несколько методов кластеризации, представляющих разные подходы к группировке данных [23]:

— метод K-средних эффективен для компактных, сферических кластеров, но чувствителен к выбросам и может не находить сложные структуры;

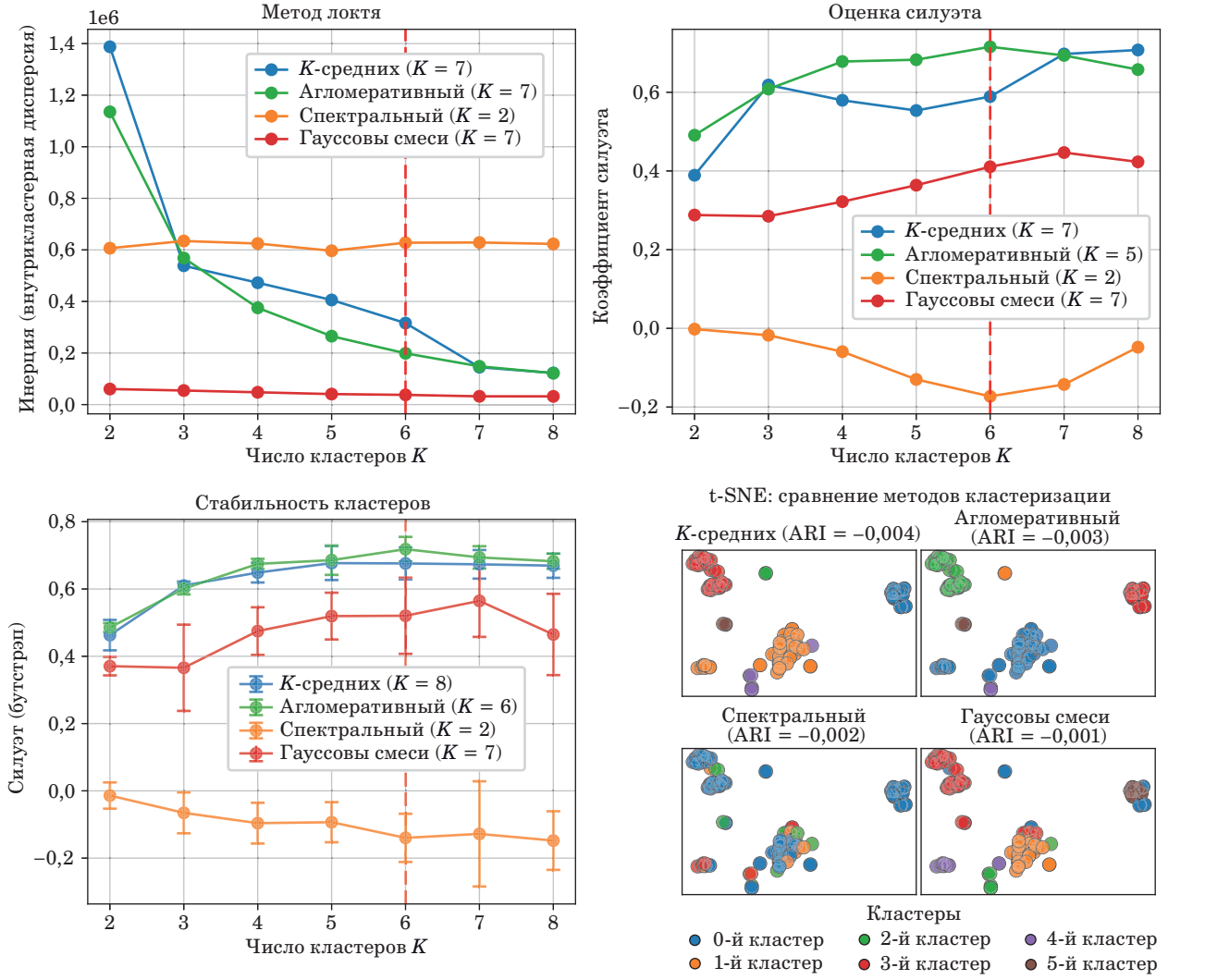
— метод агломеративной иерархической кластеризации (Hierarchical Agglomerative Clustering) позволяет выделять кластеры произвольной формы и анализировать иерархию группировок;

— метод спектральной кластеризации (Spectral Clustering) основан на спектральном разложении матрицы сходства, эффективен для кластеров с нелинейной структурой, но чувствителен к начальным параметрам;

— модель гауссовых смесей (Gaussian Mixture Model) позволяет представить данные как смесь многомерных распределений, за счет чего применима при наличии перекрытий и сложной внутренней структуры.

Оптимальное число кластеров определяется на основе вычисления внутрикластерной дисперсии, коэффициентов силуэта и оценки стабильности кластеризации на бутстрэп-выборках по скорректированному индексу Ренда (Adjusted Rand Index, ARI [24]) при различных значениях K. Результаты и визуализация кластеров через t-SNE [25] приведены на рис. 1.

Как видно из рисунка, рекомендованное количество кластеров различается для разных мето-



■ ***Рис. 1.*** Выбор оптимального количества кластеров

■ ***Fig. 1.*** Determining the optimal number of clusters

дов. Применение оценок компактности и разделенности кластеров тоже не позволяет однозначно сделать выбор K. Все методы демонстрируют низкую стабильность при увеличении числа кластеров. Коэффициент устойчивости ARI близок к нулю или отрицательный для всех методов. Это указывает на то, что признаковое пространство избыточно, структура данных не выражена явно, кластеры не компактны и не разделены.

Для уменьшения размерности признакового пространства применен метод PCA. Построена система обобщенных признаков в виде линейных комбинаций исходных переменных, объясняющих заданную долю общей дисперсии данных (не менее 95 %). На основе полученного сжатого представления выполнена кластеризация с использованием описанных выше алгоритмов (рис. 2).

Анализ результатов кластеризации (после обработки данных методом PCA) показал пересечение кластеров и их низкую стабильность (оценки ARI на бутстрэп-подвыборках близки к нулю). Разные алгоритмы кластеризации демонстрировали значительное расхождение в результатах: то, что один метод выделял как отдельный кластер и формировал компактные и изолированные группы, другой объединял с соседними группами, выделяя более протяженные структуры. Такая несогласованность затрудняет выбор единого, предпочтительного разбиения. Для оценки качества кластеризации выделенные группы проецировались обратно в исходное признаковое пространство, и их содержательная однородность анализировалась экспертами предметной области. Семантическая оценка осмысленности кластеров показала, что схожие с точки зрения предметной области сессии нередко оказывались в разных кластерах, в то время как существенно различающиеся сетевые события объединялись в один класс. Причиной этого эффекта является высокая степень зависимости между признаками, включая корреляцию, мультиколлинеарность и функциональную взаимозависимость. Такие признаки вносят избыточный вклад в отдельные направления признакового пространства, что искажает его геометрию и приводит к формированию некорректных кластеров.

Для устранения этого эффекта автором предложен новый метод отбора признаков PFF-SS (PF^{2}S), основанный на оценке стабильности и значимости (вклада) признаков в структуру данных. В методе введены метрики, которые позволяют последовательно исключать избыточные признаки при минимальной потере информативности, обеспечивая сохранение ключевых свойств признакового пространства на каждом этапе преобразования.

PF^{2}S — метод сокращения размерности признакового пространства

Цель метода PF^{2}S — пошагово сократить множество признаков $\mathbf{P} = \{p_j \mid j \in \mathbf{J}\}$ до подмножества $\mathbf{P}^H \subseteq \mathbf{P}$, удовлетворяющего критериям информативности, устойчивости и независимости. Для его применения необходимо обеспечить выполнение в матрице наблюдений $\mathbf{X}$ (1) следующих условий.

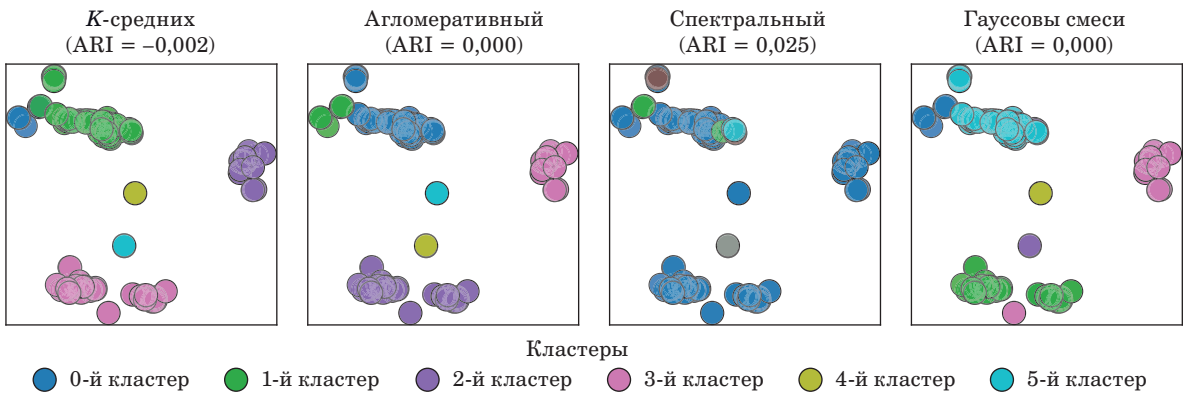
1. Все элементы матрицы $\mathbf{X}$ являются числовыми и определенными:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{I}\left(x_{ij} = \varnothing\right) = 0, \tag{5}$$

где $\mathrm{I}(\cdot)$ — индикаторная функция; x_{ij} — (i, j) значение в матрице $\mathbf{X}$.

2. Все признаки имеют дисперсию, превышающую заданный порог:

$$\forall j \ \sigma_j(p_j) \geq \tau_\sigma, \tag{6}$$



■ ***Рис. 2.*** Результат кластеризации в пространстве главных компонент

■ ***Fig. 2.*** Clustering results in the principal component space

где σ_j — стандартное отклонение; τ_σ — порог дисперсии.

3. Матрица является центрированной. Для этого построим матрицу $\tilde{\mathbf{X}} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$, $\tilde{x}_i = \left(\tilde{x}_{i1}, \tilde{x}_{i2}, \ldots, \tilde{x}_{im}\right)$, $\tilde{x}_{ij} = \dfrac{x_{ij} - \bar{x}_j}{\sigma_j}$. Это требование позволит построить ковариационную матрицу $\mathbf{\Sigma}_{\mathbf{X}} = \dfrac{1}{n-1}\tilde{\mathbf{X}}^{\mathrm{T}} \cdot \tilde{\mathbf{X}}$ и вычислить собственные числа $\lambda_1, \lambda_2, \ldots, \lambda_m$, собственные векторы $v_1, v_2, \ldots, v_m$ и сингулярные числа $\delta_1, \delta_2, \ldots, \delta_m$ ($\delta_i = \sqrt{\lambda_i}$).

Введем целевую функцию оценки признакового пространства

$$L(\tilde{\mathbf{X}}) = \alpha_1 \cdot K\left(\tilde{\mathbf{X}}\right) + \alpha_2 \cdot I\left(\tilde{\mathbf{X}}\right) + \sum_{l=1}^{L} \alpha_{3,l} \cdot R\left(\mathbf{F}, \tilde{\mathbf{X}}\right), \quad (7)$$

где K — число обусловленности матрицы $\tilde{\mathbf{X}}$, отражающее степень мультиколлинеарности; I — средняя взаимная информация между признаками, оценивающая нелинейную зависимость; R — мера сложности для семейства функций $\mathbf{F}$; $\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_l$ — весовые коэффициенты, позволяющие настраивать приоритеты между компонентами (по умолчанию $\alpha_i = 1$).

Минимизация L соответствует улучшению структуры признакового пространства за счет снижения зависимости между признаками и повышения устойчивости результата. Распишем каждое из слагаемых в L. Число обусловленности матрицы $\tilde{\mathbf{X}}$ вычисляется как $K = \delta_{\max}/\delta_{\min}$, где $\delta_{\max}$ — наибольшее сингулярное число, $\delta_{\max} \neq 0$ — наименьшее сингулярное число. K показывает, насколько данные устойчивы к малым изменениям (для линейной зависимости). Средняя взаимная информация I отвечает за оценку меры нелинейной зависимости и избыточности признаков:

$$I\left(\tilde{\mathbf{X}}\right) = \frac{2}{m(m-1)} \sum_{j=1}^{m} \sum_{k=j+1}^{m} I_{jk}, \quad (8)$$

где элемент I_{jk} определяет величину взаимной информации между P_j и P_k:

$$I_{jk} = \sum_{\tilde{x}_{ij} \in p_j} \sum_{\tilde{x}_{ik} \in p_k} f(\tilde{x}_{ij}, \tilde{x}_{ik}) \cdot \ln\left(\frac{f(\tilde{x}_{ij}, \tilde{x}_{ik})}{f(\tilde{x}_{ij}) \cdot f(\tilde{x}_{ik})}\right), \quad (9)$$

где $f(\tilde{x}_{ij}, \tilde{x}_{ik})$ – частота совместного появления значений признаков p_j и p_k для i-го объекта; $f(\tilde{x}_{ij})$ и $f(\tilde{x}_{ik})$ – частота появления каждого значения признака в отдельности. Данная формула определена для эмпирических частот [26].

В качестве меры сложности R будет применяться оценка Радемахера

$$R\left(F, \tilde{\mathbf{X}}\right) = E_{\varepsilon_i, \tilde{\mathbf{X}}}\left[\sup_{F \in \mathbf{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i F\left(\tilde{\mathbf{X}}_i\right) \right|\right], \quad (10)$$

где ε_i — случайная величина (переменная Радемахера, принимающая значения +1 и –1 с вероятностью 1/2); $F \in \mathbf{F}$ — функция, $\mathbf{F}_{\tilde{l}}$ — семейство функций; E — среднее по всем ε_i и $\tilde{\mathbf{X}}$. R — теоретическая мера, которая показывает способность аппроксимировать данные и избегать переобучения. В работе [27] обосновано применение меры сложности Радемахера (Rademacher complexity) для оценки обобщающей способности моделей, обученных на немаркированных данных.

Для метода PF^{2}S необходимо определить, какие признаки рассматривать в качестве кандидатов на удаление и по каким критериям делать выбор. Для каждого признака $p_j \in \mathbf{P}$, где $j \in \mathbf{J}$, определим три типа подмножеств, включающих признаки, связанные с p_j различными видами зависимости: $\mathbf{C}_j \subseteq \mathbf{P}$ — коррелирующих с p_j признаков, $\mathbf{V}_j \subseteq \mathbf{P}$ — мультиколлинеарных с p_j признаков и $\mathbf{I}_j \subseteq \mathbf{P}$ — взаимозависимых с p_j признаков. Для каждого признака p_j формируется подмножество: $\mathbf{D}_j = (\mathbf{C}_j, \mathbf{V}_j, \mathbf{I}_j)$, включающее все признаки, находящиеся в зависимости с p_j. Объединенное множество всех признаков используется для последующей фильтрации:

$$\mathbf{D} = \bigcup_{j=1}^{m} \mathbf{D}_j. \quad (11)$$

Ниже описаны правила построения каждого из этих подмножеств. Коррелирующими с p_j считаются признаки, коэффициент корреляции с которыми превышает заданный порог:

$$\mathbf{C}_j = \left\{ p_k \in \mathbf{P} \setminus \{p_j\} \,\middle|\, |\rho_{jk}| \geq \tau_\rho \right\}, \quad (12)$$

где ρ_{jk} — коэффициент корреляции между признаком p_j и p_k; $\tau_\rho \in [0, 1]$ — пороговое значение корреляции.

Для оценки мультиколлинеарности признака p_j выполняется построение линейной модели его восстановления по всем остальным признакам: $p_j = \sum_{k \neq j} \beta_k p_k + \zeta_j$, где p_j — вектор значений j-го признака; β_k — коэффициенты, найденные методом наименьших квадратов; p_k — векторы значений остальных признаков; ζ_j — вектор остатков. Поскольку данные предварительно центрированы, свободный член модели β_0 отсутствует.

Признак $p_j \in \mathbf{P}$ мультиколлинеарный, если коэффициент инфляции дисперсии (Variance Inflation Factor, VIF [28]) Vif_j превышает пороговое значение τ_V, т. е.

$$Vif_j = \frac{1}{1-\mathcal{R}_j^2} \geq \tau_V, \quad (13)$$

где $\mathcal{R}_j^2$ — коэффициент детерминации признака p_j, который вычисляется по формуле

$$\mathcal{R}_j^2 = \frac{\sum_{i=1}^{n}(\hat{x}_{ij} - \bar{x}_j)^2}{\sum_{i=1}^{n}(\tilde{x}_{ij} - \bar{x}_j)^2}, \quad (14)$$

где $\tilde{x}_{ij}$ – истинное значение j-го признака у i-го объекта; $\hat{x}_{ij} = \sum_{k \neq j} \beta_k \tilde{x}_{ik}$ – предсказанное значение, $\tilde{x}_{ik}$ – значение k-го признака у i-го объекта; $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} \tilde{x}_{ij}$ – среднее значение.

Для каждого признака, удовлетворяющего условиям (13), (14), анализируется структура зависимости. Выбираются признаки p_k, входящие в модель (13) с коэффициентами β_k, превышающими заданный порог. Формируется подмножество таких признаков

$$\mathbf{V}_j = \{p_k \in \mathbf{P} \setminus \{p_k\} \mid \beta_k > \tau_\beta\}, \quad (15)$$

где $\tau_\beta \geq 0$ — порог учета признака.

Взаимозависимыми считаются признаки, между которыми коэффициент взаимной информации превышает заданный порог. Для каждого p_j определим множество, отражающее нелинейные зависимости между признаками:

$$\mathbf{I}_j = \{p_k \in \mathbf{P} \setminus \{p_k\} \mid I_{jk} > \tau_I\}, \quad (16)$$

где I_{jk} — величина взаимной информации между p_j и p_k, вычисляемая по (8); τ_I — порог сильной зависимости.

Для принятия решения об исключении признака, входящего в множество $\mathbf{D}$, для каждого множества $\mathbf{D}_j \subseteq \mathbf{D}$ сформируем расширенное множество $\mathbf{D}'_j = \mathbf{D}_j \cup \{p_j\}$ и введем метрики стабильности и значимости.

Стабильность признака $p_k \in \mathbf{D}'_j$ оценивается через его воспроизводимость на случайных подвыборках:

$$S(p_k) = \frac{1}{T-1}\sum_{t=1}^{T-1} M(p_k^{(t)}, p_k^{(t+1)}), \quad (17)$$

где T — количество случайных выборок, полученных из $\tilde{\mathbf{X}}$; $M(p_k^{(t)}, p_k^{(t+1)})$ – взаимная информация, вычисленная по (9) для признака p_k на t-й и $(t+1)$-й подвыборках. Чем выше $S(p_k)$, тем стабильнее признак.

Значимость признака $p_k \in \mathbf{D}'_j$ вычисляется как вклад в объясненную дисперсию:

$$W(p_k) = \sum_{i=1}^{d} \lambda_i \cdot v_i(p_k)^2, \quad (18)$$

где $d = \min\left\{k \middle| \frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{m}\lambda_j} \geq \varsigma\right\}$ – глубина редуцированного пространства, содержащего ς дисперсии исходных данных, $\varsigma = (0, 1]$ — порог; $\lambda_1, \lambda_2, \ldots, \lambda_d$ — собственные значения ковариационной матрицы; $v_i(p_k)$ — компонента собственного вектора v_i, соответствующая признаку p_k. Чем выше $W(p_k)$, тем важнее признак с точки зрения структуры данных.

Для сложных структур данных предлагается дополнительно ввести метрику нелинейной значимости, которая включает веса, формируемые с использованием моделей машинного обучения (например, Lasso), оценивающих предсказательную способность признака относительно других признаков, рассматриваемых в роли целевых переменных по всем остальным признакам.

Построим множество кандидатов на удаление:

$$\mathbf{G}_j = \{p_k \in \mathbf{D}'_j \mid (S(p_k) < \tau_S) \wedge (W(p_k) < \tau_W)\}, \quad (19)$$

где τ_S, τ_W — пороги стабильности и значимости соответственно.

Если $\mathbf{G}_j = \varnothing$, группа $\mathbf{D}_j$ считается устойчивой и информативной — удаление признаков не производится.

Если $\mathbf{G}_j \neq \varnothing$ и существует $p^* \in \mathbf{D}'_j$ такой, что

$$p^* = \arg\min_{p_k \in \mathbf{G}_j}\left(\alpha\frac{S(p_k)}{S_{\max}} + \beta\frac{W(p_k)}{W_{\max}}\right), \quad (20)$$

где $S_{\max} = \max_{p_k \in \mathbf{G}_j} S(p_k)$, $W_{\max} = \max_{p_k \in \mathbf{G}_j} W(p_k)$, α, $\beta \geq 0$, $\alpha + \beta = 1$ — веса, позволяющие настраивать приоритет, то $\mathbf{P} = \mathbf{P} \setminus p^*$ и $\mathbf{D} = \mathbf{D} \setminus \mathbf{D}_j$.

Алгоритм итеративной фильтрации признаков заключается в пошаговом построении множества $\mathbf{D}$, выборе признаков для удаления из признакового пространства $\mathbf{P}$ и его исключении при условии, что это не ухудшает целевую функцию L. Пусть $\mathbf{A}$ — множество операций фильтрации признакового пространства. На каждом шаге $h = \{0, 1, \ldots, H\}$ из множества $\mathbf{A}$ выбирается операция $A^h \in \mathbf{A}$ исключения признака, удовлетворяющая условию

$$A^h = \arg\min_{A' \in \mathbf{A}} L(A'(\mathbf{P}^{s-1})), \quad (21)$$

где A' — операция, для которой $L(A'(\mathbf{P}^{h-1})) \leq L(\mathbf{P}^{h-1})$.

Если такая A^h существует, она применяется к текущему множеству признаков:

$$\mathbf{P}^h = A^h(\mathbf{P}^{h-1}). \quad (22)$$

Пусть A^h выполняет исключение признака $p \in \mathbf{P}^{h-1}$ из признакового пространства, тогда $\mathbf{P}^h = \mathbf{P}^{h-1} \backslash p$, $\mathbf{A} = \mathbf{A} \backslash A^h$. Итоговое признаковое пространство на шаге H определяется композицией всех преобразований

$$\mathbf{P}^H = A^H \circ A^{H-1} \circ \ \dots \ \circ A^1(\mathbf{P}), \qquad (23)$$

где $\mathbf{P}$ — исходное признаковое пространство; «$\circ$» — операция композиции.

Фильтрация признакового пространства завершается при выполнении хотя бы одного из условий

$$|\mathbf{P}^h| \le m_{\min},$$

$$\min_{A' \in \mathbf{A}} L(A'(\mathbf{P}^h)) > L(\mathbf{P}^h), \ h = H. \qquad (24)$$

Условия (24) определяют, что число признаков достигло заданного минимума, или ни одна операция из $\mathbf{A}$ не приводит к улучшению L, или достигнуто максимальное число итераций H. Применение многофакторной целевой функции, включающей меры мультиколлинеарности, нелинейной зависимости и сложности модели, позволяет выявлять как явные, так и скрытые избыточности в данных. Благодаря гибкости в выборе порогов и весов, PF^{2}S может быть адаптирован под различные типы данных и задачи.

Применение метода PF^{2}S для признакового пространства IoT

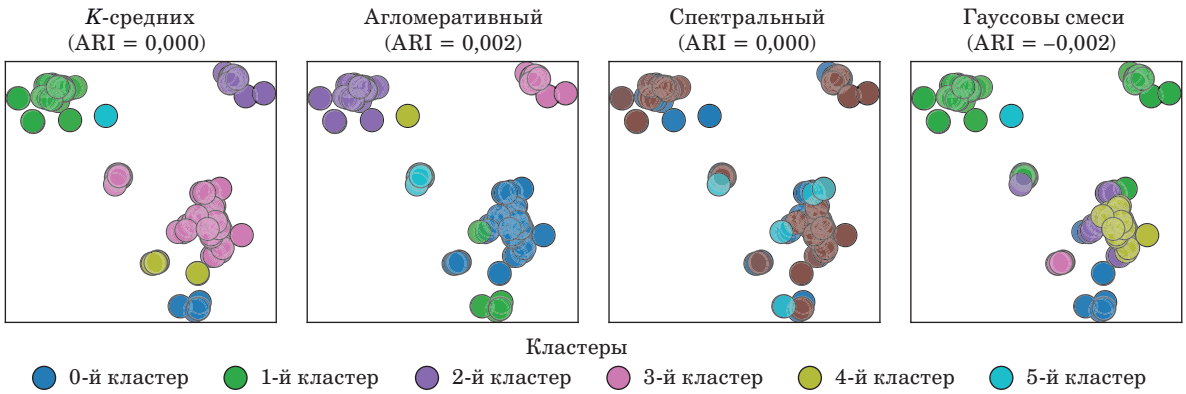
Для оценки эффективности метода PF^{2}S проведено сравнение с двумя базовыми подходами: методом PCA и методом RFE. Метод рекурсивного исключения признаков применялся в двух конфигурациях: с фиксированным числом отбираемых признаков и с порогом объясненной дисперсии 95 %. После сокращения признакового пространства выполнялась кластеризация с использованием четырех методов: K-средних, агломеративного, спектрального и модели гауссовых смесей. Для визуализации результатов применялся метод t-SNE (рис. 3). Стабильность кластеризации оценивалась с помощью скорректированного индекса Рэнда при бутстрэп-повторениях.

Значения ARI, полученные для методов кластеризации после преобразования признакового пространства методами PCA и RFE, не показывают высокой стабильности, что характеризует низкую воспроизводимость кластеров и свидетельствует о неустойчивости результатов при небольших изменениях в данных.

Для сравнения методов PF^{2}S и RFE были рассчитаны ключевые характеристики: время выполнения, итоговая размерность признакового пространства, число обусловленности, наличие коррелирующих и взаимозависимых признаков, среднее значение силуэтного коэффициента и сложность по Рademахеру. Расчет сложности выполнялся для нескольких моделей, для сравнения выбраны значения, полученные на линейной модели со случайными весами, которые позволяют оценить склонность метода к переобучению на шум.

Результаты, приведенные в таблице, показывают, что PF^{2}S обеспечивает более высокое качество кластеризации (силуэтный коэффициент — 0,82) по сравнению с RFE (силуэтный коэффициент — 0,68). Это объясняется тем, что PF^{2}S выполняет поэтапное удаление признаков с контролем стабильности и значимости на каждом шаге, что снижает риск переобучения и повышает воспроизводимость результатов.

Полученное с помощью PF^{2}S признаковое пространство является компактным (17 признаков) и обладает высокой численной устойчивостью (число обусловленности — 2,83), что указывает на отсутствие мультиколлинеарности. В отличие от



■ ***Рис. 3.*** Кластеризация в пространстве RFE
■ ***Fig. 3.*** Clustering in the RFE feature space

- Сравнение PF²S и RFE
- Comparison of PF²S and RFE

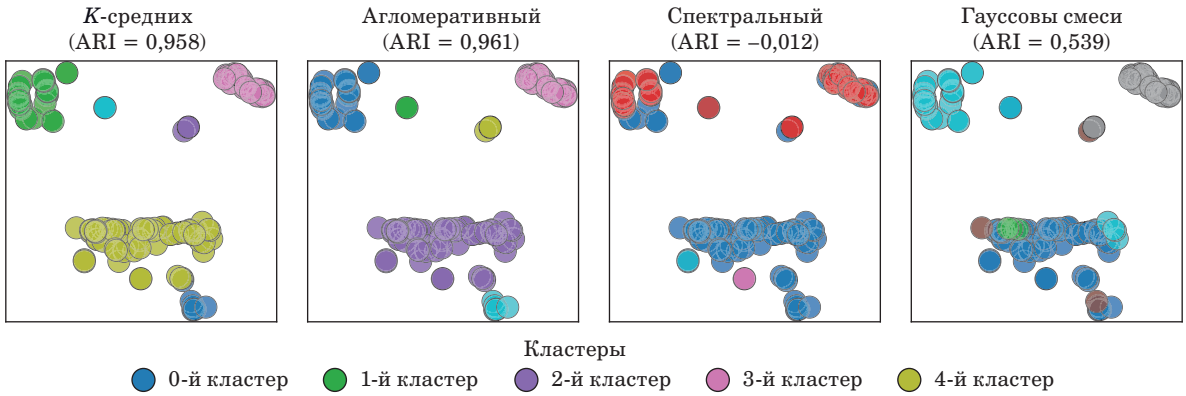
Выполненные действия	Число признаков	Сложность по Радемахеру	Силуэтный коэффициент
Сбор, парсер, загрузка	347	–	–
Предобработка, очистка	275	0,417	0,62
Метод прогрессивной фильтрации признаков Progressive Feature Filtering with Stability and Significance (PF²S)			
1. Коррелирующие, нестабильные	223	0,415	0,62
2. Коррелирующие, незначительные	79	0,233	0,64
3. Мультиколлинеарные, нестабильные	25	0,145	0,72
4. Взаимные, нестабильные	22	0,117	0,77
5. Взаимные, незначительные	**17**	**0,087**	**0,82**
Время выполнения: **6,72 с** (~1000 строк), **58,46 с** (~10 200 строк) Коррелирующие: **0** Взаимозависимые: **0** Число обусловленности: **2,83**			
Метод рекурсивного сокращения размерности (95 % дисперсии) Recursive Feature Elimination (RFE)			
Выбор подмножества признаков RFE	**19**	**0,083**	**0,68**
Время выполнения: **41,42 с** (~1000 строк), **147,42 с** (~10 200 строк) Коррелирующие: **8** Взаимозависимые: **5675,17** Число обусловленности: **75,17**			

него, признаковое пространство, сформированное с помощью RFE, характеризуется высоким числом обусловленности (75,17), свидетельствующим о сильной мультиколлинеарности и потенциальной неустойчивости модели.

К построенному PF²S признаковому пространству были применены методы кластеризации, выполнена визуализация t-SNE и расчет ARI, аналогично предыдущим подходам. Анализ количества кластеров и их структуры показал, что для нового признакового пространства эффективным является разделение данных на пять кластеров (рис. 4).

Для оценки качества кластеризации была выполнена семантическая интерпретация выделенных групп. В признаковом пространстве PF²S остались информативные и устойчивые признаки: минимальный размер пакета, интервалы между пакетами, скорости передачи данных, количество пакетов с флагами и другие, релевантные для анализа сетевого трафика.

Сравнение разбиений на пять и шесть кластеров показало, что 5-кластерная структура является более интерпретируемой: каждый кластер четко соответствует определенному типу сетевой активности. В случае шести кластеров один из них оказывается малочисленным и дублирует другие, что указывает на избыточность разбиения. Выделенные пять кластеров интерпретируются следующим образом: 0-й кластер содержит



- ***Рис. 4.*** Кластеризация в пространстве PF²S
- ***Fig. 4.*** Clustering in the PF²S feature space

одиночные пакеты (SYN, ACK, RST) — сканирование портов и фоновый трафик, в 1-й кластер вошли средние сессии с двусторонним обменом, 2-й кластер объединил высокоскоростные сессии (потоковые передачи), в 3-й кластер выделился трафик с признаками сетевой перегрузки (потенциальные DDoS-атаки), 4-й кластер собрал очень короткие сессии.

Методы *K*-средних и агломеративной кластеризации показали схожие результаты: они эффективно разделили трафик по объему данных, длительности сессий и наличию специфических флагов. Спектральный метод, чувствительный к глобальной структуре данных, выделил редкие и слабо выраженные события (например, сессии с флагом ECE). Метод гауссовых смесей продемонстрировал распределения данных по тем же типам трафика, что и в других методах. Его кластеры имеют четкую структуру и учитывают разделение по вероятностным признакам.

Заключение

Описанный в работе метод PFF-SS (PF^2S) представляет собой систематический подход к сокращению признакового пространства, сочетающий анализ линейных и нелинейных зависимостей с оценкой стабильности и информативности признаков. В отличие от традиционных методов, например метода PCA, новый метод не преобразует исходные признаки, а последовательно исключает избыточные, коррелирующие, мультиколлинеарные и нестабильные компоненты, сохраняя семантическую интерпретируемость оставшегося набора. Это особенно важно в прикладных задачах, таких как анализ сетевого трафика, где физический смысл признаков критичен для интерпретации выделенных паттернов.

В сравнении с методом RFE предложенный подход продемонстрировал существенно лучшие характеристики результирующего признакового пространства. PF^2S обеспечивает более высокое качество кластеризации (видно из расчета силуэтного коэффициента), значительно меньшую сложность модели (по Рademахеру) и формирует численно устойчивое пространство с низким числом обусловленности. Благодаря поэтапному контролю над стабильностью и значимостью признаков на каждом этапе отбора PF^2S снижает риск переобучения и повышает воспроизводимость результатов анализа.

Полученное признаковое пространство компактно и позволяет четко интерпретировать кластеры в соответствии с типами сетевого трафика: фоновые сессии, сканирование портов, веб-запросы, DNS-трафик и признаки сетевой перегрузки. Применение методов *K*-средних и агломеративной кластеризации показало схожие результаты, что подтверждается высокой стабильностью разбиений на бутстрэп-подвыборках (ARI близок к 1,0).

Таким образом, PF^2S представляет собой эффективный, быстрый и интерпретируемый инструмент для подготовки данных в задачах, для которых важны как качество кластеризации, так и понимание природы выделенных событий. Дальнейшее исследование будет направлено на выбор оптимального метода кластеризации для полученного признакового пространства и использование построенных меток для классификации трафика в сетях интернета вещей, а также на адаптацию PF^2S для потоковой обработки данных в реальном времени.

Финансовая поддержка

Работа поддержана Красноярским математическим центром, финансируемым Министерством науки и высшего образования Российской Федерации в рамках мероприятий по созданию и развитию региональных НОМЦ (Соглашение № 075-02-2025-1606).

Литература

1. **Gandomi A., Haider M.** Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 2015, vol. 35, iss. 2, pp. 137–144. doi:10.1016/j.ijinfomgt.2014.10.007
2. **Choudhary A.** Internet of Things: A comprehensive overview, architectures, applications, simulation tools, challenges and future directions. *Discov Internet Thing*, 2024, vol. 4, iss. 31. doi:10.1007/s43926-024-00084-3
3. **Ray P. P.** A survey on Internet of Things architectures. *Journal of King Saud University — Computer and Information Sciences*, 2018, vol. 30, iss. 3, pp. 291–319. doi:10.1016/j.jksuci.2016.10.003
4. **Schiller E., Aidoo A., Fuhrer J., Stahl J., Ziörjen M., Stiller B.** Landscape of IoT security. *Computer Science Review*, 2022, vol. 44, pp. 100467. doi:10.1016/j.cosrev.2022.100467
5. **Татарникова Т. М., Богданов П. Ю.** Обнаружение атак в сетях интернета вещей методами машинного обучения. *Информационно-управляющие системы*, 2021, № 6, с. 42–52. doi:10.31799/1684-8853-2021-6-42-52
6. **Ado A., Hamayadji A., Arouna N. N., Moussa A., Asside D., Ousmane T., Alidou M.** Data collection in IoT networks: Architecture, solutions, protocols and challenges. *IET Wireless Sensor Systems*, 2024, vol. 14, iss. 4, pp. 85–110. doi:14.10.1049/wss2.12080

7. **Исаева О. С., Кулясов Н. В., Исаев С. В.** Инфраструктура сбора данных и имитации угроз безопасности сети интернета вещей. *Сибирский аэрокосмический журнал*, 2025, т. 26, № 1, с. 8–20. doi:10.31772/2712-8970-2025-26-1-8-20, EDN: OPICJJ
8. **Andy S., Rahardjo B., Hanindhito B.** Attack scenarios and security analysis of MQTT communication protocol in IoT system. *4th International Conference on Electrical Engineering, Computer Science and Informatics*, 2017, pp. 1–6. doi:10.1109/EECSI.2017.8239179
9. **Исаева О. С., Исаев С. В., Кулясов Н. В.** Формирование адаптивных рассылок брокера данных интернета вещей. *Информационно-управляющие системы*, 2022, № 5, с. 23–31. doi:10.31799/1684-8853-2022-5-23-31, EDN: DNOSCW
10. **Isaeva O. S., Kulyasov N. V., Isaev S. V.** Semantic modeling of the scheme “Publisher-Subscriber” for data analysis of the Internet of Things. *AIP Conference Proceeding*, 2025, no. 3268, pp. 070025. doi:10.1063/5.0257199
11. **Алексеев А. А., Попова Ю. Б., Шестопалов М. Ю.** Алгоритмы нечеткой кластеризации в задачах диагностики технических систем. *Известия вузов. Северо-Кавказский регион. Серия: Технические науки*, 2012, № 3, с. 3–7. EDN: OYZUBP
12. **Xueyi C.** A comprehensive study of feature selection techniques in machine learning models. *Insights in Computer, Signals and Systems*, 2024, no. 1, pp. 65–78. doi:10.70088/xpf2b276
13. **Omaimah A. H., Andrew S.** Assesing the stability and selection performance of feature selection methods under different data complexity. *The International Arab Journal of Information Technology*, 2022, vol. 19, no. 3A, pp. 442–455. doi:10.34028/iajit/19/3A/4
14. **Liu Y., Mu Y., Chen K., Li Y., Guo J.** Daily activity feature selection in smart homes based on pearson correlation coefficient. *Neural Process Lett*, 2020, vol. 51, pp. 1771–1787. doi:10.1007/s11063-019-10185-8
15. **Czajkowski M., Jurczuk K., Kretowski M.** Steering the interpretability of decision trees using lasso regression — an evolutionary perspective. *Information Sciences*, 2023, vol. 638, pp. 118944. doi:10.1016/j.ins.2023.118944
16. **Kamalov F., Sulieman H., Alzaatreh A., Emarly M., Chamlal H., Safaraliev M.** Mathematical methods in feature selection: A review. *Mathematics*, 2025, no. 13, pp. 996. doi:10.3390/math13060996
17. **Cynthia R., Chaofan C., Zhi C., Haiyang H., Semenova L., Zhong C.** Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 2022, no. 16. doi:10.1214/21-SS133
18. **Priyatno A., Widiyaningtyas T.** A systematic literature review: recursive feature elimination algorithms. *Jurnal Ilmu Pengetahuan dan Teknologi Komputer*, 2024, no. 9. pp. 196–207. doi:10.33480/jitk.v9i2.5015
19. **Barbiero P., Squillero G., Tonda A.** Predictable features elimination: an unsupervised approach to feature selection. *Lecture Notes in Computer Science*, 2022, no. 13163, pp. 399–412. doi:10.1007/978-3-030-95467-3_29
20. **Aker Y.** Comparison of PCA and RFE-RF algorithm in bankruptcy prediction. *Gümüşhane Üniversitesi Sosyal Bilimler Dergisi*, 2022, vol. 13, iss. 3, pp. 1001–1008.
21. **Гайнетдинова А. А., Воробьев А. В.** Сравнение методов отбора значимых признаков для классификации геомагнитных данных. *Прикладная математика и вопросы управления*, 2023, № 4, с. 46–54. doi:10.15593/2499-9873/2023.4.02, EDN: LGWAOR
22. **Исаева О. С., Кулясов Н. В., Исаев С. В.** Создание инструментов сбора данных для анализа аспектов безопасности Интернета вещей. *Информационные и математические технологии в науке и управлении*, 2022, № 3(27), с. 113–125. doi:10.38028/ESI.2022.27.3.011, EDN: UKFFWD
23. **Булыга Ф. С., Курейчик В. М.** Алгоритмы агломеративной кластеризации применительно к задачам анализа лингвистической экспертной информации. *Известия ЮФУ. Технические науки*, 2021, № 6(223), с. 73–88. doi:10.18522/2311-3103-2021-6-73-88, EDN: UVKNNZ
24. **Santos J., Embrechts M.** On the use of the Adjusted Rand Index as a metric for evaluating supervised classification. *Lecture Notes in Computer Science*, 2009, no. 5769, pp. 175–184. doi:10.1007/978-3-642-04277-5_18
25. **Бодров А. О.** Применение метода t-SNE для визуализации и кластеризации многомерных данных. *Сборник трудов IV Международного научно-технического форума «Современные технологии в науке и образовании»*, 2021, т. 6, с. 66–69.
26. **Guoping Z.** A unified definition of mutual information with applications in machine learning. *Mathematical Problems in Engineering*, 2015, no. 201874, pp. 1–12. doi:10.1155/2015/201874
27. **Oneto L., Ghio A., Ridella S., Anguita D.** Local Rademacher Complexity: Sharper risk bounds with and without unlabeled samples. *Neural Networks*, 2015, vol. 65, pp. 115–125. doi:10.1016/j.neunet.2015.02.006
28. **Моисеев Н. А.** Сравнительный анализ эффективности методов устранения мультиколлинеарности. *Учет и статистика*, 2017, № 2 (46), с. 62–77. EDN: ZELDIN

UDC 004.6
doi:10.31799/1684-8853-2025-6-15-27
EDN: ERTCQY

Feature filtering method based on stability and significance criteria

O. S. Isaeva[a], Dr. Sc. Tech., Senior Researcher, orcid.org/0000-0002-5061-6765, isaeva@icm.krasn.ru
[a]Institute of Computational Modelling SB RAS, 50/44, Akademgorodok St., 660036, Krasnoyarsk, Russian Federation

Introduction: Network traffic analysis in the Internet of Things (IoT) is complicated by high dimensionality, feature redundancy, and instability. Strong correlation, multicollinearity, and noise degrade clustering quality and hinder interpretation. Moreover, legitimate and anomalous traffic often overlap, making it difficult to formalize class boundaries. Therefore, a feature selection method that ensures stability, compactness, and semantic interpretability is required. **Purpose:** To develop and experimentally evaluate a new method for constructing a stable and interpretable feature space in network traffic clustering tasks — Progressive Feature Filtering with Stability and Significance (PFF-SS, PF^{2}S). **Methods:** We describe a step-by-step PF^{2}S algorithm that combines analysis of linear dependencies (correlation, VIF) and nonlinear dependencies (mutual information) with assessment of feature stability and significance. At each stage, redundant, weakly significant, or unstable features are removed. **Results:** Applying PF^{2}S to an IoT network traffic dataset has reduced the number of features from over 300 to 17 while preserving high informativeness. The comparison with feature spaces reduced by Principal Component Analysis (PCA) and Recursive Feature Elimination shows that PF^{2}S achieves higher metrics in stability, interpretability, and clustering quality. Unlike Principal Component Analysis, PF^{2}S does not transform features but preserves their original semantics. Compared to Recursive Feature Elimination, PF^{2}S eliminates multicollinearity, reduces model complexity, and achieves a silhouette coefficient 17.6% higher. Clusters built on the PF2S-derived feature space are stable (high Adjusted Rand Index) and semantically interpretable. **Practical relevance:** PF^{2}S produces a compact and robust feature space suitable for anomaly detection systems in IoT network traffic. **Discussion:** Promising directions include adapting PF^{2}S for streaming data processing and integrating it with signature-based anomaly detection methods and network traffic ontologies.

Keywords — Internet of Things, feature stability, feature significance, *K*-means clustering, agglomerative clustering, spectral clustering, Gaussian Mixture Model, Principal Component Analysis, Recursive Feature Elimination, network traffic analysis, anomaly detection.

For citation: Isaeva O. S**.** Feature filtering method based on stability and significance criteria. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2025, no. 6, pp. 15–27 (In Russian). doi:10.31799/1684-8853-2025-6-15-27, EDN: ERTCQY

Financial support

This work is supported by the Krasnoyarsk Mathematical Center and financed by the Ministry of Science and Higher Education of the Russian Federation in the framework of the establishment and development of regional Centers for Mathematics Research and Education (Agreement No. 075-02-2025-1606).

References

1. Gandomi A., Haider M. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 2015, vol. 35, iss. 2, pp. 137–144. doi:10.1016/j.ijinfomgt.2014.10.007
2. Choudhary A. Internet of Things: A comprehensive overview, architectures, applications, simulation tools, challenges and future directions. *Discov Internet Thing*, 2024, vol. 4, iss. 31. doi:10.1007/s43926-024-00084-3
3. Ray P. P. A survey on Internet of Things architectures. *Journal of King Saud University — Computer and Information Sciences*, 2018, vol. 30, iss. 3, pp. 291–319. doi:10.1016/j.jksuci.2016.10.003
4. Schiller E., Aidoo A., Fuhrer J., Stahl J., Ziörjen M., Stiller B. Landscape of IoT security. *Computer Science Review*, 2022, vol. 44, pp. 100467. doi:10.1016/j.cosrev.2022.100467
5. Tatarnikova T. M., Bogdanov P. Yu. Intrusion detection in internet of things networks based on machine learning methods. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2021, no. 6, pp. 42–52 (In Russian). doi:10.31799/1684- 8853-2021-6-42-52
6. Ado A., Hamayadji A., Arouna N. N., Moussa A., Asside D., Ousmane T., Alidou M. Data collection in IoT networks: Architecture, solutions, protocols and challenges. *IET Wireless Sensor Systems*, 2024, vol. 14, iss. 4, pp. 85–110. doi:14.10.1049/wss2.12080
7. Isaeva O. S., Kulyasov N. V., Isaev S. V. Infrastructure for collecting data and simulating security threats in the Internet of Things network. *Siberian Aerospace Journal*, 2025, vol. 26, no. 1, pp. 8–20 (In Russian). doi:10.31772/2712-8970-2025-26-1-8-20, EDN: OPICJJ
8. Andy S., Rahardjo B., Hanindhito B. Attack scenarios and security analysis of MQTT communication protocol in IoT system. *4th International Conference on Electrical Engineering, Computer Science and Informatics*, 2017, pp. 1–6. doi:10.1109/EECSI. 2017.8239179
9. Isaeva O. S., Isaev S. V., Kulyasov N. V. Formation of adaptive publications from the Internet of Things data broker. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2022, no. 5, pp. 23–31 (In Russian). doi:10.31799/1684-8853-2022-5-23-31, EDN: DNOSCW
10. Isaeva O. S., Kulyasov N. V., Isaev S. V. Semantic modeling of the scheme “Publisher-Subscriber” for data analysis of the Internet of Things. *AIP Conference Proceeding*, 2025, no. 3268, pp. 070025. doi:10.1063/5.0257199
11. Alekseev A. A., Popova Yu. B., Shestopalov M. Yu. Algorithms fuzzy clustering algorithms in technical systems diagnostics problems. *Bulletin of Higher Educational Institutions. North Caucasus Region. Technical Sciences*, 2012, no. 3, pp. 3–7 (In Russian). EDN: OYZUBP
12. Xueyi C. A comprehensive study of feature selection techniques in machine learning models. *Insights in Computer, Signals and Systems*, 2024, no. 1, pp. 65–78. doi:10.70088/xpf2b276
13. Omaimah A. H., Andrew S. Assesing the stability and selection performance of feature selection methods under different data complexity. *The International Arab Journal of Information Technology*, 2022, vol. 19, no. 3A, pp. 442–455. doi:10.34028/iajit/19/3A/4
14. Liu Y., Mu Y., Chen K., Li Y., Guo J. Daily activity feature selection in smart homes based on pearson correlation coefficient. *Neural Process Lett*, 2020, vol. 51, pp. 1771–1787. doi:10.1007/s11063-019-10185-8
15. Czajkowski M., Jurczuk K., Kretowski M. Steering the interpretability of decision trees using lasso regression — an evolutionary perspective. *Information Sciences*, 2023, vol. 638, pp. 118944. doi:10.1016/j.ins.2023.118944
16. Kamalov F., Sulieman H., Alzaatreh A., Emarly M., Chamlal H., Safaraliev M. Mathematical methods in feature selection: A review. *Mathematics*, 2025, no. 13, pp. 996. doi:10.3390/math13060996
17. Cynthia R., Chaofan C., Zhi C., Haiyang H., Semenova L., Zhong C. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 2022, no. 16. doi:10.1214/21-SS133
18. Priyatno A., Widiyaningtyas T. A systematic literature review: recursive feature elimination algorithms. *Jurnal Ilmu*

Pengetahuan dan Teknologi Komputer, 2024, no. 9. pp. 196–207. doi:10.33480/jitk.v9i2.5015
19. Barbiero P., Squillero G., Tonda A. Predictable features elimination: an unsupervised approach to feature selection. *Lecture Notes in Computer Science*, 2022, no. 13163, pp. 399–412. doi:10.1007/978-3-030-95467-3_29
20. Aker Y. Comparison of PCA and RFE-RF algorithm in bankruptcy prediction. *Gümüşhane Üniversitesi Sosyal Bilimler Dergisi*, 2022, vol. 13, iss. 3, pp. 1001–1008.
21. Gainetdinova A. A., Vorobev A. V. Comparison of features elimination methods for geomagnetic data classification. *Applied Mathematics and Control Sciences*, 2023, no. 4, pp. 46–54 (In Russian). doi:10.15593/2499-9873/2023.4.02, EDN: LGWAOR
22. Isaeva O. S., Kulyasov N. V., Isaev S. V. Creating data collection tools to analyze security aspects Internet of Things. *Information and Mathematical Technologies in Science and Management*, 2022, no. 3(27), pp. 113–125 (In Russian). doi:10.38028/ESI.2022.27.3.011, EDN: UKFFWD
23. Bulyga Ph. S., Kureichik V. M. Agglomerative clusterization algorithms for the problems of analysis of linguistic expert information. *Izvestiya SFedU. Engineering Sciences*, 2021, no. 6(223), pp. 73–88 (In Russian). doi:10.18522/2311-3103-2021-6-73-88, EDN: UVKNNZ
24. Santos J., Embrechts M. On the use of the Adjusted Rand Index as a metric for evaluating supervised classification. *Lecture Notes in Computer Science*, 2009, no. 5769, pp. 175–184. doi:10.1007/978-3-642-04277-5_18
25. Bodrov A. O. Application of the t-SNE method for visualization and clustering of multidimensional data. *Sbornik trudov IV Mezhdunarodnogo nauchno-tekhnicheskogo foruma "Sovremennye tekhnologii v nauke i obrazovanii"*. [Proceedings of the IV International Scientific and Technical Forum "Modern Technologies in Science and Education"], 2021, no. 6, pp. 66–69 (In Russian).
26. Guoping Z. A unified definition of mutual information with applications in machine learning. *Mathematical Problems in Engineering*, 2015, no. 201874, pp. 1–12. doi:10.1155/2015/201874
27. Oneto L., Ghio A., Ridella S., Anguita D. Local Rademacher Complexity: Sharper risk bounds with and without unlabeled samples. *Neural Networks*, 2015, vol. 65, pp. 115–125. doi:10.1016/j.neunet.2015.02.006
28. Moiseev N. A. Comparative analysis of the effectiveness of methods for eliminating multicollinearity. *Uchet i statistika*, 2017, no. 2 (46), pp. 62–77 (In Russian). EDN: ZELDIN

УВАЖАЕМЫЕ АВТОРЫ!

Научные базы данных, включая Scopus и Web of Science, обрабатывают данные автоматически. С одной стороны, это ускоряет процесс обработки данных, с другой — различия в транслитерации ФИО, неточные данные о месте работы, области научного знания и т. д. приводят к тому, что в базах оказывается несколько авторских страниц для одного и того же человека. В результате для всех по отдельности считаются индексы цитирования, что снижает рейтинг ученого.

Для идентификации авторов в сетях Thomson Reuters проводит регистрацию с присвоением уникального индекса (ID) для каждого из авторов научных публикаций.

Процедура получения ID бесплатна и очень проста, есть возможность провести регистрацию на 12 языках, включая русский (чтобы выбрать язык, кликните на зеленое поле вверху справа на стартовой странице): https://orcid.org